Pollution Analysis

Matthew Dunne

Instructions

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.

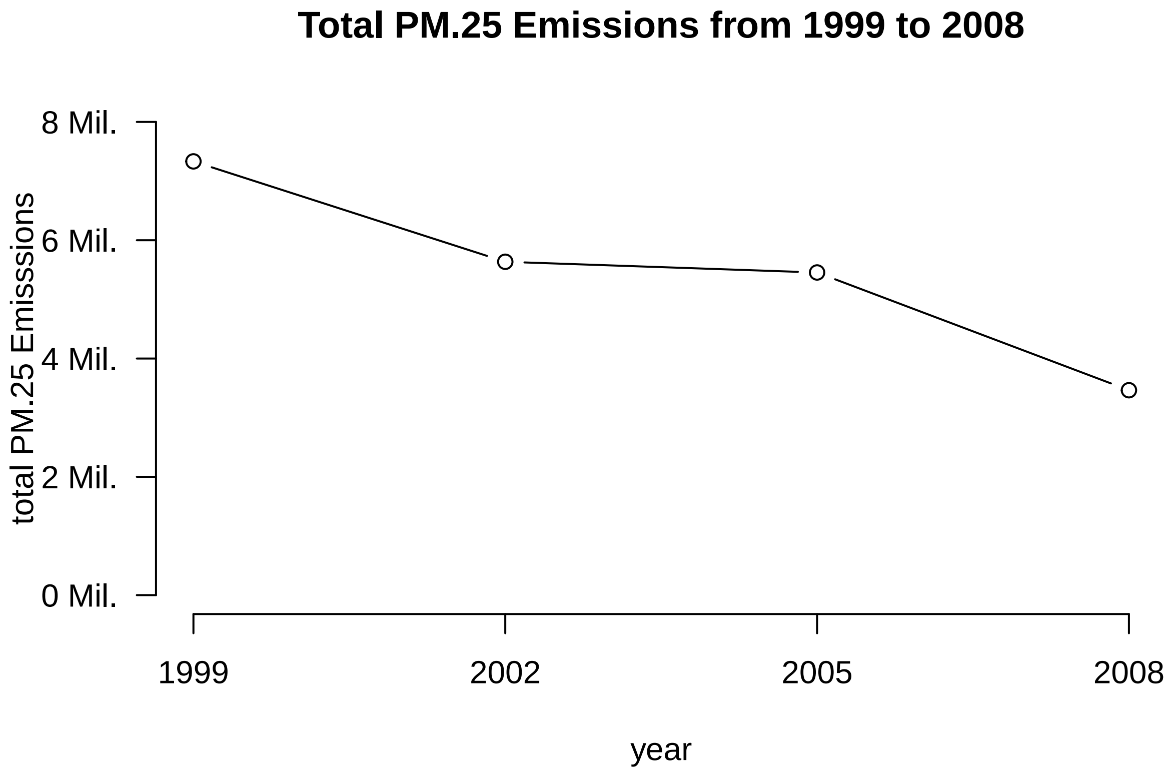
For each year and for each type of PM source, the NEI records how many tons of PM2.5 were emitted from that source over the course of the entire year. The data that you will use for this assignment are for 1999, 2002, 2005, and 2008.

```
setwd("C:/Users/mjdun/Desktop/Coursera/Data Science Specialization/Course 4 Exploratory Data Analysis/C
## read in emissions data
NEI<-readRDS("summarySCC_PM25.rds")
## read in source class code table
SCC<-readRDS("Source_Classification_Code.rds")
```

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.

```
##total emissions from all sources in each year. Returns number vector with years as attributes
one<-with(NEI, tapply(Emissions, year, sum, na.rm=TRUE))
## read into a dataframe you can use to plot
data1<-data.frame(year=names(one), total=one)
##convert total column to a numeric vector (don't know how much it helps)
data1$total<-as.numeric(data1$total)
##convert year from factor into character vector
data1$year<-as.character(data1$year)
## open graphics file
#png(file="plot1.png")
##put in plot. Divide total by 1M so the scale works
with(data1, plot(year, total/1000000, pch=1, main="Total PM.25 Emissions from 1999 to 2008", type="b", 
##add/change x axis
axis(1, at=c(1999, 2002, 2005, 2008), labels = c(1999, 2002, 2005, 2008))
## change y axis to add "Mil." for million, make labels horizontal
axis(2, at=axTicks(2), labels=paste(axTicks(2), "Mil.", sep = " "), las=1)
```



