How Institutional & Economic Factors Affect Scientific Output

Omar Hassan
Nichola Millman
Zeeshan Javed

April 17, 2021

Contents

1 Introduction

Scientific output has seen a tremendous rise over the past few years, given this exponential growth, we aim to study the economic and institutional factors which affect scientific output across various countries over the years, and we also aim to determine which countries have the greatest scientific output. To do this, we first define scientific output qualitatively and quantitatively, using these two definitions, we explore various economic and institutional aspects which would affect the scientific output by country and across the years. Our economic factors are GDP per capita and research & development expenditure, while our institutional factors are researchers per million and the education index. We then compare these factors with our two functions for scientific output, with the qualitative function being the number of citations per document, whereas the quantitative function is the number of documents per million.

2 Data & Methods

2.1 Data Collection

The majority of our data was collected using an API called WDI. WDI is the API used by the world bank to export all of its data to anyone that wants access. Each data set on the world bank has an indicator. Entering this indicator into the API would return the corresponding data set. An example of an indicator is SP.POP.TOTL; this indicator returns the total population of each country. The API also allowed us to collect data for multiple years at the same time by giving the API a start and an end date. All of the data returned was already formatted into a data frame. We used the world bank API to collect data on population, GDP, Research & development expenditure, Science & technical articles, and researchers per million of each country. We collected all this data over the years of 2012 to 2018. Although the majority of our data was collected from the world bank, we did use other sources to manually collect the citations and education index data. The citations data was collected from SCImago Institutions Rankings and the education index data was collected from the United Nations Development Programme.

2.2 Data Cleaning

Even though collecting the data was relatively simple, cleaning the data was not such an easy job. A substantial amount of the data that we received from the world bank was missing important entries. Typically countries with a lower population were missing the most amount of data, to counter this issue we decided to narrow our data set down to the countries that had a population greater than one million. This did remove a lot of the countries with missing data but there were still some missing entries, our solution was to calculate the missing values using the sample mean of all the previous years where we did have the data. This resolved all of the remaining missing values. Next, we decided to rename all of the columns in the data set to give them more descriptive and usable names. Finally, some of the data points were corresponding to regions of the world or class status rather than countries. So, we removed all regional and class data and kept only the data for the countries.

2.3 Methods

Our most effective tool was data visualizations and correlations. Data visualization often allows us to not only see the relationship between two variables but also understand the limitations of the data (we can limit our data to gain a deeper understanding of the relationship subject to some condition). Our main methods for data visualizations involved scatter, line, bar, and boxplots. Besides data visualizations, we also look at the correlations to determine the strength of the relationships between the variables. When there's an uncertainty regarding the statistical significance of the relationship, we perform a regression analysis (linear regression) and check the confidence intervals to determine the significance.

3 Results

3.1 Scientific Output

We consider two main factors as a measure of scientific output; documents published per million people, and citations per document. The number of documents published per million is a quantitative measure that accounts for the total amount of academic articles being published without being distorted by a given country's population, while the number of citations per article is a qualitative measure. We first see how the number of documents per million change by year:

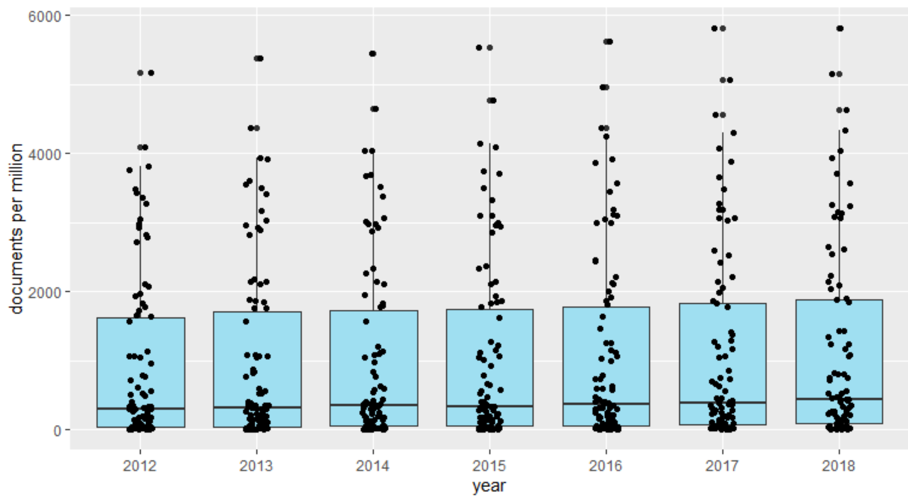


Figure 1: Documents per million by year

As can be seen by figure 1, there appears to have been a slight increase in the third quartile over time, however, after performing a linear regression, it is shown to not be statistically significant.

