Exploratory Data Analysis Project 2

This is the final project for "Exploratory Data Analysis"

The case study we are doing is to examine the emissions of PM2.5 across year 1999~ 2008 within US, looking for specific city such as Baltimore, LA, and also looking for specific type, such as On-road, off-road.

This case study not only help me understand how to initiate a basic exploratory data analysis, but also assist me review all of the previous lecture related to basic R programming.

The following is the basic skill included in this case study:

1. Basic plotting system
2. Basic ggplot plotting system
3. Subsetting with various conditions
4. grepl: Grasping for specific text and returing a factor(condition to apply to your origianl data set)
5. tapply: applying function to a variable grouping by a factor, returing a numeric vector
6. aggregate: applying function to a variable grouping by a list, returing a data frame

Data Insturctions (copy from Coursera):

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: 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.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".

```
#Basic: Reading data and setting working directory
library(ggplot2)
NEI <- readRDS("/Users/andrewhu/Documents/GitHub/Coursera_DataScience_JHU/Exploratory Data Analysis/summ
SCC <- readRDS("/Users/andrewhu/Documents/GitHub/Coursera_DataScience_JHU/Exploratory Data Analysis/Sou
```

Question 1:

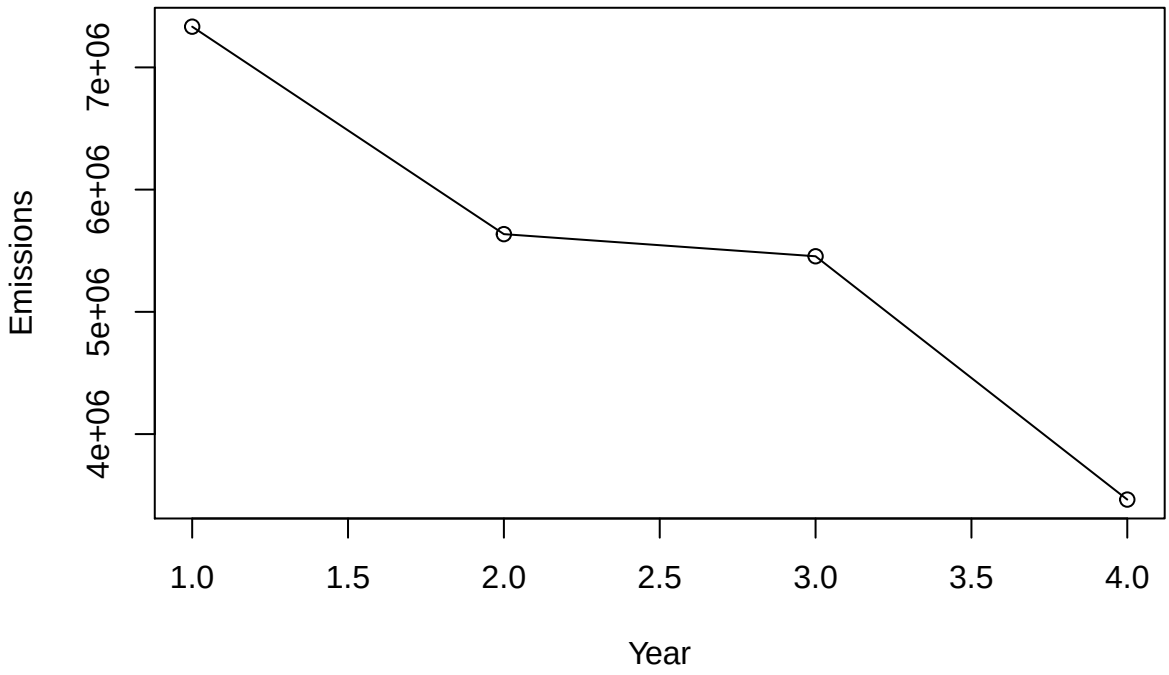
Have total emissions from PM2.5 decreased in the United States from 1999 to 2008? Using the base plotting system, make a plot showing the total PM2.5 emission from all sources for each of the years 1999, 2002, 2005, and 2008

Solution: we use tapply to calculate total emissions by year, and it will return a numeric vector indicating the sum of emissions for each year.

