Springboard Data Science Career Track
Capstone Project 1: Milestone Report

Using Machine Learning to Forecast Air Quality in Beijing

Kevin Limkrailassiri

1 Introduction

Poor air quality in major Chinese cities such as Beijing is a well-known concern that has been drawing significant attention from news outlets, social media, and the Chinese government. It has been linked to several adverse health effects including heart attacks, asthma attacks and bronchitis, with more acute effects observed in the elderly, children, and those with existing health conditions. Poor air quality is a problem affecting the quality of life of every person living in major cities in China.

In an effort to monitor air quality, the US Embassy established a measurement center in Beijing to detect the hourly concentration of atmospheric particulate matter with diameter less than 2.5 micrometers, known as PM2.5. Later on, the Chinese government followed suit by setting up multiple measurement centers and began recording PM2.5 data in 2013.

The PM2.5 measurements reported by these measurement centers provide citizens a valuable gauge of the instantaneous air quality. However, citizens may benefit even more if they were provided with a forecast that can predict the air quality with reasonable accuracy several days into the future. In the same way that a weather forecast helps citizens to arrange their weekly plans based on the predicted weather, a forecast of the air quality can also provide citizens a means of responding to days when the air quality is predicted to be poor. Moreover, this study can provide an understanding of weather trends corresponding to higher PM2.5 level, which can help the government proactively curtail contributions from pollution emitted by factories and public transportation when air quality is predicted to be poor.

In brief, the approach of this study is first to perform data wrangling to assemble all the data into one DataFrame with each row containing an observation and each column containing a parameter. Next, we will perform exploratory data analysis to observe any possible trends and relationships between sets of data. The insights from this step will help us perform supervised machine learning in order to construct a model that can forecast PM2.5 concentration based on patterns in the weather and transportation usage. A training and testing split will be used to train the machine learning model and test the accuracy of its forecasts.

2 Data Acquisition and Cleaning

The data set containing measurements of air quality in Beijing based on PM2.5 concentration are obtained from the UCI Machine Learning Repository in .csv format. The data set contains hourly measurements of air quality for the time period of January 1, 2010 to December 31, 2015 from four measurement centers in Beijing located at most 5 kilometers from one another. Among these measurement centers, `PM_Dongsi`, `PM_Dongsihuan,` and `PM_Nongzhanguan` are maintained and operated by the Beijing Municipal Environmental Monitoring Center and `PM_US Post` is maintained and operated by the US Embassy. Since `PM_US Post` contains the least number of NaNs, only the `PM_US Post` data series is employed for the purpose of this study. In addition to the air quality measurements, the data set also contains hourly measurements of temperature (Celsius), pressure (hPa), humidity (%), dew point (Celsius), wind speed (m/s and combined wind direction (NW, NE, SE, or SW), and hourly and cumulative precipitation (mm).

To evaluate the reliability of PM_US Post in providing air quality measurements, the frequency and consecutive instances of NaN are studied.

Supplementary transportation data from the National Bureau of Statistics of China are obtained from Quandl in .csv format. Railway, highway, waterway, and civil aviation transportation in units of passenger-kilometers, which is the number of passengers multiplied by the distance of transportation, are recorded monthly from January 2005 to February 2016. The few missing points in the data series for highways and waterways are resolved through linear interpolation.

The hourly air quality and weather data are formed into one dataframe called df, while the monthly transportation data set are formed into another dataframe called df_transport. Machine learning models are first trained, tested, and evaluated using the data contained in df. The machine learning models are then reevaluated after incorporating the data from df_transport to see if the additional transportation data improves or worsens the accuracy of air quality predictions. For this, df and df_transport are merged together into one dataframe.

3 Exploratory Data Analysis

We begin exploratory data analysis by visualizing the PM2.5 concentration data as a function of time and observe trends across years and by month and day of the week.

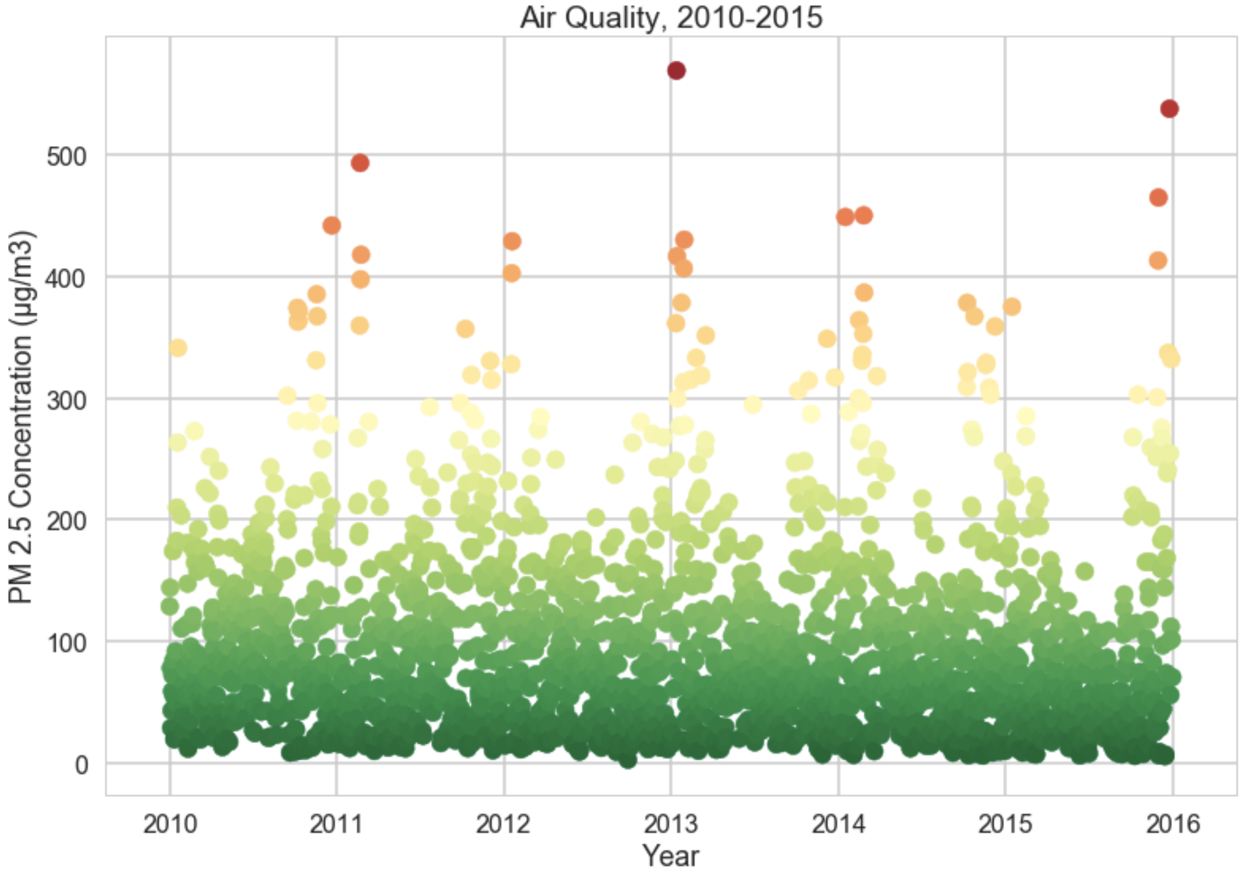

Figure 3.1: Plot of daily average PM2.5 concentration from 2010 through 2015 shows annual peaks near the start of each calendar year.

