**Project V2 – Data Acquisition & Understanding**

**Air Pollution Forecast**

**AIDI 1002 AI Algorithm**

**Course Facilitator: Marcos Bittencourt**

**Prepared By:**

Yali Huang -100788425

November 30, 2020

Table of Contents

[1. Executive Summary 1](#_Toc57631297)

[2. Exploratory Data Analysis 1](#_Toc57631298)

[Examining the dataset: 1](#_Toc57631299)

[Graphs and Data characteristics: 1](#_Toc57631300)

[Assumptions 6](#_Toc57631301)

[Constraints 6](#_Toc57631302)

[3. Algorithm Evaluation 6](#_Toc57631303)

# Executive Summary

In recent years, air pollution has become a global concern, Especially in some large cities with large populations like Beijing, London, Mumbai. It refers to the release of pollutants into the air that are detrimental to human health and the planet as a whole. Car emissions, chemicals from factories, dust, and pollen and mold spores may be suspended as particles. Among all those gauges of air pollution, PM2.5 concentration is no doubt the most famous and effective.In this project, we use data set from UCI Repository (with time series data from 2010 to 2014) to explore the factors influencing PM2.5 in Beijing, perform machine learning algorithm to predict the PM2.5 values. Suggestion on how to improve air quality in Beijing will be given at the end.

# Exploratory Data Analysis

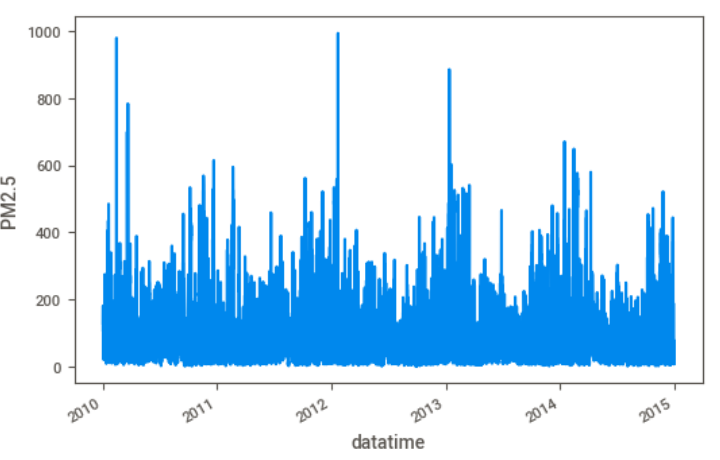
## Examining the dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Datasets** | **Features** | **Shape**  **(No. of rows)** | **Missing Values** | **Duplicate Rows** |
| Beijing+PM2.5+Data | 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 pm2.5: PM2.5 concentration (ug/m^3) DEWP: Dew Point (â„ƒ) TEMP: Temperature (â„ƒ) PRES: Pressure (hPa) cbwd: Combined wind direction Iws: Cumulated wind speed (m/s) Is: Cumulated hours of snow Ir: Cumulated hours of rain | 43824 | pm2.5 has 2067 missing values, delete these data | 0 |

## Graphs and Data characteristics:

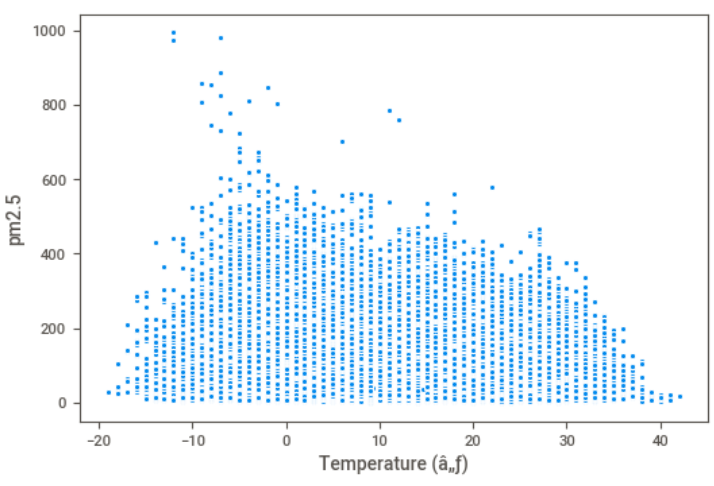
**Distribution of Label:**

We can see clear see that the PM2.5 rises and falls periodically, we need to analysis what factors make it rise and falls, then predict the next day pm2.5

