Project 1: Exploratory Data Analysis

MaterDolorosa Osueke - SDS348

03/14/2020

Introduction The two datasets that I have chosen to use for my analysis are Temp_increase_1850_2018_and co2_concentration_1850_2018_. These two datasets contain information on global temperature increase and global Co2 emissions. The first dataset looks into global temperature increase from 1850-2017. The second dataset looks into global Co2 emissions as well as emissions from some of the world's largest producers, the US and China from the years 1850-2017. I chose these datasets because I was interested in seeing the relationship between carbon emissions both globally and from the biggest producers and the way it correlates with global temperature increase. This information is important to understand global warming, changing climate, and physical as well as health impacts that we see today. The information in these datasets were acquired through "Our World Data" which collects this type of information. Some potential associations I expect to see are higher global temperatures as US and China Co2 emissions increase. I also expect to see a positive correlation between global Co2 emissions and China an US Co2 emissions. —

```
library(tidyverse)
library(tidyr)
library(dplyr)
library(devtools)
library(readxl)

tinytex::install_tinytex()
```

Tidying: Rearranging Wide/Long

```
Temp<-read_excel("Temp.xlsx")
Carbon<-read_excel("Carbon.xlsx")

longdata <- Carbon%>%pivot_longer(cols = c("US", "China"), names_to = "US and China", values_to = "ppm")

glimpse(longdata)
```

```
## Observations: 256
## Variables: 4
## $ Year           <dbl> 1850, 1850, 1851, 1851, 1854, 1854, 1855, 1855, 1857...
## $ CO2_global     <dbl> 284.00, 284.00, 287.13, 287.13, 288.05, 288.05, 285....
## $ `US and China` <chr> "US", "China", "US", "China", "US", "China", "US", "...
## $ ppm            <dbl> 19792928, 0, 24633072, 0, 26791168, 0, 30162048, 0, ...
```

From the Carbon dataset, I decided to tidy the columns called "US" and the column called "China" into a column named "US and China". I chose to use pivot longer because I thought it would make the dataset look nicer and shorter due to it looking a little wide with the previous US and China sections. I also thought it would make it easier to compare the Co2 emissions produced by China and the US over the years.

Joining/Merging

```
Global <- Temp%>%
  full_join(Carbon, by = 'Year')

Global_TC <- na.omit(Global)
glimpse(Global_TC)
```

```
## Observations: 128
## Variables: 5
## $ Year        <dbl> 1850, 1851, 1854, 1855, 1857, 1859, 1862, 1863, 1864, 1...
## $ Global_Temp <dbl> -0.373, -0.218, -0.248, -0.272, -0.461, -0.284, -0.524,...
## $ CO2_global  <dbl> 284.00, 287.13, 288.05, 285.57, 283.16, 286.63, 287.17,...
## $ US          <dbl> 19792928, 24633072, 26791168, 30162048, 33159200, 38160...
## $ China       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

The Type of join I performed on my 2 datasets was a full join. This join was performed because it returns all rows from both tables and joins matching keys in both the left and right datasets. Using this type of join also allows me to retain more cases. No cases where dropped with this join. Potential problems with this join is that not all the rows in each dataset have info needed for the other resulting in many rows with NA values. With that, I used na.omit to get rid of the rows containing NA values.

Wrangling

```
Global_TC%>%filter(Year=="1910")%>%glimpse()
```

```
## Observations: 1
## Variables: 5
## $ Year        <dbl> 1910
## $ Global_Temp <dbl> -0.49
## $ CO2_global  <dbl> 297.87
## $ US          <dbl> 235393680
## $ China       <dbl> 22731456
```

```
Global_TC%>%select(Year, US, China)%>%head(10)
```

```
## # A tibble: 10 x 3
##     Year       US China
##    <dbl>    <dbl> <dbl>
##  1  1850 19792928     0
##  2  1851 24633072     0
##  3  1854 26791168     0
##  4  1855 30162048     0
##  5  1857 33159200     0
##  6  1859 38160560     0
##  7  1862 40036528     0
##  8  1863 41055120     0
##  9  1864 41648688     0
## 10  1867 45320016     0
```

```
Global_TC%>%arrange(desc(Year), desc(Global_Temp), desc(CO2_global)%>%head(128))
```

