Final Project Writeup
Stat 370

Airbeam3 Sensors

By: Victoria Olejarz, Olivia Roeger, Olivia Pardo

The AirBeam is a low-cost sensor that measures particulate matter. It captures a lot of important information such as PM, temperature, humidity, longitude, latitude, and timestamps. Particulate matter consists of solid particles and liquid droplets found in the air. The AirBeam directly measures PM2.5, which are inhalable particles that are 2.5 micrometers and smaller. Additionally, it provides information on PM1 and PM10, which are calculated from PM2.5. With data from seven sensors, our goal was to examine the data to test the reliability of the AirBeams.

Another class was in charge of collecting the data using the AirBeams. We provided them with a Microsoft Form to collect additional information not captured by the AirBeam. The class was divided into seven groups for the seven different sensors. Each student was required to record twenty hours of air quality data. At the end of the project, they had conducted 169 sessions; however, only 62 of the sessions had the corresponding form filled out. After receiving the data, all of it was combined into one CSV, which was then used to test the reliability of the sensors. This process included delving into the NA values and examining maps of where the PM levels peaked. It is important to determine the reliability of the sensors so that Loyola knows if they are a worthwhile product to purchase in the future. With our research, we were able to gain a better understanding of the reliability of the sensors.

When attempting to look into the validity of the sensors, the main challenge we encountered throughout our consulting project was investigating the amount of zero values the Airbeam3 sensors recorded and where they are in the data.

Some restrictions we faced in aiming to find the answer as to why there are so many zero values are that each session was recorded in a different nature. PM levels indoors and outdoors would present different values, but how different would these values really be? We also could not confirm what the actual PM levels were at the time of each individual session which was difficult to infer whether the non zero values were accurate readings. Another restriction that made it more difficult to analyze where the zeros are in the data was the fact that some of the sessions did not record any data at some points. Some sessions presented in many breaks in the map which showed a type of disconnection where the sensor stopped working and did not collect any data at the specific time. We had to consider these limitations that could have possibly explained the constant zero values in the data before looking into overall patterns and trends.

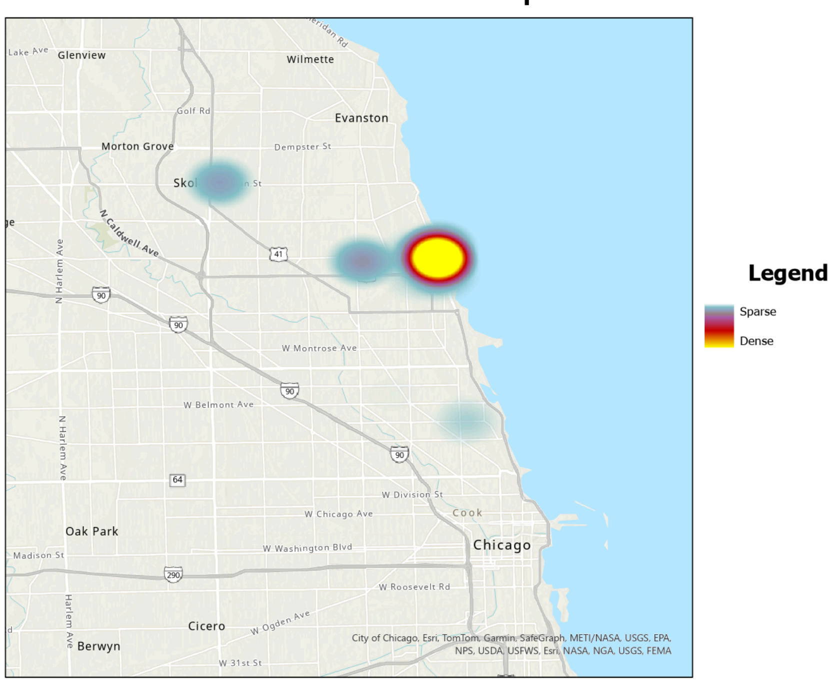
We collected our data from the Airbeam3 sensors from a website called HabitatMap. From this website, we were able to see the usernames of the people who recorded the sessions, the session names, the time frame of the session, and the location tracked throughout the session. We were not aware of which session belonged to the students from the ENVS class, so we emailed the class asking what each student's username is. We then created a list of all the usernames we should look out for on the HabitatMap website so that we could collect the data from the appropriate sessions. Since we were aware of which sessions to look for, we were able to collect all the data from the HabitatMap by exporting all of the appropriate sessions to our emails. From there, we downloaded all of the csv files that included each individual session each student recorded. All of the data was exported individually, so we combined all of the csv files into one file. The final file we analyzed includes each session name from all the students, the timestamp which includes the date and specific time (to the second) the measurement was taken, latitude, longitude, temperature, humidity, PM1, PM10, PM2.5, and the specific Airbeam3 sensor that was used to record the according data.

