Final Presentation

Prafful Patel

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MLmetrics)
```

```
##
## Attaching package: 'MLmetrics'
```

```
## The following object is masked from 'package:base':
##
##     Recall
```

The Dataset is about the Carbon Dioxide emissions from various types of fuels and other sources per year, per nation which amounts to the increase in CO2. This dataset is collected from the Carbon Dioxide Analysis Center(CDAC). These surveys were conducted from the year 1751 to 2014. The data spans over one table that contains 17232 observations and 10 variables that contain varied information. The types of data used are of integer, character and numeric types. The CO2 emission data is present in million metric ton of Carbon.

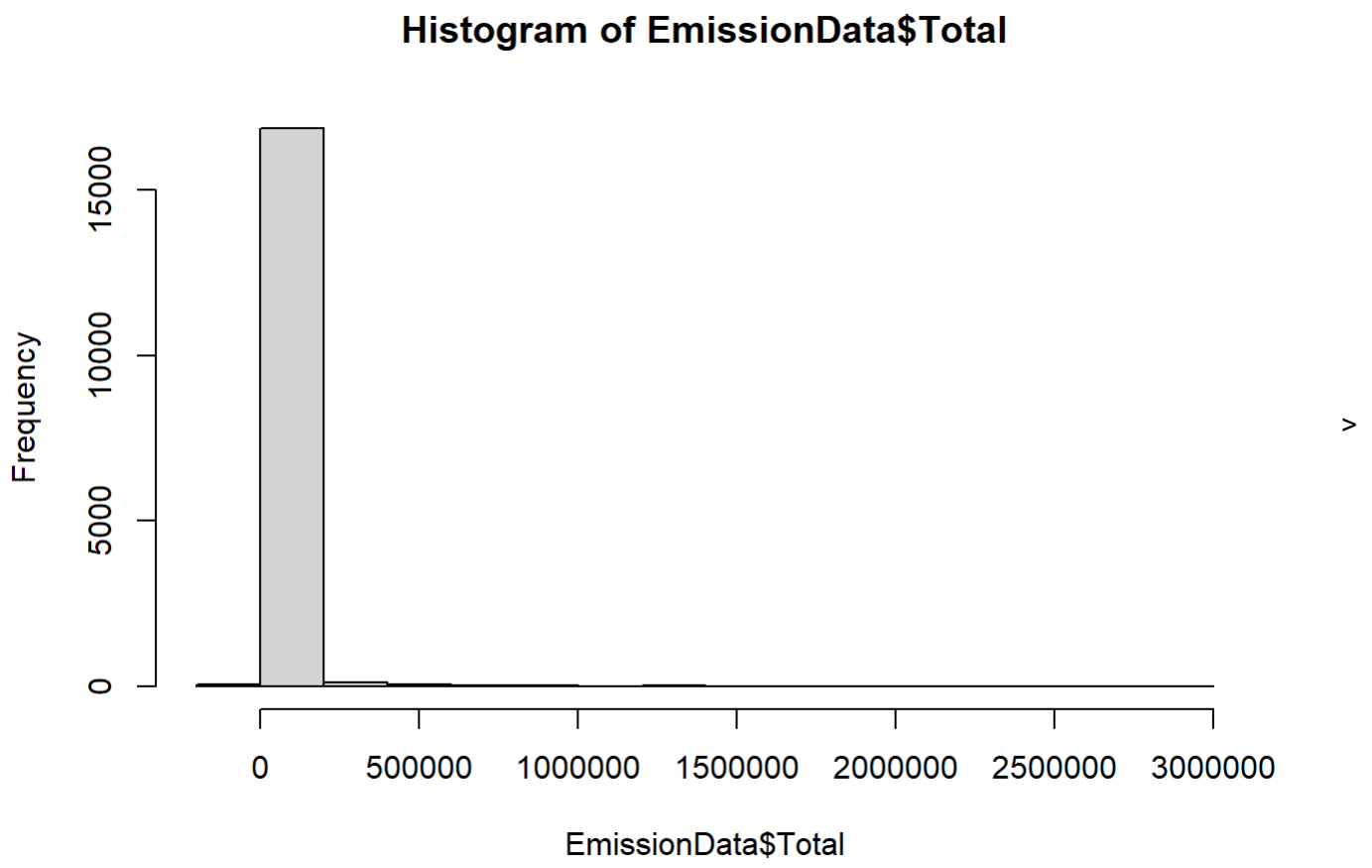
```
EmissionData<- read.csv("F:/Advance_Data_Analytics/Project/CO2/yearviseemissiondata.csv", header
=TRUE)
str(EmissionData)
```

```
## 'data.frame':    17232 obs. of  10 variables:
##  $ Year        : int  1751 1752 1753 1754 1755 1756 1757 1758 1759 1760 ...
##  $ Country     : chr  "UNITED KINGDOM" "UNITED KINGDOM" "UNITED KINGDOM" "UNITED KINGDOM" ...
##  $ Total       : int  2552 2553 2553 2554 2555 2731 2732 2733 2734 2734 ...
##  $ Solid_Fuel  : int  2552 2553 2553 2554 2555 2731 2732 2733 2734 2734 ...
##  $ Liquid_Fuel : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Gas_Fuel    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Cement      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Gas_Flaring : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Per_Capita  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Bunker_fuels: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
summary(EmissionData$Total)
```

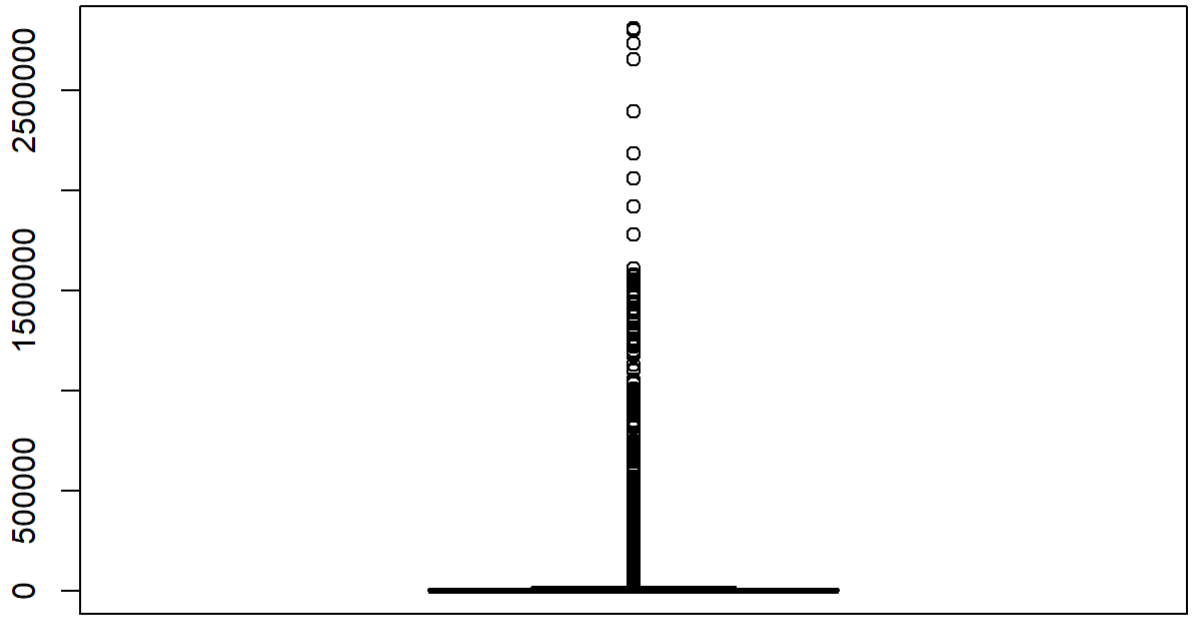
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   -1473.0     117.0     964.5   22687.1    8059.2 2806634.0
```

```
hist(EmissionData$Total)
```



