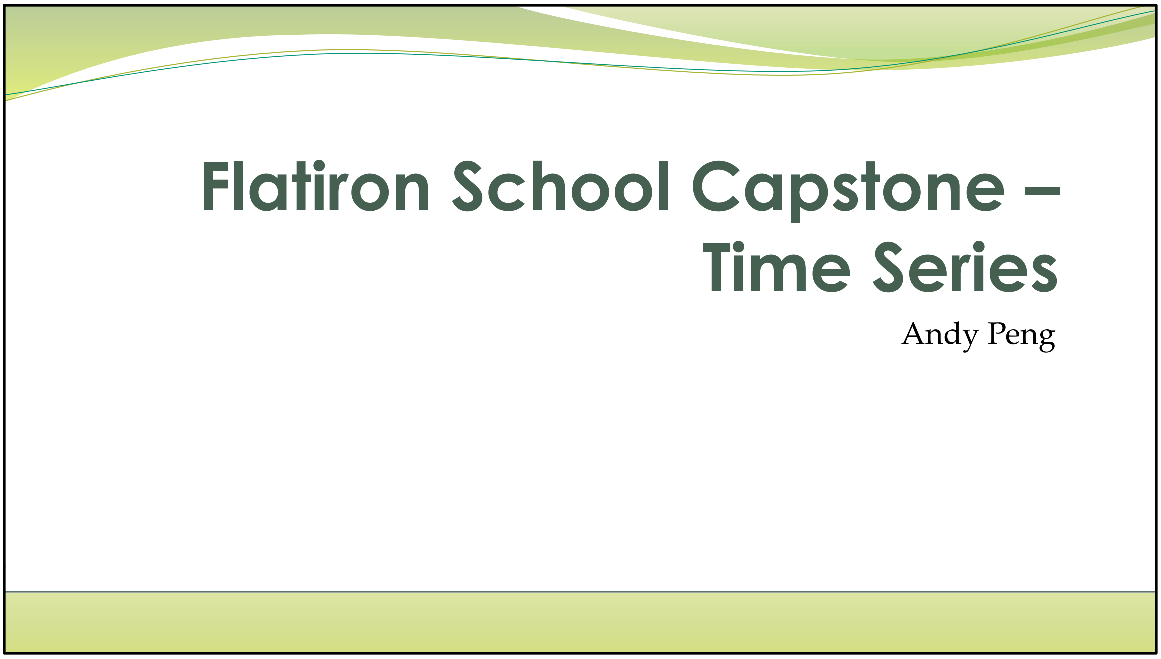
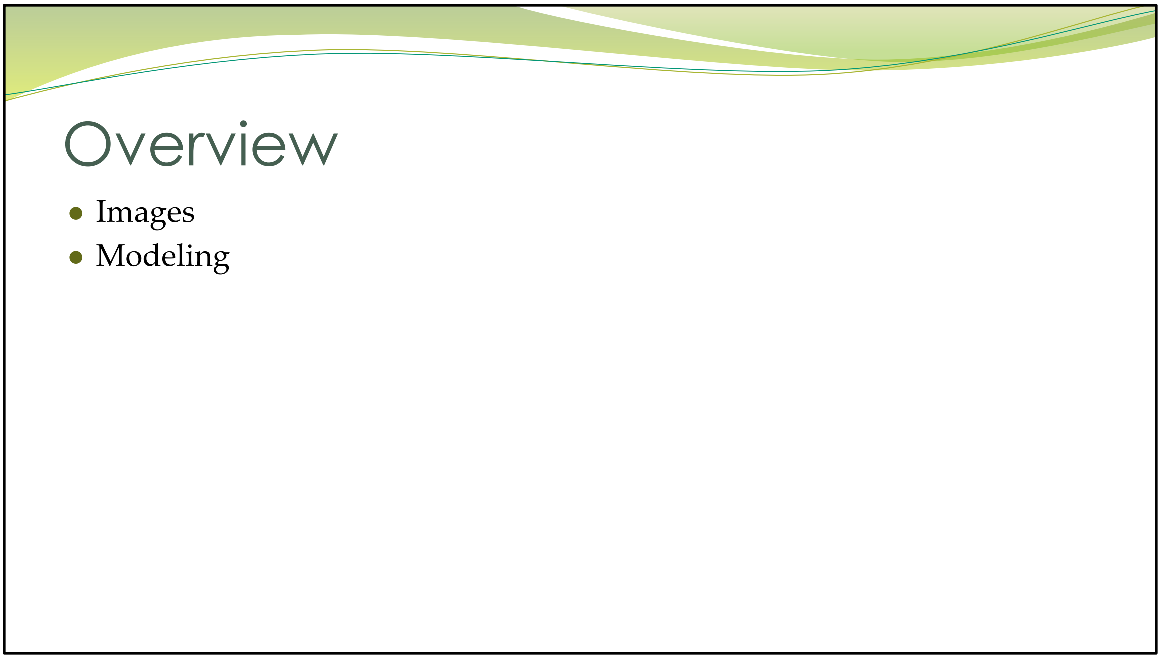