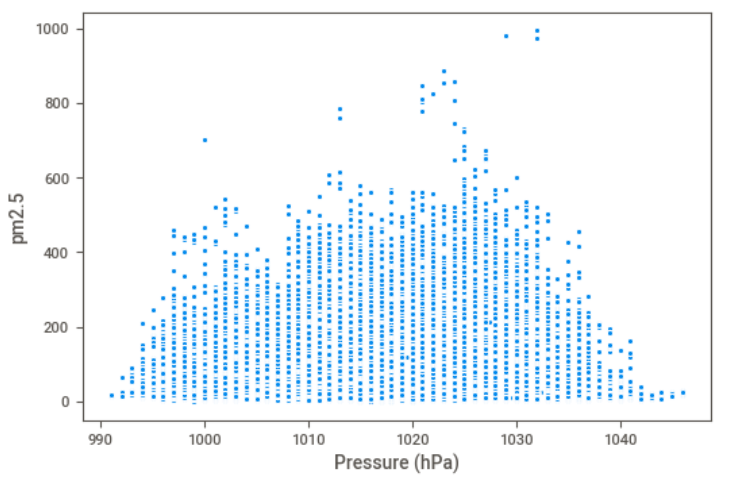


**The correlation with each features:**

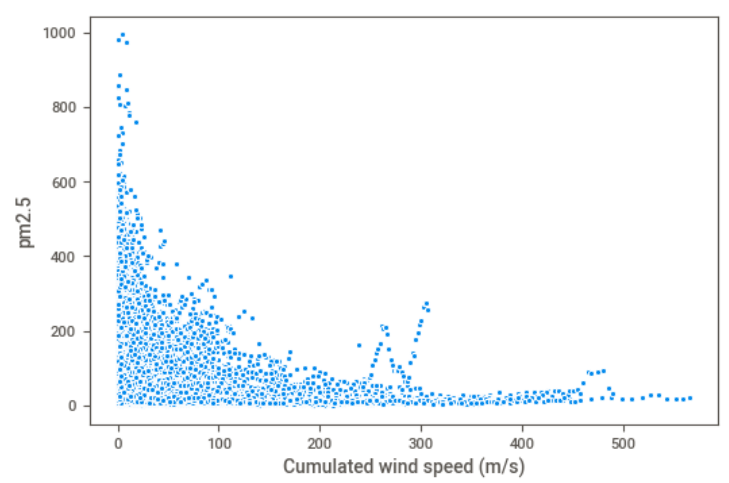
1. We can see that when the temperature at around 0 to -10, the PM2.5 are higher than others, the PM2.5 is decreasing when temperature is less than -10



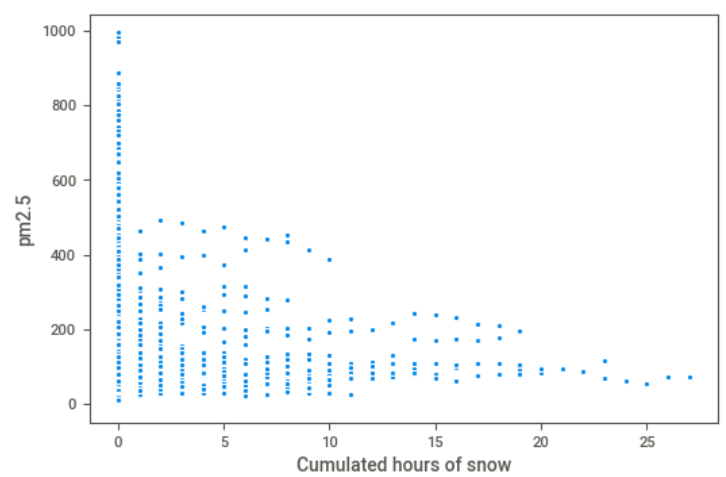
1. We can’t see very clearly correlation with pressure, seems when the pressure is between 1000-1030, it has more chance of high pm2.5



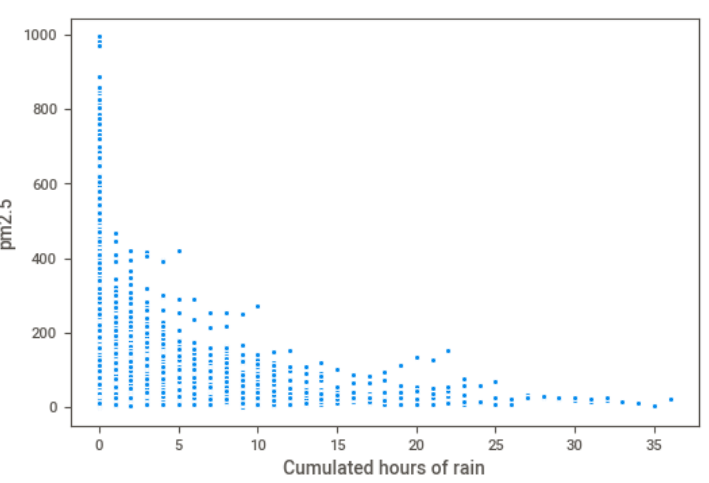
1. We can see that accumulated wind speed has negative correlation with PM2.5



