Exploratory Data Analysis Project 2

Halley Wang

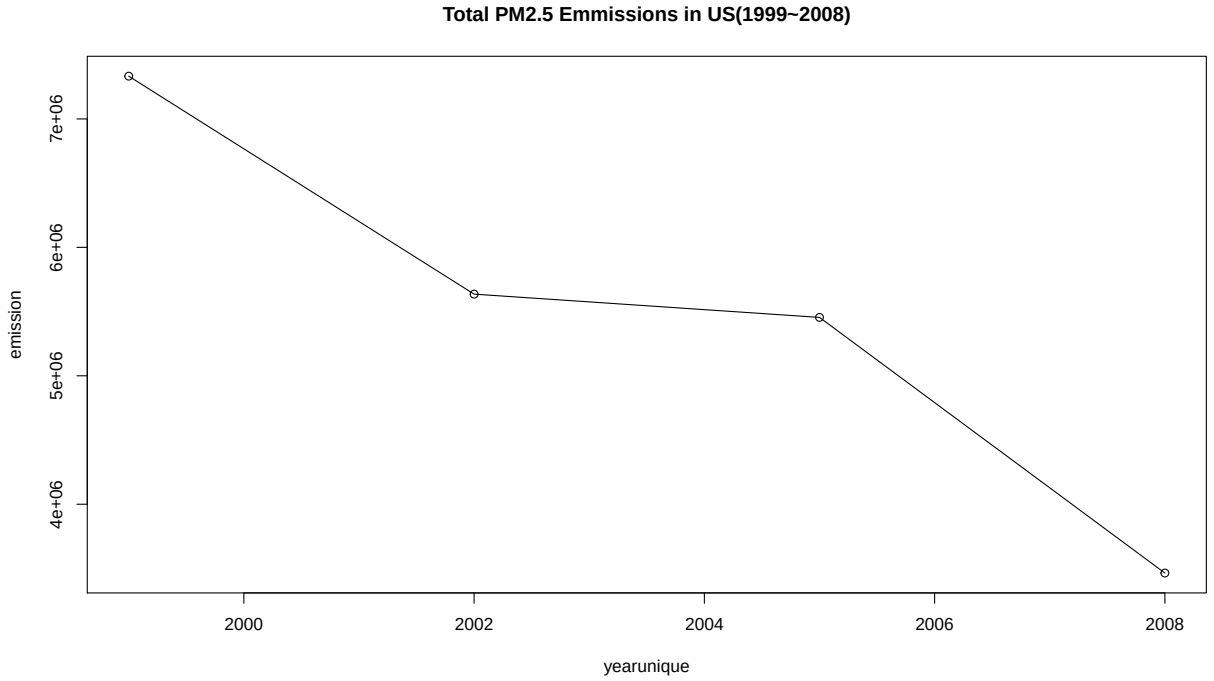
September 4, 2015

```
## Loading the necessary packages
library(ggplot2)
```

