PROJECT REPORT

ECONOMETRIC ANALYSIS OF MOVIE DATASET

AKSHAY ANANDBABU

RITVIK REDDY

DA-5020 COLLECT, STORE, RETRIEVE DATA

SUMMER-FULL 2017

TABLE OF CONTENTS

1. Project Description

As movie enthusiasts, we were immediately drawn to the idea of using movies as our topic. The main motive of a movie producer is to earn more money than was invested in creating the movie. Through our analysis, we tried to gain insight on the thoughts of a producer and what are the keys to making a successful movie. We decided to work on a movie dataset to address the factors that affect the success of a movie worldwide. In our project, we base the success of a movie by how much it earns in the domestic and worldwide box office. Thus, the theme of our project is based around the worldwide box office collection and the various components that affect it. We perform this econometric analysis using plots in R and creating a linear regression model to see the correlation between the different variables and the worldwide box office collection.

2. Selection Rationale:

To perform our analysis, we needed data on the production budget and worldwide box office collection and ratings for a variety of movies released internationally. Moreover, we needed a mix of movies with different genres and content ratings so as to widen our range of analysis. The ***the-numbers.com*** website provided a comprised list of the highest grossing movies of all time released globally. The following were the major reasons for choosing the website for our project:

- The website is open-source and does not require access authorization.
- The entire data was available in a single page as a search result
- It provided a comprehensive profile for each movie that included the domestic, international and worldwide income for the movie

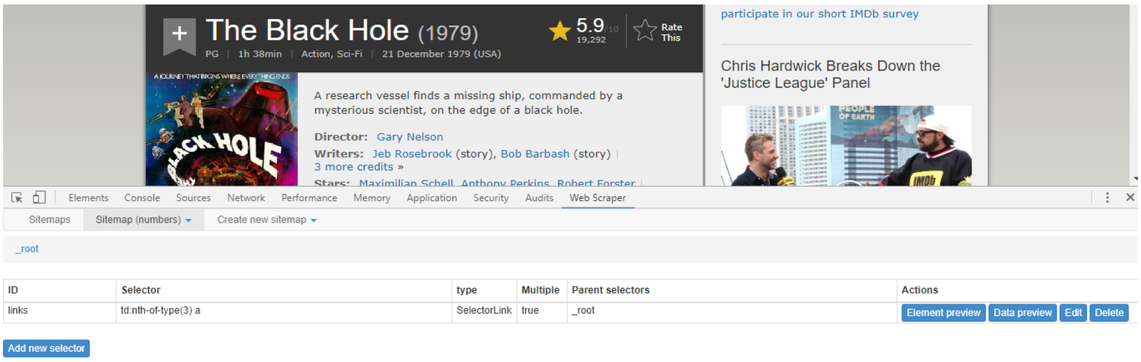
- It provided both the domestic and international release dates separately so if the international release date for a movie was missing, we could make the assumption that the movie was only available in the country of release.

3. Process Implemented

1. Scrape the movie data from “the-numbers.com” using WebScraper extension in Google Chrome.
2. Download it as a CSV file and save it.
3. Import the CSV into the R environment.
4. Clean the data-set for abnormalities and store it in a new table.
5. Store the tidy dataset into a SQLLite database.
6. Retrieve the data using queries and perform analysis (using ggplot).
7. Create a linear regression model to find the linear relation between the variables.

3.1 Scraping the Dataset

We used the WebScraper chrome extension to perform scraping. The WebScraper provides a cohesive node selection process and allows the user to explore a link on a page to get data inside the link. This feature was not available in other tools like import.io, grepsr, dexi.io and we could not use the *rvest* package as we had to traverse through multiple links and it would be uneconomical to store each link and perform extraction. WebScraper provides an interesting option to browse the data while the scraper is running. This allowed us to pause the scraping process if there were discrepancies and check the current movie that is scraping. It also provides the option of downloading the scraped data as a structured CSV file.



The Black Hole (1979)

PG | 1h 38min | Action, Sci-Fi | 21 December 1979 (USA)

5.9/10 19,292 Rate This

A research vessel finds a missing ship, commanded by a mysterious scientist, on the edge of a black hole.

