
Northeastern University

College of Professional Studies, **Toronto**

# Boxoffice Revenue Analysis

Guided By:
**Dr.Prof. M. Shafiqul Islam**

Presentation By:
**Manan Soni**
Nuid: **002982645**

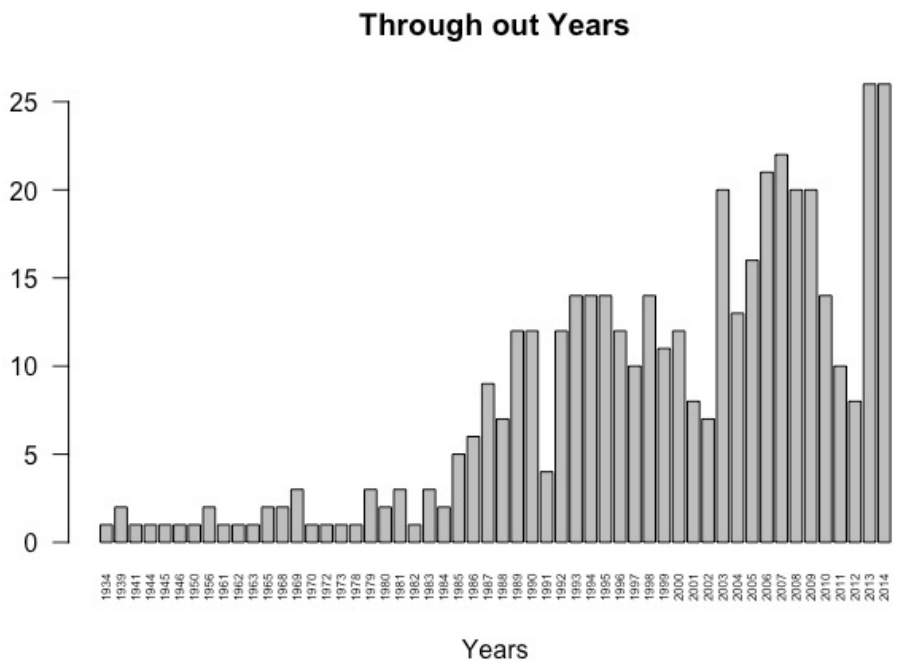
# Data Source

- The Data is from a repository of **fivethirtyeight** (https://github.com/fivethirtyeight)
- Data is of **Box Office Collection of Movies that are Biographies of personalities** .
- Timeline of Data is from **1940 to 2015** in which data before 2000 is quite noisy. By noisy it has **missing values**. To add to that, Data has redundancy as **duplication** of samples.
- Actual data coming from is IMDB so **IMDB is the actual source of the data**.
- After asking to Data, there are surprising facts.

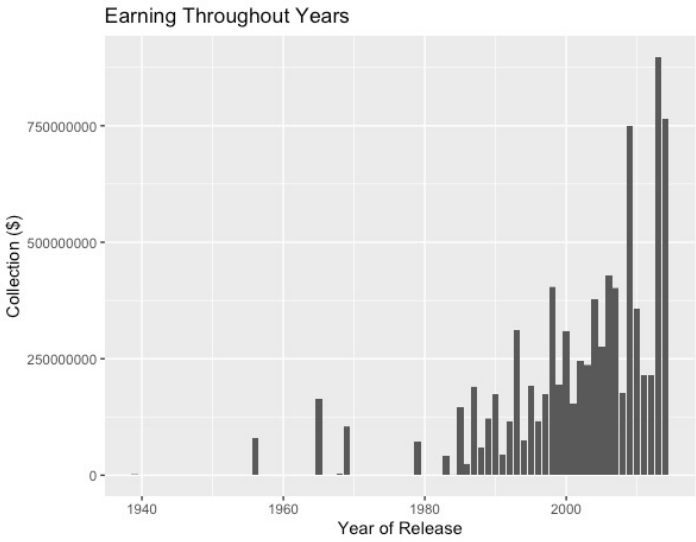
# Processing of Data

- Correlation of year and revenue is not relative as value is near to 0.
- We can’t guess the revenue so removing samples was only options to past movies.

