



## **AMERICAN INTERNATIONAL UNIVERSITY- BANGLADESH**

**Project Title: Web Scraping**

**Submitted By:**

**Name: Sayem, Easinul Abedin**

**ID:19-40291-1**

**Section:[C]**

**Subject: INTRODUCTION TO DATA SCIENCE**

**Submitted To:**

**Teachers Name: DR.AKINUL ISLAM JONY**

**Submission Date:10/12/2022**

## **Project overview:**

This Project is based on Web Scraping. This Project is collect data from a website as well as store the data into a csv file. This project we will analyse the dataset contains statistics in top rated indian movies (250). There are sequent, moviename, year as well as rating. Firstly we are install this packages [ `install.packages("rvest")` and `install.packages("dplyr")`] in RStudio. Then call this library fuction. After that change this chrome extensions as well as install and click for SelectorGadget. Then we are copy this url link for website.

## **Project solution design:**

The automated technique of extracting data from webpages is called "web scraping." Web scrapers, a type of software used for web scraping, are used to complete this procedure. According to user needs, they automatically load and extract data from the websites.

The process of cleaning up raw data and turning it into usable information is known as data preparation. The data can then be used for a variety of tasks, including reporting and decision making. Data pre- processing techniques are used when the data is inconsistent, which indicates that the data is not recorded in accordance with the constraints on the column, noisy, which may contain numerous errors or outliers, and incomplete, which indicates that some attribute value is missing.

The database may have missing data for a number of reasons, such as that the column needs a value to be entered or that the data was not captured at the time of data collection. Several methods can be used to fill in the missing values.

For obtaining summary statistics, R has a large variety of functions. The `sapply()` function can be used in conjunction with a given summary statistic to produce descriptive statistics. Mean, sd, var, min, max, median, range, and quantile are examples of potential functions that can be utilized in `sapply`.

Data visualization is a method for presenting insights in data through the use of visual signals like graphs, charts, maps, and many others. This is beneficial because it makes it simple and intuitive to interpret complex data sets, which enables users to make wiser decisions.

Data visualization is a technique for displaying data insights by utilizing visual cues like graphs, charts, maps, and more. This helps people make better decisions since it makes it easy and intuitive to analyze complex data sets.

R is a programming language intended for scientific research, graphical data analysis, and statistical computing. It is typically chosen for data visualization since it provides flexibility and requires little coding thanks to its packages.

## **Web Scraping:**

### **install csv file and Dataset:**

```
install.packages("rvest"
```

```
t")
```

```
install.packages("dplyr"
```

```
r")
```

```
library(rvest
```

```
)
```

```
library(dplyr
```

```
)
```

```
link = "https://www.imdb.com/india/top-rated-indian-
```

```
movies/" page = read_html(link)
```

```
MName = page %>% html_nodes(".titleColumn a") %>%
```

```
html_text() View(MName)
```

```
Year = page %>% html_nodes(".secondaryInfo") %>%
```

```
html_text() View(Year)
```

```
Rating = page %>% html_nodes("strong") %>%
```

```
html_text() View(Rating)
```

```
imdb = data.frame(MName,Year, Rating, stringsAsFactors = FALSE)
```

```
write.csv(imdb, "imdb.csv")
```

The screenshot displays the RStudio environment. The main editor shows a data frame 'imdb' with 250 entries and 3 columns: MName, Year, and Rating. The console shows the following R code:

```
R 4.2.1 ~/  
> Data$Year <- gsub("[()]", "", Data$Year)  
> library(tidyverse)  
>  
>  
> Data <- read.csv("imdb.csv");  
> Data$Year <- gsub("[()]", "", Data$Year)  
> View(Data)  
> View(imdb)
```

The Environment pane on the right shows the 'Data' environment with 250 observations of 4 variables. The Files pane shows the file explorer with various files and folders.

