MICROSOFT MOVIE PROJECT

1. BUSINESS UNDERSTANDING

1.1 :OVERVIEW

Microsoft , an American multinational technology company, wishes to create original video content , which has become quite popular . The company wishes to achieve this by creating a new movie studio tasked with movie production.However, they do not know anything about creating movies and need insight on the kind of genres they will need to produce based on certain factors. This project will entail the analysis of data provided from various sources such as IMDB and rotten tomatoes to help generate insight on this project so that the genres recommended at the end of the analysis will be the best fit in terms of Box office performance .

1.2: PROBLEM STATEMENT

This analysis will revolve around finding the best genre of movies that the new Microsoft Movie studio can produce. The project will centre its analysis based on movie ratings , the gross money returns as well as do an in depth analysis on the relationship between Box office performance and the production costs to get an approximated project budget.

2. DATA UNDERSTANDING

2.1: DATA COLLECTION

The data that will be key in the analysis that is about to begin has been gathered from various movie review and summary websites.These websites entail Box Office Mojo (https://www.boxofficemojo.com/) IMDB (https://www.imdb.com/) Rotten Tomatoes (https://www.rottentomatoes.com/) TheMovieDB (https://www.themoviedb.org/) The Numbers(https://www.the-numbers.com/) However, i will only include the dataset from IMDB and The Numbers.

2.2: DATA DESCRIPTION.

The datasets that will be in use are the IMDB and The Numbers datasets. The IMDB dataset is in the form of an sqlite database while The Numbers dataset is in the form of a CSV file.The IMDB database contains two tables each containing a common column which is the movie id column. The Numbers dataset, assigned the variable name movie_budget_df has 5782 rows and 6 columns, with the column names as id ,release_date , movie ,production_budget ,domestic_gross and worldwide_gross. The IMDB dataset assigned the variable name Im.db_df has two relevant tables to the analysis, namely ; movie_ratings table and movie_basics table. The two tables have been converted to dataframes and the movie_ratings table, assigned the variable name ratings_df has a total of 73856 rows and 3 columns. The column names are movie_id, averagerating and numvotes.The average rating is affected by the numvotes column. The movie_basics table has been converted to a dataframe with the variable name movie_basics_df. All these datasets have been merged , beginning with the dataframe movie_budget_df and the the movie_basics_df and the merged dataframe assigned the variable name merge_df. The final merge has been conducted between the merge_df dataframe and the ratings_df and the final dataframe which will be in use for the remaining part of the analysis assigned the variable name final_merge_df.
The final_merge_df dataframe contains some missing values.

2.3 : DATA PREPARATION AND CLEANING

The final_merge_df dataframe has a total of 118 missing values in the runtime_minutes and 8 missing values in the genres column. The approach taken to clean this dataset is to remove the missing values. This has been achieved by dropping the entire runtime_minutes column since it will not be essential in my analysis as well as dropping the rows which have missing values in the genres column. The next step is to check for any duplicate values and in the final_merge_df dataframe , there are no duplicated values. There are also no outliers in the dataframe as well as extraneous values in the dataframe.

3.DATA ANALYSIS