USC ID: 5360857307

DSCI 510 HW5 PROJECT DESCRIPTION

DESCRIPTION

For the final project, I have created a dataset in the form of a csv file called 'Movie_final.csv'. This dataset is created using three different data sources listed below in the Data Sources Section. Following files: 'IMDB_Movies', 'Apidata_movies', Dataset_ movie are intermediate csv files that have led to the creation of the final dataset. The first file 'IMDB_Movies.csv' contains movie data from IMDB including movie title, release year, director, cast, genre, budget, revenue, and ratings. The second file 'Dataset_movies.csv' contains additional movie data including popularity, runtime, tagline, keywords, production companies, and vote count. The third file 'Apidata_movies.csv' is likely another source of movie data, but without additional context it is unclear what data it contains. The merged data is saved as 'merged_data1.csv', which is then concatenated with 'Apidata_movies.csv' and saved as 'Final_combined_dataset.csv'. The latter file is then modified by dropping unnecessary columns, filling NaN values, and saved as 'Movies_file.csv'.

The project involves analysing a dataset of movies using various techniques such as data visualization, clustering, and regression analysis. The aim is to gain insights into the preferences of the general public regarding movies, as well as the relationships between movie ratings, genre, runtime, budget, revenue, and popularity. The results of this analysis can be used to inform decision-making in the movie industry, such as movie production and distribution strategies.

MOTIVATION

Understanding audience preferences - Analysing movie data can provide valuable insights into the types of movies that audiences prefer, which can inform marketing strategies and production decisions for movie studios.

Identifying successful movie characteristics - By analysing data on successful movies, common characteristics that contribute to a movie's success can be identified. This information can be used to guide decision-making in the production of future movies to increase their chances of success.

Improving revenue - Analysing movie data can identify opportunities for revenue growth by understanding box office trends and audience preferences. This can help studios make more informed decisions on how to allocate resources to maximize revenue.

DATA SOURCES

1. **IMDB Data:**
   (https://www.imdb.com/search/title/?count=100&groups=top_1000&sort=user_rating). This URL is already present in the code and is not to be given as an input by the user. It consists of 5 columns and 100 rows. It extracts information such as original_title, year of release, movie rating, gross collection, and director from the IMDb website.The first few rows are displayed as follows

original_title	Year of release	Movie Rat	Director	Gross collection
The Shawshank Redempti	1994	9.3	Frank Darabont	$28.34M
The Godfather	1972	9.2	Francis Ford Coppola	$134.97M
The Dark Knight	2008	9	Christopher Nolan	$534.86M
Schindler's List	1993	9	Steven Spielberg	$96.90M
The Lord of the Rings: The	2003	9	Peter Jackson	$377.85M

2. **TMDB Data:**
   URL=f'https://api.themoviedb.org/3/discover/movie, Using api key I have obtained other columns related to movie analysis. This uses the TMDb (The Movie Database) API to retrieve movie data from the period between 1980 and 2022. It retrieves the top 1000 movies by popularity, sorting them in descending order. The API key is used to authenticate the requests made to the API.

   The code retrieves data such as the movie name, release year, gross collection, movie rating, director name, and popularity. It saves the collected data into a pandas dataframe and then into a CSV file named "Apidata_movies.csv". The code will make a total of 10 requests to the API, with each request returning data for 20 movies (the default page size for TMDb API is 20). Displaying the first few rows and columns of Apidata_movies .

   Note: Due to the unavailability of data for columns gross collection and director we will drop those columns while merging for final data set called Movies_final.csv

Name of movie	Year of release	Movie Rating	Popularity	Gross collection	Director
Avatar: The Way of Water	2022	7.7	2244.6	Unknown	Unknown
Puss in Boots: The Last Wis	2022	8.3	1081.528	Unknown	Unknown
Adrenaline	2022	6	728.154	Unknown	Unknown

3. **Movie set data from Kaggle:**
   I downloaded csv from Kaggle called dataset _movies.csv, and this contains additional columns like revenue, budget, runtime so we can combine and use for analyses. It contains 22 columns and displaying the few rows and columns below.

