R&D EXPENDITURES

A Study of Patents

Group VI

DANYAL JAMAL KHAKWANI	23110040
HAMZA RASHID INAM	23110210
MALIK M. MUSSAB	23110229
RUBAB MUKHTAR	23110226
NAMRA ASGHAR	23110034
FARAN MAOOD	23110100

Table of Contents

Abstract

This report aims to employ data science essentials on a Patent Data set by examining how government spending on R&D affects the number of patents in a country. The Data was first cleaned, structured, explored, and then made ready for analysis. We also used complementary datasets to include control variables to obtain accurate regression results from our main model. The control variables are given year-wise and for each country, as follows: Tertiary school enrollment, Government spending per student, GDP per capita, labor force participation rate, patent applications filed by residents, and tax revenue to. Hence, the main goal of the report is to analyze the effect of government spending on Research and development on the number of patents registered, by controlling other macroeconomic factors.

Introduction

Government spending is characterized by the money spent by the public sector in avenues such as education, social protection, defence, and research and development etc. in order for the acquisition of goods and services. (Guellec 2000). The purpose of government spending includes (but is not limited to):

1) providing services, goods and benefits that are not provided by the private sector.

2) providing subsidies to industries that are struggling.

3) promoting social welfare.

Point 1, however, relays a very important concept. Government spending helps fulfil the gaps that are left by the private sector.

Research Question

A case can be made for research and development here. Although most companies engage in Research and development in order to improve the overall investment prospects within the country, firms are hesitant if there is uncertainty around government support related to this matter. Here, government spending on research and development (combined with tax incentives), produces long term economic benefits in terms of technological development and investment opportunity within the country.

Consistent and continuous efforts on research and development hence create a knowledge capital which is influential to achieve long term economic goals. This, combined with the availability of physical capital, gives rise to new inventions and technologies (Ramesh 2020). With the increase in creation of knowledge (in technological terms), these efforts are further protected by the approval of patent rights. Hence, a case can be made regarding the impact of government spending on R&D and the readiness with which Patents are registered (Ramesh 2020). Therefore, out tentative research question is:

"Does Government R&D Expenditure influence the development of patents?"

Variables

Dependent Variable

1. Number of patents filed by a country in a year

Independent Variables

Our **main independent** variables are:

1. Research and development expenditure (% of GDP)
   R&D is expressed as a percentage of total GDP. The R&D expenditure includes investment into business enterprise, government, higher education, private and non-profit avenues. R&D includes various types of research including but not limited to basic research, applied research, and experimental development.

Here are other **supporting independent variables** to account for other factors which may also influence the development of patents in a country:

2. School enrollment, tertiary (% gross)
   School enrollment in in tertiary relates to people enrolled in advanced qualifications and can be directly associated with advanced skills and human capital, which improves productivity.

3. Government expenditure per student, tertiary (% of GDP per capita)
   This variable measures the amount the government allocates per student for tertiary education expressed as a percentage of GDP per capita. Tertiary education is important because it is directly related to skill and human capital development.

4. GDP per capita (current US$)
   GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy and can be used to gauge how well people are doing overall in an economy. Data are in current U.S. dollars.

5. Labor force participation rate, total (% of total population ages 15-64)
   The proportion of population that is between the ages of 15-64 and is participating in the economy and contributing towards the formation of capital.

6. Patent applications, residents
   Patent application provides a person an exclusive right to a procedure or a device for about 20 years.

7. Tax revenue (% of GDP)
   Tax revenue is the money collected by the government from the people to be used for welfare, investment into various avenues like R&D, defence, healthcare and education. It is expressed as a percentage of GDP.

Data Cleaning

Introduction

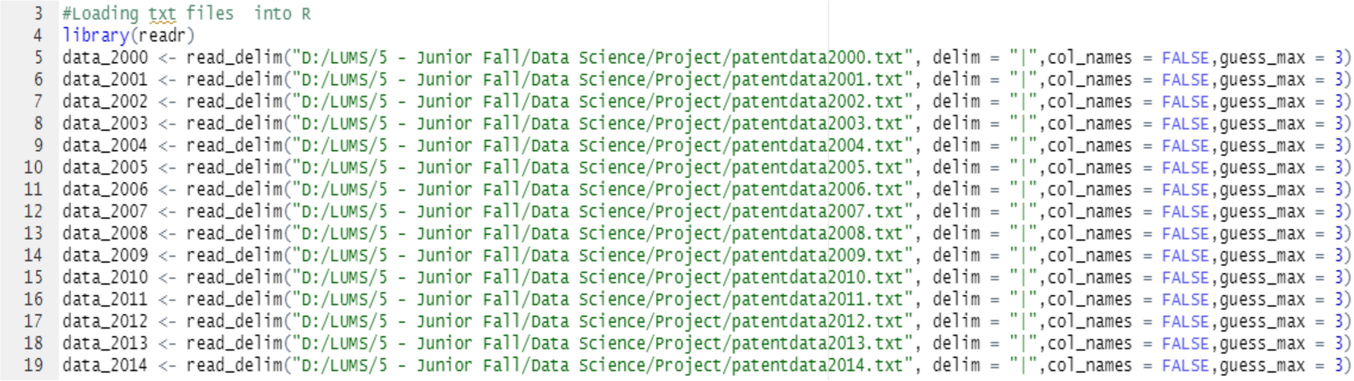
Data cleaning is the most time-consuming, cumbersome task of the project and the steppingstone from where exploratory data analysis and statistical analysis begin. Data cleaning is not just a one-step process, rather a continuous process as it involves setting up data for different analyses at every step. The massive scale of the data and which was scattered across different files made the task of compiling and then cleaning even more difficult. The Data cleaning segment was divided into two parts, data compiling and data formatting for analysis. The latter being dependent upon the former.

Data Description

Before starting off on the data cleaning aspect, it is important to have a thorough understanding of the structure of the data. The data involved was related to patents taken from the US Patents Office UPTO. The fields included were the year of the patent, patent number, assignee name (the company which registered the patent), city of the first inventor, the country where the patent was issued, class of the patent (industry/area of research it belonged to), the subclass of the patent (the niche of the class, which means certain specific industry within the class or certain area of research within the section). Yearly data was available for this from 1994 onwards till 2014. For our research question involving R&D spending and its relationship with the number of patents, we considered the time horizon from 2000-2014 to be adequate for our analysis purpose. We used an external source of data, the **World Bank**, to get the data on R&D spending as a percentage of the GDP for all the countries.

Data Loading Process

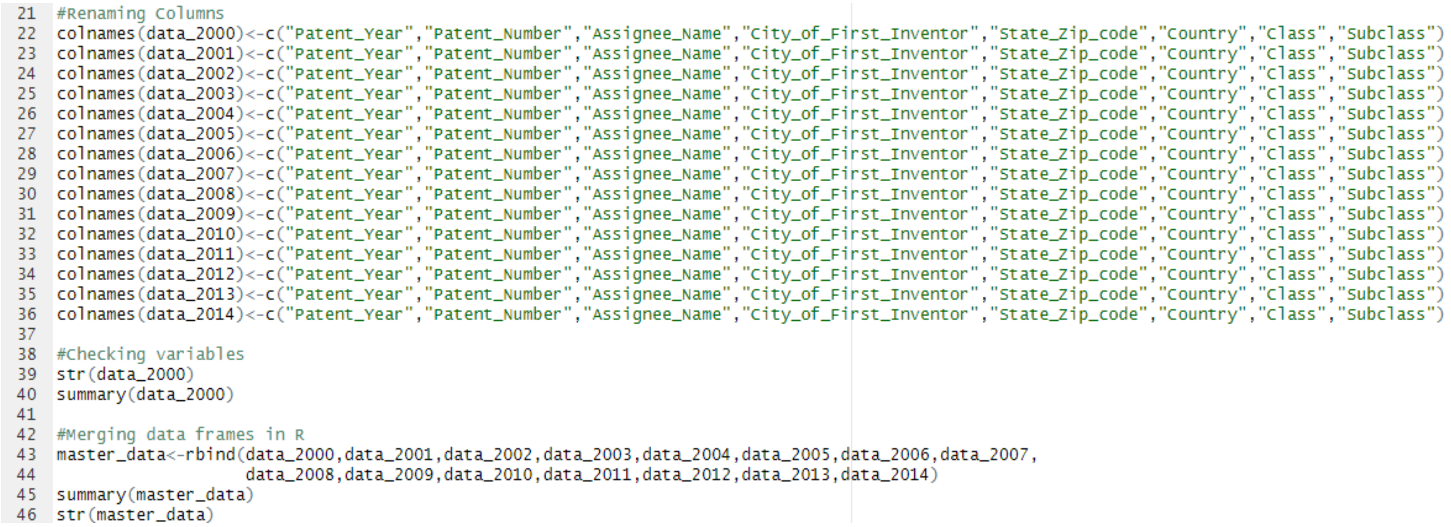
This part of the report will give you a walkthrough of the data compiling process. Firstly, patent data for each year was downloaded from 2000-2014. The data was in the **".txt"** format with missing headers. In addition to this, the size of **".txt"** files were relatively large in comparison to the ones we are accustomed to working with. Hence, the base R functions for reading delimited files were not used as it would have made the process very time-consuming. Instead, **"readr"** package was loaded and the "read_delim" function was used. The **"read_delim"** function is very fast in reading .txt files and made the whole process quite fast. In the **".txt files"**, each field was separated by **"|"** delimiter, with missing headings, and another command **"guess_max"** which is built into **"read_delim"** was used. The **"guess_max"** command identifies the data type in the rows specified and keeps it constant.



```
#Loading txt files  into R
library(readr)
data_2000 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2000.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2001 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2001.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2002 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2002.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2003 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2003.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2004 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2004.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2005 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2005.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2006 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2006.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2007 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2007.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2008 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2008.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2009 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2009.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2010 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2010.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2011 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2011.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2012 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2012.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2013 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2013.txt", delim = "|",col_names = FALSE,guess_max = 3)
data_2014 <- read_delim("D:/LUMS/5 - Junior Fall/Data Science/Project/patentdata2014.txt", delim = "|",col_names = FALSE,guess_max = 3)
```

Post-Loading and Data Handling Process

After the data was loaded, it had missing column headings so the data in the data frame with just the values and missing field names. **"colnames"**, which is a base R function, was used to add appropriate column headings to each **".txt"** file that was loaded. Following this, the data was loaded into R, but data for each year was in a separate data frame and we needed it to be consolidated in a single data frame for our analysis purposes. To tackle this **"rbind"** another base R function was used, this function allows several data frames to be combined by rows, which we saved into a variable called.



```r
#Renaming Columns
colnames(data_2000)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2001)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2002)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2003)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2004)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2005)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2006)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2007)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2008)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2009)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2010)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2011)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2012)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2013)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")
colnames(data_2014)<-c("Patent_Year","Patent_Number","Assignee_Name","City_of_First_Inventor","State_Zip_code","Country","Class","Subclass")

#Checking variables
str(data_2000)
summary(data_2000)

#Merging data frames in R
master_data<-rbind(data_2000,data_2001,data_2002,data_2003,data_2004,data_2005,data_2006,data_2007,
                   data_2008,data_2009,data_2010,data_2011,data_2012,data_2013,data_2014)
summary(master_data)
str(master_data)
```

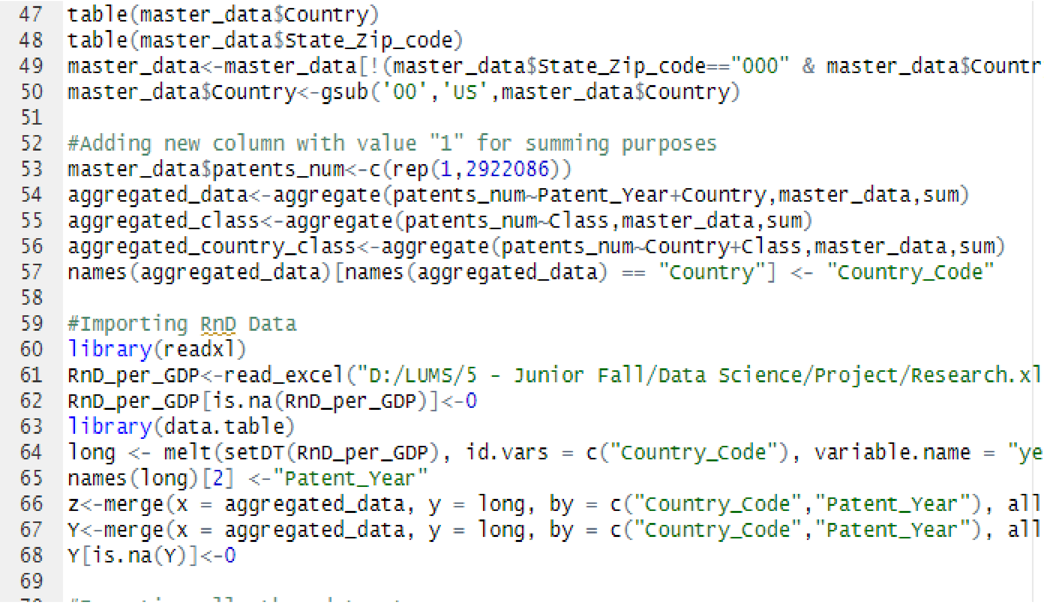
Handling Missing Values

Now we moved on to handling missing values in the data, specifically for the Country column in the **"master_data"**. Firstly, we removed columns that had both the **state zip code** and **country** missing as the missing data could not be justified, so, the only appropriate way to handle was them was through removing them, using **subsetting** available in base R. Moving on, we then renamed rows in the country column that had missing field and could be identified by **"00"**, from the table it could be observed the zip code for these countries was in the US format hence all the values with "00" were renamed to the US.

