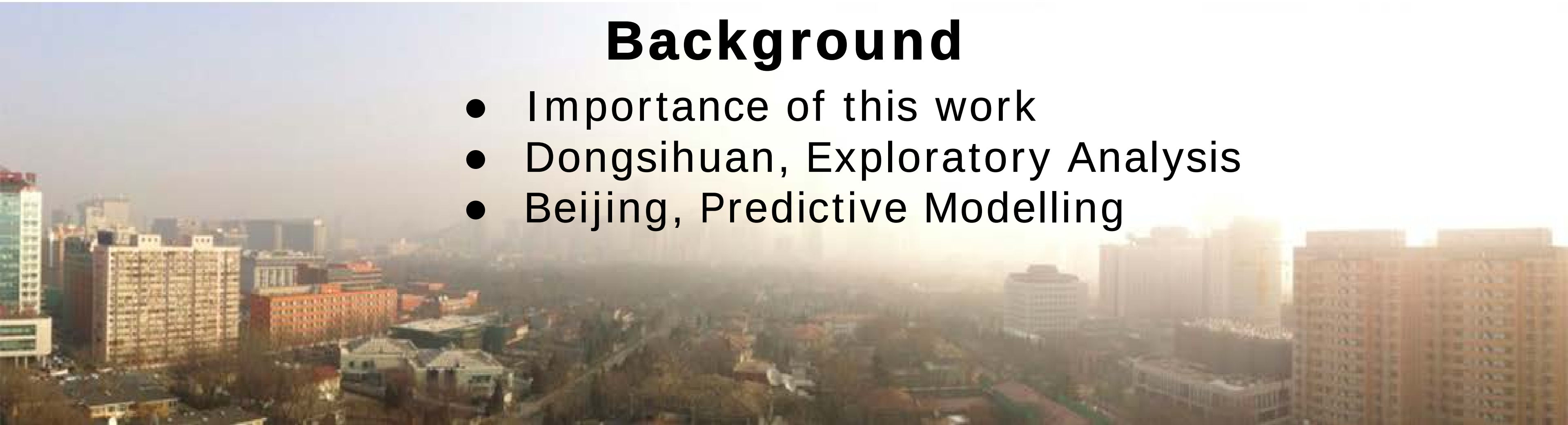
Project - Air Pollution Prediction in Beijing

Background

- Importance of this work
- Dongsihuan, Exploratory Analysis
- Beijing, Predictive Modelling