A plot of daily air quality for the period of 2010 to 2015 shows a high density of points up to PM2.5 concentration of 150 µg/m^3 along with peaks occurring annually near the start of each year. The data points composing these peaks are loosely distributed. These peaks could suggest a correlation between the weather or some human-related factor such as increased carbon emissions during cold months and the resulting air quality. Throughout the entire time period of 2010 to 2015, there is a dense concentration of data points for PM2.5 concentration up to 150 µg/m^3, which makes it difficult to visualize the overall distribution of air quality measurements. Therefore, we plot the distribution next.

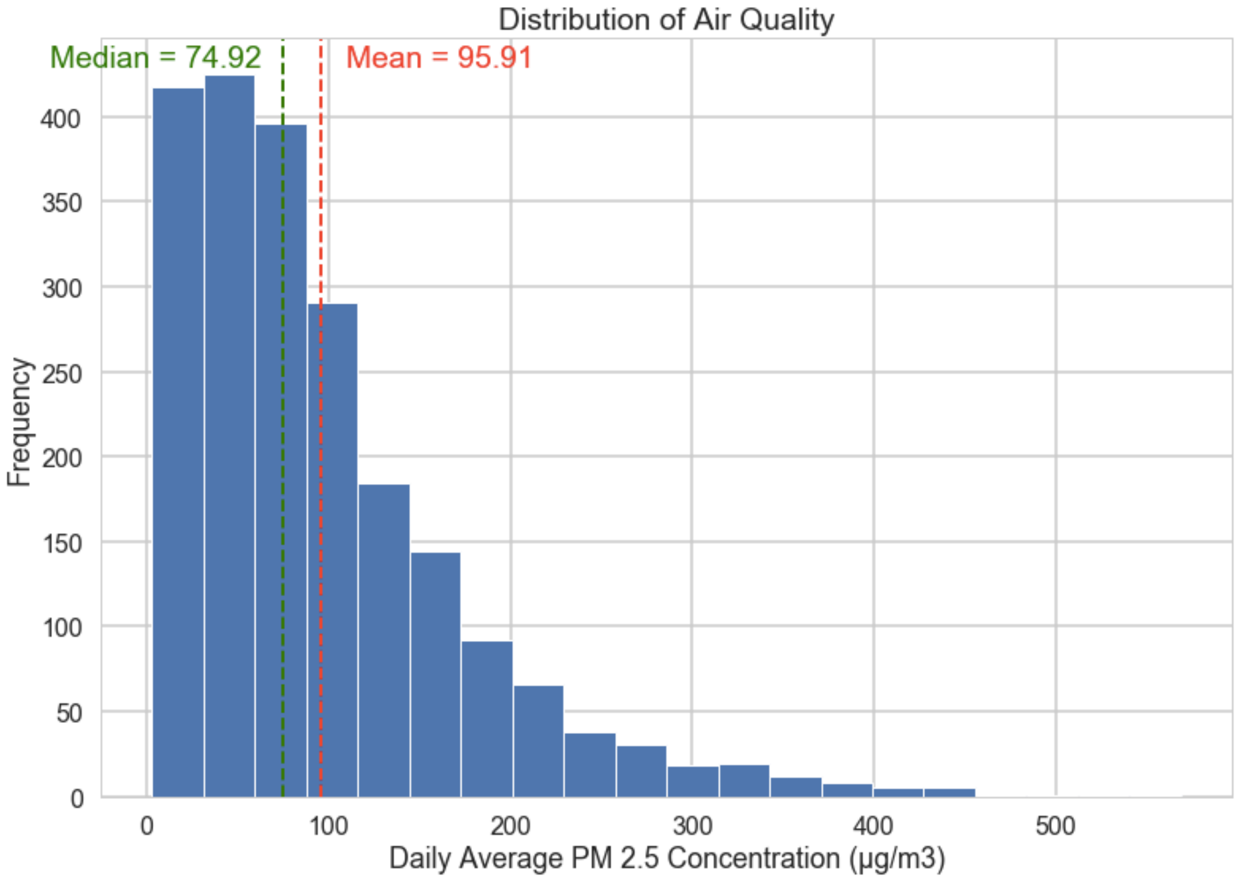

Figure 3.2: Distribution of daily average PM2.5 concentration with a mean of 95.91 µg/m^3 and median of 74.92 µg/m^3.

The distribution of daily average PM2.5 concentration is right-skewed with a peak at around 50 µg/m^3, an average of 95.91 µg/m^3, and a median of 74.92 µg/m^3. According to the interpretation of Air Quality Index (AQI) by China's Ministry of Environmental Protection, a value of 95.91 µg/m^3 is within the 'Moderate' classification, close to the 'Unhealthy for Sensitive Groups' classification which spans 101-150 µg/m^3. A table of each of the AQI classifications is provided below.

Air Quality Index (µg/m^3)	**Level of Health Concern**	**Health Implications**
0 – 50	Excellent	No health implications

51 – 100	Good	Few hypersensitive individuals should reduce outdoor exercise.
101 – 150	Lightly Polluted	Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
151 – 200	Moderately Polluted	Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
201 – 300	Heavily Polluted	Healthy people will be noticeably affected. People with breathing or heart problems will experience reduced endurance in activities. These individuals and elders should remain indoors and restrict activities.
301 – 500	Severely Polluted	Healthy people will experience reduced endurance in activities. There may be strong irritations and symptoms and may trigger other illnesses. Elders and the sick should remain indoors and avoid exercise. Healthy individuals should avoid outdoor activities.

Table 3.1: Air quality level and health implications according to Air Quality Index (AQI).

According to the above classification of air quality, the percentage of days falling under each classification is illustrated in the following pie chart.