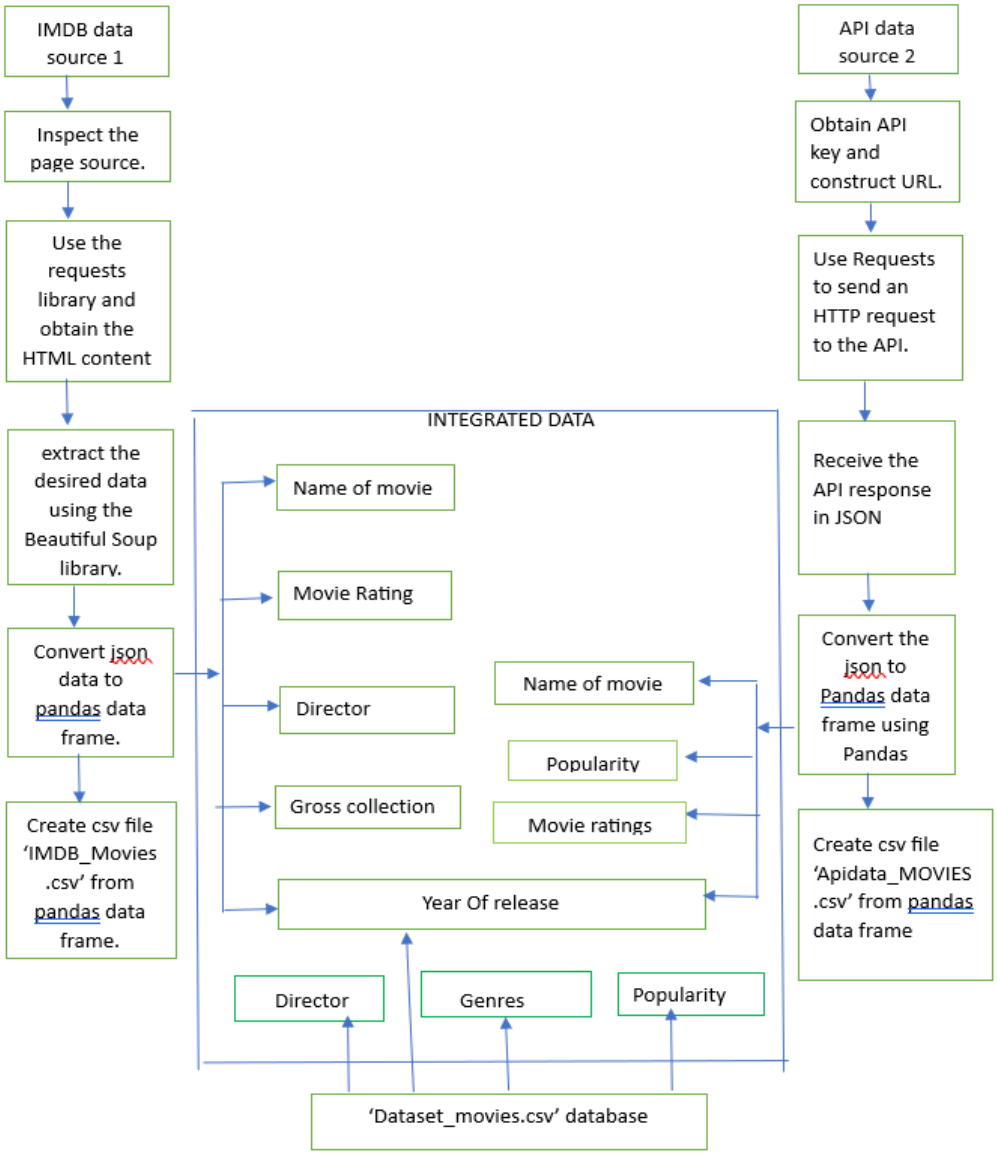
popularity	budget	revenue	original_ti	cast	homepage	director	tagline	keywords	overview	runtime	genres
32.98576	1.5E+08	1.51E+09	Jurassic W	Chris Prat	http://ww	Colin Trev	The park i	monster|(	Twenty-tv	124	Action|Ad
28.41994	1.5E+08	3.78E+08	Mad Max:	Tom Hard	http://ww	George Mi	What a Lo	future|ch	An apocal	120	Action|Ad
13.11251	1.1E+08	2.95E+08	Insurgent	Shailene V	http://ww	Robert Sc	One Choic	based on	Beatrice P	119	Adventure

