# PM2.5 level in Chengdu - Exploratory Data Analysis

*Pierre Baudin*

*March 17, 2018*

Note: the complete code will be available on my github repository to support reproducible research. (Full code repo)

## Synopsis

Cities accross the world are facing air pollution problems that are generated by human activity. Air pollution has been recognized to be a factor in the increase of respiratory diseases in urban population.

Increasingly, cities are developing programs to reduce the air pollution and decrease the health risk on the population.

In this explanatory data analysis, we will look at the city of Chengdu and its PM2.5 pollution level. PM2.5 stands for Particule Matter and describes fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller.

The data for this project is downloaded from the American embassy in Chengdu. (see: http://www.stateair.net/web/post/1/2.html). We are grateful for the U.S. Department of State to open access to the data. The website also warns that these data are not fully verified or validated.

The objective of this project is to explore the dataset, identify trends and finally propose a forecasting model.

## Data preparation

### Collection and combining

The historical data from the American embassy in Chengdu are available for download. For each year from 2013 to 2017, there is one comma separated value file (.csv). All of the files have been downloaded.

A preliminary check (not shown here) has revealed that the data formating in the csv file has been changed between 2014 and 2015. For this reason, the import and binding is done in two stage.

```
## [1] "Structure of the dataset"
```

```
## 'data.frame':    39408 obs. of  11 variables:
##  $ Site      : Factor w/ 1 level "Chengdu": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Parameter : Factor w/ 1 level "PM2.5": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Date..LST.: POSIXct, format: "2013-01-01 00:00:00" "2013-01-01 01:00:00" ...
##  $ Year      : Factor w/ 5 levels "2013","2014",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Month     : Factor w/ 12 levels "Jan","Feb","Mar",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Day       : Factor w/ 31 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Hour      : Factor w/ 24 levels "0","1","2","3",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Value     : num  129 135 -999 -999 -999 -999 -999 -999 -999 -999 ...
##  $ Unit      : Factor w/ 1 level "µg/m³": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Duration  : Factor w/ 1 level "1 Hr": 1 1 1 1 1 1 1 1 1 1 ...
##  $ QC.Name   : Factor w/ 2 levels "Missing","Valid": 2 2 1 1 1 1 1 1 1 1 ...
```

/newpage

```
## [1] "First 6 rows of the dataset"
```

```
##      Site Parameter          Date..LST. Year Month Day Hour Value  Unit
## 1 Chengdu     PM2.5 2013-01-01 00:00:00 2013   Jan   1    0   129 µg/m³
## 2 Chengdu     PM2.5 2013-01-01 01:00:00 2013   Jan   1    1   135 µg/m³
## 3 Chengdu     PM2.5 2013-01-01 02:00:00 2013   Jan   1    2  -999 µg/m³
## 4 Chengdu     PM2.5 2013-01-01 03:00:00 2013   Jan   1    3  -999 µg/m³
## 5 Chengdu     PM2.5 2013-01-01 04:00:00 2013   Jan   1    4  -999 µg/m³
## 6 Chengdu     PM2.5 2013-01-01 05:00:00 2013   Jan   1    5  -999 µg/m³
##   Duration QC.Name
## 1     1 Hr   Valid
## 2     1 Hr   Valid
## 3     1 Hr Missing
## 4     1 Hr Missing
## 5     1 Hr Missing
## 6     1 Hr Missing
```

**Interpretation of the dataset**

```
##       Site       Parameter       Date..LST.                    Year
##  Chengdu:39408   PM2.5:39408   Min.   :2013-01-01 00:00:00   2013:8760
##                                1st Qu.:2014-02-15 11:45:00   2014:8760
##                                Median :2015-04-01 23:30:00   2015:8760
##                                Mean   :2015-04-01 23:30:00   2016:8784
##                                3rd Qu.:2016-05-16 11:15:00   2017:4344
##                                Max.   :2017-06-30 23:00:00
##
##      Month             Day             Hour           Value
##  Jan    : 3720   1      : 1296   3      : 1647   Min.   :-999.00
##  Mar    : 3720   2      : 1296   0      : 1642   1st Qu.:  39.00
##  May    : 3720   3      : 1296   1      : 1642   Median :  62.00
##  Apr    : 3600   4      : 1296   4      : 1642   Mean   :  23.72
##  Jun    : 3600   5      : 1296   5      : 1642   3rd Qu.:  97.00
##  Feb    : 3384   6      : 1296   6      : 1642   Max.   : 688.00
##  (Other):17664   (Other):31632   (Other):29551
##     Unit       Duration        QC.Name
##  µg/m³:39408   1 Hr:39408   Missing: 2036
##                             Valid  :37372
##
##
##
##
##
```

We can observe that the dataset is only for Chengdu as a site and the PM2.5 as a parameter. Observations are logged on an hourly basis. Interestingly we note that the number of observations for the year 2017 is incomplete. For the year 2013 to 2016, the number of observations correspond exactly to the number of days in the year multiplied by 24 hours.

Regarding the value measured, we see that the minimum is -999.00 which has no meaning. There are also 2036 observations with QC.Name as “Missing”.

```
##
## Missing   Valid
##    2036       0
```

We can now confirm that the value recorded for missing QC is -999.

We choose to replace these missing value using a linear interpolation method. From then on, we will ignore the QC.Name variable.

