Final_Assignment_Q6.Rmd

Please note that this code requires dplyr, reshape2 and ggplot2 packages and will attempt to install them if they are missing.

This assignment asked us to create a graph, from the PM25 pollution data posted by the EPA, to answer the following question :

Which city has seen greater changes over time in motor vehicle emissions (Baltimore or LA)?

Let us first begin by downloading the data and reading the data frames into R.

```
if(!require(dplyr)){
    install.packages("dplyr")
}

if(!require(reshape2)){
    install.packages("reshape2")
}

if(!require(ggplot2)){
    install.packages("ggplot2")
}

library(dplyr)

library(reshape2)

library(ggplot2)

URL<- "https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip"

download.file(url = URL, "PMdata.zip")

unzip("PMdata.zip")

pmdata<- readRDS("summarySCC_PM25.rds")

scc<- readRDS("Source_Classification_Code.rds")

head(pmdata)
```

```
##     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
```

```
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 Ve
nts, 2,4-D Recovery: Filtration",..: 3283 3284 3293 3291 3290 3294 3295 3296 3292 3289 ...
##  $ EI.Sector          : Factor w/ 59 levels "Agriculture - Crops & Livestock Dust",..: 18
 18 18 18 18 18 18 18 18 ...
##  $ 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 32 32 32 32 32 32 ...
##  $ SCC.Level.Three    : Factor w/ 1061 levels "","100% Biosolids (e.g., sewage sludge, man
ure, mixtures of these matls)",..: 88 88 156 156 156 156 156 156 156 156 ...
##  $ SCC.Level.Four     : Factor w/ 6084 levels "","(NH4)2 SO4 Acid Bath System and Evaporat
or",..: 4455 5583 4466 4458 1341 5246 5584 5983 4461 776 ...
##  $ 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 hood
s, filtrate tanks, vacuum pump exhausts",..: 1 1 1 1 1 1 1 1 1 1 ...
```

As seen from above **pmdata** contains the actual information on PM25 pollution, while **scc** is a code book, listing the various sources of said pollution.

Our first task is to work with the ***scc*** data frame to isolate all categories of emissions which relate to motor vehicle combustion. First we filter out all “Mobile Sources” of combustion in variable “SCC.Level.One”. Then we exclude Border Crossings, aircraft, marine vessels, railroad equipment and unknown in variable “SCC.Level.Two”. Finally, we exclude commercial, lawn and garden, logging (as these were petrol powered devices, not vehicles) from SCC.Level.Three.

```
code1<- scc[scc$SCC.Level.One=="Mobile Sources",]

l1<- unique(code1$SCC.Level.Two)[-c(2, 9, 10, 11, 12, 13, 14, 15, 16)] #exclude Border Crossi
ngs, aircraft, marine vessels, railroad equipmen and unknown.

code2<-code1[code1$SCC.Level.Two %in% l1,]

l2<-unique(code2$SCC.Level.Three)[-c(4, 21, 9, 19, 22)]    ##exclude commercial, lawn and gar
den, logging, as these were pertrol powered devices, not vehicles.

code3<-code2[code2$SCC.Level.Three %in% l2,]
```

Now that we have a list of codes related to vehicle sources, lets create two new data tables for PM25 emissions in Baltimore and Los Angeles.

```
baltimore<- pmdata[pmdata$fips == "24510",]

LA<- pmdata[pmdata$fips == "06037",]
```

Then, filter out emissions from vehicle sources

```
PM25a<- baltimore[baltimore$SCC %in% code3$SCC,]

PM25b<- LA[LA$SCC %in% code3$SCC,]
```

Next, we summarize and merge the data to get annual values of PM25 from motor vehicles for each city.

```
PM25a<- group_by(PM25a, year)

PM25b<- group_by(PM25b, year)

plota<-summarise(PM25a, PM25=sum(Emissions))

plotb<-summarise(PM25b, PM25=sum(Emissions))

plot<- merge(plota, plotb, by = "year")

final<-melt(plot, id.vars = "year")

final
```

```
##   year variable     value
## 1 1999   PM25.x  374.1000
## 2 2002   PM25.x  157.3189
## 3 2005   PM25.x  151.2089
## 4 2008   PM25.x  105.2857
## 5 1999   PM25.y 5823.4900
## 6 2002   PM25.y 6803.7902
## 7 2005   PM25.y 6930.0549
## 8 2008   PM25.y 6071.9665
```

Lastly, we plot the results.

