# Fine Particulate Matter Emission (PM2.5) in United States

# Exploratory Data Analysis - Course Project 2

**NOTE: My work and answers to the questions are at the bottom of this document.**

# Introduction

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 you will use for this assignment are for 1999, 2002, 2005, and 2008.

# Data

The data for this assignment are available from the course web site as a single zip file:

- Data for Peer Assessment [29Mb]

The zip file contains two files:

PM2.5 Emissions Data (`summarySCC_PM25.rds`): This file contains a data frame with all of the PM2.5 emissions data for 1999, 2002, 2005, and 2008. For each year, the table contains number of tons of PM2.5 emitted from a specific type of source for the entire year. Here are the first few rows.

```
##     fips      SCC Pollutant Emissions  type year
## 4  09001 10100401  PM25-PRI    15.714 POINT 1999
## 8  09001 10100404  PM25-PRI   234.178 POINT 1999
## 12 09001 10100501  PM25-PRI     0.128 POINT 1999
## 16 09001 10200401  PM25-PRI     2.036 POINT 1999
## 20 09001 10200504  PM25-PRI     0.388 POINT 1999
## 24 09001 10200602  PM25-PRI     1.490 POINT 1999
```

- `fips`: A five-digit number (represented as a string) indicating the U.S. county
- `SCC`: The name of the source as indicated by a digit string (see source code classification table)
- `Pollutant`: A string indicating the pollutant

- `Emissions`: Amount of PM2.5 emitted, in tons
- `type`: The type of source (point, non-point, on-road, or non-road)
- `year`: The year of emissions recorded

Source Classification Code Table (`Source_Classification_Code.rds`): This table provides a mapping from the SCC digit strings int he Emissions table to the actual name of the PM2.5 source. The sources are categorized in a few different ways from more general to more specific and you may choose to explore whatever categories you think are most useful. For example, source “10100101” is known as “Ext Comb /Electric Gen /Anthracite Coal /Pulverized Coal”.

You can read each of the two files using the `readRDS()` function in R. For example, reading in each file can be done with the following code:

```
## This first line will likely take a few seconds. Be patient!

NEI <- readRDS("summarySCC_PM25.rds")

SCC <- readRDS("Source_Classification_Code.rds")
```

as long as each of those files is in your current working directory (check by calling `dir()` and see if those files are in the listing).

# Assignment

The overall goal of this assignment is to explore the National Emissions Inventory database and see what it say about fine particulate matter pollution in the United states over the 10-year period 1999-2008. You may use any R package you want to support your analysis.

## Introduction

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 used are for 1999, 2002, 2005, and 2008.

## Data Preparation

The first step is to maker sure the data file is downloaded and extracted

```
# Download archive file, if it does not exist


if(!(file.exists("summarySCC_PM25.rds") &&

    file.exists("Source_Classification_Code.rds"))) {

    archiveFile <- "NEI_data.zip"

    if(!file.exists(archiveFile)) {

        archiveURL <- "https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip"
        download.file(url=archiveURL,destfile=archiveFile,method="curl")
    }
    unzip(archiveFile)
}
```

We now load the NEI and SCC data frames from the .rds files.

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

View the data imported

```
head(NEI)
```

```
##     fips      SCC Pollutant Emissions  type year
## 4  09001 10100401  PM25-PRI    15.714 POINT 1999
## 8  09001 10100404  PM25-PRI   234.178 POINT 1999
## 12 09001 10100501  PM25-PRI     0.128 POINT 1999
## 16 09001 10200401  PM25-PRI     2.036 POINT 1999
## 20 09001 10200504  PM25-PRI     0.388 POINT 1999
## 24 09001 10200602  PM25-PRI     1.490 POINT 1999
```

```
head(SCC)
```

```
##        SCC Data.Category
## 1 10100101         Point
## 2 10100102         Point
## 3 10100201         Point
## 4 10100202         Point
## 5 10100203         Point
## 6 10100204         Point
##                                                                   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
##                                 EI.Sector Option.Group Option.Set
## 1 Fuel Comb - Electric Generation - Coal
## 2 Fuel Comb - Electric Generation - Coal
## 3 Fuel Comb - Electric Generation - Coal
## 4 Fuel Comb - Electric Generation - Coal
## 5 Fuel Comb - Electric Generation - Coal
## 6 Fuel Comb - Electric Generation - Coal
##                 SCC.Level.One       SCC.Level.Two
## 1 External Combustion Boilers Electric Generation
## 2 External Combustion Boilers Electric Generation
## 3 External Combustion Boilers Electric Generation
## 4 External Combustion Boilers Electric Generation
## 5 External Combustion Boilers Electric Generation
## 6 External Combustion Boilers Electric Generation
##                 SCC.Level.Three
## 1               Anthracite Coal
## 2               Anthracite Coal
## 3 Bituminous/Subbituminous Coal
## 4 Bituminous/Subbituminous Coal
## 5 Bituminous/Subbituminous Coal
## 6 Bituminous/Subbituminous Coal
##                                  SCC.Level.Four Map.To Last.Inventory.Year
## 1                               Pulverized Coal     NA                  NA
## 2             Traveling Grate (Overfeed) Stoker     NA                  NA
## 3 Pulverized Coal: Wet Bottom (Bituminous Coal)     NA                  NA
## 4 Pulverized Coal: Dry Bottom (Bituminous Coal)     NA                  NA
## 5             Cyclone Furnace (Bituminous Coal)     NA                  NA
## 6             Spreader Stoker (Bituminous Coal)     NA                  NA
##   Created_Date Revised_Date Usage.Notes
## 1
## 2
## 3
## 4
## 5
## 6
```