We now examine how citations per document changes over time:

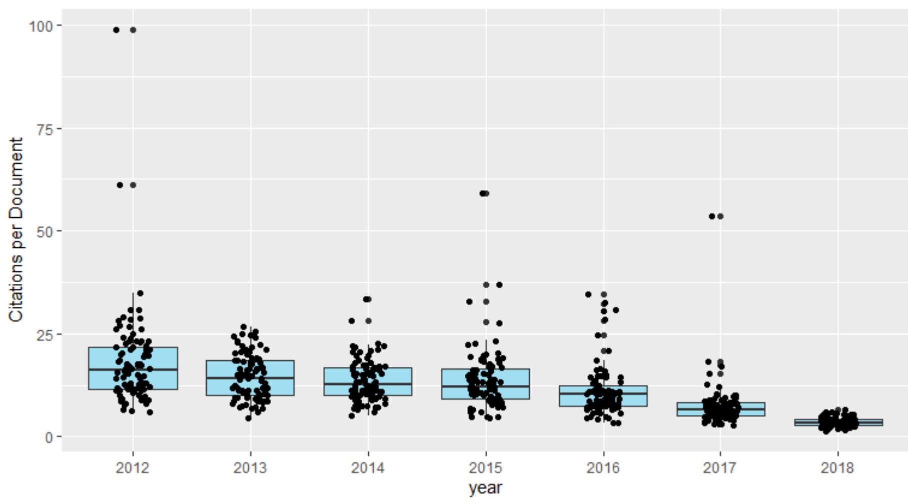


Figure 2: Citations per document by year

Figure 2 shows that there is a clear decrease over the years of the average ratio of citations per document. After performing a linear regression, and having none of the years contain zero in their respective confidence intervals, it is also statistically significant. This suggests that even though the quantity of articles being published around the world has not changed significantly, the amount of citations they have received has decreased. One possible explanation for this would be that older documents have more time to be cited than newer ones.

We now look at the relationship between documents per million and citations per document; we expect to have a strong correlation:

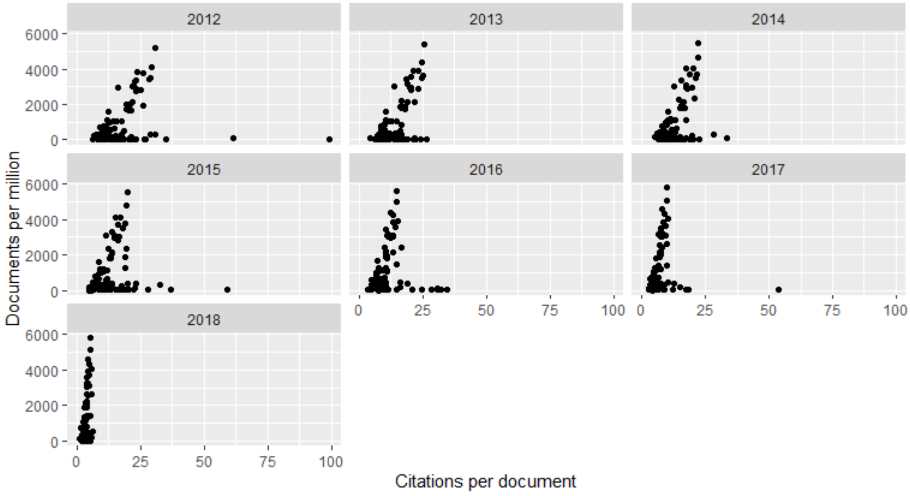
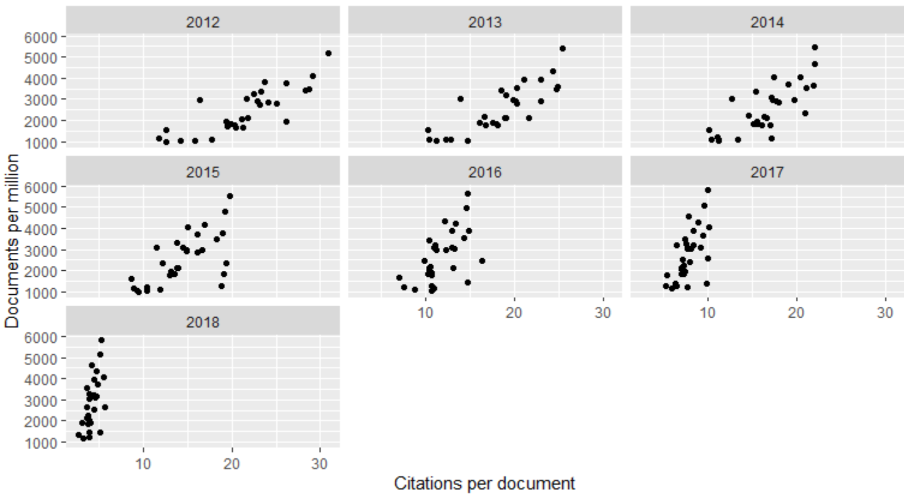


Figure 3: Documents per million vs Citations per document

There appears to be a slightly L shaped relationship between these two variables. With a very low linear correlation value of 0.08713071, this implies that there is no linear relationship. However, one explanation for this would be that a few well cited articles from a country which produces relatively few documents each year would have a large effect on the country's value of citations per document. When considering the countries that have an above average number of documents per million, there appears to be a much stronger correlation:



Year	Correlation
2012	0.847
2013	0.837
2014	0.806
2015	0.700
2016	0.637
2017	0.702
2018	0.639

Figure 4: Documents per million (above average) vs Citations per document

As can be seen by figure 4, there appears to be a solid correlation between documents per million and citations per document among the countries which produced an above average amount of documents per million. It is also notable that the slope increases over time, which can be attributed to the decreasing value of citations per document over time.