The histogram is right skewed and it shows the total carbon emission between 0 to 250k million metric ton that is occured at a frequency of greater than 15000 times.

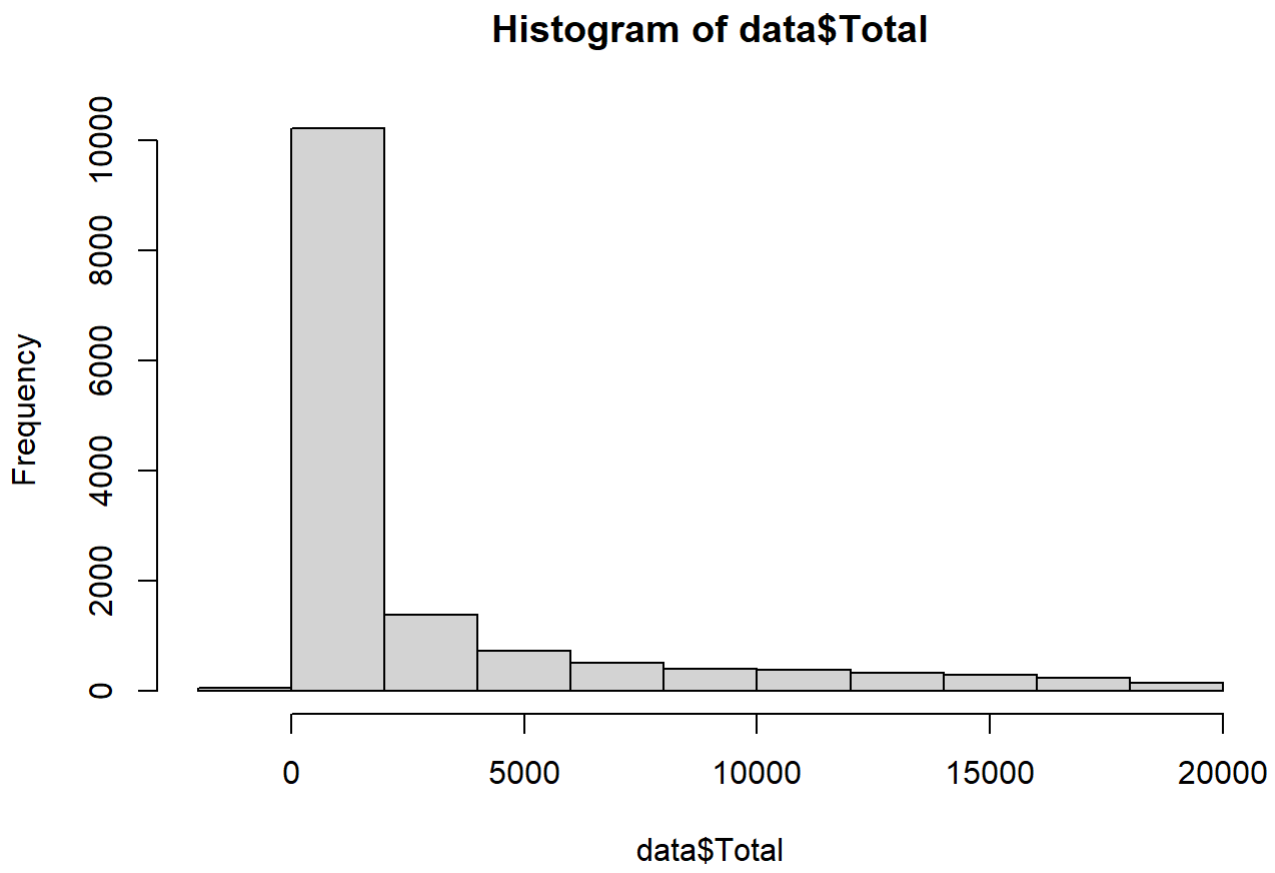
```
boxplot(EmissionData$Total)
```



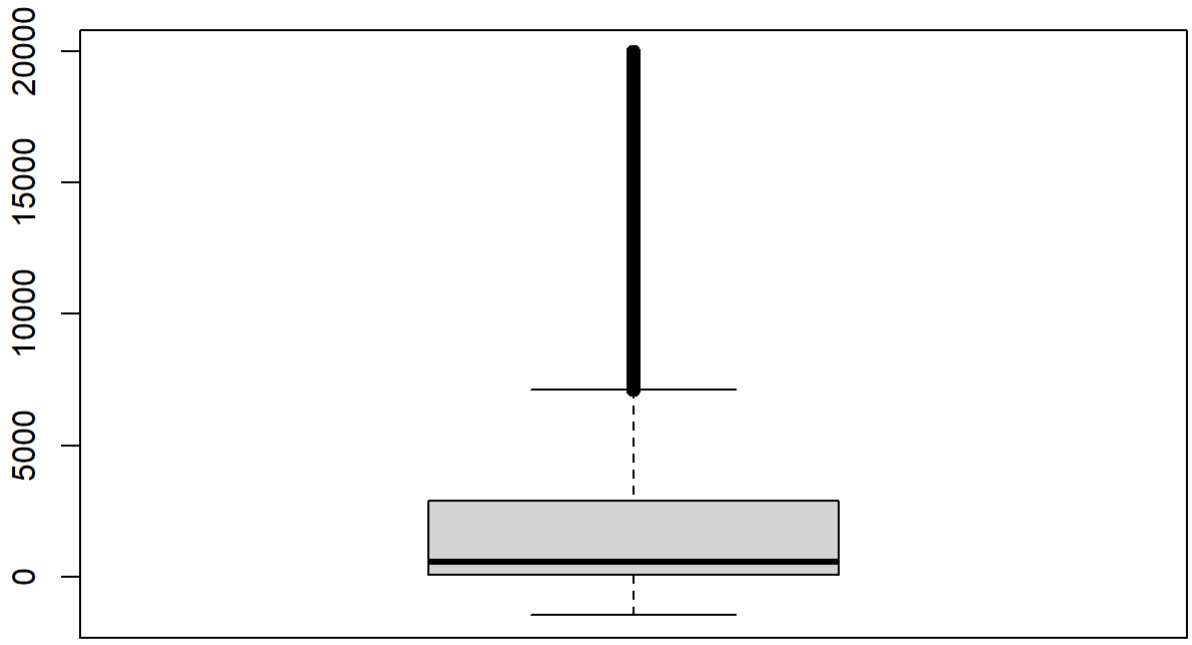
By looking at the boxplot we cannot conclude any results as there are too many outliers with data being very compact.

```
response_outliers<- boxplot.stats(EmissionData$Total)$out
data<- subset(EmissionData,!Total %in% response_outliers) #data is without outlier

hist(data$Total)
