Analysis of Air Quality and Weather in the Baltimore Area

SPRINGBOARD INTERMEDIATE DATA SCIENCE: PYTHON

AUTHOR: FAISAL MAHMOOD

MARCH 2019

Introduction

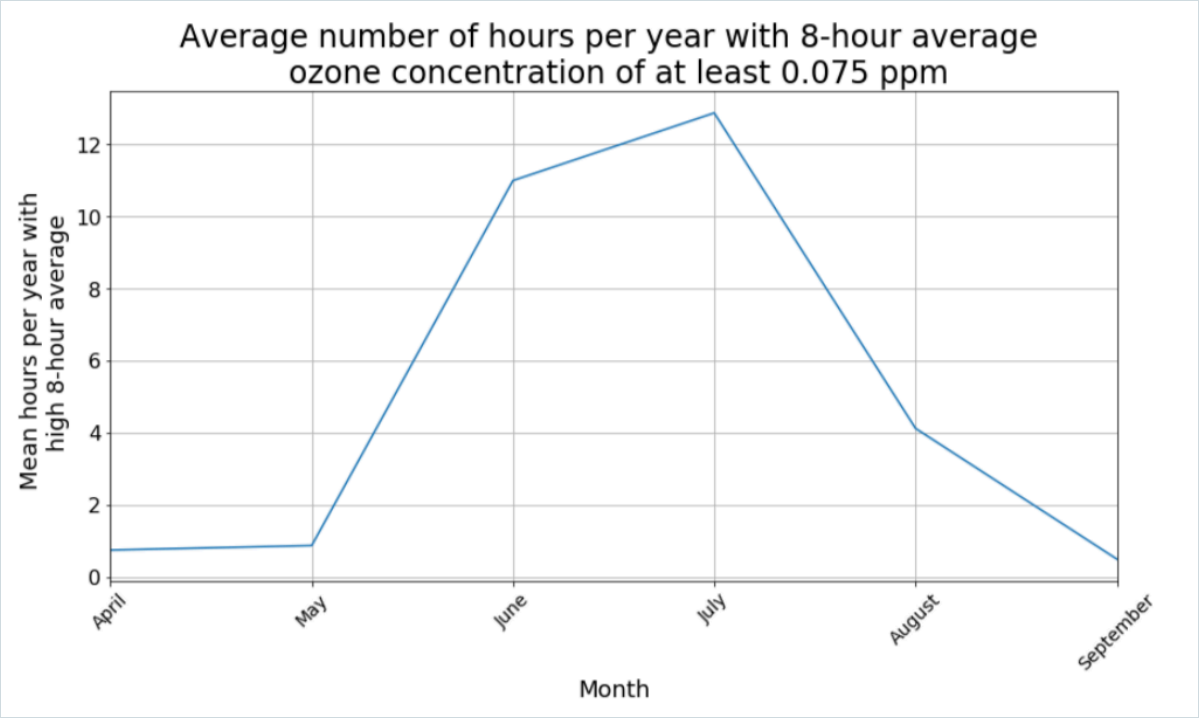
- Air quality affects the livelihood of millions of people, especially the vulnerable.
- Ground-level ozone is one of six pollutants being regularly tracked for determining air quality.
- Predictive model will be built to predict ozone levels based on hourly weather observations.
- Model is to be used by potential stakeholders, such as health or wellness firms, or summer tourism businesses, to determine if ozone levels will be at a healthy level. An effective model would benefit the clients.
- Ozone concentration of 0.075 parts per million (ppm) averaged over a period of 8 hours can affect sensitive groups.

Data Acquisition and Wrangling

- Two sets of hourly data ranging from 2006 to 2015 were obtained.
  - One set from the Environmental Protection Agency (EPA) consisted of hourly ozone concentration data for a station in Essex, MD.
  - The other set consisted of hourly weather observations for Martin State Airport, MD, located 3 miles from the Essex station where ozone levels were measured.
- The goal was to combine the datasets to ensure that hourly observations would be available for the duration of the period being studied.
- The merged dataset consisted of hourly data from 2008 to 2015.

