BCU BCA 5TH SEM — DATA ANALYTICS (DCCA512) SOLVED QUESTION BANK

Solved Question Bank by Nikhil P

PART I: SECTION A — 2 MARKS QUESTIONS (PAPERS 1 TO 5)

MODEL QUESTION PAPER – 1

Q1. Define data analytics.

Data Analytics is the systematic computational process of inspecting, cleaning, transforming, and modeling raw data to discover meaningful patterns, draw conclusions, extract actionable insights, and support informed decision-making. It utilizes statistical, mathematical, and algorithmic techniques to solve business and research problems.

Key Phases: Data Collection → Data Preprocessing → Statistical/Algorithmic Modeling → Insight Generation.

Q2. What is descriptive analytics?

Descriptive Analytics is the preliminary phase of data analysis that answers the fundamental question: **"What has happened?"** It summarizes and interprets historical (past) data using statistical measures (such as mean, median, mode, and standard deviation) and data visualization tools (charts, graphs, dashboards) to identify trends, patterns, and anomalies.

Core Objective: Summarize historical events without making predictions or suggesting actions.

Q3. What is text analytics?

Text Analytics (also known as *Text Mining*) is the automated process of converting large volumes of **unstructured textual data** into quantitative, structured data to discover patterns, trends, and relationships. It leverages **Natural Language Processing (NLP)**, computational linguistics, and machine learning for tasks like information extraction, categorization, and sentiment analysis.

Key Applications: Sentiment analysis of customer reviews, spam email filtering, social media opinion mining.

Q4. What is meant by Correlation?

Correlation is a bivariate statistical metric that quantifies the **degree (strength) and direction of a linear relationship** between two continuous variables (X and Y). It is most commonly measured using the **Pearson Correlation Coefficient (r)**, which is bounded between -1 and $+1$:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

Interpretation:

- $r = +1$: Perfect positive linear relationship.
- $r = -1$: Perfect negative linear relationship.
- $r = 0$: No linear relationship.

(Note: Correlation implies association, not causation.)

Q5. What is Regression?

Regression is a supervised statistical modeling technique used to evaluate and describe the mathematical relationship between a dependent variable (Y, target/response) and one or more **independent variables** (X, predictors/features). Its primary objectives are **prediction, forecasting, and inferring causal effects**.

Simple Linear Regression Equation:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

(where β_0 is the Y-intercept, β_1 is the slope coefficient, and ϵ is the random error term).

Q6. Define population and sample space.

- **Population (N):** In statistics, a population represents the complete set of all individuals, items, observations, or measurements possessing common observable characteristics that are of interest in a statistical study.
- **Sample Space (S or Ω):** In probability theory, the sample space is the exhaustive set of all possible outcomes of a random experiment.
  Example: For tossing a fair coin once, $S = \{\text{Heads}, \text{Tails}\}$.

Concept	Domain	Key Characteristic
Population	Statistics	Totality of all subjects/units under investigation (size N).
Sample Space	Probability	Set of all theoretical possible outcomes of an experiment (S).

Q7. Define conditional probability.

Conditional Probability is the probability of an event A occurring given that another event B has already occurred. It is denoted as $P(A \mid B)$ and is mathematically defined as the ratio of the joint probability of both events to the marginal probability of the conditioning event, provided $P(B) > 0$:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{where } P(B) > 0$$

Property: If events A and B are independent, then $P(A \mid B) = P(A)$.

Q8. Define Business Intelligence (BI).

Business Intelligence (BI) is an integrated, technology-driven framework and infrastructure that collects, integrates, cleanses, stores, and analyzes raw business data to transform it into actionable strategic and operational insights. It encompasses **ETL processes, data warehousing, OLAP (Online Analytical Processing), reporting, and interactive dashboards** to support data-driven decision-making.

Key Deliverables: Executive summary reports, KPI tracking cards, real-time interactive dashboards.

Q9. What is Power BI?

Power BI is a cloud-based suite of business analytics and data visualization software tools developed by **Microsoft**. It connects to hundreds of heterogeneous data sources, transforms and models data, and generates interactive reports and 360-degree dashboards that can be published and shared securely across an organization.

Key Architecture Components:

1. **Power BI Desktop:** Windows authoring tool for data modeling and report design.
2. **Power BI Service:** Cloud-based SaaS platform for report distribution and collaboration.
3. **Power BI Mobile:** Mobile client application for monitoring KPIs on portable devices.

Q10. What is Power Query?

Power Query is an advanced **ETL (Extract, Transform, and Load)** data preparation and data connectivity engine integrated within Microsoft Excel, Power BI, and Analysis Services. It provides an intuitive graphical interface alongside a functional, case-sensitive scripting language called **M code** to ingest data from diverse sources, clean, reshape, pivot, and merge datasets before loading them into the data model.

Key Functions: Removing duplicates, splitting columns, filtering nulls, merging/appending tables, and unpivoting data without altering the original source data.

MODEL QUESTION PAPER – 2

Q1. Mention the types of data analytics.

Data analytics is broadly categorized into four progressive types based on analytical objectives and complexity:

1. **Descriptive Analytics:** Explores and summarizes historical data to answer "*What happened?*" (e.g., standard summary reports, KPI cards).
2. **Diagnostic Analytics:** Examines data anomalies and relationships to answer "*Why did it happen?*" (e.g., drill-down analysis, root cause discovery).
3. **Predictive Analytics:** Employs statistical models, machine learning, and trend forecasting to answer "*What is likely to happen?*" (e.g., regression forecasting, demand prediction).
4. **Prescriptive Analytics:** Utilizes optimization algorithms and decision simulation engines to recommend actions answering "*What should we do?*" (e.g., dynamic price optimization, route optimization).

Q2. Define web analytics.

Web Analytics is the measurement, collection, analysis, and reporting of internet and website data for the purpose of understanding, evaluating, and optimizing web page usage and visitor behavior. It tracks visitor traffic sources, navigational clickstreams, bounce rates, session durations, and conversion funnels to improve website usability, enhance user experience (UX), and drive digital marketing strategies.

Key Metrics: Unique Visitors, Pageviews, Bounce Rate, Average Session Duration, Conversion Rate.

Q3. What is Positive and Negative Correlation?

Correlation quantifies the direction and strength of the linear association between two variables (X and Y):

- **Positive Correlation ($r > 0$):** Occurs when both variables move in the same direction. As variable X increases, variable Y also increases; as X decreases, Y also decreases (e.g., study hours and exam performance; height and weight). Perfect positive correlation is denoted by $r = +1$.
- **Negative Correlation ($r < 0$):** Occurs when two variables move in opposite directions. As variable X increases, variable Y decreases (e.g., driving speed and remaining travel time; interest rates and borrowing demand). Perfect negative correlation is denoted by $r = -1$.

Q4. What is a Regression Coefficient?

A **Regression Coefficient** is a statistical parameter in a regression equation that measures the expected average rate of change in the dependent variable per one-unit change in the independent variable.

In bivariate linear regression, there are two regression coefficients:

1. **Regression coefficient of Y on X (b_{yx}):** Slope of the regression line of Y on X:

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{\text{Cov}(X,Y)}{\sigma_x^2}$$

2. **Regression coefficient of X on Y (b_{xy}):** Slope of the regression line of X on Y:

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2} = \frac{\text{Cov}(X,Y)}{\sigma_y^2}$$

Fundamental Property: The correlation coefficient (r) is the geometric mean of the two regression coefficients:

$$r = \pm\sqrt{b_{yx} \cdot b_{xy}}$$

(Both regression coefficients and r must always share the same algebraic sign).

Q5. Distinguish between Correlation and Regression.

Basis of Comparison	Correlation	Regression
Meaning & Objective	Measures the degree and direction of linear association between two variables.	Models and predicts the mathematical/functional dependence of one variable on another.
Variable Nature	Symmetrical: Both variables are treated equally without dependent/independent classification ($r_{xy} = r_{yx}$).	Asymmetrical: Distinguishes between a dependent (response) variable Y and independent (predictor) variable X.
Cause & Effect	Does **not** indicate or establish a cause-and-effect relationship.	Attempts to establish a functional relationship for forecasting and causal inference.
Numerical Bound	Bounded: $-1 \leq r \leq +1$.	Unbounded: $-\infty < b < +\infty$.
Units of Measure	Dimensionless / relative measure (independent of units of scale and origin).	Absolute measure expressed in physical units of Y per unit of X.

Q6. Define a random variable.

A **Random Variable** (X) is a real-valued mathematical function defined on the sample space (S) of a random experiment that assigns a unique real number to every possible outcome in S:

$$X : S \to \mathbb{R}$$

Types of Random Variables:

1. **Discrete Random Variable:** Takes on isolated, countable finite or countably infinite integer values (e.g., number of defective items in a batch; characterized by a Probability Mass Function, PMF).

2. **Continuous Random Variable:** Takes on any real value within an unbroken continuous interval (e.g., height, temperature, execution time; characterized by a Probability Density Function, PDF).

Q7. Define Binomial Distribution.

A **Binomial Distribution** is a discrete probability distribution that models the probability of obtaining exactly k successes in n independent and identical Bernoulli trials, with a constant probability of success p in each trial and probability of failure $q = (1 - p)$.

Probability Mass Function (PMF):

$$P(X = k) = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, 2, \ldots, n$$

Key Parameters & Properties:

- **Parameters:** n (number of trials) and p (probability of success).
- **Mean:** $\mu = np$
- **Variance:** $\sigma^2 = npq$
- **Condition:** $\text{Mean} > \text{Variance}$ (since $q < 1$).

Q8. Distinguish between point estimation and interval estimation.

Feature	Point Estimation	Interval Estimation
Definition	Calculates a **single numerical value** from sample data as the best estimate of an unknown population parameter.	Constructs a **numerical range/interval** bounded by lower and upper limits expected to contain the true parameter.
Mathematical Form	A single statistic $\hat{\theta}$ (e.g., sample mean $\bar{x} = 50$ estimating population mean μ).	An interval $[L, U]$ (e.g., 95% Confidence Interval: $45 \leq \mu \leq 55$).
Confidence Level	Does **not** incorporate a level of confidence or probability statement.	Explicitly includes a confidence coefficient $(1 - \alpha)$, typically 90%, 95%, or 99%.
Margin of Error	Provides no indication of sampling error or uncertainty.	Explicitly accounts for standard error and margin of error ($\bar{x} \pm Z \cdot \frac{\sigma}{\sqrt{n}}$).
Reliability	Low: The probability of a continuous parameter matching a single point estimate exactly is virtually zero.	High: Provides an informative probabilistic guarantee regarding parameter enclosure.

Q9. What is Power BI Desktop?

Power BI Desktop is a free, standalone Windows-based client application developed by Microsoft that serves as the primary authoring, modeling, and report creation environment in the Power BI architecture. It allows data analysts to connect to diverse data sources, cleanse and shape data, create complex relational models, and design interactive data visualizations.

Core Integrated Components:

1. **Power Query:** Data extraction, ingestion, and transformation (ETL) using M code.
2. **Data Modeling (Power Pivot / VertiPaq Engine):** Defining relationships, star schemas, hierarchies, and calculated measures using DAX (Data Analysis Expressions).

3. **Report View:** Canvas for creating interactive multi-page visual reports, charts, and dashboards prior to publishing to Power BI Service.

Q10. What is Drill through features in Power BI?

The **Drillthrough** feature in Power BI is an advanced interactive report navigation capability that enables report consumers to right-click on a specific visual data element (such as a category, customer, product, or geographic region) on a summary page and navigate to a focused, dedicated destination report page that is automatically filtered to that selected entity.

Working Mechanism: The destination page includes a defined **Drillthrough field**; when initiated, the filter context of the selected data point is passed directly to the target page without requiring manual slicer adjustments.

Key Advantages:

- Keeps high-level executive summary dashboards clean and uncluttered.
- Allows end-users to effortlessly transition from macro-level KPI summaries down to micro-level, transactional details with a single click.

MODEL QUESTION PAPER – 3

Q1. What is predictive analytics?

Predictive Analytics is an advanced branch of data analytics that utilizes historical data, statistical algorithms, machine learning techniques, and predictive modeling to estimate the likelihood of future outcomes, events, and trends.

Core Question: Answers "*What is likely to happen?*"

Key Techniques: Linear and logistic regression, time-series forecasting, decision trees, neural networks, and pattern recognition.

Common Applications: Credit scoring, customer churn prediction, disease outbreak forecasting, and predictive equipment maintenance.

Q2. Name the three key stages in data analytics.

The three fundamental, sequential stages in a standard data analytics lifecycle are:

1. **Data Ingestion & Preparation (Data Preprocessing):** Gathering raw data from heterogeneous sources, handling missing/null values, filtering outliers, and transforming/normalizing data into a structured schema.
2. **Data Analysis & Modeling (Data Exploration & Processing):** Applying exploratory data analysis (EDA), statistical methods, mathematical formulas, and machine learning algorithms to discover patterns, correlations, and relationships.
3. **Data Interpretation & Visualization (Insight Delivery):** Translating analytical results into actionable business insights using dashboards, charts, and executive reports to support strategic decision-making.

Q3. Mention any four challenges in data analytics.

The four prominent challenges encountered in data analytics are:

1. **Data Quality & Inconsistency:** Dealing with noisy, incomplete, duplicate, unstandardized, or missing data across disparate organizational data silos.
2. **Data Privacy, Governance & Security:** Complying with regulatory standards (e.g., GDPR, HIPAA) while protecting confidential enterprise and user data from security breaches.

3. **Data Volume and Scalability (Big Data Complexity):** Handling rapid scaling, high storage overhead, and computationally intensive processing of massive volumes of structured and unstructured data.
4. **Shortage of Skilled Talent & Tool Integration:** Scarcity of professionals skilled in statistics, domain knowledge, and modern analytics tooling, combined with the difficulty of integrating legacy enterprise systems with modern BI platforms.

Q4. What is Probable Error of the coefficient of correlation?

Probable Error (P.E.) is a statistical measure used to test the reliability, sampling stability, and statistical significance of the calculated Pearson correlation coefficient (r) in a sample of size N.

Mathematical Formula:

$$P.E. = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$$

(where r is the correlation coefficient, N is the total number of pairs of observations, and $0.6745 \approx \frac{2}{3}$ of the standard error).

Rules for Interpretation & Significance:

- If $|r| < \mathbf{P.E.}$, there is no evidence of correlation (the correlation is statistically insignificant).
- If $|r| > 6 \times \mathbf{P.E.}$, the correlation is definitely significant and reliable.
- The limits of the true population correlation coefficient are given by: $r \pm \mathbf{P.E.}$

Q5. What is a Coefficient of Correlation? State its range.

The **Coefficient of Correlation** (most commonly Pearson's product-moment correlation coefficient, denoted by r) is a dimensionless, normalized numerical measure that quantifies the **degree, strength, and direction of a linear relationship** between two quantitative continuous variables.

Numerical Range: It is strictly bounded within the closed interval:

$$-1 \leq r \leq +1$$

Significance of Range Boundaries:

- $r = +1$: Perfect positive linear correlation.
- $r = -1$: Perfect negative linear correlation.
- $r = 0$: No linear correlation between the two variables.

Q6. Distinguish between independent and dependent events.

Basis of Comparison	Independent Events	Dependent Events
Definition	The occurrence of one event has **no effect** on the probability of occurrence of the other event.	The occurrence of one event **directly affects or changes** the probability of occurrence of the subsequent event.
Probability Condition	$P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$	$P(A \mid B) \neq P(A)$ and $P(B \mid A) \neq P(B)$
Multiplication Rule	$P(A \cap B) = P(A) \cdot P(B)$	$P(A \cap B) = P(A) \cdot P(B \mid A)$
Example	Tossing a fair coin twice (the outcome of the second toss is unaffected by the first).	Drawing two playing cards consecutively from a standard deck **without replacement**.

Q7. Define hypothesis testing.

Hypothesis Testing is a formal, inferential statistical procedure used to determine whether there is sufficient empirical evidence in a sample of data to infer that a certain condition holds true for the entire target population.

Core Procedure: It involves formulating two opposing statistical hypotheses (Null H_0 and Alternative H_1), selecting an appropriate significance level (α), computing a test statistic (z, t, F, or χ^2), and comparing the resulting test statistic / p-value against a critical value to decide whether to **reject** H_0 in favor of H_1.

Q8. Distinguish between null and alternative hypotheses.

Parameter / Criterion	Null Hypothesis (H_0)	Alternative Hypothesis (H_1 or H_a)
Definition & Claim	A baseline statement of **no effect, no difference, or no relationship** between variables.	A statement that claims the presence of an **actual effect, significant difference, or relationship**.
Mathematical Nature	Always contains an equality sign ($=$, $\leq$, or $\geq$).	Never contains an equality sign ($\neq$, $<$, or $>$).
Testing Objective	The hypothesis that is directly subjected to statistical testing to be either rejected or failed to be rejected.	The research hypothesis accepted when the null hypothesis is rejected based on statistical evidence.
Burden of Proof	Assumed true until proven false beyond a reasonable significance threshold (innocent until proven guilty).	Requires convincing sample evidence to be accepted.
Example	$H_0 : \mu = 50$ (A new drug has no effect on blood pressure).	$H_1 : \mu \neq 50$ or $\mu < 50$ (A new drug reduces blood pressure).

Q9. What is Power BI Service?

Power BI Service (often referred to as Power BI Online) is a secure, cloud-based **Software-as-a-Service (SaaS)** platform hosted by Microsoft. It serves as the enterprise collaboration, distribution, and consumption layer of the Power BI ecosystem.

Core Capabilities:

1. **Sharing & Collaboration:** Publishing reports created in Power BI Desktop to cloud workspaces, sharing dashboards, and managing workspace access permissions.
2. **Automated Scheduled Refresh:** Connecting to on-premises and cloud data sources via On-premises Data Gateways to ensure data models stay up to date.

3. **Enterprise Dashboarding & Alerts:** Pinning visual tiles from various reports into centralized executive dashboards and configuring automated threshold alerts.

Q10. What is Power Query Editor?

Power Query Editor is the dedicated graphical workspace window and data transformation interface embedded within Microsoft Power BI Desktop and Excel. It serves as a visual **ETL (Extract, Transform, and Load)** workbench where users connect to disparate data sources, inspect schema profiles, and transform dirty, unstructured data into clean tabular formats without modifying the underlying source files.

Core Functionality: Column renaming, filtering rows, removing duplicates, splitting/merging columns, data type conversions, pivoting/unpivoting, and joining/appending tables.

Underlying Engine: Every transformation applied via the ribbon UI is recorded sequentially as a step in Microsoft's functional query language, **M code**, enabling fully automated, reproducible data refreshes.

MODEL QUESTION PAPER – 4

Q1. What is prescriptive analytics?

Prescriptive Analytics is the final, most sophisticated frontier of data analytics that recommends specific, optimal courses of action to achieve desired business outcomes and demonstrates the likely implications of each decision.

Core Question: Answers "*What should we do?*" and "*How can we make it happen?*"

Underlying Techniques: Mathematical optimization, linear programming, simulation modeling, multi-criteria decision analysis (MCDA), and AI decision algorithms.

Examples: Dynamic airline and ride-sharing pricing algorithms, automated supply chain routing, and autonomous vehicle navigation.

Q2. List any four features of data analytics.

The four fundamental features of data analytics are:

1. **Data-Driven Decision Making:** Transforms subjective intuition into objective, empirical evidence to support strategic enterprise decisions.
2. **Pattern & Trend Discovery:** Employs statistical algorithms to automatically identify underlying correlations, seasonal trends, and behavioral anomalies across large datasets.
3. **Scalability & Data Integration:** Capable of ingesting, cleaning, and synthesizing disparate, heterogeneous data types (structured, semi-structured, and unstructured) from diverse sources at scale.
4. **Predictive & Proactive Capability:** Enables forward-looking forecasting and optimization rather than solely relying on historical retrospective reporting.

Q3. If $b_{xy} = 0.8$ and $b_{yx} = 0.6$, find r.

Given:

- Regression coefficient of X on Y: $b_{xy} = 0.8$
- Regression coefficient of Y on X: $b_{yx} = 0.6$

Formula: The Pearson correlation coefficient (r) is the geometric mean of the two regression coefficients:

$$r = \pm\sqrt{b_{xy} \cdot b_{yx}}$$

Step-by-Step Calculation:

$$r = \sqrt{0.8 \times 0.6} = \sqrt{0.48} \approx +0.6928$$

Sign Rule: Since both regression coefficients (b_{xy} and b_{yx}) are positive, the correlation coefficient r must also be **positive**.

Conclusion: $r \approx +0.6928$ (indicating a moderate-to-strong positive linear correlation).

Q4. What is Simple Correlation?

Simple Correlation is a statistical measure that studies and quantifies the mutual relationship, co-variation, and association between **only two variables** at a time (one independent variable and one dependent variable, e.g., height and weight, or advertising expenditure and sales revenue).

Key Distinction: Unlike *Multiple Correlation* (which studies three or more variables simultaneously) or *Partial Correlation* (which studies two variables while controlling the effect of others), simple correlation isolates only the pairwise association between two variables.

Metric: Evaluated using the Pearson correlation coefficient (r) or Spearman rank correlation (ρ).

Q5. What is Spearman's Rank Correlation Coefficient?

Spearman's Rank Correlation Coefficient (denoted by ρ or r_s) is a non-parametric statistical measure that quantifies the strength and direction of monotonic association between two ranked or ordinal variables.

Mathematical Formula (without ties):

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

(where $d_i = R_{x_i} - R_{y_i}$ is the difference between ranks assigned to each pair of observations, and n is the total number of paired observations).

Range & Interpretation: $-1 \leq \rho \leq +1$

- $\rho = +1$: Perfect agreement in rankings.
- $\rho = -1$: Perfect disagreement (complete inversion) in rankings.
- $\rho = 0$: No association between ranks.

Advantage: Ideal for qualitative attributes (e.g., beauty contests, honesty, intelligence) where variables cannot be measured quantitatively but can be ranked.

Q6. Define Poisson Distribution.

A **Poisson Distribution** is a discrete probability distribution that expresses the probability of a given number of rare events (k) occurring within a fixed interval of time, area, distance, or volume, given that these events occur independently at a known constant average rate (λ).

Probability Mass Function (PMF):

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \text{for } k = 0, 1, 2, \dots$$

(where $\lambda > 0$ is the average event rate, $e \approx 2.71828$, and $k!$ is the factorial of k).

Fundamental Property:

$$\text{Mean} = \text{Variance} = \lambda$$

(The equality of mean and variance is the unique characterizing property of a Poisson distribution).

Q7. State any four properties of Normal Distribution.

Four essential mathematical and graphical properties of a **Normal (Gaussian) Distribution** are:

1. **Symmetry & Bell-Shaped Curve:** The probability curve is perfectly symmetrical and bell-shaped about the vertical axis passing through the mean (μ), with zero skewness ($\beta_1 = 0$) and mesokurtic kurtosis ($\beta_2 = 3$).
2. **Coincidence of Central Tendencies:** The mean, median, and mode of the distribution are identical and coincide at the exact center:

$$\text{Mean} = \text{Median} = \text{Mode} = \mu$$

3. **Total Area Under the Curve:** The total area bounded by the normal curve and the horizontal axis is exactly equal to unity (1 or 100%), with 0.5 area lying on each side of the mean.
4. **Empirical Rule (Three-Sigma Limits):**
   - $\mu \pm 1\sigma$ encloses 68.27% of all observations.
   - $\mu \pm 2\sigma$ encloses 95.45% of all observations.
   - $\mu \pm 3\sigma$ encloses 99.73% of all observations.

Q8. What is ANOVA?

ANOVA (Analysis of Variance) is an inferential statistical technique developed by Sir R. A. Fisher used to test whether there are statistically significant differences among the means of **three or more independent groups** simultaneously, without inflating the Type I error rate.

Underlying Principle: It partitions the total variation observed in the dataset into two components:

1. **Between-Group Variance (SSB):** Variation attributable to differences between the treatment groups.
2. **Within-Group Variance (SSW / Error Variance):** Random variation or error within individual groups.

Test Statistic (F-ratio):

$$F = \frac{\text{Mean Square Between (MSB)}}{\text{Mean Square Within (MSW)}} = \frac{SSB / df_b}{SSW / df_w}$$

If calculated $F > F_{\text{critical}}$, the null hypothesis ($H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$) is rejected.

Q9. Define a data source in Power BI.

In Power BI, a **Data Source** refers to any external storage system, file, database, or online service from which Power BI ingests, extracts, or queries data to construct analytical models, reports, and dashboards. Power BI connects via built-in connector drivers through **Import Mode**, **DirectQuery**, or **Live Connection**.

Categories of Data Sources:

1. **File-based Sources:** Microsoft Excel workbooks (`.xlsx`), CSV/Text files, XML, JSON, and PDF files.
2. **Relational & Enterprise Databases:** Microsoft SQL Server, Oracle, MySQL, PostgreSQL, IBM DB2, and SAP HANA.
3. **Cloud & Web Services:** Azure SQL Database, Google BigQuery, SharePoint Online lists, Salesforce, Web APIs (REST/OData), and Google Analytics.

Q10. What are drill-through and bookmarks in Power BI?