Flatiron School Capstone – Time Series

Andy Peng

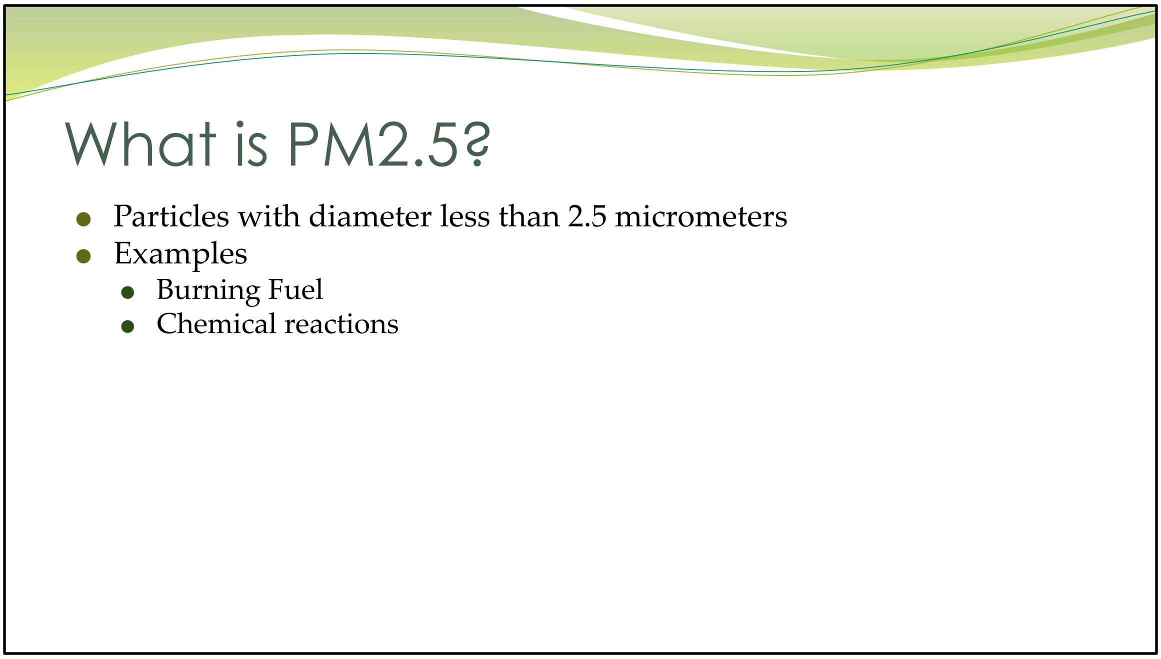
Hi my name is Andy Peng. Welcome to my presentation for Flatiron School Capstone – Time Series. For the extent of this project, our stakeholders are a company that focuses on predicting the air pollution in Beijing, China. We will mainly be focused on one of the air quality monitoring site located in Gucheng county.



Overview

- Images
- Modeling

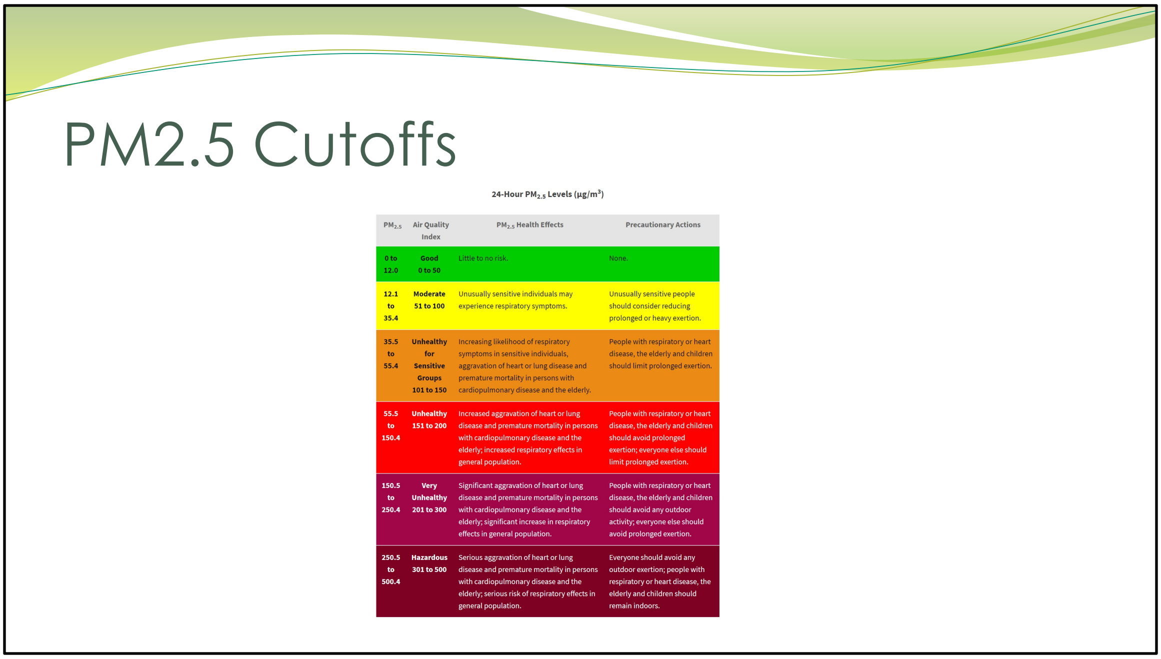
We will be displaying images from our data exploration and talk about the results from our modeling. But before we do that we will first discuss about the air quality index PM2.5.



What is PM2.5?

- Particles with diameter less than 2.5 micrometers
- Examples
  - Burning Fuel
  - Chemical reactions

What is the air quality index PM2.5? PM2.5 represents particles that have diameter less than 2.5 micrometers which is more than 10 times thinner than a human hair. These particles are formed as a result of burning fuel and chemical reactions that take place in the atmosphere. But what level of PM2.5 index is considered normal and what level is considered healthy?



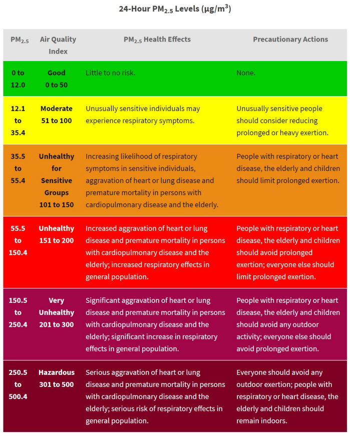
PM2.5 Cutoffs

24-Hour $PM_{2.5}$ Levels (µg/m^3)

$PM_{2.5}$	Air Quality Index	$PM_{2.5}$ Health Effects	Precautionary Actions
0 to 12.0	Good 0 to 50	Little to no risk.	None.
12.1 to 35.4	Moderate 51 to 100	Unusually sensitive individuals may experience respiratory symptoms.	Unusually sensitive people should consider reducing prolonged or heavy exertion.
35.5 to 55.4	Unhealthy for Sensitive Groups 101 to 150	Increasing likelihood of respiratory symptoms in sensitive individuals, aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly.	People with respiratory or heart disease, the elderly and children should limit prolonged exertion.
55.5 to 150.4	Unhealthy 151 to 200	Increased aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; increased respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid prolonged exertion; everyone else should limit prolonged exertion.
150.5 to 250.4	Very Unhealthy 201 to 300	Significant aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; significant increase in respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid any outdoor activity; everyone else should avoid prolonged exertion.
250.5 to 500.4	Hazardous 301 to 500	Serious aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; serious risk of respiratory effects in general population.	Everyone should avoid any outdoor exertion; people with respiratory or heart disease, the elderly and children should remain indoors.

