American International University – Bangladesh
Department of Computer Science & Engineering

Project Title: Data Pre-processing and Analysis on Worldwide University Rank
Course: Introduction to Data Science
Section: B

Submitted by-

Name	ID
Pritom Debnath	20-42414-1
Saad Nasir	20-42897-1
Md. Monjurul Islam Arman	19-39482-1
Somiya Akter Nehat	19-39615-1

Submitted to-

Dr. Akinul Islam Jony

PROJECT OVERVIEW:

To create a dataset on which to execute data processing, data visualization, and descriptive statistics, we had to scrape data from a selected website. For this case, we selected the website "cwur.org" and scraped information from it for the web scraping portion. There are 6 variables and 1995 observations in the data set.

The following data pre-processing operations must be carried out using the R programming language in order to produce a cleaned dataset. Cleaning of data, integration and transformation of Data etc. is used for the pre-processing portion.

This phase will result in a process dataset that is prepared for usage. Based on the data descriptive analysis is done. Finally, A reactive web page is made using shiny app.

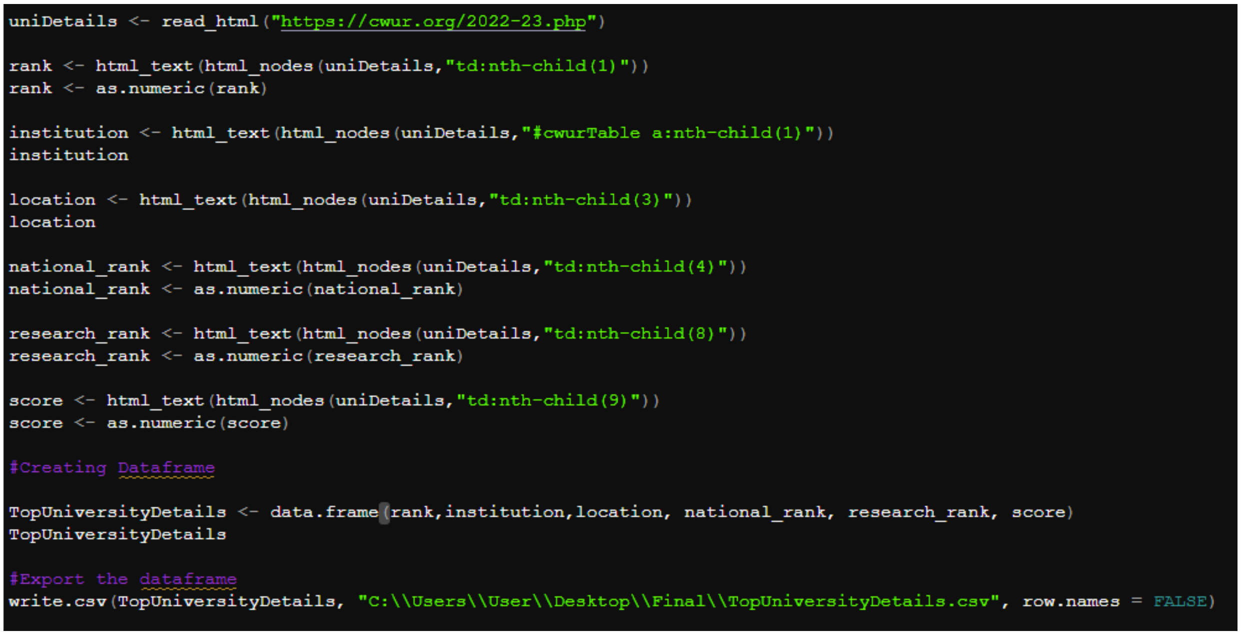
PROJECT SOLUTION DESIGN:

Dataset is generated using web scraping. The dataset contains information about the ranking of universities based on different criteria. The dataset first be created as a CSV file. The dataset must then be imported into RStudio in order for the R language to be used for data processing. The report focuses on preparing the dataset for data analysis by applying various pre-processing techniques such as data cleaning, integration, transformation and discretization using R language. The report describes the data by applying descriptive statistics and visualizing the data through different graphs. Finally, the dataset and different types of plots are used to make a reactive dashboard.

DATA COLLECTION VIA WEB SCRAPING:

To create a dataset on which to execute data processing, data visualization, and descriptive statistics, we had to scrape data from a selected website.

