# PM2.5 Emissions Data Analysis

Anandu R

8/1/2020

## PM2.5 Emissions Data Analysis

Fine particulate matter (PM2.5) is an ambient air pollutant for which there is strong evidence that it is harmful to human health. In the United States, the Environmental Protection Agency (EPA) is tasked with setting national ambient air quality standards for fine PM and for tracking the emissions of this pollutant into the atmosphere. Approximatly every 3 years, the EPA releases its database on emissions of PM2.5. This database is known as the National Emissions Inventory (NEI). You can read more information about the NEI at the EPA National Emissions Inventory web site.

For each year and for each type of PM source, the NEI records how many tons of PM2.5 were emitted from that source over the course of the entire year. The data that is used in the analysis are for 1999, 2002, 2005, and 2008.

### Loading the data

```
fileUrl = "https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip"
if(!file.exists(
  "./data/summarySCC_PM25.rds"
  )||!file.exists(
    "./data/Source_Classification_Code.rds"
    )
  ){
  download.file(fileUrl,"data.zip")
  unzip("data.zip",exdir = "./data")
  unlink("data.zip")
}
```

#### Getting the data

```
NEI <- readRDS("./data/summarySCC_PM25.rds")
SCC <- readRDS("./data/Source_Classification_Code.rds")
```

#### Reading the data

**Preliminary analysis of data** NEI

```
head(NEI)[,c(1,2,4:6)]
```

```
##     fips      SCC Emissions  type year
## 4  09001 10100401    15.714 POINT 1999
## 8  09001 10100404   234.178 POINT 1999
## 12 09001 10100501     0.128 POINT 1999
## 16 09001 10200401     2.036 POINT 1999
## 20 09001 10200504     0.388 POINT 1999
## 24 09001 10200602     1.490 POINT 1999
```

SCC

```
head(SCC)[,c(1,3)]
```

```
##        SCC
## 1 10100101
## 2 10100102
## 3 10100201
## 4 10100202
## 5 10100203
## 6 10100204
##                                                                 Short.Name
## 1                   Ext Comb /Electric Gen /Anthracite Coal /Pulverized Coal
## 2 Ext Comb /Electric Gen /Anthracite Coal /Traveling Grate (Overfeed) Stoker
## 3       Ext Comb /Electric Gen /Bituminous Coal /Pulverized Coal: Wet Bottom
## 4       Ext Comb /Electric Gen /Bituminous Coal /Pulverized Coal: Dry Bottom
## 5                   Ext Comb /Electric Gen /Bituminous Coal /Cyclone Furnace
## 6                   Ext Comb /Electric Gen /Bituminous Coal /Spreader Stoker
```

