Exploratory Data Project 2

Sherri Verdugo

July 26, 2014

Contents

Instruction

This is the peer assessment project number two for the following Coursera course:

- Institution: John Hopkins Bloomberg School of Public Health
- Class: Exploratory Data Analysis
- Part of the data scientist specialty track

The overall goal of this assigment is to explore the National Emissions Inventory database and see what it says about fine particulate matter pollution in the United States over the 10-year period 1999 to 2008. The necesary library packages are documented in each r code section.

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. Approximately 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.

Operating System and Environment Specs

- Using Mac OS X version 10.9.4
- R Studio version 0.98.953 – © 2009-2013 RStudio, Inc.
- R Project version R 3.1.0 GUI 1.64 Mavericks build

Data for Peer Assessment [29Mb]

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

- summarySCC_PM25.rds
- Source_Classification_Code.rds

Downloading the data

This is achieved by setting the working directory and using the 1_download_data.R script. Since the items are downloaded previously, we don't need to evaluate this part of the r code.

```
## Script Name: 1_download_data.R
## Version: 1.0_14

## Libraries needed
library(utils)

## Set working directory
setwd("~/Google Drive/Coursera_R_studio/exploratory_data/project2")

## Download the data
my.file = "expdata_prj2.zip"
if (!file.exists(my.file)) {
  retval = download.file("
  https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip",
  destfile = my.file, method = "curl")
}
```

List of Questions

1. Have total emissions from PM2.5 decreased in the United States from 1999 to 2008? Using the base plotting system only, make a plot showing the total PM2.5 emissions from all sources for each of the years 1999, 2002, 2005, and 2008.
2. Have total emissions from PM_2.5 decreased in the Baltimore City, Maryland (`fips == "24510"`) from 1999 to 2008? Use the base plotting system to make a plot answering this question.
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 to 2008 for **Baltimore City**? Which have seen increases in emissions from 1999 to 2008? Use the **ggplot2** plotting system to make a plot to answer this question.
4. Across the United States, how have emissions from coal combustion-related sources changed from 1999 to 2008?
5. How have emissions from motor vehicle sources changed from 1999 - 2008 in *Baltimore City*?
6. Compare emissions from motor vehicle sources in Baltimore City with emissions from motor vehicle sources in *Los Angeles County*, California, California (`fips == "06037"`). Which city has seen greater changes over time in motor vehicle emissions?

Question 1: plot1.r

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

R Code for question 1:

```
## Script Name: plot1.R
## Version: 1.0_14

## Libraries needed: no special libraries needed for this part

## Pre-Step One: Set working directory
setwd("~/Google Drive/Coursera_R_studio/exploratory_data/project2")

## Step 1: read in the data
NEI <- readRDS("expdata_prj2/summarySCC_PM25.rds")
SCC <- readRDS("expdata_prj2/Source_Classification_Code.rds")

length(NEI$Emissions)
```

```
## [1] 6497651
```

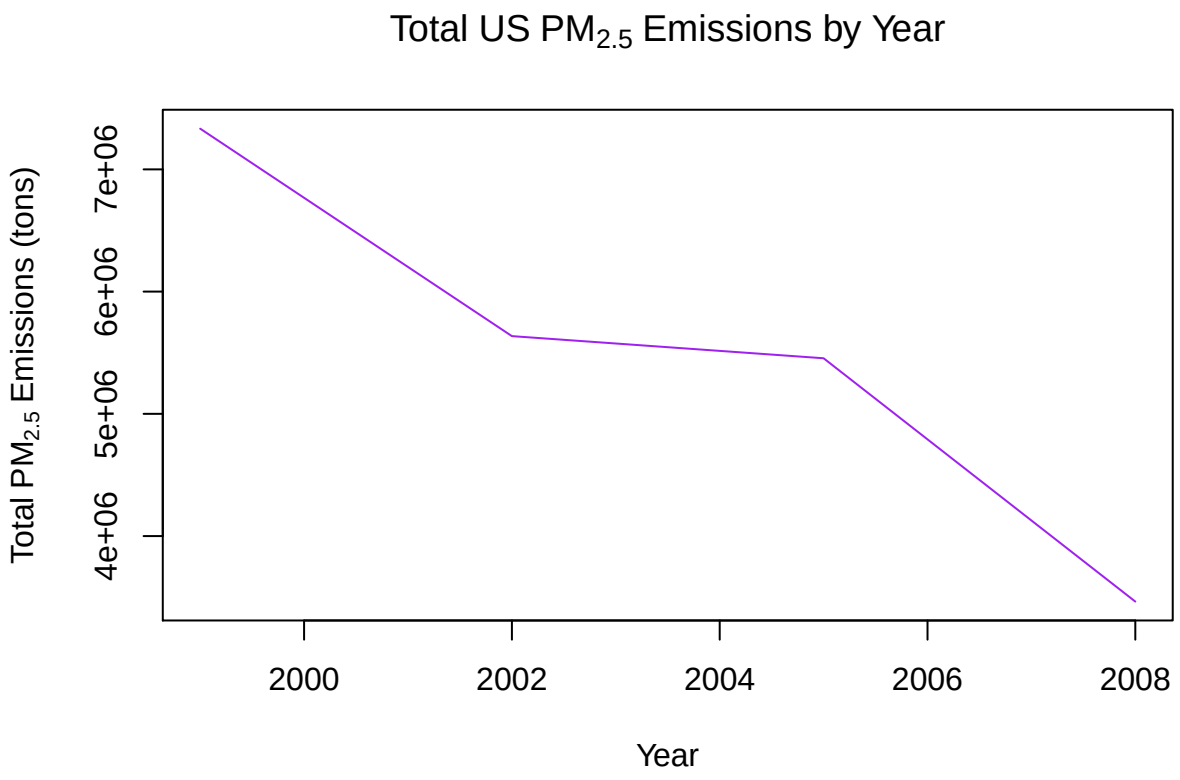
```
length(NEI$year)
```

```
## [1] 6497651
```

```r
tot.PM25yr <- tapply(NEI$Emissions, NEI$year, sum)

###Step 2: prepare to plot to png
png("plot1.png")
plot(names(tot.PM25yr), tot.PM25yr, type="l", xlab = "Year", ylab = expression
     ("Total" ~ PM[2.5] ~"Emissions (tons)"), main = expression("Total US" ~
     PM[2.5] ~ "Emissions by Year"), col="Purple")
dev.off()
```

```
## pdf
##   2
```

```r
###Step 3: prepare to plot to the markdown file
plot(names(tot.PM25yr), tot.PM25yr, type="l", xlab = "Year", ylab = expression ("Total" ~ PM[2.5] ~"Emis
```