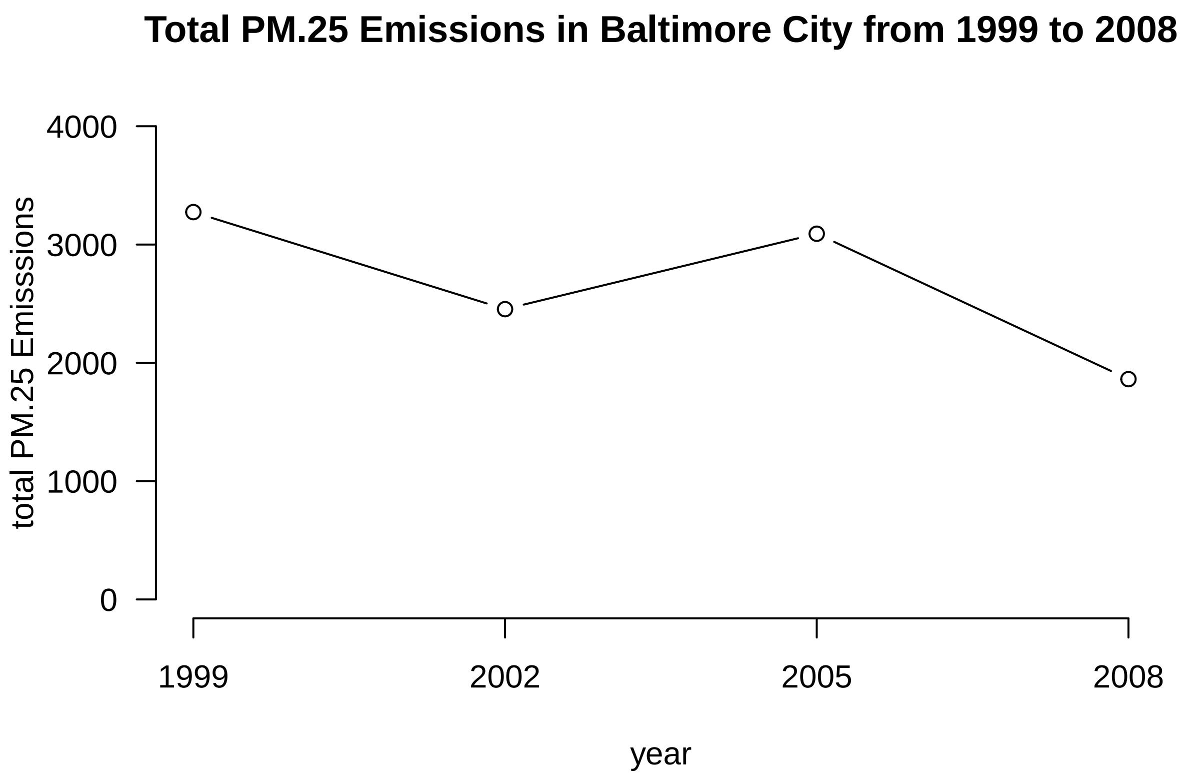
```
##close graphics device
#dev.off()
```

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 to Baltimore City data
balt1<-subset(NEI, fips=="24510")
##total emissions from all sources in each year. Returns number vector with years as attributes
two<-with(balt1, tapply(Emissions, year, sum, na.rm=TRUE))
## read into a dataframe you can use to plot
data<-data.frame(year=names(two), total=two)
##convert total column to a numeric vector (don't know how much it helps)
data$total<-as.numeric(data$total)
##convert year from factor into character vector
data$year<-as.character(data$year)
## open graphics file
#png(file="plot2.png")
##put in plot. Set y limit based on max, min values in total
with(data, plot(year, total, pch=1, main="Total PM.25 Emissions in Baltimore City from 1999 to 2008", ty
##add/change x axis
axis(1, at=c(1999, 2002, 2005, 2008), labels = c(1999, 2002, 2005, 2008))
##add/change y axis , make labels horizontal
```

