**Assignment 1**

**Data Exploration and Regression Analysis**

**Semester 1 2022**

**Student Name: Ruiyang Xu**

**Student ID: 18017865**

## **PAPER NAME:** Foundations of Data Science

**PAPER CODE:** COMP615

**Due Date:** Friday 15 April 2022 (midnight)

**TOTAL MARKS:** 100

**Instructions:**

1. **The following actions may be deemed to constitute a breach of the General Academic Regulations Part 7: Academic Discipline,**

* Communicating with or collaborating with another person regarding the Assignment
* Copying from any other student work for your Assignment
* Copying from any third-party websites unless it is an open book Assignment
* Uses any other unfair means

1. **Please email** [**DCT.EXAM@AUT.AC.NZ**](mailto:DCT.EXAM@AUT.AC.NZ) **if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
2. **Attach your code for all the datasets in the appendix section**.

Assignment 1

Data Exploration and Regression Analysis

Introduction

**This Assignment will use the dataset Beijing PM2.5 data as the research material. And the reason which is why I chose this dataset is the increase concentration of PM2.5 has already led to the serious problem of haze in northern China. Especially in winter, the air in northern China is often heavily polluted, which is seriously harmful to human health. The concentration of PM2.5 in the air is mainly determined by the emission of pollution sources and different meteorological conditions, such as temperature, wind speed, air pressure, etc. The concentration of PM2.5 sources in Beijing are mainly automobile exhaust emissions and industrial areas. High density energy consumption leads to serious air pollution problems in Beijing. The PM2.5 are suspended in the air for a long time, which is the main reason for the formation of haze weather. Besides, in northern China, the total emission of air pollutants is large and concentrated, which is an important reason for the high PM2.5 concentration in the air, and the adverse meteorological conditions make the pollution more serious. Meteorological factors can affect the dilution, diffusion, accumulation and removal of air pollutants, lead to the transmission of pollutants between regions under the action of atmospheric flow field, and quickly form air pollution characterized by high concentration of PM2.5 under the condition of high humidity and static weather. Therefore, this assignment will mainly focus on the relationship between the PM2.5 and the other weather factor such as Dew Point, Temperature, the wind direction and speed, pressure, raining time and different time periods would affect the concentration of PM2.5. With the result of this assignment. I would assume and find the concentration of PM2.5 will have some relationship with weather and various time. For the question that will be discussed, the research area will be in the entire Beijing city and the resource for this research have umber of PM2.5, Dew Point, Temperature, wind direction, wind speed, pressure, raining time and a select period time (hours) between the March 1 2010 to December 31 2014.**

Task 2:

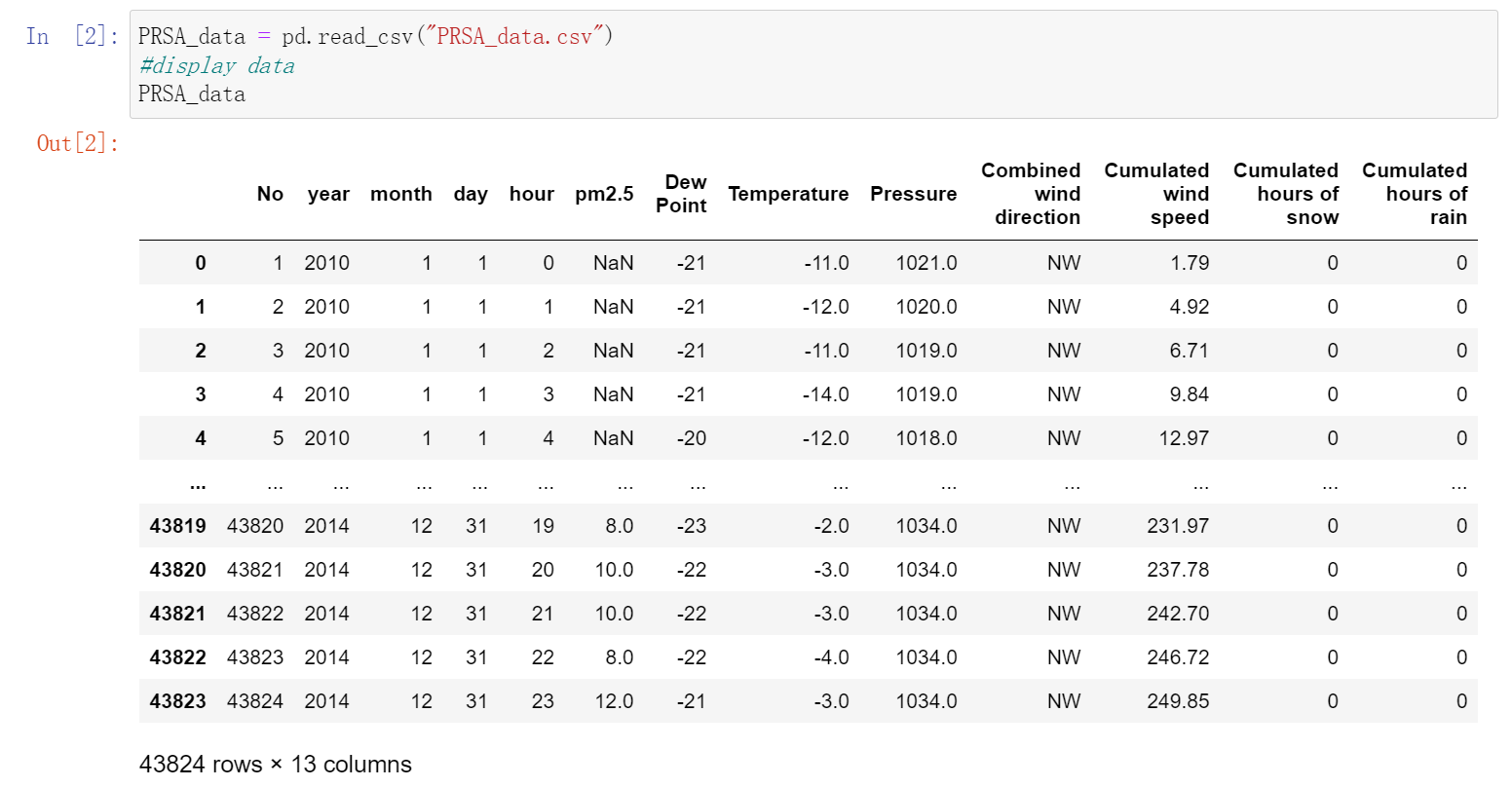
1. What is the data about?

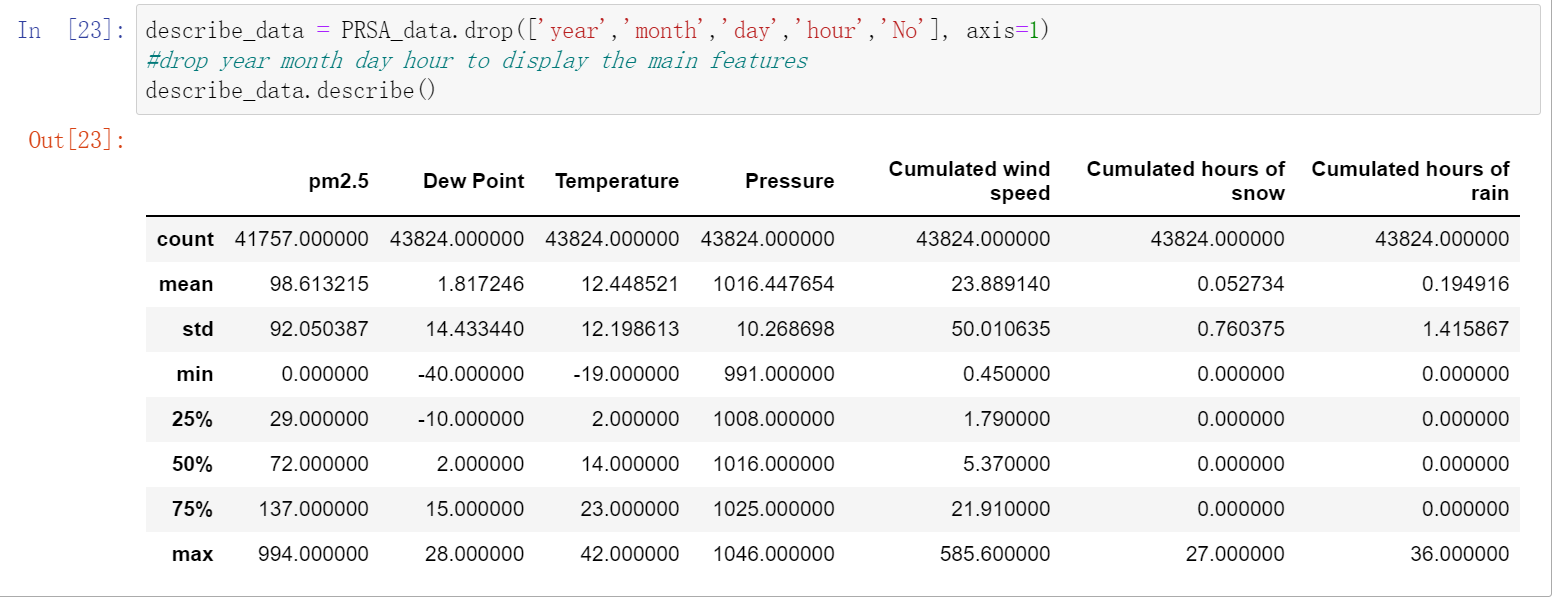
**These data sets are collected from different time period of Beijing, it has recorded the main features pm2.5 and other data that might affect the pm2.5, such as Dew Point, Temperature, wind direction, wind speed, pressure, raining time and a select period time (hours). For example, the method of human kind’s production, life activities and industrial activities. Different times might also affect the concentration of pm2.5 and strong wind could disperse the pm2.5 in the atmosphere.**

1. How many features (attributes), instances and what data types are these?

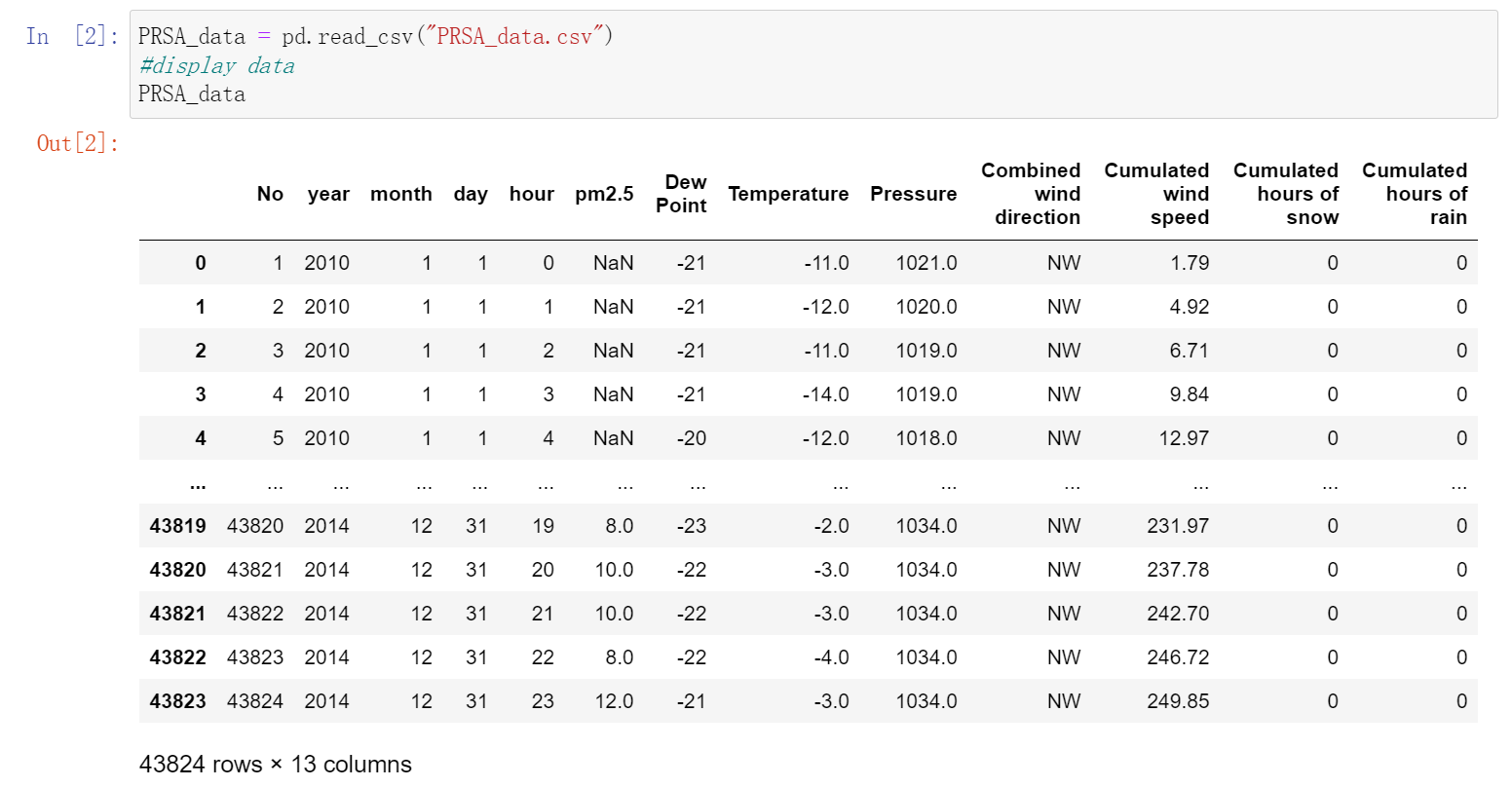
**The main features are: concentration of pm2.5(ug/m^3), Dew Point (â„ƒ), Temperature (â„ƒ), Pressure (hPa), Combined wind direction, Cumulated wind speed (m/s), Cumulated hours of snow, Cumulated hours of rain and the hour, day, month and year.**

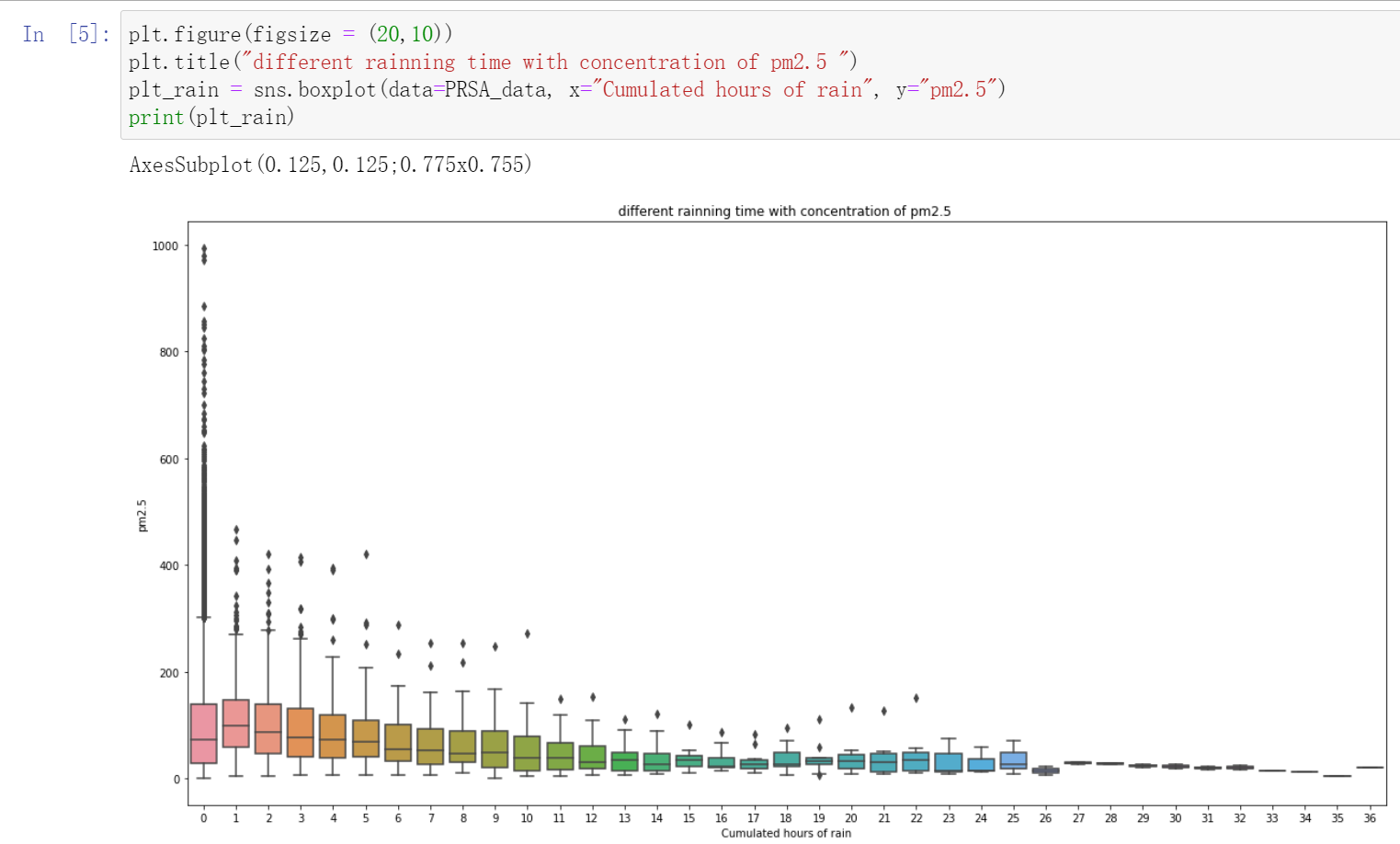
1. Provide summary statistics of the continuous numerical features.