Answer:

Yes, they sharply declined from 1999 to 2002. Then a slower decline between 2002 and 2005. Finally, they sharply declined from 2005 to 2008.

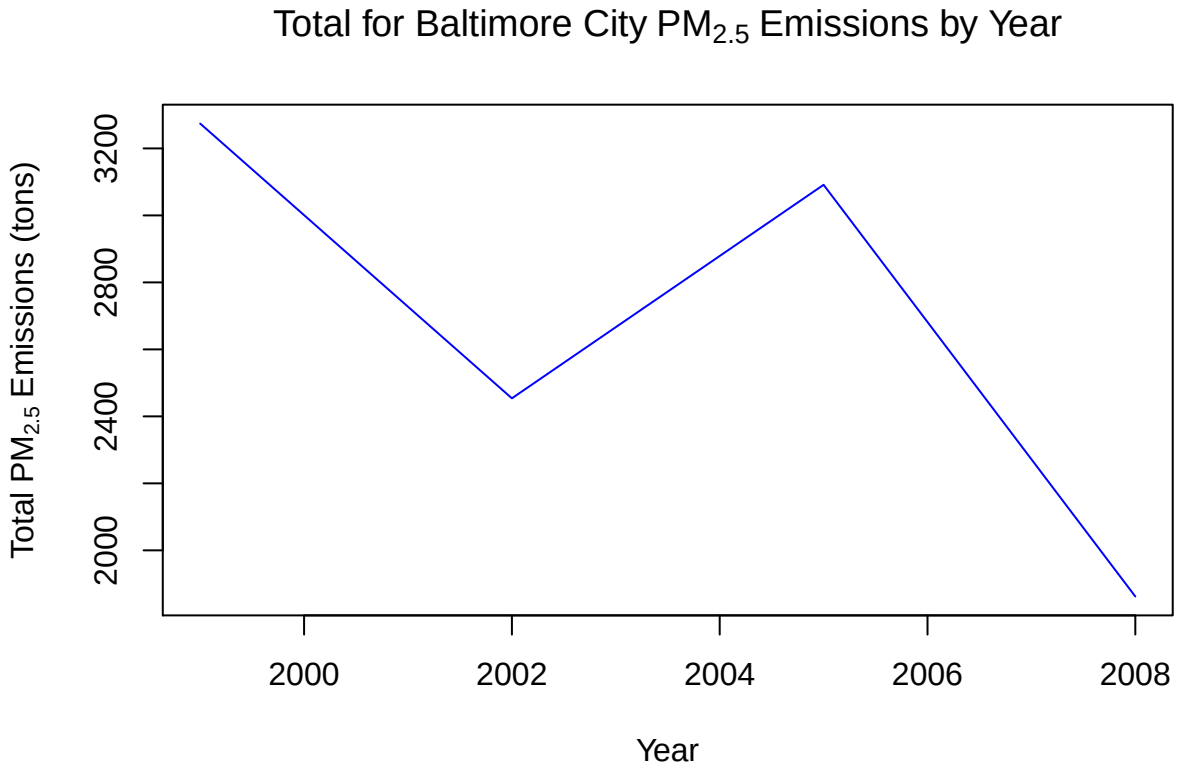
Question 2: plot2.r

Have total emissions from PM_2.5 decreased in the Baltimore City, Maryland (`fips == "24510"`) from 1999 to 2008? Use the base plotting system to make a plot answering this question.

```r
## Script Name: plot2.R

## Version: 1.0_14

## Libraries needed: no special libraries needed for this part

## Pre-Step One: Set working directory
setwd("~/Google Drive/Coursera_R_studio/exploratory_data/project2")

## Step 1: read in the data
NEI <- readRDS("expdata_prj2/summarySCC_PM25.rds")
SCC <- readRDS("expdata_prj2/Source_Classification_Code.rds")

## Step 2: obtain the subsets to plot
baltimore <- subset (NEI, fips == "24510")
total.PM25yr <- tapply(baltimore$Emissions, baltimore$year, sum)

## Step 3: plot prepare to plot to png
png("plot2.png")
plot(names(total.PM25yr), total.PM25yr, type = "l", xlab="Year", ylab= expression("Total" ~ PM[2.5] ~ "
dev.off()
```

```
## pdf
##   2
```

```r
## Step 4: plot to markdown file
plot(names(total.PM25yr), total.PM25yr, type = "l", xlab="Year", ylab=expression("Total" ~ PM[2.5] ~ "Em
```



Answer:

The data indicate a sharp decline between 1999 and 2002. A sharp increase occurred from 2002 to 2005. Finally, another sharp decrease occurred from 2005 to 2008.

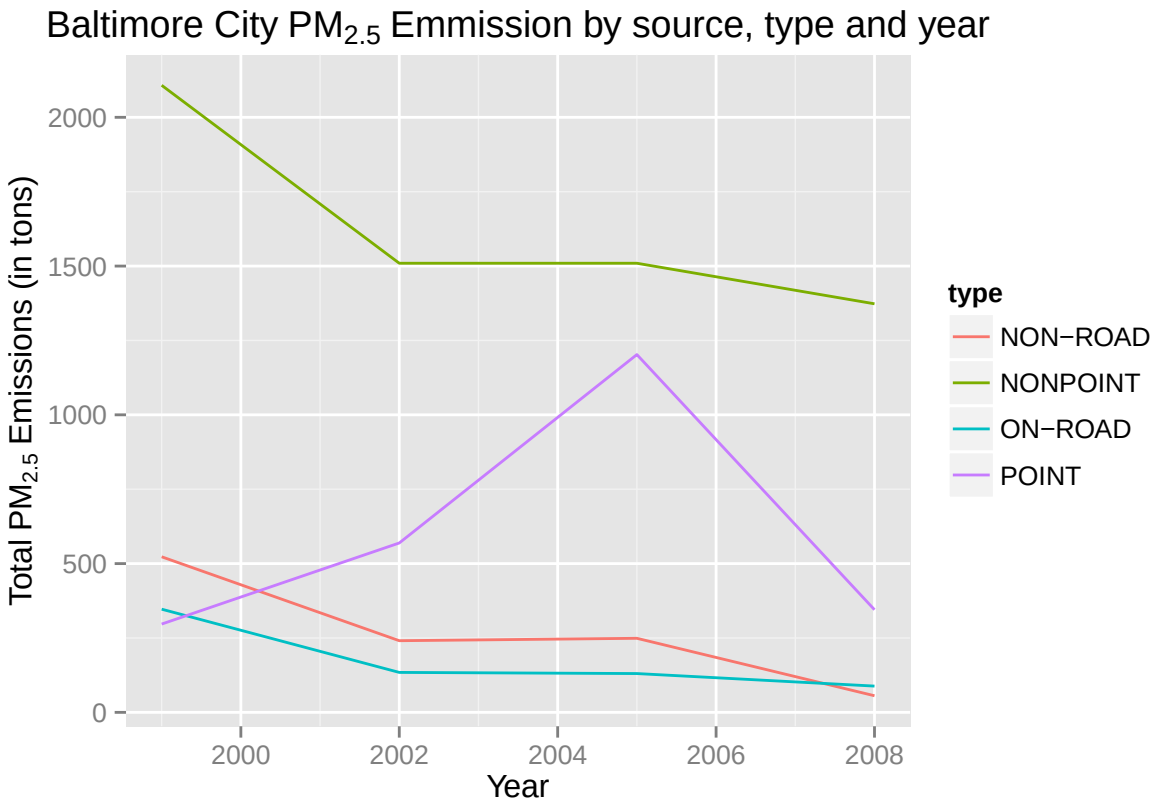
Question 3: plot3.r

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 to 2008 for **Baltimore City**? Which have seen increases in emissions from 1999 to 2008? Use the **ggplot2** plotting system to make a plot to answer this question.

```
## Script Name: plot3.R
## Version: 1.0_14

## Libraries needed:
library(ggplot2)
library(plyr)

## Step 1: read in the data
NEI <- readRDS("expdata_prj2/summarySCC_PM25.rds")
SCC <- readRDS("expdata_prj2/Source_Classification_Code.rds")

## Step 2: subset our data
baltimore <- subset (NEI, fips == "24510")
typePM25.year <- ddply(baltimore, .(year, type), function(x) sum(x$Emissions))

## Rename the col: Emissions
colnames(typePM25.year)[3] <- "Emissions"

## Step 3: prepare to plot to png
png("plot3.png")
qplot(year, Emissions, data=typePM25.year, color=type, geom ="line") + ggtitle(expression("Baltimore Ci
dev.off()
```

```
## pdf
##   2
```

```
## Step 4: prepare to plot to markdown
qplot(year, Emissions, data=typePM25.year, color=type, geom ="line") + ggtitle(expression("Baltimore Ci
```