Director: Gary Nelson

Writers: Jeb Rosebrook (story), Bob Barbash (story) | 3 more credits »

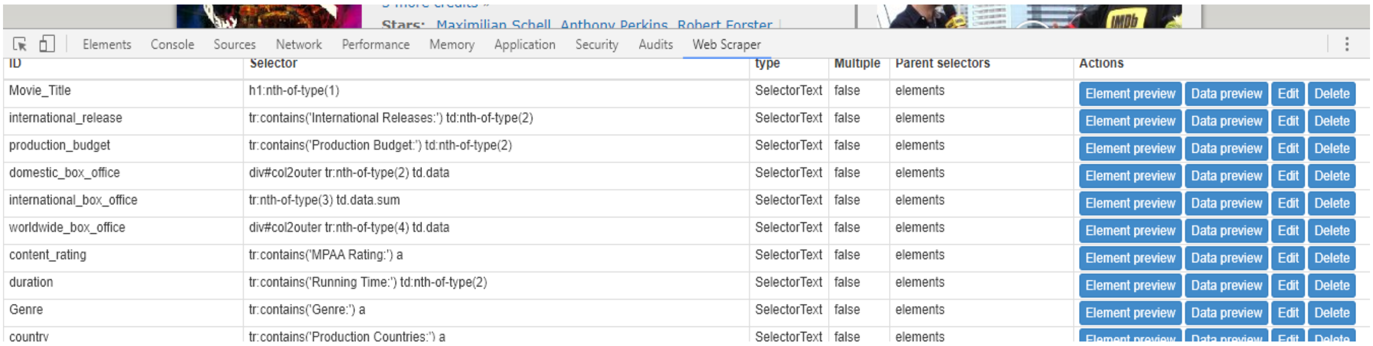
Stars: Maximilian Schell, Anthony Perkins, Robert Forster |

participate in our short IMDb survey

Chris Hardwick Breaks Down the 'Justice League' Panel

ID	Selector	type	Multiple	Parent selectors	Actions
links	td:nth-of-type(3) a	SelectorLink	true	_root	Element preview Data preview Edit Delete

Fig 3.1.1: Selecting nodes to be scraped



Stars: Maximilian Schell, Anthony Perkins, Robert Forster

ID	Selector	type	Multiple	Parent selectors	Actions
Movie_Title	h1:nth-of-type(1)	SelectorText	false	elements	Element preview Data preview Edit Delete
international_release	tr:contains('International Releases:') td:nth-of-type(2)	SelectorText	false	elements	Element preview Data preview Edit Delete
production_budget	tr:contains('Production Budget:') td:nth-of-type(2)	SelectorText	false	elements	Element preview Data preview Edit Delete
domestic_box_office	div#col2outer tr:nth-of-type(2) td.data	SelectorText	false	elements	Element preview Data preview Edit Delete
international_box_office	tr:nth-of-type(3) td.data.sum	SelectorText	false	elements	Element preview Data preview Edit Delete
worldwide_box_office	div#col2outer tr:nth-of-type(4) td.data	SelectorText	false	elements	Element preview Data preview Edit Delete
content_rating	tr:contains('MPAA Rating:') a	SelectorText	false	elements	Element preview Data preview Edit Delete
duration	tr:contains('Running Time:') td:nth-of-type(2)	SelectorText	false	elements	Element preview Data preview Edit Delete
Genre	tr:contains('Genre:') a	SelectorText	false	elements	Element preview Data preview Edit Delete
country	tr:contains('Production Countries:') a	SelectorText	false	elements	Element preview Data preview Edit Delete

Fig 3.1.2: Elements created after selecting nodes for each link.