Manipulating Data For Analysis

After the master data was compiled, there was no way, to sum up, the patents, hence, an additional column named weights was added with value 1, which made summing the patents easier. This enabled the creation of smaller data sets for microanalysis without any hassle as summing patents for grouping purposes could be done without any problem. 2,922,086 rows with value 1 were added these were the number of total patents between 2000 and 2014. This value was obtained when **"View"** function was used on the master data and the table format showed the number of rows which equaled the number of patents. Afterward, the **"aggregate"** function was used which belongs to the **"dplyr"** package and can be used to manipulate data easily for various analyses. Aggregate function splits the data into subsets, computes summary statistics for each, and returns the result in a convenient form. So, a variable was formed named **"aggregated data"** in which patents were summed and were grouped by country and year. Afterward, column names were renamed in the **"aggregated data",** so data was standardized across tables. Moving on, as the patents data was organized it was time to work on the R&D data which included the country's R&D spending as a percentage of GDP from the year 2000 to 2014 and was taken in

".xlsx" format from the world bank website. The data on R&D had missing values as certain years there was no allocation for R&D in several countries, so they were replaced by 0 in the data frame using **"is.na"** function. The R&D data was in the wide format while the rest of the data was in long format, so the data in the R&D table had to be changed to log format so tables could be merged for analysis. **"Reshape"** package was loaded and **"melt"** function was used which converts data

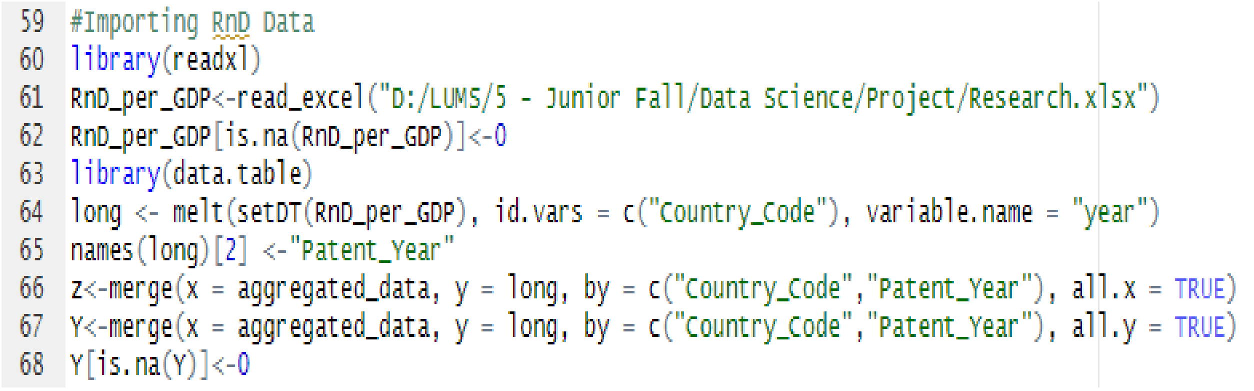


```
table(master_data$Country)
table(master_data$State_Zip_code)
master_data<-master_data[!(master_data$State_Zip_code=="000" & master_data$Country=="00"),]
master_data$Country<-gsub('00','US',master_data$Country)

#Adding new column with value "1" for summing purposes
master_data$patents_num<-c(rep(1,2922086))
aggregated_data<-aggregate(patents_num~Patent_Year+Country,master_data,sum)
aggregated_class<-aggregate(patents_num~Class,master_data,sum)
aggregated_country_class<-aggregate(patents_num~Country+Class,master_data,sum)
names(aggregated_data)[names(aggregated_data) == "Country"] <- "Country_Code"

#Importing RnD Data
library(readxl)
RnD_per_GDP<-read_excel("D:/LUMS/5 - Junior Fall/Data Science/Project/Research.xlsx")
RnD_per_GDP[is.na(RnD_per_GDP)]<-0
library(data.table)
long <- melt(setDT(RnD_per_GDP), id.vars = c("Country_Code"), variable.name = "year")
names(long)[2] <-"Patent_Year"
Z<-merge(x = aggregated_data, y = long, by = c("Country_Code","Patent_Year"), all.x = TRUE)
Y<-merge(x = aggregated_data, y = long, by = c("Country_Code","Patent_Year"), all.y = TRUE)
Y[is.na(Y)]<-0

```

from wide to long format and was applied on the R&D dataset and stored in the variable **"long".**

Data Consolidation for EDA and Statistical Analysis

Afterward, the **"merge"** function was used which is part of **"data.table"** function and is used to join the R&D table with aggregated data in two different ways. Firstly, both the tables were joined on the basis of country and year field in aggregated data for exploratory data analysis purposes. Later on, they were merged on the basis of data in year and country field in R&D so the data could be prepared for panel data and regression analysis could be performed.

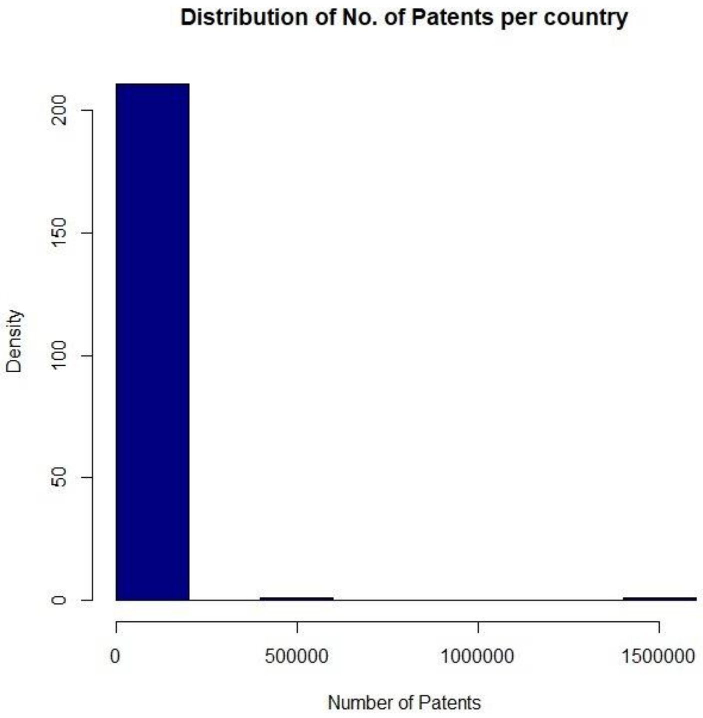


```
#Importing RnD Data
library(readxl)
RnD_per_GDP<-read_excel("D:/LUMS/5 - Junior Fall/Data Science/Project/Research.xlsx")
RnD_per_GDP[is.na(RnD_per_GDP)]<-0
library(data.table)
long <- melt(setDT(RnD_per_GDP), id.vars = c("Country_Code"), variable.name = "year")
names(long)[2] <-"Patent_Year"
Z<-merge(x = aggregated_data, y = long, by = c("Country_Code","Patent_Year"), all.x = TRUE)
Y<-merge(x = aggregated_data, y = long, by = c("Country_Code","Patent_Year"), all.y = TRUE)
Y[is.na(Y)]<-0
```

Exploratory Data Analysis

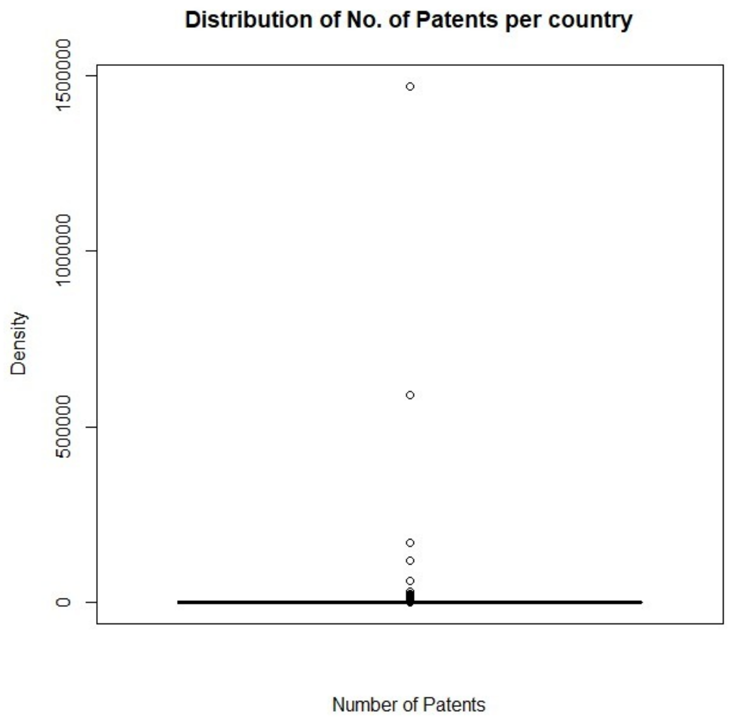
We took the approach of understanding the macro concepts of our research first and then moved towards the micro aspects to prove our understandings through individual examples. For the choice of each analysis tool, we preferred conceptualization and better understanding over making complex graphs that are harder to interpret and might result in incorrect assumptions or comprehension. We used a **funnel approach**, starting our exploration from a pool of **214 countries**, and eventually based on our findings, cutting down the data to the **top 10 patent holding countries**.

Distribution of No. of Patents per Country (Histogram)



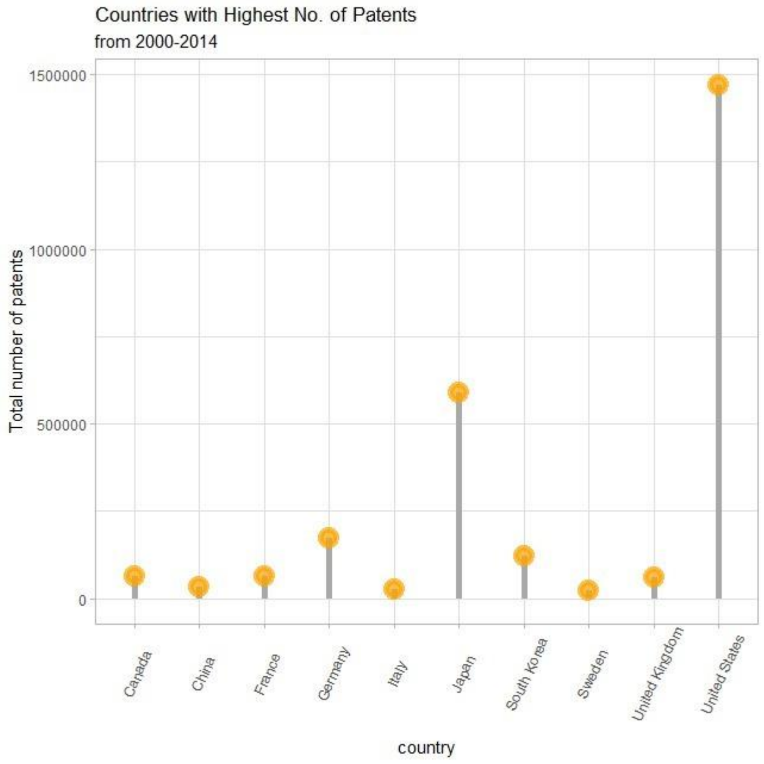
To understand basic trends, properties, patterns, and set an approach, we plotted a histogram using the **basic plotting system** for Density against the Number of Patents. Upon analysis, we found that the majority of the countries have **0 patents** whereas very few had a high number of patents.

Distribution of No. of Patents per Country (Box Plot)



Next, we plotted a **Box Plot** of Density against Number of patents to better visualize the **distribution of Patents** over all the countries. This allowed us to get more clarity on the fact that most countries had next to no patents as could be seen in the very densely packed plots **close to the zero lines**. This also showed the significant disparity between the top countries (based on the number of patents) attributing to the large difference in the number of patents between countries.

