Project Report

DISC 326-S2
Data Science for Decision Making

Group 8

Abdul Hannan Chaudhry	23110058
Hassan Imran	23110287
Arsalan Awan	23110281
Sameer Raj	23110027
Mantasha Meraj	23110245

PRESENTED TO DR. SALMA ZAMAN

Table of Contents

Introduction

Rationale*:* For this project, we wanted to focus on the patents released from a specific sector and then use that to make a prediction. We originally planned to test whether the number of patents in the Illumination Devices sector predicts Light Pollution in any way. Unfortunately, we did not find an extensive enough dataset on light pollution no matter where we looked, including GitHub and even on websites like kaggle. So instead, we looked at other sectors and decided to focus on the patents in the pharmaceuticals and biotechnology class.

The main focus of our project is to analyze whether a higher number of yearly patents released in the pharmaceutical and biotechnology class can be used to predict changes in average life expectancy, neonatal mortality rates, and infant mortality rates. Mortality rate is generally defined as units of deaths per 1,000 individuals, the data that we have is for the number of deaths per 1,000 live births (Neonatal and Infant).

Most patents from the Pharmaceutical and Biotechnology sector belong to the USA, perhaps it has something to do with the fact that the dataset belonging to USPTO or the intellectual prowess USA has, especially with its large population. Owing to this, we will focus a little more on this country. However, for the purpose of this study, we have created two groups: developing nations and developed nations

Even though some developing nations have little to no patents, we might still include them in order to keep the dataset representative. The rationale behind this is that there are obviously a lot of external variables that influence mortality rates, including a lot of socioeconomic factors such as: health education, health diseases, nutrition, pollution, natural calamities and war e.t.c. It is not to say that developed countries will not suffer from these problems, but they may be more significant in developing countries, and these factors may also affect research in the medical field.

Perhaps the biggest assumption and the basis of our project is that the number of patents a country gets per year reflects how medically advanced that country is, and thus its health services are comparatively better than other countries. This leads to lower mortality rates and higher life expectancies.

Yearly number of patents is chosen as our independent variable. The reason for this is, if the medical field of a country is highly consistent and their number of patents each year either increases or stays constant we would expect the mortality rates to go down and average life expectancy to increase. It is a big assumption on our part to expect that the medical field will deteriorate if the number of their patents decreases leading to higher mortality rates and lower average life expectancies. We generally expect that staying consistent with their work means that the medical field is highly competent, but if they do not stay consistent or their patents start decreasing, the medical field may be becoming less competent and health services are thus deteriorating.

Independent Variable: Yearly Number of Patents

Dependent Variables: Infant Mortality (per 1,000 live births), Neonatal Mortality (per 1,000 live births), Average Life Expectancy

Research Question: Do Yearly Number of Patents predict Infant Mortality, Neonatal Mortality and Average Life Expectancy?

(Note: For the purpose of this project, we tried to incorporate as many teachings of the course as we possibly could, however some things were left out, especially Social Network Analysis)

Getting and Cleaning the Data

As mentioned previously, we divided the countries amongst developing and developed countries through the use of Human Development Index ranking.
(http://hdr.undp.org/en/content/latest-human-development-index-ranking)

Twenty countries were randomly selected and allocated to the developing or developed nations data sets. The division of countries is as follows:

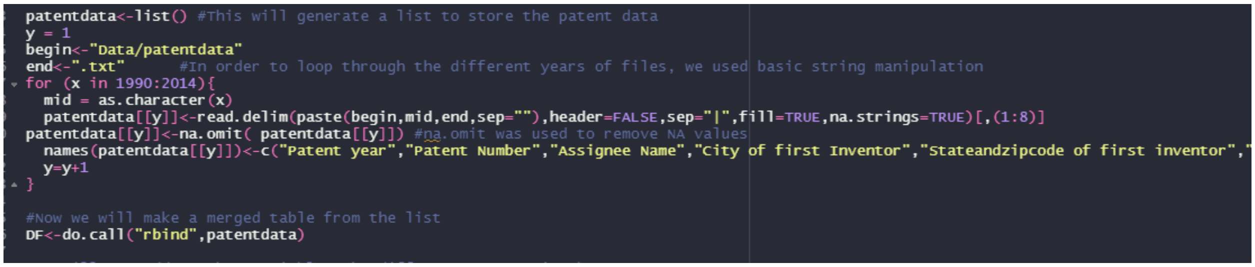
Developed: Norway, Ireland, Switzerland, Germany, Sweden, Australia, Netherlands, Denmark, Finland, Singapore, United Kingdom, Belgium, New Zealand, Canada, United States,, Israel, Japan, France, Italy
Developing: India, Bangladesh, Pakistan, Nepal, Zimbabwe, Uganda, Nigeria, Afghanistan, Kenya, Libya, Lebanon, Kuwait, Madagascar, Cambodia, Cameroon, Syria, Papua New Guinea, Namibia, Bhutan, Tajikistan, China (China is considered a developing nation, GDP per capita is really low)

Patent Data:

The provided dataset of the Patents was divided into text files from 1990 to 2014 . For ease, we merged the dataset. This was done by first generating a list which stores patent data of different years. A variable named 'begin' was created. The data was stored in a folder called "/Data". The variable 'begin' represents this folder, followed by the initial part of the name of the files from the different datasets. The variable 'end' represents the ending part, the ".txt". The part that comes in the middle of the name, the year, is achieved by setting a for loop using a x variable which loops from 1990 to 2014. .as(character) was used on 'x' to get 'mid', which represented the year as a string instead of integers and the paste function was then used along with read.delim to read the patent data for a specific year. Using the variable 'y' we stored the data of the individual years into the list.

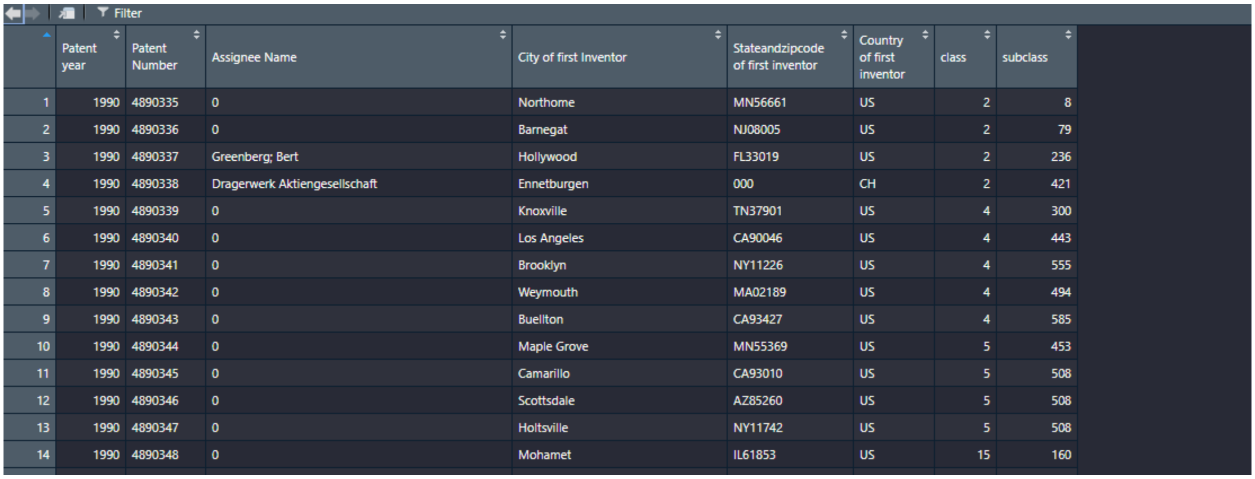
To create a merged table from the list, we used the do.call function which performs the rbind function on the list, which is used to combine rows (or vectors depending on what you use it on).

Figure 1:



```
patentdata<-list() #This will generate a list to store the patent data
y = 1
begin<-"Data/patentdata"
end<-".txt"      #In order to loop through the different years of files, we used basic string manipulation
for (x in 1990:2014){
  mid = as.character(x)
  patentdata[[y]]<-read.delim(paste(begin,mid,end,sep=""),header=FALSE,sep="|",fill=TRUE,na.strings=TRUE)[,(1:8)]
patentdata[[y]]<-na.omit( patentdata[[y]]) #na.omit was used to remove NA values
  names(patentdata[[y]])<-c("Patent year","Patent Number","Assignee Name","City of first Inventor","Stateandzipcode of first inventor","
  y=y+1
}

#Now we will make a merged table from the list
DF<-do.call("rbind",patentdata)
```

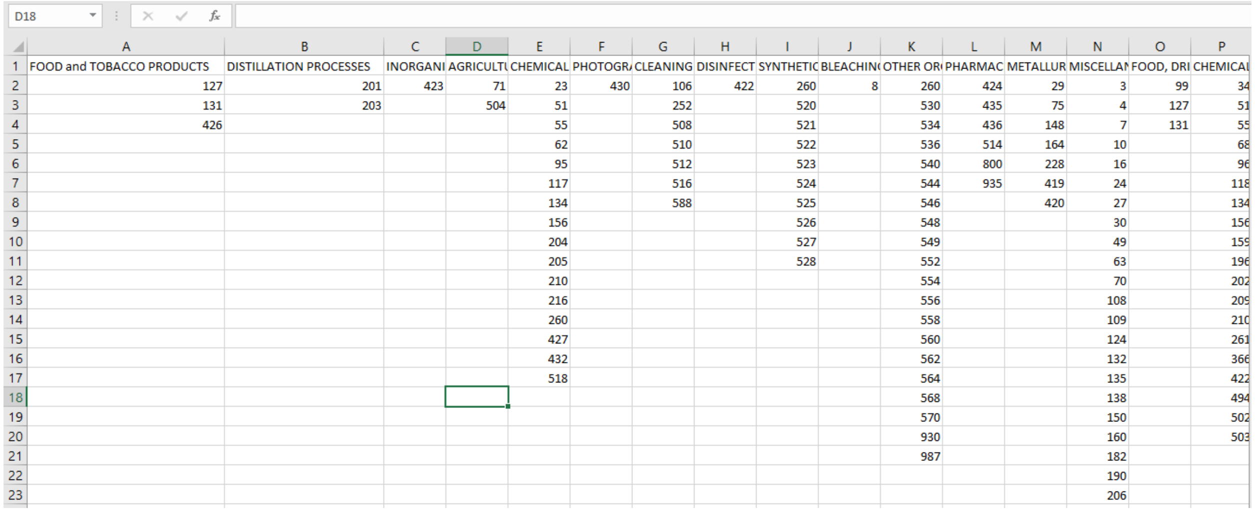
Figure 2:



	Patent year	Patent Number	Assignee Name	City of first Inventor	Stateandzipcode of first inventor	Country of first inventor	class	subclass
1	1990	4890335	0	Northome	MN56661	US	2	8
2	1990	4890336	0	Barnegat	NJ08005	US	2	79
3	1990	4890337	Greenberg; Bert	Hollywood	FL33019	US	2	236
4	1990	4890338	Dragerwerk Aktiengesellschaft	Ennetburgen	000	CH	2	421
5	1990	4890339	0	Knoxville	TN37901	US	4	300
6	1990	4890340	0	Los Angeles	CA90046	US	4	443
7	1990	4890341	0	Brooklyn	NY11226	US	4	555
8	1990	4890342	0	Weymouth	MA02189	US	4	494
9	1990	4890343	0	Buellton	CA93427	US	4	585
10	1990	4890344	0	Maple Grove	MN55369	US	5	453
11	1990	4890345	0	Camarillo	CA93010	US	5	508
12	1990	4890346	0	Scottsdale	AZ85260	US	5	508
13	1990	4890347	0	Holtsville	NY11742	US	5	508
14	1990	4890348	0	Mohamet	IL61853	US	15	160

Next, we created a .csv file which represents a specific sector along with their class numbers from the word document provided named ‘list of 56 fields,’ we left the sub-class numbers out because we will not be using them for this research.

Figure 3:

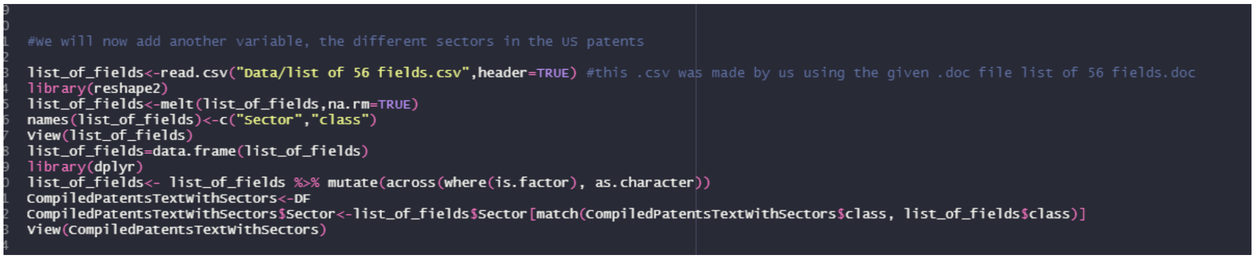


D18

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	FOOD and TOBACCO PRODUCTS	DISTILLATION PROCESSES	INORGANI	AGRICULTU	CHEMICAL	PHOTOGR/	CLEANING	DISINFECT	SYNTHETIC	BLEACHIN	OTHER OR	PHARMAC	METALLUR	MISCELLAN	FOOD, DRI	CHEMICAL
2	127	201	423	71	23	430	106	422	260	8	260	424	29	3	99	34
3	131	203		504	51		252		520		530	435	75	4	127	51
4	426				55		508		521		534	436	148	7	131	55
5					62		510		522		536	514	164	10		68
6					95		512		523		540	800	228	16		96
7					117		516		524		544	935	419	24		118
8					134		588		525		546		420	27		134
9					156				526		548			30		156
10					204				527		549			49		159
11					205				528		552			63		196
12					210						554			70		202
13					216						556			108		209
14					260						558			109		210
15					427						560			124		261
16					432						562			132		366
17					518						564			135		422
18											568			138		494
19											570			150		502
20											930			160		503
21											987			182		
22														190		
23														206		

We corrected the first column's name after reading the CSV file. Next, we used **reshape2**'s melt function to convert from a wide format data to a long format. The resultant columns were named the "Sector" and "class" in order to match the dataframe to the patents data frame (see figure 6). With the help of the **dplyr** library, we mutated the dataframe to make sure any factor values were turned into character values preparing the dataset for the use of the match function (code visible in figure 4) .

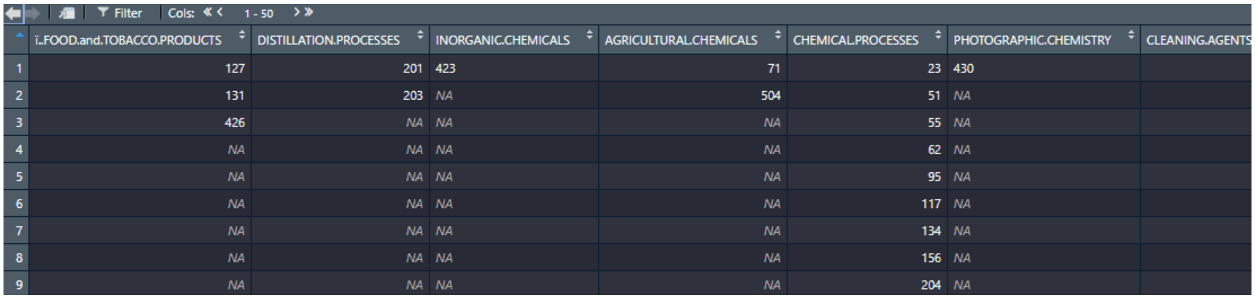
Figure 4:



```
#We will now add another variable, the different sectors in the US patents

list_of_fields<-read.csv("Data/list of 56 fields.csv",header=TRUE) #this .csv was made by us using the given .doc file list of 56 fields.doc
library(reshape2)
list_of_fields<-melt(list_of_fields,na.rm=TRUE)
names(list_of_fields)<-c("Sector","class")
View(list_of_fields)
list_of_fields=data.frame(list_of_fields)
library(dplyr)
list_of_fields<- list_of_fields %>% mutate(across(where(is.factor), as.character))
CompiledPatentsTextWithSectors<-DF
CompiledPatentsTextWithSectors$Sector<-list_of_fields$Sector[match(CompiledPatentsTextWithSectors$class, list_of_fields$class)]
View(CompiledPatentsTextWithSectors)
```

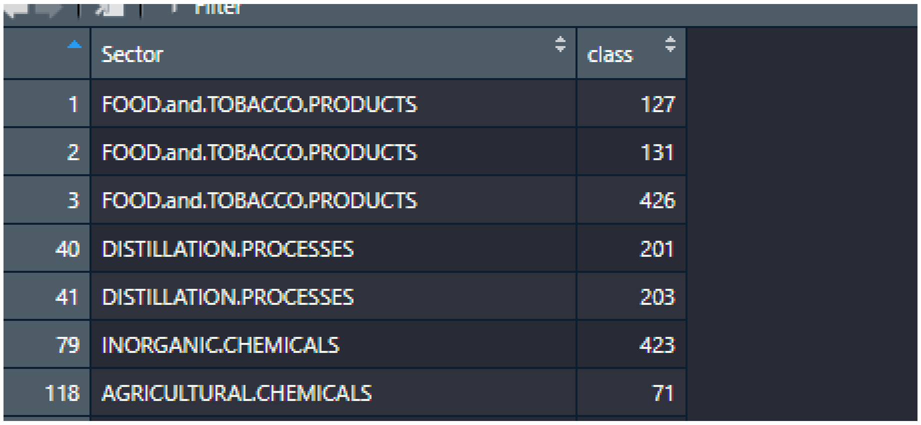
Subsequently, we created a column named 'Sector' in the 'CompiledPatentsWithSectors' dataframe, which adds the sector from the 'list_of_fields' depending on where the class number from the patent data matches with the class number in the list of fields data (see figure 5).

Figure 5:



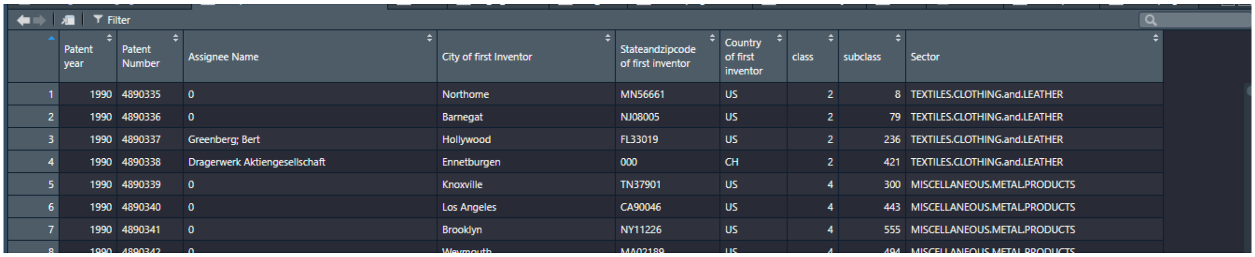
	i..FOOD.and.TOBACCO.PRODUCTS	DISTILLATION.PROCESSES	INORGANIC.CHEMICALS	AGRICULTURAL.CHEMICALS	CHEMICAL.PROCESSES	PHOTOGRAPHIC.CHEMISTRY	CLEANING.AGENTS
1	127	201	423	71	23	430	
2	131	203	NA	504	51	NA	
3	426	NA	NA	NA	55	NA	
4	NA	NA	NA	NA	62	NA	
5	NA	NA	NA	NA	95	NA	
6	NA	NA	NA	NA	117	NA	
7	NA	NA	NA	NA	134	NA	
8	NA	NA	NA	NA	156	NA	
9	NA	NA	NA	NA	204	NA	

Figure 6:



	Sector	class
1	FOOD.and.TOBACCO.PRODUCTS	127
2	FOOD.and.TOBACCO.PRODUCTS	131
3	FOOD.and.TOBACCO.PRODUCTS	426
40	DISTILLATION.PROCESSES	201
41	DISTILLATION.PROCESSES	203
79	INORGANIC.CHEMICALS	423
118	AGRICULTURAL.CHEMICALS	71

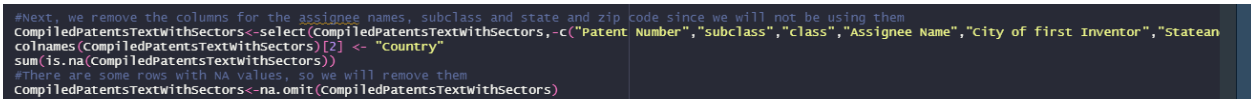
Figure: 7



	Patent year	Patent Number	Assignee Name	City of first Inventor	Stateandzipcode of first inventor	Country of first inventor	class	subclass	Sector
1	1990	4890335	0	Northome	MN56661	US	2	8	TEXTILES.CLOTHING.and.LEATHER
2	1990	4890336	0	Barnegat	NJ08005	US	2	79	TEXTILES.CLOTHING.and.LEATHER
3	1990	4890337	Greenberg; Bert	Hollywood	FL33019	US	2	236	TEXTILES.CLOTHING.and.LEATHER
4	1990	4890338	Dragerwerk Aktiengesellschaft	Ennetburgen	000	CH	2	421	TEXTILES.CLOTHING.and.LEATHER
5	1990	4890339	0	Knoxville	TN37901	US	4	300	MISCELLANEOUS.METAL.PRODUCTS
6	1990	4890340	0	Los Angeles	CA90046	US	4	443	MISCELLANEOUS.METAL.PRODUCTS
7	1990	4890341	0	Brooklyn	NY11226	US	4	555	MISCELLANEOUS.METAL.PRODUCTS
8	1990	4890342	0	Weymouth	MA02189	US	4	494	MISCELLANEOUS.METAL.PRODUCTS

The resulting table is visible above (figure 7).Once this table was created the next step was to remove data that would not directly contribute to our analysis. Columns for the assignee names, subclass, state and zip code, patent number, and class were removed from the table. Next the data was searched for "NA" values using na.omit to scrape any cells with said value. For simplicity, the column 'Country of first inventor' was renamed to 'Country.' The code for this is visible in figure 8.