This is now the summary of our final dataset ready for exploratory data analysis.

```
##       Site       Parameter       Date..LST.                    Year
##  Chengdu:39408   PM2.5:39408   Min.   :2013-01-01 00:00:00   2013:8760
##                                1st Qu.:2014-02-15 11:45:00   2014:8760
##                                Median :2015-04-01 23:30:00   2015:8760
##                                Mean   :2015-04-01 23:30:00   2016:8784
##                                3rd Qu.:2016-05-16 11:15:00   2017:4344
##                                Max.   :2017-06-30 23:00:00
##
##      Month            Day             Hour           Value
##  Jan    : 3720   1      : 1296   3      : 1647   Min.   :  0.00
##  Mar    : 3720   2      : 1296   0      : 1642   1st Qu.: 42.00
##  May    : 3720   3      : 1296   1      : 1642   Median : 65.00
##  Apr    : 3600   4      : 1296   4      : 1642   Mean   : 79.84
##  Jun    : 3600   5      : 1296   5      : 1642   3rd Qu.:101.00
##  Feb    : 3384   6      : 1296   6      : 1642   Max.   :688.00
##  (Other):17664   (Other):31632   (Other):29551
##     Unit       Duration
##  µg/m³:39408   1 Hr:39408
##
##
##
##
##
##
```

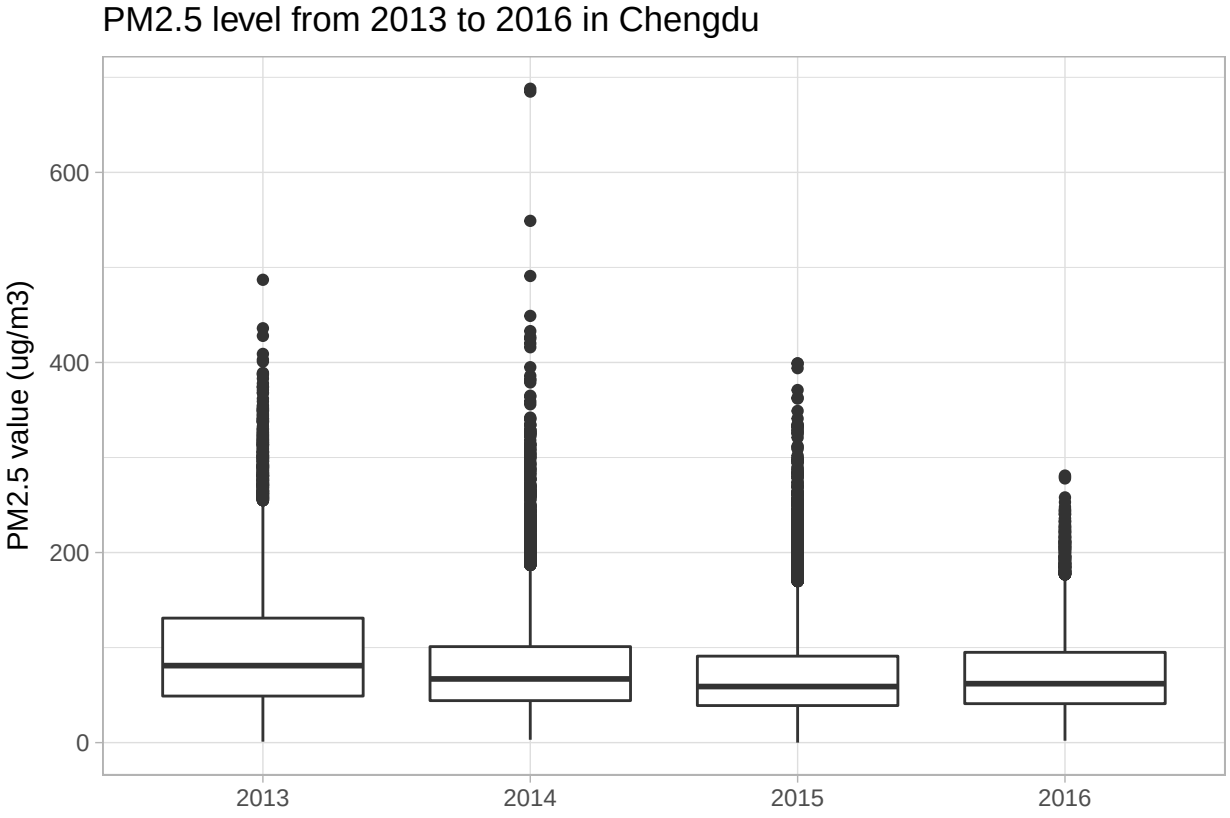
For the period of 2013 to mid-2017, the PM2.5 levels in Chengdu have a mean of 79.84 ug/m3, a median of 65.00 ug/m3. Peaks are observed at a maximum of 688.00 ug/m3 and a minimum of 0.00 ug/m3. We note that these value are extreme and need to be taken with caution. Elements such a measuring equipment, location and calibration could influence such readings.

## Visualization and Trends

### By Year

In this section, we will explore the PM2.5 level on a yearly basis and answer the following question: overall, have the PM2.5 pollution levels decrease over time in Chengdu?

Note: the year 2017 being incomplete, it will be ignored in this part of the analysis to avoid misinterpretation.




```
## # A tibble: 4 x 6
##   Year  YearMean YearMedian YearStd YearMax YearMin
##   <fct>    <dbl>      <dbl>   <dbl>   <dbl>   <dbl>
## 1 2013      97.8       81.0    65.1     487    1.00
## 2 2014      81.1       67.0    54.1     688    3.00
## 3 2015      72.6       59.0    48.8     399    0
## 4 2016      72.7       62.0    42.8     281    2.00
```

**Interpretation:**

We can see that over the year, the average level of PM2.5 in the air has decreased overall. It seems to stabilize in 2016. However, the standard deviation indicates that there are less extreme readings in 2016 which the maximum value recorded in that year confirms.

