Initial Analysis

Akshi Chaudhary

2017-12-11

1. Load the required packages

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.4.2
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
## Warning: package 'tibble' was built under R version 3.4.2
```

```
## Warning: package 'tidyr' was built under R version 3.4.2
```

```
## Warning: package 'readr' was built under R version 3.4.2
```

```
## Warning: package 'purrr' was built under R version 3.4.2
```

```
library(ggplot2)
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.4.2
```

2. Downloading data from a url

```
times <- read.csv(url('https://raw.githubusercontent.com/akshi8/University_rankings/v1.1/data/external/t
expenditure <- read.csv(url('https://raw.githubusercontent.com/akshi8/University_rankings/v1.1/data/exte
```

3. Data overview

   University ranking data from Times Higher ranking:

```
head(times)
```

```
##   world_rank                    university_name
## 1          1                  Harvard University
## 2          2   California Institute of Technology
## 3          3 Massachusetts Institute of Technology
## 4          4                  Stanford University
## 5          5                 Princeton University
## 6          6               University of Cambridge
##                     country teaching international research citations
## 1 United States of America     99.7          72.4     98.7      98.8
## 2 United States of America     97.7          54.6     98.0      99.9
## 3 United States of America     97.8          82.3     91.4      99.9
## 4 United States of America     98.3          29.5     98.1      99.2
## 5 United States of America     90.9          70.3     95.4      99.9
## 6          United Kingdom     90.5          77.7     94.1      94.0
##   income total_score num_students student_staff_ratio
## 1   34.5        96.1       20,152                 8.9
## 2   83.7        96.0        2,243                 6.9
## 3   87.5        95.6       11,074                 9.0
## 4   64.3        94.3       15,596                 7.8
```

```
## 5      -        94.2        7,929                 8.4
## 6   57.0        91.2       18,812                11.8
##   international_students female_male_ratio year
## 1                    25%                   2011
## 2                    27%          33 : 67 2011
## 3                    33%          37 : 63 2011
## 4                    22%          42 : 58 2011
## 5                    27%          45 : 55 2011
## 6                    34%          46 : 54 2011
```

Country-wise education expenditure data across public, private institutes by institute types over the years

```
head(expenditure)
```

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

4. Data cleaning, changing data formats and treating Null values

- We can see missing values for expenditures for many countries in some years, replacing numeric values with 0
- Also for the hypothesis testing we have to look at the average expenditure by countries in various education institutions

```
colnames(expenditure)[4] <- "y1995"
colnames(expenditure)[5] <- "y2000"
colnames(expenditure)[6] <- "y2005"
colnames(expenditure)[7] <- "y2009"
colnames(expenditure)[8] <- "y2010"
colnames(expenditure)[9] <- "y2011"
```

Replacing Null values with 0

```
expenditure <- expenditure %>% mutate(y1995 = ifelse(is.na(y1995),0,y1995)
                                      ,y2000 = ifelse(is.na(y2000),0,y2000)
                                      ,y2005 = ifelse(is.na(y2005),0,y2005)
                                      ,y2009 = ifelse(is.na(y2009),0,y2009)
                                      ,y2010 = ifelse(is.na(y2010),0,y2010)
                                      ,y2011 = ifelse(is.na(y2011),0,y2011))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.2
```

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

```
times$total_score <- as.numeric(times$total_score)
```

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

```
Name_mapping <- c("Ireland", "Korea, Republic of", "United States" )

