Movie Truths or Tropes:
An Analysis of Award Winning Movies

By: Reynaldo Bonita Jr.

Introduction

Movies, apart from being a US$43B industry annually (Robb, 2018), is an integral part of culture that transcends generations, backgrounds and interests. It is reflective of culture while at the same time, influences it in return. Many can identify the films that move them or made them laugh and most associate these with the factors that can be seen on camera. However, aspects of the movie industry remain hidden from the front of the camera. This project explores characteristics of award-winning movie, on camera and off camera. What is the gender distribution among the directors, writers and producers? Do visual factors like the dominant colors and number of faces in a poster have a relationship with the quality and the popularity of the film?

Data Wrangling

The main data sources for this project are: a) the IMDb dataset which contains a comprehensive list of movies, TV series, documentaries and other video art forms (see https://www.imdb.com/interfaces/), b) the list of Academy Award Winning movies available in Wikipedia (refer to https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films), and c) the Open Movie Database (OMDb) API to retrieve movie posters and other (see http://www.omdbapi.com/)

The tools I used are:

- *awk* – for filtering large TSV files quickly
- Python – specifically *pandas* for reading and merging dataframes*, genderize* to predict the gender based on names, *requests* to make REST API requests to OMDb, *json* to parse JSON files, *opencv* to count the number of faces in an image
- Microsoft Excel – for further checking and wrangling
- Tableau Public – for visualization

Here is a summary of the steps I took to merge and wrangle a dataset for my visualization:

- IMDb publishes de-normalized datasets containing a separate TSV file for the major objects. For my analysis, I was mainly concerned of the following: titles with basic attributes, movie crew, and ratings.
- The *titles* TSV file contains 5,808,063 records of movies, documentaries, TV series and other videos forms. It would take too long to load in R or Python. To make it faster, I used *awk* via the command line to limit the file to movies and I ended up with 1,249,233 records. The list was still relatively large so I made a subset of movies on 1927, the year the Academy Award began, and onwards. I ended up with a list of 560,116 movies. This then was loaded to a data frame. From this, I made a list of about 5,000 of the most recent films in English.
- With the details from this IMDb movies data frame, I populated the details of the list of Oscar winners from Wikipedia. It contains 1,298 award winning movies from 1927, or about 91 years of the Academy Award. Among the details I retrieved were the IMDb ID, year, genre, director's name, IMDb rating, runtime, etc. It was challenging to match them by name because half of the movie titles from the Wikipedia list does not match exactly the titles in the IMDb database. I managed to match about 53% of the records, or about 698 while I queried the rest via the OMDb API. I limited my use of the OMDb API because it has a daily limit of 1,000 requests.
- I retrieved the count of awards and nominations from the OMDb API. They were concatenated inside one field containing the Oscar wins, other awards and nominations. Through a combination of filtering and converting text to columns, I separated Oscar wins and I classified the rest as other wins.

I grouped all Oscar nominations and other nominations from other awards as just nominations. I tried retrieving rating from other websites like Rotten Tomatoes and Metacritic. However, only 875 and 523 ratings are available respectively. IMDb

- Next, the movie posters of the award winning movies were retrieved and saved locally using the OMDb API. These were processed to retrieve the number of faces in the poster.
- I made a list of the actors, directors, writers and producers of the award winning movies. There was no available gender in the IMDb database so I used Genderize.io API to get the probability of gender based on the name.

Data Checking

Below is a summary of the checks performed on the key attributes of the dataset. The imdbID, imdbRating, title and genre have no missing or erroneous values. I thoroughly checked the values of the attributes I plan to visualize.

Attribute	Missing	Erroneous	Fix and Remarks
imdbID	0	0	
imdbRating	0	0	
title	0	0	
genre	0	0	
year	0	6	The *year* values are correct so I just deleted the special characters
director	23	0	There were 4 values found by searching online. These are short videos or documentaries with no listed director. I left them blank since they are only about 1.8% of the data set.
runtime	3	0	Manually searched for the values online

I did a spot check on the results of the facial recognition algorithm. I noticed that it has a difficulty in identifying side profiles or partially covered faces in the images. For the gender predictions, genderize tend to classify Alexis and Jamie as male (0.52% and 53% probability respectively) and Casey as female (56%) even if these names are common gender-neutral names.

Data Exploration

This section details the trends and insights found in the data set. This section explores the relationships of various movie attributes and how they vary over time. I compared the characteristics of award winning movies versus the larger corpus of movies. I focused on making a subset of movies which contains movies that were given recognition by the The Academy of Motion Picture Arts and Sciences, or more affectionately called The Academy. I chose the Academy award because it is regarded as the most prestigious in the industry. The Academy, is composed of about 6,000 actors, directors, producers and other movie industry professionals who nominate and vote their peers for recognition. (Littlejohn, 2019) Directors vote the Best Director, actors vote for Best Actor and Best Actress etc.

