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 1950 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.

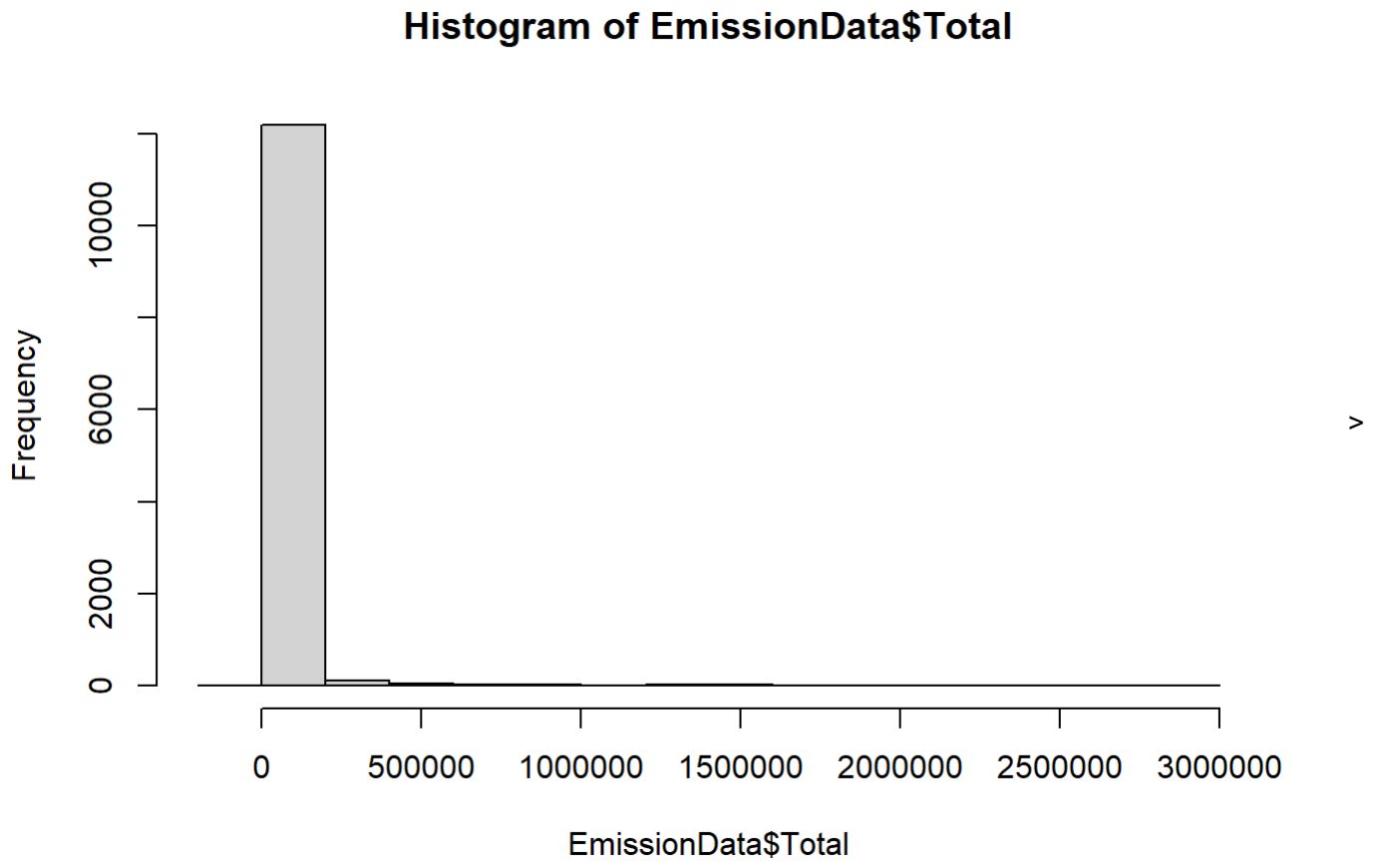
```
EmissionData<- read.csv("F:/Advance_Data_Analytics/Project/CO2/yearviseemissiondatafrom1950.csv"
, header=TRUE)
str(EmissionData)
```

```
## 'data.frame':    12462 obs. of  10 variables:
##  $ Year        : int  1950 1950 1950 1950 1950 1950 1950 1950 1950 1950 ...
##  $ Country     : chr  "AFGHANISTAN" "ALBANIA" "ALGERIA" "ANGOLA" ...
##  $ Total       : int  23 81 1033 51 8168 14941 5704 15 377 20 ...
##  $ Solid_Fuel  : int  6 12 514 16 972 12028 4744 0 0 1 ...
##  $ Liquid_Fuel : int  18 68 475 34 6982 2739 532 15 377 18 ...
##  $ Gas_Fuel    : int  0 0 0 0 0 0 253 0 0 1 ...
##  $ Cement      : int  0 2 44 0 214 174 175 0 0 0 ...
##  $ Gas_Flaring : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Per_Capita  : num  0 0.07 0.12 0.01 0.48 1.83 0.82 0.19 3.26 0.1 ...
##  $ Bunker_fuels: int  0 0 612 0 124 758 0 3 554 0 ...
```

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

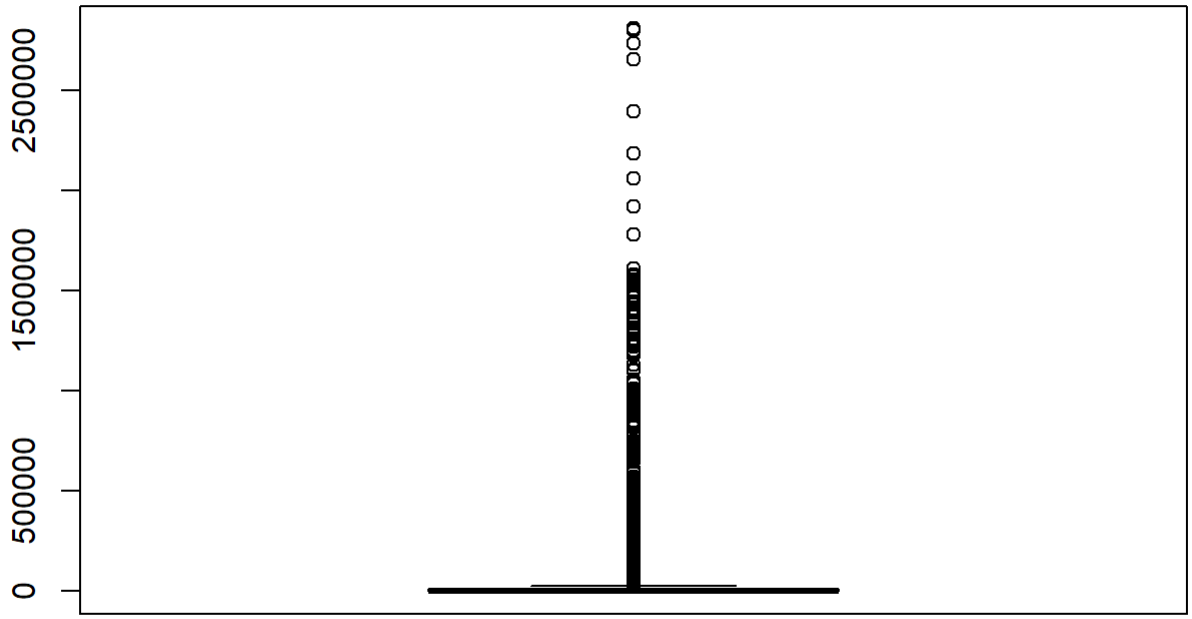
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   -1473.0     106.2     844.0   26426.9    9425.5 2806634.0
```

```
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 10000 times.