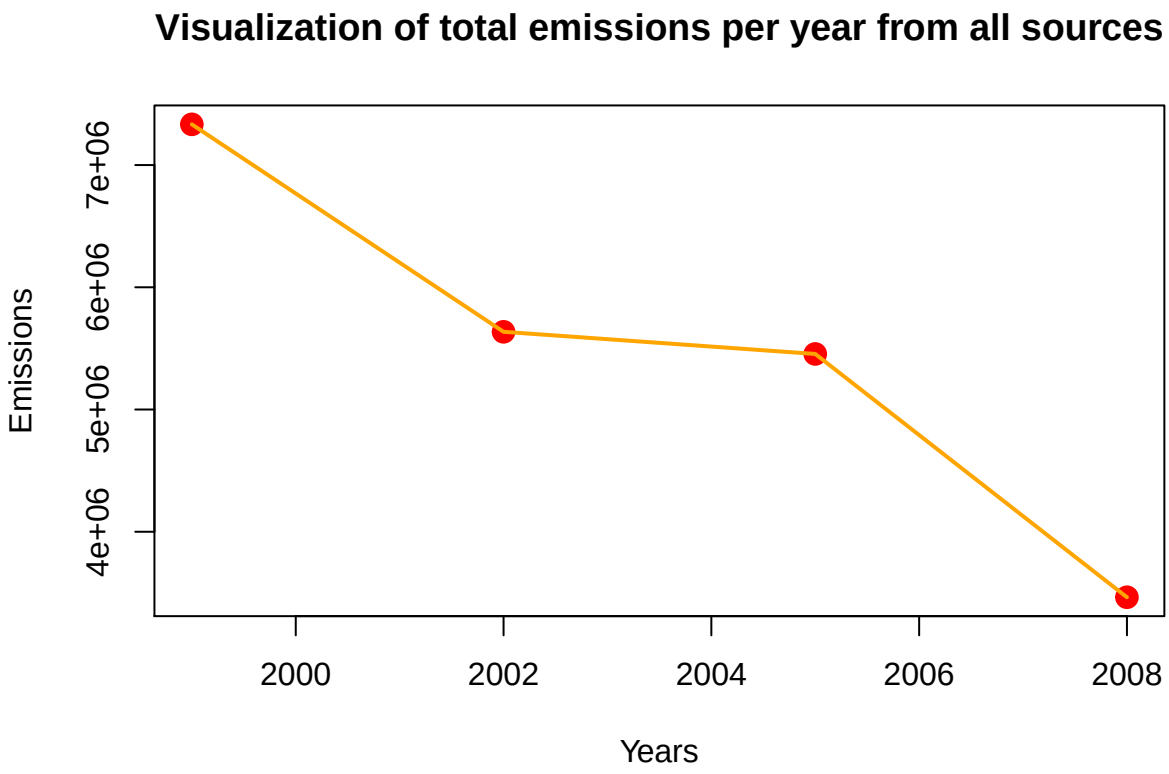
## Exploratory Analysis

**Visualization of total emissions per year from all sources** Using the base plotting system, to make a plot showing the total PM2.5 emission from all sources for each of the years 1999, 2002, 2005, and 2008.

```
dat_total = tapply(NEI$Emissions,factor(NEI$year),sum)
dates = c(1999,2002,2005,2008)
```

Visualising the data

```
plot(dates,
     dat_total,
     cex=1.5,
     col = 'red',
     pch =19,
     xlab = "Years",
     ylab = "Emissions",