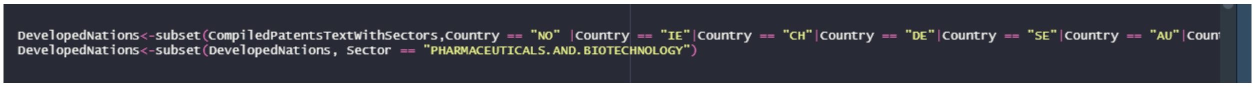
Figure 8:



```
#Next, we remove the columns for the assignee names, subclass and state and zip code since we will not be using them
CompiledPatentsTextWithSectors<-select(CompiledPatentsTextWithSectors,-c("Patent Number","subclass","class","Assignee Name","City of first Inventor","Statean
colnames(CompiledPatentsTextWithSectors)[2] <- "Country"
sum(is.na(CompiledPatentsTextWithSectors))
#There are some rows with NA values, so we will remove them
CompiledPatentsTextWithSectors<-na.omit(CompiledPatentsTextWithSectors)
```

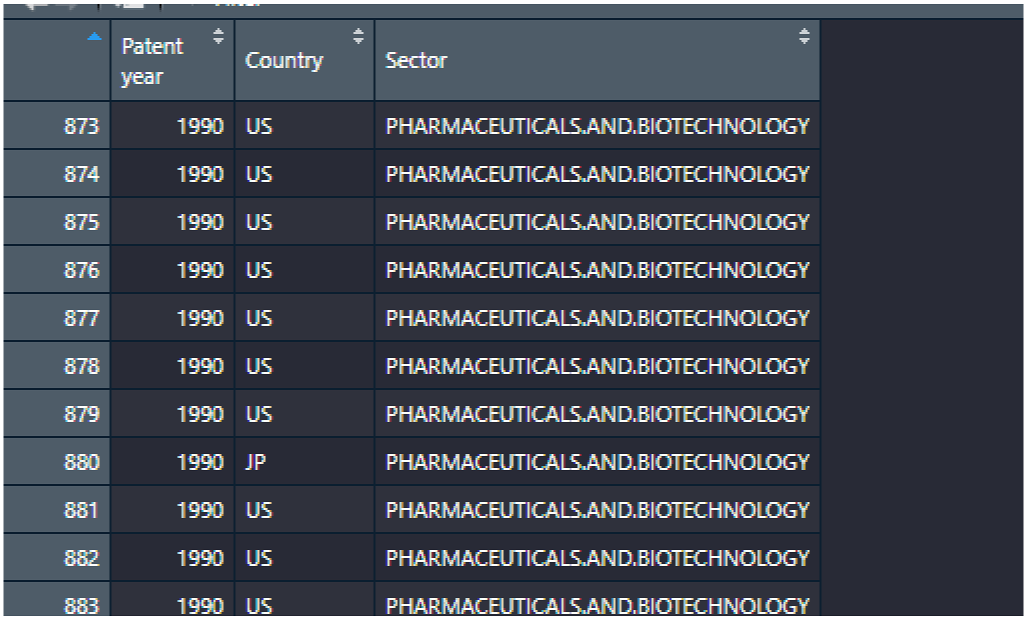
At this point, the data was ready to be divided into developing or developed datasets. This was done through subsetting by country and then by setting patent type to the 'pharmaceutical and biotechnology' sector (see figure 9) .We began by creating a subset for developed nations.

Figure 9:



```
DevelopedNations<-subset(CompiledPatentsTextWithSectors,Country == "NO" |Country == "IE"|Country == "CH"|Country == "DE"|Country == "SE"|Country == "AU"|Coun
DevelopedNations<-subset(DevelopedNations, Sector == "PHARMACEUTICALS.AND.BIOTECHNOLOGY")
```

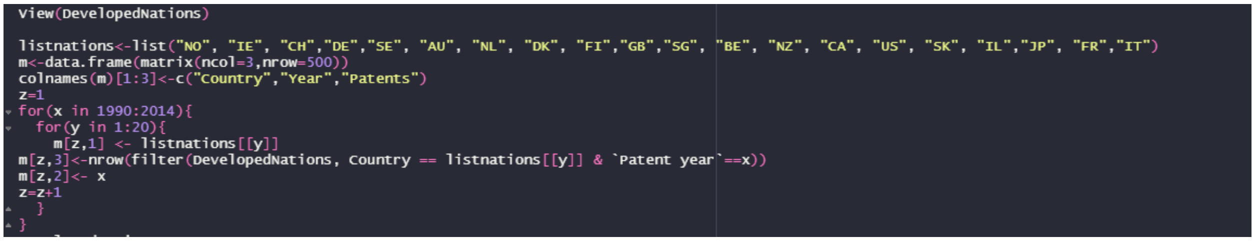
Figure 10:



	Patent year	Country	Sector
873	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY
874	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY
875	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY
876	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY
877	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY
878	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY
879	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY
880	1990	JP	PHARMACEUTICALS.AND.BIOTECHNOLOGY
881	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY
882	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY
883	1990	US	PHARMACEUTICALS.AND.BIOTECHNOLOGY

The resulting data set is visible in the figure above. Next, we wanted a count of how many patents were released per year, per country. So, we created an empty matrix with 3 columns and 500 rows (20 countries and data for 25 years, so 500 rows). We made a list which contained all of the country codes for the nation, and then we named the first column as 'Country', second as 'Year' and third as 'Patents' (table visible in figure 12) . A temp variable named z was also used. A nested for loop was created, the outer loop ran from 1990 to 2014, the inner loop from 1 to 20. The inner loop stores the name of the country depending on 'y' (y is used as an index from the list of nations) in the first column. Z represents the current row which actions are being performed on. 'Nrow' was used to count the number of rows depending on the year (1990 to 2014) and the country name. This returned a count of the patent which was stored in the third column. In the second column year was stored (from the variable in the outer loop). The code of the loop is visible in figure 11.

Figure 11:

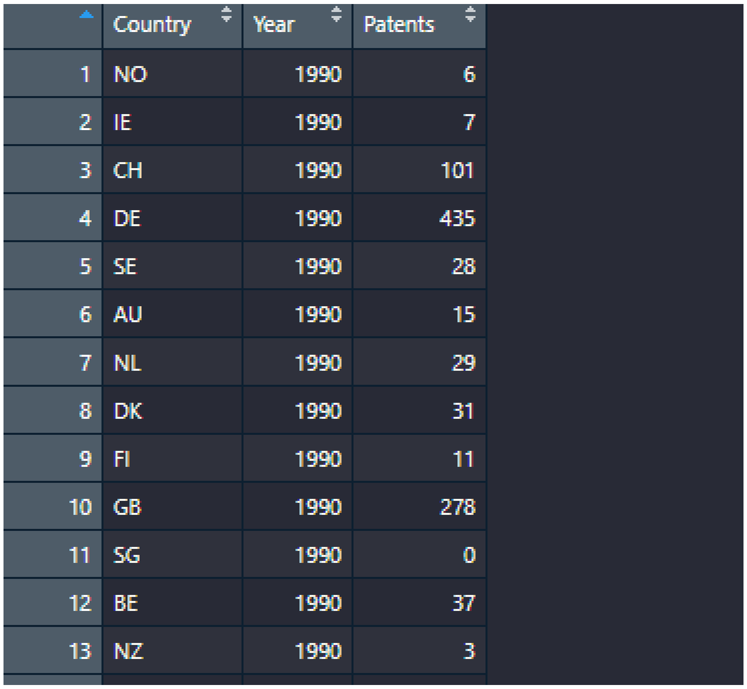


```
View(DevelopedNations)

listnations<-list("NO", "IE", "CH","DE","SE", "AU", "NL", "DK", "FI","GB","SG", "BE", "NZ", "CA", "US", "SK", "IL","JP", "FR","IT")
m<-data.frame(matrix(ncol=3,nrow=500))
colnames(m)[1:3]<-c("Country","Year","Patents")
z=1
for(x in 1990:2014){
  for(y in 1:20){
    m[z,1] <- listnations[[y]]
m[z,3]<-nrow(filter(DevelopedNations, Country == listnations[[y]] & `Patent year`==x))
m[z,2]<- x
z=z+1
  }
}
```

The resultant data set is visible in figure 12. The 'Year' represents the year in which the number of 'Patents' were issued; the 'Country' is the alpha-2 country code for the respective country.

This process was replicated to create the developing nations data set.

Figure 12:



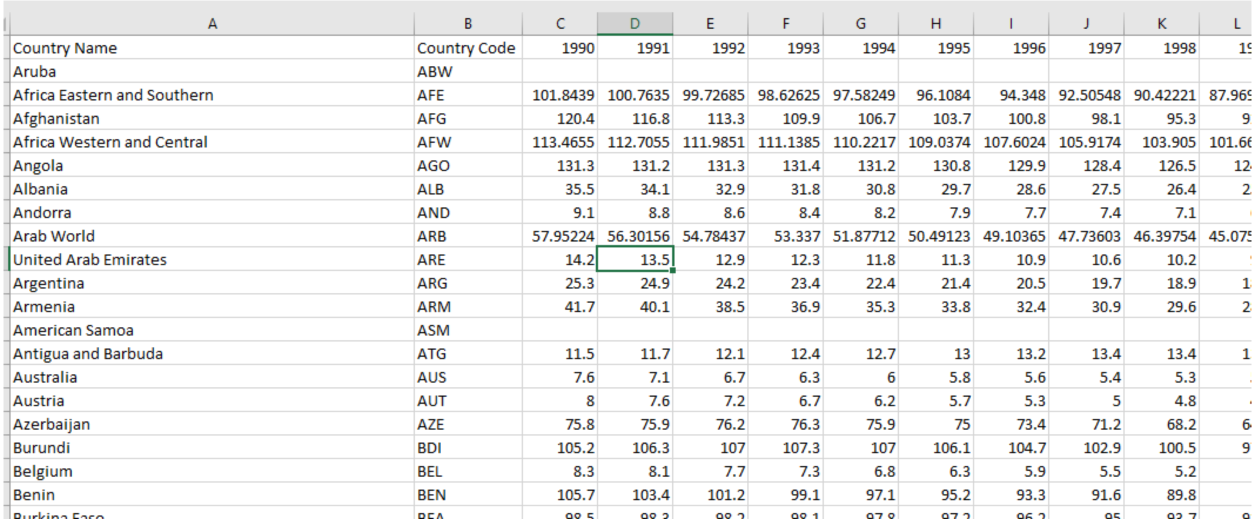
	Country	Year	Patents
1	NO	1990	6
2	IE	1990	7
3	CH	1990	101
4	DE	1990	435
5	SE	1990	28
6	AU	1990	15
7	NL	1990	29
8	DK	1990	31
9	FI	1990	11
10	GB	1990	278
11	SG	1990	0
12	BE	1990	37
13	NZ	1990	3

Infant Mortality Data:

For the infant mortality data, we downloaded a .csv file from World Bank Open Data ("Mortality Rate, Infant (per 1,000 Live Births)." *Data*, https://data.worldbank.org/indicator/SP.DYN.IMRT.IN).

This data was cleaned by manually removing extra rows, column variables such as indicator name and indicator code, and all the years apart from the range 1990 - 2014.

Figure 13:

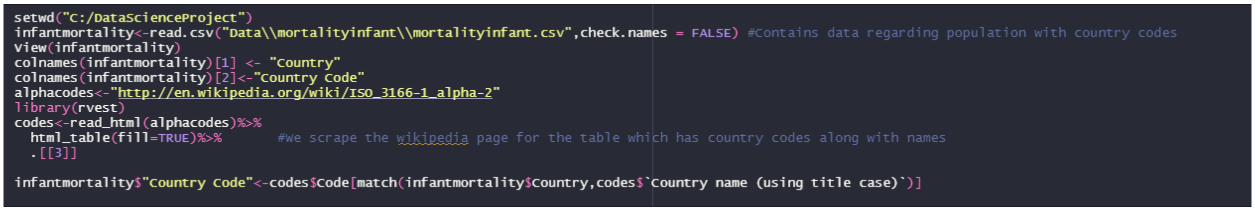


A	B	C	D	E	F	G	H	I	J	K	L
Country Name	Country Code	1990	1991	1992	1993	1994	1995	1996	1997	1998	19
Aruba	ABW										
Africa Eastern and Southern	AFE	101.8439	100.7635	99.72685	98.62625	97.58249	96.1084	94.348	92.50548	90.42221	87.96
Afghanistan	AFG	120.4	116.8	113.3	109.9	106.7	103.7	100.8	98.1	95.3	9
Africa Western and Central	AFW	113.4655	112.7055	111.9851	111.1385	110.2217	109.0374	107.6024	105.9174	103.905	101.66
Angola	AGO	131.3	131.2	131.3	131.4	131.2	130.8	129.9	128.4	126.5	12
Albania	ALB	35.5	34.1	32.9	31.8	30.8	29.7	28.6	27.5	26.4	2
Andorra	AND	9.1	8.8	8.6	8.4	8.2	7.9	7.7	7.4	7.1	
Arab World	ARB	57.95224	56.30156	54.78437	53.337	51.87712	50.49123	49.10365	47.73603	46.39754	45.07
United Arab Emirates	ARE	14.2	13.5	12.9	12.3	11.8	11.3	10.9	10.6	10.2	
Argentina	ARG	25.3	24.9	24.2	23.4	22.4	21.4	20.5	19.7	18.9	1
Armenia	ARM	41.7	40.1	38.5	36.9	35.3	33.8	32.4	30.9	29.6	2
American Samoa	ASM										
Antigua and Barbuda	ATG	11.5	11.7	12.1	12.4	12.7	13	13.2	13.4	13.4	1
Australia	AUS	7.6	7.1	6.7	6.3	6	5.8	5.6	5.4	5.3	
Austria	AUT	8	7.6	7.2	6.7	6.2	5.7	5.3	5	4.8	
Azerbaijan	AZE	75.8	75.9	76.2	76.3	75.9	75	73.4	71.2	68.2	6
Burundi	BDI	105.2	106.3	107	107.3	107	106.1	104.7	102.9	100.5	9
Belgium	BEL	8.3	8.1	7.7	7.3	6.8	6.3	5.9	5.5	5.2	
Benin	BEN	105.7	103.4	101.2	99.1	97.1	95.2	93.3	91.6	89.8	
Burkina Faso	BFA	98.5	98.3	98.2	98.1	97.8	97.2	96.2	95	93.7	9

Unfortunately, this data set used the alpha-3 format (for example 'USA') country codes and in our data set used the alpha-2 format (for example 'US'). To convert the country codes from alpha-3 to alpha-2, we scraped a Wikipedia page("ISO 3166-1 Alpha-2." *Wikipedia*, Wikimedia Foundation, 17 Dec. 2021, http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2), containing the alpha-2 codes along with their respective country names.

First we read the .csv file, and then we saved the link to the wiki in a variable named 'alphacodes'. Using read_the **rvest** package, the webpage was read and a table (stored in codes) was created (code visible in figure 14). The '.[[3]]' is used to get the third table from the wikipedia page. Afterwards, in the population table we changed the ‘country.code’ column’s values to alpha-2, by matching the country name from the ‘infant mortality’ data set to the ‘country name’ from the table stored in ‘codes’.

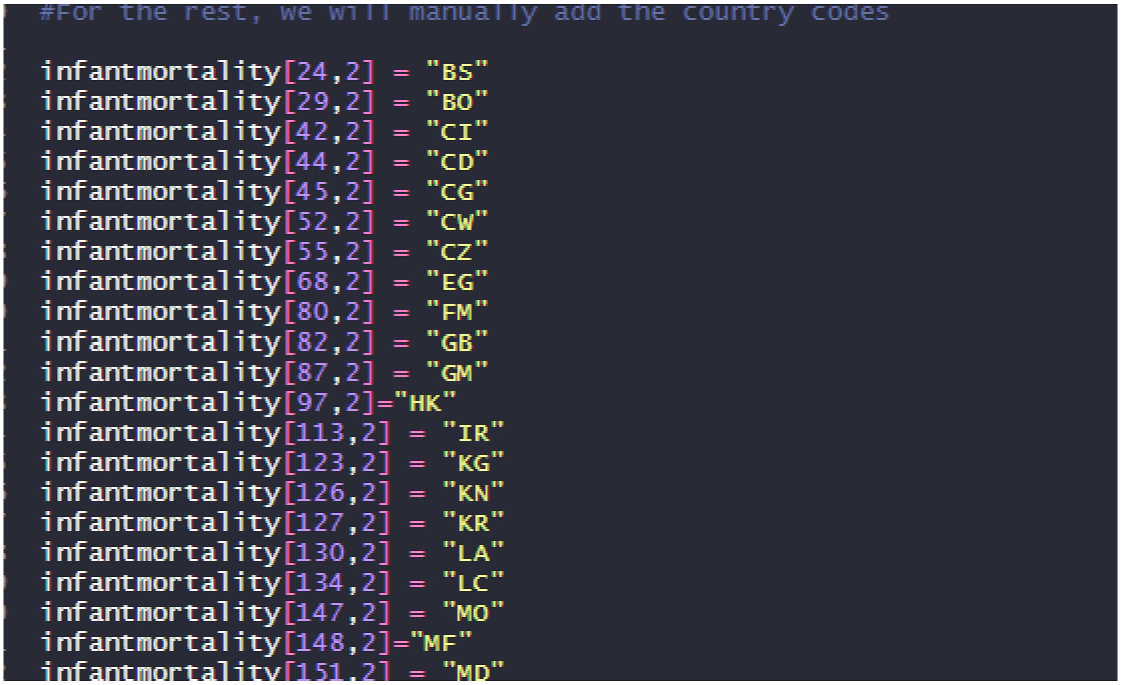
Figure 14:



```
setwd("C:/DataScienceProject")
infantmortality<-read.csv("Data\\mortalityinfant\\mortalityinfant.csv",check.names = FALSE) #Contains data regarding population with country codes
View(infantmortality)
colnames(infantmortality)[1] <- "Country"
colnames(infantmortality)[2]<-"Country Code"
alphacodes<-"http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2"
library(rvest)
codes<-read_html(alphacodes)%>%
  html_table(fill=TRUE)%>%       #We scrape the wikipedia page for the table which has country codes along with names
  .[[3]]

infantmortality$"Country Code"<-codes$Code[match(infantmortality$Country,codes$`Country name (using title case)`)]
```

In the infant mortality table, there are still some countries with NA alpha-2 codes, the reason is that the wikipedia page may use a different name for the same country, for example in the population dataset “Congo Dem Rep.” may be used but on wikipedia “DR Congo” may be used, and other formatting problems. So for these countries, we manually entered the alpha-2 codes.

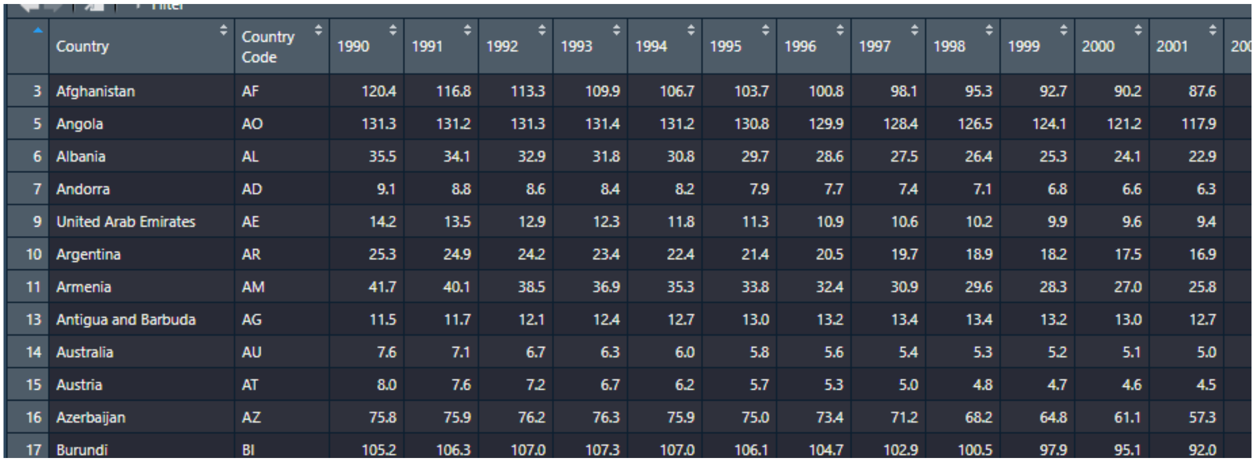
Figure 15:



```
#For the rest, we will manually add the country codes

infantmortality[24,2] = "BS"
infantmortality[29,2] = "BO"
infantmortality[42,2] = "CI"
infantmortality[44,2] = "CD"
infantmortality[45,2] = "CG"
infantmortality[52,2] = "CW"
infantmortality[55,2] = "CZ"
infantmortality[68,2] = "EG"
infantmortality[80,2] = "FM"
infantmortality[82,2] = "GB"
infantmortality[87,2] = "GM"
infantmortality[97,2]="HK"
infantmortality[113,2] = "IR"
infantmortality[123,2] = "KG"
infantmortality[126,2] = "KN"
infantmortality[127,2] = "KR"
infantmortality[130,2] = "LA"
infantmortality[134,2] = "LC"
infantmortality[147,2] = "MO"
infantmortality[148,2]="MF"
infantmortality[151,2] = "MD"
```

As we can see the data is in wide format, so we converted it to long using reshape2’s melt function.

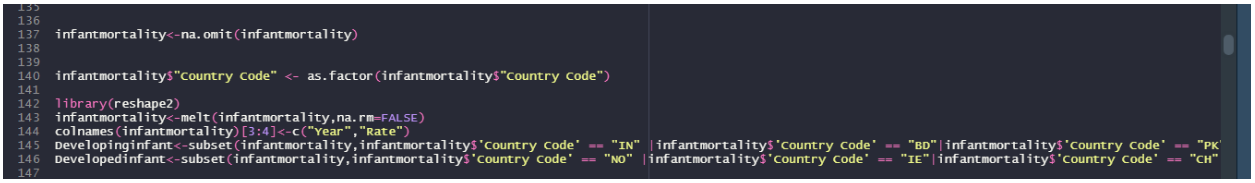
Figure 16:



	Country	Country Code	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
3	Afghanistan	AF	120.4	116.8	113.3	109.9	106.7	103.7	100.8	98.1	95.3	92.7	90.2	87.6
5	Angola	AO	131.3	131.2	131.3	131.4	131.2	130.8	129.9	128.4	126.5	124.1	121.2	117.9
6	Albania	AL	35.5	34.1	32.9	31.8	30.8	29.7	28.6	27.5	26.4	25.3	24.1	22.9
7	Andorra	AD	9.1	8.8	8.6	8.4	8.2	7.9	7.7	7.4	7.1	6.8	6.6	6.3
9	United Arab Emirates	AE	14.2	13.5	12.9	12.3	11.8	11.3	10.9	10.6	10.2	9.9	9.6	9.4
10	Argentina	AR	25.3	24.9	24.2	23.4	22.4	21.4	20.5	19.7	18.9	18.2	17.5	16.9
11	Armenia	AM	41.7	40.1	38.5	36.9	35.3	33.8	32.4	30.9	29.6	28.3	27.0	25.8
13	Antigua and Barbuda	AG	11.5	11.7	12.1	12.4	12.7	13.0	13.2	13.4	13.4	13.2	13.0	12.7
14	Australia	AU	7.6	7.1	6.7	6.3	6.0	5.8	5.6	5.4	5.3	5.2	5.1	5.0
15	Austria	AT	8.0	7.6	7.2	6.7	6.2	5.7	5.3	5.0	4.8	4.7	4.6	4.5
16	Azerbaijan	AZ	75.8	75.9	76.2	76.3	75.9	75.0	73.4	71.2	68.2	64.8	61.1	57.3
17	Burundi	BI	105.2	106.3	107.0	107.3	107.0	106.1	104.7	102.9	100.5	97.9	95.1	92.0

Before using melt, we removed all the NA values, we performed melt and subset the data according to Developing and Developed Nations.