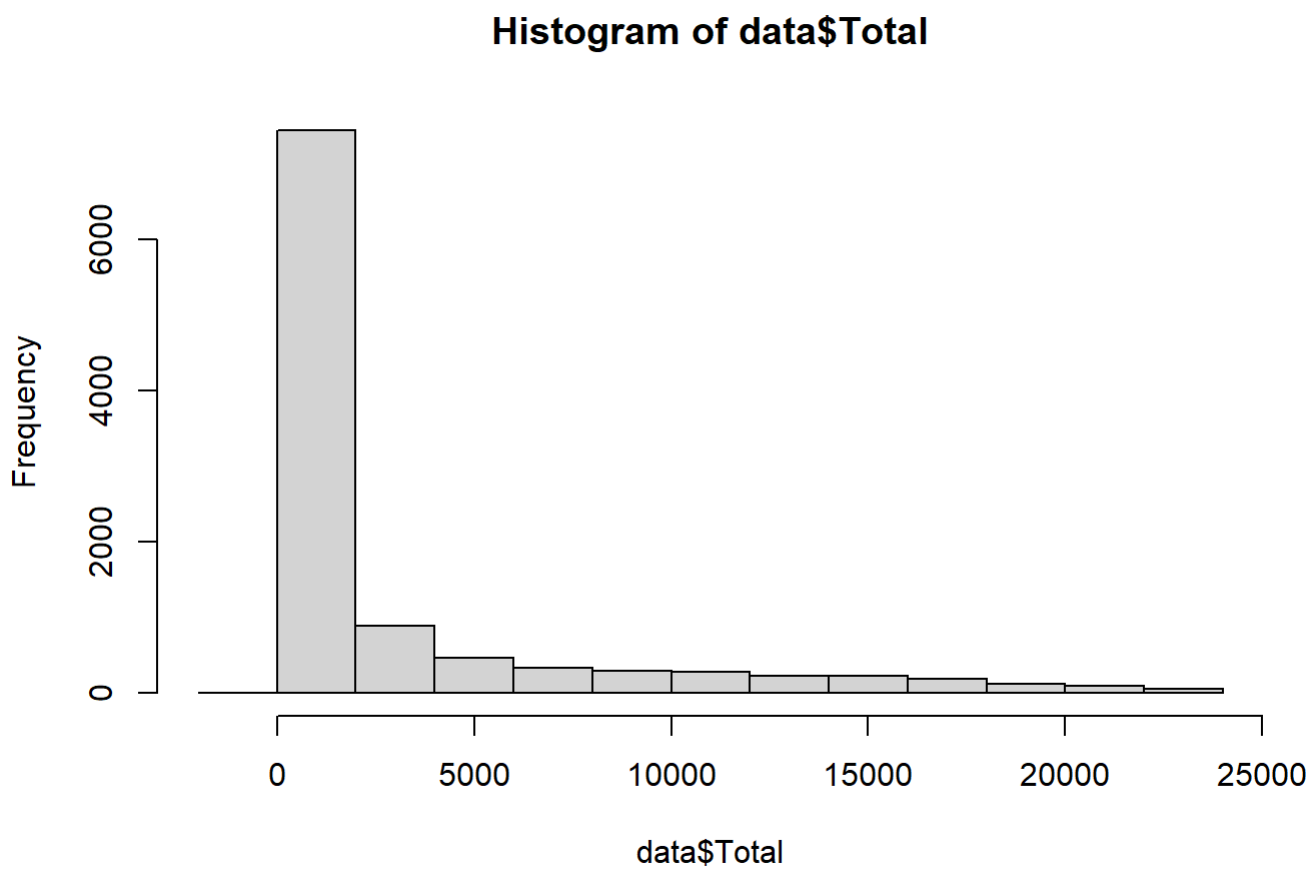
```
boxplot(EmissionData$Total)
```



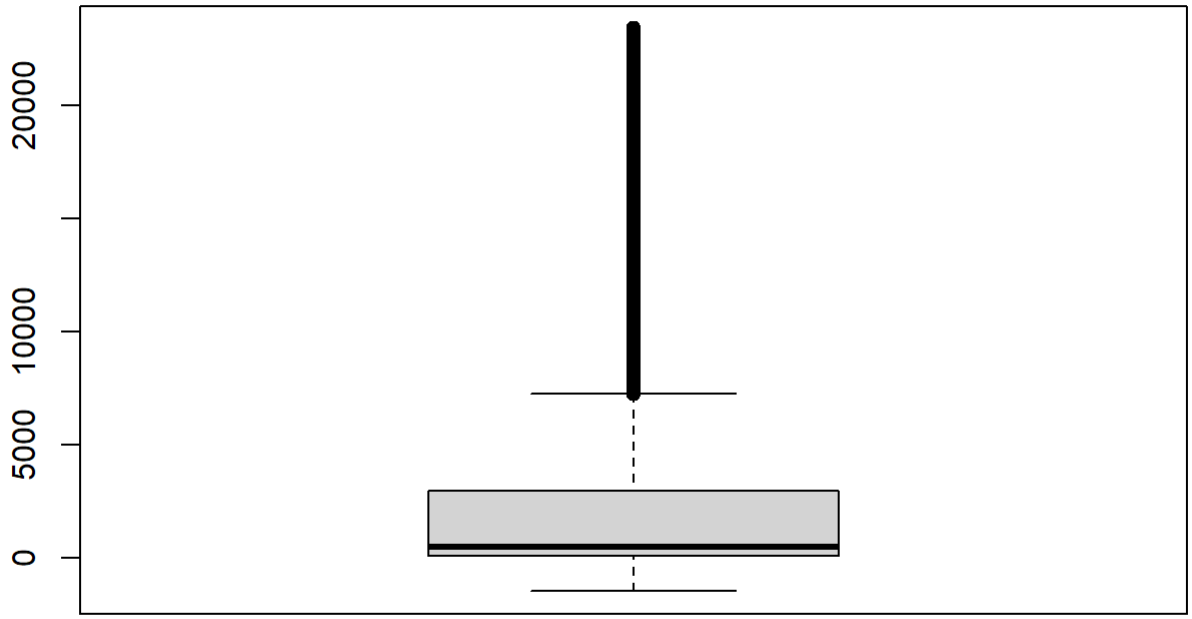
By looking at the boxplot we cannot conclude any results as there are too many outliers with data being very compact.