<u>Hence, we now focused our analysis on the Top 10 countries according to the number of patents</u>

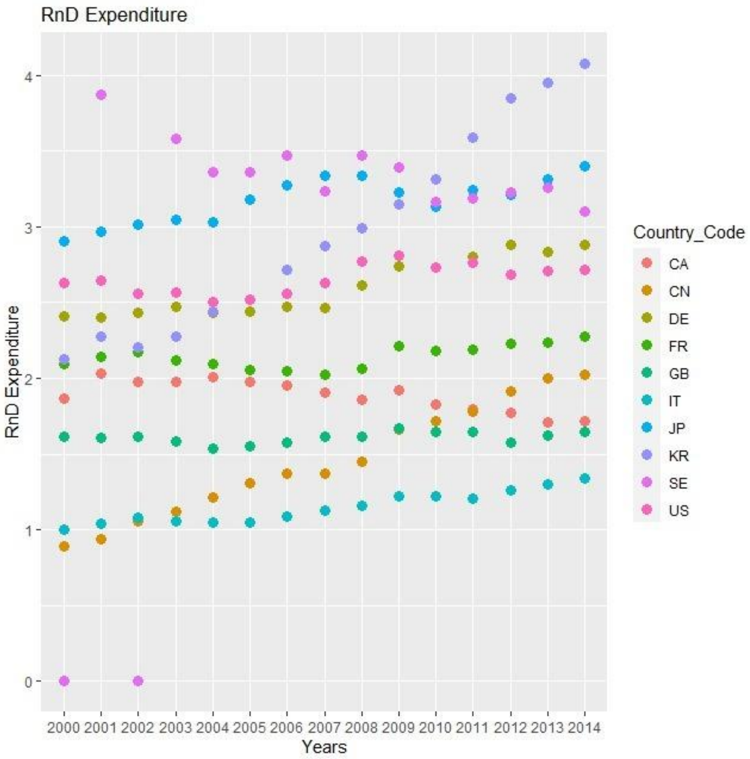


The lollipop graph depicts the number of patents in the top 10 countries with the **highest patents.**

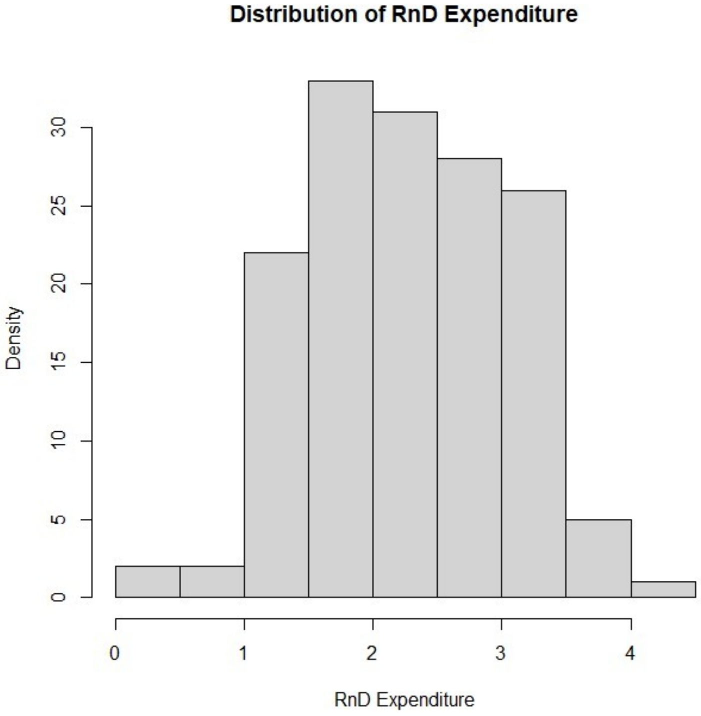
1. United States
2. Japan
3. Germany
4. South Korea
5. Canada
6. France
7. United Kingdom
8. China
9. Italy
10. Sweden

Examining the Trend and Distributions of all the Independent Variables for Top 10 Countries:

- **RnD Expenditure**

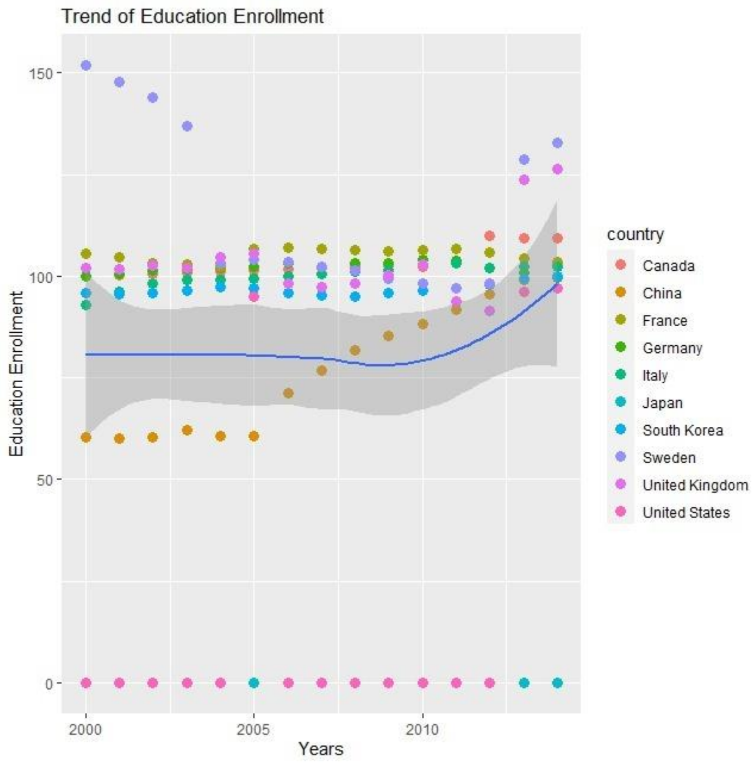


Further, we used the **ggplot2 library** to plot a scatterplot of RnD Expenditure against Years for the Top 10 countries to understand the trends in RnD expenditure. We noticed here that for the majority of the countries, the RnD expenditure **either stayed stable** or increased with **slight deviations**. Furthermore, a noticeable increase was seen in the RnD expenditure of South Korea (Code:KR) over the years with a sharp increase from being in the middle of the pack in the year 2000 to the highest RnD expenditure in 2014.

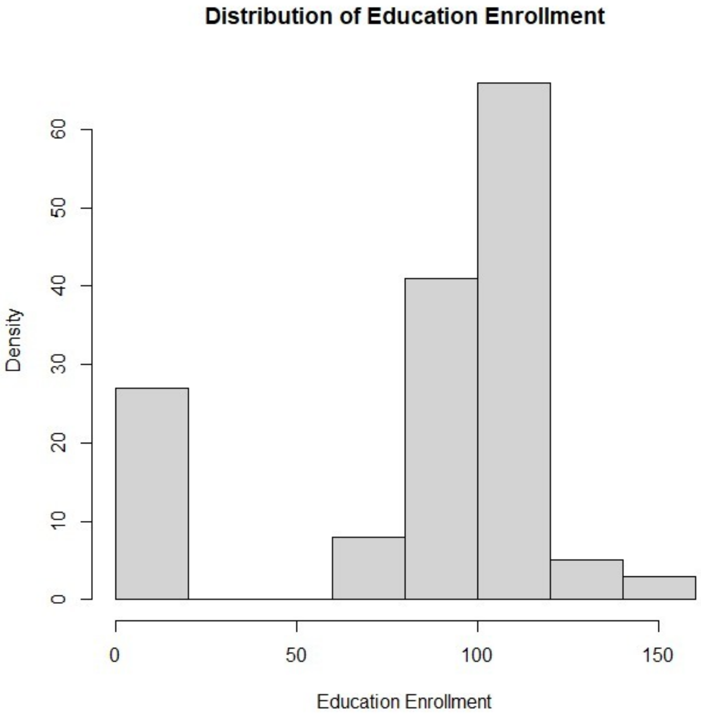


Next, we looked at a **Histogram** through the **base plotting system** of Density against RnD expenditure for the top 10 countries to judge the overall scale of RnD expenditure by the countries and found that very few countries spent the **highest and the lowest** on RnD while the majority spent along the mean at a reasonable deviation.

- **Education Enrollment**

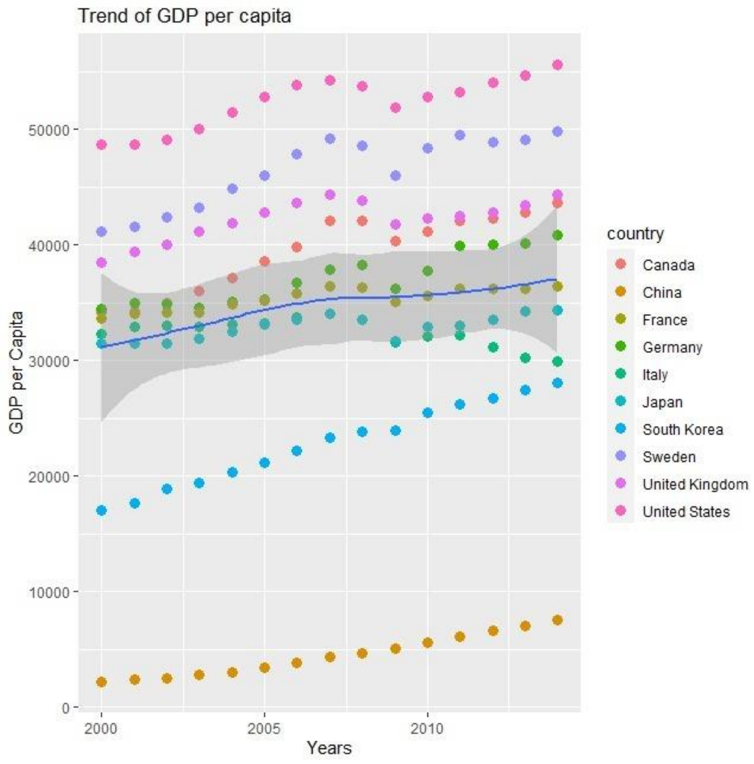


We further analyzed other Macro variables to judge their effects on the number of patents of a country. We noticed that although most of the countries lied **fairly high** on the graph with a few such as **China having a sharp increase in value**, there were still countries like the United States and South Korea that lied at the very bottom of the graph. So, through the visualization of this graph, it seems that the Education enrollment **does not have an effect** on the number of patents.

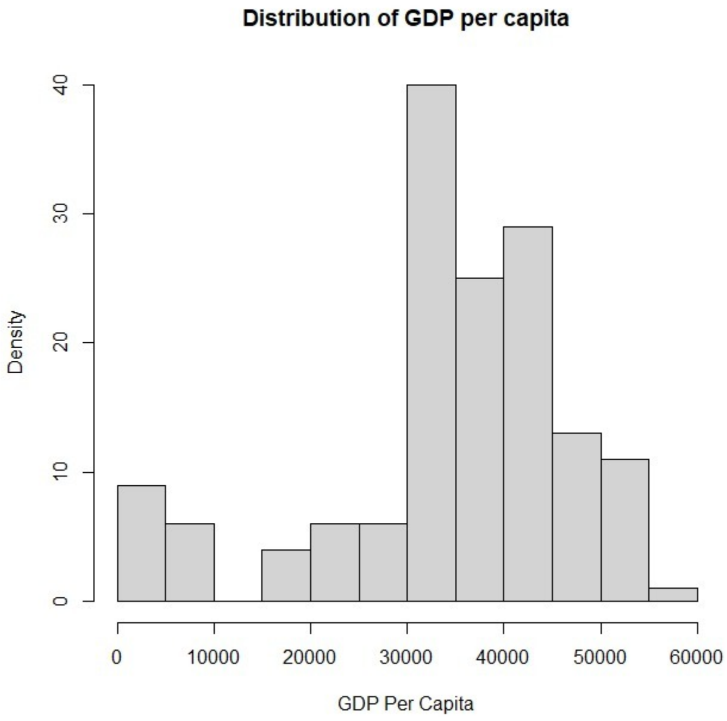


To further analyze this distribution of Education Enrollment, we plotted a histogram in the base plotting system. This made us visualize that indeed a few countries have very low education enrollment while the majority have **fairly high** Education Enrollment levels.

- **GDP per Capita**

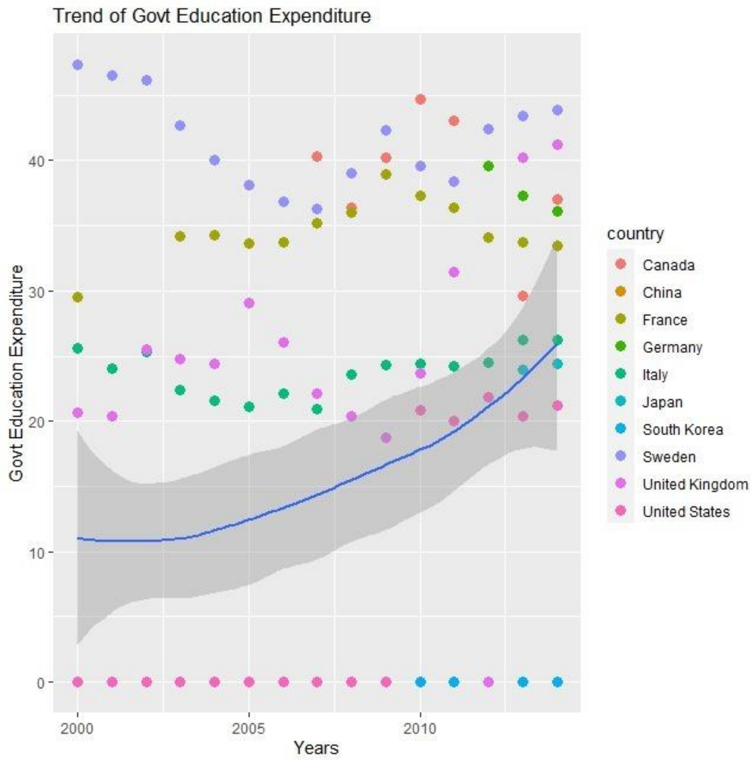


Next, we looked at the ggplot2's scatterplot of the macro factor of trend of GDP per capita. This, again, showed a majority of the countries in the upper part of the graph while **China seemed to be at the very bottom** of the graph.

