Course Project 2

Student: Frank H Jung
Last saved: 16/05/2015, 08:40:44

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 (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 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 (https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip) [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 in the 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?

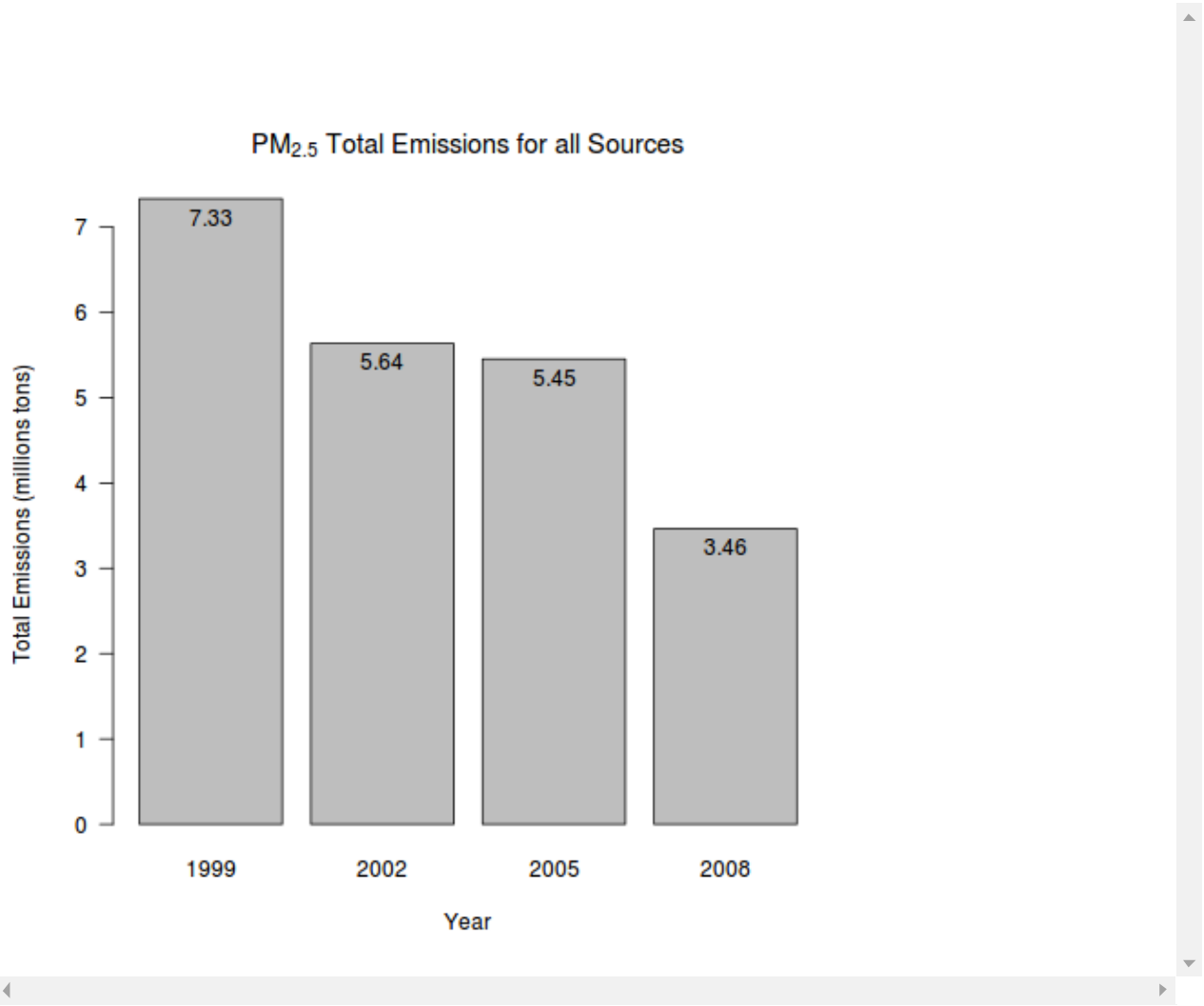
Making and Submitting Plots

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 must 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.

