Week 10

Name: Siddharth Dudugu
(siddharthdudugu@gmail.com)

Country: USA

College: Drexel University

Specialisation: Data Science

EDA

1. **Visual Inspection of Data:**
    - **Objective:** Understand the structure and content of the dataset.
    - **Code:**

pythonCopy code

df.head() df.info()

    - **Explanation:** Displays the first few rows of the dataset and provides information about column data types, non-null counts, and memory usage.

2. **Histogram and Box Plot for 'Sales':**
    - **Objective:** Explore the distribution and identify potential outliers in the 'Sales' column.
    - **Code:**

pythonCopy code

df['Sales'].plot(kind='hist') df['Sales'].plot(kind='box')

    - **Explanation:** Visualizes the distribution of 'Sales' using a histogram and identifies potential outliers through a box plot.

3. **Correlation Matrix Visualization:**
    - **Objective:** Understand the relationships between numerical variables.
    - **Code:**

pythonCopy code

correlation_stat = df.drop(['date', 'hour', 'dayofweek'], axis=1) sns.heatmap(correlation_stat.corr(), cmap='crest')

    - **Explanation:** Generates a correlation matrix and visualizes it as a heatmap, allowing for the identification of correlations between different features.

4. **Boolean Column Analysis:**
    - **Objective:** Assess the distribution and frequency of boolean columns.
    - **Code:**

pythonCopy code

df.groupby('Product')['date'].nunique()

    - **Explanation:** Grouping by 'Product' and analyzing the number of unique dates for each product, specifically exploring boolean columns ('V_DAY,' 'EASTER,' 'CHRISTMAS').

5. **Product Code Analysis:**

- **Objective:** Examine the distribution of unique products and their occurrence.
- **Code:**

pythonCopy code

df['Product'].nunique() df['Product'].unique()

- **Explanation:** Determines the number of unique products and lists their unique codes, providing an overview of the product diversity.

6. **Correlation Analysis and Removal of Features:**
    - **Objective:** Identify and potentially remove features with low correlation.
    - **Code:**

pythonCopy code

dfe = dfe.drop(['V_DAY', 'EASTER', 'CHRISTMAS'], axis=1)

- **Explanation:** Drops boolean columns ('V_DAY,' 'EASTER,' 'CHRISTMAS') due to their limited occurrence and potential low correlation.