PHASE 1: INDIVIDUAL PROJECT.

1.0:PROJECT INTRODUCTION.

This project entails research that uses exploratory data analysis from ("*zippedData*") to generate insights on creating a new movie studio by Microsoft.

1.1: PROJECT OVERVIEW.

In the conceptualization of this project, I assumed the role of a Data Scientist at Microsoft. This project entails a comprehensive understanding and analysis of movies from ("*zippedData*") to come up with business-related recommendations for the head of Microsoft on the creation of a new movie studio. The main purpose of this project is to make use of exploratory data analysis to acquire information required to recommend appropriate actions by Microsoft on the creation of the new movie studio. This project is supplemented by visualizations to assist in the conceptualization of the findings.

1.2: BUSINESS UNDERSTANDING.

The key objective of this project is to formulate data-backed recommendations for Microsoft in the creation of the new movie studio. Throughout this project, as a Data Scientist, I intend to formulate three key recommendations that will answer the following business questions.

1. Which directors should Microsoft Studio employ in creation of movies?
2. Which writers should Microsoft Studio employ in creation of movies?
3. What is the best runtime range for movies?
4. Which are the most preferred movie genres among movie fans?
5. In which regions should Microsoft studio focus its movie marketing efforts?
6. How does the movie ordering affect the number of movies watched?

Using the data sets given, I intend to use data understanding and analysis to answer the questions above which will form the foundation of my recommendations to Microsoft. I have formulated the questions above to be a guide to the data sets I require and to make judgments on the type of analysis I would need to do.

2.0:PROJECT BODY

This section of the project entails a detailed understanding and analysis of the data collected in an attempt of answering the questions in the introduction. The analysis done in this section will act as a foundation for the recommendations of my project.

2.1:DATA UNDERSTANDING.

2.1.1:DATA COLLECTION.

The data used in this project was collected from various sources including Box Office Mojo, IMDB, Rotten Tomatoes, TheMovieDB, and The Numbers, and compiled into a folder (*"zippedData"*). From the folder, I chose to use the recommended data sets which include: The ("*zippedData/im.db*") database and ("zippedData/bom.movie_gross.csv.gz") CSV file.

2.1.2:DATA DESCRIPTION.

The ("zippedData/im.db")database contains 8 tables including: movie_basics, directors, known_for, movie_akas, movie_ratings, persons, principals and writers.
The tables contain numerous data from various sources in the following shapes;

1. Principals - It has 1028186 rows and 6 columns.
   The columns are: 1.movie_id - This is a unique identifier for each movie.
   2. ordering - This is the studio requesting a level of production in a movie.
   3.person_id - This is a unique identifier for each person.
   4. category - This shows the occupation of each person.
   5. job - This shows the current job for each person.
   6. characters- This shows the various characters played by each person.
2. Persons - It has 606648rows and 5 columns.
   The columns are: 1.person_id - This is a unique identifier for each person.
   2.primary_name - This is the name associated with the person_id
   3.birth_year - This is the year of birth for each person.
   4.death_year - This is the year of death for each person(if deceased).
   5.primary_profession - This shows the main occupation of each person.
3. Known_for - It has 1638260 rows and 2 columns
   The columns are: 1.person_id - This is the unique identifier for each person.
   2.movie_id - This is the unique identifier for each movie.
4. Directors - It has 291174 rows and 2 columns.

The columns are: 1.person_id - This is the unique identifier for each person.
2.movie_id - This is the unique identifier for each movie.

5. Writers - It has 255873 rows and 2 columns.
The columns are: 1.person_id - This is the unique identifier for each person.
2.movie_id - This is the unique identifier for each movie.
6. Movie_basics - It has 146144 rows and 6 columns.
The columns are: 1.movie_id - This is a unique identifier for each movie.
2.primary_title - This is the main title related to the movie_id.
3.original_title- This is the initial title related to the movie_id.
4.start_year - This is the year of conception of the movie.
5.runtime_minutes- This is the duration of the movie in minutes.
6. genres- This is the type of movie.
7. Movie_ratings - It has 73856 rows and 3 columns.
The columns are: 1.movie_id - This is the unique identifier for each movie.
2.averagerating - This is the rating of each movie out of 10 as per feedback.
3.numvotes - This is the number of feedbacks for each movie's averagerating
8. Movie_akas - It has 331703 rows and 8 columns.
The columns are: 1.movie_id - This is a unique identifier for each movie.
2. ordering - This is the studio requesting a level of production in a movie.
3. title- This is the title related to the movie_id.
4. region - This is the geographical area where the movie is popular.
5. language- This is the audio language of the movie.
6. types- This is the version of the movie.
7. attributes - This shows the unique features of each movie.
8.is_original_title - This shows if the current title is the original title.

