Course Project 2

Student: Dipanjan Sarkar
Last saved: 5/22/2014 1:55:30 PM

Introduction

Fine particulate matter ($PM_{2.5}$) 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 $PM_{2.5}$. 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 (http://www.epa.gov/ttn/chief/eiinformation.html).

For each year and for each type of PM source, the NEI records how many tons of $PM_{2.5}$ 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 (https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip) [29Mb]

The zip file contains two files:

$PM_{2.5}$ Emissions Data (`summarySCC_PM25.rds`): This file contains a data frame with all of the $PM_{2.5}$ emissions data for 1999, 2002, 2005, and 2008. For each year, the table contains number of **tons** of $PM_{2.5}$ 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 $PM_{2.5}$ 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 $PM_{2.5}$ 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 $PM_{2.5}$ decreased in the United States from 1999 to 2008? Using the **base** plotting system, make a plot showing the *total* $PM_{2.5}$ emission 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 1999–2008 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

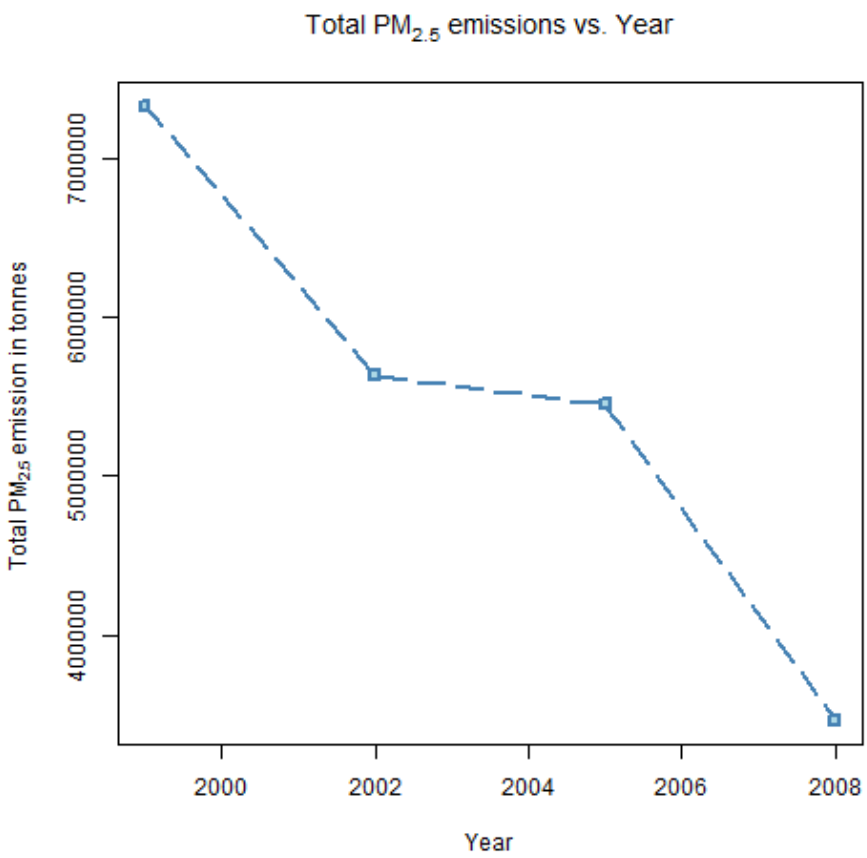
For each plot you should

- 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 $PM_{2.5}$ decreased in the United States from 1999 to 2008? Using the **base** plotting system, make a plot showing the *total* $PM_{2.5}$ emission from all sources for each of the years 1999, 2002, 2005, and 2008.
Upload a PNG file containing your plot addressing this question.



From the above plot, we see that the total $PM_{2.5}$ emissions show a downward trend as the years progress, making it clear that the total emissions have decreased.

