Johnpaul Ogah (jogah3@gatech.com)

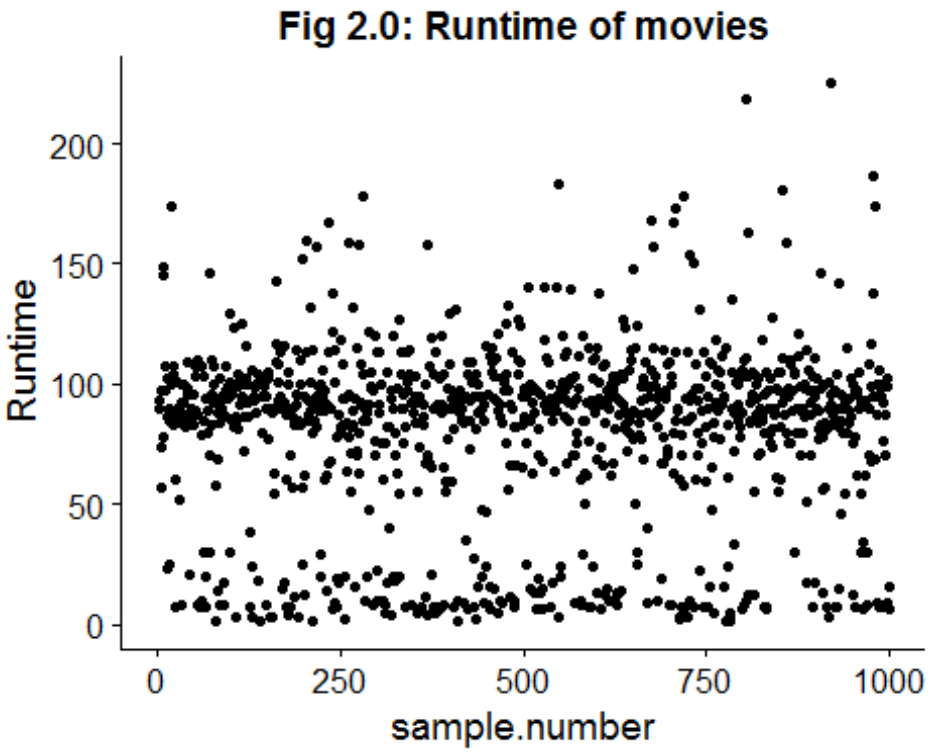
1) The variable Type captures whether the row is a movie, a TV series, or a game. Remove all rows that do not correspond to movies. How many rows did you remove?

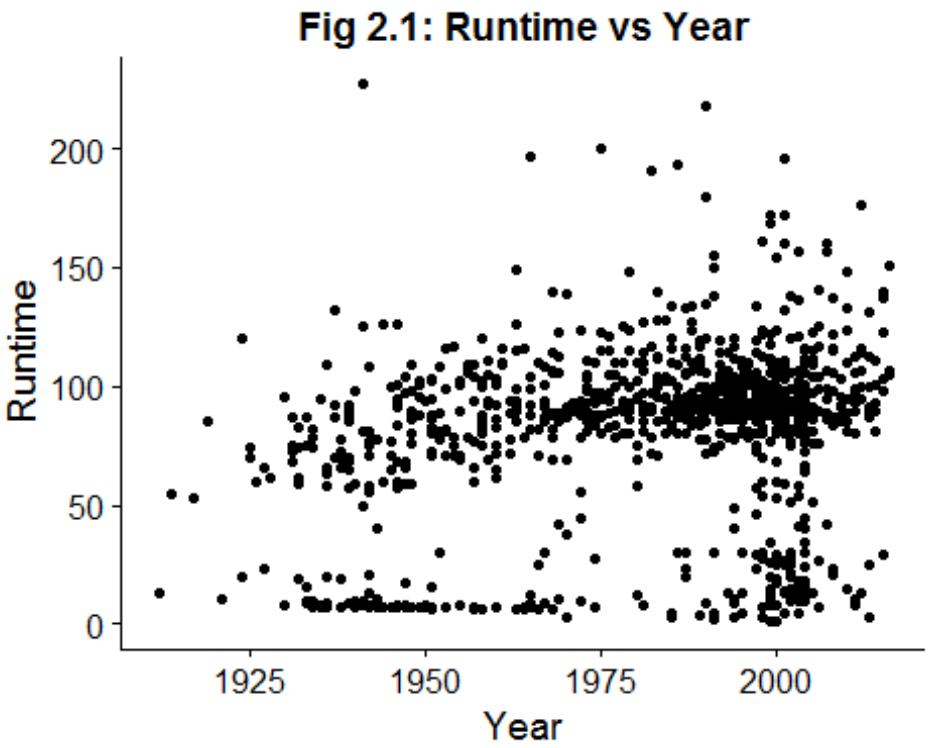
The code for this section is shown in Listing 1.0 in jogah3_project1.R script:

As seen from the output from the output of code, the total number of rows removed is 789.