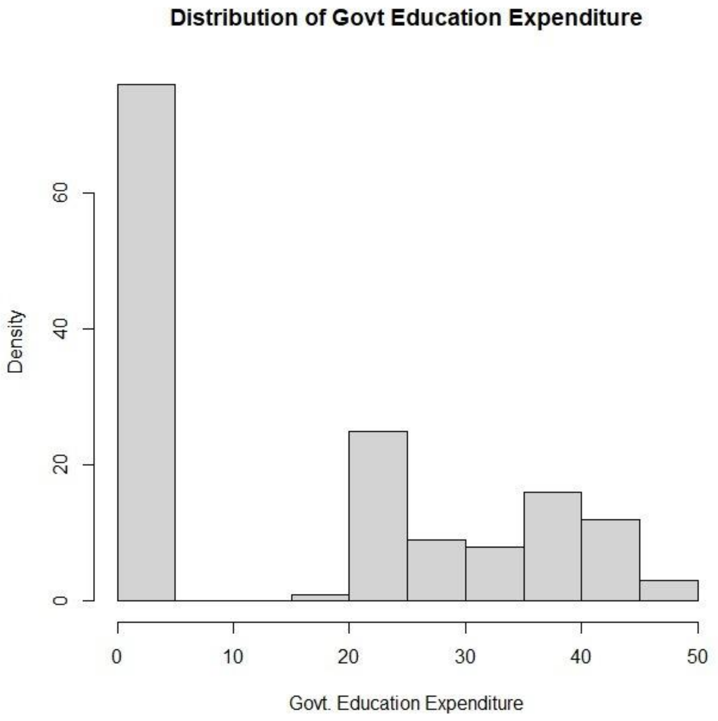


To visualize this factor better, we looked at the distribution of the GDP per capita through a base plot histogram. This again showed that the majority of the top 10 countries did **fairly well** whereas a few countries performed quite worse.

- **Government Education Expenditure**

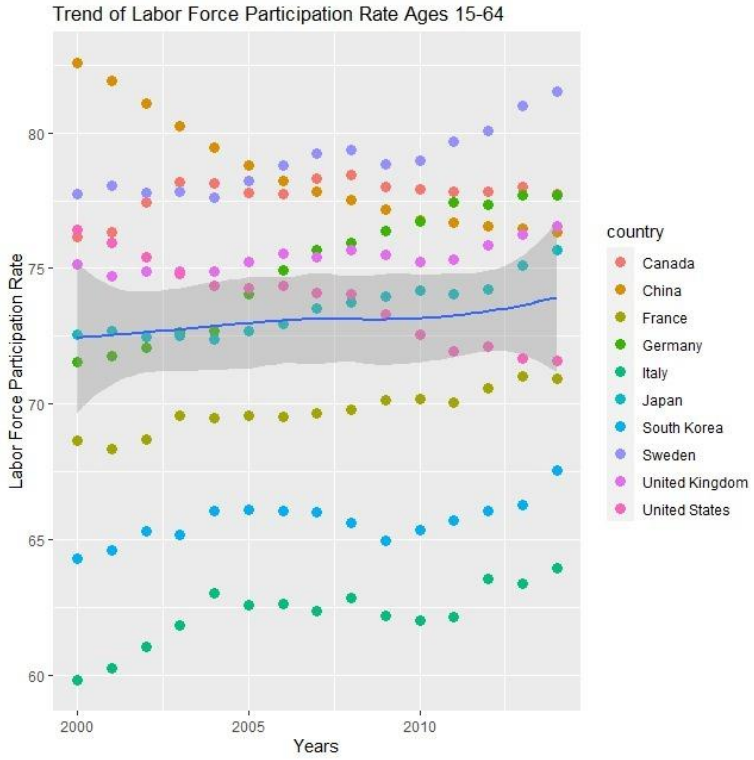


We then moved to the next variable of Trend of Government Education Expenditure and visualized Government Education Expenditure by Year for the top 10 countries through a ggplot2 scatterplot. This again showed a majority of the countries performing **fairly well** whereas a few were at the very bottom of the graph.

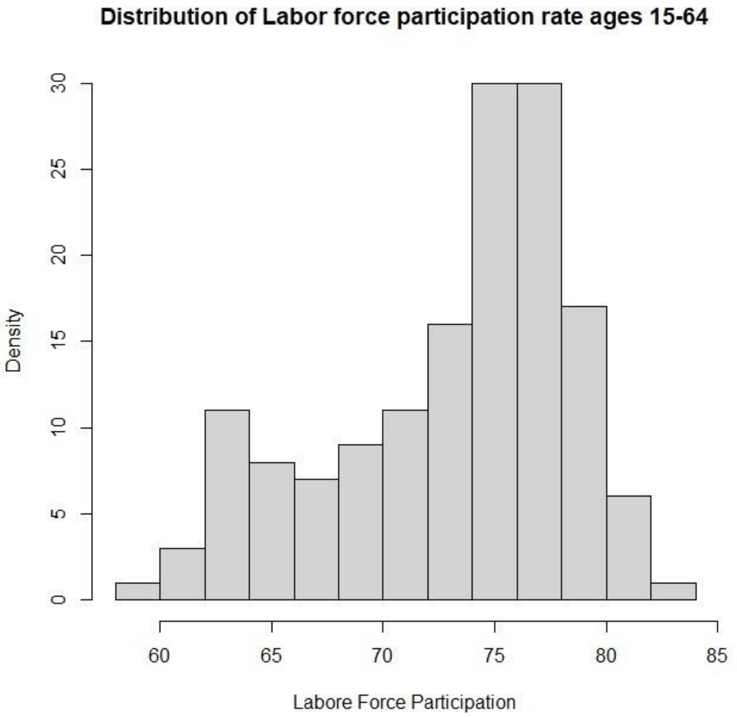


To analyze this variable further, we looked at a histogram in the base plotting system of the Density against the Distribution of Government Education Expenditure. This showed a contrasting view of a majority of the countries having very low Govt. Education expenditure whereas a **very few spent more**.

- **Labor Force Participation Rate Ages 15-64**

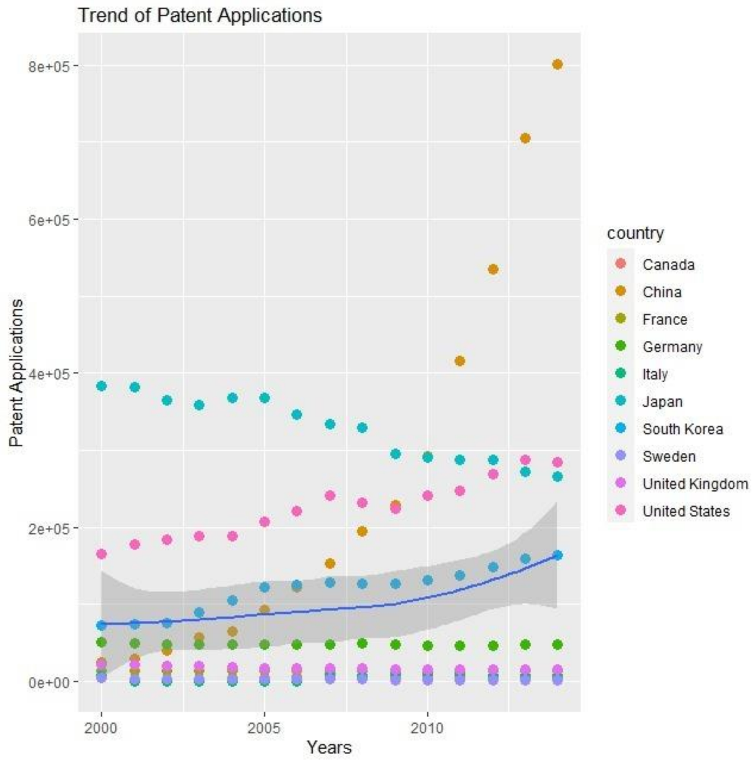


The graph depicts the average labor force participation rates in countries with top 10 highest patents from the year 2000 to 2014. The **blue line depicts** the **mean** labor participation rate, which is approximately **72.5%.** It can be seen there has been a **general increase** in labor force participation from the year 2000 to 2014.

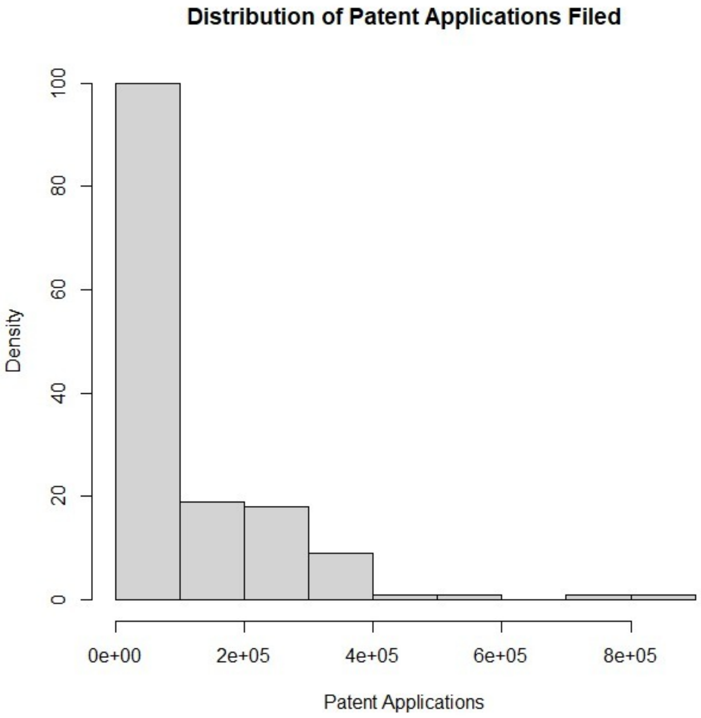


The graph shows distribution of average labour force participation by age in the top 10 countries. It can be seen the graph is **left skewed.**

- **Number of Patent Applications**

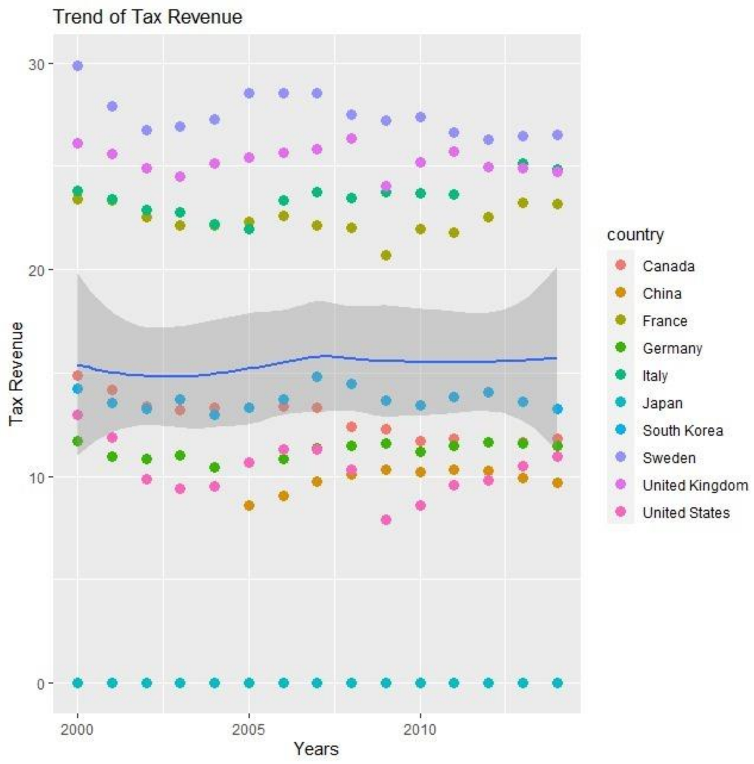


The graph depicts the number of patent applications in the top 10 countries from 2000 to 2014. The average number of patent applications depicted by the blue line can be seen to be in an **upward trend from 2000 onwards**, which means the number of applications each year has increased at a **steady pace.**

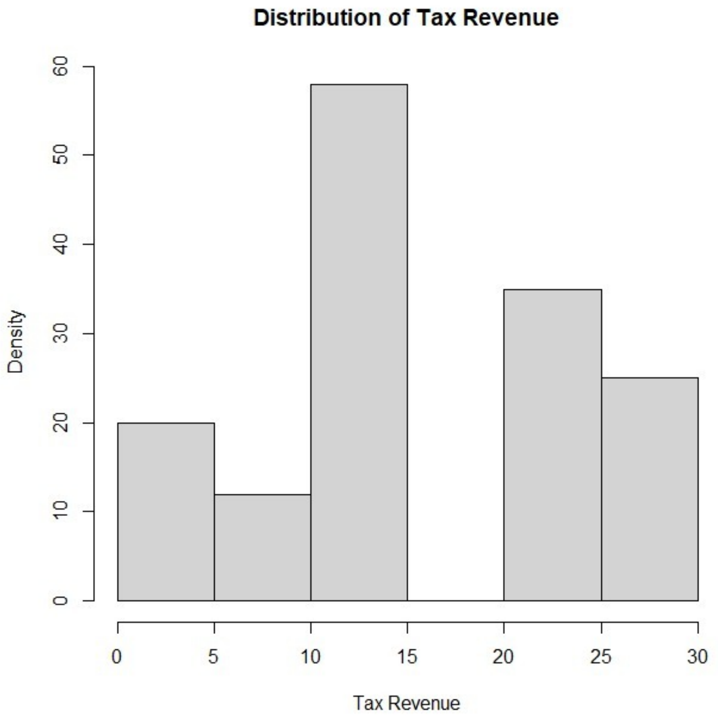


The graph shows that the distribution is **highly skewed towards the right side** which means that in some years there have been **abnormally high** patent applications.