Then step by step , Data Cleaning >Data Integration > Data Transformation >Data Reduction  
>Data Discretization will be solved:

## **Data pre-processing:**

The process of cleaning up raw data and turning it into usable information is known as data preparation. The data can then be used for a variety of tasks, including reporting and decision making. Data pre- processing techniques are used when the data is inconsistent, which indicates that the data is not

recorded in accordance with the constraints on the column, noisy, which may contain numerous errors or outliers, and incomplete, which indicates that some attribute value is missing.

## **Data Cleaning:**

### **Smooth Noisy Data:**

Since the year is negative as this column is noisy.

**Before:**

The screenshot shows the RStudio interface. The main editor displays a data frame with the following columns: X, MName, Year, and Rating. The data is as follows:

X	MName	Year	Rating
1	Ramayana: The Legend of Prince Rama	-1993	8.5
2	Rocketry: The Nambi Effect	-2022	8.4
3	Golmaal	-1979	8.4
4	777 Charlie	-2022	8.4
5	Nayakan	-1987	8.4
6	Anbe Sivam	-2003	8.4
7	Jai Bhim	-2021	8.4
8	Pariyerum Perumal	-2018	8.4
9	3 Idiots	-2009	8.4

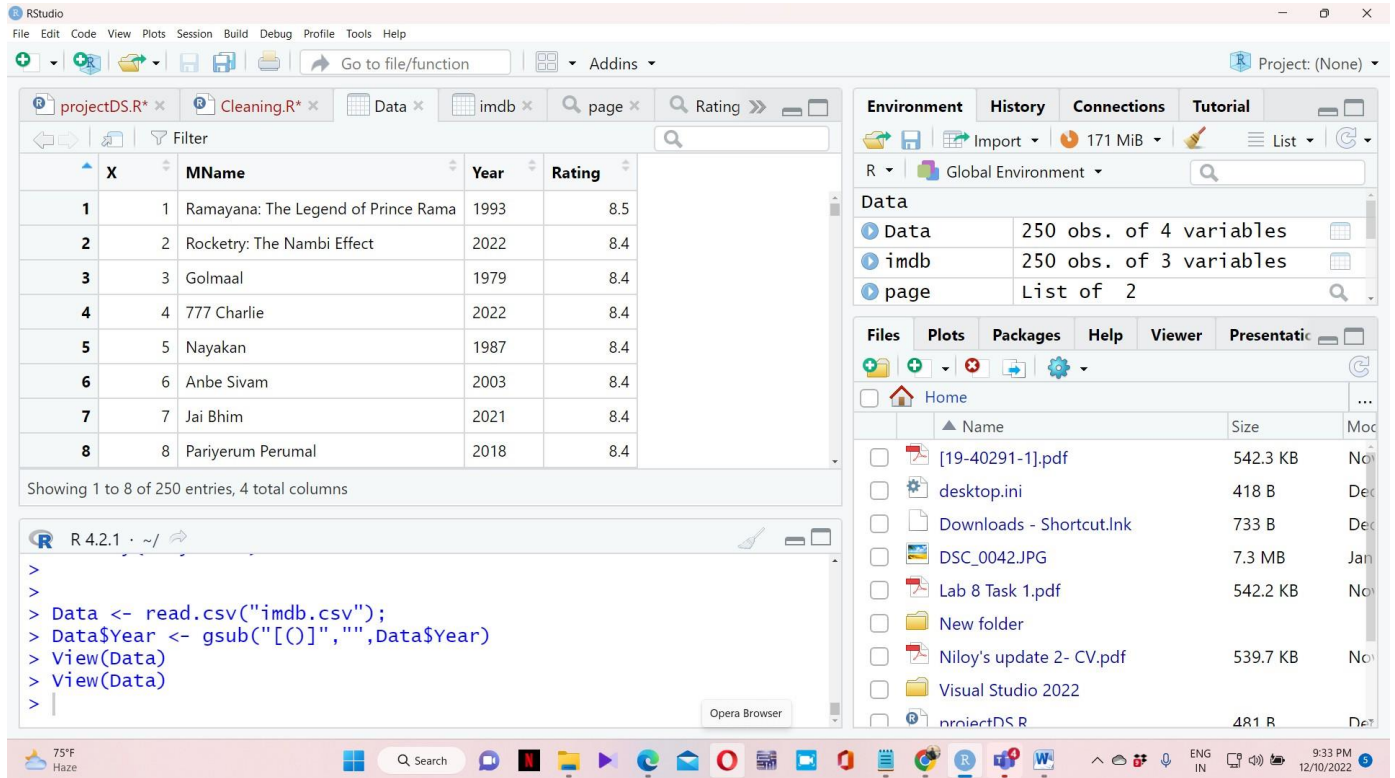
The console shows the following R code:

```
> write.csv(imdb, "imdb.csv")
> View(Data)
> View(imdb)
> View(page)
>
```