```




```
boxplot(data$Total)
```

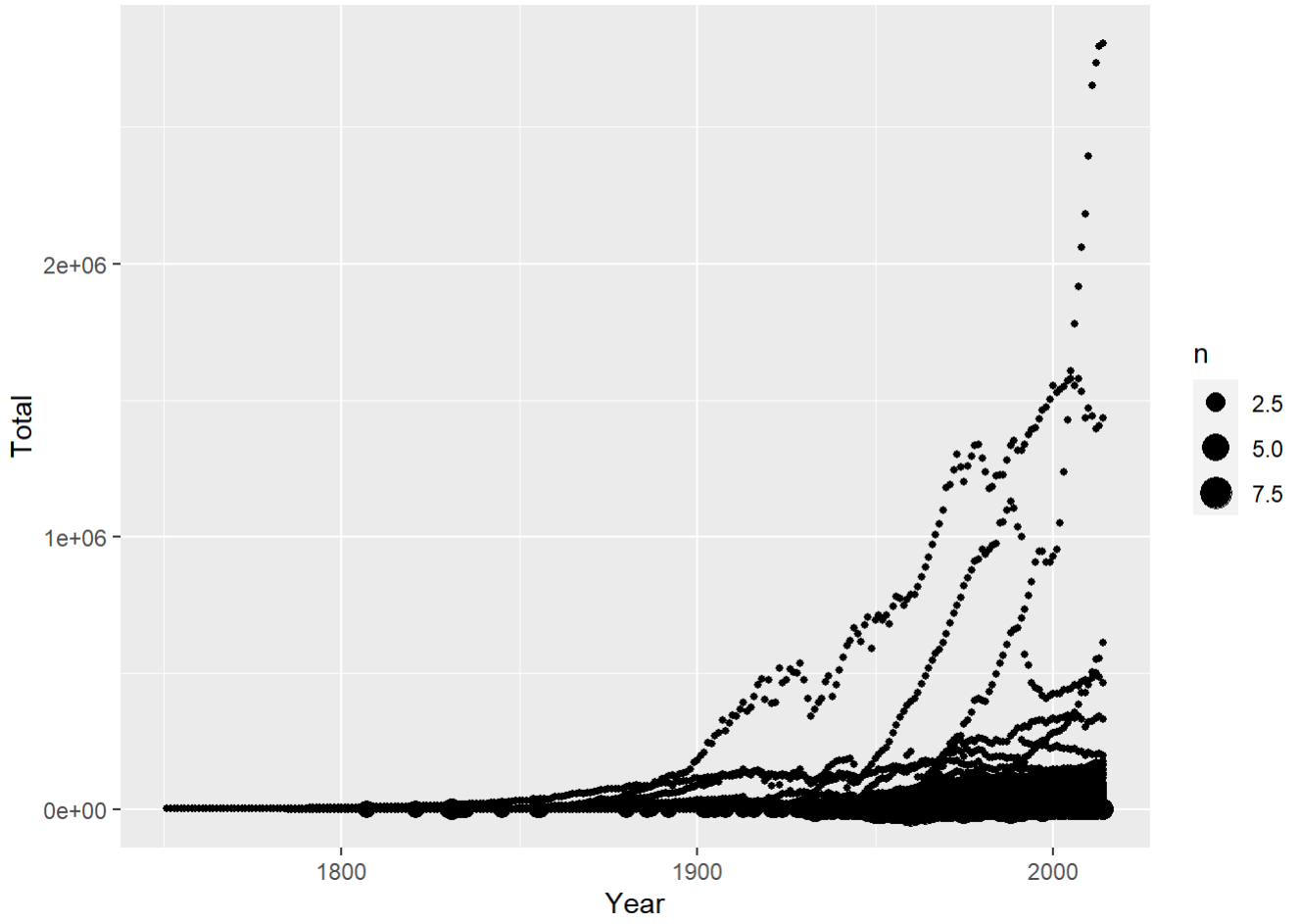



```
summary(data$Total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1473      78     551    2654    2900   19959
```

Above is the code that constructs the histogram and boxplot when the outliers are eliminated for better representation of the data due to the data being too compact.

```
ggplot(data = EmissionData) +
  geom_count(mapping = aes(x = Year, y = Total))
```



The above scatter plot indicates the carbon emission is continously increasing with respect to the year.

