

## Project Overview

For this project, I will use exploratory data analysis to generate insights for a business stakeholder.

## Business Problem

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.

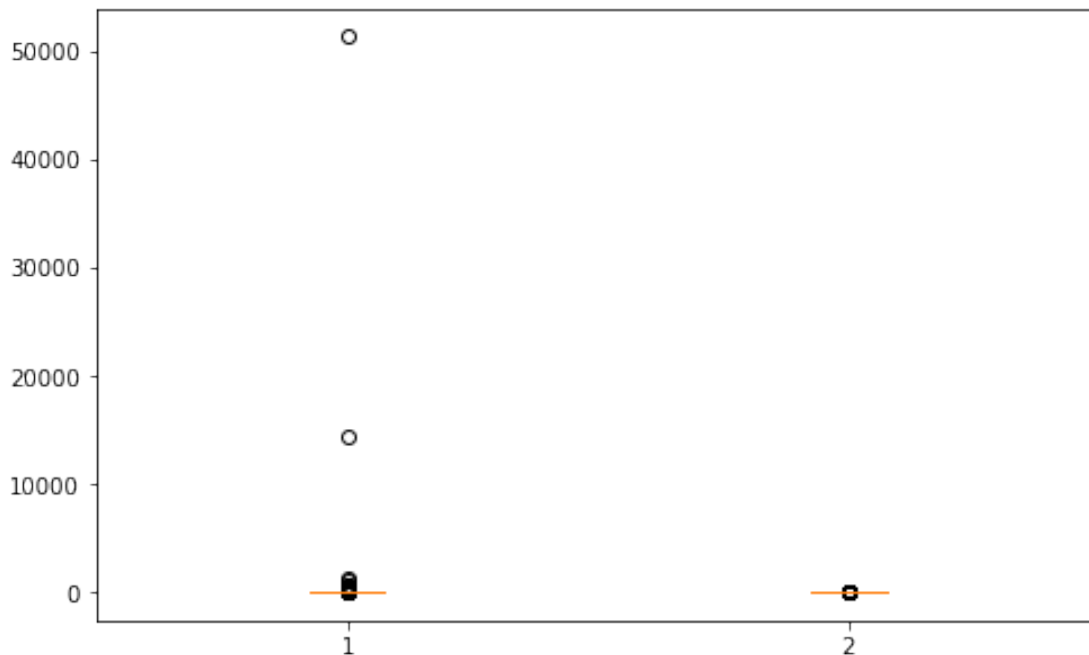
## Research Questions

1. *Is there a relationship between run-time in minutes and average movie rating?*
2. *Which movie genres are the most popular?*
3. *Which movies are currently doing well locally and internationally?*

The variables, averagerating and numvotes have the highest percentage of missing values. A half of the data for those two variables are missing. Dropping these columns is not an option since one of our objectives will not be attained without them.