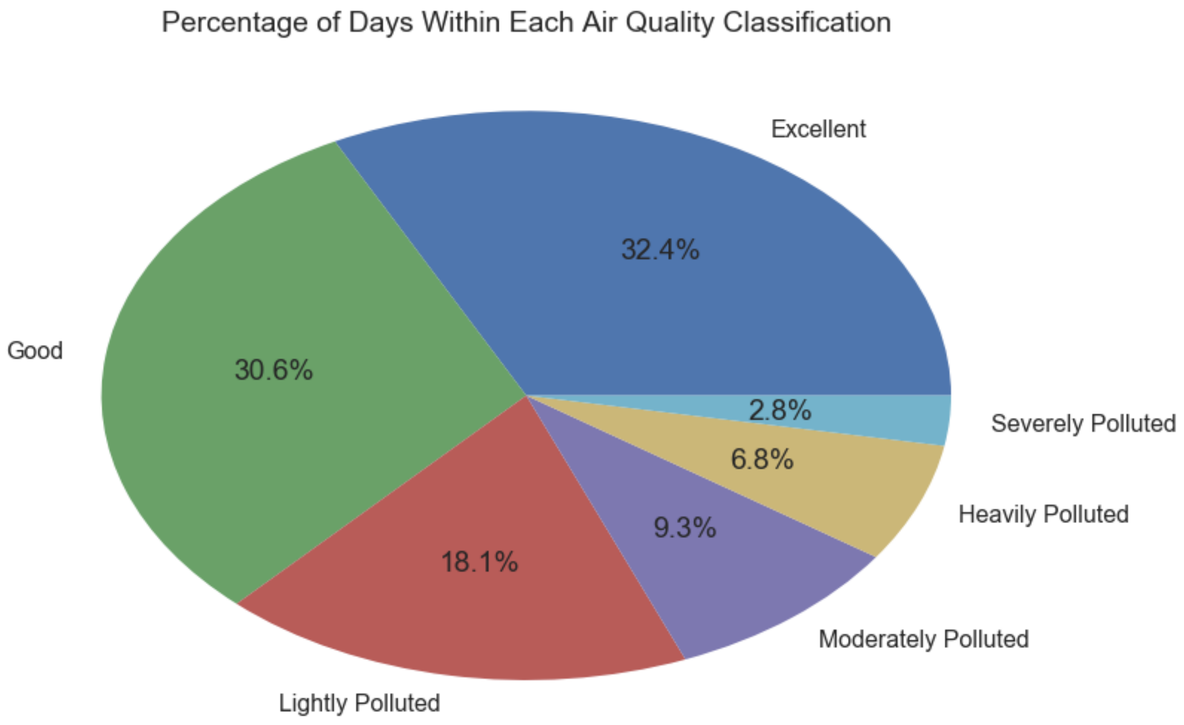

Figure 3.3: Percentage of days classified by each category of the Air Quality Index (AQI).

This pie chart shows that 37% of the days from 2010 through 2015 are characterized as 'Lightly Polluted' or worse. This figure highlights the fact that poor air quality has become a familiar experience in the everyday life of Beijing citizens, impinging on their well-being and quality of life. A study of the trends and parameters correlating with poor air quality may help Beijing citizens anticipate and prepare for days when air quality is forecasted to be poor.

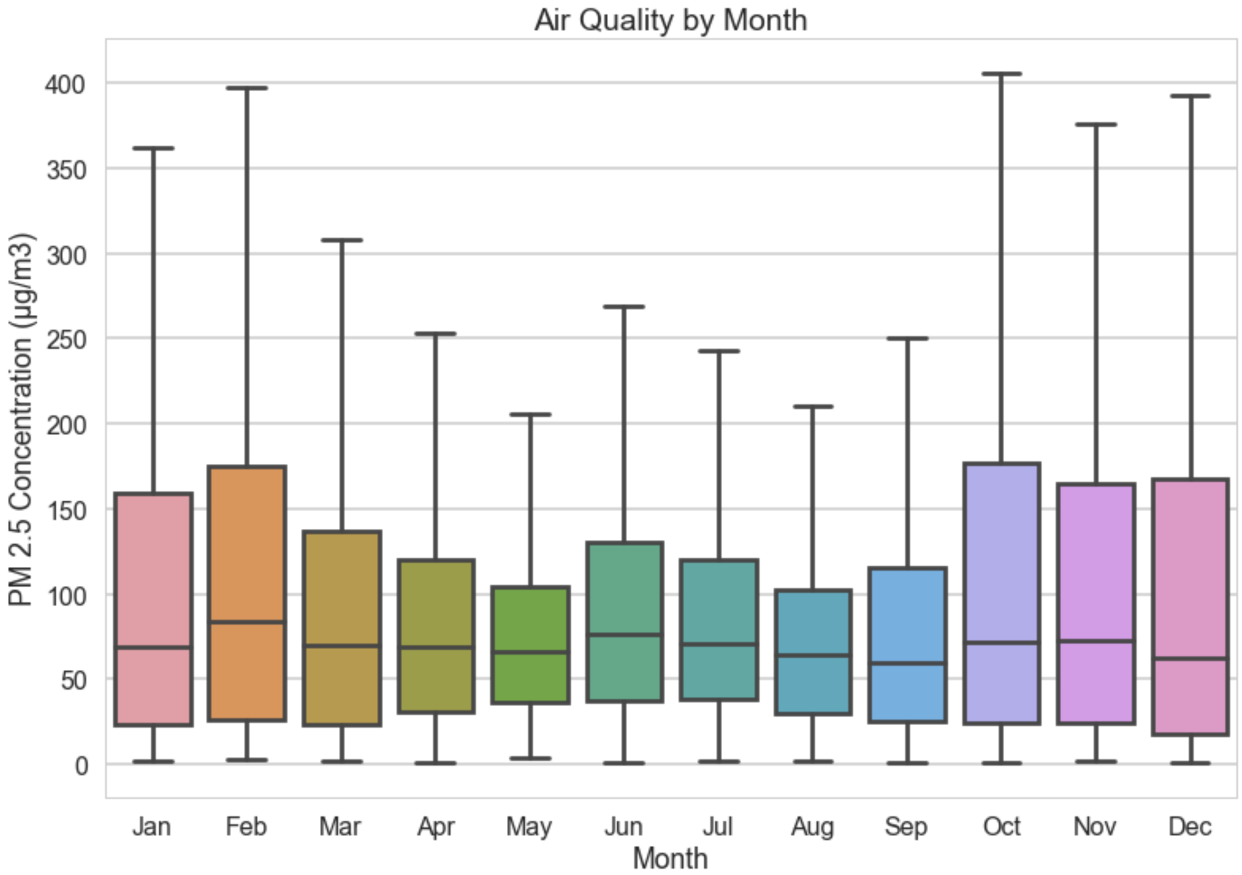

Figure 3.4: Distribution of air quality concentration by month of the year for the years 2010 through 2015.

The plot of monthly air quality from 2010 through 2015 shows two slight peaks in the median PM2.5 concentration during the months of February and June and a slight dip for the month of September. In terms of the range of PM2.5 concentration per month, the autumn and winter months show considerably more variation compared to the spring and summer months. It will be helpful to pinpoint the parameters that encourage lower PM2.5 concentration and tighter range as observed in the months of May, August, and September.

