Capstone Milestone Report

I. Defining the Problem

Air quality is an issue that affects the livelihood of millions of people, especially those with respiratory issues, and at times it can affect the ability for people to spend time outdoors. In this report I will attempt to build a model to predict ozone concentrations in the air, based on hourly weather observations. Ozone is one of six pollutants whose atmospheric concentration is regularly tracked for determining air quality. Although ozone in the stratosphere plays a crucial role in protecting humans from the harmful effects of ultraviolet radiation from the sun, ozone near the ground is harmful for public health, especially among individuals with asthma or other respiratory issues.

II. Clients

For this report, my clients would be health and wellness firms, companies that rely on summer tourism, and others who have a vested interest in maintaining a healthy air quality. If an effective model from this analysis can be used to reliably predict the ozone concentration, then my clients will benefit financially and potentially save people from the health problems associated with poor air quality.

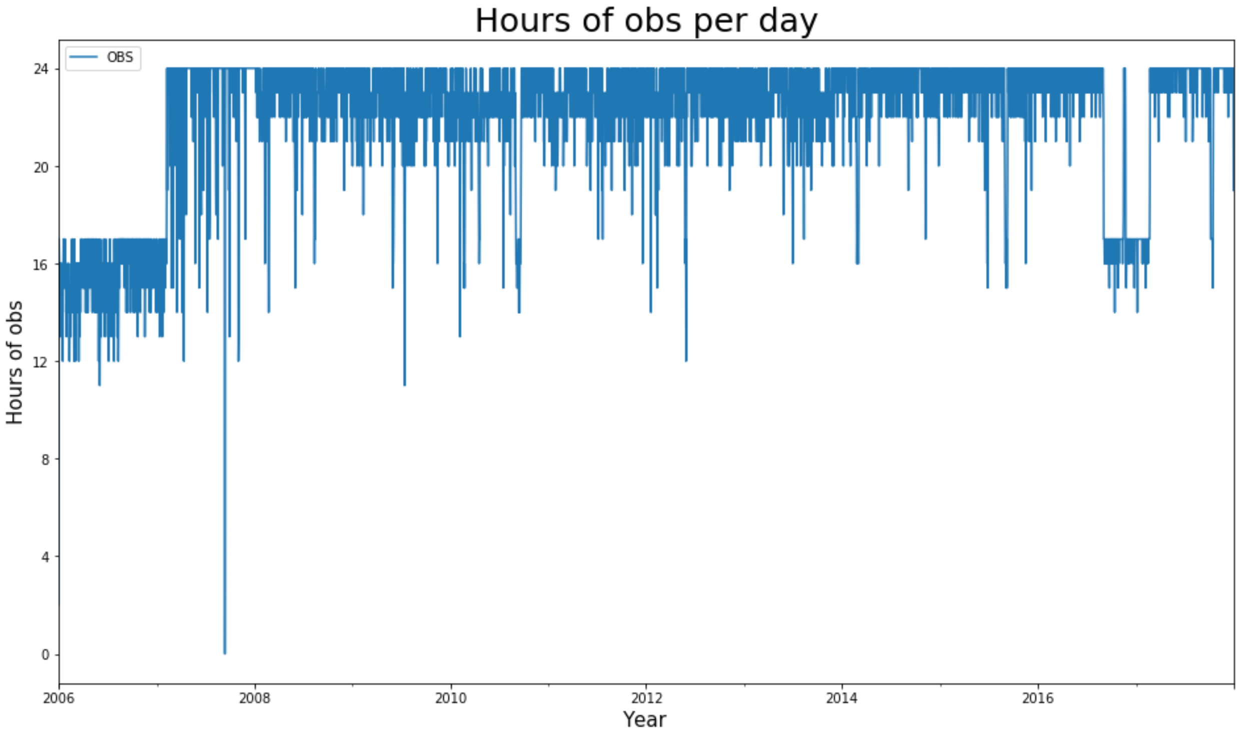
III. Dataset and Data Wrangling

Two sets of data were obtained for this project, each of them with data ranging from 2005 to 2017. One set, obtained from the Environmental Protection Agency (EPA), consisted of air quality data that had hourly ozone concentrations for a measuring station in Essex, MD, measured in parts per million (ppm). The other set, obtained from the National Oceanic and Atmospheric Administration (NOAA), consisted of hourly weather observations for Martin State Airport in Maryland, which located 3 miles from Essex.

