BDA503 Exam

Efehan Danisman

27 Aralik 2018

Short and Simple

1. What is your opinion about Python vs R debate? To what extent do you agree with the post on https://www.dataschool.io/python-or-r-for-data-science/? Be honest, you won't be penalized or rewarded for stating your opinions; only by the quality your arguments.

I started by trying both R and Python to Data Analysis before starting this MA programme. So far my experience with R is enriching while Python becomes a demotivating nightmare time to time due to package and environment issues. While I didn't dive into everything that is said in this article (such as Machine Learning), I have had some experience in analyzing exploratory data, visualization and text mining with at least one of these tools.

For the machine learning part of the Python which I only dipped my toes so far, nonetheless I have high expectations. At the Data Science survey of the Kaggle or at the job requirements, Python leads significantly comparing to R. For that reason, I think there should be something at Python that I could not enjoy so far. Hence at the second term of this programme, I am hoping to understand Python's perks more by using it regularly.

2. What is your exploratory data analysis workflow? Suppose you are given a data set and a research question. Where do you start? How do you proceed? For instance, you are given the task to distribute funds from donations to public welfare projects in a wide range of subjects (e.g. education, gender equality, poverty, job creation, healthcare) with the objective of maximum positive impact on the society in general. Assume you have almost all the data you require. How do you measure impact? How do you form performance measures? What makes you think you find an interesting angle? Would you present an argument for a policy that you are more inclined to (e.g. suppose you are more inclined to allocate budget to fix gender inequality than affordable healthcare) or would you just present what data says? In other words, would the (honest) title of your presentation be "Gender Inequality - The Most Important Social Problem Backed by Data" or "Pain Points in Our Society and Optimal Budget Allocation"?

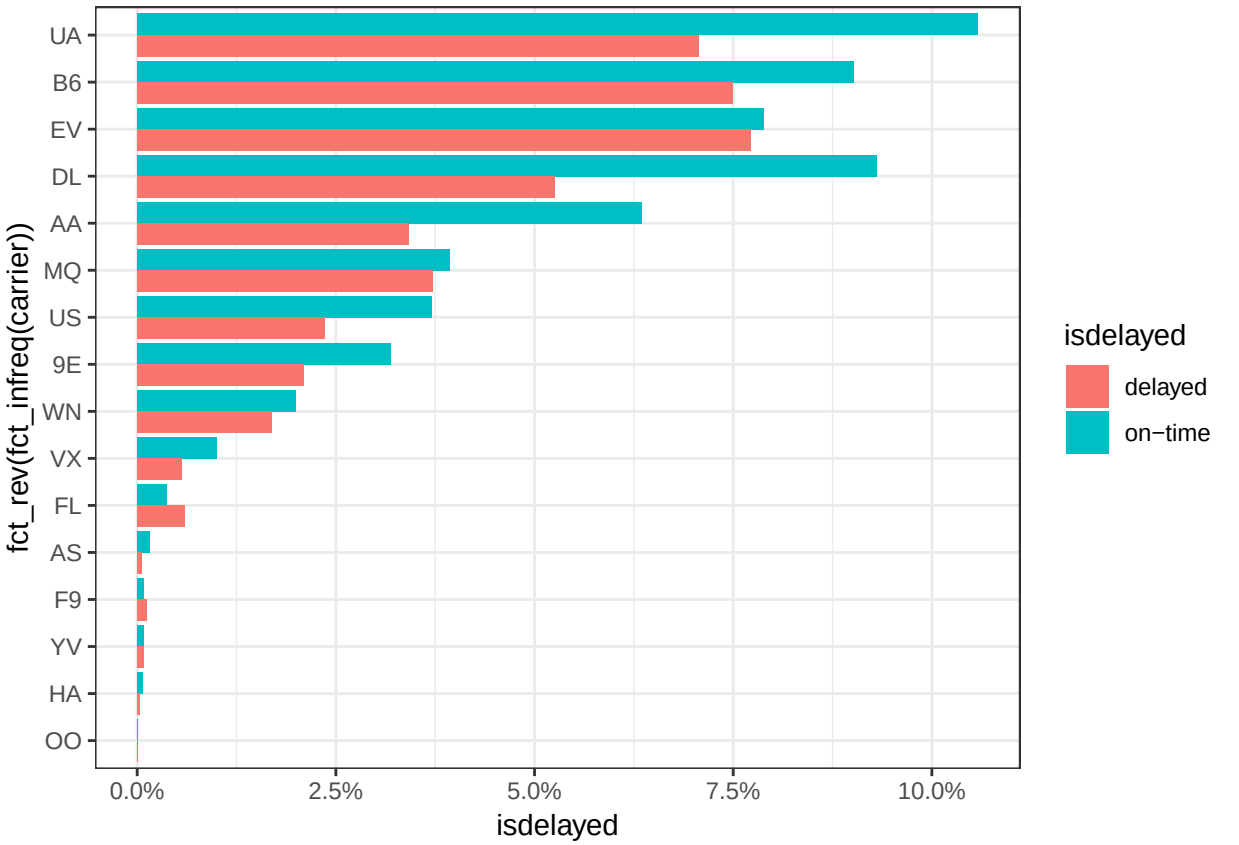
As a straightforward start, distribution of funds to the different fields can not be objective. All of the issues such as education, gender, poverty, healthcare are a priority for some groups in society and data alone can not prioritize them. One way would be looking at the weakest field in the society according to the data we have however we do not know that investing in that area would yield highest positive impact.

