USA Greenhouse Gas Emissions

2023-03-11

Introduction

The complete, precise emission totals in this dataset will serve as the basis for the 2021 USA Statewide Greenhouse Gas Emissions Report, a crucial component of the State's climate change policy. This emission inventory's objectives are to satisfy the requirements of the Climate Leadership and Community Protection Act (CLCPA), monitor the development of the effort to reduce greenhouse gas emissions, and broaden public access to information about greenhouse gases. This dataset contains the most current estimate of annual emissions from 1990 to the most recent year for which data are accessible.

Data Source Link

https://data.ny.gov/api/views/5i6e-asw6/rows.csv?accessType=DOWNLOAD

Mission Goals:

To ascertain the total quantity of gases released into the atmosphere, the polluting industry, and the yearly tendency. We will therefore extract this information from the aforementioned data collection.

The Findings From This Study Are:

Statistics, linear regression, polynomial regression, and data plotting details were collected for each rule.

1) Total Gas Emissions Yearly.
2) Gases Releases From Various Sectors.
3) Various Gases Are Emitted Each Year.

Main Packages And Libraries:

```
install.packages("tidyr", repos = "http://cran.us.r-project.org")
install.packages("factoextra", repos="http://cran.us.r-project.org")
install.packages("dplyr", repos = "http://cran.us.r-project.org")
install.packages("knitr", repos = "http://cran.us.r-project.org")
install.packages("ggplot2", repos = "http://cran.us.r-project.org")

library(factoextra)
library(tidyr)
library(dplyr)
library(lubridate)
library(stringr)
library(magrittr)
library(knitr)
library(ggplot2)
```

Reading The Dataset:

```
core_df =  read.csv("C:/Users/Public/USA_Greenhouse_Gas_Emissions.csv")
head(core_df)
```

```
##   Gross Net Conventional.Accounting Economic.Sector Sector
## 1   Yes Yes                     Yes        Industry Energy
## 2    No Yes                     Yes   Net Emissions  AFOLU
## 3    No Yes                     Yes   Net Emissions  AFOLU
## 4    No Yes                     Yes   Net Emissions  AFOLU
## 5    No Yes                     Yes   Net Emissions  AFOLU
## 6    No Yes                     Yes   Net Emissions  AFOLU
##                Category Sub.Category.1 Sub.Category.2           Sub.Category.3
## 1 Other Fossil Fuel Use     Industrial Not Applicable              Natural Gas
## 2 Net Emission Removals       Land Use         Forest Forests Remaining Forests
## 3 Net Emission Removals       Land Use         Forest Forests Remaining Forests
## 4 Net Emission Removals       Land Use         Forest Forests Remaining Forests
## 5 Net Emission Removals       Land Use         Forest Forests Remaining Forests
## 6 Net Emission Removals       Land Use         Forest Forests Remaining Forests
##   Year Gas MT.CO2e.AR5.20.yr MT.CO2e.AR4.100.yr
## 1 1990 CH4                 0                  0
## 2 1990 CO2         -27910000          -27910000
## 3 1991 CO2         -27810000          -27810000
## 4 1992 CO2         -27700000          -27700000
## 5 1993 CO2         -27570000          -27570000
## 6 1994 CO2         -27450000          -27450000
```

```
str(core_df)
```

```
## 'data.frame':    14162 obs. of  13 variables:
##  $ Gross                  : chr  "Yes" "No" "No" "No" ...
##  $ Net                    : chr  "Yes" "Yes" "Yes" "Yes" ...
```

```
##  $ Conventional.Accounting: chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Economic.Sector        : chr  "Industry" "Net Emissions" "Net Emissions" "Net Emissions" ...
##  $ Sector                 : chr  "Energy" "AFOLU" "AFOLU" "AFOLU" ...
##  $ Category               : chr  "Other Fossil Fuel Use" "Net Emission Removals" "Net Emission Remova
##  $ Sub.Category.1         : chr  "Industrial" "Land Use" "Land Use" "Land Use" ...
##  $ Sub.Category.2         : chr  "Not Applicable" "Forest" "Forest" "Forest" ...
##  $ Sub.Category.3         : chr  "Natural Gas" "Forests Remaining Forests" "Forests Remaining Forest
##  $ Year                   : int  1990 1990 1991 1992 1993 1994 1995 1996 1997 1998 ...
##  $ Gas                    : chr  "CH4" "CO2" "CO2" "CO2" ...
##  $ MT.CO2e.AR5.20.yr      : int  0 -27910000 -27810000 -27700000 -27570000 -27450000 -27320000 -2720
##  $ MT.CO2e.AR4.100.yr     : int  0 -27910000 -27810000 -27700000 -27570000 -27450000 -27320000 -2720
```

