Causes of Death Across the World
And the Factors that Influence Them
Stat 3355.501

Team Probable Dead - Group 4
Andy Boulos, James Boyer, Omshravan Manickavelu
April 8, 2020

Introduction

The purpose of this project was to use statistics to gain further insight into the causes of life expectancy across the globe. Our group obtained and analyzed large datasets from three different sources. The largest dataset came from the World Health Organization [1]. This dataset contained the number of deaths from different causes for multiple countries. The values that were taken from the dataset included deaths by communicable diseases (*comm*), non-communicable diseases (*non_comm*), and injuries (*injury*). It also had thirty-two different more specific causes of death. Given that there were so many different causes, we chose the top five leading causes of death. Our group chose cardiovascular diseases (*cardio*), cancer (*cancer*), musculoskeletal diseases (*musculo*), neonatal diseases (*neonatal*), and diarrhea (*diarrhea*). Our second source was

the World Bank, which had economic factors such as GDP (*GDP*), GDP per capita (*GDP_cap*), and GDP per employed person (*GDP_PEP*) [2]. The third source was Our World in Data [3]. This contains data for other factors that we wanted to look into. Among these factors were unemployment (*unemployment*), health expenditure as a percentage of GDP (*spending*), and Deaths per 100,00 from unsafe sanitation (*sanitation*). Since the datasets were formatted differently, and sometimes incomplete, we decided for the purposes of this report to use 16 countries that were common from all the datasets from 1996 to 2015. Furthermore, we tried to choose these countries so that they were uniformly distributed with respect to geographic location and GDP. We created this dataset because we wanted to explore the relation between causes of death, economic factors, and region.

Initially, when we began our research, we intended to focus strictly on the differences of how income levels affect causes of death. As we progressed, we realized there were many aspects that were not related to economics. We realized that these factors could have a much higher correlation and importance when it comes to the cause of death. In response, we chose to focus on analyzing a wide variety of factors that might distinguish countries from each other. Some of the factors were the type of government, level of inequality, unemployment rate, and healthcare system of these countries. We also focused on identifying outliers and the causes for these statistical anomalies. Our group also hoped to identify how different factors correlated with each other. Initially, we had to create our own dataset containing all the information from the sources in a format that was feasible. We began by plotting the change over time for income and life expectancy to identify general trends of each country. Next, we looked at which factors had the highest correlation and found that they typically related more to sanitation than to GDP. Finally, our group plotted the correlation for all these relationships. Through our analysis we were able to relate certain causes of death to other factors and identified outliers for these cases, detailed in Section 3 of this report.

Data Cleaning

As mentioned before, we constructed our dataset using the various charts from *Our World in Data, The World Bank* and *The World Health Organization* [1, 2, 3]. We choose 20 years of data (1996-2015) for 16 randomly chosen countries. We randomly picked these countries so that they were uniformly distributed across various continents and income levels. We filtered each variable for the selected countries and years and then added them to a google sheet document so we could keep all the information in one place. When trying to explore the dataset, we simply converted the google sheet document into a .csv file and used R Studio. There were three main ways we modified the data so it was more manageable.

First, we converted some continuous variables into factor variables. We found a political regime variable that assigned a numerical value to each country based on how autocratic or democratic the government was each year. Since most of the countries fell to one extreme or the other, we converted these numbers into three categories: Mostly Democratic, Middle, and Mostly Autocratic. Furthermore, we created a factor variable based on the percentile the country fell into when it came to income levels. We also converted the variable of the proportion by broad age group into a factor variable of the largest broad age group. Finally, we wanted to see if there was any relationship between certain variables like GDP or Unemployment and continent. Therefore, we made a region variable based on the part of the world the country was in and then found the average for all the countries in that region for each variable.

Second, we extrapolated data when certain years were missing in the original data set. For example, the GINI inequality measure for some countries was only available for the years 1990 and 2015. Based on the countries that did have an inequality measure for every year, we observed that the number was relatively constant. Using constancy as an assumption, we used the seq() function in R to create a linear trend between the 1990 and 2015 numbers and then used the results from 1996-2015 in our dataset. Additionally, for Health Expenditure as a % of GDP, the year 2015 was missing so we duplicated the 2014 number to fill in the gap. For the political regime variable, there were no numbers assigned to Luxembourg. To fill this gap, we looked at how countries with similar governments were measured and used those numbers for Luxembourg.

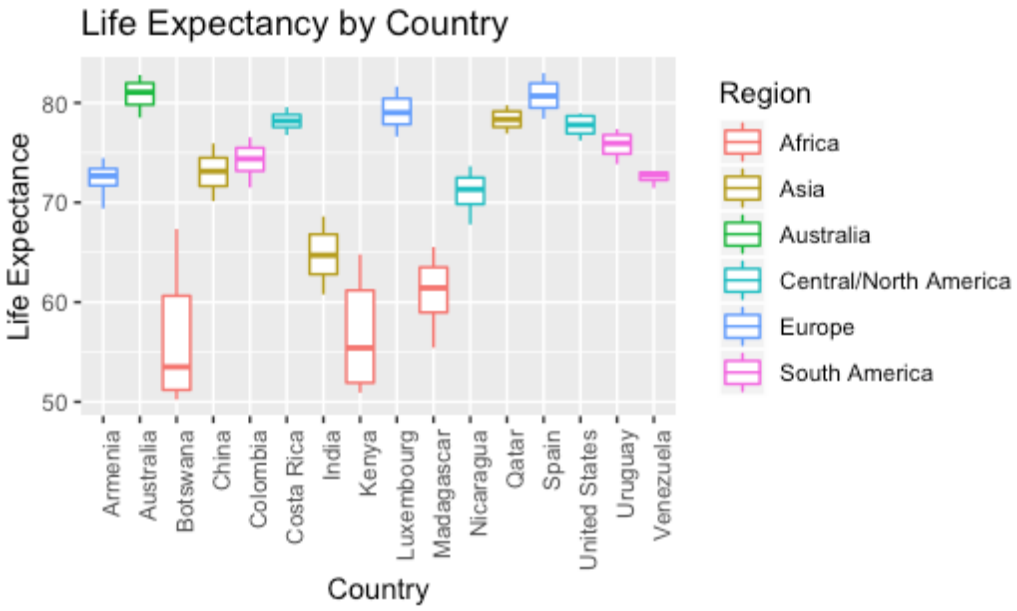
Third, we normalized certain variables so that countries of different populations were more comparable. For instance, we wanted to examine the relationship between the total number of deaths by communicable disease and certain causes of death like diarrhea. Due to their larger populations, the total number of deaths in China and India were so much larger than the number of deaths in other countries that we could not reasonably infer any information from it. In response, we computed the percentage of deaths by communicable disease compared to injury and non-communicable disease. This number was easily comparable between countries of different population sizes.