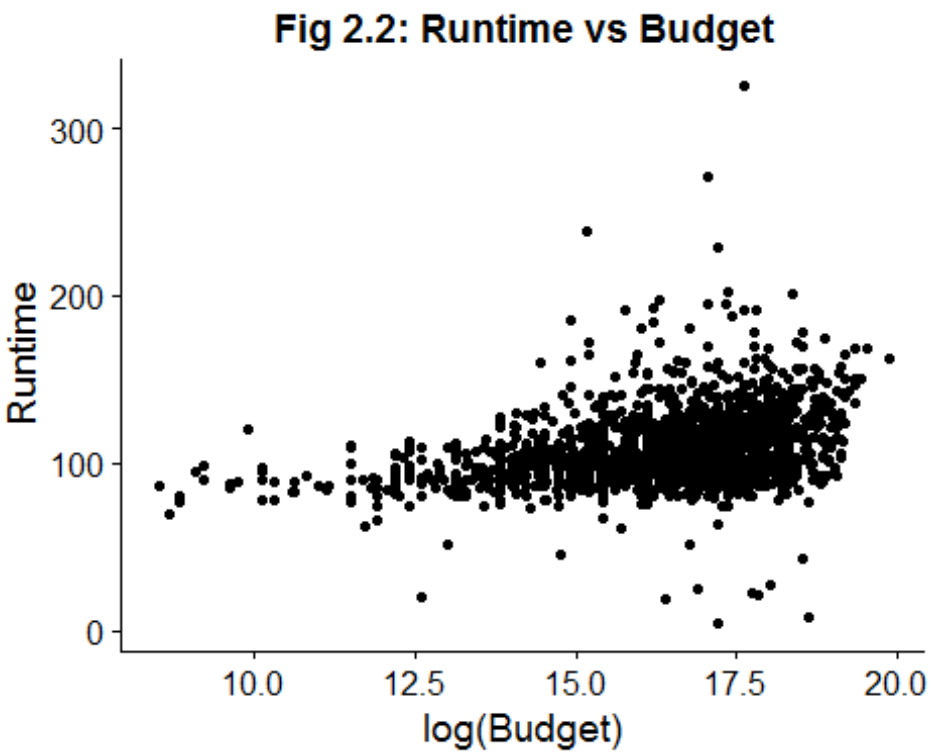
2) The variable Runtime represents the length of the title as a string. Write R code to convert it to a numeric value (in minutes) and replace Runtime with the new numeric column. Investigate and describe the distribution of that value and comment on how it changes over years (variable Year) and how it changes in relation to the budget (variable Budget).

The code for this question is shown in Listing 2.0 in the R script. As seen from the fig 2.0, we see that movies have two typical runtime of 0-25min and 75-125min with majority of the movies having a runtime in the range of 75-100 minute.





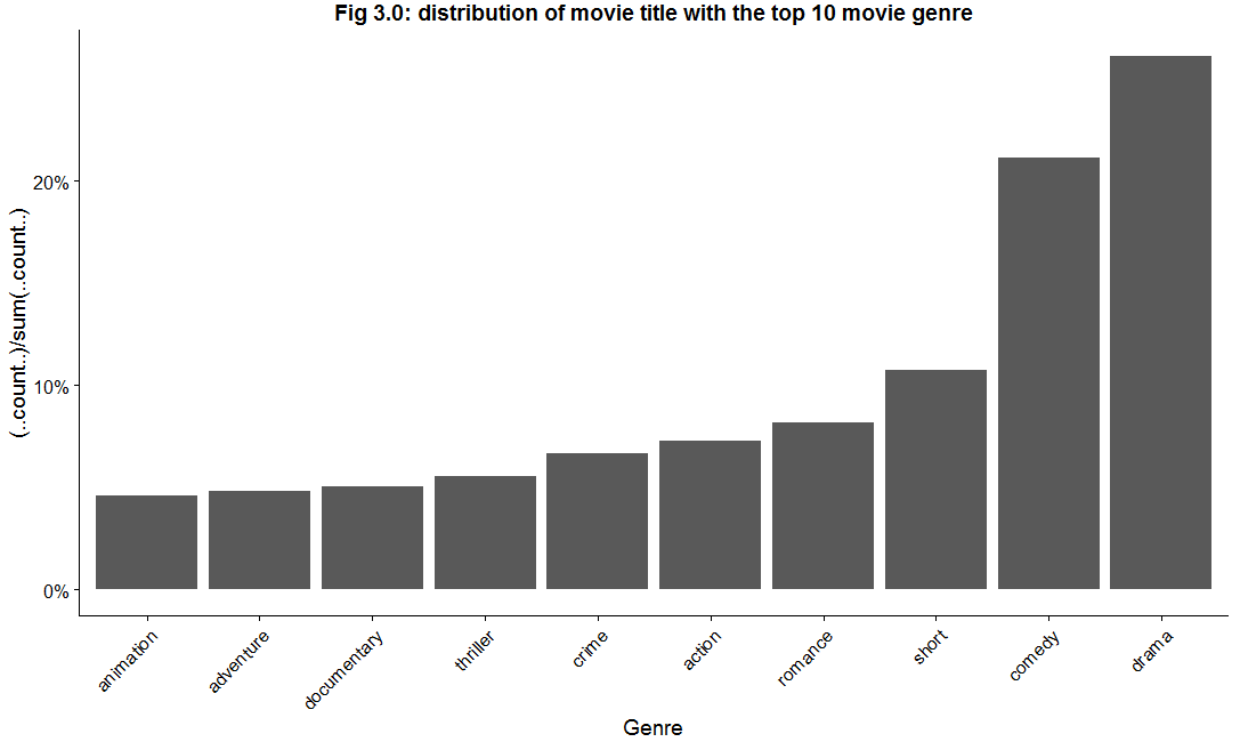
As seen from the second fig 2.1 above, two typical runtime of 0-25 min and 75-125 minute exist across various years which is consistent with our previous result. Again majority of movies have a runtime between 75 and 100 minute, this is also consistent with our previous result.