Replacing Spaces With NAs:

```
core_df[core_df==''] = NA
```

Renaming The Columns:

```
core_df = core_df %>%select_all(~gsub("\\s+|\\.", "_", .))
str(core_df)
```

```
## 'data.frame':    14162 obs. of  13 variables:
##  $ Gross                  : chr  "Yes" "No" "No" "No" ...
##  $ Net                    : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Conventional_Accounting: chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Economic_Sector        : chr  "Industry" "Net Emissions" "Net Emissions" "Net Emissions" ...
##  $ Sector                 : chr  "Energy" "AFOLU" "AFOLU" "AFOLU" ...
##  $ Category               : chr  "Other Fossil Fuel Use" "Net Emission Removals" "Net Emission Remova
##  $ Sub_Category_1         : chr  "Industrial" "Land Use" "Land Use" "Land Use" ...
##  $ Sub_Category_2         : chr  "Not Applicable" "Forest" "Forest" "Forest" ...
##  $ Sub_Category_3         : chr  "Natural Gas" "Forests Remaining Forests" "Forests Remaining Forest
##  $ Year                   : int  1990 1990 1991 1992 1993 1994 1995 1996 1997 1998 ...
##  $ Gas                    : chr  "CH4" "CO2" "CO2" "CO2" ...
##  $ MT_CO2e_AR5_20_yr      : int  0 -27910000 -27810000 -27700000 -27570000 -27450000 -27320000 -2720
##  $ MT_CO2e_AR4_100_yr     : int  0 -27910000 -27810000 -27700000 -27570000 -27450000 -27320000 -2720
```

Transforming The Data:

```
#Removing the NA from the data frame.
core_dft1 = na.omit(core_df)

#Selecting required columns

core_dft2 = select(core_dft1,Gross,MT_CO2e_AR5_20_yr,Economic_Sector,MT_CO2e_AR4_100_yr,Year,Category,N

head(core_dft2)
```

```
##   Gross MT_CO2e_AR5_20_yr Economic_Sector MT_CO2e_AR4_100_yr Year
## 1   Yes                 0        Industry                  0 1990
## 2    No         -27910000   Net Emissions          -27910000 1990
## 3    No         -27810000   Net Emissions          -27810000 1991
## 4    No         -27700000   Net Emissions          -27700000 1992
## 5    No         -27570000   Net Emissions          -27570000 1993
## 6    No         -27450000   Net Emissions          -27450000 1994
##                Category Net Gas
## 1 Other Fossil Fuel Use Yes CH4
## 2 Net Emission Removals Yes CO2
## 3 Net Emission Removals Yes CO2
## 4 Net Emission Removals Yes CO2
## 5 Net Emission Removals Yes CO2
## 6 Net Emission Removals Yes CO2
```

Total Gas Emissions Yearly:

```
core_dft3 = filter(core_dft2, Gross== "Yes")

core_dft4 = aggregate(core_dft3$MT_CO2e_AR5_20_yr, list(core_dft3$Year), FUN=sum)

core_dft4 = rename(core_dft4,"MT_CO2e_AR5_20_yr" = "x", "Year" = "Group.1")

kable(summary(select(core_dft4,MT_CO2e_AR5_20_yr,Year)),row.names = FALSE,caption = "Total Gases Emissi
```

Table 1: Total Gases Emission Statistics

MT_CO2e_AR5_20_yr	Year
Min. :373250919	Min. :1990
1st Qu.:405862658	1st Qu.:1997
Median :419285484	Median :2004
Mean :422904313	Mean :2004
3rd Qu.:445625892	3rd Qu.:2012
Max. :463424415	Max. :2019

Total Gases Emission Linear Regression For Each Year:

```
corelm1_dft = lm(MT_CO2e_AR5_20_yr~Year, data = core_dft4)
summary(corelm1_dft)
```

```
##
## Call:
## lm(formula = MT_CO2e_AR5_20_yr ~ Year, data = core_dft4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -38905426 -17788614  -1476419  18414018  36284603
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2985888176  948476034   3.148  0.00388 **
## Year          -1278615     473169  -2.702  0.01157 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22430000 on 28 degrees of freedom
## Multiple R-squared:  0.2068, Adjusted R-squared:  0.1785
## F-statistic: 7.302 on 1 and 28 DF,  p-value: 0.01157
```