Given these variables that the sensors themselves recorded, we wanted to explore more into the background of each individual session. We wanted to know what the students were doing and what was going on at the time of their sessions. We decided to create a microsoft form that would help us understand the nature behind each session. In the microsoft form, we asked the students from the ENVS class to tell us their initials so we can link their sessions to their names in case we had further questions regarding their data. We asked what the date of their session was, as well as whether their session was fixed or mobile. For context, the Aircasting App included different types of sessions to record. The main difference between the types of sessions was the type of wifi or overall connection that the sensor used at the time of the session. We wanted to specify what type of session they used in case connection got lost, and to see if data was presented differently in different types of sessions. We also asked on the microsoft form for the students to specify the level of traffic at the time of the session, varying from no traffic, light, moderate, and heavy traffic. We wanted to explore if there was a correlation between traffic levels and PM levels. We as well asked about what neighborhood/city the students recorded in, what the weather was like (sunny, cloudy, rainy, windy), and whether or not there was fire burning at the time of the session. We asked about what the students mode of transportation was as well as all of these were potential causes for varying PM levels that the sensor could have picked up. We then sent out this form to the entire ENVS class for the students to fill out as they completed their sessions. Our motive for the microsoft form was for us to understand what was happening at the time of the students sessions, and to potentially explain any discrepancies or patterns within the data.

As stated previously, a feature of the airbeams included the latitude and longitude of each measurement. This feature allowed us to visualize where measurements were being taken and how they might change depending on the area.

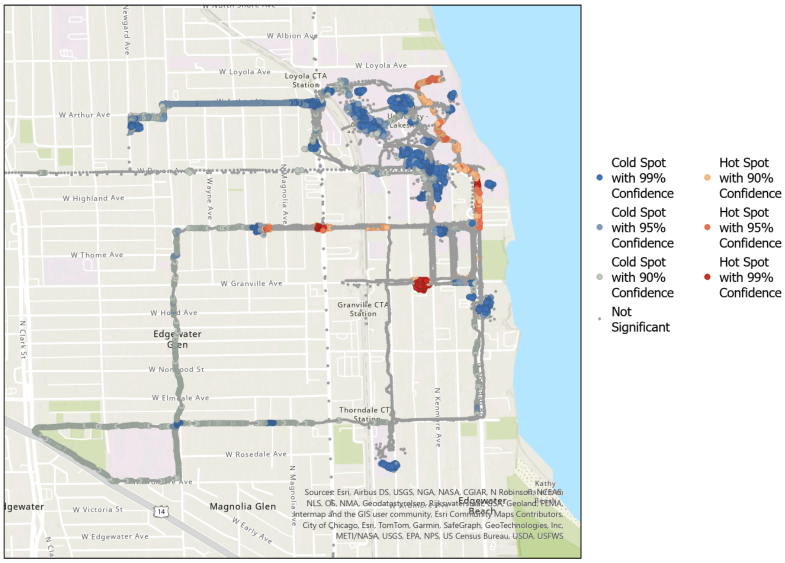
To visualize this, we used ArcGIS, which is an information system geographic software. We created maps that allowed us to visualize where data was being collected, hotspots and cold spots of PM levels, and how PM levels change over a specific session.

Airbeam Location Heat Map



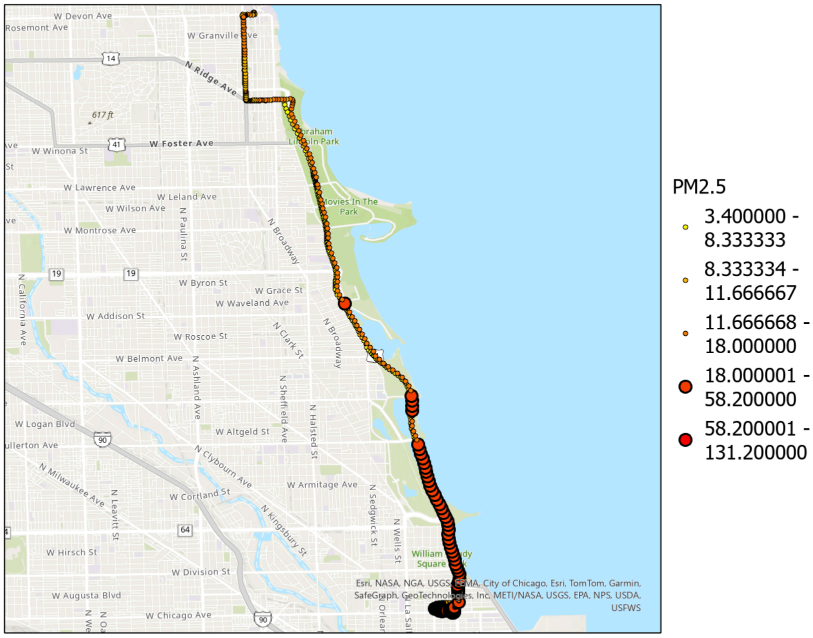
This map here shows a heat map of where students recorded PM levels. It is very obvious that the majority of records were taken in the Rogers Park/Edgewater area, which we assumed would happen considering that this is the area where students live and go to class.