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

hist(data$Total)
```




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

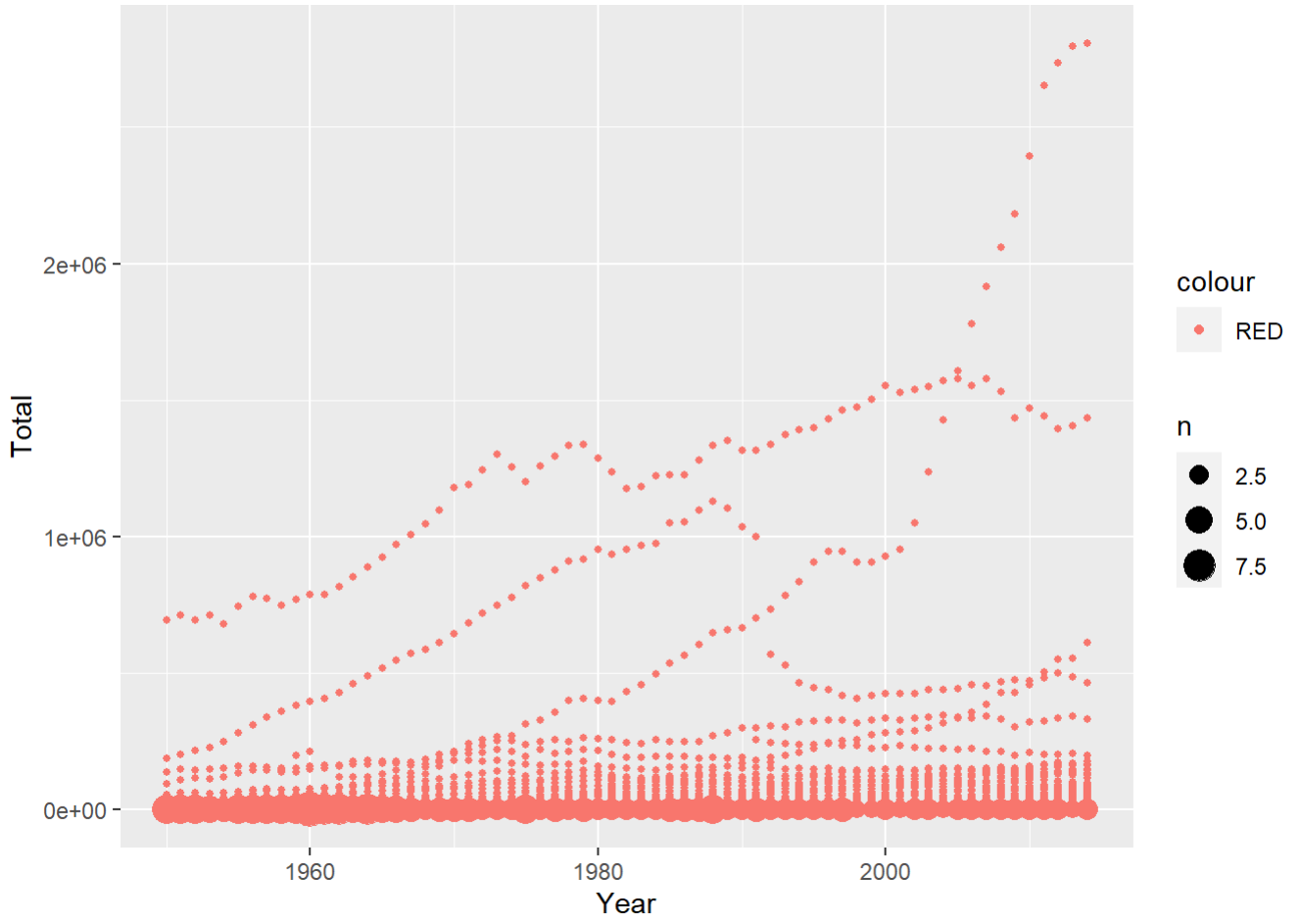



```
summary(data$Total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1473      71     484    2911    2950   23404
```

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, color = 'RED'))
```



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

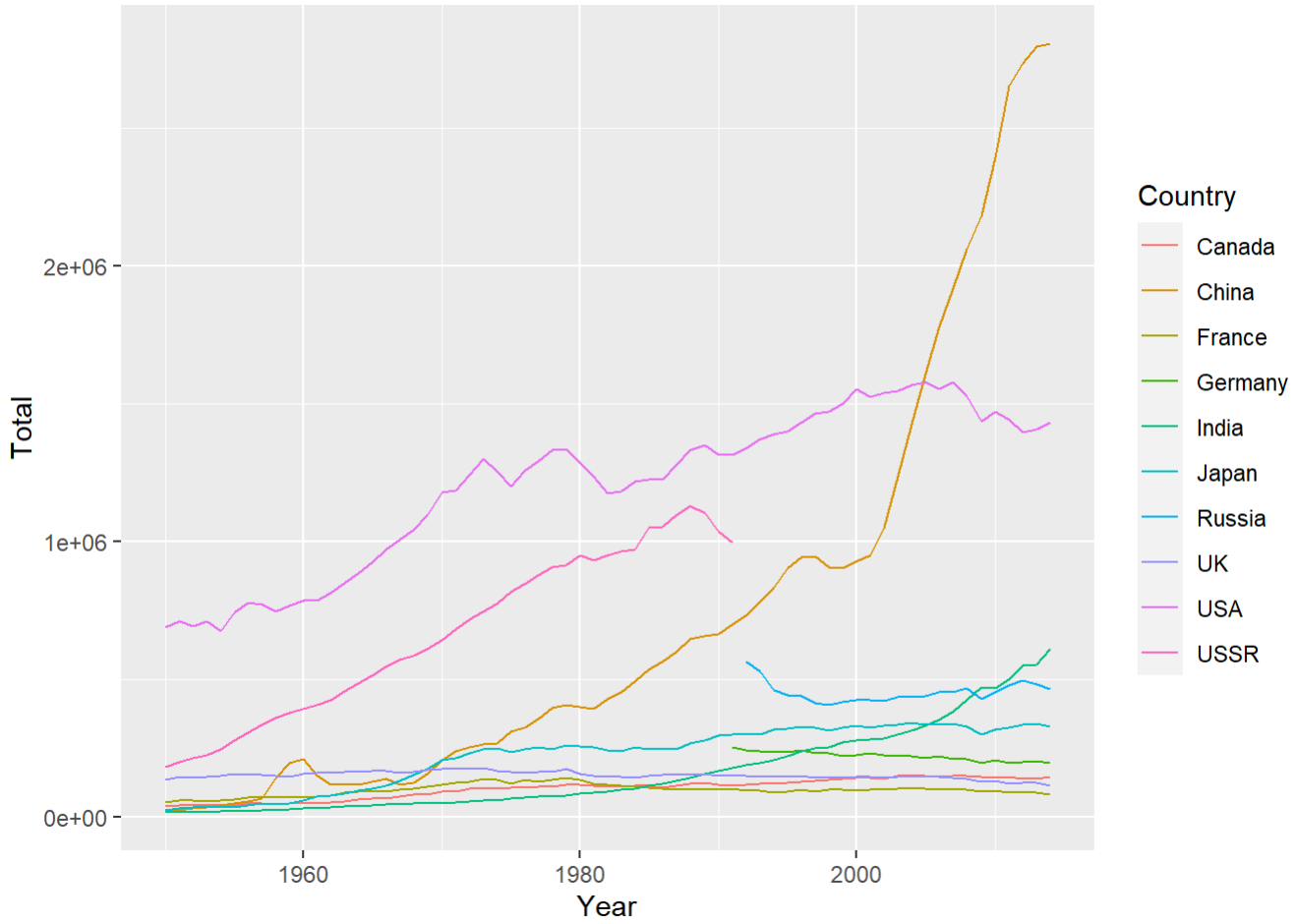
```
Top<- read.csv("F:/Advance_Data_Analytics/Project/CO2/Top10yearvise.csv", header=TRUE)
str(Top)
```

```
## 'data.frame':    544 obs. of  10 variables:
##  $ Year        : int  1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 ...
##  $ Country     : chr  "Canada" "Canada" "Canada" "Canada" ...
##  $ Total       : int  42070 44402 43510 43838 44482 46258 51807 49900 49721 50356 ...
##  $ Solid_Fuel  : int  26424 26442 24522 22636 21857 19709 21420 17719 15629 14614 ...
##  $ Liquid_Fuel : int  14314 16480 17421 19063 20518 23405 26530 27326 28539 29383 ...
##  $ Gas_Fuel    : int  970 1114 1167 1313 1623 2030 2331 2965 3665 4723 ...
##  $ Cement      : int  361 367 400 480 484 543 620 746 759 775 ...
##  $ Gas_Flaring : int  0 0 0 347 0 571 906 1144 1129 861 ...
##  $ Per_Capita  : num  3.06 3.14 3 2.94 2.9 2.94 3.2 3 2.91 2.88 ...
##  $ Bunker_fuels: int  1230 1335 1369 1489 1335 1480 1535 1802 1541 1828 ...
```