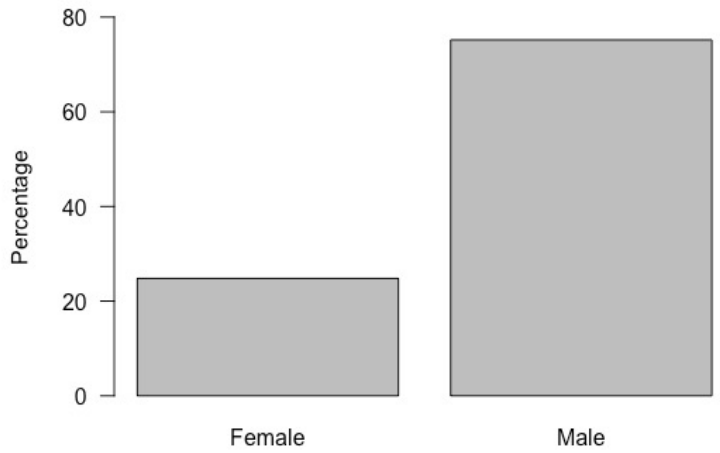
# How much earning through out years




- As seen trend is going up through out years and some falls can be seen as well.
- Movies earned well after 2008 as we can see in the bar plot.

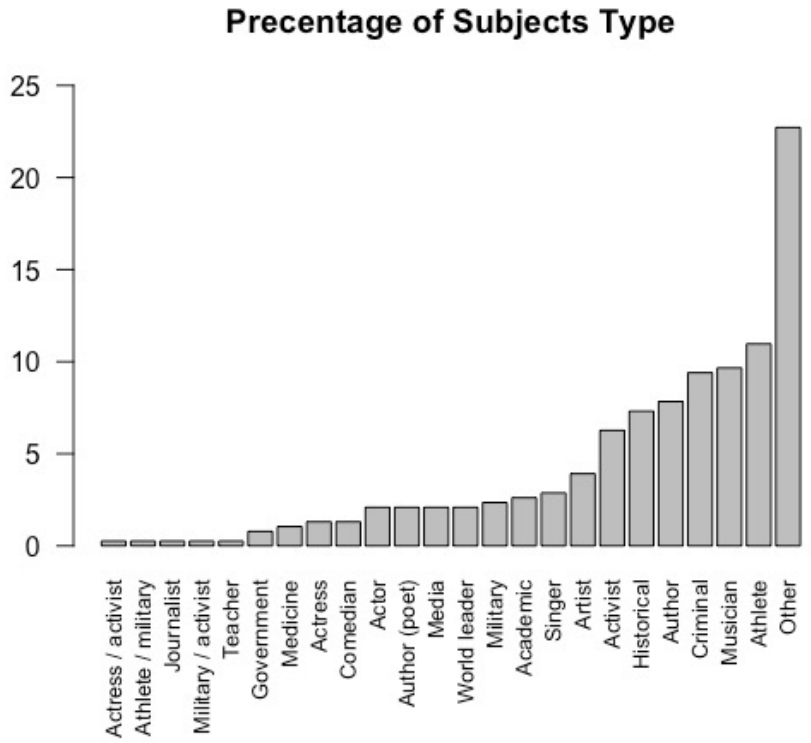


# Subject Sex affects **BOR**



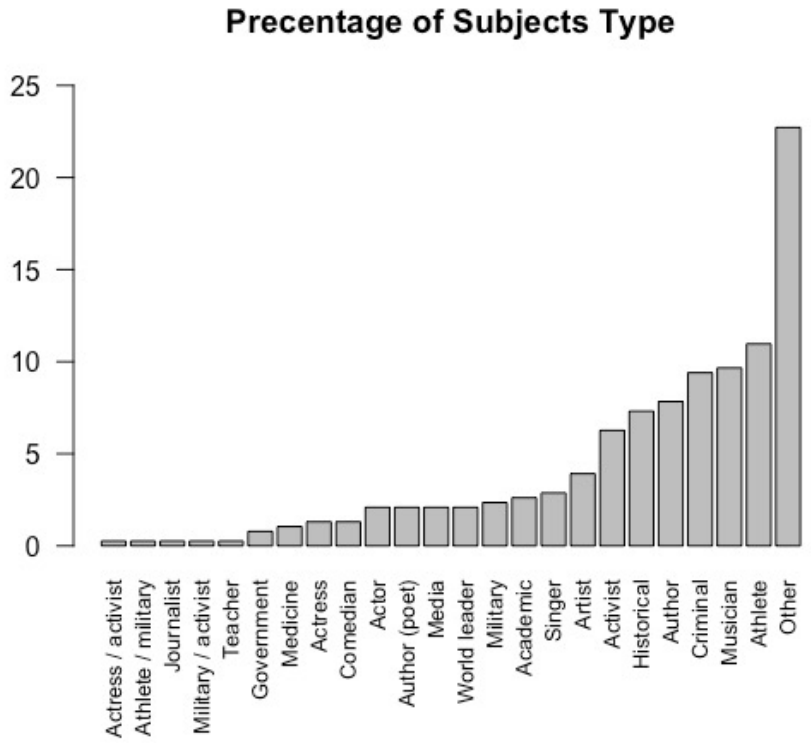