Load the packages used in the exploratory analysis

```
library(ggplot2)
library(plyr)
```

Further Pre-processing of the data is done.

```
## Converting "year", "type", "Pollutant", "SCC", "fips" to factor
colToFactor <- c("year", "type", "Pollutant","SCC","fips")
NEI[,colToFactor] <- lapply(NEI[,colToFactor], factor)


head(levels(NEI$fips))
```

```
## [1] "   NA" "00000" "01001" "01003" "01005" "01007"
```

```
## The levels have NA as "   NA", so converting that level back to NA
levels(NEI$fips)[1] = NA
NEIdata<-NEI[complete.cases(NEI),]
colSums(is.na(NEIdata))
```

```
##      fips       SCC Pollutant Emissions      type      year 
##         0         0         0         0         0         0
```

# Questions

Following questions and tasks are targetted by the exploratory analysis

## Question 1

### *Have total emissions from PM2.5 decreased in the United States from 1999 to 2008? Using the base plotting system, make a plot showing the total PM2.5 emission from all sources for each of the years 1999, 2002, 2005, and 2008.*

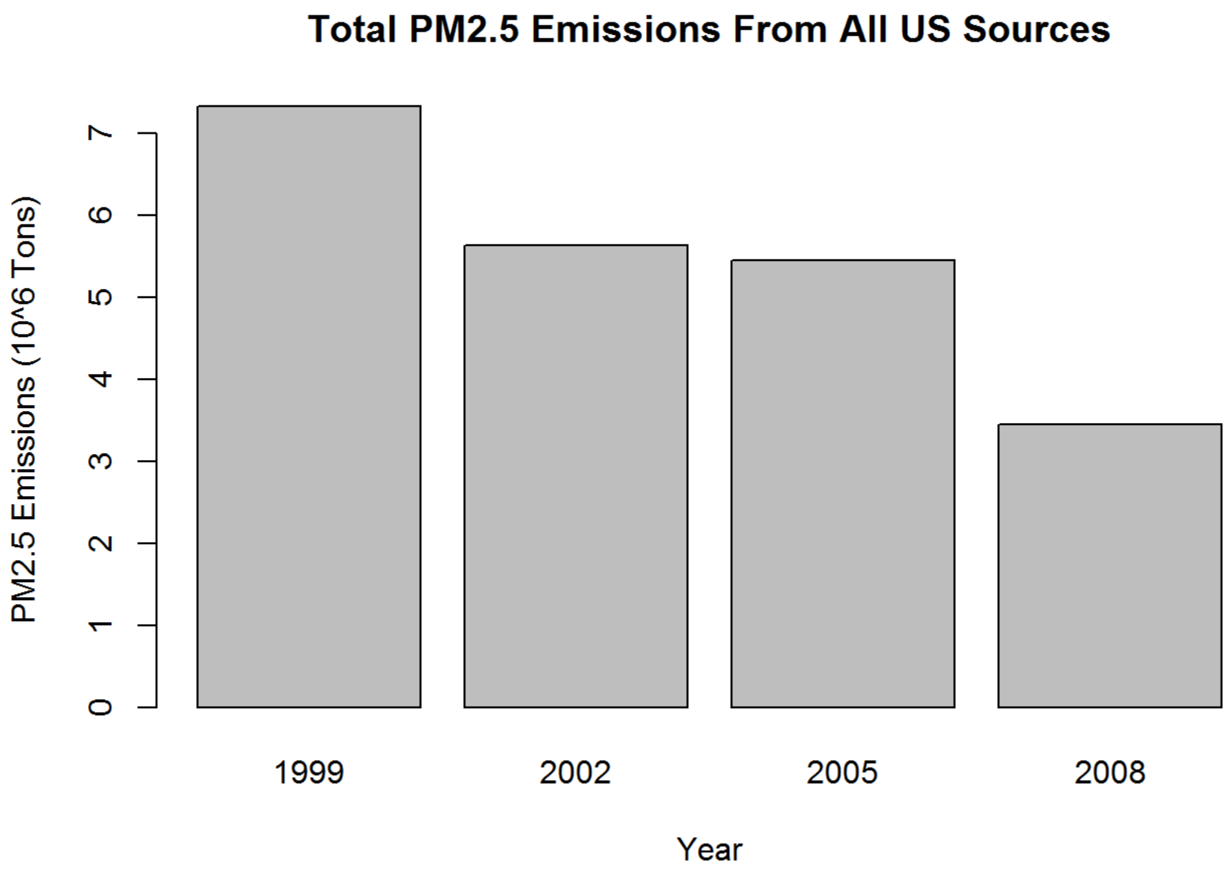
First aggregate the data in

```
totalEmission <- aggregate(Emissions ~ year, NEIdata, sum)
totalEmission
```

```
##   year Emissions
## 1 1999   7332967
## 2 2002   5635780
## 3 2005   5454703
## 4 2008   3456273
```