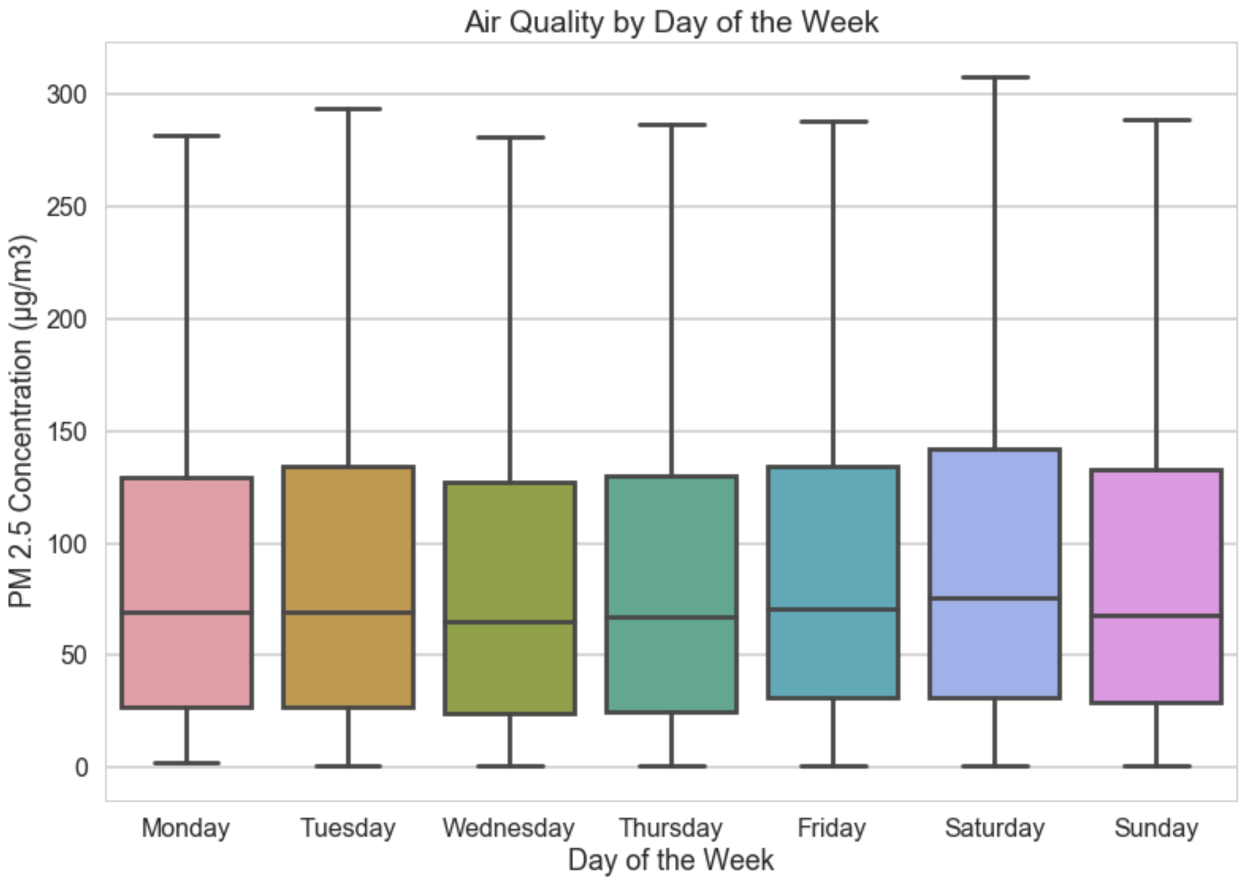

Figure 3.5: Distribution of air quality concentration by day of the week for the years 2010 through 2015.

The plot of daily PM2.5 concentration from 2010 through 2015 shows very similar median and range across the entire week. Therefore, air quality appears to be insensitive to the particular day of the week.

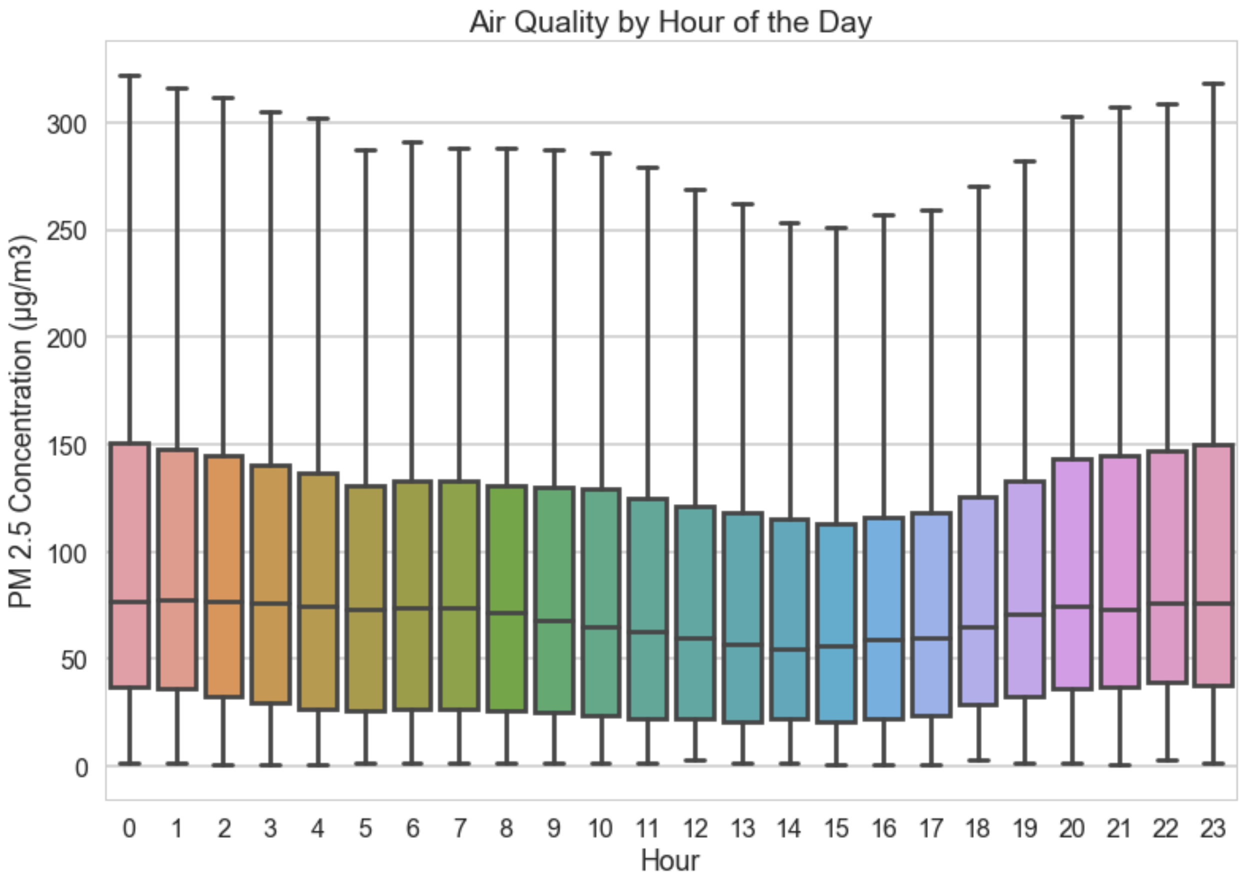

Figure 3.6: Distribution of air quality concentration by day of the week for the years 2010 through 2015.

