INVESTIGATE WORLD UNIVERSITY RANKING

Task :1 generate a bar chart of the distribution students on the country by using ggplot.
.

First what is “ggplot2” so it is a one kind of package which is used for data visualization and also used for semantic grammar of graphics.

so now we have to install ggplot2 package by below code

```
>install.packages("ggplot2")
```

```
>library(ggplot2)
```

now we have to load the data by below code

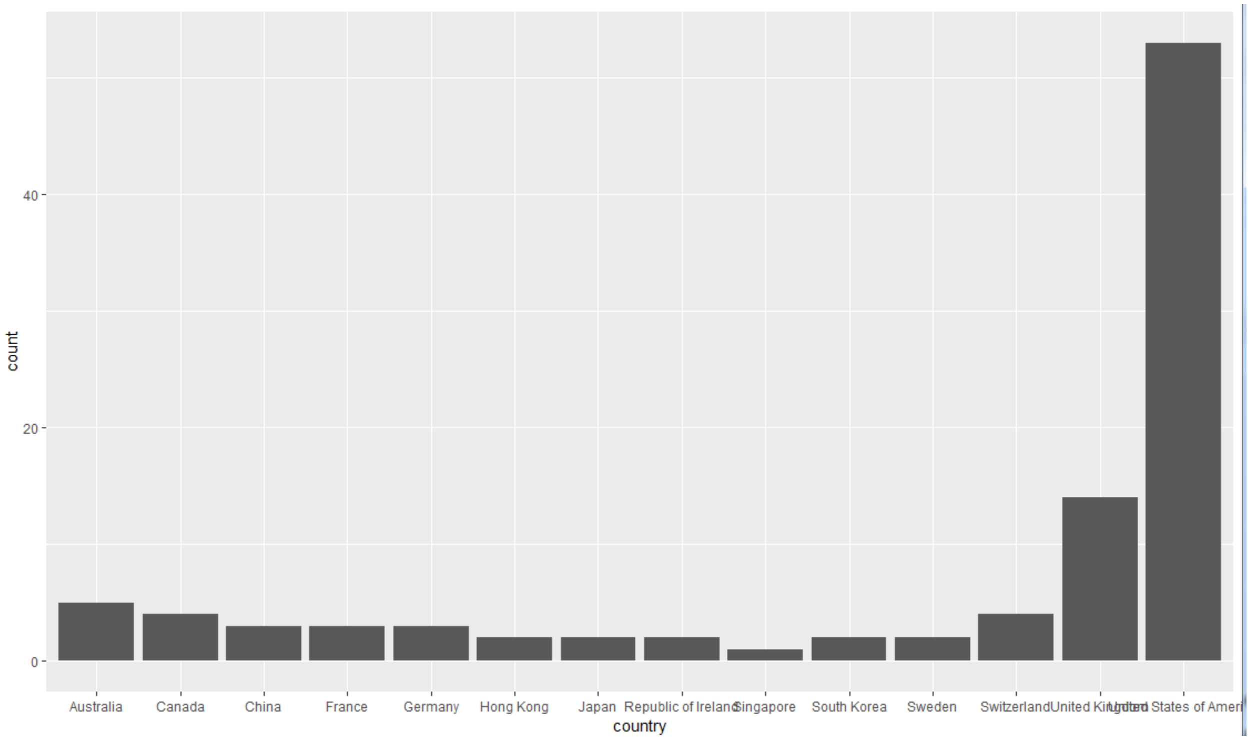
```
>mydata <- read.csv("D:/LAKEHEAD WINTER 2019/DATA SCI/project/timesData.csv")
```

```
>View(mydata)
```

Now bar chart of the distribution of students on the country .

```
>ggplot(mydata[1:100,]) + geom_bar(aes(country))
```