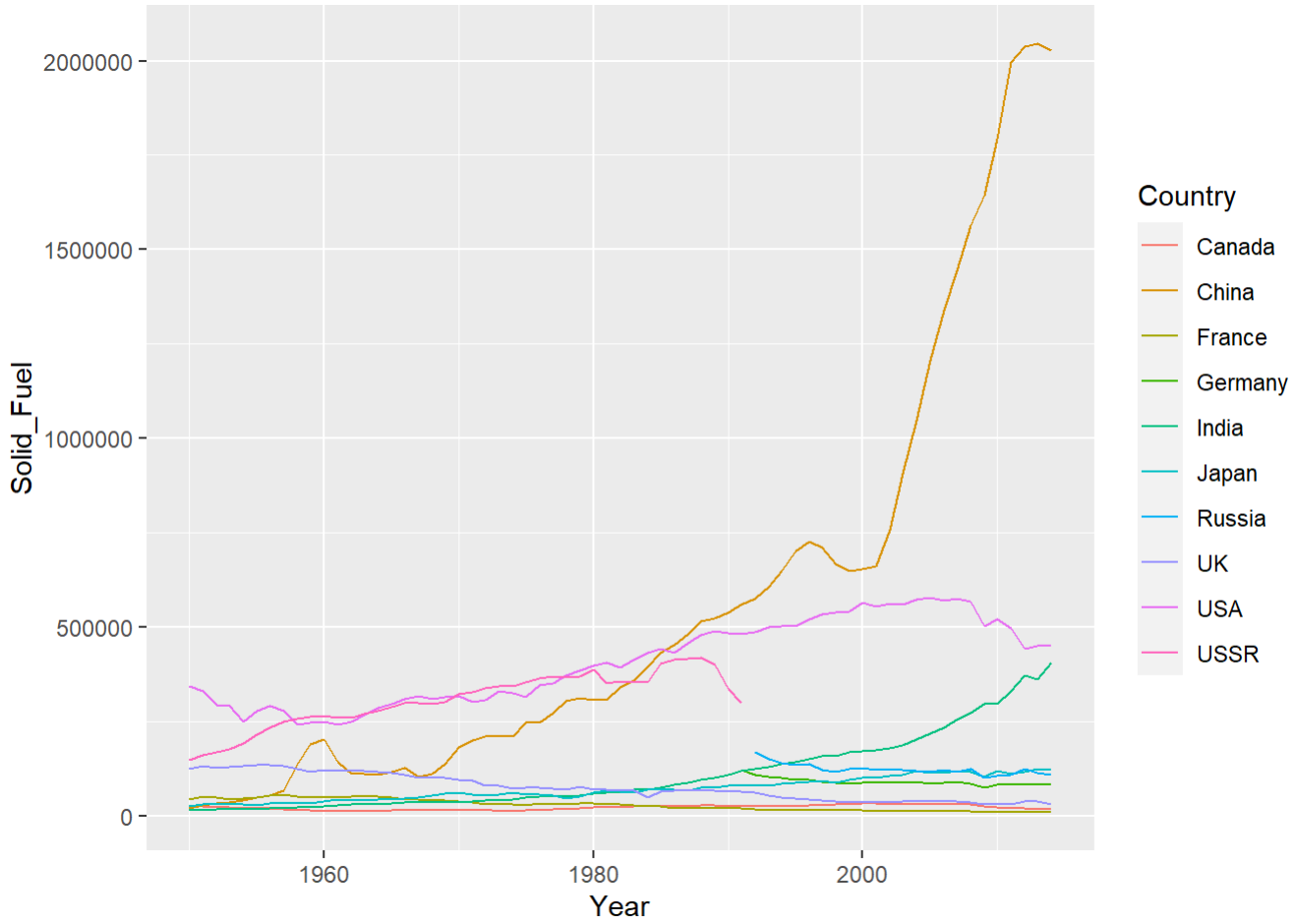
The above dataset consist data of top 10 countries to closely examine the data. By this we can figure which country is emitting more CO2 in the air. And to know which Country is responsible for Most and Least CO2 emission we can find out this using two methods. First is by using Data Visualization technique and another one is by using data frame operation. You can see both methods below.