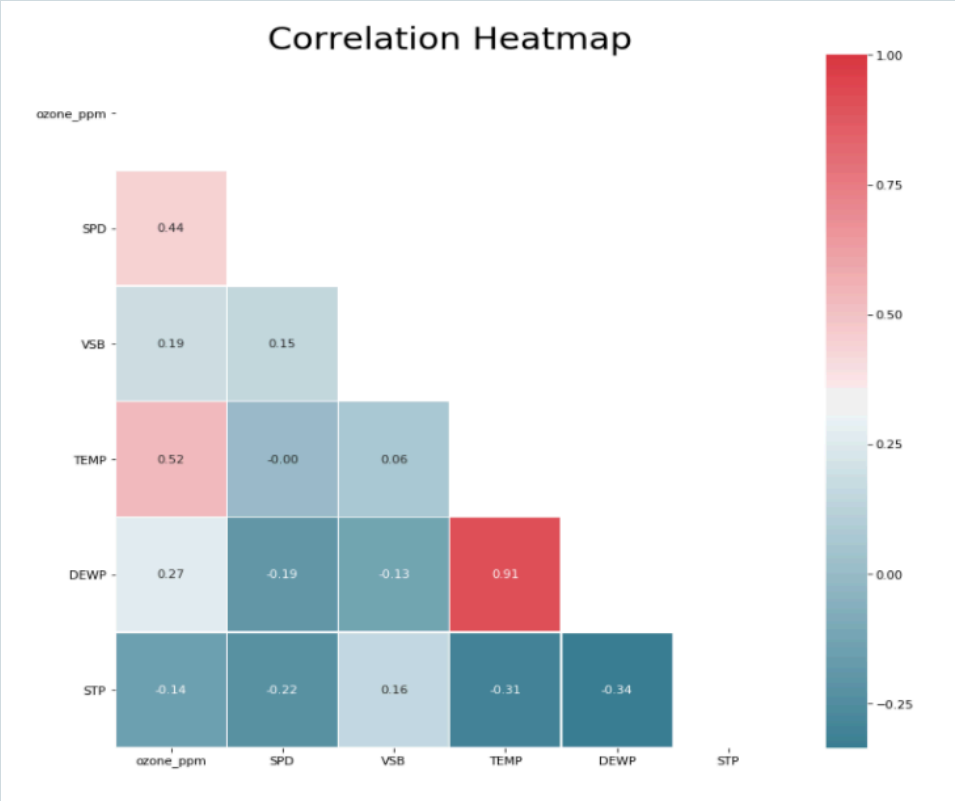
Exploratory Analysis

- Since ozone levels of 0.075 ppm or higher (averaged over an 8-hour period) can affect sensitive groups, the prevalence of those levels was plotted on a monthly basis.
- Hazardous ozone levels were found to be the most common from June to August
- Spring and early fall also had occasional days with high ozone levels.



