CMR
INSTITUTE OF
TECHNOLOGY

USN | | | | | | | | | | |

Internal Assessment Test 1 – Nov 2025

Sub:	Business Data Analytics						Sub Code:		MMCA311C
Date:	06.11.25	Duration:	90 min's	Max Marks:	50	Sem:	III	Branch:	MCA

Note : Answer FIVE FULL Questions, choosing ONE full question from each Module

	PART I	MARKS	OBE	
			CO	RBT
1	Explain how data-driven decision making improves organizational performance. OR	[10]	CO1	L1
2	Compare structured and unstructured data with examples of business applications.	[10]	CO2	L1
	PART II			
3	What are the key skills and tools required by a business analyst? Discuss with examples. OR	[10]	CO2	L2
4	Describe the steps involved in data wrangling. Why is it considered the most time-consuming phase in data analytics?	[10]	CO2	L4
	PART III			
5	Explain the use of correlation and covariance in identifying relationships among business variables. OR	[10]	CO3	L3
6	Illustrate how histograms, boxplots, and heatmaps help in exploratory data analysis. Provide examples.	[10]	CO3	L3
	PART IV			
7	Explain the assumptions and applications of simple linear regression in business analytics. OR	[10]	CO3	L3
8	Compare linear regression and logistic regression in terms of objectives and outputs.	[10]	CO3	L3
	PART V			
9	Discuss how model performance metrics like accuracy and recall help assess a model's reliability in real-world cases. OR	[10]	CO3	L3
10	Write a detailed note on customer churn prediction models and how predictive analytics supports business retention strategies.	[10]	CO3	L3

Answers:

1. Explain how data-driven decision making improves organizational performance.

   Data-driven decision making (DDDM) improves organizational performance by enabling decisions that are based on facts, trends, and analytical insights rather than intuition or assumptions. When organizations collect and analyze data from multiple sources—such as customer behavior, operational processes, sales performance, and market trends—they gain a clearer understanding of what is happening and why. This allows leaders to identify patterns, diagnose problems accurately, and predict future outcomes.
   By using data analytics, organizations can evaluate the effectiveness of strategies, reduce risks, and allocate resources more efficiently. For example, analyzing customer data helps businesses identify target segments, personalize services, and improve customer satisfaction. Operational data analysis can reveal inefficiencies and guide process improvements, reducing costs and increasing productivity.
   Additionally, DDDM supports continuous improvement through performance monitoring and real-time feedback. It encourages an evidence-based culture where decisions are justified and measurable. As a result, organizations become more competitive, adaptive to change, and capable of identifying new opportunities in the market. Overall, data-driven decision making enhances strategic planning, operational efficiency, innovation, and customer value, leading to improved organizational performance.

2. Compare structured and unstructured data with examples of business applications

   Structured and unstructured data are two major categories of data used in business analytics, differing in format, storage, processing, and applications.

   Structured Data
   Structured data is highly organized and stored in a **tabular format** (rows and columns) such as databases or spreadsheets. It follows a fixed schema, making it easy to search, sort, analyze, and process using traditional data tools (SQL, Excel).
   Characteristics:
- Predefined format
- Easy to store and query
- Suitable for quantitative analysis

   Examples in Business Applications:

Business Area	Example of Structured Data	Application
Sales	Customer purchase records (date, product, price)	Sales forecasting and revenue analysis
Banking	Transaction history (account number, amount, timestamp)	Fraud detection and credit scoring
HR	Employee records (ID, salary, attendance)	Performance evaluation and payroll processing

   Unstructured Data
   Unstructured data has **no predefined format or schema**. It is mostly textual, audio, video, or image-based. Analyzing it requires advanced analytics techniques like **NLP, computer vision, or deep learning**.
   Characteristics:
- Free-form and non-tabular
- Difficult to store and search directly
- Suitable for qualitative insights

   Examples in Business Applications:

Business Area	Example of Unstructured Data	Application
Marketing	Customer product reviews, social media comments	Sentiment analysis for brand improvement
Healthcare	Medical images (X-rays, MRIs)	Diagnosis using image recognition models
Customer Service	Call center voice recordings	Chatbots and automated support systems

3. What are the key skills and tools required by a business analyst? Discuss with examples.
   A business analyst requires a combination of **technical skills**, **analytical abilities**, and **communication skills** to effectively interpret data and support decision making. These skills and tools help in understanding business problems, analyzing data, and recommending appropriate solutions that align with organizational goals.

