Analysis of PM2.5 in the U.S. and its Impact on Public Health

Jeffrey Hunter

18 May, 2019

Contents

Course Project

Exploratory Data Analysis Course Project 2

Peer-graded Assignment

- This course project is available on GitHub

  Analysis of PM2.5 in the United States and its Impact on Public Health

Synopsis

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.

Assignment

The overall goal of this assignment 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 1999–2008. You may use any R package you want to support your analysis.

The deliverable for this assignment is to address the questions and tasks provided in the Questions section of this report.

Environment Setup

Load packages used in this analysis

```
if (!require(ggplot2)) {
    install.packages("ggplot2", repos = "http://cran.us.r-project.org")
    library(ggplot2, warn.conflicts = FALSE)
}
```

```
## Loading required package: ggplot2
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
if (!require(dplyr)) {
    install.packages("dplyr", repos = "http://cran.us.r-project.org")
    library(dplyr, warn.conflicts = FALSE)
}
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
if (!require(scales)) {
    install.packages("scales", repos = "http://cran.us.r-project.org")
    library(scales, warn.conflicts = FALSE)
}
```

```
## Loading required package: scales
```

```
if (!require(stringi)) {
    install.packages("stringi", repos = "http://cran.us.r-project.org")
    library(stringi, warn.conflicts = FALSE)
}
```

```
## Loading required package: stringi
```

Session information

```
sessionInfo()
```

```
## R version 3.6.0 (2019-04-26)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
```

```
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] stringi_1.4.3 scales_1.0.0  dplyr_0.8.1   ggplot2_3.1.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.1       knitr_1.22       magrittr_1.5     tidyselect_0.2.5
##  [5] munsell_0.5.0    colorspace_1.4-1 R6_2.4.0         rlang_0.3.4
##  [9] stringr_1.4.0    plyr_1.8.4       tools_3.6.0      grid_3.6.0
## [13] packrat_0.5.0    gtable_0.3.0     xfun_0.7         withr_2.1.2
## [17] htmltools_0.3.6  assertthat_0.2.1 yaml_2.2.0       lazyeval_0.2.2
## [21] digest_0.6.18    tibble_2.1.1     crayon_1.3.4     purrr_0.3.2
## [25] glue_1.3.1       evaluate_0.13    rmarkdown_1.12   compiler_3.6.0
## [29] pillar_1.4.0     pkgconfig_2.0.2
```

Load Data

Download the compressed data file from the source URL (if not found locally), uncompress it and then load the two serialized R objects via the `readRDS` function. Prior to processing the data, validate the downloaded data file and loaded dataset by checking the file size and dimensions respectively.

```r
setwd("~jhunter/repos/coursera/data-science-specialization/exploratory-data-analysis-course-project-2")
extNeiDataFileURL <- "https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip"
extNeiDataFile <- "data/exdata_data_NEI_data.zip"
neiDataFile <- "data/summarySCC_PM25.rds"
sccDataFile <- "data/Source_Classification_Code.rds"
if (!file.exists('data')) {
    dir.create('data')
}
if (!file.exists(extNeiDataFile)) {
    download.file(url = extNeiDataFileURL, destfile = extNeiDataFile)
    unzip(extNeiDataFile, exdir = "data")
}
stopifnot(file.size(extNeiDataFile) == 30643310)
NEI <- readRDS(neiDataFile)
SCC <- readRDS(sccDataFile)
stopifnot(dim(NEI) == c(6497651, 6))
stopifnot(dim(SCC) == c(11717, 15))
```