We now examine which countries have the greatest scientific output both qualitatively and quantitatively; we only plot the top 10 countries for clarity:

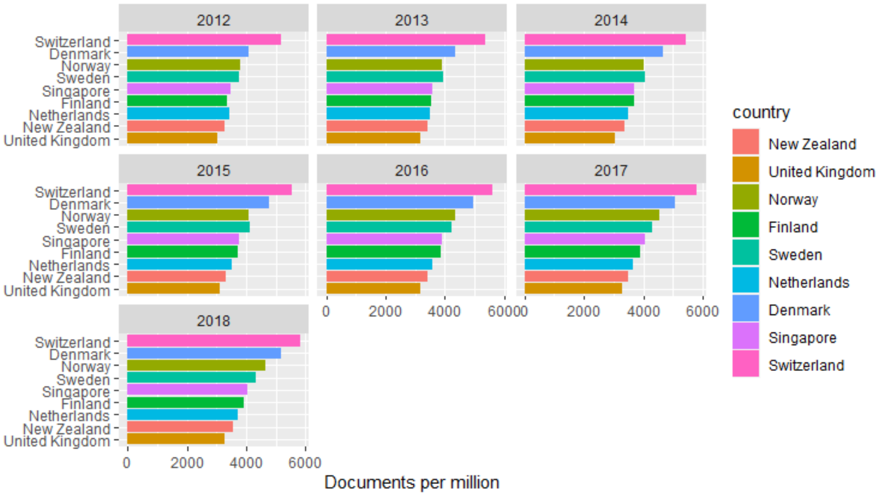


Figure 5: Top 10 Countries by Documents per million by year

It is clear that Switzerland is the top performer across all the years. Followed by Denmark and then Norway/Sweden. There seems to be a greater presence of Nordic countries compared to any other region.

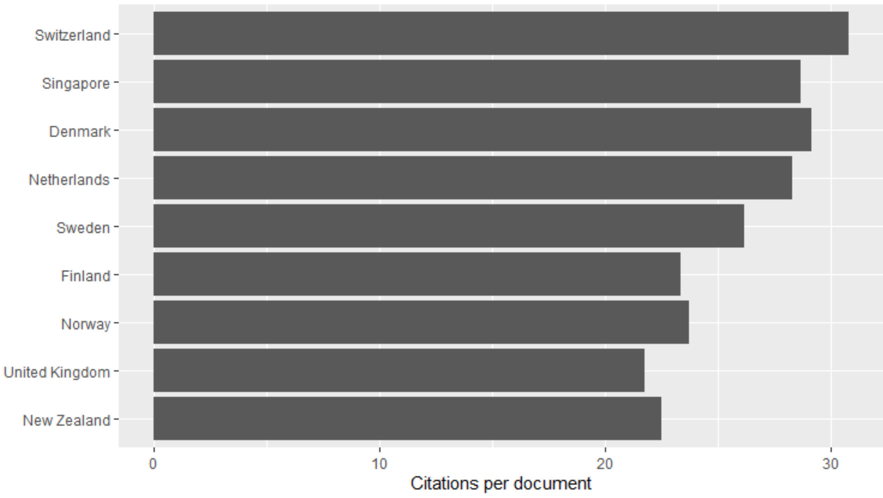
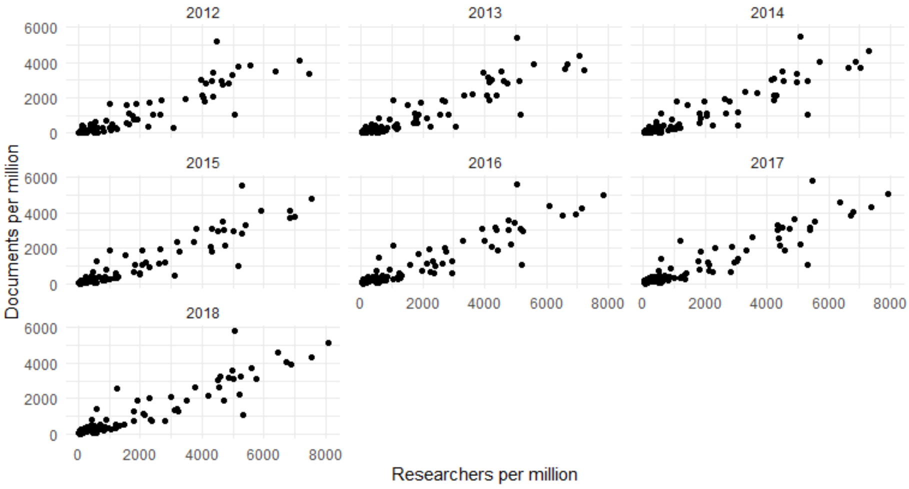


Figure 6: Top 10 Countries by Total Citations per document across 2012-2018

Switzerland is the top performer when it comes to the total citations per document across the years, and the Nordic regions maintain a strong presence in the top 10. We can confidently say that Switzerland is the country with the greatest scientific output, while the Nordic region has the most scientific output when compared to other regions. We now look to the institutional and economic factors which may explain why these countries have such a high scientific output.

3.2 Institutional Factors

We explore two main relationships, the relationship between documents per million and both researchers per million and the education index. Starting with the former, we perform a simple scatter plot for each year giving us:



Year	Correlation
2012	0.914
2013	0.920
2014	0.923
2015	0.925
2016	0.922
2017	0.925
2018	0.922

Figure 7: Documents per million vs Researchers per million

As can be expected, there is a strong correlation between the number of documents published and the number of researchers per million. This is visible from figure 7, the correlation confirms this relationship and can be found to be on average about 0.92. This correlation remains fairly constant throughout the years.

Figure 8: Documents per million vs Education Index

We now explore the relationship between documents per million and education index, we expect countries with a greater education index to produce more documents per million. The relationship is given in figure 8. It is quite clear that there is no relation up until the education index is around 0.6, to get a more accurate investigation of this relationship, we can filter all data points with an education index of less than 0.6. This way, we can get a meaningful correlation between education index greater than 0.6 and documents per million.

