World University Ranking Analysis

The information in this pdf is related to a brief analysis of universities and their world rankings.

Two data sets have been primarily used CWUR data (Center for World University Rankings) and Times data (Times Higher Education World University Ranking).

The first part of the pdf deals with a simple analysis CWUR data and the later half deals with a brief Times data analysis.

Rstudio has been used for the data analysis.

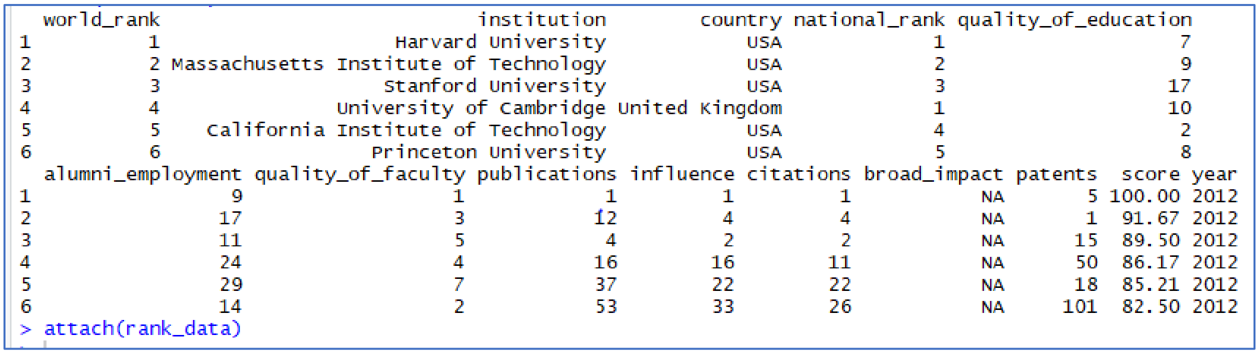
1. Center for World University Rankings (CWUR) data analysis:

Importing required libraries :

```
library(ggplot2)          #Baisc Data visualization
library(readr)
library(dplyr)
library(tidyr)
library(ggrepel)
library(RColorBrewer)
library(rworldmap)         #For map visualizations
library(rpart)             # for decision trees
library(rattle)
library(tidyverse)
library(stringr)
```

Loading data set into Rstudio

```
rank_data <- read.csv("cwurData.csv")
attach(rank_data)
head(rank_data)
```



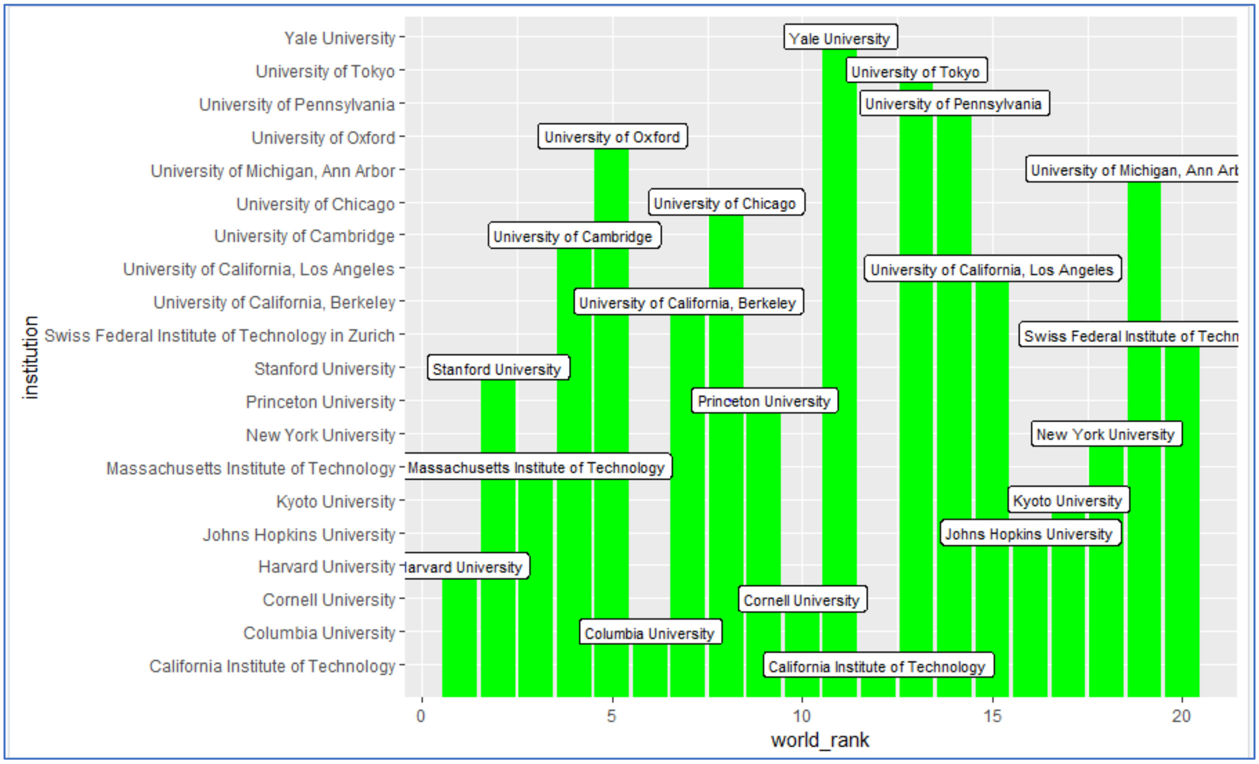
```
  world_rank                                  institution        country national_rank quality_of_education
1          1                           Harvard University            USA             1                    7
2          2 Massachusetts Institute of Technology            USA             2                    9
3          3                          Stanford University            USA             3                   17
4          4                     University of Cambridge United Kingdom             1                   10
5          5          California Institute of Technology            USA             4                    2
6          6                         Princeton University            USA             5                    8
  alumni_employment quality_of_faculty publications influence citations broad_impact patents  score year
1                 9                  1            1         1         1           NA       5 100.00 2012
2                17                  3           12         4         4           NA       1  91.67 2012
3                11                  5            4         2         2           NA      15  89.50 2012
4                24                  4           16        16        11           NA      50  86.17 2012
5                29                  7           37        22        22           NA      18  85.21 2012
6                14                  2           53        33        26           NA     101  82.50 2012
> attach(rank_data)
```

As we can see, the CWUR has different attributes contributing to the world ranking of universities like national rank, quality of education, quality of faculty, university score etc. This data is for the years 2012-2015.

1a) Viewing the top 20 universities in the latest year in the dataset i.e 2015:

rank_data %>% filter(world_rank <= 20 & year == 2015) %>% select(world_rank, institution, year) %>% arrange(world_rank) %>%

ggplot(aes(x = world_rank, y = institution)) + geom_bar(stat = "identity",fill ="green") + geom_label(aes(label = institution),size = 3)

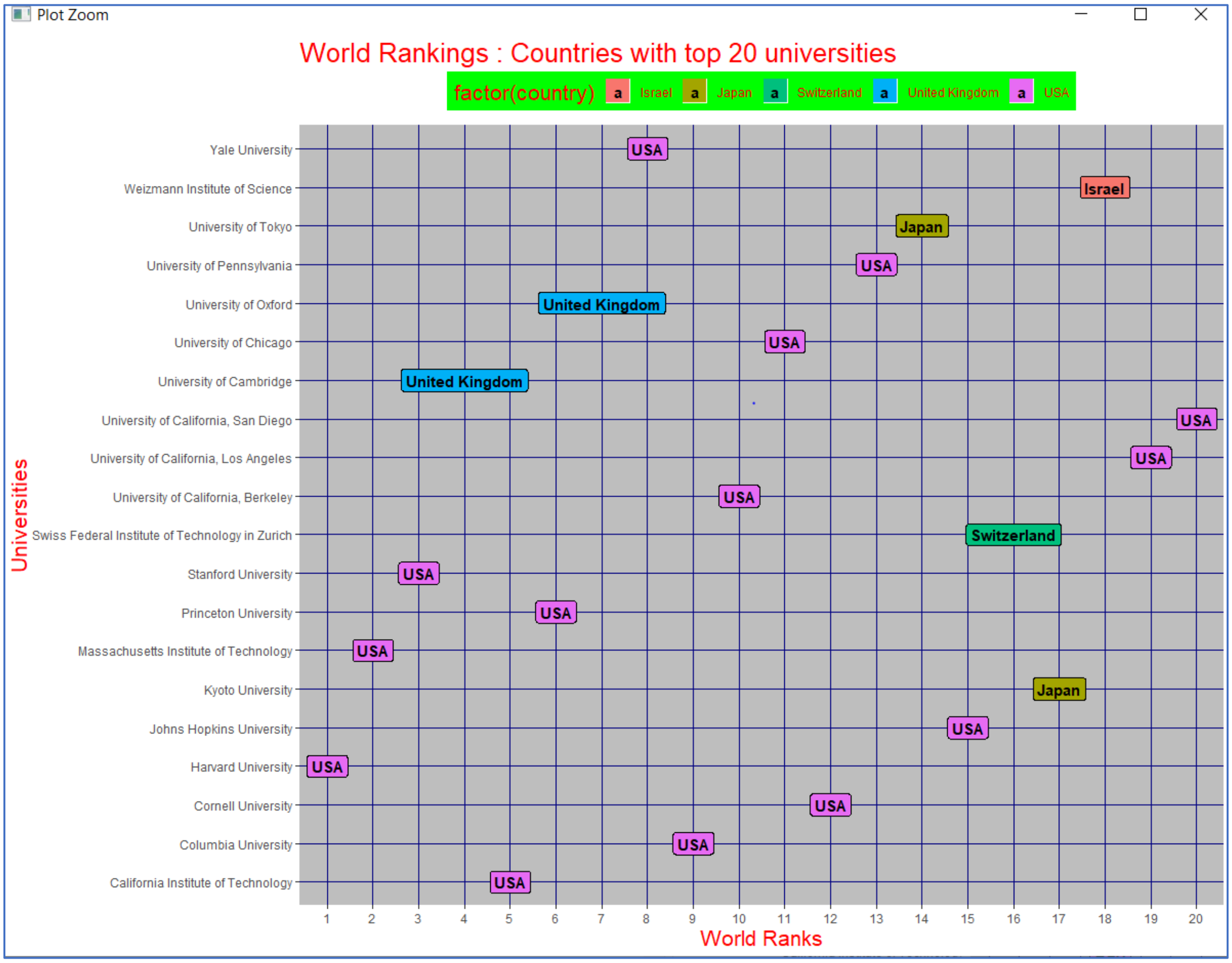


As we can see, Harvard, Stanford, Massachusetts, Cambridge and Oxford are the top 5 universities in world as of 2015.