```
## # A tibble: 128 x 5
##     Year Global_Temp CO2_global         US      China
##    <dbl>       <dbl>      <dbl>      <dbl>      <dbl>
##  1  2017       0.677       407. 5269529513 9838754028
##  2  2016       0.797       404. 5310861406 9704479432
##  3  2015       0.763       401. 5420804127 9716467840
##  4  2014       0.579       399. 5568759258 9820360492
##  5  2013       0.514       397. 5519612557 9796527160
##  6  2012       0.47        394. 5366730281 9633899303
##  7  2011       0.425       392. 5570706560 9388199234
##  8  2010       0.56        390. 5701075808 8500542695
##  9  2009       0.506       387. 5495394958 7758811768
## 10  2008       0.395       386. 5932775281 7375189907
## # ... with 118 more rows
```

```
Global_TC%>% group_by(Year)%>%summarize(avg_CO2 = mean(CO2_global, na.rm = T), sd_co2 = sd(CO2_global, r
```

```
## # A tibble: 128 x 3
##     Year avg_CO2 sd_co2
##    <dbl>   <dbl>  <dbl>
##  1  1850    284      NA
##  2  1851    287.     NA
##  3  1854    288.     NA
##  4  1855    286.     NA
##  5  1857    283.     NA
##  6  1859    287.     NA
##  7  1862    287.     NA
##  8  1863    285.     NA
##  9  1864    287.     NA
## 10  1867    285.     NA
## # ... with 118 more rows
```

```
Global_TC%>%mutate(USplusChina = US+China)%>%head(10)
```

```
## # A tibble: 10 x 6
##     Year Global_Temp CO2_global       US China USplusChina
```

```
##    <dbl>       <dbl>      <dbl>    <dbl> <dbl>       <dbl>
##  1  1850      -0.373       284  19792928     0    19792928
##  2  1851      -0.218       287. 24633072     0    24633072
##  3  1854      -0.248       288. 26791168     0    26791168
##  4  1855      -0.272       286. 30162048     0    30162048
##  5  1857      -0.461       283. 33159200     0    33159200
##  6  1859      -0.284       287. 38160560     0    38160560
##  7  1862      -0.524       287. 40036528     0    40036528
##  8  1863      -0.278       285. 41055120     0    41055120
##  9  1864      -0.494       287. 41648688     0    41648688
## 10  1867      -0.321       285. 45320016     0    45320016
```

```
Global_TC%>%summarize(as_US = sd(US, na.rm = T))
```

```
## # A tibble: 1 x 1
##         as_US
##         <dbl>
## 1 2231493862.
```

```
Global_TC%>%slice(1:10)
```

```
## # A tibble: 10 x 5
##     Year Global_Temp CO2_global       US China
##    <dbl>       <dbl>      <dbl>    <dbl> <dbl>
##  1  1850      -0.373       284  19792928     0
##  2  1851      -0.218       287. 24633072     0
##  3  1854      -0.248       288. 26791168     0
##  4  1855      -0.272       286. 30162048     0
##  5  1857      -0.461       283. 33159200     0
##  6  1859      -0.284       287. 38160560     0
##  7  1862      -0.524       287. 40036528     0
##  8  1863      -0.278       285. 41055120     0
##  9  1864      -0.494       287. 41648688     0
## 10  1867      -0.321       285. 45320016     0
```

```
Global_TC%>% group_by(Year)%>%summarize(avg_TEMP = mean(Global_Temp, na.rm = T), sd_temp = sd(Global_Ter
```

```
## # A tibble: 128 x 3
##     Year avg_TEMP sd_temp
##    <dbl>    <dbl>   <dbl>
##  1  1850   -0.373      NA
##  2  1851   -0.218      NA
##  3  1854   -0.248      NA
##  4  1855   -0.272      NA
##  5  1857   -0.461      NA
##  6  1859   -0.284      NA
##  7  1862   -0.524      NA
##  8  1863   -0.278      NA
##  9  1864   -0.494      NA
## 10  1867   -0.321      NA
## # ... with 118 more rows
```

```
Global_TC%>% group_by(Year)%>%summarize(med_US = median(US, na.rm = T), max_US = max(US, na.rm = T)%>%h
```

