Emre Kemerci / 311802017 / BDA 503 Fall 2018 Final

Part I

Q1

I noted my insights for both languages below:

Python

- I have not enough experience but according to what I heard commonly, Python is more capable in modelling & machine learning
- I think syntax of Python is easier to read and write comparing the R. R has disadvantage especially due to paranthesis
- As far as I know, Python is preferred more among data analists/scientists

R

- Better for charming visualisation
- Better reporting capabilities and customization opportunities with Markdown comparing Jupyter notebook
- R Studio looks more friendly for new starters comparing PyCharm
- My insights are that R have variety of libraries, on the other hand coming with consistency problems of the libraries

Opinions stated in the dataschool website commonly overlap with my thoughts but at the end of the day, both languages have advantages and disadavntages. So I think that a data scientist should have capabilities to an extend in both languages and should prefer the one better satisfing the need.

Q2

First, the objective - question to be identified - should be clarified. Then I determine what I need to answer the question and continue with understanding what I have on hand. I design the execution of analysis considering the outputs from understanding the data on hand and execute accordingly to reach a conclusion answering the initial objective.

Following the example in the question; I can measure the impact of each project through finding the correlations of overall wellfare with each project. Delta of overall welfare per unit of improvement project. Most probably marginal benefit to overall welfare will decrease and considering the marginal benefit curve can be used to allocate funds efficiently. If I found a clear picture showing which projects should be focused for each welfare levels, info becomes valuable for me.

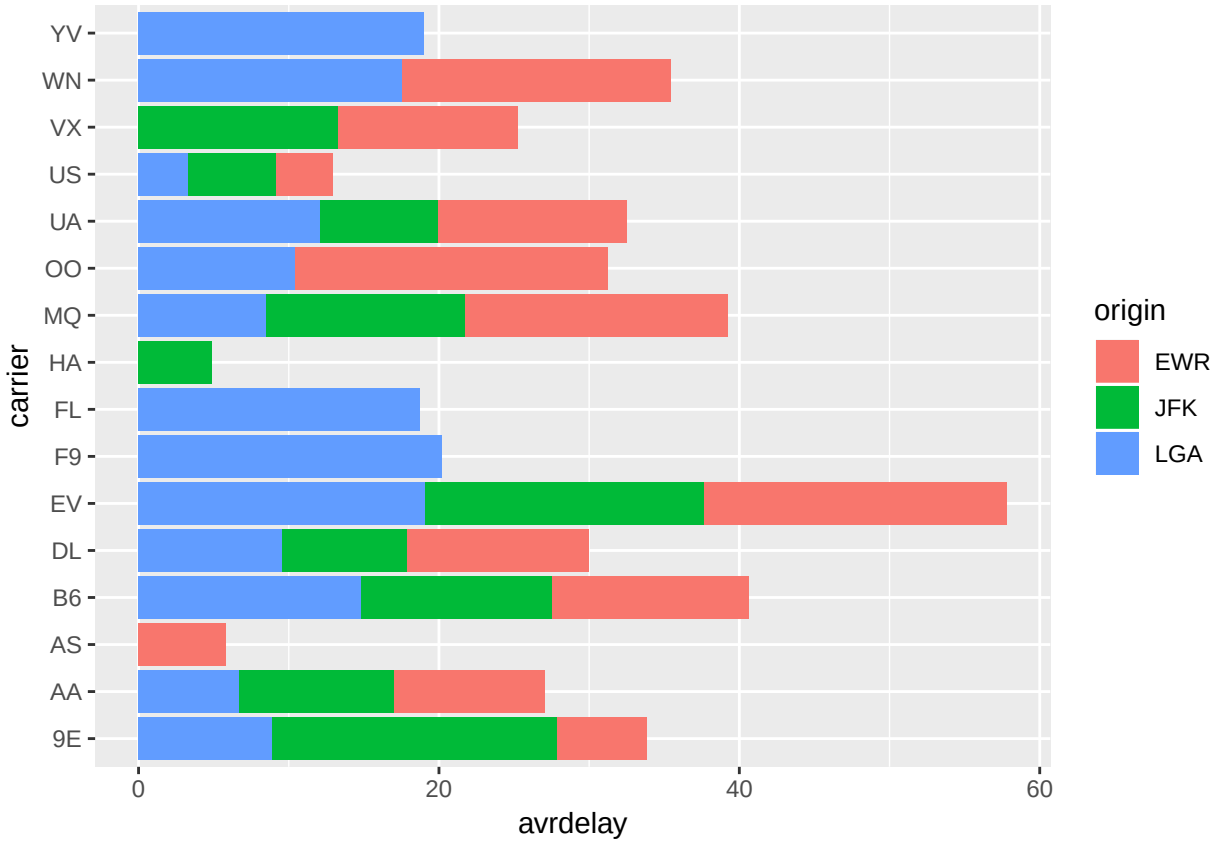
I prefer to present what the data says but I must explain the results of the analysis in a meaningful way.