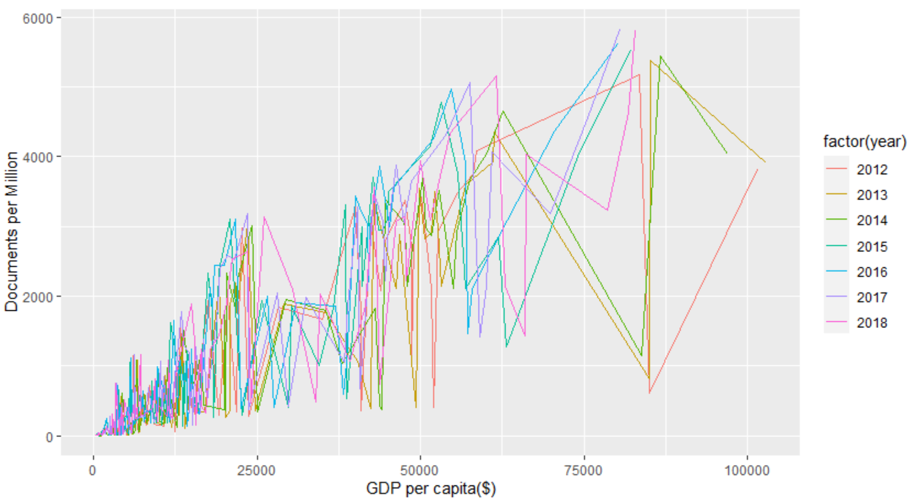
Year	Correlation
2012	0.787
2013	0.792
2014	0.777
2015	0.777
2016	0.775
2017	0.769
2018	0.773

Figure 9: Documents per million vs Education Index (> 0.6)

After filtering, we can see that the relationship between the two variables is quite strong, the exponential relation appears to be more linear when focusing on education index values of > 0.6. The correlation further proves the strength of this relationship with a somewhat constant value across the years, averaging out at around 0.77.

3.3 Economic Factors

To measure the effects of economic factors on scientific output, we decided to measure the effects of 2 factors on scientific output, GDP per capita and research & development expenditure. Starting with the former:

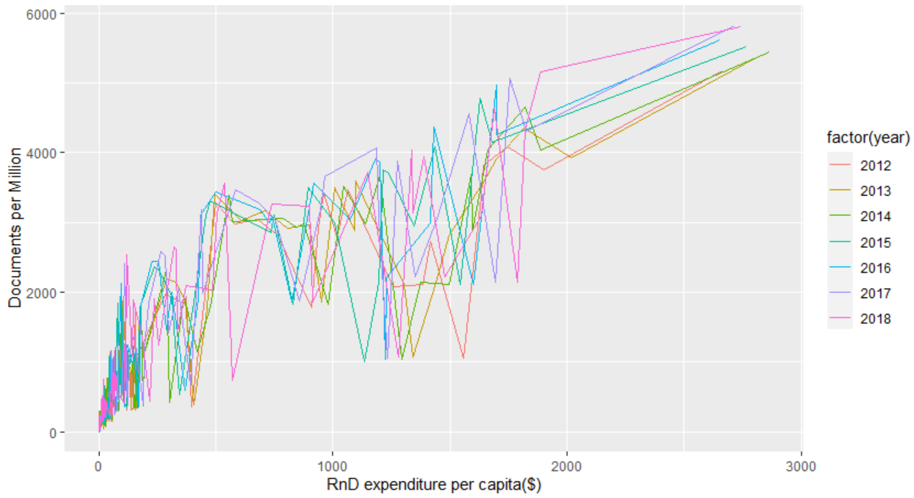


Year	Correlation
2012	0.816
2013	0.833
2014	0.852
2015	0.868
2016	0.882
2017	0.879
2018	0.876

Figure 10: Documents per million vs GDP per capita

As we can clearly see from the above figure, as GDP per capita of a country increases, the amount of scientific output also increases. There seems to be a very strong correlation between GDP per capita and Scientific output. The correlation also seems to be increasing over the years.

We now look to research & development expenditure when compared to documents per million, we expect to have quite a strong correlation given that the more a country spends on research and development, the more documents it should produce:



Year	Correlation
2012	0.873
2013	0.879
2014	0.884
2015	0.874
2016	0.865
2017	0.866
2018	0.865

Figure 11: Documents per million vs Research & Development Expenditure

We can also see there's a strong correlation from figure 11. The correlation between the years has stayed relatively the same. One possible explanation could be that the research & development expenditure did not change significantly through the years of 2012 to 2018.

4 Discussion & Conclusion

After having performed the analysis, we have found a positive correlation between the number of documents published each year and the number of citations per document for countries which publish documents above the average mean. We also discovered that the country with the largest documents per million was consistently Switzerland from the years 2012-2018. In addition, we discovered a positive correlation between GDP per capita and documents per million, and an exponential relationship between the education index and documents per million. The strongest correlation we found was between researchers per million and documents per million.

Some of the limitations that we experienced were the lack of available data for many of the countries in the world. In order to offset this, some of the countries with a few missing data points used the mean average over the years to miss in filling values, but some countries were missing values altogether. Additionally, the numbers provided in the SCImango dataset for citations and number of documents published each year are most likely an underestimation as they were compiled through a collection of academic journals, which was most likely not comprehensive.

