Exploratory Data Analysis Course Work

Zanin Pavel

March 5, 2016

Link to project on GitHUB

Synopsis

The overall goal of this assigment 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 to 2008.

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. Approximately 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.

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.

Questions:

1. Have total emissions from $PM_{2.5}$ decreased in the United States from 1999 to 2008?

2. Have total emissions from $PM_{2.5}$ decreased in the Baltimore City, Maryland from 1999 to 2008?

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?

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

Data [29Mb]

The data for this assignment are available from the course web site as a single zip file initially containing two files:

- summarySCC_PM25.rds

- Source_Classification_Code.rds

Operating System and Environment Specs

- Windows 8
- R Studio version 0.99.489 - © 2009-2015 RStudio, Inc.
- R Project version R 3.3.0

Data processing

Downloading the data

```
if(!file.exists("./data")){dir.create("./data")}
fileUrl <- "https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip"
download.file(fileUrl,destfile="./data/exdata-data-NEI_data.zip")

# Unzip dataSet to /data directory
unzip(zipfile="./data/exdata-data-NEI_data.zip",exdir="./data")
```

Reading the data

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

Question 1 - Have total emissions from $PM_{2.5}$ decreased in the United States from 1999 to 2008?

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.3
```

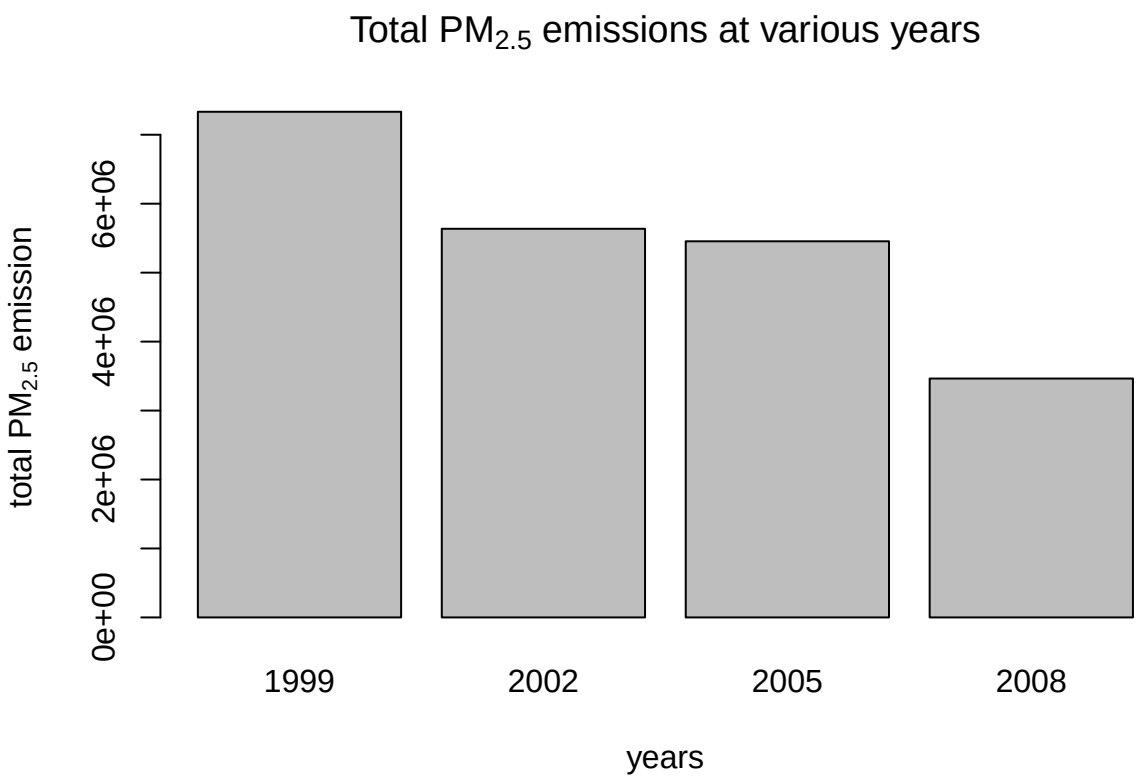
```
##
## Attaching package: 'dplyr'
```

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

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