Nonetheless,if I had all the data I would first look for its structure and ask questions like what kind of data we have, to what extend it is reliable, is there any discrepancy at it. Then from that data, I would look for the field that would bring the most sustainable benefit to the society in the long term. Short term gains may bring more benefits and could be more beneficial if we are a politican but in this case I suppose we are not. My title would be the second one is this case.

3. If you had to plot a single graph using the flights data what would it be? Why? Make your argument, actually code the plot and provide the output. (You can find detailed info about the movies data set in its help file. Use ?flights, after you load nycflights13 package.)

```
flights <- flights %>% mutate(isdelayed=ifelse(arr_delay >= 0 ,"delayed","on-time")) %>%
  drop_na(isdelayed)

ggplot(flights,aes(x=fct_rev(fct_infreq(carrier)),y=isdelayed,fill=isdelayed))+
```

```
    geom_bar(aes(y = (..count..)/sum(..count..)),stat="count",position="dodge")+
    theme_bw()+
    scale_y_continuous(labels = scales::percent)+
    coord_flip()
```



I selected this because I would like to see which airlines delays most for the flights originating from New York so that I can select the airline to book my ticket. From this graph some airlines delays at almost half of their flights such as the EV. I dropped around 9000 values which have no information about delay times and it is a small percentage of total dataset(approximately %3) thus hereby we can see airlines performance.