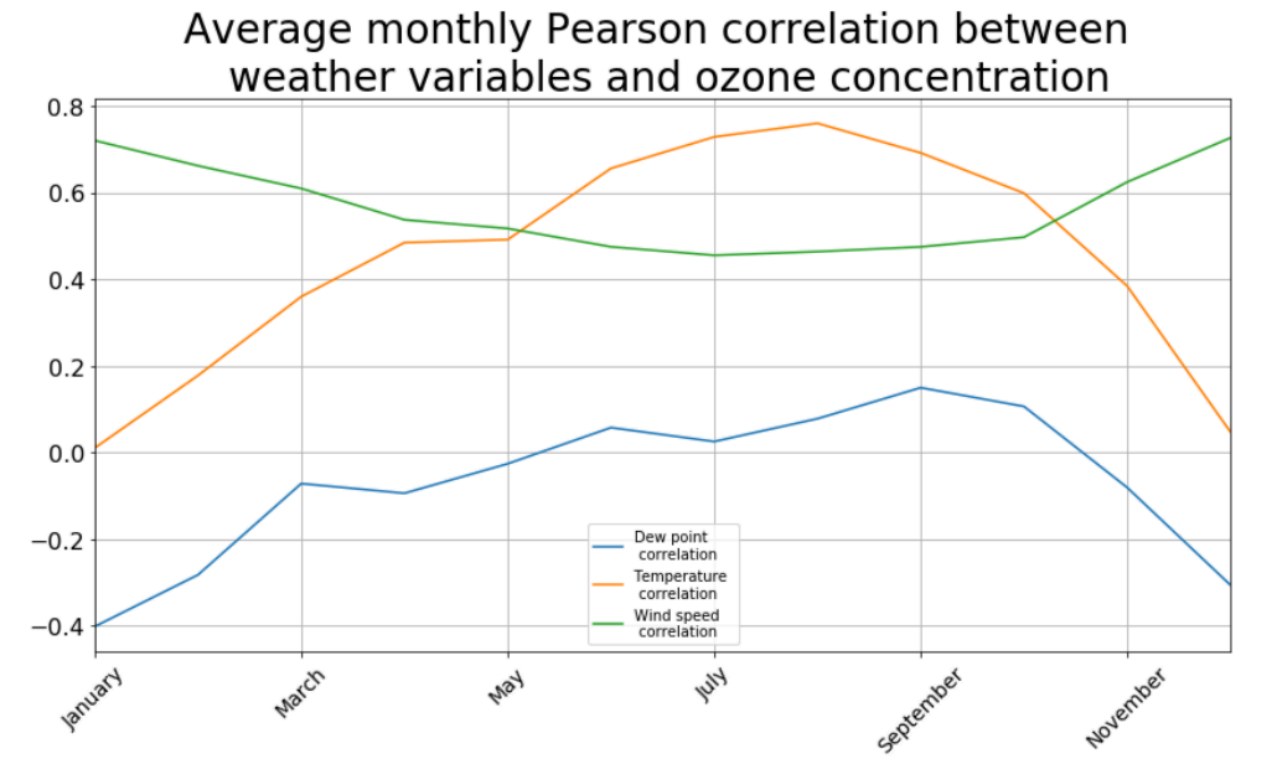
Exploratory Analysis

- Correlations among the variables were evaluated. Correlations of ozone levels to the remaining variables were of particular interest.
- Temperature and wind speed had the highest correlations with ozone concentration.
- What are the seasonal variations of these correlations?



Exploratory Analysis

- Temperature has the highest correlation with ozone in the summer and early fall, while wind speed has the highest correlation in the cooler months.
- Are these correlations statistically significant?



Inferential Statistics

- Average monthly pearson correlations with ozone concentration were calculated for all weather variables.
- Those correlations were evaluated for statistical significance.
- Null hypothesis assumed no relationship, and alpha was set to 0.01 due to large size of dataset.
- 54 out of 60 tests (12 monthly tests for all 5 weather variables) had p-values below the value of alpha.
- Results suggest that for 54 of those tests, a statistically significant relationship exists with the ozone concentration, even among correlations that were very low.
- With such a large dataset, a statistically significant model can be constructed even with very low predictability. Statistical significance does not always reveal much information about *practical* significance.

Baseline Modeling

- Dataframe was split to training and test sets.
  - Training set consisted of data from January 2008 to November 2015.
  - Test set consisted of data for December 2015.
- A baseline linear regression model was built to attempt to predict ozone levels in the test set.
- Baseline model had a very low R^2 accuracy score for the test set, so a better model was needed.

Baseline Model Accuracy Scores			
Score Metric	R^2 (correlation)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)
Training Set Score	0.582197679433	0.0113865390885	0.00903829888978
Test Set Score	0.0432025252226	0.00903532790804	0.00742406213061

Extended Analysis

- Lasso and Ridge regularization were both attempted, and Random Forest regressors were investigated. However, even the best available models did not offer any meaningful improvements from the baseline model.
- Perhaps an analysis with test set focused on a single month was not going to produce an effective model.

Baseline and Regularized Model Accuracy Scores			
Score Metric	R^2 (correlation)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)
Baseline Training Score	0.582197679433	0.0113865390885	0.00903829888978
Baseline Test Score	0.0432025252226	0.00903532790804	0.00742406213061
Lasso Test Score	0.0432025252194	0.00903532790806	0.00742406213071
Ridge Test Score	0.0432025252226	0.00903532790804	0.00742406213061
Random Forest Test Score	0.089999847083	0.00881159729491	0.00678782734193

Extended Analysis: Summer months

- Similar analysis was done with an exclusive focus on the summer months (June to September for the purpose of this analysis).
- Training set consisted of data for the summers of 2008 through 2014. Test set consisted of data for summer 2015.
- Baseline model produced R^2 scores near 0.62 for both the summer training and test sets. Mean absolute error (MAE) and root mean square error (RMSE) were both similar.
- Additional adjustments to the baseline model produced no meaningful improvements, because there was no overfitting.
- For the baseline model, 95% of residual values ranged from -0.019 to +0.023 ppm. No other regularized model had any improved residual spread.

Baseline and Regularized Model Accuracy Scores for Summer Period			
Score Metric	R^2 (correlation)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)
Baseline Training Score	0.61954291927	0.0120153737564	0.0093943006521
Baseline Test Score	0.616979835465	0.0108932691041	0.00867100523515
Lasso Test Score	0.616979835465	0.0108932691041	0.00867100523515
Ridge Test Score	0.616979835465	0.0108932691041	0.00867100523515
Random Forest Test Score	0.594303997616	0.0112110884086	0.00875663544491

Conclusion and Ideas for Future Work

- Summer-only model was far more effective than a model trained on data from throughout the year, based on R^2 accuracy scores. Due to seasonal changes in weather, predictive models that account for those seasonal differences will produce far better results
- Ideas for future work may include the following:
  - Creating similar predictive models for other cities or regions.
  - Evaluating the concentration of pollutants other than ground-level ozone, and creating a similar model based on that data.
  - Fully analyzing all pollutants tracked by the EPA to create a classification model that predicts the general AQI (air quality index) by color code.

Recommendations for Client

- Baseline model for summer months is recommended, since summer is the most likely time for hazardous air quality.
- High correlation and relatively low error rate can help the clients use weather forecasts to devise a more effective way to predict air quality levels.
- If a general margin of error of 0.02 ppm is satisfactory for clients, then this model will be highly effective.