Exploratory Data Analysis Assignment 2

Myron Keith Gibert Jr

1/4/2021

Contents

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.

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 I used for this assignment are for 1999, 2002, 2005, and 2008.

*Dataset

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

Data for Peer Assessment [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.

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".

Set-up

Loading relevant packages.

```r
knitr::opts_chunk$set(echo = TRUE)

if (!require("tidyverse")) install.packages("tidyverse")
library("tidyverse")

if (!require("ggplot2")) install.packages("ggplot2")
library("ggplot2")

if (!require("ggthemes")) install.packages("ggthemes")
library("ggthemes")
```

Reading in Data

I used the following commands to read in the data. The chunk first checks for the presence of the dataset zip file. If it is not present, then the file is downloaded. Then, the chunk checks for the presence of the individual datasets. If any of them are not present in the working directory, the chunk will extract the original zip file. After this has been completed, I read the data in using the final two lines.

```r
if(!file.exists("exdata_data_NEI_data.zip")){
  download.file(
"https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip",
"exdata_data_NEI_data.zip", quiet = TRUE)}

if(!file.exists("summarySCC_PM25.rds")|!file.exists("Source_Classification_Code.rds")){
  unzip("exdata_data_NEI_data.zip")
}

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

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. I was allowed to use any R package I wanted to support my analysis.

For each plot, I did the following:

- 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.

Completing the Assignment

Now, we're ready to start answering some questions in the next few pages.

(CONTINUED ON NEXT PAGE)

Problem 1

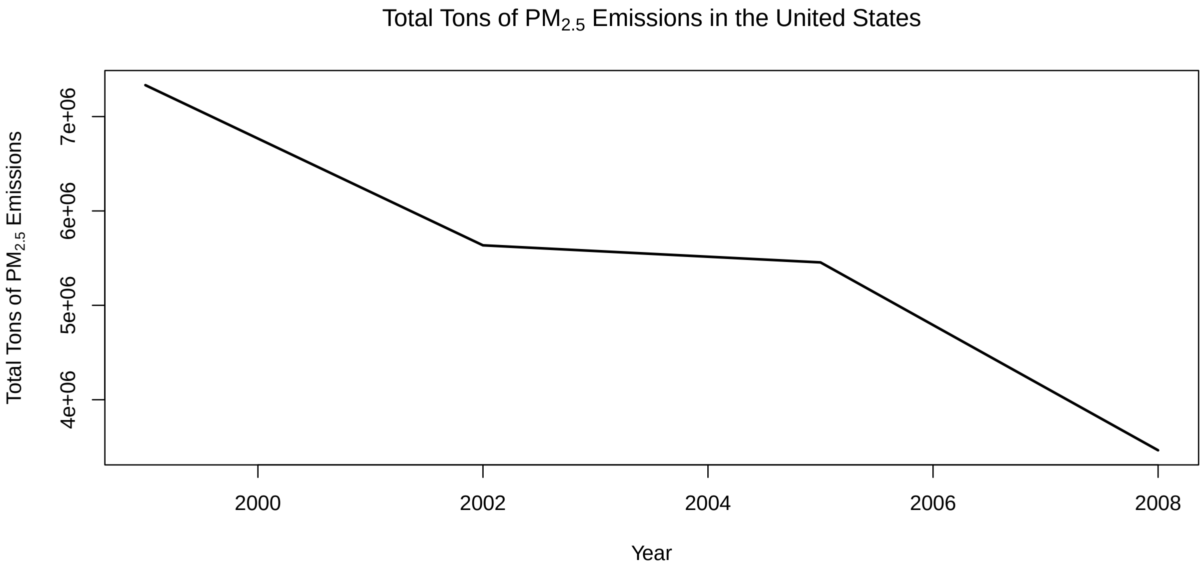
Have total emissions from PM2.5 decreased in the United States from 1999 to 2008? Using the base plotting system, make a plot showing the total PM2.5 emission from all sources for each of the years 1999, 2002, 2005, and 2008.