1b) Viewing the top 20 universities along with the countries they belong to:

```
worldrankplot1 <- ggplot(rank_data[1:20,], aes(x=as.factor(world_rank), y=institution,
label=country))+geom_point() + geom_label( aes(fill = factor(country)), colour = "black", fontface =
"bold")+labs(title = "World Rankings : Countries with top 20 universities ")

worldrankplot1 + theme(  text = element_text(size=16, colour = "red"),
 axis.text = element_text(size = 9),  axis.text.x = element_text(size = 9),
 axis.text.y = element_text(size = 9),  legend.key = element_rect(fill = "white"),
 legend.background = element_rect(fill = "green"),  legend.position = "top",
 legend.text =  element_text(size = 9),  legend.direction = "horizontal",
 panel.grid.major = element_line(colour = "navy"),  panel.grid.minor = element_blank(),
 panel.background = element_rect(fill = "grey")) + xlab("World Ranks") + ylab("Universities")
```

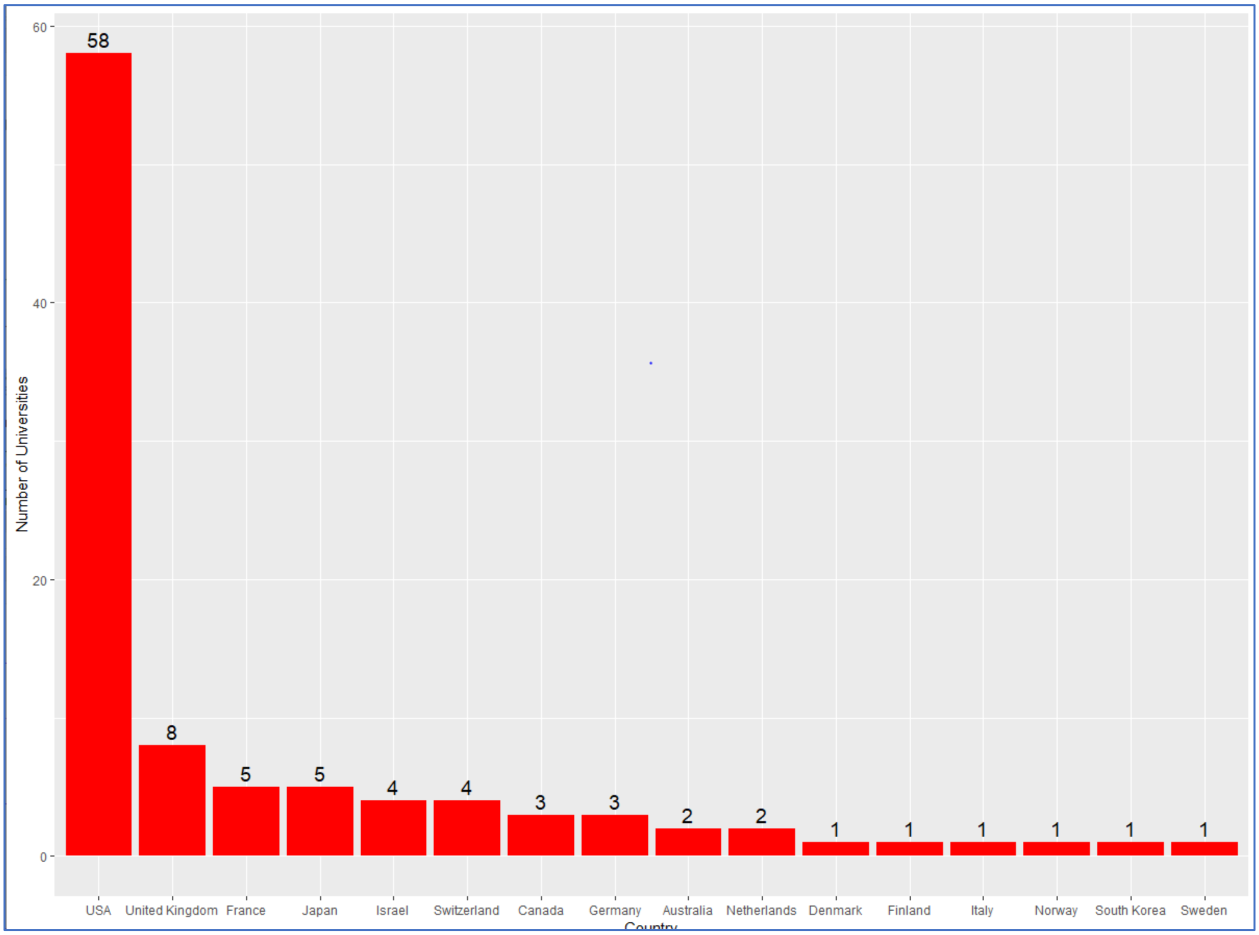


Its evident that among the top 20 universities, most of them belong to the country USA. UK, Japan, Israel and Switzerland are the other countries having their universities in the top 20.

1c) Number of Universities in a country 2012 vs 2015 comparison

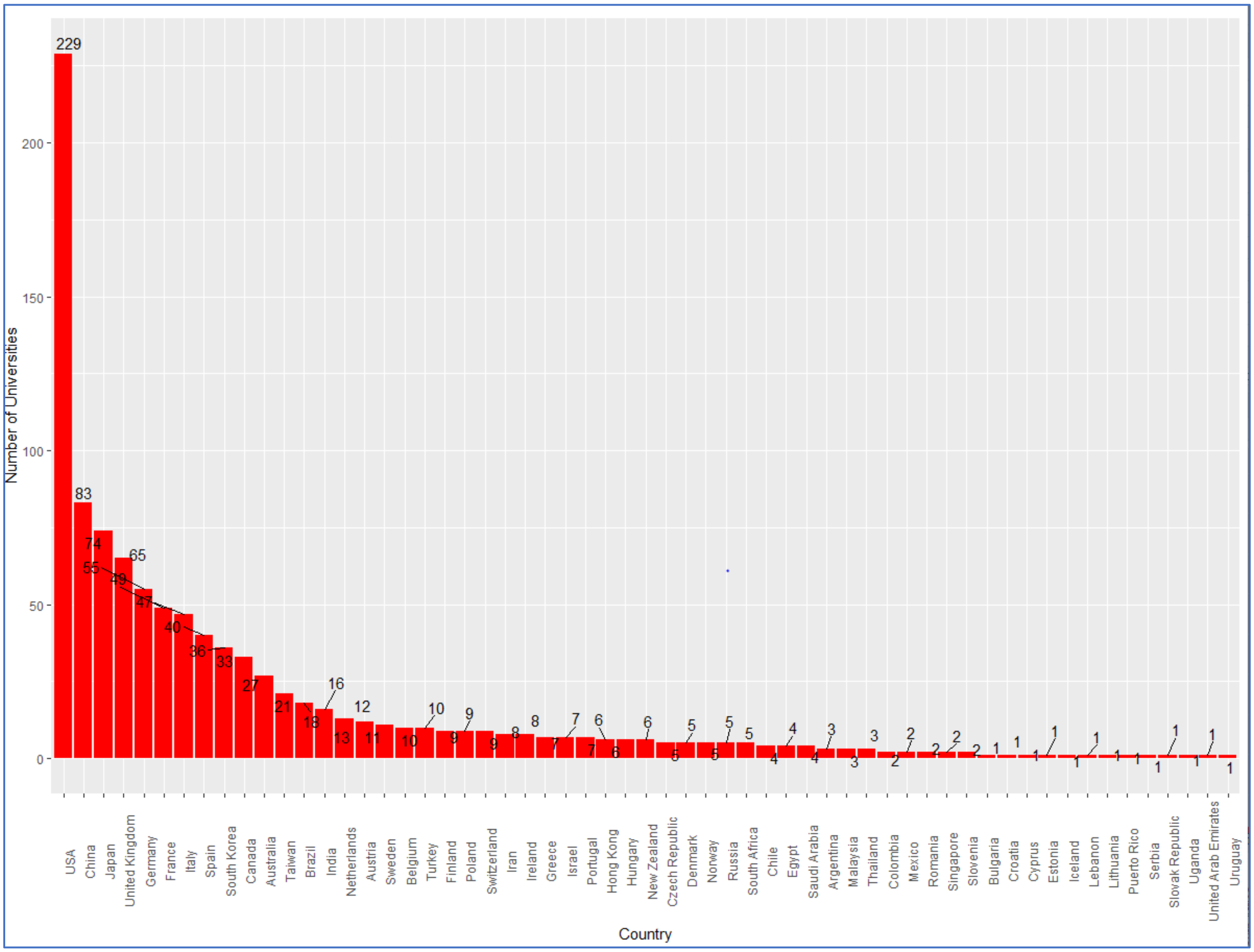
For the year 2012:

```
rank_data %>% filter(year == 2012) %>%  group_by(country) %>%
  summarise(stat_by_country = n_distinct(institution)) %>%  arrange(desc(stat_by_country)) %>%
  ggplot(aes(x = reorder(country,-stat_by_country), y= stat_by_country)) +
  geom_bar(stat = "identity", fill = "red")  +  geom_text_repel(aes(label = stat_by_country), vjust = -.4,
size = 5) +  xlab("Country") + ylab("Number of Universities")
```



For the year 2015:

```
  rank_data %>% filter(year == 2015) %>%
  group_by(country) %>%  summarise(stat_by_country = n_distinct(institution)) %>%
  arrange(desc(stat_by_country)) %>%  ggplot(aes(x = reorder(country,-stat_by_country), y=
stat_by_country)) +  geom_bar(stat = "identity", fill = "red")  +  geom_text_repel(aes(label =
stat_by_country), vjust = -.4, size = 5) +  xlab("Country") + ylab("Number of Universities")
```



All through the years 2012 to 2015, USA is the country with most number of years. The number of universities in all the countries have increased significantly. Most notably USA has almost 4 times the number of universities in 2015 when compared to 2012, China didn't have universities in 2012, but by 2015, it's the country with second highest number of universities (83). The number of universities in Japan has increased by a staggering 14 times almost, from 5 to 74.

Overall, there's been a huge increase in the number of universities in 2015 when compared to 2012.

