Final Project: Understanding what Factors are Associated with Higher Life Expectancy

Mathew Joseph Jacob

Introduction

Many countries are interested in improving the life expectancy of citizens. However, it is important to know which factors are associated with life expectancy. Using data from the Gapminder data set and data from the World Bank, this analysis will look to find if improvements in water access, decreasing exposure to pollutants (using PM 2.5 as a proxy), or improving access to proper sanitation is associated with a larger improvement in life expectancy. The gapminder package data was used to provide data on a countries Life Expectancy and GDP per Capita (which was used experimented with as a control variable in the regression). Water access, polution exposure, and sanitation access data were obtained from the World Bank. By going to this *link* and selecting these indicators, the necessary data was downloaded as one single csv file which was renamed as worldbankdata.csv. The data used in this analysis has not been used for other classes or research projects.

- *Find the water access data here.*
- *Find the sanitation access data here.*
- *Find the PM 2.5 emissions access data here .*

A *study* on a similar topic was done in 1990 that looked into other factors that corressponded with Life Expectancy. However, aside from water access, this study focused on differeent factors such as energy consumption/capita, literacy rates, and family planning. I worked on this project alone.

Data wrangling

Gapminder Data Cleaning

First, several of the country names in the gapminder dataset were changed to match the names used by the World Bank to prevent these countries from being removed when the data sets were merged. The GDP per Capita (control variable) and Life Expectancy (dependent variable) data were put into their own data frames and pivoted. Plotting Life Expectancy against GDP per Capita resulted in a log pattern and so a log transformation was applied to the GDP per Capita data.

World Bank Data Cleaning

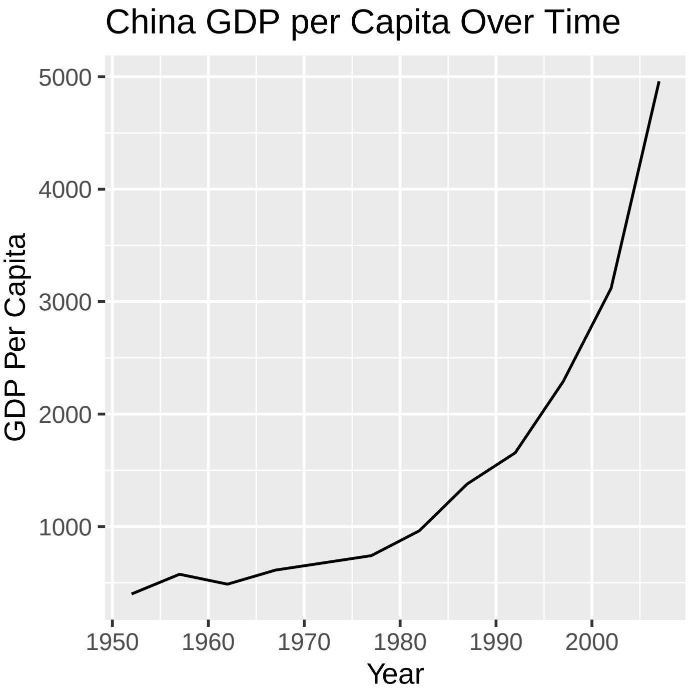
For each variable (water access, sanitation access, and polution exposure), the same steps were applied to clean the data. First, the data relevant to one particular variable was filtered from the worldbankdata.csv. Unnecessary columns were dropped, the numerical values were converted from Strings to numbers and absent values were set to NA. Finally, the columns were renamed to make it easier to access the data for a particular year. In order to make creating the line plots for these three variables easier, the variable dataframes were pivoted.