- We can see that movie made on male have 75% chances to earn more than movie made on women

# Subject Type affects **BOR**



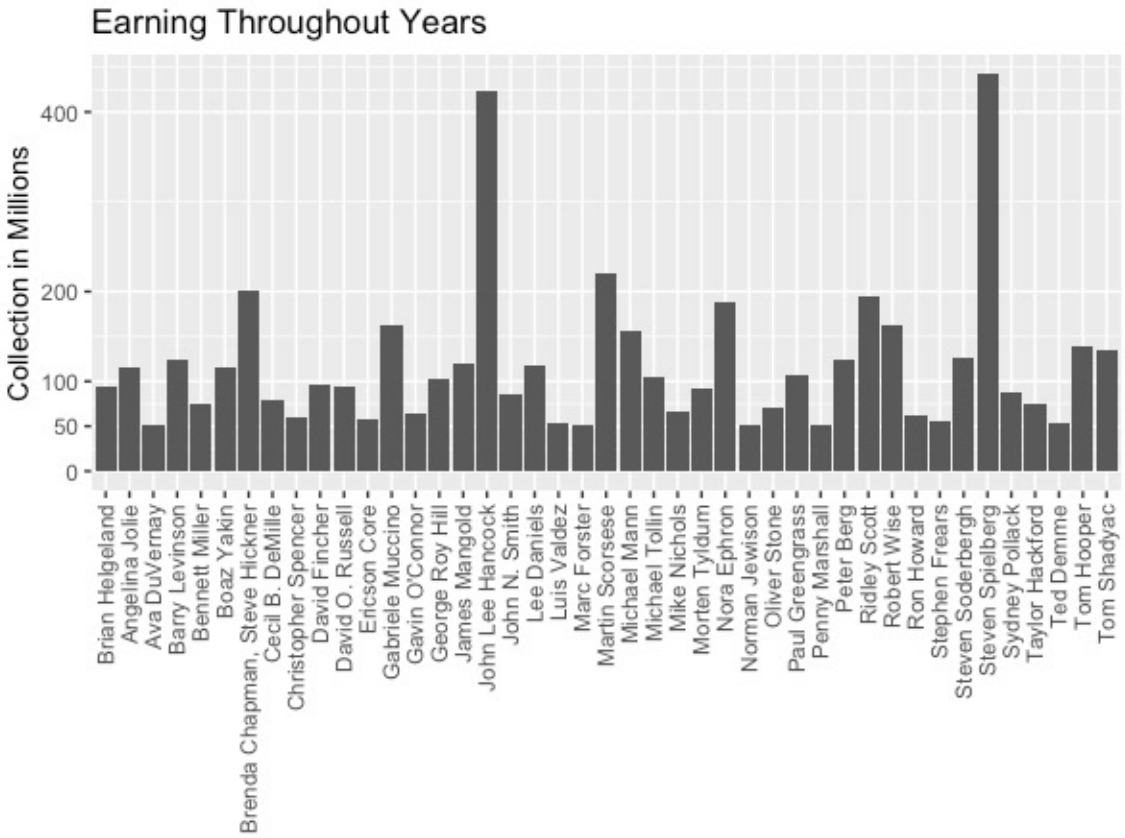

- 80% of the movies have categorized in **other.**
- Movies made on actresses/activists have lowest percentage.
- We can say people are more interested in Historical, based on author, Crime related, based on athlete.

