Springboard Data Science Career Track
Capstone Project 1: Final Report

Using Machine Learning to Forecast Air Quality in Beijing

Kevin Limkrailassiri

1 Introduction

Poor air quality in major cities in China including Beijing is a well-known concern that has been drawing significant attention from news outlets, social media, and the Chinese government [1]. 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 [2]. 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 [3].

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 a weather forecast helps citizens arrange their weekly plans based on the predicted weather, a forecast of air quality can also provide citizens a means of anticipating the predicted air quality several days ahead. Moreover, the development of such a forecast can provide an understanding of weather trends corresponding to higher PM2.5 level, which can help the government proactively curtail pollution contributions from 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 in the data. The insights from this step will help us perform linear regression and supervised machine learning in order to construct a model that can forecast PM2.5 concentration based on weather data. A training and testing split will be used to train the machine learning model and evaluate 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 is 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 and combined wind direction (NW, NE, SE, or SW), and precipitation (mm). To evaluate the reliability

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

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.

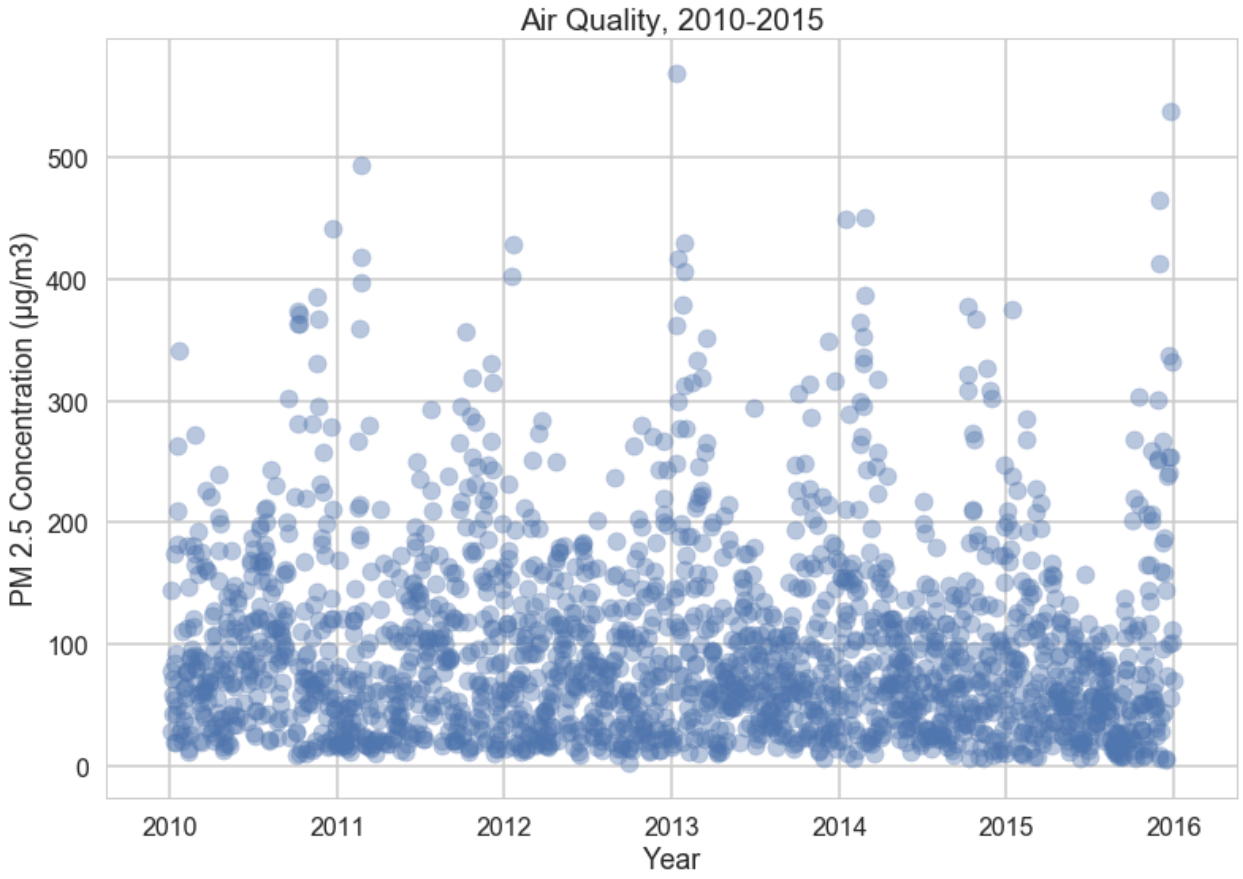

Figure 3.1: Plot of daily average PM2.5 concentration from 2010 through 2015 shows annual peaks near the start of each calendar year.

A plot of daily average air quality for the period of 2010 to 2015 in Figure 3.1 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 cold 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.

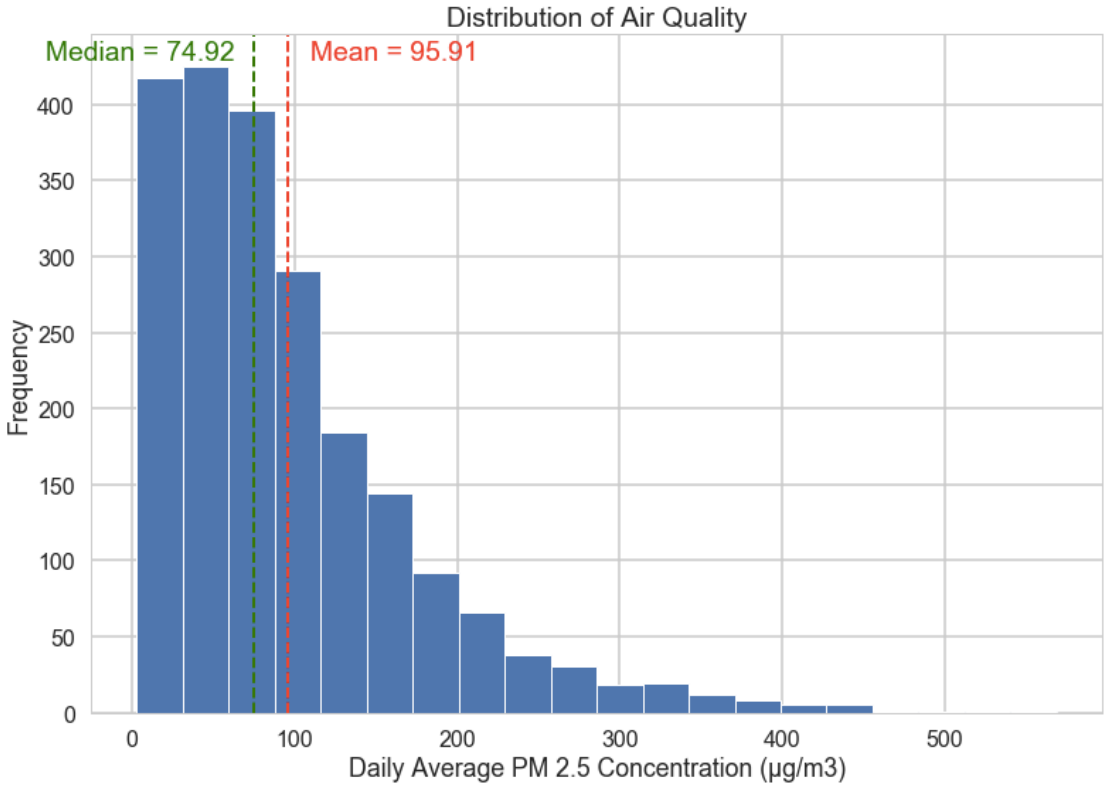

Figure 3.2: Distribution of daily average PM2.5 concentration with a mean of 95.91 µg/m^3 and median of 74.92 µg/m^3.

As shown in Figure 3.2, 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 description of each AQI classification is provided below in Table 3.1.

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).

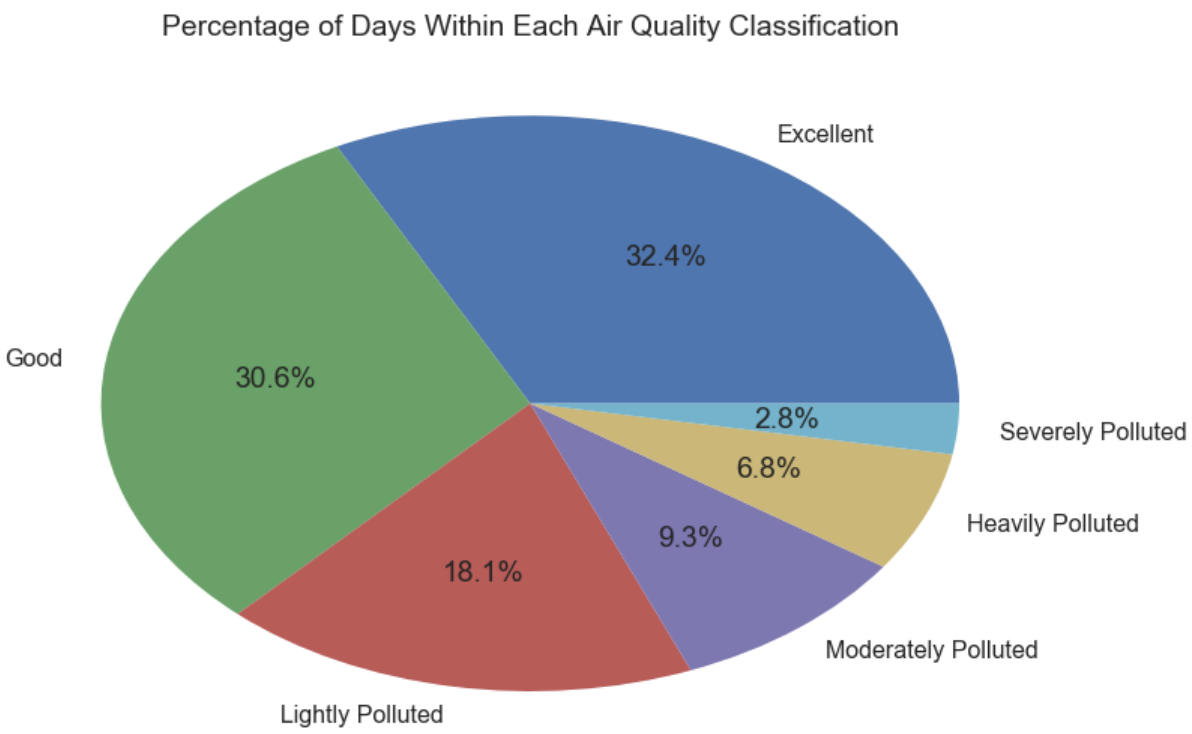

Figure 3.3: Percentage of days classified by each category of the Air Quality Index (AQI)