Merged Data

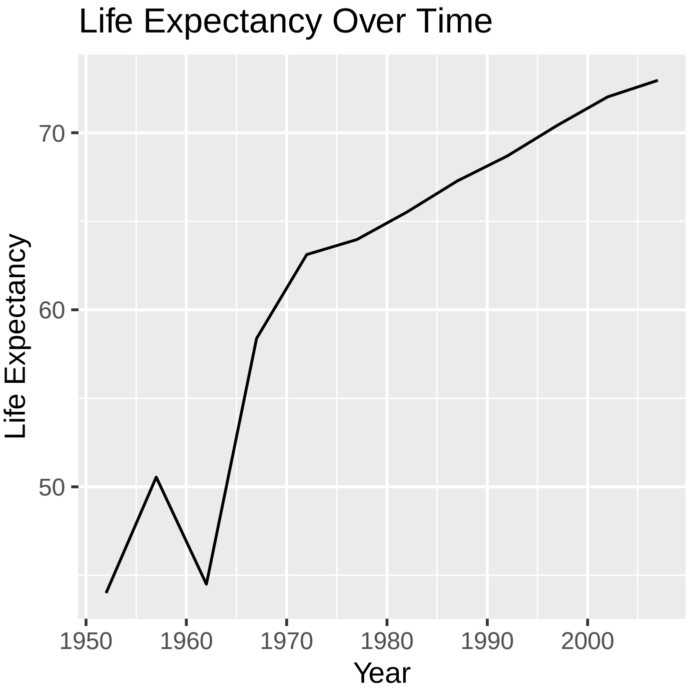
The GDP per Capita, Life Expectancy, water access, and sanitation access data from 2007 and the polution exposure data from 2010 (the closest year available) were merged into one data set for the regression analysis.

Visualizations: Interest variables over time and understanding GDP per Capita

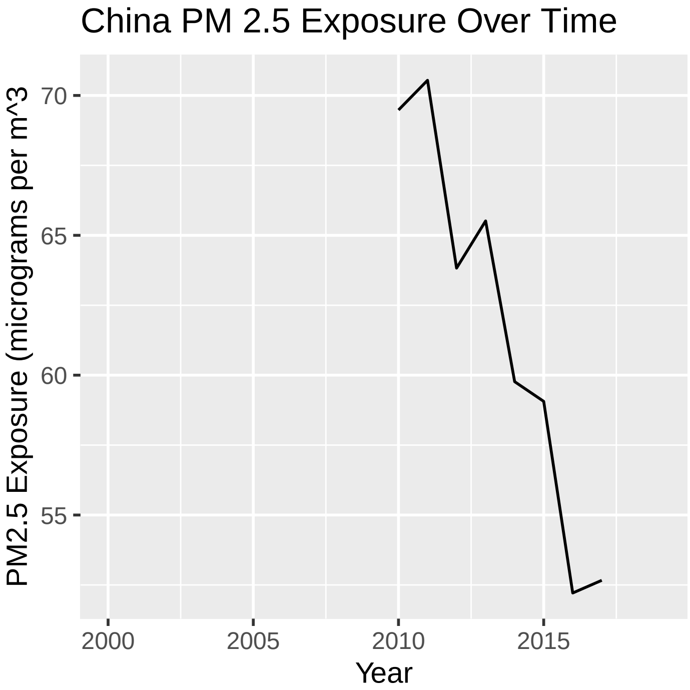
```
# Plots China GDP Over Time
ggplot(Chinagapminder) + geom_line(aes(x = year, y = gdpPercap)) +
    ggtitle("China GDP per Capita Over Time") + xlab("Year") +
    ylab("GDP Per Capita")
```



