IMDB MOVIE REVIEW DETAILS DATASET

Team Names and ID:

Sai Rohith Pasala; 1001873156

Sri Naga Venkata Pavan Kalyan Sirigibattula; 1001886149

Rutvik Naga Sai Dondapati; 1001879177

INTRODUCTION:

Movie Review details dataset has been taken from https://www.kaggle.com/preetviradiya/imdb-movies-ratings-details. This data consists of all the details of Imdb details of movie ratings, reviews, votes etc.

Content:

Contains the information about the movie such as:

1. Name
2. Short storyline
3. Box-office Collection
4. IMDB ratings
5. IMDB votes
6. IMDB metascore

Acknowledgements

IMDB .

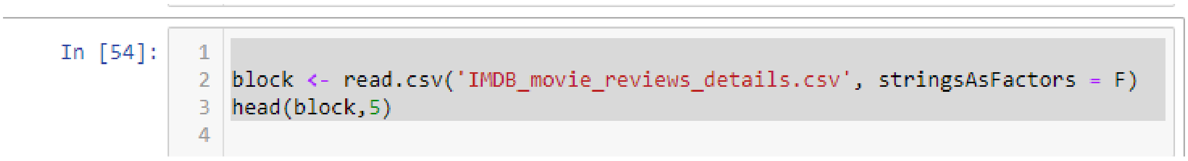
Retrieving the Data:

We will be using the R programming language in Anaconda / Rstudio to analyze this dataset.

R Programming Language:

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

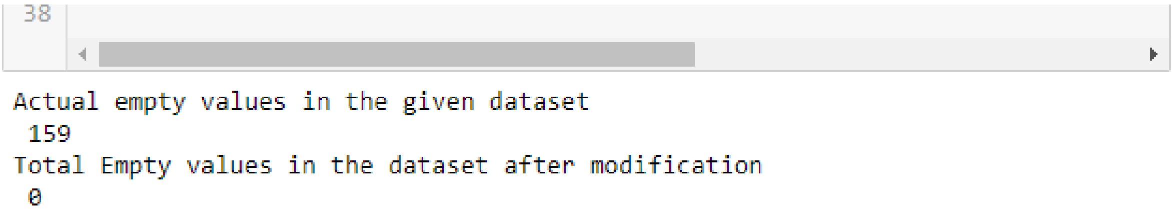
The dataset is imported to R notebook using R data frame named “block”. Following code explains about the retrieve the csv dataset file and print first 5 rows in the dataset.



```
In [54]:
block <- read.csv('IMDB_movie_reviews_details.csv', stringsAsFactors = F)
head(block,5)
```

Note:-
For execution of this file , the dataset 'IMDB_movie_reviews_details.csv' must be in the same folder as this ipynb file.

Preprocessing of Data:
The dataset given has many redundant values, noisy data, and null/empty values. We must clean them before proceeding to start visualizing/ analyzing the dataset.

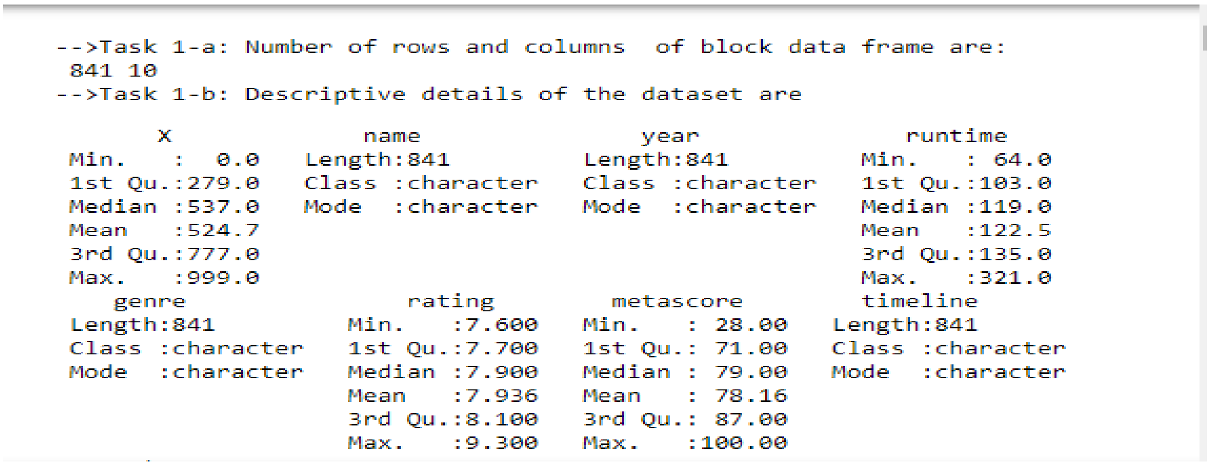


```
Actual empty values in the given dataset
 159
Total Empty values in the dataset after modification
 0
```