We selected "cwur.org" as the website to scrape data from for the web scraping portion. 1995 observations make up the data set, which also includes 6 variables (Rank, Institution, Location, National Rank, Research Rank, and Score). After creating the data frame, it is stored in the local machine.



```
uniDetails <- read_html("https://cwur.org/2022-23.php")

rank <- html_text(html_nodes(uniDetails,"td:nth-child(1)"))
rank <- as.numeric(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 <- as.numeric(national_rank)

research_rank <- html_text(html_nodes(uniDetails,"td:nth-child(8)"))
research_rank <- as.numeric(research_rank)

score <- html_text(html_nodes(uniDetails,"td:nth-child(9)"))
score <- as.numeric(score)

#Creating Dataframe

TopUniversityDetails <- data.frame(rank,institution,location, national_rank, research_rank, score)
TopUniversityDetails

#Export the dataframe
write.csv(TopUniversityDetails, "C:\\Users\\User\\Desktop\\Final\\TopUniversityDetails.csv", row.names = FALSE)
```

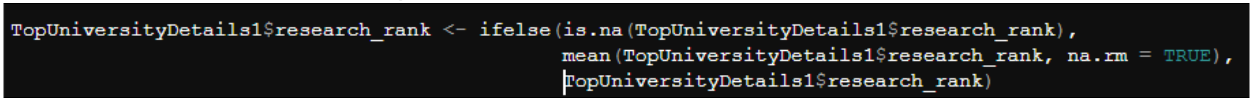
Figure-1: Web scraping, data frame creation and saving the dataset

DATA PRE-PROCESSING:

1. Data cleaning:

- *Dealing with Missing Values:*

  According to the dataset, there is a missing value (NA) in the research_rank column. We can see how many missing values are there by using this: *sum(is.na(TopUniversityDetails1$research_rank)).* We can replace the missing value with the column's mean value.



```
TopUniversityDetails1$research_rank <- ifelse(is.na(TopUniversityDetails1$research_rank),
                                              mean(TopUniversityDetails1$research_rank, na.rm = TRUE),
                                              TopUniversityDetails1$research_rank)
```

Figure-2: Removing the null values

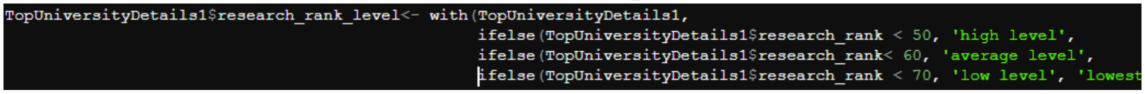
- *Data formatting:*

  After replacing the null value data should be formatted correctly. For this job we have used this code: *TopUniversityDetails1$research_rank <- as.numeric(format(round(TopUniversityDetails1$research_rank, 0)))*

2. Data transformation and reduction:

This dataset does not need any data transformation and reduction steps.

3. **Data Discretization and Data Integration:**
   We frequently work with data that is gathered through continuous procedures. But there are situations when it's necessary to split up these continuous numbers into smaller chunks. Discrete mapping is the term for this process. Using logic, we may discretize the data into category kinds and include the column into our dataset. R code for this step-

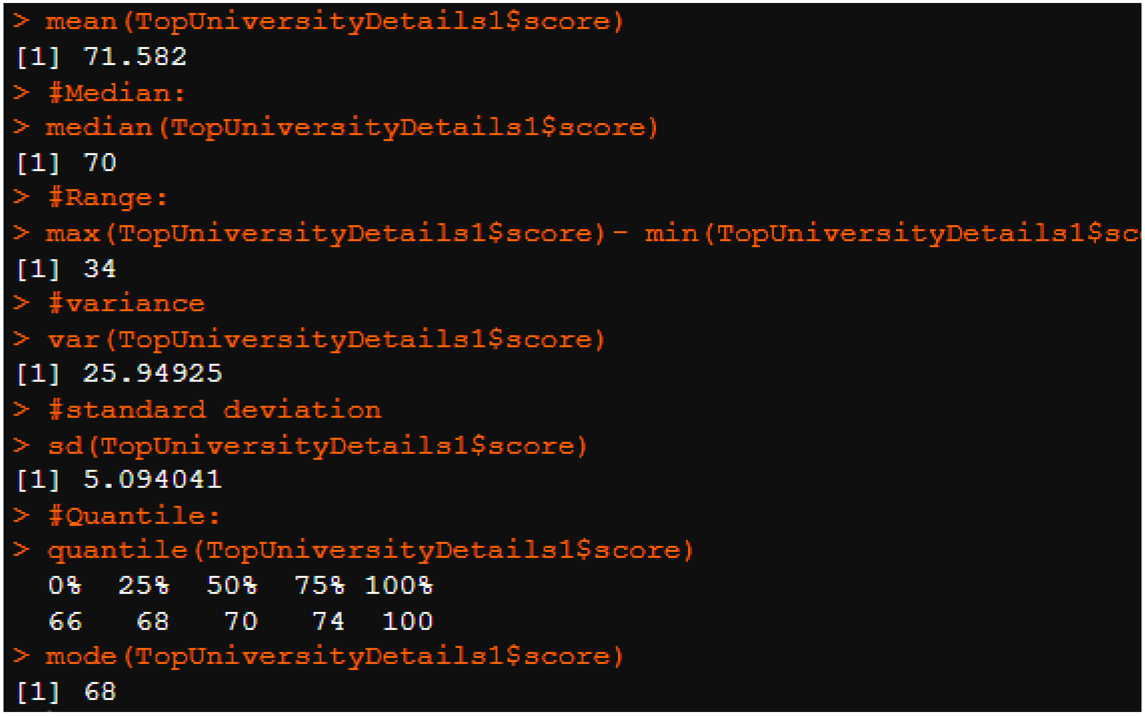


```
TopUniversityDetails1$research_rank_level<- with(TopUniversityDetails1,
                                                 ifelse(TopUniversityDetails1$research_rank < 50, 'high level',
                                                 ifelse(TopUniversityDetails1$research_rank< 60, 'average level',
                                                 ifelse(TopUniversityDetails1$research_rank < 70, 'low level', 'lowest
```

