Global Fossil Fuel Consumption Project Summary

9/15/2019

Nicole Villanueva [email address]
Kanning Wu [kwu130@syr.edu]
Jackson Shands [jfshands@syr.edu]

Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019

Table of Contents

Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019

Project Abstract

In this project, we applied the K-means clustering algorithm and Elbow Method, Silhouettes Methods to test the best number of clusters. For the K-means clustering algorithm, we first cleaned data and standardize it. Then, we tested the algorithm with different numbers of clusters from 2 to 10. For Elbow Method and Silhouettes Methods, we implemented both functions. With these two functions, we reached a result of having 10 as our optimal number of clusters. After we had the best number of clusters, we ran our code and got the final conclusion. The datasets we used were from datahub.com

Contribution Statement

Nicole Villanueva: Gathered and completed the K-mean cluster Algorithm with Linna Yu. Finished the presentation slides

Jackson Shands: Revised the code, made modifications, ran the coeffie inserted plots on project summary

Kanning Wu: Write the Elbow and Silhouettes method function for the best number of clustering.

Introduction

With the worsening climate every year, we tried to build a more globally collaborative plan to combat the global warming crisis devastating the planet, it is important to survey the recent patterns of carbon emissions from fossil fuels of countries all over the world (249) to identify similar patterns. The U.S. can then accurately measure its contribution to the current state of the atmosphere and identify other leading countries with lower carbon emission levels to learn from, potentially collaborate with, and hopefully mirror in the coming years as an effort to significantly reduce the U.S.'s contribution to carbon emissions and global warming. In this project, we investigated the first phrase of the U.S.'s climate action plan by grouping together similar fossil fuel consumption patterns across yearly totals of countries all over the globe and the research questions are how do countries across the globe compare in terms of carbon emissions from burning fossil fuels and what are the major differences between the U.S.'s fossil fuel/carbon emissions and that of countries with lower levels of carbon emissions/fossil fuel consumption.

The Dataset

The data set that we used is composed of per Country CO2 Emissions from fossil-fuels annually since 1751 till 2014. It outlines each country's carbon emissions from the consumption of fossil fuels by year and type. We didn't need the data to go as far back as 1751, so we chose to cut it off at 1958 to match with data sets that would be relevant for phase 2 of this project. (This is the climate data set that merged together global atmospheric CO2 levels by year and average global temperature by year.)

Data 1	
CO2 PPM Trends in Atmospheric Carbon Dioxide	https://datahub.io/core/co2-ppm
Data 2	
Global Temperature Time Series	https://datahub.io/core/global-temp
Data 3	
CO2 Emissions from Fossil Fuels since 1751, By nation	https://datahub.io/core/co2-fossil-by-nation

Data 1 Description

Date	YYYY-MM-DD
Decimal Date	
Average	The monthly mean CO2 mole fraction determined from daily averages. If there are missing days concentrated either early or late in the month, the monthly mean is corrected to the middle of the month using the average seasonal cycle. Missing months are denoted by -99.99.
Interpolated	Values from the average column and interpolated values where data are missing. Interpolated values are computed in two steps. First, we compute for each month the average seasonal cycle in a 7-year window around each monthly value. In this way the seasonal cycle is allowed to change slowly over time. We then determine the trend value for each month by removing the seasonal cycle; this result is shown in the trend column. Trend values are linearly interpolated for missing months. The interpolated monthly mean is then the sum of the average seasonal cycle value and the trend value for the missing month.
Trend	Seasonally corrected
Number of Days	-1 denotes no data for number of daily averages in the month.

Data 2 Description

Source	
Year	YYYY

Mean	Average global mean temperature anomalies in degrees Celsius relative to a base period. GISTEMP base period: 1951-1980. GCAG base period: 20th century average.

Data 3 Description

Year	YYYY
Country	Nation
Total	Total carbon emissions from fossil fuel consumption and cement production (million metric tons of C)
Solid Fuel	Carbon emissions from solid fuel consumption
Liquid Fuel	Carbon emissions from liquid fuel consumption
Gas Fuel	Carbon emissions from gas fuel consumption
Cement	Carbon emissions from cement fuel consumption
Gas Flaring	Carbon emissions from gas flaring
Per Capita	Per capita carbon emissions (metric tons of carbon; after 1949 only)
Bunker fuels (Not in Total)	Carbon emissions from bunker fuels (not included in total)

Data Analysis Methods

After we gathered the data and chose one dataset from each data packages. We deleted some unused columns and merged datasets into different parts for each different steps. After we cleaned the dataset, we could start to analyze the data by K-mean Clustering algorithm.

Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019

Our datasets contains large number of observations and based on our research questions, we used K-means cluster analysis. The K-means algorithm is a method by which observations are grouped or categorized around a centroid so that the mean difference of each data point's distance from the centroid is as small as possible and clustering is a broad set of techniques of finding subgroups of observations. It is an unsupervised method, which implies that it isn't required being trained by a response variable. *The classification of observations into groups requires some methods for computing the distance or the similarity between each pair of observations.*

To identify the best number of clusters, we used the elbow method to identify where the minimization of the within sum of squares leveled off. The elbow method results in a curve of the within-clusters sum of squares, and the "elbow" of the curve is generally the ideal number of clusters because there are diminishing returns as we increase the number of clusters. If we were to keep going far enough, the within-clusters sum of squares would eventually be 0, when the number of clusters equals the number of data points.

Analysis Summary

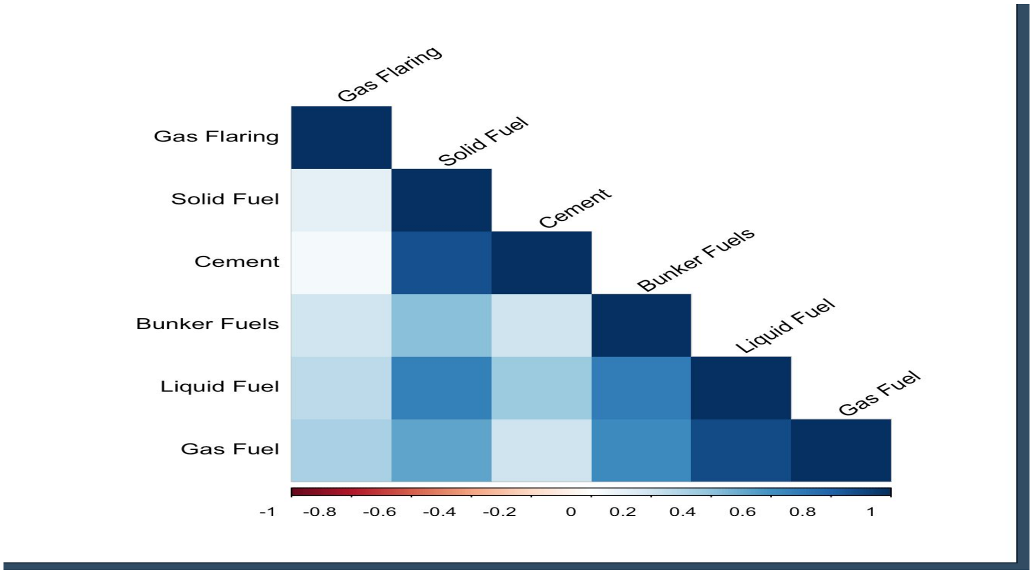
In the exploratory phase of the fossil fuel data, we created a map to display the emissions released by each country. This visualization aligns is showing the United States of America and Russia have the same pattern of CO2 emissions

Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019

After running a Pearson Correlation method on the fossil fuel data to identify the similarities between the fossil fuel types we can identify that the following attributes are statistically significant. Cement and Solid Fuel,Liquid Fuel and Gas Fuel, Cement and Solid Fuel. The factors mentioned before are statistically significant and it can be determined that they have a relationship because they have a strong coefficient correlation (R) with it.