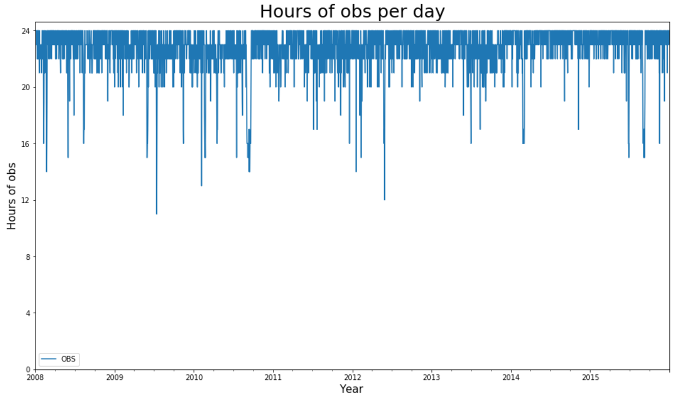
The objective of this data wrangling process was to ensure that for both datasets, observations at each hour would be available for the duration of the time period being analyzed, so that they can be merged into one dataframe containing both the measured ozone levels and the weather observations for each hour.

The weather dataframe was cleaned up first, and that one was the more involved of the two. The dataframe was filtered so that only the datetime (hour and date, labeled **YR--MODAHRMN**, then renamed to **TIME**), wind speed (**SPD**, measured in miles per hour), visibility (**VSB**, measured in miles), temperature (**TEMP**, measured in degrees Fahrenheit), dew point (**DEWP**, also measured in degrees Fahrenheit), and surface air pressure (**STP**, measured in millibars).

Other than the **TIME** column, the columns appeared to show no clear datatype. All of the variables are numeric in nature, so they were converted using the pd.to_numeric() function, with the errors coerced into NaN values. Also, the **TIME** column was set as the DatetimeIndex for this dataframe. The null values were identified by converting all columns to float data types, and then the missing values were counted. In the datetime column, the times were initially given as Greenwich time, so they were converted to Eastern US time. Next, the observations were resampled into a new dataframe so they could be cleanly shown for every hour, and then they were counted hourly to see whether each hour had an observation. Each hour was shown, and there was a column in which the number of observations per hour were counted. Many hours had zero observations, and those zero values were converted to null values so that the number of hours with observations for each day could be counted. A plot was done, and it showed that for the earlier and later parts of the dataset, significant numbers of observations were consistently missing.



Therefore, the dataset was truncated to only include 2008 to 2015, inclusive. This time period had occasional missing values (as would most datasets), but no consistent period of missing hours was found. As a result, the main weather observation dataframe was similarly truncated.



Then, an interpolation was done. To interpolate this data, a new dataframe ‘wx2’ consisting of every hour from 2008 to 2015 was created, under the singular column **TIME**. The index of the initial ‘wx’ dataframe was reset, so that the datetimes could be used as a column. A for loop was used to replicate all the column names of the wx dataframe, but without any data. However, the data types of those columns were set to float, so that any data that gets appended would recognized as numeric.

Afterwords, a new column 'stamped' in wx2 was created, with the value being set to 1, so that every row of this original wx2 dataframe would be recognized once the data from the original wx dataframe is appended. Another dataframe, wx_int, was then created to include all the rows and data from both the original wx and the newly created wx2. The 'stamped' column indicated which dataframe the row originated from, with values of 1 associated with the wx2 dataframe, and values of 0 associated with the original wxdataframe. The rows were then sorted by time.

With all the rows of the wx and wx2 dataframes sorted in chronological order, with the rows from wx2 having null values but clean hourly timeframes, the dataframe was finally ready for interpolation.

The interpolation was done in both directions, and all the hourly rows of the wx2 dataframe were successfully filled in with values based on the data from the original wx dataframe. The values from wx were deleted, and then wx_int was made to be the new weather dataframe.