Q3

I prefer to draw average delay time according to carriers and origins of flight. I choose delay time instead of arrival time because intuitively we know that departure delays are more than arrival delays. Carriers and the airports are key variables to analyze delay times, that is why I choose these variables.This plot can be used as starting point for following detailed analysis.

```
library(dplyr)
library(nycflights13)
library(ggplot2)
```

```
fl <- flights %>%
  select(dep_delay, carrier, origin) %>%
  group_by(carrier,origin) %>%
  summarize(avrdelay=mean(dep_delay, na.rm=TRUE))

ggplot (data=fl, aes(x=carrier, y=avrdelay)) +
  geom_bar(stat='identity', aes(fill=origin)) +
  coord_flip()
```



Part II

I choose to extend our group term project analysis through comparing the YoY percentage change of total export and import values with change in value of currencies of each countries. The steps of my analysis are as follows:

Loading the required libraries.

```
library(tidyverse)
library(readxl)
library(reshape2)
```

Reading our group project's cleaned the RDS files.

```
githubURL_export <- ("https://github.com/MEF-BDA503/gpj18-group_four/blob/master/export_jTable.rds?raw=
githubURL_import <- ("https://github.com/MEF-BDA503/gpj18-group_four/blob/master/import_jTable.rds?raw=
githubURL_exportMELTED <- ("https://github.com/MEF-BDA503/gpj18-group_four/blob/master/export_jTableMEL
```

```
githubURL_importMELTED <- ("https://github.com/MEF-BDA503/gpj18-group_four/blob/master/import_jTableMEL

export_jTable<- readRDS(url(githubURL_export))
import_jTable<- readRDS(url(githubURL_import))
export_jTableMELTED<- readRDS(url(githubURL_exportMELTED))
import_jTableMELTED<- readRDS(url(githubURL_importMELTED))

rm(githubURL_export)
rm(githubURL_import)
rm(githubURL_exportMELTED)
rm(githubURL_importMELTED)
```

I run the report on TCMB Website to gather closing FX rates of each year from 2013 to 2018 and save as Excel file. Then I read this file and manipulated.

```
Fxtable <- read_excel("C:/Users/emrek/Google Drive/BDA/503-EssentialsOfDataAnalytics/GitHub/pj18-EmreKe
Fxtable <- Fxtable %>% melt(id=c("year"))
colnames(Fxtable) <-  c("Year","Fx","FxYoYChg")
```

I also manipulated trade stat data of our porject

```
tableExport <- export_jTableMELTED %>%
  group_by(Country, year) %>%
  summarise(totalExpV=sum(Values,na.rm=TRUE)) %>%
  arrange(Country, year) %>%
  mutate(ExpValChng = (totalExpV - lag(totalExpV))/lag(totalExpV)*100)

tableImport <- import_jTableMELTED %>%
  group_by(Country, year) %>%
  summarise(totalImpV=sum(Values,na.rm=TRUE)) %>%
  arrange(Country, year) %>%
  mutate(ImpValChng = (totalImpV - lag(totalImpV))/lag(totalImpV)*100)

Tradetable <- full_join(tableExport,tableImport) %>%
  select(Country, year, Export="ExpValChng", Import = "ImpValChng") %>%
  melt(id=c("Country", "year")) %>%
  filter(year!=2013)

colnames(Tradetable) <-  c("Country","Year","TradeType","TradeYoYChg")

Tradetable <- Tradetable %>% mutate(Fx = ifelse (Country =="Canada", "CAD",
                                    ifelse(Country=="Russian", "RUB",
                                     ifelse(Country=="China", "CNY",
                                     ifelse(Country=="India", "INR",
                                     ifelse(Country=="Japan", "JPY",
                                     ifelse(Country=="UK", "GBP",
                                     ifelse(Country=="USA", "USD", "EUR")))))))

Tradetable <- transform(Tradetable, Year = as.numeric(Year))
```