As for the plot of hourly air quality in the span of a 24-hour day, there appears to be a smooth decrease through the early morning leading to a minimum median PM2.5 concentration at 2pm, followed by a smooth increase hitting a maximum median around midnight. It is interesting that PM2.5 concentration is minimized during working hours and maximized during non-working hours, which may suggest that pollution released by automobiles, trains, and other modes of transportation during working hours are not strong influencers of air quality compared to environmental factors. This hypothesis will be revisited later through the study of several machine learning models.

Now, we explore PM2.5 concentration as a function of several weather-related parameters.

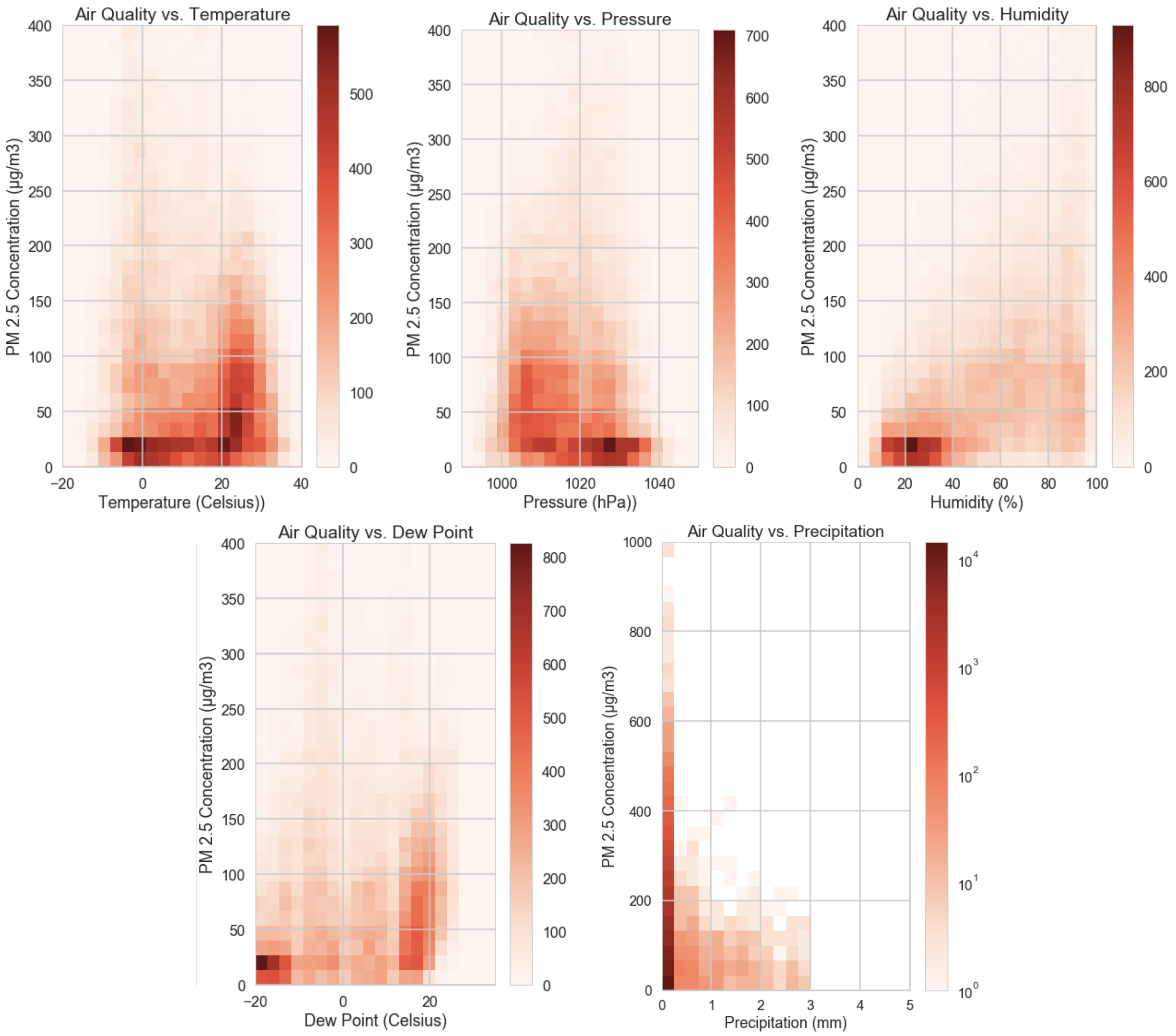

Figure 3.7: Heat maps correlating PM2.5 concentration to temperature, pressure, humidity, dew point, and precipitation.

The heat maps of Figure 3.7 show the correlation between PM2.5 concentration and weather parameters temperature, pressure, humidity, dew point, and precipitation. Darker red spots indicate that a specific level of PM2.5 concentration correlates frequently with a particular value of the weather-related parameter. The heat map of PM2.5 concentration correlated to temperature appears to show a weak positive correlation with two highly concentrated centers at PM2.5 concentration near 20 μg/m^3 with one centered around -5 °C and the other around 20 °C. It appears that the darker spots around 20 °C are distributed across a wider range of PM2.5 concentrations, which matches our observation from Figure 3.4. It would be interesting to see if there is another parameter in effect that causes the PM2.5 concentration to vary more widely at higher temperatures. In contrast, the heat map correlating PM2.5 concentration to pressure shows the opposite effect with the PM2.5 concentration shrinking as pressure increases. As for humidity, dark red spots are concentrated in the humidity range of 10% to 35%, corresponding to PM2.5 concentration up to around 25 μg/m^3. Dew point shares a similar correlation to PM2.5 concentration as observed with temperature, while the heat map for precipitation shows the majority of darker spots in the first column of bins, close to 0 mm of precipitation. This suggests that not only is there very little precipitation on most days, but this is correlated with lower values of PM2.5 concentration.

The last parameter we will study is wind speed and direction. Distributions of wind speed for winds coming from the northwest, northeast, southeast, and southwest directions are plotted below.