Answer:

- **Nonpoint (green line):** From the plot, we see that nonpoint (green line) sharply decreased from 1999 to 2002. It remained steady from 2002 to 2005 with 1,500 Total $PM_{2.5}$ emissions. Finally, a slight decrease occurred between 2005 and 2008 from 1,500 Total $PM_{2.5}$ emissions.
- **Point (purple line):** From the plot, we see that the point (purple line) slightly increased from 1999 to 2002. It then sharply increased in $PM_{2.5}$ emissions from 2002 to 2005. Finally, from 2005 to 2008, the $PM_{2.5}$ emissions sharply decreased.
- **Onroad (blue line):** From the plot, we see that the onroad (blue line) slightly decreased from 1999 to 2002. It remained approximately steady from 2002 to 2005 and continued this trend from 2005 to 2008. In comparison to the nonroad values, this over all trend was lower compared to the nonroad values.
- **Nonroad (red line):** From the plot, we see that the nonroad (red line) followed the same path as the onroad values only slightly higher in $PM_{2.5}$ emissions values. slightly decreased from 1999 to 2002. It remained approximately steady from 2002 to 2005 and continued this trend from 2005 to 2008.

Question 4: plot4.r

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

```
## Script Name: plot4.R
## Version: 1.0_14

## Libraries needed:
```

```
library(plyr)
library(ggplot2)

## Step 1: read in the data
NEI <- readRDS("expdata_prj2/summarySCC_PM25.rds")
SCC <- readRDS("expdata_prj2/Source_Classification_Code.rds")

## Step 2: subset our data for only coal-combustion
coalcomb.scc <- subset(SCC, EI.Sector %in% c("Fuel Comb - Comm/Instutional - Coal",
  "Fuel Comb - Electric Generation - Coal", "Fuel Comb - Industrial Boilers, ICEs -
  Coal"))

## Step 3: comparisons so that we didn't ommit anything weird
coalcomb.scc1 <- subset(SCC, grepl("Comb", Short.Name) & grepl("Coal", Short.Name))

nrow(coalcomb.scc) #evaluate: rows 0
```

```
## [1] 35
```

```
nrow(coalcomb.scc1) #evaluate: rows 91
```

```
## [1] 91
```

```
## Step 4: set the differences
dif1 <- setdiff(coalcomb.scc$SCC, coalcomb.scc1$SCC)
dif2 <- setdiff(coalcomb.scc1$SCC, coalcomb.scc$SCC)

length(dif1)#0
```

```
## [1] 6
```

```
length(dif2)#91
```

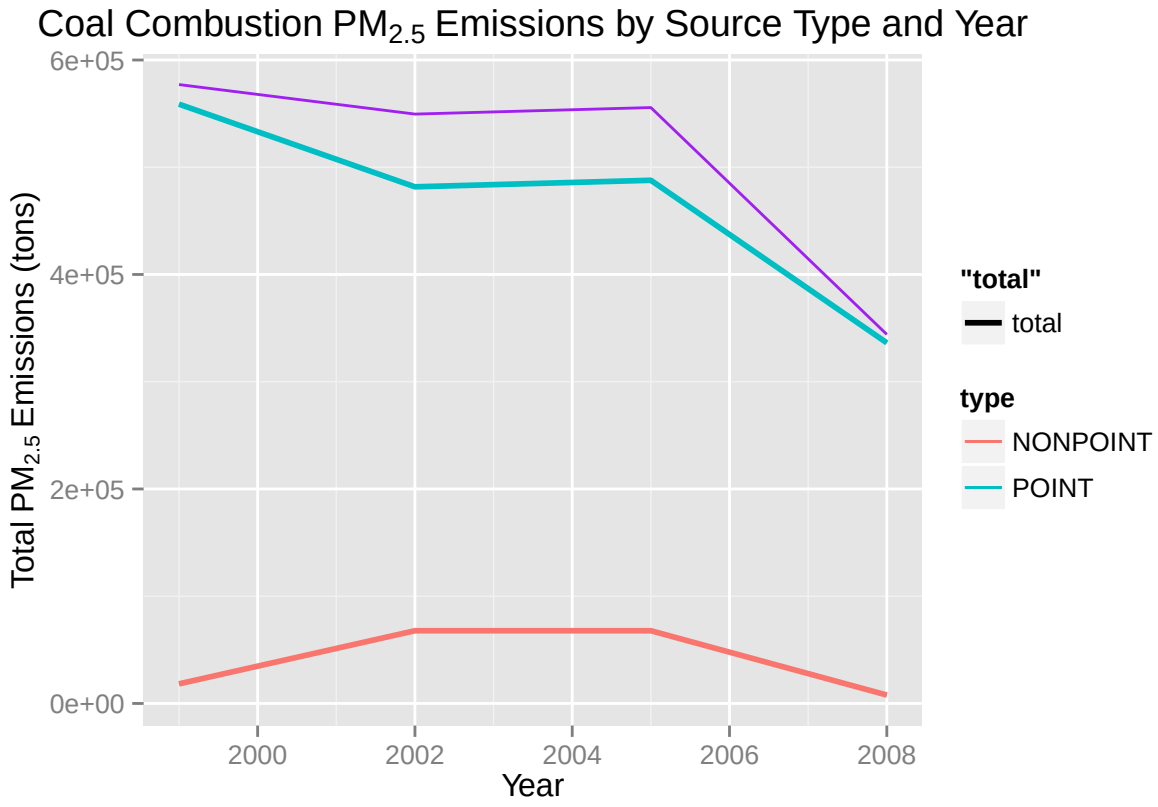
```
## [1] 62
```

```
##Based on other coursera courses and previous history...
###it's time to look at the union of these sets
coalcomb.codes <- union(coalcomb.scc$SCC, coalcomb.scc1$SCC)
length(coalcomb.codes) #91
```

```
## [1] 97
```