By using Data Visualization

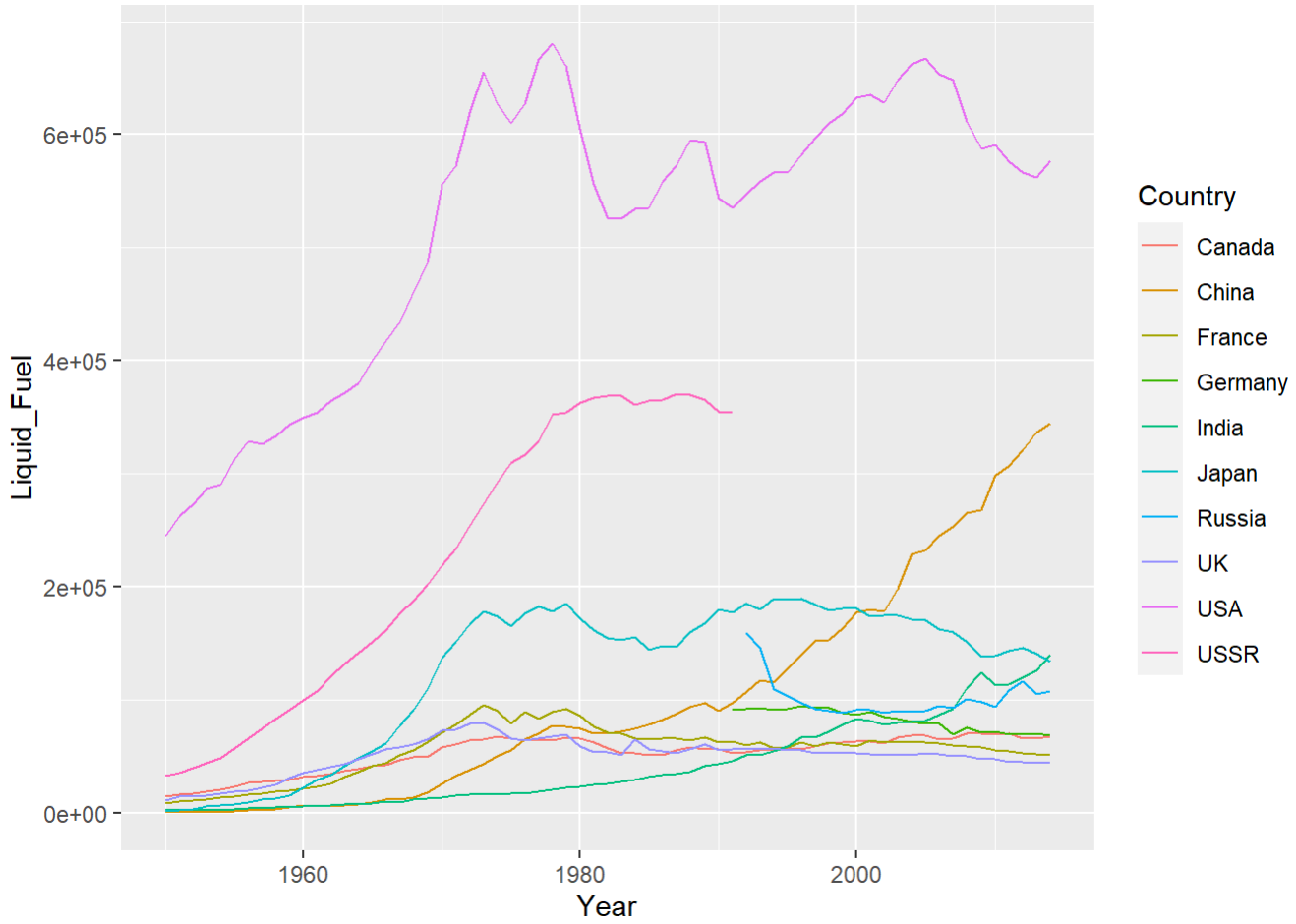
```
ggplot(data = Top, mapping = aes(x = Year)) +
  geom_line(aes(y = Total,colour = Country))
```



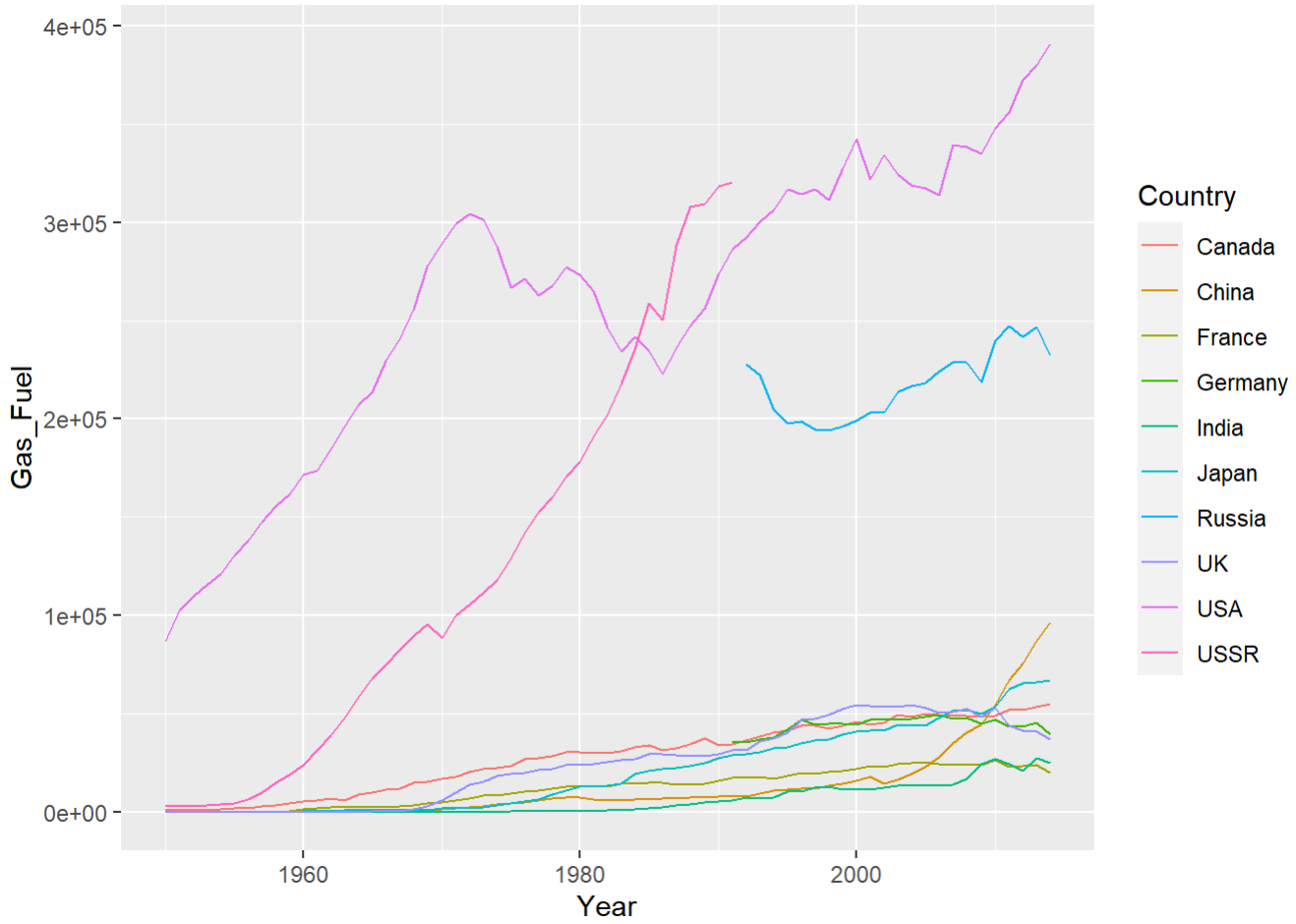

```
ggplot(data = Top, mapping = aes(x = Year)) +
  geom_line(aes(y = Solid_Fuel,colour = Country))
```



