Analysis on Chronic Obstructive Pulmonary Disease Deaths, Percentage Spent on Health Care based on Gross Domestic Product (GDP), and Particulate Matter 2.5 Concentration Levels by Country

KEERTI SUNDARAM, Rensselaer Polytechnic Institute, USA

While there are many non-communicable diseases impacting countries today, one of the most prevalent is Chronic Obstructive Pulmonary Disease (COPD). COPD is the third leading cause of death worldwide (WHO). Chronic Obstructive Pulmonary Disease refers to a group of diseases that cause airflow blockage and breathing-related problems. It is typically caused by long-term exposure to irritating gases or particulate matter such as particulate matter 2.5 (PM2.5). According to the World Health Organization, low- and middle-income countries are disproportionately affected by COPD, with nearly 90% of COPD deaths in those under 70 occurring in these countries. The goal of the following analysis is to determine whether government spending on healthcare and pollution levels contribute significantly to the number of COPD deaths.

1 INTRODUCTION

According to the World Health Organization (WHO), COPD, a common lung disease caused 3.23 million deaths in 2019. Therefore, it is important to analyze how different factors contribute towards COPD deaths. While the most common causes of COPD are smoking and air pollution, this study aims to look at how access to health care impacts COPD fatalities.

The goal of health equity is to ensure that across the world, everyone has access to necessary and quality health care. As per WHO, "health equity is a priority in the post-2015 sustainable development agenda and other major health initiatives" (WHO). Health financing is an important aspect to health equity and can offer insight into the efficacy and influence of government spending on health care. Although the goal of this analysis is to observe the effects of access to health care on COPD deaths, it is important that the other contributing factors of COPD are considered.

The following study will observe two hypotheses:
i. Countries with a higher percent of GDP spent on health will have lower COPD deaths.
ii. Air pollution, quantified here by PM2.5, is also a contributing factor that accounts for an increased amount of COPD deaths despite a high percentage of GDP spent on health care.

Additional Key Words and Phrases: healthcare, gross domestic product, chronic obstructive pulmonary disease, health equity, pm2.5

ACM Reference Format:
Keerti Sundaram. 2023. Analysis on Chronic Obstructive Pulmonary Disease Deaths, Percentage Spent on Health Care based on Gross Domestic Product (GDP), and Particulate Matter 2.5 Concentration Levels by Country. 1, 1 (April 2023), 30 pages. https://doi.org/XXXXXXX.XXXXXXX

Author's address: Keerti Sundaram, sundak3@rpi.edu, Rensselaer Polytechnic Institute, 110 8th St, Troy, New York, USA, 12180.

---

2 THE DATA AND EXPLORATORY DATA ANALYSIS

Three World Health Organization data sets are utilized to perform the analysis:

(1) Deaths by sex and age group for a selected country or area and year for chronic obstructive pulmonary disease
(2) Domestic general government health expenditure (GGHE-D) as percentage of gross domestic product (GDP) (%)
(3) Concentrations of Fine Particulate Matter (PM2.5)

For simplicity, the data sets will be referred to as COPD, GDP, and PM2.5 respectively. Initial data cleaning, which will be elaborated on further in section 3, resulted in data for 104 countries. This includes COPD and GDP data for 21 years and PM2.5 data for 10 years. Each of the data sets were chosen for specific purposes. The COPD data allows one to determine how many individuals died due to COPD based on country and year. The GDP data will be used to observe how access to healthcare impacts COPD deaths and offers insight into the efficacy and influence of government spending on health care. As mentioned previously, particulate matter, such as PM2.5, is a significant risk factor for COPD. Particulate matter 2.5, an air pollutant, refers to tiny particles or droplets in the air that are two and one half microns or less in width. These particles are able to travel deep into the respiratory tract, reaching the lungs (NY State Department of Health). To analyze how pollution levels, represented here by particulate matter concentrations, contributes to countries with higher COPD cases, data on particulate matter 2.5 (PM2.5) by country will be utilized.

2.1 COPD Data

The following code and figures are for the exploratory data analysis done on the COPD data set.
R Code:

```
copd <- read.csv("/Users/keertisundaram/Dropbox/Data Analytics/
mortality_csv.csv", header=T)
str(copd)
summary(copd)
#find number of NAs
sum(is.na(copd))
#filtering out country labels and grand totals for boxplots
no_labels_copd <- copd[-1]
boxplot(no_labels_copd[-length(no_labels_copd)], main = "COPD
cases of all Countries over 2000-2020")
hist(copd$Grand.Total, main = "Grand Total Deaths due COPD")
```

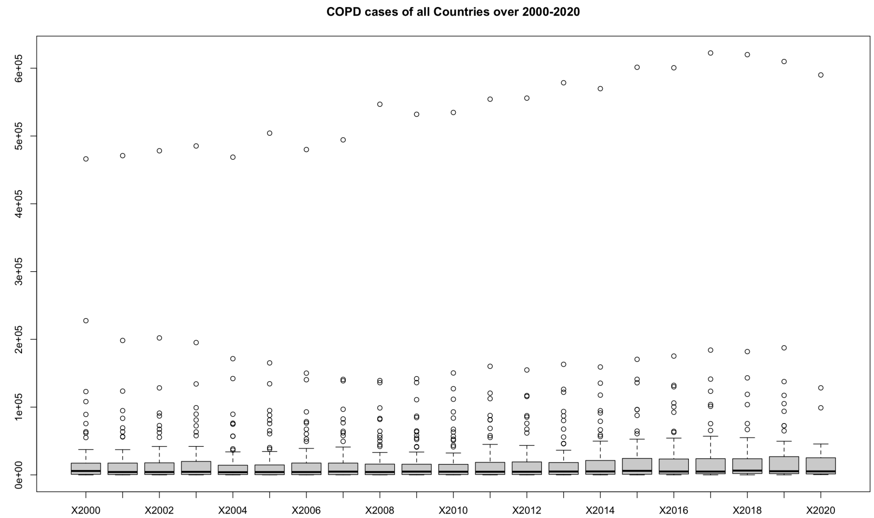


Fig. 1. COPD all Countries Boxplot

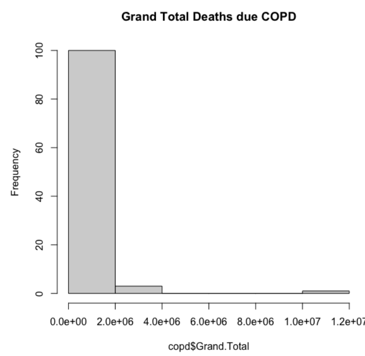


Fig. 2. COPD Histogram

2.2 GDP Data

The following code and figures are for the exploratory data analysis done on the GDP data set.

