cap5771sp25
February 26, 2025

Exam 1 part 1 version e

Question 1. 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 enables real-time streaming of data by mimicking a network file interface for pd.read_json.`

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

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

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

Question 2. 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 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.`

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

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

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

Question 3. According to the lecture on attribute types in Data Exploration, which of the following best describes an ordinal attribute?

(a) A numeric attribute with real values that can be measured continuously, such as temperature.

(b) A categorical attribute with distinct labels but no meaningful order, such as hair color.

(c) A binary attribute with only two states, such as positive and negative test results .

(d) A special type of numeric attribute that always follows a normal distribution.

(e) A categorical attribute with a meaningful order but without a known magnitude between successive values, such as rankings.

Question 4. 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) Structured data can only be stored in relational databases, while semi-structured and unstructured data must be stored in NoSQL databases.

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

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

Question 5. According to the Data Exploration lecture and demo, which of the following best describes the internal structure of a pandas DataFrame as used in Python?

(a) It is a list of Excel sheet objects where each column has its own JSON file with a tiny SQLite database running secret queries.

(b) It is implemented as a JSON object, where each key is a column name and each value is a list containing the column's data.

(c) It is a data field, representing a characteristic or feature of a data object.

(d) It is a dictionary where each key is a column name and each value is a Series object representing that column.

(e) `It is a one-dimensional labeled array capable of holding any data type (integers, strings, floating point numbers).`

Question 6. According to the lecture on data exploration(page 25), which of the following is an example of an asymmetric binary attribute?

(a) `A temperature measurement in Celsius.`

(b) `A hair color attribute with values like black, brown, and blonde.`

(c) `A gender attribute with values male and female.`

(d) `A medical test result where positive is more significant than negative.`

(e) `A zip code attribute representing different geographic regions.`

Question 7. According to the lecture on data exploration, which one of the following is not apart of "The 5 Vs of Big Data"?

(a) `Visualization`

(b) `Volume`

(c) `Veracity`

(d) `Velocity`

(e) `Variety`

Question 8. According to slide 10 of the Intro to Data Science lecture, which of the following best describes the difference between Machine Learning and Data Science?

(a) `Machine learning is the study of building neural networks to generate new data, such as images and text. Data science uses simpler models to analyze existing data and learn from it.`

(b) `Machine learning is about working with structured data, such as from databases and spreadsheets. Data science often deals with that data, but can deal with unstructured data as well, such as images and videos.`

(c) `Data science encompasses the process of gathering and cleaning data, while machine learning only focuses on the specific algorithms used to analyze that data.`

(d) `Machine learning deals with the creation of new models to predict based on training data. Data science deals with exploring data and finding trends, in part by using those models.`

(e) `Machine learning is purely theoretical, while data science is purely practical.`

Question 9. According to the lecture on data exploration, how do traditional databases stores load data into a table?

(a) `Schemamore.`

(b) `Schemaless.`

(c) `Schema-On-Read.`

(d) `Schema-On-Write.`

(e) `noSQL.`

Question 10. Scenario: A sports analytics company is analyzing player performance statistics from multiple basketball games. Each game is recorded with numerical values such as points scored, assists, and rebounds for each player. According to the lecture on data exploration from slide 27, which type of numeric attribute best describes the number of points a player scores in a game?

(a) `Discrete numeric attribute`

(b) `Continuous numeric attribute`

(c) `Ordinal attribute`

(d) `Binary attribute`

(e) `Nominal attribute`

Question 11. 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 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.`

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

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

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

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

Question 12. According to the lecture and materials, During the data exploration phase, which of the following methods is suitable for identifying patterns of missing values in a dataset?

(a) `Use a classification model to predict missing values.`

(b) `Calculate the mean and median for each variable.`

(c) `Use a scatter plot to visualize the relationship between two variables.`

(d) `Use a missing value matrix or heatmap to display the locations of missing values.`

(e) `Perform Principal Component Analysis (PCA) to reduce dimensionality.`

Question 13. According to the lecture on probability distributions, where do most data points lie in a normal distribution?

(a) `The data points are evenly distributed across all values, with no concentration near the mean.`

(b) `Most data points align exactly with the standard deviation, making` σ `the most frequent value.`

(c) `Most data points are found in the extreme tails, far from the mean.`

(d) `Most data points cluster around the mean, with fewer values appearing as you move further away.`

(e) `Most data points lie outside the interquartile range (IQR), making the middle range of data less significant.`

Question 14. 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 makes sure that the connection is automatically closed after the code block completes.`

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

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

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

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

Question 15. According to data exploration, which is an advantage of using a quantile-quantile (Q-Q) plot in data analysis?

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

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

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

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

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

Question 16. 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 retrieves only one row from the SQL output`

(c) `It will crash your whole system`

(d) `It forms a new SQL query`

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

Question 17. According to the lectures on data exploration, which of the following does NOT describe the attribute of height, as measured in whole inches?

(a) `Ratio-scaled.`

(b) `Discrete.`

(c) `Quantitative.`

(d) `Ordinal.`

(e) `Numeric.`

Question 18. Which of the following best describes a data object in the context of data science?

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

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

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

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

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

Question 19. According to lecture material on database design paradigms, which approach requires defining data structure constraints before storing information in a database?

(a) `Puppy-Driven Data Modeling`

(b) `Semi-structured Data Model`

(c) `Schema-on-Write`

(d) `Relational Model`

(e) `Schema-on-Read`

Question 20. According to the data exploration demo, assume we are given a table called movies with columns (title, director, genre, box office record), which SQL query tells us the genres that have more than 100 movies that have achieved a box office record of $500,000 or more?

(a)
```
SELECT genre, count(*) FROM movies WHERE boxofficerecord >= 500000
```

(b)
```
SELECT genre, count(*) FROM movies WHERE boxofficerecord >= 500000 GROUP BY genre
  HAVING count(*) > 100
```

(c)
```
SELECT title, count (*) FROM movies WHERE boxofficerecord >= 500000 GROUP BY title
  HAVING count(*) > 100
```

(d)
```
SELECT genre, count(*) FROM movies GROUP BY genre HAVING count(*) > 100
```

(e)
```
SELECT genre, count(*) FROM movies GROUP BY genre HAVING count(*) > 100 WHERE
  boxofficerecord >= 500000
```

Question 21. In the context of data visualization, which of the following best describes the key difference between a histogram and a bar chart?

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

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

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

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

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

Question 22. A dataset has a highly right-skewed (positively skewed) distribution. How will the mean and median compare in this case?

(a) `The median will be more affected by extreme values than the mean.`

(b) `The mean will be greater than the median.`

(c) `The mean and median will be equal.`

(d) `The mean will be less than the median.`

(e) `The mean is always a more reliable measure of central tendency in skewed data.`

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

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

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

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

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

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

Question 24. 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) `Instance`

(b) `Database`

(c) `Attribute`

(d) `Schema`

(e) `Column`

Question 25. According to the data exploration lecture, what is the primary reason for detecting and handling missing data during the data preparation phase?

(a) `To reduce the file size of the dataset for faster processing.`

(b) `To make the dataset look more professional in presentations.`

(c) `To automatically generate new data to replace the missing values without validation .`

(d) `To improve data visualization aesthetics by eliminating blank spaces.`

(e) `To ensure the accuracy and reliability of statistical analyses and machine learning models.`