Popular vs Award Winning Genres

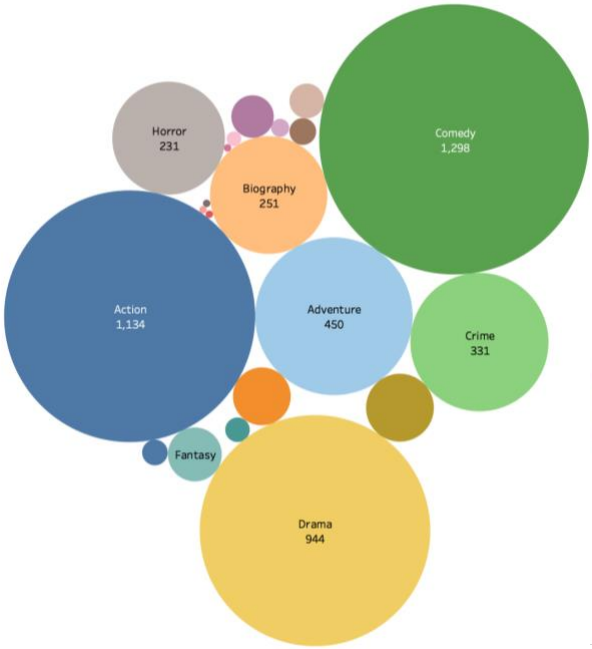

Figure 2 Popular Movie Genres

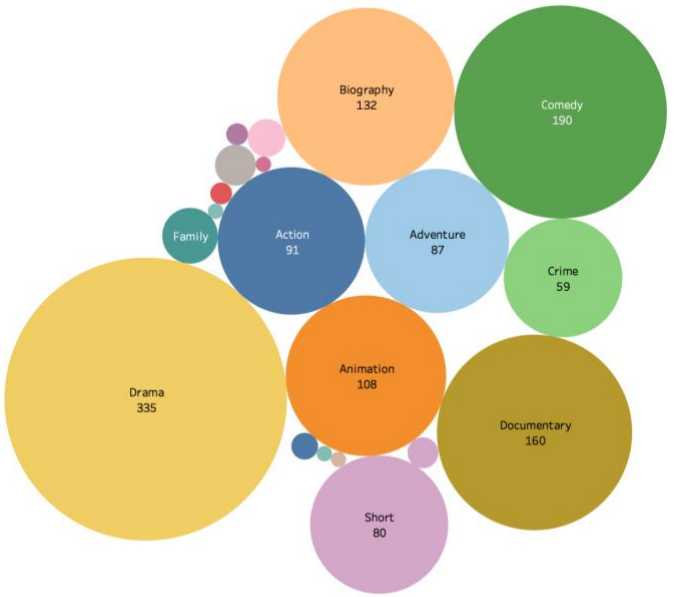

Figure 1 Award Winning Movies per Genre

Comedy, Action, and Drama are the top three popular genres for the past 9 decades. However, the distribution is quite different for Oscar winning movies. The Academy prefers Drama and Comedy movies. It seems logical as an actor and director can show more of their skill. Best actors and actresses shine in these genres. However, what surprised me is that the number of awards to Drama films are more than twice of any genre. And it is more than four times compared to those awarded to Action films.