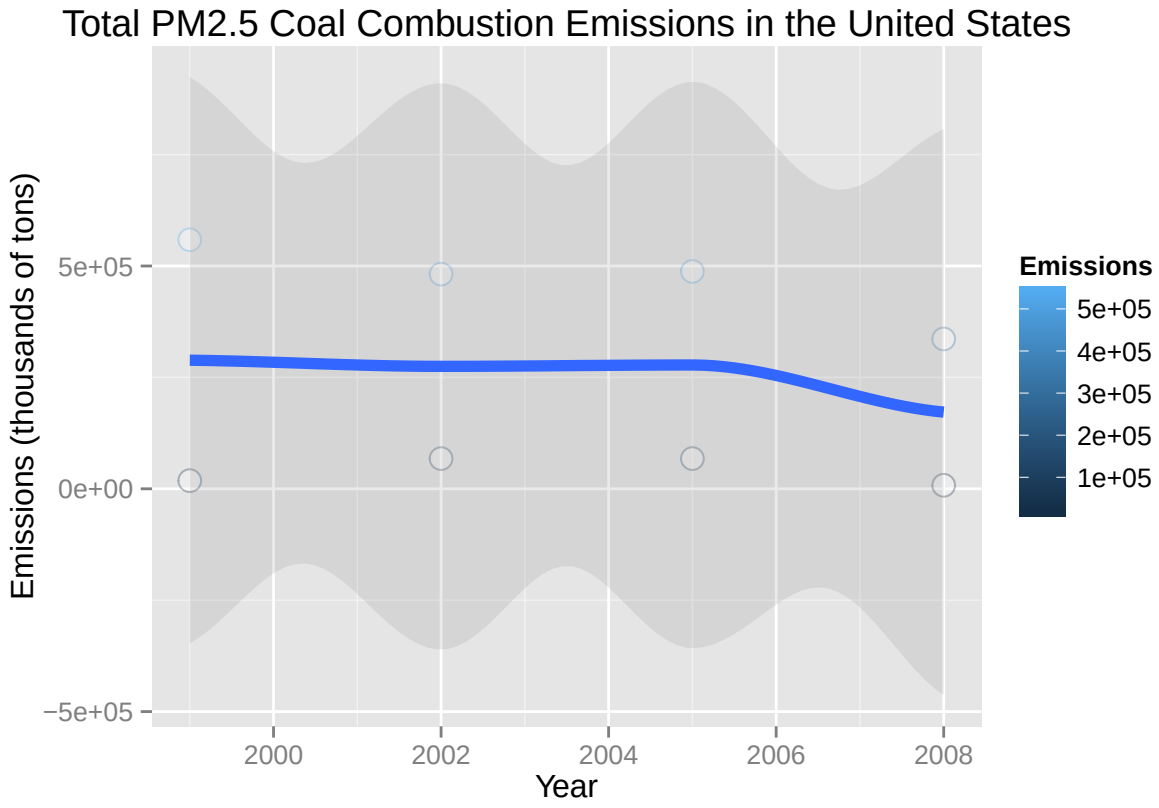
```
## Step 5: subset again for what we want
coal.comb <- subset(NEI, SCC %in% coalcomb.codes)

##Step 6: get the PM25 values as well
coalcomb.pm25year <- ddply(coal.comb, .(year, type), function(x) sum(x$Emissions))

#rename the col
colnames(coalcomb.pm25year)[3] <- "Emissions"
```

```
##Step 7: finally plot4.png prepare to plot to png
#png("plot4.png")
qplot(year, Emissions, data=coalcomb.pm25year, color=type, geom="line") + stat_summary(fun.y = "sum", fu
```



