Collegio Carlo Alberto
UNIVERSITÀ DEGLI STUDI DI TORINO

Predicting Air Quality in Milan

Dan Herweg

May 2019

MADAS
MASTER in DATA SCIENCE for COMPLEX ECONOMIC SYSTEMS

The Air Quality in Milan is over 80% predictable using ensemble models and 5 variables, 4 of which are weather

Summary

- Milan struggles with air quality, an important health issue. Decision makers could benefit from being able to predict the Air Quality Index on an hourly basis

- Using sensor data for weather and traffic to predict the index, it has been found that the AQI is over 80% predictable, mostly using weather variables. They are:
  - Atmospheric Pressure
  - Wind Speed
  - Temperature
  - Relative Humidity
  - Traffic

- The random forest and bagging model were most accurate, but the bagging model presented fewer large errors and should be preferred

- Data collection should continue, and the model should be retrained regularly to overcome initial weaknesses in analysis

Contents

I. Background

I. Motivation

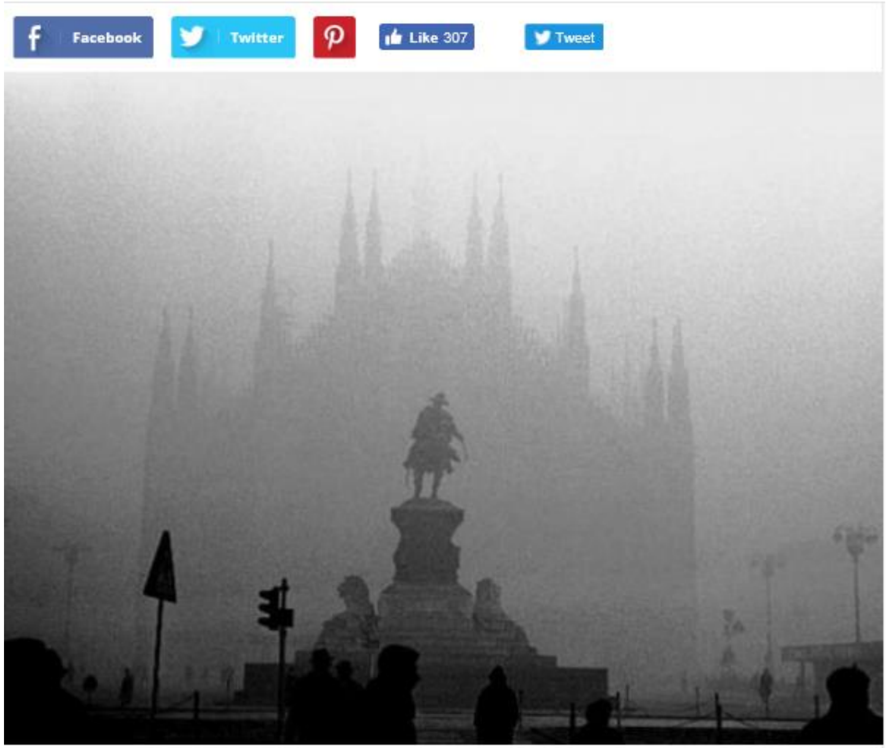
II. Methodology

Milan has poor air quality

Motivation

Milan has second worst smog in Europe – WHO

30 Jan, 2018



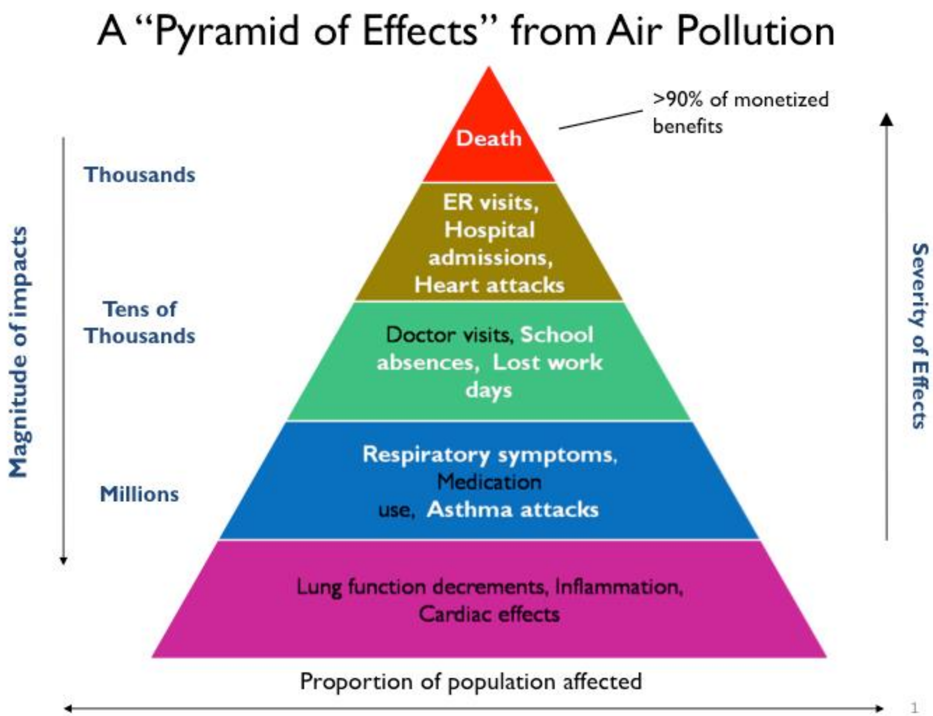
A report by the World Health Organization has placed Milan just behind Turin and just before Naples as the three European cities with the worst levels of atmospheric pollution.

Article based on WHO report using 2016 data

https://www.wantedinmilan.com/news/milan-has-second-worst-smog-in-europe-who.html

Poor air quality affects citizens' health

Motivation



Fine particles can enter deep into the lungs and enter the blood stream. **Health impacts from particles include:**

- **Premature death**
- **Non-fatal heart attacks**
- **Aggravated asthma**

US Environmental Protection Agency

https://www.epa.gov/benmap/how-benmap-ce-estimates-health-and-economic-effects-air-pollution

Predicting AQI could help decision makers introduce interventions that are predicted to manage air quality in real time improving citizens' health

I. **Background**

   I. Motivation

   II. **Methodology**

We will take sensor data from weather and traffic to predict AQI using machine learning methods, selecting on accuracy

Methodology

- First, data from weather stations, pollution sensors, and traffic gates will be cleaned and explored. The Air Quality Index target variable will be constructed

- Next, the numerical and categorical AQI will be computed directly and by first predicting pollutants

- The best features will be selected and the best model will be chosen based on accuracy in predicting the categorical AQI

II. Data

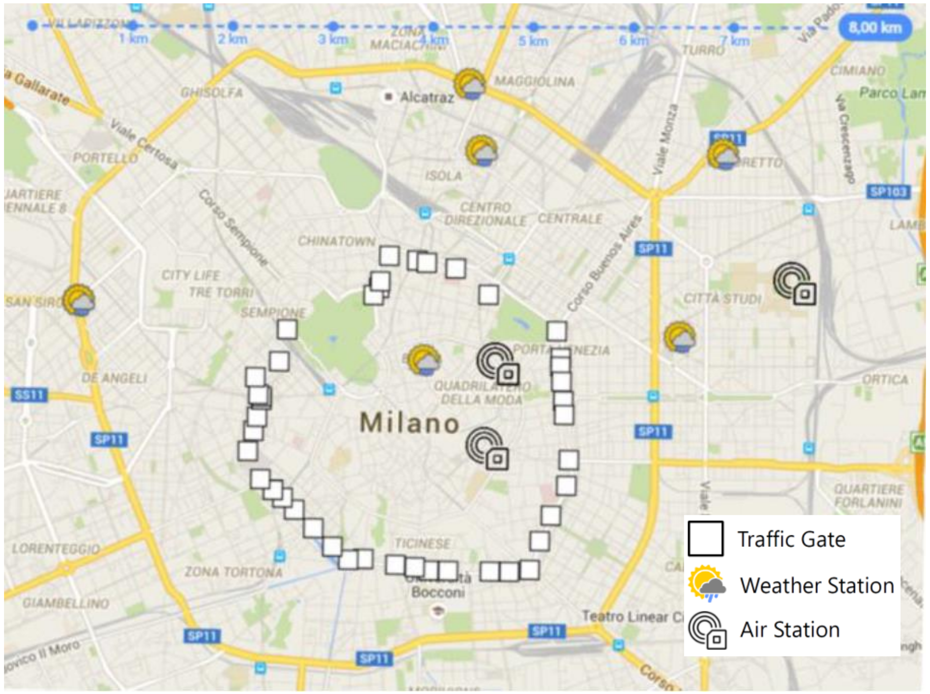
I. Feature Construction

II. Data Exploration

Data had to be extracted from files generated by sensors, cleaned and merged

Feature Construction

Milan Sensor Locations



- The data had three main sources
  - Weather Sensors
  - Traffic Gates
  - Pollution Sensors (used to calculate AQI)
- Sensors measuring the same data were averaged
- Temporal data was added based on the timestamp

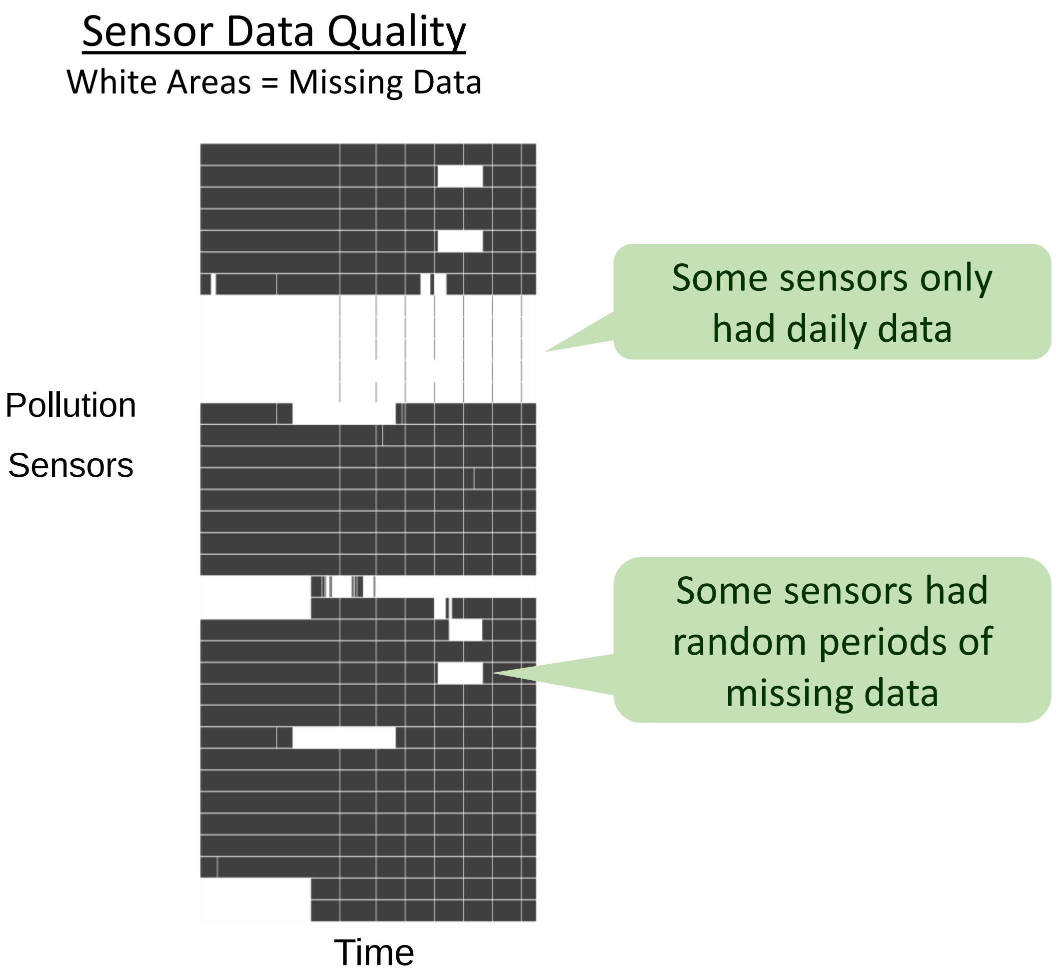
The quantity of initial features was very large

Feature Construction

Weather Features	Traffic Features
Wind Direction Wind Speed Temperature Relative Humidity Precipitation Global Radiation Atmospheric Pressure Net Radiation	Total Traffic Count Count by: -Vehicle Type -Fuel Type -Vehicle Length -Emissions Standard -Presence of a Diesel Particulate Filter
Pollutant Features (For prediction/imputation of AQI components and calculation of AQI only) PM10* PM2.5 Ammonia Benzene Ozone* Sulphur Dioxide Total Nitrogen*	**Temporal Features** Time of Day Weekend/Weekday

