Project Overview:

This project aims to provide actionable insights to Microsoft's new movie studio as they venture into the film industry. Microsoft, a technology giant, is looking to tap into the world of original video content creation, but they lack experience in the film industry. To make informed decisions on the types of films to produce, they seek to understand the current landscape of successful movies at the box office. This project leverages exploratory data analysis (EDA) to analyze and extract valuable information from a dataset of box office movie data.

Business Understanding:

Microsoft recognizes the growing trend of major companies producing original video content, particularly in the film industry. They perceive an opportunity to enter this competitive market, but they lack the domain knowledge required to navigate the complexities of the movie industry. To address this, Microsoft aims to conduct a thorough analysis of successful films to gain insights that will guide their decision-making process.

The primary objectives of this project are as follows:

1. **Identify Successful Film Trends:** Analyze historical box office data to identify trends, genres, and attributes associated with successful movies. This will involve understanding which genres, release dates, and production budgets tend to yield the highest box office returns.

2. **Audience Preferences:** Explore audience preferences by examining factors such as movie ratings, runtime, and popular actors and directors. Understanding what appeals to viewers can guide content creation.

3. **Budget Allocation:** Provide insights into the relationship between production budgets and box office performance. This will help Microsoft allocate resources effectively.

4. **Competitive Landscape:** Analyze the competitive landscape by identifying key players, studios, and distribution strategies in the movie industry.

By addressing these objectives, the project aims to equip Microsoft's movie studio head and stakeholders with actionable insights. These insights will enable them to make informed decisions regarding the types of films to produce, when to release them, and how to allocate resources effectively to maximize their success in the movie industry. Through data-driven decision-making, Microsoft intends to position itself as a strong contender in the world of original video content creation.

DATA PREPARATION

We begin by importing all the necessary libraries that we will be working with in this analysis. These are:

1. Pandas
2. Sqlite3
3. Numpy
4. Matplotlib

We then need to establish a connection to the im.db database to get the list of tables that are in the database.

The tables are as follows:

1. Movie_basics
2. Directors
3. Known_for
4. Movie_akas
5. Movie_ratings
6. Persons
7. Principals
8. Writers

Main focus will be on the Movie_basics Table and movie_ratings table.

We then load the csv data to get an understanding of all the datasets and how they may relate with each other.

DATA UNDERSTANDING

We run the .shape function to find how many columns and rows are contained in the movie_basics table. The table has 146144 rows and 6 columns.

The columns and data types in the table are:

```
#   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   movie_id         146144 non-null  object
 1   primary_title    146144 non-null  object
 2   original_title   146123 non-null  object
 3   start_year       146144 non-null  int64
 4   runtime_minutes  114405 non-null  float64
 5   genres           140736 non-null  object
```

Movie_ratings Table

The movie-details table has 73856 rows and 3 columns

The columns and data types are as follows:

```
#   Column         Non-Null Count  Dtype
```

```
---  ------          -------------  -----
 0   movie_id        73856 non-null  object
 1   averagerating   73856 non-null  float64
 2   numvotes        73856 non-null  int64
```

The CSV file has 3387 rows and 5 columns

The columns and data types are as follows:

```
#   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   title            3387 non-null   object
 1   studio           3382 non-null   object
 2   domestic_gross   3359 non-null   float64
 3   foreign_gross    2037 non-null   object
 4   year             3387 non-null   int64
```

Note that the data type for foreign_gross is object yet it should be numeric. We proceed to convert it into a float data type.

We will proceed to join the datasets to get a comprehensive dataset with more details.

We first write a query to merge the movie_basics table to the movie_ratings using the movie_id as the common key.

Change the column name ‘title’ in the csv dataset to ‘primary_title so us to enable us merge the datasets and get more comprehensive data.

DATA CLEANING

Now that we have looked at our data, we need to start cleaning the data i.e to removing duplicates, remove/replace missing data.

We begin by running the .duplicated function to check for duplicates. The data has no duplicate rows.

Null Values and Data Types

We run the .isna function to check for null values per column. Results are as follows

```
movie_id              0
primary_title         0
original_title        0
start_year            0
runtime_minutes      47
genres                7
averagerating         0
numvotes              0
```

```
studio              3
domestic_gross     22
foreign_gross    1195
year                0
```

The columns runtime_minutes and genres, domestic_gross and foreign_gross have missing values