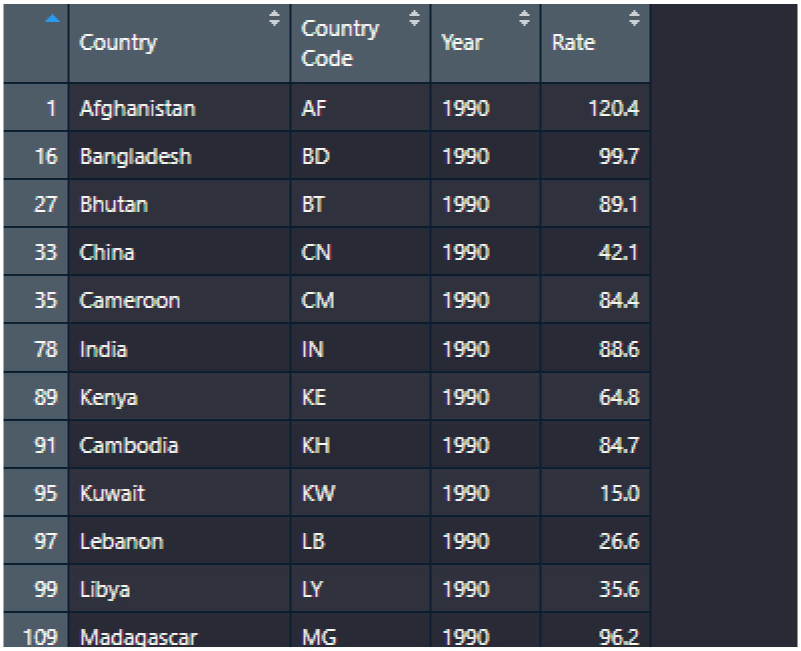
Figure 17:



```
infantmortality<-na.omit(infantmortality)


infantmortality$"Country Code" <- as.factor(infantmortality$"Country Code")

library(reshape2)
infantmortality<-melt(infantmortality,na.rm=FALSE)
colnames(infantmortality)[3:4]<-c("Year","Rate")
Developinginfant<-subset(infantmortality,infantmortality$'Country Code' == "IN" |infantmortality$'Country Code' == "BD"|infantmortality$'Country Code' == "PK'
Developedinfant<-subset(infantmortality,infantmortality$'Country Code' == "NO" |infantmortality$'Country Code' == "IE"|infantmortality$'Country Code' == "CH"
```

Here is the final data for developing nations as an example:

Figure 18:



	Country	Country Code	Year	Rate
1	Afghanistan	AF	1990	120.4
16	Bangladesh	BD	1990	99.7
27	Bhutan	BT	1990	89.1
33	China	CN	1990	42.1
35	Cameroon	CM	1990	84.4
78	India	IN	1990	88.6
89	Kenya	KE	1990	64.8
91	Cambodia	KH	1990	84.7
95	Kuwait	KW	1990	15.0
97	Lebanon	LB	1990	26.6
99	Libya	LY	1990	35.6
109	Madagascar	MG	1990	96.2

Neonatal Mortality & Average Life Expectancy Data Set:

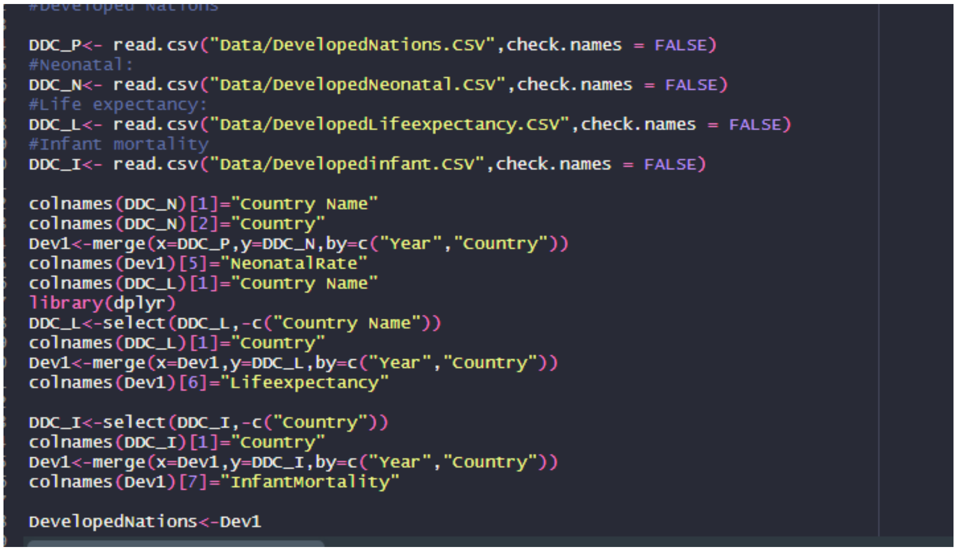
For Neonatal Mortality and average Life Expectancy data set,, the data source was from the same website, and the formatting of the data was also the same, World Bank, (https://data.worldbank.org/indicator/SH.DYN.NMRT , https://data.worldbank.org/indicator/SP.DYN.LE00.IN) , so we performed the same steps as earlier used on Infant Mortality on these datasets.

Merging the data:

Finally, since it is needlessly time consuming to plot with independent variables in one dataset and dependent variables in another datasets, we merged the datasets for infant mortality, neonatal mortality, average life expectancy and nations to create two final datasets, one for developing nations and one for developed.

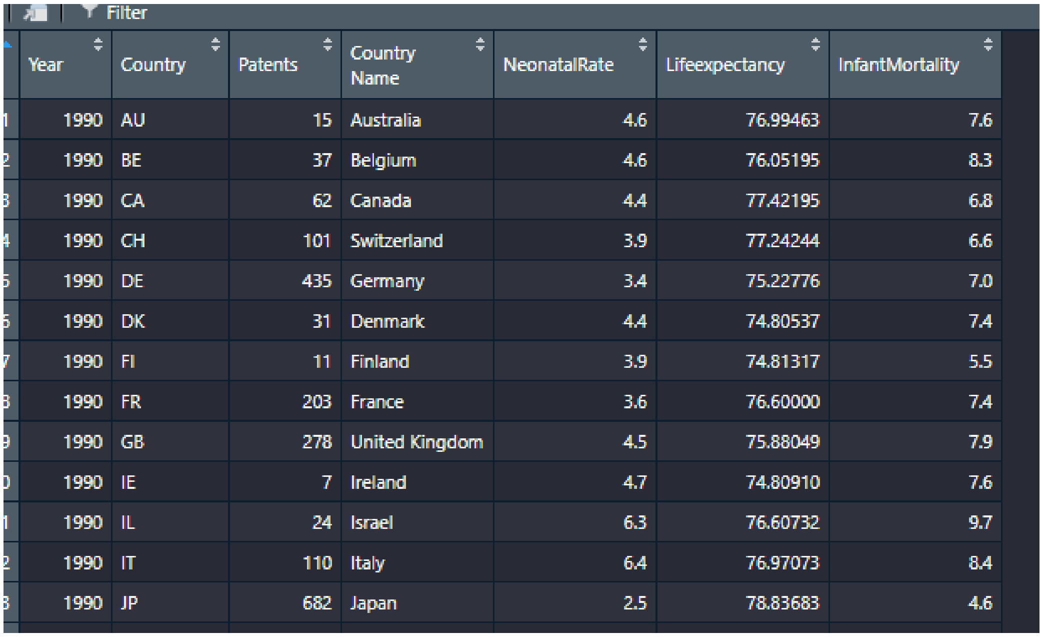
This all was done by iteratively using an inner join, matched using Year and Country (we renamed country codes to country). We also removed some extra columns that we did not need.

Figure 19:



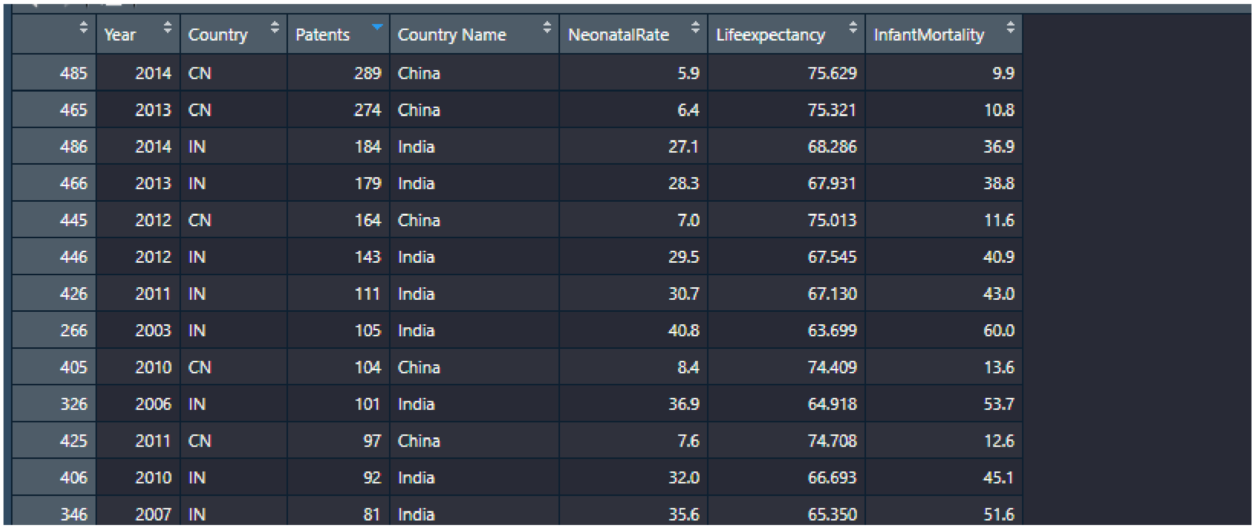
```
#Developed Nations

DDC_P<- read.csv("Data/DevelopedNations.CSV",check.names = FALSE)
#Neonatal:
DDC_N<- read.csv("Data/DevelopedNeonatal.CSV",check.names = FALSE)
#Life expectancy:
DDC_L<- read.csv("Data/DevelopedLifeexpectancy.CSV",check.names = FALSE)
#Infant mortality
DDC_I<- read.csv("Data/Developedinfant.CSV",check.names = FALSE)

colnames(DDC_N)[1]="Country Name"
colnames(DDC_N)[2]="Country"
Dev1<-merge(x=DDC_P,y=DDC_N,by=c("Year","Country"))
colnames(Dev1)[5]="NeonatalRate"
colnames(DDC_L)[1]="Country Name"
library(dplyr)
DDC_L<-select(DDC_L,-c("Country Name"))
colnames(DDC_L)[1]="Country"
Dev1<-merge(x=Dev1,y=DDC_L,by=c("Year","Country"))
colnames(Dev1)[6]="Lifeexpectancy"

DDC_I<-select(DDC_I,-c("Country"))
colnames(DDC_I)[1]="Country"
Dev1<-merge(x=Dev1,y=DDC_I,by=c("Year","Country"))
colnames(Dev1)[7]="InfantMortality"

DevelopedNations<-Dev1
```

Figure 20:



	Year	Country	Patents	Country Name	NeonatalRate	Lifeexpectancy	InfantMortality
1	1990	AU	15	Australia	4.6	76.99463	7.6
2	1990	BE	37	Belgium	4.6	76.05195	8.3
3	1990	CA	62	Canada	4.4	77.42195	6.8
4	1990	CH	101	Switzerland	3.9	77.24244	6.6
5	1990	DE	435	Germany	3.4	75.22776	7.0
6	1990	DK	31	Denmark	4.4	74.80537	7.4
7	1990	FI	11	Finland	3.9	74.81317	5.5
8	1990	FR	203	France	3.6	76.60000	7.4
9	1990	GB	278	United Kingdom	4.5	75.88049	7.9
0	1990	IE	7	Ireland	4.7	74.80910	7.6
1	1990	IL	24	Israel	6.3	76.60732	9.7
2	1990	IT	110	Italy	6.4	76.97073	8.4
3	1990	JP	682	Japan	2.5	78.83683	4.6

Figure 21:



	Year	Country	Patents	Country Name	NeonatalRate	Lifeexpectancy	InfantMortality
485	2014	CN	289	China	5.9	75.629	9.9
465	2013	CN	274	China	6.4	75.321	10.8
486	2014	IN	184	India	27.1	68.286	36.9
466	2013	IN	179	India	28.3	67.931	38.8
445	2012	CN	164	China	7.0	75.013	11.6
446	2012	IN	143	India	29.5	67.545	40.9
426	2011	IN	111	India	30.7	67.130	43.0
266	2003	IN	105	India	40.8	63.699	60.0
405	2010	CN	104	China	8.4	74.409	13.6
326	2006	IN	101	India	36.9	64.918	53.7
425	2011	CN	97	China	7.6	74.708	12.6
406	2010	IN	92	India	32.0	66.693	45.1
346	2007	IN	81	India	35.6	65.350	51.6

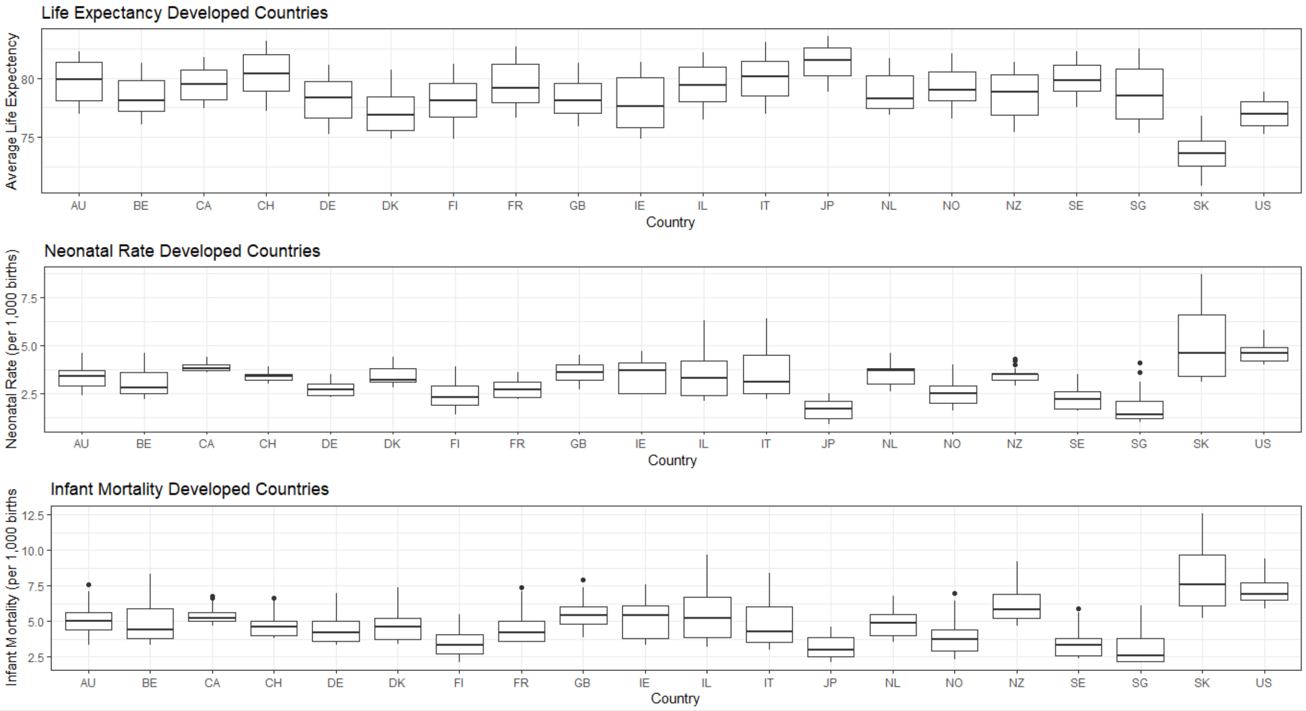
The final dataset for both developing and developed countries is visible in figures 20 and 21. The 'Year' represents the years the patents came out. Country represents the alpha-2 code for the country. 'Patents' represents the count of patents that were released that year for the country. `Country Name` represents the name of the country. `NeonatalRate` is the Neonatal mortality rate per 1,000 live births of the country, `Life Expectancy` is average Life Expectancy of the country and `InfantMortality` is the infant mortality rate per 1,000 live births of the country.

Data Exploration

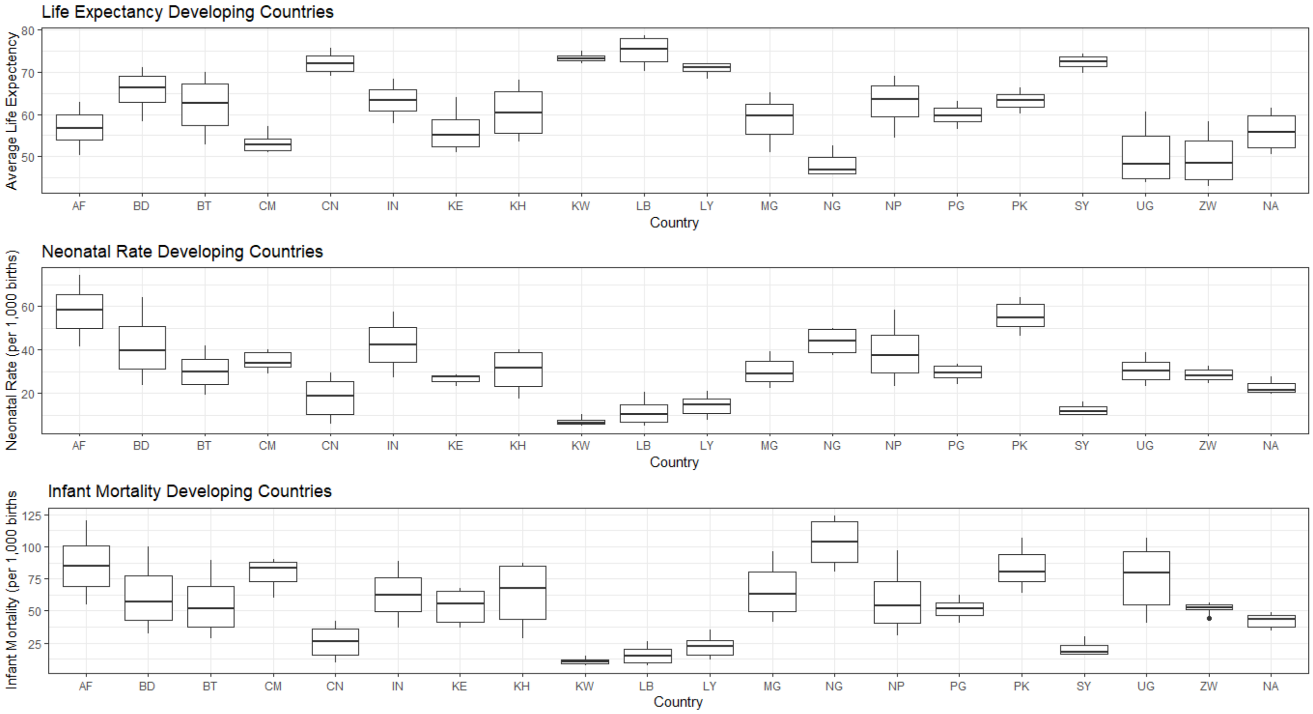
Before we begin, we will look at the general spread of dependent variables for both Developed and Developing Nations. This will be done using boxplot (**ggplot2**). The purpose of this was to just get a general idea of the type of data we will be working with.

First, we plot Infant Mortality on the x-axis against Average Life Expectancy on the y-axis (**using base R**). We made a similar graph for Neonatal mortality on the x-axis against Average Life Expectancy on the y-axis. It can be inferred from both of the graphs that generally you would expect your Average life expectancy to go down as either Infant Mortality or Neonatal mortality increase. For both of the predictors, we generally see that Developing countries exhibit more scatter, this can be attributed to a variety of external factors such as famine or war (refer to graphs 3 and 4).