4. Movies final dataset:

From the above 3 datasets files, I have performed pandas' operations and created a final file called Movies_final.csv. This file we will be using it for analysis of project and it contains all the required columns and rows. The first few rows and columns are displayed as follows

original_title	cast	director	tagline	keywords	overview	runtime	genres	vote_cour	Year of rel	Movie Rat	Gross collection
The Shawshank R	Tim Robbi	Frank Dar	Fear can h	prison|co	Framed in	142	Drama|Cr	5754	1994	9.3	$28.34M
The Godfather	Marlon Br	Francis Fo	An offer y	italy|love	Spanning	175	Drama|Cr	3970	1972	9.2	$134.97M
The Dark Knight	Christian I	Christoph	Why So Se	dc comics	Batman ra	152	Drama|Ac	8432	2008	9	$534.86M

FLOWCHART FOR DATASET GENERATION:



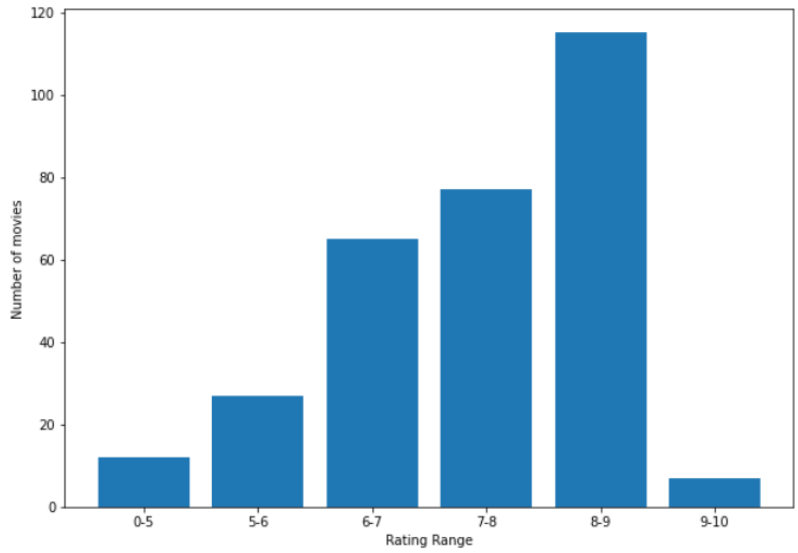
ANALYSIS PERFORMED

With the help of integrated data set called movies_final.csv, we can first analyse the top-rated movies. The table is show as below.

	original_title	Movie Rating	Year of release
0	The Shawshank Redemption	9.3	1994
1	The Godfather	9.2	1972
2	The Dark Knight	9.0	2008
3	The Lord of the Rings: The Return of the King	9.0	2003
4	Schindler's List	9.0	1993
5	12 Angry Men	9.0	1957
6	The Godfather Part II	9.0	1974
7	Pulp Fiction	8.9	1994
12	The Lord of the Rings: The Two Towers	8.8	2002
16	Rocketry: The Nambi Effect	8.8	2022

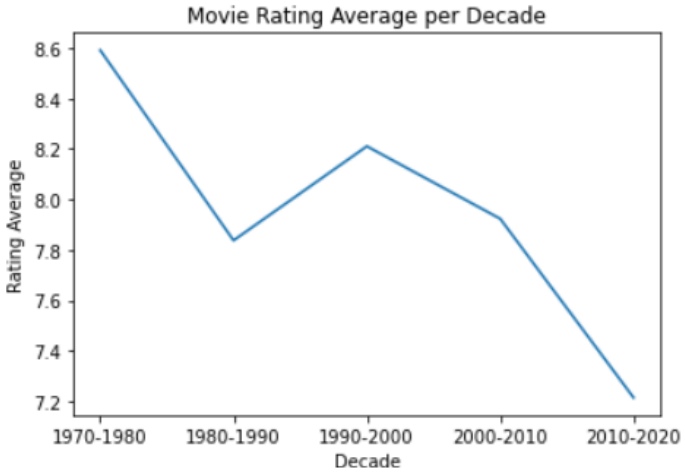
Figure: Top 10 rated movies

The table provides information about the top 10 movies based on their public ratings, which can be useful for people who are looking for popular and highly-rated movies to watch. The rankings can also provide insights into which types of movies are currently popular among the public. Additionally, the table could be used as a reference point for movie industry professionals, such as producers and distributors, who are interested in understanding the public's movie preferences and trends.