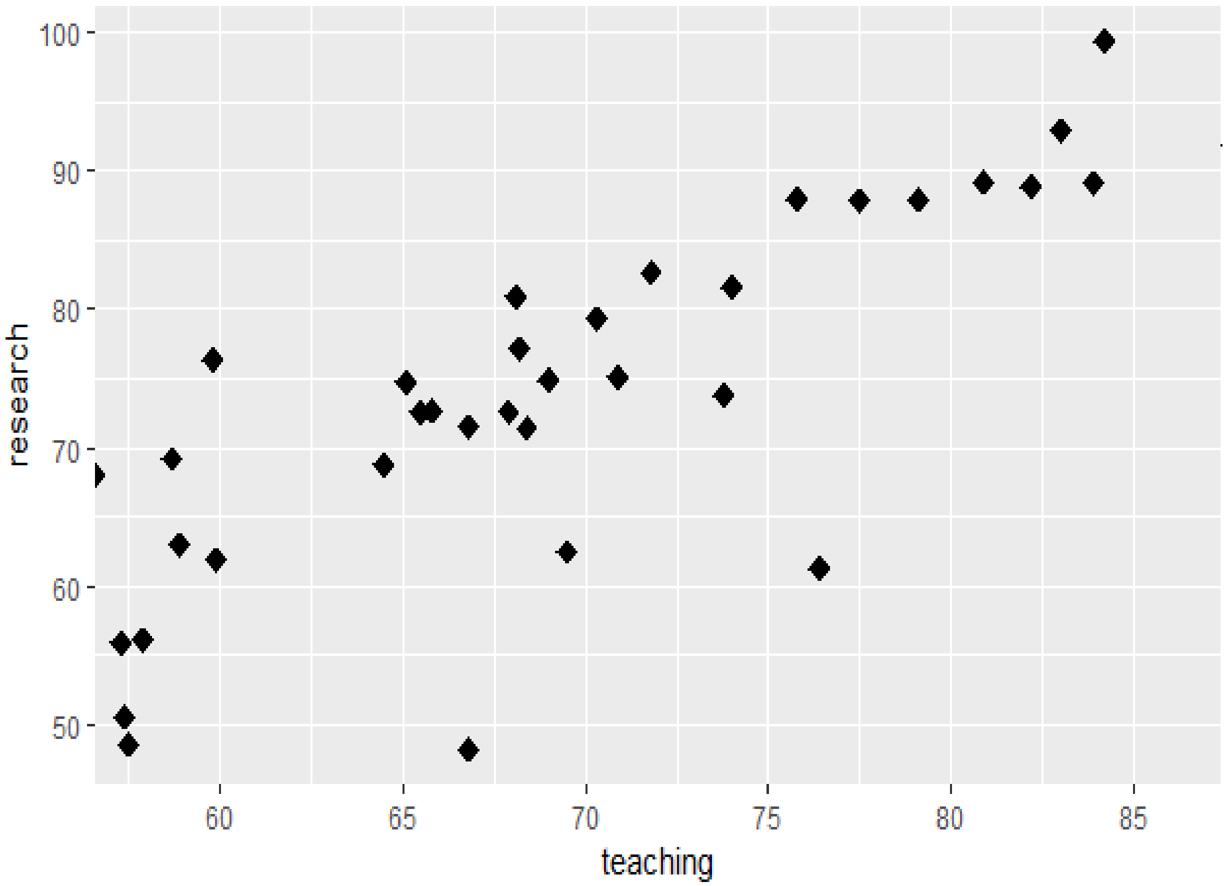
here we are using geom_bar() which is used to define geometric object for barchart.



Task 2: using scatter plot compare students's teaching and research ratio.

```
>ggplot(mydata[1:50,], aes(teaching, research)) + geom_point(size=3.5,shape=18) + coord_cartesian(xlim = c(58,86)) + scale_x_continuous(breaks = seq(60,85,5))
```

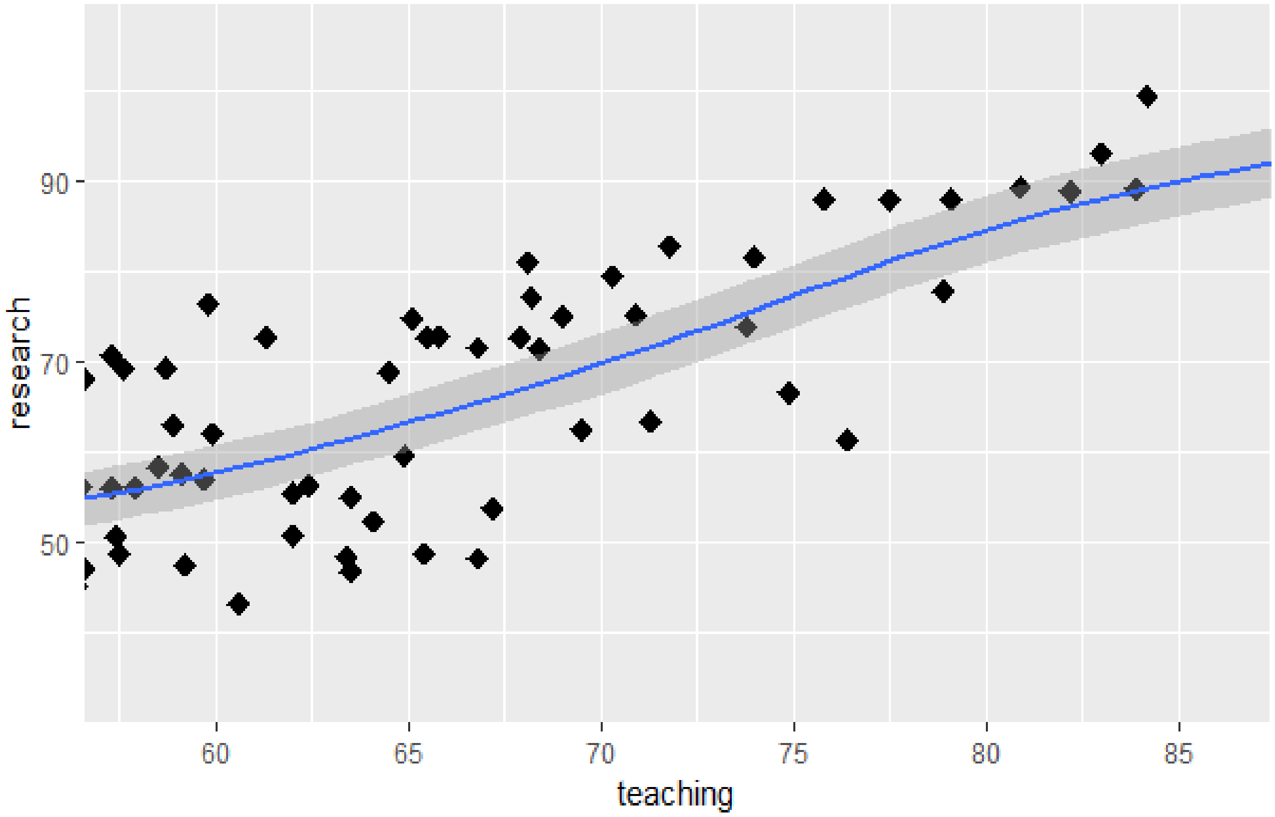
Here geom_point is define object of scatter plot.