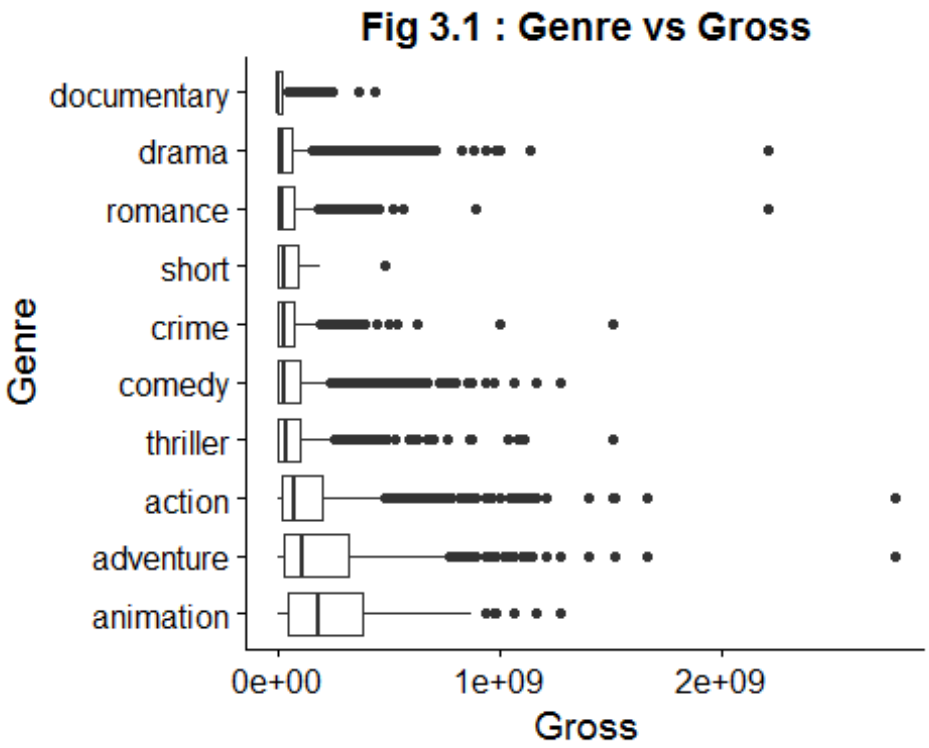
From the fig 2.2 above, majority of movies with a budget have a runtime in the range of 75-125 minute. This is also consistent with previous result.

3. The column Genre represents a list of genres associated with the movie in a string format. Write code to parse each text string into a binary vector with 1s representing the presence of a genre and 0s the absence and add it to the data frame as additional columns. For example, if there are a total of 3 genres: Drama, Comedy, and Action a movie that is both Action and Comedy should be represented by a binary vector (0, 1, 1). Note that you need to first compile a dictionary of all possible genres and then figure out which movie has which genres (you can use the R tm package to create the dictionary). Graph and describe the relative proportions of titles having the top 10 genres and examine how the distribution of gross revenue (variable Gross) changes across genres

Fig 3.0: distribution of movie title with the top 10 movie genre



The code is shown in Listing 3.0 in the R script's seen from fig 3.0 above, the top 10 genre are(in this order) :1)drama 2)comedy 3)short 4) romance 5)action 6)crime 7)thriller 8)documentary 9) adventure 10)animation .Further the histogram plot shows the proportion of title with the top 10 genre , we see 27% of the titles have drama as Genre and ~4.5% of titles have animation as Genre. We can see the percentage of the other top 10 genre as well from the graph.



As seen from the fig 3.1 above, the Genre with highest gross revenue is animation and lowest gross revenue is documentary

4. Find and remove all rows where you suspect a merge error occurred based on a mismatch between these two variables. What is your precise removal logic and how many rows did you end up removing? The data frame was put together by merging two different sources of data and it is possible that the merging process was inaccurate in some cases (the merge was done based on movie title, but there are cases of different movies with the same title). The first source's release time was represented by the column Year (numeric representation of the year) and the second by the column Release (string representation of release date). Find and remove all rows where you suspect a merge error occurred based on a mismatch between these two variables. To make sure subsequent analysis and modeling work well, avoid removing more than 10% of the rows that have a present Gross variable. What is your precise removal logic and how many rows did you end up removing?

After researching imdb database and comparing it with our project dataset(movie_merged), I realized that “Year” and “Released” column in our project dataset are from the same source i.e imdb . I queried for different titles using the omdb api and the result I got from each confirms this conclusion. The “Year” represent the year of first public screening for a particular movie and “Released” represent the actual released date. The year in both variable need not be the same for the same movie .There is no missing value in the “Year” column but there are some missing value in “Released” column. Missing value in release column can indicated the release date is unknown or not yet released. I removed “0” rows for this question.

5. An important question is when to release a movie. Investigate the relationship between release date and gross revenue and comment on what times of year are most high revenue movies released in. Does your answer changes for different genres? Based on the data, can you formulate a genre-based recommendation for release date that is likely to increase the title's revenue? If you have a recommendation motivate it with the appropriate disclaimers, or otherwise explain why you are unable to produce a recommendation.

The code for this question is shown in Listing 5.0 in jogah3_project1.R script.

As seen from fig 5.1, high revenue movie are released in the 51st week. That is, movies release few days before the Christmas holidays are more likely to generate higher revenue than other release date.