Plotting the total Emissions over time using a base plotting

```
barplot(

  (totalEmission$Emissions)/10^6,

  names.arg=totalEmission$year,

  xlab="Year",

  ylab="PM2.5 Emissions (10^6 Tons)",

  main="Total PM2.5 Emissions From All US Sources"

)
```



As observed from the plot, the total emissions have decreased in the US from 1999 to 2008

# Question 2

*Have total emissions from PM2.5 decreased in the Baltimore City, Maryland (fips == "24510") from 1999 to 2008? Use the base plotting system to make a plot answering this question.*

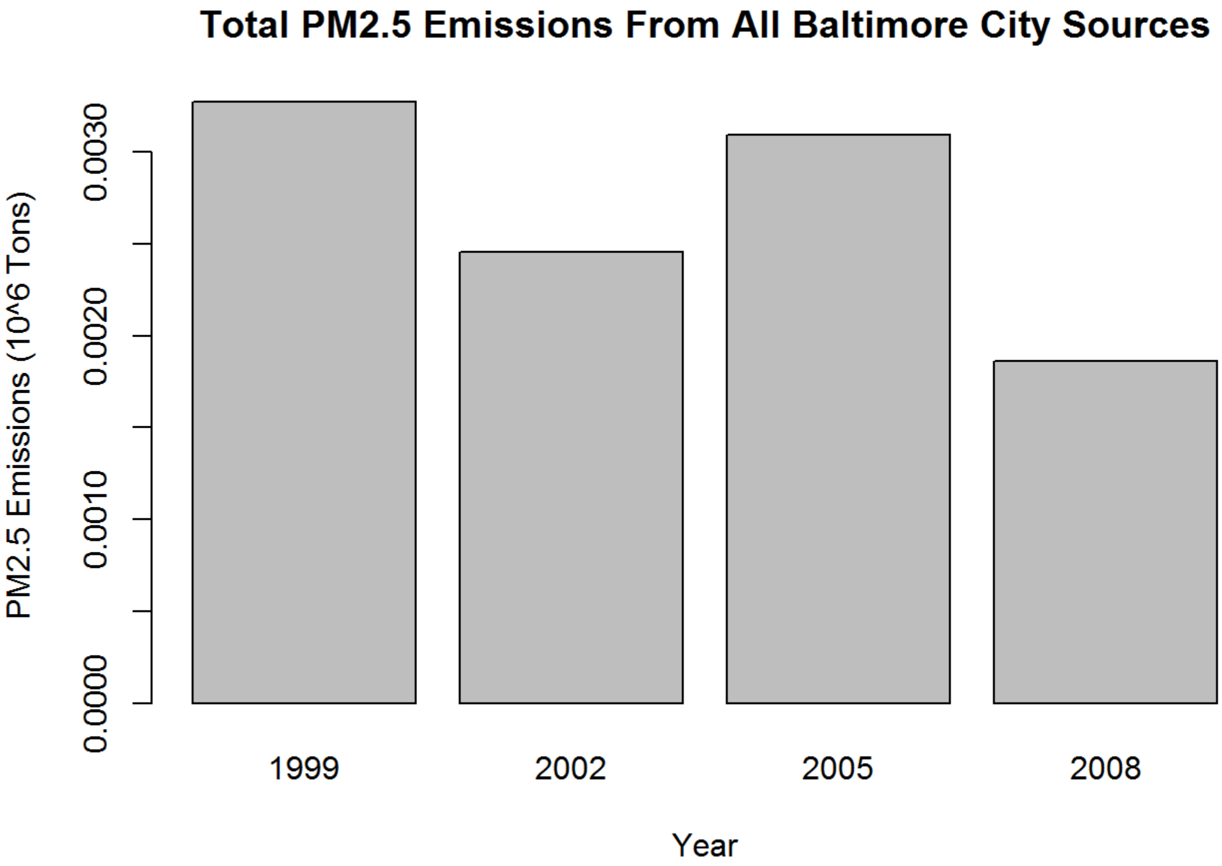
Subset the data for fips == "24510" and then aggregate them by summing the Emissions per years

```
NEIdataBaltimore<-subset(NEIdata, fips == "24510")

totalEmissionBaltimore <- aggregate(Emissions ~ year, NEIdataBaltimore, sum)

totalEmissionBaltimore
```

```
##   year Emissions
## 1 1999      3274
## 2 2002      2454
## 3 2005      3091
## 4 2008      1862
```

