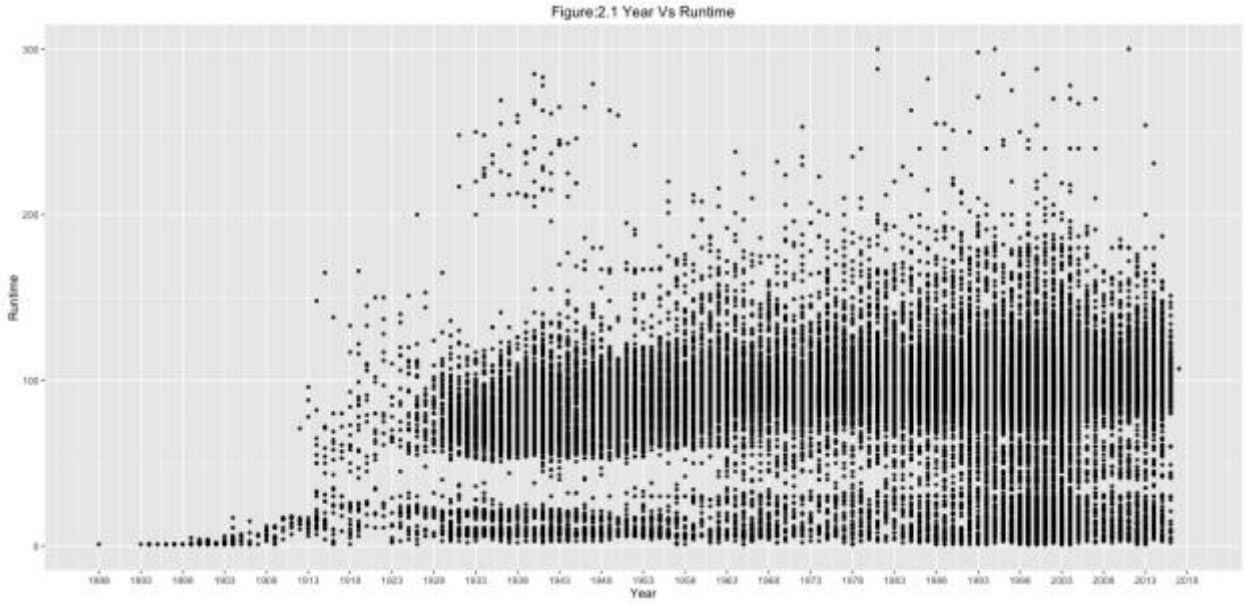
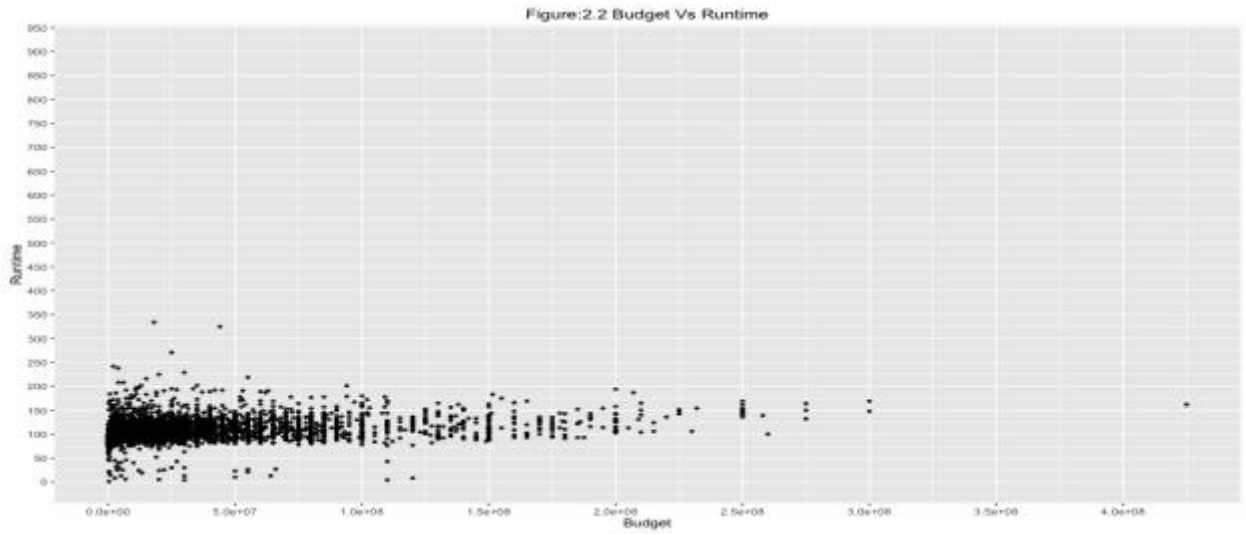
Project: 1
Name: Ajay Joshi

1. 798 rows has been removed.

2. Figure:2.1 Year Vs Runtime

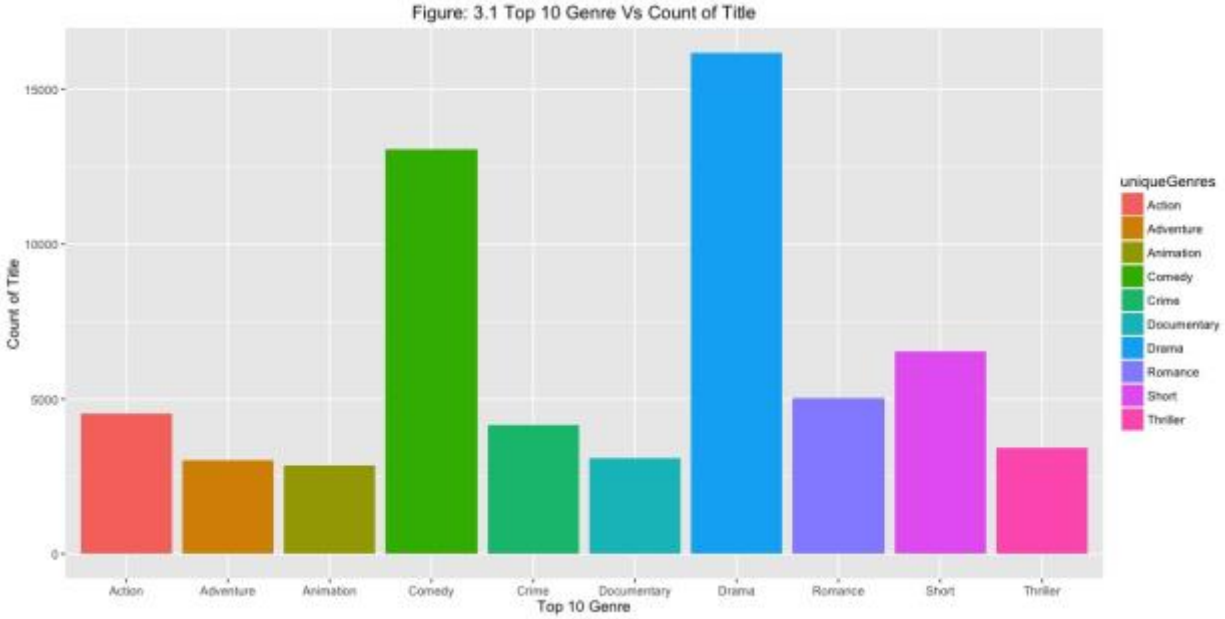


The figure 2.1 shows that the runtime values are the lowest from 1888 till around 1912 and there's gradual increase of runtime from a certain point (around 1925). some of the high runtime movies are made in the range of 2010-2018.

