Coursera: Data Science: Exploratory Data Analysis: Project 2

John W. Tiede

09/06/2014

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 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 [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).

```
## Reading emission sources file

## Reading emission data file

## [1] "System RAM                           : 4096.00 MBi"

## [1] "Emission Sources Memory Requirements:    1.34 MBi"

## [1] "Emission Data Memory Requirements:     297.44 MBi"
```

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.

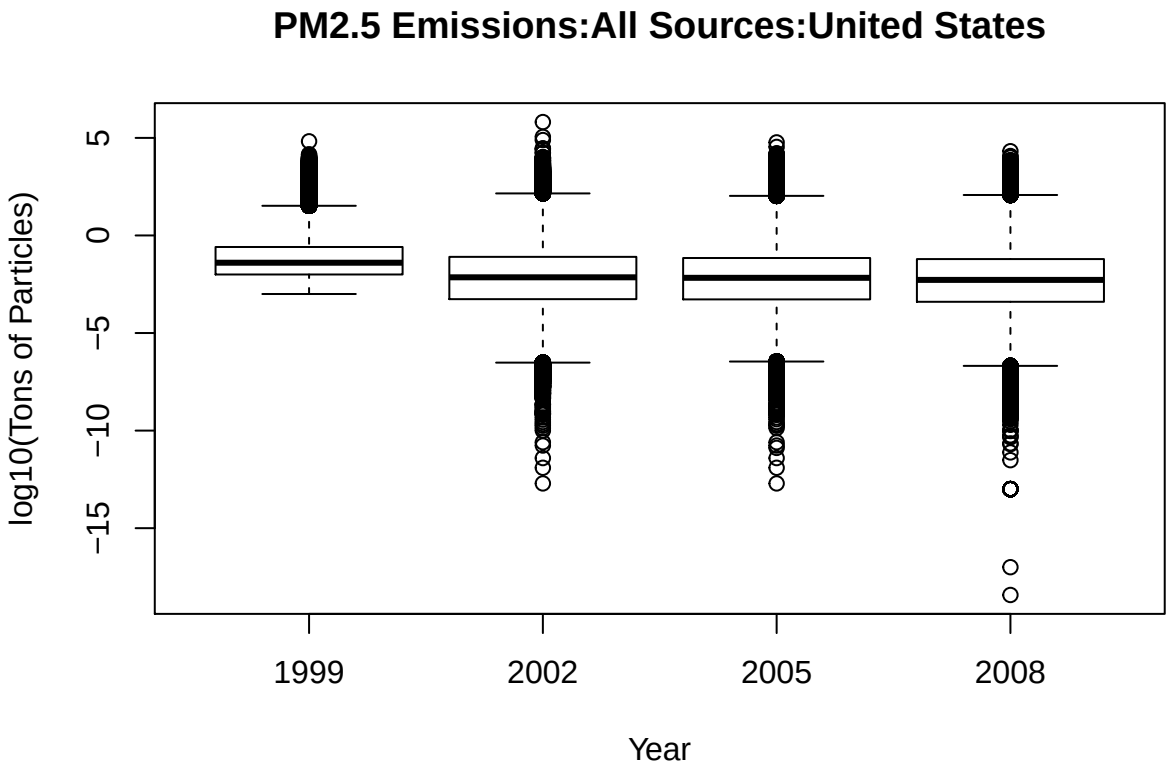
Questions

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.

```
# If conditional is TRUE, plot graph to window system, else plot to file:
if (TRUE) {
  # Create the graph in the 'base' package:
  boxplot(log10(Emissions) ~ year, dat=edata,
          main="PM2.5 Emissions:All Sources:United States",
          xlab="Year", ylab="log10(Tons of Particles)")
} else {
  # Create the 'png' file:
  png(filename=paste0(working.directory, "/", "Question01.png"))
  boxplot(log10(Emissions) ~ year, dat=edata,
          main="PM2.5 Emissions:All Sources:United States",
          xlab="Year", ylab="log10(Tons of Particles)")
  dev.off()
}
```

```
## Warning: Outlier (-Inf) in boxplot 1 is not drawn
## Warning: Outlier (-Inf) in boxplot 2 is not drawn
## Warning: Outlier (-Inf) in boxplot 3 is not drawn
## Warning: Outlier (-Inf) in boxplot 4 is not drawn
```

