WB Board Class 12 Data Science Question Paper with Solutions(Memory Based)

Time Allowed :3 Hour	Maximum Marks :60	Total Questions :24

General Instructions

Read the following instructions very carefully and strictly follow them:

- Answers to this Paper must be written on the paper provided separately.
- You will not be allowed to write during the first 15 minutes
- This time is to be spent in reading the question paper.
- The time given at the head of this Paper is the time allowed for writing the answers,
- The paper has four Sections.
- Section A is compulsory - All questions in Section A must be answered.
- You must attempt one question from each of the Sections B, C and D and one other question from any Section of your choice.

1. Explain the steps involved in Data Wrangling and why it is essential before analysis.

Correct Answer: Data wrangling involves collecting, cleaning, transforming, and structuring raw data into a usable format. It is essential because high-quality, well-structured data ensures accurate analysis and reliable model performance.

Solution: Concept: Data Wrangling (or data preprocessing) is the process of converting raw, messy data into a clean and structured format suitable for analysis or machine learning. Since real-world data is often incomplete or inconsistent, wrangling is a critical preparation step.
Step 1: Data Collection
Gather data from multiple sources:

- Databases, APIs, surveys, logs
- Structured and unstructured datasets

The goal is to consolidate relevant data for analysis.
Step 2: Data Cleaning
Remove errors and inconsistencies:

- Handle missing values
- Remove duplicates
- Correct formatting errors

This improves data quality.
Step 3: Data Transformation
Convert data into usable formats:

- Normalization or scaling
- Encoding categorical variables
- Aggregation or feature engineering

Step 4: Data Integration
Combine data from multiple sources:

- Merge datasets
- Resolve schema conflicts

This creates a unified dataset.
Step 5: Data Structuring
Organize data into analysis-ready formats:

- Tables, matrices, or data frames
- Proper labeling and indexing

Step 6: Why Data Wrangling is Essential
It is crucial because:

- Poor-quality data leads to incorrect insights
- Improves model accuracy and reliability
- Reduces bias and noise in analysis

Quick Tip

Good analysis starts with good data — data wrangling ensures your data is clean, consistent, and ready for meaningful insights.

---

2. What is the difference between Structured, Semi-structured, and Unstructured data?

Correct Answer: Structured data is highly organized in tabular formats, semi-structured data has partial organization with flexible schemas, and unstructured data lacks a predefined format and is more complex to analyze.

Solution: Concept: Data can be categorized based on how it is organized and stored. Understanding the differences between structured, semi-structured, and unstructured data is important for selecting appropriate storage, processing, and analysis techniques.
Step 1: Structured Data
Structured data is highly organized and follows a fixed schema:

- Stored in rows and columns (tables)
- Easily searchable and analyzable

Examples:

- Relational databases (SQL tables)
- Spreadsheets (Excel)

Step 2: Semi-structured Data
Semi-structured data has some organizational properties but no rigid schema:

- Uses tags or key-value pairs
- Flexible structure

Examples:

- JSON and XML files
- Emails and log files

Step 3: Unstructured Data
Unstructured data has no predefined format:

- Difficult to store in traditional databases
- Requires advanced processing techniques

Examples:

- Images, videos, audio files
- Social media posts and documents

Step 4: Key Differences

- **Schema:** Fixed (structured) vs flexible (semi-structured) vs none (unstructured)
- **Ease of Analysis:** Easy → Moderate → Difficult
- **Storage:** Relational DB → NoSQL → Data lakes

Quick Tip

Structured = Organized tables,
Semi-structured = Flexible tagged data,
Unstructured = Raw data like images, videos, and text.

---

3. Describe the Data Science Lifecycle from data collection to deployment.

Correct Answer: The Data Science Lifecycle includes stages such as data collection, data preparation, exploration, modeling, evaluation, and deployment, ensuring a systematic approach from raw data to real-world implementation.