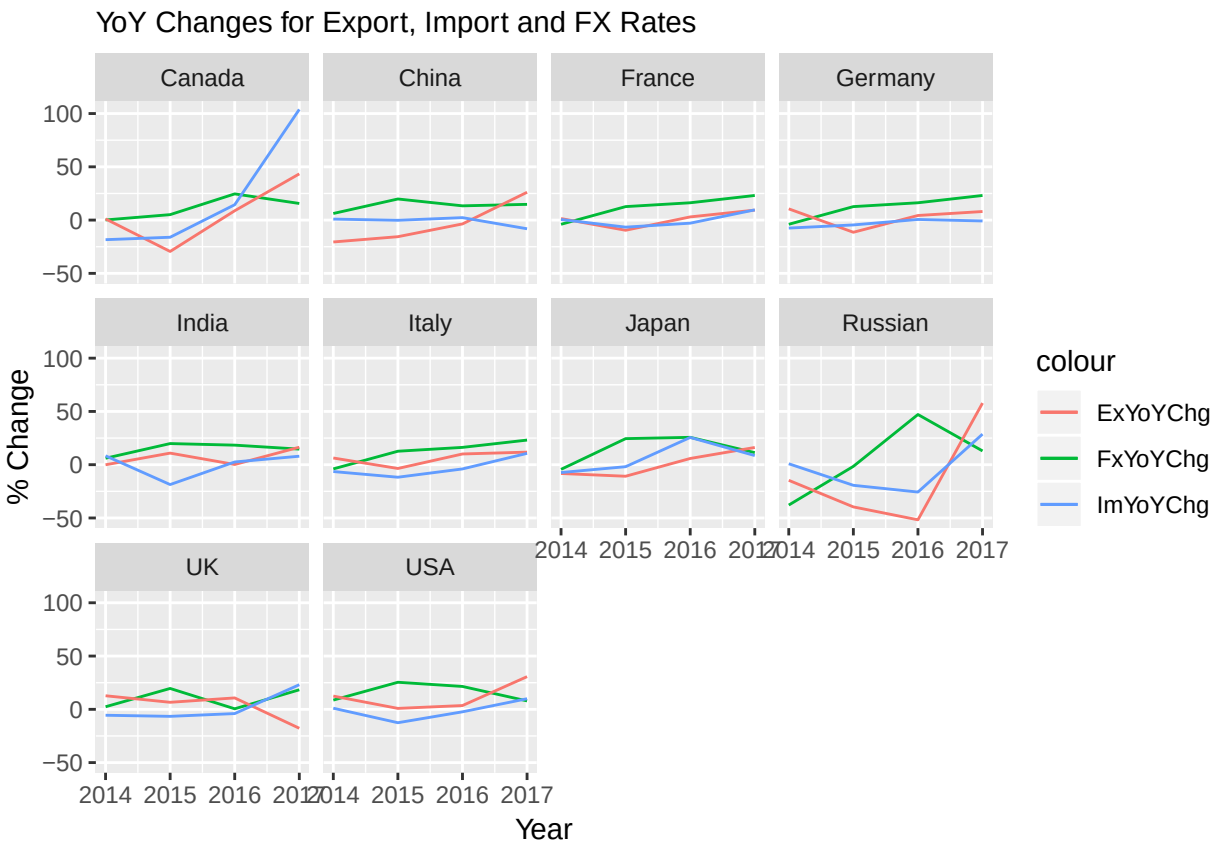
And join the trade stat data with FX data.

```
df <- left_join(Tradetable,Fxtable)

df <- df %>% select (Country, Fx, Year, FxYoYChg, TradeType, TradeYoYChg) %>%
  dcast(Country+Fx+Year+FxYoYChg~TradeType)
```

```
colnames(df) <-  c("Country", "Fx", "Year", "FxYoYChg", "ExYoYChg","ImYoYChg")
```

And visualized change in export value, change in import value and change of each countries' FX rate over years according to trade partner countries.

```
ggplot(df, aes(x=Year)) +
  geom_line(aes(y=FxYoYChg, col="FxYoYChg")) +
  geom_line(aes(y=ExYoYChg, col="ExYoYChg")) +
  geom_line(aes(y=ImYoYChg, col="ImYoYChg")) +
  facet_wrap(~Country) +
  labs(subtitle="YoY Changes for Export, Import and FX Rates",
       y="% Change")
```



Normally, I expect that in times TL depreciated against each countries currencies (hike in green line), export to these countries rises while imports declines.