*AQI Component

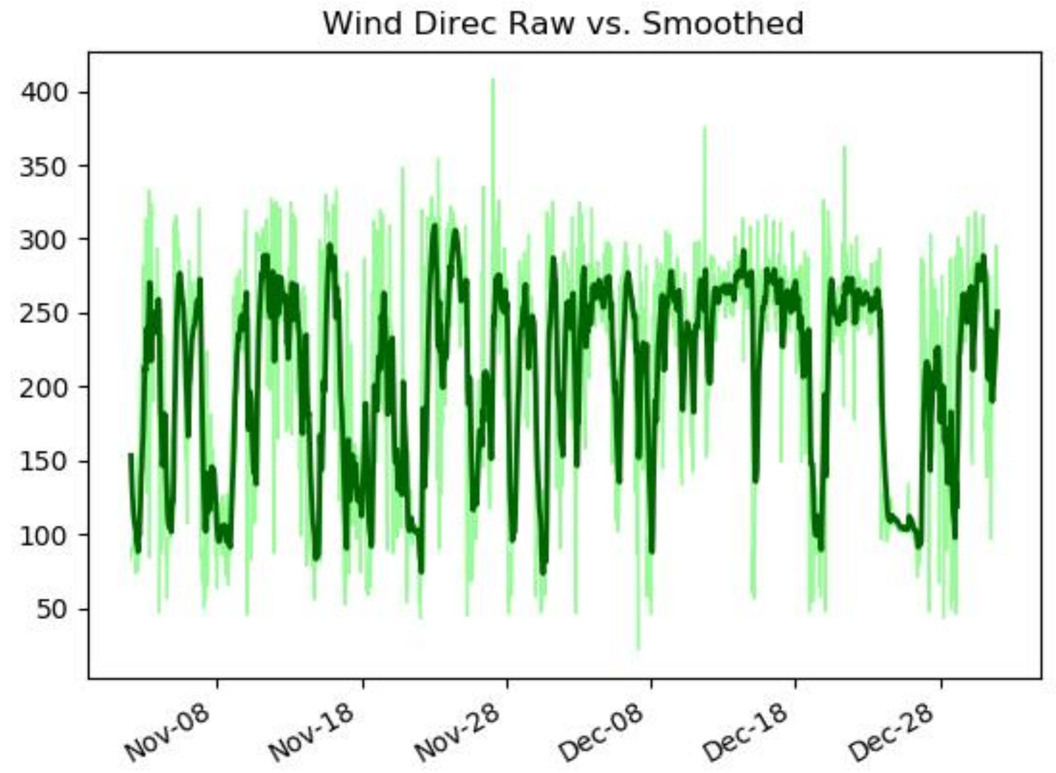
Missing data was an issue
Feature Construction



- Values for missing data were imputed using random forest on the type of data that had missing values

Noisy data was an issue
Feature Construction

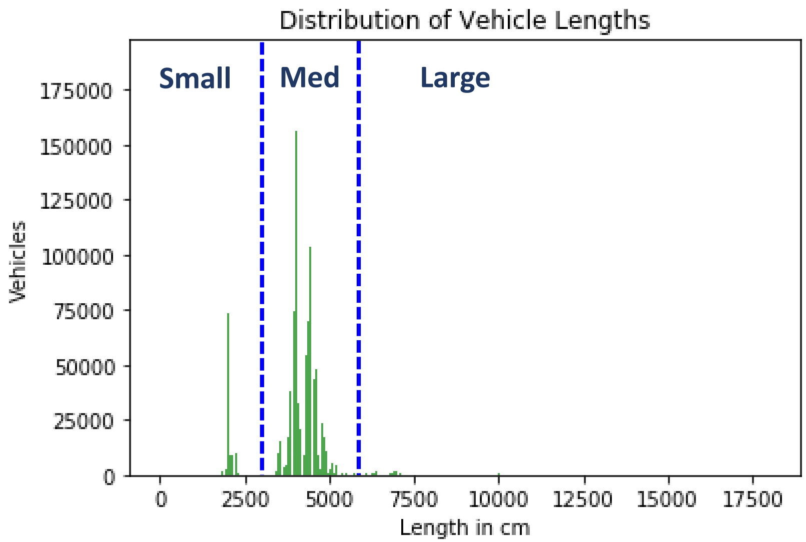
- Noisy time series data with potential for measurement errors was smoothed



Some features were constructed by grouping raw data

Feature Construction

- The vehicle length categories were selected by visually inspecting the distribution to identify a “small” mode, a “medium” mode, and a “large” tail



- A time of day variable was constructed by grouping hours as the following:
  - Morning: 5-11am
  - Mid day = 11am-3pm
  - Evening = 3-8pm
  - Night = 8pm-5am

AQI had to be calculated and then transformed to a classifier to serve as the target variables

Feature Construction

- The AQI numerical calculation:

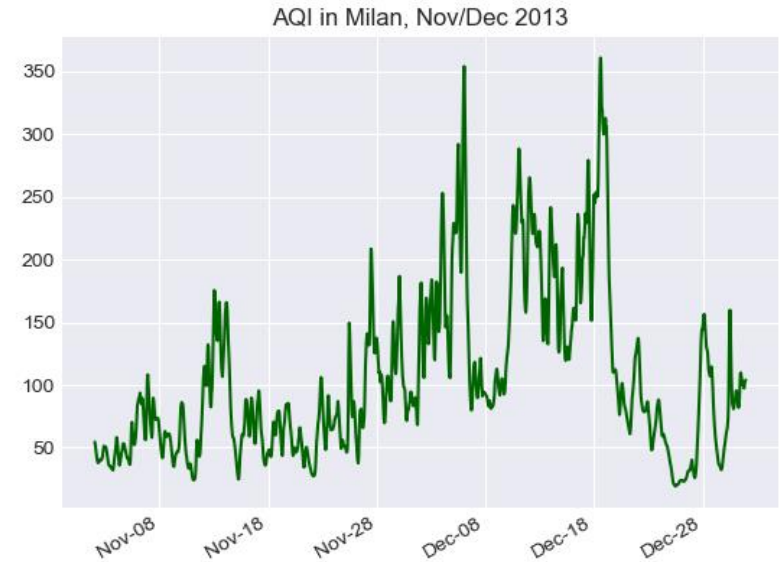
$$I_{IQA} = \frac{I_{PM10} + \max(I_{NO2}, I_{O3})}{2}$$

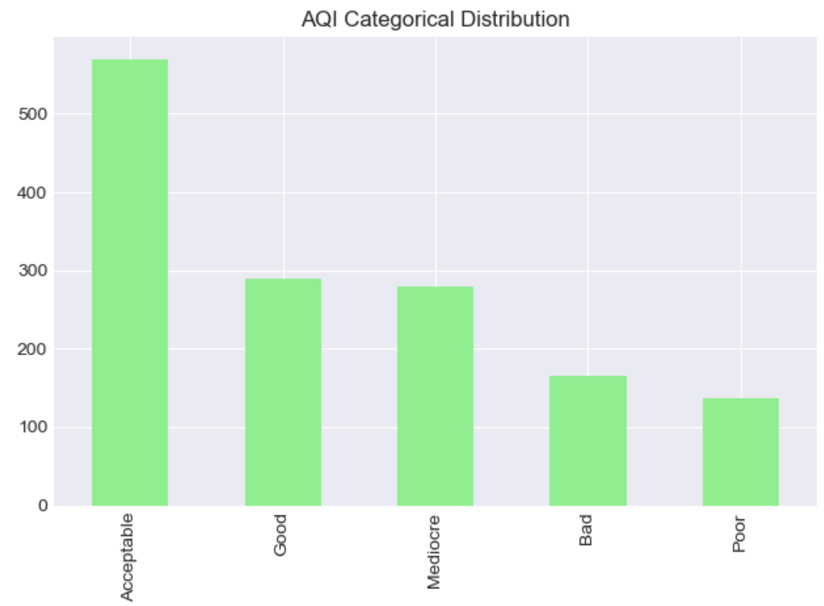
- Categorical classification of AQI values:

Valori dell'indice	Cromatismi	Qualità dell'aria
< 50	🟢	Buona
50-99	🟡	Accettabile
100-149	🟠	Mediocre
150-199	🔴	Scadente
> 200	🟣	Pessima

Air Quality Index score and class distribution

Feature Construction



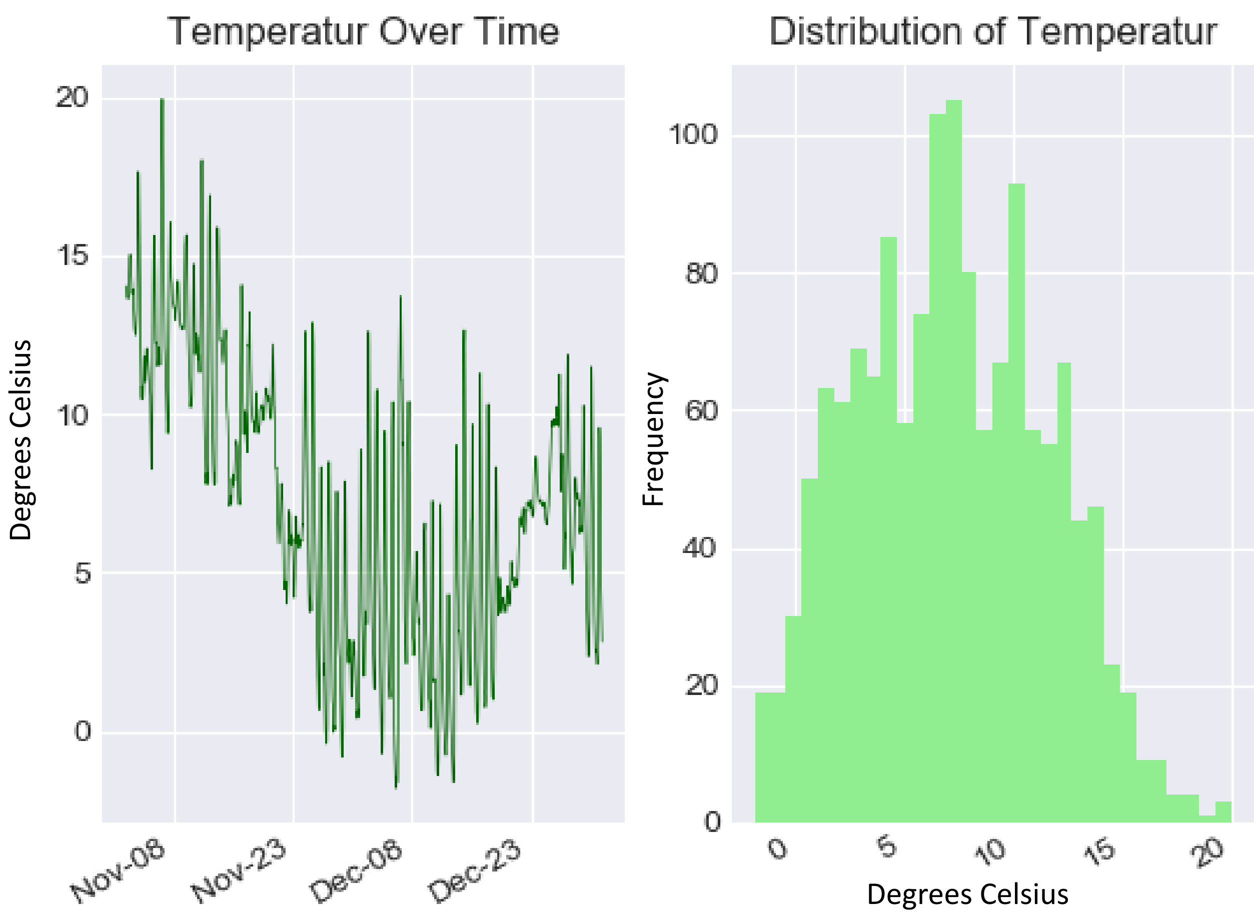


II. Data

- I. Feature Construction
- II. Data Exploration
  - I. Univariate
  - II. Multivariate

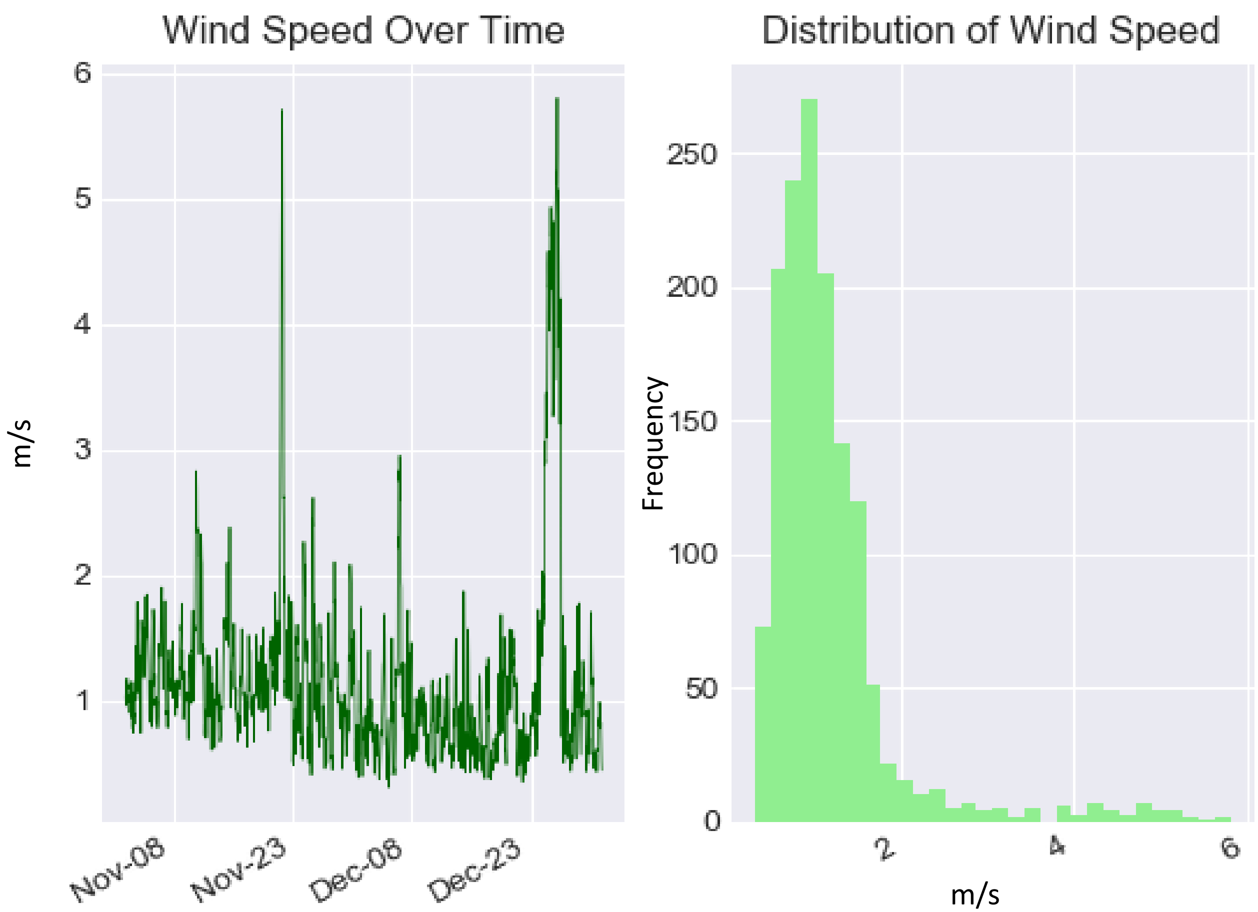
Temperature had a normal distribution and declined into year end