This chart displays the cutoffs for our PM2.5 index.

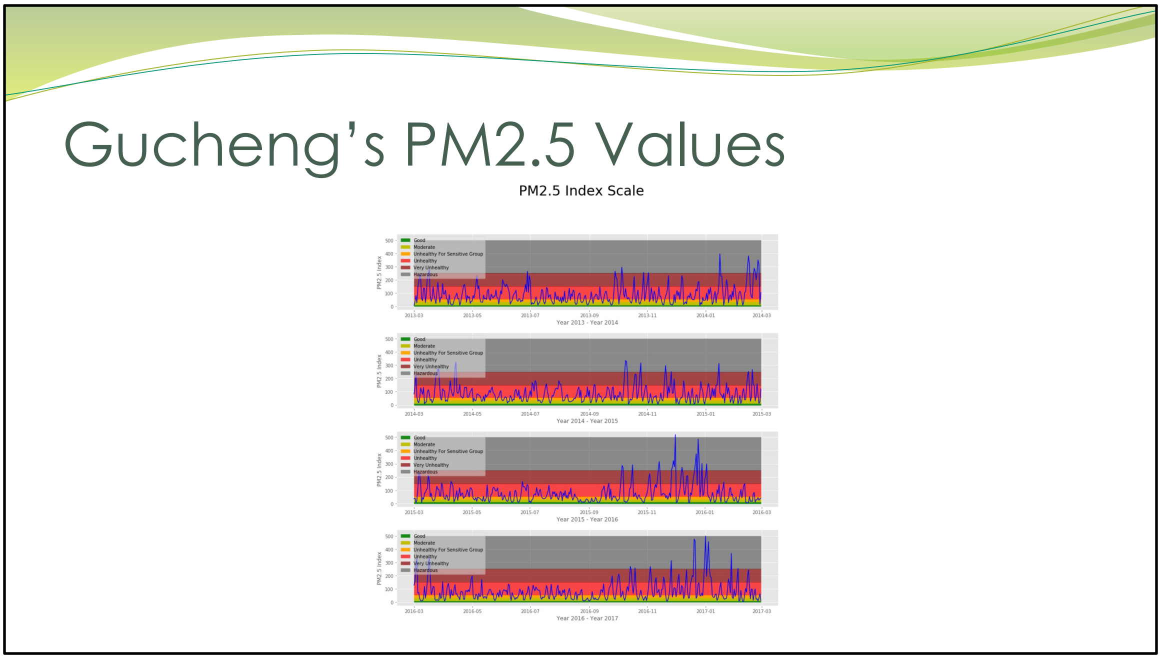
PM2.5 Cutoffs



24-Hour $PM_{2.5}$ Levels (µg/m^3)

$PM_{2.5}$	Air Quality Index	$PM_{2.5}$ Health Effects	Precautionary Actions
0 to 12.0	**Good 0 to 50**	Little to no risk.	None.
12.1 to 35.4	**Moderate 51 to 100**	Unusually sensitive individuals may experience respiratory symptoms.	Unusually sensitive people should consider reducing prolonged or heavy exertion.
35.5 to 55.4	**Unhealthy for Sensitive Groups 101 to 150**	Increasing likelihood of respiratory symptoms in sensitive individuals, aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly.	People with respiratory or heart disease, the elderly and children should limit prolonged exertion.
55.5 to 150.4	**Unhealthy 151 to 200**	Increased aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; increased respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid prolonged exertion; everyone else should limit prolonged exertion.
150.5 to 250.4	**Very Unhealthy 201 to 300**	Significant aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; significant increase in respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid any outdoor activity; everyone else should avoid prolonged exertion.
250.5 to 500.4	**Hazardous 301 to 500**	Serious aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; serious risk of respiratory effects in general population.	Everyone should avoid any outdoor exertion; people with respiratory or heart disease, the elderly and children should remain indoors.

These are the four levels that we care about because this is when PM2.5 levels become harmful for us. It starts with recommending individuals with certain respiratory or heart disease to limit their prolonged exertion to everyone avoiding outdoor exertion with certain people remaining indoors for safety reasons. But how is Gucheng’s PM2.5 values?



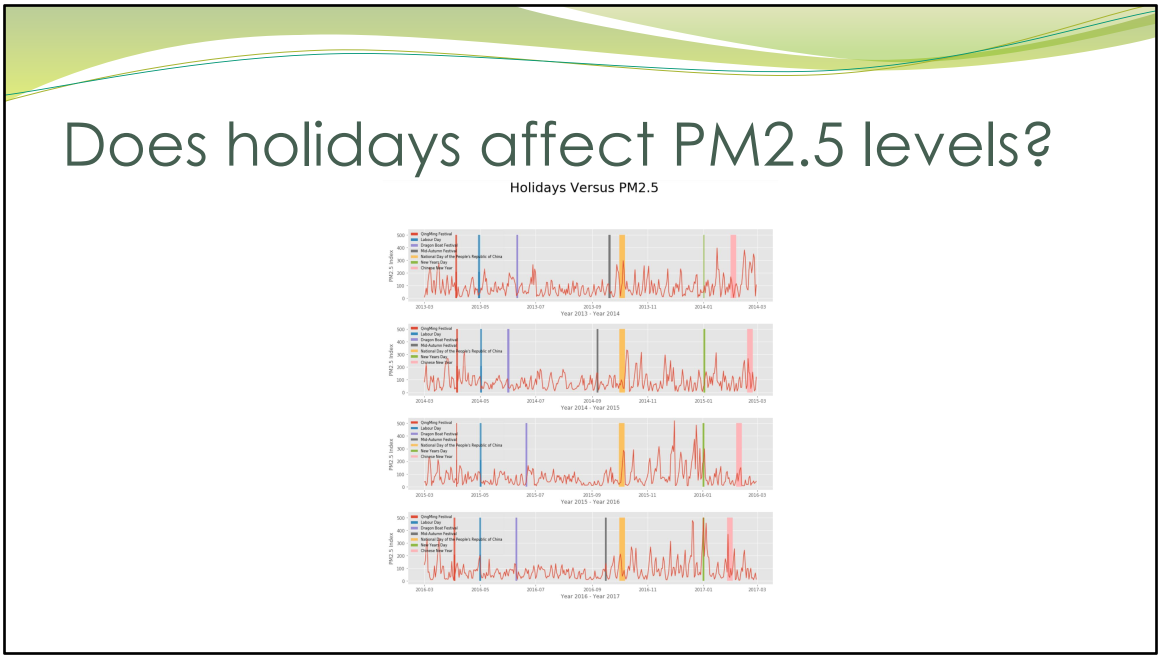
Gucheng's PM2.5 Values

PM2.5 Index Scale

This is a graph of the PM2.5 values measured at Gucheng's air quality monitoring site. The time range of this data ranges from March 01, 2013 to February 28, 2017. As you can see in this picture majority of the days in each year have values in the unhealthy for sensitive groups to very unhealthy region. PM2.5 value increases when there is burning fuel or chemical reactions happening in the atmosphere. What do you think would lead to a result of this in Gucheng?

