Per Capita Metrics as Predictors of Tuberculosis Infection Rates: Data Analysis

Author: James Topor

Contents

Introduction **2**

Research Questions **2**

Data to be Used **3**

Instructions for Use: How to Reproduce the Results of the Data Collection Process **4**

R Packages Used for Analysis **5**

Finding "Hot Spots": High TB Case Counts & Infection Rates, 1995 - 2013 **6**

Average Annual Number of TB Cases per Country, 1995 - 2013 6

Average Annual TB Infection Rates per Country, 1995 - 2013 9

Cause For Alarm: High Case Counts + High Infection Rates 13

Trend Analysis: Changes in TB Case Counts & Infection Rates, 1995 - 2013 **14**

Trends in Total TB Cases Worldwide, 1995 - 2013 15

Trends in Annual Counts of TB Cases per Country, 2001 vs. 2012 16

Trends in Annual TB Infection Rates per Country, 2001 vs. 2012 21

Trends in Annual TB Case Counts & Infection Rates for Select Countries, 1995 - 2013 26

Per Capita Metrics & Their Relationship to TB Infection Rates **33**

Life Expectancy at Birth 34

Annual Health Care Expenditure Per Capita 38

Gross National Income (GNI) Per Capita 42

Access to Electricity (percentage of population) 46

Average Years of Schooling 50

"Describing" Countries Using Per Capita Metrics **54**

A Multivariate Linear Model for TB Infection Rates? **55**

Conclusion **58**

Introduction

The World Health Organization (WHO) states that tuberculosis (TB) is one of the leading bacterial-based causes of death worldwide. It is one of the top 5 causes of death among women aged 15-44, and in 2014 alone killed approximately 1.5 million people throughout the world. Furthermore, the WHO states that "About one-third of the world's population has latent TB, which means people have been infected by TB bacteria but are not (yet) ill with the disease and cannot transmit the disease" and "Over 95% of TB deaths occur in low- and middle-income countries".

In fact, 22 countries alone account for roughly 80% of the world's TB cases:

http://www.stoptb.org/countries/tbdata.asp

Based on these facts, the world's health authorities clearly face a major ongoing challenge in managing and preventing the spread of TB infections. The purpose of this study will be to explore the relationship between TB infection rates and other country-specific "per capita" metrics for 100 countries throughout the world and attempt to quantify some common characteristics of countries having either low or high rates of TB infections.

Research Questions

In this study we will attempt to address the following questions:

- Do countries that have high incidence of TB share certain ***measurable*** attributes?
- Do countries that have relatively lower incidence of TB share certain ***measurable*** attributes that countries with high incidences of TB should perhaps strive to emulate?

For purposes of this study, the "measurable attributes" we will consider will be country-specific "per capita" metrics.

In addition to the gross number of TB cases recorded within a country for a given year, the "per capita" metrics we will examine are as follows:

- TB infection rates per capita per country;
- Life Expectancy at Birth (in years);
- Health Care Expenditure per capita;
- Gross National Income (GNI) per capita;
- The percentage of a country's population having access to electricity.
- The average number of years of schooling per capita.

Using these data we will attempt to "describe" the attributes of countries that have varying degrees of TB incidence via the use of tools such as regression models and boxplots. The results of that analysis will then be used to identify benchmarks that countries with high incidence of TB might need to achieve if they wish to reduce their overall incidence of TB infections.

Data to be Used

The Week 3 Assignment of the Spring 2016 MSDA 607 class made use of a collection of TB and population data for 100 countries for the years 1995 - 2013. We'll make use of that data as well as several other sources of per capita metrics. The complete list of the data sources used herein is as follows:

Tuberculosis Infection Counts

- A MySQL database table that tells us how many individuals of a given age category ('child', 'adult', or 'elderly') of a given gender were diagnosed with tuberculosis within a given country (100 countries in total) for a given year for the years 1995 - 2013.

Country Population Counts

- A CSV file containing population counts for 100 countries for the years 1995 - 2013.
- https://raw.githubusercontent.com/jtopor/CUNY-MSDA-607/master/Final%20Project/Data/population.csv

Life Expectancy at Birth

- The average life expectancy for the populations of 100 countries for the years 1995 - 2013.
- http://data.worldbank.org/indicator/SP.DYN.LE00.IN

Annual Health Care Expenditure Per Capita

- Total annual health expenditure per capita is the sum of public and private health expenditures as a ratio of total population of a country. Data are in U.S. dollars and cover the years 1995 - 2013.
- http://data.worldbank.org/indicator/SH.XPD.PCAP

Gross National Income (GNI) Per Capita

- The per capita Gross National Income of 100 countries for the years 1995 - 2013.
- http://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD

Access to electricity (percentage of population)

- The percentage of a country's population having access to electricity in their homes for 100 countries for the years 2000, 2010, 2012.
- http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?page=1

Average years of Schooling

- The average years of schooling for the populations of 100 countries for the years 2000, 2005 - 2012.
- http://hdr.undp.org/en/content/mean-years-schooling-adults-years

Instructions for automatically collecting and organizing data from these sources are provided below.

Instructions for Use: How to Reproduce the Results of the Data Collection Process

The data collection process is comprised of two key components:

- Creating the required MySQL database schema and tables via a provided MySQL script file;
- Running a separately provided R Markdown file named **"FP__Data__Loader.Rmd"** to automatically load the relevant per capita metric data.

*The steps outlined below provide detailed instructions that you **MUST** adhere to for purposes of reproducing the results of the data collection process relied upon herein. Failure to follow these steps will prevent the user from reproducing the results we've generated here.*

Step 1:

Ensure you have access to the following Rmd and .csv files:

- **FP__Data__Loader.Rmd**: An R Markdown file that performs the data acquisition, data cleansing, and database loading tasks necessary as a prerequisite for running ***this*** R Markdown file.

https://raw.githubusercontent.com/jtopor/CUNY-MSDA-607/master/Final%20Project/FP__Data__Loader.Rmd

- **tb__db__createscript.sql**: An SQL script used to create the required schema and database tables. It is accessible at the following GitHub link:

https://raw.githubusercontent.com/jtopor/CUNY-MSDA-607/master/Final%20Project/tb__db__createscript.sql

The following six files can be found in a folder at this GitHub page:

https://github.com/jtopor/CUNY-MSDA-607/tree/master/Final%20Project/Data

- ElecAccess.csv
- GNI.csv
- HC_Spending.csv
- LifeExpec.csv
- population.csv
- tb.csv

Step 2:

Ensure you have MySQL Server up and running properly on your local machine. Failure to do so will prevent the SQL script and various R code modules described below from functioning properly.

Step 3:

Load the '**tb_db_createscript.sql**' script into your local MySQL environment and execute it. If successful, you will find a new schema by the name of '**tb_prediction**' within your MySQL Server environment. Within that schema you will find several tables that have been purpose-built for this application.

Step 4: Load and execute the **FP_Data_Loader.Rmd** file within your local R environment to automatically load the relevant data from the data sources described in the **Data Used** section (see above) into the MySQL database.

IMPORTANT!!!
Please note that you will need to open the **FP_Data_Loader.Rmd** file and set the MySQL server reference name used by your own local instance of MySQL server within the "*con <- odbcConnect("local_server")*" function call. Failure to do so will prevent the **FP_Data_Loader.Rmd** code from functioning properly.

Upon completion of these steps you will have reproduced the data set that serves as the basis for the analysis discussed below.

R Packages Used for Analysis

Several R packages are required for the analysis performed herein. Specifically, the **tidyr**, **dplyr**, **rworldmap**, **maptools**, and **RODBC** packages must be installed within your local R environment prior to running the code contained within this R Markdown file. The **knitr** package is also required for purposes of properly formatting the on-screen appearance of portions of the output of the processes contained herein. The packages are loaded as follows:

```r
# Load Packages
library(knitr)
library(RODBC)
library(ggplot2)
suppressMessages(library(rworldmap))
suppressMessages(library(maptools))
library(dplyr)
library(tidyr)
options(stringsAsFactors = FALSE)
```

Prior to starting the analysis we'll establish a connection with the MySQL server:

```r
# establish connection to local SQL server
# NOTE: Be sure to set the server reference appropriately to comport with your own local
# computing environment.
con <- odbcConnect("local_server")

# select a database to use - in this case, the 'tb_prediction' database
x <- sqlQuery(con, "use tb_prediction")
rm(x)
```

Finding “Hot Spots”: High TB Case Counts & Infection Rates, 1995 - 2013

As discussed earlier, the tuberculosis case count data set we are using covers 100 countries throughout the world for the years 1995 - 2013. We’ll start our analysis by identifying what we’ll refer to as TB “hot spots”: countries that have relatively high TB case counts and / or high relative rates of TB infection for the 19 year timeframe encompassed by the data set.

Specifically, we’ll calculate the average number of TB cases recorded annually for the years 1995 - 2013 for each of the 100 countries for which we have data. We’ll also calculate the ***rate*** of TB infections per capita within each of those countries by deriving the average number of TB cases per 100,000 people as recorded within the data set for the years 1995 - 2013.

The results of these calculations will allow us to generate geoplots that highlight TB “hot spot” countries. We’ll also be able to compare the results of both sets of calculations to see how well the countries with the highest average gross incidence of TB cases match up to the countries with the highest average ***rates*** of infection per capita. Countries that have both high TB case counts and high rates of infection throughout their populations would be of particular interest to the world’s various health agencies.

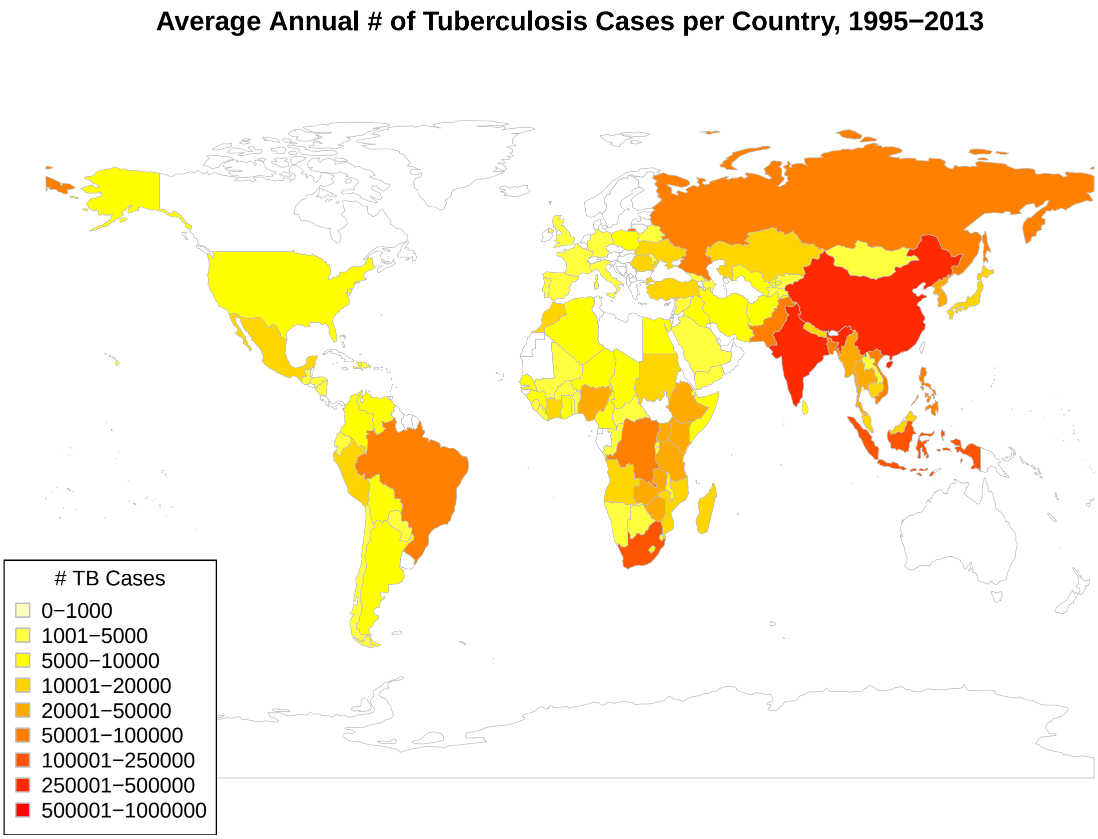
We can retrieve the number of TB cases for each of our 100 countries for the years 1995 - 2013 as follows:

```r
# get counts of cases by country, year
tb_df <- sqlQuery(con, "SELECT country, year, SUM(child + adult + elderly)
                        FROM tb
                                    GROUP BY country, year
                        ORDER BY country, year", stringsAsFactors=F)

# rename third column to meaningful name
colnames(tb_df)[3] <- "cases"

# Ensure country names match those used by R's geomapping tools:
# first, fetch R country names from database lookup table
rmap_df <- sqlQuery(con, "SELECT * FROM rmap_lookup", stringsAsFactors=F)

# then update non-matching country names within data frame
for(i in 1:nrow(rmap_df)) {
  tb_df$country[tb_df$country == rmap_df$tb_country[i]] <- rmap_df$rmap_country[i]
}
```

Average Annual Number of TB Cases per Country, 1995 - 2013

With the raw TB case counts loaded we can now calculate the average annual number of cases for each of the 100 countries. The results of those calculations are then plotted in a color-coded geoplot map that highlights the average annual case counts for each country.

```r
tb_means <- tb_df %>%
            group_by(country) %>%
            summarise(avg_cases = mean(cases, na.rm = TRUE))

# Sort the averages in descending order
tbm_sorted <- arrange(tb_means, desc(avg_cases))
```



The plot indicates that countries such as Russia, China, India, South Africa, Indonesia, Brazil, and many of the countries in sub-Saharan Africa have had relatively high absolute numbers of TB cases during the 1995 - 2013 time period.

In fact, the 20 countries with the highest average numbers of TB cases for the period are as follows:

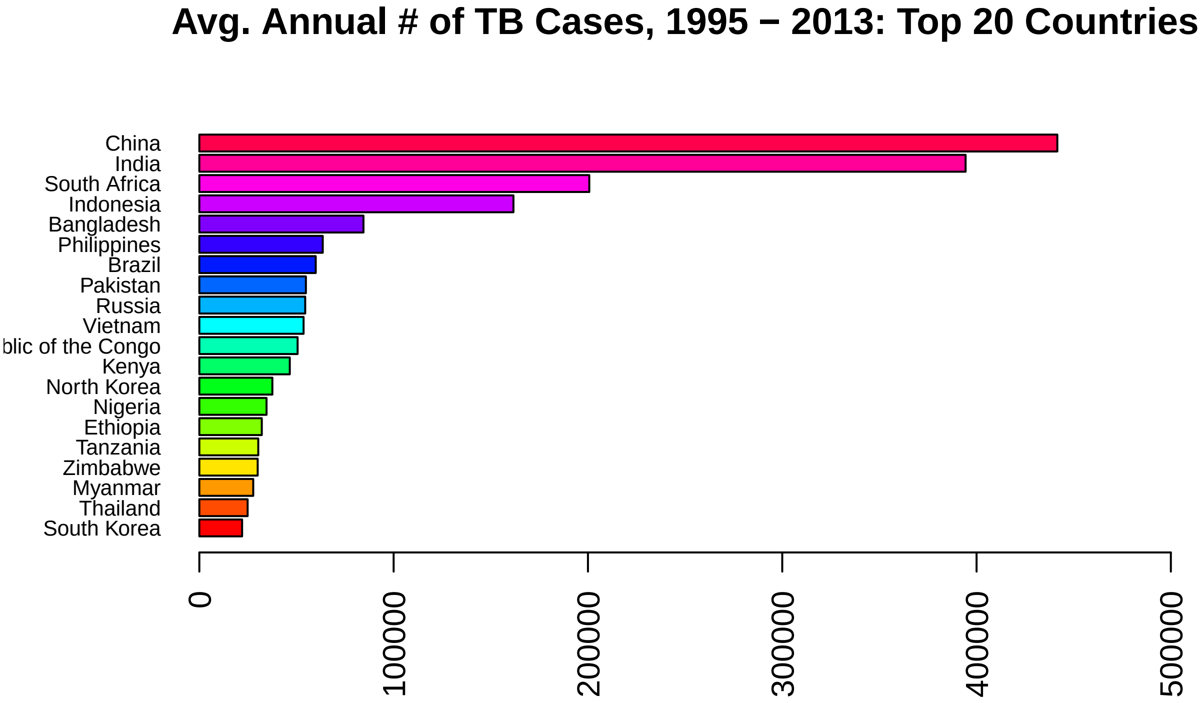
Table 1: Top 20 Countries by Avg # of TB Cases, 1995 - 2013

	country	avg_cases
1	China	441570.47
2	India	394364.00
3	South Africa	200684.80
4	Indonesia	161662.50
5	Bangladesh	84476.00
6	Philippines	63521.87
7	Brazil	59910.33
8	Pakistan	54844.75
9	Russia	54484.47
10	Vietnam	53648.06
11	Democratic Republic of the Congo	50573.79
12	Kenya	46554.84
13	North Korea	37596.94
14	Nigeria	34588.11

	country	avg_cases
15	Ethiopia	32143.59
16	Tanzania	30335.94
17	Zimbabwe	30034.58
18	Myanmar	27740.22
19	Thailand	24854.94
20	South Korea	22013.68

As we can see, China, India, and South Africa top the list, with Russia, Indonesia, and Brazil each placing within the top 10. However, the top three countries alone accounted for, on average, more than 1 million TB cases each year during the 1995 - 2013 time period.

We can plot the average case counts for the top 20 countries to get a sense of how widely those averages might vary from one another:



The barplot above clearly shows that the top 4 countries' average case counts (those of China, India, South Africa, and Indonesia) are dramatically higher than even those of the other 16 countries identified within the top 20 list, with both China and India having average case counts that are at least double those of any other countries in the list.

To get an idea of whether the average case counts for the top 20 countries might be extreme relative to those of the other 80 countries in our data set, we can use R's **summary** function to calculate the mean, median, and quantiles of the average case counts for all 100 countries:

```
summary(tbm_sorted$avg_cases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1438    3381    6681   24560   16680  441600
```

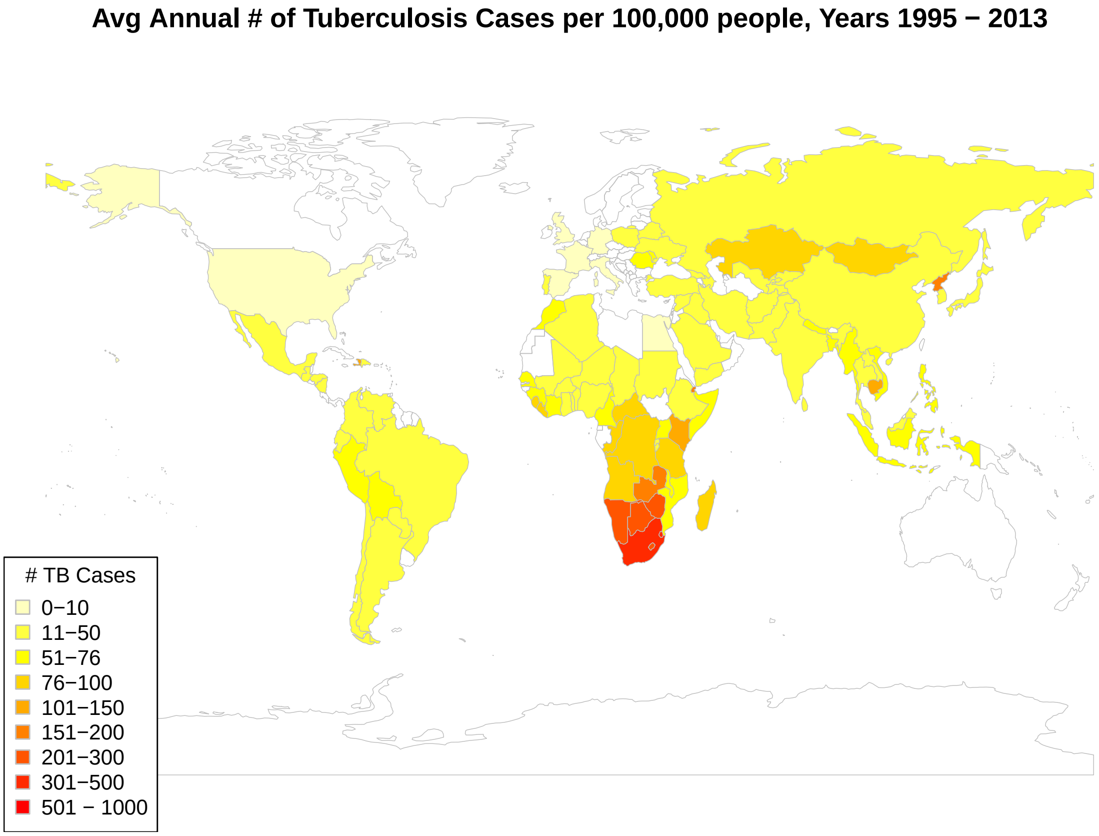
The output above shows that for the 100 countries in our data set we have a median annual number of cases of **6681** and a mean of **24560**. A mean value so much larger than the median indicates that we have a severely right-skewed distribution. Furthermore, we have an interquartile range (IQR) of **(3381, 16680)**, which indicates that 50% of the annual TB cases counts are within the range of **(3381, 16680)** and 75% of the average annual TB case counts are less than **16,680**. Therefore, the relatively low median value combined with the IQR appears to be supportive of the statement that the top 20 countries listed above have unusually high average TB case counts.

