Final Project

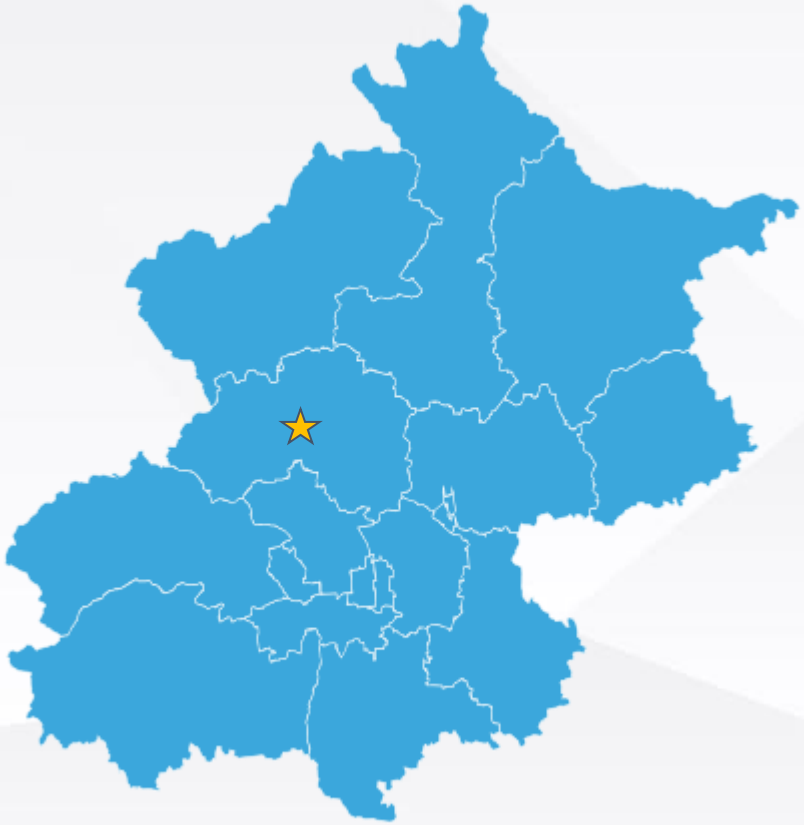
Group 8: Hung Wen Chen, Wei Lin Liao

Instructor: Daya Rudhramoorthi

CONTENTS

01 Introduction

- **Dataset Introduction**



- Our dataset is the hourly air pollutants data from the Changping district, near Beijing.
- The time period is from 03/01/2013 to 02/28/2017. Missing data are denoted as NA.
- Our research focused on two parts,
  - PM2.5 with time series
  - Relationships between each pollutants and weather conditions.

01 Introduction

- **Dataset Cleaning & Preperation**

```
# Check NaN
df.isnull().sum()
```

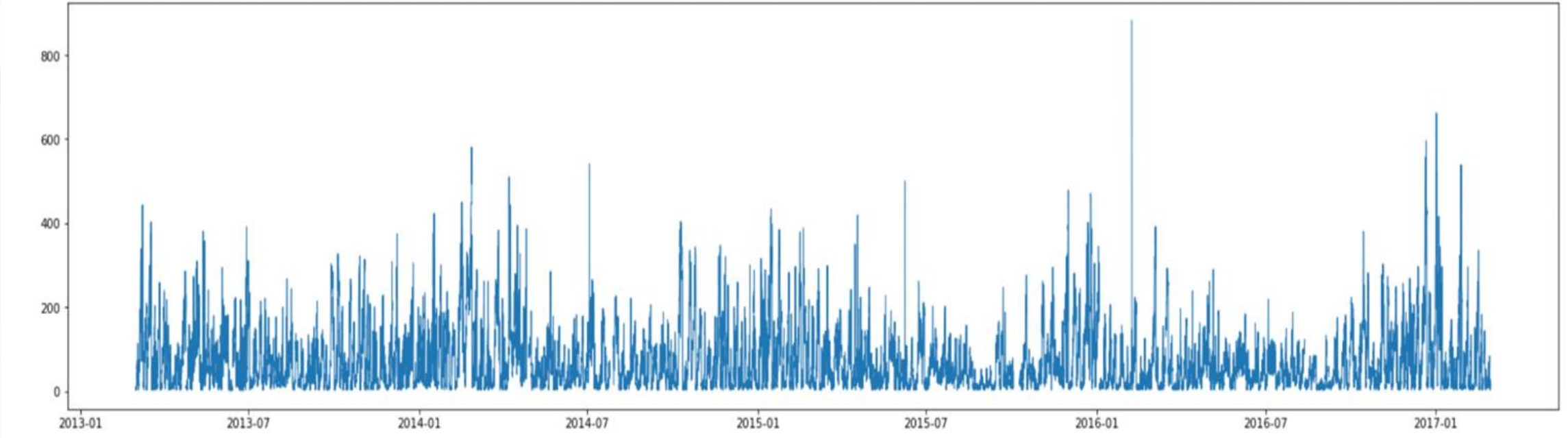
```
No          0
year        0
month       0
day         0
hour        0
PM2.5     774
PM10      582
SO2       628
NO2       667
CO       1521
O3        604
TEMP       53
PRES       50
DEWP       53
RAIN       51
wd        140
WSPM       43
station     0
dtype: int64
```

```
# Data cleaning & preparation
df = df.interpolate()

pmseries = df.copy()
pmseries['date'] = df[['month', 'day', 'hour', 'year']].astype(str).agg('-'.join, axis=1)
pmseries['date'] = pd.to_datetime(pmseries['date'], format='%m-%d-%H-%Y')

pm25 = pd.DataFrame(df, columns=['year', 'month', 'PM2.5'])
```

- Most of the missing values occured in pollutants columns, we use interpolate to insert value as equally spaced.
- For the data analysis, we add a datetime column and subset a pm2.5 dataframe.
- For the data modeling, we merged hourly dataset to daily dataset by mean.