The environment pane on the right shows the 'page' variable with a list of 2 values:

```
l1: "8.5"
l2: "8.4"
l3: "8.4"
l4: "8.4"
l5: "8.4"
l6: "8.4"
l7: "8.4"
l8: "8.4"
l9: "8.4"
```

## **After:**



The screenshot displays the RStudio interface. The main window shows a data table with 8 rows and 4 columns: X, MName, Year, and Rating. The data is as follows:

X	MName	Year	Rating
1	Ramayana: The Legend of Prince Rama	1993	8.5
2	Rocketry: The Nambi Effect	2022	8.4
3	Golmaal	1979	8.4
4	777 Charlie	2022	8.4
5	Nayakan	1987	8.4
6	Anbe Sivam	2003	8.4
7	Jai Bhim	2021	8.4
8	Pariyerum Perumal	2018	8.4

The code editor shows the following R code:

```
> Data <- read.csv("imdb.csv");
> Data$Year <- gsub("[()]", "", Data$Year)
> View(Data)
> View(Data)
```

The environment pane on the right shows the following data objects:

- Data: 250 obs. of 4 variables
- imdb: 250 obs. of 3 variables
- page: List of 2

## **Handling Missing Data:**

No Handling Missing is necessary for the data set.

## **Data Munging:**

No munging or wrangling is necessary for the data set.

## **Data Integration:**

No Data Integration is necessary for the data set.

## **Data Transformation:**

As is well known, the data transformation process includes one or more of the following steps: normalization, summarization, noise removal, smoothing, and noise removal from data. We shall use smoothing in this example because it is easier than summarizing and normalizing.

## **Data Reduction:**

Data reduction aims to produce a reduced representation of the dataset that may be utilized to produce the same or comparable analytical results is what data reduction aims to produce.

## **Data Discretization:**

All of the attributes included in our dataset are continuous type, as can be seen (values in real numbers). The attribute values may need to be discretized into binary or categorical categories, though, depending on the model you wish to create.