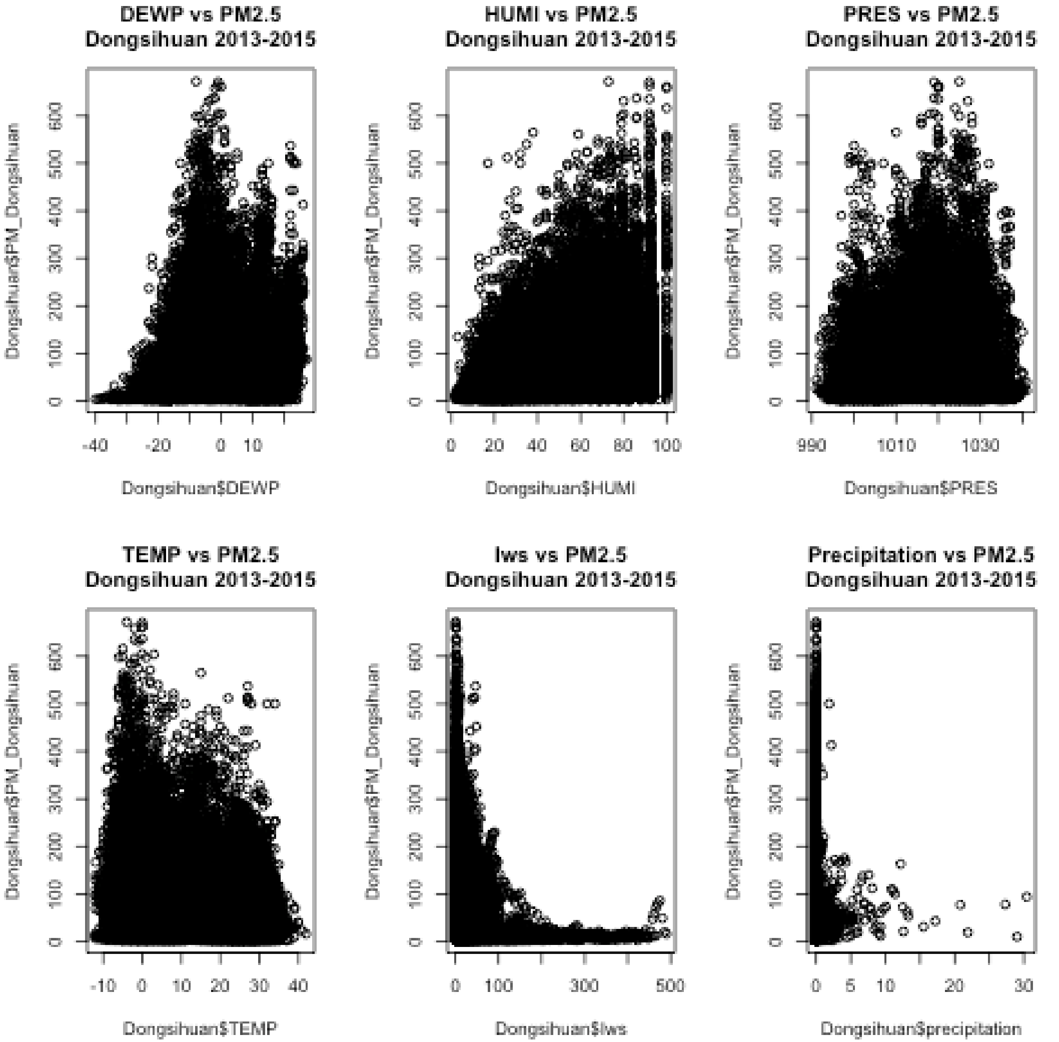


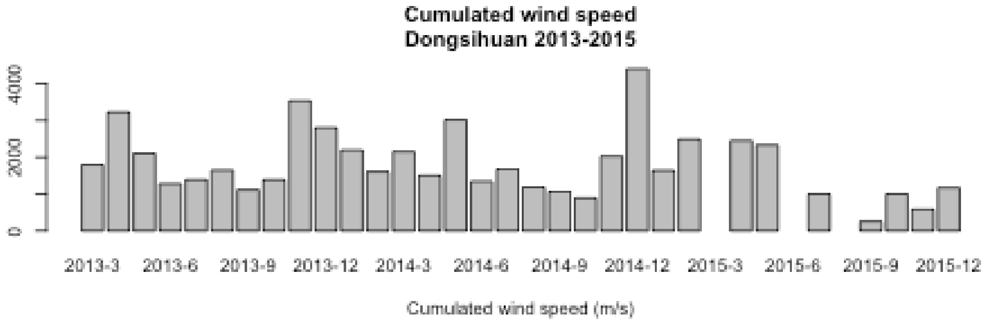
Dongsihuan Exploratory Analysis

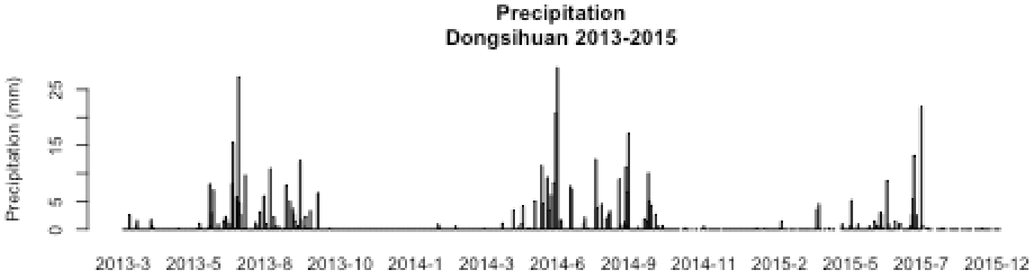
Atmospheric Correlates

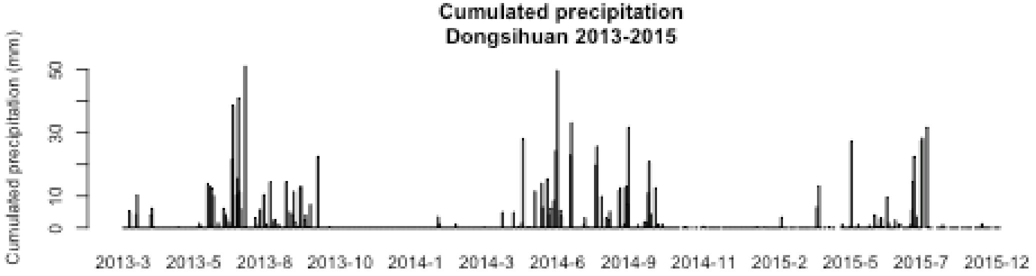
Higher PM2.5 appear to be associated with;

1) Increasing;
   a) Dewpoint (DEWP) and
   b) Humidity (HUMI)
2) Decreasing
   a) Precipitation
   b) Cumulated wind speed (Iws)
   c) Temperature (TEMP)




Cumulated wind speed
Dongsihuan 2013-2015


Precipitation
Dongsihuan 2013-2015

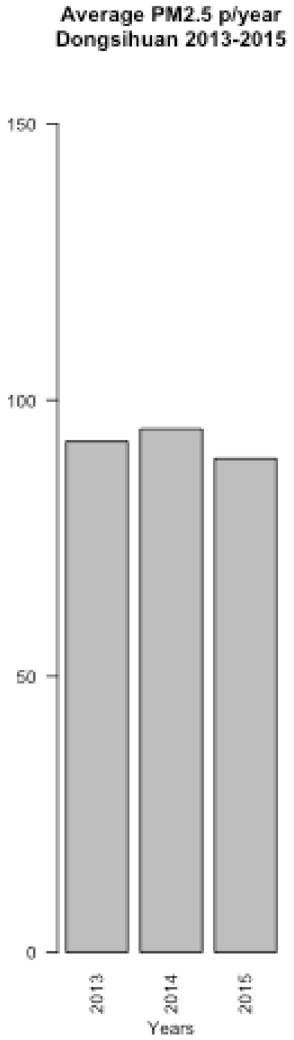

Cumulated precipitation
Dongsihuan 2013-2015

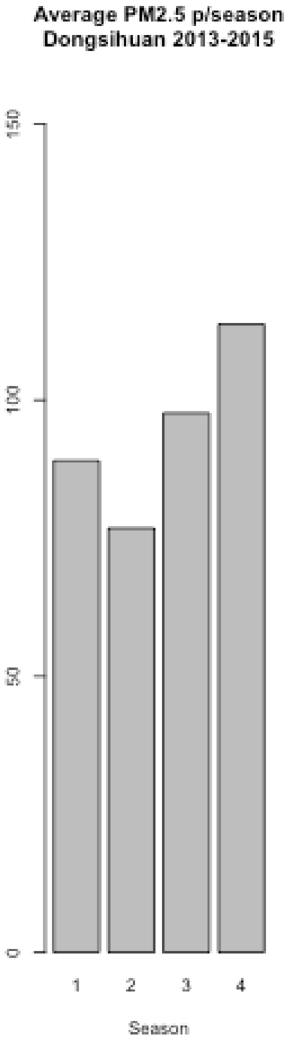
Dongsihuan Exploratory Analysis

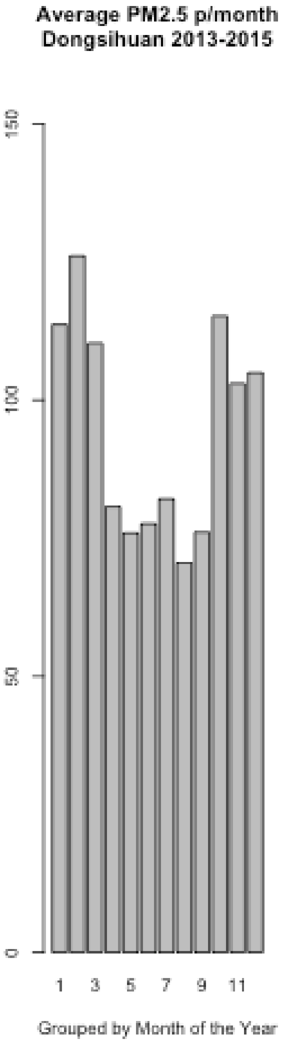
Season ~ Atmospheric correlates

This slide shows seasonal variation in cumulated wind speed and precipitation.

Note the higher precipitation in Beijing’s mid-year warmer months.







Dongsihuan Exploratory Analysis

Seasonal Correlates

Note relatively lower Average PM2.5 p/month and p/season in the same warmer months as the higher precipitation in the earlier slide.