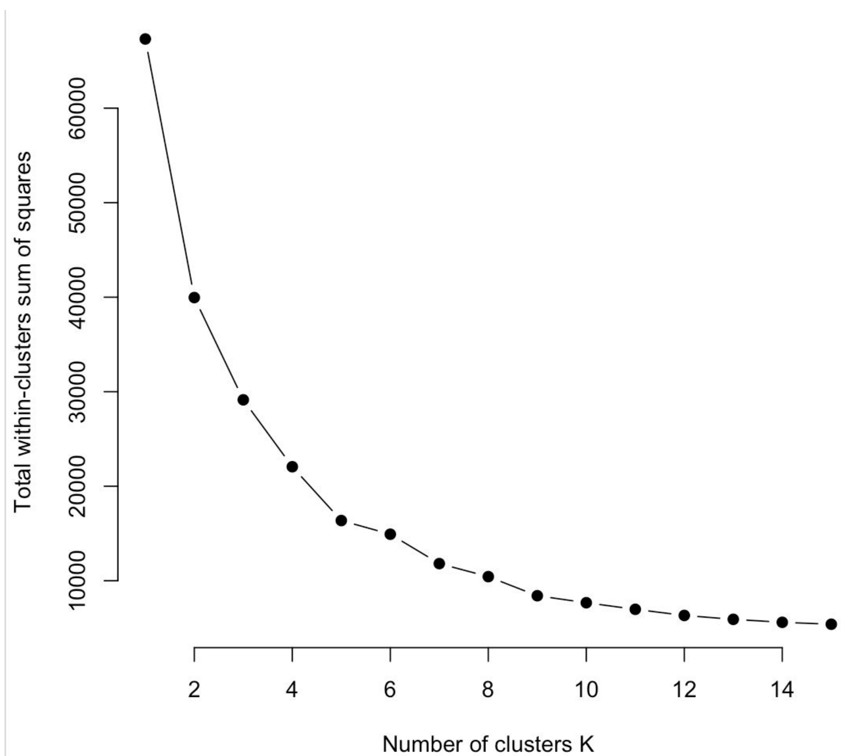
```
> ResCor
             Solid Fuel Liquid Fuel  Gas Fuel     Cement Gas Flaring Bunker Fuels
Solid Fuel    1.0000000   0.6732223 0.5265631 0.87746533  0.11969293    0.4168783
Liquid Fuel   0.6732223   1.0000000 0.8965653 0.36039787  0.26708333    0.6941348
Gas Fuel      0.5265631   0.8965653 1.0000000 0.20457697  0.32887213    0.6346965
Cement        0.8774653   0.3603979 0.2045770 1.00000000  0.04972761    0.2067740
Gas Flaring   0.1196929   0.2670833 0.3288721 0.04972761  1.00000000    0.1902904
Bunker Fuels  0.4168783   0.6941348 0.6346965 0.20677403  0.19029040    1.0000000
>
```

Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019



The comparison on New Zealand and Ireland. These countries have similar or better HDI (which correlates with a strong economy) than the U.S., so they are highly likely to be the countries to emulate in terms of sustainability practices. The U.S. should study how these countries keep their fossil fuel consumption so low and adopt best practices where possible.

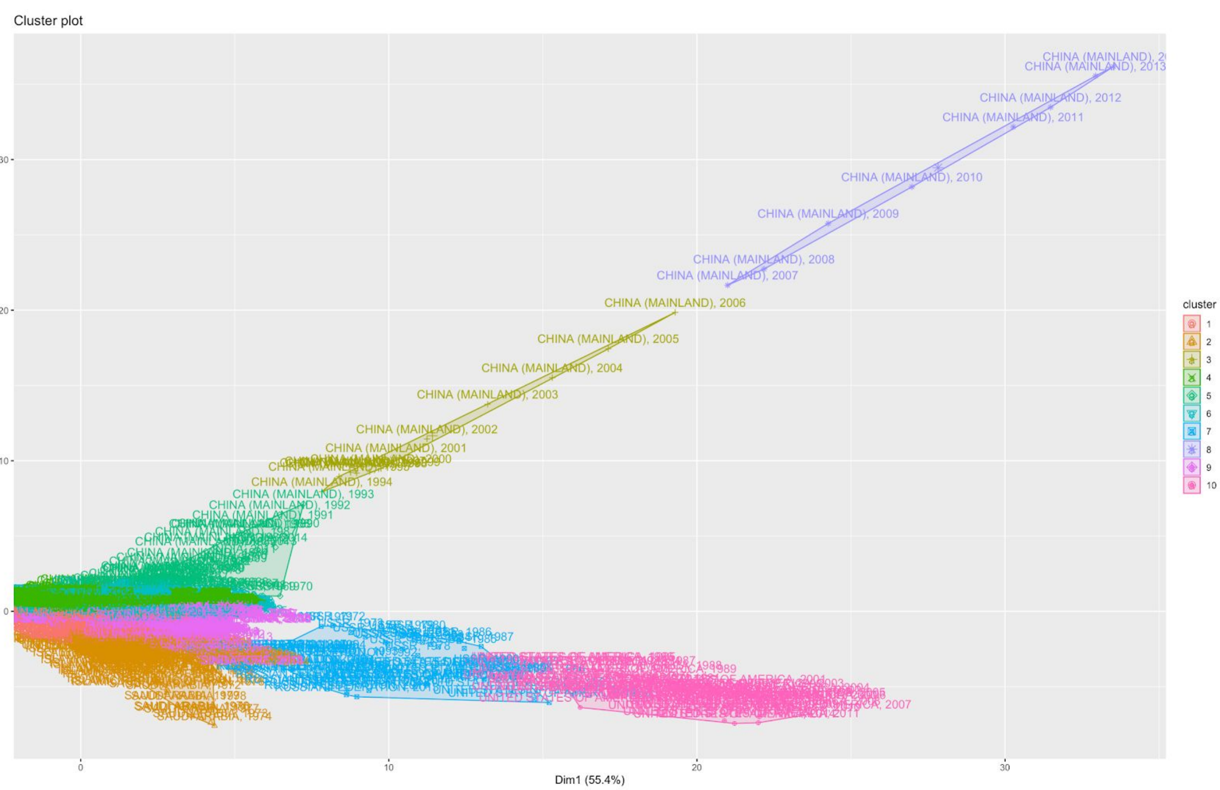
By Elbow method, we found that the best number of clusters is 10.