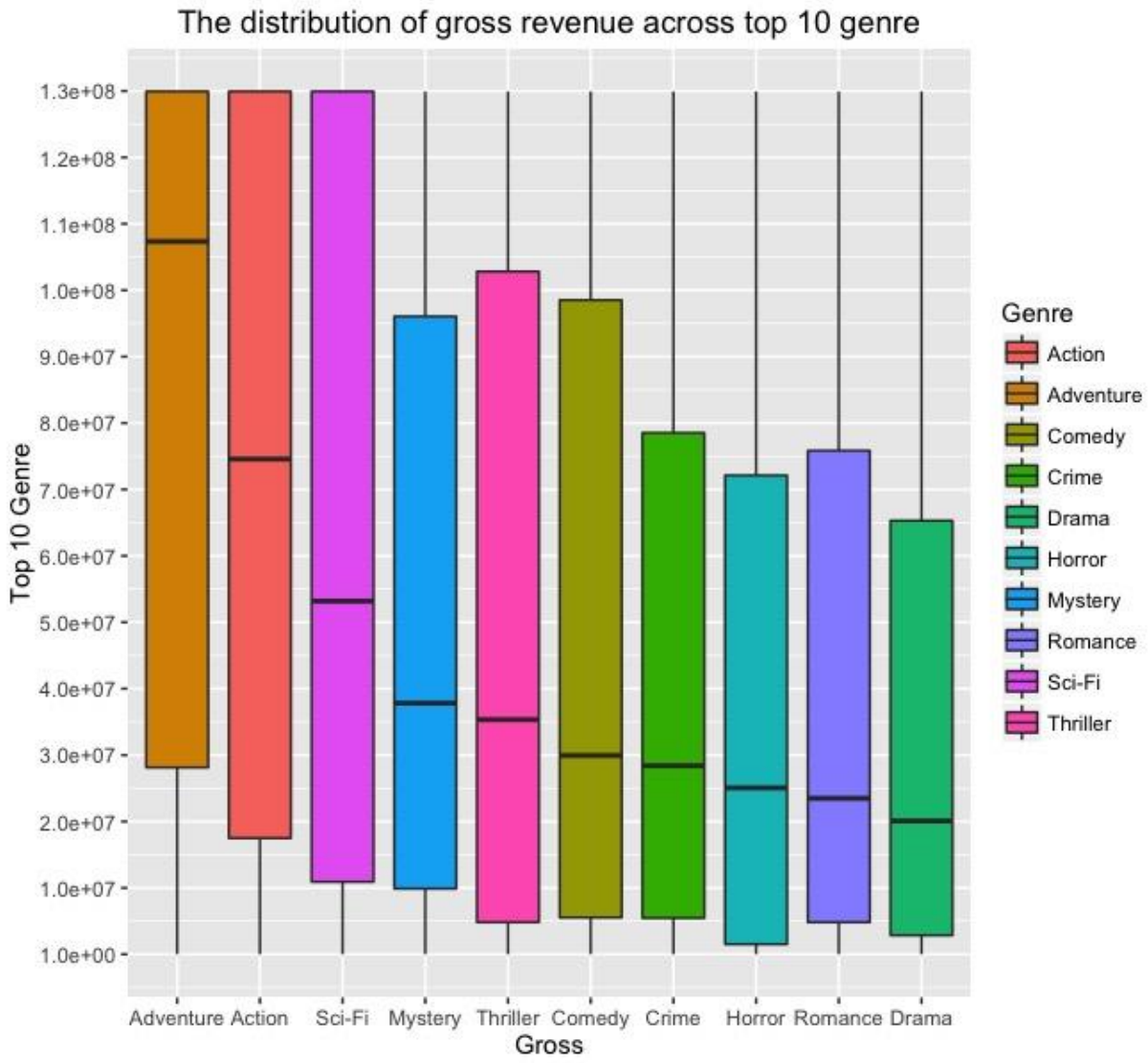


According to the figure 2.2, the runtime over the budget of the movie is relatively linear.

3. Figure: 3.1-Use bar plot [Top 10 Genre Vs Count of Title]



According to the figure 3.1, Drama and comedy movies have highest nuumber of title count wheareas animation, adventure, and documentary genred movies have lowest title count.

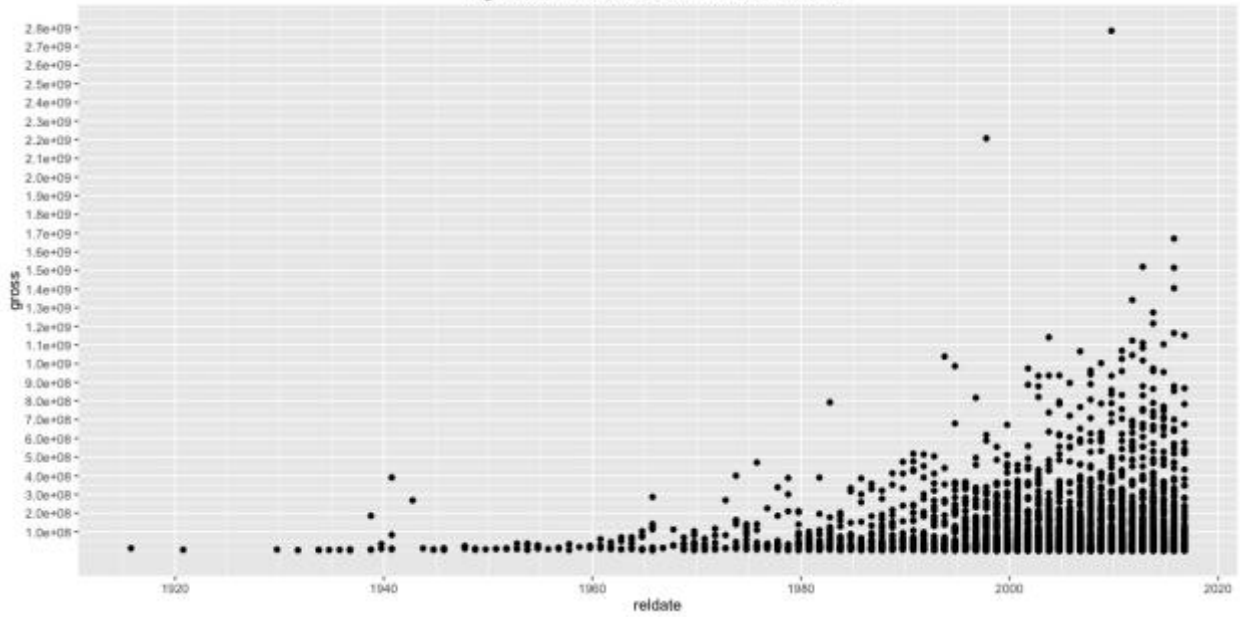


According to the figure 3.2, the highest median value is adventure genre and the lowest is drama. Action, adventure, and Sci-fi genres have approx. maximum values but the least mimimum values among the three is of Sci-fi.

