Report

Collin Carmichael; Jennings Cheng; Zhengtao Xu

7/9/2021

# Introduction

In this project we are going to construct models to predict PM2.5. Data comes from <https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>. This data set includes PM2.5 (fine particulate matter), concentrations with a number of meteorological variables for the city of Beijing during the years 2010 to 2014. PM2.5 can be used to monitor PM2.5 pollution in a location and the meteorological conditions can influence the PM2.5 concentrations.

To archive the prediction goal, we will first conduct the EDA analysis, in this part we will summary the basic statistics on the time series of PM2.5 data, as well as the relevant impact factor or signal factors.

After this, the we will construct prediction models, from the multiply linear regression to other regression analysis.

Finally, we will summary the result.

# EDA

Exploratory Data Analysis will be conducted in this section, we make some graphical displays of the data and numerical summaries of the data.

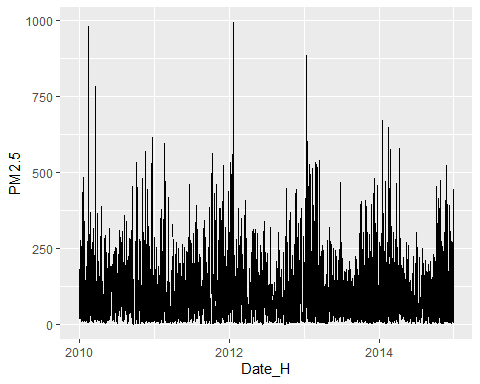
## No Year Month Day Hour PM2.5 DEWP TEMP PRES CBWD LWS LS LR  
## 1 1 2010 1 1 0 NA -21 -11.00 1021.00 NW 1.79 0 0  
## 2 2 2010 1 1 1 NA -21 -12.00 1020.00 NW 4.92 0 0  
## 3 3 2010 1 1 2 NA -21 -11.00 1019.00 NW 6.71 0 0  
## 4 4 2010 1 1 3 NA -21 -14.00 1019.00 NW 9.84 0 0  
## 5 5 2010 1 1 4 NA -20 -12.00 1018.00 NW 12.97 0 0  
## 6 6 2010 1 1 5 NA -19 -10.00 1017.00 NW 16.10 0 0

## [1] "No" "Year" "Month" "Day" "Hour" "PM2.5" "DEWP" "TEMP" "PRES"   
## [10] "CBWD" "LWS" "LS" "LR"

## No Year Month Day   
## Min. : 1 Min. :2010 Min. : 1.000 Min. : 1.00   
## 1st Qu.:10957 1st Qu.:2011 1st Qu.: 4.000 1st Qu.: 8.00   
## Median :21913 Median :2012 Median : 7.000 Median :16.00   
## Mean :21913 Mean :2012 Mean : 6.524 Mean :15.73   
## 3rd Qu.:32868 3rd Qu.:2013 3rd Qu.:10.000 3rd Qu.:23.00   
## Max. :43824 Max. :2014 Max. :12.000 Max. :31.00   
##   
## Hour PM2.5 DEWP TEMP   
## Min. : 0.00 Min. : 0.00 Min. :-40.000 Length:43824   
## 1st Qu.: 5.75 1st Qu.: 29.00 1st Qu.:-10.000 Class :character   
## Median :11.50 Median : 72.00 Median : 2.000 Mode :character   
## Mean :11.50 Mean : 98.61 Mean : 1.817   
## 3rd Qu.:17.25 3rd Qu.:137.00 3rd Qu.: 15.000   
## Max. :23.00 Max. :994.00 Max. : 28.000   
## NA's :2067   
## PRES CBWD LWS LS   
## Length:43824 Length:43824 Min. : 0.45 Min. : 0.00000   
## Class :character Class :character 1st Qu.: 1.79 1st Qu.: 0.00000   
## Mode :character Mode :character Median : 5.37 Median : 0.00000   
## Mean : 23.89 Mean : 0.05273   
## 3rd Qu.: 21.91 3rd Qu.: 0.00000   
## Max. :585.60 Max. :27.00000   
##   
## LR   
## Min. : 0.0000   
## 1st Qu.: 0.0000   
## Median : 0.0000   
## Mean : 0.1949   
## 3rd Qu.: 0.0000   
## Max. :36.0000   
##

