Assignment Exploratory Data Analysis

Manousos Emmmanouil Theodosiou [6686311], Group: GG Force

19-11-2020

Import Libraries

```
library(ISLR)
library(tidyverse)
library(haven)
library(gridExtra)
```

Import dataset from CWUR (Center for World University Rankings).

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

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   world_rank = col_double(),
##   institution = col_character(),
##   country = col_character(),
##   national_rank = col_double(),
##   quality_of_education = col_double(),
##   alumni_employment = col_double(),
##   quality_of_faculty = col_double(),
##   publications = col_double(),
##   influence = col_double(),
##   citations = col_double(),
##   broad_impact = col_double(),
##   patents = col_double(),
##   score = col_double(),
##   year = col_double()
## )
```

```
## Warning: 4 parsing failures.
##  row         col           expected actual                file
## 1010 institution delimiter or quote      A 'data/cwurData.csv'
## 1010 institution delimiter or quote      B 'data/cwurData.csv'
## 1829 institution delimiter or quote      A 'data/cwurData.csv'
```

```
## 1829 institution delimiter or quote      0 'data/cwurData.csv'
```

```
head(cwur_data)
```

```
## # A tibble: 6 x 14
##   world_rank institution country national_rank quality_of_educ~ alumni_employme~
##        <dbl> <chr>       <chr>           <dbl>            <dbl>             <dbl>
## 1          1 Harvard Un~ USA                 1                7                 9
## 2          2 Massachuse~ USA                 2                9                17
## 3          3 Stanford U~ USA                 3               17                11
## 4          4 University~ United~             1               10                24
## 5          5 California~ USA                 4                2                29
## 6          6 Princeton ~ USA                 5                8                14
## # ... with 8 more variables: quality_of_faculty <dbl>, publications <dbl>,
## #   influence <dbl>, citations <dbl>, broad_impact <dbl>, patents <dbl>,
## #   score <dbl>, year <dbl>
```

The dataset

The above dataset comes from the Center for World University Rankings (CWUR), which is an organisation that provides consulting to universities (in this particular case), to achieve higher goals in academia and research. The CWUR has its headquarters in the United Arab Emirates. The above dataset enlists universities out of 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.

Summary statistics

For the summary statistics of the above dataset, I will compare the final score given to universities per country and I will calculate measures of spread (variance), measures of location (mean, median) and max-min score of each country for the years 2012-2015.

Year 2012

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

summarise(by_country, mean_score = mean(score),
          variance_score = var(score), min_score = min(score),
          max_score = max(score)) %>%
  arrange(desc(mean_score))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 16 x 5
##    country        mean_score variance_score min_score max_score
##    <chr>               <dbl>          <dbl>     <dbl>     <dbl>
##  1 United Kingdom       58.0      289.           43.8      86.2
##  2 USA                  57.6      186.           43.9     100
##  3 Japan                56.5      116.           45.8      69.5
##  4 Israel               56.2       64.5          48.8      65.1
##  5 Switzerland          51.7      102.           44.5      66.7
##  6 Canada               51.0        8.52         47.7      53.4
##  7 France               47.9        7.20         43.4      50.4
##  8 Sweden               47.6       NA            47.6      47.6
##  9 South Korea          46.7       NA            46.7      46.7
## 10 Italy                46.3       NA            46.3      46.3
## 11 Germany              45.0        0.304        44.3      45.3
## 12 Finland              44.4       NA            44.4      44.4
## 13 Netherlands          44.3        1.38         43.5      45.1
## 14 Norway               44.3       NA            44.3      44.3
## 15 Australia            44.2        0.00125      44.1      44.2
## 16 Denmark              44.2       NA            44.2      44.2
```

Year 2013

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 18 x 5
##    country        mean_score standard_deviation_score min_score max_score
##    <chr>               <dbl>                    <dbl>     <dbl>     <dbl>
##  1 United Kingdom       62.8                   19.9        46.5      92.5
##  2 USA                  57.8                   14.7        44.3     100
##  3 Japan                55.8                   13.7        44.5      76.2
##  4 Israel               52.6                    5.75       47.4      60.0
##  5 Switzerland          51.5                    9.07       45.7      65.0
##  6 South Korea          51.3                   NA          51.3      51.3
##  7 Canada               50.0                    4.84       44.5      56.1
##  8 Sweden               48.0                   NA          48.0      48.0
##  9 France               47.9                    3.19       44.4      51.7
## 10 Italy                47.8                   NA          47.8      47.8
## 11 Denmark              47.1                   NA          47.1      47.1
## 12 Germany              46.7                    0.714      46.2      47.2
## 13 Norway               46.1                   NA          46.1      46.1
## 14 Netherlands          45.7                   NA          45.7      45.7
## 15 Singapore            45.2                   NA          45.2      45.2
## 16 Russia               44.9                   NA          44.9      44.9
## 17 Australia            44.6                    0.191      44.5      44.8
## 18 Finland              44.4                   NA          44.4      44.4
```

Year 2014

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 59 x 5
##    country        mean_score standard_deviation_score min_score max_score
##    <chr>               <dbl>                    <dbl>     <dbl>     <dbl>
##  1 Israel               52.1                     8.28      44.6      66.8
##  2 Switzerland          51.7                     8.31      44.6      72.2
##  3 Singapore            51.4                     2.95      49.4      53.5
##  4 USA                  50.6                    10.9       44.3     100
##  5 Russia               49.1                     6.43      44.4      56.4
##  6 United Kingdom       48.4                     9.48      44.4      97.6
##  7 Netherlands          48.4                     2.47      44.8      52.4
##  8 Denmark              48.4                     3.29      44.8      52.9
##  9 Sweden               48.3                     2.85      44.7      53.6
## 10 Canada               47.3                     4.03      44.3      60.9
## # ... with 49 more rows
```