It is important to note here that case counts by themselves are not necessarily indicative of the pervasiveness of TB within a country. Absolute case counts do not account for the size of a country's population, so a country with a relatively high average TB case count might in fact have a relatively low overall rate of TB infection if those cases are spread over a relatively large population. However, our data set includes countries with both relatively large populations (e.g, the United states, Argentina, etc..) and relatively small populations. As such, it is reasonable to conclude that the averages of the annual TB case counts for the top 20 countries listed above are unusually high.

Average Annual TB Infection Rates per Country, 1995 - 2013

The SQL query and R code shown below calculates the average annual TB infection rates per 100,000 people for each of the 100 countries for the years 1995 - 2013. The results of those calculations are then plotted in a color-coded geoplot map that highlights the average TB infection rates for each country.

```r
tbrate_means <- sqlQuery(con, "SELECT DISTINCT country, AVG(rate)
                              FROM tb_rates
                              GROUP BY country
                              ORDER BY AVG(rate) DESC", stringsAsFactors=F)

# rename second column to meaningful name
colnames(tbrate_means)[2] <- "avg_rate"

# multiply rate by 100K to get TB infection rate per 100K people
tbrate_means$avg_rate <- tbrate_means$avg_rate * 100000
```


Avg Annual # of Tuberculosis Cases per 100,000 people, Years 1995 − 2013
# TB Cases
0−10
11−50
51−76
76−100
101−150
151−200
201−300
301−500
501 − 1000

The plot shows very clearly that the countries with the highest rates of TB infection per 100,000 people are located in sub-Saharan Africa, with the countries of North Korea, Kazakhstan, Mongolia, and Cambodia also appearing to have very high average rates of infection during the years 1995 - 2013.

The 20 countries with the highest average rates of TB infection for the period are as follows:

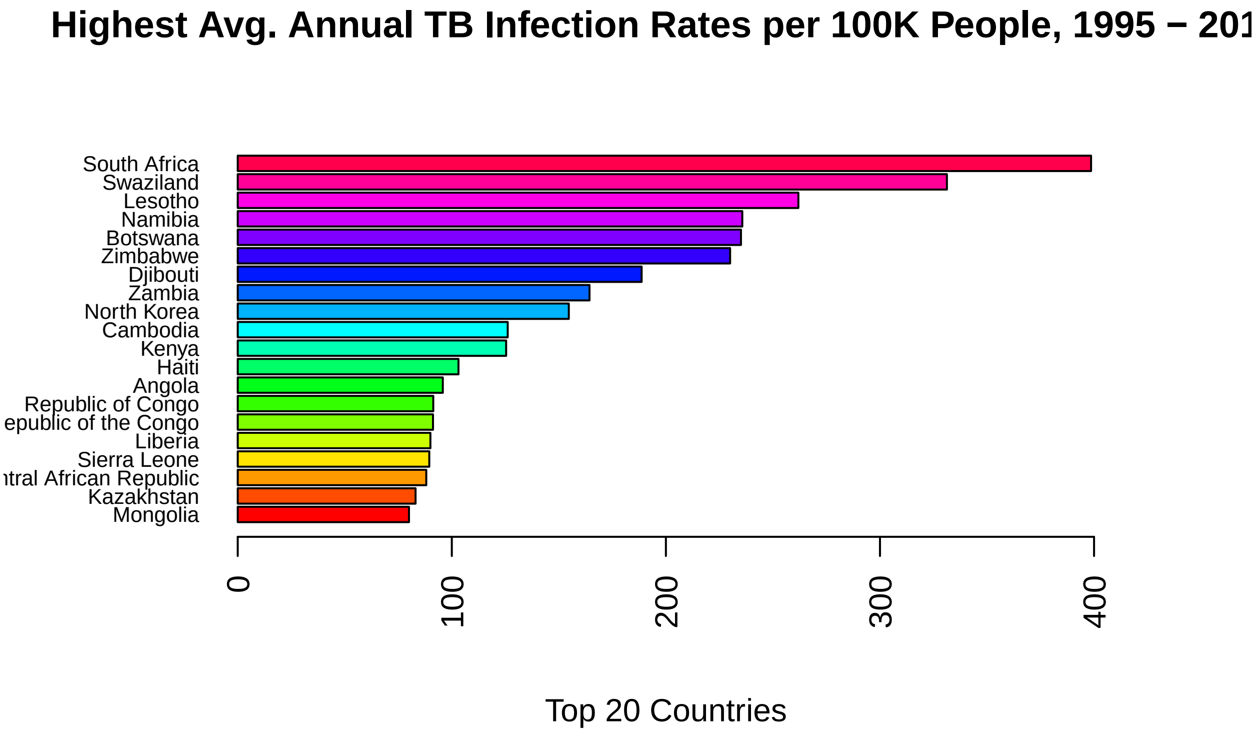
Table 2: Top 20 Countries by Avg Infection Rate per 100K People, 1995 - 2013

	country	avg_rate
1	South Africa	398.60000
2	Swaziland	331.26667
3	Lesotho	261.88235
4	Namibia	235.68421
5	Botswana	235.06250
6	Zimbabwe	230.00000
7	Djibouti	188.66667
8	Zambia	164.30769
9	North Korea	154.64706
10	Cambodia	126.11111
11	Kenya	125.36842
12	Haiti	103.11765
13	Angola	95.78947

	country	avg_rate
14	Republic of Congo	91.33333
15	Democratic Republic of the Congo	91.21053
16	Liberia	90.00000
17	Sierra Leone	89.47059
18	Central African Republic	88.06250
19	Kazakhstan	83.06250
20	Mongolia	80.00000

This list shows that we appear to have some overlap between the two "Top 20" country lists we've derived thus far, with countries such as South Africa and The Democratic Republic of the Congo appearing in both.

We'll have more to say about this "overlap" in a moment, but first we'll plot the average rates of TB infection for the top 20 countries listed above to get a sense of how widely those averages might vary from one another:



The barplot above clearly shows that the TB infection rates for the top 9 countries are dramatically higher than even those of the other 11 countries identified within the top 20 list, with the top 6 countries (South Africa, Swaziland, Lesotho, Namibia, Botswana, and Zimbabwe) having average rates of infection per 100K people that are more than double those of the countries 12-20 in the list.

To get an idea of whether the average TB infection rates for the top 20 countries shown here might be extreme relative to those of the other 80 countries in our data set, we can use R's **summary** function to calculate the mean, median, and quantiles of the average case counts for all 100 countries:

```
summary(tbrate_means$avg_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.842  24.480  41.680  62.010  69.260 398.600
```

The output above shows that for the 100 countries in our data set we have a median annual TB infection per 100K people of **41.68** and a mean of **62.01**. A mean value so much larger than the median indicates that we have a severely right-skewed distribution. Furthermore, we have an interquartile range (IQR) of **(24.48,69.26)**, which indicates that 50% of the annual TB cases counts are within the range of **(24.48, 69.26)** and 75% of the average annual TB case counts are less than **69.26**. Therefore, the relatively low median value combined with the IQR appears to be supportive of the statement that the top 20 countries listed above have unusually high average TB infection rates. In fact, none of the top 20 have an infection rate that falls below either the mean value or the upper bound of the IQR.

Given the relatively very high rates of infection observed for the top 9 countries within the list it may be prudent to see if we can determine how often any particular country may have registered the highest rates of infection per 100K people during the years covered by our data set. For example, we can aggregate the twenty highest TB rates of infection across all nineteen years of our data and list them by ‘country’ and ‘year’:

```
# get tb rates for all countries
tbr_df <- sqlQuery(con, "SELECT * FROM tb_rates", stringsAsFactors=F)

tbr_df$rate <- tbr_df$rate * 100000

# Sort the rates in descending order for ALL years
highrates <- arrange(tbr_df, desc(rate), country, year)
```

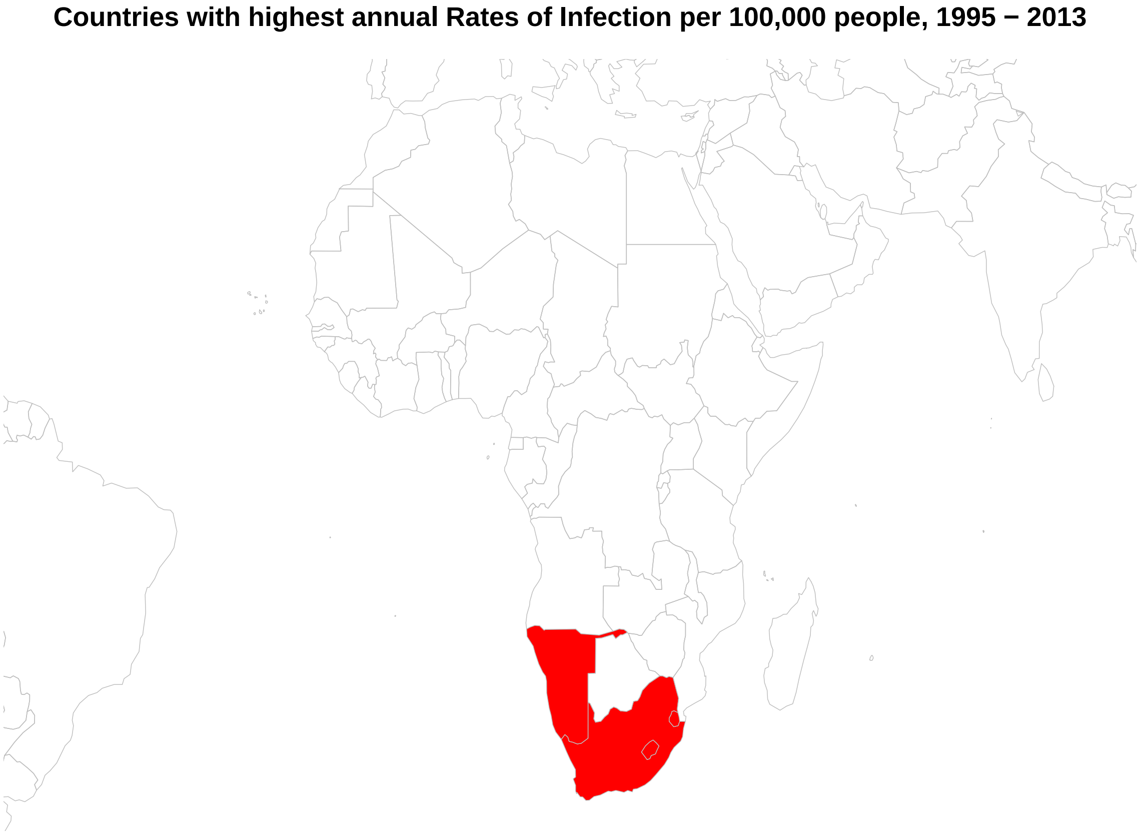
Table 3: Top 20 Country/Year TB Infection Rates per 100K People, 1995 - 2013

	country	year	rate
1	Swaziland	2010	813
2	South Africa	2009	754
3	South Africa	2010	709
4	Swaziland	2006	708
5	South Africa	2008	637
6	South Africa	2013	592
7	Lesotho	2009	580
8	South Africa	2007	579
9	South Africa	2012	567
10	Swaziland	2012	558
11	South Africa	2011	557
12	South Africa	2006	556
13	Lesotho	2010	555
14	Lesotho	2011	539
15	Swaziland	2013	535
16	Lesotho	2012	503
17	Swaziland	2011	465
18	Namibia	2012	438
19	Lesotho	2013	430
20	Namibia	2013	409

When isolating the ‘country’ / ‘year’ pairs in this manner, we can see that a handful of countries appear to have repeatedly experienced the highest observed annual rates of TB infection per 100K people for the years covered by our data set. Consolidating the country names from the top 20 list yields the following distinct country names:

	country
1	Swaziland
2	South Africa
3	Lesotho
4	Namibia

Plotting these countries within a geomap shows that they are all located in southern Africa and are, in fact, adjacent to one another.


Countries with highest annual Rates of Infection per 100,000 people, 1995 − 2013

The fact that these countries are so proximal to one another may be indicative of a particularly acute geographical TB "hotspot" within southern Africa. Such information can be of great use to global and regional health authorities as they attempt to focus their TB prevention and treatment efforts.

Cause For Alarm: High Case Counts + High Infection Rates

While high absolute TB case counts and high rates of infection per 100,000 people can each be of concern in and of their own right, a country having **both** high absolute TB case counts and relatively high rates of infection should be of particular concern to health authorities. In fact, it would be reasonable to state that such countries should be considered “very high risk” areas for TB infections relative to other countries.

We can quickly identify which of the top 20 countries for absolute TB case counts also appear in the top 20 list of countries for TB infection rates using data we derived earlier herein:

```
rateT20 <- head(tbrate_means, n = 20)
casesT20 <-  head(tbm_sorted, n = 20)

high_cr <- data.frame()

for(i in 1:nrow(rateT20)) {
  # find country names that occur in both top 20 case counts and top 20 infection rates
  high_cr <- rbind(high_cr,
                   data.frame(casesT20$country[casesT20$country == rateT20$country[i]]))
}

# rename second column to meaningful name
colnames(high_cr)[1] <- "country"
```

Table 5: Countries With High TB Case Counts & High Infection Rates

	country
1	South Africa
2	Zimbabwe
3	North Korea
4	Kenya
5	Democratic Republic of the Congo

As shown above, five countries appear in both “Top 20” lists, with four of them being located in sub-Saharan Africa. In other words, for each of these countries during the years 1995 - 2013, the TB infection rate was relatively high within their populations ***and*** a relatively large number of people had the disease. Therefore, people living within these five countries likely had a greater chance of contracting TB than did people living within the other 95 countries within our data set.

Trend Analysis: Changes in TB Case Counts & Infection Rates, 1995 - 2013

While calculating averages of TB case counts and infection rates over a span of years can prove useful for purposes of identifying likely TB “hotspots” throughout the world, the averages themselves can hide trend-related information that will also be relevant to health authorities. For example, while we were able to identify the 20 countries having the highest TB infection rates for the years 1995 - 2013, that analysis provides no insight as to whether infection rates have been rising or falling within those countries over the course of that period, nor does it give us any sense as to the rate at which the infection rates might be changing over time.

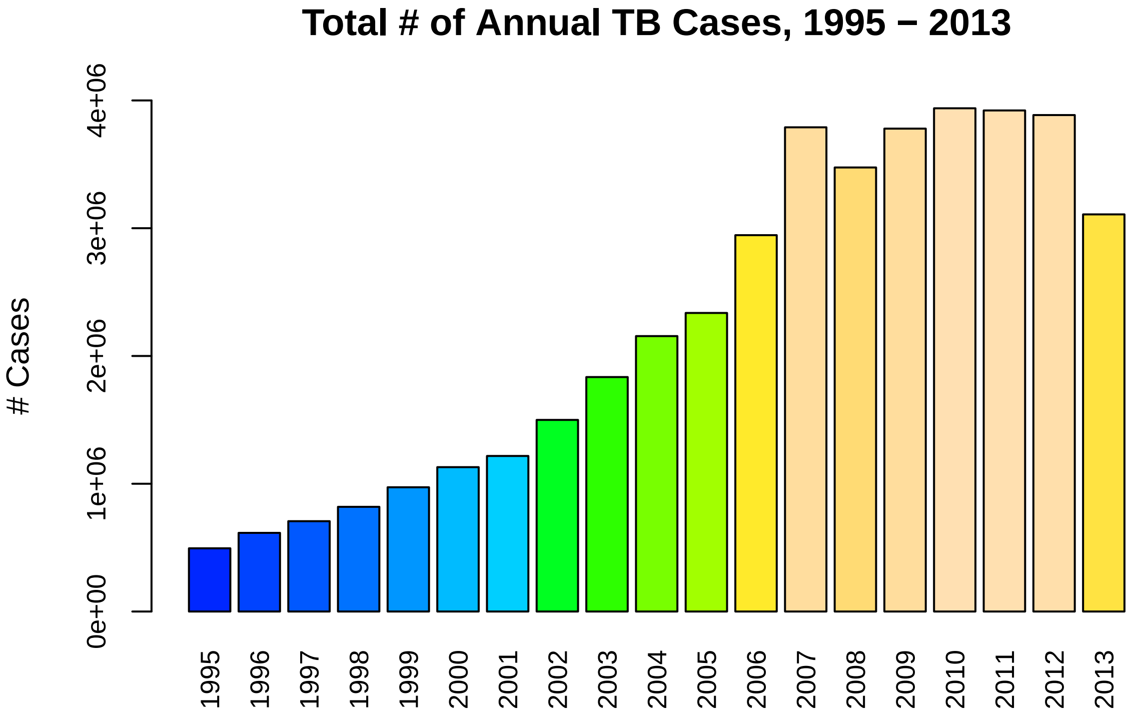
In other words, are some countries experiencing a rapid increase or decrease in the rate of TB infections over time? Similarly, are absolute case counts building up or trailing off within a country over time? The averages we've calculated can't tell us. However, the data we've collected allow us to isolate TB infection metrics for individual calendar years. Therefore, we can easily compare the metrics for any pair or sequence of years for the period 1995 - 2013 to see whether any trends make themselves obvious.

While an exhaustive trend analysis is beyond the scope of this study, we will briefly examine the metrics for global TB case counts for the entire 1995 - 2013 period as well as for a pair of years from our data set to see if any obvious trends can be noted. Specifically, for our pair of years we'll examine metrics for 2001 and 2012. The use of this eleven year timeframe is somewhat arbitrary, but it will provide an opportunity to compare data from the chronologically earlier part of the data set (2001) with data from close to the chronological end of the data set (2012), thereby allowing for an opportunity to ascertain whether any trend may have occurred during that timeframe. We'll also plot the case counts and infection rates for several select countries in an attempt to identify TB infection trends within those selected countries.

Trends in Total TB Cases Worldwide, 1995 - 2013

As a first step in our trend analysis effort, we can sum and plot the aggregate number of annual TB cases reported throughout the 100 countries represented within our data set for the years 1995 - 2013:

```
y_cases <- tb_df %>%
           group_by(year) %>%
           summarise(tot_cases = sum(cases, na.rm = TRUE))
```



The plot clearly shows a rather alarming increase in the number of TB cases recorded during the period from 1995 to 2007. During that time the number of recorded cases appears to have increased from approximately

500,000 annually to nearly 4,000,000 annually. The case counts appear to level off through the years 2008 - 2012 before declining by approximately 500,000 or so in 2013.

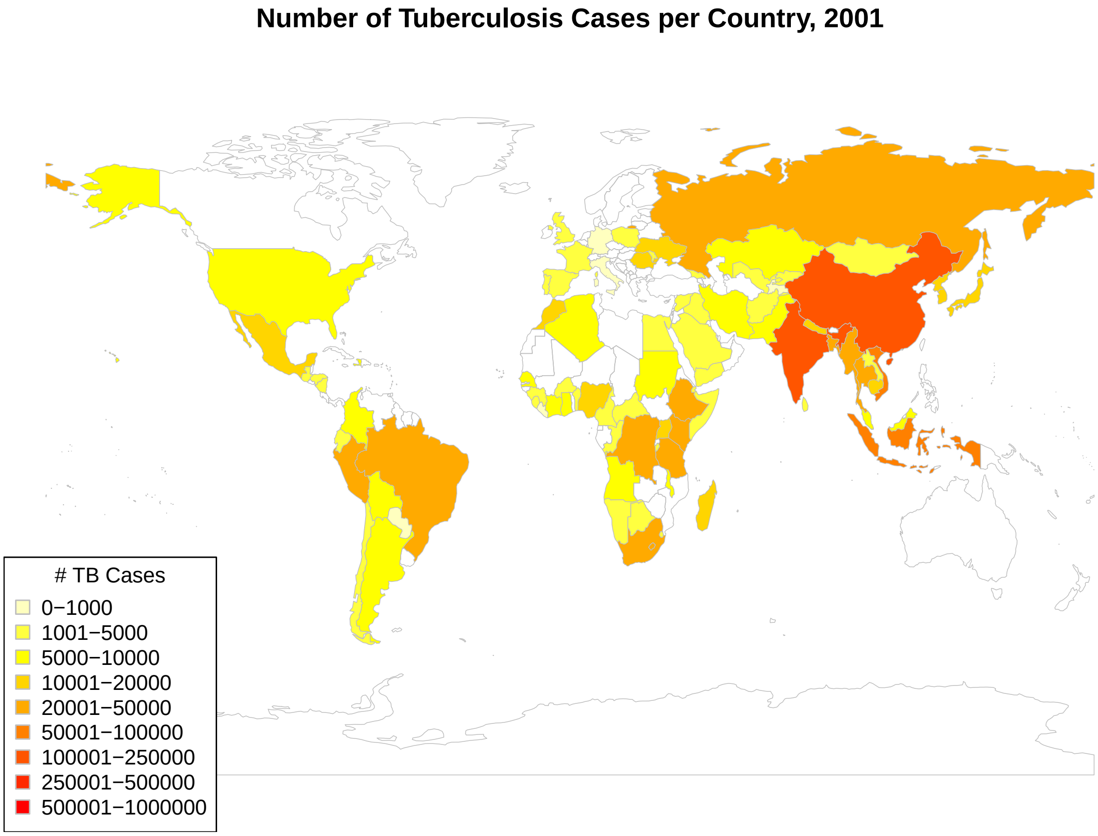
Unfortunately, the data themselves to not provide any insight as to the exact cause of that rapid increase, e.g., was it due solely to a rapid spread of the TB virus or was it in part due to improved TB diagnosis and record keeping techniques? We'll need to keep such questions in mind as we continue with the analysis herein.