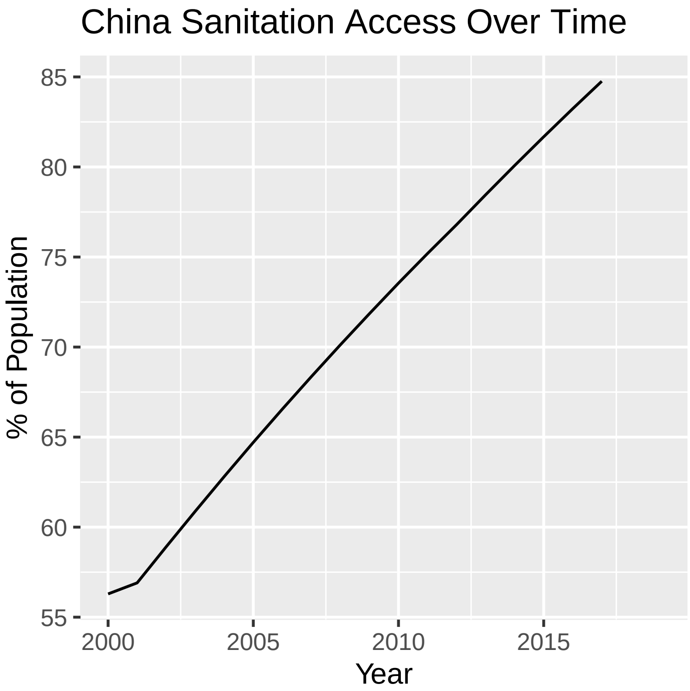
```
# Plots China Life Expectancy Over Time
ggplot(Chinagapminder) + geom_line(aes(x = year, y = lifeExp)) +
    ggtitle("Life Expectancy Over Time") + xlab("Year") + ylab("Life Expectancy")
```



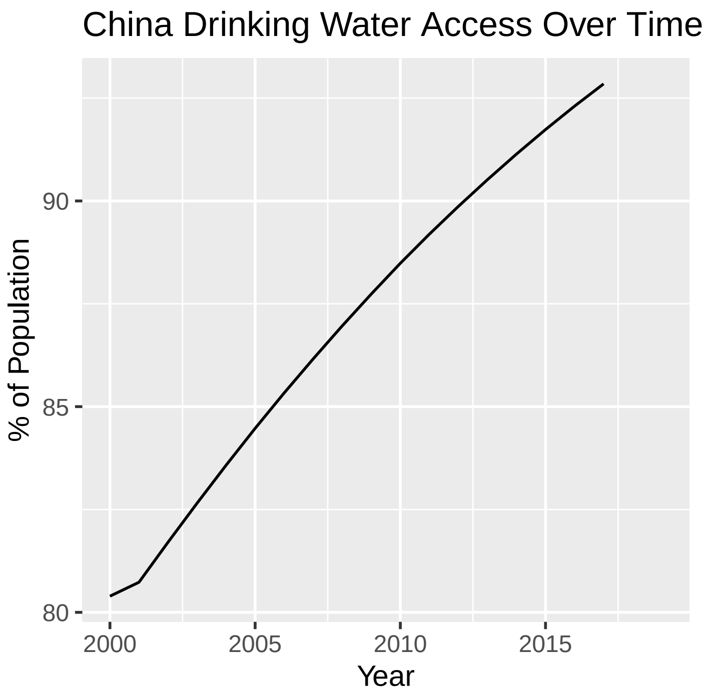
```r
# Plots China PM 2.5 Exposure Over Time
ggplot(Chinapollution) + geom_line(aes(x = year, y = value)) +
    ggtitle("China PM 2.5 Exposure Over Time") + xlab("Year") +
    ylab("PM2.5 Exposure (micrograms per m^3")
```