```
# Defining total emissions per year
TotalEmissions <-
                  NEI %>%
                  group_by(year) %>%
                  summarise(SumEmissions = sum(Emissions))
```

```r
# Plotting total emissions per year
barplot(height = TotalEmissions$SumEmissions,
        names.arg = TotalEmissions$year,
        xlab="years",
        ylab=expression('total PM'[2.5]*' emission'),
        main=expression('Total PM'[2.5]*' emissions at various years')
        )
```



Answer: Yes, they sharply declined from 1999 to 2002. Then a slower decline between 2002 and 2005. Finally, they sharply declined from 2005 to 2008.

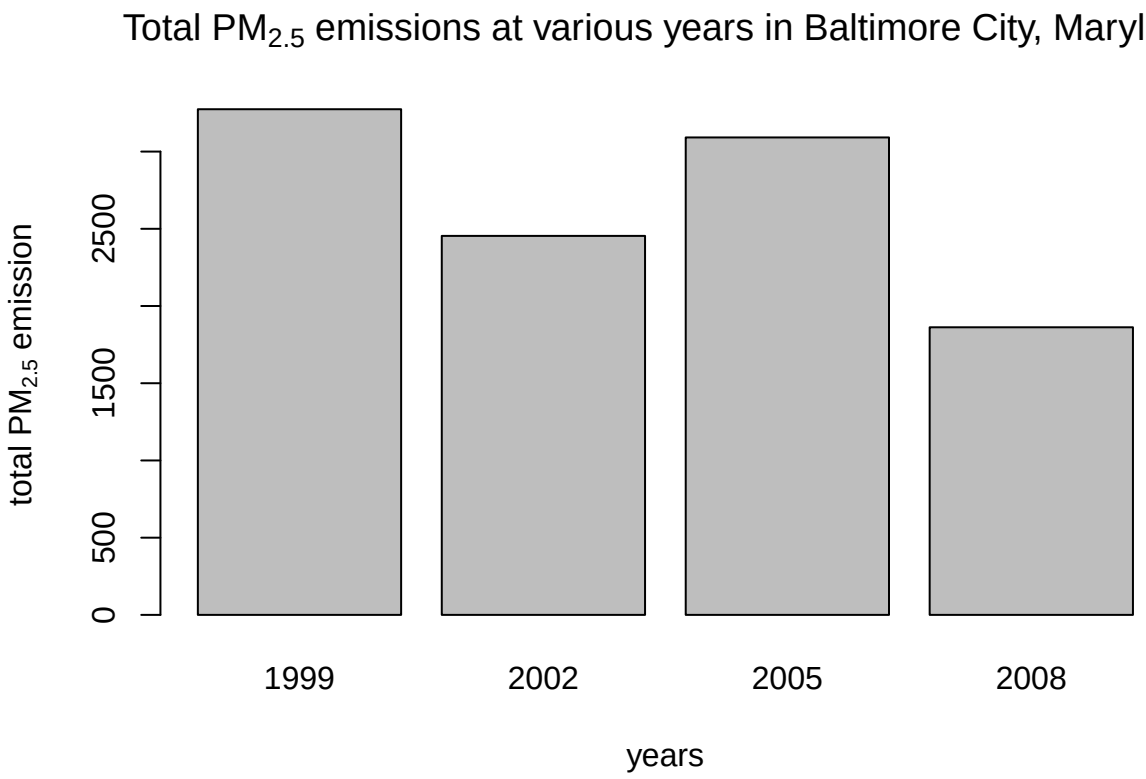
Question 2 - Have total emissions from $PM_{2.5}$ decreased in the Baltimore City, Maryland from 1999 to 2008?