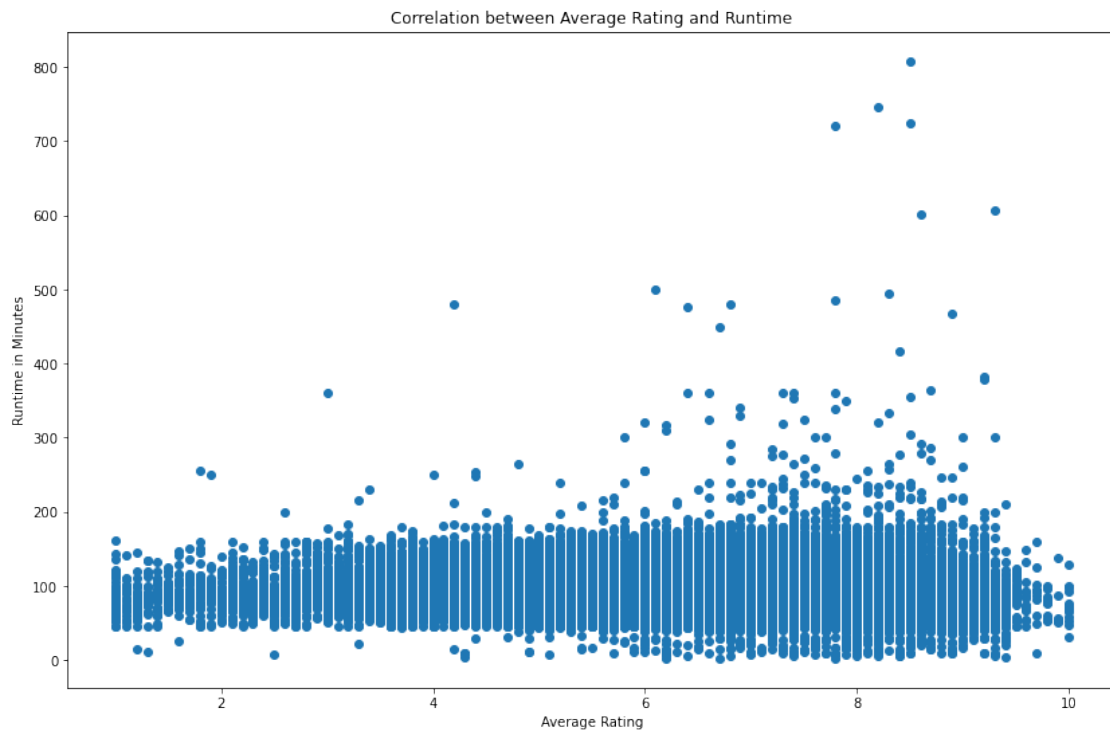
## 1. Relationship between run-time in minutes and average movie rating

Here, we want to establish the relationship between run-time in minutes and average movie rating. This will help us understand how the rating of a movie varies with the length (in minutes) of a movie. The outcome will help advise Microsoft whether to concentrate more on short duration movies or long duration movies.



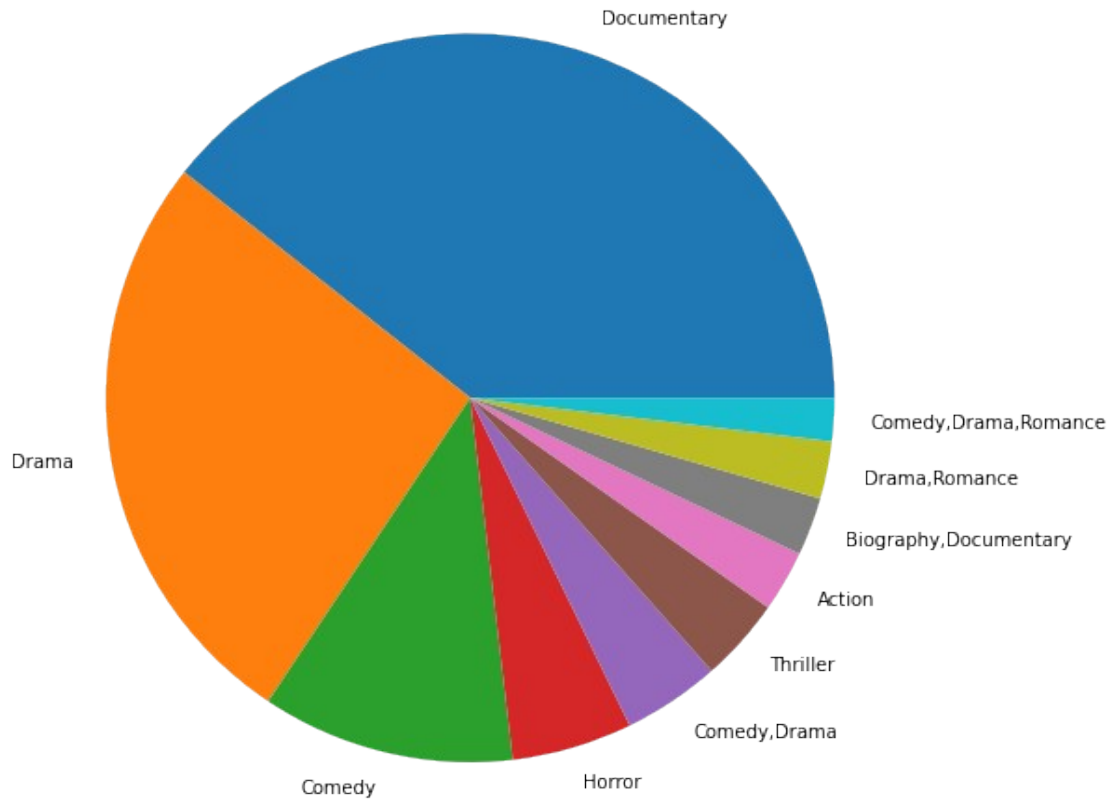
As can be seen from the boxplot diagram, there are two data points which seem to be outliers in the runtime\_minutes variable. These two points can be removed from the data before we proceed.