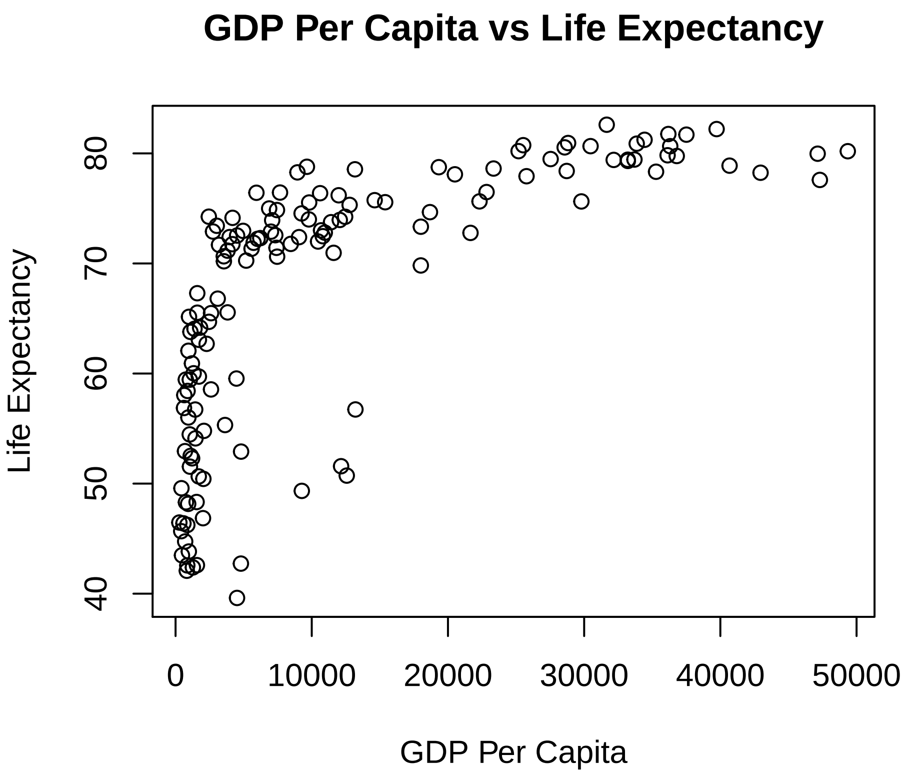
```r
# Plots China Sanitation Access Over Time
ggplot(Chinasanitation) + geom_line(aes(x = year, y = value)) +
    ggtitle("China Sanitation Access Over Time") + xlab("Year") +
    ylab("% of Population")
```



```
# Plots China Drinking Water Access Over Time
ggplot(Chinawater) + geom_line(aes(x = year, y = value)) + ggtitle("China Drinking Water Access Over Ti
    xlab("Year") + ylab("% of Population")
```



```
# Plot GDP per Capita vs LFE
plot(gdpPC$gdpPercap.2007, lfExp$lifeExp.2007, main = "GDP Per Capita vs Life Expectancy",
    xlab = "GDP Per Capita", ylab = "Life Expectancy")
```



Visualization Discussion

To get a better idea of the variables, the following plots were created using China as an example country. In the first two plots the data available went back to 1952. It shows that aside from a drop in 1960, China has been improving in these two categories over time. For the next three plots of Pollution Exposure, Sanitation Access, and Water Access, data was only available from 2000 onwards and the plots were restricted to this time frame to make it easier to understand. China saw improvements in Pollution Exposure (decreasing), and Sanitation and Water Access (increasing)

The final plot of GDP per Capita against LFE was used to show the logarithmic pattern in order to justify a log transformation of this variable.

Analyses: Regression Models

First Set of Regression Models

```r
model1 <- lm(lifeExp2007 ~ pollution2010, data = merged)

model2 <- lm(lifeExp2007 ~ water2007, data = merged)

model3 <- lm(lifeExp2007 ~ sanitation2007, data = merged)

stargazer(model1, model2, model3, header = F, title = "Table 1", omit.stat=c("f"), float = FALSE)
```

		Dependent variable:	
		lifeExp2007	
	(1)	(2)	(3)
pollution2010	-0.234^{***} (0.054)		
water2007		0.502^{***} (0.026)	
sanitation2007			0.319^{***} (0.016)
Constant	74.182^{***} (1.911)	26.199^{***} (2.169)	45.872^{***} (1.159)
Observations	136	137	137
R^2	0.125	0.736	0.754
Adjusted R^2	0.119	0.734	0.752
Residual Std. Error	11.210 (df = 134)	6.168 (df = 135)	5.957 (df = 135)
Note:			*p<0.1; **p<0.05; ***p<0.01