The (*"zippedData/bom.movie_gross.csv.gz*") CSV file contains a table with data about movies.
The table has 3387 rows and 5 columns.
The columns are:1.title - This is the title of a movie.
2. studio- This shows which studio conceptualized the movie.
3.domestic_gross- This is the revenue earned by a movie in the country of conception.
4.foreign_gross- This is the revenue earned by a movie outside the country of conception.
5. year -This shows the year in which the movie gained the revenue streams.
The (*"zippedData/tmdb.movies.csv.gz"*) CSV file contains a table with data about movies.
The table has 26517 rows and 10 columns.
The columns are:1. Unnamed:0- This is a repetition of the index.
2. genre_ids- This is the unique identifier for each genre.

3. ids- This is a unique identifier for each movie.
4. Original language- This is the initial language of a movie in its conception.
5. Original_title - This is the initial title for the movie when it was created
6. popularity-This is the count of the number of times a movie is watched.
7. Release_date- This is the date when a movie was released from a studio.
8. Title- This is the name that a movie goes by currently.
9. vote_average-This is the average rating of the votes by fans on a specific movie.
10. Vote_count- This is the number of votes cast by fans on a movie.

The ("zippedData/tn.movie_budgets.csv.gz") CSV file contains a table with data about movies.
The table has 5782 rows and 6 columns.
The columns are:1.id - This is a unique identifier for each movie.
2.Release_date- This is the date when a movie was released from a studio.
3. movie -This is the name of a movie.
4. production_budget-This is the cost of creating a movie.
5. domestic_gross-This is the amount of revenue earned by a movie in its origin country.
6. worldwide_gross-This is the amount of revenue earned by a movie globally.

.

2.2:DATA ANALYSIS.

2.2.1:DATA CLEANING.

The main purpose of data cleaning is to prepare the data for analysis which will then be useable to answer our project questions. Data cleaning was also done to ensure the validity, accuracy, completeness, consistency, and uniformity of the data.I then selected the tables that I would use as the basis for my recommendations. I then cleaned the tables selected using the following steps:

TASK1:The first task was to check if the column names were uniform and readable.

TASK2: The second task was to check for duplicated rows.

TASK3:The next task was to check if there were missing values.

TASK4:The next task was to decide how to deal with the missing values. (Either drop if they are unnecessary or replace the missing values with the best fit.