The values in this interpolated dataframe were all floats, and every column needed for its values to either be rounded or expressed as integers. Once that was done, and once every column was confirmed to have no missing values, the data cleaning of the weather dataframe was finally complete.

The ozone dataframe was then cleaned. The ozone dataset was uploaded as a dataframe. Although it was a plain text file, it had commas separating each value, so no delimiter needed to be specified. The date format was odd, so that column needed to be cleaned before setting it as an index. The initialized dataframe contained 20 columns and 102,249 rows. The only columns of importance were the 'datetime' column and the 'value' column, the latter of which contains the ozone concentration measured in parts per million (ppm). Fortunately, there were no missing values for those two columns, with the exception of the final row, which is void of any values.

After a brief inspection, the dataframe was filtered so that only the 'datetime' and 'value' columns were selected. The 'datetime' column still needed to be cleaned. The reported times were local, so no time zone needed to be assigned. After the 'datetime' column was cleaned, it was ready to be converted to a datetime object. The datetime values were then arranged in chronological order, and the index was set to that column. The value column was renamed to ozone_ppm, so that the column name could be more descriptive of the data.

After the dataframe was confirmed to have no missing values, the ozone data was merged with the weather data into the dataframe df. After the merge, the new dataframe df was confirmed to have no missing values, and was pickled so it could easily be loaded again in another notebook. The data wrangling was finally complete.

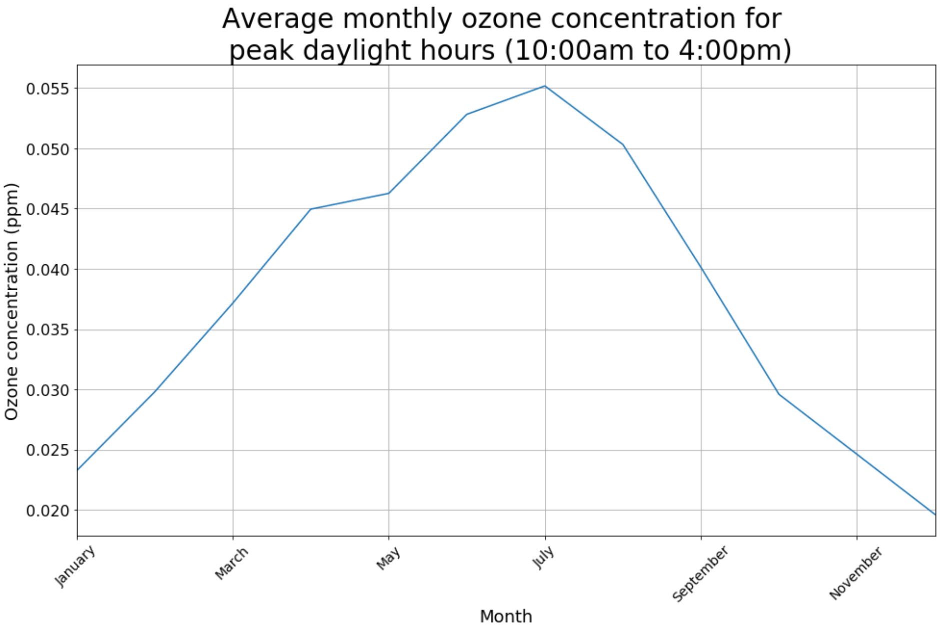
IV. Other potential datasets

This type of analysis can be done with any location in which reliable weather data and air quality data are available on a regular timely basis (either hourly or in increments of a few hours), with both sets of data being recorded reasonably close to each other. Other sites could have been chosen, but Baltimore was chosen due to personal familiarity.