```r
library(dplyr)

# Defining total emissions in the Baltimore City, Maryland per year
TotalEmissions <-
                  NEI %>%
                  filter(fips == 24510) %>%
                  group_by(year) %>%
                  summarise(SumEmissions = sum(Emissions))

# Plotting the result
```

```
barplot(height = TotalEmissions$SumEmissions,
                names.arg = TotalEmissions$year,
                xlab="years", ylab=expression('total PM'[2.5]*' emission'),
                main=expression('Total PM'[2.5]*' emissions at various years in Baltimore City, Maryla
        )
```



Answer: The data indicate a sharp decline between 1999 and 2002. A sharp increase occurred from 2002 to 2005. Finally, another sharp decrease occurred from 2005 to 2008.

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?

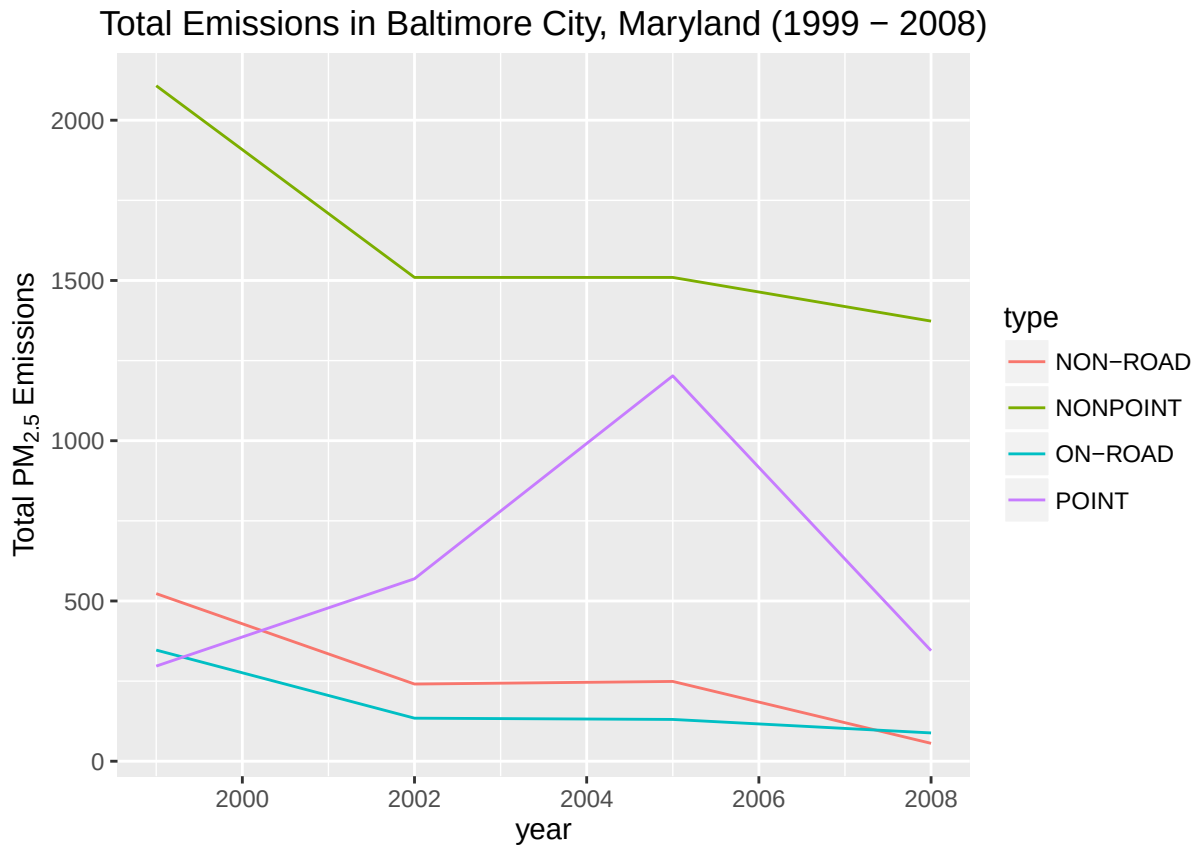
```
library(dplyr)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

```
# Defining total emissions in the Baltimore City, Maryland by types per year
TotalEmissions <-
  NEI %>%
  filter(fips == 24510) %>%
  group_by(year, type) %>%
```

```r
  summarise(SumEmissions = sum(Emissions))

# Plotting the result
g <- ggplot(TotalEmissions,
            aes(x = year, y = SumEmissions, color = type))
g + geom_line() +
  xlab("year") +
  ylab(expression('Total PM'[2.5]*" Emissions")) +
  ggtitle('Total Emissions in Baltimore City, Maryland (1999 - 2008)')
```



