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.

Questions

You must address the following questions and tasks in your exploratory analysis. For each question/task you will need to make a single plot. Unless specified, you can use any plotting system in R to make your plot.

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.
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.
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.
4. Across the United States, how have emissions from coal combustion-related sources changed from 1999–2008?
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 (fips == "06037"). Which city has seen greater changes over time in motor vehicle emissions?

Making and Submitting Plots

For each plot you should

- Construct the plot and save it to a PNG file.
- Create a separate R code file (plot1.R, plot2.R, etc.) that constructs the corresponding plot, i.e. code in plot1.R constructs the plot1.png plot. Your code file should include code for reading the data so that the plot can be fully reproduced. You should also include the code that creates the PNG file. Only include the code for a single plot (i.e. plot1.R should only include code for producing plot1.png)
- Upload the PNG file on the Assignment submission page
- Copy and paste the R code from the corresponding R file into the text box at the appropriate point in the peer assessment.

```
Have total emissions from PM2.5 decreased in the United States from 1999 to 2008? Using the base plottin
```

```
Upload a PNG file containing your plot addressing this question.
```

Solution

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

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

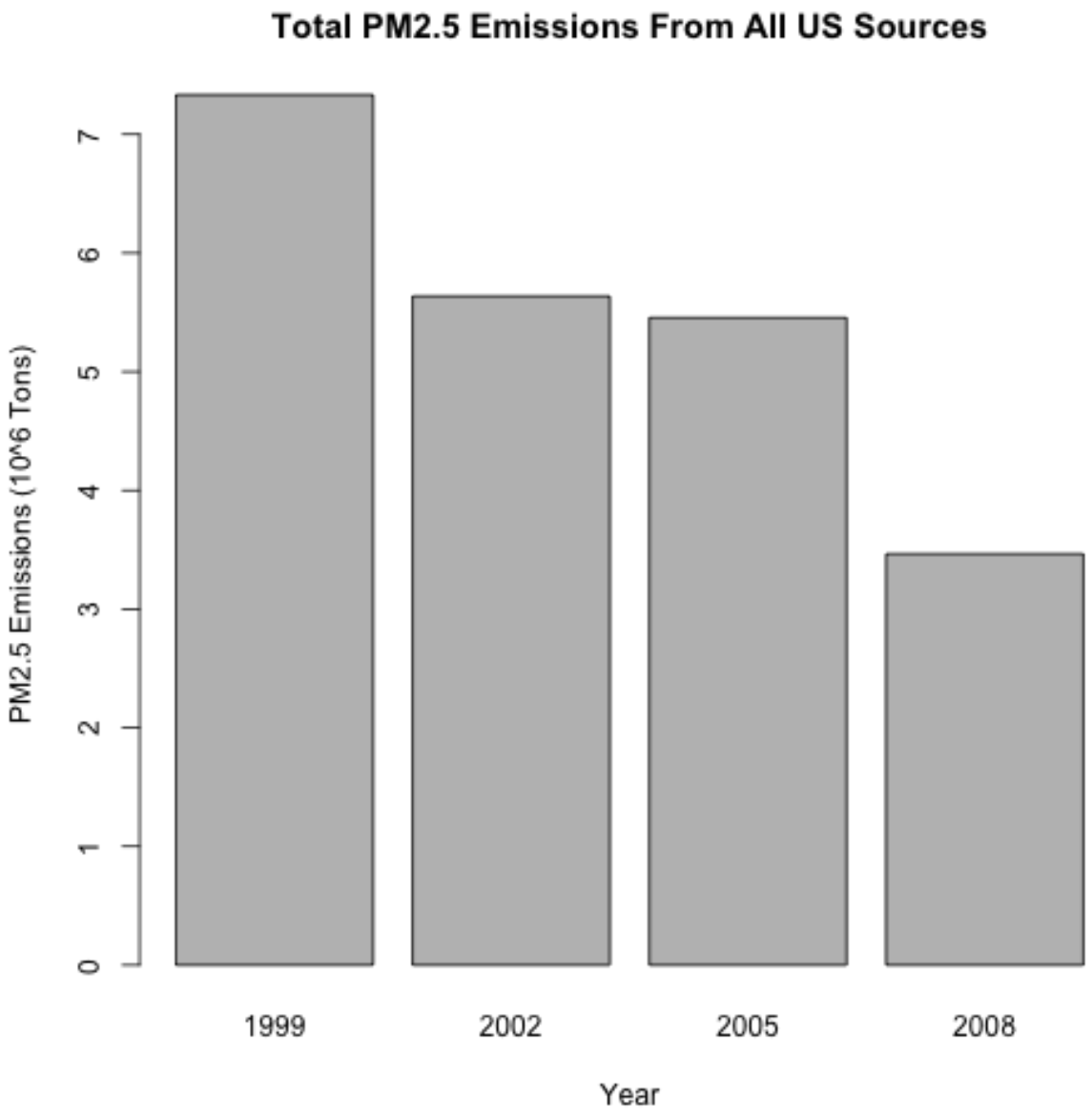
Question 1

First we'll aggregate the total PM2.5 emission from all sources for each of the years 1999, 2002, 2005, and 2008.

```
aggTotals <- aggregate(Emissions ~ year,NEI, sum)
```

Using the base plotting system, now we plot the total PM2.5 Emission from all sources,

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



Have total emissions from PM2.5 decreased in the United States from 1999 to 2008?

As we can see from the plot, total emissions have decreased in the US from 1999 to 2008.

