CAPSTONE PROJECT

1. **The project consists of two sections.**
2. **Section-1 consists of Netflix Dataset whereas section-2 consists of Zomato Dataset. Both the datasets will be provided.**
3. **Hits given are only for reference purpose; you can create your own graph.**
4. **Follow the below questions to perform EDA on both datasets & follow the reference document provided to submit the project.**

SECTION-1

Netflix Dataset

Exploratory Data Analysis (EDA)

Prerequisites

You should have Python installed along with the necessary libraries: pandas, matplotlib, and seaborn.

If not, you can install the necessary libraries

pip install pandas matplotlib seaborn

Following are the steps to be followed to perform EDA.

Q1: How can you load a Netflix dataset into a Pandas Data Frame, and what method will you use to get basic information about the dataset?

Load the Dataset

This section initiates the EDA process by importing necessary Python libraries, such as pandas for data manipulation and analysis.

This step is crucial as it sets the foundation for subsequent analysis.

Hint: Utilize the **pd.read_csv()** function to load the dataset into a Pandas Data Frame.

Basic Information about the Dataset

Following dataset loading, this section prints basic information about the structure of the dataset. It displays column data types and non-null counts, providing an overview of the dataset's composition. Understanding the data types is essential for subsequent data cleaning and analysis.

Hint: Use the info() method to get basic information.

Q2: What Pandas method helps you identify missing values in the dataset, and how can you count the number of missing values for each column?

Handling Missing Values

Identifying and handling missing values is crucial for ensuring data quality. This section examines the dataset for any missing values and implements strategies to handle them appropriately, ensuring the integrity of the analysis.

Hint: Employ the isnull() method to create a Data Frame of Boolean values indicating missing values. Sum the Boolean values to get the count of missing values for each column.

Q3: To get summary statistics for numerical columns, what Pandas method should you use, and what insights can you gather from these statistics?

Summary Statistics of Numerical Columns:
To gain insights into the numerical aspects of the dataset, this section generates summary statistics. These statistics include measures such as mean, standard deviation, minimum, maximum, and quartiles for numerical columns. This information aids in understanding the central tendencies and variability within the dataset.
Hint: Use the describe() method on the Data Frame to obtain summary statistics for numerical columns.

Q4: Identify categorical columns using a specific Pandas function, and explain the purpose of the loop iterating through these columns.
Explore Categorical Variables:
Exploration of categorical variables is vital for understanding the diversity of data. This section examines the unique values within categorical columns, providing insights into the categorical distribution and supporting subsequent analysis involving these variables.
Hint: Identify categorical columns using the select_dtypes() method. Iterate through these columns and print the unique values for each.

Q5: Create any plot to visualize the distribution of your choice . Which Seaborn function will you use, and what information can you obtain from the plot?

Visualize Distribution of Target Variable:
Visualizing the distribution of the target variable in Netflix dataset, is essential for understanding the class distribution. This section employs visualizations, like bar plots or pie charts, to represent the frequency distribution of the target variable.
Hint: Example: Create a count plot using Seaborn's countplot() function. Specify the appropriate column for the x-axis.(you can create a plot of your choice).

Q6: Visualize the distribution of the 'Runtime' column using a Seaborn plot with specific parameters. What insights can you gather from this visualization?

Visualize Distribution of Numerical Column
Visualizing the distribution of a numerical column, like 'Duration,' helps in understanding its spread. This section creates visualizations, such as histograms or kernel density plots, to illustrate the distribution characteristics of the chosen numerical column.
Hint: Use Seaborn's histplot() function to create a histogram of the 'Runtime' column. Set appropriate parameters like 'bins' and 'kde'

Q7: Create any plot of choice to explore the relationship between two numerical columns. Which Seaborn function should you use, and what does the scatter plot reveal?

Explore Relationships Between Numerical Columns
Understanding relationships between numerical columns is valuable for uncovering patterns or correlations. This section employs visualizations, such as scatter plots or heatmaps, to explore relationships between numerical variables, facilitating deeper insights into the dataset.
Hint:Example: Generate a scatter plot using Seaborn's scatterplot() function. Specify 'Rating' for the x-axis, 'Runtime' for the y-axis.

Q8: Introduce a different type of visualization to gain more insights. What Seaborn or Matplotlib function allows you to create this visualization?

Hint: Try using a different Seaborn or Matplotlib function, like a bar plot or box plot, to explore data from a different perspective.

Q9: How can you display the generated plots? Which Matplotlib function will you use to showcase these visualizations?

Hint: Use plt.show() after creating each plot to display them.

Q10: Based on the visualizations, what conclusions or recommendations can you make about the Netflix dataset? Consider patterns, trends, or interesting observations.
Hint: Look for patterns or interesting observations in the visualizations. Consider how different variables relate to each other and what insights can be drawn.

SECTION-2

ZOMATO DATASET

Exploratory Data Analysis (EDA)

Q1: How do you load the Zomato dataset into a Pandas Data Frame, and which Pandas method provides basic information about the dataset?

Hints: Use **pd.read_csv()** to load the dataset and **info()** to display basic information.

Q2: Identify missing values in the Zomato dataset using a Pandas method. How can you count the number of missing values for each column?

Hints: Utilize **isnull()** to create a Data Frame of Boolean values and then use **sum()** to count missing values.

Q3: What Pandas method provides summary statistics for numerical columns in the Zomato dataset, and what insights can you derive from these statistics?

Hints: Apply **describe()** to obtain summary statistics; look for patterns, central tendencies, and potential outliers.

Q4: How do you identify categorical columns in the Zomato dataset, and why is it useful to iterate through these columns?

Hints: Use **select_dtypes()** to find categorical columns; iterate to understand unique values for insights.

Q5: Create a Seaborn plot to visualize the distribution of cuisine types. Which Seaborn function is suitable, and what insights can be gained?

Hints: Use **countplot()** to visualize the distribution of categorical cuisine types; observe popular cuisines.

Q6: Visualize the distribution of restaurant ratings using a Seaborn plot. What insights can you gather from this visualization?

Hints: Use **histplot()** to display the distribution of numerical ratings; observe the spread and frequency.

Q7: Create any plot of choice to explore the relationship between two numerical columns. Which Seaborn function is appropriate, and what does the scatter plot reveal?

Hints: Use **scatterplot()** to visualize the relationship; observe patterns or correlations between two numerical variables.

Q8: Introduce a different type of visualization. What Seaborn or Matplotlib function can be used?

Hints: Consider using a bar plot **barplot()** to explore a different perspective of the data.

Q9: How can you display the generated plots for the Zomato dataset, and which Matplotlib function should be used?

Hints: Utilize **plt.show()** after creating each plot to display them.

Q10: Based on the visualizations of the Zomato dataset, what conclusions or recommendations can be made? Look for patterns, trends, or interesting observations.