Naveen Sanka | Data Visualization | May 26, 2017

Movie Data Analysis

Assistance for producers



Table of Contents

[Background of the Project: 2](#_Toc483584544)

[Claim for this project: 2](#_Toc483584545)

[How are we going to help? 3](#_Toc483584546)

[Data about cast/actors: 3](#_Toc483584547)

[Data of movies: 3](#_Toc483584548)

[Metadata of API Call: 3](#_Toc483584549)

[How did we Python Script to do Data Wrangling? 4](#_Toc483584550)

[Why we choose this visualization to represent data? 6](#_Toc483584551)

[Critique: 9](#_Toc483584552)

[Forward Path: 10](#_Toc483584553)

[Conclusion: 10](#_Toc483584554)

# **Background of the Project:**

Movie... Movie... It’s always a fascinating, glamorous and alluring field with lots of efforts, dedication and most importantly passion involved. I was always curious about the way filmmakers compress their imaginative thoughts into a reel and portray their vision and feelings on the screen. However, this field involves a lot of investment with lots of competition and high risk of failure. So, everyone in this field should be careful in choosing the right opportunity and delivering the best one can, this will get a better opportunity and fame.

In Filmmaking, though there are interesting scripts with the best director and great cast, success is not guaranteed and a single failure can make a producer insolvent. So, today let’s step into a producer's shoes and think from their point of view to invest in optimal way and make huge from Box Office.

As a producer, my priorities in movie making is success which can be denoted by awards and Box Office. To achieve that I have a formula that can increase the probability of getting a blockbuster.

Blockbuster = Paying the cast what they deserve +

Walkthrough lead actors past performance and movie success +

Perfect Director who can portray the script in the best possible way +

Right time to release a movie +

Picking the genre which can drag maximum viewers to theaters

# **Claim for this project:**

In this project, we are not having any claim to support or reject, we will help producers to meet their goals by using our mantra.

# **How are we going to help?**

We try to consider historical data of movie industry and get some insights which reflect our elements of the success formula. To achieve this, we need lots of data from various sources. We have tried to collect data and blend them to extract info we wanted to plug into the formula.

# **Data about cast/actors:**

We have extracted highest paid actor’s data from Statista and focused on year 2016 as it’s the most recent data[[1]](#footnote-1).

# **Data of movies:**

We found a source where they had a collect of 40,000+ imdb movie ids[[2]](#footnote-2).

These ids were used to extract data using imdb API[[3]](#footnote-3).

# **Metadata of API Call:**

Fields which we are going to receive while making API Call.

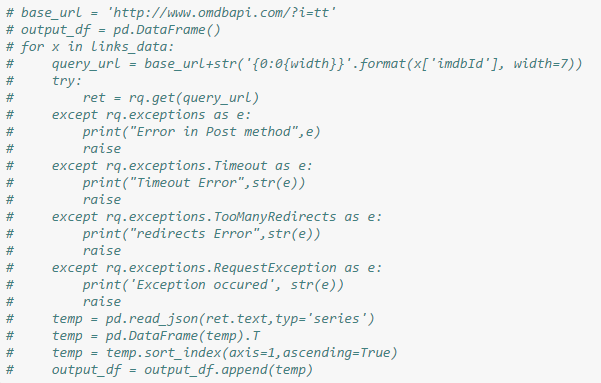
Actors, Awards, Box Office, Country, DVD, Director, Episode, Genre, Language, Metascore, Plot, Poster, Production, Rated, Ratings, Released, Response, Runtime, Season, Title, Type, Website, Writer, Year, imdbID, imdbRating, imdbVotes, seriesID, totalSeasons

For the convenience of project, we have exported the API response to a JSON file to stop API Abuse :) as it’s now a big issue.

# **How did we Python Script to do Data Wrangling?**

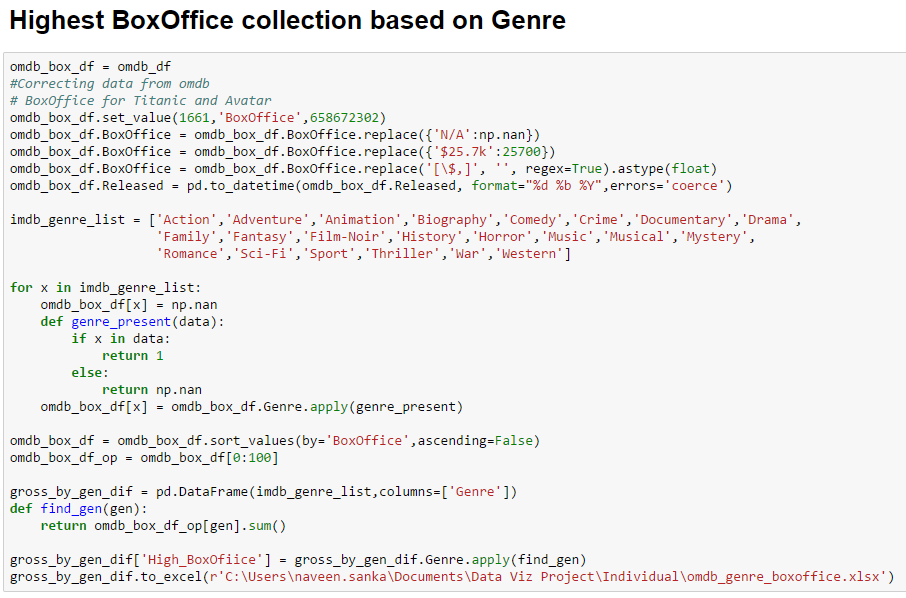
1. To Make API calls and extract data from imdb API:

This piece of code will iterate over all the movie ids and make call using POST request. We receive a JSON reply to this call and append the data, we repeat the same for all the 40,000+ movies and form a huge Pandas DataFrame.

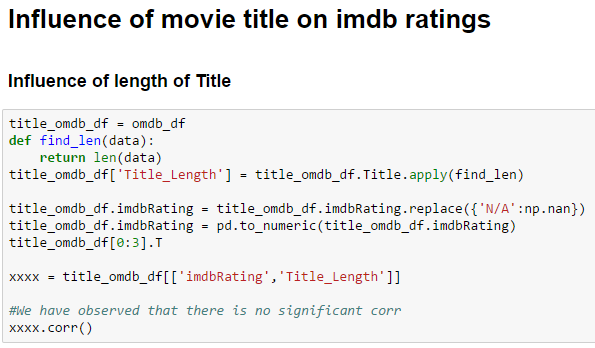


1. Deriving data to depict High box office based on Genre

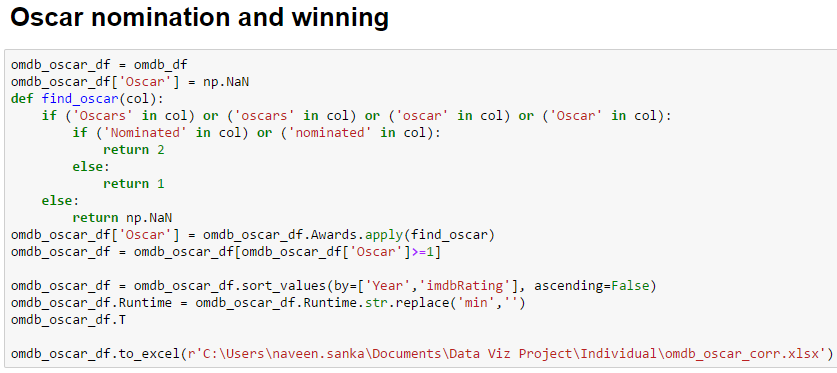
We have identified the list of genres and iterated over them to collect box office, we will be using top 100 box office to represent the viz for better understanding.



1. We are quite curious about the influence of movie length on the movie success and we found that there is no real correlation. So, producers don’t need to worry about the length of the title they are using.

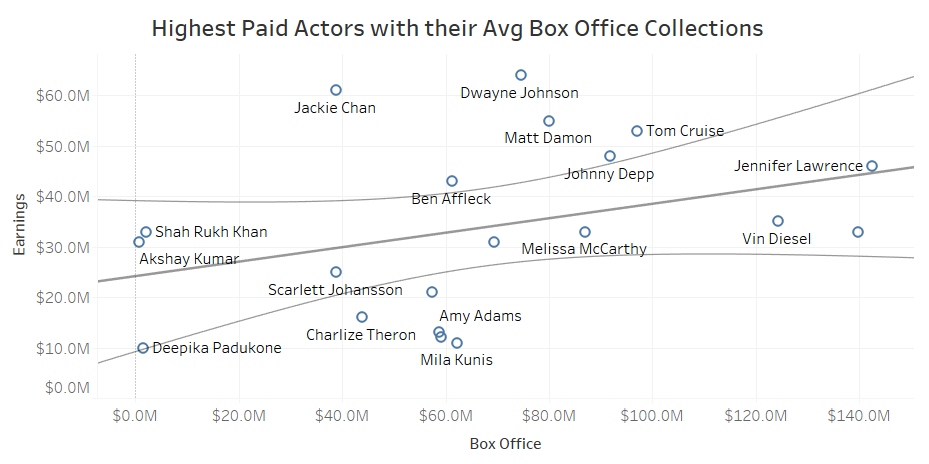


1. This code snippet is used to identify Oscar nominated and winning movies. We have enable flags to denote ‘1’ as won and ‘2’ as nominated.