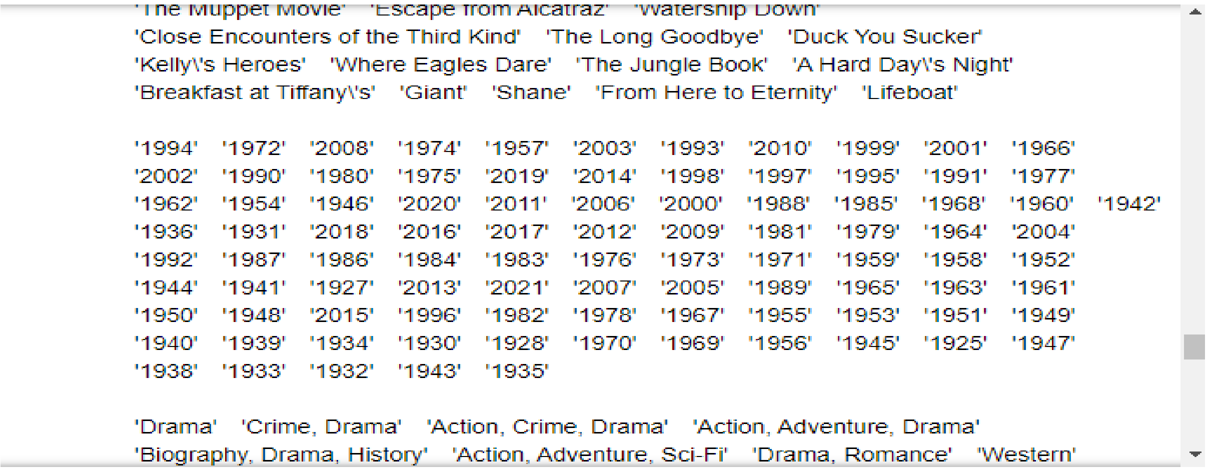
Data Exploration:

Task 1: Statistical Exploratory Data Analysis

In this we display number of rows and columns , print descriptive details, print unique values of movies , years, and genre of given dataset.



```
-->Task 1-a: Number of rows and columns  of block data frame are:
 841 10
-->Task 1-b: Descriptive details of the dataset are

       X               name               year              runtime
 Min.   :  0.0   Length:841         Length:841         Min.   : 64.0
 1st Qu.:279.0   Class :character   Class :character   1st Qu.:103.0
 Median :537.0   Mode  :character   Mode  :character   Median :119.0
 Mean   :524.7                                         Mean   :122.5
 3rd Qu.:777.0                                         3rd Qu.:135.0
 Max.   :999.0                                         Max.   :321.0
    genre               rating         metascore        timeline
 Length:841         Min.   :7.600   Min.   : 28.00   Length:841
 Class :character   1st Qu.:7.700   1st Qu.: 71.00   Class :character
 Mode  :character   Median :7.900   Median : 79.00   Mode  :character
                    Mean   :7.936   Mean   : 78.16
                    3rd Qu.:8.100   3rd Qu.: 87.00
                    Max.   :9.300   Max.   :100.00
```

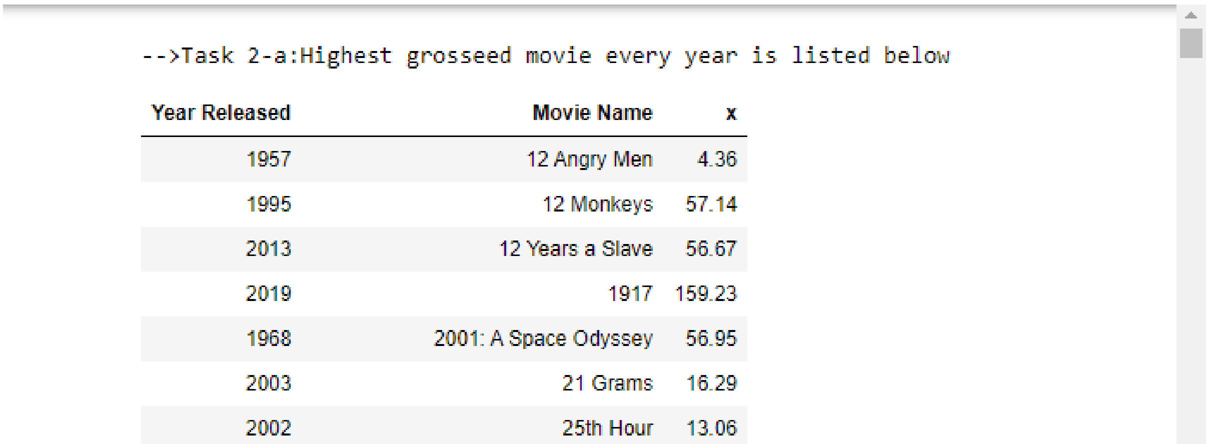


```
'The Muppet Movie'   'Escape from Alcatraz'   'Watership Down'
'Close Encounters of the Third Kind'   'The Long Goodbye'   'Duck You Sucker'
'Kelly\'s Heroes'   'Where Eagles Dare'   'The Jungle Book'   'A Hard Day\'s Night'
'Breakfast at Tiffany\'s'   'Giant'   'Shane'   'From Here to Eternity'   'Lifeboat'

'1994'  '1972'  '2008'  '1974'  '1957'  '2003'  '1993'  '2010'  '1999'  '2001'  '1966'
'2002'  '1990'  '1980'  '1975'  '2019'  '2014'  '1998'  '1997'  '1995'  '1991'  '1977'
'1962'  '1954'  '1946'  '2020'  '2011'  '2006'  '2000'  '1988'  '1985'  '1968'  '1960'  '1942'
'1936'  '1931'  '2018'  '2016'  '2017'  '2012'  '2009'  '1981'  '1979'  '1964'  '2004'
'1992'  '1987'  '1986'  '1984'  '1983'  '1976'  '1973'  '1971'  '1959'  '1958'  '1952'
'1944'  '1941'  '1927'  '2013'  '2021'  '2007'  '2005'  '1989'  '1965'  '1963'  '1961'
'1950'  '1948'  '2015'  '1996'  '1982'  '1978'  '1967'  '1955'  '1953'  '1951'  '1949'
'1940'  '1939'  '1934'  '1930'  '1928'  '1970'  '1969'  '1956'  '1945'  '1925'  '1947'
'1938'  '1933'  '1932'  '1943'  '1935'

'Drama'   'Crime, Drama'   'Action, Crime, Drama'   'Action, Adventure, Drama'
'Biography, Drama, History'   'Action, Adventure, Sci-Fi'   'Drama, Romance'   'Western'
```