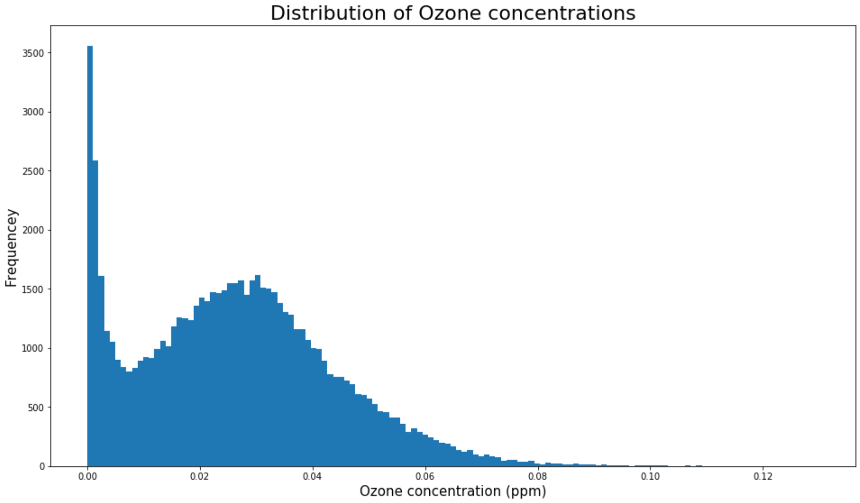
V. Initial findings

The first step was to import all the relevant packages and upload the dataframe from the pickle. Afterwards, there was a great deal of possibilities to explore with the data, to see how it appears and what insights can be gathered from it. Since ozone concentration (ppm) is the target variable, the first step was to simply visualize the entire scope of the data and see how it appears. This dataframe included 68,231 rows, and it consisted of every hour of observation from 2008 to 2015, inclusive.

A time series plot showing ozone concentrations throughout the whole time period was displayed, but was later zoomed in to 2011 and then specifically July 2011, for a better view of the cyclical nature of the data. Then, for a cleaner, more simple visualization of the seasonal differences in ozone levels, the monthly mean of ozone levels during peak daylight hours was plotted for every month. The dataframe was filtered to only include data for the hours from 10:00am to 4:00pm, and then the average monthly levels were taken. Average ozone levels during those hours ranged from under 0.02 ppm in December, to over 0.05 ppm in the summer months.

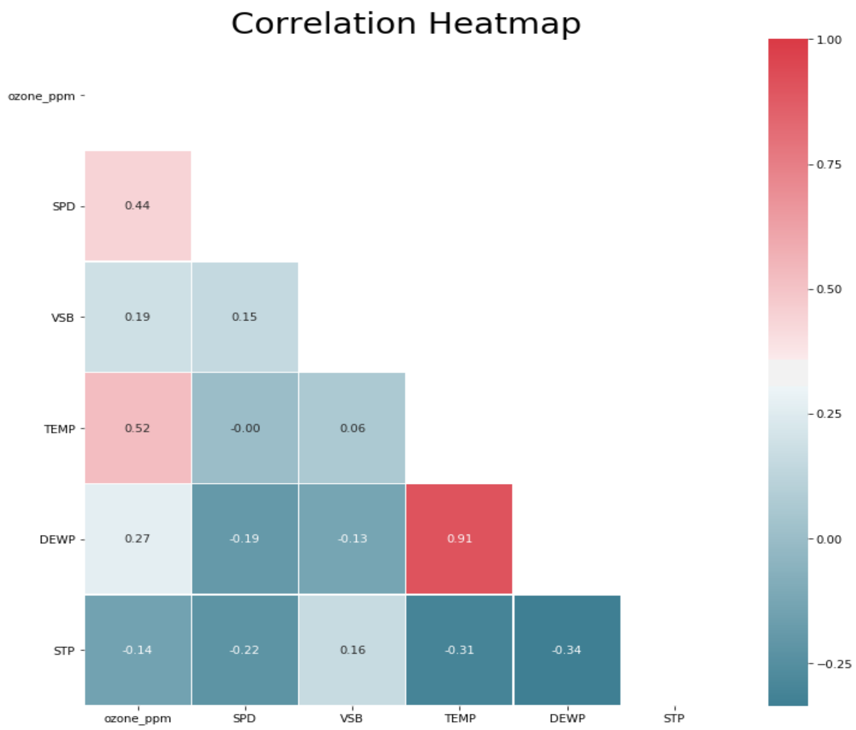


The frequency of ozone concentrations was then plotted in a histogram, with 131 bins chosen, and the plot appeared as a smooth curve, shown below.



Next, the weather variables were evaluated to see how much they correlate with ozone levels.

