Project2

D Togashi

11 November 2015

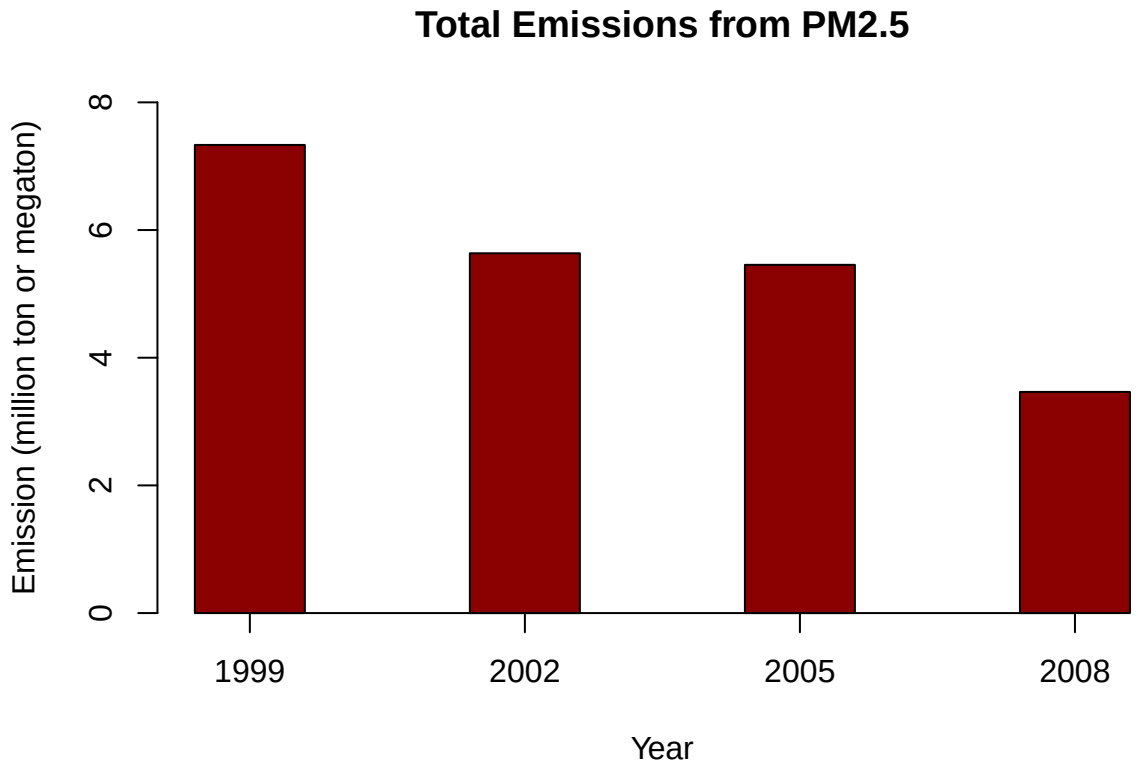
Overview: Fine particulate matter (PM2.5) is an ambient air pollutant for which there is strong evidence that it is harmful to human health. In the United States, the Environmental Protection Agency (EPA) is tasked with setting national ambient air quality standards for fine PM and for tracking the emissions 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 (http://www.epa.gov/ttn/chief/eiinformation.html). 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. Please note, this is part of the Project Assigment of Coursera Data Scientist Certificate.