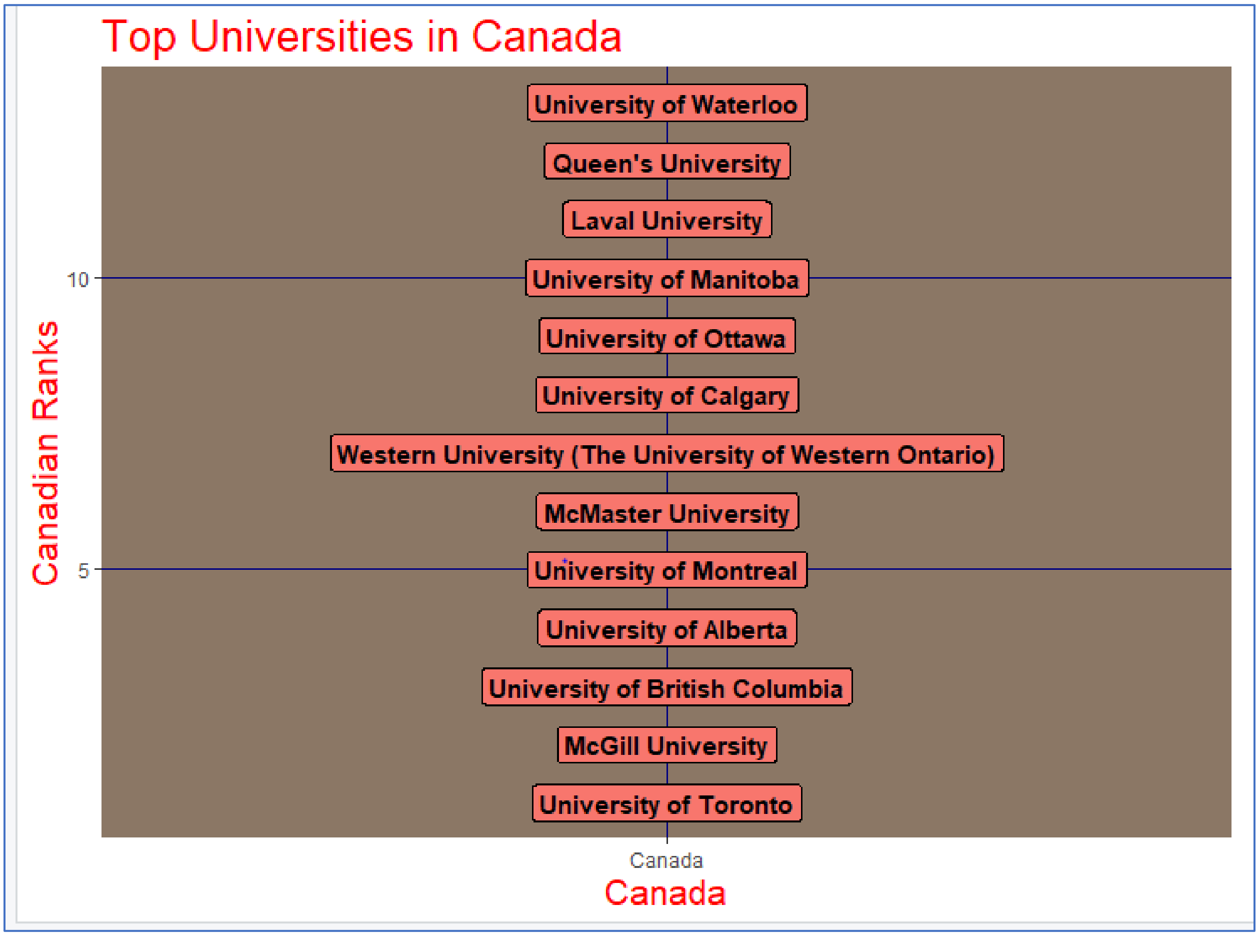
1d) *Quick look at the top universities of Canada*

Now let’s have a look at the top 20 universities of our country i.e Canada

```
canada_ranks <- rank_data[country == "Canada",]
```

```
canada_plot1 <- ggplot(canada_ranks[1:20,], aes(x=country, y=national_rank,
label=institution))+geom_point() + geom_label( aes(fill = factor(country)), colour = "black", fontface =
"bold")+labs(title = "Top Universities in Canada ")
```

```
canada_plot1 + theme(  text = element_text(size=16, colour = "red"),
  axis.text = element_text(size = 9),  axis.text.x = element_text(size = 9),
  axis.text.y = element_text(size = 9),  panel.grid.major = element_line(colour = "navy"),
  panel.grid.minor = element_blank(),  panel.background = element_rect(fill = "peachpuff4")
)+ xlab("Canada") + ylab("Canadian Ranks")
```



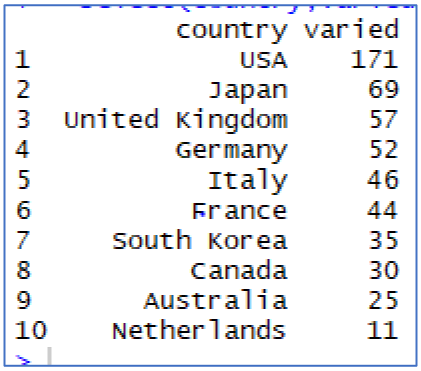
The top 5 universities in Canada are University of Toronto, McGill University, British Columbia and University of Alberta.

1e) *Countries with the highest increase in number of universities from 2012 to 2015*

```
temp <- rank_data %>% filter(year %in% c(2012, 2013, 2014, 2015)) %>%
group_by(country,year)%>%  summarise(stat_by_country = n_distinct(institution)) %>%
spread(year,stat_by_country)

colnames(temp) <- c("country","Y_2012","Y_2013","Y_2014","Y_2015")
temp <- as.data.frame(temp)

temp$varied <- temp$Y_2015 - temp$Y_2012
temp %>% arrange(desc(varied)) %>%  head(10) %>% filter(varied > 0) %>%  select(country,varied)
```



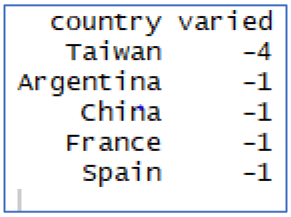
```
          country varied
1             USA    171
2           Japan     69
3  United Kingdom     57
4         Germany     52
5           Italy     46
6          France     44
7     South Korea     35
8          Canada     30
9       Australia     25
10    Netherlands     11
```

USA has 171 more universities in 2015 when compared to 2012, Japan has the second highest increase in number i.e 69.

1f) *Which countries saw a decrease in the number of universities in 2015 compared to the previous year i.e 2014*

```
temp$varied <- temp$Y_2015 - temp$Y_2014

temp %>% arrange(varied) %>%  head(10) %>% filter(varied < 0) %>%  select(country,varied)
```



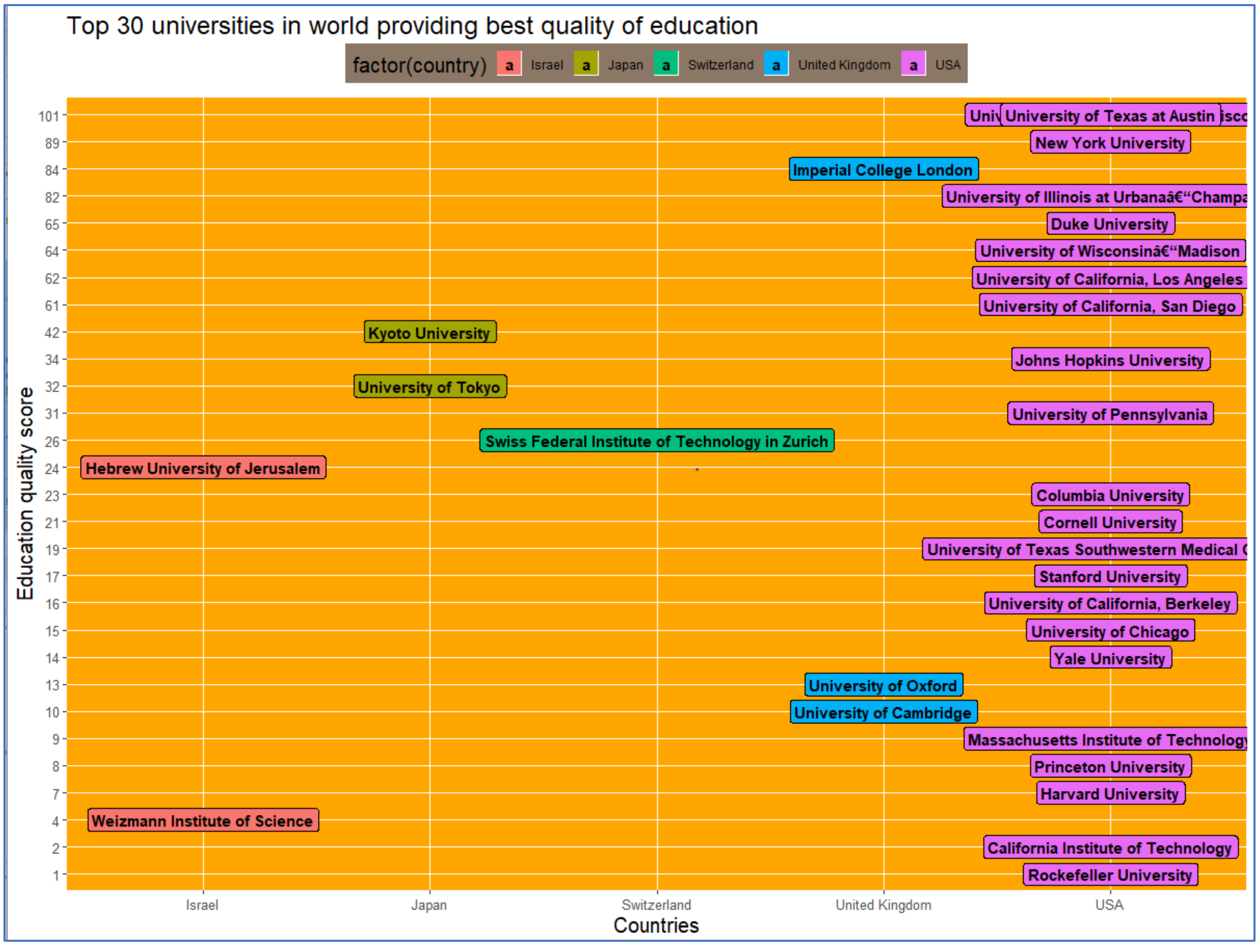
```
   country varied
    Taiwan     -4
 Argentina     -1
     China     -1
    France     -1
     Spain     -1
```

Taiwan, Argentina are some of the countries in which the number of universities has decreased.