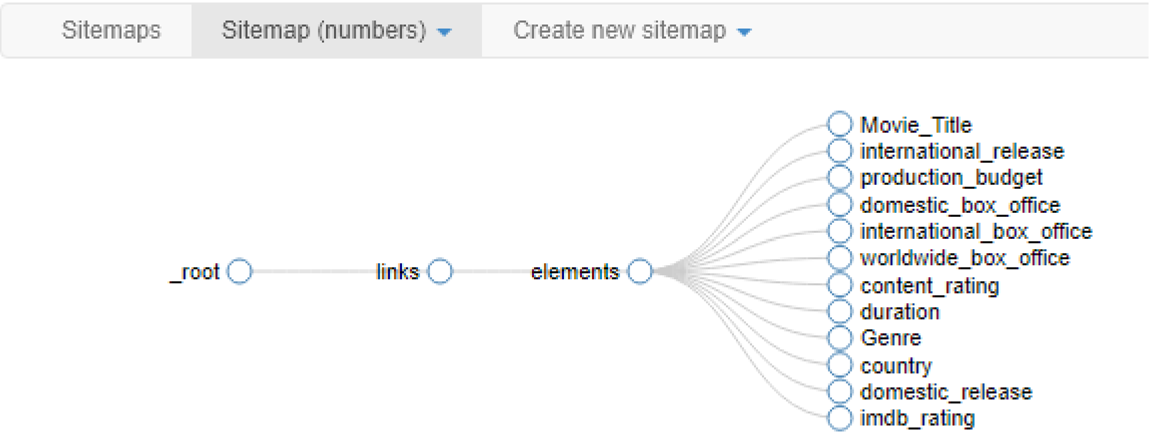
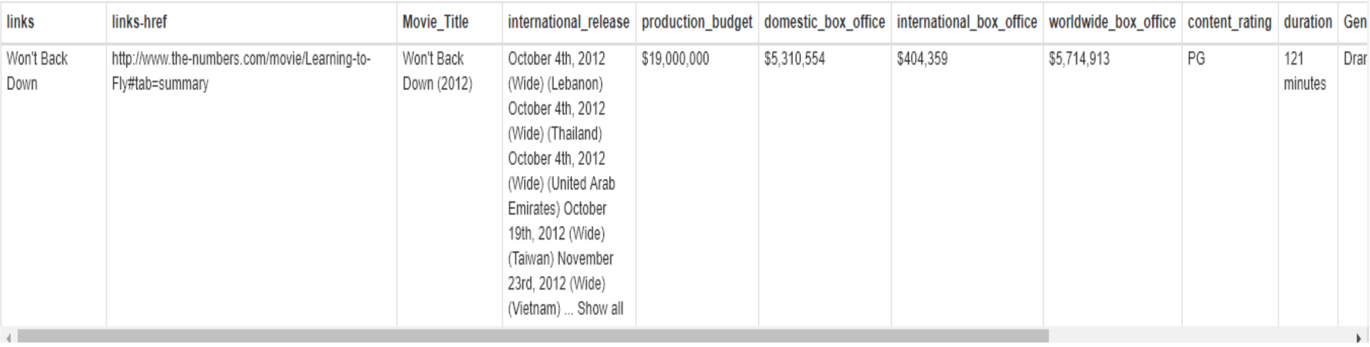