4. I think, the year is the first released date and the released column is the release date for some region/country. Using “Date” or another data column will be more appropriate to analyze the mismatch. If there's the merge from two different sources based on title, its likely to see “title” with slightly different names displays on multiple rows but the same “Released” and “Year” but it doesnt happen on the data provided. Hence, there’s no discrepancy between "Year" and "Release".

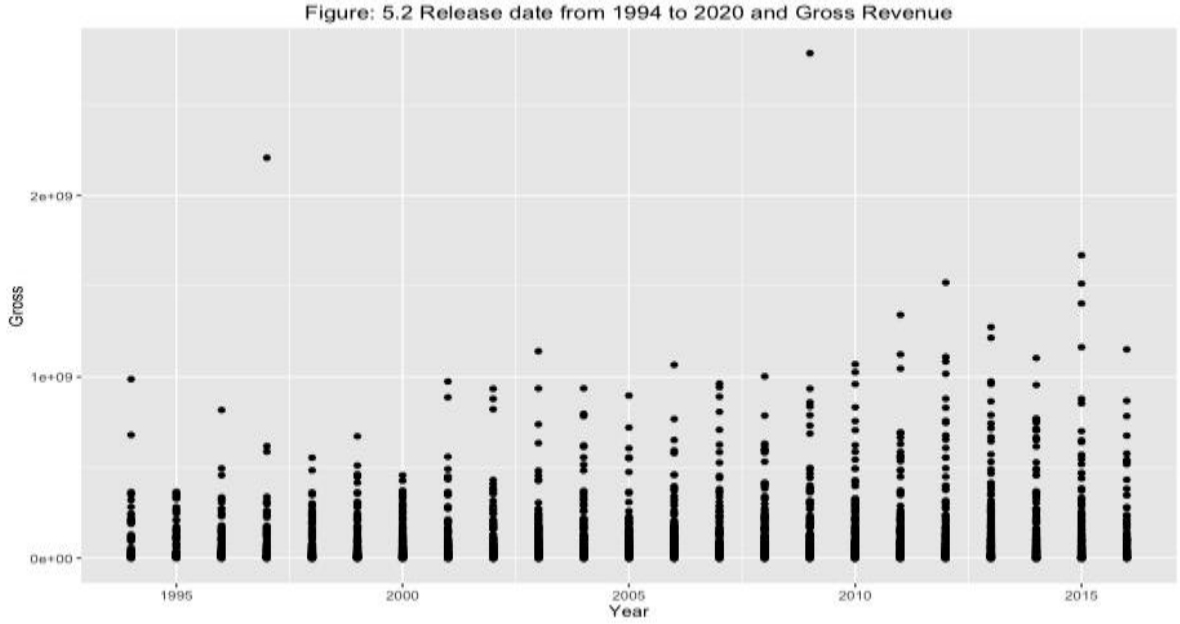
5.

Figure: 5.1 Release date and Gross Revenue



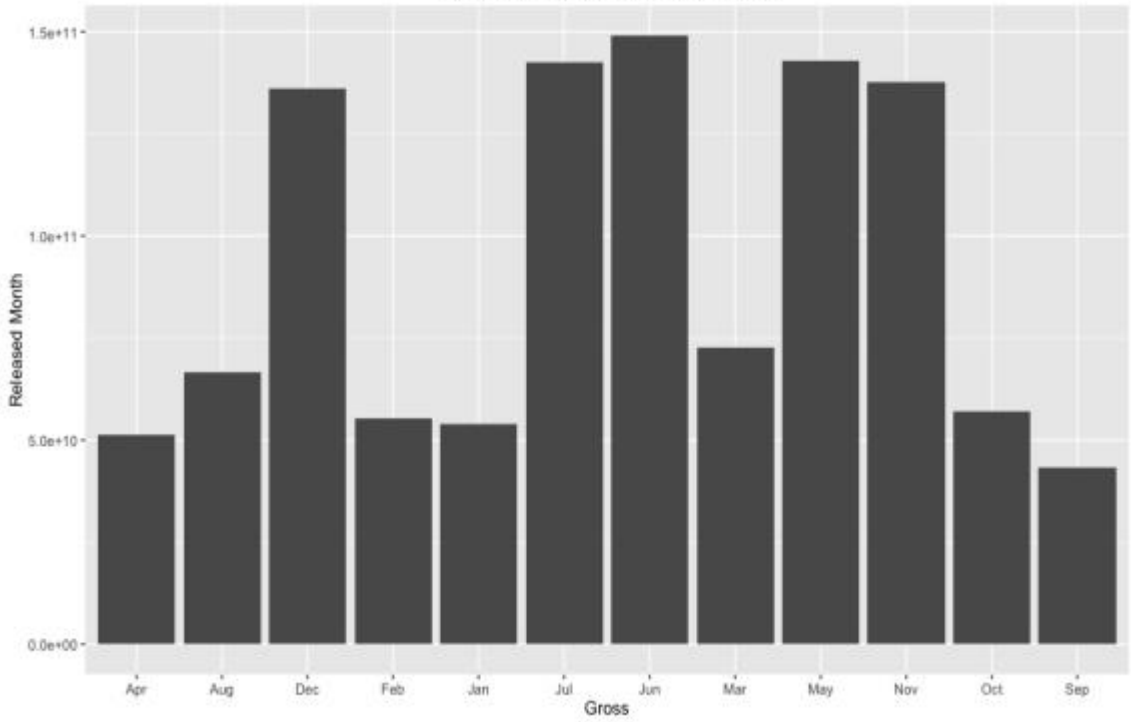
By looking at *figure 5.1* graph, roughly, we can see that the highest revenue movies released in the range of 2010 to 2020. After around 1980, there's gradual increase in movies' budget. The highest revenue movies were released in 2009. However, it is not clear enough to investigate the on what times of year are highest revenue movies released in. Figure 5.2 shows that the relations between the Gross revenue and year that is between 1994 and 2020. The reason I picked the year between 1994 and 2020 is because those years have most movies with released date and gross data. By looking at the *figure 5.2*, we can conclude that the highest revenue movies were usually released in and around 2015.

Figure: 5.2 Release date from 1994 to 2020 and Gross Revenue



By looking at the figure 5.2, we can conclude that the highest revenue movies were usually released in and around 2015.

Figure 5.3: Released Month Vs Gross



By looking at the figure 5.3, the best month to release a movie is in the month of Jun, May, July and the worst months are sep, april, jan and feb.

