Assignment 1: EDA

GG Force Manousos Emmmanouil Theodosiou [6686311]
Philipp Schmitz-Justen [6870945] David Vinchansky
Koen Berendsen Asher van der Schelde

23-11-2020

```
library(ISLR)
library(tidyverse)
library(GGally)
library(gridExtra)
```

```
cwur <- read_csv("data/cwurData.csv")

shanghai <- read.csv("data/shanghaiData.csv")

times <-  read_csv("data/timesData.csv") %>%
  mutate(total_score = as.numeric(total_score)) %>%
  mutate(world_rank = as.numeric(world_rank)) %>%
  drop_na()



View(cwur)
View(times)
```

Explanation of the Data Set

For this assignment we have used the World University Ranking data set found on kaggle. (https://www.kaggle.com/mylesoneill/world-university-rankings?select=cwurData.csv)

The data set contains information about three World University Rankings: The Times Higher Education Ranking, the Shanghai Ranking, and the lesser known Center for World University Rankings. By comparing these different data sets, we can see if there are disambiguities in their rankings or if they lean towards specific measures or insitutions.

For simiplicity we generally focused on the CWUR rankings. The Center for World University Rankings (CWUR) organisation enlists more than 20,000 universities globally, and provides

information about: the ranking globally, ranking nationally, country, quality of education, employment of alumni, publications, influence, citations and of course the score for years 2012-2015. The data was generally well structured and contains only a few NA's in the broad_impact column.

Tables of Relevant Summary Statistics

Overview

Since most of the metrics in the CWUR table are rankings, a direct summary of them does not make a lot of sense, however we can take a look at the scores over the years:

```
cwur_scores <- cwur %>%
  group_by(year) %>%
  summarise(
    mean = mean(score),
    variance = var(score),
    min = min(score),
    max = max(score),
    med = median(score),
    size = n())

times_scores <- times %>%
  group_by(year) %>%
  summarise(
    mean = mean(total_score),
    variance = var(total_score),
    min = min(total_score),
    max = max(total_score),
    med = median(total_score),
    size = n())



cwur_scores
```

```
## # A tibble: 4 x 7
##    year  mean variance   min   max   med  size
##   <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl> <int>
## 1  2012  54.9    159.   43.4   100  50.2   100
## 2  2013  55.3    183.   44.3   100  49.7   100
## 3  2014  47.3     44.5  44.2   100  45.1   999
## 4  2015  46.9     42.4  44.0   100  44.8   999
```

```
times_scores
```

```
## # A tibble: 6 x 7
##    year  mean variance   min   max   med  size
##   <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl> <int>
## 1  2011  60.1     148.  46.2  96    56.4   177
## 2  2012  57.6     192.  41.4  94.8  53.6   179
## 3  2013  61.2     165.  46.2  95.5  56.8   179
## 4  2014  57.4     158.  44.3  94.9  52.4   179
## 5  2015  58.5     146.  45.6  94.3  53.9   182
## 6  2016  66.8     182.  48.8  95.2  63.5    99
```

When looking at the progression over the years we see that for CWUR the first two years were much smaller in terms of sample size than the later two years, explaining the difference in variance and mean scores. Meanwhile the metrics remained relatively constant for the times data set across the years.

We can also look at the individual years by country as in the following tables:

Year 2012

```
by_country <-
  filter(cwur, year == 2012) %>%
  group_by(country) %>%
  select(institution, country, score)

summarise(by_country, mean = mean(score),
    variance = var(score),
    min = min(score),
    max = max(score),
    med = median(score),
    size = n()) %>%
  arrange(desc(mean))
```

```
## # A tibble: 16 x 7
##    country         mean  variance   min   max   med  size
##    <chr>          <dbl>     <dbl> <dbl> <dbl> <dbl> <int>
##  1 United Kingdom  58.0 289.       43.8  86.2  51.8     8
##  2 USA             57.6 186.       43.9 100    51.9    58
##  3 Japan           56.5 116.       45.8  69.5  54.4     5
##  4 Israel          56.2  64.5      48.8  65.1  55.4     4
##  5 Switzerland     51.7 102.       44.5  66.7  47.8     4
```

```
##  6 Canada          51.0   8.52     47.7  53.4  51.7     3
##  7 France          47.9   7.20     43.4  50.4  48.3     5
##  8 Sweden          47.6  NA        47.6  47.6  47.6     1
##  9 South Korea     46.7  NA        46.7  46.7  46.7     1
## 10 Italy           46.3  NA        46.3  46.3  46.3     1
## 11 Germany         45.0   0.304    44.3  45.3  45.2     3
## 12 Finland         44.4  NA        44.4  44.4  44.4     1
## 13 Netherlands     44.3   1.38     43.5  45.1  44.3     2
## 14 Norway          44.3  NA        44.3  44.3  44.3     1
## 15 Australia       44.2   0.00125  44.1  44.2  44.2     2
## 16 Denmark         44.2  NA        44.2  44.2  44.2     1
```