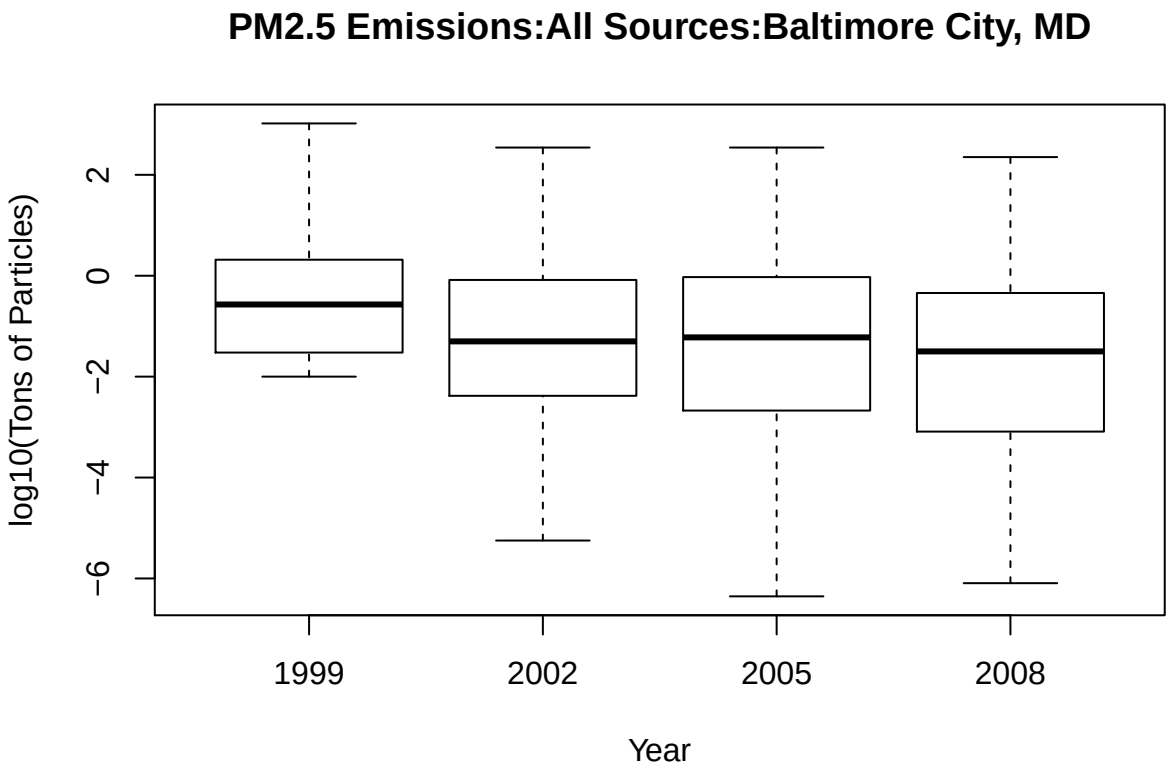


2. Have total emissions from PM2.5 decreased in the **Baltimore City**, Maryland (fips == "24510") from 1999 to 2008? Use the base plotting system to make a plot answering this question.

```
# Subset edata for Baltimore City, Maryland:
edata.24510 <- edata[edata$fips == "24510", ]
# Make "fips" column values to be factors.
# Make "type" column values to be factors.
# Make "year" column values to be factors.
edata.24510$fips <- as.factor(edata.24510$fips)
edata.24510$type <- as.factor(edata.24510$type)
edata.24510$year <- as.factor(edata.24510$year)
#str(edata.24510)

# If conditional is TRUE, plot graph to window system, else plot to file:
if (TRUE) {
  # Create the graph in the 'base' package:
  boxplot(log10(Emissions) ~ year, dat=edata.24510,
          main="PM2.5 Emissions:All Sources:Baltimore City, MD",
          xlab="Year", ylab="log10(Tons of Particles)")
} else {
  # Create the 'png' file:
  png(filename=paste0(working.directory, "/", "Question02.png"))
  boxplot(log10(Emissions) ~ year, dat=edata.24510,
          main="PM2.5 Emissions:All Sources:Baltimore City, MD",
```

```
          xlab="Year", ylab="log10(Tons of Particles)")
  dev.off()
}
```

```
## Warning: Outlier (-Inf) in boxplot 1 is not drawn
## Warning: Outlier (-Inf) in boxplot 2 is not drawn
## Warning: Outlier (-Inf) in boxplot 3 is not drawn
## Warning: Outlier (-Inf) in boxplot 4 is not drawn
```