```
axis(2, at=axTicks(2), las=1)
```



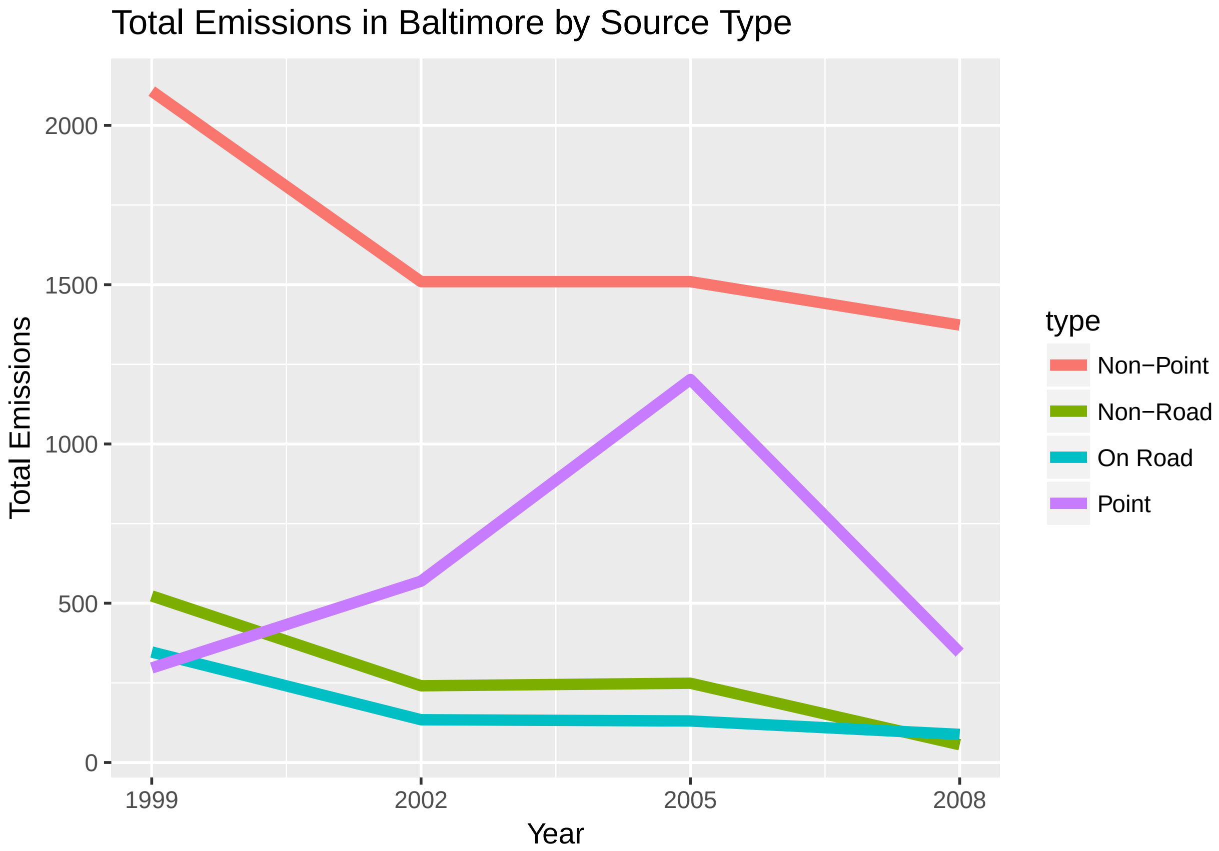
```
##close graphics device
#dev.off()
```

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.