Data Exploration - Univariate



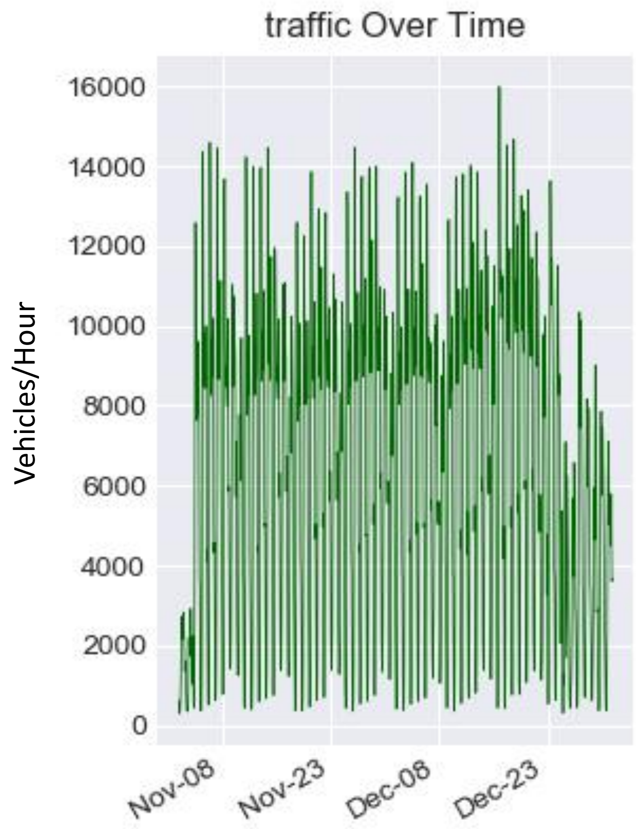
Wind speed was very right skewed, with high peaks

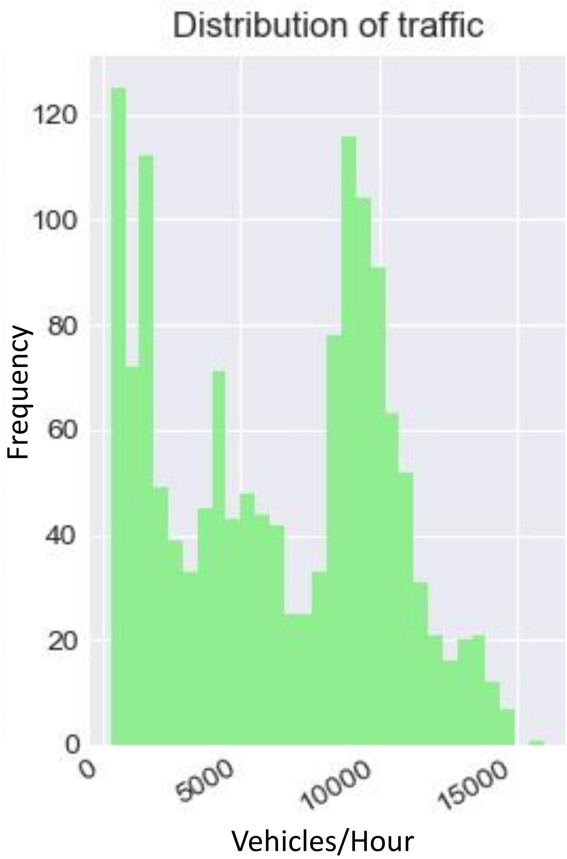
Data Exploration - Univariate



Traffic had daily and weekly patterns and a distribution with multiple peaks

Data Exploration - Univariate

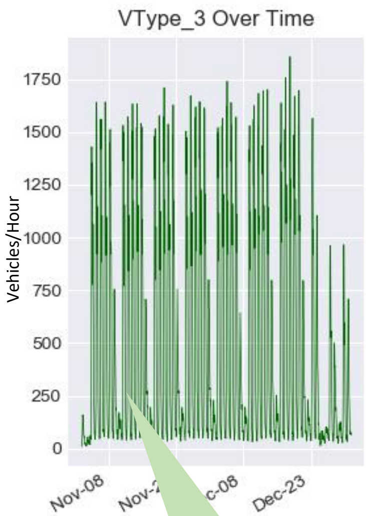


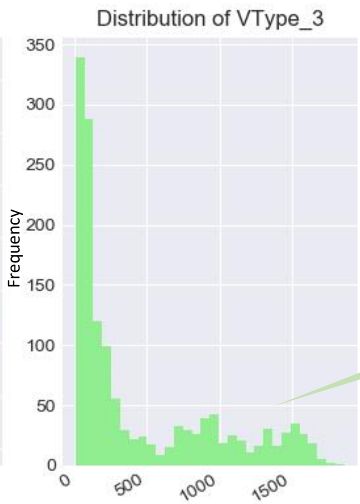


Within traffic, different vehicle types had distinct behaviors

Data Exploration - Univariate

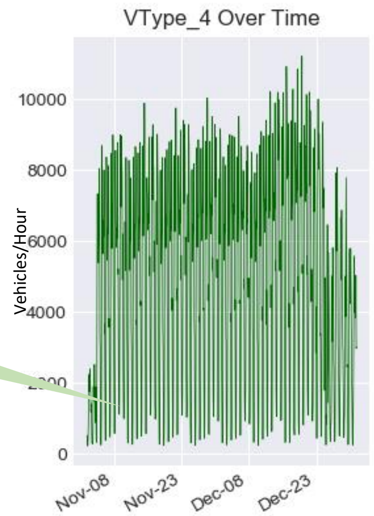
Freight Vehicles

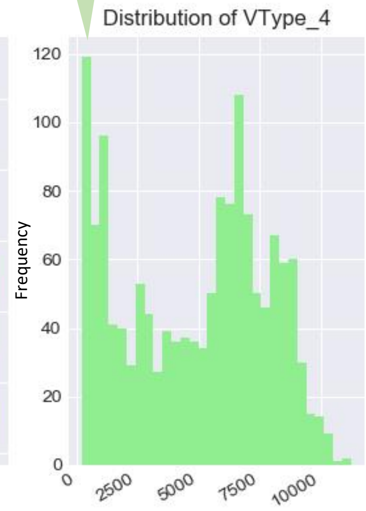




Right skewed distribution vs. multi-modal distribution

People Vehicles

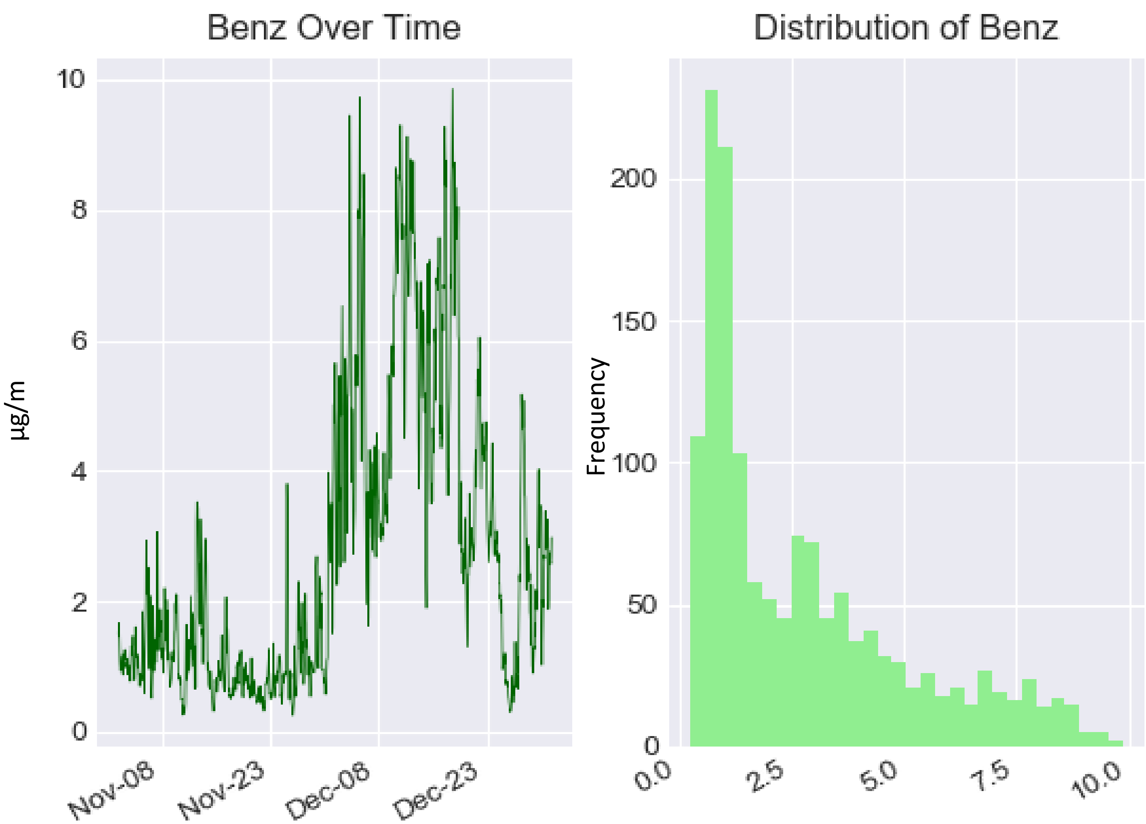




Pronounced weekend troughsvs. less visible weekends

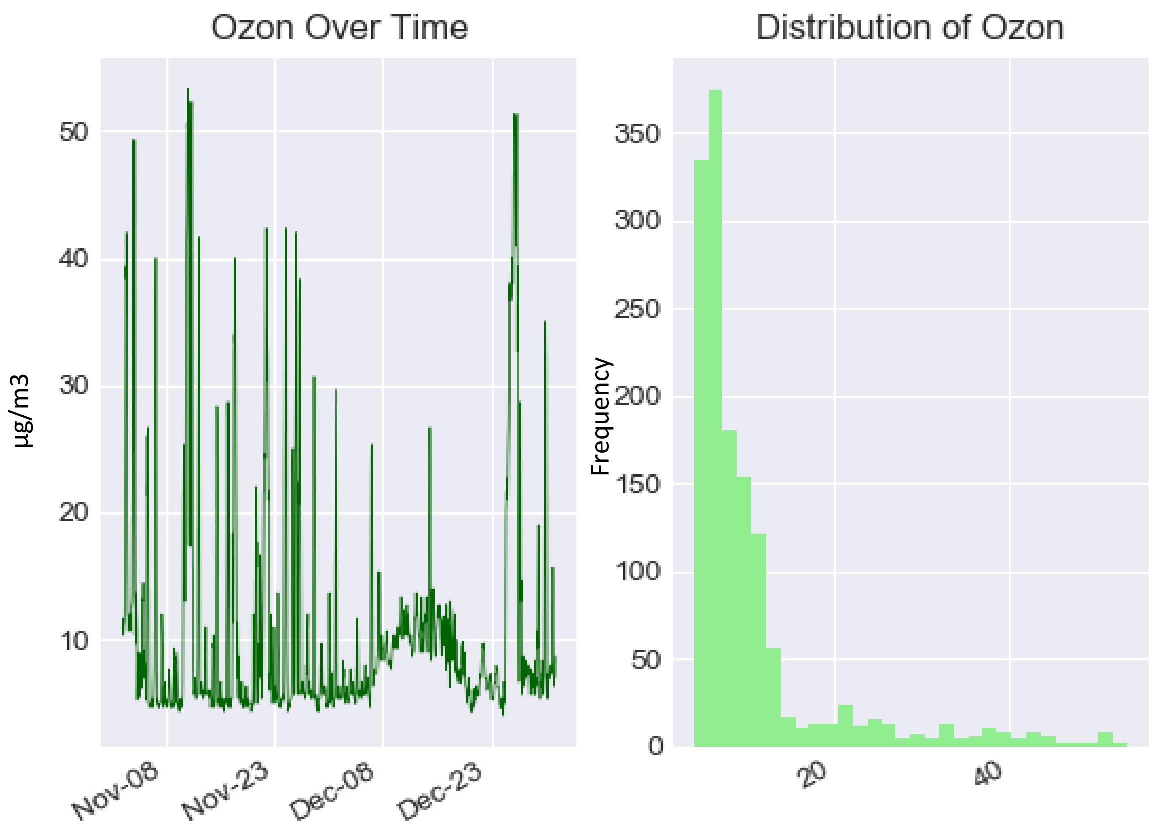
Benzine, a pollutant not in the AQI calculation, looked similar to the AQI over time