Figure-3: Adding a new column

DESCRIPTIVE STATISTICS:

Descriptive Statistics is the process of summarizing and describing the dataset using various statistical measures such as mean, median, mode, standard deviation, and range. The descriptive statistics of the dataset are as follows:



```
> mean(TopUniversityDetails1$score)
[1] 71.582
> #Median:
> median(TopUniversityDetails1$score)
[1] 70
> #Range:
> max(TopUniversityDetails1$score)- min(TopUniversityDetails1$sc
[1] 34
> #variance
> var(TopUniversityDetails1$score)
[1] 25.94925
> #standard deviation
> sd(TopUniversityDetails1$score)
[1] 5.094041
> #Quantile:
> quantile(TopUniversityDetails1$score)
  0%  25%  50%  75% 100% 
  66   68   70   74  100 
> mode(TopUniversityDetails1$score)
[1] 68
```

Figure-4: code of Descriptive Statistics

Mean: 71.582, Median: 70, Mode: 68, Standard Deviation: 5.09, Range: 34

From the above descriptive statistics, we can observe that the mean rank of the universities is 71.582, with a standard deviation of 5.09, smaller standard deviation indicates that the values are closer to the mean. Median 70 is the most frequently occurring value in the data set. The minimum rank is 66, and the maximum rank is 100, with a range of 34 it is the difference between the largest value and the smallest value in the data set. It provides an indication of how much variation there is in the data.

DATA VISUALIZATION:

From the data two types of graphs are plotted. Then the data are shown in a web-based application.

Bar Graph: The bar graph shows the national rank of the universities. It is evident from the bar graph that the majority of the top-ranked universities are from the USA, followed by the China, United Kingdom and Canada. R code and the diagram as follows: -

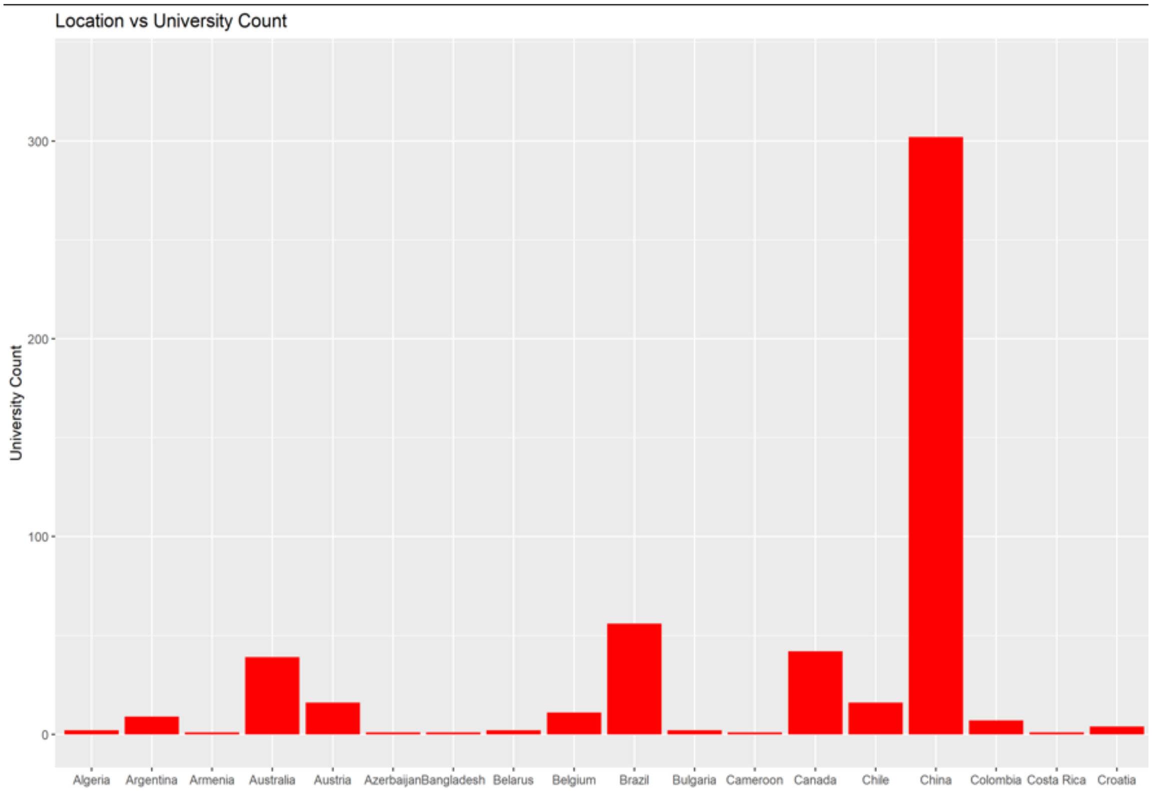
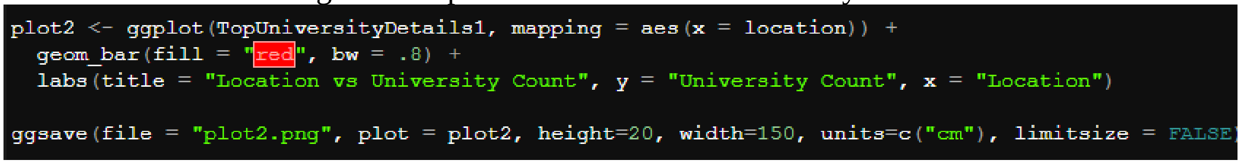


