Module 6 Writeup:

The Salt Lake City basin is well known as a growing hub of activity with quick access to great skiing and climbing. Unfortunately, the geographic shape of the region leads to “sinkholes” of pollution, particularly particulate matter (PM) and ozone. In the winter it is common to see inversion set in after long storm cycles. With reduced weather and wind, warm air traps cold air in the Salt Lake basin resulting in cooler temperatures closer to the earth. The byproducts of manufacturing and transportation are also caught in the cold air layer. These trapped particulates are dangerous to human health and careful monitoring is essential to preventing toxic exposure. Particulate matter 2.5 (PM2.5) is robustly associated with many negative disease outcomes including non-fatal heart attacks, asthma and increased cancer risk. Within Utah, the tracking, analysis and forecasting of PM2.5 is critical in protecting public health.

Currently, my final project focuses on time-series data – the collection of PurpleAir and Airnow data to explore air pollution across the state. In order to perform a robust analysis, this module represents an initial foray into the PurpleAir data for a 24 hour cycle on April 8th. Data was collected from the PurpleAir website, by scraping all measurements at hourly increments utilizing an Airflow workflow. This automated scheduler required the building of a directed acyclic graph to run the scraping. Once created, this task collects both the “legacy” and “experimental” data from PurpleAir, orients them in a data frame and saves the data as a .pkl file. This script cannot backfill or log entries from the past.

Several cleaning steps were required to clean and standardize the data. All sensor data was loaded and Utah data was selected by utilizing a Lat/Long bounding box. This resulted in ~25,000 data points for 1030 unique sensors spread across Utah. More sensors may exist but did not send signal on April 8th from 12:00AM to 12:00PM. Examination of the data yielded roughly ~1400 null PM2.5 measurements and ~12,000 null measurements for temperature, humidity and pressure. All other legacy measurements were considered too sparse for accurate filling. Since these data contain both temporal and spatial components, a “nearest-neighbor” method was used for null filling. For each sensor with missing values, the nearest non-null sensor was found by minimizing Euclidean distance. Then, a time-match was performed to fill the null value with the closest sensor both spatially and temporally. Additionally, several sensors were found to push high values often. Thus, any sensor recording over 500, the expected max for PM2.5 was removed.

Next, the sensor locations were visualized. As expected, highly metropolitan areas contained the highest number of sensors (Salt Lake, Utah, David and Weber). All sensor data was then aggregated by county location to increase the interpretability. Plotting the PM2.5 data over a 24 hour period showcased consistently low, non-dangerous levels of PM2.5. Perhaps the most prominent spikes in PM2.5 occur in Sanpete around 1:00 and 8:00PM. To visualize the average PM2.5 across each county, a heatmap was created, to showcase counties with increased PM2.5 As expected, these daily averages were higher for the areas where the population density is high. Note, the upregulations were minimal on a relative scale where above 100 is considered hazardous for the average person.

Decomposition was attempted for Salt Lake County and Utah County. This analysis was limited as the cycle is short and representative of a single day. County separation was maintained in order to account for geographical differences and minimize noise. For April 8th in Salt Lake PM2.5 generally increased throughout the day while Utah county saw a spike around 12:00PM with a fall shortly after. When modeled with basic linear regression, the fit was poor with only 48.6% and 54.5% of the variance being explained by a simple model. Adding in temperature, humidity and pressure boosted model performance with multi-linear regression resulting in an R2 value of 0.63 and 0.70 for each county respectively.

In addition to modeling, auto and partial correlation were calculated on the Salt Lake and Utah county data. These plots showed very few of the cycles being predictive of the previous. Interesting, for both, it seemed that those measurements closest to the 24 hour mark were considered the most predicative. These analysis will become more interesting with an increased amount of data with some cyclical pattern. Finally, prophet was utilized to make some predictions about the PM2.5 levels in Salt Lake in the following 48 hours. Visualization showed predictions increasing PM2.5 linearly with time. This trend is obviously incorrect as PM2.5 will vary based upon a plethora of factors (or perhaps this truly is the end of the world!). Nonetheless, this speaks on the data “hungry” nature of predictive models to become accurate.

These analysis were most limited by data size. The modeling was inaccurate and cleaning ungrounded as an acceptable level of context could not be established from the data. This work acts as a great example of a “cold-start” algorithm, where the initial predictions are poor because the body of data is simplistic and uniform. Of note, the PurpleAir sensors are notorious for their inaccuracies. Several points were found in the data which contained readings above 5000 while the remainder of the nearby (<10km) sensors showcased values <10. This calls into question the accuracy of these sensors and bias of location. In this case, it may be better to overestimate the PM2.5 in order to ensure inaccuracies do not increase risk of human harm. Combined, the level of inaccuracy makes these results non-actionable as bias exists due to the small sample size and measurement error.