To summarize, the overall process was not difficult. Once we found a way to filter the variable for the selected countries and years, we were able to put the dataset together relatively quickly. The most difficult part of the process arose from the fact that the various variables we wanted to use were often measured by different organizations and therefore were measured in different units and for different years. In order to create a cohesive dataset where variables were comparable, we had to use a variety of methods to complete and normalize the dataset.

Analysis and Findings

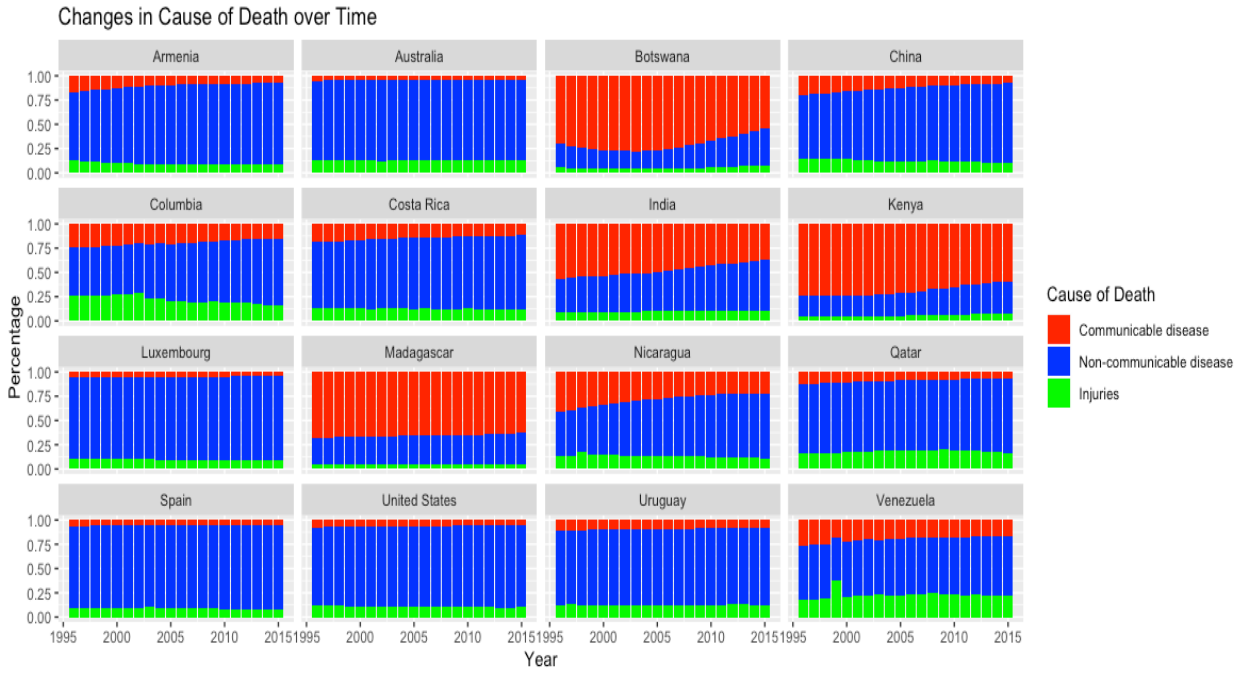
In this section, we will discuss all the questions that our group had with regard to our data and our findings. We also include possible causes along with notable findings. These questions are designed to show that health and disease can be affected by a variety of factors. The questions with more interesting findings will be focused on in greater detail.

3.1 Does geography have an effect on life expectancy?



One of the first possibilities we explored was the effect of geographical location on life expectancy. This helped to identify any glaring outliers and give a general idea of the life expectancy for each of the countries that we used for our analysis. The most obvious outlier from this visual is that African countries have the lowest life expectancy, around 50 to 60 years. When looking at the change in life expectancy over time, we observed that the life expectancy has increased in varying amounts for most of these regions. For instance, African countries also tended to have the largest variation in life expectancy. This is likely due to the fact that living situations have improved drastically. For example, Botswana had a range of 15 years whereas Venezuela had a range of 3 years. Following African countries, Asian countries had the next highest variation in life expectancy. Similarly, we predict this is due to the fact that life standards regarding health have improved greatly in this region. Aside from these two, variation was typically under 5 years and life expectancy tended to hover between 70 and 80. These observations serve more to lay a foundation for investigating the specific causes for death in each country and the reasons for variations.

