Climate Change

Stephen Smitshoek

19/05/2022

Import and Clean Data

Each of the five files will be imported into R using 'read.csv'

The global_temp dataframe will have any data before the year 1753 deleted as there are many N/As. Then the mean global temperature for each year will be calculated using the aggregate function. Finally the column names will be updated to something more readable.

The country_temp dataframe will have its yearly means aggregated and column names updated.

The co2_atmo dataframe will have its yearly means aggregated and column names updated.

The co2_emiss data frame will have the sum the total emissions each year aggregated from each country and stored as a total number.

The sea_level dataframe will have its yearly means aggregated and column names updated.

All of the five databases will be joined together into one dataframe, climate_change_df, based on the year. This will allow the data to be easily accessed and compared.

Final Dataset

```
##   year  AvgTemp AvgTempUncer CO2_ppm CO2_ppm_Seas_Adj co2_tonnes GMSL
## 1 1753 8.388083     3.176000      NA               NA   28062576   NA
## 2 1754 8.469333     3.494250      NA               NA   28073568   NA
## 3 1755 8.355583     3.850333      NA               NA   28084560   NA
## 4 1756 8.849583     3.262333      NA               NA   30019152   NA
## 5 1757 9.022000     4.026000      NA               NA   30030144   NA
## 6 1758 6.743583     3.362917      NA               NA   30041136   NA
##   GMSL.uncertainty
## 1               NA
## 2               NA
## 3               NA
## 4               NA
## 5               NA
## 6               NA
```

What Information is Not Self Evident

The uncertainties in the average temperature and global mean seal level (GMSL) along with the seasonally adjusted CO2 ppm will need to plotted to ensure that even if the levels are rising it is not entirely due to a rising uncertainty.

The temperature as a global level may be on the rise but to ensure that it is not a few regions skewing the data it will be necessary to pick a few random countries from different regions and check if they are rising as well and if that is similar to the global rate. If one region stands out it may be necessary to see if that matches with the CO2 emissions in that area.

What Are Different Ways You Could Look at This Data

The data will be looked at using scatter plots, histograms, and linear models. Through these avenues it should be possible to determine if there is a a global temperature change occurring and if so, if it is correlated to a rise in CO2 emissions caused by humans.

Data Summary to Answer Key Questions

The data was summarized into yearly averages and sums for each piece of data then joined into a single data frame for ease of use. Some cleaning of the data will still need to be done to account for missing data in specific years.

Types of Plots and Tables

Scatter plots of temperatures, sea level, CO2 levels and emissions will need to be plotted over time to see if there is truly a change.

A histogram of a temperature delta over a few decades of different regions and the global average will help to show if the global change is being driven by a few select regions.

Tables of the different coefficients and their respective p values in each linear model that is created will help to highlight if there is a strong relationship between the variables used in the model, for example, global temperature and sea level.

Machine Learning

If machine learning can be used to improve the model of predicting how the global temperature is changing then it will be implemented. If the basic level of machine learning does not provide an improvement over a linear model it will not be considered.

Questions for Future Steps

- Can machine learning be implemented to improve the accuracy of a linear model?
- If CO2 emissions and temperature change is correlated can causation be proved?

```
---
title: "Climate Change"
author: "Stephen Smitshoek"
date: "19/05/2022"
output: pdf_document
---

# Import and Clean Data