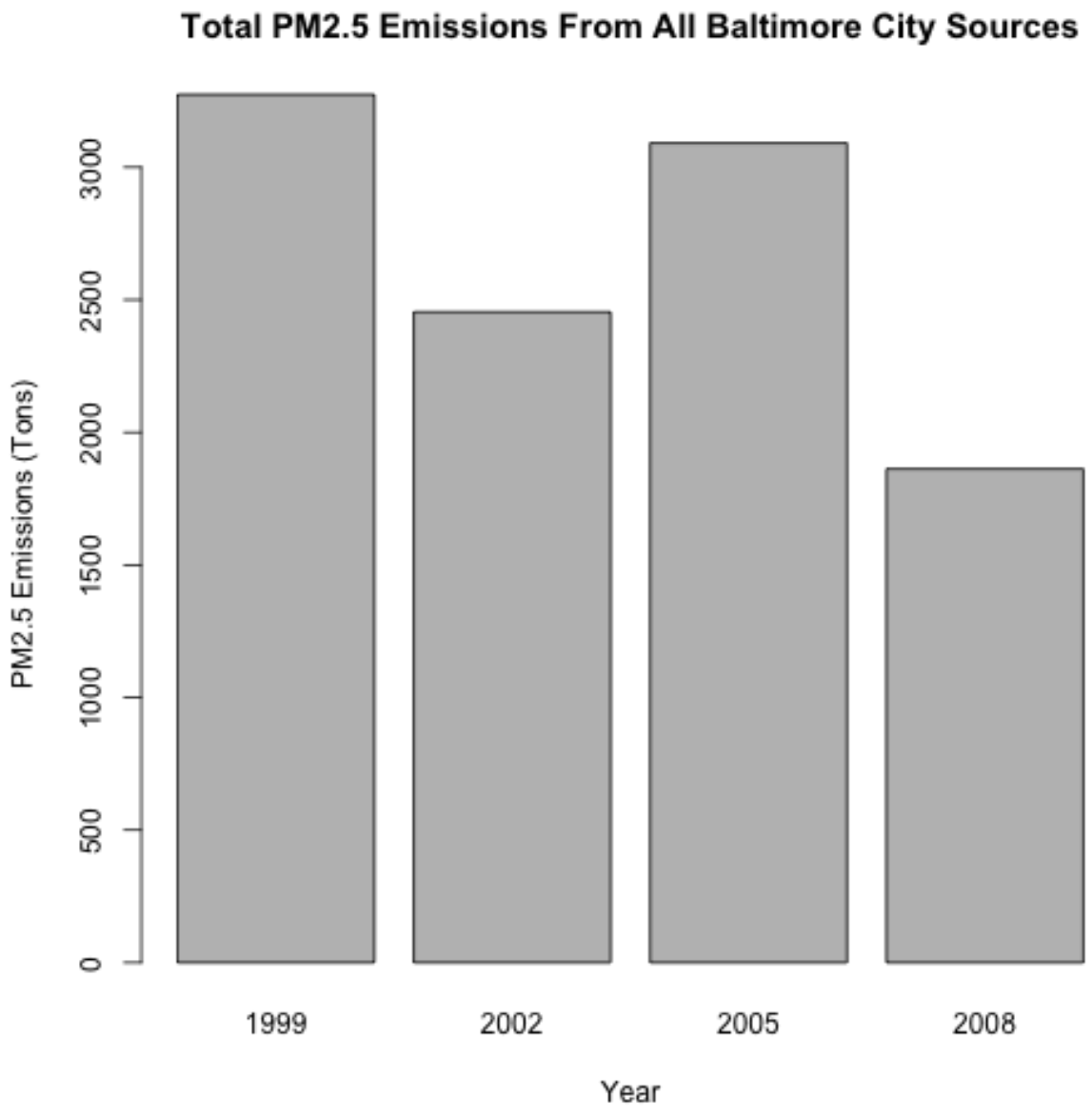
Question 2

First we aggregate total emissions from PM2.5 for Baltimore City, Maryland (fips="24510") from 1999 to 2008.

```
baltimoreNEI <- NEI[NEI$fips=="24510",]
aggTotalsBaltimore <- aggregate(Emissions ~ year, baltimoreNEI,sum)
```

Now we use the base plotting system to make a plot of this data,

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



Have total emissions from PM2.5 decreased in the Baltimore City, Maryland (fips == "24510") from 1999 to 2008?

Overall total emissions from PM2.5 have decreased in Baltimore City, Maryland from 1999 to 2008.