3.2 What are the primary causes of death by broad category in each country?

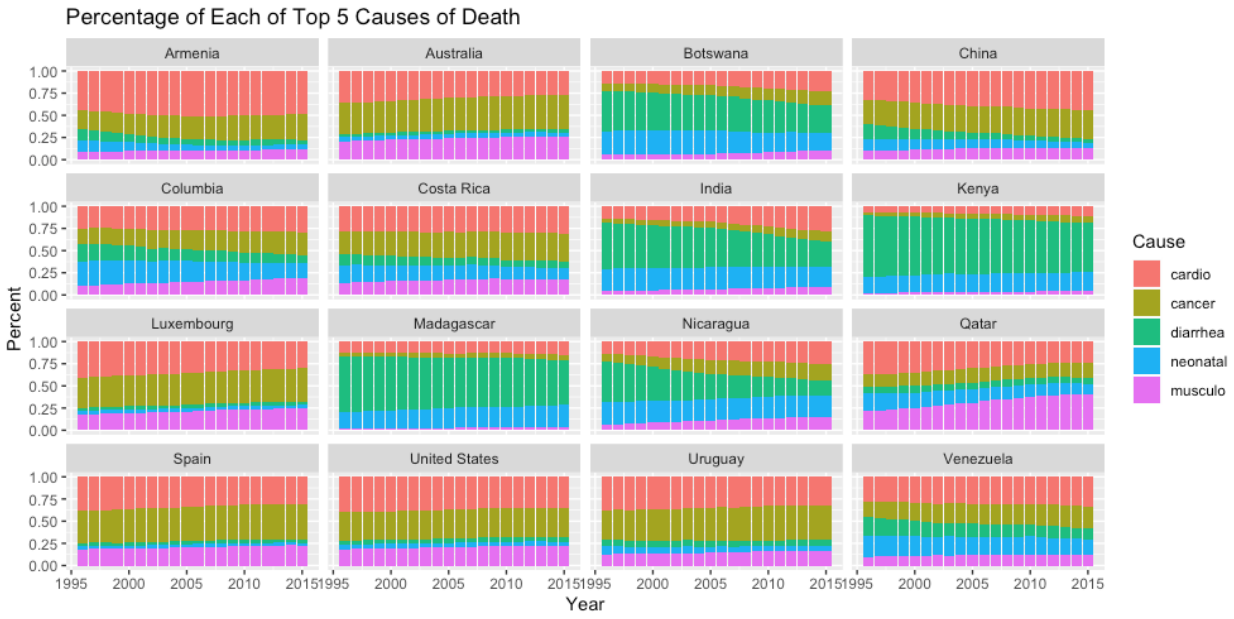


We then examined the causes of deaths in a country when separated by the broad categories of communicable disease, non-communicable disease, and injuries. We noticed that the majority of deaths in most countries is caused by non-communicable diseases such as heart disease or cancer.

We observed that death by injury was fairly steady across all the countries. However, we noticed that Colombia, Qatar, and Venezuela had above average levels of death by injury. After some research, we predict this is due to the prevalence of violence and conflict in these regions. For example, Colombia has experienced over 220,000 deaths due to conflicts with the guerrilla movement FARC [4]. In Venezuela, there have been a variety of political conflicts and revolutions between 1996-2016 [5]. Finally in Qatar, there have been regional conflicts with Saudi Arabia since 1995 [6]. Finally, one other deviation we noticed was the 1999 spike of injury related deaths in Venezuela. After some brief online research, we found that there was a tragic flood that took the lives of about 20,000 Venezuelans in 1999 [7]. Our main takeaway from these observations is that data analysis on social and economic variables cannot be practiced in isolation. Apart from some external research into current events and political issues, there is no easy way we could have grasped these deviations and understood the patterns.

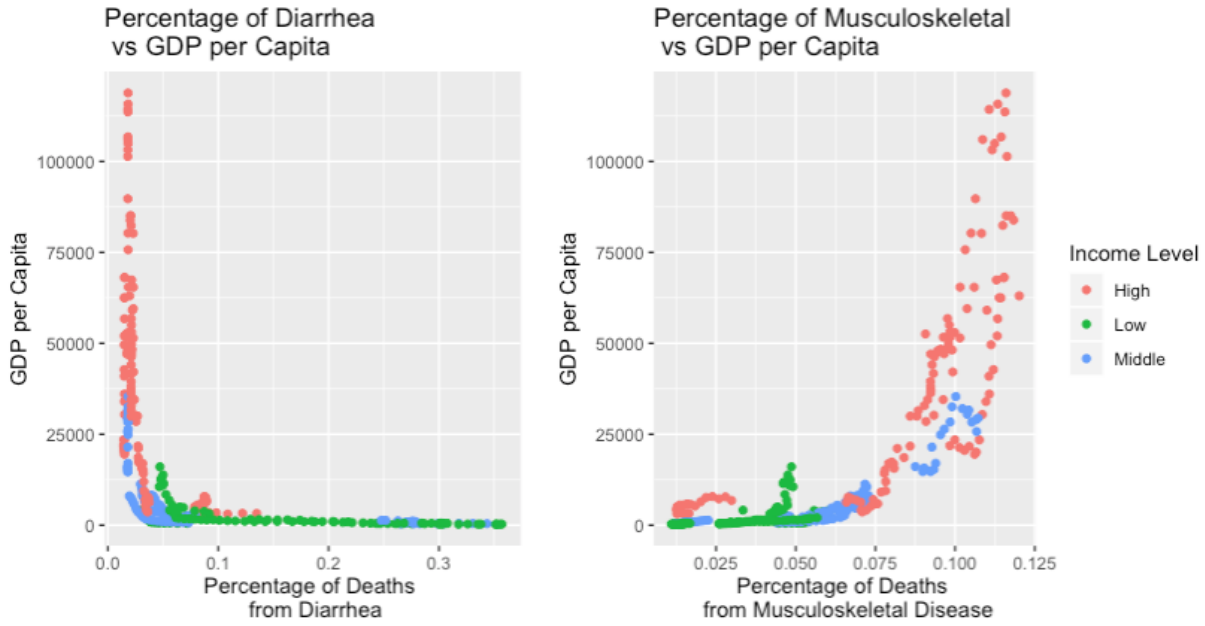
The final main observation was that several countries (Botswana, India, Kenya, Nicaragua, and Madagascar) had a disproportionately high percentage of deaths caused by communicable diseases compared to the majority of countries.

3.3 What are the top specific causes of death worldwide?



Upon discovering the broad causes of death in each country, we wanted to learn more about what specifically are the leading causes of death. The top 5 causes of death among the 16 countries are cardiovascular diseases, cancer, musculoskeletal disorders, diarrhea, and neonatal disorders. In the figure above, the proportion of causes of death within each country varied. For most countries, cancer and cardiovascular diseases were the top two most common causes of death. We predict these are the top two causes because these countries are well-developed, have a normal age distribution, and have relatively high life expectancies. Cancer and cardiovascular diseases are expected because they are both causes of death that occur most predominantly in older people. In countries such as Botswana, India, and Madagascar, we predict life expectancies are shorter because communicable diseases affect all ages and thus are disproportionately represented in the top five causes of death. As expected, diarrhea and neonatal disorders are relatively high. Over time in each country, the proportions only slightly vary. In India and Botswana, diarrhea deaths decrease in proportion while deaths due to cancer increase. This is anticipated due to their respective increases in life expectancy and GDP per capita over time. In countries with low poverty rates and high government health expenditures, the proportions remained relatively constant. We anticipate that proper sanitation and convservative lifestyles in these countries provide a stability to the causes of death.

