
**AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH**

## Prepared and Submitted by


| NAME | ID |
|---|---|
| SAYED ATAULLAH MANSUR | 19-41032-2 |
| TANJEMUL HOQUE TUHIN | 19-41020-2 |
| MUSFIQUR RAHMAN | 19-41056-2 |
| PURAVI DEBNATH NITU | 19-41371-3 |


Course Name:Introduction To Data Science

Section:B[Fall 22-23]

# Project Title: Apply Web Scraping and Data Pre-processing

## Submitted to

Dr. Akinul Islam Jony

Department of Computer Science

Faculty of Science and Technology,AIUB

## Date of Submission

12-12-2022

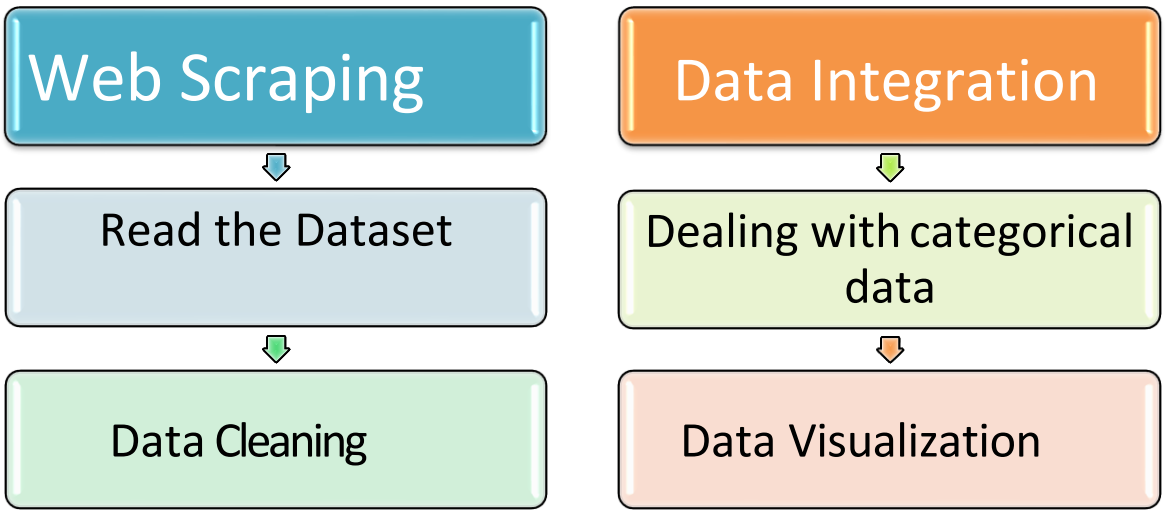
# Project Outline :

Web scraping is an automatic method to obtain large amounts of data from websites. A spreadsheet or an API, for example. There are many techniques for web scraping. we have used r studio tool to scraping our data table. We used a IMDB websites for our Scrapping after that pre-processed the data.

Data pre-processing techniques are used when the data is inconsistent, which indicates that the data is not recorded in accordance with the restrictions on the column, noisy, which may contain a variety of mistakes or outliers, and incomplete, which indicates that some attribute value is missing. In our given dataset First,we tried to fix our missing values. After changing format, we have added a new column using another column value what had been given in our question condition. After adding column then we handled the categorical value to numerical value for our discretization part. At last,we tried to visualize these data.

Data Visualization is the approach used to offer patterns in the data using visual cues such as graphs, charts, maps, and many more. we have used scatter plot to represent our data visualization.

# Project Solution Design:




After creating new project in RStudio read the .CSV file and then done all the data pre-processing systems.

# Web Scraping:

Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications.