There is no meaningful linear association between the length of a movie and the rating. This means that the movies were given high or low rating regardless of its length.



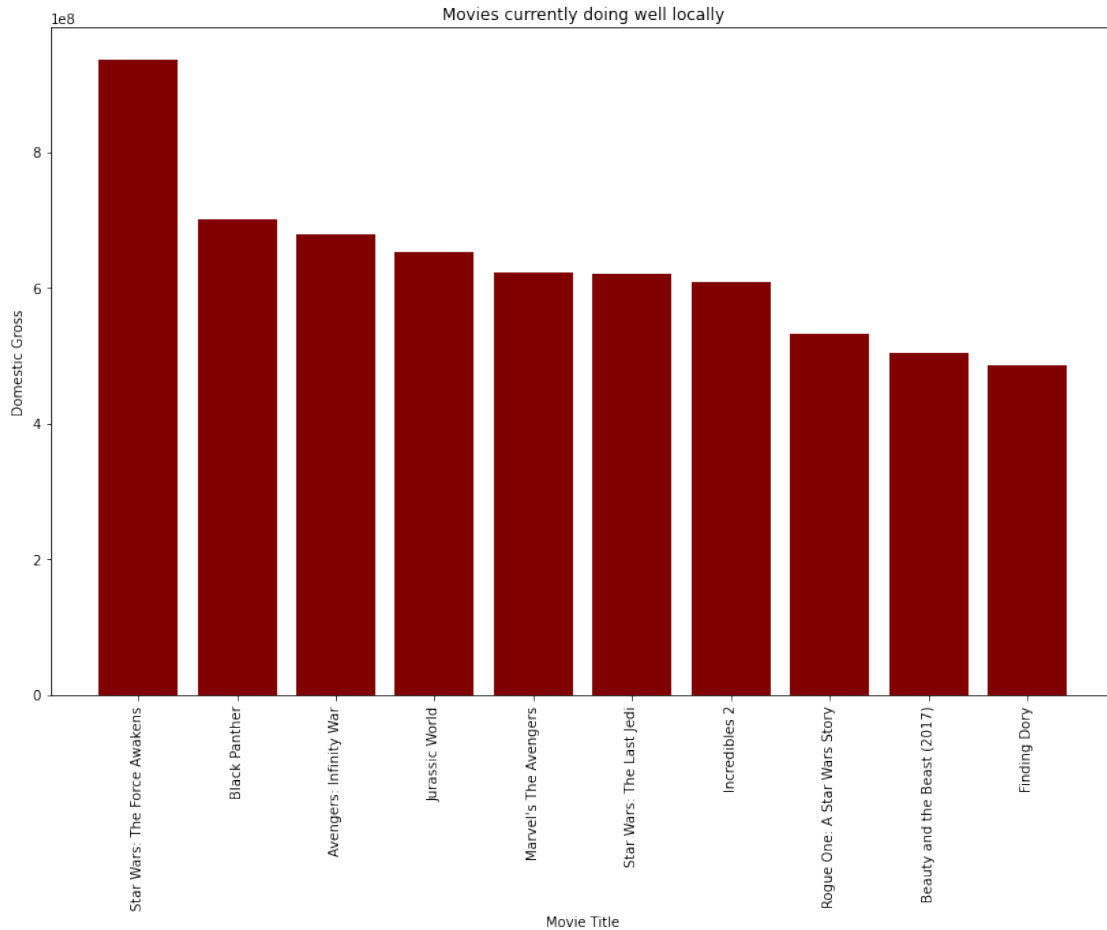
There is no significant relationship between the average rating and runtime