Data Exploration - Univariate



However Ozone, a component of the AQI, showed a different pattern

Data Exploration - Univariate



II. Data

I. Feature Construction

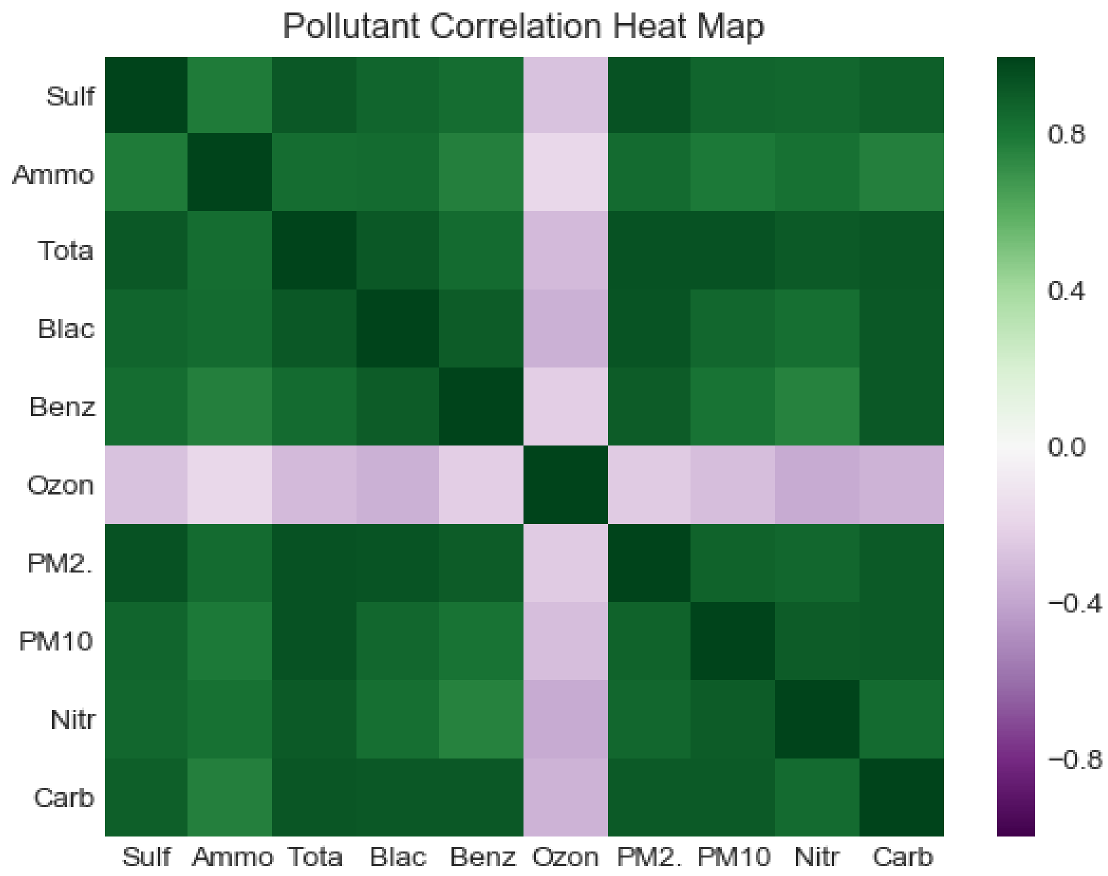
II. Data Exploration

I. Univariate

II. Multivariate

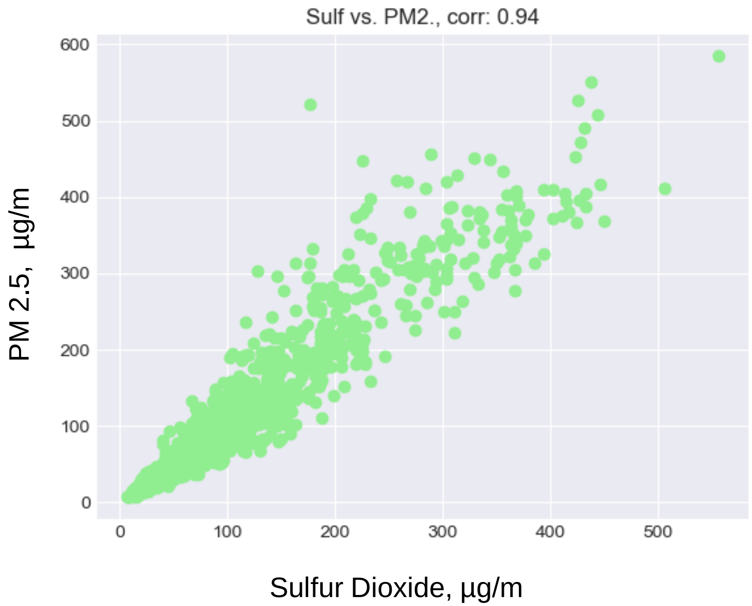
Pollutants are very correlated except ozone

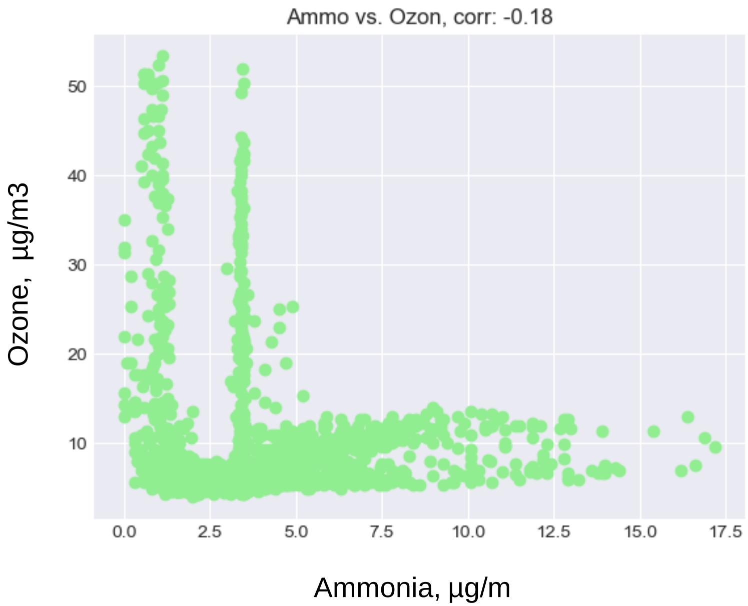
Data Exploration - Multivariate



Pollutants are very correlated except ozone

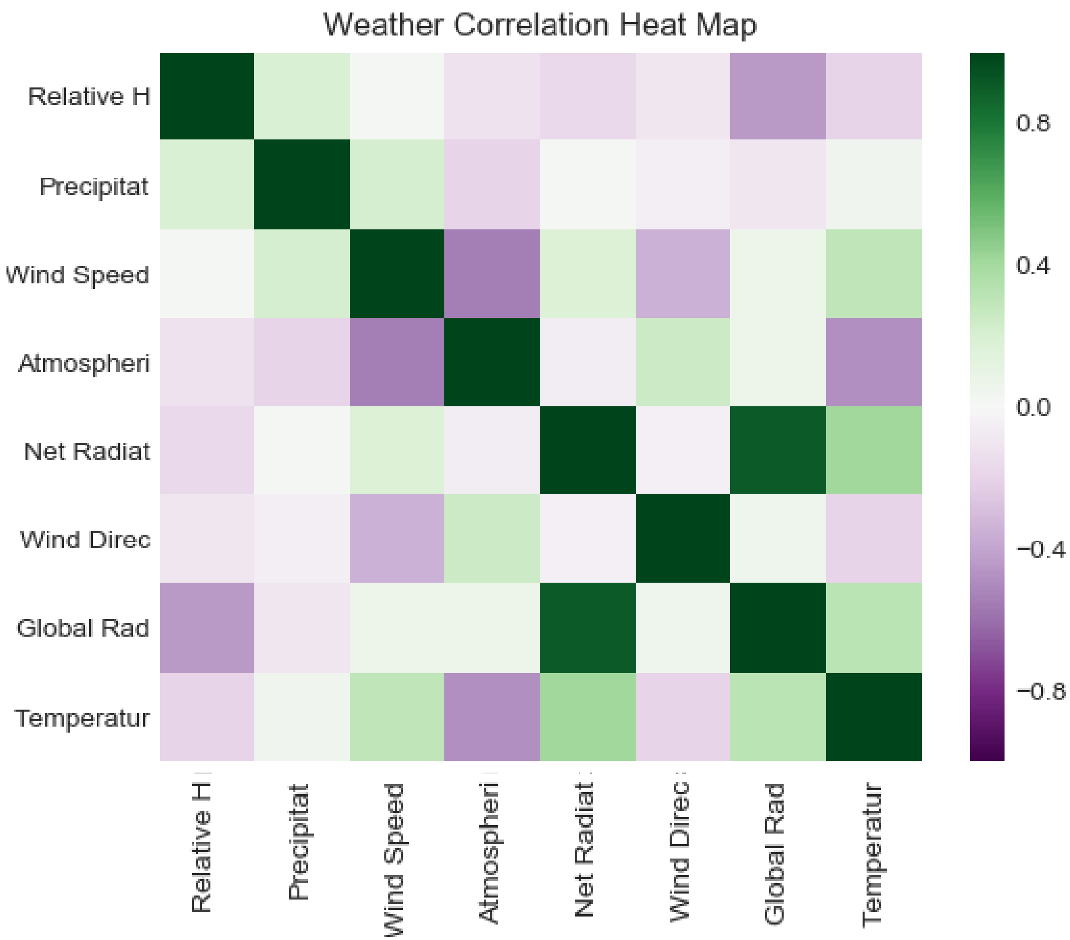
Data Exploration - Multivariate





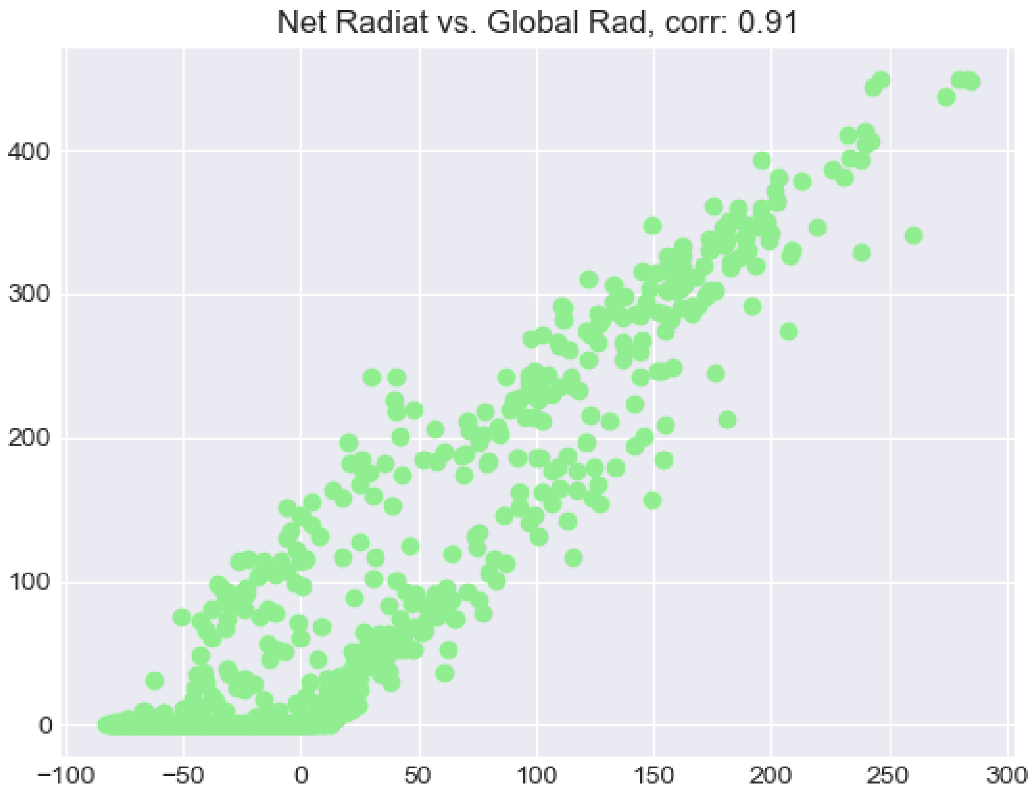
Weather data is not very correlated except the two radiation measures

Data Exploration - Multivariate



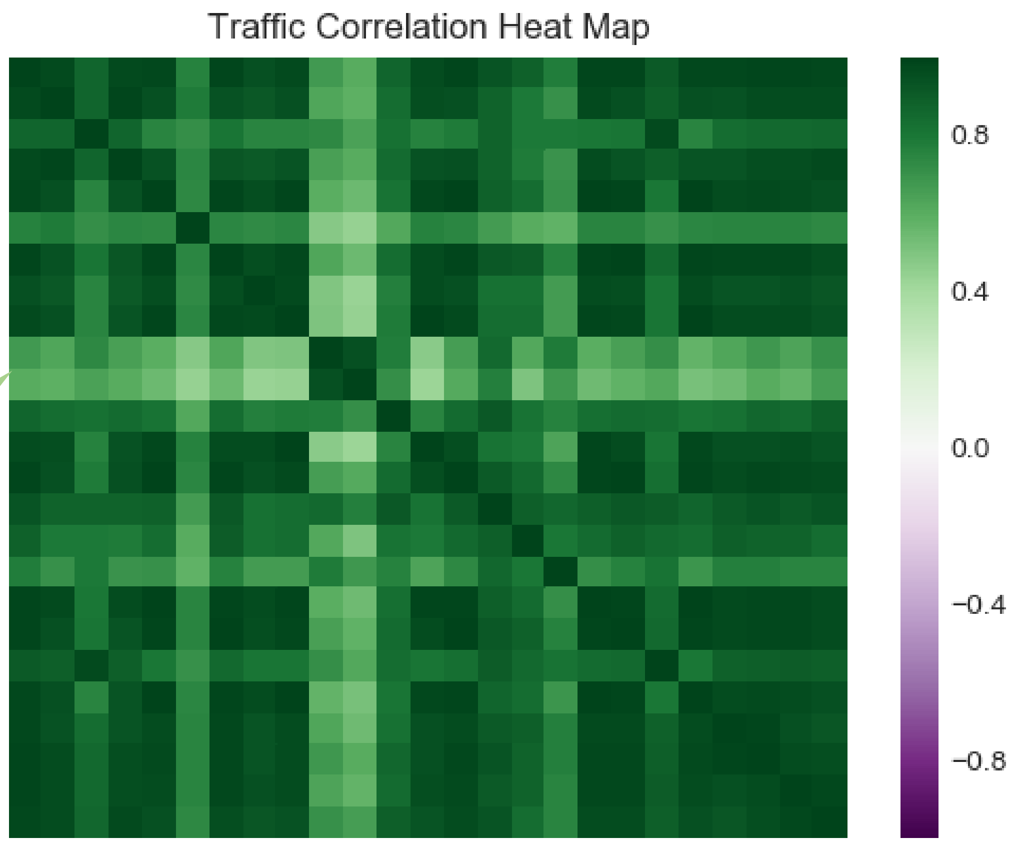
Weather data is not very correlated except the two radiation measures

