AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh

Assignment Title:DATA SCIENCE FINAL REPORT	
Assignment No:02	Date of Submission:03/05/2023
Course Title:DATA SCIENCE	
Course Code:4180	Section:B
2023-2024	Course Teacher: AKINUL ISLAM JONY

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*

** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.:

No	Name	ID	Program	Signature
1	**TAJKIA FARUK**	**19-40009-1**	BSC in CSE	TAJKIA
2	**EFAT SAYRA**	**19-40475-1**	BSC in CSE	EFAT
3	**FAIRUZ ZAHIN SNEHA**	**20-41912-1**	BSC in CSE	SNEHA
4	**ASMAUL HUSNA JARIN**	**20-42363-1**	BSC in CSE	JARIN
5				
6				
7				
8				
9				
10				

Faculty use only

FACULTYCOMMENTS	**Marks Obtained**	

Content

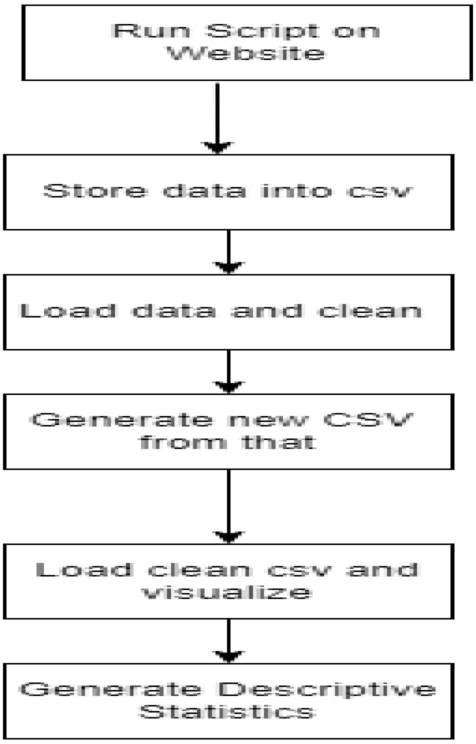
1.1Project Overview

IMDB is one of the most popular movie rating sites in the world. Where we can get proper idea about movies and check user reviews and judge based on them. The IMDB data scraping project involved extracting information about 150 movies from the Internet Movie Database (IMDB) website using R programming language. The goal of the project was to gather data on a wide range of movies, including their title, runtime, number of reviews and ratings. This data was then analyzed and visualized to gain insights into movie trends and patterns. The first step in the project was to identify the data that was to be scraped from the IMDB website. The data was to be scraped from the movie pages on the website, using web scraping techniques to extract the relevant information from the HTML code. The data scraping process involved using R packages like rvest, dplyr, tidyr, and ggplot2. The rvest package was used for web scraping, while the dplyr and tidyr packages were used for data cleaning and manipulation. The ggplot2 package was used for data visualization. Once the data was scraped, it was cleaned and prepared for analysis. This involved removing any missing or duplicate data, converting data types, and creating new variables. The data was then analyzed using descriptive statistics, such as mean, median, standard deviation, and frequency distributions. The results of the analysis showed that movies with medium range of runtime are the most popular among them. The average runtime of the movies was found to be around 135 minutes, with a standard deviation of 30 minutes. The ratings of the movies ranged from 8 to 9.3, with an average rating of 8.4. Finally, the data was visualized using various graphs and charts, such as bar charts, histograms. These visualizations helped to highlight the trends and patterns in the data and make it easier to interpret.

Overall, the IMDB data scraping project provided a valuable insight into the world of movies and demonstrated the power of web scraping and data analysis with R programming language.

1.2Solution Design

The project solution consists of six steps. After which we can do a proper data analysis. Belowa flowchart is given by which we can achieve our goal.

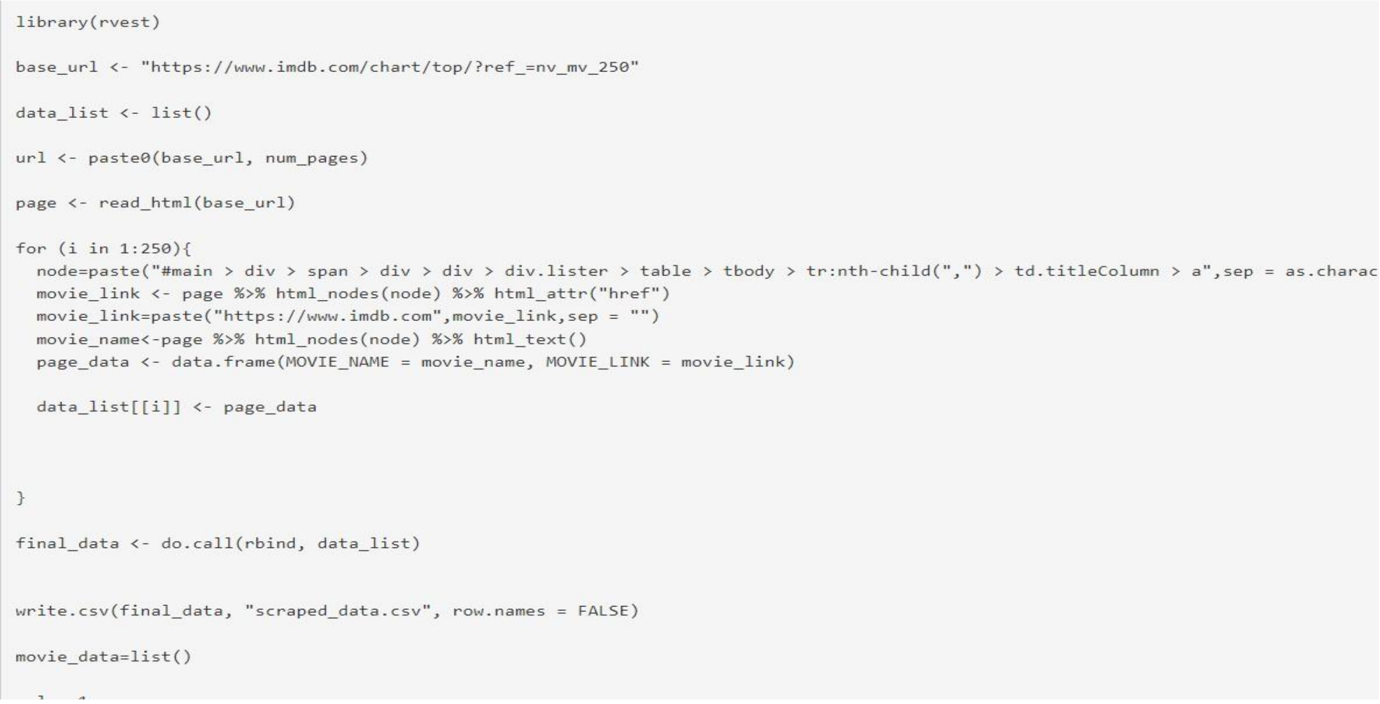

Run Script on Website
Store data into csv
Load data and clean
Generate new CSV from that
Load clean csv and visualize
Generate Descriptive Statistics

As the initial step we must write the R script which will let us scrap all the data from the website and after that we have to store that data into a CSV file. After that we must load thatCSV file into another R script. Then we must use data pre-processing methods to clean the data. After the cleaning process we must generate a new csv for that clean fresh dataset. Nowwe can import it and do the visualization part. After visualization we must do the descriptivestatistical analysis from that dataset again. Finally, we can give some verdict on that.

1.3Data Collection via Web Scraping

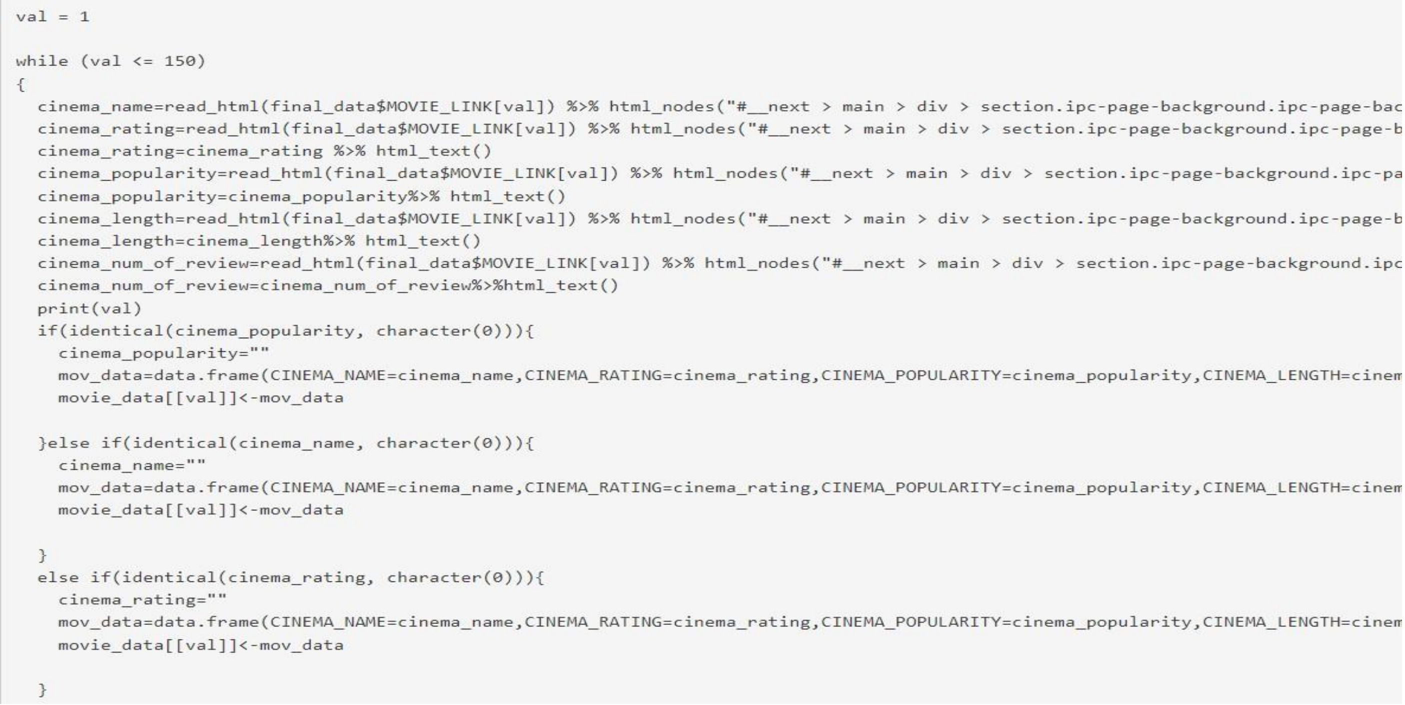
In this project to complete the task of web scraping the “rvest” library was used. The websitescraped in this project was https://www.imdb.com/chart/top/?ref_=nv_mv_250 which is a movie review website.