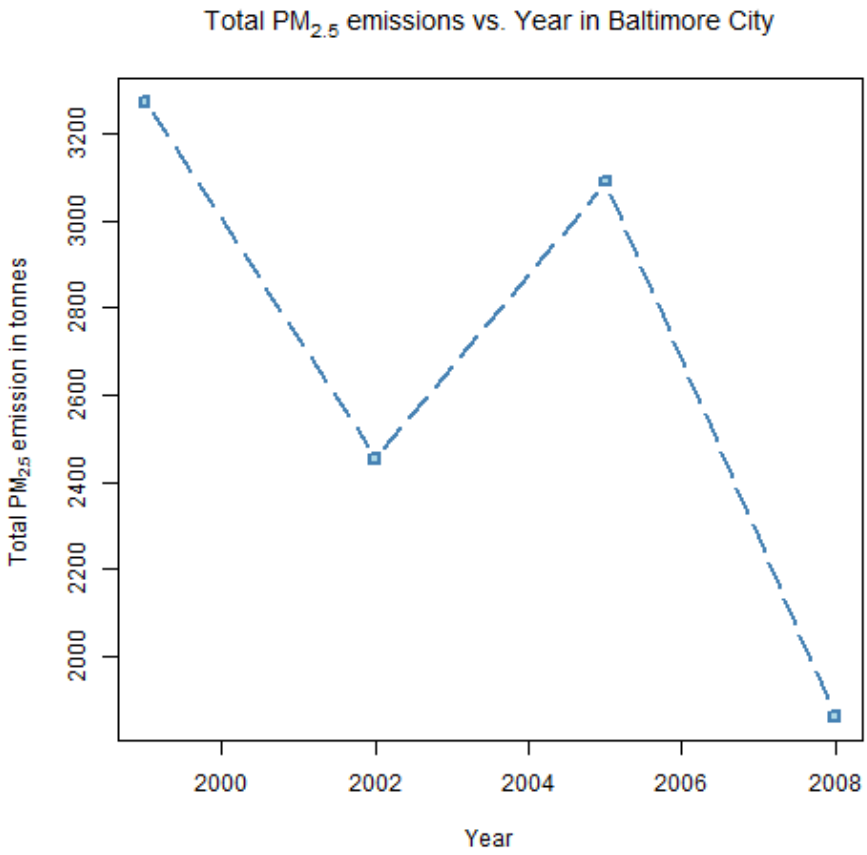
Upload the R code file for the plot uploaded in the previous question.

Please refer to the code segment below

```
# set the working directory to the directory where the data is present
# the path in the following line is just an example,
# uncomment replace with your own path
# setwd('E:/MOOCs/Coursera/Data Science - Specialization/Exploratory Data Analysis/Cour
se Project 2')

# read in the two datasets
pmed <- readRDS("summarySCC_PM25.rds")
scc <- readRDS("Source_Classification_Code.rds")

# get the total emissions for each year
totalpm <- aggregate(Emissions~year, pmed, sum)

# disabling scientific notation
options(scipen=999)

# plotting required graph
png(filename="plot1.png")
plot(totalpm$year,totalpm$Emissions,type='o', main = expression('Total '* PM[2.5] * ' e
missions vs. Year'),
     xlab="Year", ylab= expression('Total '* PM[2.5] * ' emission in tonnes'), col='ste
elblue', pch=22, lwd=2, cex=1, lty=5, bg='lightblue')
dev.off()
```

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.

Upload a PNG file containing your plot addressing this question.



From the above plot, we see that the total $PM_{2.5}$ emissions show a downward trend as the years progress, making it clear that the total emissions have decreased even though there was a peak increase in the emissions in the year 2005 but the general trend shows that with time it has been decreasing..

