Air Quality Prediction for Major Cities of China

Jing Zhao
Capstone Project
Springboard Data Science
Oct, 2018

Air pollution Problem in China

- **Appears as thick haze and smog**
  - PM2.5 (ultrafine particles < 2.5 microns in diameter) accountable for > 2/3 of the severely polluted days
  - Due to rapid industrialization and high energy consumption

- **Raises intense public concerns and media attentions**
  - Emergence alerts and school shutdowns
  - Headlines of national and worldwide news

- **Puts a threat on human health**
  - Linked to increased occurrence of various respiratory diseases (Asthma, Bronchitis, Lung Disease)
  - Number of hospital admission due to respiratory problems dramatically increased

- **Status**
  - PM2.5 monitoring and reporting are introduced since 2013, and extended to 388 cities by 2015
  - Problem improves over the past five years, but still significant most days of the year

The Approach

- **Identify patterns of PM2.5**
  - Trends (time-dependent, city-dependent, etc.)
  - Correlations to weather conditions (temperature, pressure, wind, precipitation, etc.)

- **Predict future PM2.5**
  - Statistical models using StatsModels
  - Machine learning models using Scikit-Learn
  - Feature engineering to improve models

The Client

A PM2.5 predictive model is beneficial to people at all levels to reduce exposure to extremely polluted air

- **Government**
  - Provide outdoor activity guideline for adults and children
  - Announce appropriate warnings in advance

- **Local organization**
  - Plan outdoor events
  - Schedule emergency shutdowns

- **Individual citizen**
  - Schedule commute choices
  - Plan daily outdoor activities
  - Wear personal protective equipment

The Dataset

- **Source:** Two public datasets hosted by UCI Machine Learning Repository

- **Link:** https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities
  https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data

- **Dataset #1**
  Hourly PM2.5 readings for five Chinese cities (Beijing, Chengdu, Guangzhou, Shanghai, Shenyang)
  Meteorological readings
  (dew point, temperature, humidity, pressure, wind direction and speed, precipitation)
  Period of 2010-2015

- **Dataset #2**
  Similar to dataset #1, but for city Beijing alone during the period of 2010-2014.
  Precipitation is specified as ‘snow’ and ‘rain’

Data Wrangling

- **Introduce new columns**
  ‘date_time’: datetime format by parsing time-related columns 'year', 'month', 'day', 'hour'
  ‘pm_average’: the average pm2.5 level computed from multiple stations of the same city
  ‘ws’: hourly wind speed derived from the cumulated wind speed (‘iws’) in the raw data to eliminate artifacts

- **Divide into three subsets for different purposes**
  pm_clean: main dataset containing measurement data from all five cities, use ’pm_average’ as PM2.5 levels
  pm_stations: supplementary dataset containing station-specific individual PM2.5 readings, only serving the purpose of validating measurement consistency among stations in each city
  pm_sr: side dataset containing information on cumulated hours of snows and rains for city 'beijing’, only serving the purpose of comparing the impact of snow versus rain on PM2.5

- **Missing & Outlier**
  Missing: 36% NaNs (93% of which is due to the missing pm2.5 values), dropped
  Outlier: 0.025%, replaced when appropriate or dropped

Cleaned Data

DataFrame : pm_clean Index : RangeIndex

Column name	Data Type	Unit	Description
year	int64	N/A	year of observation in current row
month	int64	N/A	month of observation in current row
day	int64	N/A	day of observation in current row
hour	int64	N/A	hour of observation in current row
season	int64	N/A	season of observation in current row
date_time	datetime64	N/A	datetime format of observation in current row
dewp	float64	˚C	dew point
humi	float64	%	humidity
pres	float64	hPa	pressure
temp	float64	˚C	temperature
iws	float64	m/s	cumulated wind speed
ws	float64	m/s	hourly wind speed
precipitation	float64	mm	hourly precipitation
iprec	float64	mm	cumulated precipitation
pm_average	float64	ug/m^3	average PM2.5 concentration of nearby stations
cbwd	object (string)	N/A	combined wind direction
city	object (string)	N/A	associated city of observation in current row

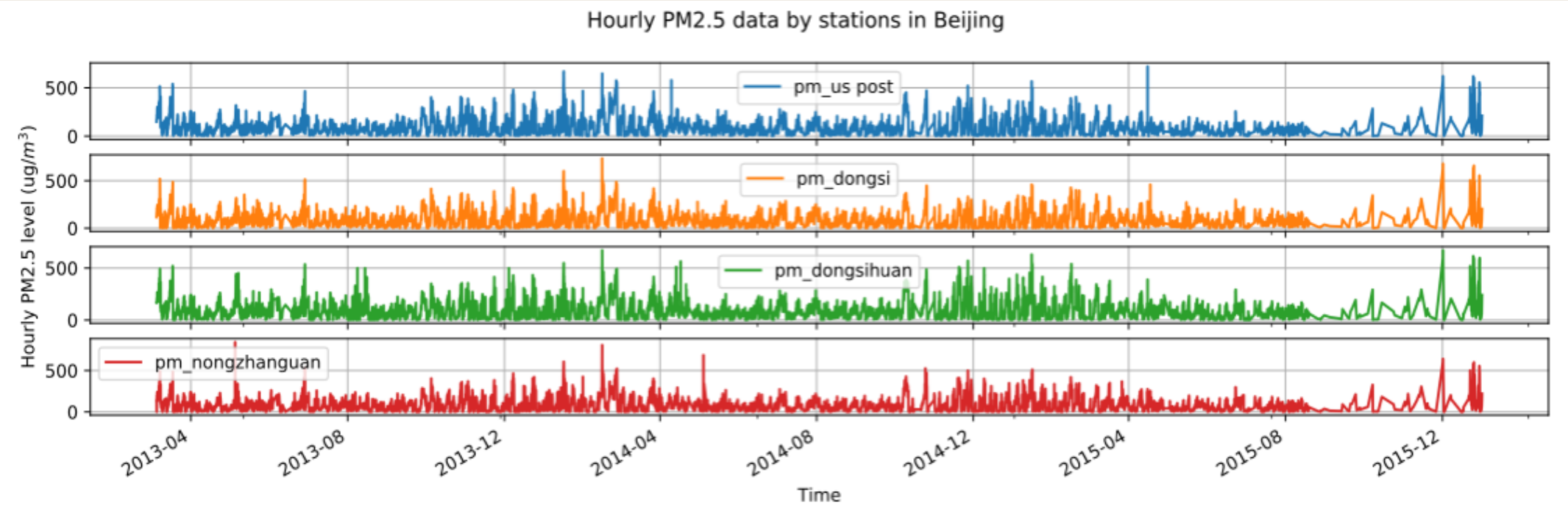
DataFrame : pm_stations Index : DatetimeIndex