To collect data from that website we scraped the movie names and their links into a data frame to make our work easier. After that from that data frame we accessed each of the linksand fetched each movie’s name, runtime, popularity and rating. Finally after collecting all thedata, we stored them into separate data frame and exported that file into a csv file.



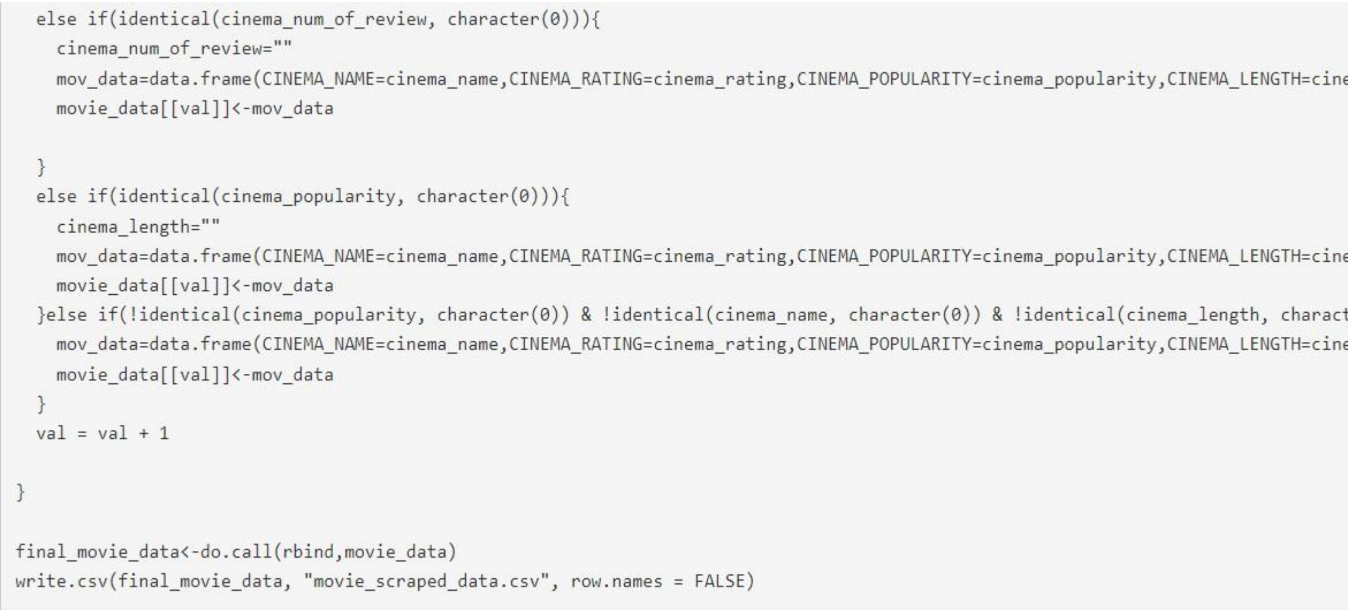
```
library(rvest)

base_url <- "https://www.imdb.com/chart/top/?ref_=nv_mv_250"

data_list <- list()

url <- paste0(base_url, num_pages)

page <- read_html(base_url)

for (i in 1:250){
  node=paste("#main > div > span > div > div > div.lister > table > tbody > tr:nth-child(",") > td.titleColumn > a",sep = as.charac
  movie_link <- page %>% html_nodes(node) %>% html_attr("href")
  movie_link=paste("https://www.imdb.com",movie_link,sep = "")
  movie_name<-page %>% html_nodes(node) %>% html_text()
  page_data <- data.frame(MOVIE_NAME = movie_name, MOVIE_LINK = movie_link)

  data_list[[i]] <- page_data



}

final_data <- do.call(rbind, data_list)


write.csv(final_data, "scraped_data.csv", row.names = FALSE)

movie_data=list()
```

Figure 2:Web Scrapping Part 1



```
val = 1

while (val <= 150)
{
  cinema_name=read_html(final_data$MOVIE_LINK[val]) %>% html_nodes("#__next > main > div > section.ipc-page-background.ipc-page-bac
  cinema_rating=read_html(final_data$MOVIE_LINK[val]) %>% html_nodes("#__next > main > div > section.ipc-page-background.ipc-page-b
  cinema_rating=cinema_rating %>% html_text()
  cinema_popularity=read_html(final_data$MOVIE_LINK[val]) %>% html_nodes("#__next > main > div > section.ipc-page-background.ipc-pa
  cinema_popularity=cinema_popularity%>% html_text()
  cinema_length=read_html(final_data$MOVIE_LINK[val]) %>% html_nodes("#__next > main > div > section.ipc-page-background.ipc-page-b
  cinema_length=cinema_length%>% html_text()
  cinema_num_of_review=read_html(final_data$MOVIE_LINK[val]) %>% html_nodes("#__next > main > div > section.ipc-page-background.ipc
  cinema_num_of_review=cinema_num_of_review%>%html_text()
  print(val)
  if(identical(cinema_popularity, character(0))){
    cinema_popularity=""
    mov_data=data.frame(CINEMA_NAME=cinema_name,CINEMA_RATING=cinema_rating,CINEMA_POPULARITY=cinema_popularity,CINEMA_LENGTH=cinem
    movie_data[[val]]<-mov_data

  }else if(identical(cinema_name, character(0))){
    cinema_name=""
    mov_data=data.frame(CINEMA_NAME=cinema_name,CINEMA_RATING=cinema_rating,CINEMA_POPULARITY=cinema_popularity,CINEMA_LENGTH=cinem
    movie_data[[val]]<-mov_data

  }
  else if(identical(cinema_rating, character(0))){
    cinema_rating=""
    mov_data=data.frame(CINEMA_NAME=cinema_name,CINEMA_RATING=cinema_rating,CINEMA_POPULARITY=cinema_popularity,CINEMA_LENGTH=cinem
    movie_data[[val]]<-mov_data

  }
```

Figure 3: Web Scrapping Part 2

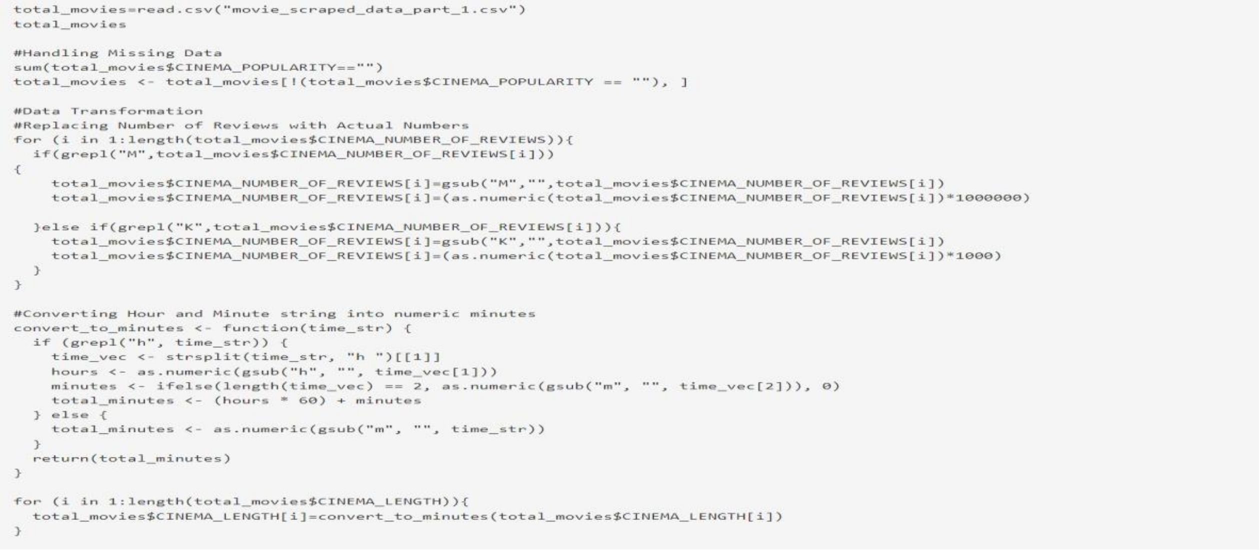


```
  else if(identical(cinema_num_of_review, character(0))){
    cinema_num_of_review=""
    mov_data=data.frame(CINEMA_NAME=cinema_name,CINEMA_RATING=cinema_rating,CINEMA_POPULARITY=cinema_popularity,CINEMA_LENGTH=cine
    movie_data[[val]]<-mov_data

  }
  else if(identical(cinema_popularity, character(0))){
    cinema_length=""
    mov_data=data.frame(CINEMA_NAME=cinema_name,CINEMA_RATING=cinema_rating,CINEMA_POPULARITY=cinema_popularity,CINEMA_LENGTH=cine
    movie_data[[val]]<-mov_data
  }else if(!identical(cinema_popularity, character(0)) & !identical(cinema_name, character(0)) & !identical(cinema_length, charact
    mov_data=data.frame(CINEMA_NAME=cinema_name,CINEMA_RATING=cinema_rating,CINEMA_POPULARITY=cinema_popularity,CINEMA_LENGTH=cine
    movie_data[[val]]<-mov_data
  }
  val = val + 1

}

final_movie_data<-do.call(rbind,movie_data)
write.csv(final_movie_data, "movie_scraped_data.csv", row.names = FALSE)
```