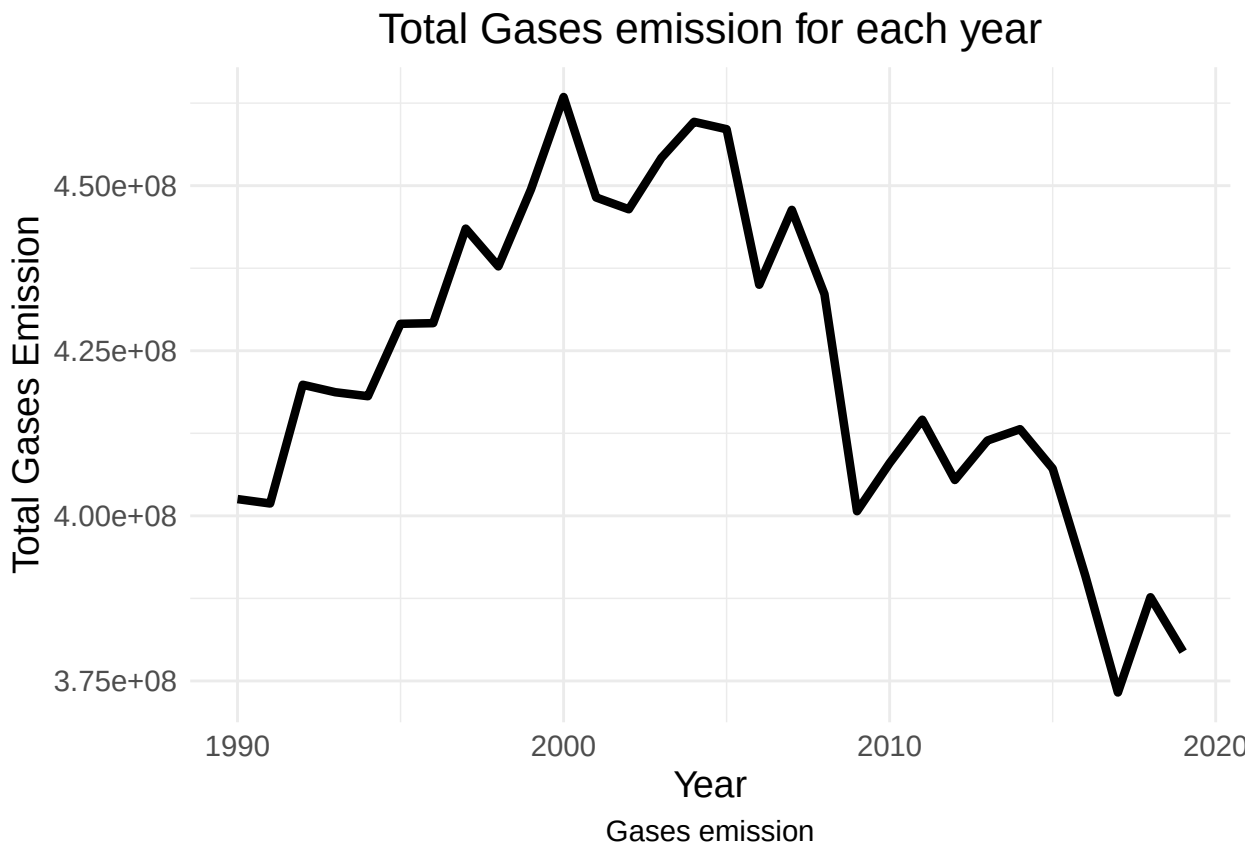
Total Gases Emission Polynomial Regression For Each Year:

```
corelm1_dft = lm(MT_CO2e_AR5_20_yr ~ poly(Year, 2, raw = TRUE),data = core_dft4)
summary(corelm1_dft)
```

```
##
## Call:
## lm(formula = MT_CO2e_AR5_20_yr ~ poly(Year, 2, raw = TRUE), data = core_dft4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -31772440  -6105386  -1076177   7278485  19441794
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.123e+12  1.256e+11  -8.947 1.46e-09 ***
## poly(Year, 2, raw = TRUE)1  1.123e+09  1.253e+08   8.961 1.42e-09 ***
## poly(Year, 2, raw = TRUE)2 -2.803e+05  3.125e+04  -8.971 1.38e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11450000 on 27 degrees of freedom
## Multiple R-squared:  0.8007, Adjusted R-squared:  0.786
## F-statistic: 54.25 on 2 and 27 DF,  p-value: 3.483e-10
```

