cap5771sp25
February 26, 2025

Exam 1 part 1 version c

Question 1. According to data exploration demo, what is the meaning of the cursor.fetchall() in this code block?

```
cursor.execute("SELECT tracks.Name, albums.Title, artists.Name
            FROM tracks JOIN albums ON tracks.AlbumId = albums.AlbumId
              JOIN artists ON albums.ArtistId = artists.ArtistId")
tracks_albums_artists = cursor.fetchall()
for item in tracks_albums_artists:
    print(item)
```

(a) `It retrieves the output as the a of tuples`

(b) `It creates a new SQL table and stores the data`

(c) `It will crash your whole system`

(d) `It forms a new SQL query`

(e) `It retrieves only one row from the SQL output`

Question 2. According to lecture materials on exploratory data analysis, a data scientist observes two histograms with identical five-number summaries but differing bin structures during a quality control audit of pharmaceutical batch processing times. Which fundamental analytical limitation of boxplots does this scenario best demonstrate?

(a) `Boxplots obscure distribution modality and density variations across value ranges`

(b) `Boxplots exaggerate outlier impacts through 1.5xIQR whisker thresholds`

(c) `Boxplots confuse stakeholders by using medians instead of means`

(d) `Boxplots inaccurately represent continuous data as discrete quartiles`

(e) `Boxplots require minimum sample sizes exceeding 100 observations`

Question 3. Which of the following is an advantage of using a box plot over a histogram when analyzing a dataset?

(a) `A box plot shows the exact frequency of individual data points, which a histogram cannot.`

(b) `A box plot provides a summary of the dataset's distribution, including median, quartiles, and outliers, without requiring bin selection.`

(c) `A histogram is more effective for comparing multiple distributions than a box plot.`

(d) `A box plot is better for visualizing the overall shape of the distribution compared to a histogram.`

(e) `A histogram is less affected by outliers than a box plot.`

Question 4. In the context of data visualization, which of the following best describes the key difference between a histogram and a bar chart?

(a) `A histogram requires an equal number of values in each bin, while a bar chart does not.`

(b) `A bar chart always has bars touching each other, whereas a histogram never does.`

(c) `A histogram represents categorical data, whereas a bar chart represents numerical data.`

(d) `A histogram groups data into bins to show distribution, while a bar chart compares distinct categories.`

(e) `A bar chart is only used for time-series data, while a histogram is used for all types of data.`

Question 5. According to data visualization best practices, which of the following would NOT be a good reason to use a box plot?

(a) `To show and compare distribution values.`

(b) `To show correlations between two variables.`

(c) `To show data distribution shapes such as asymmetry and skewness.`

(d) `To identify and display outliers in a dataset.`

Question 6. According to the lectures on data exploration, which of the following does NOT describe the attribute of height, as measured in whole inches?

(a) `Discrete.`

(b) `Ratio-scaled.`

(c) `Numeric.`

(d) `Ordinal.`

(e) `Quantitative.`

Question 7. According to data exploration, which is an advantage of using a quantile-quantile (Q-Q) plot in data analysis?

(a) `It shows how two categorical variables are related.`

(b) `It replaces the need for boxplots and histograms in statistical analysis.`

(c) `It measures the dispersion of a dataset by identifying extreme values.`

(d) `It compares the distribution of a dataset against a theoretical distribution.`

(e) `It helps visualize the central tendency of a dataset.`

Question 8. According to lecture data exploration demo, when using Python to interact with a database, the cursor.execute(...) method is commonly used to execute SQL queries or commands. Which of the following statements correctly describes the usage of cursor.execute(...)?

(a) `cursor.execute(...) takes an SQL query as an argument and executes it.`

(b) `cursor.execute(...) is used to create new database tables.`

(c) `cursor.execute(...) is used to close the database connection.`

(d) `cursor.execute(...) is used to commit changes to the database.`

(e) `cursor.execute(...) returns a database connection object.`

Question 9. In data classification, how are attributes like ID numbers and ZIP codes categorized?