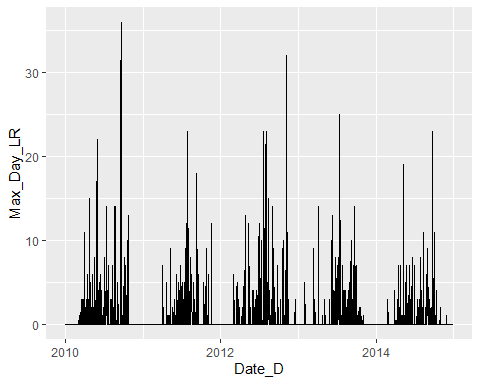
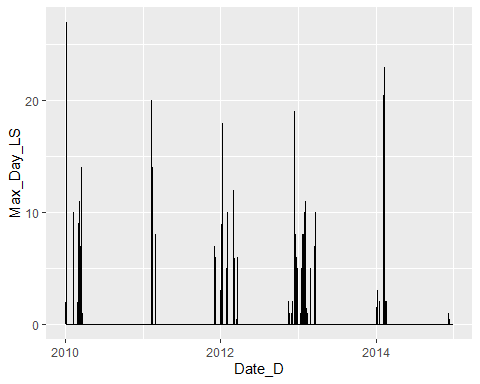
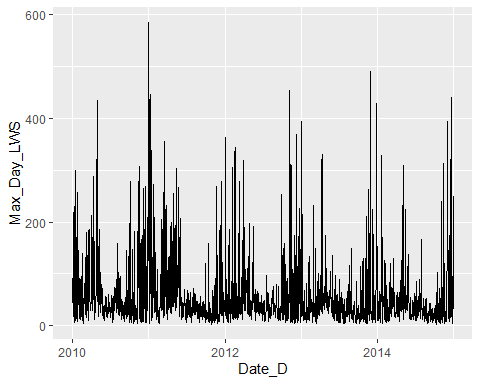
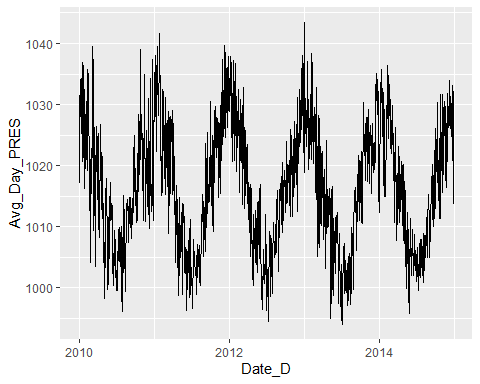
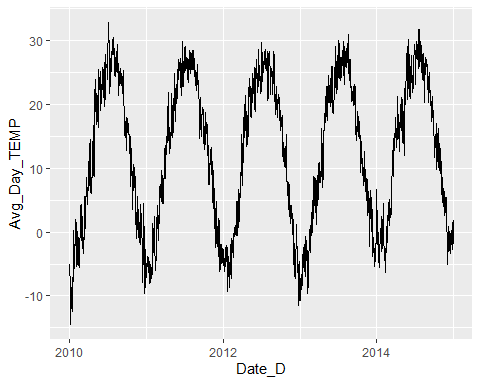
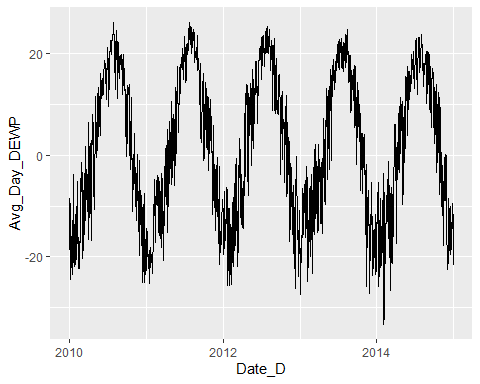
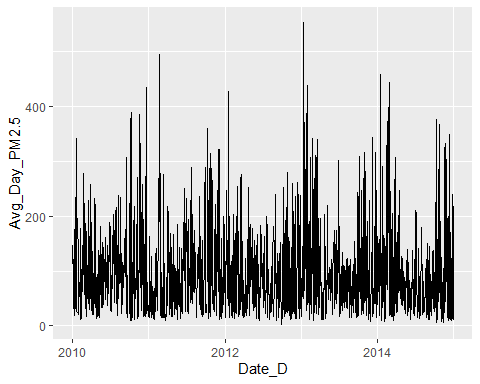
## No Year Month Day   
## Min. : 1 Min. :2010 Min. : 1.000 Min. : 1.00   
## 1st Qu.:10957 1st Qu.:2011 1st Qu.: 4.000 1st Qu.: 8.00   
## Median :21913 Median :2012 Median : 7.000 Median :16.00   
## Mean :21913 Mean :2012 Mean : 6.524 Mean :15.73   
## 3rd Qu.:32868 3rd Qu.:2013 3rd Qu.:10.000 3rd Qu.:23.00   
## Max. :43824 Max. :2014 Max. :12.000 Max. :31.00   
##   
## Hour PM2.5 DEWP TEMP   
## Min. : 0.00 Min. : 0.00 Min. :-40.000 Min. :-19.00   
## 1st Qu.: 5.75 1st Qu.: 29.00 1st Qu.:-10.000 1st Qu.: 2.00   
## Median :11.50 Median : 72.00 Median : 2.000 Median : 14.00   
## Mean :11.50 Mean : 98.61 Mean : 1.817 Mean : 12.45   
## 3rd Qu.:17.25 3rd Qu.:137.00 3rd Qu.: 15.000 3rd Qu.: 23.00   
## Max. :23.00 Max. :994.00 Max. : 28.000 Max. : 42.00   
## NA's :2067 NA's :2   
## PRES CBWD LWS LS   
## Min. : 991 Length:43824 Min. : 0.45 Min. : 0.00000   
## 1st Qu.:1008 Class :character 1st Qu.: 1.79 1st Qu.: 0.00000   
## Median :1016 Mode :character Median : 5.37 Median : 0.00000   
## Mean :1016 Mean : 23.89 Mean : 0.05273   
## 3rd Qu.:1025 3rd Qu.: 21.91 3rd Qu.: 0.00000   
## Max. :1046 Max. :585.60 Max. :27.00000   
## NA's :2   
## LR   
## Min. : 0.0000   
## 1st Qu.: 0.0000   
## Median : 0.0000   
## Mean : 0.1949   
## 3rd Qu.: 0.0000   
## Max. :36.0000   
##