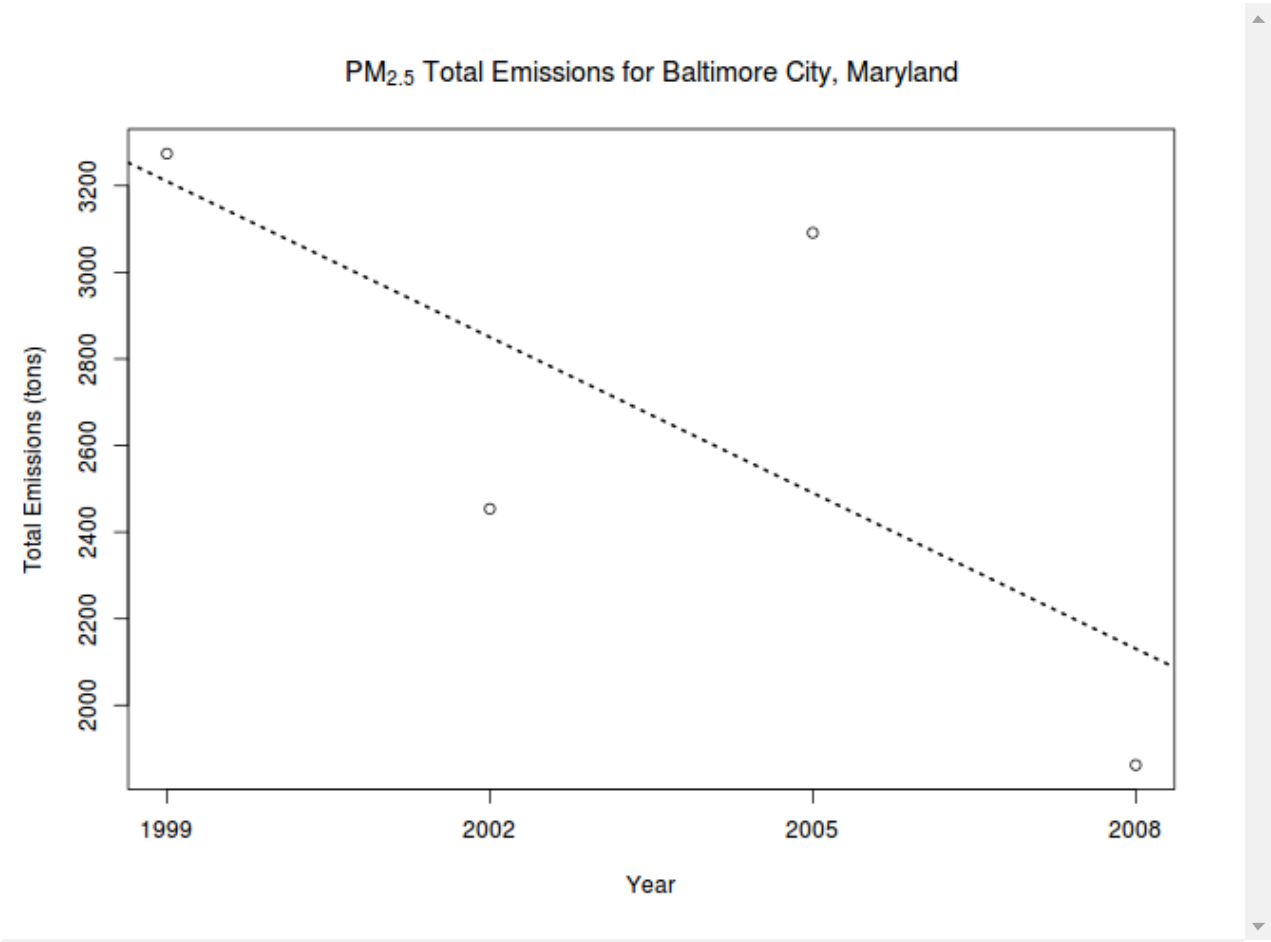


Copy and paste the R code file for the plot uploaded in the previous question.

```
library(dplyr)

nei <- readRDS("data/summarySCC_PM25.rds")

# aggregate emission by year
totals <- nei %>%
    select(year, Emissions) %>%
    arrange(year) %>%
    group_by(year) %>%
    summarise(total = sum(Emissions))

# report total emissions in millions of tons
totals <- transform(totals, total = total / 10^6)

# plot bar chart
png(filename = "plot1a.png", width = 480, height = 480, units = "px")