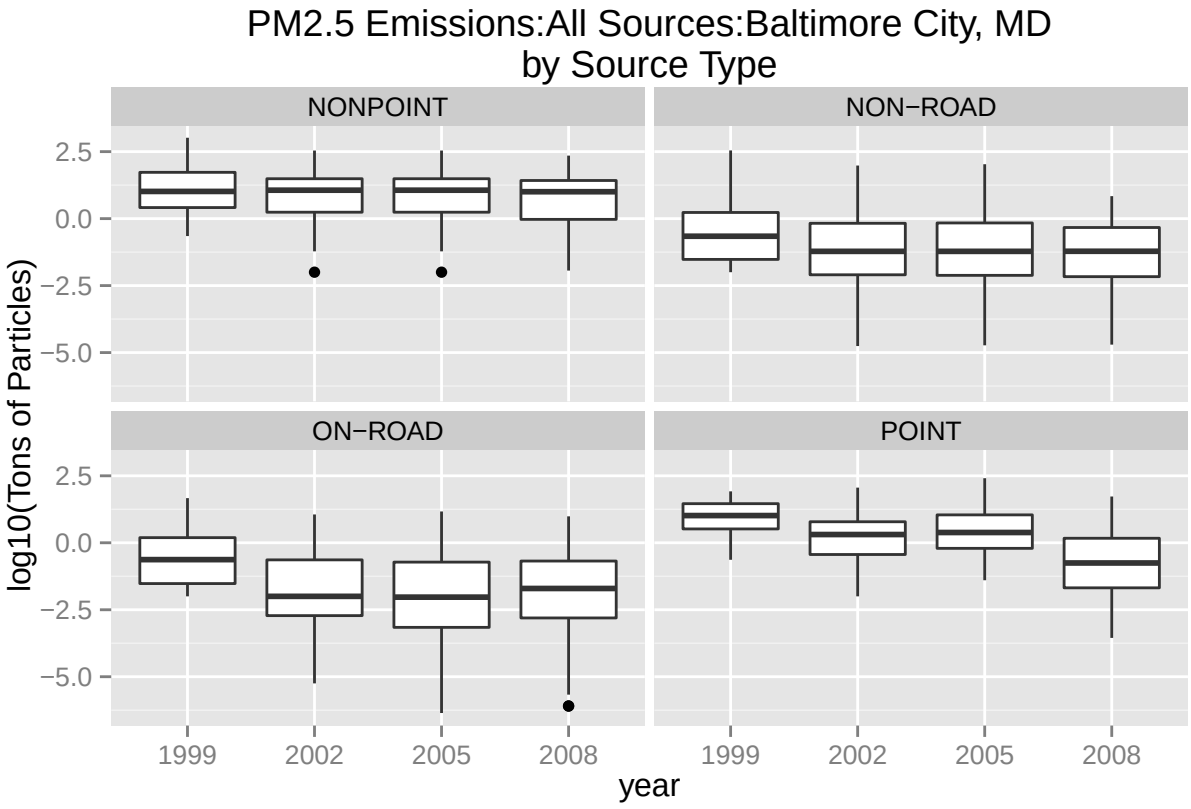


3. Of the four types of sources indicated by the type (point, nonpoint, onroad, nonroad) variable, which of these four sources have seen decreases in emissions from 1999–2008 for **Baltimore City**? Which have seen increases in emissions from 1999–2008? Use the ggplot2 plotting system to make a plot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

```
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(plyr))
suppressPackageStartupMessages(library(reshape2))

# Because we are using "ggplot2", make data long & tidy (one observation per row):
edata.24510.melt <- edata.24510[, c("fips", "SCC", "year", "type", "Emissions")]
edata.24510.melt <- melt(edata.24510, measure.vars="Emissions")

# If conditional is TRUE, plot graph to window system, else plot to file:
if (TRUE) {
  q3.plot <- ggplot(edata.24510.melt, aes(x=year, y=log10(value), group=year))
  q3.plot <- q3.plot + geom_boxplot()
  q3.plot <- q3.plot + facet_wrap(~ type)
```

```
  q3.plot <- q3.plot + ggtitle("PM2.5 Emissions:All Sources:Baltimore City, MD\nby Source Type")
  q3.plot <- q3.plot + ylab("log10(Tons of Particles)")
  q3.plot
} else {
  # Create the 'png' file:
  png(filename=paste0(working.directory, "/", "Question03.png"))
  q3.plot <- ggplot(edata.24510.melt, aes(x=year, y=log10(value), group=year))
  q3.plot <- q3.plot + geom_boxplot()
  q3.plot <- q3.plot + facet_wrap(~ type)
  q3.plot <- q3.plot + ggtitle("PM2.5 Emissions:All Sources:Baltimore City, MD\nby Source Type")
  q3.plot <- q3.plot + ylab("log10(Tons of Particles)")
  q3.plot
  dev.off()
}
```

```
## Warning: Removed 18 rows containing non-finite values (stat_boxplot).
## Warning: Removed 19 rows containing non-finite values (stat_boxplot).
## Warning: Removed 91 rows containing non-finite values (stat_boxplot).
```