## No Year Month Day Hour PM2.5 DEWP TEMP PRES CBWD LWS LS LR  
## 1 1 2010 1 1 0 NA -21 -11 1021 NW 1.79 0 0  
## 2 2 2010 1 1 1 NA -21 -12 1020 NW 4.92 0 0  
## 3 3 2010 1 1 2 NA -21 -11 1019 NW 6.71 0 0  
## 4 4 2010 1 1 3 NA -21 -14 1019 NW 9.84 0 0  
## 5 5 2010 1 1 4 NA -20 -12 1018 NW 12.97 0 0  
## Date\_H Date\_D  
## 1 2010-01-01 00:00:00 2010-01-01  
## 2 2010-01-01 01:00:00 2010-01-01  
## 3 2010-01-01 02:00:00 2010-01-01  
## 4 2010-01-01 03:00:00 2010-01-01  
## 5 2010-01-01 04:00:00 2010-01-01



## 'data.frame': 1826 obs. of 9 variables:  
## $ Date\_D : Date, format: "2010-01-01" "2010-01-02" ...  
## $ Avg\_Day\_PM2.5: num NaN 146 78.8 31.3 42.5 ...  
## $ Avg\_Day\_DEWP : num -18.8 -8.5 -10.1 -20.9 -24.6 ...  
## $ Avg\_Day\_TEMP : num -6.75 -5.12 -8.54 -11.5 -14.46 ...  
## $ Avg\_Day\_PRES : num 1017 1025 1023 1029 1034 ...  
## $ Max\_Day\_LWS : num 43.8 55.4 127.8 198.4 218.6 ...  
## $ Max\_Day\_LS : num 0 4 27 0 0 0 0 0 0 0 ...  
## $ Max\_Day\_LR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Max\_Day\_CBWD : chr "NW" "SE" "SE" "NW" ...

## Date\_D Avg\_Day\_PM2.5 Avg\_Day\_DEWP Avg\_Day\_TEMP   
## Min. :2010-01-01 Min. : 2.958 Min. :-33.333 Min. :-14.458   
## 1st Qu.:2011-04-02 1st Qu.: 42.231 1st Qu.:-10.083 1st Qu.: 1.542   
## Median :2012-07-01 Median : 79.167 Median : 2.021 Median : 13.896   
## Mean :2012-07-01 Mean : 98.653 Mean : 1.817 Mean : 12.448   
## 3rd Qu.:2013-09-30 3rd Qu.:131.167 3rd Qu.: 15.073 3rd Qu.: 23.167   
## Max. :2014-12-31 Max. :552.478 Max. : 26.208 Max. : 32.875   
## NA's :37   
## Avg\_Day\_PRES Max\_Day\_LWS Max\_Day\_LS Max\_Day\_LR   
## Min. : 994 Min. : 2.69 Min. : 0.0000 Min. : 0.0000   
## 1st Qu.:1008 1st Qu.: 20.68 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Median :1016 Median : 36.23 Median : 0.0000 Median : 0.0000   
## Mean :1016 Mean : 59.97 Mean : 0.2218 Mean : 0.9617   
## 3rd Qu.:1025 3rd Qu.: 64.37 3rd Qu.: 0.0000 3rd Qu.: 0.0000   
## Max. :1043 Max. :585.60 Max. :27.0000 Max. :36.0000   
##   
## Max\_Day\_CBWD   
## Length:1826   
## Class :character   
## Mode :character   
##   
##   
##   
##



## Date\_D Avg\_Day\_PM2.5 Avg\_Day\_DEWP Avg\_Day\_TEMP Avg\_Day\_PRES Max\_Day\_LWS  
## 1 2010-01-01 NaN -18.75000 -6.750000 1017.083 43.82  
## 2 2010-01-02 145.95833 -8.50000 -5.125000 1024.750 55.43  
## 3 2010-01-03 78.83333 -10.12500 -8.541667 1022.792 127.84  
## 4 2010-01-04 31.33333 -20.87500 -11.500000 1029.292 198.45  
## 5 2010-01-05 42.45833 -24.58333 -14.458333 1033.625 218.57  
## 6 2010-01-06 56.41667 -23.70833 -12.541667 1033.750 59.00  
## Max\_Day\_LS Max\_Day\_LR Max\_Day\_CBWD Year Month Day  
## 1 0 0 NW 2010 1 1  
## 2 4 0 SE 2010 1 2  
## 3 27 0 SE 2010 1 3  
## 4 0 0 NW 2010 1 4  
## 5 0 0 NW 2010 1 5  
## 6 0 0 NE 2010 1 6