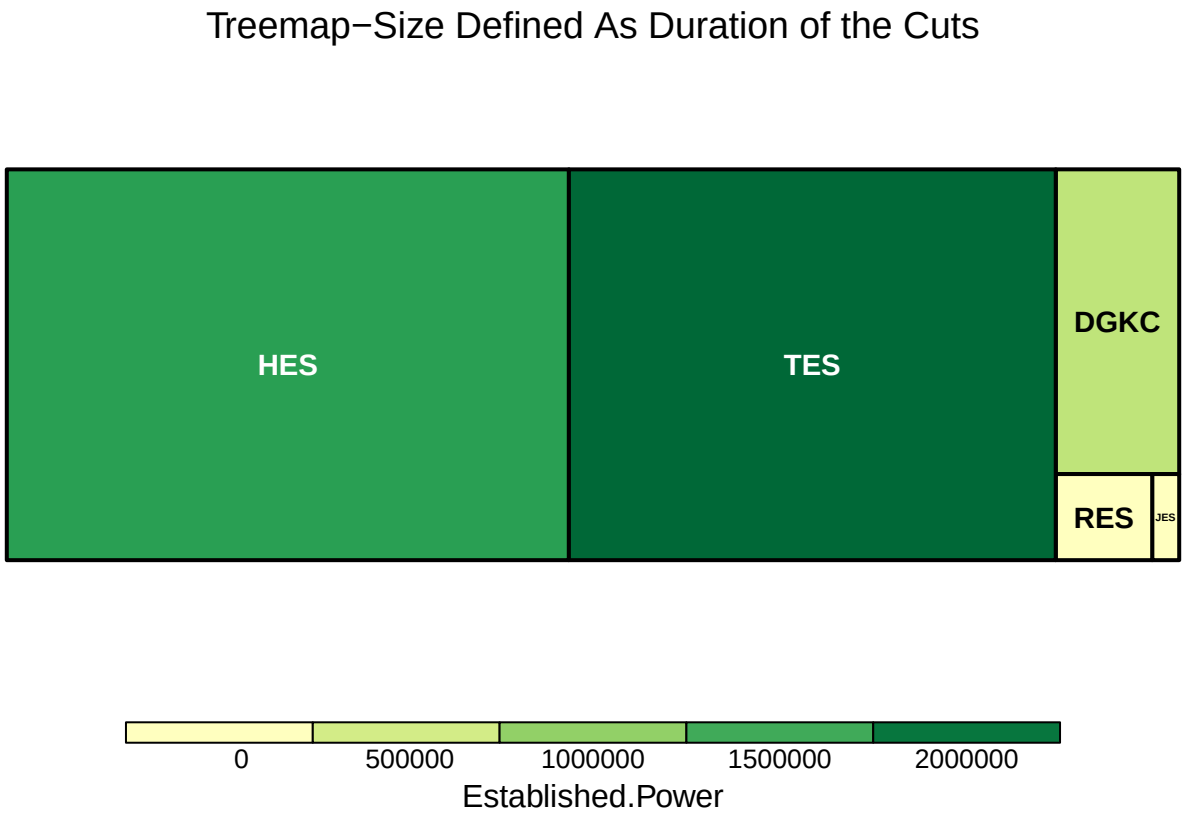
Extending Your Group Project

Wrapping-up My Group Project

First part of my single analysis would be adding a visualization that will summarize the most points from the dataset. I think a treemap would be a good way to do it. Below you can see that most of the power is at the TES' even though duration of the cut was significantly higher at the HES's. In order to get a visualization of this heat map, I omitted Other plant types which are the ones we could not classify.

```
cutswithoutother <- cuts %>% filter(Plant.Type != "Other")
cutswithoutother$Duration <- as.numeric(cutswithoutother$Duration)
treemap(cutswithoutother,
        index="Plant.Type",
        vSize="Duration",
```

```
        vColor="Established.Power",
        type="value",title="Treemap-Size Defined As Duration of the Cuts",aspRatio=3)
```