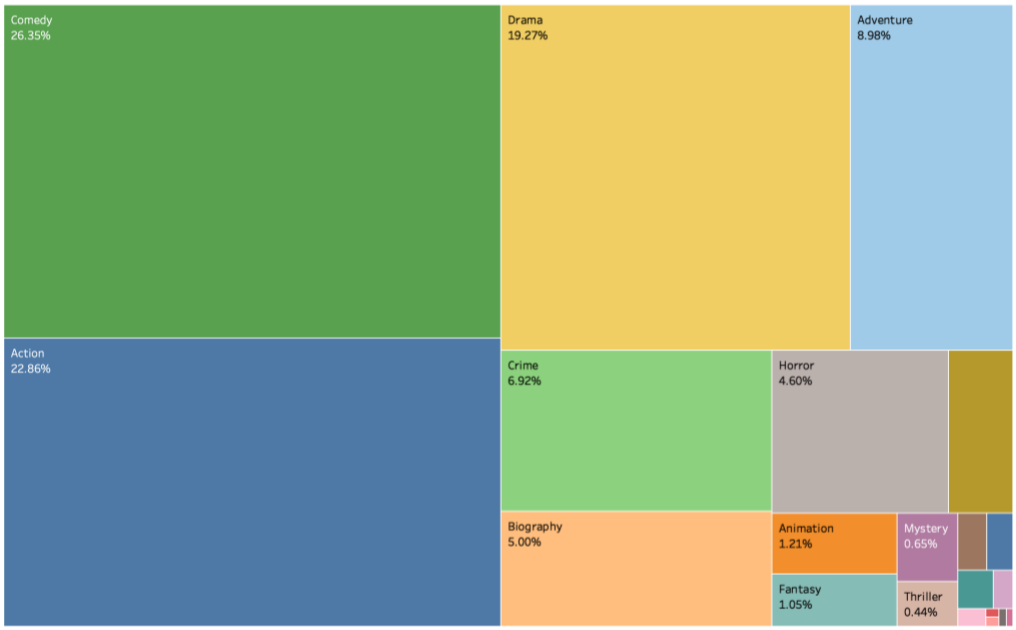

Figure 3 Popular Movie Genres

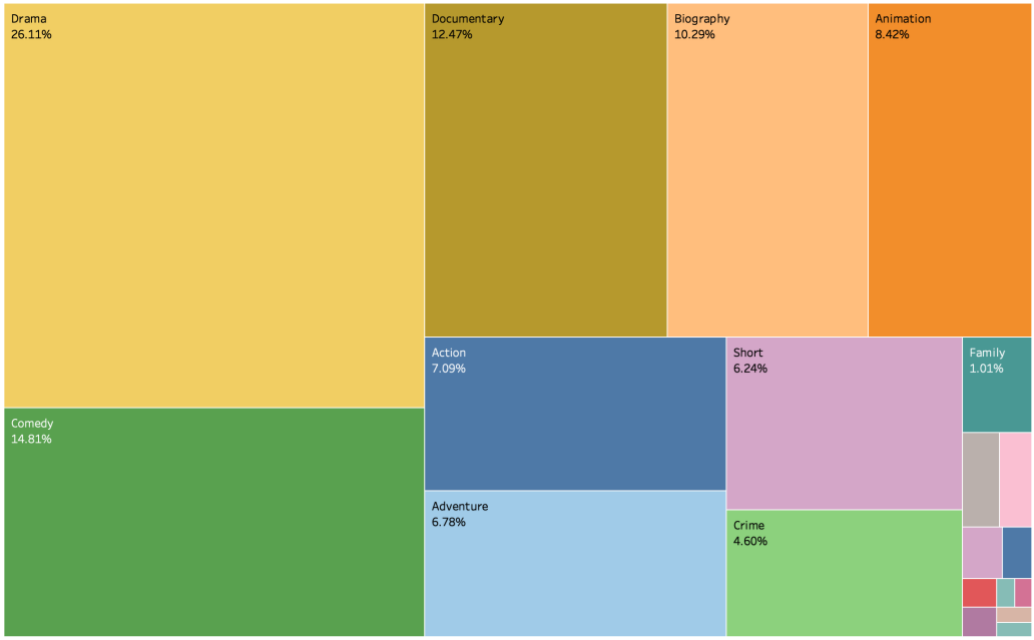

Figure 4 Award Winning Movies per Genre

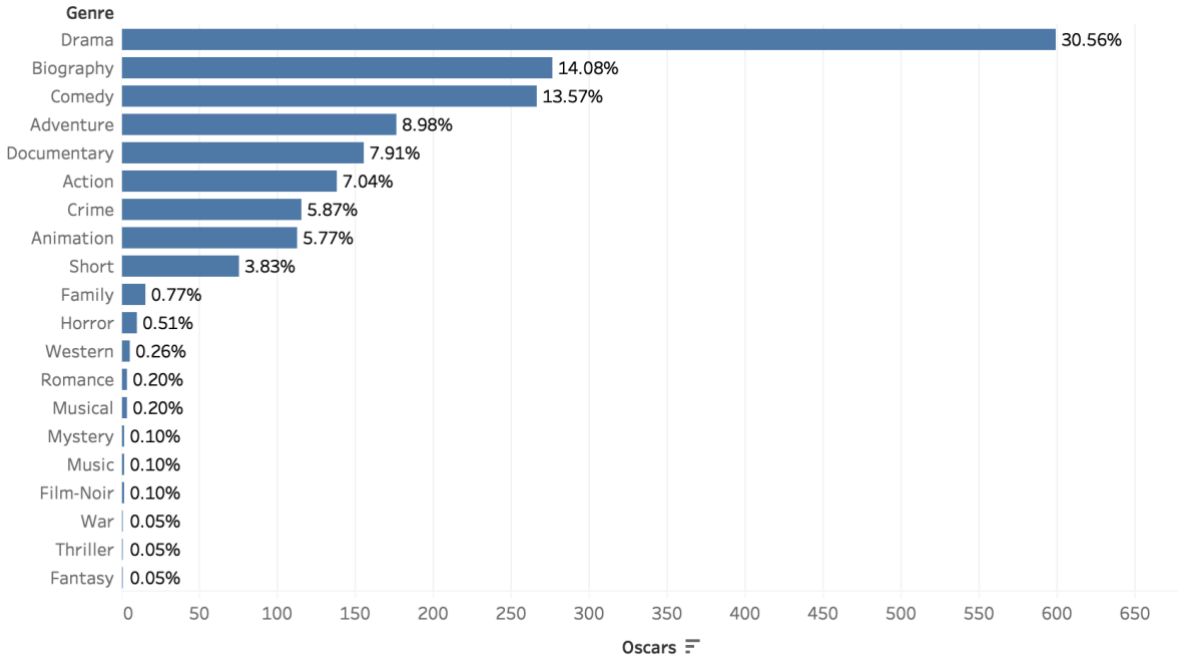

Figure 5 Awards per Genre

The Academy has a consistent preference for awarding Drama, Comedy and Documentary films since it began in 1927.

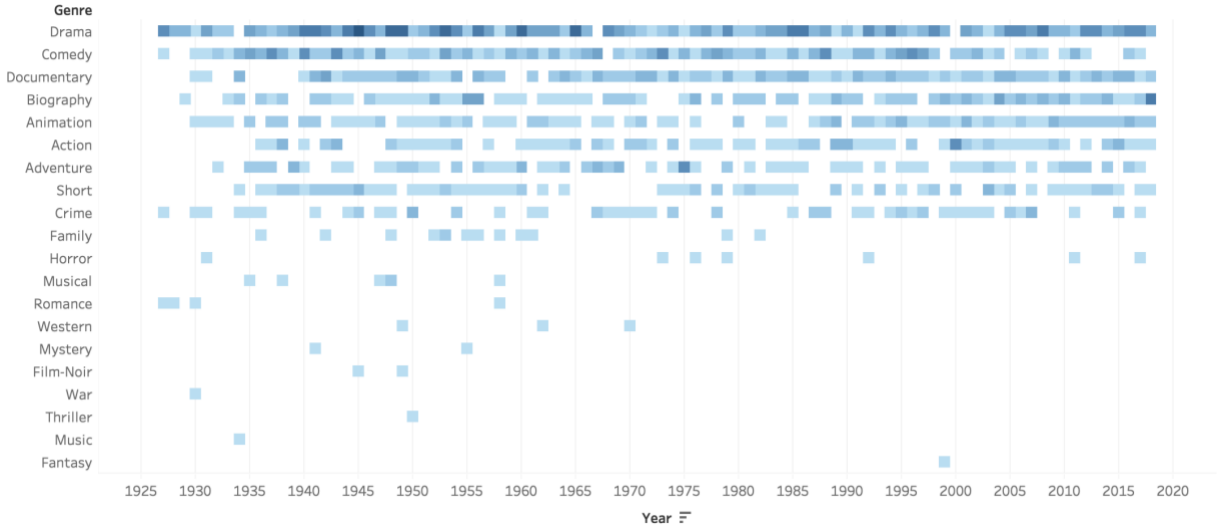

Figure 6 Awards per Genre over Time

Drama films overtake every other move genre in bagging Oscars since the 1920s. About one out of three Oscar awards are given to Drama films compared to one out of 13 for Action films. Movie professionals question why, for example, superhero films rarely win awards (Jones, 2019). Despite displaying excellent directorial, acting and technical skills, almost none of the action superhero films make it to the Oscars. With the exception of Star Wars and The Lord of the Rings Trilogy, almost no other adventure, fantasy or action film get nominated for nor awarded an Oscar. Critics argue that these films are our modern-day myths, mirroring human nature and the problems we face in society now (Jones, 2019).

Nonetheless, there is an increase in the number of action and adventure films produced since the 2000s. Despite the lack of recognition from the Academy, films from these genres are among the highest grossing. With action and adventure films earning more than triple compared to Drama films.