Fig 3.1.3: Selector graph displaying the schema of the scraping process

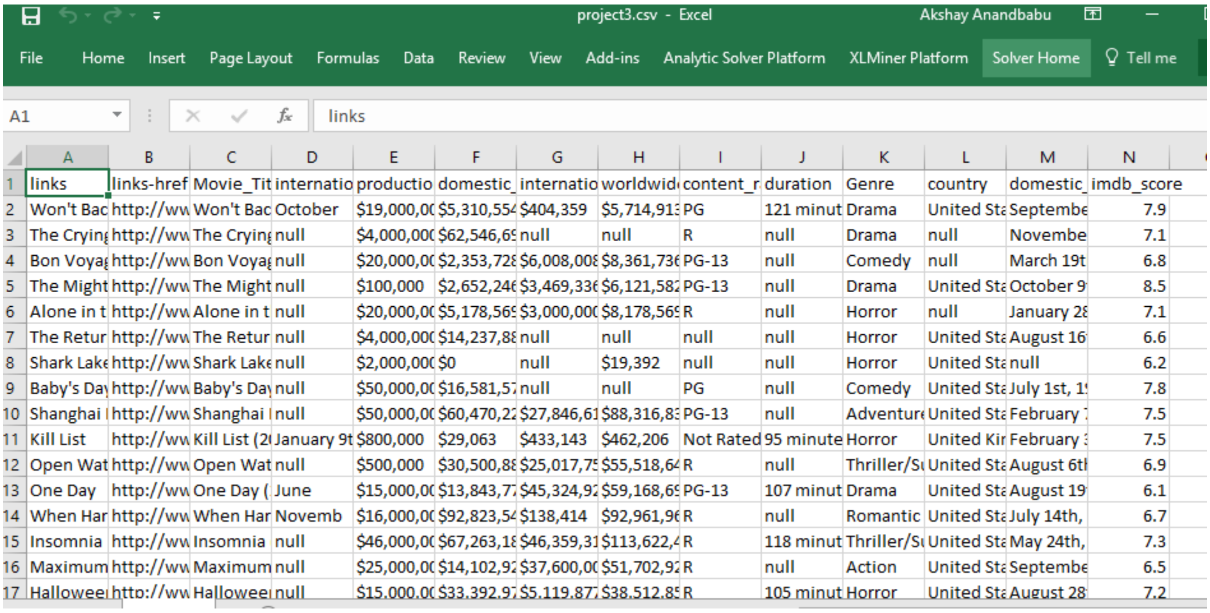


links	links-href	Movie_Title	international_release	production_budget	domestic_box_office	international_box_office	worldwide_box_office	content_rating	duration	Gen
Won't Back Down	http://www.the-numbers.com/movie/Learning-to-Fly#tab=summary	Won't Back Down (2012)	October 4th, 2012 (Wide) (Lebanon) October 4th, 2012 (Wide) (Thailand) October 4th, 2012 (Wide) (United Arab Emirates) October 19th, 2012 (Wide) (Taiwan) November 23rd, 2012 (Wide) (Vietnam) ... Show all	$19,000,000	$5,310,554	$404,359	$5,714,913	PG	121 minutes	Drar

Fig 3.1.4: Preview of the final data

3.2 Importing the CSV file into the R environment

Once the scraping process was completed and downloaded as a CSV file, we imported the file into the R environment using the read_csv() function and stored it in a suitable object as a data frame.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	links	links-href	Movie_Tit	internatio	productio	domestic_	internatio	worldwid	content_r	duration	Genre	country	domestic_	imdb_score
2	Won't Bac	http://ww	Won't Bac	October	$19,000,0C	$5,310,55	$404,359	$5,714,91	PG	121 minut	Drama	United Sta	Septembe	7.9
3	The Crying	http://ww	The Crying	null	$4,000,00C	$62,546,6	null	null	R	null	Drama	null	Novembe	7.1
4	Bon Voyag	http://ww	Bon Voyag	null	$20,000,0C	$2,353,72	$6,008,00	$8,361,73	PG-13	null	Comedy	null	March 19t	6.8
5	The Might	http://ww	The Might	null	$100,000	$2,652,24	$3,469,33	$6,121,58	PG-13	null	Drama	United Sta	October 9	8.5
6	Alone in t	http://ww	Alone in t	null	$20,000,0C	$5,178,56	$3,000,00C	$8,178,56	R	null	Horror	null	January 2	7.1
7	The Retur	http://ww	The Retur	null	$4,000,00C	$14,237,8	null	null	null	null	Horror	United Sta	August 16	6.6
8	Shark Lake	http://ww	Shark Lake	null	$2,000,00C	$0	null	$19,392	null	null	Horror	United Sta	null	6.2
9	Baby's Day	http://ww	Baby's Day	null	$50,000,0C	$16,581,5	null	null	PG	null	Comedy	United Sta	July 1st, 1	7.8
10	Shanghai	http://ww	Shanghai	null	$50,000,0C	$60,470,2	$27,846,61	$88,316,8	PG-13	null	Adventure	United Sta	February	7.5
11	Kill List	http://ww	Kill List (2	January 9t	$800,000	$29,063	$433,143	$462,206	Not Rated	95 minute	Horror	United Kir	February	7.5
12	Open Wat	http://ww	Open Wat	null	$500,000	$30,500,8	$25,017,7	$55,518,6	R	null	Thriller/S	United Sta	August 6t	6.9
13	One Day	http://ww	One Day (	June	$15,000,0C	$13,843,7	$45,324,9	$59,168,6	PG-13	107 minut	Drama	United Sta	August 19	6.1
14	When Har	http://ww	When Har	Novemb	$16,000,0C	$92,823,5	$138,414	$92,961,9	R	null	Romantic	United Sta	July 14th,	6.7
15	Insomnia	http://ww	Insomnia	null	$46,000,0C	$67,263,1	$46,359,31	$113,622,	R	118 minut	Thriller/S	United Sta	May 24th,	7.3
16	Maximum	http://ww	Maximum	null	$25,000,0C	$14,102,9	$37,600,0C	$51,702,9	R	null	Action	United Sta	Septembe	6.5
17	Halloweel	http://ww	Halloweel	null	$15,000,0C	$33,392,9	$5,119,87	$38,512,8	R	105 minut	Horror	United Sta	August 28	7.2

