# **Introduction**

When venturing into a new thing, a business has to conduct research to assess their target audience, understand the existing markets, if they are there, needs of the target market, make an estimate budget, source of funding and predict revenues among others. Microsoft wants to venture into film creation, and are using data from different databases to come up with an informed strategy for their new venture.

# **Background**

The first step involves opening the already provided data, to see what data is available and what shape it takes. The next step involves cleaning the data. In data cleaning, it is important to check for any inconsistencies with the data. Some of the things to look for include missing data and duplicate data. When presented with missing data, there are different ways to deal with the missing data that vary from removing, replacing and keeping the data. When removing data, one may decide to drop rows or columns, depending on the extent of the missing data. When replacing data, this mostly works with continuous and not categorical day. The reason is because with continuous data, it is possible to get a common value, say the mean or average of the day, and use that to replace the missing values. You can choose to keep the missing data based on circumstance and this mostly gives an understanding of other factors such as unavailability of the data, which helps raise other questions that may give different perspective to the research.

# **Objectives**

There are several project objectives which include:

1. To identify the most popular journals to put Microsoft on the movie creation radar.
2. To make projections on what genres can generate the most revenue for Microsoft
3. To evaluate what studios Microsoft could use for a higher revenue.
4. To establish the studios that Microsoft can collaborate with for high quality films, or see who their greatest competitors would be.

# **Data cleaning**

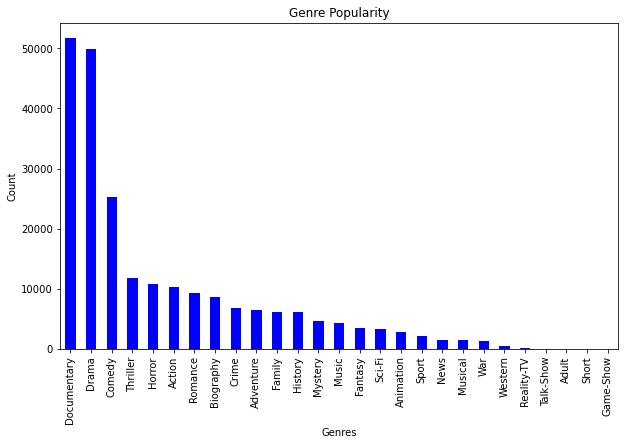
From the given data, there were several rows and columns with missing values from the different dataframes. With the first dataframe, movie basics, there were missing values in the original\_title, runtime\_minutes and genre columns. For the original\_title and genres, the missing values were quite few, with a low percentage of 0.01% and 3.7 percent, and it was easier to drop the rows with the missing values. However, with the runtime\_minutes column, there was 21.72% of missing data. Deleting the rows with the missing data would greatly affect the entire dataset. It is also one of the columns needed for analysis, and so it was impossible to drop the whole column entirely. For the runtime\_minutes, the best way to clean the missing values was to replace the missing values with the mean of the existing runtime\_minutes mean, or median.

Something else to look out for is duplicate values in the data. Duplicate values can greatly affect the data and outcome, so it is important to remove any duplicate values and entries before performing further executions. An example is when performing statistical measures such as the mean, median and standard deviations, the duplicates can affect the outcome, which will in turn affect the decision-making.

# **Data Analysis and Visualization**

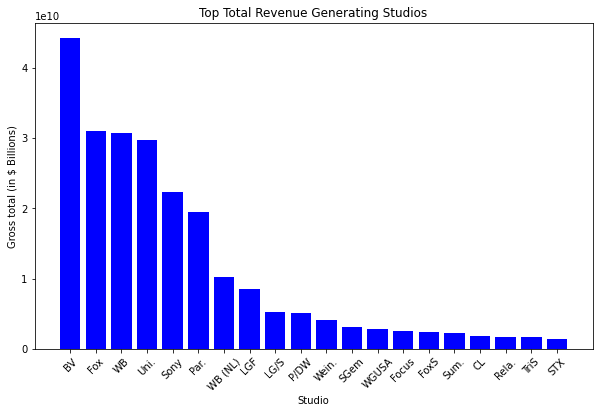
## **Genre popularity**

One movie can fall under one more than one genre. The first step, therefore is to break down all movies and the genres they fall in. That means that one movie can fall under more than one genre. The next step is getting the genre totals to get the popularity, that helps show which genres that movie creators are inclined to creating.



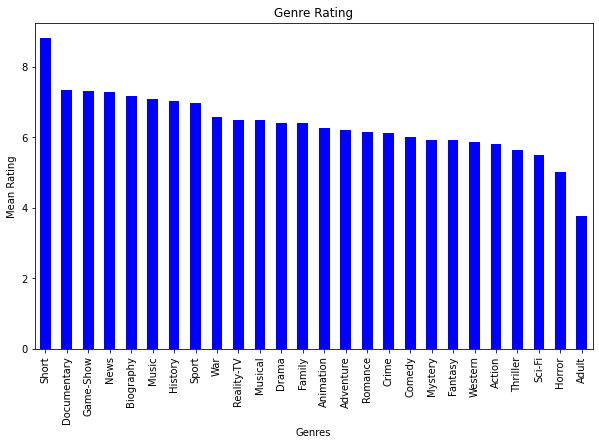
## **Top Total Revenue Generating Studios**

The plot is a bar graph that shows the top total (gross) generating studios. It is used to show how much revenue (gross) revenue each studio brings.



## **Genre Rating**

The genre rating is an analysis that is based on the cumulative mean ratings of films in different genres. Since, as earlier mentioned a film may fall in more than one genre, this is a cumulative mean, and is broken down to the ratings of the top 20. In this table, the shorts have the highest mean. The assumption is that since they are shorter films based on the runtime, more viewers will watch more short films in one sitting in comparison to watching the several movies in one sitting.



# **Recommendations**

In the beginning, Microsoft should not mostly focus on the revenue, but rather work towards being on the radar when it comes to film creation. They can achieve this by creating films that are in the already popular genre for a start. After a while, especially after they have gained popularity as a film creation firm, they can then focus on the less popular genres, research on what makes them popular and create a plan on how they can create films in these genres, but in have their films stand out to grow the genres’ popularity. After a while, they maybe recognized as the film firm that changed the content consumption of certain film genres.

The next step that Microsoft needs to take is to collaborate with the studios that generate high revenue. Initially, the primary reason for collaborating with these studios is to have quality films, and also from the viewers perspective, they may have a preference of watching films produced in certain studios. The collaboration will give them access to a larger audience and increase their popularity and visibility as a film creation. In the long-run, this also means more revenue for Microsoft.

Alternatively, if Microsoft would prefer to set up their own studio(s), they can gain insight from the high revenue studios to see what they can borrow from them, and what they can do differently to give their films a uniqueness from films produced in the already existing studios.

# **Assumptions and Further Research**

There were several assumptions while working on the data, one being that for the null values in the foreign gross column of the movie gross dataframe were for movies with only local sales. The assumption helps raise other questions such as why would movie creators choose to only sell their content locally? Do movie promoters have recommendations on movie genres that do better locally and why they don’t do so well in foreign areas? These, and more questions would give more insight to Microsoft so they make better investment choices.

Based on the genre popularity research, Microsoft needs to find out if the movie popularity, which is based on creators, whether it is viewer-led or creators-led. In a creators-led, it means that viewers are watching movies based on what the creator offers, but it does not necessarily mean that that’s what the viewers are interested in. They can, therefore, bridge the gap by creating viewer-led movies. Alternatively, if this is false, they can create their movies based on the existing trend.