PM2.5 Hotspots and Cold Spots




This map shows areas with hot and cold spots of PM2.5 levels. Areas that are considered hot spots (in red), are considered to have a pattern of higher PM2.5 levels across all sessions. Areas with cold spots (in blue), are considered to have a pattern of lower PM2.5 levels across all sessions. Areas in gray are insignificant. One thing interesting to note is that some of the areas that are considered cold spots are actually buildings on Loyola's campus, and many of the areas in red are on roads where there might be traffic/cars.

PM2.5 Changing Across One Session




The final map created shows how PM2.5 levels change over one session. This session shows a student traveling from the Lakeshore Campus to the Water Tower Campus, then back to the Lakeshore campus. This route takes the major road of Lakeshore Drive. We can see that within this session, PM2.5 levels are lower towards Rogers Park, but then they increase as they get closer to downtown. This could be because there are more cars/traffic towards the downtown area, thus resulting in higher PM2.5 levels.

Along with visualizing where the data was recorded, as we mentioned before, one of our main challenges was recognizing exactly where the data was collected as zero or null. Before taking a deeper dive into this, we first looked at an overall summary of our data.

Excluding null/zero values, the average levels for PM1, PM2.5, and PM10 were 10.38, 11.70, and 15.74 in respective order. Upon research we found that is an overall “normal” reading to have for particulate matter. The overall average temperature (excluding zeros) was 75.78

degrees fahrenheit- which we know is a high value considering the weather from February to April in Chicago, and the overall average for relative humidity (excluding zeros) was 26.95.

Along with looking at averages, we also wanted to take a look at the variability between PM2.5 measurements across all sessions (still excluding the zero values). We felt that there were definitely significant outliers in some measurements, but to be sure we conducted an F-test between all of the data, and the middle 25-75 percentile of the data. Our null hypothesis is that the two variances would be equal, and the alternative was that the variance for the middle data was much smaller than for all data. Upon calculation, we found that the variance for all of the data was 432.3, while the variance for the middle was only 12.26. The F-test resulted in a p-value = 0. Thus we reject our null hypotheses and conclude that the two variances are statistically significant.

After taking a look at the average measurements and the variability, we started looking into where the zeros/nulls were in a general sense. The following graphic shows the number of zeros/nulls for each field that the airbeam collects:

```
Number of zeros for each field:
Session_Name         0
Timestamp            0
Latitude             0
Longitude            0
Temperature      27117
PM1             175656
PM10            174783
PM25            175668
Humidity         27112
Airbeam_name         0
dtype: int64
```

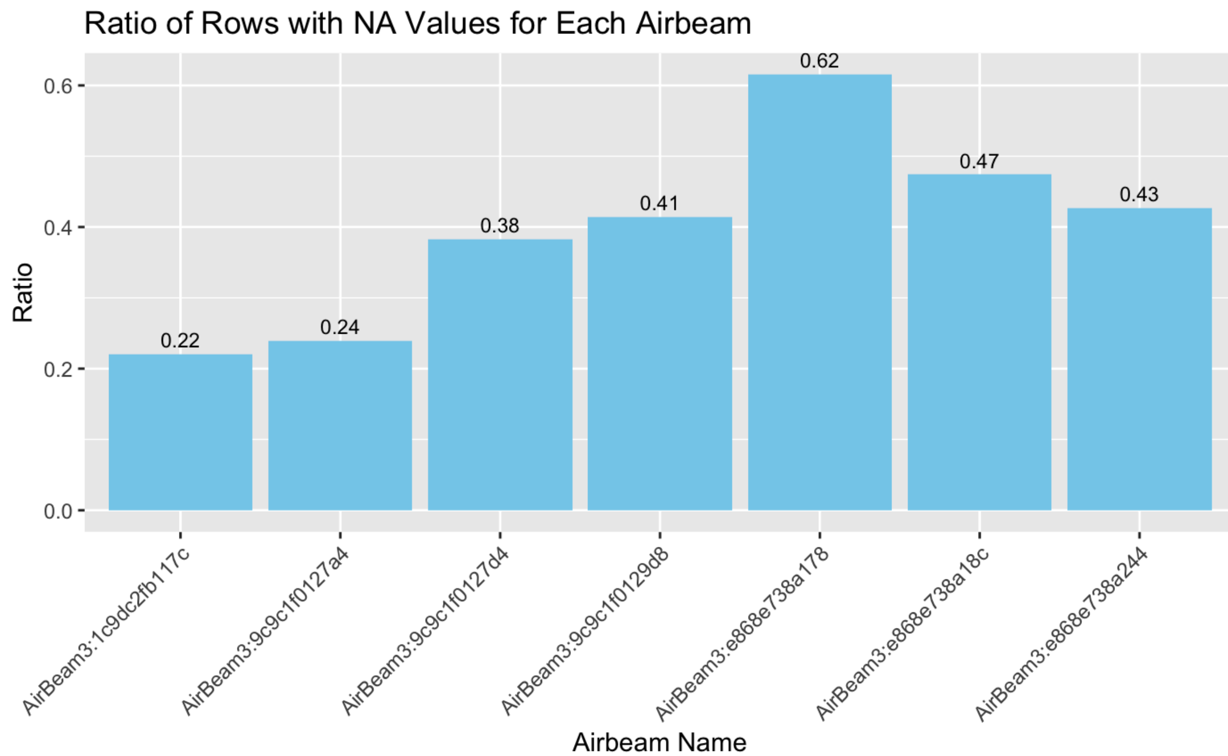
We can see that the majority of the zeros fall under the PM measurements. Upon further exploration, we actually found that 165202 out of 433887 records had zero/null values for every PM level. That means that 38% of all records had no PM level recorded. We did look into if there was some sort of calibration equation to troubleshoot for the zero values, but there was nothing provided by the Airbeam website.Thus it would be safe to assume that if one PM level was not recorded, then no PM level was recorded.