Answer:

- **Nonpoint (green line):** From the plot, we see that nonpoint (green line) sharply decreased from 1999 to 2002. It remained steady from 2002 to 2005 with 1,500 Total $PM_{2.5}$ emissions. Finally, a slight decrease occurred between 2005 and 2008 from 1,500 Total $PM_{2.5}$ emissions.
- **Point (purple line):** From the plot, we see that the point (purple line) slightly increased from 1999 to 2002. It then sharply increased in $PM_{2.5}$ emissions from 2002 to 2005. Finally, from 2005 to 2008, the $PM_{2.5}$ emissions sharply decreased.
- **Onroad (blue line):** From the plot, we see that the onroad (blue line) slightly decreased from 1999 to 2002. It remained approximately steady from 2002 to 2005 and continued this trend from 2005 to 2008. In comparison to the nonroad values, this over all trend was lower compared to the nonroad values.
- **Nonroad (red line):** From the plot, we see that the nonroad (red line) followed the same path as the onroad values only slightly higher in $PM_{2.5}$ emissions values. slightly decreased from 1999 to 2002. It remained approximately steady from 2002 to 2005 and continued this trend from 2005 to 2008.

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

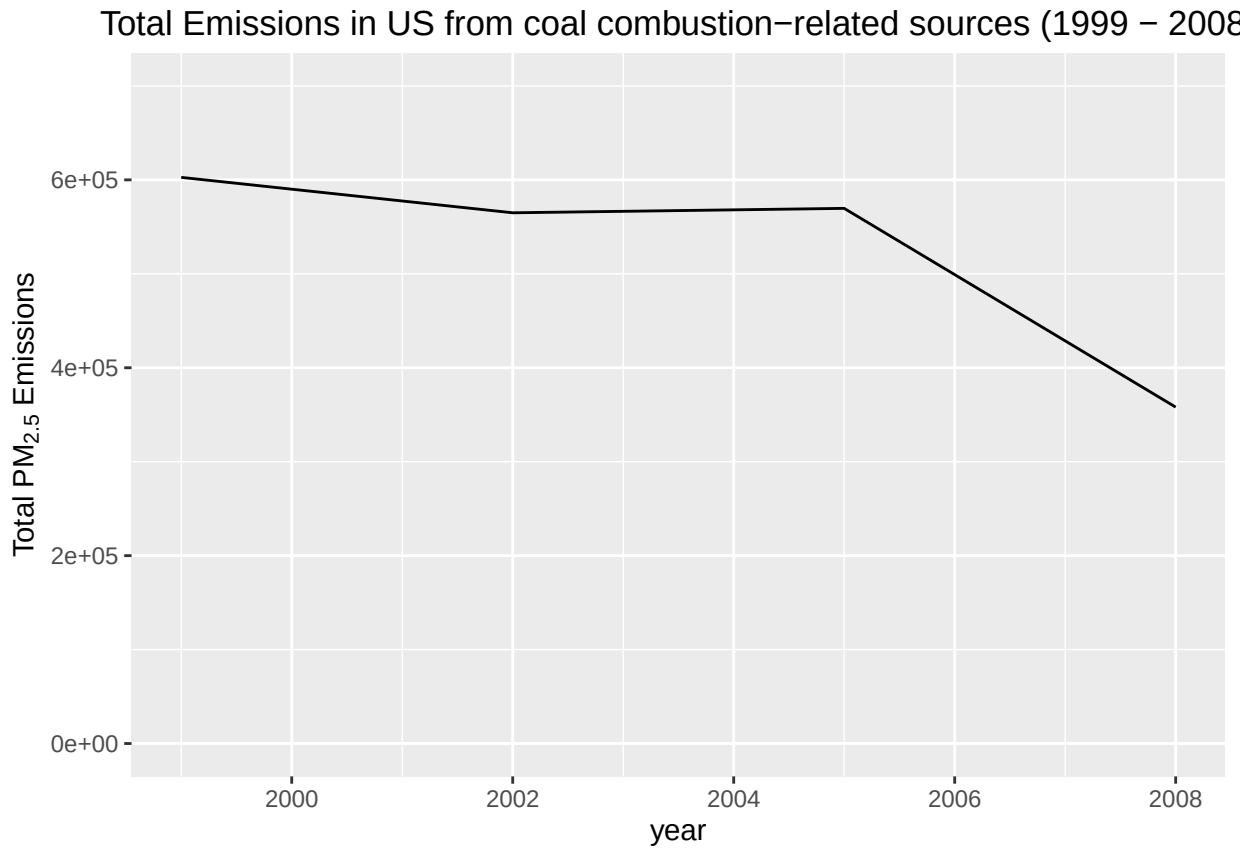
```
library(dplyr)
library(ggplot2)

# Merging datasets
mrg <- merge(NEI, SCC, by = "SCC")

# Defining total emissions from coal combustion-related sources per year
CoalMatches <- grepl("coal", mrg$Short.Name, ignore.case=TRUE)
mrgFiltered <- mrg[CoalMatches, ]

TotalEmissions <-
  mrgFiltered %>%
  group_by(year) %>%
  summarise(SumEmissions = sum(Emissions))

# Plotting the result
g <- ggplot(TotalEmissions,
            aes(x = year, y = SumEmissions))
g + geom_line() +
  xlab("year") +
  ylab(expression('Total PM'[2.5]*" Emissions")) +
  ggtitle('Total Emissions in US from coal combustion-related sources (1999 - 2008)') +
  ylim(0, 700000)
```