(a) `Only ZIP codes are quantitative, while ID numbers are qualitative.`

(b) `Quantitative attributes, because they consist of numerical values that can be used in mathematical operations.`

(c) `Both qualitative and quantitative attributes, depending on the context in which they are used.`

(d) `Qualitative attributes, because they represent categorical information without inherent numerical meaning.`

(e) `Neither qualitative nor quantitative, as they do not fit into standard data classification categories.`

Question 10. According to the lecture, in the context of data attribute types, which of the following statements best describes a key difference between interval-scaled and ratio-scaled attributes?

(a) `Interval-scaled attributes are always discrete, while ratio-scaled attributes are always continuous.`

(b) `Interval-scaled attributes can only be used in qualitative analysis, whereas ratio-scaled attributes are used in quantitative analysis.`

(c) `Ratio-scaled attributes cannot be negative, while interval-scaled attributes must always be positive.`

(d) `Both interval-scaled and ratio-scaled attributes are used only for nominal data types.`

(e) `Ratio-scaled attributes have a true zero point, allowing meaningful multiplicative comparisons, whereas interval-scaled attributes do not.`

Question 11. According to the lecture, if you have a table named Students with columns (sid, name, age, gpa), then in relational database terminology, each row in the Students table is referred to as which of the following?

(a) `Column`

(b) `Database`

(c) `Instance`

(d) `Schema`

(e) `Attribute`

Question 12. According to the lecture and slides on 2/7 (data-wrangling-part1), which of the following is never part of the data cleaning process?

(a) `Data integration`

(b) `Data smoothing`

(c) `Data discrepancy detection using rules`

(d) `Outlier detection`

(e) `Iteration`

Question 13. In the context of data management, what is a data dictionary?

(a) `A type of dictionary data structure in Python used to store key-value pairs.`

(b) `A collection of data facts stored in a csv format.`

(c) `A programming library used for managing and manipulating databases.`

(d) `A visualization tool used to represent data in graphical formats.`

(e) `A centralized repository that contains metadata about data, including its structure , definitions, and relationships.`

Question 14. In the code snippet from the data exploration (for the unstructured data) where the function fetchdata is used to retrieve weather data, the following lines appear:
data = fetchdata("https://api.weather.gov/gridpoints/LOX/155,38")
rawtemps = data["properties"]["temperature"]["values"]"
df = pd.read_json(io.StringIO(json.dumps(rawtemps)))
What is the primary purpose of using io.StringIO with json.dumps(rawtemps) before calling pd.read_json?

(a) `It converts the JSON data directly into CSV format before importing it into the DataFrame.`

(b) `It automatically validates the JSON format to ensure there are no syntax errors before reading.`

(c) `It converts the Python data (often a list or dictionary) into a JSON-formatted string and wraps it as a file-like object so that pd.read_json can parse it.`

(d) `It enables real-time streaming of data by mimicking a network file interface for pd .read_json.`

(e) `It compresses the raw JSON data to reduce memory usage when loading it into the DataFrame.`

Question 15. According to the Data Exploration PDF (page 13), which of the following best describes the purpose of covariance in statistical analysis?

(a) `Covariance is the same as correlation.`

(b) `Covariance measures the direction of the relationship between two variables.`

(c) `Covariance always provides a standardized value between -1 and 1.`

(d) `A covariance value of zero always indicates that two variables are independent.`

(e) `Covariance is used to calculate the mean of a dataset.`

Question 16. According to the lecture, what is the primary difference between structured, semi-structured, and unstructured data?

(a) `Semi-structured data lacks any form of organization and cannot be queried efficiently.`

(b) `Structured data cannot be transformed into unstructured data, whereas semi-structured data can be converted into both structured and unstructured formats.`

(c) `Unstructured data always consists of multimedia files, while structured data includes only text and numerical values.`

(d) `Structured data follows a fixed schema, semi-structured data has a flexible schema, and unstructured data has no predefined schema.`

(e) `Structured data can only be stored in relational databases, while semi-structured and unstructured data must be stored in NoSQL databases.`

Question 17. According to the lecture on Data Exploration, Which statement correctly describes the cardinality and arity of this data.