| | Movie_name | Realese_date | Director_name | PgRating | Genre | RunTime | Rating | Gross |
|---|---|---|---|---|---|---|---|---|
| 20 | The Magnificent Seven | (2016) | Antoine Fuqua | PG-13 | Action, Adventure, Western | 132 min | 6.8 | $5.02M |
| 21 | Kimi no Na wa. | (2016) | Makoto Shinkai | PG-13 | Animation, Drama, Fantasy | 106 min | 8.4 | $100.01M |
| 22 | Passengers | (I) (2016) | Morten Tyldum | PG-13 | Drama, Romance, Sci-Fi | 116 min | 7.0 | $87.24M |
| 23 | Miss Peregrine's Home for Peculiar Children | (2016) | Tim Burton | R | Adventure, Drama, Family | 127 min | 6.7 | $330.36M |
| 24 | Batman v Superman: Dawn of Justice | (2016) | Zack Snyder | R | Action, Adventure, Sci-Fi | 151 min | 6.4 | $0.18M |
| 25 | The Bad Batch | (2016) | Ana Lily Amirpour | PG-13 | Action, Horror, Mystery | 118 min | 5.2 | $113.26M |
| 26 | Bad Moms | (2016) | Jon Lucas | R | Comedy | 100 min | 6.2 | $234.04M |
| 27 | Fantastic Beasts and Where to Find Them | (2016) | David Yates | PG-13 | Adventure, Family, Fantasy | 132 min | 7.2 | $47.70M |
| 28 | Manchester by the Sea | (2016) | Kenneth Lonergan | PG-13 | Drama | 137 min | 7.8 | $47.37M |
| 29 | Warcraft | (2016) | Duncan Jones | R | Action, Adventure, Fantasy | 123 min | 6.7 | $2.13M |
| 30 | Busanhaeng | (2016) | Sang-ho Yeon | R | Action, Horror, Thriller | 118 min | 7.6 | $232.64M |
| 31 | Doctor Strange | (2016) | Scott Derrickson | R | Action, Adventure, Fantasy | 115 min | 7.5 | $52.85M |
| 32 | 13 Hours | (2016) | Michael Bay | PG-13 | Action, Drama, History | 144 min | 7.3 | $5.88M |
| 33 | Captain Fantastic | (2016) | Matt Ross | R | Comedy, Drama | 118 min | 7.8 | $0.02M |
| 34 | Better Watch Out | (II) (2016) | Chris Peckover | PG | Comedy, Horror, Thriller | 89 min | 6.5 | $72.08M |
| 35 | 10 Cloverfield Lane | (2016) | Dan Trachtenberg | PG-13 | Drama, Horror, Mystery | 103 min | 7.2 | $27.85M |
| 36 | Moonlight | (I) (2016) | Barry Jenkins | PG-13 | Drama | 111 min | 7.4 | $368.38M |
| 37 | The Secret Life of Pets | (2016) | Chris Renaud | PG-13 | Animation, Adventure, Comedy | 87 min | 6.5 | $158.85M |
| 38 | Below Her Mouth | (2016) | April Mullen | PG-13 | Drama, Romance | 94 min | 5.5 | $162.43M |
| 39 | Star Trek Beyond | (2016) | Justin Lin | PG-13 | Action, Adventure, Sci-Fi | 122 min | 7.0 | $127.44M |
| 40 | Jason Bourne | (I) (2016) | Paul Greengrass | R | Action, Thriller | 123 min | 6.6 | $51.74M |
| 41 | Central Intelligence | (2016) | Rawson Marshall Thurber | R | Action, Comedy, Crime | 107 min | 6.3 | $155.44M |
| 42 | Lion | (2016) | Garth Davis | PG-13 | Biography, Drama | 118 min | 8.0 | $14.43M |
| 43 | X-Men: Apocalypse | (2016) | Bryan Singer | R | Action, Adventure, Sci-Fi | 144 min | 6.9 | $1.13M |
| 44 | Mashina lyubvi | (2016) | Pavel Ruminov | R | Drama, Romance | 79 min | 4.0 | $128.34M |
| 45 | The Edge of Seventeen | (2016) | Kelly Fremon Craig | R | Comedy, Drama | 104 min | 7.3 | $97.69M |
| 46 | Lady Macbeth | (2016) | William Oldroyd | NA | Drama, Romance | 89 min | 6.8 | $4.21M |
| 47 | Ghostbusters | (2016) | Paul Feig | NA | Action, Comedy, Fantasy | 117 min | 6.9 | $8.11M |
| 48 | Sausage Party | (2016) | Greg Tiernan | NA | Animation, Adventure, Comedy | 89 min | 6.1 | NA |
| 49 | Swiss Army Man | (2016) | Dan Kwan | NA | Comedy, Drama, Fantasy | 97 min | 6.9 | NA |
| 50 | A Cure for Wellness | (2016) | Gore Verbinski | NA | Drama, Fantasy, Horror | 146 min | 6.4 | NA |

```
library(rvest)
library(dplyr)
link ="https://www.imdb.com/search/title/?year=2016&title_type=feature&"
page=read_html(link)
Movie_name = page %>% html_nodes(".lister-item-header a") %>% html_text()
Director_name =page %>% html_nodes(".text-muted+ p a:nth-child(1)") %>% html_text()
```

```
Rating= page %>% html_nodes(".ratings-imdb-rating strong") %>% html_text()
Realese_date =page %>% html_nodes(".text-muted.unbold") %>% html_text()
Gross =page %>% html_nodes(".ghost~ .text-muted+ span") %>% html_text()
PgRating =page %>% html_nodes(".certificate") %>% html_text()
Genre = page %>% html_nodes(".genre") %>% html_text()
RunTime = page %>% html_nodes(".runtime") %>% html_text()
length(PgRating)=length(Movie_name)
length(Gross)=length(Movie_name)
movie_list=
data.frame(Movie_name,Realese_date,Director_name,PgRating=PgRating,Genre,RunTime,Rating,Gross=G
ross,stringsAsFactors = FALSE)
View(movie_list)
write.csv(movie_list, file ="moviesList.cs
```