1g) Having a look at the universities ranked best in providing Quality of education

```
QoE <- ggplot(rank_data[1:30,], aes(x=country, y=as.factor(quality_of_education),
label=institution))+geom_point() + geom_label( aes(fill = factor(country)), colour = "black", fontface =
"bold")+labs(title = "Top 30 universities in world providing best quality of education ")

QoE + theme(  text = element_text(size=15, colour = "black"),
axis.text = element_text(size = 10),  axis.text.x = element_text(size = 10),
axis.text.y = element_text(size = 10),  legend.key = element_rect(fill = "white"),
legend.background = element_rect(fill = "peachpuff4"),  legend.position = "top",  legend.text =
element_text(size = 9),  legend.direction = "horizontal",  panel.grid.minor = element_blank(),
panel.background = element_rect(fill = "orange"))+ xlab("Countries") + ylab("Education quality score")
```



As the first 30 rows of the dataset has been considered, this plot depicts the top 30 universities which have provided best quality of education in the year 2012. Rockfeller University is ranked 1 when it comes to education quality.

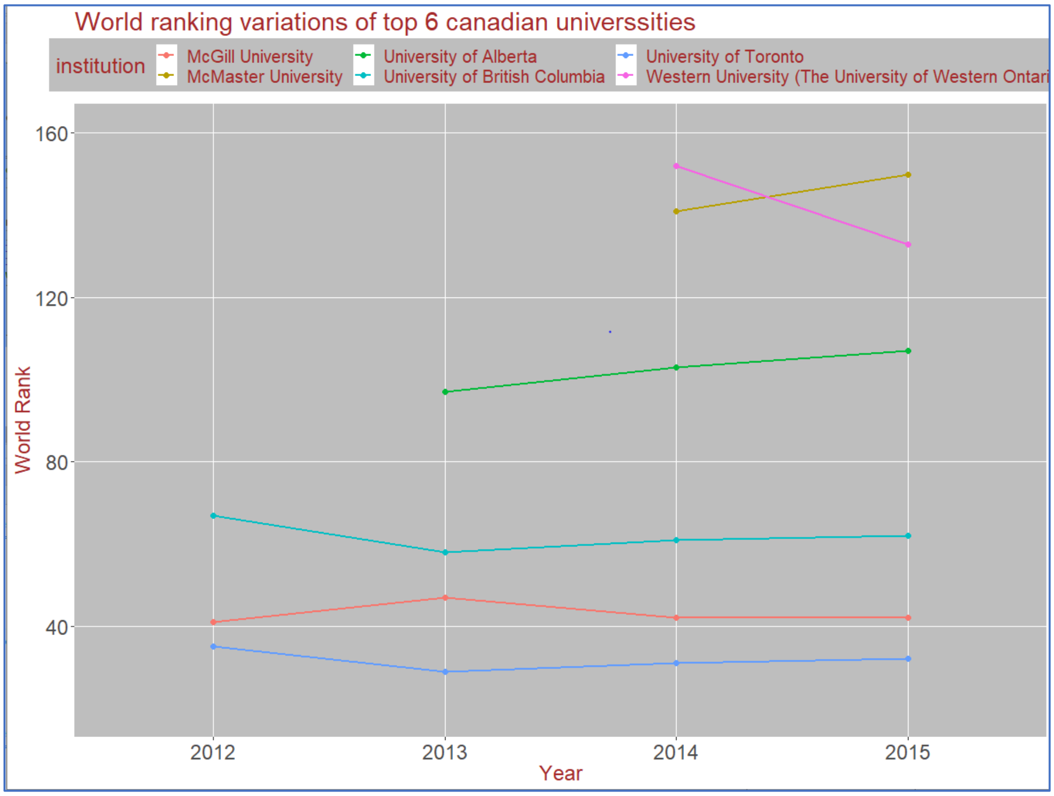
1h) Variations in world rankings for top 6 canadian universities

Lets have a look at the top 6 canadian universities and how their overall world ranking has varied from 2012 to 2015.

```
Uni_Toronto <- rank_data[grep("Toronto", rank_data$institution), ]
Uni_McGill <- rank_data[grep("McGill", rank_data$institution), ]
Uni_BC <- rank_data[grep("British Columbia", rank_data$institution), ]
Uni_Alberta <- rank_data[grep("Alberta", rank_data$institution), ]
Uni_WU <- rank_data[grep("Western University", rank_data$institution), ]
Uni_MM <- rank_data[grep("McMaster", rank_data$institution), ]
summary1 <- rbind(Uni_Toronto, Uni_McGill, Uni_BC, Uni_Alberta, Uni_WU, Uni_MM )

canada_plot2 <- ggplot(summary1, aes(x=as.factor(year), y=world_rank, color=institution,
group=institution)) +   geom_line(size=1) +geom_point(size=2) +ggtitle("World ranking variations of top
6 canadian universsities") +xlab("Year") +   ylab("World Rank") +ylim(20, 160)

canada_plot2+theme(  text = element_text(size=18, colour = "brown"), axis.text = element_text(size =
18),  axis.text.x = element_text(size = 18), axis.text.y = element_text(size = 18),  legend.key =
element_rect(fill = "white"), legend.background = element_rect(fill = "grey"),  legend.position = "top",
legend.text =  element_text(size = 15),  panel.grid.minor = element_blank(), panel.background =
element_rect(fill = "grey"))
```



We can see that University of Toronto, Western Ontario and British Columbia have been the top 3 canadian universities in all the four years, although there have been some fluctuations in their world rankings. University of Alberta was not in the top 6 in 2012, it was ranked among top 6 in 2013, although its world rank has slightly decreased over the next two years. McGill University and McMaster university have been ranked among the top 6 in the year 2014. However, one's national and world rank has improved while another one's national and world rank has diminished in 2015.

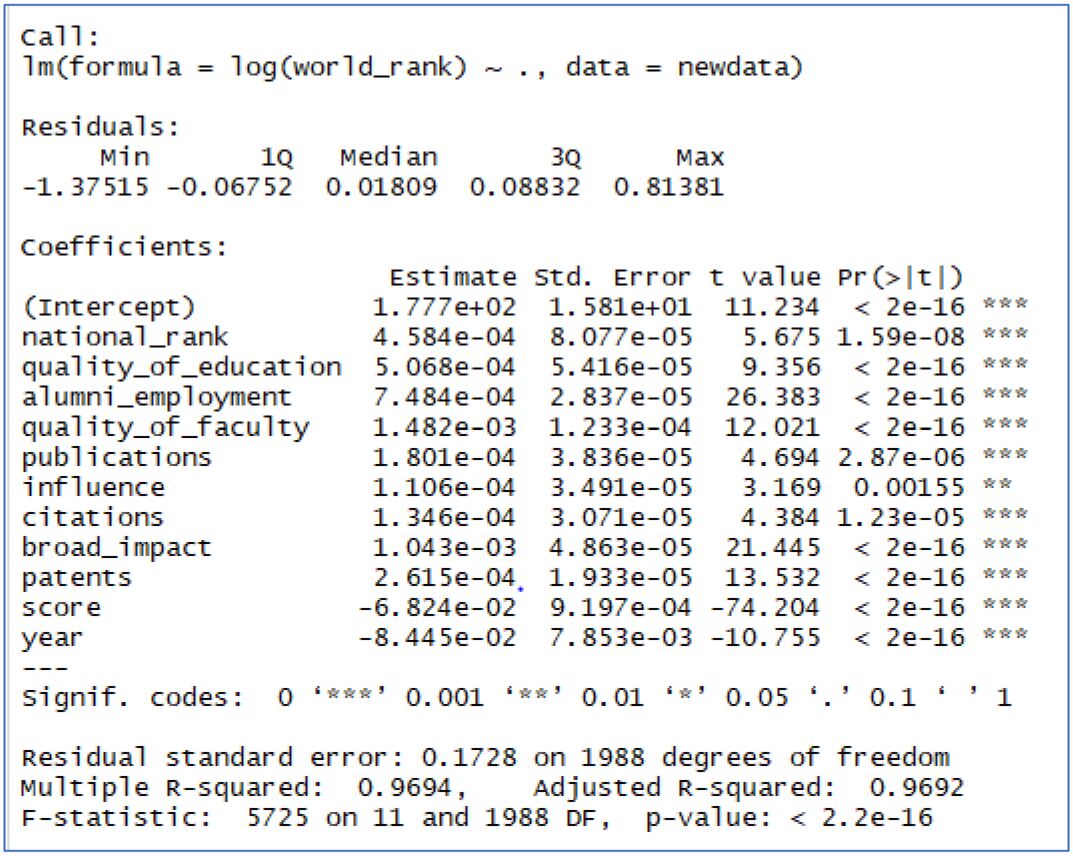
1i) Building a linear regression model for the world ranks of universities

Since the target and attribute variables should be continuous in order to make a linear regression model, first step is to convert all values of dataset into numeric values and store in a new data frame.

```
newdata <- sapply(rank_data, is.numeric)
newdata <- rank_data[,newdata]
newdata <- na.omit(newdata)
```

step 2 is building the LM model using the newly created data frame.

```
model1 <- lm(log(world_rank)~., newdata)
summary(model1)
```



```
Call:
lm(formula = log(world_rank) ~ ., data = newdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.37515 -0.06752  0.01809  0.08832  0.81381 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.777e+02  1.581e+01  11.234  < 2e-16 ***
national_rank         4.584e-04  8.077e-05   5.675 1.59e-08 ***
quality_of_education  5.068e-04  5.416e-05   9.356  < 2e-16 ***
alumni_employment     7.484e-04  2.837e-05  26.383  < 2e-16 ***
quality_of_faculty    1.482e-03  1.233e-04  12.021  < 2e-16 ***
publications          1.801e-04  3.836e-05   4.694 2.87e-06 ***
influence             1.106e-04  3.491e-05   3.169  0.00155 ** 
citations             1.346e-04  3.071e-05   4.384 1.23e-05 ***
broad_impact          1.043e-03  4.863e-05  21.445  < 2e-16 ***
patents               2.615e-04  1.933e-05  13.532  < 2e-16 ***
score                -6.824e-02  9.197e-04 -74.204  < 2e-16 ***
year                 -8.445e-02  7.853e-03 -10.755  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1728 on 1988 degrees of freedom
Multiple R-squared:  0.9694,    Adjusted R-squared:  0.9692 
F-statistic:  5725 on 11 and 1988 DF,  p-value: < 2.2e-16
```

- 2e-16 means almost zero

We have created a liner regression model for predicting world rank based on all other attributes in our cwur dataset. For a variable to play a significant role in prediction of target, the acceptable p-value limit is near to 0.05.