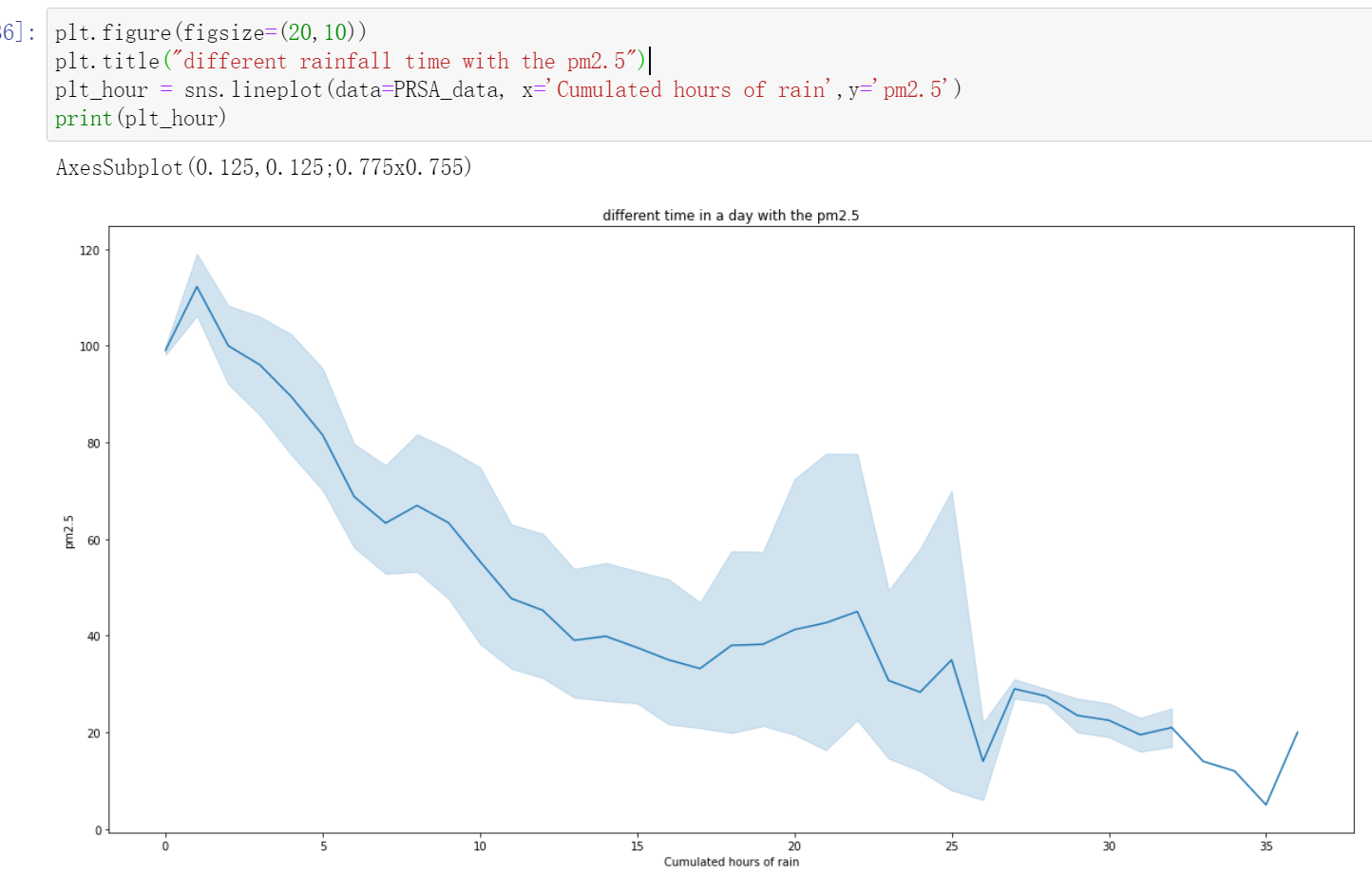


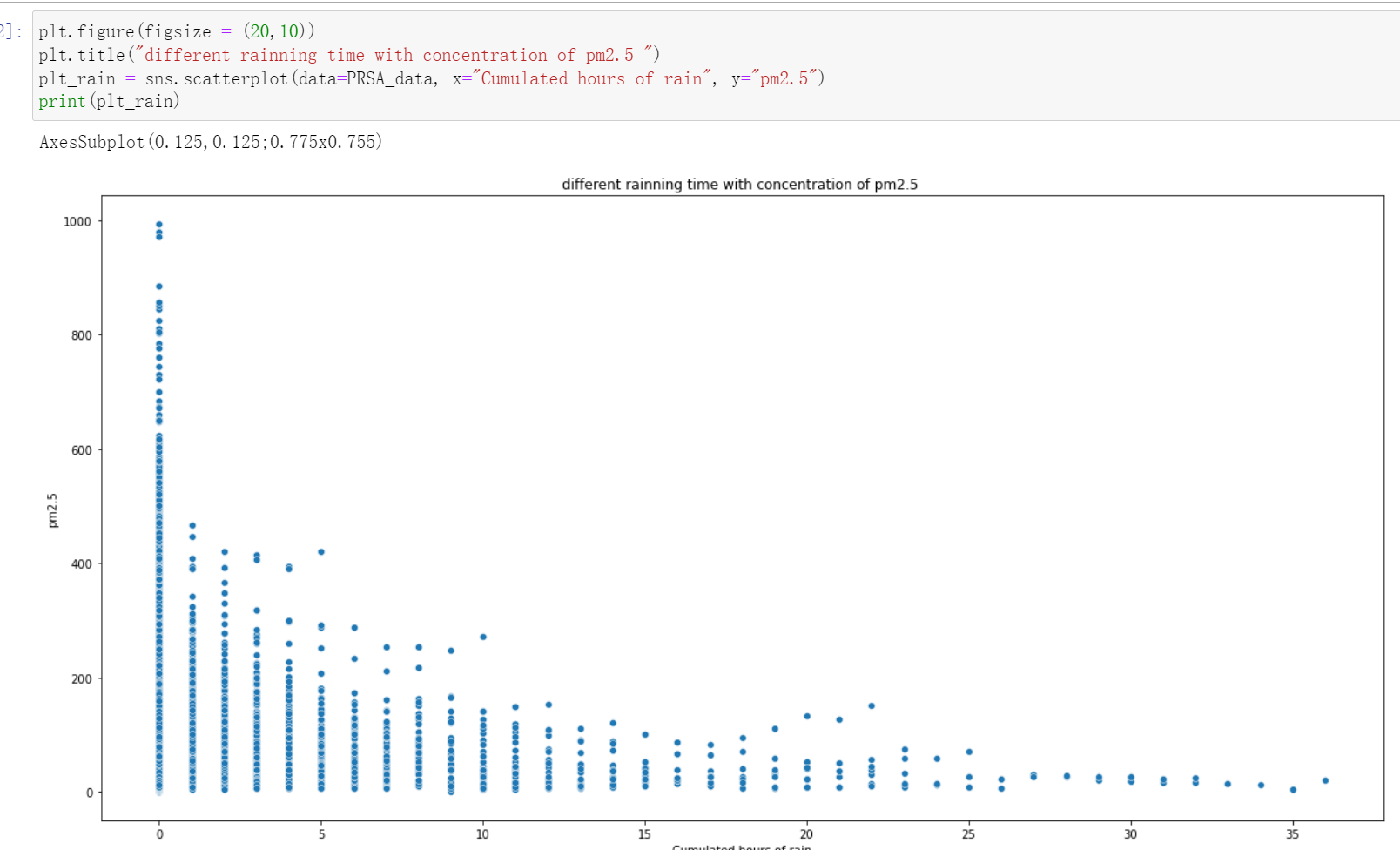


1. Illustrate the features of your dataset using meaningful boxplots, histograms and grouped scatter plots (remember, these plots allow you to analyses the individual distribution of features, as well as the relationship between them).

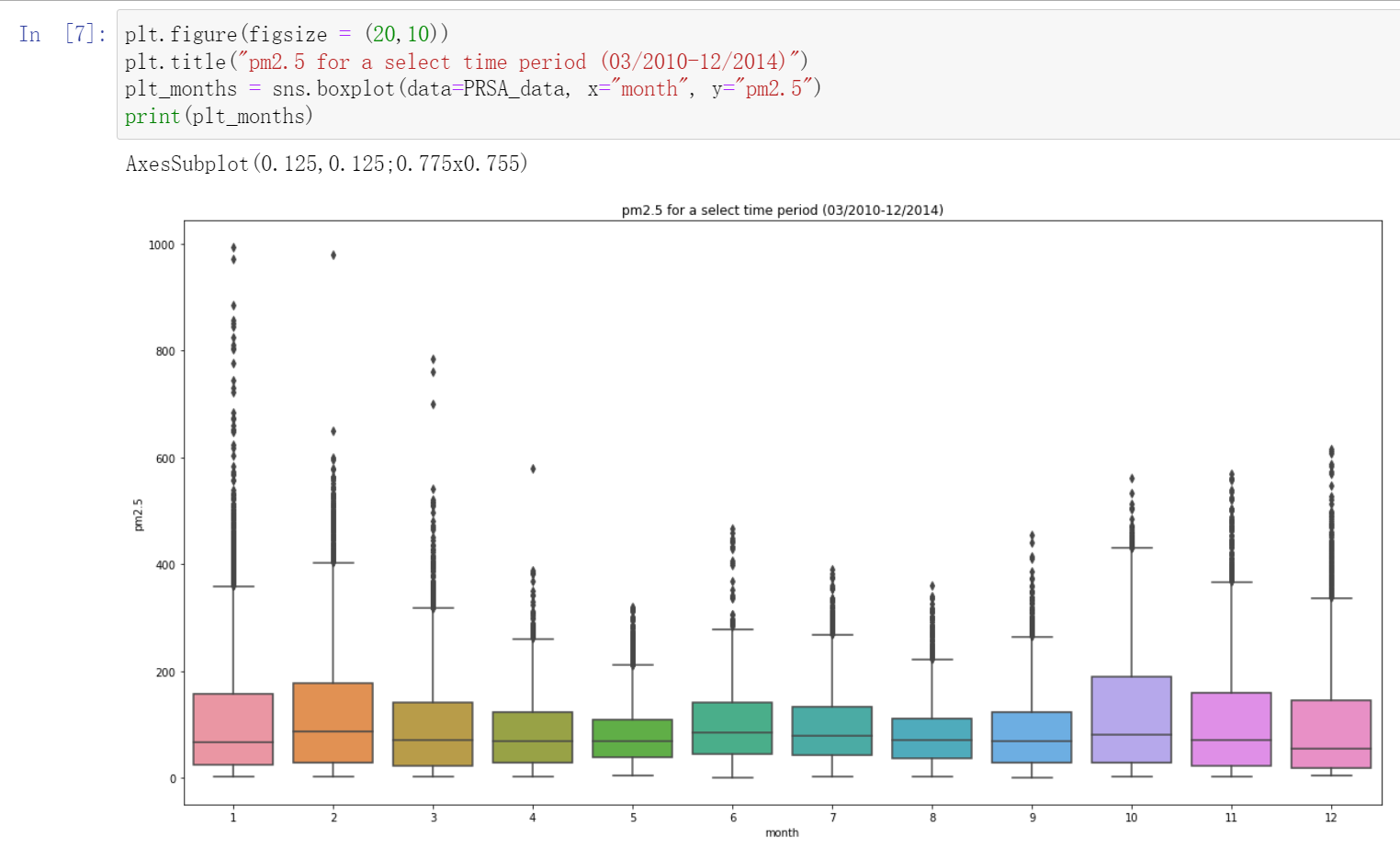




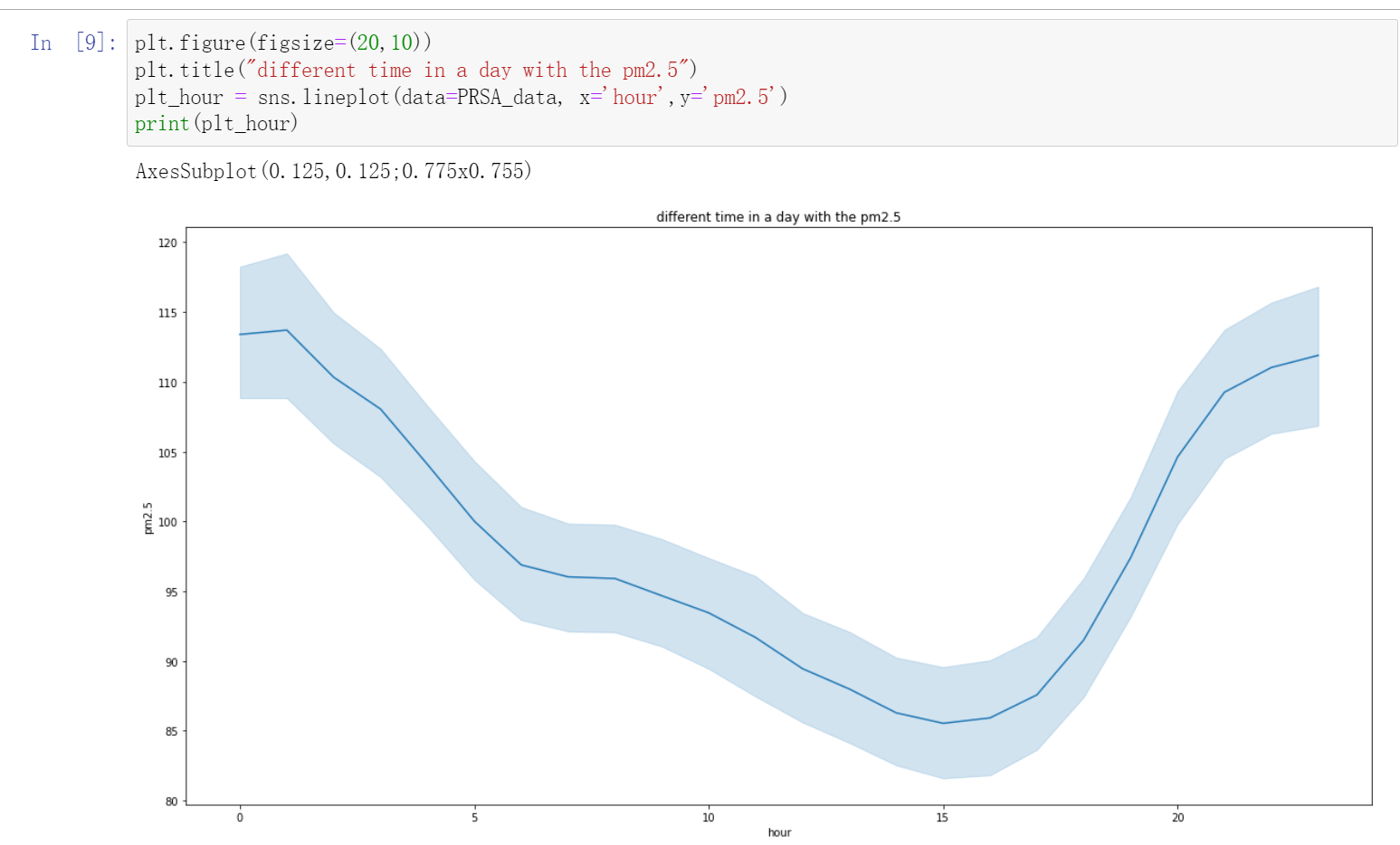




**Depending on the time of rainfall, the pm2.5 concentration would have two high concentration area. In the 0-4 hours rainfall, concentration of pm2.5 will remain at a high value (100ug/m^3-300ug/m^3), especially no raining time (0 hour), it has the highest concentration of pm2.5. However, With the increase of rainfall time, it will make the pm2.5 decrease, almost under 100 ug/m^3.**