Chinese New Year

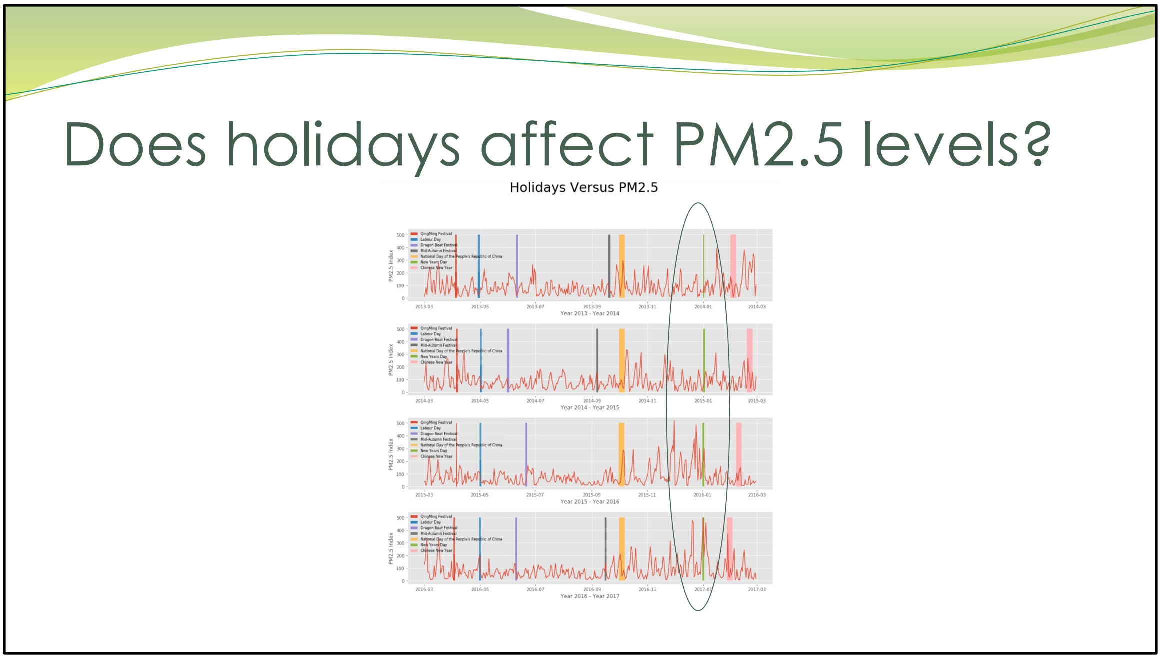
It would be Chinese New Year. On Chinese New Year, people would be driving their cars to visit families and lighting fireworks to celebrate the holiday.



Does holidays affect PM2.5 levels?

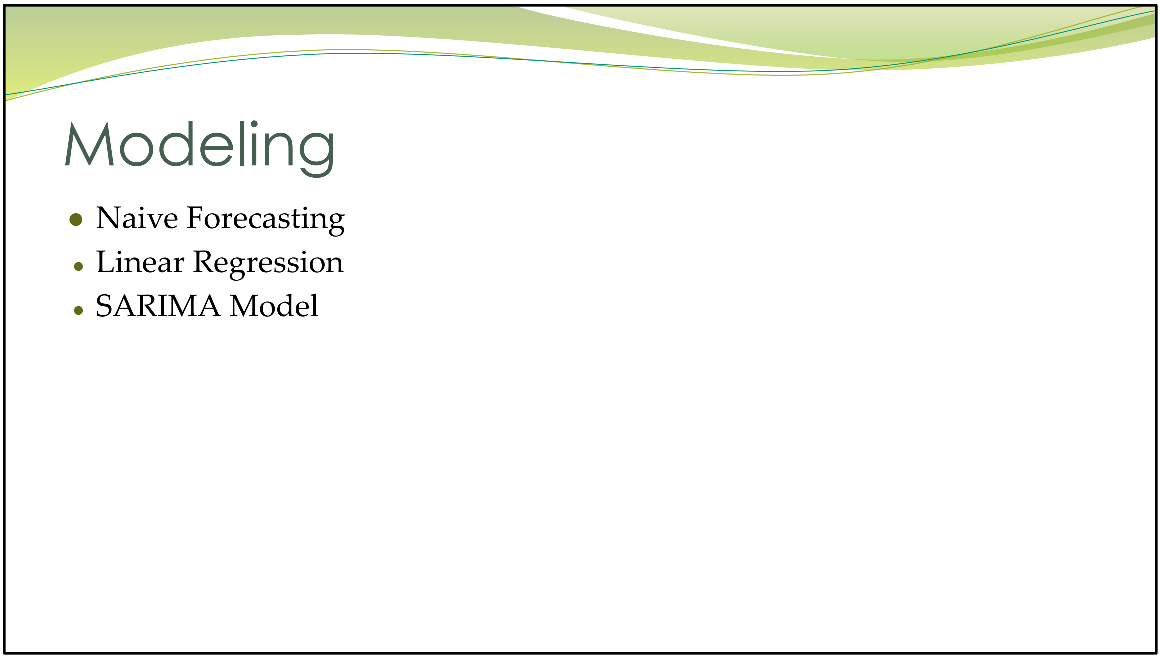
Holidays Versus PM2.5

Does holidays affect PM2.5 levels? We graphed the daily data and the different holiday ranges to see if there is a relationship between public holidays in China and PM2.5 levels.



Does holidays affect PM2.5 levels?