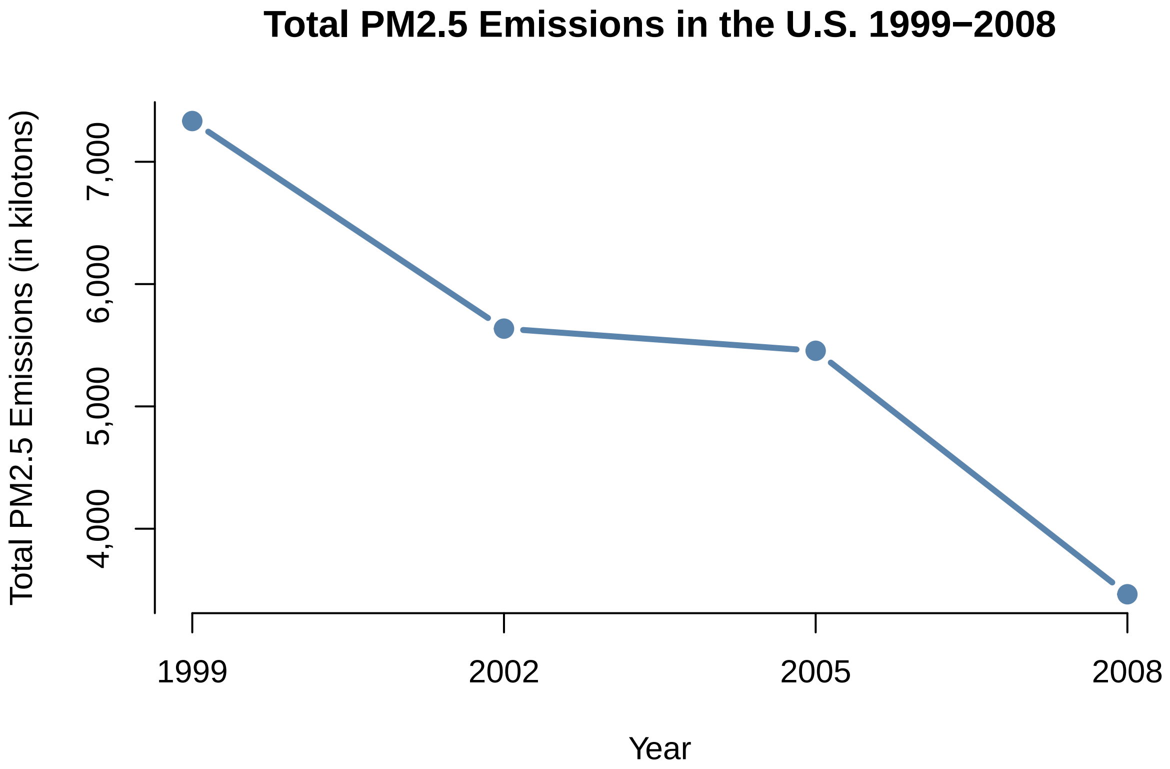
Questions

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.

   Answer

   The answer to the question is yes. The plot below shows that total emissions from PM2.5 has decreased in the United States from 1999 to 2008.

```r
# summarize total emissions by year for U.S.
totalByYear <- NEI %>%
               group_by(year) %>%
               filter(year == 1999|2002|2005|2008) %>%
               summarize(Total.Emissions = sum(Emissions))

# manually specify axis parameters
xYears <- c(1999, 2002, 2005, 2008)
yEmiss <- pretty(totalByYear$Total.Emissions/10^3, n = 4)

# create plot with custom axes
plot(totalByYear$year,
     totalByYear$Total.Emissions/10^3,
     type = "b",
     bty = "l",
     lwd = 3,
     pch = 19,
     col = rgb(0.2, 0.4, 0.6, 0.8),
     axes = FALSE,
     yaxt = "n",
     xlab = "Year",
     ylab = "Total PM2.5 Emissions (in kilotons)",
     main = "Total PM2.5 Emissions in the U.S. 1999-2008")
axis(1, at = xYears, labels = xYears)
axis(2, at = yEmiss, labels = prettyNum(yEmiss, big.mark = ","))
```