Structure for tapply

tapply(the var you want to calculate, by xxx factor, the function)**

```
#calculate the total emissions by year
Emi_over_year_total<- with(NEI, tapply(Emissions, year, sum, na.rm=T))
#basic plot
plot(Emi_over_year_total, type="o", xlab="Year", ylab= "Emissions")
```



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? Use the ggplot2 plotting system to make a plot answer this question.

Solution: Now we want to examine four types of sources within Baltimore city. First, we need to use the aggregate function, to return a data frame that has three variables we are interested: Type, year and Emissions.

Structure for aggregate function: aggregate(df with a specific var you want to calculate, list of factor variables, function)

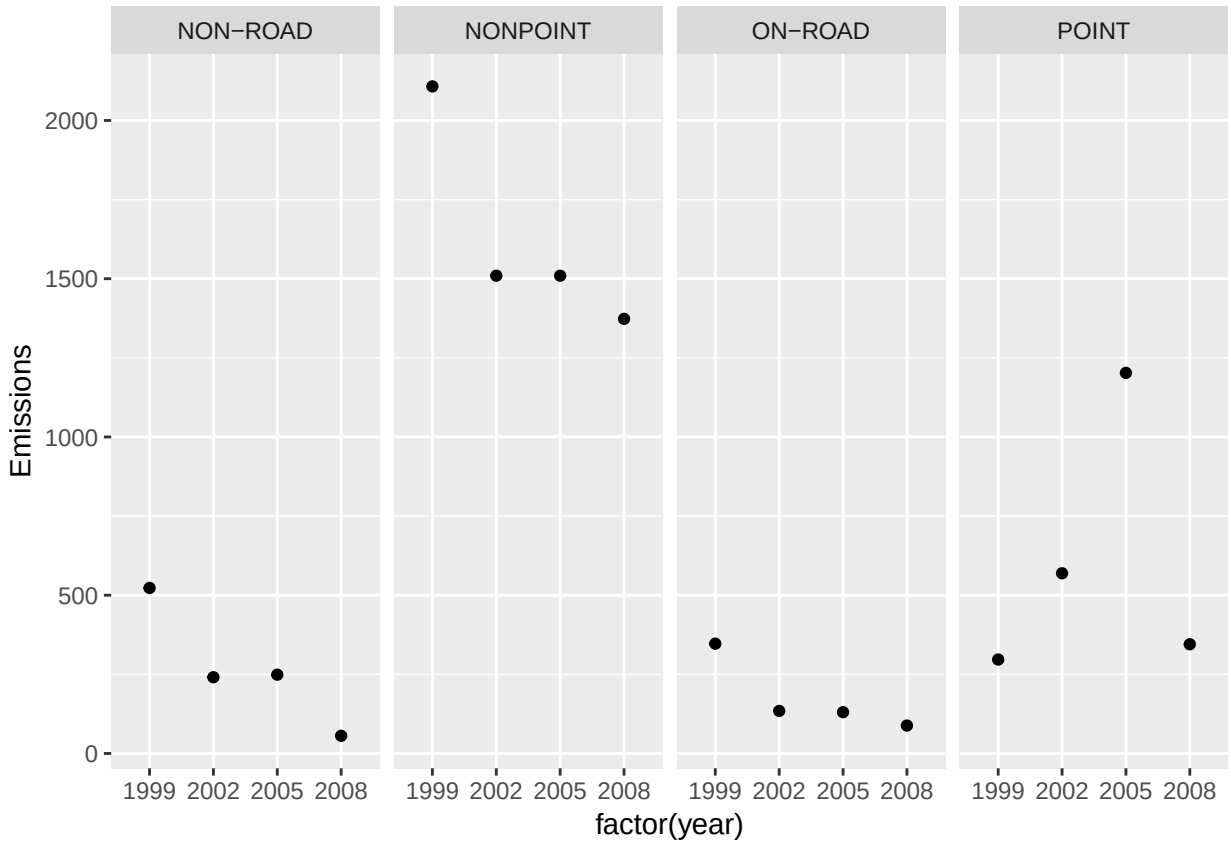
Structure for ggplot: Basic: ggplot(df, aes(x,y)) Addings: geom_point(), geom_bar(stat= "identity").. Categorize: facet_grid(.~ the factor)

```
#Subset a df: sum for Emissions, by type and by year
Emi_over_year_BA_df <- aggregate(NEIBA[c("Emissions")], list(type= NEIBA$type, year= NEIBA$year), sum )

##ggplot

#base
g<- ggplot(Emi_over_year_BA_df, aes(factor(year),Emissions) )

#
g + geom_point()+ facet_grid(. ~type)
```