Throughout the analysis, there were both systematic and random sources of error. By filling in the missing values of some of the columns with the mean average value across the years, the impact of outliers in the calculations of correlation and linear regression may have been underestimated, in particular for year by year comparisons. Additionally, some academic articles and citations being excluded in the scimango dataset could be a source of random error, as, for the sake of comparison, the impact of the articles and citations that were not included could either increase or decrease a given country's value for citations per document.

We would like to explore this data further using better methods of correcting our missing data, we also lost a significant amount of countries in the cleaning process. We could potentially do fitted values rather than the sample mean and compare the results with this report, and coupling this with finding more comprehensive datasets could provide us with the data to explore our questions to a greater extent. It would also be worth exploring scientific output by region rather than by country, we saw that the Nordic region seemed to be over-represented in the high output countries, we could explore this idea with various other regions as well.

In conclusion, we have found that researchers per million, GDP, research spending, and the education index of a given country all influence its scientific output, with both the quantity and the quality of the countries with a larger number of documents per million being closely related.

5 Appendix

5.1 Tweet

Why do Switzerland & Denmark have the highest scientific output globally? We explored the institutional and economic factors which lead to a high scientific output, and the prevailing factors are the number of researchers, the GDP per capita, and the research expenditure.

5.2 Datasets & remove.txt

The datasets and the remove.txt file can be found on: https://github.com/S-tuberosum/Scientific-Output