   Key Skills Required

1. **Analytical Thinking and Problem-Solving**
   A business analyst should be able to break down complex business issues, identify the root cause, and propose feasible solutions.
   Example: An analyst identifies why sales dropped by examining customer purchase patterns.
2. **Communication and Documentation Skills**
   Business analysts interact with clients, management, and development teams. Clear documentation of requirements and reports is essential.
   Example: Preparing Business Requirement Documents (BRDs) and presenting dashboard insights.
3. **Domain Knowledge**
   Understanding the specific business environment (e.g., banking, healthcare, retail) helps in accurate interpretation of data.
   Example: Knowing financial indicators is essential for analyzing banking transactions.
4. **Critical Thinking and Decision Making**
   Ability to evaluate alternative solutions and choose the best course of action.
   Example: Choosing between cost reduction strategy vs. customer acquisition strategy.

   Key Tools Used

Tool / Category	Example Tools	Purpose	Example Use Case
Spreadsheets	Microsoft Excel, Google Sheets	Data cleaning, filtering, and basic analytics	Creating sales trend charts
Data Visualization Tools	Tableau, Power BI	Create interactive dashboards	Visualizing customer segmentation
Database Management	SQL, MySQL, Oracle	Query and retrieve structured data	Extracting monthly sales from a database
Statistical & Analytics Tools	Python, R, SAS	Advanced analysis, predictive modeling	Predicting customer churn
Documentation & Collaboration Tools	MS Word, Confluence, Jira	Requirement gathering and task management	Writing requirement specs and tracking project progress

4. Describe the steps involved in data wrangling. Why is it considered the most time-consuming phase in data analytics?

   Steps Involved in Data Wrangling
   Data wrangling (also called **data cleaning or data preprocessing**) is the process of converting raw, unstructured, or messy data into a clean and structured format suitable for analysis. The key steps involved are:

1. **Data Collection**
   Gathering data from different sources such as databases, spreadsheets, web APIs, sensors, or surveys.
   Example: Downloading sales data from an ERP system.
2. **Data Inspection and Profiling**
   Examining the dataset to understand its structure, missing values, inconsistencies, and overall quality.
   Example: Checking data types, ranges, and unique values.
3. **Data Cleaning**
   This involves handling missing values, removing duplicates, correcting inaccurate entries, and standardizing formats.
   Example: Replacing empty cells in a salary column with the average salary value.
4. **Data Transformation**
   Converting data into a suitable format for analysis. This may include normalization, encoding categorical data, or applying mathematical transformations.
   Example: Changing “High/Medium/Low” responses into numeric codes (2, 1, 0).

5. **Data Integration and Merging**
   Combining data from multiple sources into one unified dataset.
   Example: Merging customer demographics with purchase history data.
6. **Data Validation**
   Ensuring that the cleaned data is accurate and logically consistent with business expectations.
   Example: Verifying that dates are in chronological order and numeric values fall within expected ranges.

Why Data Wrangling Is Time-Consuming

Data wrangling is considered the **most time-consuming phase** in data analytics (often taking **60–80%** of the project time) because:

- **Real-world data is rarely clean.** It often contains missing values, errors, duplicates, and outliers that must be corrected before meaningful analysis can occur.
- **Multiple data sources may use different formats.** These need to be standardized to create a consistent dataset.
- **Understanding data requires domain knowledge.** Analysts must interact with business teams to interpret ambiguous entries.
- **Quality and accuracy are critical.** Incorrect or unclean data can produce misleading analysis and wrong decisions, so thorough verification is necessary.
- **Iterative Process.** Data wrangling often requires going back and forth between cleaning and validation until the dataset is usable.

5. Explain the use of correlation and covariance in identifying relationships among business variables.

Correlation and covariance are statistical measures used to examine the relationship between two business variables. Both help businesses understand how one variable changes with respect to another, which is essential for forecasting, planning, and decision-making.

Covariance

Covariance measures the **direction of the relationship** between two variables. It indicates whether the variables move together (positive covariance) or move in opposite directions (negative covariance).

- **Positive Covariance:** When sales of laptops increase, the sales of laptop accessories (like bags or mouse) may also increase.
- **Negative Covariance:** When product prices are increased, customer demand may decrease.

However, covariance **does not show the strength** of the relationship; it only shows whether variables move together.