According to the AQI classification of air quality, the percentage of days falling under each classification is illustrated in the pie chart of Figure 3.3. 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.

The plot of monthly air quality from 2010 through 2015 in Figure 3.4 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.

The plot of daily PM2.5 concentration from 2010 through 2015 in Figure 3.5 shows very similar median and range across the entire week. Therefore, air quality appears to be insensitive to the particular day of the week.

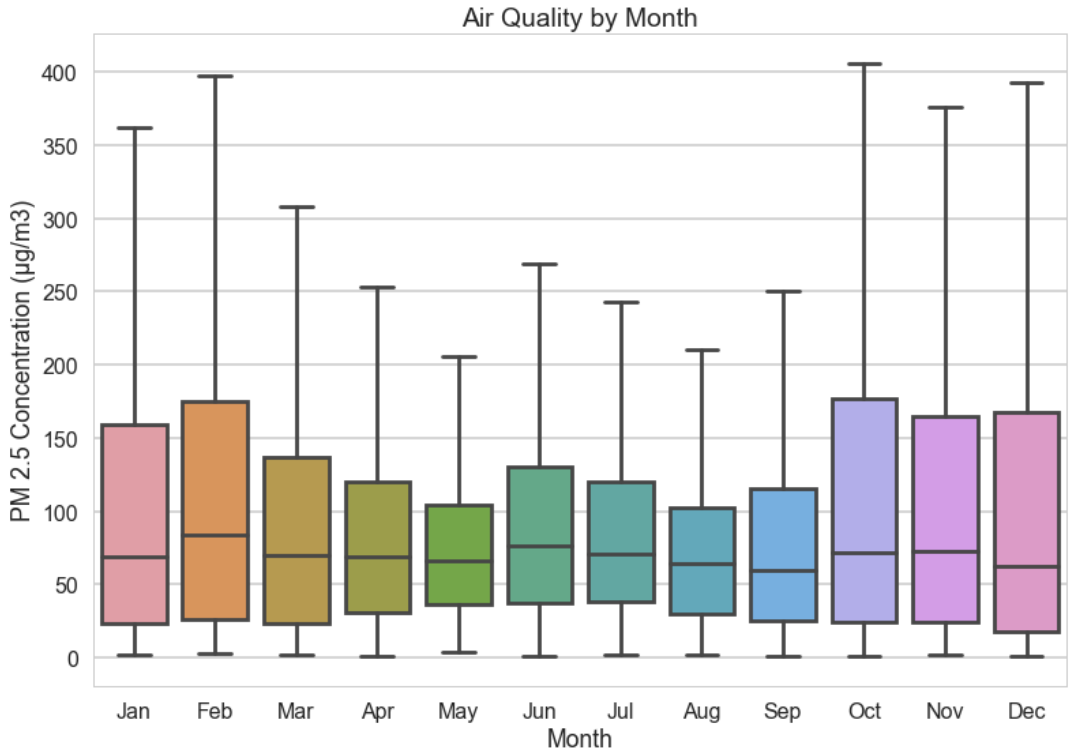

Figure 3.4: Distribution of air quality concentration by month of the year for the years 2010 through 2015.

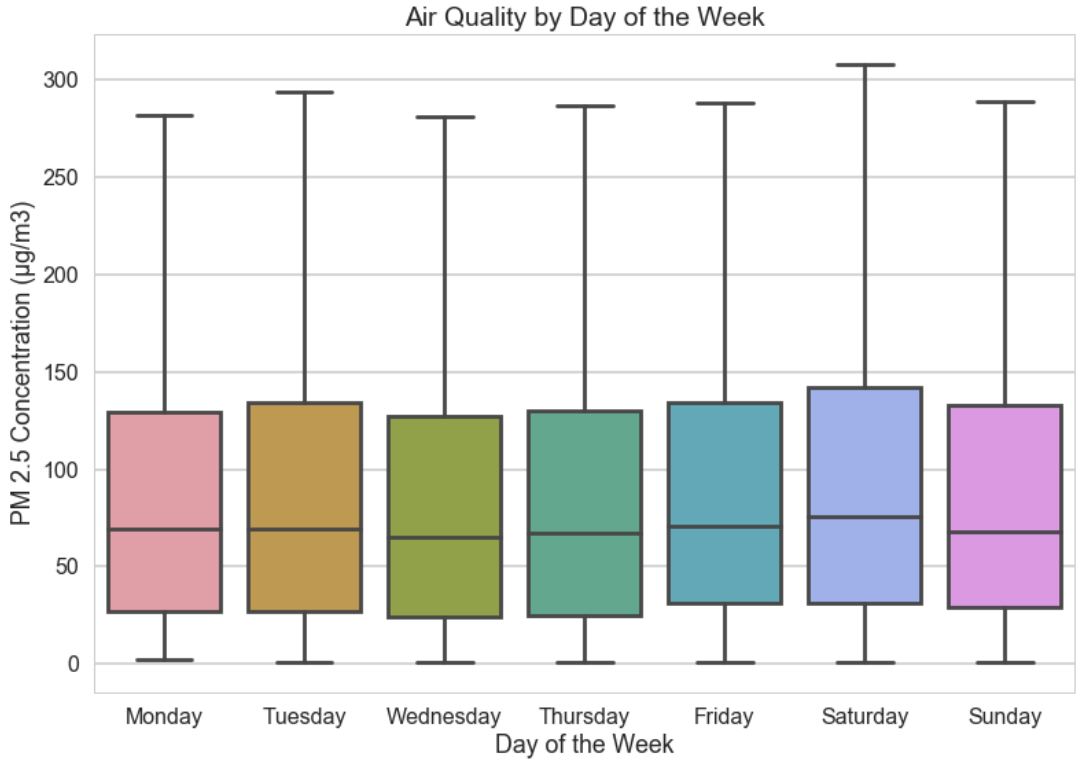

Figure 3.5: Distribution of air quality concentration by day of the week for the years 2010 through 2015.

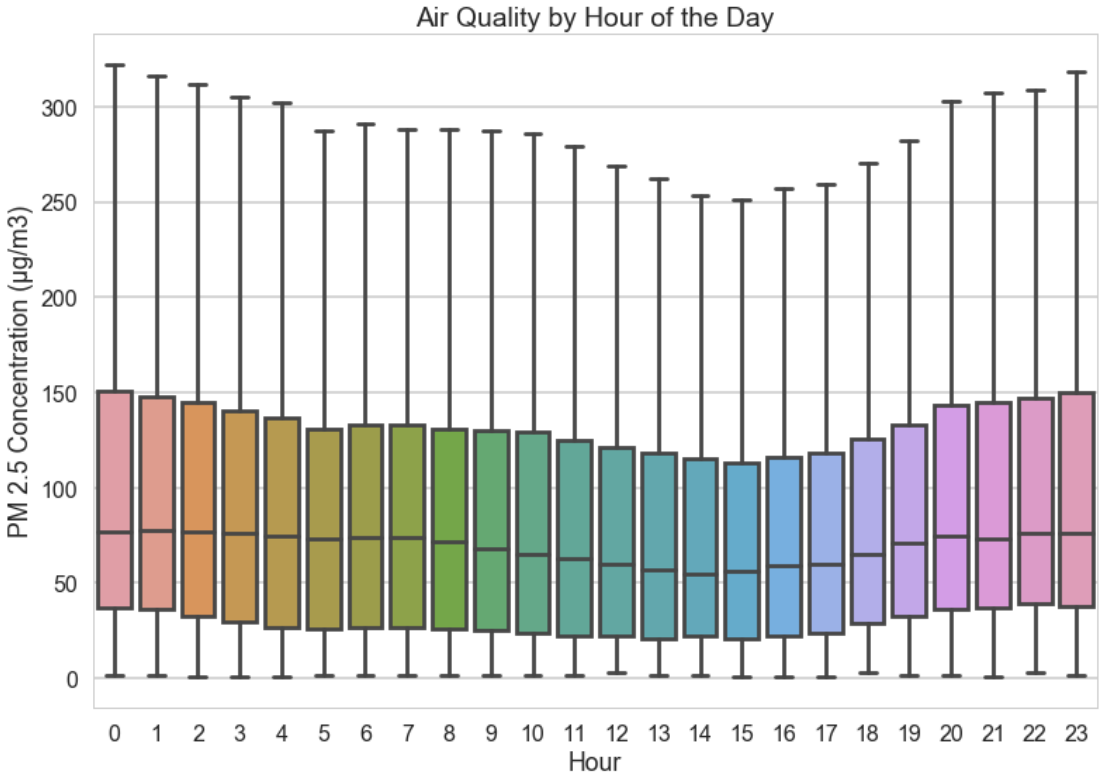

Figure 3.6: Distribution of air quality concentration by day of the week for the years 2010 through 2015.