```
## # A tibble: 128 x 3
##     Year   med_US   max_US
##    <dbl>    <dbl>    <dbl>
##  1  1850 19792928 19792928
##  2  1851 24633072 24633072
##  3  1854 26791168 26791168
##  4  1855 30162048 30162048
##  5  1857 33159200 33159200
##  6  1859 38160560 38160560
##  7  1862 40036528 40036528
##  8  1863 41055120 41055120
##  9  1864 41648688 41648688
## 10  1867 45320016 45320016
## # ... with 118 more rows
```

```
Global_TC%>% group_by(Year)%>%summarize(med_china = median(China, na.rm = T), min_china = min(China, na
```

```
## # A tibble: 128 x 3
##     Year med_china min_china
##    <dbl>     <dbl>     <dbl>
##  1  1850         0         0
##  2  1851         0         0
##  3  1854         0         0
##  4  1855         0         0
##  5  1857         0         0
##  6  1859         0         0
##  7  1862         0         0
##  8  1863         0         0
##  9  1864         0         0
## 10  1867         0         0
## # ... with 118 more rows
```

```
Global_TC%>%summarize(as_Temp = sd(Global_Temp, na.rm = T))
```

```
## # A tibble: 1 x 1
##   as_Temp
##     <dbl>
## 1   0.311
```

```
Global_TC%>%summarize(as_Co2 = sd(CO2_global, na.rm = T))
```

```
## # A tibble: 1 x 1
##   as_Co2
##    <dbl>
## 1   33.1
```

```
Global_TC%>%summarize(as_China = sd(China, na.rm = T))
```

```
## # A tibble: 1 x 1
##      as_China
##         <dbl>
## 1 2641838000.
```

```
Global_TC%>%mutate(co2.dat = case_when(CO2_global>400 ~ "high", CO2_global<400 & CO2_global>300 ~ "med"
```

```
## # A tibble: 10 x 6
##     Year Global_Temp CO2_global       US China co2.dat
##    <dbl>       <dbl>      <dbl>    <dbl> <dbl> <chr>
##  1  1850      -0.373       284  19792928     0 low
##  2  1851      -0.218       287. 24633072     0 low
##  3  1854      -0.248       288. 26791168     0 low
##  4  1855      -0.272       286. 30162048     0 low
##  5  1857      -0.461       283. 33159200     0 low
##  6  1859      -0.284       287. 38160560     0 low
##  7  1862      -0.524       287. 40036528     0 low
##  8  1863      -0.278       285. 41055120     0 low
##  9  1864      -0.494       287. 41648688     0 low
## 10  1867      -0.321       285. 45320016     0 low
```

```
Global_TC%>%mutate(Temp.dat = case_when(Global_Temp>.5 ~ "high", Global_Temp<.5 & Global_Temp>0 ~ "med"
```

