Collegio Carlo Alberto
UNIVERSITÀ DEGLI STUDI DI TORINO

Predicting Air Quality in Milan

blurb

Dan Herweg

May 2019

Contents

Summary sentence

Summary

- A summary of results
- Pursuing hypothesis that it will work

- This deck is a brief documentation of the project

II. Background

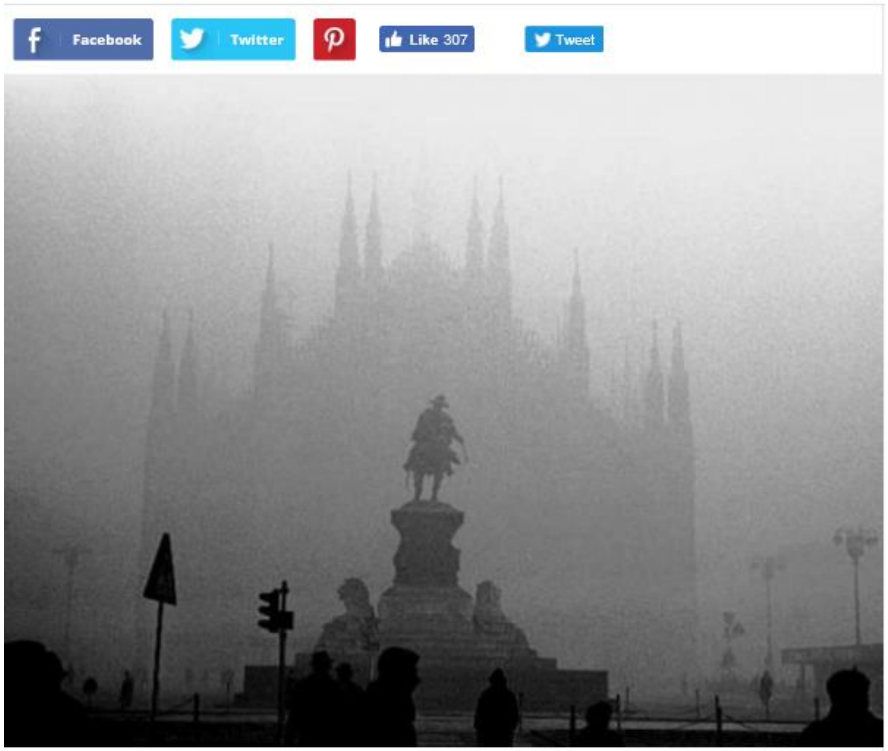
I. Motivation

Milan has poor air quality

Motivation

Milan has second worst smog in Europe – WHO

30 Jan, 2018



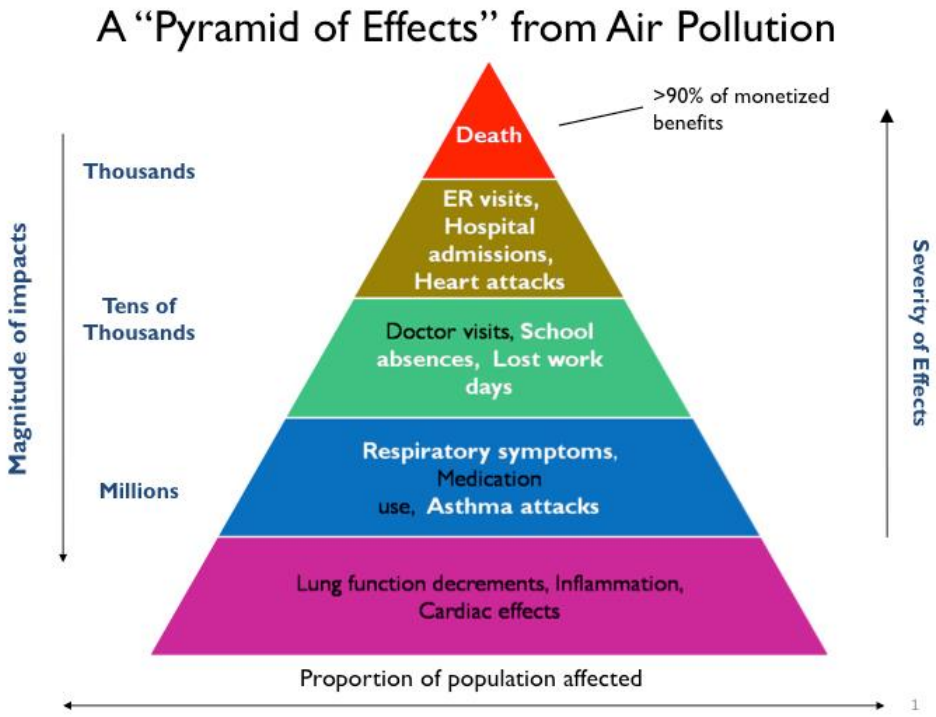
A report by the World Health Organization has placed Milan just behind Turin and just before Naples as the three European cities with the worst levels of atmospheric pollution.

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

III. Data

I. Methodology
II. Feature Construction
III. Data Exploration

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

Methodology

- The process will be the following:
  1. Features will be constructed, including the target
  2. Features will be selected based on a variety of dimensionality reduction techniques and intuition
  3. Train and test sets (Validation?) will be created
  4. Multiple prediction models will be trained and tested with these variables
  5. The model with the most accurate predictions in the test period will be selected

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

Feature Construction

- The data had three main sources
  - Weather Sensors
  - Traffic Gates
  - Pollution Sensors (used to calculate AQI)

- Sensors measuring the same thing were averaged

- Source of this data was X

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

Missing data was an issue

Feature Construction

- We wished to predict on the hourly level of data, which was not always available. Additionally, in some periods data was just missing. The data was either recorded hourly for the whole period, or imputed to hourly with the appropriate covariates

- Noisy time series data with potential for measurement errors was smoothed (alpha=.2)

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

Some features were constructed somewhat arbitrarily

Feature Construction

- The vehicle length was selected by visually guessing where a tri modal distribution was best separated into small, medium and large cars