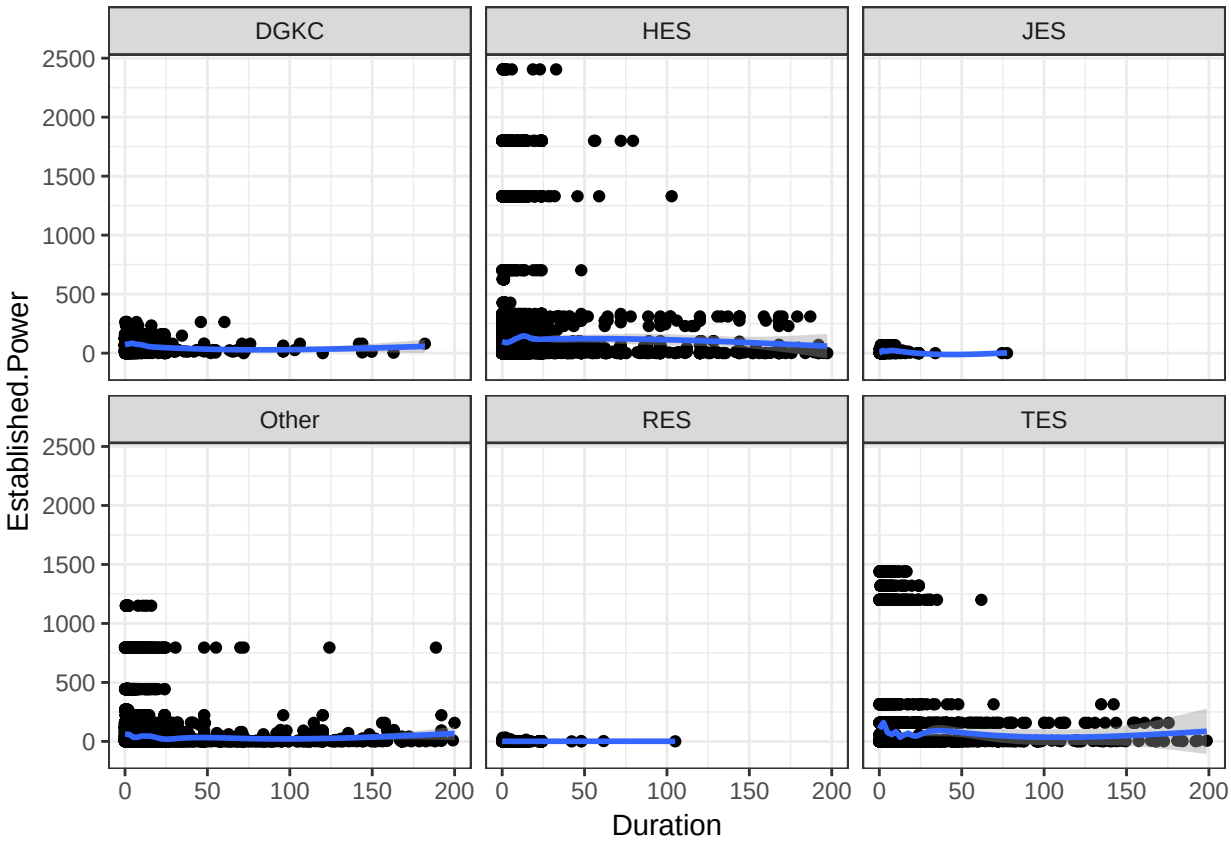
As a next step, I would like to check relationship between duration of the cut and established power to see whether cuts tend to be longer at plants with larger capacity. To check this I will employ linear regression and see the results. Then I will plot it with a scatterplot for each plant type to check if there is a relationship for a specific plant type.

```
cuts$Duration <- as.numeric(cuts$Duration)

linearreg <-  lm(cuts$Duration ~ cuts$Established.Power)
summary(linearreg)
```

```
##
## Call:
## lm(formula = cuts$Duration ~ cuts$Established.Power)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -11.9   -8.1   -7.8   -4.5 8045.7
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              9.119418   0.297070  30.698   <2e-16 ***
## cuts$Established.Power -0.002886   0.001189  -2.428   0.0152 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.36 on 73311 degrees of freedom
## Multiple R-squared:  8.044e-05,  Adjusted R-squared:  6.68e-05
## F-statistic: 5.897 on 1 and 73311 DF,  p-value: 0.01517
```

```
ggplot(cuts,aes(Duration,Established.Power))+
  geom_point()+
  geom_smooth()+
  xlim(c(0,200))+
  facet_wrap(~Plant.Type)+
  theme_bw()
```



Result of this is straightforward. There is no relationship between duration of the cuts and established power of the plants. This is clear from the model and also from the plot which I looked in order to see any relationship for a plant type.

