Part 3: Preparing for the Interview Questions

1. What are missing values and how do you handle them?

- **Answer:** Missing values are data points that are absent or null in a dataset. They can be handled by:
    - **Deletion:** Removing rows or columns with missing values (e.g., using dropna()), though this can lead to loss of information.
    - **Imputation:** Filling the missing values with a substitute value, such as the mean, median, or mode of the column (e.g., using fillna()).

2. How do you treat duplicate records?

- **Answer:** Duplicate records are identical rows in a dataset. They are typically removed to avoid biasing the analysis. In Pandas, this is done using the

  .drop_duplicates() method.

3. Difference between dropna() and fillna() in Pandas?

- **Answer:** dropna() is used to *remove* rows or columns containing missing values (NaNs). fillna() is used to *replace* missing values with a specified value (like the mean, median, or a constant).

4. What is outlier treatment and why is it important?

- **Answer:** Outlier treatment is the process of identifying and managing data points that are significantly different from the rest of the data. It's important because outliers can skew statistical analyses and machine learning models, leading to inaccurate results.

5. Explain the process of standardizing data.

- **Answer:** Data standardization is the process of transforming data into a common format to ensure consistency. This includes standardizing text values (e.g., converting 'USA' and 'United States' to a single format), ensuring consistent date formats (e.g., 'dd-mm-yyyy'),

and making column headers uniform.

6. How do you handle inconsistent data formats (e.g., date/time)?

- **Answer:** Inconsistent formats are handled by converting the data into a single, consistent type. For dates and times, this usually means converting string representations (like '01/01/2022' or 'Jan 1, 2022') into a standard datetime object.

7. What are common data cleaning challenges?

- **Answer:** Common challenges include structural errors like typos and inconsistent capitalization, handling missing data in a way that doesn't bias the results, dealing with outliers, and resolving conflicting or duplicate data entries across different sources.

8. How can you check data quality?

- **Answer:** Data quality can be checked by profiling the data. This involves using descriptive statistics, checking for null values (`.isnull()`), counting unique values in columns to spot inconsistencies, and creating visualizations like histograms and box plots to identify outliers and understand data distributions.