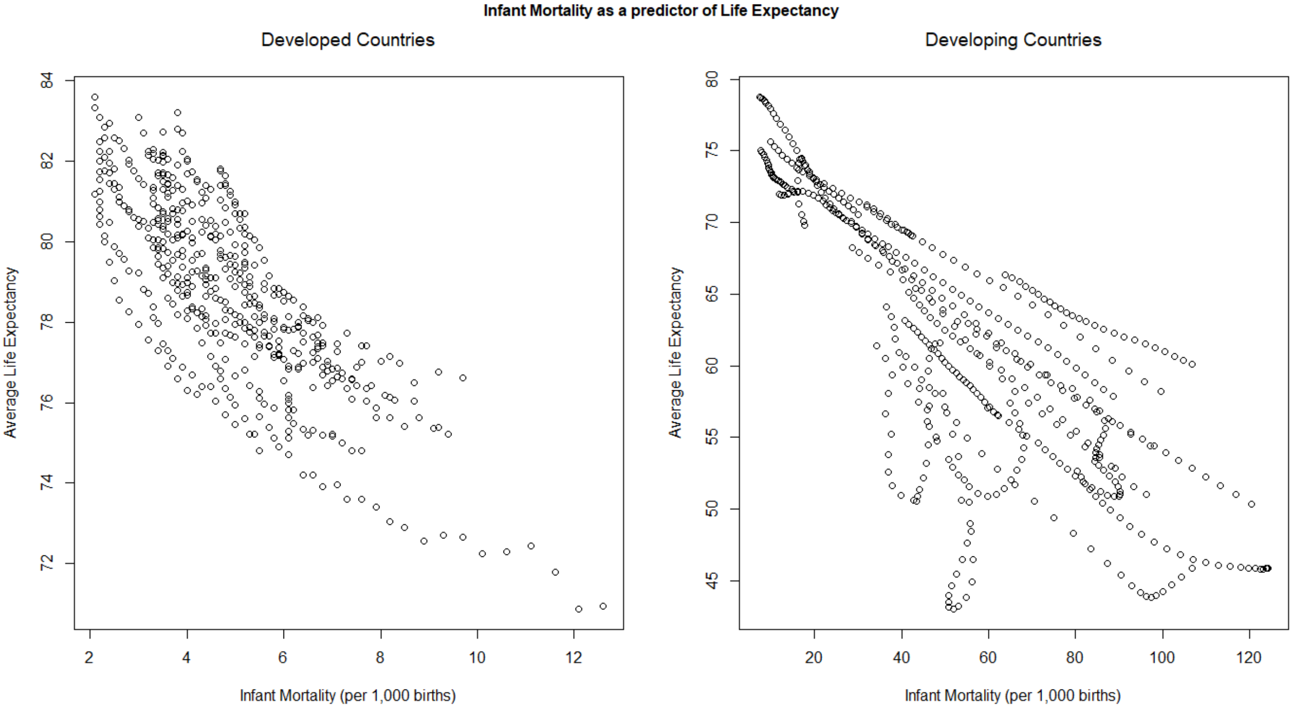
Graph 1:



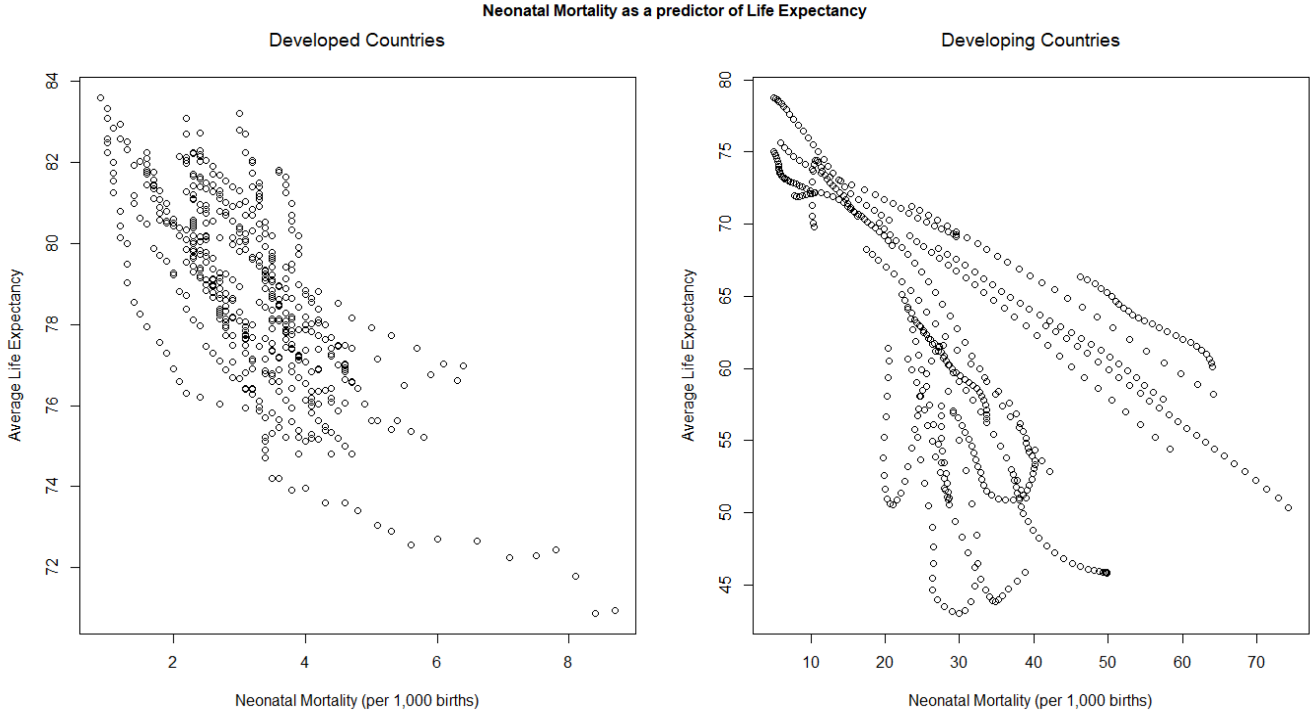
Graph 2:



Graph 3:

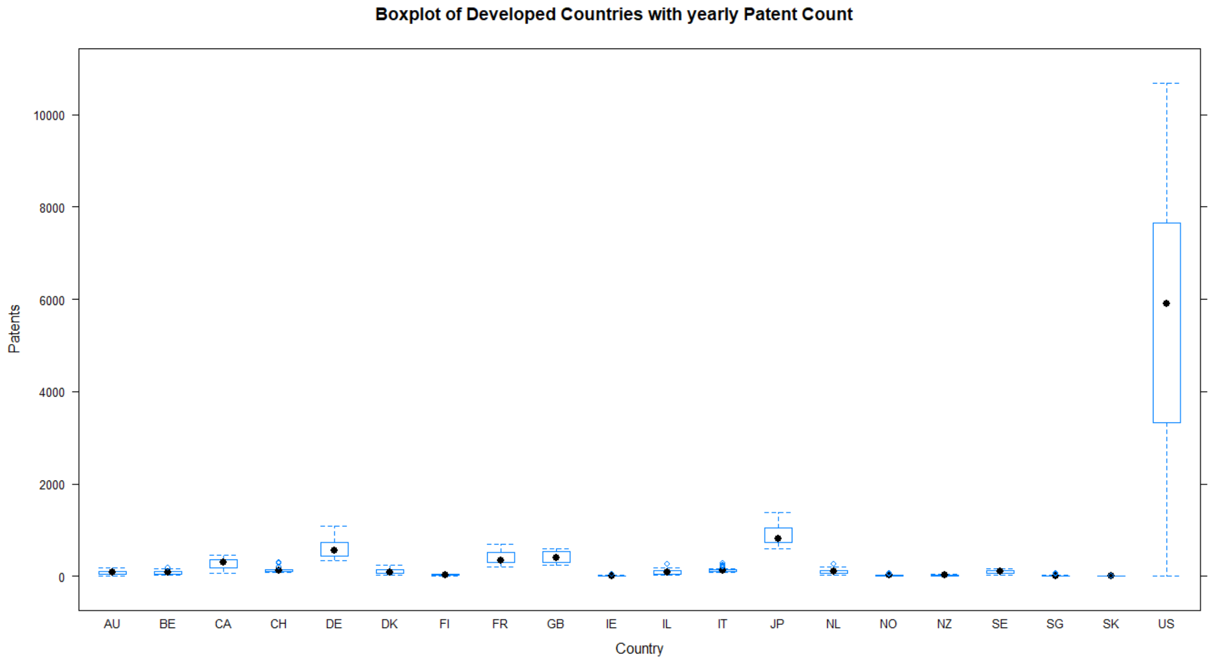


Graph 4:



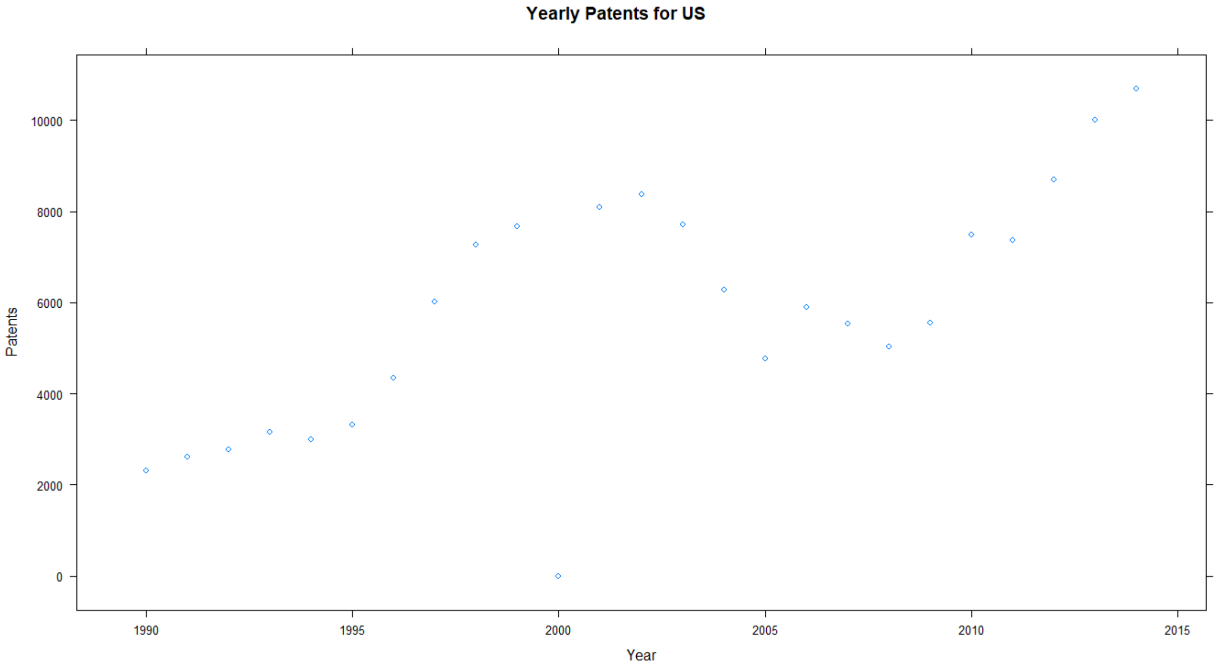
Before we proceed, to get an idea of how many patents a country generally gets per year we created box plots of Developing and Developed Countries (**lattice**). From the boxplot (graphs 4 and 5) of Developed Countries, we can see that most of the countries have gotten some number of patents over the years. Perhaps the most significant for us to look at is the US. It shows that the mean value of the US is the highest and from the size of it we can see there is a lot of standard deviation.

Graph 5:



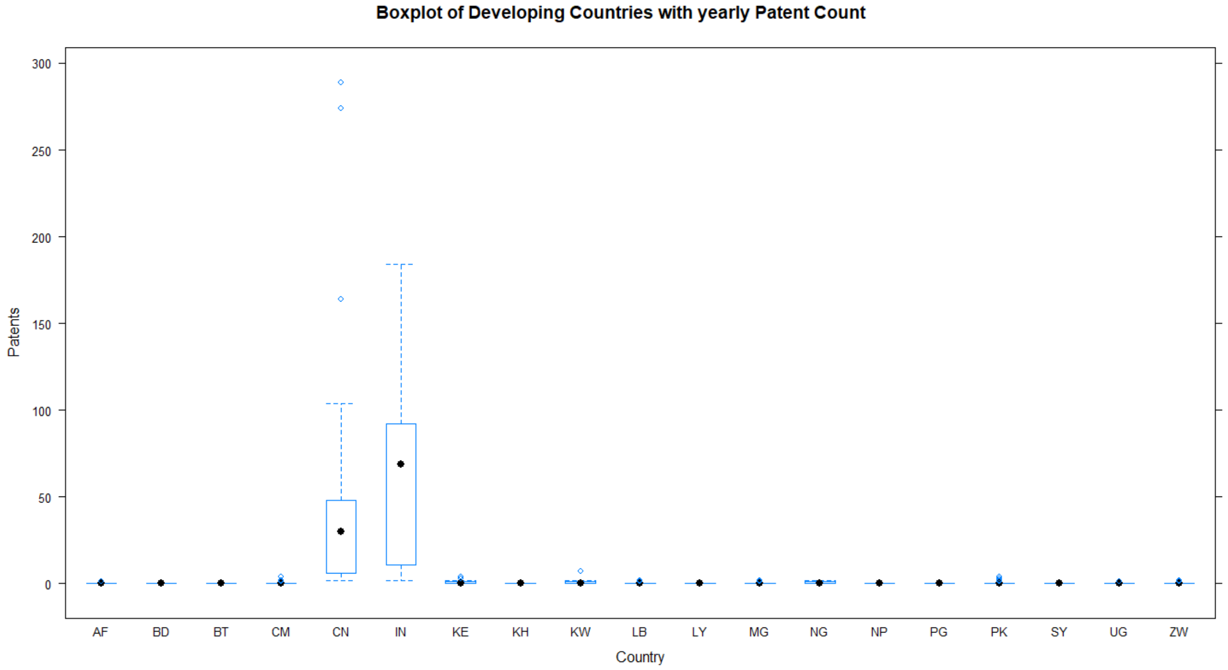
If we plot the graph for yearly patents of US against Year **(lattice)**, we can see that there is a generally increasing trend but there is a lot of variation.

Graph 6:



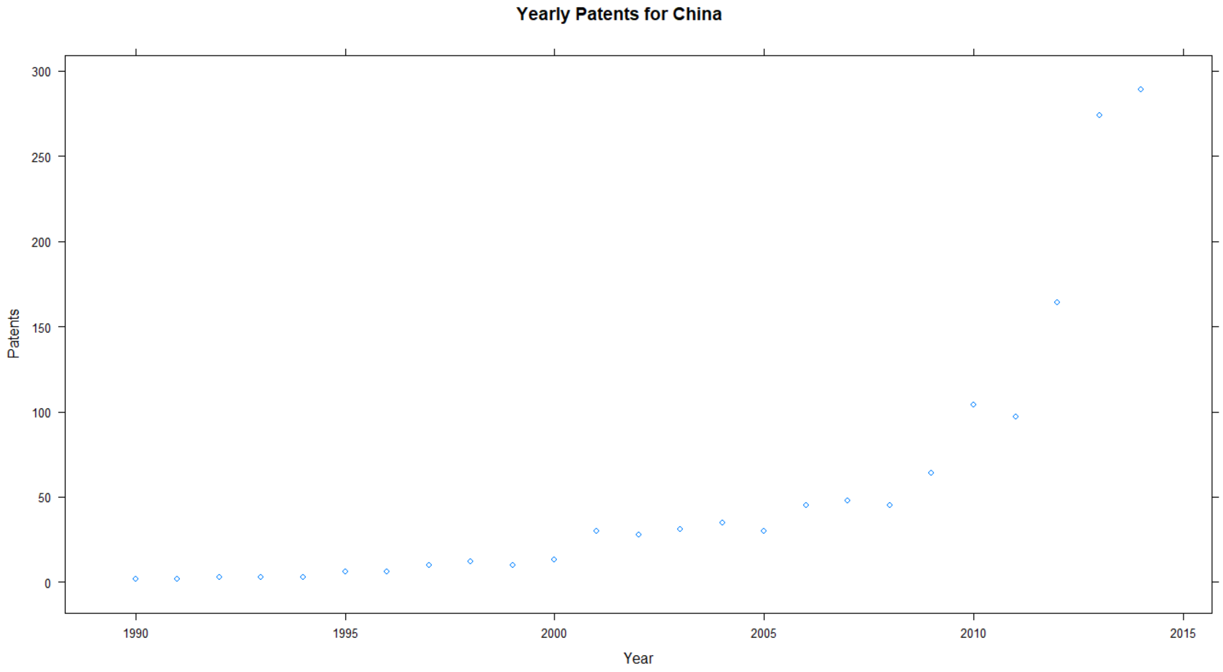
For the Developing Countries dataset, we can see that generally apart from a handful of countries, most countries have a really small number or no patents in this sector. China and India, the biggest boxplots(**lattice**) in this graph, even their values are generally magnitudes smaller compared to the patents US has. It is to be noted that China does seem to have a lot of outliers. It may be possible that in recent years due to their rapidly growing economy they may have started focusing more on this sector.

Graph 7:



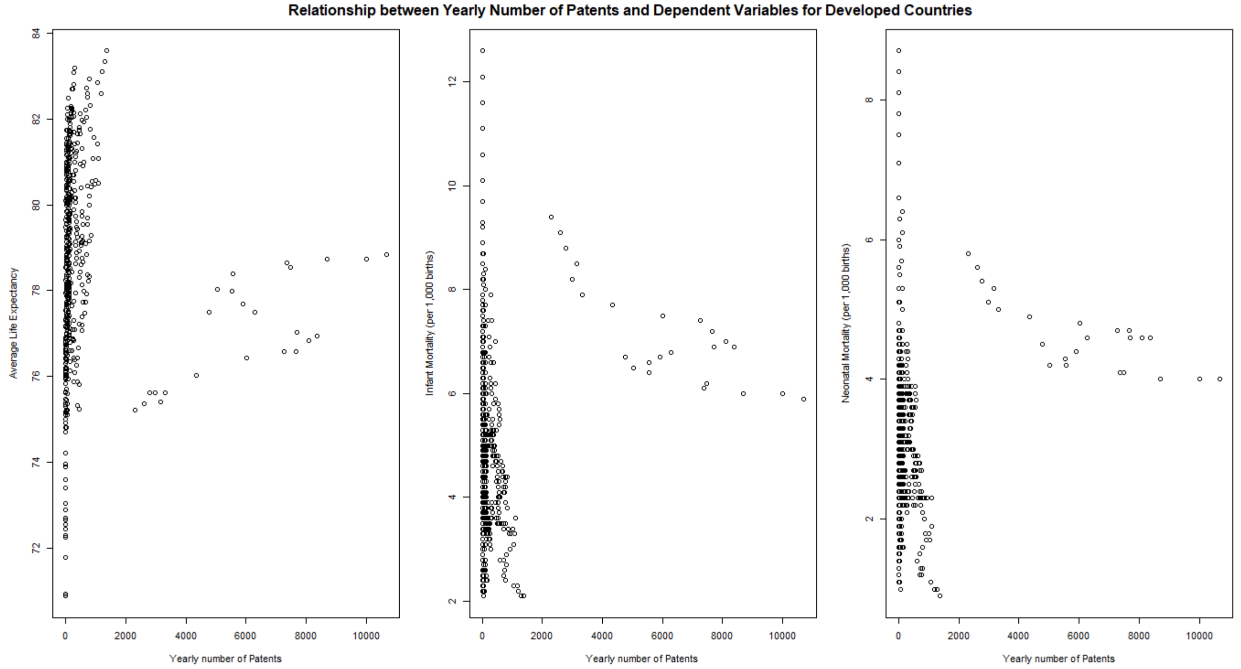
If we plot(**lattice**) the patents released in China against Year, we do see that there is an increase in recent years, this may be correlated to China's rapidly increasing economy.

Graph 8:

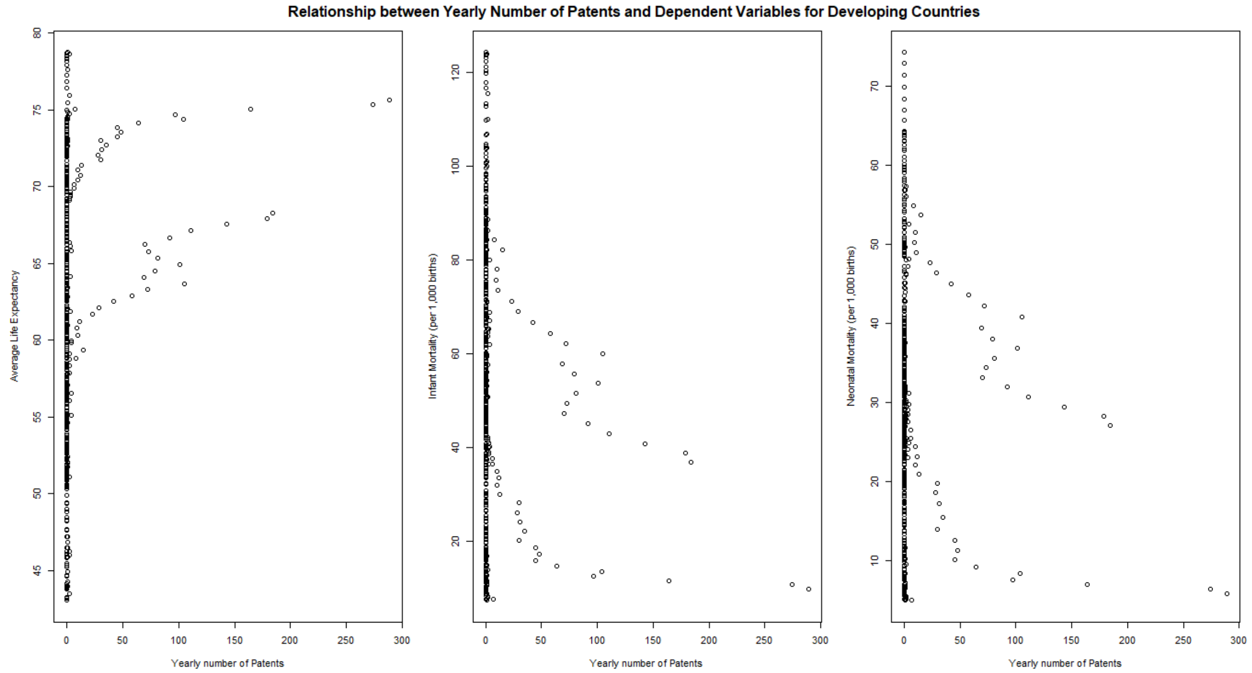


Next, we plotted(base R) the various dependent variables with our independent variable in base R together to see whether there exists an obvious trend. Generally, we see that there is not an obvious trend present. This is as different dots represent different countries, so we will need to look at each of these graphs individually and add colour to them.