```
## # A tibble: 128 x 6
## # Groups:   Year [128]
##     Year Global_Temp CO2_global       US China Temp.dat
##    <dbl>       <dbl>      <dbl>    <dbl> <dbl> <chr>
##  1  1850      -0.373       284  19792928     0 low
##  2  1851      -0.218       287. 24633072     0 low
##  3  1854      -0.248       288. 26791168     0 low
##  4  1855      -0.272       286. 30162048     0 low
##  5  1857      -0.461       283. 33159200     0 low
##  6  1859      -0.284       287. 38160560     0 low
##  7  1862      -0.524       287. 40036528     0 low
##  8  1863      -0.278       285. 41055120     0 low
##  9  1864      -0.494       287. 41648688     0 low
## 10  1867      -0.321       285. 45320016     0 low
## # ... with 118 more rows
```

With the filter code, I set the dataset to filter by Year specifically 1910. This showed summary statistics for global temperature increase, global co2 emissions and co2 emissions in the US and China at that time. With the select function, I was able to select for the Year, US, and China information. With the arrange code, I was able to arrange Year, Global Temp, and Co2 global in descending order. I used the group by function to allow the dataset to correspond by Year and the summarize was used to summarize this data giving mean and standard deviation of global Co2 emissions. I performed similar group by functions for each variable summarizing with functions such as median, mean, max, min and standard deviation. I used mutate to create a new column called "USplusChina" where I was able to get the sum of US and China Co2 emissions in a new column. I also used the code slice which showed observations 1-10 of the dataset. The summarize function was also used to summarize standard deviation information for US co2 emissions. Mutate was also used to create new columns showing low, medium, and high values for different variables.

Visualizing

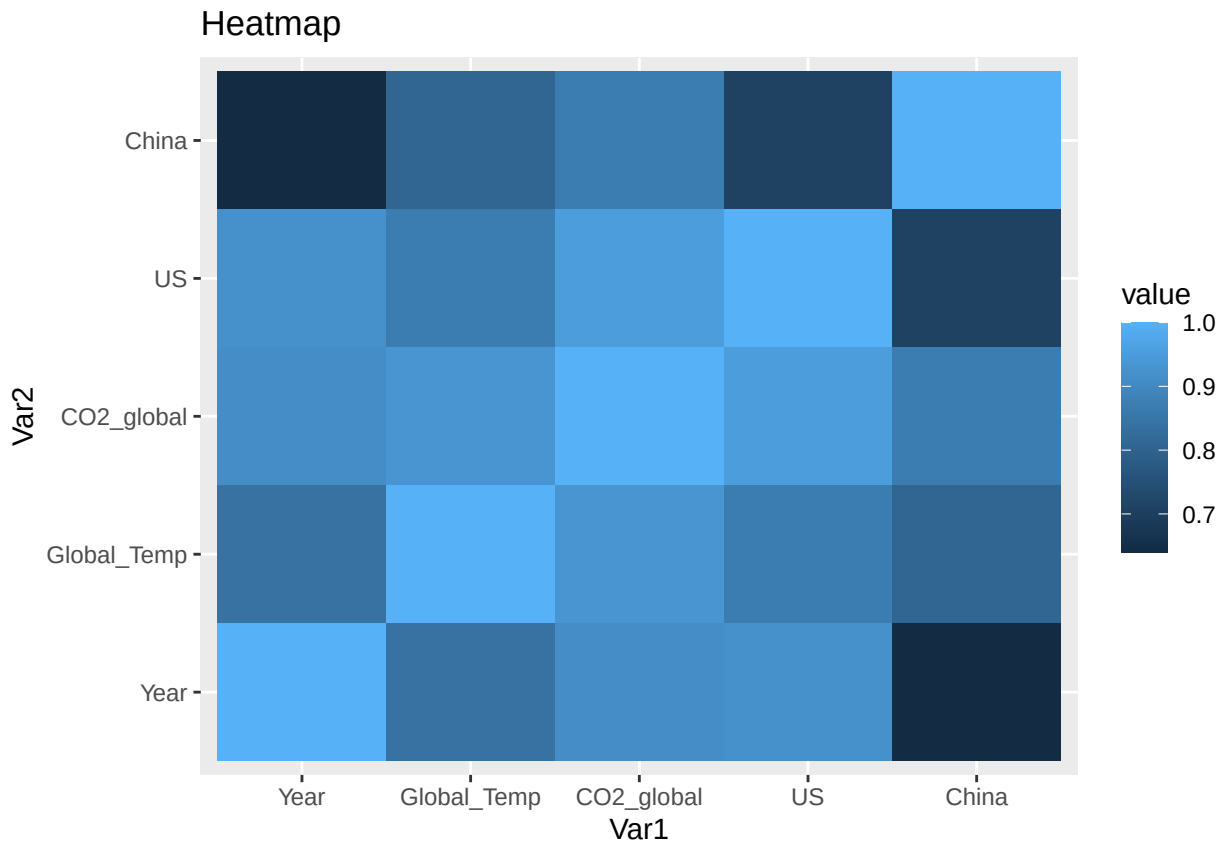
```
library(ggplot2)
carbtemp <- round(cor(Global_TC),2)
head(carbtemp)
```