Task 2: Aggregation and Filtering and rank

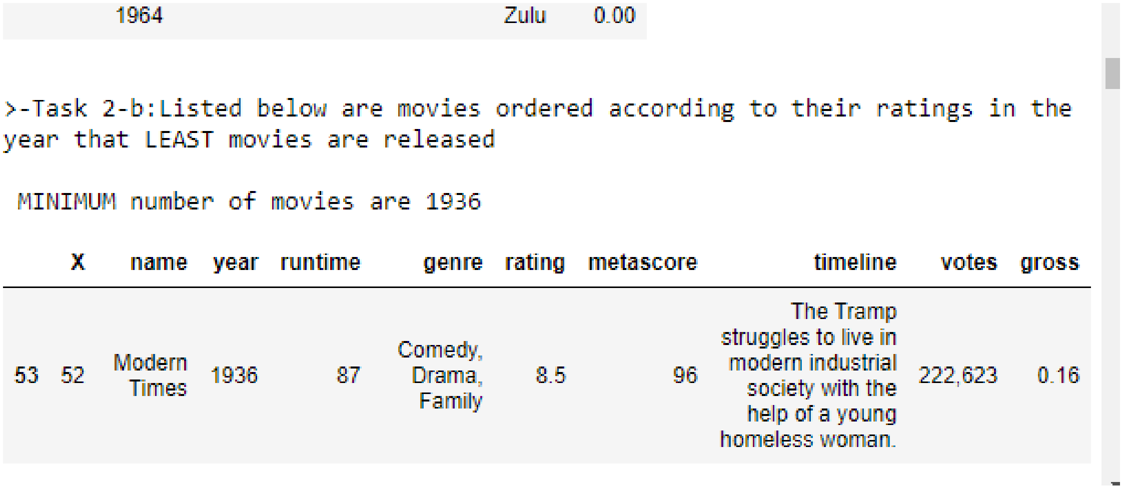
In this task , we will perform some very high-level aggregation and filtering operations. Then, we will apply ranking on the results for some tasks.

We estimate the highest gross money achieved every year, rank all movies based on rating in a year where minimum number of movies are released and maximum number of movies are released.



```
-->Task 2-a:Highest grosseed movie every year is listed below
```