R Code:

```
gdp <- read.csv("/Users/keertisundaram/Dropbox/Data Analytics/gdp_csv.csv", header=T)
str(gdp)
summary(gdp)
#find number of NAs
sum(is.na(gdp))
no_labels_gdp <- gdp[-1]
```

```
boxplot(no_labels_gdp[-length(no_labels_gdp)], main = "Percent of GDP Spent on Health of
all Countries over 2000-2020")
hist(gdp$Grand.Total, main = "Grand Total Percent of GDP Spent on Health")
```

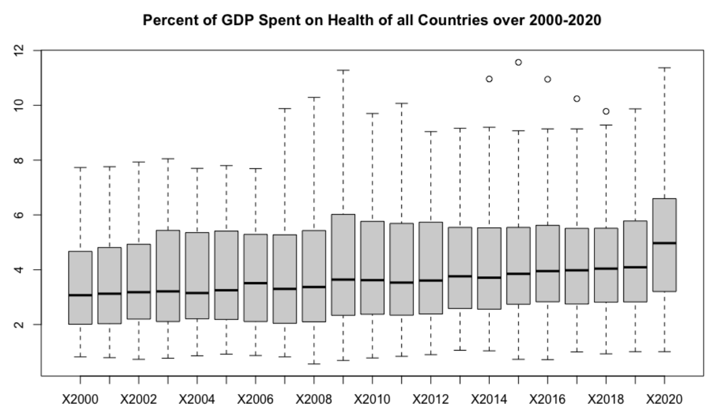


Fig. 3. GDP all Countries Boxplot

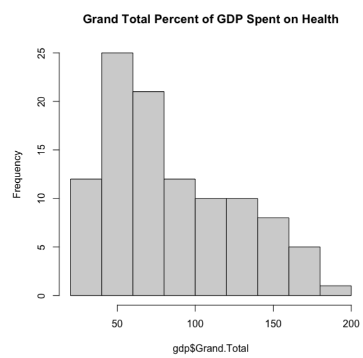


Fig. 4. GDP Histogram

2.3 PM2.5 Data

The following code and figures are for the exploratory data analysis done on the PM2.5 data set.
R code:

```
pm2.5 <- read.csv("/Users/keertisundaram/Dropbox/Data Analytics/Final_Project/pm2.5.csv", header=T)
str(pm2.5)
summary(pm2.5)
#adding a grand total column to maintain same structure as copd and gdp datasets
```

```
grand_total <- rowSums(pm2.5[-1])
pm2.5$Grand.Total <- grand_total
#find number of NAs
sum(is.na(pm2.5))
no_labels_pm2.5 <- pm2.5[-1]
boxplot(no_labels_pm2.5[-length(no_labels_pm2.5)], main = "PM2.5 Concentrations of all
Countries over 2010-2019")
hist(pm2.5$Grand.Total, main = "Grand Total PM2.5 Concentrations")
```

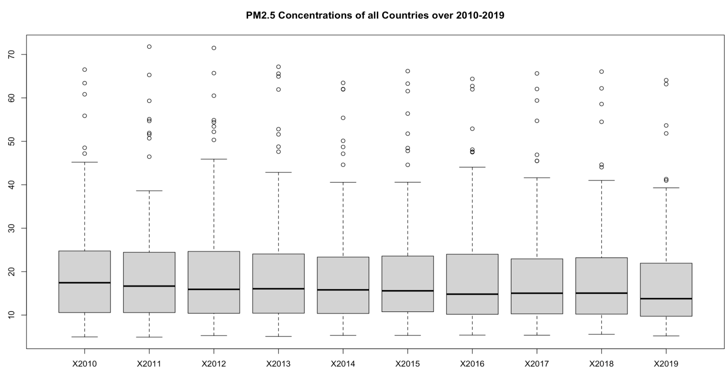


Fig. 5. PM2.5 Concentrations of all Countries Boxplot

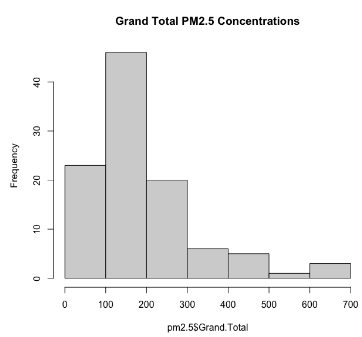


Fig. 6. PM2.5 Histogram

2.4 Interpretation

For readability the outputs for the str and summary function calls are not included. However, each of the data frames are set up the same way, with one country column, the years columns, and the grand total column. Additionally, analysis identified that the COPD data contains 418 NA values, the GDP data contains 22 NA values, and the PM2.5 data contains 0 NA values. For each of the data sets, a boxplot was created for each year, meaning each box is representative of all 104 countries for that year. A histogram was also created for each data set based off the grand totals for each country (i.e. sum of numbers over the years for each country). Based on a visual analysis of the figures, many of the countries have a similar amount of total COPD cases, however, on the boxplot, there are outliers (countries with higher cases). The GDP data has more variance than the COPD data and there are fewer outliers that are revealed by the boxplot. Lastly, the PM2.5 data also has more variance than the COPD data when comparing the histograms. Based on the boxplots, the boxes are bigger than compared to the COPD and smaller than the GDP boxes. This means that there is more variance (i.e. larger ranges between the 1st and 3rd quartiles) between the PM2.5 data than the COPD but not the GDP data.

3 ANALYSIS OF DATA AND PRE PROCESSING

Each of the data sets needed a significant amount of formatting to create uniform rows and columns. Several Excel pivot tables were used to format the raw data into the current format. The COPD raw data was formatted by age, sex, country, and year. The data was then formatted as the sum of all deaths over a year for each country. The GDP data was formatted by country and year but required reformatting of the columns and rows to ensure the format matched the COPD data. Similarly, the PM2.5 data had to be organized by country and year but required the columns to be aggregated due to the data being split up by location (i.e. initially split up into Urban, Rural, and Cities and then combined into one value). Additionally, the data was filtered to ensure the analysis only focused on countries with available data, this was done by matching country names. It is important to note that the GDP data was filtered based on the years included in the COPD data (2000-2020), but the PM2.5 data is from years (2010-2019). This was done to ensure the cluster analyses using the GDP and COPD data would have as much data as possible. Whereas the regression analysis on the COPD and PM2.5 data was limited to the years of 2010-2019. As revealed in the exploratory data analysis, both the COPD and GDP data had missing values. To move forward with the analysis, it was decided that data imputation would be used to fill in missing values. The nearest-neighbor algorithm in the VIM package was utilized to accomplish this. The grand total column was also updated based on the imputed values.