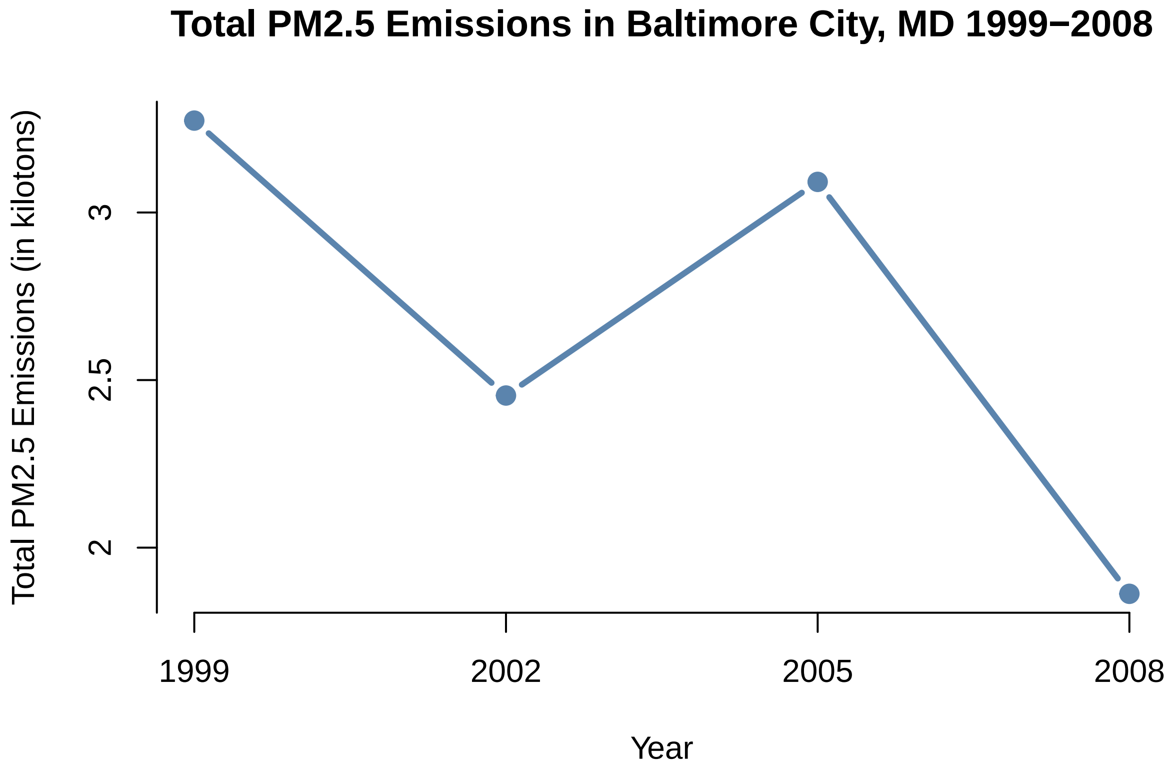


2. Have total emissions from PM2.5 decreased in the **Baltimore City**, Maryland (`fips == "24510"`) from 1999 to 2008? Use the **base** plotting system to make a plot answering this question.

   Answer

The answer to the question is yes. The plot below shows that total emissions from PM2.5 has decreased in Baltimore City, MD from 1999 to 2008. The plot also shows that the decrease in emissions from 1999 to 2008 did not follow a consistent negative trend. A sharp rise in emissions was observed between 2002 and 2005.

```r
# summarize total emissions by year for Baltimore (fips == "24510")
totalBaltimore <- NEI %>%
                  filter(fips == "24510") %>%
                  group_by(year) %>%
                  filter(year == 1999|2002|2005|2008) %>%
                  summarize(Total.Emissions = sum(Emissions))

# manually specify axis parameters
xYears <- c(1999, 2002, 2005, 2008)
yEmiss <- pretty(totalBaltimore$Total.Emissions/10^3, n = 4)

# create plot with custom axes
plot(totalBaltimore$year,
     totalBaltimore$Total.Emissions/10^3,
     type = "b",
     bty = "l",
     lwd = 3,
     pch = 19,
     col = rgb(0.2, 0.4, 0.6, 0.8),
     axes = FALSE,
     xlab = "Year",
     ylab = "Total PM2.5 Emissions (in kilotons)",
     main = "Total PM2.5 Emissions in Baltimore City, MD 1999-2008")
axis(1, at = xYears, labels = xYears)
axis(2, at = yEmiss, labels = yEmiss)
```