Plotting the Total Emissions for baltimore over Time

```
barplot(
  (totalEmissionBaltimore$Emissions)/10^6,
  names.arg=totalEmissionBaltimore$year,
  xlab="Year",
  ylab="PM2.5 Emissions (10^6 Tons)",
  main="Total PM2.5 Emissions From All Baltimore City Sources"
)
```

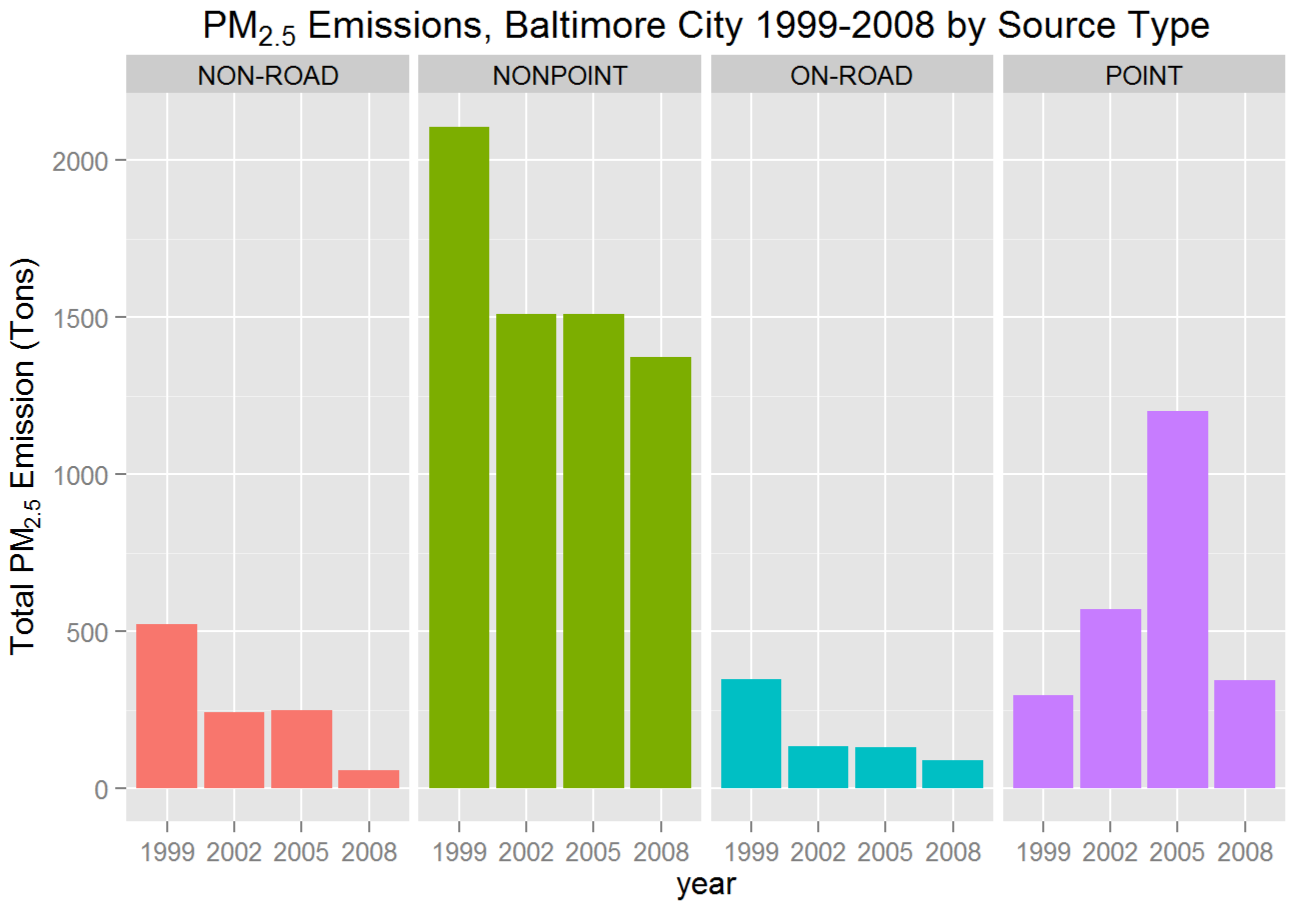


As Observed, The total PM2.5 have not continously decreased, They decreased from 1999 to 2002, but have increasedin 2005 and then decreased.

# Question 3

*Of the four types of sources indicated by the type (point, nonpoint, onroad, nonroad) variable, which of these four sources have seen decreases in emissions from 1999-2008 for Baltimore City? Which have seen increases in emissions from 1999-2008? Use the ggplot2 plotting system to make a plot answer this question.*