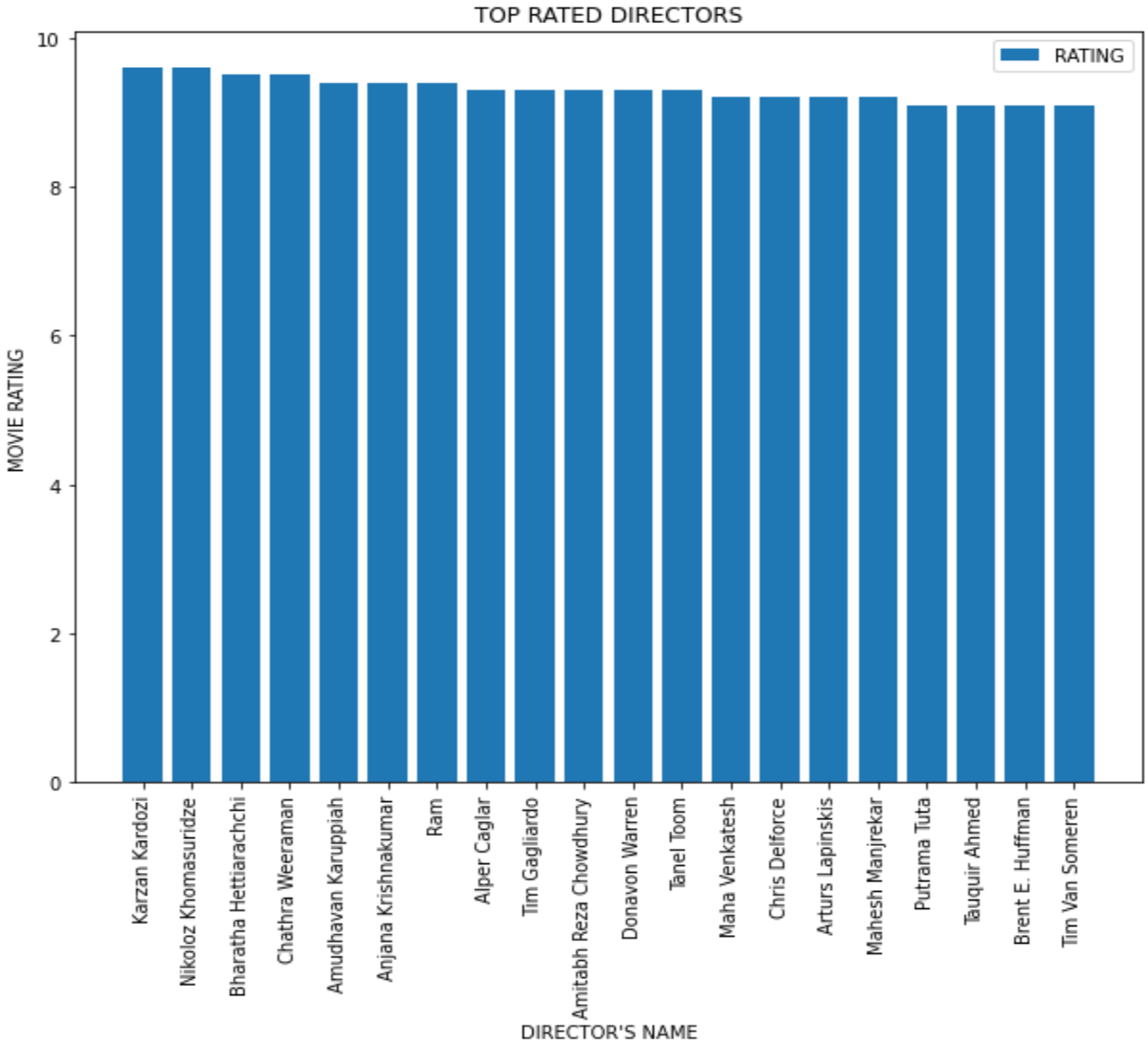
TASK5: The last task was to check if the data types of the columns were correct.(If the data type was wrong I corrected it.

After I had cleaned all the data that I needed. The resultant data frames were ready for analysis.

2.2.2: EXPLORATORY DATA ANALYSIS.