Below is the code for the data imputation performed.

```
R code:
#data imputation, nearest-neighbor imputation
library(VIM)
#copd data
copd_imp <- kNN(copd[-length(copd)], k=5)
summary(copd_imp)
#removing extra imputation columns
copd_imp <- subset(copd_imp, select=Country:X2020)
#updating grand total column
grand_total <- rowSums(copd_imp[-1])
```

```
copd_imp$Grand.Total <- grand_total

#gdp data
gdp_imp <- kNN(gdp[-length(gdp)], k=5)
#removing extra imputation columns
gdp_imp <- subset(gdp_imp, select=Country:X2020)
#updating grand total column
grand_total <- rowSums(gdp_imp[-1])
gdp_imp$Grand.Total <- grand_total
```

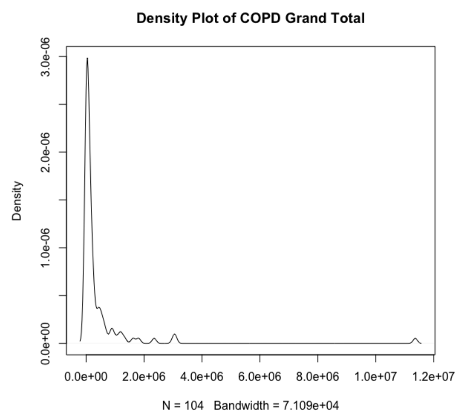
While data imputation was decided as the best route for this application, it is possible it may lead to a possible source of error. The data is not exact and is based on the closest values, therefore it creates limitations on the analysis.

As a next step density plots were created using the data with the imputed values. Density plots may offer better insight into the distribution that histograms with the default bin width cannot.

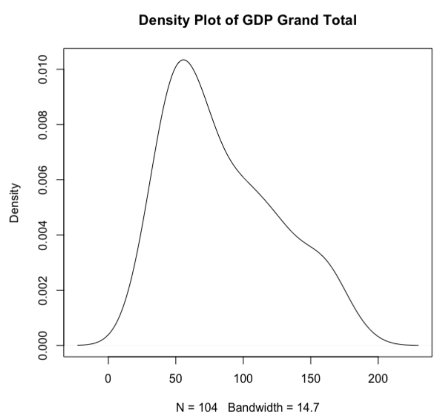
Below is the code used to create the density plots for each of the data sets. It is important to note that the imputed COPD and GDP data frames were utilized to create the density plots.

R code:

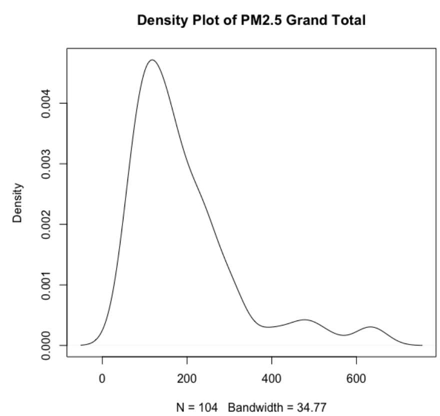
```
plot(density(copd_imp$Grand.Total), main="Density Plot of COPD Grand Total")
plot(density(gdp_imp$Grand.Total), main= "Density Plot of GDP Grand Total")
plot(density(pm2.5$Grand.Total), main= "Density Plot of PM2.5 Grand Total")
```



(a) Density Plot of COPD



(b) Density Plot of GDP



(c) Density Plot of PM2.5

Fig. 7. Density Plots

Based on a visual analysis of the density plots, all three of the data sets follow different distributions. The COPD data appears to follow an exponential distribution, whereas the GDP data appears to follow a weibull distribution, and the PM2.5 data appears to follow an F distribution. It is important to note however that the visual analysis of these plots is limited and may not be accurate in categorizing the actual distribution.

4 MODEL DEVELOPMENT AND APPLICATION

4.1 K-Means Clustering

K-Means Clustering was performed on both the COPD and GDP data sets. The goal of clustering was to determine which countries are grouped together for both data sets. Clustering was initially done on the entire data set. To determine the number of groups, or the value of k, elbow plots were created. Based on these figures, k was set to 4 for both data sets. Below is the code used to create the elbow plots.

```
R code:
library(factoextra)
fviz_nbclust(copd_imp[-1], kmeans, method="wss")
fviz_nbclust(gdp_imp[-1], kmeans, method="wss")
```