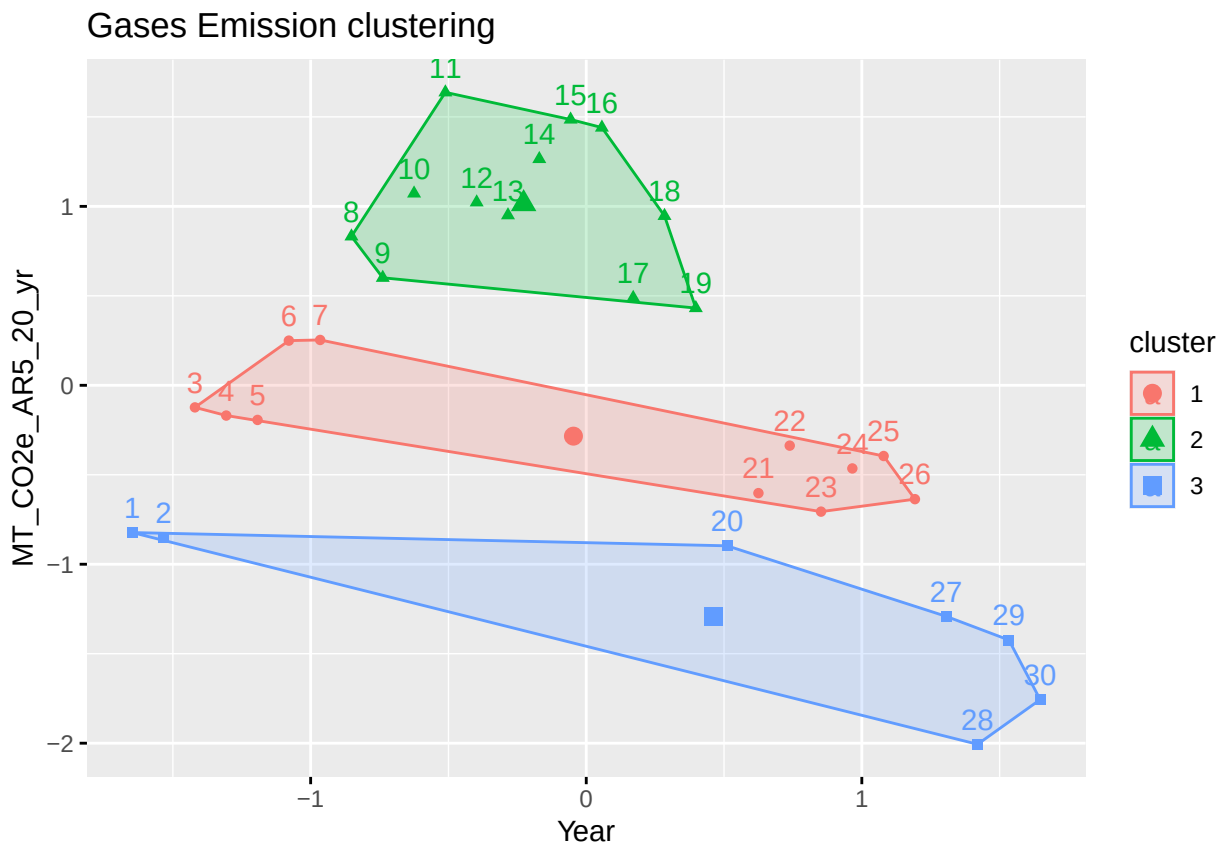
Total Gases Emission Plotting:

```
ggplot(core_dft4,aes(x=Year,y=MT_CO2e_AR5_20_yr)) +geom_line(size=1.5) +
    labs(x = "Year", y = "Total Gases Emission",caption="Gases emission") +
    ggtitle(paste0("Total Gases emission for each year"))+
    theme_minimal()+
    theme(legend.position = "right",
          plot.caption = element_text(hjust = 0.5),
          plot.title = element_text(hjust = 0.5,size=16),
          text = element_text(size=14))+
    scale_fill_brewer(palette="Set3")
```



Total Gases Emission Clustering:

```
coreclst1_dft = select(core_dft4,Year,MT_CO2e_AR5_20_yr)
coreclst1_dft1 = kmeans(coreclst1_dft, centers = 3)
fviz_cluster(coreclst1_dft1, data = coreclst1_dft,title="Gases Emission clustering")
```



Gases Releases From Various Sectors:

```
core_dft5 = filter(core_dft2, Gross== "Yes")

core_dft6 = aggregate(core_dft5$MT_CO2e_AR5_20_yr, list(core_dft5$Year,core_dft5$Economic_Sector), FUN=s

core_dft6 = rename(core_dft6,"Gas_Emissions" = "x", "Year" = "Group.1","Economic_Sector" = "Group.2")

kable(summary(select(core_dft6,Gas_Emissions,Year)),row.names = FALSE,caption = "Gases Releases From Va
```

Table 2: Gases Releases From Various Sectors Statistics

Gas_Emissions	Year
Min. : 15007936	Min. :1990
1st Qu.: 43086676	1st Qu.:1997
Median : 58625092	Median :2004
Mean : 70484052	Mean :2004
3rd Qu.:107441117	3rd Qu.:2012
Max. :141492310	Max. :2019

Checking The Linear Regression For Gases Releases From Various Sectors:

```
corelm2_dft = lm(Gas_Emissions~Year, data = core_dft6)
summary(corelm2_dft)
```

```
##
## Call:
## lm(formula = Gas_Emissions ~ Year, data = core_dft6)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -58249908 -26853948 -12144371  38600599  70901707
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 497648029  653447618   0.762    0.447
## Year          -213103     325987  -0.654    0.514
##
## Residual standard error: 37860000 on 178 degrees of freedom
## Multiple R-squared:  0.002395,   Adjusted R-squared:  -0.003209
## F-statistic: 0.4273 on 1 and 178 DF,  p-value: 0.5141
```