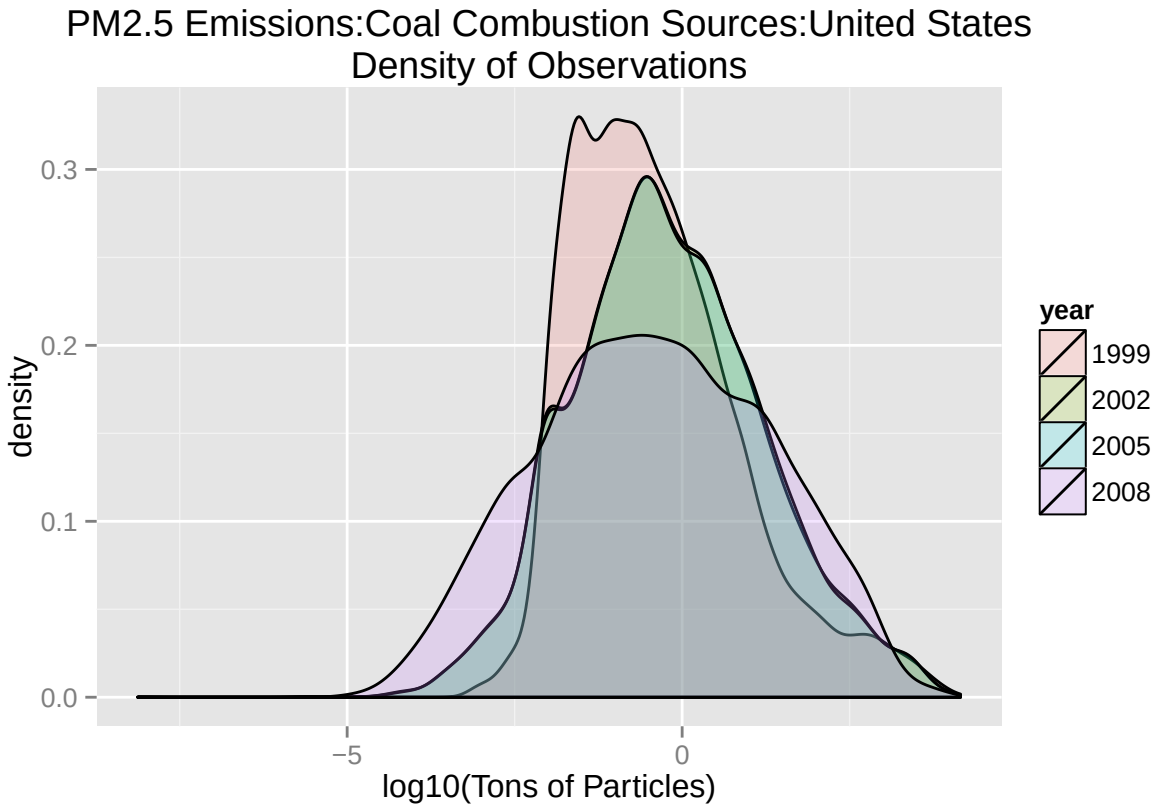


4. Across the **United States**, how have emissions from coal combustion-related sources changed from 1999–2008?

```
# Subset Emission Sources to get "coal" Source Classification Codes (SCC):
esrcs.coal <- as.character(esrcs$SCC[grepl("coal$", as.character(esrcs$SCC.Level.Three), ignore.case=TRU
#class(esrcs.coal)
#length(esrcs.coal)
```

```r
# Subset Emission Data to get only coal sources:
# Reference:
#   http://stackoverflow.com/questions/15227887/how-can-i-subset-rows-in-a-data-frame-in-r-based-on-a-v
edata.coal <- edata[edata$SCC %in% esrcs.coal, c("fips", "year", "type", "Emissions")]
edata.coal.melt <- melt(edata.coal, measure.vars="Emissions")
edata.coal.melt$fips <- as.factor(edata.coal.melt$fips)
edata.coal.melt$type <- as.factor(edata.coal.melt$type)
edata.coal.melt$year <- as.factor(edata.coal.melt$year)

# Testing (to be removed):
#dim(edata.coal)
#summary(edata.coal)
#summary(edata.coal[edata.coal$year == 1999, "Emissions"])
#sum(edata.coal$year == 1999)
#summary(edata.coal[edata.coal$year == 2002, "Emissions"])
#sum(edata.coal$year == 2002)
#boxplot(log10(edata.coal[edata.coal$year == 2002, "Emissions"]))
#hist(edata.coal[edata.coal$year == 2002, "Emissions"], breaks=20)
#plot(density(log10(edata.coal[edata.coal$year == 2002, "Emissions"])))
#summary(edata.coal[edata.coal$year == 2005, "Emissions"])
#sum(edata.coal$year == 2005)
#summary(edata.coal[edata.coal$year == 2008, "Emissions"])