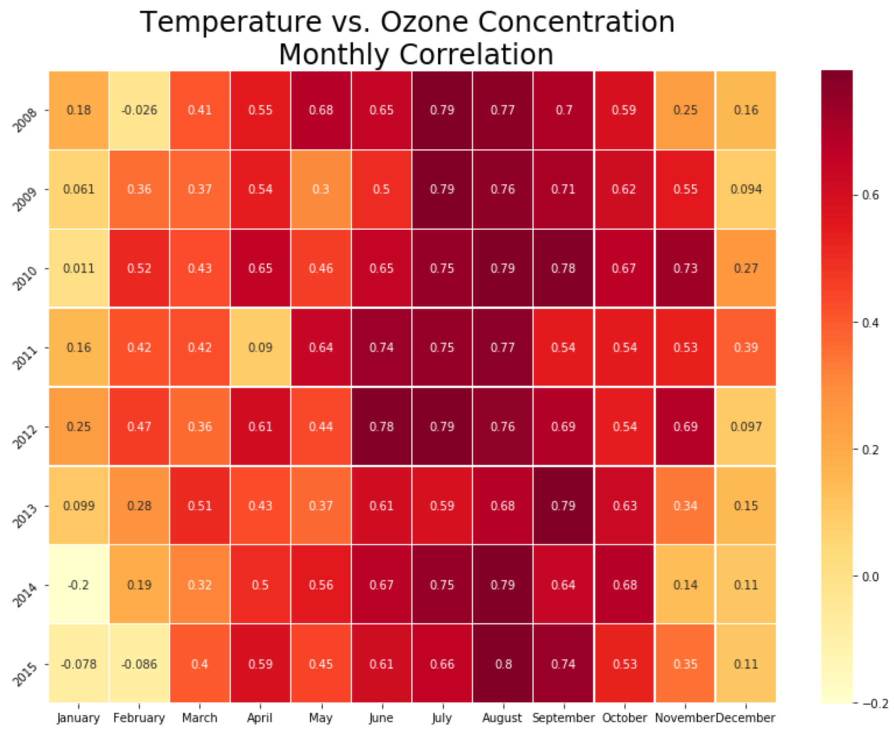
The correlation heatmap shows the pairwise correlations between all of the variables of the dataframe. The dew point is regarded as a proxy for humidity, and not surprisingly, the dew point and temperature were very highly correlated. However, since ozone concentration was the target variable, any correlations of the other variables with ozone levels were be of particular interest. The far left column shows the correlations of each of the feature variables with ozone concentration, with temperature and wind speed having the strongest correlation with the ozone level. The dew point has a weaker relationship with ozone levels, but even that variable was chosen for more exploration.



Since the ozone time series plots showed seasonal variation in the ozone concentration, a breakdown of monthly correlations with ozone was done for temperature, dew point, and wind speed, which showed the highest overall correlation with ozone levels, with heatmaps being shown for all three feature variables.

As seen below, the temperature vs. ozone heatmap shows a dramatic seasonal difference in the relationship between temperature and ozone level. That relationship is much stronger in the warmer months, especially from June to September, than it is in the winter months, when correlations near zero are common. Since the summer months are the most likely time for heat waves that affect air quality, the temperature is a powerful predictor of ozone levels when they are the most relevant.