Year 2013

```
## # A tibble: 18 x 7
##    country         mean variance   min   max   med  size
##    <chr>          <dbl>    <dbl> <dbl> <dbl> <dbl> <int>
##  1 United Kingdom  62.8 397.      46.5  92.5  56       7
##  2 USA             57.8 215.      44.3 100    51.1    57
##  3 Japan           55.8 188.      44.5  76.2  49.9     6
##  4 Israel          52.6  33.0     47.4  60.0  51.5     4
##  5 Switzerland     51.5  82.3     45.7  65.0  47.7     4
##  6 South Korea     51.3  NA       51.3  51.3  51.3     1
##  7 Canada          50.0  23.5     44.5  56.1  49.6     4
##  8 Sweden          48.0  NA       48.0  48.0  48.0     1
##  9 France          47.9  10.2     44.4  51.7  47.7     5
## 10 Italy           47.8  NA       47.8  47.8  47.8     1
## 11 Denmark         47.1  NA       47.1  47.1  47.1     1
## 12 Germany         46.7   0.510   46.2  47.2  46.7     2
## 13 Norway          46.1  NA       46.1  46.1  46.1     1
## 14 Netherlands     45.7  NA       45.7  45.7  45.7     1
## 15 Singapore       45.2  NA       45.2  45.2  45.2     1
## 16 Russia          44.9  NA       44.9  44.9  44.9     1
## 17 Australia       44.6   0.0365  44.5  44.8  44.6     2
## 18 Finland         44.4  NA       44.4  44.4  44.4     1
```

Year 2014

```
## # A tibble: 59 x 7
##    country         mean variance   min   max   med  size
##    <chr>          <dbl>    <dbl> <dbl> <dbl> <dbl> <int>
##  1 Israel          52.1    68.5   44.6  66.8  50.9     7
##  2 Switzerland     51.7    69.1   44.6  72.2  49.2     9
##  3 Singapore       51.4     8.69  49.4  53.5  51.4     2
```

```
##  4 USA             50.6   119.    44.3 100    46.4   229
##  5 Russia          49.1    41.3   44.4  56.4  46.5     3
##  6 United Kingdom  48.4    89.9   44.4  97.6  45.9    64
##  7 Netherlands     48.4     6.13  44.8  52.4  49.5    13
##  8 Denmark         48.4    10.8   44.8  52.9  48.0     5
##  9 Sweden          48.3     8.11  44.7  53.6  48.3    11
## 10 Canada          47.3    16.3   44.3  60.9  45.6    32
## # ... with 49 more rows
```