- We took the 15 minute traffic counts

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

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

Feature Construction

- Everybody's favorite equation

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

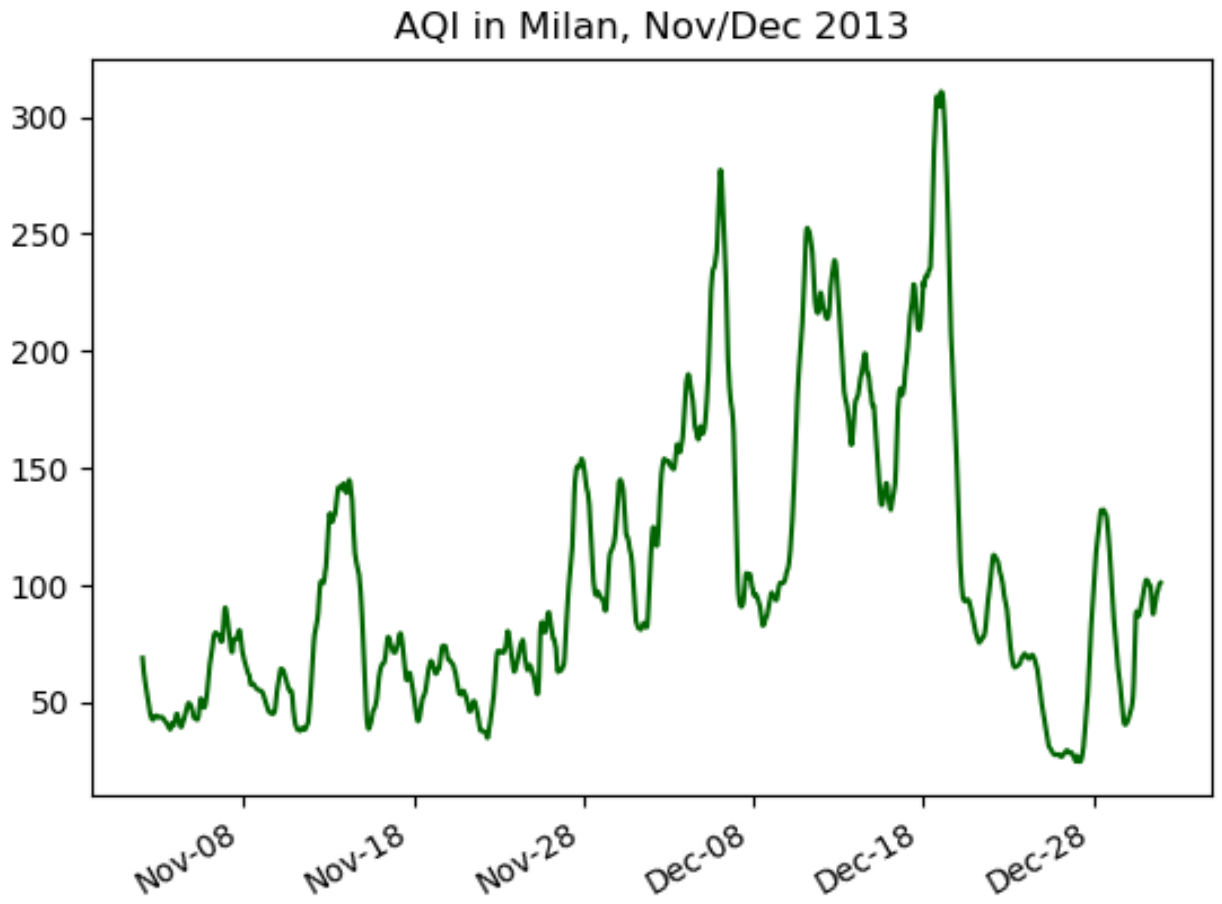
- Graphic of Scale from poor to good

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

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

Air Quality Index score

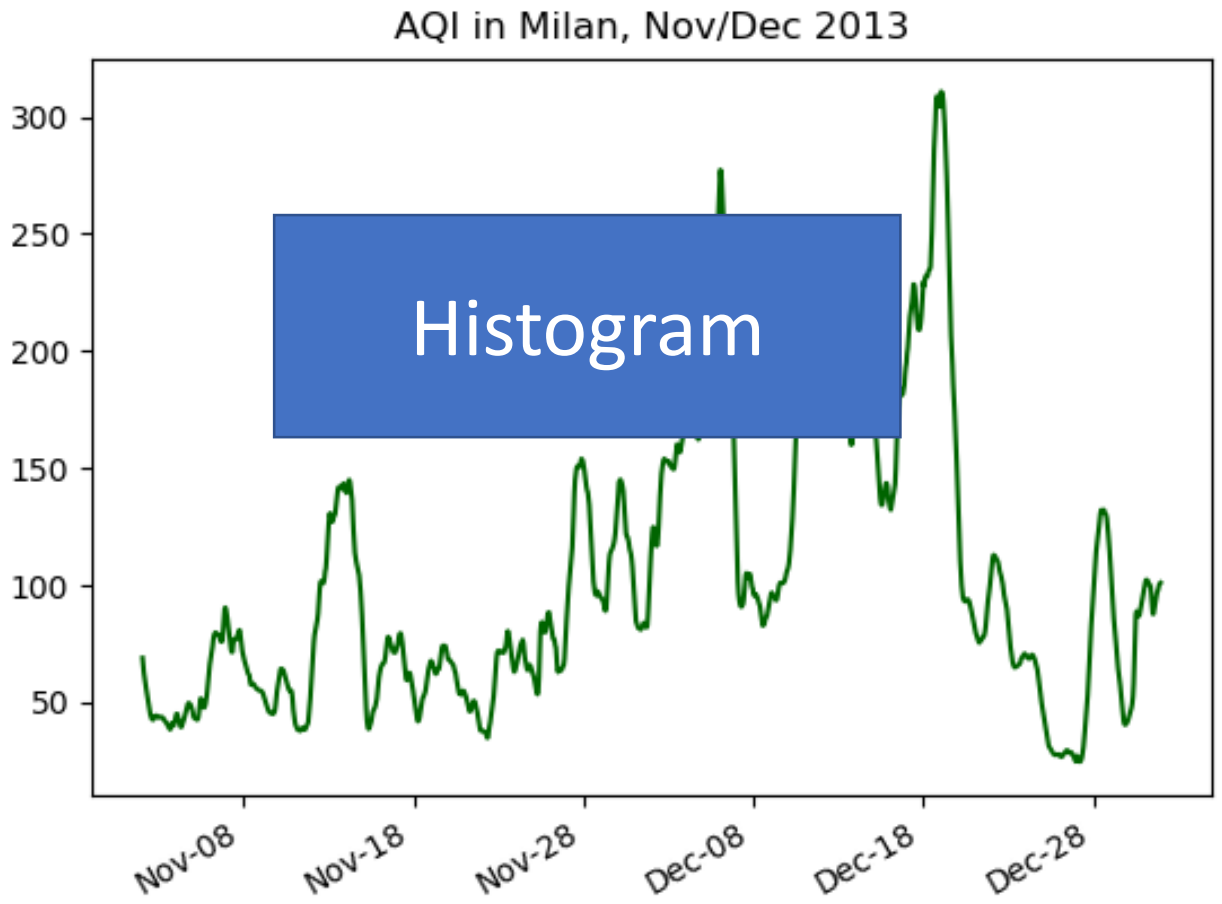
Feature Construction



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