Capstone Milestone Report

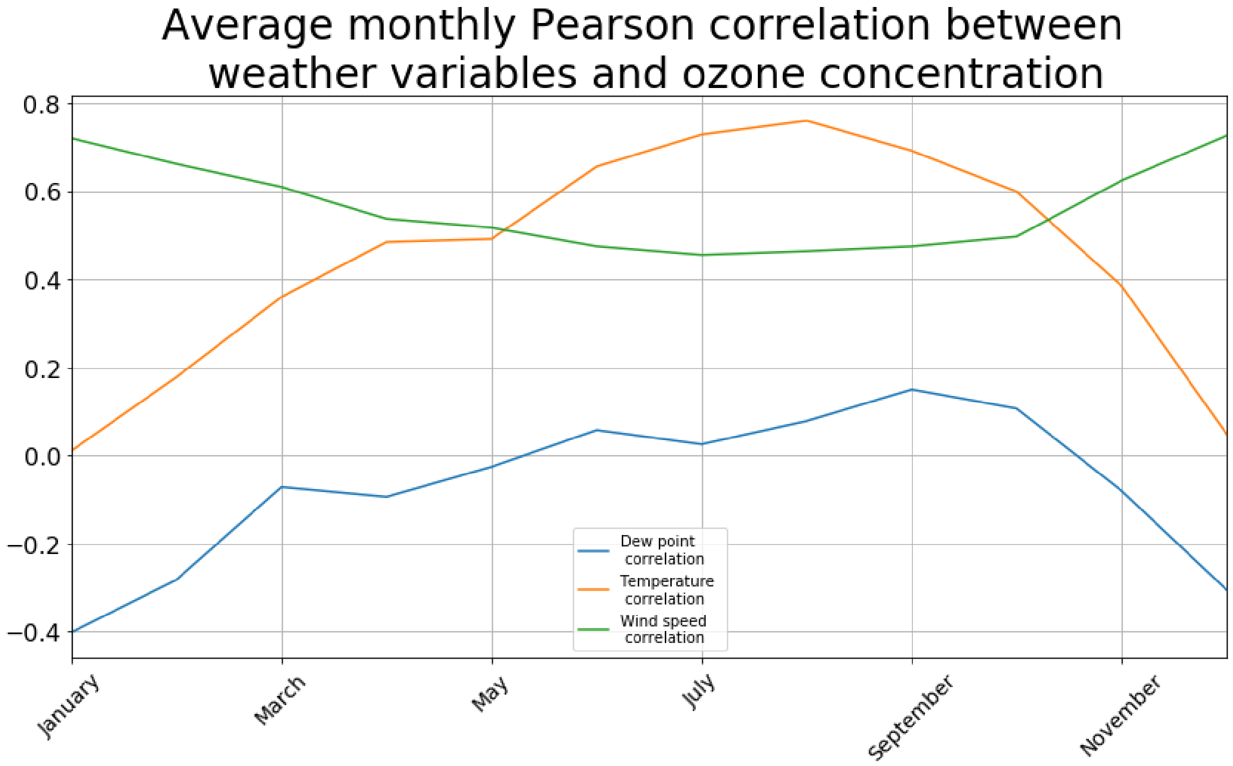


The next variable whose correlation with ozone was evaluated was the dew point. Although the temperature and dew point are highly correlated among each other, the dew point does not correlate with ozone levels as well as the temperature does. In fact, even in the summer months, average monthly correlations near zero were found to be common. However, in the winter months, dew points had a moderate negative correlation with ozone levels. In other words, higher dew points in the winter are moderately associated with lower ozone concentrations, which is a somewhat surprising finding.

Finally, the wind speed correlation with ozone was evaluated. Wind speed had an overall correlation of 0.44 with ozone level, making it the second most correlated variable after temperature. This was a fascinating yet unexpected finding, exactly the type of discovery that is worthy of being further explored. The monthly breakdown of this data shows that wind speed correlates strongly with ozone levels in the winter months, but has a moderate association with ozone even in the summer months. Although ozone levels in the winter are too low to have a significant impact on human health, the strong association with wind speed is still an interesting and very persistent finding.