Both are powerful interactive user-navigation features in Power BI that enhance user experience, but they serve distinctly different purposes:

Feature	Drill-Through	Bookmarks
Primary Purpose	Contextual navigation from a high-level summary to an entity-specific detailed target page.	State capture to save and recall specific report configurations, views, or filter states.
How it Works	Passes the active filter context of a selected visual data element (e.g., a specific product ID) to a dedicated destination page.	Captures a snapshot of the current report page state (visual visibility, slicer settings, drill status, and sorting).
Trigger Mechanism	Right-click on a data point → Select "Drill through" → Target Page.	Linked to clickable buttons, shapes, or image icons via the "Bookmark" action.
Common Use Case	Moving from a regional sales overview to an individual store's detailed sales transactions.	Creating interactive storyboards, toggling between different charts (e.g., Bar vs Pie), or building navigation menus.

MODEL QUESTION PAPER – 5

Q1. What is diagnostic analytics?

Diagnostic Analytics is a form of advanced analytics that examines historical data to determine the underlying reasons and root causes behind observed trends, patterns, or anomalies.

Core Question: Answers "*Why did it happen?*"

Key Techniques: Drill-down analysis, data discovery, correlation analysis, data mining, and root-cause analysis (RCA).

Example: Investigating why product sales dropped by 20% in a particular quarter (e.g., discovering supplier delivery delays or regional competitor discounts).

Q2. Name the three major areas of web analytics.

Web analytics (or Web Mining) is divided into three primary research and operational domains:

1. **Web Content Analysis (Web Content Mining):** The extraction and analysis of useful information, text, multimedia, and structured data embedded directly within web pages (e.g., search keywords, product descriptions).
2. **Web Structure Analysis (Web Structure Mining):** The examination of link architecture, hyperlinks, and relational topology connecting web pages and domains to determine page authority, popularity, and navigational structure (e.g., Google PageRank algorithm).
3. **Web Usage Analysis (Web Usage Mining / Clickstream Analytics):** The tracking, extraction, and evaluation of user behavioral patterns and navigation trails generated during user interaction with a website, recorded in server logs, cookies, and session data (e.g., click patterns, bounce rates, visit paths).

Q3. If $r = 0.6$ and $N = 64$, find probable error.

Given:

- Pearson correlation coefficient: $r = 0.6$
- Sample size: $N = 64$

Formula:

$$\text{P.E.} = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$$

Step-by-Step Calculation:

1. $r^2 = (0.6)^2 = 0.36$
2. $1 - r^2 = 1 - 0.36 = 0.64$
3. $\sqrt{N} = \sqrt{64} = 8$
4. $\frac{1-r^2}{\sqrt{N}} = \frac{0.64}{8} = 0.08$
5. $\text{P.E.} = 0.6745 \times 0.08 = 0.05396 \approx 0.054$

Significance Test:

$$6 \times \text{P.E.} = 6 \times 0.05396 = 0.32376$$

Since $r = 0.6 > 6 \times \text{P.E.}$ ($0.6 > 0.3238$), the correlation is **statistically significant and highly reliable**.

Q4. What is Non-linear Correlation?

Non-linear (or Curvilinear) Correlation is a statistical relationship between two variables where the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

Graphical Representation: When plotted on a scatter diagram, the paired data points form a curved line (such as a parabola, exponential curve, or logarithmic curve) rather than a straight line.

Example: Crop yield and fertilizer dosage (yield increases rapidly with initial fertilizer amounts, levels off, and eventually declines due to over-fertilization).

Note: The standard Pearson correlation coefficient (r) fails to accurately capture non-linear relationships and may yield a value near 0 even when a strong curved relationship exists.

Q5. What is Chi-Square Test?

The **Chi-Square (χ^2) Test** is a non-parametric statistical hypothesis test developed by Karl Pearson used to determine whether there is a statistically significant difference between expected frequencies and observed frequencies across one or more categorical variables.

Mathematical Formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

(where O_i is the Observed frequency, E_i is the Expected frequency, and $\sum$ denotes summation over all cells/categories).

Two Major Applications:

1. **Test of Goodness of Fit:** Tests whether observed sample data conforms to a theoretical probability distribution.
2. **Test of Independence of Attributes:** Determines whether two categorical variables organized in a contingency table are statistically independent (e.g., testing if vaccination status is independent of disease incidence).

Q6. Define probability.

Probability is a mathematical measure of the likelihood, chance, or degree of certainty that a specific event will occur as a result of a random experiment.

Classical (Mathematical) Definition: If a random experiment results in $n(S)$ mutually exclusive, exhaustive, and equally likely outcomes, and $n(E)$ of these outcomes are favorable to the occurrence of event E, the probability of event E is defined as:

$$P(E) = \frac{n(E)}{n(S)} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Axiomatic Range: For any event E, its probability strictly satisfies:

$$0 \leq P(E) \leq 1$$

(where $P(E) = 0$ represents an impossible event, and $P(E) = 1$ represents a certain event).

Q7. Define probability distribution.

A **Probability Distribution** is a mathematical function or rule that describes the probabilities of all possible values or outcomes that a random variable can take within a specified sample space.

Classification:

1. **Discrete Probability Distribution:** Defined for discrete random variables, where probabilities are assigned to isolated values using a Probability Mass Function (PMF), satisfying $P(X = x_i) \geq 0$ and $\sum P(X = x_i) = 1$ (e.g., Binomial, Poisson).
2. **Continuous Probability Distribution:** Defined for continuous random variables, where probabilities over intervals are modeled using a continuous curve known as a Probability Density Function (PDF), satisfying $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$ (e.g., Normal distribution).

Q8. What is a t-test? When is a t-test generally used?

A **t-test** (Student's t-test, formulated by William Sealy Gosset) is a parametric inferential statistical test used to determine whether there is a statistically significant difference between the means of two groups, or between a sample mean and a known population mean.

Formula (One-Sample t-test):

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \quad \text{with } df = n - 1$$

(where $\bar{x}$ is sample mean, μ is population mean, s is sample standard deviation, and n is sample size).

When a t-test is generally used:

1. When the sample size is **small ($n < 30$)**.
2. When the true population standard deviation (σ) is **unknown** and estimated using the sample standard deviation (s).

3. When the underlying population from which the sample is drawn is approximately **normally distributed**.

Q9. What is the Get Data option in Power BI?

The **Get Data** option is the primary data ingestion gateway and connectivity feature prominently located on the Home ribbon of Microsoft Power BI Desktop. It serves as the initial entry point for data analysts to discover, connect to, authenticate, and import data from hundreds of heterogeneous sources into Power BI.

Connectivity Categories: File (Excel, CSV, JSON, XML, PDF), Database (SQL Server, Oracle, MySQL, PostgreSQL), Microsoft Fabric / Power Platform, Azure Cloud databases, Online Services (SharePoint, Salesforce, Google Analytics), and Other (Web URLs, OData feed, Blank Query).

Ingestion Modes Available:

1. **Import Mode:** Loads and compresses a copy of data into memory using VertiPaq.
2. **DirectQuery Mode:** Keeps data in source and issues live queries on demand.

Q10. What are slicers and filters in Power BI?

Both slicers and filters are data-filtering mechanisms in Power BI that narrow down visual report data, but they differ fundamentally in implementation, scope, and user interface:

Dimension / Criterion	Slicers	Filters (Filters Pane)
Location & UI	Placed directly **on the report canvas** as an interactive, visible visual element (dropdown, list, slider, buttons).	Resides in the collapsible **Filters Pane** on the right side of the canvas (can be hidden from viewers).
Intended User	Designed for **end-users and report consumers** for intuitive, visible, one-click interactive exploration.	Primarily used by **report authors/designers** for granular, complex, or silent background filtering rules.
Scope of Impact	Affects only visuals on the current page that share the same data model (unless configured with *Sync Slicers* across pages).	Can be scoped precisely at three distinct levels: **Visual-level**, **Page-level**, and **Report-level** (all pages).
Advanced Logic	Limited to basic selection (single, multi-select, relative date/time).	Supports advanced filtering logic: Top N, text contains/starts with, empty/null checks, and complex boolean rules.

PART II: SECTION B — 6 MARKS QUESTIONS (PAPERS 1 TO 5)

MODEL QUESTION PAPER – 1

Q11. Explain the importance and benefits of data analytics to organisations.

1. Definition

Data Analytics (DA) is the systematic computational process of inspecting, cleansing, transforming, and modeling raw organizational data to discover meaningful patterns, correlations, and actionable insights. It transforms enterprise data into a strategic asset, replacing subjective intuition with evidence-based decision-making.

2. Workflow Diagram

```
[Raw Data] ---> [ETL & Modeling] ---> [Actionable Insights] ---> [Business Value & ROI]
```

3. Key Points (Importance & Benefits)

- **Evidence-Based Decision-Making:** Eliminates guesswork by providing quantifiable metrics, allowing executives to validate hypotheses and allocate capital with statistical certainty.
- **Operational Efficiency & Cost Reduction:** Identifies workflow bottlenecks and automates routine processes, significantly reducing operational waste and overhead costs.
- **Customer Personalization & Retention:** Segments customers by behavior and purchasing history to deliver personalized offerings, increasing retention and Customer Lifetime Value (CLV).
- **Proactive Risk Mitigation & Fraud Prevention:** Detects anomalous transaction patterns in real time, preventing financial fraud, credit defaults, and regulatory non-compliance.
- **Accurate Demand Forecasting:** Employs predictive time-series models on historical trends to optimize supply chains and inventory levels.

4. Example

- **Example:** E-commerce platforms analyze user browsing logs to power recommendation engines, boosting product sales conversions by over 25%.

Q12. Describe the important features and sources of text analytics.

1. Definition

Text Analytics (or Text Mining) is the automated discipline of converting unstructured natural language text into structured, quantitative data using Natural Language Processing (NLP) and machine learning to uncover hidden trends, opinions, and concepts.

2. Process Flowchart

```
[Unstructured Text] ---> [NLP Preprocessing] ---> [Feature Extraction] ---> [Structured Insights]
```

3. Key Points (Features & Sources)

- **Information Extraction (NER):** Automatically recognizes and extracts named entities such as person names, organizations, dates, and locations from raw text.
- **Sentiment Analysis:** Evaluates subjective customer tone, classifying text polarity into positive, negative, or neutral categories.
- **Text Classification & Summarization:** Categorizes documents into predefined thematic buckets (e.g., spam detection) and generates concise extractive or abstractive summaries.

- **Internal Enterprise Sources:** Ingests customer support tickets, CRM sales logs, call center transcripts, employee surveys, and corporate PDF reports.
- **External Web Sources:** Extracts real-time sentiment from social media posts (X, Reddit), customer review portals (Amazon, Google), and online news feeds.

4. Example

- **Example:** Airlines use text analytics on social media tweets to identify passenger complaints and dispatch real-time customer service support.

Q13. From the following table, find the correlation coefficient between age and playing habits of students.

Age (in years)	Number of Students	Regular Players
15	250	200
16	200	150
17	150	90
18	120	48
19	100	30
20	80	12

1. Given Data & Formulation

- Variable X = Age of students (in years).
- Variable Y = Playing habit percentage = $\left(\frac{\text{Regular Players}}{\text{Total Students}}\right) \times 100$.
- Number of observations $N = 6$.

$$Y_{15} = \frac{200}{250} \times 100 = 80\%, \quad Y_{16} = \frac{150}{200} \times 100 = 75\%, \quad Y_{17} = \frac{90}{150} \times 100 = 60\%$$

$$Y_{18} = \frac{48}{120} \times 100 = 40\%, \quad Y_{19} = \frac{30}{100} \times 100 = 30\%, \quad Y_{20} = \frac{12}{80} \times 100 = 15\%$$

2. Formula: Karl Pearson's Correlation Coefficient (r)

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \cdot \sum (Y - \bar{Y})^2}} = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}}$$

- Mean of X: $\bar{X} = \frac{\sum X}{N} = \frac{105}{6} = 17.5$
- Mean of Y: $\bar{Y} = \frac{\sum Y}{N} = \frac{300}{6} = 50.0$

3. Calculation Table

Age (X)	Students	Players	Y (%)	$x = X - 17.5$	$y = Y - 50$	x^2	y^2	xy
15	250	200	80	-2.5	$+30$	6.25	900	-75.0
16	200	150	75	-1.5	$+25$	2.25	625	-37.5
17	150	90	60	-0.5	$+10$	0.25	100	-5.0
18	120	48	40	$+0.5$	-10	0.25	100	-5.0
19	100	30	30	$+1.5$	-20	2.25	400	-30.0
20	80	12	15	$+2.5$	-35	6.25	1225	-87.5
Total	—	—	$\sum Y = 300$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 17.50$	$\sum y^2 = 3350$	$\sum xy = -240.0$

4. Step-by-Step Computation

$$r = \frac{-240.0}{\sqrt{17.50 \times 3350.0}} = \frac{-240.0}{\sqrt{58625.0}} = \frac{-240.0}{242.126} \approx -0.9912$$

5. Boxed Final Answer

$$\boxed{r \approx -0.9912}$$

- **Interpretation:** There is an exceptionally strong negative linear correlation ($r \approx -0.99$). As student age increases from 15 to 20, regular playing habit drops drastically from 80% to 15%.

Q14. State and explain Bayes' Theorem and describe its applications.

1. Definition / Statement

Bayes' Theorem is a fundamental probability law that calculates the updated (posterior) probability of an event given that related evidence has occurred, combining prior knowledge with observed likelihoods.

2. Formula & Core Components

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j) \cdot P(A_j)}$$

Component	Term	Meaning
$P(A_i \mid B)$	**Posterior**	Probability of hypothesis A_i given evidence B.
$P(B \mid A_i)$	**Likelihood**	Probability of evidence B given hypothesis A_i.
$P(A_i)$	**Prior**	Initial baseline probability of hypothesis A_i.
$P(B)$	**Evidence**	Total probability of observing evidence B.

3. Key Points (Applications)

- **Medical Diagnostics:** Calculates true disease probability after a positive screening test, resolving the base-rate fallacy.
- **Spam Filtering:** Classifies emails as spam or legitimate based on word frequencies using Naive Bayes.
- **Fraud Detection:** Updates transaction risk scores dynamically based on unusual purchasing velocity or location.

- **Machine Learning:** Underpins Bayesian classifiers, parameter estimation, and probabilistic decision networks.

4. Example

- **Example:** In a screening with 1% disease rate and 95% test accuracy, Bayes' theorem reveals a positive test yields ~16.1% disease probability.

Q15. Explain data loading in Power BI and the common approaches used for loading data.

1. Definition

Data Loading in Power BI is the ingestion phase where external data is extracted, transformed via the **Power Query (M engine)** ETL pipeline, and loaded into the **VertiPaq** columnar model or queried live.

2. Comparison Table of Loading Approaches

Storage Mode	Storage Location	Performance	Data Freshness
Import	Local RAM (VertiPaq)	Very Fast (In-memory)	Scheduled Refresh
DirectQuery	Source Database	DB/Network bound	100% Real-Time
Composite	Hybrid (RAM + DB)	Optimized for Aggs	Mixed
Live Connection	External SSAS	Very Fast (Server)	Centrally Managed

3. Key Points (Approaches & Mechanism)

- **Import Mode (Default):** Loads a compressed data snapshot into memory, providing full DAX and transformation functionality.
- **DirectQuery Mode:** Stores no local data; converts report clicks into real-time SQL queries executed at the source database.
- **Composite Mode:** Blends Import (for small dimension tables) and DirectQuery (for massive fact tables) in one model.
- **Data Refresh Architecture:** Synchronizes imported datasets via scheduled refresh using the On-Premises Data Gateway.

4. Example

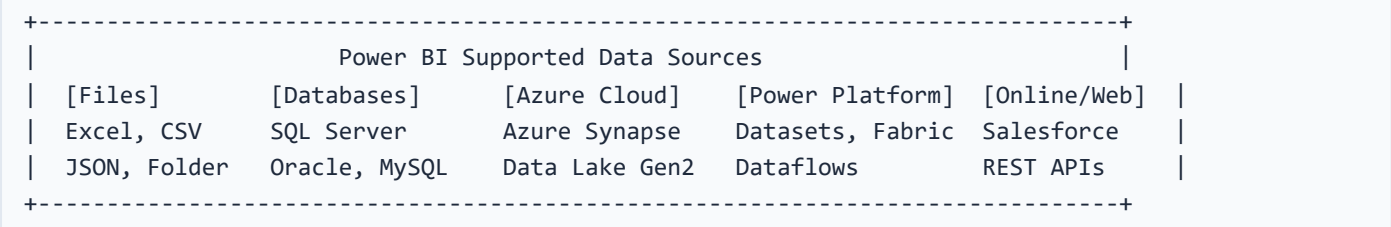
- **Example:** An enterprise queries a 500-million-row sales ledger via DirectQuery while importing date and customer dimension tables in memory.

Q16. Explain the different categories of data sources supported by Power BI.

1. Definition

Power BI provides extensive enterprise connectivity via its **"Get Data"** interface, supporting hundreds of heterogeneous data repositories across on-premises files, enterprise databases, cloud platforms, SaaS solutions, and web feeds.

2. Taxonomy Diagram



3. Key Points (Source Categories)

- **File Sources:** Imports local and shared files including Excel (.xlsx), CSV, Text, JSON, XML, PDF tables, and SharePoint folders.
- **Relational Databases:** Connects natively to enterprise engines like Microsoft SQL Server, Oracle, MySQL, PostgreSQL, IBM DB2, and SAP HANA.
- **Azure Cloud Sources:** Ingests from hyperscale cloud solutions including Azure SQL Database, Azure Synapse Analytics, and Azure Data Lake Storage Gen2.
- **Power Platform & Fabric:** Connects to certified Power BI semantic models, Dataflows, Microsoft Dataverse, and Fabric OneLake for enterprise reuse.
- **Online Services & Web:** Ingests from SaaS applications (Salesforce, Google Analytics, SharePoint Lists) and live web URLs/REST APIs.

4. Example

- **Example:** A retail company connects Power BI to Azure Synapse for sales analytics and a SharePoint folder for department budget spreadsheets.

Model Question Paper - 2 | Section B Solutions (Exam-Length Edition)

Total Marks: 24 (Answer any 4 out of 6 Questions, 6 Marks each — All 6 Solved)
Format: Strict 4-Part Structure (Definition, Diagram/Table, Key Points, Example/Formula) calibrated for a 10-minute handwritten exam response (strictly 150–220 words per theory answer). Strictly NO memory tips.

MODEL QUESTION PAPER – 2

Q11. Explain the complete data analytics process with a neat diagram.

Answer:

1. Definition (1 Mark)

The **Data Analytics Process** is an iterative lifecycle where raw data is systematically collected, cleaned, transformed, modeled, and visualized to discover actionable business insights and support data-driven decision-making.

2. Workflow Diagram (2 Marks)



3. Key Points (2–3 Marks)

- **Problem Definition & Ingestion:** Define analytical KPIs and extract structured/unstructured data from SQL databases, APIs, and flat files.
- **Cleaning & Preprocessing:** Impute missing values, eliminate duplicate records, filter outliers, and standardize date and numeric fields.
- **Exploratory Data Analysis (EDA):** Compute descriptive statistics (mean, variance) and inspect distributions, correlations, and anomalies using graphical plots.
- **Statistical Modeling:** Apply supervised or unsupervised algorithms (e.g., linear regression, classification, clustering) to discover patterns or forecast trends.
- **Validation & Deployment:** Assess performance (R^2, RMSE), build interactive visualization dashboards, and share actionable findings with stakeholders.

4. Practical Example (1 Mark)

- **Example:** A retail chain analyzes transaction logs to forecast seasonal inventory demands and reduce stockout rates by 20%.

Q12. Explain words, phrases and topics in text analytics with suitable examples.

Answer:

1. Definition (1 Mark)

In **Text Analytics**, unstructured text is decomposed into three granularities—words, phrases, and topics—to extract structured features, contextual sentiment, and thematic patterns.

2. Comparison Table (2 Marks)

Level	Linguistic Unit	Extraction Method	Practical Example
Word	Lexical token	Tokenization, Lemmatization	`"battery"`, `"fast"`
Phrase	Syntactic N-gram	POS Tagging, Bigrams	`"long battery life"`
Topic	Semantic theme	Latent Dirichlet Allocation (LDA)	`{battery, charge, drain}`

3. Key Points (2–3 Marks)

- **Words (Token Level):** Atomic units produced by splitting text, stripping punctuation, filtering stop-words ("*the*"), and lemmatizing inflections ("*running*" $\rightarrow$ "*run*").
- **Phrases (N-gram Level):** Contiguous word sequences (bigrams, trigrams) that preserve local grammatical context lost by isolated words.
- **Sentiment Modifiers:** Phrases capture polarity shifters (e.g., distinguishing "*not good*" from "*good*") vital for sentiment scoring.
- **Topics (Corpus Level):** Latent themes discovered across a document collection without manual annotations.
- **Topic Modeling Mechanics:** Algorithms like LDA model documents as topic mixtures and topics as keyword probability distributions.

4. Practical Example (1 Mark)

- **Example:** In smartphone reviews, "*battery*" is a word, "*fast charging*" is a phrase, and "*Battery Performance*" is the discovered topic.

Q13. Define Regression Analysis. Explain its importance and applications in Data Analysis.

Answer:

1. Definition (1 Mark)

Regression Analysis is a supervised statistical modeling technique used to investigate, quantify, and model the functional relationship between a dependent variable (Y, target) and one or more **independent variables** (X, predictors).

2. Mathematical Formula & Flow (2 Marks)

$$Y = \beta_0 + \beta_1 X + \epsilon$$

(Where β_0 is the intercept, β_1 is the slope coefficient, and ϵ is the random error term).

```
[Predictor X (e.g., Ad Spend)] ---> [Model: Y = b0 + b1*X] ---> [Target Y (e.g., Sales)]
```

3. Key Points: Importance in Data Analysis (2–3 Marks)

- **Predictive Forecasting:** Generates reliable equations to forecast future values of continuous business outcomes based on historical trends.
- **Assessing Relationship Strength:** Regression coefficients (β) measure the magnitude and direction (positive or negative) of each predictor's effect on Y.
- **Controlling Confounders:** Multiple regression isolates the unique impact of one variable while holding all other independent factors constant (*ceteris paribus*).
- **Statistical Validation:** Uses hypothesis testing (t-test, p-values) and R^2 (explained variance) to verify relationship significance.
- **Optimization:** Guides resource allocation decisions by identifying predictors that yield the highest marginal return.

4. Applications (1 Mark)

- **Applications:** Sales forecasting based on ad spend, real estate price valuation based on square footage, and credit risk default scoring.

Q14. Numerical: Continuous RV $f(x) = k$ for $5 \leq x \leq 15$, find k and $P(7 \leq X \leq 12)$.

Solution:

1. Given Data

A continuous random variable X has the probability density function (PDF):

$$f(x) = \begin{cases} k, & \text{for } 5 \leq x \leq 15 \\ 0, & \text{otherwise} \end{cases}$$

2. Step 1: Find the Value of k

For any valid continuous PDF, the total area under the density curve must equal 1:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \implies \int_{5}^{15} k\,dx = 1$$

Integrate with respect to x:

$$k[x]_{5}^{15} = 1 \implies k(15 - 5) = 1$$

$$10k = 1 \implies k = \frac{1}{10} = 0.1$$

3. Step 2: Calculate $P(7 \leq X \leq 12)$

The probability that X lies between 7 and 12 is given by the definite integral:

$$P(7 \leq X \leq 12) = \int_{7}^{12} f(x)\,dx = \int_{7}^{12} \frac{1}{10}\,dx$$

$$= \frac{1}{10}[x]_{7}^{12} = \frac{1}{10}(12 - 7) = \frac{5}{10} = \frac{1}{2} = 0.5$$

4. Final Boxed Answers

$$k = \frac{1}{10} = 0.1$$

$$P(7 \leq X \leq 12) = \frac{1}{2} = 0.5$$

Q15. Explain the Power BI Desktop interface and its important parts.

Answer:

1. Definition (1 Mark)

Power BI Desktop is Microsoft's free application for data preparation (Power Query), data modeling (VertiPaq/DAX), and creating interactive analytical visual reports.

2. Interface Layout Diagram (2 Marks)

```
+-------------------------------------------------------------------------+
| TOP: Ribbon Menu (Home, Insert, Modeling, View, Help)                   |
+----------+-----------------------------------+--------------------------+
| LEFT:    | CENTER:                           | RIGHT:                   |
| - Report |                                   | - Filters Pane           |
| - Data   |         Report Canvas             | - Visualizations Pane    |
| - Model  |                                   | - Data / Fields Pane     |
+----------+-----------------------------------+--------------------------+
| BOTTOM: Page Tabs (Page 1, Page 2, +) | Canvas Zoom Slider              |
+-------------------------------------------------------------------------+
```

3. Key Parts of Interface (2–3 Marks)

- **Ribbon Menu:** Top command bar hosting `Get Data`, `Transform Data`, `New Measure`, and `Publish`.
- **Three Views (Left Bar):** *Report View* (visual canvas), *Data View* (tabular data preview), and *Model View* (ER schema design).
- **Report Canvas:** Central design workspace where charts, cards, and slicers are arranged across pages.
- **Visualizations Pane:** Palette of charts (bar, line, pie) with field wells (*Axis*, *Values*) and formatting.
- **Data & Filters Panes:** *Data Pane* lists tables/measures; *Filters Pane* sets visual-, page-, and report-level filters.