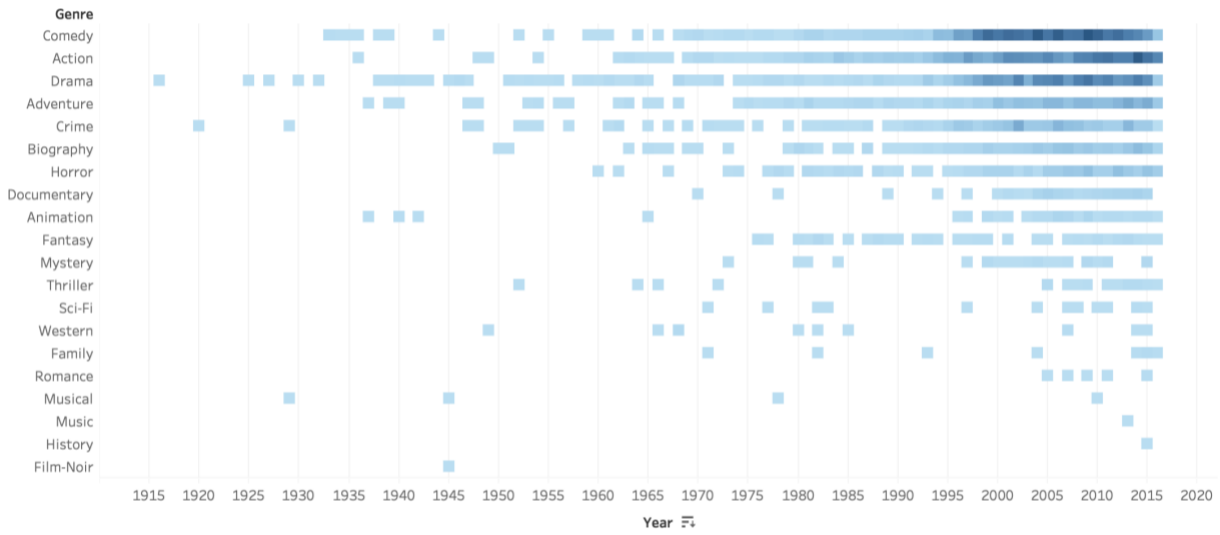

Figure 7 Movies per Genre over Time

Action, adventure and animation movies have a higher average box office earning than most genres.

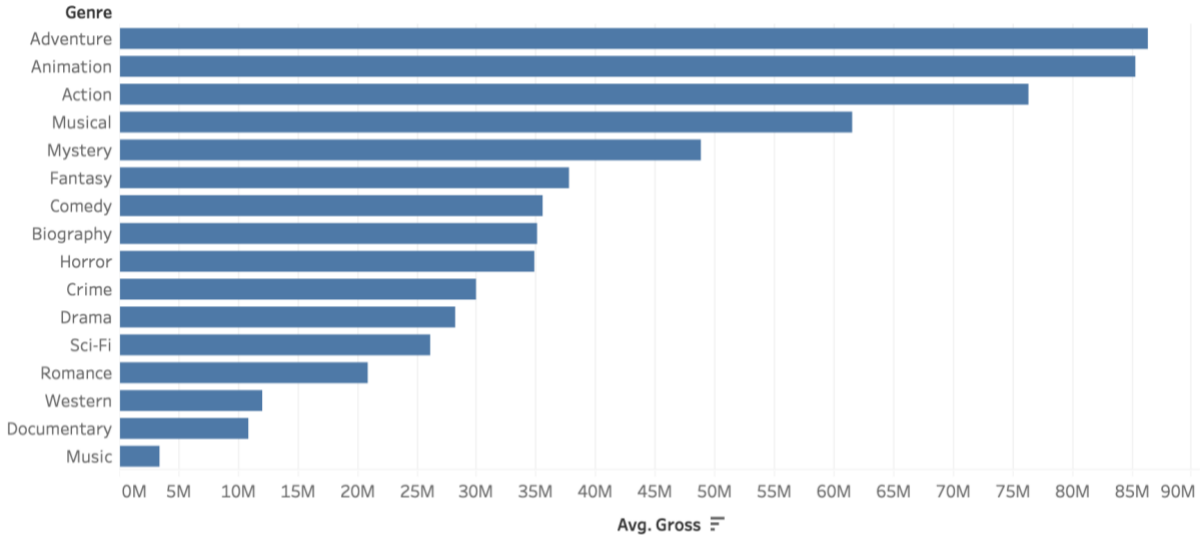

Figure 8 Average Box Office Earning per Genre

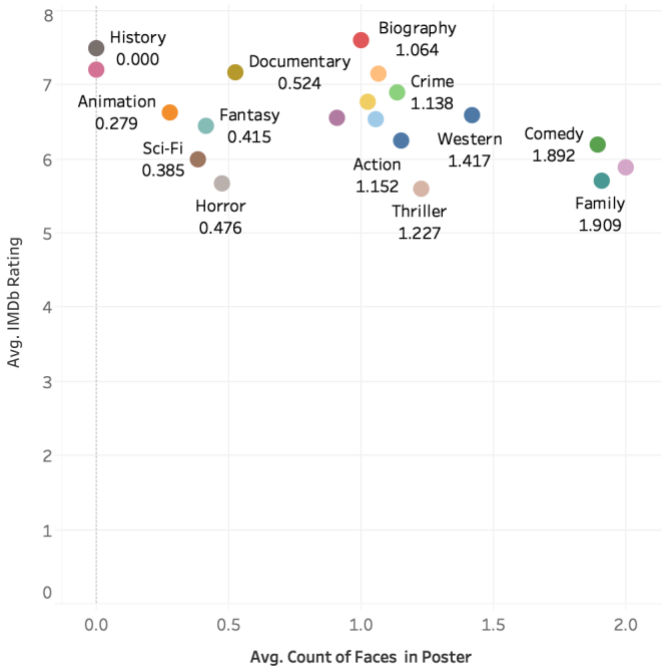

Figure 9 Average Count of Faces in the Movie Poster

Hoping to confirm stereotypes in visual design of movie posters, I compared the number of faces found in the poster and averaged them per movie genre. Predictably, Romance movies have an average of two faces. Comedy and Family movies have 1.8 and 1.9 faces on average respectively. Biography films usually feature only one person and the 1.06 average face count confirms that. Fantasy and Animation films usually feature non-human characters which may explain the low face count.