- Winter (4)
- Spring (1)
- Summer (2)
- Autumn (3)

Dongsihuan Exploratory Analysis

Anthropological Correlates

Our analysis also considered effects working hours in Beijing. Where working hours are

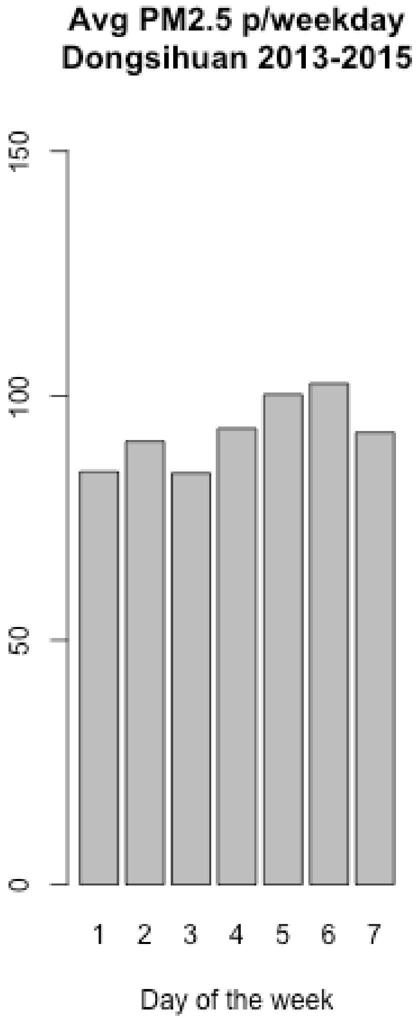
Monday to Friday (8:00 – 18:00) [5]

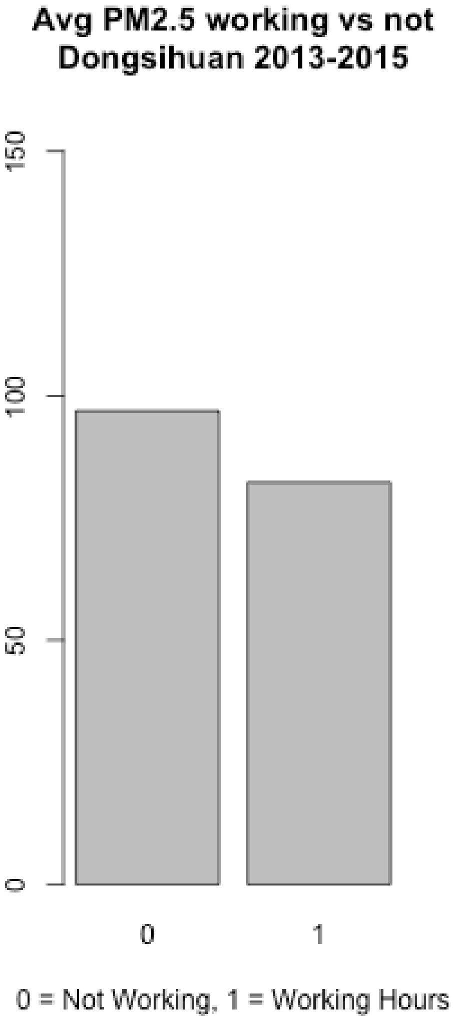
In this figure we find a statistically significant difference between working (1) and non working hours (0).

Welch Two Sample t-test

t(20508) = 148.47, $p < 2.2 \times 10^{-16}$,

99.99% CI [88.87218, 93.65620]


Avg PM2.5 p/weekday
Dongsihuan 2013-2015


Avg PM2.5 working vs not
Dongsihuan 2013-2015

Issues in Data

1. Data type identification

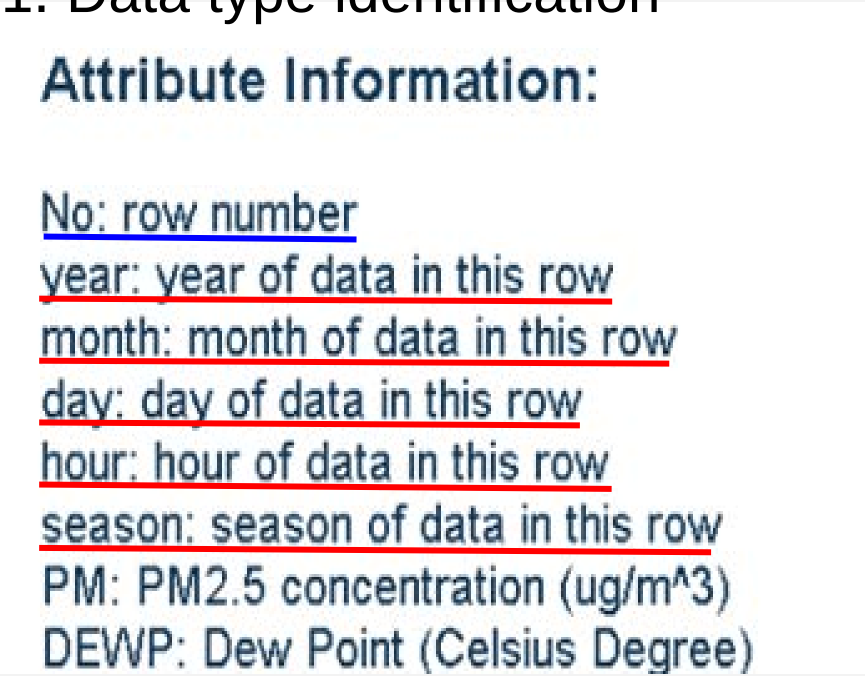
2. NA values

3. Outliers

4. Negative values and right skewed distribution

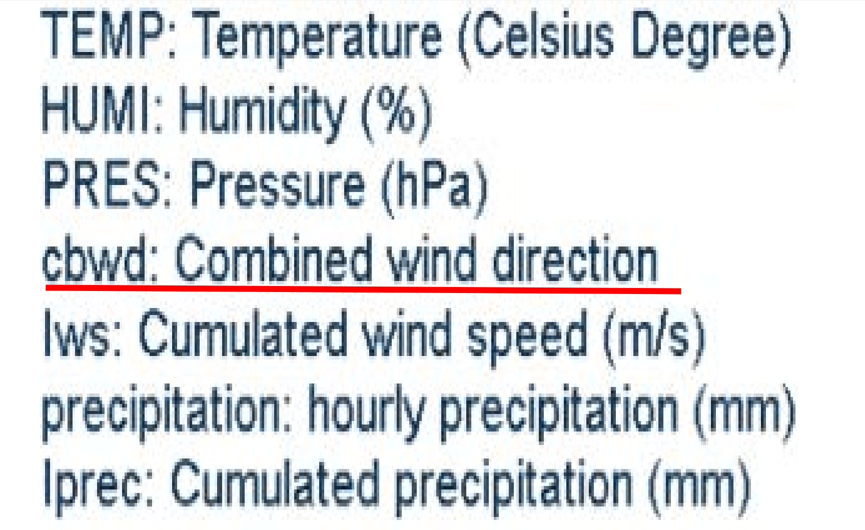
Issues in Data

1. Data type identification



Attribute Information:

No: row number
year: year of data in this row
month: month of data in this row
day: day of data in this row
hour: hour of data in this row
season: season of data in this row
PM: PM2.5 concentration (ug/m^3)
DEWP: Dew Point (Celsius Degree)



TEMP: Temperature (Celsius Degree)
HUMI: Humidity (%)
PRES: Pressure (hPa)
cbwd: Combined wind direction
Iws: Cumulated wind speed (m/s)
precipitation: hourly precipitation (mm)
Iprec: Cumulated precipitation (mm)

+

~~Weekday: Mon~Fri:1, Sat~Sun:0~~

~~Working hour: Weekday 8:00~17:59:1,~~ others:0

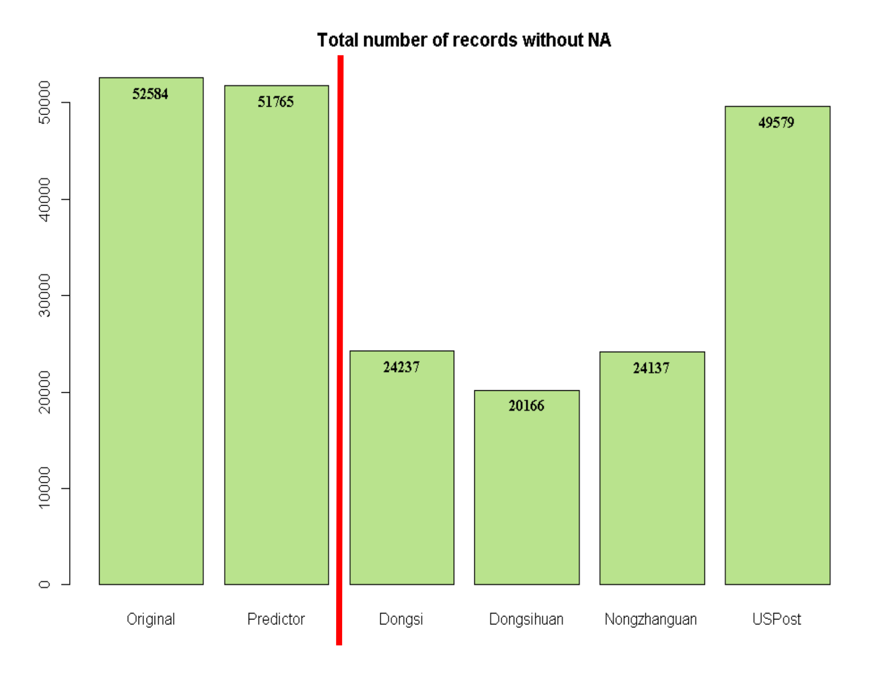
https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities

🟦 :remove 🟥 :factor

Issues in Data

2. NA values

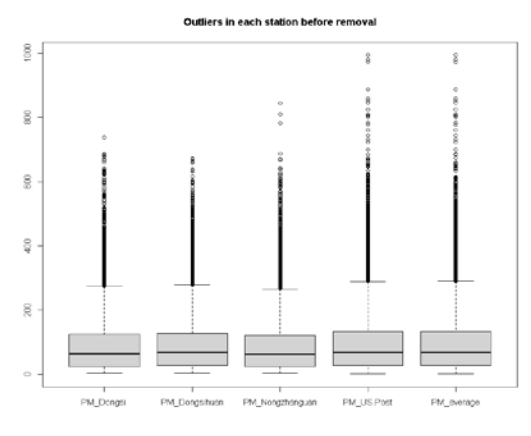
- Removal of records having NA
  - Predictors: small reduction (52,584 → 51,765 [98%])
  - Responses: large reduction in majority stations
    - Dongsi (52,584 → 24,237 [46%])
    - Dongsihuan (52,584 → 20,166 [38%])
    - Nongzhanguan (52,584 → 24,137 [46%])
    - USPost (52,584 → 49,579 [94%])
- Solutions
  - (1) Use USPost
  - (2) Aggregate data

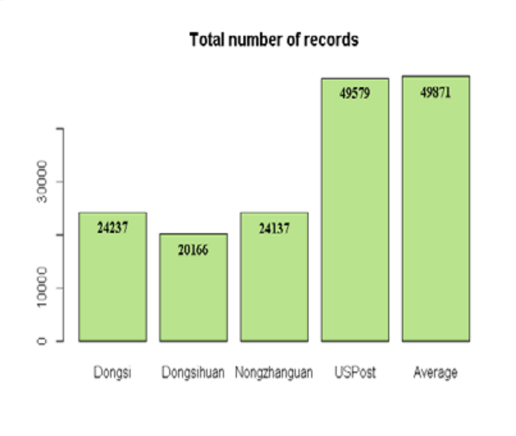


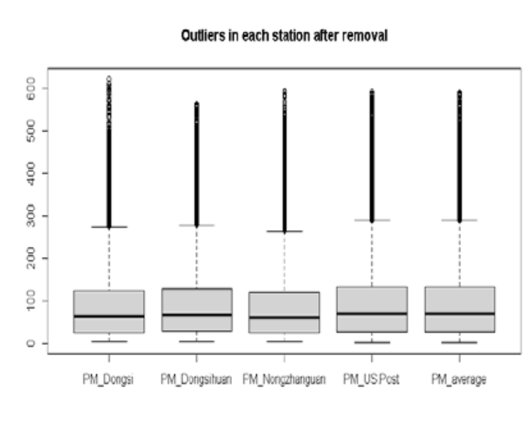
Issues in Data

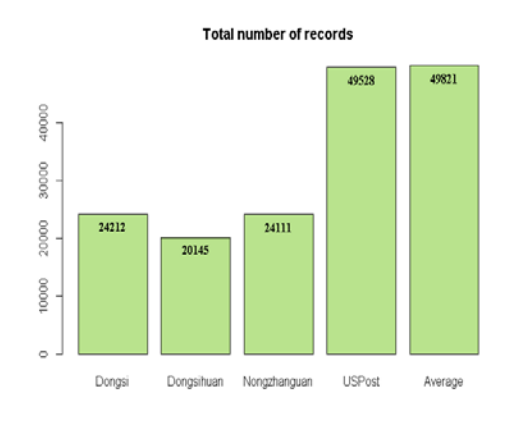
3. Outliers

- Observed many outliers
  - Removing all → lose an enormous amount of data
- Solution
  - Set threshold at 0.999
    - Preserve many records
    - Extremely large values → removed







