Report of AI Programming – Project

Instructor: شروق المحمدي

مجموعة ٥:

شعبة F5

زينب حسين النخلي -٤٤٥٣٧٥٣ , لما فايز البلادي - ٤٤٥٣٤٢٣, شهد الحميدي الرشيدي - ٤٤٥٣٣٠١

شعبة F7

روز فهد الأحمدي- ٤٤٥٣٢٢٧, ريمان محمد الصاعدي- ٤٤٥٣٠٤٨ ,شذى علي الحربي - ٤٤٥٣٠٨٦

Our data set focuses on Disney movies by providing descriptive data about each movie such as the movie title, released date, MPAA rating, and the movie's corresponding genre, as well as numeric data as the total gross and the inflation–adjusted gross.

We have been working on the dataset through three key stages:

a. **Data Processing**

b. **Data Analysis**

c. **Data Visualization**

d. **Which Disney princess are you? (additional)**

It is a quiz-like survey that asks the user questions and based on the user's answer it will output the princess picture (Using the PIL library)which is the princess that will be more likely to share common interests and traits with the user and hence will help the user to know what type of Disney movies align with his interest best.

Data Processing :

process step	Description	student performed it
Filling or removing missing values	We used df.isnull().sum().sum() function to make sure if there is any missing values.after that, we found missing values labeled as 'Not Rated' or 'Unknown,' so we adjusted them to make the data more informative.	زينب النخلي.
Tokenizing text data	Split the phrase 'movie title' into single words.	روز الأحمدي.
Normalizing or scaling data	Normalizing using minimax to the date released .	ريمان الصاعدي.
Validating data	We validate data by checking for duplicated values in the data set using duplicated().sum().sum(), and then checked for date formats using the valid_date(date) method, after that we make sure that all numeric values are of the correct data type.	شهد الرشيدي.
Removing duplicate records.	Using df. duplicated(), to check if there is any duplicated values remove it using drop() and just take the first match.	شذا الحربي.
Reformatting data	Rewrite the movie title and genre to lowercase, and format the release date as (year–month–day).	لما البلادي

Data Analysis :

process step	Description	student performed it
Finding central values (mean, median, mode).	Finding central values using the functions from pandas library for the numerical values (Total Gross, Inflation Adjusted Gross), and the mode for the others(Date Released, Movie Title, MPAA rating, Genre)	لما البلادي.
Calculating the correlation between variables.	For the correlation we used the built-in corr() method to find the correlation between direct numeric values in our data set the total gross and the inflation-adjusted gross, and derived undirected relationships from the data set since we only have two numeric columns to work with, the first relationship is between genre means to the total gross and month released also to the total gross which checks how does the genre means and the released month influence the total gross.	شهد الرشيدي.
Finding common words in text data (e.g., word frequency analysis).	Based on the results of both tokenizing and reformatting, We found common words using methods counter() and most common()	ريمان الصاعدي.
Grouping and aggregating data	We take the Genre values and calculate the max, min, count , mean, and sum for just the numerical values using .agg() and display it using .describe()	روز الأحمدي.
Finding the movies that have remake storylines.	Using df['Movie Title'].value_counts() to count the number of remakes for each movie, then process it using loops, and list, then return the movie edition with the highest total gross	شذا الحربي.
finding the most profitable genre	calculate the total gross for each genre using dictionary then return the top three genres according to its total gross	زينب النخلي.

Data Visualization :

process step	Description	student performed it
Histograms.	1- We created a histogram to compare Total Gross and Inflation Adjusted Gross , adds labels, a legend, and a title, then displays the plot . 2- We used histograms to show MPAA Rating and frequency of MPAA Rating, with specific labels, a legend, and a title.	١. روز الأحمدي. ٢. لما البلادي
scatter plots.	We created a dictionary with genres as keys and the number of movies in each genre as values. Then, we represented the keys on the X-axis and the values on the Y-axis using a scatter plot.	ريمان الصاعدي.
Heatmaps to show correlations.	We used a heatmap to visualize the correlation relationship between genre means and released month and inflation-adjusted gross to the total gross	شهد الرشيدي.
Word clouds for text data visualization.	We used a word cloud for 'Movie Titles' by splitting it into single words, after that generated a word cloud with specific colors, width, and height using the WordCloud library	شذا الحربي.
Bar charts.	We use it to illustrate the total gross in billions for each genre along the years histograms	زينب النخلي.