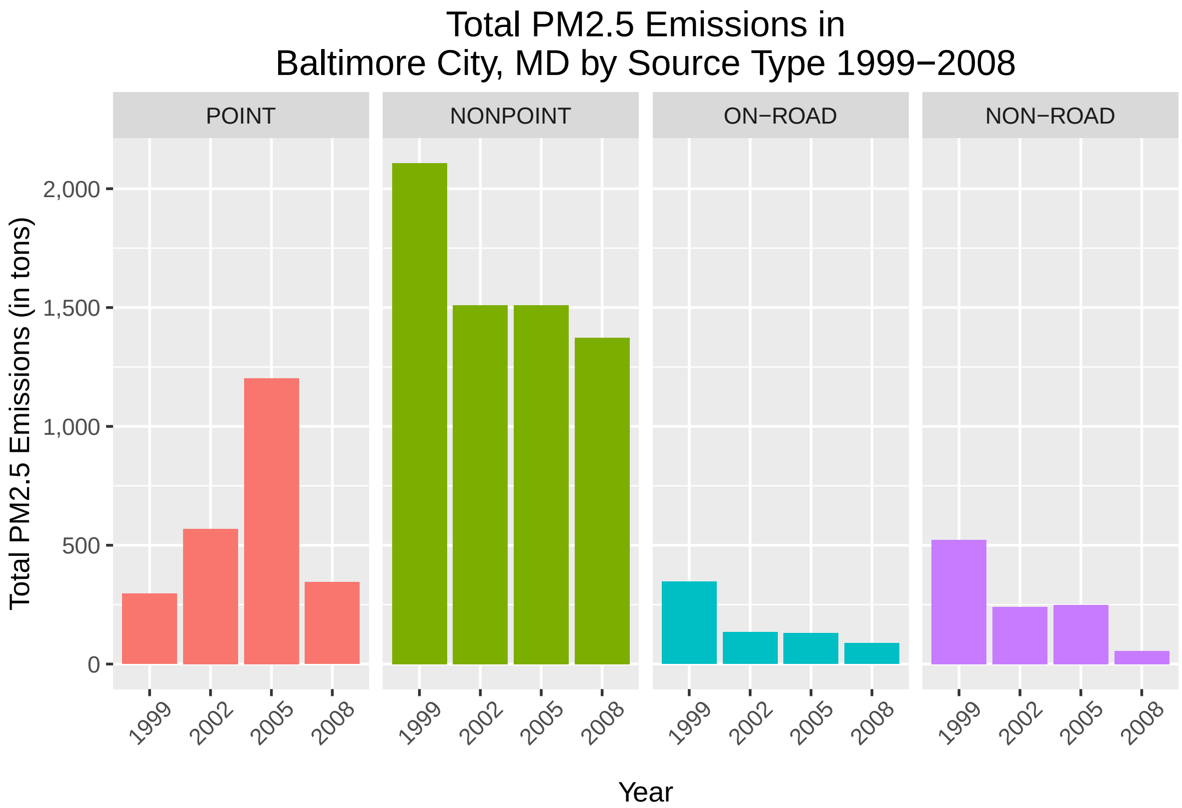


3. Of the four types of sources indicated by the `type` (point, nonpoint, onroad, nonroad) variable, which of these four sources have seen decreases in emissions from 1999–2008 for **Baltimore City**? Which have seen increases in emissions from 1999–2008? Use the **ggplot2** plotting system to make a plot answer this question.

   Answer

   The plot below shows that total emissions from PM2.5 has decreased in Baltimore City, MD from 1999 to 2008 for the *nonpoint*, *on-road* and *non-road* source types whereas the *point* source type has seen an increase in emissions during the same time period.

```
# summarize total emissions by source type and year for Baltimore (fips == "24510")
totalBaltimoreBySource <- NEI %>%
                          filter(fips == "24510") %>%
                          group_by(type, year) %>%
                          filter(year == 1999|2002|2005|2008) %>%
                          summarize(Total.Emissions = sum(Emissions))

# optional: specify the order of the plots
totalBaltimoreBySource$type <- factor(totalBaltimoreBySource$type,
                                      levels = c("POINT", "NONPOINT", "ON-ROAD", "NON-ROAD"))

# create plot
gTotBalSrc <- ggplot(data = totalBaltimoreBySource, aes(x = factor(year),
                                                        y = Total.Emissions,
                                                        fill = type)) +
     geom_bar(stat = "identity") +
     facet_grid(. ~ type) +
     xlab("Year") +
     ylab("Total PM2.5 Emissions (in tons)") +
     guides(fill = FALSE) +
     theme(plot.title = element_text(size = 14, hjust = 0.5, vjust = 0.5),
           axis.text.x = element_text(angle = 45,
                                      hjust = 0.5,
                                      vjust = 0.5,
                                      margin = margin(b = 10))) +
     scale_y_continuous(labels = comma) +
     ggtitle("Total PM2.5 Emissions in\nBaltimore City, MD by Source Type 1999-2008")
print(gTotBalSrc)
```