I see this positive collerated pattern only for India but the other countries do not support this thesis. So, change of FX rate with countries does not enough to explain changes in export and import values.

Part III

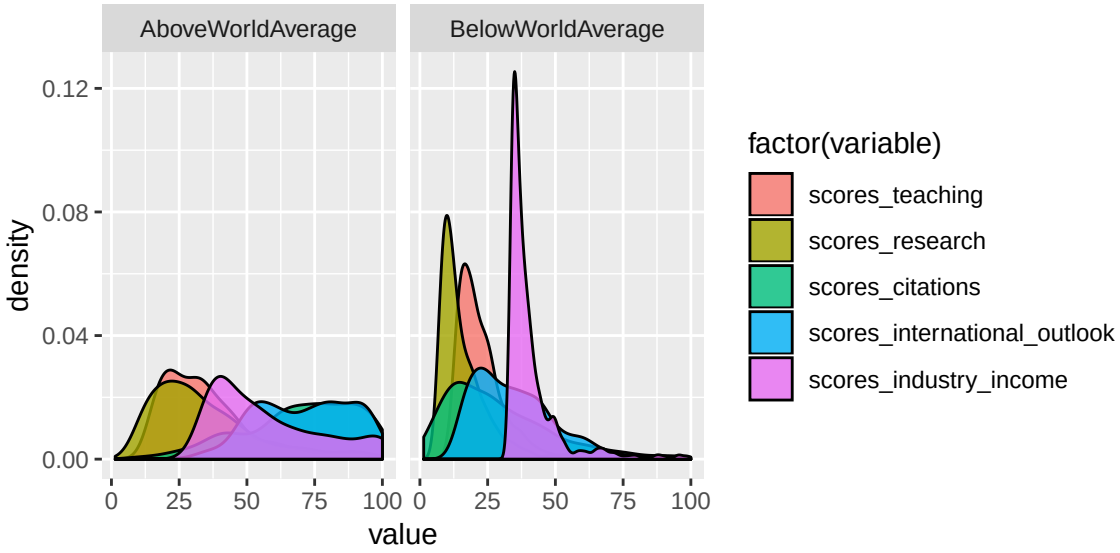
```
library(jsonlite)
url <- ("https://www.timeshighereducation.com/sites/default/files/the_data_rankings/world_university_ra

jsonfile <- fromJSON(url)
```

```
WUR_2019 <- jsonfile[[1]]

saveRDS(WUR_2019,file="C:/Users/emrek/Google Drive/BDA/503-EssentialsOfDataAnalytics/GitHub/pj18-EmreKe

rm(url)
rm(jsonfile)
rm(data)
```

I want to found out the universities which has high scores although their countries are not successful considering the average scores of all universities in this country.The data is extracted from timeshighereducation website and shows the 2019 survey.

First, I calculate world average score for each score type and average scores of each country per score type. Then I marked the countries as "Above Average" or "Below Average" considering whether each of their scores is above or below world average (per each score type).

