Getting and Cleaning Data assignment

Oleksii Yehorchenkov

30 11 2020

This assignment is based on materials from Coursera course Exploratory Data Analysis

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 could be downloaded by the link:

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.

- **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:

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

```
library(ggplot2)
library(RColorBrewer)
dir.create("./AirPollutionData")
NEI <- readRDS("./AirPollutionData/summarySCC_PM25.rds")
SCC <- readRDS("./AirPollutionData/Source_Classification_Code.rds")
str(NEI)
```

```
## 'data.frame':    6497651 obs. of  6 variables:
##  $ fips     : chr  "09001" "09001" "09001" "09001" ...
##  $ SCC      : chr  "10100401" "10100404" "10100501" "10200401" ...
##  $ Pollutant: chr  "PM25-PRI" "PM25-PRI" "PM25-PRI" "PM25-PRI" ...
##  $ Emissions: num  15.714 234.178 0.128 2.036 0.388 ...
##  $ type     : chr  "POINT" "POINT" "POINT" "POINT" ...
##  $ year     : int  1999 1999 1999 1999 1999 1999 1999 1999 1999 1999 ...
```

```
str(SCC)
```

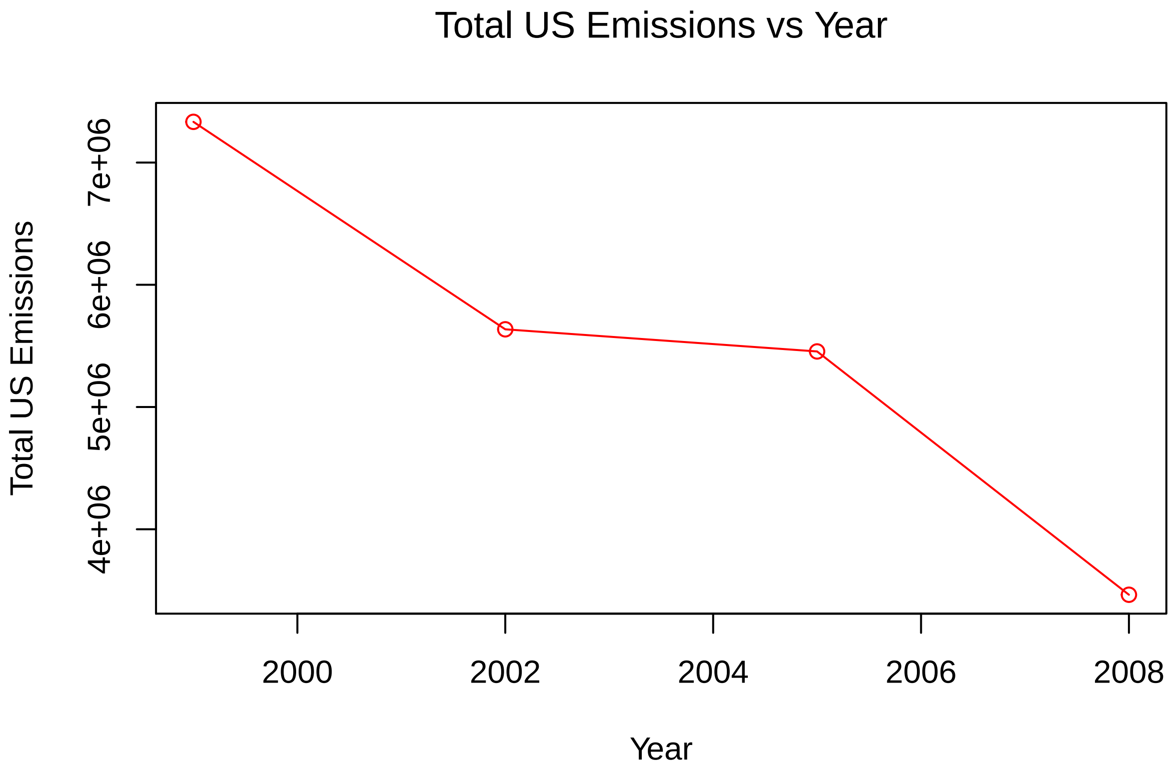
```
## 'data.frame':    11717 obs. of  15 variables:
##  $ SCC                : Factor w/ 11717 levels "10100101","10100102",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Data.Category      : Factor w/ 6 levels "Biogenic","Event",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ Short.Name         : Factor w/ 11238 levels "","2,4-D Salts and Esters Prod /Process Vents, 2,4-D
##  $ EI.Sector          : Factor w/ 59 levels "Agriculture - Crops & Livestock Dust",..: 18 18 18 18 1
##  $ Option.Group       : Factor w/ 25 levels "","C/I Kerosene",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Option.Set         : Factor w/ 18 levels "","A","B","B1A",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ SCC.Level.One      : Factor w/ 17 levels "Brick Kilns",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ SCC.Level.Two      : Factor w/ 146 levels "","Agricultural Chemicals Production",..: 32 32 32 32
##  $ SCC.Level.Three    : Factor w/ 1061 levels "","100% Biosolids (e.g., sewage sludge, manure, mixtu
##  $ SCC.Level.Four     : Factor w/ 6084 levels "","(NH4)2 SO4 Acid Bath System and Evaporator",..: 44
##  $ Map.To             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Last.Inventory.Year: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Created_Date       : Factor w/ 57 levels "","1/27/2000 0:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Revised_Date       : Factor w/ 44 levels "","1/27/2000 0:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Usage.Notes        : Factor w/ 21 levels ""," ","includes bleaching towers, washer hoods, filtrate
```