```
##subset to Baltimore City data
balt<-subset(NEI, fips=="24510")
##run tapply indexing by year and type on the sum of emissions
three<-with(balt, tapply(Emissions, list(year, type), sum, na.rm=TRUE))
## convert to data frame
data3<-data.frame(three)
## create tranpose of data so years are columns (easier to work with in rearranging)
data3<-t(data3)
##create separate data frames for each year
df1<-data.frame(Total=data3[,1], type=c("Non-Road", "Non-Point", "On Road", "Point"), year=1999)
df2<-data.frame(Total=data3[,2], type=c("Non-Road", "Non-Point", "On Road", "Point"), year=2002)
df3<-data.frame(Total=data3[,3], type=c("Non-Road", "Non-Point", "On Road", "Point"), year=2005)
df4<-data.frame(Total=data3[,4], type=c("Non-Road", "Non-Point", "On Road", "Point"), year=2008)
##merge into one data frame
```

```
data<-rbind(df1, df2, df3, df4)
row.names(data)<-NULL
## open graphics file
#png(file="plot3.png")
library(ggplot2)
## set up plot with year on x-axis, Total emissions on y-axis
g<-ggplot(data, aes(year, Total))
##line plot differentiated by type. X-axis modified. Re-labelled
g+geom_line(aes(color=type), size=2)+scale_x_continuous(breaks = c(1999, 2002, 2005, 2008), label=c("19
```



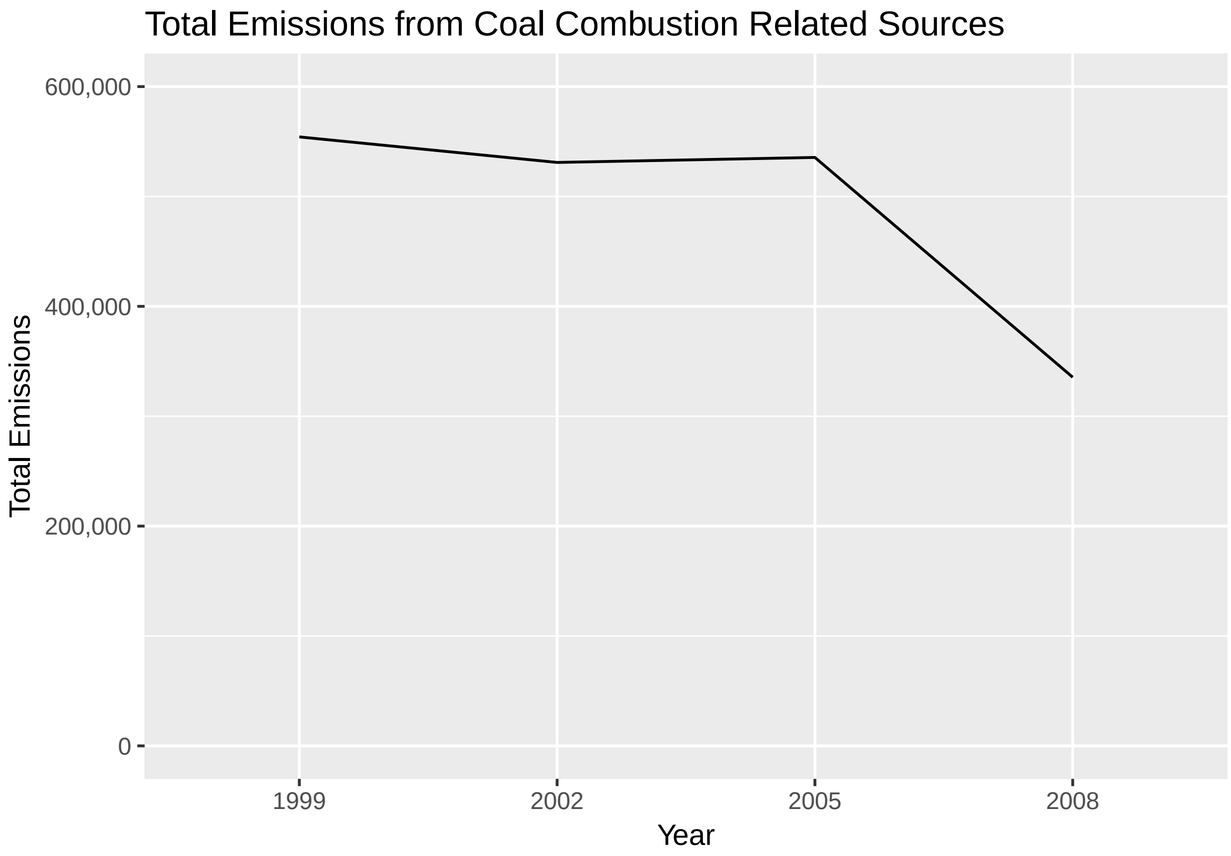
```
##close graphics device
#dev.off()
```

Question 4

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