Data Exploration - Multivariate



Traffic data is more correlated, and all positively correlated

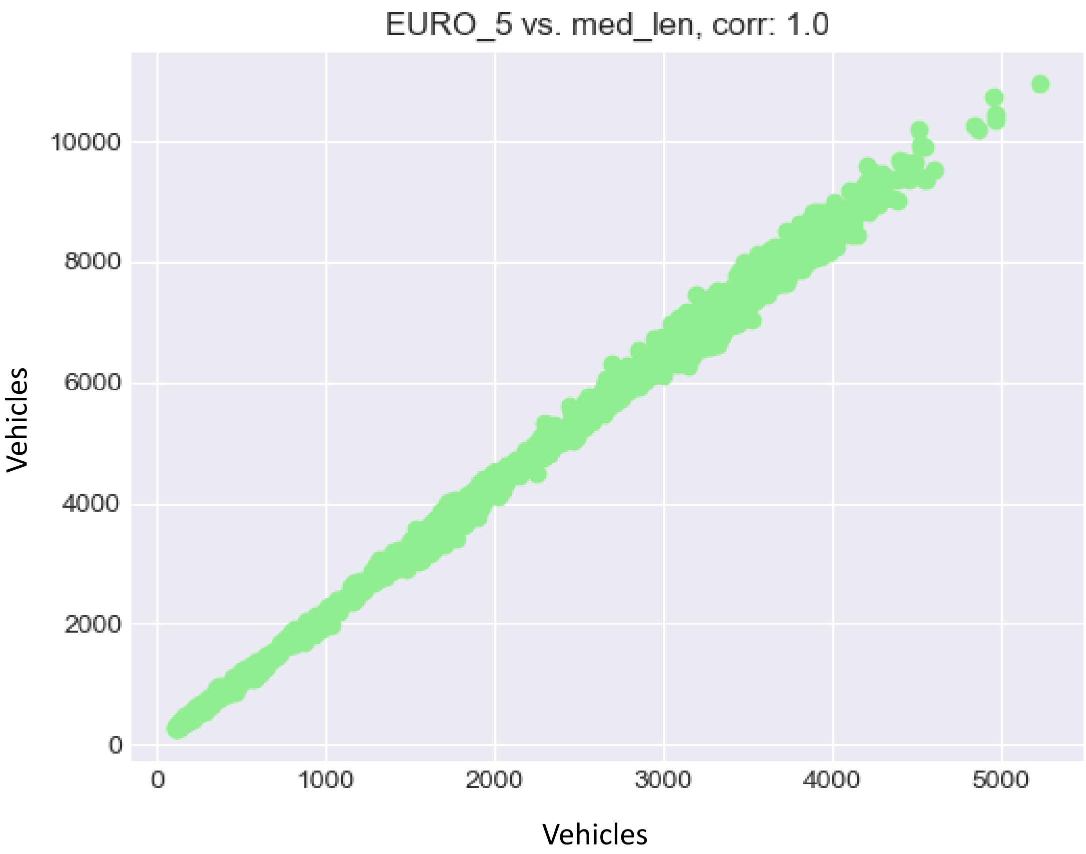
Data Exploration - Multivariate



The only variables that seem uncorrelated are “Other” and Freight Vehicles, which are correlated with each other

Traffic data is more correlated, and all positively correlated

Data Exploration - Multivariate



- The correlation of the number of Euro 5 emissions designated vehicles and the number of medium length vehicles rounds to 1!

III. Model

I. Train/Test Split

II. Predicting AQI

III. Feature Selection

IV. Best Model

Data limitations forced tradeoffs in choosing train and test sets

Train/Test Split

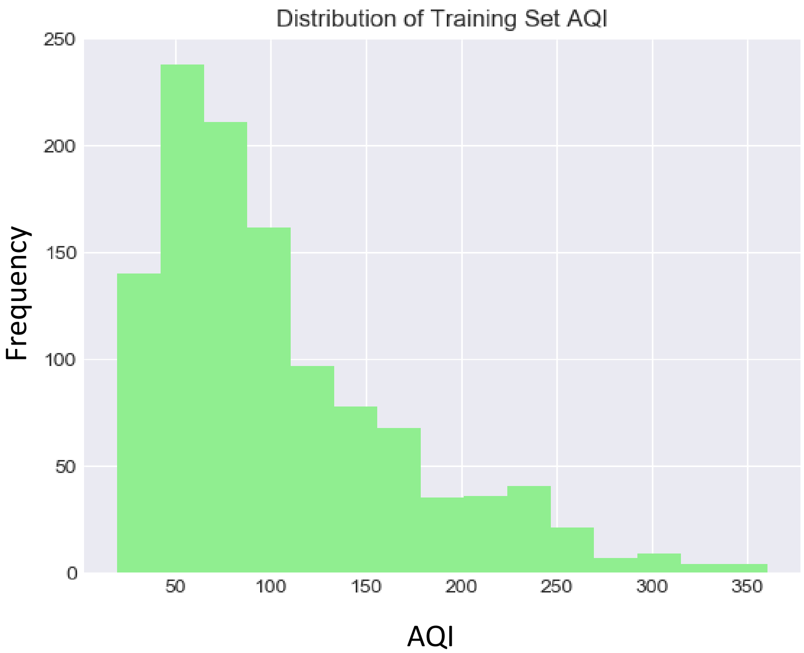
- We do not have a lot of data

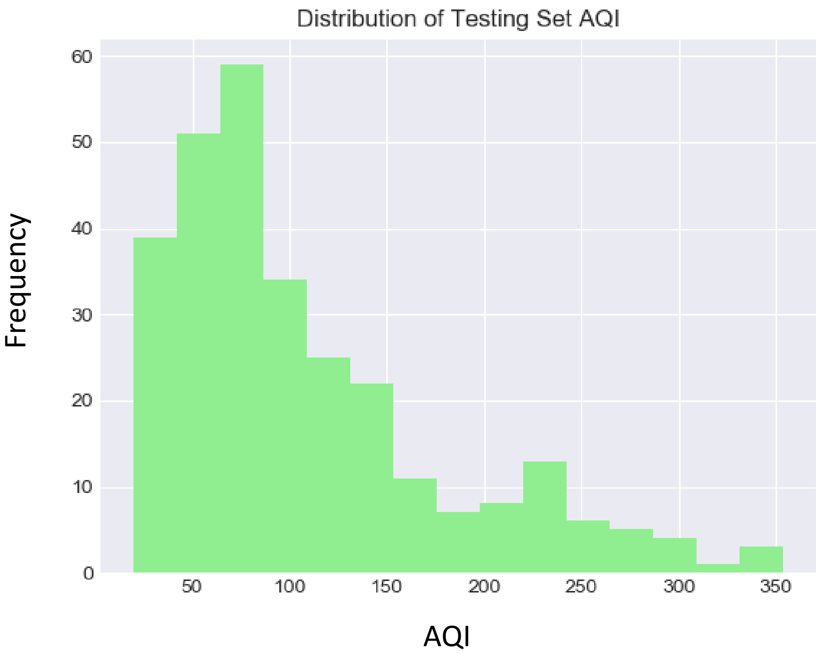
- The data ends in the end of December, an unrepresentative period
  - Intuitively, traffic patterns are expected to be anomalous
  - Technically, the target variable distribution is unbalanced
  - There is a secular decrease in temperatures and not a full year of data to compare the whole cycle

- **Since a sequential train/test split may not be feasible at this time, a stratified sampling technique was used to split the data**
  - As more data is collected, this should be revisited to reduce dependence between the training and testing set, especially when using smoothed data
  - A validation set could be added to account for the number of models used

The distributions of the classes in the train and test sets

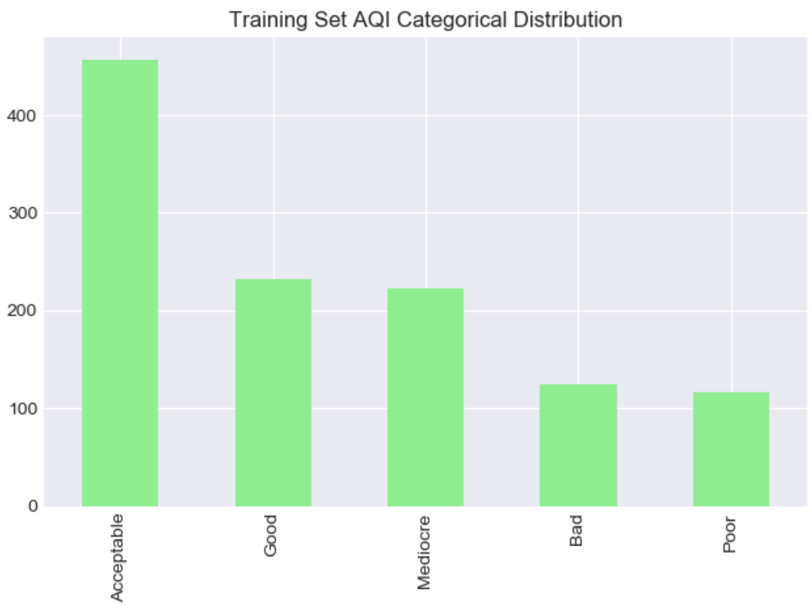
Train/Test Split

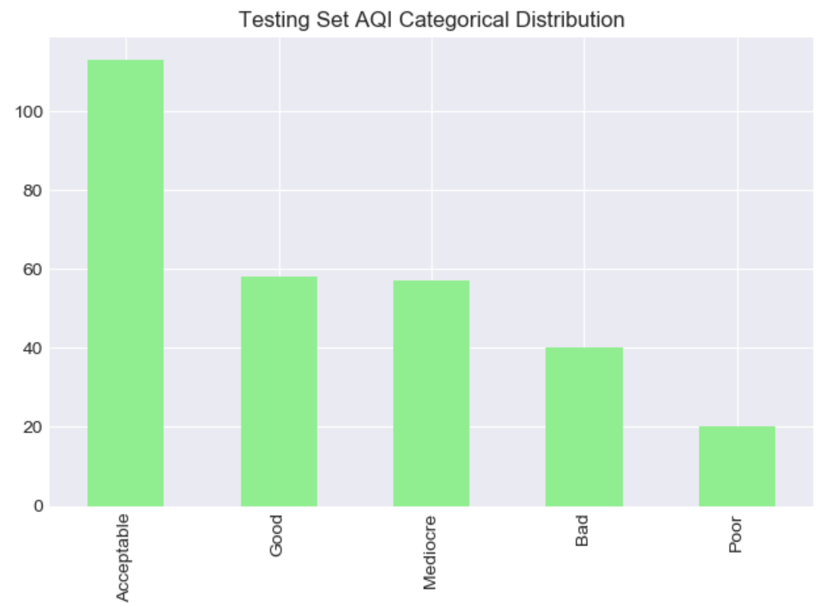




The distributions of the classes in the train and test sets

Train/Test Split





III. Model

I. Train/Test Split

II. Predicting AQI

III. Feature Selection

IV. Final Model

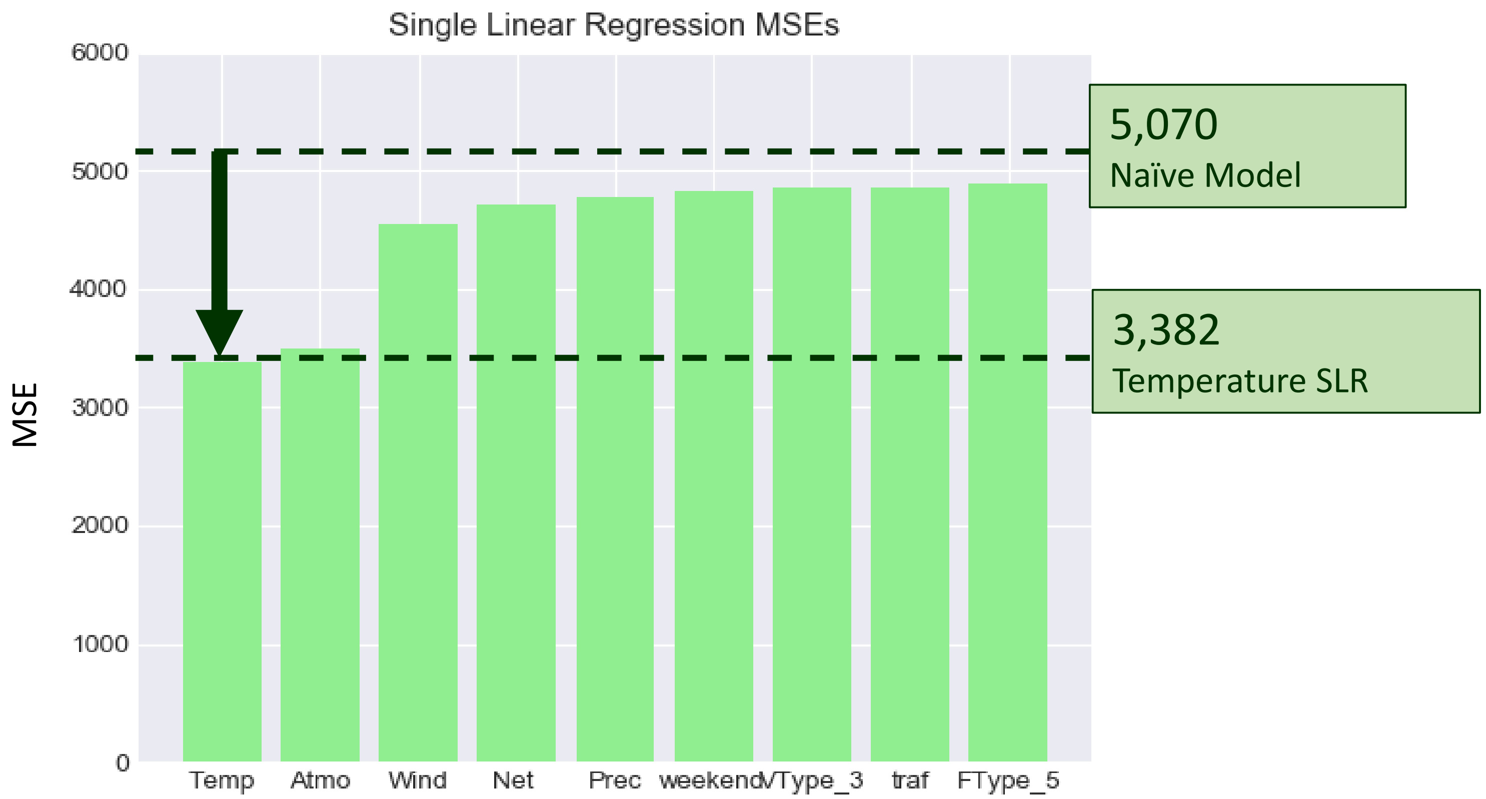
A naïve result serves as a benchmark

Predicting AQI

- Guessing the average AQI gave a mean squared error of **5,070**
- Guessing the modal AQI class “Acceptable” yielded an accuracy of **39%**