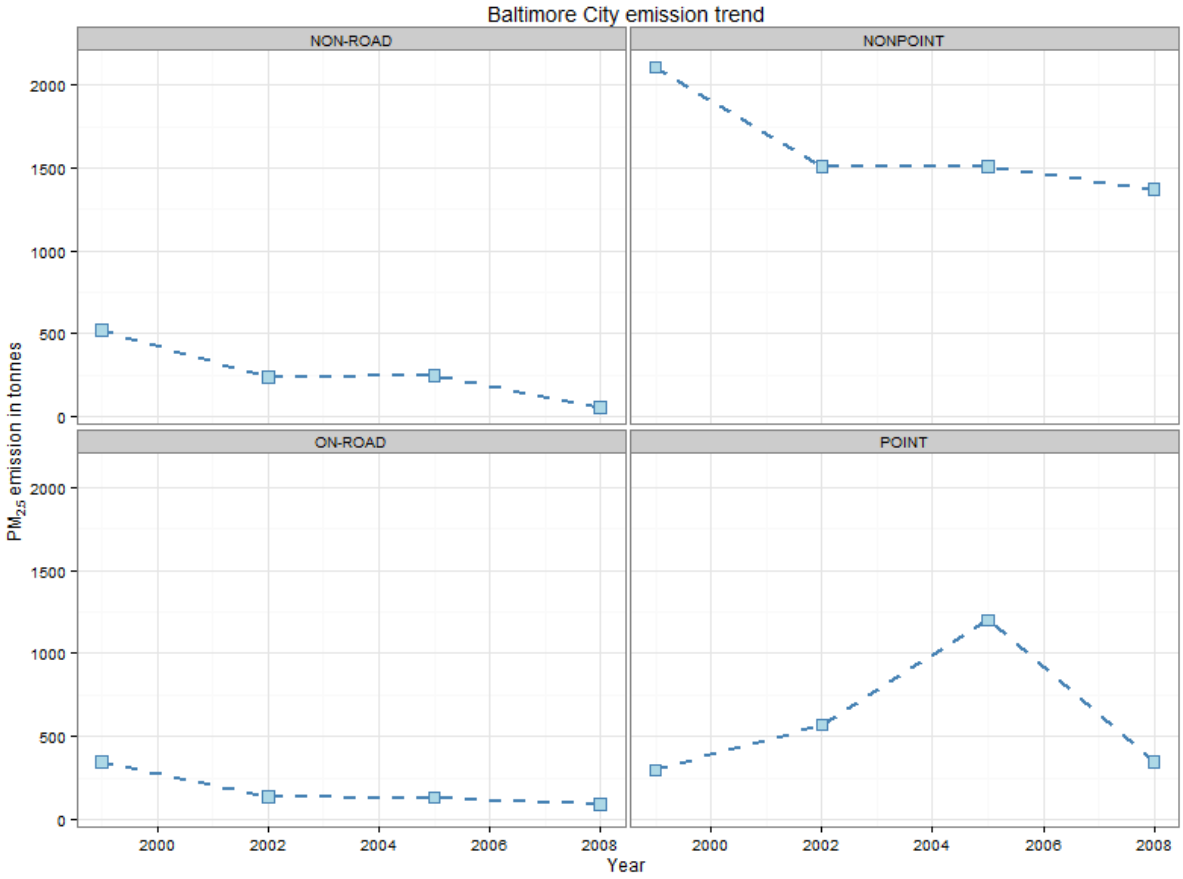
Upload the R code file for the plot uploaded in the previous question.

Please refer to the code segment below

```
# set the working directory to the directory where the data is present
# the path in the following line is just an example,
# uncomment replace with your own path
# setwd('E:/MOOCs/Coursera/Data Science - Specialization/Exploratory Data Analysis/Cour
se Project 2')

# read in the two datasets
pmed <- readRDS("summarySCC_PM25.rds")
scc <- readRDS("Source_Classification_Code.rds")

# subset out data for baltimore city
baltimorePMed <- pmed[pmed$fips == "24510",]

# get the total emissions for each year
totalpm <- aggregate(Emissions~year, baltimorePMed, sum)

# disabling scientific notation
options(scipen=999)

# plotting required graph
png(filename="plot2.png")
plot(totalpm$year,totalpm$Emissions,type='b', main = expression('Total '* PM[2.5] * ' e
missions vs. Year in Baltimore City'),
     xlab="Year", ylab=expression('Total '* PM[2.5] * ' emission in tonnes'), col='stee
lblue', pch=22, lwd=2, cex=1, lty=5, bg='lightblue')
dev.off()
```

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.

Upload a PNG file containing your plot addressing this question.