Trends in Annual Counts of TB Cases per Country, 2001 vs. 2012

We can extract data from our TB data set for specific individual years for purposes of identifying changes in annual TB case counts across all 100 countries for a given time period. As discussed above, we'll examine the period from 2001 to 2012 to see if any significant changes in the "top 20" countries (as defined by their total annual TB case counts) occurred during that 11 year period.

Number of TB Cases per Country for 2001

```r
# subset cases for a single year (valid years = 1995 - 2013)
tb_2001 <- subset(tb_df, year == 2001, select = c(country, cases))

# Sort the number of cases in descending order
highcases2001 <- arrange(tb_2001, desc(cases), country)
```



According to our map, it appears as though the countries with the highest TB case counts in 2001 included China, India, Indonesia, Russia, Brazil, and several countries in sub-Saharan Africa. However, it also appears as though we have no data for 2001 for several countries in Africa as well as some countries in Asia, including the Phillipines which was identified as having the sixth-highest average TB case count in the table titled **"Top 20 Countries by Avg # of TB Cases, 1995 - 2013"** shown earlier. This lack of data serves as somewhat of an impediment in our attempt to identify a trend for the 2001 - 2012 time period. However, we should still be able to note trends in other countries for which we do have data.

Let's take a look at a list of the 20 countries (for which we have data) with the highest TB case counts for 2001:

Table 6: Top 20 TB Case Count Countries, 2001

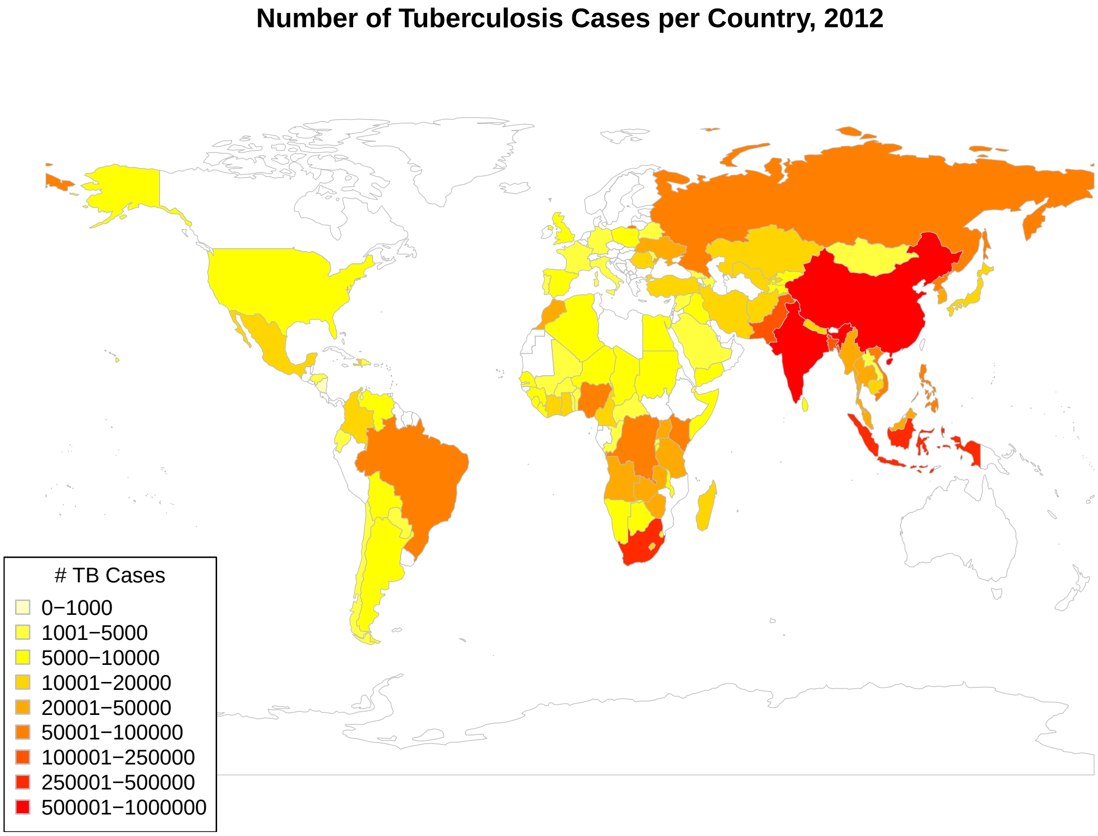
	country	cases
1	China	212766
2	India	185277
3	Vietnam	54202
4	Indonesia	53965
5	Bangladesh	40487
6	Democratic Republic of the Congo	40450
7	Brazil	37491
8	Ethiopia	33028
9	Kenya	31307

	country	cases
10	Thailand	28542
11	Russia	26606
12	Tanzania	24685
13	Peru	21685
14	Myanmar	20686
15	South Africa	20544
16	Nigeria	18097
17	Uganda	17283
18	Mexico	15103
19	North Korea	14428
20	Cambodia	14361

As expected, we see China, India, and Indonesia at or near the top of the list. Case counts for the top 20 span the range **(14361, 212766)**, which seems to indicate that we have a narrower distribution of case counts for 2001 than we found for the 19-year average case count distribution we derived earlier (recall from above that the "top 20" 19-year average case counts span the range **(22013, 441570)**).

Number of TB Cases per Country for 2012

```r
# subset cases for a single year (valid years = 1995 - 2013)
tb_2012 <- subset(tb_df, year == 2012, select = c(country, cases))

# Sort the number of cases in descending order
highcases2012 <- arrange(tb_2012, desc(cases), country)
```



According to our map, it appears as though the countries with the highest TB case counts in 2012 included China, India, Indonesia, South Africa, Russia, Brazil, and several countries in sub-Saharan Africa. Regarding data completeness, note that we appear to have data for the Phillipines for 2012, while we apparently lack data for Peru (which was represented in the 2001 data). However, we are still able to note trends in other countries for which we do have data for both years.

Let's take a look at a list of the 20 countries (for which we have data) with the highest TB case counts for 2012:

Table 7: Top 20 TB Case Count Countries, 2012

	country	cases
1	China	858861
2	India	629589
3	Indonesia	322882
4	South Africa	296996
5	Bangladesh	161790
6	Pakistan	132757
7	Russia	97542
8	Philippines	93983
9	North Korea	85184
10	Kenya	81449
11	Democratic Republic of the Congo	71124

	country	cases
12	Brazil	71072
13	Nigeria	52901
14	Vietnam	51033
15	Myanmar	42909
16	South Korea	39513
17	Zambia	38869
18	Zimbabwe	34391
19	Thailand	30998
20	Tanzania	30063

The case counts for several countries, including China, India, Indonesia, and South Africa appear to have skyrocketed relative to 2001. Case counts for the top 20 span the range **(30063, 858861)**, which is a drastic change from the 2001 "top 20" case count distribution range of **(14361, 212766)**. Furthermore, the 2012 range is much broader than the **(22013, 441570)** range we found for the "top 20" 19-year average case counts. Therefore, there appears to be a clear and rather alarming trend of increasing TB case counts in many of these countries in 2012 when compared to their TB case counts for 2001.

The 2001 and 2012 top 20 lists also appear to include many of the same countries:

```r
T20_2001 <- head(highcases2001, n = 20)
T20_2012 <-  head(highcases2012, n = 20)

high_cr <- data.frame()

for(i in 1:nrow(T20_2012)) {
  # find country names that occur in both top 20 case counts and top 20 infection rates
  high_cr <- rbind(high_cr,
                   data.frame(T20_2001$country[T20_2001$country == T20_2012$country[i]]))
}

# rename second column to meaningful name
colnames(high_cr)[1] <- "country"
```

Table 8: Countries Appearing in Top 20 for both 2001 and 2012

	country
1	China
2	India
3	Indonesia
4	South Africa
5	Bangladesh
6	Russia
7	North Korea
8	Kenya
9	Democratic Republic of the Congo
10	Brazil
11	Nigeria
12	Vietnam
13	Myanmar
14	Thailand
15	Tanzania

As shown above, fifteen countries appear in the "Top 20" lists for both 2001 and 2012. In fact, as can be seen in the top 20 lists shown above, for most of these countries the total number of TB cases in 2012 was larger, in some instances dramatically larger, than it was in 2001.

The five countries appearing in the 2001 top 20 but not the 2012 top 20 are as follows:

Table 9: Countries in 2001 Top 20 but not 2012 Top 20

Country
Ethiopia
Peru
Uganda
Mexico
Cambodia

Referring to the 2012 TB case count map shown above, it appears as though there are no data for either Ethiopia or Peru for 2012, a fact which would fully explain their disappearance from the top 20 list for 2012. For Cambodia, Mexico, and Uganda it appears as though their annual TB case counts for 2001 and 2012 are fairly similar as indicated by the color codings as indicated on the maps shown above.

The five countries appearing in the top 20 for 2012 but not in the top 20 for 2001 are as follows:

Table 10: Countries in 2012 Top 20 but not 2001 Top 20

Country
Pakistan
Philippines
South Korea
Zambia
Zimbabwe

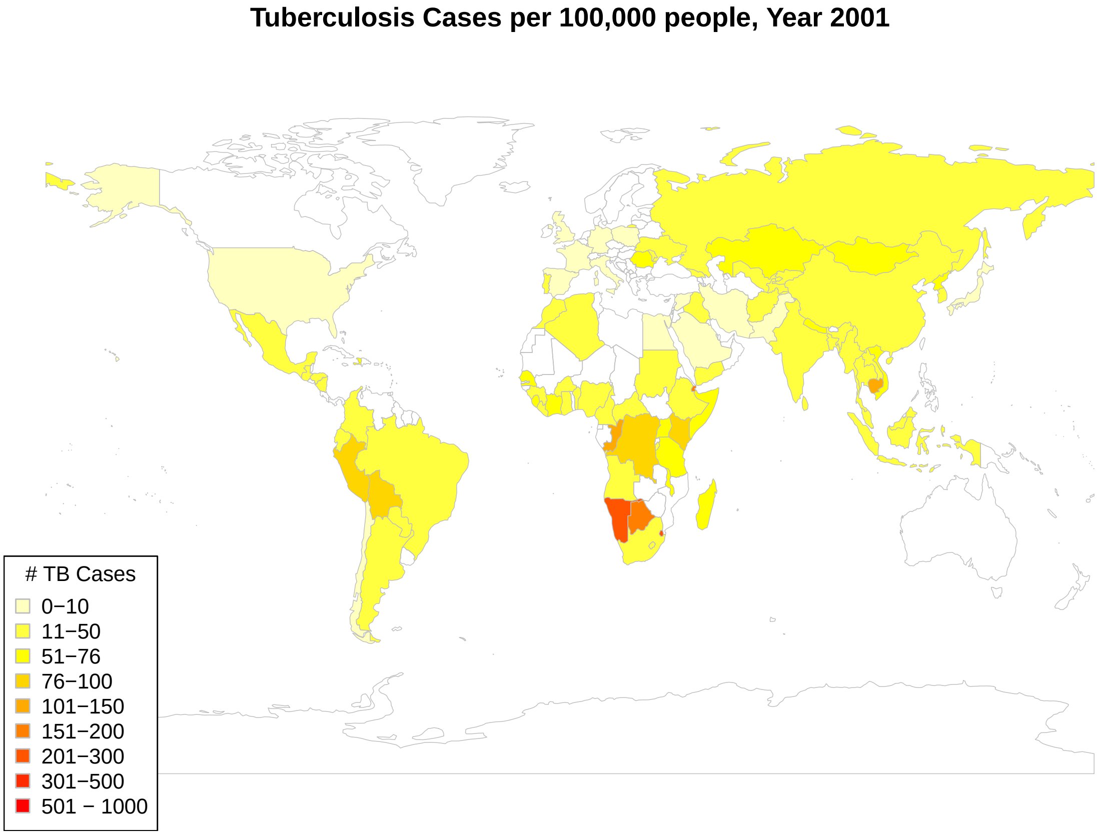
Referring to the 2001 TB case count map shown above, it appears there are no data for the countries of the Phillipines, Zambia, and Zimbabwe for 2001. As such, we would need to examine data for other years for these countries if we wish to draw any conclusions regarding trends in their TB case counts. For both Pakistan and South Korea there appears to have been a noticable increase in their annual TB case counts in 2012 as compared to 2001 as indicated by the color codings as indicated on the maps shown above.

Trends in Annual TB Infection Rates per Country, 2001 vs. 2012

We can also extract data from our TB data set for specific individual years for purposes of identifying changes in annual TB rates of infection across all 100 countries for a given time period. As with the case count data, we'll examine data for the years 2001 and 2012 to see if any significant changes in the "top 20" countries (as defined by their total annual TB rates of infection per 100K people) occurred during that 11 year period.

TB Infection Rates per Country for 2001

```
# subset cases for a single year (valid years = 1995 - 2013)
tbr_2001 <- subset(tbr_df, year == 2001, select = c(country, rate))

# Sort the number of cases in descending order
highrates2001 <- arrange(tbr_2001, desc(rate), country)
```



According to the map, it appears as though the countries with the highest TB rates of infection in 2001 were several countries in sub-Saharan Africa, Cambodia, Peru, and Bolivia. However, as discussed earlier, we have no data for 2001 for several countries in Africa as well as some countries in Asia, including the Phillipines which was identified as having the sixth-highest average TB case count in the table titled **"Top 20 Countries by Avg # of TB Cases, 1995 - 2013"** shown earlier. However, we should still be able to note trends in TB infection rates in other countries for which we do have data.

Let's take a look at a list of the 20 countries (for which we have data) with the highest TB infection rates for 2001:

Table 11: Top 20 TB Infection Rates per 100K People, 2001

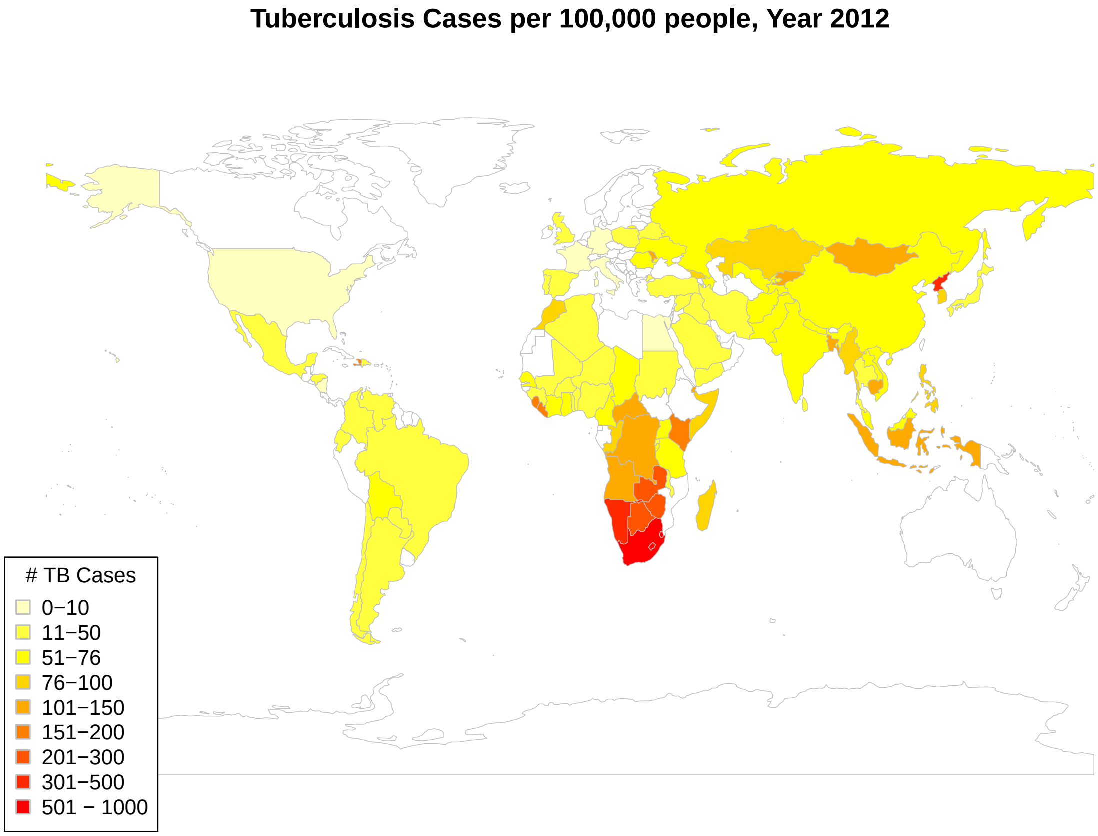
	country	rate
1	Swaziland	237
2	Namibia	235
3	Djibouti	179

	country	rate
4	Botswana	172
5	Republic of Congo	130
6	Cambodia	115
7	Kenya	97
8	Democratic Republic of the Congo	84
9	Peru	82
10	Bolivia	77
11	Malawi	71
12	Tanzania	71
13	Uganda	69
14	Madagascar	68
15	Mongolia	67
16	Vietnam	66
17	Haiti	64
18	North Korea	63
19	Sierra Leone	63
20	Kazakhstan	62

Many of the countries in the list are, in fact, located in Africa. TB infection rates per 100K people for the top 20 for 2001 span the range **(62, 237)**.

TB Infection Rates per Country for 2012

```r
# subset cases for a single year (valid years = 1995 - 2013)
tbr_2012 <- subset(tbr_df, year == 2012, select = c(country, rate))

# Sort the number of cases in descending order
highrates2012 <- arrange(tbr_2012, desc(rate), country)
```



The map indicates that the countries with the highest rates of TB infection per 100K people in 2012 have changed relative to 2001, with most countries on the list having recorded drastically higher rates of infection than were recorded in 2001. While countries in much of sub-Saharan Africa remain at high rates of infection, countries in both Central Asia and the Far East are also appear to have higher rates of infection than in 2001. In particular, we see the countries of Mongolia, North Korea, South Africa, and Indonesia, amongst others, showing a clear and alarming increase in TB rates of infection relative to 2001.

Let's take a look at a list of the 20 countries (for which we have data) with the highest TB rates of infection for 2012:

Table 12: Top 20 TB Infection Rates per 100K People, 2012

	country	rate
1	South Africa	567
2	Swaziland	558
3	Lesotho	503
4	Namibia	438
5	North Korea	344
6	Botswana	289
7	Zambia	276
8	Zimbabwe	251
9	Liberia	192
10	Kenya	189

	country	rate
11	Sierra Leone	162
12	Haiti	156
13	Mongolia	141
14	Djibouti	136
15	Indonesia	131
16	Democratic Republic of the Congo	108
17	Moldova	108
18	Kyrgyzstan	107
19	Bangladesh	105
20	Central African Republic	103

South Africa tops the list for 2012, despite not even appearing in the top 20 for 2001. Furthermore, the rates of infection for several countries appear to have skyrocketed relative to 2001.

Rates of infection per 100K people for the top 20 countries span the range **(103, 567)**, which is a drastic change from the 2001 "top 20" case count distribution range of **(62, 237)**. Therefore, there appears to be a clear and rather alarming trend of increasing annual TB infection rates for many countries in 2012 when compared to the 2001 data. This trend appears to be reflective of the trend in annual case counts we documented earlier despite the fact that many of the countries with the largest overall case counts (e.g, China, India, etc..) do not, in fact, have relatively high rates of infection per 100K people due to their relatively large populations.

As with the annual case count data, the 2001 and 2012 top 20 lists derived here for TB infection rates also appear to include many of the same countries:

```
T20_2001r <- head(highrates2001, n = 20)
T20_2012r <- head(highrates2012, n = 20)

high_ir <- data.frame()

for(i in 1:nrow(T20_2012r)) {
  # find country names that occur in both top 20 case counts and top 20 infection rates
  high_ir <- rbind(high_ir,
                   data.frame(T20_2001r$country[T20_2001r$country == T20_2012r$country[i]]))
}

# rename second column to meaningful name
colnames(high_ir)[1] <- "country"
```

Table 13: Countries Appearing in Top 20 for both 2001 and 2012

	country
1	Swaziland
2	Namibia
3	North Korea
4	Botswana
5	Kenya
6	Sierra Leone
7	Haiti
8	Mongolia
9	Djibouti
10	Democratic Republic of the Congo

As shown above, only ten countries appear in the "Top 20" lists for both 2001 and 2012. Most of these countries experienced a large increase in their annual TB infection rate for 2012 relative to 2001 as evidenced in the geoplots and tables shown above.

