PROJECT REPORT ON ANALYSIS OF THE MOVIE DATABASE

N.B – My findings investigated in this Dataset are Tentative

Introduction

The Dataset being analyzed is The Movie Database which gives a comprehensive details of movies produced from 1960 to 2015 with movie specifications such as their cast, the genres of the movie, the popularity of the movie, the director, the cost of movie production, the revenue generated, original movie title, release date, release year, vote count as well as vote average.

The questions posed by me regarding the movie database are stated below:

- Which genres of movies were mostly produced over the years?
- What are the movie details with the highest Popularity?
- What are the movie details with the highest Revenue?
- What are the movie details with the highest Budget?
- What are the movie details with the highest Profit?
- What are the movie details with the highest vote count?
- What are the properties associated with movies with highest Revenue?
- What are the properties associated with movies with the highest Budget?
- What was the average 2010 revenue for each genre of movies produced?
- What was the average popularity for each genre of movies produced?
- What are the movies with the highest top ten revenues?
- What are the movies with the highest top ten budgets?
- Which movies are the most popular top ten movies?
- What are the genres of movies with the highest budget for movie production?
- What are the genres of movies with the highest revenue from movie sales?
- What is the movie title with the highest budget amongst highest revenue movies?
- What is the movie title with the highest revenue amongst highest budget movies?

After asking the following questions as stated above, I proceeded to set up my environment by importing the packages required for the project analysis which includes pandas, numpy, matplotlib, seaborn. I then proceeded to load my data using the pandas read_csv and the shape of the dataset was obtained which contained 10,866 rows and 21 columns. I used the describe functions to analyze the numerical data which produced the count, mean, standard deviation, minimum, (25th, 50th and 75th) percentile as well as their maximum respectively. It was then discovered from the analysis of the describe function that the columns; budget, revenue, runtime, budget_adj, revenue_adj had zeros as their min, 25th and 50th percentile values respectively. These are as a result of some errors in the data in

which most of them are zeros which are Nan values. These errors were addressed using numpy (np.nan) which were used to replace the zero values. The affected columns were passed as a list (zero_cols) and the necessary code was passed as shown in the Jupyter notebook to replace the zero values. The describe function was then called again and the errors were fixed. The info function was used to determine the non-null counts of each column as well as their data types respectively. However, the isnull function was used to determine the count of all null values in each column. It was observed that columns; homepage, keywords, tagline and production companies have high null values.

Data Cleaning/Wrangling

The column names were determined using df.columns. I dropped the columns with very high null values from the isnull functions which are columns; imdb_id, homepage, tagline, keywords, production_companies and overview respectively. These were set up as a list in cols_to_drop and were then called up in the drop function to drop the columns and using inplace=True. After dropping, I re-confirmed the columns had been dropped using the df.columns again to check the column names excluded. The duplicated function was used to check for duplicates in the datasets and the drop_duplicates function was used to drop the duplicates and this was re-confirmed using the duplicated function again. The rename function was used to rename the last two columns from "_adj" to "_2010" since it was budget and revenue for 2010 dollars as stated in the dataset options overview and notes section. The genres of movie columns had a pipe which was used to cluster the genres together. These needed to be splitted into separate genres and this was achieved using the split genre string, at |, explode the list into rows. The genres were then splitted into different rows with the same index number which was a remarkable feat. The fillna function was used to fill all null values in the cast, director and genres columns with "Not Applicable" since they are not numerical data types where the mean values can be used. The isnull function was then used to re-confirm that all null values are zero except the numerical columns in which we replaced their zero values with np.nan in zero_cols list.

Exploratory Data Analysis

The questions posed earlier in the introductory sections were provided with answers using different codes for analysis. These answers can be obtained in the Jupyter notebook cells where for instance, the genres of movies mostly produced over the years were Comedy, Drama and Documentary as shown in the code cell. Other questions posed such as what are the movie details with the highest Popularity?, What are the movie details with the highest Revenue?, What are the movie details with the highest Budget?, What are the movie details with the highest Profit? What are the movie details with the highest vote count? were also provided answers to in different code cells as stated in the Jupyter notebook. In addition to these, analyses were made by grouping the genres of movies and calculating the average

2010 revenue for each genre of movies. The average popularity for each genre of movies were also obtained by grouping the genres of movies and then calculating the mean of the popularity for each grouped genre of movies.