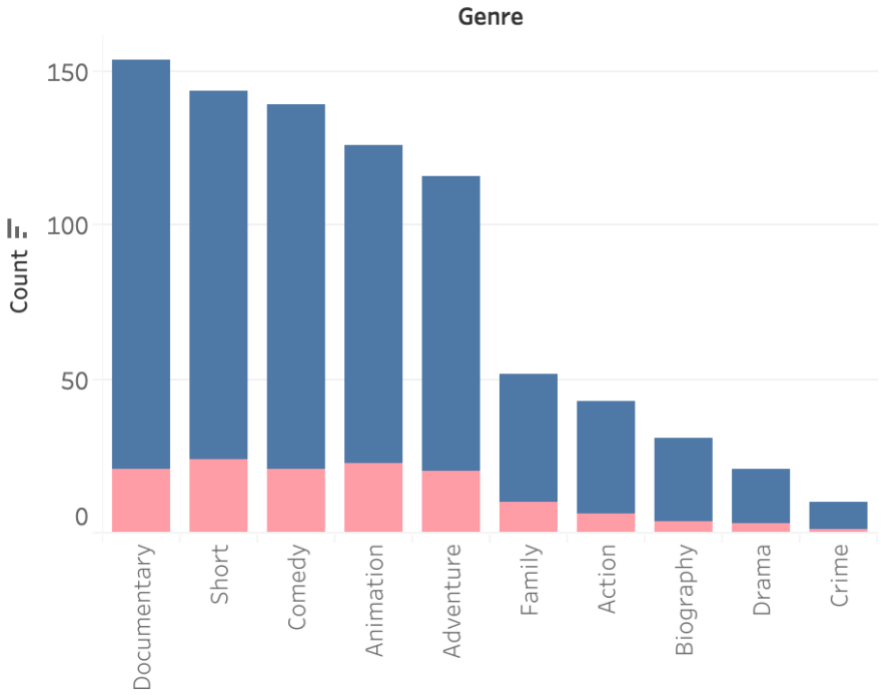

Figure 10 Gender Parity among Film Professionals

The film industry is male-dominated, even for the professionals behind the camera like producers, screenplay staff and writers.

Prolific and Award Winning Directors

This section explores data regarding the people on camera and those behind the camera. The graphs below show the most prolific directors and those who made the most films that received awards from the Academy.