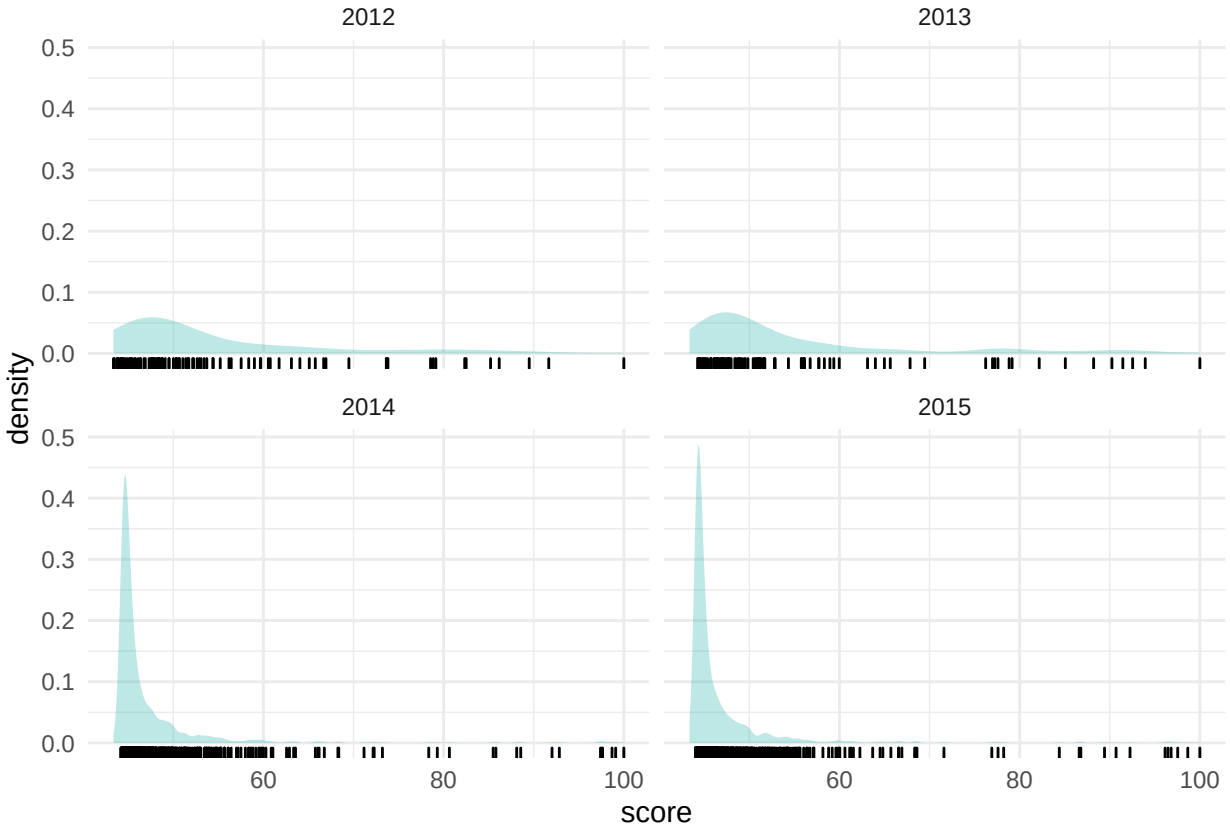
Year 2015

```
## # A tibble: 59 x 7
##    country         mean variance   min   max   med  size
##    <chr>          <dbl>    <dbl> <dbl> <dbl> <dbl> <int>
##  1 Singapore       51.4     8.61  49.3  53.4  51.4     2
##  2 Israel          51.2    62.1   44.4  65.7  49.1     7
##  3 Switzerland     50.4    45.8   44.3  66.9  48.5     9
##  4 USA             50.1   116.    44.1 100    46.0   229
##  5 Netherlands     48.2     6.48  44.4  51.8  49.3    13
##  6 Denmark         48.1    10.3   44.6  52.5  48.1     5
##  7 United Kingdom  48.0    87.7   44.1  96.8  45.5    65
##  8 Sweden          47.5     6.90  44.4  52.8  46.8    11
##  9 Russia          46.8    17.8   44.0  54.2  45.1     5
## 10 Canada          46.8    13.9   44.0  60.0  45.4    33
## # ... with 49 more rows
```

Again we see that there are no real outliers, probably due to the way the data was constructed, as very bad universities would not be included in the ranking at all. The variance varies across countries based on the sample size.

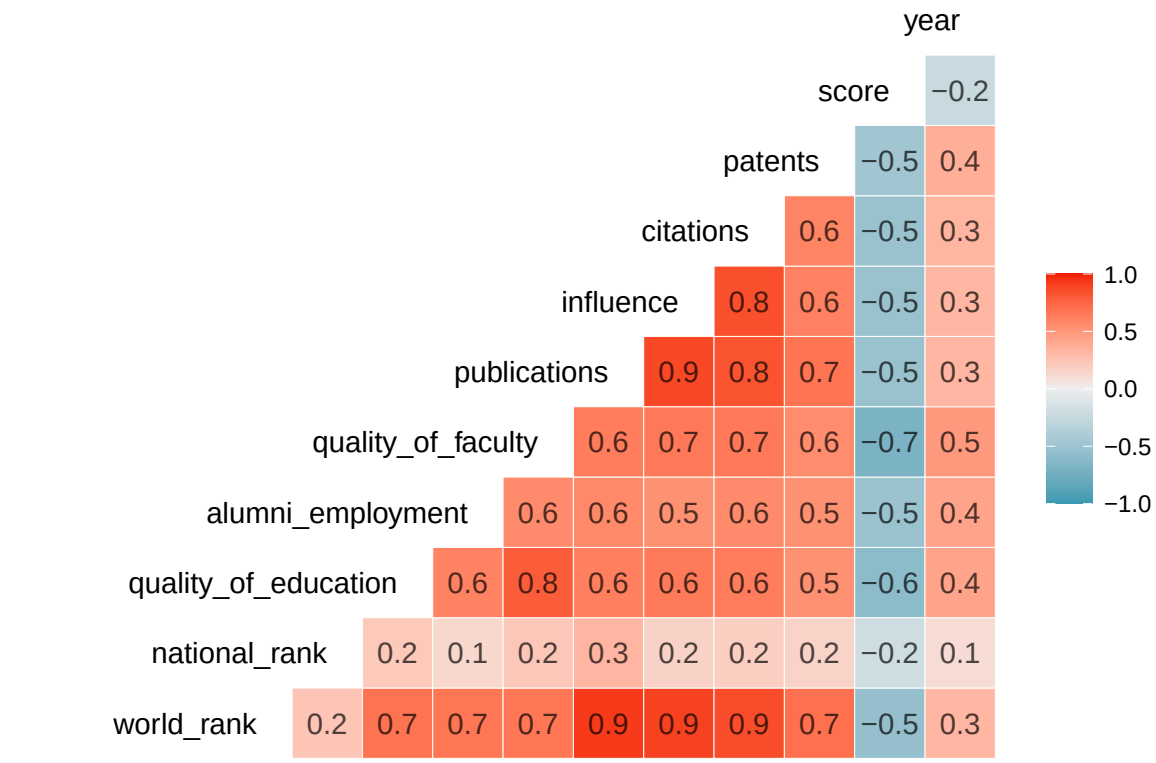
Descriptive Plots

```
ggplot(cwur, aes(x= score)) +
  geom_density(alpha = 0.3, fill = "light seagreen", color = NA) +
  geom_rug() +
  theme_minimal() +
  facet_wrap(vars(year))
```