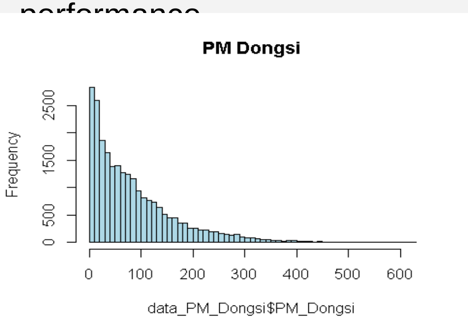


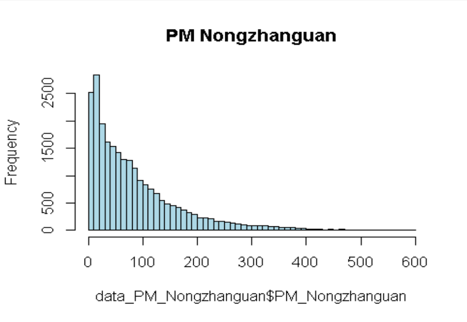
Issues in Data

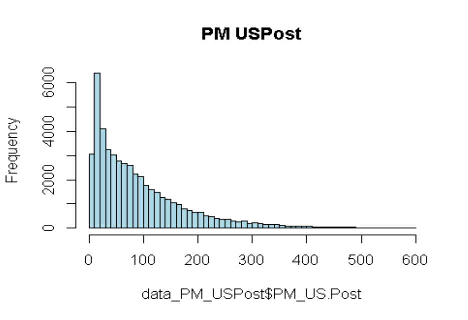
4. Negative values and right skewed distribution

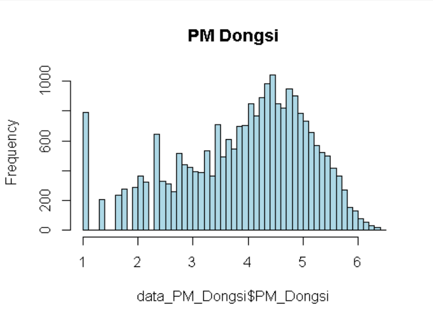
- Responses: must be negative & right skewed
    - If predictions = negative → lowers performance
    - If distribution = not normally distributed → lowers performance

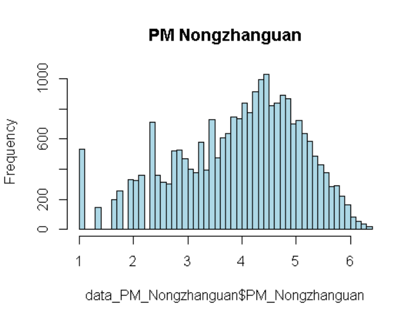
- Solution
    - Change the distribution
    - Log transformation

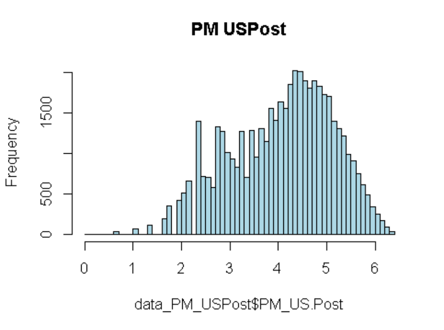










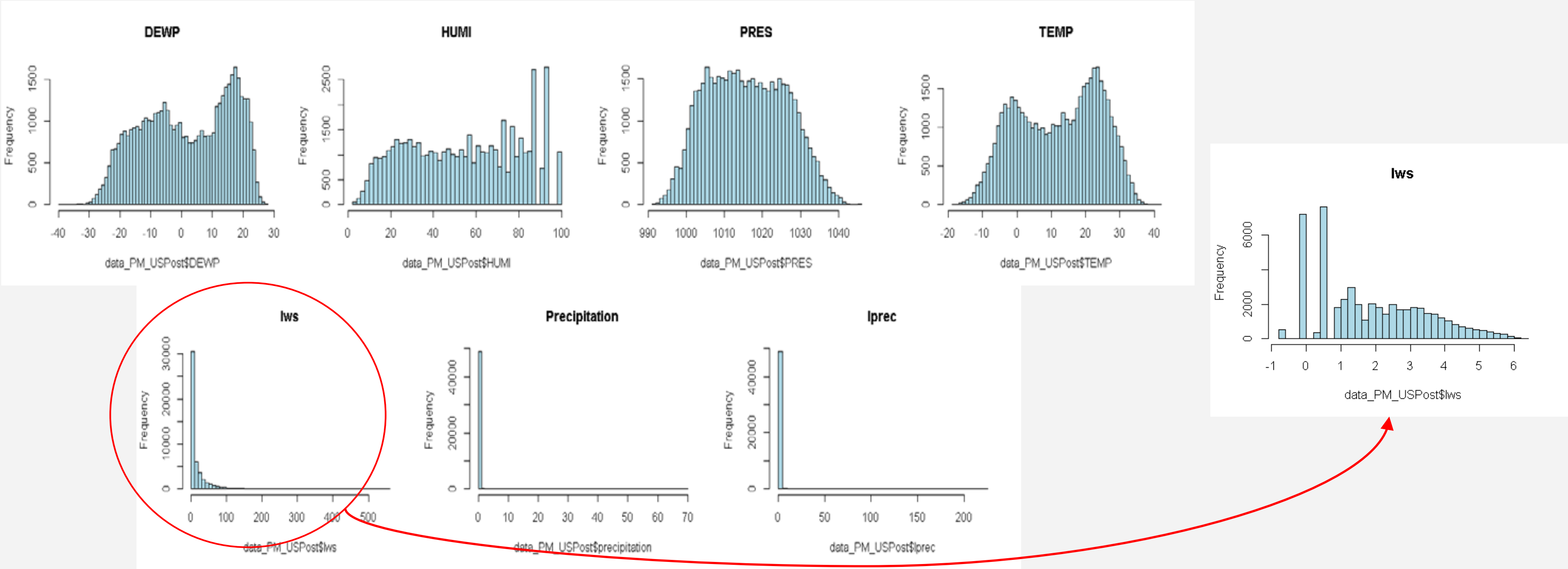


Issues in Data

4. Negative values and right skewed distribution (cont.)

- Continuous predictors: some are right skewed
    - Iws
- If distribution = not normally distributed → lowers performance

- Solution
    - Change the distribution
    - Log transformation



Linear Regression - Overview

Linear Regression has following form:

$$y = \beta_0 + \sum_{i=1}^{p} (\beta_i x_i) + \varepsilon$$

Such that: x_i are predictors, y is the response and ε is normal distributed white noise.

