MICROSOFT MOVIES PROJECT

1. BUSINESS UNDERSTANDING

1.1 UNDERSTANDING THE PROBLEM

The movie industry is one of the most lucrative and thriving industries around the world and as a result of this companies are trying to step into the Industry to seek the benefits that comes along with the business. Microsoft as a company wants to start venturing into a new movie studio but due to the inadequacy of the knowledge about movies, they want to know about what types of films are currently doing the best at the box office. As an enthusiast Data Scientist coming up with data that will help the head of Microsoft's new movie studio on what type of films to create is a big step in that will assist the company. The task is to process data from different sources from which the movies are found in order to explore what types of movies they should create.

1.2 PROBLEM STATEMENT.

The problem statements is to define what movies types are currently doing well in the market .To do this ,I will have to use the explatory data analysis on several datasets to generate insights for the head of Microsoft movie studio and to tell exactly what movie type of films Is doing well.

2. DATA UNDERSTANDING

2.1. DATA COLLECTION

The data was collected from various sources like Box office IMDB which contains different movies/films types.

2.2.DATA DESCRIPTION

This project contains 2 dataframes that is bom.movie_gross and im.db.

Bom.movie_gross;

This dataset contains 3387 rows and 5 columns which represents different distinct features of the dataset. From this we find that the dataset entries are majorly object datatypes except the column year which is an integer. The columns that will be given much weight interms of investigation will be studio, domestic gross and foreign gross.

Studios- this refers to the major entertainment company or motion picture company that has its own privately owned studio facility or facilities that are used to make films ,which is handled by the production company.

Domestic Gross - this refers to the monetary value of final goods and services that are bought by the final user produced in a given period of time for example annually, semi-anually or quartely. It counts all the output generated within the boarders of a country.

Foreign Gross- this refers to the monetary value of final goods and services that are bought by the final user produced in a given country. It counts all the output generated within foreign country.

im.db

This is the second dataset that will be looking into.under this we are going to dig deep into two aspects of this dataset that is movie_ratings and movie_basics. For it to bring out the clear picture and meet its intended purpose we have to join the two and work with it as one. That bring us into movie_rating_basics.

movie_ratings- this has different aspects that has to be looked into for it to be complete . this include;

Rated G : this is for General audiences - all pages admitted

Rated PG: Parental guidance suggested for some material might not be suitable for childrens view.

Rated PG-13: Parents strongly cautioned against some material which are inappropriate for children under the age of 13.

Rated R: Restricted - persons under the age of 17 requires parents accommpaniment in order to watch .

Movie_basics; these are filmmaking basics that every filmmaker needs to know.from shot sizes and angles to the stages of production.

3. DATA PREPARATION

3.1 SELECTING DATA

In data selection we are going to use all the columns that are relevant to the task. The next thing was to load data which is to be used for the analysis.

3.2 DATA CLEANING

Here the first thing was to to check the null values after which we also checked for the missing values and we found out that indeed there were missing values in some of the columns for example studio had 5 missing values,domestic gross had 28 missing values and lastly the foreign gross which had the biggest number of missing values at 1350. looked at the dataset to find out if there were any duplicated values and found out none meaning the data was found to be consistent .also checked to see if there was any inappropriate datatypes. once that was done we replaced the missing values. we later looked at the outliers within the dataset.

4.DATA ANALYSIS

4.1 EXPLORATORY DATA ANALYSIS

From our dataset we find that the mean of the foreign gross is greater than the mean of the domestic gross .

The median of the foreign gross is also greater than the median of the domestic gross.

It is also evident that the median of the movie_rating_basics is greater than its mean.

5. CONCLUSION

In conclusion ,Drama is the most popular genre, followed by comedy, comedy/Drama with the comedy,Romance and Horror Thriller being the least popular. In studio performance the IFC

studio is the leading studio and Bv being the last according to our dataset.

6. RECOMMENDATION

The above information about movies on the two datasets that is bom_movie gross and the im.db .This suggest that the movies foreign gross , domestic gross and movie_rating_basics relate to the studio and how they can influence the setting up of a film studio and what to expect. The mentioned aspects has a greater impact to the company’s involvement in a movie industry.