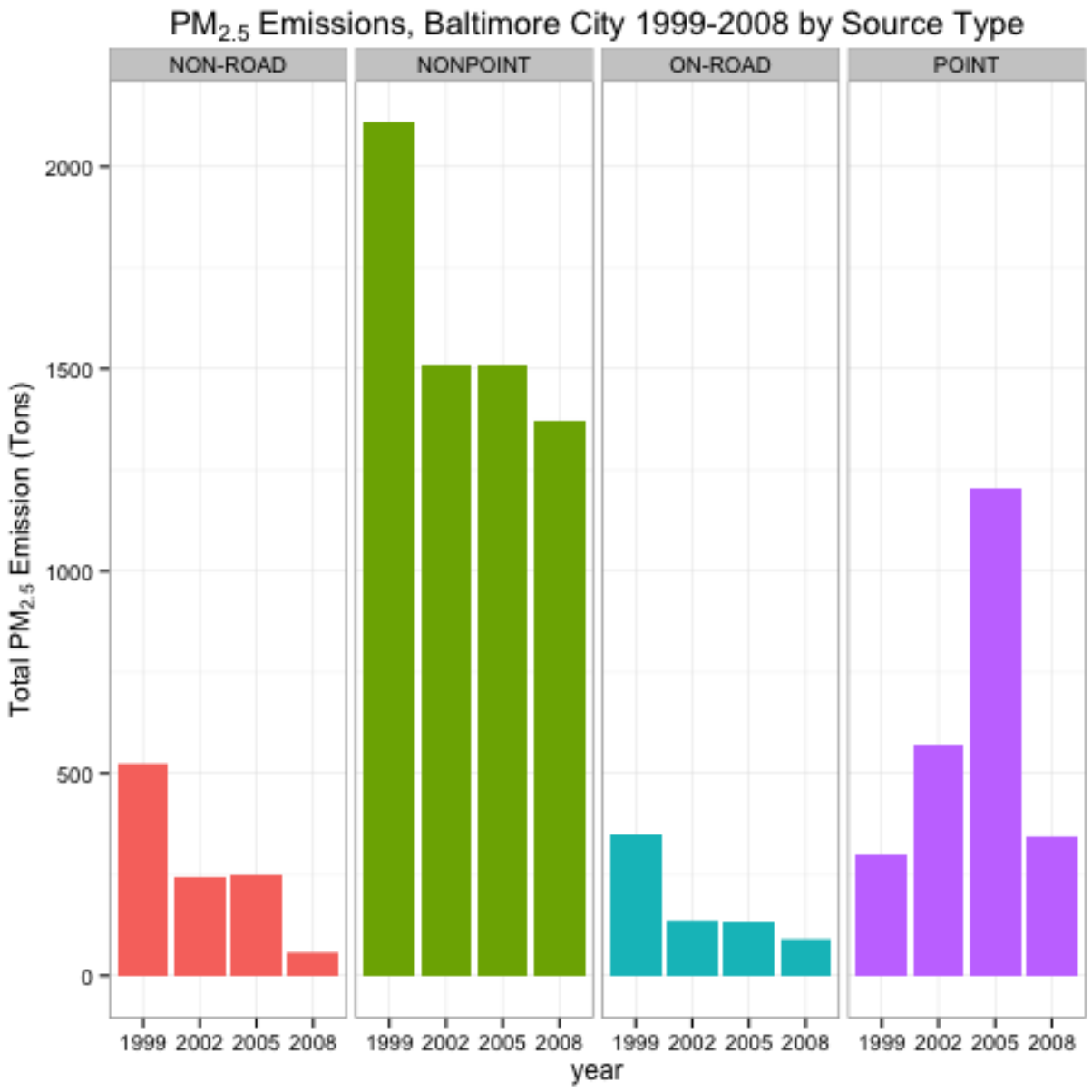
Question 3

Using the ggplot2 plotting system,

```
library(ggplot2)

ggp <- ggplot(baltimoreNEI,aes(factor(year),Emissions,fill=type)) +
  geom_bar(stat="identity") +
  theme_bw() + guides(fill=FALSE)+
  facet_grid(.~type,scales = "free",space="free") +
  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"))

print(ggp)
```



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?

The `non-road`, `nonpoint`, `on-road` source types have all seen decreased emissions overall from 1999-2008 in Baltimore City.

Which have seen increases in emissions from 1999–2008?

The `point` source saw a slight increase overall from 1999-2008. Also note that the `point` source saw a significant increase until 2005 at which point it decreases again by 2008 to just above the starting values.

(Note that I did not catch this originally as I started off with a log scale on Emissions)

Question 4

First we subset coal combustion source factors NEI data.