Figure-5: A portion of location vs university count

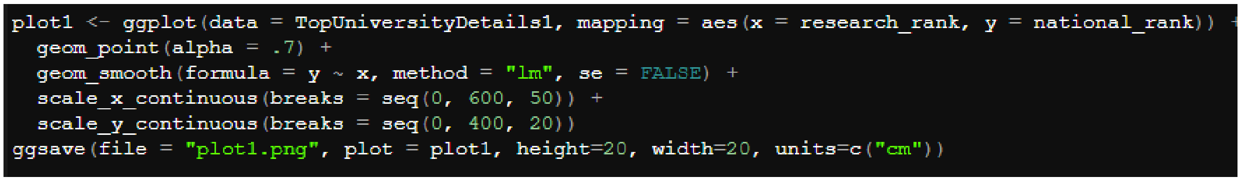


```
plot2 <- ggplot(TopUniversityDetails1, mapping = aes(x = location)) +
  geom_bar(fill = "red", bw = .8) +
  labs(title = "Location vs University Count", y = "University Count", x = "Location")

ggsave(file = "plot2.png", plot = plot2, height=20, width=150, units=c("cm"), limitsize = FALSE)
```

Figure-6: Code of the location vs university count graph

Scatter Plot: The scatter plot shows the relationship between research rank and national rank. It is evident from the scatter plot that there is a positive correlation between research national rank and research rank, which means that universities with higher research rank also have a higher national rank. R code and the diagram as follows: -



```
plot1 <- ggplot(data = TopUniversityDetails1, mapping = aes(x = research_rank, y = national_rank)) +
  geom_point(alpha = .7) +
  geom_smooth(formula = y ~ x, method = "lm", se = FALSE) +
  scale_x_continuous(breaks = seq(0, 600, 50)) +
  scale_y_continuous(breaks = seq(0, 400, 20))
ggsave(file = "plot1.png", plot = plot1, height=20, width=20, units=c("cm"))
```