4. Practical Example (1 Mark)

- **Example:** Linking `Sales` and `Regions` in Model View, then dragging `Revenue` onto the canvas for a sales chart.

Q16. Explain the procedure for removing unwanted columns and rows and renaming columns in Power Query.

Answer:

1. Definition (1 Mark)

Power Query is Power BI's ETL engine that cleans and reshapes data, logging each step in the **Applied Steps** pane using **M code**.

2. Summary Procedure Table (2 Marks)

Operation	Ribbon Path	Direct Method	M Function
Remove Columns	Home → Remove Columns	Select → Delete	Table.RemoveColumns()
Keep Other Columns	Home → Remove Other Columns	Right-click → Remove Other	Table.SelectColumns()
Remove Rows	Home → Remove Rows	Top/Bottom Rows	Table.Skip(), Table.Distinct()
Rename Column	Transform → Rename	Double-click header	Table.RenameColumns()

3. Step-by-Step Procedures (2–3 Marks)

- **Remove Columns:** Select headers (Ctrl + click) → Home → click Remove Columns (or Delete).
- **Remove Other Columns:** Select required columns → click *Remove Other Columns* to exclude unwanted fields.
- **Remove Rows:** Home → Remove Rows → select *Remove Top Rows* or *Remove Duplicates*.
- **Rename Columns:** Double-click header (or right-click → *Rename*), type new name, press Enter.
- **Apply Changes:** Click Close & Apply to load cleaned records into the model.

4. Practical Example (1 Mark)

- **Example:** Dropping 2 header rows via *Remove Top Rows* and renaming cust_nm to Customer Name.

Model Question Paper - 3 | Section B Solutions (Exam Length)

Exam Standard: 10-minute writing limit per question (strictly 150–220 words for theory; compact, step-by-step arithmetic for numericals). Strictly NO Memory Tips.

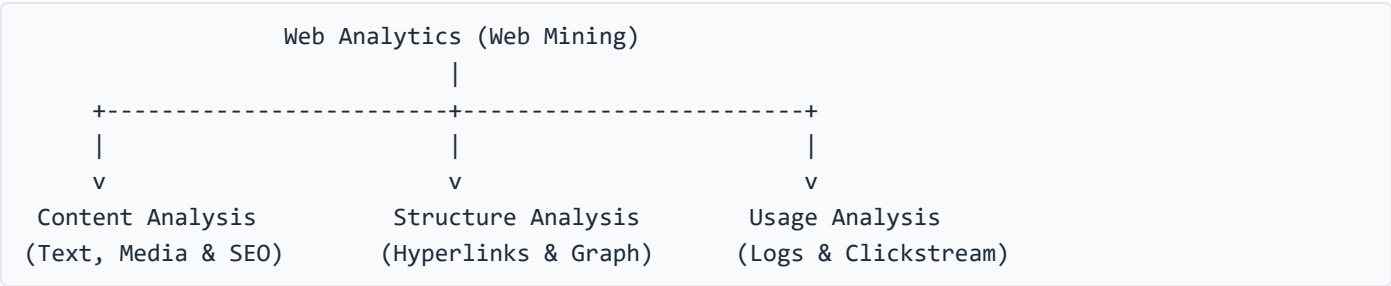
MODEL QUESTION PAPER – 3

Q11. What is Web Analytics? Explain content analysis, structure analysis and usage analysis in web analytics. (6 Marks)

1. Definition (1 Mark)

Web Analytics is the measurement, collection, analysis, and reporting of web data to understand, evaluate, and optimize website usage, visitor experience, and business return on investment (ROI).

2. Architecture / Taxonomy Diagram (2 Marks)



3. Key Points (2 Marks)

- **Web Content Analysis:** Evaluates the actual substance of web pages (unstructured text, media, metadata) using Natural Language Processing (NLP) and TF-IDF to determine content relevance and optimize Search Engine Optimization (SEO).
- **Web Structure Analysis:** Examines the hyperlink topology connecting pages using graph theory (modeled as nodes and directed edges); utilizes algorithms like Google PageRank and Kleinberg's HITS (Hubs and Authorities) to analyze link authority.
- **Web Usage Analysis:** Tracks user behavioral clickstream data gathered from web server logs, JavaScript page tags, and cookies to analyze navigation patterns, session durations, and drop-off points.

- **Core Business Objective:** Harmonizes informational quality (Content), website crawlability (Structure), and user engagement (Usage) to minimize bounce rates and maximize conversion rates.

4. Real-World Example (1 Mark)

- **Example:** In an e-commerce platform, PageRank analyzes external backlinks (Structure), NLP monitors keyword relevance (Content), and Google Analytics tracks cart abandonment (Usage).

Q12. Find out Karl Pearson's Coefficient of Correlation from the following data between Income and Weight. (6 Marks)

1. Given Data

Number of pairs (N) = 6:

Income in ₹ (X)	100	200	300	400	500	600
Weight in Kgs (Y)	120	130	140	150	160	170

2. Arithmetic Means

- Mean of X: $\bar{X} = \frac{\sum X}{N} = \frac{2100}{6} = 350$
- Mean of Y: $\bar{Y} = \frac{\sum Y}{N} = \frac{870}{6} = 145$

3. Formula

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}}$$

where $x = X - \bar{X} = X - 350$ and $y = Y - \bar{Y} = Y - 145$.

4. Calculation Table

X	$x = X - 350$	x^2	Y	$y = Y - 145$	y^2	$xy = x \cdot y$
100	-250	$62,500$	120	-25	625	$+6,250$
200	-150	$22,500$	130	-15	225	$+2,250$
300	-50	$2,500$	140	-5	25	$+250$
400	$+50$	$2,500$	150	$+5$	25	$+250$
500	$+150$	$22,500$	160	$+15$	225	$+2,250$
600	$+250$	$62,500$	170	$+25$	625	$+6,250$
Total	$\sum x = 0$	$\sum x^2 = 175,000$	Total	$\sum y = 0$	$\sum y^2 = 1,750$	$\sum xy = 17,500$

5. Computation Steps

$$r = \frac{17,500}{\sqrt{175,000 \times 1,750}} = \frac{17,500}{\sqrt{306,250,000}} = \frac{17,500}{17,500} = +1.00$$

6. Final Answer

$$\boxed{r = +1 \quad \text{(Perfect Positive Linear Correlation)}}$$

Interpretation: There is a perfect positive linear relationship ($Y = 110 + 0.1X$).

> **Q13. In a factory, 0.5% of the components are defective. Components are packed in boxes of 20. Using Poisson approximation, find the expected number of boxes containing exactly one defective component in 10,000 boxes. (6 Marks)**

1. Given Data

- Defect probability (p): $p = 0.5\% = \frac{0.5}{100} = 0.005$
- Box capacity (n): $n = 20$
- Total boxes (N): $N = 10,000$

2. Poisson Parameter (λ)

Since n is finite and p is very small, Binomial approximates Poisson:

$$\lambda = n \cdot p = 20 \times 0.005 = 0.10$$

3. Formula

Poisson Probability Mass Function:

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

(Given table value: $e^{-0.10} \approx 0.904837$)

4. Probability of Exactly One Defective ($k = 1$)

$$P(X = 1) = \frac{e^{-0.10} \cdot (0.10)^1}{1!} = 0.10 \times 0.904837 = 0.0904837$$

5. Expected Number of Boxes (E)

$$E = N \times P(X = 1) = 10,000 \times 0.0904837 = 904.837 \approx 905 \text{ boxes}$$

6. Final Answer

$$\boxed{\text{Expected Boxes} = 904.84 \approx 905 \text{ boxes}}$$

> **Q14. Explain data loading in Power BI and the common approaches used for loading data. (6 Marks)**

1. Definition (1 Mark)

Data Loading in Power BI is the ETL process of connecting to external data sources, applying transformations in Power Query, and loading the resulting tables into the analytical data model (VertiPaq engine).

2. Comparison Table of Loading Modes (2 Marks)

Storage Mode	Data Location	Query Speed	Data Freshness	DAX Support
Import (Default)	In-Memory (VertiPaq)	Blazing fast	Scheduled Refresh	100% full DAX
DirectQuery	Source Database	Source dependent	Real-time / Live	Restricted DAX
Live Connection	External Model (SSAS)	High performance	Real-time	Visualization only
Composite	Hybrid (Cache + Live)	Optimized	Mixed	Supported

3. Key Points (2 Marks)

- **Import Mode:** Ingests and compresses all source data locally in the `.pbix` file; provides sub-second query speeds but requires scheduled refreshes (capped at 1 GB per dataset on Pro).
- **DirectQuery Mode:** Stores zero data in memory; translates visual interactions dynamically into native source queries (e.g., SQL), ideal for multi-terabyte enterprise databases.
- **Live Connection Mode:** Connects directly to an external pre-built semantic model (SSAS or Power BI Service Dataset); completely bypasses Power Query and acts purely as a reporting canvas.
- **Composite / Dual Mode:** Combines Import and DirectQuery tables in a single data model, setting dimension tables to "Dual" for optimized local caching.
- **Loading Procedure:** User clicks Get Data → selects connector → previews data in Navigator → transforms in Power Query → clicks Close & Apply.

4. Real-World Example (1 Mark)

- **Example:** Loading a 50 MB Excel sales target file via Import Mode, while connecting to a multi-terabyte SAP HANA database via DirectQuery.

Q15. Explain Merge Queries and Append Queries with suitable examples. (6 Marks)

1. Definition (1 Mark)

In Power Query, **Merge Queries** combines two tables horizontally by joining columns based on a common matching key, whereas **Append Queries** combines two or more tables vertically by stacking rows with identical schemas.

2. Structural Comparison Table (2 Marks)

Feature	Merge Queries	Append Queries
Operation Orientation	**Horizontal** (Adds columns side-by-side)	**Vertical** (Stacks rows top-to-bottom)
Database Equivalent	SQL `JOIN` (Inner, Left Outer, Full Outer)	SQL `UNION ALL`
Matching Requirement	Requires a **Common Key Column**	Requires **Matching Column Header Names**
Table Multiplicity	Two tables per merge operation	Two or multiple (3+) tables simultaneously

3. Key Points (2 Marks)

- **Merge Queries Functionality:** Compares keys across tables and supports 6 join types: Left Outer (default), Right Outer, Full Outer, Inner, Left Anti, and Right Anti.
- **Merge Expansion Step:** Generates a nested table column that must be explicitly "Expanded" to expose the desired lookup attributes.
- **Append Queries Functionality:** Combines records vertically; column alignment depends strictly on exact column header names (case-sensitive) rather than column positions.
- **Mismatched Schema Handling:** In an append operation, unique columns not shared across tables are retained, populating `null` values for rows lacking that column.

- **Query Modification Options:** Both operations support in-place alteration (*Merge/Append*) or outputting to a brand new query (*Merge/Append as New*).

4. Concrete Examples (1 Mark)

- **Merge Example:** Joining `Employees (EmpID, Name, DeptID)` with `Departments (DeptID, DeptName)` on `DeptID` to add `DeptName`.
- **Append Example:** Stacking `Sales_Jan (Date, Amount)` and `Sales_Feb (Date, Amount)` into a unified `Sales_Total` table.

Q16. Explain data types and formatting operations available in Power Query. (6 Marks)

1. Definition (1 Mark)

Data Types in Power Query define the classification, memory allocation, and valid operations for column values, while **Formatting Operations** clean, restructure, and transform the appearance and layout of that data.

2. Classification Matrix (2 Marks)

```
+-------------------------------------------------------------------------+
|                      Power Query Data Types & Formatting                |
+------------------+--------------------+--------------------+------------------+
| Numeric          | Temporal           | Text & Logic       | Complex / Binary |
| Decimal (1.2)    | Date, Time         | Text (ABC)         | List, Record     |
| Fixed Dec ($)    | Date/Time, Duration | True/False (T/F)  | Table, Binary    |
| Whole No. (123)  | Timezone           | Any                |                  |
+------------------+--------------------+--------------------+------------------+
```

3. Key Points (2 Marks)

- **Numeric Data Types & Math:** Supports Whole Number, Decimal Number, and Fixed Decimal (Currency to avoid rounding errors); provides operations like Rounding (Floor, Ceiling) and standard arithmetic.
- **Text Formatting Operations:** Case conversion (`lowercase`, `UPPERCASE`, `Capitalize Each Word`), `Trim` (removes leading/trailing spaces), and `Clean` (strips non-printable ASCII characters).
- **Text Splitting & Extraction:** Splits text by delimiter or character count; extracts string length, prefix, suffix, and substrings before/after delimiters.
- **Date & Time Operations:** Extracts components (Year, Month name, Quarter, Week, Day of week) and computes elapsed `Age` against current local time (`DateTime.LocalNow()`).
- **Locale-Specific Parsing:** "Change Type with Locale" handles regional differences, such as distinguishing `DD/MM/YYYY` from `MM/DD/YYYY` or parsing comma decimals.

4. Real-World Example (1 Mark)

- **Example:** Setting `Salary` to Fixed Decimal Number (`$`) and applying `Trim` to `CustomerName` to eliminate trailing spaces before executing table merges.

Model Question Paper – 4 | Section B Solutions
Total Marks: 24 (Answer any 4 out of 6 Questions, 6 Marks each) | Comprehensive Solutions

Total Marks: 24 (Answer any 4 out of 6 Questions; All 6 Solved Here × 6 Marks each)
Standard: Exam-length solutions (10-minute writing limit, 150–220 words per theory question, strict 4-part structure, and step-by-step mathematical numericals). Exclusively formal academic solutions.

MODEL QUESTION PAPER – 4

Q11. Explain the major techniques used in data analytics.

1. Definition:
Data analytics techniques are computational, statistical, and algorithmic methods used to inspect, clean, transform, and model raw data to discover meaningful patterns and support business decisions.

2. Architecture & Taxonomy Diagram:

```
Data Analytics Techniques
├── Statistical Inference (Hypothesis testing, ANOVA)
├── Regression Modeling (Linear, Logistic)
├── Data Mining & Clustering (K-Means, Association Rules)
├── Machine Learning (Decision Trees, Random Forest)
└── Text & Time Series Analytics (NLP, ARIMA)
```

3. Key Analytics Techniques:
* **Statistical Analysis:** Applies descriptive measures and hypothesis tests (t-test, ANOVA, Chi-square) to infer population properties from sample data. * **Regression Modeling:** Quantifies mathematical dependencies between a dependent target variable (Y) and independent predictor variables (X) for predictive forecasting. * **Clustering & Data Mining:** Discovers natural groupings in unlabeled observations (K-Means) and extracts transaction affinities via Association Rule Mining. * **Machine Learning Algorithms:** Trains supervised and ensemble models (Decision Trees, Random Forests) to recognize complex, non-linear relationships. * **Time Series Forecasting:** Models sequential chronological observations using moving averages, exponential smoothing, or ARIMA models to project future trends.

4. Real-World Example:
* *Example:* An e-commerce platform uses association rule mining (Apriori algorithm: $\text{Confidence} = \frac{P(A \cap B)}{P(A)}$) to generate automated cross-sell product recommendations.

Q12. Write the skills required for business analytics.

1. Definition:
Business analytics skills represent an interdisciplinary blend of quantitative reasoning, computational tools, and domain expertise required to convert raw organizational data into strategic business decisions.

2. Skill Triad Diagram:

```
Business Analytics Skills Triad
┌──────────────────────────────────────────────────────────┐
│ Technical & Tools   | Statistical & Math   | Domain & Comm  |
│ (SQL, Python, BI)   | (Probability, Reg)   | (KPIs, Story)  |
└──────────────────────────────────────────────────────────┘
```

3. Key Skill Dimensions:
* **Database & SQL Querying:** Proficient in writing complex multi-table joins, aggregations, CTEs, and window analytical functions to extract transaction data. * **BI & Visual Storytelling:** Skilled in Power BI or Tableau to build interactive dashboards, data models, and DAX measures for executive reporting. * **Statistical & Mathematical Modeling:** Grounded in probability distributions, hypothesis testing, linear regression, and error estimation to validate business patterns. * **Business Domain Acumen:** Deep understanding of industry-specific KPIs (e.g., CAC, Churn Rate, LTV, ROI) to structure open-ended business problems. * **Stakeholder Communication:** Ability to translate complex algorithmic insights into clear, persuasive executive summaries for non-technical leadership.

4. Practical Application:
* *Application:* A retail analyst queries sales using SQL, builds a customer churn model in Python, and presents retention strategies via a Power BI dashboard.

Q13. Obtain two regression equations from the following data, where deviations are taken from actual means:

$$N = 20, \quad \bar{x} = 4, \quad \bar{y} = 2, \quad \sum x^2 = 1680, \quad \sum y^2 = 320, \quad \sum xy = 480$$

1. Given Data:
* Number of observations (N) = 20 * Actual mean of X ($\bar{x}$) = 4 * Actual mean of Y ($\bar{y}$) = 2 * Deviations taken from actual means: $x = (X - \bar{x})$ and $y = (Y - \bar{y})$ * $\sum x^2 = \sum (X - \bar{x})^2 = 1680$ * $\sum y^2 = \sum (Y - \bar{y})^2 = 320$ * $\sum xy = \sum (X - \bar{x})(Y - \bar{y}) = 480$

2. Formulas:
* Regression coefficient of Y on X: $b_{yx} = \dfrac{\sum xy}{\sum x^2}$ * Regression coefficient of X on Y: $b_{xy} = \dfrac{\sum xy}{\sum y^2}$ * Regression equation of Y on X: $(Y - \bar{y}) = b_{yx}(X - \bar{x})$ * Regression equation of X on Y: $(X - \bar{x}) = b_{xy}(Y - \bar{y})$

3. Step-by-Step Calculation:

- **Step 1: Calculate Regression Coefficients (b_{yx} and b_{xy}):**

$$b_{yx} = \frac{480}{1680} = \frac{48}{168} = \frac{2}{7} \approx 0.2857$$

$$b_{xy} = \frac{480}{320} = \frac{48}{32} = \frac{3}{2} = 1.5$$

- **Step 2: Obtain Regression Equation of Y on X:**

$$(Y - \bar{y}) = b_{yx}(X - \bar{x})$$

$$(Y - 2) = \frac{2}{7}(X - 4)$$

$$7(Y - 2) = 2(X - 4)$$

$$7Y - 14 = 2X - 8$$

$$2X - 7Y + 6 = 0 \quad \text{or} \quad Y = \frac{2}{7}X + \frac{6}{7} \quad (Y = 0.2857X + 0.8571)$$

- **Step 3: Obtain Regression Equation of X on Y:**

$$(X - \bar{x}) = b_{xy}(Y - \bar{y})$$

$$(X - 4) = 1.5(Y - 2)$$

$$X - 4 = 1.5Y - 3$$

$$X = 1.5Y + 1 \quad \text{or} \quad 2X - 3Y - 2 = 0$$

- **Step 4: Consistency Verification:**

$$r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\frac{2}{7} \times \frac{3}{2}} = \sqrt{\frac{3}{7}} \approx +0.6547 \quad (-1 \leq r \leq +1 \checkmark)$$

4. Final Boxed Answer:

Regression Equation of Y on X:	$2X - 7Y + 6 = 0 \quad (or Y = 0.2857X + 0.8571)$
Regression Equation of X on Y:	$7X - 2Y - 14 = 0 \quad (or X = 0.2857Y + 2)$

Q14. A Binomial Distribution has mean 4 and variance 2. Find n and p. Also find $P(X = 2)$.

1. Given Data:
* Mean (μ) = $np = 4$ --- (1) * Variance (σ^2) = $npq = 2$ --- (2) * Condition: $p + q = 1 \implies q = 1 - p$

2. Formulas:
* Binomial Probability Mass Function (PMF):

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad \text{where } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

3. Step-by-Step Calculation:

- **Step 1: Find Probability of Failure (q) and Probability of Success (p):**
  Dividing equation (2) by equation (1):

$$\frac{\text{Variance}}{\text{Mean}} = \frac{npq}{np} = \frac{2}{4}$$

$$q = \frac{1}{2} = 0.5$$

  Since $p = 1 - q$:

$$p = 1 - 0.5 = 0.5 = \frac{1}{2}$$

- **Step 2: Find Number of Trials (n):**
  Substitute $p = 0.5$ into equation (1):

$$n \cdot p = 4 \implies n(0.5) = 4 \implies n = \frac{4}{0.5} = 8$$

  Thus, parameters are: $n = 8$, $p = 0.5$, and $q = 0.5$.

- **Step 3: Calculate $P(X = 2)$:**

$$P(X = 2) = \binom{8}{2} (0.5)^2 (0.5)^{8-2} = \binom{8}{2} (0.5)^8$$

- Combinations: $\binom{8}{2} = \frac{8 \times 7}{2 \times 1} = 28$
- Power term: $(0.5)^8 = \left(\frac{1}{2}\right)^8 = \frac{1}{256}$

- Probability:

$$P(X=2) = 28 \times \frac{1}{256} = \frac{28}{256} = \frac{7}{64} = 0.109375$$

4. Final Boxed Answer:

Parameters of Binomial Distribution:	$n = 8, \quad p = 0.5 \ \left(\frac{1}{2}\right), \quad q = 0.5 \ \left(\frac{1}{2}\right)$
Probability of Exactly 2 Successes:	$P(X=2) = \frac{28}{256} \approx 0.1094$

Q15. Explain the Business Intelligence process and its different stages.

1. Definition:
The Business Intelligence (BI) process is an integrated operational framework comprising tools, architectures, and data practices that transform raw transactional data into structured, actionable business insights.

2. BI Process Flowchart:

```
Data Sources ──▶ ETL Pipeline ──▶ Data Warehouse ──▶ OLAP Engine ──▶ Dashboards ──▶ Decision
(ERP, CRM, DB)  (Clean, Standard) (Fact/Dimensions) (Slices, DAX)  (Power BI)    (Action/ROI)
```

3. Key Stages of the BI Process:
* **Data Sourcing:** Gathers raw data from heterogeneous operational systems including ERP, CRM, relational SQL databases, flat files, and web APIs. * **ETL (Extract, Transform, Load):** Ingests source data, cleanses errors (deduplication, null handling), applies business logic, and standardizes formats. * **Data Warehousing:** Stores sanitized historical data in dimensional repositories (EDW, Data Marts) structured via Star or Snowflake schemas. * **OLAP & Semantic Modeling:** Enables multidimensional analysis (slice, dice, drill-down, roll-up) and KPI computations using analytical engines like DAX. * **Visualization & Reporting:** Delivers interactive executive dashboards, scorecards, and scheduled reports using tools like Power BI or Tableau.

4. Real-World Example:
* *Example:* A retail chain integrates daily point-of-sale data through ETL into an enterprise data warehouse to monitor regional store performance in real time.

Q16. Explain the meaning and important features of web analytics.

1. Definition:
Web analytics is the systematic measurement, collection, analysis, and reporting of Internet data to understand visitor behavior, evaluate website efficacy, and optimize digital marketing performance.

2. Web Analytics Architecture Diagram:

```
Web Analytics Architecture
├── Traffic Channels (Organic Search, Direct, Referral, Paid)
├── User Behavior (Pageviews, Bounce Rate, Session Duration)
├── Funnel Tracking (Shopping Cart Checkout Drop-offs)
└── Technical Metrics (Page Load Speed, Device & OS Breakdown)
```

3. Key Features of Web Analytics:
* **Traffic Acquisition Tracking:** Identifies user traffic channels (organic search, direct URLs, referral links, and paid campaigns using UTM parameters). * **Audience Profiling:** Analyzes user demographics, geographic location, device category (desktop vs mobile), browser type, and new versus returning visitors. * **Behavioral Engagement Metrics:** Evaluates user interactions via total pageviews, average session duration, exit rates, and bounce rate. * **Conversion & Funnel Analysis:** Maps multi-step user flows (e.g., cart to checkout) to identify friction points causing user drop-off. *

Real-Time Web Monitoring: Provides immediate live tracking of active concurrent visitors, current pages viewed, and instant goal completions.

4. Formula & Application:
* *Formula:* $\text{Bounce Rate} = \left(\frac{\text{Single-Page Sessions}}{\text{Total Sessions}}\right) \times 100\%$; highlights visitor engagement and landing page relevance.

End of Model Question Paper - 4 Section B Solutions.

Model Question Paper – 5 | Section B Solutions
Total Marks: 24 (Answer any 4 out of 6 Questions, 6 Marks each) | Comprehensive Solutions

Total Marks: 36 (6 Questions × 6 Marks each)
Format: Strict 10-minute exam-length standard (150–220 words per theory question, ~1.5 handwritten pages).
Structure per Theory Question: 1. Definition (1M) | 2. Diagram / Flowchart / Table (2M) | 3. Key Points (4-5 bullets) (2-3M) | 4. Example / Formula (1M).
Numericals: Given Data, Exact Formulas, Calculation Steps, Verification, and Boxed Answers. Strictly NO Memory Tips.

MODEL QUESTION PAPER – 5

Q11. Explain the steps involved in preparing text for analysis (Text Preprocessing Pipeline).

1. Definition (1 Mark)

Text Preprocessing is the data-preparation pipeline that cleans, standardizes, and transforms raw, noisy natural language text into a structured numerical feature representation suitable for machine learning.