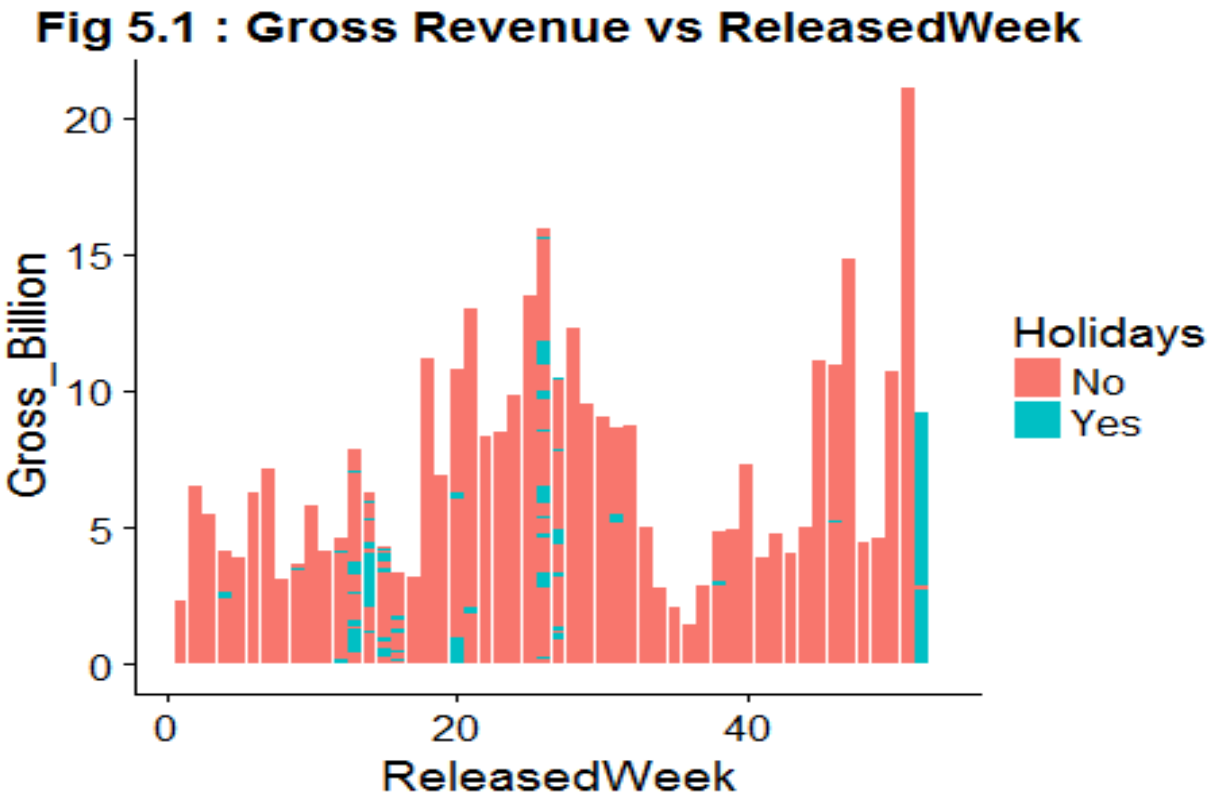
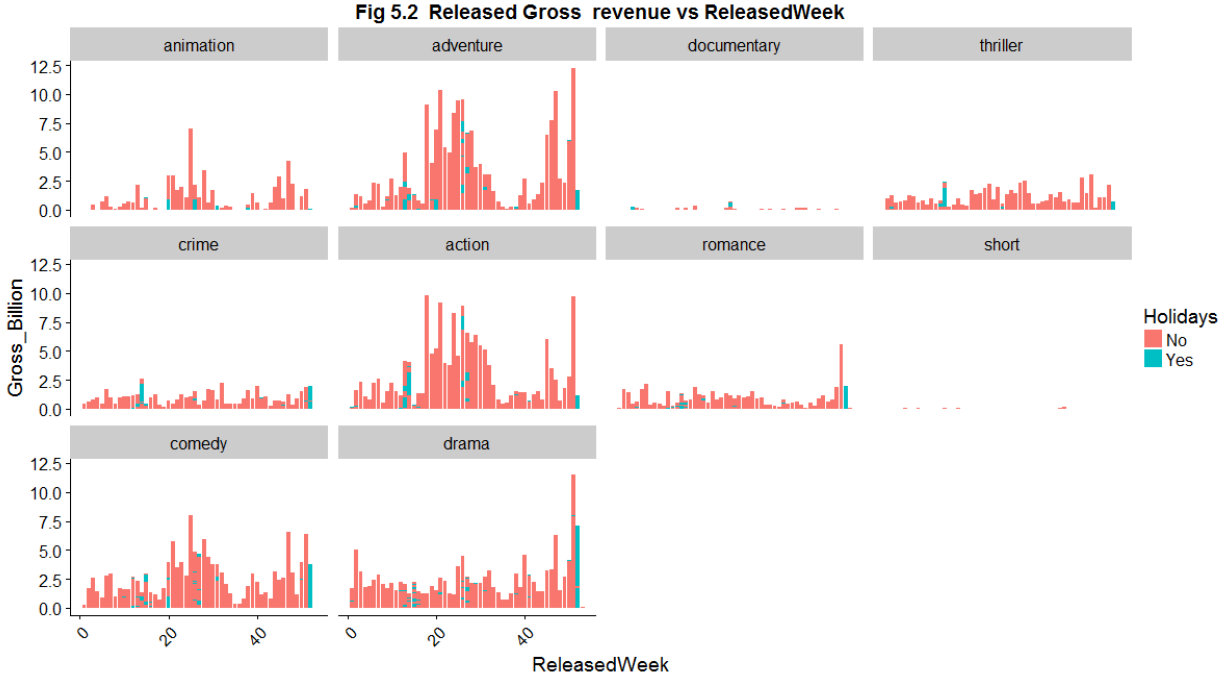


Fig 5.2 Released Gross revenue vs ReleasedWeek



As seen from fig 5.2 , the gross revenue distribution for different weeks changes with different genre. The genre adventure has the highest gross revenue and the genre short has the lowest gross revenue.

