**Project Title: Web Scraping**

**Project Overview:**

We had to scrape data from a preferred website and make a dataset based on which we have to perform Data processing, Data visualization and Descriptive statistics.

For the web scrapping part, we chose a website called “MovieHaat” and scraped data from this website. The data set consists of 1323 observations and 7 variables.  
To prepare a cleaned dataset, we have to perform the following tasks of data pre-processing using R language-

1. Data cleaning

2. Data Integration

3. Data Transformation

4. Data Reduction

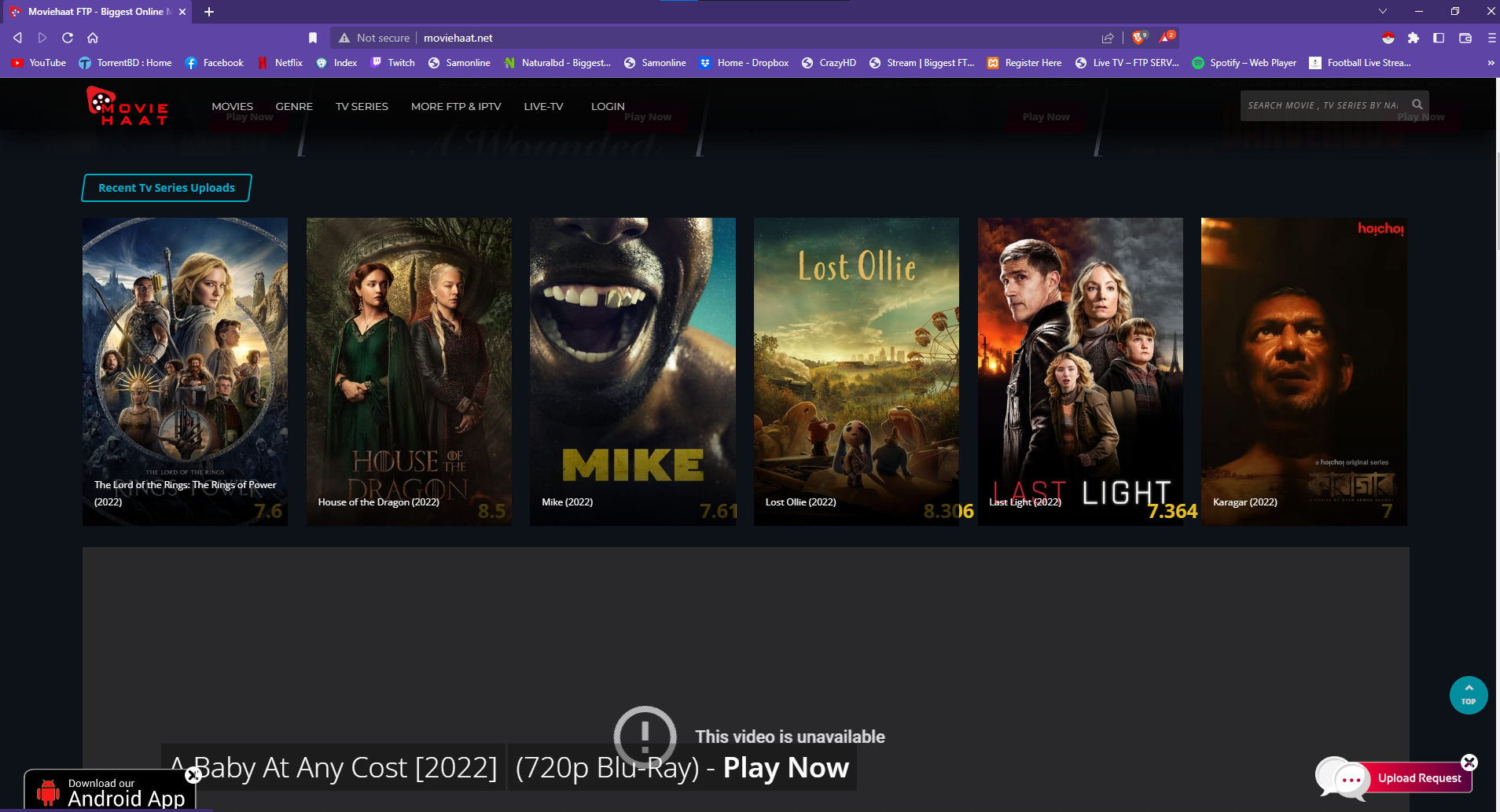
5. Data Discretization

After performing this step, we will have a process dataset ready to use.