```
##             Year Global_Temp CO2_global   US China
## Year        1.00        0.84       0.91 0.92  0.64
## Global_Temp 0.84        1.00       0.93 0.87  0.81
## CO2_global  0.91        0.93       1.00 0.95  0.87
## US          0.92        0.87       0.95 1.00  0.71
## China       0.64        0.81       0.87 0.71  1.00
```

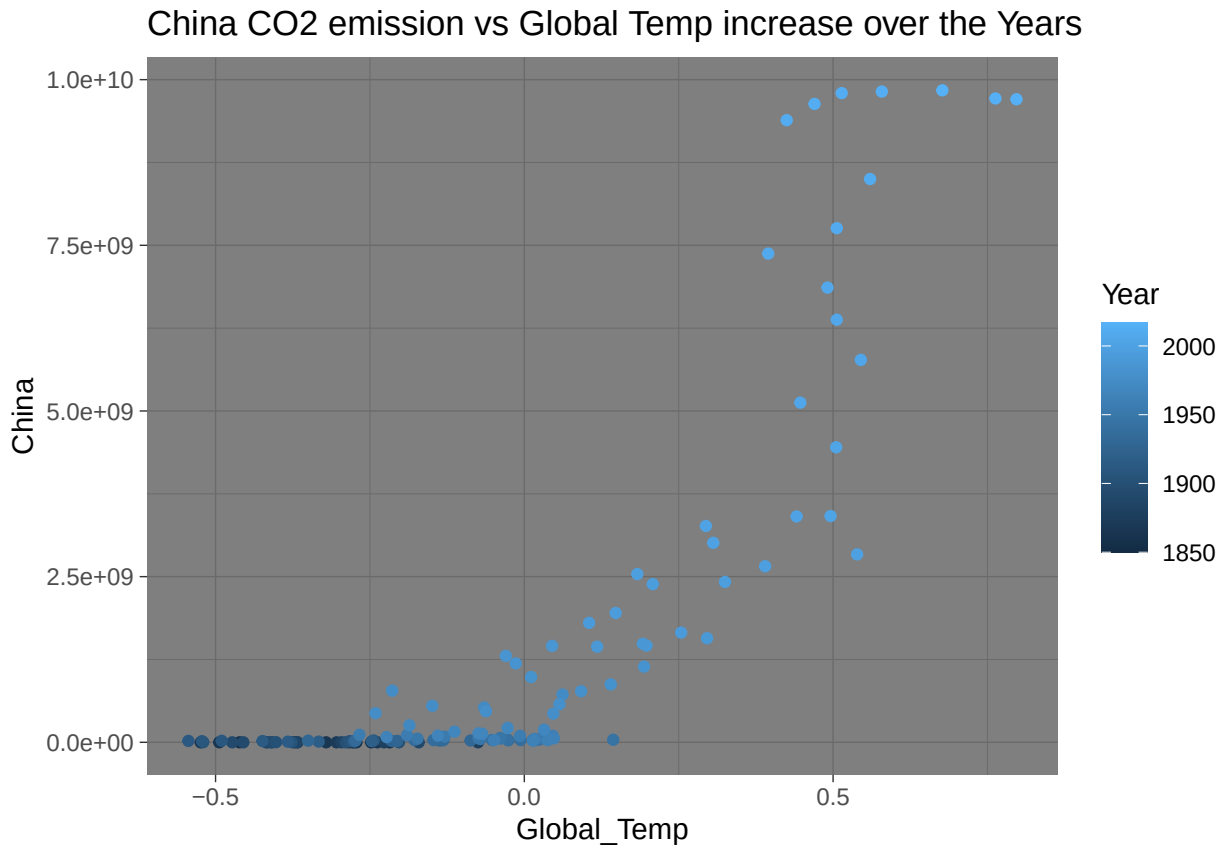
```
library(reshape2)
melted_carbtemp <- melt(carbtemp)
head(melted_carbtemp)
```