1. We can see that snow can help to reduce PM2.5, the more snow the lower pm2.5

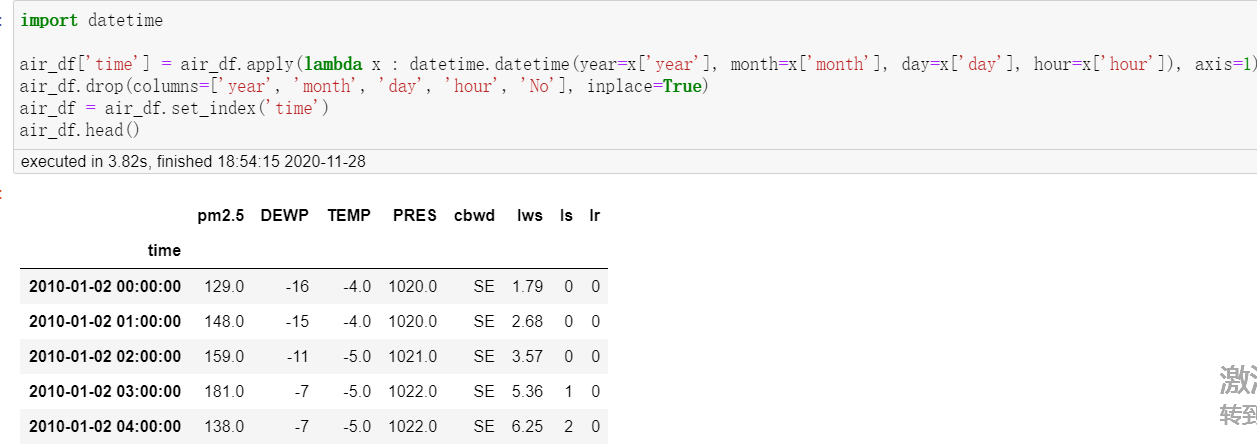


1. Same to snow , We can see that rain can help to reduce PM2.5, the more rain the lower pm2.5

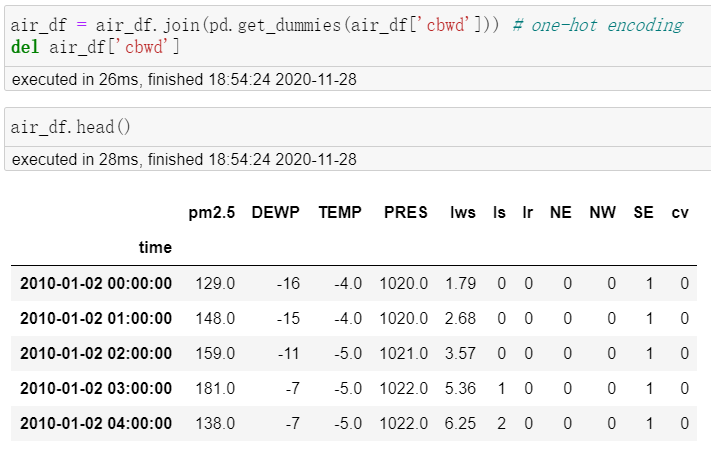


**Data Featuring:**

1. We combine year,month,day, hour to datetime as the index of data

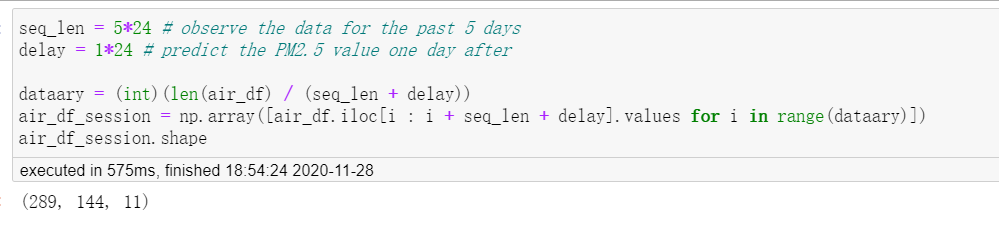


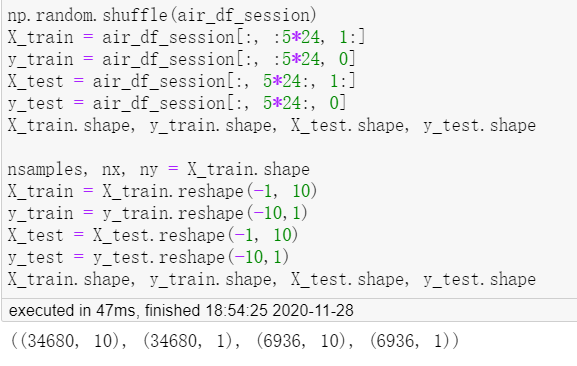
1. Use one-hot encoding to split the values of cbwd to different columns.



1. Split the dataset to train and test the dataset.

put the data for the past 5 days to train  
put the data one day after to test set and predict the PM2.5 value one day after





## Assumptions

* The dataset is valid and comes from a reliable source
* The dataset does not represent a sudden extreme air condition.
* We assume the next day's air condition based on the last five days' air condition.

## Constraints

* The additional feature or attributes cannot be added in the dataset