x <- with(totals, barplot(total, width = 4, names.arg = year, las = 1, yaxs = "i"))
with(totals, text(x, total, labels = round(total, 2), pos = 1, offset = 0.5))
title(xlab = "Year")
title(ylab = "Total Emissions (millions tons)")
title(main = expression(PM[2.5] * " Total Emissions for all Sources"))
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.

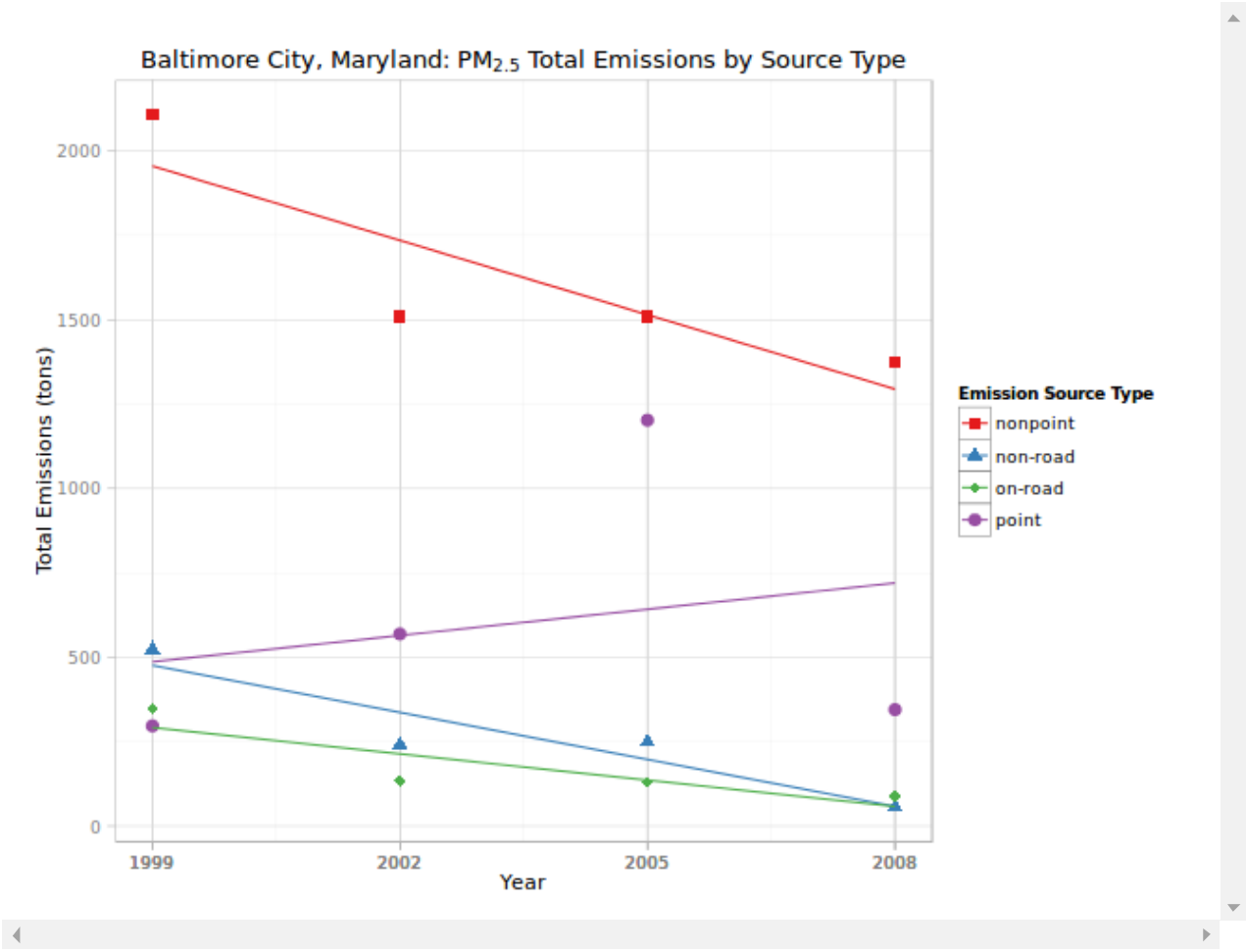


Copy and paste the R code file for the plot uploaded in the previous question.