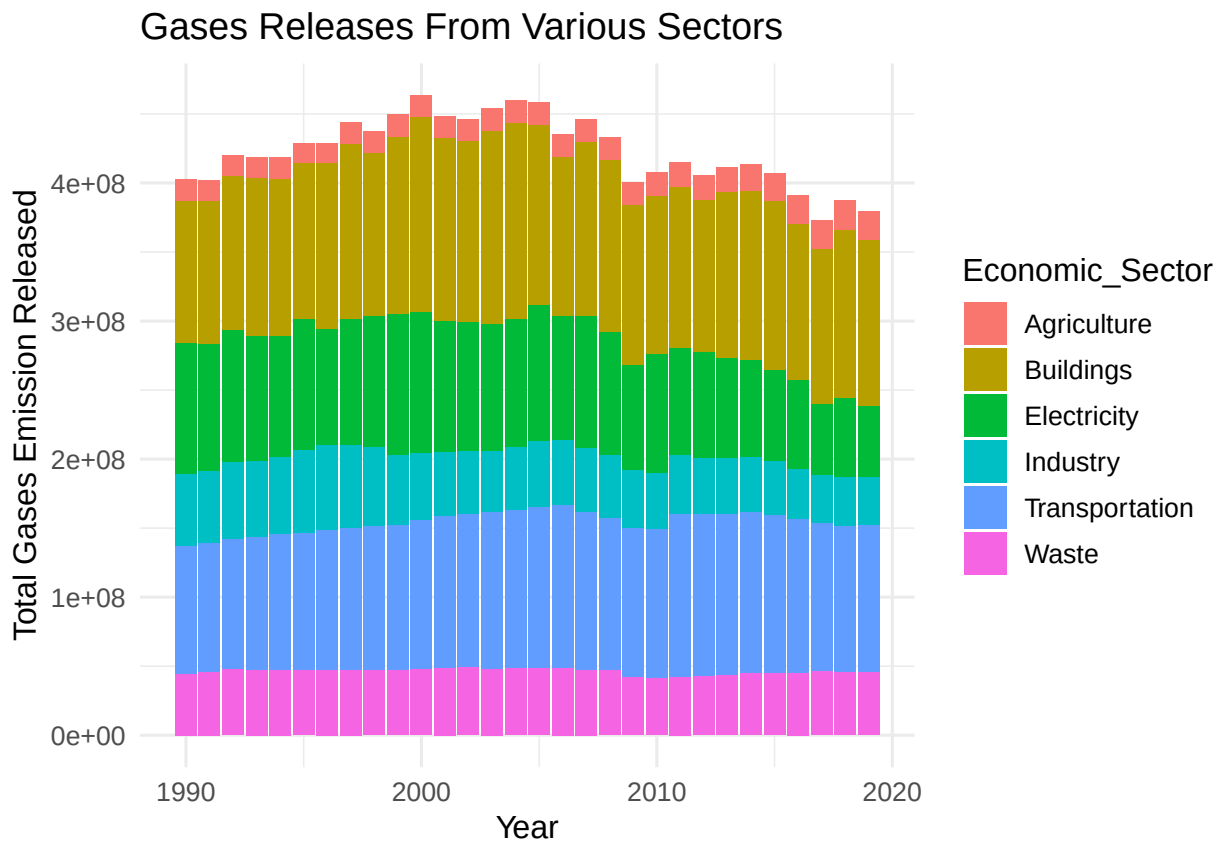
Checking the polynomial regression for Gases Releases From Various Sectors:

```
coreplm2_dft = lm(Gas_Emissions ~ poly(Year, 2, raw = TRUE),data = core_dft6)
summary(coreplm2_dft)
```

```
##
## Call:
## lm(formula = Gas_Emissions ~ poly(Year, 2, raw = TRUE), data = core_dft6)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -58341781 -27476809 -11276953  36324823  67413193
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.872e+11  1.694e+11  -1.105    0.270
## poly(Year, 2, raw = TRUE)1  1.871e+08  1.690e+08   1.107    0.270
## poly(Year, 2, raw = TRUE)2 -4.672e+04  4.215e+04  -1.108    0.269
##
## Residual standard error: 37830000 on 177 degrees of freedom
## Multiple R-squared:  0.009272,   Adjusted R-squared:  -0.001923
## F-statistic: 0.8282 on 2 and 177 DF,  p-value: 0.4385
```