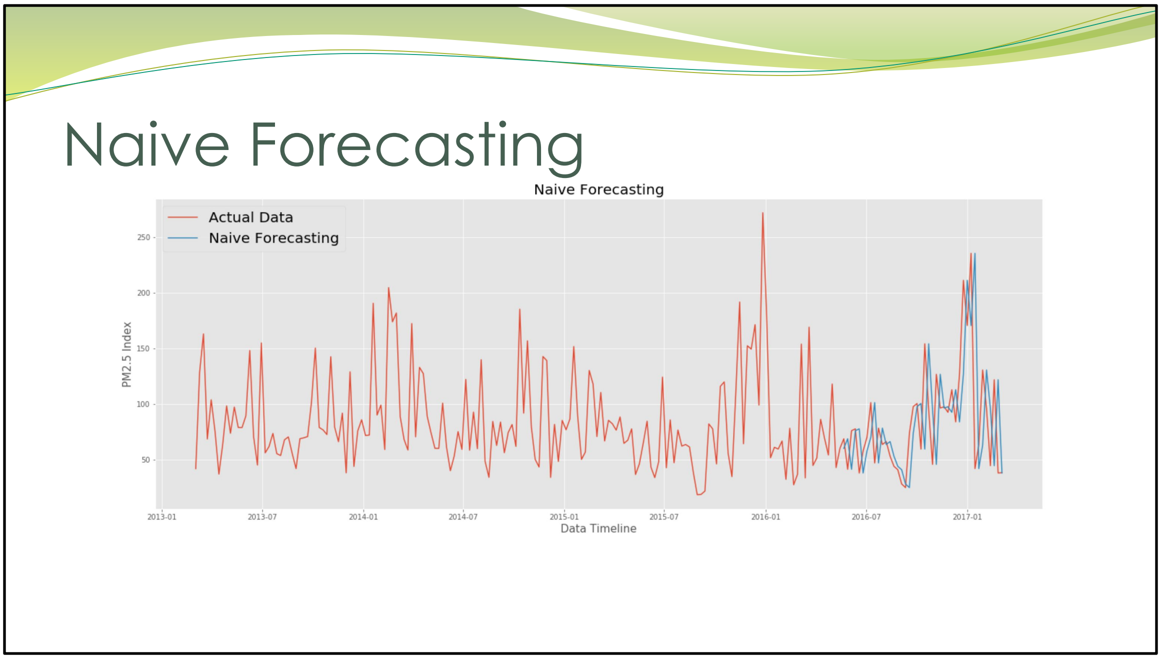
Based off this picture we can see a slight increase during and after the New Years holiday. Other than that holiday, for all the other public holidays, we can't really tell if there is any increase or decrease in PM2.5 levels. Now let's move onto the modeling part of our presentation.



Modeling

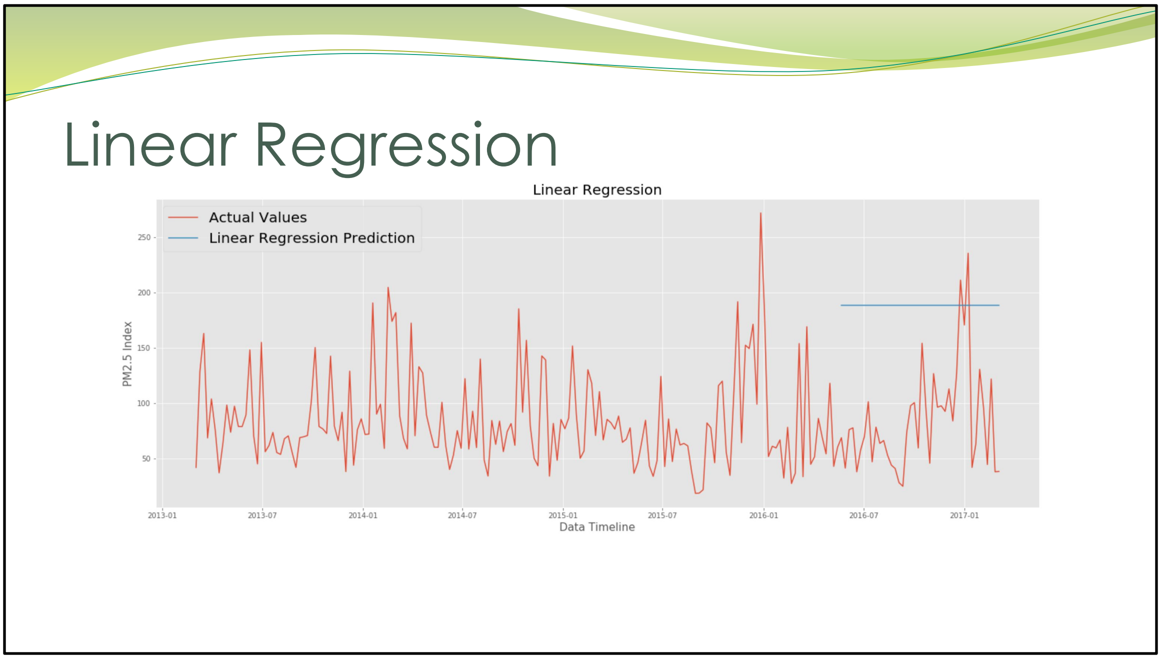
- Naive Forecasting
- Linear Regression
- SARIMA Model

For modeling we used Naive Forecasting, linear regression and SARIMA model. And here are the results.