```
nei <- readRDS("data/summarySCC_PM25.rds")

totals <- aggregate(Emissions ~ year, data = subset(nei, fips == "24510"), sum)
lmfit <- lm(Emissions ~ year, totals)

png(filename = "plot2-1.png", width=640, height=480, units="px")
plot(totals$year, totals$Emissions,
     xaxt = "n",
     xlab = "Year",
     ylab="Total Emissions (tons)",
     main = expression(PM[2.5] * " Total Emissions for Baltimore City, Maryland"))
axis(1, at = totals$year)
abline(lmfit, lty = 3, lwd = 2)
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.

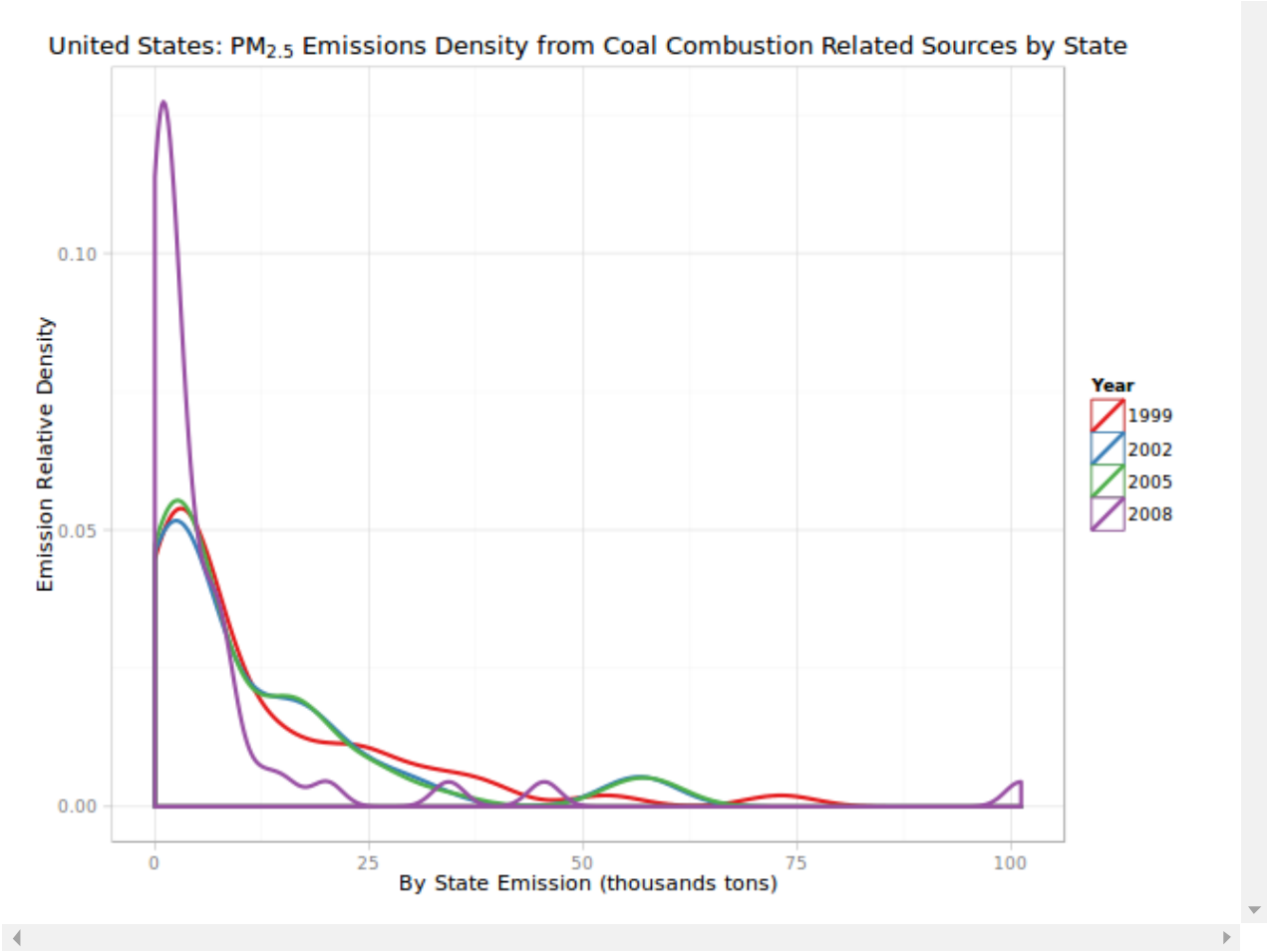


Copy and paste the R code file for the plot uploaded in the previous question.

```
library(dplyr)
library(ggplot2)

nei <- readRDS("data/summarySCC_PM25.rds")