Figure-7: code of the national rank vs research rank graph

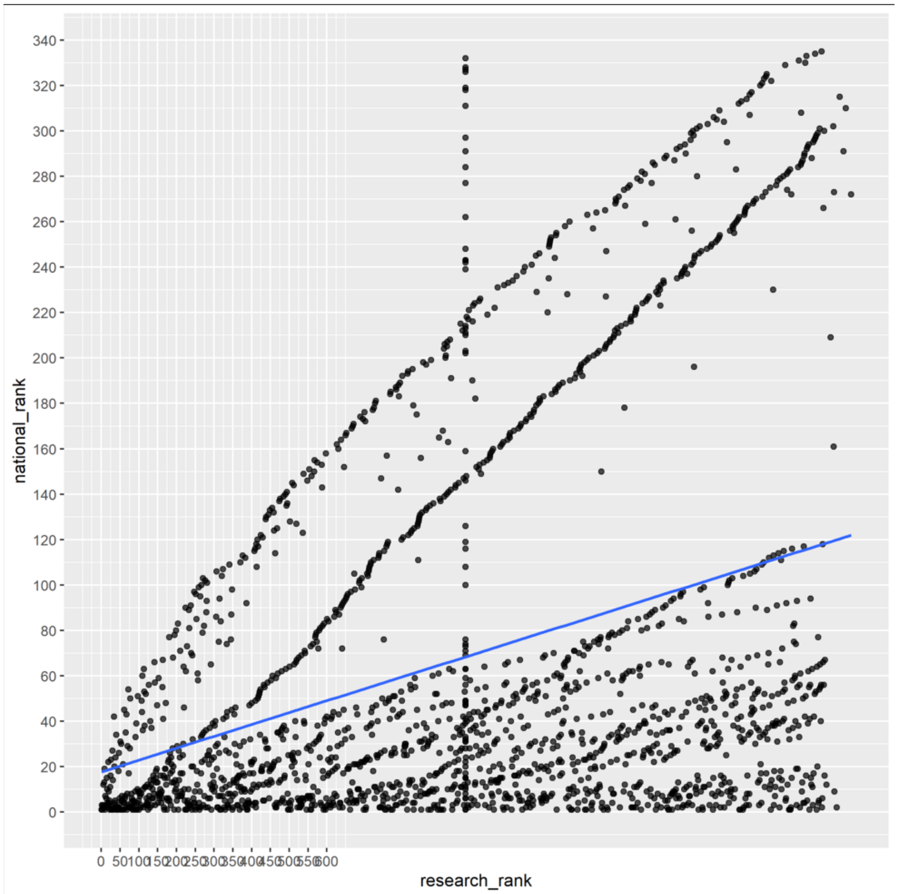
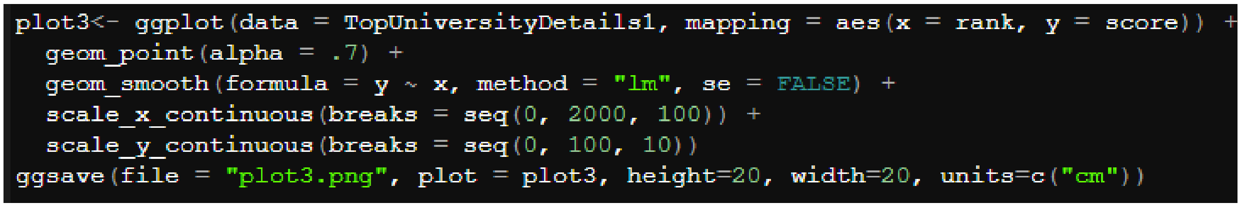


Figure-8: Graph of national rank vs research rank

Another scatter plot shows the relationship between rank and score. It is evident from the scatter plot that there is a negative correlation between rank and score, which means that universities with higher score have a lower rank. R code and the diagram as follows: -



```
plot3<- ggplot(data = TopUniversityDetails1, mapping = aes(x = rank, y = score)) +
  geom_point(alpha = .7) +
  geom_smooth(formula = y ~ x, method = "lm", se = FALSE) +
  scale_x_continuous(breaks = seq(0, 2000, 100)) +
  scale_y_continuous(breaks = seq(0, 100, 10))
ggsave(file = "plot3.png", plot = plot3, height=20, width=20, units=c("cm"))
```