If the p-value of a variable is more than 0.05, then changes in the variable does not impact the target variable implying they are not a meaningful addition in our model.

Based on the above linear regression model, we can conclude that quality of education, alumni employment, quality of faculty, broad impact, patents, score etc almost all the variables are significant for the world ranking of the university.

1j) Decision tree to find if a university can be in top 20 or not

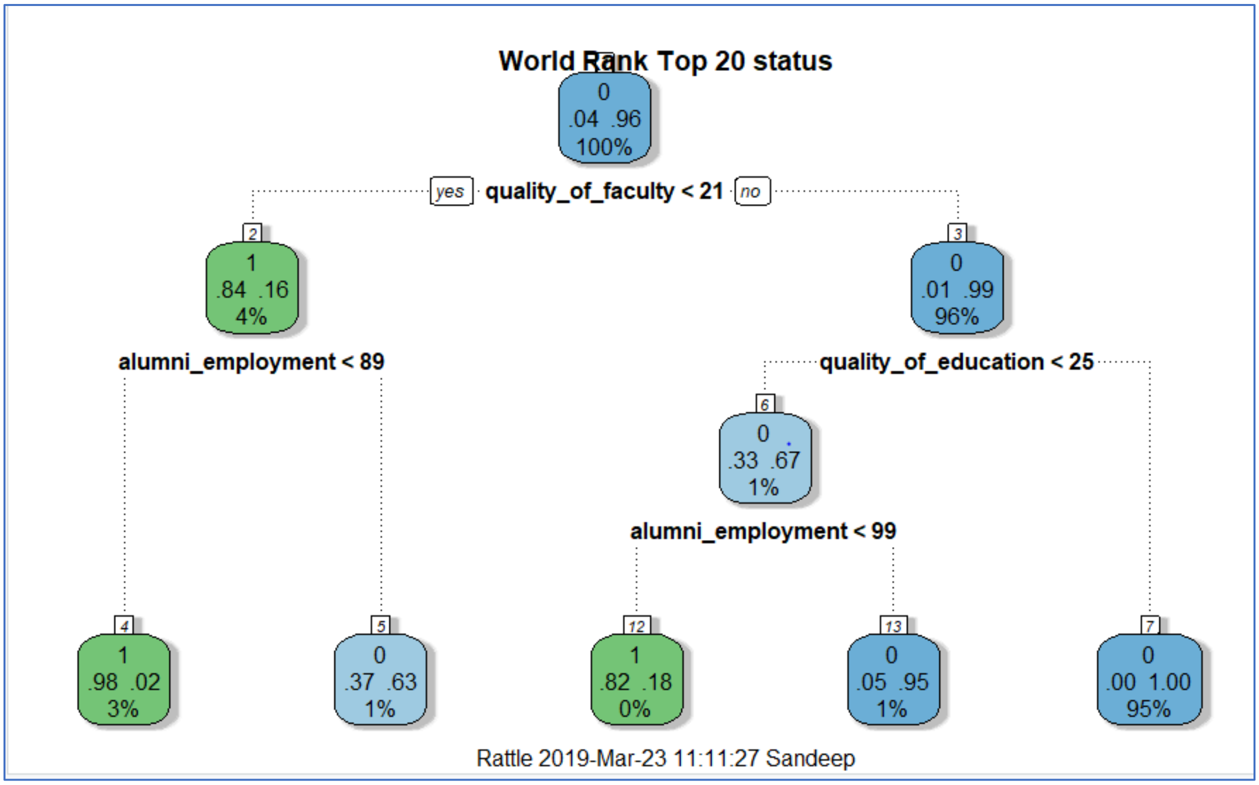
Let us categorise all the world rankings of universities into two categories i.e
1 implying the university is in top 20 and 0 implying the university is not in top 20.
For this we create a new variable and add it to our dataset

```
top20status <-cut(rank_data$world_rank, breaks=c(0,20,1000), labels=c("1", "0"))
rank_data$top20status <- top20status
```

Now construct a decision tree. This tree has top 20 rank status mapped against quality of education, quality of faculty and alumni employment.

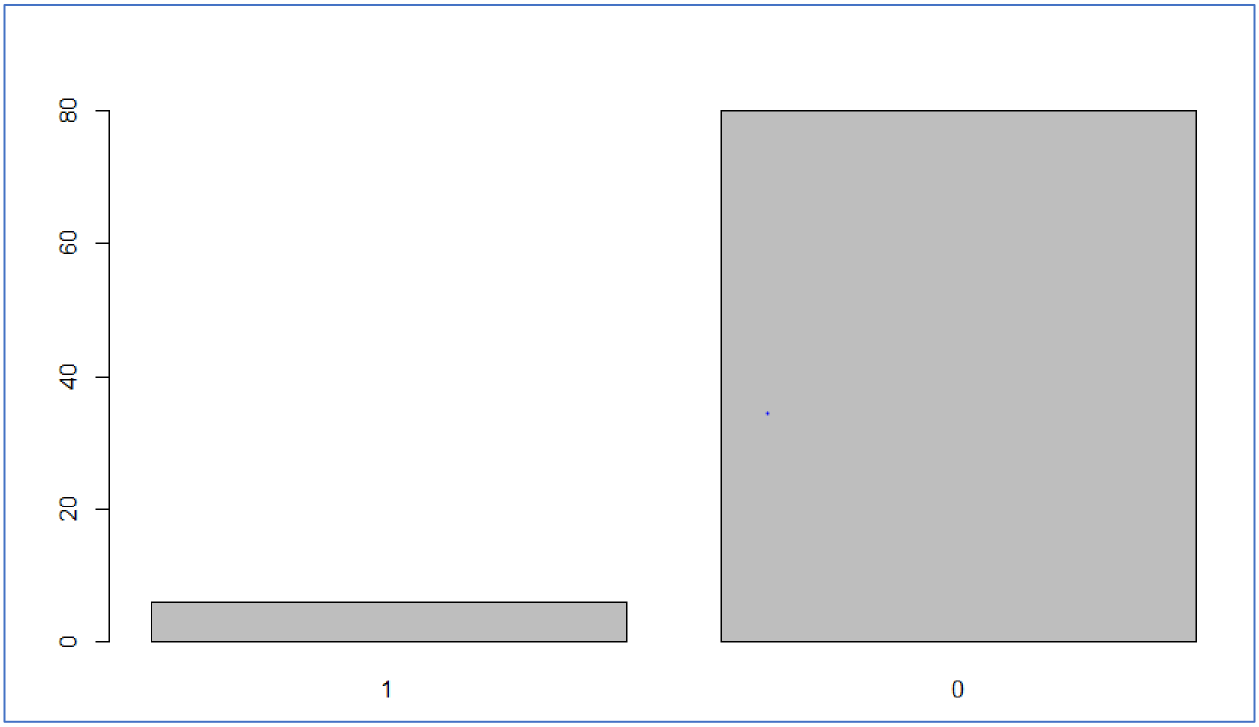
```
fitree <- rpart(top20status~ quality_of_education+quality_of_faculty+alumni_employment , data =
rank_data, method = "class")
fancyRpartPlot(fitree, main = "World Rank Top 20 status")
```

From the below decision tree, consider the first leaf node. If the quality of faculty ranking is less than 21 and if the alumni employment ranking is less than 89, then there is 98% chance that the university is in top 20. For the input data from the dataset, 3% of the universities fall under this category. Similarly, we can analyse other nodes as well.



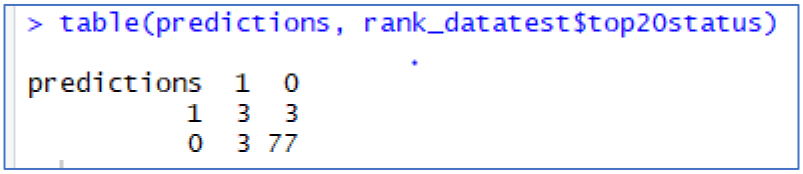
Now lets use this decision tree to predict the top 20 rank status of universities from 15th row to 100th row of our dataset.

```
predictions<-predict(fitree, rank_data[15:100,], type = "class")
plot(predictions)
```



Lets tabulate the prediction results with that of the actual data from dataset and see how accurate the prediction was:

```
rank_datatest <- rank_data[15:100,]
table(predictions, rank_datatest$top20status)
```



```
> table(predictions, rank_datatest$top20status)

predictions  1  0
          1  3  3
          0  3 77
```

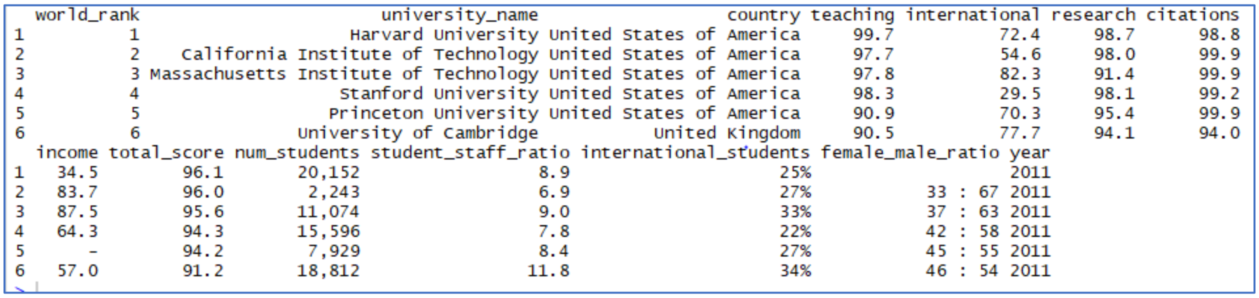
Out of the 86 input rows, 6 have been incorrectly predicted. The model is 93% accurate in predicting if a university falls under the world top 20 category or not.