Figure 5.4: Released Month Vs Gross Vs Genre

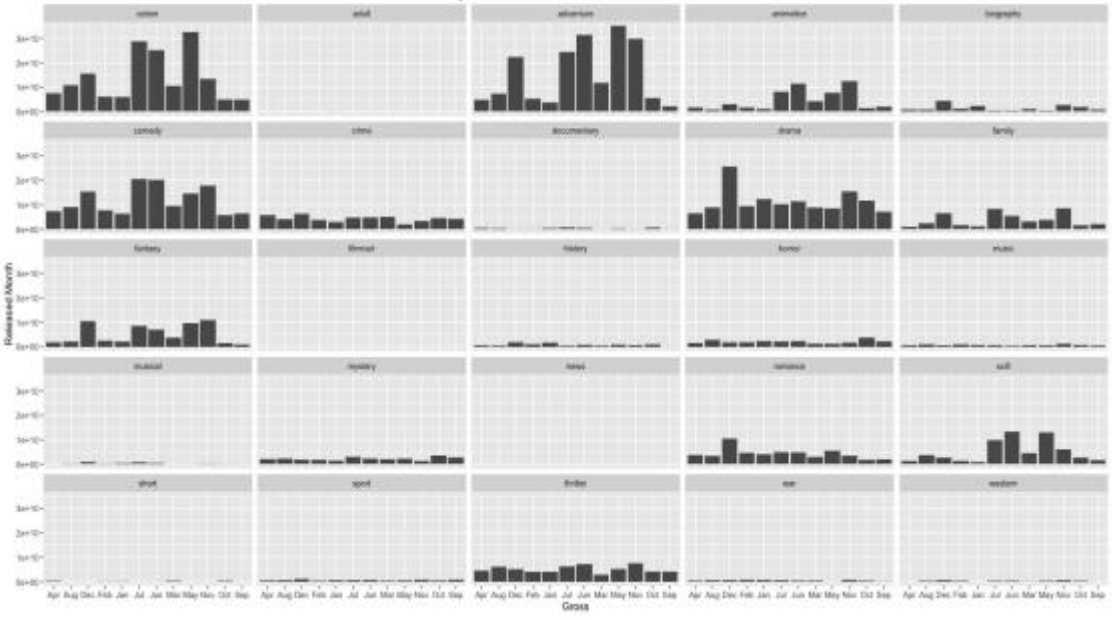


Figure5.4 is a graphical representation of a genre-based recommendation for release date that is likely to increase the title's revenue. May month is best month for action, adventure, and western genre movies.

6.