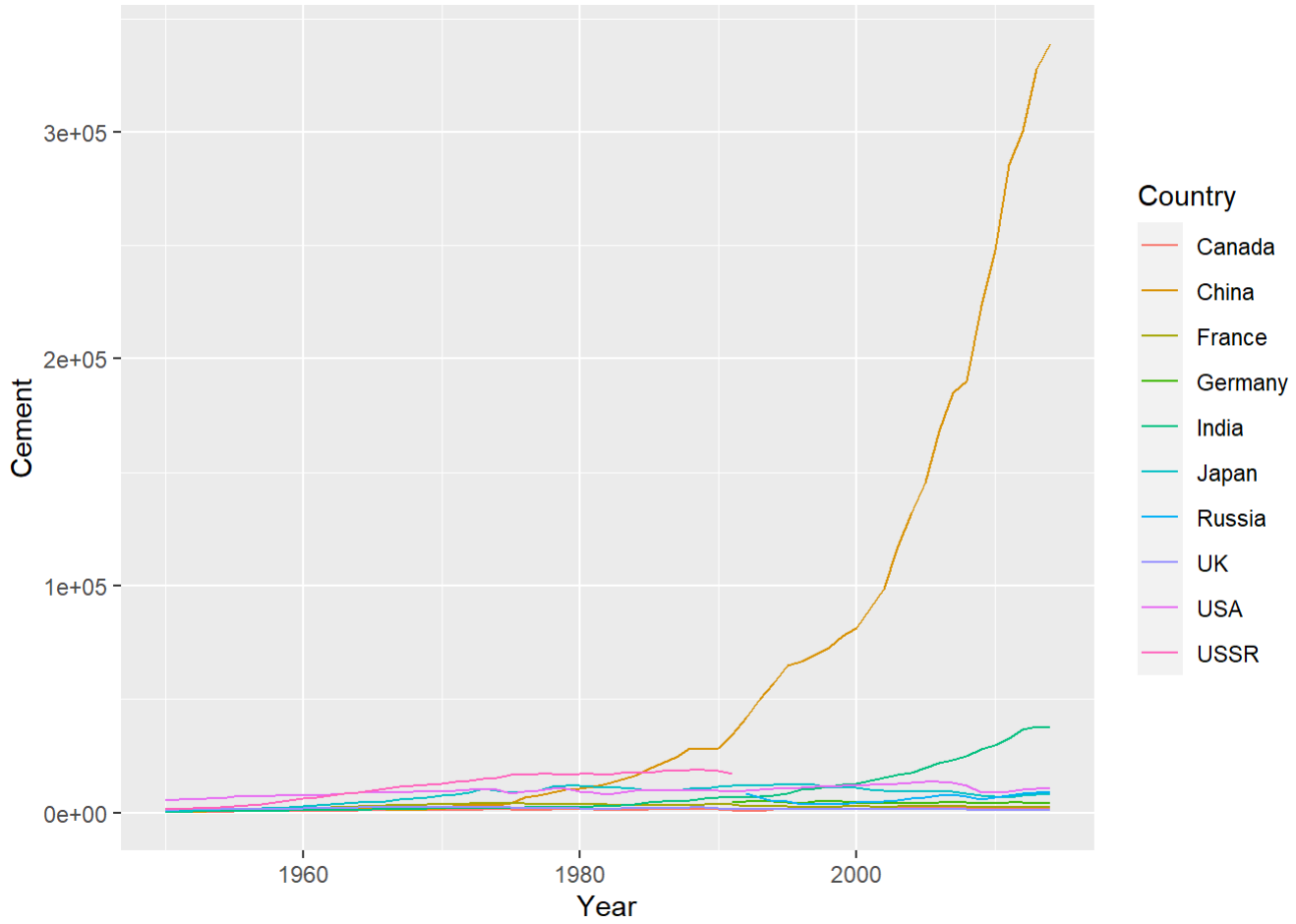

```
ggplot(data = Top, mapping = aes(x = Year)) +
  geom_line(aes(y = Liquid_Fuel,colour = Country))
```



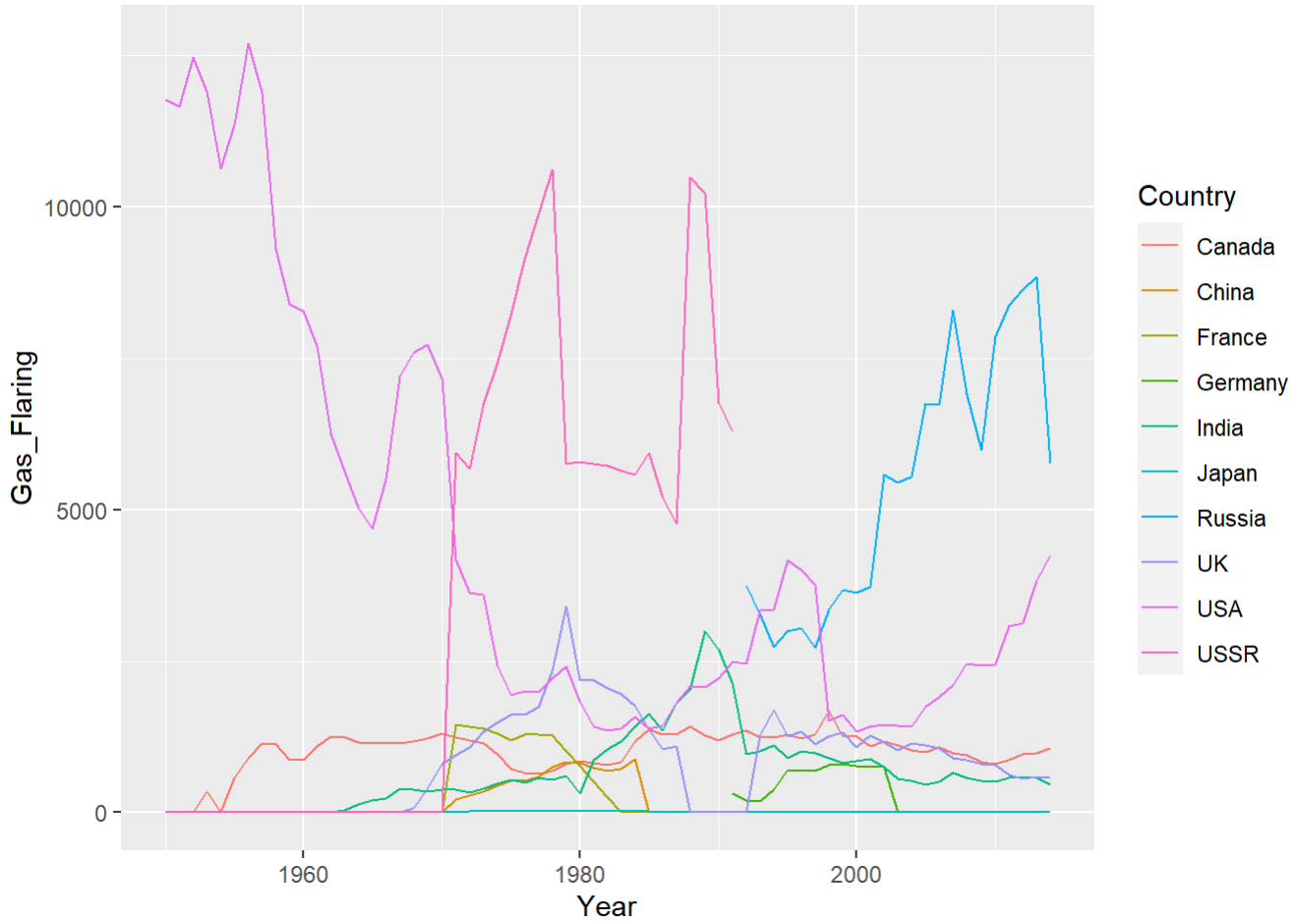

```
ggplot(data = Top, mapping = aes(x = Year)) +
  geom_line(aes(y = Gas_Fuel,colour = Country))
```