Solution: Concept: The Data Science Lifecycle is a structured process that guides data-driven projects from gathering raw data to deploying actionable solutions. It ensures systematic development, validation, and implementation of data science models.
Step 1: Data Collection
Gather raw data from various sources:

- Databases, APIs, sensors, web scraping
- Internal and external data sources

Step 2: Data Preparation (Wrangling)
Clean and preprocess the data:

- Handle missing values and duplicates
- Normalize and transform features

This ensures data quality and usability.
Step 3: Exploratory Data Analysis (EDA)
Understand patterns and relationships:

- Visualizations and summary statistics
- Detect trends, correlations, and anomalies

Step 4: Feature Engineering
Create meaningful input variables:

- Feature selection and extraction
- Encoding categorical variables

This improves model performance.
Step 5: Model Building
Develop predictive or analytical models:

- Select algorithms (regression, classification, clustering)
- Train models on prepared data

Step 6: Model Evaluation
Assess model performance:

- Use metrics like accuracy, precision, RMSE
- Validate using test data

Step 7: Deployment
Implement the model in real-world applications:

- Integrate into software systems or dashboards
- Enable real-time predictions

Step 8: Monitoring and Maintenance
Ensure long-term effectiveness:

- Track model performance

- Update with new data when needed

Quick Tip

Data Science Lifecycle: Collect → Clean → Explore → Model → Evaluate → Deploy → Monitor.

---

4. What is Data Cleaning, and how do you handle missing values in a dataset?

Correct Answer: Data cleaning is the process of correcting or removing inaccurate, incomplete, or inconsistent data. Missing values can be handled by deletion, imputation (mean/median/mode), or predictive methods depending on the context.

Solution: Concept: Data Cleaning is a key preprocessing step in data science that ensures datasets are accurate, consistent, and suitable for analysis. Real-world data often contains errors, duplicates, and missing values that can negatively affect model performance.
Step 1: What is Data Cleaning?
Data cleaning involves:

- Removing duplicates
- Fixing formatting errors
- Handling missing or inconsistent data

It improves data reliability and quality.
Step 2: Understanding Missing Values
Missing data may occur due to:

- Data entry errors
- Sensor failures
- Incomplete surveys

Handling them correctly is essential to avoid biased analysis.
Step 3: Method 1 — Deletion
Remove rows or columns with missing values:

- Useful when missing data is minimal
- Risky if large portions of data are removed

Step 4: Method 2 — Imputation
Fill missing values using statistical measures:

- Mean (numerical data)
- Median (robust to outliers)
- Mode (categorical data)

Step 5: Method 3 — Advanced Techniques
More sophisticated approaches include:

- Predictive modeling (e.g., regression)
- Interpolation for time-series data
- KNN or machine learning-based imputation

Step 6: Choosing the Right Method
The strategy depends on:

- Amount of missing data
- Data type and distribution
- Impact on analysis goals

Quick Tip

Clean data leads to reliable results — always analyze the pattern of missing values before choosing a handling method.

---

5. Explain the concept of Data Normalization and why it is used in scaling features.

Correct Answer: Data normalization is the process of rescaling numerical features to a common range (often 0–1). It is used to ensure all features contribute equally to model training and to improve convergence in machine learning algorithms.

Solution: Concept: Data normalization is a feature scaling technique used to bring different numerical variables onto a similar scale. Since real-world datasets often contain features with varying units and magnitudes, normalization ensures fair contribution during model training.
Step 1: What is Data Normalization?
Normalization rescales values to a fixed range, typically:

- Between 0 and 1 (Min-Max normalization)
- Sometimes between -1 and 1

It preserves relationships while changing scale.
Step 2: Common Normalization Formula
Min-Max normalization is given by:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where:

- x = original value
- x' = normalized value

Step 3: Why Feature Scaling is Needed
Different feature scales can cause issues:

- Large-value features dominate smaller ones
- Slower convergence in optimization algorithms

Step 4: Benefits of Normalization
Normalization helps:

- Improve training speed
- Ensure equal feature contribution
- Enhance performance of distance-based algorithms (e.g., KNN, clustering)

Step 5: When to Use Normalization
It is especially useful for:

- Gradient descent-based models
- Neural networks
- Algorithms sensitive to scale

Quick Tip

Normalization rescales features to a common range, preventing large-value features from dominating model learning.