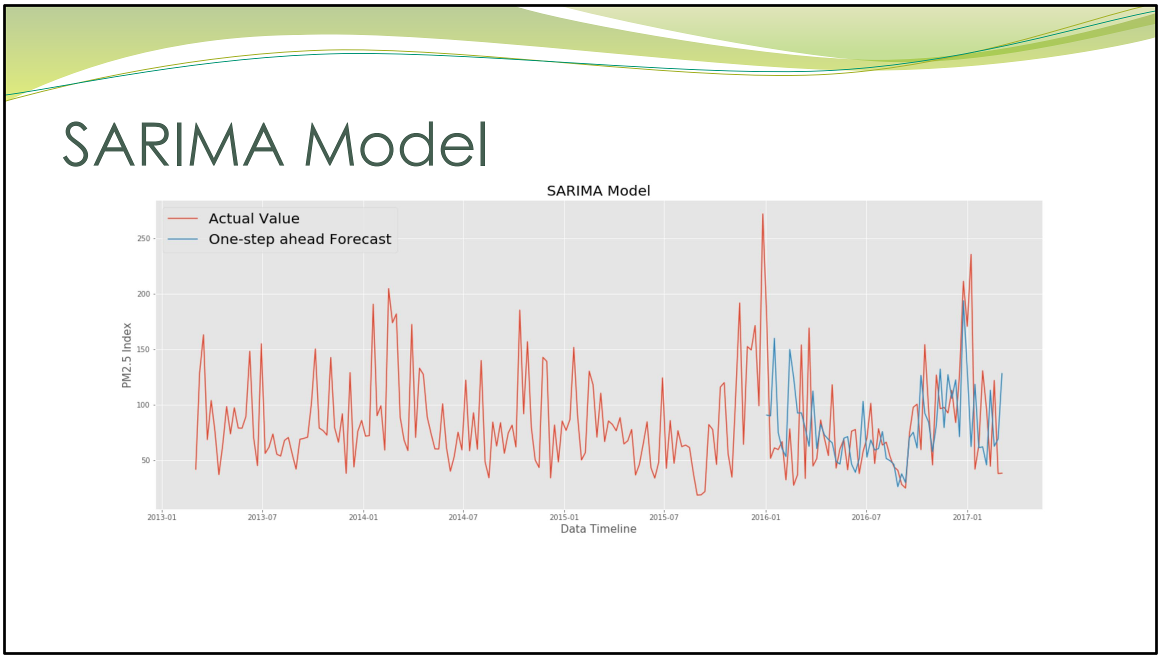
Naive Forecasting

Using naive forecasting, we were able to predict the PM2.5 values within a 51.65 range. Here is a graph demonstrating how our prediction would look like compare to our data.



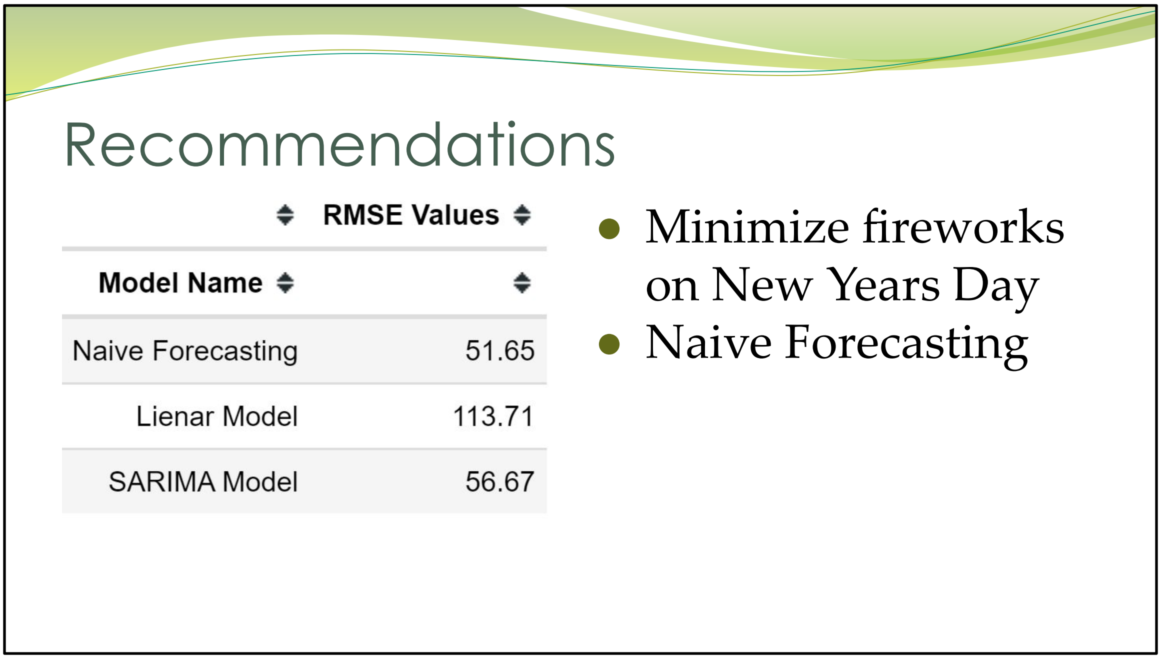
Linear Regression

Using linear regression, we were able to predict the PM2.5 values within a 113.71 range. The graph here demonstrates on well our prediction for linear regression is doing compare to the actual data. As you can see the linear regression prediction is a horizontal line and therefore would perform worser than the naive forecasting.



SARIMA Model

Using SARIMA Model, we were able to predict the PM2.5 values within a 56.67 range. Here is a graph demonstrating how our prediction would look like compare to our data. As you can see this model is better than the linear regression model, but a performs slightly worser than the naive forecasting.

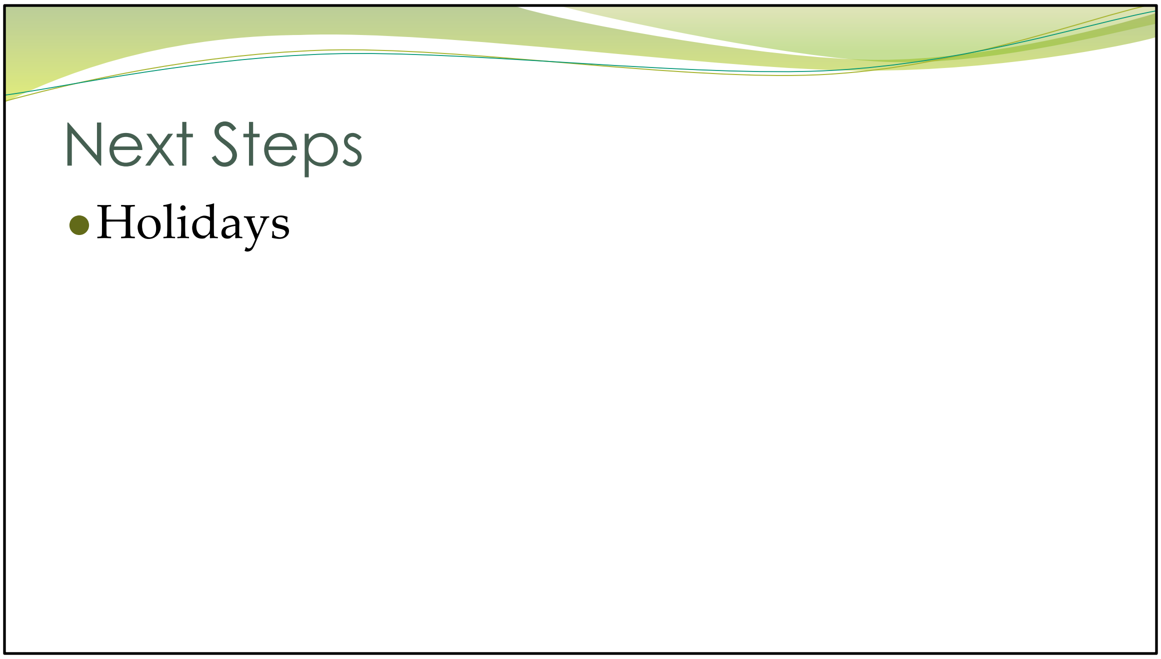


Recommendations

	RMSE Values
Model Name	
Naive Forecasting	51.65
Lienar Model	113.71
SARIMA Model	56.67