Second Set of Regression Models

```r
model4 <- lm(lifeExp2007 ~ pollution2010 + gdpPerCap2007 , data = merged)
model4$AIC <- AIC(model4)
```

```r
model4$BIC <- BIC(model4)

model5 <- lm(lifeExp2007 ~ water2007 + gdpPerCap2007 , data = merged)
model5$AIC <- AIC(model5)
model5$BIC <- BIC(model5)

model6 <- lm(lifeExp2007 ~ sanitation2007 + gdpPerCap2007 , data = merged)
model6$AIC <- AIC(model6)
model6$BIC <- BIC(model6)

model7 <- lm(lifeExp2007 ~ water2007 + sanitation2007 + gdpPerCap2007 , data = merged)
model7$AIC <- AIC(model7)
model7$BIC <- BIC(model7)

stargazer(model4, model5, model6, model7, header = F, title = "Table 2", omit.stat=c("f"), float = FALS
```

	Dependent variable:			
	lifeExp2007			
	(1)	(2)	(3)	(4)
pollution2010	−0.025			
	(0.036)			
water2007		0.328***		0.221***
		(0.041)		(0.049)
sanitation2007			0.234***	0.134***
			(0.031)	(0.036)
gdpPerCap2007	7.086***	3.212***	2.340***	1.717**
	(0.475)	(0.611)	(0.747)	(0.711)
Constant	6.812	12.685***	31.395***	25.379***
	(4.669)	(3.245)	(4.759)	(4.636)
Observations	136	137	137	137
R^2	0.672	0.781	0.771	0.802
Adjusted R^2	0.668	0.778	0.767	0.797
Akaike Inf. Crit.	915.691	867.550	874.067	856.232
Bayesian Inf. Crit.	927.342	879.230	885.747	870.832
Residual Std. Error	6.885 (df = 133)	5.636 (df = 134)	5.772 (df = 134)	5.389 (df = 133)
Note:				*p<0.1; **p<0.05; ***p<0.01

Conclusion

Based on the first set of regressions without controling for GDP per Capita (Table 1), as shown by the 3 stars (a p-value below 0.01) next to the coefficients of each model. The coefficient signs also make sense as increasing the amount of pollution leads to a decrease in life expectancy for the first model. On the other hand increasing the percent of the population that has access to sanitation services and clean water leads to an increase in life expectancy for the second and third model. However, while all coefficients are significant, they do not explain a comparable amount. Comparing the R^2 is a proper indication of this since we are not concerned about overfitting when we are only using one variable values (making adjusted R^2, AIC, and BIC less necessary to utilize). The R^2 for model 1 shows that only 0.125 percent of the variability in life expectancy is explained by pollution. Water and sanitation access on the other hand explain 0.736 percent and 0.754 percent respecctively of the variability in life expectancy.

However, because countries can be very different, adding a control variable such as GDP per Capita can help provide more information. When this variable is added to the models (Table 2), it can be seen that pollution is no longer significant. However, water and sanitation access are still significant at the 0.01 level and the R^2 increases as well for both models. Finally, creating a model with water and sanitation access and GDP per Capita resulted in a model where all three coefficients were signficant at least at the 0.05 level. The R^2 increased as well compared to the models with only sanitation or water access and GDP per Capita. However, since multiple variables are being used measures such as adjusted R^2, AIC, and BIC are better for making sure this addition is meaningful. All three scores indicate this final model is the best at explaining the variability of life expectancy.

This analysis shows that increasing water and sanitation access are associated with a higher life expectancy and that exposure to pollution (using PM 2.5 as a proxy) is not as signficant of a predictor. An important distinction is that this does not necessarily indicate causality. Experiments or using more complicated analyses (such as setting up a difference in differences study) are possible future areas to better understand whether water and sanitation access are a direct causal factor of a countries' life expectancy.