```
##          Var1        Var2 value
## 1        Year        Year  1.00
## 2 Global_Temp        Year  0.84
## 3  CO2_global        Year  0.91
## 4          US        Year  0.92
## 5       China        Year  0.64
## 6        Year Global_Temp  0.84
```

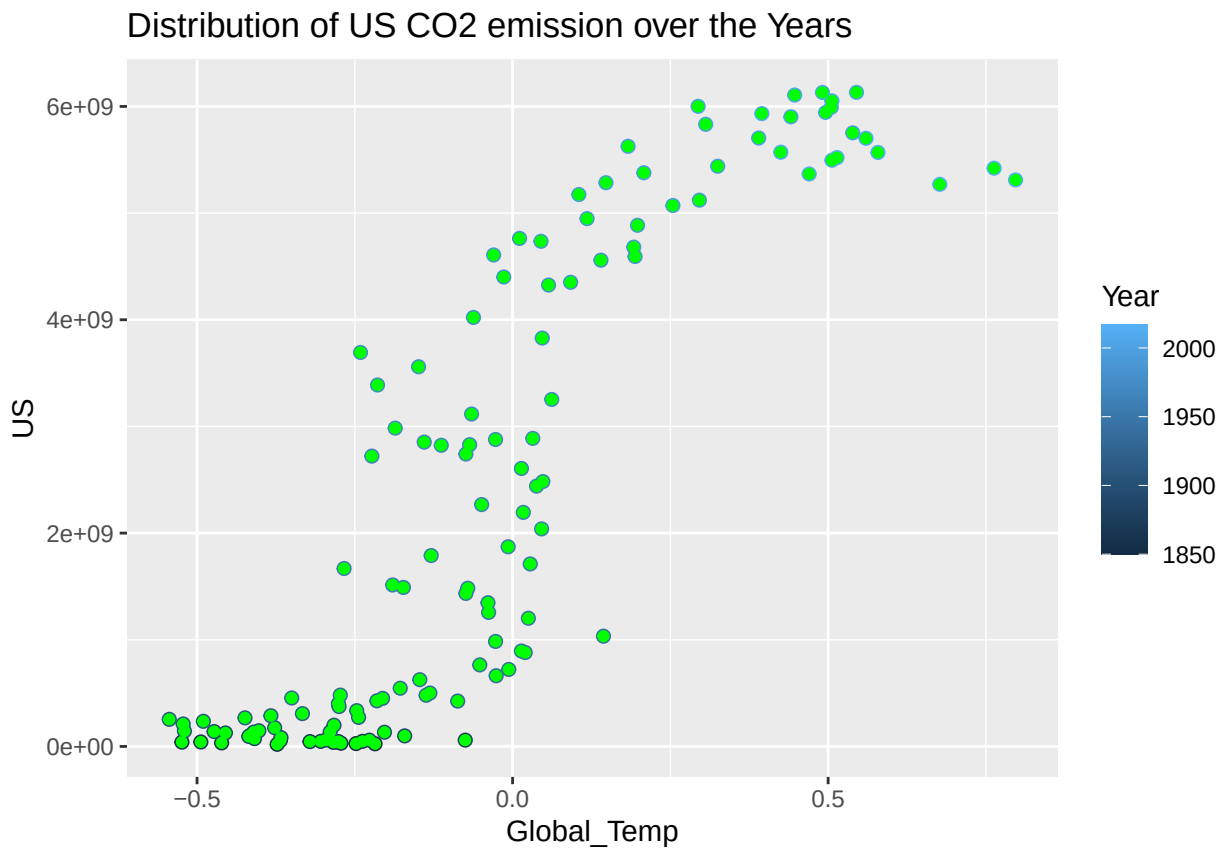
```
ggplot(data = melted_carbtemp, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()+ggtitle("Heatmap")
```



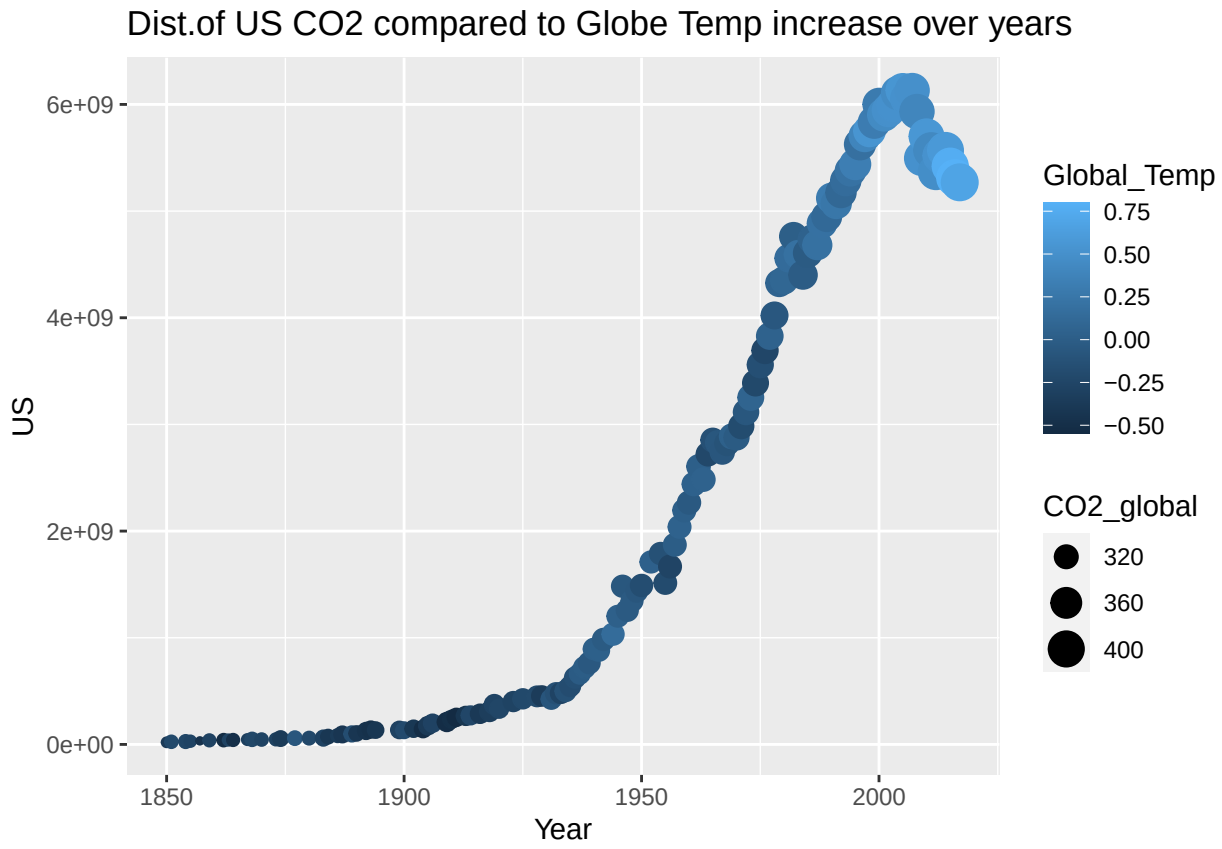