HYPOTHESIS

We will be doing a hypothesis on two datasets 1. Datafrom1994to2003 this data set includes data from year 1994 to 2003. 2. After2003 this data set inlcudes data from year 2004 to 2014. By the Hypothesis testing we will figure out the Total CO2 emission from years 1994 to 2003 and from years 2004 to 2014 is increasing or not.

$$H_0 : \mu 1 = \mu 2,$$
$$H_1 : \mu 1 \neq \mu 2.$$

```
datafrom1994to2003<-subset.data.frame(EmissionData,EmissionData$Year > 1993 & EmissionData$Year
 < 2004 )
summary(datafrom1994to2003)
```

```
##       Year        Country             Total           Solid_Fuel
##  Min.   :1994   Length:2145        Min.   :      1   Min.   :     0
##  1st Qu.:1996   Class :character   1st Qu.:    179   1st Qu.:     0
##  Median :1999   Mode  :character   Median :   1328   Median :     4
##  Mean   :1999                      Mean   :  29783   Mean   : 11345
##  3rd Qu.:2001                      3rd Qu.:  13408   3rd Qu.:  1437
##  Max.   :2003                      Max.   :1552682   Max.   :905917
##   Liquid_Fuel        Gas_Fuel          Cement        Gas_Flaring
##  Min.   : -4663   Min.   :     0   Min.   :     0   Min.   :    0.0
##  1st Qu.:   147   1st Qu.:     0   1st Qu.:     0   1st Qu.:    0.0
##  Median :   794   Median :     0   Median :    55   Median :    0.0
##  Mean   : 11396   Mean   :  5815   Mean   :  1032   Mean   :  194.9
##  3rd Qu.:  6242   3rd Qu.:  2449   3rd Qu.:   427   3rd Qu.:    0.0
##  Max.   :648067   Max.   :342282   Max.   :117243   Max.   :12207.0
##    Per_Capita      Bunker_fuels
##  Min.   : 0.000   Min.   :    0
##  1st Qu.: 0.170   1st Qu.:    8
##  Median : 0.750   Median :   60
##  Mean   : 1.331   Mean   :  966
##  3rd Qu.: 1.980   3rd Qu.:  394
##  Max.   :19.340   Max.   :40072
```

```
after2003<-subset.data.frame(EmissionData,EmissionData$Year > 2003)
summary(after2003)
```

```
##       Year        Country             Total              Solid_Fuel
##  Min.   :2004   Length:2395        Min.   :      1.0   Min.   :      0
##  1st Qu.:2006   Class :character   1st Qu.:    243.5   1st Qu.:      0
##  Median :2009   Mode  :character   Median :   1848.0   Median :     13
##  Mean   :2009                      Mean   :  38905.8   Mean   :  16626
##  3rd Qu.:2012                      3rd Qu.:  15080.5   3rd Qu.:   1542
##  Max.   :2014                      Max.   :2806634.0   Max.   :2045156
##   Liquid_Fuel        Gas_Fuel          Cement          Gas_Flaring
##  Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :    0.0
##  1st Qu.:   183   1st Qu.:     0   1st Qu.:     0.0   1st Qu.:    0.0
##  Median :  1077   Median :     9   Median :   112.0   Median :    0.0
##  Mean   : 12473   Mean   :  7532   Mean   :  1979.0   Mean   :  295.6
##  3rd Qu.:  6474   3rd Qu.:  3182   3rd Qu.:   599.5   3rd Qu.:    0.0
##  Max.   :667143   Max.   :390719   Max.   :338912.0   Max.   :12662.0
##    Per_Capita      Bunker_fuels
##  Min.   : 0.000   Min.   :    0.0
##  1st Qu.: 0.215   1st Qu.:   10.0
##  Median : 0.830   Median :   92.0
##  Mean   : 1.417   Mean   : 1340.0
##  3rd Qu.: 1.940   3rd Qu.:  610.5
##  Max.   :17.690   Max.   :45630.0
```

```
t.test(datafrom1994to2003$Total,after2003$Total, var.equal = FALSE, conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  datafrom1994to2003$Total and after2003$Total
## t = -1.9096, df = 4230.8, p-value = 0.05625
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18489.6660    243.1931
## sample estimates:
## mean of x mean of y
##  29782.61  38905.84
```

We are using t.test because we have two means and the variance are unknown and are not equal. The p-value is greater than alpha i.e., 0.05. So we can accept the null hypothesis H_0 and agree that the increase of Total CO2 emission from year 1994 to 2003 is equal to the increase of Total CO2 emission from year 2004 to 2014.

REGRESSION

Total is Response and Solid_Fuel,Liquid_Fuel, Gas_Fuel, Cement, Gas_Flaring, Bunker_fuels is Predictor.

```
fitlm <- lm(Total ~.-Country-Year, data=EmissionData)
summary(fitlm)
```

```
##
## Call:
## lm(formula = Total ~ . - Country - Year, data = EmissionData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01001  0.00672  0.00853  0.00877  2.01331
##
## Coefficients:
##                Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  -8.791e-03  3.675e-03 -2.392e+00   0.0168 *
## Solid_Fuel    1.000e+00  1.332e-07  7.506e+06   <2e-16 ***
## Liquid_Fuel   1.000e+00  2.458e-07  4.068e+06   <2e-16 ***
## Gas_Fuel      1.000e+00  3.839e-07  2.605e+06   <2e-16 ***
## Cement        1.000e+00  1.001e-06  9.986e+05   <2e-16 ***
## Gas_Flaring   1.000e+00  3.289e-06  3.041e+05   <2e-16 ***
## Per_Capita    1.092e-03  1.566e-03  6.980e-01   0.4853
## Bunker_fuels  8.400e-07  1.951e-06  4.310e-01   0.6668
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4362 on 17224 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.659e+14 on 7 and 17224 DF,  p-value: < 2.2e-16
```