5.3 Code

```
# Code written by Zeeshan Javed, Nichola Millman, and Omar Hassan
#--------------------------------------------
#             READ ME
#--------------------------------------------
#Before you do anything, make sure you have all of the libraries installed
#Also make sure all the provided .xlsx files are in the correct directory
#After doing the above, run the entire script at once and all the data sets will be created

#We are only using data for countries with a population greater than 1 million

#There is missing data for some countries

#Run any of your experiments at the very bottom of the script
#so that it doesn't mess with the created datasets

#The data frame labeled 'FinalData' contains all the data for all the years

#The data frame labeled 'FinalDataYYYY' contains all the data for the year YYYY

library(WDI)
library(tidyverse)
library(dplyr)
library(readxl)
library(dslabs)
library(broom)
library(cowplot)


#--------------------------------------------
#             GET THE DATA
#--------------------------------------------


#population of each country from 2012 to 2018
```

```
population = WDI(indicator='SP.POP.TOTL', start=2012, end=2018)

#GDP of each country from 2012 to 2018
gdp = WDI(indicator='NY.GDP.MKTP.CD', start=2012, end=2018)

#research and development expenditure (% of GDP) from 2012 to 2018
RnD_expenditure = WDI(indicator='GB.XPD.RSDV.GD.ZS', start=2012, end=2018)

#Scientific and technical journal articles from 2012 to 2018
Sci_and_tech_journals = WDI(indicator = 'IP.JRN.ARTC.SC', start=2012, end=2018)

#researchers per million from 2012 to 2018
researchers_million = WDI(indicator = 'SP.POP.SCIE.RD.P6', start =2012, end =2018)

#Citations per country / add the year column to the data
Citations2012 <- read_excel("./Data/scimagojr country rank 2012.xlsx")
Citations2012$year = 2012

Citations2013 <- read_excel("./Data/scimagojr country rank 2013.xlsx")
Citations2013$year = 2013

Citations2014 <- read_excel("./Data/scimagojr country rank 2014.xlsx")
Citations2014$year = 2014

Citations2015 <- read_excel("./Data/scimagojr country rank 2015.xlsx")
Citations2015$year = 2015

Citations2016 <- read_excel("./Data/scimagojr country rank 2016.xlsx")
Citations2016$year = 2016

Citations2017 <- read_excel("./Data/scimagojr country rank 2017.xlsx")
Citations2017$year = 2017

Citations2018 <- read_excel("./Data/scimagojr country rank 2018.xlsx")
Citations2018$year = 2018

#Education index per country / add the year column to the data, change col names

EducationIndex2012 <- read_excel("./Data/EducationIndex2012.xlsx")
EducationIndex2012$year = 2012

EducationIndex2013 <- read_excel("./Data/EducationIndex2013.xlsx")
EducationIndex2013$year = 2013

EducationIndex2014 <- read_excel("./Data/EducationIndex2014.xlsx")
EducationIndex2014$year = 2014

EducationIndex2015 <- read_excel("./Data/EducationIndex2015.xlsx")
EducationIndex2015$year = 2015

EducationIndex2016 <- read_excel("./Data/EducationIndex2016.xlsx")
EducationIndex2016$year = 2016

EducationIndex2017 <- read_excel("./Data/EducationIndex2017.xlsx")
EducationIndex2017$year = 2017

EducationIndex2018 <- read_excel("./Data/EducationIndex2018.xlsx")
```

```r
EducationIndex2018$year = 2018
names(EducationIndex2018)[names(EducationIndex2018) == 'education_vaue'] <- 'education_value'

#now bind all the Citations and EducationIndex into 1 data set
citations <- rbind(Citations2012,Citations2013,Citations2014,Citations2015,Citations2016,
                   Citations2017,Citations2018)

eduindex <- rbind(EducationIndex2012,EducationIndex2013,EducationIndex2014,
                  EducationIndex2015,EducationIndex2016,EducationIndex2017,
                  EducationIndex2018)

#now delete some columns from the Citations & eduindex and rename some columns
drop <- c('Rank','Region')
citations = citations[!(names(citations) %in% drop)]
names(citations)[1] <- 'country'

names(eduindex)[names(eduindex) == 'Country'] <- 'country'

#--------------------------------------------
#          Create the Super data set
#--------------------------------------------


FinalData <- inner_join(population, gdp, by=c('country','year','iso2c'))
FinalData <- inner_join(FinalData, RnD_expenditure, by=c('country','year','iso2c'))
FinalData <- inner_join(FinalData, Sci_and_tech_journals, by=c('country','year','iso2c'))
FinalData <- inner_join(FinalData, researchers_million, by=c('country','year','iso2c'))

#rename some columns then add the citation data
names(FinalData)[1] <- 'symbol'
names(FinalData)[3] <- 'population'
names(FinalData)[5] <- 'gdp'
names(FinalData)[6] <- 'RnD_Expenditure'
names(FinalData)[7] <- 'sci_tech_articles'
names(FinalData)[8] <- 'researchers_per_million'

#now add the citations data
FinalData <- inner_join(FinalData, citations, by=c('country','year'))
FinalData <- inner_join(FinalData, eduindex, by=c('country','year'))

names(FinalData)[13] <- 'cit_per_doc'
names(FinalData)[15] <- 'education_index'

#-------------------------------------------------------
#              Clean the Super data set
#-------------------------------------------------------


#now remove all countries with a population less than 1 million
FinalData <- subset(FinalData, population > 1000000)

#create data frame of countries with NA across all the years and remove them
remove <- read.delim("remove.txt")
FinalData <- anti_join(FinalData, remove, by='country')


#replacing all NA values with the sample mean of the column
```

```
FinalData <- group_by(FinalData, country)
FinalData <- mutate(FinalData, mean_rnd = mean(RnD_Expenditure, na.rm = TRUE),
                mean_rpm = mean(researchers_per_million, na.rm = TRUE))

FinalData$RnD_Expenditure <- ifelse(is.na(FinalData$RnD_Expenditure),
                                FinalData$mean_rnd, FinalData$RnD_Expenditure)

FinalData$researchers_per_million <- ifelse(is.na(FinalData$researchers_per_million),
                                FinalData$mean_rpm, FinalData$researchers_per_million)

#removing columns we don't need & formatting
FinalData <- FinalData[-c(1, 14, 16:17)]
FinalData$education_index <- as.double(FinalData$education_index)

#adding our scientific output function

FinalData <- FinalData %>%
  mutate(doc_per_mil =  (Documents*1000000)/population)

#add gdp_per_capita to the Final data
FinalData = FinalData%>%
  mutate(gdp_per_capita =  (gdp/population))

#add RnD expenditure dollar amount
FinalData = FinalData%>%
  mutate(RnD_dollar_amount =  (RnD_Expenditure/100) * gdp)

#add RnD expenditure per capita to the FinalData
FinalData = FinalData%>%
  mutate(RnD_per_capita =  (RnD_dollar_amount/population))

#--------------------------------------------------------
# Delete all the 'extra' data sets that we don't need
#--------------------------------------------------------


rm(Citations2012)
rm(Citations2013)
rm(Citations2014)
rm(Citations2015)
rm(Citations2016)
rm(Citations2017)
rm(Citations2018)
rm(citations)
rm(EducationIndex2012)
rm(EducationIndex2013)
rm(EducationIndex2014)
rm(EducationIndex2015)
rm(EducationIndex2016)
rm(EducationIndex2017)
rm(EducationIndex2018)
rm(eduindex)
rm(gdp)
rm(drop)
rm(population)
rm(RnD_expenditure)
rm(Sci_and_tech_journals)
```

```r
rm(researchers_million)


#---------------------------------------------
# Create the separate data sets for each year
#---------------------------------------------


FinalData2012 <- FinalData %>%
  filter(year == 2012)

FinalData2013 <- FinalData %>%
  filter(year == 2013)

FinalData2014 <- FinalData %>%
  filter(year == 2014)

FinalData2015 <- FinalData %>%
  filter(year == 2015)

FinalData2016 <- FinalData %>%
  filter(year == 2016)

FinalData2017 <- FinalData %>%
  filter(year == 2017)

FinalData2018 <- FinalData %>%
  filter(year == 2018)


#---------------------------------------------
#                 analyses
#---------------------------------------------

#--------------------------------------------------
#      Scientific Output
#--------------------------------------------------

#First lets take a look at the relationship between number of documents published per million per

output_over_time = FinalData%>%
  group_by(year)%>%
  summarize(sum(Documents),mean(Documents))

#boxplot no jitter
FinalData$year = factor(FinalData$year)
ggplot(FinalData,aes(x= year, y = doc_per_mil)) +
  geom_boxplot( fill = "#9FDFF1") +
  ylab("Documents per million")

#boxplot with jitter
FinalData$year = factor(FinalData$year)
ggplot(FinalData,aes(x= year, y = doc_per_mil)) +
  geom_boxplot( fill = "#9FDFF1") +
  geom_jitter(height =0.10,width =0.10) +