- Minimize fireworks on New Years Day
- Naive Forecasting

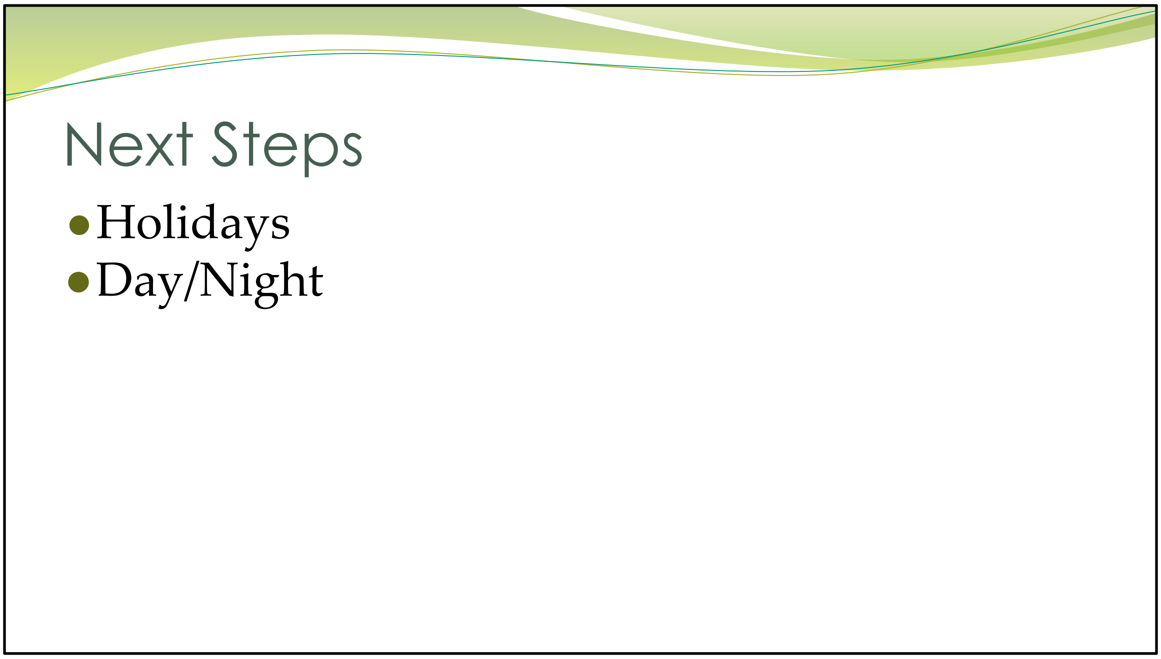
To summarize all that we just talked about, we found out that there seems to be an increase in PM2.5 levels during and after New Years day. We could minimize the amount of fireworks lit on New Year's day so that the PM2.5 levels wouldn’t increase as much. For modeling, I would use Naive Forecasting because it yielded the best result in predicting PM2.5 levels.



Next Steps

- Holidays

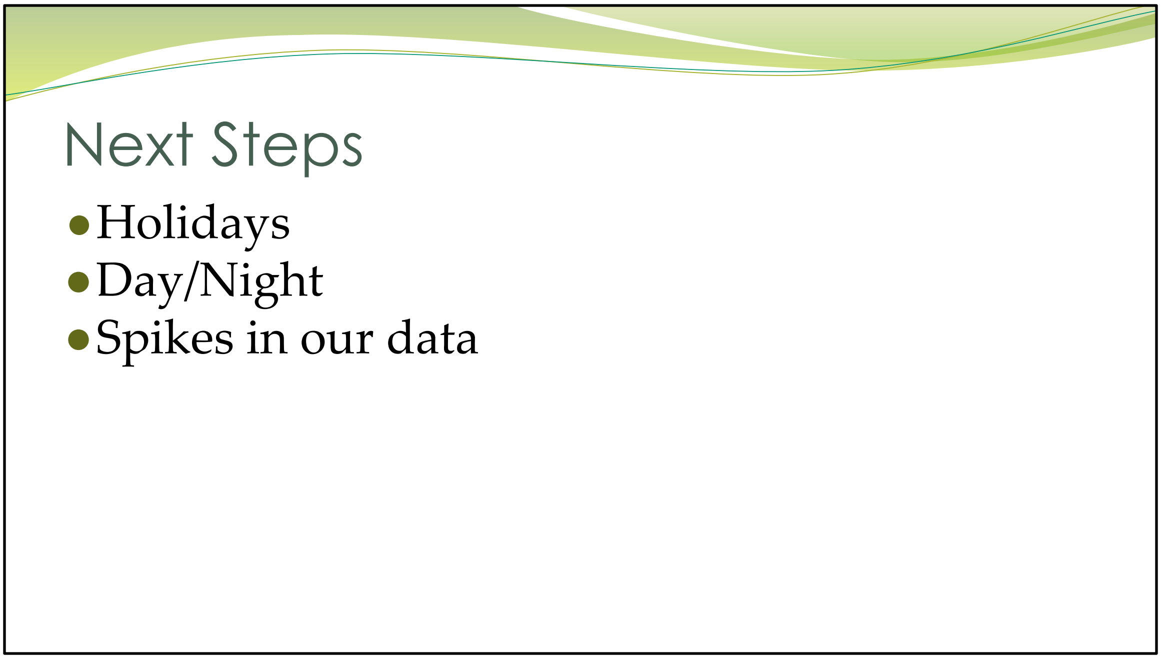
For our next steps we can investigate more on holiday effects on PM2.5. I only graphed public holidays in China on the graph we seen earlier, but there might be certain holidays that appear specifically in Gucheng that we do not know of.