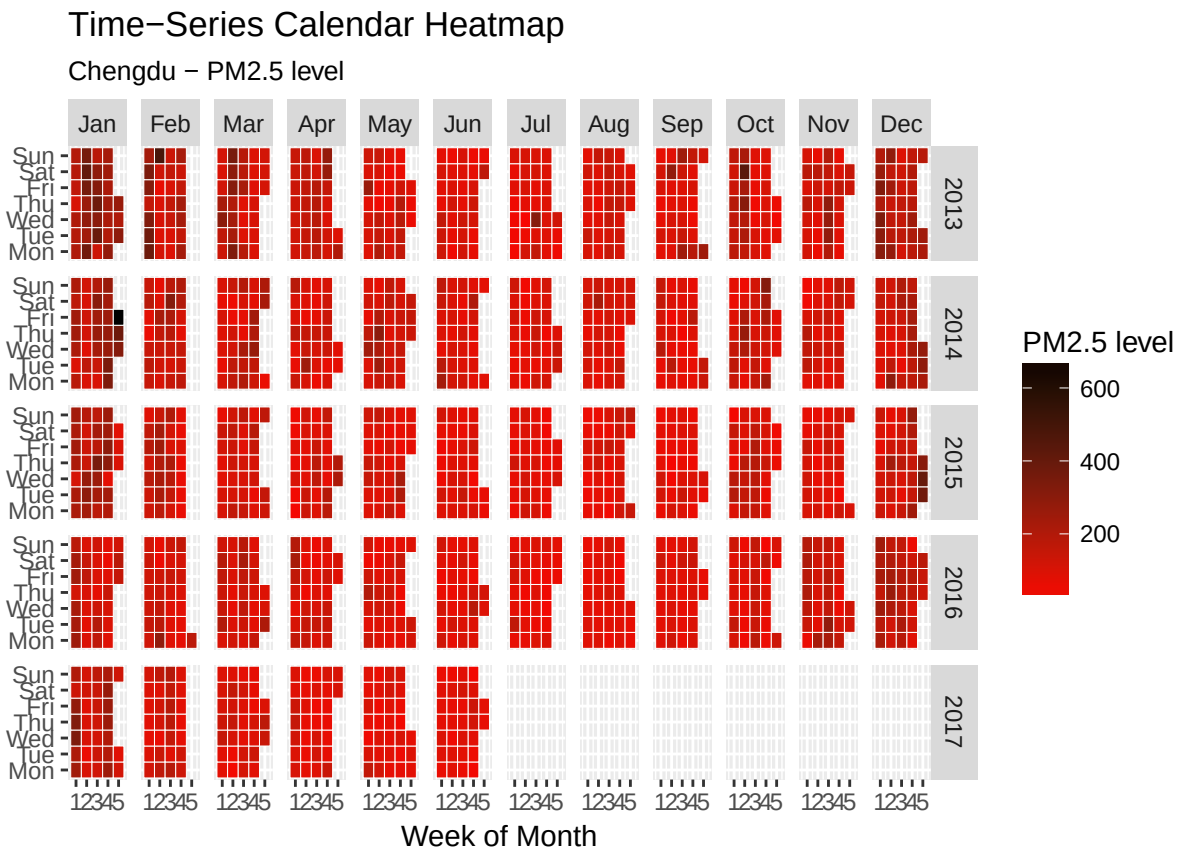
### Seasonality

In this section, we will look at the seasonality of the PM2.5 level accross year, months and weeks. The aim is to identify patterns and trends that will be useful in the construction of a model.

### Heatmap

The heatmap shows for each week day, each month and each year, the maximum PM2.5 level measured during that day.

The year 2017 although incomplete is included in this analysis to make use of the monthly data.




### Interpretation:

The months of from January to April and October to December record higher pollution level compare to the period from June to September. January and December appear to be the worst month of the year each year.

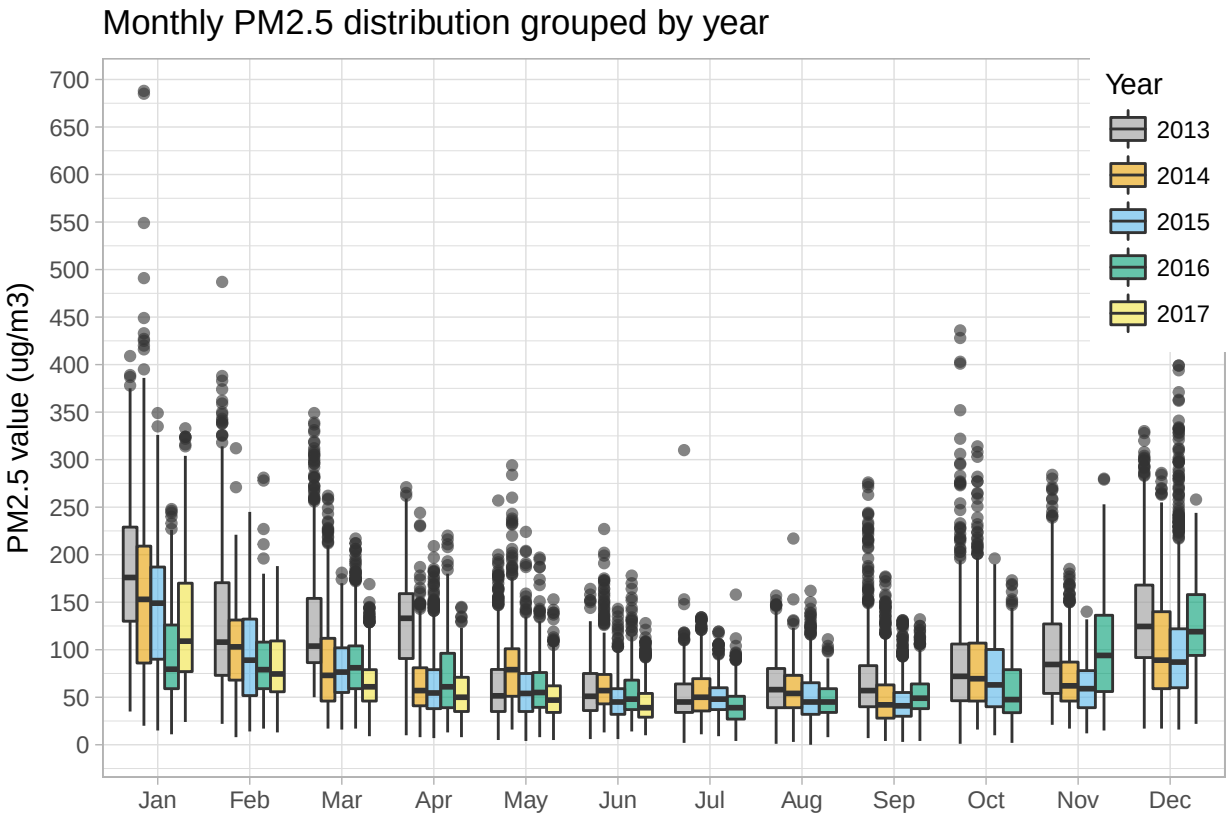
We can also observe that over the year it seems that the highest level of PM2.5 recorded have decreased i.e. extreme values are less extreme.