In this chart we see that most of the universities have lower scores, especially with the larger samples 2014 and 2015.

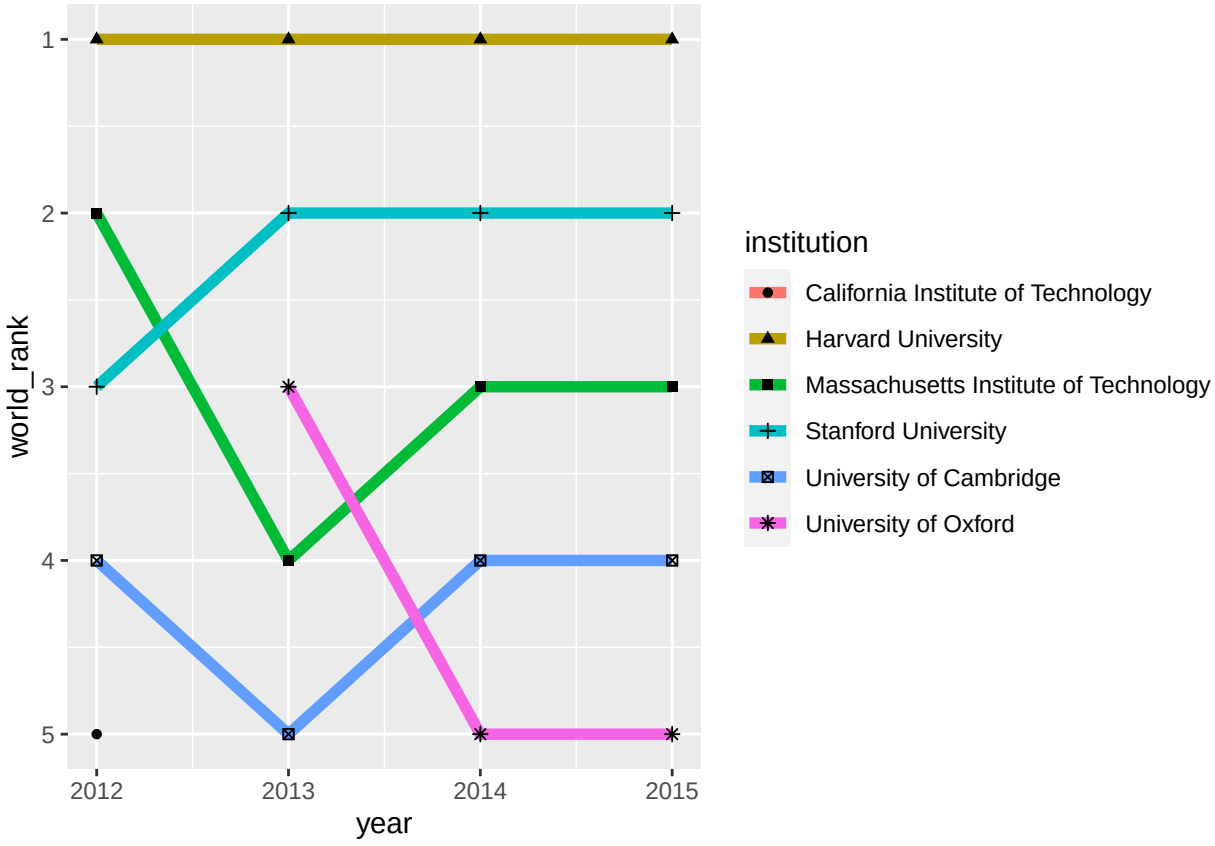
```
ggcorr(cwur[,-11], method = c("everything", "pearson"), label_alpha = 0.7, hjust =1, lab
  ggplot2::labs(title = "Correlation of various ranking aspects")
```