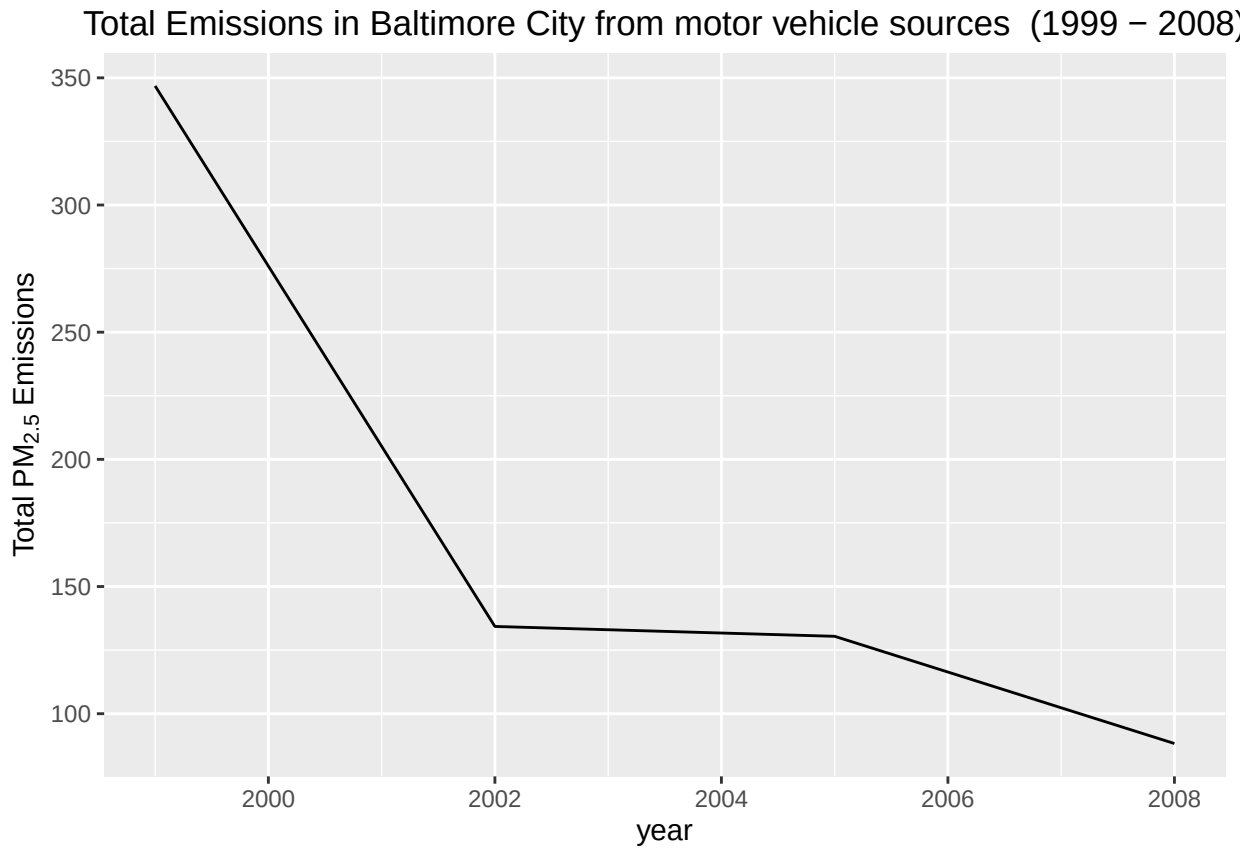
Answer: From the plot, we see that the purple line for total slightly declines from 1999 to 2002. From 2002 to 2005 the line has a marginal increase. Finally, from 2005 to 2008, the overall trend has a sharp decrease.