In the following paragraph, we will investigate these findings.

### Trend and peaks

In this section, we will look at the monthly data to investigate trends over the year.

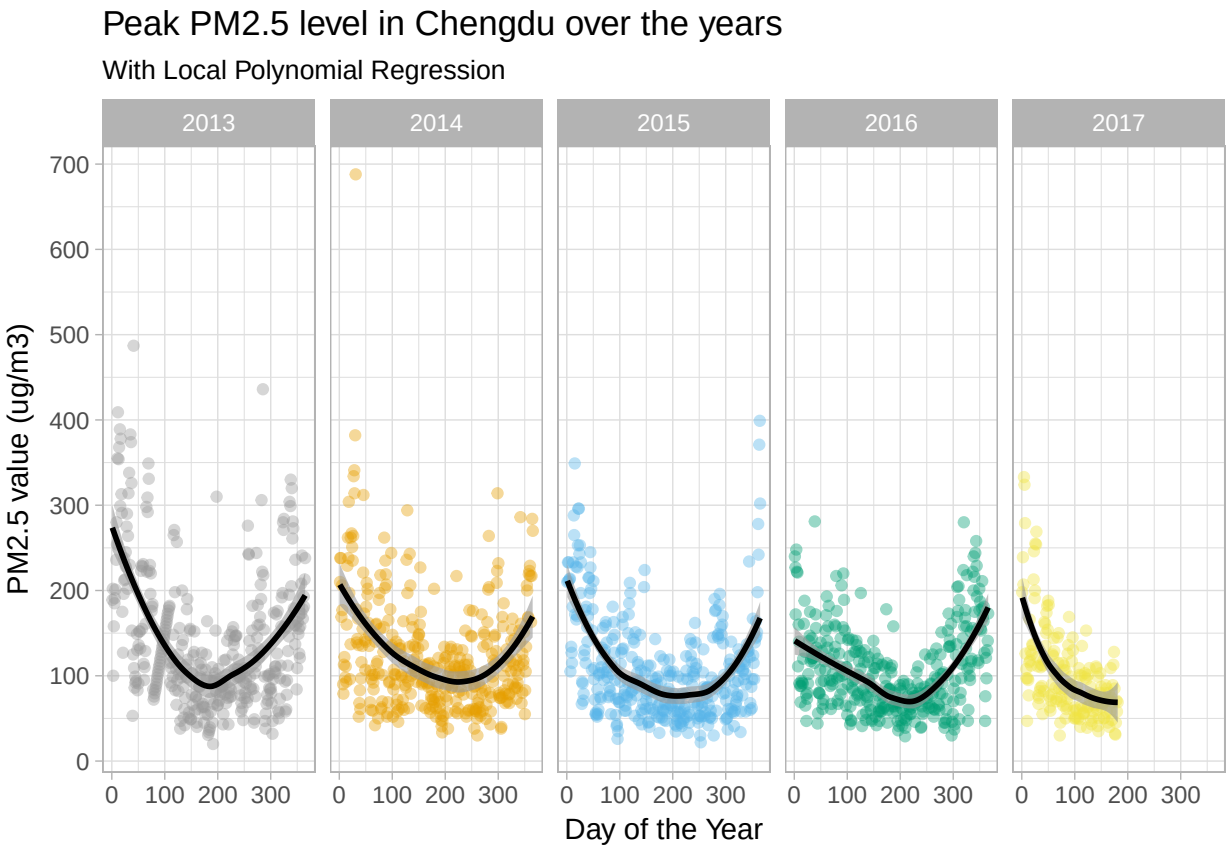
The table containing the monthly summary is available in the annexes section.