Task: 3 modify the teaching and research ratio by including smooth regression.

```
> ggplot(mydata[1:100,], aes(teaching, research) ) + geom_point(size=3.5,shape=18) +
coord_cartesian(xlim = c(58,86)) + scale_x_continuous(breaks = seq(60,85,5)) +
geom_smooth()
```

Geom_smooth is used for adding extra line on scatter plot.

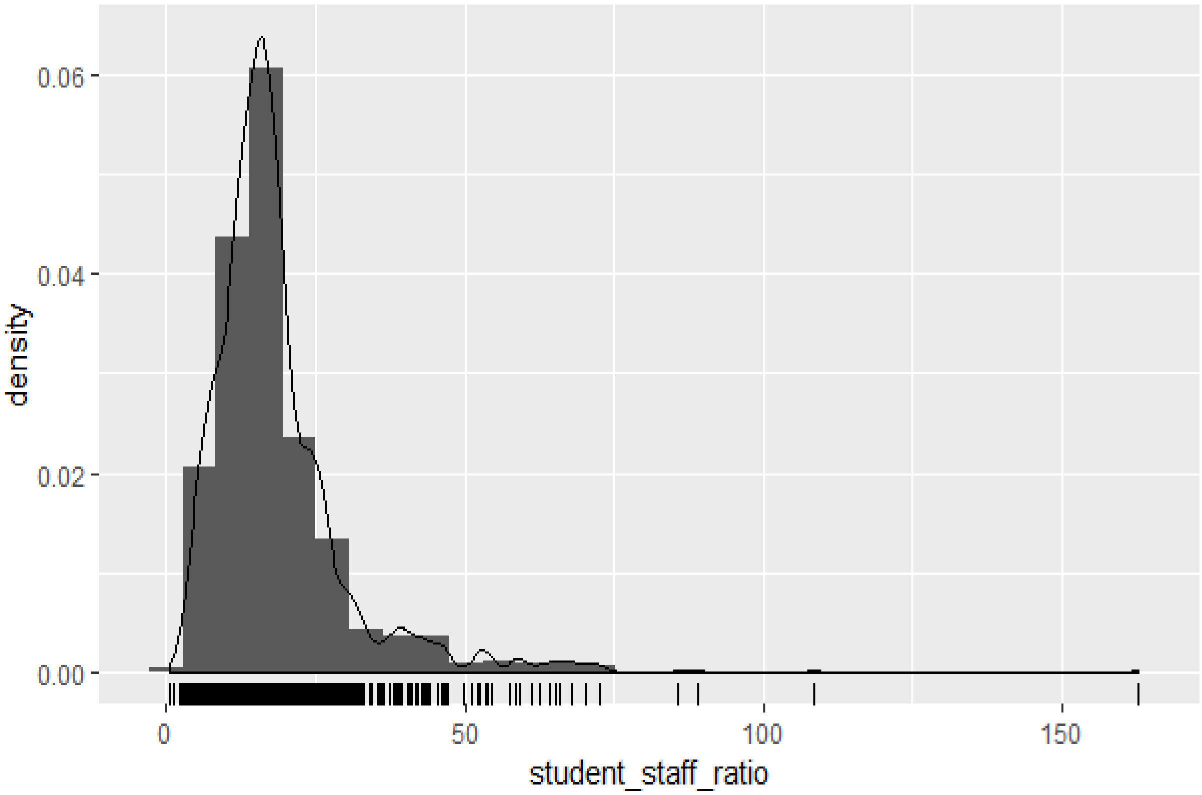


Task 4: visualize student_staff_ratio by using a histogram with density distribution function.

We are using geom_histogram for defining histrogram. Also we are using geom_density for define density curve as object.

Geom_rug can be used for tick marks on x-axes.

```
#Histogram with density distribution function

>mydata <- read.csv("D:/LAKEHEAD WINTER 2019/DATA SCI/project/timesData.csv")

>View(mydata)

>a<- ggplot(data=mydata, aes(student_staff_ratio)) + geom_histogram(aes(y=..density..)) +

  geom_density(aes(x=student_staff_ratio))  + geom_rug(aes(x=student_staff_ratio))

>a
```



Task: 5 find out how many universities have teaching ratio less than 50.