Reflection

I thought this project was a very helpful exercise in applying the concepts we learned in class. I felt that creating and comparing the regressions and visualizing the data for China as an example over time went well. I felt the most difficult part was cleaning the data because there were countries with different names from the different sources. Reshaping the data to make it easier for merging or creating a time series was also tricky. I also tried experimenting with other factors such as environmental health but did not include these in the final write-up because the data I found did not have as many of the countries that the gapminder package had and would have resulted in models comparing variables with a significantly different number of data points. I spent 5.5 hours on the analysis and 2 hours writing up the report.

Appendix

```r
# Complete code for the entire analysis.

#### SETUP ######################################################################

# Imports necessary libraries
library("gapminder")
library("ggplot2")
library("dplyr")

# Imports necessary data sources
data("gapminder")
gapminder <- data.frame(gapminder)
worldbankdata <- read.csv("worldbankdata.csv", as.is = T)



#### DATA CLEANING ##############################################################

# Fixes country names in gapminder data to match the World Bank Country Names
```

```r
corrections <- read.csv("corrections.csv", as.is = T)
countries = c()
for (country in gapminder$country){
  if(country %in% corrections$wrong){
    correct <- corrections[corrections$wrong == country,]$right
    countries <- c(countries, correct)
  }else{
    countries <- c(countries, country)
  }
}
gapminder$country <- countries

# Gets list of all country names and regions
countryinfo <- gapminder[!duplicated(gapminder$country),]
countryinfo <- countryinfo[ ,-c(3:6)]

# Creates GDP dataframe from Gapminder Data
gdpPC <- gapminder
gdpPC[, c(2,4,5)] <- NULL
gdpPC <- reshape(gdpPC, direction = "wide", idvar = "country", timevar = "year")
gdpPC <- merge(countryinfo, gdpPC, by = "country", all.x = T, all.y = T)
rownames(gdpPC) <- NULL

# Creates Life Expectancy dataframe from Gapminder Data
lfExp <- gapminder
lfExp[, c(2,5,6)] <- NULL
lfExp <- reshape(lfExp, direction = "wide", idvar = "country", timevar = "year")
lfExp <- merge(countryinfo, lfExp, by = "country", all.x = T, all.y = T)
rownames(lfExp) <- NULL

# Creates PM 2.5 Exposure (microgram/m^3) dataframe from Worldbank Data
pollution <- worldbankdata[worldbankdata$Series.Code == "EN.ATM.PM25.MC.M3", ]
pollution[, c(2,3,4)] <- NULL
pollution <- data.frame(pollution[, 1], apply(pollution[,-1], 2, as.numeric))
colnames(pollution) <- c("country", as.character(1960:2019))

# Creates Sanitation Access dataframe from Worldbank Data
sanitation <- worldbankdata[worldbankdata$Series.Code == "SH.STA.BASS.ZS", ]
sanitation[, c(2,3,4)] <- NULL
sanitation <- data.frame(sanitation[, 1], apply(sanitation[,-1], 2, as.numeric))
colnames(sanitation) <- c("country", as.character(1960:2019))

# Creates Drinking Water Access dataframe from Worldbank Data
water <- worldbankdata[worldbankdata$Series.Code == "SH.H2O.BASW.ZS", ]
water[, c(2,3,4)] <- NULL
water <- data.frame(water[, 1], apply(water[,-1], 2, as.numeric))
colnames(water) <- c("country", as.character(1960:2019))



#### DATA MERGING ##############################################################

# Gets 2007 Sanitation Data
```

```r
sanitation2007 <- data.frame(sanitation$country, as.numeric(sanitation$`2007`))
colnames(sanitation2007) <- c("country","sanitation2007")

# Gets 2010 Pollution Data (nearest year available)
pollution2010 <- data.frame(pollution$country, pollution$`2010`)
colnames(pollution2010) <- c("country","pollution2010")