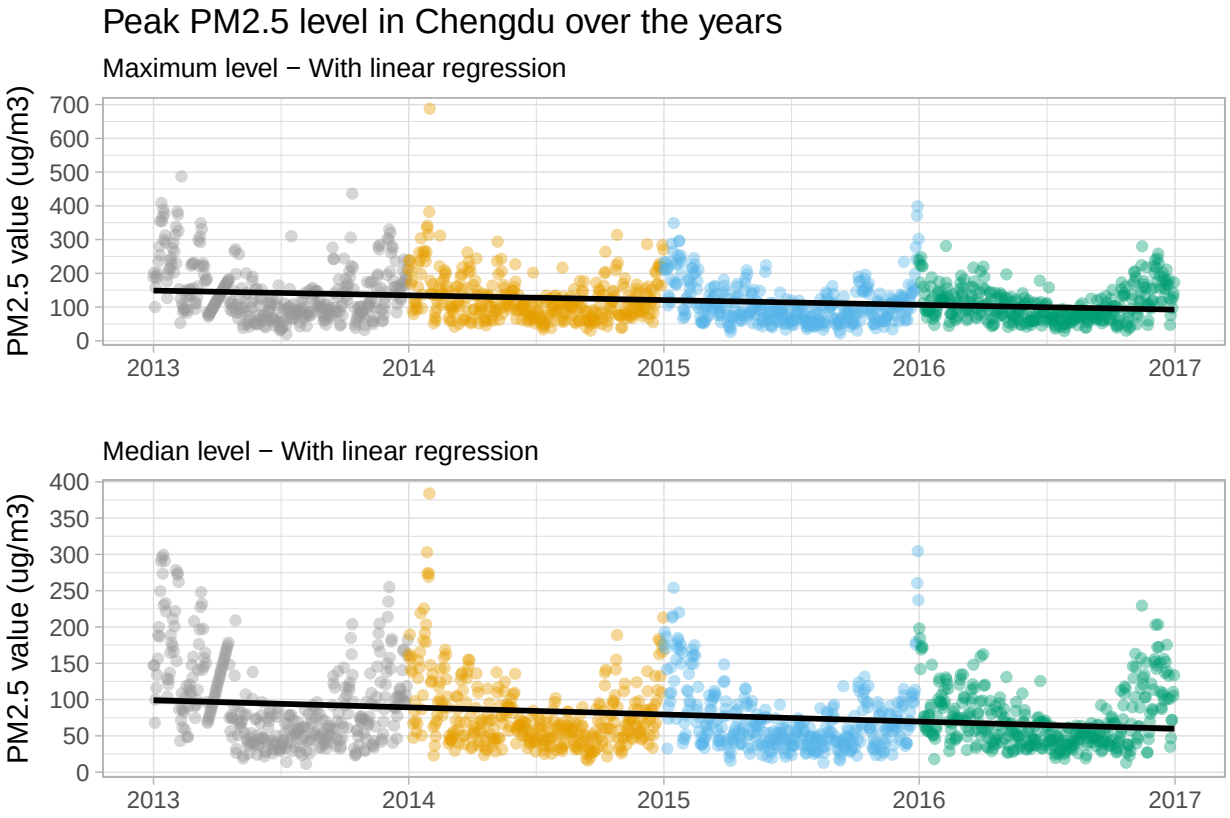

**Interpretation:**

We can now confirm the observations from the heatmap. The period from October to March shows a rise in pollution level with a peak in January, followed by a decline afterwards. The months with the lowest pollution level and lowest extreme values are July and August.

These trends are visible accross years. We now look at the yearly trend and fit a regression line to highlight the trends. The folowing graph describes the maximum values recorded each day and identify the trends for each year and over the years.




The regression model helps to confirm the visual interpretation of the data. We can now confirm that hyperbolic shape of the PM2.5 level for each years.

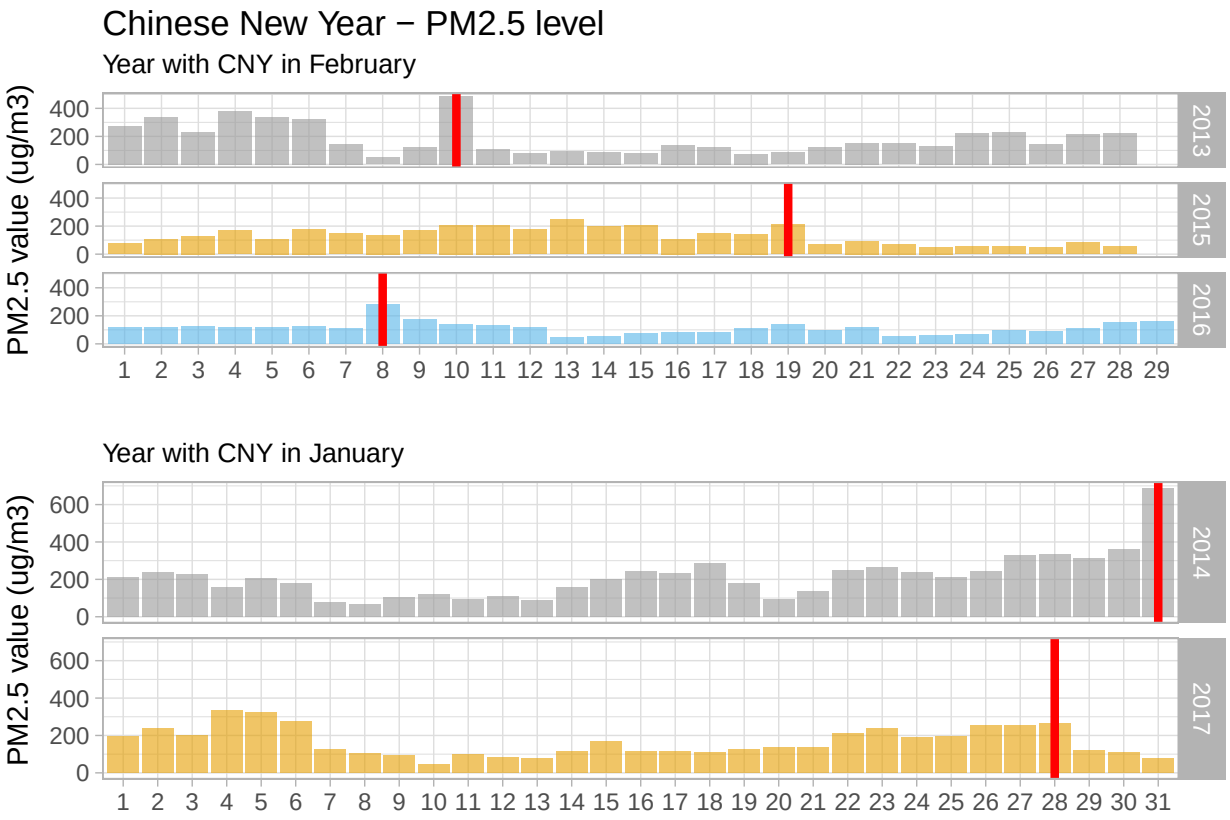



For this part, the year 2017 has been ignored to avoid skewed the model.

The linear regression line over the entire dataset from 2013 to 2016 presents a negtive slope for both the maximum values and median values. This indicates a reduction over time of median and extreme PM2.5 pollution level in Chengdu.

### Chinese New Year and PM2.5 level

In China, an interesting period to study is the Chinese New Year. This period of the year is describe as a modern migration phenomenon as Chinese people uses transport to enjoy this time in their hometown or travel to visit places. This is also the time where fireworks are used to celebrate the New Year. The following graph identify the Chinese New Year day and the corresponding maximum pollution level recorded.