2. Process Flowchart (2 Marks)

```
[Raw Text] -> [Noise Removal] -> [Case Folding] -> [Tokenization]
                                                        |
[Feature Matrix] <- [Vectorization] <- [Lemmatize] <- [Stop Words]
```

3. Key Pipeline Steps (2-3 Marks)

- **Noise Removal:** Strips HTML/XML tags, URLs, and punctuation using regular expressions (`re.sub()`) to eliminate non-textual distractions.
- **Case Normalization:** Converts text to lowercase to prevent duplicate features (`"Data"` vs `"data"`) and expands contractions (`"don't"` -> `"do not"`).
- **Tokenization:** Splits continuous text into atomic linguistic units called tokens (individual words, sentences, or n-grams).
- **Stop Words Filtering:** Removes high-frequency grammatical words (`"is"` , `"the"` , `"at"`) that convey negligible domain semantics.
- **Stemming & Lemmatization:** Reduces word variants to root forms; stemming chops affixes heuristically (`"studies"` -> `"studi"`), while lemmatization uses dictionaries (`"better"` -> `"good"`).

4. Example (1 Mark)

- **Example:** Raw: `"Data Analytics isn't hard!"` -> Tokens: `['data', 'analytics', 'not', 'hard']` .

Q12. Explain the applications, benefits, and limitations of text analytics.

1. Definition (1 Mark)

Text Analytics is the automated process of extracting meaningful patterns, quantitative metrics, and actionable intelligence from unstructured natural language text using NLP and machine learning.

2. Summary Comparison Table (2 Marks)

Dimension	Core Technique	Business Impact
Applications	Sentiment Analysis & Chatbots	Classifies customer sentiment; routes support tickets.
Benefits	High Scalability & Dark Data	Analyzes millions of documents without human fatigue.
Limitations	Sarcasm & Ambiguity	Polysemy (`"bank"`) and sarcasm degrade basic models.
Mitigation	Domain Transformers & Lexicons	Uses context-aware BERT with custom dictionaries.

3. Key Points (2-3 Marks)

- **Core Applications:** Drives brand sentiment tracking, automated support ticketing, clinical mining, and resume screening.
- **High Scalability:** Processes massive text volumes in minutes, unlocking dormant enterprise "dark data".
- **Objective Consistency:** Standardizes document classification, eliminating human cognitive fatigue, bias, and inconsistency.
- **Linguistic Ambiguity:** Struggles with polysemy (multiple word meanings) and sarcasm where literal words contradict intent.
- **Domain Overhead:** Demands costly retraining and labeled corpora to parse specialized medical or legal terminology.

4. Example (1 Mark)

- **Example:** Mining 50,000 product reviews to quantify customer sentiment (82% Positive, 18% Negative).

Q13. From the following data, obtain the two regression equations:

- **Series X:** Mean = $\mathbf{20}$, Variance = $\mathbf{4}$
- **Series Y:** Mean = $\mathbf{25}$, Variance = $\mathbf{9}$
- **Correlation coefficient (r):** $\mathbf{0.75}$

1. Given Data

- Mean of X: $\bar{X} = 20$
- Variance of X: $\sigma_X^2 = 4 \implies \sigma_X = \sqrt{4} = 2$
- Mean of Y: $\bar{Y} = 25$
- Variance of Y: $\sigma_Y^2 = 9 \implies \sigma_Y = \sqrt{9} = 3$
- Correlation coefficient: $r = 0.75 = \frac{3}{4}$

2. Formulas

- **Regression Coefficient of Y on X:** $b_{yx} = r \cdot \frac{\sigma_Y}{\sigma_X}$
- **Regression Equation of Y on X:** $(Y - \bar{Y}) = b_{yx}(X - \bar{X})$
- **Regression Coefficient of X on Y:** $b_{xy} = r \cdot \frac{\sigma_X}{\sigma_Y}$
- **Regression Equation of X on Y:** $(X - \bar{X}) = b_{xy}(Y - \bar{Y})$

3. Step-by-Step Calculation

A. Computation of Regression Coefficients

- b_{yx}:

$$b_{yx} = 0.75 \times \frac{3}{2} = 0.75 \times 1.5 = 1.125 \quad \left(\text{or } \frac{9}{8}\right)$$

- b_{xy}:

$$b_{xy} = 0.75 \times \frac{2}{3} = \frac{3}{4} \times \frac{2}{3} = \frac{2}{4} = 0.5 \quad \left(\text{or } \frac{1}{2}\right)$$

B. Derivation of Regression Equation of Y on X

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X})$$

$$(Y - 25) = 1.125(X - 20)$$

$$Y - 25 = 1.125X - 22.5$$

$$Y = 1.125X - 22.5 + 25$$

$$Y = 1.125X + 2.5 \quad \text{or} \quad 9X - 8Y + 20 = 0$$

C. Derivation of Regression Equation of X on Y

$$(X - \bar{X}) = b_{xy}(Y - \bar{Y})$$

$$(X - 20) = 0.5(Y - 25)$$

$$X - 20 = 0.5Y - 12.5$$

$$X = 0.5Y - 12.5 + 20$$

$$X = 0.5Y + 7.5 \quad \text{or} \quad 2X - Y = 15$$

4. Mathematical Verification

1. Geometric Mean Check: $b_{yx} \cdot b_{xy} = 1.125 \times 0.5 = 0.5625 = r^2 = (0.75)^2 \quad (\checkmark)$
2. Intersection at Means $(\bar{X} = 20, \bar{Y} = 25)$:
3. $Y = 1.125(20) + 2.5 = 22.5 + 2.5 = 25 \quad (\checkmark)$
4. $X = 0.5(25) + 7.5 = 12.5 + 7.5 = 20 \quad (\checkmark)$

5. Boxed Final Answer

Regression Equation of Y on X:	$2X - 7Y + 6 = 0 \quad (or Y = 0.2857X + 0.8571)$
Regression Equation of X on Y:	$7X - 2Y - 14 = 0 \quad (or X = 0.2857Y + 2)$

Q14. Explain Classical, Empirical, and Axiomatic approaches to probability.

1. Definition (1 Mark)

Probability quantifies the likelihood of occurrence of random events, conceptualized through deductive symmetry (Classical), relative frequency (Empirical), or set-theoretic foundations (Axiomatic).

2. Comparative Approaches Table (2 Marks)

Approach	Formulation Principle	Mathematical Formula	Underlying Condition
Classical	Deductive symmetry	$P(A) = \frac{m}{n}$	Finite, equiprobable outcomes.
Empirical	Relative frequency	$P(A) = \lim_{n \to \infty} \frac{f_n(A)}{n}$	Repeatable independent trials.
Axiomatic	Measure & set theory	Kolmogorov Axioms	Non-negativity, $P(S) = 1$, Additivity.

3. Key Points (2-3 Marks)

- **Classical (Laplace):** Computes $P(A) = m/n$ *a priori*; flawed by circularity ("*equally likely*") and fails for biased or continuous events.
- **Empirical (von Mises):** Computes probability *a posteriori* as limiting relative frequency; limited because infinite trials are impossible.
- **Axiomatic (Kolmogorov):** Defines probability as a real-valued function $P : \mathcal{F} \to \mathbb{R}$ on sample space S satisfying:
- *Non-negativity:* $P(A) \geq 0$.
- *Certainty:* $P(S) = 1$.
- *Additivity:* For disjoint events, $P(\bigcup A_i) = \sum P(A_i)$.
- **Universal Rigor:** Unifies classical and empirical models across discrete and continuous distributions.

4. Example (1 Mark)

- **Example:** Fair die: Classical $P(\text{Even}) = 3/6 = 0.5$; Empirical: 10,000 rolls with ratio $\frac{4,992}{10,000} \approx 0.499$.

Q15. Explain visualizations, reports, and dashboards in Power BI and distinguish between them.

1. Definition (1 Mark)

In Power BI, information delivery is structured across three presentation tiers: individual charts (**Visualizations**), multi-page analytical workbooks (**Reports**), and single-page executive canvases (**Dashboards**).

2. Systematic Comparison Table (2 Marks)

Feature	Visualization	Report	Dashboard
Canvas	Single visual	Multi-page canvas	Single-page only
Authoring	Desktop & Service	Desktop & Service (.pbix)	Power BI Service ONLY
Data Scope	Single visual fields	Single dataset	Multiple datasets & reports
Interactivity	Local tooltips	High (cross-filter, drill)	Pinned tiles (opens report)

3. Key Points (2-3 Marks)

- **Visualizations:** Atomic graphical charts (bar, line, KPI card) displaying aggregated metrics visually.
- **Reports:** Multi-page exploratory workbooks built on a single dataset, offering cross-filtering, slicers, and drill hierarchies.
- **Dashboards:** Single-page executive canvases in Power BI Service combining pinned tiles across multiple reports.

- **Natural Language Q&A:** Dashboards feature a search bar ("*Ask a question about your data*"), generating instant dynamic charts.
- **Data Alerts:** Dashboards support automated notifications when KPI targets are breached (unsupported in reports).

4. Workflow Hierarchy (1 Mark)

- **Architecture:** Dataset ⟶ Build Visuals ⟶ Assemble Multi-Page Report ⟶ Pin Tiles to Single-Page Dashboard.

Q16. Explain the process of data cleaning and transformation in Power Query.

1. Definition (1 Mark)

Power Query is the self-service ETL (Extract, Transform, Load) engine in Power BI and Excel that extracts raw data, executes cleaning transformations, and loads normalized tables into the data model.

2. Process Flowchart (2 Marks)

```
[Raw Data] -> [Profiling] -> [Row/Col Cleaning] -> [Type Casting]
                                                        |
[Data Model] <- [Merge / Append] <- [Unpivot Columns] <- [Trim / Clean]
```

3. Key Cleaning & Transformation Stages (2-3 Marks)

- **Non-Destructive Applied Steps:** Built on the functional M language; every transformation step is recorded sequentially in the "Applied Steps" pane without altering source files.
- **Structural Cleaning:** Promotes headers (`Use First Row as Headers`), removes unneeded columns (`Remove Other Columns`), and purges blank, error, or duplicate rows.
- **Data Typing & String Cleansing:** Explicitly casts scalar types (Whole Number, Decimal, Date, Text) and applies `Trim` (removes spaces) and `Clean` (strips unprintable characters).
- **Table Reshaping (Unpivot):** Converts wide, cross-tabulated reporting spreadsheets into tall, normalized tables (`Unpivot Columns`) required for star-schema models.
- **Data Mashup (Merge & Append):** Combines tables horizontally via key joins (`Merge Queries`) or stacks identical transaction tables vertically (`Append Queries`).

4. Practical Example (1 Mark)

- **Example:** Unpivoting monthly columns `[Jan, Feb, Mar]` into two normalized attributes: `[Month]` and `[Sales Amount]` .

PART III: SECTION C — 8 MARKS QUESTIONS (PAPERS 1 TO 5)

MODEL QUESTION PAPER – 1

Q17. Explain the evolution of Data Analytics.

1. Core Concept & Definition

The evolution of **Data Analytics (DA)** traces the transformation of raw quantitative recording into automated, real-time cognitive intelligence. Over six decades, organizational decision-making has progressed from manual historical bookkeeping to cloud-native, prescriptive machine-learning systems that autonomously guide global business actions.

2. Evolutionary Timeline Flowchart

```
[1970s: Manual Stats & RDBMS] ---> [1990s: Data Warehousing & BI]
            |                                   |
            v                                   v
[2000s: Big Data & Analytics 2.0] ---> [2020s+: Cloud AI & Prescriptive]
```

3. Key Phases in the Evolution of Data Analytics

- **Descriptive Foundations (Pre-1970s):** Grounded in classical inferential statistics (Gauss, Pearson), using manual tabulating machines and punch cards for census, accounting, and inventory records.
- **RDBMS & Structured Querying (1970s–1980s):** Edgar Codd's Relational Database Management System and SQL enabled structured Online Transaction Processing (OLTP) and scheduled batch mainframe reporting.
- **Data Warehousing & BI (1990s):** Organizations built centralized Enterprise Data Warehouses (EDW) and OLAP cubes (Inmon, Kimball), answering "*What happened?*" through historical dashboards and KPI scorecards.
- **Big Data Explosion (2000s–2010s):** Emergence of web platforms generated massive unstructured data (3Vs: Volume, Velocity, Variety), handled by distributed frameworks like Hadoop HDFS and Apache Spark.
- **Predictive Analytics & Data Science (2010s):** Focus shifted from retrospective reporting to forward-looking modeling ("*What will happen?*") using machine learning (regression, random forests) and Python/R.
- **Autonomous & Prescriptive Analytics (2020s+):** Characterized by serverless cloud lakehouses (Snowflake, BigQuery), streaming telemetry (Kafka), and automated ML (AutoML) answering "*What action to take?*"
- **Self-Service Democratization:** Transitioned from centralized IT-driven reports to low-code BI tools (Power BI, Tableau) empowering business stakeholders.

4. Concrete Example

- **Example:** Banking evolved from manual paper ledger balancing (1970s) to batch month-end EDW reporting (1990s), to real-time machine-learning fraud scoring blocking unauthorized charges in milliseconds (present).

Q18. Find out the correlation coefficient for the following X and Y series, using assumed mean.

X	45	55	56	58	60	65	68	70	75	80
Y	56	50	48	60	62	64	65	70	74	82

1. Given Data & Methodology

- Number of pairs of observations: $N = 10$.
- Method required: **Assumed Mean (Shortcut) Method**.
- Let Assumed Mean for X be $A = 60$.

- Let Assumed Mean for Y be $B = 60$.
- Deviations: $d_x = X - A = X - 60$ and $d_y = Y - B = Y - 60$.

2. Formula: Karl Pearson's Correlation Coefficient (r)

$$r = \frac{N\sum d_x d_y - (\sum d_x)(\sum d_y)}{\sqrt{[N\sum d_x^2 - (\sum d_x)^2]\,[N\sum d_y^2 - (\sum d_y)^2]}}$$

3. Complete Calculation Table

S.No	X	$d_x = X - 60$	d_x^2	Y	$d_y = Y - 60$	d_y^2	$d_x d_y$
1	45	-15	225	56	-4	16	$+60$
2	55	-5	25	50	-10	100	$+50$
3	56	-4	16	48	-12	144	$+48$
4	58	-2	4	60	0	0	0
5	60	0	0	62	$+2$	4	0
6	65	$+5$	25	64	$+4$	16	$+20$
7	68	$+8$	64	65	$+5$	25	$+40$
8	70	$+10$	100	70	$+10$	100	$+100$
9	75	$+15$	225	74	$+14$	196	$+210$
10	80	$+20$	400	82	$+22$	484	$+440$
Total	—	$\sum d_x = 32$	$\sum d_x^2 = 1084$	—	$\sum d_y = 31$	$\sum d_y^2 = 1085$	$\sum d_x d_y = 968$

4. Step-by-Step Substitution & Arithmetic

1. **Numerator:**

$$\text{Numerator} = N\sum d_x d_y - (\sum d_x)(\sum d_y) = (10 \times 968) - (32 \times 31) = 9680 - 992 = 8688$$

2. **Denominator Terms:**

$$\text{Term}_X = N\sum d_x^2 - (\sum d_x)^2 = (10 \times 1084) - (32)^2 = 10840 - 1024 = 9816$$

$$\text{Term}_Y = N\sum d_y^2 - (\sum d_y)^2 = (10 \times 1085) - (31)^2 = 10850 - 961 = 9889$$

3. **Combined Denominator:**

$$\text{Denominator} = \sqrt{\text{Term}_X \times \text{Term}_Y} = \sqrt{9816 \times 9889} = \sqrt{97,070,424} \approx 9852.432$$

4. **Coefficient Calculation:**

$$r = \frac{8688}{9852.432} \approx +0.8818$$

5. Boxed Final Answer & Interpretation

$$\boxed{r \approx +0.8818 \quad (\text{or } +0.882)}$$

- **Statistical Interpretation:** There is a high degree of positive linear correlation ($r = +0.8818$) between variable X and variable Y. As values in series X increase, values in series Y exhibit a strong, consistent tendency to increase proportionally.

Q19. What is Simple Linear Regression? Explain its meaning, features and applications with suitable examples.

1. Core Concept & Definition

Simple Linear Regression (SLR) is a fundamental supervised parametric statistical modeling technique used to quantify and predict the linear relationship between a single continuous **independent predictor variable (X)** and a continuous **dependent response variable (Y)** using Ordinary Least Squares (OLS).

2. Mathematical Model & Geometry

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Estimated Sample Equation: $\hat{Y} = a + bX$
- Slope Coefficient: $b = r\frac{\sigma_y}{\sigma_x} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$
- Intercept: $a = \bar{Y} - b\bar{X}$

```
Y (Response)
 ^              .  * Actual Data Point (Y)
 |            .   /|
 |          .    / | Residual e = Y - Yhat
 |        .     *  v (Predicted Point Yhat)
 |      .  Line of Best Fit: Yhat = a + bX
 +-----------------------------------------> X (Predictor)
```

3. Key Features and Statistical Properties

- **OLS Criterion:** Minimizes the sum of squared vertical residuals: $\sum e_i^2 = \sum(Y_i - \hat{Y}_i)^2 = \mathbf{Minimum}$.
- **Centroid Intersection:** The regression line strictly passes through the mean coordinates $(\bar{X}, \bar{Y})$.
- **Asymmetry:** Unlike correlation ($r_{xy} = r_{yx}$), regression is directional; regressing Y on X differs mathematically from X on Y.
- **Goodness of Fit (R^2):** Measures the proportion of total variance in Y explained by X ($R^2 = r^2$, bounded between 0 and 1).
- **Gauss-Markov Assumptions:** Errors have zero mean, constant variance (homoscedasticity), normality, and independence.
- **Predictive Forecasting:** Enables quantitative estimation of unknown target values for specified feature inputs.

4. Real-World Applications & Concrete Example

- **Applications:** Sales forecasting from ad spend, crop yield estimation from rainfall, and house pricing from square footage.
- **Example:** If a sales model is $\hat{Y} = 40 + 3X$ (sales Y and ad spend X in Lakhs), spending $X = 15$ Lakhs yields $\hat{Y} = 40 + 3(15) = Rs.\,85\ \text{Lakhs}$.

Q20. A Normal variate has mean 60 and standard deviation 10. Find the probability that X lies: (i)

between 50 and 75, (ii) above 80.

1. Given Data

- Population Mean $\mu = 60$.
- Population Standard Deviation $\sigma = 10$.
- Random Variable $X \sim N(\mu = 60, \sigma^2 = 100)$.
- Standard Normal Transformation Formula:

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 60}{10}$$

2. Case (i): Probability that X lies between 50 and 75 [$P(50 \leq X \leq 75)$]

1. **Transform endpoints to Standard Normal Variates (Z):**
2. For $X_1 = 50$: $Z_1 = \frac{50-60}{10} = \frac{-10}{10} = -1.0$
3. For $X_2 = 75$: $Z_2 = \frac{75-60}{10} = \frac{+15}{10} = +1.5$
4. **Formulate Probability Statement:**

$$P(50 \leq X \leq 75) = P(-1.0 \leq Z \leq +1.5)$$

5. **Decompose Area under Normal Curve (Symmetry Property):**

$$P(-1.0 \leq Z \leq +1.5) = \text{Area from } Z = -1.0 \text{ to } 0 + \text{Area from } Z = 0 \text{ to } +1.5$$

$$= \text{Area}(0 \leq Z \leq 1.0) + \text{Area}(0 \leq Z \leq 1.5)$$

6. **Values from Standard Normal Distribution Table:**
7. $\text{Area}(0 \leq Z \leq 1.0) = 0.3413$
8. $\text{Area}(0 \leq Z \leq 1.5) = 0.4332$
9. **Compute Total Probability:**

$$P(50 \leq X \leq 75) = 0.3413 + 0.4332 = 0.7745 \quad (\text{or } 77.45\%)$$

3. Case (ii): Probability that X lies above 80 [$P(X > 80)$]

1. **Transform $X = 80$ to Standard Normal Variate (Z):**
2. For $X = 80$: $Z = \frac{80-60}{10} = \frac{+20}{10} = +2.0$
3. **Formulate Probability Statement:**

$$P(X > 80) = P(Z > +2.0)$$

4. **Compute Tail Area (Total right half area = 0.5):**

$$P(Z > +2.0) = 0.5000 - \text{Area}(0 \leq Z \leq 2.0)$$

5. **Value from Standard Normal Distribution Table:**
6. $\text{Area}(0 \leq Z \leq 2.0) = 0.4772$

7. **Compute Tail Probability:**

$$P(X > 80) = 0.5000 - 0.4772 = 0.0228 \quad (\text{or } 2.28\%)$$

4. Summary Table & Boxed Final Answers

Condition	Z-Score Range	Area Calculation	Final Probability
(i) Between 50 and 75	$-1.0 \le Z \le +1.5$	$0.3413 + 0.4332$	0.7745 (77.45%)
(ii) Above 80	$Z > +2.0$	$0.5000 - 0.4772$	0.0228 (2.28%)

$$\boxed{\text{(i) } P(50 \le X \le 75) = 0.7745 \quad (77.45\%)}$$

$$\boxed{\text{(ii) } P(X > 80) = 0.0228 \quad (2.28\%)}$$

> **Q21. A manufacturer claims that the average life of a battery is 50 hours. A random sample of 16 batteries has an average life of 54 hours with a standard deviation of 8 hours. Test whether the average battery life is significantly different from 50 hours at the 5% level of significance. (Given: Critical t-value for 15 degrees of freedom at the 5% level of significance for a two-tailed test = 2.131).**

1. Given Data

- Hypothesized Population Mean $\mu_0 = 50 \text{ hours}$.
- Sample Size $n = 16$ ($n < 30 \implies \text{Small Sample Student's } t\text{-test}$).
- Sample Mean $\bar{x} = 54 \text{ hours}$.
- Sample Standard Deviation $s = 8 \text{ hours}$.
- Level of Significance $\alpha = 0.05$ (5%).
- Degrees of Freedom $df = n - 1 = 16 - 1 = 15$.
- Two-tailed Critical Value: $t_{\text{critical}} = 2.131$.

2. Hypothesis Formulation

- **Null Hypothesis (H_0):** $\mu = 50$
  (The average battery life is not significantly different from the manufacturer's claim of 50 hours).
- **Alternative Hypothesis (H_1):** $\mu \neq 50$
  (Two-tailed test: The average battery life is significantly different from 50 hours).

3. Test Statistic Formula (Student's Single-Sample t-Test)

$$t = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

4. Step-by-Step Calculation

1. **Calculate Standard Error ($SE_{\bar{x}}$):**

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{8}{\sqrt{16}} = \frac{8}{4} = 2.0 \text{ hours}$$

2. **Compute Calculated t-Statistic:**

$$t_{\text{calculated}} = \frac{54 - 50}{2.0} = \frac{4}{2.0} = +2.000$$

$$|t_{\text{calculated}}| = 2.000$$

5. Statistical Decision Rule & Conclusion

- **Decision Rule:** For a two-tailed test at $\alpha = 0.05$ with $df = 15$:
- If $|t_{\text{calculated}}| > t_{\text{critical}} \implies$ Reject H_0.
- If $|t_{\text{calculated}}| \leq t_{\text{critical}} \implies$ Fail to reject (Accept) H_0.
- **Comparison:**

$$|t_{\text{calculated}}| = 2.000 < t_{\text{critical}} = 2.131$$

- **Statistical Inference:** Since the absolute calculated value of t (2.000) is strictly less than the critical tabulated value (2.131), we fail to reject (accept) the Null Hypothesis (H_0) at the 5% significance level.
- **Final Conclusion:** There is no statistically significant evidence to reject the manufacturer's claim. The average battery life is not significantly different from 50 hours. The observed sample difference of 4 hours ($54 - 50$) is attributable purely to random sampling fluctuation.

6. Boxed Final Answer

$$\boxed{|t_{\text{calculated}}| = 2.000 < t_{\text{critical}} = 2.131 \implies \text{Accept } H_0 \text{ (Difference is Not Significant)}}$$

Q22. Explain and Compare Power BI Desktop, Power BI Service, and Power BI Mobile Apps.

1. Core Architecture & Organizational Roles

Microsoft Power BI delivers a complete enterprise business intelligence platform structured into three integrated components: **Desktop** (authoring), **Service** (cloud sharing & governance), and **Mobile** (on-the-go consumption).

```
[Power BI Desktop]  ===(Publish)===>  [Power BI Service]  ===(Sync)===>  [Power BI Mobile]
(Authoring & Modeling)                 (Cloud SaaS Sharing)              (KPI Consumption)
```

2. Comprehensive Comparison Table

Feature	Power BI Desktop	Power BI Service (Cloud)	Power BI Mobile Apps
Deployment	Free Windows desktop app.	Cloud SaaS via browser.	iOS, Android, Windows app.
Primary Role	Data modeling, ETL & report authoring.	Report sharing, governance & collaboration.	KPI monitoring & mobile exploration.
Target User	BI developers & data analysts.	Business managers & consumers.	Executives & field workforce.
Power Query ETL	Full M engine transformation capabilities.	Limited (Dataflows creation only).	None (Read-only access).
Data Modeling	Full star schemas, relationships & DAX.	Minimal web modeling; no complex DAX.	None (Cannot alter models).
Dashboards	Not supported (Reports only, .pbix).	Full support (Pins visual tiles to dashboards).	Responsive phone dashboard views.
Data Refresh	Manual refresh on local PC.	Automated scheduled refresh via Gateways.	Consumes refreshed cloud data.
Security	Local file saving; no sharing.	Workspaces, apps, Row-Level Security (RLS).	Biometric lock, push metric alerts.