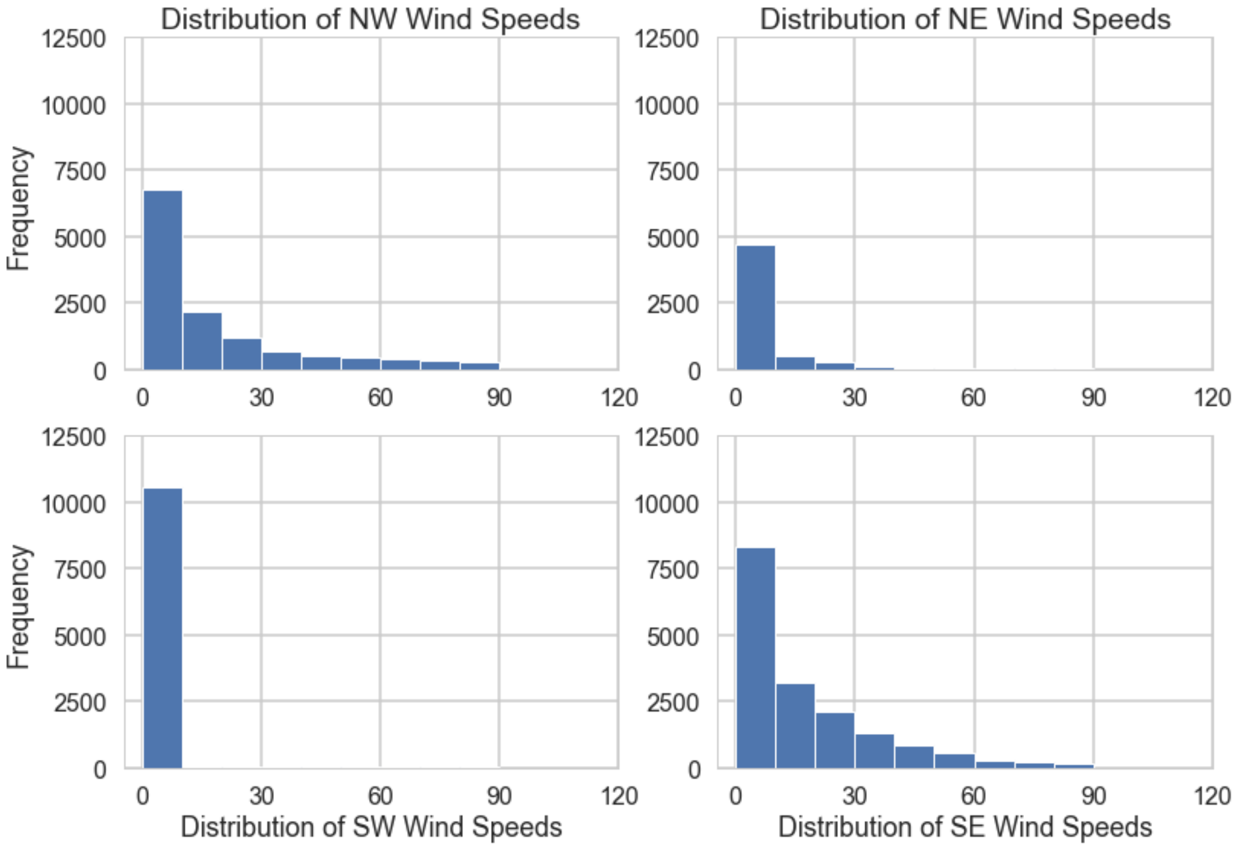

Figure 3.8: Distribution of wind speed for winds coming from the northwest (*upper left*), northeast (*upper right*), southeast (*bottom right*), and southwest (*bottom left*) directions.

All four plots show strongly right-skewed distributions with winds coming from the northwest and southeast directions showing a significantly broader distribution of wind speeds. If PM2.5 concentration is correlated to wind direction, we would expect to see a distribution of PM2.5 concentration levels linked to winds coming primarily from the northwest and southeast directions. We plot this relationship next in the form of heat maps.

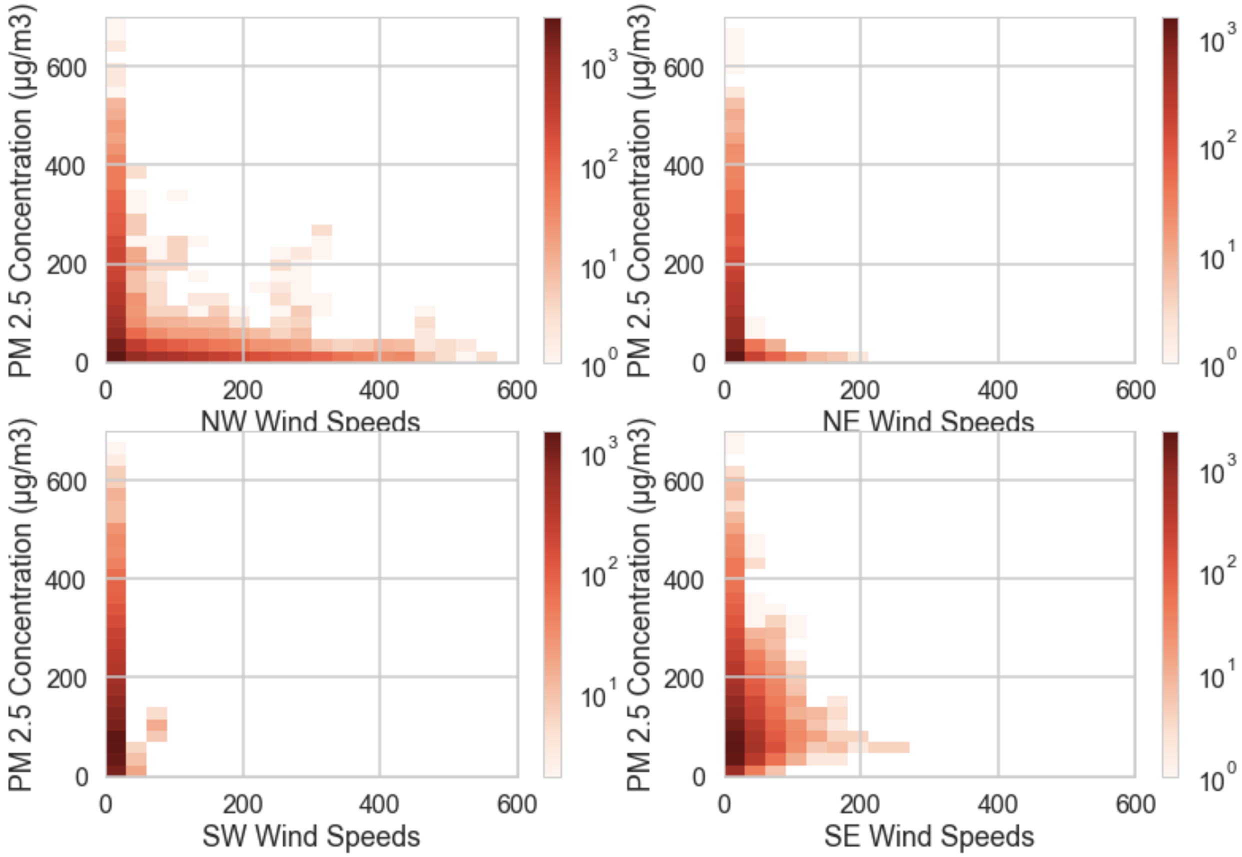

Figure 3.9: Heat maps showing the correlation of PM2.5 concentration to wind speed for winds flowing in the northwest (*upper left*), northeast (*upper right*), southeast (*bottom right*), and southwest (*bottom left*) directions.

Indeed, there is clearly stronger correlation to PM2.5 concentration in the heat maps for winds coming from the northwest and southeast directions than from the northeast and southwest directions. It appears that northwest winds tend to drive the PM2.5 concentration down as depicted in the row of red boxes nearest 0 μg/m^3, while southeast winds show a similar trend but with a narrower range of wind speeds. Wind speed and direction will be parameters worth examining closely as we create a machine learning model to forecast PM2.5 concentration.