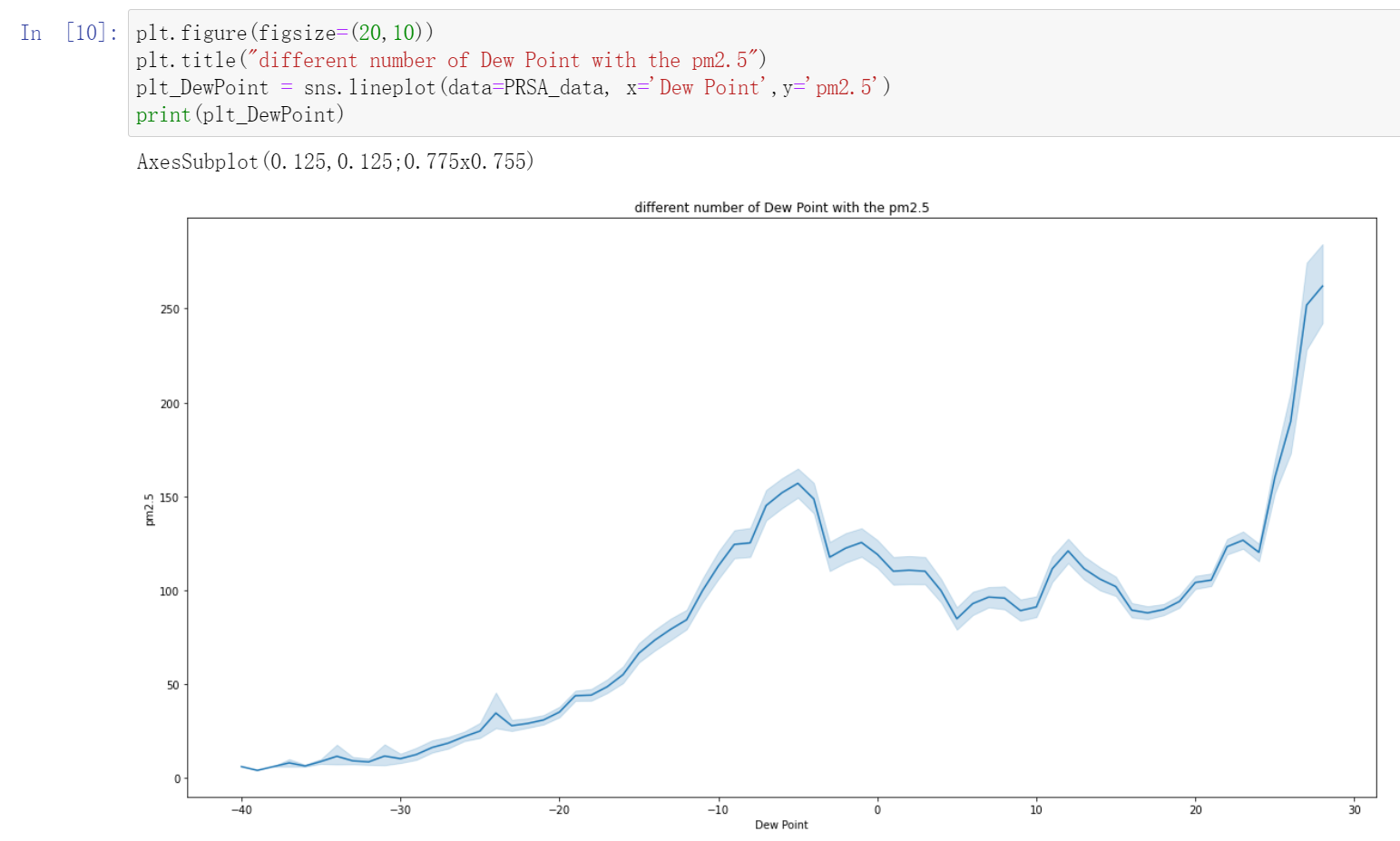


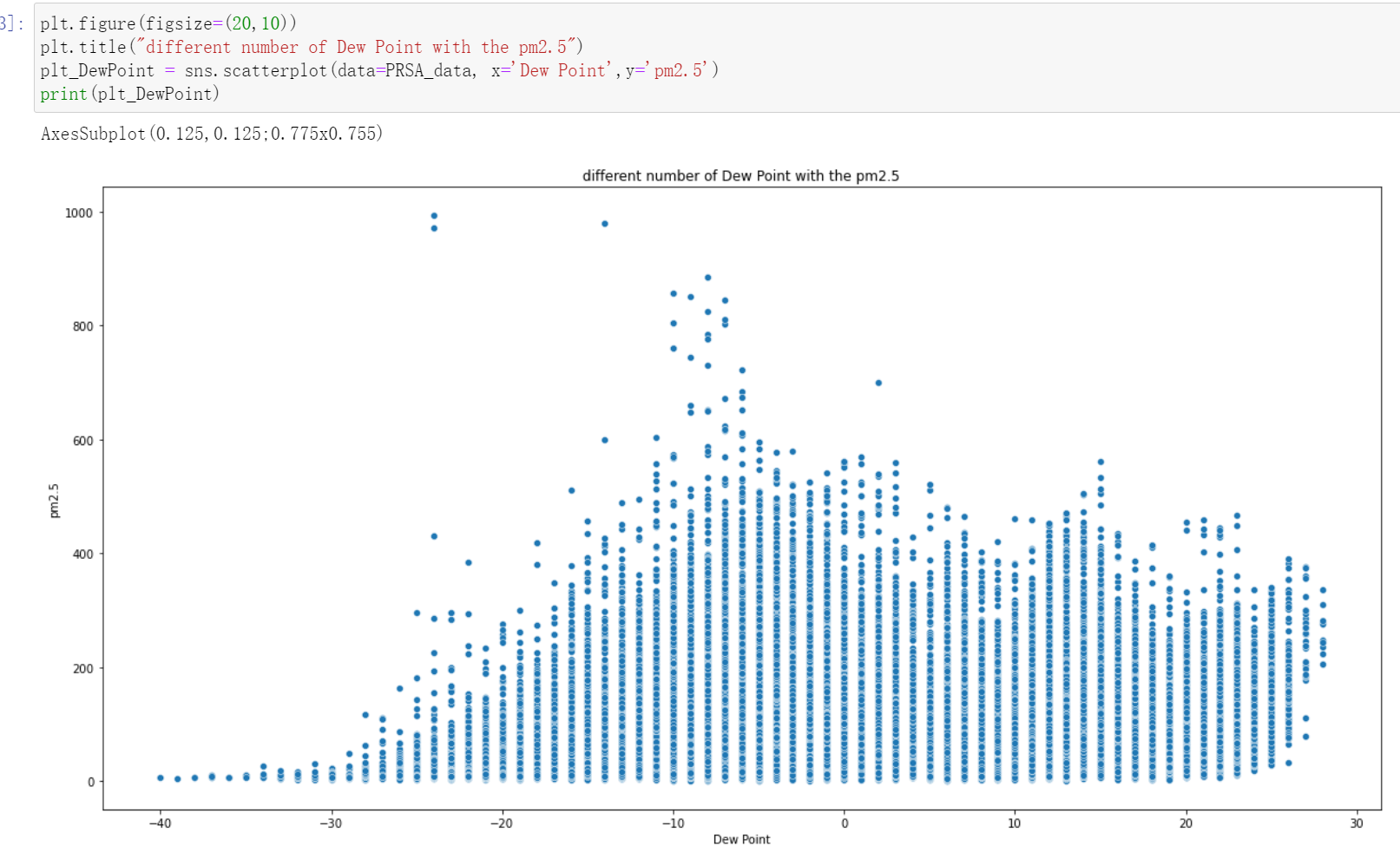
**Only have 3 months (May, July. August), the concentration of pm2.5 would maintain a stable and low value; however, the other time will keep a high value of pm2.5, especially the January and February, it always maintains a vary level of concentration of pm2.5.**



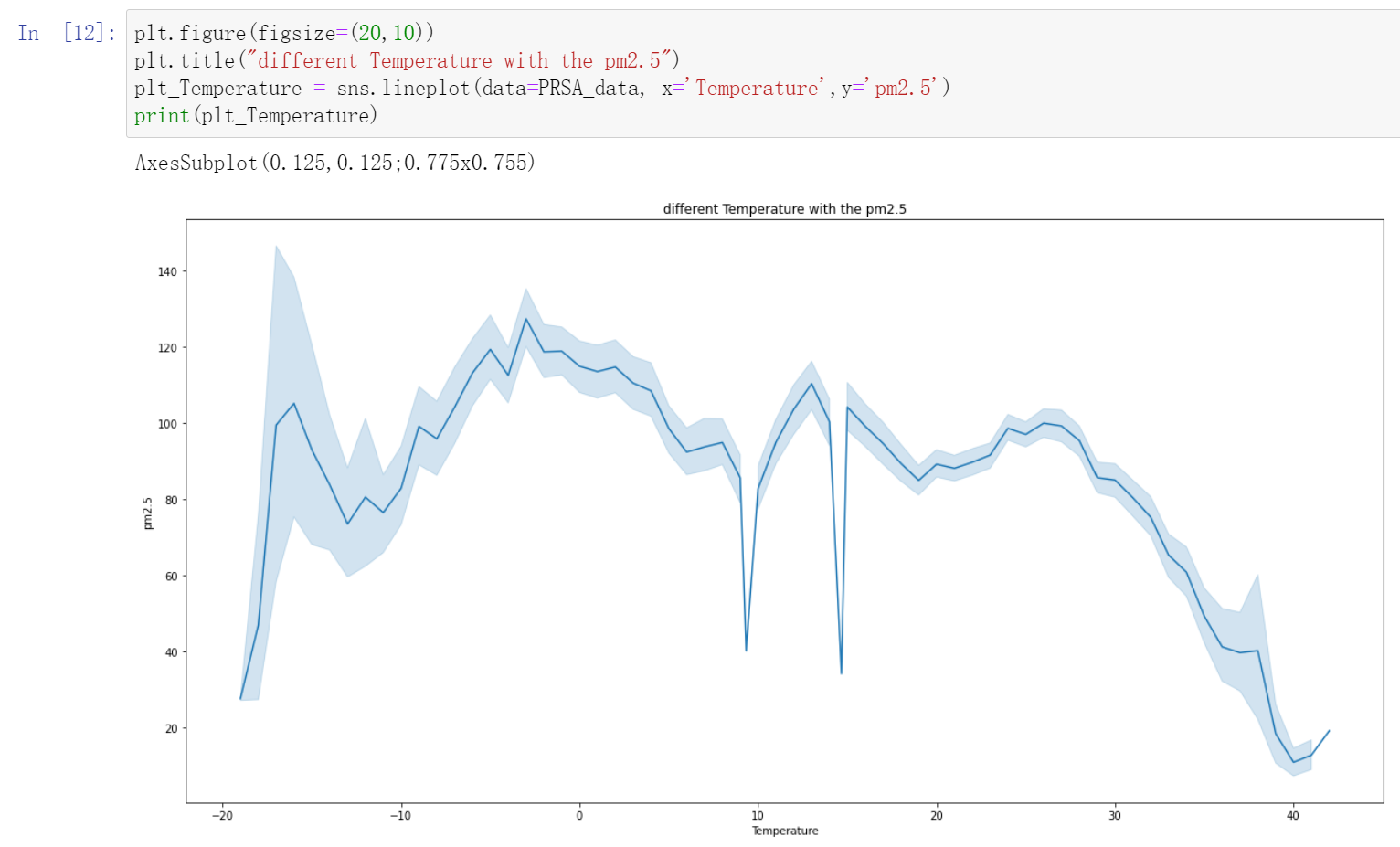


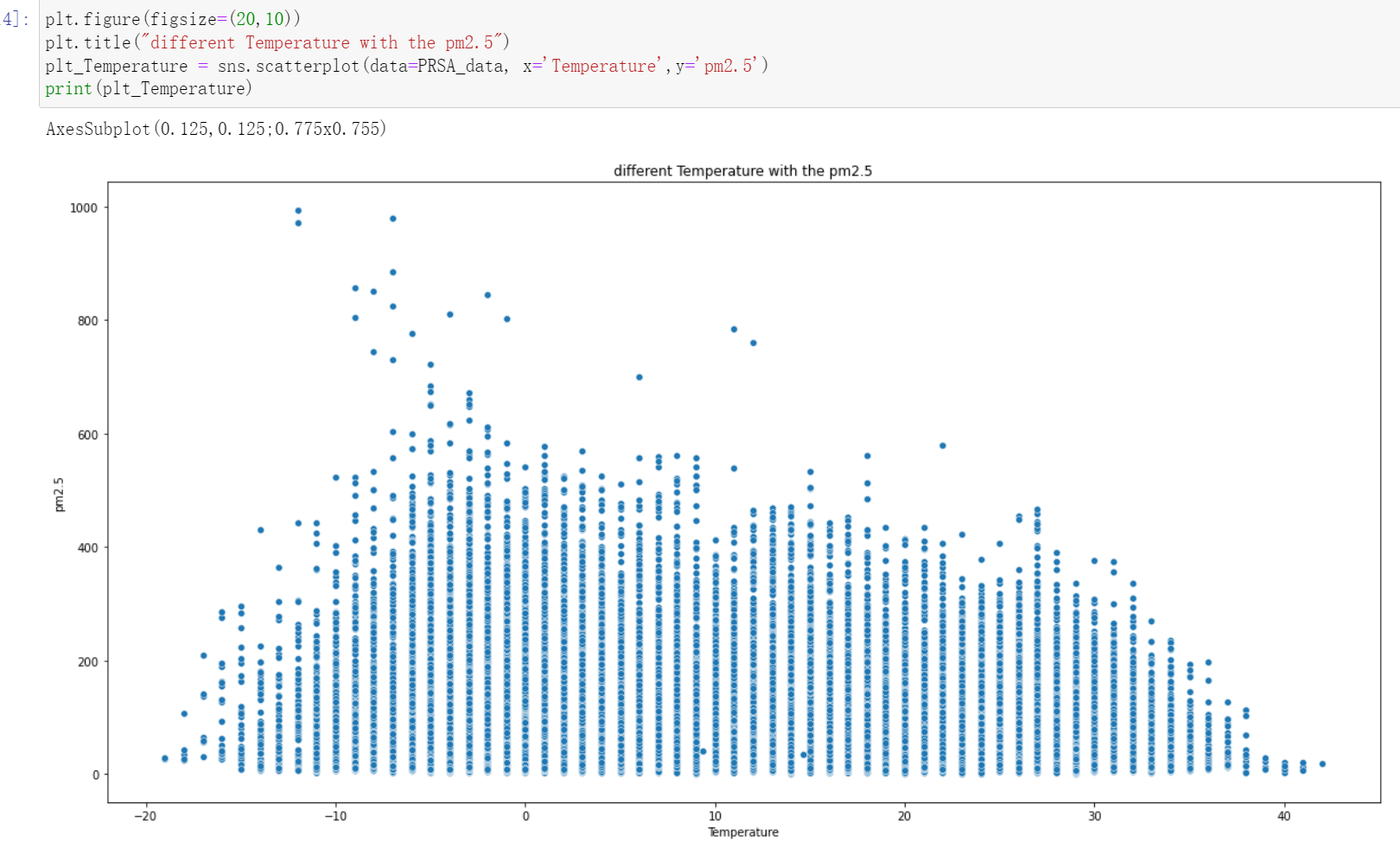
**According to this plot, the high level pm2.5 will appear in the midnight and it will keep decrease to a low concentration until to 15pm, after 15pm it will increase again.**