Looking further into the zero values, a common pattern we spotted was that some sensors recorded more zero values than other sensors, which led us to believe the sensors themselves are recording the variables differently. We decided to explore this more by grouping the data according to each individual Airbeam3 sensor.

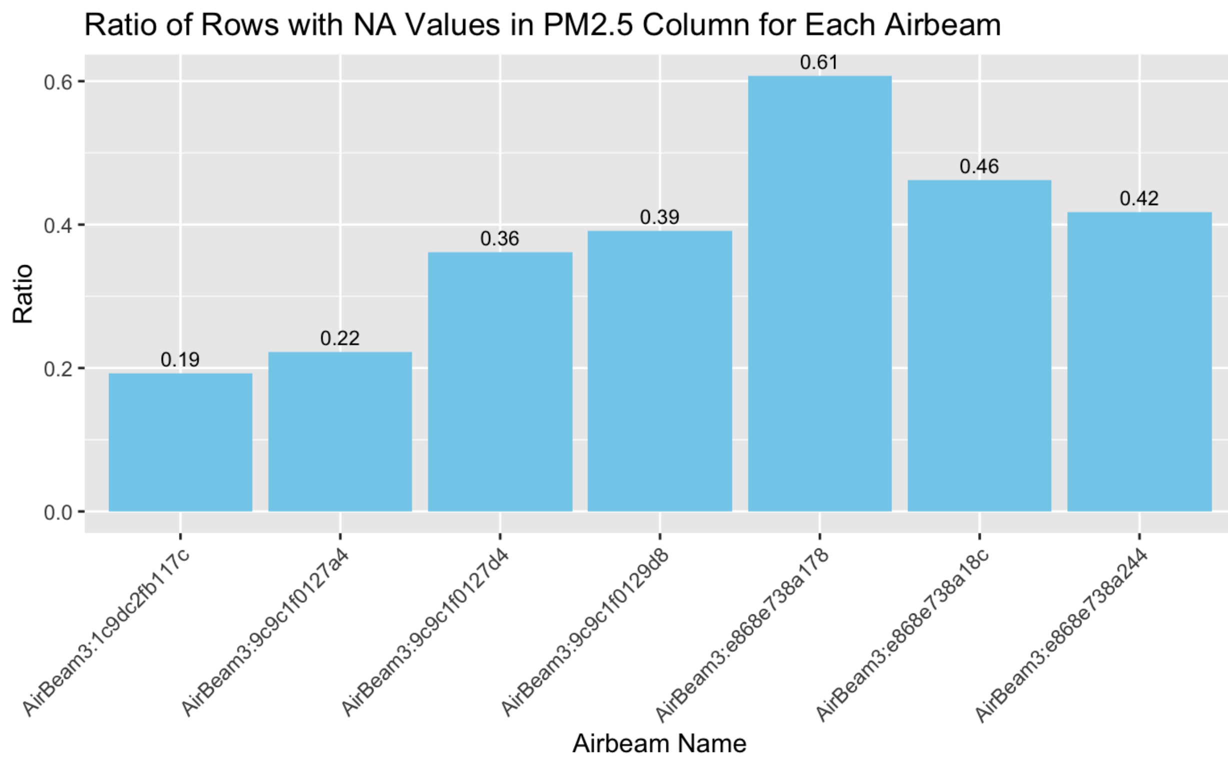
It is important to group the sensors and examine the differences between the Airbeam outputs to test for reliability. Every CSV exported from the habitat map includes the name of the Airbeam. In R, this Airbeam name was used to add a column called "Airbeam_Name" to the combined data for easy grouping. Once the Airbeam column was created, we then made graphs using ggplot2 that visualized the ratio of the total number of rows the Airbeam recorded to the number of rows that included at least one NA value or 0 in the row. We included when the sensor recorded 0 because it is not logical for the PM value, temperature, or humidity to be 0. We then examined the graphs to compare the reliability between the sensors.