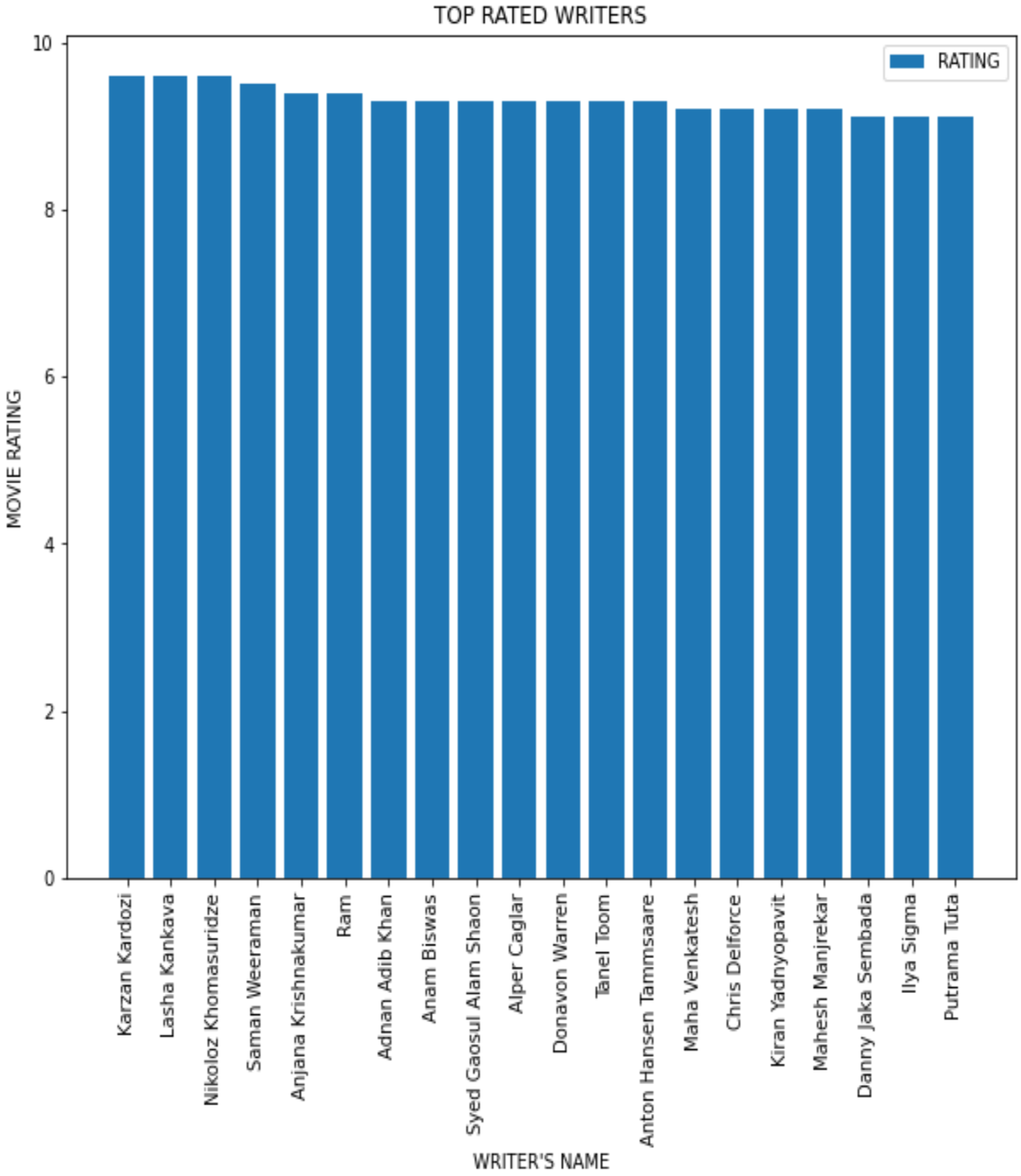
In this section, I investigated the data sets I intended to use in order to create visualizations that would act as the base of my recommendations.

2.2.2.1:Graph1.