#sum(edata.coal$year == 2008)

# If conditional is TRUE, plot graph to window system, else plot to file:
# Reference:
#   http://stackoverflow.com/questions/21563864/ggplot2-overlay-density-plots-r
#   http://stackoverflow.com/questions/12944357/overlay-10-density-plots-in-r-with-colour-proportional-
if (TRUE) {
  # Create the graph in the 'ggplot2' package:
  q4.plot <- ggplot(edata.coal.melt, aes(x=log10(value), fill=year, group=year))
  q4.plot <- q4.plot + geom_density(alpha=0.20)
  q4.plot <- q4.plot + ggtitle("PM2.5 Emissions:Coal Combustion Sources:United States\nDensity of Obser
  q4.plot <- q4.plot + xlab("log10(Tons of Particles)")
  q4.plot
} else {
  # Create the 'png' file:
  png(filename=paste0(working.directory, "/", "Question04.png"))
  q4.plot <- ggplot(edata.coal.melt, aes(x=log10(value), fill=year, group=year))
  q4.plot <- q4.plot + geom_density(alpha=0.20)
  q4.plot <- q4.plot + ggtitle("PM2.5 Emissions:Coal Combustion Sources:United States\nDensity of Obser
  q4.plot <- q4.plot + xlab("log10(Tons of Particles)")
  q4.plot
  dev.off()
}
```

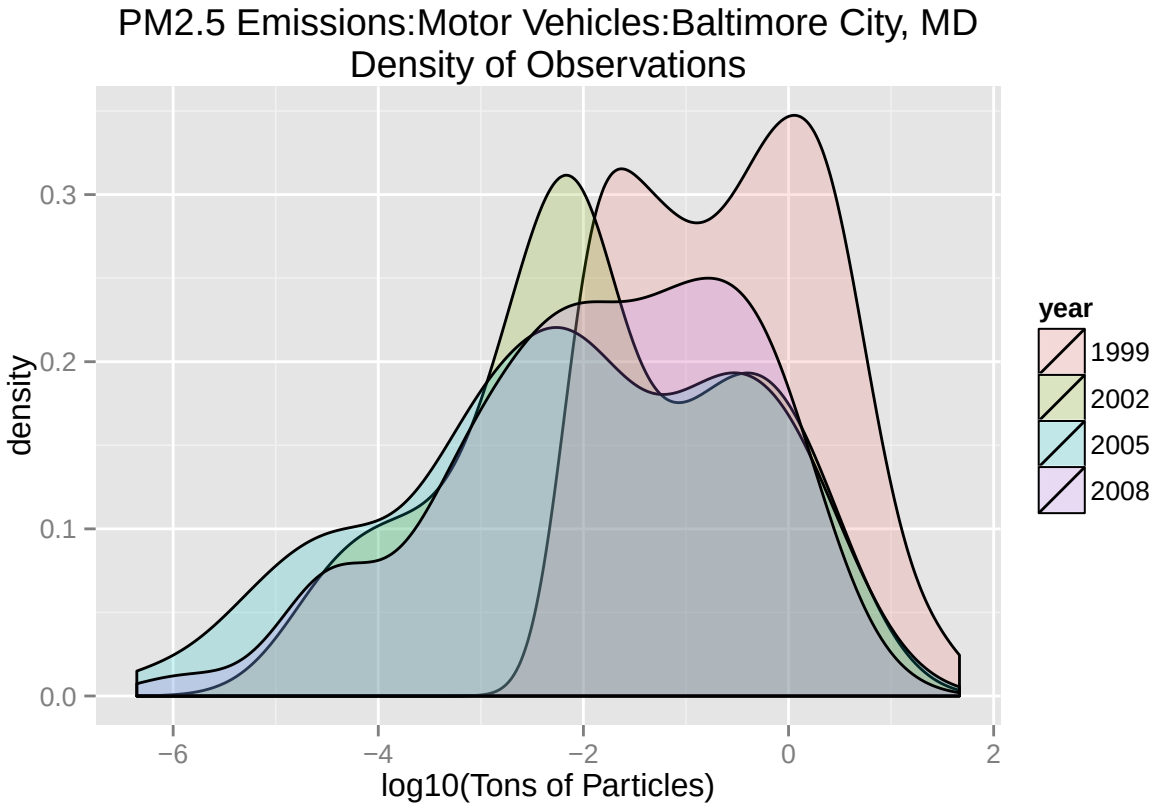
```
## Warning: Removed 34 rows containing non-finite values (stat_density).
## Warning: Removed 2842 rows containing non-finite values (stat_density).
## Warning: Removed 2828 rows containing non-finite values (stat_density).
## Warning: Removed 6025 rows containing non-finite values (stat_density).
```