Fig 3.2: Final Scraped CSV file

3.3 Cleaning the Data

- The movie title included the year of release and since we have release dates in separate columns, it was unnecessary to have the year of release along with the movie name. We used the gsub() function and general expressions to remove the movie year.
- The international and domestic release dates were not in the default R date-format so we converted them using the mdy() function in the lubridate package. The release dates also contained the place of release inside quotes that had to be removed, so we used the str_extract() function with general expressions to extract the date. We stored the clean variables in the same table.
- Cleaning the budget and box office columns were fairly uncomplicated as the values for production budget and box office collections for movies that were released outside the united states were already converted into USD. We cleaned the columns by removing the dollar sign "$" and comma "," from each row and stored it as a numeric to perform analysis.
- We converted the duration column into a numeric and extracted the number of minutes, removing unnecessary characters using general expressions.

3.4 Storing the clean data into SQLLite

We used ***dbConnect*** to create a connection R to SQLLite and created a database named "AR_project.sqlite". The clean dataset was written inside this database using the ***dbWriteTable*** *query.*

3.5 Retrieving the data and performing analysis

The clean data was retrieved using ***dbGetQuery*** and the required variables were selected using the select query in SQLLite. Our analysis consists of three parts distribution, econometric analysis and linear regression modelling.

- In our distribution analysis, we check for consistencies in the data and check if there are certain variables that skew the results. Using ggplot to create a histogram, we found that the distribution of IMDb ratings follow a near normal distribution with the mean lying around the 6.5-7.5 range. We also created a distribution plot for content ratings. From the plot, we see that movies that are rated R occur the highest in the data-set. Since the data set is a list of movies with highest box office collection, we can make a possible conclusion that R rated movies fare well in box office collections.