Column name	Data Type	Unit	Description
year	int64	N/A	year of observation in current row
month	int64	N/A	month of observation in current row
day	int64	N/A	day of observation in current row
hour	int64	N/A	hour of observation in current row
season	int64	N/A	season of observation in current row
city	object (string)	N/A	associated city of observation in current row
pm_us post	float64	ug/m^3	PM2.5 concentration measured at local US Post
pm_dongsi	float64	ug/m^3	PM2.5 concentration measured at Dongsi station in Beijing
pm_dongsihuan	float64	ug/m^3	PM2.5 concentration measured at Dongsihuan station in Beijing
pm_nongzhanguan	float64	ug/m^3	PM2.5 concentration measured at Nongzhanguan station in Beijing
pm_jingan	float64	ug/m^3	PM2.5 concentration measured at Jingan station in Shanghai
pm_xuhui	float64	ug/m^3	PM2.5 concentration measured at Xuhui station in Shanghai
pm_city station	float64	ug/m^3	PM2.5 concentration measured at City station in Guangzhou
pm_5th middle school	float64	ug/m^3	PM2.5 concentration measured at 5th Middle School station in Guangzhou
pm_caotangsi	float64	ug/m^3	PM2.5 concentration measured at Caotangsi station in Chengdu
pm_shahepu	float64	ug/m^3	PM2.5 concentration measured at Shahepu station in Chengdu
pm_taiyuanjie	float64	ug/m^3	PM2.5 concentration measured at Taiyuanjie station in Shenyang
pm_xiaoheyan	float64	ug/m^3	PM2.5 concentration measured at Xiaoheyan station in Shenyang

DataFrame : pm_sr. Index : RangeIndex

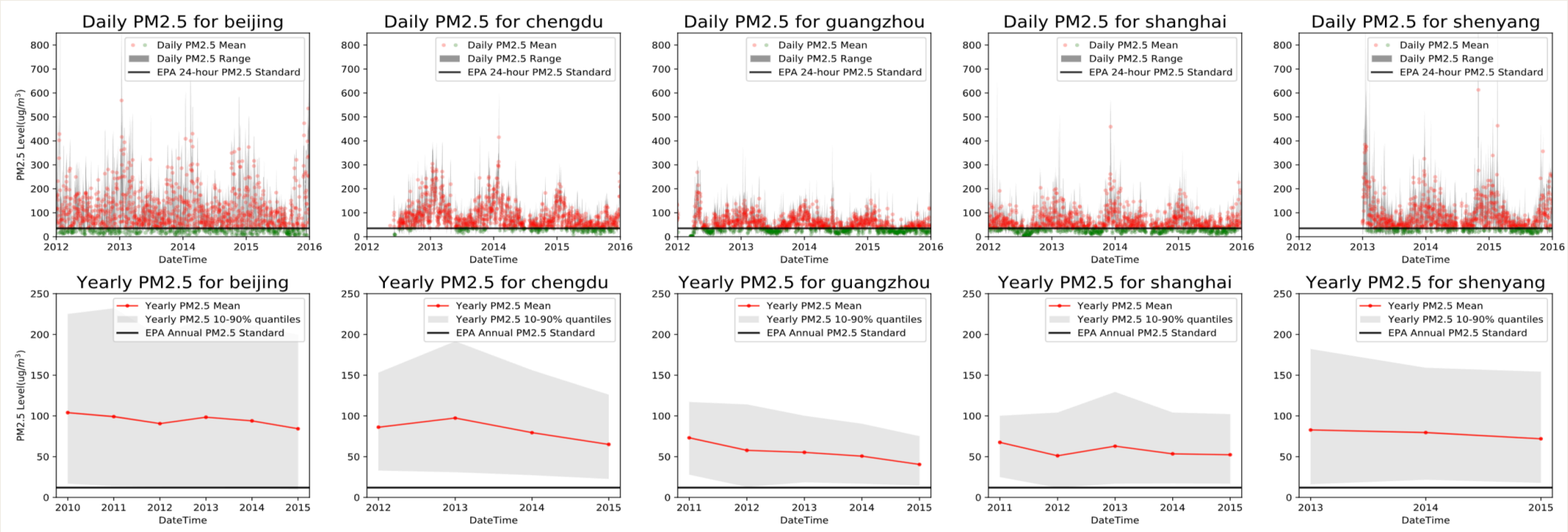
Column name	Data Type	Unit	Description
year	int64	N/A	year of observation in current row
month	int64	N/A	month of observation in current row
day	int64	N/A	day of observation in current row
hour	int64	N/A	hour of observation in current row
dewp	float64	˚C	dew point
pres	float64	hPa	pressure
temp	float64	˚C	temperature
iws	float64	m/s	cumulated wind speed
is	float64	hour	cumulated hours of snow
ir	float64	hour	cumulated hours of rain
pm2.5	float64	ug/m^3	PM2.5 concentration
cbwd	object (string)	N/A	combined wind direction
city	object (string)	N/A	Beijing