# Subject Type for **BOR**



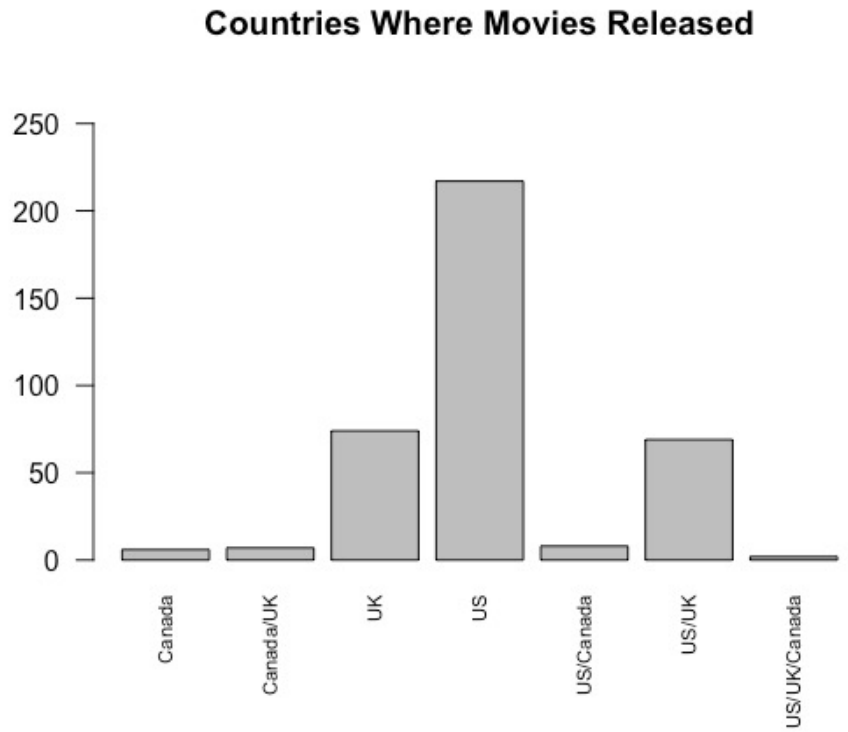

- 80% of the movies have categorized in **other.**
- Movies made on actresses/activists have lowest percentage.
- We can say people are more interested in Historical, based on author, Crime related, based on athlete.