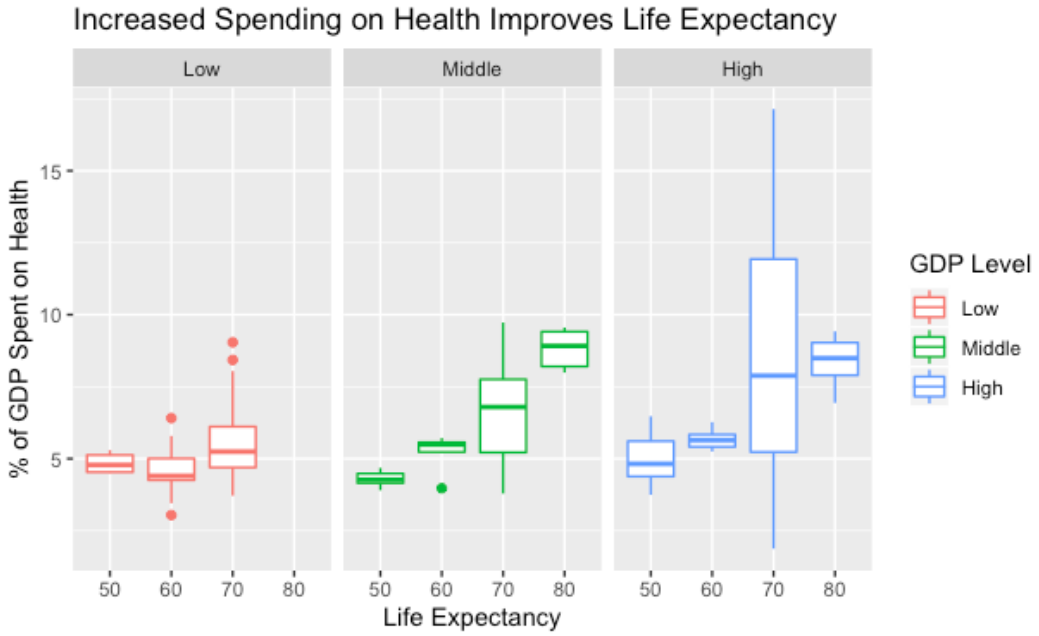
3.4 Which diseases are affected by economic factors? Why might this correlation exist?



After we had identified the top causes of death in each country, we wanted to know what impact did economic factors play for life expectancy. Initially, our group was trying to compare overall GDP to the total number of deaths from these diseases. The issue with this approach is that GDP largely depends on the size of the country and says very little about the actual economy of a country. Similarly, the total deaths from any cause is not an accurate indicator because it highly depends on the population of the country. Therefore, we decided to change our variables so that they would more accurately show correlation. We chose GDP per capita since it takes GDP and divides it by the population. In order to represent deaths, we divided the number of deaths from the cause by the total number of deaths in that country. Both of these categories allow us to reduce the effect of the size of the population. As shown on the left, as GDP per capita decreases, the percentage of deaths increases. The most notable shift happens when GDP per capita falls below 25,000 and income level begins to shift from high to middle and low income. The percentage of deaths from diarrhea increases up to 35% when the GDP per capita falls low enough. Diarrhea typically comes from gastrointestinal infections and leads to severe dehydration which is why it remains a leading global killer. While many of these infections are treatable, treatment costs money. These infections essentially come from food poison such as Salmonella which results from eating undercooked or contaminated food. In higher income countries, properly cooked, safe food is hardly an issue. However, in impoverished countries, more people struggle to feed their families and will eat whatever is available. Comparatively, musculoskeletal disorders come from obesity, repetitive motions, and poor fitness. These conditions are far more common in wealthy countries and seem to basically describe an office environment. People eat whatever food is convenient (typically fast food) and live a stagnant life

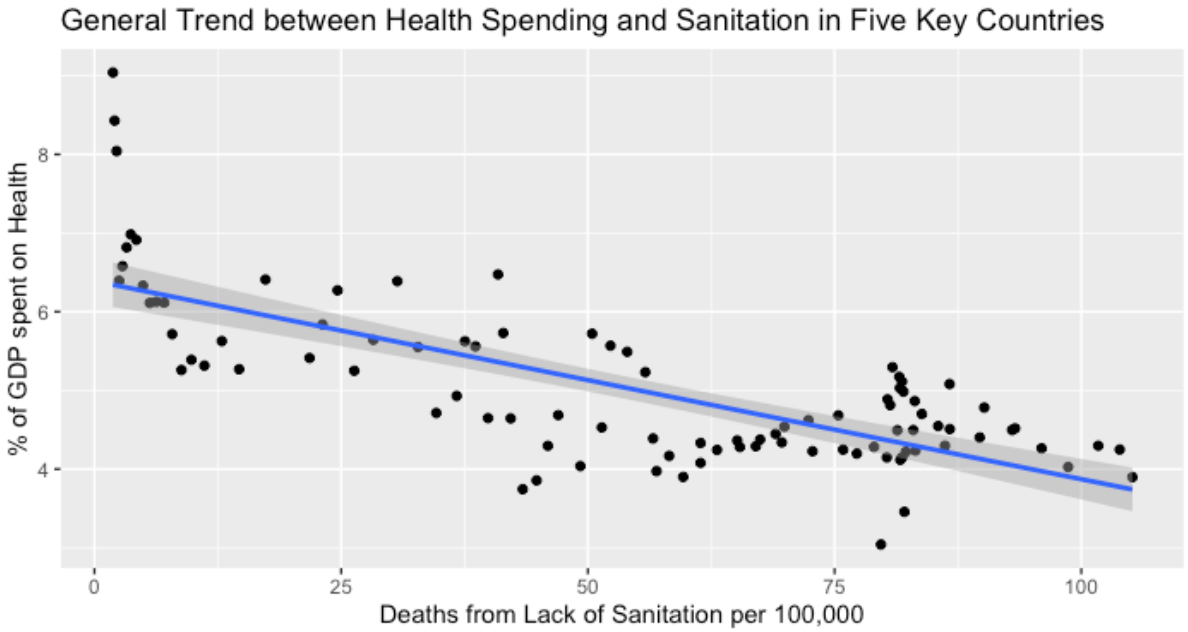
with little exercise. In lower income countries, fewer people work in a repetitive environment or can afford a car. This means that they have to walk more. As GDP per capita increases, the percentage of deaths from musculoskeletal diseases increases substantially. However, even in the most extreme cases, musculoskeletal diseases only account for about 13% of the deaths as opposed to the 35% from diarrhea, despite being more difficult to treat. This may be due to the fact that people in high income countries are far less likely to die from easily treatable diseases and disorders.