Air Quality Index score distribution

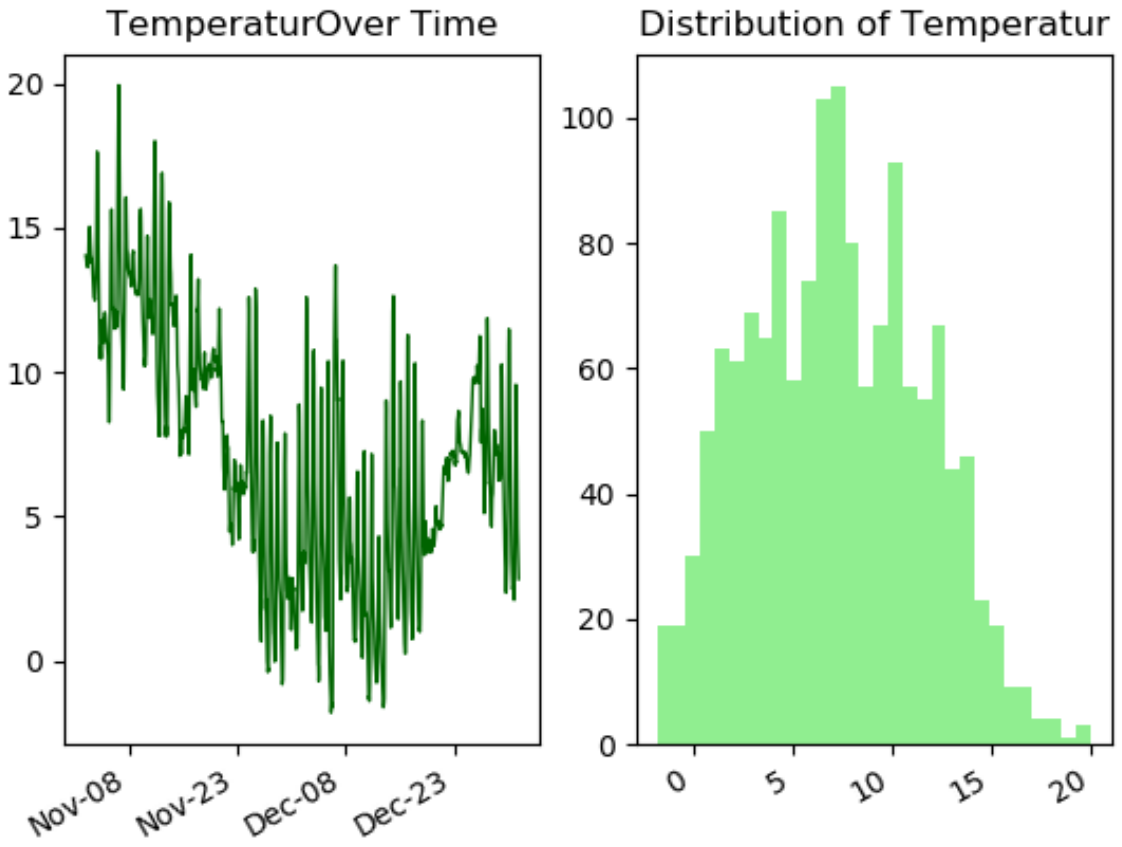
Feature Construction



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

Temperature had a normal distribution and declined into year end

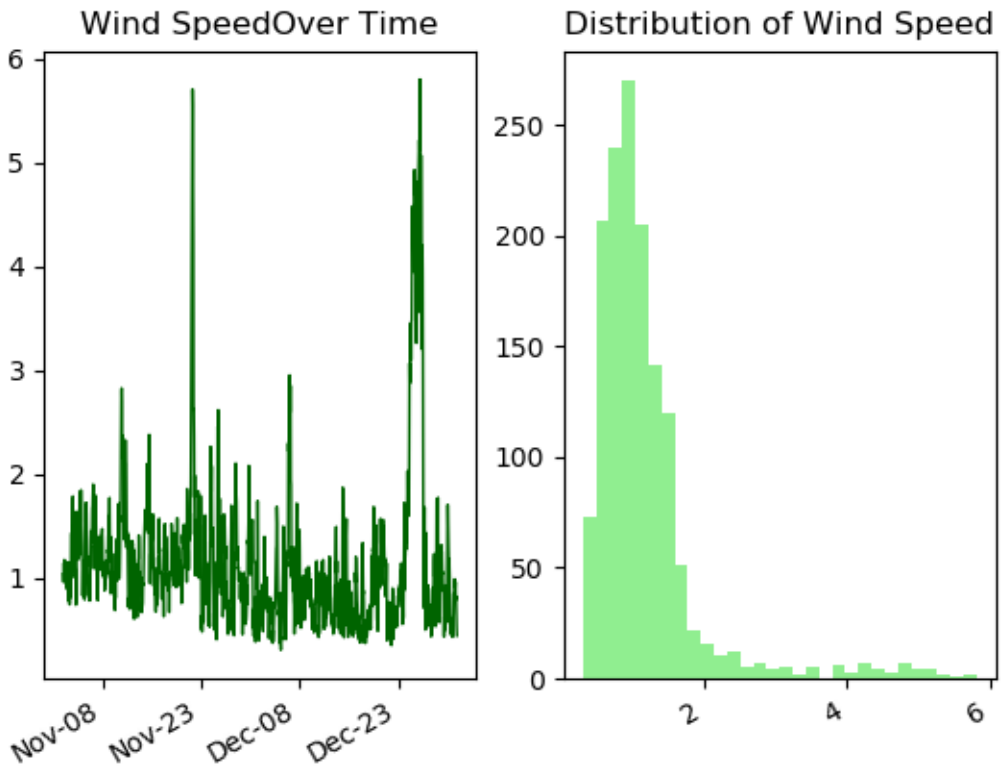
Data Exploration



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

Wind speed was very right skewed, with large peaks

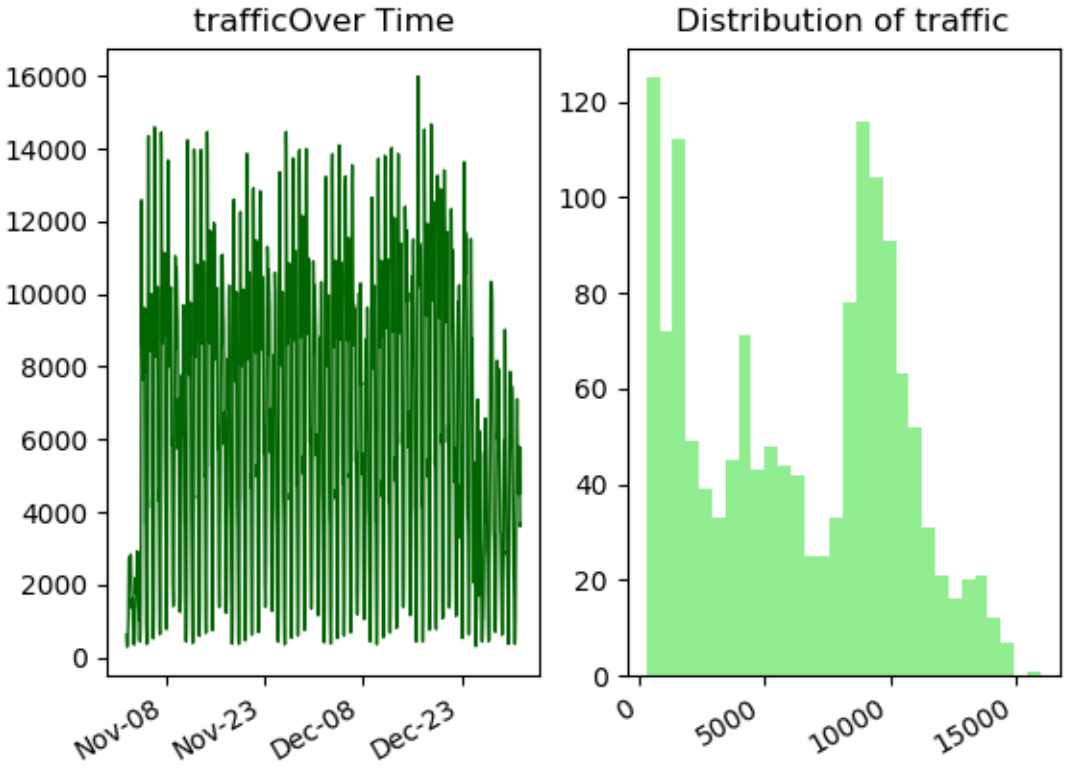
Data Exploration



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

Traffic had obvious daily/weekly pattern and a “tri-modal” distribution