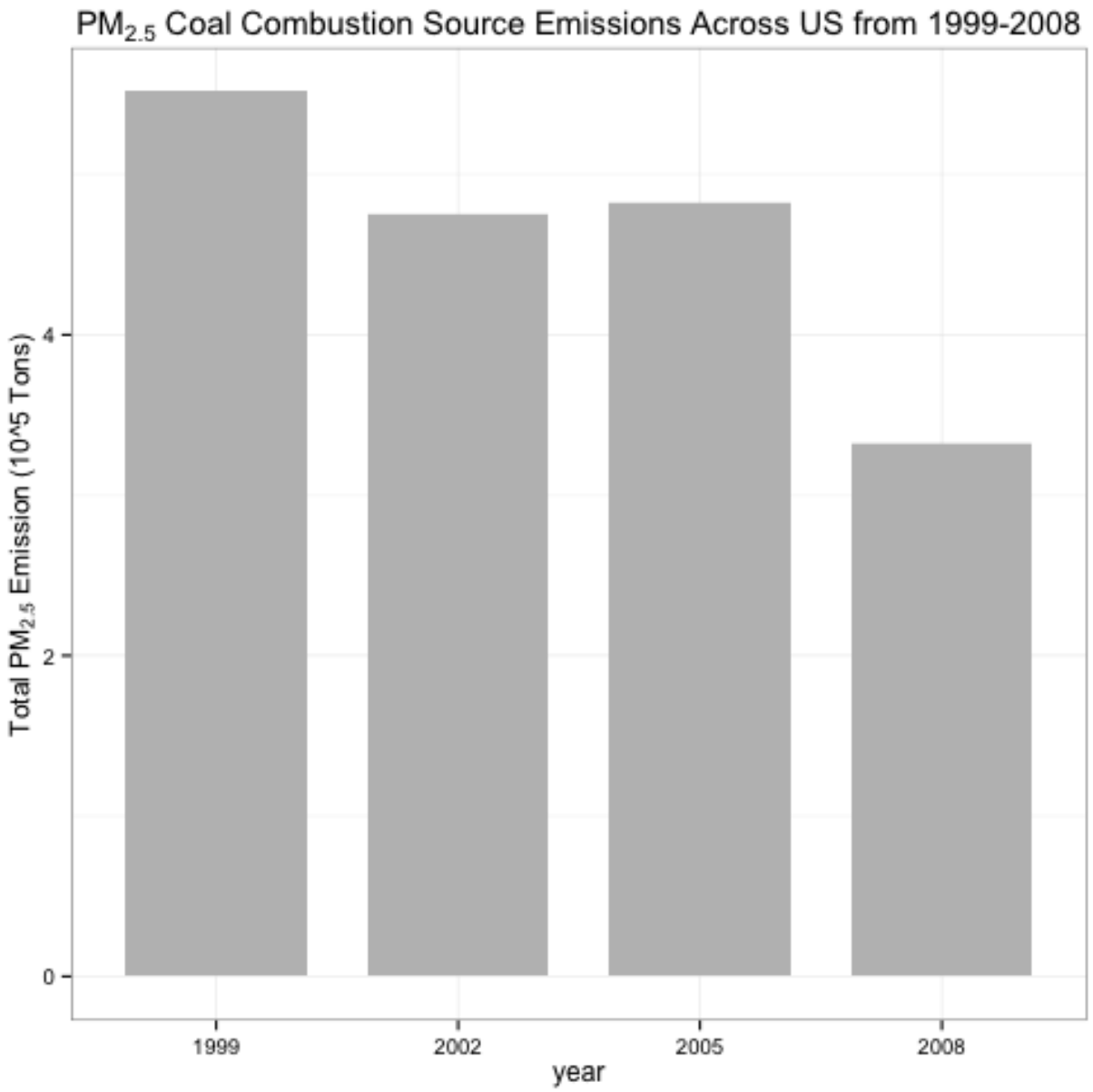
```
# Subset coal combustion related NEI data
combustionRelated <- grepl("comb", SCC$SCC.Level.One, ignore.case=TRUE)
coalRelated <- grepl("coal", SCC$SCC.Level.Four, ignore.case=TRUE)
coalCombustion <- (combustionRelated & coalRelated)
combustionSCC <- SCC[coalCombustion,]$SCC
combustionNEI <- NEI[NEI$SCC %in% combustionSCC,]
```

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'.

```
library(ggplot2)

ggp <- ggplot(combustionNEI,aes(factor(year),Emissions/10^5)) +
  geom_bar(stat="identity",fill="grey",width=0.75) +
  theme_bw() +  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"))

print(ggp)
```



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

Emissions from coal combustion related sources have decreased from 6 * 10^6 to below 4 * 10^6 from 1999-2008.

Eg. Emissions from coal combustion related sources have decreased by about 1/3 from 1999-2008!

Question 5

First we subset the motor vehicles, which we assume is anything like Motor Vehicle in SCC.Level.Two.

```
vehicles <- grepl("vehicle", SCC$SCC.Level.Two, ignore.case=TRUE)
vehiclesSCC <- SCC[vehicles,]$SCC
vehiclesNEI <- NEI[NEI$SCC %in% vehiclesSCC,]
```