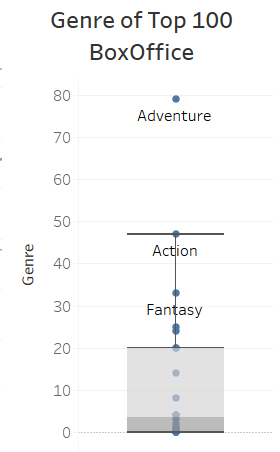


# **Why we choose this visualization to represent data?**

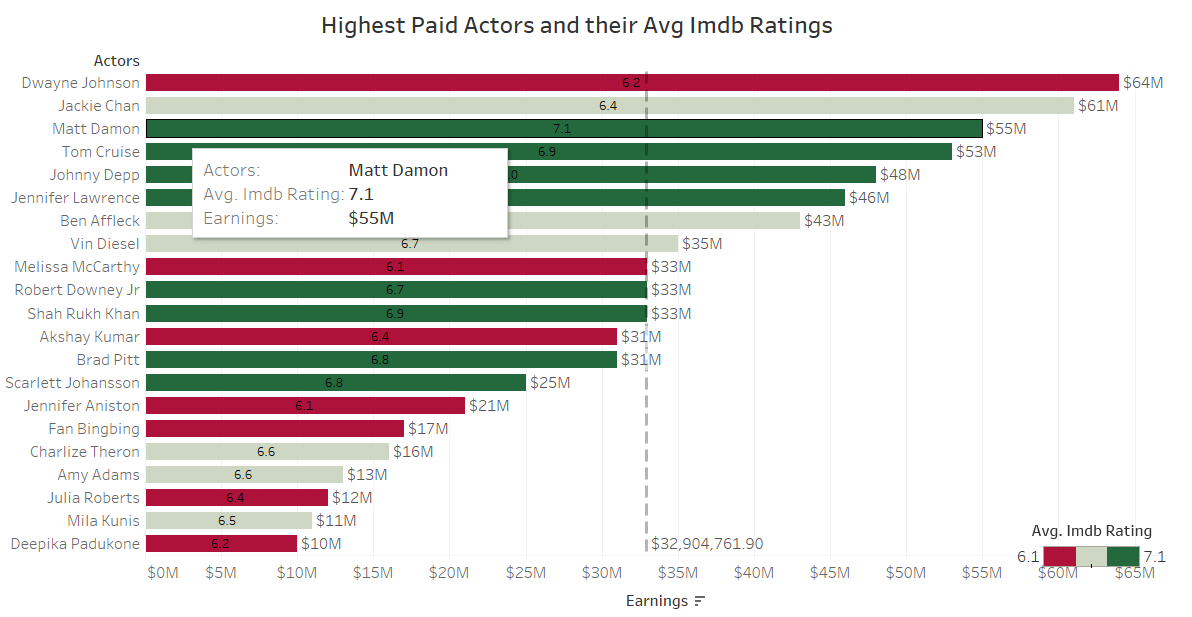
1. How much an actor make and how much their movie makes is a great way to evaluate the right amount we pay to the cast of a movie. So, we are using a scatter plot with a linear fit which can clearly show the actors who are highly paid with low box office such as Dwayne Johnson. Secondly, we can notice that women are underpaid though they contribute equally to some of the highly-paid actors.



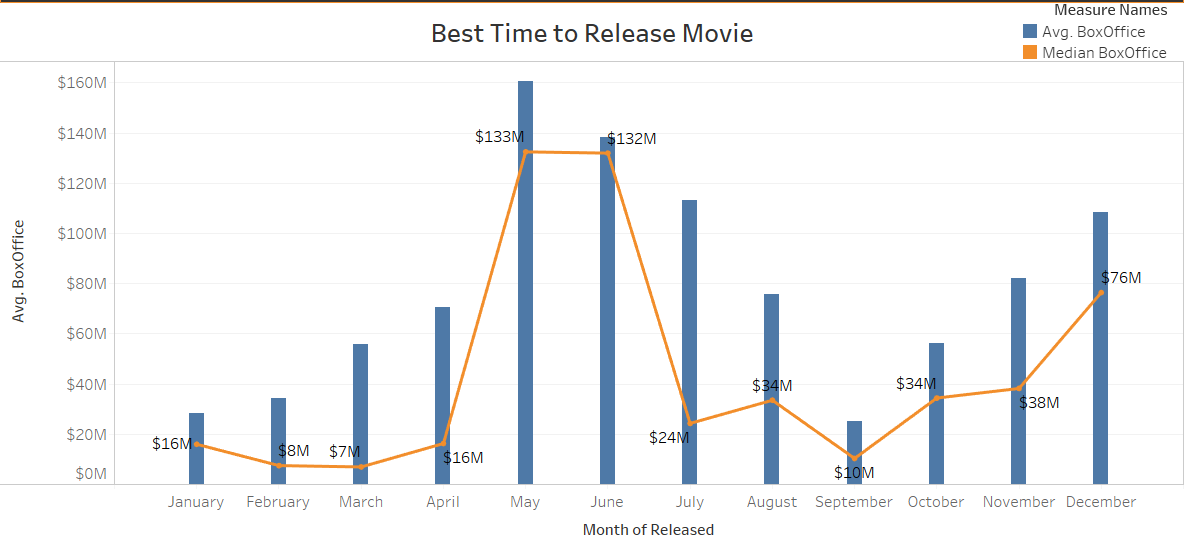
1. Genre plays a important role in producing movie, we have complied all time highest box office and filtered top 100 movies and found that a producer should choose adventure, and action as their first priority.