# Data Integration:

Data integration is the process of combining data from different sources into a single, unified view. As we collected this dataset from a single source. So, data integration isn't needed in this case. Here,we prepared the dataset to integrate a new column (named Criteria) based on the IMDB gross margin of 2016 popular movies .Converting the gross margin into some types.For example, Flop (<$80M), Average (<$130M), Superhit (<$250M), and Blockbuster $(250M and above).]

| | Genre | RunTime | Rating | Gross | Criteria |
|---|---|---|---|---|---|
| 1 | \nAction, Adventure, Sci-Fi | 133 min | 7.8 | 532.18 | Block Blaster |
| 2 | \nHorror, Thriller | 85 min | 5.6 | 54.77 | flop |
| 3 | \nComedy | 105 min | 5.9 | 325.10 | Block Blaster |
| 4 | \nAction, Adventure, Fantasy | 123 min | 5.9 | 56.25 | flop |
| 5 | \nDrama, Romance | 106 min | 7.4 | 138.29 | Super Hit |
| 6 | \nHorror, Thriller | 117 min | 7.3 | 341.27 | Block Blaster |
| 7 | \nAnimation, Adventure, Comedy | 108 min | 8.0 | 2.01 | flop |
| 8 | \nDrama, Romance, Thriller | 145 min | 8.1 | 363.07 | Block Blaster |
| 9 | \nAction, Adventure, Comedy | 108 min | 8.0 | 169.61 | Super Hit |
| 10 | \nBiography, Drama, History | 127 min | 7.8 | 10.66 | flop |
| 11 | \nDrama, Thriller | 116 min | 7.5 | 151.10 | Super Hit |
| 12 | \nComedy, Drama, Music | 128 min | 8.0 | 100.55 | Average |
| 13 | \nDrama, Mystery, Sci-Fi | 116 min | 7.9 | 67.21 | flop |
| 14 | \nBiography, Drama, History | 139 min | 8.1 | 26.86 | flop |
| 15 | \nCrime, Drama, Thriller | 102 min | 7.6 | 270.40 | Block Blaster |
| 16 | \nAnimation, Comedy, Family | 108 min | 7.1 | 408.08 | Block Blaster |
| 17 | \nAction, Adventure, Sci-Fi | 147 min | 7.8 | 36.26 | flop |
| 18 | \nAction, Comedy, Crime | 116 min | 7.3 | 248.76 | Super Hit |
| 19 | \nAnimation, Adventure, Comedy | 107 min | 7.6 | 93.43 | Average |
| 20 | \nAction, Adventure, Western | 132 min | 6.8 | 5.02 | flop |
| 21 | \nAnimation, Drama, Fantasy | 106 min | 8.4 | 100.01 | Average |
| 22 | \nDrama, Romance, Sci-Fi | 116 min | 7.0 | 87.24 | Average |
| 23 | \nAdventure, Drama, Family | 127 min | 6.7 | 330.36 | Block Blaster |
| 24 | \nAction, Adventure, Sci-Fi | 151 min | 6.4 | 0.18 | flop |
| 25 | \nAction, Horror, Mystery | 118 min | 5.2 | 113.26 | Average |
| 26 | \nComedy | 100 min | 6.2 | 234.04 | Super Hit |
| 27 | \nAdventure, Family, Fantasy | 132 min | 7.2 | 47.70 | flop |
| 28 | \nDrama | 137 min | 7.8 | 47.37 | flop |
| 29 | \nAction, Adventure, Fantasy | 123 min | 6.7 | 2.13 | flop |
| 30 | \nAction, Horror, Thriller | 118 min | 7.6 | 232.64 | Super Hit |
| 31 | \nAction, Adventure, Fantasy | 115 min | 7.5 | 52.85 | flop |
| 32 | \nAction, Drama, History | 144 min | 7.3 | 5.88 | flop |
| 33 | \nComedy, Drama | 118 min | 7.8 | 0.02 | flop |
| 34 | \nComedy, Horror, Thriller | 89 min | 6.5 | 72.08 | flop |
| 35 | \nDrama, Horror, Mystery | 103 min | 7.2 | 27.85 | flop |
| 36 | \nDrama | 111 min | 7.4 | 368.38 | Block Blaster |
| 37 | \nAnimation, Adventure, Comedy | 87 min | 6.5 | 158.85 | Super Hit |
| 38 | \nDrama, Romance | 94 min | 5.5 | 162.43 | Super Hit |
| 39 | \nAction, Adventure, Sci-Fi | 122 min | 7.0 | 127.44 | Average |
| 40 | \nAction, Thriller | 123 min | 6.6 | 51.74 | flop |
| 41 | \nAction, Comedy, Crime | 107 min | 6.3 | 155.44 | Super Hit |
| 42 | \nBiography, Drama | 118 min | 8.0 | 14.43 | flop |
| 43 | \nAction, Adventure, Sci-Fi | 144 min | 6.9 | 1.13 | flop |
| 44 | \nDrama, Romance | 79 min | 4.0 | 128.34 | Average |
| 45 | \nComedy, Drama | 104 min | 7.3 | 97.69 | Average |

```
> |
```

```
dataset$Criteria <-rep (NA, nrow(dataset))

dataset[dataset$Gross>0 & dataset$Gross<80,]["Criteria"] <-"flop"

dataset[dataset$Gross>=80 &dataset$Gross<=130,]["Criteria"] <- "Average"

dataset[dataset$Gross>130 & dataset$Gross<250,]["Criteria"] <- "Super Hit"

dataset[dataset$Gross>=250,]["Criteria"] <- "Block Blaster"

dataset
```