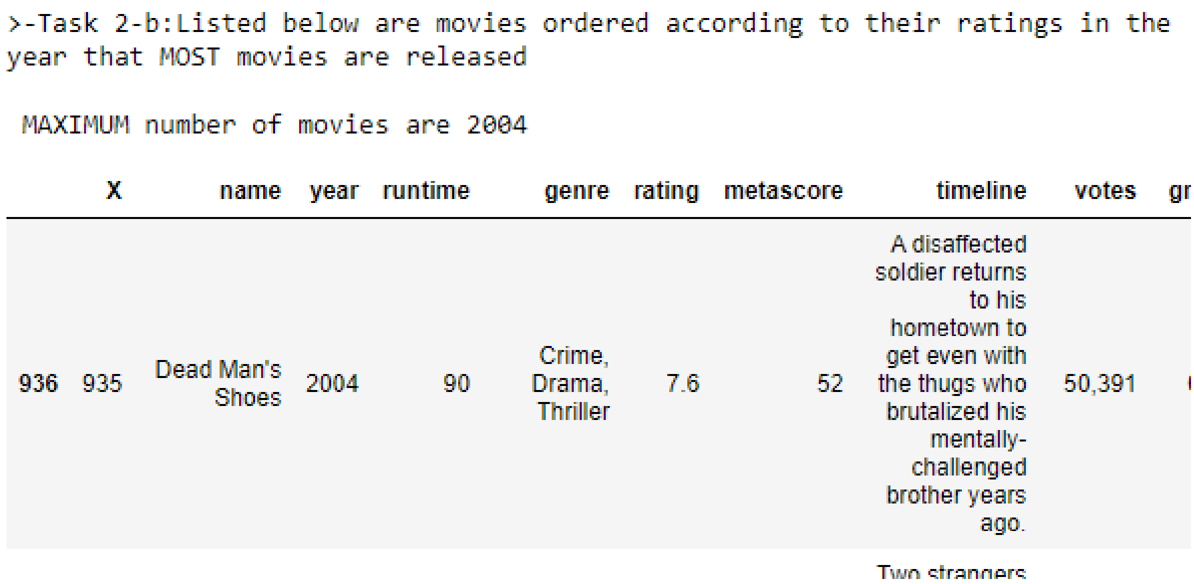
Year Released	Movie Name	x
1957	12 Angry Men	4.36
1995	12 Monkeys	57.14
2013	12 Years a Slave	56.67
2019	1917	159.23
1968	2001: A Space Odyssey	56.95
2003	21 Grams	16.29
2002	25th Hour	13.06



1964	Zulu	0.00

```
>-Task 2-b:Listed below are movies ordered according to their ratings in the
year that LEAST movies are released

 MINIMUM number of movies are 1936
```

	X	name	year	runtime	genre	rating	metascore	timeline	votes	gross
53	52	Modern Times	1936	87	Comedy, Drama, Family	8.5	96	The Tramp struggles to live in modern industrial society with the help of a young homeless woman.	222,623	0.16



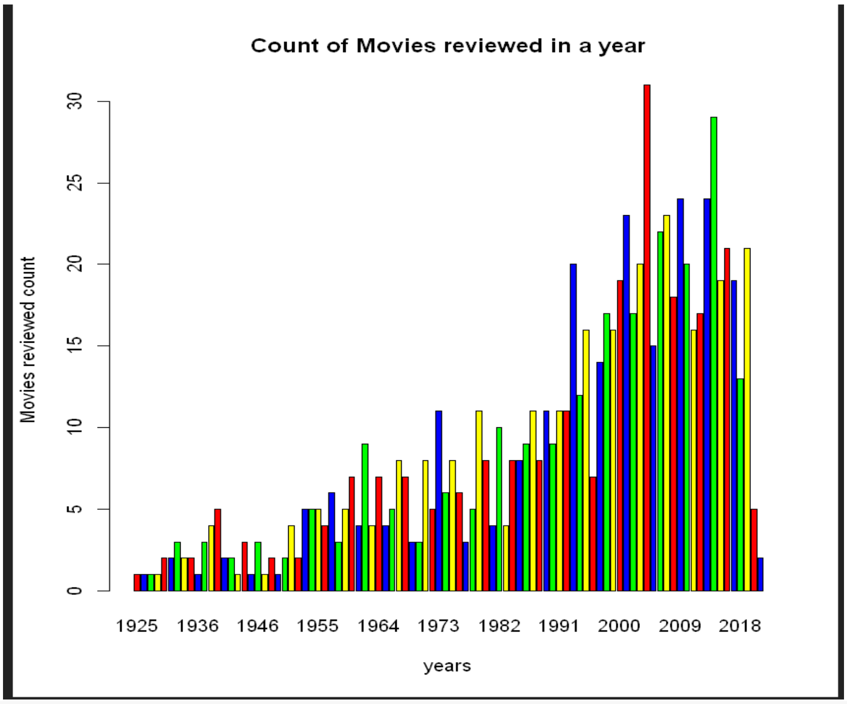
```
>-Task 2-b:Listed below are movies ordered according to their ratings in the
year that MOST movies are released

 MAXIMUM number of movies are 2004
```

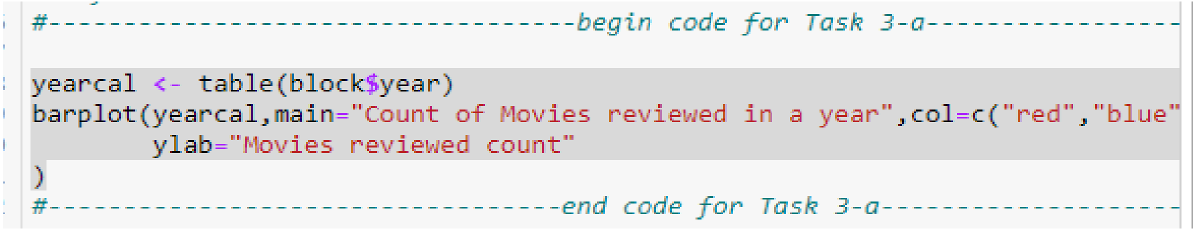
	X	name	year	runtime	genre	rating	metascore	timeline	votes	gr
936	935	Dead Man's Shoes	2004	90	Crime, Drama, Thriller	7.6	52	A disaffected soldier returns to his hometown to get even with the thugs who brutalized his mentally-challenged brother years ago.	50,391	(
								Two strangers		

Task 3: Visualization:

We can use different libraries like ggplot and tidyverse or use the Default R inbuilt statistical graphs syntax.