EDA - PM2.5 consistency across stations



- PM2.5 measured at nearby stations are consistent with each other
- The averaged PM2.5 is more representative and used hereafter

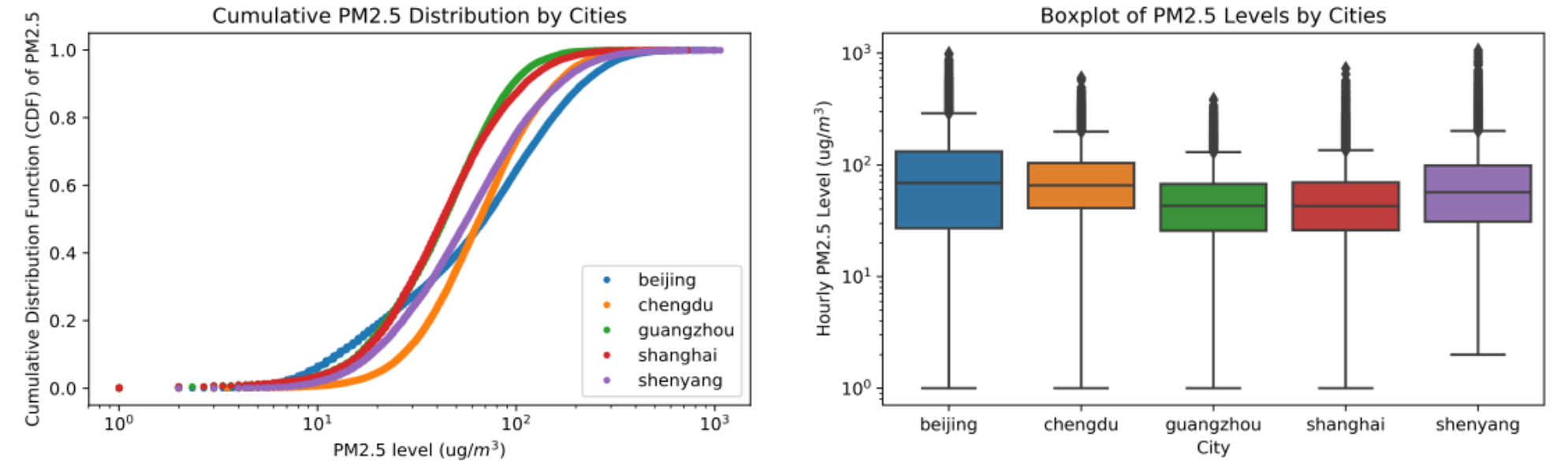
EDA - how bad is it ?



All five cities fail to meet the EPA standards:

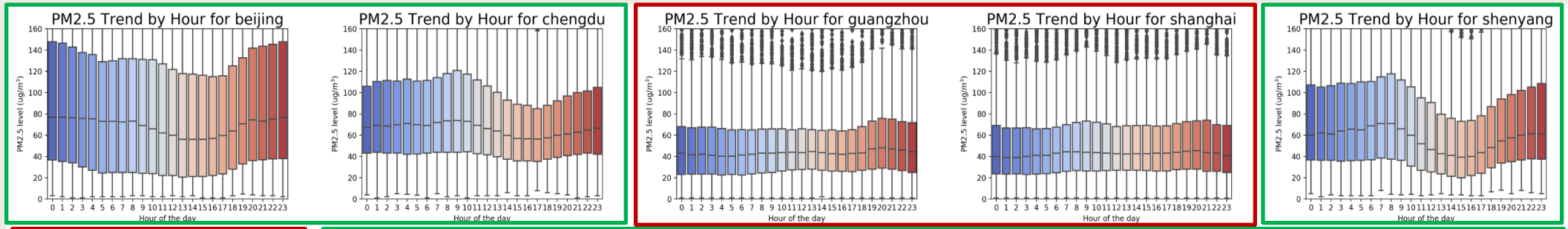
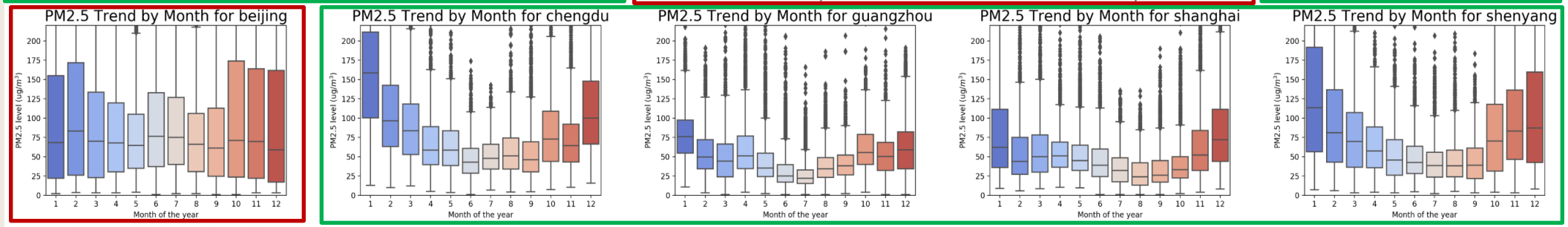
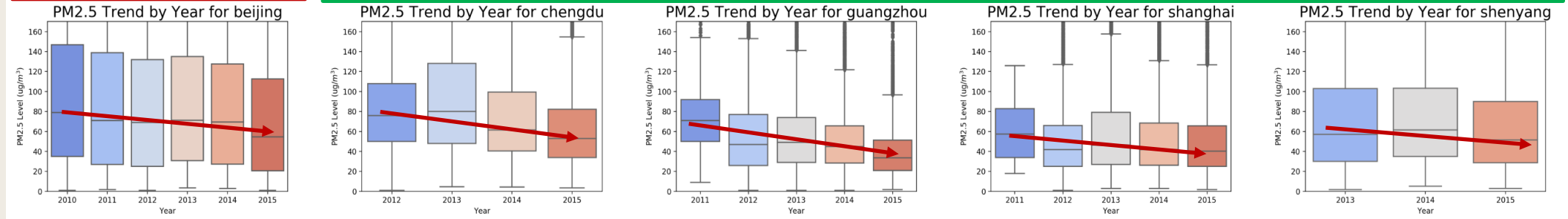
- Most daily average PM2.5 exceeds EPA recommendation for 24-hour PM2.5 level (35 μg/m^3)
- None of the annual PM2.5 levels meet EPA guidance for annual PM2.5 level (12 μg/m^3)