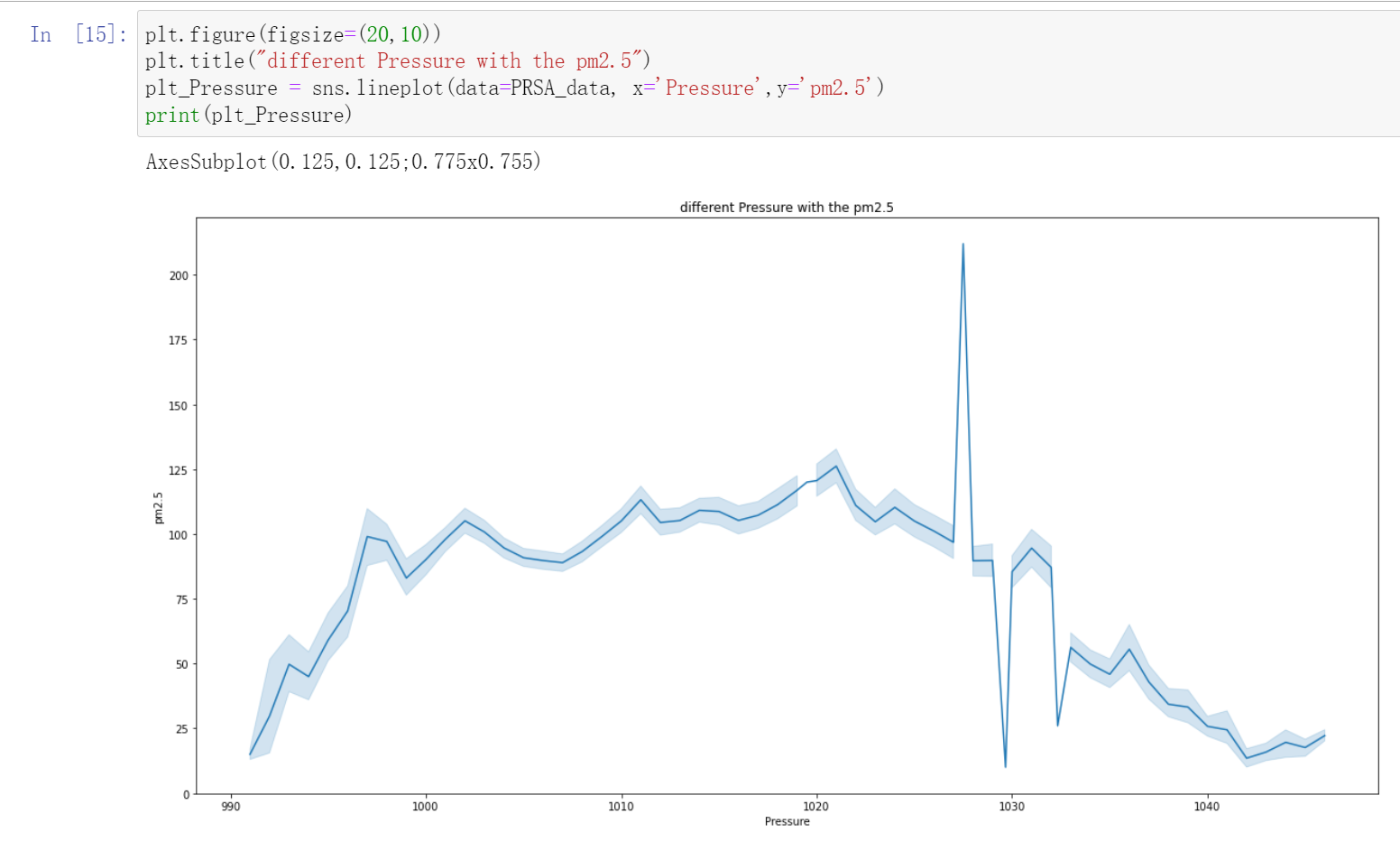


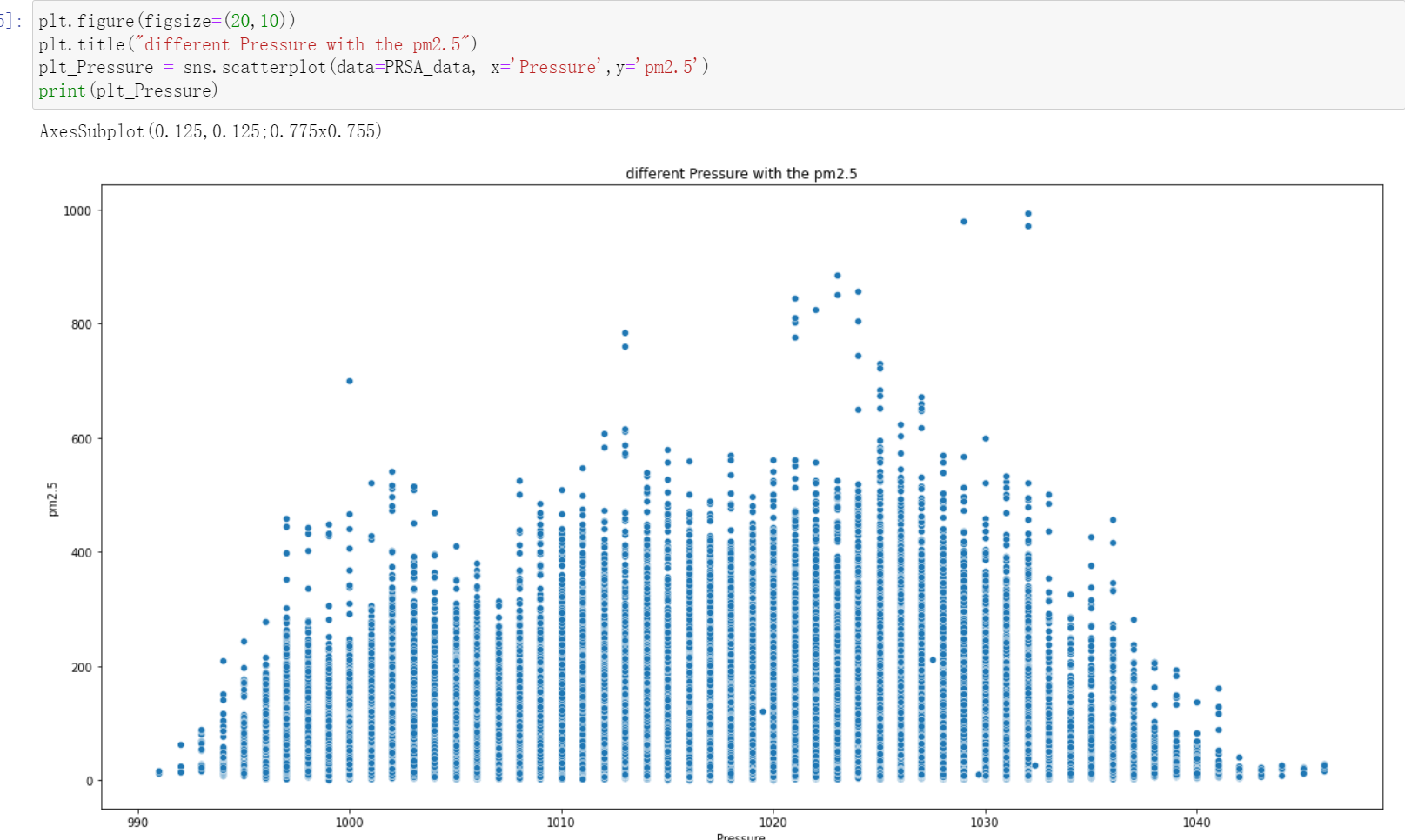
**According to the plot, the pm2.5 has shown an increase correlation trend. Higher dew point will bring a higher concentration of pm2.5, but when the Dew Point reach -6 â„ƒ to 5 â„ƒ, it will have an obvious downward trend. After that, the pm2.5 will keep going increase.**



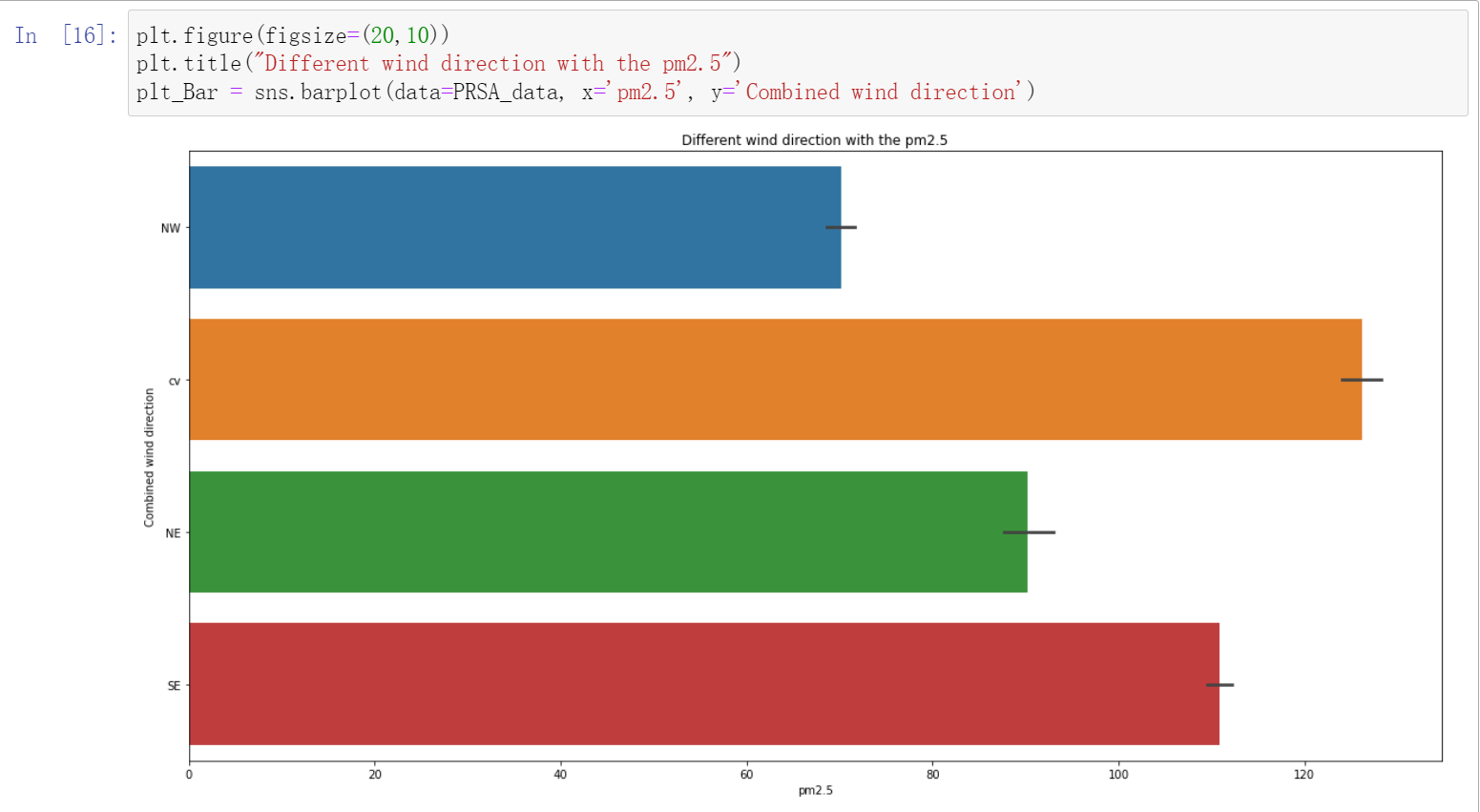


**According to this plot, the pm2.5 will keep increase when the temperature has increase from-20℃ to -2℃ even there is a small decrease trend in near -15℃, but when the temperature is almost 0℃ and it will keep an obvious decrease.**