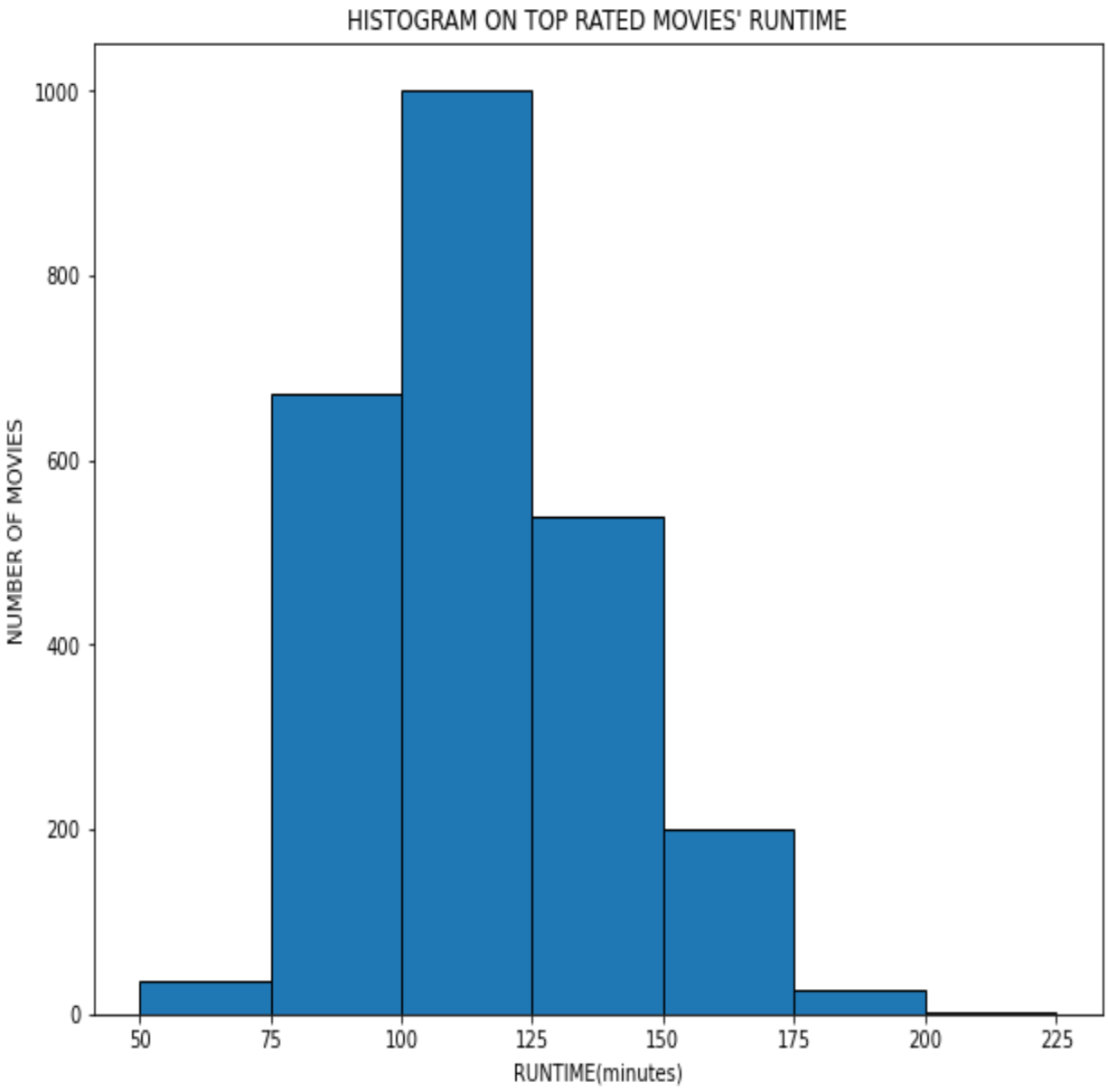
This bar chart shows the best directors in the movie industry. I only took into consideration movies that had above 1000 votes so as to ensure that the average rating was a clear indication of what movie fans think about a certain movie.

2.2.2.2:Graph2.