02 Explementory Data Analysis

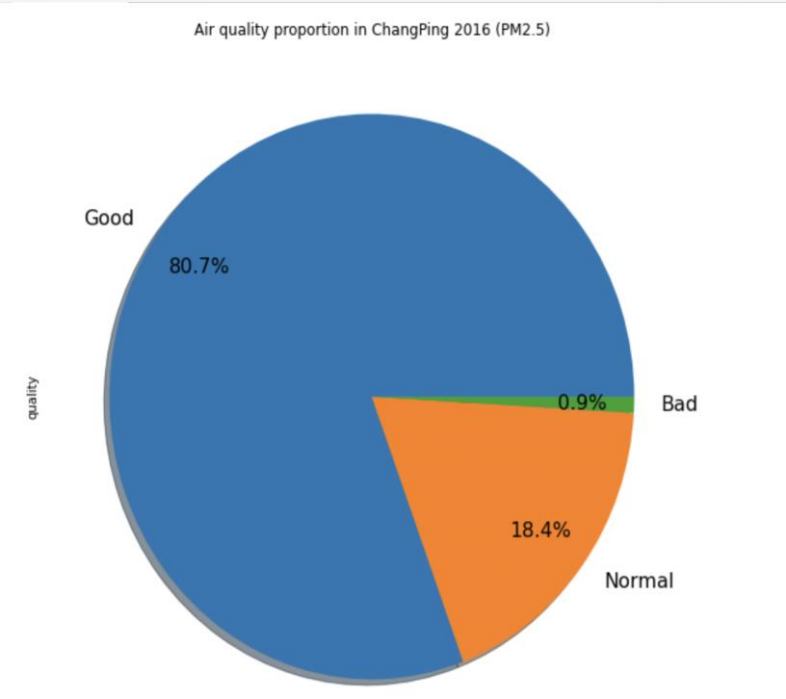
1. Variation characteristics of Particulate Matter in Changping. (PM2.5)



- The PM2.5 content in the air in Changping decreased significantly in the second half of 2016, the reason may be affected by the environment and policies.

02 Explementory Data Analysis

2. In the whole year of 2016, what is the proportion of days with good air quality (PM2.5<=100) and the worst air quality (PM2.5>300)?



2016

- Good air quality : 80.7%
- Bad air quality : 0.9%

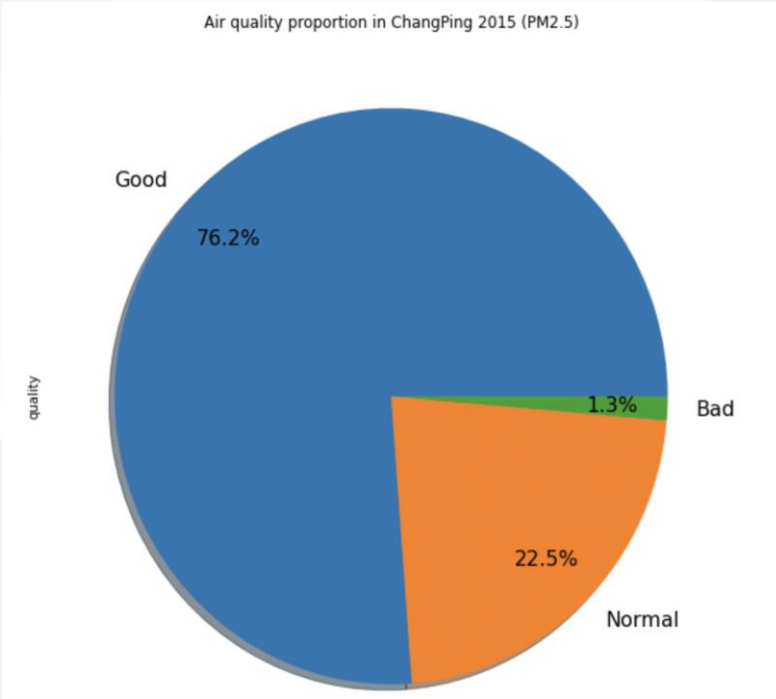
2015

- Good air quality : 76.2%
- Bad air quality : 1.3%

Conclusion

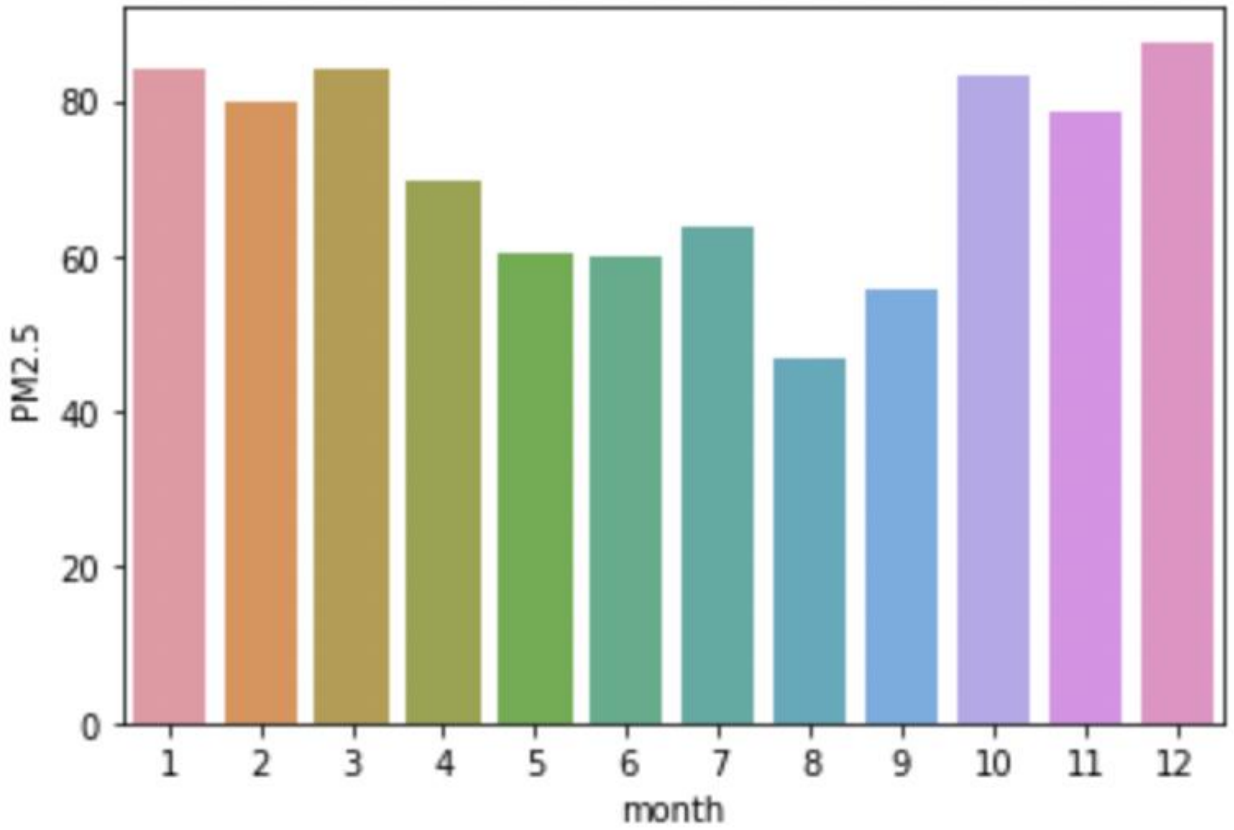
- The air quality in Changping in 2016 was generally better than in 2015.