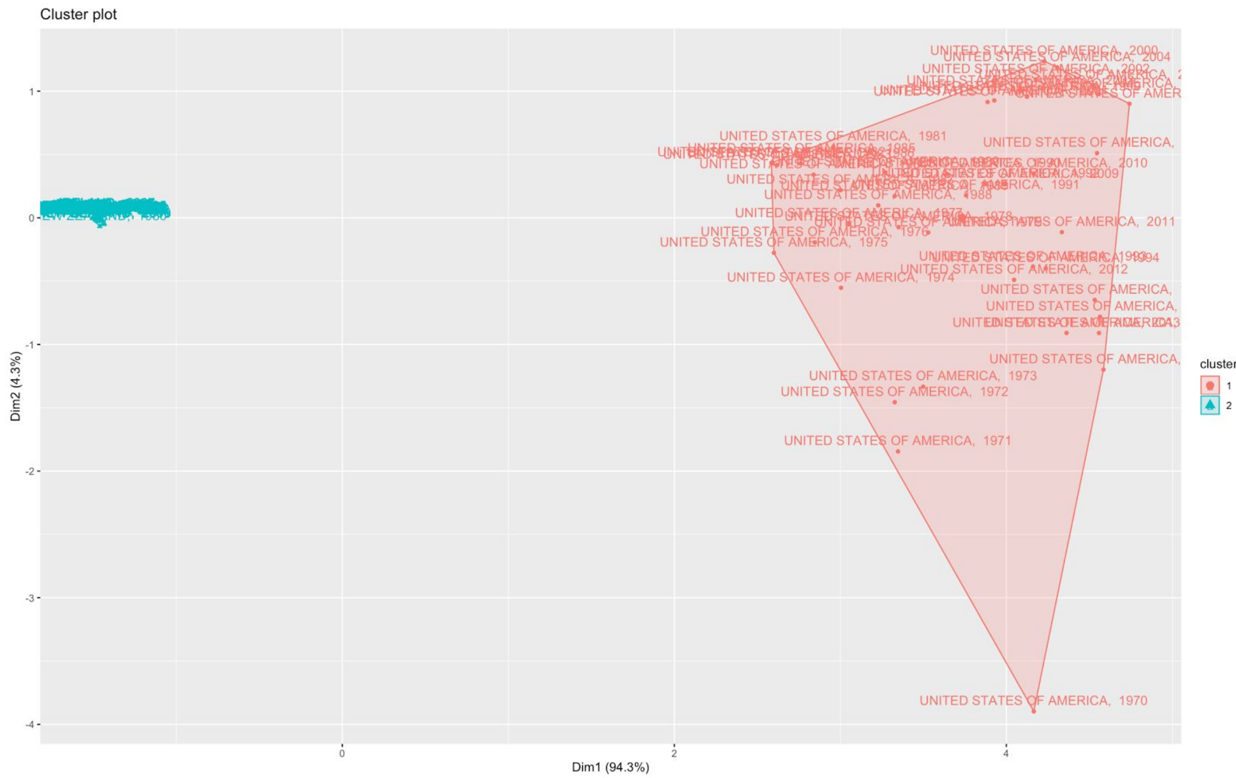
Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019