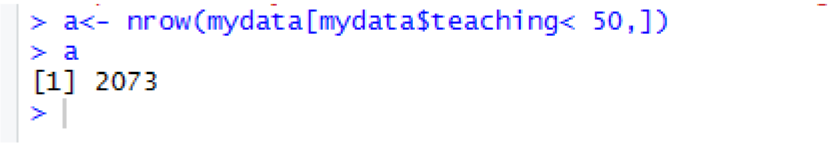
> **mydata <- read.csv("D:/LAKEHEAD WINTER 2019/DATA SCI/project/timesData.csv")**

>View(mydata)

#data has teching < 50

>a<- nrow(mydata[mydata$teaching< 50,])

>a



```
> a<- nrow(mydata[mydata$teaching< 50,])
> a
[1] 2073
> 
```

Task: 6 find out canadian institution names with world rank and national rank in year 2015.

> **install.packages("tidyverse")**

>library(tidyverse)

>x <- read.csv("D:/LAKEHEAD WINTER 2019/DATA SCI/project/cwurData.csv")

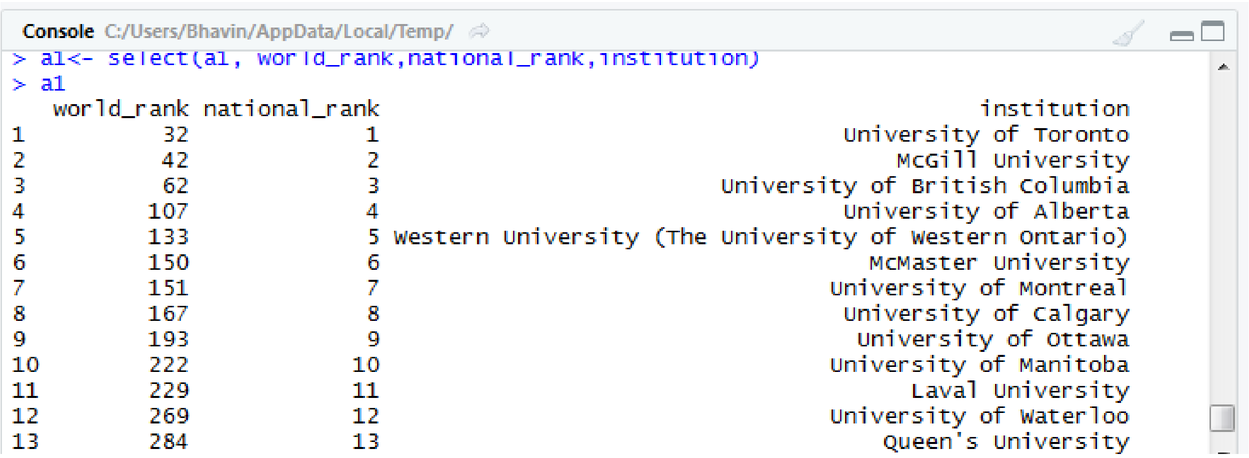
>x

>data_canada<- filter(x, country=="Canada")

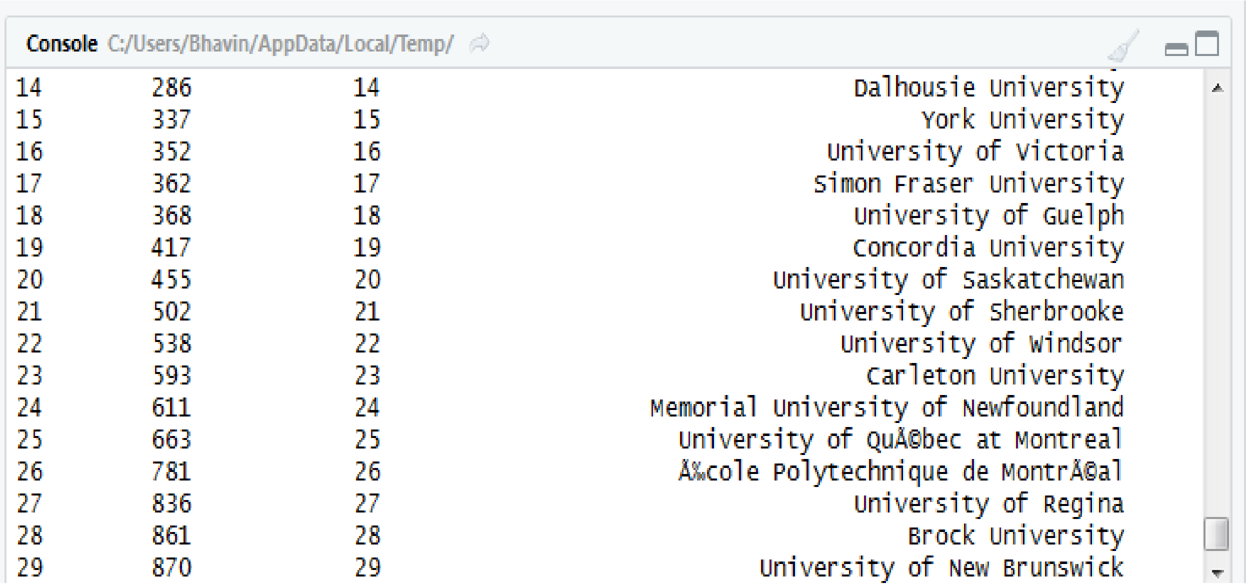
>a1<-filter(data_canada, year=="2015")