3. In the whole year of 2015, what is the proportion of the days with good air quality (PM2.5<=100) and the worst air quality (PM2.5>300)?



02 Explementory Data Analysis

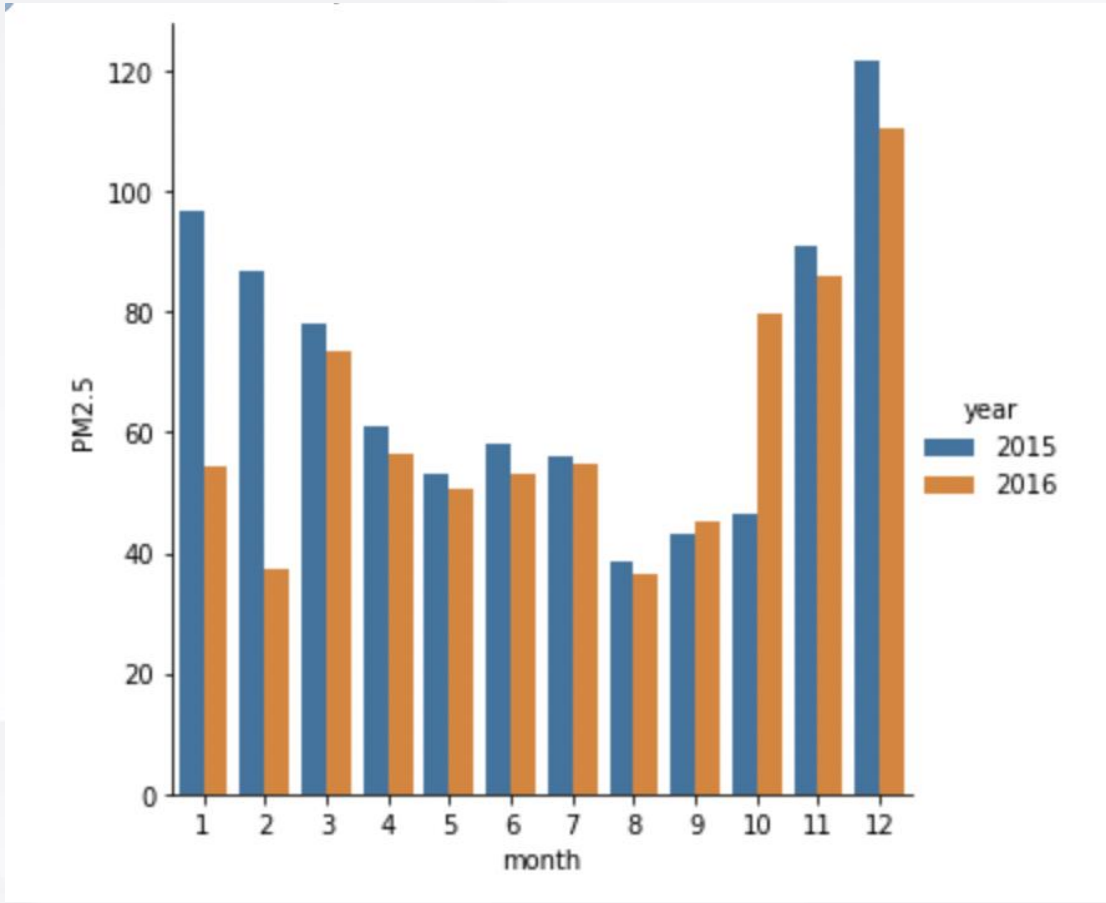
4. What is the relationship between air quality and seasonal (month) changes?



- For the PM2.5 pollution level in average, August is the lowest, and the December is the highest.
- PM2.5 will be affected by seasons, with lower PM2.5 levels in summer and autumn, and higher in winter and spring.