Figure 4: Web Scrapping Part 3

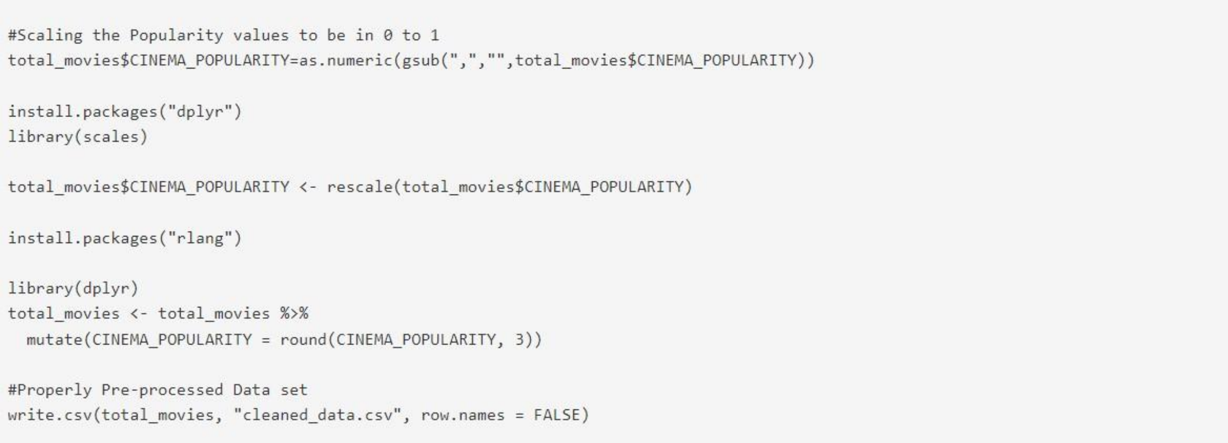
1.4Data Pre-Processing

For data-preprocessing first we moved those rows which had some missing records such as rating or popularity. Then we did the data transformation. Here the number of reviews is stored as a string such as “2.7M” we converted it to numeric format for all of them. Then weconverted all the movie length string to minute numeric format. The cinema popularity hadhuge range of values, so those values were scaled down to 0 to 1. Finally after all the cleaningwe stored them into a new csv file.



```
total_movies=read.csv("movie_scraped_data_part_1.csv")
total_movies

#Handling Missing Data
sum(total_movies$CINEMA_POPULARITY=="")
total_movies <- total_movies[!(total_movies$CINEMA_POPULARITY == ""), ]

#Data Transformation
#Replacing Number of Reviews with Actual Numbers
for (i in 1:length(total_movies$CINEMA_NUMBER_OF_REVIEWS)){
  if(grepl("M",total_movies$CINEMA_NUMBER_OF_REVIEWS[i]))
{
    total_movies$CINEMA_NUMBER_OF_REVIEWS[i]=gsub("M","",total_movies$CINEMA_NUMBER_OF_REVIEWS[i])
    total_movies$CINEMA_NUMBER_OF_REVIEWS[i]=(as.numeric(total_movies$CINEMA_NUMBER_OF_REVIEWS[i])*1000000)

  }else if(grepl("K",total_movies$CINEMA_NUMBER_OF_REVIEWS[i])){
    total_movies$CINEMA_NUMBER_OF_REVIEWS[i]=gsub("K","",total_movies$CINEMA_NUMBER_OF_REVIEWS[i])
    total_movies$CINEMA_NUMBER_OF_REVIEWS[i]=(as.numeric(total_movies$CINEMA_NUMBER_OF_REVIEWS[i])*1000)
  }
}

#Converting Hour and Minute string into numeric minutes
convert_to_minutes <- function(time_str) {
  if (grepl("h", time_str)) {
    time_vec <- strsplit(time_str, "h ")[[1]]
    hours <- as.numeric(gsub("h", "", time_vec[1]))
    minutes <- ifelse(length(time_vec) == 2, as.numeric(gsub("m", "", time_vec[2])), 0)
    total_minutes <- (hours * 60) + minutes
  } else {
    total_minutes <- as.numeric(gsub("m", "", time_str))
  }
  return(total_minutes)
}

for (i in 1:length(total_movies$CINEMA_LENGTH)){
  total_movies$CINEMA_LENGTH[i]=convert_to_minutes(total_movies$CINEMA_LENGTH[i])
}
```

Figure 5: Code for Pre-Processing Part 1

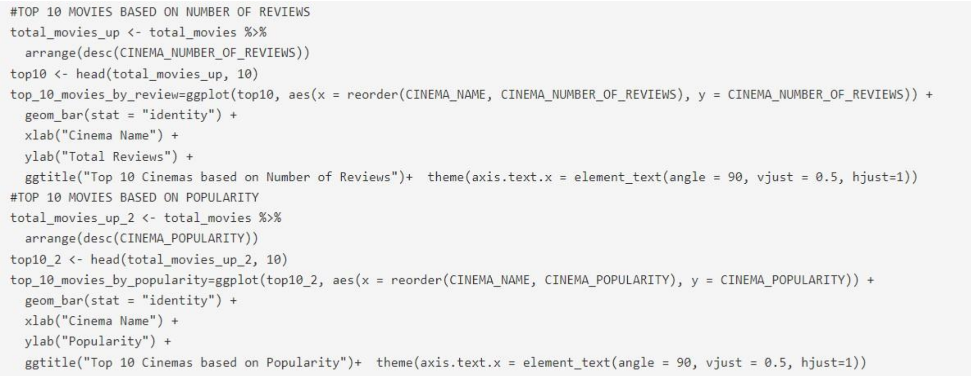


```
#Scaling the Popularity values to be in 0 to 1
total_movies$CINEMA_POPULARITY=as.numeric(gsub(",","",total_movies$CINEMA_POPULARITY))

install.packages("dplyr")
library(scales)

total_movies$CINEMA_POPULARITY <- rescale(total_movies$CINEMA_POPULARITY)

install.packages("rlang")

library(dplyr)
total_movies <- total_movies %>%
  mutate(CINEMA_POPULARITY = round(CINEMA_POPULARITY, 3))

#Properly Pre-processed Data set
write.csv(total_movies, "cleaned_data.csv", row.names = FALSE)
```