```
#Short.Name is Coal combo of use and source
#EI.Sector is Coal what for
#Level.Three is Coal where from
## select the two relevant columns
data<-SCC[ ,c(1,9)]
##subsets to rows where "Coal" in SCC.Level.Three column
data1<-subset(data, data$SCC.Level.Three %in% grep("Coal", data$SCC.Level.Three, value=TRUE))
## but "Coal Mining" is not
```

```
data2<-subset(data1, !(data1$SCC.Level.Three %in% grep("Coal Mining", data1$SCC.Level.Three, value=TRUE
## and "Methane" is not either
data3<-subset(data2, !(data2$SCC.Level.Three %in% grep("Methane", data2$SCC.Level.Three, value=TRUE)))
##subset NEI on the SCC values in data3 (which is best estimate of coal combustion related sources)
data<-subset(NEI, NEI$SCC %in% data3$SCC)
##get total Emissions by year from the subsetted data
four<-with(data, tapply(Emissions, year, sum, na.rm=TRUE))
## convert years from a dimension (name of columns) to a column in a data frame
data<-data.frame(year=names(four), total=four)
## lose the dimension names which are now row names
row.names(data)<-NULL
##open graphics device and load ggplot2
#png(file="plot4.png")
library(ggplot2)
## plot with year on x axis, total on y axis. Connect all points (group=1)
g<-ggplot(data, aes(year, total, group=1))
## plot as a line. Set limits of y-axis with set tickmarks and labels. Add labels
g+geom_line()+scale_y_continuous(limits=c(0, 600000), breaks = c(0, 200000, 400000, 600000), labels = c
```

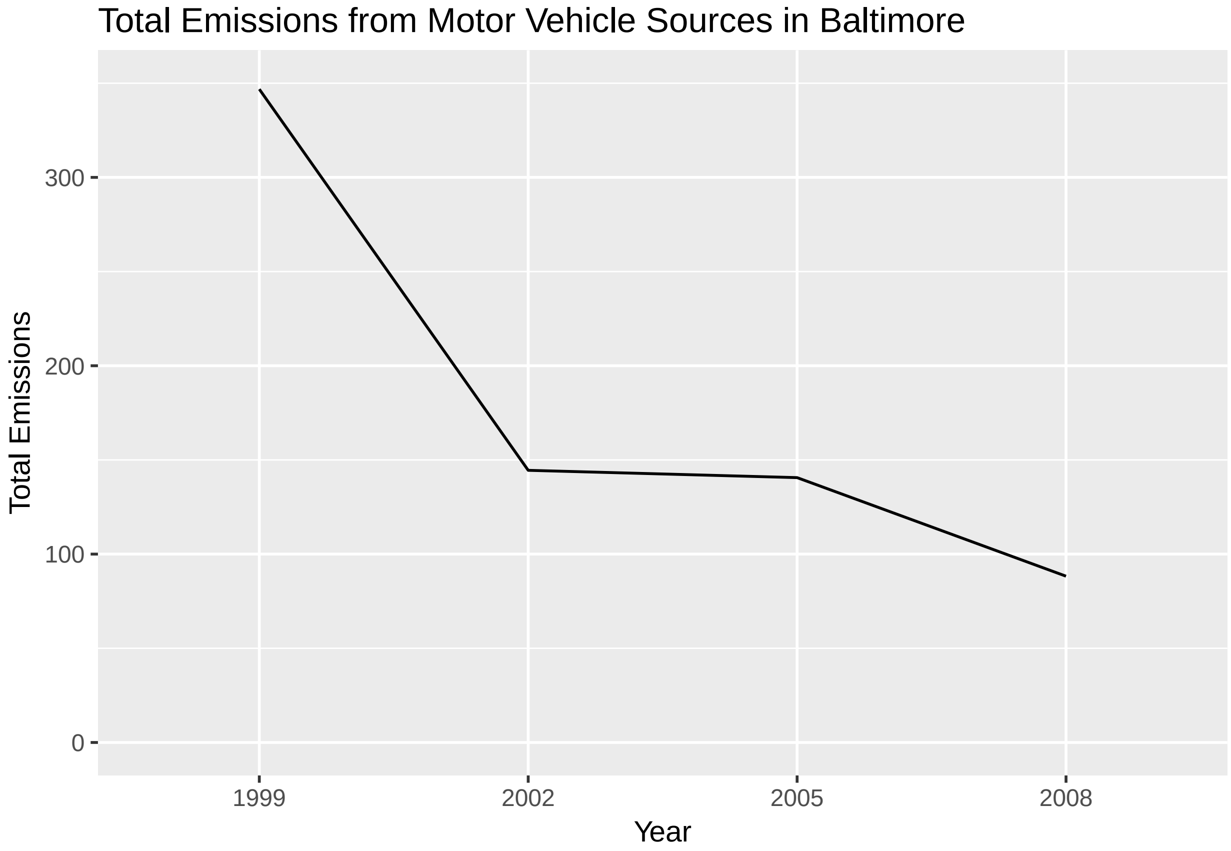