02 Explementory Data Analysis

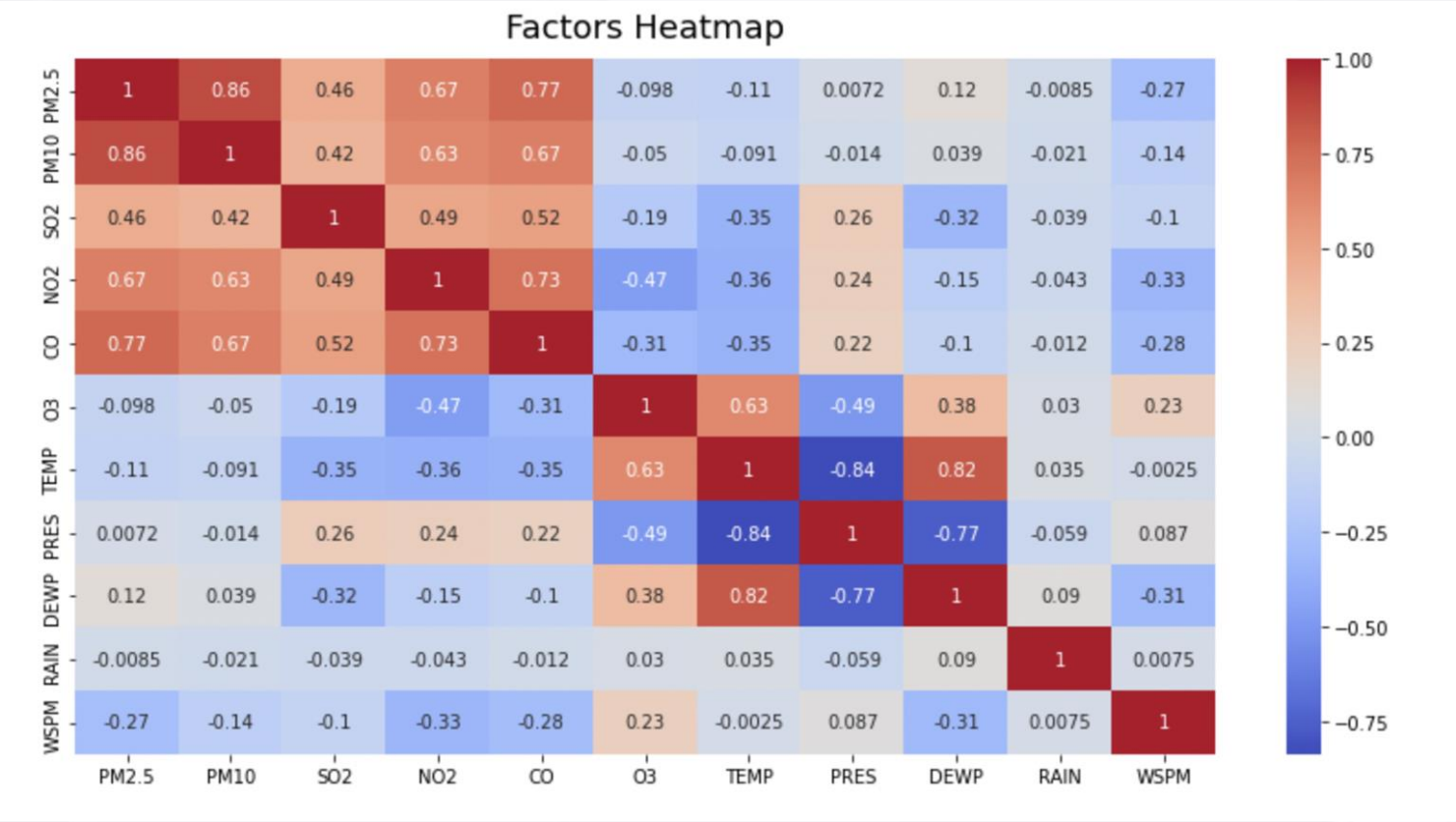
5. Comparing 2016 with 2015 at the same time, how often was the air quality better or worse?



- In most of the months, the air quality in 2016 is better than 2015, September and October are the exceptions.

02 Explementory Data Analysis

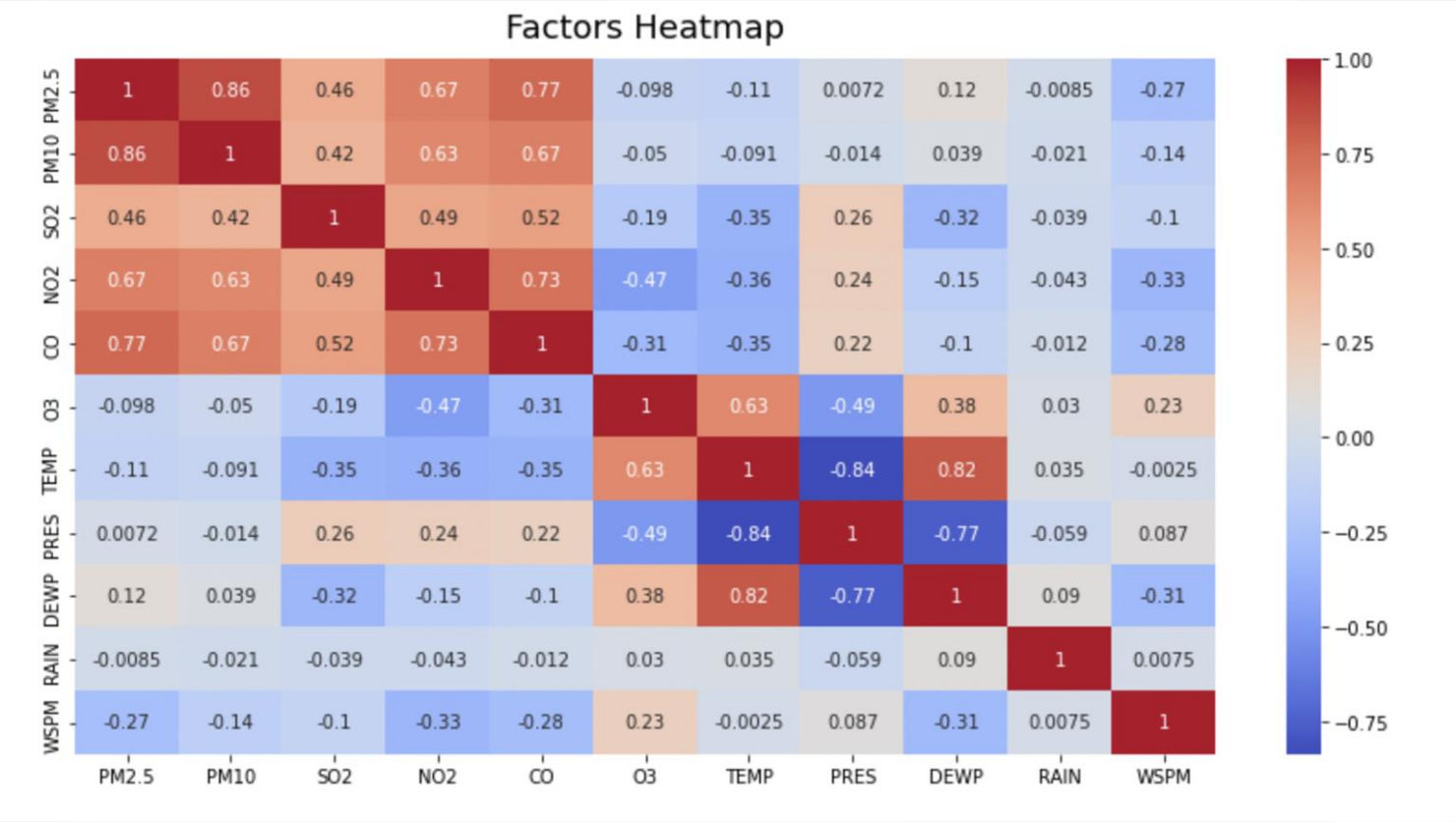
6. Pollutants and meteorological environment correlation coefficients in Changping.