Furthermore, in the course of exploratory analysis, the movies with the highest top ten revenues were ascertained using codes in the cell and were stored as 'mhr'. On the other hand, the highest top ten budgets expended on movie production as well as the highest top ten most popular movies were also ascertained and stored as 'mhb' and 'mhp' respectively. These stored values (mhr, mhb and mhp) were used to establish relationships such as the genres of movies with highest budget for movie production, the genres of movies with the highest revenue from movie sales, the movie titles with the highest popularity, the movie titles budget for production of highest revenue movies as well as the movie title with revenue relationships for highest budget movies. These relationships were established using bar charts and pie charts respectively.

Visualizations/Results

From the exploratory data analysis, it can be shown that Comedy, Drama and Documentary genres of movies were mostly produced over the years. Jurassic World which is in the Action|Adventure|Science Fiction|Thriller genre was the movie with the highest popularity. Avatar which is in the Action|Adventure|Fantasy|Science Fiction genre was the movie with the highest revenue. Warrior's way which is in the Action|Adventure|Fantasy|Western|Thriller genre was the movie with the highest budget. Profit columns were created for the original budget and revenue as well as 2010 budget and revenue by subtracting their revenues from their budgets. Similarly, Avatar was the movie with the highest profit. Inception which is in the Action|Mystery|Science Fiction|Thriller|Adventure genre was the movie with the highest vote count. The results for the top ten highest revenue movies, top ten highest budget movies and the top ten most popular movies were also shown in the Jupyter notebook code cells.

However, from the visualizations, it can be shown that Action followed by Fantasy are on average the genre of movies with the highest 2010 revenue. It can also be shown that Adventure followed by Science Fiction are on average the genre of movies with the highest popularity. A function was created to avoid repetitive coding to be able to call out the plot for the Visualizations of the budget and revenue vs their release years. It was shown from the visualization of revenue vs the release year using scatterplot that the revenue realized from movies increased significantly over the years. Similarly, from the visualization of budget vs release year, it was shown that the budget expended for movie production also increased significantly over the years. It was also observed that the frequency distribution of movies by release year was more skewed to the left while the frequency distribution for budget and revenue were skewed to the right with budget being more skewed to the right

than revenue. The distribution for the entire datasets columns were made to show the relationships between them. From visualization, there was a huge difference between the profit obtained from the 2010 inflation budget and revenue data and the original profit from the normal budget and revenue data. It was shown that the genres of movies with the highest revenue was Action|Adventure|Fantasy|Science Fiction followed by Drama|Romance|Thriller. It was also shown that the genres of movies with the highest budget for production was Adventure|Fantasy|Action|Western|Thriller followed by Adventure|Action|Fantasy. It was also shown that Jurassic World was the movie with the most popularity followed by Mad Max: Fury road and Interstellar. In addition, Avengers: Age of Ultron was the movie with the highest budget amongst the top ten movies that made the highest revenues. Similarly, Avengers: Age of Ultron was also the movie that generated the highest revenue amongst the top ten highest budget movies.

Conclusion

It can be concluded from the analysis of the movie database after thorough investigation of the dataset that Action genre followed by Fantasy are on average the genre of movies with the highest 2010 revenue. It can also be concluded that Adventure followed by Science Fiction are on average the genre of movies with the highest popularity. It can be concluded that the genres of movies with the highest revenue was Action|Adventure|Fantasy|Science Fiction followed by Drama|Romance|Thriller. It can also be concluded that the genres of movies with the highest budget for production was Adventure|Fantasy|Action|Western|Thriller followed by Adventure|Action|Fantasy. It can be concluded that Jurassic World was the movie with the most popularity followed by Mad Max: Fury road and Interstellar. In addition, Avengers: Age of Ultron was the movie with the highest budget amongst the top ten movies that made the highest revenues. Similarly, Avengers: Age of Ultron was also the movie that generated the highest revenue amongst the top ten highest budget movies. It can be concluded that there was a huge difference between the profit obtained from the 2010 inflation budget and revenue data and the original profit from the normal budget and revenue data.

Limitations

1. I have used the TMDB Movies dataset for my analysis and worked with popularity, revenue and runtime. My analysis is limited to only the provided dataset. For example, the dataset does not confirm that every release of every director is listed.
2. There is no normalization or exchange rate or currency conversion is considered during this analysis and our analysis is limited to the numerical values of revenue.
3. Dropping missing or Null values from variables of our interest might skew our analysis and could show unintentional bias towards the relationship being analyzed.

Reference: N/A