Team 2 Deliverable 1 Report

1. Create a data set for FY21 that overlays the data from these 2 vendors into a single file.

(a) Account for factors such as summer break in analysis.

In general from combined graphs we can witness that summer break periods from May to September are usually represented with shorter spikes with gradually longer rising periods to reach max PSI. It indicates that in summer it takes from 2 to 3 times longer to reach peak level PSI and peaks are lower meaning time between pickups are longer and amount of trash generated is smaller. It proves an obvious correlation that the summer period generates the least amount of trash.

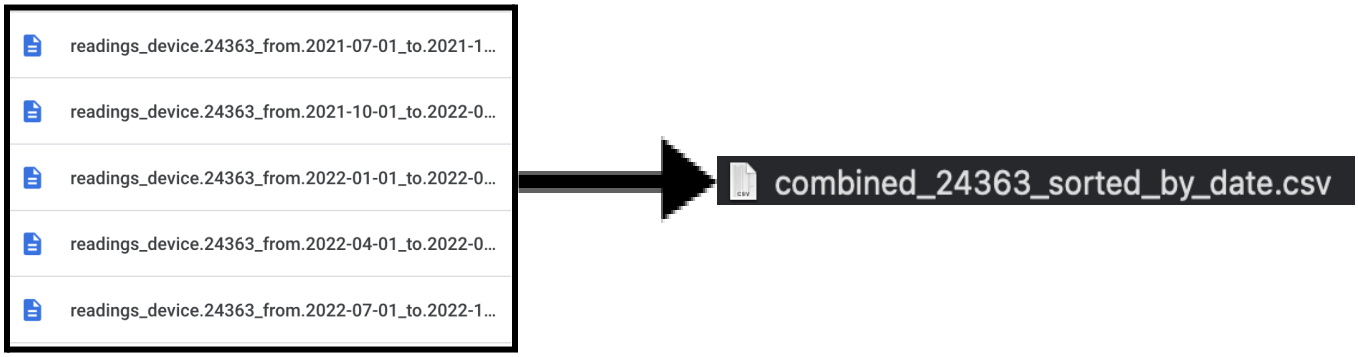
2. Answer a series of questions through data analysis to help BU sustainability implement the Zero Waste plan:

(a) Does temperature impact waste generation?

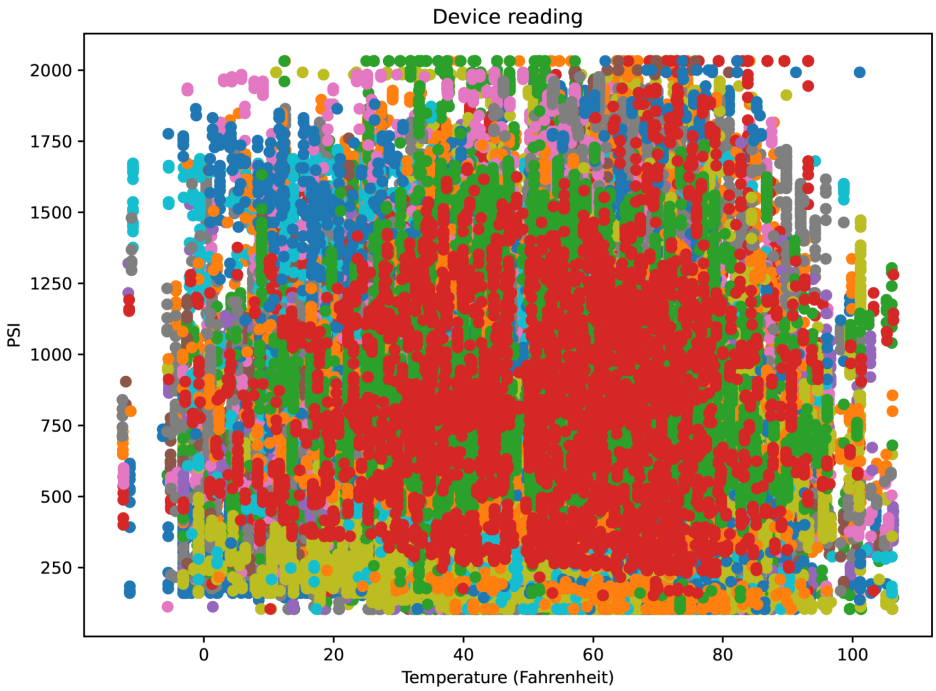
(All analysis below is assuming the higher the PSI value, the more waste it's generated)

Here, our strategy is to analyze the mass data. We did some pre-process for the **reading_device.#serial#_from.*.csv** files. (* means the date range of the files)

First, we observed for each device, the documents are separated by a certain time frame, so we want to combine the files together based on their device number:

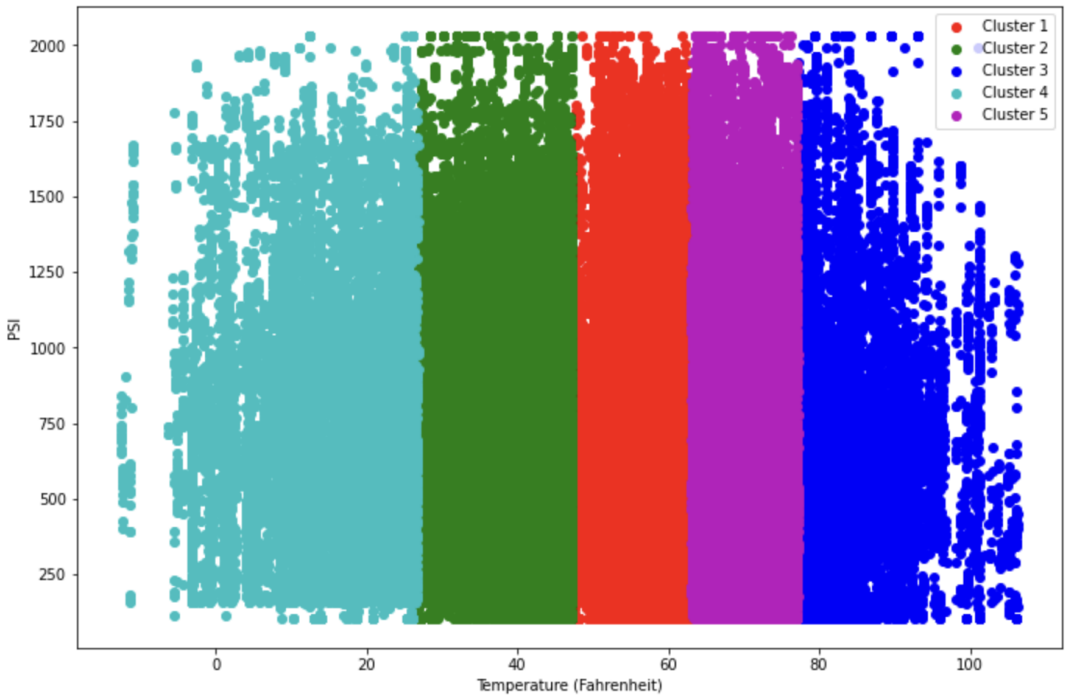


Second, we try to make scatter plots for the data based on temperature(fahrenheit) and PSI values. The color represents different **combined_#serial#_sorted_by_date.csv files:**



As we can see from the graph above, it is very hard to tell if there's a relationship between temperature and the PSI value simply by plotting them together.

Third, we try to make k means clustering for these points, the graph below is when k=5:



Clearly this graph is still not able to tell us the relationship between the temperature and the PSI.

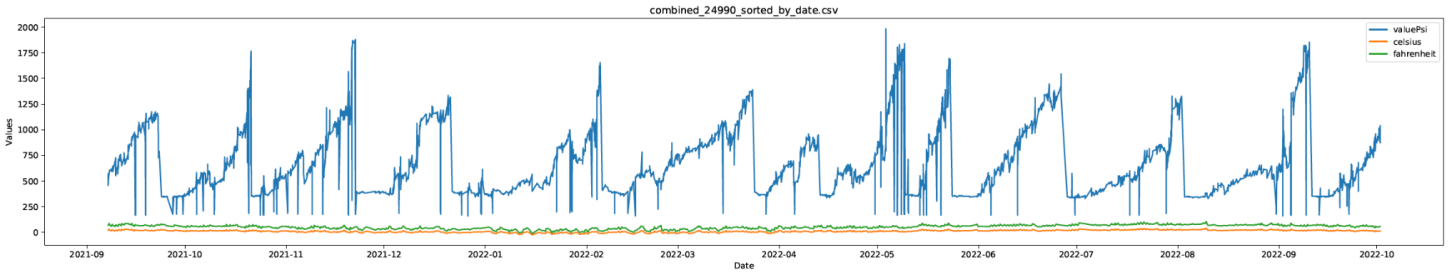
Then we try to make the correlation index between fahrenheit and valuePsi:

```
correlation = combined_df_clean['valuePsi'].corr(combined_df_clean['fahrenheit'])
print(f"The correlation between psi value and temperature is: {correlation}")
```