As for the plot of hourly air quality in the span of a 24-hour day depicted in Figure 3.6, 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 reaching 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 do not strongly influence air quality compared to coal burned for heating during the evenings and early mornings.

Now, we explore PM2.5 concentration as a function of several weather-related parameters. 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 shows a weak positive correlation with two highly concentrated centers at PM2.5 concentration near 20 $\mu g/m^3$ 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. In contrast, the heat map correlating PM2.5 concentration to pressure shows the opposite effect with the PM2.5 concentration mildly decreasing 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 $\mu g/m^3$. 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 no or very little precipitation on most days, but this is correlated with lower values of PM2.5 concentration.

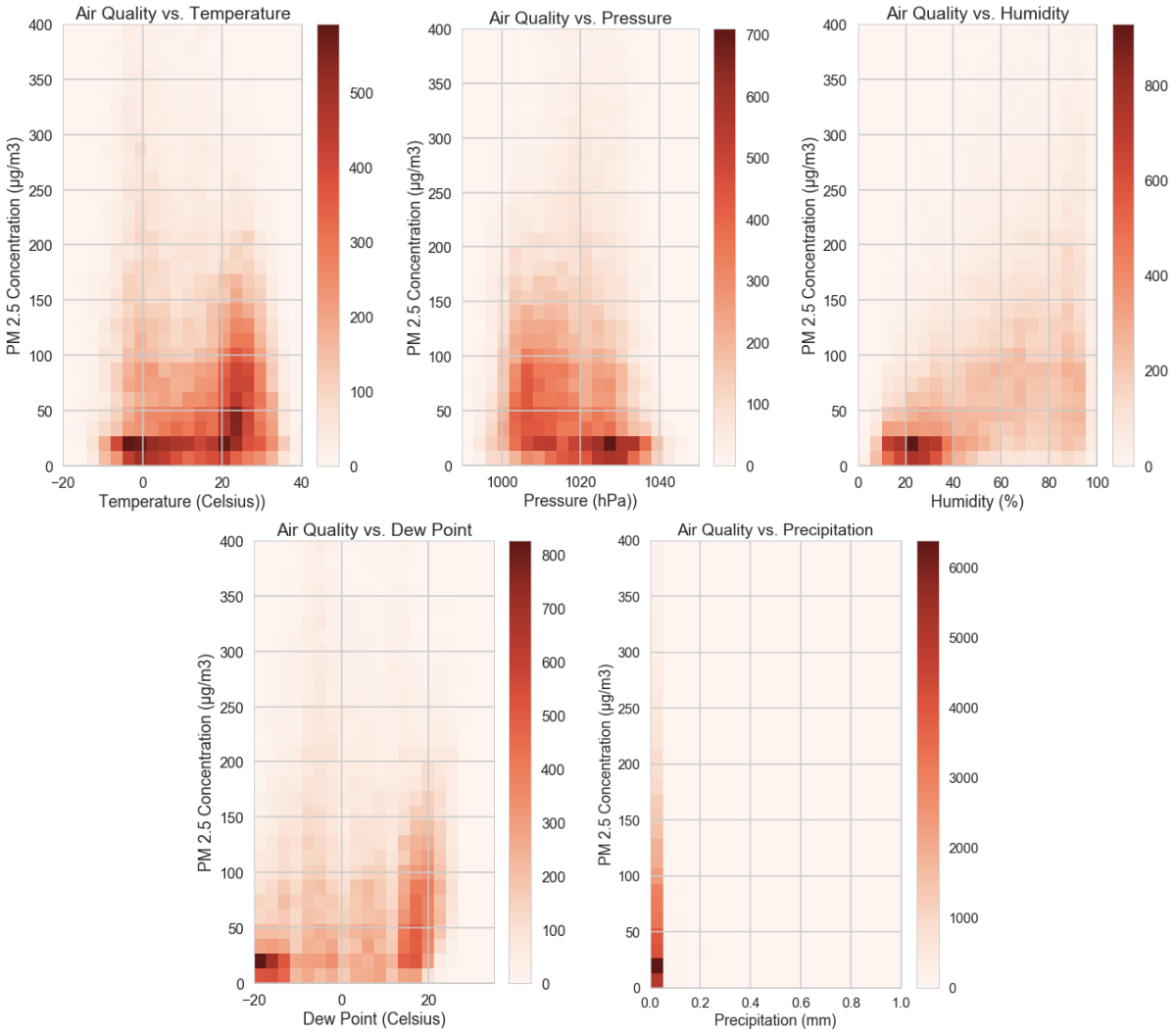

Figure 3.7: Heat maps correlating PM2.5 concentration to temperature, pressure, humidity, dew point, and precipitation.

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 in Figure 3.8. 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 in Figure 3.9.

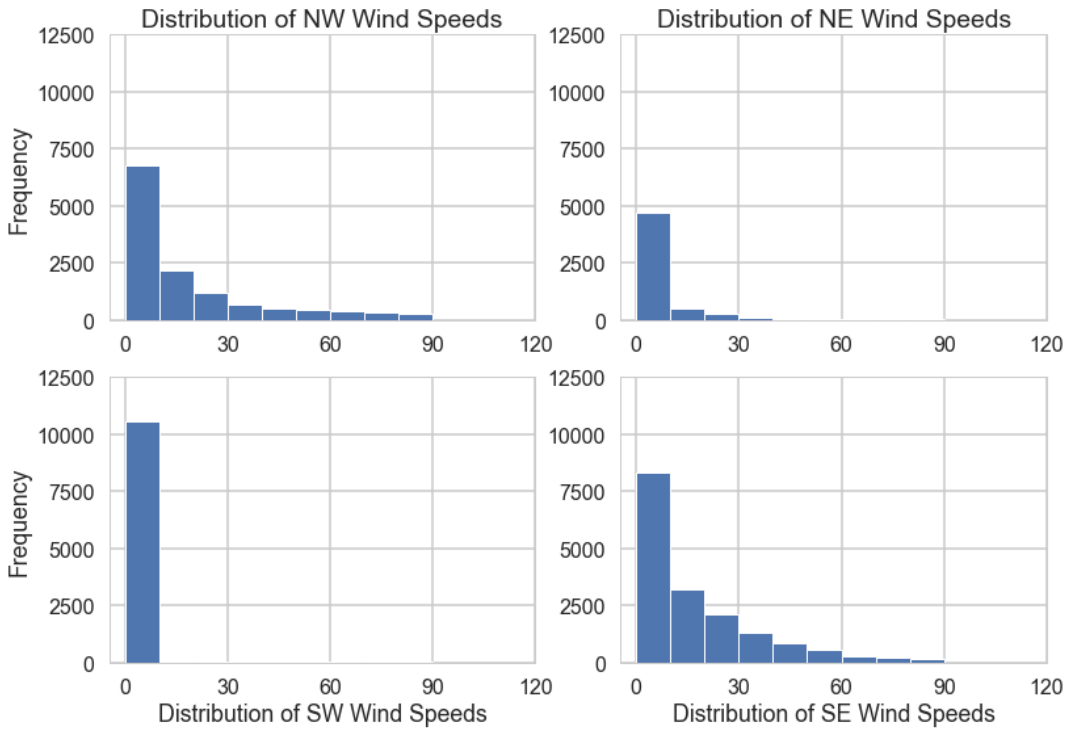

Figure 3.8: Distribution of wind speed for winds coming from the northwest (*upper left*), northeast (*upper right*), southeast (*bottom right*), and southwest (*bottom left*) directions.

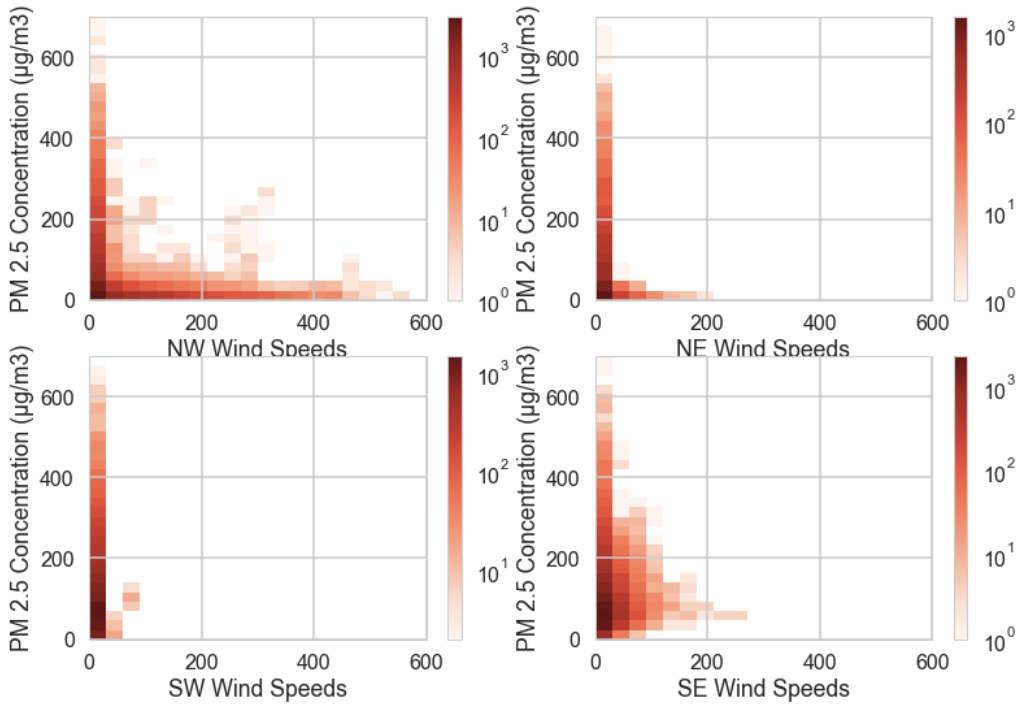

Figure 3.9: Heat maps showing the correlation of PM2.5 concentration to wind speed for winds flowing in the northwest (*upper left*), northeast (*upper right*), southeast (*bottom right*), and southwest (*bottom left*) directions.