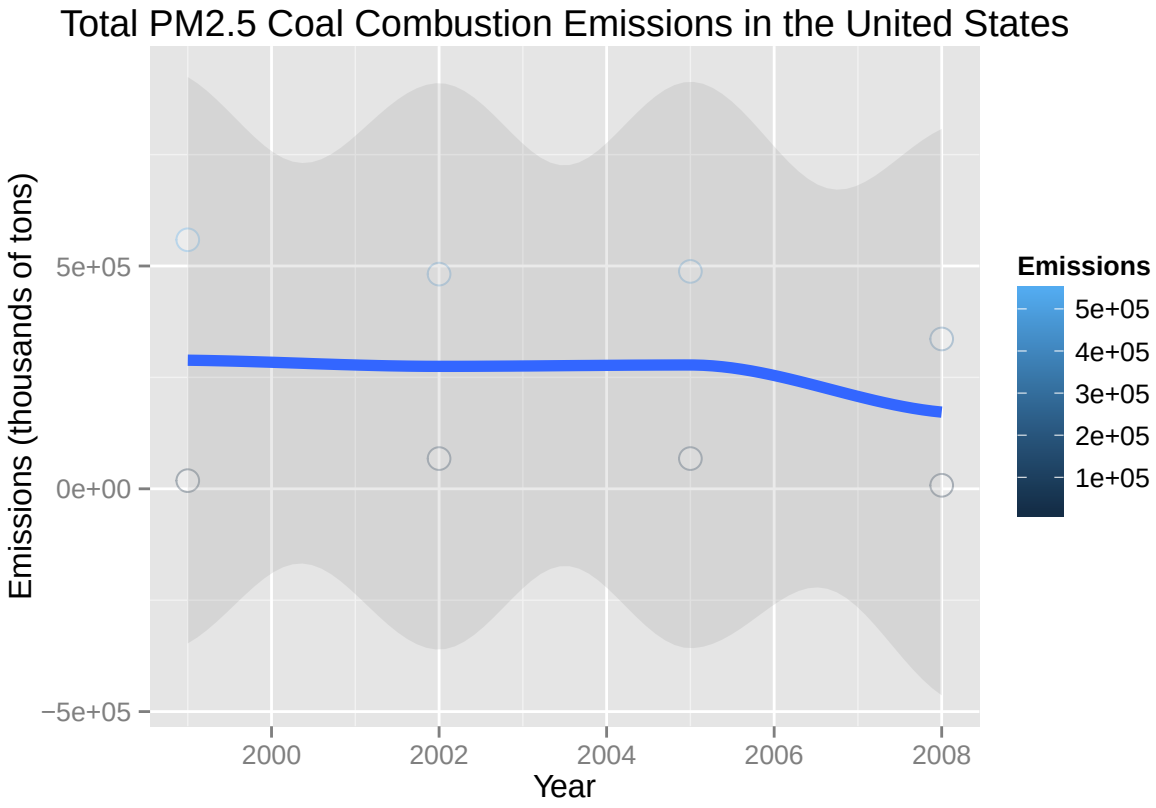

```
#dev.off()

##step nine: plot
#png("plot4a.png", width = 490, height = 490, units="px",pointsize = 12, bg = "white")
##Step 8: plot of combined coal values using smooth
plot.new <- ggplot(coalcomb.pm25year, aes(x = year, y = Emissions)) +
  geom_point(alpha = .3, aes(colour=Emissions), size=4, shape=21, fill="white") + xlab("Year") + ylab("
  geom_smooth(alpha = .2, size=2, pch=2, aes(colour=Emissions)) +
  ggtitle("Total PM2.5 Coal Combustion Emissions in the United States")

print(plot.new)
```

```
## geom_smooth: method="auto" and size of largest group is <1000, so using loess. Use 'method = x' to cl
```




```
print(plot.new)
```

```
## geom_smooth: method="auto" and size of largest group is <1000, so using loess. Use 'method = x' to cl
```




```
#close devise
#dev.off()
```

Answer:

- **Total (Purple Line):** From the plot, we see that the purple line for total slightly declines from 1999 to 2002. From 2002 to 2005 the line has a marginal increase. Finally, from 2005 to 2008, the overall trend has a sharp decrease.
- **Point (Blue Line):** From the plot, we see that the blue line for point is slightly similar in shape to the total purple line. From 1999 to 2002 the point blue line has a steeper decrease. From 2002 to 2005, the point blue line increases only slightly. Finally, from 2005 to 2008, the overall trend has a sharp decrease.
- **Nonpoint (Red Line):** This line is remarkably different from the other two lines. From 1999 to 2002 it has an increase (although it starts from a much lower level than the other two lines at just above zero). From 2002 to 2005 it remains nearly level and does not appear to increase or decrease much. Finally, from 2005 to 2008, another symmetrical decrease occurs to end just below the initial levels for the 1999 values.
- **LastPlot (Blue Line):** This line shows a slight steady trend from 1999 to around 2004. The line decreases more dramatically from 2004 to 2008.