Also as seen from fig 5.2 , the genre combination of adventure, action and drama for release date of 51st week or a few days before Christmas will likely result in increased gross revenue.

Disclaimer: This recommendation was based on the data used for analysis and may be inaccurate and as a result may not give the expected outcome.

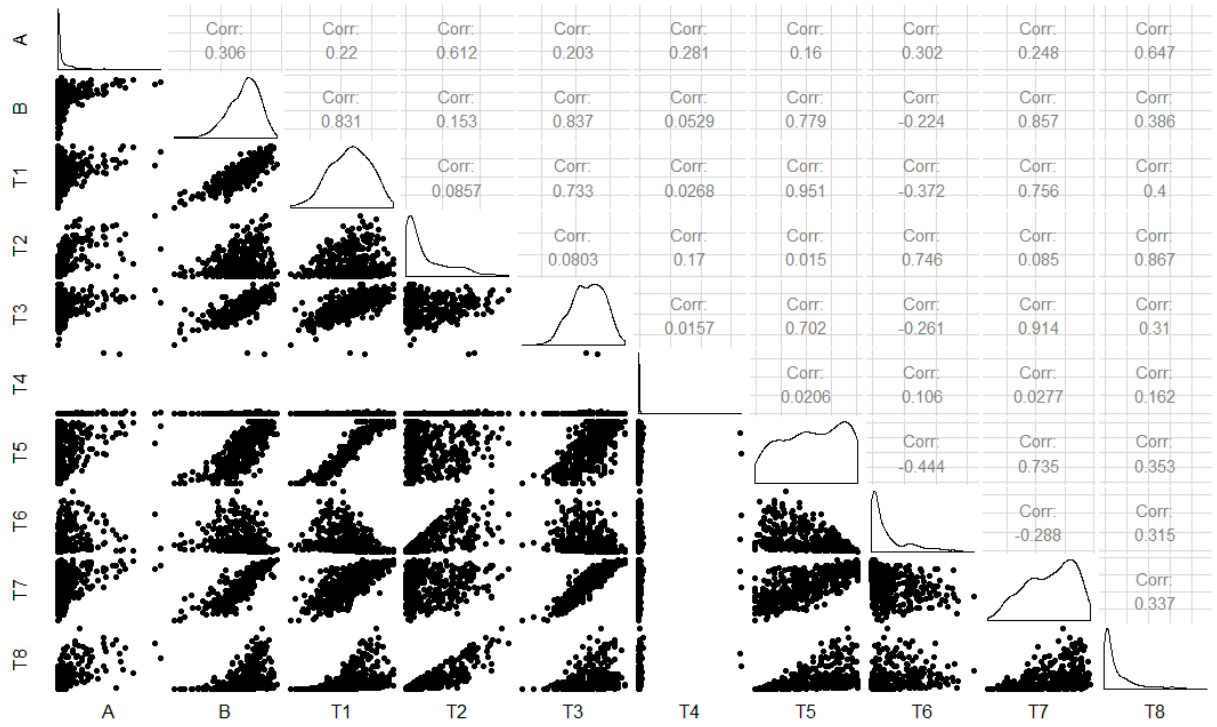
6. There are several variables that describe ratings including IMDb ratings (imdbRating represents average user ratings and imdbVotes represents the number of user ratings) and multiple Rotten Tomatoes ratings (represented by several variables pre-fixed by tomato). Read up on such ratings on the web (for example rottentomatoes.com/about and *http://* www.imdb.com/help/show_leaf?votestopfaq) and investigate the pairwise relationships between these different descriptors using graphs. Comment on similarities and differences between the user ratings of IMDb and the critics ratings of Rotten Tomatoes. Comment on the relationships between these variables and the gross revenue. Which of these ratings are the most highly correlated with gross revenue (use the R function cor and remove rows with missing values)?

The code for this question is in Listing 6.0 in jogah3_project1.R script . To avoid the variable from overlapping in x-axis and y-axis I have renamed the variables as below:

imdbRating -> A, imdbVotes -> B, tomatoRating -> T1, tomatoReviews -> T2

tomatoUserRating -> T3, tomatoUserReviews ->T4

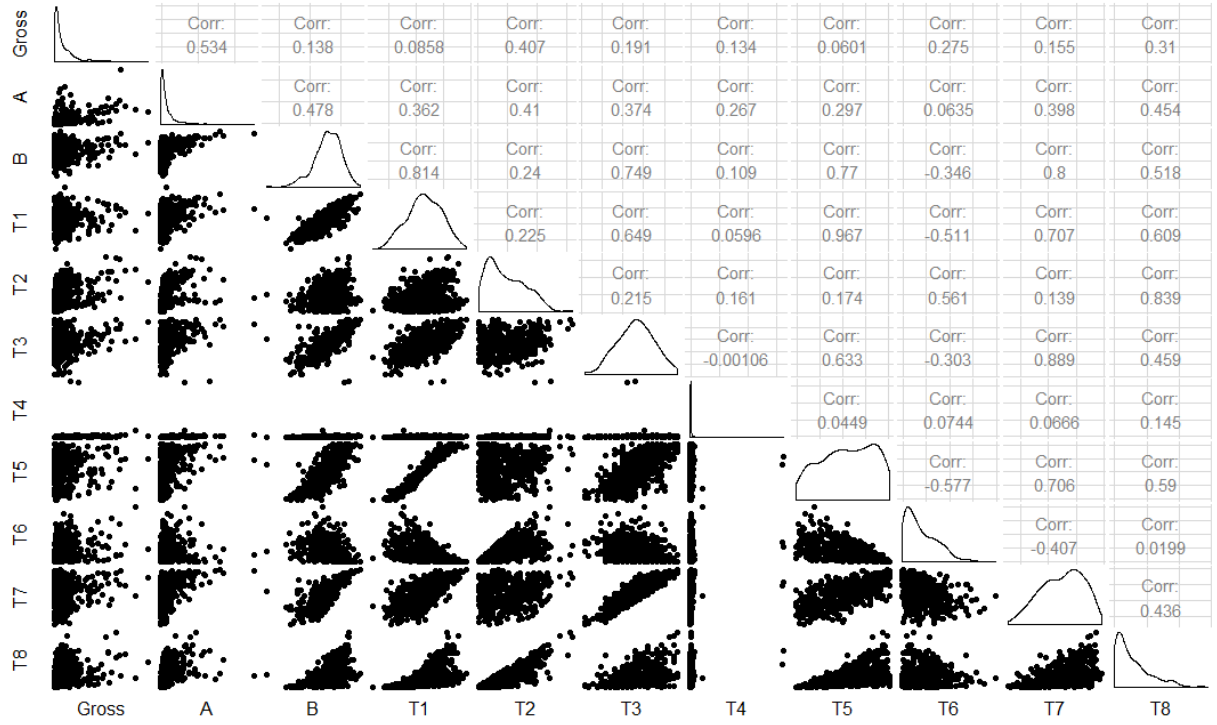
tomatoMeter -> T5 , tomatoRotten -> T6 , tomatoUserMeter -> T7 , tomatoFresh -> T8



To investigate the pairwise relationship between the variables, I used the ggpairs function available in GGally packages. As shown in the figure above, we can see that "imdbVotes"" has a linear relationship with "tomatoReviews" and the correlation between this two varoiables is 0.625. We see imdbVotes increasing with increasing tomatoReviews. Similarly imdbRating has a linear relationship with tomatoRating with a correllation of 0.821 between both variables.imdbRating is also related with tomatoUserRating with a correlation of 0.844.Imdbrating also has linear relationship with tomatoMeter with a correlation of 0.763.Imdbrating also has linear relationship with tomatoUserMeter with a correlation of 0.856. tomatoFresh is related to tomatoReviews with a correlation of 0.879 between variables.tomatoUserMetre is related with tomatoUserReview with a correlation of 0.654.tomatoUserMetre related to tomatoUserRating with correllation of 0.912.tomatoUserMetre related to tomatoRating with correlation of 0.725.tomatoRotten is related with tomatoReveiw with correlation 0.724.tomattometter is related to tomatoUserrating with correlation 0.654 tomattometter is related to tomatorating with correlation 0.948 tomatoRating is related to tomatoUserRating with correlation 0.713 imdbVotes is related to tomatoReveiw with correlation 0.625.

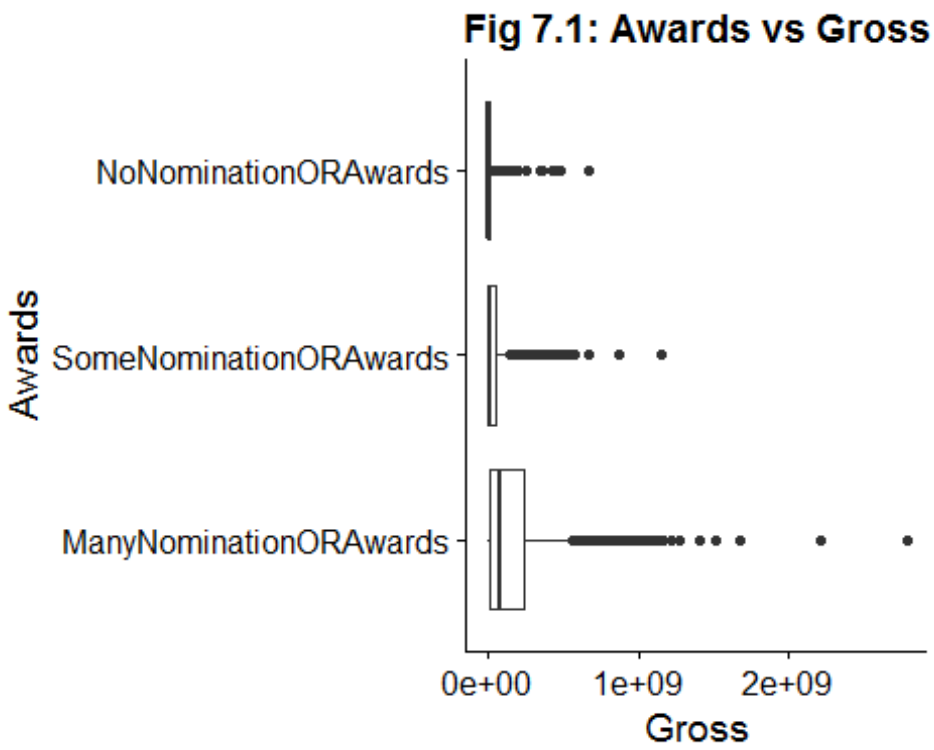
We can see that the imbdRating has no relationship with imdbVotes, whereas the imdbVotes has a linear relation with one of the tomato rating variables "tomatoReviews" and why the the tomato rating have a pairwise relation between its variables, the two imdb rating variables are not related. Also imdbRating has a linear relationship with several tomato rating variables.