EDA – Cross-city Comparison

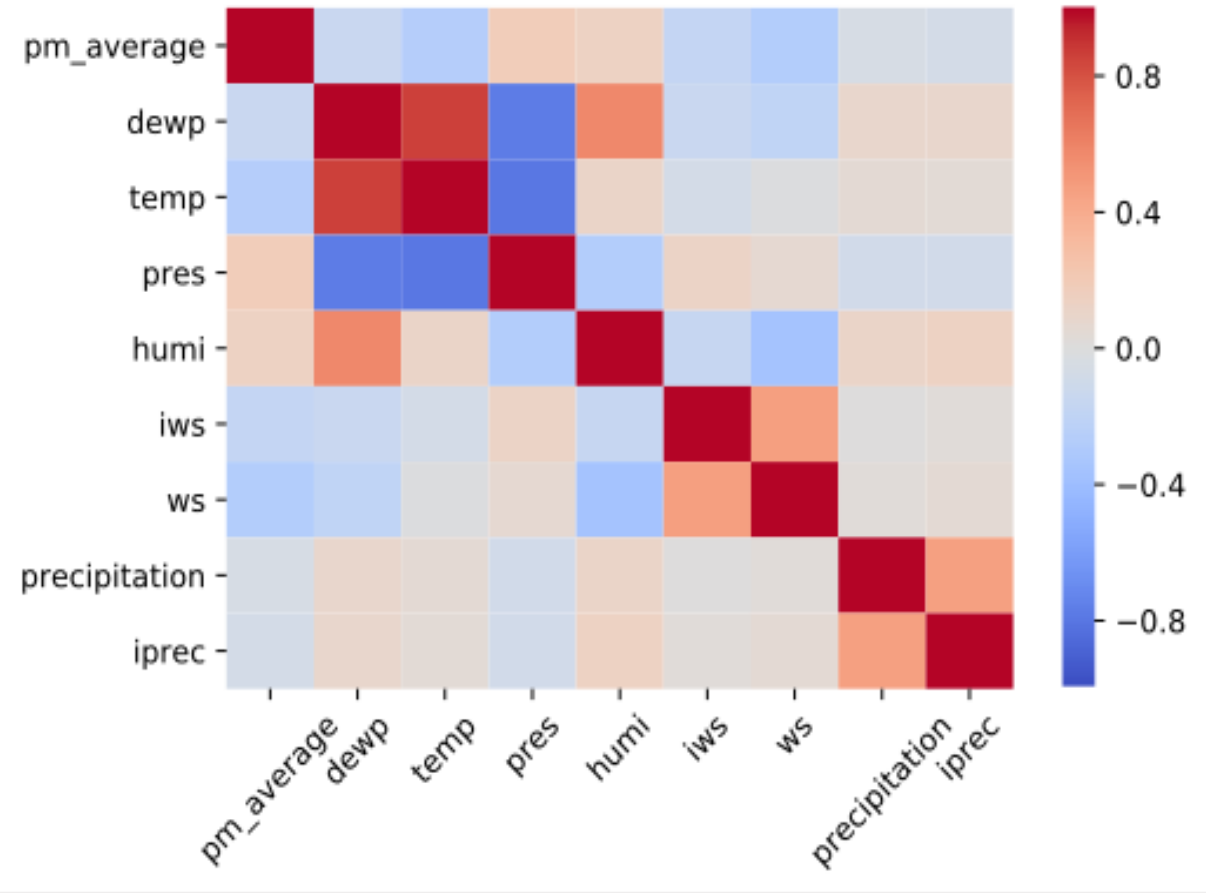


- Air quality ranking (best → worst): Guangzhou, Shanghai, Chengdu/Shenyang, Beijing
- Five cities all have distinct non-overlapping PM2.5 spikes, as a result of unique local environment
- Five cities share similar time trends:
  higher PM2.5 with larger fluctuations in winter, and lower PM2.5 with smaller fluctuations in summers

EDA - Trends over Time

	Beijing	Chengdu	Guangzhou	Shanghai	Shenyang
Hourly:					
Monthly:					
Yearly:					

EDA – Correlation Matrix with Weather



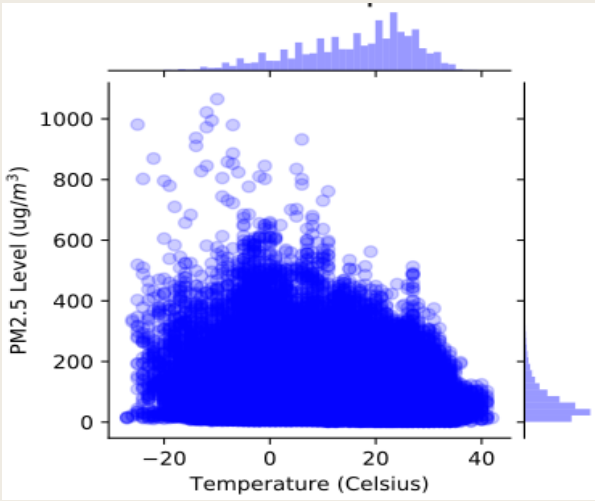
Correlation with PM2.5

temperature: negative
dew point: negative
wind speed: negative
precipitation: negative