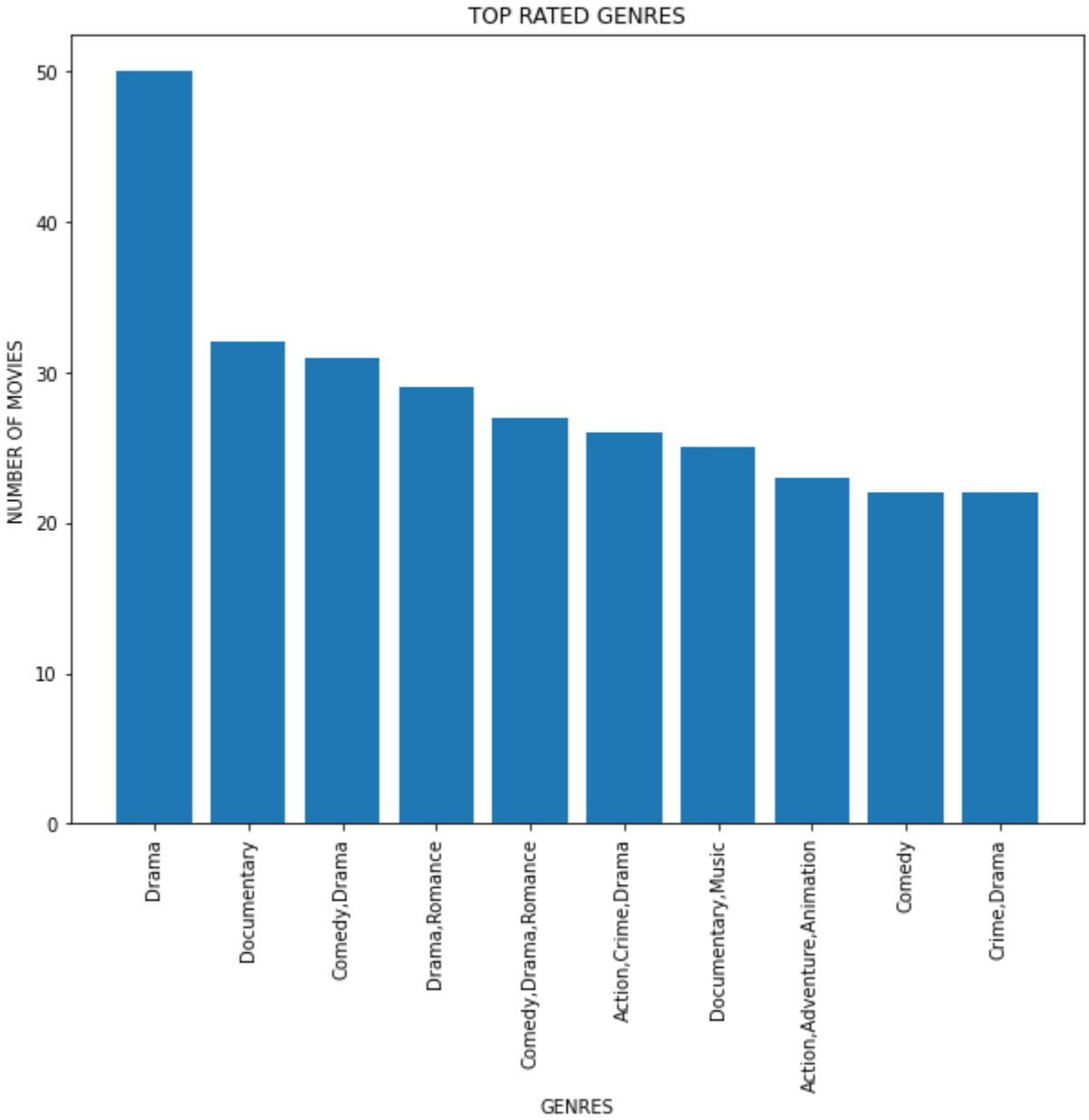
This bar chart shows the best writers in the movie industry. I only took into consideration movies that had above 1000 votes so as to ensure that the average rating was a clear indication of what movie fans think about a certain movie.

2.2.2.3:Graph3.