```
#
#   THIS SECTION BUILDS Plot 1 FOR QUESTION 1
#   ALSO IT LOADS ALL THE LIBRARY AND DATA USED IN THIS PROJECT
#
#
#   if you don't have the package use:
#   install.packages("dplyr")
#   install.packages("ggplot2")
#
    library("dplyr")
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

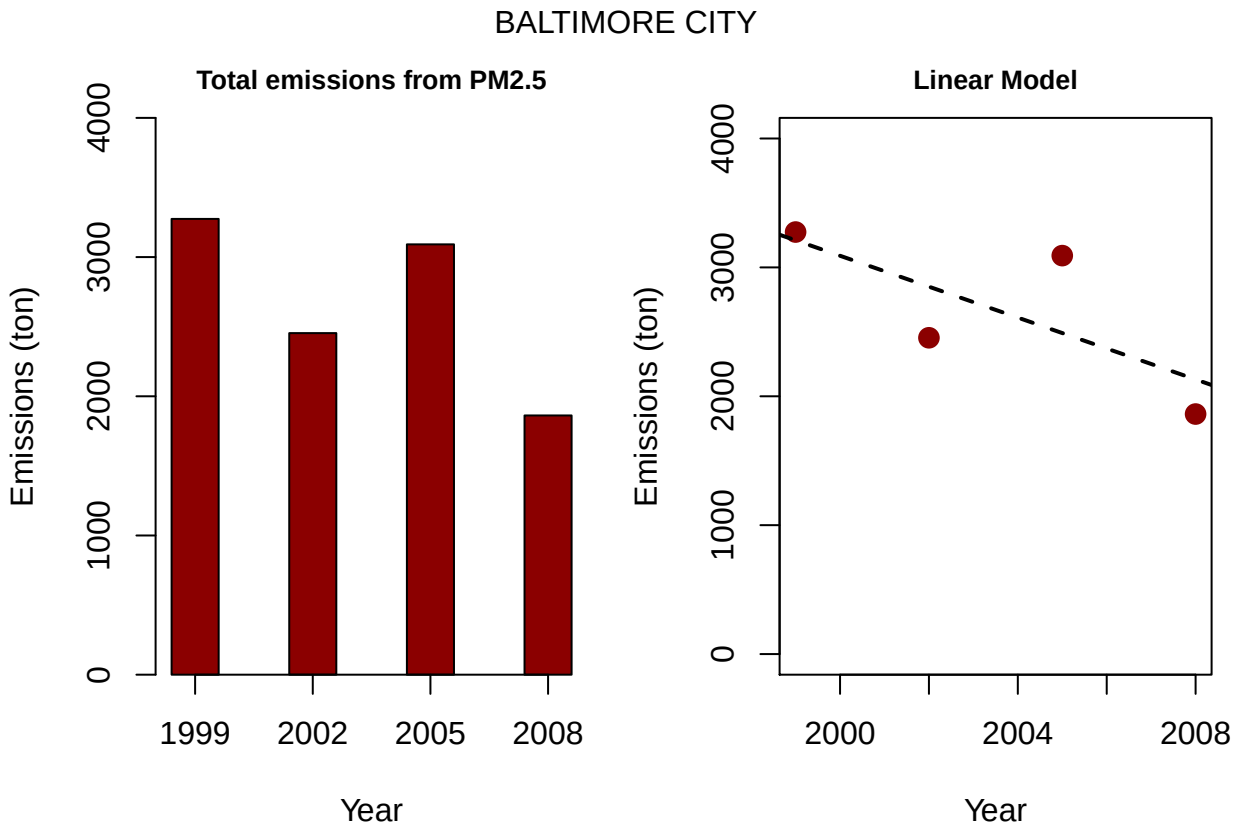
```
    library("ggplot2")

#   Dowloading the datafile
#
    fileUrl <- "http://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip"
    download.file(fileUrl, destfile="./NEI_data.zip", mode="wb")
    unzip("./NEI_data.zip",exdir = ".")