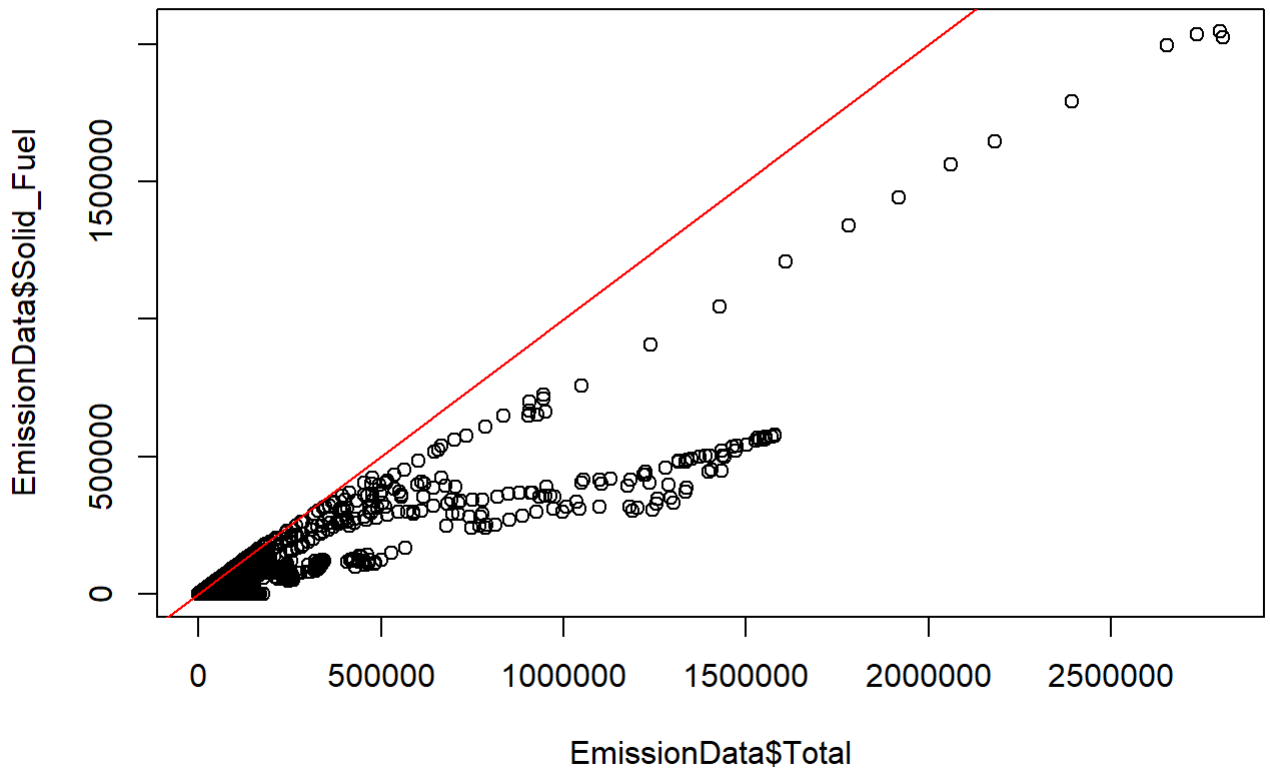
By performing the Multiple Linear Regression we found that there is strong relationship between Response and all predictors because the p-value of each model is close to 0 except Per_Capita and Bunker_fuels because they don't have direct relationship with Total CO2 emission.

The relationship between Response and all Predictors is Positive because the coefficient value is positive which means Response is directly propotional to the Predictor.

The model is a very fit model because the R-squared value is 1 and the RSE value is very close to 0.

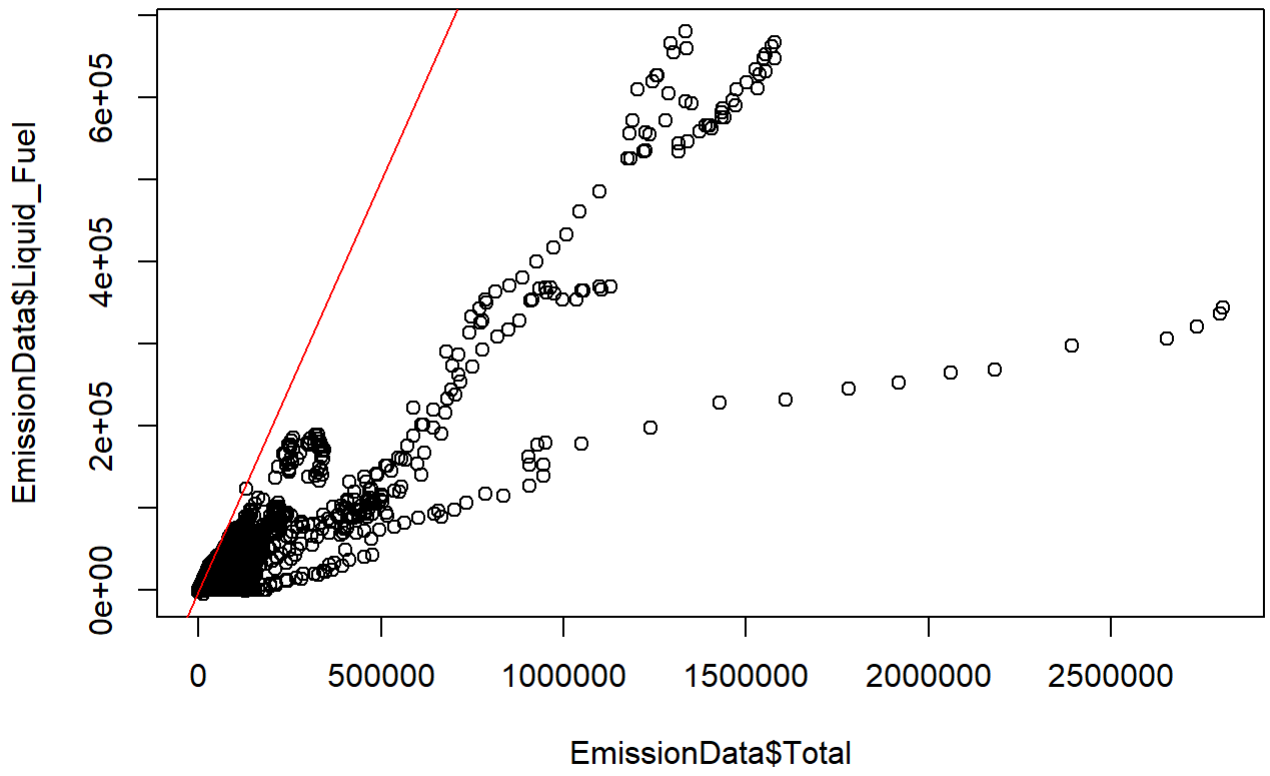
```
plot(EmissionData$Total, EmissionData$Solid_Fuel)
abline(fitlm, col="red")
```

```
## Warning in abline(fitlm, col = "red"): only using the first two of 8 regression
## coefficients
```



