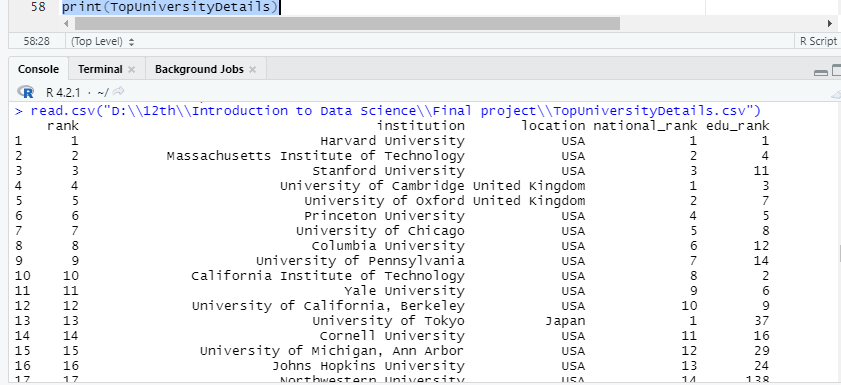
**Project Title: Web Scraping**

**Project Overview:**

The introduction of the project started with web scraping where scarping data on a specific website was performed by rvest library. The idea was to do a realistic project which will help us to know and to choose the right path for our career. Many of us hold the vision to accomplish their further degrees abroad and so the idea of the whole top ranked university details appeared to us. Most of the websites were unavailable for scraping data but our patience and dedication finally took us to the website which goes by the name of cwur.org.  
  
After scraping, generated the data in data frame then took in the .xls format and afterwards we converted the .xls into .CSV in order to load it in the RStudio.  
  
  
  
  
  
  
  
  
  
To prepare clean dataset we will perform the principles of data preprocessing:  
  
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:**

**Data Collection via Web Scraping:**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 “cwur.org” and scraped data from this website. The data set consists of 2000 observations and 9 variables(Rank, Institution, Location, national\_rank, edu\_rank, emp\_rank, faculty\_rank, research\_rank, score).  
  
library(rvest)

uniDetails <- read\_html("https://cwur.org/2022-23.php")

rank <- html\_text(html\_nodes(uniDetails,"td:nth-child(1)"))

rank

institution <- html\_text(html\_nodes(uniDetails,"#cwurTable a:nth-child(1)"))

institution

location <- html\_text(html\_nodes(uniDetails,"td:nth-child(3)"))

location

national\_rank <- html\_text(html\_nodes(uniDetails,"td:nth-child(4)"))

national\_rank

edu\_rank<- html\_text(html\_nodes(uniDetails,"td:nth-child(5)"))

edu\_rank

emp\_rank <- html\_text(html\_nodes(uniDetails,"td:nth-child(6)"))

emp\_rank

faculty\_rank <- html\_text(html\_nodes(uniDetails,"td:nth-child(7)"))

faculty\_rank

research\_rank <- html\_text(html\_nodes(uniDetails,"td:nth-child(8)"))

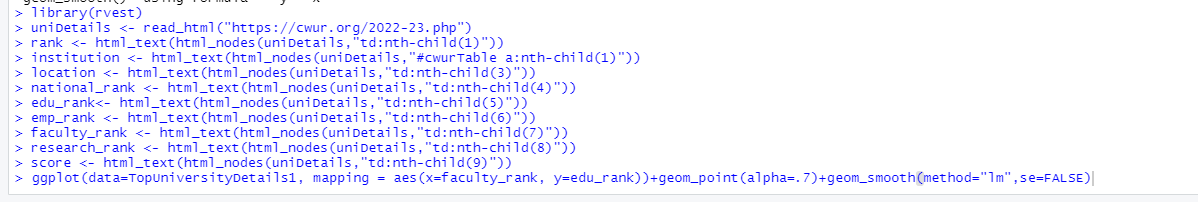
research\_rank

score <- html\_text(html\_nodes(uniDetails,"td:nth-child(9)"))