3. Key Operational Highlights

- **Desktop:** Performs heavy data extraction, column cleaning, schema relationships, and advanced DAX measure authoring.
- **Service:** Manages central governance, automated gateway refresh cycles, and multi-department workspace collaboration.
- **Mobile:** Supports low-latency executive decisions via touchscreen KPI cards, offline caching, and threshold alerts.

4. Concrete Example

- **Example:** A developer models data in Desktop and publishes it to Service; regional managers track live sales cards on their smartphones via Mobile.

Q23. (a) Explain different applications of Data Analytics. [4 Marks]

1. Core Concept & Value

Data Analytics (DA) transforms raw quantitative data into actionable strategic intelligence, enabling enterprises across industries to optimize operations, mitigate risks, and maximize profitability.

2. Major Industry Applications

- **Banking & Finance (BFSI):** Real-time credit scoring models evaluate loan default risks; anomaly detection algorithms inspect transaction velocity and geolocation to block fraudulent credit card charges.
- **Healthcare & Medicine:** Predictive diagnostics analyze patient vital trends for early sepsis warnings; hospital analytics optimize ICU bed allocation and medical staffing schedules.
- **Retail & E-Commerce:** Dynamic pricing algorithms optimize profit margins based on demand; recommendation engines personalize user feeds to drive cross-selling.
- **Manufacturing & Supply Chain:** Predictive maintenance analyzes industrial IoT sensor telemetry (vibration, heat) to service machinery before breakdown, minimizing factory downtime.
- **Telecommunications:** Churn prediction algorithms identify subscribers with declining usage, triggering automated retention discounts.
- **Example:** E-commerce platforms deploy collaborative filtering recommendation engines to personalize product recommendations, boosting average order values by 18%.

Q23. (b) Explain the Extract-Transform-Load (ETL) process in Power BI. [4 Marks]

1. Core Concept

In Microsoft Power BI, the **Extract-Transform-Load (ETL)** pipeline is the data preparation process managed by **Power Query (M language)**. It ingests messy raw data, cleanses and shapes it, and loads it into the in-memory **VertiPaq** columnar engine.

```
[Raw Sources] ---> [Extract: Connectors] ---> [Transform: Power Query] ---> [Load: VertiPaq Model]
```

2. The Three ETL Stages

- **1. Extract (Get Data):** Ingests raw data from over 150 sources (Excel, CSV, SQL Server, Azure, REST APIs), previewing schemas and importing tables.
- **2. Transform (Power Query Editor):** Cleanses, standardizes, and reshapes data without altering the underlying source. Common tasks include removing nulls/duplicates, splitting columns, unpivoting cross-tabs, changing data types, and merging/appending tables.
- **3. Load (Close & Apply):** Commits the curated data into Power BI's tabular model. VertiPaq compresses tables into columnar RAM structures, enabling fast DAX calculations and visual reporting.
- **Example:** Connecting to 12 monthly sales CSV files, removing null rows via Power Query, and loading one consolidated table into VertiPaq.

Model Question Paper - 2 | Section C Solutions (Exam-Length Edition)

Total Marks: 40 (Answer any 5 out of 7 Questions, 8 Marks each — All 7 Solved for Complete Preparation)
Format: Strictly adheres to 15-minute exam writing limits (~250–350 words for full 8-mark theory questions; ~120–150 words each for 4M sub-questions). Numericals include Given Data, Exact Formulas, Complete Step-by-Step Calculation Tables, Statistical Decision Rules, and Boxed Final Answers. Strictly NO memory tips.

MODEL QUESTION PAPER – 2

Q17. Explain the types of Data Analytics with examples.

Answer:

1. Core Concept & Definition

Data Analytics is structured into four progressive evolutionary tiers characterized by increasing algorithmic complexity and business value: **Descriptive**, **Diagnostic**, **Predictive**, and **Prescriptive Analytics**. Together, they form an analytical maturity continuum from retrospective reporting to forward-looking decision optimization.

2. Comparative Taxonomy Table

Analytics Type	Question Answered	Temporal Horizon	Key Statistical / ML Techniques
Descriptive	*What happened?*	Historical (Past)	Data aggregation, summary statistics, KPI reports
Diagnostic	*Why did it happen?*	Historical (Root cause)	Drill-down, anomaly detection, correlation analysis
Predictive	*What will happen?*	Future Forecast	Regression, time-series forecasting, classification
Prescriptive	*What should we do?*	Future Optimization	Linear programming, simulation, heuristic optimization

3. In-Depth Explanation & Key Points

- **Descriptive Analytics (Historical Baseline):** Summarizes historical data using metrics (mean, median, standard deviation) and visual reports to provide operational visibility without inferring causes.
- **Diagnostic Analytics (Root-Cause Analysis):** Explores anomalies using drill-down and correlation matrices to isolate why specific outcomes deviated from expectations.

- **Predictive Analytics (Future Forecasting):** Uses statistical models, regression, and machine learning to forecast future trends and probabilities from past patterns.
- **Prescriptive Analytics (Action Optimization):** Evaluates alternative decision paths using optimization algorithms and simulation models to recommend actions that maximize returns.
- **Algorithmic Progression:** Advances from basic arithmetic summaries (descriptive) to inferential statistics (diagnostic), predictive machine learning, and mathematical optimization solvers.
- **Business Value Generation:** Higher tiers automate decisions and directly drive competitive business advantages.

4. End-to-End Concrete Example: Retail Chain

- **Descriptive:** Dashboard reports festive sales fell \%$ year-over-year.
- **Diagnostic:** Drill-down reveals an out-of-stock issue for top 5 electronic products.
- **Predictive:** Forecasting models project a \%$ demand surge for upcoming holidays.
- **Prescriptive:** Supply chain solver automatically issues reorders to prevent stockouts.

Q18. Calculate Rank correlation from the following data:

X	60	34	40	50	45	41	22	43	42	66
Y	75	32	35	40	46	33	45	50	43	42

Solution:

1. Given Data & Identification

- Number of paired observations: $N = 10$.
- Both series X and Y contain distinct values with **no tied (repeated) ranks**.
- **Formula:** Spearman's Rank Correlation Coefficient without ties is given by:

$$\rho = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

(Where R_X and R_Y are ranks assigned in descending order, $d = R_X - R_Y$, and $\sum d = 0$).

2. Ranking Scheme

- Ranks for X (Descending order):
  $66 \to 1,\ 60 \to 2,\ 50 \to 3,\ 45 \to 4,\ 43 \to 5,\ 42 \to 6,\ 41 \to 7,\ 40 \to 8,\ 34 \to 9,\ 22 \to 10$.
- Ranks for Y (Descending order):
  $75 \to 1,\ 50 \to 2,\ 46 \to 3,\ 45 \to 4,\ 43 \to 5,\ 42 \to 6,\ 40 \to 7,\ 35 \to 8,\ 33 \to 9,\ 32 \to 10$.

3. Step-by-Step Calculation Table

Sl. No.	X	Rank R_X	Y	Rank R_Y	$d = R_X - R_Y$	d^2
1	60	2	75	1	$2 - 1 = +1$	1
2	34	9	32	10	$9 - 10 = -1$	1
3	40	8	35	8	$8 - 8 = 0$	0
4	50	3	40	7	$3 - 7 = -4$	16
5	45	4	46	3	$4 - 3 = +1$	1
6	41	7	33	9	$7 - 9 = -2$	4
7	22	10	45	4	$10 - 4 = +6$	36
8	43	5	50	2	$5 - 2 = +3$	9
9	42	6	43	5	$6 - 5 = +1$	1
10	66	1	42	6	$1 - 6 = -5$	25
Total	-	-	-	-	$\sum d = 0$	$\sum d^2 = 94$

Arithmetic Verification: $\sum d = (+1 - 1 + 0 - 4 + 1 - 2 + 6 + 3 + 1 - 5) = 0$. (Rank differences sum to zero, verifying ranking consistency).

4. Computation of Rank Correlation Coefficient (ρ)

Substitute $\sum d^2 = 94$ and $N = 10$ into Spearman's formula:

$$\rho = 1 - \frac{6 \times 94}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{564}{10(100 - 1)}$$

$$\rho = 1 - \frac{564}{10 \times 99}$$

$$\rho = 1 - \frac{564}{990}$$

$$\rho = 1 - 0.5697 = +0.4303$$

5. Statistical Interpretation & Final Boxed Answer

The calculated rank correlation coefficient is $+0.4303$, indicating a **moderate positive monotonic relationship** between variables X and Y.

$$\rho = +0.4303 \quad \left(\text{or } \frac{71}{165} \right)$$

Q19. The following information about Advertising (AD) and Sales is given:

Particulars	AD Expenditure (X), Rs. Lakhs	Sales (Y), Rs. Lakhs
Mean	$\bar{X} = 10$	$\bar{Y} = 90$
Standard Deviation	$\sigma_X = 3$	$\sigma_Y = 12$
Correlation Coefficient (r) = 0.8		

Find: 1. The two regression equations. 2. The likely sales when AD expenditure is ₹15 lakhs. 3. The AD expenditure required if the company wants to attain a sales target of ₹120 lakhs.

Solution:

1. Identification of Given Parameters

- Mean of X: $\bar{X} = 10$
- Mean of Y: $\bar{Y} = 90$
- Standard deviation of X: $\sigma_X = 3$
- Standard deviation of Y: $\sigma_Y = 12$
- Pearson's correlation coefficient: $r = 0.8$

2. Sub-Part (i): Obtain the Two Regression Equations

A. Regression Equation of Y on X (Used to estimate Y for given X)

- **Regression Coefficient of Y on X (b_{yx}):**

$$b_{yx} = r \times \frac{\sigma_Y}{\sigma_X} = 0.8 \times \frac{12}{3} = 0.8 \times 4 = 3.2$$

- **Equation:**

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 90 = 3.2(X - 10)$$

$$Y - 90 = 3.2X - 32$$

$$Y = 3.2X + 58$$

B. Regression Equation of X on Y (Used to estimate X for given Y)

- **Regression Coefficient of X on Y (b_{xy}):**

$$b_{xy} = r \times \frac{\sigma_X}{\sigma_Y} = 0.8 \times \frac{3}{12} = 0.8 \times 0.25 = 0.2$$

- **Equation:**

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 10 = 0.2(Y - 90)$$

$$X - 10 = 0.2Y - 18$$

$$X = 0.2Y - 8$$

Consistency Verification: $b_{yx} \times b_{xy} = 3.2 \times 0.2 = 0.64 = r^2 = (0.8)^2 = 0.64$. (Coefficients satisfy statistical properties).

3. Sub-Part (ii): Likely Sales when AD Expenditure is ₹15 Lakhs ($X = 15$)

Substitute $X = 15$ into the regression equation of Y on X:

$$Y = 3.2(15) + 58$$

$$Y = 48 + 58 = 106 \text{ lakhs}$$

* **Conclusion:** The likely sales revenue is **₹106 Lakhs**.

4. Sub-Part (iii): AD Expenditure Required for Sales Target of ₹120 Lakhs ($Y = 120$)

Substitute $Y = 120$ into the regression equation of X on Y:

$$X = 0.2(120) - 8$$

$$X = 24 - 8 = 16 \text{ lakhs}$$

* **Conclusion:** The required advertising outlay is **₹16 Lakhs**.

5. Final Boxed Answers

Equation of Y on X:	$Y = 3.2X + 58$
Equation of X on Y:	$X = 0.2Y - 8$
Likely Sales for $X = 15$:	Rs. 106 Lakhs
Required AD for $Y = 120$:	Rs. 16 Lakhs

> **Q20. The number of calls received by an office follows a Poisson Distribution with mean 3 calls per minute. Find the probability of:**

1. No call
2. Exactly 2 calls
3. More than 3 calls during a minute.
   (Use $e^{-3} = 0.0498$)

Solution:

1. Given Parameters & PMF

- Let X denote the number of telephone calls received per minute.
- X follows a Poisson distribution with parameter $\lambda = 3$ calls/minute.
- **Poisson Probability Mass Function (PMF):**

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad \text{for } k = 0, 1, 2, \ldots$$

- Given constant: $e^{-3} = 0.0498$.

2. Sub-Part (i): Probability of No Call ($X = 0$)

Substitute $k = 0$ into the PMF:

$$P(X = 0) = \frac{e^{-3} \times 3^0}{0!}$$

Since $3^0 = 1$ and $0! = 1$:

$$P(X = 0) = \frac{0.0498 \times 1}{1} = 0.0498$$

3. Sub-Part (ii): Probability of Exactly 2 Calls ($X = 2$)

Substitute $k = 2$ into the PMF:

$$P(X = 2) = \frac{e^{-3} \times 3^2}{2!}$$

Since $3^2 = 9$ and $2! = 2 \times 1 = 2$:

$$P(X = 2) = \frac{0.0498 \times 9}{2} = \frac{0.4482}{2} = 0.2241$$

4. Sub-Part (iii): Probability of More Than 3 Calls ($X > 3$)

Using the complement rule of probability:

$$P(X > 3) = 1 - P(X \leq 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

- **Evaluate Individual Terms:**
- $P(X = 0) = 0.0498$
- $P(X = 1) = \frac{e^{-3} \times 3^1}{1!} = \frac{0.0498 \times 3}{1} = 0.1494$
- $P(X = 2) = 0.2241$
- $P(X = 3) = \frac{e^{-3} \times 3^3}{3!} = \frac{0.0498 \times 27}{6} = 0.0498 \times 4.5 = 0.2241$
- **Sum Cumulative Probability $P(X \leq 3)$:**

$$P(X \leq 3) = 0.0498 + 0.1494 + 0.2241 + 0.2241 = 0.6474$$

- **Calculate Complement Probability:**

$$P(X > 3) = 1 - 0.6474 = 0.3526$$

5. Final Boxed Answers

$$P(\text{No Call}) = 0.0498$$

$$P(\text{Exactly 2 Calls}) = 0.2241$$

$$P(\text{More Than 3 Calls}) = 0.3526$$

> **Q21. A college wants to compare the performance of students taught using two different teaching methods. The following information is obtained:**

Particulars	Method A	Method B
Number of Students	$n_1 = 10$	$n_2 = 10$
Mean Marks	$\bar{X}_1 = 74$	$\bar{X}_2 = 68$
Standard Deviation	$s_1 = 5$	$s_2 = 4$

Test whether there is a significant difference between the mean marks of the two groups at the 5% level of significance. Assume that the population variances are equal.
(Given: Critical t-value for 18 degrees of freedom at the 5% level of significance for a two-tailed test = 2.101).

Solution:

1. Formulation of Statistical Hypotheses

- **Null Hypothesis (H_0):** $\mu_1 = \mu_2$
  (There is no statistically significant difference between the mean performance of students taught using Method A and Method B).
- **Alternative Hypothesis (H_1):** $\mu_1 \neq \mu_2$
  (There is a significant difference between the two teaching methods — two-tailed test).

2. Degrees of Freedom (df)

$$df = n_1 + n_2 - 2 = 10 + 10 - 2 = 18$$

3. Pooled Sample Variance (s_p^2) & Standard Error (SE)

Since population variances are assumed equal ($\sigma_1^2 = \sigma_2^2$), we compute the pooled sample variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Substitute given values:

$$s_p^2 = \frac{(10 - 1)(5^2) + (10 - 1)(4^2)}{10 + 10 - 2} = \frac{9(25) + 9(16)}{18}$$

$$s_p^2 = \frac{225 + 144}{18} = \frac{369}{18} = 20.5$$

- **Standard Error of the Difference Between Means (SE):**

$$SE = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{20.5 \left(\frac{1}{10} + \frac{1}{10} \right)}$$

$$SE = \sqrt{20.5 \times \frac{2}{10}} = \sqrt{20.5 \times 0.2} = \sqrt{4.1} \approx 2.0248$$

4. Computation of Test Statistic ($t_{\text{calculated}}$)

$$t_{\text{cal}} = \frac{\bar{X}_1 - \bar{X}_2}{SE} = \frac{74 - 68}{2.0248} = \frac{6}{2.0248} \approx 2.9632$$

5. Statistical Decision Rule & Conclusion

- **Calculated Value:** $|t_{\text{cal}}| = 2.9632$
- **Critical / Table Value (t_{tab}):** At $\alpha = 0.05$ for $df = 18$ (two-tailed test), $t_{\text{crit}} = 2.101$.

$$|t_{\text{cal}}| = 2.9632 > t_{\text{crit}} = 2.101$$

- **Decision:** Since the absolute calculated t-statistic exceeds the critical value, we reject the null hypothesis (H_0) at the 5% level of significance.
- **Conclusion:** There is a statistically significant difference between the mean marks of students taught by the two methods. Students taught under Method A ($\bar{X}_1 = 74$) performed significantly better than those under Method B ($\bar{X}_2 = 68$).

6. Final Boxed Summary

$$t_{\text{calculated}} = 2.9632, \quad t_{\text{critical}} = 2.101$$

$$\text{Decision: Reject } H_0 \quad \text{(Significant difference exists)}$$

Q22. Explain Bar Chart, Column Chart, Pie Chart, and Line Chart and state their uses.

Answer:

1. Core Concept & Definition

Data visualization charts translate quantitative datasets into intuitive graphical representations. Selecting the appropriate chart depends on whether the analytical goal is discrete category comparison, proportional part-to-whole composition, or continuous temporal trend analysis.

2. Visual Comparative Taxonomy

Chart Type	Visual Orientation	Analytical Purpose	Key Strength
Bar Chart	Horizontal bars ($-axis categories)	Comparing discrete categorical groups	Ideal for long text labels
Column Chart	Vertical columns ($-axis categories)	Comparing items over discrete intervals	Natural left-to-right flow
Pie Chart	Circular radial slices	Showing part-to-whole proportions	Instantly displays percentage shares
Line Chart	Connected line vertices	Continuous time-series trend analysis	Visualizes continuity and momentum

3. In-Depth Explanation & Key Uses

- **Bar Chart (Horizontal Design):** Displays category labels on the vertical axis with horizontal lengths representing metric values.
- *Uses:* Ideal for comparing items with lengthy text descriptions (e.g., department names) or displaying ranked leaderboards (e.g., top 10 sales executives).
- **Column Chart (Vertical Design):** Plots discrete categories along the horizontal $-axisandvaluesalongthevertical$-axis.
- *Uses:* Comparing discrete metrics across categories (e.g., revenue across four regions) or discrete periodic comparisons (e.g., quarterly profits).
- **Pie Chart (Proportional Composition):** A circular chart divided into radial slices totaling strictly \%$.
- *Uses:* Visualizing high-level proportional shares with few categories (< 5), such as gender demographics or asset allocations.
- **Line Chart (Continuous Trends):** Connects chronological data points with continuous line segments.
- *Uses:* Tracking continuous trends, fluctuations, and velocity over time (e.g., monthly sales growth, annual GDP, or stock prices).

4. Practical Example

- In an Executive Sales Dashboard, a **Column Chart** compares regional sales, a Horizontal Bar Chart ranks sales reps, a Pie Chart shows category market share, and a Line Chart tracks monthly revenue growth.

Q23.

(a) Differentiate between descriptive, diagnostic, predictive and prescriptive analytics. [4 Marks]

(b) Explain the Get Data facility and the procedure for selecting a data source in Power BI. [4 Marks]

Answer:

(a) Differentiate between Descriptive, Diagnostic, Predictive and Prescriptive Analytics [4 Marks]

1. Comparative Matrix

Parameter	Descriptive	Diagnostic	Predictive	Prescriptive
Question	*What happened?*	*Why did it happen?*	*What will happen?*	*What should we do?*
Horizon	Past / Historical	Past / Root-cause	Future Forecast	Future Optimization
Technique	Aggregations, KPIs	Drill-down, OLAP	Regression, ML	Linear programming
Outcome	Static reports	Identified drivers	Probabilistic trends	Optimal action paths

2. Key Distinctions

- **Analytical Focus:** Descriptive summarizes past data; Diagnostic explains variances; Predictive models future likelihoods; Prescriptive identifies optimal choices.
- **Complexity & Value:** Advances from basic summary statistics to machine learning and mathematical optimization, delivering exponential business value.
- **Example:** Descriptive shows sales dropped 10%; Diagnostic attributes it to shipping delays; Predictive forecasts stockouts; Prescriptive reroutes supply logistics.

(b) Get Data Facility & Procedure for Selecting a Data Source in Power BI [4 Marks]

1. Meaning & Core Connectors

The **Get Data** facility is Power BI Desktop's primary ingestion gateway, providing over 150 pre-built connectors to import or query data from: * **Files:** Excel (.xlsx), CSV/Text, JSON, XML, Folders. * **Databases:** SQL Server, Oracle, MySQL, PostgreSQL. * **Cloud & Services:** Azure SQL, SharePoint, Salesforce, Web URLs.

2. Step-by-Step Procedure

1. **Open Gateway:** On the Home Tab, click Get Data in the *Data Group*.
2. **Choose Source:** Select the connector category and click Connect.
3. **Configure Connection:** Enter server/filepath details and choose storage mode (Import or DirectQuery).
4. **Authenticate & Load:** Enter credentials, preview tables in the Navigator, and click Load (or Transform Data for Power Query cleaning).

Model Question Paper - 3 | Section C Solutions (Exam Length)

Exam Standard: 15-minute writing capacity per question (strictly 250–350 words for 8-mark theory; ~120–150 words each for 4-mark sub-questions; rigorous step-by-step arithmetic, ANOVA tables, decision rules, and boxed answers for numericals). Strictly NO Memory Tips.

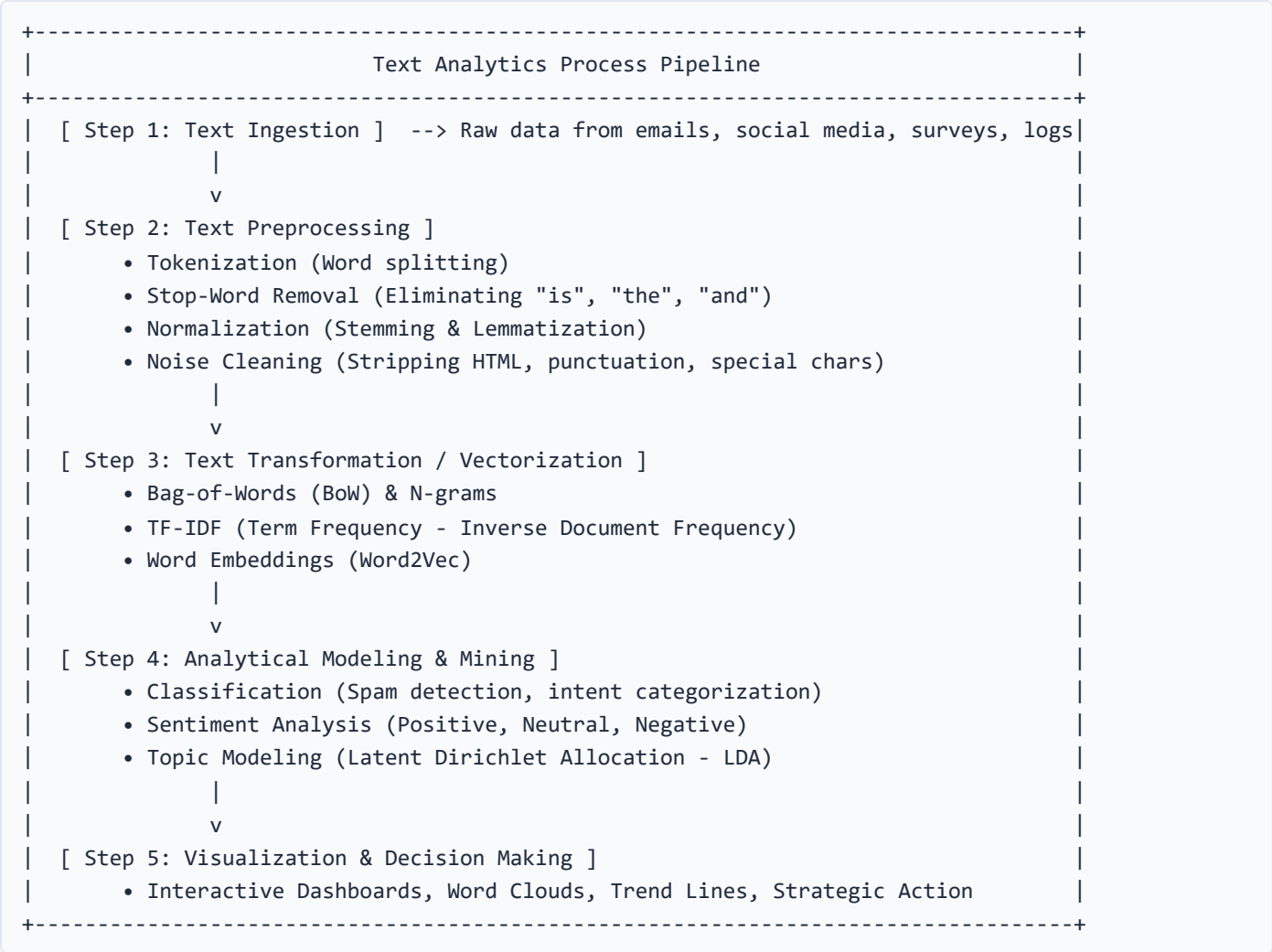
MODEL QUESTION PAPER – 3

Q17. Explain the complete text analytics process with a neat diagram. (8 Marks)

1. Core Concept and Definition

Text Analytics (or *Text Mining*) is the automated process of converting large volumes of unstructured textual data into quantitative, structured business intelligence. It integrates **Natural Language Processing (NLP)**, data mining, and machine learning to discover hidden semantic patterns, thematic trends, and sentiment polarity.