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

```
############################
####    Question 1 : plot 1
############################

    TotalEmissionYear <-with(NEI, tapply(Emissions,year,sum))

#    png(filename="plot1.png",width = 480, height = 480)

    barplot(TotalEmissionYear/1000000,col ='darkred',
        space = 1.5, axis.lty = 1,
        main='Total Emissions from PM2.5',
        xlab="Year",ylab="Emission (million ton or megaton)", ylim=c(0,8))
```



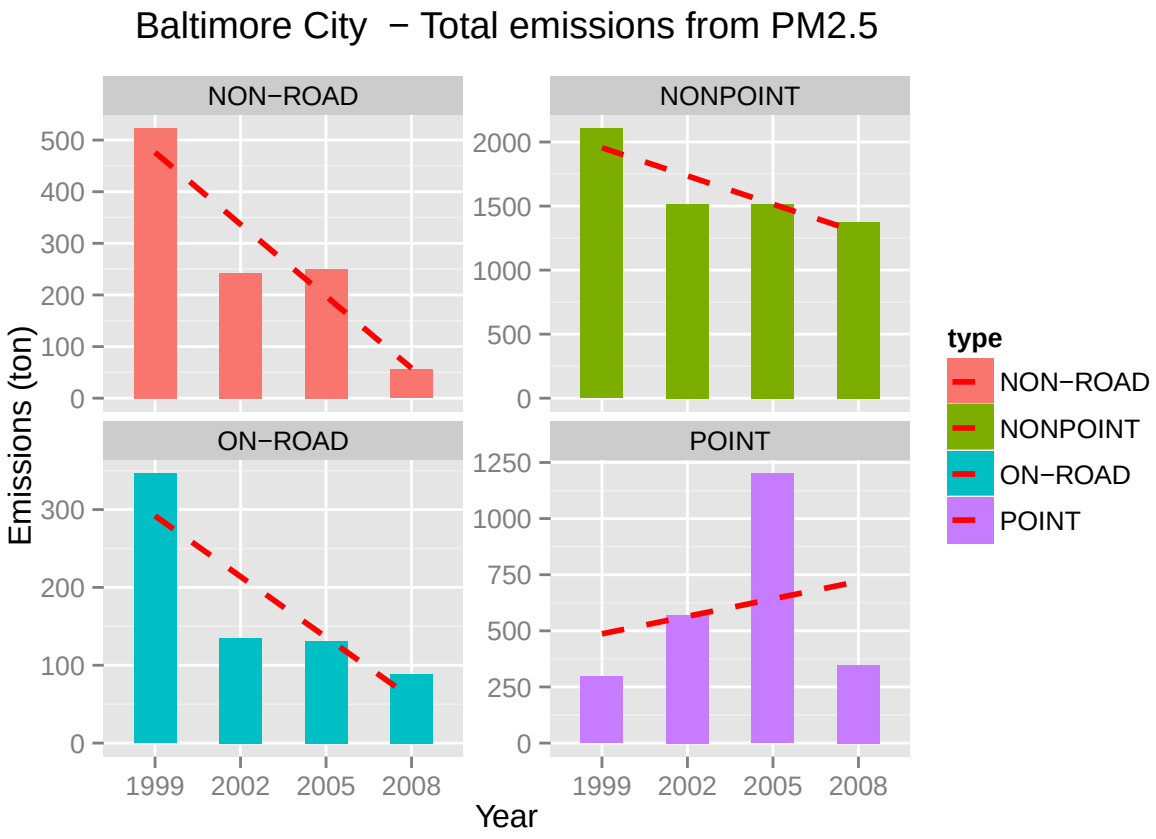

```
#    dev.off()

#   End of Question 1.
#

############################
####    Question 2 : plot 2
############################
#
#   ASSUMING THAT THE QUESTION 1 CODE WAS EXECUTED
#   THE QUESTION 2 CODE IS DESCRIBED BELOW.
```

```r
    county <- split(NEI,NEI$fips)
    Baltimore <- county$"24510"
    BaltimoreTotalEmission <- with(Baltimore,aggregate(Emissions,list(year),sum))
    names(BaltimoreTotalEmission) = c('year','Emission')
    model <- lm(Emission~year,BaltimoreTotalEmission)

#    png(filename="plot2.png",width = 480, height = 480)

    par(mfrow=c(1,2), mar= c(4,4,2,1),oma=c(0,0,2,0))
    barplot(t(BaltimoreTotalEmission)[2,],
            names.arg=t(BaltimoreTotalEmission)[1,],
            col='darkred', space = 1.5, axis.lty = 1,
            main='Total emissions from PM2.5', cex.main=0.8,
            xlab="Year",ylab="Emissions (ton)", ylim=c(0,4000))
    plot(BaltimoreTotalEmission,
            main='Linear Model',cex.main=0.8,
            pch=16, cex=1.5, col='darkred',
            xlab="Year",ylab="Emissions (ton)", ylim=c(0,4000))
    abline(model,lty=2,lwd=2)
    mtext('BALTIMORE CITY', cex.main=1.5, outer = TRUE)
```




```r
#    dev.off()
#
#   End of Question 2.
```