Question 5 - How have emissions from motor vehicle sources changed from 1999-2008 in Baltimore City?

```
library(dplyr)
library(ggplot2)

# Defining total emissions from motor vehicle sources per year in Baltimore City
TotalEmissions <-
  NEI %>%
  filter(fips == 24510, type == "ON-ROAD") %>%
  group_by(year) %>%
  summarise(SumEmissions = sum(Emissions))

# Plotting the result
g <- ggplot(TotalEmissions,
            aes(x = year, y = SumEmissions))
g + geom_line() +
  xlab("year") +
  ylab(expression('Total PM'[2.5]*" Emissions")) +
  ggtitle('Total Emissions in Baltimore City from motor vehicle sources  (1999 - 2008)')
```



Answer: Starting with 1999, the $PM_{2.5}$ emissions was just below 350, the levels fell sharply until 2002. From 2002 to 2005 the levels plateaued. Finally from 2005 to 2008, the $PM_{2.5}$ emissions drop to below 100 $PM_{2.5}$ emissions

Question 6 - Compare emissions from motor vehicle sources in Baltimore City with emissions from motor vehicle sources in Los Angeles County, California. Which city has seen greater changes over time in motor vehicle emissions?

```
library(dplyr)

# Defining total emissions from motor vehicle sources per year in Baltimore City
TotalEmissionsBaltimore <-
  NEI %>%
  filter(fips == "24510", type == "ON-ROAD") %>%
  group_by(year) %>%
  summarise(SumEmissions = sum(Emissions))

# Defining total emissions from motor vehicle sources per year in Los Angeles County, California
TotalEmissionsLosAngeles <-
  NEI %>%
  filter(fips == "06037", type == "ON-ROAD") %>%
  group_by(year) %>%
  summarise(SumEmissions = sum(Emissions))
```

```r
# Plotting results
rng <- c(0, 5000)

par(mfrow = c(1, 2))
barplot(height = TotalEmissionsBaltimore$SumEmissions,
        names.arg = TotalEmissionsBaltimore$year,
        xlab="years",
        ylab=expression('Emissions PM'[2.5]*' emission'),
        main=expression('Total emissions from motor vehicle sources
                        Baltimore City, Maryland'),
        ylim = rng
       )
```