2. Process Flowchart Diagram



3. Key Phases of the Process

- **Data Ingestion & Collection:** Gathers raw, unstructured text from internal and external repositories (CRM tickets, Twitter feeds, PDF documents, online customer reviews).
- **Tokenization:** Segments continuous streams of written text into discrete linguistic units called tokens (individual words, n-grams, or phrases).
- **Text Cleansing & Stop-Word Removal:** Eliminates non-informative noise including HTML/XML tags, punctuation, numbers, and high-frequency stopwords (*a, the, in, at, is*).
- **Morphological Normalization (Stemming & Lemmatization):** Reduces inflectional variants to base dictionary forms (e.g., *analyzing, analyzed, analyzer* $\rightarrow$ *analyze*) using Porter Stemmer or WordNet Lemmatizer.
- **Feature Extraction / Vectorization:** Converts preprocessed text strings into structured numerical matrices using algorithms like $\mathbf{TF\text{-}IDF}$, representing term importance relative to the entire document collection.
- **Pattern Discovery & Machine Learning Modeling:** Applies algorithms for sentiment polarity scoring, document clustering (K-Means), automated topic modeling (LDA), and named entity recognition (NER).
- **Interpretation & Business Visualization:** Surfaces quantitative insights using word clouds, sentiment dashboards, and automated alert systems to guide executive decision-making.

4. Real-World Example

- **Example:** An airline ingesting 15,000 monthly passenger reviews, using NLP preprocessing and sentiment modeling to identify baggage-handling complaints, and surfacing alerts on an executive Power BI dashboard.

Q18. From the following data:

X	10	12	13	17	18
Y	5	6	7	9	13

(i) Calculate the two regression equations.
(ii) Estimate the value of X when Y = 20.
(iii) Determine the value of the correlation coefficient through regression coefficients. (8 Marks)

1. Given Data and Arithmetic Means

Number of observations: $N = 5$.

- **Mean of X:**

$$\sum X = 10 + 12 + 13 + 17 + 18 = 70 \implies \bar{X} = \frac{\sum X}{N} = \frac{70}{5} = 14$$

- **Mean of Y:**

$$\sum Y = 5 + 6 + 7 + 9 + 13 = 40 \implies \bar{Y} = \frac{\sum Y}{N} = \frac{40}{5} = 8$$

Since both means are integers, we use the **Actual Mean Deviation Method** ($x = X - 14$, $y = Y - 8$).

2. Calculation Table

X	$x = X - 14$	x^2	Y	$y = Y - 8$	y^2	$xy = x \cdot y$
10	-4	16	5	-3	9	$(-4) \times (-3) = +12$
12	-2	4	6	-2	4	$(-2) \times (-2) = +4$
13	-1	1	7	-1	1	$(-1) \times (-1) = +1$
17	$+3$	9	9	$+1$	1	$(+3) \times (+1) = +3$
18	$+4$	16	13	$+5$	25	$(+4) \times (+5) = +20$
Total	$\sum x = 0$	$\sum x^2 = 46$	Total	$\sum y = 0$	$\sum y^2 = 40$	$\sum xy = 40$

3. Step (i): Calculation of the Two Regression Equations

- **Regression Coefficient of Y on X (b_{yx}):**

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{40}{46} = \frac{20}{23} \approx 0.8696$$

- **Regression Equation of Y on X:**

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 8 = \frac{20}{23}(X - 14)$$

$$23(Y - 8) = 20(X - 14)$$

$$23Y - 184 = 20X - 280$$

$$20X - 23Y - 96 = 0 \quad \text{or} \quad Y = 0.8696X - 4.1739$$

- **Regression Coefficient of X on Y (b_{xy}):**

$$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{40}{40} = 1.00$$

- **Regression Equation of X on Y:**

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 14 = 1.0(Y - 8)$$

$$X = Y + 6 \quad \text{or} \quad X - Y - 6 = 0$$

4. Step (ii): Estimate the Value of X when $Y = 20$

To predict the dependent variable X from an independent variable Y, we use the **Regression Equation of X on Y**:

$$X = Y + 6$$

Substitute $Y = 20$:

$$X = 20 + 6 = 26$$

$$\boxed{\text{Estimated } X = 26 \quad (\text{when } Y = 20)}$$

5. Step (iii): Determine Correlation Coefficient (r) Through Regression Coefficients

The coefficient of correlation is the geometric mean of the two regression coefficients:

$$r = \pm\sqrt{b_{yx} \cdot b_{xy}}$$

Since both $b_{yx} = +0.8696$ and $b_{xy} = +1.00$ are positive, r is positive:

$$r = +\sqrt{\frac{20}{23} \times 1.0} = \sqrt{\frac{20}{23}} = \sqrt{0.869565} \approx +0.9325$$

$$\boxed{r = +0.9325 \quad \text{(High Degree of Positive Correlation)}}$$

Q19. Explain Spearman's Rank Correlation Coefficient. State the situations in which this method is suitable. (8 Marks)

1. Core Concept and Definition

Spearman's Rank Correlation Coefficient (ρ or r_s), developed by British psychologist Charles Spearman in 1904, is a non-parametric statistical measure of the strength and direction of monotonic association between two ranked or ordinal variables.

Unlike Karl Pearson's correlation, Spearman's rank correlation does not assume that data is normally distributed, nor does it require a strictly linear relationship; it requires only that the relationship be **monotonic**.

2. Mathematical Formulas and Range

- **Case 1: When Ranks are Non-Tied (Distinct Values):**

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

  where:

- $d = R_X - R_Y$ (Difference between paired ranks for each individual item)
- n = Number of paired observations
- **Case 2: When Ranks are Tied (Repeated Values):**

$$\rho = 1 - \frac{6 \left[\sum d^2 + \sum \frac{m(m^2-1)}{12} \right]}{n(n^2 - 1)}$$

  where m is the number of times a particular rank repeats (tied count), and tied items receive the arithmetic mean of the rank positions they occupy.

- **Bounded Range:** $-1.00 \leq \rho \leq +1.00$ (-1 = Perfect Negative Monotonic Association, $+1$ = Perfect Positive Monotonic Association, 0 = No Monotonic Association).

3. Situations in Which Spearman's Method is Suitable

- **Qualitative and Subjective Attributes:** Highly suitable when variables cannot be measured on an exact quantitative scale but can be sorted or ranked hierarchically (e.g., assessing beauty, intelligence, morality, taste, or leadership qualities).
- **Non-Normal and Heavily Skewed Data:** Indispensable when sample distributions violate the assumption of normality, as rank transformations neutralize distribution skewness.
- **Presence of Extreme Outliers:** Resistant to extreme outliers because converting raw metric values into discrete sequential ranks ($1, 2, \ldots, n$) bounds the distorting impact of extreme data points.
- **Monotonic Non-Linear Relationships:** Measures whether variables increase or decrease together, even when the underlying mathematical function is non-linear (e.g., exponential or logarithmic curves).
- **Ordinal Scale Survey Data:** Ideal for analyzing Likert-scale questionnaire responses (*Strongly Agree* to *Strongly Disagree*) in social science and market research.
- **Judge and Inter-Rater Reliability Testing:** Standard tool for evaluating the level of agreement between two independent adjudicators evaluating contestants in competitions.
- **Directly Ranked Input Data:** Directly applicable when input data is pre-ranked by observers without raw metric readings.

4. Real-World Example

- **Example:** Two human resource managers independently evaluating and ranking 10 job candidates from 1 to 10 based on interview performance; Spearman's ρ measures the consistency of their hiring judgment.

> **Q20. The probability that a salesperson makes a successful sale is 0.6. If 8 independent customers are contacted, find the probability of:**

(i) exactly 5 successful sales
(ii) at least one successful sale
(iii) at most 2 successful sales. (8 Marks)

1. Given Data and Distribution Parameters

- Probability of a successful sale (p): $p = 0.60$
- Probability of failure ($q = 1 - p$): $q = 1 - 0.60 = 0.40$
- Number of independent customers contacted (n): $n = 8$
- Random variable X: Number of successful sales, where $X \sim B(n = 8, p = 0.60)$.

2. Binomial Probability Mass Function (PMF)

$$P(X = k) = \binom{n}{k} p^k q^{n-k} = \binom{8}{k} (0.6)^k (0.4)^{8-k}, \quad k = 0, 1, 2, \ldots, 8$$

3. Calculation Steps

(i) Probability of Exactly 5 Successful Sales ($P(X = 5)$)

$$P(X = 5) = \binom{8}{5} (0.6)^5 (0.4)^{8-5} = \binom{8}{3} (0.6)^5 (0.4)^3$$

$$\binom{8}{5} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$$

$$(0.6)^5 = 0.07776$$

$$(0.4)^3 = 0.064$$

$$P(X = 5) = 56 \times 0.07776 \times 0.064 = 56 \times 0.00497664 = 0.27869 \approx 0.2787$$

$$\boxed{P(X = 5) = 0.2787 \quad (27.87\%)}$$

(ii) Probability of At Least One Successful Sale ($P(X \geq 1)$)

Using the complement rule:

$$P(X \geq 1) = 1 - P(X = 0)$$

$$P(X = 0) = \binom{8}{0} (0.6)^0 (0.4)^8 = 1 \times 1 \times (0.4)^8 = 0.00065536$$

$$P(X \geq 1) = 1 - 0.00065536 = 0.99934464 \approx 0.9993$$

$$\boxed{P(X \geq 1) = 0.9993 \quad (99.93\%)}$$

(iii) Probability of At Most 2 Successful Sales ($P(X \leq 2)$)

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

- $P(X = 0) = 0.00065536$
- $P(X = 1) = \binom{8}{1}(0.6)^1(0.4)^7 = 8 \times 0.6 \times 0.0016384 = 0.00786432$
- $P(X = 2) = \binom{8}{2}(0.6)^2(0.4)^6 = 28 \times 0.36 \times 0.004096 = 0.04128768$

Summing the terms:

$$P(X \leq 2) = 0.00065536 + 0.00786432 + 0.04128768 = 0.04980736 \approx 0.0498$$

$$\boxed{P(X \leq 2) = 0.0498 \quad (4.98\%)}$$

> **Q21. Three different teaching methods are used for three groups of students. Their test scores are given below:**

Method A	Method B	Method C
8	12	10
10	11	9
9	13	11

Using One-Way ANOVA, test whether there is a significant difference among the mean scores of the three teaching methods at the 5% level of significance.
Given: Critical F-value for df = (2, 6) at the 5% level of significance = 5.14. (8 Marks)

1. Hypothesis Formulation

- **Null Hypothesis (H_0):** $\mu_A = \mu_B = \mu_C$ (There is no significant difference among the mean test scores of the three teaching methods).
- **Alternative Hypothesis (H_1):** At least one teaching method mean is significantly different.
- **Level of Significance (α):** 0.05 (5%).

2. Group Totals, Sample Sizes, and Grand Total

- Method A: $T_1 = 8 + 10 + 9 = 27$, sample size $n_1 = 3$, mean $\bar{X}_1 = 9.0$
- Method B: $T_2 = 12 + 11 + 13 = 36$, sample size $n_2 = 3$, mean $\bar{X}_2 = 12.0$
- Method C: $T_3 = 10 + 9 + 11 = 30$, sample size $n_3 = 3$, mean $\bar{X}_3 = 10.0$
- Number of groups (k) = 3
- Total observations (N) = $n_1 + n_2 + n_3 = 3 + 3 + 3 = 9$
- Grand Total (G):

$$G = T_1 + T_2 + T_3 = 27 + 36 + 30 = 93$$

3. Step-by-Step Sum of Squares Calculations

- **Step 1: Correction Factor (CF):**

$$CF = \frac{G^2}{N} = \frac{(93)^2}{9} = \frac{8,649}{9} = 961$$

- **Step 2: Total Sum of Squares (TSS):**

$$\sum X^2 = (8^2 + 10^2 + 9^2) + (12^2 + 11^2 + 13^2) + (10^2 + 9^2 + 11^2)$$

$$\sum X^2 = (64 + 100 + 81) + (144 + 121 + 169) + (100 + 81 + 121) = 245 + 434 + 302 = 981$$

$$TSS = \sum X^2 - CF = 981 - 961 = 20$$

- **Step 3: Between-Methods Sum of Squares (SSB):**

$$SSB = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} \right) - CF = \left(\frac{27^2}{3} + \frac{36^2}{3} + \frac{30^2}{3} \right) - 961$$

$$SSB = \left(\frac{729}{3} + \frac{1296}{3} + \frac{900}{3} \right) - 961 = (243 + 432 + 300) - 961 = 975 - 961 = 14$$

- **Step 4: Within-Methods Sum of Squares (SSW / Error):**

$$SSW = TSS - SSB = 20 - 14 = 6$$

4. One-Way ANOVA Summary Table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Sum of Squares (MS)	Variance Ratio (F_{calc})	Critical Value (F_{crit} at 5%)
Between Methods	$SSB = 14$	$k - 1 = 3 - 1 = 2$	$MSB = \frac{14}{2} = 7.0$	$F = \frac{MSB}{MSW} = \frac{7.0}{1.0} = 7.00$	$F_{(2,6)} = 5.14$
Within Methods (Error)	$SSW = 6$	$N - k = 9 - 3 = 6$	$MSW = \frac{6}{6} = 1.0$	—	—
Total	$TSS = 20$	$N - 1 = 9 - 1 = 8$	—	—	—

5. Statistical Decision and Inference

- **Decision Rule:** If $F_{\text{calc}} > F_{\text{crit}}$, reject H_0. Otherwise, accept H_0.
- **Comparison:**

$$F_{\text{calculated}} = 7.00 > F_{\text{critical}}(2, 6) = 5.14$$

- **Conclusion:** Since the calculated F-value (7.00) exceeds the critical table value (5.14) at the 5% level of significance, the Null Hypothesis (H_0) is REJECTED.
- **Interpretation:** There is a statistically significant difference among the mean test scores of the three teaching methods (Method B with $\bar{X} = 12$ is significantly superior).

$$\boxed{F_{\text{calculated}} = 7.00 > 5.14 \implies \text{Reject } H_0 \text{ at } 5\% \text{ significance level}}$$

Q22. Explain Map, Matrix, Card, and Gauge visuals and state their uses. (8 Marks)

1. Core Concept and Definition

Power BI Visualizations are graphical representations of aggregated model data designed to communicate operational and strategic metrics. **Map, Matrix, Card, and Gauge** visuals represent four distinct functional categories: spatial, multi-dimensional tabular, headline KPI, and radial target-tracking visualizations.

2. Comparative Visual Matrix

Visual Type	Primary Purpose	Key Configuration Fields	Primary Real-World Use Case
Map	Spatial distribution & geographic clustering	Location (City/State), Latitude/Longitude, Bubble Size	Regional sales territory penetration
Matrix	Multi-dimensional tabular cross-tabulation	Rows, Columns, Values, Drill-down Hierarchies	Financial Profit & Loss (P&L) statements
Card	Highlighting a solitary critical metric	Single Aggregated Field, Callout Value, Label	Total Gross Revenue, Active User count
Gauge	Progress tracking against a target threshold	Current Value, Min, Max, Target Goal Value	Quarterly sales quota achievement

3. Detailed Visual Features and Uses

- **A. Map Visuals (Bubble & Filled Maps):**
- *Features:* Utilizes built-in Bing and Azure Maps integration to plot geographical coordinates or administrative boundaries (Country, State, Postal Code). In Bubble Maps, bubble diameters scale proportional to metric magnitudes, while Filled (Choropleth) Maps colorize geographic territories based on performance intensity.
- *Uses:* Logistics route monitoring, regional revenue mapping, and pinpointing untapped geographical sales territories.
- **B. Matrix Visuals:**
- *Features:* An advanced, multi-dimensional grid resembling an Excel PivotTable. Supports simultaneous row and column groupings, stepped indentation layouts, expandable/collapsible hierarchical levels (e.g., Year → Quarter → Month), and multi-metric conditional formatting (heatmaps, data bars).
- *Uses:* Financial balance sheets, cross-category performance tables, and detailed multi-variable drill-down reporting.
- **C. Card Visuals (Single-Value & Multi-Row):**
- *Features:* Renders a single aggregated numerical figure using prominent display typography. Designed for immediate executive scanning without tabular clutter.
- *Uses:* Top-level dashboard KPI banners displaying vital headline metrics such as *Total Revenue ($5.4M)*, *Year-over-Year Growth (+14%)*, or *Net Promoter Score (78)*.
- **D. Gauge Visuals (Radial Speedometer Display):**
- *Features:* A circular arc showing progress from a minimum value to a maximum value, featuring a dynamic callout needle and a designated target milestone marker.
- *Uses:* Tracking operational KPIs against quotas, project completion percentages, and monitoring actual expenditure against annual budgetary limits.

4. Real-World Dashboard Example

- **Example:** An executive retail sales dashboard featuring a Card for *Total Sales (*
  $10M)*, a * *Gauge * *tracking * Progresstoward$*12M Target*, a Map visualizing *State-wise Store Sales*, and a Matrix cross-tabulating *Product Category across Fiscal Quarters*.

Q23. (a) Discuss the important features of data analytics. (4 Marks)

1. Concept and Meaning

Data Analytics is the computational and statistical science of examining raw datasets to extract actionable insights, detect non-obvious patterns, and support strategic decision-making.

2. Important Features (Key Characteristics)

- **Scalability and Big Data Capability:** Efficiently ingests, processes, and queries massive volumes of structured (SQL), semi-structured (JSON), and unstructured (text, media) data streams.
- **Predictive & Proactive Intelligence:** Transcends historical retrospective reporting by using statistical forecasting and machine learning algorithms to model future outcomes.
- **Automated Data Integration (ETL):** Connects to diverse, heterogeneous source systems, executing programmatic cleaning, deduplication, and schema modeling.
- **Interactive Visual Exploration:** Translates complex mathematical models into dynamic visual dashboards featuring drill-through, slicers, and cross-filtering.
- **Algorithmic Decision Automation:** Empowers real-time automated workflows, including algorithmic credit scoring, high-frequency trading, and fraud prevention alerts.

3. Real-World Example

- **Example:** Financial institutions deploying real-time analytics to detect fraudulent transactions within milliseconds using predictive anomaly scoring.

Q23. (b) Explain the capabilities and scope of data analytics. (4 Marks)

1. Concept and Meaning

The **Scope of Data Analytics** encompasses the entire spectrum of analytical capabilities across organizational functions, ranging from retrospective diagnostic analysis to autonomous prescriptive optimization.

2. Core Capabilities and Industry Scope

- **Four-Tier Analytical Capabilities:**
- *Descriptive Analytics:* Answers "*What happened?*" through historical summaries and KPI reporting.
- *Diagnostic Analytics:* Answers "*Why did it happen?*" using root-cause drill-downs and correlation analysis.
- *Predictive Analytics:* Answers "*What is likely to happen?*" using machine learning regression and classification.
- *Prescriptive Analytics:* Answers "*What action should be taken?*" using optimization algorithms and simulation models.
- **Commercial & Operational Scope:** Optimizes inventory levels, predicts customer churn, and personalizes dynamic marketing funnels in retail and e-commerce.
- **Healthcare Scope:** Powers clinical patient risk stratification, epidemiological outbreak tracking, and genomic sequencing analytics.
- **Banking and Risk Governance Scope:** Manages credit underwriting, loan default prediction, and regulatory compliance reporting.
- **Smart Cities & Public Infrastructure:** Optimizes municipal traffic routing, energy grid load distribution, and public resource management.

3. Real-World Example

- **Example:** An e-commerce platform using predictive analytics to forecast festival demand and prescriptive analytics to dynamically adjust inventory replenishment and pricing rules.

Model Question Paper – 4 | Section C Solutions
Total Marks: 32 (Answer any 4 out of 7 Questions, 8 Marks each) | Comprehensive Solutions

Total Marks: 40 (Answer any 5 out of 7 Questions; All 7 Solved Here × 8 Marks each)
Standard: Rigorous, point-wise academic standard for BCU examinations (strictly 15-minute writing limit, 250–350 words for theory, ~120–150 words for sub-questions, and full step-by-step mathematical calculations for numericals). Exclusively formal academic solutions.

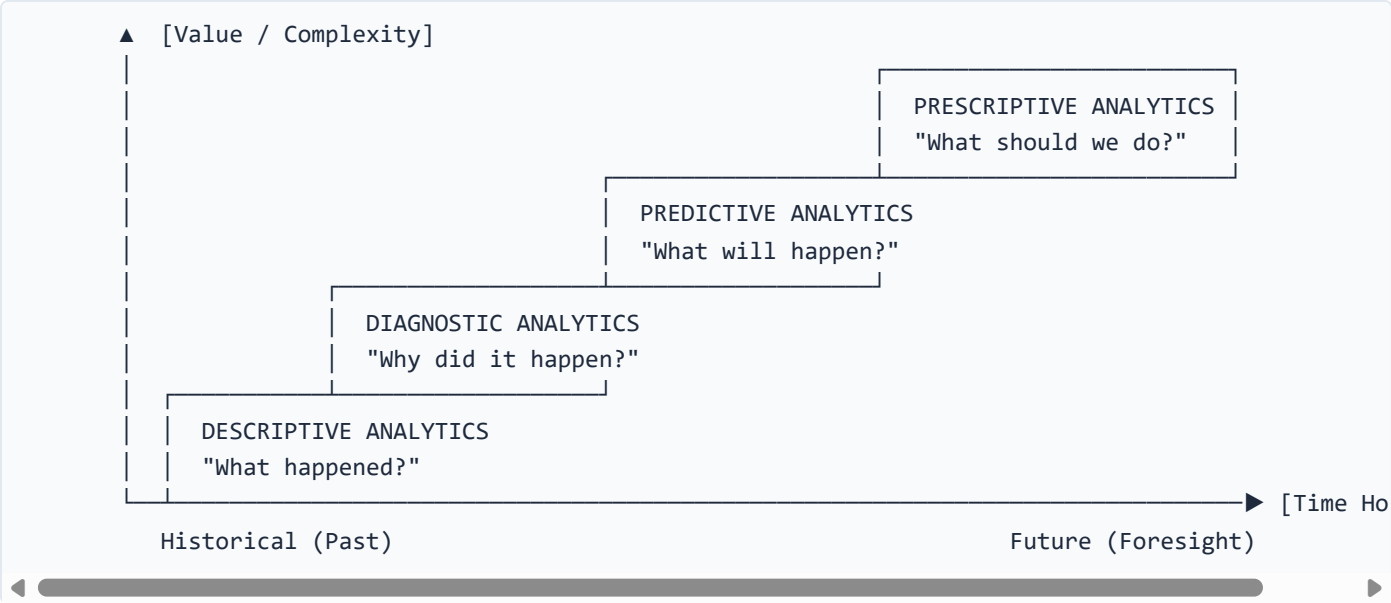
MODEL QUESTION PAPER – 4

Q17. Compare and contrast different types of data analytics with suitable examples.

Marks: 8 Marks
Subject Focus: Data Analytics Taxonomy, Analytical Maturity Spectrum, and Business Value Hierarchy

1. Core Concept & Definition

Data analytics is classified into four progressive evolutionary tiers—**Descriptive, Diagnostic, Predictive, and Prescriptive**—representing increasing levels of analytical maturity, business value, and algorithmic complexity. Organizations leverage this analytical continuum to transition from reactive historical reporting to autonomous forward-looking decision systems.



2. Comparative Matrix of Analytics Types

Feature / Dimension	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Primary Question	What has happened?	Why did it happen?	What will happen?	What should we do?
Time Orientation	Past (Historical)	Past to Present	Future (Foresight)	Future (Prescriptive Action)
Analytical Rigor	Aggregations & Summaries	Root-cause & Drilling	Machine Learning & Stats	Optimization & Simulations
Business Value	Basic situational awareness	Explanatory understanding	Forward-looking estimates	Automated strategic choices
Core Techniques	Mean, charts, KPI scorecards	Slicing, dicing, correlation	Regression, ARIMA, trees	Linear programming, heuristics

3. In-Depth Breakdown of Analytics Types

- **Descriptive Analytics (Historical Baseline):**
- Focuses on summarizing and interpreting historical transaction data to understand past operational performance.
- Employs measures of central tendency, data aggregation pipelines, and interactive executive dashboards (e.g., total quarterly sales, web pageviews, customer churn rate).
- **Diagnostic Analytics (Root-Cause Discovery):**
- Investigates underlying drivers, correlations, and anomalies observed in descriptive summaries.
- Uses multidimensional OLAP cubes (drill-down, drill-through, slice and dice), correlation analysis, hypothesis testing, and data discovery techniques to isolate root causes.
- **Predictive Analytics (Future Trajectory Estimation):**
- Utilizes statistical modeling, historical patterns, and supervised machine learning algorithms to project future event probabilities and trends.
- Employs techniques including linear/logistic regression, time-series forecasting (ARIMA), decision trees, random forests, and gradient boosting.
- **Prescriptive Analytics (Strategic Optimization):**
- Synthesizes predictive forecasts with operational constraints, business rules, and objective functions to recommend optimal business actions.
- Relies on linear programming, genetic algorithms, Monte Carlo simulation, and automated decision-tree engines to maximize profits or minimize operational risks.
- **Maturity & Value Dependency:**
- The four types form an interdependent value chain; diagnostic analysis requires solid descriptive data, while prescriptive optimization depends entirely on reliable predictive accuracy.