     main = "Visualization of total emissions per year from all sources")
lines(dates,as.numeric(dat_total),lwd=2, col = "orange")
```




As we can see from the visualization above the emission levels have dropped from 1999 to 2008

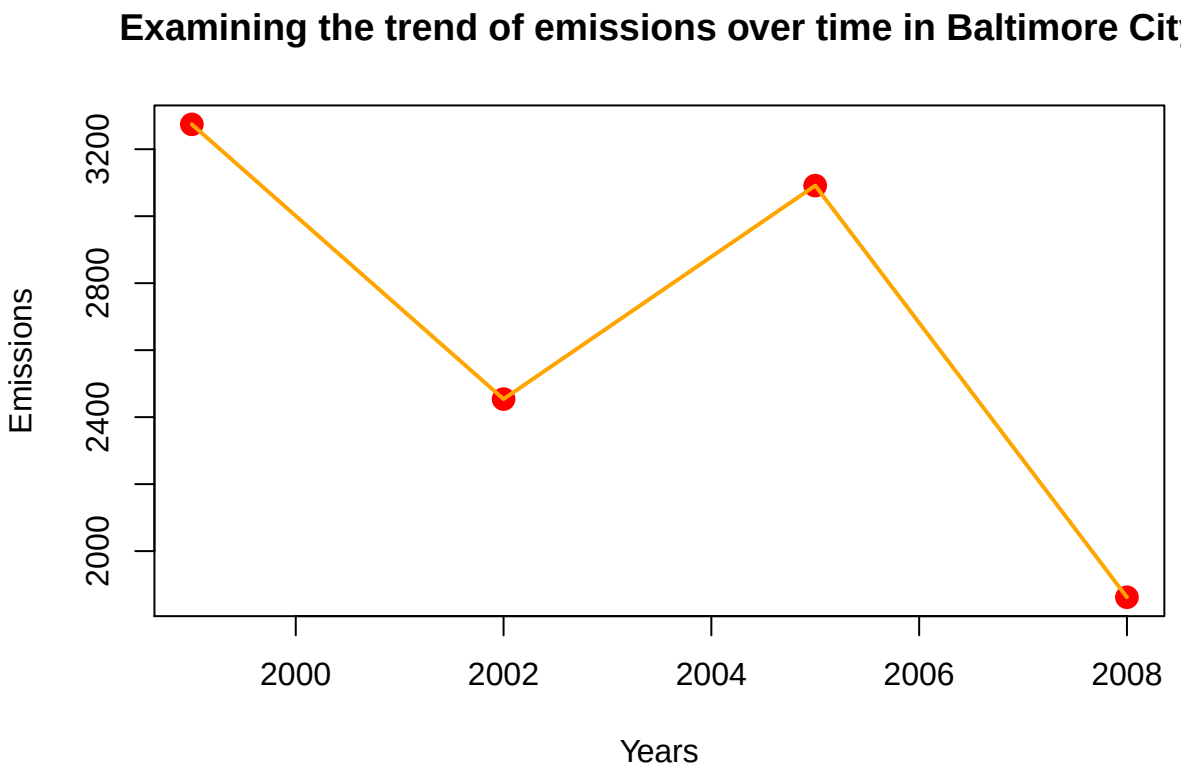
**Examining the trend of emissions over time in Baltimore City** Using the base plotting system, to make a plot showing the total PM2.5 emission in Baltimore City alone for each of the years 1999, 2002, 2005, and 2008.