Graph 9:

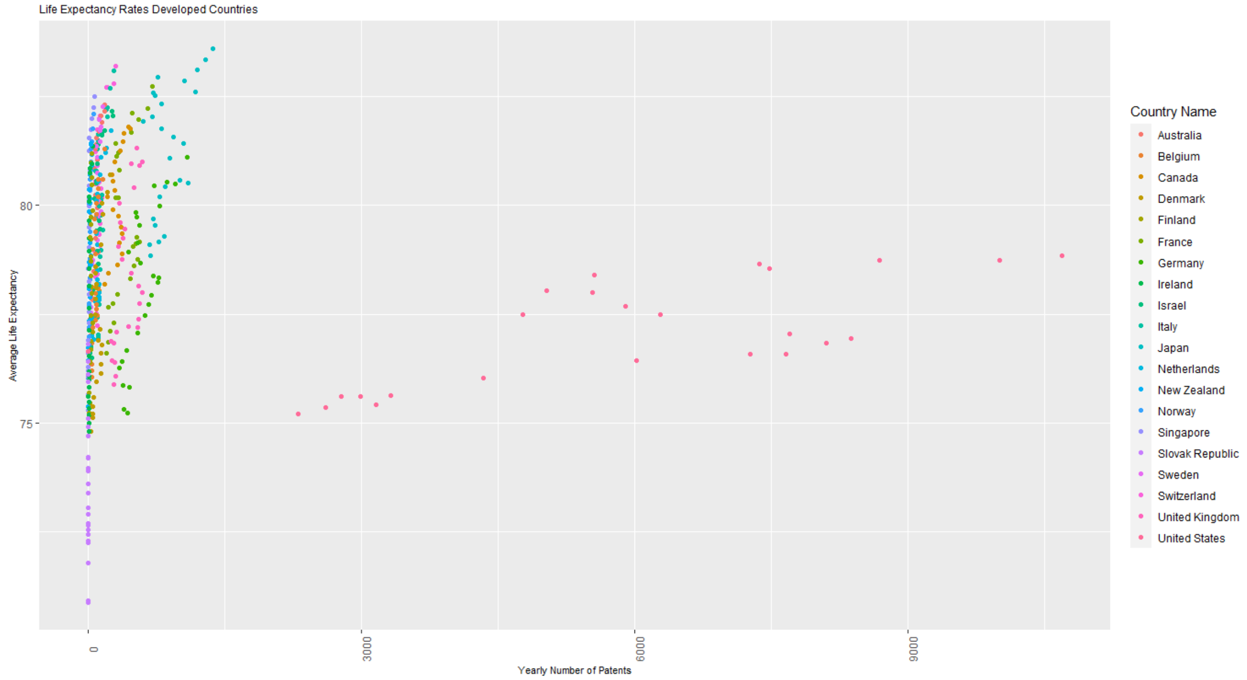


Graph 10:



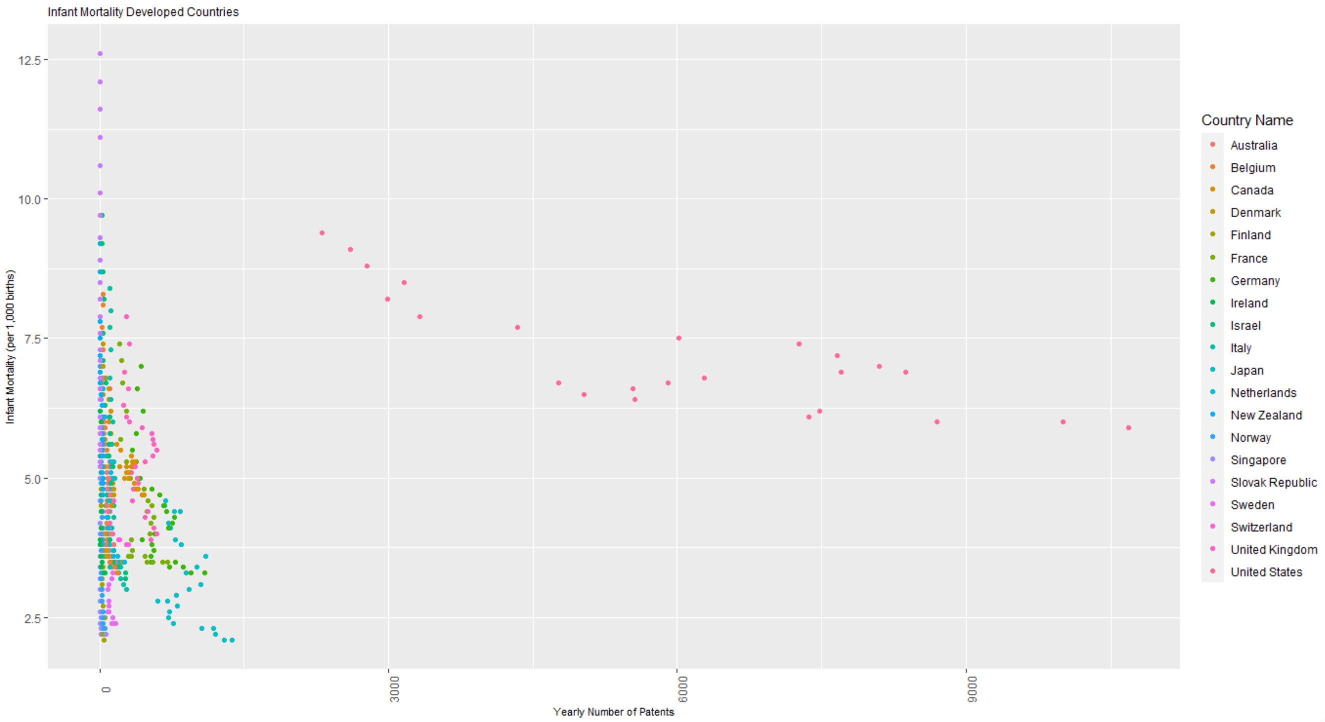
If we look at the average life expectancy rates of developed countries, plotted against yearly number of patents(**ggplot**), we generally see that there does seem to be some kind of trend present. As the number of patents increases, average life expectancy is increasing. There are some countries however where there is a general increase in average life expectancy without any change in the yearly number of patents.

Graph 11:

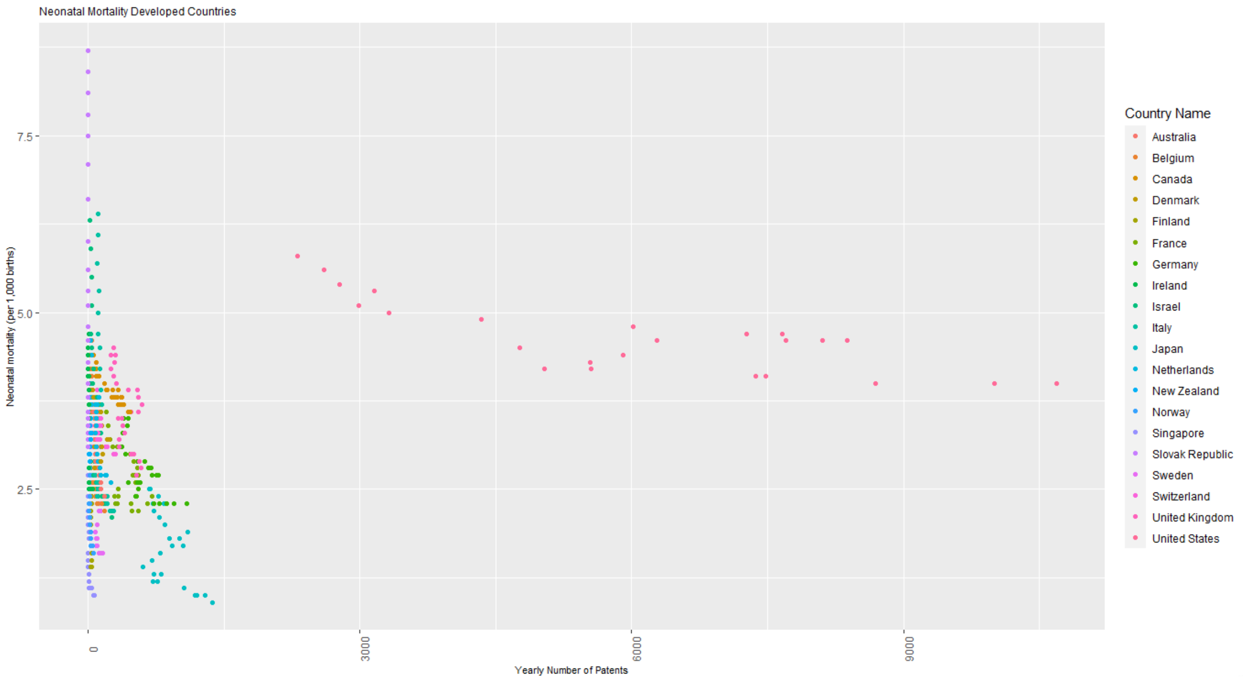


Similarly, for Infant Mortality rates and for Neonatal mortality rates, we see a similar pattern as earlier. As the number of yearly patents increases Infant and Neonatal mortality is decreasing, however, some countries have had a decrease in their mortality rates with a really small or 0 amount of patents. The reason for this may be the effect of external variables, as mentioned in the introduction. (**ggplot**)

Graph 12:

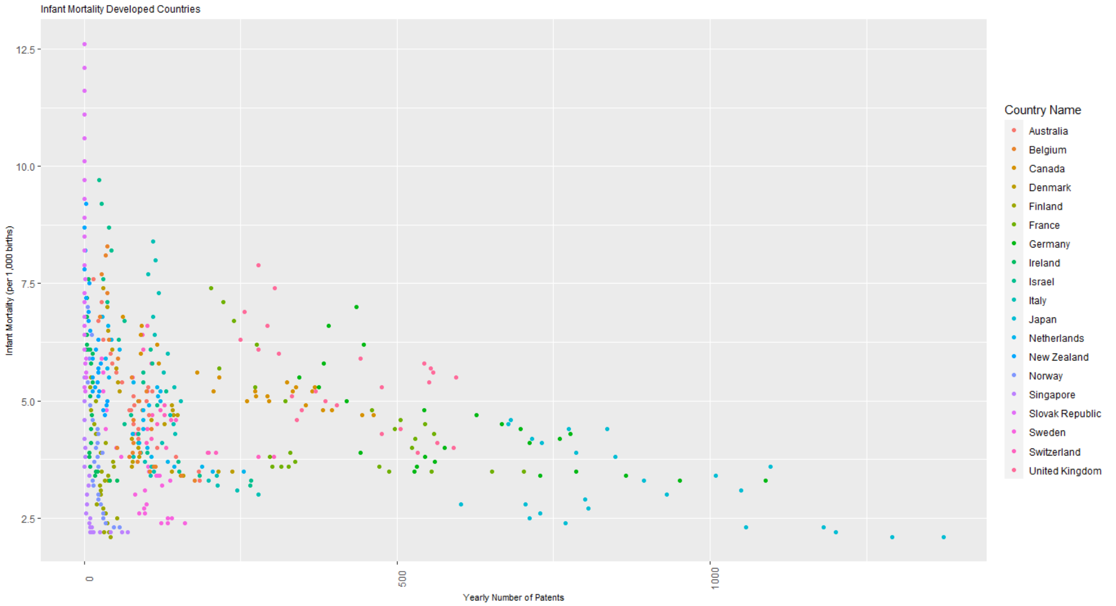


Graph 13:

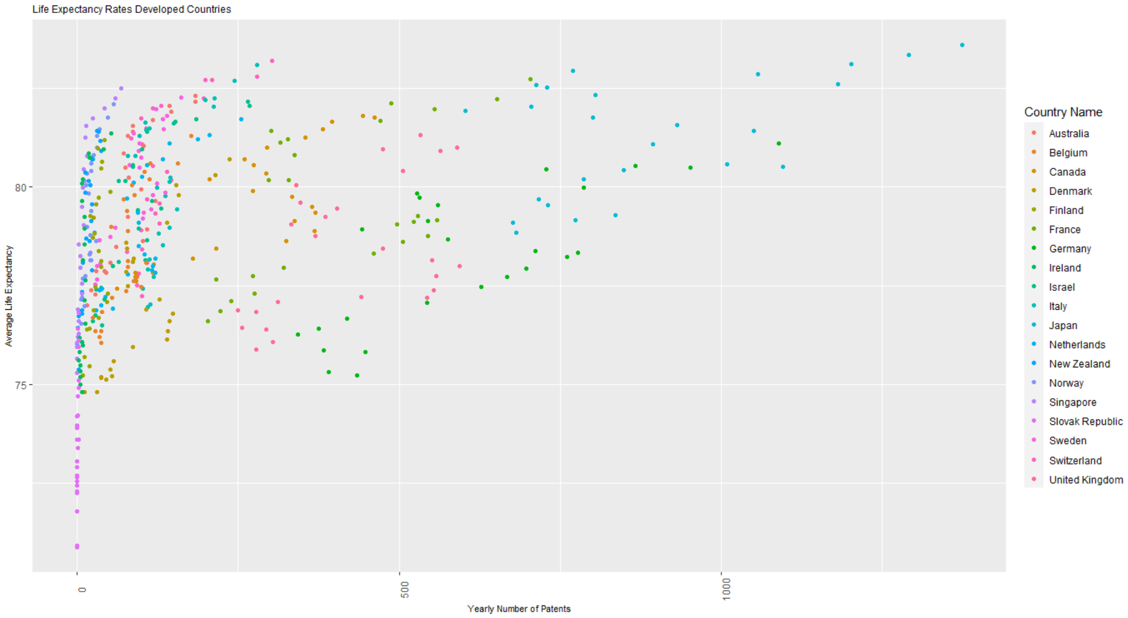


We will make the scatter plots (**ggplot**) again for developing countries, this time temporarily removing the United States as the number of patents they release yearly is really high, making it hard to see a trend. As expected, plotting developing countries without the United States stays consistent with our expectations, we see that as the yearly number of patents increases average Life Expectancy increases, Neonatal mortality decreases and Infant Mortality Decreases

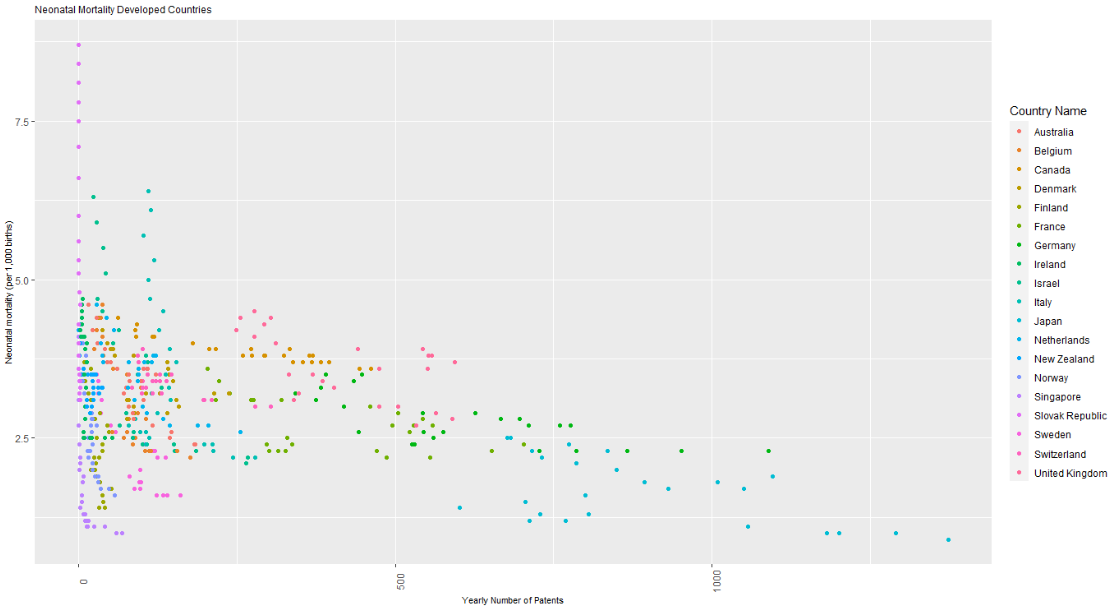
Graph 13:



Graph 14:.

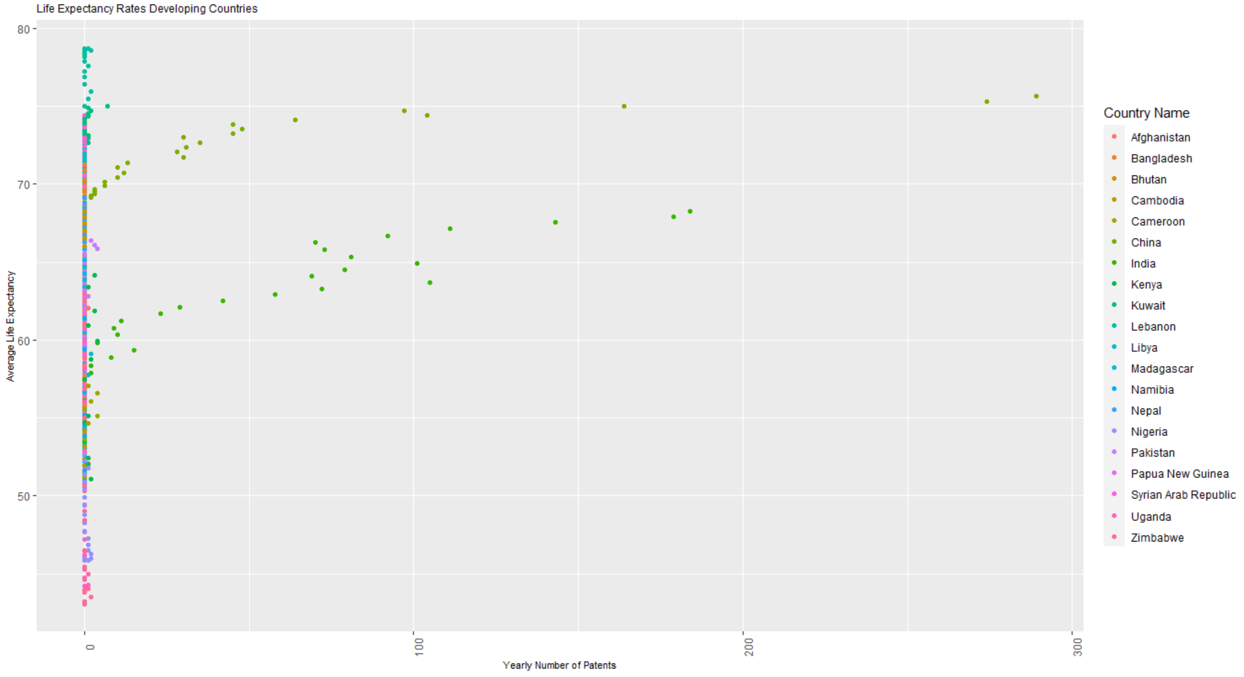


Graph 15:

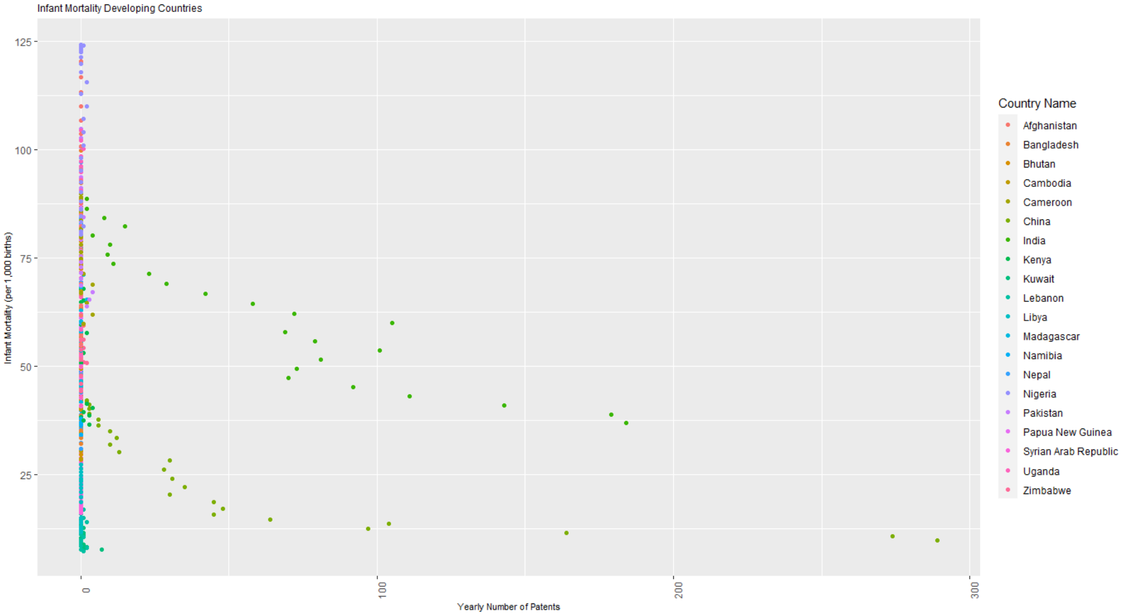


For Developing Countries, we expected a similar result. However, Average life expectancy increased and Neonatal and Infant mortality decreased for a lot of countries without there being any significant (or any!) number of patents.