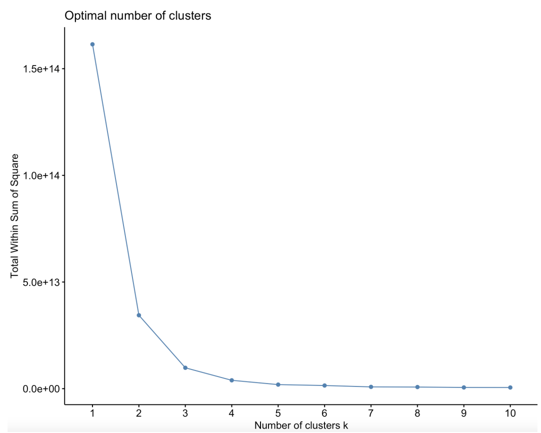


Fig. 8. COPD Elbow Plot

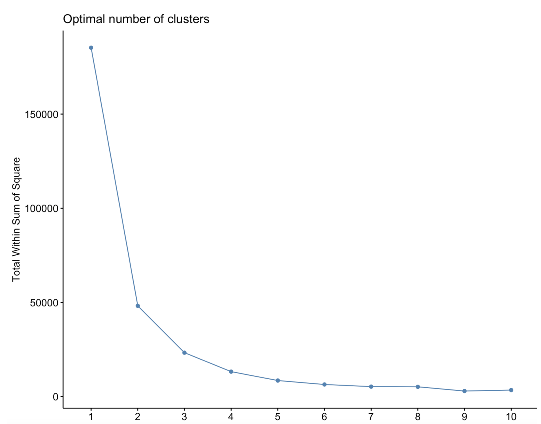
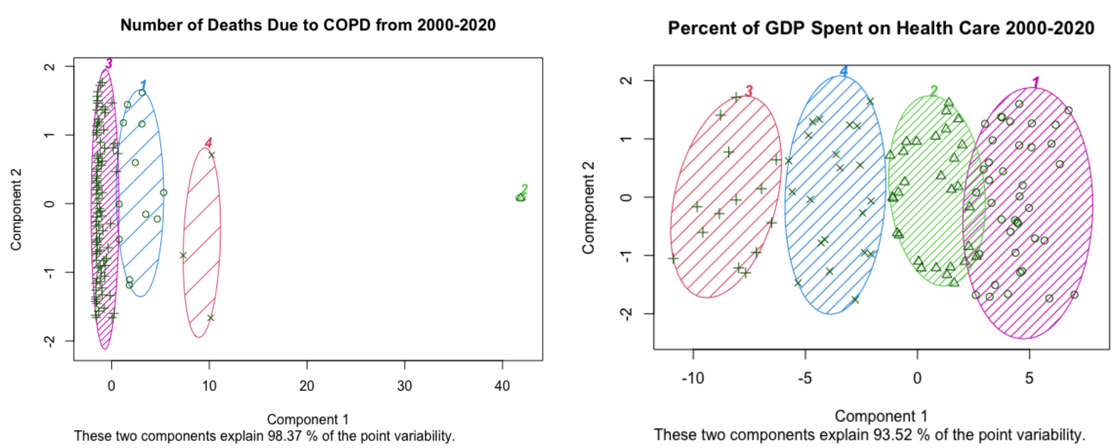


Fig. 9. GDP Elbow Plot

The results from the initial clustering is shown in figure 10, 11, and 12.
Below is the code utilized to perform K-means clustering and plot the clusters.
R Code:

```
library(ISLR)
set.seed(101)
library(cluster)
totalClusters <- kmeans(copd_imp[-1], 4, nstart = 20)
#nstart is the number of random start
print(totalClusters$cluster)
clusplot(copd_imp,totalClusters$cluster, color = TRUE,
shade = TRUE, labels = 4, lines = 0,  main = "Number of Deaths Due
to COPD from 2000-2020")
#gdp
totalClusters <- kmeans(gdp_imp[-1], 4, nstart = 20)
print(totalClusters$cluster)
clusplot(gdp_imp,totalClusters$cluster, color = TRUE,
shade = TRUE, labels = 4, lines = 0, main = "Percent of GDP Spent on
Health Care 2000-2020")
```



(a) COPD clusters (b) GDP clusters

Fig. 10. Cluster Plots

Cluster 1	Canada, Colombia, France, Italy, Japan, Kazakhstan, Mexico, Philippines, South Africa, Spain, Turkey Ukraine
Cluster 2	United States of America
Cluster 3	Albania, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahamas, Bahrain, Barbados, Belarus, Belgium, Belize, Bosnia and Herzegovina, Brunei Darussalam, Bulgaria, Cabo Verde, Chile, "China, Hong Kong SAR", Costa Rica, Croatia, Cuba, Cyprus, Czechia, Denmark, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, Estonia, Fiji, Finland, Georgia, Greece, Grenada, Guatemala, Guyana, Hungary, Iceland, Iran (Islamic Republic of), Iraq, Ireland, Israel, Jamaica, Jordan, Kuwait, Kyrgyzstan, Latvia, Lebanon, Lithuania, Luxembourg, Maldives, Malta, Mauritius, Mongolia, Montenegro, Netherlands, New Zealand, Nicaragua, Norway, Panama, Paraguay, Peru, Poland, Portugal, Republic of Korea, Republic of Moldova, Romania, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Serbia, Seychelles, Singapore, Slovakia, Slovenia, Sri Lanka, Suriname, Sweden, Switzerland, Syrian Arab Republic, Tajikistan, Thailand, Trinidad and Tobago, Turkmenistan, Uruguay, Uzbekistan
Cluster 4	Brazil, Germany, Russian Federation

Fig. 11. COPD Clusters

Cluster 1	Albania, Antigua and Barbuda, Armenia, Azerbaijan, Bahamas, Bahrain, Brunei Darussalam, China, Cyprus, Dominican Republic, Egypt, Fiji, Georgia, Grenada, Guatemala, Guyana, Iran (Islamic Republic of), Iraq, Kazakhstan, Kuwait, Kyrgyzstan, Mauritius, Mexico, Mongolia, Paraguay, Peru, Philippines, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Singapore, Sri Lanka, Suriname, Syrian Arab Republic, Tajikistan, Thailand, Trinidad and Tobago, Turkmenistan, Uzbekistan
Cluster 2	Barbados, Belarus, Belize, Brazil, Bulgaria, Cabo Verde, Chile, Dominica, Ecuador, El Salvador, Estonia, Israel, Jamaica, Jordan, Latvia, Lebanon, Lithuania, Maldives, Nicaragua, Panama, Poland, Republic of Korea, Republic of Moldova, Romania, Russian Federation, Seychelles, South Africa, Switzerland, Turkiye, Ukraine
Cluster 3	Austria, Belgium, Canada, Cuba, Denmark, Finland, France, Germany, Iceland, Japan, New Zealand, Norway, Sweden, United States of America
Cluster 4	Argentina, Australia, Bosnia and Herzegovina, Colombia, Costa Rica, Croatia, Czechia, Greece, Hungary, Ireland, Italy, Luxembourg, Malta, Montenegro, Netherlands, Portugal, Serbia, Slovakia, Slovenia, Spain, Uruguay

Fig. 12. GDP Clusters

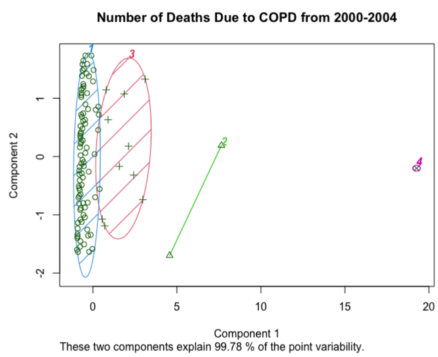
Based on the results of the initial clustering (shown in figures 10, 11, and 12), a decision was made to investigate the COPD clusters further. Interestingly, most of the countries are contained in one cluster. Since this clustering was done for all of the years (2000-2020), it was decided that further analysis would involve splitting up the years into groups to perform additional clustering. The goal was to determine if there were countries that had a different clustering pattern throughout the years. In other words, this clustering was performed to find countries that "changed" clusters at some point throughout the years.