## **Data Discretization**:

Dealed with categorical data.We wanted to discretize the types(flop,average,superhit and Blockbuster) into four categories 10,20,30 and 40.

| | Genre | Runtime | Rating | Gross | Criteria |
|---|---|---|---|---|---|
| 1 | \nAction, Adventure, Sci-Fi | 133 min | 7.8 | 532.18 | 40 |
| 2 | \nHorror, Thriller | 85 min | 5.6 | 54.77 | 10 |
| 3 | \nComedy | 105 min | 5.9 | 325.10 | 40 |
| 4 | \nAction, Adventure, Fantasy | 123 min | 5.9 | 56.25 | 10 |
| 5 | \nDrama, Romance | 106 min | 7.4 | 138.29 | 30 |
| 6 | \nHorror, Thriller | 117 min | 7.3 | 341.27 | 40 |
| 7 | \nAnimation, Adventure, Comedy | 108 min | 8.0 | 2.01 | 10 |
| 8 | \nDrama, Romance, Thriller | 145 min | 8.1 | 363.07 | 40 |
| 9 | \nAction, Adventure, Comedy | 108 min | 8.0 | 169.61 | 30 |
| 10 | \nBiography, Drama, History | 127 min | 7.8 | 10.66 | 10 |
| 11 | \nDrama, Thriller | 116 min | 7.5 | 151.10 | 30 |
| 12 | \nComedy, Drama, Music | 128 min | 8.0 | 100.55 | 20 |
| 13 | \nDrama, Mystery, Sci-Fi | 116 min | 7.9 | 67.21 | 10 |
| 14 | \nBiography, Drama, History | 139 min | 8.1 | 26.86 | 10 |
| 15 | \nCrime, Drama, Thriller | 102 min | 7.6 | 270.40 | 40 |
| 16 | \nAnimation, Comedy, Family | 108 min | 7.1 | 408.08 | 40 |
| 17 | \nAction, Adventure, Sci-Fi | 147 min | 7.8 | 36.26 | 10 |
| 18 | \nAction, Comedy, Crime | 116 min | 7.3 | 248.76 | 30 |
| 19 | \nAnimation, Adventure, Comedy | 107 min | 7.6 | 93.43 | 20 |
| 20 | \nAction, Adventure, Western | 132 min | 6.8 | 5.02 | 10 |
| 21 | \nAnimation, Drama, Fantasy | 106 min | 8.4 | 100.01 | 20 |
| 22 | \nDrama, Romance, Sci-Fi | 116 min | 7.0 | 87.24 | 20 |
| 23 | \nAdventure, Drama, Family | 127 min | 6.7 | 330.36 | 40 |
| 24 | \nAction, Adventure, Sci-Fi | 151 min | 6.4 | 0.18 | 10 |
| 25 | \nAction, Horror, Mystery | 118 min | 5.2 | 113.26 | 20 |
| 26 | \nComedy | 100 min | 6.2 | 234.04 | 30 |
| 27 | \nAdventure, Family, Fantasy | 132 min | 7.2 | 47.70 | 10 |
| 28 | \nDrama | 137 min | 7.8 | 47.37 | 10 |
| 29 | \nAction, Adventure, Fantasy | 123 min | 6.7 | 2.13 | 10 |
| 30 | \nAction, Horror, Thriller | 118 min | 7.6 | 232.64 | 30 |
| 31 | \nAction, Adventure, Fantasy | 115 min | 7.5 | 52.85 | 10 |
| 32 | \nAction, Drama, History | 144 min | 7.3 | 5.88 | 10 |
| 33 | \nComedy, Drama | 118 min | 7.8 | 0.02 | 10 |
| 34 | \nComedy, Horror, Thriller | 89 min | 6.5 | 72.08 | 10 |
| 35 | \nDrama, Horror, Mystery | 103 min | 7.2 | 27.85 | 10 |
| 36 | \nDrama | 111 min | 7.4 | 368.38 | 40 |
| 37 | \nAnimation, Adventure, Comedy | 87 min | 6.5 | 158.85 | 30 |
| 38 | \nDrama, Romance | 94 min | 5.5 | 162.43 | 30 |
| 39 | \nAction, Adventure, Sci-Fi | 122 min | 7.0 | 127.44 | 20 |
| 40 | \nAction, Thriller | 123 min | 6.6 | 51.74 | 10 |
| 41 | \nAction, Comedy, Crime | 107 min | 6.3 | 155.44 | 30 |
| 42 | \nBiography, Drama | 118 min | 8.0 | 14.43 | 10 |
| 43 | \nAction, Adventure, Sci-Fi | 144 min | 6.9 | 1.13 | 10 |
| 44 | \nDrama, Romance | 79 min | 4.0 | 128.34 | 20 |
| 45 | \nComedy, Drama | 104 min | 7.3 | 97.69 | 20 |

```
>
```