- All kinds of pollutants except O3 are positively correlated.
- NO2 is strongly negatively correlated with O3.
- Temperature and dew point temperature are strongly positively correlated.

02 Explementory Data Analysis

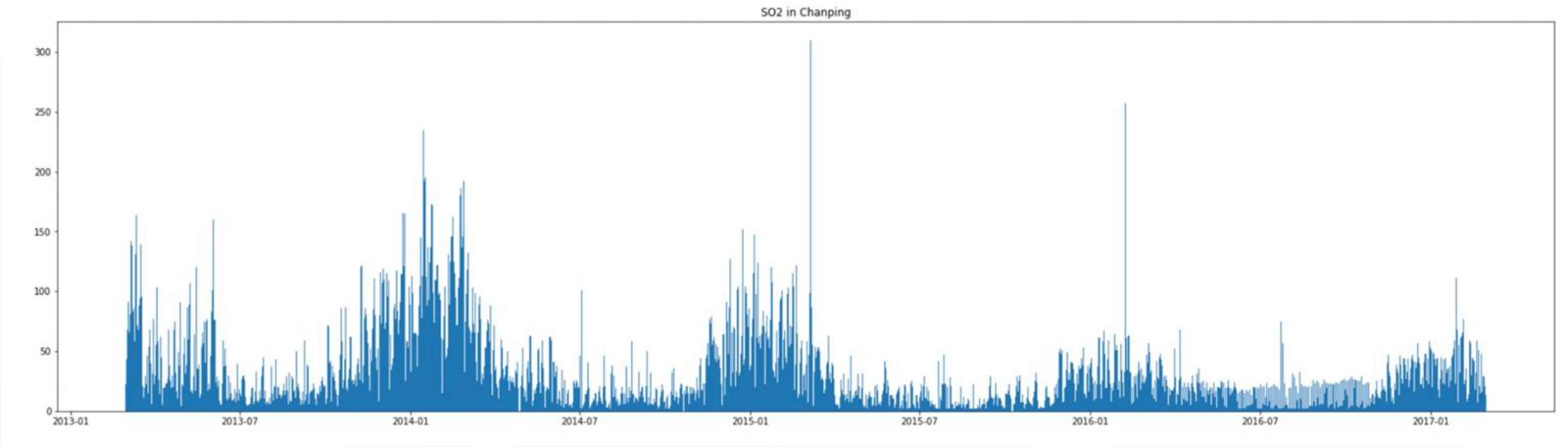
6. Pollutants and meteorological environment correlation coefficients in Changping.



- Temperature and atmospheric pressure are strongly negatively correlated.
- Raining seems to have little effect with PM2.5.