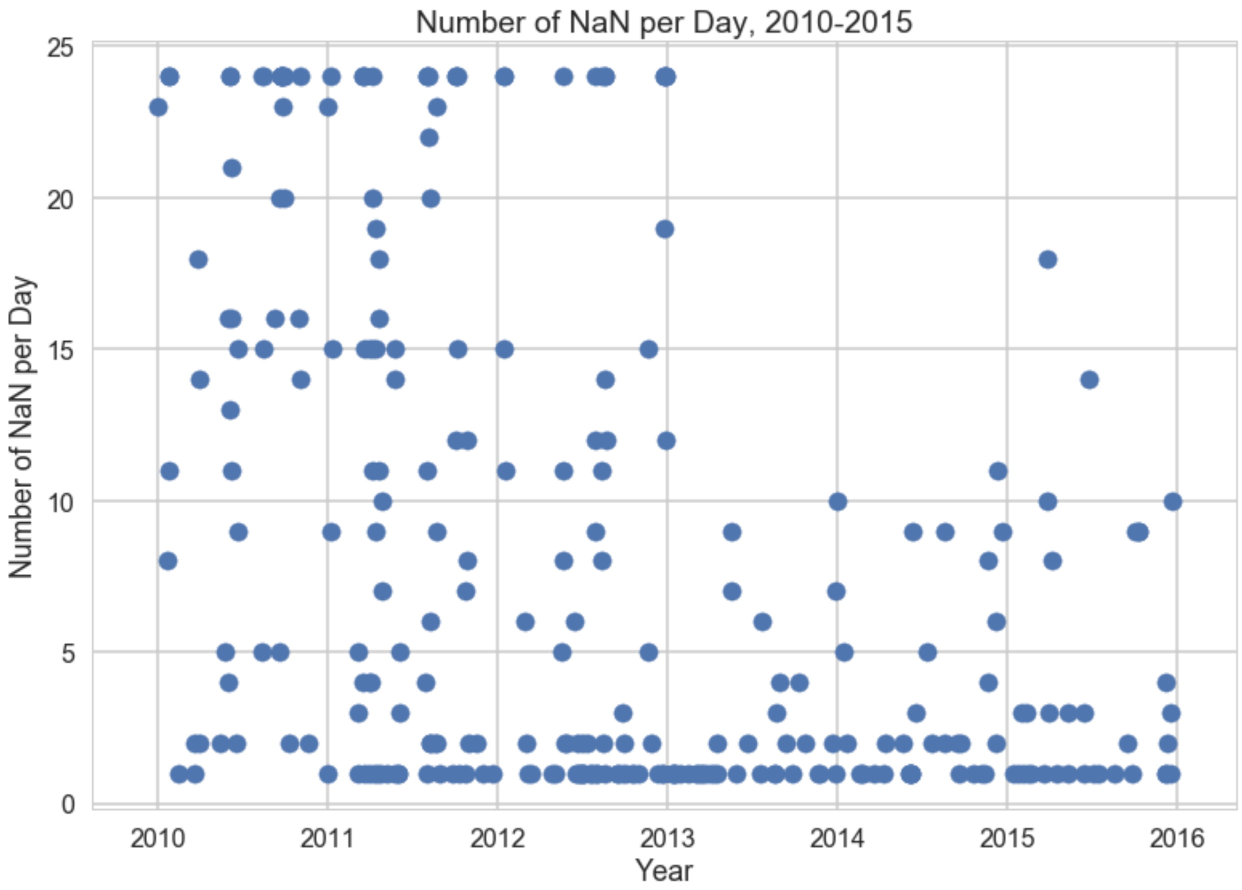

Figure 3.10: Count of NaNs per day from 2010 through 2015.

In addition to studying the parameters correlated to PM2.5 concentration, it is worthwhile to examine how reliably the measurement center records hourly data. Figure 3.10 plots the number of NaNs recorded each day from 2010 through 2015. The plot shows a fairly random distribution of points through the end of 2013 with many bunched together at one per day and also 24 per day, which signifies an entire day of NaNs. Starting from 2013, the number of NaNs exceeding 10 per day significantly drops. It is worth noting that `PM_Dongsi`, `PM_Dongsihuan`, and `PM_Nongzhanguan` began recording data in 2013, so there may be some relationship between the operation of these centers and the improvement in uptime for `PM_US Post`. A clearer depiction of the distribution of NaNs per day is provided in the histogram of Figure 3.11, which shows that the majority of these NaNs occur only once per day.

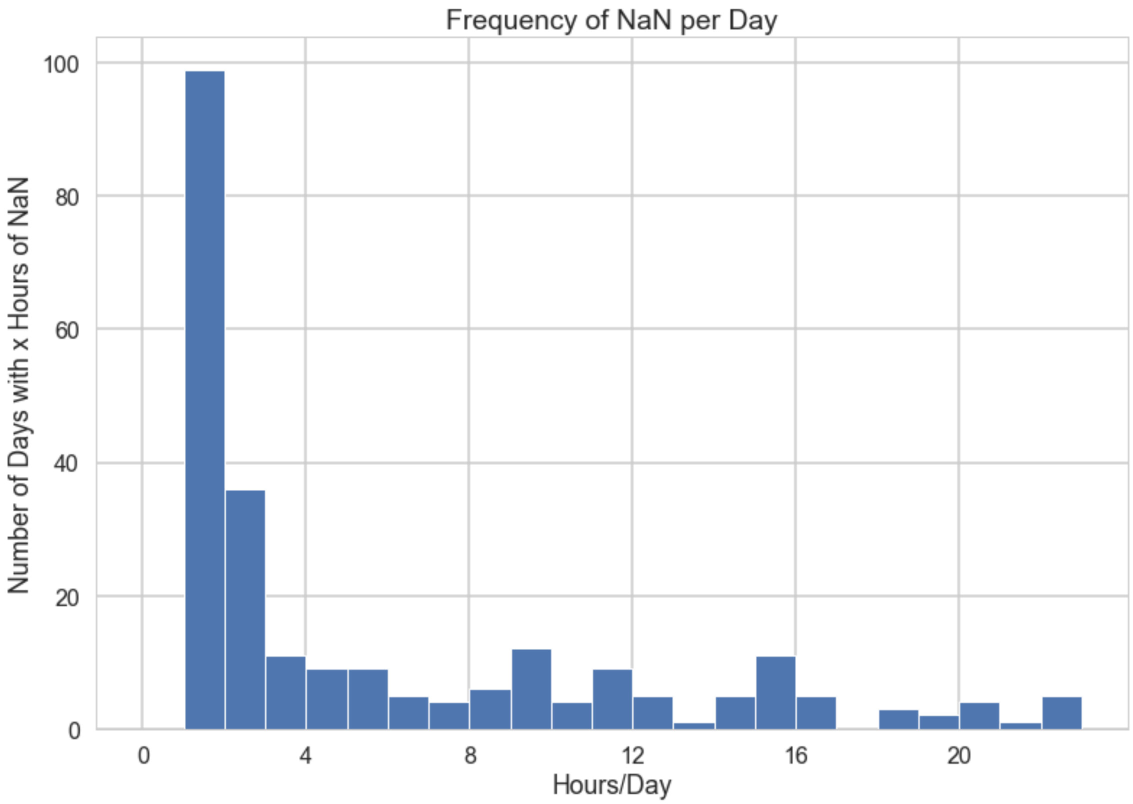

Figure 3.11: Distribution of NaNs per day.