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 $\mu g/m^3$, 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.

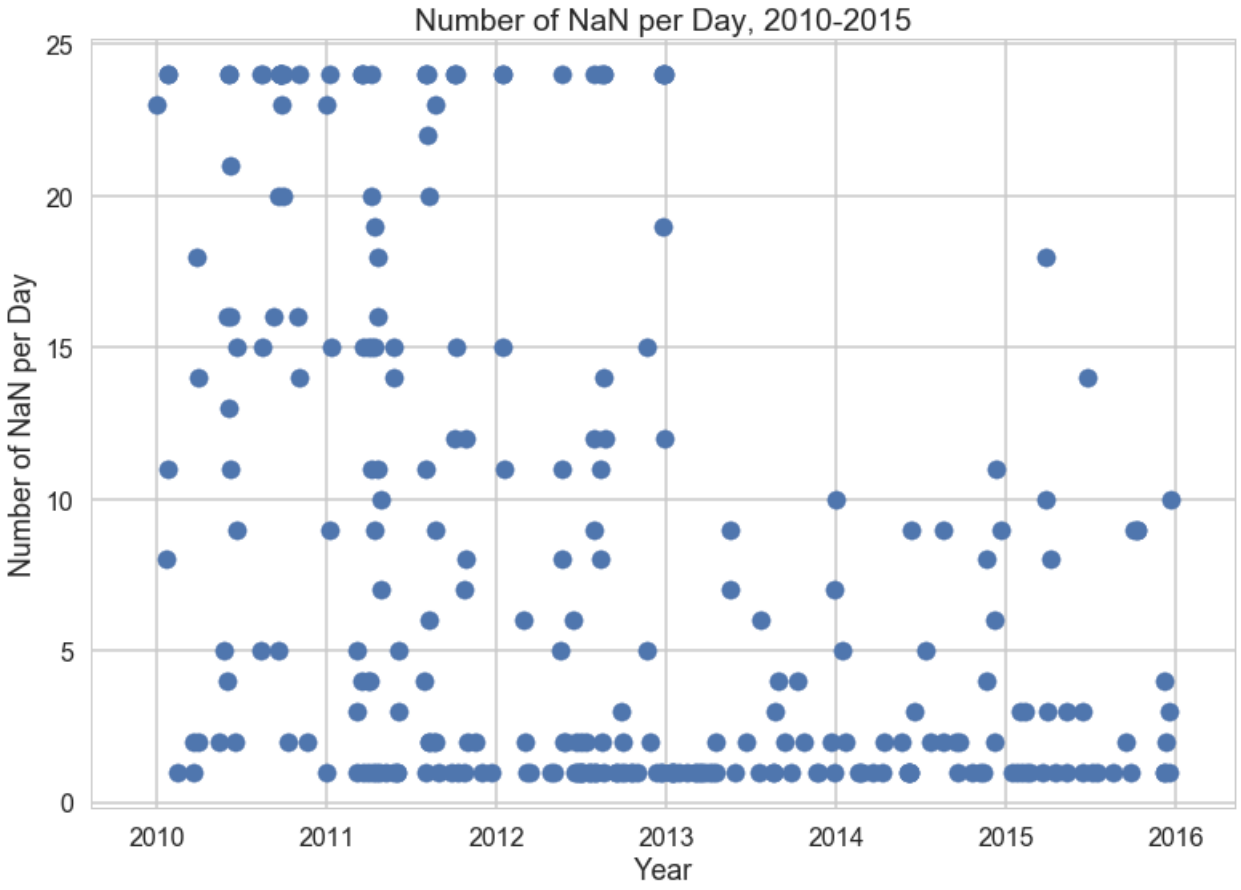

Figure 3.10: Count of NaNs per day from 2010 through 2015.

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.

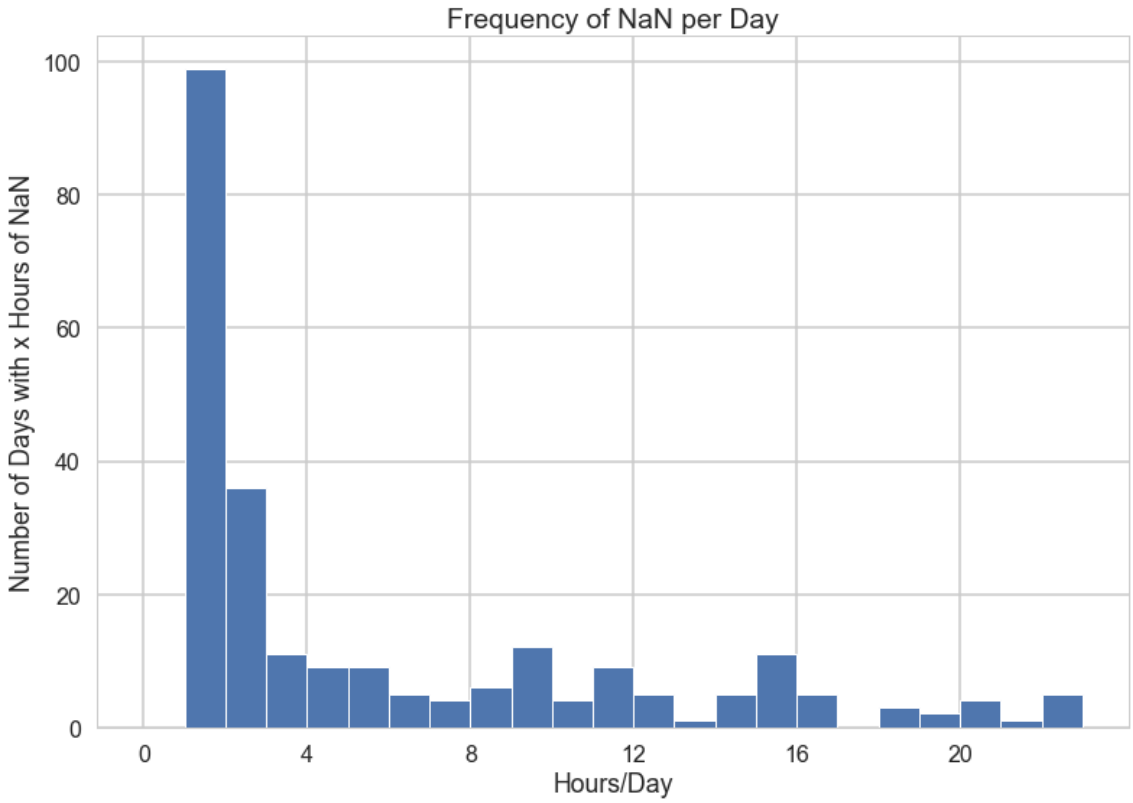

Figure 3.11: Distribution of NaNs per day.

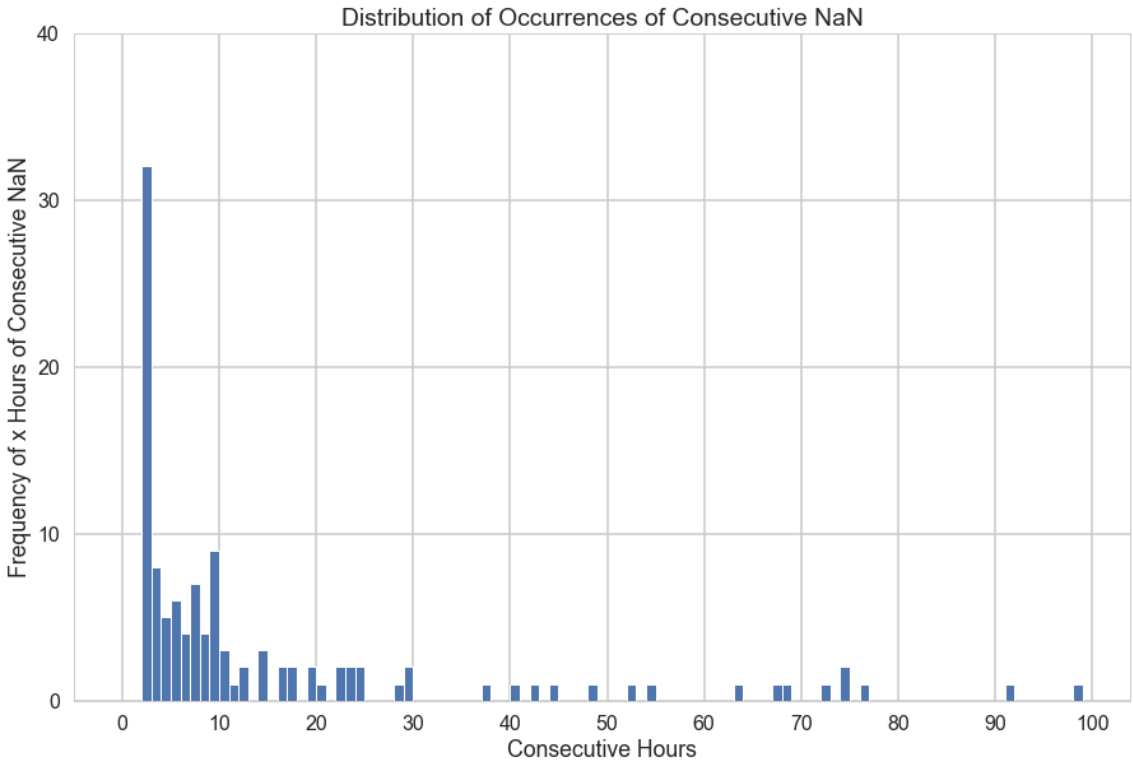

Figure 3.12: Distribution of consecutive hours of NaN.