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

5. Summarizing data based on the input for hypothesis testing and visualization

- For this the total expenditures through 1995-2011 have been averaged for each institution type

```
school_exp <- expenditure %>% filter(direct_expenditure_type !='Total') %>%
  mutate(avg_exp = (y1995+ y2000 +y2005+y2009+y2010+y2011)/6) %>% group_by(country,direct_expenditure_ty
    summarise(total_exp = round(sum(avg_exp),2)) %>% arrange(desc(total_exp))
head(school_exp)
```

```
## # A tibble: 6 x 4
## # Groups:   country, direct_expenditure_type [6]
##   country direct_expenditure_type   institute_type total_exp
##    <fctr>                  <fctr>           <fctr>     <dbl>
## 1 Denmark                 Public All Institutions      7.05
## 2  Norway                 Public All Institutions      6.53
## 3 Iceland                 Public All Institutions      6.43
## 4  Sweden                 Public All Institutions      6.37
## 5 Finland                 Public All Institutions      6.17
## 6 Belgium                 Public All Institutions      5.85
```

- Taking the score of the best ranking institute of each country using Times ranking data
- We are assuming this to be the proxy for ranking the higher education system for each country

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

- Arranging best_scores for each country from highest to lowest

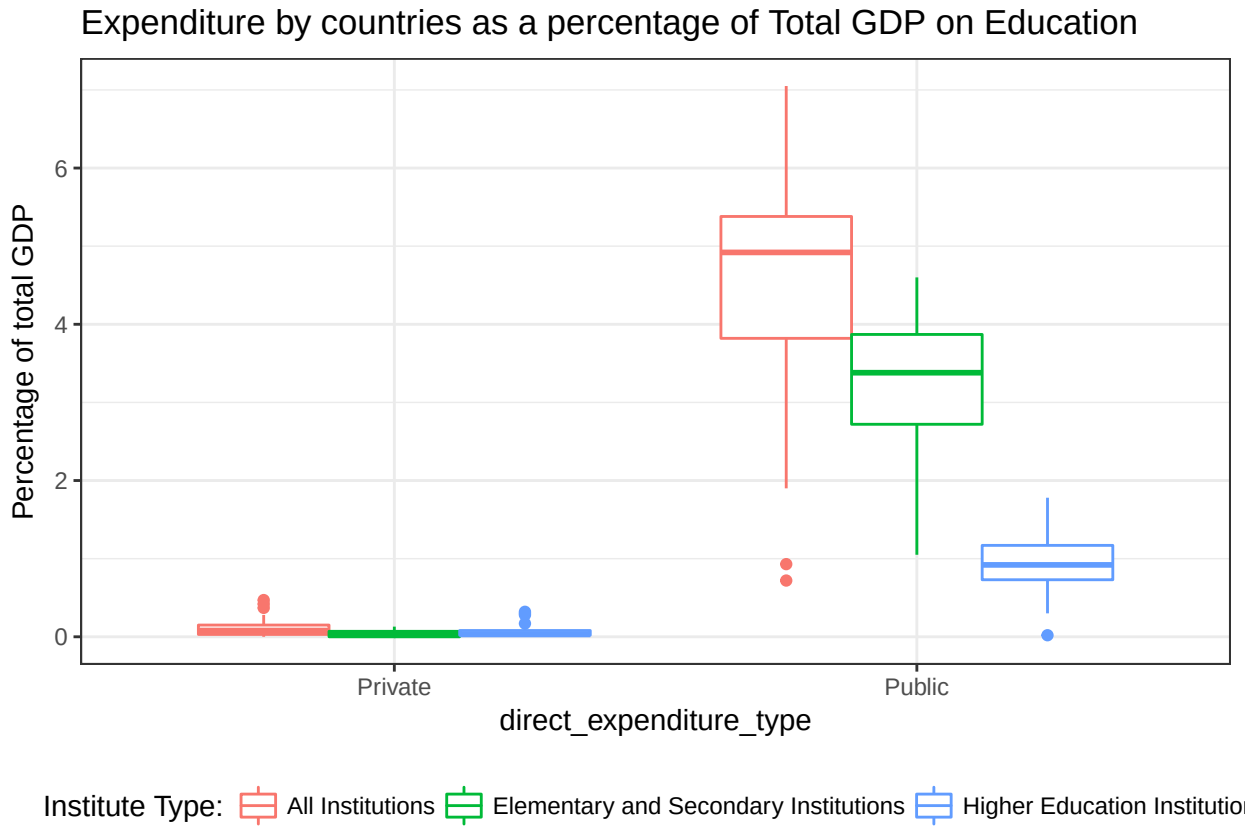
```
country_score$country <- factor(country_score$country , levels = country_score$country [order(country_s

head(country_score)
```

```
## # A tibble: 6 x 2
##          country best_score
##           <fctr>      <dbl>
## 1  United States        415
## 2 United Kingdom        407
## 3    Switzerland        376
## 4         Canada        350
## 5      Hong Kong        324
## 6      Singapore        324
```