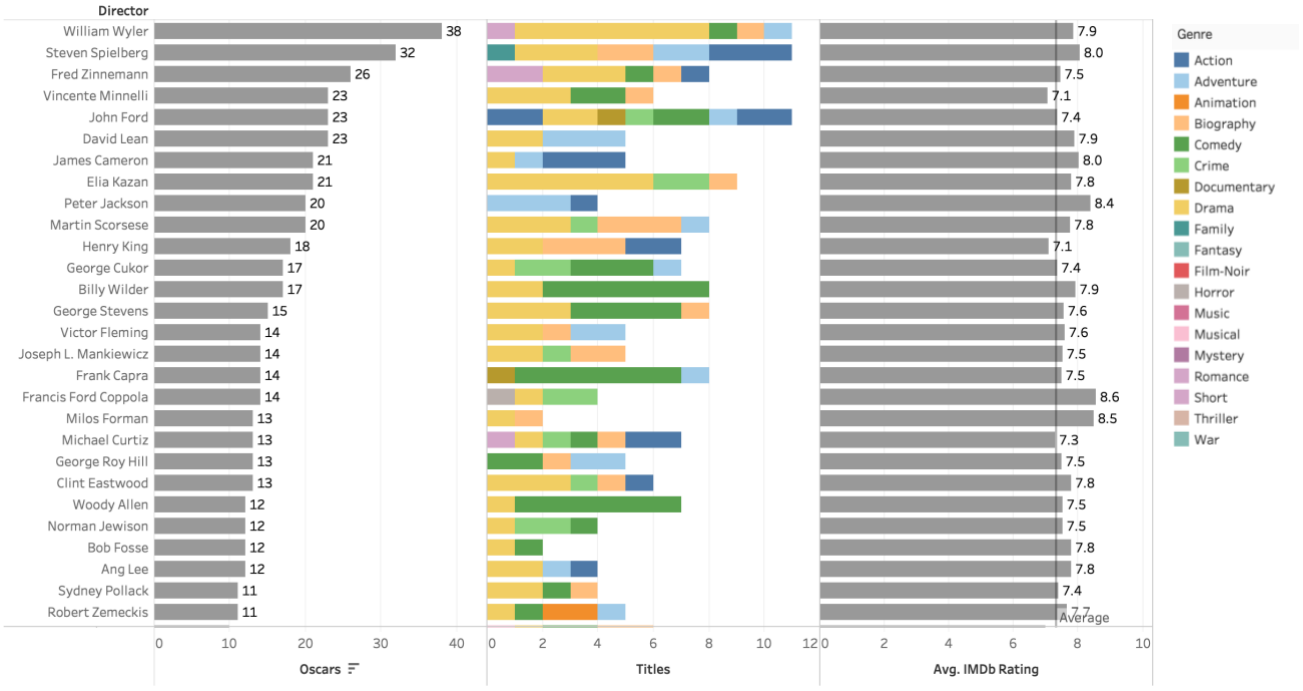

Figure 11 Directors of Award Winning Films

Many prolific directors like Steven Spielberg and Martin Scorsese have movies which won several awards. Having a larger body of work from a variety of genres helps a director hone his craft, thus leading to higher quality films and eventually getting an award. Most directors of the award winning movies in the dataset diversify although there are a few who specialize on certain genres. Woody Allen, Billy Wilder and Frank Capra focus on making comedies while William Wyler and Elia Kazan focus on dramas. All the top award

winning directors though, are men. Further data and analysis is needed to see the effect of that in the various aspects of their work.