```
NEI_summary <- NEI %>%
  group_by(year) %>%
  summarise(Emissions=sum(Emissions))

png(file="plot1.png")

plot(NEI_summary$year,NEI_summary$Emissions, type = "l", lwd = 2,
     xlab = "Year",
     ylab = expression("Total Tons of PM"[2.5]*" Emissions"),
     main = expression("Total Tons of PM"[2.5]*" Emissions in the United States"))

dev.off()
```



Emissions have decreased in the United States from 1999 to 2008.

Problem 2

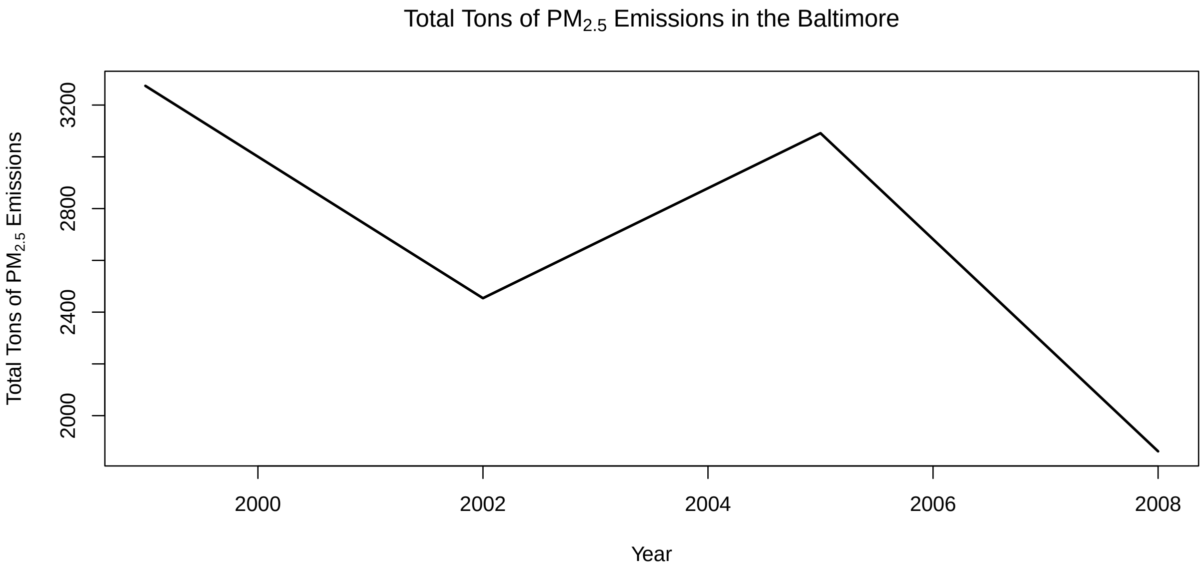
Have total emissions from PM2.5 decreased in the Baltimore City, Maryland (fips == "24510") from 1999 to 2008? Use the base plotting system to make a plot answering this question. Of the four types of sources indicated by the (point, nonpoint, onroad, nonroad) variable, which of these four sources have 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.

```
NEI_baltimore <- NEI %>%
  filter(fips == 24510) %>%
  group_by(year) %>%
  summarise(Emissions=sum(Emissions))

png(file="plot2.png")

plot(NEI_baltimore$year,NEI_baltimore$Emissions, type = "l", lwd = 2,
     xlab = "Year",
     ylab = expression("Total Tons of PM"[2.5]*" Emissions"),
     main = expression("Total Tons of PM"[2.5]*" Emissions in the Baltimore"))