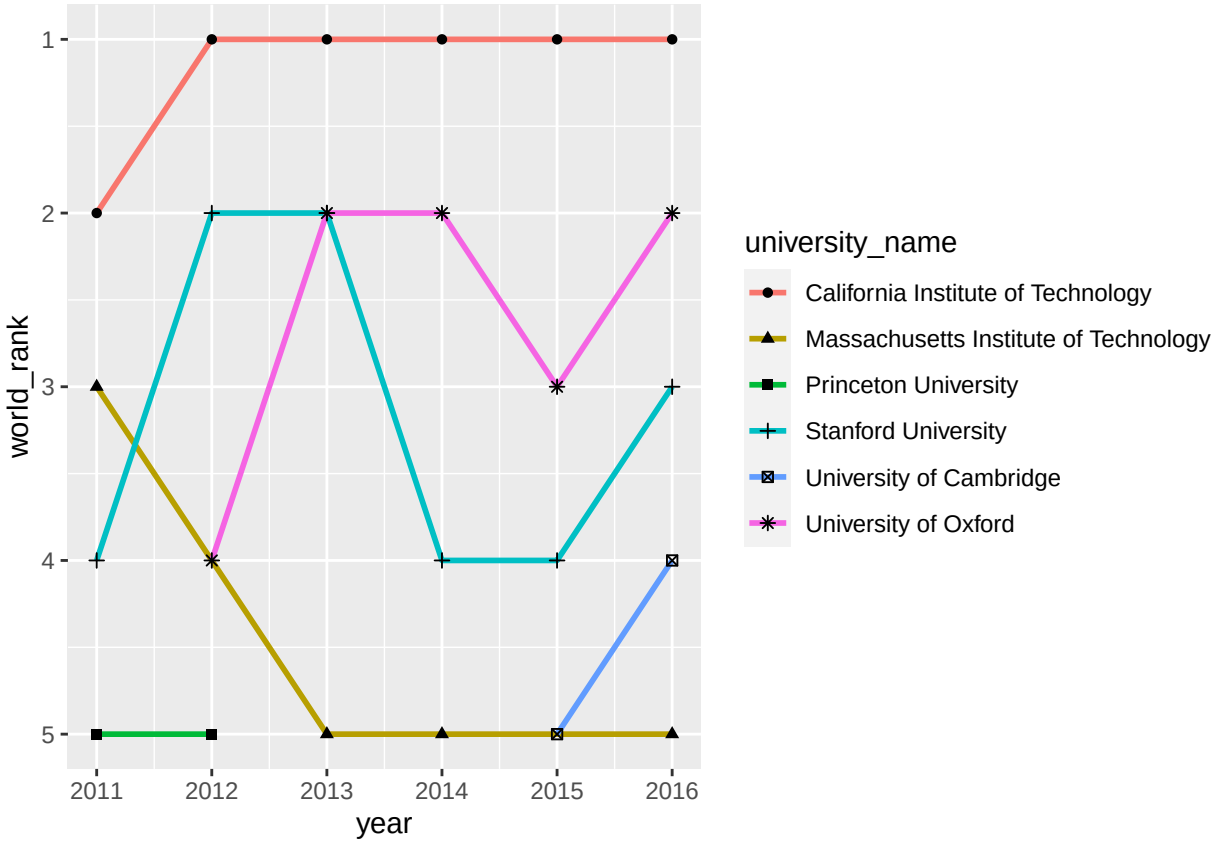
This correlation heat map shows us that certain factors are very strongly correlated such as publications, influence and world rank, while others such as national rank seem not to correlate strongly with the world_rank.

```
top_cwur <- cwur %>%
  filter( world_rank < 6) %>%
  arrange(world_rank)

top_times <- times %>%
  filter( world_rank < 6) %>%
  arrange(world_rank)
```

```
ggplot(top_cwur, map = aes(x = year, y = world_rank)) +
  geom_line(aes(colour= institution), size = 2) +
  geom_point(aes(shape = institution)) +
  scale_y_continuous(trans = "reverse")
```




```
ggplot(top_times, map = aes(x = year, y = world_rank)) +
  geom_line(aes(colour= university_name), size = 1) +
  geom_point(aes(shape = university_name)) +
  scale_y_continuous(trans = "reverse")
```