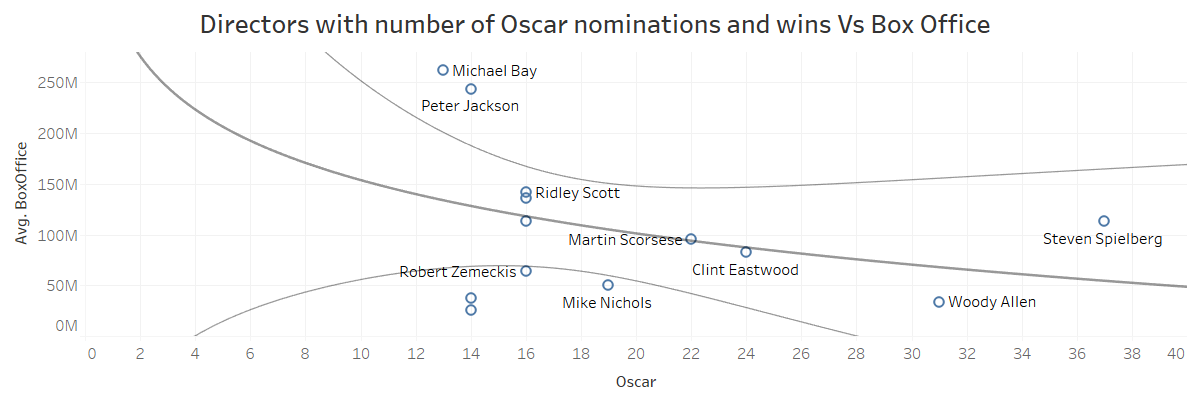
1. Inflation can mislead Box Office so let’s evaluate actors to their pay using imdb ratings, which can be considered a credible source of movie ratings. We have distributed ratings in three zones as the legend depicts. So, we can see that people on the top and boot are not having great movies that they worked in past.



1. Time to release a movie play a criticl role from commercial point of view. Summer time where students have done for the year and Christmas time are perfect for releases.



1. There's always a tradeoff to choose between producing a commercial with biggest Box Office and making an artistic movie which gets you awards like Oscar. But, choosing them with perfect balance can fetch you both the Oscar and Box Office. If a producer selects a director like Martin Scorsese they can counteract between Oscar and Box Office.



# **Critique:**

We will walk through the visualization that we can do it in a better way:

1. Directors with number of Oscar nominations and wins Vs Box Office:

In this viz there are few directors that are missing because, the movies they worked doesn’t have Box Office data and we are not able provide enough data for use for better choice.

1. Highest Paid Actors and their Avg Imdb Ratings:

We have applied this on lead actors and actress but it would be better if we can apply the same of the entire cast so that producer spend optimally. This is process is like what we have seen in Moneyball.

# **Forward Path:**

There’s lots of data that we can work on from the sources I have mentioned, due to time constraints we had to reduce the scope of the project. Things we can do in future with the same data:

1. Sentiment analysis of reviews of movie.
2. Influence of movie poster to its success.
3. If we can collect movie profits, Box Office and budget spent on a movie we can create a predictive model with cast, director and time to release as parameters.

# **Conclusion:**

Movie making is a passion but without enough monetary support one cannot produce a movie. On the other hand, when we invest in movies it’s better that we do our ground work before we start investing. The formula that we derived may be useful and reduce the risk of failure for producers and expect better returns from movies.

1. Actor: <https://www.statista.com/statistics/451516/best-paid-actors/>

   Actress: https://www.statista.com/statistics/594482/best-paid-actress/ [↑](#footnote-ref-1)
2. Source: <https://grouplens.org/datasets/movielens/> [↑](#footnote-ref-2)
3. API Link: https://www.omdbapi.com/ [↑](#footnote-ref-3)