### Interpretation

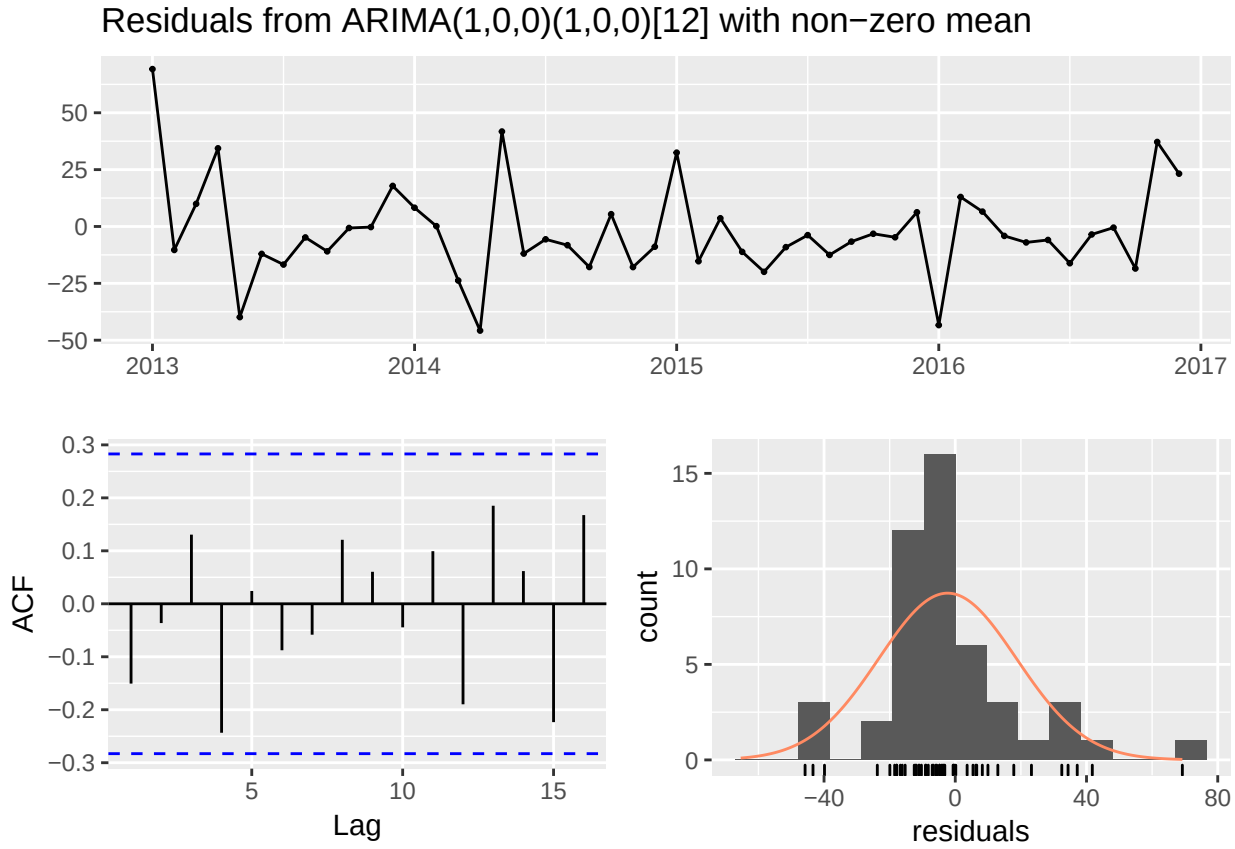
The graph seems to indicate that after 2014, the PM2.5 pollution level during Chinese New Year are comparable to average level in the city of Chengdu. This might be the result of a policy change by the Chengdu government.

### Forecasting model

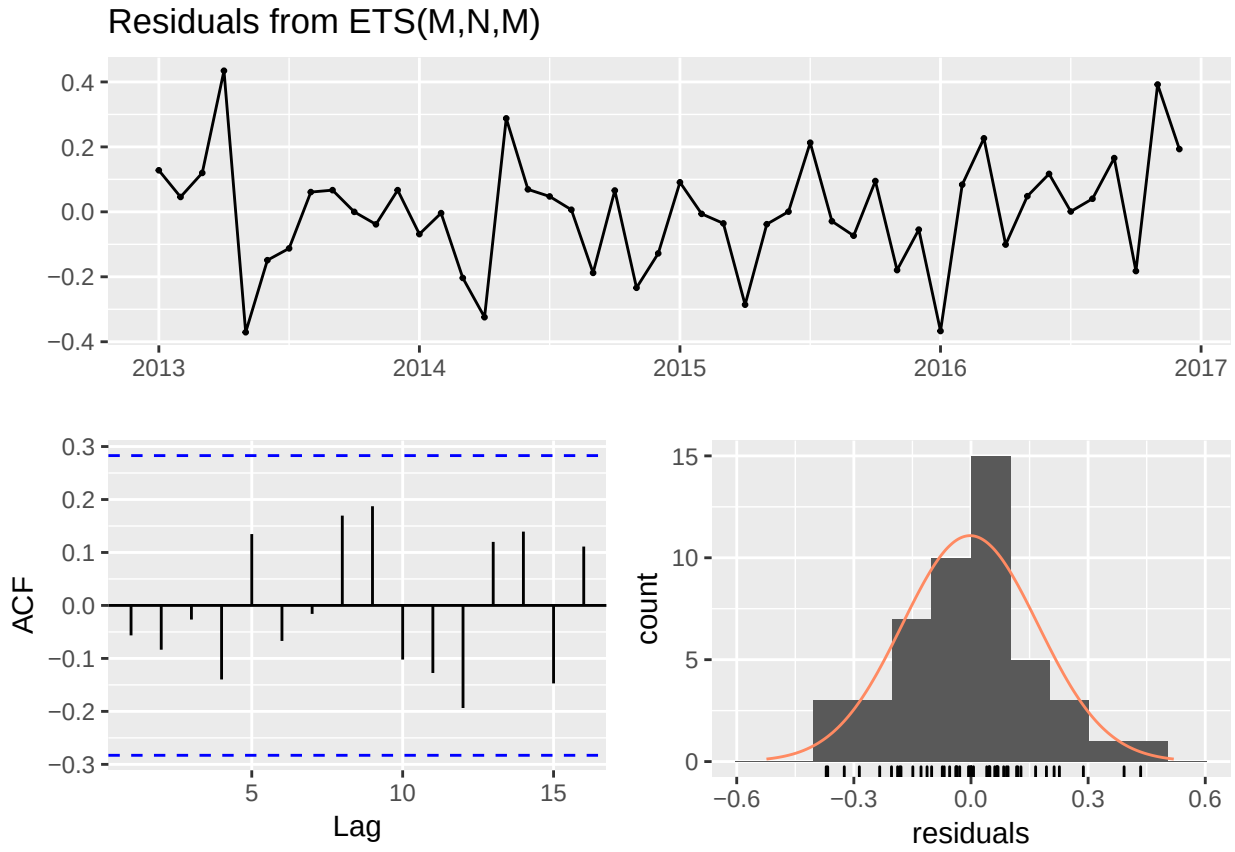
In this section, we will use the historical data from 2013 to 2016 and train a model to predict the median pollution level in the first six months of 2017. The data from 2017 will be kept as a test set to evaluate the model accuracy and the model will be trained on monthly median pollution level

Two modeling techniques are tested and compared for the forecasting of the data: - ARIMA model: an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model - Exponential smoothing model: Exponential Smoothing is a technique to make forecasts by using a weighted mean of past values, wherein more recent values are given higher weights.

