**Summary**

There is variance in recorded PM2.5 values across stations. Numerous spikes are observed indicating periods of significantly increased PM2.5 concentration. Station GC-5 shows the highest readings. Variance tests indicate that there are significant differences in average recorded PM2.5 levels between stations.

**Methods**

Available data aggregated to 10-minute averages from installed stations was downloaded from April 1st to July 21st using the PurpleAir API.

Data analysis was performed in Python and R.

**Results and Discussion**

There is apparent variance in recorded PM2.5 values across stations. Spikes in the 10-minute average are observed, with a very large spike of 1974 ug/m3 recorded by GC-5 on 4/26/24 at 5:20 AM. Some of these spikes are quite short in duration and could be due to the sensor detecting short bursts of high concentration PM, such as someone smoking nearby or vehicular exhaust, while others are more sustained. I’ll work on analyzing the data using different time aggregations to identify more sustained increases. There appears to be a general increase in average PM2.5 across stations during the first couple weeks of June. (figure 1, 2)

A group of blue lines

Description automatically generated

Figure 1 - PM2.5 concentration (ug/m3) 10-minute average from April to July. Note observed spikes in recorded PM2.5. Stations GC-1 - GC-4.

A graph of a graph

Description automatically generated with medium confidence

Figure 2 - PM2.5 concentration (ug/m3) 10-minute average from April to July, stations GC-6 – GC-8.

Boxplots indicate a large number of outliers, especially in the data from stations GC-5 and GC-6. GC-5 shows the greatest number of outliers compared to the other stations, with several instances of recorded 10-minute PM2.5 averages that are orders of magnitude greater than readings from the other stations. (figure 3, Table 1).

A graph of a number of objects

Description automatically generated

Figure 3 - Distributions of pm2.5 (ug/m3) 10-minute average readings across stations. Note the large number of outliers in station GC-5, which shows readings orders of magnitude larger than the highest recorded concentration at other stations.

Table 1 - Descriptive statistics



After removing outliers, the data distribution is more even and less dominated by GC-5, which no longer shows the highest readings. There is apparent variance in the distribution of 10-minute PM2.5 readings across stations, with GC-3 showing the highest average PM2.5 concentration over the observation period. While variance is observed, the data distributions from all stations are within each other’s IQR range (there is no station in which the recorded mean is above the 75th percentile of another). (figure 4 Table 2).

A graph of different colored boxes

Description automatically generated

Figure 4 - Distributions of pm2.5 (ug/m3) 10-minute average readings across stations, outliers removed.

Table 2 - Descriptive statistics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **GC-1** | **GC-2** | **GC-3** | **GC-4** | **GC-5** | **GC-6** | **GC-7** | **GC-8** |
| count | 9935 | 14655 | 14518 | 10854 | 8719 | 7374 | 7255 | 13238 |
| mean | 0.96008 | 1.54336 | 1.68419 | 1.37012 | 0.921996 | 0.93315 | 0.82491 | 1.55604 |
| std | 1.04925 | 1.54537 | 1.62916 | 1.40877 | 0.838207 | 0.95987 | 0.74645 | 1.61116 |
| min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 0.086 | 0.299 | 0.445 | 0.248 | 0.268 | 0.177 | 0.2375 | 0.267 |
| 50% | 0.5875 | 0.955 | 1.06 | 0.8105 | 0.69 | 0.593 | 0.621 | 0.92 |
| 75% | 1.489 | 2.426 | 2.524 | 2.131 | 1.2795 | 1.398 | 1.1745 | 2.449 |
| max | 4.191 | 6.516 | 6.797 | 5.858 | 3.485 | 3.956 | 3.115 | 6.684 |

Normality tests indicate that data distributions from all stations are significantly non-normal. Distributions are significantly right skewed, and there is apparent variance in distributions between stations, including distribution shape. GC-1 shows the greatest skew, with the largest number of low PM2.5 concentration readings. There is again indication of variance in readings across stations, with differences in mean readings. GC-3 shows the highest mean PM2.5 readings (figures 5, 6). A Kruskal-Wallace variance test indicates that this variance is statistically significant (Table 3).

A graph of different colored bars

Description automatically generated

Figure 5 - Histograms of pm2.5 10-minute averages from stations overlaid on top of each-other.

A graph of a number of people

Description automatically generated with medium confidence

Figure 6 - Histograms of pm2.5 10-minute averages from stations. Note variance in mean recorded pm2.5 between stations.

Table 3 - Kruskal-Wallace variance test results. Results indicate statistically significant difference in mean-rank recorded 10-minute pm2.5 between at least one pair of stations. Outliers removed.

|  |  |
| --- | --- |
| Kruskal-Wallace Result |  |
| H-statistic | 3103.17 |
| p-value | **0** |

**Work going forward**

I will analyze the data using different time aggregations to further investigate sustained increases in recorded PM2.5 concentration.