```
ggplot(data = Global_TC, aes(x=Global_Temp, y=China, color=Year))+geom_point()+ggtitle("China CO2 emissi
```




```
ggplot(data = Global_TC, aes(x=Global_Temp, y=US, color=Year))+geom_point(size=2, stat = "summary")+ gg
```




```
ggplot(Global_TC, aes(Year,US))+geom_point(aes(color=Global_Temp, size=CO2_global))+ggtitle("Dist.of US
```



In the first graph a heat map was created maping the different variables in the dataset. The cells show the degree of correlation ranging from 0 to 1. In the next graph titled "China CO2 emission vs Global Temp increase over the Years", it showed the relationship between Global Temperature increase across the world and the amount of CO2 emissions produced by China from the years 1850-2017. Based on the graph we see that global temperature increase (>0) began to rise around the time that China's CO2 emissions increased dramatically. There seems to be a positive correlation as the graph has an upward positive slope.

In the second graph titled "Distribution of US CO2 emission over the Years", it graphed the global temperature increase on the x-axis, the US Co2 emissions on the Y-axis and analyzed those variables over the years. Based on this graph the we see that most data is around 2,000000000 ppm of Co2 emission from the US. In addition, we can see that there is a positive correlation between global temperature increase and US Co2 emissions over the years from 1850-2017. The mapping of the third graph allows us to see that ther is an increase in global temp as US emmissions of CO2 have increased over the years. As we look past the year 2000 we see a slight decrease in the US co2 emmisison however the global temperature still shows increase.

Dimensionality Reduction

```r
library(cluster)
kmeans1 <- Global_TC %>% kmeans(3)
kmeans1
```

```
## K-means clustering with 3 clusters of sizes 14, 82, 32
##
## Cluster means:
##     Year Global_Temp CO2_global         US      China
## 1 2010.5   0.5482143   391.3021 5684149717 8262128099
## 2 1918.5  -0.2129878   303.3965  844935631   37321549
## 3 1987.5   0.1522187   350.1984 4810263720 1686608067
##
## Clustering vector:
##   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [ reached getOption("max.print") -- omitted 28 entries ]
##
## Within cluster sum of squares by cluster:
## [1] 3.860966e+19 7.377304e+19 5.754559e+19
##  (between_SS / total_SS =  88.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
kmeans1$cluster
```

