Project Overview

This project aims to use exploratory data analysis to generate insights for a business stakeholder aiming to venture into a new line of business

Business Understanding

Microsoft Corporation is an American multinational technology corporation which produces computer software, consumer electronics, personal computers, and related services. They are looking to diversify and venture into the movie scene by opening up a movie studio. Before making the big leap, they need to have a better understanding of the movie industry. To this end, I have been tasked with exploring what types of films are currently doing the best at the box office and then translate those findings into actionable insights that the head of Microsoft's new movie studio can use to help decide what genre of films to produce.

The main question to answer will be what genre of movies perform well in the movie industry and what are the production cost/vs profits. We will also aim to answer the question of what are the names of the film crew associated with the best performing movies.

Problem Statement

The movie industry is without a doubt just as competitive and as dependent on internal and external factors as any other industry. A knowledge management system on the industry is therefore essential in order to navigate the industry successfully. The main challenge is the uncontrollability of the product, there is now way to predict how the audience will interact with it. This large financial gamble therefore necessitates an in-depth analysis on which genres of movies are preferred by majority of the audience and those that generate high returns on investments.

Data Understanding

- Find the dimensions of the dataset; After loading the required python modules, I first opened each of the data sets to see what they contained by loading the first few rows using the .head() pandas method. The goal of any business is making profits so an ideal data set should have production budgets and revenue generated from the movies. Data from the different data sets were loaded and examined to see which database contained relevant information for our analysis.
- The tmdb.movies, rt.movie_info datasets did not have any financial information on the movies therefore I felt it would not adequately answer my question on profitability of the different genres of movies.
- The bom.movie_gross dataset had information on the gross income on both the domestic and world market but no information on production costs therefore it would not be possible to calculate the profits. The data set however has information on the different movie studios and this would have been helpful if our investigation was about best performing studios.
- For the csv data sets provided, I settled on tn.movie_budgets as it had the information I was looking for, production costs and gross profits. For the SQL data set, I joined movie_basics, movie_ratings in order to get the names of the movie crew associated with the high performing movies. Finally after selecting these tables, I used pd.join() to join the dataframe from the csv data and that from the SQL data.

Data Cleaning

Next step was to check the data for any missing values, duplicated values and the type of data contained in the data frames.
Using .info() it was noted the the numeric columns were stored as objects/strings.
The production cost and gross income columns for example contained "$" implying the figures were keyed in as strings and therefore .replace() method was used to clean the data and then the columns were converted to floats. The figures were converted to millions by dividing by 1000000 for readability. Data cleaning involved dropping the missing(null) values as they were too many in the data set. Anew column for world profits was calculated by subtracting the production cost from the world profits.
We ordered the dataframe in descending order using world profits in order to pick the best performing movies by genre. Majority of these were "Action,Crime,Drama" and 'Drama' movies.

Data Visualization

The data visualization was mainly in form of bar plots. The different genres of movies were plotted against world profits. Since most of the top performers were in the two genres "Action,Crime,Drama" and 'Drama', these are the graphs that were produced.
A bar graph based on the value counts of the different genres is also shown.

Findings

Movies under the genre 'Action, Crime, Drama' had the highest ROI in terms on net world earnings with a maximum of 2351 million and a minimum of 98 million. The production budget had a maximum value of 425 million and a minimum of 2.5 million. We can infer from the figures that it's a heavy investment but the returns are indeed rewarding. As seen from the boxplot, half of the profits from this genre above the median value. Comparing the production cost and the net world profits, the ROI in this category is five times over the production cost.
The highest profitable movies are: **Avatar, Avengers: Infinity Wars Star, Wars Ep. VII: The Force Awakens, Avengers: Age of Ultron** The crew behind these movies are **Gulzar(writer), Naushad(composer), Balraj Sahni(actor), David Harbour(actor),Cary-Hiroyuki Tagawa(actor) Abrar Alvi(writer), Liam Neeson(actor)**
The second most profitable genre was the drama category with a maximum of 1328 million almost half that of 'Action, Crime, Drama'.
The third profitable genre was "Comedy, Drama, Romance' with a maximum ROI of 1047 million. It is not an expected result since the average rating for these movies is a 7.9. Genres such as Family and Animation returned negative ROIs but this could be due to the fact that the sample size for these genres was very small compared to the rest.

Areas of Further Study

This analysis was mainly based on measuring successful movies based on ROI but in reality there are many other compounding factors that affect the movie industry and these should be explored before Microsoft makes the big leap. These factors include but are not limited to: actor's role, the uncontrollability of the industry, influence of social trends on movie viewership, marketing especially online trailers, effects of piracy on the movie industry and critiques reception.

Recommendations

Filmmaking is both a commercial and artistic venture. With the revolution in visual F/X in the movie industry, it is expected that Action, Crime movies would be the ones with the highest ROI since they heavily make use of this technique. Microsoft is among the industry leaders in the tech industry from advanced cloud and AI tools to build their own digital solutions, it will be a wise move to produce Action, Crime, Drama movies with a production budget of around 250 million. Data has shown higher production costs more than triples the worldwide ROI. I would therefore recommend heavy financial aspect in areas such as special visual effects such as Pyrotechnics, prosthetic makeup, animatronics, and live-action weather elements. Movies such as Avatar, Avengers performed well because of their special effects. Since these are usually technology intensive and Microsoft has the capacity.

Microsoft should also consider opening a production studio as we recognize the role the major Hollywood studios have in shaping the movie industry today. The studios generate marketing and publicity.

The film-making process is a machine with many running parts. If you want to successfully make a film, you will need a great film production team and therefore Microsoft should consider working with the top writers, directors, composers and actors in this genre.