02 Explementory Data Analysis

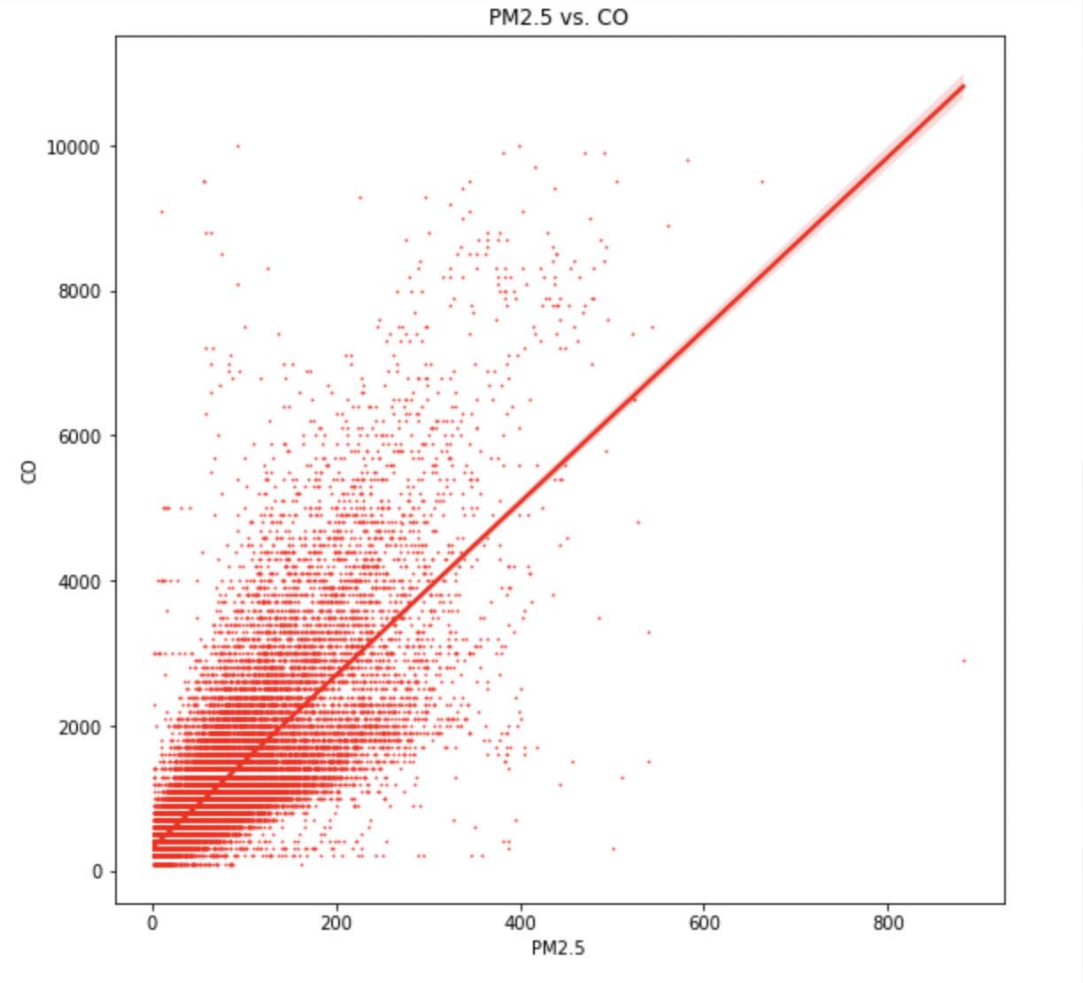
7. The fluctuation of SO2 content around Changping over time.



- The content of SO2 reaches a peak around January and downs to bottom around July every year.

02 Explementory Data Analysis

8. PM2.5 vs. CO

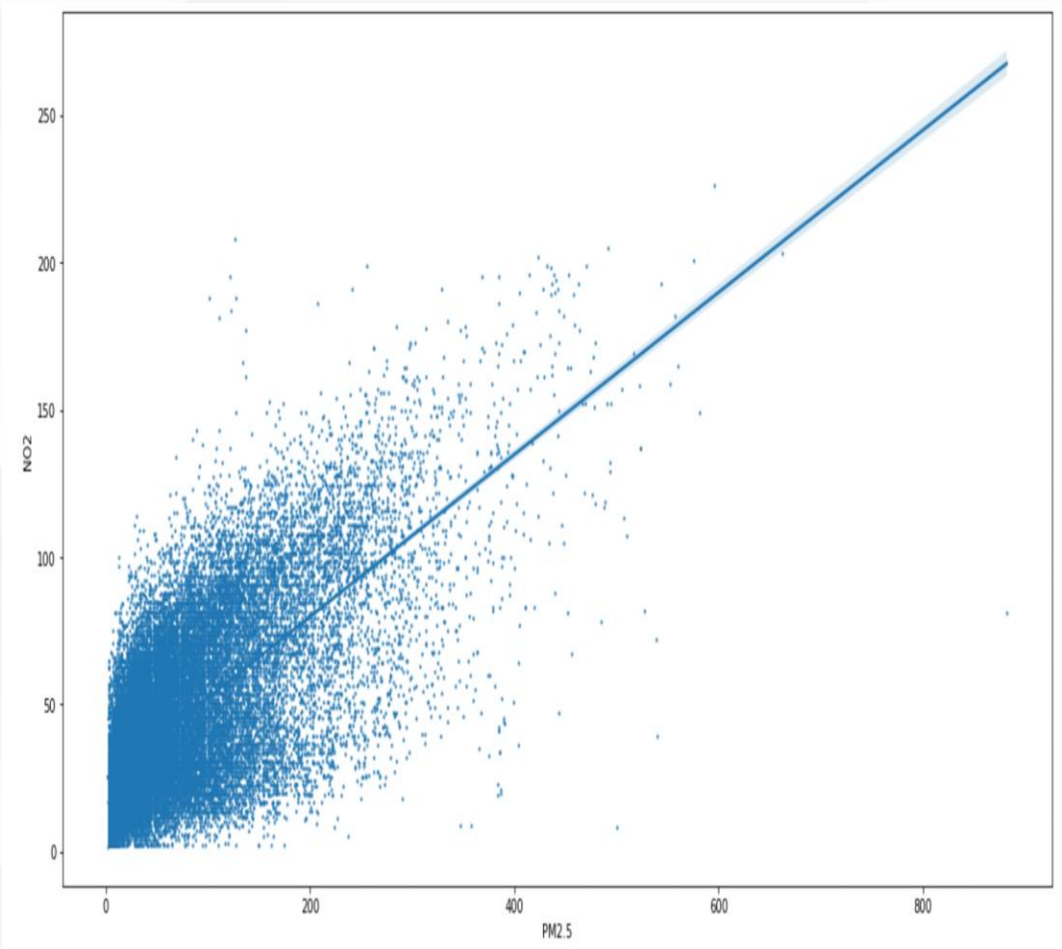


- CO and PM2.5 have positive correlation.