---

6. Define Mean, Median, and Mode and explain which is most affected by outliers.

Correct Answer: Mean is the arithmetic average, median is the middle value in an ordered dataset, and mode is the most frequent value. The mean is most affected by outliers.

Solution: Concept: Mean, median, and mode are measures of central tendency used to summarize a dataset. Each provides a different perspective on the typical value depending on data distribution.
Step 1: Mean
The mean is the arithmetic average:

$$\text{Mean} = \frac{\sum x_i}{n}$$

where:

- x_i = data values
- n = number of observations

It uses all values in the dataset.
Step 2: Median
The median is the middle value when data is sorted:

- If n is odd $\rightarrow$ middle value
- If n is even $\rightarrow$ average of two middle values

It is less sensitive to extreme values.
Step 3: Mode
The mode is the most frequently occurring value:

- A dataset may have one, multiple, or no modes

It is useful for categorical data.
Step 4: Effect of Outliers
Outliers are extreme values that differ significantly from others:

- Mean is highly affected (uses all values)
- Median is resistant to outliers
- Mode is usually unaffected

Step 5: Conclusion
Among the three measures:

- Mean changes significantly with outliers
- Median is the most robust

Quick Tip

Outliers distort the mean the most, while the median remains stable and is preferred for skewed data.

---

7. What is a Hypothesis Test, and what do the Null and Alternative hypotheses represent?

Correct Answer: A hypothesis test is a statistical method used to make decisions based on data. The null hypothesis represents no effect or status quo, while the alternative hypothesis represents a significant effect or difference.

Solution: Concept: Hypothesis testing is a statistical framework used to evaluate assumptions about a population using sample data. It helps determine whether observed results are statistically significant or due to random chance.
Step 1: What is a Hypothesis Test?
A hypothesis test involves:

- Formulating assumptions about a population
- Using sample data to evaluate those assumptions
- Making a decision using statistical evidence

Step 2: Null Hypothesis (H_0)
The null hypothesis represents:

- No effect, no difference, or status quo
- A baseline assumption to test against

Example: A new drug has no effect compared to the old one.
Step 3: Alternative Hypothesis (H_1 or H_a)
The alternative hypothesis represents:

- A significant effect or difference
- What the researcher aims to support

Example: The new drug is more effective than the old one.
Step 4: Decision Making
Based on statistical evidence:

- Reject H_0 if strong evidence exists
- Fail to reject H_0 if evidence is insufficient

Step 5: Importance
Hypothesis testing is widely used in:

- Scientific research
- A/B testing
- Quality control and data analysis

Quick Tip

H_0 = No effect (default assumption),
H_a = There is an effect (research claim).

---

8. What is a DataFrame in Pandas, and how does it differ from a Series?

Correct Answer: A DataFrame in Pandas is a two-dimensional labeled data structure with rows and columns, while a Series is a one-dimensional labeled array. A DataFrame can contain multiple Series as columns.

Solution: Concept: In the Pandas library (Python), Series and DataFrames are core data structures used for data manipulation and analysis. They differ primarily in dimensionality and structure.
Step 1: What is a Series?
A Series is a one-dimensional labeled array:

- Contains a single column of data
- Has an index for labeling values

Example: A column of student marks.
Step 2: What is a DataFrame?
A DataFrame is a two-dimensional table:

- Multiple rows and columns
- Each column can have different data types

It resembles a spreadsheet or SQL table.
Step 3: Key Differences

- **Dimensionality:** Series = 1D, DataFrame = 2D
- **Structure:** Series = single column, DataFrame = multiple columns
- **Complexity:** DataFrame can store heterogeneous data

Step 4: Relationship Between Them
A DataFrame is essentially a collection of Series:

- Each column in a DataFrame is a Series
- Series share a common index

Step 5: Use Cases

- Series → Single-variable analysis
- DataFrame → Tabular datasets and data analysis workflows

Quick Tip

Series = 1D labeled data,
DataFrame = 2D table made of multiple Series.

---

9. How do you perform a Join and Merge operation on two different datasets in Python?

Correct Answer: In Python (Pandas), datasets can be combined using `merge()` for database-style joins on keys and `join()` for index-based alignment. These operations allow combining rows and columns from multiple datasets.