# aggregate emission by year
totals <- nei %>%
    filter(fips == "24510") %>%
    select(year, type, Emissions) %>%
    arrange(year, type) %>%
    group_by(year, type) %>%
    summarise(total = sum(Emissions))

# for legend lowercase the emissions source types
totals <- transform(totals, type = factor(tolower(type)))

# points
png(filename = "plot3a.png", width = 640, height = 480, units = "px")
g <- ggplot(data = totals, aes(year, total))
g + geom_point(aes(color = type, shape = type), size = 3) +
    scale_shape_manual(values = c(15, 17, 18, 19)) +
    geom_smooth(method = "lm", se = FALSE, aes(color = type)) +
    theme_light(base_family = "Avenir", base_size = 11) +
    scale_color_brewer(palette = "Set1") +
    scale_x_continuous(name = "Year", breaks = totals$year) +
    labs(shape = "Emission Source Type", color = "Emission Source Type") +
    labs(y = "Total Emissions (tons)") +
    ggtitle(expression("Baltimore City, Maryland: " * PM[2.5] * " Total Emissions by
Source Type"))
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.

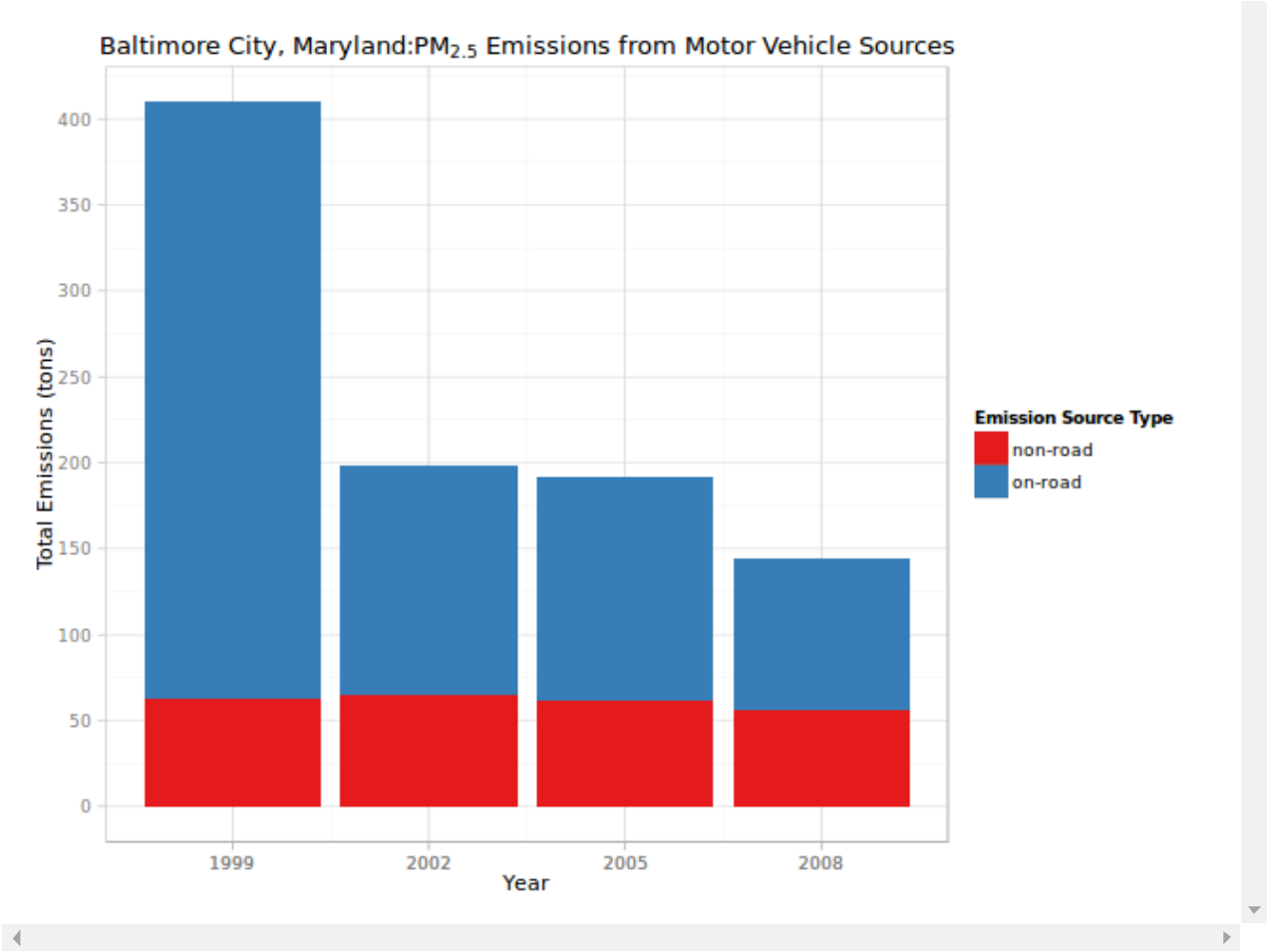


Copy and paste the R code file for the plot uploaded in the previous question.

```
library(dplyr)
library(ggplot2)
library(scales)

nei <- readRDS("data/summarySCC_PM25.rds")
scc <- readRDS("data/Source_Classification_Code.rds")

# get SCC (source code classification) digits for coal combustion related sources
coalscc <- as.character(scc[grepl("(?=.*Comb)(?=.*Coal)", scc$EI.Sector, perl = T),
"SCC"])