We created graphs based on the data grouped on the sensor. The first bar graph showed the differences between the proportions for each Airbeam. The Airbeam with the smallest ratio was AirBeam3:1c9dc2fb117c. This Airbeam also had the fewest number of sessions at three.

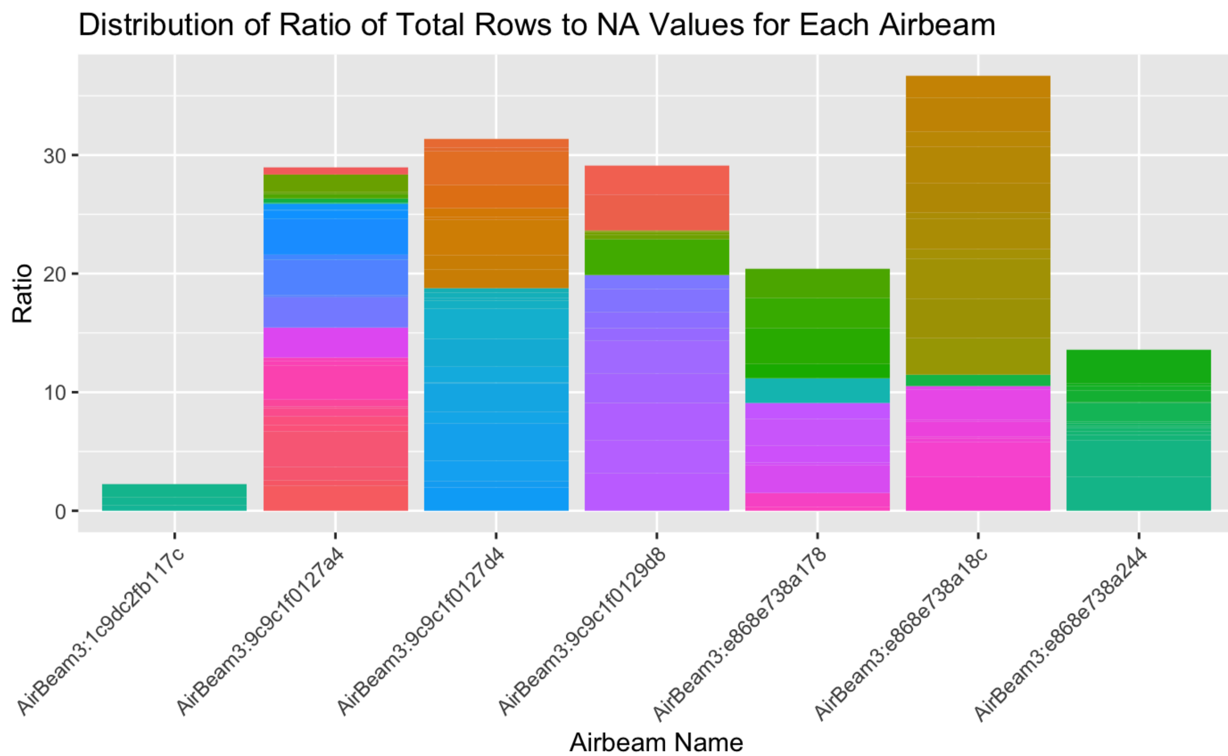
AirBeam3:e868e738a178 was the sensor with the highest proportion of NA values at 62% and a total of 39477 rows with at least one NA value in them.




Following the bar graph for the entire data frame, a bar graph that just visualized the PM2.5 column was created. PM2.5 is the most important value to look at because this is what the Airbeam censors actually measure, while the other PM values are calculated based on PM2.5. The PM2.5 graph revealed that most of the missing data comes from this column. From the graph that visualized all the data to the one that only visualized PM2.5, each Airbeam's proportion only dropped one to three percent.