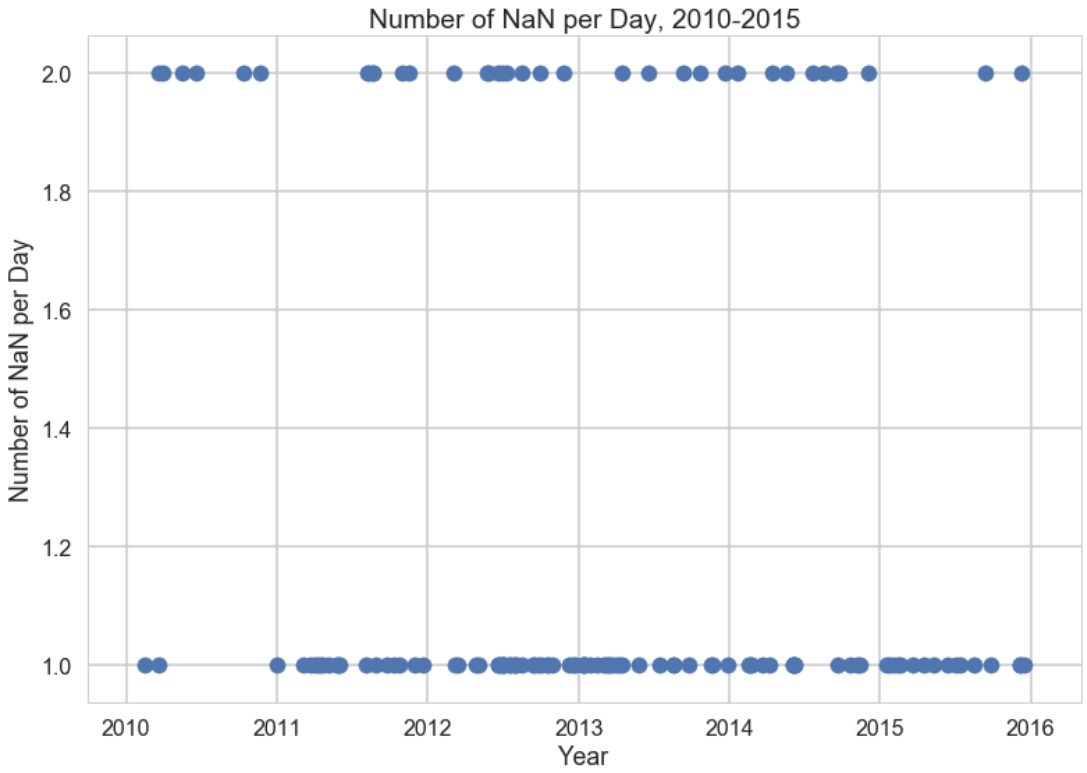

Figure 3.13: Distribution of NaNs per day for the cleaned dataframe `df`.

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.

4 Modeling

4.1 Linear Regression

We begin generating predictions of air quality concentration by first using regularized linear regression, namely ridge and lasso regression. The `Ridge` and `Lasso` modules from sci-kit learn were employed for this study. The first test used temperature as the lone feature. With each subsequent test, an additional feature was added to observe its influence on the coefficient of determination R^2 for both ridge and lasso regression. The R^2 score is a measure of how well the regression model explains the variability of the data. A summary of the tests and R^2 scores is provided in Table 4.1.

Features	**Ridge R^2**	**Lasso R^2**
`temp`	0.0185	0.0187
`temp, humidity`	0.197	0.197
`temp, humidity, pressure`	0.207	0.207
`temp, humidity, pressure, dew_point`	0.207	0.207
`temp, humidity, pressure, dew_point, precipitation`	0.211	0.211
`temp, humidity, pressure, dew_point, precipitation, [NW,NE,SE,SW]`	0.242	0.242

Table 4.1: Features included in the regression tests and corresponding ridge and lasso scores.

The addition of humidity shows by far the strongest increase in R^2 scores from among all the weather-related parameters. The fact that humidity strongly correlates with air quality concentration is consistent with the physicochemical phenomenon of water vapor trapping air pollutants and keeping them in the atmosphere [4]. Apart from humidity, the rest of the features weakly correlate with air quality concentration, and the R^2 score of 0.242 when all features are incorporated into the model shows that the variability of the data is difficult to describe based on this set of features. Supplementing the model with additional features that correlate strongly with air quality concentration may allow the model to track the air quality concentration more closely.

To visualize how the model compares with the actual air quality concentration, the actual and predicted data are plotted together in Figure 4.1. While the actual data exhibits several sharp peaks, the predicted data is mostly bound between 0 to 200 μg/m^3. This is due to the coefficients that were used to fit the regression model to the data. The regression model likely yields a better overall fit when it tracks well with air quality concentration up to 200 μg/m^3 than if it were to prioritize values over 200 μg/m^3. Lastly, the lone negative value has no practical significance; it is an erroneous prediction caused by the coefficients that were selected by the model. The root-mean-square-error between the actual and predicted data is 80.69. In other words, for every observation there is an average error of 80.69 μg/m^3 greater than or less than the actual air quality concentration.

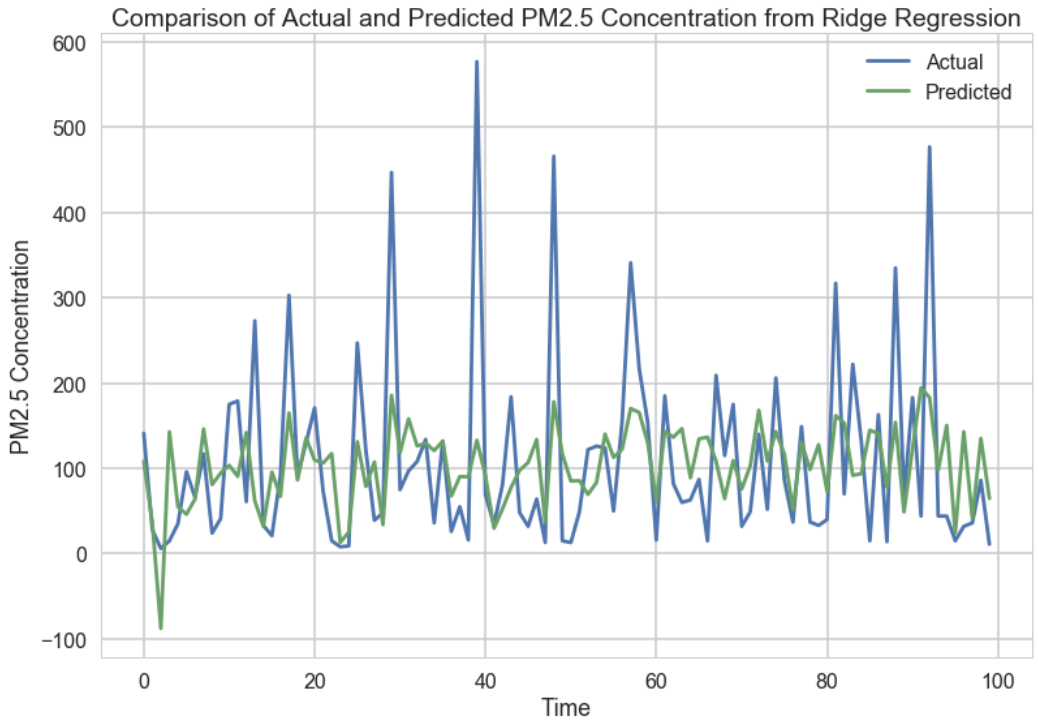

Figure 4.1: Comparison of actual and predicted PM2.5 concentration provided by ridge regression.

4.2 Machine Learning with Long Short-Term Memory (LSTM)

A limitation of the above linear regression analysis is that it does not consider the time-dependence of the features and target variable. From the earlier exploratory data analysis, we observed that air quality and weather do follow a cyclical trend through the span of one day and also one year. Therefore, the time-dependence of air quality and weather measurements can be leveraged to possibly generate more accurate air quality predictions rather than treating every observation of air quality and weather as independent of all other observations.

In this vein, we employ machine learning through a specific kind of Recurrent Neural Network (RNN) called Long Short-Term Memory (LSTM). Recurrent Neural Networks perform a sequence of the same operation wherein the output of one operation is retained as the input of the next operation, thereby affording the model the capacity to retain memory as it is being trained by a set of data. Long Short-Term Memory improves upon this concept by minimizing an effect known as vanishing gradients, in which memory recorded several iterations prior is lost in favor of memory recorded recently. This is what allows LSTM to retain extremely long sequences of information [5].