Below is the code executed to split the data frame into four groups and then perform K-means clustering on each group. Figure 13 displays the plots that resulted from clustering.

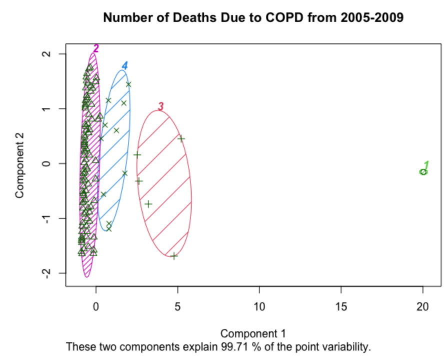
R code:

```
group1<- data.frame(copd_imp$Country, copd_imp[,c(2:6)])
group2<- data.frame(copd_imp$Country, copd_imp[,c(7:11)])
group3<- data.frame(copd_imp$Country, copd_imp[,c(12:16)])
group4<- data.frame(copd_imp$Country, copd_imp[,c(17:22)])

group1Clusters <- kmeans(group1[-1], 4, nstart = 20)
clusplot(group1,group1Clusters$cluster, color = TRUE, shade = TRUE,
labels = 4, lines = 0, main = "Number of Deaths Due to COPD from 2000-2004")
```

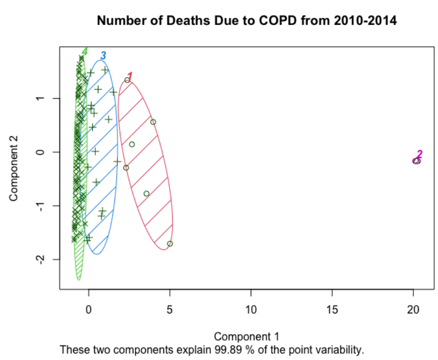
```
group2Clusters <- kmeans(group2[-1], 4, nstart = 20)
clusplot(group2,group2Clusters$cluster, color = TRUE, shade = TRUE,
labels = 4, lines = 0, main = "Number of Deaths Due to COPD from 2005-2009")

group3Clusters <- kmeans(group3[-1], 4, nstart = 20)
clusplot(group3,group3Clusters$cluster, color = TRUE, shade = TRUE,
labels = 4, lines = 0, main = "Number of Deaths Due to COPD from 2010-2014")

group4Clusters <- kmeans(group4[-1], 4, nstart = 20)
clusplot(group4,group4Clusters$cluster, color = TRUE, shade = TRUE,
labels = 4, lines = 0, main = "Number of Deaths Due to COPD from 2015-2020")
```



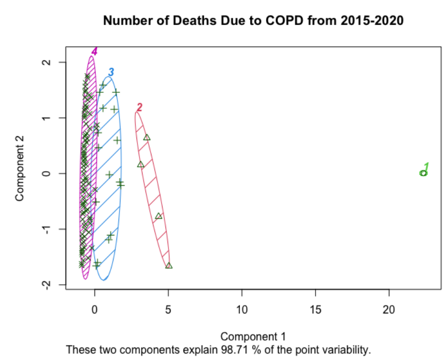
(a) COPD 2000-2004



(b) COPD 2005-2009



(a) COPD 2010-2014



(b) COPD 2015-2020

Fig. 13. COPD Clustering in Groups

While many of the countries remained in similar clusters (i.e. grouped with the same countries throughout the years), several countries switched their groupings at some point throughout the years. These countries were: Argentina, Australia, France, Germany, Hungary, Italy, Kazakhstan, Mexico, Netherlands, Poland, Republic of Korea, Romania, Thailand, and Turkey. Since there was not a large difference between splitting the countries into groups compared to using all of the data at once, it was decided to move forward with the analysis using the clustering that utilized the entire data set.

As mentioned, hypothesis i was that countries with a higher percent of GDP spent on health will have lower COPD deaths. To analyze the clustering in relation to the hypothesis, the groupings of the countries were observed. Assuming that the hypothesis is true, the groupings of both clusterings (i.e. COPD and GDP clustering) would be similar if not the same. In other words, countries that spent a higher percent of their GDP on health care would be clustered together in the GDP clustering. These same countries would be grouped together in the COPD clustering, more specifically would be grouped as countries with low COPD deaths. When comparing the clusters, many of the countries are grouped similarly (i.e. grouped with the same countries). The countries that are not: Brazil, Canada, Colombia, France, Germany, Italy, Japan, Kazakhstan, Mexico, Philippines, Russian Federation, South Africa, Spain, Turkey, Ukraine, United States of America. These countries, the ones that were not grouped similarly, will be utilized in the next regression analysis in regards to the impact of PM2.5 concentrations.

4.2 Regression Analysis

In order to determine if PM2.5 concentration levels have a significant relationship to COPD deaths, regression analysis was done using COPD and PM2.5 data. The analysis was performed using the 16 countries identified for further analysis in the previous section. The predictor variable is PM2.5 concentration levels and the response variable is the number of COPD deaths.

4.2.1 Year Based Approach: Initially, ten linear models were created for each year, for which data was available (2010-2019), and analysis was performed for all 16 countries. In other words, each model represented all 16 countries in a given year.
The following code snippet below is how the year based linear models were defined.

```
R code:
df_2010 <- data.frame(copd$Country, pm2.5$X2010, copd$X2010)
colnames(df_2010) <- c('Country', 'pm2.5', 'copd')
lm_2010 <- lm(copd ~ pm2.5, data = df_2010)
summary(lm_2010)
```