>a1<- select(a1, world_rank,national_rank,institution)

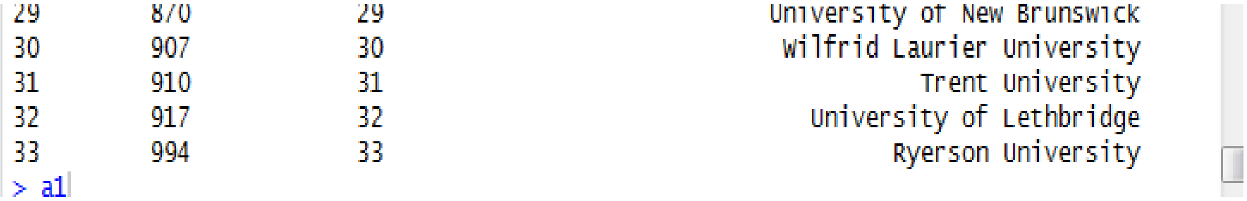
>a1



```
> a1<- select(a1, world_rank,national_rank,institution)
> a1
   world_rank national_rank                                             institution
1          32             1                                   University of Toronto
2          42             2                                       McGill University
3          62             3                         University of British Columbia
4         107             4                                   University of Alberta
5         133             5 Western University (The University of Western Ontario)
6         150             6                                     McMaster University
7         151             7                                  University of Montreal
8         167             8                                   University of Calgary
9         193             9                                    University of Ottawa
10        222            10                                  University of Manitoba
11        229            11                                        Laval University
12        269            12                                  University of Waterloo
13        284            13                                      Queen's University
```



```
14          286            14                                  Dalhousie University
15          337            15                                       York University
16          352            16                                University of Victoria
17          362            17                                Simon Fraser University
18          368            18                                  University of Guelph
19          417            19                                  Concordia University
20          455            20                          University of Saskatchewan
21          502            21                             University of Sherbrooke
22          538            22                                University of Windsor
23          593            23                                   Carleton University
24          611            24                Memorial University of Newfoundland
25          663            25                   University of QuÃ©bec at Montreal
26          781            26                   Ã‰cole Polytechnique de MontrÃ©al
27          836            27                                 University of Regina
28          861            28                                      Brock University
29          870            29                         University of New Brunswick
```

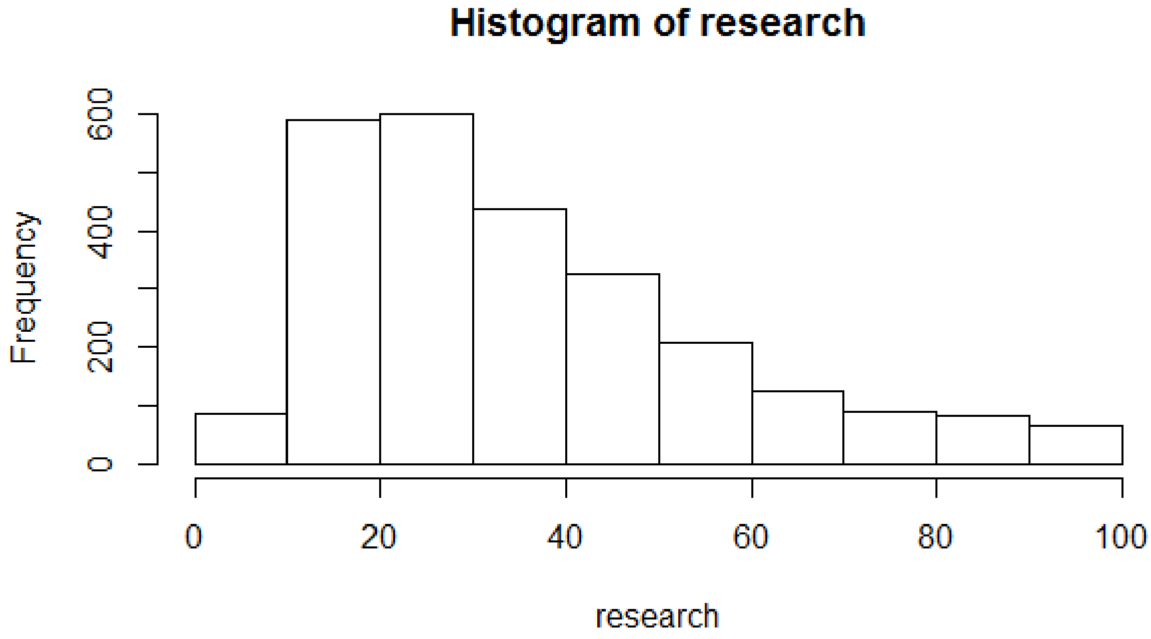


```
29          870            29                         University of New Brunswick
30          907            30                          Wilfrid Laurier University
31          910            31                                     Trent University
32          917            32                           University of Lethbridge
33          994            33                                    Ryerson University
> a1
```

Task 7 : Make histograms of the variable "research" using different numbers of bars.

```
>hist(mydata$research, breaks = 10, main = 'Histogram of research', xlab = "research")
```

Here x-axes will be research and it will break out at interval 10.