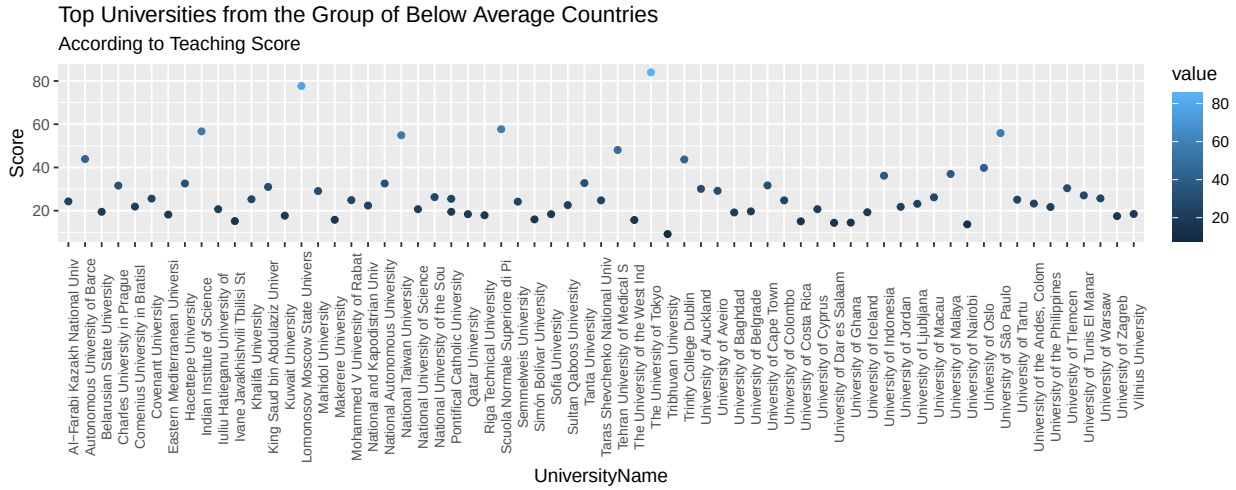
```
wur <- readRDS("C:/Users/emrek/Google Drive/BDA/503-EssentialsOfDataAnalytics/GitHub/pj18-EmreKemerci/F

wur <- wur %>% select (name, location, scores_teaching, scores_research, scores_citations, scores_inter
  melt(id=c("name", "location"))

wur$value <- as.numeric(wur$value)

wur <- wur %>% group_by(variable) %>% mutate(world_avr_score_per_score_type=mean(value))
wur <- wur %>% group_by(location, variable) %>% mutate(location_avr_score_per_score_type=mean(value))
wur <- wur %>% mutate(CountryGroup = ifelse(location_avr_score_per_score_type > world_avr_score_per_sco

ggplot(wur, aes(value)) +
  geom_density(aes(fill=factor(variable)), alpha=0.8) +
  facet_wrap(~CountryGroup)
```



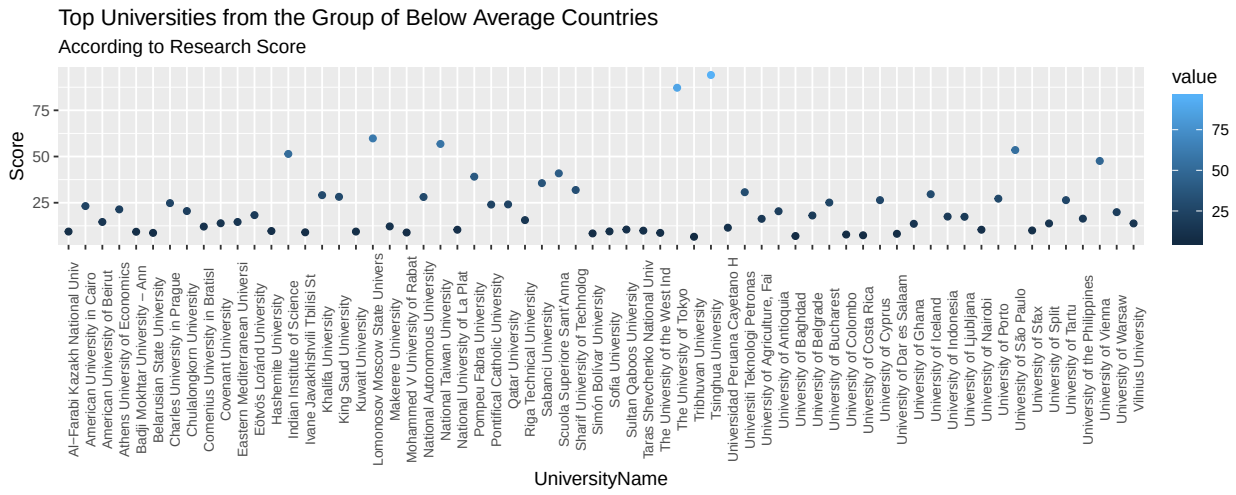
As seen above; "above world average"" group's scores spreads in a wide range although they are tighten in low scores for "below world average" group. Since I want to find the universities diverging their countries, I focus to find the universities in right tail of "below world average" group density graph.