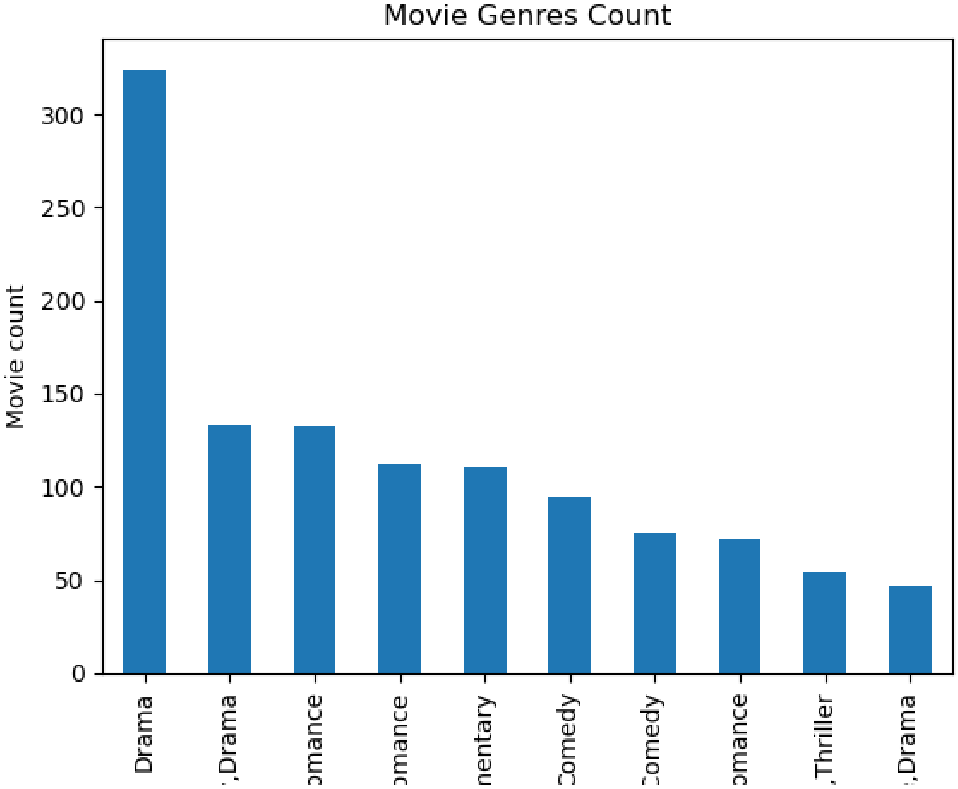
1. Replace the missing values in runtime_minutes with the mean
2. Replace the missing values in domestic_gross with the mean
3. Replace the missing values in genres with the mode
4. Replace the missing values in studio with the model

Summary Statistics

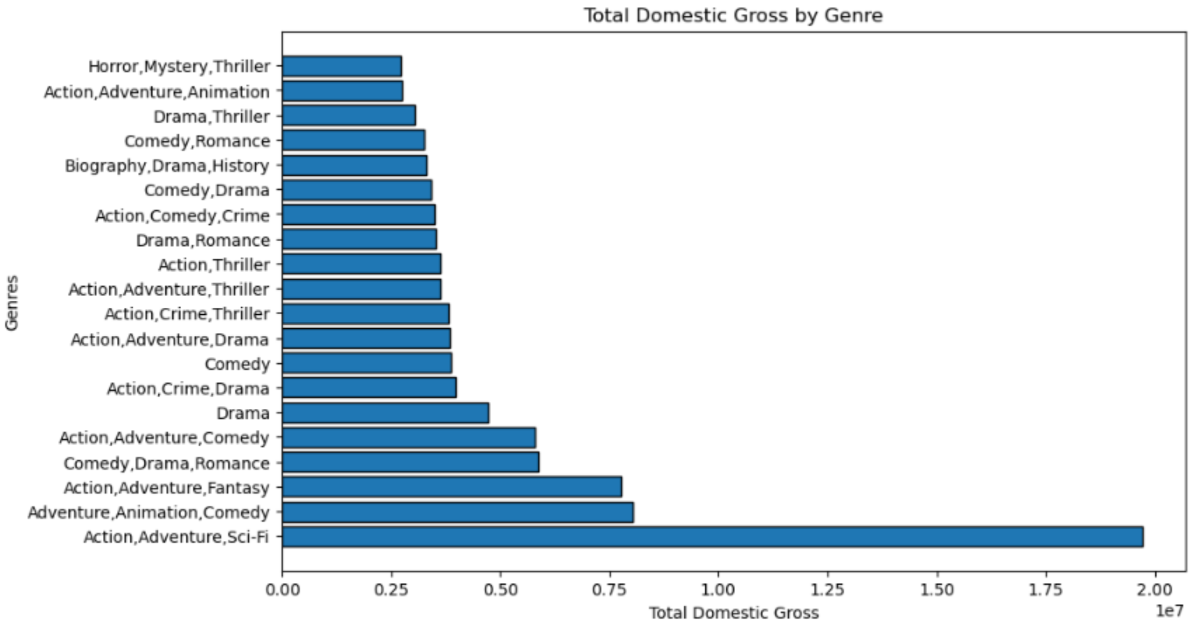
Run the describe function on the dataframe to get the summary statistics.

DATA ANALYSIS

Drama genre has the most productions though it's not the most popular among viewers



Visualization to show the total domestic gross by individual genres



RECOMMENDATIONS

1. Microsoft team should focus on producing videos in the Action, Adventure, Sci-Fi category as it generates quite a high revenue and is also popular among viewers.
2. The team should also focus on the Adventure, Animation, Comedy and Action, Adventure, Fantasy as they are the second and third highest revenue generating consecutively.