```
#dev.off()
```

Question 5

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

```
## select the two relevant columns
data<-SCC[ ,c(1,3)]
##subset SCC to vehicles
data1<-subset(data, data$Short.Name %in% grep("Veh", data$Short.Name, value=TRUE))
##subset to Baltimore City data
balt<-subset(NEI, fips=="24510")
##subset to where there is a match of SCC (motor vehicle sources from Baltimore)
data<-subset(balt, balt$SCC %in% data1$SCC)
##get total Emissions by year from the subsetted data
five<-with(data, tapply(Emissions, year, sum, na.rm=TRUE))
## convert years from a dimension (name of columns) to a column in a data frame
data<-data.frame(year=names(five), total=five)
## lose the dimension names which are now row names
row.names(data)<-NULL
##open graphics device and load ggplot2
#png(file="plot5.png")
library(ggplot2)
## plot with year on x axis, total on y axis. Connect all points (group=1)
g<-ggplot(data, aes(year, total, group=1))
## plot as a line. Set limits of y-axis with set tickmarks and labels. Add labels
g+geom_line()+ylim(0, 350) +labs(title="Total Emissions from Motor Vehicle Sources in Baltimore", x="Yea
```

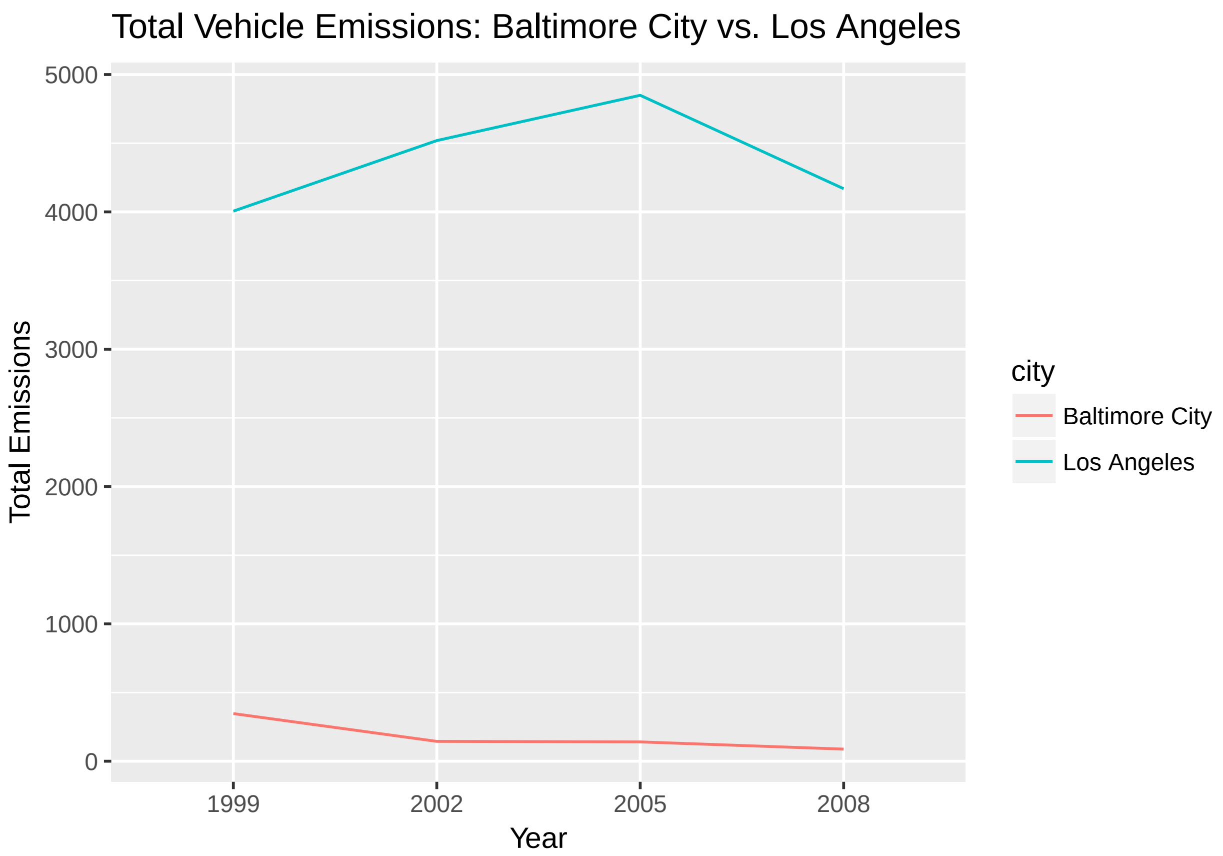


```
#dev.off()
```

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?

```r