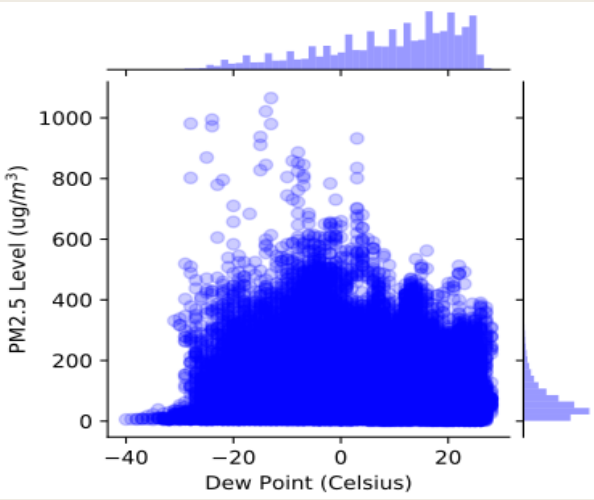
pressure: positive
humidity: positive

EDA – A Closer Look at Correlation
(Temperature, Dew point, Pressure, Humidity)

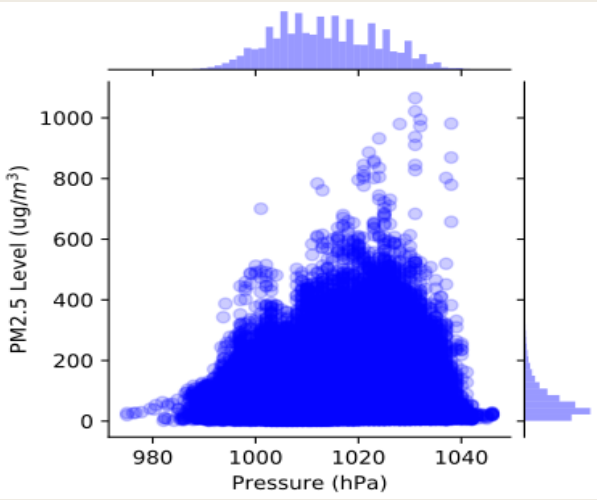
PM2.5 vs. Temperature



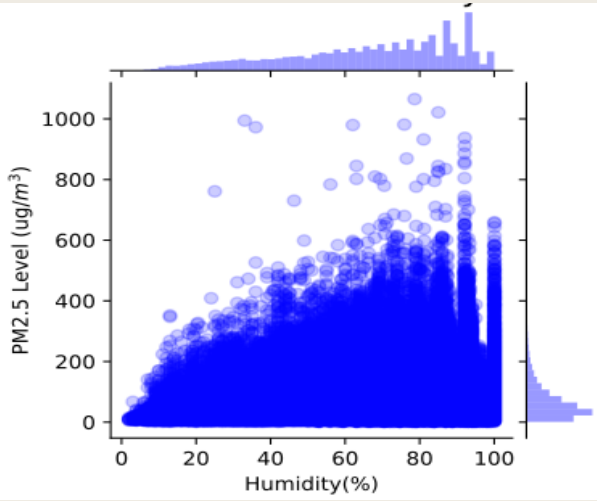
PM2.5 vs. Dew Point



PM2.5 vs. Pressure



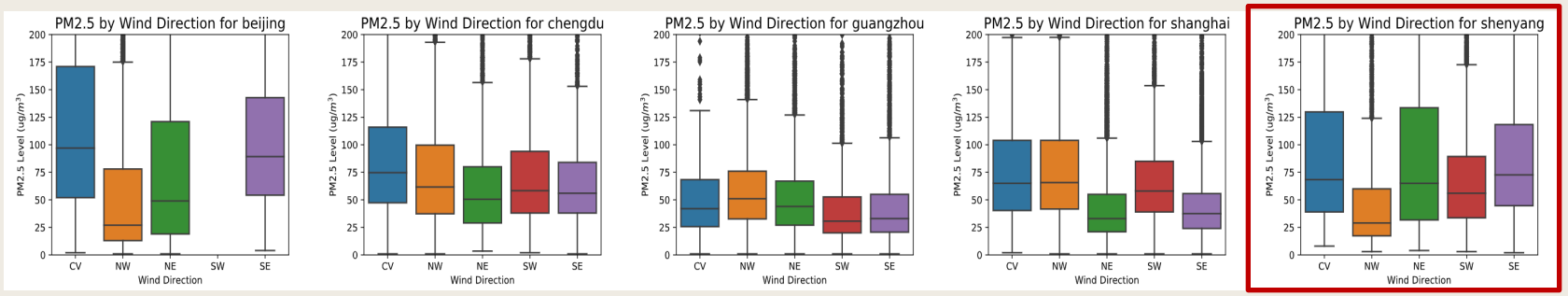
PM2.5 vs. Humidity



Temperature: negative correlation, higher PM2.5 mostly associated with T < 10˚C (cold weather)
Dew point: negative correlation, higher PM2.5 mostly associated with T < 5˚C
Pressure: positive correlation, higher PM2.5 mostly associated with higher atmospheric pressures.
Humidity: weak positive correlation (statistically significant), higher PM2.5 more likely to occur at higher humidity

EDA – A Closer Look at Correlation (Wind)

