EDA Dataset content Interview Questions

(Practice Project)

PW SKILLS

Easy Questions

Question 1:
What is the purpose of loading a dataset and getting an overview of its structure?

Answer:
The purpose of loading a dataset and getting an overview of its structure is to understand the basic characteristics of the data, including the number of rows and columns, the data types of each column, and the general distribution of the data.

Question 2:
How can you identify numerical and categorical columns in a dataset using Python?

Answer:
One can identify numerical columns by using data.select_dtypes(include=['int64', 'float64']).columns and categorical columns by using data.select_dtypes(include=['object']).columns. This separation is crucial for determining which analyses and visualizations are suitable for each type of data.

Question 3:
What is a data dictionary, and why is it important?

Answer:
A data dictionary is a table or document that describes the structure, content, and format of the data in a dataset. It typically includes information like column names, data types, and a brief description of each column. The data dictionary is important because it helps users quickly understand the dataset, ensuring that they use the data correctly and efficiently.

Question 4:
Why is it important to check for missing values in a dataset?

Answer:
Checking for missing values is important because missing data can skew the results of your analysis, lead to biased conclusions, or cause errors in machine learning models. Identifying and addressing missing values ensures the integrity and accuracy of the data.

Question 5:
What is the difference between mean and median?

Answer:
The mean is the average of all values in a dataset, while the median is the middle value when the data is sorted. The median is less sensitive to outliers than the mean, making it a better measure of central tendency in skewed distributions.

Medium Questions

Question 6:
How can you check for duplicates in a dataset, and why should they be removed?

Answer:
One can check for duplicates in a dataset using the data.duplicated().sum() function in Python. Duplicates should be removed because they can skew the analysis by over-representing certain observations, leading to inaccurate conclusions.

Question 7:
Explain the significance of the standard deviation in a dataset.

Answer:
The standard deviation measures the amount of variation or dispersion in a dataset. A low standard deviation indicates that the data points are close to the mean, while a high standard deviation indicates that the data points are spread out over a wider range. It helps in understanding the consistency of the data and is essential in comparing the spread between different datasets.

Question 8:
What does skewness tell you about the distribution of data?

Answer:
Skewness indicates the asymmetry of a data distribution. A positive skewness suggests that the data distribution tails off to the right, with more data points on the left, while a negative skewness indicates a left-tailed distribution. Skewness helps in understanding the shape of the data distribution and whether it deviates from a normal distribution.

Question 9:
Describe the purpose of a box plot in exploratory data analysis.

Answer:
A box plot is used to visualize the distribution of data based on five summary statistics: minimum, first quartile, median, third quartile, and maximum. It helps in identifying outliers, understanding the spread, and comparing distributions across different categories. Box plots are particularly useful for spotting deviations from normality.

Question 10:
How can a heatmap be useful in data analysis?

Answer:
A heatmap is useful for visualizing the correlation matrix of a dataset, showing the strength and direction of relationships between variables. It helps identify highly correlated variables, which is valuable for feature selection in machine learning models or understanding potential multicollinearity issues.

Hard Questions

Question 11:
What steps would you take to address inconsistencies in categorical data?

Answer:

To address inconsistencies in categorical data, I would:

1. Identify unique values in the categorical columns using functions like data[col].unique() to spot any variations in spelling, case, or format.
2. Standardize the values by correcting any inconsistencies (e.g., "Male" vs. "male" should be made consistent).
3. If necessary, map similar categories to a single standard category.
4. Document the changes to ensure transparency and reproducibility in the analysis.

Question 12:
Why is it important to check for outliers, and how can they affect your analysis?

Answer:
Outliers can significantly impact statistical analyses and model performance. They can skew mean values, increase variance, and distort the results of correlation and regression analyses. Identifying outliers using techniques like box plots or z-scores is important so you can decide whether to remove, transform, or investigate them further based on the context of the data.

Question 13:
Explain how variance and standard deviation are related, and why both are important.

Answer:
Variance measures the average squared deviation of each data point from the mean, while standard deviation is the square root of the variance, bringing it back to the original units of the data. Both metrics are important because variance provides a raw measure of spread, while standard deviation gives a more interpretable measure that is easier to compare across datasets with different scales.

Question 14:
Describe the impact of kurtosis on data distribution analysis.

Answer:
Kurtosis measures the "tailedness" of a data distribution. High kurtosis indicates heavy tails, meaning there are more outliers than in a normal distribution, while low kurtosis indicates light tails, with fewer outliers. Understanding kurtosis helps assess the likelihood of extreme values and can influence the choice of statistical tests or models, especially those sensitive to outliers.

Question 15:
How would you interpret a right-skewed distribution in the context of the Titanic dataset's fare column?

Answer:
A right-skewed distribution in the Titanic dataset's fare column indicates that most passengers paid lower fares, while a smaller number paid significantly higher fares. This skewness suggests that while affordable ticket prices were common, there were some outliers where passengers paid much more, possibly for higher-class accommodations or luxury services.

Question 16:
When does multicollinearity in a dataset occur, and why is it problematic?

Answer:
Multicollinearity occurs when two or more independent variables are highly correlated, leading to redundancy in the model. This can inflate the variance of the coefficient estimates, making them unstable and difficult to interpret.

Question 17:
What are some potential issues if you find a high correlation between "Age" and "Fare" in the Titanic dataset?

Answer:
If there's a high correlation between "Age" and "Fare," it could indicate that older passengers generally paid more for their tickets, which might be due to factors like travel class or family size. This could lead to issues in multivariate analysis or predictive modeling, where multicollinearity could affect the stability and interpretability of the model. It's important to further investigate the cause of this correlation and decide if any adjustments are needed.

Question 18:
How would you deal with a situation where your dataset has both categorical and continuous variables that need to be visualized together?

Answer:

To visualize both categorical and continuous variables together, we can use:

1. **Box plots:** To show the distribution of continuous variables across different categories.
2. **Violin plots:** To combine a box plot with a density plot, providing more insight into the distribution.
3. **Scatter plots with color coding:** To represent continuous variables, while categories can be indicated using different colors or shapes.
4. **Facet grids:** To create separate plots for different categories. These methods help in effectively communicating the relationship between different types of data.