2. Short analysis of Times data and expenditure data

Loading data in Rstudio:

```
time_data <- read.csv("timesData.csv"
expenditure_data <- read.csv("expenditure.csv")
```

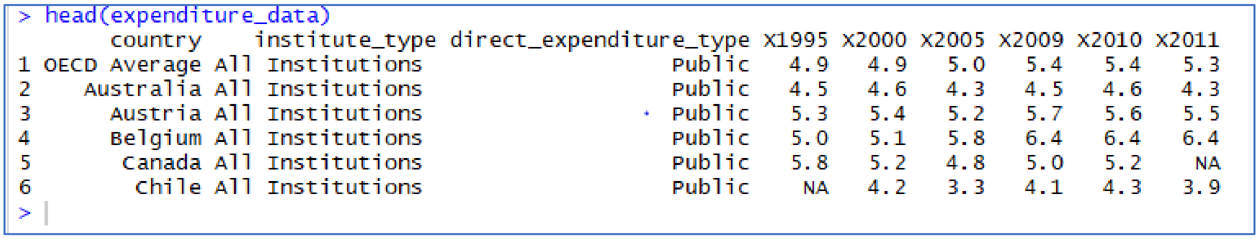
head(time_data)



```
  world_rank                       university_name                  country teaching international research citations
1          1                    Harvard University United States of America     99.7          72.4     98.7      98.8
2          2    California Institute of Technology United States of America     97.7          54.6     98.0      99.9
3          3 Massachusetts Institute of Technology United States of America     97.8          82.3     91.4      99.9
4          4                   Stanford University United States of America     98.3          29.5     98.1      99.2
5          5                  Princeton University United States of America     90.9          70.3     95.4      99.9
6          6               University of Cambridge           United Kingdom     90.5          77.7     94.1      94.0
  income total_score num_students student_staff_ratio international_students female_male_ratio year
1   34.5        96.1       20,152                 8.9                    25%                   2011
2   83.7        96.0        2,243                 6.9                    27%           33 : 67 2011
3   87.5        95.6       11,074                 9.0                    33%           37 : 63 2011
4   64.3        94.3       15,596                 7.8                    22%           42 : 58 2011
5      -        94.2        7,929                 8.4                    27%           45 : 55 2011
6   57.0        91.2       18,812                11.8                    34%           46 : 54 2011
```

The times data has world rankings, scores of universities in teaching, research, student staff ratio, international students percentage etc.

head(expenditure_data)

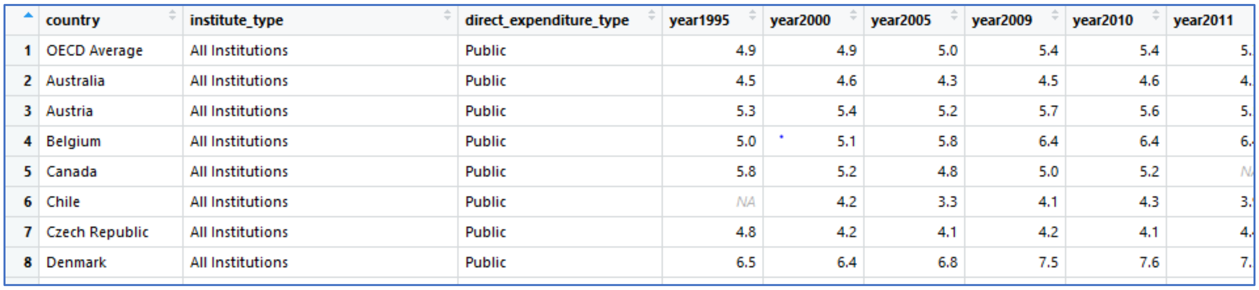


```
> head(expenditure_data)
        country     institute_type direct_expenditure_type X1995 X2000 X2005 X2009 X2010 X2011
1 OECD Average All Institutions                     Public   4.9   4.9   5.0   5.4   5.4   5.3
2     Australia All Institutions                     Public   4.5   4.6   4.3   4.5   4.6   4.3
3       Austria All Institutions                     Public   5.3   5.4   5.2   5.7   5.6   5.5
4       Belgium All Institutions                     Public   5.0   5.1   5.8   6.4   6.4   6.4
5        Canada All Institutions                     Public   5.8   5.2   4.8   5.0   5.2    NA
6         Chile All Institutions                     Public    NA   4.2   3.3   4.1   4.3   3.9
> 
```

The expenditure data has expenditure information for years 1995, 2000, 20005, 2009, 2011

We can see missing values for expenditures for many countries in some years.
Lets replace those null values with zeroes

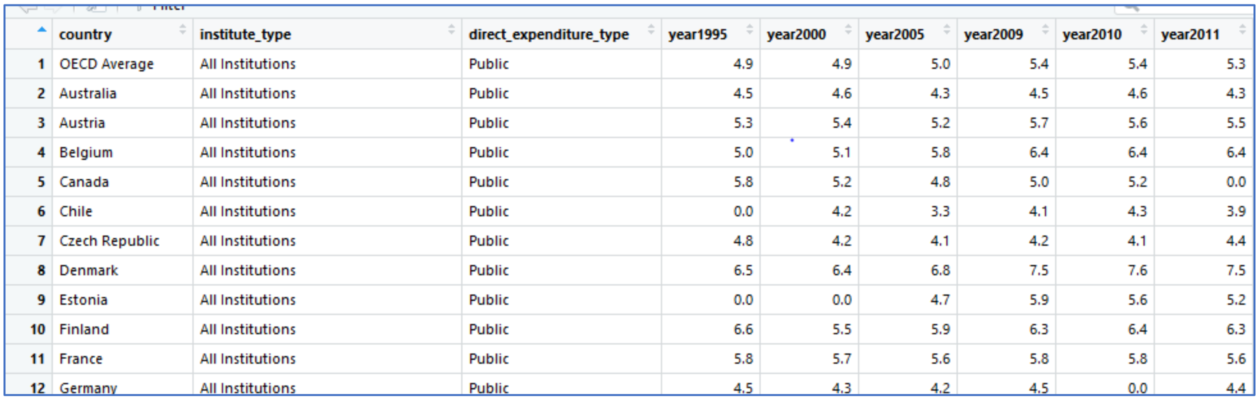
Initially:



	country	institute_type	direct_expenditure_type	year1995	year2000	year2005	year2009	year2010	year2011
1	OECD Average	All Institutions	Public	4.9	4.9	5.0	5.4	5.4	5.
2	Australia	All Institutions	Public	4.5	4.6	4.3	4.5	4.6	4.
3	Austria	All Institutions	Public	5.3	5.4	5.2	5.7	5.6	5.
4	Belgium	All Institutions	Public	5.0	5.1	5.8	6.4	6.4	6.
5	Canada	All Institutions	Public	5.8	5.2	4.8	5.0	5.2	NA
6	Chile	All Institutions	Public	NA	4.2	3.3	4.1	4.3	3.
7	Czech Republic	All Institutions	Public	4.8	4.2	4.1	4.2	4.1	4.
8	Denmark	All Institutions	Public	6.5	6.4	6.8	7.5	7.6	7.

Replacing Null values with 0 so that there no errors when plotting the data.

```
expenditure_data <- expenditure_data %>% mutate(year1995 = ifelse(is.na(year1995),0,year1995)
            ,year2000 = ifelse(is.na(year2000),0,year2000),year2005 = ifelse(is.na(year2005),0,year2005)
            ,year2009 = ifelse(is.na(year2009),0,year2009),year2010 = ifelse(is.na(year2010),0,year2010)
            ,year2011 = ifelse(is.na(year2011),0,year2011))
```



	country	institute_type	direct_expenditure_type	year1995	year2000	year2005	year2009	year2010	year2011
1	OECD Average	All Institutions	Public	4.9	4.9	5.0	5.4	5.4	5.3
2	Australia	All Institutions	Public	4.5	4.6	4.3	4.5	4.6	4.3
3	Austria	All Institutions	Public	5.3	5.4	5.2	5.7	5.6	5.5
4	Belgium	All Institutions	Public	5.0	5.1	5.8	6.4	6.4	6.4
5	Canada	All Institutions	Public	5.8	5.2	4.8	5.0	5.2	0.0
6	Chile	All Institutions	Public	0.0	4.2	3.3	4.1	4.3	3.9
7	Czech Republic	All Institutions	Public	4.8	4.2	4.1	4.2	4.1	4.4
8	Denmark	All Institutions	Public	6.5	6.4	6.8	7.5	7.6	7.5
9	Estonia	All Institutions	Public	0.0	0.0	4.7	5.9	5.6	5.2
10	Finland	All Institutions	Public	6.6	5.5	5.9	6.3	6.4	6.3
11	France	All Institutions	Public	5.8	5.7	5.6	5.8	5.8	5.6
12	Germany	All Institutions	Public	4.5	4.3	4.2	4.5	0.0	4.4

Similarly, we have to convert the university scores in each country to numeric, as total_score is not numeric in raw data

```
time_data$total_score <- as.numeric(time$total_score)
```

Some country names are different in both data sources so we'll have to keep uniform country names, for summarized data

```
name_matching <- c("Ireland", "Korea, Republic of", "United States" )
time_data["country"] <- str_replace(time_data$country,pattern = "Republic of Ireland", name_matching[1])
time_data["country"] <- str_replace(time_data$country,pattern = "South Korea", name_matching[2])
time_data["country"] <- str_replace(time_data$country,pattern = "United States of America", name_matching[3])
```

Replacing missing value with 0 for expenditures in all years and storing all expenditures in an object

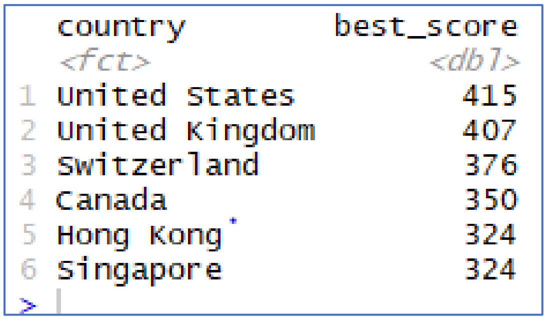
```
overall <- expenditure_data %>% mutate(year1995 = ifelse(is.na(year1995),0,year1995)
,year2000 = ifelse(is.na(year2000),0,year2000) ,year2005 = ifelse(is.na(year2005),0,year2005)
,year2009 = ifelse(is.na(year2009),0,year2009)  ,year2010 = ifelse(is.na(year2010),0,year2010)
,year2011 = ifelse(is.na(year2011),0,year2011))
```