# Algorithm Evaluation

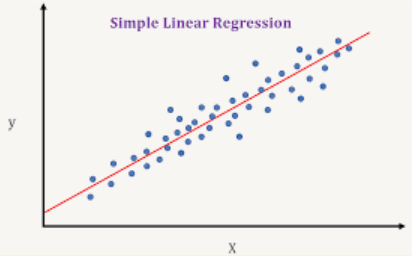
The project is to predict the value of pm2.5, so it’s a regression problem. We’ll narrow down the model to linear regression and polynomial regression.

1. **Linear Regression**

Linear regression is simply a linear method to model the relationship between your independent variables and your dependent variables. Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.

**Equation**

Y= aX +b



**Pros and Cons**

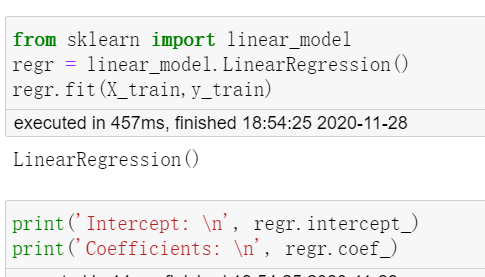
Pros: has the least complex to compared to other algorithms

**Cons**:

There aren’t many problems in the world that exhibit a clear relationship between the independent and dependent variables.

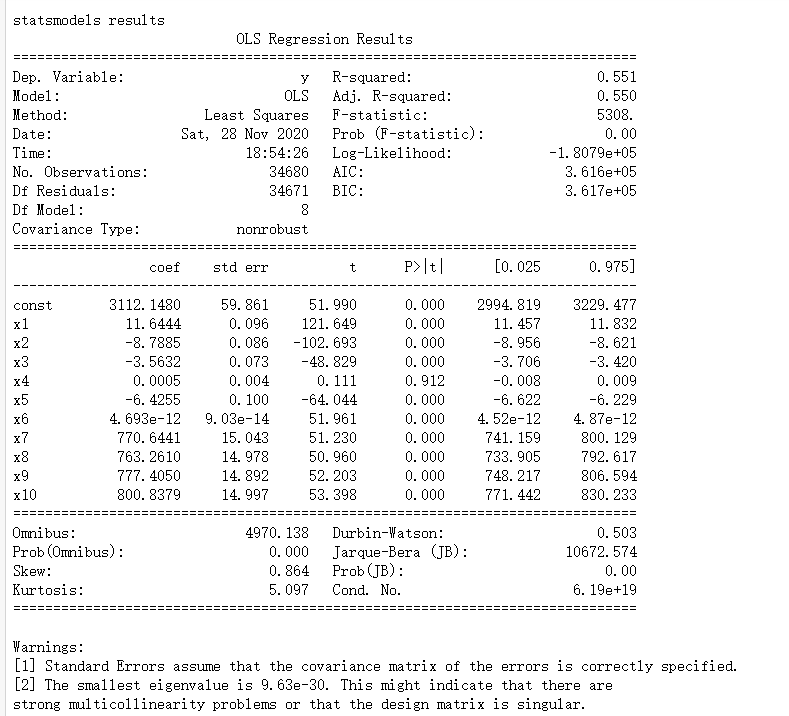
Outliers are other that make linear regression more limited

**Code**



**Evaluation**

We add statsmodels to evaluate our mode, we can see the R-Squared Score is only 55, so the linear model is not fit for our dataset.