2 models are fit with selection criteria being:

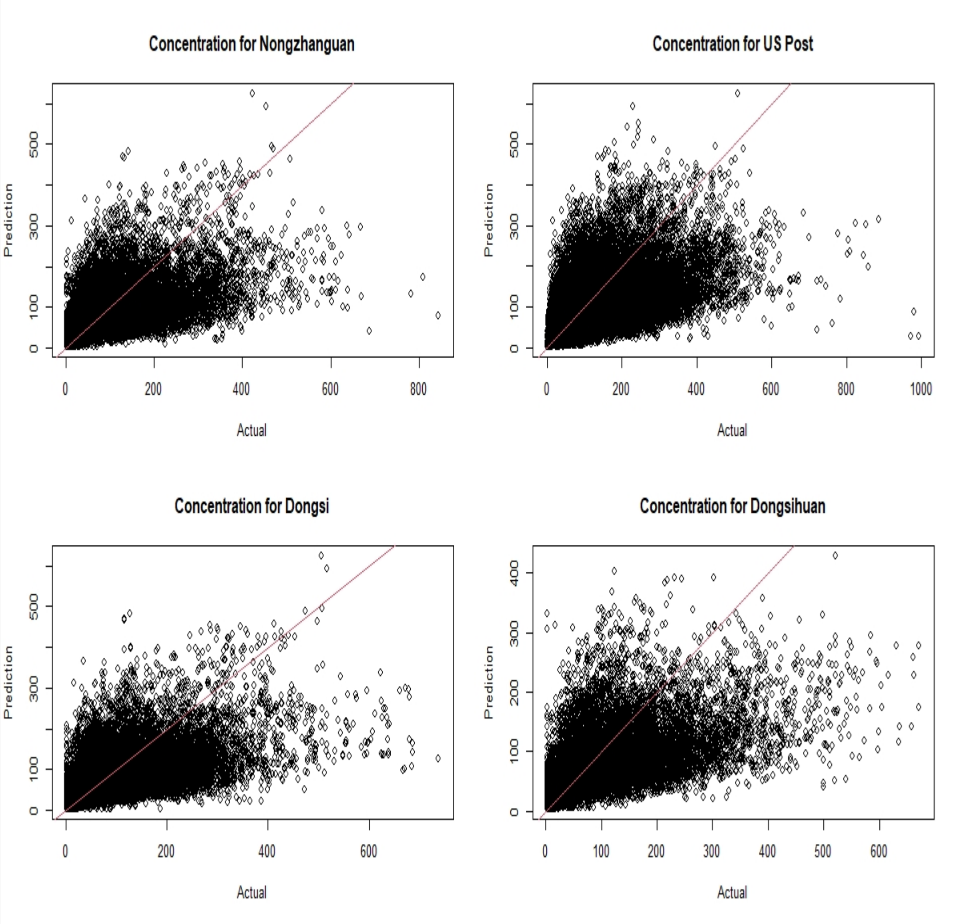
- AIC to penalise more complex models
- Variance explained by fitted model
- Root Mean Square Error (RMSE) for predictions against actual values

Linear Regression without aggregation

Beijing Model

Fitting data using linear regression without aggregation it is found:

- 54.63% of variance explained from fitted model.
- Lowest AIC is -22,429.26 using parameters.
- RMSE is 69.2294, 71.1151, 69.2994 and 75.0562 respectively for each prediction.

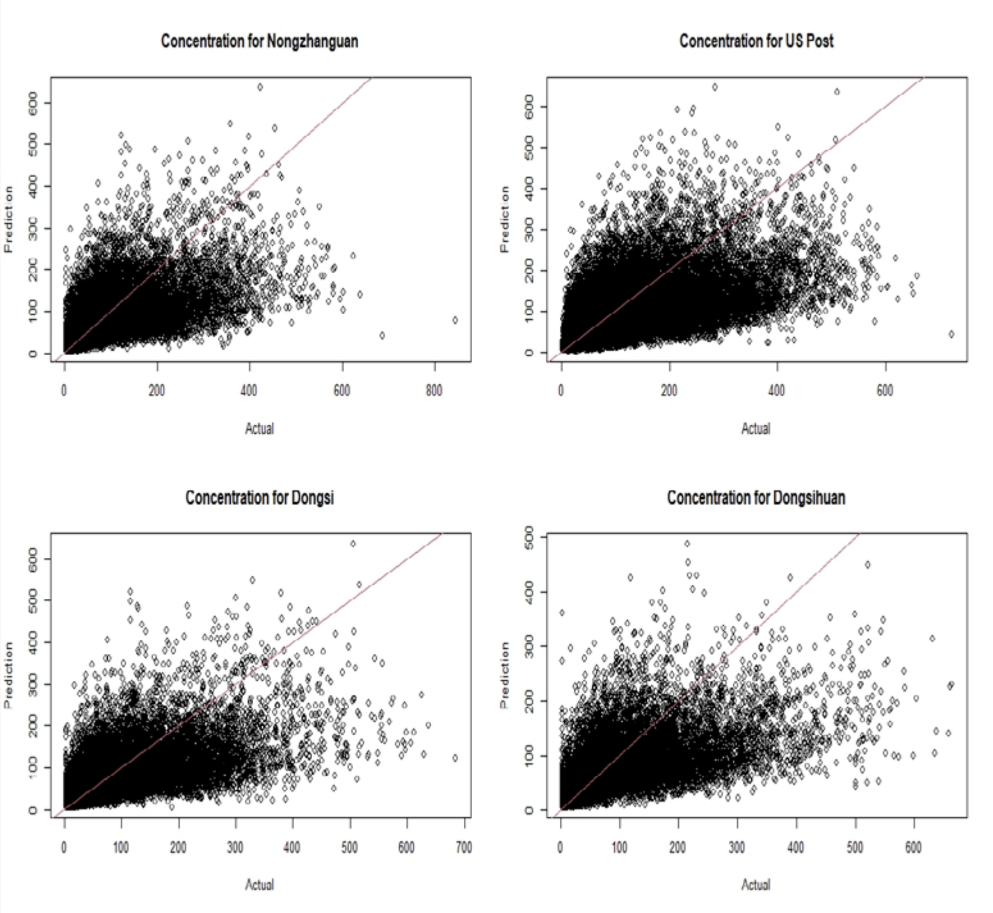


Linear Regression with aggregation

Beijing Model

Fitting data using linear regression with aggregation it is found:

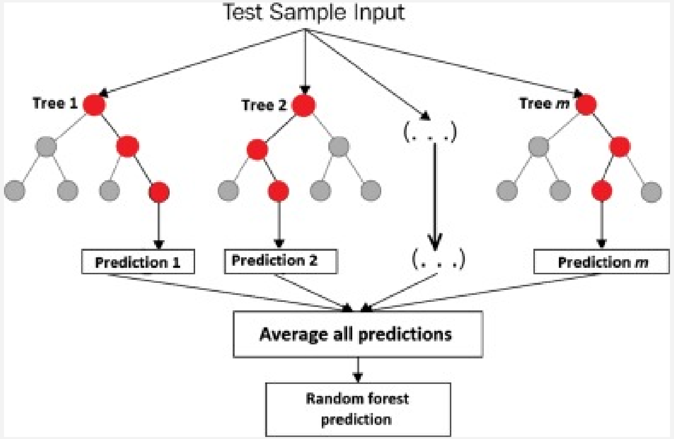
- 56.77% of variance explained from fitted model.
- Lowest AIC is -25,464.58 using parameters.
- RMSE is 69.63813, 71.2218, 68.8664 and 73.04727 respectively for each prediction.



Random Forest - Overview

- Built on Decision Tree
- **Training dataset -> Bootstrap -> Ensemble of trees -> Aggregation**
- Strengths & Weaknesses analysis

Strengths	Weaknesses
Rely on many predictions instead of a single prediction. It has higher accuracy than decision tree. Low risk of overfitting	It is relatively slow and ineffective for predictions because of a bunch of trees
Used to handle unbalanced data. It can be used for both regression and classification problems	
Robust to outliers and non-linear data	



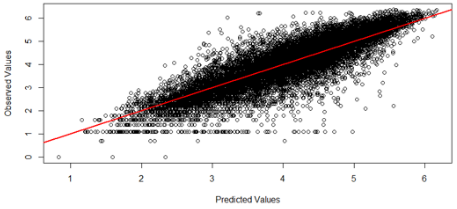
Source from [4]

Random Forest Without Aggregation

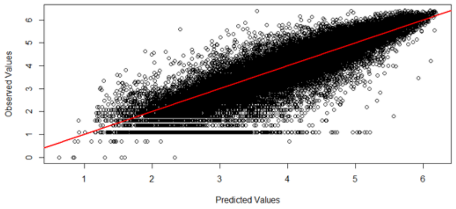
- Develop a model based on Dongsihuan replacing NA values with PM_US Post

Fitting data using random forest without aggregation it is found (training/test sets):

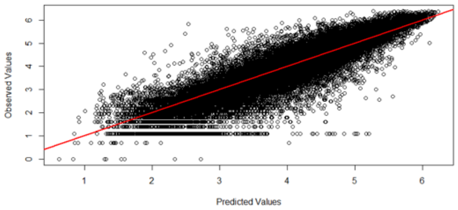
- 75.64% of variance explained from fitted model
- RMSE is 0.5312, 0.4691, 0.4345 and 0.3911 respectively for each prediction
- Lowest RMSE for the US Post prediction



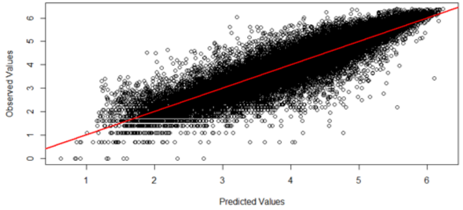
Dongsihuan



Dongsi



Nongzhanguan



US Post

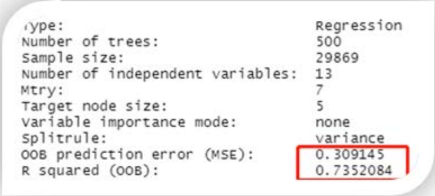
Random Forest Without Aggregation

- Develop four models based on different locations

Fitting data using random forest without aggregation it is found (training/test sets):

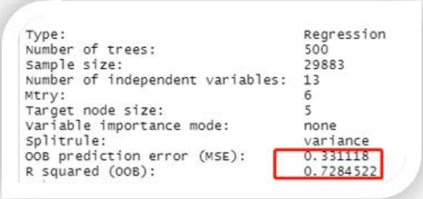
Metrics/Locations	Dongsihuan	Dongsi	Nongzhanguan	US Post
R squared (OOB)	0.7352	0.7285	0.7354	0.7484
RMSE	0.5508	0.5724	0.5602	0.5304

- Highest value of R squared and lowest value of RMSE for the US Post prediction



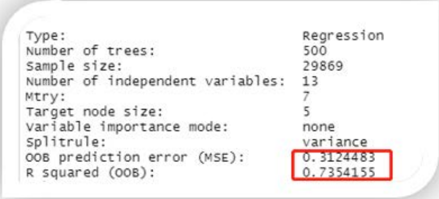
```
ype:                              Regression
Number of trees:                  500
Sample size:                      29869
Number of independent variables:  13
Mtry:                             7
Target node size:                 5
Variable importance mode:         none
Splitrule:                        variance
OOB prediction error (MSE):       0.309145
R squared (OOB):                  0.7352084
```

Dongsihuan (Targeted location)



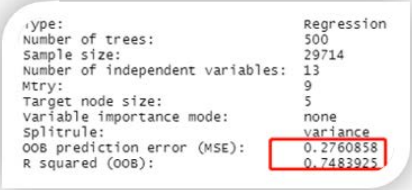
```
Type:                             Regression
Number of trees:                  500
Sample size:                      29883
Number of independent variables:  13
Mtry:                             6
Target node size:                 5
Variable importance mode:         none
Splitrule:                        variance
OOB prediction error (MSE):       0.331118
R squared (OOB):                  0.7284522
```

Dongsi



```
Type:                             Regression
Number of trees:                  500
Sample size:                      29869
Number of independent variables:  13
Mtry:                             7
Target node size:                 5
Variable importance mode:         none
Splitrule:                        variance
OOB prediction error (MSE):       0.3124483
R squared (OOB):                  0.7354155
```