Questions

You must address the following questions and tasks in your exploratory analysis. For each question/task you will need to make a single **bar** plot. You can use any plotting system in R to make your plot.

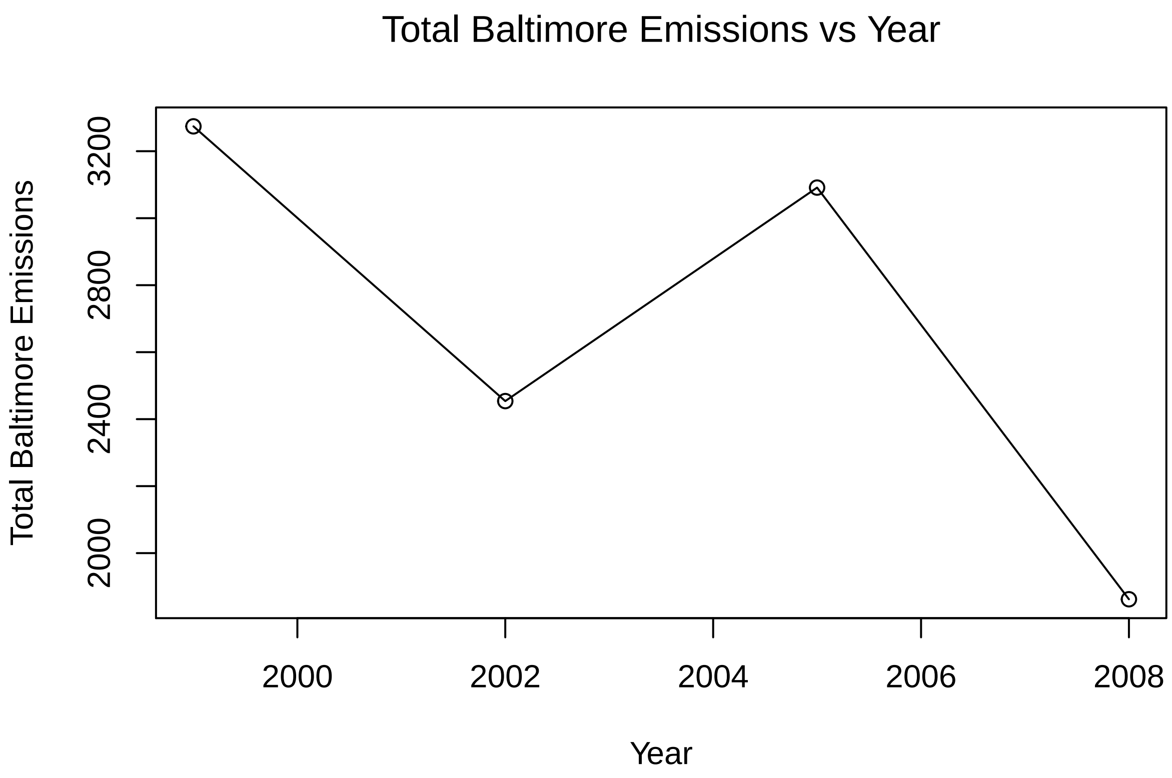
1. Have total emissions from PM2.5 decreased in the United States from 1999 to 2008? Make a plot showing the **total** PM2.5 emission from all sources for each of the years 1999, 2002, 2005, and 2008.

```
total_NEI <- aggregate(Emissions ~ year, NEI, sum)

plot(total_NEI$year, total_NEI$Emissions, type = "o", col = "red", main = expression("Total US Emission
```