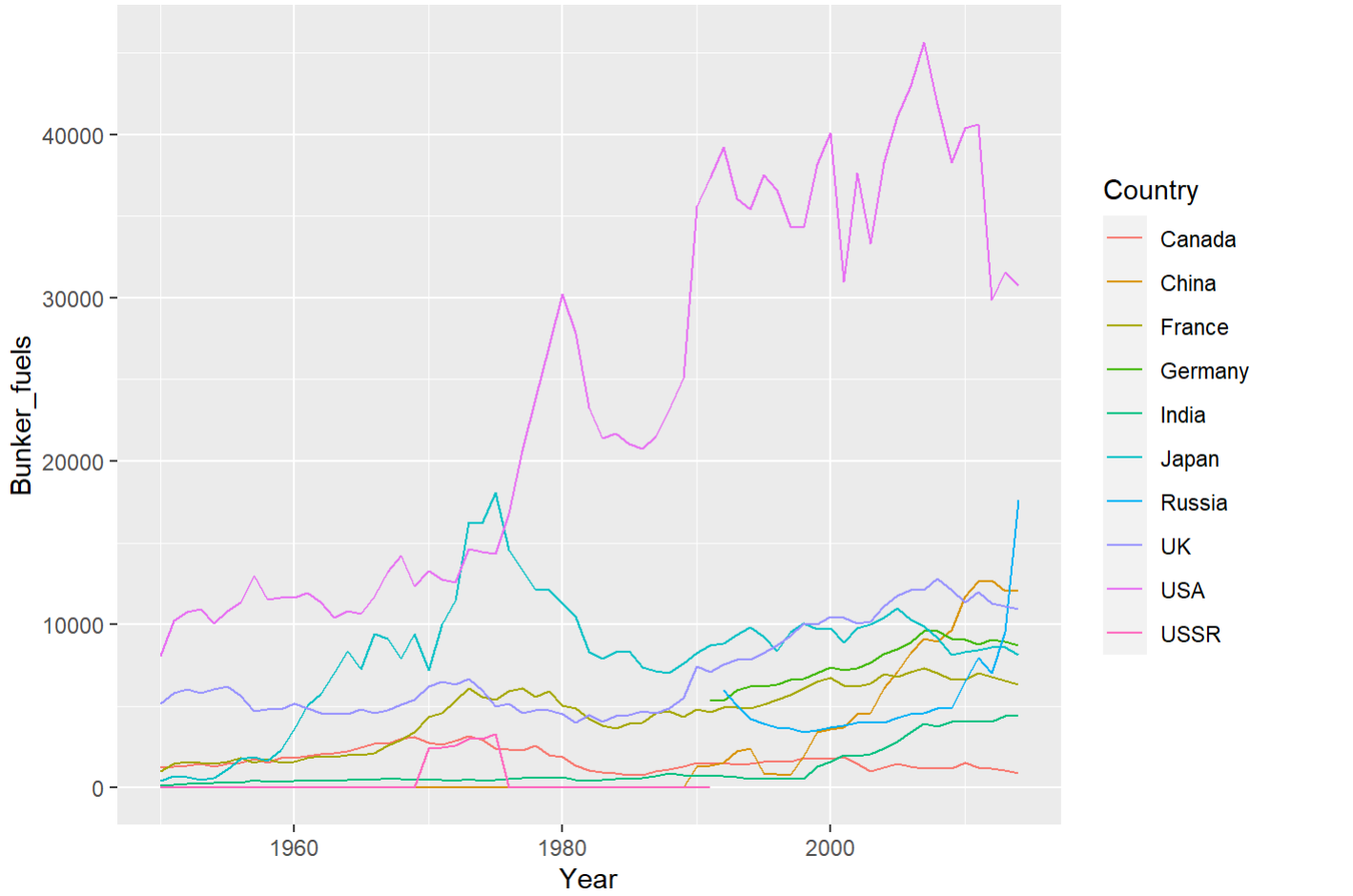

```
ggplot(data = Top, mapping = aes(x = Year)) +
  geom_line(aes(y = Cement,colour = Country))
```




```
ggplot(data = Top, mapping = aes(x = Year)) +
  geom_line(aes(y = Gas_Flaring,colour = Country))
```




```
ggplot(data = Top, mapping = aes(x = Year)) +
  geom_line(aes(y = Bunker_fuels,colour = Country))
```



By using Data frame operations. For using Data frame technique we have prepared a data frame where data is segreated countryvise. Where the data of which is distributed into years of CO2 emission from various Fuels is combined into one Final column which respect to each country.

```
country<- read.csv("F:/Advance_Data_Analytics/Project/CO2/countryviseemissiondata.csv", header=T
RUE)
str(country)
```

```
## 'data.frame':    256 obs. of  9 variables:
##  $ Country     : chr  "AFGHANISTAN" "ALBANIA" "ALGERIA" "ANDORRA" ...
##  $ Total       : int  39133 70600 1003007 3332 142188 706 42 5478 2013085 25463 ...
##  $ Solid_Fuel  : int  9379 17499 30472 0 437 0 0 0 139429 198 ...
##  $ Liquid_Fuel : int  22427 44172 345945 3332 66887 706 42 5478 1145352 5369 ...
##  $ Gas_Fuel    : int  5003 4299 424263 0 8756 0 0 0 630756 18638 ...
##  $ Cement      : int  696 4634 51390 0 4737 0 0 0 50231 1257 ...
##  $ Gas_Flaring : int  1625 0 150948 0 61375 0 0 0 47304 0 ...
##  $ Per_Capita  : num  2.12 26.22 38.81 47.68 10.64 ...
##  $ Bunker_fuels: int  284 406 28381 0 16913 0 1253 2744 36509 905 ...
```

```
country$Country[which.max(country$Total)]
```

```
## [1] "UNITED STATES OF AMERICA"
```

```
country$Country[which.min(country$Total)]
```

```
## [1] "MARSHALL ISLANDS"
```

> From the above operation we can see United States of America emitted the higest CO2 from all the fuels combined and Marshalls Islands emitted the least.

```
country$Country[which.max(country$Solid_Fuel)]
```

```
## [1] "UNITED STATES OF AMERICA"
```

```
country$Country[which.min(country$Solid_Fuel)]
```

```
## [1] "ANDORRA"
```

> United States of America emitted the higest CO2 from all the SOlid Fuels combined and Andorra emitted the least.

```
country$Country[which.max(country$Liquid_Fuel)]
```

```
## [1] "UNITED STATES OF AMERICA"
```

```
country$Country[which.min(country$Liquid_Fuel)]
```

```
## [1] "LIECHTENSTEIN"
```

> United States of America emitted the higest CO2 from all the Liquid Fuels combined and Liechtenstein emitted the least.

```
country$Country[which.max(country$Gas_Fuel)]
```

```
## [1] "UNITED STATES OF AMERICA"
```

```
country$Country[which.min(country$Gas_Fuel)]
```

```
## [1] "ANDORRA"
```