**Project Solution Design:**

At first, we have to Create the dataset as CSV file. Then we have to import the dataset into RStudio so that we can perform the data processing operations using R language. We have to start with the Data cleaning process. Here we will clean the data such as we have to deal with missing values or smooth noisy data.Then we have to do transformation. Such as converting values into numerical or other types. After that data reduction here, we reduced the data which is unnecessary. And in data discretization we have to make the data set discrete.

**Data Collection via Web Scraping:**

****

For the web scraping part, we chose a website called “MovieHaat” and scraped data from this website. The data set consists of 1323 observations and 7 variables. We scraped about 23 pages and collected the information of the movies in those pages. Such as Movie Name, Year, Ratings, Quality, Industry, Genre and total number of Downloads. We used and web extension called instant data scrapper which allowed us to scrape data smoothly and easily. But there were some issues with those data. Which we handled in Data pre-processing that is elaborated below.

**Data pre-processing:**

**1.Data cleaning:**

At first, we have handled missing data in the given dataset. We have removed every missing data completely from the dataset.

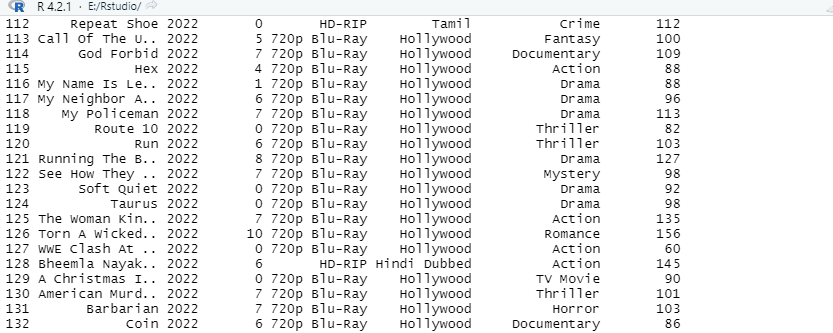
**RCODE: Removing Missing Values:**

movies\_1 <- movies

movies\_1[!apply(movies\_1 == "", 1, all),]

movies\_1

After Removing the missing values:



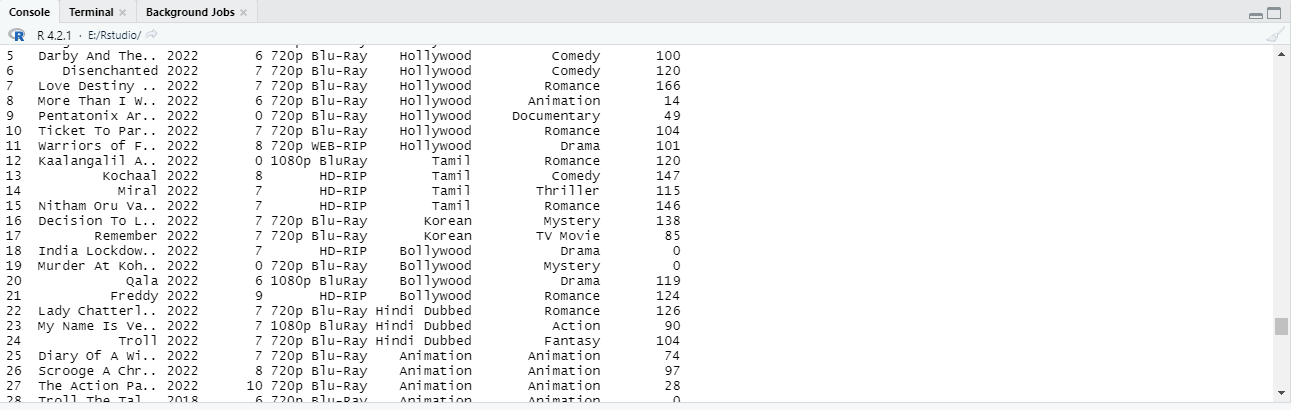
**2. Data Transformation:**

In This step we transformed the column Ratings as Numeric value.

After removing the missing values, we can see the column has now 4-digit decimal values. So, we have to format that as numeric.

**RCODE: #Data Formatting... To round up the Murder and Assault variable**

movies\_1$Ratings = as.numeric(format(round(movies\_1$Ratings, 0)))



After data transformation the dataset looks like this.