```
############################
####    Question 3 : plot 3
############################
#
#   ASSUMING THAT THE QUESTION 2 CODE WAS EXECUTED
#   THE QUESTION 3 CODE IS DESCRIBED BELOW.
#
####### Function Faux ########
#
#   This auxiliar function subsects the list 'dlist' by the variable 'type'.
#   The output is a dataframe that contains: year, the sum of Emissions, and type.
#   This is useful to create a separated data frame to use the linear model
#   This function is specfic to the this project.
#
    Faux <-function(type,dlist) {
        temp<-NULL
        nyears<-4
        for (i in 1:length(type)) {
            v <- rep(type[i],nyears)
            temp1 <- cbind(with(dlist[[type[i]]],aggregate(Emissions,list(year),sum)),v)
            temp <- rbind(temp,temp1)}
        names(temp)<-c("year","Emissions","type")
        temp}

    BaltimoreEmissionType <- split(Baltimore,Baltimore$type)
    EmissionType <- names(BaltimoreEmissionType)
    BaltimoreTotalEmissionType <- Faux(EmissionType,BaltimoreEmissionType)

    g <- ggplot(Baltimore, aes(x=factor(year),y=Emissions,fill=type))
    plot3 <- g + geom_bar(stat = "identity", width=.5)+
                facet_wrap(~ type,scales = 'free_y') +
                xlab('Year') + ylab('Emissions (ton)')+
                ggtitle('Baltimore City  - Total emissions from PM2.5')+
                theme(plot.title = element_text(vjust=2))+

#   Collating the dashed line from linear model fitting of BaltimoreTotalEmissionType:
                geom_smooth(data=BaltimoreTotalEmissionType,
                        method = "lm",aes(group = type),se=FALSE,
                        colour = "red", size = 1,linetype=2)

#    png(filename="plot3.png",width = 480, height = 480)
    plot3
```




```
#    dev.off()
#
#   Note: The question refers to TOTAL EMISSION and not the average.
#   Linear model applied to Baltimore data frame will use the average values.
#
#   End of Question 3.


############################
####    Question 4 : plot 4
############################
#
#   ASSUMING THAT THE QUESTION 3 CODE WAS EXECUTED
#   THE QUESTION 4 CODE IS DESCRIBED BELOW.
#
#
#
####### FUNCTION word_filter
#
#   This function search a specific array of string, word, in the data frame, df.
#   Create an unique list of row number in which "word" appears, at least once,
#   in any column of df.
#   arguments: word, df and ignore.case (TRUE or FALSE)

    word_filter <- function(word,df,ignore.case) {
        list1 <- sapply(df[,1:dim(df)[2]],function(x){sum(grepl(word,x,ignore.case))})
```

```
        temp <- NULL
        for (i in 1:dim(df)[2]) {temp[[i]] <- isTRUE(list1[[i]]!=0)}
        list2 <- lapply(df[,temp],function(x){grep(word,x,ignore.case)})
#       Only elements that do not repeat,
        Reduce(union,list2) %>% sort}
#
#   end of function.
#
#
#   Much of controversy was found in this question. Here, I considered only emissions
#   that appears 'Coal' AND 'combustion'. I ingnored the case for 'combution' and,
#   considered the case for 'Coal' - the word "Charcoal" is removed.
#
    all_coal <- word_filter('Coal',SCC,FALSE)
    all_combustion <- word_filter('comb',SCC,TRUE)

#   Creating a vector that contains coal AND combustion related words
    all_coal_comb <- intersect(all_coal,all_combustion)

#   Creating a list of SCC index of the source name in SCC table
    SCC_coal_comb <- SCC[all_coal_comb,]$SCC
    coal_comb_total_emission <- with(filter(NEI,SCC %in% SCC_coal_comb),
                                tapply(Emissions,year,sum))
#    png(filename="plot4.png",width = 480, height = 480)
    barplot(coal_comb_total_emission/1000, col ='darkred',
        main='US Total coal combustion-related Emissions',xlab="Year",
        space = 1.5, axis.lty = 1,
        ylab="Emission (thousands ton or kiloton)",ylim=c(0,600))
    abline(h=550, lty=2, col='blue')
#    dev.off()

#   End of Question 4.
#