Data Exploration

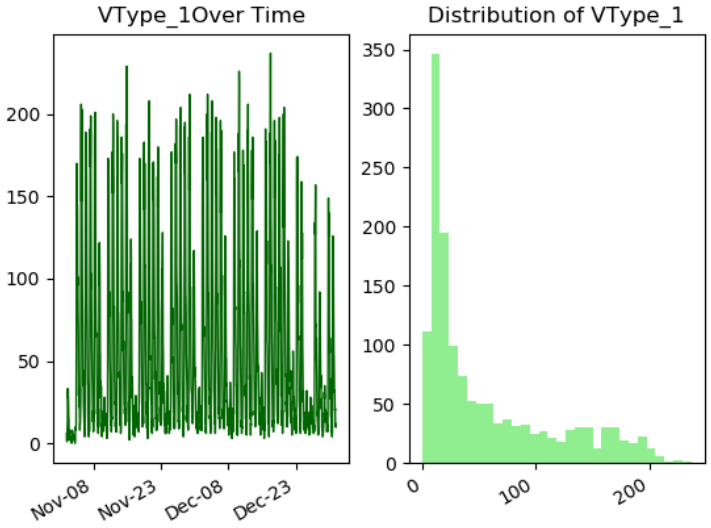


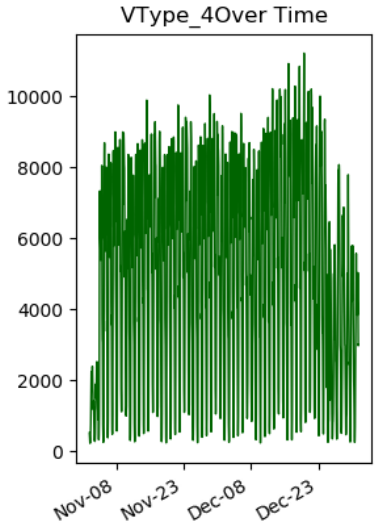
Picture: https://www.123rf.com/photo_20237672_turtle-on-money.html

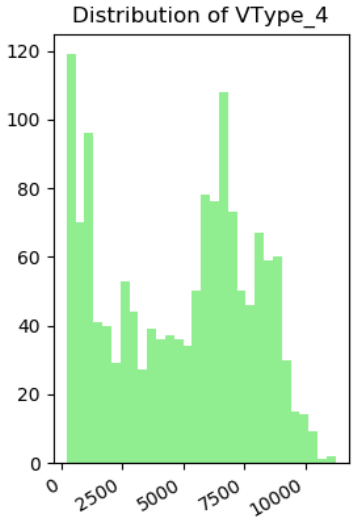
Different Vehicle types had different distributions

Data Exploration

- Can see weekends in 1
- Dist more bimodal or uniform in 4, more power lawish in 1
- What are the vtypes and can we tell a story?





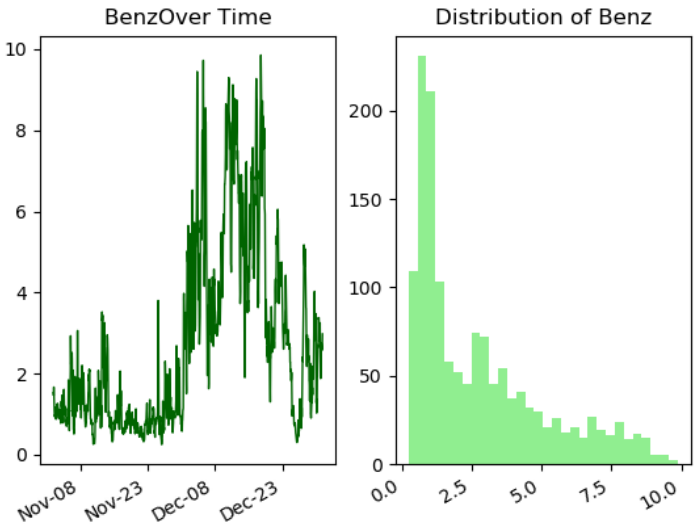


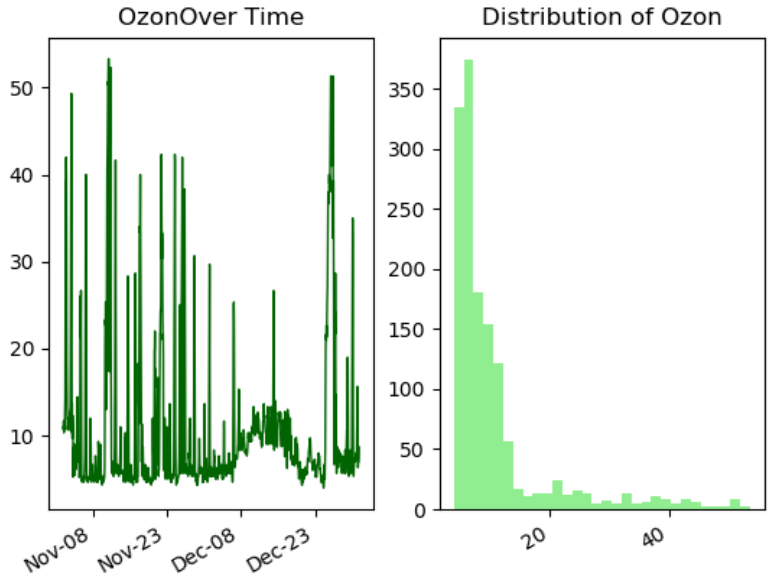
Picture: https://www.123rf.com/photo_20237672_turtle-on-money.html