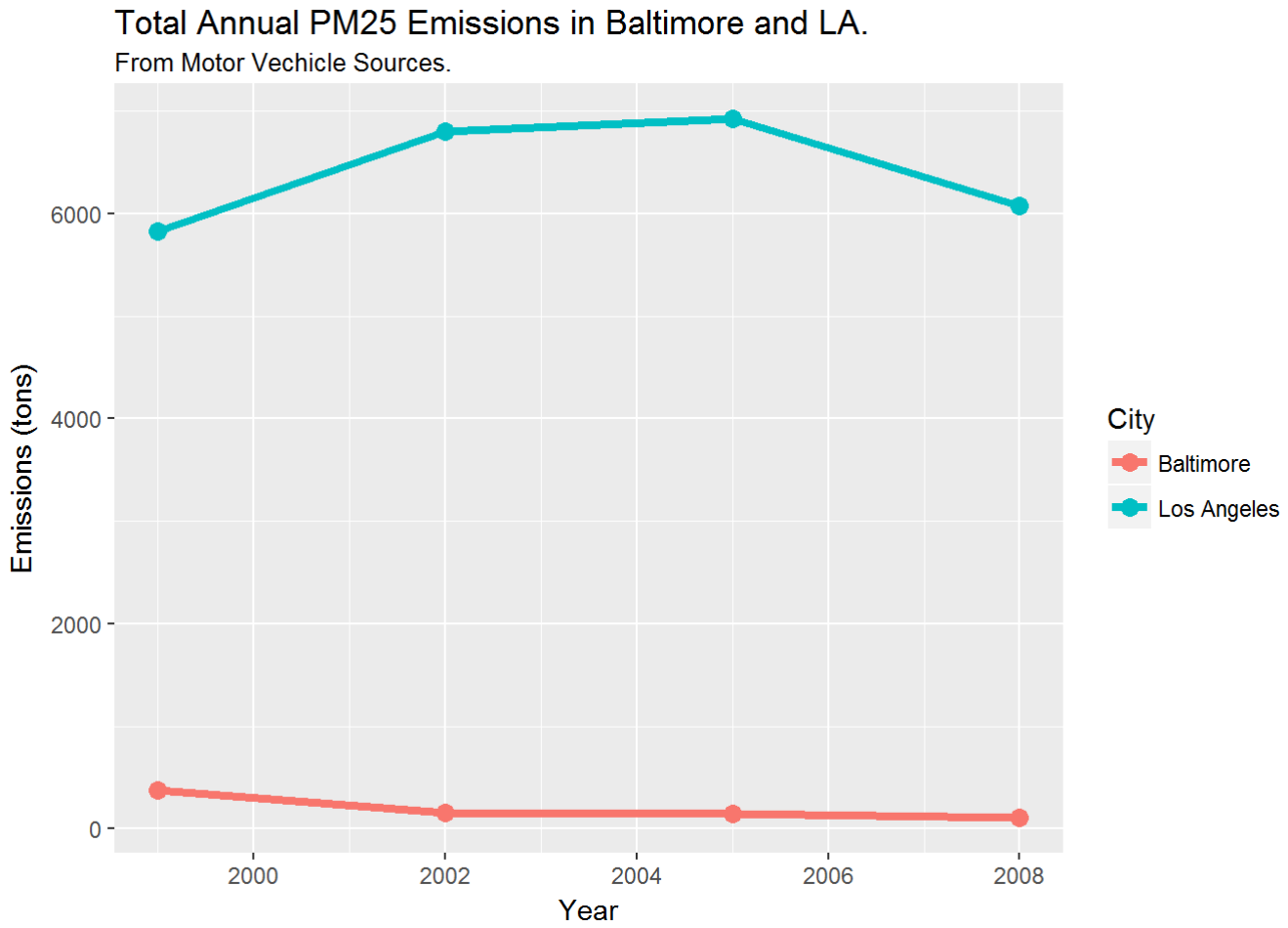
```
library(ggplot2)

ggplot(final, aes(year, value)) + geom_point(aes(col = variable), size=3)+

    geom_line(aes(col = variable), size=1.5)+

    labs(title="Total Annual PM25 Emissions in Baltimore and LA.", subtitle="From Motor Vechi
cle Sources.")+

    xlab("Year") + ylab("Emissions (tons)") + scale_colour_discrete(name="City", labels =
c("Baltimore", "Los Angeles"))
```



As can be seen from the graph above, Los Angeles has much higher levels and a greater change over the years in PM25 pollution from motor vehicle sources as compared to Baltimore.