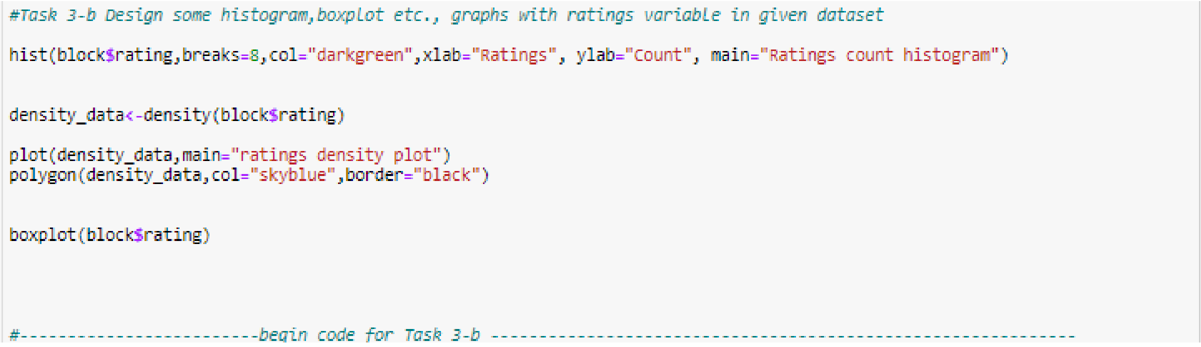
Plotting the number of reviews each year against bar graph:-