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

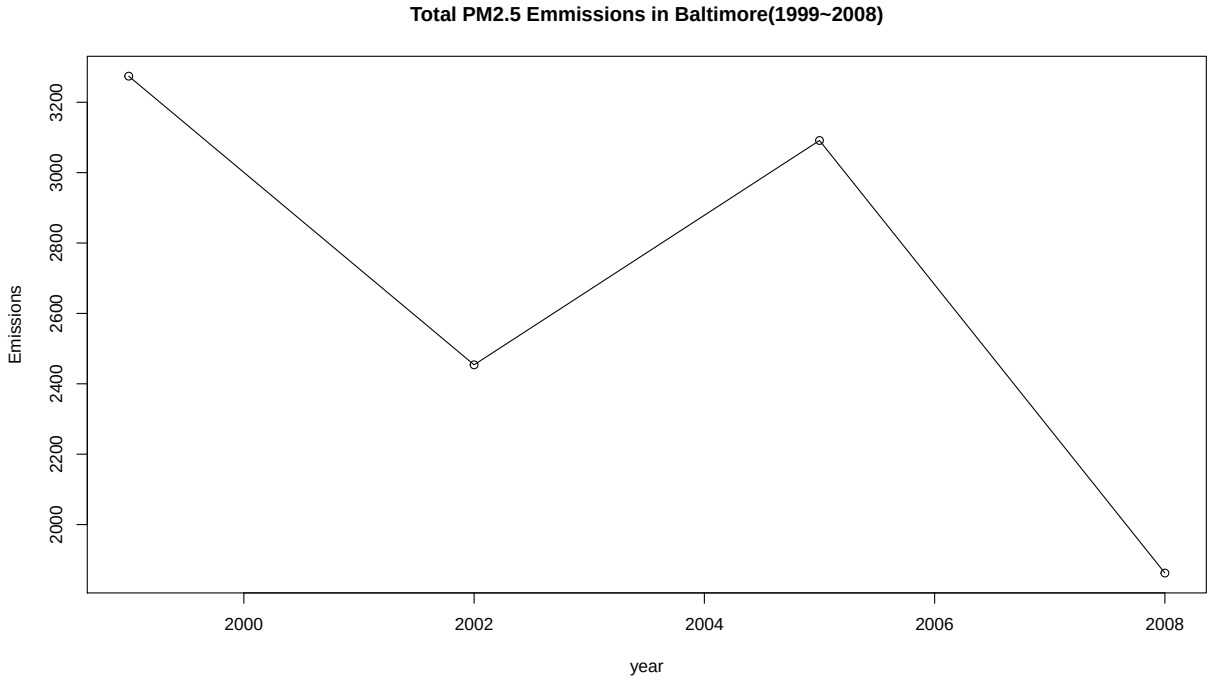
1. Have total emissions from PM2.5 decreased in the United States from 1999 to 2008? Using the base plotting system, make a plot showing the total PM2.5 emission from all sources for each of the years 1999, 2002, 2005, and 2008.

```
## tapply method. however this method can hardly retrieve the year element. Have to input it manually.
emissionSum = with(NEI, tapply(Emissions, year, sum))
yearunique = names(emissionSum)
# yearunique = unique(NEI$year)


emissionT1 = data.frame(year = yearunique, emission = emissionSum)

## Initial the points
with(emissionT1, plot(yearunique, emission))
## Connect by line:
with(emissionT1, lines(yearunique, emission))
## Add a title:
title("Total PM2.5 Emmissions in US(1999~2008)")
```



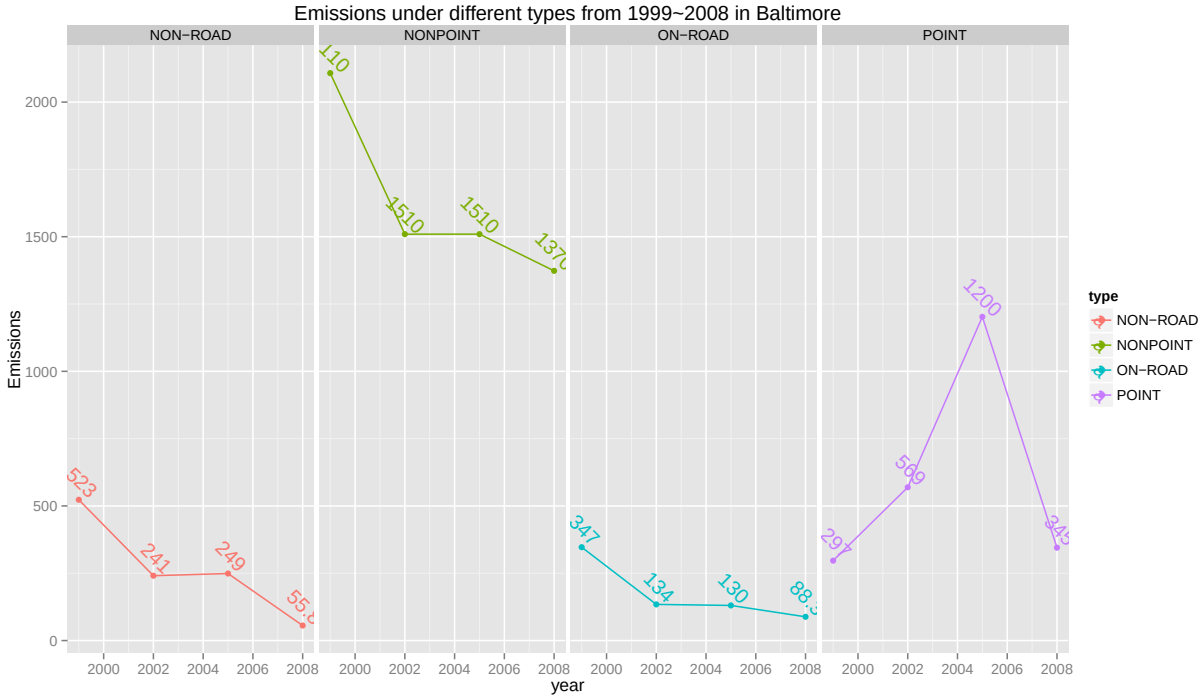
2. Have total emissions from PM2.5 decreased in the Baltimore City, Maryland (fips == 24510) from 1999 to 2008? Use the base plotting system to make a plot answering this question.

```
## Aggregate method. Easy and feasible.
BaltimoreT = subset(NEI, fips == 24510)
emissionT2 = aggregate(Emissions~year, data = BaltimoreT, sum)

## Initial the points
with(emissionT2, plot(year, Emissions))
## Connect by line:
with(emissionT2, lines(year, Emissions))
## Add a title:
title("Total PM2.5 Emmissions in Baltimore(1999~2008)")
```