############################
####    Question 5 : plot 5
############################
#
#   ASSUMING THAT THE QUESTION 4 CODE WAS EXECUTED
#   THE QUESTION 5 CODE IS DESCRIBED BELOW.
#
#
#   It can find observables in SCC table for "Motor" and "Vehicle"

    all_motor <- word_filter('motor',SCC,TRUE)
    all_vehicle <- word_filter('vehicle',SCC,TRUE)

    motor_OR_vehicle <- union(all_motor,all_vehicle)
    motor_AND_vehicle <- intersect(all_motor,all_vehicle)

#   Creating the list of SCC index of the source name in SCC table
    SCC_motor_OR_vehicle <- SCC[motor_OR_vehicle,]$SCC
```

```
    SCC_motor_AND_vehicle <- SCC[motor_AND_vehicle,]$SCC
    SCC_vehicle <- SCC[all_vehicle,]$SCC

#   Subsecting for Baltimore city and Motor and/or Vehicle, and Vehicle standading alone
#   Remember in code  2:
#   county <- split(NEI,NEI$fips)
#   Baltimore <- county$"24510"

    motor_AND_vehicle.baltimore<-filter(Baltimore, SCC %in% SCC_motor_AND_vehicle)
    motor_OR_vehicle.baltimore <- filter(Baltimore, SCC %in% SCC_motor_OR_vehicle)
    vehicle.baltimore <- filter(Baltimore, SCC %in% SCC_vehicle)

#   By using dim, vehicle.baltimore and motor_OR_vehicle.baltimore have same size (1395 by 6).
#   The two data.frames are the same, by using:

    identical(motor_OR_vehicle.baltimore,vehicle.baltimore)
```

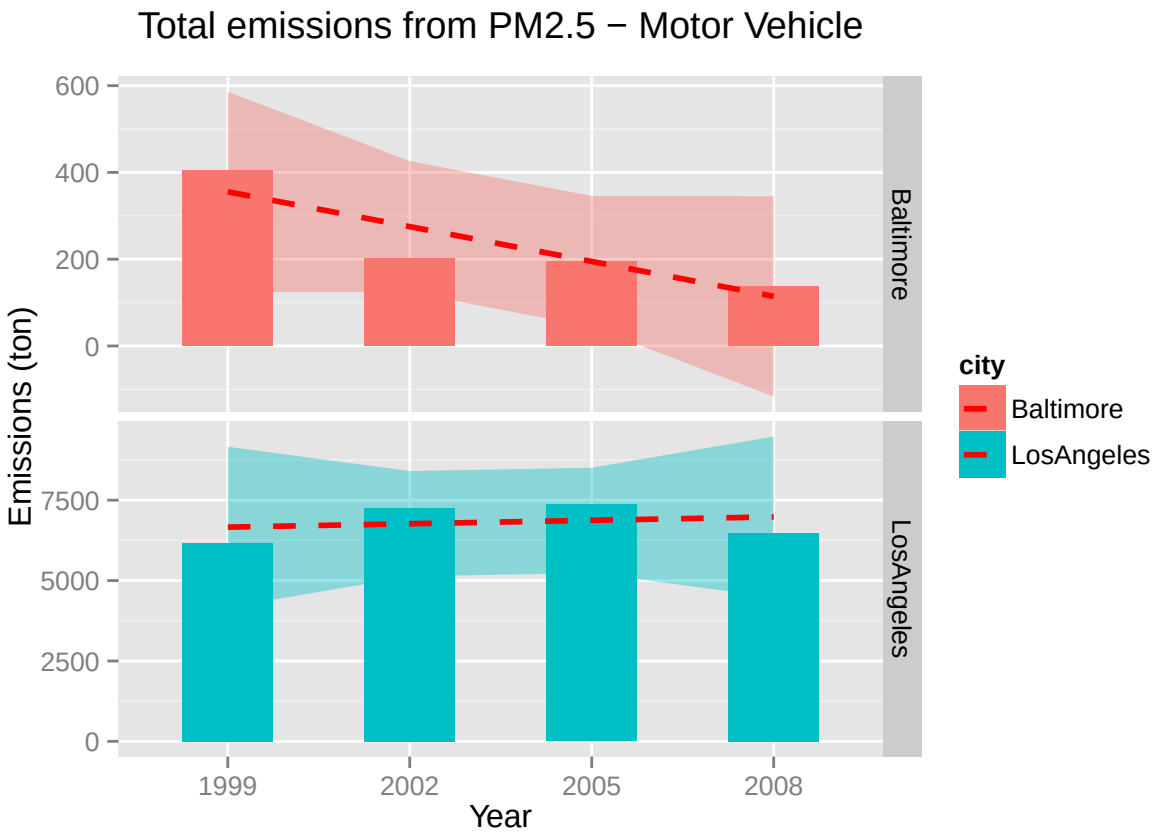
```
## [1] TRUE
```