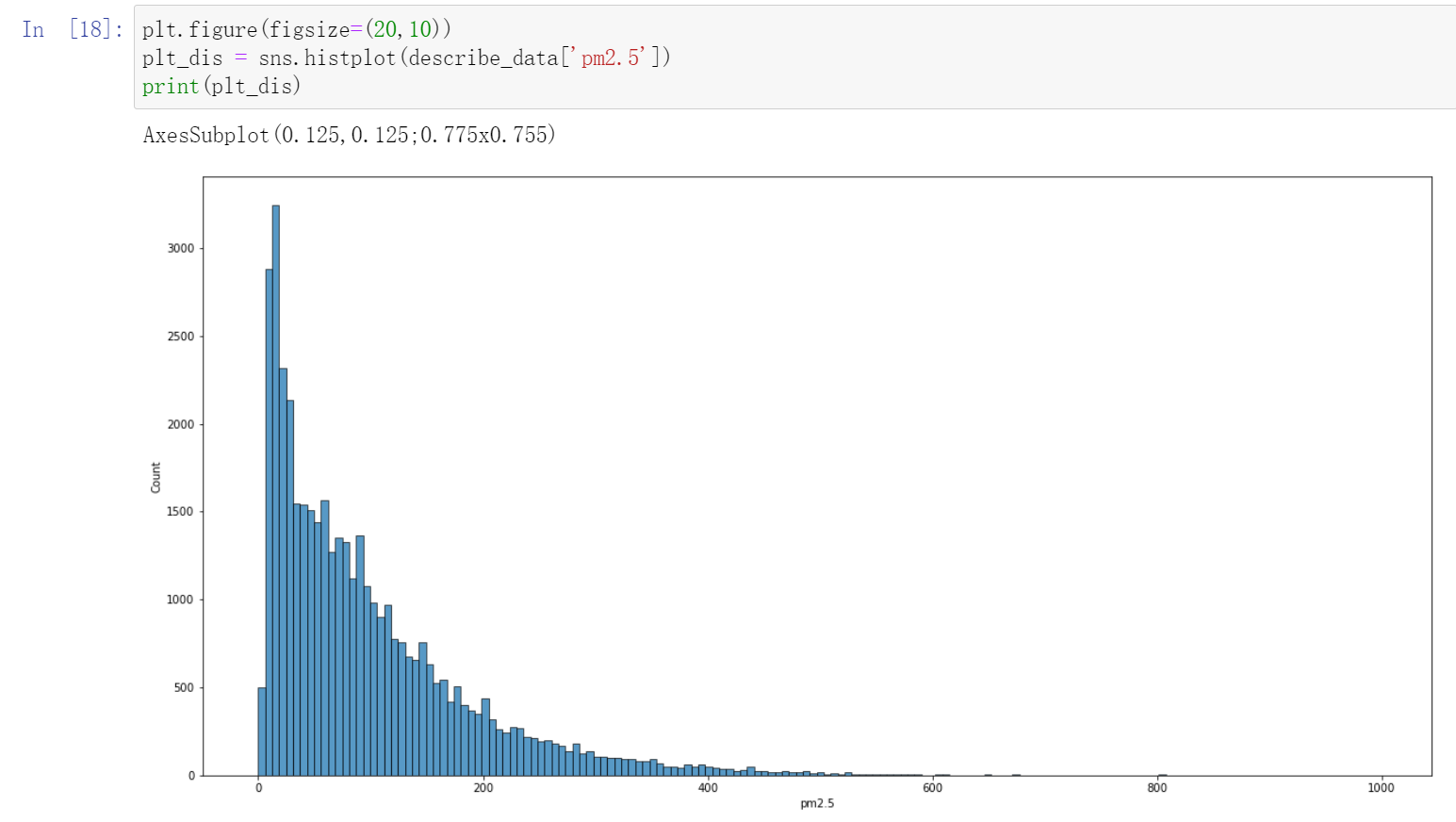




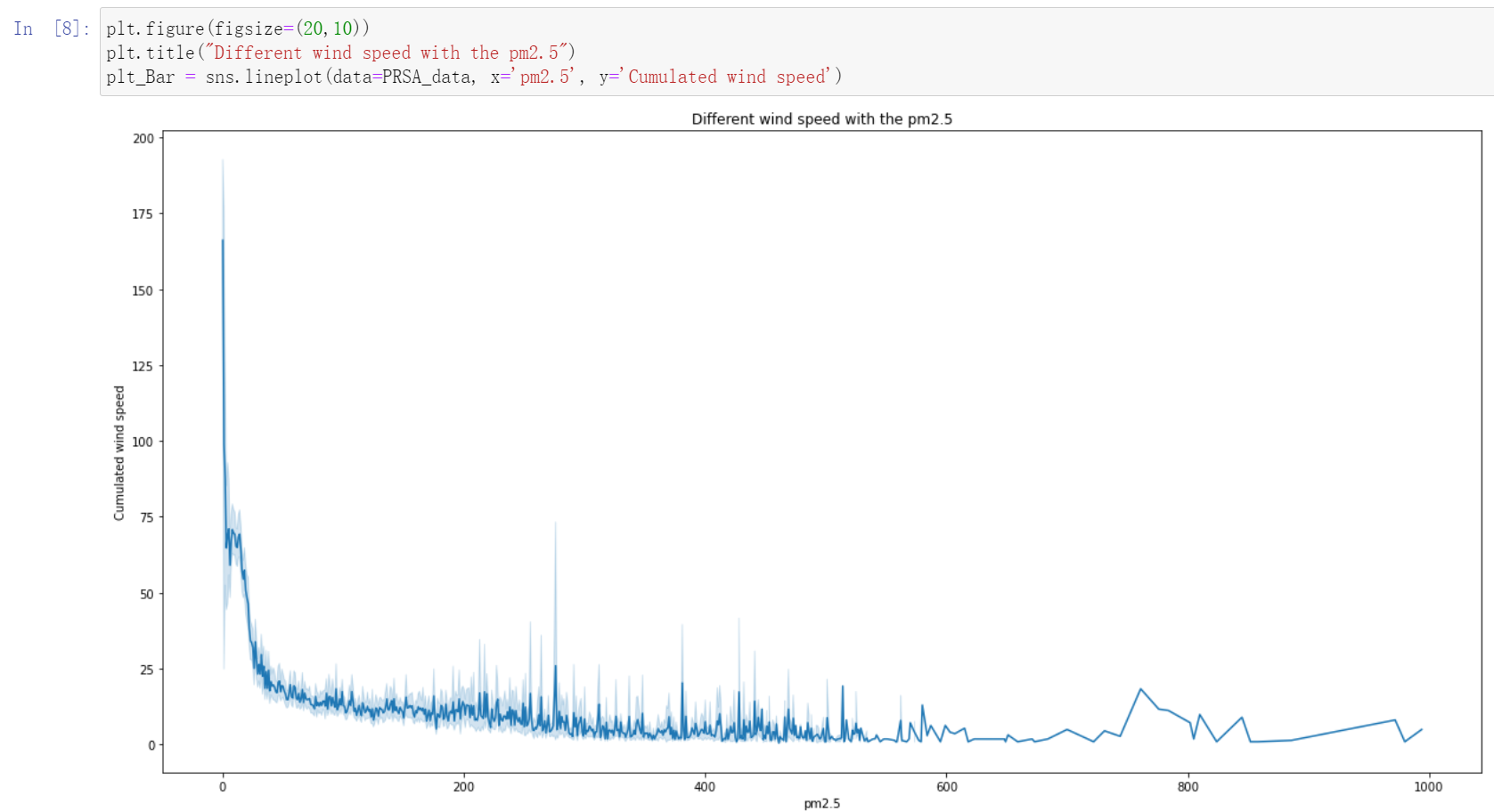
**According to this plot, the pm2.5 will increase when the pressure is 990 to 1020, however after that value, the pm2.5 will decrease with the increasing pressure.**

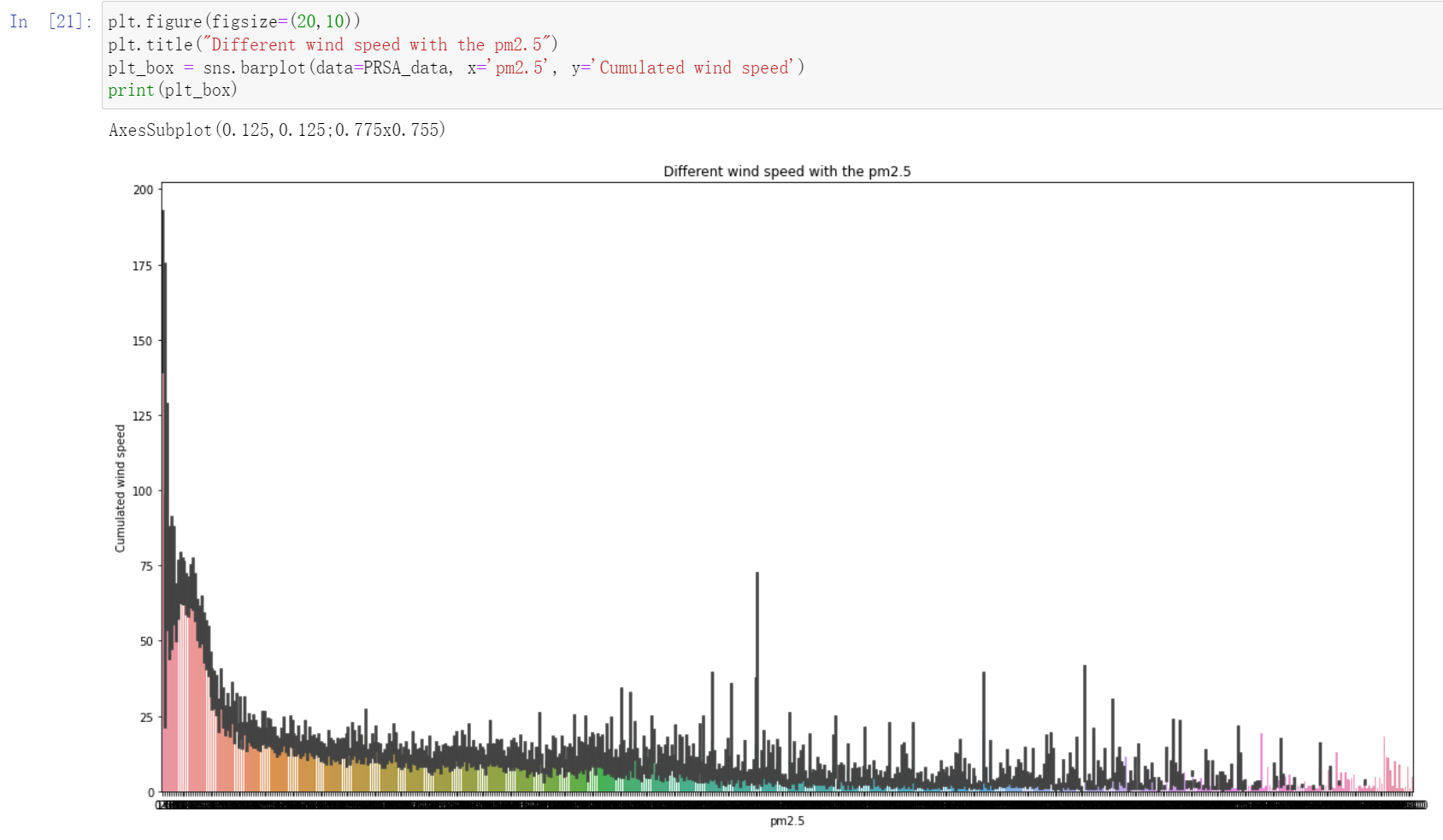


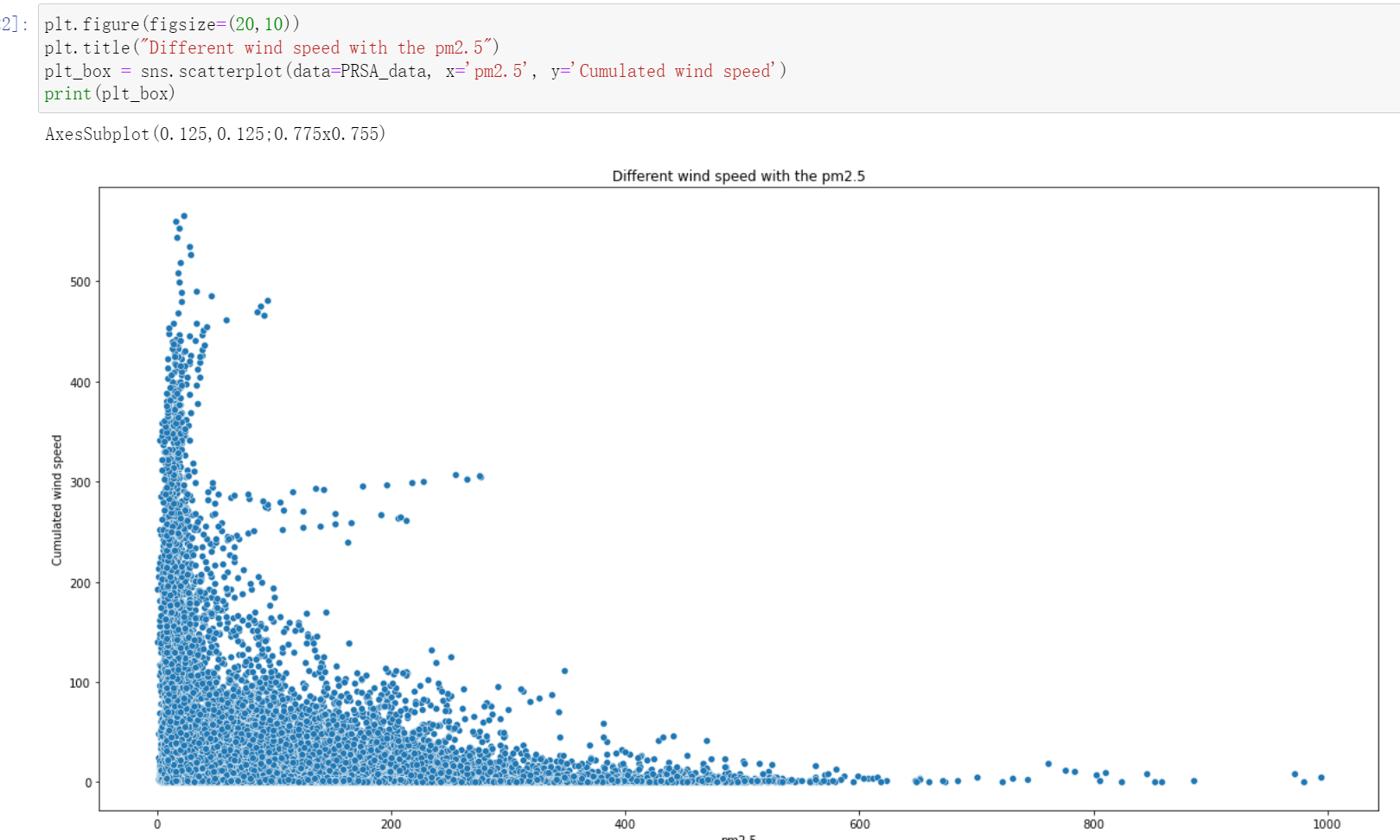
**According to this plot, different wind direction might affect the pm2.5. the direction NW would bring a low value of pm2.5 than is NE direction. The wind direction with highest pm2.5 is cv.**



**According to this plot, the most of value of pm2.5 concentration distribute in the range of 0 to 200 ug/m^3. The values which are over 200 ug/m^3 is maintain a low numbers.**







**According to plot, when the wind speed is under 50m/s, the value of pm2.5 is high and with the increase of wind speed, the concentration of pm2.5 will decrease at a high speed.**

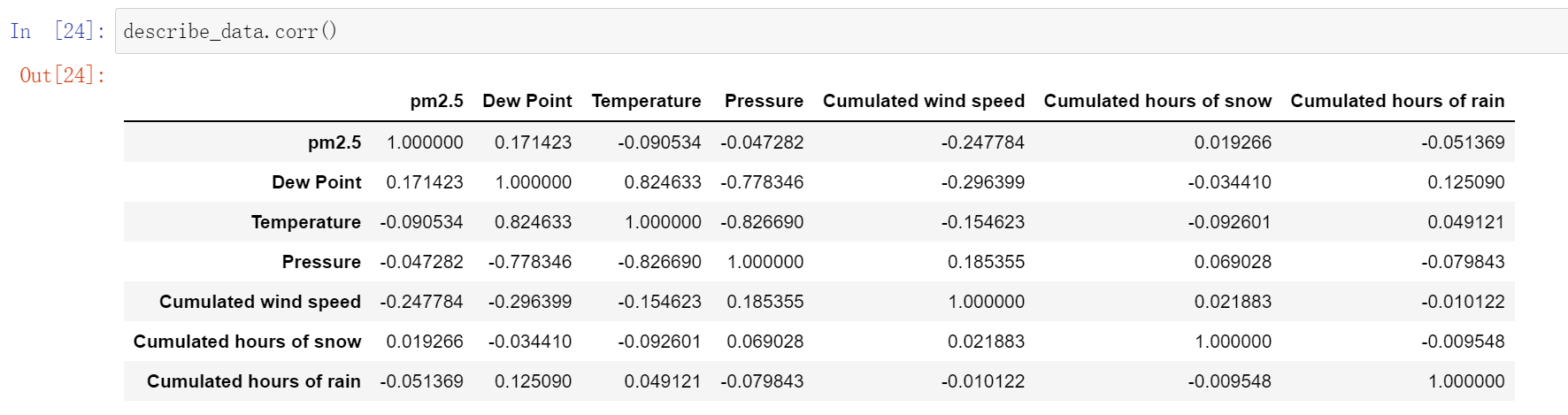
1. Explain what you can learn about the dataset from the diagrams. Is there any sign of violation of assumption(s)? If yes, explain your approach to handle it before moving to the next tasks?

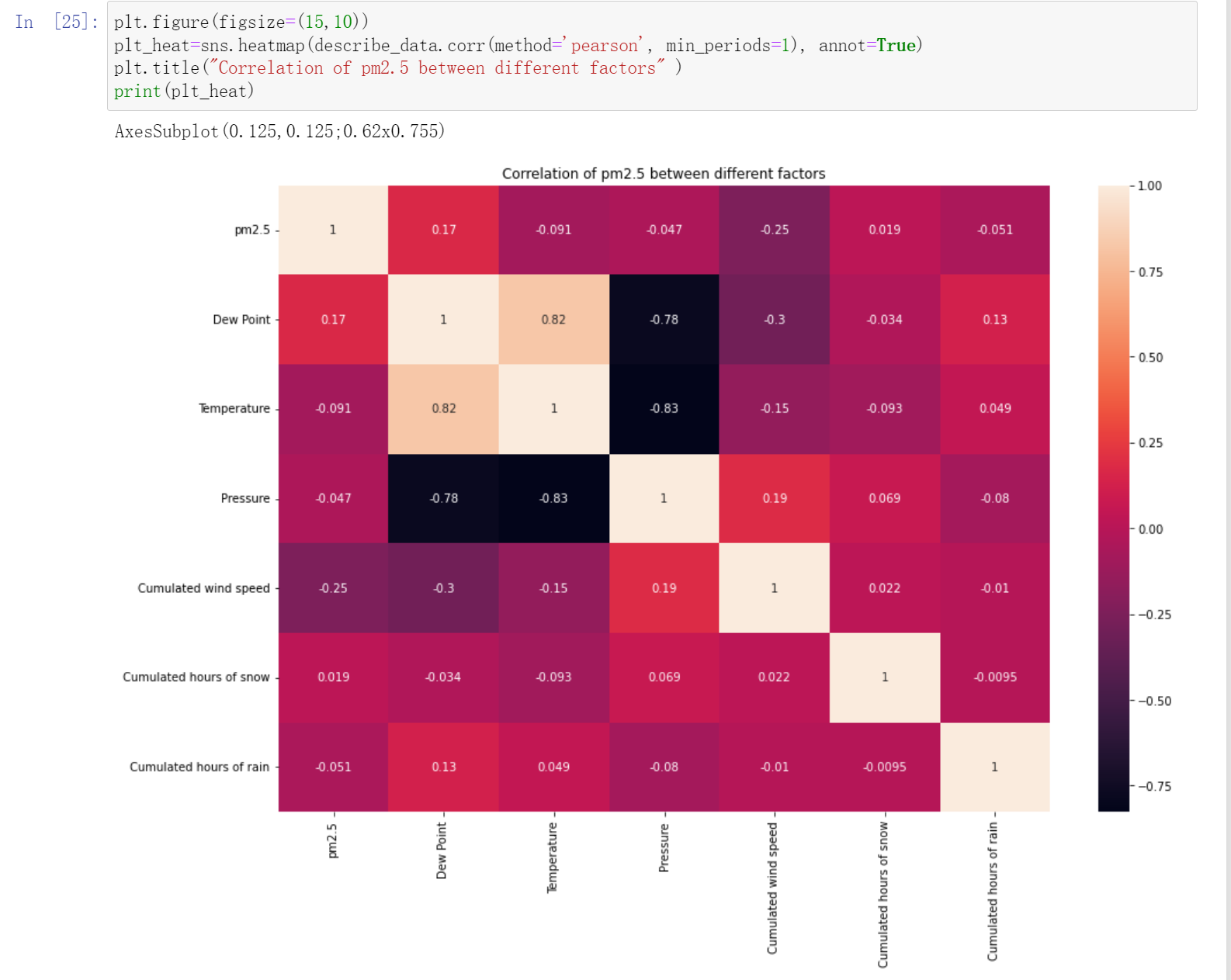
**Based on the obtained image analysis results, the concentration of pm2.5 has different relationship between the dew point, Temperature, pressure, different time period, wind speed and wind direction. No obvious violation of assumptions was found.**