dataset$Criteria = factor(dataset$Criteria,levels = c('flop','Average','Super Hit','Block Blaster'), labels = c(10,20,30,40))

dataset

## Data Cleaning:

Data in the real world is often dirty. Data is in need of being cleaned up before it can be used for a desired purpose. This is often called data preprocessing. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.Here we have cleaned the N/A(Not Available) data.

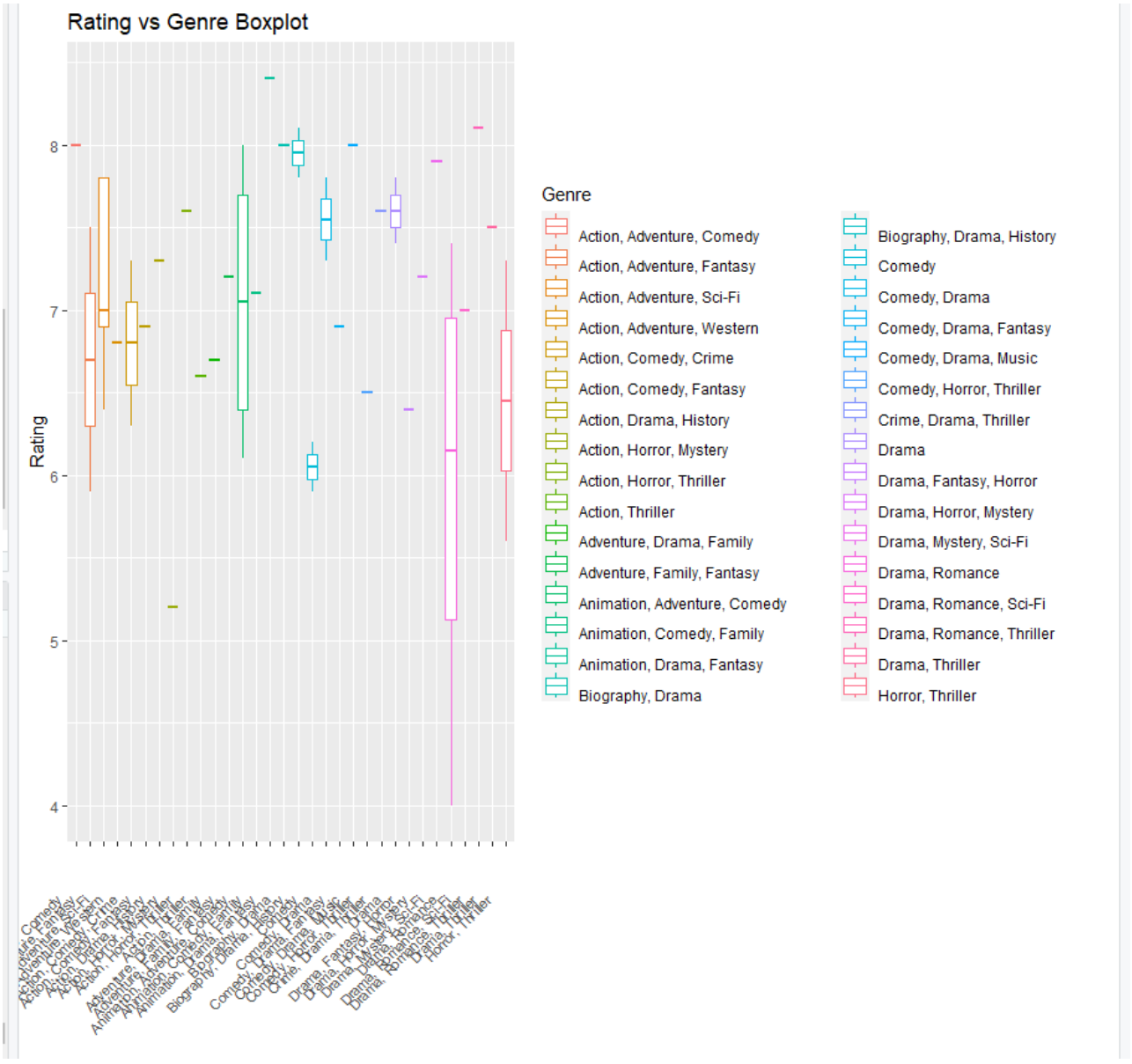
| X | Movie_name | Release_date | Director_name | PgRating |
|---|---|---|---|---|
| 1 | Rogue One | -2016 | Gareth Edwards | PG-13 |
| 2 | Terrifier | -2016 | Damien Leone | R |
| 3 | Office Christmas Party | -2016 | Josh Gordon | PG-13 |
| 4 | Suicide Squad | -2016 | David Ayer | PG-13 |
| 5 | Me Before You | -2016 | Thea Sharrock | PG-13 |
| 6 | Split | (IX) (2016) | M. Night Shyamalan | PG |
| 7 | Zootopia | -2016 | Byron Howard | R |
| 8 | Ah-ga-ssi | -2016 | Park Chan-wook | PG |
| 9 | Deadpool | -2016 | Tim Miller | R |
| 10 | Hidden Figures | -2016 | Theodore Melfi | PG-13 |
| 11 | Nocturnal Animals | -2016 | Tom Ford | PG-13 |
| 12 | La La Land | -2016 | Damien Chazelle | R |
| 13 | Arrival | (II) (2016) | Denis Villeneuve | R |
| 14 | Hacksaw Ridge | -2016 | Mel Gibson | PG |
| 15 | Hell or High Water | (II) (2016) | David Mackenzie | PG-13 |
| 16 | Sing | -2016 | Garth Jennings | R |
| 17 | Captain America: Civil War | -2016 | Anthony Russo | PG |
| 18 | The Nice Guys | -2016 | Shane Black | PG-13 |
| 19 | Moana | (I) (2016) | Ron Clements | PG |
| 20 | The Magnificent Seven | -2016 | Antoine Fuqua | PG-13 |