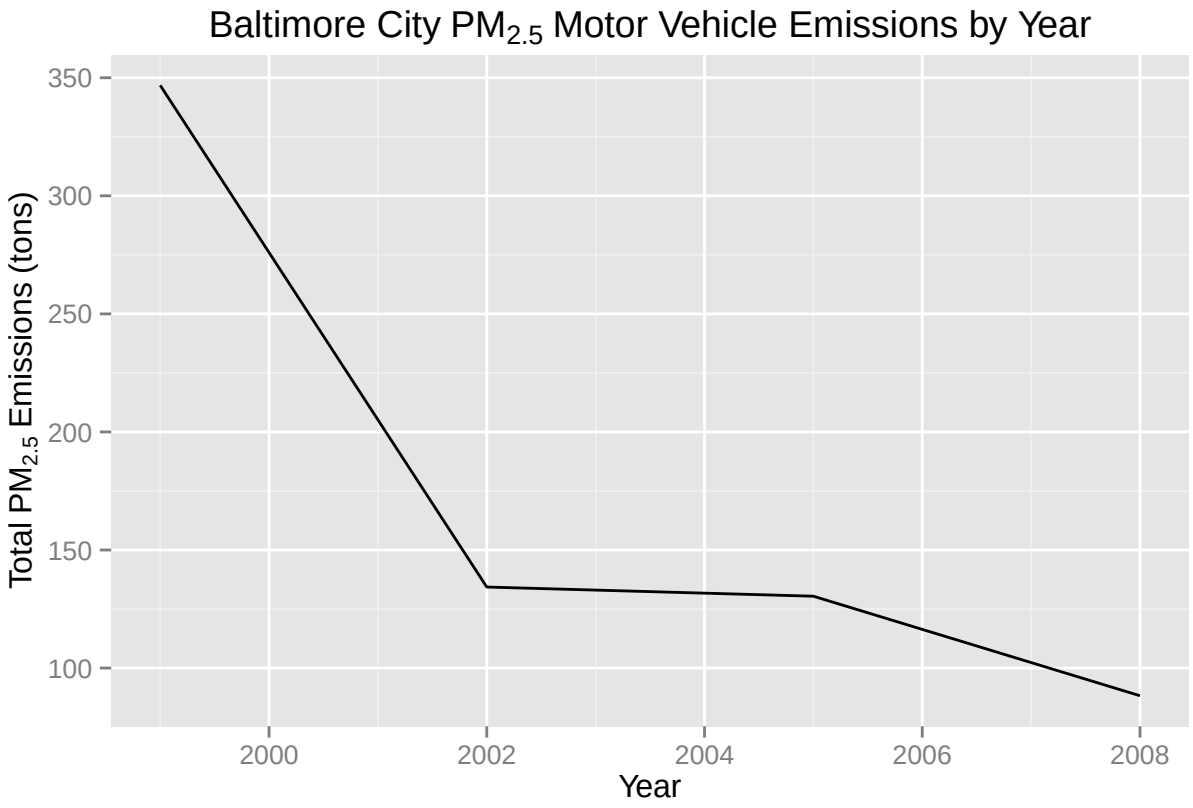
Question 5: plot5.r

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

```
## Script Name: plot5.R
## Version: 1.0_14

## Libraries needed:
library(plyr)
library(ggplot2)

## Step 1: read in the data
NEI <- readRDS("expdata_prj2/summarySCC_PM25.rds")
SCC <- readRDS("expdata_prj2/Source_Classification_Code.rds")

## Step 2A: subset our data
## Assumptions: motor vehicles = On and
###check the levels for types of vehicles defined
mv.sourced <- unique(grep("Vehicles", SCC$EI.Sector, ignore.case = TRUE, value = TRUE))

mv.sourcec <- SCC[SCC$EI.Sector %in% mv.sourced, ]["SCC"]

##Step 2B: subset the emissions from motor vehicles from
##NEI for Baltimore, MD.
emMV.ba <- NEI[NEI$SCC %in% mv.sourcec$SCC & NEI$fips == "24510",]

## Step 3: find the emissions due to motor vehicles in Baltimore for every year
balmv.pm25yr <- ddply(emMV.ba, .(year), function(x) sum(x$Emissions))
colnames(balmv.pm25yr)[2] <- "Emissions"

## Step 4: Plot to png
png("plot5.png")
qplot(year, Emissions, data=balmv.pm25yr, geom="line") + ggtitle(expression("Baltimore City" ~ PM[2.5] 
dev.off()
```

```
## pdf
##   2
```

```
## Step 5: Plot to markdown
qplot(year, Emissions, data=balmv.pm25yr, geom="line") + ggtitle(expression("Baltimore City" ~ PM[2.5] 
```



Answer:

Starting with 1999, the $PM_{2.5}$ emissions was just below 350, the levels fell sharply until 2002. From 2002 to 2005 the levels plateaued. Finally from 2005 to 2008, the $PM_{2.5}$ emissions drop to below 100 $PM_{2.5}$ emissions.