```
##   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [ reached getOption("max.print") -- omitted 28 entries ]
```

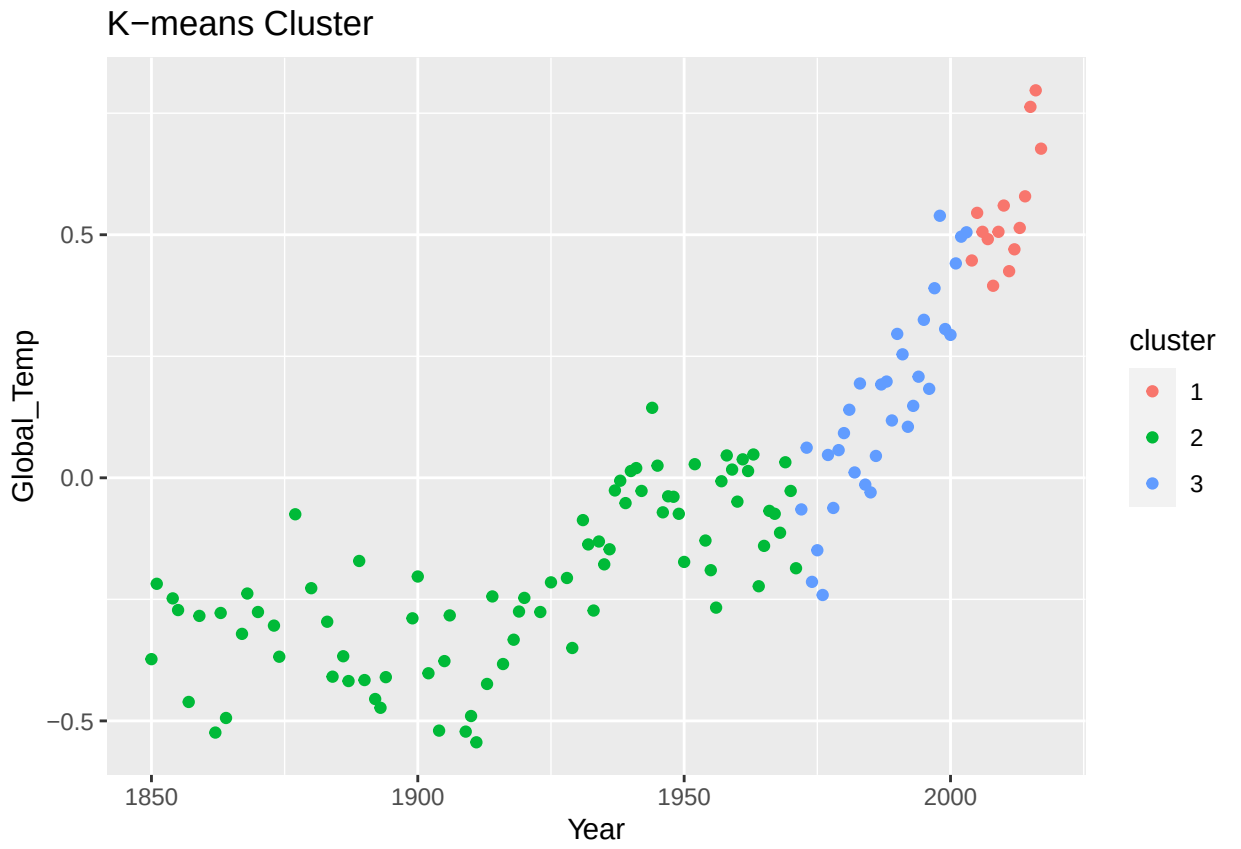
```r
kmeans1$centers
```

```
##     Year Global_Temp CO2_global         US      China
## 1 2010.5   0.5482143   391.3021 5684149717 8262128099
## 2 1918.5  -0.2129878   303.3965  844935631   37321549
## 3 1987.5   0.1522187   350.1984 4810263720 1686608067
```

```r
kmeans1$size
```

```
## [1] 14 82 32
```

```r
kmeansclust <- Global_TC %>% mutate(cluster=as.factor(kmeans1$cluster))
kmeansclust %>% ggplot(aes(Year,Global_Temp,CO2_global,US,China, color=cluster)) + geom_point()+ggtitle
```



For my data I performed a K-means cluster where I asked it to find 3 clusters using variables: Year, Global_Temp, CO2_global, US,China in my dataset. K means works by picking three initial cluster points and assigning points to each cluster based on the distances. With my graph we are able to see these three cluster groups. The first group is from year 1850-1866 with its center being ~ 1918.5 years and its cluster size being 82,the second cluster is from 1866-2000 with its center being ~ 1987.5 years and cluster size 32, and the third cluster is from 2000 up with its center being ~2010.5 and cluster size being 14. The number of clusters, cluster size, and cluster centers could be seen through the use of function: kmeans1$cluster, kmeans1$centers, kmeans1$size.