4. Real-World Industry Application (E-Commerce Use Case)

- **Descriptive:** The analytics dashboard highlights a 15% month-over-month decrease in shopping cart checkouts.
- **Diagnostic:** Drill-down analysis discovers that 80% of cart drop-offs occurred on mobile devices following a payment gateway update that caused checkout errors.
- **Predictive:** Machine learning models predict that cart abandonment will increase by an additional 25% next quarter if mobile payment latency is not resolved.
- **Prescriptive:** The recommendation engine automatically switches mobile checkouts to a secondary payment gateway and delivers a personalized 10% discount notification to mobile users to restore conversion rates.

Q18. Find the coefficient of correlation between the following two variables by using an assumed mean for deviations. Comment on the result through the probable error.

X	6	8	12	15	18	20	24	28	31
Y	10	12	15	15	18	25	22	26	28

Marks: 8 Marks
Subject Focus: Karl Pearson's Correlation Coefficient, Short-Cut (Assumed Mean) Method, Probable Error

1. Given Data & Assumed Means

- Total number of paired observations: $N = 9$
- Let Assumed Mean for variable X be $A = 18$ (central value of series X)
- Let Assumed Mean for variable Y be $B = 18$ (central value of series Y)
- Deviations: $dx = X - A = X - 18$ and $dy = Y - B = Y - 18$

2. Formula for Karl Pearson's Correlation Coefficient (Assumed Mean Method)

$$r = \frac{N\sum dxdy - (\sum dx)(\sum dy)}{\sqrt{[N\sum dx^2 - (\sum dx)^2]\,[N\sum dy^2 - (\sum dy)^2]}}$$

Formula for **Probable Error (P.E.)**:

$$P.E. = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$$

3. Complete Calculation Table

X	$dx = (X - 18)$	dx^2	Y	$dy = (Y - 18)$	dy^2	$dx \cdot dy$
6	-12	144	10	-8	64	$(-12)(-8) = 96$
8	-10	100	12	-6	36	$(-10)(-6) = 60$
12	-6	36	15	-3	9	$(-6)(-3) = 18$
15	-3	9	15	-3	9	$(-3)(-3) = 9$
18	0	0	18	0	0	$(0)(0) = 0$
20	+2	4	25	+7	49	$(2)(7) = 14$
24	+6	36	22	+4	16	$(6)(4) = 24$
28	+10	100	26	+8	64	$(10)(8) = 80$
31	+13	169	28	+10	100	$(13)(10) = 130$
$N = 9$	$\sum dx = 0$	$\sum dx^2 = 598$	-	$\sum dy = 9$	$\sum dy^2 = 347$	$\sum dx\,dy = 431$

4. Step-by-Step Mathematical Calculation

Step 1: Compute Numerator

$$\text{Numerator} = N\sum dxdy - (\sum dx)(\sum dy)$$

$$\text{Numerator} = 9(431) - (0)(9) = 3879 - 0 = 3879$$

Step 2: Compute Denominator Factors

$$\text{Factor}_1 = N\sum dx^2 - (\sum dx)^2 = 9(598) - (0)^2 = 5382 - 0 = 5382$$

$$\text{Factor}_2 = N\sum dy^2 - (\sum dy)^2 = 9(347) - (9)^2 = 3123 - 81 = 3042$$

Step 3: Compute Denominator

$$\text{Denominator} = \sqrt{\text{Factor}_1 \times \text{Factor}_2} = \sqrt{5382 \times 3042} = \sqrt{16,372,044} \approx 4046.238$$

Step 4: Calculate Correlation Coefficient (r)

$$r = \frac{3879}{4046.238} \approx +0.9587$$

5. Calculation of Probable Error (P.E.) & Significance Test

- **Coefficient of Determination (r^2):**

$$r^2 = (0.9587)^2 \approx 0.9191$$

- **Evaluating $1 - r^2$:**

$$1 - r^2 = 1 - 0.9191 = 0.0809$$

- **Square Root of Sample Size ($\sqrt{N}$):**

$$\sqrt{N} = \sqrt{9} = 3$$

- **Probable Error (P.E.):**

$$P.E. = 0.6745 \times \frac{0.0809}{3} = 0.6745 \times 0.02697 \approx 0.0182$$

- **Evaluation of $6 \times P.E.$:**

$$6 \times P.E. = 6 \times 0.0182 = 0.1092$$

6. Statistical Comment on the Result

1. **Degree of Correlation:** The value of $r = +0.9587$ indicates an exceptionally strong positive linear relationship between variables X and Y.
2. **Significance Decision Rule:**
3. If $r < P.E.$, correlation is non-significant.
4. If $r > 6 \times P.E.$, correlation is considered **definitely significant and reliable**.
5. **Conclusion:** Since $r = 0.9587$ is substantially greater than $6 \times P.E. = 0.1092$ ($0.9587 > 0.1092$), the correlation between X and Y is statistically highly significant and dependable.

7. Final Boxed Answer

Coefficient of Correlation (r):	$r = +0.9587$
Probable Error ($P.E.$):	$P.E. = 0.0182$
$6 \times P.E.$**:**	$6 \times P.E. = 0.1092$
Statistical Significance:	$r > 6 \times P.E. \implies$ **Highly Significant & Reliable**

Q19. Explain the different types of Correlation.

Marks: 8 Marks
Subject Focus: Bivariate & Multivariate Statistics, Degree and Direction of Correlation, Taxonomical Classification

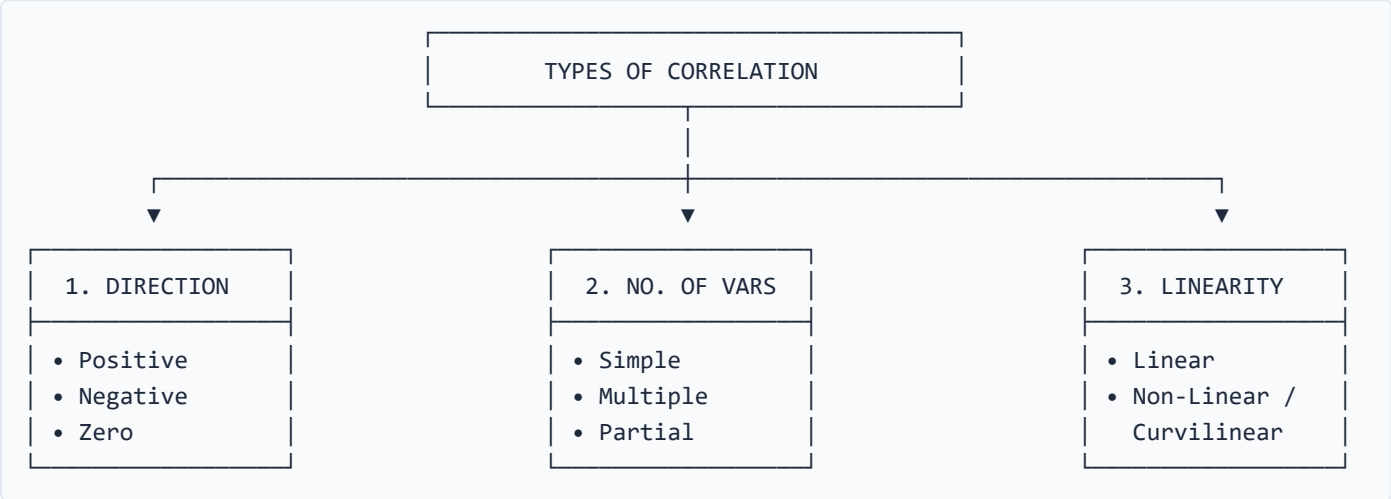
1. Core Concept & Definition

Correlation is a fundamental statistical measure that describes and quantifies the strength, degree, and direction of association between two or more random variables. Formulated mathematically by Karl Pearson, the correlation coefficient (r) ranges strictly between -1 and $+1$:

$$-1 \leq r \leq +1$$

Correlation indicates mutual co-variation but does not imply causal dependency.

2. Comprehensive Taxonomy of Correlation



3. Detailed Explanation of Correlation Types

A. Classification Based on Direction of Co-Movement

- **Positive Correlation ($0 < r \leq +1$):**
- Occurs when two variables change simultaneously in the same direction; an increase in variable X corresponds to an increase in variable Y, and a decrease in X corresponds to a decrease in Y.
- On a scatter plot, observations form a cluster that slopes upward from bottom-left to top-right.
- *Example:* Advertising expenditure and sales revenue; height and weight of individuals.
- **Negative Correlation (Inverse Correlation, $-1 \leq r < 0$):**
- Occurs when two variables move in opposite directions; an increase in variable X corresponds to a decrease in variable Y.
- Scatter points slope downward from top-left to bottom-right.
- *Example:* Unit price of a commodity and market demand; vehicle speed and travel time.
- **Zero Correlation ($r = 0$):**
- Indicates complete absence of any linear relationship between variables; variations in X provide zero information regarding variations in Y.
- Scatter points appear completely random and circular across the graph.
- *Example:* Shoe size and individual IQ; student roll number and exam marks.

B. Classification Based on Number of Variables Under Study

- **Simple Correlation:**
- Examines the linear association between strictly **two variables** (e.g., studying the relationship between rainfall and agricultural yield in isolation).
- **Multiple Correlation:**

- Evaluates the combined, joint relationship between **one dependent variable and two or more independent variables** simultaneously (denoted by $R_{1.23}$).
- *Example:* Examining how agricultural yield (Y) is jointly associated with rainfall (X_1) and fertilizer amount (X_2).
- **Partial Correlation:**
- Measures the net linear relationship between two variables while mathematically **controlling and holding constant** the confounding influence of one or more additional variables (denoted by $r_{12.3}$).
- *Example:* Measuring the correlation between crop yield and fertilizer while keeping rainfall and temperature constant.

C. Classification Based on Linearity / Ratio of Variation

- **Linear Correlation:**
- Occurs when the amount of change in one variable bears a **constant ratio** to the amount of change in the other variable throughout the entire range of values.
- When plotted on a Cartesian plane, points form an exact or near straight line ($Y = a + bX$).
- **Non-Linear (Curvilinear) Correlation:**
- Occurs when the ratio of change between variables is **not constant** but fluctuates at different intervals.
- Scatter plots display curved trajectories, parabolic arcs, or U-shaped curves.
- *Example:* Crop yield versus chemical fertilizer application; yield increases up to an optimal dosage, after which excess fertilizer burns crops and reduces yield.

4. Comparative Summary Table

Category	Type of Correlation	Value of r / Notation	Graphical Scatter Pattern	Real-World Benchmark Example
Direction	Positive	$0 < r \leq +1$	Upward sloping diagonal line	Household income vs. monthly savings
Direction	Negative	$-1 \leq r < 0$	Downward sloping diagonal line	Temperature vs. winter apparel sales
Direction	Zero	$r = 0$	Dispersed circular cloud	Atmospheric pressure vs. stock returns
Variables	Simple	r_{xy}	Single 2D bivariate scatter plot	Study hours vs. examination score
Variables	Multiple	$R_{y.x_1x_2}$	Multidimensional hyperplane	House price vs. (area + rooms + age)
Variables	Partial	$r_{12.3}$	Controlled partial residual plot	Blood pressure vs. age (holding salt intake constant)
Linearity	Linear	Constant slope	Exact straight line	Distance travelled vs. fuel consumed at constant speed
Linearity	Non-Linear	Variable slope	Parabolic or U-shaped curve	Machine output vs. continuous operating hours (fatigue)

Q20. A discrete random variable X has the following probability distribution:

X	0	1	2	3
$P(X)$	k	$2k$	$3k$	$4k$

Find k, the mean, variance and standard deviation.

Marks: 8 Marks
Subject Focus: Discrete Probability Distributions, Expectation, Variance, Standard Deviation

1. Mathematical Foundations

For any mathematically valid discrete probability distribution of a random variable X: 1. Every individual probability must be non-negative:

$$P(X = x_i) \geq 0, \quad \forall i$$

2. The total sum of all probabilities over the complete sample space must equal unity:

$$\sum_{i} P(X = x_i) = 1$$

3. **Mathematical Expectation (Mean μ):**

$$\mu = E(X) = \sum X \cdot P(X)$$

4. **Second Raw Moment ($E(X^2)$):**

$$E(X^2) = \sum X^2 \cdot P(X)$$

5. **Variance (σ^2 or $\text{Var}(X)$):**

$$\sigma^2 = \text{Var}(X) = E(X^2) - [E(X)]^2$$

6. **Standard Deviation (σ):**

$$\sigma = \sqrt{\text{Var}(X)}$$

2. Step-by-Step Derivation

Step 1: Find the Value of k

Using the axiom of total probability:

$$\sum P(X) = 1$$

$$P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1$$

$$k + 2k + 3k + 4k = 1$$

$$10k = 1$$

$$k = \frac{1}{10} = 0.1$$

- **Formulated Probability Distribution:**
- $P(X = 0) = k = 0.1$
- $P(X = 1) = 2k = 0.2$

- $P(X = 2) = 3k = 0.3$
- $P(X = 3) = 4k = 0.4$
- *Check:* $0.1 + 0.2 + 0.3 + 0.4 = 1.0$ ✓

3. Calculation Table for Moments

X	$P(X)$	$X \cdot P(X)$	X^2	$X^2 \cdot P(X)$
0	0.1	$0 \times 0.1 = 0.0$	0	$0 \times 0.1 = 0.0$
1	0.2	$1 \times 0.2 = 0.2$	1	$1 \times 0.2 = 0.2$
2	0.3	$2 \times 0.3 = 0.6$	4	$4 \times 0.3 = 1.2$
3	0.4	$3 \times 0.4 = 1.2$	9	$9 \times 0.4 = 3.6$
Total	$\sum P(X) = 1.0$	$\sum XP(X) = 2.0$	-	$\sum X^2P(X) = 5.0$

Step 2: Compute Mean ($E(X)$)

$$\mu = E(X) = \sum X \cdot P(X)$$

$$E(X) = 0(0.1) + 1(0.2) + 2(0.3) + 3(0.4)$$

$$E(X) = 0 + 0.2 + 0.6 + 1.2 = 2.0$$

Step 3: Compute Second Raw Moment ($E(X^2)$)

$$E(X^2) = \sum X^2 \cdot P(X)$$

$$E(X^2) = 0^2(0.1) + 1^2(0.2) + 2^2(0.3) + 3^2(0.4)$$

$$E(X^2) = 0 + 0.2 + 4(0.3) + 9(0.4)$$

$$E(X^2) = 0 + 0.2 + 1.2 + 3.6 = 5.0$$

Step 4: Compute Variance (σ^2)

$$\sigma^2 = \text{Var}(X) = E(X^2) - [E(X)]^2$$

$$\sigma^2 = 5.0 - (2.0)^2$$

$$\sigma^2 = 5.0 - 4.0 = 1.0$$

Step 5: Compute Standard Deviation (σ)

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{1.0} = 1.0$$

4. Final Boxed Answer

Constant (k):	$k = \frac{1}{10} = 0.1$
Mean [$\mu = E(X)$]:	$\mu = 2.0$
Variance [$\sigma^2 = \text{Var}(X)$]:	$\sigma^2 = 1.0$
Standard Deviation (σ):	$\sigma = 1.0$

Q21. A supermarket records the preferred payment methods of 200 customers as follows:

Payment Method	Cash	Card	UPI	Net Banking
Number of Customers	40	55	65	40

Test whether customers have an equal preference for the four payment methods at the 5% level of significance. Given: Critical χ^2-value for 3 degrees of freedom at the 5% level of significance = 7.815.

Marks: 8 Marks
Subject Focus: Chi-Square (χ^2) Goodness-of-Fit Test, Hypothesis Testing, Decision Rule

1. Hypothesis Formulation

- **Null Hypothesis (H_0):** Customers have an equal preference for all four payment methods in the supermarket:

$$p_1 = p_2 = p_3 = p_4 = \frac{1}{4} = 0.25$$

- **Alternative Hypothesis (H_1):** Customers do not have an equal preference for the four payment methods (at least one method differs significantly).

2. Test Statistic & Formula

The **Chi-Square (χ^2) Goodness-of-Fit** test statistic is defined as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where: * O_i = Observed frequency in category i * E_i = Expected frequency under H_0 in category i * k = Total number of categories ($k = 4$)

Expected Frequency Calculation:

Under the null hypothesis of equal customer preference across $k = 4$ categories:

$$E_i = \frac{\text{Total Sample Size } (N)}{k} = \frac{200}{4} = 50 \quad \text{for each payment method}$$

3. Complete Calculation Table

Payment Method	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Cash	40	50	$40 - 50 = -10$	$(-10)^2 = 100$	$\frac{100}{50} = 2.000$
Card	55	50	$55 - 50 = +5$	$(5)^2 = 25$	$\frac{25}{50} = 0.500$
UPI	65	50	$65 - 50 = +15$	$(15)^2 = 225$	$\frac{225}{50} = 4.500$
Net Banking	40	50	$40 - 50 = -10$	$(-10)^2 = 100$	$\frac{100}{50} = 2.000$
Total	$N = 200$	**200**	$\sum(O - E) = 0$	-	$\chi^2_{\text{calc}} = 9.000$

4. Degrees of Freedom & Critical Value

- **Degrees of Freedom (df):**

$$df = k - 1 = 4 - 1 = 3$$

- **Level of Significance (α):**

$$\alpha = 0.05 \ (5\%)$$

- **Critical Value:**

$$\chi^2_{\text{critical}} \ (0.05, 3) = 7.815 \quad \text{(Given)}$$

5. Statistical Decision Rule & Comparison

- **Decision Criterion:**
- If $\chi^2_{\text{calculated}} \leq \chi^2_{\text{critical}}$, fail to reject H_0.
- If $\chi^2_{\text{calculated}} > \chi^2_{\text{critical}}$, reject H_0.
- **Comparison:**

$$\chi^2_{\text{calculated}} = 9.000$$

$$\chi^2_{\text{critical}} = 7.815$$

$$9.000 > 7.815 \implies \chi^2_{\text{calculated}} > \chi^2_{\text{critical}}$$

6. Final Interpretation & Conclusion

Since the calculated test statistic ($\chi^2 = 9.000$) falls in the critical rejection region: 1. We **reject the Null Hypothesis (H_0)** at the 5% level of significance. 2. **Business Conclusion:** There is statistically significant evidence to conclude that customers **do not have an equal preference** for the four payment methods. UPI (65) and Card (55) are significantly more preferred by supermarket shoppers than Cash (40) and Net Banking (40).

7. Final Boxed Answer

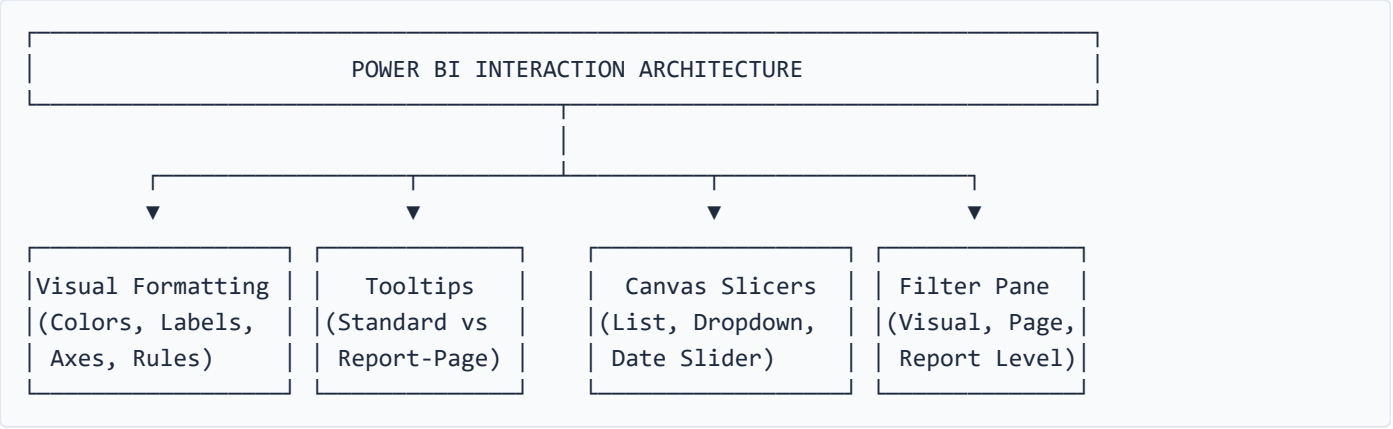
Expected Frequency for Each Method (E):	$E = 50$
Degrees of Freedom (df):	$df = 3$
Calculated Chi-Square Value (χ^2_{calc}):	$\chi^2_{calc} = 9.000$
Critical Chi-Square Value ($\chi^2_{0.05,3}$):	$\chi^2_{crit} = 7.815$
Statistical Decision:	Reject H_0 $(9.000 > 7.815)$
Final Inference:	Customer preferences are NOT equal

Q22. Explain formatting visuals, tooltips, slicers, and filters used for creating interactive Power BI reports.

Marks: 8 Marks
Subject Focus: Power BI Interactivity, Report UX/UI Design, Visual Formatting, Dynamic Exploration

1. Core Concept of Interactivity in Power BI

Interactivity in Microsoft Power BI transforms static graphical reports into dynamic, multidimensional analytical applications. By configuring visual formatting, informative tooltips, on-canvas slicers, and structured filter hierarchies, report designers empower business stakeholders to explore data self-service, drill into anomalies, and answer ad-hoc questions without modifying underlying queries.



2. Core Interactive Components in Power BI

A. Formatting Visuals

Visual formatting modifies the aesthetic and structural presentation of chart components via the **Format Pane**: * **Axes and Gridlines:** Customizing X and Y axis scale ranges, font size, title orientation, and reference lines (e.g., Target Line, Average Line). * **Data Colors & Conditional Formatting:** Dynamically coloring data points based on business rules (e.g., formatting profit bars as Green for > 0 and Red for < 0). * **Data Labels & Totals:** Enabling explicit numerical values on data points, bar segments, and stacked column totals to eliminate visual ambiguity. * **Containers & Cards:** Customizing backgrounds, borders, drop shadows, visual padding, and header icons for high visual clarity.

B. Tooltips (Standard vs. Report Page Tooltips)

Tooltips provide context-sensitive details on-demand when a user hovers their cursor over any visual element: * **Default Tooltips:** Automatically displays the underlying categorical dimension and exact numerical measure associated with the hovered data point. * **Custom Tooltip Fields:** Users can drag additional data metrics into the visual's "Tooltips" field well to show extra context without cluttering chart real estate. * **Report Page Tooltips (Mini-Pages):** Designers can design an entirely separate hidden report canvas, configure its page type as "Tooltip", and bind it to a visual. Hovering over a chart element renders rich miniature visuals, sparklines, and KPI breakdown cards.

C. On-Canvas Slicers

Slicers are specialized interactive controls placed directly onto the visual report canvas to enable intuitive, one-click dataset filtering: * **Slicer Formats:** * *List / Checkbox:* Displays all available categorical options for multi-selection. * *Dropdown:* Compact, space-efficient control ideal for dense dashboard canvases. * *Between / Slider:* Numerical and date range sliders allowing consumers to drag start and end bounds. * *Relative Date Slicer:* Filters data dynamically by rolling timeframes (e.g., "Last 30 Days", "This Quarter", "Next 2 Weeks"). * **Sync Slicers:** Allows a single on-canvas slicer on one page to synchronize its state seamlessly across multiple designated report pages.

D. Filters Pane Architecture

The **Filters Pane** provides a structured, hierarchical mechanism for scoping data across three distinct operational levels: 1. **Visual-Level Filters:** Applies filtering logic strictly to the single selected visual without influencing neighboring visuals. 2. **Page-Level Filters:** Restricts data globally across every visual present on the active report page. 3. **Report-Level Filters (All Pages):** Enforces universal data criteria across all pages throughout the entire Power BI report file (e.g., restricting the entire report to `Year = 2024`).

E. Cross-Filtering and Cross-Highlighting

- Clicking on any data category in a chart natively cross-filters or cross-highlights related data elements in all other visuals on the canvas simultaneously, enabling instantaneous multi-variable correlation analysis.

3. Summary of Interaction Components

Feature	Primary Function	User Interaction Method	Typical Enterprise Application
Formatting Visuals	Aesthetic styling & rule highlights	Visual presentation & conditional rules	Highlighting revenue deficit targets in red
Report-Page Tooltips	Deep contextual drill-down on demand	Hover cursor over data point	Hovering store pin shows top 5 items sold
On-Canvas Slicers	Direct, user-driven data slicing	Click checkboxes, dropdowns, date sliders	Slicing financial reports by Region or Fiscal Year
Filters Pane	Structured hierarchical scope control	Expand right-hand pane & select criteria	Enforcing report-wide filter to active product lines

---

Q23.