The ten countries appearing in the 2001 top 20 but not the 2012 top 20 are as follows:

Table 14: Countries in 2001 Top 20 but not 2012 Top 20

	Country
1	Republic of Congo
2	Cambodia
3	Peru
4	Bolivia
5	Malawi
6	Tanzania
7	Uganda
8	Madagascar
9	Vietnam
10	Kazakhstan

Of these ten, we know that we lack data for Peru for 2012. For the other nine, a comparison of their color codes across the two geoplots shown above for 2001 and 2012 shows that they recorded either similar or rather modest increases in TB rates of infection in 2012 relative to 2001.

The ten countries appearing in the top 20 for 2012 but not in the top 20 for 2001 are as follows:

Table 15: Countries in 2012 Top 20 but not 2001 Top 20

	Country
1	South Africa
2	Lesotho
3	Zambia
4	Zimbabwe
5	Liberia
6	Indonesia
7	Moldova
8	Kyrgyzstan
9	Bangladesh
10	Central African Republic

Examining the 2001 map shown above shows we lack data for that year for the countries of Lesotho, Zambia, and Zimbabwe. For the other seven countries there appears to have been a noticeable increase in their annual TB rates of infection in 2012 relative to 2001 as indicated by the color codings on the maps shown above.

Trends in Annual TB Case Counts & Infection Rates for Select Countries, 1995 - 2013

The trend analysis we've performed thus far has primarily been focused on examining the differences in reported TB case counts and TB infection rates across all 100 countries represented in our data set for two specific years: 2001 and 2012. The resulting geoplots and "top 20" lists give us some limited insight into how those metrics may have changed for various countries between 2001 and 2012. Furthermore, as we discovered at the start of our trend analysis, there was a clear and significant upward trend in global TB case counts for the period 1995 - 2007, after which case counts seemed to level off for several years before declining a bit in 2013.

However, the year-by-year trends in TB case counts and infections rates could potentially ebb and flow in an unpredictable manner within individual countries due to factors such as an occurrence of a new TB outbreak or improved TB treatment protocols being administered within a country. Therefore, we will now plot the year-by-year case count and infection rate metrics for the entire 1995 - 2013 period for several countries in an attempt to visually present a sample of these country-specific trends.

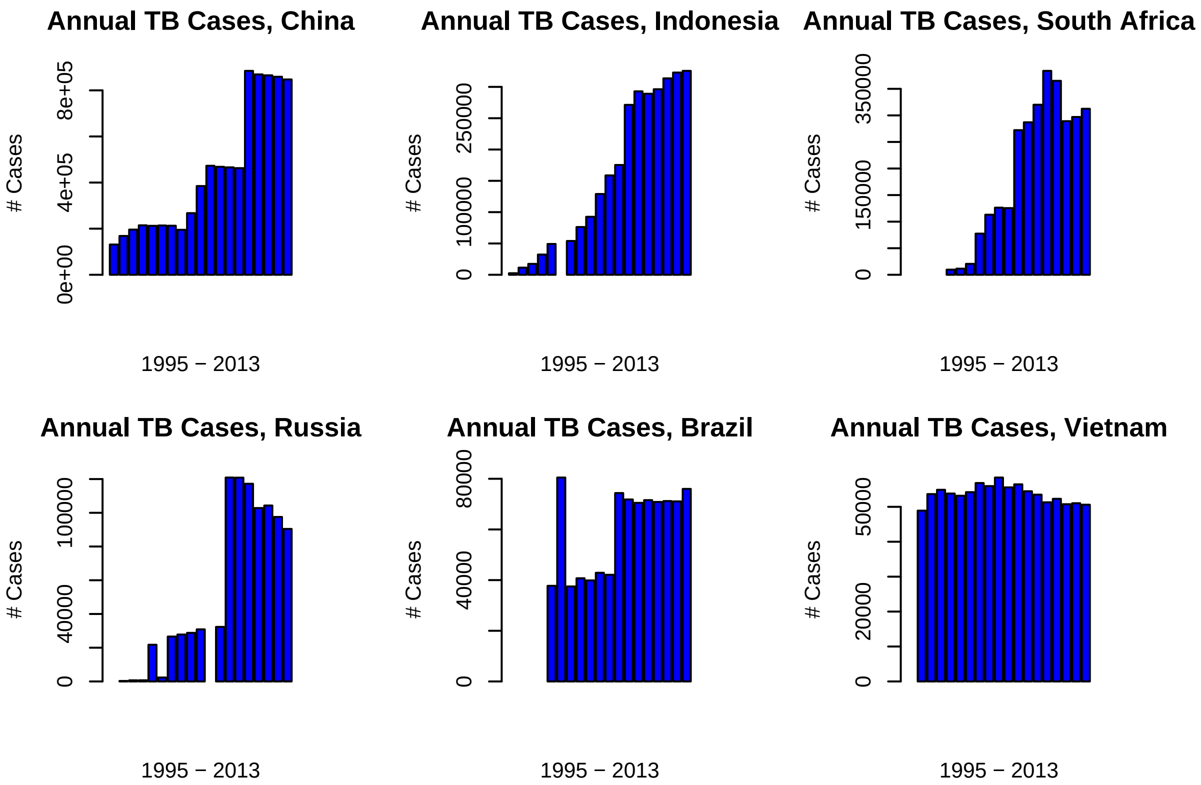
Since we will not be performing a comprehensive, exhaustive trend analysis in this study, plotting metrics for all 100 countries from our data set simply won't be feasible. Instead, we can select a handful of countries from our data having varying levels of TB case counts and infection rates and compare and contrast the resulting plots.

TB Case Count Trends: 18 Selected Countries

PLEASE NOTE: In some instances, case counts for a specific country/year are not available. Missing data are indicated by the lack of a vertical bar for a particular year in the plots shown below.

Six Countries with High Case Counts, 1995 - 2013

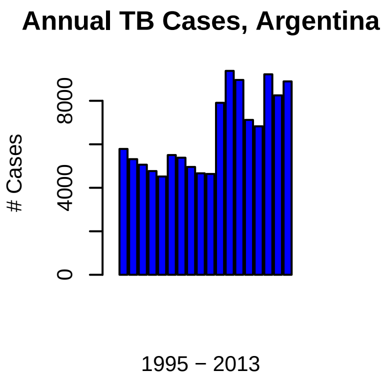
In our earlier analysis on average TB case counts we identified 20 countries as having the highest average case counts throughout the 1995 - 2013 time period. From that list we'll examine China, Indonesia, South Africa, Russia, Brazil, and Vietnam as examples of countries having relatively high overall TB case counts.

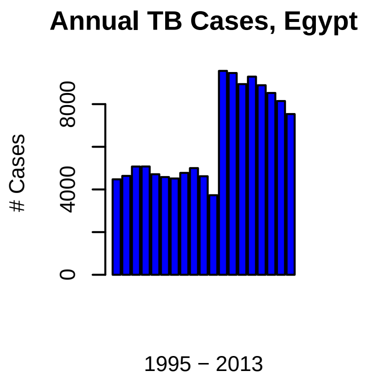


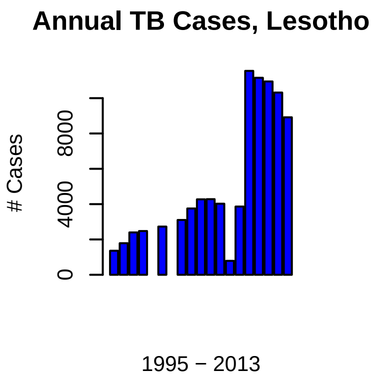
The plots show that five of these countries appear to have experienced a significant increase in their TB case counts during the 2006 - 2008 time period. Of the six, only Vietnam maintained a relatively stable (though also relatively high) case count during the entire 1995 - 2013 period.

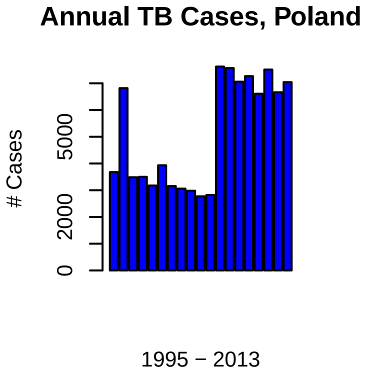
Six Countries with "Near Median" Case Counts, 1995 - 2013

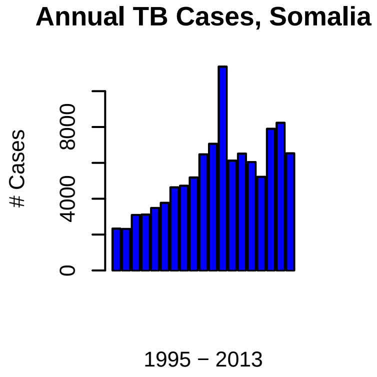
To find "average" TB case count countries, recall from earlier that we determined the median number of average annual TB cases across all 100 countries to be **6681**. We can use that median value as the basis of selecting a few countries with "middling" average TB case counts. Specifically, we'll look at Argentina, Egypt, Lesotho, Poland, Somalia, and the USA.

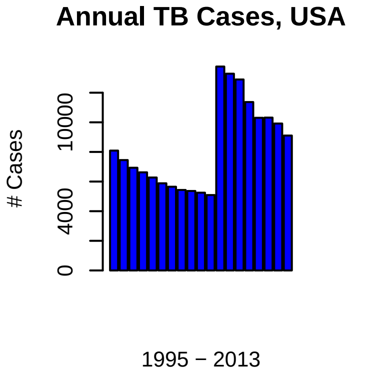








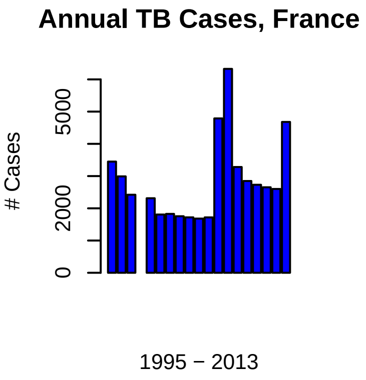


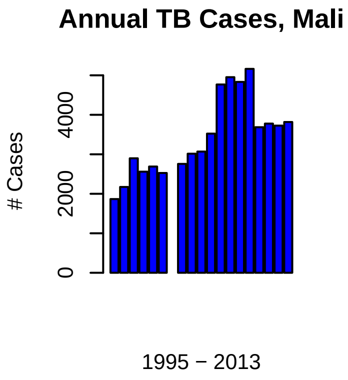


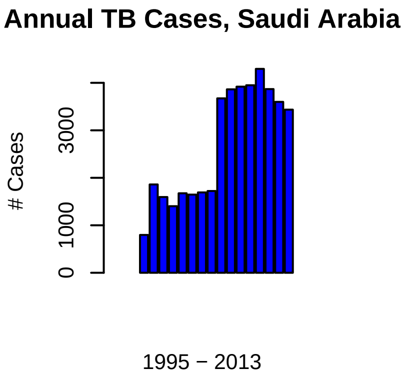
The plots show that each of these six countries appears to have experienced a significant increase in their TB case counts during the 2006 - 2008 time period. Of the six, the USA, Egypt, and Lesotho appear to have subsequently experienced gradual declines in TB case counts, while both Poland and Argentina experienced have experienced relatively recent upticks in case counts after modest declines from the 2006 - 2008 spike in case counts. Somalia's data appear rather unusual in that, while the country appears to have experienced a brief spike in TB case counts in the 2006 - 2008 time period, the spike was short-lived and case counts rapidly dropped off from their 1-year peak before rising again in 2011 - 2012.

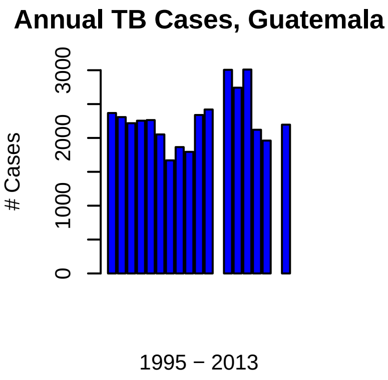
Six Countries with Low Case Counts, 1995 - 2013

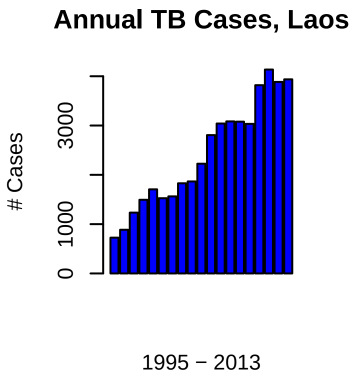
To find countries with relatively low average TB case counts, recall from earlier that we found the IQR of the average annual TB case counts to be **(3381, 16686)**. Therefore, roughly 25% of the countries in our data set will have average annual case counts of 3381 or less. From that group we'll look at France, Mali, Saudi Arabia, Guatemala, Laos, and Rwanda.

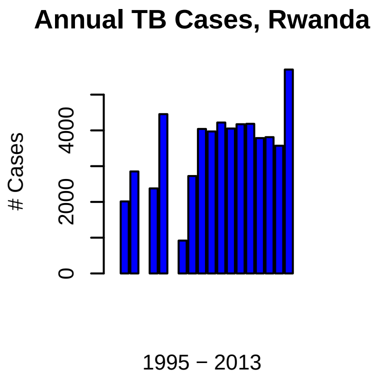












Each of these six countries appears to have experienced a surge in TB case counts during the 2006 - 2008 period. Of the six, only Laos appears to have experienced a rather continuous and steady upward trend in TB case counts throughout the entire 1995 - 2013 period. France experienced a rather sharp upturn in case counts during the 2006 - 2008 period followed by a rapid decrease, only to experience yet another sharp uptick in 2013.

In fact, the plot for France's case counts appears somewhat similar to that of Somalia's. Could it be that France's historical colonial relationship with Somalia has resulted in similar TB case count dynamics within both countries? Possibly, but such questions are beyond the scope of this study.

Country-by-Country TB Case Count Trend Comments

While we haven't examined case counts for each of the 100 countries in our data set, there appears to be one general comment we can make regarding the data for the 18 countries we've examined: in each of our three categories of countries explored above, we see clear evidence of a significant increase in annual TB case counts during the 2006 - 2008 period. This increase reflects the surge in global TB case counts we identified earlier in the section titled **"Trends in Total TB Cases Worldwide, 1995 - 2013"**.

In most instances annual case counts then leveled off or declined a bit, though in some countries such as Laos, Indonesia, Argentina, and Brazil the increase has either not abated or has continued

Though beyond the scope of this study, it might be worthwhile to investigate the reasons some countrys' TB case counts have declined a bit in recent years while others have not. However, we will explore aspects of this question later when we delve into the relationship between incidence of TB within countries and the various "per capita" metrics we described earlier.

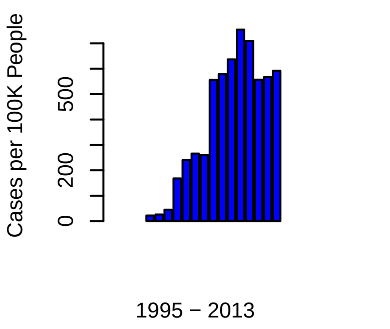
TB Infection Rate Trends: 18 Selected Countries

PLEASE NOTE: In some instances, TB infection rates for a specific country/year are not available. Missing data are indicated by the lack of a vertical bar for a particular year in the plots shown below.

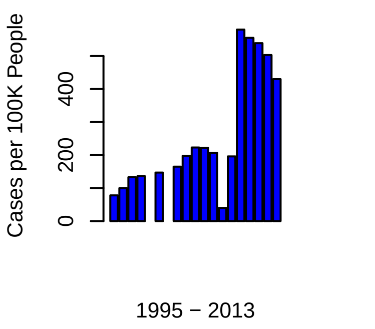
Six Countries with High TB Infection Rates, 1995 - 2013

During our earlier analysis on average TB infection rates per 100K people we identified a list of 20 countries having the highest average TB infection rates for the 1995 - 2013 time period. From that list we'll examine South Africa, Lesotho, Botswana, North Korea, Haiti, and Kazakhstan as examples of countries having relatively high overall TB infection rates.

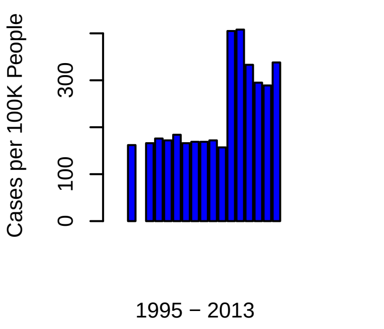
TB Infection Rates, South Afric



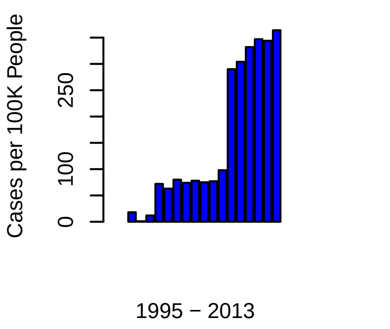
TB Infection Rates, Lesotho



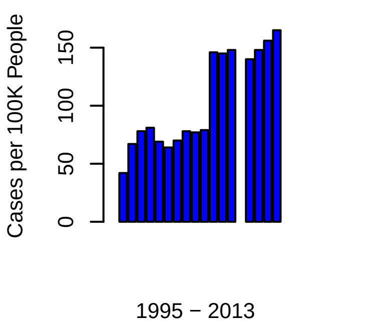
TB Infection Rates, Botswana



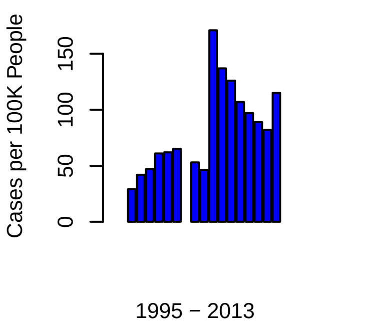
TB Infection Rates, North Kore



TB Infection Rates, Haiti



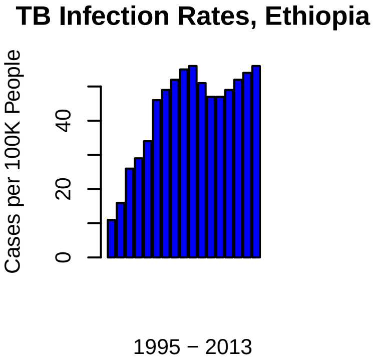
TB Infection Rates, Kazakhstan

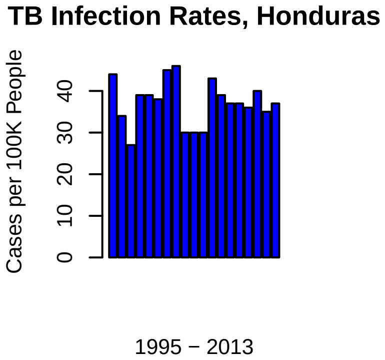


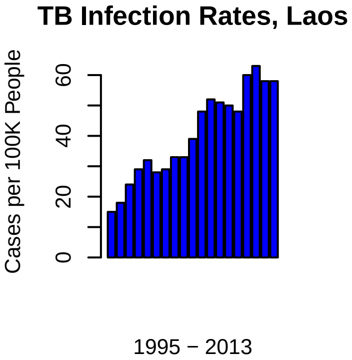
The plots show that all six of these countries appear to have experienced a significant increase in their TB infection rates during the 2006 - 2008 time period. Of the six, both North Korea and Haiti appear to have experienced an ongoing increase in TB infection rates, while the other four countries saw their infection rates decline, though that decline appears to have ended for South Africa, Botswana, and Kazahkstan in 2013.

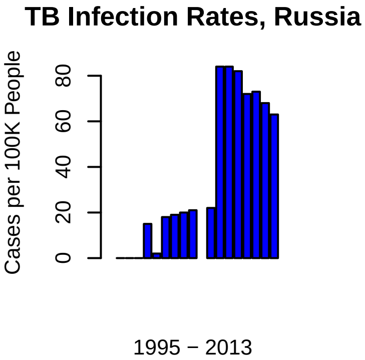
Six Countries with "Near Median" TB Infection Rates, 1995 - 2013

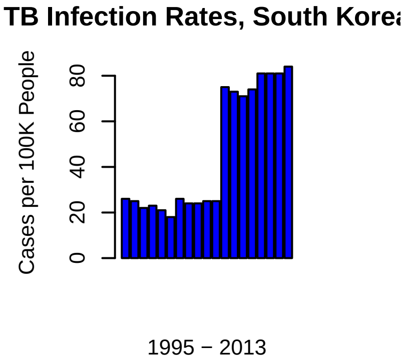
To find "average" TB infection rate countries, recall from earlier that we determined the median number of average annual TB cases across all 100 countries to be **41.68**. We can use that median value as the basis of selecting a few countries with "middling" infection rates. Specifically, we'll look at Ethiopia, Honduras, Laos, Russia, South Korea, and Thailand.