| 21 | Kimi no Na wa. | -2016 | Makoto Shinkai | PG-13 |
| 22 | Passengers | (I) (2016) | Morten Tyldum | PG-13 |
| 23 | Miss Peregrine's Home for Peculiar Children | -2016 | Tim Burton | R |
| 24 | Batman v Superman: Dawn of Justice | -2016 | Zack Snyder | R |
| 25 | The Bad Batch | -2016 | Ana Lily Amirpour | PG-13 |
| 26 | Bad Moms | -2016 | Jon Lucas | R |
| 27 | Fantastic Beasts and Where to Find Them | -2016 | David Yates | PG-13 |
| 28 | Manchester by the Sea | -2016 | Kenneth Lonergan | PG-13 |
| 29 | Warcraft | -2016 | Duncan Jones | R |
| 30 | Busanhaeng | -2016 | Sang-ho Yeon | R |
| 31 | Doctor Strange | -2016 | Scott Derrickson | R |
| 32 | 13 Hours | -2016 | Michael Bay | PG-13 |
| 33 | Captain Fantastic | -2016 | Matt Ross | R |
| 34 | Better Watch Out | (II) (2016) | Chris Peckover | PG |
| 35 | 10 Cloverfield Lane | -2016 | Dan Trachtenberg | PG-13 |
| 36 | Moonlight | (I) (2016) | Barry Jenkins | PG-13 |
| 37 | The Secret Life of Pets | -2016 | Chris Renaud | PG-13 |
| 38 | Below Her Mouth | -2016 | April Mullen | PG-13 |
| 39 | Star Trek Beyond | -2016 | Justin Lin | PG-13 |
| 40 | Jason Bourne | (I) (2016) | Paul Greengrass | R |
| 41 | Central Intelligence | -2016 | Rawson Marshall Thurber | R |
| 42 | Lion | -2016 | Garth Davis | PG-13 |
| 43 | X-Men: Apocalypse | -2016 | Bryan Singer | R |
| 44 | Mashina lyubvi | -2016 | Pavel Ruminov | R |
| 45 | The Edge of Seventeen | -2016 | Kelly Fremon Craig | R |

```
dataset<-read.csv("moviesList.csv")
dataset
dataset<-na.omit(dataset)
dataset
```