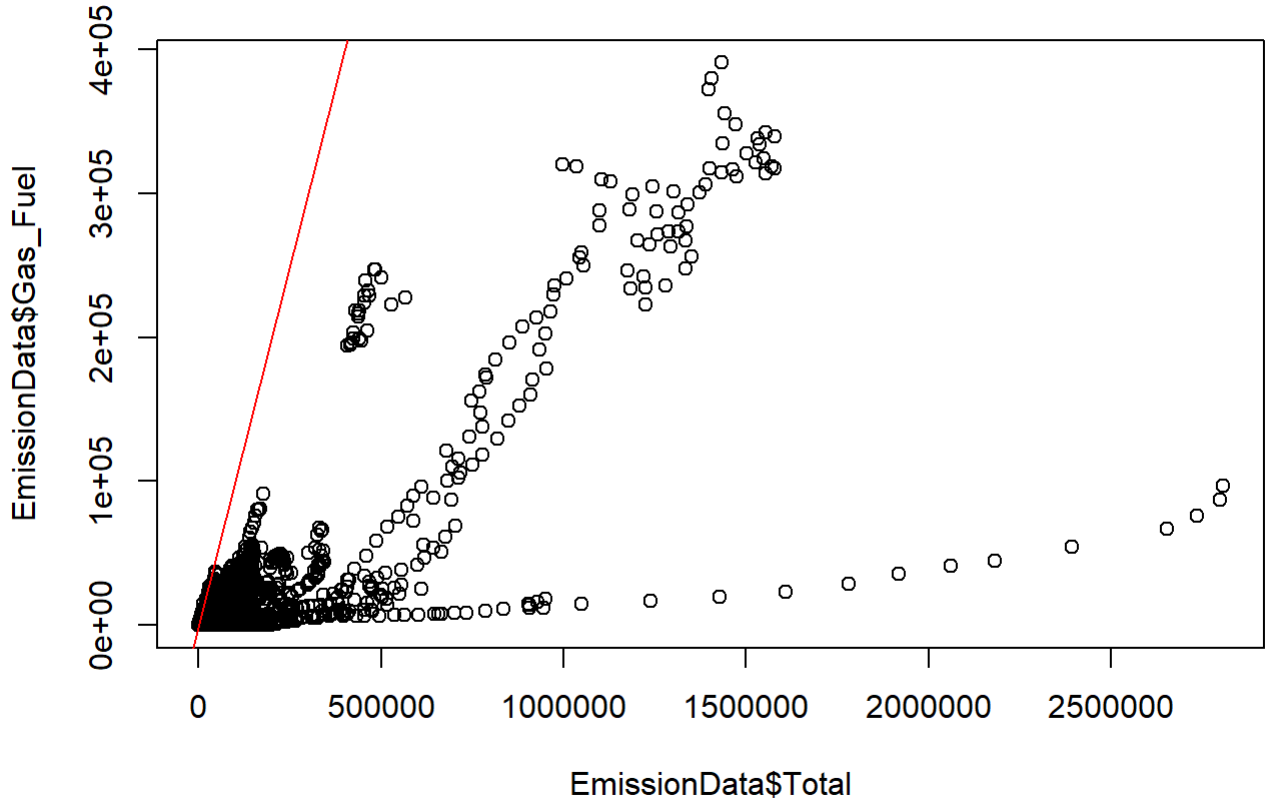

```
plot(EmissionData$Total, EmissionData$Liquid_Fuel)
abline(fitlm, col="red")
```

```
## Warning in abline(fitlm, col = "red"): only using the first two of 8 regression
## coefficients
```



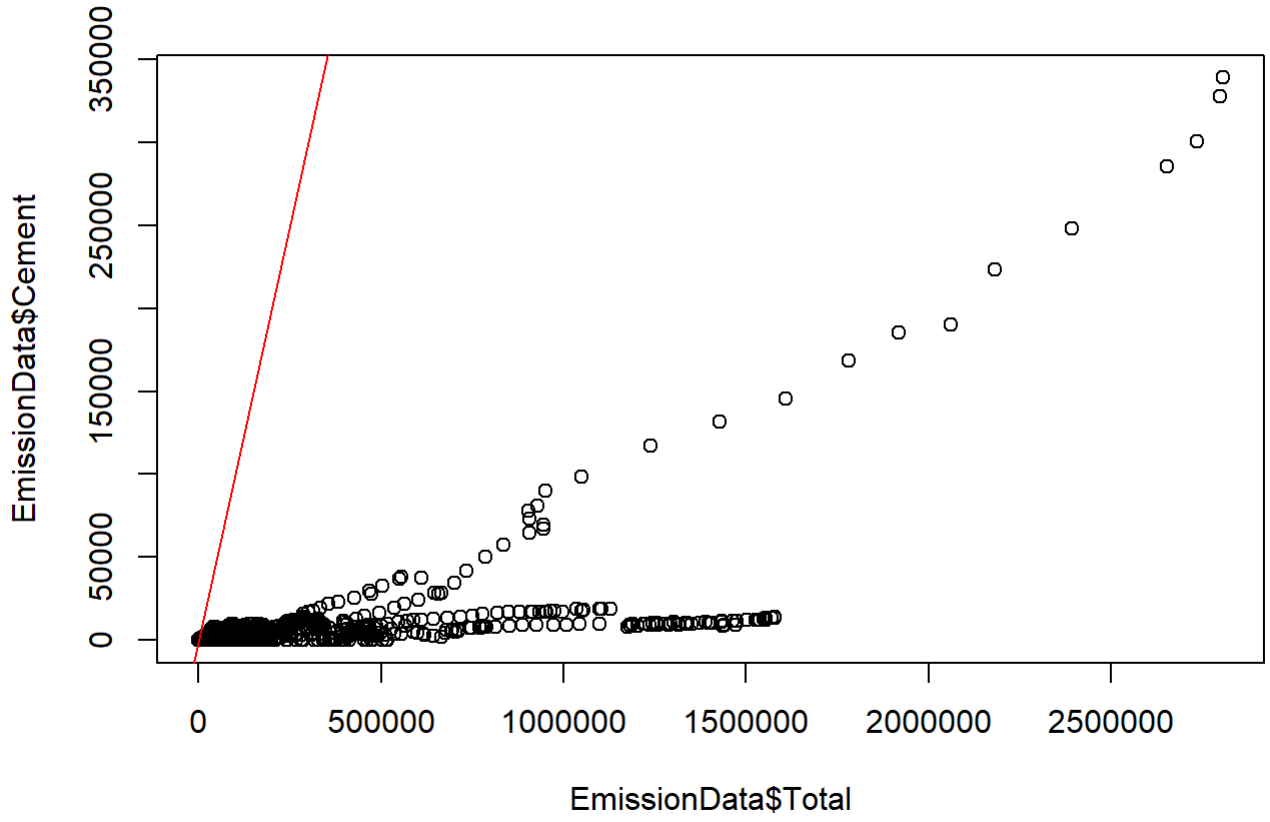

```
plot(EmissionData$Total, EmissionData$Gas_Fuel)
abline(fitlm, col="red")
```

```
## Warning in abline(fitlm, col = "red"): only using the first two of 8 regression
## coefficients
```