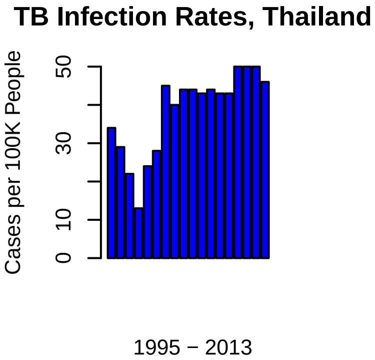












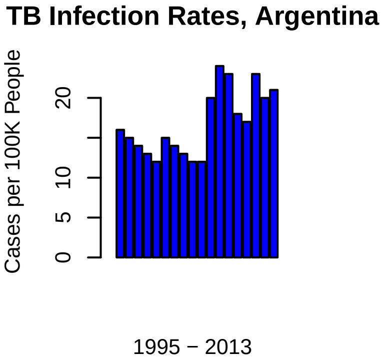
Of these six countries, Russia and South Korea show the most obvious evidence of a sudden and significant upturn in TB infection rates during the 2006 - 2008 time period. Both countries saw their TB infection rates jump from approximately 20 cases per 100K people to approximately 80 cases per 100K people during the 1995 - 2013 time period, a fourfold increase.

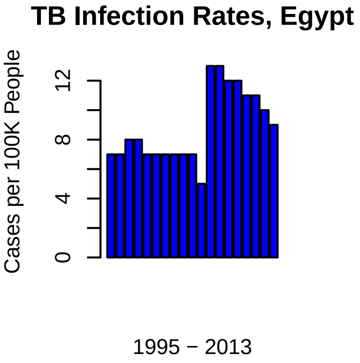
Ethiopia, Laos, and Thailand each show evidence of a steady increase in TB infection rates dating back to the late 1990's, and both Ethiopia and Laos saw a nearly fivefold increase in TB infection rates during the 1995 - 2013 time period.

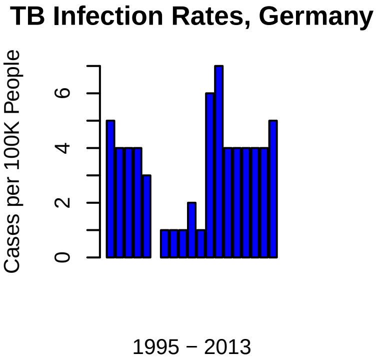
Honduras appears to have had the most stable rate of TB infection rates during the 10995 - 2013 time period relative to the other five countries in this group, though it too shows evidence of a sudden upturn in TB infection rates during the 2006 - 2008 period.

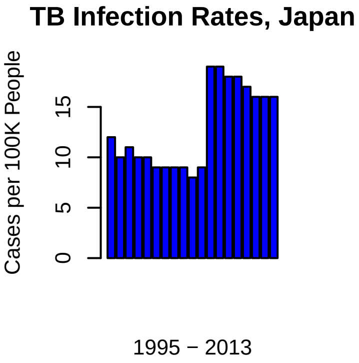
Six Countries with Low Average TB Infection Rates, 1995 - 2013

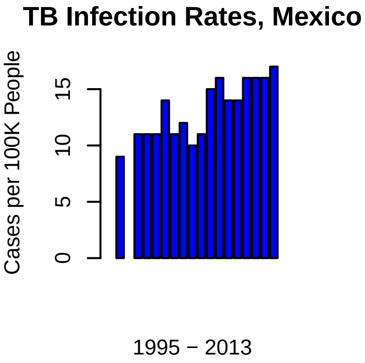
To find countries with relatively low average TB infection rates, recall from earlier that we found the IQR of the average annual TB infection rates to be **(24.48, 69.26)**. Therefore, roughly 25% of the countries in our data set will have average annual case counts of 24.48 or less. From that group we'll look at Argentina, Egypt, Germany, Japan, Mexico, and Nigeria.

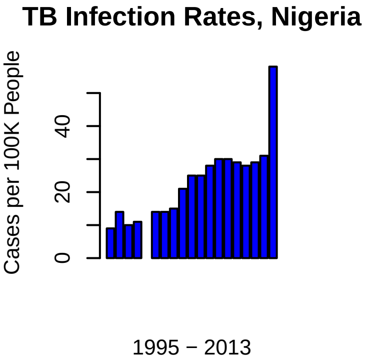












Despite their relatively low TB infection rates, each of these six countries shows evidence of having experienced an upturn in infection rates during the 2006 - 2008 time period. Germany in particular appears to have experienced a sevenfold increase in its TB infection rates during the 2005 - 2007 time period. Infection rates subsequently declined a bit in Egypyt, Germany, Argentina, and Japan. However, after a brief period of decline both Argentina and Germany saw their infection rates creep up again.

Both Mexico and Nigeria appear to have experienced a steadier rate of increase in their TB infection rates, with both achieving their highest rates of infection in 2013 for the 1995 - 2013 time period, and Nigeria experiencing a near doubling of its infection rate in 2013 alone.

Country-by-Country TB Infection Rate Trend Comments

As with the country-by-country TB case count data examined above, there appears to be one general comment we can make regarding the TB infection rates for the 18 countries we've examined here: in each of the three categories of countries, we see clear evidence of a significant increase in annual TB infection rates during the 2006 - 2008 period. This increase reflects the surge in global TB case counts we identified earlier in the section titled **"Trends in Total TB Cases Worldwide, 1995 - 2013"**.

In most instances the infection rates then leveled off or declined a bit, though in some countries such as North Korea, Haiti, Ethiopia, Laos, South Korea, Mexico, and Nigeria the increase has either not abated or has continued.

Per Capita Metrics & Their Relationship to TB Infection Rates

Our analysis thus far has focused solely on TB case counts and infection rates. We now turn our attention to the question of whether TB infection rates might be influenced by how well a country measures up against other countries in terms of a variety of metrics. As mentioned earlier, the metrics to be considered are:

- Life expectancy at birth
- Annual per capita healthcare expenditure per capita
- Per capita gross national income (GNI)
- The percentage of the population having access to electricity in their homes
- Average years of schooling

We'll examine each of these metrics independently of one another to determine how strongly each metric relates to TB infection rates in general across all 100 countries from our data set. We'll also examine how well the countries with the highest average TB infection rates compare to the countries having the lowest average TB infection rates relative to each individual metric. Such comparisons will allow us to then "describe" a relatively high TB infection rate country in terms of the entire group of metrics.

The description we develop can then be compared to a similar description we'll develop for countries having relatively low TB infection rates for purposes of explaining some possible disparities that the "high TB rate" countries may need to resolve if they are to succeed in reducing their own TB infection rates. In other words, by "profiling" both sets of countries we may be able to define a collection of benchmarks that can perhaps be used for measuring some aspects of progress against TB infection rates in "high TB rate" countries.

Finally, we'll attempt to derive a mulitivariate linear model that relates TB infection rates to all of the metrics collectively. The results of that modeling effort will provide some insight into whether the metrics we've chosen to evaluate might actually be predictive of TB infection rates within a given country.

Before we proceed with the analysis we'll create two dataframes:

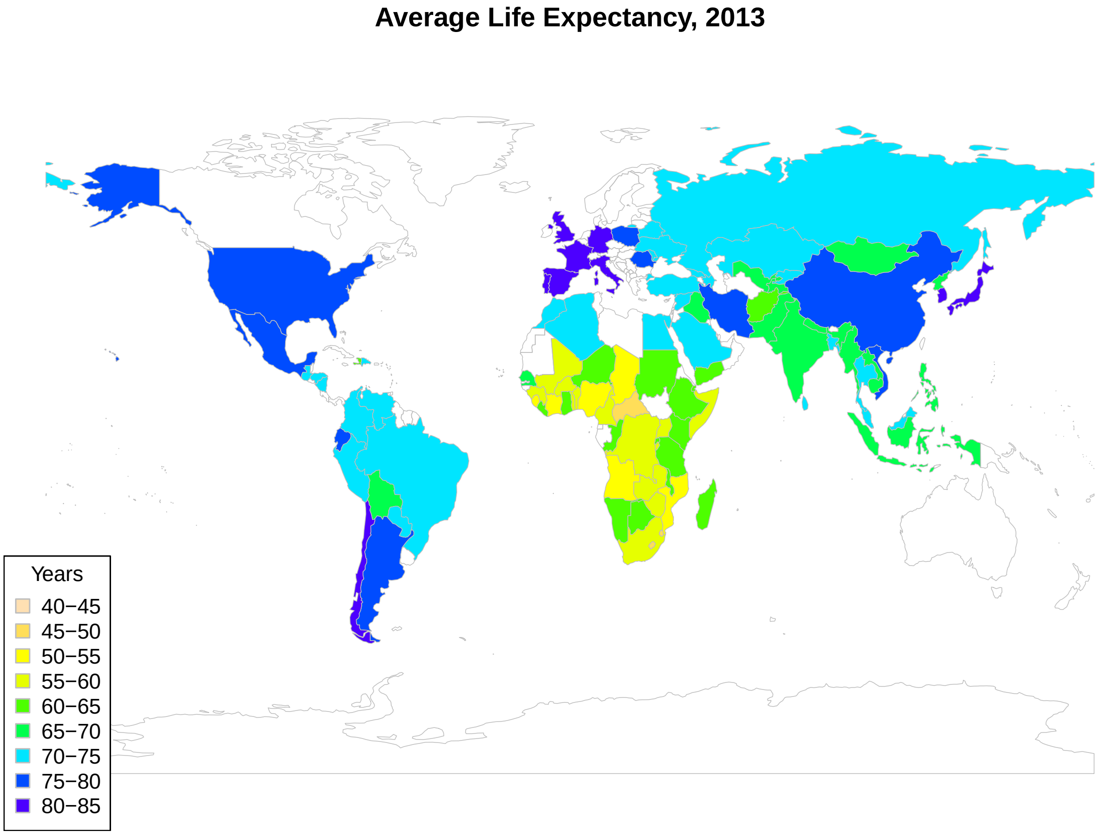
- A dataframe containing the 1995 - 2013 TB infection data for the 20 countries having the ***highest*** average TB infection rates during the 1995 - 2013 timeframe;
- A dataframe containing the 1995 - 2013 TB infection data for the 20 countries having the ***lowest*** average TB infection rates during the 1995 - 2013 timeframe.

```r
# get country names for top 20 tb infection rate countries
countrys <- head(tbrate_means$country, n = 20)

# create data frame with only those 20 countries' data
top20 <- subset(tbr_df, country %in% countrys)


# get country names for bottom 20 tb infection rates
bot_countrys <- tail(tbrate_means$country, n = 20)

# create data frame with only those 20 countries' data
bot20 <- subset(tbr_df, country %in% bot_countrys)
```

Life Expectancy at Birth

The life expectancy data we've collected describes ***the average life expectancy for the populations of 100 countries for the years 1995 - 2013***. We'll start our analysis of the life expectancy data by creating a geoplot of life expectancies for the countries in our data set for the year 2013 (the most recent year available in our data set):

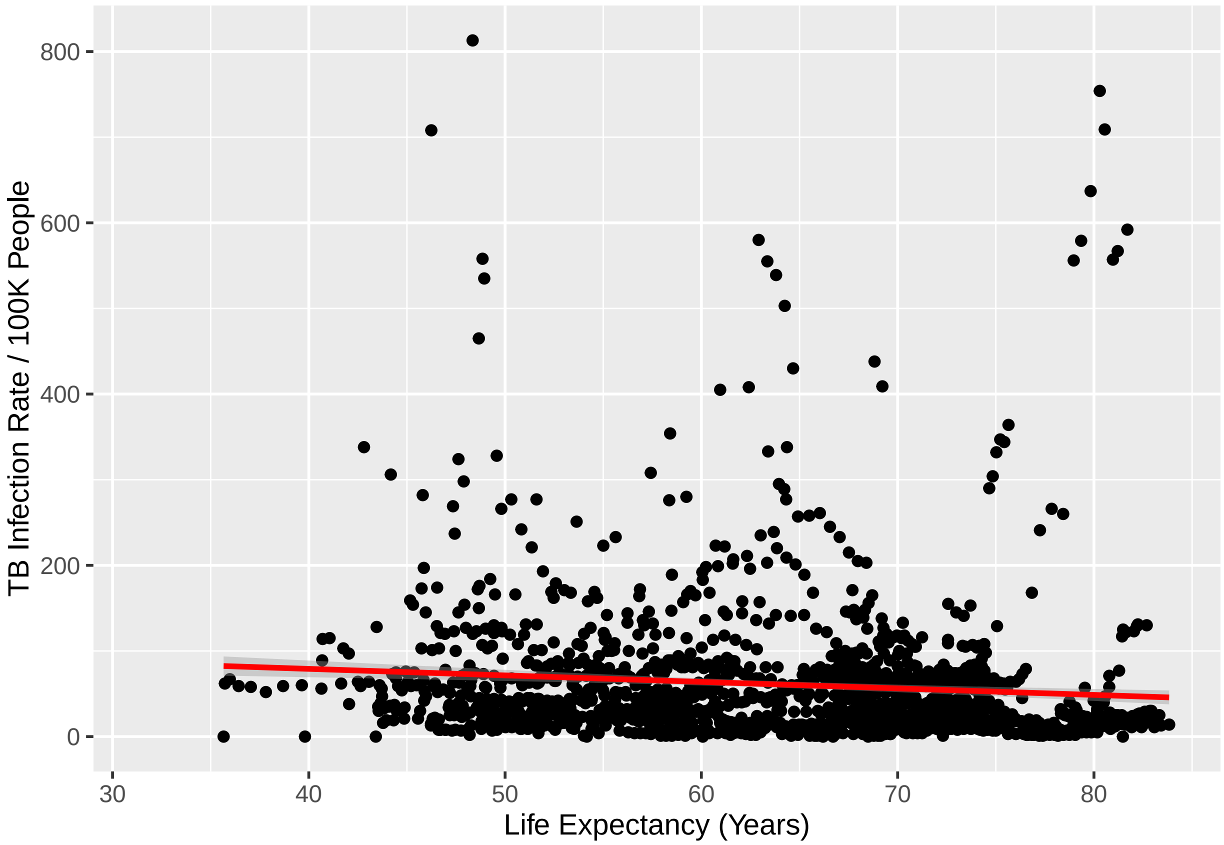
```r
# get life expectancy data for 2013
lifexp_df <- sqlQuery(con, "SELECT * FROM life_exp WHERE year = 2013", stringsAsFactors=F)
```



The geoplot shows that countries in North America and western Europe as well as the countries of Japan, Chile, Argentina, China, Iran, and a few others appear to have the highest life expectancies as of 2013. Many countries in Africa appear to have relatively low life expectancies.

To see whether we can find a relationship between life expectancies and TB infection rates we'll first make use of a scatterplot and add a plot of the results of a linear regression model wherein we attempt to predict TB infection rates on the basis of a country's life expectancy. The scatterplot is created using all TB infection rate data for the years 1995 - 2013 from all 100 countries contained within our data set.

```r
# get life expectancy data set
lifexp_df <- sqlQuery(con, "SELECT * FROM life_exp", stringsAsFactors=F)
```



The scatterplot appears to show a slight negative relationship between TB infection rates and life expectancies. This makes sense from an intuitive standpoint: if a country has high rates of TB infection we would expect the population to have a lower life expectancy than if the country had lower TB infection rates.

We can use R's **cor** function to compute the correlation coefficient of the relationship between TB infection rates and life expectancies:

```
# correlation test
cor(t.df$rate, t.df$life_exp, use="complete")
```

```
## [1] -0.09803323
```

The output of R's **cor** function tells us that the variables have a correlation of -0.098, which is reflective of the negative slope of the regression line shown in the scatterplot.

R's **lm** function provides us with the components of a characteristic equation for the relationship between the two variables:

```
# fit a model & plot for rate ~ life_exp
fit1 <- lm(rate ~ life_exp, data = t.df)

# output of lm function
summary(fit1)
```

```
##
## Call:
```

```
## lm(formula = rate ~ life_exp, data = t.df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -82.39 -40.73 -21.19  12.60 740.25 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 109.4983    12.2353   8.949  < 2e-16 ***
## life_exp     -0.7602     0.1878  -4.048 5.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 79.75 on 1689 degrees of freedom
##   (209 observations deleted due to missingness)
## Multiple R-squared:  0.009611,   Adjusted R-squared:  0.009024 
## F-statistic: 16.39 on 1 and 1689 DF,  p-value: 5.39e-05
```

The characteristic equation is:

TB Infection Rate = 109.4983 - 0.7602 * Life Expectancy

In other words, for each year of life expectancy, the TB infection rate per 100,000 people should decline by an amount of 0.76. However, according to the output of R's **lm** function shown above, this equation explains only 0.009 of the variability we find in TB infection rates as evidenced by the R-Squared value. So while life expectancy is, in fact, a statistically significant predictor of TB infection rates as evidenced by its p-value of approximately zero, a linear least squares model of the relationship between the two variables doesn't actually help us all that much in explaining the variability in TB infection rates that we see throughout our data set.

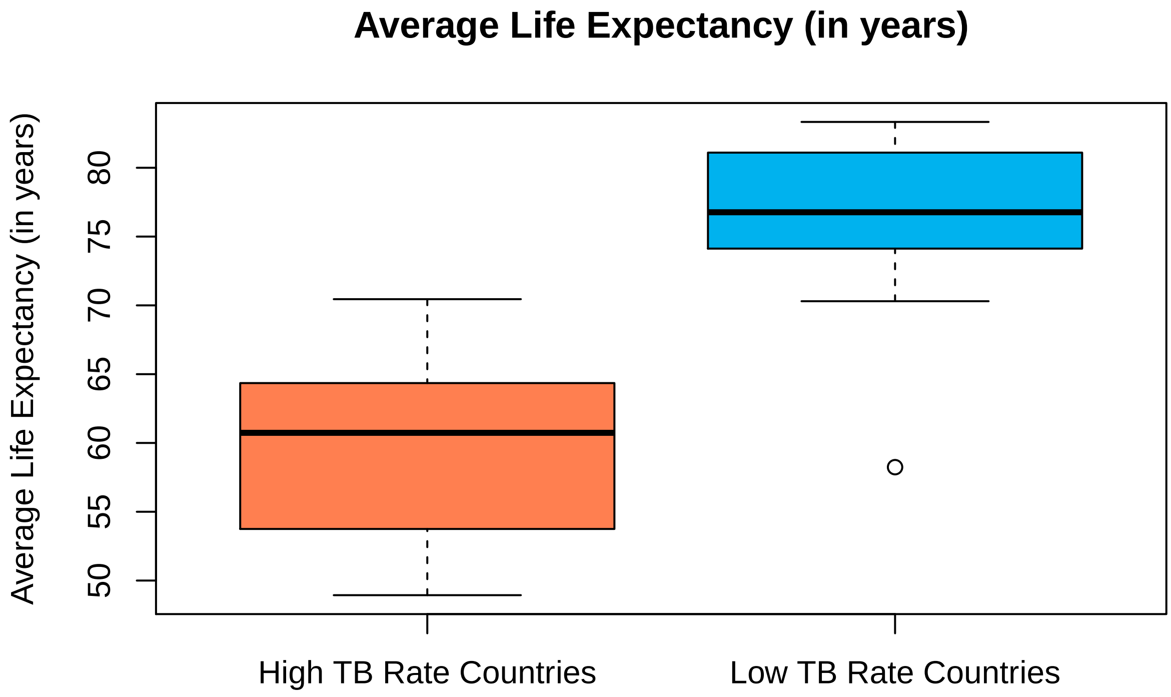
However, we can instead examine differences in life expectancies between the countries having the (on average) highest TB infection rates and those having the (on average) lowest TB infection rates. We'll make use of R's **summary** function as well as side-by-side boxplots of life expectancy data for the year 2013:

```
# get life expectancy data for 20 highest TB infection rate countries
top20.le <- subset(lifexp_df, country %in% countrys & year == 2013)
summary(top20.le$life_exp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   48.94   54.69   60.74   59.66   64.35   70.45 
```

```
# get life expectancy data for 20 lowest TB infection rate countries
bot20.le <- subset(lifexp_df, country %in% bot_countrys & year == 2013)
summary(bot20.le$life_exp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   58.24   74.15   76.77   76.63   81.05   83.33 
```



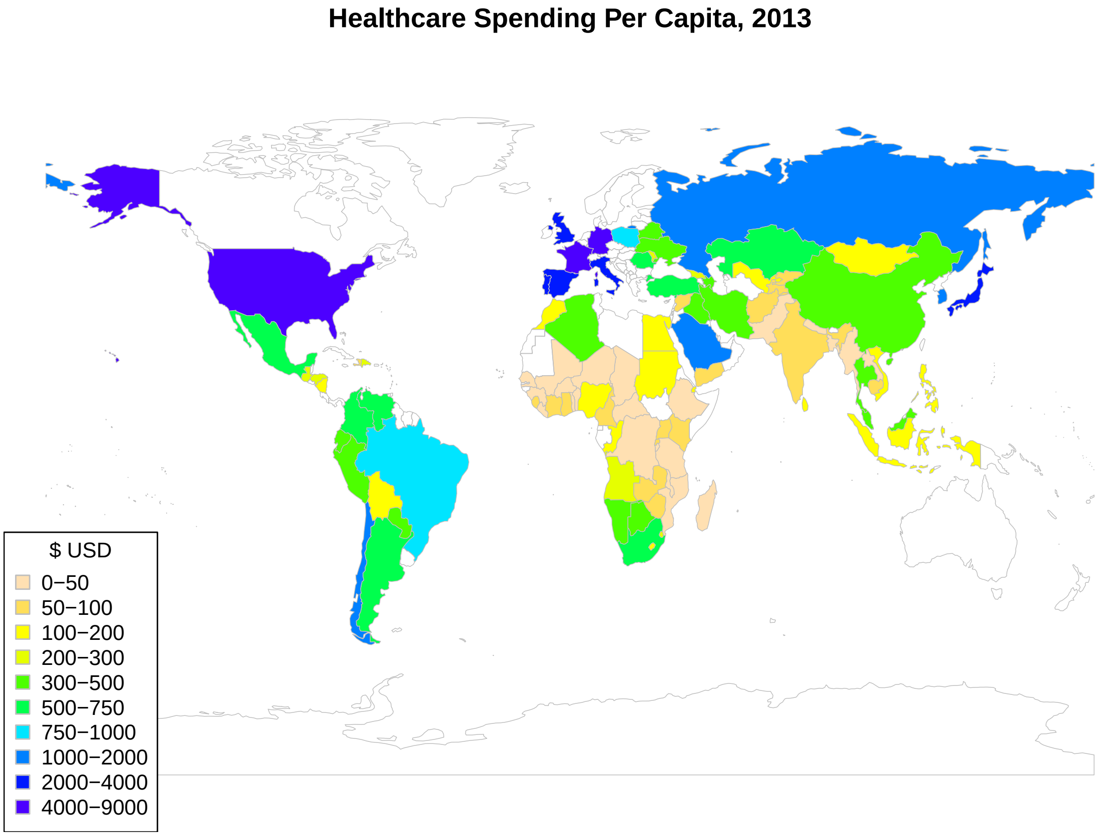
The boxplots shown above provide clear evidence of a striking difference in life expectancies between "high TB rate" countries and those with relatively low TB infection rates. The average life expectancy in a high TB rate country is approximately **60 years** while in low TB rate countries the average is over **76 years**. In fact, the life expectancies of all but one low TB rate country appear to exceed the life expectancies of each high TB rate country.