score

#Creating Dataframe

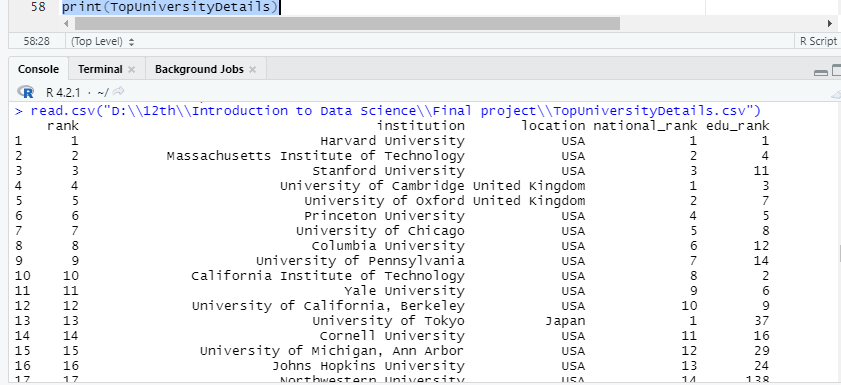
TopUniversityDetails <- data.frame(rank,institution,location, national\_rank, edu\_rank, emp\_rank, faculty\_rank, research\_rank, score)

TopUniversityDetails  
  


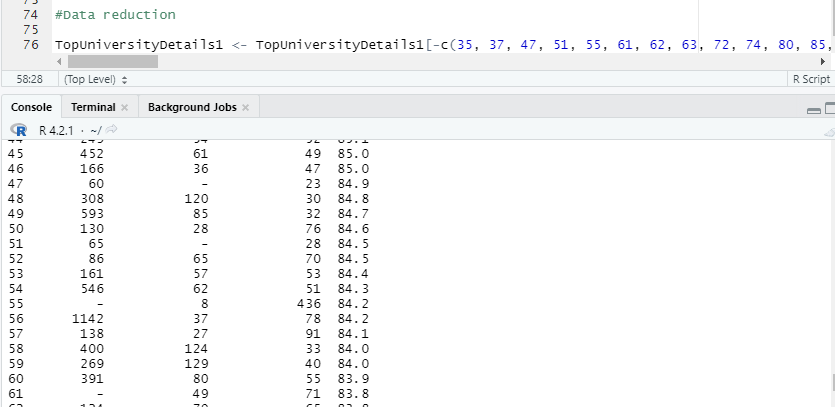
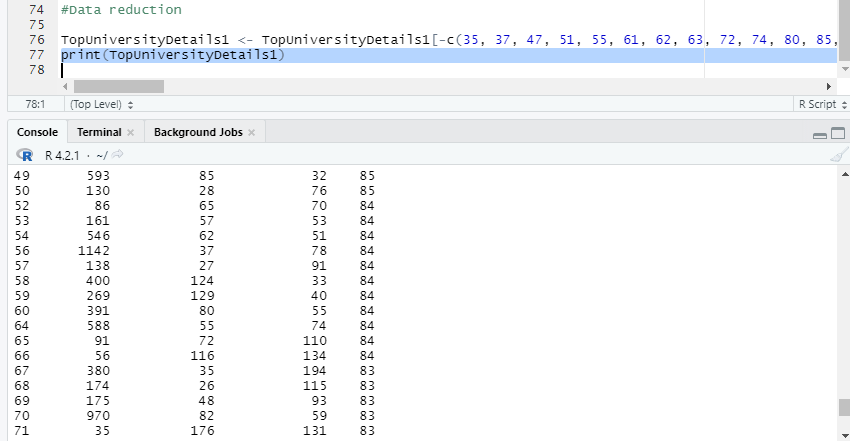
Export the dataframe

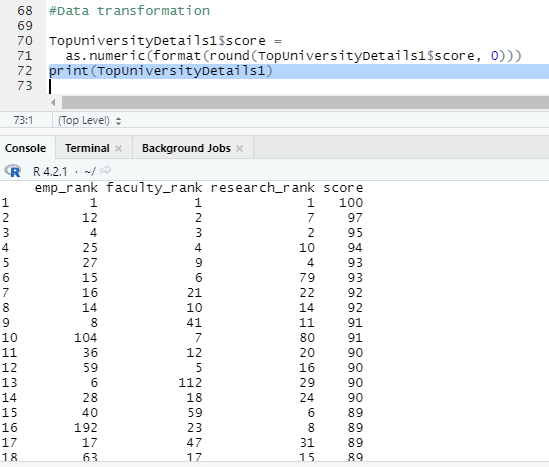
TopUniversityDetails1 <- read.csv("D:\\12th\\Introduction to Data Science\\Final project\\TopUniversityDetails.csv")

print(TopUniversityDetails1)