Figure 6: Code for Pre-Processing Part 2

1.5Data 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.



```
#TOP 10 MOVIES BASED ON NUMBER OF REVIEWS
total_movies_up <- total_movies %>%
  arrange(desc(CINEMA_NUMBER_OF_REVIEWS))
top10 <- head(total_movies_up, 10)
top_10_movies_by_review=ggplot(top10, aes(x = reorder(CINEMA_NAME, CINEMA_NUMBER_OF_REVIEWS), y = CINEMA_NUMBER_OF_REVIEWS)) +
  geom_bar(stat = "identity") +
  xlab("Cinema Name") +
  ylab("Total Reviews") +
  ggtitle("Top 10 Cinemas based on Number of Reviews")+  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
#TOP 10 MOVIES BASED ON POPULARITY
total_movies_up_2 <- total_movies %>%
  arrange(desc(CINEMA_POPULARITY))
top10_2 <- head(total_movies_up_2, 10)
top_10_movies_by_popularity=ggplot(top10_2, aes(x = reorder(CINEMA_NAME, CINEMA_POPULARITY), y = CINEMA_POPULARITY)) +
  geom_bar(stat = "identity") +
  xlab("Cinema Name") +
  ylab("Popularity") +
  ggtitle("Top 10 Cinemas based on Popularity")+  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

Figure 7: TOP 10 MOVIES BASED ON NUMBER OF REVIEWS

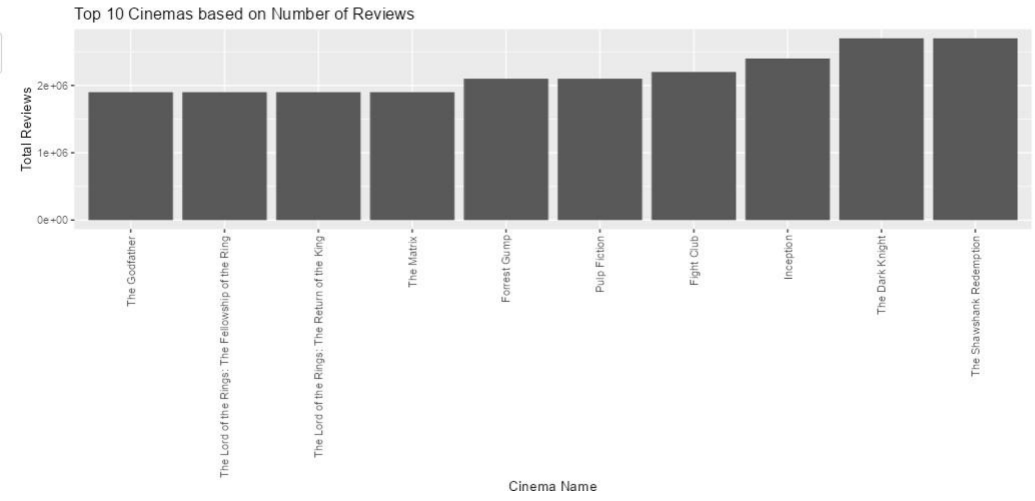
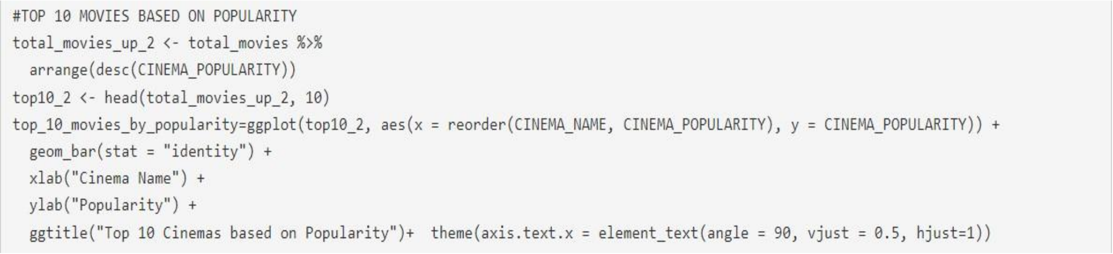


Figure 8: PLOTTING OF TOP 10 MOVIES BASED ON NUMBER OF REVIEWS



```
#TOP 10 MOVIES BASED ON POPULARITY
total_movies_up_2 <- total_movies %>%
  arrange(desc(CINEMA_POPULARITY))
top10_2 <- head(total_movies_up_2, 10)
top_10_movies_by_popularity=ggplot(top10_2, aes(x = reorder(CINEMA_NAME, CINEMA_POPULARITY), y = CINEMA_POPULARITY)) +
  geom_bar(stat = "identity") +
  xlab("Cinema Name") +
  ylab("Popularity") +
  ggtitle("Top 10 Cinemas based on Popularity")+  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

Figure 9: TOP 10 MOVIES BASED ON POPULARITY

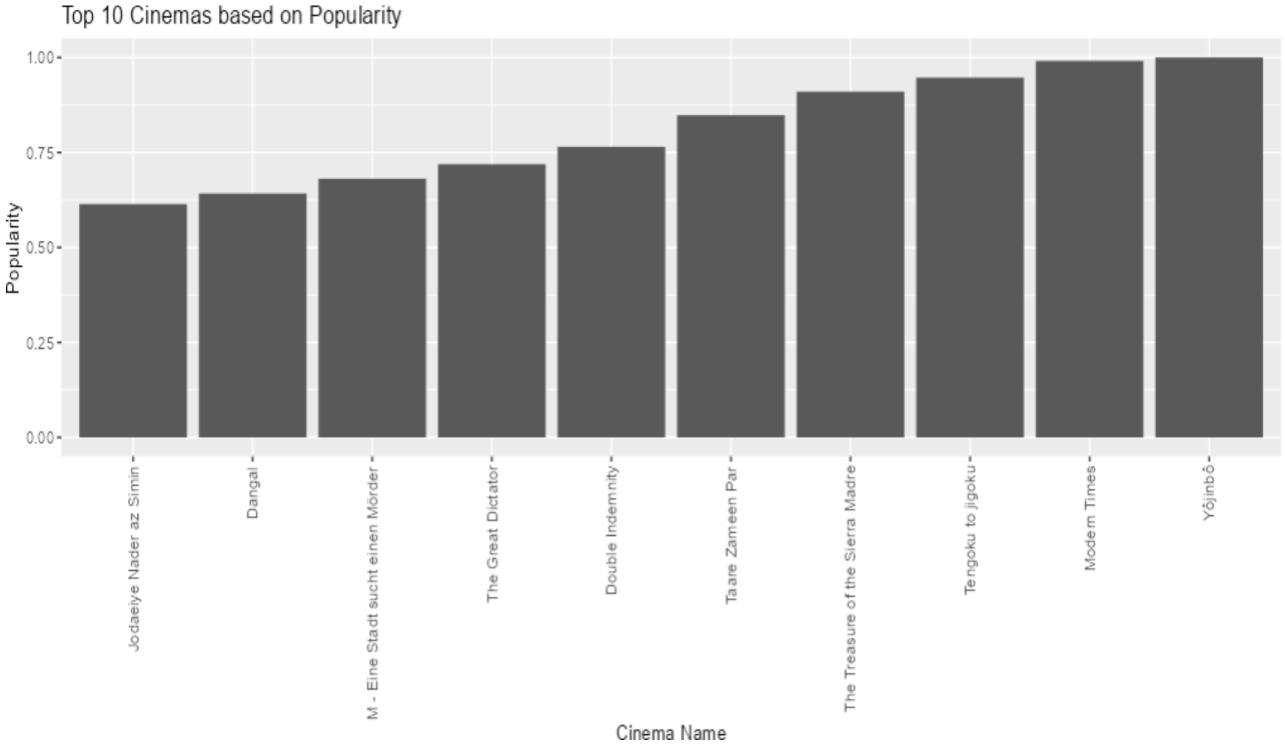
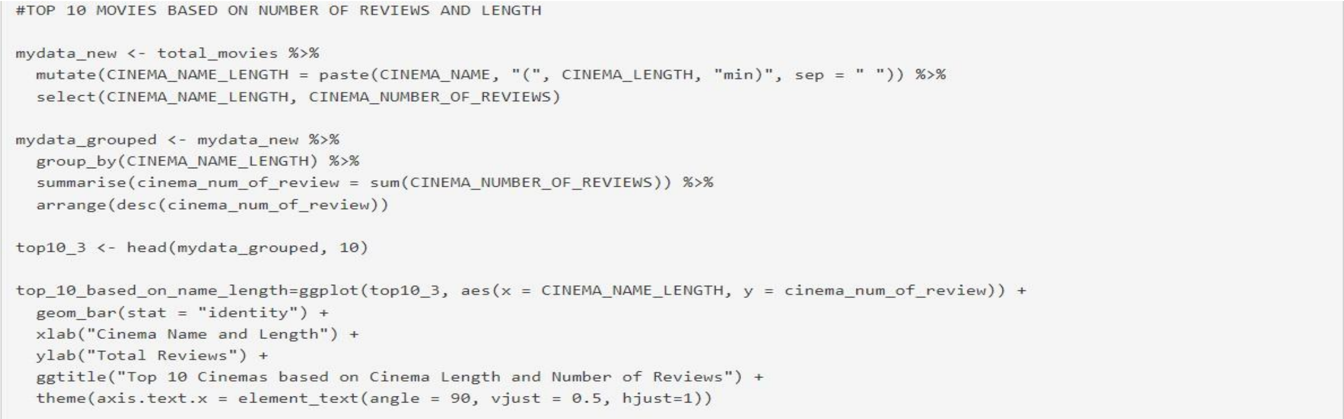