Subsetting the dataset

```
dat_balt = subset(NEI, fips == "24510")
em_balt = as.numeric(tapply(dat_balt$Emissions,factor(dat_balt$year),sum))
dat_balt_df = data.frame(dates, Emissions = em_balt)
```

Visualising the data

```
plot(
  dat_balt_df,
  cex=1.5,
  col = 'red',
  pch =19,
  xlab = "Years",
  ylab = "Emissions",
  main = "Examining the trend of emissions over time in Baltimore City")
lines(dat_balt_df,lwd=2, col = "orange")
```




We observe that within Baltimore City, the emission levels decrease from 1999 to 2002 then the graph shows an increase from 2002 to 2005 but the emission levels decrease considerably from 2005 to 2008

**Indicating trends in emissions of different Source Types in Baltimore City** There are various source types - point, non-point, onroad, non-road, we want to see which of these source types have had their emissions lowered over the years and which of them had increase in their emission levels.

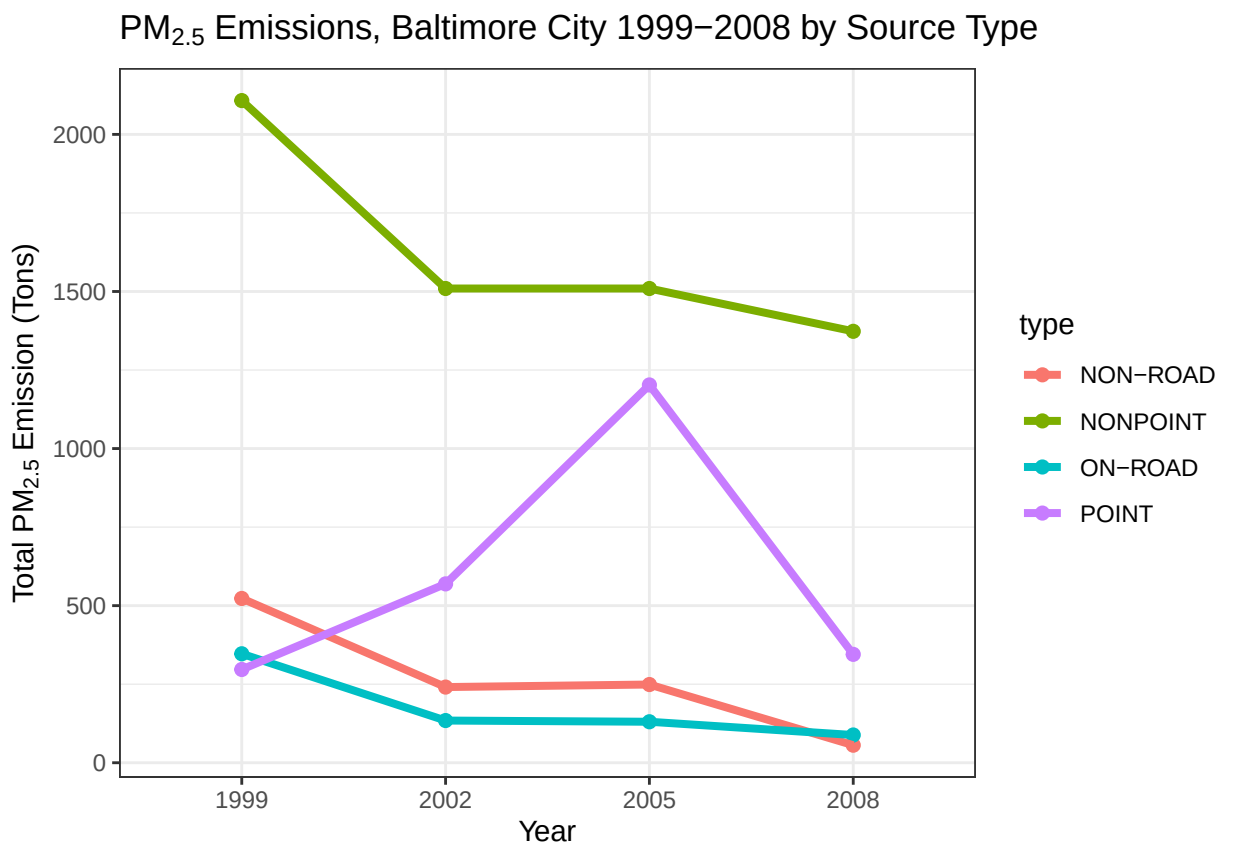
Recasting the data based on need

```
library(reshape2)
data_melt = reshape2::melt(dat_balt[,4:6],id.vars = c('year','type'))
data_cast = reshape2::dcast(data_melt, year+type~ variable, fun.aggregate = sum)
```

Visualising the data

```
library(ggplot2)
ggplot(
  data_cast,
  aes(
    x = factor(year),
    y = Emissions,
    color = type,
    group = type
  ),
) + geom_point(cex = 2) +
  geom_line(cex =1.4) +
```

```
  theme_bw() + guides(fill=FALSE)+
  labs(x="Year", y=expression("Total PM"[2.5]*" Emission (Tons)")) +
  labs(title=expression("PM"[2.5]*" Emissions, Baltimore City 1999-2008 by Source Type"))
```




From the graph above we observe that the emission levels go down overtime for the non-point, non-road and on-road souces, while that of point source has increased from 1999 to 2008

**Emissions from coal combustion-related sources** We are to interpret how emissions from coal combustion-related sources have changed from 1999–2008 across the United States.

We identify the SCC value(s) that represent the coal combustion related sources

We need to identify souces that employ combustion of coal, the process - combustion, is identified in the SCC.Level.One whereas the fuel - coal, is identified by looking at the SCC.Level.Four variable. For the purpose we identify the SCC codes that represent the required data from the SCC dataset.