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

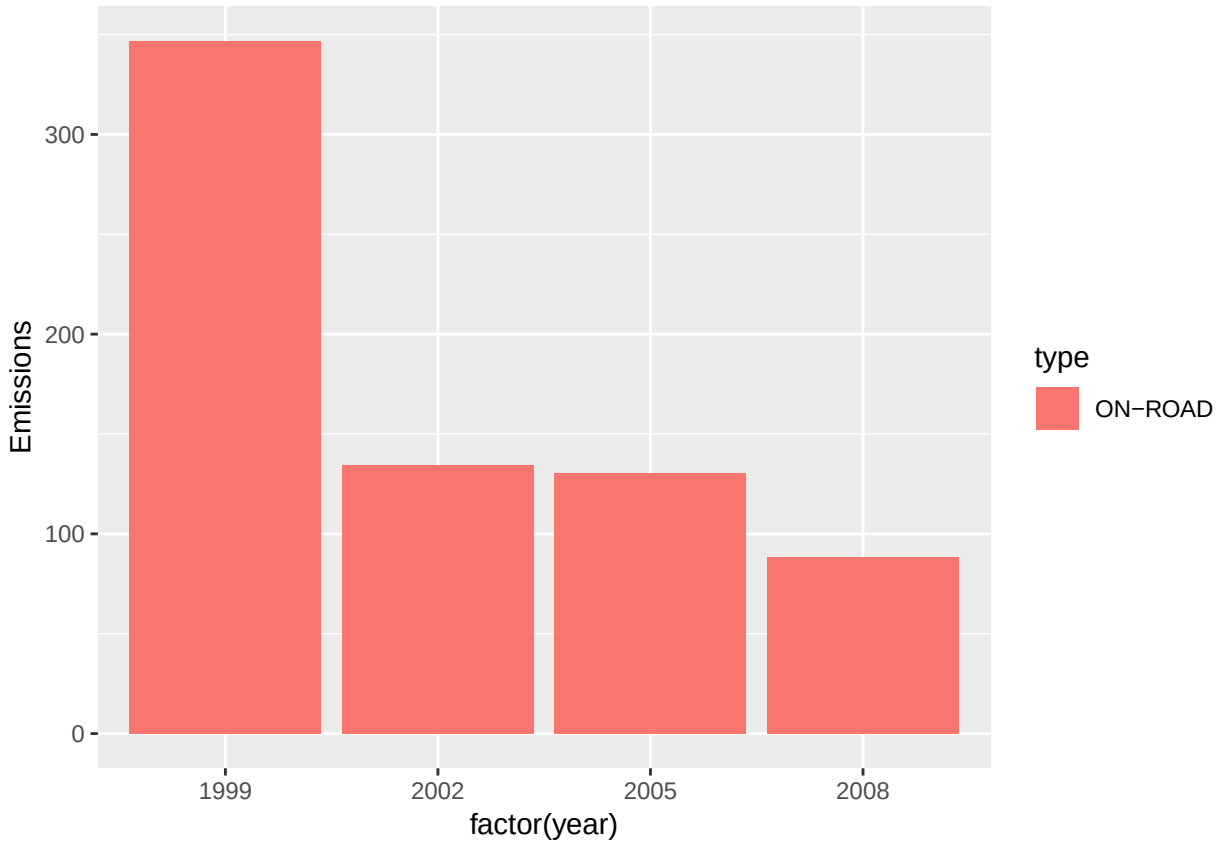
Solution: 1. Simply subset from the origianl data with BA and type of "ONROAD" 2. Aggregate the data just as we've done previously 3. Plot

```
NEIBAONROAD <- subset(NEI, fips=="24510" & type == "ON-ROAD")


NEI.onroad <- aggregate(NEIBAONROAD[c("Emissions")], list(type= NEIBAONROAD$type, year= NEIBAONROAD$yea

#base
g<- ggplot(NEI.onroad , aes(x= factor(year),y= Emissions, fill=type))

#
g + geom_bar(stat= "identity")
```



Question 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?