Year 2015

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 59 x 5
##    country        mean_score standard_deviation_score min_score max_score
##    <chr>               <dbl>                    <dbl>     <dbl>     <dbl>
##  1 Singapore            51.4                     2.93      49.3      53.4
##  2 Israel               51.2                     7.88      44.4      65.7
##  3 Switzerland          50.4                     6.77      44.3      66.9
##  4 USA                  50.1                    10.8       44.1     100
##  5 Netherlands          48.2                     2.54      44.4      51.8
##  6 Denmark              48.1                     3.21      44.6      52.5
##  7 United Kingdom       48.0                     9.37      44.1      96.8
##  8 Sweden               47.5                     2.63      44.4      52.8
##  9 Russia               46.8                     4.22      44.0      54.2
## 10 Canada               46.8                     3.73      44.0      60.0
## # ... with 49 more rows
```

Descriptive plots

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

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

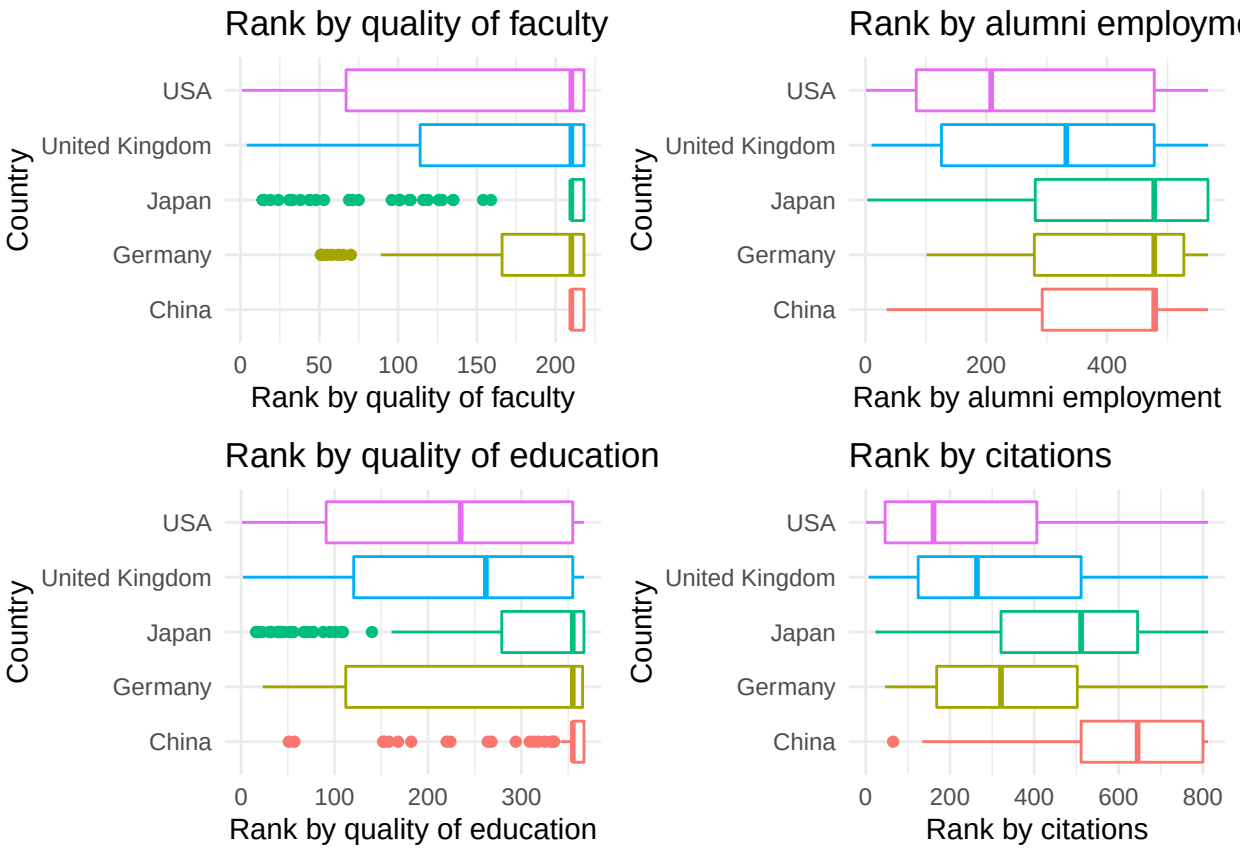
```
cwur_data %>% filter(country %in% d$country) %>%
ggplot(aes(x=country, y=quality_of_faculty, col=country)) + guides(col=FALSE) +
```

```
  geom_boxplot() +  theme_minimal() + coord_flip() +
  labs(x="Country", y="Rank by quality of faculty",
      title="Rank by quality of faculty") -> plot_1

cwur_data %>% filter(country %in% d$country) %>%
ggplot(aes(x=country, y=alumni_employment, col=country)) + guides(col=FALSE) +
  geom_boxplot() +  theme_minimal() + coord_flip() +
  labs(x="Country", y="Rank by alumni employment",
      title="Rank by alumni employment") -> plot_2

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

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



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