Welcome to Real Life

Dipping the toes to analyze Times Higher Education ratings

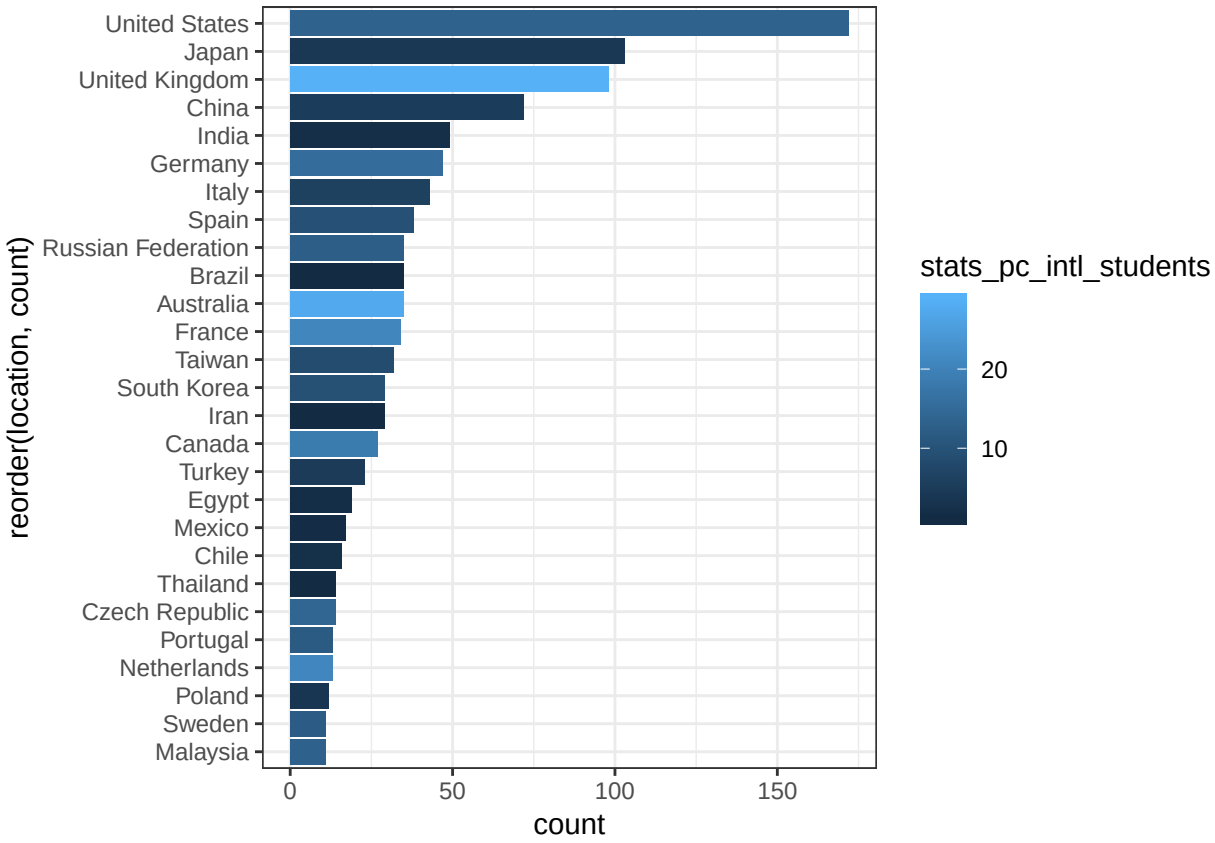
Times Higher Education rankings are one of the most prestigious rankings regarding universities performance on various matters. As a short analysis, I would like to look at varios things. First what top universities of the world taught, secondly which country is more internationalized, thirdly women's participation and lastly variables relationship between each other.

Here as the first thing I would like to see what subjects are thought at top universities in different countries by counting number of words at subjects offered. It is clear that engineering is leading with a significant margin in each country.

```
subjects <- ranking2019 %>% unnest_tokens(subjects, subjects_offered, token = "ngrams", n = 1)
subject_count <- subjects %>% group_by(subjects,location) %>% count(subjects,sort=TRUE) %>%
  na.omit()

#Removing irrelevant words
subject_count <- subject_count %>%
filter(subjects!="studies" & subjects!= "science" & subjects !="sciences")
#It looks like engineering leads by far.
head(subject_count,10)
```

