CBSE Class 10 Data Science Question Paper 2026 with Solution Memory Based

Time Allowed :3 Hours	**Maximum Marks :**80	**Total Questions :**38

General Instructions

1. Please check that this question paper contains 23 printed pages.
2. Q.P. Code given on the right hand side of the question paper should be written on the title page of the answer-book by the candidate.
3. Please check that this question paper contains 37 questions.
4. 15 minute time has been allotted to read this question paper. The question paper will be distributed at 10.15 a.m. From 10.15 a.m. to 10.30 a.m., the candidates will read the question paper only and will not write any answer on the answer-book during this period.

1. Which Python library is primarily used for Data Manipulation and Data Frames?

Correct Answer: Pandas

Solution:

Step 1: Understanding the Concept:
In Python's data science ecosystem, different libraries serve specific, distinct purposes.
The requirement for "Data Manipulation" and handling "Data Frames" points specifically to the most popular library designed for structured data.

Step 3: Detailed Explanation:
Pandas is an open-source Python library providing high-performance, easy-to-use data structures.
The primary data structure in Pandas is the **DataFrame**, which is a two-dimensional, size-mutable, and potentially heterogeneous tabular data structure with labeled axes.
It allows for operations like merging, reshaping, selecting, cleaning, and wrangling data efficiently.

Step 4: Final Answer:
The Python library primarily used for Data Manipulation and Data Frames is **Pandas**.

Quick Tip

Remember: NumPy is primarily for numerical computations (arrays), while Pandas is built on top of NumPy for working with tabular data (DataFrames).

2. Identify the process of dividing a dataset into 100 equal parts.

Correct Answer: Percentiles

Solution:

Step 1: Understanding the Concept:
In statistics, quantiles are values that divide a sorted dataset into continuous intervals with equal probabilities.

Step 2: Key Formula or Approach:
Common quantile divisions include:
1. Quartiles: Divide data into 4 equal parts.
2. Deciles: Divide data into 10 equal parts.
3. Percentiles: Divide data into 100 equal parts.

Step 3: Detailed Explanation:
The process of dividing a sorted dataset into exactly 100 equal parts is known as calculating **Percentiles**.
The n^{th} percentile is the specific value below which $n\%$ of the data points fall.
For example, the 50^{th} percentile is identical to the Median, dividing the dataset into two perfectly equal halves.

Step 4: Final Answer:
The correct process is known as **Percentiles**.

Quick Tip

The 25^{th} percentile corresponds to the 1^{st} Quartile (Q_1), and the 75^{th} percentile corresponds to the 3^{rd} Quartile (Q_3).

---

3. Name the type of join used to combine rows from two tables based on a related column between them.

Correct Answer: Inner Join

Solution:

Step 1: Understanding the Concept:
Relational database management systems (RDBMS) store data across different tables.
To retrieve meaningful, combined information from multiple tables, we use "Join" operations

based on common keys.

Step 3: Detailed Explanation:
An **Inner Join** (often referred to simply as a Join) is the most fundamental and commonly used type of join.
It returns only the records that have matching values in the related column of both tables.
If a row in Table A does not have a corresponding, matching row in Table B, that specific row is entirely excluded from the final result set.

Step 4: Final Answer:
The type of join used to combine rows based on related, matching columns is an **Inner Join**.

Quick Tip

If you need to keep all records from the left table regardless of whether there is a match in the right table, you should use a **Left Join**.

---

4. In a dataset, if the mean is significantly higher than the median, what does it indicate about the distribution?

Correct Answer: Positively Skewed (Right-Skewed)

Solution:

Step 1: Understanding the Concept:
Skewness refers to the asymmetry or distortion that deviates from the symmetrical bell curve in a dataset.

Step 2: Key Formula or Approach:
- Symmetrical Distribution: Mean = Median = Mode
- Right-Skewed (Positive Skewness): Mean > Median > Mode
- Left-Skewed (Negative Skewness): Mean < Median < Mode

Step 3: Detailed Explanation:
When the Mean is significantly higher than the Median, it implies that extreme high values (outliers) exist on the upper end of the scale.
These large outliers disproportionately pull the Mean towards the right, while the Median remains resistant to them.
This creates an elongated "tail" that extends towards the right side of the distribution graph.