```
g<-ggplot(aes(x = year, y = Emissions, fill=type), data=NEIdataBaltimore)

g+geom_bar(stat="identity")+

  facet_grid(.~type)+

  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"))+

  guides(fill=FALSE)
```



As seen from the graphs, the “NON-ROAD”, “NONPOINT” and “ON-ROAD” type of sources have shown a decrease in the total PM2.5 Emissions. “POINT” type of source, shows the increase in the total PM2.5 emissions from 1999-2005 but again a decrease in 2008

# Question 4

*Across the United States, how have emissions from coal combustion-related sources changed from 1999-2008?*

```
## making the names in the SCC dataframe pretty by removing \\. in all the names
```

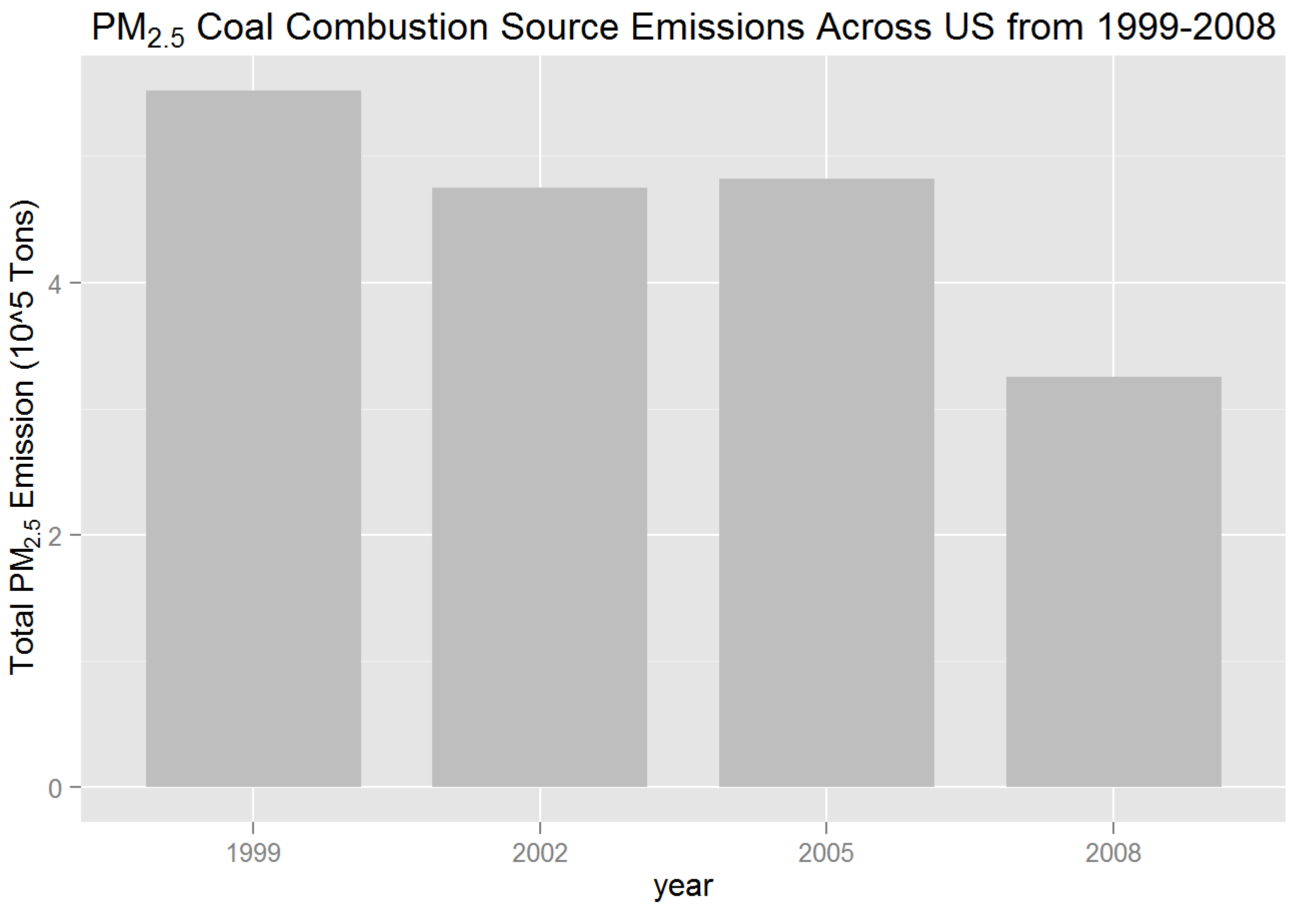
```
names(SCC)<-gsub("\\.","", names(SCC))
```