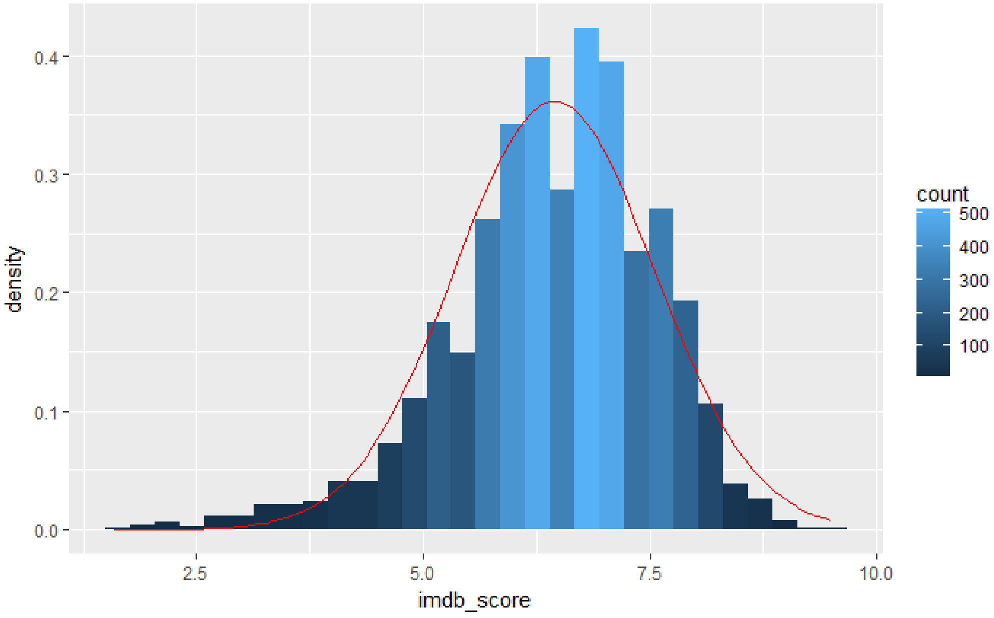


Fig 3.5.1: Normal distribution of IMDb ratings

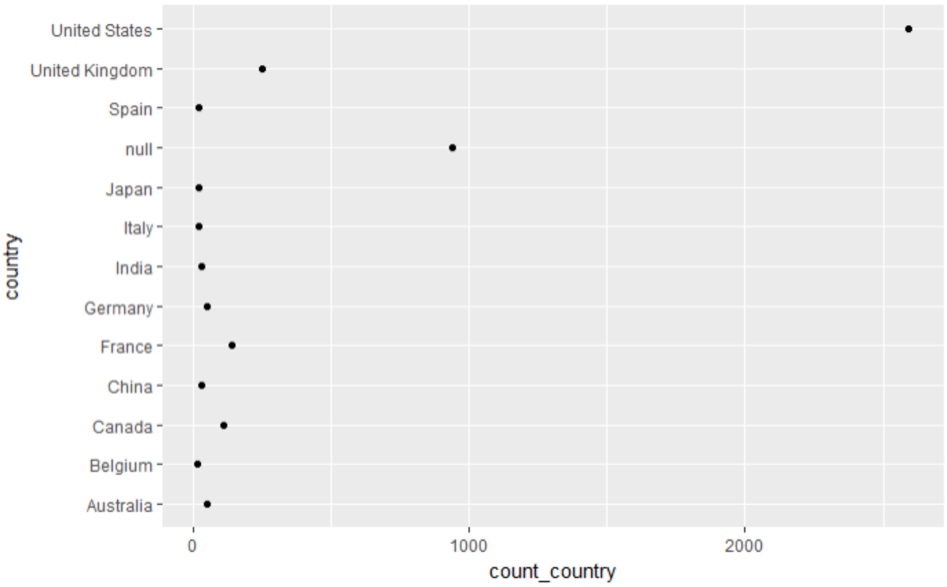


Fig 3.5.2: Scatterplot showing the number of movies released in each country

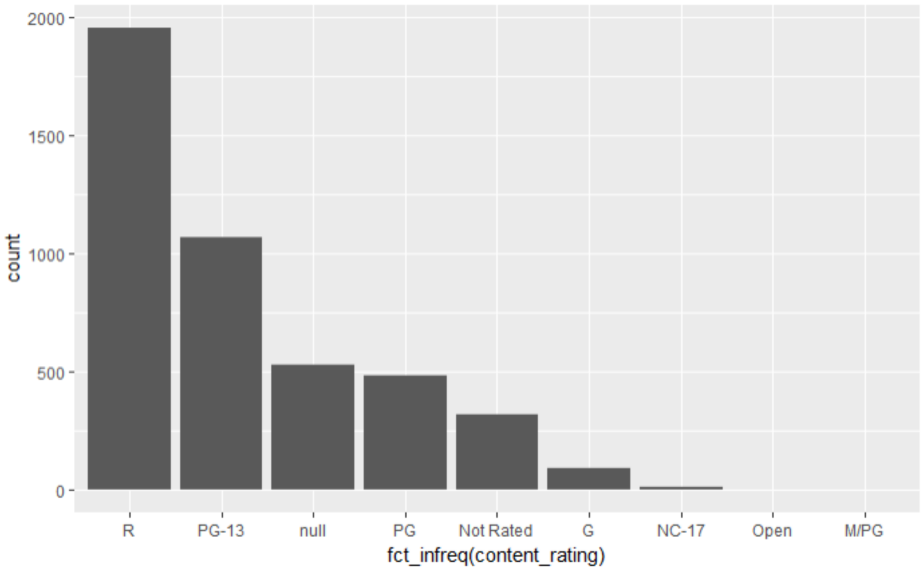


Fig 3.5.3 Barplot showing the number of movies based on content ratings

- Our econometric analysis involves finding the effect of IMDb rating, genre on the worldwide box office. We used a scatterplot and a smoothening line to find a reliable trend. A boxplot was created to see the effects of genre on the worldwide box office. We decided a boxplot would provide an ideal visualisation as it would provide the range of income of each genre and the median line to assist in finding trends.