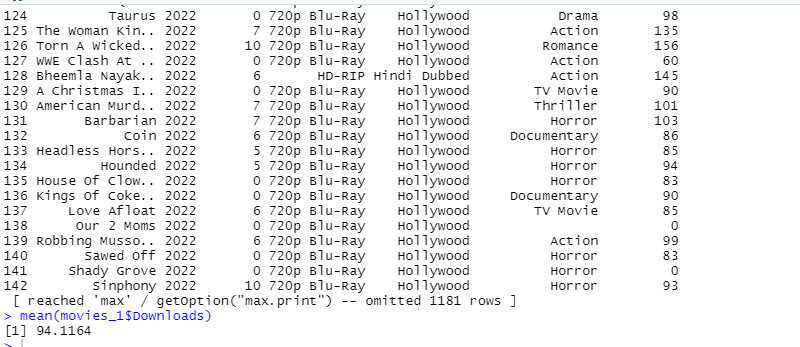
**Descriptive Statistics:**

**1. Descriptive Statistics - Mean:**

In descriptive Statistics first we did mean of movies\_1 dataset. In the data set we calculated the mean of Downloads column.

**RCODE: Mean of Downloads:**

mean(movies\_1$Downloads)

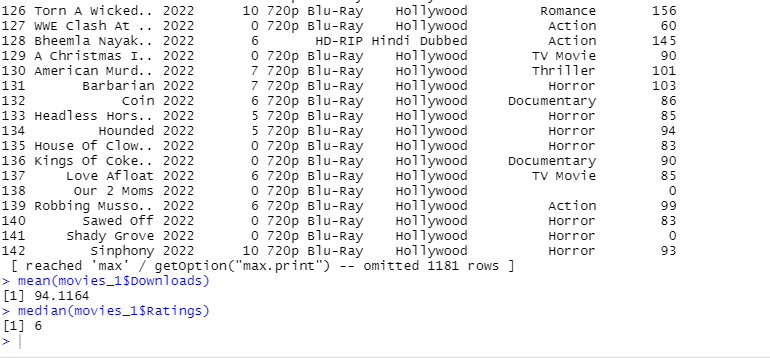


**2. Descriptive Statistics - Median:**

In descriptive Statistics secondly, we did median of movies\_1 dataset. In the data set we calculated the median of Ratings column.

**RCODE: Median of Downloads:**

median(movies\_1$Ratings)

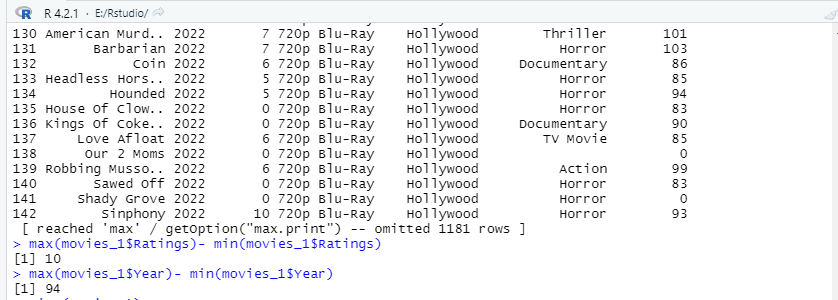


**3. Descriptive Statistics - Range:**

After that, we calculated the range of years the movies were downloaded from. In the data set we calculated the range of Year column.

**RCODE: Range of Year:**

max(movies\_1$Year)- min(movies\_1$Year)

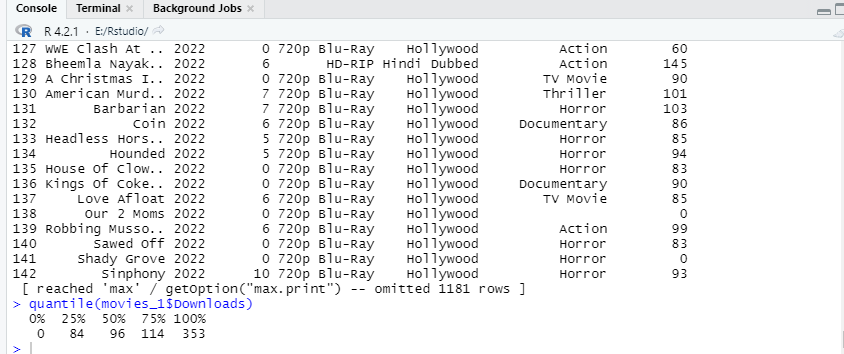
****

**4. Descriptive Statistics - Quantile:**

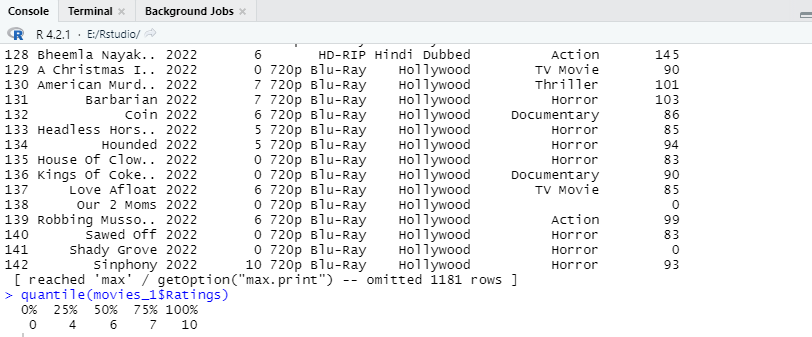
And lastly, we determined the quantile range of the dataset based on the column Downloads and Ratings from movies\_1 data frame.