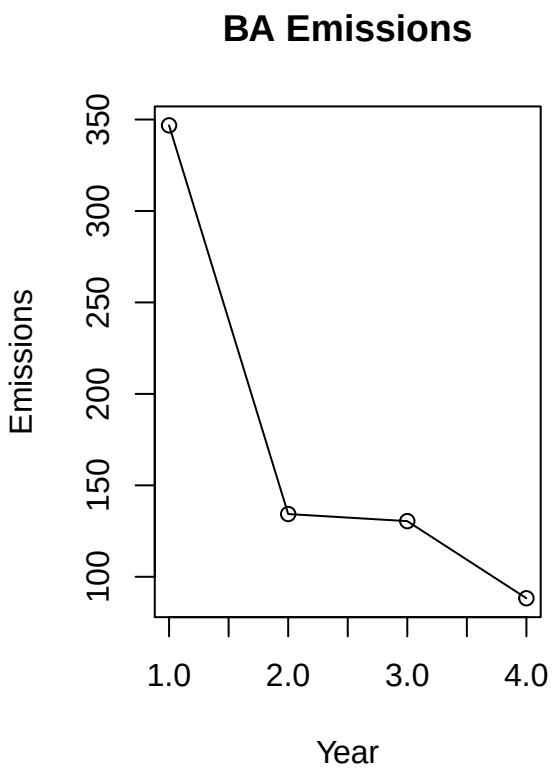
Solution: Simply plot BA / LA and compare. I use basic plotting system so as to set plotting parameters to plot two column at once.

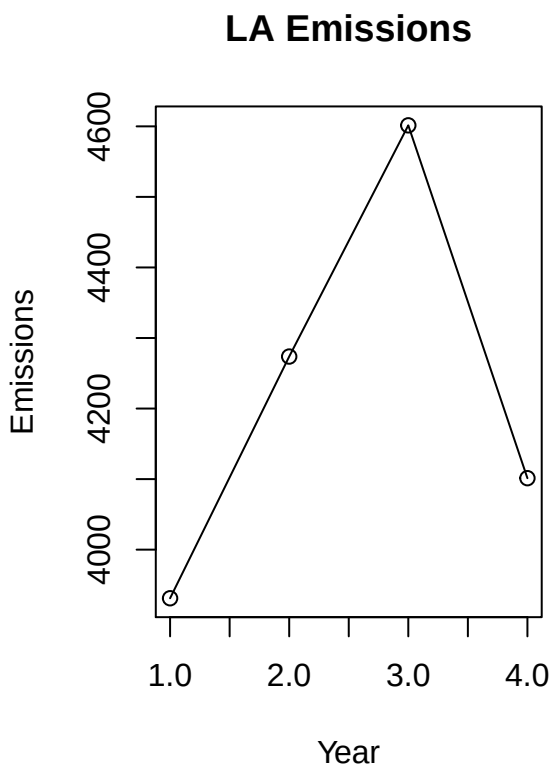
```
NEIBAONROAD <- subset(NEI, fips=="24510" & type == "ON-ROAD")
NEILAONROAD <- subset(NEI, fips=="06037" & type == "ON-ROAD")


par(mfrow=c(1,2))

ba <- with(NEIBAONROAD, tapply(Emissions, year, sum, na.rm= TRUE))
la <- with(NEILAONROAD, tapply(Emissions, year, sum, na.rm= TRUE))


plot(ba, type ="o", xlab="Year", ylab= "Emissions", main="BA Emissions")
plot(la, type ="o", xlab="Year", ylab= "Emissions", main= "LA Emissions")
```


BA Emissions
Emissions
100
150
200
250
300
350
1.0
2.0
3.0
4.0
Year


LA Emissions
Emissions
4000
4200
4400
4600
1.0
2.0
3.0
4.0
Year