dev.off()
```



Emissions have decreased in Baltimore City, Maryland from 1999 to 2008.

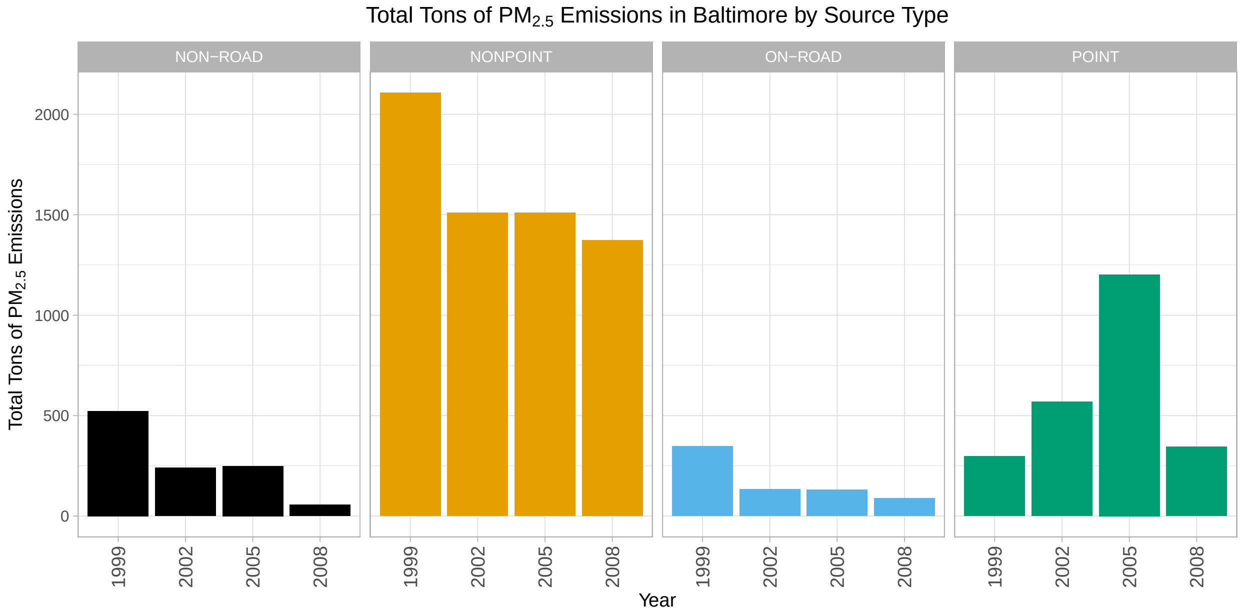
Problem 3

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.

```
NEI_baltimore_type <- NEI %>%
  filter(fips == 24510) %>%
  group_by(year,type) %>%
  summarise(Emissions=sum(Emissions))

p <- ggplot(NEI_baltimore_type,
            aes(x = factor(year), y = Emissions, fill = type)) +
  geom_bar(stat = "identity") +
  facet_grid(. ~ type) +
  xlab("Year") +
  ylab(expression("Total Tons of PM"[2.5]*" Emissions")) +
  ggtitle(expression("Total Tons of PM"[2.5]*" Emissions in Baltimore
                     by Source Type")) +
  theme_light() +
  scale_color_colorblind() +
  scale_fill_colorblind() +
  theme(axis.text.x=element_text(size=rel(1.2),angle = 90, vjust = 0.5, hjust = 1),
        plot.title = element_text(size=rel(1.2), face="bold",hjust = 0.5),
        legend.position = "none")