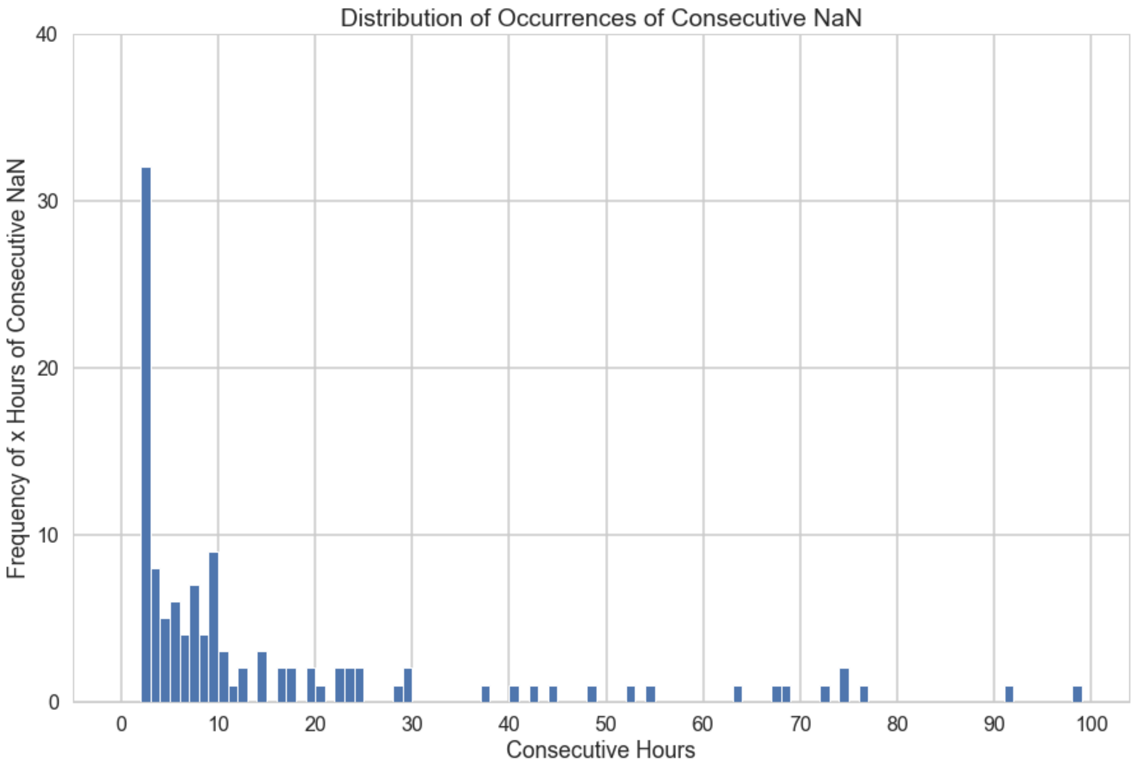

Figure 3.12: Distribution of consecutive hours of NaN.

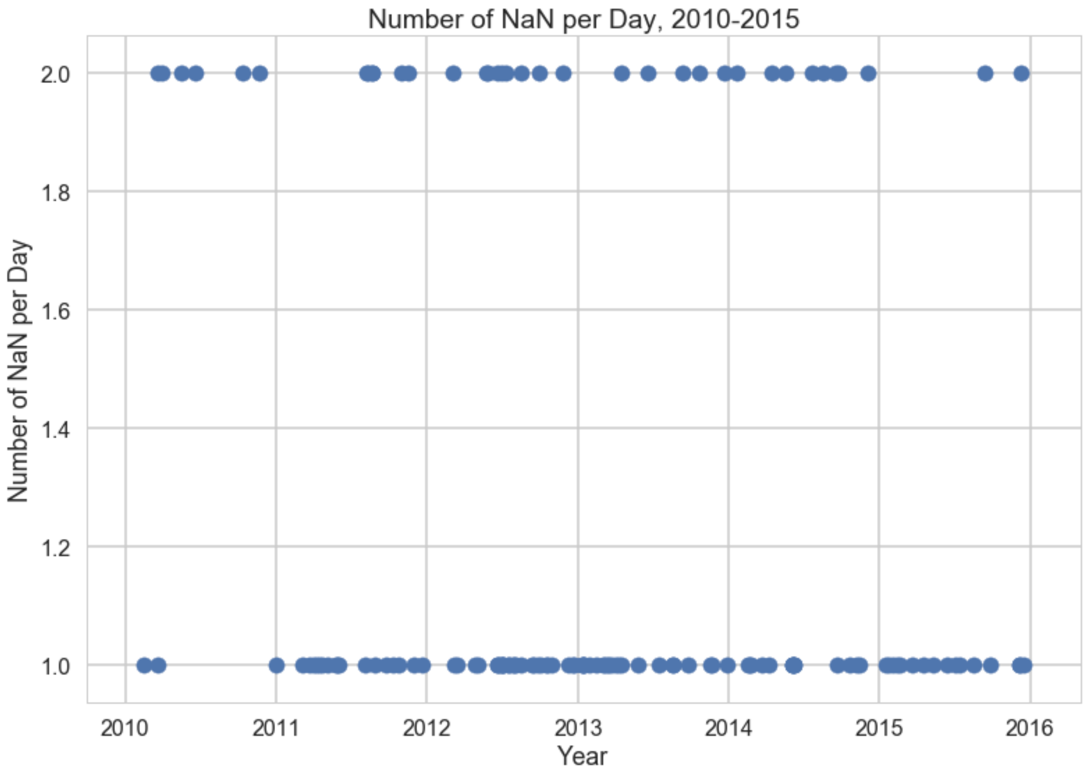

Figure 3.13: Distribution of NaNs per day for the cleaned dataframe `df`.

In order to distinguish between NaNs caused by random measurement error and outages due to equipment failure or maintenance, the NaNs are grouped according to the number of consecutive hours that they appear. Strings of NaNs spanning several consecutive hours are more likely to be outages rather than random errors. A distribution of consecutive hours of NaN is plotted in the histogram of Figure 3.12.

While there are numerous instances of NaNs lasting up to 10 hours consecutively, the number of instances beyond 10 hours drops to about 2 or 3, after which most of the outages exceeding 30 hours occurred only once from 2010 to 2015. While PM2.5 concentration data is provided for the majority of the time, the occurrence of these outages does reinforce the need for some means of forecasting future PM2.5 concentration when the measurement centers are out of order.

In terms of cleaning the data, especially when daily average PM2.5 concentration is needed, days containing an excessive number of NaNs are concerning since the mean may not be accurately represented by the rest of the measurements for that day. Therefore, all days containing 3 or more NaNs will be discarded. After this was done, the plot of Figure 3.10 was replotted in Figure 3.13, and the presence of days with only one or 2 NaNs per day confirm that this cleaning step successfully discarded all days with 3 or more NaNs.