02 Explementory Data Analysis

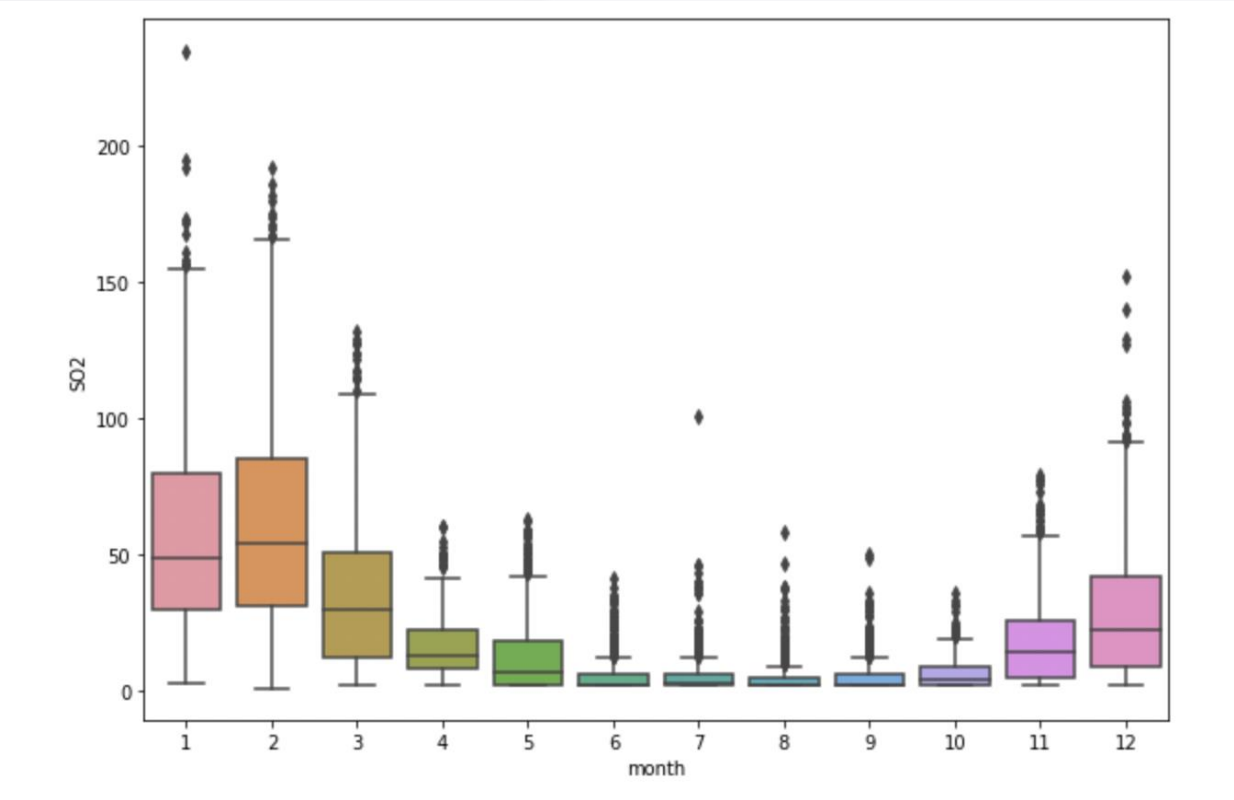
- NO2 and PM2.5 have positive correlation as well.