2a) Let us consider the score of the best ranking institute of each country from the Times data.

```
countrywise_scores <- time_data %>% filter(total_score != '') %>%  group_by(country) %>%  summarise(best_score
= max(total_score)) %>%  select(country,best_score) %>% arrange(desc(best_score))
```

Now rearrange to best scores from highest to lowest to find which country has the best scoring university.

```
countrywise_scores$country <- factor(countrywise_scores$country , levels = countrywise_scores$country
[order(countrywise_scores$best_score)])
head(countrywise_scores)
```

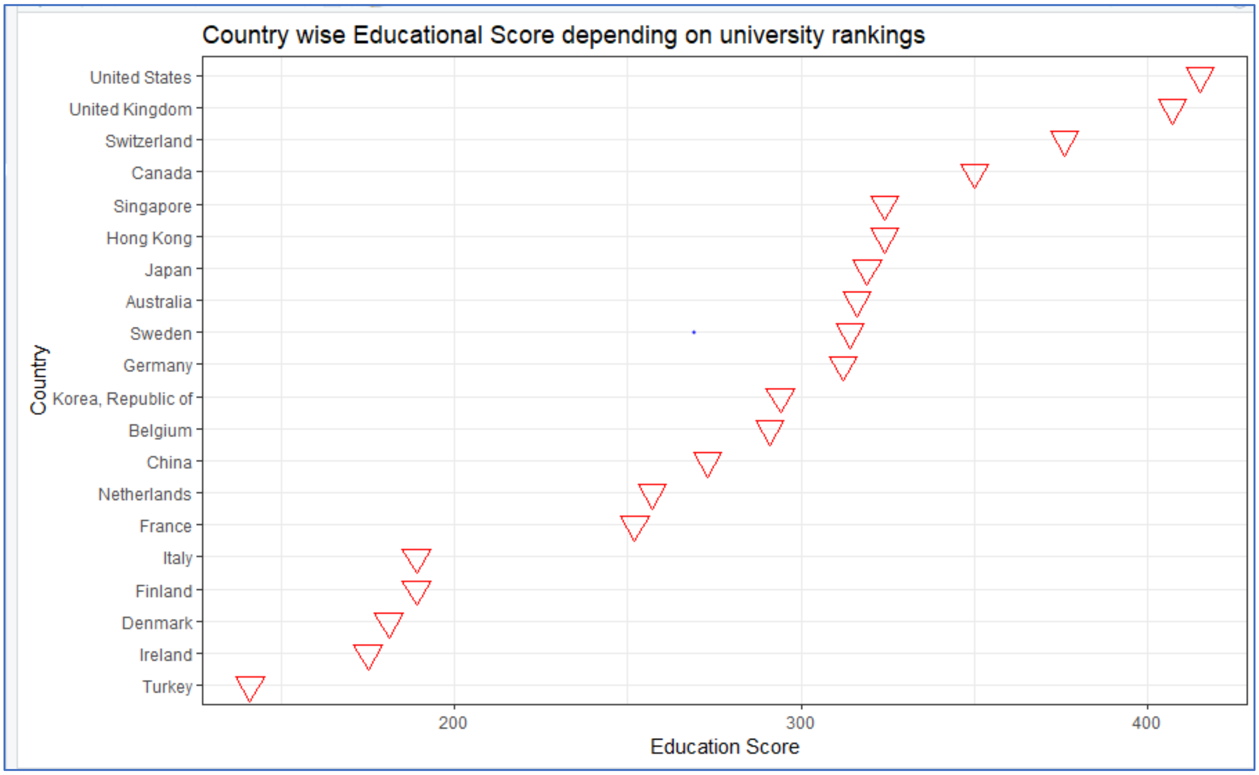


```
  country        best_score
  <fct>               <dbl>
1 United States         415
2 United Kingdom        407
3 Switzerland           376
4 Canada                350
5 Hong Kong             324
6 Singapore             324
> 
```

We can see that USA is the country with best scoring university.
It’s obvious that USA, UK, Switzerland, Canada and Singapore are the top 5 countries with highest university scores

2b) Scatterplot of best scores in each country

scores_plot <- countrywise_scores %>% top_n(20) %>% arrange(desc(best_score)) %>% ggplot(aes(x = best_score, y = country)) + geom_point(color ='red', size = 5, shape = 6) + labs(title = 'Country wise Educational Score depending on university rankings', x = 'Education Score', y ='Country') + heme(axis.text=element_text(size=10), axis.title=element_text(size=16,face="bold")) + theme_bw()



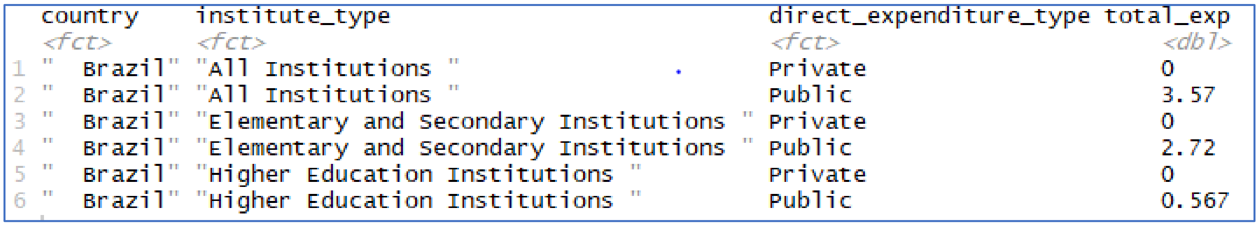
.

2c) Now let us examine the Overall expenditure trends of various countries and the various levels of educations in those countries

Calucating the average of the total expenditures through 1995-2011 for every institution type is expenditure data

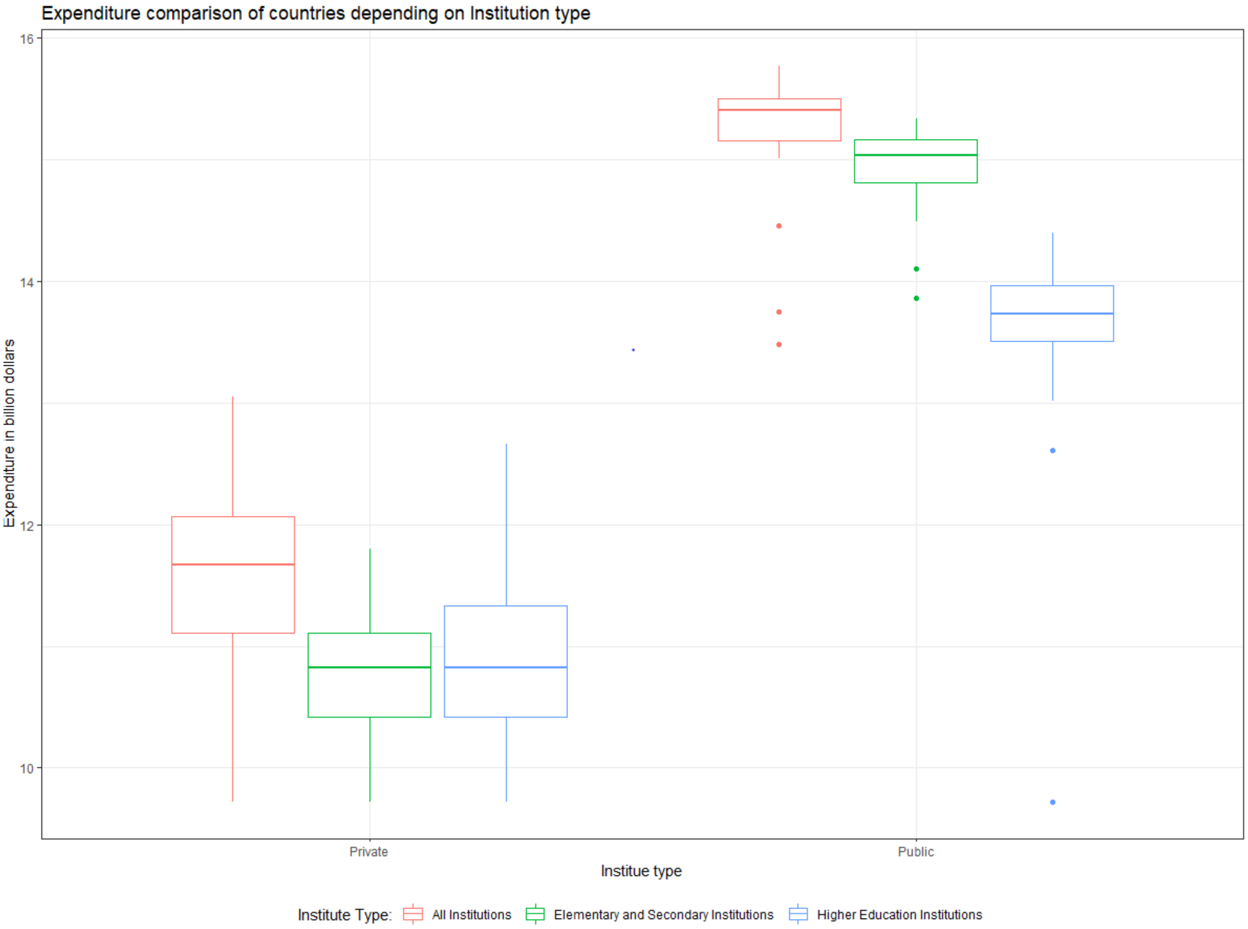
```
institute_expenditure <- overall %>% mutate(avg_exp = (year1995+ year2000 +year2005+year2009+year2010+year2011)/6) %>%   group_by(country,institute_type, direct_expenditure_type) %>% summarise(total_exp = sum(avg_exp)) %>% filter(direct_expenditure_type != "Total")
```

```
head(institute_expenditure)
```



```
  country     institute_type                                 direct_expenditure_type total_exp
  <fct>       <fct>                                          <fct>                       <dbl>
1 "   Brazil" "All Institutions "                            Private                     0    
2 "   Brazil" "All Institutions "                            Public                      3.57 
3 "   Brazil" "Elementary and Secondary Institutions "       Private                     0    
4 "   Brazil" "Elementary and Secondary Institutions "       Public                      2.72 
5 "   Brazil" "Higher Education Institutions "               Private                     0    
6 "   Brazil" "Higher Education Institutions "               Public                      0.567
```

boxplot1 <- institute_expenditure %>% ggplot(aes(direct_expenditure_type,log(total_exp * 10^6))) + geom_boxplot(aes(color = institute_type)) + labs(title = 'Expenditure by countries based on Institute type', y = 'expenditure in billion dollars', x = "Institute type") + theme(axis.text=element_text(size=8), axis.title=element_text(size=10,face="bold")) + theme_bw() +   theme(legend.position = "bottom") + scale_color_discrete("Institute Type:")



The public institutions expenditure is more when compared to private institutions expenditure.
In private institutions, the average expenditure is almost same for elementary,secondary institutions and higher education institutions. But when it comes to public universities, higher education institutions expenditure is significantly less compared to that of elementary and secondary institutions.