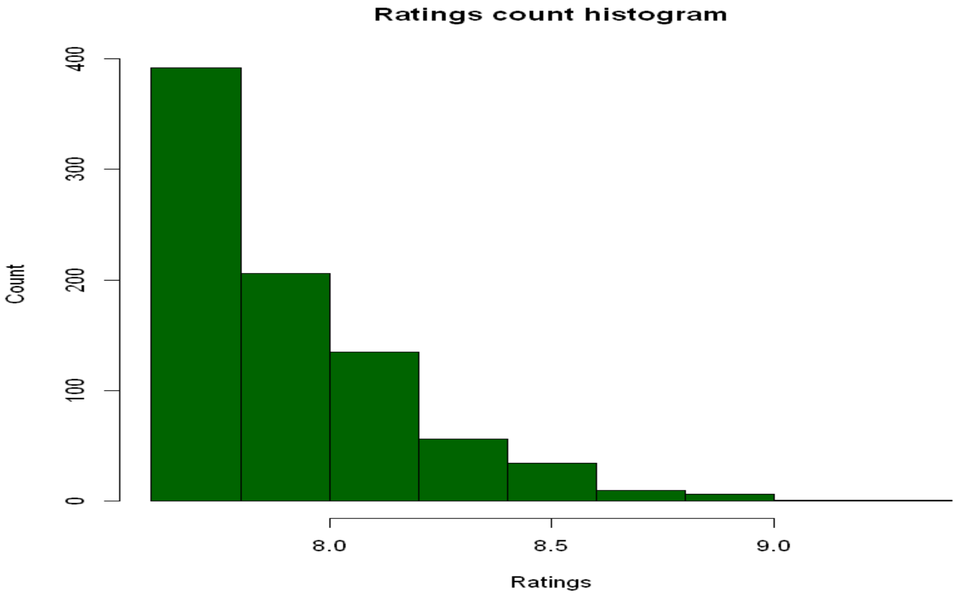


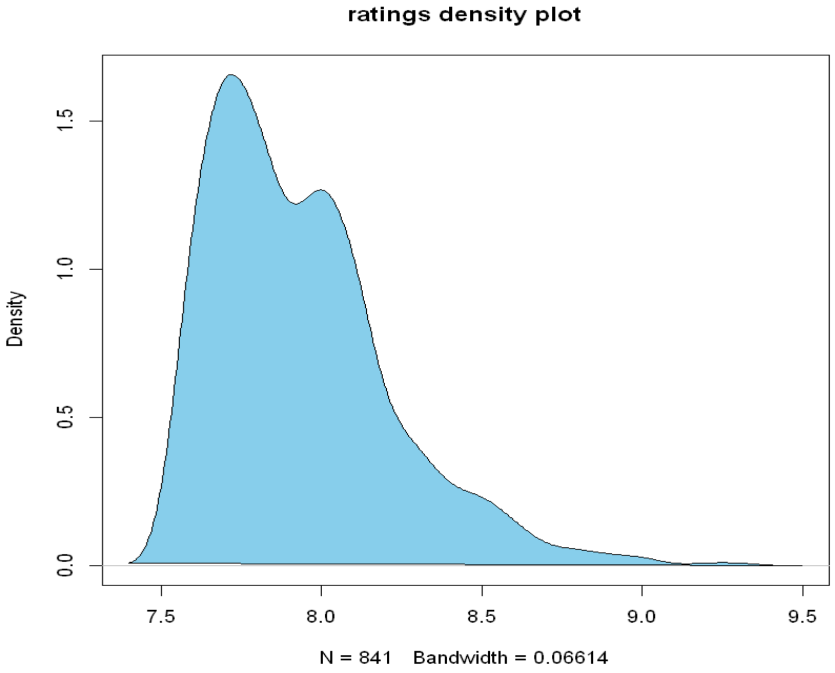
```
#----------------------------------begin code for Task 3-a----------------

yearcal <- table(block$year)
barplot(yearcal,main="Count of Movies reviewed in a year",col=c("red","blue"
        ylab="Movies reviewed count"
)
#---------------------------------end code for Task 3-a------------------
```

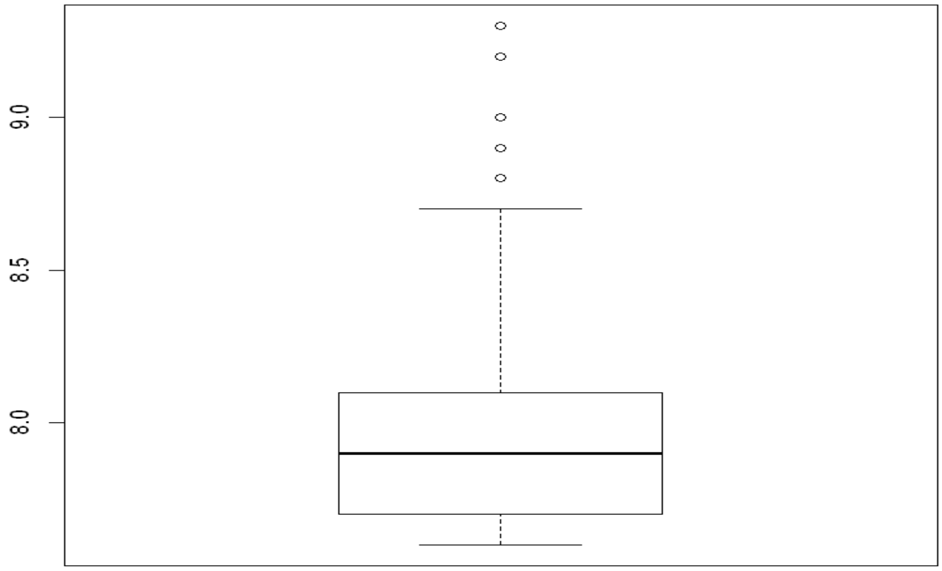
Plotting Histogram, Boxplot, etc. Graphs for ratings variable:



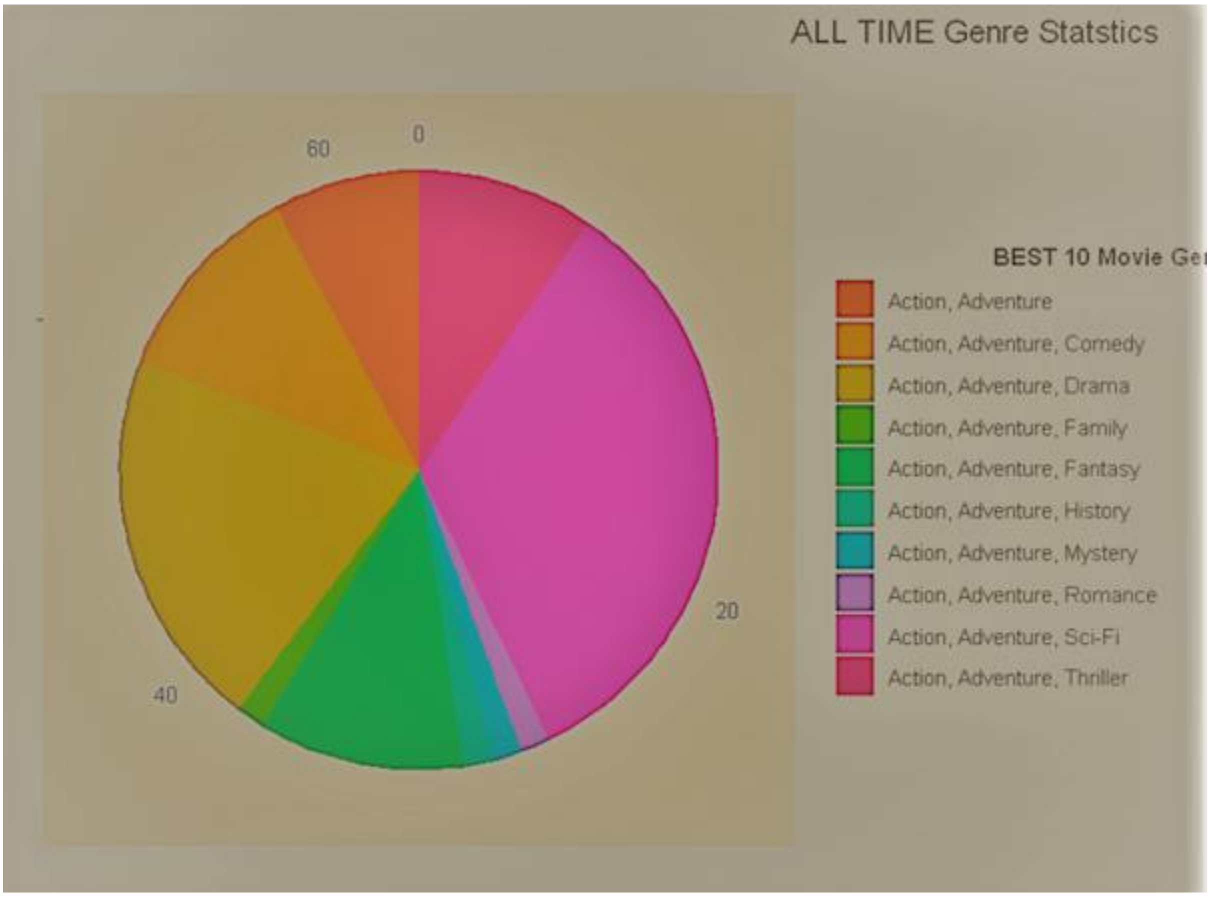
```
#Task 3-b Design some histogram,boxplot etc., graphs with ratings variable in given dataset
hist(block$rating,breaks=8,col="darkgreen",xlab="Ratings", ylab="Count", main="Ratings count histogram")


density_data<-density(block$rating)

plot(density_data,main="ratings density plot")
polygon(density_data,col="skyblue",border="black")


boxplot(block$rating)



#------------------------begin code for Task 3-b ------------------------------------------------------------
```