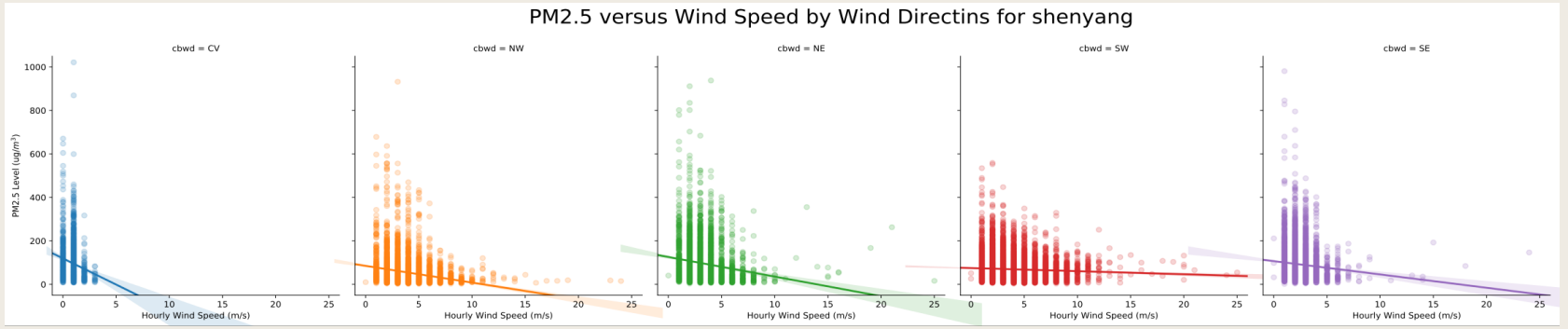
Wind Direction[1]



[1] northwest(NW), northeast(NE), southeast(SE), southwest (SW), static wind (CV)

- Strong dependency on direction
- Dependence differs by city:
  location, neighbor environment

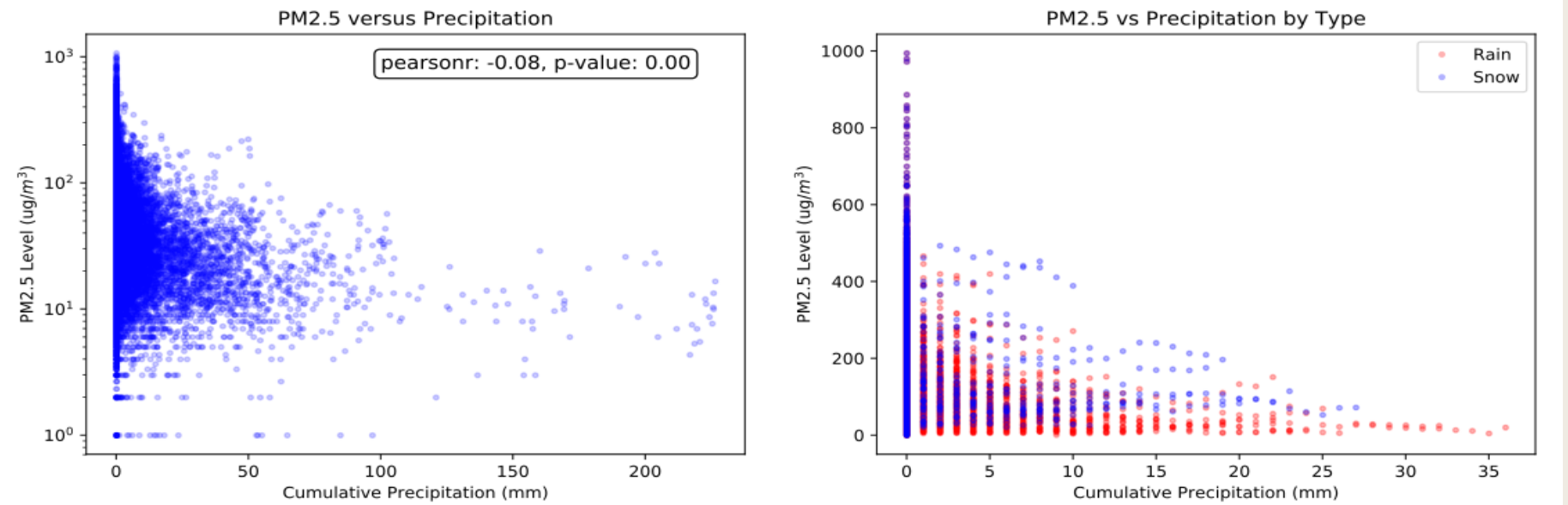
Wind Speed



low speed | higher speed | stronger correlation | higher speed | stronger correlation

- Higher speed reduces PM2.5:
  all directions, all cities
- Direction with higher speed:
  smaller PM2.5 median value
  narrows inter-quantile range
- Direction with stronger correlation:
  broader inter-quantile range

EDA – A Closer Look at Correlation (Precipitation)



- Precipitation reduces high PM2.5 levels regardless of the type
- Rain is slightly more effective at reducing PM2.5 than snow
- Large or extended periods of precipitation (high cumulative precipitation)
  very effective at eliminating high PM2.5 (>100)
  only very limited impact for low PM2.5 (< 50)

Modeling – Time Series by ARIMA/SARIMA

Time series PM2.5 by itself (without the weather features) is first modeled using ARIMA/SARIMA from StatsModels
Training set: year 2010-2014; Test set: year 2015; Optimization: grid search