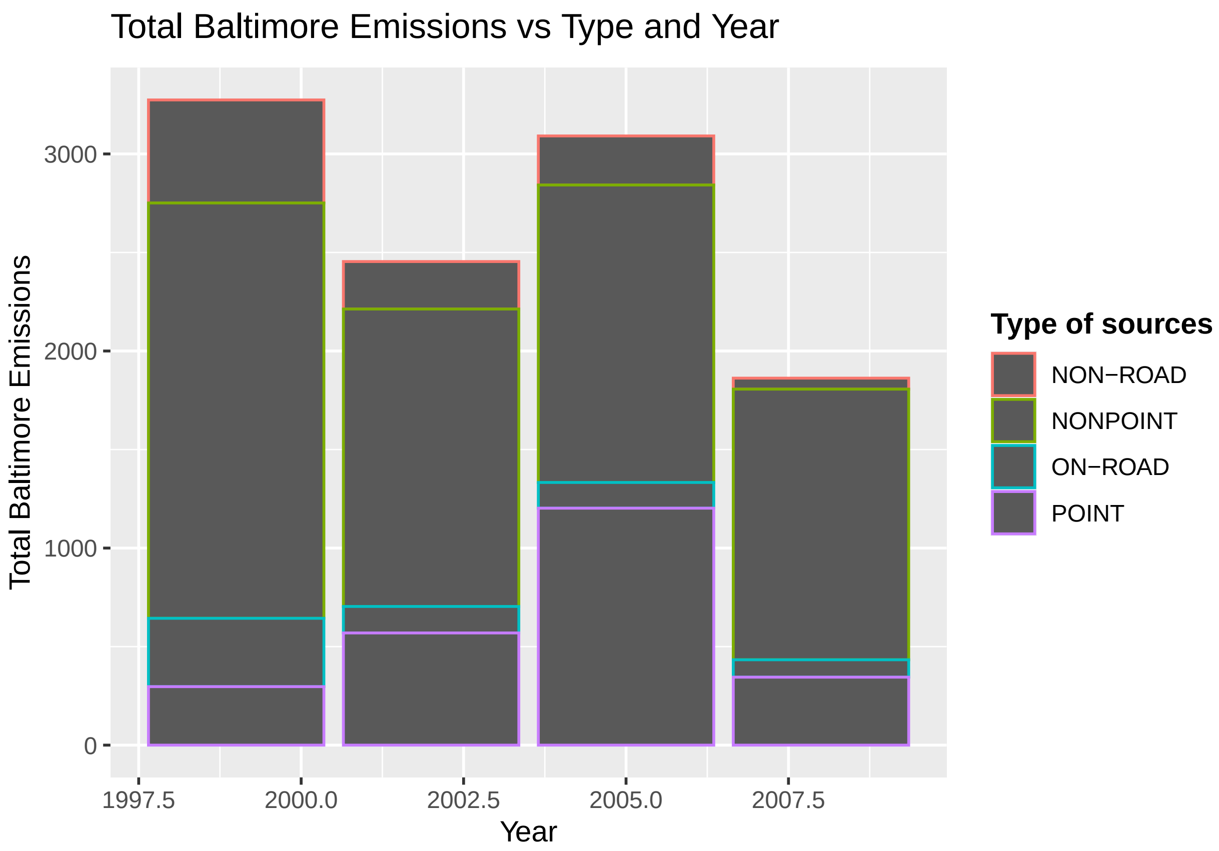
2. Have total emissions from PM2.5 decreased in the **Baltimore City**, Maryland (`fips == "24510"`) from 1999 to 2008?

```
baltimore_city <- subset(NEI, NEI$fips == "24510")

total_baltimore_city <- aggregate(Emissions ~ year, baltimore_city, sum)

plot(total_baltimore_city$year, total_baltimore_city$Emissions, type = "o", main = expression("Total Bal
```