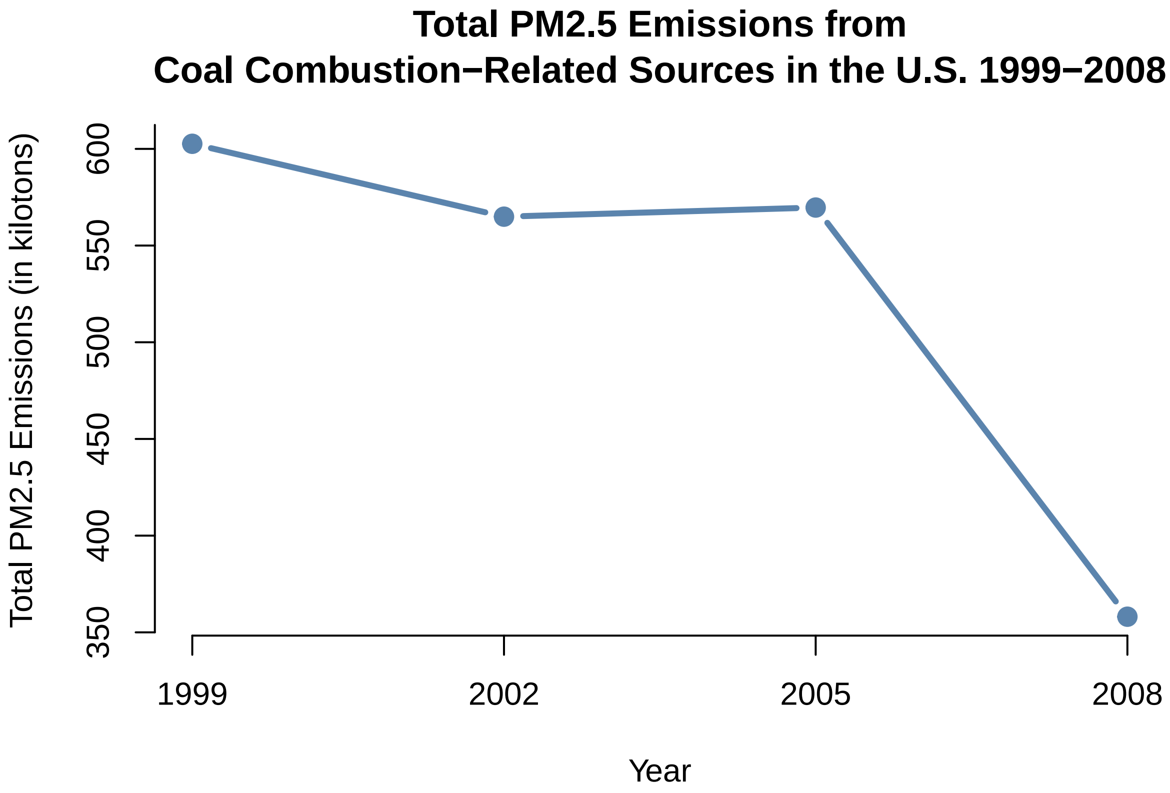


4. Across the United States, how have emissions from coal combustion-related sources changed from 1999–2008?

   Answer

   The plot below shows that emissions from coal combustion-related sources has decreased from 1999-2008.

```r
# select SCC values in SCC where Short.Name includes "[Cc]oal"
coalSCC <- subset(SCC,
                  stri_detect_regex(Short.Name, "Coal", case_insensitive = TRUE),
                  select = c(SCC))

# convert to numeric factor
coalSCC <- coalSCC$SCC

# select year, Emissions from NEI where SCC in (filtered list of SCC values)
coalNEI <- subset(NEI, SCC %in% coalSCC, select = c(Emissions, year))

# summarize total coal emissions by year for U.S.
totalCoalNEI <- coalNEI %>%
                group_by(year) %>%