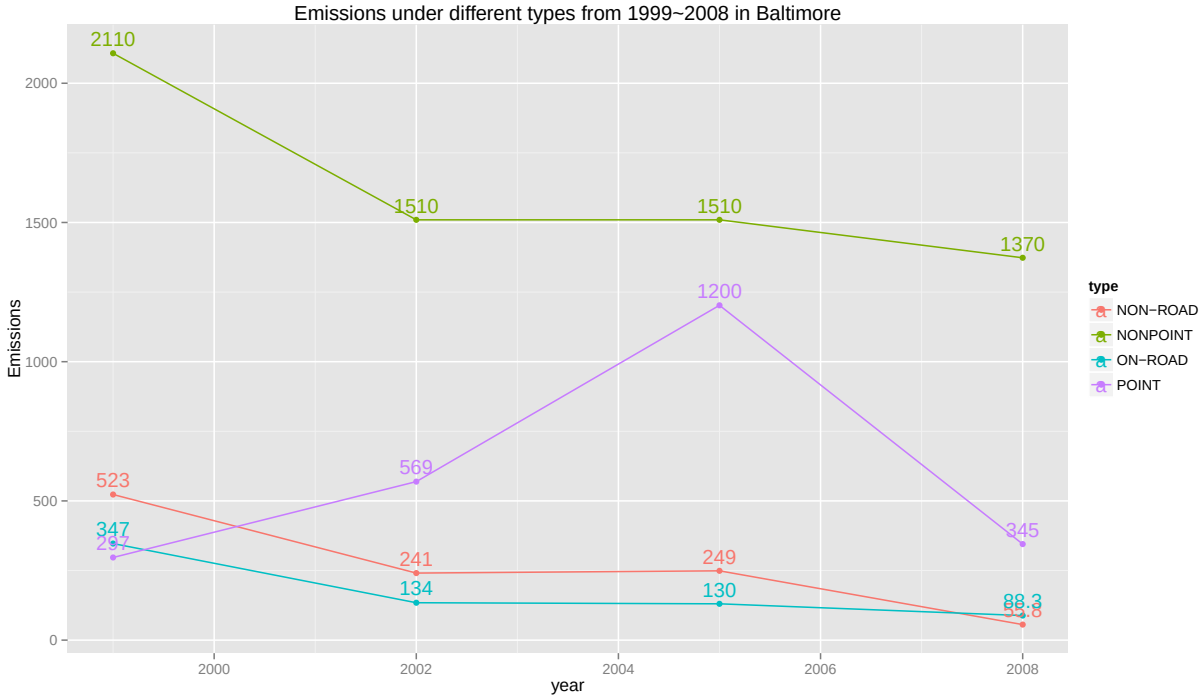
3. Of the four types of sources indicated by the type (point, nonpoint, onroad, nonroad) variable, which of these four sources have seen decreases in emissions from 1999-2008 for Baltimore City? Which have seen increases in emissions from 1999-2008? Use the ggplot2 plotting system to make a plot answer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

```
## Aggregate the table to see emissions distribute along years and types
emissionT3 = aggregate(Emissions~year+type, data = BaltimoreT, sum)

## Establish the ggplot object
p3 = ggplot(emissionT3, aes(year, Emissions, label = signif(Emissions, 3), color = type))

## Add point and line layers(Separating type by colors)
## Add facets along the types
## Add a title
p3 + geom_point() + geom_line() + facet_grid( .~type) + labs(title = "Emissions under different types fi
```