  ylab("documents per million")
```

```
#linear model
fit <- lm(doc_per_mil~year, data = FinalData)
fit
confint(fit)


#Now, let's look at the relationship between citations and articles published


#boxplot not jitter
FinalData$year = factor(FinalData$year)
ggplot(FinalData,aes(x= year, y = cit_per_doc)) +
  geom_boxplot( fill = "#9FDFF1") +
  ylab("Citations per Document")

#boxplot with jitter
FinalData$year = factor(FinalData$year)
ggplot(FinalData, aes(x = year, y = cit_per_doc)) +
  geom_boxplot(fill = "#9FDFF1") +
  geom_jitter(height =0.15, width =0.15) +
  ylab("Citations per document")

#linear model
fit2 <- lm(cit_per_doc~year, data = FinalData)
fit2
confint(fit2)

#Now let's look at output vs population

#point graph over the years

FinalData%>%
  ggplot()+
  geom_point(aes(x= population/100000, y = Documents/1000))+
  xlab("Population")+
  ggtitle("Population vs Published Documents")+
  facet_wrap(~year)+
  theme(plot.title = element_text(hjust = 0.5))

#point graph zoomed in (outliars out of view)
FinalData2018%>%
  ggplot()+
  geom_point(aes(x= population/1000000, y = Documents/100))+
  coord_cartesian(ylim = c(0,3000))+
  ggtitle("Population vs Published Documents 2018")+
  theme(plot.title = element_text(hjust = 0.5))

#linear model and correlation
cor(FinalData$population,FinalData$Documents,method = "pearson")
fit3 <- lm(population~Documents, data = FinalData)
fit3
confint(fit3)

#Now lets look at the top performers

#first, lets graph citations per document vs documents per million
FinalData%>%
```

```
  ggplot()+
  geom_point(aes(x= cit_per_doc, y = doc_per_mil))+
  xlab("Citations per document")+
  ylab("Documents per million")+
  facet_wrap(~year)

cor(FinalData$cit_per_doc,FinalData$doc_per_mil,method = "pearson")
fit4 <- lm(cit_per_doc~doc_per_mil, data = FinalData)
fit4
confint(fit4)

#Now lets look at above average countries

#above mean average docs per mil dataset
above_avg2012 = FinalData2012%>%
  filter(doc_per_mil>mean(FinalData2012$doc_per_mil))

above_avg2013 = FinalData2013%>%
  filter(doc_per_mil>mean(FinalData2013$doc_per_mil))

above_avg2014 = FinalData2014%>%
  filter(doc_per_mil>mean(FinalData2014$doc_per_mil))

above_avg2015 = FinalData2015%>%
  filter(doc_per_mil>mean(FinalData2015$doc_per_mil))

above_avg2016 = FinalData2016%>%
  filter(doc_per_mil>mean(FinalData2016$doc_per_mil))

above_avg2017 = FinalData2017%>%
  filter(doc_per_mil>mean(FinalData2017$doc_per_mil))

above_avg2018 = FinalData2018%>%
  filter(doc_per_mil>mean(FinalData2018$doc_per_mil))

above_avg = full_join(above_avg2012,above_avg2013)%>%
  full_join(above_avg2014)%>%
  full_join(above_avg2015)%>%
  full_join(above_avg2016)%>%
  full_join(above_avg2017)%>%
  full_join(above_avg2018)

#graph of above average docs per mil vs citations per document

above_avg%>%
  ggplot()+
  geom_point(aes(x= cit_per_doc, y = doc_per_mil))+
  xlab("Citations per document")+
  ylab("Documents per million")+
  facet_wrap(~year)

above_avg_cor = above_avg%>%
  group_by(year)%>%
  summarize(cor(cit_per_doc,doc_per_mil,method = "pearson"))

#performing a linear regression to see if the above mean average docs_per_mil are increasing over t
fit4 <- lm(year~doc_per_mil, data = above_avg)
```

```r
fit4
confint(fit4)

#Now, lets consider the top ten countries for docs per million over the years
#first let's filter the dataset for the top ten each year
top_performers2012 = FinalData2012%>%
  filter(doc_per_mil>3000)

top_performers2013 = FinalData2013%>%
  filter(doc_per_mil>3100)

top_performers2014 = FinalData2014%>%
  filter(doc_per_mil>3050)

top_performers2015 = FinalData2015%>%
  filter(doc_per_mil>3104.5)

top_performers2016 = FinalData2016%>%
  filter(doc_per_mil>3150)

top_performers2017 = FinalData2017%>%
  filter(doc_per_mil>3200)

top_performers2018 = FinalData2018%>%
  filter(doc_per_mil>3250)


top_performers = full_join(top_performers2012,top_performers2013)%>%
  full_join(top_performers2014) %>%
  full_join(top_performers2015) %>%
  full_join(top_performers2016) %>%
  full_join(top_performers2017) %>%
  full_join(top_performers2018)

#now let's do a bar graph of each year

top_performers%>%
  ggplot(aes(y= doc_per_mil,x= reorder(country,doc_per_mil), fill = country))+
  geom_bar(stat = "identity",position = "dodge")+
  ylab("Documents per million")+
  xlab("")+
  coord_flip() +
  facet_wrap(~year)

#Now lets look at citations per doc of the top ten countries

top_performers %>%
  ggplot(aes(y= cit_per_doc,x = country)) +
  geom_bar(stat = "identity",position = "dodge") +
  ylab("Citations per document") +
  xlab("") +
  coord_flip()


#-------------------------------------------------
#      Institutional factors vs scientific output
#-------------------------------------------------
```

```r


# scatter doc/mil vs researchers/mil
FinalData %>%
  ggplot() +
  geom_point(aes(x= researchers_per_million, y = doc_per_mil)) +
  xlab("Researchers per million") +
  ylab("Documents per million") +
  facet_wrap(~year) +
  theme_minimal()

# correlations for doc/mil vs researchers/mil
temp1 <- cor.test(FinalData2012$researchers_per_million,
                  FinalData2012$doc_per_mil, method = 'pearson')
temp1

temp2 <- cor.test(FinalData2013$researchers_per_million,
                  FinalData2013$doc_per_mil, method = 'pearson')
temp2

temp3 <- cor.test(FinalData2014$researchers_per_million,
                  FinalData2014$doc_per_mil, method = 'pearson')
temp3

temp4 <- cor.test(FinalData2015$researchers_per_million,
                  FinalData2015$doc_per_mil, method = 'pearson')
temp4

temp5 <- cor.test(FinalData2016$researchers_per_million,
                  FinalData2016$doc_per_mil, method = 'pearson')
temp5

temp6 <- cor.test(FinalData2017$researchers_per_million,
                  FinalData2017$doc_per_mil, method = 'pearson')
temp6

temp7 <- cor.test(FinalData2018$researchers_per_million,
                  FinalData2018$doc_per_mil, method = 'pearson')
temp7

# scatter doc/mil vs education index
FinalData %>%
  filter(education_index > 0.6) %>%
  ggplot() +
  geom_point(aes(x= education_index, y = doc_per_mil)) +
  xlab("Education Index") +
  ylab("Documents per million") +
  facet_wrap(~year) +
  theme_minimal()

# correlations for doc/mil vs education index
FinalData %>%
  filter(education_index > 0.6) %>%
  group_by(year)

temp8 <- cor.test(filter(FinalData2012, education_index >0.6)$education_index,
                  filter(FinalData2012, education_index >0.6)$doc_per_mil,
```

```
                  method = 'pearson')
temp8

temp9 <- cor.test(filter(FinalData2013, education_index >0.6)$education_index,
                  filter(FinalData2013, education_index >0.6)$doc_per_mil,
                  method = 'pearson')
temp9

temp10 <- cor.test(filter(FinalData2014, education_index >0.6)$education_index,
                   filter(FinalData2014, education_index >0.6)$doc_per_mil,
                   method = 'pearson')
temp10

temp11 <- cor.test(filter(FinalData2015, education_index >0.6)$education_index,
                   filter(FinalData2015, education_index >0.6)$doc_per_mil,
                   method = 'pearson')
temp11

temp12 <- cor.test(filter(FinalData2016, education_index >0.6)$education_index,
                   filter(FinalData2016, education_index >0.6)$doc_per_mil,
                   method = 'pearson')
temp12

temp13 <- cor.test(filter(FinalData2017, education_index >0.6)$education_index,
                   filter(FinalData2017, education_index >0.6)$doc_per_mil,
                   method = 'pearson')
temp13

temp14 <- cor.test(filter(FinalData2018, education_index >0.6)$education_index,
                   filter(FinalData2018, education_index >0.6)$doc_per_mil,
                   method = 'pearson')
temp14
#--------------------------------------------------
#      Economic factors vs scientific output
#--------------------------------------------------


#graph gdp per capita vs documents per million over the years
FinalData %>%
  ggplot(aes(x = gdp_per_capita, y= doc_per_mil,group=year, color = factor(year))) +
  xlab("GDP per capita($)") +
  ylab("Documents per Million") +
  geom_line()

#Find the correlation
temp15 <- cor.test(FinalData2012$gdp_per_capita, FinalData2012$doc_per_mil, method = 'pearson')
temp15

temp16 <- cor.test(FinalData2013$gdp_per_capita, FinalData2013$doc_per_mil, method = 'pearson')
temp16

temp17 <- cor.test(FinalData2014$gdp_per_capita, FinalData2014$doc_per_mil, method = 'pearson')
temp17

temp18 <- cor.test(FinalData2015$gdp_per_capita, FinalData2015$doc_per_mil, method = 'pearson')
temp18