Figure:6.1 Density plot: Imdb Rating Vs Tomato Rating

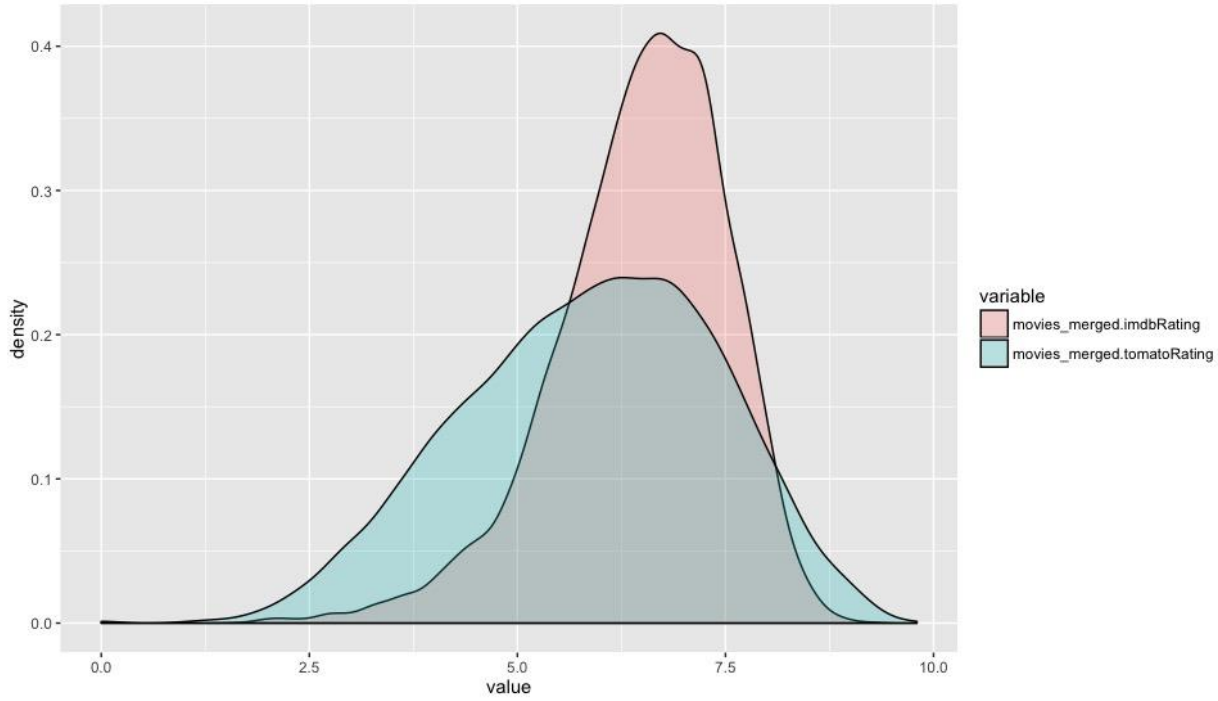


Figure:6.2 Density Difference: Imdb Rating Vs Tomato Rating

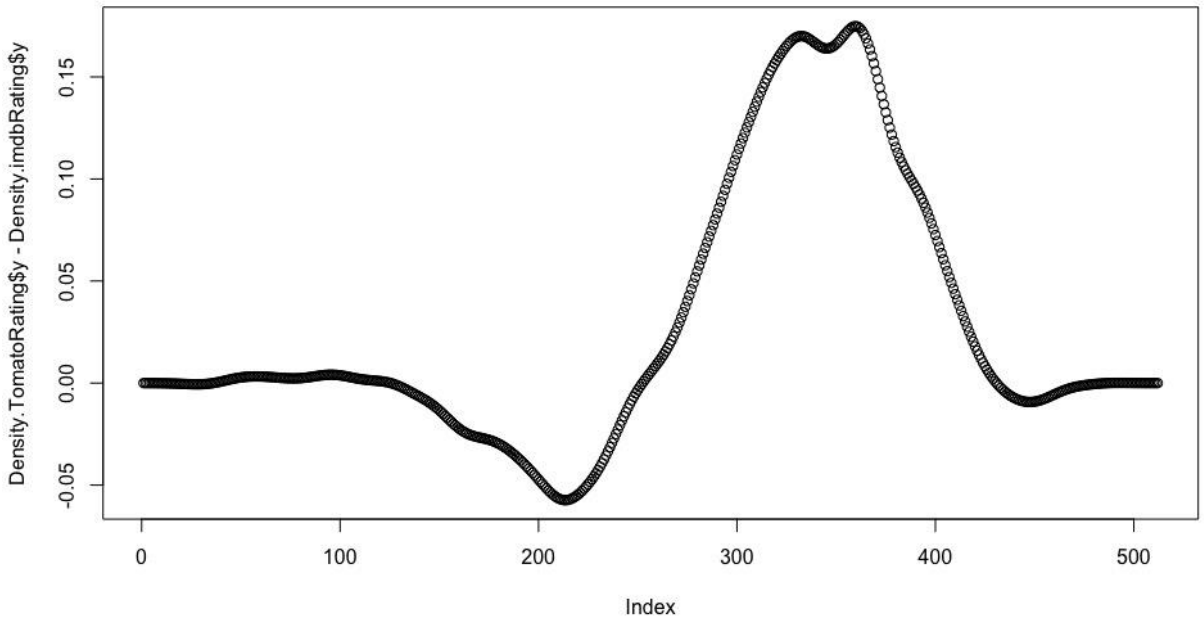


Figure: 6.3 correlation between imdb rating and tomato rating

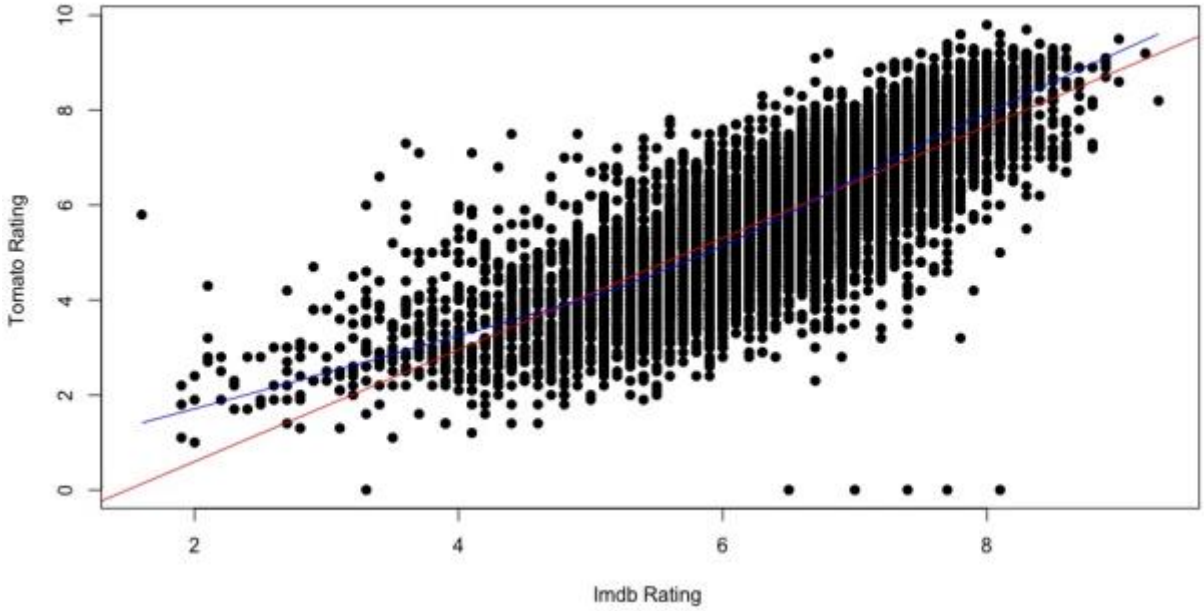
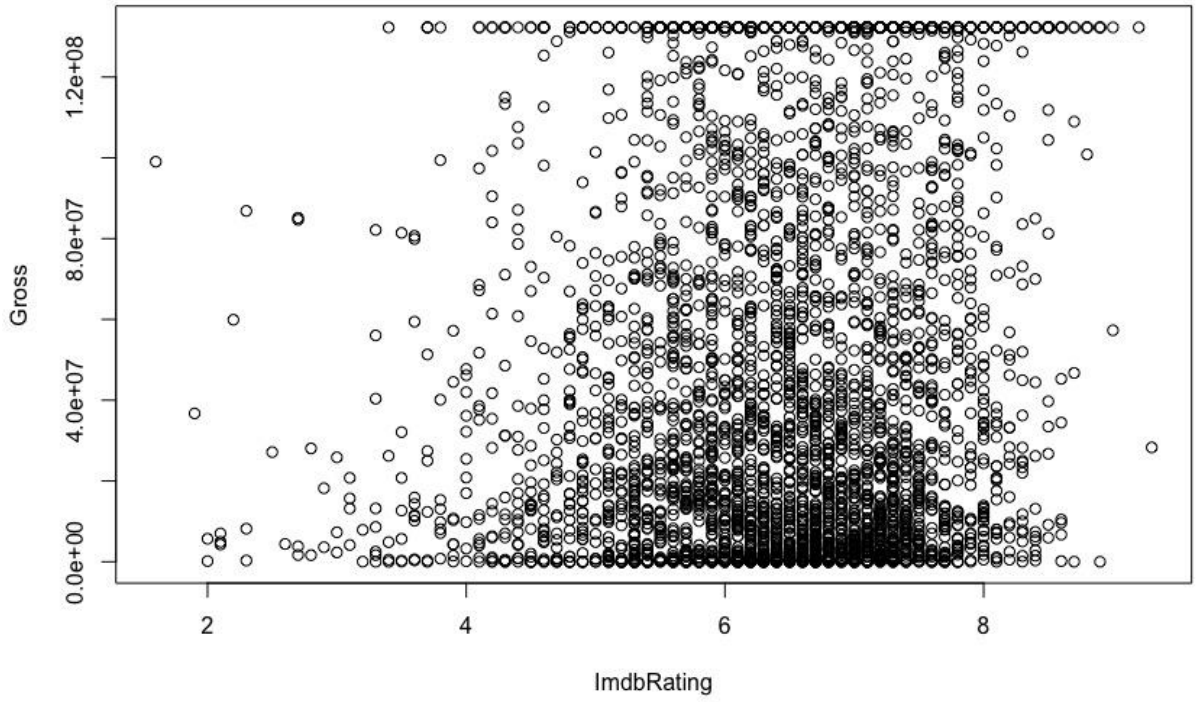