ggsave(file="plot3.png",plot = print(p), width = 10, height = 5, dpi = 300)
```



On-road, Off-road, and Nonpoint emissions decreased in Baltimore City from 1999-2008.

Problem 4

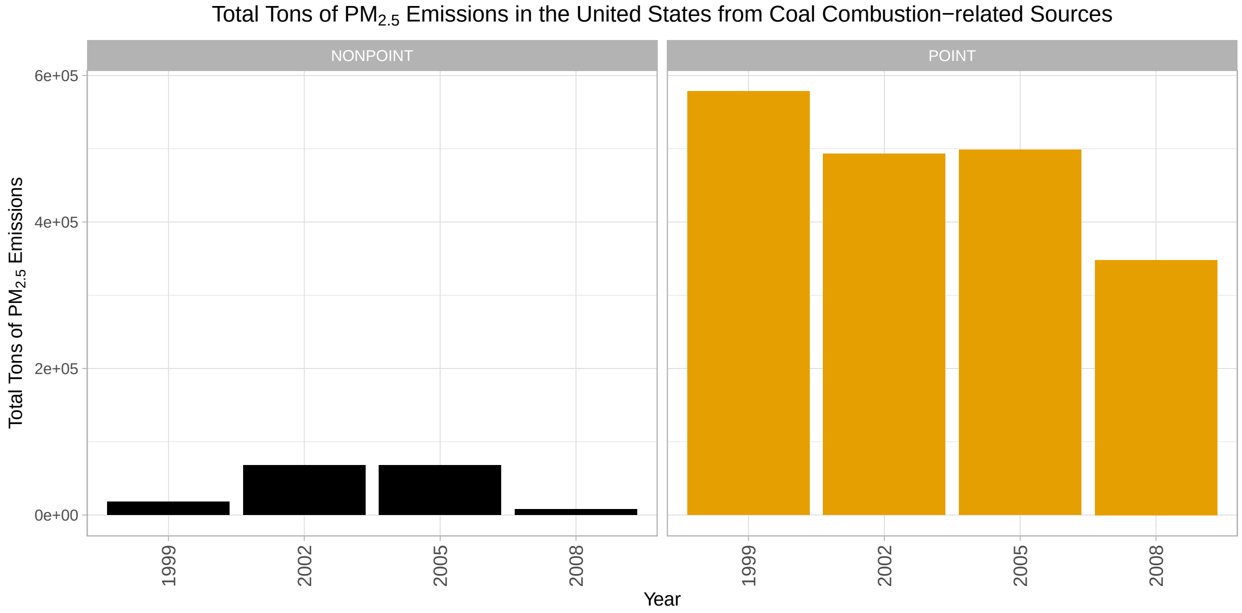
Across the United States, how have emissions from coal combustion-related sources changed from 1999–2008?

```
SCC_coal <- SCC[grep("Coal",SCC$Short.Name),]
SCC_coal_SCC <- unique(SCC_coal$SCC)
NEI_coal <- subset(NEI,SCC %in% SCC_coal_SCC)

NEI_coal_type <- NEI_coal %>%
  group_by(year,type) %>%
  summarise(Emissions=sum(Emissions))

p <- ggplot(NEI_coal_type, aes(x = factor(year), y = Emissions, fill = type)) +
  geom_bar(stat = "identity") +
  facet_grid(. ~ type) +
  xlab("Year") +
  ylab(expression("Total Tons of PM"[2.5]*" Emissions")) +
  ggtitle(expression("Total Tons of PM"[2.5]*" Emissions in the United States
                     from Coal Combustion-related Sources")) +
  theme_light() +
  scale_color_colorblind() +
  scale_fill_colorblind() +
  theme(axis.text.x=element_text(size=rel(1.2),angle = 90, vjust = 0.5, hjust = 1),
        plot.title = element_text(size=rel(1.2), face="bold",hjust = 0.5),
        legend.position = "none")