```{r libraries, echo=FALSE, warning=FALSE, message=FALSE}
library(dplyr)
```

Each of the five files will be imported into R using 'read.csv'
```{r setup, echo=FALSE}
setwd("C:\\Users\\sksmi\\PeytoAccess\\Personal\\Bellevue\\DSC520\\dsc520")
global_temp <- read.csv("final_project\\data\\GlobalTemperatures.csv")
country_temp <- read.csv("final_project\\data\
\GlobalLandTemperaturesByCountry.csv")
co2_atmo <- read.csv("final_project\\data\\co2_atmo.csv")
co2_emiss_country <- read.csv("final_project\\data\\co2_emission.csv")
sea_level <- read.csv("final_project\\data\\sea_levels_2015.csv")
```

The global_temp dataframe will have any data before the year 1753 deleted as
there are many N/As. Then the mean global temperature for each year will be
calculated using the aggregate function. Finally the column names will be
updated to something more readable.
```{r global_temp Clean Up, echo=FALSE}
global_temp$dt <- as.Date(global_temp$dt, "%Y-%m-%d")
global_temp <- global_temp[global_temp$dt >= as.Date("1753-01-01", "%Y-%m-
%d"),]
global_temp <- subset(global_temp, select = c(dt,
                                              LandAverageTemperature,
                                              LandAverageTemperatureUncertainty))
global_temp <- aggregate(global_temp[,c(2,3)], by=list(format(global_temp$dt,
format="%Y")), FUN=mean)
colnames(global_temp) <- c("year", "AvgTemp", "AvgTempUncer")
```

The country_temp dataframe will have its yearly means aggregated and column
names updated.
```{r country_temp Clean Up, echo=FALSE}
country_temp$dt <- as.Date(country_temp$dt, "%Y-%m-%d")
country_temp <- aggregate(cbind(AverageTemperature,
                                AverageTemperatureUncertainty) ~
                            format(country_temp$dt, format="%Y") + Country,
                          data = country_temp,
                          FUN=mean)
colnames(country_temp)[1] <- "year"
```

The co2_atmo dataframe will have its yearly means aggregated and column names
updated.
```{r co2_atmo Clean Up, echo=FALSE}
```
```

```
co2_atmo$date <- as.Date(paste0(co2_atmo$Year, "-", co2_atmo$Month, "-",
"01"),
                         "%Y-%m-%d")
co2_atmo <- subset(co2_atmo, select = c(date,
                                        Carbon.Dioxide..ppm.,
                                        Seasonally.Adjusted.CO2..ppm.))
co2_atmo <- aggregate(co2_atmo[,c(2,3)], by=list(format(co2_atmo$date,
format="%Y")), FUN=mean, na.action=na.omit)
colnames(co2_atmo) <- c("year","CO2_ppm", "CO2_ppm_Seas_Adj")
```

The co2_emiss data frame will have the sum the total emissions each year aggregated from each country and stored as a total number.

```{r co2_emiss Clean Up, echo=FALSE}
colnames(co2_emiss_country)[4] <- "co2_tonnes"
co2_emiss_global <- aggregate(co2_emiss_country$co2_tonnes,
                              by=list(co2_emiss_country$Year),
                              FUN=sum)
colnames(co2_emiss_global) <- c("year", "co2_tonnes")
co2_emiss_global$year <- as.character(co2_emiss_global$year)
```

The sea_level dataframe will have its yearly means aggregated and column names updated.

```{r sea_level Clean Up, echo=FALSE}
sea_level$date <- as.Date(paste0(substring(sea_level$Time, 1, 4), "-",
                                 "01", "-", "01"), "%Y-%m-%d")
sea_level <- aggregate(sea_level[,c(2,3)], by=list(format(sea_level$date,
format="%Y")), FUN=mean)
colnames(sea_level)[1] <- "year"
```

All of the five databases will be joined together into one dataframe, climate_change_df, based on the year. This will allow the data to be easily accessed and compared.

```{r climate_change_df Creation, echo=FALSE}
climate_change_df <- left_join(global_temp, co2_atmo, by="year")
climate_change_df <- left_join(climate_change_df, co2_emiss_global,
by="year")
climate_change_df <- left_join(climate_change_df, sea_level, by="year")
```

Final Dataset

```{r climate_change_df, echo=FALSE}
head(climate_change_df)
```

What Information is Not Self Evident

The uncertainties in the average temperature and global mean seal level (GMSL) along with the seasonally adjusted CO2 ppm will need to plotted to ensure that even if the levels are rising it is not entirely due to a rising uncertainty.

The temperature as a global level may be on the rise but to ensure that it is not a few regions skewing the data it will be necessary to pick a few random countries from different regions and check if they are rising as well and if that is similar to the global rate. If one region stands out it may be necessary to see if that matches with the CO2 emissions in that area.

What Are Different Ways You Could Look at This Data
The data will be looked at using scatter plots, histograms, and linear models. Through these avenues it should be possible to determine if there is a a global temperature change occurring and if so, if it is correlated to a rise in CO2 emissions caused by humans.

Data Summary to Answer Key Questions
The data was summarized into yearly averages and sums for each piece of data then joined into a single data frame for ease of use. Some cleaning of the data will still need to be done to account for missing data in specific years.

Types of Plots and Tables
Scatter plots of temperatures, sea level, CO2 levels and emissions will need to be plotted over time to see if there is truly a change.

A histogram of a temperature delta over a few decades of different regions and the global average will help to show if the global change is being driven by a few select regions.

Tables of the different coefficients and their respective p values in each linear model that is created will help to highlight if there is a strong relationship between the variables used in the model, for example, global temperature and sea level.

Machine Learning
If machine learning can be used to improve the model of predicting how the global temperature is changing then it will be implemented. If the basic level of machine learning does not provide an improvement over a linear model it will not be considered.

Questions for Future Steps
* Can machine learning be implemented to improve the accuracy of a linear model?
* If CO2 emissions and temperature change is correlated can causation be proved?