United States of America emitted the higest CO2 from all the Gas Fuels combined and Andorra emitted the least.

```
country$Country[which.max(country$Cement)]
```

```
## [1] "CHINA (MAINLAND)"
```

```
country$Country[which.min(country$Cement)]
```

```
## [1] "ANDORRA"
```

China emitted the higest CO2 from Cement and Andorra emitted the least.

```
country$Country[which.max(country$Gas_Flaring)]
```

```
## [1] "ISLAMIC REPUBLIC OF IRAN"
```

```
country$Country[which.min(country$Gas_Flaring)]
```

```
## [1] "ALBANIA"
```

IRAN emitted the higest CO2 from Gas Flaring and Albania emitted the least.

```
country$Country[which.max(country$Bunker_fuels)]
```

```
## [1] "UNITED STATES OF AMERICA"
```

```
country$Country[which.min(country$Bunker_fuels)]
```

```
## [1] "ANDORRA"
```

United States of America emitted the higest CO2 from the Bunker Fuels and Andorra emitted the least.

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
## -1.99515  0.00694  0.00969  0.01021  2.01263
##
## Coefficients:
##                Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  -1.028e-02  4.640e-03 -2.216e+00   0.0267 *
## Solid_Fuel    1.000e+00  1.981e-07  5.048e+06   <2e-16 ***
## Liquid_Fuel   1.000e+00  2.680e-07  3.732e+06   <2e-16 ***
## Gas_Fuel      1.000e+00  4.048e-07  2.470e+06   <2e-16 ***
## Cement        1.000e+00  1.397e-06  7.156e+05   <2e-16 ***
## Gas_Flaring   1.000e+00  3.422e-06  2.923e+05   <2e-16 ***
## Per_Capita    1.327e-03  1.667e-03  7.960e-01   0.4260
## Bunker_fuels  1.178e-06  2.041e-06  5.770e-01   0.5639
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4524 on 12454 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.454e+14 on 7 and 12454 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.

To see the coorelation matrix of response and the predictors we have to make a small change in the dataset. We have to remove all thode Predictors which are other than Integer because coorelation matrix will only generate if the response and all the predictors are Integer.

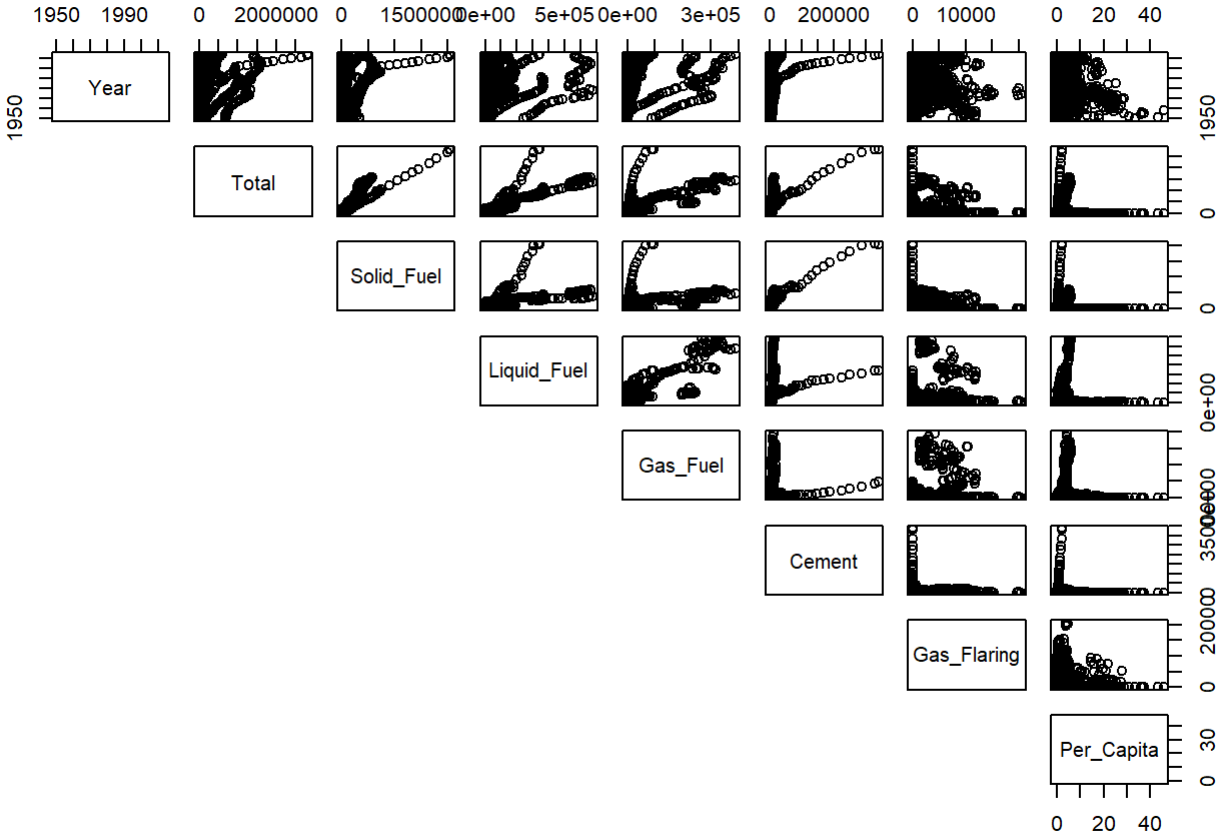
```
EmissionData1<- select(EmissionData,-Country)
str(EmissionData1)
```

```
## 'data.frame':    12462 obs. of  9 variables:
##  $ Year        : int  1950 1950 1950 1950 1950 1950 1950 1950 1950 1950 ...
##  $ Total       : int  23 81 1033 51 8168 14941 5704 15 377 20 ...
##  $ Solid_Fuel  : int  6 12 514 16 972 12028 4744 0 0 1 ...
##  $ Liquid_Fuel : int  18 68 475 34 6982 2739 532 15 377 18 ...
##  $ Gas_Fuel    : int  0 0 0 0 0 0 253 0 0 1 ...
##  $ Cement      : int  0 2 44 0 214 174 175 0 0 0 ...
##  $ Gas_Flaring : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Per_Capita  : num  0 0.07 0.12 0.01 0.48 1.83 0.82 0.19 3.26 0.1 ...
##  $ Bunker_fuels: int  0 0 612 0 124 758 0 3 554 0 ...
```

```
library(ISLR)


i <- sample(2, nrow(EmissionData1), replace=TRUE, prob=c(0.8,0.2))
EmissionDataTraining <- EmissionData1[i==1,]
EmissionDataTesting <- EmissionData1[i==2,]

pairs(EmissionDataTraining[,1:8],lower.panel =NULL)
```



By the coorelation matrix you can see the relation between Response and all the Predictors using plot points.

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

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   -1473.0     106.2     844.5   26426.9    9426.2 2806633.3
```

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

```
## [1] 0.2103703
```

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

```
## [1] 0.2045449
```

And the Mean Absolute Error and Mean Squared Error are very low.The model is the best model.