                filter(year == 1999|2002|2005|2008) %>%
                summarize(Total.Emissions = sum(Emissions))

# manually specify axis parameters
xYears <- c(1999, 2002, 2005, 2008)
yEmiss <- pretty(totalCoalNEI$Total.Emissions/10^3, n = 4)

# create plot with custom axes
plot(totalCoalNEI$year,
     totalCoalNEI$Total.Emissions/10^3,
     type = "b",
     bty = "l",
     lwd = 3,
     pch = 19,
     col = rgb(0.2, 0.4, 0.6, 0.8),
     axes = FALSE,
     xlab = "Year",
     ylab = "Total PM2.5 Emissions (in kilotons)",
     main = "Total PM2.5 Emissions from\nCoal Combustion-Related Sources in the U.S. 1999-2008")
axis(1, at = xYears, labels = xYears)
axis(2, at = yEmiss, labels = yEmiss)
```



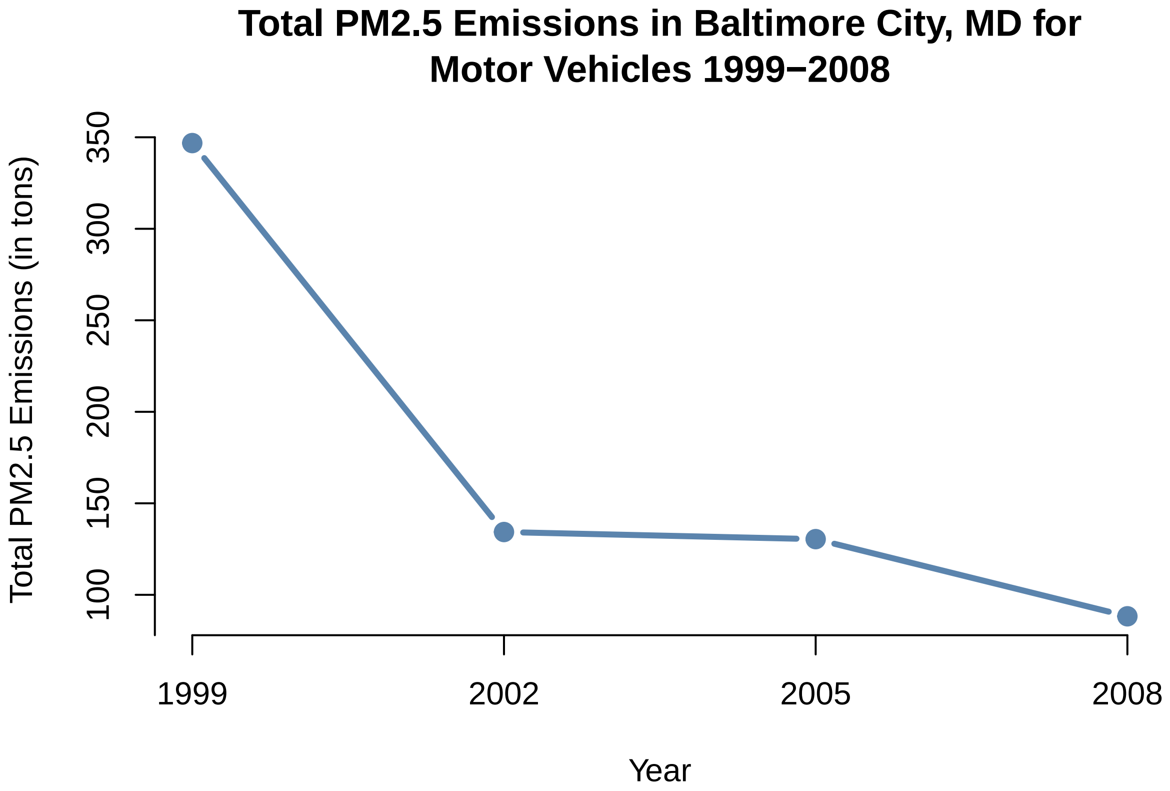
5. How have emissions from motor vehicle sources changed from 1999–2008 in **Baltimore City**?