We begin with a simple LSTM model and plot the evolution of RMSE over 100 training epochs as well as a sample period of 100 time-steps to visualize how well the model follows the actual air quality concentration, shown in Figure 4.2.

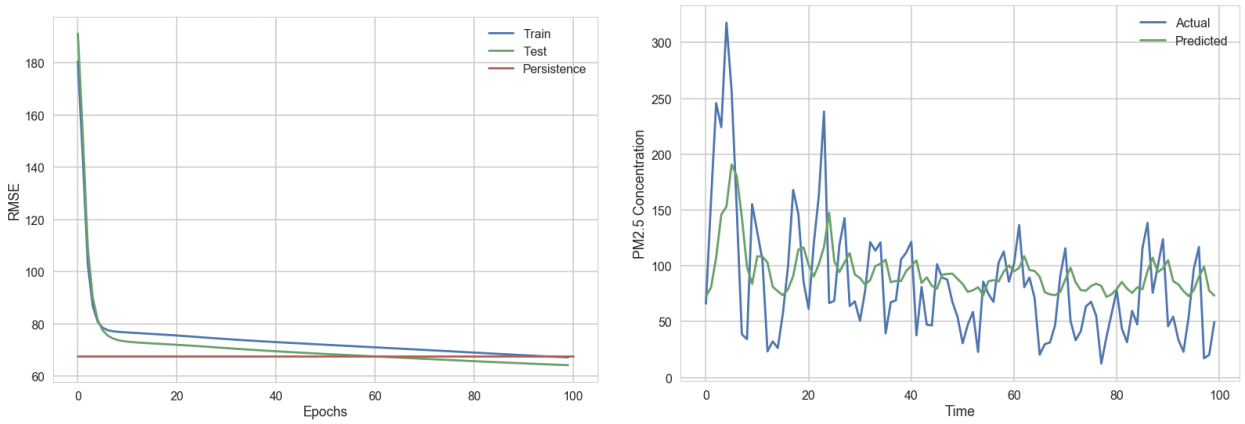

Figure 4.2: Results of LSTM model with 1 feature (`pm25`), 1 target (`pm25`), 5 prior time steps, 70:30 train-test ratio, batch size of 10, and 1 neuron in 1 LSTM stack trained over 100 epochs. Test RMSE: 64.26.

For all plots, a persistence model is included as a standard of reference to evaluate the skill of the LSTM model. The persistence model simply sets the value at the next time-step as the value at the time-step immediately preceding it. For example, the prediction for tomorrow's air quality concentration is simply set as the known air quality concentration today. If the LSTM model does not exhibit skill better than the persistence model, it would be moot to employ the LSTM model at all. The first result shows that the test RMSE falls just below the persistence model by the end of 100 epochs. It is interesting to note that the train set normally shows smaller RMSE compared to the test set, but in this case the LSTM model fits the test set better than the train set. This is possibly due to the fact that the train set contains more volatile changes in the air quality concentration compared to the test set, making it harder for the LSTM model to fit the train set. Using some form of cross validation or incorporating more prior time steps may smooth over this effect. We observe also for the sampling of 100 time-steps that the predicted values are constrained within a tighter range of air quality concentration value compared to the actual values. We look to improve upon this first attempt by increasing the capacity for retention through increasing the number of neurons from 1 to 5.

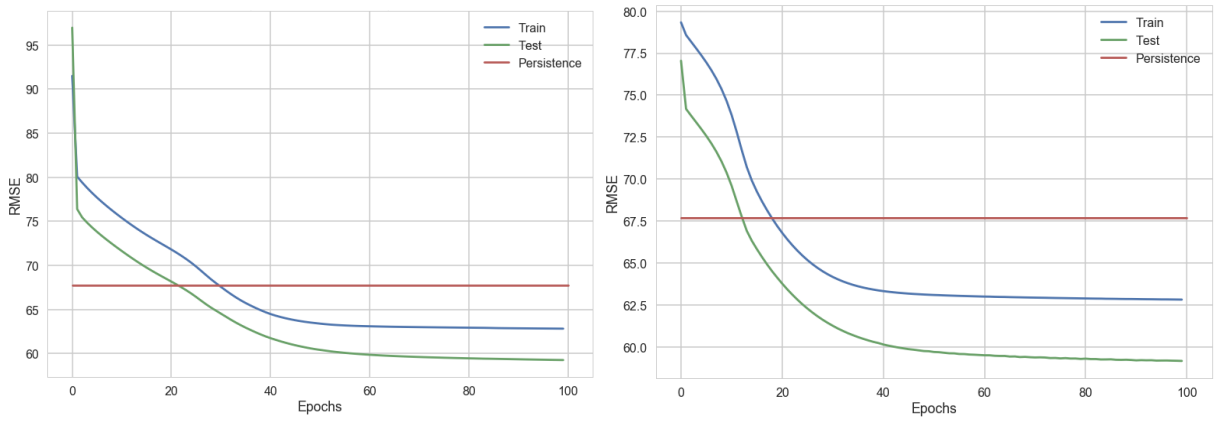

Figure 4.3: **Observing effect of varying prior time steps**. (*Left*) Model trained on 5 prior time steps. Test RMSE: 59.26. (*Right*) Model trained on 20 prior time steps. Test RMSE: 59.19. Both LSTM models were designed with 1 feature (`pm25`), 1 target (`pm25`), 70:30 train-test ratio, batch size of 10, and 5 neurons in 1 LSTM stack trained over 100 epochs.

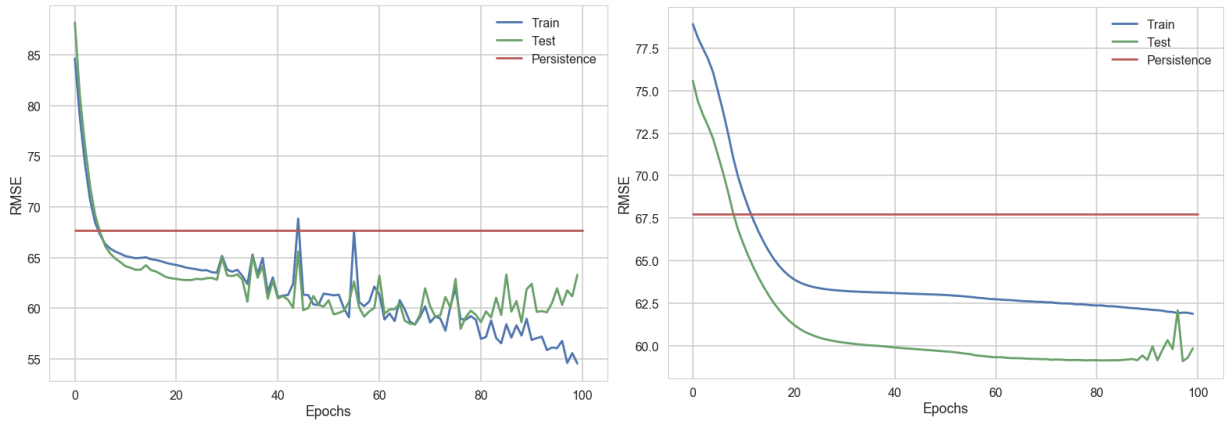

Figure 4.4: **Observing effect of varying batch size**. (*Left*) Model trained with batch size of 1. Test RMSE: 57.97. (*Right*) Model trained with batch size of 10. Test RMSE: 59.08. Both LSTM models were designed with 1 feature (`pm25`), 1 target (`pm25`), 20 prior time steps, 70:30 train-test ratio, and 20 neurons in 1 LSTM layer trained over 100 epochs.

The increase of neurons from 1 to 5 causes a clear improvement in the skill of the LSTM model, as it now easily outperforms the persistence model. The comparison of the LSTM model with 5 prior time steps and 10 prior time steps in Figure 4.3 shows that the plots reach roughly the same RMSE by the end of 100 epochs, but the model with 10 prior time steps exhibits a faster descent to this final value. This suggests that more time steps allow the model to learn the trends in the data more quickly.

Varying the batch size between 1 and 10 produces distinctly different trends in the RMSE with increasing epochs. The model with batch size 10 (Figure 4.4 *right*) looks similar to the plots of Figure 4.3, but the model with batch size 1 (Figure 4.4 *left*) exhibits trends in the RMSE typical of overfitting. Namely, as training epochs increase, the RMSE for the train set and test set deviate significantly as the model overfits the train set at the expense of fitting the test set. Typically, training is interrupted once the RMSE of the test set begins to increase, in this case at around the 80th epoch.