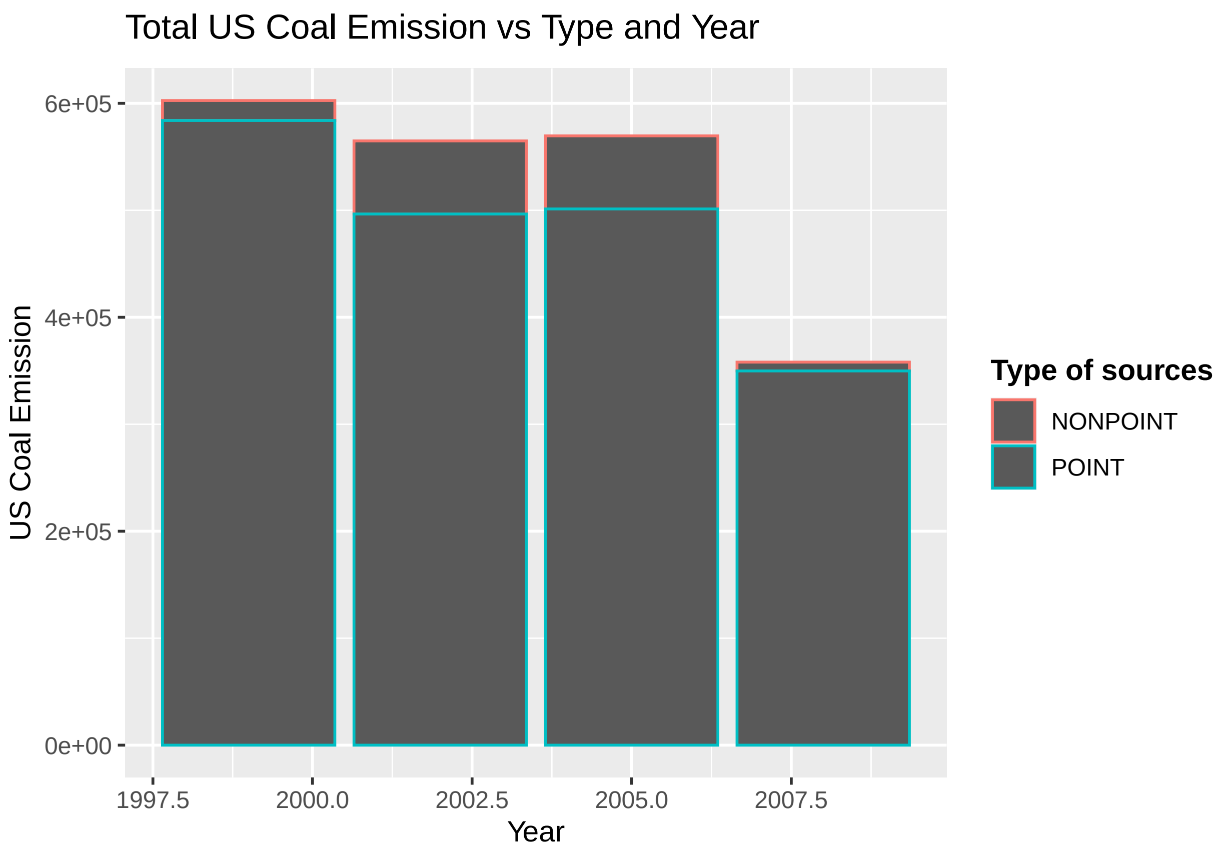
3. Of the four types of sources indicated by the `type` (point, nonpoint, onroad, nonroad) variable, which of these four sources have seen decreases in emissions from 1999–2008 for **Baltimore City**? Which have seen increases in emissions from 1999–2008?

```r
baltimore_city_3 <- subset(NEI, NEI$fips == "24510")
baltimore_Type <- aggregate(Emissions ~ year + type, baltimore_city_3, sum)

ggplot(baltimore_Type, aes(year, Emissions, col = type)) +geom_bar(stat = "identity")+
      ggtitle(expression("Total Baltimore Emissions vs Type and Year")) +
      ylab(expression("Total Baltimore Emissions")) +
      xlab("Year") +
      scale_colour_discrete(name = "Type of sources") +
      theme(legend.title = element_text(face = "bold"))
```