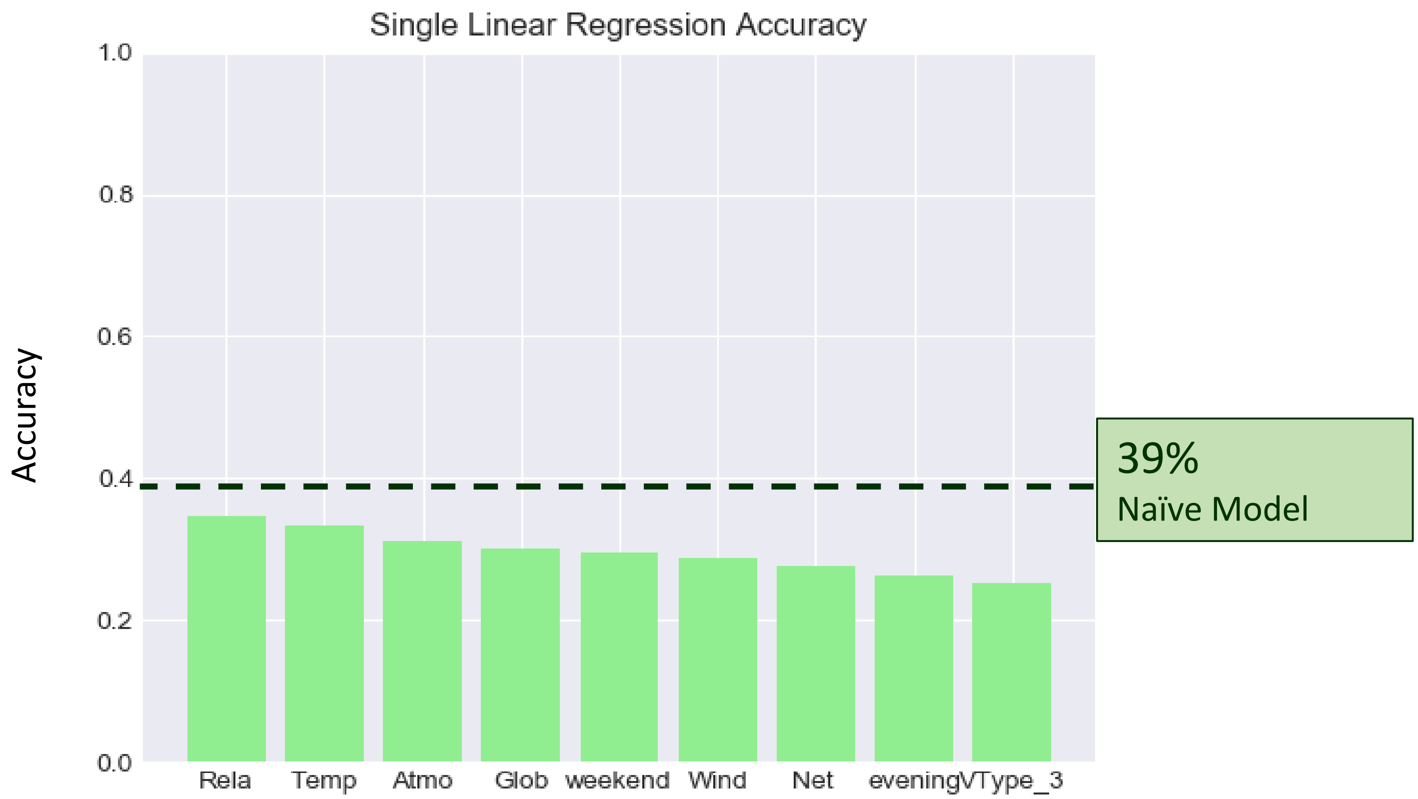
Regressing each variable on AQI yielded better MSE than the naïve model by construction

Predicting AQI



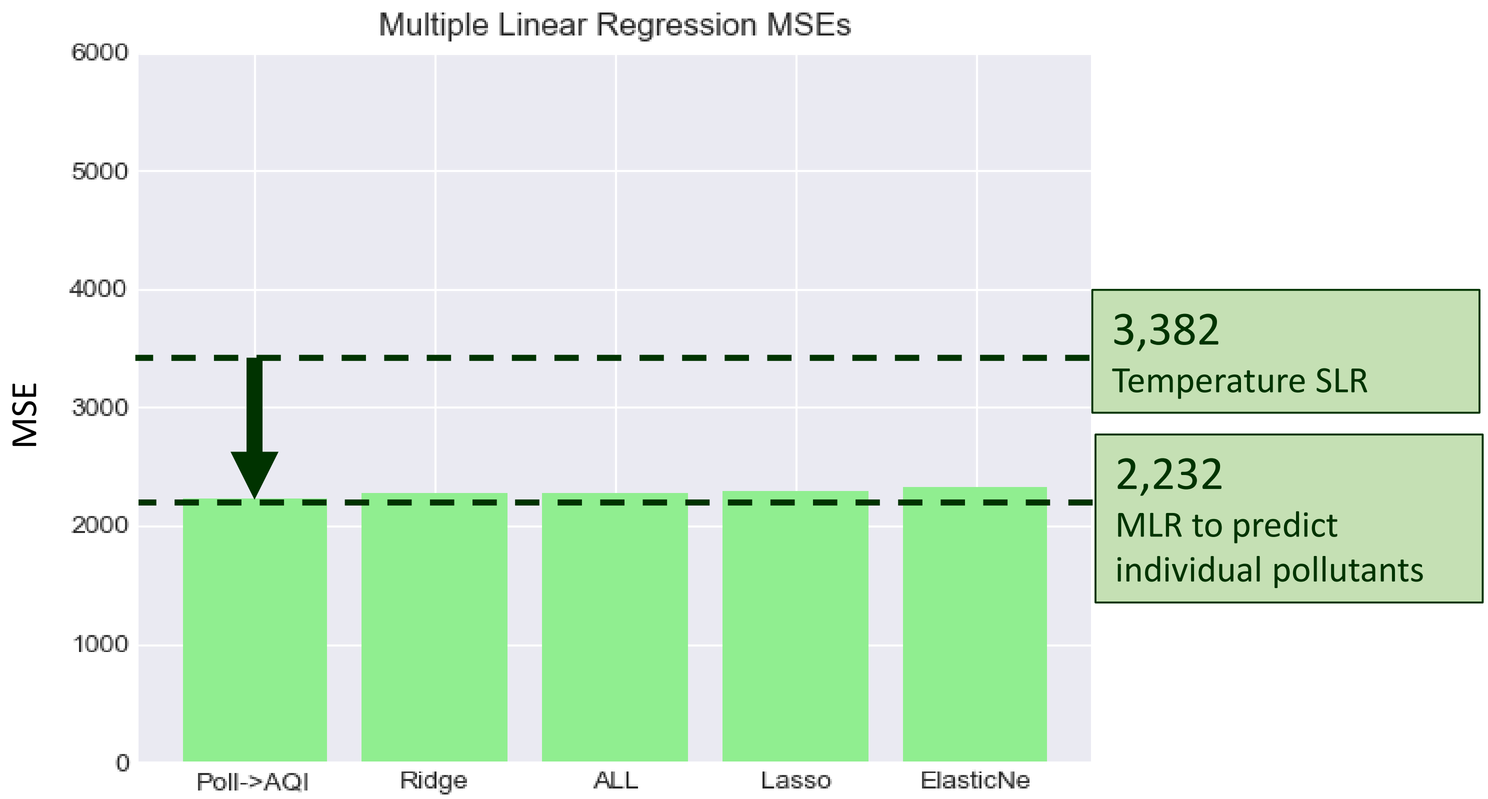
Single linear regressions did not improve accuracy over the naïve model

Predicting AQI



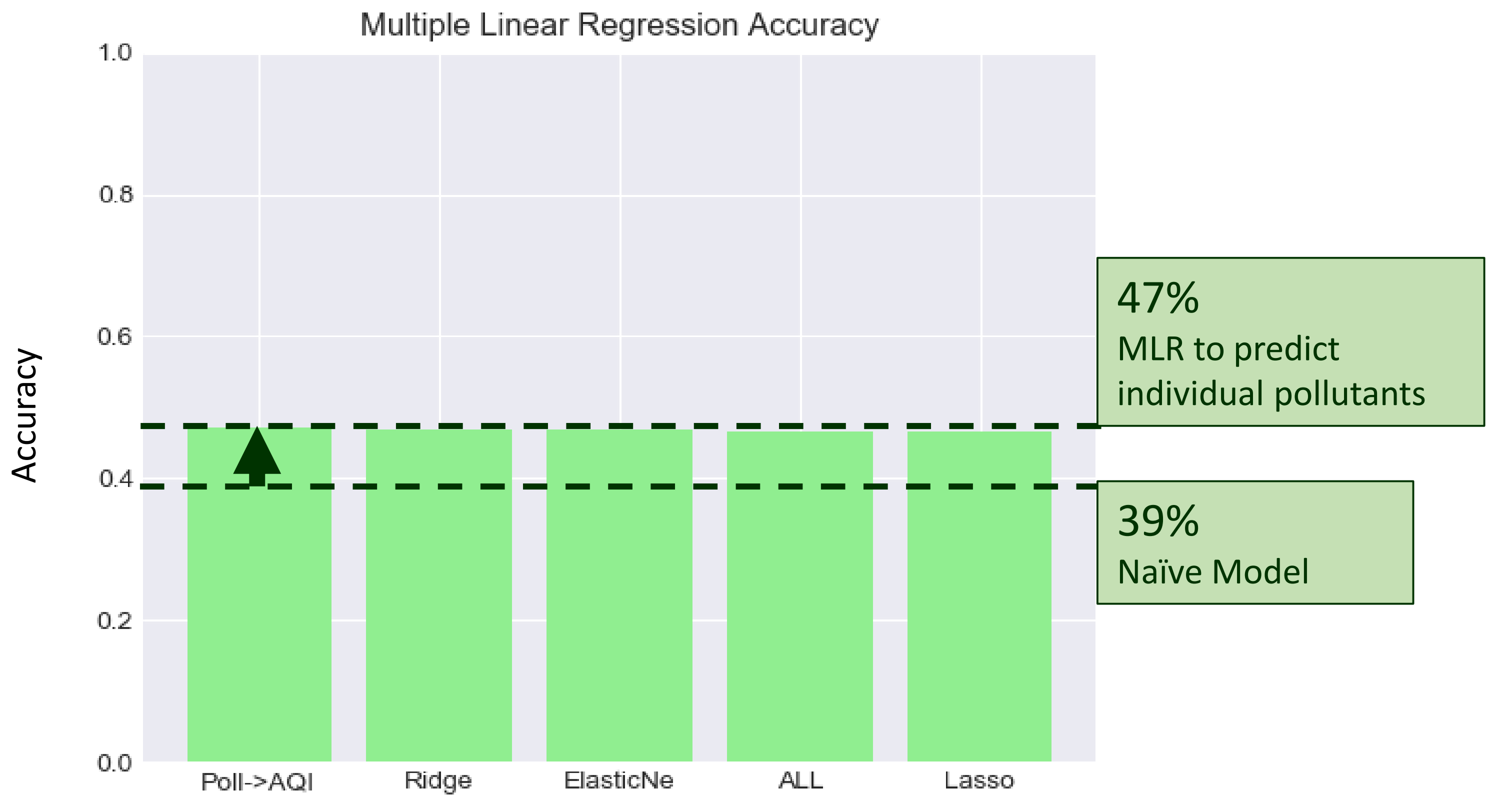
Multiple regression methods further reduced MSE vs. SLR. The best model predicted individual pollutants to compute AQI