Through K-means clustering, we found that the U.S.'s fossil fuel consumptions from 1958 - 1969 are highly similar to Russia's fuel patterns from 1970-2014 (Cluster 7). The U.S. was only present in 2 clusters, which are 7 and 10. In cluster 7, we could see that the U.S.'s carbon emissions from fossil fuel consumption were highly similar to Russia's carbon emissions from fossil fuel consumption from 1969 to 2014. After 1969, the US's carbon emissions from fossil fuel consumption are unique. This indicates a distinct shift in 1969 on how the country consumed fossil fuel. It could be an excellent place to start looking to better understand what has driven fuel consumption in the U.S. There was a distinct shift in carbon emissions from fossil fuel consumption in the U.S. between 1969 to 1970.

The U.S fossil fuel consumption patterns from 1969 to current are unique (cluster 10)



Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019



New Zealand and Ireland's fossil fuel consumption from the entirety of the time period in question fall into cluster 4, but only Singapore's fossil fuel patterns from 1958-1977 match up for Singapore. New Zealand, Ireland, and Singapore are countries of similar or better Human Development Index (HDI) than the U.S. with lower fossil fuel consumption levels and they were all in one cluster.

Conclusions

The conclusion is that Congress agree to form a plan to combat climate change while further research is conducted on the fossil fuel practices and economic profiles of other developed countries with lower carbon emissions. It is clear from the U.S.-specific cluster that our fossil fuel consumption practices are not only unique to the rest of the world, but that it has changed and varied widely since the 1970s and that it is a major detrimental contributor to the health of the planet. If we have a long history of fluctuations in our carbon emissions from fossil fuel consumption, we have the capacity and ability to continue changing and potentially move in a direction that is less harmful to the planet.

Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019

References

https://uc-r.github.io/kmeans_clustering

https://datahub.io/core/co2-ppm

https://datahub.io/core/global-temp

https://datahub.io/core/co2-fossil-by-nation#r

Appendix

```
### Climate Project

# Install packages

#install.packages("ggplot2")

library(ggplot2)

library(tidyverse)  # data manipulation

library(cluster) # clustering algorithms

library(factoextra) # clustering algorithms & visualization

# Read in the csv data for global co2 levels, global temperature, and fossil fuel emmissions by nation.

atmos <- read.csv("co2-mm-mlo_csv.csv")

globaltemp <- read.csv("annual_csv.csv")

fossilread <- read.csv("fossilfuelnation.csv")

### Clean up the datasets so they can knit together neatly.
```

Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019

```
# Separate GCAG and GISTEMP in globaltemp

gistemp <- globaltemp[globaltemp$Source=="GISTEMP",]

gcagtemp <- globaltemp[globaltemp$Source=="GCAG",]


# Extract the year from atmos$Date and put it in a new column

atmos$Date <- as.Date(atmos$Date,"%Y-%m-%d")

atmos$Year <- as.numeric(format(atmos$Date,"%Y"))

atmos <- atmos[,c(1,7,3,4,5)]


# Find the yearly average of each variable in co2data

annualatmos <-
data.frame("Year"=unique(atmos$Year),"Average"=aggregate(Average~Year,atmos,mean),


"Interpolated"=aggregate(Interpolated~Year,atmos,mean),"Trend"=aggregate(Trend~Year,atmo
s,mean))

annualatmos <- annualatmos[,c(-2,-4,-6)]

colnames(annualatmos) <- c("Year","AvgC02","InterpolatedCO2","TrendCO2")


# Merge Global Temp and CO2 emmissions data

climatedata <- merge(annualatmos,gistemp,by="Year")

colnames(climatedata)[colnames(climatedata)=="Mean"] <- "Avg GISTEMP"

climatedata <- climatedata[,-5]
```

Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019

```
# Clean fossil fuel data by removing extraneous columns, limiting time frame to match climatedata

fossil <- fossilread[,c(-4:-10)]

recentfossil <- fossil[fossil$Year>="1958",]

colnames(recentfossil) <- c("Year","Country","Total") ## switch order


# Add global sums to climatedata

climatedata <- climatedata[which(climatedata$Year<=2014),]

a <- aggregate(Total~Year,recentfossil,sum)

climatedata$TotalGlobalCE <- a[,2]

# climatedata <- climatedata[,-6]


### Analyze the data


# Create a multiline visual an see if there are any trends

colnames(climatedata) <- c("Year","AvgCE","InterpolatedCE","TrendCE","AvgTemp","TotalCE")

# Carbon emissions trends multiline:

ggplot() +

  geom_line(data = climatedata, aes(x = Year, y = AvgCE, group=1), color = "blue") +

  geom_line(data = climatedata, aes(x = Year, y = TrendCE, group=1), color = "green") +

  ylab("Carbon Emissions")

# Total global carbon emissions line

ggplot() + geom_line(data=climatedata, aes(x = Year, y = TotalCE, group=1), color = "red") +
```

```
  ylab("Carbon Emissions")

# Average GIS temp anomaly plotline

ggplot() + geom_line(data=climatedata, aes(x = Year, y = AvgTemp, group=1), color = "orange")
+

  ylab("Average Temperature Anomalies")



# Create a fossil types dataframe

fossiltypes <- fossilread[,-9]

fossiltypes$CountryYear <- paste(fossiltypes$Country,fossiltypes$Year, sep=", ")

# Keep years 1958 - 2014 to match climatedata

fossiltypes <- fossiltypes[which(fossiltypes$Year>=1958),]

fossiltypes <- fossiltypes[which(fossiltypes$Year<=2014),]

# Reorder columns and keep relevant columns

fossiltypes <- fossiltypes[,c(10,1:9)]

rownames(fossiltypes) <- fossiltypes[,1]

fossiltypes <- fossiltypes[,-1]


# Create GGplot to explore World data

library(ggplot2)

library(dplyr)


# Clean map data to match world data

MapFossilType <- fossiltypes

MapFossilType$Country <- tolower(MapFossilType$Country)
```

```
MapFossilType$Country <- gsub("united states of america","usa",MapFossilType$Country)

MapFossilType$Country <- gsub("russian federation","russia",MapFossilType$Country)

MapFossilType$Country <- gsub("china (mainland)","china",MapFossilType$Country)




require(maps)

world_map <- map_data("world")

PrepMapNames <- unique(world_map$region)

PrepMapNames <- as.data.frame.character(PrepMapNames,stringsAsFactors = F)

colnames(PrepMapNames) <- c("country")

str(PrepMapNames)


FossilPrepName <- unique(fossiltypes$Country)

FossilPrepName <- as.data.frame.character(fossiltypes$Country)


MapFossilTypeMatch <-(merge(fossiltypes,PrepMapNames,by.x = "Country",by.y = "country",all
= F))




WorldData <- map_data('world') %>% filter(region != "Antarctica") %>% fortify

WorldData$region <- tolower(WorldData$region)


p <- ggplot() +

  geom_map(data = WorldData, map = WorldData,
```

```
        aes(x = long, y = lat, group = group, map_id=region),

        fill = "white", colour = "#7f7f7f", size=0.2) +

  geom_map(data = MapFossilType, map=WorldData,

        aes(fill=Total, map_id=Country),

        colour="#7f7f7f", size=0.2) +

  coord_map("rectangular", lat0=0, xlim=c(-180,180), ylim=c(-60, 90)) +

  scale_fill_continuous() +

  scale_y_continuous(breaks=c()) +

  scale_x_continuous(breaks=c()) +

  labs(fill="legend", title="Fossil Fuel Totals", x="", y="") +

  theme_bw()

p
```

```
# Correlation test

library(Hmisc)

CorrFossil <-fossiltypes

CorrFossil <- CorrFossil[,c(-1:-4)]

colnames(CorrFossil) <- c("Solid Fuel","Liquid Fuel","Gas Fuel","Cement","Gas Flaring","Bunker
Fuels")
```

```
ResCor <- cor(as.matrix(CorrFossil))

ResCor
```

```
library(corrplot)

corrplot(ResCor, method = "pie",

        order = "hclust",addrect = 3,

        tl.col = "black",rect.col = "orange",

        bg="grey")
```

```
corrplot(ResCor,method = "shade",order = "hclust",

        type="lower", addrect = 3, bg = "grey",

        cl.pos = "b",cl.ratio = 0.1, cl.align = "r",

        tl.col = "black",tl.srt = 40)
```

```
# Standardize it

stndfossil <- fossiltypes[,4:9]

#stndfossil <- as.numeric(stndfossil)
```

```
stndfossil <- scale(stndfossil)


# Run a cluster with 2 centers

k2 <- kmeans(stndfossil, centers = 2, nstart = 25)

str(k2)

fviz_cluster(k2, data = stndfossil)


# Different clusters

k3 <- kmeans(stndfossil, centers = 3, nstart = 25)

fviz_cluster(k3, data = stndfossil)

k4 <- kmeans(stndfossil, centers = 4, nstart = 25)

fviz_cluster(k4, data = stndfossil)

k5 <- kmeans(stndfossil, centers = 5, nstart = 25)

fviz_cluster(k5, data = stndfossil)


# Compare clusters

p1 <- fviz_cluster(k2, geom = "point", data = stndfossil) + ggtitle("k = 2")

p2 <- fviz_cluster(k3, geom = "point",  data = stndfossil) + ggtitle("k = 3")

p3 <- fviz_cluster(k4, geom = "point",  data = stndfossil) + ggtitle("k = 4")

p4 <- fviz_cluster(k5, geom = "point",  data = stndfossil) + ggtitle("k = 5")


library(gridExtra)

grid.arrange(p1, p2, p3, p4, nrow = 2)
```