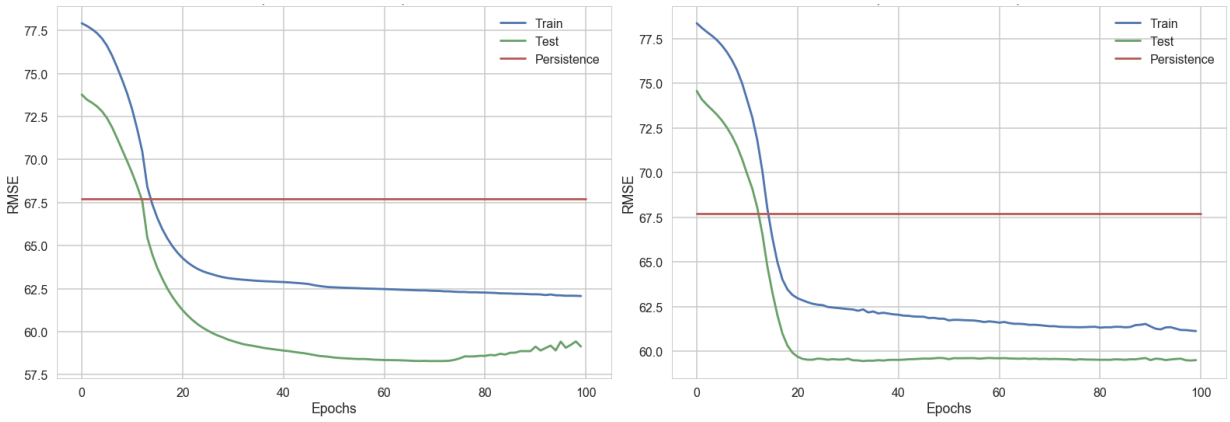

Figure 4.5: **Observing effect of varying neurons**. (*Left*) Two LSTM layers with 5 neurons each. Test RMSE: 58.29. (*Right*) Two LSTM layers with 10 neurons each. Test RMSE: 59.46. Both LSTM models were designed with 1 feature (`pm25`), 1 target (`pm25`), 20 prior time steps, 70:30 train-test ratio, batch size of 10, trained over 100 epochs.

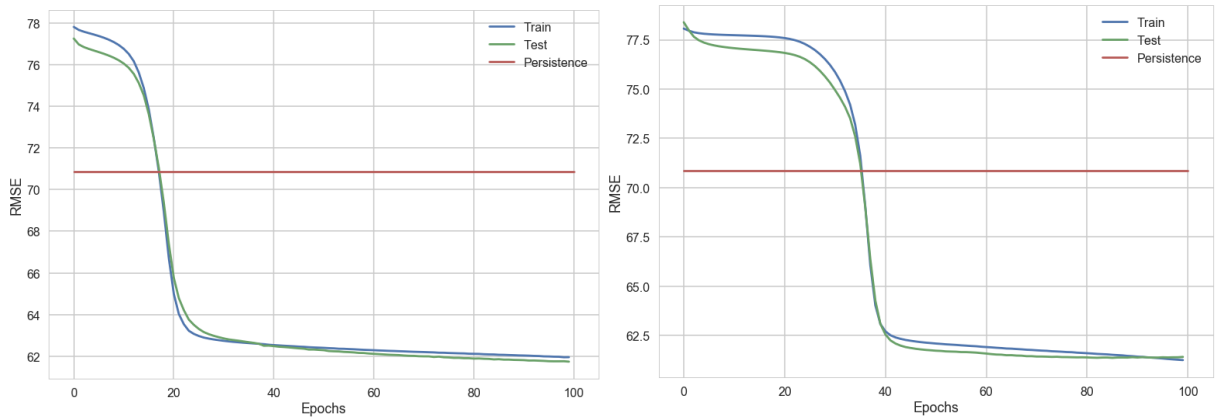

Figure 4.8: **Increasing prior time steps to 365 days.** (*Left*) Two LSTM layers with 365 prior time steps. Test RMSE: 61.74. (*Right*) Three LSTM layers with 365 prior time steps. Test RMSE: 61.37. Both LSTM models were designed with 1 feature (`pm25`), 1 target (`pm25`), 70:30 train-test ratio, batch size of 10, and 10 neurons trained over 100 epochs.

The presence of overfitting possibly signals the need for either more training examples or greater learning capacity for the LSTM model. We will increase the number of LSTM layers from 1 to 2 to see what kind of effect it will have in the skill of the LSTM model. Figure 4.5 shows an LSTM model of 2 layers with either 5 neurons or 10 neurons in each layer. The model with 10 neurons shows a convergence of RMSE for the train and test sets. This seems promising as it suggests that the model is doing a better job of fitting the train set without overfitting either the train or test set.

Considering that LSTM models have capacity to retain long-term dependencies, what if we significantly increased the number of prior time steps? Figure 4.8 shows 2 LSTM models with 365 prior time steps, denoting 1 year's worth of training data, with 1 model containing 2 LSTM layers and the other 3 LSTM layers. It appears that the model with 2 LSTM layers is able to learn the trends in the data more quickly, but the 2 models end with nearly the same RMSE. There

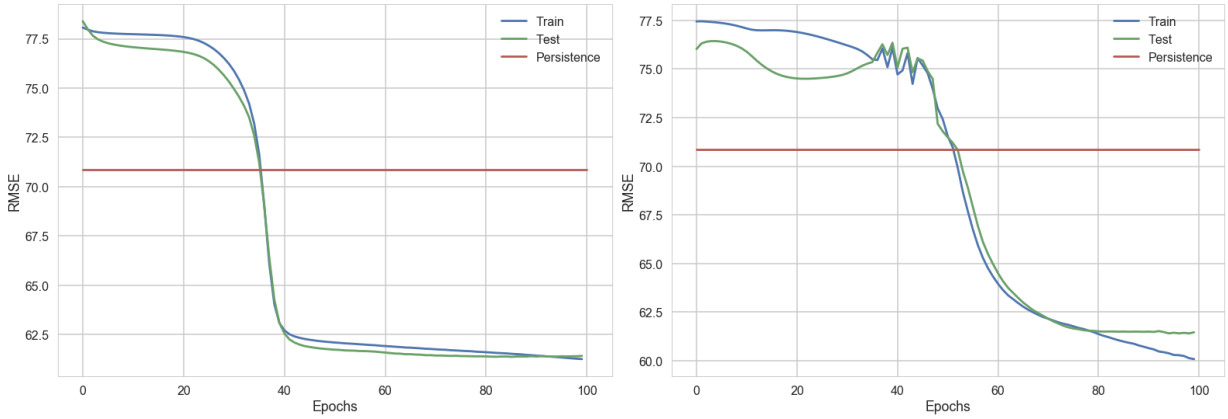

Figure 4.9: **Adding more features to the model.** (*Left*) Three LSTM layers with 1 feature (`pm25`) and 1 target (`pm25`). Test RMSE: 61.37. (*Right*) Three LSTM layers with 3 features (`pm25, dew_point, humidity`) and 1 target (`pm25`). Test RMSE: 61.40. Both LSTM models were designed with 365 prior time steps, 70:30 train-test ratio, batch size of 10, and 10 neurons in each layer trained over 100 epochs.

appears to be minimal further benefit to continuing to increase the number of neurons or layers at this point. The variation in day-to-day air quality concentration may be too difficult to accurately model using LSTM.

We recall that all the LSTM models evaluated up to this point accounted for only 1 feature, `pm25`, used to predict the target variable, which is also `pm25`. Since humidity was revealed to be strongly correlated to air quality concentration, what if we expanded the LSTM model to account for more features? Figure 4.9 includes `dew_point` and `humidity` as additional features into the LSTM model, and the resulting test RMSE score is roughly identical to the model with only 1 feature, `pm25`. The addition of features increases the complexity of the model and therefore the number of training epochs required before the RMSE curve drops sharply. Additionally, overfitting appears at around the 80th epoch when the train RMSE continues to drop while test RMSE plateaus.

Comparing all the RMSE values among the models may lead one to wonder why more neurons, more prior time steps, or more layers are needed if the RMSE value does not appear to change much. In fact, even the simpler models have slightly lower RMSE than the more complex models with more neurons, prior time steps, and layers. It is helpful first to note that the low RMSE values of all the single-layer LSTM models and some of the double-layer LSTM models are a bit misleading since the test RMSE outperformed the train RMSE. These test RMSE values are probably too optimistic and should be taken with a grain of salt, as it is expected that the test RMSE should always be less than the train RMSE since the LSTM model trains and fits to the train set. If these models were trained with a larger number of prior time steps, the train and test RMSE would likely converge. Given this, it is not enough to evaluate the performance of an LSTM model based on test RMSE alone. The evolution of RMSE as the model is trained gives clues to the quality of the model's fit to the data. In the LSTM models of Figure 4.8, the train and test RMSE follow one another closely and finally settle on very similar RMSE scores. This is an example of a model exhibiting good fit to the data. A summary of all LSTM models tested is provided Table 4.2.

Lastly, a comparison of the actual and predicted air quality concentration of Test #9 is shown in Figure 4.10. The predicted values seem to track more closely with the actual data compared to the linear regression model. Although the LSTM model does not seem to follow the actual values

Test #	LSTM layers	Neurons	Prior Time Steps	Batch size	Features	Profile of fit	Test RMSE
1	1	1	5	10	1	Underfitted	64.26
2	1	5	5	10	1	Underfitted	59.26
3	1	5	20	10	1	Underfitted	59.19
4	1	20	20	1	1	Overfitted	57.97
5	1	20	20	10	1	Underfitted	59.08
6	2	5	20	10	1	Underfitted	58.29
7	2	10	20	10	1	Underfitted	59.46
8	2	10	365	10	1	Good fit	61.74
9	3	10	365	10	1	Good fit	61.37
10	3	10	365	10	3	Overfitted	61.40

Table 4.2: Summary of LSTM models.