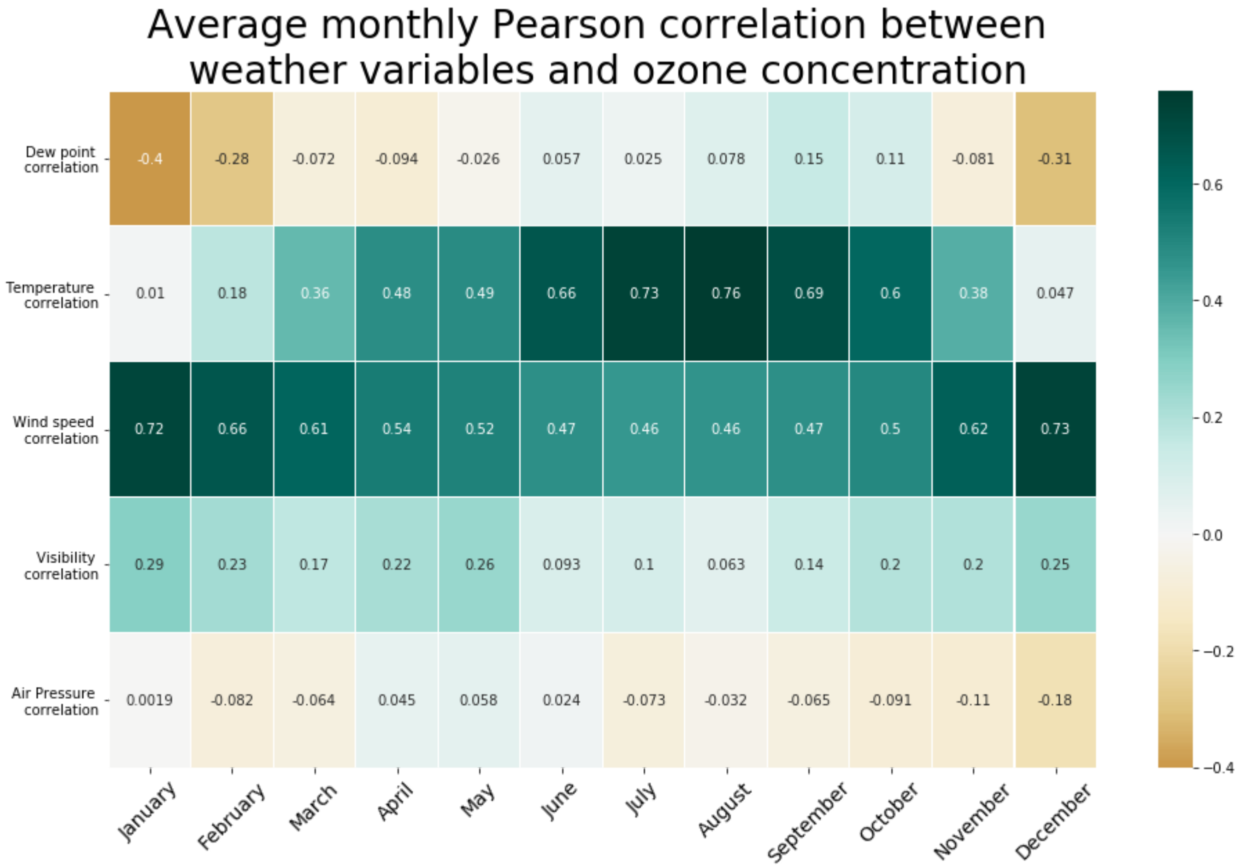
After the individual monthly plots, the monthly pearson correlations were plotted (displayed below) for a direct comparison, and the seasonal differences can clearly be seen. Wind speed has the strongest relationship with ozone levels in the winter months, while temperature is the strongest predictor in the summer.