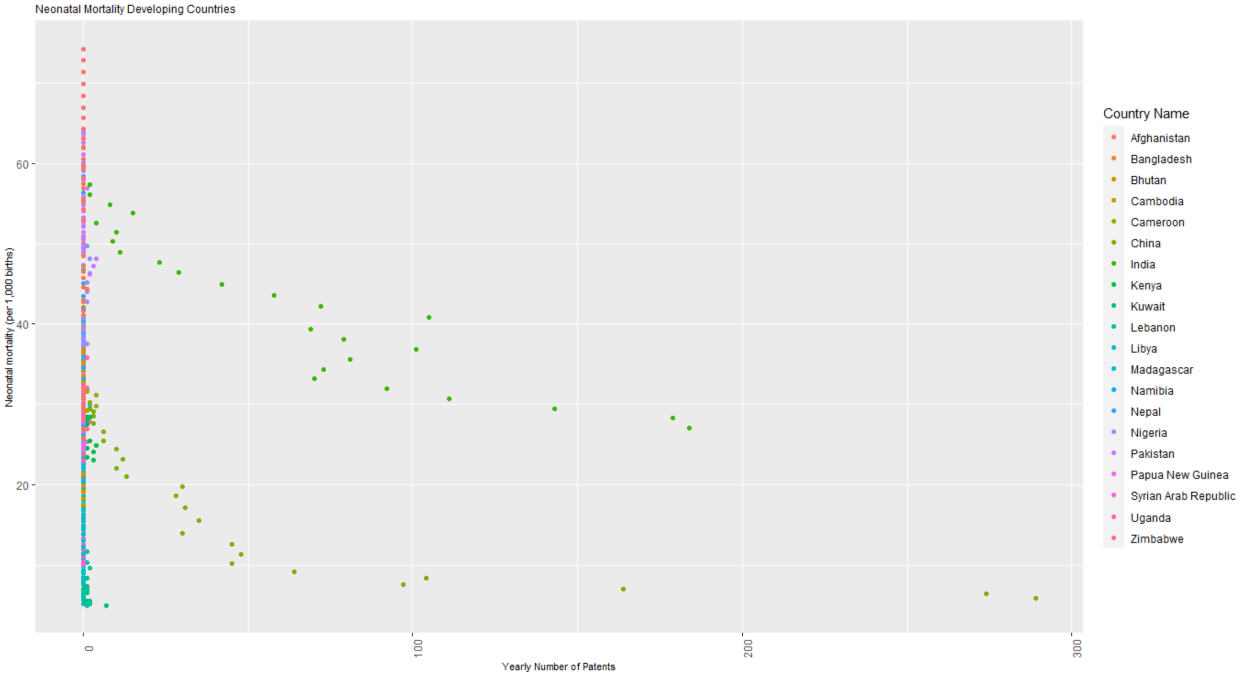
Graph 16:



Graph 17:

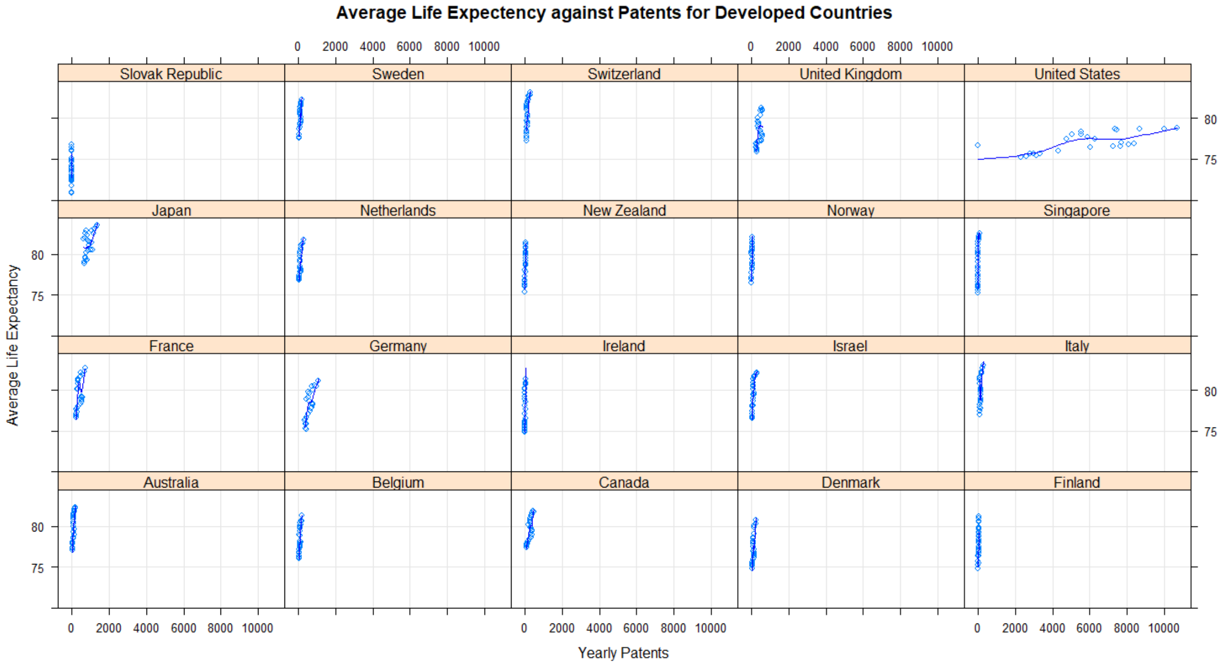


Graph 18:

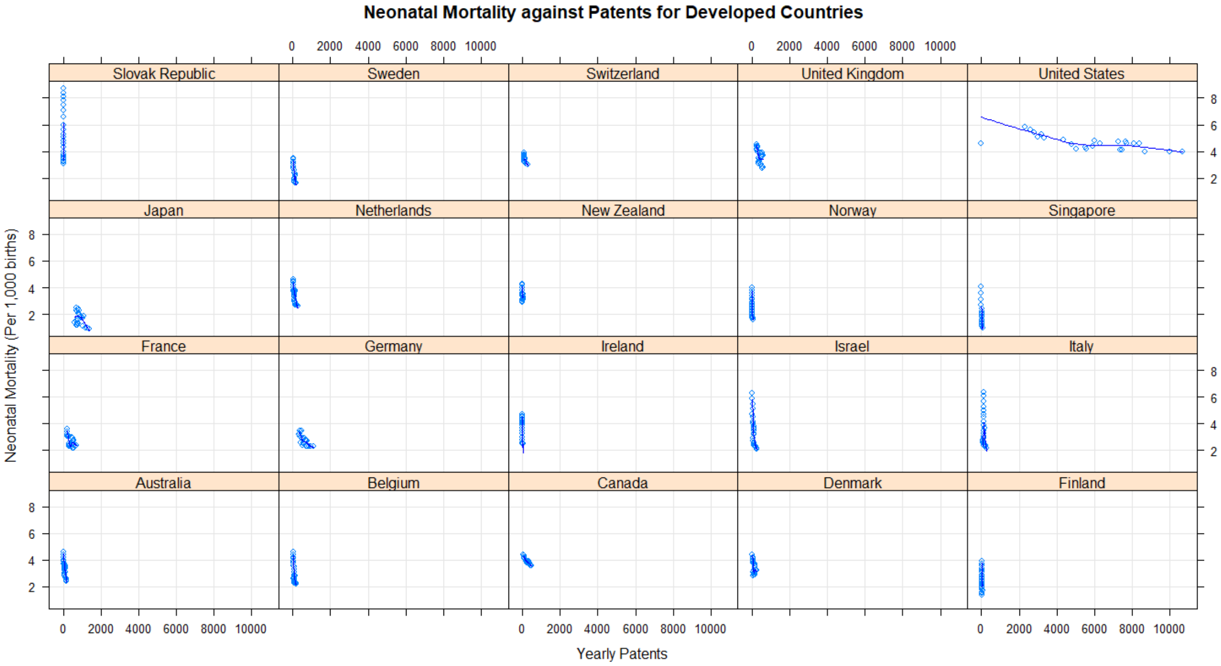


Next, we will look at the Data country-wise. First we look at Developed Nations. For most of the countries, average Life Expectancy increases, Neonatal mortality decreases and Infant Mortality decreases as the yearly number of patents increases. This trend seems to be the most obvious in the United States. It is harder to see in some other countries, due to the fact that the yearly number of patents that they have is really small.But if we look closely we can see that the trend holds true.

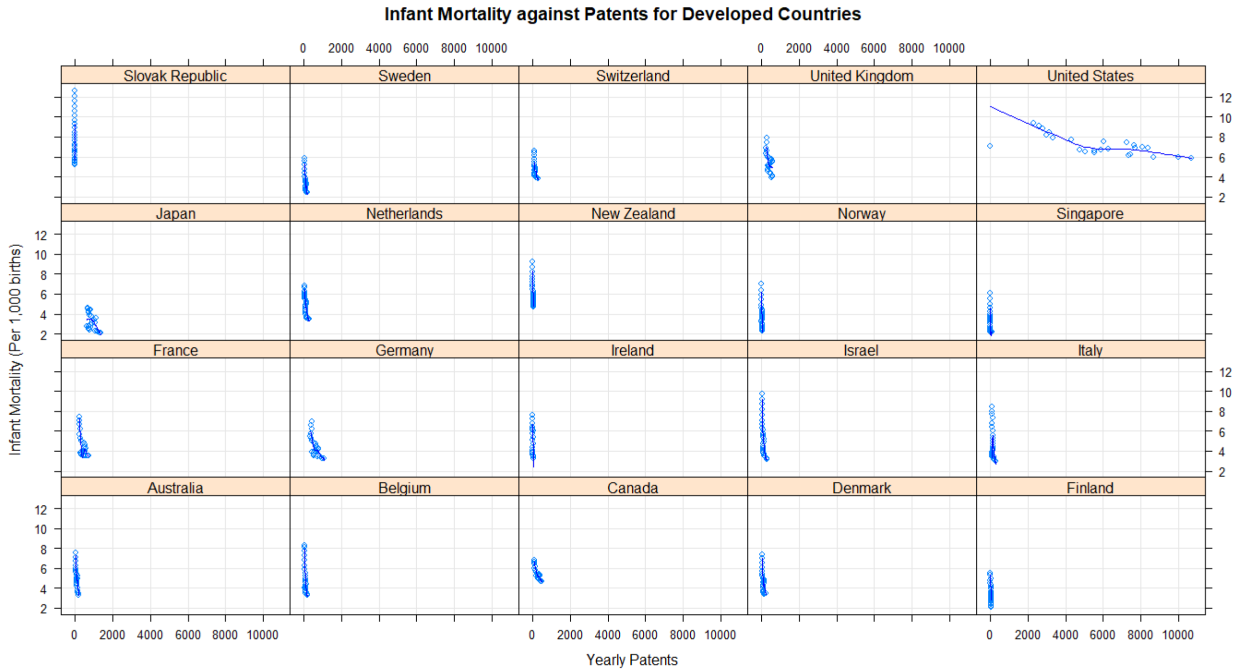
Graph 19:



Graph 20:

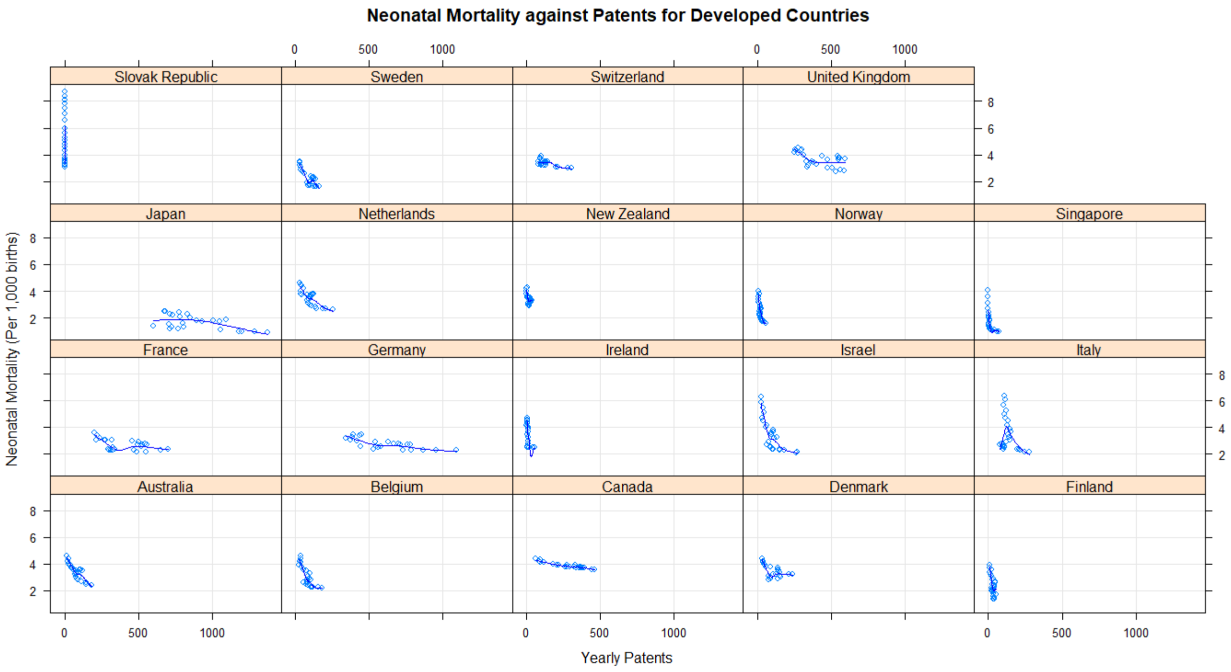


Graph 21:

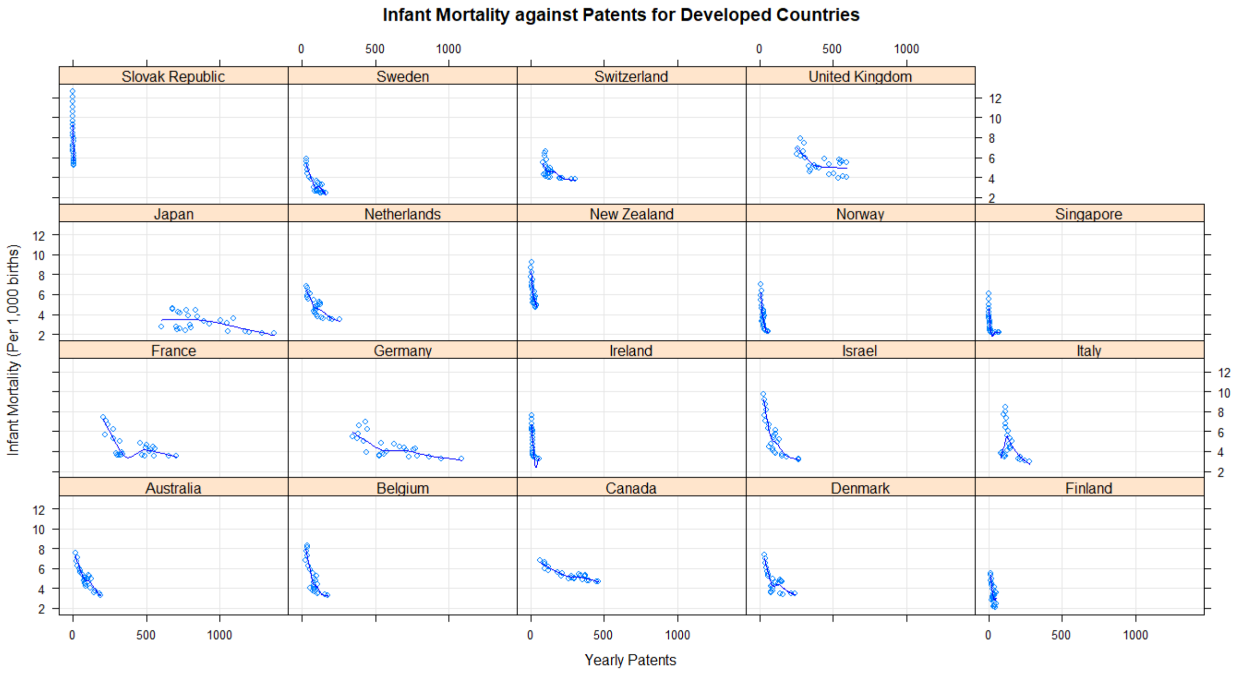


If we plot the same graphs without the United States, we see that the trend is more obvious, especially in Japan, France, Germany, Israel, United Kingdom, Switzerland,Canada, Italy and Denmark. The only difference in these countries with United States is that they have a lower number of yearly patents, if we assume a hypothetical world where the only possible predictor of average Life Expectancy, Neonatal mortality and Infant Mortality is the yearly number of patents, it would seem as if the United States has a really high number of patents for no reason, they might be focusing on Quantity instead of quality, but this situation is not reflective of reality in any way.

Graph 22:

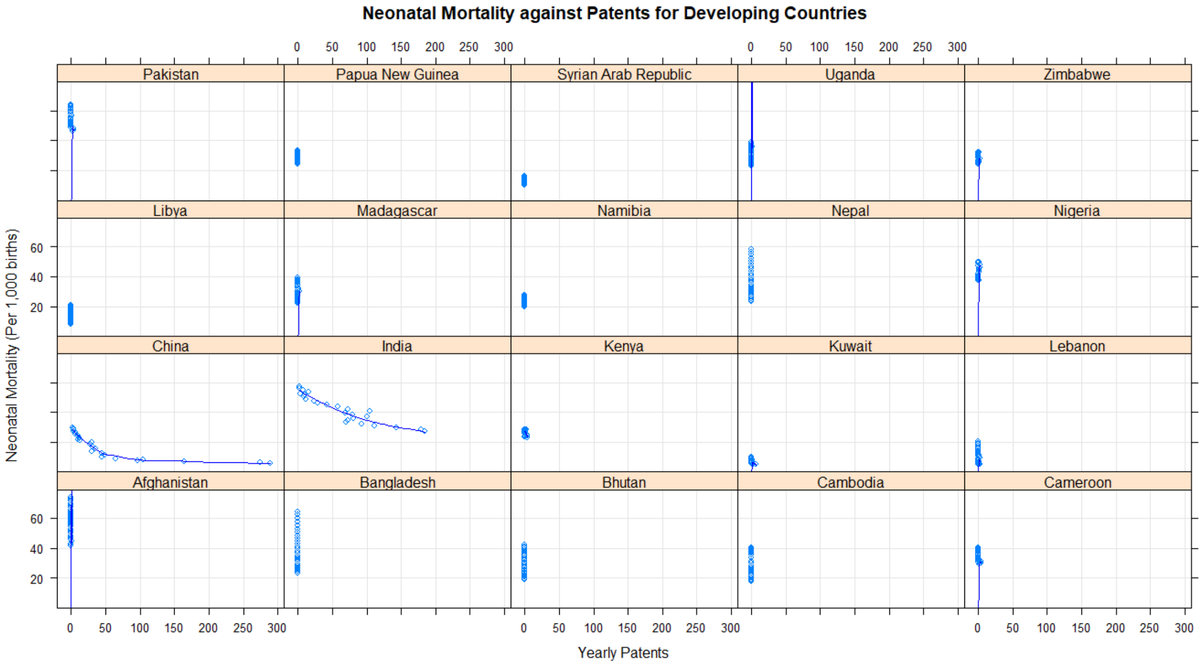


Graph 23:

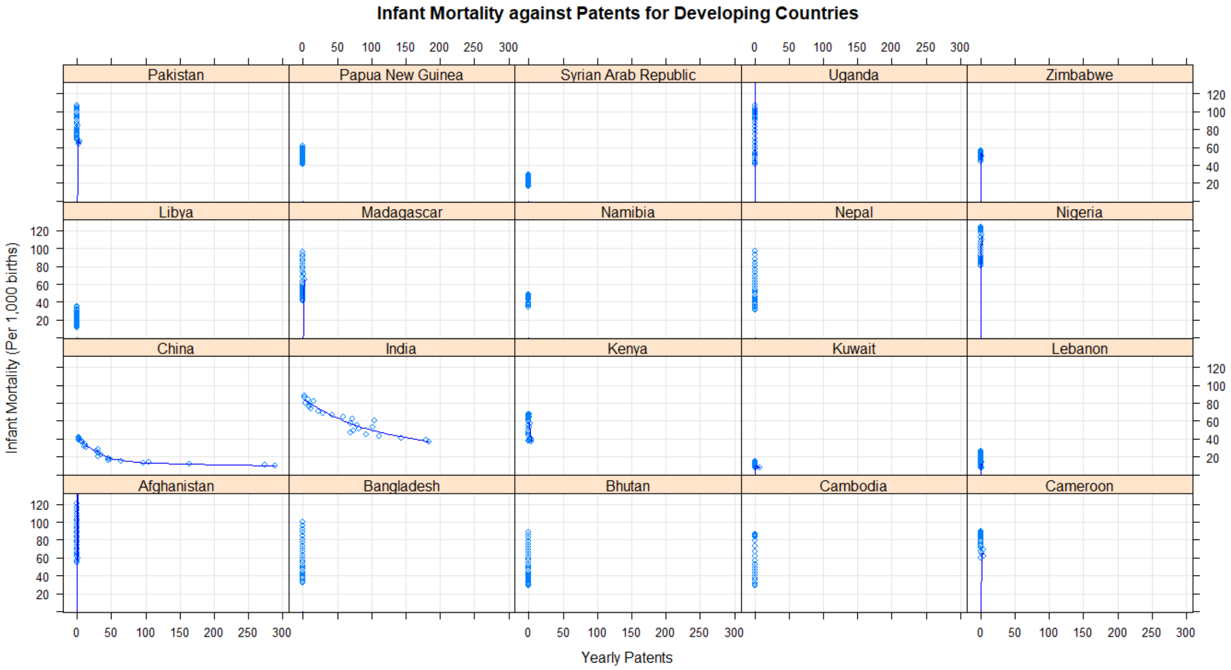


Plotting similar graphs for Developing Nations, we see that the trend is not visible in alot of countries. It is clearly visible in China and India, but some of the other countries do not even have any patents yet there is still variance in their average Life Expectancy, Neonatal rate and infant mortality rate. We will now plot these graphs without China and India to see whether the predicted trend holds.

Graph 24:

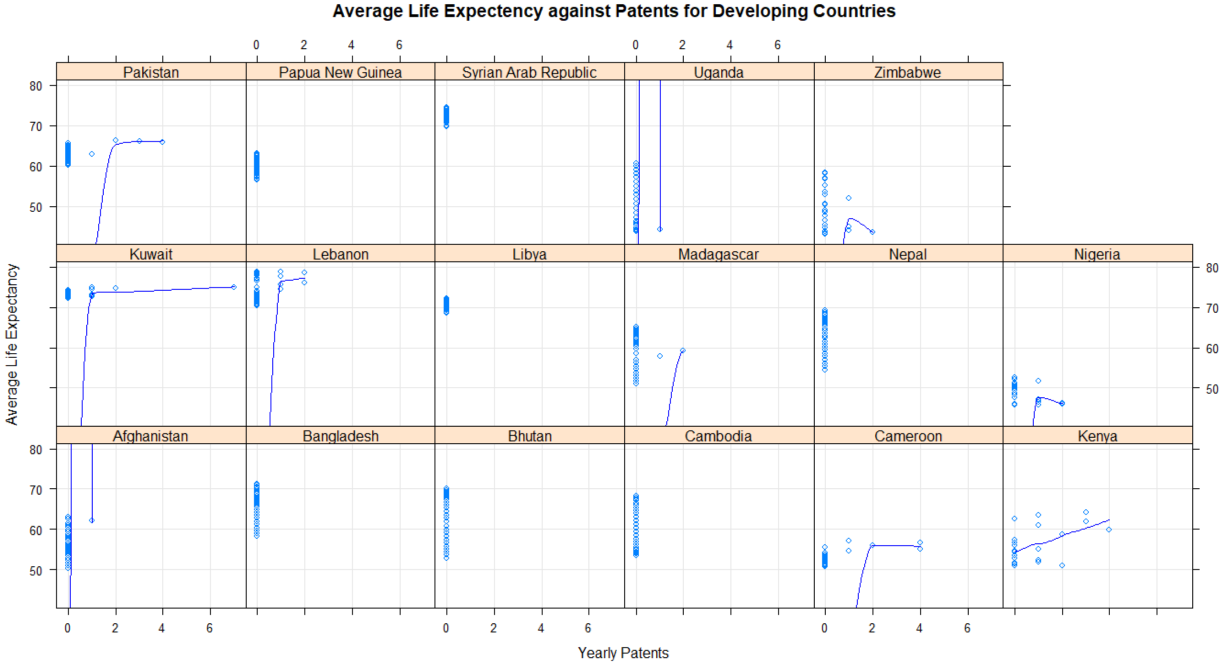


Graph 25:

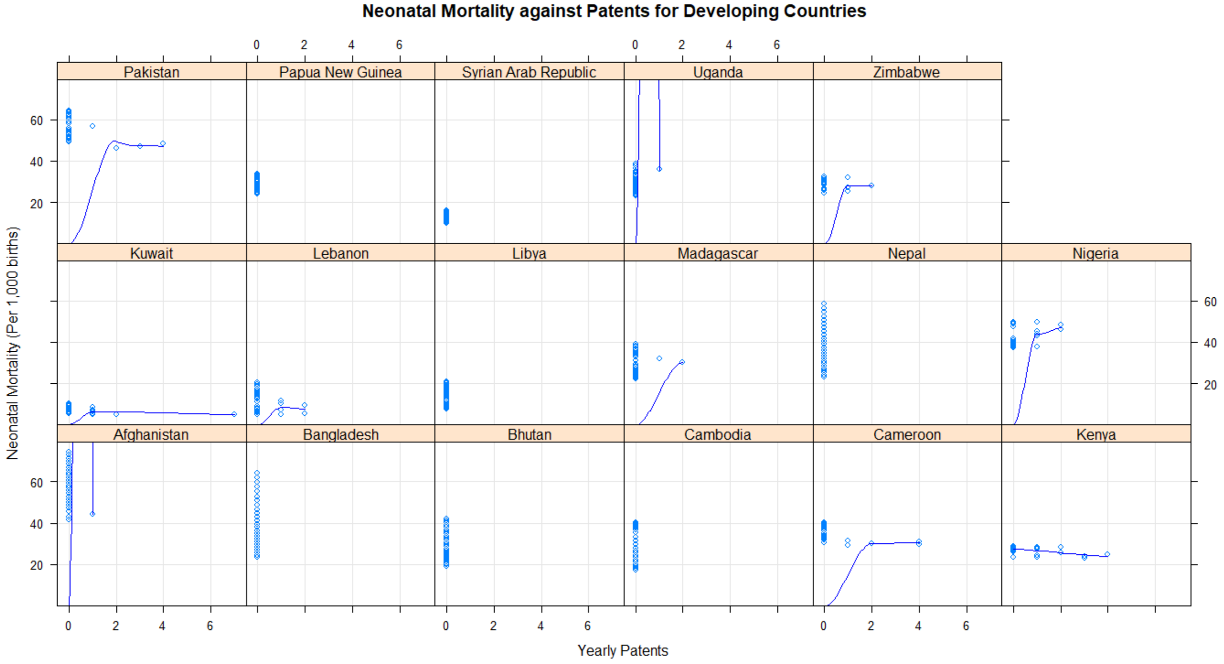


Plotting without China and India, we see that the trend does not hold true for most of the countries. An exception is Kenya, where the trend is perfectly as we predict. A possible reason for the average life expectancy increasing and the mortality rate decreasing for the countries without any patents might be that they have imported a lot of the medical technology that they use from developed countries.