Or

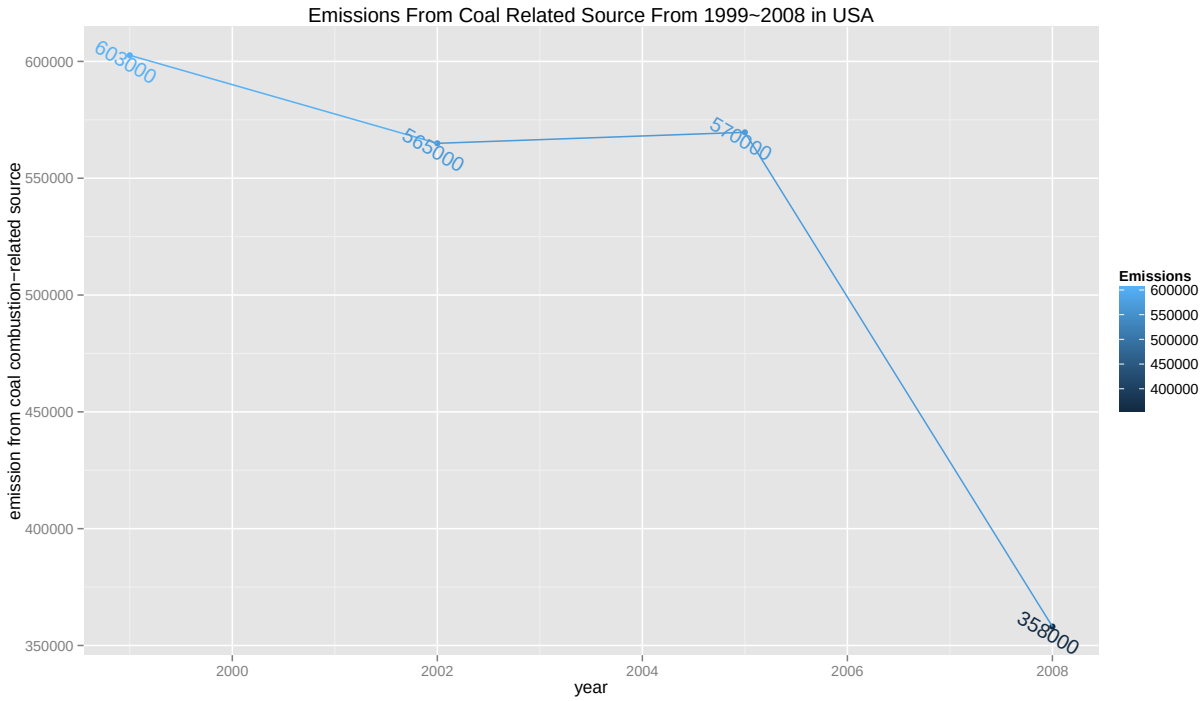
```
## Establish the ggplot object
pp4 = ggplot(emissionT3, aes(year, Emissions, color = type, label = signif(Emissions, 3)))

## Add points and lines layers(Separating type by colors)
## Add texts and labels
pp4 + geom_point() + geom_line() + geom_text(vjust = -0.5) + labs(title = "Emissions under different ty
```



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

```
## Merge the dataframes:
## Set all = True in order to get a complete table:
EmiUSA = merge(NEI, SCC, by.x = "SCC", by.y = "SCC", all = TRUE)

## Specifying coal related emission by grepl function:
coalEmiUSA = EmiUSA[with(EmiUSA, grepl("coal", Short.Name, ignore.case = TRUE)), ]

## Aggregate for the Emission along years:
coalDeltaUSA = aggregate(Emissions ~ year, data = coalEmiUSA, sum)

## Tapply cannot be used to generate plot, because it comes out with an array instead of a dataframe:
#coalDeltaUSA = with(coalEmiUSA, tapply(Emissions, year, sum))
#coalDeltaUSA
```