From the above plot of emissions in Baltimore City, it is clear that from 1999 - 2008,

- NON-ROAD, NONPOINT and ON-ROAD have seen clear cut decreases in emissions
- POINT has a general trend of increase in emissions but it has decreased rapidly from 2005 - 2008

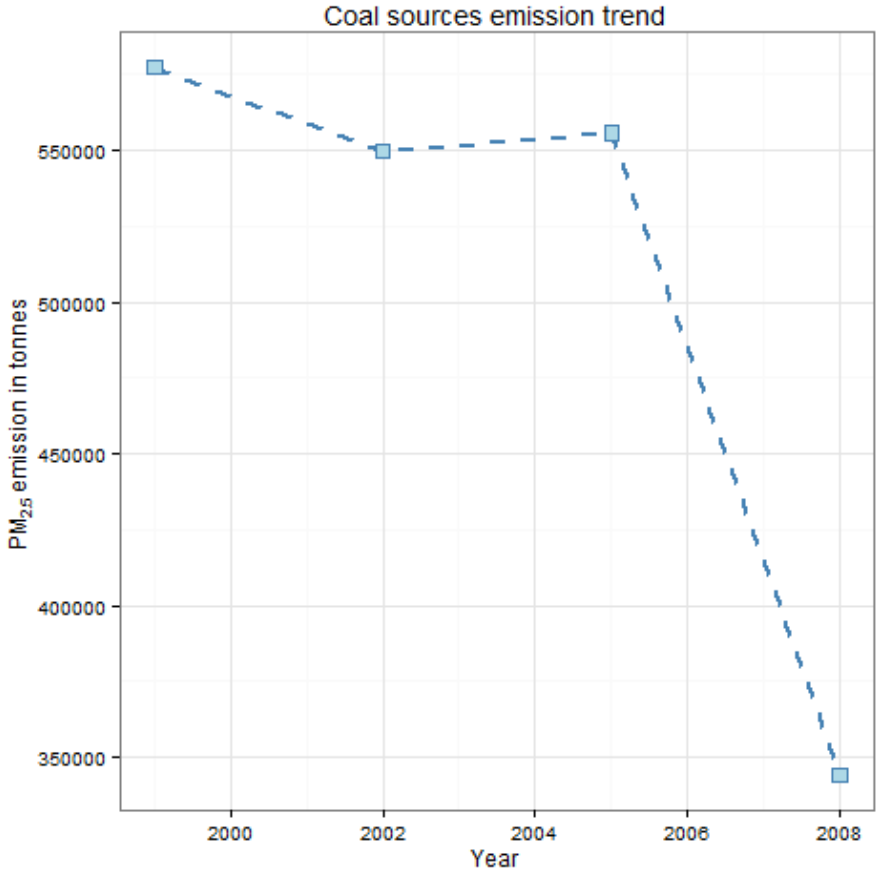
Upload the R code file for the plot uploaded in the previous question.

Please refer to the code segment below

```
# set the working directory to the directory where the data is present
# the path in the following line is just an example,
# uncomment replace with your own path
# setwd('E:/MOOCs/Coursera/Data Science - Specialization/Exploratory Data Analysis/Cour
se Project 2')

# read in the two datasets
pmed <- readRDS("summarySCC_PM25.rds")
scc <- readRDS("Source_Classification_Code.rds")

# subset out data for baltimore city
baltimorePMed <- pmed[pmed$fips == "24510",]

# disabling scientific notation
options(scipen=999)

# loading ggplot2 package
library(ggplot2)

# getting emissions for each year and type
pmedByType <- aggregate(Emissions ~ year + type, sum, data=baltimorePMed)

# plotting required graph
png(filename="plot3.png",width=800, height=600, units="px")
g <- ggplot(pmedByType, aes(year,Emissions)) +
     geom_line(color='steelblue', size=0.75, linetype='dashed') +
     geom_point(shape=22, size=4, color='steelblue',bg='lightblue') +
     facet_grid(type ~ .) +
     facet_wrap(~ type,nrow=2) +
     theme_bw() +
     labs(title = 'Baltimore City emission trend') +
     labs(x = 'Year') +
     labs(y = expression(PM[2.5]*' emission in tonnes'))

print(g)
dev.off()
```

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

Upload a PNG file containing your plot addressing this question.