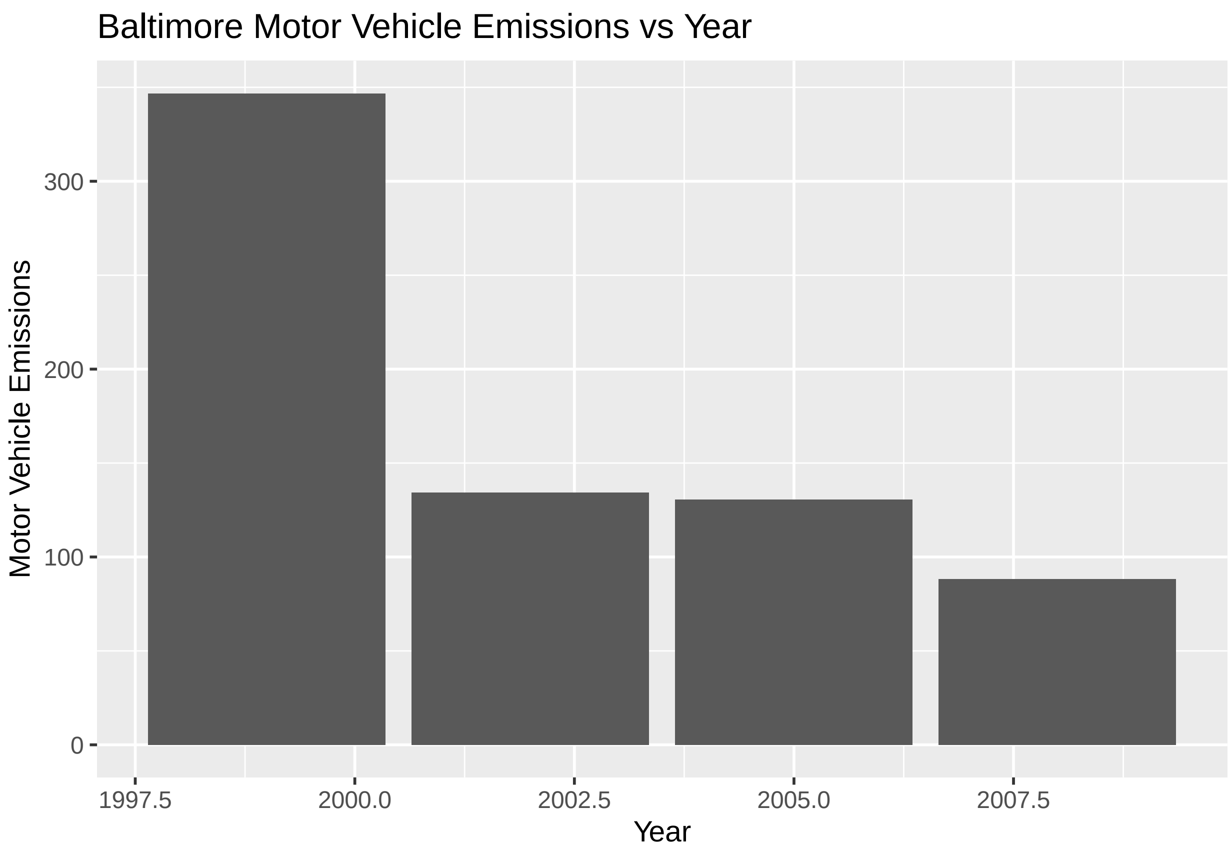
4. Across the United States, how have emissions from coal combustion-related sources changed from 1999–2008?

```
SCC_coal <- SCC[grepl("coal", SCC$Short.Name, ignore.case = T),]
NEI_coal <- NEI[NEI$SCC %in% SCC_coal$SCC,]
total_Coal <- aggregate(Emissions ~ year + type, NEI_coal, sum)

ggplot(total_Coal, aes(year, Emissions, col = type)) +geom_bar(stat = "identity")+
      ggtitle(expression("Total US Coal Emission vs Type and Year")) +
      xlab("Year") +
      ylab(expression("US Coal Emission")) +
      scale_colour_discrete(name = "Type of sources") +
      theme(legend.title = element_text(face = "bold"))
```