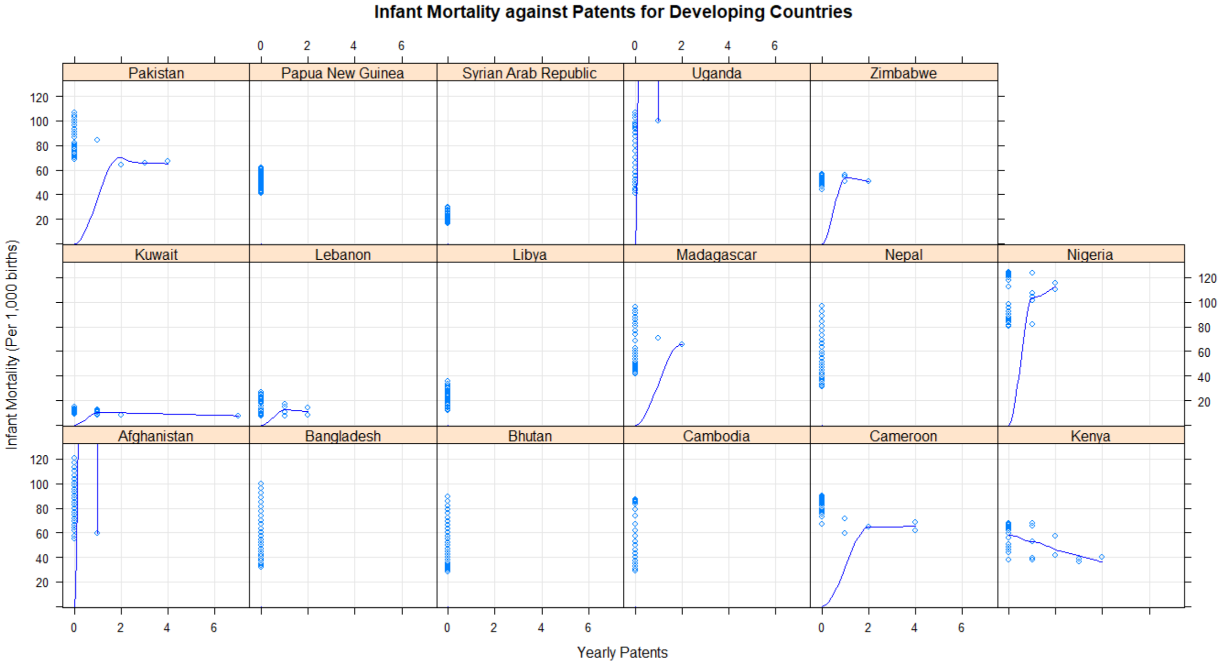
Graph 26:



Graph 27:



Graph 28:



Moving on, we will keep in mind that the United States (Due to its extremely large number of yearly patents) and India and China (same reason) may have a detrimental effect on our statistical testing. So, we will keep these 3 countries in mind while conducting our statistical inference because they are acting as outliers.

Statistical Inference

For our hypothesis testing, the statistical test that we chose was simple linear regression. The reason for that is that due to the nature of our variables and the fact that we only have one independent variable, we will individually run simple linear regression against each of the dependent variables. We feel like other statistical tests would not necessarily be useful for us.

We will run our statistical test on Developed Countries, Developing Countries, Developed Countries (Without the United States) and Developing Countries (without India and China).

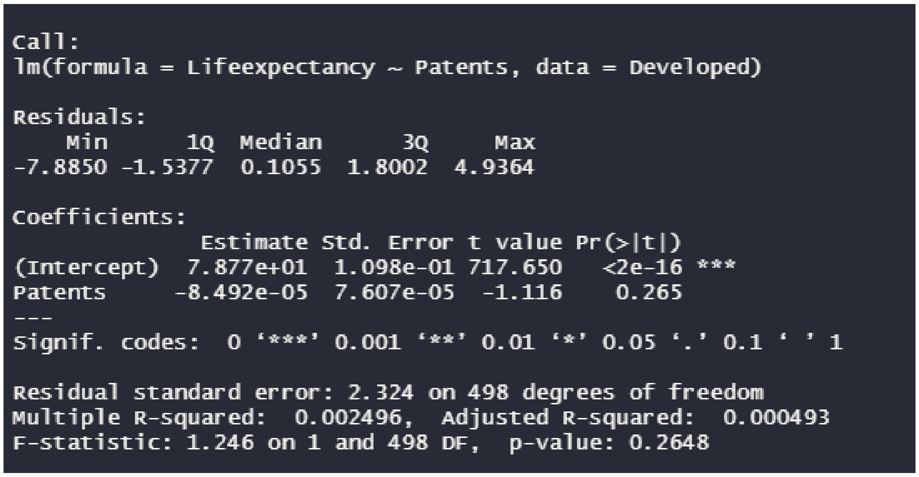
Developed Countries:

H0 = There will be no significant prediction of average Life Expectancy by Yearly Number of Patents

H1 = There is a significant prediction of average Life Expectancy by Yearly Number of Patents

Result: H0 is true. The p value is 0.2648 which is greater than 0.05, not significant at all.

Figure 22:



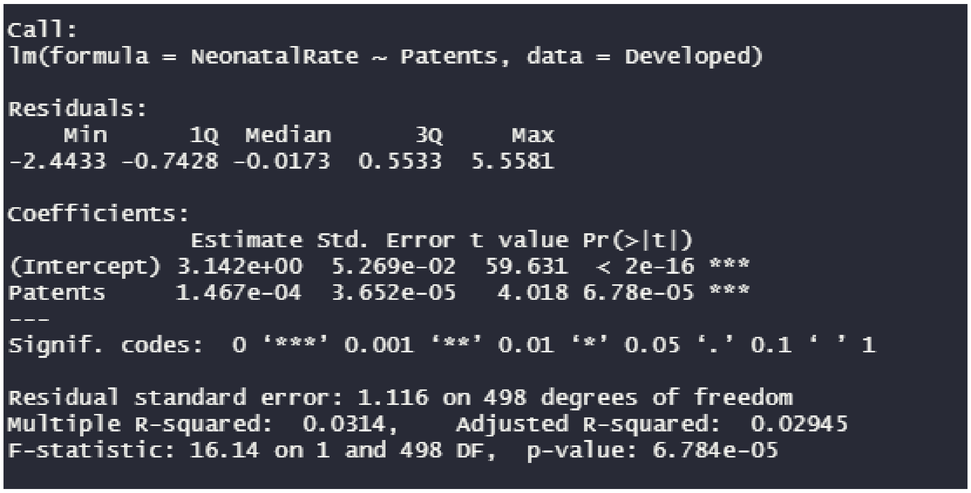
```
Call:
lm(formula = Lifeexpectancy ~ Patents, data = Developed)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.8850 -1.5377  0.1055  1.8002  4.9364 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.877e+01  1.098e-01 717.650   <2e-16 ***
Patents     -8.492e-05  7.607e-05  -1.116    0.265    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.324 on 498 degrees of freedom
Multiple R-squared:  0.002496,	Adjusted R-squared:  0.000493 
F-statistic: 1.246 on 1 and 498 DF,  p-value: 0.2648
```

H0 = There will be no significant prediction of Neonatal Mortality Rate by Yearly Number of Patents

H1 = There is a significant prediction of Neonatal Mortality Rate by Yearly Number of Patents

Result: The p-value is less than 0.05, so the prediction is highly significant meaning that H1 is true.

Figure 23:



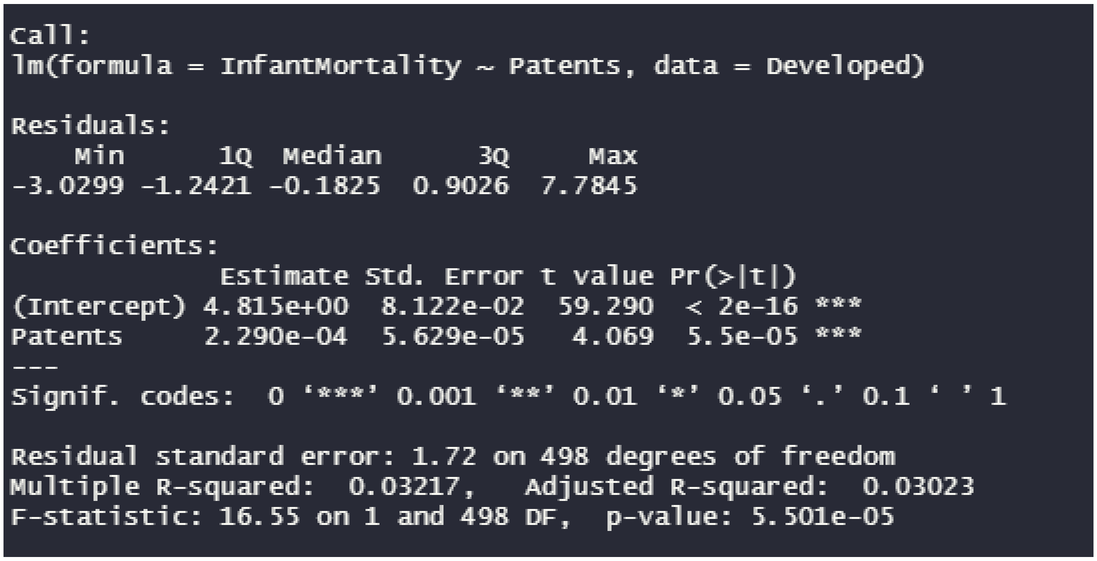
```
Call:
lm(formula = NeonatalRate ~ Patents, data = Developed)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4433 -0.7428 -0.0173  0.5533  5.5581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.142e+00  5.269e-02  59.631  < 2e-16 ***
Patents     1.467e-04  3.652e-05   4.018 6.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.116 on 498 degrees of freedom
Multiple R-squared:  0.0314,	Adjusted R-squared:  0.02945 
F-statistic: 16.14 on 1 and 498 DF,  p-value: 6.784e-05
```

H0 = There will be no significant prediction of Infant Mortality by Yearly Number of Patents

H1 = There is a significant prediction of Infant Mortality by Yearly Number of Patents

Result: H1 is true, there is a significant prediction of Infant mortality by Yearly number of Patents.

Figure 24:



```
Call:
lm(formula = InfantMortality ~ Patents, data = Developed)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0299 -1.2421 -0.1825  0.9026  7.7845 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.815e+00  8.122e-02  59.290  < 2e-16 ***
Patents     2.290e-04  5.629e-05   4.069  5.5e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.72 on 498 degrees of freedom
Multiple R-squared:  0.03217,	Adjusted R-squared:  0.03023 
F-statistic: 16.55 on 1 and 498 DF,  p-value: 5.501e-05
```

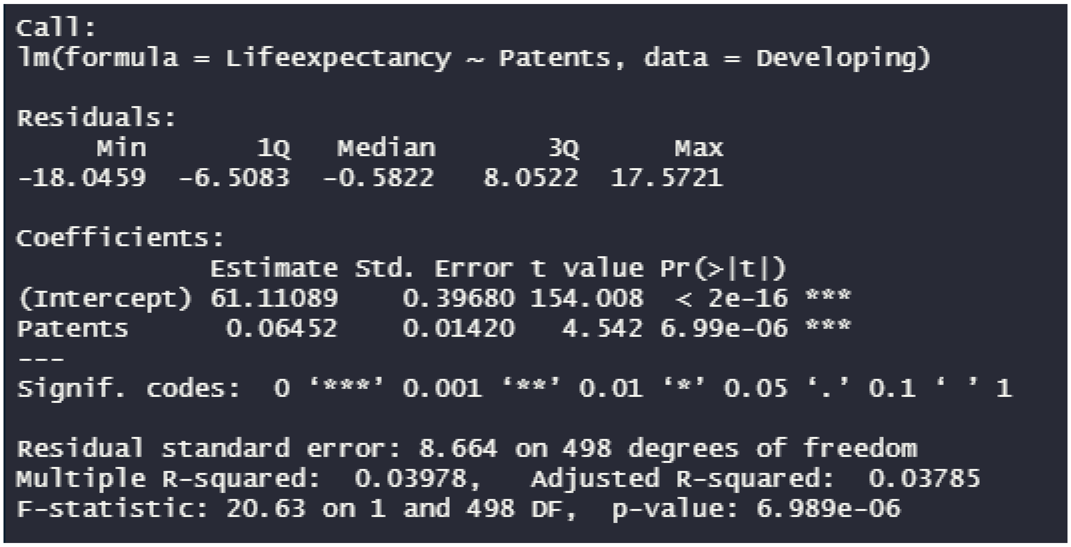
Developing Countries:

H0 = There will be no significant prediction of average Life Expectancy by Yearly Number of Patents

H1 = There is a significant prediction of average Life Expectancy by Yearly Number of Patents

Result: The p-value is smaller than 0.05, so H1 is true, average Life Expectancy is predicted by Yearly Number of Patents.

Figure 25:



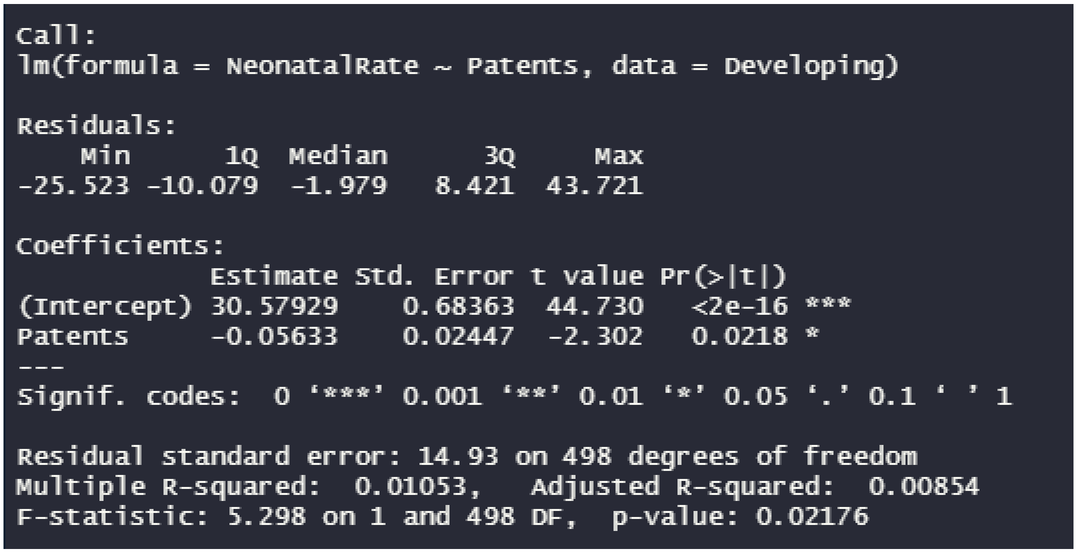
```
Call:
lm(formula = Lifeexpectancy ~ Patents, data = Developing)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.0459  -6.5083  -0.5822   8.0522  17.5721 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 61.11089    0.39680 154.008  < 2e-16 ***
Patents      0.06452    0.01420   4.542 6.99e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.664 on 498 degrees of freedom
Multiple R-squared:  0.03978,	Adjusted R-squared:  0.03785 
F-statistic: 20.63 on 1 and 498 DF,  p-value: 6.989e-06
```

H0 = There will be no significant prediction of Neonatal Mortality Rate by Yearly Number of Patents

H1 = There is a significant prediction of Neonatal Mortality Rate by Yearly Number of Patents

Result: The p-value is less than 0.05, so H1 is true, there is a significant prediction of Neonatal Mortality Rate by Yearly Number of Patents.

Figure 26:



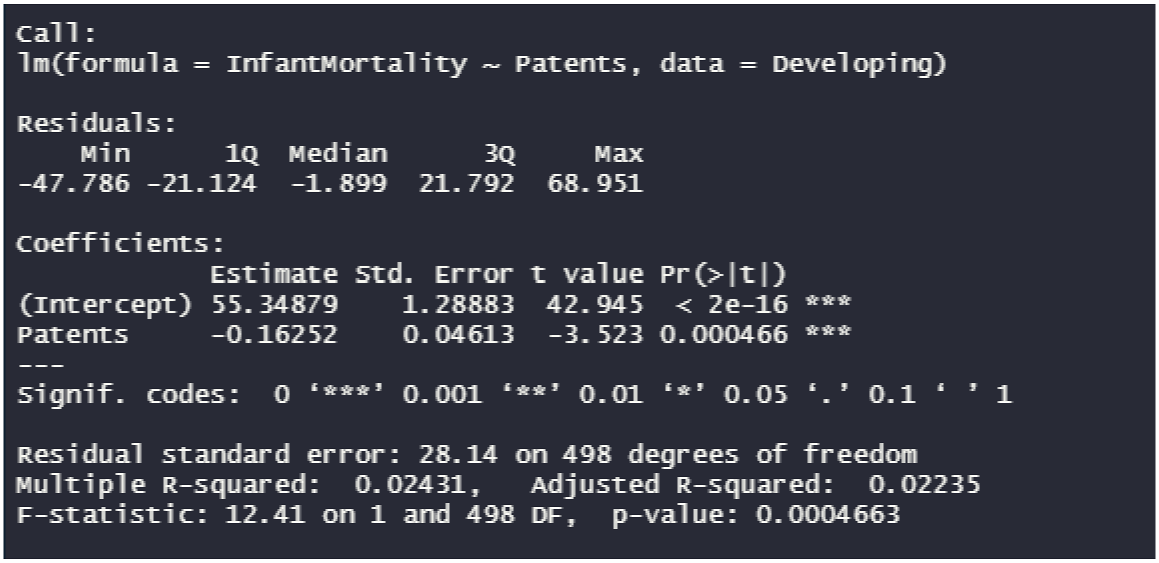
```
Call:
lm(formula = NeonatalRate ~ Patents, data = Developing)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.523 -10.079  -1.979   8.421  43.721 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.57929    0.68363  44.730   <2e-16 ***
Patents     -0.05633    0.02447  -2.302   0.0218 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.93 on 498 degrees of freedom
Multiple R-squared:  0.01053,	Adjusted R-squared:  0.00854 
F-statistic: 5.298 on 1 and 498 DF,  p-value: 0.02176
```

H0 = There will be no significant prediction of Infant Mortality by Yearly Number of Patents

H1 = There is a significant prediction of Infant Mortality by Yearly Number of Patents

Result: The p-value is less than 0.05, H1 is true, there is a significant prediction of Infant Mortality by Yearly Number of Patents.

Figure 27:



```
Call:
lm(formula = InfantMortality ~ Patents, data = Developing)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.786 -21.124  -1.899  21.792  68.951 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 55.34879    1.28883  42.945  < 2e-16 ***
Patents     -0.16252    0.04613  -3.523 0.000466 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.14 on 498 degrees of freedom
Multiple R-squared:  0.02431,	Adjusted R-squared:  0.02235 
F-statistic: 12.41 on 1 and 498 DF,  p-value: 0.0004663
```

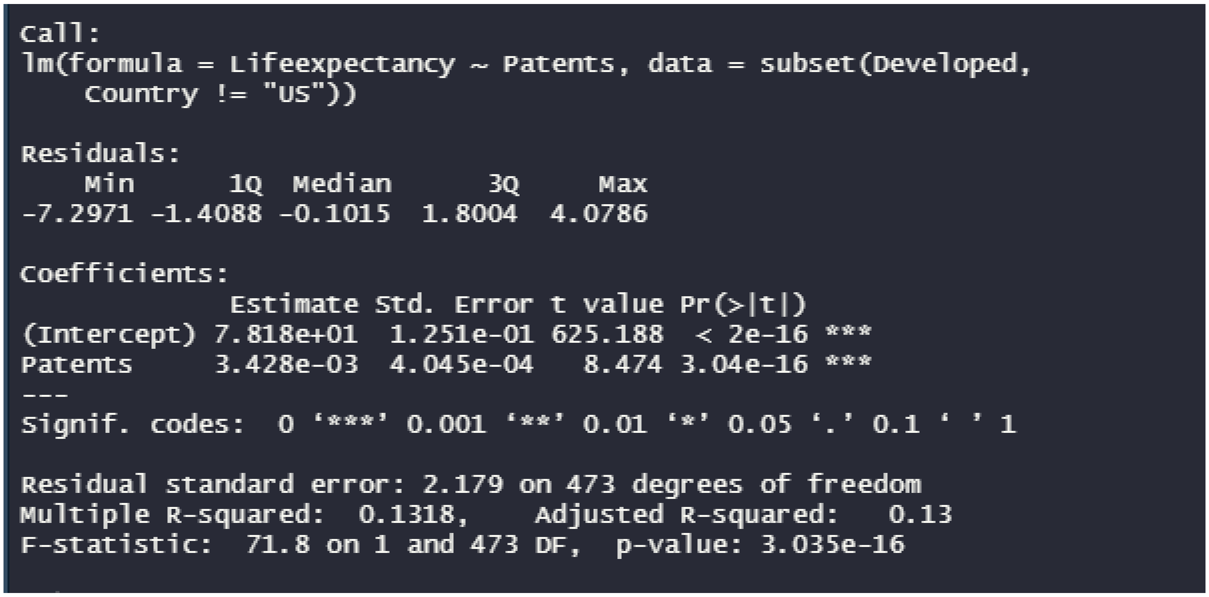
Developed Countries Without United States:

H0 = There will be no significant prediction of average Life Expectancy by Yearly Number of Patents

H1 = There is a significant prediction of average Life Expectancy by Yearly Number of Patents

Result: H1 is true as the p-value is less than 0.05, there is a significant prediction of average Life Expectancy by Yearly Number of Patents.