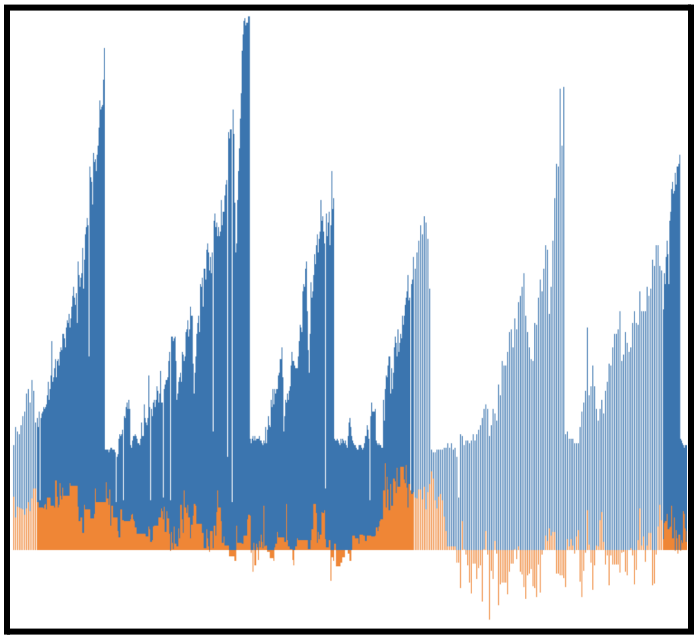
```
The correlation between psi value and temperature is: -0.05108003761451929
```

The correlation coefficient ranges from -1 to 1, with values closer to 1 indicating a strong positive correlation between two variables, values closer to -1 indicating a strong negative correlation, and values close to 0 indicating no correlation. Since the correlation coefficient is close to 0, there is no clear relationship between psi value and temperature.

Then we try to look at each device separately. We plotted line charts for each of them, and you can see from the file **lineChart_for_each_device.pdf**, and here is an example of showing device #24990:

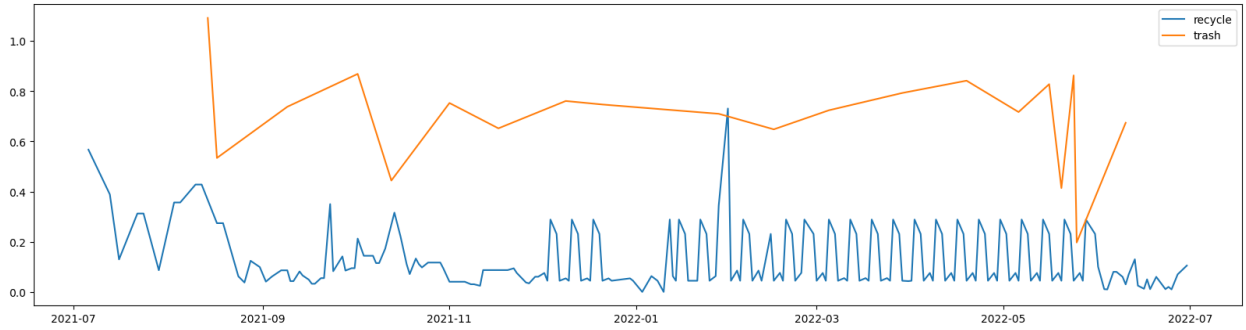


It is difficult to draw conclusions since we are still working with data and some parts of it are missing and we can't pin locations to data id's. But preliminary analysis on some data shows positive trend of temperature and PSI correlation like for data from device #24990 where negative temperatures show slightly lower spikes rather than positive temperatures. So we can assume that negative temperature affects the PSI. However we need to do more in depth analysis possibly month by month.



(b) If so, in what ways (i.e. more recycling, more of all materials, less recycling, etc.).

Since we can't fully connect data there is no certain way to say if there are more recycling materials or not. However we can graph some data such as one of the largest data from **BU_Daily_Weights_FY22 31769** where in general more trash than recycling material is produced. It can also be noticed that in colder periods of time there is less recycling material produced whereas warmer periods produce very symmetrical output like from march to may.



We are still looking for different strategies to figure out the relationship between waste generation and weather conditions.

(c) Can we use temperature as a predictor of waste generation and service level requirements?

Deliverable 1 - Analysis

Sufficient data should have been collected to perform a preliminary analysis of the data and attempt to answer one question relevant to your project proposal which you will submit as a pull request.

If data has already been collected for your project you must answer two questions.

1. Collect and pre-process a preliminary batch of data
2. Perform a preliminary analysis of the data
3. Answer 1-2 key questions
4. Submit all of the following information (code, notebooks, answers to questions) as a PR to your team's branch on github. (Add your PM and TE as reviewers!)
5. Submit the Weekly Scrum report to the gradescope and upload to google drive.
    a. Make sure to identify which team member is doing which tasks on the scrum report