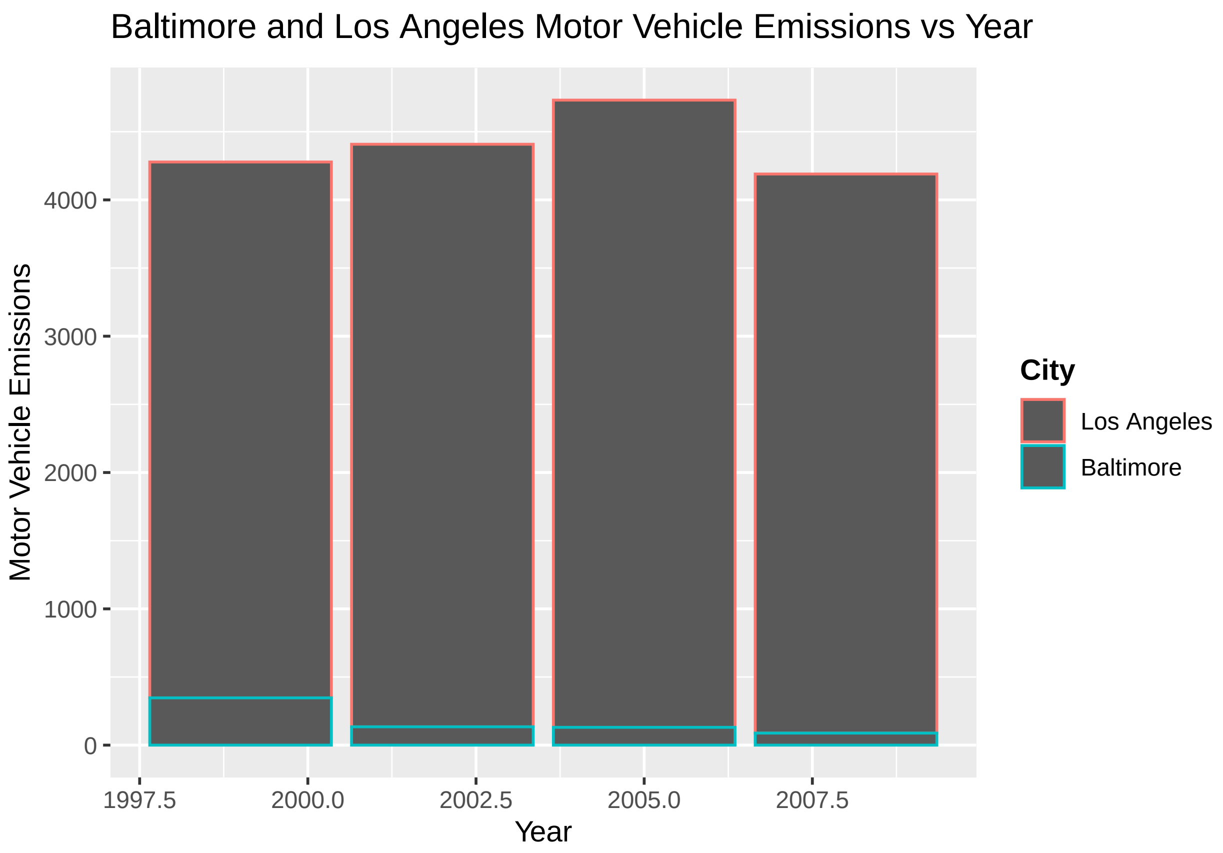
5. How have emissions from motor vehicle sources changed from 1999–2008 in **Baltimore City** (EI.Sector starts from "Mobile")?

```
baltimore_Motor <- subset(NEI, NEI$fips == "24510" & NEI$type == "ON-ROAD")
baltimore_Motor_AGG <- aggregate(Emissions ~ year, baltimore_Motor, sum)

ggplot(baltimore_Motor_AGG, aes(year, Emissions)) +geom_bar(stat = "identity")+
      ggtitle(expression("Baltimore Motor Vehicle Emissions vs Year")) +
      xlab("Year") +
      ylab(expression("Motor Vehicle Emissions"))
```



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?

```r
baltimore_LosAngeles_Motors <- subset(NEI, NEI$fips %in% c("24510","06037") & NEI$type == "ON-ROAD")
baltimore_LosAngeles_Motors_AGG <- aggregate(Emissions ~ year + fips, baltimore_LosAngeles_Motors, sum)

ggplot(baltimore_LosAngeles_Motors_AGG, aes(year, Emissions, col = fips)) +geom_bar(stat = "identity")+
      ggtitle(expression("Baltimore and Los Angeles Motor Vehicle Emissions vs Year")) +
      labs(x = "Year", y = expression("Motor Vehicle Emissions") ) +
      scale_colour_discrete(name = "City", labels = c("Los Angeles", "Baltimore")) +
      theme(legend.title = element_text(face = "bold"))
```


Baltimore and Los Angeles Motor Vehicle Emissions vs Year
Motor Vehicle Emissions
4000
3000
2000
1000
0
1997.5
2000.0
2002.5
2005.0
2007.5
Year
City
Los Angeles
Baltimore