Note: The SCC levels go from generic to specific. We assume that coal combustion related SCC records are those where SCC.Level.One contains the substring ‘comb’ and SCC.Level.Four contains the substring ‘coal’.

```
SCCcombustion<-grepl(pattern = "comb", SCC$SCCLevelOne, ignore.case = TRUE)
SCCCoal<-grepl(pattern = "coal", SCC$SCCLevelFour, ignore.case = TRUE)


## extracting the SCC in
SCCCoalCombustionSCC<-SCC[SCCcombustion & SCCCoal,]$SCC
NIECoalCombustionValues<-NEIdata[NEIdata$SCC %in% SCCCoalCombustionSCC,]
NIECoalCombustionTotalEm<-aggregate(Emissions~year, NIECoalCombustionValues, sum)
```

Plotting the subset of NEI data with SCC matched with coal and combustion.

```
g<-ggplot(aes(year, Emissions/10^5), data=NIECoalCombustionTotalEm)
g+geom_bar(stat="identity",fill="grey",width=0.75) +
  guides(fill=FALSE) +
  labs(x="year", y=expression("Total PM"[2.5]*" Emission (10^5 Tons)")) +
  labs(title=expression("PM"[2.5]*" Coal Combustion Source Emissions Across US from 1999-2008"))
```



As is viewed in the graph, Coal cumbustion is showing a decreasing trend with a sligh increase from 2002-2005, and then a decrease after