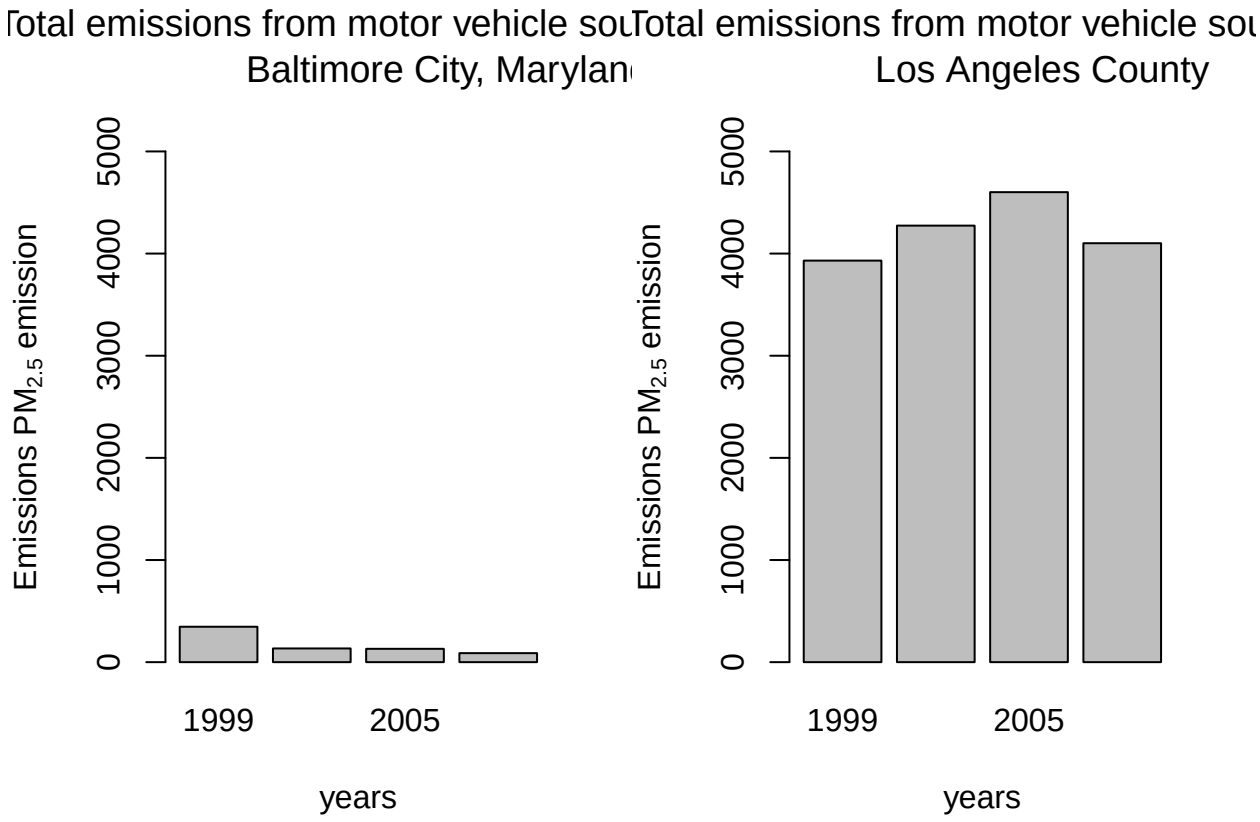
```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## font metrics unknown for character 0xa

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## font metrics unknown for character 0xa
```

```r
barplot(height = TotalEmissionsLosAngeles$SumEmissions,
        names.arg = TotalEmissionsLosAngeles$year,
        xlab="years",
        ylab=expression('Emissions PM'[2.5]*' emission'),
        main=expression('Total emissions from motor vehicle sources
                        Los Angeles County'),
        ylim = rng
       )
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## font metrics unknown for character 0xa

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## font metrics unknown for character 0xa
```



Answer:

- $PM_{2.5}$ emissions from motor vehicle sources in Baltimore City much more lower than $PM_{2.5}$ emissions from motor vehicle sources in Los Angeles County, California.
- **Baltimore, MD [city] (Left plot):** Emissions from motor vehicle sources starts marginally above zero and below 1,000 $PM_{2.5}$ emission values. Between 1999 and 2002, it slowly declines and remains nearly static between 2002 and 2008.
- **Los Angeles, CA [county] (Right plot):** Emissions from motor vehicle sources starts slightly below 4,000 $PM_{2.5}$ emisions and steadily increases to 2005. The value of $PM_{2.5}$ emissions for 2005 hits a peak at approximately 4,500 $PM_{2.5}$ emision levels and then decreases between 2005 and 2008 with an ending value point of slightly above 4,000 $PM_{2.5}$ emissions.