- **Tax Revenue**

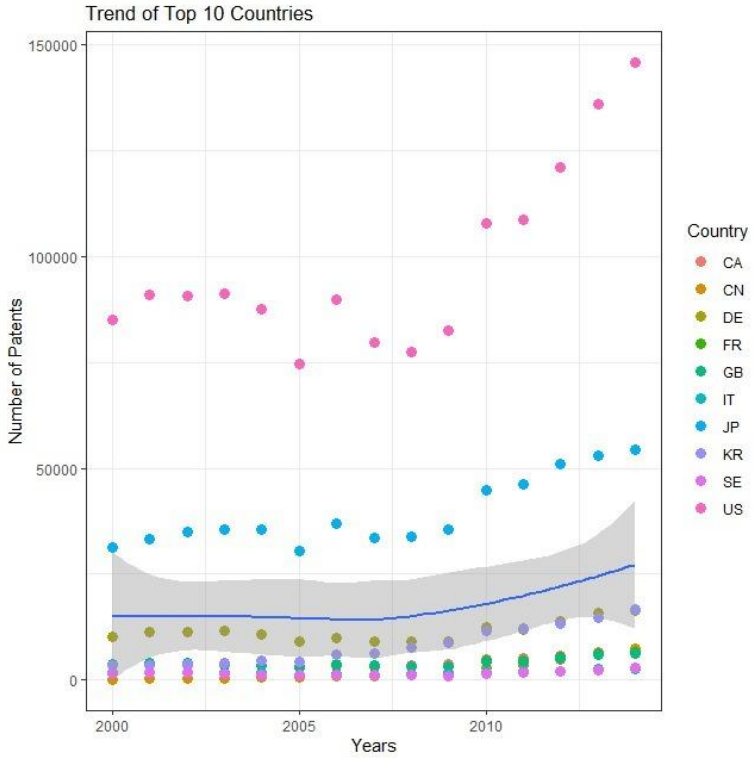


The graph shows the trend in tax revenue as a percentage of GDP from the year 2000 to 2014. The tax revenue collection as a percentage of GDP has remained **fairly constant** from year 2000 to 2014. There are significant outliers towards the upper and lower side as well.



The distribution of tax revenue can be said to be **slightly skewed towards the left side** with higher tax revenues in the later years. The tax revenue collection as a percentage of GDP has been generally on the higher side.

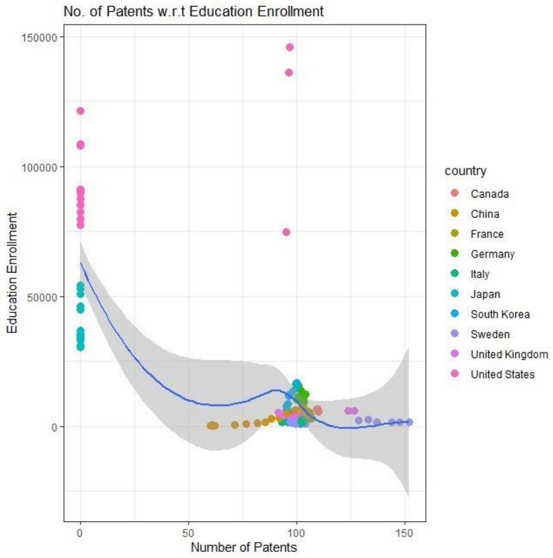
- **Trend of Top 10 countries patents**



The graph shows that the average number of patents in the top 10 countries with highest patents has shown a **steady upward trend** from year 2000 to 2014, with the US and Japan being outliers.

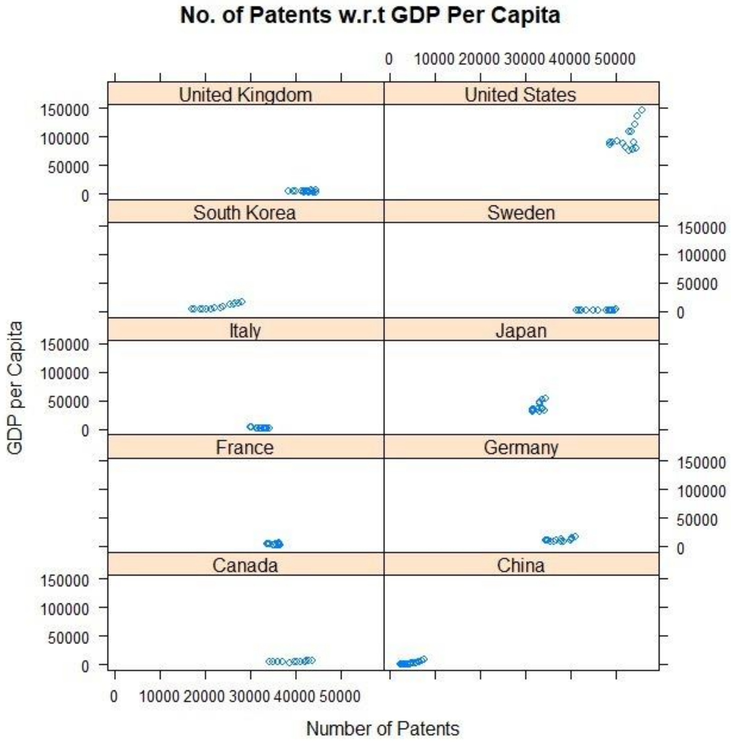
Examining the relation of Number of Patents against all independent variables for Top 10 Countries:

- **Number of Patents with respect to education enrollment**



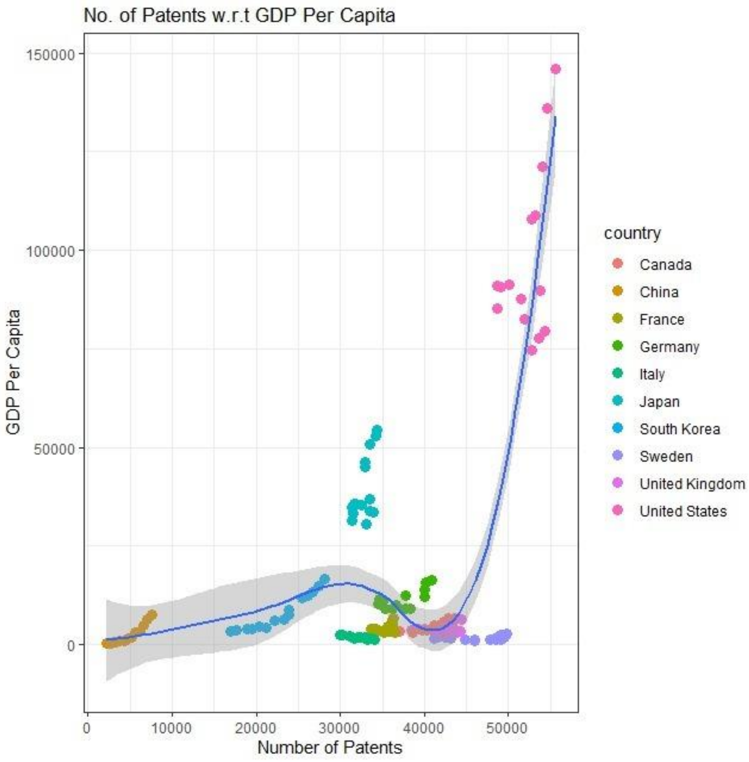
The graph shows that as education enrollment depicted on the y-axis decreases, the number of patents decreases, with few countries being **outliers** to this trend like **the United States**, in the case of Japan, the education enrollment has shown a downward trend but there has **not been any significant change** in the number of patents.

- **Number of patents w.r.t to GDP Per Capita (Lattice)**



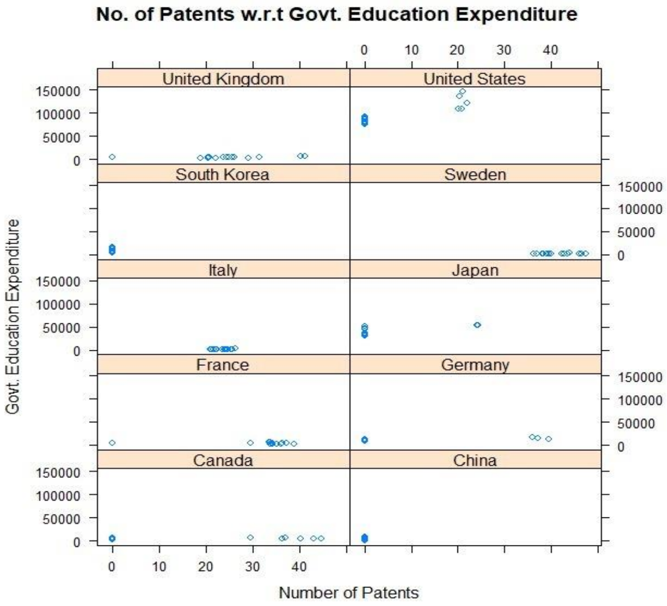
The United Kingdom has more than 40,000 patents, however, the GDP per capita is very low, compared to the United States, which not only has the highest number of patents but also has a very high GDP per capita. Compared to these, within the top 10 countries with the most patents, China ranks as the lowest – relatively lowest in terms of the number of patents and the GDP per Capita.

- **Number of patents w.r.t to GDP Per Capita (GGPlot)**



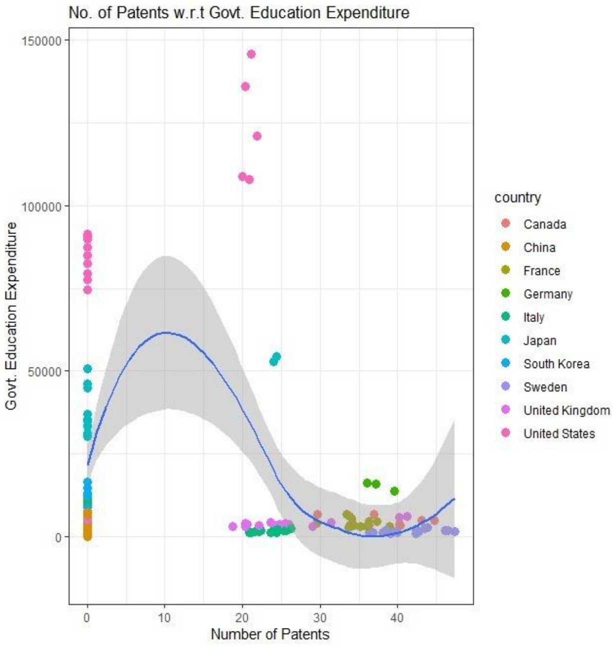
This shows that the US has the highest number of patents and also has the highest GDP per Capita, while China has the lowest, in either of the markers. The use of ggplot makes it easier to determine this when compared to an xyplot, and also provides a better way of comparing the countries.

- **Number of patents w.r.t to Government Education Expenditure (Lattice)**



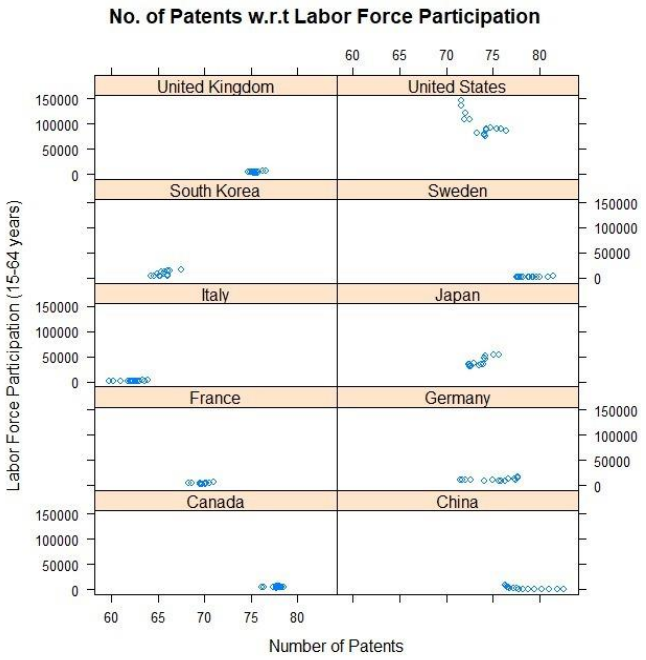
Relative to the amount of GDP that the government spends on education, the number of patents that they have is analyzed. The United States ranks the highest in terms of the expenditure and the number of patents is near 20. Sweden has the most patents but the very low expenditure.

- **Number of patents w.r.t to Government Education Expenditure (GGPlot)**



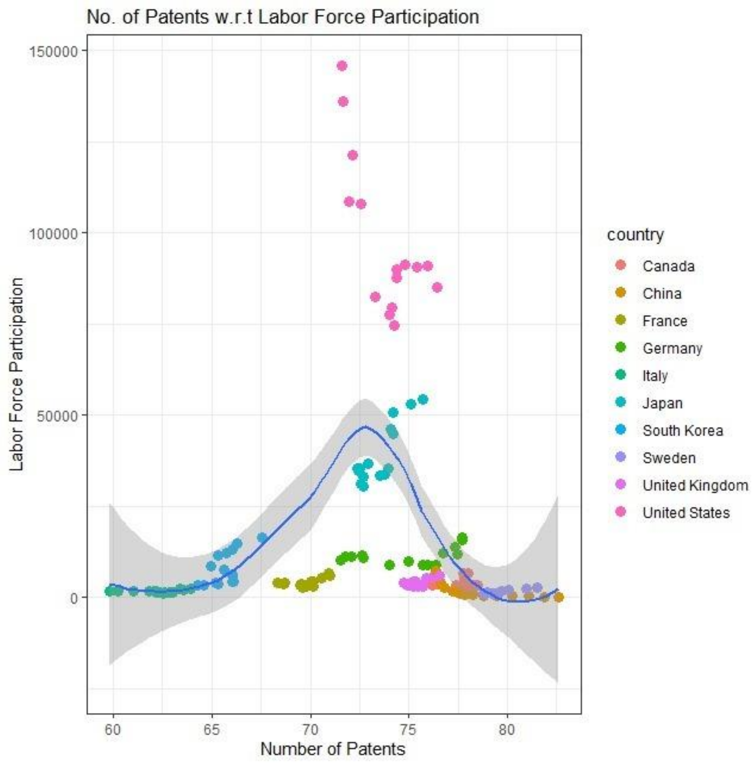
The United States has the highest government expenditure on education. Germany, Sweden, France and China have relatively similar expenditure, but Sweden has the highest number of patents while China has the lowest.