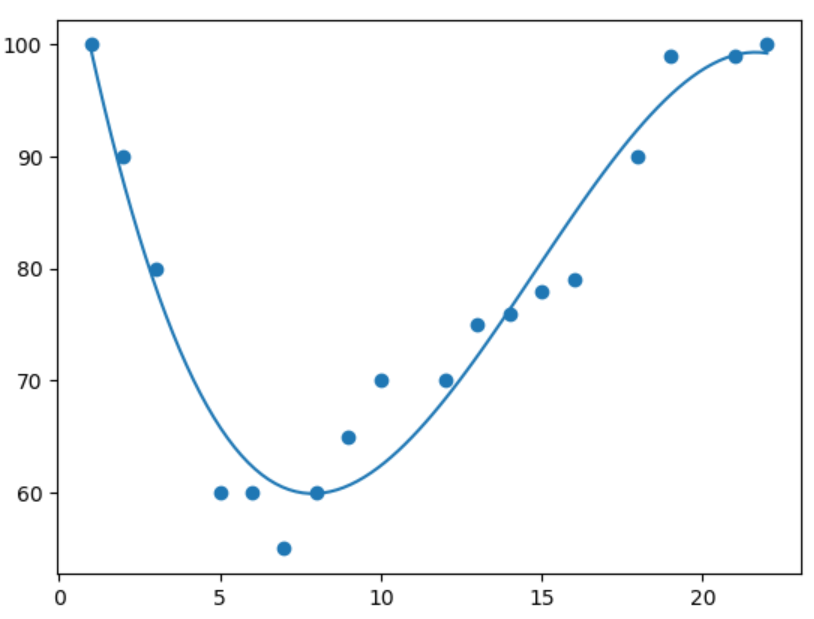


**Polynomial Regression**

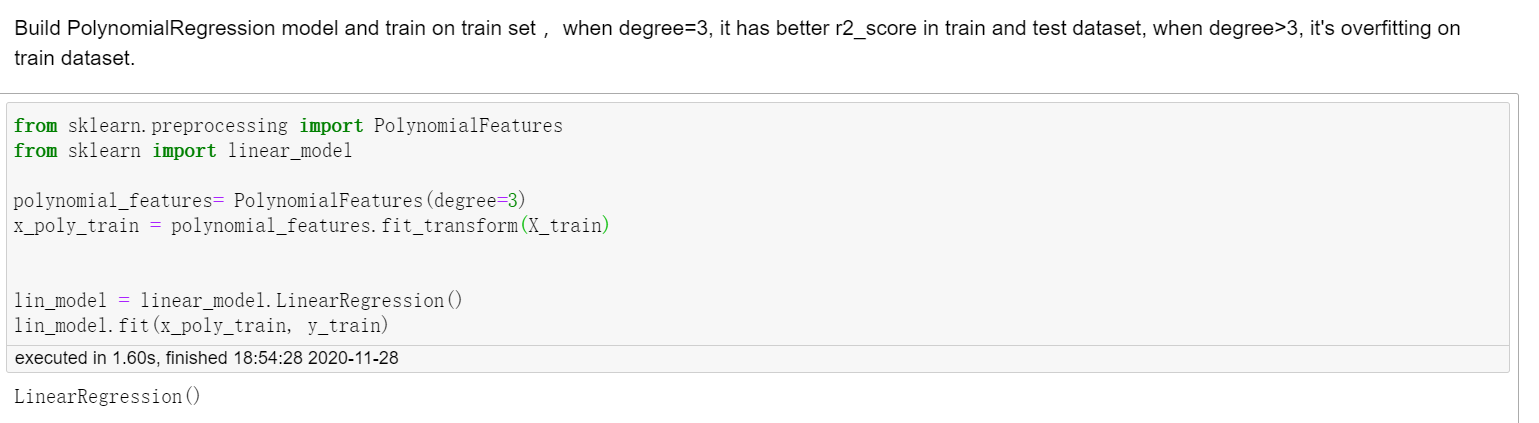
Polynomial regression, like linear regression, uses the relationship between the variables x and y to find the best way to draw a line through the data points.But the relation is no liner.

**Equation**





**Code**



**Evaluation**

**We use R-Square, Mean root error, Root mean squared error to evaluate our model, we can see the R-Square score is 81, it improved a lot than the linear regression. But still need to improve,   maybe later we can try natural network models to predict.**