Solution: Concept: Combining datasets is a common task in data analysis. In Pandas, the two primary methods are `merge()` and `join()`, which allow combining data based on keys or indices, similar to SQL joins.
Step 1: Using Merge Operation
The `merge()` function combines datasets based on one or more common columns:

- Similar to SQL joins
- Allows inner, left, right, and outer joins

Example:

```
import pandas as pd

df1 = pd.DataFrame({'id': [1, 2], 'name': ['A', 'B']})
df2 = pd.DataFrame({'id': [1, 2], 'score': [90, 85]})

merged = pd.merge(df1, df2, on='id', how='inner')
```

Step 2: Types of Merge Joins

- **Inner Join:** Common rows only
- **Left Join:** All rows from left dataset
- **Right Join:** All rows from right dataset
- **Outer Join:** All rows from both datasets

Step 3: Using Join Operation
The `join()` method combines datasets based on index alignment:

- Default is left join
- Useful when indices represent relationships

Example:

```
df1 = df1.set_index('id')
df2 = df2.set_index('id')

joined = df1.join(df2)
```

Step 4: Key Differences

- **merge():** Column-based joining, more flexible
- **join():** Index-based joining, simpler syntax

Step 5: When to Use Which

- Use `merge()` for relational-style joins
- Use `join()` when working with indexed data

Quick Tip

Use `merge()` for SQL-like joins on columns and `join()` for combining datasets based on index alignment.

---

10. What are the key principles of Effective Data Visualization?

Correct Answer: Effective data visualization relies on clarity, simplicity, accuracy, proper chart selection, meaningful use of color, and focus on storytelling to communicate insights clearly.

Solution: Concept: Effective data visualization is the practice of presenting data in a graphical format that communicates insights clearly and efficiently. Good visualizations simplify complex information and help audiences understand patterns, trends, and relationships.
Step 1: Clarity and Simplicity
A good visualization should:

- Avoid clutter and unnecessary elements
- Focus on the main message

Simple visuals are easier to interpret.
Step 2: Choose the Right Chart Type
Selecting appropriate visuals is essential:

- Bar charts for comparisons
- Line charts for trends
- Pie charts for proportions

Wrong chart types can mislead viewers.
Step 3: Accuracy and Integrity
Ensure:

- Proper scaling of axes
- No distortion of data

Misleading visuals reduce credibility.
Step 4: Effective Use of Color
Colors should:

- Highlight key insights
- Maintain contrast and readability
- Avoid excessive or distracting palettes

Step 5: Labeling and Context
Provide clear context:

- Titles, axis labels, and legends
- Annotations for key points

Step 6: Storytelling with Data
Visualization should:

- Guide the audience through insights
- Emphasize patterns and conclusions

A narrative improves engagement and understanding.

Quick Tip

Good visualizations are clear, honest, and purposeful — choose the right chart, reduce clutter, and tell a compelling data story.

---

11. When should you use a Box Plot versus a Bar Chart?

Correct Answer: Use a Box Plot to visualize data distribution, spread, and outliers, while a Bar Chart is best for comparing categorical values or aggregated quantities.

Solution: Concept: Box plots and bar charts serve different purposes in data visualization. Choosing the correct chart depends on whether you want to analyze data distribution or compare categories.
Step 1: When to Use a Box Plot
A box plot (box-and-whisker plot) is used to show statistical distribution:

- Displays median, quartiles, and range
- Highlights outliers
- Useful for comparing distributions across groups

Use cases:

- Salary distribution across departments
- Exam score variability

Step 2: When to Use a Bar Chart
A bar chart compares values across categories:

- Shows counts, sums, or averages
- Easy to interpret categorical comparisons

Use cases:

- Sales by region
- Number of students per course

Step 3: Key Differences

- **Purpose:** Distribution (box plot) vs comparison (bar chart)
- **Outliers:** Visible in box plots, not in bar charts
- **Data Type:** Continuous data vs categorical aggregates

Step 4: Choosing the Right Chart

- Use box plots for statistical insights and spread
- Use bar charts for simple comparisons and summaries

Quick Tip

Box Plot = Distribution and outliers,
Bar Chart = Comparing category values.