## 2. Most popular movie genres



The most popular genres are Documentary, Drama and Comedy movies.

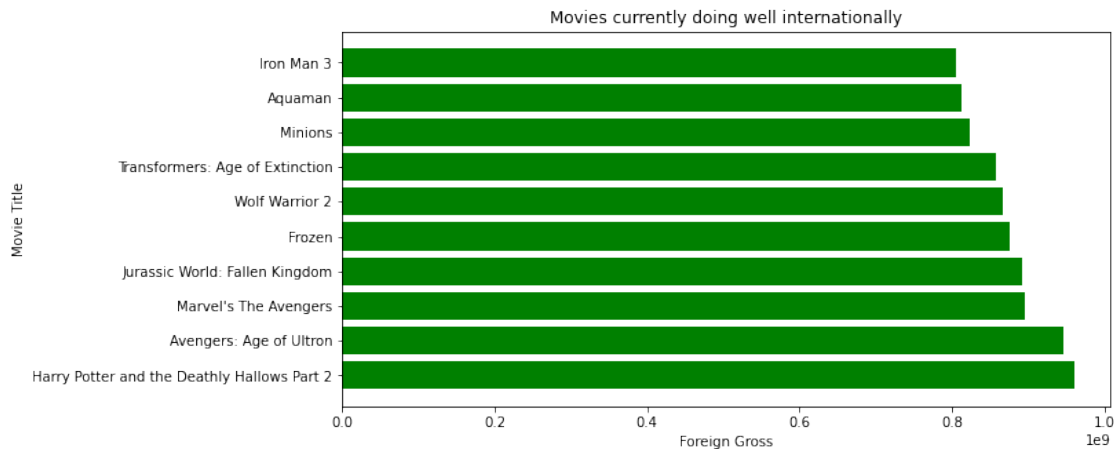
### 3. Movies that are currently doing well locally and internationally



We can see that the movie *Star Wars: The Force Awakens* made more money domestically. Surprisingly, the top ten movies that made more money domestically were all (except one) produced by BV studios which was discontinued a few years ago and now run by Walt Disney Pictures.

I tried sorting by `foreign_gross` and noticed it does not make sense. Later, I dicovered that the column `foreign_gross` needs to be cleaned because some vales have comas and others do not and the datatype is object.

After cleaning, the sorting makes sense now. We can see that *Harry Potter and the Deathly Hallows Part 2* made more money internationally. The movie was produced by Warner Bros. Entertainment Inc.



## Recommendations

1. Should Microsoft produce short movies or long movies? This question was investigated through seeking a relationship between run-time in minutes and average movie rating. There was no meaningful linear relation. Therefore, microsoft is free to produce both single movies and series movies without fear of losing customers.
2. Which Movie genres should Microsoft be concentrating on? The analysis revealed that the most popular genres are documentary, drama and comedy. Microsoft should give priority to the production of these genres. Also, the genres can be combined with other genres for more interesting outcomes.
3. Further investigation revealed that **Harry Potter and the Deathly Hallows Part 2** which made more money internationally and **Star Wars: The Force Awakens** which made more money locally were both produced on a budget of over 250 million USD. Microsoft should be ready to put more money by purchasing the best resources and hiring te best actors if it wants to produced best-selling movies.