Figure-9: Code of the rank vs score graph

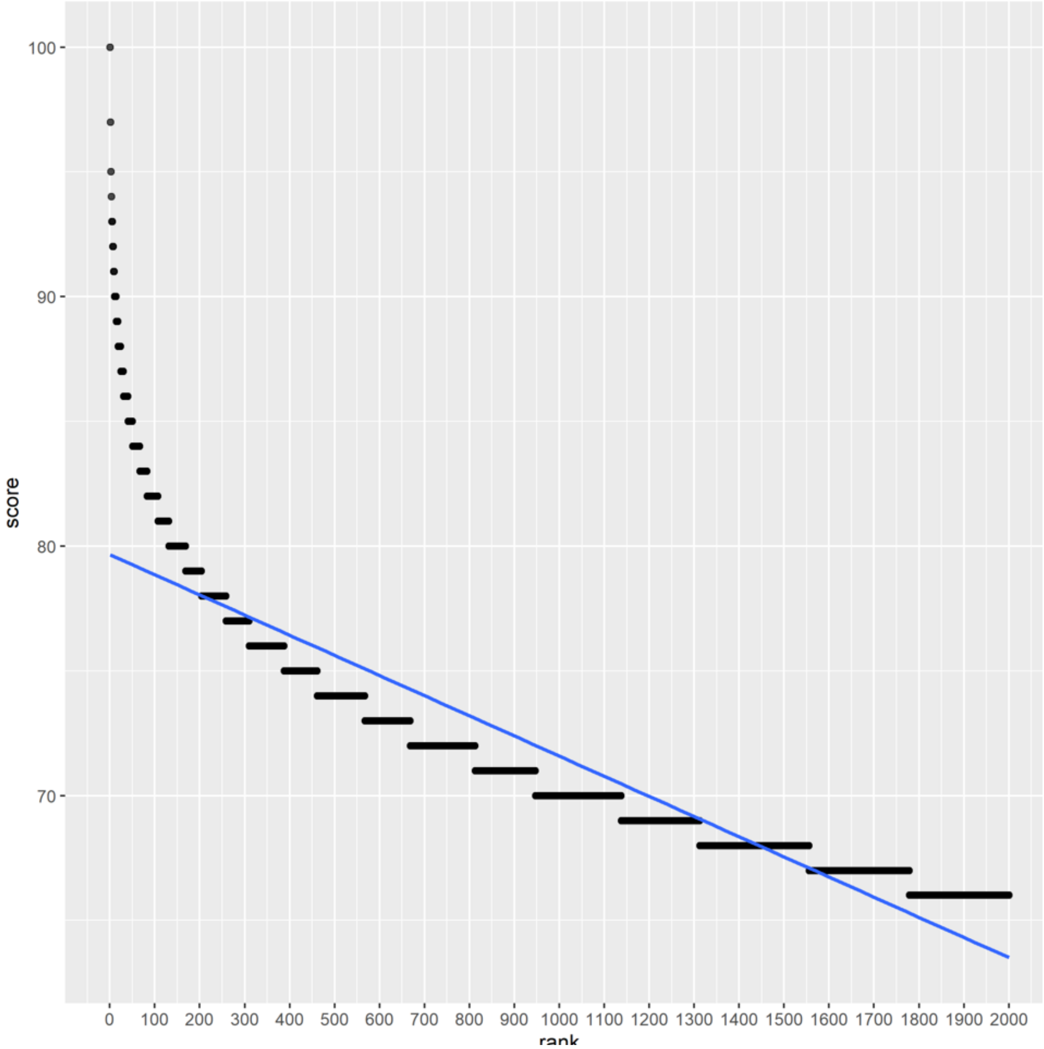
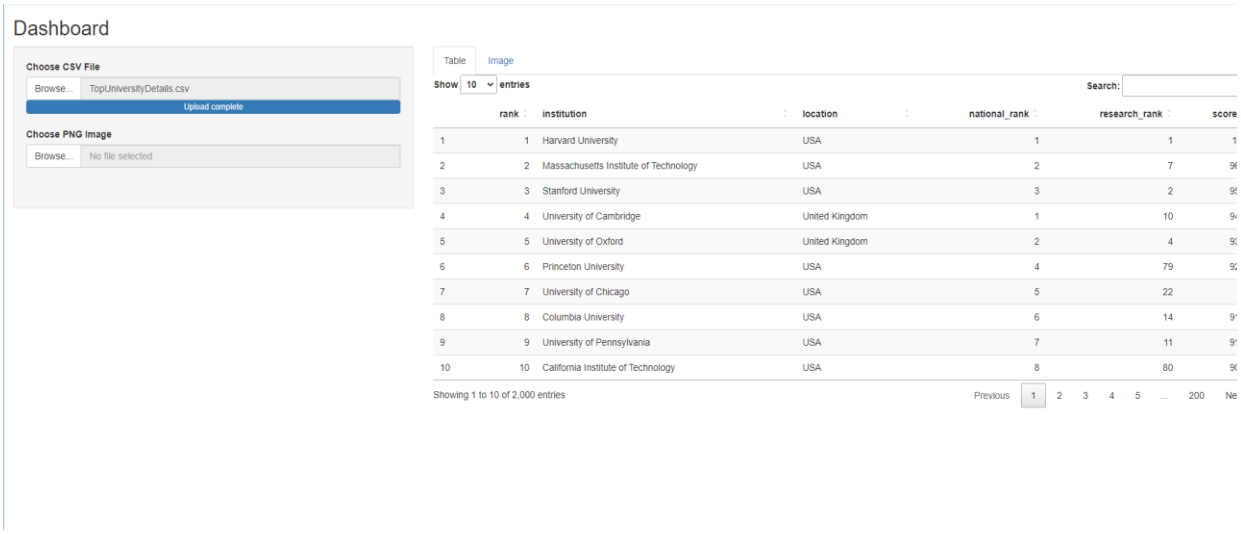


Figure-10: Graph of rank vs score

DISCUSSION AND CONCLUSION:

In this report, we have applied various pre-processing techniques such as data cleaning, integration and discretization to prepare the dataset for analysis. We have also described the data by applying descriptive statistics and visualized the data through different graphs. The analysis shows that the majority of the top-ranked universities are from the USA, followed by the China, United Kingdom and Canada. Also, there is a positive correlation between research rank and score, which means that universities with higher research rank also have a higher score. And there is a negative correlation between rank and score, which means that universities with higher score have a lower rank.