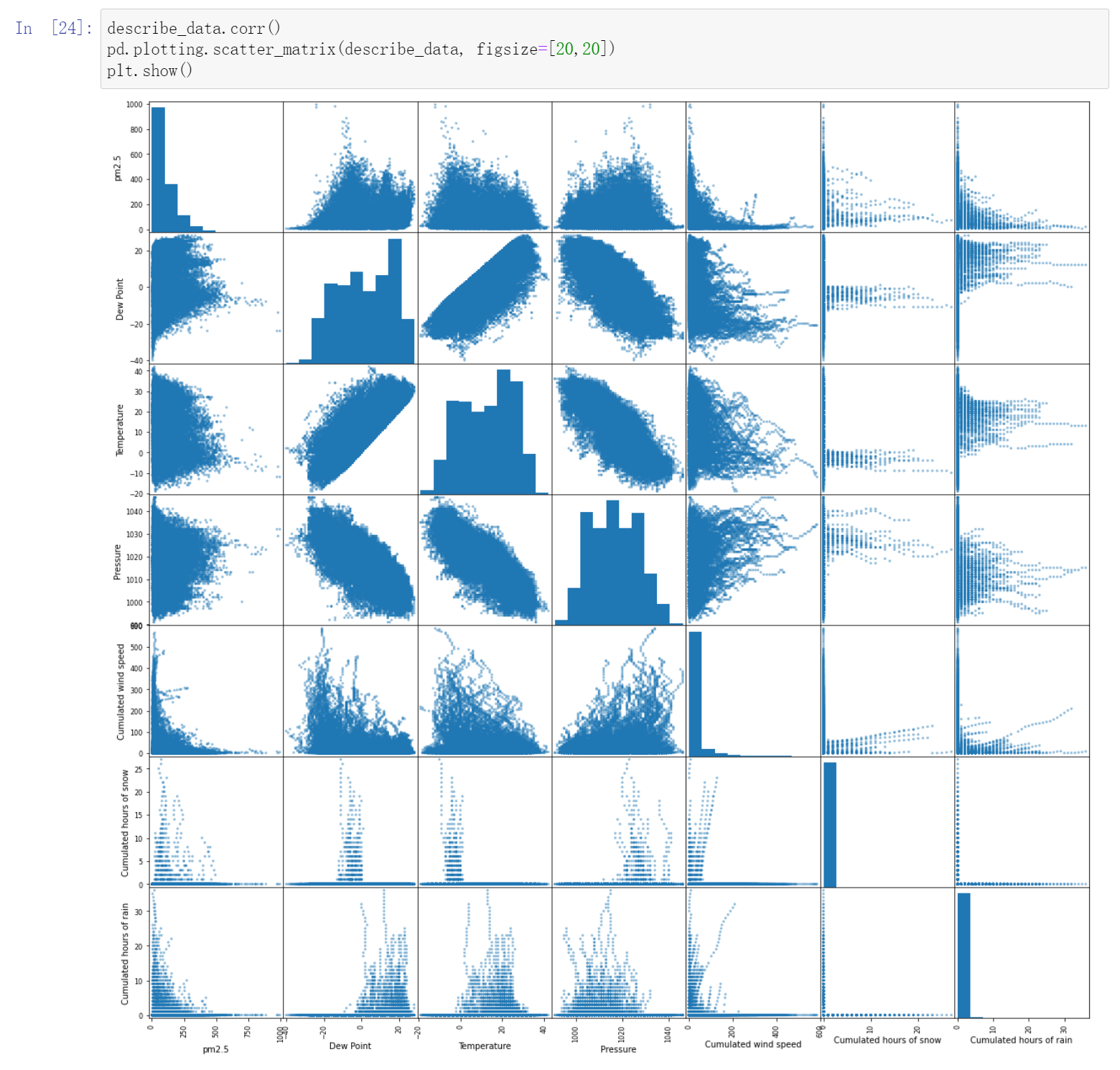
Task 3

Perform correlation analysis and provide correlation matrixes and plots. Discuss your findings in terms of

1. correlation between the independent variables (multi-Collinearity)





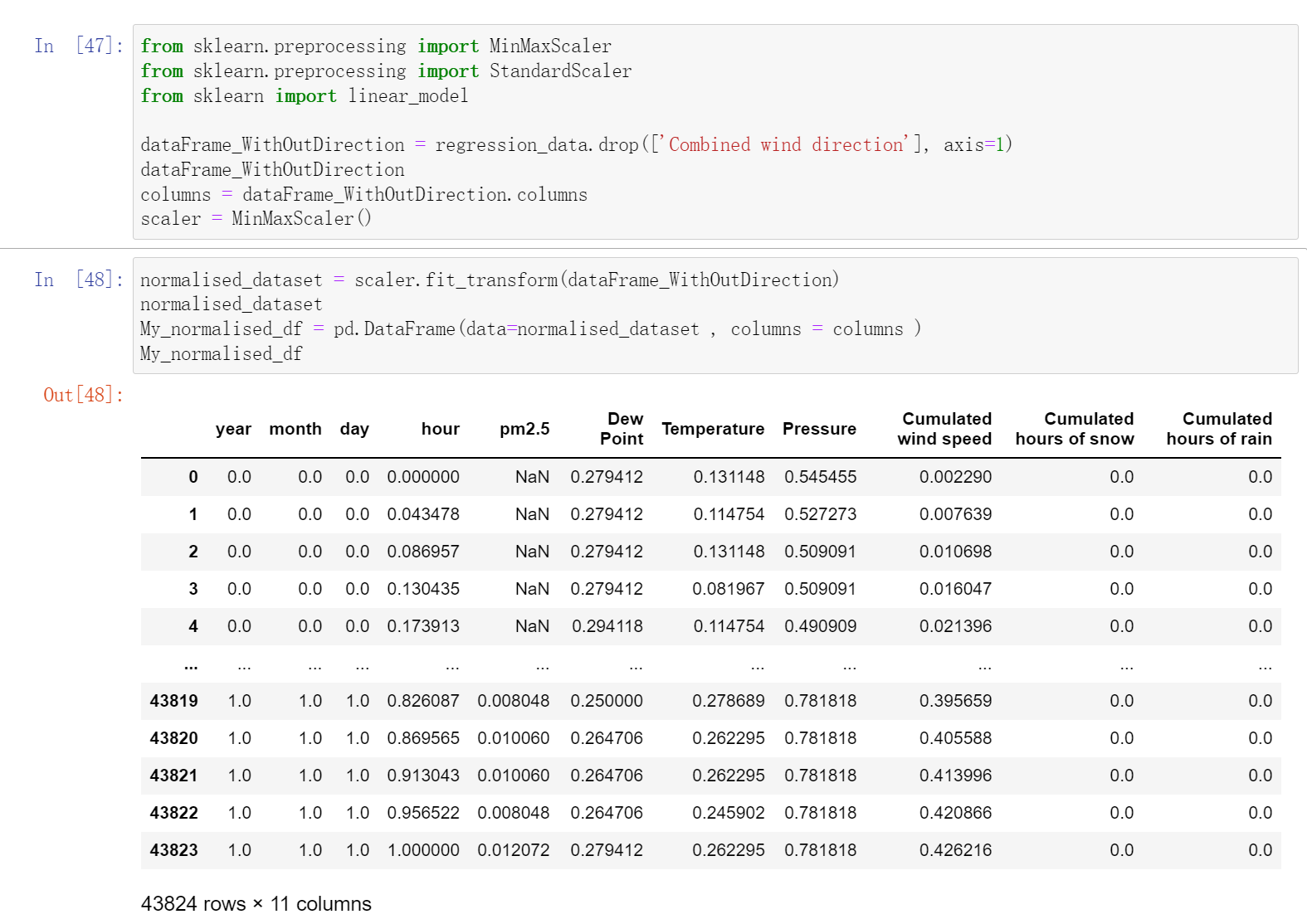


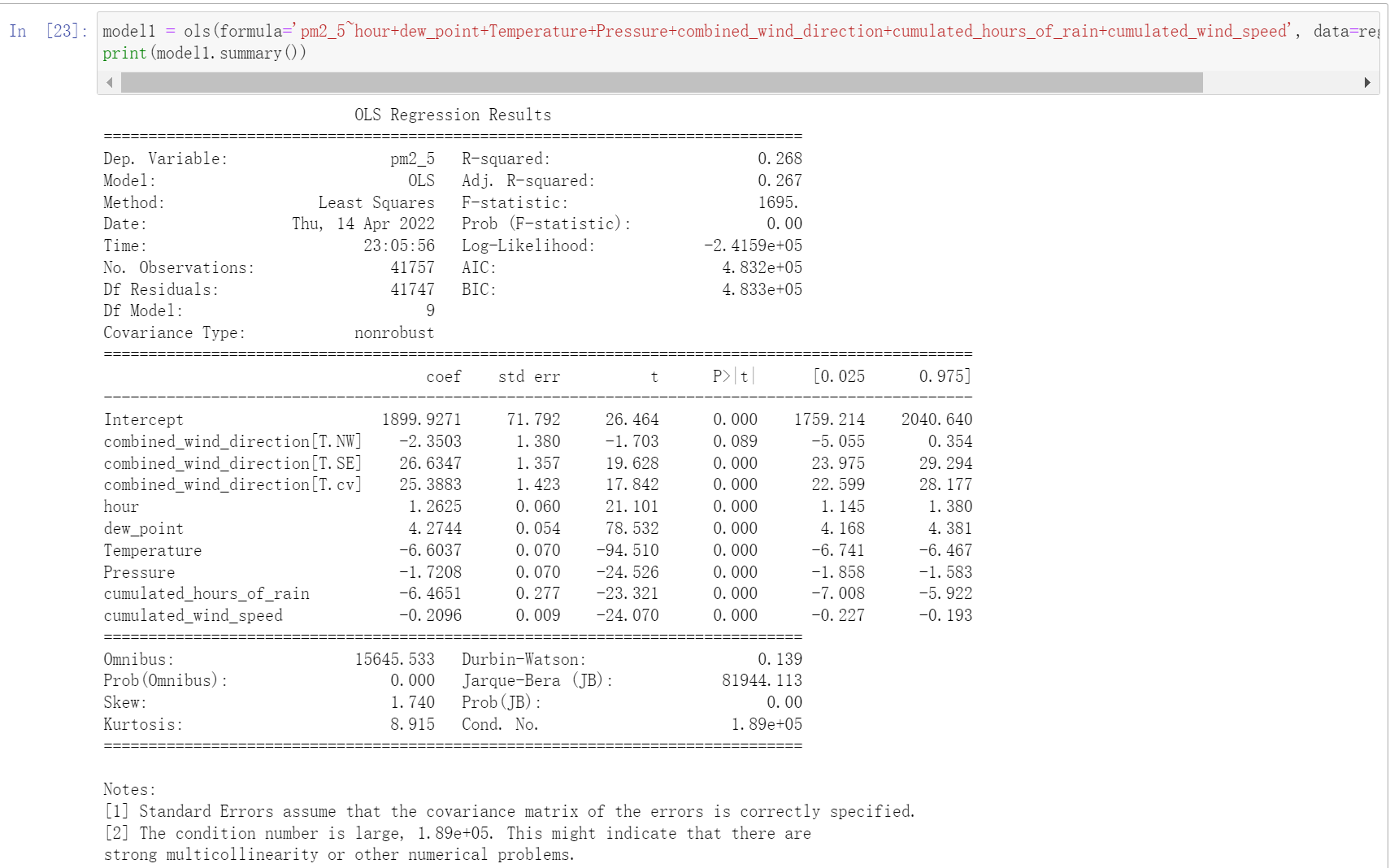
1. between independent variables and dependent variables. Is there any sign of violation of assumption(s)? If yes, explain your approach to handling it before moving to the next task?

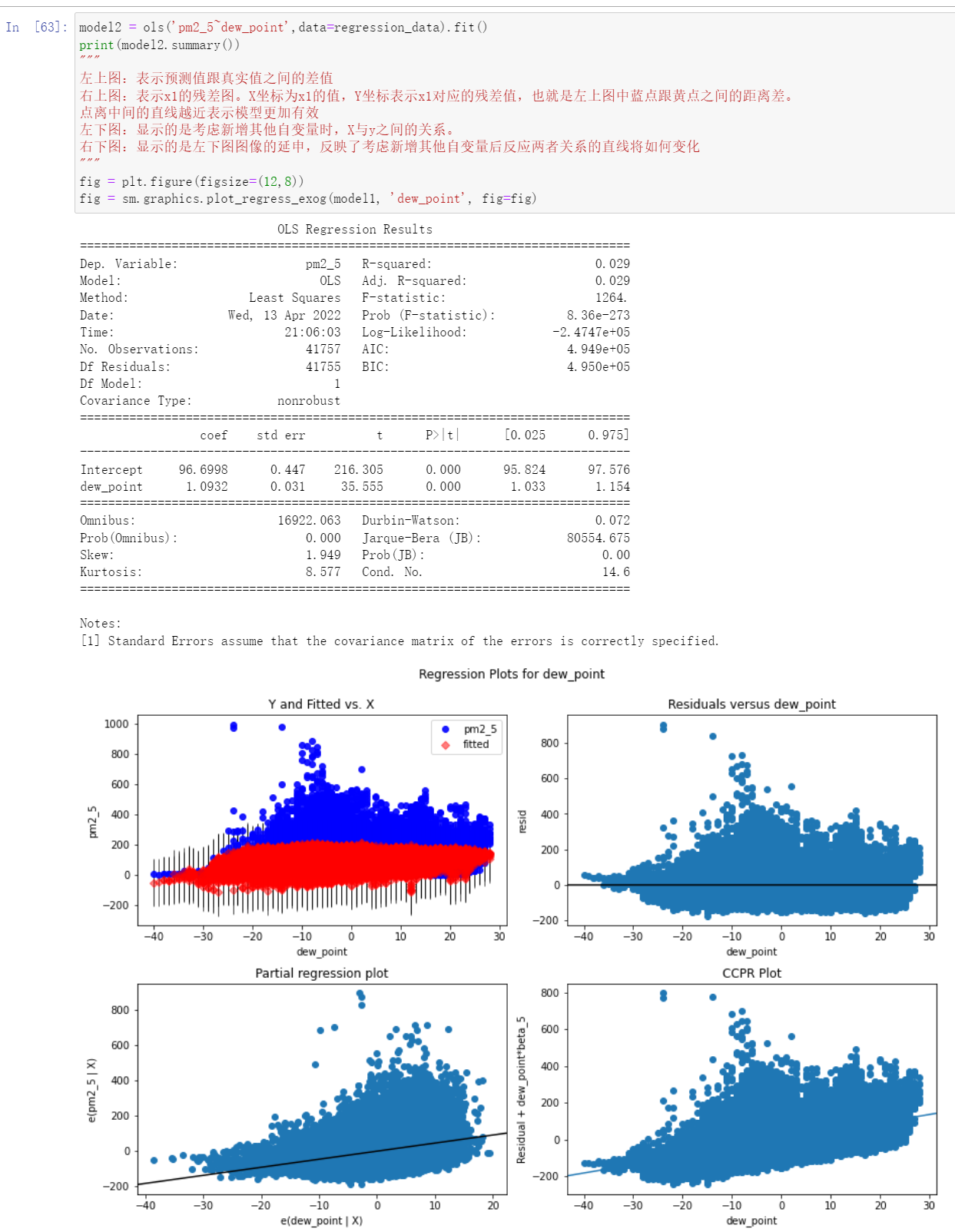
**Based on the obtained image analysis, the concentration of pm2.5 would have a low effect between different features such as: dew point, Temperature, pressure, different time period rainfall time, wind speed and wind direction. But these features of possibility of impact is relatively small. Because the correlation coefficient number of dew point and wind speed only have 0.17 and -0.25. Although there is a certain degree of correlation, however, the degree of correlation is not strong enough, the remaining correlation coefficient number are both less than 0.1.**