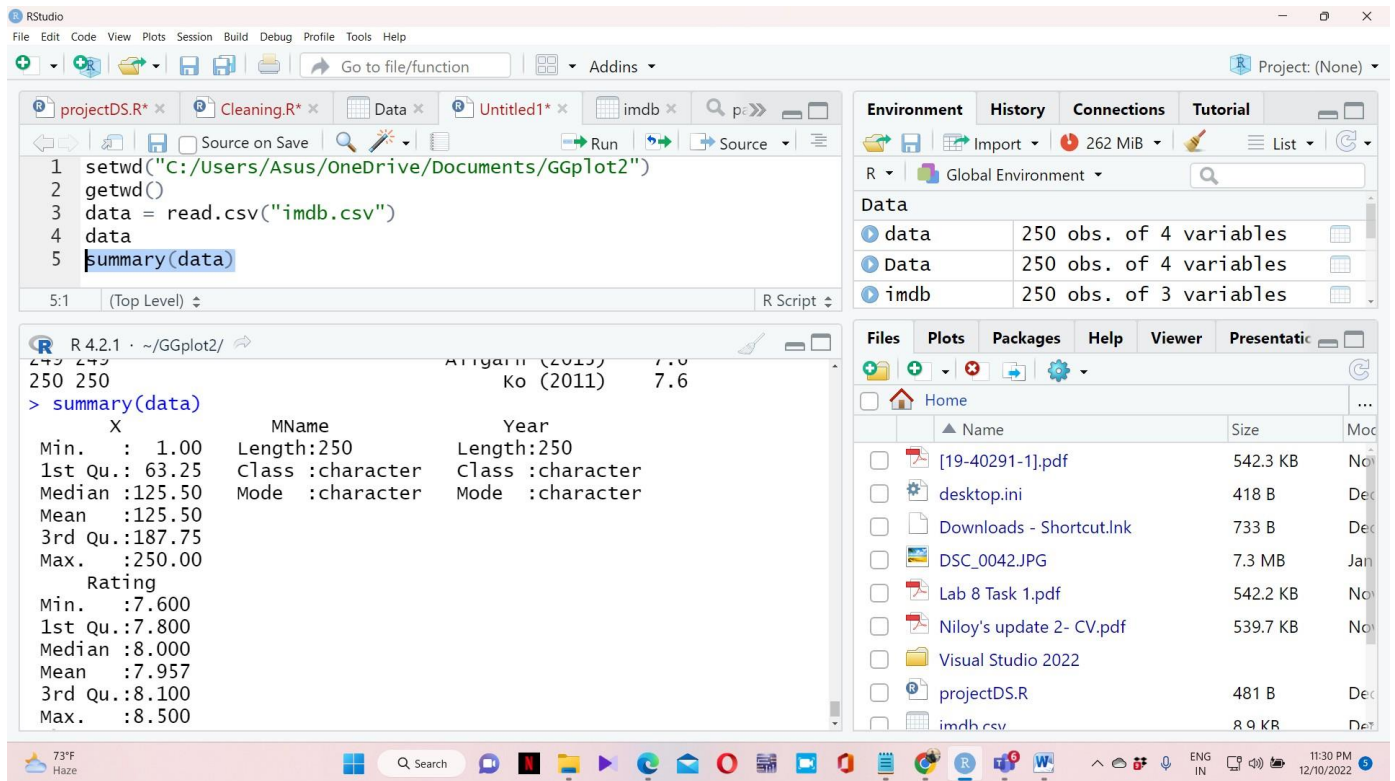
## **Descripting statistics:**

```
setwd("C:/Users/Asus/OneDrive/Documents/GGplo  
t2") getwd()  
data = read.csv("imdb.csv")
```

data

summary(dat

a)



## Data visualization:

setwd("C:/Users/Asus/OneDrive/Documents/GGplo

t2") getwd()

data =

read.csv("imdb.csv")

data

summary(dat

a) str(data)



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

projectDS.R\* Cleaning.R\* Data x Untitled1\* imdb x

```

1 setwd("C:/Users/Asus/OneDrive/Documents/GGplot2")
2 getwd()
3 data = read.csv("imdb.csv")
4 data
5 summary(data)
6 str(data)

```

6:1 (Top Level) R Script

R 4.2.1 ~ /GGplot2/

```

> str(data)
'data.frame': 250 obs. of 4 variables:
 $ X : int 1 2 3 4 5 6 7 8 9 10 ...
 $ MName : chr "Ramayana: The Legend of Prince Rama" "Rocketry: The Nambi Effect" "Golmaal" "777 Charlie" ...
 $ Year : chr "(1993)" "(2022)" "(1979)" "(2022)" ...
 $ Rating: num 8.5 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 ...
>

```

Environment History Connections Tutorial

R Global Environment

Data

Object	Size
data	250 obs. of 4 variables
Data	250 obs. of 4 variables
imdb	250 obs. of 3 variables

Files Plots Packages Help Viewer Presentation

Home

Name	Size
[19-40291-1].pdf	542.3 KB
desktop.ini	418 B
Downloads - Shortcut.lnk	733 B
DSC_0042.JPG	7.3 MB
Lab 8 Task 1.pdf	542.2 KB
Niloy's update 2- CV.pdf	539.7 KB
Visual Studio 2022	
projectDS.R	481 B
imdb.csv	8.9 KB