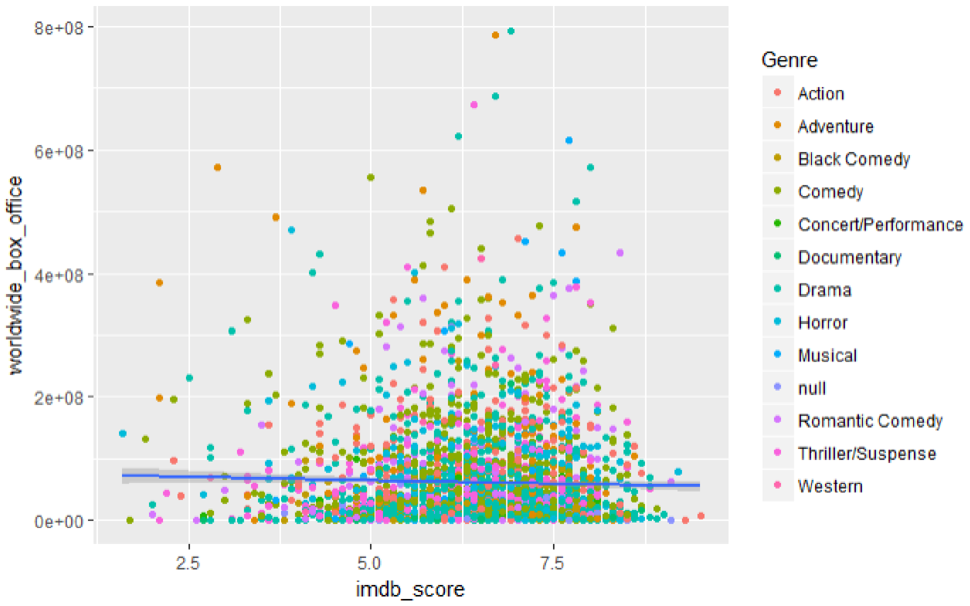


Fig 3.5.4: Scatterplot showing the trend between IMDb rating and worldwide box office collection.

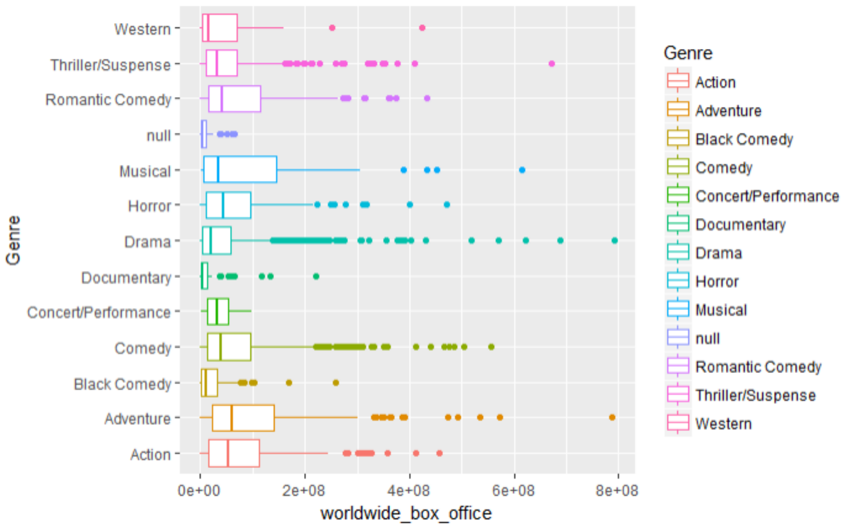
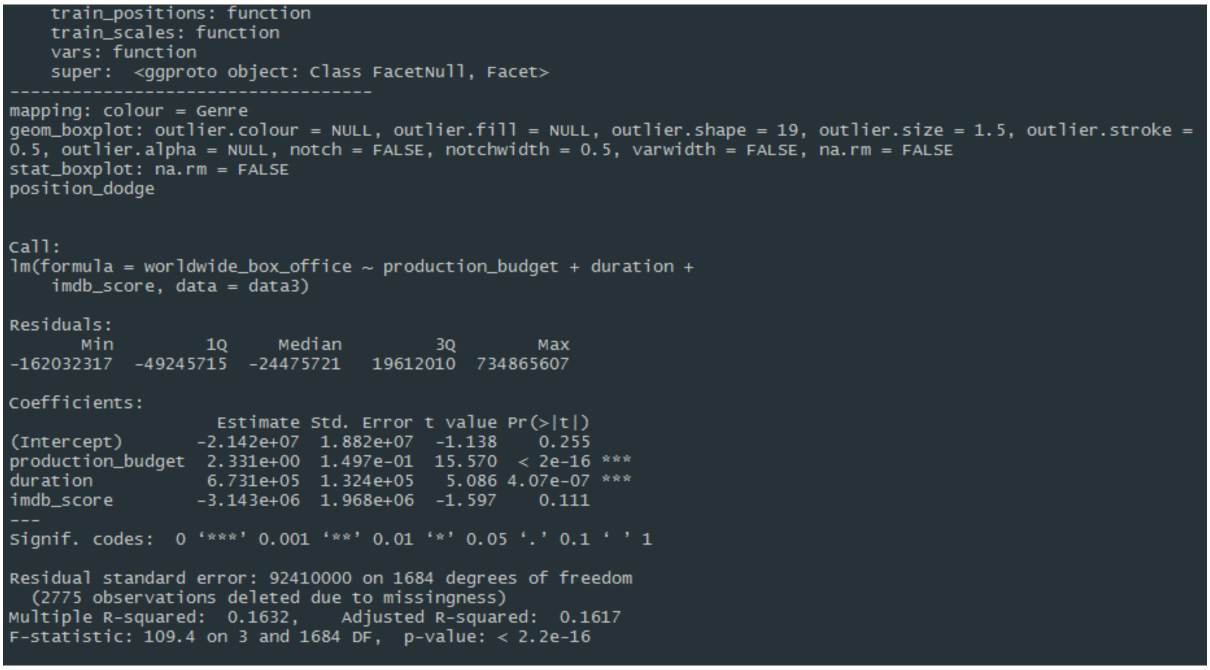