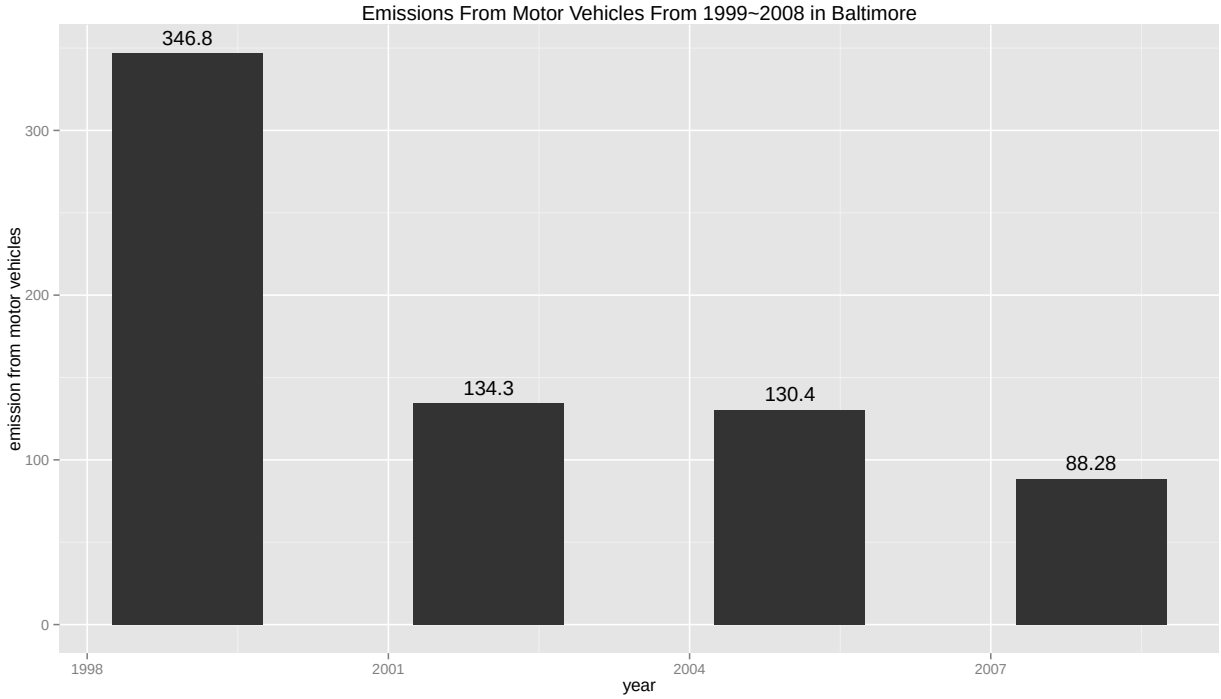
```
## Establish the ggplot object
p4 = ggplot(coalDeltaUSA, aes(year, Emissions, color = Emissions))

## Add point and line layers
## Add texts
## Add titles
p4 + geom_point() + geom_line() + labs(title = "Emissions From Coal Related Source From 1999~2008 in US
```



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

```
## Assume all "onroad" type belongs to motor vehicles

## Merge the Baltimore table
## Set all = FALSE because we don't want the entire table:
EmiBalt = merge(BaltimoreT, SCC, by.x = "SCC", by.y = "SCC")

## Specifying motor vehicle related emission by grepl function:
mvEmiBalt = subset(EmiBalt, type == "ON-ROAD")

## Aggregate for the Emission along years:
mvDeltaBalt = aggregate(Emissions ~ year, data = mvEmiBalt, sum)

#mvDeltaBalt
```