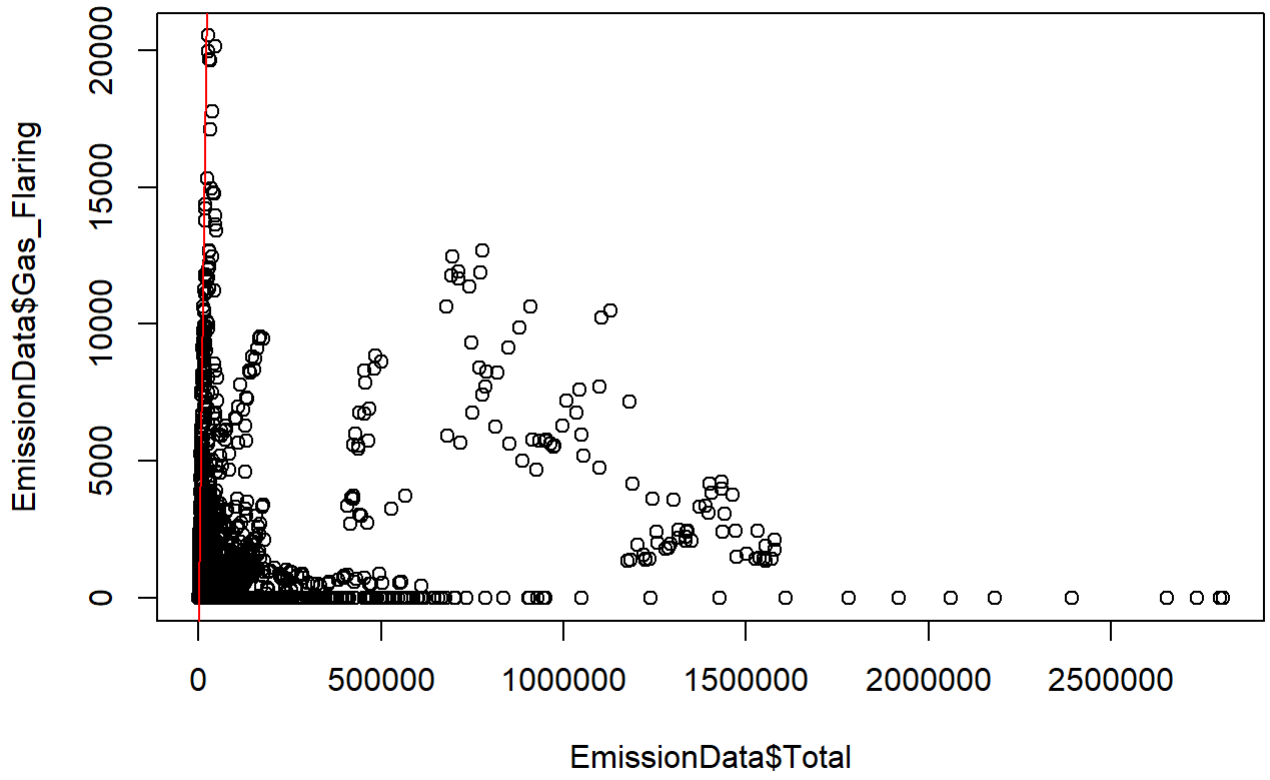

```
plot(EmissionData$Total, EmissionData$Cement)
abline(fitlm, col="red")
```

```
## Warning in abline(fitlm, col = "red"): only using the first two of 8 regression
## coefficients
```