# Question 5

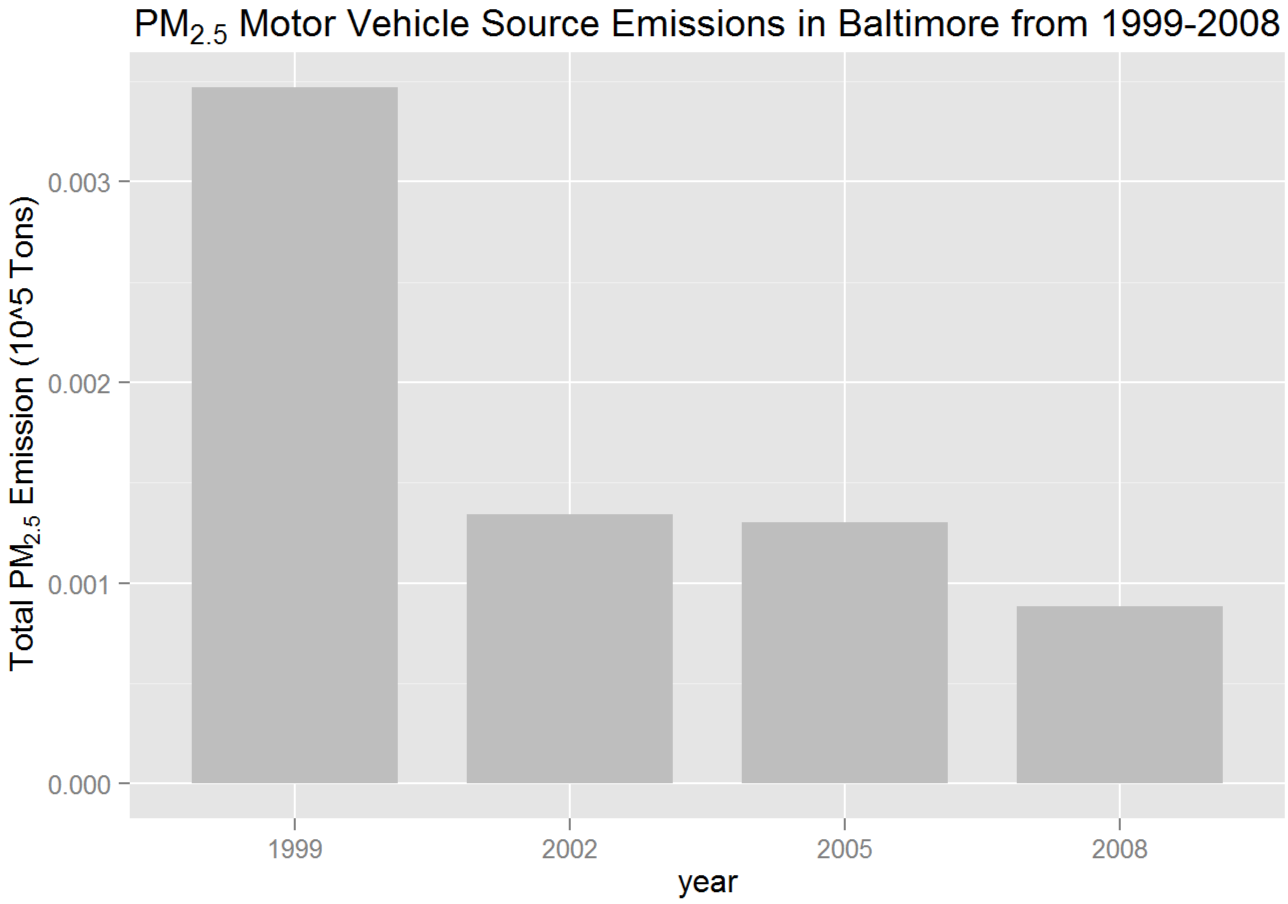
### *How have emissions from motor vehicle sources changed from 1999-2008 in Baltimore City?*

First we subset the motor vehicles, which we assume is anything like Vehicle in EISector column

```
SCCvehicle<-grepl(pattern = "vehicle", SCC$EISector, ignore.case = TRUE)
SCCvehicleSCC <- SCC[SCCvehicle,]$SCC

## using this boolean vector get the interested rows from the baltimore data
NEIvehicleSSC <- NEIdata[NEIdata$SCC %in% SCCvehicleSCC, ]
NEIvehicleBaltimore <- subset(NEIvehicleSSC, fips == "24510")
NIEvehicleBaltimoreTotEm<-aggregate(Emissions~year, NEIvehicleBaltimore, sum)
```

Plotting the year-Emissions

```
g<-ggplot(aes(year, Emissions/10^5), data=NIEvehicleBaltimoreTotEm)
g+geom_bar(stat="identity",fill="grey",width=0.75) +
  guides(fill=FALSE) +
  labs(x="year", y=expression("Total PM"[2.5]*" Emission (10^5 Tons)")) +
  labs(title=expression("PM"[2.5]*" Motor Vehicle Source Emissions in Baltimore from 1999-2008"))
```