To do this, I filtered "below world average" countries, then filtered top universities in each score type for each country. And visulazied as follows:

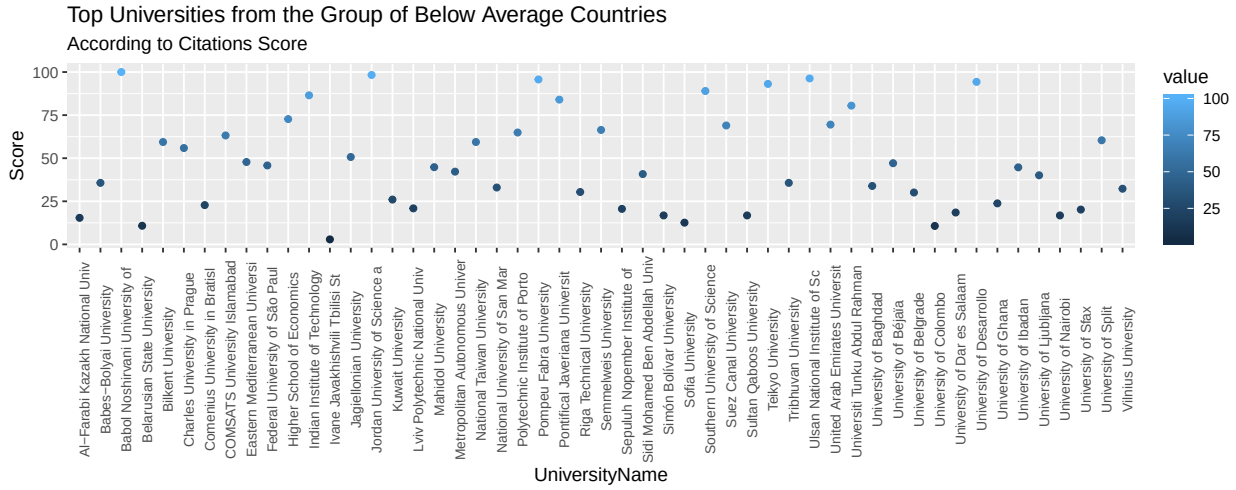
```
wur2 <- wur %>% filter(CountryGroup=="BelowWorldAverage")

wur2 <- wur2 %>% select (name, location, variable, value, location_avr_score_per_score_type)

wur2 <- wur2 %>% group_by(location, variable) %>% top_n(value, n=1)

wur_teaching <- wur2 %>% filter(variable=="scores_teaching")
wur_research <- wur2 %>% filter(variable=="scores_research")
wur_citations <- wur2 %>% filter(variable=="scores_citations")
wur_outlook <- wur2 %>% filter(variable=="scores_international_outlook")
wur_income <- wur2 %>% filter(variable=="scores_industry_income")

ggplot(wur_teaching, aes(x=substr(name, start = 1, stop = 30))) +
  geom_point(aes(y=value, color=value)) +
  theme(axis.text.x = element_text(size=8, angle = 90)) +
  labs(title="Top Universities from the Group of Below Average Countries", subtitle="According to Teachi
```



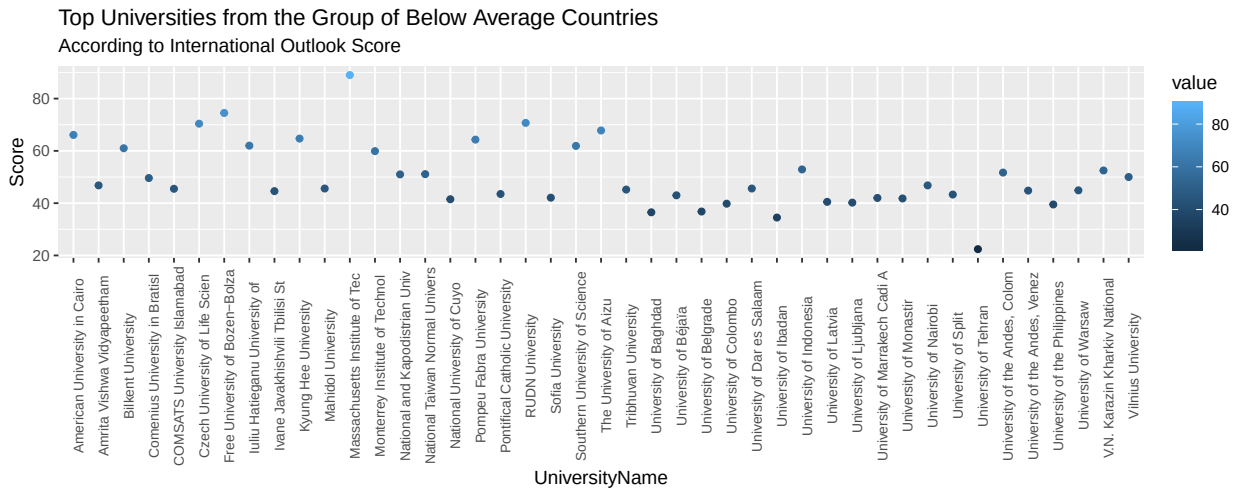

```
ggplot(wur_research, aes(x=substr(name, start = 1, stop = 30))) +
  geom_point(aes(y=value, color=value)) +
  theme(axis.text.x = element_text(size=8, angle = 90)) +
  labs(title="Top Universities from the Group of Below Average Countries", subtitle="According to Resea
```