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

```
# Subset Emission Sources to get "motor vehicle" Source Classification Codes (SCC):
esrcs.mv <- as.character(esrcs$SCC[grepl(".*Mobile - On-Road.*",
                                         as.character(esrcs$EI.Sector),
                                         ignore.case=TRUE, perl=TRUE)])

# Subset Emission Data to get only motor vehicles & Baltimore City sources:
edata.24510.mv <- edata.24510[edata.24510$SCC %in% esrcs.mv, c("fips", "SCC", "year", "type", "Emissions
#summary(edata.24510.mv[edata.24510.mv$year == 1999, "Emissions"])
#summary(edata.24510.mv[edata.24510.mv$year == 2002, "Emissions"])
#summary(edata.24510.mv[edata.24510.mv$year == 2005, "Emissions"])
#summary(edata.24510.mv[edata.24510.mv$year == 2008, "Emissions"])
edata.24510.mv.melt <- melt(edata.24510.mv, measure.vars="Emissions")

# If conditional is TRUE, plot graph to window system, else plot to file:
if (TRUE) {
  # Create the graph in the 'ggplot2' package:
  q5.plot <- ggplot(edata.24510.mv.melt, aes(x=log10(value), fill=year, group=year))
  q5.plot <- q5.plot + geom_density(alpha=0.20)
  q5.plot <- q5.plot + ggtitle("PM2.5 Emissions:Motor Vehicles:Baltimore City, MD\nDensity of Observatio
  q5.plot <- q5.plot + xlab("log10(Tons of Particles)")
  q5.plot
} else {
  # Create the 'png' file:
  png(filename=paste0(working.directory, "/", "Question05.png"))
  q5.plot <- ggplot(edata.24510.mv.melt, aes(x=log10(value), fill=year, group=year))
```

```
  q5.plot <- q5.plot + geom_density(alpha=0.25)
  q5.plot <- q5.plot + ggtitle("PM2.5 Emissions:Motor Vehicles:Baltimore City, MD\nDensity of Observati
  q5.plot <- q5.plot + xlab("log10(Tons of Particles)")
  q5.plot
  dev.off()
}
```