For simplicity only one model is shown, similar code was written for years 2011-2019.
Output:

```
> summary(lm_2010)


Call:
lm(formula = copd ~ pm2.5, data = df_2010)

Residuals:
```

```
    Min      1Q  Median      3Q     Max 
-138067  -52316   -9984   25707  368613 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   247622      90450   2.738    0.016 *
pm2.5          -8777       5007  -1.753    0.101  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 114700 on 14 degrees of freedom
Multiple R-squared:   0.18,Adjusted R-squared:  0.1214 
F-statistic: 3.072 on 1 and 14 DF,  p-value: 0.1015

> summary(lm_2011)

Call:
lm(formula = copd ~ pm2.5, data = df_2011)

Residuals:
    Min      1Q  Median      3Q     Max 
-126806  -57560  -19342   27344  395518 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   228018      87382   2.609   0.0206 *
pm2.5          -7396       4733  -1.562   0.1405  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 121000 on 14 degrees of freedom
Multiple R-squared:  0.1485,Adjusted R-squared:  0.08766 
F-statistic: 2.441 on 1 and 14 DF,  p-value: 0.1405

> summary(lm_2012)

Call:
lm(formula = copd ~ pm2.5, data = df_2012)

Residuals:
    Min      1Q  Median      3Q     Max 
```

```
-116927  -61826  -17677   21208  405380

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   212628      82721   2.570   0.0222 *
pm2.5          -6849       4630  -1.479   0.1612
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 122400 on 14 degrees of freedom
Multiple R-squared:  0.1352,Adjusted R-squared:  0.07339
F-statistic: 2.188 on 1 and 14 DF,  p-value: 0.1612

> summary(lm_2013)

Call:
lm(formula = copd ~ pm2.5, data = df_2013)

Residuals:
    Min      1Q  Median      3Q     Max
-123329  -59762  -20241   30285  418637

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   222089      87398   2.541   0.0235 *
pm2.5          -7106       4923  -1.444   0.1709
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 127300 on 14 degrees of freedom
Multiple R-squared:  0.1296,Adjusted R-squared:  0.06739
F-statistic: 2.084 on 1 and 14 DF,  p-value: 0.1709

> summary(lm_2014)

Call:
lm(formula = copd ~ pm2.5, data = df_2014)

Residuals:
    Min      1Q  Median      3Q     Max
-121723  -58712  -15730   30066  412259
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   220541      89171   2.473   0.0268 *
pm2.5          -7298       5248  -1.391   0.1860
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 125900 on 14 degrees of freedom
Multiple R-squared:  0.1214,Adjusted R-squared:  0.05862
F-statistic: 1.934 on 1 and 14 DF,  p-value: 0.186

> summary(lm_2015)

Call:
lm(formula = copd ~ pm2.5, data = df_2015)

Residuals:
    Min      1Q  Median      3Q     Max
-129008  -73070  -15202   32016  433151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   233264      90824   2.568   0.0223 *
pm2.5          -7824       5374  -1.456   0.1675
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132500 on 14 degrees of freedom
Multiple R-squared:  0.1315,Adjusted R-squared:  0.06944
F-statistic: 2.119 on 1 and 14 DF,  p-value: 0.1675

> summary(lm_2016)

Call:
lm(formula = copd ~ pm2.5, data = df_2016)

Residuals:
    Min      1Q  Median      3Q     Max
-125745  -67890  -20395   35213  435807
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   216227      84269   2.566   0.0224 *
pm2.5          -6814       5092  -1.338   0.2022
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132600 on 14 degrees of freedom
Multiple R-squared:  0.1134,Adjusted R-squared:  0.05008
F-statistic: 1.791 on 1 and 14 DF,  p-value: 0.2022

> summary(lm_2017)

Call:
lm(formula = copd ~ pm2.5, data = df_2017)

Residuals:
    Min      1Q  Median      3Q     Max
-137038  -65281  -33761   21070  457394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   231248      96071   2.407   0.0305 *
pm2.5          -8311       5923  -1.403   0.1823
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 141800 on 14 degrees of freedom
Multiple R-squared:  0.1233,Adjusted R-squared:  0.06068
F-statistic: 1.969 on 1 and 14 DF,  p-value: 0.1823

> summary(lm_2018)

Call:
lm(formula = copd ~ pm2.5, data = df_2018)

Residuals:
    Min      1Q  Median      3Q     Max
-120124  -78056  -26569   28722  459085

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   233677      94967   2.461   0.0275 *
pm2.5          -8944       5983  -1.495   0.1571
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140500 on 14 degrees of freedom
Multiple R-squared:  0.1376,Adjusted R-squared:  0.07605
F-statistic: 2.235 on 1 and 14 DF,  p-value: 0.1571

> summary(lm_2019)

Call:
lm(formula = copd ~ pm2.5, data = df_2019)

Residuals:
    Min      1Q  Median      3Q     Max
-120389  -69940  -36537   30261  446397

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   220110      89232   2.467   0.0272 *
pm2.5          -7886       5823  -1.354   0.1971
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 137300 on 14 degrees of freedom
Multiple R-squared:  0.1158,Adjusted R-squared:  0.05266
F-statistic: 1.834 on 1 and 14 DF,  p-value: 0.1971
```

All of the p-values for the pm2.5 variable are above 0.05, this means one can not say that pm2.5 concentrations attribute to the variance in COPD deaths. Additionally, each of the models have large residual standard errors which means they do not fit the data well. Each model also has a low multiple R-squared value which reveals that the model may explain some of the variance in the data but not as much as expected. Lastly, the p-value for all of the models is above 0.05 which means they can not be used to declare a significant relationship between the variables.

4.2.2 Country Based Approach: Based on prior research, it was expected that there would be a significant relationship between the number of COPD deaths and PM2.5 concentration levels (Wen and Gao). Since the summary statistics of the models revealed that these models were not the best fit for the data, it was decided that the data should be reformatted to perform additional linear regressions. For the following models, the regression was performed using data over the years 2010-2019 from only a specific country.

The following code snippet below is how the country based linear models were defined.

```
R code:
df <- data.frame(t(pm2.5[pm2.5$Country == 'Brazil', ]),
t(copd[copd$Country == 'Brazil', ]))
colnames(df) <- c('pm2.5', 'copd')
df <- df[-1,]
df$pm2.5 <- as.numeric(df$pm2.5)
df$copd <- as.numeric(df$copd)
brazil_lm <- lm(copd ~ pm2.5, data = df)
summary(brazil_lm)
```

For simplicity only one model is shown, similar code was written for the following countries: Canada, Colombia, France, Germany, Italy, Japan, Kazakhstan, Mexico, Philippines, Russian Federation, South Africa, Spain, Turkey, Ukraine, and United States of America.