ggsave(file="plot4.png",plot = print(p), width = 10, height = 5, dpi = 300)
```



Point emissions in the United States have steadily decreased since 1999, while nonpoint emissions spiked between 1999-2005 before decreasing from 2005-2008.

Problem 5

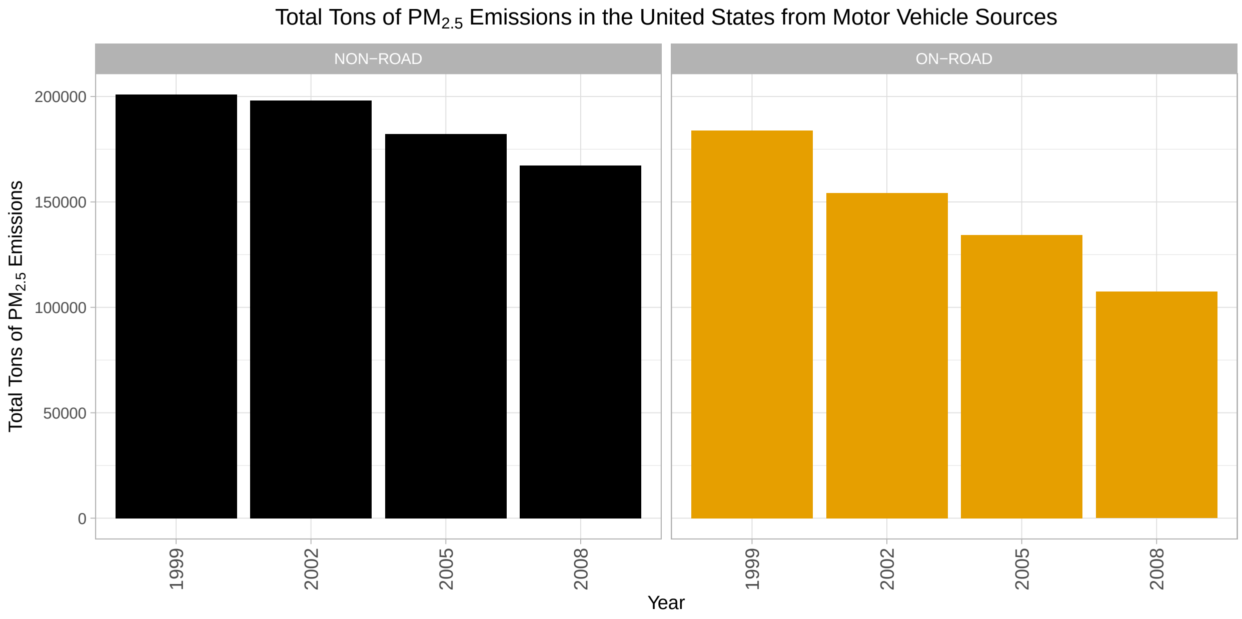
Compare emissions from motor vehicle sources in Baltimore City with emissions from motor vehicle sources in Los Angeles County, California (fips == “06037”). Which city has seen greater changes over time in motor vehicle emissions?

```
SCC_dmv <- SCC[grep("Vehicle",SCC$SCC.Level.Two),]
SCC_dmv_SCC <- unique(SCC_dmv$SCC)
NEI_dmv <- subset(NEI,SCC %in% SCC_dmv_SCC)

NEI_dmv_type <- NEI_dmv %>%
  filter(type=="ON-ROAD"|type=="NON-ROAD") %>%
  group_by(year,type) %>%
  summarise(Emissions=sum(Emissions))

p <- ggplot(NEI_dmv_type, aes(x = factor(year), y = Emissions, fill = type)) +
  geom_bar(stat = "identity") +
  facet_grid(. ~ type) +
  xlab("Year") +
  ylab(expression("Total Tons of PM"[2.5]*" Emissions")) +
  ggtitle(expression("Total Tons of PM"[2.5]*" Emissions in the United States
                     from Motor Vehicle Sources")) +
  theme_light() +
  scale_color_colorblind() +
  scale_fill_colorblind() +
  theme(axis.text.x=element_text(size=rel(1.2),angle = 90, vjust = 0.5, hjust = 1),
        plot.title = element_text(size=rel(1.2), face="bold",hjust = 0.5),
        legend.position = "none")