## Data Visualization:

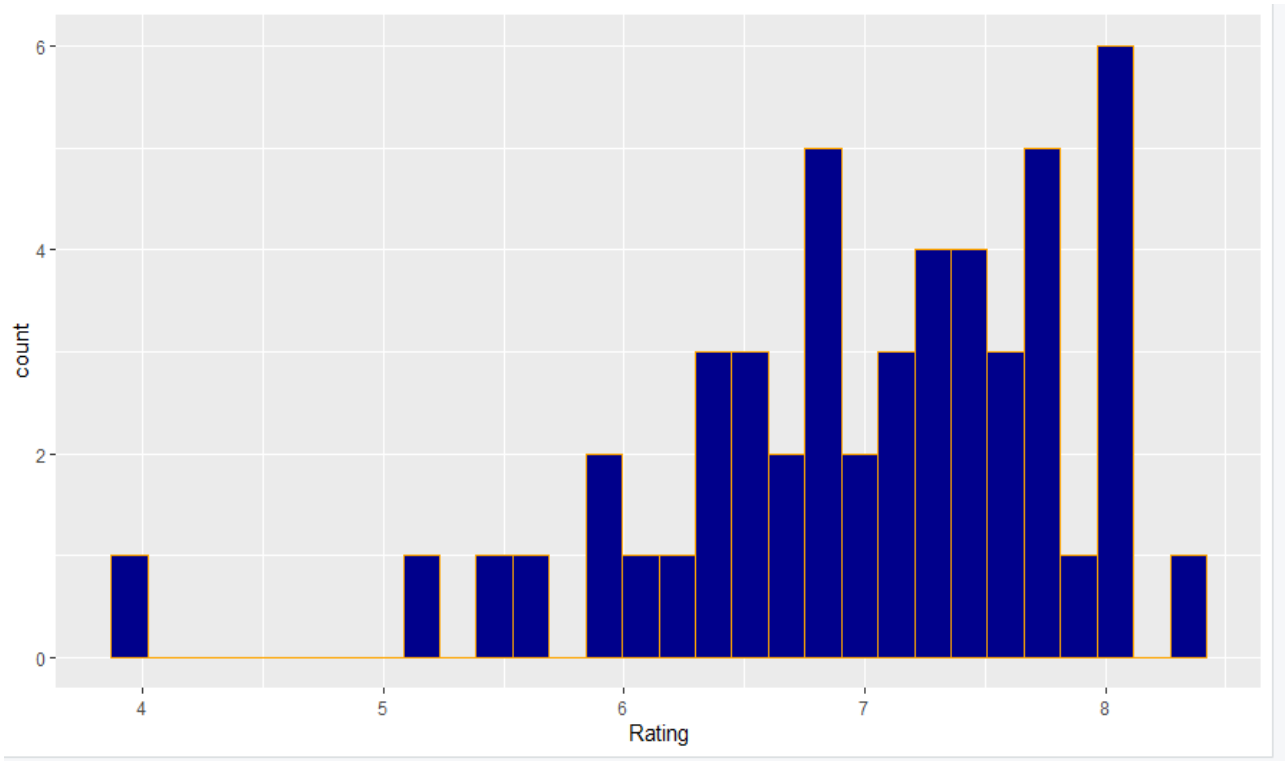
Data Visualization is the approach used to offer patterns in the data using visual cues such as graphs, charts, maps, and many more. This is helpful because it facilitates intuitive and simple understanding of the vast amounts of data, allowing for better decision-making.