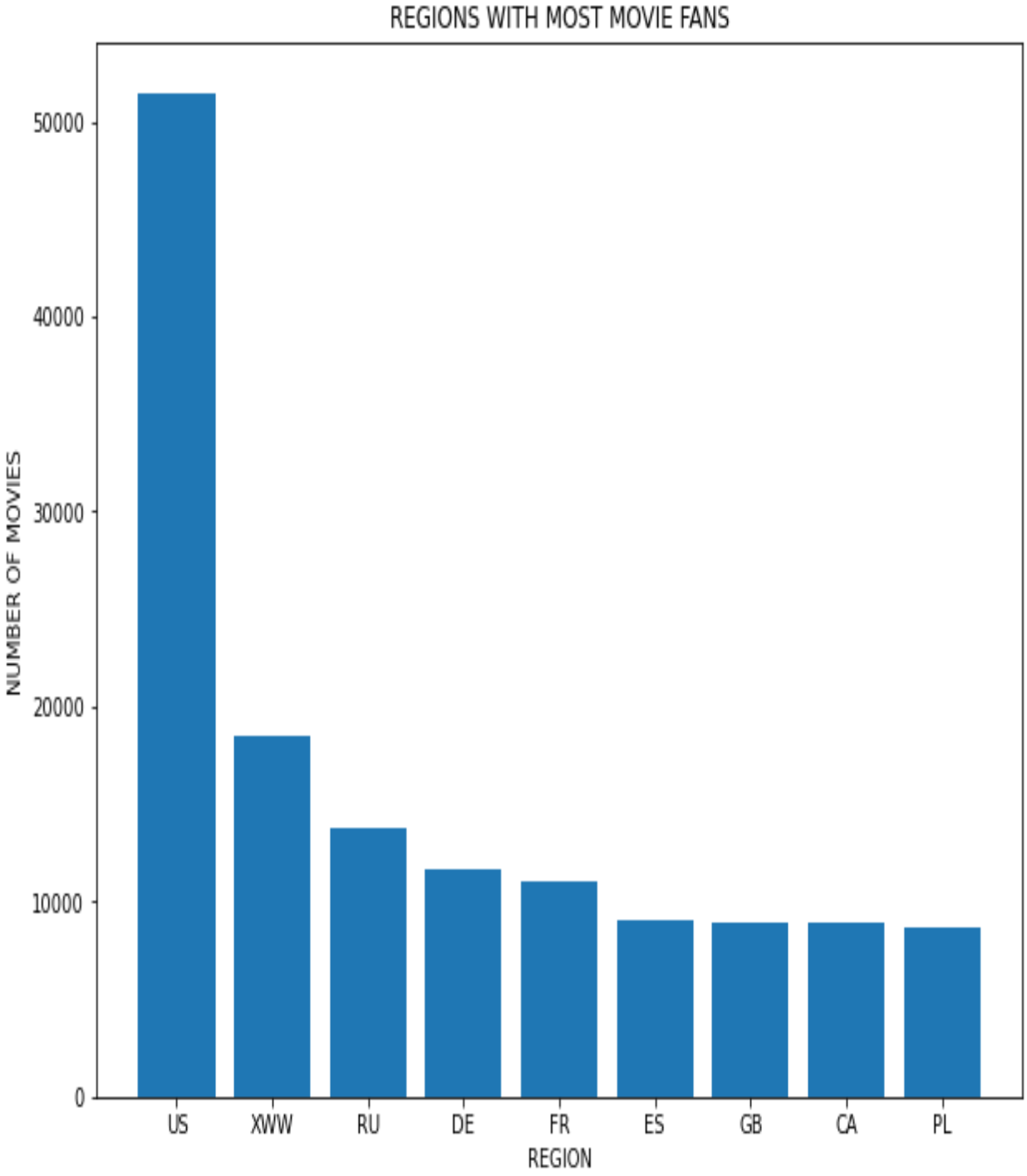
The histogram above shows the runtime in minutes of the movies that had an average rating of above 7 and had a minimum of 1001 votes. This was to show what is the best runtime of a movie especially movies that have a high rating.

2.2.2.4:Graph4.



