# 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.

# Background

Microsoft sees all the big companies creating original video content and they want to get in on the fun. They have decided to create a new movie studio, but they don’t know anything about creating movies. You are charged with exploring what types of films are currently doing the best at the box office. You must then translate those findings into actionable insights that the head of Microsoft's new movie studio can use to help decide what type of films to create.

# Rationale/ Objectives

# There are several project objectives which include:

To identify the most popular journals to put Microsoft on the movie creation radar. (genre rating and genre popularity)

To make projections on what genres can generate the most revenue for Microsoft

To evaluate what studios Microsoft could use for a higher revenue.

To establish the studios that Microsoft can use in their movie creation, or see who their greatest competitors are.

# Method and Design

Method: How would you collect the data and why?

Design: What would the stimuli look like and why?

Procedure

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.

Data cleaning - fixing data inconsistencies and handling missing values

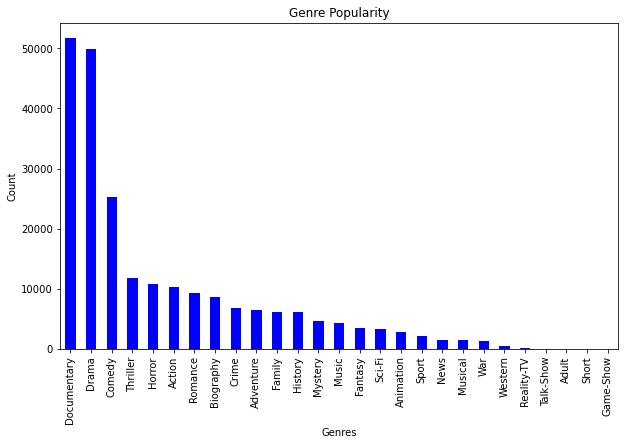
Feature engineering – selecting and dropping certain features and manipulating others to make them more meaningful than the raw data

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. On calculating the mean and median, the difference between the two was 0.2, and the choice on whether to use the mean or median scales down to preference.

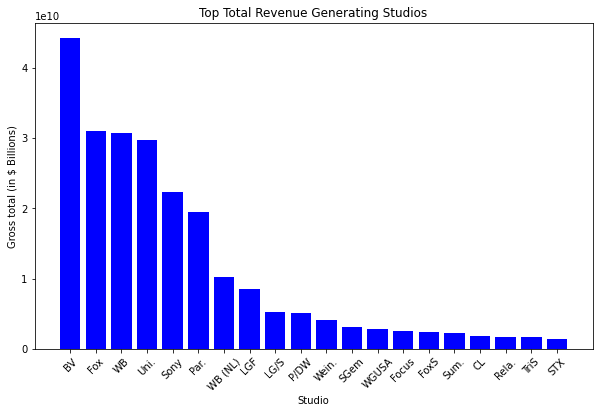
Data Analysis and Visualization

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.



The plot is a bar graph that shows the top total (gross) generating studios. There are two ways in which the information can help Microsoft. The first one is that Microsoft can target to use specific studios to increase their chances of making a higher revenue. Alternatively, they can find out what makes the specific studios stand out from the rest, then they can use the information they get to create a studio that is an equal competitor, or an even better studio.



# Recommendation

The first one is that Microsoft can target to use specific studios to increase their chances of making a higher revenue. Alternatively, they can find out what makes the specific studios stand out from the rest, then they can use the information they get to create a studio that is an equal competitor, or an even better studio.

Assumptions

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