Task 8: Print out from times_teaching_score object all the observations which have teaching>88 and student_staff_ratio >10.

> install.packages("tidyverse")

>library(tidyverse)

#create the object times_teaching_tscore containing the column name income and students staff ratio

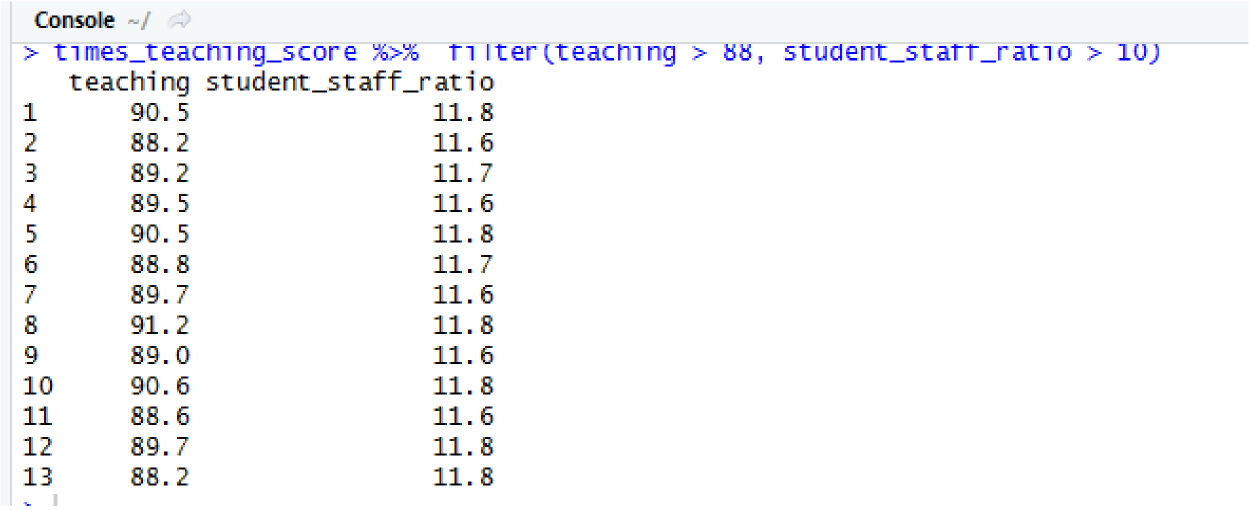
>times_teaching_score = mydata %>%

> select(teaching , student_staff_ratio)

>times_teaching_score

#print out from times_income_tscore object all the observation having teaching > 88 and student staff ratio> 10

>times_teaching_score %>% filter(teaching > 88, student_staff_ratio > 10)

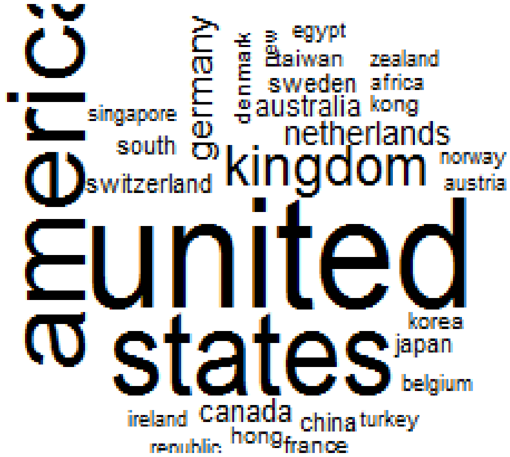


```
> times_teaching_score %>%  filter(teaching > 88, student_staff_ratio > 10)
   teaching student_staff_ratio
1      90.5                11.8
2      88.2                11.6
3      89.2                11.7
4      89.5                11.6
5      90.5                11.8
6      88.8                11.7
7      89.7                11.6
8      91.2                11.8
9      89.0                11.6
10     90.6                11.8
11     88.6                11.6
12     89.7                11.8
13     88.2                11.8
```

Task 9: Use the wordcloud library to visualize which country has mentioned most in 2011.