Rather than testing and plotting an increasing number of clusters, let's use the elbow method to find a a solid number.

Elbow method

Compute and plot wss (within cluster sum of squares) for k = 2 to k = 15. We want a low sum of squares.

```
k.max <- 15

wss <- sapply(1:k.max,

        function(k){

                kmeans(stndfossil, k, nstart=50,iter.max = 15 )$tot.withinss

                })

wss

plot(1:k.max, wss,

        type="b", pch = 19, frame = FALSE,

        xlab="Number of clusters K",

        ylab="Total within-clusters sum of squares")
```

It looks like the sum of squares levels off around 10, so we can rely on a model with 10 centers.

```
k10 <- kmeans(stndfossil, centers = 10, nstart = 25)

fviz_cluster(k10, data = stndfossil)


avg_sil <- function(k) {
```

```
  km.res <- kmeans(na.omit(stndfossil), centers = k, nstart = 25)

  ss <- silhouette(km.res$cluster, dist(stndfossil))

  mean(ss[, 3])

}

# Compute and plot wss for k = 2 to k = 15

k.values <- 2:15

# extract avg silhouette for 2-15 clusters
stndfossil = data.frame(stndfossil)

avg_sil_values <- map_dbl(k.values, avg_sil)

plot(k.values, avg_sil_values,

       type = "b", pch = 19, frame = FALSE,
```

```
    xlab = "Number of clusters K",

    ylab = "Average Silhouettes")

##################################################

gap_stat <- clusGap(stndfossil, FUN = kmeans, nstart = 25,

        K.max = 25, B = 50)

print(gap_stat, method = "firstmax")

fviz_gap_stat(gap_stat)

# Let's put the cluster data into a data frame so we can see where the US maps out in all of this.
kdata <- data.frame(k10$cluster)
colnames(kdata) <- "cluster"
kdata$countryYear <- rownames(kdata)
kdata <- kdata[,c(2,1)]
rownames(kdata) <- NULL
# library(tidyr)
kdata2 <- separate(kdata , countryYear , c("Country" , "Year") , ",")
kdf <- data.frame(kdata2)
```

```
# Separate out US data

UScluster <- kdf[which(kdf$Country=="UNITED STATES OF AMERICA"),]


# Looks like the US is only in cluster 7 and 10. Let's see what countries we clustered with, if any

c7 <- data.frame(kdata[which(kdata$cluster==7),])

c10 <- data.frame(kdata[which(kdata$cluster==10),])

# Cluster 7 contains the US and Russia (1958-1969 US; 1971-2014 Russia)

# Cluster 10 only contains the US (1970-2014)


# Let's look at the fossil fuel data for each cluster

fossiltypes$row <- rownames(fossiltypes)

fossiltypes <- fossiltypes[,c(10,1:9)]

rownames(fossiltypes) <- NULL

allcfd <- merge(fossiltypes,kdata,by.x="row",by.y="countryYear")

c1fd <- allcfd[which(allcfd$cluster==1),]

c2fd <- allcfd[which(allcfd$cluster==2),]

c3fd <- allcfd[which(allcfd$cluster==3),]

c4fd <- allcfd[which(allcfd$cluster==4),]

c5fd <- allcfd[which(allcfd$cluster==5),]

c6fd <- allcfd[which(allcfd$cluster==6),]

c7fd <- allcfd[which(allcfd$cluster==7),]