9. PM2.5 vs. NO2



02 Explementory Data Analysis

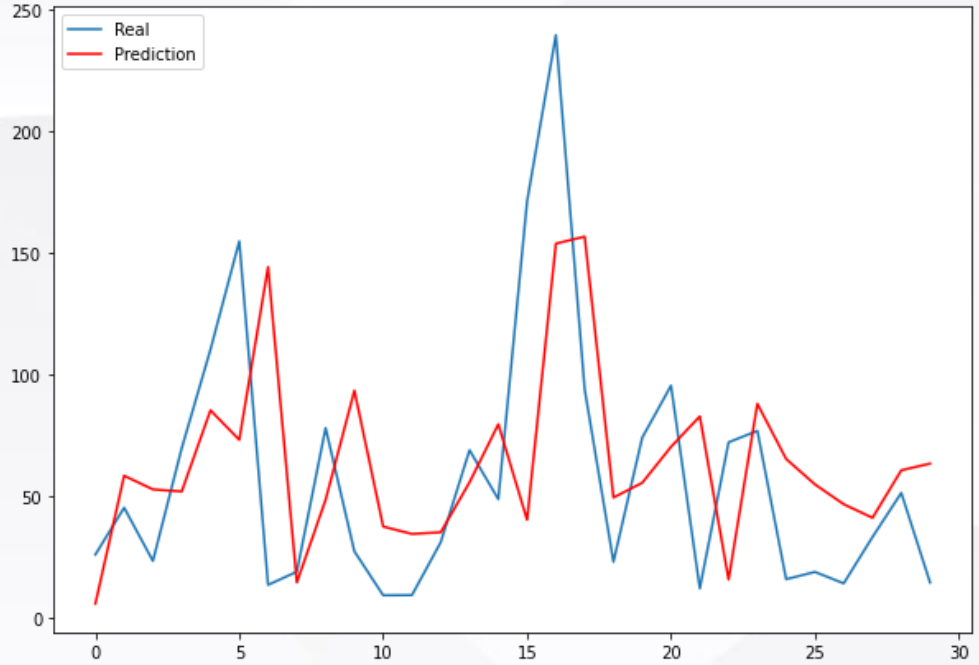
10. SO2 boxplot

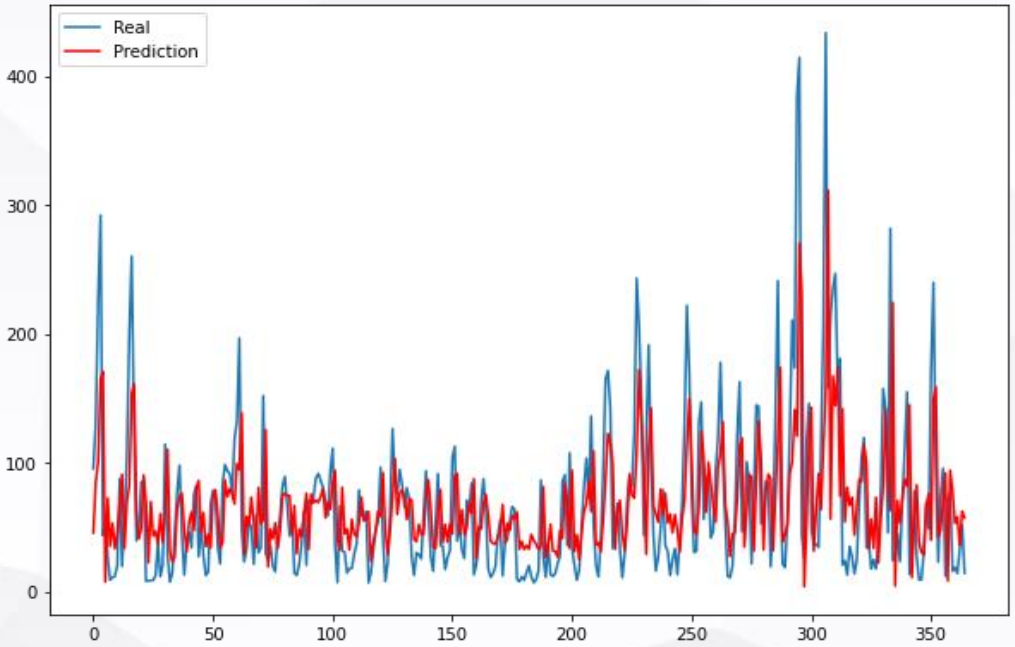


- We can see outliers of every month by boxplot.
- There are some high peaks in January and December.

03 Predictive Models

Autoregressive model: PM2.5 by time series.

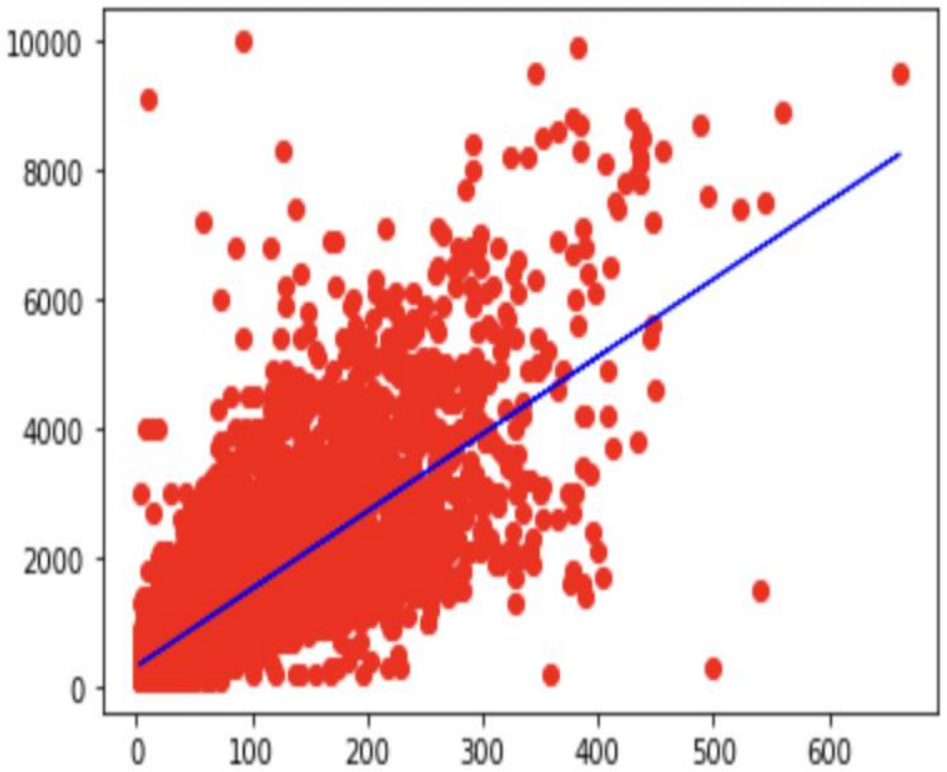




- Left: 30 days prediction, Right: 365 days prediction.
- Use last 30 days to predict the next day (lag=30)
- The prediction follows the same trend but weak at the peak value.

03 Predictive Models

Linear Regression: PM2.5 vs. CO



- Blue line : linear regression line
- Red scatter points : data of the test set.
- Regression coefficient : 12
- Accuracy of the training set : 58.5%
- Accuracy of the test set : 59.6%

04 Conclusion

- Seasonal factors will affect air quality, PM2.5 will increase in winter and decline in summer
- The air quality is improving in 2016 compared to 2015
- The heatmap find the highly connected factors that are PM2.5 vs CO and PM2.5 vs NO2, we create scatter plots and regression lines to approve the positive correlation
- The models are used to know the relationship between PM2.5 and CO and the prediction of PM2.5 future value.
- The government can plan their energy production management by the forecasting model, dedicated to improving the air quality

Reference

1. EPA. (2022, August 1). Particulate Matter (PM2.5) Trends. Retrieved November 12, 2022 from https://www.epa.gov/air-trends/particulate-matter-pm25-trends
2. Mackenzie, J., & Turrentine, J. (2021, June 22). *Air pollution: Everything you need to know*. NRDC. Retrieved November 12, 2022, from https://www.nrdc.org/stories/air-pollution-everything-you-need-know
3. Kumar, A. (2022, August 23). *Different types of time-series forecasting models*. Data Analytics. Retrieved November 13, 2022, from https://vitalflux.com/different-types-of-time-series-forecasting-models/
4. Song Xi Chen. (2019, September 20). Beijing Multi-Site Air-Quality Data Data Set. November 12, 2022, from https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data
5. htuhxf. (n.d.). Time series analysis 1: Analyzing and predicting time series with ar autoregressive model in Python. Retrieved December 11, 2022, from https://blog.csdn.net/htuhxf/article/details/105382451

THANKS YOU