Further to see which variables are most highly correlated with the Gross variables, again I used the ggpairs function. As seen in the graph below imdbVotes, tomatoReviews and tomatoFresh are the variables that are most correlated with the gross revenue.



7. The variable Awards describes nominations and awards in text format. Convert it to a three dimensional binary vector whose first component represents no nomination or awards, the second component represents some nominations/awards, and the third component represents many nominations or awards. The relationship between the second and the third categories should be close to 5:1 (not precisely - this is a broad guideline to help you avoid creating a third category that is useless due to being extremely small and to encourage consistency). How did you construct your conversion mechanism? How does the gross revenue distribution changes across these three categories.

The code for this question is shown in List 7.0 in the jogah3_project1.R code.



As shown in the code, I replaced missing values with 0 in the Awards column. Then I created a function called "extract.numbers" , it extract the digit numbers from a string and sum its values. If the sum is 0, it returns 1 for no nomination, if the sum is greater than 0 but <= 11 it returns 1 for some nominations and if the sum is greater than 11 it returns 1 for many nominations.

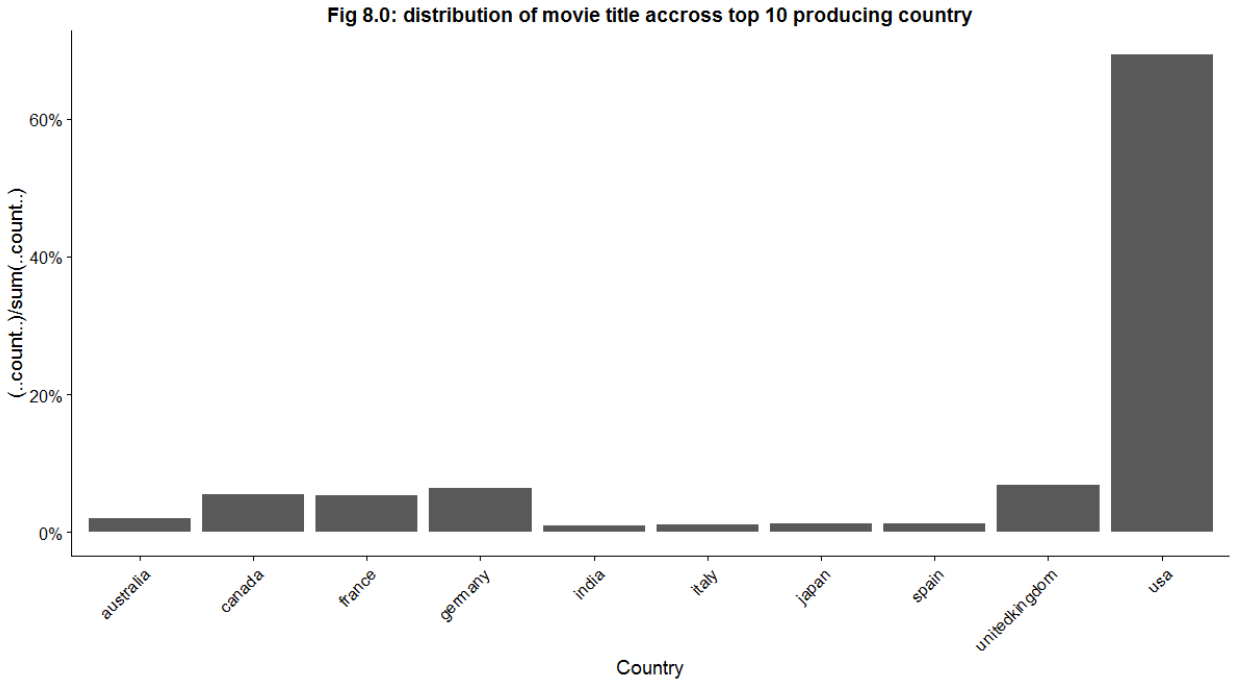
As seen in fig 7.1 , manyNominationORAwards has the highest median gross revenue and NoNominationORAwards has the lowest median gross revenue. We see that the number of awards affect the gross revenue.

8. Come up with two new insights (backed up by the data and graphs) that are expected, and one new insight (backed up by data and graphs) that is unexpected at first glance and do your best to motivate it. By "new" here I mean insights that are not an immediate consequence of one of the above assignments