73°F Haze

Search

11:33 PM 12/10/2022

plot(data)

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

projectDS.R\* Cleaning.R\* Data x Untitled1\* imdb x

```

1 setwd("C:/Users/Asus/OneDrive/Documents/GGplot2")
2 getwd()
3 data = read.csv("imdb.csv")
4 data
5 summary(data)
6 str(data)
7 plot(data)

```

7:1 (Top Level) R Script

R 4.2.1 ~ /GGplot2/

```

'data.frame': 250 obs. of 4 variables:
 $ X : int 1 2 3 4 5 6 7 8 9 10 ...
 $ MName : chr "Ramayana: The Legend of Prince Rama" "Rocketry: T he Nambi Effect" "Golmaal" "777 Charlie" ...
 $ Year : chr "(1993)" "(2022)" "(1979)" "(2022)" ...
 $ Rating: num 8.5 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 ...
> plot(data)
>

```

Environment History Connections Tutorial

R Global Environment

Data

data 250 obs. of 4 variables

Files Plots Packages Help Viewer Presentation

Zoom Export

73°F Haze

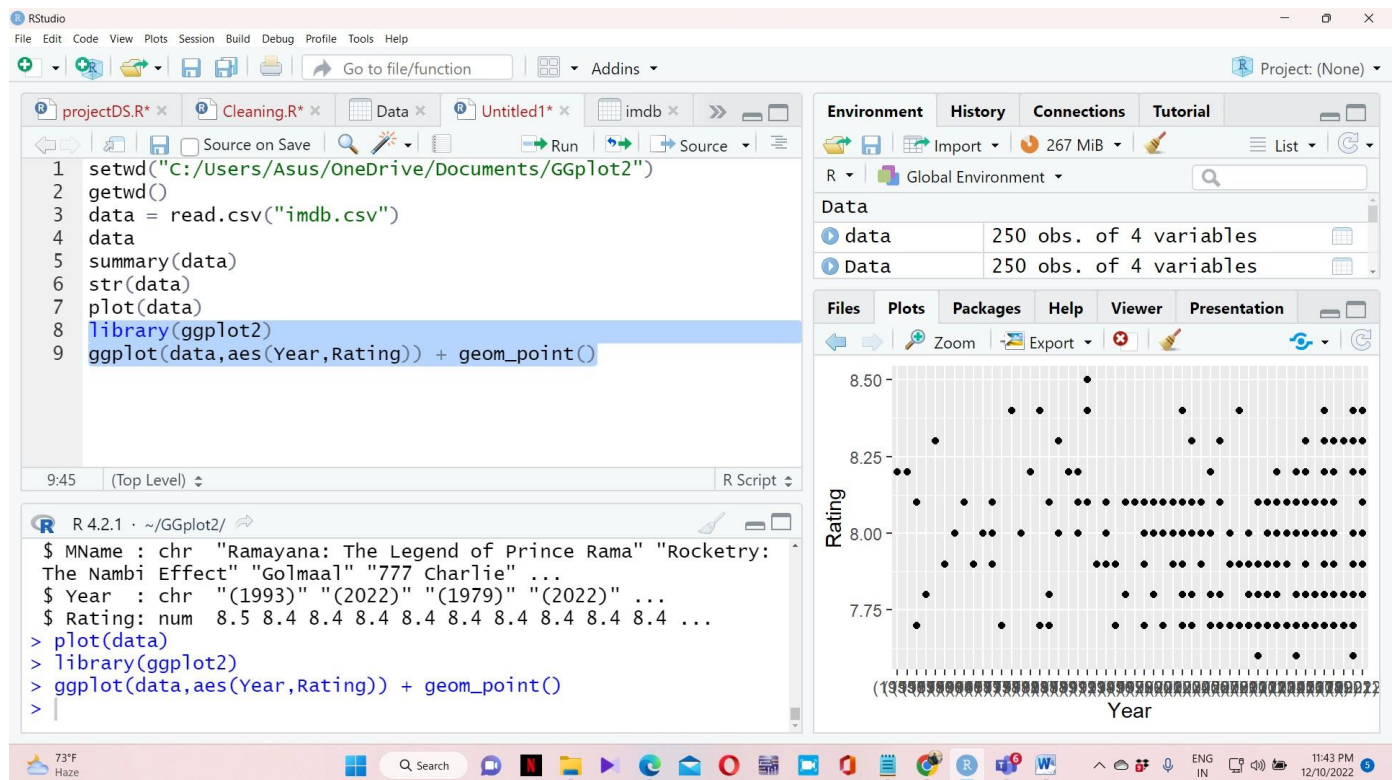
Search

11:34 PM 12/10/2022

```
library(ggplot2)

ggplot(data,aes(Year,Rating)) +

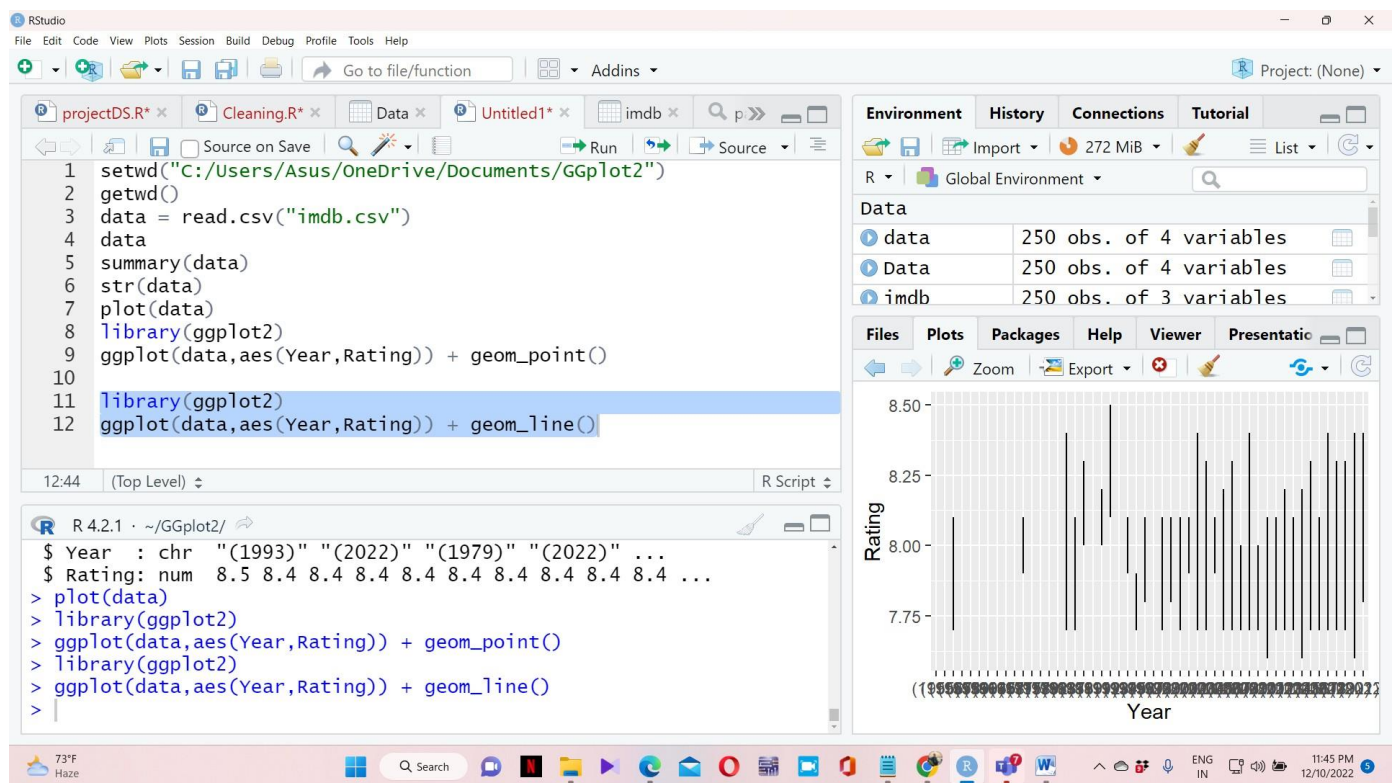
geom_point()
```



```
library(ggplot2)
```