Figure 10: PLOTTING OF TOP 10 MOVIES BASED ON POPULARITY



```
#TOP 10 MOVIES BASED ON NUMBER OF REVIEWS AND LENGTH

mydata_new <- total_movies %>%
  mutate(CINEMA_NAME_LENGTH = paste(CINEMA_NAME, "(", CINEMA_LENGTH, "min)", sep = " ")) %>%
  select(CINEMA_NAME_LENGTH, CINEMA_NUMBER_OF_REVIEWS)

mydata_grouped <- mydata_new %>%
  group_by(CINEMA_NAME_LENGTH) %>%
  summarise(cinema_num_of_review = sum(CINEMA_NUMBER_OF_REVIEWS)) %>%
  arrange(desc(cinema_num_of_review))

top10_3 <- head(mydata_grouped, 10)

top_10_based_on_name_length=ggplot(top10_3, aes(x = CINEMA_NAME_LENGTH, y = cinema_num_of_review)) +
  geom_bar(stat = "identity") +
  xlab("Cinema Name and Length") +
  ylab("Total Reviews") +
  ggtitle("Top 10 Cinemas based on Cinema Length and Number of Reviews") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

Figure 11: TOP 10 MOVIES BASED ON POPULARITY AND LENGTH

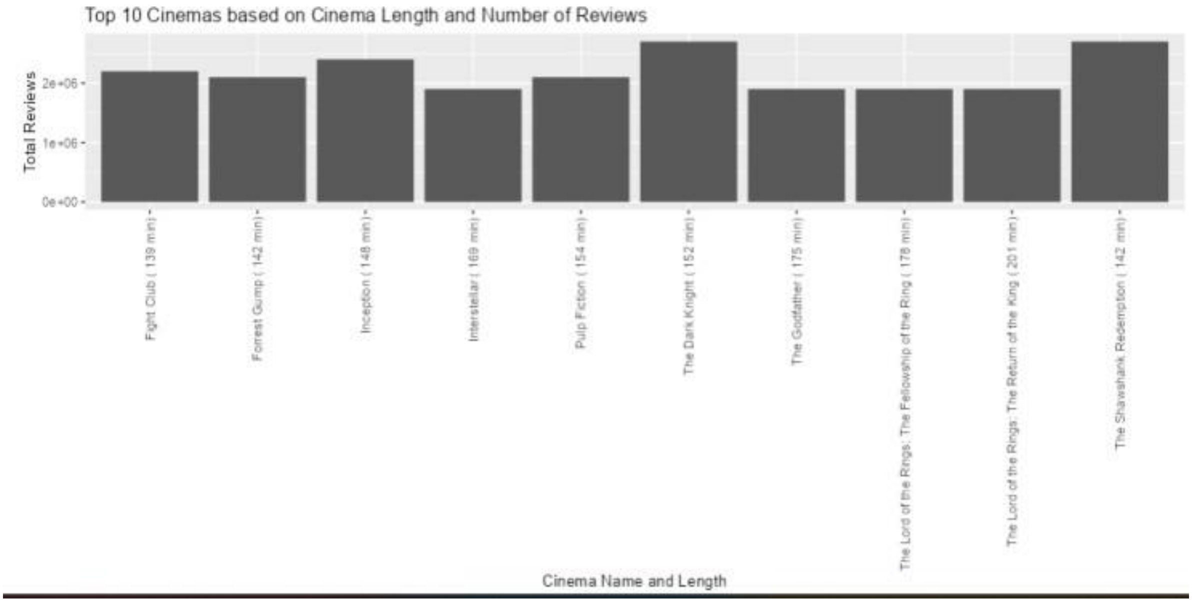
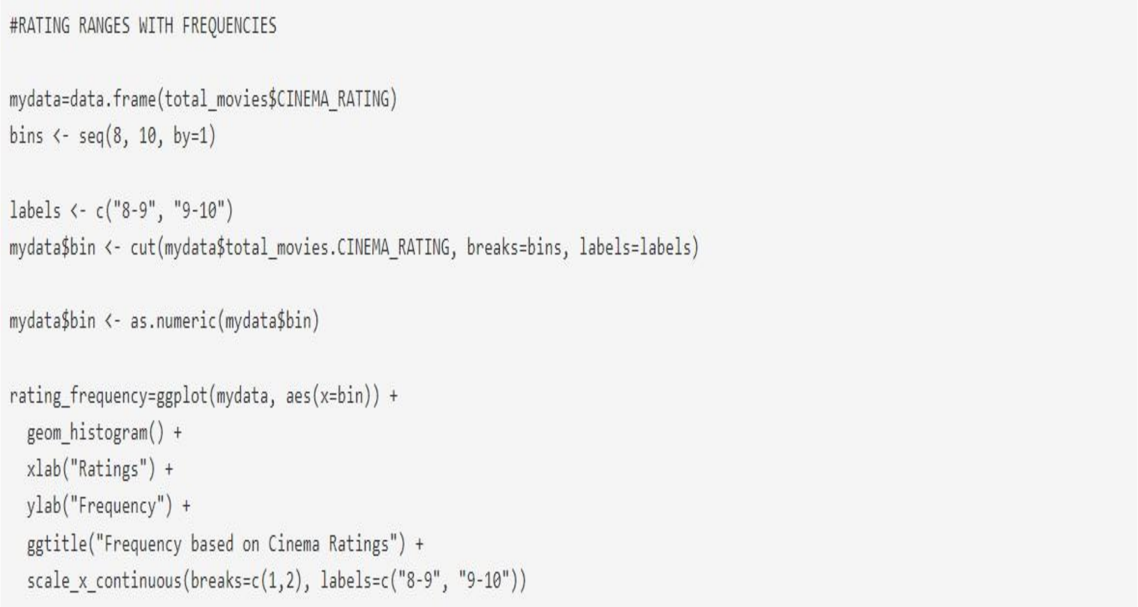


Figure 12: PLOTTING OF TOP 10 MOVIES BASED ON NUMBER OF REVIEWS



```
#RATING RANGES WITH FREQUENCIES

mydata=data.frame(total_movies$CINEMA_RATING)
bins <- seq(8, 10, by=1)

labels <- c("8-9", "9-10")
mydata$bin <- cut(mydata$total_movies.CINEMA_RATING, breaks=bins, labels=labels)

mydata$bin <- as.numeric(mydata$bin)

rating_frequency=ggplot(mydata, aes(x=bin)) +
  geom_histogram() +
  xlab("Ratings") +
  ylab("Frequency") +
  ggtitle("Frequency based on Cinema Ratings") +
  scale_x_continuous(breaks=c(1,2), labels=c("8-9", "9-10"))
```

Figure 13: FREQUENCY BASED ON CINEMA RATINGS

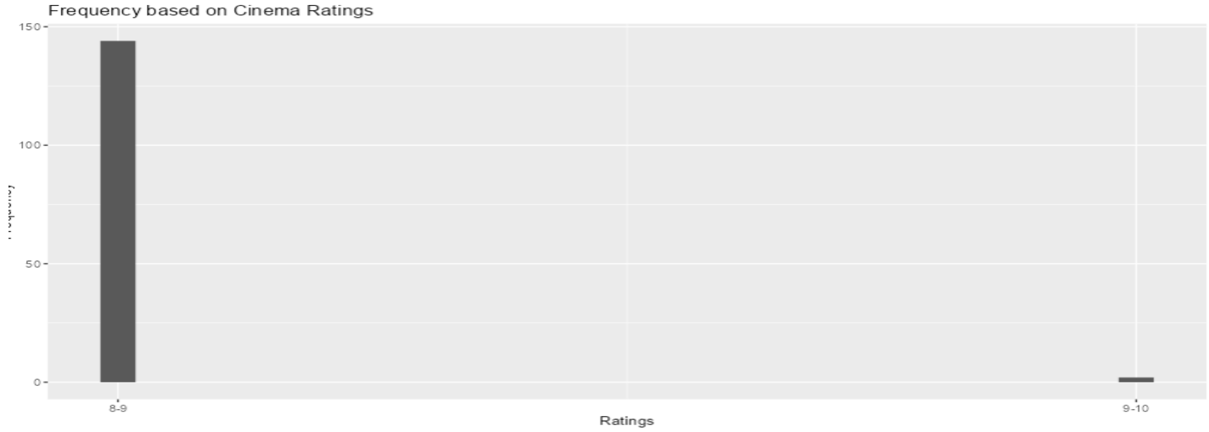
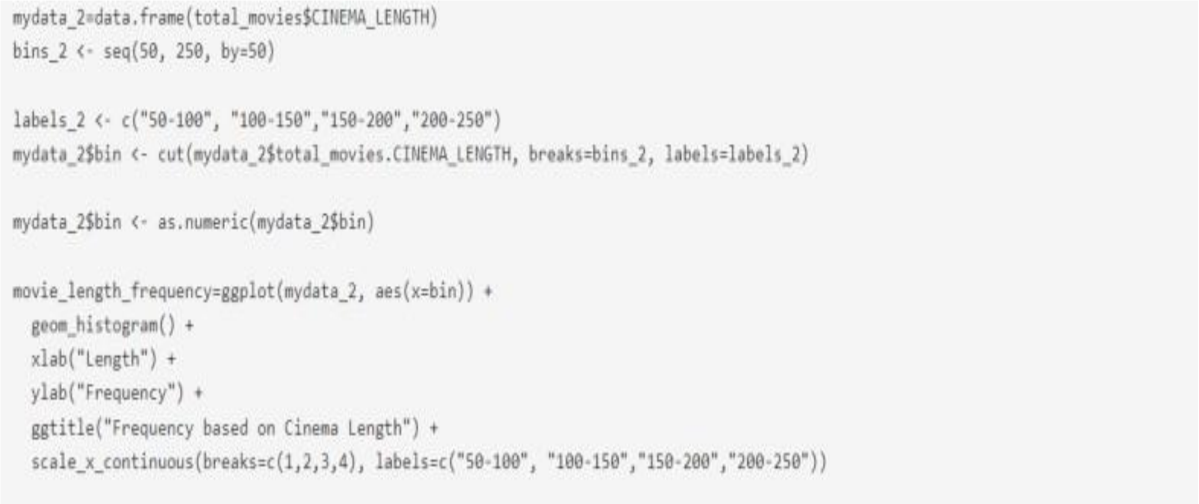


Figure 14: PLOTTING OF FREQUENCY BASED ON CINEMA RATINGS