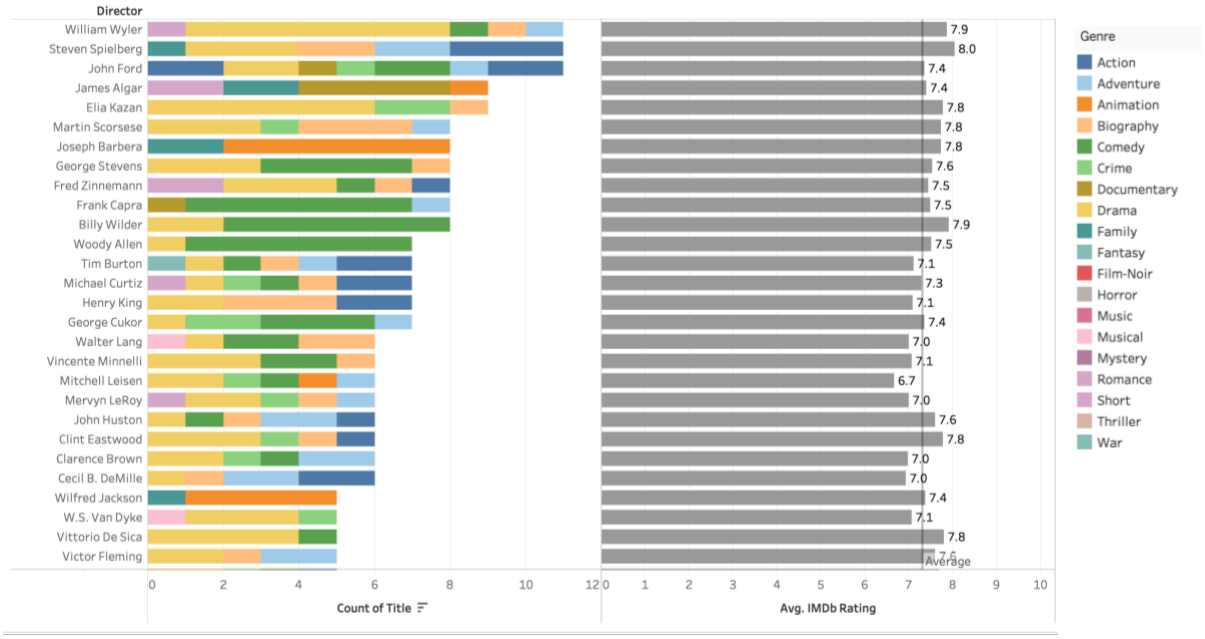

Figure 12 Award Winning Directors

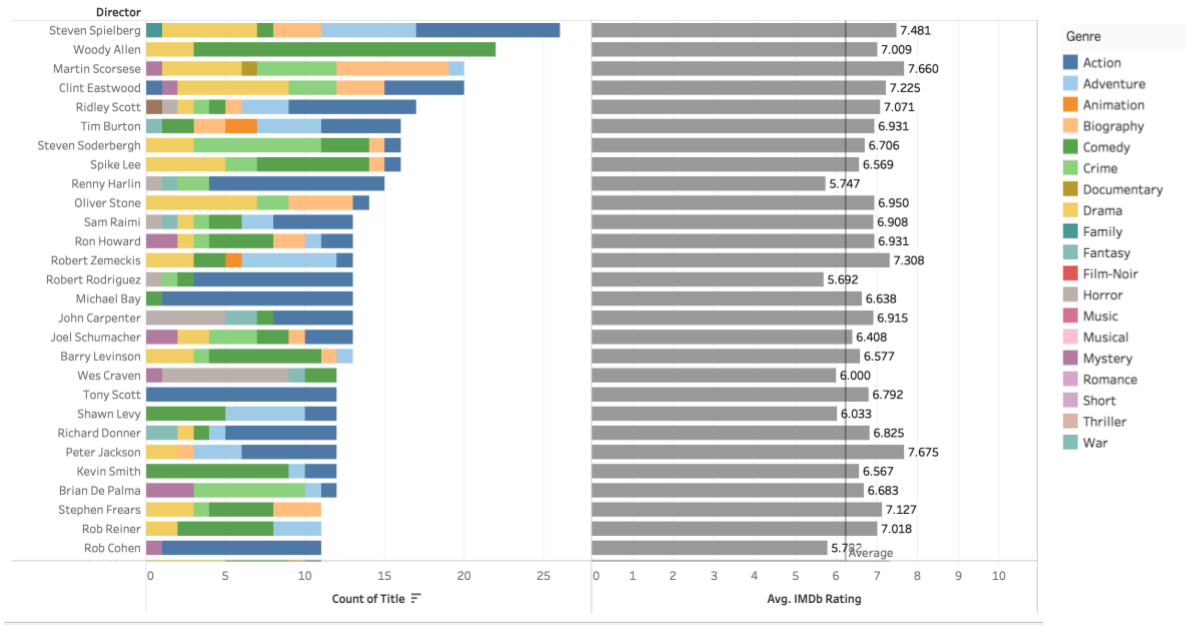

Figure 13 Most Prolific Directors

Conclusion

After exploring the dataset of the recent movies and the award winning ones, Drama and Comedy movies have a higher chance of being recognized in the Oscars while Action, Animation and Adventure films have a better change of earning in the box office on average. There was a rise in the number of while Action, Animation and Adventure films since the early 200s. Film is a male-dominated industry where all the most prolific and award winning directors are male. Furthermore, majority of the film professionals behind the camera like screenwriters, producers and writers are men.

Reflection

While doing this project, I felt enjoyment and pressure simultaneously. I would like to think that I enjoyed researching a topic I really liked, films. However, as with most realities of the world, it is multi-faceted and condensing its properties into attributes and data points seem inadequate. I felt the pressure of representing the realities of the industry as accurately as I can and to paint an adequate picture of some issues. As I was researching, I tried to find qualitative information that provide context to the industry. These are aspects of the industry that a mere movie database cannot provide. These insights, combined with logic and common sense, helped inform my decision on which attributes to explore and to attempt an explanation of what the data is showing. There are aspects of the issues that I do not have the data for, partly because it is difficult to access it at scale, like the ethnicity of the actors, writers, producers and directors.

Doing this project taught me several things. Firstly, as with any research endeavor, context is king. Before embarking on a data exploration program like this, I should perform a deeper research on the issues and factors that are involved. This should help me ask the right questions and determine what kind of data is needed to answer them. Furthermore, I should have processed the data during the proposal phase. In that case, I would know beforehand which leads and angles of the problem are worth pursuing. Lastly, although I had a notion that more than three quarters of the task will be dedicated to wrangling the data, I did not realize that I will spend several days gathering, merging, cleaning and correcting the dataset.

References

Jones, E. (2019). *Why don't superhero films win awards?*. *Bbc.com*. Retrieved 28 April 2019, from http://www.bbc.com/culture/story/20180306-why-dont-superhero-films-win-awards

Littlejohn, G. (2019). *What's the difference between the Golden Globes and the Oscars and why are the Academy Awards more prestigious?*. *The Sun*. Retrieved 28 April 2019, from https://www.thesun.co.uk/tvandshowbiz/2589715/oscars-golden-globes-difference-academy-awards-prestigious/

Robb, D. (2018). *U.S. Film Industry Topped $43 Billion In Revenue Last Year, Study Finds, But It's Not All Good News*. *Deadline*. Retrieved 27 April 2019, from https://deadline.com/2018/07/film-industry-revenue-2017-ibisworld-report-gloomy-box-office-1202425692/

Santos, N. (2017). *Python OpenCV: Face detection and counting*. *techtutorialsx*. Retrieved 27 April 2019, from https://techtutorialsx.com/2017/05/02/python-opencv-face-detection-and-counting/

Tiwari, S. (2019). *Face Recognition with Python, in Under 25 Lines of Code – Real Python*. *Realpython.com*. Retrieved 27 April 2019, from https://realpython.com/face-recognition-with-python/