   Answer

   The plot below shows that emissions from motor vehicle sources has decreased from 1999-2008.

```
# select SCC values in SCC where EI.Sector includes "[Vv]ehicle"
vehicleSCC <- subset(SCC,
                     stri_detect_regex(EI.Sector, "Vehicle", case_insensitive = TRUE),
                     select = c(SCC))

# convert to numeric factor
vehicleSCC <- vehicleSCC$SCC

# select year, Emissions, fips from NEI where SCC in (filtered list of SCC values)
vehicleNEI <- subset(NEI, SCC %in% vehicleSCC, select = c(Emissions, year, fips))
```

```r
# summarize vehicle emissions by year for Baltimore (fips == "24510")
totalVehicleNEI <- vehicleNEI %>%
                   filter(fips == "24510") %>%
                   group_by(year) %>%
                   filter(year == 1999|2002|2005|2008) %>%
                   summarize(Total.Emissions = sum(Emissions))

# manually specify axis parameters
xYears <- c(1999, 2002, 2005, 2008)
yEmiss <- pretty(totalVehicleNEI$Total.Emissions, n = 4)

# create plot with custom axes
plot(totalVehicleNEI$year,
     totalVehicleNEI$Total.Emissions,
     type = "b",
     bty = "l",
     lwd = 3,
     pch = 19,
     col = rgb(0.2, 0.4, 0.6, 0.8),
     axes = FALSE,
     xlab = "Year",
     ylab = "Total PM2.5 Emissions (in tons)",
     main = "Total PM2.5 Emissions in Baltimore City, MD for\nMotor Vehicles 1999-2008")
axis(1, at = xYears, labels = xYears)
axis(2, at = yEmiss, labels = yEmiss)
```



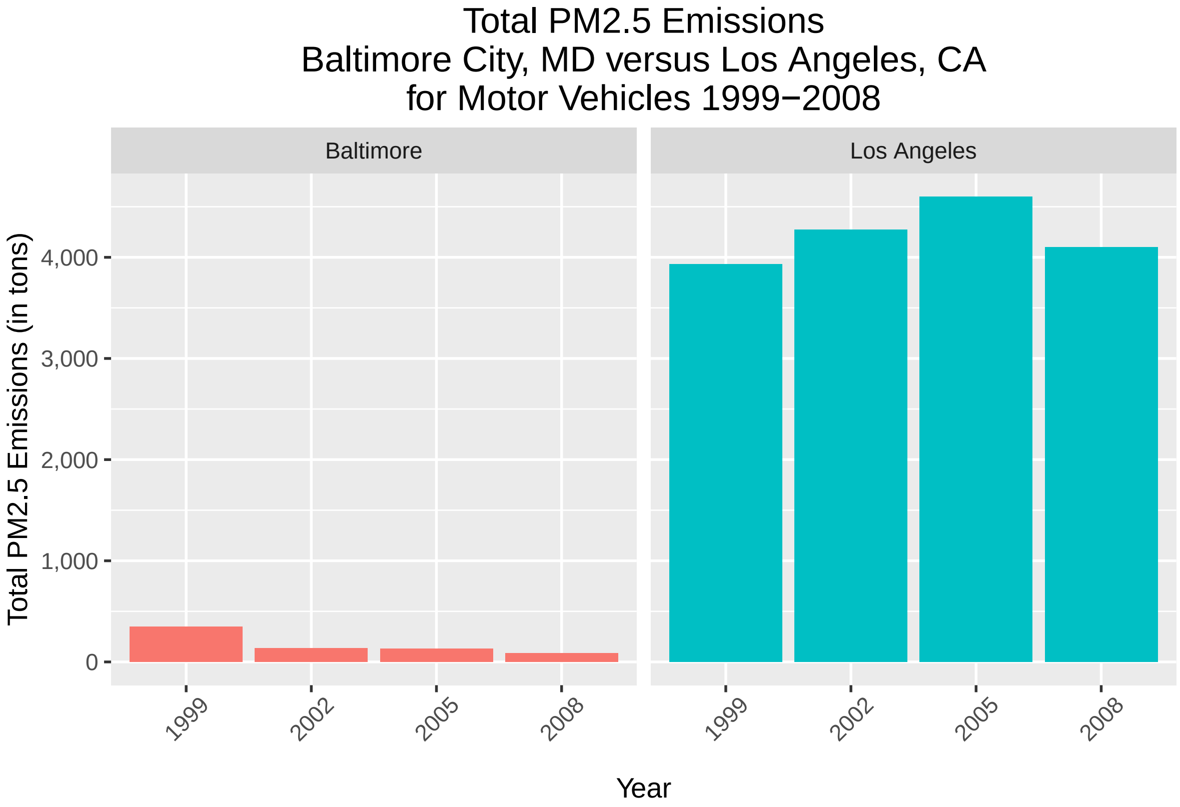
6. Compare emissions from motor vehicle sources in **Baltimore City** with emissions from motor vehicle sources in **Los Angeles County**, California (`fips == "06037"`). Which city has seen greater changes over time in motor vehicle emissions?

