Title Goes Here

Keerti Sundaram (sundak3@rpi.edu), (¹Rensselaer Polytechnic Institute 110 8th St., Troy, NY, 12180 United States)

Abstract

While there are many communicable and non-communicable diseases impacting countries today, one of the most prominent is Chronic Obstructive Pulmonary Disease (COPD), which is the third leading cause of death worldwide (WHO). Chronic Obstructive Pulmonary Disease (COPD) 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 COPD deaths.

Hypotheses

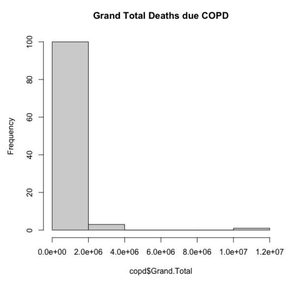
1. Countries with a higher percent of GDP spent on health will have lower COPD deaths.
2. Air pollution, quantified here by PM2.5, is also a contributing factor that accounts for an increased amount of COPD deaths despite a high percentage of GDP spent on health care.

Data Description

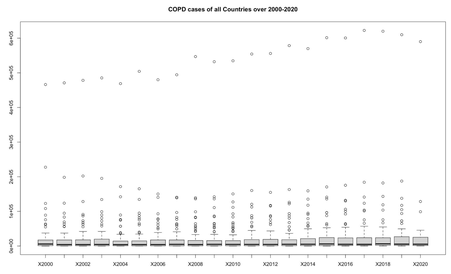
1. Utilized data from the World Health Organization on COPD deaths, percent of GDP spent on health, and PM2.5 Concentrations
2. Performed data imputation on the COPD and GDP datasets using the VIM library and knn-imputation (COPD code shown below)

```
library(VIM)
#copd data
copd_imp <- kNN(copd[-length(copd)], k=5)
#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
```

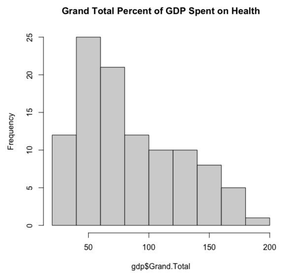
Exploratory Data Analysis



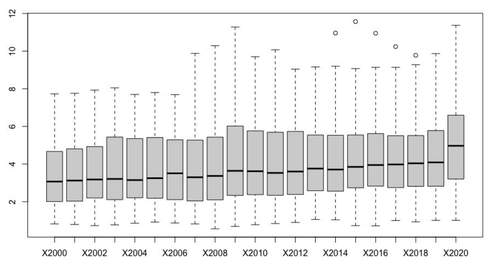
a. COPD histogram



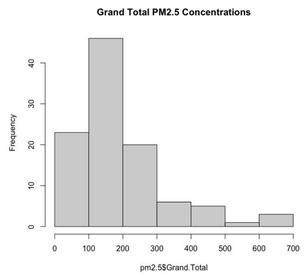
b. COPD boxplot



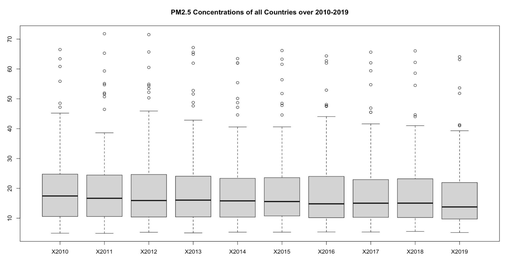
c. GDP histogram



d. GDP boxplot



e. PM2.5 Histogram



f. PM2.5 Boxplot

Plot Descriptions:
a. Histogram of countries' total deaths due to COPD
b. Boxplots of countries' deaths due to COPD by year
c. Histogram of countries' total percent of GDP spent on health care

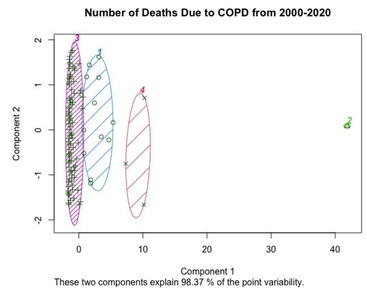
d. Boxplots of countries' percent of GDP spent on health by year
e. Histogram of countries' total concentration levels of PM2.5
f. Boxplots of countries' concentration levels of PM2.5 by year

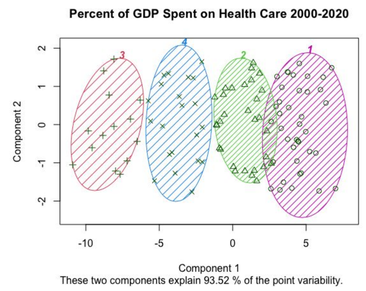
Model Development and Application

1. K-Means Clustering

K-Means Clustering was performed on both the COPD and GDP datasets. The goal of clustering was to determine which countries are grouped together for both datasets. To determine the value of k, elbow plots were used, these are omitted for space.

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





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

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

GDP clusters

2. Linear Regression

Linear regression was performed on countries that were grouped with different countries when comparing the COPD and GDP clusters. These countries were: Brazil, Canada, Colombia, France, Germany, Italy, Japan, Kazakhstan, Mexico, Philippines, Russian Federation, South Africa, Spain, Turkey, Ukraine, and United States of America. The goal of the regression analysis was to determine whether PM2.5 concentration levels have a signficant relationship to COPD deaths. An example code snippet is shown below, only one country is shown, similar code was written for the other 15 countries.

```
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)
```

The summary calls for each of the models revealed that there is a significant relationship between PM2.5 concentrations and deaths due to COPD when examining the data by country. Additionally, each of the 16 countries had a positive estimate coefficient for pm2.5, this means copd has a positive increase when pm2.5 increases.

Conclusions

1. Based on the above models, hypothesis 1 can not be proven or disproven. If hypothesis 1 was true, comparison of the clusters would reveal similar groupings of the countries. 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.
2. The regression analysis proved that there is a significant positive relationship between COPD deaths and PM2.5 concentrations. 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 but this does not prove that it is the only factor that accounts for differences between the clusters of the data sets.. In the future, it would be worthwhile to perform the regression analysis utilizing all of the countries in the data set.

References