3.5 Does spending on health improve life expectancy?



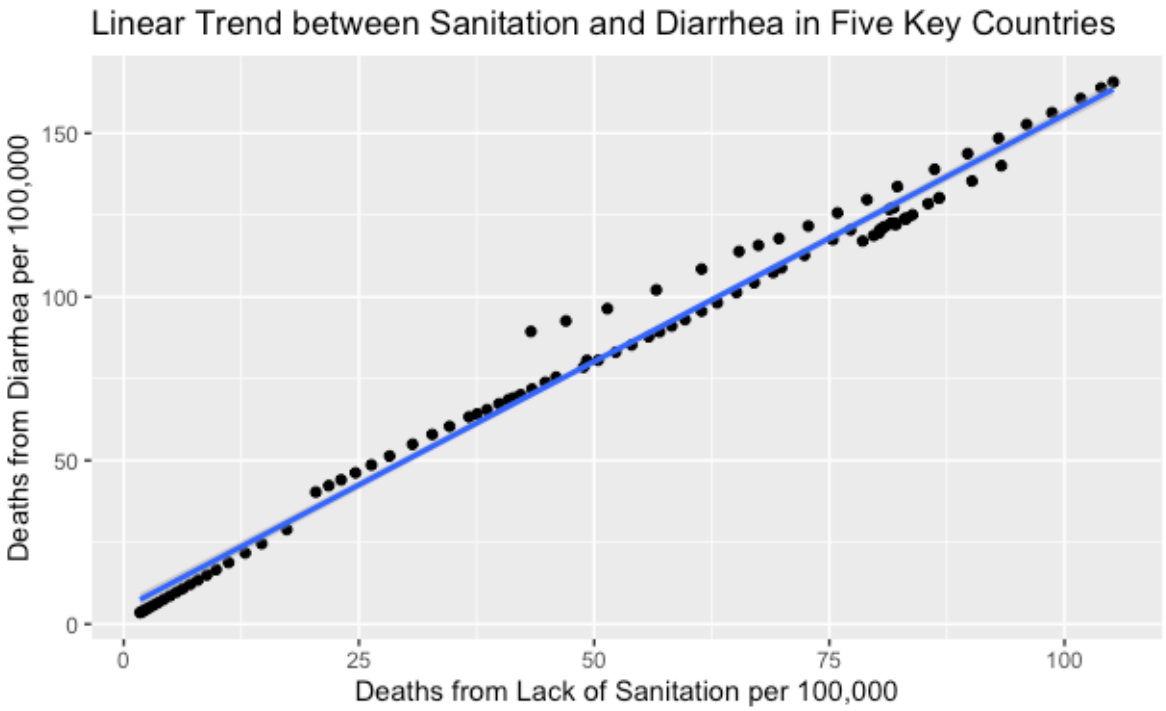
Upon confirming that economic factors like GDP per Capita influence life expectancy, we explored potential explanations for this correlation. One variable we explored was the percentage of a country's GDP spent on health. This variable seemed to have explanatory power for these observations because increased spending on health would ideally improve medical care for the population. Specifically, we investigated the relationship between health expenditure as a percentage of GDP and life expectancy. We observed that correlation depends on the overall GDP of the country itself as shown above. When the GDP of a country is greater, spending on health has a greater impact on improving life expectancy. When creating a linear model between the two variables, we found that a 3% increase in health expenditure correlates a 20 year increase in life expectancy. While not providing exact causation, this simple correlation demonstrates the importance of wealth in providing the resources needed to keep a population healthy.

3.6 Does spending on health improve sanitation levels?



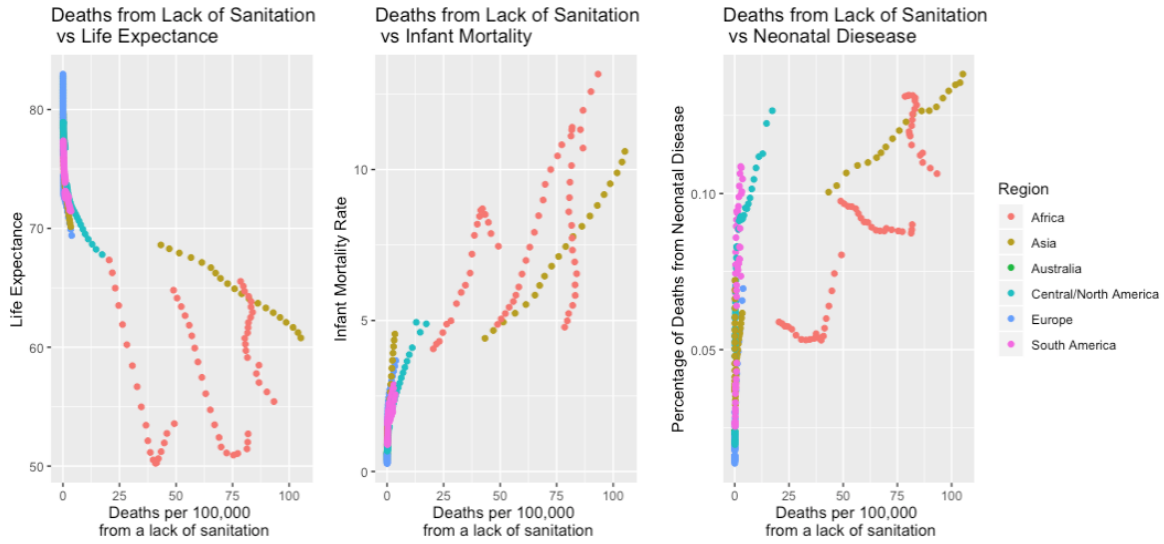
At this point, we wanted to learn more about why spending on health improves life expectancy. Since this question is rather broad and nuanced, we decided to focus on communicable diseases since we had gathered several variables that showed promising correlations. Furthermore, we temporarily narrowed our attention to the top five countries that suffered from communicable diseases: Botswana, India, Kenya, Nicaragua and Madagascar. As demonstrated in the graph above, we observed immediately that greater expenditures on health lead to a decline in the number of deaths due to lack of sanitation. Individually, each country's spending on health had varying levels of impact on the number of deaths from lack of sanitation. A preliminary guess leads us to believe that much of this relationship depends on the size of the population. For example, spending on health in India and Kenya has a much weaker benefit than spending in Nicaragua and Botswana. Theoretically, this relationship makes sense because larger populations seem to make healthcare harder to administer and therefore less efficient. In the end, these five countries are a small sample size and therefore only future studies of more countries in similar situations can highlight whether population has a genuine impact.