## Question 6

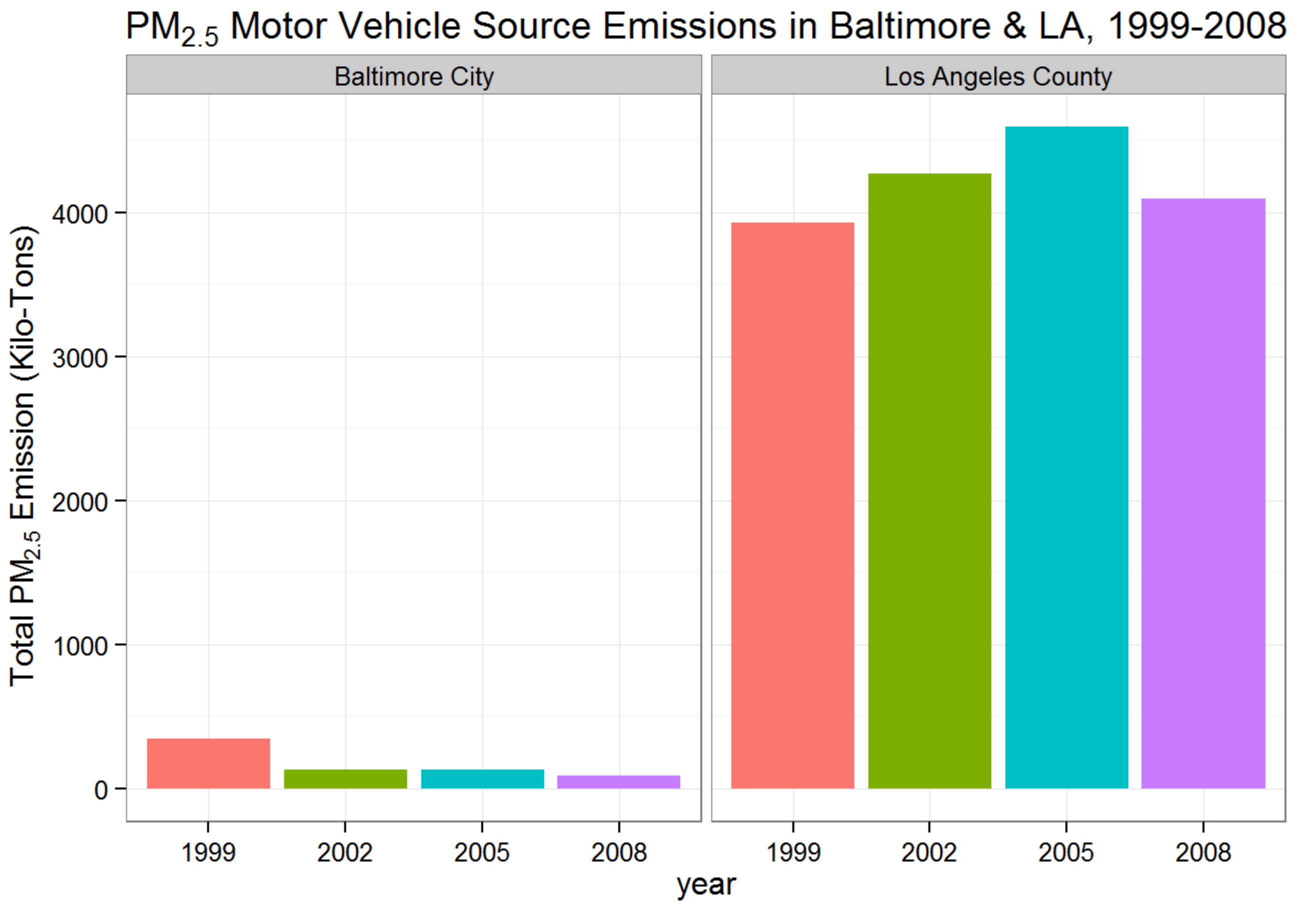
*Compare emissions from motor vehicle sources in Baltimore City with emissions from motor vehicle sources in Los Angeles County, California (fips == “06037”). Which city has seen greater changes over time in motor vehicle emissions?*

Subset the VehiclesSSC based on fips and add a column for the city name, then combine the two data frame to generate data for both cities

```
NEIvehicleBalti<-subset(NEIvehicleSSC, fips == "24510")

NEIvehicleBalti$city <- "Baltimore City"

NEIvehiclela<-subset(NEIvehicleSSC, fips == "06037")

NEIvehiclela$city <- "Los Angeles County"

NEIBothCity <- rbind(NEIvehicleBalti, NEIvehiclela)
```

Plot the result by facets for each city

```
 ggplot(NEIBothCity, aes(x=year, y=Emissions, fill=city)) +

  geom_bar(aes(fill=year),stat="identity") +

 facet_grid(.~city) +

 guides(fill=FALSE) + theme_bw() +
 labs(x="year", y=expression("Total PM"[2.5]*" Emission (Kilo-Tons)")) +
 labs(title=expression("PM"[2.5]*" Motor Vehicle Source Emissions in Baltimore & LA, 1999-2008"))
```



To View the maximum change in the emission levels

```
aggregateEmissions <- aggregate(Emissions~city+year, data=NEIBothCity, sum)
aggregate(Emissions~city, data=aggregateEmissions, range)
```

```
##                 city Emissions.1 Emissions.2
## 1     Baltimore City       88.28      346.82
## 2 Los Angeles County     3931.12     4601.41
```

So it is observed that Los Angeles County has seen greater changes over time in vehicle emissions