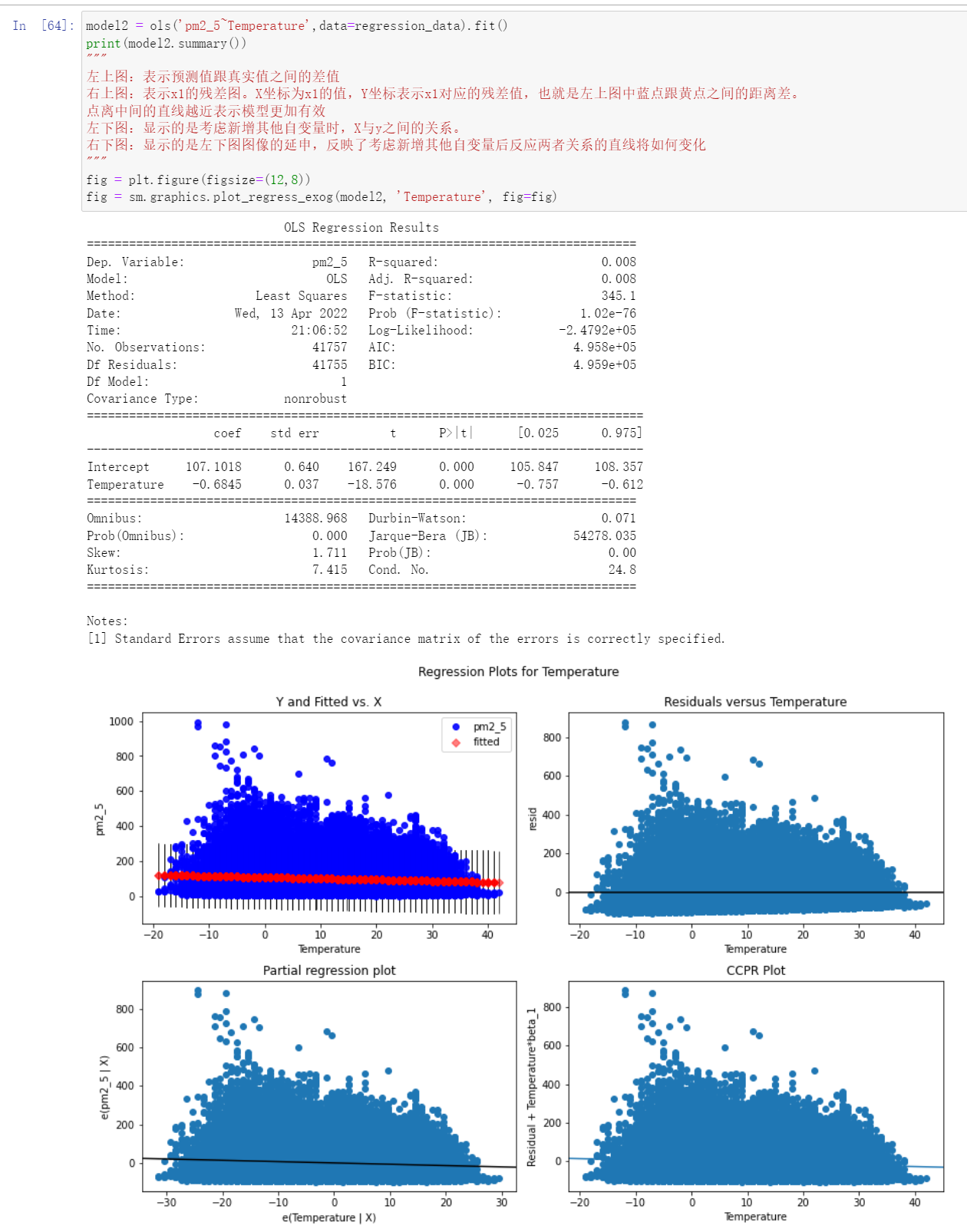
Task 4

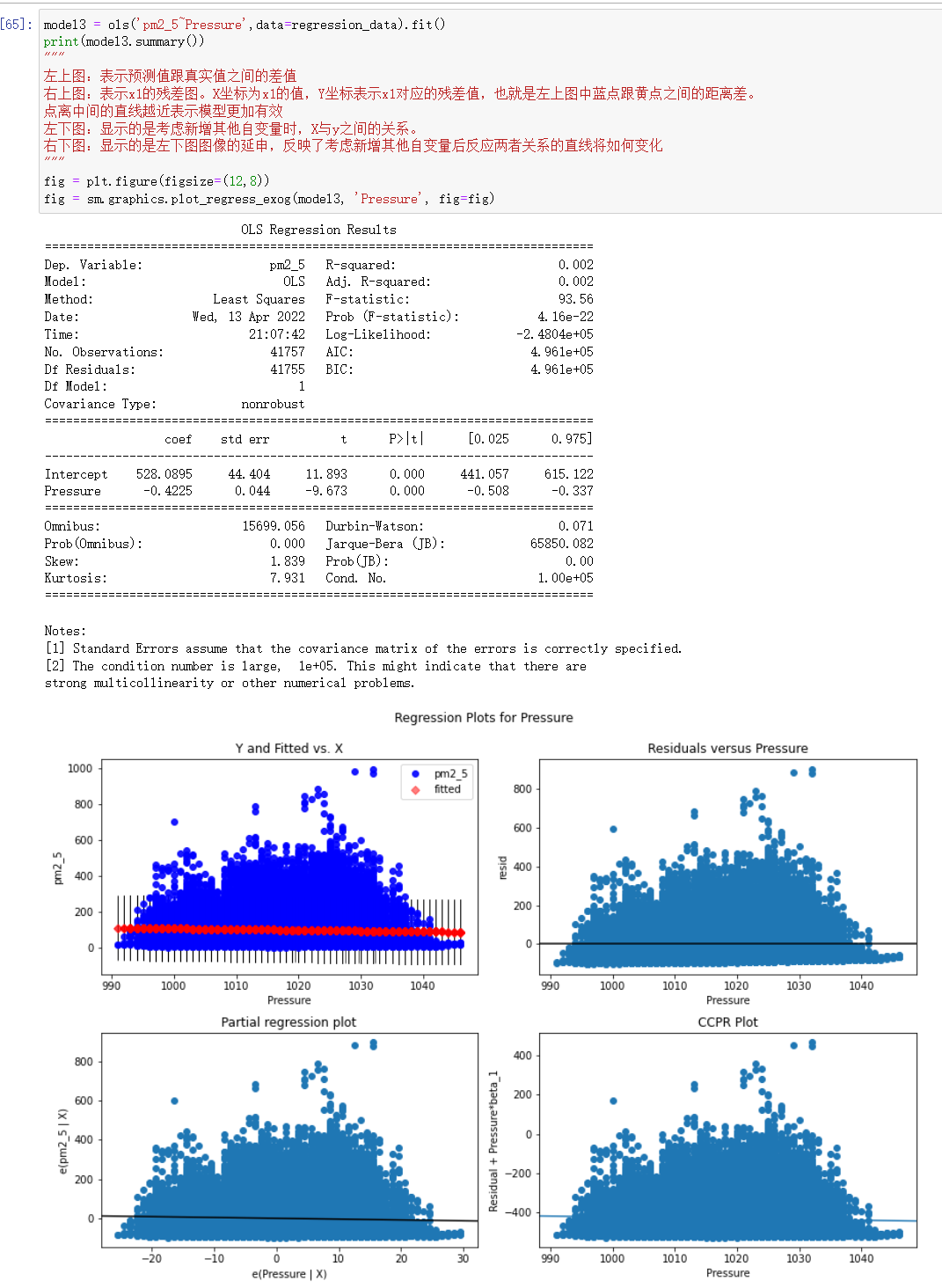
Provide the results including regression results, statistical significance metrics, and coefficients tables from this model.