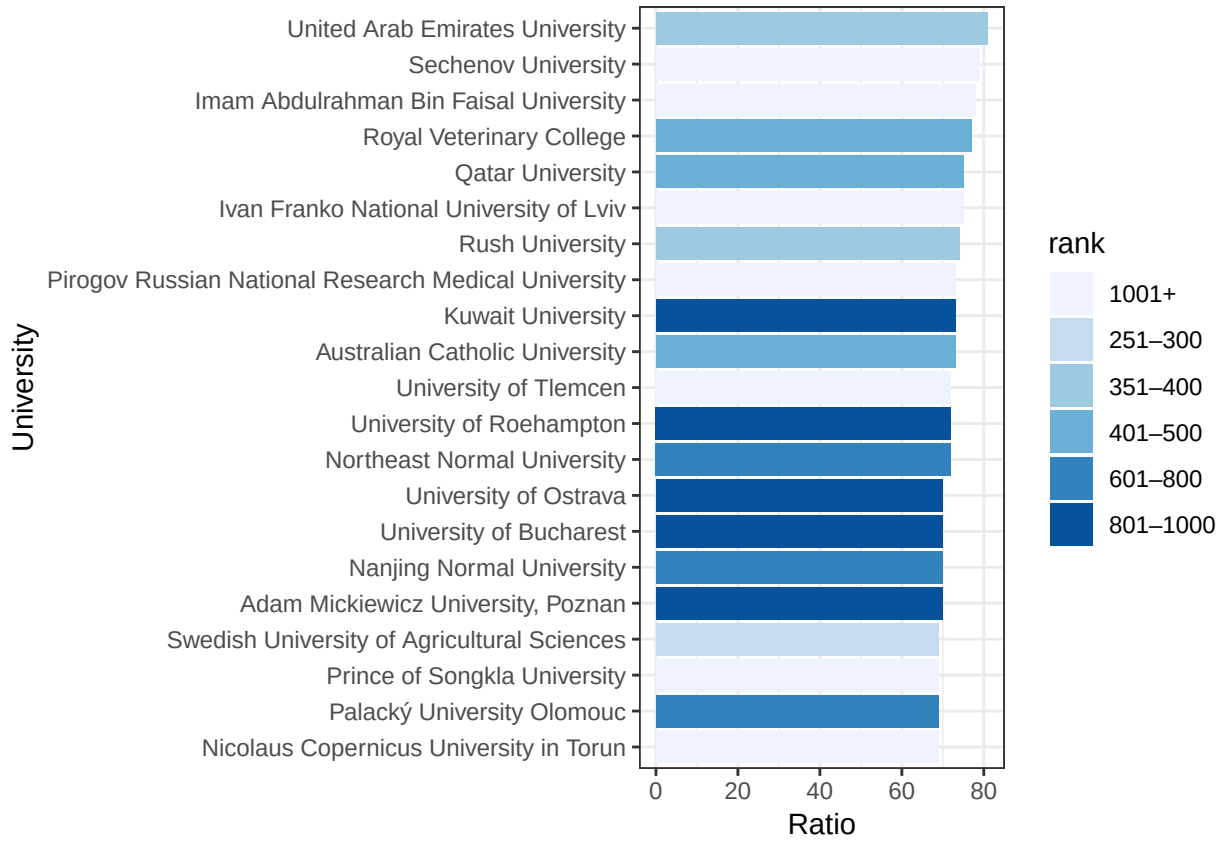
```
## # A tibble: 10 x 3
## # Groups:   subjects, location [10]
##    subjects    location           n
##    <chr>       <fct>          <int>
##  1 engineering United States    692
##  2 engineering Japan            349
##  3 engineering United Kingdom   331
##  4 engineering China            245
##  5 engineering India            177
##  6 biological  United States    169
##  7 chemistry   United States    167
##  8 mathematics United States    167
##  9 statistics  United States    167
## 10 astronomy   United States    166
```

```
ranking2019$stats_pc_intl_students <- as.numeric(sub("%", "", ranking2019$stats_pc_intl_students))

#For the sake of analysis let's see countries with at least ten universities.
#Drop NA is needed because some universities does not disclose percentage of international students.
ranking2019 <- ranking2019 %>% drop_na(stats_pc_intl_students)

ranking10 <- ranking2019 %>%
  group_by(location) %>%
  summarize(count=n(),stats_pc_intl_students=mean(stats_pc_intl_students))%>%
  filter(count > 10)


ggplot(ranking10,aes(reorder(location,count),count,fill=stats_pc_intl_students))+
  geom_col()+
  coord_flip()+
  theme_bw()
```



Secondly we see number of universities at each country and their internationalization level.USA is leading by far as number of universities in the rankings but not at internationalization which is an important pillar nowadays due to race to attract talent globally.Here Netherlands, the UK and Australia are top notch while some countries with high number of universities in the rankings such as India and Japan are not so international. Though we can not say that internationalization is the core parameter here to be in the list for a university. Now let's see how universities perform at gender perspective.