Output:

```
> summary(brazil_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-37832 -19824  -6458  17774  46324

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    697.7    10766.4   0.065     0.95
pm2.5        12842.6      259.6  49.468 2.83e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29220 on 9 degrees of freedom
Multiple R-squared:  0.9963,Adjusted R-squared:  0.9959
F-statistic:  2447 on 1 and 9 DF,  p-value: 2.828e-12

> summary(canada_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-8676.2 -4336.9   443.7  4667.4  8039.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)     98.6     2290.4   0.043    0.967    
pm2.5         6446.6      100.7  63.993   2.8e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6216 on 9 degrees of freedom
Multiple R-squared:  0.9978,Adjusted R-squared:  0.9976 
F-statistic:  4095 on 1 and 9 DF,  p-value: 2.804e-13

> summary(Colombia_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-11415.7  -9069.9   -371.9   4943.9  21930.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   182.11    3975.67   0.046    0.964    
pm2.5        3053.53      75.43  40.482 1.71e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10790 on 9 degrees of freedom
Multiple R-squared:  0.9945,Adjusted R-squared:  0.9939 
F-statistic:  1639 on 1 and 9 DF,  p-value: 1.705e-11

> summary(France_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
```

```
     Min       1Q   Median       3Q      Max 
-25965.8  -2079.5   -181.9   6211.3  13167.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -40.43    4315.81  -0.009    0.993    
pm2.5        2336.78     110.25  21.195 5.44e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11710 on 9 degrees of freedom
Multiple R-squared:  0.9804,Adjusted R-squared:  0.9782 
F-statistic: 449.2 on 1 and 9 DF,  p-value: 5.439e-09

> summary(Germany_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-40379 -13411   3011  19307  31103 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    669.2     9641.8   0.069    0.946    
pm2.5         9871.0      237.2  41.622 1.33e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26170 on 9 degrees of freedom
Multiple R-squared:  0.9948,Adjusted R-squared:  0.9943 
F-statistic:  1732 on 1 and 9 DF,  p-value: 1.33e-11

> summary(Italy_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
```

```
-56473  -2774   1550   7664  31068

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    91.58    8622.86   0.011    0.992
pm2.5        4454.44     157.55  28.274 4.21e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23410 on 9 degrees of freedom
Multiple R-squared:  0.9889,Adjusted R-squared:  0.9876
F-statistic: 799.4 on 1 and 9 DF,  p-value: 4.212e-10

> summary(Japan_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-10490  -6472  -4355   5678  16165

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   189.11    3567.42   0.053    0.959
pm2.5        5458.85      89.32  61.113 4.24e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9682 on 9 degrees of freedom
Multiple R-squared:  0.9976,Adjusted R-squared:  0.9973
F-statistic:  3735 on 1 and 9 DF,  p-value: 4.24e-13

> summary(Kazakhstan_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-23271  -9155   7395   9939  13032
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   179.44    5618.37   0.032    0.975
pm2.5        1272.44      57.51  22.126 3.72e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15250 on 9 degrees of freedom
Multiple R-squared:  0.9819,Adjusted R-squared:  0.9799
F-statistic: 489.6 on 1 and 9 DF,  p-value: 3.719e-09

> summary(Mexico_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-19545  -9974  -3096  10057  23845

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   265.56    5420.29   0.049    0.962
pm2.5        4544.31      80.57  56.403 8.71e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14710 on 9 degrees of freedom
Multiple R-squared:  0.9972,Adjusted R-squared:  0.9969
F-statistic:  3181 on 1 and 9 DF,  p-value: 8.712e-13

> summary(Philippines_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-16083.5  -7440.7    341.2   7781.2  15930.7
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    75.81    4172.21   0.018    0.986
pm2.5        2522.47      58.22  43.325 9.28e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11320 on 9 degrees of freedom
Multiple R-squared:  0.9952,Adjusted R-squared:  0.9947
F-statistic:  1877 on 1 and 9 DF,  p-value: 9.284e-12

> summary(RF_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-24150 -12287   1308  10339  21068

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    332.3     5783.7   0.057    0.955
pm2.5        12713.6      185.5  68.553 1.51e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15700 on 9 degrees of freedom
Multiple R-squared:  0.9981,Adjusted R-squared:  0.9979
F-statistic:  4700 on 1 and 9 DF,  p-value: 1.511e-13

> summary(S_Afr_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-10791.2   -856.5     21.4   2291.8   4597.0

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.75    1599.85   0.003    0.998
pm2.5        1762.76      26.53  66.446    2e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4341 on 9 degrees of freedom
Multiple R-squared:  0.998,Adjusted R-squared:  0.9977
F-statistic:  4415 on 1 and 9 DF,  p-value: 2e-13

> summary(Spain_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8975.5  -547.5  1318.1  3211.8  3970.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    97.83    1764.63   0.055    0.957
pm2.5        5106.01      50.73 100.658 4.78e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4790 on 9 degrees of freedom
Multiple R-squared:  0.9991,Adjusted R-squared:  0.999
F-statistic: 1.013e+04 on 1 and 9 DF,  p-value: 4.783e-15

> summary(Turkey_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-67038  -5617   7933  21844  29545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)    26.93   11961.33   0.002    0.998
pm2.5        3141.24     164.01  19.153 1.33e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32460 on 9 degrees of freedom
Multiple R-squared:  0.9761,Adjusted R-squared:  0.9734
F-statistic: 366.8 on 1 and 9 DF,  p-value: 1.329e-08

> summary(Ukraine_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12636.1  -6954.1   -275.6   6262.6  21143.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -61.25    3921.89  -0.016    0.988
pm2.5        2419.34      81.97  29.513 2.87e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10640 on 9 degrees of freedom
Multiple R-squared:  0.9898,Adjusted R-squared:  0.9886
F-statistic:   871 on 1 and 9 DF,  p-value: 2.874e-10

> summary(US_lm)

Call:
lm(formula = copd ~ pm2.5, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-111011  -52143    6671   61892  110498

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     1482      28880   0.051     0.96
```

```
pm2.5          69346       1084  63.962 2.82e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78380 on 9 degrees of freedom
Multiple R-squared:  0.9978,Adjusted R-squared:  0.9976
F-statistic:  4091 on 1 and 9 DF,  p-value: 2.816e-13
```

Based on the above results, these models are better fit for the data. All of the p-values for the pm2.5 attribute are less than 0.05 which means the variable is significant to the model and attributes to variance in the COPD data. There is also a small residual standard error for each model as compared to the previous yearly models. Therefore, these models fit the data better compared to the other models. Each model also has a high multiple R-squared value which would indicate that pm2.5 concentrations explain a high amount of variation within the COPD data. Lastly, the p-value of all the models is less than 0.05 which means they can be used to declare a significant relationship between the variables. Therefore, utilizing these models allows one to come to the conclusion that there is a significant relationship between PM2.5 concentrations and deaths due to COPD when examining the data by country.

Referring back to hypothesis ii: Air pollution, quantified here by PM2.5, is also a contributing factor that accounts for an increased amount of COPD deaths despite a high percentage of GDP spent on health care.
The linear regression models have proved part of this hypothesis, that PM2.5 is a contributing factor to COPD deaths. In order to prove or disprove the rest of this hypothesis, it is necessary to look specifically at the countries used for the regression models. In the next section, the conclusions of hypothesis i will be restated and then followed by further discussion of hypothesis ii.