- **Number of patents w.r.t to Labor Force participation for Age Bracket 15-64 (Lattice)**



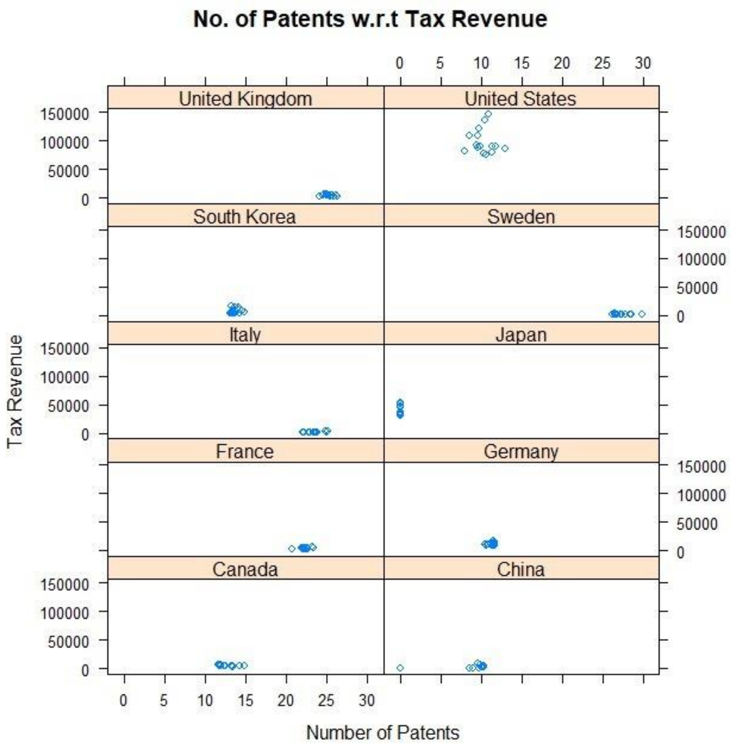
The United States and Japan have a higher rate of labor force participation and also a good number of patents. Compared to these, China has the most patents and the lowest labor force participation, while Italy ranks the lowest in both.

- **Number of patents w.r.t to Labor Force participation for Age Bracket 15-64 (GGPlot)**



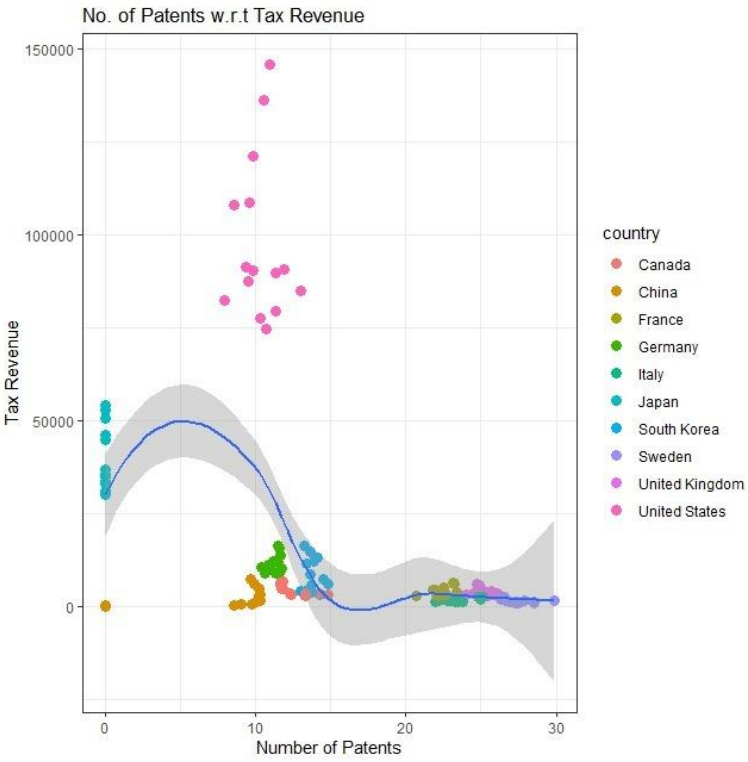
The United States has the highest number of labor force participation, while China and Italy have the lowest. However, Italy has the lowest number of patents while China has the most.

- **Number of patents w.r.t to Tax Revenue (Lattice)**



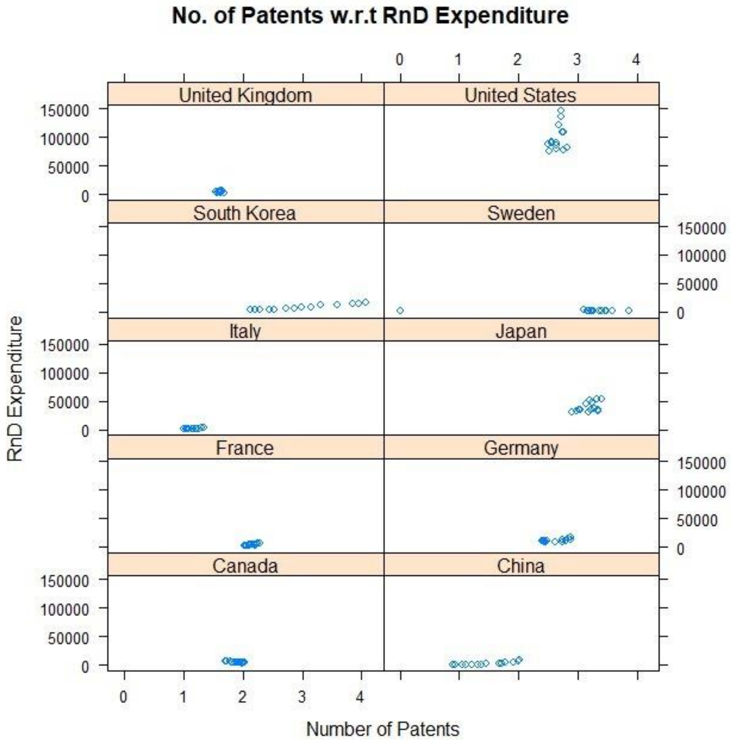
Sweden, Italy, the United Kingdom and France have relatively greater numbers of patents while the tax revenue is low. Comparatively, the United States has the most tax revenue and their number of patents lies between 5 and 15.

- **Number of patents w.r.t to Tax Revenue (GGPlot)**



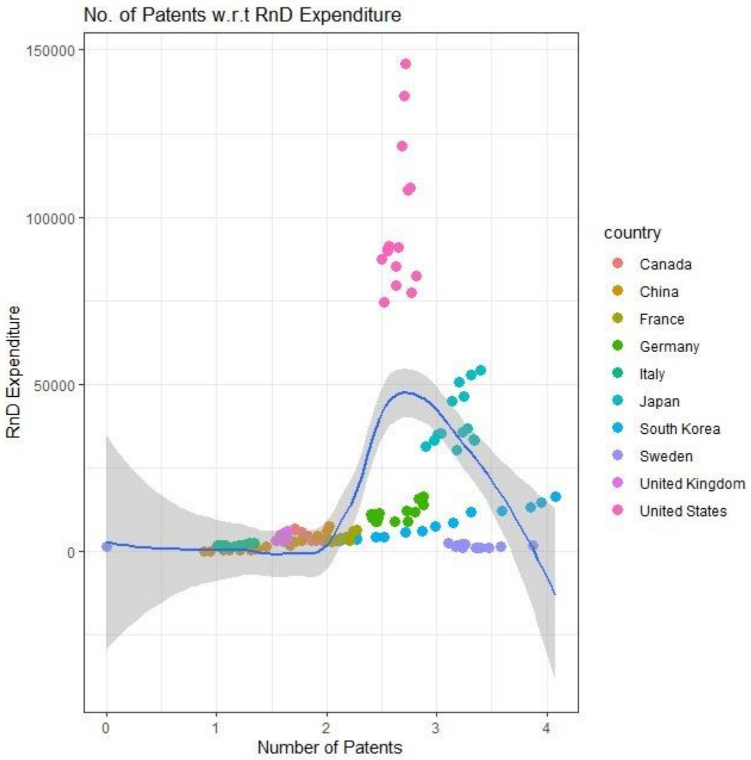
The United States has the highest tax revenue and Sweden has the greatest number of patents. China has the least number of patents and has the least tax revenue as well.

- **Number of patents w.r.t to RnD Expenditure (Lattice)**



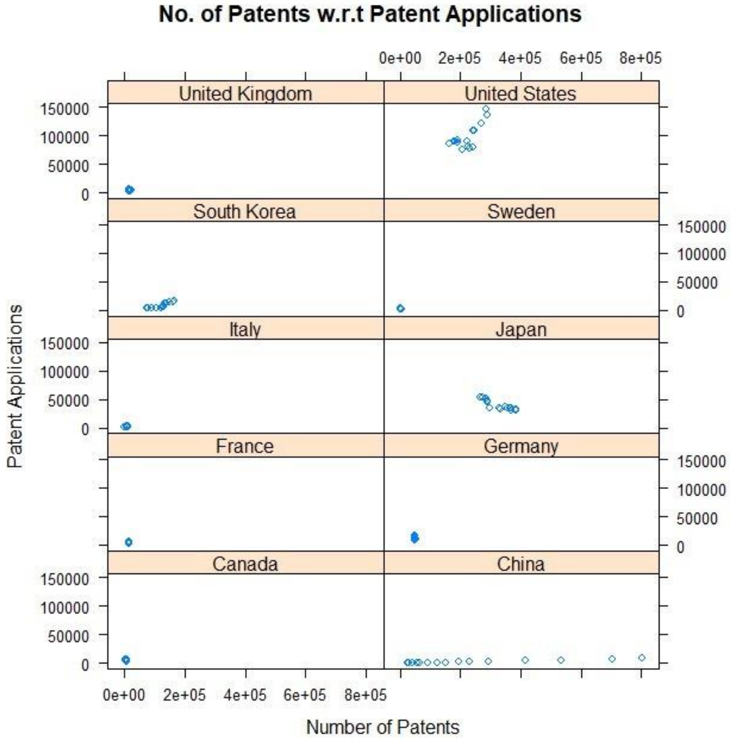
Italy has the greatest spread over the number of patents while Canada, Italy, United Kingdom and France have the lowest and likewise, the RnD expenditure in these countries is relatively lower.

- **Number of patents w.r.t to RnD Expenditure (GGPlot)**



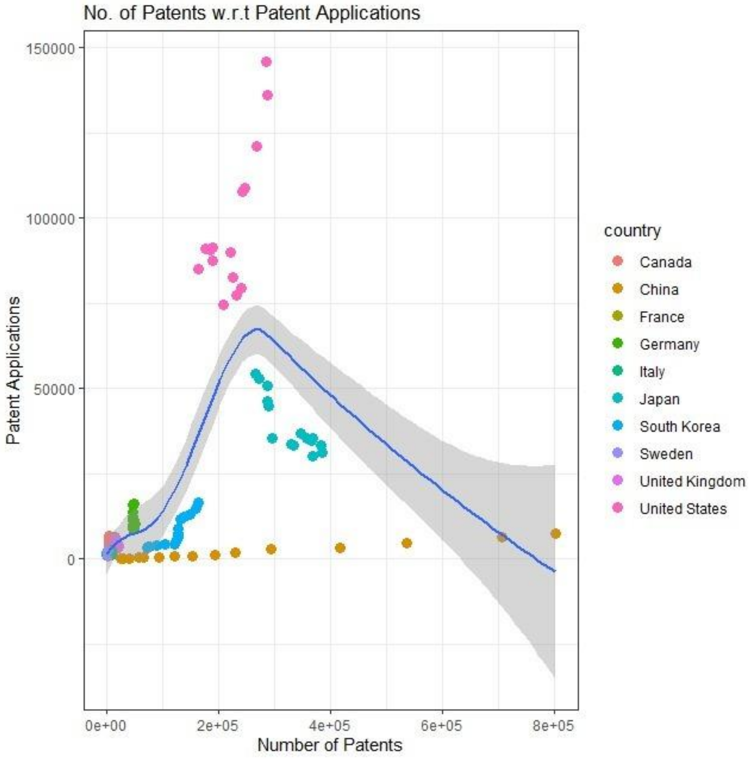
Here it can be seen that the number of patents and RnD expenditure have an increasing rate in Japan, South Korea and Sweden, while the concentration is between 1 and 2.

- **Number of patents w.r.t to patents applications (Lattice)**



China has the most spread over the number of patents and compared to it, the United Kingdom, Italy, France, Canada, Sweden, and Germany have the lowest number of patent applications and the number of patents, while the United States has the highest patent applications.

- **Number of patents w.r.t to patents applications (GGPlot)**



The United States has the highest number of patent applications and China has the lowest, however, there is an increase near the end. Most of the applications are concentrated within the 0e+00 and 4e+05 region.

Statistical Analysis

To test our main control variable that is government spending on research and development, we converted the dataset into a **panel data**, since it contained collection of variable values across 10 countries over an even time intervals

Taking **log function**:

By taking the summary of our dataset, and gauging distribution of control variable values by plotting **histogram**, we observed that our dependent variable (Number of Patents) and a control variable (Patent applications) had an abnormal distribution of values. This was due to the **large range** of the data; therefore, we took the log function to eliminate the effects of outliers and control the range.