Step 4: Final Answer:
It heavily indicates that the dataset possesses a **Positively Skewed** or **Right-Skewed** distribution.

Quick Tip

A classic real-world example of positive skewness is national income distribution, where a few ultra-wealthy individuals pull the average (mean) income far above the typical (median) salary.

---

5. Define Encryption in the context of data security.

Correct Answer: Converting readable data into an unreadable format to prevent unauthorized access.

Solution:

Step 1: Understanding the Concept:
Data security involves protecting digital information from unauthorized access, corruption, or theft throughout its entire lifecycle.

Step 3: Detailed Explanation:
Encryption is the core cryptographic process of encoding sensitive information.
It specifically converts the original representation of data, known as **plaintext**, into a scrambled, alternative form known as **ciphertext**.
Only explicitly authorized parties possessing the correct cryptographic **key** can decrypt the ciphertext back into readable plaintext.
This vital process ensures absolute confidentiality, even if malicious actors intercept the data during transmission.

Step 4: Final Answer:
Encryption is defined as the process of converting readable data into an unreadable format to prevent unauthorized access.

Quick Tip

Modern encryption techniques generally fall into two broad categories: Symmetric (using the exact same key to lock and unlock) and Asymmetric (using distinct Public and Private keys).

---

6. State whether Selection Bias occurs during data collection or during model deployment.

Correct Answer: During data collection

Solution:

Step 1: Understanding the Concept:
Statistical bias in data science refers to systematic errors that occur when certain elements of a target dataset are heavily overrepresented or underrepresented.

Step 3: Detailed Explanation:
Selection bias occurs specifically when the sample data used to train a machine learning model is not representative of the actual real-world population.
This fundamental error happens early during the **data collection** phase.
If researchers fail to properly randomize the selection of individuals, groups, or data points, the resulting sample becomes inherently flawed.
Consequently, any model built upon this skewed data will produce inaccurate, biased predictions when deployed.

Step 4: Final Answer:
Selection Bias definitively occurs during the **data collection** phase.

Quick Tip

To aggressively prevent selection bias, data scientists must utilize robust random sampling techniques, ensuring every member of a population holds an equal chance of being selected.

---

7. Which of the following best differentiates Structured and Unstructured data?

Correct Answer: Structured data follows a pre-defined format (e.g., SQL tables); Unstructured data does not (e.g., images).

Solution:

Step 1: Understanding the Concept:
Digital data is broadly classified based on its structural organization and how easily standard computing machines can process it.

Step 3: Detailed Explanation:
Structured Data:
- It is highly organized and strictly adheres to a predefined, rigid schema.
- It is typically stored in traditional tabular formats comprising distinct rows and columns.
- It is straightforward to search, query, and analyze using standard tools like SQL.
- **Example:** Relational databases containing detailed spreadsheets of student names and exam marks.

Unstructured Data:
- It fundamentally lacks a predefined data model or strict organizational framework.
- It is frequently rich in raw text, audio, or complex multimedia.
- It demands advanced, specialized tools like AI algorithms or Natural Language Processing for deep analysis.
- **Example:** Raw emails, social media feeds, digital images, and video files.

Step 4: Final Answer:
Structured data follows a pre-defined format (e.g., SQL tables), whereas Unstructured data lacks this format completely (e.g., images and videos).

> Quick Tip
>
> Semi-structured data (like JSON or XML formats) operates as a middle ground, possessing organizational tags and hierarchies but lacking a rigid tabular structure.

---

8. Why is the Median considered a more robust measure of central tendency than the Mean when outliers are present?

Correct Answer: Because the Median is completely unaffected by extreme numerical values (outliers).

Solution:

Step 1: Understanding the Concept:
Robustness in statistical terms refers to the distinct ability of a mathematical measure to remain highly reliable even when the dataset contains significant errors or extreme values.