Correlation

Correlation measures **both the direction and the strength of the relationship** between variables. It is the standardized form of covariance and ranges from **-1 to +1**.

- **+1**: Perfect positive relationship (as one increases, the other increases proportionally)
- **-1**: Perfect negative relationship (as one increases, the other decreases proportionally)
- **0**: No relationship between variables

Example in Business Applications:

- A high positive correlation between **advertising expenditure and sales revenue** means increasing advertisement budgets can boost sales.
- A negative correlation between **delivery time and customer satisfaction** shows that longer delivery times reduce customer satisfaction.

Importance in Business Decision-Making

Measure	What It Shows	Business Usage
Covariance	Whether variables move together	To understand basic directional trends in data
Correlation	Direction **and strength** of the relationship	To make predictions, set strategies, and build models

6. Illustrate how histograms, boxplots, and heatmaps help in exploratory data analysis. Provide examples.

Exploratory Data Analysis (EDA) helps in understanding patterns, trends, and relationships in a dataset before applying any analytical or predictive techniques. Visualizations such as **histograms, boxplots, and heatmaps** play a key role in revealing hidden characteristics of data.

1. Histogram
A **histogram** displays the **distribution of a numerical variable** by showing the frequency of data values in different ranges (bins).
How it Helps:
- Identifies whether data is **normally distributed**, skewed, or has outliers.
- Shows **concentration** of values — where most data points lie.

Example:
A retail company analyzing the **age distribution of customers**:
- If the histogram shows most customers are between **25 and 35**, marketing campaigns can be targeted at that age group.

2. Boxplot
A **boxplot** visualizes the **spread and skewness** of data using five summary statistics: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. It also identifies **outliers**.
How it Helps:
- Quickly shows if data is **symmetrical or skewed**.
- Highlights **extreme values** (outliers), which may require correction.

Example:
A company comparing **monthly sales revenue across different branches**:
- A branch with many extreme outliers may indicate inconsistent performance or data entry issues.
- Boxplots help compare performance across branches in one single chart.

3. Heatmap
A **heatmap** is a 2D graphical representation where values are shown as colors. Heatmaps are often used to visualize **correlation matrices**.
How it Helps:
- Shows **relationship strength** between variables.
- Helps identify **which variables are useful for prediction**.

Example:
In a telecom company analyzing **customer churn**:
- A heatmap may show a strong positive correlation between *contract length* and *customer retention*.
- It may also show a strong negative correlation between *monthly charges* and *customer satisfaction*, helping the business revise pricing strategies.

7. Explain the assumptions and applications of simple linear regression in business analytics.

Simple Linear Regression
Simple Linear Regression is a statistical technique that models the relationship between two variables:
- **Independent Variable (X)** – the predictor or cause
- **Dependent Variable (Y)** – the outcome or effect

It predicts Y based on X using a straight-line equation:
$Y=a+bX$
where **a** = intercept and **b** = slope.

Assumptions of Simple Linear Regression
1. **Linearity**
   The relationship between X and Y should be linear. Changes in X lead to proportional changes in Y.
2. **Independence of Observations**
   Each observation in the dataset should be independent of the others.

3. **Homoscedasticity (Constant Variance)**
   The spread of errors (residuals) should remain constant across all values of X. No pattern should be visible in residual plots.
4. **Normality of Residuals**
   The residuals (differences between actual and predicted values) should follow a normal distribution.
5. **No Multicollinearity**
   Although this applies mainly to multiple regression, in simple regression the single predictor should not be correlated with other variables influencing Y.

Applications in Business Analytics
Simple Linear Regression is widely used to **forecast outcomes**, **identify trends**, and **support decision-making**. Some common business applications include:

Business Domain	Application	Purpose
Sales & Marketing	Predicting future sales based on advertising expenditure	Budget planning & marketing strategy
Finance	Estimating stock prices based on market indices	Portfolio and risk management
Human Resources	Predicting employee performance using experience or training hours	Recruitment and training evaluation
Retail & Inventory	Forecasting product demand using past sales data	Efficient inventory planning
Operations	Estimating production output based on machine running time	Scheduling and resource utilization

Example
A company wants to assess the effect of advertising budget (X) on monthly sales revenue (Y). Using past data, linear regression can estimate how much additional sales can be expected for every extra unit of money spent on advertising.