3.7 Does improved sanitation reduce deaths by communicable disease?



Since spending on health shows some correlation to sanitation levels, we wanted to tie this observation back into our original investigation into causes of death. Specifically, we wanted to find out what are the underlying causes between such a high rate of communicable disease related death in these four specific countries. After exploring the relationships among these five countries, we discovered there is 99.5% correlation between the number of deaths caused by sanitation per 100,000 people and the number of deaths caused by diarrhea per 100,000 people. This makes sense since diarrhea related deaths are communicable diseases and greater sanitation practices hinder the spread of these diseases. When plotting deaths from diarrhea vs. deaths from diarrhea in these countries, we observed a strong linear trend and found that the y-intercept of the best fit line is 4.847 while the slope is 1.508. Since the y-intercept is sufficiently close to zero, this implies that virtually all deaths from poor sanitation are diarrhea related and consequently about two-thirds of deaths caused by diarrhea are a result of poor sanitation practices.

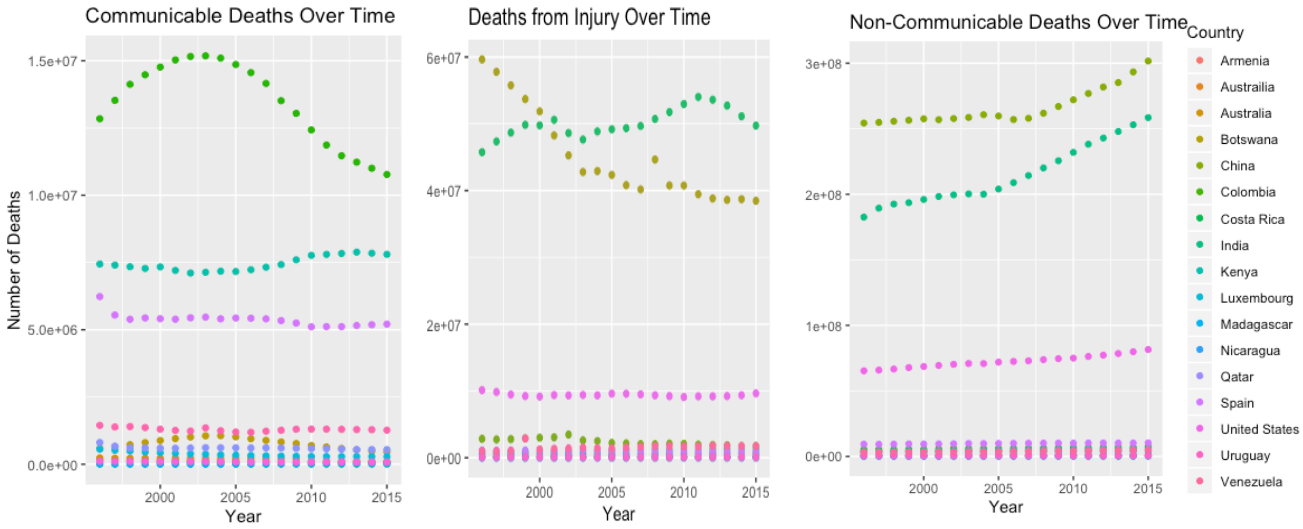
3.8 How does sanitation relate to the age distribution? Does region play a role in this?



Now that we established a solid connection between deaths from sanitation and deaths from diarrhea, we explored our dataset more to discover what are the problems that result due to a lack of sanitation and thus harm life expectancy levels. Figure A shows the relation between sanitation and the Infant Mortality Rate, life expectancy, and percentage of deaths from neonatal disease. Even at a glance, there seems to be correlation. The first graph details that life expectancy and the number of deaths from sanitation are inversely correlated. In regions where more people die from living in unsanitary conditions, the life expectancy is much lower. This is partially because the two groups with the highest potential for immunodeficiencies are the elderly and infants. If fewer elderly people are able to survive and more people are dying as infants, then the life expectancy is significantly lower. It is also evident that when the number of deaths per 100,000 exceeds 13 people, the life expectancy is always below 70. In the second graph, we see that countries in which there are more deaths from a lack of sanitation, more people tend to die as infants. Formally, the Infant Mortality Rate is defined as the number of deaths of children under the age of one compared to the number of live births. Since infants have very weak immune systems, being in a sanitary environment is vital. It is also clear that the highest rates are in Africa and Asia. Many of these countries in these regions lack the proper resources to establish and maintain a system for sanitation. On the third chart, we see the number of deaths from neonatal disease as a percentage of the total annual deaths for each country. While this graph is similar to the second, it details an important distinction. For countries that tend to be more unsanitary, not only are there higher Infant Mortality Rates, but also that infant deaths represent a far greater portion of the deaths in that country. In some extreme cases, neonatal disease represents nearly 13% of all deaths in a country. This means that for every ten people that die, one did not live to be a year old. We also see that past 50 deaths from sanitation for every 100,000 people, there is a clear spike in the percentage of deaths from neonatal disease. This all contributes to the fact that there is a relation between countries with high rates of neonatal

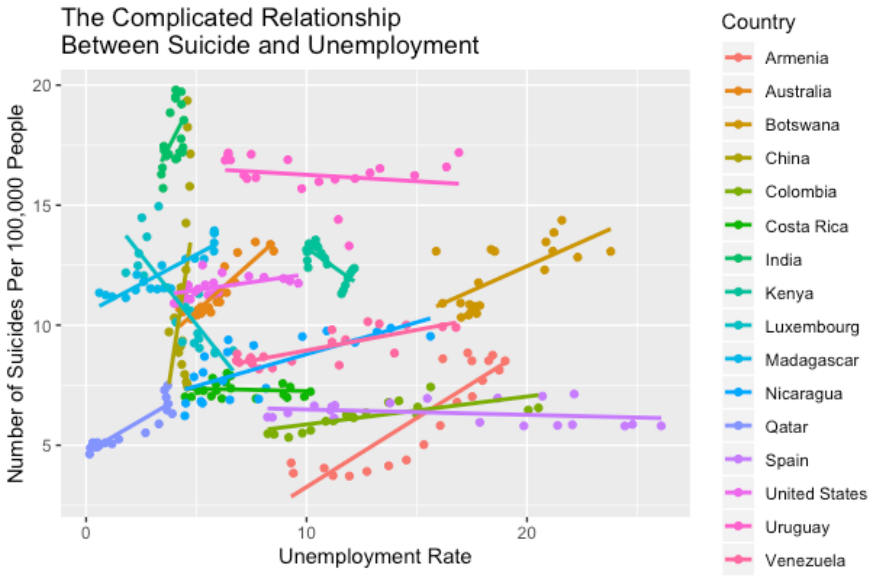
disease and countries with low life expectancy. If a tenth of the population dies before they turn one year old, then the life expectancy will obviously be much lower.