Pollutants looked a lot like AQI pollutants, Ozone least so

Data Exploration

- Show example of correlated, least correlated, most anticorrelated pollutants

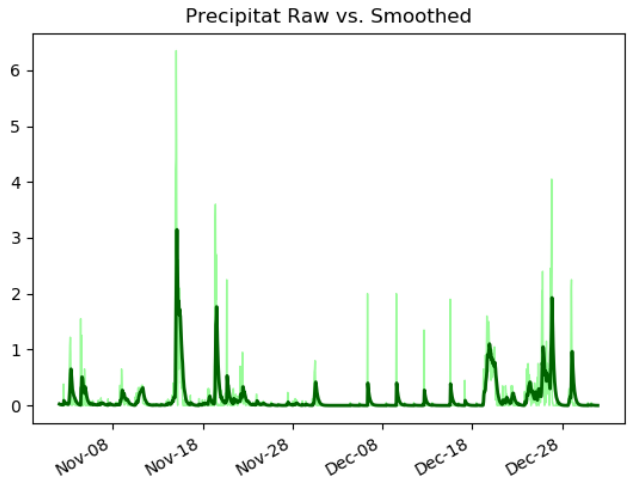


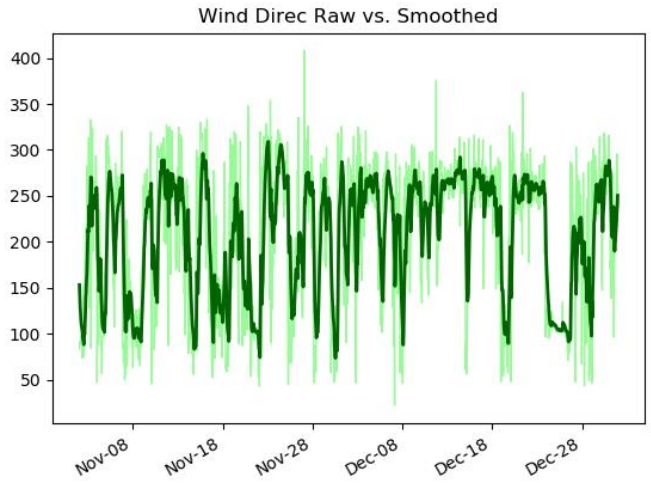


Picture: https://www.123rf.com/photo_20237672_turtle-on-money.html

Smoothing eliminates noise from the features

Data Exploration



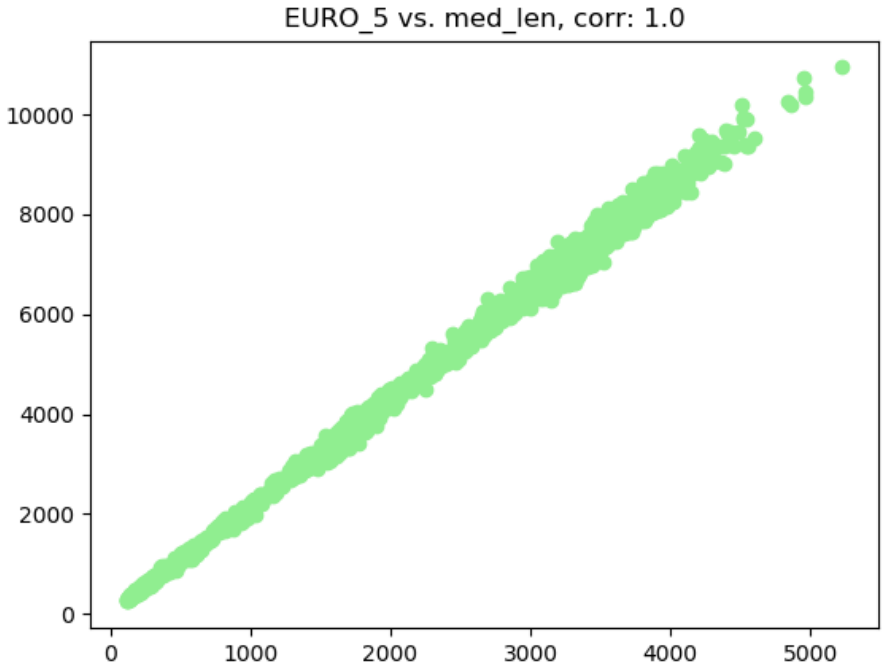


Picture: https://www.123rf.com/photo_20237672_turtle-on-money.html

Apparently Euro5 vehicles tend to be medium length

Data Exploration

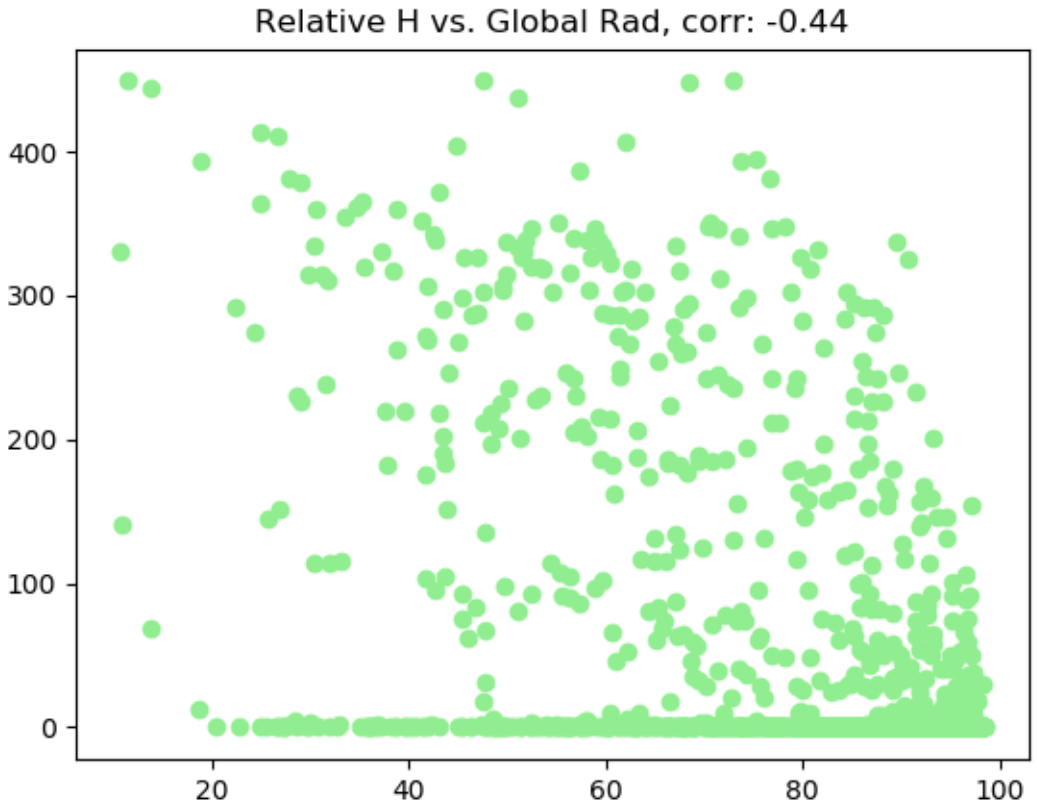
- Though correlation rounds to 1, coefficient is about .5 so not 1:1
- Still, a good model should probably not have both of these variables