**RCODE: Quantile(Downloads & Ratings):**

quantile(movies\_1$Downloads)



quantile (movies\_1$Ratings



**Data Visualization:**

* **Data Visualization - Scatterplot:**

We plotted a scatterplot based on our dataset. We used X axis as Ratings and Y axis as Downloads.   
How many downloads were made based on the ratings is shown in this graph we drawn using r language

**RCODE: ScatterPlot:**

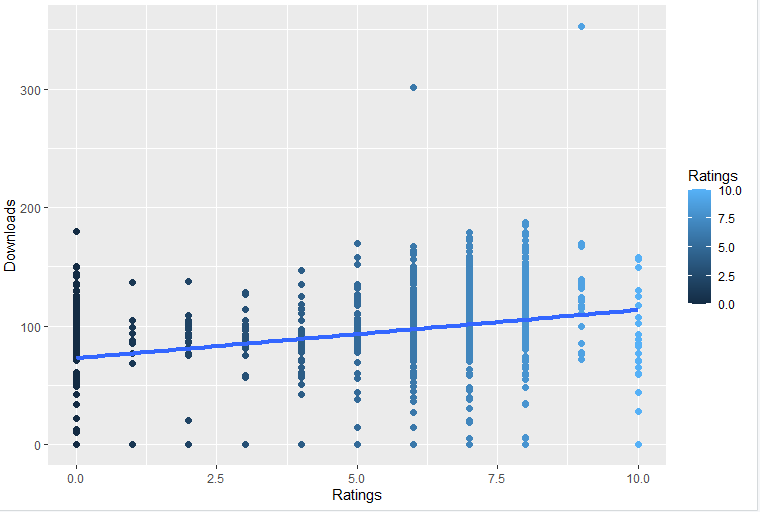
library(ggplot2)

ggplot(movies\_1 ,aes(x = Ratings, y = Downloads,

color=Ratings )) +

geom\_point(alpha = 5, size = 2) +

geom\_smooth(method = "lm", se = FALSE, size = 1.5)



* **Data Visualization - Bar Diagram:**

We plotted a Bar Diagram based on our dataset. We used X axis as Industry.  
How many downloads were made based on the Industry of the movies is shown in this graph we drawn using r language.

**RCODE: Bar Diagram:**

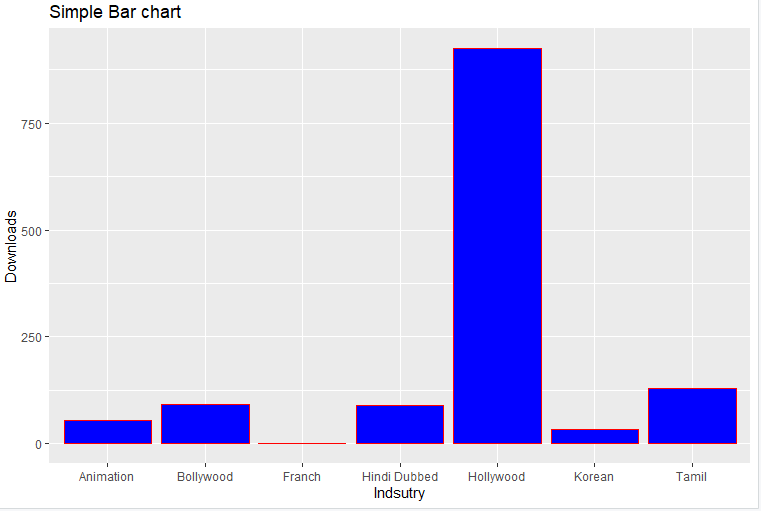
library(ggplot2)

ggplot(movies\_1, aes(x=Industry)) + geom\_bar(fill="blue", color="red") +

labs(title="Simple Bar chart",

x="Indsutry",

y="Downloads")



**Discussion and Conclusion:**

After doing date pre-processing operations in dataset we scraped from MovieHaat website, we can perform these steps in any datasets when we need. And data pre-processing helps AI or machine learning to easily analyze the data. We can also get an easy-to-understand dataset after doing these operations.We can also analysis the data by doing descriptive statistic methods and also clearly visualize the dataset based on what we want by doing scatterplot, bar diagram and so on. It makes machines to understand a huge data easily and properly without facing any problem and errors.