```
mydata_2=data.frame(total_movies$CINEMA_LENGTH)
bins_2 <- seq(50, 250, by=50)

labels_2 <- c("50-100", "100-150","150-200","200-250")
mydata_2$bin <- cut(mydata_2$total_movies.CINEMA_LENGTH, breaks=bins_2, labels=labels_2)

mydata_2$bin <- as.numeric(mydata_2$bin)

movie_length_frequency=ggplot(mydata_2, aes(x=bin)) +
  geom_histogram() +
  xlab("Length") +
  ylab("Frequency") +
  ggtitle("Frequency based on Cinema Length") +
  scale_x_continuous(breaks=c(1,2,3,4), labels=c("50-100", "100-150","150-200","200-250"))
```

Figure 15: FREQUENCY BASED ON CINEMA LENGTHS

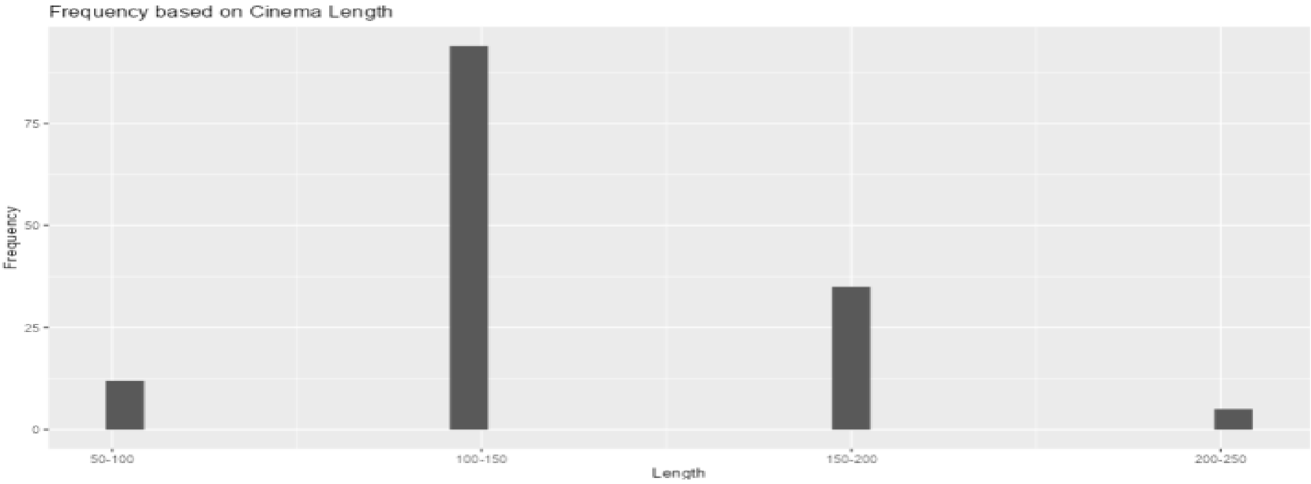
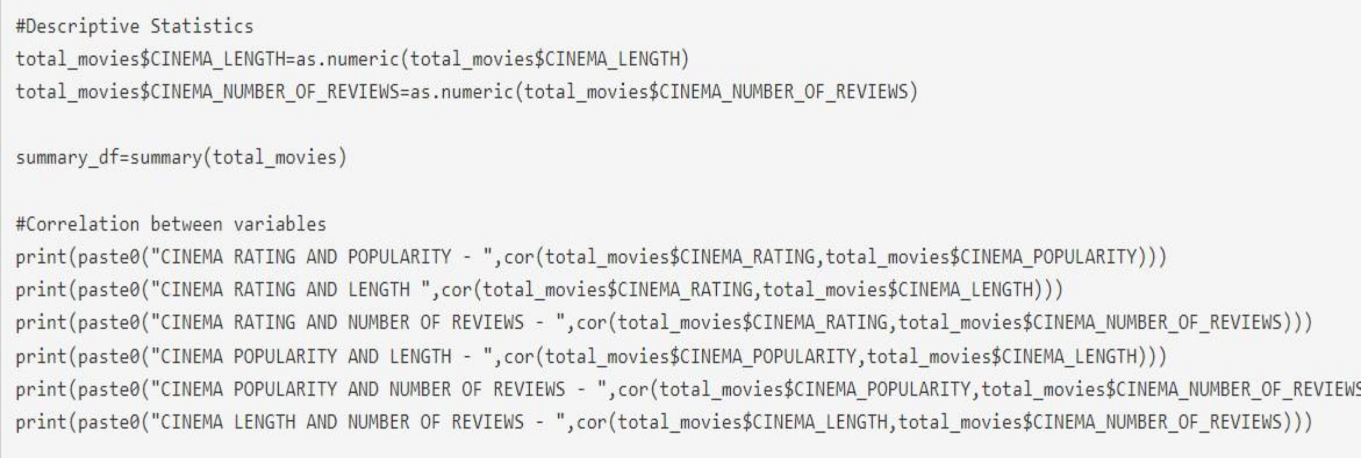


Figure 16: PLOTTING OF FREQUENCY BASED ON CINEMA LENGTHS

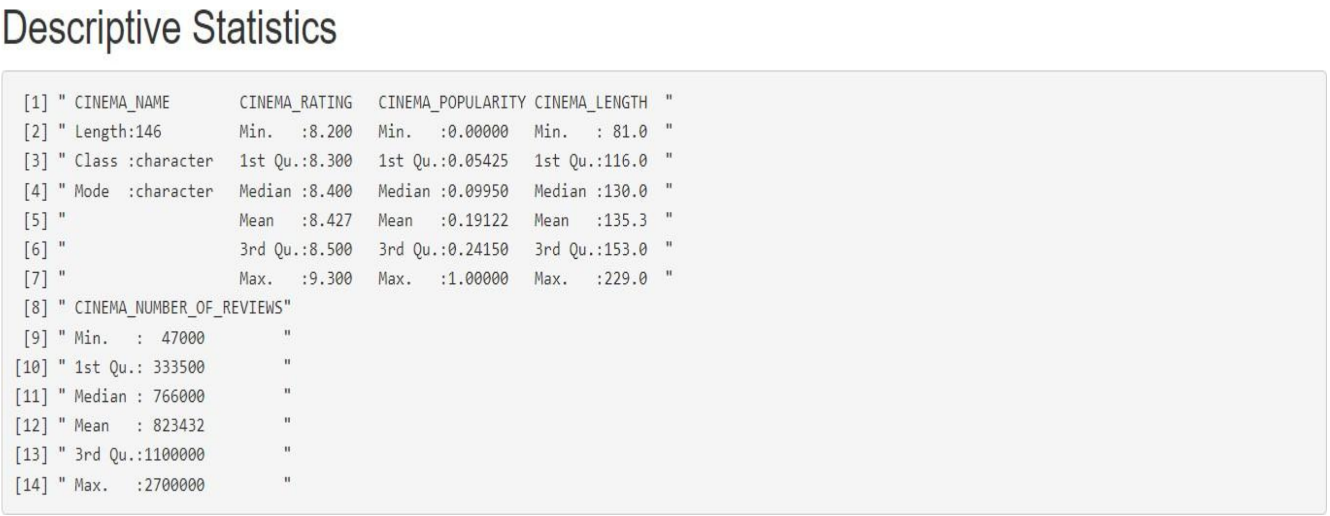
1.6Descriptive Statistics

Data can be described or summarized using descriptive statistics in relevant and practical ways.



```
#Descriptive Statistics
total_movies$CINEMA_LENGTH=as.numeric(total_movies$CINEMA_LENGTH)
total_movies$CINEMA_NUMBER_OF_REVIEWS=as.numeric(total_movies$CINEMA_NUMBER_OF_REVIEWS)

summary_df=summary(total_movies)

#Correlation between variables
print(paste0("CINEMA RATING AND POPULARITY - ",cor(total_movies$CINEMA_RATING,total_movies$CINEMA_POPULARITY)))
print(paste0("CINEMA RATING AND LENGTH ",cor(total_movies$CINEMA_RATING,total_movies$CINEMA_LENGTH)))
print(paste0("CINEMA RATING AND NUMBER OF REVIEWS - ",cor(total_movies$CINEMA_RATING,total_movies$CINEMA_NUMBER_OF_REVIEWS)))
print(paste0("CINEMA POPULARITY AND LENGTH - ",cor(total_movies$CINEMA_POPULARITY,total_movies$CINEMA_LENGTH)))
print(paste0("CINEMA POPULARITY AND NUMBER OF REVIEWS - ",cor(total_movies$CINEMA_POPULARITY,total_movies$CINEMA_NUMBER_OF_REVIEWS
print(paste0("CINEMA LENGTH AND NUMBER OF REVIEWS - ",cor(total_movies$CINEMA_LENGTH,total_movies$CINEMA_NUMBER_OF_REVIEWS)))
```

Figure 17:Descriptive Statistics Code

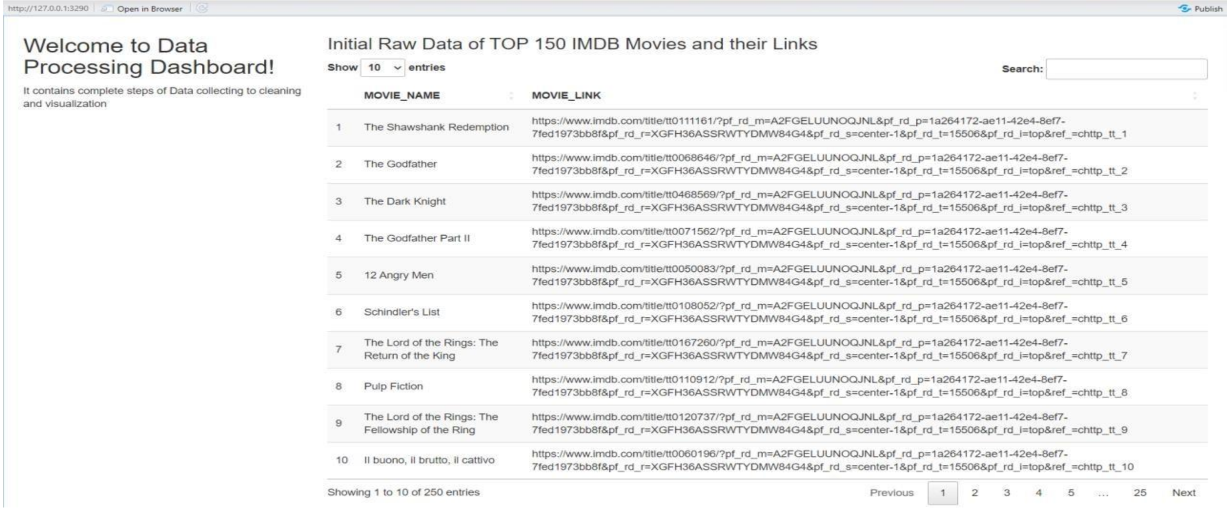


Descriptive Statistics