Figure 28:



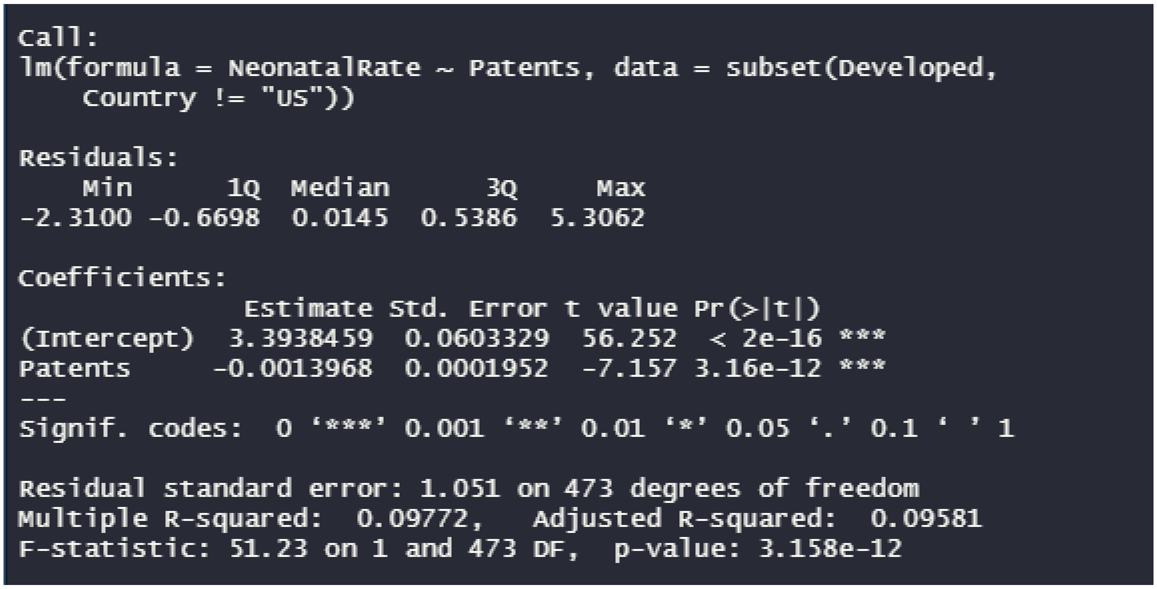
```
Call:
lm(formula = Lifeexpectancy ~ Patents, data = subset(Developed,
    Country != "US"))

Residuals:
    Min      1Q  Median      3Q     Max
-7.2971 -1.4088 -0.1015  1.8004  4.0786

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.818e+01  1.251e-01 625.188  < 2e-16 ***
Patents     3.428e-03  4.045e-04   8.474 3.04e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.179 on 473 degrees of freedom
Multiple R-squared:  0.1318,    Adjusted R-squared:   0.13
F-statistic:  71.8 on 1 and 473 DF,  p-value: 3.035e-16
```

H0 = There will be no significant prediction of Neonatal Rate by Yearly Number of Patents

H1 = There is a significant prediction of Neonatal Rate by Yearly Number of Patents

Result: H1 is true as the p-value is less than 0.05, there is a significant prediction of Neonatal Rate by Yearly Number of Patents.

Figure 29:



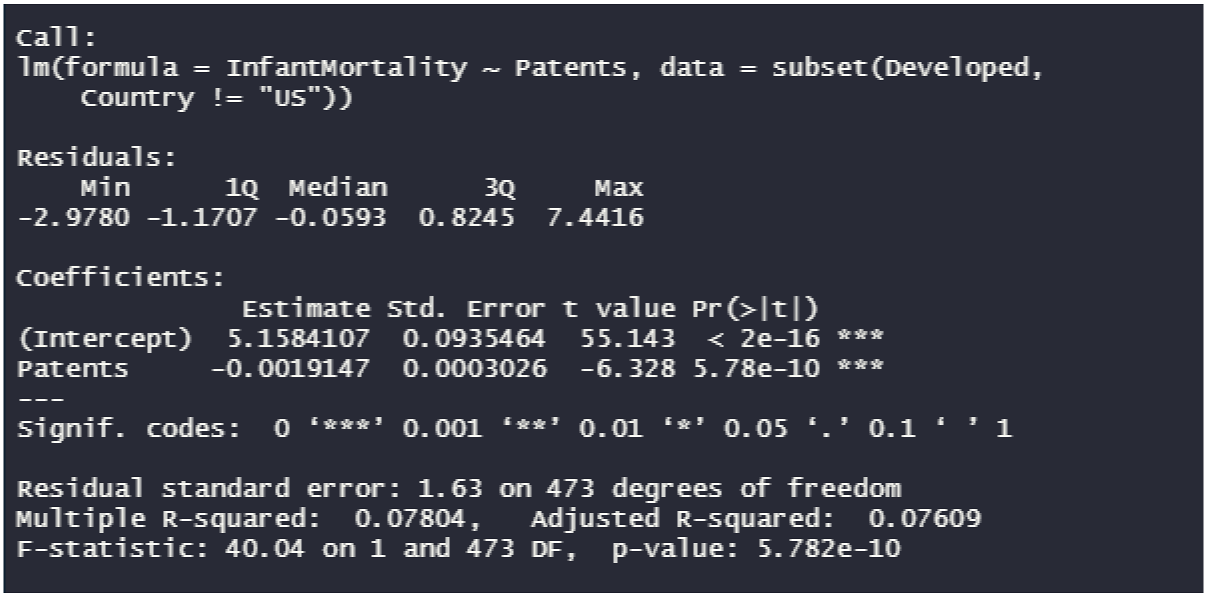
```
Call:
lm(formula = NeonatalRate ~ Patents, data = subset(Developed,
    Country != "US"))

Residuals:
    Min      1Q  Median      3Q     Max
-2.3100 -0.6698  0.0145  0.5386  5.3062

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.3938459  0.0603329  56.252  < 2e-16 ***
Patents     -0.0013968  0.0001952  -7.157 3.16e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.051 on 473 degrees of freedom
Multiple R-squared:  0.09772,   Adjusted R-squared:  0.09581
F-statistic: 51.23 on 1 and 473 DF,  p-value: 3.158e-12
```

H0 = There will be no significant prediction of Infant Mortality by Yearly Number of Patents

H1 = There is a significant prediction of Infant Mortality by Yearly Number of Patents

Result: H1 is true as the p-value is less than 0.05, there is a significant prediction of Infant Mortality by Yearly Number of Patents.

Figure 30:



```
Call:
lm(formula = InfantMortality ~ Patents, data = subset(Developed,
    Country != "US"))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9780 -1.1707 -0.0593  0.8245  7.4416 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.1584107  0.0935464  55.143  < 2e-16 ***
Patents     -0.0019147  0.0003026  -6.328 5.78e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.63 on 473 degrees of freedom
Multiple R-squared:  0.07804,	Adjusted R-squared:  0.07609 
F-statistic: 40.04 on 1 and 473 DF,  p-value: 5.782e-10
```

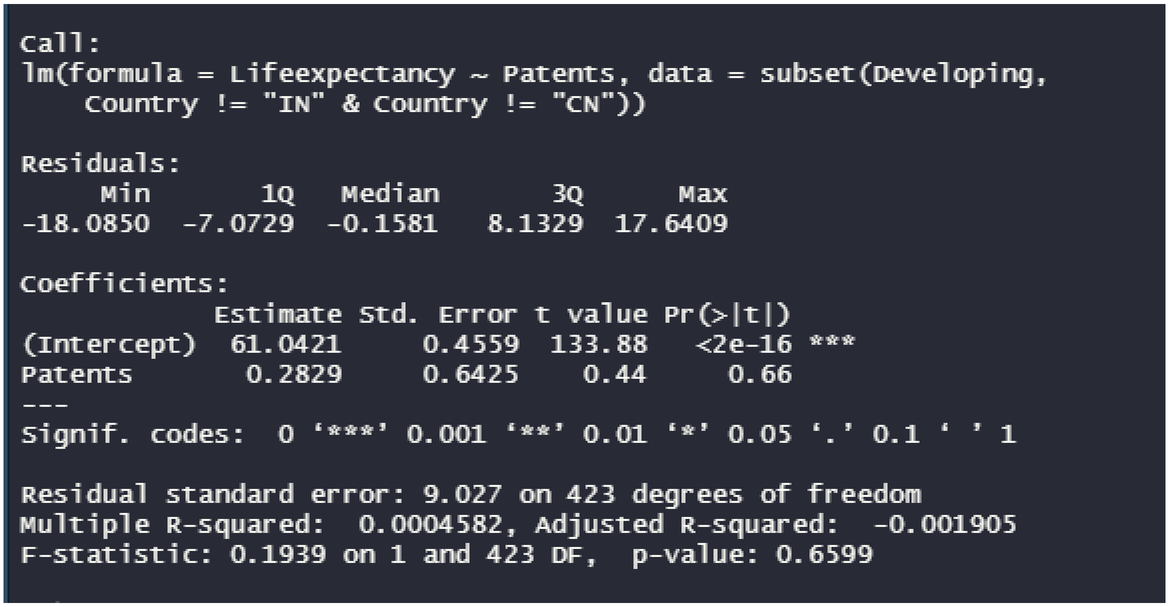
Developing Countries without China and India:

H0 = There will be no significant prediction of average Life Expectancy by Yearly Number of Patents

H1 = There is a significant prediction of average Life Expectancy by Yearly Number of Patents

Result: H0 is true as the p value is greater than 0.05, there is no significant prediction of average Life Expectancy by Yearly Number of Patents.

Figure 31:



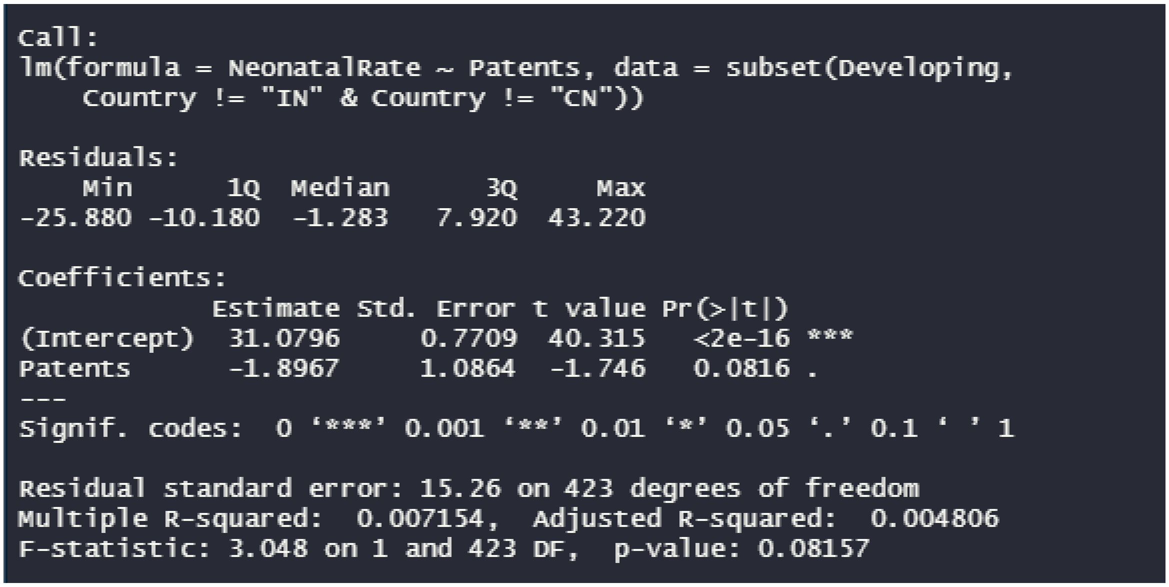
```
Call:
lm(formula = Lifeexpectancy ~ Patents, data = subset(Developing,
    Country != "IN" & Country != "CN"))

Residuals:
     Min       1Q   Median       3Q      Max 
-18.0850  -7.0729  -0.1581   8.1329  17.6409 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  61.0421     0.4559  133.88   <2e-16 ***
Patents       0.2829     0.6425    0.44     0.66    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.027 on 423 degrees of freedom
Multiple R-squared:  0.0004582, Adjusted R-squared:  -0.001905 
F-statistic: 0.1939 on 1 and 423 DF,  p-value: 0.6599
```

H0 = There will be no significant prediction of Neonatal Rate by Yearly Number of Patents

H1 = There is a significant prediction of Neonatal Rate by Yearly Number of Patents

Result: H0 is true as p-value is greater than 0.05, there will be no significant prediction of Neonatal Rate by Yearly Number of Patents.

Figure 32:



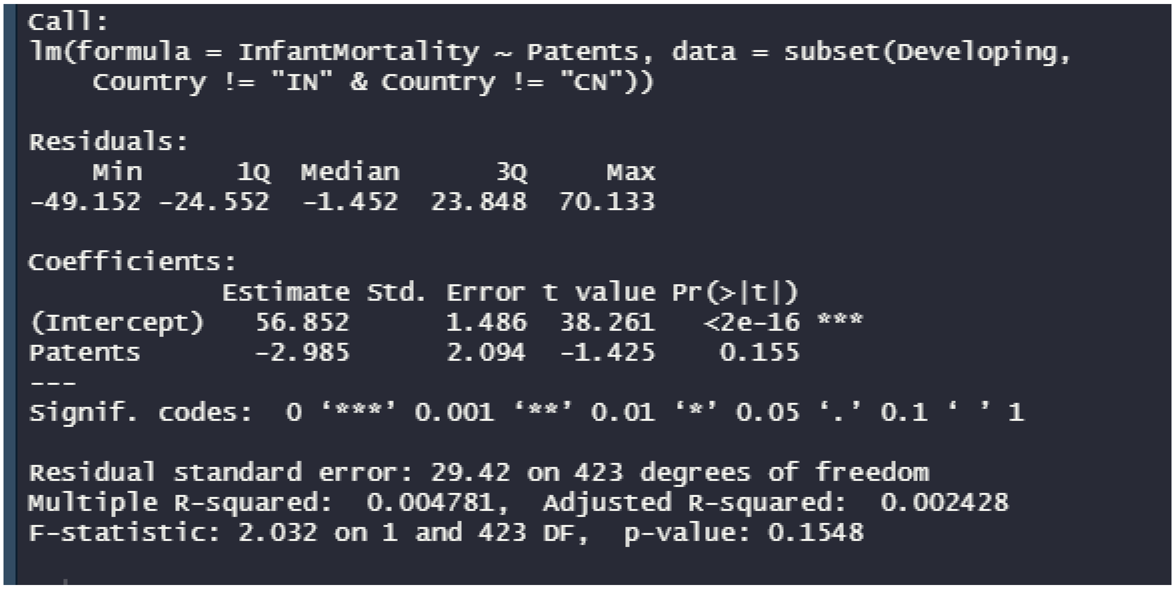
```
Call:
lm(formula = NeonatalRate ~ Patents, data = subset(Developing,
    Country != "IN" & Country != "CN"))

Residuals:
    Min      1Q  Median      3Q     Max 
-25.880 -10.180  -1.283   7.920  43.220 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.0796     0.7709  40.315   <2e-16 ***
Patents      -1.8967     1.0864  -1.746   0.0816 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.26 on 423 degrees of freedom
Multiple R-squared:  0.007154,  Adjusted R-squared:  0.004806 
F-statistic: 3.048 on 1 and 423 DF,  p-value: 0.08157
```

H0 = There will be no significant prediction of Infant Mortality by Yearly Number of Patents

H1 = There is a significant prediction of Infant Mortality by Yearly Number of Patents

Result: H0 is true as the p-value is greater than 0.05, there is no significant prediction of Infant Mortality by Yearly Number of Patents.

Figure 33:



```
Call:
lm(formula = InfantMortality ~ Patents, data = subset(Developing,
    Country != "IN" & Country != "CN"))

Residuals:
    Min      1Q  Median      3Q     Max
-49.152 -24.552  -1.452  23.848  70.133

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   56.852      1.486  38.261   <2e-16 ***
Patents       -2.985      2.094  -1.425    0.155
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.42 on 423 degrees of freedom
Multiple R-squared:  0.004781,  Adjusted R-squared:  0.002428
F-statistic: 2.032 on 1 and 423 DF,  p-value: 0.1548
```

Limitations

1. We only had 20 countries (Possible Selection Bias) each for the developing and developed nations datasets, even though they were picked as randomly as possible, there is generally some human error and bias in picking countries (For example, we may have subconsciously chosen Pakistan since we are Pakistanis). Other than that, there are 195 total countries on earth. Only 40 countries may not necessarily reflect the trends for the entire world. So another analysis with big data consisting of all the countries may be a better method
2. Even if there are new medical advancements (new medicine, machinery e.t.c), they may not necessarily be affordable by a lot of a country's population or hospitals
3. Medical advancements may be imported and exported from one country to another (Flow of knowledge/information)
4. Our dataset might not have been extensive enough, patents from European Patent Office, Japan Patent Office, Canadian Intellectual property office, Korean Intellectual Property Office e.t.c were not included (Analytics Bias)

Summary of the hypothesis testing

For developed countries, the yearly number of patents predicts Infant and Neonatal mortality, however it does not predict average life expectancy. For developing countries, the yearly number of patents predicts Infant, Neonatal mortalities and average life expectancy. We will not use these for our conclusion due to the extensive limitations of our project.

In Developed countries, United States held way too many yearly number of patents compared to the other nations (this conclusion was drawn from our exploratory graph plotting phase; we did not use a T-Test to compare the mean number of patents with other Nations as we believe the data will not be normally distributed (n<30)). The possible reasons:

1. United States is the only country which is focusing a lot on the medical field
2. Our dataset was not extensive enough, a lot of developed countries are european, so maybe we should have gathered data from the European Patent Office too
3. United States has the 3rd largest population, meaning for researchers in the medical field and thus more patents

For the Developing countries, similar to the United States, China and India held a lot more yearly patents when compared to the other developing nations. The possible reasons are:

1. They both are in the top countries population wise
2. The other nations may have relative to China and India harsher socioeconomic factors which may not necessarily support medical advancements.

Thus, we ran the test again, excluding the United States from the developed country dataset and China and India from the developing country dataset.

For developed countries, the Yearly Number of Patents significantly predicted average Life Expectancy and Neonatal and Infant Mortality rates. However, for developing countries, we see that the yearly number of patents do not predict average life expectancy, neonatal mortality and infant mortality.

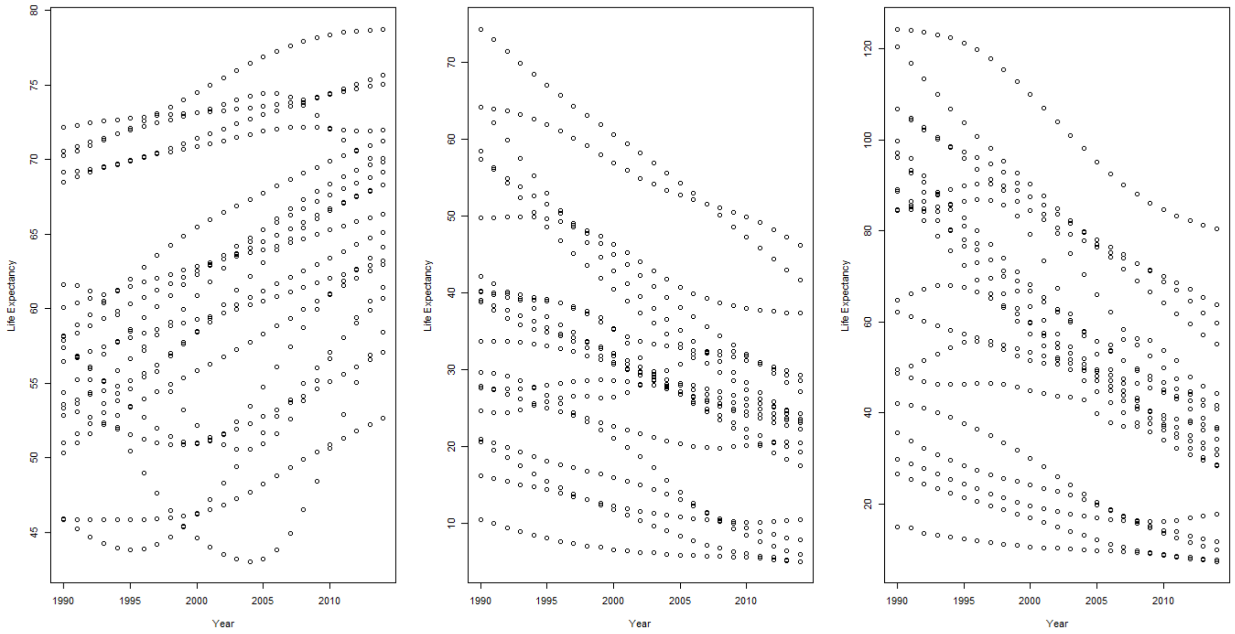
Conclusion

From the statistical tests, we can see that our prediction holds true for most of the Developed Countries however it does not hold true for most of the developing countries. If we wholly look at the developed countries dataset, completely ignoring the developing countries, it makes sense that the average life expectancy is increasing and neonatal and infant mortality rates are going down. This can easily be explained by thinking that as the medical and biotechnology field improves the population of a country will have better access to more health services.

For the developing countries, most of the countries did not have any significant number of patents in the medical field (if any at all) so there is no reason as to why that would be able to significantly predict average life expectancy, neonatal mortality and infant mortality. However, if we plot the year against life expectancy, neonatal mortality and infant mortality for the developing countries, we can see that even if the baseline differs, the average life expectancy for most countries, if not all, is generally increasing and the Neonatal and Infant mortality rates are going down.

If we assume that these countries have no significant patents from any other Patent Offices, it reflects another limitation of our project. New medical advancements are generally imported/exported by countries too, for example the US might have a new invention that may be used by Syria and/or even by Russia.

Graph 29:


Life Expectancy
80
75
70
65
60
55
50
45
1990
1995
2000
2005
2010
Year
Life Expectancy
70
60
50
40
30
20
10
1990
1995
2000
2005
2010
Year
Life Expectancy
120
100
80
60
40
20
1990
1995
2000
2005
2010
Year