Step 3: Detailed Explanation:
- The **Mean** is calculated mathematically by summing all data values and dividing by the total count ($\frac{\sum x}{n}$).
- Because absolutely every value directly contributes to this sum, a single massively large or small value (an outlier) will drastically warp the final result.
- Conversely, the **Median** simply represents the middle positional value of an ascendingly sorted dataset.
- Its value relies entirely on the middle position of the data rather than the actual numerical magnitude of the surrounding numbers.
- If you alter the absolute largest value in a set to something astronomically larger, the middle position remains unchanged, allowing the Median to stay perfectly stable.

Step 4: Final Answer:
The Median is widely considered robust because it is strictly positional and thus actively ignores the disruptive magnitude of extreme outliers.

Quick Tip

Always utilize the Mean for normal, bell-shaped distributions without outliers, but switch to the Median for heavily skewed distributions (such as analyzing global house prices or salaries).

---

9. Explain the concept of the Central Limit Theorem and its significance in data analysis.

Correct Answer: It dictates that sample means approach a normal distribution as the sample size increases.

Solution:

Step 1: Understanding the Concept:
The Central Limit Theorem (CLT) is a foundational, critical theorem in probability theory and advanced statistics.

Step 3: Detailed Explanation:
The CLT definitively states that if you take sufficiently large random samples from any given population (completely regardless of the population's original underlying distribution), the resulting distribution of those sample means will invariably approximate a normal, bell-shaped distribution.
The overall mean of these collected sample means will strictly equal the original population mean.
Furthermore, the standard deviation of this newly formed distribution (known as the standard error) actively decreases as your chosen sample size increases.

Significance in Data Analysis:
1. It safely empowers statisticians to perform parametric tests (such as t-tests and Z-tests) even when the underlying population is visibly skewed or not normally distributed.
2. It serves as the mathematical bedrock for constructing accurate confidence intervals.
3. It vastly simplifies complex predictive modeling by granting the assumption of normality when dealing with massive datasets.

Step 4: Final Answer:
The CLT establishes that the distribution of sample means becomes normal for large samples, thereby enabling widespread parametric statistical methodologies.

Quick Tip

In the vast majority of practical, real-world scenarios, a baseline sample size of $n \geq 30$ is widely considered fully sufficient for the CLT to properly take effect.

10. What is Data Merging? Mention one scenario where it is required.

Correct Answer: The computational process of combining multiple datasets based on a shared common identifier.

Solution:

Step 1: Understanding the Concept:
Data merging is a crucial data preprocessing technique actively utilized to integrate fragmented data from disparate sources.

Step 3: Detailed Explanation:
Data merging fundamentally involves horizontally joining distinct data frames or database tables by leveraging a shared common key or overlapping column.
Real-World Scenario:
Consider a large-scale retail company managing its digital databases.
They might maintain one specialized table exclusively for "Customer Details" (housing Name, Address, and Phone Number) and a completely separate table for "Transaction Records" (housing Item Purchased, Price, and Date).
To accurately analyze which specific geographic region purchases the most expensive items, the company urgently needs to **merge** these two independent tables using a unique 'Customer ID' column as the central linking mechanism.

Step 4: Final Answer:
Data Merging is defined as the process of linking separate datasets via common columns, heavily required when related target information is stored across multiple independent tables.

Quick Tip

Always double-check that the common column used for the merging process features the exact same data type (e.g., integer or string) in both target datasets to entirely avoid compilation errors.

11. Which of the following represents the correct sequence of the four steps of the Statistical Problem-Solving Process?

Correct Answer: Formulate, Collect, Analyze, Interpret

Solution:

Step 1: Understanding the Concept:
The Statistical Problem-Solving Process is a standardized, logical framework (frequently referred to as the GAISE framework) methodically utilized to investigate phenomena using data.

Step 3: Detailed Explanation:
The structured process rigorously follows four explicit steps:
1. **Formulate a Statistical Investigative Question:**
You must define the core problem with extreme clarity.
The targeted question must be tangibly answerable through data gathering and must inherently account for natural variability (e.g., "Do local students actively study more on weekdays compared to weekends?").
2. **Collect Data:**
You must thoughtfully design and implement a strategic plan to gather relevant data.
This stage heavily involves choosing appropriate sampling methodologies, launching surveys, or running controlled experiments while actively ensuring the procedure thoroughly minimizes unwanted bias.
3. **Analyze Data:**
You must utilize clear graphical displays (such as histograms or box plots) alongside precise numerical summaries (like mean, median, and standard deviation) to actively identify underlying patterns and trends within the gathered data.
4. **Interpret Results:**
You must draw logical, evidence-based conclusions rooted purely in the prior analysis.
You must directly relate all findings back to the original formulated question and transparently acknowledge any statistical limitations or mathematical uncertainty inherent in the data.