```
 [1] " CINEMA_NAME        CINEMA_RATING   CINEMA_POPULARITY CINEMA_LENGTH  "
 [2] " Length:146         Min.   :8.200   Min.   :0.00000   Min.   : 81.0  "
 [3] " Class :character   1st Qu.:8.300   1st Qu.:0.05425   1st Qu.:116.0  "
 [4] " Mode  :character   Median :8.400   Median :0.09950   Median :130.0  "
 [5] "                    Mean   :8.427   Mean   :0.19122   Mean   :135.3  "
 [6] "                    3rd Qu.:8.500   3rd Qu.:0.24150   3rd Qu.:153.0  "
 [7] "                    Max.   :9.300   Max.   :1.00000   Max.   :229.0  "
 [8] " CINEMA_NUMBER_OF_REVIEWS"
 [9] " Min.   :  47000        "
[10] " 1st Qu.: 333500        "
[11] " Median : 766000        "
[12] " Mean   : 823432        "
[13] " 3rd Qu.:1100000        "
[14] " Max.   :2700000        "
```

Figure 18: Descriptive Statistics Output

1.7Shiny Dashboard



Welcome to Data Processing Dashboard!

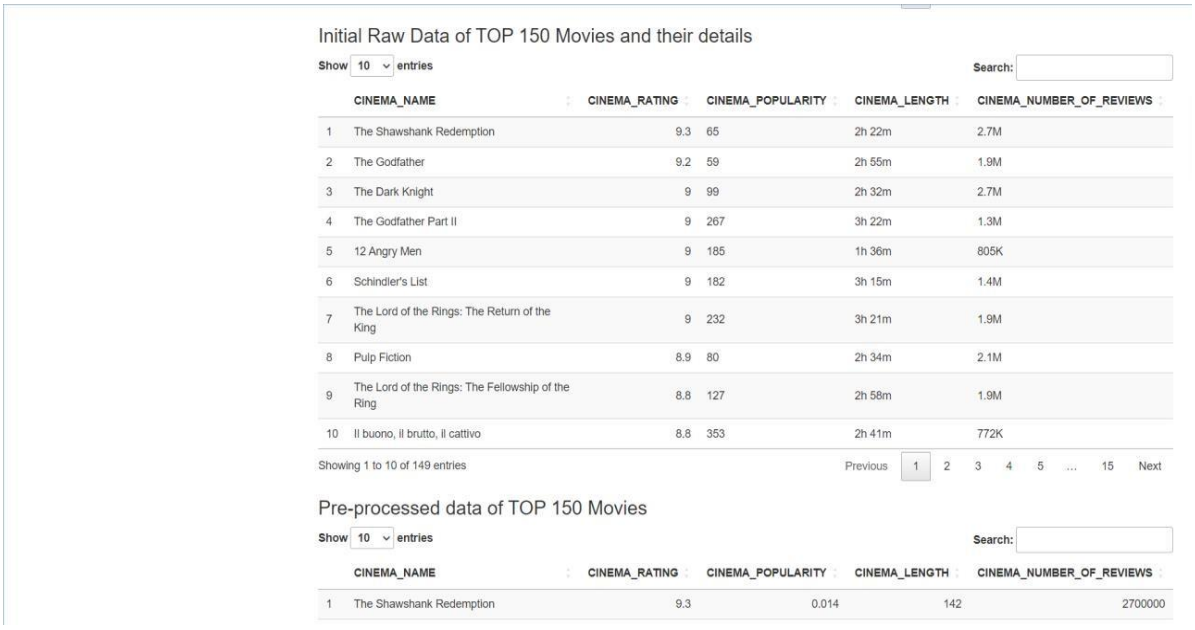
It contains complete steps of Data collecting to cleaning and visualization

Initial Raw Data of TOP 150 IMDB Movies and their Links

Show 10 entries　　Search:

	MOVIE_NAME	MOVIE_LINK
1	The Shawshank Redemption	https://www.imdb.com/title/tt0111161/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_1
2	The Godfather	https://www.imdb.com/title/tt0068646/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_2
3	The Dark Knight	https://www.imdb.com/title/tt0468569/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_3
4	The Godfather Part II	https://www.imdb.com/title/tt0071562/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_4
5	12 Angry Men	https://www.imdb.com/title/tt0050083/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_5
6	Schindler's List	https://www.imdb.com/title/tt0108052/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_6
7	The Lord of the Rings: The Return of the King	https://www.imdb.com/title/tt0167260/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_7
8	Pulp Fiction	https://www.imdb.com/title/tt0110912/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_8
9	The Lord of the Rings: The Fellowship of the Ring	https://www.imdb.com/title/tt0120737/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_9
10	Il buono, il brutto, il cattivo	https://www.imdb.com/title/tt0060196/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=1a264172-ae11-42e4-8ef7-7fed1973bb8f&pf_rd_r=XGFH36ASSRWTYDMW84G4&pf_rd_s=center-1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_10

Showing 1 to 10 of 250 entries



Initial Raw Data of TOP 150 Movies and their details

Show 10 entries

Search:

	CINEMA_NAME	CINEMA_RATING	CINEMA_POPULARITY	CINEMA_LENGTH	CINEMA_NUMBER_OF_REVIEWS
1	The Shawshank Redemption	9.3	65	2h 22m	2.7M
2	The Godfather	9.2	59	2h 55m	1.9M
3	The Dark Knight	9	99	2h 32m	2.7M
4	The Godfather Part II	9	267	3h 22m	1.3M
5	12 Angry Men	9	185	1h 36m	805K
6	Schindler's List	9	182	3h 15m	1.4M
7	The Lord of the Rings: The Return of the King	9	232	3h 21m	1.9M
8	Pulp Fiction	8.9	80	2h 34m	2.1M
9	The Lord of the Rings: The Fellowship of the Ring	8.8	127	2h 58m	1.9M
10	Il buono, il brutto, il cattivo	8.8	353	2h 41m	772K

Showing 1 to 10 of 149 entries

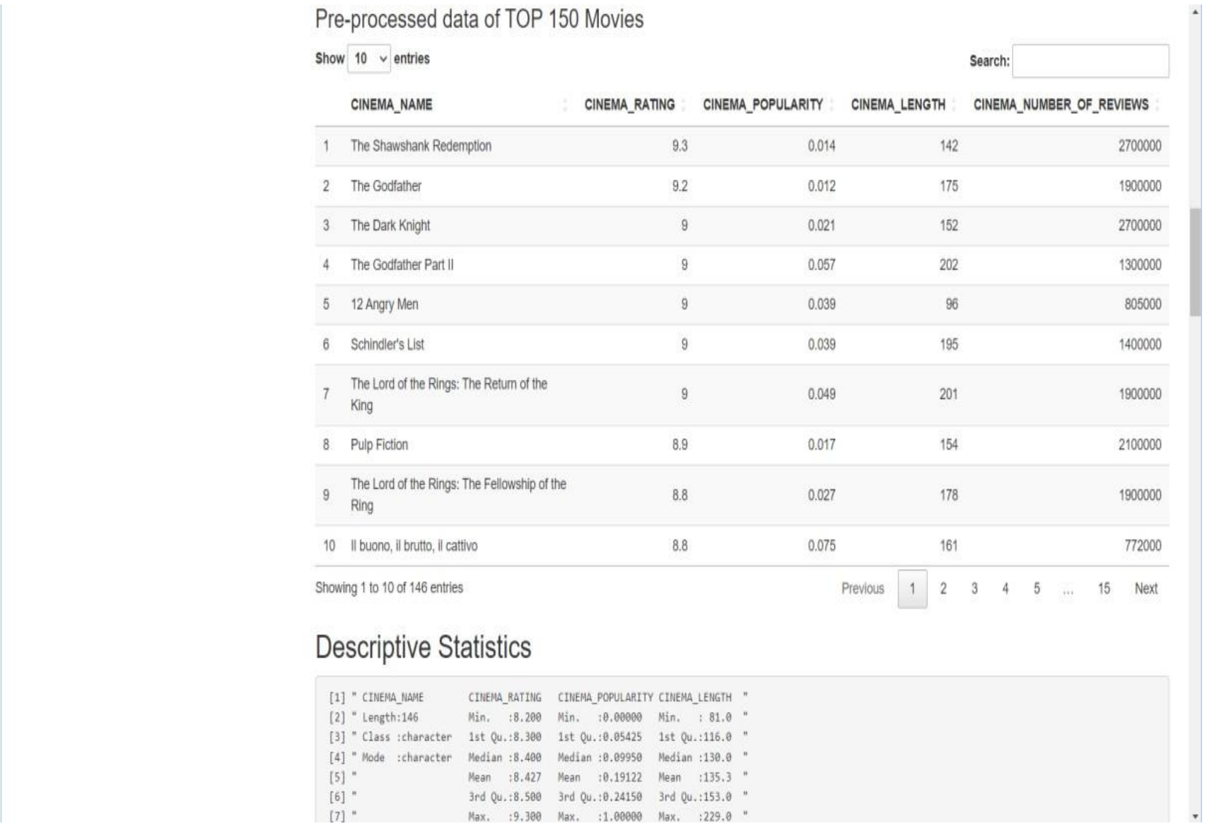
Previous 1 2 3 4 5 … 15 Next

Pre-processed data of TOP 150 Movies

Show 10 entries

Search:

	CINEMA_NAME	CINEMA_RATING	CINEMA_POPULARITY	CINEMA_LENGTH	CINEMA_NUMBER_OF_REVIEWS
1	The Shawshank Redemption	9.3	0.014	142	2700000



Pre-processed data of TOP 150 Movies

Show 10 entries

Search:

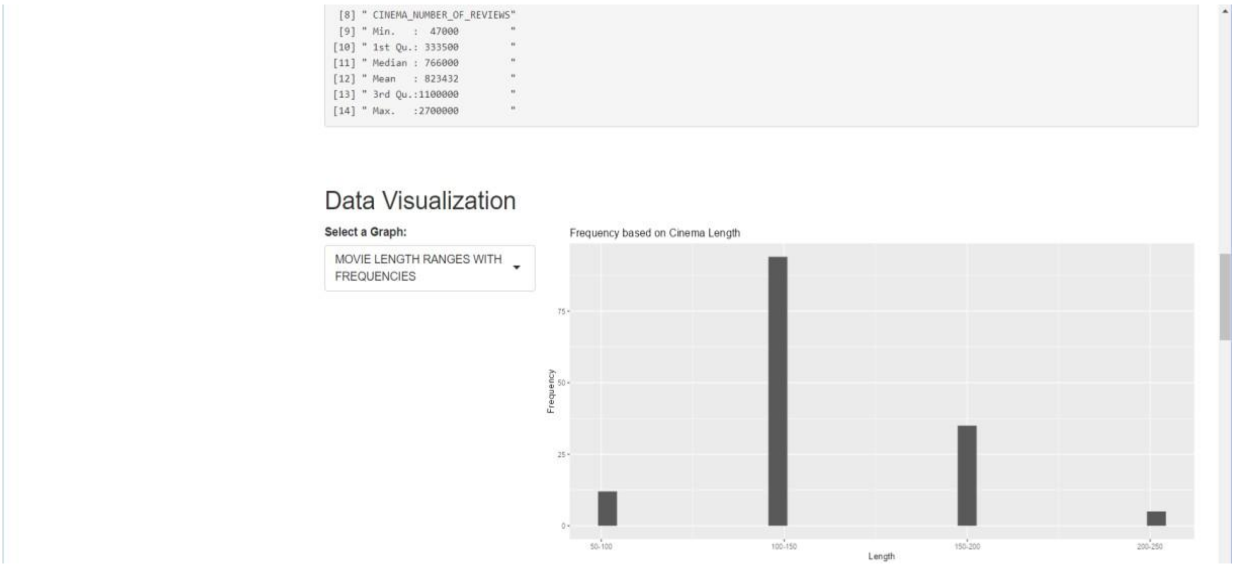
	CINEMA_NAME	CINEMA_RATING	CINEMA_POPULARITY	CINEMA_LENGTH	CINEMA_NUMBER_OF_REVIEWS
1	The Shawshank Redemption	9.3	0.014	142	2700000
2	The Godfather	9.2	0.012	175	1900000
3	The Dark Knight	9	0.021	152	2700000
4	The Godfather Part II	9	0.057	202	1300000
5	12 Angry Men	9	0.039	96	805000
6	Schindler's List	9	0.039	195	1400000
7	The Lord of the Rings: The Return of the King	9	0.049	201	1900000
8	Pulp Fiction	8.9	0.017	154	2100000
9	The Lord of the Rings: The Fellowship of the Ring	8.8	0.027	178	1900000
10	Il buono, il brutto, il cattivo	8.8	0.075	161	772000

Showing 1 to 10 of 146 entries

Previous 1 2 3 4 5 … 15 Next

Descriptive Statistics