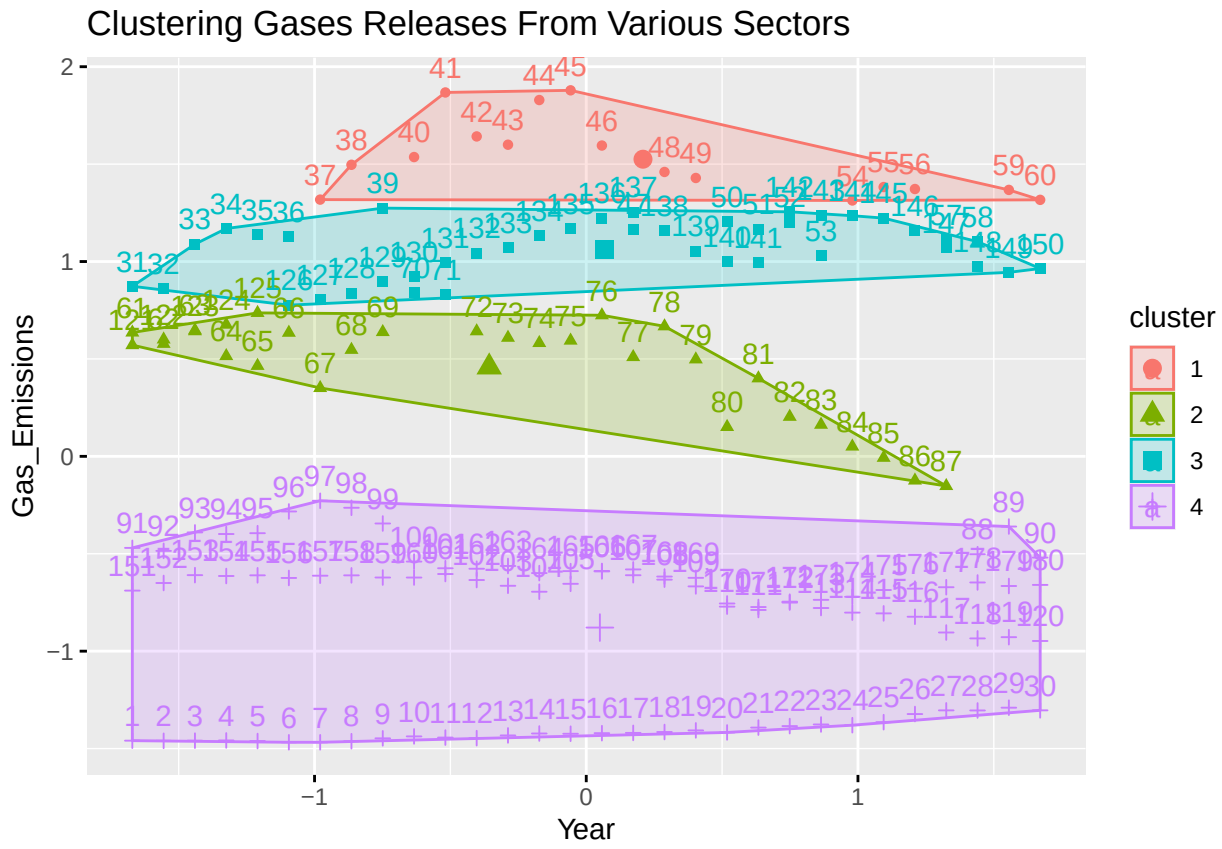
Plotting for Gases Releases From Various Sectors:

```
ggplot(core_dft6,aes(x=Year, y=Gas_Emissions,fill=Economic_Sector)) +
  ggtitle("Gases Releases From Various Sectors") +
  xlab("Year") +
  ylab("Total Gases Emission Released") +
  theme_minimal(base_size = 12) +
  geom_bar(stat="identity") +
  scale_color_discrete(name = "Economic SectorS")
```



Clustering for Gases Releases From Various Sectors:

```
coreclst2_dft = select(core_dft6,Year,Gas_Emissions)
coreclst2_dft1 = kmeans(coreclst2_dft, centers = 4)
fviz_cluster(coreclst2_dft1, data = coreclst2_dft,title="Clustering Gases Releases From Various Sectors
```



Checking The Various Gases Are Emitted Each Year:

```
core_dft7 = filter(core_dft2, Gross== "Yes")

core_dft8 = aggregate(core_dft7$MT_CO2e_AR5_20_yr, list(core_dft7$Year,core_dft7$Gas), FUN=sum)

core_dft8 = rename(core_dft8,"Gas_Emissions" = "x", "Year" = "Group.1","Gas" = "Group.2")

kable(summary(select(core_dft8,Gas_Emissions,Year)),row.names = FALSE,caption = "Statistics Of Various
```

Table 3: Statistics Of Various Gases Are Emitted Each Year

Gas_Emissions	Year
Min. : 489	Min. :1990
1st Qu.: 333114	1st Qu.:1997
Median : 3672371	Median :2004
Mean : 52863039	Mean :2004
3rd Qu.: 48835146	3rd Qu.:2012
Max. :279668952	Max. :2019