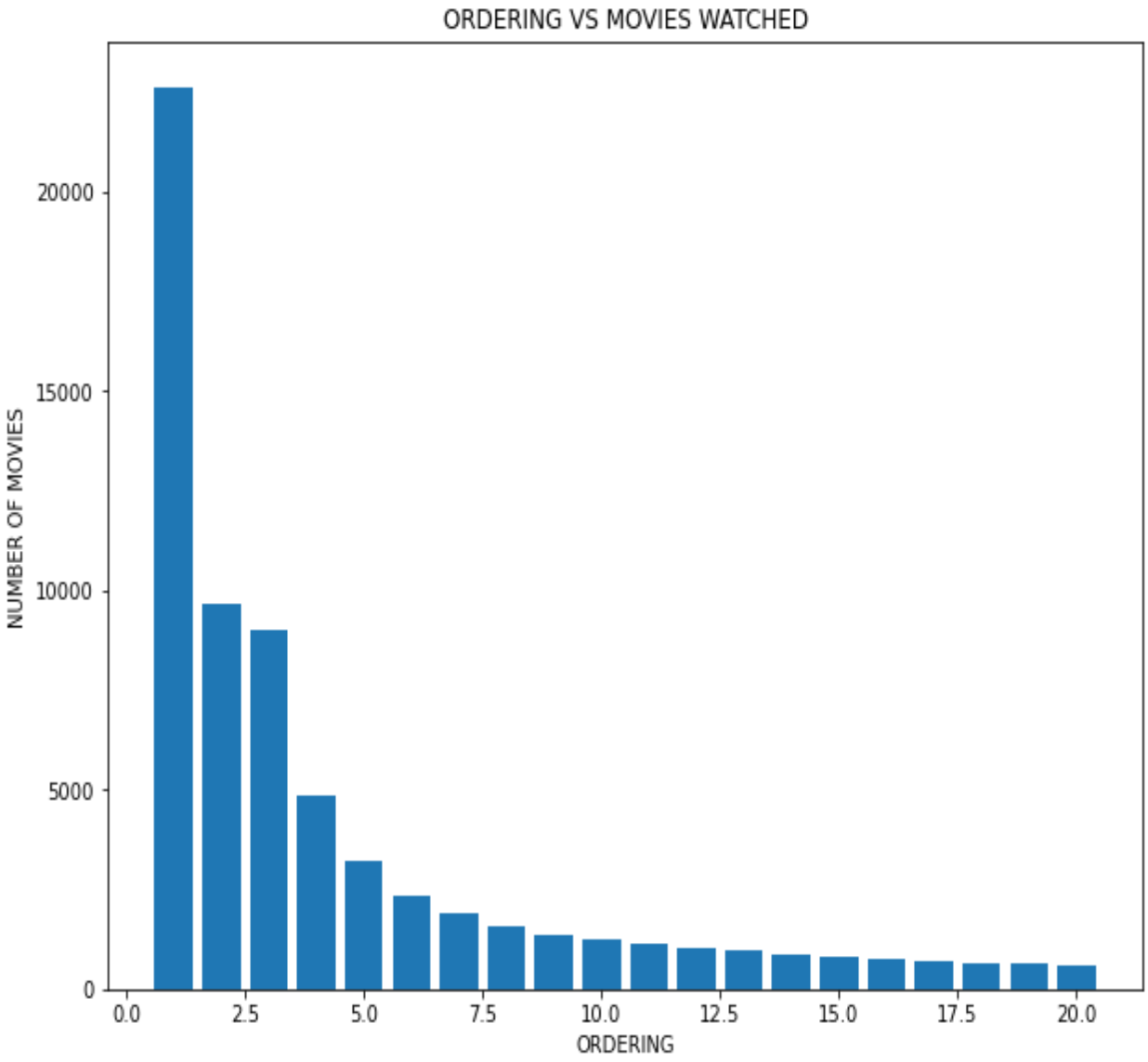
In the bar chart above, I have shown the relationship between the Top rated genres against the number of movies that had an average rating of above 7 and at least 1001 votes. I did this to show which are the most watched among the top-rated genres.

2.2.2.5:Graph5.



The bar chart above shows the regions with the most movie fans based on the number of movies watched per region. I did this visualization to show the relationship between each region and the popularity of movies in the regions.

2.2.2.6:Graph6.



The bar chart above shows the relationship between movie ordering and the number of movies watched. It shows that the number of movies watched reduces as the ordering increases.

3.0:PROJECT CONCLUSION.

RECOMENDATIONS.

1. From graph 2.2.2.1, I recommend that Microsoft Studio employ directors who feature in the figure based on their availability and rating showed by the figure. This is because they have directed the movies with the best rating among movie fans.
2. From graph 2.2.2.2, I recommend that Microsoft Studio employ writers who feature in the figure based on their availability and rating showed by the figure. This is because they have helped create the movies with the best rating among movie fans.
3. From graph 2.2.2.3, I recommend that Microsoft Studio should create movies that have a runtime in the range of (100-125 minutes) . This is because my analysis has concluded that most of the top rated movies have a runtime in this range.
4. From graph 2.2.2.4, I recommend that Microsoft Studio should focus on the following movie genres (Drama, Documentary, Comedy-Drama, Drama-Romance, Comedy-Drama-Romance). This is because among the top rated movies this genres feature the most.
5. From graph 2.2.2.5, I recommend that Microsoft Studio should focus their movie marketing in the following regions (US, XWW, RU, DE, FR). This is because they are the regions with the most movie fans.
6. From graph 2.2.2.6, I recommend that Microsoft Studio should create movies with as little ordering as possible. This is because I found out in my analysis that the number of movies watched is inversely proportional to increase in ordering.

CONCLUSION.

During this project, I dedicated my time and efforts to answer various business questions that have helped me come up with recommendations for Microsoft Movie Studio. This project is backed by analysed data to prove the accuracy and correctness of my recommendations.