Clearly we've found our first "measurable" difference between countries with high TB rates and those with low TB rates. We'll now turn our attention to the health care spending metric.

Annual Health Care Expenditure Per Capita

Total annual health expenditure is ***the sum of public and private health expenditures as a ratio of total population of a country***. Data are in current U.S. dollars and cover the years 1995 - 2013. As we did with the life expectancy data, we'll start our analysis by creating a geoplot of annual health care expenditure per capita for the countries in our data set for the year 2013 (the most recent year available in our data set):

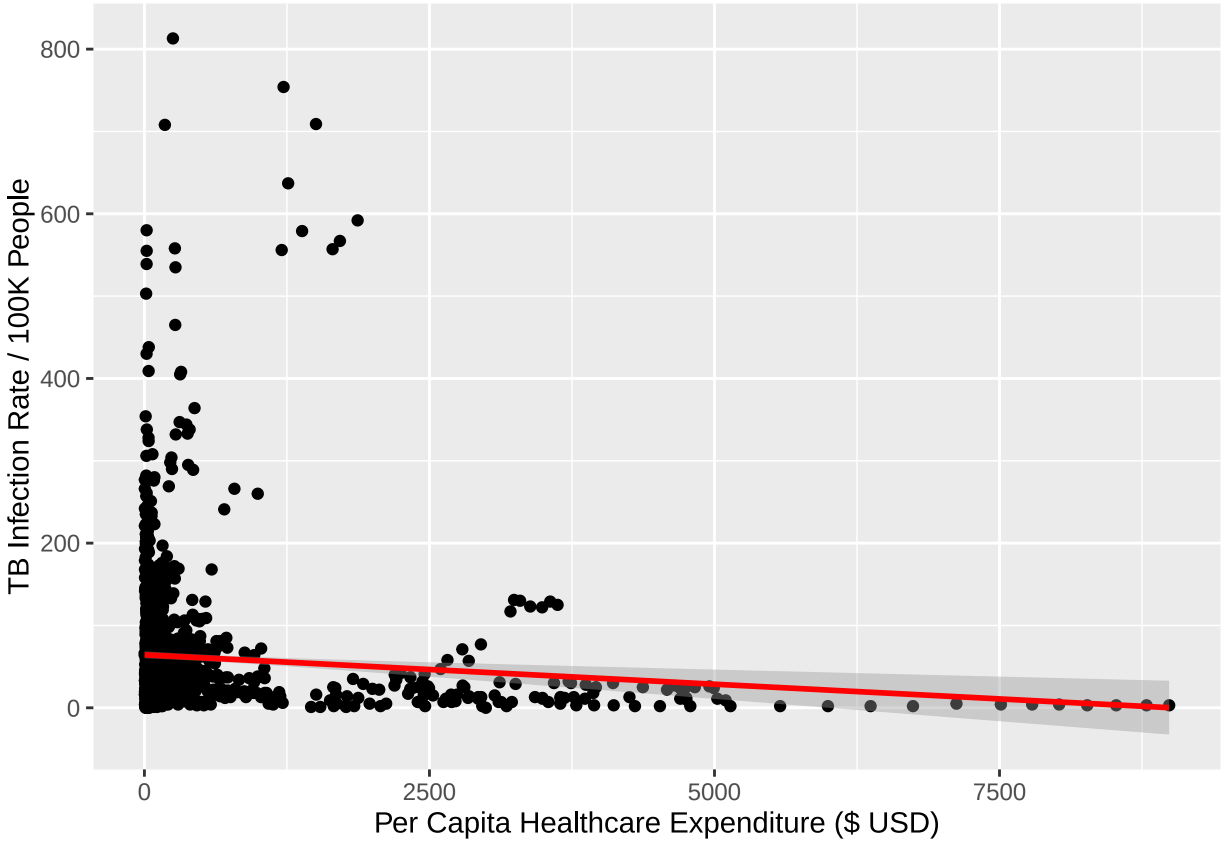
```
# get healthcare expenditure data  for 2013
hc_df <- sqlQuery(con, "SELECT * FROM percap_hc WHERE year = 2013", stringsAsFactors=F)
```



The geoplot shows that countries in western Europe as well as the countries of the USA, Japan, Chile, Russia, South Korea, and Saudi Arabia appear to have the highest healthcare expenditure per capita as of 2013. Many countries in Africa and south-central Asia appear to have relatively low amounts of healthcare expenditure per capita.

To see whether we can find a relationship between per capita health care expenditures and TB infection rates we'll first make use of a scatterplot and add a plot of the results of a linear regression model wherein we attempt to predict TB infection rates on the basis of a country's per capita healthcare expenditure. The scatterplot is created using all TB infection rate data for the years 1995 - 2013 from all 100 countries contained within our data set.

```r
# get healthcare data set
hc_df <- sqlQuery(con, "SELECT * FROM percap_hc", stringsAsFactors=F)
```



The scatterplot shows very clearly that countries that have relatively low per capita healthcare expenditures also tended to experience the highest rates of TB infections during the 1995 - 2013 time period. The regression line echoes this by showing a clear negative relationship between the two variables.

We can again use R's **cor** function to compute the correlation coefficient of the relationship between TB infection rates and per capita healthcare expenditures:

```
# correlation test
cor(t.df$rate, t.df$hc, use="complete")
```

```
## [1] -0.0912598
```

The output of R's **cor** function tells us that the variables have a correlation of -0.091, which is reflective of the negative slope of the regression line shown in the scatterplot.

R's **lm** function provides us with the components of a characteristic equation for the relationship between the two variables:

```
# fit a model & plot for rate ~ healthcare spending
fit1 <- lm(rate ~ hc, data = t.df)

# output of lm function
summary(fit1)
```

```
##
## Call:
```

```
## lm(formula = rate ~ hc, data = t.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64.28 -41.24 -21.16   8.91 750.44
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.347957   2.154577  29.866  < 2e-16 ***
## hc          -0.007142   0.001931  -3.698 0.000225 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.84 on 1628 degrees of freedom
##   (270 observations deleted due to missingness)
## Multiple R-squared:  0.008328,   Adjusted R-squared:  0.007719
## F-statistic: 13.67 on 1 and 1628 DF,  p-value: 0.0002248
```

The characteristic equation is:

TB Infection Rate = 64.348 - 0.007142 * per capita healthcare expenditure

In other words, for each US dollar of per capita healthcare expenditure, the TB infection rate per 100,000 people should decline by an amount of 0.007142. However, according to the output of R's **lm** function shown above, this equation explains only 0.008328 of the variability we find in TB infection rates as evidenced by the R-Squared value. So while per capita healthcare expenditure is, in fact, a statistically significant predictor of TB infection rates as evidenced by its p-value of approximately zero, a linear least squares model of the relationship between the two variables doesn't actually help us all that much in explaining the variability in TB infection rates that we see throughout our data set.

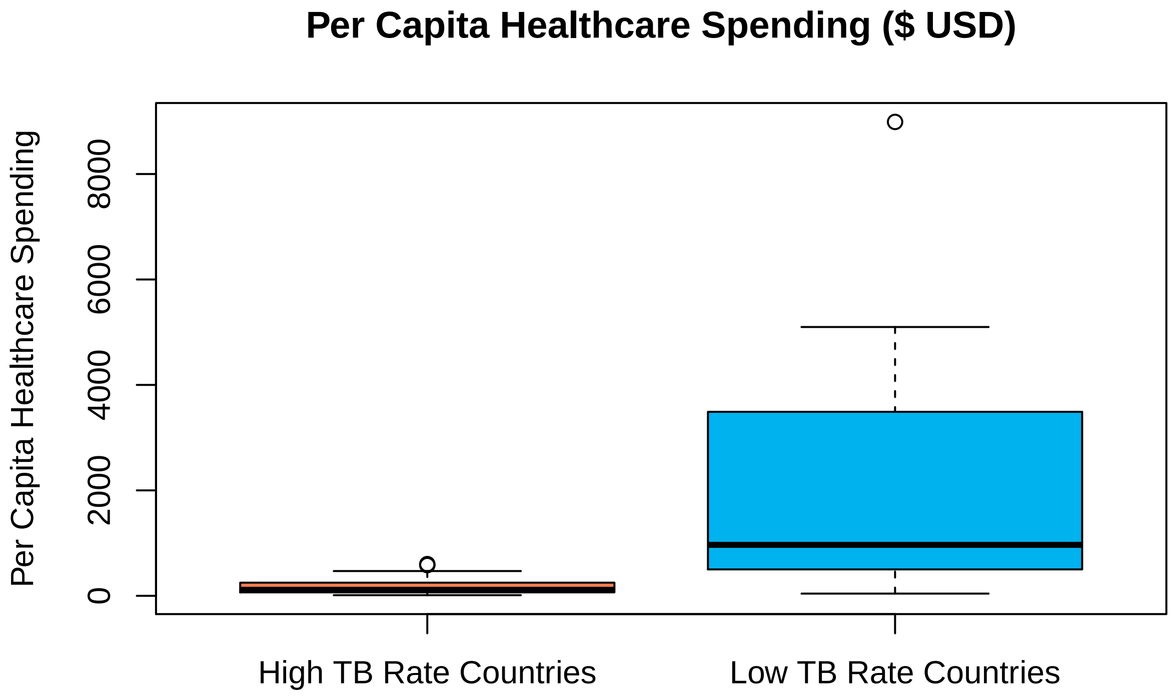
However, we can instead examine differences in per capita healthcare expenditure between the countries having the (on average) highest TB infection rates and those having the (on average) lowest TB infection rates. We'll make use of R's **summary** function as well as side-by-side boxplots of per capita healthcare expenditure data for the year 2013:

```r
top20.hc <- subset(hc_df, country %in% countrys & year == 2013)
summary(top20.hc$percap_hc, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   12.53   66.15  114.20  194.20  249.20  601.40       1
```

```r
bot20.hc <- subset(hc_df, country %in% bot_countrys & year == 2013)
summary(bot20.hc$percap_hc, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   42.21  544.80  966.90 2066.00 3392.00 8988.00
```



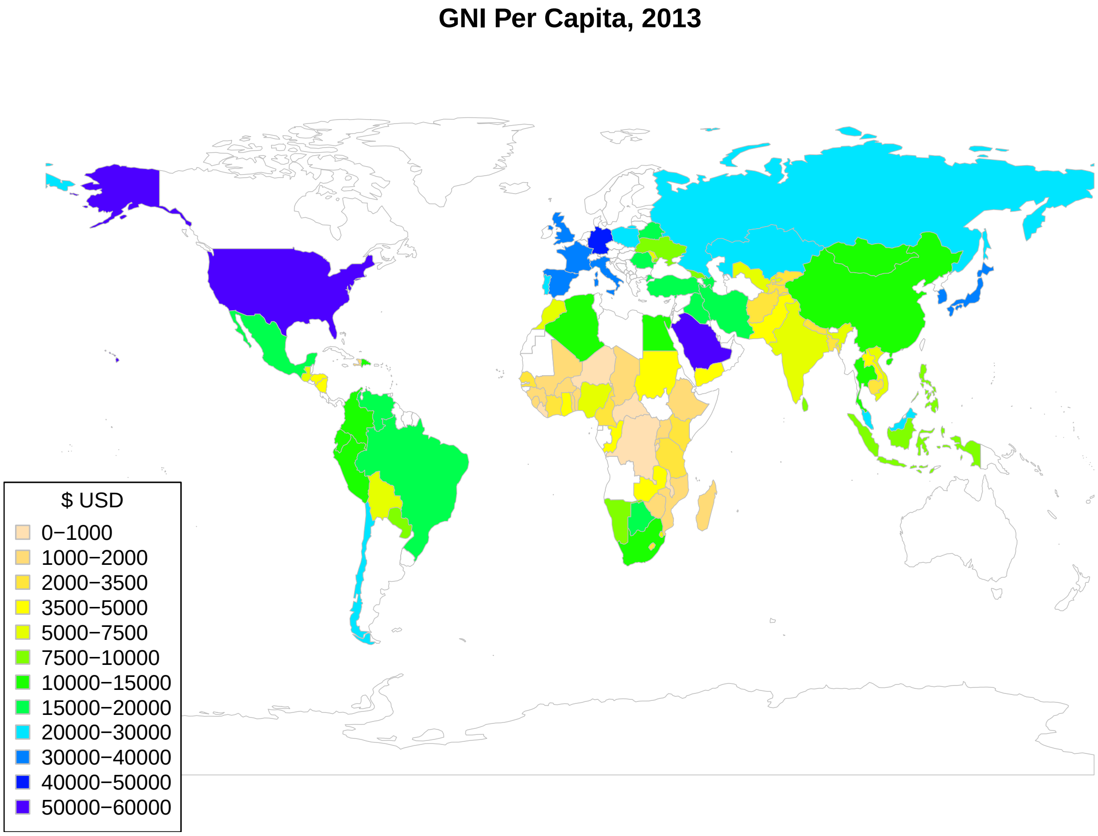
The boxplots shown above provide clear evidence of a striking difference in per capita healthcare expenditure between "high TB rate" countries and those with relatively low TB infection rates. The median per capita healthcare expenditure in a high TB rate country is approximately **$114.20 USD** while in low TB rate countries the median is **$966.90 USD**, or nearly 8.5 times as high.

Therefore, we've clearly found another "measurable" difference between countries with high TB rates and those with low TB rates. We'll now turn our attention to the per capita gross national income metric.

Gross National Income (GNI) Per Capita

The per capita gross national income data we've collected describes ***the per capita Gross National Income of 100 countries for the years 1995 - 2013.***. We'll start our analysis of the GNI data by creating a geoplot of per capita GNI for the countries in our data set for the year 2013 (the most recent year available in our data set):

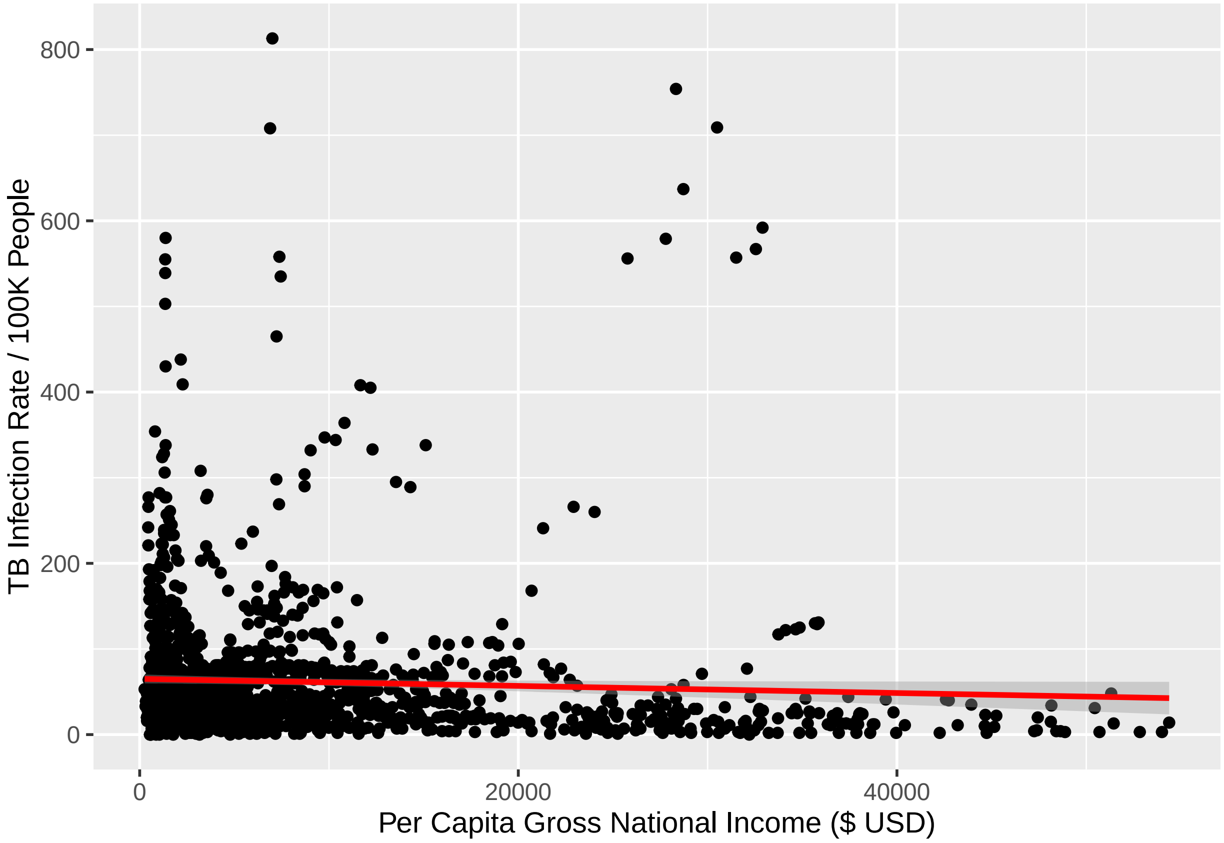
```
# get gni data for 2013
gni_df <- sqlQuery(con, "SELECT * FROM percap_gni WHERE year = 2013", stringsAsFactors=F)
```



The geoplot shows that countries in western Europe as well as the countries of the USA, Japan, and Saudi Arabia appear to have the highest per capita GNI as of 2013. As with the per capita healthcare expenditure data, many countries in Africa and south-central Asia appear to have relatively low amounts of per capita GNI.

To see whether we can find a relationship between per capita GNI and TB infection rates we'll first make use of a scatterplot and add a plot of the results of a linear regression model wherein we attempt to predict TB infection rates on the basis of a country's per capita GNI. The scatterplot is created using all TB infection rate data for the years 1995 - 2013 from all 100 countries contained within our data set.

```
# get gni data set
gni_df <- sqlQuery(con, "SELECT * FROM percap_gni", stringsAsFactors=F)
```



The scatterplot shows very clearly that countries that have relatively low per capita GNI also tended to experience the highest rates of TB infections during the 1995 - 2013 time period, while countries having per capita GNI in excess of roughly $38,000 USD consistently experienced realtively low rates of TB infection. The regression line shows a slight negative relationship between the two variables.