# aggregate emissions for each year by state
# only for state codes 01 ... 56, see http://www.epa.gov/envirofw/html/codes/state.ht
ml (http://www.epa.gov/envirofw/html/codes/state.html)
totals <- nei %>%
    filter(SCC %in% coalscc) %>%
    mutate(state = as.integer(substr(fips, 1, 2))) %>%
    filter(state < 56) %>%
    select(year, state, Emissions) %>%
    arrange(year, state) %>%
    group_by(year, state) %>%
    summarise(total = sum(Emissions))
totals <- transform(totals, state = factor(state), total = total / 1000, year = facto
r(year))

png(filename = "plot4e.png", width = 640, height = 480, units = "px")
attach(totals)
g <- ggplot(data = totals, aes(x = total))
g + geom_density(aes(group = year, color = year), size = 1) +
    theme_light(base_family = "Avenir", base_size = 11) +
    scale_color_brewer(palette = "Set1") +
    xlab(label = "By State Emission (thousands tons)") +
    scale_y_continuous(name = "Emission Relative Density") +
    labs(color = "Year") +
    ggtitle(expression("United States: " * PM[2.5] * " Emissions Density from Coal Co
mbustion Related Sources by State"))
detach(totals)
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.

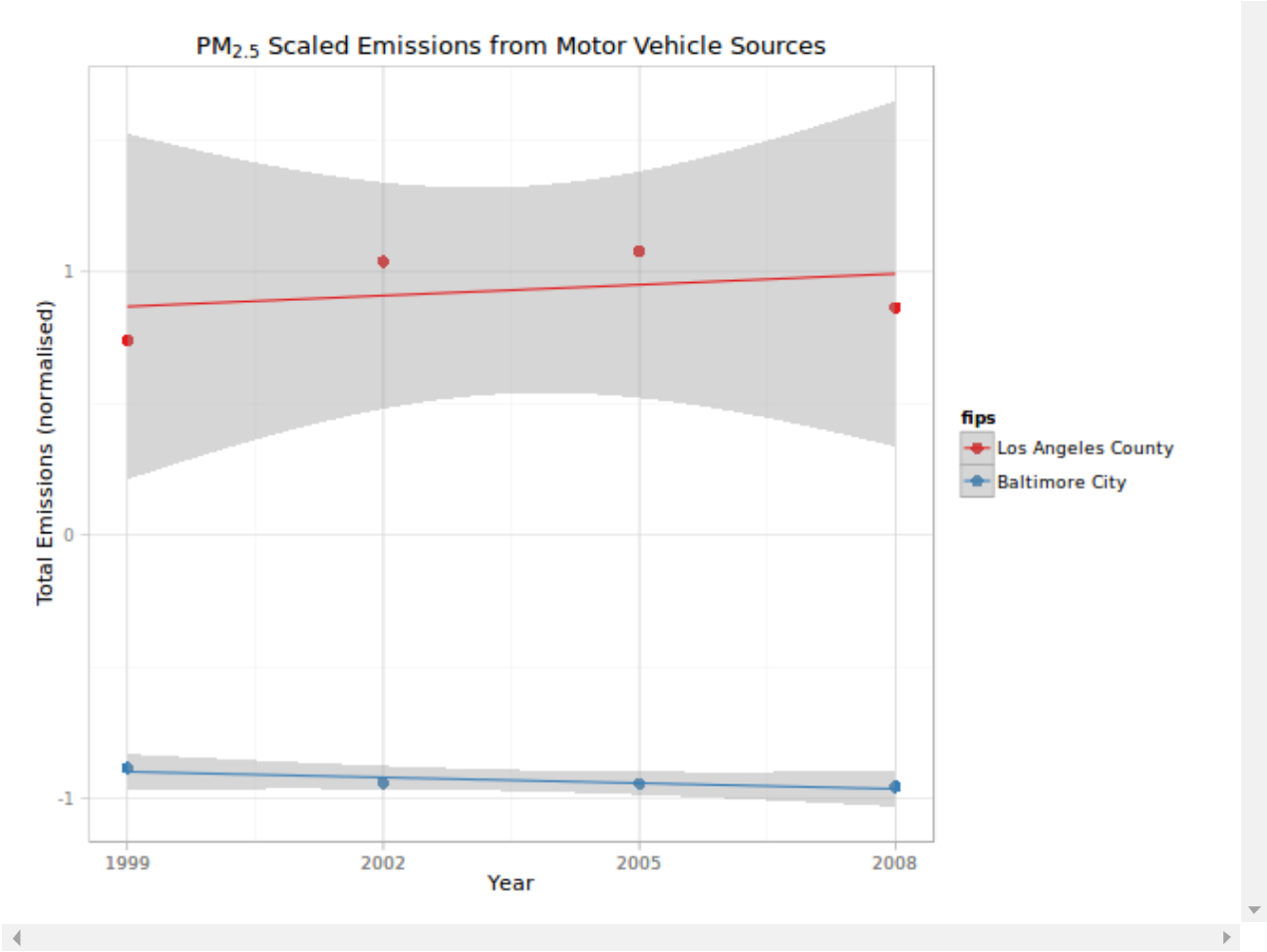


Copy and paste the R code file for the plot uploaded in the previous question.

```
library(dplyr)
library(ggplot2)
library(scales)

nei <- readRDS("data/summarySCC_PM25.rds")
scc <- readRDS("data/Source_Classification_Code.rds")

# get SCC (source code classification) digits for mobile sources
vehiclescc <- as.character(scc[grepl("(?=.*Mobile - )(?=.*-Road)", scc$EI.Sector, per
l = T), "SCC"])

# aggregate emissions by year
totals <- nei %>%
    filter(fips == "24510") %>%
    filter(SCC %in% vehiclescc) %>%
    select(year, type, Emissions) %>%
    arrange(year, type) %>%
    group_by(year, type) %>%
    summarise(total = sum(Emissions))
totals <- transform(totals, type = factor(tolower(type)))

# plot bar graph
png(filename = "plot5b.png", width = 640, height = 480, units = "px")