Predicting AQI



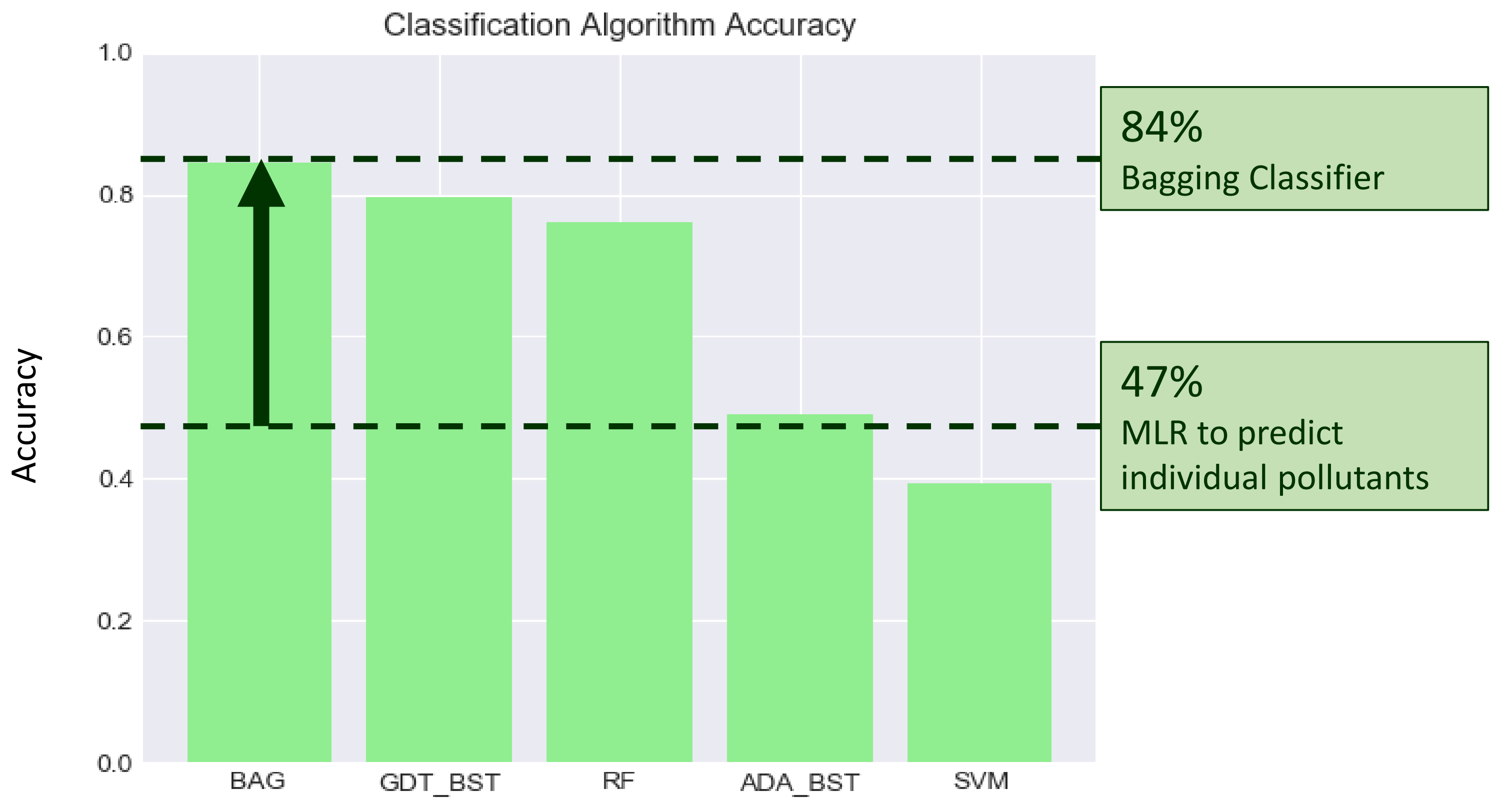
Predicting individual pollutants increased accuracy to 47%

Predicting AQI



Ensemble classification algorithms were most accurate, with bagging at 84%, gradient boosting at 80% and random forest at 76%

Predicting AQI



III. Model

I. Train/Test Split
II. Predicting AQI
III. Feature Selection
IV. Final Model

Several methods that have been used can inform our final choice of variables

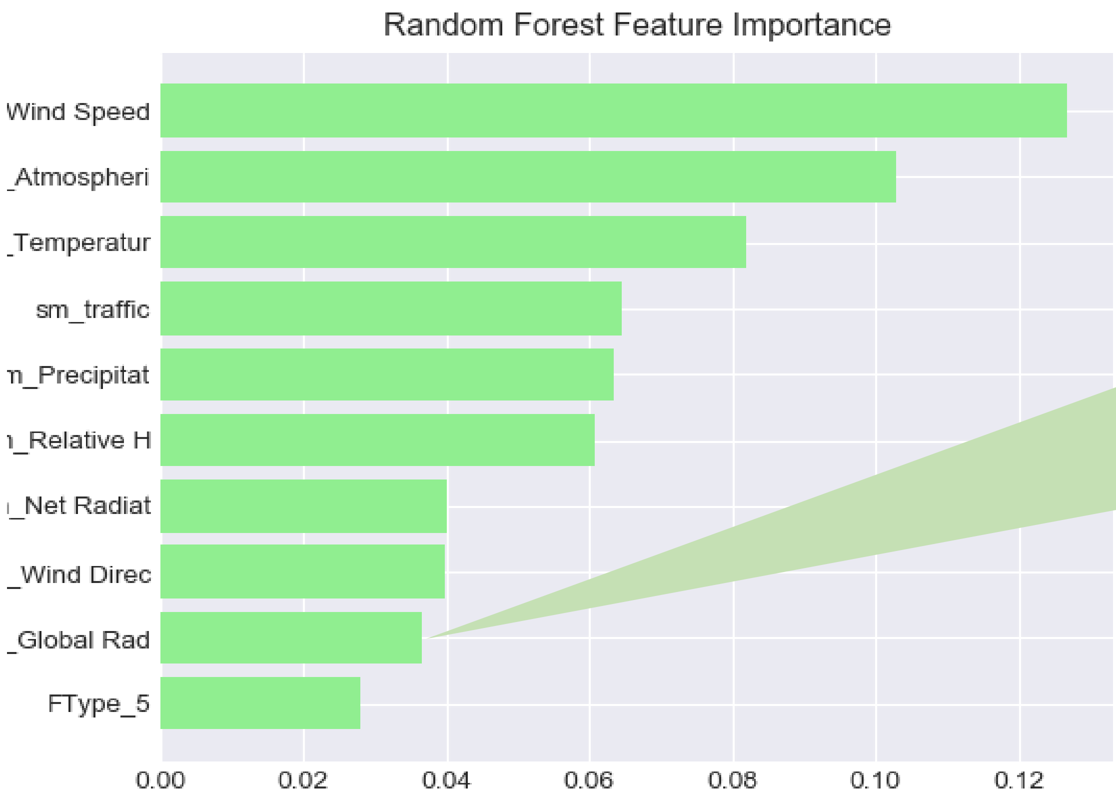
Feature Selection

- Embedded methods: **Random Forest Feature Importance**
  - RF was by far the better model so this will be weighted more heavily

- Penalization methods: **Lasso Coefficients**

Random forest rates wind speed, atmospheric pressure, and temperature as the most important features

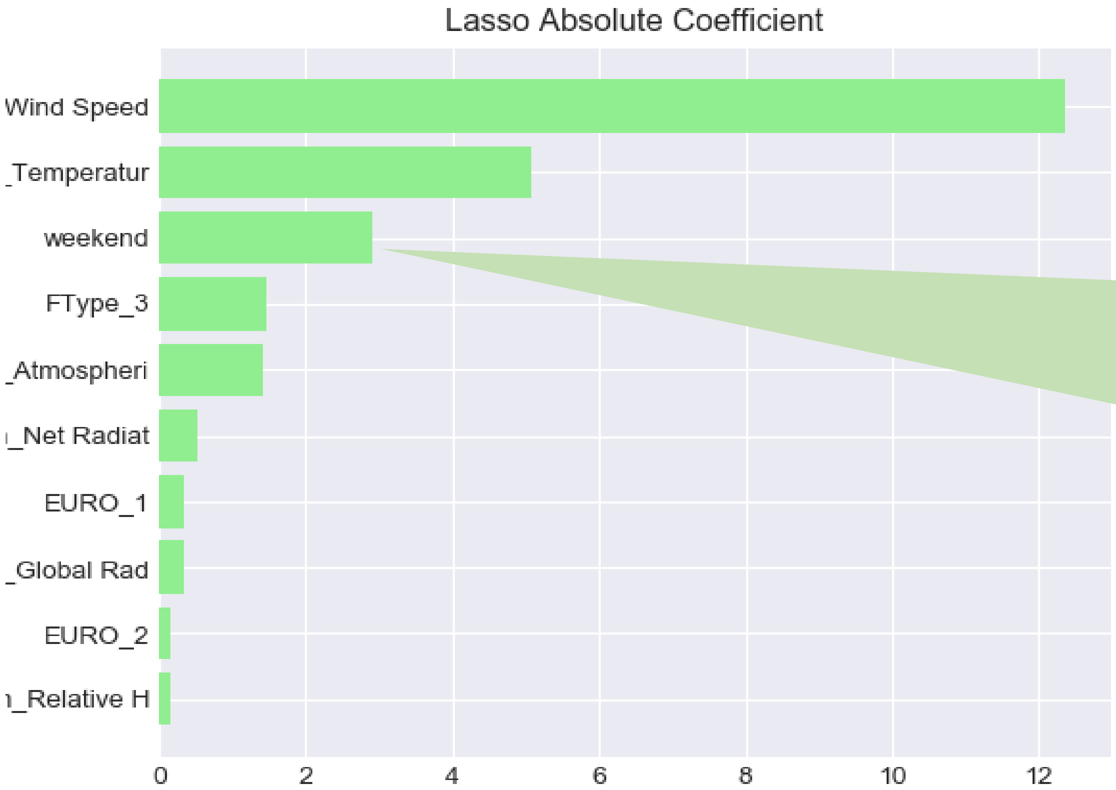
Feature Selection



Global Radiation was highly correlated with Net Radiation, had the highest MSE in solo linear regression, and universally rates below Net Radiation, so we will drop it

Lasso coefficients also indicate wind speed is most important

Feature Selection



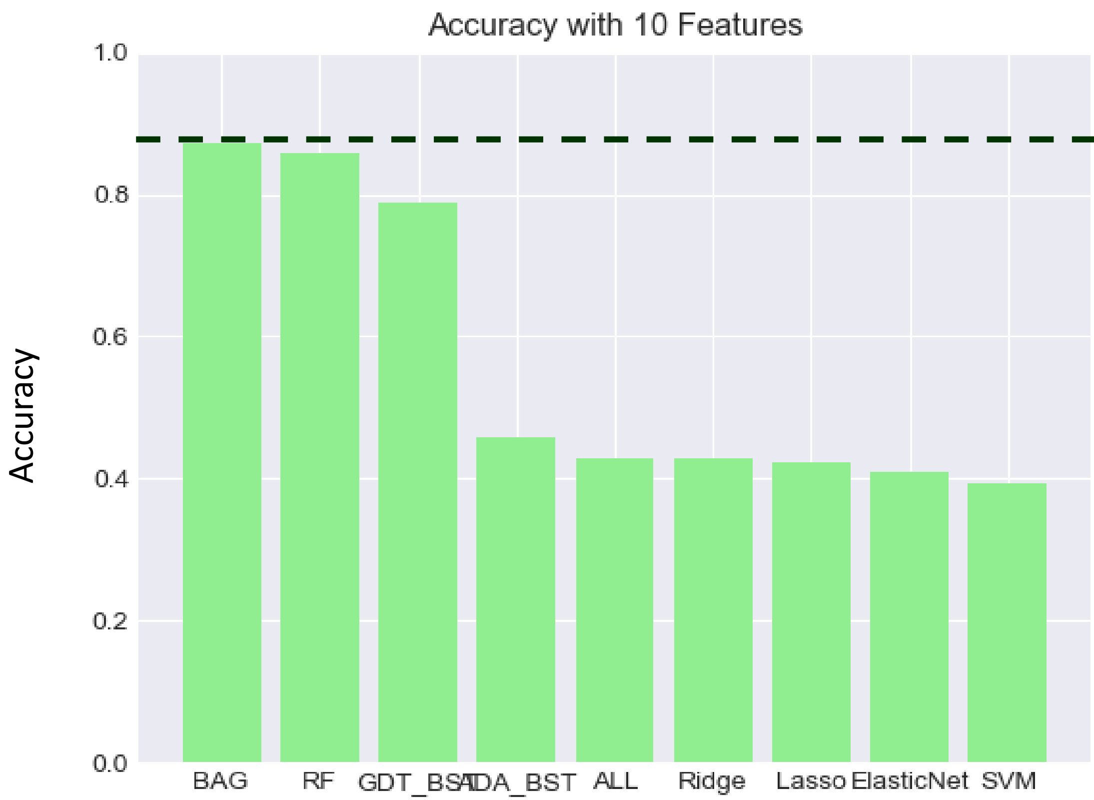
Weekend is surprisingly high here even though it was ranked very low in feature importance. It is also the only temporal variable. It will be included in the final analysis

III. Model

I. Train/Test Split
II. Predicting AQI
III. Feature Selection
IV. Final Model

Using a set of 10 features only improves accuracy from 84% to 86%, but parsimony is enhanced