For the above analysis we used combustible coal sources and from the trend we see that overall the emissions have decreased over time even though there was a slight increase from 2002 - 2005, it decreased rapidly from 2005 - 2008

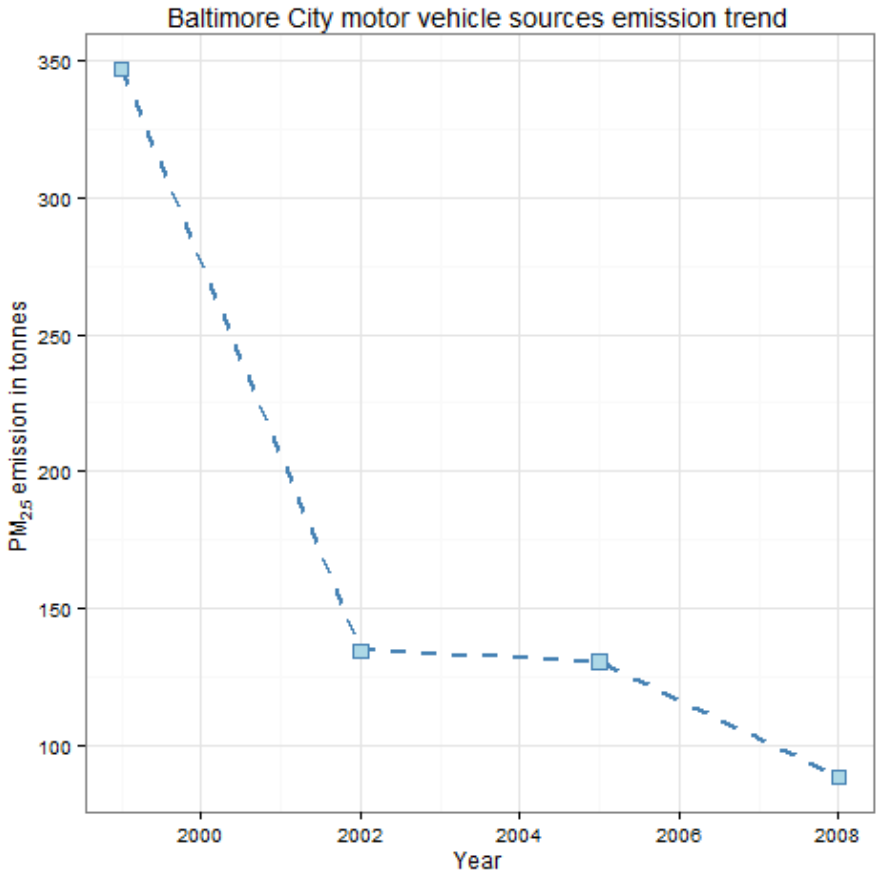
Examine the submitted R code file. Does the R code appear to construct the plot shown in the previous question? NOTE: Do not run the code on your own computer.

Please refer to the code segment below

```
# set the working directory to the directory where the data is present
# the path in the following line is just an example,
# uncomment replace with your own path
# setwd('E:/MOOCs/Coursera/Data Science - Specialization/Exploratory Data Analysis/Cour
se Project 2')

# read in the two datasets
pmed <- readRDS("summarySCC_PM25.rds")
scc <- readRDS("Source_Classification_Code.rds")

# getting source code classification details for coal combustion related sources
scc_coal_comb <- scc[
                     grepl("combustion", scc$SCC.Level.One, ignore.case=TRUE) &
                     !grepl("charcoal", scc$SCC.Level.Three, ignore.case=TRUE) &
                     (grepl("coal", scc$SCC.Level.Three, ignore.case=TRUE) |
                      grepl("lignite", scc$SCC.Level.Three, ignore.case=TRUE) |
                      grepl("anthracite", scc$SCC.Level.Three, ignore.case=TRUE)),
                   ]
# subsetting out data for emissions due to coal sources
pmed_coal_comb <- subset(pmed, SCC %in% scc_coal_comb$SCC)

# getting the total emissions for each year
totalpm <- aggregate(Emissions~year, pmed_coal_comb, sum)

# disabling scientific notation
options(scipen=999)

# loading ggplot2 package
library(ggplot2)

# plotting required graph
png(filename="plot4.png")
g <- ggplot(totalpm, aes(year,Emissions)) +
     geom_line(color='steelblue', size=0.75, linetype='dashed') +
     geom_point(shape=22, size=4, color='steelblue',bg='lightblue') +
     theme_bw() +
     labs(title = 'Coal sources emission trend') +
     labs(x = 'Year') +
     labs(y = expression(PM[2.5]*' emission in tonnes'))

print(g)
dev.off()
```

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

Upload a PNG file containing your plot addressing this question.