To prepare a cleaned dataset, we have to perform the following tasks of data pre-processing using R language-

**Data pre-processing:**Now we will be performing the principles of data preprocessing which include   
  
Data Reduction: Reducing missing values   
  
  
   
After:  
  


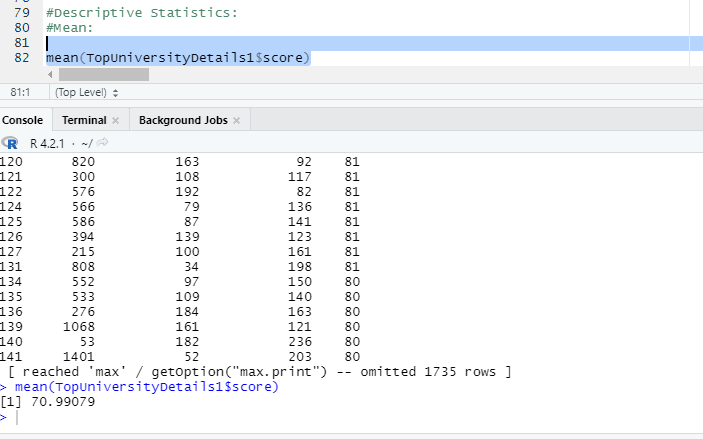
Converted the data type into numeric and rounded up the score column.  
  


**Descriptive Statistics:**

**1.Mean:**

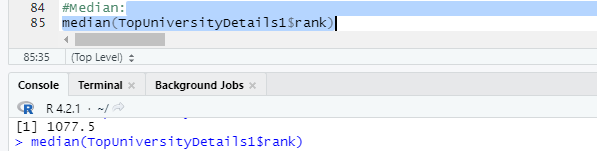
In descriptive Statistics first we did mean of TopUniversityDetails1 dataset. In the data set we calculated the mean of score column.

mean(TopUniversityDetails1$score)



**2. Median:**

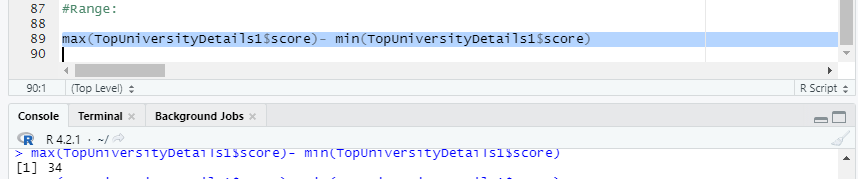
In descriptive Statistics secondly, we did median of TopUniversityDetails1 dataset. In the data set we calculated the median of Ratings column.

median(TopUniversityDetails1$rank)  
  


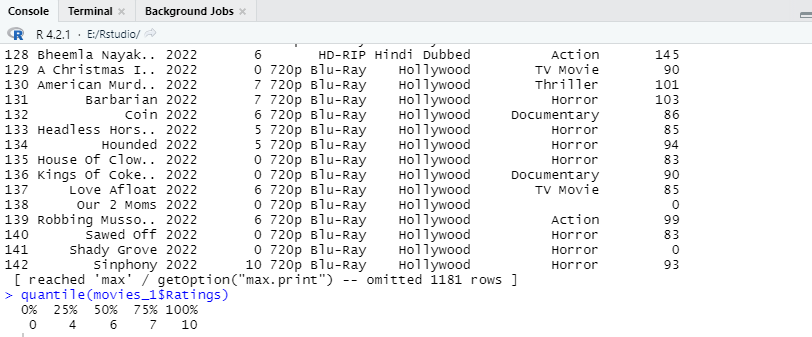
**3. Range:**

After that, we calculated the range of score of the top notch universities abroad. In the dataset we calculated the range of score column.

max(TopUniversityDetails1$score)- min(TopUniversityDetails1$score)

****

**4. Quantile:**

And lastly, we determined the quantile range of the dataset based on the column Downloads and Ratings from TopUniversityDetails1 data frame.quantile(TopUniversityDetails1$education\_rank)

**Discussion and Conclusion:**

After doing date pre-processing operations in dataset we scraped from cwur website, we can perform these steps in any datasets when we need. And data pre-processing helps. We can also get an easy-to-understand dataset after doing these operations. We can also analyze 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.