```
[1] " CINEMA_NAME       CINEMA_RATING   CINEMA_POPULARITY CINEMA_LENGTH  "
[2] " Length:146        Min.   :8.200   Min.   :0.00000   Min.   : 81.0  "
[3] " Class :character  1st Qu.:8.300   1st Qu.:0.05425   1st Qu.:116.0  "
[4] " Mode  :character  Median :8.400   Median :0.09950   Median :130.0  "
[5] "                   Mean   :8.427   Mean   :0.19122   Mean   :135.3  "
[6] "                   3rd Qu.:8.500   3rd Qu.:0.24150   3rd Qu.:153.0  "
[7] "                   Max.   :9.300   Max.   :1.00000   Max.   :229.0  "
```

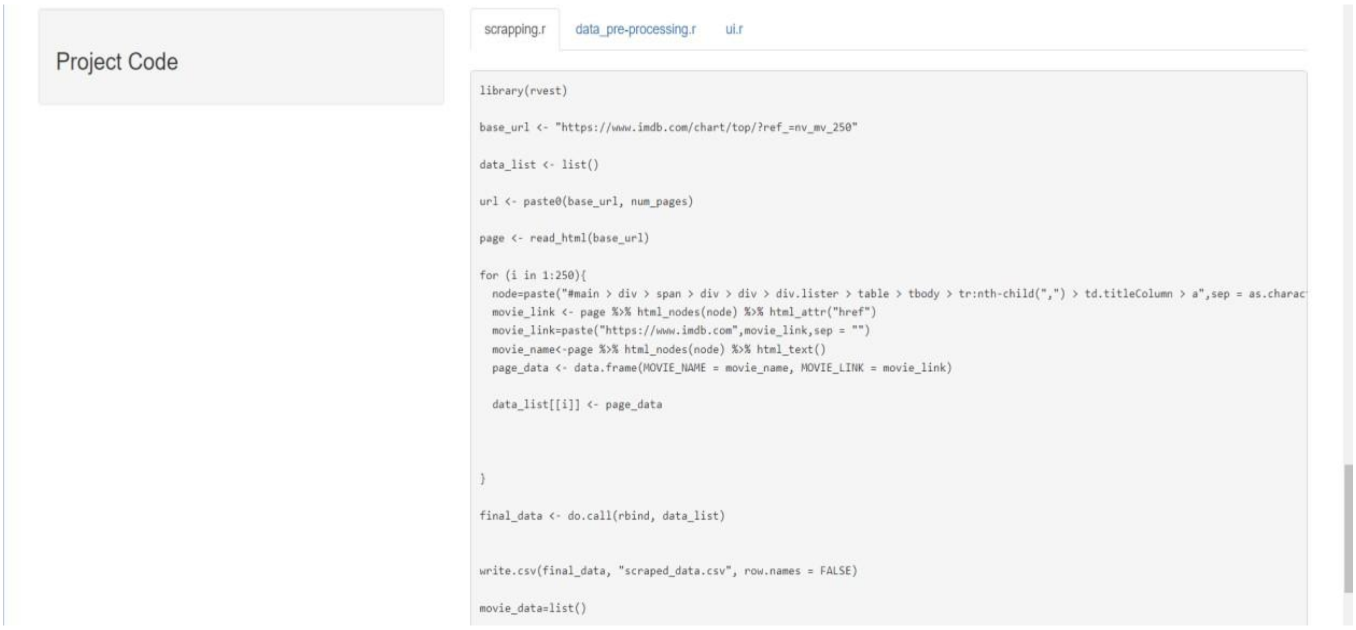


```
 [8] " CINEMA_NUMBER_OF_REVIEWS"
 [9] " Min.    :  47000        "
[10] " 1st Qu.: 333500        "
[11] " Median : 766000        "
[12] " Mean   : 823432        "
[13] " 3rd Qu.:1100000        "
[14] " Max.   :2700000        "
```

Data Visualization

Select a Graph: MOVIE LENGTH RANGES WITH FREQUENCIES

Frequency based on Cinema Length



Project Code

```
library(rvest)

base_url <- "https://www.imdb.com/chart/top/?ref_=nv_mv_250"

data_list <- list()

url <- paste0(base_url, num_pages)

page <- read_html(base_url)

for (i in 1:250){
  node=paste("#main > div > span > div > div > div.lister > table > tbody > tr:nth-child(",") > td.titleColumn > a",sep = as.charac
  movie_link <- page %>% html_nodes(node) %>% html_attr("href")
  movie_link=paste("https://www.imdb.com",movie_link,sep = "")
  movie_name<-page %>% html_nodes(node) %>% html_text()
  page_data <- data.frame(MOVIE_NAME = movie_name, MOVIE_LINK = movie_link)

  data_list[[i]] <- page_data



}

final_data <- do.call(rbind, data_list)


write.csv(final_data, "scraped_data.csv", row.names = FALSE)

movie_data=list()
```

1.8Discussion and Conclusion

This project was all about scraping data from imdb and do a complete analysis with that. The whole analysis has helped us gain knowledge with data and their workings which can be used in our further data analysis. As we analyzed the data of imdb, we saw the most popular movies are in between 1-2hr movies which means the audience like shorter lengths movies more and also, they gained the most rating which we can analyze from the graphs. With all the analysis we can conclude that if the movie makers focus of making moderate length movies, they can be more successful and generate more profit but also, they have to look intothe stories as well.