From the above plot of emissions in Baltimore City from motor vehicle sources, it shows a clear decreasing trend with progress of years.

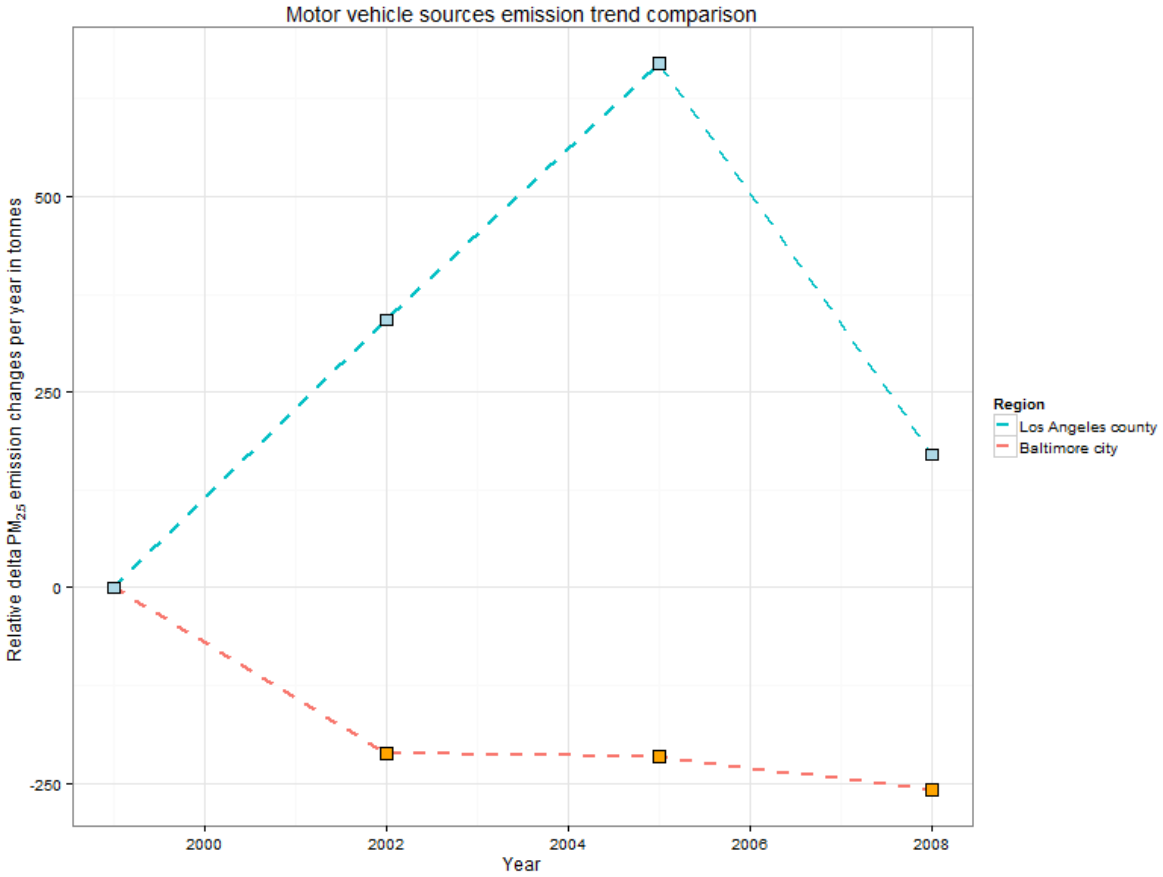
Upload the R code file for the plot uploaded in the previous question.