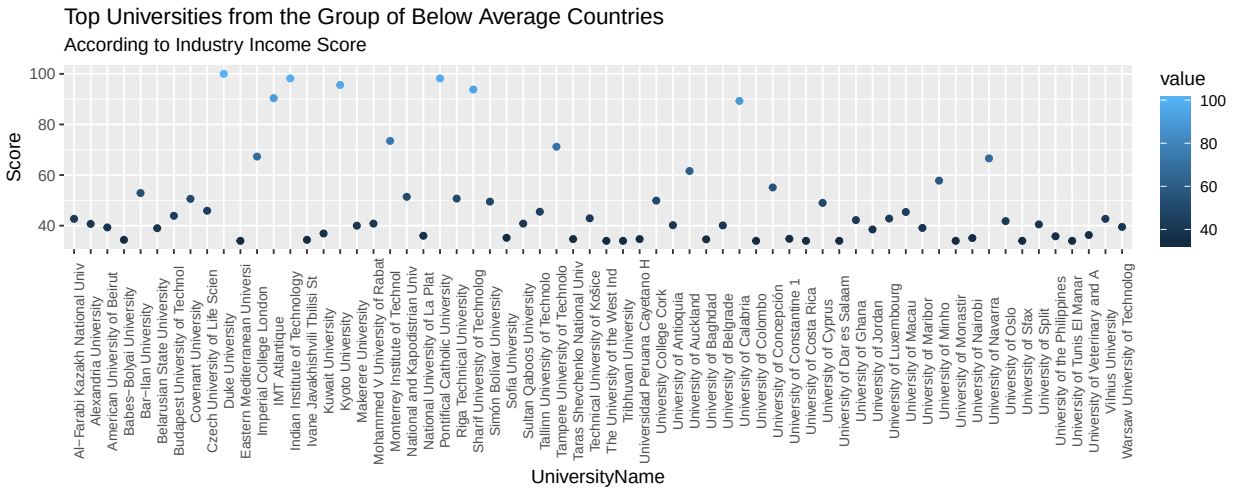
```
ggplot(wur_citations, aes(x=substr(name, start = 1, stop = 30))) +
  geom_point(aes(y=value, color=value)) +
  theme(axis.text.x = element_text(size=8, angle = 90)) +
  labs(title="Top Universities from the Group of Below Average Countries", subtitle="According to Citati
```




```
ggplot(wur_outlook, aes(x=substr(name, start = 1, stop = 30))) +
  geom_point(aes(y=value, color=value)) +
  theme(axis.text.x = element_text(size=8, angle = 90)) +
  labs(title="Top Universities from the Group of Below Average Countries", subtitle="According to Interr
```




```
ggplot(wur_income, aes(x=substr(name, start = 1, stop = 30))) +
  geom_point(aes(y=value, color=value)) +
  theme(axis.text.x = element_text(size=8, angle = 90)) +
  labs(title="Top Universities from the Group of Below Average Countries", subtitle="According to Indus
```



the universities which has high scores although their countries are not successful considering the average scores of all universities in this country are:

- University of Tokyo in “Teaching Score”" although Japan is below the world average in teaching score.
- Tsinghua University in “Research Score” although China is below the world average in research score.
- Babol Noshirvani University in “Citation Score” although Iran is below the world average in citation score.
- Massachusetts Inst. of Tech in “International Outlook Score” although USA is below the world average in international outlook score.
- Duke University in “Industry Income Score” although USA is below the world average in international outlook score.