Capstone Milestone Report



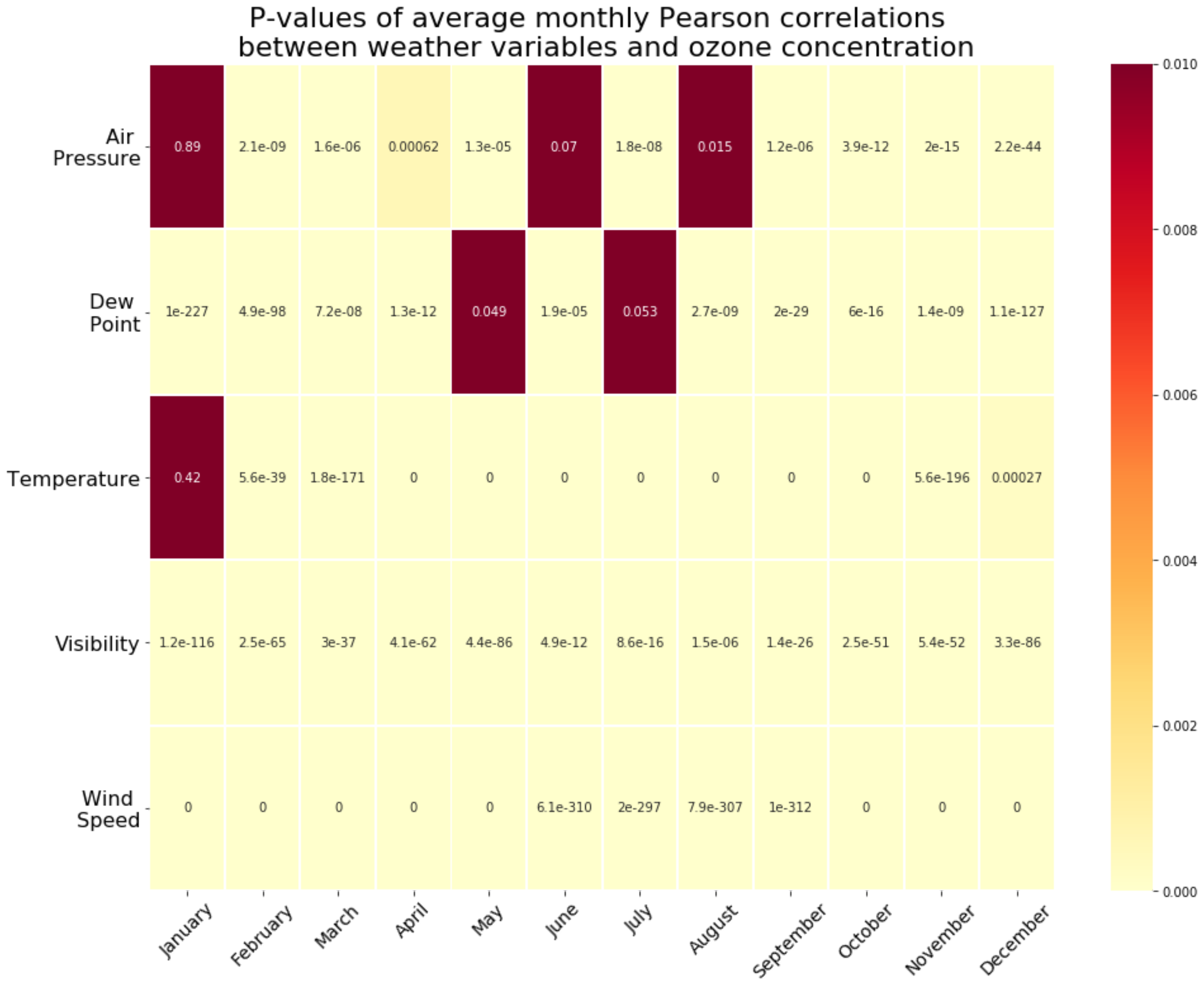
The dataframe with which the above plot was created was then given correlation columns for air pressure and visibility, for the sake of the inferential statistical analysis. Then the dataframe was pickled, so that it could be downloaded in the inferential statistics notebook.

Using the dataframe that was pickled at the end of the storytelling portion, a full monthly Pearson correlation heatmap with all of the variables was created, with temperature and wind speed having the highest correlation with ozone concentration, as was also shown in the storytelling portion. The visibility and air pressure were added to the comparison, after being disregarded in the storytelling portion due to their weak correlations. They both have very little relationship with ozone levels during the summer months, while visibility has a moderate relationship with ozone levels in the winter months. Air pressure is a weak variable throughout the year.



While the weather variables had various levels of correlation with the ozone concentration, they were evaluated on a monthly basis to determine the statistical significance of that correlation. The null hypothesis assumed no relationship, and the value of alpha was be set to 0.01, which means that the null hypothesis could be rejected for p-values that are under 0.01.

The p-values for the Pearson correlations of each of the variables with ozone concentration can be seen in the heatmap below. For all the values under 0.01, the null hypothesis can be rejected, and a statistically significant relationship exists between the relevant variable and ozone concentration for the given month.



For the vast majority of values, a statistically significant relationship exists with the ozone levels, and the null hypothesis can be rejected. This includes values that are associated with very low correlations, and illustrates the difference between statistical significance and *practical* significance. Only the months shaded in a dark maroon color have a p-value above 0.01 where the null is not rejected.

One possible reason for the discrepancy between statistical significance and practical significance is the large size of this dataset. With tens of thousands of values, a statistically significant model can be constructed even if it has very low predictability.

More analysis using predictive models will be shown in the final report.