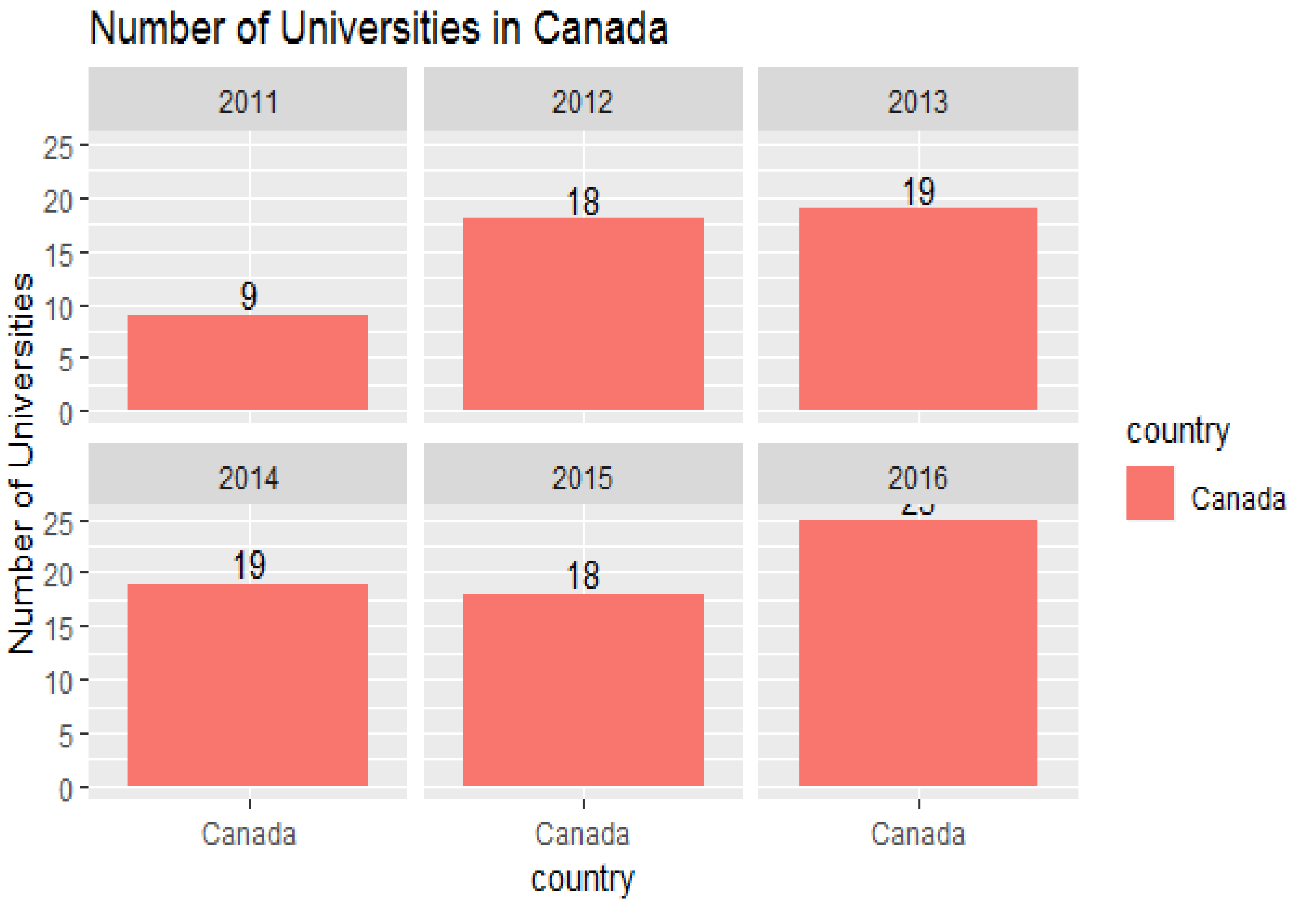
Here we are using two packages called “tm” ,“wordcloud”.with the help of wordcount words can visually show up in content information such as word scatterd around the figure. Words seeming all the more frequently in the text are appeared in a bigger text style, while less basic terms are appeared littler textual styles.

```
> install.packages("tm")

>install.packages("wordcloud")

>library(wordcloud)

#use the wordcloud library to visualize which country is mentioned most in year 0f 2011

>a <- mydata$country[mydata$year == 2011]

>wordcloud(a, min.freq = 500, random.order = FALSE)
```



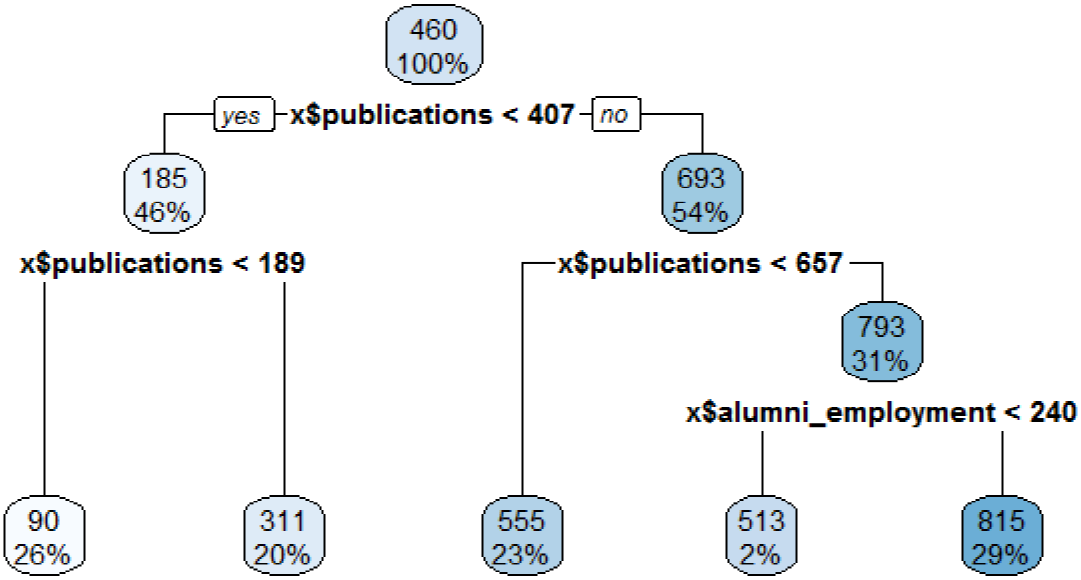
Task 10: find out distribution for number of universities in Canada with respect to the year.

```
> mydata <- read.csv("D:/LAKEHEAD WINTER 2019/DATA SCI/project/timesData.csv")

>View(mydata)

>mydata %>% group_by(year, country) %>% summarise(count = n()) %>% filter(country ==
"Canada") %>%

>ggplot(aes(country, count, fill = country)) + geom_bar(stat = "identity") + facet_wrap(~year)
+ geom_text(aes(label = count), vjust = -0.2) + labs(title = "Number of Universities in Canada",
y = "Number of Universities")
```