6. Overall expenditure trends across countries in various levels of educations

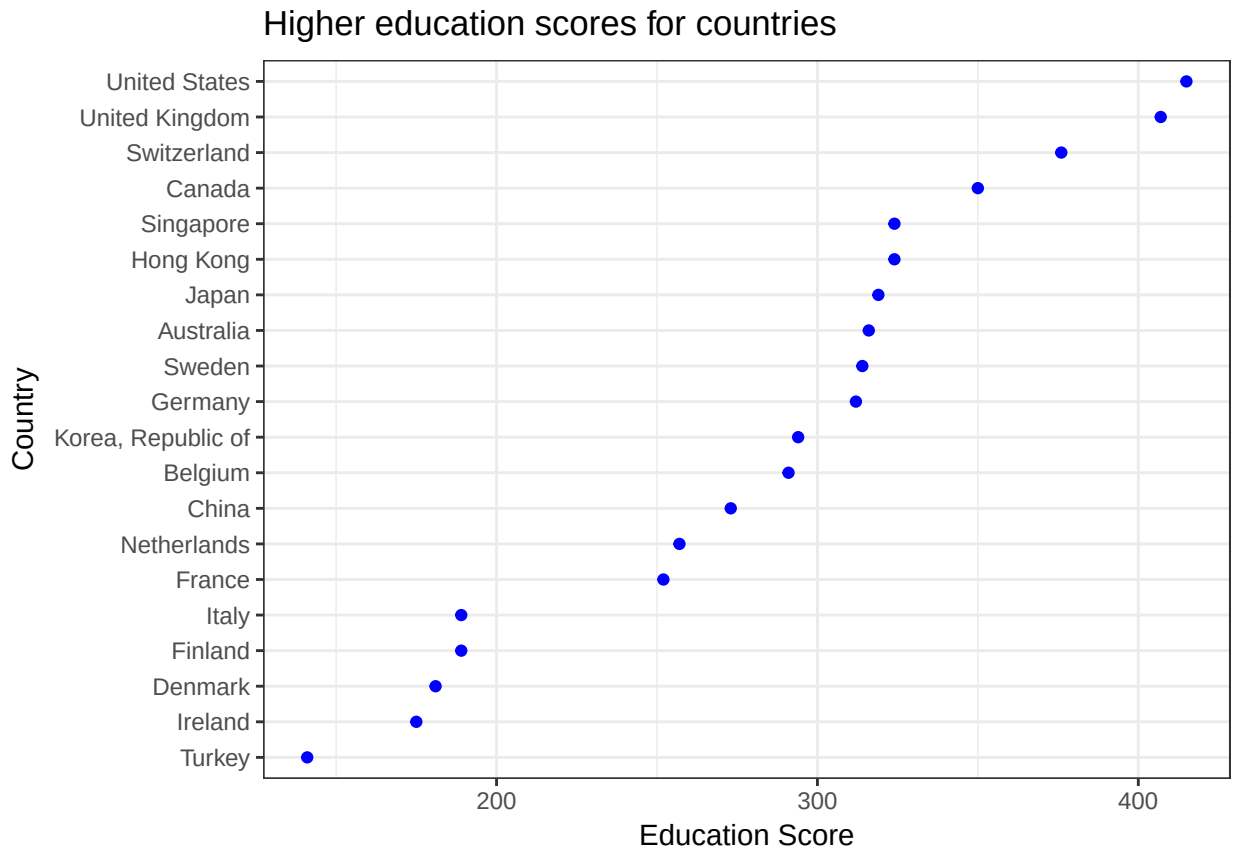
```
school_exp %>% ggplot(aes(direct_expenditure_type,total_exp )) + geom_boxplot(aes(color = institute_type
  theme(axis.text=element_text(size=8),axis.title=element_text(size=10,face="bold" )) + theme_bw() + the
```