Fig 3.5.5 Boxplot showing the trend between Genre and worldwide box office

- Our linear model finds a linear relation between the dependent variable(worldwide box office)and independent variables (IMDb rating, production budget, domestic box office, international box office). Through our model we find the correlation between these variables and the coefficient of determination which is given by the r-squared value.



```
    train_positions: function
    train_scales: function
    vars: function
    super:  <ggproto object: Class FacetNull, Facet>
-----------------------------------
mapping: colour = Genre
geom_boxplot: outlier.colour = NULL, outlier.fill = NULL, outlier.shape = 19, outlier.size = 1.5, outlier.stroke =
0.5, outlier.alpha = NULL, notch = FALSE, notchwidth = 0.5, varwidth = FALSE, na.rm = FALSE
stat_boxplot: na.rm = FALSE
position_dodge


Call:
lm(formula = worldwide_box_office ~ production_budget + duration +
    imdb_score, data = data3)

Residuals:
       Min         1Q     Median         3Q        Max
-162032317  -49245715  -24475721   19612010  734865607

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.142e+07  1.882e+07  -1.138    0.255
production_budget  2.331e+00  1.497e-01  15.570  < 2e-16 ***
duration           6.731e+05  1.324e+05   5.086 4.07e-07 ***
imdb_score        -3.143e+06  1.968e+06  -1.597    0.111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92410000 on 1684 degrees of freedom
  (2775 observations deleted due to missingness)
Multiple R-squared:  0.1632,    Adjusted R-squared:  0.1617
F-statistic: 109.4 on 3 and 1684 DF,  p-value: < 2.2e-16
```

Fig 3.5.6 R output displaying the results of the linear model

4. Complications faced during the process

- **Long run times of WebScraper:**

Web-Scraper requires a long time to scrape data because it prints the results of each movie before moving on to the next movie. The scraper has to go through each link, scrape the data and then go back to the start link and repeat the process for each movie. The entire scraping process ran for upto 6 hours, during which we had to leave our computers on or it ended up in pausing the scrape.

- **Multiple international dates:**

  The data scraped has multiple international release dates for each country that was released globally though the release dates for each country were about equal. We solved this problem by extracting the first release date by using ***str_extract().***

- **Cleaning Process:**

  There were a few intricate problems we faced while cleaning the scraped data set. The majority of the data types that we encountered were factors so we could not perform any string operations on it. We used ***lapply[]*** to change the variables into characters and then performed string operations on the variables to clean them. We also had to convert the world-wide box office collection given in dollars to numeric befor using it for analysis.

5. Learning Outcomes and Future Work:

- Web Scraping using web-tools.
- Cleaning and formatting data using general expressions.
- Storing data in a database using SQLLite.
- Retrieving data from the database using queries.
- Creating plots and learning ways to make the plot look more informative using ***stat_function().***
- Creating a linear regression model to find correlations between variables.

Future work for this project would be to create and implement an algorithm which would predict the overall box office collection of a movie by knowing its genre, production budget and the domestic box office collection. The project can be expanded by performing a sentiment analysis on movie reviews and see the effect of movie reviews on the rating of a movie which would in turn affect the box office collection.

6. References

- http://www.the-numbers.com/movie/budgets/all
- https://stackoverflow.com/questions/
- https://www.google.com
- http://webscraper.io/