Answer

The plot below shows that emissions from motor vehicle sources in Baltimore City has decreased from 1999-2008 whereas Los Angeles County has seen an increase from motor vehicle sources during the same time period.

```
# select SCC values in SCC where EI.Sector includes "[Vv]ehicle"
vehicleSCC <- subset(SCC,
                     stri_detect_regex(EI.Sector, "Vehicle", case_insensitive = TRUE),
                     select = c(SCC))

# convert to numeric factor
vehicleSCC <- vehicleSCC$SCC

# select year, Emissions, fips from NEI where SCC in (filtered list of SCC values)
vehicleNEI <- subset(NEI, SCC %in% vehicleSCC, select = c(Emissions, year, fips))

# summarize vehicle emissions by year for Baltimore (fips == "24510")
vehicleBaltimoreNEI <- vehicleNEI %>%
                       filter(fips == "24510") %>%
                       mutate(city = "Baltimore") %>%
                       group_by(city, year) %>%
                       filter(year == 1999|2002|2005|2008) %>%
                       summarize(Total.Emissions = sum(Emissions))

# summarize vehicle emissions by year for Los Angeles (fips == "06037")
vehicleLosAngelesNEI <- vehicleNEI %>%
                        filter(fips == "06037") %>%
                        mutate(city = "Los Angeles") %>%
                        group_by(city, year) %>%
                        filter(year == 1999|2002|2005|2008) %>%
                        summarize(Total.Emissions = sum(Emissions))

# combine both datasets
totalVehicleNEI <- rbind(vehicleBaltimoreNEI, vehicleLosAngelesNEI)

# optional: specify the order of the plots
totalVehicleNEI$city <- factor(totalVehicleNEI$city, levels = c("Baltimore", "Los Angeles"))

# create plot
gtotalVehicleNEI <- ggplot(data = totalVehicleNEI, aes(x = factor(year),
                                                       y = Total.Emissions,
                                                       fill = city)) +
     geom_bar(stat = "identity") +
     facet_grid(. ~ city) +
     xlab("Year") +
     ylab("Total PM2.5 Emissions (in tons)") +
     guides(fill = FALSE) +
     theme(plot.title = element_text(size = 14, hjust = 0.5, vjust = 0.5),
           axis.text.x = element_text(angle = 45,
                                      hjust = 0.5,
                                      vjust = 0.5,
                                      margin = margin(b = 10))) +
     scale_y_continuous(labels = comma) +
     ggtitle("Total PM2.5 Emissions\nBaltimore City, MD versus Los Angeles, CA\nfor Motor Vehicles
```

```r
print(gtotalVehicleNEI)
```


Total PM2.5 Emissions
Baltimore City, MD versus Los Angeles, CA
for Motor Vehicles 1999−2008
Baltimore
Los Angeles
Total PM2.5 Emissions (in tons)
Year