ggsave(file="plot5.png",plot = print(p), width = 10, height = 5, dpi = 300)
```



Both non-road and on-road motor vehicle emissions have decreased in the United States from 1999-2008.

Problem 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?

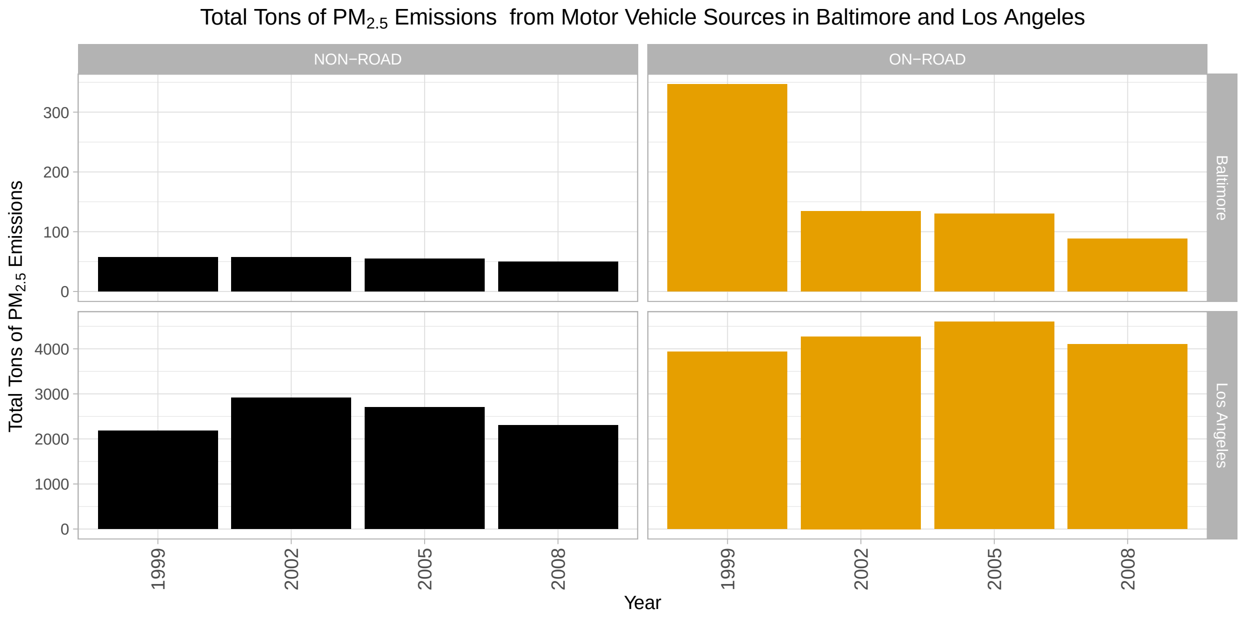
```r
SCC_dmv_2 <- SCC[grep("Vehicle",SCC$SCC.Level.Two),]
SCC_dmv_SCC_2 <- unique(SCC_dmv_2$SCC)
NEI_dmv_2 <- subset(NEI,SCC %in% SCC_dmv_SCC_2)

NEI_dmv_type_2 <- NEI_dmv_2 %>%
  filter(fips == "24510" | fips == "06037") %>%
  filter(type=="ON-ROAD"|type=="NON-ROAD") %>%
  group_by(year,type,fips) %>%
  summarise(Emissions=sum(Emissions))

NEI_dmv_type_2$fips[NEI_dmv_type_2$fips == "24510"] <- "Baltimore"
NEI_dmv_type_2$fips[NEI_dmv_type_2$fips == "06037"] <- "Los Angeles"


p <- ggplot(NEI_dmv_type_2, aes(x = factor(year), y = Emissions, fill = type)) +
  geom_bar(stat = "identity") +
  facet_grid(fips ~ type, scales = "free") +
  xlab("Year") +
  ylab(expression("Total Tons of PM"[2.5]*" Emissions")) +
  ggtitle(expression("Total Tons of PM"[2.5]*" Emissions  from Motor Vehicle
                     Sources in Baltimore and Los Angeles")) +
  theme_light() +
  scale_color_colorblind() +
  scale_fill_colorblind() +
  theme(axis.text.x=element_text(size=rel(1.2),angle = 90, vjust = 0.5, hjust = 1),
        plot.title = element_text(size=rel(1.2), face="bold",hjust = 0.5),
        legend.position = "none")

ggsave(file="plot6.png",plot = print(p), width = 10, height = 5, dpi = 300)
```



Both non-road and on-road motor vehicle emissions have decreased from 1999-2008 in Baltimore, but not in Los Angeles.

Clean-up

Removes unnecessary files after analysis is completed.

```r
if(file.exists("exdata_data_NEI_data.zip")){
  unlink("exdata_data_NEI_data.zip")
}

if(file.exists("summarySCC_PM25.rds")){
  unlink("summarySCC_PM25.rds")
}

if(file.exists("Source_Classification_Code.rds")){
  unlink("Source_Classification_Code.rds")
}
```