We can observe that the majority of movies fall in the rating range of 7-9, with the highest number of movies falling in the rating range of 8-9. This indicates that the general public tends to rate movies between 7 and 9 more frequently. Additionally, we can see that there are relatively few movies with a rating of 9-10, indicating that it is difficult for movies to achieve such high ratings. This information can help movie producers adjust their strategies to align with the preferences of the general public. For example, producing more movies in the 7-9 rating range may be a good strategy. Additionally, knowing the expected rating range can help in predicting the popularity of a movie, which can inform promotion and distribution decisions.

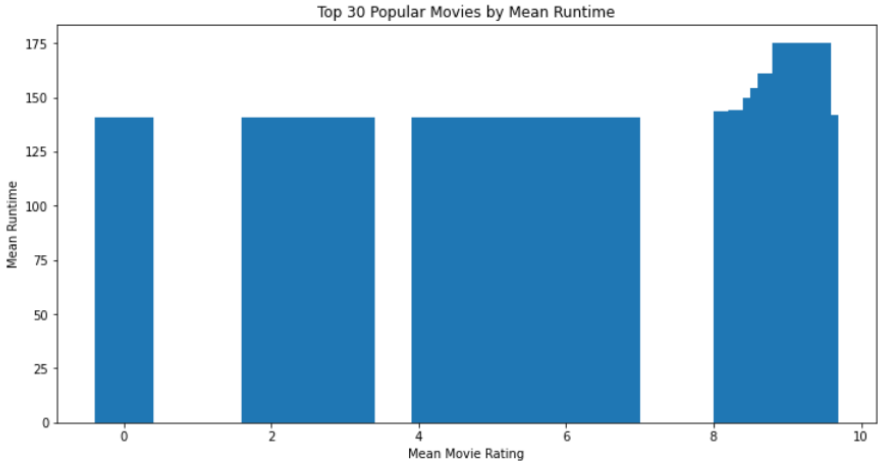
Plot loading for Movie Rating Average per Decade



This analysis provides insight into the trend of movie ratings over time. The high rating in 1970-1980 could be due to the popularity of classic movies from that era. The decrease in rating in the following decade could be due to changes in the movie industry or changes in audience preferences. The gradual increase in rating from 1990 to 2000 suggests that the movie industry may have adapted to the changing preferences of the audience and produced movies that were more appealing.

The gradual decrease in rating from 2000 to 2020 could be due to several factors, such as an increase in the number of movies produced, a decrease in the quality of movies, changes in audience preferences, or increased competition from other forms of entertainment.

Based on this analysis, movie producers can adjust their strategies to improve the ratings of their movies. For example, they could focus on producing movies that appeal to the current audience preferences or invest more in improving the quality of their movies. Additionally, this analysis could be used to identify the factors that contributed to the decline in movie ratings over time and address them to improve the overall quality of movies.



The movies with the highest ratings (9-10) have a longer average runtime compared to the movies with lower ratings (0-6). This may indicate that movies with a longer runtime are more

likely to be highly rated by viewers. However, this analysis is only based on the top 30 most popular movies, so the results may not be representative of the entire population of movies.

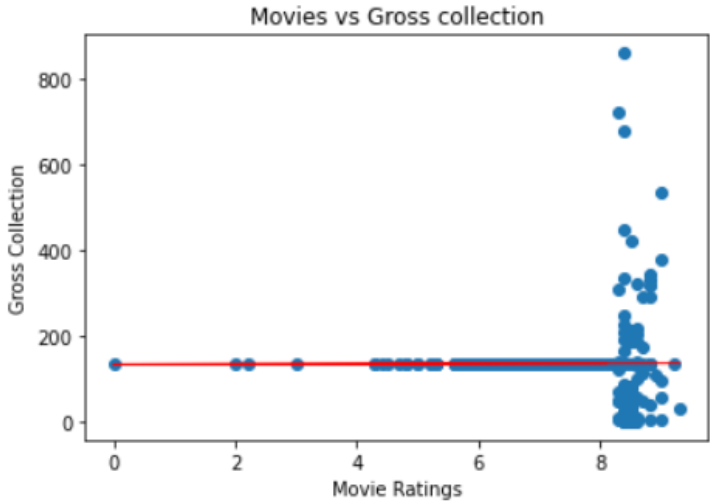
GENERES	AVERAGE RATINGS
Fantasy	8.8
Crime	8.71
Adventure	8.63
Romance	8.6
History	8.6
Action	8.56
Drama	8.55
Science Fiction	8.53
Thriller	8.51
War	8.5