Please refer to the code segment below

```
# set the working directory to the directory where the data is present
# the path in the following line is just an example,
# uncomment replace with your own path
# setwd('E:/MOOCs/Coursera/Data Science - Specialization/Exploratory Data Analysis/Cour
se Project 2')

# read in the two datasets
pmed <- readRDS("summarySCC_PM25.rds")
scc <- readRDS("Source_Classification_Code.rds")

# subset out data for baltimore city and motor vehicles
baltimorePMed <- pmed[pmed$fips == "24510" & pmed$type=='ON-ROAD',]

# get the total emissions for each year
totalpm <- aggregate(Emissions~year, baltimorePMed, sum)

# disabling scientific notation
options(scipen=999)

# loading ggplot2 package
library(ggplot2)

# plotting required graph
png(filename="plot5.png")
g <- ggplot(totalpm, aes(year,Emissions)) +
     geom_line(color='steelblue', size=0.75, linetype='dashed') +
     geom_point(shape=22, size=4, color='steelblue',bg='lightblue') +
     theme_bw() +
     labs(title = 'Baltimore City motor vehicle sources emission trend') +
     labs(x = 'Year') +
     labs(y = expression(PM[2.5]*' emission in tonnes'))

print(g)
dev.off()
```

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?

Upload a PNG file containing your plot addressing this question.



Here by delta $PM_{2.5}$ emissions we mean the relative change in emissions per year. Since the two cities had different scales of emission, to bring them to a same scale we used some transformation where we took the baseline in 1999 as zero and calculated the relative change in emissions since then on a similar scale.

From the above plot we see that Baltimore City has seen a gradual decreasing trend in emissions. But the city which has seen greater changes over time has to be Los Angeles County because from the graph it is clear that the emissions rise rapidly till 2005 and then start dropping rapidly from 2005-2008.

Upload the R code file for the plot uploaded in the previous question.

Please refer to the code segment below

```
# set the working directory to the directory where the data is present
# the path in the following line is just an example,
# uncomment replace with your own path
# setwd('E:/MOOCs/Coursera/Data Science - Specialization/Exploratory Data Analysis/Cour
se Project 2')

# read in the two datasets
pmed <- readRDS("summarySCC_PM25.rds")
```

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

# subset out data for baltimore city and motor vehicles
baltimorePMed <- pmed[pmed$fips == "24510" & pmed$type=='ON-ROAD',]

# subset out data for los angeles county and motor vehicles
losangelesPMed <- pmed[pmed$fips == "06037" & pmed$type=='ON-ROAD',]

# get the total emissions for each year in baltimore city
totalpmBaltimore <- aggregate(Emissions~year, baltimorePMed, sum)

# get the total emissions for each year in los angeles county
totalpmLosAngeles <- aggregate(Emissions~year, losangelesPMed, sum)

# merge the above emissions into a single data frame
comparepm <- merge(x=totalpmBaltimore,y=totalpmLosAngeles,by.x='year',by.y='year')
names(comparepm) <- c('Year', 'baltimoreEmissions', 'laEmissions')

# calculate relative delta pm2.5 emission change per year by setting baseline
# emission in 1999 as zero and bringing both emissions to the same scale
comparepm$baltimoreEmissions <- comparepm$baltimoreEmissions - comparepm$baltimoreEmiss
ions[comparepm$Year==1999]
comparepm$laEmissions <- comparepm$laEmissions - comparepm$laEmissions[comparepm$Year==
1999]

# load required packages
library(ggplot2)
require(grid)

# plotting required graph
png(filename="plot6.png",width=800, height=600, units="px")
g <- ggplot(comparepm, aes(Year)) +
      geom_line(aes(y=baltimoreEmissions, color='baltimoreEmissions'), size=0.75,linety
pe='dashed') +
      geom_line(aes(y=laEmissions, color='laEmissions',), size=0.75,linetype='dashed')
+
      geom_point(aes(y=baltimoreEmissions),size=4,shape=22, bg='orange') +
      geom_point(aes(y=laEmissions),size=4,shape=22, bg='lightblue') +
      theme_bw() +
      theme(legend.position="right",legend.key.size = unit(0.5, "cm")) +
      scale_colour_discrete(name  ="Region",
                        breaks=c("laEmissions", "baltimoreEmissions"),
                        labels=c("Los Angeles county", "Baltimore city")) +
      labs(title = 'Motor vehicle sources emission trend comparison') +
      labs(x = 'Year') +
      labs(y = expression('Relative delta '*PM[2.5]*' emission changes per year in tonn
es'))



print(g)
dev.off()
```