3.9 Which countries are outliers within the data? What are some potential causes?



In the graph displaying the changes in cause of death over time , Kenya and Botswana are both outliers. Their proportion of Communicable Diseases are noticeably higher than most other countries. This figure indicates that in these low-income African countries, communicable diseases such as HIV are prominent. This is due to Kenya and Botswana being in the region Africa where HIV/AIDS is prominent. There are also many other mosquito-spread communicable diseases that occured during this time frame such as malaria and West Nile virus. In the figure regarding life expectancy, the countries Kenya and Botswana are again outliers. This is probably due to low standards of sanitation and low child mortality rates every year in addition to communicable diseases. Both Kenya and Botswana bring the overall data averages down dramatically. In another graph showing the amount of non-communicable deaths over time, China and India are outliers. The reason that these countries are outliers is mainly due to their populations. Datasets that measure number of deaths and number of incidents will be skewed to China and India. In this figure, you can clearly see that both China and India don't align with the rest of the data. Their respective deaths due to non-communicable diseases such as cancer and heart disease are distinctively high. This is not due to their government health expenditure or any other factor as shown in the data. It is purely due to their large populations. Every other country measured has an average population relative to the global population. As can be seen, these countries lie toward the bottom of the graph and are basically non-distinguishable with the exception of the United States.

3.10 Is there a correlation between unemployment and suicide? Are there any limitations?



As a final tangent from our investigations above, we were interested in investigating was whether unemployment has an impact on suicide. We plotted the unemployment rate against the number of suicides per 100,000 people. The resulting graph is inconclusive because half of the countries show positive correlation between while the remaining countries have either negative or zero correlation. Additionally, we tried comparing suicide rates to economic measures such as GDP and GDP per Employed Person but unfortunately these did not reveal anything insightful. Since we are trying to find patterns across many different nations and cultures, we anticipate suicide is too complicated of an issue to simply to a simple association such as unemployment. However, the fact that half of our countries do exhibit some form of positive correlation implies that there is likely some common attribute we were missing. For future study, we anticipate that adding more countries beyond our meager sample of sixteen countries will reveal greater correlations between unemployment and suicide. In the end, the mixed results are unsurprising considering that official health organizations like the CDC list numerous factors that contribute to suicide rates such as family history and mental illness which cannot be easily reduced to nationwide measurements [8].

Conclusion

Upon going through multiple spreadsheets and aligning up the data to fit our report, we found many connections between several variables in our data. We were mainly trying to find relationships between a country's national economic status and the overall health of its' citizens. After building graphs and exploring the interrelations between variables, we were able to find a lot of interesting things. One of the strongest and most expected connections we found was the negative correlation between life expectancy and infant mortality rate. This is intuitive and

makes sense because as more children die before birth, the average expectancy of a country will fall. Another steady correlation we noticed was between government spending on health and deaths from levels of poor sanitation. Countries whose governments placed more of their GDP into healthcare had lower death rates from poor sanitation in addition to a lower number of deaths due to communicable diseases. We wanted to figure out which kinds of diseases were influenced by Income level and we found that GDP per capita has a correlation with both diarrhea and musculoskeletal diseases. As the GDP per capita increased, the rate of death due to diarrhea increased while the rate of death due to musculoskeletal diseases decreased. A potential reason for this is higher-income countries have well cooked, uncontaminated food which would indicate low levels of diarrhea, while these countries also have high obesity rates due to a plethora of food indicating higher levels of musculoskeletal diseases. Something that we found that had little to no correlation was the relationship between unemployment rates between each country and time. This is something that we expected to see a correlation because GDP per capita increased steadily for almost all countries over time. In addition, more and more jobs are opening as the use of technology increases and more countries are urbanizing which would theoretically bring about the decrease of unemployment.

The data we gathered was rather small compared to the many other variables that could have potentially been factored in. Despite this, we still managed to come to some fair and concise conclusions. If we were to continue and go deeper with this study, we would collect data from more than just 3 countries per region and this would allow us to have more concrete evidence. We could also try and predict how the future would look like if the trends we analyzed continued. Many more interesting findings could be made with this topic by incorporating more categorical data and exploring the connections between things like language, religion, driving side, and demographics.

Appendix

[1] https://www.who.int/gho/database/en/
[2] https://data.worldbank.org/
[3] https://ourworldindata.org/
[4] https://www.chicagotribune.com/news/ct-xpm-1999-12-22-9912220106-story.html
[5] https://www.cfr.org/backgrounder/colombias-civil-conflict
[6] https://www.bbc.com/news/world-latin-america-19652436
[7] https://www.aljazeera.com/indepth/features/2017/06/timeline-qatar-gcc-disputes-170605110356982.html
[8] https://www.cdc.gov/violenceprevention/suicide/riskprotectivefactors.html
[9] https://www.populationpyramid.net/world/2020/

Code

3.1 Plot

```
#Life Expectancy by Country
#Compared by Continent
ggplot(data = stat) + geom_boxplot(mapping = aes(x = Country, y = exp, color = Region)) +
  labs(title = 'Life Expectancy by Country', x = 'Country',
       y = 'Life Expectancy') + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

3.2 Plot

```
top_1 <- melt(data = stat, id.vars = "Year",
          measure.vars = c("comm_perc", "non_comm_perc", "injury_perc"))

topc_1 <-rep(rep(c("Armenia", "Australia", "Botswana", "China",
                "Columbia", "Costa Rica", "India", "Kenya",
                "Luxembourg", "Madagascar", "Nicaragua", "Qatar",
                "Spain", "United States", "Uruguay", "Venezuela"), each = 20), times = 3)
top2_1 <- cbind(top_1, topc_1)

# Plots the data into a four by four grid using facet_wrap()
# To make the colors match the three broad causes, scale_fill_manual() corrects the labels
ggplot(data = top2_1) +
  geom_bar(mapping = aes(x = Year, y = value, fill = variable), stat = "identity") +
  facet_wrap(topc _1~ . ) + labs(x = "Year",
                        y = "Percentage",
                        fill = "Cause of Death",
                        title = "Changes in Cause of Death over Time") +
  scale_fill_manual(labels = c("Communicable disease",
```

```
                "Non-communicable disease", "Injuries"),
        values = c("comm_perc" = "red",
                   "non_comm_perc" = "blue",
                   "injury_perc" = "green"))
```

3.3 Plot