# Directors for **BOR**



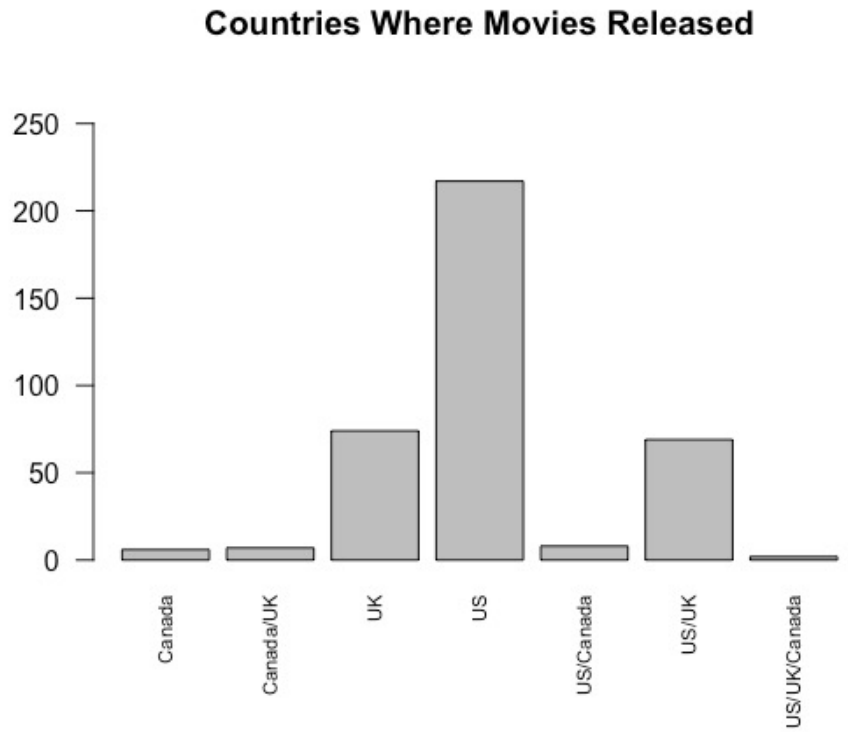

- These are the directors who were successfully captured audience through their direction.
- Steven Spielberg and John Lee Hancock have earned highest amongst all.

# Country who played good role in **BOR**



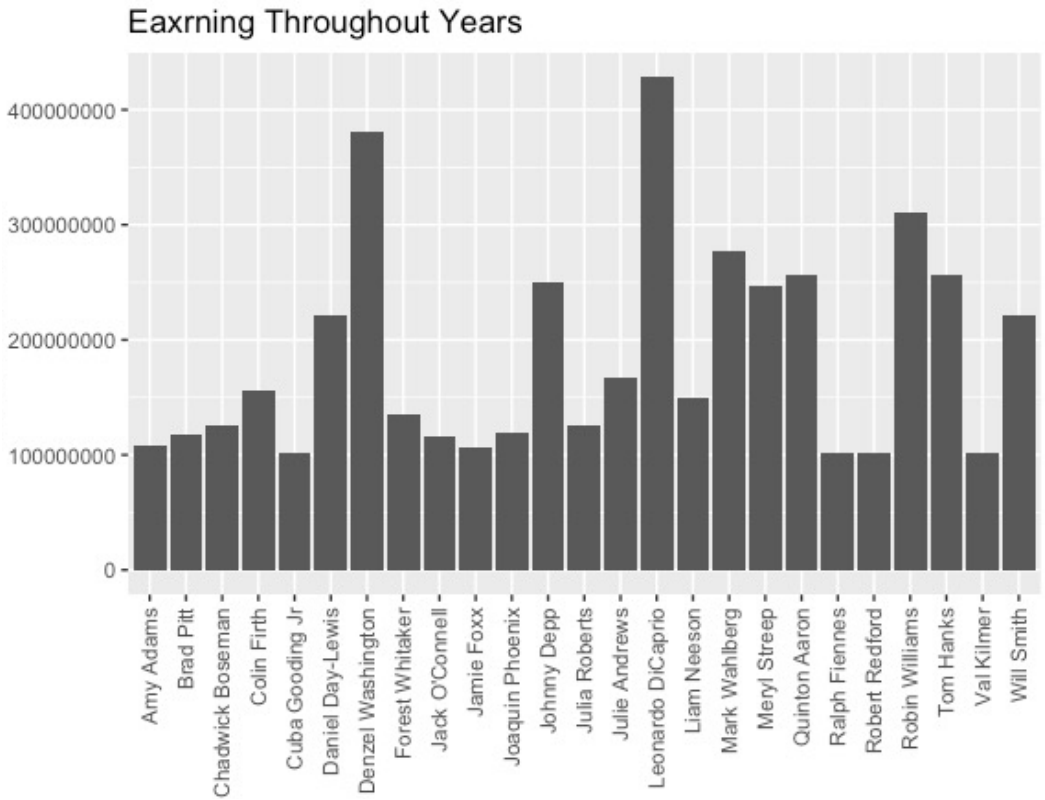

- Us has the highest frequency. To add to that, 200 plus movies releases in US.