```
#   There are 1395 observables for vehicle, in which 88 motor are included. Therefore,
#   vehicle.baltimore subset is the most completed for the question.

    g <- ggplot(vehicle.baltimore, aes(x=factor(year),y=Emissions,fill=type))
    plot5 <- g + geom_bar(stat = "identity", width=.5) +
            xlab('Year') + ylab('Emissions (ton)') +
            ggtitle('Baltimore City - Total Emissions from PM2.5 - Vehicle')+
            theme(plot.title = element_text(vjust=2))

#   plots <- plot5 + facet_wrap(~ type,nrow=3,ncol=1,scales = 'free_y')

#    png(filename="plot5.png",width = 480, height = 480)
    plot5
#    dev.off()

#
#   NOTE: Because of the controversy observed in the course FORUM,
#   I have included all type ('ON-ROAD' and the other types).
#   'POINT' type does not have any contribution.
#   It can observe that 'ON-ROAD' is the highest contributor for the total Emissions in
#   Baltimore city. Likewise 'ON-ROAD' type, the total emissions decreased ca. 50%
#   from 1999 to 2002. Small decrease (almost steady) on 2005, and a futher decrease in 2008.
#
#   End of Question 5.
#

############################
####    Question 6 : plot 6
############################
#
#   ASSUMING THAT THE QUESTION 5 CODE WAS EXECUTED
#   THE QUESTION 5 CODE IS DESCRIBED BELOW.
#
    LosAngeles <- county$"06037"
```

```
    motor_AND_vehicle.LosAngeles  <-filter(LosAngeles, SCC %in% SCC_motor_AND_vehicle)
    motor_OR_vehicle.LosAngeles <- filter(LosAngeles, SCC %in% SCC_motor_OR_vehicle)
    vehicle.LosAngeles <- filter(LosAngeles, SCC %in% SCC_vehicle)

#   By using dim, vehicle.baltimore and motor_OR_vehicle.baltimore have same size (1395 by 6).
#   The two data.frames are the same, by using:

    identical(vehicle.LosAngeles,motor_OR_vehicle.LosAngeles)
```

```
## [1] TRUE
```

```
#   There are 1328 observables for vehicle, in which 94 motor are included. Therefore,
#   vehicle.LosAngeles subset is the most completed for the question.

    vehicle.both <- rbind(vehicle.baltimore,vehicle.LosAngeles)

    city <- rep('Baltimore',4)
    total_emission_vehicle.baltimore <- cbind(city,with(vehicle.baltimore,aggregate(Emissions,list(year

#   BASED ON one of the evalutions:
#   test <- aggregate(Emissions ~ year + type + fips, vehicle.both, sum)

    city <- rep('LosAngeles',4)
    total_emission_vehicle.LosAngeles <- cbind(city,with(vehicle.LosAngeles,aggregate(Emissions,list(ye

    total.emission.vehicle.both <- rbind(total_emission_vehicle.baltimore,total_emission_vehicle.LosAng

    names(total.emission.vehicle.both) <- c('city','year','Emissions')

    g <- ggplot(total.emission.vehicle.both, aes(x=factor(year),y=Emissions,fill=city))
    plot6 <-    g + geom_bar(stat = "identity", width=.5) +
                facet_grid(city~., scales = 'free')+
                xlab('Year') + ylab('Emissions (ton)') +
                ggtitle('Total emissions from PM2.5 - Motor Vehicle')+
                theme(plot.title = element_text(vjust=2))+
                geom_smooth(method = "lm",aes(group = 1),se=TRUE,
                            colour = "red", size = 1,linetype=2)

#    png(filename="plot6.png",width = 480, height = 480)
    plot6
#    dev.off()

#
#   End of Question 6.
#
```


Total emissions from PM2.5 − Motor Vehicle
Emissions (ton)
600
400
200
0
7500
5000
2500
0
1999
2002
2005
2008
Year
Baltimore
LosAngeles
city
Baltimore
LosAngeles