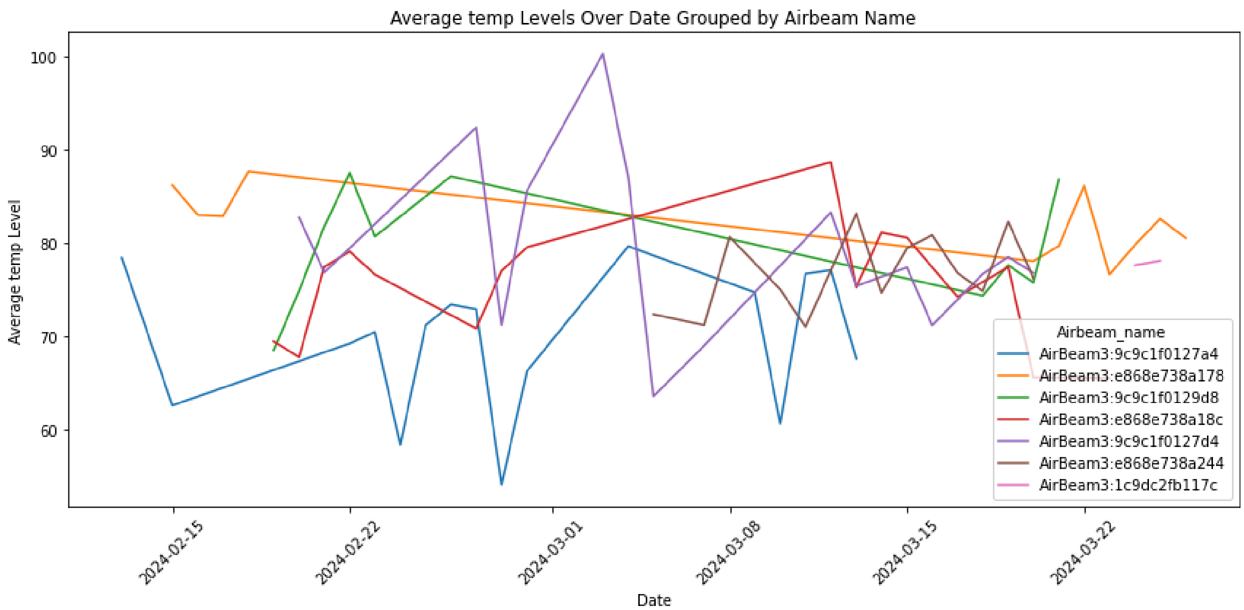

To diagnose why there are so many rows with 0s and NA values, we decided to look into individual sessions. A stacked bar graph was created to visualize each session for each air sensor. The stacked bar graph displays that some sessions recorded most of the data while some had a proportion of missing values of 100%. We could not identify a pattern based on session, as even two sessions recorded by the same person could have significantly different ratios of NA values.




The overall summary statistics for the temperature data showed decently consistent patterns across all of the individual sensors. Each of the seven sensors presented commonalities with the overall data, proving to us the sensors recorded temperature relatively the same way. This is also proved in the summary statistics provided for the humidity variable. The main problem we spotted when grouping the data by sensors were the summary statistics when looking at PM1, PM2.5, and PM10 levels. Because of all the zero values in all of the PM recordings, the boxplots centered extremely close to the very bottom of the plot. Some sensors recorded PM2.5 levels ranging from 0 to 20, and other sensors ranged from 0 to 310. These are drastically different measurements that made us question why some sensors recorded much higher values than others when the average PM2.5 level for the entire dataset was 11.75. The issue was that all five of the seven sensors recorded drastically high measurements, which were presented as outliers in the plots. As the mean for PM2.5 levels is 11.75, most of the PM2.5 data lies between 0 and 13. It was clear that the five of the seven sensors collected less reliable data than the other two. It is important to note that across all the boxplots for each PM level measurement, Airbeam sensor (AirBeam3:e868e738a178) recorded the most and the highest outliers out of all the sensors.