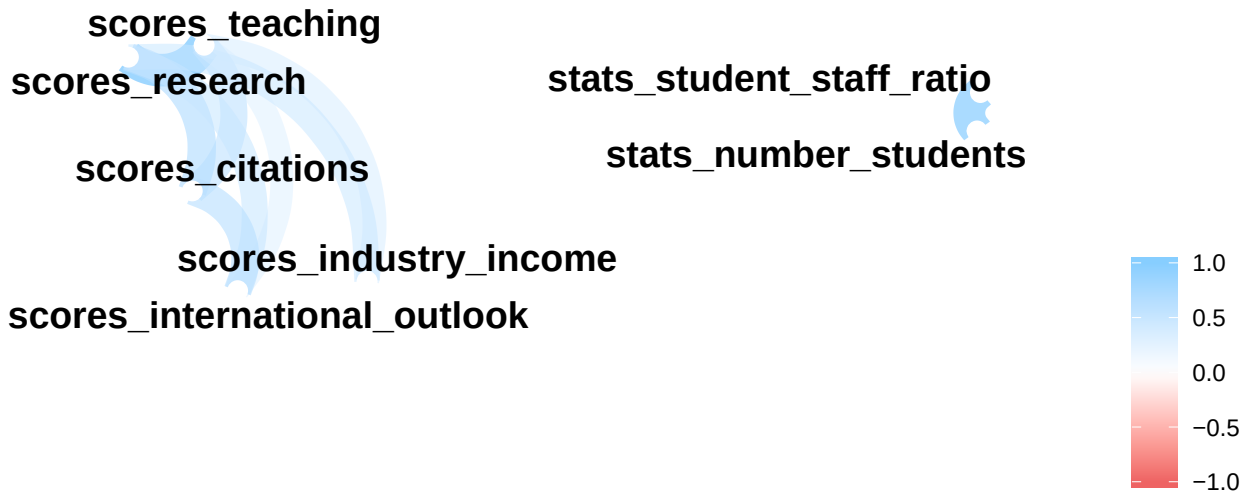
```
#Let's separate percentage of female ratio in universities.
ranking2019$female_ratio <- as.numeric(substr(ranking2019$stats_female_male_ratio,1,2))
#Now let's see which universities have over 50% females.
ranking2019topfemale <- ranking2019 %>% filter(female_ratio > 50) %>% top_n(20)
cbPalette <-c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

ggplot(ranking2019topfemale,aes(x=reorder(name,female_ratio),y=female_ratio,fill=rank))+
  geom_bar(stat="identity")+
  coord_flip()+
  theme_bw()+
  scale_fill_brewer()+
  ylab("Ratio")+
  xlab("University")
```



Universities with highest ratio of females points out interesting results. These are the countries at the Gulf region and Saudi Arabia. And none of those universities are ranked at the top 250 of the Times Higher Education which can be seen from label. With a short google search, I have seen that some of those are the universities where gender segregation takes place.

As last point of this anaylsis, I would like to visualize the relationship between different indicators at the data set.First we need to prepare the data.

```
# Turn them into numeric and prepare for plotting.
cols = c(6,8,10,12,14,22)
ranking2019[,cols] <- ranking2019[,cols] %>% lapply(function(x) as.numeric(as.character(x)))
ranking2019$stats_number_students <- as.numeric(gsub("," ,"", ranking2019$stats_number_students))
correlationdf <- ranking2019[,c(6,8,10,12,14,21,22)]

correlationdf$female_ratio <- ranking2019$female_ratio
# Create the dataset

correlationdf %>% correlate() %>% network_plot()
```



Network plot is a great tool to show correlation between different variables. Color and thickness of the connection implies direction and strength of the correlation. So here is my two cents to interpret this. Teaching and research scores are closely related to each other but not international outlook and industry income. Hence we can once again confirm that top universities first priority is research rather than equipping students with necessary tools for the job market. May be in the long term we can see a division of labour on reserach and job-readyness between universities and new generational educational establishments such as bootcamps, MOOC providers etc. Lastly, but not least, having a high women's participation has no correlation with any success indicators at the Times Higher Education ratings while internationalization has a relatively weak relationship. I think this also in a way confirms the output of the previous plots on internationalizaton and female ratio.