Task 11 : generate a decision tree for CWURDATA using rpart and tree packages.

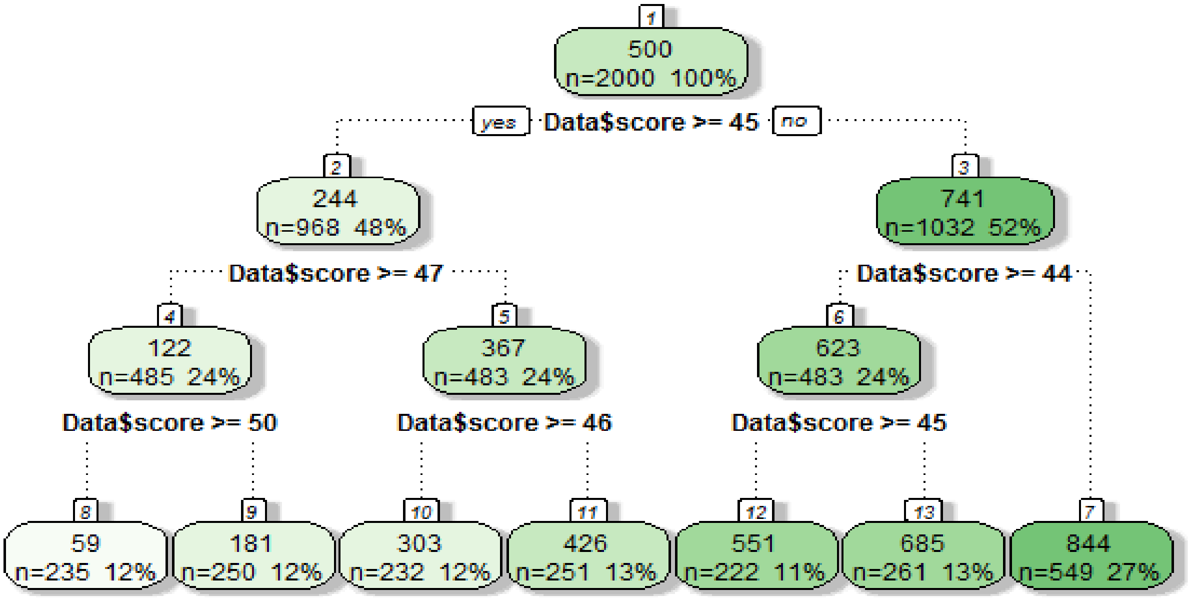
```
> x <- read.csv("D:/LAKEHEAD WINTER 2019/DATA SCI/project/cwurData.csv")

>f <- x[sample(nrow(x)),]

>x_train <- f[1:2000, ]

>x_test <- f[2001:2200, ]

>library(rpart)

>model <- rpart(x$world_rank ~ x$patent   +x$alumni_employment   +x$citations
+x$publications , data = x_train)

>install.packages("rpart.plot")

>install.packages("tree")

>library(tree)

>library(rpart.plot)

>rpart.plot(model, digits = 2)
```



Task 12: produce decision tree using “rpart”,”tree”,”rattle”,”RColorBrewer” library for CWURDATA.

```
> x <- read.csv("D:/LAKEHEAD WINTER 2019/DATA SCI/project/cwurData.csv")

>Data<-na.omit(x)
```

```
>library(rpart)

>library(tree)

>set.seed(101)

>alpha<-0.7

>train<-sample(1:nrow(Data),alpha*nrow(Data))

>train.set<-Data[train,]

>test.set<-Data[-train,]

>tree.model<-rpart(Data$world_rank ~
Data$score+Data$quality_of_education+Data$national_rank)

>install.packages("rattle")

>install.packages("RColorBrewer")

>library(rattle)

>library(RColorBrewer)

>plot(tree.model)

>text(tree.model)

>fancyRpartPlot(tree.model)
```



Rattle 2019-Mar-24 11:28:47 Bhavin