Step 4: Final Answer:
The proper sequence is: 1. Formulate Question, 2. Collect Data, 3. Analyze Data, and 4. Interpret Results.

> Quick Tip
>
> Never skip the formulation stage; without a sharply defined objective, subsequent data collection rapidly becomes unfocused, aimless, and highly inefficient.

---

12. Which of the following best defines and explains Recall Bias and Survivor Bias?

Correct Answer: Recall bias is a human memory-based error; Survivor bias is an analytical error focusing solely on "surviving" data.

Solution:

Step 1: Understanding the Concept:
Cognitive and statistical biases represent systematic, repeated errors that routinely lead re-

searchers to an incorrect, flawed estimate of a targeted effect or association.

Step 3: Detailed Explanation:
1. **Recall Bias:**
- This specific bias vividly occurs when participating individuals in a retrospective study fail to remember past historical events accurately or accidentally omit critical details.
- It is incredibly common in subjective surveys inquiring heavily about the past.
- **Example:** Patients who have actively suffered a severe disease might intensely remember their past dietary habits far more vividly than entirely healthy people, rapidly leading to heavily skewed, unreliable data.

2. **Survivor Bias:**
- This logical error heavily occurs when a focused researcher examines only the specific people or entities that successfully "survived" a rigorous process, while systematically ignoring all those that unfortunately failed.
- It directly leads to highly distorted, overly optimistic analytical conclusions.
- **Example:** Passionately studying only massively successful tech startups to isolate the ultimate "secret to success," while dangerously ignoring the thousands of completely failed startups that utilized those exact same methods.

Step 4: Final Answer:
Recall bias represents a memory-related cognitive distortion, while Survivor bias is a severe logical error directly caused by improperly excluding unsuccessful or "non-surviving" data points from the final mathematical analysis.

Quick Tip

A historically famous example of survivor bias involved World War II bomber planes. Engineers heavily wanted to reinforce areas littered with bullet holes, until an astute statistician realized they were only observing the planes that successfully **returned** (the survivors), meaning the untouched areas actually represented the true fatal weak points.

---

13. Compare CSV, Spreadsheets, and SQL as data storage formats. Which one is definitively best for handling massive, interrelated datasets?
.

Correct Answer: SQL is distinctly the best format for handling massive, interrelated datasets.

Solution:

Step 1: Understanding the Concept:
Digital data can be actively stored in various distinct file formats depending directly on the required size, structural complexity, and real-time access requirements.

Step 3: Detailed Explanation:
- **CSV (Comma Separated Values):**
- It is an incredibly simple, plain text file format boasting a very small file size.
- It is remarkably easy to quickly share, but it completely lacks built-in security and physically cannot handle complex tabular relationships.
- **Spreadsheets (e.g., Microsoft Excel):**
- It heavily supports visual formatting, automated formulas, and interactive charts.
- It is functionally excellent for manual data entry and small-scale visual analysis.
- However, it rapidly becomes sluggish, unresponsive, and highly unstable when confronted with extremely large datasets (such as those exceeding one million rows).
- **SQL (Relational Databases):**
- It is intricately designed from the ground up to handle massive, enterprise-level amounts of data seamlessly.
- It strictly utilizes complex primary and foreign keys to robustly manage multi-table relationships.
- It natively supports massive concurrent user access, strict data security protocols, and incredibly complex querying capabilities.

Step 4: Final Answer:
While CSV remains wonderfully simple and Spreadsheets are distinctly user-friendly for minor tasks, **SQL** clearly stands as the absolute best format for managing massive, highly interrelated datasets due to its unmatched scalability and rigid relational structure.

Quick Tip

Professionals routinely use CSV strictly for lightweight data transfer between disparate programs, but they exclusively deploy robust SQL databases for building permanent, enterprise-grade software applications.