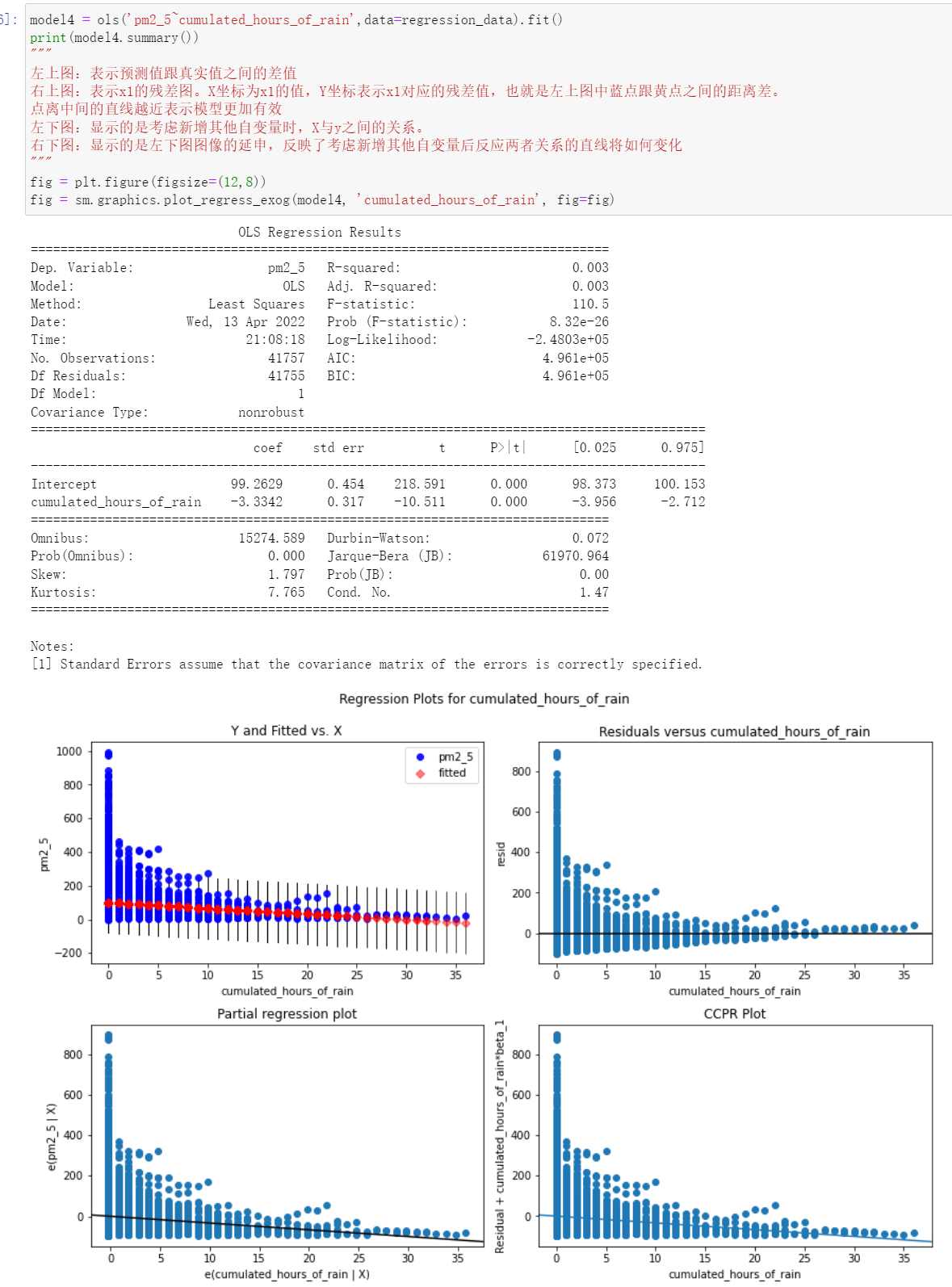


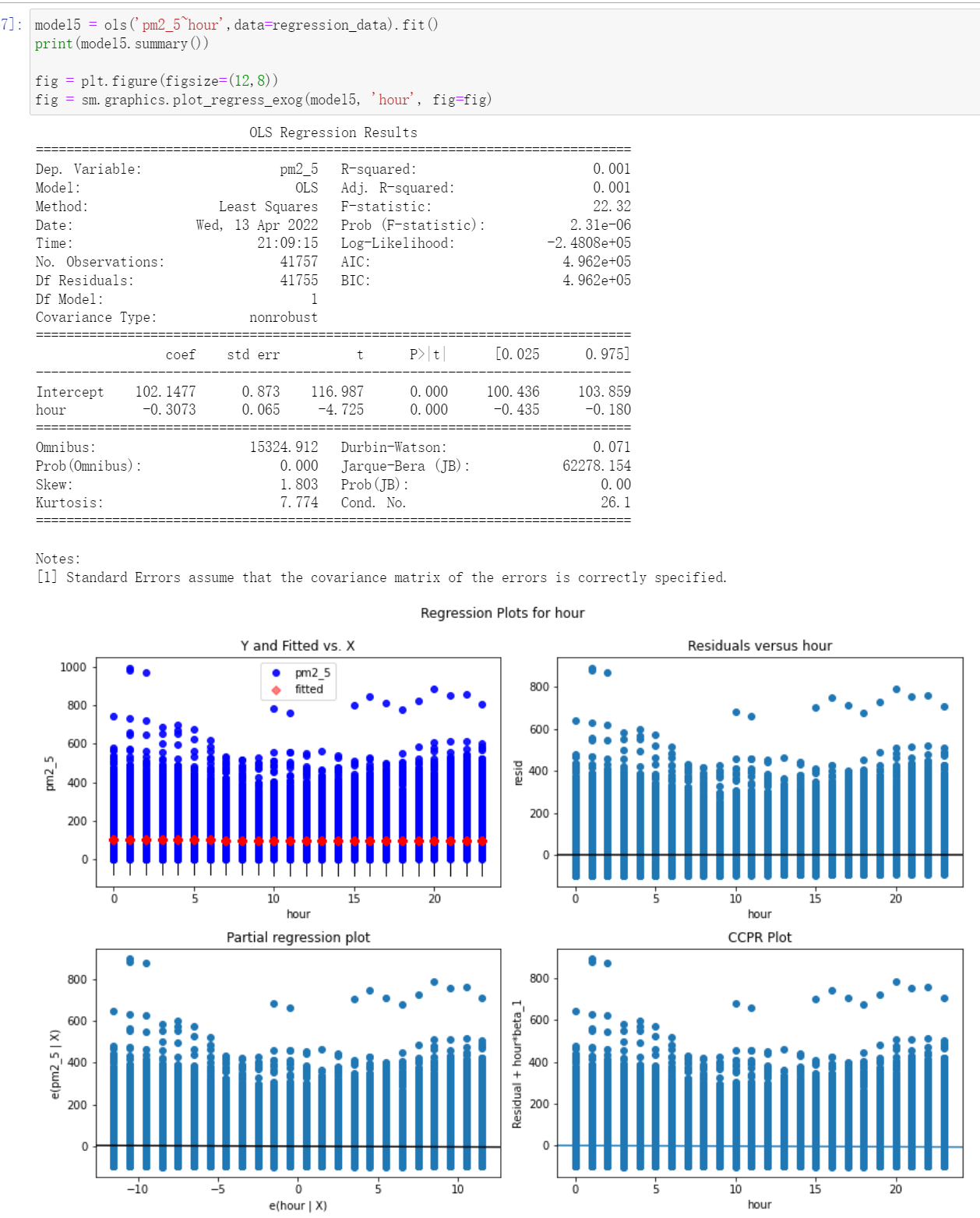


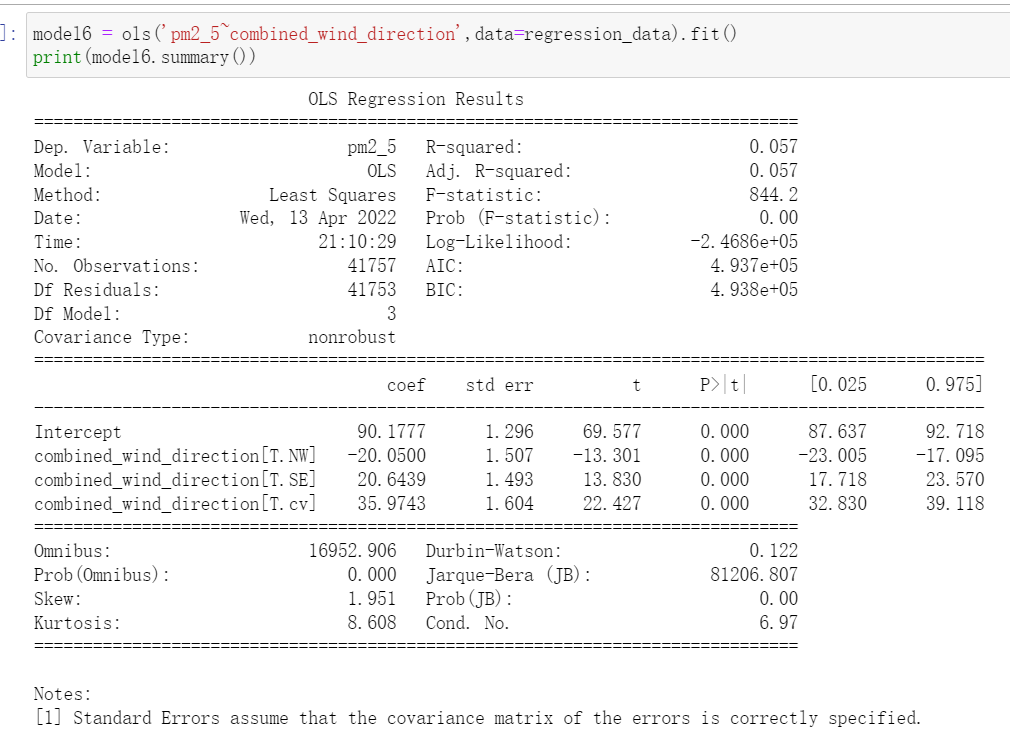


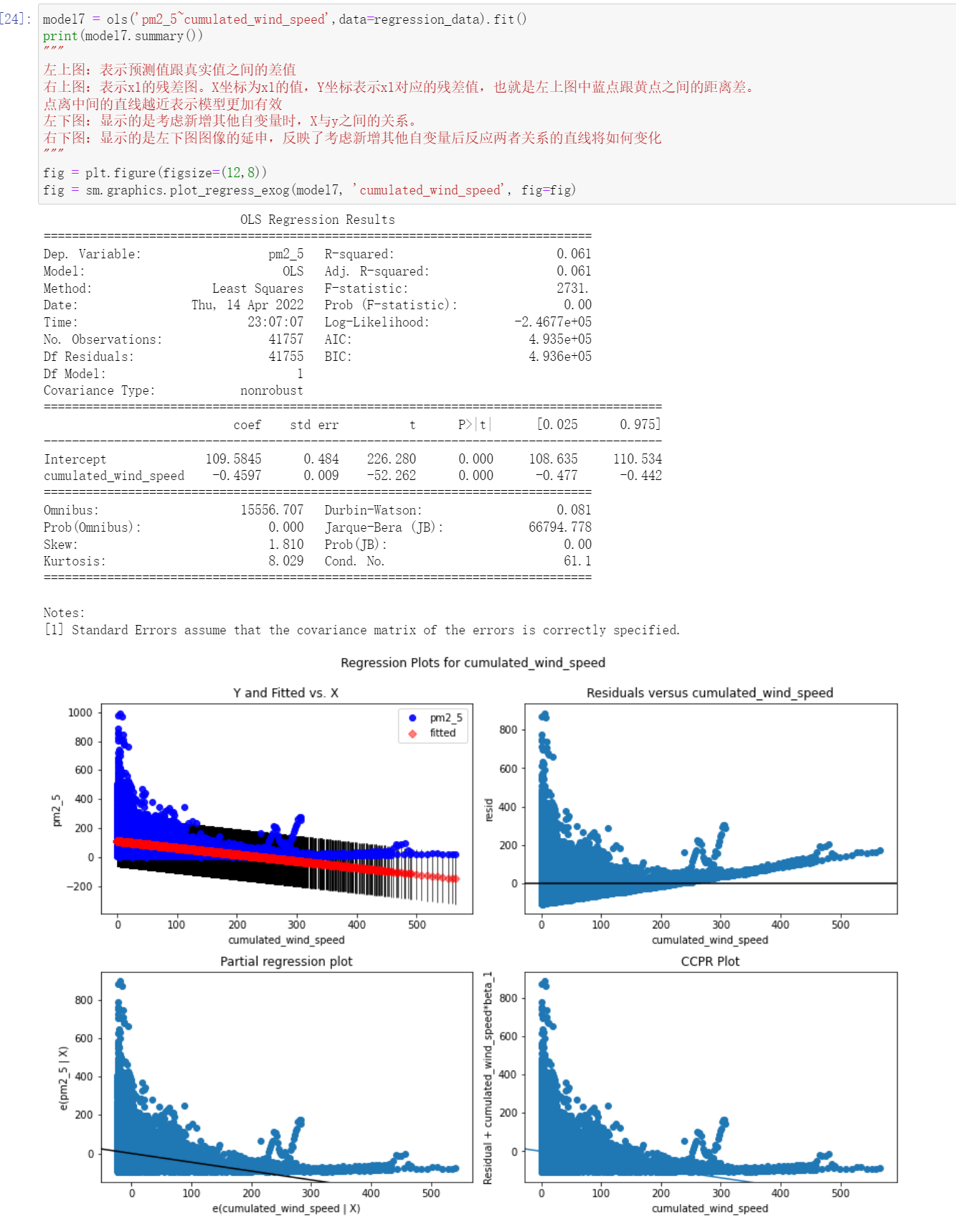












Task 5

In this section, you are required to describe and analyses the results of your regression model.

1. Are all the assumptions made in Task 1 satisfied? Provide evidence to support your answer.

**Based on the regression plots and the results, the assumption is not tenable. It means these features (Dew Point, Temperature, wind direction, wind speed, pressure, raining time and a select period time (hours)) might have no any effect and relate to the pm2.5. The hypothesis of the experiment in task 1 is not tenable. The reason is the OLS regression results has shown, the R-squared is 0.268 and Adj.R-squared only have 0.267, it is a very low number because if the regression line need a high degree of fitting to the observed values, the R-squared need to closer to 1, it means this fitting degree of the regression line to the observed value is quite poor. If it is significant, it indicates that at least one of all independent variables will have an impact on the dependent variable. And Prob (F-statistic) is used to determine whether at least one of the independent variables has an impact on the dependent variable. Except the first regression result, the rest of plot are all shown the at least one factor could impact to the number of pm2.5, but the effect is really low. Besides, the value of Durbin-Watson only has 0.139, very low number, because the Durbin-Watson statistic is about 2, which indicates that the residual is subject to normal distribution. If it deviates too far from 2 such as this test number 0.139, the interpretation ability of the constructed model will be greatly affected, the explanatory power of the model is very low. Moreover, the Jarque-Bera number is also too high: 81944.113, because the Jarque-Bera test results are always non negative. If the result is much greater than 0, it means that the data does not have normal distribution. And condition number is too large, 1.89e+05. This might indicate that there are strong multicollinearity or other numerical problems. And the number of Skewness and Kurtosis also shown these data are not subject to normal distribution, because the requirement of normal distribution is the number of Skewness and Kurtosis need to closer to 0, however, in these results, the number are Skew: 1.740, Kurtosis: 8.915, both are far from 0. The reason might be these data set only collect the weather information and the time, it is not include the other pollutants that contribute to high concentration of pm2.5, if there are some features or data, the pm2.5 could be predict better.**

1. Compare your findings in Task 3’s with the coefficients table results generated in Task 4 and discuss your findings.

**According to the compare of coefficient number between the task 3 and 4. It has some result could be found. The positive and negative values of the coefficient number are basically the similar as task3, however, these features are not quiet effect to the number of pm2.5.**

Code

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

PRSA\_data = pd.read\_csv("PRSA\_data.csv")

#display data

PRSA\_data

describe\_data = PRSA\_data.drop(['year','month','day','hour','No'], axis=1)

#drop year month day hour to display the main features

describe\_data.describe()

plt.figure(figsize = (20,10))

plt.title("different rainning time with concentration of pm2.5 ")

plt\_rain = sns.boxplot(data=PRSA\_data, x="Cumulated hours of rain", y="pm2.5")

print(plt\_rain)

plt.figure(figsize = (20,10))

plt.title("different rainning time with concentration of pm2.5 ")

plt\_rain = sns.scatterplot(data=PRSA\_data, x="Cumulated hours of rain", y="pm2.5")

print(plt\_rain)

plt.figure(figsize=(20,10))

plt.title("different rainfall time with the pm2.5")

plt\_hour = sns.lineplot(data=PRSA\_data, x='Cumulated hours of rain',y='pm2.5')

print(plt\_hour)

plt.figure(figsize = (20,10))

plt.title("pm2.5 for a select time period (03/2010-12/2014)")

plt\_months = sns.boxplot(data=PRSA\_data, x="month", y="pm2.5")

print(plt\_months)

plt.figure(figsize=(20,10))

plt.title("different time in a day with the pm2.5")

plt\_hour = sns.lineplot(data=PRSA\_data, x='hour',y='pm2.5')

print(plt\_hour)

plt.figure(figsize=(20,10))

plt.title("different number of Dew Point with the pm2.5")

plt\_DewPoint = sns.lineplot(data=PRSA\_data, x='Dew Point',y='pm2.5')

print(plt\_DewPoint)

plt.figure(figsize=(20,10))

plt.title("different number of Dew Point with the pm2.5")

plt\_DewPoint = sns.scatterplot(data=PRSA\_data, x='Dew Point',y='pm2.5')

print(plt\_DewPoint)

plt.figure(figsize=(20,10))

plt.title("different Temperature with the pm2.5")

plt\_Temperature = sns.lineplot(data=PRSA\_data, x='Temperature',y='pm2.5')

print(plt\_Temperature)

plt.figure(figsize=(20,10))

plt.title("different Temperature with the pm2.5")

plt\_Temperature = sns.scatterplot(data=PRSA\_data, x='Temperature',y='pm2.5')

print(plt\_Temperature)

plt.figure(figsize=(20,10))

plt.title("different Pressure with the pm2.5")

plt\_Pressure = sns.lineplot(data=PRSA\_data, x='Pressure',y='pm2.5')

print(plt\_Pressure)

plt.figure(figsize=(20,10))

plt.title("different Pressure with the pm2.5")

plt\_Pressure = sns.scatterplot(data=PRSA\_data, x='Pressure',y='pm2.5')

print(plt\_Pressure)

plt.figure(figsize=(20,10))

plt.title("Different wind direction with the pm2.5")

plt\_Bar = sns.lineplot(data=PRSA\_data, x='pm2.5', y='Combined wind direction')

plt.figure(figsize=(20,10))

plt.title("Different wind speed with the pm2.5")

plt\_line = sns.lineplot(data=PRSA\_data, x='pm2.5', y='Cumulated wind speed')

print(plt\_line)

plt.figure(figsize=(20,10))

plt.title("Different wind speed with the pm2.5")

plt\_box = sns.scatterplot(data=PRSA\_data, x='pm2.5', y='Cumulated wind speed')

print(plt\_box)

plt.figure(figsize=(20,10))

plt\_dis = sns.histplot(describe\_data['pm2.5'])

print(plt\_dis)

describe\_data.corr()

pd.plotting.scatter\_matrix(describe\_data, figsize=[20,20])

plt.show()

plt.figure(figsize=(15,10))

plt\_heat=sns.heatmap(describe\_data.corr(method='pearson', min\_periods=1), annot=True)

plt.title("Correlation of pm2.5 between different factors" )

print(plt\_heat)

regression\_data = PRSA\_data.drop(['No'], axis=1)

regression\_data

from sklearn.preprocessing import MinMaxScaler

from sklearn.preprocessing import StandardScaler

from sklearn import linear\_model

dataFrame\_WithOutDirection = regression\_data.drop(['Combined wind direction'], axis=1)

dataFrame\_WithOutDirection

columns = dataFrame\_WithOutDirection.columns

scaler = MinMaxScaler()

normalised\_dataset = scaler.fit\_transform(dataFrame\_WithOutDirection)

normalised\_dataset

My\_normalised\_df = pd.DataFrame(data=normalised\_dataset , columns = columns )

My\_normalised\_df

import statsmodels.api as sm

from statsmodels.formula.api import ols

#rename the var

regression\_data.rename(columns={'pm2.5':'pm2\_5', 'Dew Point':'dew\_point', 'Cumulated hours of rain':'cumulated\_hours\_of\_rain',

                                'Cumulated wind speed':'cumulated\_wind\_speed','Cumulated hours of snow':'cumulated\_hours\_of\_snow',

                               'Combined wind direction':'combined\_wind\_direction'}, inplace=True)

regression\_data

model1 = ols(formula='pm2\_5~hour+dew\_point+Temperature+Pressure+combined\_wind\_direction+cumulated\_hours\_of\_rain+cumulated\_wind\_speed', data=regression\_data).fit()

print(model1.summary())

model2 = ols('pm2\_5~dew\_point',data=regression\_data).fit()

print(model2.summary())

"""

左上图：表示预测值跟真实值之间的差值

右上图：表示x1的残差图。X坐标为x1的值，Y坐标表示x1对应的残差值，也就是左上图中蓝点跟黄点之间的距离差。

点离中间的直线越近表示模型更加有效

左下图：显示的是考虑新增其他自变量时，X与y之间的关系。

右下图：显示的是左下图图像的延申，反映了考虑新增其他自变量后反应两者关系的直线将如何变化

"""

fig = plt.figure(figsize=(12,8))

fig = sm.graphics.plot\_regress\_exog(model1, 'dew\_point', fig=fig)

model2 = ols('pm2\_5~Temperature',data=regression\_data).fit()

print(model2.summary())

"""

左上图：表示预测值跟真实值之间的差值

右上图：表示x1的残差图。X坐标为x1的值，Y坐标表示x1对应的残差值，也就是左上图中蓝点跟黄点之间的距离差。

点离中间的直线越近表示模型更加有效

左下图：显示的是考虑新增其他自变量时，X与y之间的关系。

右下图：显示的是左下图图像的延申，反映了考虑新增其他自变量后反应两者关系的直线将如何变化

"""

fig = plt.figure(figsize=(12,8))

fig = sm.graphics.plot\_regress\_exog(model2, 'Temperature', fig=fig)

model3 = ols('pm2\_5~Pressure',data=regression\_data).fit()

print(model3.summary())

"""

左上图：表示预测值跟真实值之间的差值

右上图：表示x1的残差图。X坐标为x1的值，Y坐标表示x1对应的残差值，也就是左上图中蓝点跟黄点之间的距离差。

点离中间的直线越近表示模型更加有效

左下图：显示的是考虑新增其他自变量时，X与y之间的关系。

右下图：显示的是左下图图像的延申，反映了考虑新增其他自变量后反应两者关系的直线将如何变化

"""

fig = plt.figure(figsize=(12,8))

fig = sm.graphics.plot\_regress\_exog(model3, 'Pressure', fig=fig)

model4 = ols('pm2\_5~cumulated\_hours\_of\_rain',data=regression\_data).fit()

print(model4.summary())

"""

左上图：表示预测值跟真实值之间的差值

右上图：表示x1的残差图。X坐标为x1的值，Y坐标表示x1对应的残差值，也就是左上图中蓝点跟黄点之间的距离差。

点离中间的直线越近表示模型更加有效

左下图：显示的是考虑新增其他自变量时，X与y之间的关系。

右下图：显示的是左下图图像的延申，反映了考虑新增其他自变量后反应两者关系的直线将如何变化

"""

fig = plt.figure(figsize=(12,8))

fig = sm.graphics.plot\_regress\_exog(model4, 'cumulated\_hours\_of\_rain', fig=fig)

model5 = ols('pm2\_5~hour',data=regression\_data).fit()

print(model5.summary())

fig = plt.figure(figsize=(12,8))

fig = sm.graphics.plot\_regress\_exog(model5, 'hour', fig=fig)

model6 = ols('pm2\_5~combined\_wind\_direction',data=regression\_data).fit()

print(model6.summary())

model7 = ols('pm2\_5~cumulated\_wind\_speed',data=regression\_data).fit()

print(model7.summary())

"""

左上图：表示预测值跟真实值之间的差值

右上图：表示x1的残差图。X坐标为x1的值，Y坐标表示x1对应的残差值，也就是左上图中蓝点跟黄点之间的距离差。

点离中间的直线越近表示模型更加有效

左下图：显示的是考虑新增其他自变量时，X与y之间的关系。

右下图：显示的是左下图图像的延申，反映了考虑新增其他自变量后反应两者关系的直线将如何变化

"""

fig = plt.figure(figsize=(12,8))

fig = sm.graphics.plot\_regress\_exog(model7, 'cumulated\_wind\_speed', fig=fig)