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

There were negatively correlated variables, but not as striking

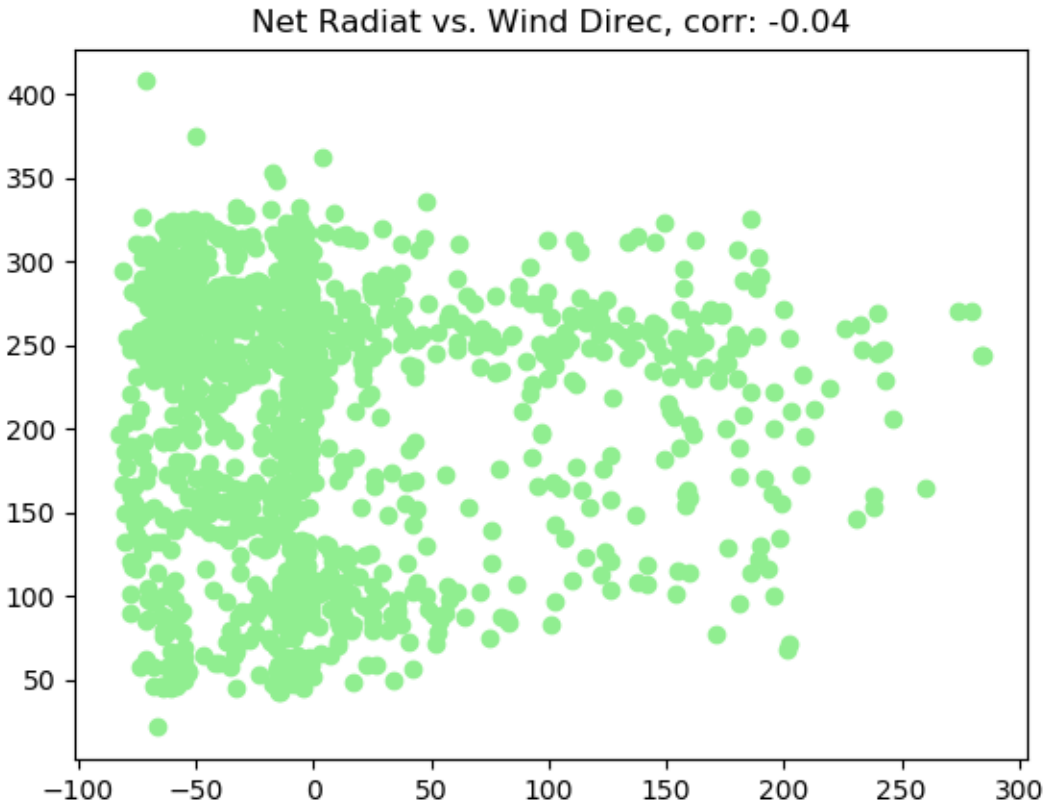
Data Exploration



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

Some variables not correlated

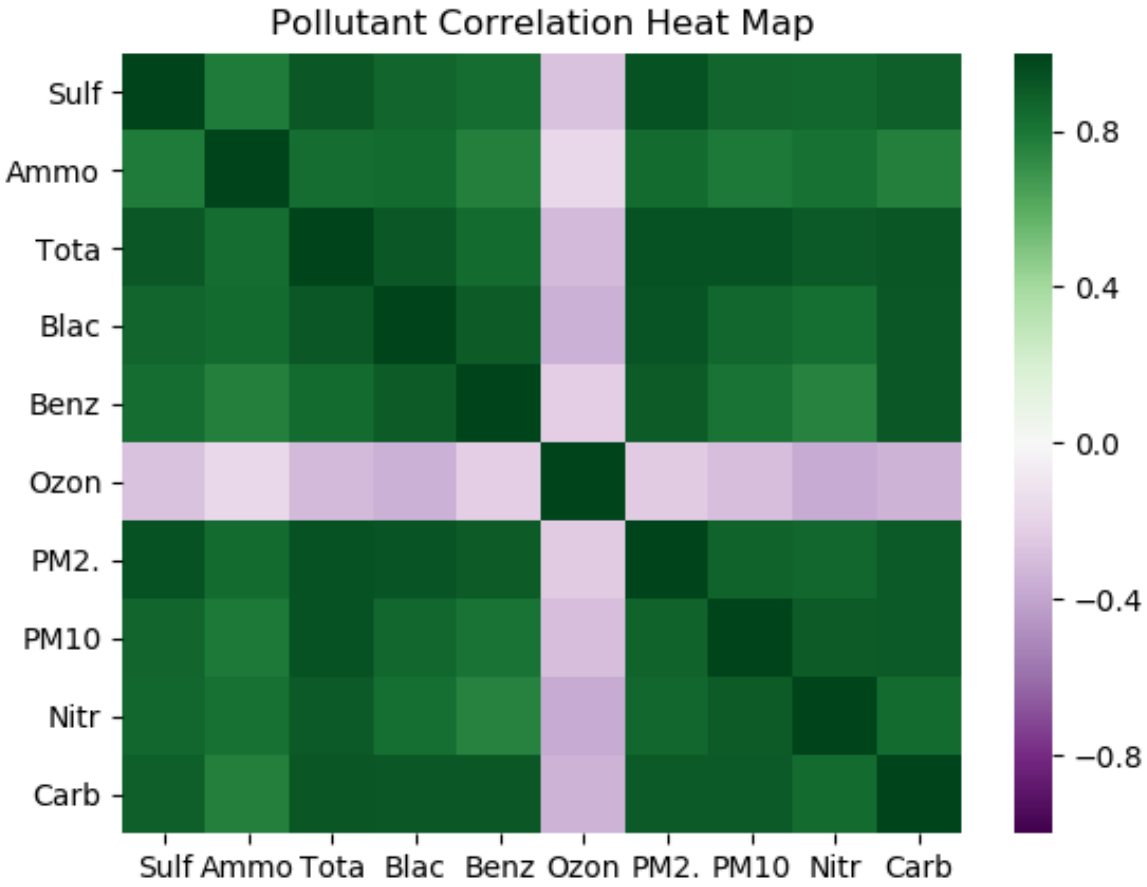
Data Exploration



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

Pollutants are very correlated except ozone. This is good for imputation

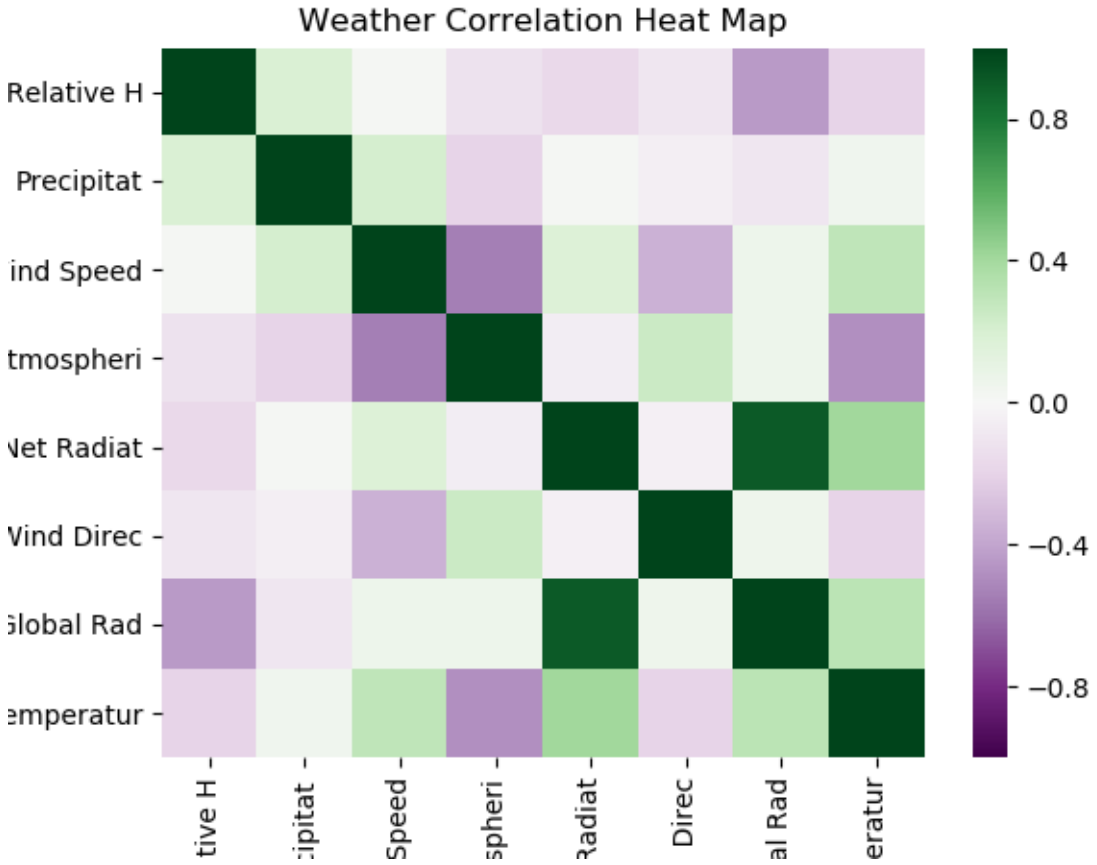
Data Analysis - Multivariate



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

Weather data is not super correlated except the two radiation measures, which is to be expected