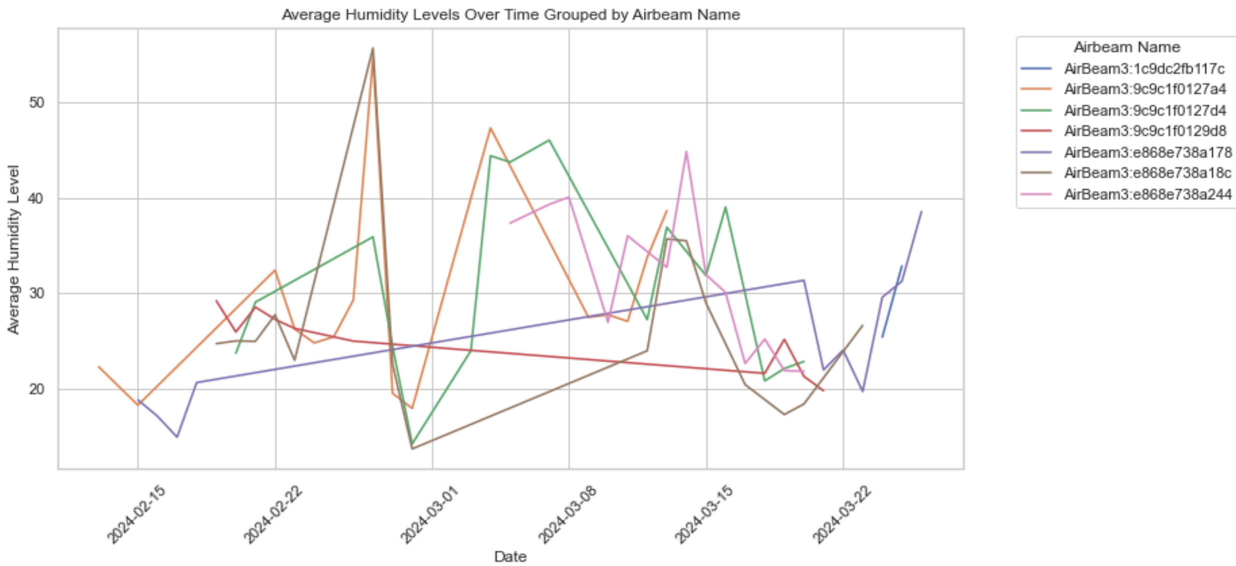
Looking at the trends of the data over time, we are able to explore the range of measurements the sensors recorded within similar time frames. First, we decided to make a graph that plotted the change in PM2.5 levels over time grouped by Airbeam3 sensor. This method allowed us to see how the sensors differed in measurements, which proved to be a significant plot. We also used the same method with the variable temperature. Theoretically, we inferred that temperature recordings should be recorded extremely similarly between each sensor given the fact that most of the sessions in our data were recorded outside. Also, all of the recordings were measured in the Chicago area, we also let us infer that the temperature recordings should be about the same for each individual sensor given similar time frames. This allowed us to investigate if the sensors indeed recorded the variables differently, and if so, which sensors do not align with the rest of the data.

When looking at the temporal plots of our dataset focusing on temperature grouped by airbeam sensor, we found clear inconsistencies of temperature recordings within similar time frames. In this graph, we plotted the average temperature level over time grouped by airbeam sensor.



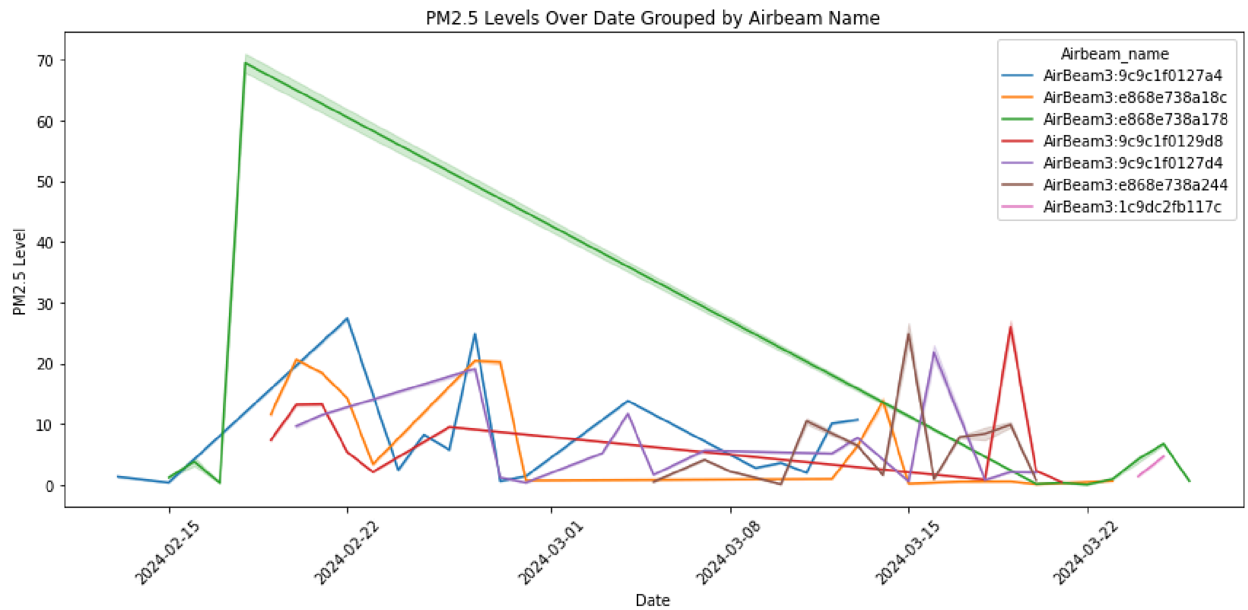

One finding that stands out the most in this graph is the fact that some sensors record much lower or much higher temperature than other sensors within the same period of time. For example, in this graph we see a sensor record an average temperature of about 40 degrees and another sensor record an average temperature of about 104 degrees. This is significant because these two average temperatures from two different airbeam sensors recorded drastically different temperatures within a time span of a few days. Given that the Chicago weather is relatively the same within a few days span, the temperatures should be more so similarly measured. This example is the most drastic difference, but the plot continues to prove this concept throughout the entire timeline.