We can again use R's **cor** function to compute the correlation coefficient of the relationship between TB infection rates and per capita GNI:

```
# correlation test
cor(t.df$rate, t.df$gni, use="complete")
```

```
## [1] -0.05094459
```

The output of R's **cor** function tells us that the variables have a correlation of -0.0509, which is reflective of the slight negative slope of the regression line shown in the scatterplot.

R's **lm** function provides us with the components of a characteristic equation for the relationship between the two variables:

```
# fit a model & plot for rate ~ gni
fit1 <- lm(rate ~ gni, data = t.df)

# output of lm function
summary(fit1)
```

```
##
```

```
## Call:
## lm(formula = rate ~ gni, data = t.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64.90 -41.49 -22.89   8.93 750.78
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.1200022  2.6522248  24.553   <2e-16 ***
## gni         -0.0004139  0.0002046  -2.023   0.0432 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.04 on 1573 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.002595,   Adjusted R-squared:  0.001961
## F-statistic: 4.093 on 1 and 1573 DF,  p-value: 0.04323
```

The characteristic equation is:

TB Infection Rate = 65.12 - 0.0004139 * per capita GNI

In other words, for each US dollar of per capita GNI, the TB infection rate per 100,000 people should decline by an amount of 0.0004139. However, according to the output of R's **lm** function shown above, this equation explains only 0.002595 of the variability we find in TB infection rates as evidenced by the R-Squared value. So while per capita GNI is, in fact, a statistically significant predictor of TB infection rates as evidenced by its p-value of approximately 0.0432, a linear least squares model of the relationship between the two variables doesn't actually help us all that much in explaining the variability in TB infection rates that we see throughout our data set.

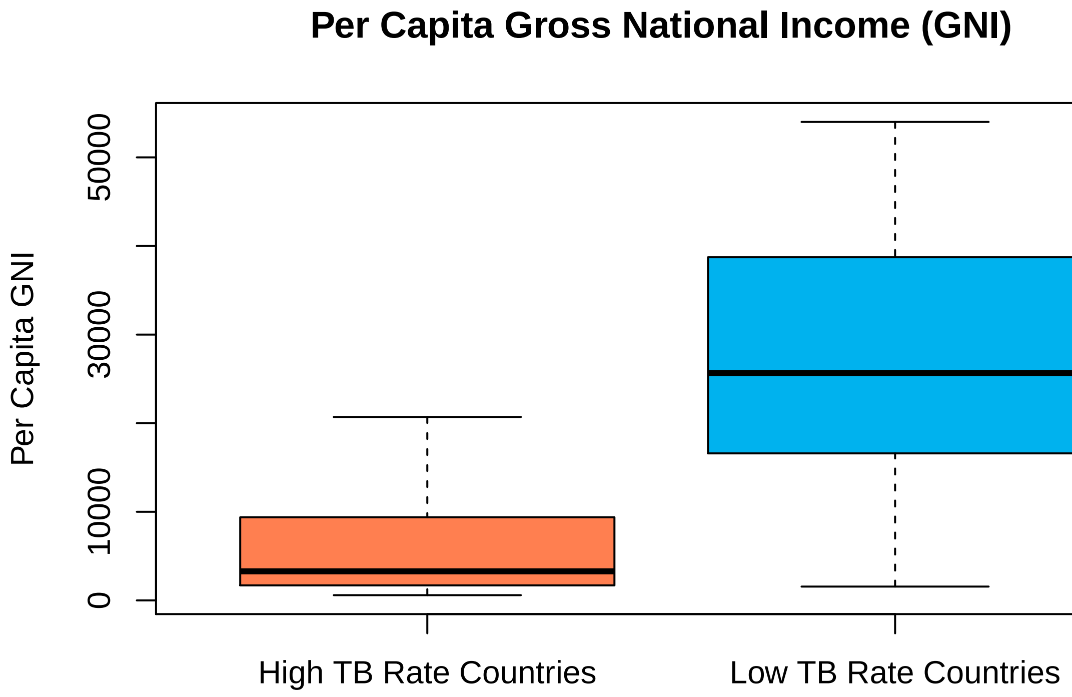
However, we can instead examine differences in per capita GNI between the countries having the (on average) highest TB infection rates and those having the (on average) lowest TB infection rates. We'll make use of R's **summary** function as well as side-by-side boxplots of per capita GNI data for the year 2013:

```
top20.gni <- subset(gni_df, country %in% countrys & year == 2013)
summary(top20.gni$percap_gni, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     590    1690    3280    5875    9380   20700       3
```

```
bot20.gni <- subset(gni_df, country %in% bot_countrys & year == 2013)
summary(bot20.gni$percap_gni, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1560   16870   25640   27500   38560   54000       2
```



The boxplots shown above provide clear evidence of a striking difference in per capita GNI between "high TB rate" countries and those with relatively low TB infection rates. The median per capita GNI in a high TB rate country is approximately **$3280 USD** while in low TB rate countries the median is **$25640 USD**, or more than 7.8 times as high.

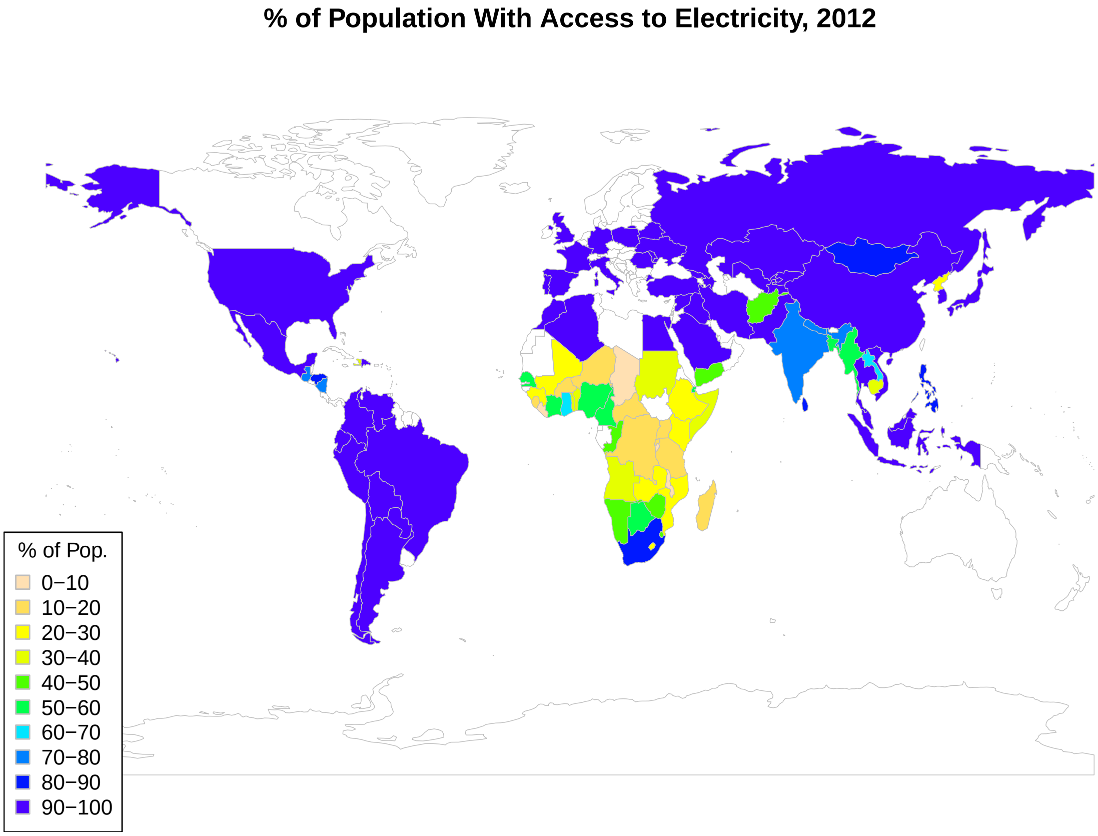
Therefore, we've clearly found another "measurable" difference between countries with high TB rates and those with low TB rates. We'll now turn our attention to the percentage of a country's population having access to electricity.

Access to Electricity (percentage of population)

The access to electricity data we've collected describes ***the percentage of a country's population having access to electricity in their homes for 100 countries for the years 2000, 2010, 2012.***. We'll start our analysis of the electricity data by creating a geoplot of the metric for the countries in our data set for the year 2012 (the most recent year available in our data set).

PLEASE NOTE: We will make use of the term "***Perc__Elec***" within this section to refer to **"percentages of a country's population having access to electricity in their homes"**.

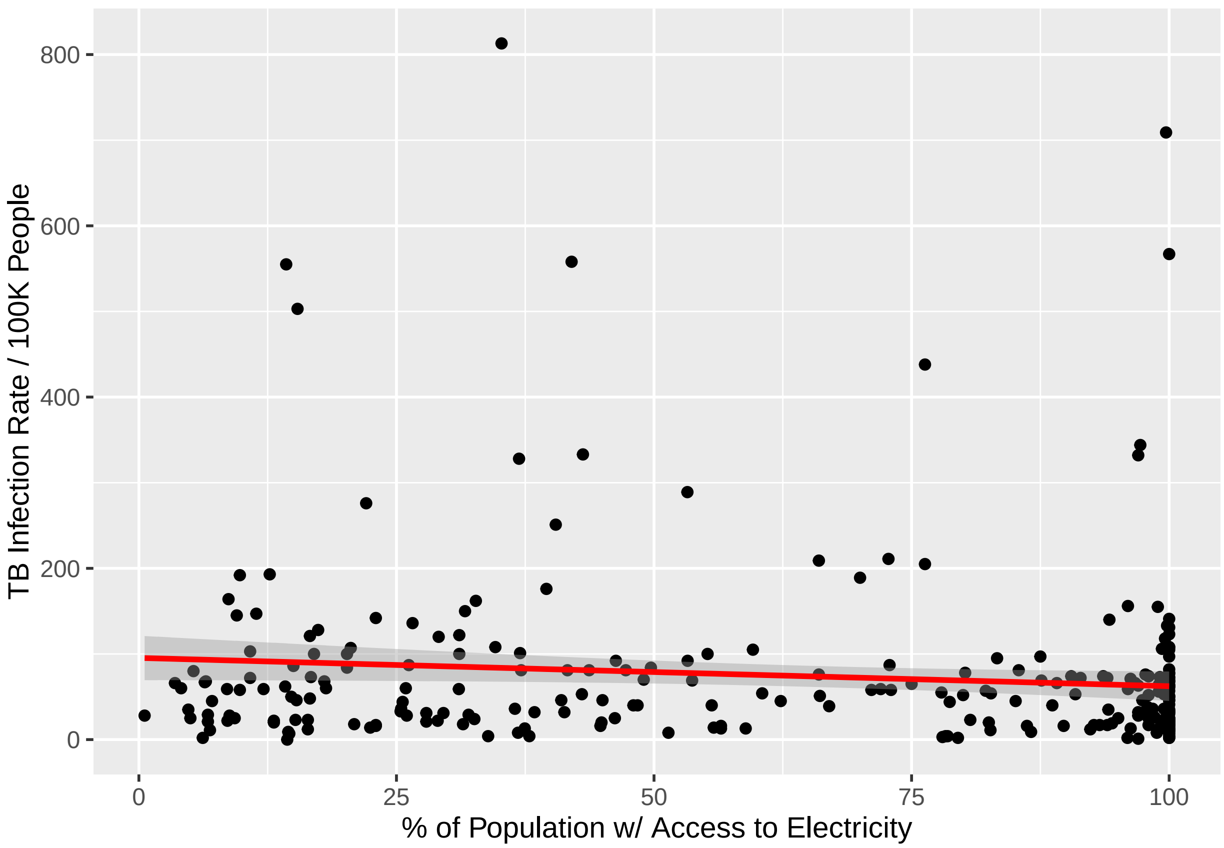
```
# get electricity data for 2012
eacc_df <- sqlQuery(con, "SELECT * FROM perc_e_acc WHERE year = 2012", stringsAsFactors=F)
```



The geoplot shows that only countries in Africa as well as the countries of Afganistan and a few countries in southeast Asia appear to have relatively low ***Perc_Elec***.

To see whether we can find a relationship between ***Perc_Elec*** and TB infection rates we'll first make use of a scatterplot and add a plot of the results of a linear regression model wherein we attempt to predict TB infection rates on the basis of a country's ***Perc_Elec***. The scatterplot is created using all TB infection rate data for the years 2000, 2010, 2012 from all 100 countries contained within our data set.

```
# get electricity data set
eacc_df <- sqlQuery(con, "SELECT * FROM perc_e_acc", stringsAsFactors=F)

# fetch TB rates data for 3 yrs only: 2000, 2010, 2012
t.df <- sqlQuery(con, "SELECT * FROM tb_rates WHERE year IN(2000, 2010, 2012)")
t.df$rate <- t.df$rate * 100000

# add electricity data to tb_rates dataframe
t.df$elec <- eacc_df$perc_e_acc
```



The scatterplot appears to show a slight negative relationship between TB infection rates and ***Perc__Elec*** for the years 2000, 2010, and 2012. However, relatively high rates of TB infection appear to have occurred in all categories of ***Perc__Elec***.

We can again use R's **cor** function to compute the correlation coefficient of the relationship between TB infection rates and ***Perc__Elec***:

```
# correlation test
cor(t.df$rate, t.df$elec, use="complete")
```

```
## [1] -0.1121609
```

The output of R's **cor** function tells us that the variables have a correlation of −0.1121609, which is reflective of the slight negative slope of the regression line shown in the scatterplot.

R's **lm** function provides us with the components of a characteristic equation for the relationship between the two variables:

```
# fit a model & plot for rate ~ electricity
fit1 <- lm(rate ~ elec, data = t.df)

# output of lm function
summary(fit1)
```

```
##
## Call:
```

```
## lm(formula = rate ~ elec, data = t.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -91.35 -50.42 -27.69   9.55 729.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  95.3987    13.1382   7.261 3.98e-12 ***
## elec         -0.3304     0.1768  -1.868   0.0628 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.7 on 274 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.01258,    Adjusted R-squared:  0.008976
## F-statistic: 3.491 on 1 and 274 DF,  p-value: 0.06278
```

The characteristic equation is:

TB Infection Rate = 95.3987 - 0.3304 * Perc_Elec

In other words, for each ***Perc_Elec***, the TB infection rate per 100,000 people should decline by an amount of 0.3304. However, according to the output of R's **lm** function shown above, this equation explains only 0.01258 of the variability we find in TB infection rates as evidenced by the R-Squared value. Furthermore, as evidenced by the p-value of 0.0628, ***Perc_Elec*** does not appear to be a statistically significant predictor of TB infection rates.

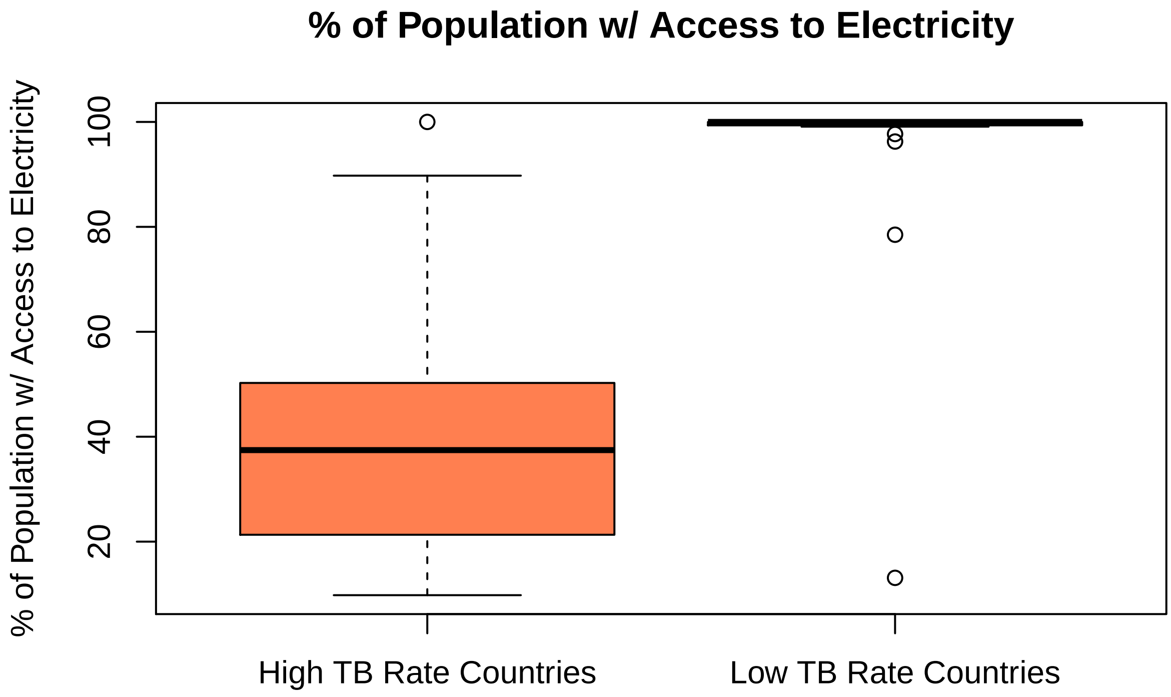
However, we can instead examine differences in ***Perc_Elec*** between the countries having the (on average) highest TB infection rates and those having the (on average) lowest TB infection rates. We'll make use of R's **summary** function as well as side-by-side boxplots of ***Perc_Elec*** data for the year 2012:

```
top20.elec <- subset(eacc_df, country %in% countrys & year == 2012)
summary(top20.elec$perc_e_acc, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   21.69   37.45   40.27   48.76  100.00
```

```
bot20.elec <- subset(eacc_df, country %in% bot_countrys & year == 2012)
summary(bot20.elec$perc_e_acc, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.10   99.47  100.00   94.20  100.00  100.00
```



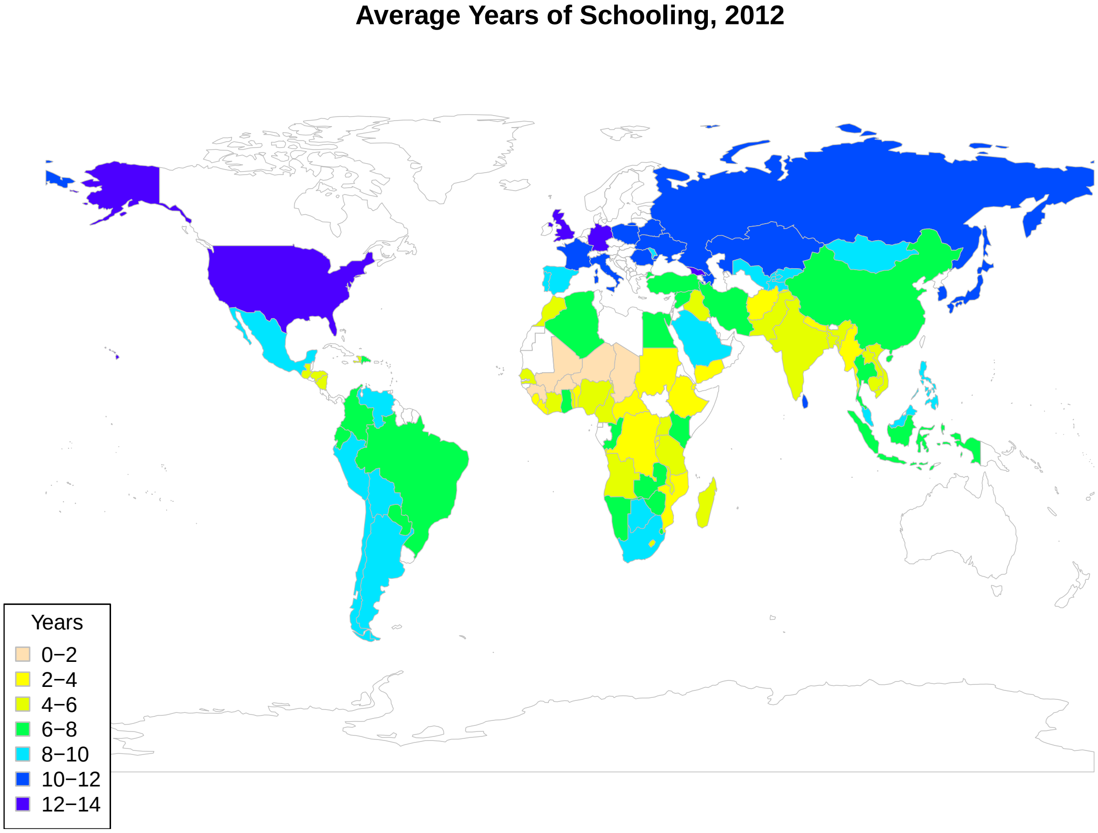
The boxplots shown above provide clear evidence of a striking difference in ***Perc_Elec*** between "high TB rate" countries and those with relatively low TB infection rates. The median ***Perc_Elec*** in a high TB rate country is approximately **37.45%** while in low TB rate countries the median is **100%**.