- Let's see the plot of countries with top `higher` education system rankings

```
country_score %>% top_n(20) %>%
  arrange(desc(best_score)) %>%
  ggplot(aes(x = best_score, y = country)) + geom_point(color ='blue') +
  labs(title = 'Higher education scores for countries', x = 'Education Score', y ='Country') +   theme(a
  theme_bw()
```

```
## Selecting by best_score
```



7. Hypothesis Testing

- Do countries who spend more in public education system(as part of their GDP) rank higher in global higher education ranking?

Null Hypothesis H_0 : Percentage GDP expenditure on public institute has no association with education score Alternate Hypothesis H_A : Percentage GDP expenditure on public institute affects the education score of a country

- Combine score data with public expenditure data and filter for public education expenditure

```
df <- left_join(school_exp, country_score, by = "country") %>% filter(direct_expenditure_type == "Publi
```

- Many countries in the expenditure data don't have very high ranking institutes and therefore their best scores are missing, let's impute missing best score with 1 as that is the least score of times ranking

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

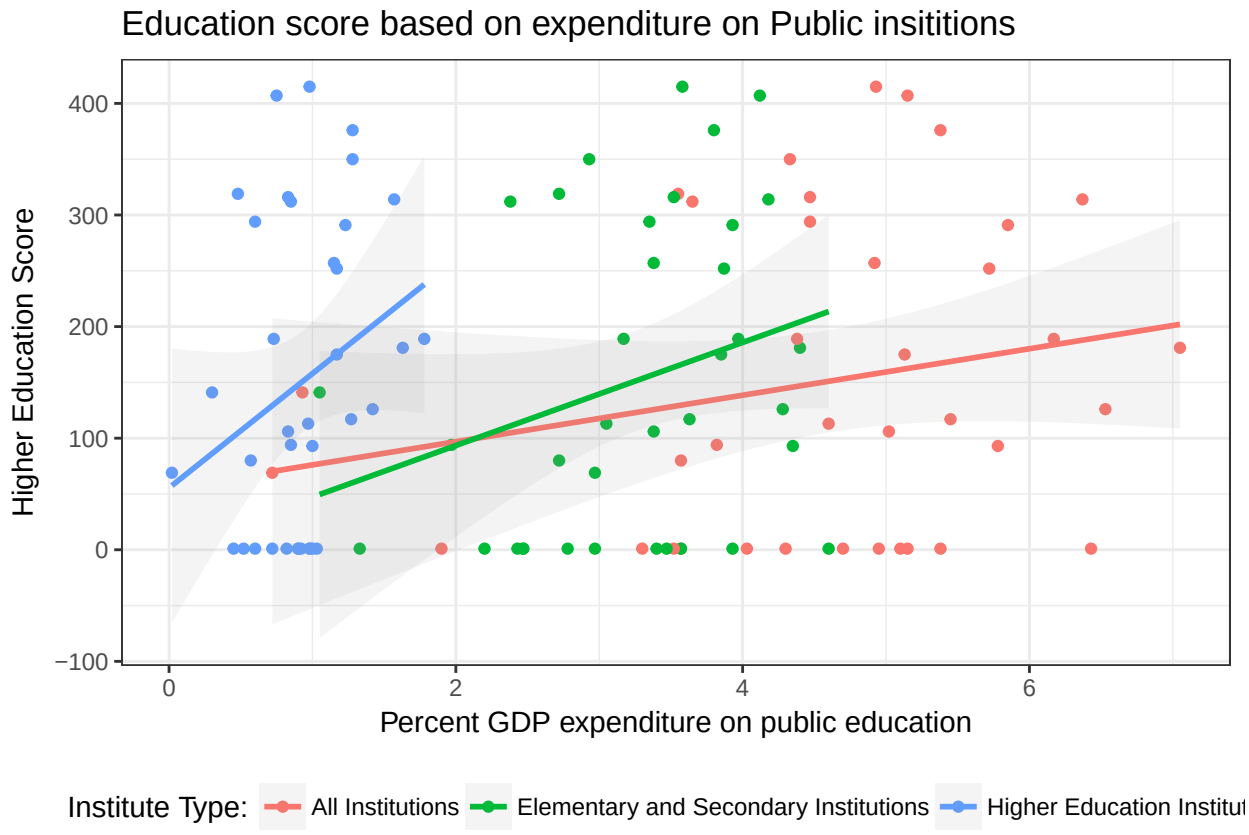
```
## # A tibble: 111 x 5
## # Groups:   country, direct_expenditure_type [37]
##        country direct_expenditure_type   institute_type total_exp
##          <chr>                  <fctr>           <fctr>     <dbl>
##  1     Denmark                  Public All Institutions      7.05
##  2      Norway                  Public All Institutions      6.53
##  3     Iceland                  Public All Institutions      6.43
##  4      Sweden                  Public All Institutions      6.37
##  5     Finland                  Public All Institutions      6.17
##  6     Belgium                  Public All Institutions      5.85
```

```
##  7 New Zealand                  Public All Institutions      5.78
##  8      France                  Public All Institutions      5.72
##  9     Austria                  Public All Institutions      5.45
## 10    Portugal                  Public All Institutions      5.38
## # ... with 101 more rows, and 1 more variables: best_score <dbl>
```