These two methods are commonly used to model time series with a seasonality component.




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,0)(1,0,0)[12] with non-zero mean
## Q* = 27.557, df = 21, p-value = 0.1532
##
## Model df: 3.   Total lags used: 24
```



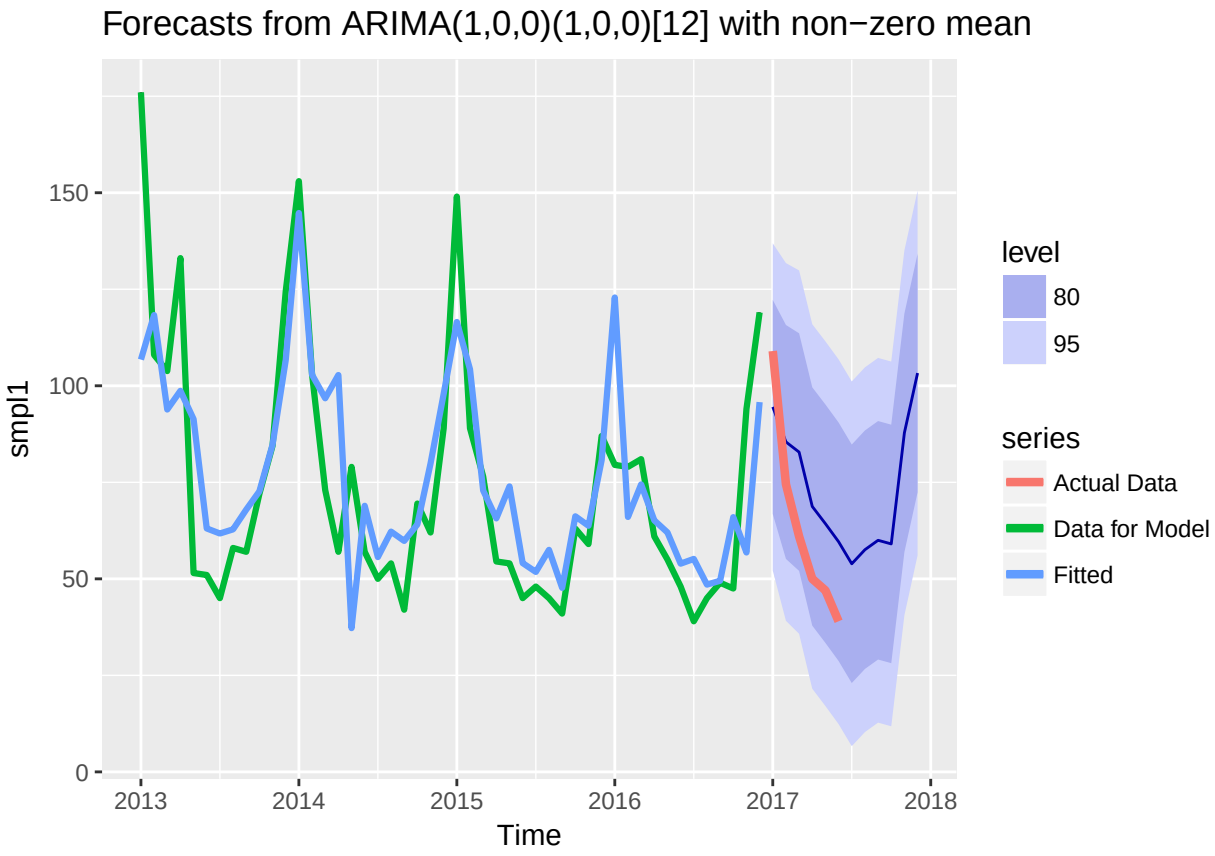

```
##
##  Ljung-Box test
##
## data:  Residuals from ETS(M,N,M)
## Q* = 26.054, df = 10, p-value = 0.003668
##
## Model df: 14.   Total lags used: 24
```

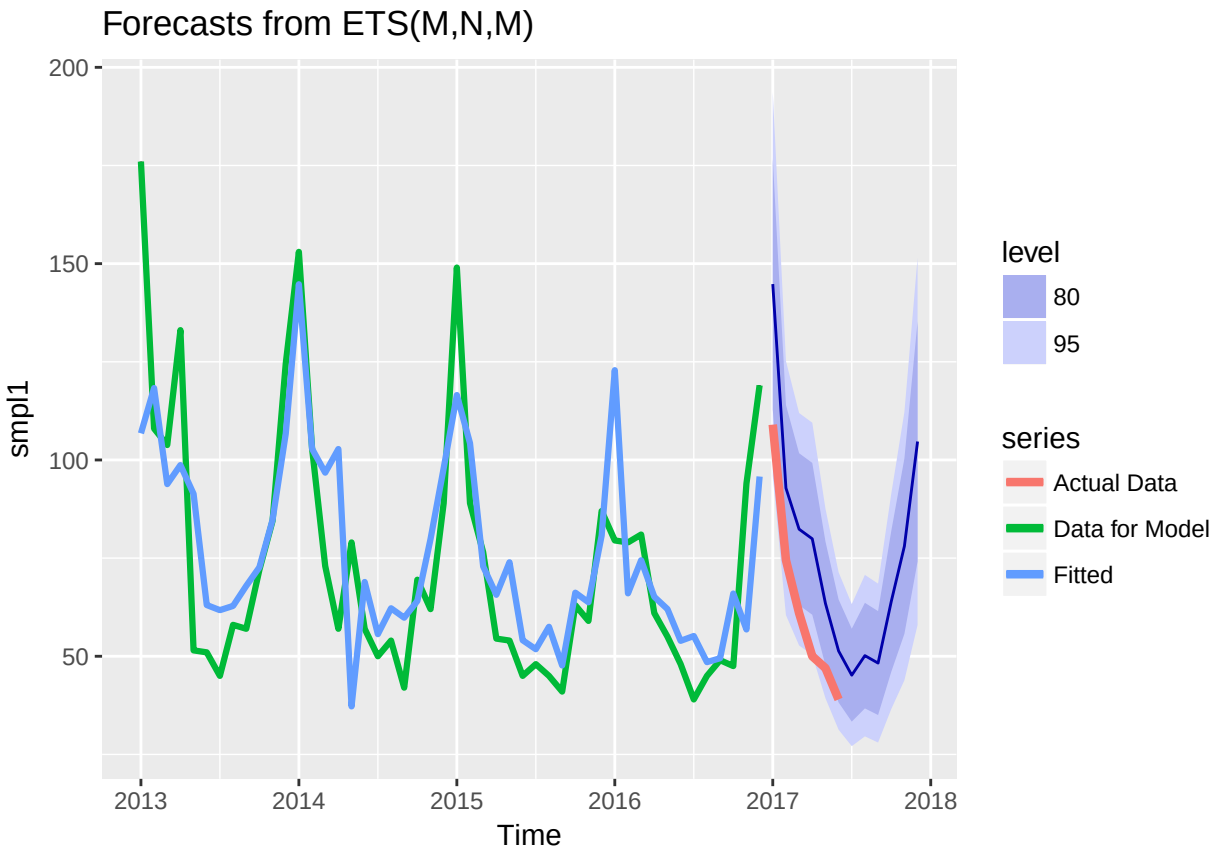
```
## [1] "Accuracy calculation for ARiMA model"
```