## select the two relevant columns
data<-SCC[ ,c(1,3)]
##subset SCC to vehicles
data1<-subset(data, data$Short.Name %in% grep("Veh", data$Short.Name, value=TRUE))
##subset to Baltimore City data
balt<-subset(NEI, fips=="24510")
## subset to Los Angeles data
la<-subset(NEI, fips=="06037")
## subset Baltimore and Los Angeles data to motor vehicle sources
balt<-subset(balt, balt$SCC %in% data1$SCC)
la<-subset(la, la$SCC %in% data1$SCC)
## get total emissions in Baltimore by year and put info into useable data frame
balt<-with(balt, tapply(Emissions, year, sum, na.rm=TRUE))
balt<-data.frame(year=names(balt), total=balt, city="Baltimore City")
## get total emissions in Baltimore by year and put info into useable data frame
la<-with(la, tapply(Emissions, year, sum, na.rm=TRUE))
la<-data.frame(year=names(la), total=la, city="Los Angeles")
##open graphics device and load ggplot2
#png(file="plot6.png")
library(ggplot2)
## instead of binding two data frames together plot each one separately
ggplot()+geom_line(data=balt, aes(year, total, group=1, color=city))+geom_line(data=la, aes(year, total
```


Total Vehicle Emissions: Baltimore City vs. Los Angeles
Total Emissions
5000
4000
3000
2000
1000
0
1999
2002
2005
2008
Year
city
Baltimore City
Los Angeles

```
#dev.off()
```