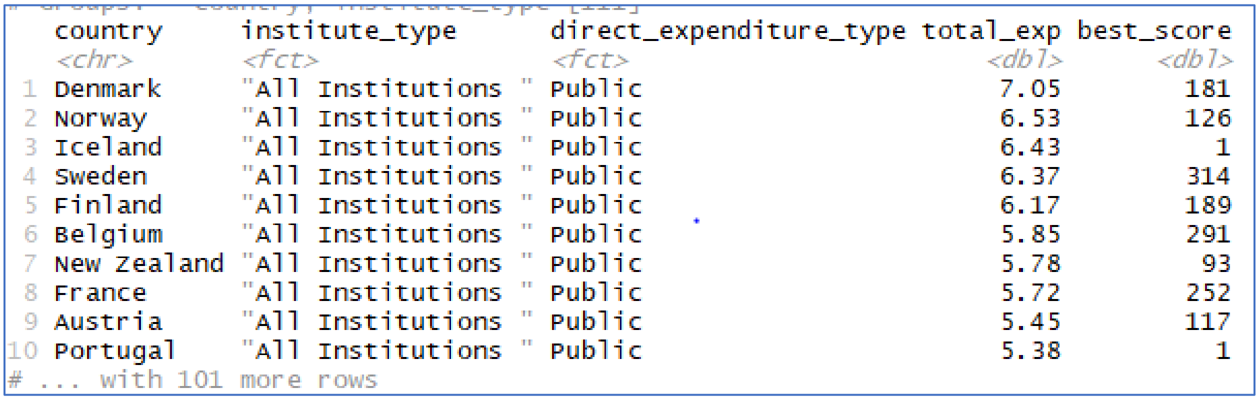
2c) Let us try to examine if the countries expenditure on public institutions have any impact on the country’s education ranking globally

First we merge the country wise score data with public expenditure data and filter for public education expenditure

```
expVscore_data <- left_join(institute_expenditure, countrywise_scores, by = "country") %>%
filter(direct_expenditure_type == "Public")
```

Replacing missing best score with 1 as that is the least score of times ranking to avoid errors while plotting.

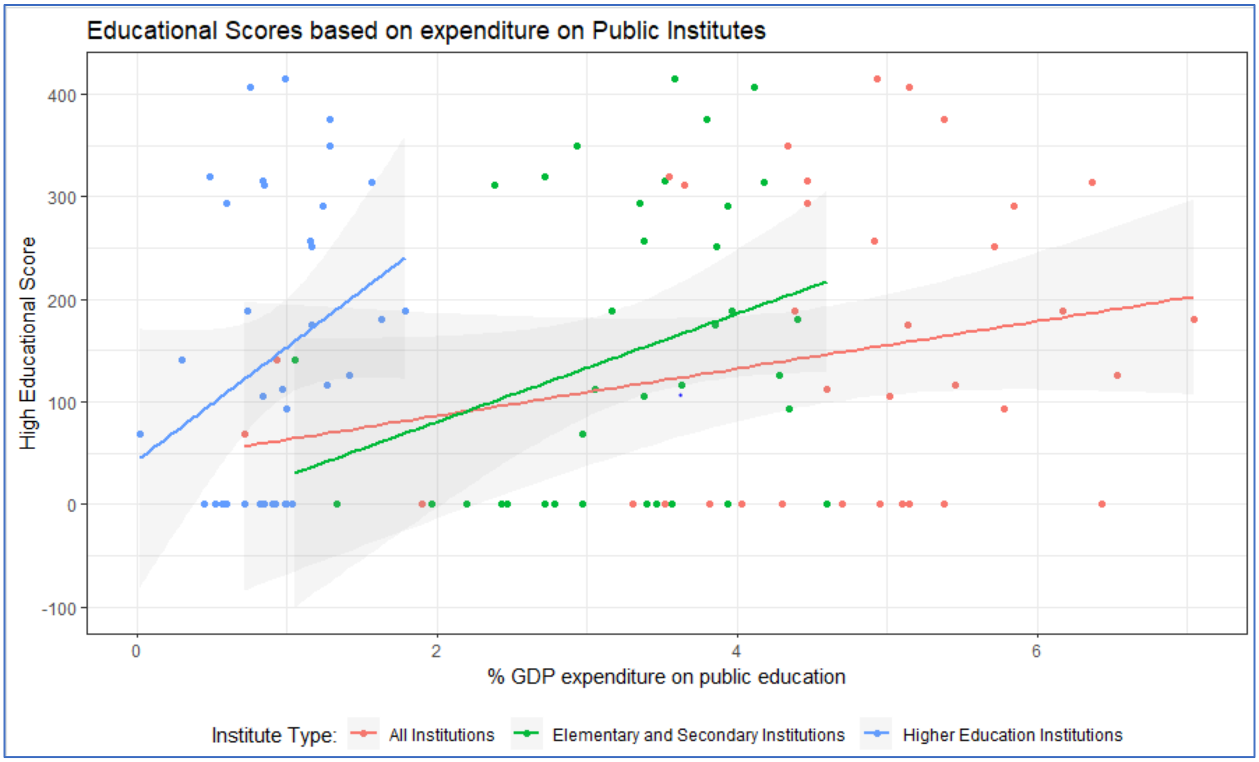
```
expVscore_data <- expVscore_data %>% mutate(best_score = ifelse(is.na(best_score),1,best_score))
expVscore_data %>% arrange(desc(total_exp))
```



```
# Groups:   country, institute_type [111]
   country     institute_type      direct_expenditure_type total_exp best_score
   <chr>       <fct>               <fct>                       <dbl>      <dbl>
 1 Denmark     "All Institutions " Public                       7.05        181
 2 Norway      "All Institutions " Public                       6.53        126
 3 Iceland     "All Institutions " Public                       6.43          1
 4 Sweden      "All Institutions " Public                       6.37        314
 5 Finland     "All Institutions " Public                       6.17        189
 6 Belgium     "All Institutions " Public                       5.85        291
 7 New Zealand "All Institutions " Public                       5.78         93
 8 France      "All Institutions " Public                       5.72        252
 9 Austria     "All Institutions " Public                       5.45        117
10 Portugal    "All Institutions " Public                       5.38          1
# ... with 101 more rows
```

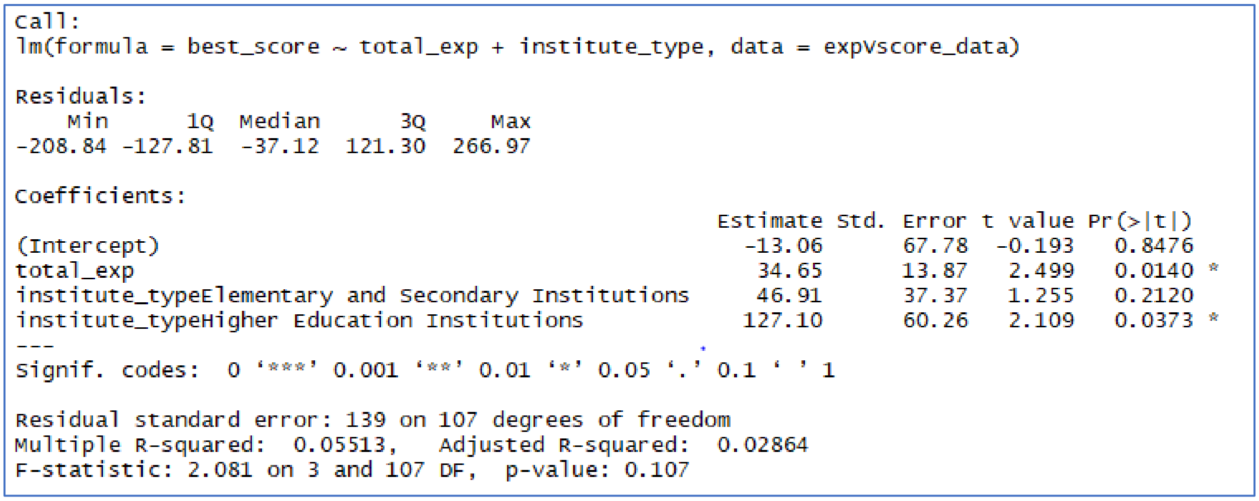
2d) Trying a Linear regression model for the above data (expenditure vs best score)

```
expVscore_data %>% ggplot(aes(total_exp,best_score)) + geom_point(aes(color = institute_type)) +
geom_smooth(method = lm, aes(group = institute_type, color = institute_type), alpha = 0.1) + labs(title =
'Educational Scores based on expenditure on Public Institutes', y = 'High Educational Score', x = '% GDP expenditure
on public education') +theme(axis.text=element_text(size=10),axis.title=element_text(size=10,face="bold" )) +
theme_bw() + theme(legend.position = "bottom") + scale_color_discrete(" Institute Type:")
```



Let us see the summary of the linear model

summary(lm(best_score ~ total_exp + institute_type , data = expVscore_data))



```
Call:
lm(formula = best_score ~ total_exp + institute_type, data = expVscore_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-208.84 -127.81  -37.12  121.30  266.97 

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)  
(Intercept)                                           -13.06      67.78  -0.193   0.8476  
total_exp                                              34.65      13.87   2.499   0.0140 *
institute_typeElementary and Secondary Institutions    46.91      37.37   1.255   0.2120  
institute_typeHigher Education Institutions           127.10      60.26   2.109   0.0373 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139 on 107 degrees of freedom
Multiple R-squared:  0.05513,	Adjusted R-squared:  0.02864 
F-statistic: 2.081 on 3 and 107 DF,  p-value: 0.107
```

It’s natural to assume that, the higher a country spends on its public institutes, the higher is education score will be world wide, but based on this linear model, the assumption is inconclusive as the p-values are not so near to zero. Hence this data should be refined more and more analysis is required before reaching a conclusion.