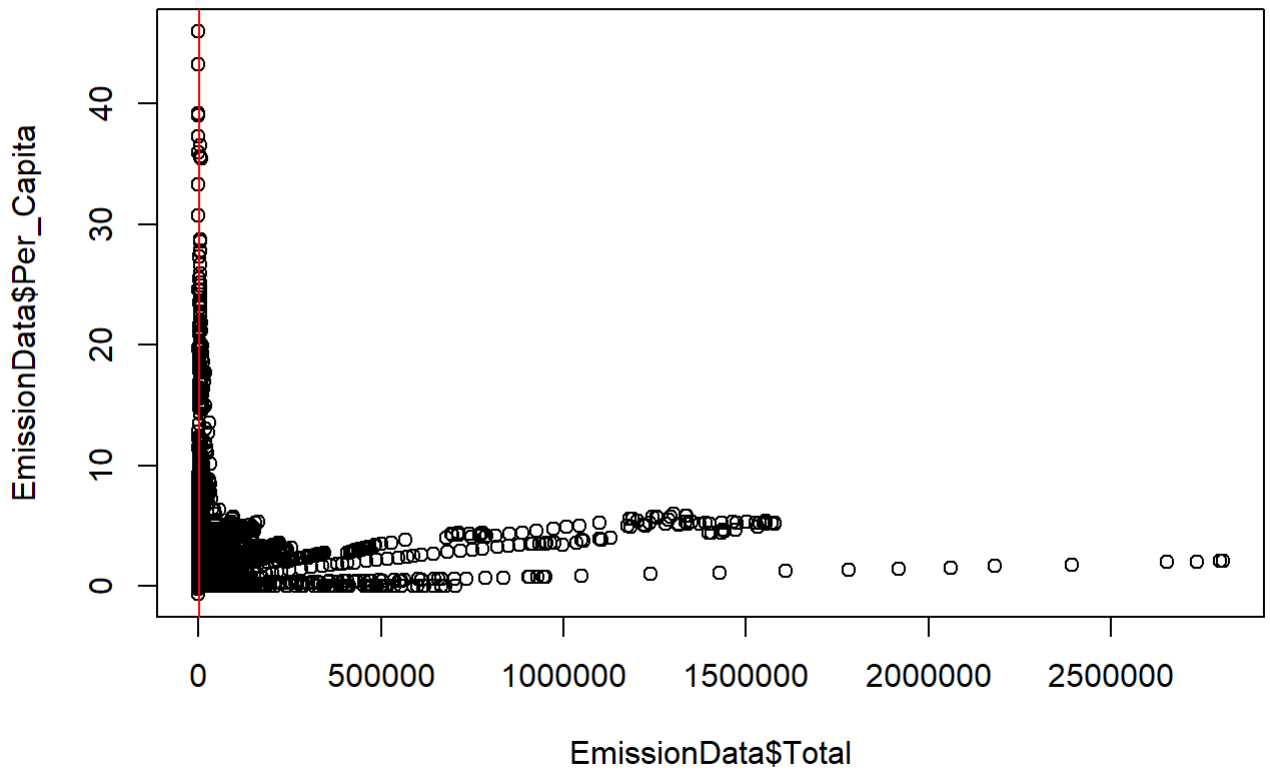

```
plot(EmissionData$Total, EmissionData$Gas_Flaring)
abline(fitlm, col="red")
```

```
## Warning in abline(fitlm, col = "red"): only using the first two of 8 regression
## coefficients
```



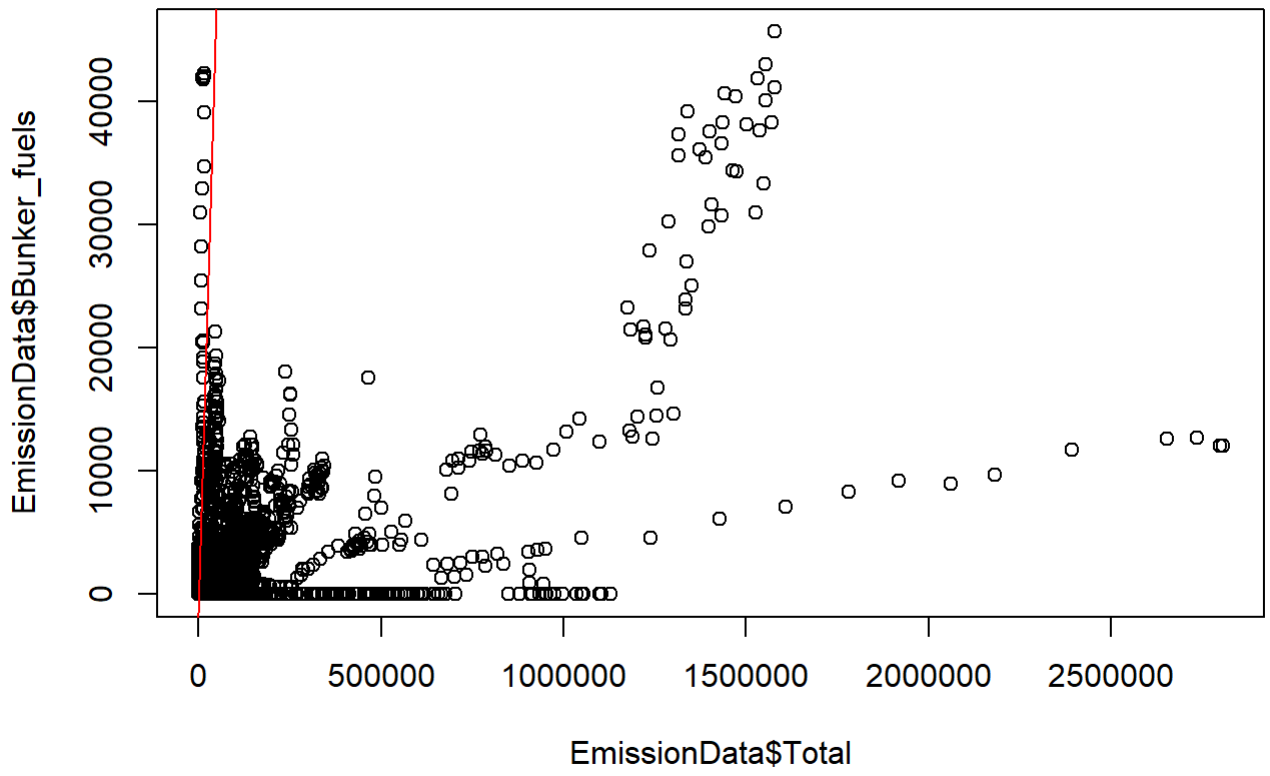

```
plot(EmissionData$Total, EmissionData$Per_Capita)
abline(fitlm, col="red")
```

```
## Warning in abline(fitlm, col = "red"): only using the first two of 8 regression
## coefficients
```




```
plot(EmissionData$Total, EmissionData$Bunker_fuels)
abline(fitlm, col="red")
```

```
## Warning in abline(fitlm, col = "red"): only using the first two of 8 regression
## coefficients
```




```
library(MLmetrics)
ypred <-predict(object = fitlm, newdata = EmissionData)
summary(ypred)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   -1473.0     117.0     964.5   22687.1    8059.2 2806633.3
```

```
MAE(y_pred = ypred, y_true = EmissionData$Total)
```

```
## [1] 0.1955349
```

```
MSE(y_pred = ypred, y_true = EmissionData$Total)
```

```
## [1] 0.1902228
```

And the Mean Absolute Error and Mean Squared Error are very low.