Nongzhanguan



```
ype:                              Regression
Number of trees:                  500
Sample size:                      29714
Number of independent variables:  13
Mtry:                             9
Target node size:                 5
Variable importance mode:         none
Splitrule:                        variance
OOB prediction error (MSE):       0.2760858
R squared (OOB):                  0.7483925
```

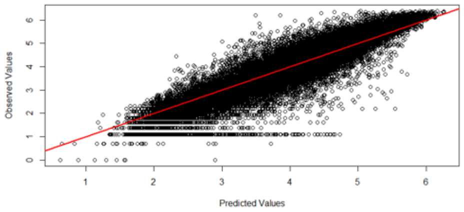
US Post

Random Forest With Aggregation

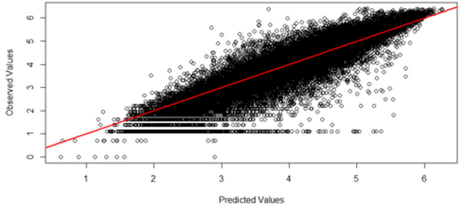
Beijing Model

Fitting data using random forest with aggregation it is found (training/test sets):

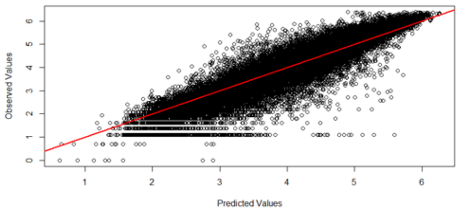
- 76.94% of variance explained from fitted model
- RMSE is 0.4042, 0.4314, 0.4066 and 0.3666 respectively for each prediction
- Lowest RMSE for the US Post prediction
- Perform better than the model without aggregation



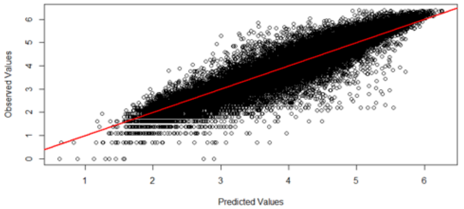
Dongsihuan



Dongsi



Nongzhanguan



US Post

Technical Thinking and Inferring

1. Surprisingly, the most important feature in predicting PM2.5 is Dew point, rather than rain.

RMSE	Without aggregation	With aggregation
Linear Regression	71.2 (Average of 4 locations)	70.7
Random Forest	0.46 (US Post values replace NA values)	
Random Forest	0.55 (Average of 4 locations)	**0.4**

RMSE = 0.4 is an excellent result

2. Compared with the 71.2 RMSE of the linear model, the random forest produces only 0.55 RMSE (without aggregation.

3. Data cleaning can reduce RMSE by approximately 15%(Random forest).

3. After aggregation, the RMSE of the model decreased by about 30% compared to before (Random forest).

4. Therefore, for this dataset, model selection is most important than other factors, random forest is the best model; cleaning data and feature engineering (aggregation) also have a positive impact on the predictions of the models.

Project Bonus Exploration

1. Additional decision trees analysis

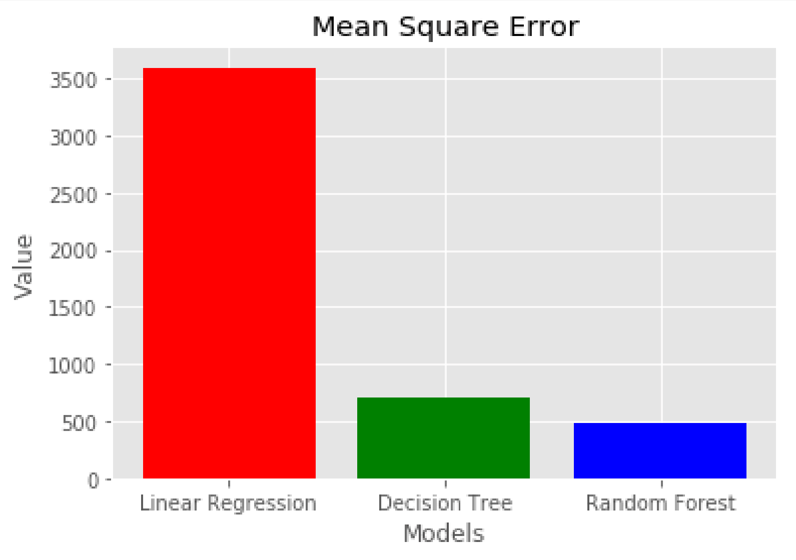
Performance: Random Forest > Decision Tree > Linear Model.

2. Additional Python Platform Test

Result: The result of Python analysis is almost no different from R. In Python, Random Forest still performs the best.

3. The correctness of our conclusions is verified from the perspective of other platforms.

Why not try it?



Conclusion

1. Pollution is most severe in the evening of winter on non-working days each year, while it is least severe in the weekday afternoon of spring and summer. This reflects the fact that human activities and natural conditions are affecting PM2.5.

2. Days with excellent, moderate and pollution air quality in Beijing accounted for 31.5%, 21.9% and 46.6% respectively [6]. The air quality in Beijing is generally almost half and half between acceptable and unacceptable.

3. Looking at the 3-years trend, there is a slight downward trend in pm2.5 in Beijing.

4. Suggestions: make related laws and regulations, strengthen air monitoring, develop green energy-saving energy, and reduce PM2.5 emissions to improve the air quality.

Project successfully completed

References

1. Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., Chen, S., 2015. Assessing Beijing's PM 2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 471, 20150257. https://doi.org/10.1098/rspa.2015.0257
2. blood - Red blood cells (erythrocytes) | Britannica [WWW Document], n.d. URL https://www.britannica.com/science/blood-biochemistry/Red-blood-cells-erythrocytes (accessed 5.7.22).
3. Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. JAMA 287, 1132–1141. https://doi.org/10.1001/jama.287.9.1132
4. Lin, Y., Zou, J., Yang, W., Li, C.-Q., 2018. A Review of Recent Advances in Research on PM2.5 in China. Int J Environ Res Public Health 15, 438. https://doi.org/10.3390/ijerph15030438
5. AsialinkBusiness, n.d. Business hours in China [WWW Document]. Asialink Business. URL https://asialinkbusiness.com.au/china/business-practicalities-in-china/business-hours-in-china?doNothing=1 (accessed 5.6.22).
6. Based on World Health Organization PM2.5 Standard 2005 & 2021, the 3 key indicators are: 75 µg/m3, 50 µg/m3 and 37.5 µg/m3 per 24-hour.
   https://www.c40knowledgehub.org/s/article/WHO-Air-Quality-Guidelines?language=en_US