Q&A

1. What is data analysis?
    - Data analysis is the process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making.
2. What is data preprocessing? Why is it important?
    - Data preprocessing is the process of preparing raw data for analysis. It's important because real-world data is often incomplete, inconsistent, and noisy, which can hinder analysis.
3. What are some common data preprocessing steps?
    - Common steps include handling missing values, data formatting, data normalization, and converting categorical variables to numerical ones.
4. What Python library is commonly used for data manipulation and analysis?
    - Pandas.
5. What is a Pandas DataFrame?
    - A DataFrame is a two-dimensional labeled data structure, similar to a table, with rows and columns.

Dataset Specific (Titanic)

6. What is the source of the Titanic dataset?
    - Kaggle (https://www.kaggle.com/c/titanic/data).
7. What type of data does the Titanic dataset contain?
    - It contains information about passengers aboard the Titanic, including survival status, passenger class, age, and other attributes.
8. What are some key variables in the Titanic dataset?
    - Key variables include PassengerId, Survived, Pclass, Name, Sex, Age, Cabin, and Embarked.
9. Which columns in the Titanic dataset had missing values?
    - Age and Cabin.

Pandas Fundamentals

10. How do you load a CSV file into a Pandas DataFrame?
    - Using the pd.read_csv() function.
11. How do you display the first few rows of a DataFrame?
    - Using the df.head() method.
12. How do you check the last few rows of a DataFrame?
    - Using the df.tail() method.
13. How do you get the number of rows and columns in a DataFrame?
    - Using the df.shape attribute.

14. How do you check the data types of the columns in a DataFrame?
    - Using the df.dtypes attribute.
15. How do you get descriptive statistics of numerical columns in a DataFrame?
    - Using the df.describe() method.
16. How do you check for missing values in a DataFrame?
    - Using the df.isnull() method, often combined with .sum() to count missing values per column.

Missing Value Handling

17. What is missing data?
    - Missing data occurs when no data value is stored for a particular attribute in an observation.
18. Why is it important to handle missing values?
    - Missing values can lead to biased or inaccurate results in data analysis and machine learning.
19. What are some common techniques for handling missing values?
    - Imputation (replacing with estimated values) and deletion (removing rows or columns with missing values).
20. How was the missing 'Age' data handled in this exercise?
    - Missing 'Age' values were imputed based on the mean age of passengers in the same 'Pclass'.
21. Why was the 'Cabin' column dropped?
    - It had a large number of missing values, making meaningful imputation difficult.

Data Formatting and Normalization

22. What is data formatting?
    - Ensuring that data is stored in the correct data type (e.g., integer, float, string).
23. What is data normalization?
    - Scaling numerical data to a standard range. (Note: The provided code didn't *perform* normalization, but the concept is in the writeup)

Categorical Data

24. What is a categorical variable?
    - A variable that represents categories or groups (e.g., Sex, Embarked).
25. How are categorical variables typically converted to numerical variables for analysis?
    - Using techniques like one-hot encoding or label encoding. (Note: The

provided code didn't *perform* this, but it's a common next step)