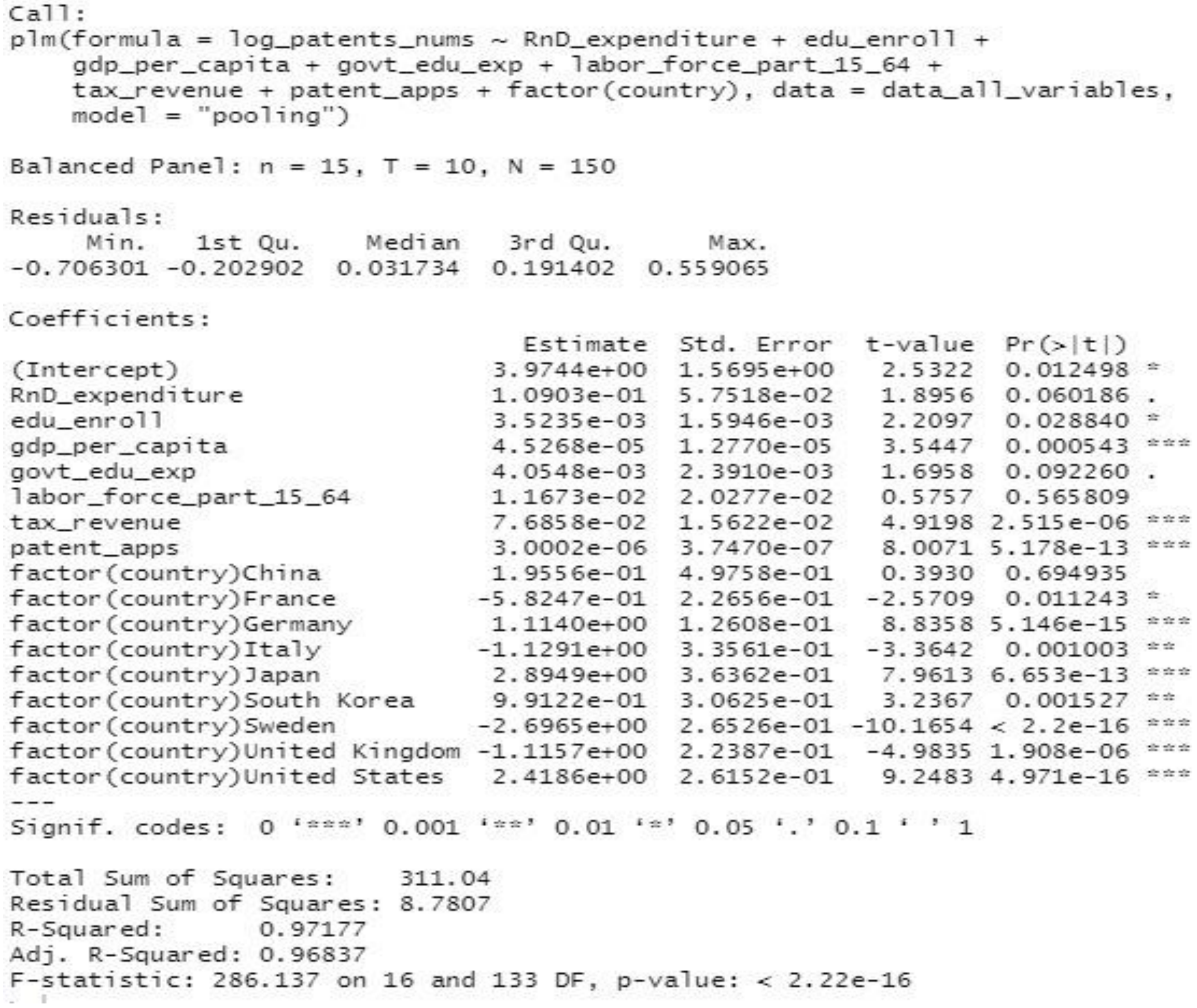
Use of correlation tests:

As a pre-regression step for our hypothesis model, we used the **correlation test** to gauge the correlation between the number of patents (dependent variable) and our control variables (independent variables) to test for perfect collinearity. Its an important step before regression since, according to **Gauss Markov assumption** of Ordinary Least Squares (OLS) method, our variables should never be perfectly correlated. The test results indicated an acceptable range of correlation; values for all control variables with respect to number of patents had a value between the range of -1 and 1. Thus, we could use all our variables in the regression model.

Regression

We ran a total of 3 regressions. One with **two-way fixed effects**, and the other two using **least squares dummy variable** model with one model taking country as a factor and the other taking year as a factor.

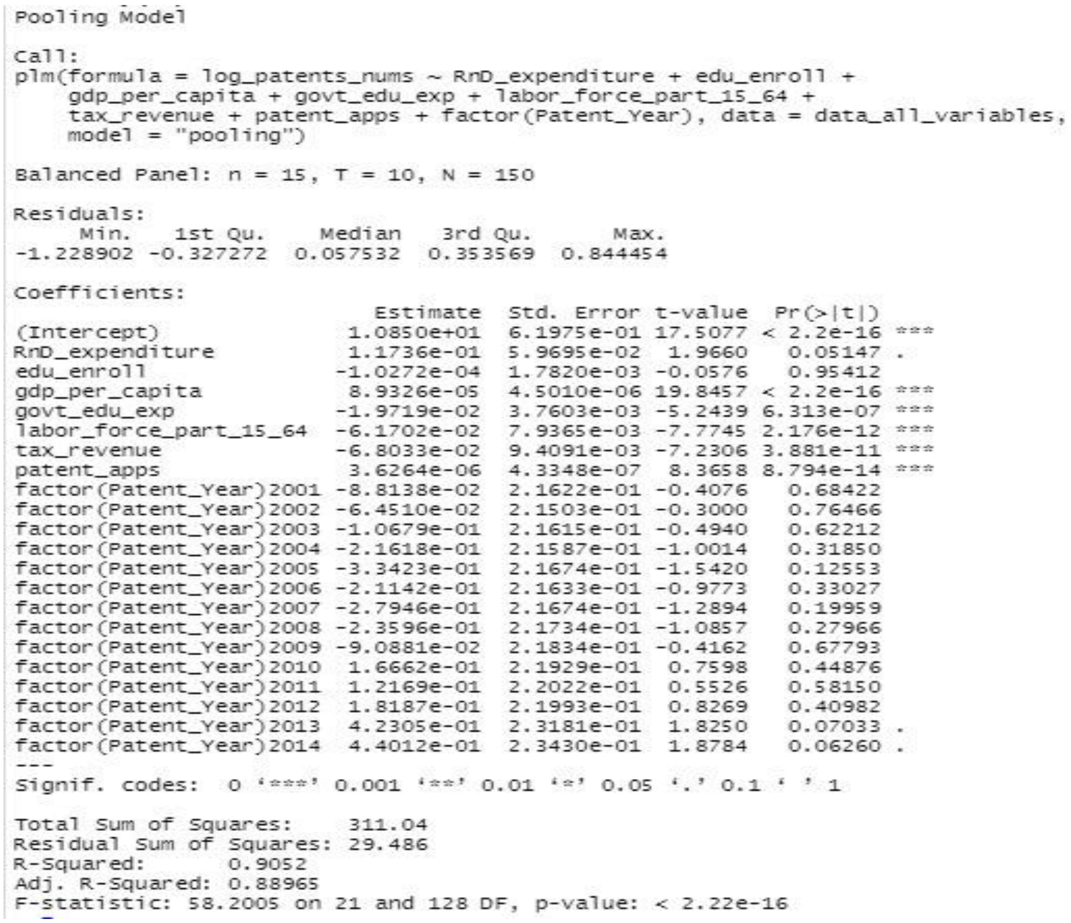
Least squares dummy variable model with country as a factor



```
Call:
plm(formula = log_patents_nums ~ RnD_expenditure + edu_enroll +
    gdp_per_capita + govt_edu_exp + labor_force_part_15_64 +
    tax_revenue + patent_apps + factor(country), data = data_all_variables,
    model = "pooling")

Balanced Panel: n = 15, T = 10, N = 150

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.706301 -0.202902  0.031734  0.191402  0.559065

Coefficients:
                                 Estimate  Std. Error  t-value  Pr(>|t|)
(Intercept)                    3.9744e+00  1.5695e+00   2.5322  0.012498 *
RnD_expenditure                1.0903e-01  5.7518e-02   1.8956  0.060186 .
edu_enroll                     3.5235e-03  1.5946e-03   2.2097  0.028840 *
gdp_per_capita                 4.5268e-05  1.2770e-05   3.5447  0.000543 ***
govt_edu_exp                   4.0548e-03  2.3910e-03   1.6958  0.092260 .
labor_force_part_15_64         1.1673e-02  2.0277e-02   0.5757  0.565809
tax_revenue                    7.6858e-02  1.5622e-02   4.9198 2.515e-06 ***
patent_apps                    3.0002e-06  3.7470e-07   8.0071 5.178e-13 ***
factor(country)China           1.9556e-01  4.9758e-01   0.3930  0.694935
factor(country)France         -5.8247e-01  2.2656e-01  -2.5709  0.011243 *
factor(country)Germany         1.1140e+00  1.2608e-01   8.8358 5.146e-15 ***
factor(country)Italy          -1.1291e+00  3.3561e-01  -3.3642  0.001003 **
factor(country)Japan           2.8949e+00  3.6362e-01   7.9613 6.653e-13 ***
factor(country)South Korea     9.9122e-01  3.0625e-01   3.2367  0.001527 **
factor(country)Sweden         -2.6965e+00  2.6526e-01 -10.1654 < 2.2e-16 ***
factor(country)United Kingdom -1.1157e+00  2.2387e-01  -4.9835 1.908e-06 ***
factor(country)United States   2.4186e+00  2.6152e-01   9.2483 4.971e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    311.04
Residual Sum of Squares: 8.7807
R-Squared:      0.97177
Adj. R-Squared: 0.96837
F-statistic: 286.137 on 16 and 133 DF, p-value: < 2.22e-16
```

Though the model absorbs the results according to each country, the results obtained are inefficient since it contains a lot of dummy variables. This clouds up the whole interpretation of the model and therefore it becomes vague and so we had to discard it.

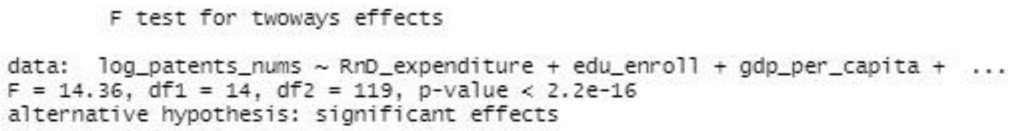
Least squares dummy variable model with year as a factor



```
Pooling Model

Call:
plm(formula = log_patents_nums ~ RnD_expenditure + edu_enroll +
    gdp_per_capita + govt_edu_exp + labor_force_part_15_64 +
    tax_revenue + patent_apps + factor(Patent_Year), data = data_all_variables,
    model = "pooling")

Balanced Panel: n = 15, T = 10, N = 150

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-1.228902 -0.327272  0.057532  0.353569  0.844454

Coefficients:
                            Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)               1.0850e+01  6.1975e-01 17.5077 < 2.2e-16 ***
RnD_expenditure           1.1736e-01  5.9695e-02  1.9660   0.05147 .
edu_enroll               -1.0272e-04  1.7820e-03 -0.0576   0.95412
gdp_per_capita            8.9326e-05  4.5010e-06 19.8457 < 2.2e-16 ***
govt_edu_exp             -1.9719e-02  3.7603e-03 -5.2439 6.313e-07 ***
labor_force_part_15_64   -6.1702e-02  7.9365e-03 -7.7745 2.176e-12 ***
tax_revenue              -6.8033e-02  9.4091e-03 -7.2306 3.881e-11 ***
patent_apps               3.6264e-06  4.3348e-07  8.3658 8.794e-14 ***
factor(Patent_Year)2001 -8.8138e-02  2.1622e-01 -0.4076   0.68422
factor(Patent_Year)2002 -6.4510e-02  2.1503e-01 -0.3000   0.76466
factor(Patent_Year)2003 -1.0679e-01  2.1615e-01 -0.4940   0.62212
factor(Patent_Year)2004 -2.1618e-01  2.1587e-01 -1.0014   0.31850
factor(Patent_Year)2005 -3.3423e-01  2.1674e-01 -1.5420   0.12553
factor(Patent_Year)2006 -2.1142e-01  2.1633e-01 -0.9773   0.33027
factor(Patent_Year)2007 -2.7946e-01  2.1674e-01 -1.2894   0.19959
factor(Patent_Year)2008 -2.3596e-01  2.1734e-01 -1.0857   0.27966
factor(Patent_Year)2009 -9.0881e-02  2.1834e-01 -0.4162   0.67793
factor(Patent_Year)2010  1.6662e-01  2.1929e-01  0.7598   0.44876
factor(Patent_Year)2011  1.2169e-01  2.2022e-01  0.5526   0.58150
factor(Patent_Year)2012  1.8187e-01  2.1993e-01  0.8269   0.40982
factor(Patent_Year)2013  4.2305e-01  2.3181e-01  1.8250   0.07033 .
factor(Patent_Year)2014  4.4012e-01  2.3430e-01  1.8784   0.06260 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    311.04
Residual Sum of Squares: 29.486
R-Squared:      0.9052
Adj. R-Squared: 0.88965
F-statistic: 58.2005 on 21 and 128 DF, p-value: < 2.22e-16
```

A similar problem exists while regressing with year as a factor. Although, the model estimates for the individual factors of each year on the increase/decrease of the number of patents, the estimates obtained are inefficient and biased.