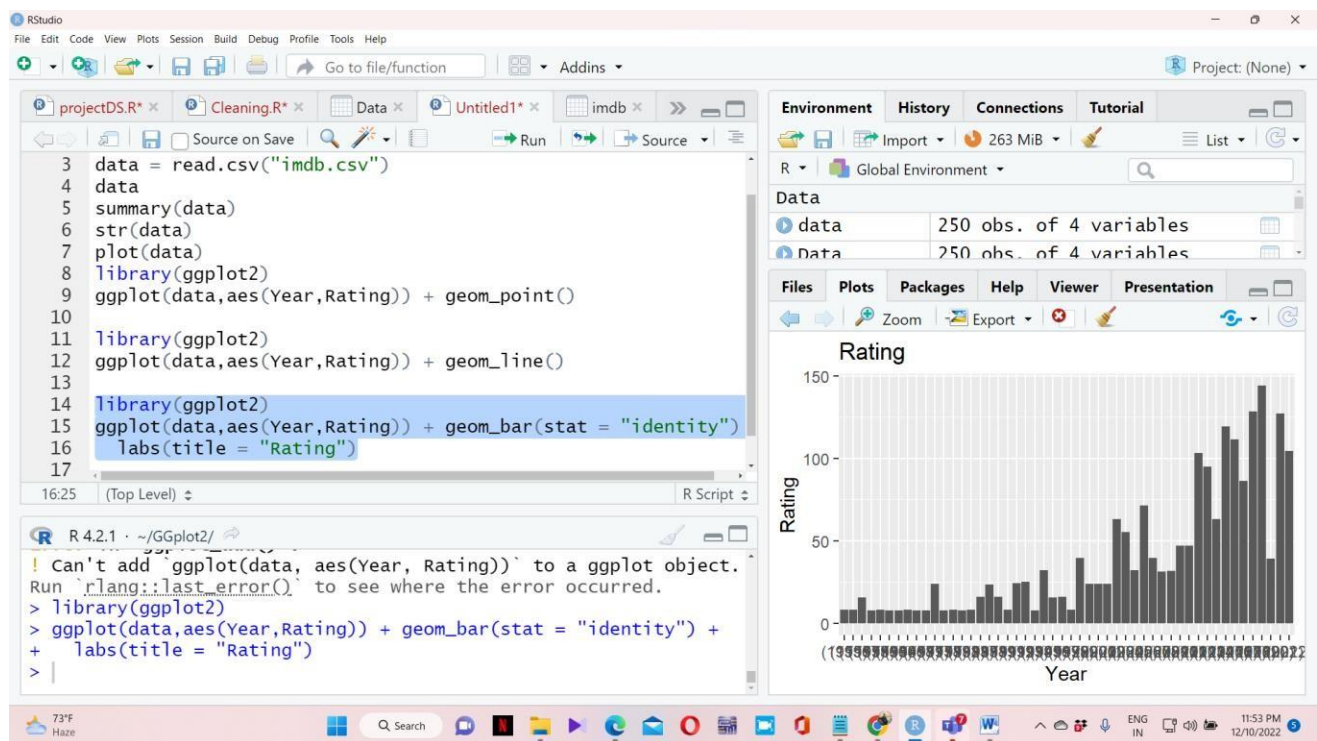
```
ggplot(data,aes(Year,Rating)) +
```

```
geom_line()
```



```
library(ggplot2)
```

```
ggplot(data,aes(Year,Rating)) + geom_bar(stat =  
"identity") + labs(title = "Rating")
```



## Discussion and Conclusion:

In order to offer the necessary dataset, pre-processing techniques are used, and all missing data and outliers are found. without data pre-processing we can not get a perfect analysis result. We need to do web scraping and then we can do pre- processing , Describing statistics and Data visualization.