Lets try the Linear model for hypotheis testing

- Lets visualize the variables first

```
df %>% ggplot(aes(total_exp,best_score)) + geom_point(aes(color = institute_type)) + geom_smooth(method
  theme(axis.text=element_text(size=8),axis.title=element_text(size=10,face="bold" )) + theme_bw() + th
```



- For this we will apply linear regression on total_exp and best_score to see the association between education ranking and public education expenditure

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

```
##
## Call:
## lm(formula = best_score ~ total_exp + institute_type, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -206.87 -125.84  -38.94  113.38  262.52
##
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                       7.949     66.399
```

```
## total_exp                                                31.091     13.585
## institute_typeElementary and Secondary Institutions   42.099     36.600
## institute_typeHigher Education Institutions          114.063     59.034
##                                                    t value Pr(>|t|)
## (Intercept)                                          0.120   0.9049
## total_exp                                            2.289   0.0241 *
## institute_typeElementary and Secondary Institutions   1.150   0.2526
## institute_typeHigher Education Institutions          1.932   0.0560 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 136.1 on 107 degrees of freedom
## Multiple R-squared:  0.04667,    Adjusted R-squared:  0.01994
## F-statistic: 1.746 on 3 and 107 DF,  p-value: 0.162
```

Observations from the plot and and linear model

- The plot shows the highest association of public higher education expenditure with higher educations ranking scores
- The higher education score of a country is best explained by the it's public expenditure on higher education institutes and that should be the case also
- The linear model however does not provide a very concrete evidence to reject the null hypothesis
- The P - value of for higher education expenditure versus score is on the margin of significance level testing i.e. 0.056
- This could mean while public expenditure is important for good higher education ranking of a country, it is not the only variable to explain it

8. Next steps would testing other factors affecting higher education scores

   Reserach funding Male-female ratio