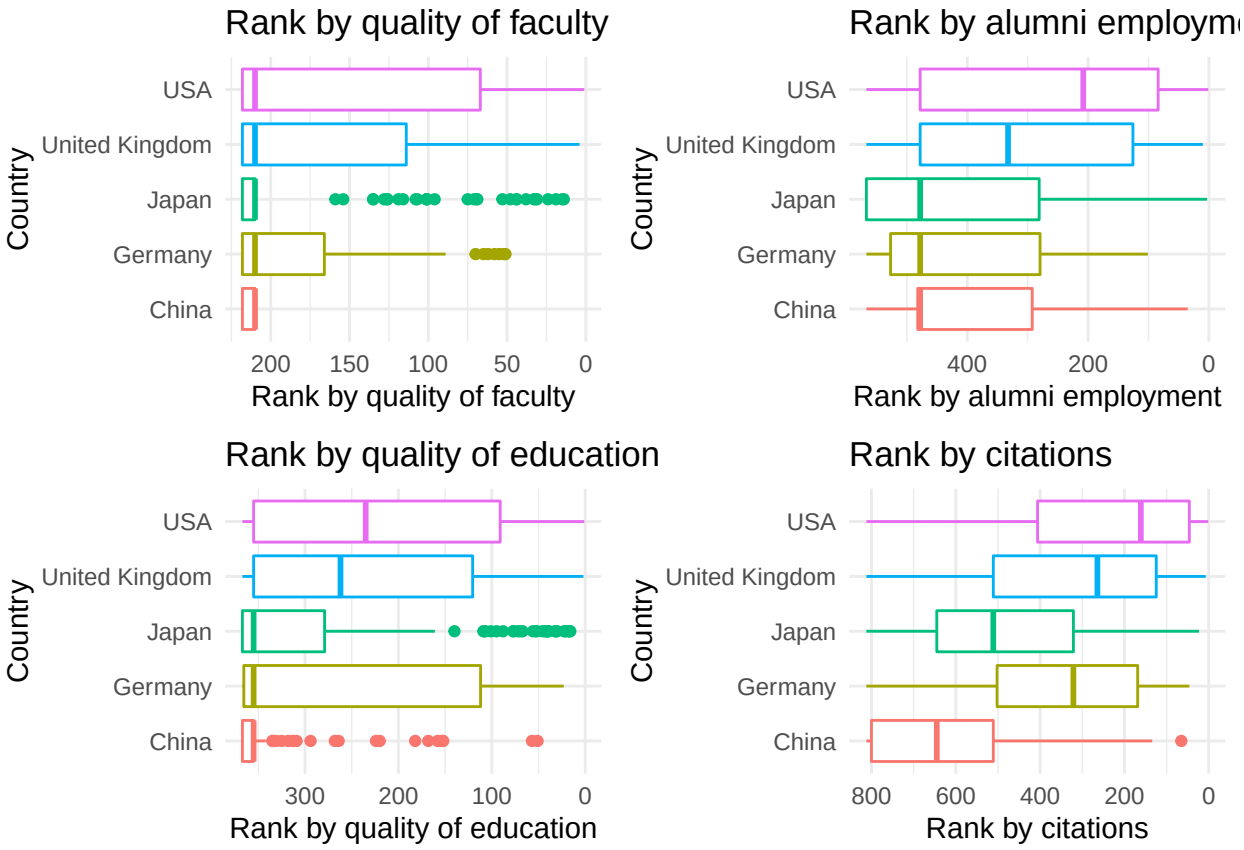


These two line charts nicely show the discrepancies between the top five schools in the two different rankings.