The below Box plot is made depending on the rating(X-axis) and genre(Y-Axis)



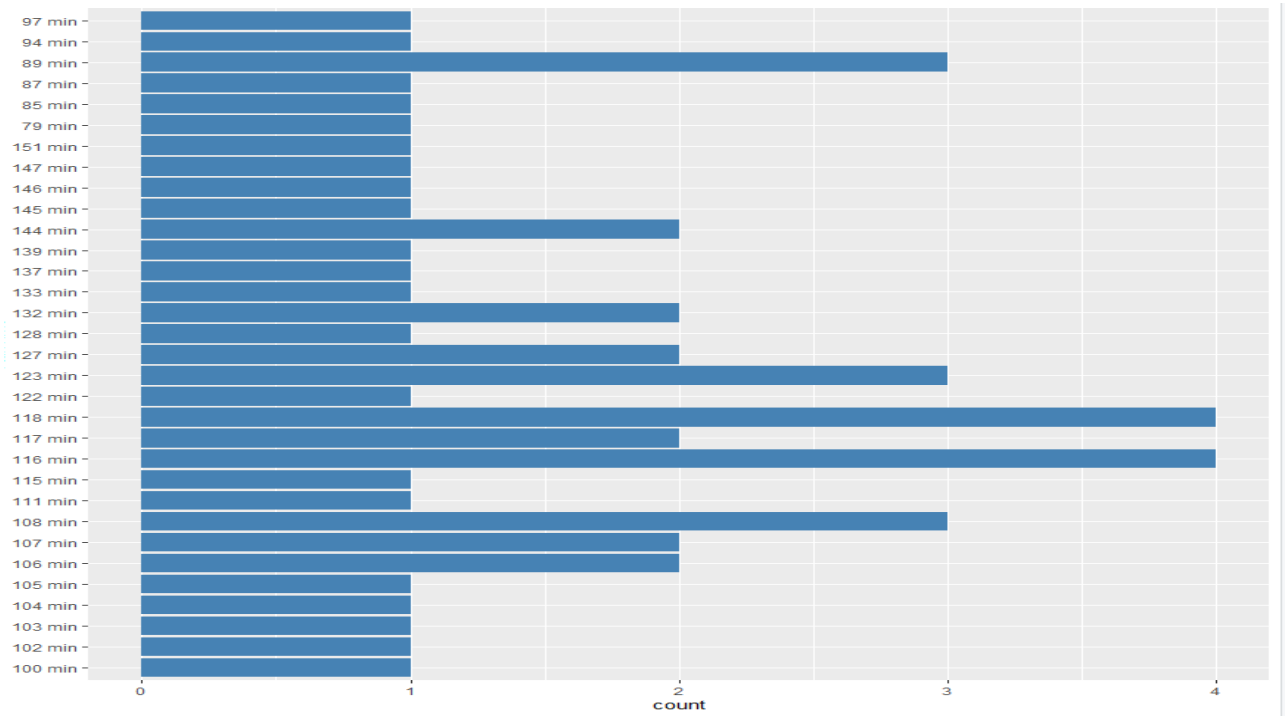

```
library(ggplot2)
ggplot(dataset, aes(x =Genre, y =Rating,color=Genre))+geom_boxplot()+labs(title="Rating vs Genre Boxplot",y="Rating",x="")+theme(axis.text.x=element_text(angle=45,hjust=1,size=8))
```

The below Histogram Plot is made depending on the Movie rating




ggplot(dataset, aes(x=Rating)) + geom_histogram(color="orange", fill="darkblue")+theme_gray()

The below Bar plot is made depending on the movie duration




```
ggplot(dataset, aes(y=RunTime)) + geom_bar(fill="steelblue")
```

## Discussion and Conclusion:

The project plays a vital role to learn how to web scrape and pre-process data. We learned about webscraping, cleaning,integration,reduction,transformation and discretization of data from a datset.I faced some kind of difficulties working with the datas using RStudio but It gives us a few knowledge how a data scientists work with datas.Now a days,every information is a data and we are surrounding by datas.so,its important to know how data can be scraped and processed.