Checking The Linear Regression For Various Gases Are Emitted Each Year:

```
corelm3_dft = lm(Gas_Emissions~Year, data = core_dft8)
summary(corelm3_dft)
```

```
##
## Call:
## lm(formula = Gas_Emissions ~ Year, data = core_dft8)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -55180040 -51446525 -48931319  -2869148 226086692
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  373236022 1321325097   0.282    0.778
## Year           -159827     659173  -0.242    0.809
##
## Residual standard error: 88390000 on 238 degrees of freedom
## Multiple R-squared:  0.000247,   Adjusted R-squared:  -0.003954
## F-statistic: 0.05879 on 1 and 238 DF,  p-value: 0.8086
```

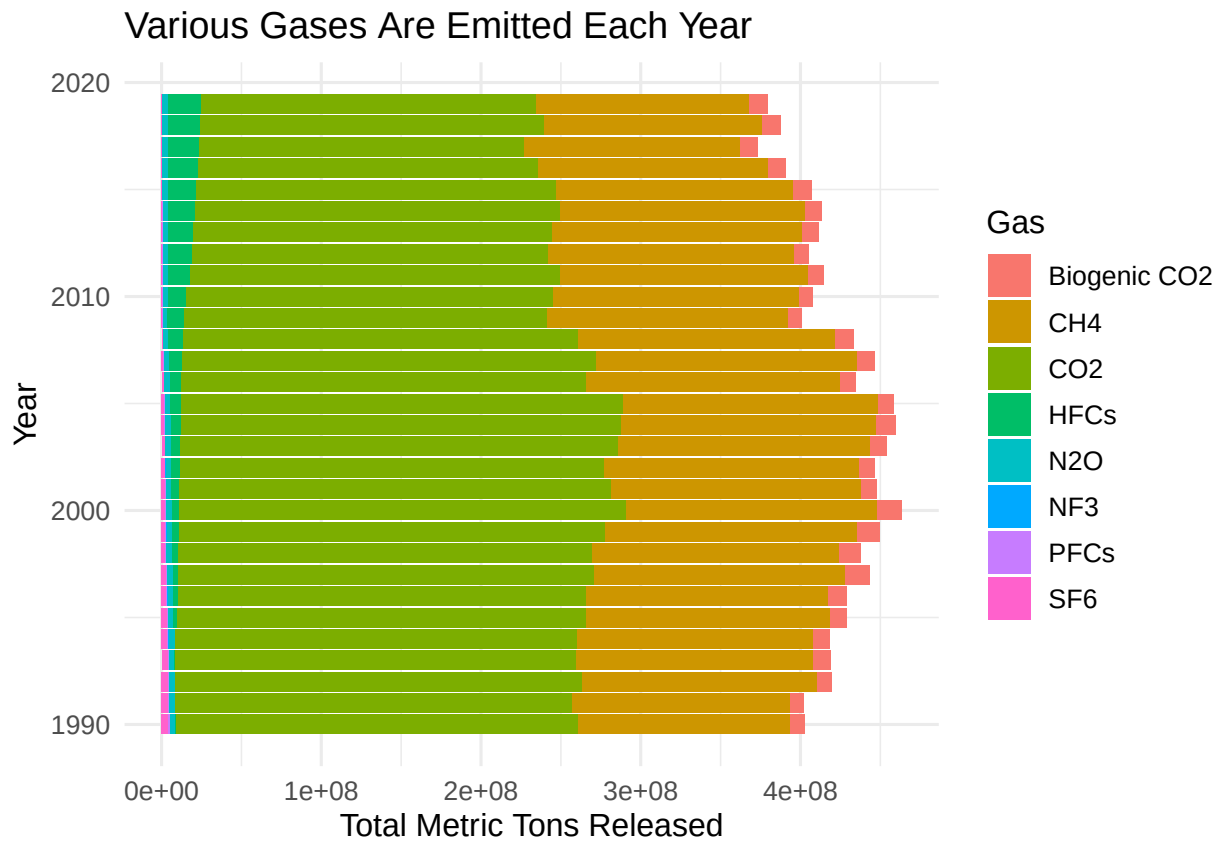
Checking The Polynomial Regression For Various Gases Are Emitted Each Year:

```
coreplm3_dft = lm(Gas_Emissions ~ poly(Year, 2, raw = TRUE),data = core_dft8)
summary(coreplm3_dft)
```

```
##
## Call:
## lm(formula = Gas_Emissions ~ poly(Year, 2, raw = TRUE), data = core_dft8)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -55660046 -52015013 -47810736   1873050 224171124
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.404e+11  3.433e+11  -0.409    0.683
## poly(Year, 2, raw = TRUE)1  1.403e+08  3.425e+08   0.410    0.682
## poly(Year, 2, raw = TRUE)2 -3.504e+04  8.544e+04  -0.410    0.682
##
## Residual standard error: 88540000 on 237 degrees of freedom
## Multiple R-squared:  0.000956,   Adjusted R-squared:  -0.007475
## F-statistic: 0.1134 on 2 and 237 DF,  p-value: 0.8928
```