Next we wanted to visualise the quality of education/citation/faculty/alumni employment for the countries with the best university rankings and used boxplots for this. In order to fit them well on the grid we used coord_flip() as it was suggested in the lecture notes. Furthermore, to increase the data to ink ratio we have used the theme_minimal(). For the visualisation, it needs to be considered that the smaller the rank (closer to 0) the better.

```
cwur %>% group_by(country) %>% summarise(n = length(publications)) %>%
  top_n(5,n) %>% ungroup() -> d

cwur %>% filter(country %in% d$country) %>%
ggplot(aes(x=country, y=quality_of_faculty, col=country)) + guides(col=FALSE) +
  geom_boxplot() +  theme_minimal() + coord_flip() +
  scale_y_continuous(trans = "reverse") +
  labs(x="Country", y="Rank by quality of faculty",
      title="Rank by quality of faculty") -> plot_1

cwur %>% filter(country %in% d$country) %>%
ggplot(aes(x=country, y=alumni_employment, col=country)) + guides(col=FALSE) +
  geom_boxplot() +  theme_minimal() + coord_flip() +
```

```
  scale_y_continuous(trans = "reverse") +
  labs(x="Country", y="Rank by alumni employment",
      title="Rank by alumni employment") -> plot_2

cwur %>% filter(country %in% d$country) %>%
ggplot(aes(x=country, y=quality_of_education, col=country)) + guides(col=FALSE) +
  geom_boxplot() +  theme_minimal() + coord_flip() +
  scale_y_continuous(trans = "reverse") +
  labs(x="Country", y="Rank by quality of education",
      title="Rank by quality of education") -> plot_3

cwur %>% filter(country %in% d$country) %>%
ggplot(aes(x=country, y=citations, col=country)) + guides(col=FALSE) +
  geom_boxplot() +  theme_minimal() + coord_flip() +
  scale_y_continuous(trans = "reverse") +
  labs(x="Country", y="Rank by citations",
      title="Rank by citations") -> plot_4
grid.arrange(plot_1, plot_2, plot_3, plot_4, ncol=2)
```



This graph clearly shows us that while most countries are equal in quality of faculty, they diverge on citations and alumni employment, where the US clearly leads the pack.

Finally we would like to look at comparing the CWUR and times data set in one visualisation and will proceed as follows:

We would like to work with only a select few columns such as name, county, world ranking, national ranking (when exists), total score and year. Thus we create new data frames, whilst renaming the column names to homogenous names. Furthermore, we add a column to reflect which data set it is from and filter based on common years (2012-2015) only.

```
cwur <- cwur %>%
  select(world_rank, institution, country, national_rank, score, year) %>%
  rename( university_name = institution, total_score = score) %>%
  filter(year %in% (2012:2015)) %>%
  add_column(publication = "cwur")

shanghai <- shanghai %>%
  select(world_rank, university_name, national_rank, total_score, year) %>%
  filter(year %in% (2012:2015)) %>%
  add_column(publication = "shanghai")

times <- times %>%
  select(world_rank, university_name, country, total_score, year) %>%
  filter(year %in% (2012:2015)) %>%
  add_column(publication = "times")
```

We now assume that each university name is stated the same way in all three of our data frames, we do this in order to add ('mutate') the columns which are missing to make all three data frames the same size (identical) by performing an 'inner join'. Note: filter to remove N/A because our assumption is not bullet-proof.

```
shanghai <- shanghai %>%
  mutate(university_name = cwur$university_name[match(shanghai$university_name, cwur$uni
  mutate(country = cwur$country[match(shanghai$university_name, cwur$university_name)])
  filter(!is.na(university_name), !is.na(country))

times <- times %>%
  mutate(university_name = cwur$university_name[match(times$university_name, cwur$univer
  mutate(country = cwur$country[match(times$university_name, cwur$university_name)]) %>%
  add_column(national_rank = NA) %>% # because no such data exists
  filter(!is.na(university_name), !is.na(country))
```

Now that we are satisfied that our dataframe are of the same (column) size, let us combine them into one giant data frame which would help some of the data analysis/visualisation work we will do later on. Also round the 'total_score' to nearest decimal point.

```
ranking <- rbind(cwur, shanghai, times)
```

Check the data types in data frame.

```
lapply(ranking, class)
```

```
## $world_rank
## [1] "character"
##
## $university_name
## [1] "character"
##
## $country
## [1] "character"
##
## $national_rank
## [1] "character"
##
## $total_score
## [1] "numeric"
##
## $year
## [1] "numeric"
##
## $publication
## [1] "character"
```

Convert 'world_rank', 'national_rank' and 'total_score' to numerical values.