```
##                      ME     RMSE      MAE       MPE     MAPE     MASE
## Training set  -2.329164 20.93455 15.21301 -10.01869 21.37439 1.030688
## Test set     -12.486736 17.70272 17.31369 -27.37857 31.80697 1.173010
##                    ACF1 Theil's U
## Training set -0.1506674        NA
## Test set      0.1632544  1.591709
```

```
## [1] "Accuracy calculation for ETS model"
```

```
##                       ME     RMSE      MAE        MPE     MAPE      MASE
## Training set  -0.6810069 14.53677 10.20087  -3.492175 14.05643 0.6911138
## Test set     -22.3730314 23.78055 22.37303 -36.504679 36.50468 1.5157828
##                     ACF1 Theil's U
## Training set -0.02260829        NA
## Test set     -0.11066192  1.633535
```


Forecasts from ARIMA(1,0,0)(1,0,0)[12] with non−zero mean
smpl1
150
100
50
0
2013
2014
2015
2016
2017
2018
Time
level
80
95
series
Actual Data
Data for Model
Fitted




**Interpretation:**

Both model are able to predict the seasonality observed in the exploratory part. The similarities between observations as a function of the time lag between them, given by the ACF does not show a significant auto-correlation (values in-between the twodashed blue lines).

Finally, we can also see that the ARIMA model is better at predicting future values as its RMSE is lower on the test set.

## Conclusion

We have demonstrated that the PM2.5 levels in Chengdu from 2013 to mid-2017 are decreasing overall and over time. We have presented the seasonality of the pollution with high level in December and January. The best period with lower risk of exposure is during the months of June and August. This trend has been succesfully captured and replicated in a preliminary forecasting model.

Air pollution is among the problems faced by cities across the world. Despite extreme values, the data shows a promising trend for the city of Chengdu.

To go further, we know that weather condition has an effect on air pollution. It would be interesting to investigate the correlation between the PM2.5 levels and parameters such as temperature and wind speed.

## Annexes

### Monthly Seasonality Table

```
## # A tibble: 54 x 7
## # Groups:   Year [?]
##    Year  Month MonthMean MonthMedian MonthStd MonthMax MonthMin
##    <fct> <fct>     <dbl>       <dbl>    <dbl>    <dbl>    <dbl>
##  1 2013  Jan       183         176       72.6      409    35.0
##  2 2013  Feb       131         108       79.2      487    22.0
##  3 2013  Mar       125         104       56.8      349    50.0
##  4 2013  Apr       125         133       47.1      271    10.0
##  5 2013  May        61.1        51.5     36.9      257     5.00
##  6 2013  Jun        56.0        51.0     27.0      164     6.00
##  7 2013  Jul        50.0        45.0     25.2      310     2.00
##  8 2013  Aug        61.2        58.0     30.2      157     1.00
##  9 2013  Sep        70.1        57.0     44.3      276     7.00
## 10 2013  Oct        84.4        72.0     57.7      436     1.00
## # ... with 44 more rows
```

### R Session Info

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 16299)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] forecast_8.2    gridExtra_2.3   bindrcpp_0.2    ggplot2_2.2.1
## [5] zoo_1.8-1       lubridate_1.7.1 dplyr_0.7.4
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.14      pillar_1.1.0      compiler_3.4.3
##  [4] plyr_1.8.4        bindr_0.1         xts_0.10-1
##  [7] tseries_0.10-43   tools_3.4.3       digest_0.6.13
## [10] evaluate_0.10.1   tibble_1.4.2      gtable_0.2.0
## [13] lattice_0.20-35   pkgconfig_2.0.1   rlang_0.1.6
## [16] cli_1.0.0         curl_3.1          parallel_3.4.3
## [19] yaml_2.1.16       stringr_1.2.0     knitr_1.19
## [22] nnet_7.3-12       lmtest_0.9-35     rprojroot_1.3-2
## [25] grid_3.4.3        glue_1.2.0        R6_2.2.2
## [28] rmarkdown_1.8     TTR_0.23-2        reshape2_1.4.3
```

```
## [31] magrittr_1.5      backports_1.1.2   scales_0.5.0
## [34] htmltools_0.3.6   quantmod_0.4-12   assertthat_0.2.0
## [37] timeDate_3042.101 colorspace_1.3-2  fracdiff_1.4-2
## [40] quadprog_1.5-5    labeling_0.3      utf8_1.1.3
## [43] stringi_1.1.6     lazyeval_0.2.1    munsell_0.4.3
## [46] crayon_1.3.4
```
‘‘‘