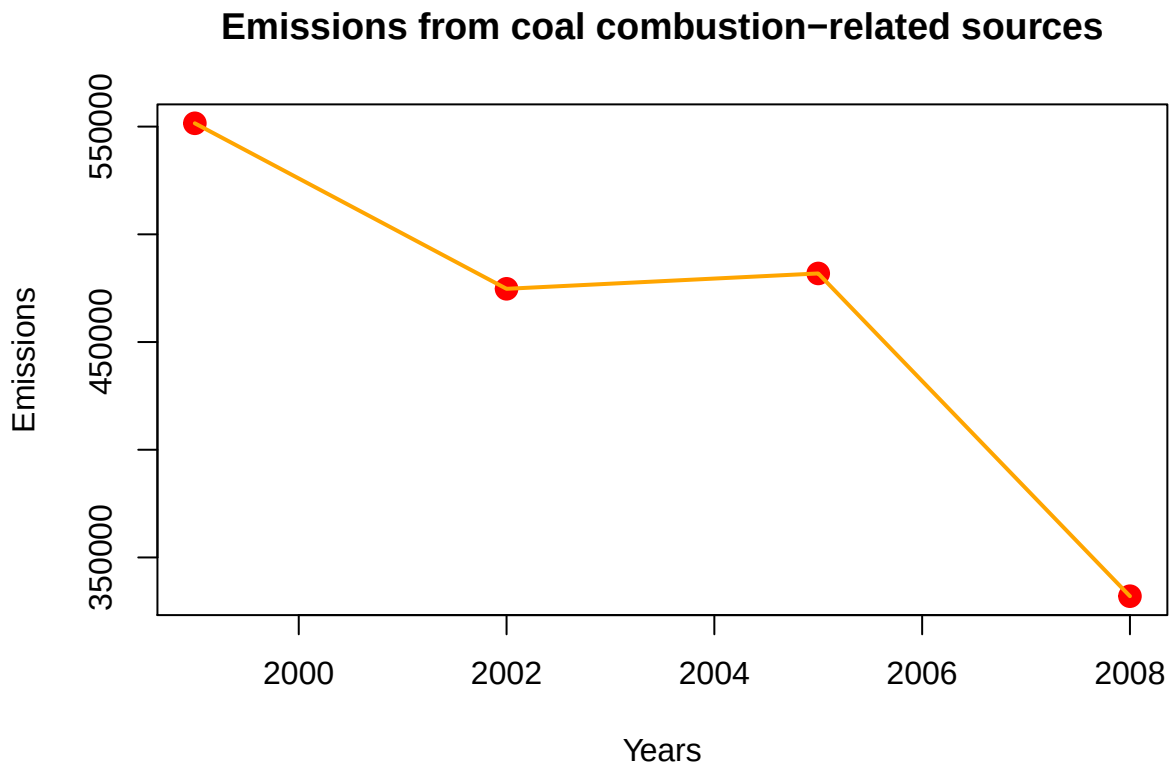
```
combustionRelated = grepl("comb", SCC[, 'SCC.Level.One'], ignore.case=TRUE)
coalRelated = grepl("coal", SCC[, 'SCC.Level.Four'], ignore.case=TRUE)
both = combustionRelated&coalRelated
reqSCC = SCC[both,'SCC']
```

Subsetting the dataset

```
dat_total_cc = subset(NEI, SCC %in% reqSCC)
dat_total_cc = tapply(dat_total_cc$Emissions,factor(dat_total_cc$year),sum)
```

Visualising the data

```
plot(dates,
     dat_total_cc,
     cex=1.5,
     col = 'red',
     pch =19,
     xlab = "Years",
     ylab = "Emissions",
     main = "Emissions from coal combustion-related sources")
lines(dates,as.numeric(dat_total_cc),lwd=2, col = "orange")
```




We observe that the overall emissions do decrease from 1999 to 2008 eventhough there is slight increase from 2002 to 2005.

**Emissions from motor vehicle sources Baltimore City** To interpret how emissions from motor vehicle sources have changed from 1999–2008 in Baltimore City, First we must identify the SCC codes that represent motor vehicle sources from the SCC dataset.