Picture of the dashboard and code is here with,



Dashboard

Choose CSV File

Browse... TopUniversityDetails.csv

Upload complete

Choose PNG Image

Browse... No file selected

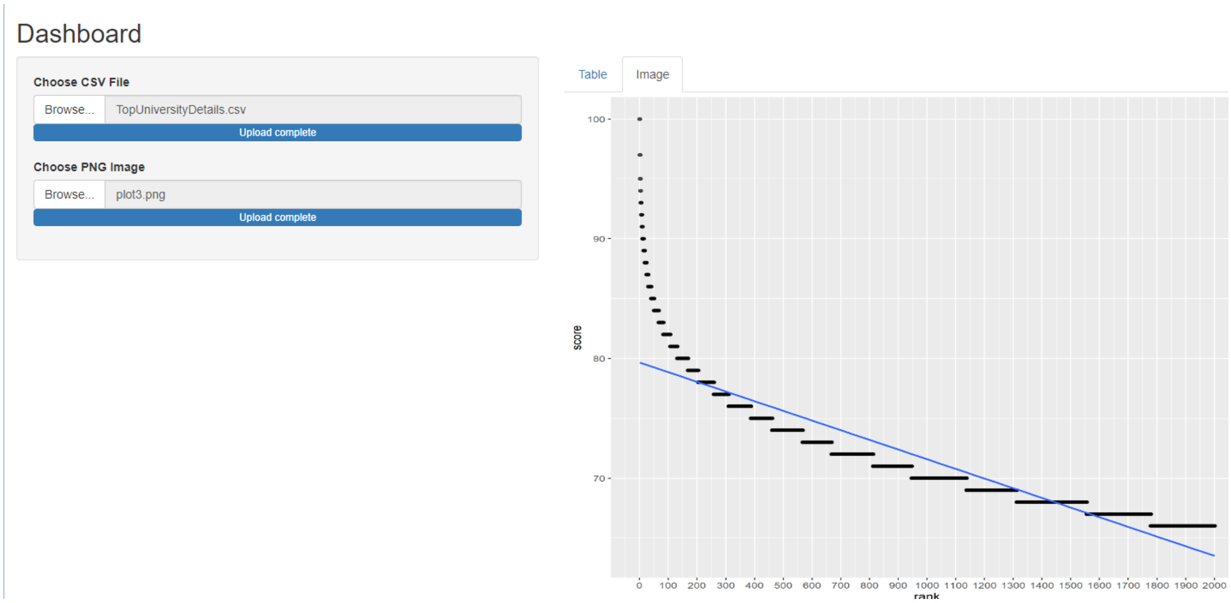
Show 10 entries

Search:

	rank	institution	location	national_rank	research_rank	score
1	1	Harvard University	USA	1	1	1
2	2	Massachusetts Institute of Technology	USA	2	7	96
3	3	Stanford University	USA	3	2	95
4	4	University of Cambridge	United Kingdom	1	10	94
5	5	University of Oxford	United Kingdom	2	4	93
6	6	Princeton University	USA	4	79	92
7	7	University of Chicago	USA	5	22	
8	8	Columbia University	USA	6	14	91
9	9	University of Pennsylvania	USA	7	11	91
10	10	California Institute of Technology	USA	8	80	90

Showing 1 to 10 of 2,000 entries

Figure-11: Dashboard of the dataset.



