Code

GR5223 Project

Team members:

Xinyi Hu: xh2383

Xinge Jia: xj2221

Nikita Tourani:nrt2117

1. What is your main question/topic of interest for the group project? Why have you chosen this question/topic?

The data set that we are going to use is downloaded from https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data (https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data), which includes the value of PM2.5 varied through time and some related predictor variables. PM2.5 refers to particulate matter in the atmosphere that have a diameter of less than 2.5 micrometers, and is widely regarded as an air pollutant that can be hazardous to human health. In this data set, hourly PM2.5 data of US Embassy in Beijing has been collected. The data set also includes seven other meteorological data attributes such as dew point, pressure, temperature, cumulated hours of rain etc.

In this group project, our aim is to understand how these factors influence (interact with) the value of PM2.5 and learn about the correlation between the value of PM2.5 and time. Based on these results, we can try to propose some policy ideas and practical solutions for PM2.5 reduction and prevention.

The reason why we are interested in these questions is that the air pollution problem has been a key cause of public concern, and has remained an unsolved issue in China for years. Two members of our team are from China and both of them have suffered from very unhealthy levels of PM2.5. Thus, we would like to examine the PM2.5 issue statistically and attempt to find a possible remedy for this issue.

2. What types of information/data are you planning on using in your project?

The data set that we are going to use is a multivariate data set including both categorial data and numerical data. Besides, this is a time-series data recorded per hour per day from 2010 to 2014. There are eight atmospheric attributes measured, including PM2.5 concentration.

3. How are your planning to carry out the analysis of the data/the project, i.e. which software/methods do you plan to use and, if applicable, what are you expecting to observe?

In this project, we will use R to perform the analysis. There are several methods that might be useful in building and selecting the appropriate model. The transformation of response variable and predictor variables, variable selection according to the Mallows Cp criterion and the value of AIC as well as multiple R-squared and hypothesis testing can be applied during our analysis. What's more, since it is a time-series data set, we will apply time series related methodologies to find the trend and make some predictions to complete the analysis. We are expecting to have the model that builds a relationship between the value of PM2.5 and other related factors.

4. Ideas:

Separate PM2.5 values into different levels according to Air Quality Guide for PM2.5 https://www.cnn.com/2017/05/04/asia/beijing-sand-storm-pollution-beyond-index/index.html (https://www.cnn.com/2017/05/04/asia/beijing-sand-storm-pollution-beyond-index/index.html), and the cluster months in each of the four years, to see the pattern of how PM2.5 values change according to the period of the year. Similarly, we will find out how the PM2.5 value changes within a 24 hour period. (Using some simple data exploring skills and sorting the data)

We will then summarize some overview statistics about PM2.5, like the number of days under each PM2.5 level.

We would also like to examine how other factors influence the levels of PM2.5 (ie. Dew, pressure, temperature, wind direction, wind speed, snow, rain). We will do this by analyzing their correlation with the PM2.5 value, do regression/classification, and train the model (cross-validation).

After understanding how these factors influence PM2.5 values, we will try to propose some approaches/policies that could lower the value of PM2.5 (ie. improve air quality).

There is some missing data, and we would like to filter the data before we start our analyses:

Hide

```
#setwd("../")
library(data.table)
pm = fread("PRSA_data_2010.1.1-2014.12.31.csv")
na_check = apply(is.na(pm),2,sum);na_check
```

```
   No  year month   day  hour pm2.5  DEWP  TEMP  PRES  cbwd   Iws    Is    Ir 
    0     0     0     0     0  2067     0     0     0     0     0     0     0 
```

Hide

```
del = which(is.na(pm$pm2.5) == TRUE)
pm_delete_na = pm[-del,]
apply(is.na(pm_delete_na),2,sum)
```

```
   No  year month   day  hour pm2.5  DEWP  TEMP  PRES  cbwd   Iws    Is    Ir 
    0     0     0     0     0     0     0     0     0     0     0     0     0 
```

Hide

```
pm25 = pm_delete_na[,-1]
dim(pm25)
```

```
[1] 41757    12
```

Hide

```
range(pm25$pm2.5)
```

```
[1]   0 994
```