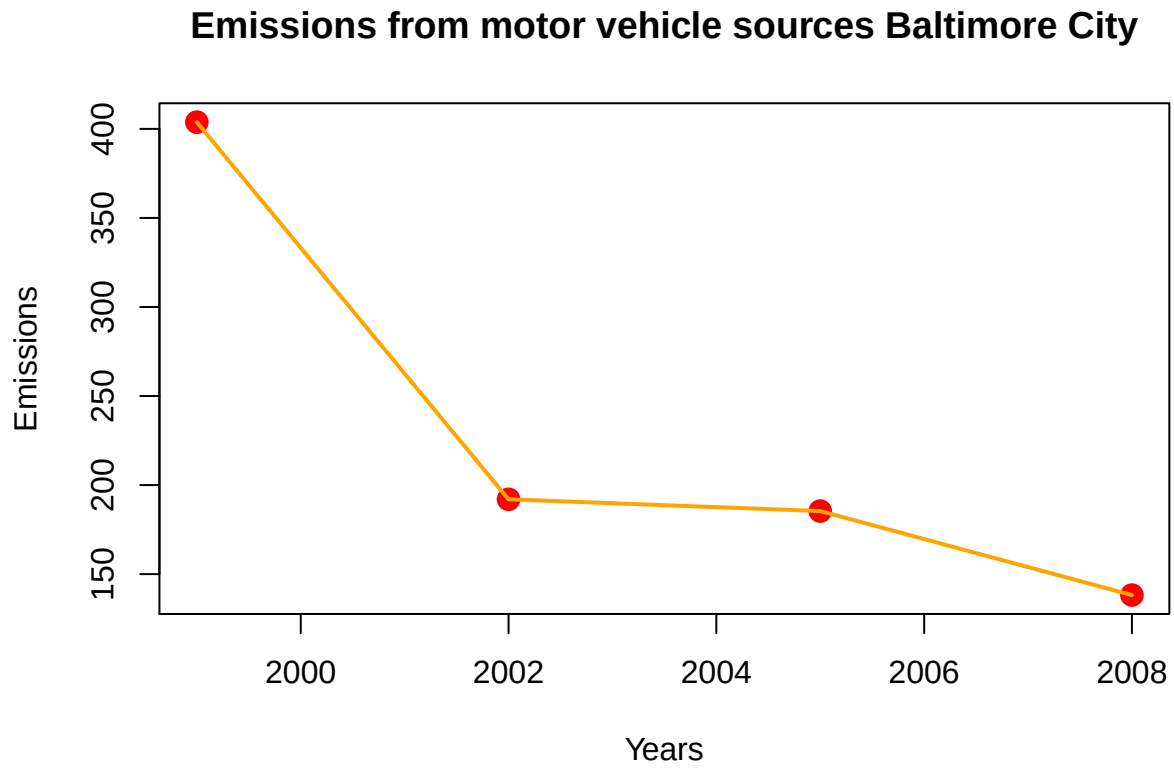
```
motoVehicleRelated = grepl("vehicle", SCC[, 'SCC.Level.Two'], ignore.case=TRUE)
reqSCC = SCC[motoVehicleRelated,'SCC']
```

Subsetting the dataset

```
dat_balt_mv = subset(dat_balt, SCC %in% reqSCC)
dat_balt_mv = tapply(dat_balt_mv$Emissions,factor(dat_balt_mv$year),sum)
```

Visualising the data

```
plot(dates,
     dat_balt_mv,
     cex=1.5,
     col = 'red',
     pch =19,
     xlab = "Years",
     ylab = "Emissions",
     main = "Emissions from motor vehicle sources Baltimore City")
lines(dates,as.numeric(dat_balt_mv), lwd = 2, col = "orange")
```




#### Comparing the emissions between Baltimore City and Los Angeles County Compare emissions from motor vehicle sources in Baltimore City with emissions from motor vehicle sources in Los Angeles County, California (fips == "06037"). To interpret Which city has seen greater changes over time in motor vehicle emissions.