Therefore, we resorted to using the two-way fixed effects model of panel data regression since it controls for the effects of fixed error terms in our model. A fixed error term includes factors that are time-invariant, meaning that they do not vary over-time. Such fixed errors in our model may include the cultural differences between countries in our dataset and population demographics etc. that changes over decades. By removing such fixed effects, we made sure that our assessment of the net effect of R&D expenditure on the number of patents is accurate. To statistically prove this, we used the pf-test:

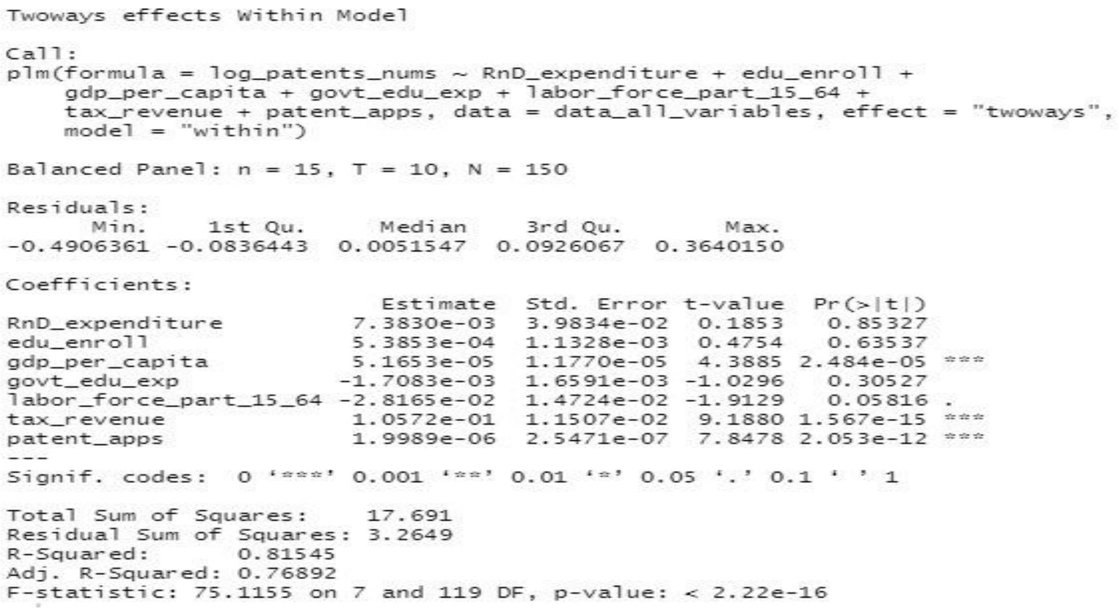


```
        F test for twoways effects

data:  log_patents_nums ~ RnD_expenditure + edu_enroll + gdp_per_capita +  ...
F = 14.36, df1 = 14, df2 = 119, p-value < 2.2e-16
alternative hypothesis: significant effects
```

*The **pf-test** rejects the null hypothesis (The two-ways fixed effects model is not a better estimator than the least squares dummy variable model)*

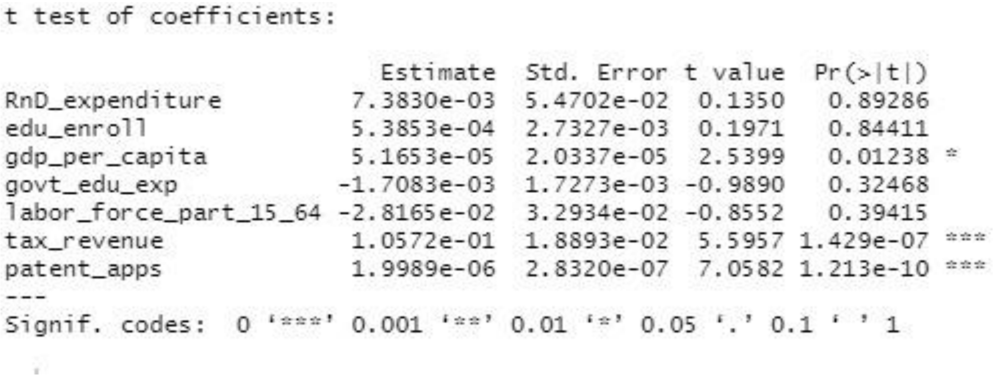
Hypothesis Testing:

H$_0$:*β$_1$*: government expenditure on R&D has no effect on the number of patents



```
Twoways effects Within Model

Call:
plm(formula = log_patents_nums ~ RnD_expenditure + edu_enroll +
    gdp_per_capita + govt_edu_exp + labor_force_part_15_64 +
    tax_revenue + patent_apps, data = data_all_variables, effect = "twoways",
    model = "within")

Balanced Panel: n = 15, T = 10, N = 150

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.4906361 -0.0836443  0.0051547  0.0926067  0.3640150

Coefficients:
                          Estimate  Std. Error t-value  Pr(>|t|)
RnD_expenditure         7.3830e-03  3.9834e-02  0.1853   0.85327
edu_enroll              5.3853e-04  1.1328e-03  0.4754   0.63537
gdp_per_capita          5.1653e-05  1.1770e-05  4.3885 2.484e-05 ***
govt_edu_exp           -1.7083e-03  1.6591e-03 -1.0296   0.30527
labor_force_part_15_64 -2.8165e-02  1.4724e-02 -1.9129   0.05816 .
tax_revenue             1.0572e-01  1.1507e-02  9.1880 1.567e-15 ***
patent_apps             1.9989e-06  2.5471e-07  7.8478 2.053e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    17.691
Residual Sum of Squares: 3.2649
R-Squared:      0.81545
Adj. R-Squared: 0.76892
F-statistic: 75.1155 on 7 and 119 DF, p-value: < 2.22e-16
```

For more accurate estimates, we used the **robust standard errors** which eliminates unbiasedness from the model and provides for more sound value for estimators.



```
t test of coefficients:

                          Estimate  Std. Error t value  Pr(>|t|)
RnD_expenditure         7.3830e-03  5.4702e-02  0.1350   0.89286
edu_enroll              5.3853e-04  2.7327e-03  0.1971   0.84411
gdp_per_capita          5.1653e-05  2.0337e-05  2.5399   0.01238 *
govt_edu_exp           -1.7083e-03  1.7273e-03 -0.9890   0.32468
labor_force_part_15_64 -2.8165e-02  3.2934e-02 -0.8552   0.39415
tax_revenue             1.0572e-01  1.8893e-02  5.5957 1.429e-07 ***
patent_apps             1.9989e-06  2.8320e-07  7.0582 1.213e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results

Since the number of patents considered in regression is in log-form, we will use a log-level interpretation to assess the results of our hypothesis. Regression results provides us with a model p-value < 0.05. It indicates that the model is valid and that no coefficients estimates of our control variables = 0. Upon analyzing the coefficient estimates calculated, the value of β_1 estimates turns out to be 0.00783. It indicates that an increase in one unit of R&D expenditure increases the number of patents by 0.00783%. However, the p-value for the estimate is less than 0.05 thus, we can conclude that the result for this variable is not statistically significant. Therefore, we fail to reject our null hypothesis, which implies that government spending on research and development is not a rightful predictor for the increase/decrease in the number of patents.

On the other hand, our results predicts that patents applications, tax revenue, and Gross domestic product (GDP) per capita of the country has a causal effect on the increase in the number of patents:

- An increase in one unit of GDP per capita increases the number of patents by 0.000052%
- An increase in one unit of tax revenue increases the number of patents by 0.11%
- A unit percentage increase in patent applications increases the number of patents by 0.000002%

Research Implications

The R&D expenditure not having a direct effect on the number of patents in a country can mean many things. One of the possibilities can be that the allocation of the R&D expenditure is not being done properly. This suggests that the allocation of R&D resources should be done in a different way that is perhaps more advantageous to individuals to promote innovative thinking which would eventually result in an increase in the number of patents in a country.

Also, our study sets up a platform for future research related to this domain where the effect of R&D expenditure on the number of patents can be tested with a time lag as the effect of the expenditure is not instantaneously portrayed in the number of patents, rather, the results of such investments show up after a few years.

Moreover, it can also mean that patents, that are generally a result of innovation are not impacted by R&D expenditure because an individual aspiring to innovate something does not benefit from the R&D costs, rather, is limited by the ease of applying for a patent.

This brings us onto our next point. A control variable that did affect the number of patents was the patents applications submitted. This can have a number of different applications. For one, it is made very evident that an increase in the number of applications increases the number of patents, hence, the ease of applying for a patent should be increased in a country. Complicated procedures and barriers to entries in the form of excessive documentation may lead to hesitation in applying for patents which may result in the wastage of several ideas that could have resulted in approved patents set up for great success. Furthermore, it can also be interpreted in the sense of creating more avenues of opportunities for people to come up with unique ideas that can result in the registration of a patent. These can be incubation centers or pitching avenues where new ideas are encouraged, and an environment is created that is conducive for innovation that would eventually result in an increased number of patents.

We also came to the conclusion that Tax Revenue had the greatest impact on the number of patents in a country. This is very obvious in terms of greater economic activity leading towards more opportunities and avenues for growth where the number of patents increases drastically. Through another perspective though, there are other implications too that create this increased tax revenue that would be conducive towards the number of patents in a country. These would be a stable system in the country that would have set governmental procedures that are accommodating for the population. This creates an environment that is easier to thrive in that not only promotes already established businesses but even the new entrants or individuals who are thinking of entering the market. These factors might seem small, but they come together to promoting a tax-giving culture that aids in many things. Not only are the people cocooned in a conductive environment, but the government also has a greater revenue to spend on facilitating businesses and new entrepreneurs as the total percentage of the country's revenue spent on these services decreases even when in absolute terms this amount increases. All this would eventually lead to a greater number of innovative ideas and correspondingly, a greater number of patents.

An increase in the GDP per capita also had an effect on the increase in the number of patents in our research. Although this factor might seem too obvious, but we believe that there are several implications that come as a result of this. Firstly, an increased GDP per capita leads to greater economic activity that results in greater confidence to businesses to innovate and take risks that eventually result in new and improved products that increase the number of patents. Furthermore, this also implies that the number of patents in a country are not just limited to the size or the population of a country but the opportunities that the country provides and the level of satisfaction that the citizens get out of it. All in all, the size of the country does not matter but the conditions of the people it inhabits do.

Overall, we believe that the implications of our research go beyond just the number of patents that a country produces but also the implicit and explicit factors that result these numbers to occur. The variables that we have established to have a relationship with the number of patents in a country promote an overall conducive business environment that would aid in the thriving of a country both in terms of the satisfaction of its citizens and the milestones that it achieves.

Limitations

1) We have drawn a relationship between our number of variables and the number of patents in an instantaneous manner, herein lag effect could be considered, to provide a more realistic visualization of how a government's expenditure affects growth in terms of patents over the long run. This, however, could not be applied due to the absence of accurate knowledge regarding the time horizons for which R&D expenditure influences the development of a generation which would result in an increase in the number of patents.

2) Research and design expenditure were taken as a percentage of GDP and was taken in relative terms. Thus, if the research and design expenditure is high compared to other countries, and the GDP is significantly higher still, it would result in the percentage of expenditure per revenue to be low. However, if it were taken in absolute terms, then it would have been higher. This would consequently mask the true relation between research and design expenditure and the number of patents that a country has.

3) An analysis of the bottom 10 or middle 10 countries could be incorporated to consider a more holistic viewpoint of what exactly does the relationship between the number of patents in a country and the percentage of GDP it sets for research and design, entail.

4) Due to the bottom 30 countries having the same number of total patents, we were unable to draw conclusions for a lack of development towards increasing the number of patents as we were not able to draw relations over time.

Conclusion

In conclusion, the original hypothesis that we set out to test could not be proved. We set R&D expenditure has no effect on the number of patents as our null hypothesis and were testing whether R&D expenditure positively impacts the development of patents. After running various statistical tests, a statistically significant relations could not be established between R&D expenditures positively impacting the number of patents, hence we were unable to reject the null hypothesis in favor of the alternate hypothesis. However, we did not limit our analysis to only this factor and observed different trends while conducting exploratory data analysis. While dwelling deep into the data and creating relationships between various macroeconomic factors, as independent variables, and the number of patents, as dependent variables, we observed various relationships in the plots that we created. This served as a steppingstone to conduct further statistical analysis between the macroeconomic factors and the number of patents. After conducting the same rigorous testing on the new macroeconomic factors that was performed between R&D expenditure and the number of patents, we were able to find numerous statically significant relations and gave us new insights on to what governments should be doing to improve innovations in their countries and how to allocate resources. All in all, the macroeconomic factors that showed a statistically significant positive relationship on the number of patents were, GDP per capita, tax revenue, and patent applications.

Work Cited

Das, Ramesh Chandra. "Interplays among R&D Spending, Patent and Income Growth: New Empirical Evidence from the Panel of Countries and Groups." *Journal of Innovation and Entrepreneurship*, vol. 9, no. 1, 2020, https://doi.org/10.1186/s13731-020-00130-8.

"Public Funding for Innovation and Research Cooperation." *Managing Open Innovation*, https://doi.org/10.4337/9781781953594.00012.

"The Impact of Public R&D Expenditure on Business R&D." *OECD Science, Technology and Industry Working Papers*, 2000, https://doi.org/10.1787/670385851815.

Data Sources:

https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS?type=shaded
https://data.worldbank.org/indicator/SE.TER.ENRR?type=shaded
https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?type=shaded
https://data.worldbank.org/indicator/GC.TAX.TOTL.GD.ZS?type=shaded
https://data.worldbank.org/indicator/SL.TLF.ACTI.ZS?type=shaded
https://data.worldbank.org/indicator/IP.PAT.RESD?type=shaded