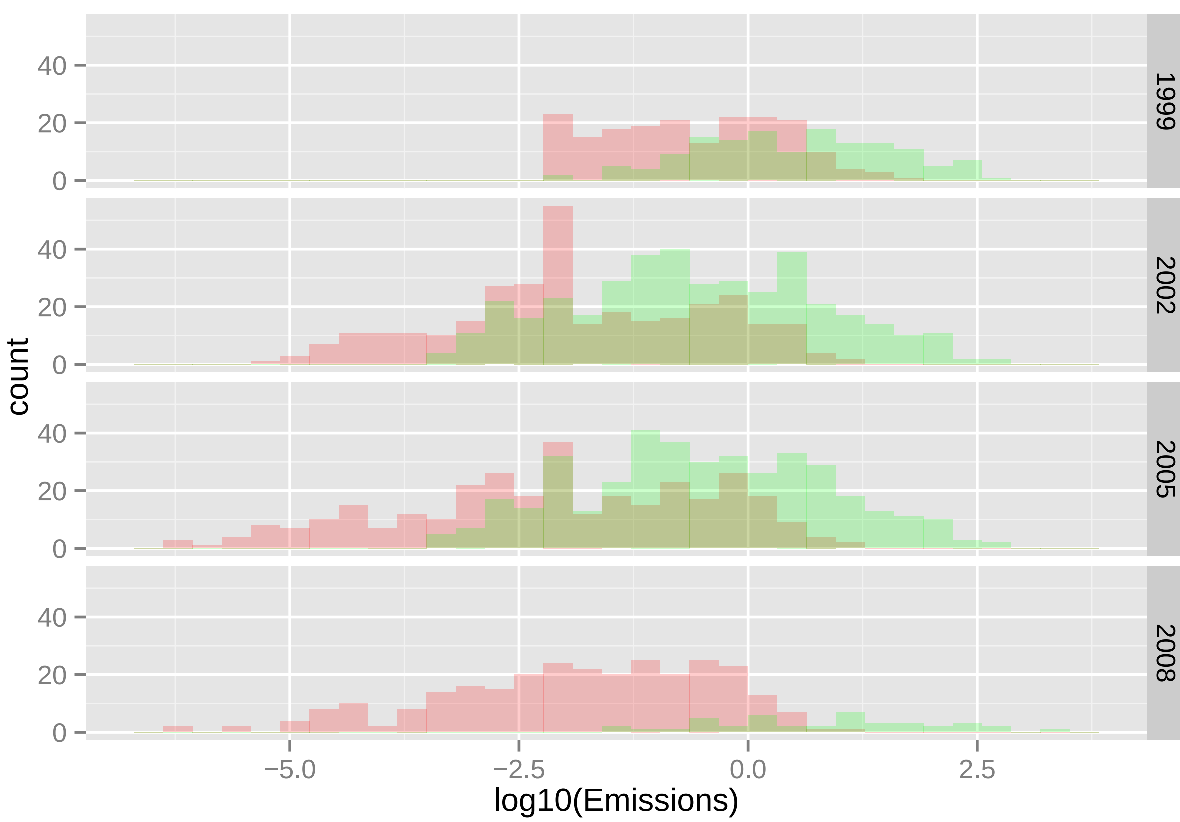


6. Compare emissions from motor vehicle sources in Baltimore City with emissions from motor vehicle sources in **Los Angeles County, California** (fips == "06037"). Which city has seen greater changes over time in motor vehicle emissions?

```
# Subset edata for Los Angeles County, California:
edata.06037 <- edata[edata$fips == "06037", ]
edata.06037$fips <- as.factor(edata.06037$fips)
edata.06037$type <- as.factor(edata.06037$type)
edata.06037$year <- as.factor(edata.06037$year)
#unique(edata.06037$fips)

# Subset Emission Data to get only motor vehicles & Los Angeles sources:
edata.06037.mv <- edata.06037[edata.06037$SCC %in% esrcs.mv, c("fips", "SCC", "year", "type", "Emissions

# Combine Los Angeles, CA with Baltimore City, MD in a data.frame & melt:
edata.both.mv <- rbind(edata.24510.mv, edata.06037.mv)
# Reference:
#   http://stackoverflow.com/questions/3472980/ggplot-how-to-change-facet-labels
# First check what the levels are:
#levels(edata.both.mv.melt$fips)
```