Student ID Name Age Major
101 Peter 24 Mathematics
102 Ram 22 Computer Science
103 James 19 Physics

(a) `The cardinality of the Student table is 12, and the arity is 3`

(b) `The Arity of the Student Table is 4, but cardinality refers to the number of relationships between different tables, as we don't have a second table cardinality cannot be defined.`

(c) `The cardinality and arity refer to the same thing, and the value for this table is 4`

(d) `The cardinality of the Students table is 4, and the arity is 3`

(e) `The cardinality of the Students table is 3, and the arity is 4`

Question 18. Which of the following is true about semi-structured data models like XML?

(a) `They represent data using a set of rigid tables with rows and columns.`

(b) `They rely on predefined schemas that must be applied before data is stored.`

(c) `They do not allow for hierarchical relationships or nesting of elements.`

(d) `They are mostly used for numeric data and require precise data types for every field.`

(e) `They are self-describing documents that can represent tree structures and free-text .`

Question 19. According to lecture on January 29th on the Codio Data Exploration Demo, what is the main purpose of using the "with" function when opening a URL request? Ex:

```
req = urllib.request.Request(url, headers=headers)
with urllib.request.urlopen(req) as response:
    data = response.read().decode('utf-8')
...
```

(a) `It allows multiple URL requests to be opened simultaneously without error.`

(b) `It makes sure that the connection is automatically closed after the code block completes.`

(c) `It continually tries to establish a connection upon failure until success.`

(d) `It is required by the urllib library to be able to open the request.`

(e) `It acts as a while loop for as long as the connection remains until terminated within the code block.`

Question 20. Given table A and table B and a result set which of the options below is the mostly likey operation that was performed.

—Table A
oid cid val
1 101 50
2 102 30
3 101 20
— Table B
cid text
101 Alice
102 Bob
103 Carol
— Result Set
text val
Alice 70
Bob 30

(a) `RIGHT JOIN`

(b) `INNER JOIN followed by GROUP BY`

(c) `LEFT JOIN without aggregation`

(d) `CROSS JOIN`

(e) `FULL OUTER JOIN`

Question 21. According to the data-exploration lecture, which of the following statements best describes the differences between schema-on-write and schema-on-read approaches in data management?

(a) `Schema-on-read is only applicable to structured data formats like relational databases, while schema-on-write is used for semi-structured data like XML.`

(b) `Schema-on-read allows for more flexibility by deferring schema application until data is read, while schema-on-write requires a predefined structure before data can be loaded, resulting in faster read times but less agility.`

(c) `Schema-on-write and schema-on-read are interchangeable terms referring to the same data loading approach in traditional databases.`

(d) `Schema-on-write provides more flexibility than schema-on-read by allowing data to be loaded without a predefined structure.`

(e) `Schema-on-read requires data transformation before loading, while schema-on-write allows data to be copied directly to the file store without transformation.`

Question 22. In Pandas, after running `data = pd.read_csv('file.csv')`, what type of object is data?

(a) `A Python dictionary, which stores key-value pairs.`

(b) `A pandas Series, which is a one-dimensional labeled array.`

(c) `A NumPy array, which is a multi-dimensional homogeneous array.`

(d) `A pandas DataFrame, which is a two-dimensional labeled data structure.`

(e) `A list of lists, where each inner list represents a row in the CSV file.`

Question 23. What do you need to be aware of when imputing missing values using the mean of an attribute?

(a) `Mean imputation works well for categorical attributes.`

(b) `Mean imputation preserves all statistical properties of the data.`

(c) `Mean imputation ensures missing values do not affect correlation calculations.`

(d) `Mean imputation is always preferable to median imputation.`

(e) `Mean imputation can distort the distribution and underestimate variance.`

Question 24. Which of the following best describes a data object in the context of data science?

(a) `A temporary dataset that is used only for intermediate calculations.`

(b) `A single attribute that defines the characteristics of an entity.`

(c) `A unique identifier used to differentiate records in a database.`

(d) `A special type of binary data structure used for machine learning models.`

(e) `A representation of an entity that consists of multiple attributes.`

Question 25. According to the lecture on data exploration, which of the following types of data is considered semi-structured?

(a) `Video files`

(b) `Excel spreadsheets`

(c) `SQL application database`

(d) `Server log files`

(e) `Customer service phone call recordings`