attach(totals)
g <- ggplot(data = totals, aes(year, total, fill = type))
g + geom_bar(stat = "identity", position = "stack") +
    theme_light(base_family = "Avenir", base_size = 11) +
    scale_fill_brewer(name = "Emission Source Type", palette = "Set1") +
    scale_x_continuous(name = "Year", breaks = year) +
    scale_y_continuous(name = "Total Emissions (tons)", breaks = pretty_breaks(n = 1
0)) +
    ggtitle(expression("Baltimore City, Maryland:" * PM[2.5] * " Emissions from Motor
Vehicle Sources"))
detach(totals)
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.



Copy and paste the R code file for the plot uploaded in the previous question.

```
library(dplyr)
library(ggplot2)

nei <- readRDS("data/summarySCC_PM25.rds")
scc <- readRDS("data/Source_Classification_Code.rds")

# get SCC (source code classification) digits for motor vehicle sources
vehiclescc <- as.character(scc[grepl("(?=.*Mobile - )(?=.*-Road)", scc$EI.Sector, per
l = T), "SCC"])

# scale emissions by year by county and type
totals <- nei %>%
    filter(fips == "06037" | fips == "24510") %>%
    filter(SCC %in% vehiclescc) %>%
    select(year, fips, Emissions) %>%
    arrange(year, fips) %>%
    group_by(year, fips) %>%
    summarise(total = sum(Emissions))

totals <- transform(totals, scale = scale(total),
                    fips = factor(fips, labels = c("Los Angeles County", "Baltimore C
ity")))

png(filename = "plot6c.png", width = 640, height = 480, units = "px")
attach(totals)
g <- ggplot(data = totals, aes(year, scale))
g + geom_point(aes(color = fips), size = 3) +
    theme_light(base_family = "Avenir", base_size = 11) +
    geom_smooth(method = "lm", se = TRUE, aes(color = fips)) +
    scale_color_brewer(palette = "Set1") +
    scale_x_continuous(name = "Year", breaks = year) +
    labs(y = "Total Emissions (normalised)") +
    ggtitle(expression(PM[2.5] * " Scaled Emissions from Motor Vehicle Sources"))
detach(totals)
dev.off()
```