Figure: 6.4 Imdb Rating Vs Gross



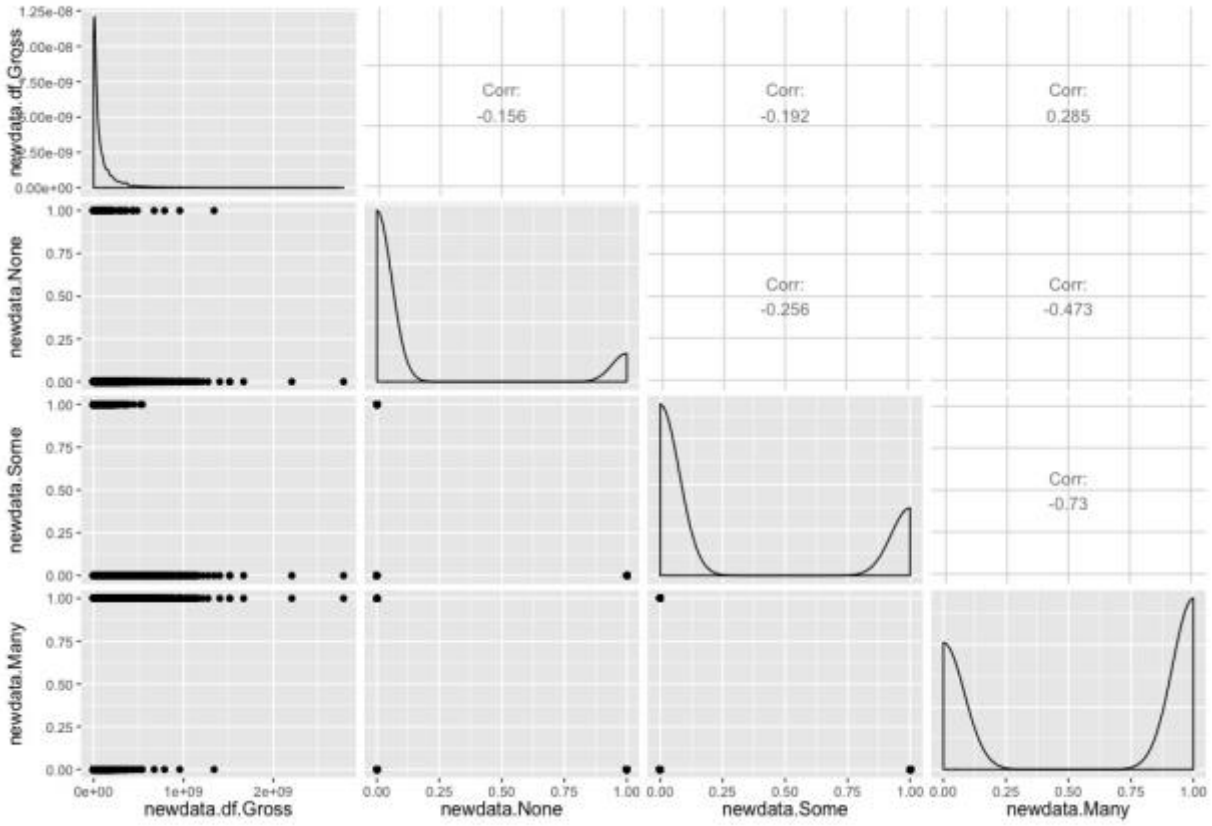
Most imdb raters tend to rate movies that they liked and skip to rate that don't like. By looking at the figure 6.4, it is likely to get high rating for lower grossed movies and vice versa. Ratings are more dispersed on Tomato rating than that in imdb.

Similarities and difference between Imdb rating and Rotten tomato ratings

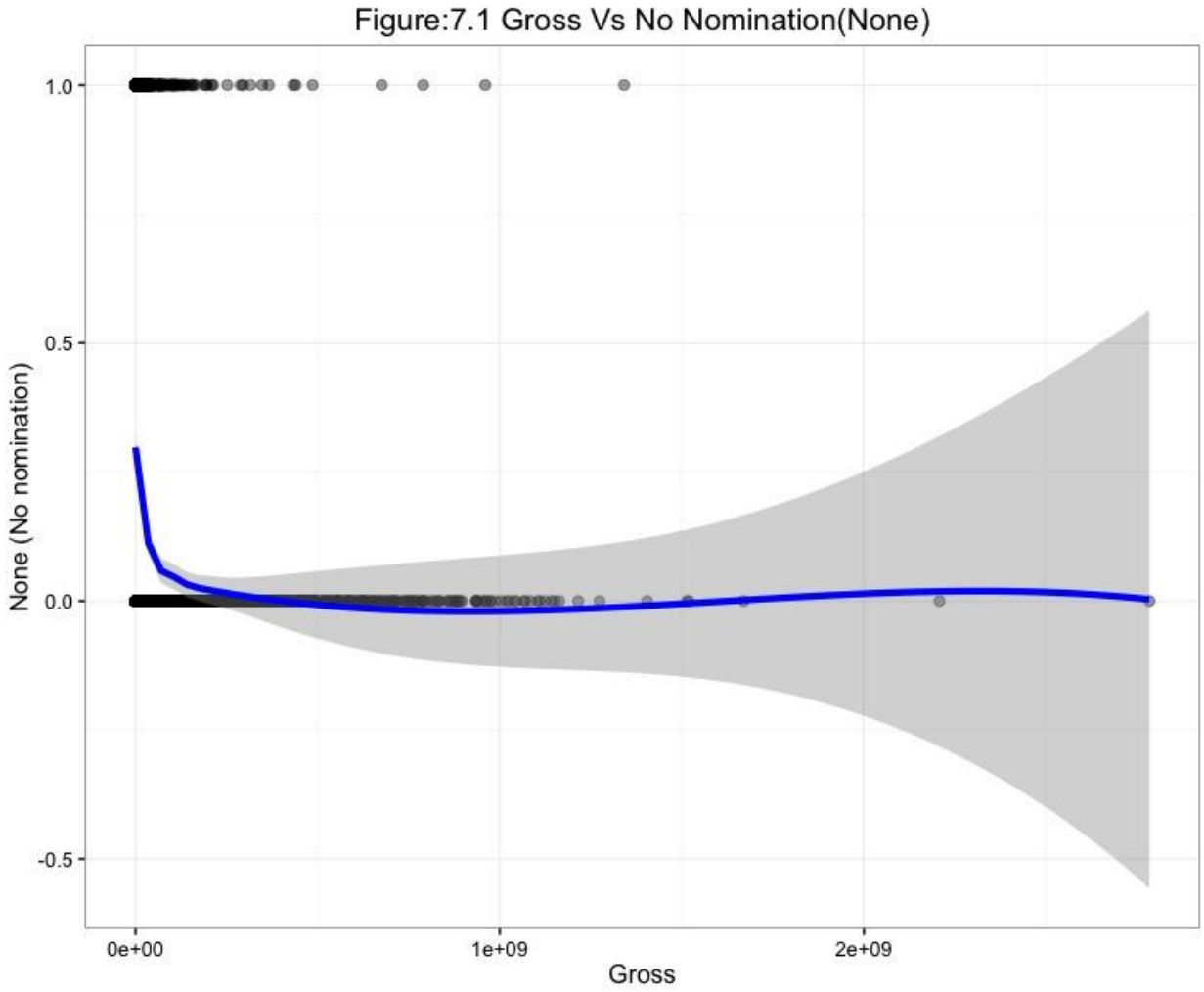
IMDB calculates rating based on the average based on how people fill out ratings and to avoid "vote stuffing" (Weighted Average Ratings). On the other hand, Rotten Tomatoes is based on scores given by critics. They do not average the rating. They categorized the rating into 'Fresh' and 'Rotten'. Any movie that receive 60% reading on the Tomatometer for that movie will be considered as 'Fresh' and if a movie receives less that 60 % reading will be considered as 'Rotten'.