Dashboard

Choose CSV File

Browse... TopUniversityDetails.csv

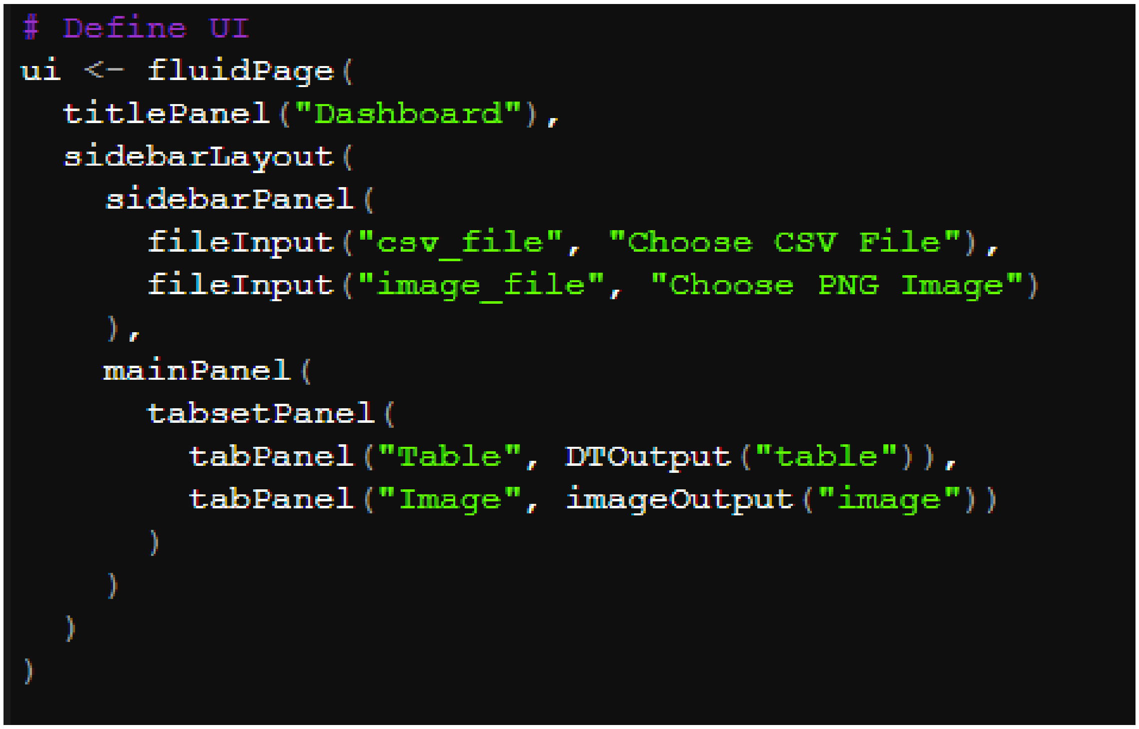
Upload complete

Choose PNG Image

Browse... plot3.png

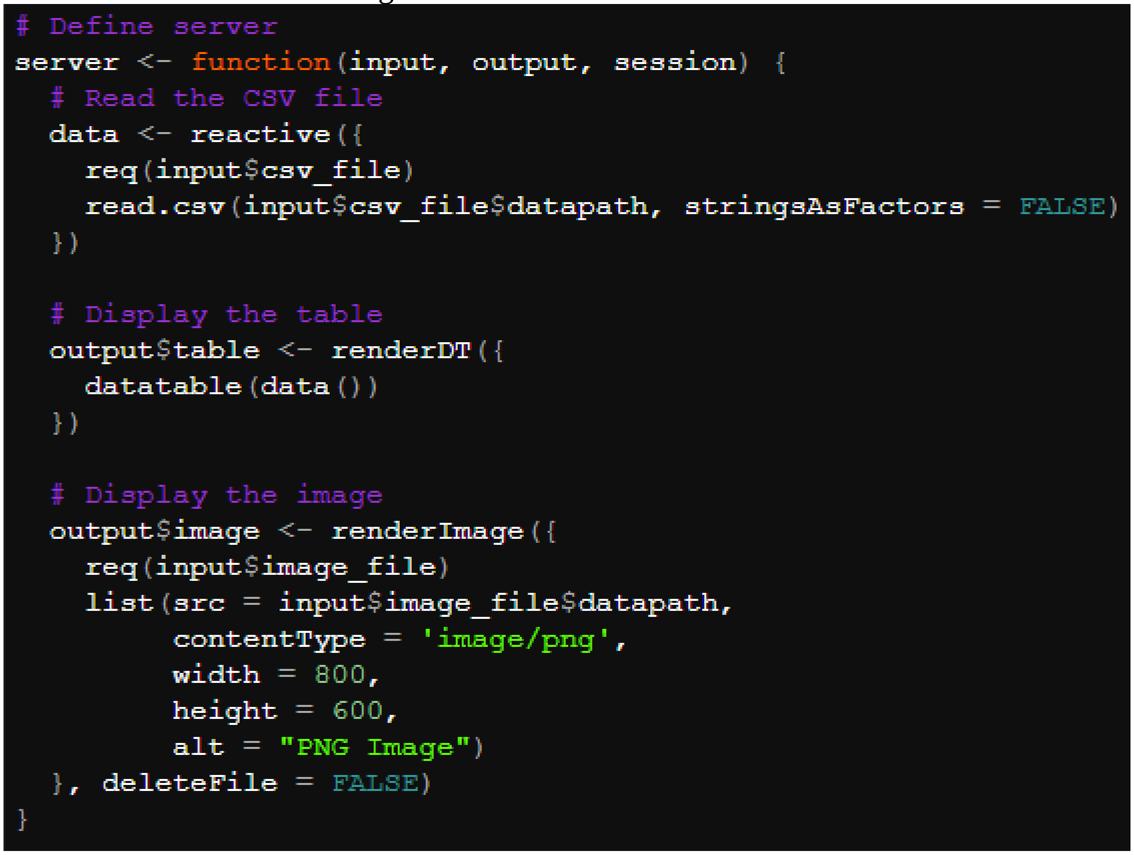
Upload complete

Figure-12: Dashboard of the graph.



```
# Define UI
ui <- fluidPage(
  titlePanel("Dashboard"),
  sidebarLayout(
    sidebarPanel(
      fileInput("csv_file", "Choose CSV File"),
      fileInput("image_file", "Choose PNG Image")
    ),
    mainPanel(
      tabsetPanel(
        tabPanel("Table", DTOutput("table")),
        tabPanel("Image", imageOutput("image"))
      )
    )
  )
)
```

Figure-13: Code of User interface



```
# Define server
server <- function(input, output, session) {
  # Read the CSV file
  data <- reactive({
    req(input$csv_file)
    read.csv(input$csv_file$datapath, stringsAsFactors = FALSE)
  })

  # Display the table
  output$table <- renderDT({
    datatable(data())
  })

  # Display the image
  output$image <- renderImage({
    req(input$image_file)
    list(src = input$image_file$datapath,
         contentType = 'image/png',
         width = 800,
         height = 600,
         alt = "PNG Image")
  }, deleteFile = FALSE)
}
```

Figure-13: Code for the backend