Now, with the same plotting method, we look at humidity levels change over time.




Looking at this plot, we expect the average humidity levels to be recorded more similarly during the same time frames, just like we expect in temperature. We see the same pattern where individual sensors record drastically different average humidity levels give or take a few days. For example, one sensor recorded an average humidity of about 60, whereas another sensors recorded and average humidity of about 20 within the same day. This go to show that the sensors do not record the same humidity values given the same timeframes when theoretically the average recordings should be a lot more similar to each other. We have the same problem looking at humidity as we did looking at temperature. This is once again a key finding we made when looking at how each airbeam3 sensor recorded data drastically differently.

In the same manner, we use this method again, but we looked at average PM2.5 levels instead of just temperature.




As mentioned before, we noticed that Airbeam3 sensor (AirBeam3:e868e738a178) contained the most problematic data given it has the most and the highest amount of outliers compared to any other sensor. We see this finding once again in the graph that plots PM2.5 levels over time grouped by airbeam sensor. There are consistent recordings that are more so based on the bottom of the graph between all of the sensors, except Airbeam3 sensor (AirBeam3:e868e738a178). This sensor is clearly inconsistent with the rest of the data given that the recordings are not the same whatsoever compared to the other sensors.

Overall, grouping the data according to airbeam sensors allowed us to find that the individual sensors themselves record data differently. Specifically, Airbeam3 sensor (AirBeam3:e868e738a178) containing the most inconsistent data is the greatest finding when we compared all of the sensors individually. This allows us to come to the overall understanding that many of the zeros values in our entire dataset are unique to which sensor recorded the measurements.

There are many things that we would’ve liked to either go back and change, or continue exploring in the future. In terms of exploring the sensor accuracy one thing we would’ve liked to have done was test the sensors with different user techniques before collecting the “final data” so that we could see how the data is affected depending on how it is collected. We also would've liked to have been able to compare the data collected from the Airbeam sensor to data collected by other sensors in the area, such as Purple Air, the Egg Sensors, or EPA sensors (which would be considered the high standard for accuracy). Another major thing that we wished we could’ve established was a connection between the Microsoft form and the Airbeam data. With how the form was set up, and since not all students filled out the form, we were unable to link the two to each other. This is something we think would have provided great insight as to why some of the data was the way it was, but unfortunately this was not a possibility. In the future we would put a greater stress on the importance of the form, and make sure the form has more specifics and features that make it easy to link to the airbeam data.

Overall, there are a few things that really stood out to us when analyzing the data from this project. For one, we know that AirBeam3:e868e738a178 (which we found to be “Group 5”), had very inconsistent measurements compared to the other sensors. Many of the records collected by this sensor were zeros, and we could see that the PM levels collected by this sensor were much farther off compared to the other sensors. Secondly, referring back to the maps that were created, we can see that the main location where this data was collected was within Rogers Park/Edgewater. This can be seen as both a pro and a con. On the pro side, we were able to see variation in measurements for this particular area, as well as validate measurements (as they should be similar since it is the same area). On the con side, it doesn't give us a variety of locations to compare to, so we aren't really able to see how PM level might change depending on the neighborhood/community/etc. Lastly, our main conclusion refers to the accuracy of the sensors. Overall, we felt that the abundance of zeros was a large concern. With 38% of the data not collecting any PM measurement, this definitely calls into question the performance of the Airbeam3 sensor. We also noticed temperature levels that were much higher than what they should be, as well as inconsistent temperatures for sessions that are days apart. Although temperature is not exactly the most vital measurement- it does call into question the accuracy of the sensor, especially if the sensor advertises that it is able to record temperature. As for humidity, we actually found the levels recorded by the sensor to be lower than the normal levels, however we do not have as much knowledge about humidity levels so this could be explored further. Overall, within our research we did not find these sensors to be completely accurate, but there could be much more testing in order to come to a definite conclusion.

**References:**


"AirCasting." *Aircasting.habitatmap.org*, aircasting.habitatmap.org/mobile_map#?map=%7B%22zoom%22:16.

"HabitatMap Is an Environmental Tech Org and Maker of AirBeam." *HabitatMap Is an Environmental Tech Org and Maker of AirBeam*, www.habitatmap.org. Accessed 27 Apr. 2024.

Lim, Michael Heimbinder & Chris. "AirBeam3 Technical Specifications, Operation & Performance." *HabitatMap Is an Environmental Tech Org and Maker of AirBeam*, 17 Nov. 2021, www.habitatmap.org/blog/airbeam3-technical-specifications-operation-performance.

https://forms.office.com/Pages/ResponsePage.aspx?id=408fApwrJEiDeLvPnsWsy-70odOGuUtLnZRzc1TV1ZBUNFlaMjZBVkNaQUZXN09JREEzM0dTMU9BVy4u