5 CONCLUSIONS AND DISCUSSIONS

5.1 Conclusions

i. In terms of hypothesis i, based on the models above, one can not say for certain that countries that spend more on healthcare have lower COPD cases. As mentioned previously, clusters of COPD and GDP data sets were compared. If hypothesis i was true one would expect to see similar groupings of the countries between both clusters. Although many countries are grouped similarly (except for the 16 used for the regression analysis), all four of the GDP clusters are split among the four COPD clusters (i.e. no GDP cluster is entirely contained in a COPD cluster). Additionally, the countries that are grouped similarly may be attributed to the fact that most countries are contained in one COPD cluster.

ii. Moving to hypothesis ii, the regression analysis proved that there is a significant relationship between COPD deaths and PM2.5 concentrations. Additionally, for the 16 countries that were analyzed, all of the models had a positive estimate coefficient for pm2.5, this means COPD cases positively increase when PM2.5 concentration levels increase. The 16 countries used for the regression analysis were also the countries with the 16 highest COPD deaths (when using the total number of deaths over 2000-2020). However, only 7 of these countries were in the top 20 for total amount spent on healthcare (i.e. sum of percent of GDP spent on healthcare over 2000-2020). Therefore, it is not necessarily true for these 16 countries that an increase in COPD deaths is due to lower spending on healthcare. However, PM2.5 concentrations have a significant relationship to COPD deaths. Nonetheless, it does not prove that PM2.5 levels are the only factor that account for differences between the clusters of the data sets. Conclusions were based on analyzing these 16 countries, to continue research, it would be worthwhile to perform the regression analysis utilizing all of the countries in the data set.

5.2 Other Considered Approaches

Multiple other approaches were considered, grouping by year was performed to see if COPD clusters would change. These clusterings did not offer any additional insight, leading to the decision to move forward with the total clustering. Another path pursued was to use a COPD death ratio instead of the number of deaths. It was suspected that the COPD clustering may have been impacted by the fact that the number of deaths was used as opposed to the ratio of deaths to total population. A country with a larger population will likely have more deaths than a country with smaller population even if the ratio of COPD deaths to total population was the same. For this analysis, population data from the World Bank was utilized and the data sets were matched using the country name to create a new data frame with the ratios. Below is the code utilized to create the deaths-population ratio data frame, followed by K-means clustering on the data frame.

```
R code:
population <- read.csv("/Users/keertisundaram/Dropbox/Data Analytics/Final_Project/population.csv"
,header=T)
View(population)
copd <- copd[order(copd$Country),]
population <- population[order(population$Country),]
copd_no_labels <- copd_imp[-1]
ratios <- cbind(population[1],
round(copd_no_labels[-length(copd_no_labels)]/population[-1],5))
View(ratios)

library(ISLR)
library(cluster)
totalClusters <- kmeans(ratios[-1], 4, nstart = 20)
print(totalClusters$cluster)
clusplot(ratios,totalClusters$cluster, color = TRUE,
shade = TRUE, labels = 4, lines = 0,  main = "Number of Deaths Due to COPD from 2000-2020")
```

Based on the output from this clustering, a decision was made not to pursue the ratio approach. Most of the countries still remained in one cluster with the exception of South Africa (cluster 2), Spain (cluster 1), Sri Lanka (cluster 3), and Mexico (cluster 3). It may be interesting to look specifically at these countries to determine why they are outliers within the data.

5.3 Further Proposed Studies

In addition to further regression analysis using all of the countries in the data set, subsequent analysis would involve looking into other contributing factors to COPD such as smoking.

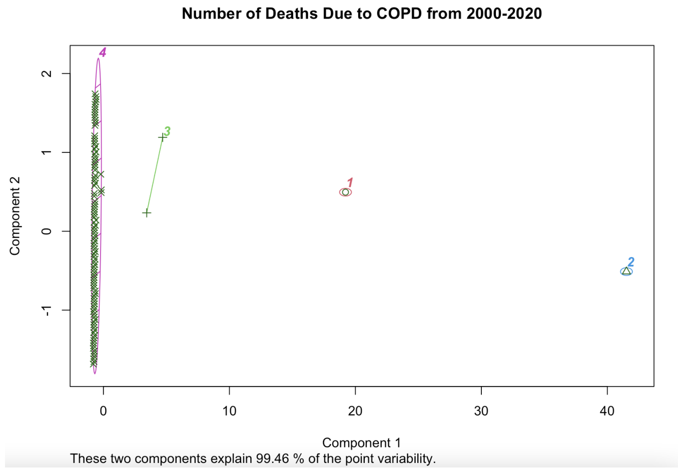


Fig. 14. COPD Clusters Using Ratios

5.4 Code Repository

The code utilized for this analysis can be accessed at the following link: GitHub Repository

6 REFERENCES

"Chronic Obstructive Pulmonary Disease (COPD)." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 8 Apr. 2022, https://www.cdc.gov/copd/index.html.

"Chronic Obstructive Pulmonary Disease (COPD)." World Health Organization, World Health Organization, 16 Mar. 2023, https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease -(copd).

"COPD." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 15 Apr. 2020, https://www.mayoclinic.org/diseases-conditions/copd/symptoms-causes/syc-20353679.

"Fine Particles (PM 2.5) Questions and Answers." *New York State Department of Health*, Feb. 2018, https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm.

"Health Financing." *World Health Organization*, World Health Organization, https://www.who.int/health-topics/health-financingtab=tab_1.

Hosseinpoor, Ahmad Reza, et al. "Promoting Health Equity: Who Health Inequality Monitoring at Global and National Levels." *Global Health Action*, vol. 8, no. 1, 18 Sept. 2015, p. 29034., https://doi.org/10.3402/gha.v8.29034.

Kowarik, Alexander, and Statistik Austria. "K-Nearest Neighbour Imputation." *R Documentation*, https://search.r-project.org/CRAN/refmans/VIM/html/kNN.html.

Kassambara, Alboukadel. “K-Means Clustering in R: Algorithm and Practical Examples.” *Datanovia*, 21 Oct. 2018, https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/.

Wen, Chi Pang, and Wayne Gao. “PM 2·5 : An Important Cause for Chronic Obstructive Pulmonary Disease?” *The Lancet Planetary Health*, vol. 2, no. 3, Mar. 2018, https://doi.org/10.1016/s2542-5196(18)30025-1.

Received 25 April 2023