Figure: 6.1 shows the relationship between the imdb rating and tomato rating. *Figure* 6.2 explains the difference between the density of tomato rating any imdb rating. *Figure* 6.3 shows the correlation graph using scatter plot. The correlation value(r) is 0.79. if the r value is between .5 to .8, it is considered as Medium positive correlation. By the looking at those graph, For the top rated movies, the disparity is actually not high at all.

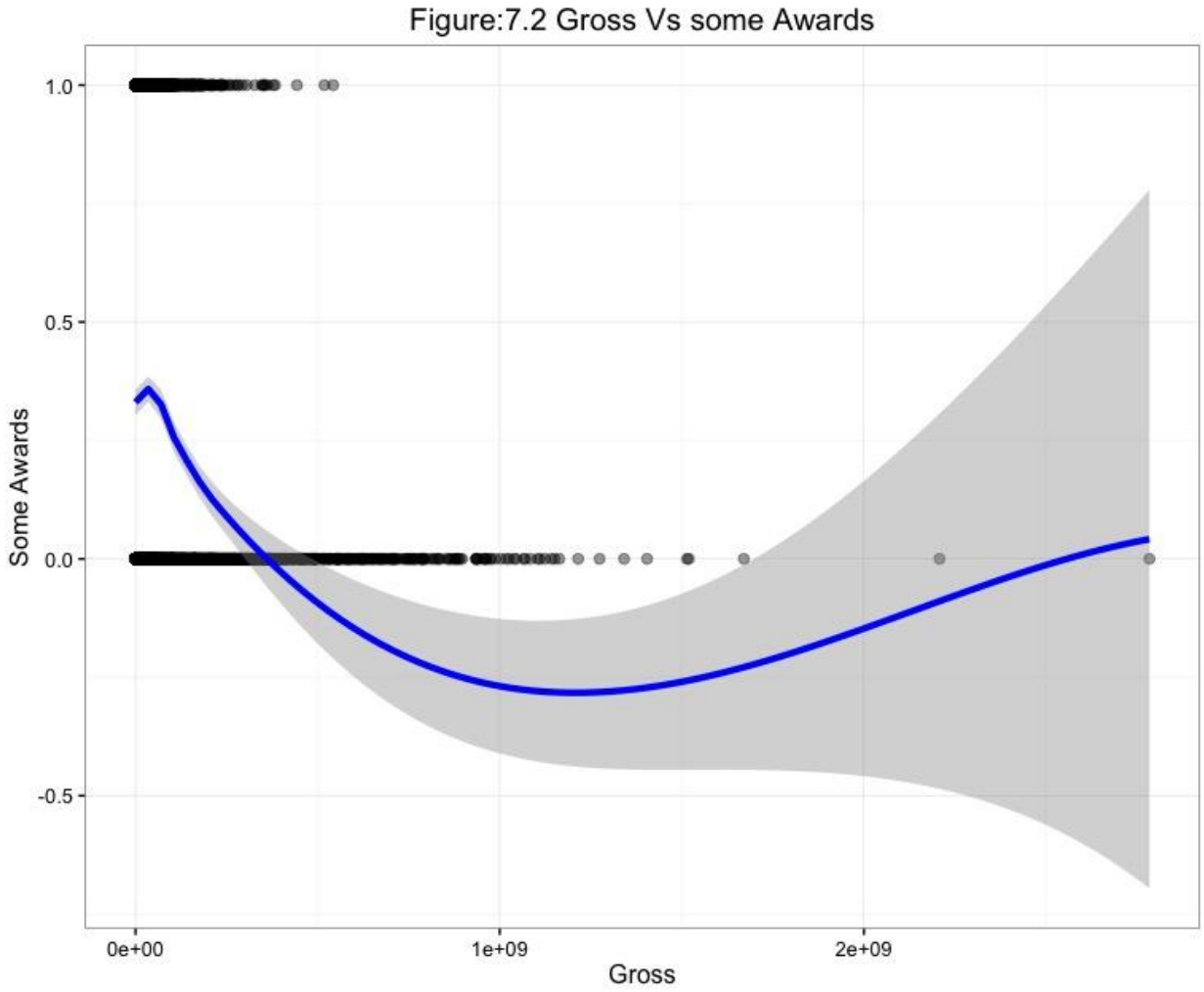
7.



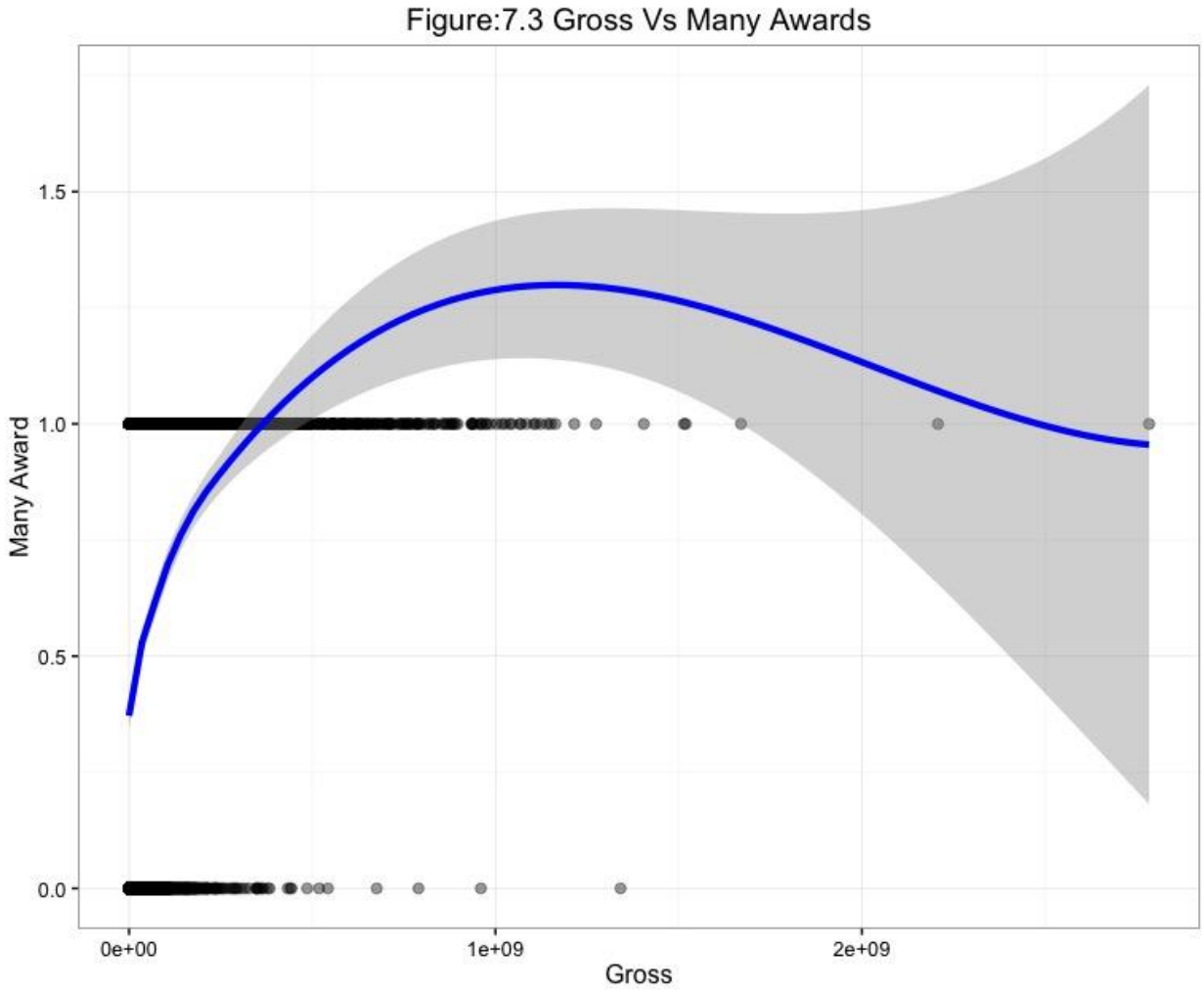
The above shows an overall correlation graph of Gross and three categories.