Creating new dataset with the required information

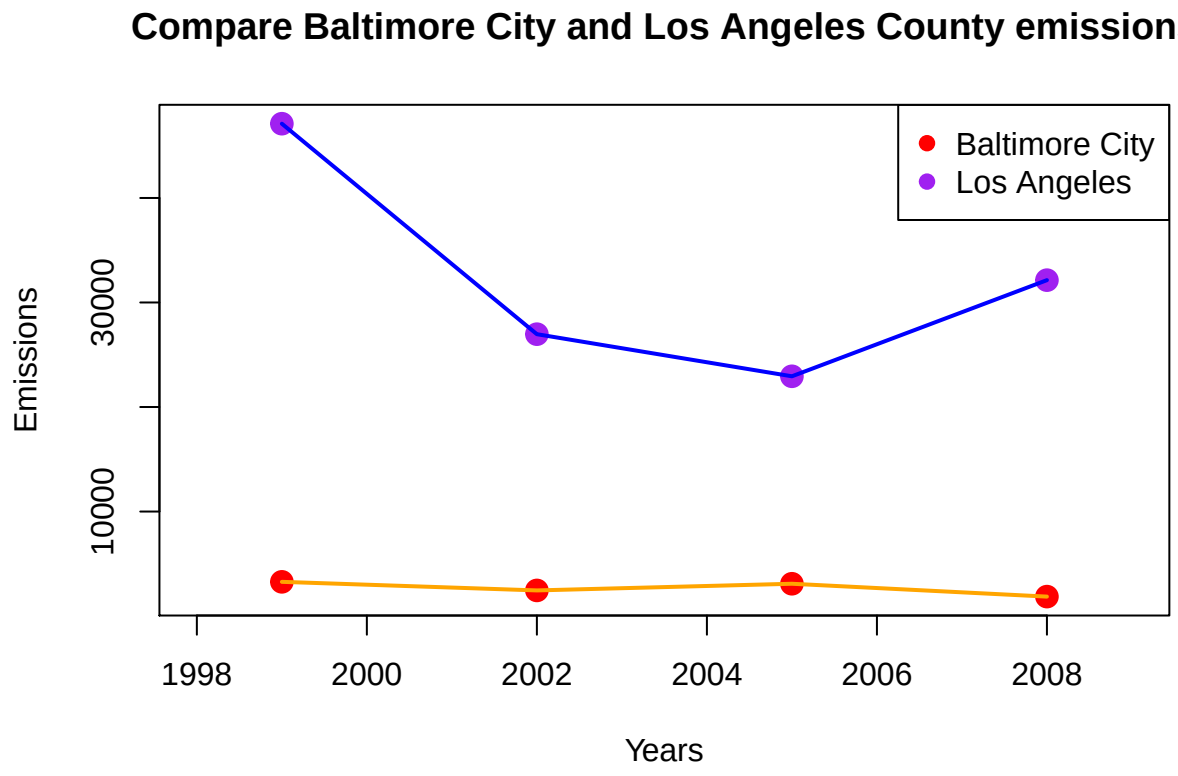
```
dat_losAngeles = subset(NEI, fips == "06037")
em_losAngeles = as.numeric(tapply(dat_losAngeles$Emissions,factor(dat_losAngeles$year),sum))
dat_balt_losAngeles_df = data.frame(dates, Emissions.Baltimore = em_balt, Emissions.LosAngeles = em_los
head(dat_balt_losAngeles_df)
```

```
##   dates Emissions.Baltimore Emissions.LosAngeles
## 1  1999            3274.180             47103.19
## 2  2002            2453.916             26968.79
## 3  2005            3091.354             22939.78
## 4  2008            1862.282             32135.48
```

```
rng = range(dat_balt_losAngeles_df$Emissions.Baltimore,dat_balt_losAngeles_df$Emissions.LosAngeles)
```

Visualising the data

```
plot(
  ylim = rng,
  x = dates,
  type = "n",
  xlim = c(1998,2009),
  xlab = "Years",
  ylab = "Emissions",
  main = "Compare Baltimore City and Los Angeles County emissions")
points(dates,dat_balt_losAngeles_df$Emissions.Baltimore, cex = 1.5, col = "red", pch=19)
points(dates,dat_balt_losAngeles_df$Emissions.LosAngeles, cex = 1.5, col = "purple", pch=19)
lines(dates,dat_balt_losAngeles_df$Emissions.Baltimore, lwd = 2, col = "orange")
lines(dates,dat_balt_losAngeles_df$Emissions.LosAngeles, lwd = 2, col = "blue")
legend("topright", legend=c("Baltimore City","Los Angeles"),col = c("red","purple"),pch =19)
```




Comparing the data between that of Baltimore City and Los Angeles we observe that the emission levels are relatively the same for Baltimore City whereas the emissions of Los Angeles has varied the most from 1999 to 2008 and has considerably decreased as well.

```r
unlink("data", recursive = T)
```

**Removing files after analyis**