Next Steps

- Holidays
- Day/Night

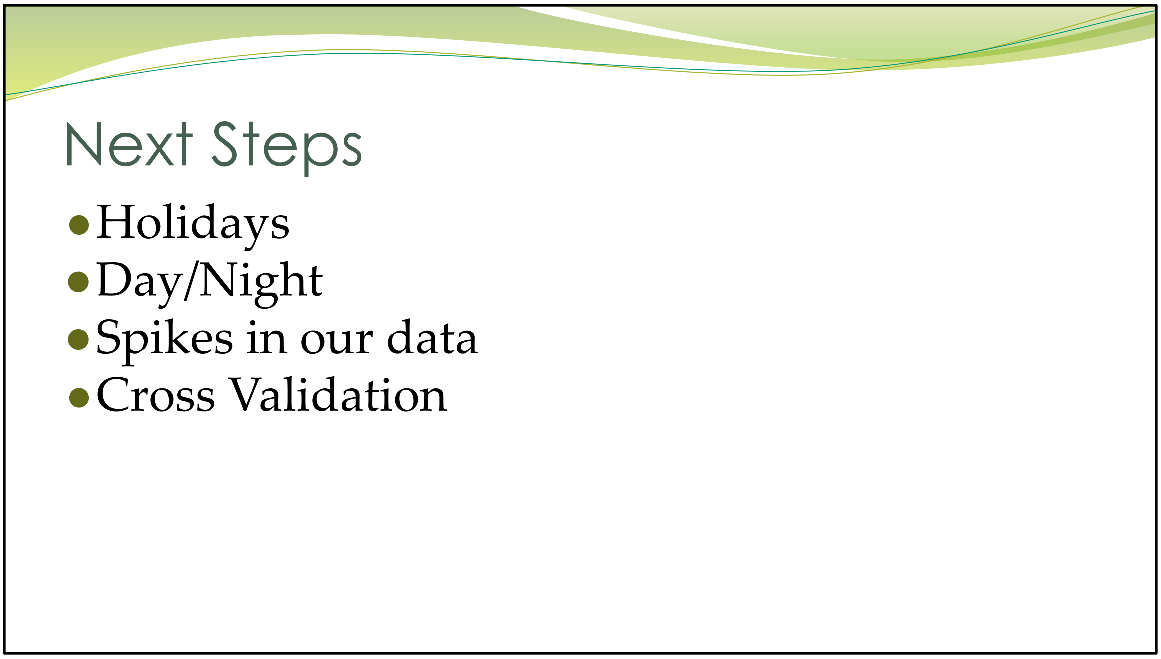
We could also investigate whether there is a difference in PM2.5 values during the day time compare to night time. Would there be a higher burning of fuels from cars during the day when people have work or would there be a higher burning of fuels from night workers at night?



Next Steps

- Holidays
- Day/Night
- Spikes in our data

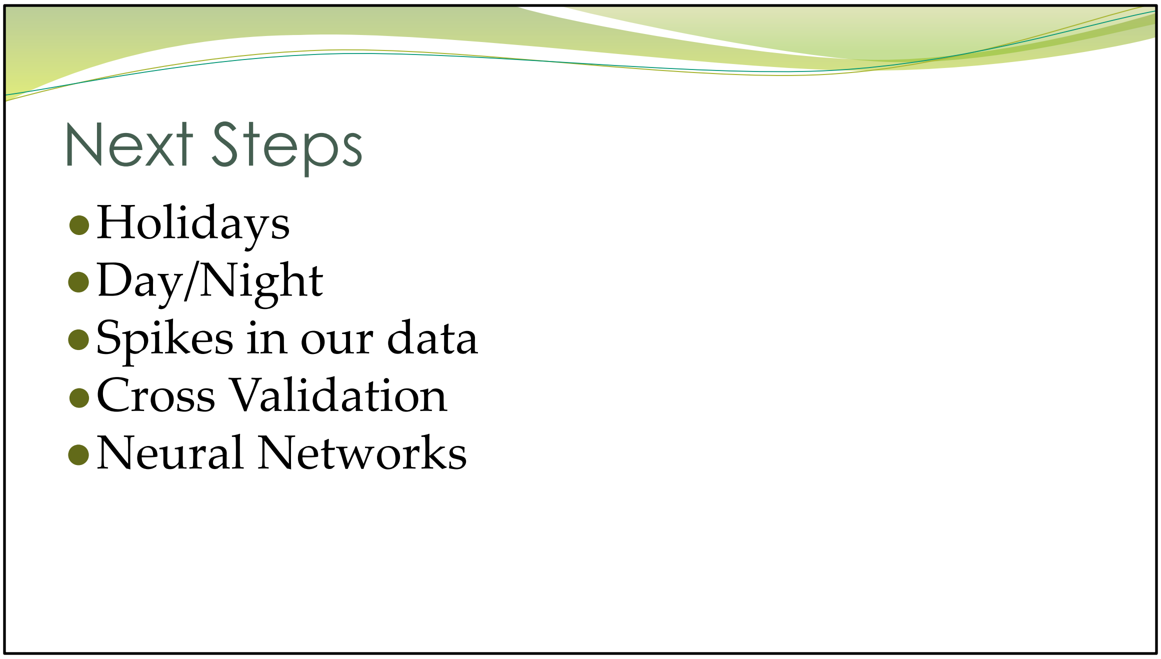
Not only can we explore day and night PM2.5 values, we could also explore why there are certain huge spikes in our data. Could there be a certain event that happen around the air quality monitoring site that led to this huge spike?



Next Steps

- Holidays
- Day/Night
- Spikes in our data
- Cross Validation

Our model can also be improve by performing cross validation on the data set.



Next Steps

- Holidays
- Day/Night
- Spikes in our data
- Cross Validation
- Neural Networks

We could also try out other time series modeling techniques that we didn’t mention in this presentation such as building a Neural Network to predict PM2.5 values.

Next Steps

- Holidays
- Day/Night
- Spikes in our data
- Cross Validation
- Neural Networks
- Gather More Data

Lastly our models can further be improved by gathering data because our data only goes up to February 28, 2017. But, we are missing a chunk of data from March 1, 2017 to present day. However, all of these next steps require more money and time to collect more data and to train our model.

Thank You

Thank You