```

```r
temp19 <- cor.test(FinalData2016$gdp_per_capita, FinalData2016$doc_per_mil, method = 'pearson')
temp19

temp20 <- cor.test(FinalData2017$gdp_per_capita, FinalData2017$doc_per_mil, method = 'pearson')
temp20

temp21 <- cor.test(FinalData2018$gdp_per_capita, FinalData2018$doc_per_mil, method = 'pearson')
temp21

#graph RnD expenditure per capita vs documents per million over the years
FinalData %>%
  ggplot(aes(x=RnD_per_capita,y=doc_per_mil,group=year, color = factor(year))) +
  xlab("RnD expenditure per capita($)") +
  ylab("Documents per Million") +
  geom_line()


#Find the correlation
temp22 <- cor.test(FinalData2012$RnD_per_capita, FinalData2012$doc_per_mil, method = 'pearson')
temp22

temp23 <- cor.test(FinalData2013$RnD_per_capita, FinalData2013$doc_per_mil, method = 'pearson')
temp23

temp24 <- cor.test(FinalData2014$RnD_per_capita, FinalData2014$doc_per_mil, method = 'pearson')
temp24

temp25 <- cor.test(FinalData2015$RnD_per_capita, FinalData2015$doc_per_mil, method = 'pearson')
temp25

temp26 <- cor.test(FinalData2016$RnD_per_capita, FinalData2016$doc_per_mil, method = 'pearson')
temp26

temp27 <- cor.test(FinalData2017$RnD_per_capita, FinalData2017$doc_per_mil, method = 'pearson')
temp27

temp28 <- cor.test(FinalData2018$RnD_per_capita, FinalData2018$doc_per_mil, method = 'pearson')
temp28
```