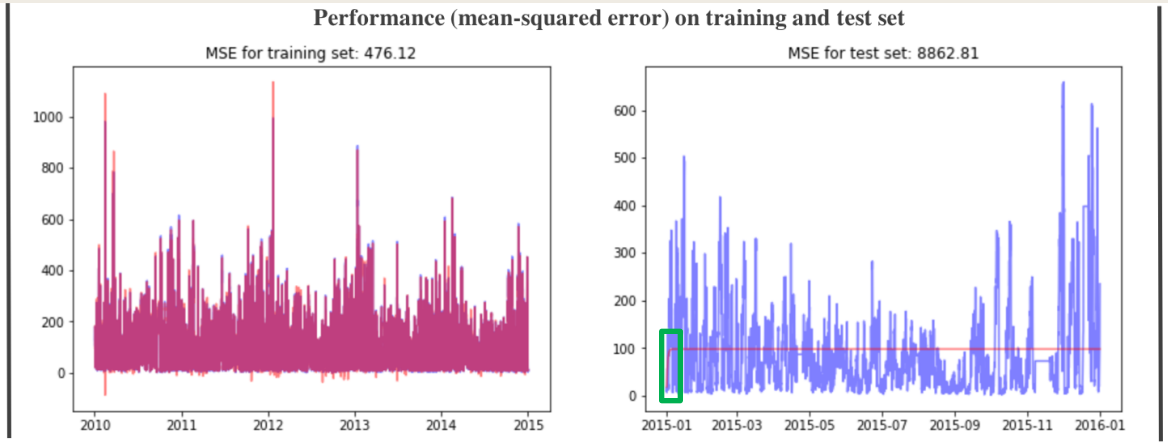
ARMA Model Results

Dep. Variable:	pm_hourly_beijing	No. Observations:	43801
Model:	ARMA(2, 0)	Log Likelihood	-197182.478
Method:	css-mle	S.D. of innovations	21.820
Date:	Thu, 20 Sep 2018	AIC	394372.956
Time:	01:02:07	BIC	394407.706
Sample:	01-01-2010 - 12-31-2014	HQIC	394383.909

	coef	std err	z	P>|z|	[0.025	0.975]
const	97.8876	2.686	36.449	0.000	92.624	103.151
ar.L1.pm_hourly_beijing	1.1899	0.005	255.801	0.000	1.181	1.199
ar.L2.pm_hourly_beijing	-0.2287	0.005	-49.157	0.000	-0.238	-0.220

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0539	+0.0000j	1.0539	0.0000
AR.2	4.1498	+0.0000j	4.1498	0.0000



First few predictions

date_time	actual	predict
2015-01-01 00:00:00	16.75	18.272462
2015-01-01 01:00:00	8.00	22.395916
2015-01-01 02:00:00	9.50	26.268130
2015-01-01 03:00:00	9.25	29.932639
2015-01-01 04:00:00	7.25	33.407459
2015-01-01 05:00:00	7.50	36.704070
2015-01-01 06:00:00	9.25	39.832010
2015-01-01 07:00:00	12.25	42.800008
2015-01-01 08:00:00	11.00	45.616264
2015-01-01 09:00:00	21.50	48.288542
2015-01-01 10:00:00	28.50	50.824204
2015-01-01 11:00:00	35.25	53.230234
2015-01-01 12:00:00	57.00	55.513261
2015-01-01 13:00:00	53.50	57.679572
2015-01-01 14:00:00	43.25	59.735135

- Acceptable performance on training sets
- Reasonable precision for the very first out-of-sample prediction into the future
- Poor predictions for further steps into the future

Time series alone is not enough
Other underlying factors!

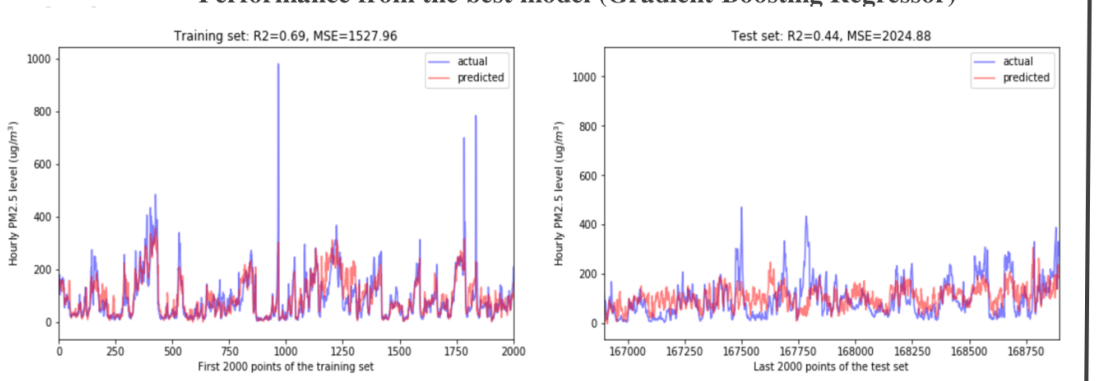
Modeling – Datetime + Weather Feature

Machine learning regression models are constructed by incorporating weather data with datetime data:
Datetime feature: year, month, day, hour, season; Weather feature: wind speed, temperature, humidity, etc.
Training set: year 2010-2014; Test set: year 2015; Optimization: grid search

Performance summary of various regression models

	r2_train	r2_test	mse_train	mse_test
LinearRegression	0.237894	0.205384	3802.51	2893.23
Lasso	0.234133	0.209052	3821.28	2879.87
Ridge	0.235065	0.213904	3816.63	2862.21
RandomForestRegressor	0.972667	0.417547	136.377	2120.73
GradientBoostingRegressor	0.693764	0.443874	1527.96	2024.88
KNeighborsRegressor	1	0.346882	0	2378.03
MLPRegressor	0.523008	0.388481	2379.94	2226.57

Performance from the best model (Gradient Boosting Regressor)



Feature importance ranking

Gradient Boosting Regressor		Random Forest Regressor	
feature	importance	feature	importance
day	0.130245	humi	0.119629
month	0.116702	day	0.112051
dewp	0.114533	ws	0.112023
pres	0.091142	dewp	0.105110
humi	0.090524	pres	0.097455
temp	0.080331	temp	0.097101
ws	0.076496	month	0.074197
year	0.075437	hour	0.063550
season	0.041055	year	0.060646
hour	0.038366	season	0.030423
city_guangzhou	0.026419	city_shanghai	0.022324
city_shanghai	0.023500	city_guangzhou	0.020598
city_shenyang	0.019203	cbwd_NW	0.015047
city_chengdu	0.019199	cbwd_SE	0.014330
cbwd_NW	0.014536	city_chengdu	0.013157
cbwd_SE	0.013775	cbwd_NE	0.012142
cbwd_NE	0.008859	city_shenyang	0.011547
iprec	0.007284	iprec	0.007346
cbwd_SW	0.007017	cbwd_SW	0.005669
precipitation	0.005377	precipitation	0.005657

- Gradient Boosting Regressor performs the best among all regressors: R^2 = 0.44 (holdout set)
- Important meteorological weather feature: wind speed, temperature, humidity, dew point, pressure
- Important datetime feature: month, day
- Separate models for individual cities: R^2 up to 0.45-0.5

Feature engineering is needed for additional improvement!