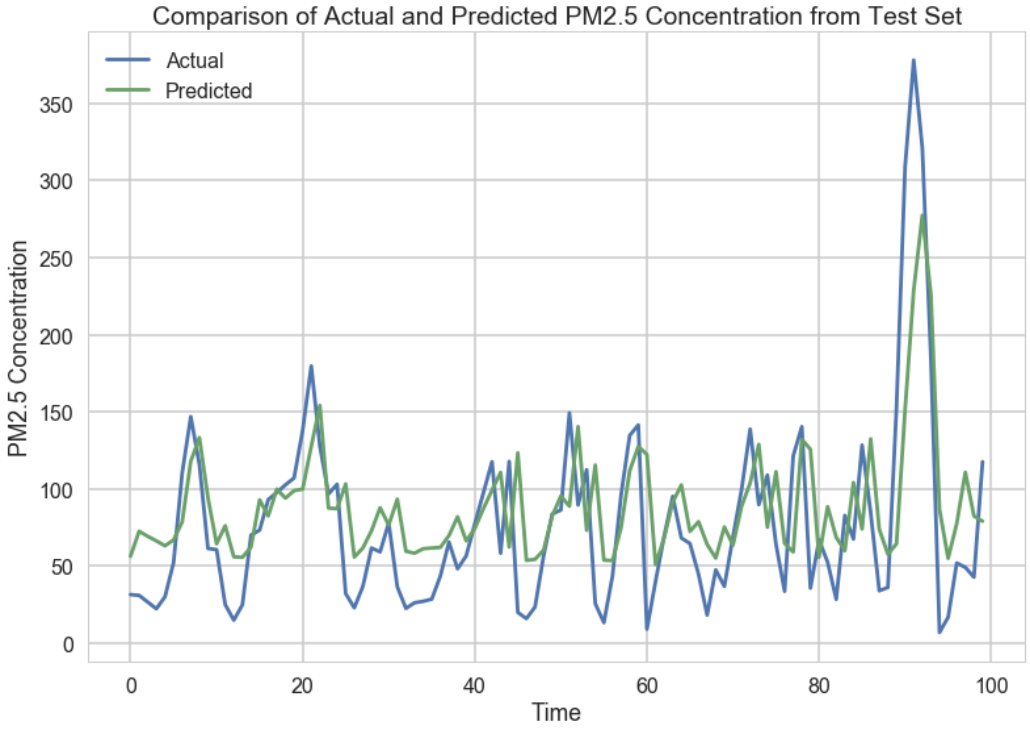

Figure 4.10: Comparison of actual and predicted PM2.5 concentration for an LSTM model designed with 365 prior time steps, 70:30 train-test ratio, batch size of 10, and 10 neurons in each of 3 layers trained over 100 epochs.

below 50 μg/m^3, the model does attempt to track sharp peaks in the actual data.

5 Discussion

From the LSTM models tested, it is clear that designing an LSTM model exhibiting good fit and highly accurate predictive capability is not straight-forward. Designing LSTM models is an empirical process requiring many iterations of testing, evaluation and redesign. It is conceivable that improvements in prediction accuracy can be gained through further tuning of LSTM layers, neurons, or prior time steps. To illustrate how accurate the existing models are, the PM2.5 air concentration is plotted in Figure 4.11, and the points are colored based on the predicted AQI classification according to the LSTM model of Test #10. The lightly-shaded regions show the ranges of the AQI classifications so that, for example, an orange point falling within the yellow region denotes an inaccurate prediction of AQI classification. Of all the points, 32.6% of the predictions fell into the correct AQI classification, which is about double the rate of random guessing, 16.7%.

Currently, air quality forecasting is available online via websites such as the World Air Quality Index Project, which relies on satellite observations to make predictions of air quality akin to weather forecasting [6]. Figure 4.12 shows a screenshot from aqicn.org taken on Friday, January 12, 2018 of the actual and forecasted AQI classifications for the previous two weeks, and of those instances when both an actual and forecasted AQI classification are provided, only 19 out of 83 or

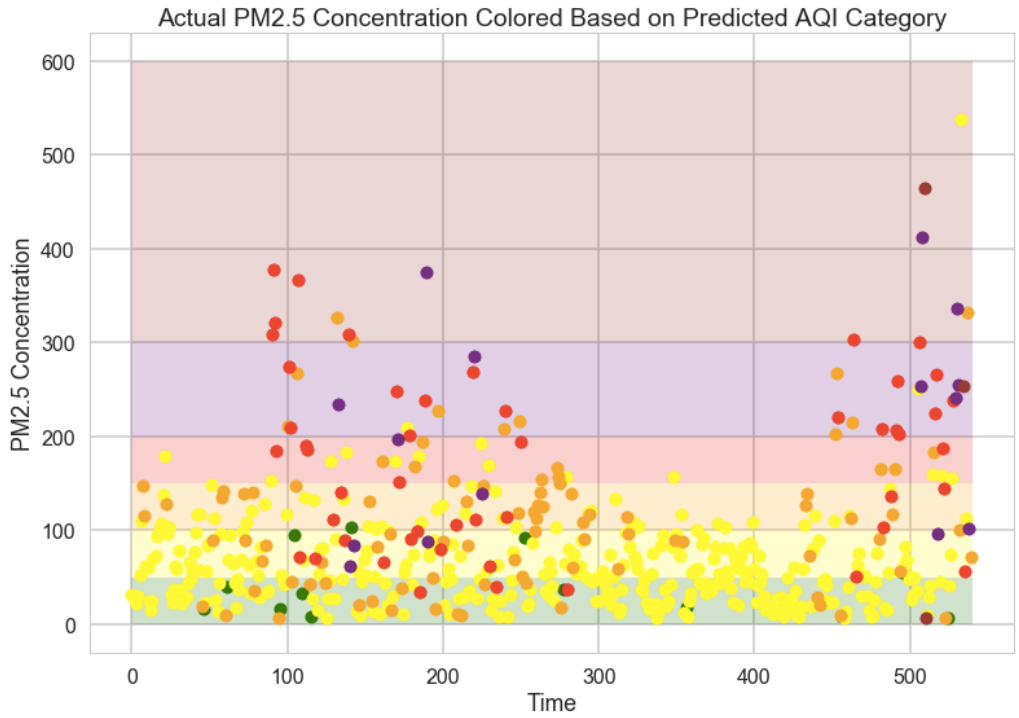

Figure 4.11: Actual PM2.5 concentration plotted with color according to predicted AQI category

22.9% of the predictions were accurate. Furthermore, a team of Microsoft researchers in Beijing developed air quality forecasting using a linear regression-based model to account for local nuances in air quality, a neural network-based model for modeling global trends, a dynamic aggregator merging these models with meteorological data, and an inflection predictor for sudden changes in air quality, and their model achieves an RMSE of 30 μg/m^3 for predictions 1-6 hours into the future and an RMSE of 64 μg/m^3 for predictions 7-12 hours into the future [7]. In comparison with these results, the predictions of the LSTM models suddenly appear more promising.

6 Conclusion and Recommendations

We have employed linear regression and a type of recurrent neural network called long short-term memory to generate predictions of air quality concentration in Beijing based on air quality and weather data recorded from 2011 through 2015. Incorporating 9 features into the linear regression model yields an R^2 score of 0.242 with RMSE 80.69 μg/m^3. In addition, several LSTM models with varying numbers of layers, neurons, prior time steps, batch sizes, and features were tested, and LSTM models with 2 or 3 layers and 365 prior time steps provided the best fit to the data. The RMSE for the 3-layer LSTM is 61.37 μg/m^3.

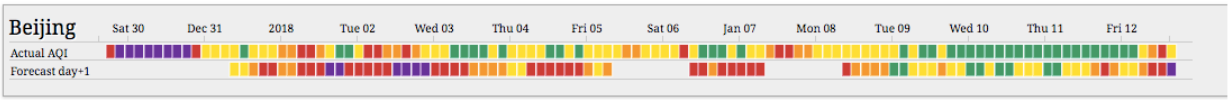

Figure 4.12: Actual and predicted AQI classification for Beijing from the World Air Quality Index Project for a two week span ending on Friday, December 12, 2018.

While the accuracy of the predictions from the LSTM model is comparable to forecasts currently available online, more improvement is needed to generate a forecast that the general public and the government can reliably depend on. Here are some suggestions to improve the LSTM model:

- More data. The features incorporated into the LSTM model can be expanded to include hourly carbon emissions data from local factories and residential areas since PM2.5 concentration peaks during the nighttime and early morning.

- More LSTM layers, neurons, and prior time steps, which means more computing power. A significant bottleneck in testing increasingly complex LSTM models is the computing power that is required to run the model. To explore LSTM models beyond 3 layers, more computing power will be needed to allow models to be run in a reasonable amount of time.

7 References

[1] Jia, H. *et al*. Peering into China's thick haze of air pollution. *American Chemical Society*, https://cen.acs.org/articles/95/i4/Peering-Chinas-thick-haze-air.html

[2] Stromberg, J. "What Does the Unbelievably Bad Air Quality in Beijing Do to the Human Body?" *Smithsonian.com*, www.smithsonianmag.com/science-nature/what-does-the-unbelievably-bad-air-quality-in-beijing-do-to-the-human-body-22655/

[3] Roberts, D. "Opinion: How the US Embassy Tweeted to Clear Beijing's Air." *Wired*, www.wired.com/2015/03/opinion-us-embassy-beijing-tweeted-clear-air/

[4] Tie, X. *et al*. Severe Pollution in China Amplified by Atmospheric Moisture. *Scientific Reports* **7** (2017).

[5] Graves, A. *et al*. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31** (2009).

[6] "Beijing Air Pollution: Real-time PM2.5 Air Quality Index (AQI)." *World Air Quality Index*, www.aqicn.org/

[7] "Urban Air." *Microsoft Corporation*, www.microsoft.com/en-us/research/project/urban-air/