cap5771sp25
February 26, 2025

Exam 1 part 1 version a

Question 1. 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) `Attribute`

(b) `Column`

(c) `Schema`

(d) `Database`

(e) `Instance`

Question 2. 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 converts the JSON data directly into CSV format before importing it into the DataFrame.`

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

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

(e) `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.`

Question 3. What do you need to be aware of when imputing missing values using the mean of an attribute?

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

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

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

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

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

Question 4. According to the lecture on Data Exploration, which type of attribute is represented by the following scenario? In a race, five athletes finish at different times. You rank them from 1st place to 5th place based on who finishes first to last. What type of attribute is used when ranking athletes from 1st to 5th?

(a) `Ratio attribute.`

(b) `Interval attribute.`

(c) `Nominal attribute.`

(d) `Binary attribute.`

(e) `Ordinal attribute.`

Question 5. According to the lecture, what is the primary difference between structured, semi-structured, and unstructured data?

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

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

Question 6. According to the lecture on attribute types in Data Exploration, which of the following best describes an ordinal attribute?

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

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

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

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

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

Question 7.

```
with open(‘sample.json’,‘r’) as f:
    data = json.load(f)
```

What is the most likely data type of data after executing this code?

(a) `An instance of a user-defined class dynamically inferred from the JSON schema.`

(b) `A hierarchical tree structure where each node is a JSON key-value pair.`

(c) `A Pandas DataFrame if the JSON contains tabular data.`

(d) `A Python dictionary or a list, depending on the JSON structure.`

(e) `A serialized JSON object stored as a Python string.`

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(...) returns a database connection object.`

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

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

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

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

Question 9. 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 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.

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

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

(d) 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.

(e) 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.

Question 10. According to the lecture on Data Exploration, what is the trouble with relying solely on summary statistics when analyzing a dataset?

(a) Summary statistics always give an incomplete picture of the data.

(b) Please don't pick me

(c) Summary statistics are only useful for small datasets.

(d) Sometimes two datasets can have identical summary statistics but very different distributions.

(e) Summary statistics are only valid if the dataset follows a normal distribution.

Question 11. 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 represents categorical data, whereas a bar chart represents numerical data.

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

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

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

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

Question 12. In a dataset with both continuous and categorical variables, which visualization method is most suitable for examining the relationship between a continuous variable and a categorical variable with more than two categories?

(a) Box plot for the continuous variable and color-coded by the categorical variable.

(b) Line chart for each category of the categorical variable.

(c) Scatter plot with different colors for each category of the categorical variable.

(d) Heatmap of correlations between all variables.

(e) Violin plot for the continuous variable and split by the categorical variable.

Question 13. 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) Ratio-scaled attributes cannot be negative, while interval-scaled attributes must always be positive.

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

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

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

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

Question 14. 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 histogram is less affected by outliers than a box plot.

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

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

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

Question 15. In Pandas, after running `data = pd.read_csv('file.csv')`, what type of object is data?

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

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

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

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

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

Question 16. 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-write provides more flexibility than schema-on-read by allowing data to be loaded without a predefined structure.

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

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

(e) 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.

Question 17. According to the lecture on data exploration, why might modern data-intensive organizations prefer a schema-on-read approach instead of schema-on-write when managing rapidly evolving datasets?

(a) Schema-on-read accelerates query performance by preloading metadata definitions.

(b) Schema-on-read enables data ingestion without transformation, supporting agile decision-making when new data types emerge.

(c) Schema-on-read enforces strict governance during data collection, ensuring higher data quality.

(d) Schema-on-write simplifies the integration of social network data with relational systems by delaying schema creation.

(e) Schema-on-read allows only structured data to be stored efficiently.

Question 18. According to data visualization best practices, which of the following would NOT be a good reason to use a box plot?

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

(b) To show and compare distribution values.

(c) To identify and display outliers in a dataset.

(d) To show correlations between two variables.

Question 19. According to the lecture 'Data Exploration', what is the purpose of using the pandas library in the context of the weather data analysis project?

(a) To load, manipulate, and analyze weather data, including converting units and calculating statistics like mean and standard deviation.

(b) To fetch weather data, such as temperature, for a specific location based on latitude and longitude.

(c) To store and query weather data using SQLite database for complex data analysis.

(d) To get city information, including latitude and longitude, from an open-source geocoding website.

(e) To visualize data using matplotlib, such as scatter plots and histograms.

Question 20. 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 a data field, representing a characteristic or feature of a data object.

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

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

(e) 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.

Question 21. 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 missing value matrix or heatmap to display the locations of missing values.

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

(c) Calculate the mean and median for each variable.

(d) Perform Principal Component Analysis (PCA) to reduce dimensionality.

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

Question 22. According to the lecture on probability distributions, where do most data points lie in a normal distribution?

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

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

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

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

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

Question 23. 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 acts as a while loop for as long as the connection remains until terminated within the code block.

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

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

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

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

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

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

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

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

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

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

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 automatically generate new data to replace the missing values without validation .`

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

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

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

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