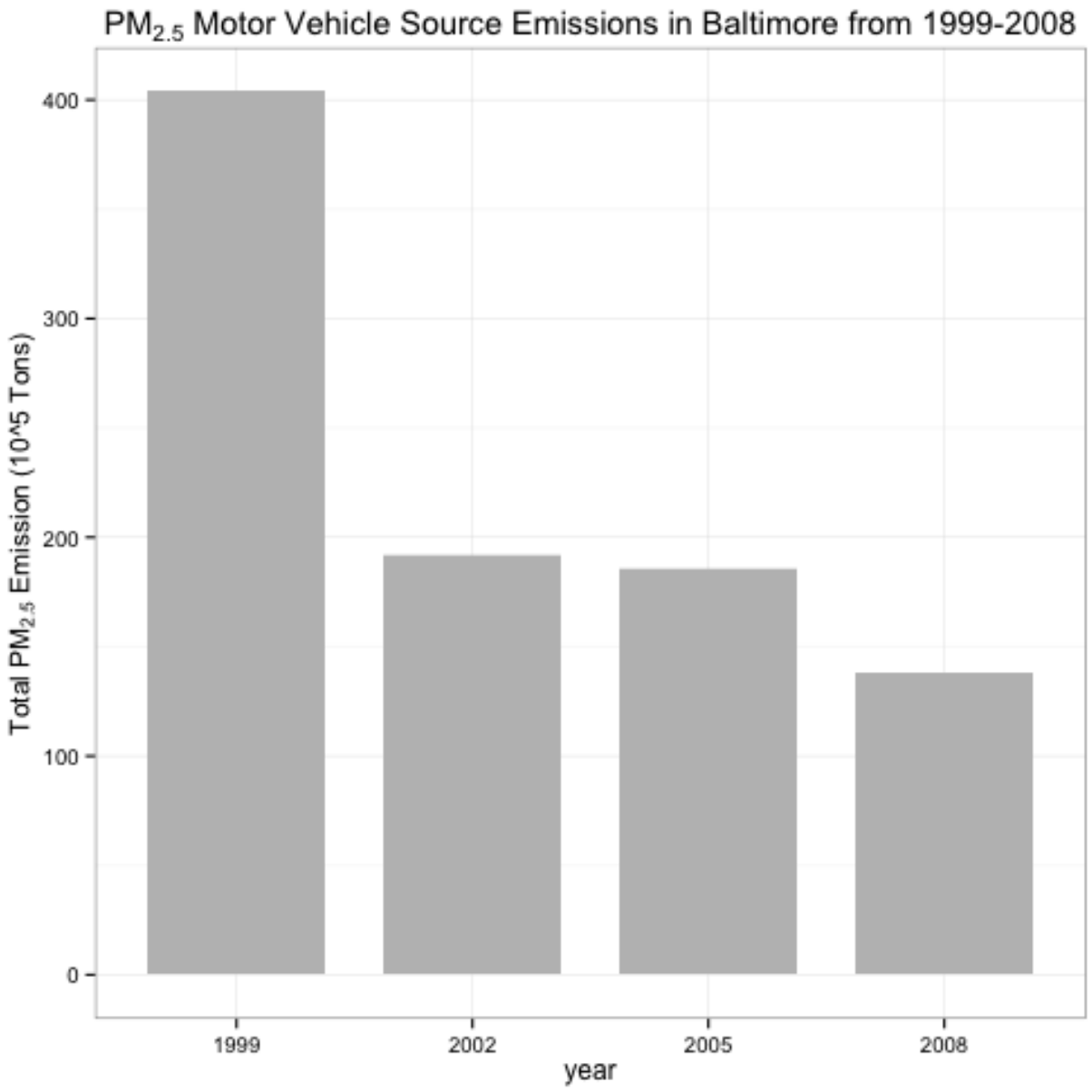
Next we subset for motor vehicles in Baltimore,

```
baltimoreVehiclesNEI <- vehiclesNEI[vehiclesNEI$fips==24510,]
```

Finally we plot using ggplot2,

```
library(ggplot2)
ggp <- ggplot(baltimoreVehiclesNEI,aes(factor(year),Emissions)) +
  geom_bar(stat="identity",fill="grey",width=0.75) +
  theme_bw() +  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"))

print(ggp)
```



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

Emissions from motor vehicle sources have dropped from 1999-2008 in Baltimore City!

Question 6

Comparing emissions from motor vehicle sources in Baltimore City (fips == "24510") with emissions from motor vehicle sources in Los Angeles County, California (fips == "06037"),