Question 6: plot6.r

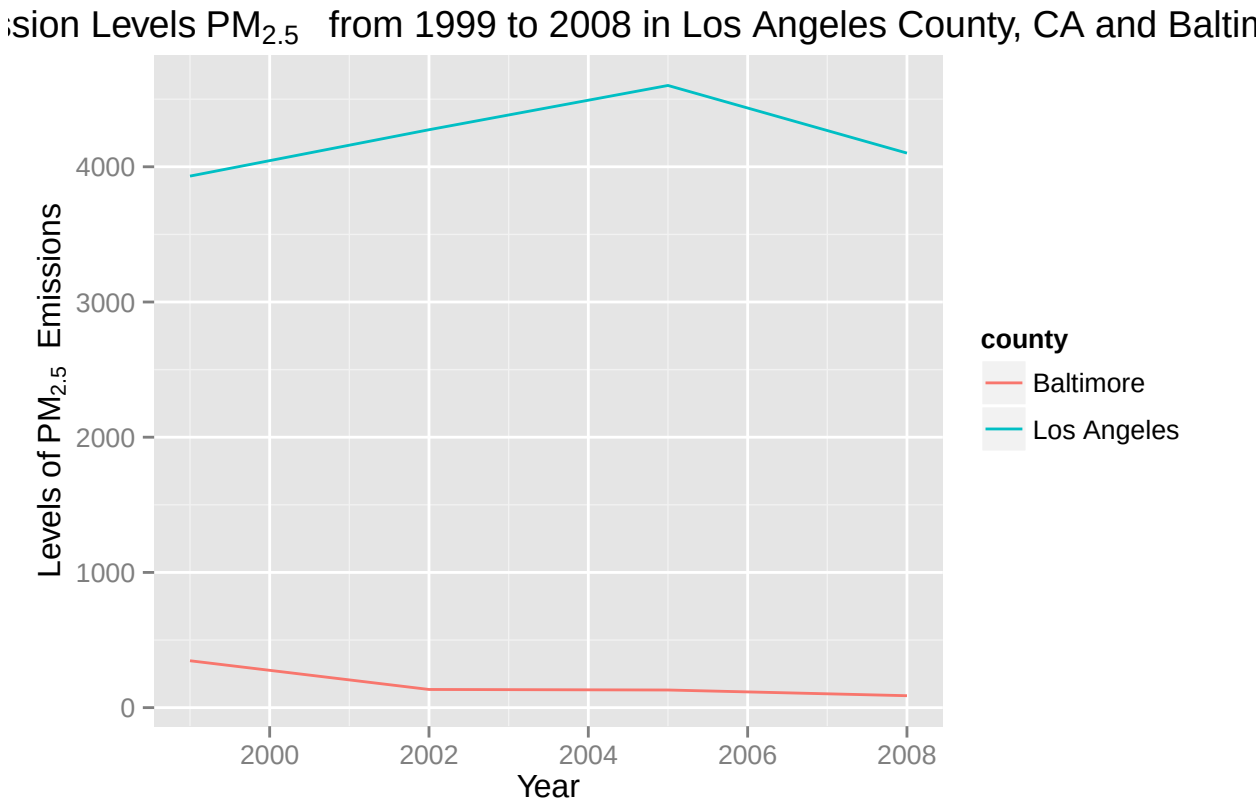
Compare emissions from motor vehicle sources in Baltimore City with emissions from motor vehicle sources in *Los Angeles County*, California, California (`fips == "06037"`). Which city has seen greater changes over time in motor vehicle emissions?

```
## Script Name: plot6.R
## Version: 1.0_14

## Libraries needed:
library(plyr)
library(ggplot2)
library(grid)

## Step 1: read in the data
NEI <- readRDS("expdata_prj2/summarySCC_PM25.rds")
SCC <- readRDS("expdata_prj2/Source_Classification_Code.rds")

## Step 2: check the levels for types of vehicles defined
mv.sourced <- unique(grep("Vehicles", SCC$EI.Sector, ignore.case = TRUE, value = TRUE))
```

```
mv.sourcec <- SCC[SCC$EI.Sector %in% mv.sourced, ]["SCC"]


## Step 3A: subset our data Baltimore City
emMV.ba <- NEI[NEI$SCC %in% mv.sourcec$SCC & NEI$fips == "24510", ]
## Step 3B: subset our data Los Angeles County
emMV.LA <- NEI[NEI$SCC %in% mv.sourcec$SCC & NEI$fips == "06037", ]

## Step 3C: bind the data created in steps 3A and 3B
emMV.comb <- rbind(emMV.ba, emMV.LA)

## Step 4: Find the emmissions due to motor vehicles in
## Baltimore (city) and Los Angeles County
tmveYR.county <- aggregate (Emissions ~ fips * year, data =emMV.comb, FUN = sum )
tmveYR.county$county <- ifelse(tmveYR.county$fips == "06037", "Los Angeles", "Baltimore")

## Step 5: plotting to png
png("plot6.png", width=750)
qplot(year, Emissions, data=tmveYR.county, geom="line", color=county) + ggtitle(expression("Motor Vehic
dev.off()
```

```
## pdf
##   2
```

```
##Plot to markdown
qplot(year, Emissions, data=tmveYR.county, geom="line", color=county) + ggtitle(expression("Motor Vehic
```



Answer:

Comparisons of $PM_{2.5}$ emissions between Baltimore, MD (a city in MD) and Los Angeles, CA (county). In this case, we are asked to compare a city to a county. In plot 6, we notice that Baltimore, MD starts considerably lower in terms of $PM_{2.5}$ emissions.

- **Baltimore, MD [city] (Red Line):** The red line starts marginally above zero and below 1,000 $PM_{2.5}$ emission values. Between 1999 and 2002, it slowly declines and remains nearly static between 2002 and 2008.
- **Los Angeles, CA [county] (Blue Line):** The blue line starts significantly higher than the Baltimore, MD line. Starting with 1999, slightly below 4,000 $PM_{2.5}$ emisions and steadily increases to 2005. The value of $PM_{2.5}$ emissions for 2005 hits a peak at approximately 4,500 $PM_{2.5}$ emision levels and then decreases between 2005 and 2008 with an ending value point of slightly above 4,000 $PM_{2.5}$ emissions.