# Country who played good role in BOR



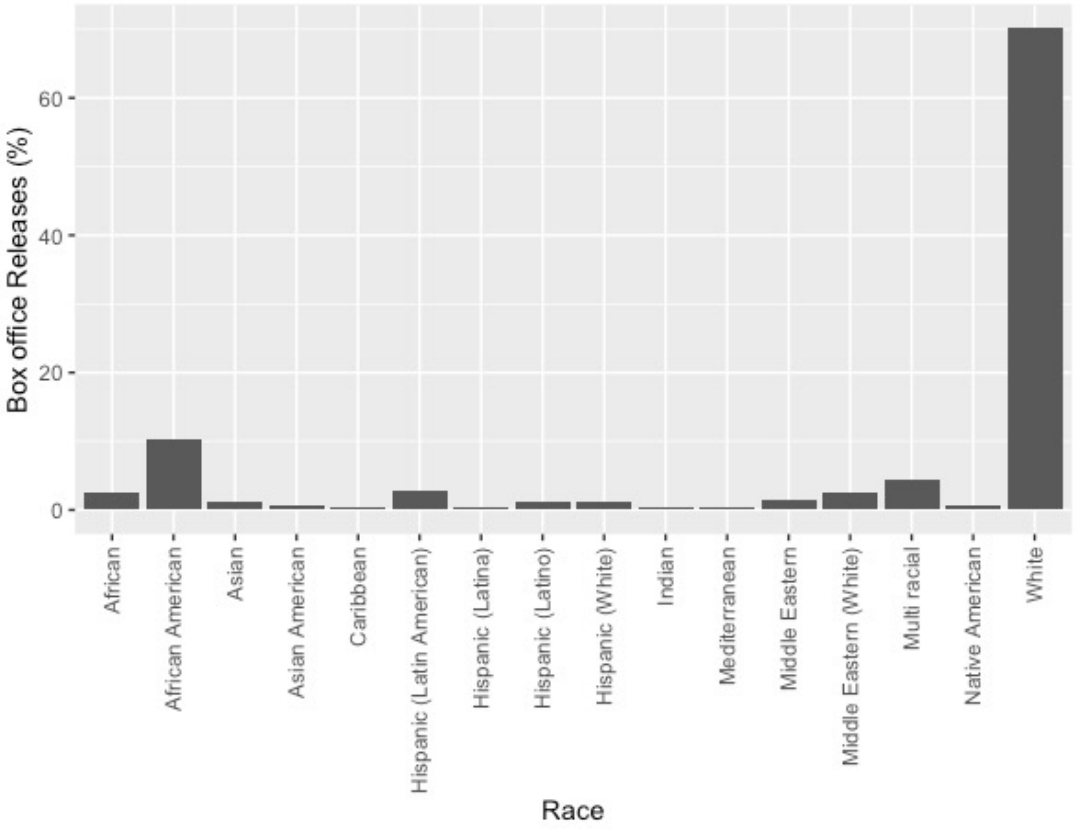

- Us has the highest frequency. To add to that, 200 plus movies releases in US.

# Lead Actors/Actresses for **BOR**




- This are the actor/actresses who made movies successful and get good revenue.
- Highest earner is female and best actresses of her time.

# If Race affects **BOR**




- Here, We can't say that people have interest in watching biography of white personalities.
- It's a possibilities that most of the movies which had released were of white and less movies of other races.

# Conclusion

- If a production house thinking to make a biography movie and to generate more revenue we need to get good director, actor (not actress), a famous personality of US, Subject can be vary and last we need to release it in US because most of the revenue will come from US.
- We can see that past movies made less money but todays inflation is 16% compared to 1960. So 1$ can be 16$ of todays.

# Additionals

- If dataset had gender of director, we can dig more like whom directed better.
- We can add a feature/variable like Multi country release, Single country release.
- If dataset had ratings of movie from imdb than accuracy of analysis can be increase.

# References


- FiveThirtyEight, IMDB:https://github.com/fivethirtyeight/data/tree/master/biopics
- Admin, & *, N. (n.d.). *12.1 million in numbers – 12.1 million written out*. Numbers in Words. Retrieved February 21, 2022, from https://numbersinwords.net/12-1-million-in-numbers
- Anand, *Group by a column and sort by another column in R*. Stack Overflow. Retrieved February 21, 2022, from https://stackoverflow.com/questions/39470731/group-by-a-column-and-sort-by-another-column-in-r
- Hadley, et al. “Subset rows using column values.” the Grammar of Data Manipulation • dplyr, RStudio, https://dplyr.tidyverse.org/index.html.

# Thank you