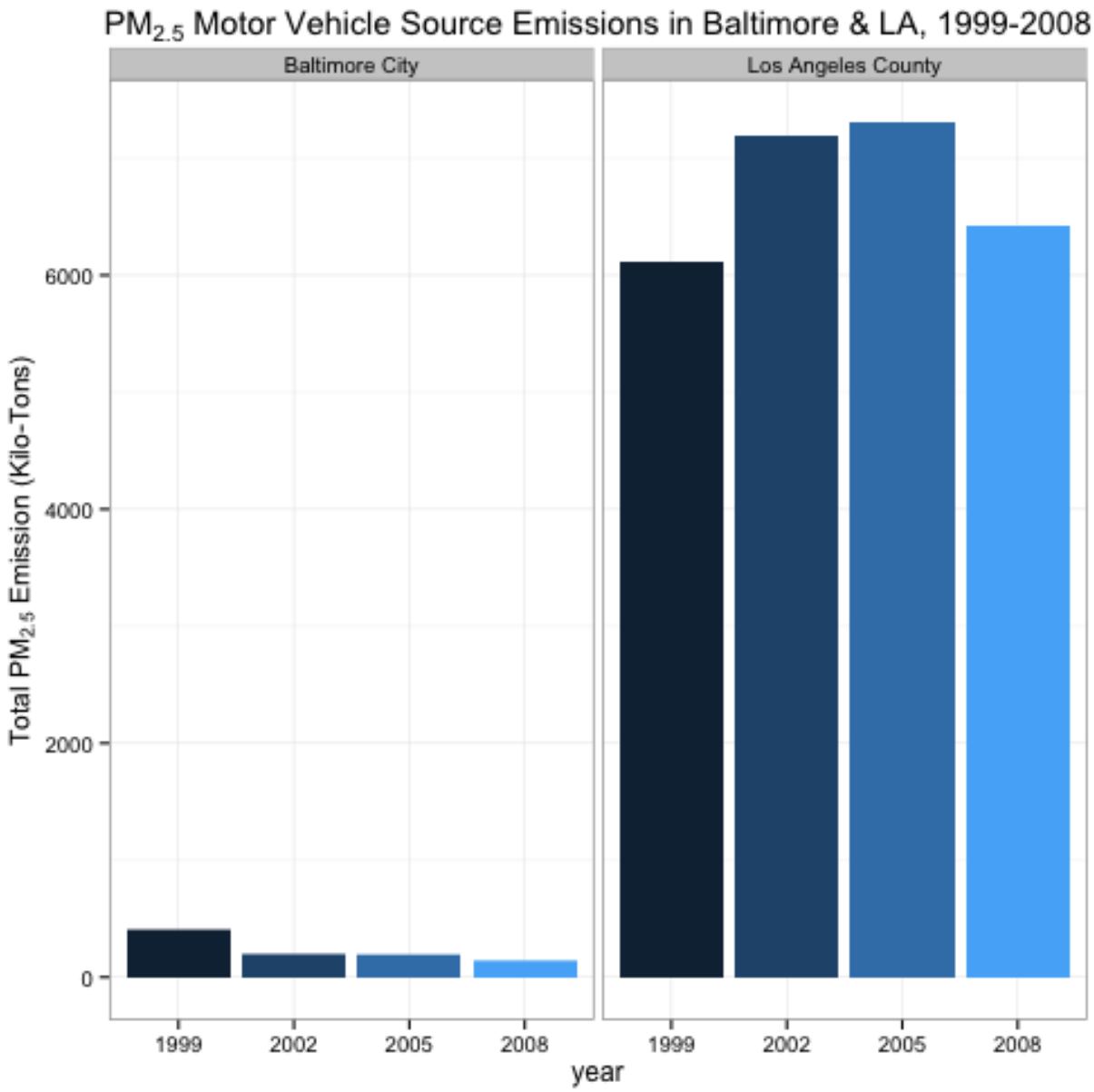
```
vehiclesBaltimoreNEI <- vehiclesNEI[vehiclesNEI$fips == 24510,]
vehiclesBaltimoreNEI$city <- "Baltimore City"
vehiclesLANEI <- vehiclesNEI[vehiclesNEI$fips=="06037",]
vehiclesLANEI$city <- "Los Angeles County"
bothNEI <- rbind(vehiclesBaltimoreNEI,vehiclesLANEI)
```

Now we plot using the ggplot2 system,

```
library(ggplot2)

ggp <- ggplot(bothNEI, aes(x=factor(year), y=Emissions, fill=city)) +
 geom_bar(aes(fill=year),stat="identity") +
 facet_grid(scales="free", space="free", .~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"))

print(ggp)
```



Which city has seen greater changes over time in motor vehicle emissions?

Los Angeles County has seen the greatest changes over time in motor vehicle emissions.