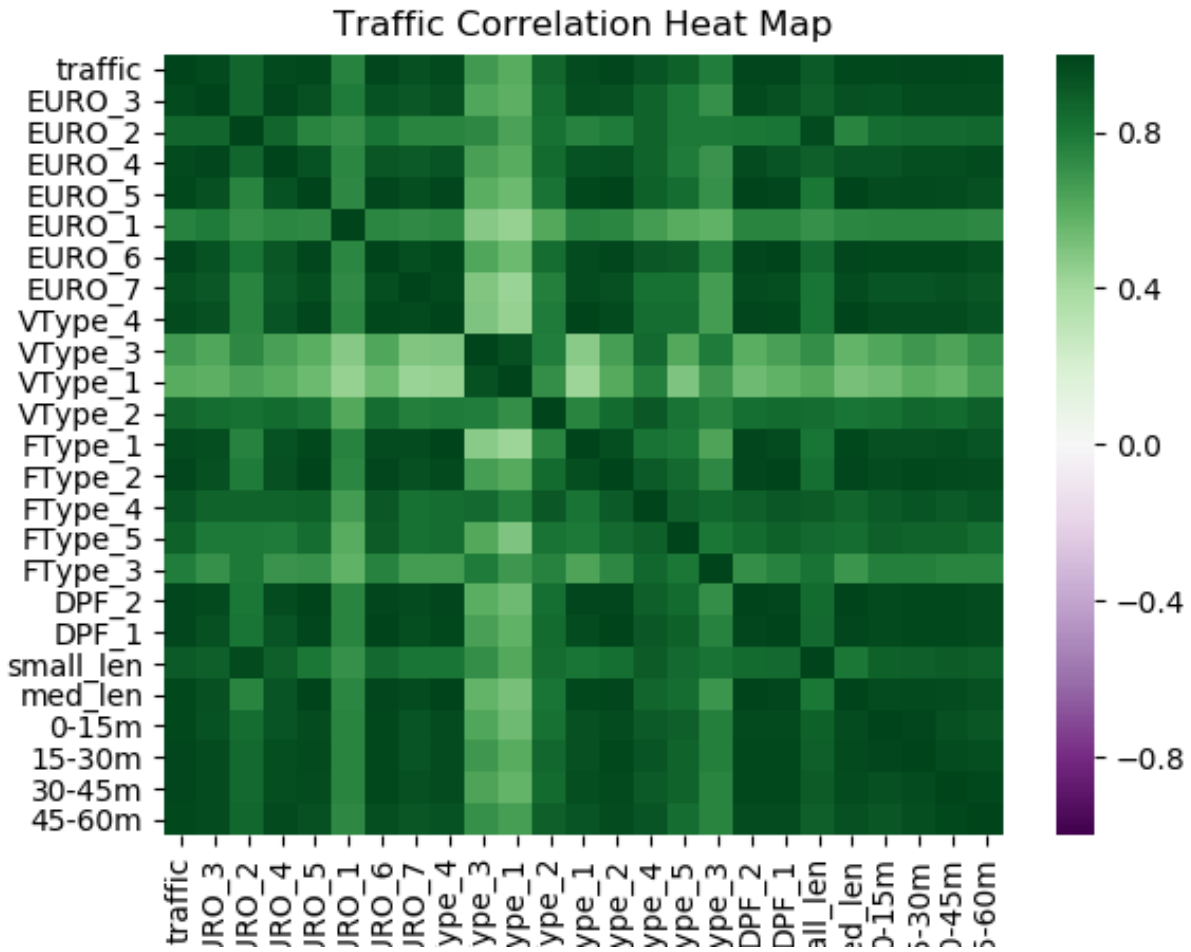
Data Exploration



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

Traffic data is more correlated, and all positively correlated

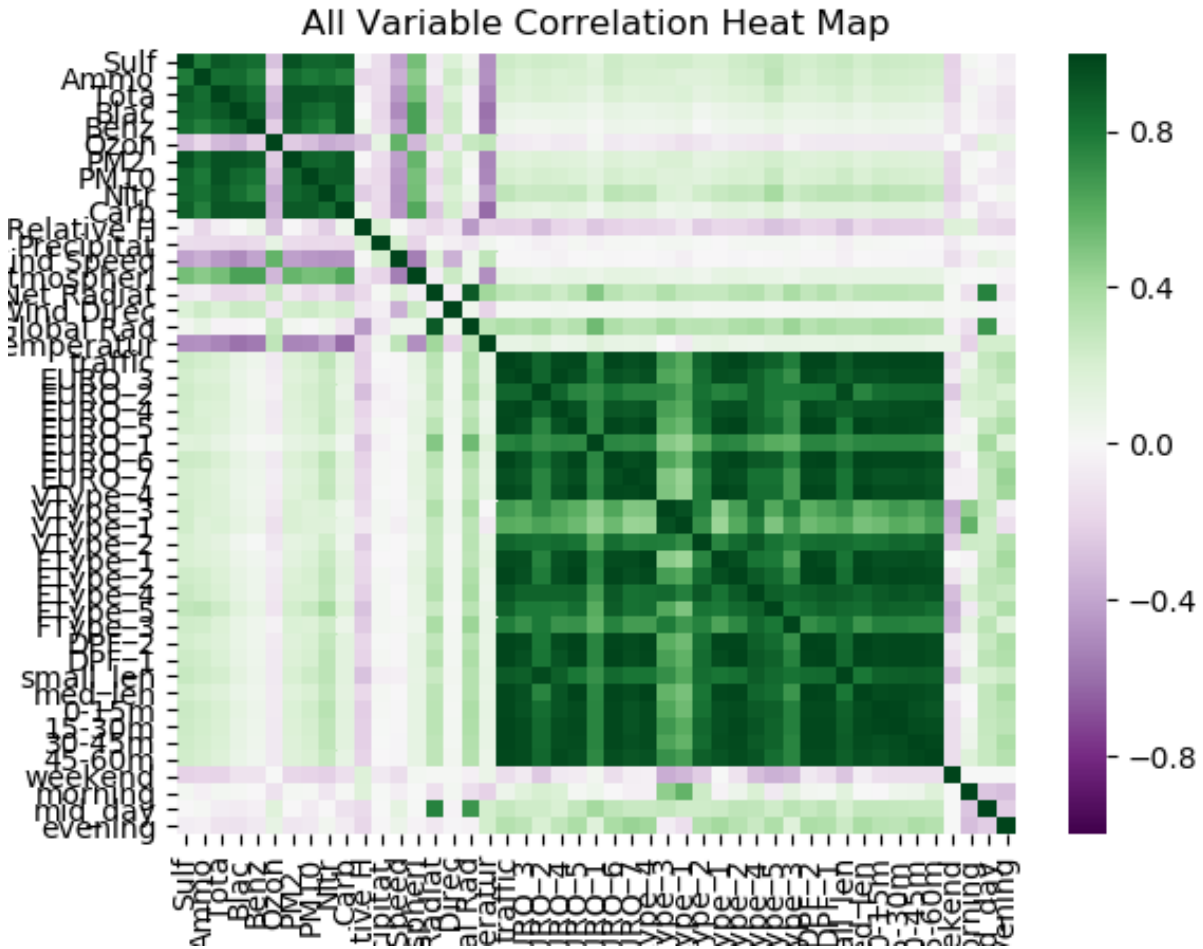
Data Exploration



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

All data (should make this just features?)

Data Exploration



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

IV. Model

I. Feature Selection

II. Prediction

III. Bonus: Is smoothed data better?

We have too many variables to drive an actionable result

Feature Selection

- Some techniques to do dimensionality reduction:
  - Penalization: Lasso
  - Embedded: Tree-based methods
  - Algorithmic: Stepwise selection
  - Heuristic: A method I made up

- After these we may apply a Voting: What variables are common to these methods?

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

Variables from each method

Feature Selection

- Table

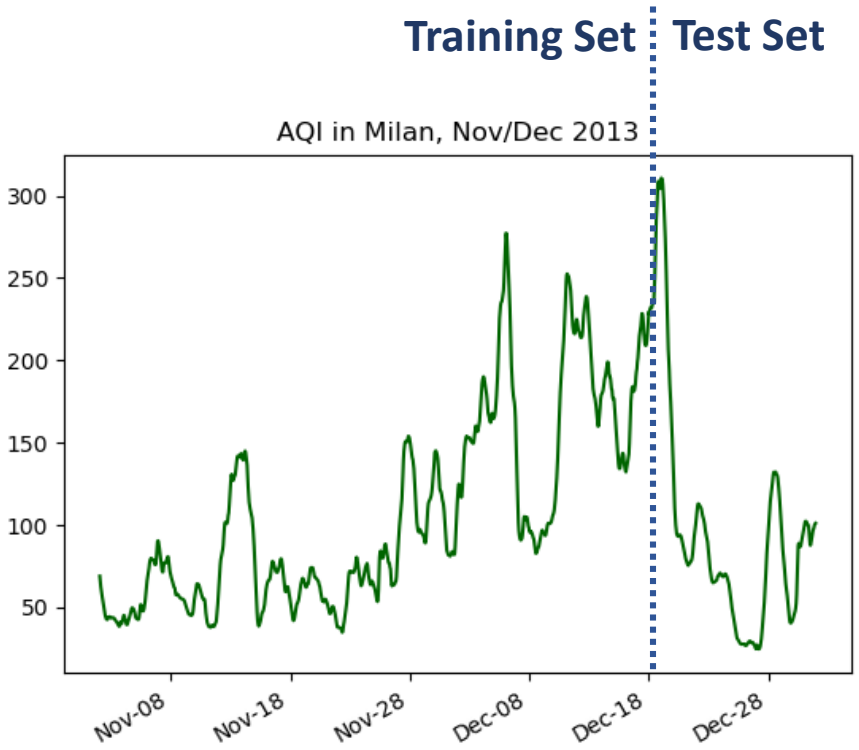
- Lets select the variables we want to use and try some methods that don't select for variables as well

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

Before we predict we need a train and test set

Prediction

- Distribution of classes in train test
- Strengths weaknesses tradeoff
- Validation?
- Visualization – histogram,
  - grey version of AQI to emphasize
  - dividing line?



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

Picking the modal training classification for the test data sets a “naive” benchmark

Prediction

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

Each of our variables alone does not match the “naive” accuracy

Prediction

- Bar chart of variables’ accuracy

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

Each of our variables alone does not match the “naive” accuracy

Prediction

- Time series visualization of test data

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

Using the variables together yields better results, best in random forest probably with X%

Prediction

- asdf

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

Using the variables together yields better results

Prediction

- Visualization of test set

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

Final selection of data and model

Prediction

- asdf

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

Rerunning the data on the nonsmoothed versions was better or worse

Noisy Data

- Wow I am surprised

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

V. Conclusion

X model with Y variables performed prediction best

Conclusion

- Why it was selected
- Weaknesses
- Noisy data conclusion

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

Resources

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