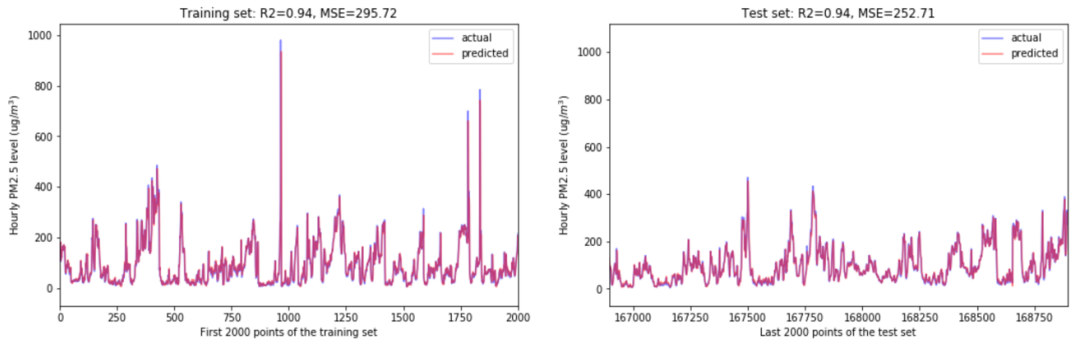
Modeling – Introduce Lag Feature

A lag feature (PM2.5 value from the last step) is included to accommodate its autoregressive nature (progress on prior value):
All previous regression models updated by adding the new lag feature
Train set: year 2010-2014 of four cities; Test set: year 2015 of four cities and all data of the fifth city (unseen city)

Performance summary of various regression models

	r2_train	r2_test	mse_train	mse_test
LinearRegression	0.939551	0.941042	295.604	252.903
Lasso	0.939526	0.941088	295.723	252.706
Ridge	0.939551	0.941042	295.604	252.903
RandomForestRegressor	0.991363	0.912035	42.2353	377.331
GradientBoostingRegressor	0.948737	0.930511	250.679	298.079
KNeighborsRegressor	1	0.841426	0	680.215
MLPRegressor	0.945827	0.939054	264.912	261.431

Performance from the best model (LASSO)



Feature importance ranking

	Gradient Boosting Regressor			Random Forest Regressor	
	feature	importance		feature	importance
12	pm_1hr	0.634249	12	pm_1hr	0.945248
11	ws	0.074355	6	humi	0.006916
3	hour	0.066486	5	dewp	0.006817
8	temp	0.032753	11	ws	0.006777
6	humi	0.029102	7	pres	0.006170
9	precipitation	0.026146	3	hour	0.006016
5	dewp	0.023794	8	temp	0.005168
7	pres	0.017015	2	day	0.005017
15	cbwd_SE	0.015304	0	year	0.002554
19	city_shanghai	0.013725	1	month	0.002412
10	iprec	0.012402	10	iprec	0.001170
14	cbwd_NW	0.009135	14	cbwd_NW	0.000939
13	cbwd_NE	0.007724	15	cbwd_SE	0.000897
16	cbwd_SW	0.007270	4	season	0.000852
0	year	0.006507	9	precipitation	0.000848
4	season	0.006468	13	cbwd_NE	0.000607
18	city_guangzhou	0.006455	19	city_shanghai	0.000476
1	month	0.005770	18	city_guangzhou	0.000387
2	day	0.003197	17	city_chengdu	0.000377
17	city_chengdu	0.001965	16	cbwd_SW	0.000354
20	city_shenyang	0.000000	20	city_shenyang	0.000000

- The introduction of lag feature is successful, R^2 scores → boosted significantly regardless of the model used
- Simple linear models → R^2 = 0.94 (holdout set), more advanced models → no further improvement
- Feature important: the new lag feature → > 60% of the contribution, weather feature → up to 25%, datetime → the rest
- The winning model not only forecast into the future, but also extends to predict an unseen city.

Limitation and Future Work

Limitation of the Winning Model

- Predicts limited steps:
  requires values in the previous step for predicting future steps
  error propagates as the prediction is used to forecast further steps
- Pipeline setup:
  log the forecasted PM2.5 values to enable predictions for further steps
- Maintenance:
  replace prediction with actual observation whenever available to reduce error

Other Potential Future Work

- Ensembles of the above machine learning models to average out bias and improve performance
- Downsampling on a daily frequency instead of using the hourly frequency

Summary

Hourly PM2.5 and weather data of five major Chinese cities in 2010-2015 are analyzed and modeled:

Patterns in PM2.5

- **Alters significantly among cities**
  - air quality from best to worse are Guangzhou, Shanghai, Chengdu and Shenyang, Beijing
  - PM2.5 levels for all five cities are way too high to be considered healthy and safe
- **Varies systematically with time**
  - daily & monthly trends are unique for individual cities
  - yearly trend shows consistent improvements over time
- **Correlates with meteorological weather data to some extent**
  - high wind, large precipitation improve air quality in all cities
  - wind direction affects air quality, but the dependency differs by city
  - PM2.5 positively correlated to temperature/dew point, and negatively correlated to pressure/humidity

Modeling PM2.5

- Historical value, datetime, weather condition all affect future PM2.5 values
- Historical value → baseline of future PM2.5, weather condition + datetime → how future PM2.5 deviates from its past
- The winning model achieves R^2 of 0.94 on the holdout set, not only forecasts future, but also extends to predict an unseen city

Acknowledgement

- Mentor: Max Sop
- UCI Machine Learning repository
- Springboard team