Therefore, we've clearly found another "measurable" difference between countries with high TB rates and those with low TB rates. We'll now turn our attention to the average years of schooling for a country's population.

Average Years of Schooling

The average years of schooling data we've collected describes ***the average years of schooling for the populations of 100 countries for the years 2000, 2005 - 2012.***. We'll start our analysis of the average years of school data by creating a geoplot of the metric for the countries in our data set for the year 2012 (the most recent year available in our data set).

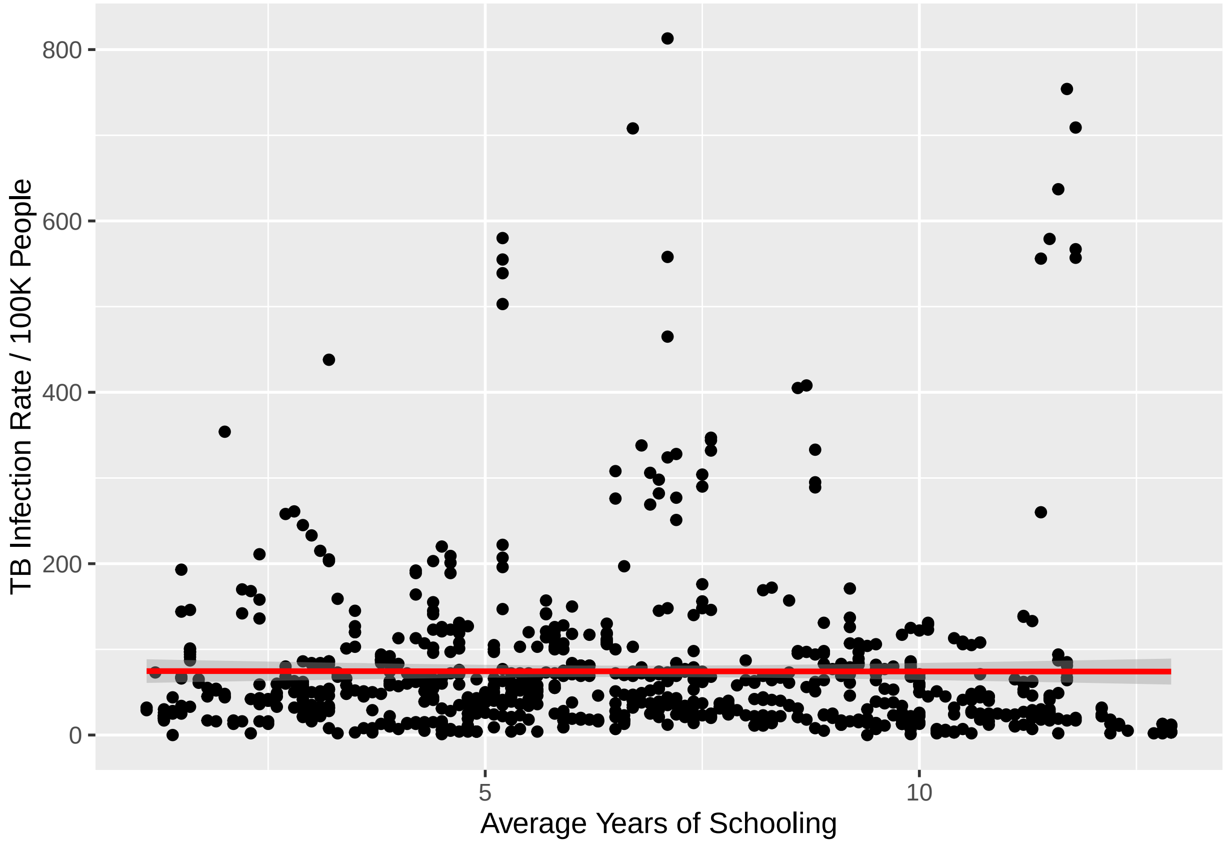
```r
# get avg yrs schooling for 2012
school_df <- sqlQuery(con, "SELECT * FROM yrs_school WHERE year = 2012", stringsAsFactors=F)
```



The geoplot shows that the populations of the USA, UK, Germany, Japan, and South Korea have the highest average years of schooling, while countries in Africa, Central America, south-central Asia, and southeast Asia appear to have relatively low levels of average educational attainment.

To see whether we can find a relationship between a country's average years of schooling and TB infection rates we'll first make use of a scatterplot and add a plot of the results of a linear regression model wherein we attempt to predict TB infection rates on the basis of a country's average years of schooling. The scatterplot is created using all TB infection rate data for the years 2000, 2005 - 2012 from all 100 countries contained within our data set.

```
# get avg yrs schooling data set
school_df <- sqlQuery(con, "SELECT * FROM yrs_school", stringsAsFactors=F)

# fetch TB rates data for only yrs: 2000, 2005 - 2012
t.df <- sqlQuery(con, "SELECT * FROM tb_rates WHERE year
                 IN(2000, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012)")

t.df$rate = t.df$rate * 100000

# add school data to tb_rates dataframe
t.df$school <- school_df$yrs_school
```



The scatterplot appears to show virtually no relationship between TB infection rates and average years of schooling for the years 2000, 2005-2012. Relatively high rates of TB infection appear to have occurred within nearly all categories of the metric.

We can again use R's **cor** function to compute the correlation coefficient of the relationship between TB infection rates and average years of school:

```
# correlation test
cor(t.df$rate, t.df$school, use="complete")
```

```
## [1] -0.001492464
```

The output of R's **cor** function tells us that the variables have a correlation of -0.001492464, which is reflective of the very flat nature of the regression line shown in the scatterplot.

R's **lm** function provides us with the components of a characteristic equation for the relationship between the two variables:

```
# fit a model & plot for rate ~ school
fit1 <- lm(rate ~ school, data = t.df)

# output of lm function
summary(fit1)
```

```
##
## Call:
```

```
## lm(formula = rate ~ school, data = t.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74.61 -51.20 -24.44   7.77 738.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.67509    8.12568   9.190   <2e-16 ***
## school      -0.04768    1.11226  -0.043    0.966
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.61 on 825 degrees of freedom
##   (73 observations deleted due to missingness)
## Multiple R-squared:  2.227e-06,  Adjusted R-squared:  -0.00121
## F-statistic: 0.001838 on 1 and 825 DF,  p-value: 0.9658
```

As evidenced by the p-value of 0.966 shown above, average years of schooling does not appear to be a statistically significant predictor of TB infection rates. Therefore, we will forgo the development of a characteristic equation for the two variables.

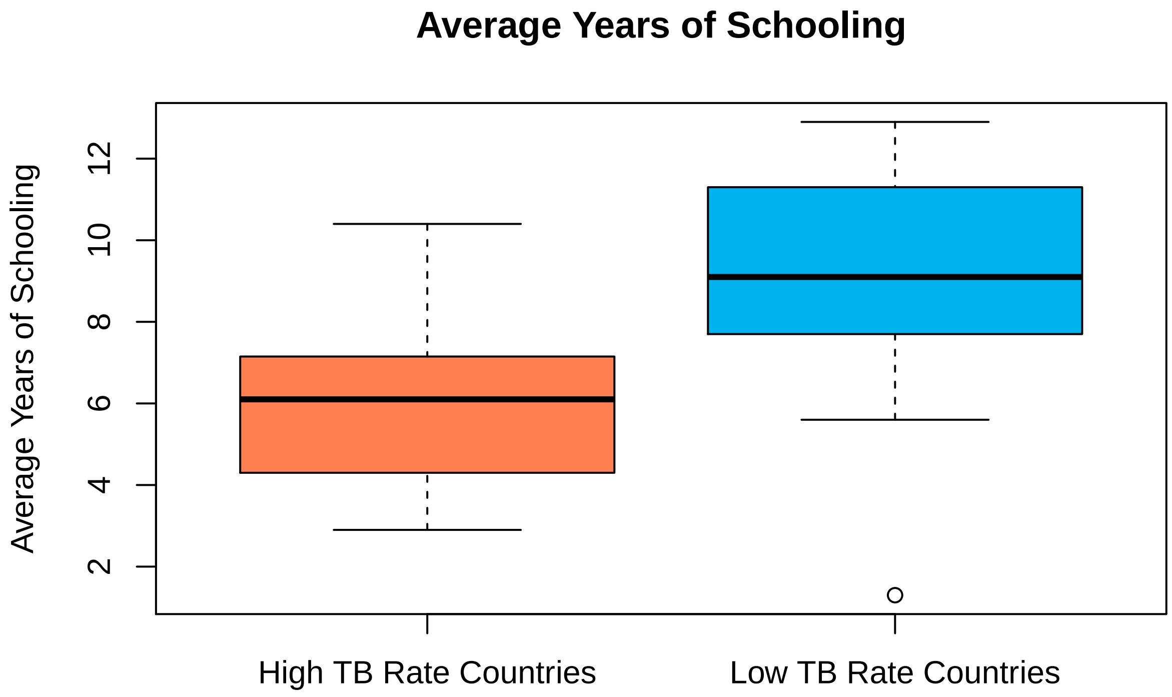
However, we can instead examine differences in average years of schooling between the countries having the (on average) highest TB infection rates and those having the (on average) lowest TB infection rates. We'll make use of R's **summary** function as well as side-by-side boxplots of average years of schooling data for the year 2012:

```
top20.school <- subset(school_df, country %in% countrys & year == 2012)
summary(top20.school$yrs_school, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.900   4.300   6.100   6.068   7.150  10.400       1
```

```
bot20.school <- subset(school_df, country %in% bot_countrys & year == 2012)
summary(bot20.school$yrs_school, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.30    7.75    9.10    9.05   11.20   12.90
```



The boxplots shown above provide clear evidence of a striking difference in the average years of schooling between "high TB rate" countries and those with relatively low TB infection rates. The average years of schooling in a high TB rate country is approximately **6.07 years** while in low TB rate countries the average is **9.05 years**.

Therefore, we've clearly found another "measurable" difference between countries with high TB rates and those with low TB rates. We'll now turn our attention to describing the characteristics of the metrics we've explored as they apply to both high rate TB countries and those with relatively low TB rates of infection.

"Describing" Countries Using Per Capita Metrics

Our analyses of the relationship between TB infection rates and individual per capita metrics revealed large disparities in those metrics between countries with relatively high rates of TB infections and those with relatively low rates of infection. The disparities are summarized in the table below.

Type of Country	Life Exp.	HC Exp.	GNI	Elec. Acc.	Schooling
Low TB Rate	76.63 yrs	$966.90	$25,640	100.00%	9.05 yrs
High TB Rate	59.66 yrs	$114.20	$ 3,280	37.45%	6.07 yrs
Difference:	16.97 yrs	$852.70	$22,360	62.55%	2.98 yrs

The table tells us that, on average, countries having the highest TB rates will have:

- An average life expectancy of **59.66 years**;

- Per capita health care expenditures of **$114.20 USD**;
- Per capita GNI of **$3,280 USD**;
- **37.45%** of their population having access to electricity in their homes;
- **6.07 years** of schooling.

By contrast, countries with the lowest TB rates will have on average:

- An average life expectancy of **76.63** years;
- Per capita health care expenditures of **$966.90** USD;
- Per capita GNI of **$25,640 USD**;
- **100%** of their population having access to electricity in their homes;
- **9.05 years** of schooling.

While beyond the scope of this study, the results of an investigation into the underlying reasons behind these disparities would likely be of great use to health authorities. For example, can these disparities be explained solely due to economic factors within specific countries or regions? Or are the economic situations within high TB rate countries the result of some set of factors that might prove to be influenceable via changes in governmental policies? There are many such questions that could be explored in future research efforts using the data we've collected for this study.

A Multivariate Linear Model for TB Infection Rates?

As we saw above, none of the individual per capita metrics we've examined appear to explain much of the variability we find in the TB infection rate data when linear least squares modeling is used. Of the linear models we examined, the model for electricity access showed the strongest correlation with the TB infection rates, with a correlation of -0.112 and an R^2 of 0.012, both of which indicate a rather weak relationship between the two variables.

Given the mimimal correlation between the per capita metrics and TB infection rates when using linear models, it seems unlikely that we'll be able to derive a multivariate linear least squares model that would prove useful in describing the variability we see in TB infection rates within the data. In fact, we would likely need to attempt to derive a model using an approach other than linear least squares if our goal is to find a useful generalized model of TB infection rates based on the five per capita metrics we've examined here. Finding such a model is beyond the scope of this study.

We can examine the results of a multivariate linear least squares model to verify whether or not our assumption regarding the likely invalidity of such a model are accurate. Specifically, we can generate a multivariate linear least squares model using data from the years 2000, 2010, 2012. Please note that we are constrained to those specific years by the data we have access to for the per capita electricity access metric.

```r
# Now run a linear model against all variables
mfit <- lm(rate ~ life_exp + hc + gni + elec + school, data = t.df)
summary(mfit)
```

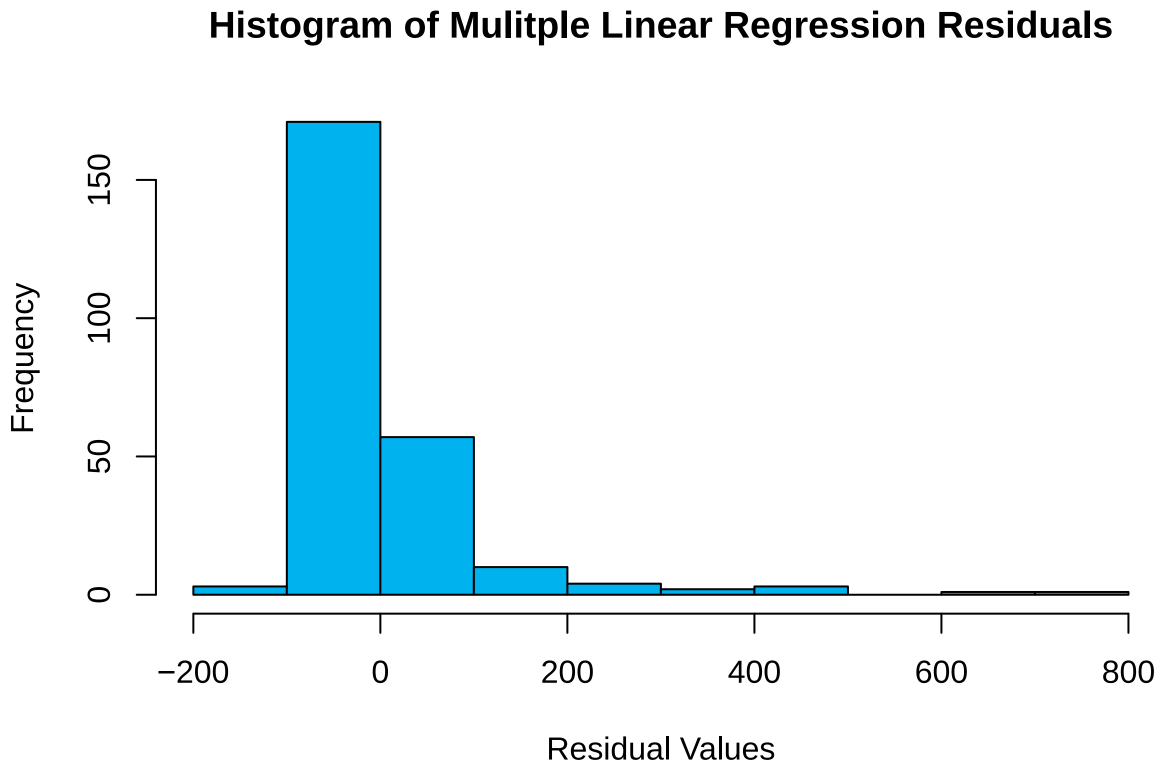
```
##
## Call:
## lm(formula = rate ~ life_exp + hc + gni + elec + school, data = t.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -105.96  -51.17  -23.20    9.20  706.12
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 56.871333  76.108310   0.747   0.4556
## life_exp     0.424866   1.429046   0.297   0.7665
## hc          -0.026182   0.011073  -2.364   0.0188 *
## gni          0.002044   0.001468   1.392   0.1652
## elec        -1.009221   0.419482  -2.406   0.0169 *
## school       8.059565   3.714982   2.169   0.0310 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 106.1 on 246 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.05002,    Adjusted R-squared:  0.03071
## F-statistic: 2.591 on 5 and 246 DF,  p-value: 0.02636
```

As evidenced by the R^2 value indicated above, the multivariate linear model we've derived explains only 5% of the variability we find in the TB infection rate data.

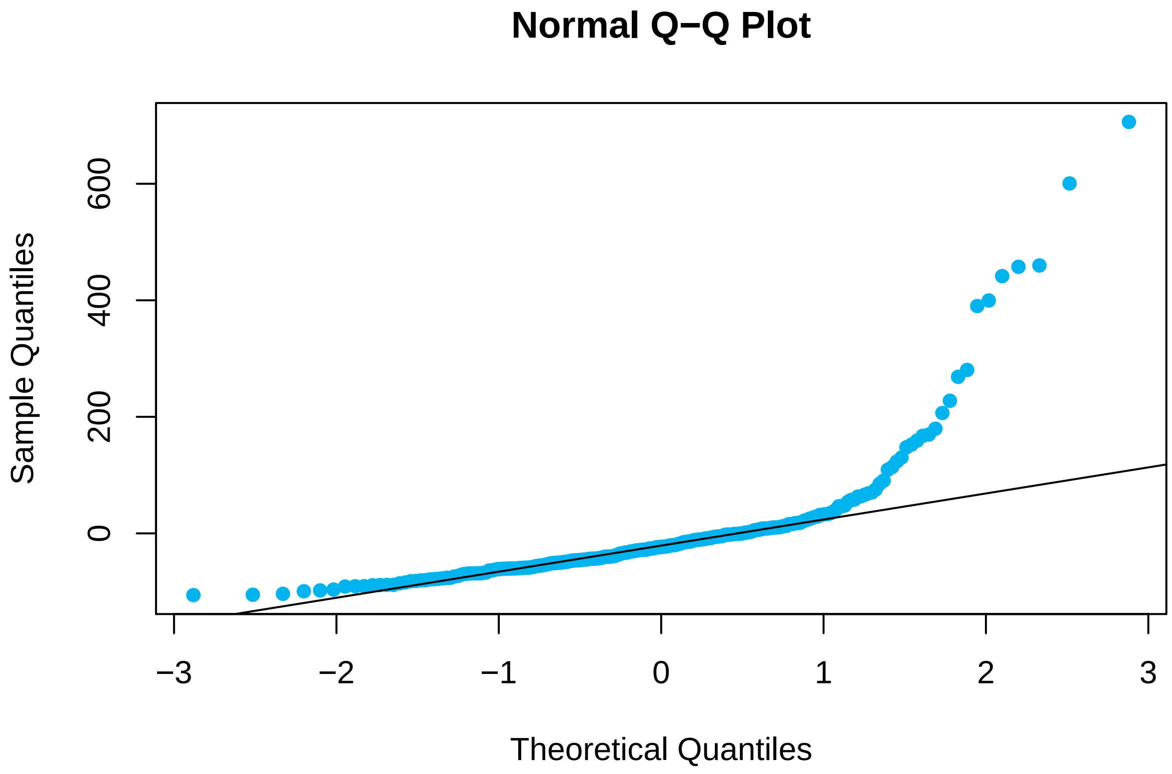
For a multiple linear least squares regression model to be considered valid we have 4 conditions that must be satisfied:

- The residuals of the model are nearly normal;
- The variability of the residuals is nearly constant;
- The residuals are independent;
- Each variable is linearly related to the outcome.

We can check the residuals of the model for normality using a histogram and a normal Q-Q plot:



Histogram of Mulitple Linear Regression Residuals



Both the histogram and the normal plot show clear evidence that the residuals of the model are **NOT** nearly normal, nor does the variability of the residuals appear nearly constant. Therefore, we cannot rely on this linear model for purposes of explaining the variability in TB infection rates we've found in our data.

Conclusion

In this study we've explored data related to TB case counts and infection rates for various countries throughout the world. Using that data we've mapped global "hotspots" where either high TB case counts or high rates of infection were found during the 1995 - 2013 time period.

We've discovered that, in general, most of the countries having relatively high rates of TB infections are located in sub-Saharan Africa and share a variety of characteristics, including relatively low per capita GNI, relatively low per capita healthcare expenditures, relatively low percentages of households with access to electricity, and lower average years of schooling and life expectancies than we find in countries having relatively low rates of TB infections.

We've also identified four adjacent countries in southern Africa that have experienced both very high TB case counts and high rates of TB infection during the 1995 - 2013 time period. These four countries (Namibia, South Africa, Swaziland, and Lesotho) therefore present a particular challenge for the world's health authorities as they work to limit the spread of TB.

Unfortunately, we were not able to derive a multivariate linear least squares model for TB infection rates using the five per capita metrics we collected due to the fact that the data failed to satisfy the requirements for a valid linear regression model. Further research using the data collected herein could perhaps include attempting to identify a descriptive model using a different approach (e.g., perhaps non-linear modeling).