# Gets 2007 Water Data
water2007 <- data.frame(pollution$country, as.numeric(water$`2007`))
colnames(water2007) <- c("country","water2007")

# Combines all data
merged <- data.frame(lfExp$country, lfExp$continent,
                     lfExp$lifeExp.2007, gdpPC$gdpPercap.2007)
colnames(merged) <- c("country","continent","lifeExp2007","gdpPerCap2007")
merged <- merge(merged, sanitation2007, by = "country")
merged <- merge(merged, pollution2010, by = "country")
merged <- merge(merged, water2007, by = "country")



#### VISUALIZATION #############################################################

# Gets China Gapminder Data
Chinagapminder <- gapminder[gapminder$country == "China", ]

# Gets China Pollution Data
Chinapollution <- pollution[pollution$country == "China", -c(2:41)]
Chinapollution <- reshape(Chinapollution, direction = "long",
                       varying = list(names(Chinapollution)[2:21]),
                       v.names = "value", idvar = c("country"),
                       timevar = "year", times = 2000:2019)

# Gets China Sanitation Data
Chinasanitation <- sanitation[sanitation$country == "China", -c(2:41)]
Chinasanitation <- reshape(Chinasanitation, direction = "long",
                       varying = list(names(Chinasanitation)[2:21]),
                       v.names = "value", idvar = c("country"),
                       timevar = "year", times = 2000:2019)

# Gets China Water Data
Chinawater <- water[water$country == "China", -c(2:41)]
Chinawater <- reshape(Chinawater, direction = "long",
                        varying = list(names(Chinawater)[2:21]),
                        v.names = "value", idvar = c("country"),
                        timevar = "year", times = 2000:2019)

# Plots China GDP Over Time
ggplot(Chinagapminder) + geom_line(aes(x = year, y = gdpPercap)) +
  ggtitle("China GDP Over Time") +
  xlab("Year") + ylab("GDP Per Capita")

# Plots China Life Expectancy Over Time
```

```r
ggplot(Chinagapminder) + geom_line(aes(x = year, y = lifeExp)) +
  ggtitle("Life Expectancy Over Time") +
  xlab("Year") + ylab("Life Expectancy")

# Plots China PM 2.5 Exposure Over Time
ggplot(Chinapollution) + geom_line(aes(x = year, y = value)) +
  ggtitle("China PM 2.5 Exposure Over Time") +
  xlab("Year") + ylab("PM 2.5 Exposure (micrograms per cubic meter")

# Plots China Sanitation Access Over Time
ggplot(Chinasanitation) + geom_line(aes(x = year, y = value)) +
  ggtitle("China Sanitation Access Over Time") +
  xlab("Year") + ylab("% of Population")

# Plots China Drinking Water Access Over Time
ggplot(Chinawater) + geom_line(aes(x = year, y = value)) +
  ggtitle("China Drinking Water Access Over Time") +
  xlab("Year") + ylab("% of Population")

# Plot GDP vs LFE and log transform GDP
plot(gdpPC$gdpPercap.2007, lfExp$lifeExp.2007,
     main = "GDP Per Capita vs Life Expectancy",
     xlab = "GDP Per Capita", ylab = "Life Expectancy")
merged$gdpPerCap2007 <- log(merged$gdpPerCap2007)


#### REGRESSION ANALYSIS #######################################################

model1 <- lm(lifeExp2007 ~ pollution2010, data = merged)

model2 <- lm(lifeExp2007 ~ water2007, data = merged)

model3 <- lm(lifeExp2007 ~ sanitation2007, data = merged)

model4 <- lm(lifeExp2007 ~ pollution2010 + gdpPerCap2007 , data = merged)
model4$AIC <- AIC(model4)
model4$BIC <- BIC(model4)

model5 <- lm(lifeExp2007 ~ water2007 + gdpPerCap2007 , data = merged)
model5$AIC <- AIC(model5)
model5$BIC <- BIC(model5)

model6 <- lm(lifeExp2007 ~ sanitation2007 + gdpPerCap2007 , data = merged)
model6$AIC <- AIC(model6)
model6$BIC <- BIC(model6)

model7 <- lm(lifeExp2007 ~ water2007 + sanitation2007 + gdpPerCap2007 , data = merged)
model7$AIC <- AIC(model7)
model7$BIC <- BIC(model7)
```