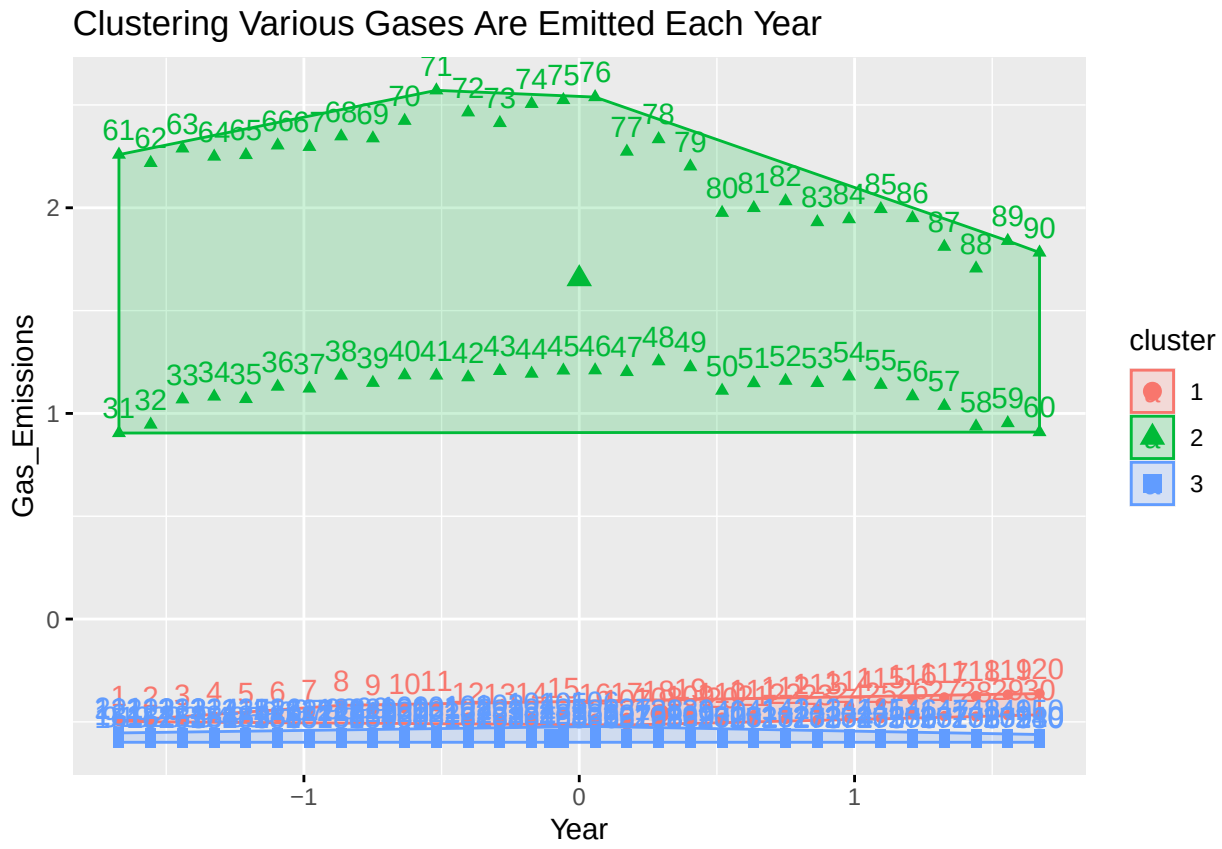
Plotting Various Gases Are Emitted Each Year:

```
ggplot(core_dft8,aes(x=Year, y=Gas_Emissions,fill=Gas)) +
  ggtitle("Various Gases Are Emitted Each Year") +
  xlab("Year") +
  ylab("Total Metric Tons Released") +
  theme_minimal(base_size = 12) +
  geom_bar(stat="identity") +
  coord_flip() +
  scale_color_discrete(name = "Different Gases")
```



Clustering Various Gases Are Emitted Each Year:

```
coreclst3_dft = select(core_dft8,Year,Gas_Emissions)
coreclst3_dft1 = kmeans(coreclst3_dft, centers =3)
fviz_cluster(coreclst3_dft1, data = coreclst3_dft,title="Clustering Various Gases Are Emitted Each Year
```


Clustering Various Gases Are Emitted Each Year
Gas_Emissions
Year
cluster
1
2
3