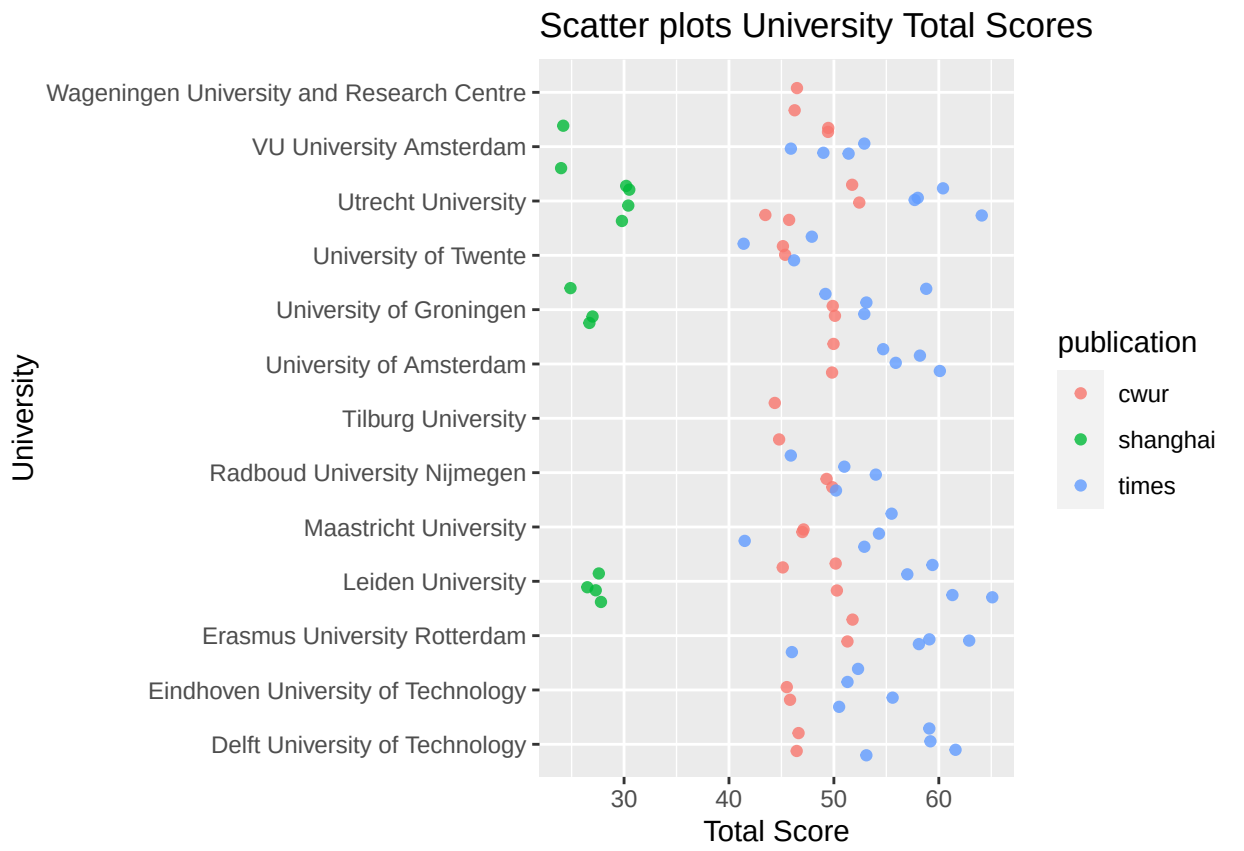
```
ranking <- ranking %>%
  mutate(world_rank=as.numeric(world_rank)) %>%
  mutate(national_rank=as.numeric(national_rank)) %>%
  mutate(total_score=as.numeric(total_score))

lapply(ranking,class)
```

```
## $world_rank
## [1] "numeric"
##
## $university_name
## [1] "character"
##
```

```
## $country
## [1] "character"
##
## $national_rank
## [1] "numeric"
##
## $total_score
## [1] "numeric"
##
## $year
## [1] "numeric"
##
## $publication
## [1] "character"
```

Plot the scores for universities in the Netherlands and rotate the x-axis.

```
ggplot(subset(ranking, country == "Netherlands"),
       aes(x = university_name, y = total_score, colour = publication)) +
  geom_point(alpha = 0.8,  position = position_jitter()) +
  coord_flip() +
  labs(title = "Scatter plots University Total Scores", x = "University", y = "Total Sco
```



This visualisation nicely shows the discrepancies in the total score for different rankings. Shanghai clearly consistly ranks Dutch universities lower than both CWUR and times, which seem to often overlap, with times tendentially giving higher scores.

Explanation

The process for looking into the indidivudal data sets was pretty straight forward, but in order to properly compare the rankings it was important to determine what data we work with. We therefore created new data frames and only included the relevant columns for the name of university, world ranking, national ranking, country, total score and year. However, because the data frames were of different sizes and had different column headings, we renamed certain columns, changed the data types where needed and matched the missing fields from other data frames by matching using the university name. Going forward it would make sense to base the tables and visualisations in the first half of the document on the refined data set used in the later half, however this did not happen due to time constraints and division of labor.

For this EDA Assignment we all took slightly different aprproaches which explains the plethora of different tables and visualisations in this document. Since most of the explanations were placed in the document it suffices to say here that generally there were no real outliers within the individual data sets, but that there was varation between the two rankings as well as between countries in the same ranking.