Initial Loading + Exploration

Pd.read_csv() - Loads the dataset into a dataframe
Input: URL
Output: Data Frame

.shape - Returns the number of rows and columns
Input: Data Frame
Output: Tuple of (rows, columns)

.dtypes -Returns the data types of each column
Input: Data Frame
Output: All datatypes in each column

.head() - Returns the first 5 rows
Output: Returns the first 5 Rows

.tail() - Note: You might know what this does…
Output?

.describe() - Returns Count, mean, std, min, max, quartiles
Input: Data Frame
Output: Returns the Count, mean, standard deviation, min, max, and quartiles

Data Cleaning + Basic Descriptive Analysis

.isnull() - detects missing values
Output: Boolean Table with true for missing values and false for non-missing

.sum() - counts missing values
Output: Total count of values

.dropna() - remove missing values
Output: Dataframe without nulls

.duplicated() - identify duplicate rows
Output: Boolean Values where true means a duplicate row

.drop_duplicates() -removes duplicates

.max()- Returns Maximum

.min() - Returns Minimum
.mean() - Returns the Average
.sum() - Returns the sum of all values in a column

Data Aggregation + Frequency Analysis

.value_counts() - count unique values within a column
Input: Column of Data Frame
Output: List of values + Their counts

.groupby() - aggregate data by categories
Input: Dataframe Columns for grouping
Output: Aggregated values for each category

Data Visualization

plt.figure (figsize=(x,y)) - create figure and set size
plt .title() - add plot title
plt.xlabel() - label x-axis
plt.ylabel() - label y-axis
plt.tight_layout() - adjust spacing
plt.show() - display plot
sns.barpot() - create bar plot
    x=_____ -sets the numeric values (counts)
    y=_____ -sets the category labels
    Palette = _____ -sets the color theme
sns.histplot()
    df[‘column_name]
    bins = x - number of intervals
    Color = _____ - bar color
    Kde = T/F - adds a smooth density curve