1.0[true] and 0[false] on the Y-axis represent the true and false variable respectively whereas x-axis represent the gross revenue in the Figure 7.1. The blue regression line accross the graph expresses the relationship between the gross revenue and Y-axis variables (true and false) . Movies with high gross revenue win some to many awards which is represented by the plot points accross the Y-axis indexes.

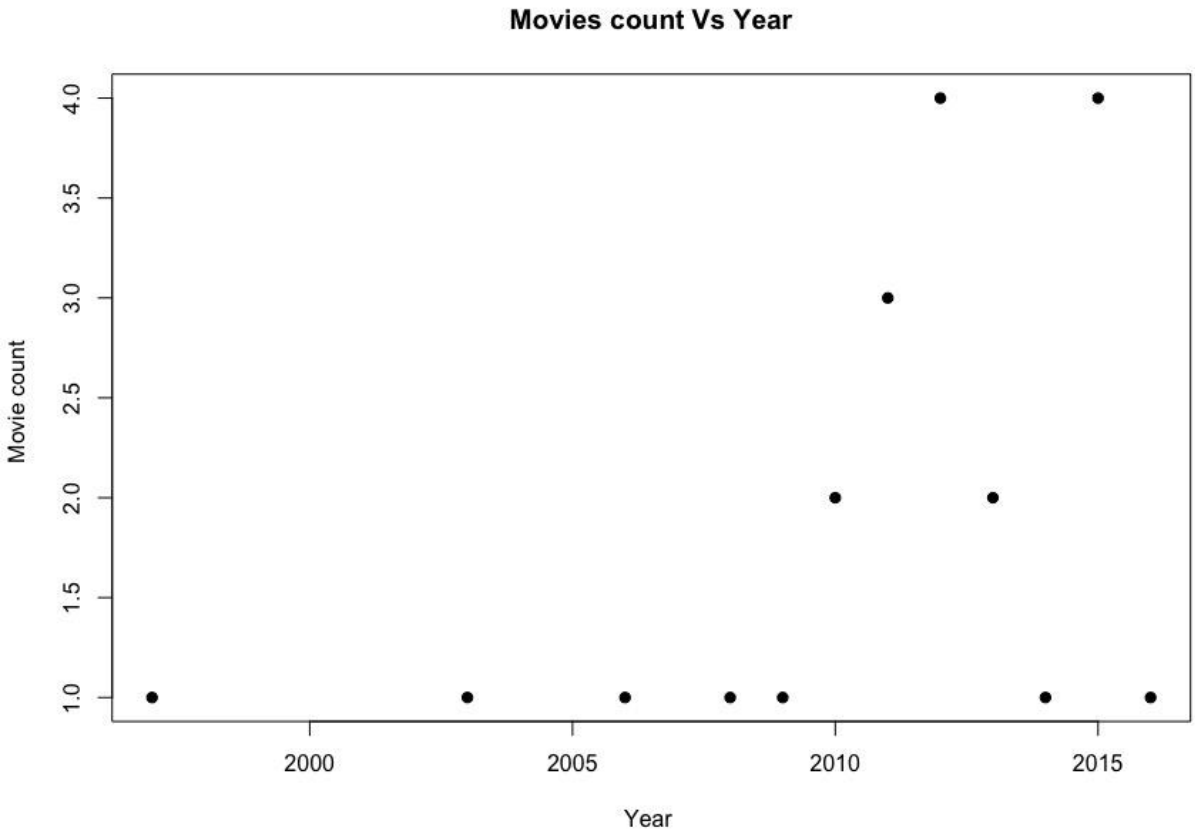


1.0[True] and 0.0[False] on the Y-axis represent the true and false variable respectively whereas x-axis represent the gross revenue in the Figure 7.2. The blue regression line across the graph expresses the relationship between the gross revenue and Y-axis variables (true and false). As the gross revenue increases beyond gross revenue value 1e+09, the movies win either many awards or no awards at all which is represented by the plot points.



#1.0 and 0.0 on the Y-axis represent the true and false variable respectively whereas x-axis represent the gross revenue in the Figure 7.3. The blue regression line across the graph expresses the relationship between the gross revenue and Y-axis variables (true and false). As the gross revenue increases approx. beyond gross revenue value 1e+09, the movies win many awards as oppose to winning some or none awards which is represented by the plot points.

8.
[Insight -1] How many movie are there that generate at least a billion gross revenue ?
[Insight -2] How many of these movies were released in which year?
Describe and investigate a relationship between the movies and its gross revenue.



There are 22 movies that grossed atleast billion dollar.

By looking at the figure 8.1, there are 4 movies released in 2012 and 2015 that grossed atleast or more than billion dollar. In year: 1997, 2003, 2006, 2008, 2009, 2014, 2016, each of those years, a billion dollar grossed movies were released.

New insight that are not expected

As the movie's gross revenue goes up, the probability of winning atleast an award goes up as well. This expected consequence is valid for majority of movies. However, this is not true for all movies since there are some movies that have grossed more than a billion dollars and won no awards, which is an unexpected consequence. For example, the movie "Harry Potter and the Deathly Hallows: Part II" has engrossed $1,341,511,219 and has won no awards. In Figure 7.3, the blue regression line expresses the relationship between the gross revenue and winning many awards. However, the number of movies winning many awards lowers as the gross revenue increases more than the gross value 2e+09 which represents an unexpected consequence.

Figure:7.3 Gross Vs Many Awards