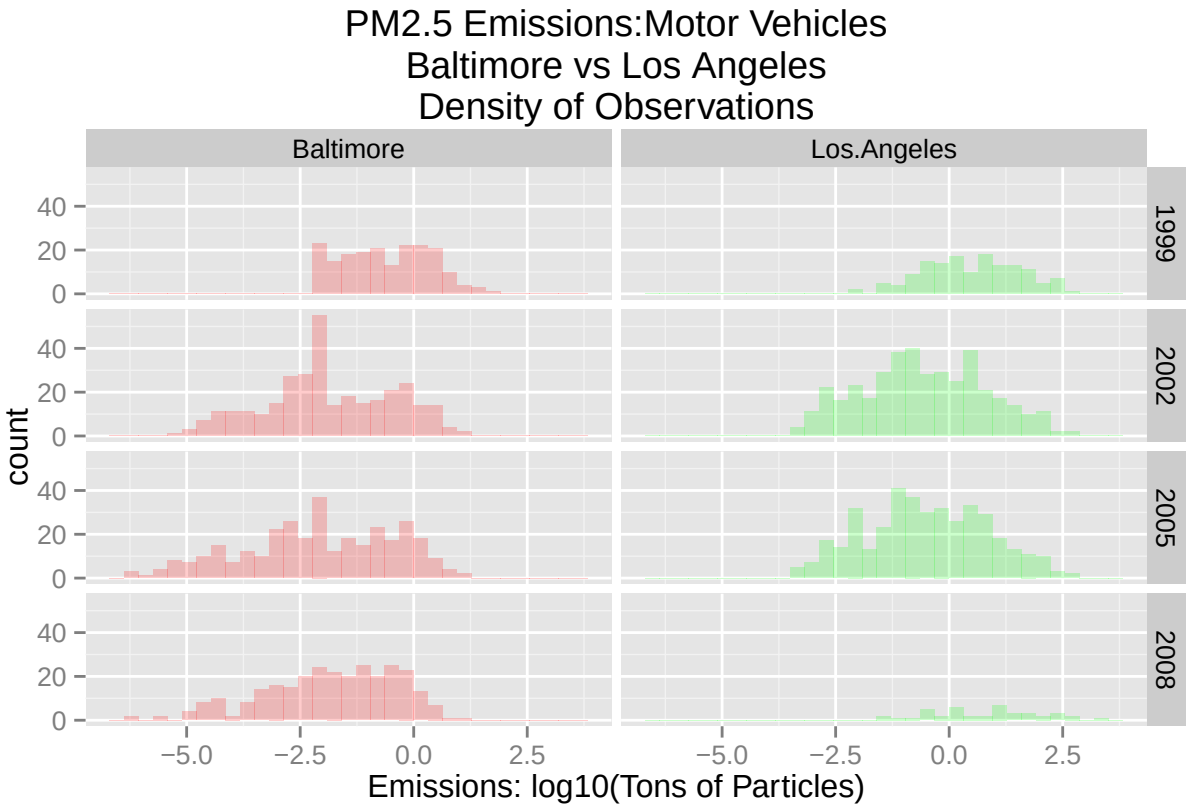
```r
# Then replace with good names for the plot:
levels(edata.both.mv$fips) <- c("Baltimore", "Los.Angeles")
edata.both.mv.melt <- melt(edata.both.mv, measure.vars="Emissions")
#unique(edata.both.mv.melt$fips)
#summary(edata.both.mv.melt)

# Incremental Testing (to be removed):
#ggplot(edata.both.mv, aes(x=log10(Emissions), group=year, fill=year)) +
#  geom_histogram(alpha=0.20)
#ggplot(edata.both.mv, aes(x=log10(Emissions), group=year, fill=year)) +
#  geom_histogram(alpha=0.20) +
#  facet_grid(year ~ .)
ggplot(edata.both.mv.melt, aes(x=log10(Emissions), group=year)) +
  geom_histogram(data=subset(edata.both.mv, fips == "Baltimore"), fill="red", alpha=0.20) +
  geom_histogram(data=subset(edata.both.mv, fips == "Los.Angeles"), fill="green", alpha=0.20) +
  facet_grid(year ~ .)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```r
# Reference:
#   http://stackoverflow.com/questions/6957549/overlaying-histograms-with-ggplot2-in-r
#
# If conditional is TRUE, plot graph to window system, else plot to file:
if (TRUE) {
  # Create the graph in the 'ggplot2' package:
  q6.plot <- ggplot(edata.both.mv.melt, aes(x=log10(Emissions), group=year))
  q6.plot <- q6.plot + geom_histogram(data=subset(edata.both.mv, fips == "Baltimore"), fill="red", alpha
  q6.plot <- q6.plot + geom_histogram(data=subset(edata.both.mv, fips == "Los.Angeles"), fill="green", a
  q6.plot <- q6.plot + facet_grid(year ~ fips)
  q6.plot <- q6.plot + ggtitle("PM2.5 Emissions:Motor Vehicles\nBaltimore vs Los Angeles\nDensity of Obs
  q6.plot <- q6.plot + xlab("Emissions: log10(Tons of Particles)")
  q6.plot
} else {
  # Create the 'png' file:
  png(filename=paste0(working.directory, "/", "Question06.png"))
  q6.plot <- ggplot(edata.both.mv.melt, aes(x=log10(Emissions), group=year))
  q6.plot <- q6.plot + geom_histogram(data=subset(edata.both.mv, fips == "Baltimore"), fill="red", alpha
  q6.plot <- q6.plot + geom_histogram(data=subset(edata.both.mv, fips == "Los.Angeles"), fill="green", a
  q6.plot <- q6.plot + facet_grid(year ~ fips)
  q6.plot <- q6.plot + ggtitle("PM2.5 Emissions:Motor Vehicles\nBaltimore vs Los Angeles\nDensity of Obs
  q6.plot <- q6.plot + xlab("Emissions: log10(Tons of Particles)")
  q6.plot
  dev.off()
}
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```


PM2.5 Emissions:Motor Vehicles
Baltimore vs Los Angeles
Density of Observations
Baltimore
Los.Angeles
1999
2002
2005
2008
count
0
20
40
−5.0
−2.5
0.0
2.5
Emissions: log10(Tons of Particles)