```
#Deaths from Top 5 causes by country
top <- melt(data = stat, id.vars = "Year", measure.vars = c("cardio", "cancer", "diarrhea", "neonatal", "musculo"))
topc <-rep(rep(c("Armenia","Australia", "Botswana", "China",
        "Columbia", "Costa Rica", "India", "Kenya",
        "Luxembourg", "Madagascar", "Nicaragua", "Qatar",
        "Spain", "United States", "Uruguay", "Venezuela"), each = 20), times = 5)
top2 <- cbind(top, topc)

#TOP 5 CAUSES
#Percentage of Each of Top 5 Causes of Death by Country
ggplot(data = top2) +
  geom_bar(mapping = aes(x = Year, y = value, fill = variable),stat="identity", position = position_fill()) +
  facet_wrap(topc~.) +
  labs(title = 'Percentage of Each of Top 5 Causes of Death', x = 'Year',
      y = 'Percent', fill = "Cause")
```

3.4 Plot

```
#Death from Musculoskeletal Disease Compared to GDP per Capita
#Comparing by Income Level
ggplot(data = stat) + geom_point(mapping = aes(x = (musculo / total), y = GDP_cap, color = inc_level)) +
  labs(title = 'Percentage of Musculoskeletal \n vs GDP per Capita', x = 'Percentage of Deaths \n from
Musculoskeletal Disease',
      y = 'GDP per Capita', color = "Income Level")

#Death from Diarrhea Compared to GDP per Capita
#Comparing by Income Level
ggplot(data = stat) + geom_point(mapping = aes(x = (diarrhea/total), y = GDP_cap, color = inc_level)) +
  labs(title = 'Percentage of Diarrhea \n vs GDP per Capita', x = 'Percentage of Deaths \n from Diarrhea',
      y = 'GDP per Capita')
```

3.5 Plot

```
# Turn the income level into an ordered factor group
stat$income <- factor(stat$income,
        levels = c("Low", "Middle", "High"),
        labels = c("Low", "Middle", "High"))
```

```
# Floor the life expectancy variable down to the nearest multiple of 10
# And create a boxplot that is faceted by income.
ggplot(data=stat,
      mapping = aes(x = factor(floor(exp / 10) * 10), y = spending, color = income)) +
geom_boxplot() +
facet_wrap(. ~ income) +
  labs(x = "Life Expectancy", y = "% of GDP Spent on Health", color = "GDP Level",
      title = "Increased Spending on Health Improves Life Expectancy")

# Find the coefficients for a linear model between life expectancy and spending on health
linearMod <- lm(exp ~ spending, data = stat)
print(linearMod)
```

3.6 Plot

```
# Plots the relationship between spending on health and
# deaths from sanitation
sanitation_countries <- c("India", "Kenya", "Botswana", "Nicaragua", "Madagascar")
ggplot(data = stat[which(stat$country %in% sanitation_countries), ],
      mapping = aes(x = sanitation, y = spending, color = country)) +
  geom_point() + stat_smooth(method = "lm") +
  labs(x = "Deaths from Lack of Sanitation per 100,000",
     y = "% of GDP spent on Health",
     title = "General Trend between Health Spending and Sanitation in Five Key Countries")
```

3.7 Plot

```
# Plots relationship between deaths from poor sanitation
# and deaths from diarrhea
ggplot(data = stat[which(stat$country %in% sanitation_countries), ],
      mapping = aes(x = sanitation,y = diarrhea_hund)) +
  geom_point() + stat_smooth(method = "lm") +
  labs(x = "Deaths from Lack of Sanitation per 100,000",
      y = "Deaths from Diarrhea per 100,000",
      title = "Linear Trend between Sanitation and Diarrhea in Five Key Countries")

x_1 <- stat[which(stat$country %in% sanitation_countries), ]$sanitation
x_2 <- stat[which(stat$country %in% sanitation_countries), ]$spending
cor(x_1[!is.na(x_2)], x_2[!is.na(x_2)])

linearMod <- lm(sanitation ~ spending,
                    data=stat[which(stat$country %in% sanitation_countries), ])
```

3.8 Plot

```
#Deaths from Lack of Sanitation Compared to Life Expectancy
#Compared by Continent
ggplot(data = stat) + geom_point(mapping = aes(x = sanitation, y = exp, color = Region)) +
  labs(title = 'Deaths from Lack of Sanitation \n vs Life Expectancy', x = 'Deaths per 100,000 \n from a lack of
sanitation',
       y = 'Life Expectancy')

#Deaths from Lack of Sanitation Compared to Infant Mortality Rate
#Compared by Continent
ggplot(data = stat) + geom_point(mapping = aes(x = sanitation, y = inf_mort, color = Region)) +
  labs(title = 'Deaths from Lack of Sanitation \n vs Infant Mortality', x = 'Deaths per 100,000 \n from a lack of
sanitation',
       y = 'Infant Mortality Rate')

#Deaths from Lack of Sanitation Compared to Deaths from Neonatal Disease
#Compared by Continent
ggplot(data = stat) + geom_point(mapping = aes(x = sanitation, y = (neonatal/total), color = Region)) +
  labs(title = 'Deaths from Lack of Sanitation \n vs Neonatal Disease', x = 'Deaths per 100,000 \n from a lack of
sanitation',
       y = 'Percentage of Deaths from Neonatal Disease')
```

3.9 Plot

```
#COMMUNICABLE, NON-COMMUNICABLE, AND INJURY
#Communicable Deaths over time
ggplot(data = stat) + geom_point(mapping = aes(x = Year, y = comm, color = Country)) +
  labs(title = 'Communicable Deaths Over Time', x = 'Year', y = 'Number of Deaths')
#Non-Communicable Deaths over time
ggplot(data = stat) + geom_point(mapping = aes(x = Year, y = non_comm, color = Country)) +
  labs(title = 'Non-Communicable Deaths Over Time', x = 'Year', y = 'Number of Deaths')
#Deaths from Injuries over time
ggplot(data = stat) + geom_point(mapping = aes(x = Year, y = injury, color = Country)) +
  labs(title = 'Deaths from Injury Over Time', x = 'Year', y = 'Number of Deaths')
```

3.10 Plot

```
# Plots deaths from suicide and unemployment rates
ggplot(data = stat,
       mapping = aes(y = suicide, x = unemployment, color = country)) +
  geom_point() + stat_smooth(method="lm", se=FALSE) +
  labs(x = "Unemployment Rate", y = "Number of Suicides Per 100,000 People",
       color = "Country", title = "The Complicated Relationship \nBetween Suicide and Unemployment")
```