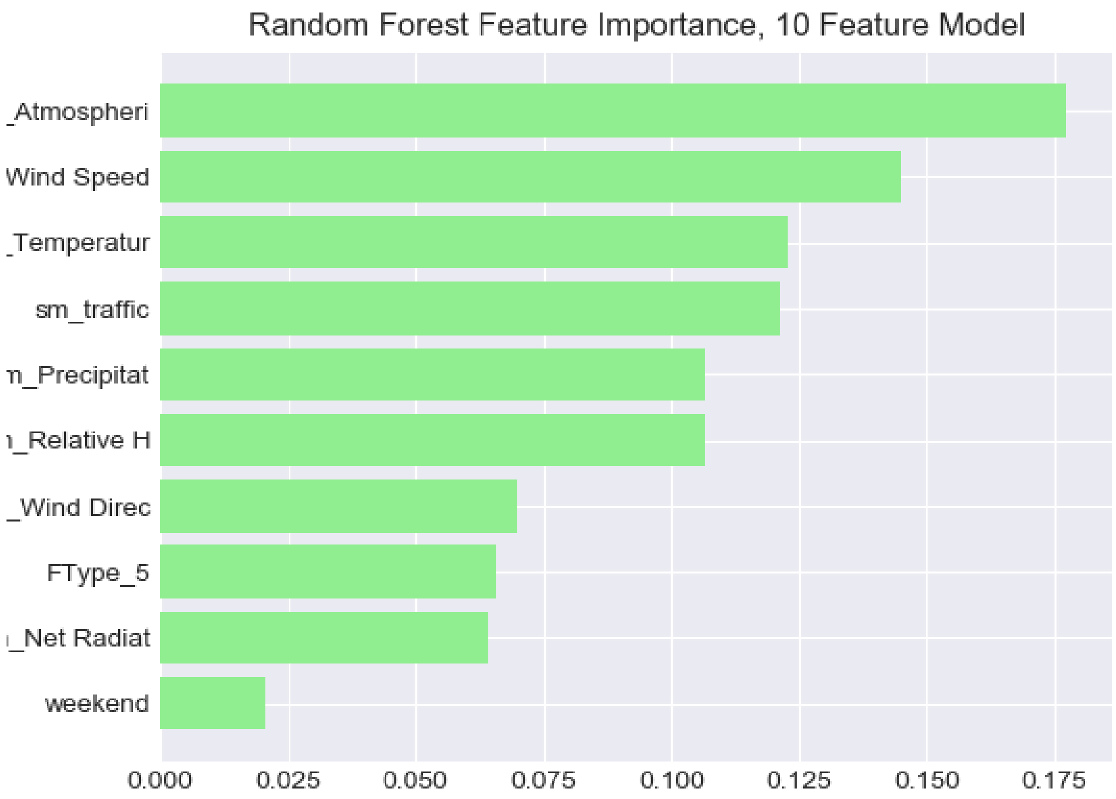
Final Model



86%
Random Forest Classifier

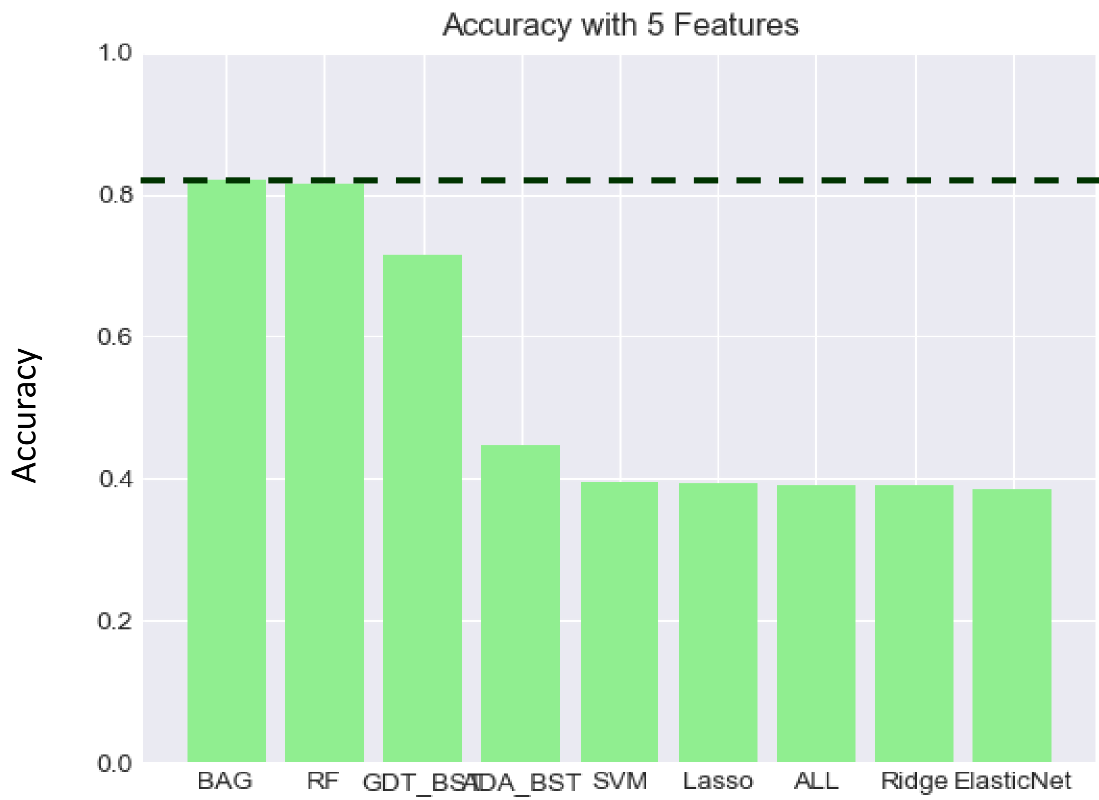
Weekend is still not considered important by the random forest model

Final Model



The model can halve its features to 5 with only a small decrease in accuracy

Final Model



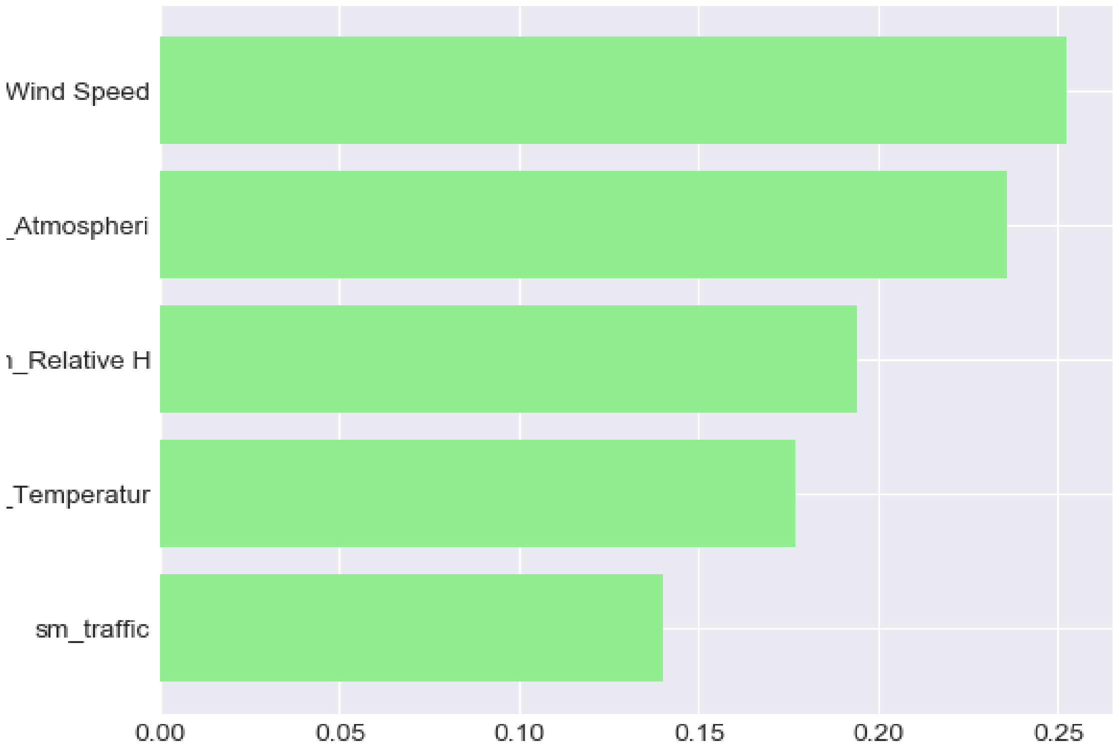
82%
Random Forest and Bagging Classifiers

Picture: https://www.123rf.com/photo_20237672_turtle-on-money.html

There is no great disparity in the importance of the remaining features

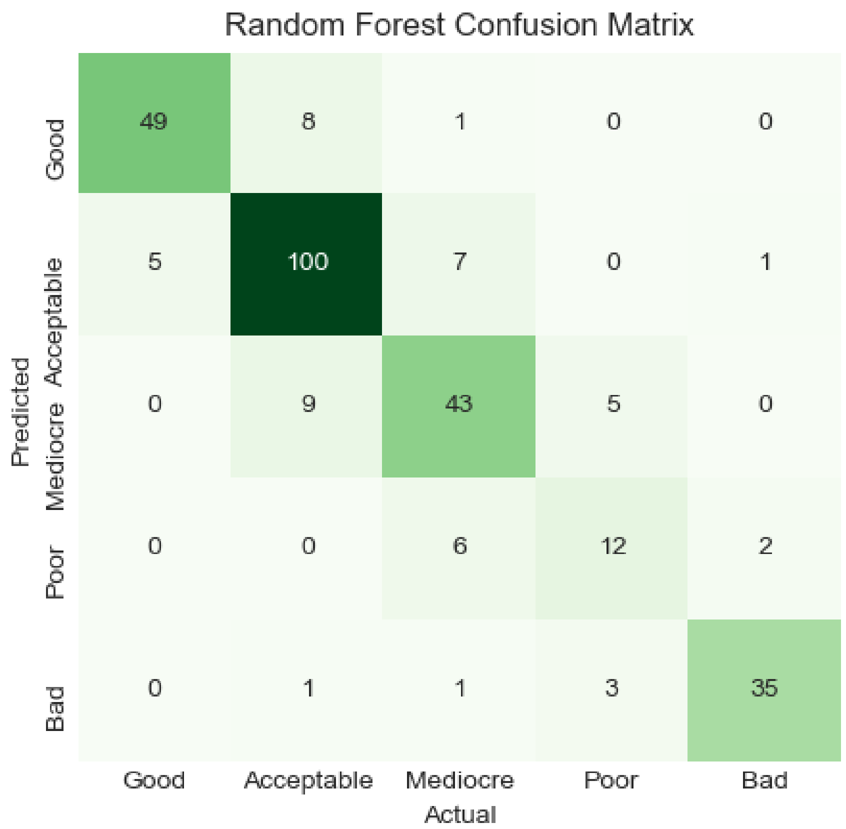
Final Model

Random Forest Feature Importance, Final Model



Random forest has 82% accuracy, misses by 2+ classes 1.4%, and missed by 3 classes 0.7%

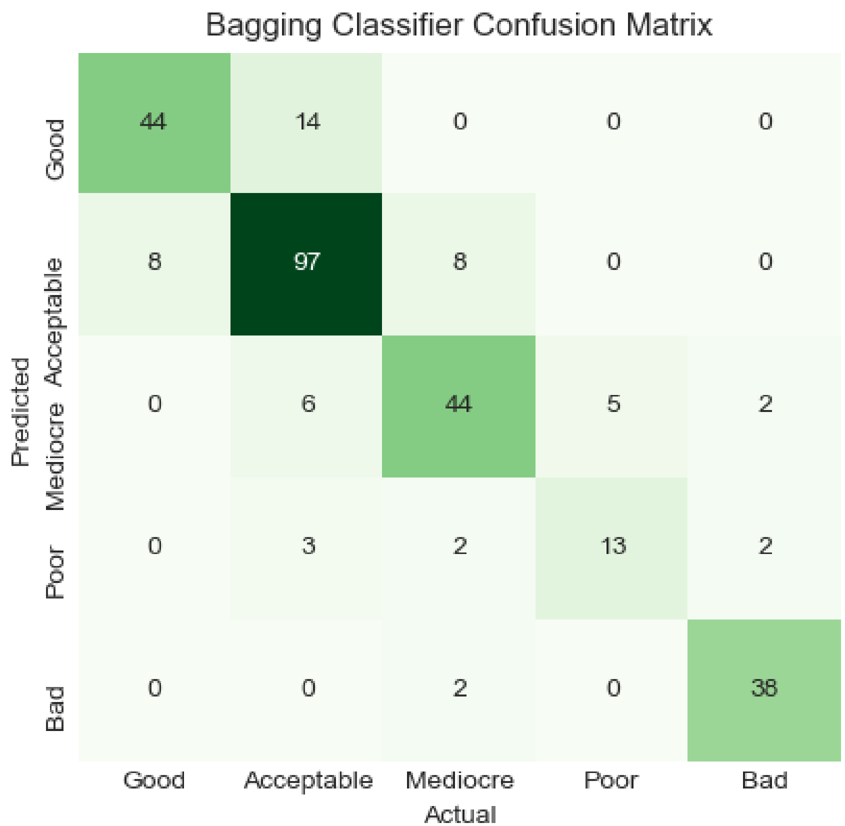
Final Model



Picture: https://www.123rf.com/photo_20237672_turtle-on-money.html

Bagging also has accuracy of 82%, misses by 2+ classes 2.4% of the time, but does not miss by 3 or more classes

Final Model



IV. Conclusion

The Air Quality Index is over 80%predictable on an hourly basis using mostly weather variables and ensemble classifiers

Conclusion

- Only one traffic variable was included in the final model, and it was about as important as each of the four weather variables present

- Ensemble models are better at predicting air quality from weather and traffic

- This does not rule out that variables like the number of hybrid vehicles and diesel particulate filters could be valuable to air quality long term

V. Recommendation

Decision makers should use a bagging model based on a set of 5 variables to predict the AQI in Milan

Recommendation

- Bagging has accuracy of nearly double that of more interpretable models using linear regression, and the incidence of large errors is less than random forest
  - Large overestimates of air quality might be more beneficial to avoid. A recommendation to go outside when air quality is Bad could be more harmful to health, for example

- Few variables reduces data requirements and makes clear what is important to the model in the absence of interpretable estimators

- The five variables are
  - Atmospheric Pressure
  - Wind Speed
  - Temperature
  - Relative Humidity
  - Traffic

Data collection should continue, and the model should be retrained regularly to overcome initial weaknesses in analysis

Recommendation

- A sequential train/test set was not practical due to small amounts of data, this approach will need to be retested to see if it is robust going forward

- The model will also need to be revisited with data for other seasons, since it has only considered November and December

- Finally, since many models have been tested, more data should allow a training, testing and validation split of the data to prevent overfitting to the test data