(a) Explain the Power Query Editor and its major parts. [4 Marks]

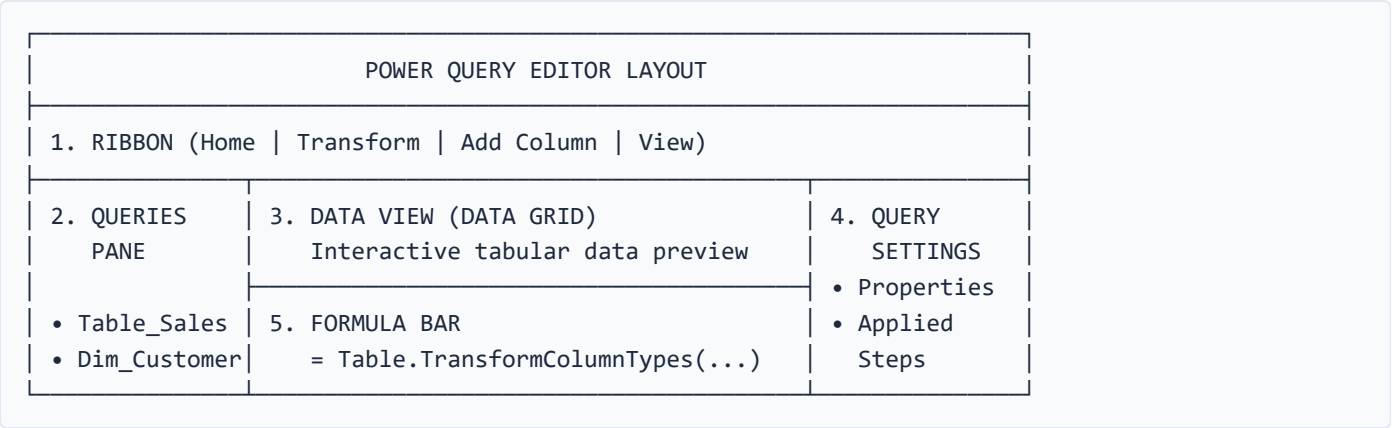
(b) Discuss the advantages and disadvantages of Power BI. [4 Marks]

Q23 (a). Explain the Power Query Editor and its major parts.

Marks: 4 Marks
Subject Focus: Power Query Interface Architecture, ETL Pipelines, M Formula Engine

1. Concept of Power Query Editor

The **Power Query Editor** is a dedicated graphical ETL (Extract, Transform, Load) environment embedded within Power BI Desktop. It enables data analysts to connect to heterogeneous external sources, ingest raw datasets, cleanse inconsistencies, reshape table layouts, and load structured tables into the analytical data model. Every user-driven transformation is automatically translated and recorded in the underlying functional **M Formula Language**.



2. Five Major Parts of the Interface

- **1. Query Ribbon:** The top navigation header organized into specialized functional tabs (Home, Transform, Add Column, and View). Houses one-click commands for data type casting, splitting columns, replacing nulls, grouping records, and merging queries.
- **2. Queries Pane:** Located on the far-left sidebar; displays an organized hierarchical list of all active tables, queries, database connections, and parameters within the project file.
- **3. Data View (Grid):** The central interactive canvas displaying a live tabular preview of the selected query. Enables users to inspect column values, column quality profiles, value distributions, and sort order.
- **4. Formula Bar:** Positioned directly above the data grid; displays and permits manual code editing of the precise M-code expression generated for the currently active transformation step.
- **5. Query Settings Pane:** Located on the right sidebar; divided into two essential sections:
- *Properties:* Renaming queries and setting description metadata.
- *Applied Steps List:* Displays a sequential, immutable audit trail of every applied transformation (e.g., `Source`, `Promoted Headers`, `Changed Type`). Users can click any step to travel back in time, reorder actions, or delete mistakes.

Q23 (b). Discuss the advantages and disadvantages of Power BI.

Marks: 4 Marks
Subject Focus: Business Intelligence Capabilities, Enterprise Architecture, Platform Evaluation

1. Overview of Microsoft Power BI

Microsoft Power BI is an industry-leading cloud-based business analytics platform used to model enterprise data and deliver interactive visual reporting. While offering robust enterprise capabilities, organizations must weigh its benefits against specific operational constraints.

2. Major Advantages of Power BI

- **1. Extensive Source Connectivity:** Natively integrates with hundreds of heterogeneous on-premises, cloud, and big data sources (Microsoft SQL Server, Oracle, Excel, Azure Synapse, Salesforce, REST APIs).
- **2. Robust Transformation & Modeling Engine:** Combines the visual ETL data-cleansing power of Power Query with high-performance dimensional data modeling and advanced calculations using DAX (Data Analysis Expressions).
- **3. Rich Interactive Visualizations:** Provides an expansive library of customizable charts, cards, matrix tables, geospatial maps, and access to the Microsoft AppSource custom visuals marketplace with native cross-filtering.
- **4. Seamless Microsoft Ecosystem Integration:** Native, frictionless integration with Microsoft 365, Teams, SharePoint, and Excel (Analyze in Excel), backed by cost-effective per-user pricing (Power BI Pro).

3. Major Disadvantages & Limitations of Power BI

- **1. In-Memory Dataset Size Limits:** Power BI Pro imposes a strict uncompressed 1 GB dataset memory limit per workspace model, necessitating expensive Premium capacity licenses for massive enterprise big data.
- **2. Steep DAX Learning Curve:** While basic drag-and-drop report creation is straightforward, mastering complex DAX formula logic (filter context transition, row context, `CALCULATE()`, time-intelligence) is intellectually demanding.
- **3. Local Hardware Resource Consumption:** Power BI Desktop is memory-intensive; importing and refreshing multi-million row datasets causes significant CPU and RAM throttling on standard client machines.
- **4. Cloud Collaboration Licensing Barrier:** Secure dashboard sharing, automated scheduled data refreshes, and workspace collaboration require paid licenses for both the content author and each viewing stakeholder.

End of Model Question Paper - 4 Section C Solutions.

Model Question Paper – 5 | Section C Solutions
Total Marks: 32 (Answer any 4 out of 7 Questions, 8 Marks each) | Comprehensive Solutions

Total Marks: 40 (5 Questions to be answered out of 7 × 8 Marks each)
Questions Covered: Q17 to Q23 (All 7 Questions Fully Solved)
Exam-Length Standard: Realistic 15-minute exam writing length (250–350 words for theory questions, ~120–150 words for 4M sub-questions).
Structure: Core Concepts, Process Workflows / Comparative Tables, Bold Bullet Points, Step-by-Step Numericals with Calculation Tables, Decision Rules, and Boxed Answers. Strictly NO Memory Tips.

MODEL QUESTION PAPER – 5

Q17. Describe the web analytics process and the major techniques used to study website data. [8 Marks]

1. Core Concept & Definition

Web Analytics is the systematic measurement, collection, analysis, and reporting of internet data to understand user behavior, evaluate website efficacy, and optimize digital channels for business performance. It converts raw browser interactions and web server logs into actionable operational and strategic insights.

2. Process Flowchart & Lifecycle

```
[1. Collection]   --> Page tags (JavaScript) and web server logs capture clicks, hits, and dwell tim
         |
[2. Processing]   --> Sessionization aggregates raw hits; purges internal IP traffic and web crawler
         |
[3. Analysis]     --> Evaluates conversion funnels, visitor drop-offs, user paths, and channel attri
         |
[4. Optimization] --> Deploys A/B split testing and UX refinements to maximize conversion rates.
```

3. Major Techniques Used to Study Website Data

- **Clickstream & Heatmap Analysis:** Tracks granular user navigational trails, cursor movements, and scroll depth across web pages to identify engaging visual elements and layout dead zones.
- **Conversion Funnel Analysis:** Maps sequential multi-stage user journeys (e.g., *Homepage* → *Product Catalog* → *Cart* → *Checkout*) to pinpoint exact operational stages where users abandon transactions.
- **Traffic Acquisition & Attribution Modeling:** Categorizes inbound traffic sources (Organic Search, Paid Search Ads, Social Media, Referrals, Direct) to evaluate marketing channel ROI and allocate marketing spend.
- **Cohort & Behavioral Segmentation:** Segments visitors by demographic, device type, geographic location, or temporal attributes (new vs. returning users) to uncover retention trends over time.
- **A/B & Multivariate Testing:** Compares two or more web page variations simultaneously against live user traffic to identify layout or call-to-action (CTA) button designs that optimize conversion rates empirically.

- **Bounce & Exit Rate Evaluation:** Measures single-page session abandonments (bounce rate) and common checkout abandonment exit pages to detect technical defects, broken links, or unengaging content.

4. Concrete Example

An e-commerce platform using funnel analytics detects a 68% drop-off at checkout due to mandatory account registration; replacing it with guest checkout decreases checkout abandonment by 24% and lifts revenue.

Q18. Calculate the coefficient of Rank correlation from the following: [8 Marks]

X	48	33	40	9	16	16	65	24	16	57
Y	13	13	24	6	15	4	20	9	6	19

1. Given Data & Identification of Repeated Ranks (Ties)

- Total number of paired observations: $N = 10$.
- **Ties in Series X:**
- The value 16 appears 3 **times** ($m_1 = 3$).
- The values of X sorted descending are: $65(1), 57(2), 48(3), 40(4), 33(5), 24(6), 16, 16, 16(7, 8, 9), 9(10)$.
- Shared rank for $16 = \frac{7+8+9}{3} = 8.0$.
- **Ties in Series Y:**
- The value 13 appears 2 **times** ($m_2 = 2$).
- The values of Y sorted descending are: $24(1), 20(2), 19(3), 15(4), 13, 13(5, 6), 9(7), 6, 6(8, 9), 4(10)$.
- Shared rank for $13 = \frac{5+6}{2} = 5.5$.
- The value 6 appears 2 **times** ($m_3 = 2$).
- Shared rank for $6 = \frac{8+9}{2} = 8.5$.

2. Formulas

- **Spearman's Rank Correlation Coefficient with Ties:**

$$\rho = 1 - \frac{6\left(\sum d_i^2 + \mathrm{CF}\right)}{N(N^2 - 1)}$$

- **Correction Factor (CF):**

$$\mathrm{CF} = \sum \frac{m(m^2 - 1)}{12} = \frac{m_1(m_1^2 - 1)}{12} + \frac{m_2(m_2^2 - 1)}{12} + \frac{m_3(m_3^2 - 1)}{12}$$

3. Complete Calculation Table

Sl. No.	X	Y	R_X	R_Y	$d = R_X - R_Y$	d^2
1	48	13	3	5.5	-2.5	6.25
2	33	13	5	5.5	-0.5	0.25
3	40	24	4	1.0	+3.0	9.00
4	9	6	10	8.5	+1.5	2.25
5	16	15	8.0	4.0	+4.0	16.00
6	16	4	8.0	10.0	-2.0	4.00
7	65	20	1	2.0	-1.0	1.00
8	24	9	6	7.0	-1.0	1.00
9	16	6	8.0	8.5	-0.5	0.25
10	57	19	2	3.0	-1.0	1.00
Total	-	-	**55**	**55**	$\sum d = 0$	$\sum d^2 = 41.00$

4. Step-by-Step Arithmetic

1. **Computation of Correction Factors (CF):**
2. For X ($m_1 = 3$): $\frac{3(3^2-1)}{12} = \frac{3(8)}{12} = \frac{24}{12} = 2.0$
3. For Y ($m_2 = 2$): $\frac{2(2^2-1)}{12} = \frac{2(3)}{12} = \frac{6}{12} = 0.5$
4. For Y ($m_3 = 2$): $\frac{2(2^2-1)}{12} = \frac{2(3)}{12} = \frac{6}{12} = 0.5$

$$\text{Total CF} = 2.0 + 0.5 + 0.5 = 3.0$$

5. **Corrected Sum of Squared Differences:**

$$\sum d^2 + \mathrm{CF} = 41.0 + 3.0 = 44.0$$

6. **Substitution into Rank Correlation Formula:**

$$\rho = 1 - \frac{6 \times 44.0}{10(10^2 - 1)} = 1 - \frac{264}{10(99)} = 1 - \frac{264}{990}$$

$$\rho = 1 - \frac{4}{15} = \frac{11}{15} \approx 0.7333$$

5. Boxed Final Answer

Sum of Squared Differences:	$\sum d^2 = 41.00$
Total Correction Factor (CF):	$\mathrm{CF} = 3.00$
Coefficient of Rank Correlation:	$\rho = \frac{11}{15} \approx +0.7333$
Interpretation:	Strong positive rank correlation between X and Y.

Q19.

(a) Calculate the two regression coefficients when $r = 0.8$, $\sigma_x = 5$ and $\sigma_y = 7$. (4 Marks)
(b) If covariance between X and Y variables is 12.5 and the variance of X and Y are respectively 16.4 and 13.8, find the coefficient of correlation between them. (4 Marks)

Part (a): Calculation of Regression Coefficients [4 Marks]

1. Given Data

- Correlation coefficient: $r = 0.8$
- Standard deviation of X: $\sigma_x = 5$
- Standard deviation of Y: $\sigma_y = 7$

2. Formulas

- Regression Coefficient of Y on X: $b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$
- Regression Coefficient of X on Y: $b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$

3. Calculation

- b_{yx}:

$$b_{yx} = 0.8 \times \frac{7}{5} = 0.8 \times 1.4 = 1.12$$

- b_{xy}:

$$b_{xy} = 0.8 \times \frac{5}{7} = \frac{4.0}{7} = \frac{4}{7} \approx 0.5714$$

4. Verification Check

$$b_{yx} \times b_{xy} = 1.12 \times \frac{4}{7} = \frac{1.12 \times 4}{7} = \frac{4.48}{7} = 0.64 = (0.8)^2 = r^2 \quad (\checkmark)$$

Regression Coefficient of Y on X:	$b_{yx} = 1.12$
Regression Coefficient of X on Y:	$b_{xy} = 0.5714 \quad \left(\text{or } \frac{4}{7}\right)$

Part (b): Calculation of Correlation Coefficient from Covariance [4 Marks]

1. Given Data

- Covariance between X and Y: $\text{Cov}(X, Y) = 12.5$
- Variance of X: $\text{Var}(X) = \sigma_x^2 = 16.4$
- Variance of Y: $\text{Var}(Y) = \sigma_y^2 = 13.8$

2. Formula

$$r = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y}$$

3. Calculation

1. **Product of Variances:**

$$\text{Var}(X) \times \text{Var}(Y) = 16.4 \times 13.8 = 226.32$$

2. **Square Root of Product:**

$$\sqrt{226.32} \approx 15.0439$$

3. **Correlation Coefficient (r):**

$$r = \frac{12.5}{15.0439} \approx 0.8309$$

Covariance:	$\text{Cov}(X, Y) = 12.5$
Denominator:	$\sqrt{16.4 \times 13.8} = 15.0439$
Coefficient of Correlation:	$r \approx +0.8309$ (High Positive Correlation)

> **Q20. A sample of 64 customers has an average monthly expenditure of ₹5,400. The population standard deviation is ₹800. Find: [8 Marks]**

(i) the standard error
(ii) the 95% confidence interval for the population mean.
Use $Z = 1.96$.

1. Given Data

- Sample size: $n = 64$
- Sample mean expenditure: $\bar{X} = Rs.\,5,400$
- Population standard deviation: $\sigma = Rs.\,800$
- Critical value for 95% confidence level: $Z = 1.96$

2. Formulas

- **(i) Standard Error of the Mean (SE):**

$$\text{SE} = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- **(ii) 95% Confidence Interval for Population Mean (μ):**

$$\text{Confidence Interval} = \bar{X} \pm Z \cdot \text{SE} = [\bar{X} - E,\ \bar{X} + E]$$

  where Margin of Error $E = Z \cdot \text{SE}$.

3. Step-by-Step Calculation

(i) Computation of Standard Error

$$\text{SE} = \frac{\sigma}{\sqrt{n}} = \frac{800}{\sqrt{64}} = \frac{800}{8} = Rs.\,100$$

(ii) Computation of the 95% Confidence Interval

1. **Margin of Error (E):**

$$E = Z \times \text{SE} = 1.96 \times 100 = Rs.196$$

2. **Lower Confidence Limit (μ_{lower}):**

$$\mu_{\text{lower}} = \bar{X} - E = 5400 - 196 = Rs.5,204$$

3. **Upper Confidence Limit (μ_{upper}):**

$$\mu_{\text{upper}} = \bar{X} + E = 5400 + 196 = Rs.5,596$$

4. Statistical Interpretation

We are **95% confident** that the true population mean monthly expenditure (μ) for all customers lies within the interval between **₹5,204 and ₹5,596**.

5. Boxed Final Answer

(i) Standard Error of the Mean:	$\text{SE} = \text{Rs. } 100$
(ii) Margin of Error (at 95%):	$E = \text{Rs. } 196$
(iii) 95% Confidence Interval for μ:	$[\text{Rs. } 5,204,\ \text{Rs. } 5,596]$ or $\text{Rs. } 5,204 \leq \mu \leq \text{Rs. } 5,596$

Q21. A college wants to determine whether year of study is related to preferred learning resource. A sample of 200 students gives the following information: [8 Marks]

Year of Study	Printed Books	E-Books	Video Tutorials	Total
First Year	45	30	25	100
Second Year	25	35	40	100
Total	**70**	**65**	**65**	**200**

Test whether year of study and preferred learning resource are associated at the 5% level of significance.
Given: Critical χ^2-value for 2 degrees of freedom at the 5% level of significance = 5.991.

1. Hypothesis Formulation

- **Null Hypothesis (H_0):** There is no association between the year of study and preferred learning resource (they are independent).
- **Alternative Hypothesis (H_1):** There is a statistically significant association between the year of study and preferred learning resource (they are dependent).
- **Level of Significance:** $\alpha = 0.05$ (5%).

2. Formulas & Degrees of Freedom

- **Expected Frequency Formula:**

$$E_{ij} = \frac{R_i \times C_j}{N}$$

where R_i = Row Total, C_j = Column Total, N = Grand Total (200).

- **Chi-Square Test Statistic:**

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- **Degrees of Freedom (df):**

$$df = (r-1)(c-1) = (2-1)(3-1) = 1 \times 2 = 2$$

3. Calculation of Expected Frequencies (E_{ij})

- $E_{11}(\text{First Year, Printed}) = \frac{100 \times 70}{200} = 35.0$
- $E_{12}(\text{First Year, E-Books}) = \frac{100 \times 65}{200} = 32.5$
- $E_{13}(\text{First Year, Video}) = \frac{100 \times 65}{200} = 32.5$
- $E_{21}(\text{Second Year, Printed}) = \frac{100 \times 70}{200} = 35.0$
- $E_{22}(\text{Second Year, E-Books}) = \frac{100 \times 65}{200} = 32.5$
- $E_{23}(\text{Second Year, Video}) = \frac{100 \times 65}{200} = 32.5$

4. Chi-Square Calculation Table

Cell (Row, Col)	O_{ij}	E_{ij}	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
(1, 1) 1st Year, Printed	45	35.0	+10.0	100.00	2.8571
(1, 2) 1st Year, E-Books	30	32.5	-2.5	6.25	0.1923
(1, 3) 1st Year, Video	25	32.5	-7.5	56.25	1.7308
(2, 1) 2nd Year, Printed	25	35.0	-10.0	100.00	2.8571
(2, 2) 2nd Year, E-Books	35	32.5	+2.5	6.25	0.1923
(2, 3) 2nd Year, Video	40	32.5	+7.5	56.25	1.7308
Total	**200**	**200.0**	**0.0**	-	$\chi^2_{\text{calc}} = 9.5604$

5. Statistical Decision Rule & Conclusion

- **Calculated Test Statistic:** $\chi^2_{\text{calc}} = 9.5604$
- **Critical Value at $\alpha = 0.05, df = 2$:** $\chi^2_{\text{crit}} = 5.991$
- **Decision Rule:** Since $\chi^2_{\text{calc}}(9.5604) > \chi^2_{\text{crit}}(5.991)$, we REJECT the Null Hypothesis (H_0).
- **Conclusion:** There is a statistically significant association between the year of study and the preferred learning resource at the 5% significance level. (First-year students exhibit a stronger preference for printed books, whereas second-year students lean distinctly toward video tutorials).

6. Boxed Final Answer

Degrees of Freedom:	$df = (2-1)(3-1) = 2$
Calculated Chi-Square Value:	$\chi^2_{\text{calc}} = 9.5604$
Critical Chi-Square Value (5%):	$\chi^2_{\text{crit}} = 5.991$
Statistical Decision:	Reject H_0 $\quad (\chi^2_{\text{calc}} > \chi^2_{\text{crit}})$
Conclusion:	Year of study and learning resource preference are significantly associated.

Q22. Explain drill-through and bookmarks and describe their role in creating interactive Power BI reports. [8 Marks]

1. Core Concept & Definition

In Microsoft Power BI, **Drill-Through** and **Bookmarks** are advanced interactivity features that transform static dashboard pages into dynamic, exploratory business intelligence applications, enabling self-service analysis and guided analytical storytelling.

2. Architectural Comparison Table

Feature	Primary Function	Trigger Mechanism	Scope of Action
Drill-Through	Contextual deep-dive to detail page	Right-click visual data point	Navigates across pages with filtered context
Bookmarks	Captures canvas display states	Button / Shape click	Toggles visual visibility, filters, and slicers

3. Key Interactivity Roles in Power BI Reports

- **Drill-Through Navigation:** Allows users to right-click a summarized metric (e.g., a specific sales region or product category) and jump to a dedicated destination report page filtered automatically for that specific entity.
- **Target Canvas Configuration:** Built by placing filtering attributes into the destination page's "Drill-through fields" well; Power BI automatically generates a "Back" navigation button to return to the parent summary view.
- **Cross-Report Drill-Through:** Enables enterprise scalability by connecting distinct Power BI reports that share common data models and dimensional keys.
- **Bookmark State Capture:** Records an exact snapshot of a report canvas, including visual visibility, filter pane selections, slicer values, and sort orders.
- **Collapsible Slicer Panels:** Pairing bookmarks with the Selection Pane enables toggling visual visibility on and off, allowing designers to build overlay menus and clean slicer pop-outs that conserve canvas real estate.
- **Guided Storytelling & Navigation:** Links bookmarks to shapes, icons, and buttons (`Action` $\rightarrow$ `Bookmark`) to guide non-technical executives through sequential presentations and reset default filter states.

4. Concrete Example

A financial report where an executive clicks a bookmark to toggle between "Revenue" and "Expense" views, then right-clicks the "North America" visual to drill through to individual ledger invoices.

Q23.

(a) What is Simple Linear Regression? Explain its meaning, features and applications with suitable examples. (4 Marks)
(b) Explain Business Intelligence and its importance in modern organizations. (4 Marks)

(a) Simple Linear Regression [4 Marks]

1. Meaning & Mathematical Model

Simple Linear Regression is a supervised parametric statistical modeling technique that models and quantifies the straight-line linear relationship between a single independent predictor variable ($\boldsymbol{X}$) and a continuous dependent response variable ($\boldsymbol{Y}$).

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \epsilon}$$

where $\boldsymbol{\beta_0}$ represents the population $\boldsymbol{Y}$-intercept, $\boldsymbol{\beta_1}$ denotes the regression slope (rate of change in $\boldsymbol{Y}$ per unit increase in $\boldsymbol{X}$), and $\boldsymbol{\epsilon}$ represents the random error term.

2. Core Features

- **Ordinary Least Squares (OLS):** Computes model coefficients by minimizing the sum of squared vertical prediction residuals ($\sum \boldsymbol{e_i^2}$).
- **Model Assumptions:** Assumes linearity, constant error variance (homoscedasticity), independence of observations, and normal distribution of residuals.
- **Goodness-of-Fit:** Evaluated using the Coefficient of Determination ($\boldsymbol{R^2}$), representing the proportion of total variance in $\boldsymbol{Y}$ explained by $\boldsymbol{X}$.

3. Applications & Practical Example

- **Applications:** Trend extrapolation, demand forecasting, and risk analysis.
- **Example:** Forecasting a retail store's annual sales revenue ($\boldsymbol{Y}$) based on its annual advertising marketing expenditure ($\boldsymbol{X}$).

(b) Business Intelligence and Its Importance in Modern Organizations [4 Marks]

1. Meaning

Business Intelligence (BI) is an integrated enterprise technology framework encompassing data extraction, ETL pipelines, data warehousing, analytical querying, and visual dashboards that convert raw transactional operational data into actionable strategic knowledge.

2. Strategic Importance in Modern Organizations

- **Informed Data-Driven Decision Making:** Replaces subjective intuition and guesswork with verified historical trends and live operational KPI metrics.
- **Single Source of Truth:** Consolidates isolated departmental silos (ERP, CRM, legacy SQL, flat files) into a unified, standardized enterprise data model.
- **Rapid Bottleneck Identification:** Visual dashboards alert managers instantly to supply chain delays, inventory shortages, or customer churn spikes.
- **Competitive Advantage & Revenue Growth:** Identifies emerging market opportunities, optimizes product pricing strategies, and enhances operational cost efficiencies.