c8fd <- allcfd[which(allcfd$cluster==8),]
```

```
c9fd <- allcfd[which(allcfd$cluster==9),]

c10fd <- allcfd[which(allcfd$cluster==10),]

# Remove cluster column

c1fd <- c1fd[,-11]

c2fd <- c2fd[,-11]

c3fd <- c3fd[,-11]

c4fd <- c4fd[,-11]

c5fd <- c5fd[,-11]

c6fd <- c6fd[,-11]

c7fd <- c7fd[,-11]

c8fd <- c8fd[,-11]

c9fd <- c9fd[,-11]

c10fd <- c10fd[,-11]


# Identify which countries (HDI > 15) have the lowest contributions to fossil fuel consumption.

HDI15 <- c("NORWAY", "SWITZERLAND", "AUSTRALIA", "IRELAND", "GERMANY",
"ICELAND", "SWEDEN", "SINGAPORE", "NETHERLANDS", "DENMARK", "CANADA",
"UNITED STATES OF AMERICA", "UNITED KINGDOM", "FINLAND", "NEW ZEALAND")

HDI15 <- as.data.frame(HDI15)

colnames(HDI15) <- "country"



# Calculate ffc totals for each country from 1958-2014

ffctotal <- function(country) {

  total <- sum(fossiltypes$Total[which(fossiltypes$Country==country)])
```

```
  return(total)

}

tot1 <- ffctotal("NORWAY")

tot2 <- ffctotal("SWITZERLAND")

tot3 <- ffctotal("AUSTRALIA")

tot4 <- ffctotal("IRELAND")

tot5 <- ffctotal("GERMANY")

tot6 <- ffctotal("SWEDEN")

tot7 <- ffctotal("SWITZERLAND")

tot8 <- ffctotal("SINGAPORE")

tot9 <- ffctotal("NETHERLANDS")

tot10 <- ffctotal("DENMARK")

tot11 <- ffctotal("CANADA")

tot12 <- ffctotal("UNITED STATES OF AMERICA")

tot13 <- ffctotal("UNITED KINGDOM")

tot14 <- ffctotal("FINLAND")

tot15 <- ffctotal("NEW ZEALAND")


# Add totals to each country in HDI15

HDI15$totals <- c(tot1,tot2,tot3,tot4,tot5,tot6,tot7,tot8,tot9,tot10,tot11,tot12,tot13,tot14,tot15)

HDI15$rank <- rownames(HDI15)

HDI15 <- HDI15[,c(3,1,2)]

HDI15$totals <- as.numeric(HDI15$totals)
```

```
# Order least to greatest

HDI15 <- HDI15[order(HDI15$totals),]

head(HDI15)

# Find the clusters that NZ, Ireland, and Singapore are in

NZclusters <- kdf[which(kdf$Country=="NEW ZEALAND"),] # Cluster 4

Irelclusters <- kdf[which(kdf$Country=="IRELAND"),] # Cluster 4

Singclusters <- kdf[which(kdf$Country=="SINGAPORE"),] # Cluster 4, 6, 9



# Cluster 4 looks interesting. Lets take a look at these top HDI countries in cluster 4 with fossilfuel data

c4top3 <- c4fd[which(c4fd$Country=="NEW ZEALAND"),]

c4top3 <- rbind(c4top3,c4fd[which(c4fd$Country=="IRELAND"),])

c4top3 <- rbind(c4top3,c4fd[which(c4fd$Country=="SINGAPORE"),])

# We can see from the years that cluster 4 contains every year (1958-2014)

# for New Zealand and Ireland, and only years 1958-1977 for Singapore.



# we can also compare each cluster's global historical (1958-2014) fossil impact

clustersum <- function(cluster) {

  fossilimpact <- sum(cluster$Total)/sum(fossiltypes$Total)

  return(fossilimpact)

}

c4sum <- clustersum(c4fd)

c7sum <- clustersum(c7fd)

c10sum <- clustersum(c10fd)
```

Global Fossil Fuel Consumption Project Summary
IST 687 Applied Data Science
Summer 2019