Figure: Table shows top 10 genres, average rating

The output values show the top 10 genres in the dataset and their average rating. It means that, on average, movies belonging to the genre "Fantasy" have the highest rating of 8.80, followed by "Crime" with 8.71 and "Adventure" with 8.67. It provides insights for movie enthusiasts who want to explore popular and highly-rated movies in different genres. Similarly, we can use this information to target specific audiences and promote movies to them.

```
T-value:  3.3163575848599707
P-value:  0.0031219152482495354
```

High budget movies can also get low gross and low budget movies can get high gross. The relationship between budget and gross is not always straightforward and can be affected by various factors such as marketing, star power, genre, release date, and more. The t-test results only show that there is a statistically significant difference between the budgets of high-grossing and low-grossing movies, but it does not imply a causal relationship.



The scatter plot shows the relationship between these two variables, and the linear regression line shows the trend in the data. The purpose of this is to visualize the relationship between movie ratings and gross collection, and to explore whether there is a linear relationship between these variables that can be modelled using linear regression. This explores the relationship between movie ratings and gross collection using a scatter plot and linear regression model. It suggests that there is a positive correlation between these variables, meaning that movies with higher ratings tend to have higher gross collections. Based on the scatter plot and linear regression model, it appears that movies with a minimum gross collection of 200 million $have a higher chance of having a higher rating.

Column	Pearson	Kendall	Spearman
budget	0.401175	0.439849	0.473388
revenue	0.565491	0.567883	0.603614
runtime	0.102310	0.155242	0.169766
vote_count	0.401043	0.448278	0.487756
Year of release	0.247230	0.238229	0.303757

The table shows the correlation coefficients (Pearson, Kendall, and Spearman) between different variables in the dataset. Correlation coefficients range from -1 to 1, with 1 indicating a perfect positive correlation, 0 indicating no correlation, and -1 indicating a perfect negative correlation.

In this case, we can see that revenue has the highest correlation with movie rating, followed by vote count and budget. This suggests that these variables may have a strong relationship with movie rating.

EXTENSIBILITY AND MAINTAINABILITY:

- The project can be extended by adding new features to the analysis. For example, additional data such as user ratings, production budgets, or box office revenues could be incorporated to provide more comprehensive insights into the movie's dataset. The code can be extended to retrieve more pages of data from the API by modifying the range in the for loop, but this would increase the time required for analysis.
- If the structure of the IMDb website changes, such as changes in class names or HTML structure, the code may fail to scrape data correctly. This may require updating the Beautiful Soup parsing logic to adapt to the changes in the website structure.
- The linear regression model can be used to predict gross collections based on movie ratings, but its accuracy may be limited by the variability in the data.

CONCLUSION

The movie analysis provides valuable insights into public movie preferences and trends. The majority of movies fall within the rating range of 7-9, indicating that producing movies in this range could be a viable strategy for movie producers. Additionally, the analysis suggests that movie ratings have decreased over time, possibly due to changes in audience preferences or increased competition. Thus, producers could focus on producing movies that appeal to the

current audience or invest in improving the quality of their movies to enhance ratings. Longer runtime movies are more likely to be highly rated by viewers, and certain genres like Fantasy, Crime, and Adventure tend to have higher ratings. This information can help movie enthusiasts explore popular and highly-rated movies in different genres, and help movie producers target specific audiences and promote movies to them. The analysis also shows a positive correlation between movie ratings and gross collections, indicating that movies with higher ratings tend to have higher gross collections. However, the relationship between budget and gross is not straightforward, and multiple factors can affect it, so movie producers should consider various factors when predicting the success of a movie. Finally, the analysis provides correlation coefficients indicating that revenue, vote count, and budget have a strong relationship with movie ratings, providing valuable information for movie producers to understand the factors that contribute to movie ratings and develop strategies to improve them