8. Compare linear regression and logistic regression in terms of objectives and outputs.

Comparison Between Linear Regression and Logistic Regression

Feature	Linear Regression	Logistic Regression
Objective	To model the relationship between a continuous dependent variable and one or more independent variables.	To model the probability of a categorical dependent variable (usually binary) based on independent variables.
Type of Dependent Variable	Continuous (e.g., sales, temperature, profit).	Categorical (e.g., Yes/No, Churn/Not Churn, Disease/No Disease).
Output	Produces a **numeric value** as output.	Produces a **probability value** between 0 and 1, which is later converted into a class label.
Model Equation	$Y = a + bX$ (predicts actual numerical values).	Uses the logistic (sigmoid) function:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(a+bX)}}$$

9. Discuss how model performance metrics like accuracy and recall help assess a model's reliability in real-world cases.

Model performance metrics such as **accuracy** and **recall** play an important role in evaluating how reliable and useful a predictive model will be when applied in real-world situations.

Accuracy

Accuracy measures the **overall correctness** of a classification model. It is calculated as the ratio of correctly predicted instances to the total instances.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

Accuracy is useful when:

- The dataset is **balanced** (i.e., both classes have similar number of observations).
- The cost of misclassification is **similar** for all classes.

Example:

In email classification (Spam vs. Not Spam), if both categories are equally represented, a model with high accuracy indicates strong overall performance.

However, accuracy alone can be **misleading** in **imbalanced datasets.** For example, if 95% of patients are healthy and 5% have a disease, predicting *everyone as healthy* will give 95% accuracy but is useless for diagnosing the disease.

Recall

Recall measures the model's ability to correctly identify **actual positive cases.**

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall is especially important in cases where:

- **Missing positive cases has serious consequences**, such as medical diagnosis, fraud detection, or identifying defective products.

Example:

In healthcare, failing to detect a disease (false negative) can be dangerous. A model with **high recall** ensures that most actual patients are correctly identified.

Real-World Reliability

- **Accuracy** ensures general correctness but is reliable only when class distribution is balanced.
- **Recall** ensures that **positive cases are not missed**, which is crucial in sensitive applications.

When used together, these metrics provide a more **complete evaluation** of the model's performance. In practice, metrics like **Precision, F1-score, and AUC-ROC** are also considered for more robust assessment.

10. Write a detailed note on customer churn prediction models and how predictive analytics supports business retention strategies.

Customer churn refers to the phenomenon where customers cease to use a company's products or services. Predicting churn is crucial for businesses because acquiring new customers is often more expensive than retaining existing ones. **Customer Churn Prediction Models** use data-driven and machine learning approaches to identify customers who are likely to leave, enabling organizations to take corrective actions on time.

Customer Churn Prediction Models

Churn prediction models typically analyze historical customer behavior patterns to determine the probability that a customer will discontinue service. These models use features such as:

- Usage patterns (e.g., call duration, internet usage, purchase frequency)
- Customer service interactions (number of complaints, support tickets)
- Demographic information (location, income, age)
- Account or subscription details (plan type, tenure, payment history)

Common predictive techniques include:

1. **Logistic Regression:**
   Used for binary classification (churn / no churn). It calculates the likelihood of churn based on weighted input features.
2. **Decision Trees and Random Forest:**
   Useful for capturing complex, non-linear relationships in customer behavior. Random Forest improves prediction accuracy by combining multiple tree outputs.
3. **Support Vector Machines (SVM):**
   Helps create a boundary between churn and non-churn classes, especially useful in high-dimensional datasets.
4. **Neural Networks & Deep Learning:**
   Used when dealing with large-scale customer behavioral datasets, especially for telecom, e-commerce, and banking sectors.

Predictive Analytics in Business Retention

Predictive analytics helps organizations **anticipate churn** rather than reacting after it occurs. It supports retention strategies in several ways:

- **Identifying At-Risk Customers:**
  The model highlights customers who exhibit behaviors similar to those who have churned previously. This allows targeted intervention.
- **Personalized Retention Campaigns:**
  Based on churn probability, businesses can offer customized discounts, loyalty rewards, plan upgrades, or direct engagement to encourage continued usage.
- **Improving Customer Experience:**
  Analysis of churn drivers helps organizations enhance product features, improve service quality, or simplify user interfaces.
- **Resource Optimization:**
  Instead of spending on all customers, businesses can **prioritize** retention efforts on high-value customers who are most likely to churn.