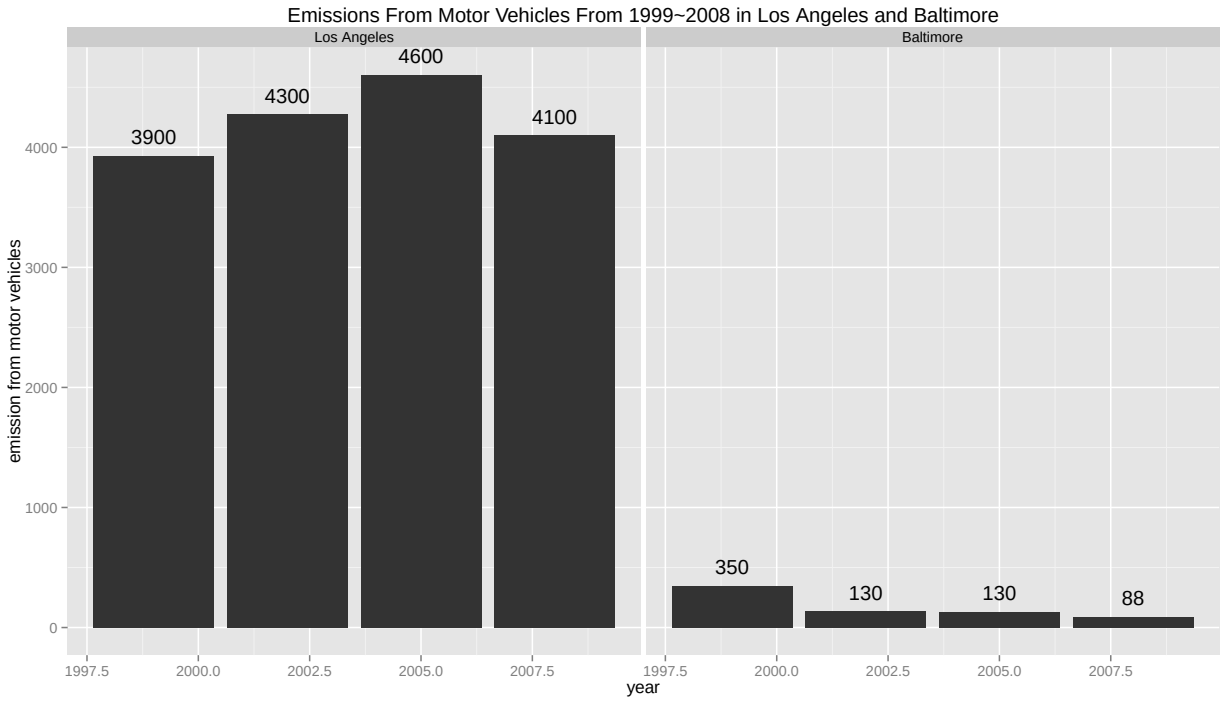
```
## Establish the ggplot object
p5 = ggplot(mvDeltaBalt, aes(year, Emissions, label = signif(Emissions,4)))

## When the data contains y values(instead of count) in a column, use stat="identity"
## Add barplots, text, labels
p5 + geom_bar(stat = "identity", width = 1.5) + geom_text(vjust = -0.6) + labs(title = "Emissions From
```



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?

```
## Assume all "On-Road" type belongs to motor vehicles

## Select for the global table by location(LA and Baltimore) and emission type(On-Road)
mvEmiLABalt = subset(EmiUSA, ((fips %in% c("06037", "24510")) & (type == "ON-ROAD")))

## Change the location information to factor level of LA and Baltimore:
mvEmiLABalt$Location = factor(mvEmiLABalt$fips, label = c("Los Angeles", "Baltimore"), order = TRUE)

## Aggregate the emission along years, type and location:
mvDeltaLABalt = aggregate(Emissions~year+Location, data = mvEmiLABalt, sum)
```

```
## Establish the ggplot object
p6 = ggplot(mvDeltaLABalt, aes(year, Emissions, label = signif(Emissions,2)))

## Add barplots, text, labels
p6 + geom_bar(stat = "identity") + facet_grid(.~Location) + geom_text(vjust = -0.8) + labs(title = "Emi
```



Or

```
pp6 = ggplot(mvDeltaLABalt, aes(year, Emissions, color = Location, label = signif(Emissions,2)))

pp6 + geom_point() + geom_line() + geom_text(vjust = -0.5) + labs(title = "Emissions From Motor Vehicles
```