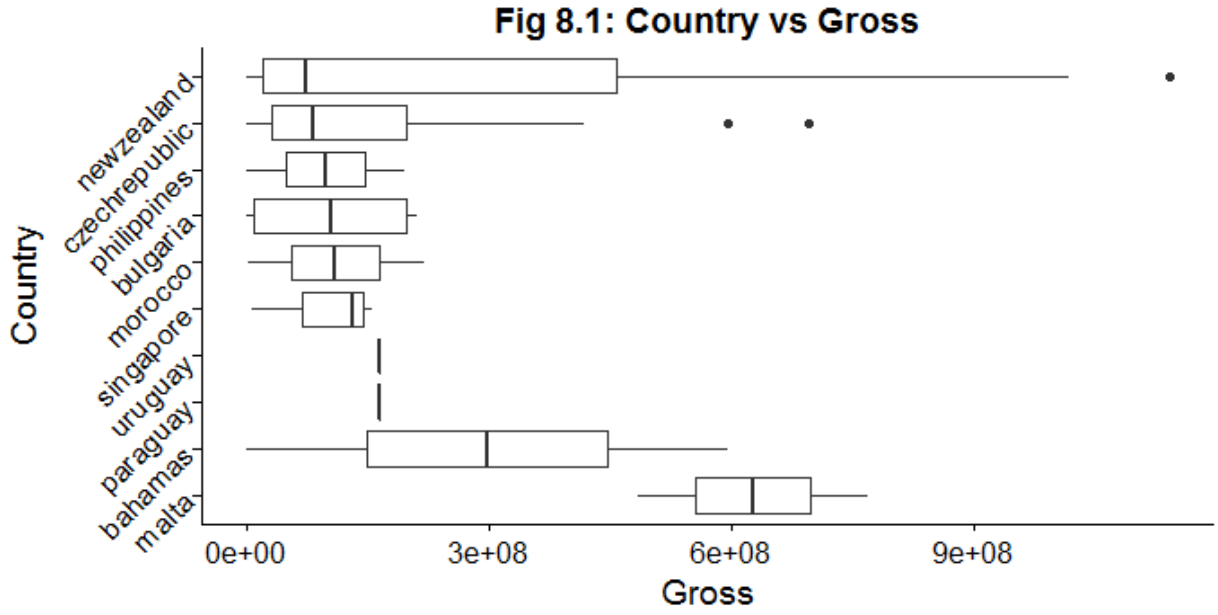
The code for this question is in Listing 8.0 in jogah3_project1.R script.

Further research of the "Country" column reveals that this column represent the country where the movie was produce/shot. A particular movie can be shot in multiple countries. I seek to investigate the relationship between the Country where the movie is produced/shot, Genre and the Gross revenue. Things I expected are: 1) majority of the movies will be produced in developed countries like USA and United Kingdom. Hollywood being the largest movie producing industry is located in the USA so I expect most movies to have USA as producing country. 2) Most popular genre for a particular country should be a high gross revenue generating genre

What I do not expect is: the country where the highest median gross revenue movie are produced are anywhere other than the USA. So I expect the movies shot in USA to have the highest median gross revenue.

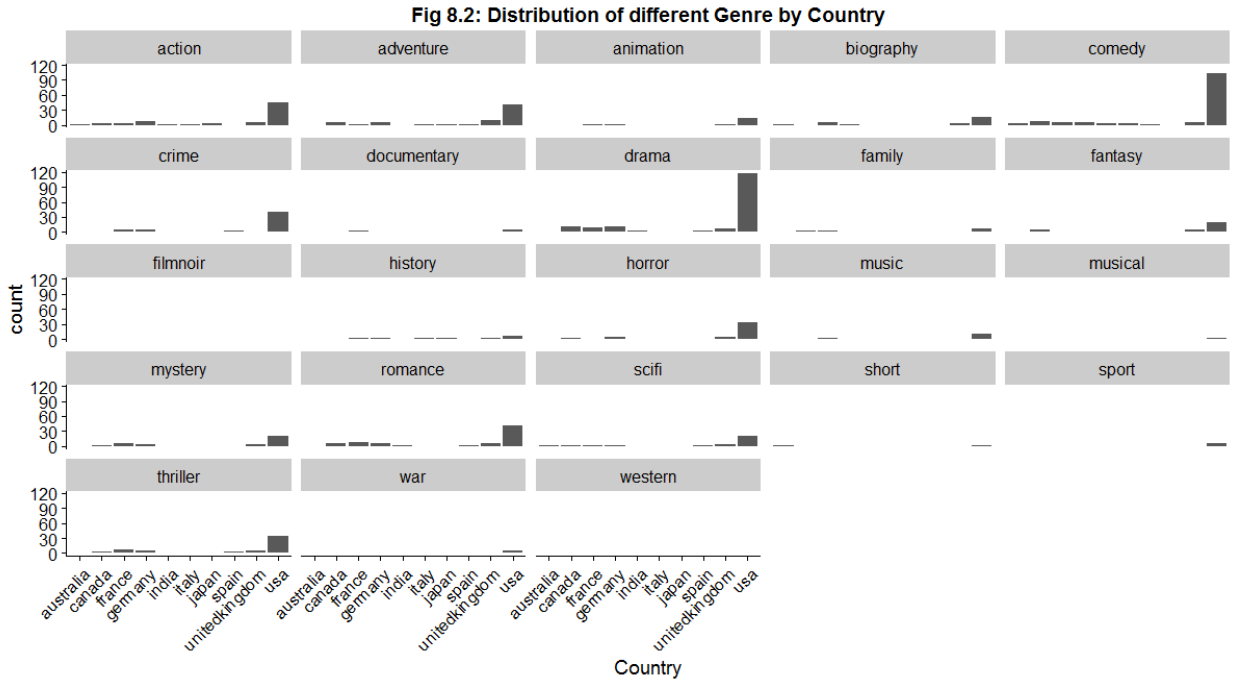


As seen in fig 8.0, about 65% of the movies have USA as country of production. United Kingdom, Germany, Canada and France each have less than 10% of movies shot there.



As seen in fig 8.1, movies that has Malta as a country of production has the highest median gross revenue. This is really unexpected result but after a little more research I found out that some very

popular high gross revenue movie titles like captain Philips, World War Z , along with others were produced in Malta. This movies were directed by Hollywood actors but they were completely shot/produced in Malta, thus the reason why a Malta has the highest median gross revenue across different countries. It's surprising to see that movies shot in USA is not among the top 10 of highest median gross revenue.



As seen in fig 8.2, the popular genre in USA are drama, comedy, adventure and action. According to previous graph in fig 3.1, these are also high revenue genre which confirm what I had expected.