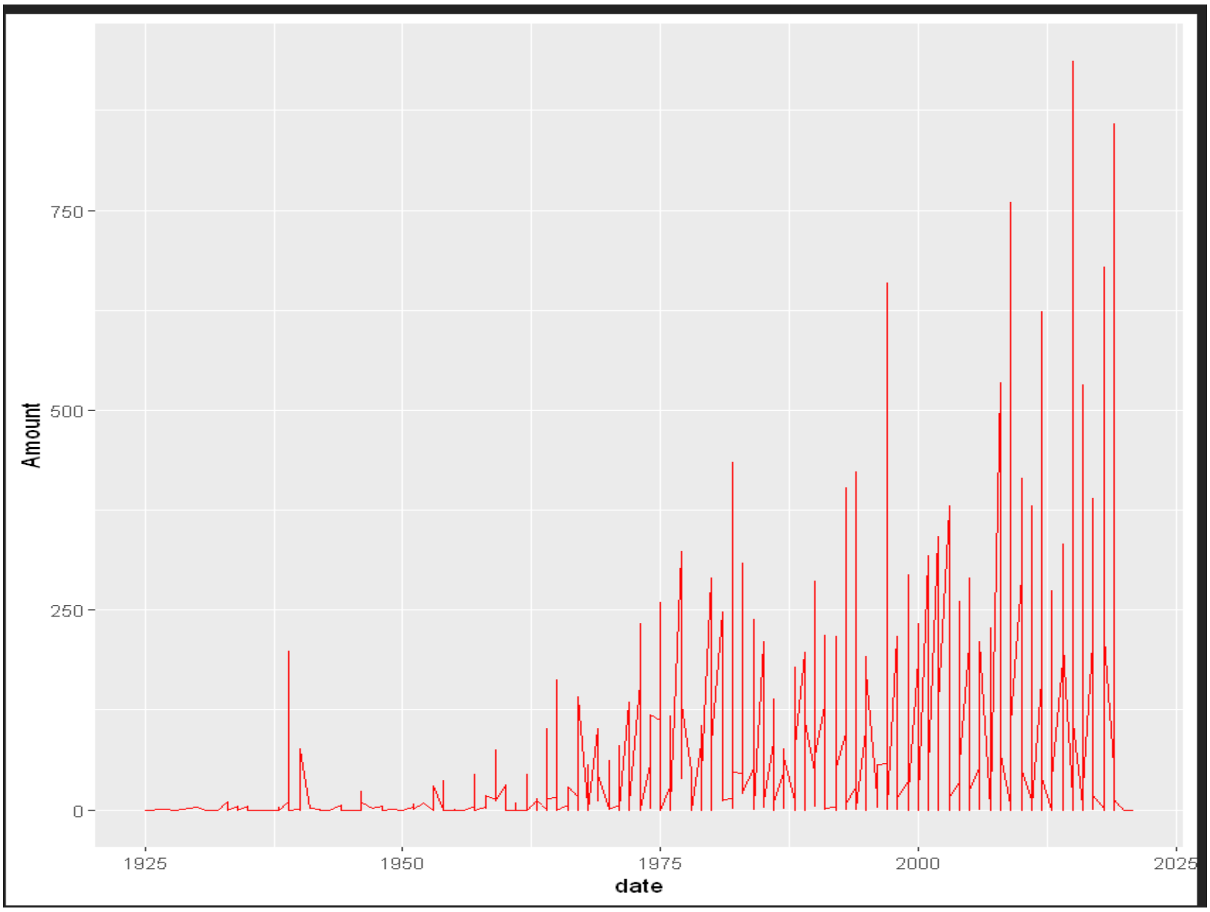


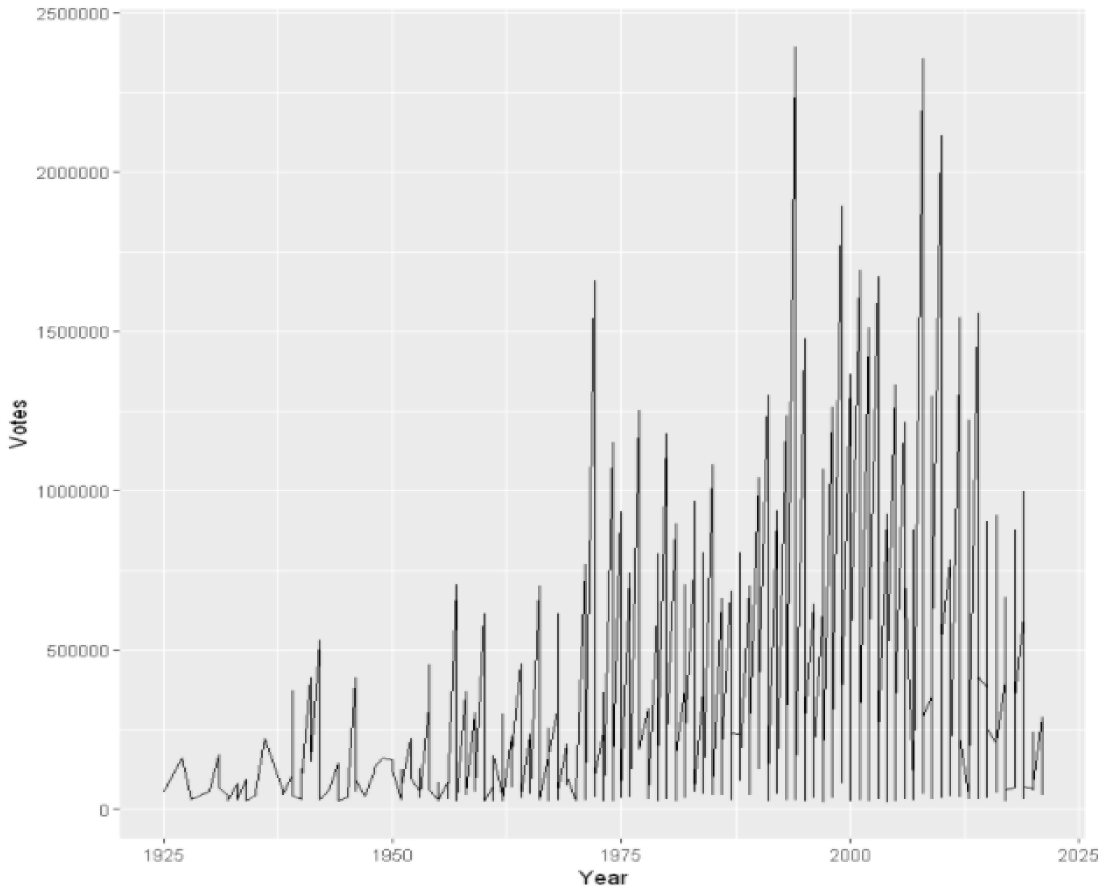
Plotting Pie chart for Genre Statistics:



Task 4: Find out relatable relationship in the dataset:

We can observe from visualizations that gross money, movie reviews and ratings relatable. We will be plotting them against each other to get a clear pattern among them.





1. After investigating and analyzing the data, we discovered that the Gross value gained by movies and the year which movies are released are correlative.

2. we also discovered that the number of votes cast, and the number of votes collected by the films are highly correlated.

3. We can clearly see that when the number of votes cast is low, the amount collected is low as well.

4. when the Gross income for movies is highest ,number of votes is highest, and the rating is also highest in within that year.