BU Sustainability: Understanding How Weather Impacts Waste

Spring 2023 CS506 Data Science

Team 1 - Deliverable 3

Team Member Details and Contributions:

1. Karan Vombatkere - 2nd year PhD, kvombat@bu.edu
    - Preliminary analysis on waste generation and compactions
    - Data audit for readings, daily weights, psi readings, alert, events files
    - Data merging process and joining across different files
    - Timeline plots for each site by different material type, and analysis by type.
    - Clustering sites by tonnage and temperature
2. Junyi Zhu - 2nd year Master, astalos@bu.edu
    - Dataset Augmentation
    - Location-based PSI correlation pattern analysis
    - Predictive model
3. Abdelazim Lokma - 4th year Undergraduate, alokma@bu.edu
    - Augmented Dataset to include weather data
    - Geolocated Container Sites in BU Daily Weights
    - Graphed relationship between Temperature, Waste tonnages (by type) and Time
    - Generated two HeatMaps
    - Generated Dynamic Graphs

Notes:

We performed a thorough audit of the data and have performed a preliminary data analysis of the various datasets provided by the client. All tasks were completed successfully. Each team member's name and Github branch is listed above their part of the work towards the deliverable. The notebook of each team member can be found in their updated GitHub branch.

Introduction - Project goal and overview

The objective of this project is to assist BU's Sustainability department in achieving their mission of transforming Boston University's planning, operations, and culture towards a sustainable and equitable future. BU Sustainability supports the transformation of Boston University's planning, operations, and culture toward a sustainable and equitable future. Specifically, the project aims to investigate the relationship between weather factors and waste production and storage.

We have detailed data sets from each monitor, overall data, temperature data, and waste generation spreadsheet by data from Casella. The analysis of these datasets will inform BU Sustainability how they can potentially improve where to store waste if there are adverse weather effects. By conducting a comprehensive analysis of the data provided by the Sustainability Department, and their third-party vendors, the project seeks to help the department better understand how external factors affect the waste collection process. Ultimately, our team hopes to provide insights that can help improve the waste collection process and support BU's sustainability efforts.

Base Analysis

Data Exploration and Audit Process (KV)

- We explored different aspects of the dataset - PSI, temperature, waste, alerts, events, notifications, compactions, etc.
- Explored preliminary correlations between temperature, weather and the trash (tonnage and PSI) for each individual device. (outlined in earlier deliverables)
- Performed a full data audit and preliminary analysis of each dataset. Here is a brief overview of compiling the multiple data sources:
    - Individual device readings - data, weather, PSI
    - Device Daily weights - Daily tonnage of waste by site and waste material type (trash, recycling, compost)
    - Alert flag history - alerts triggered by date and site
    - Hauler response, notifications and events - additional event data and notifications
    - Device compactions - total number of compactions by site, hourly breakdown
- Successfully audited and compiled temperature, PSI, and site for 21 Sites. Identified methods to merge daily weight trash type and tonnage data for available date ranges.
- Challenges faced while joining data sources
    - Serial numbers do not match across files (2 different sets of serial numbers, as well as a device ID)
    - Site names aren't listed in a consistent manner
    - Date overlap between files is minimal - not enough overlapping dates to get a comprehensive overview with all features.
    - Min and max dates are offset by several months between files such as PSI readings

Data Merging Process (KV)

As part of one of the required deliverables for the project, our task was to compile and combine various different datasets from Casella and Contelligent, to ensure all the different data sources are being incorporated correctly into a comprehensive dataset.

After the initial data exploration and audit process, and gaining a better understanding of the datasets through client meetings the following methodology was developed to generate a comprehensive dataset.

1. **Concatenate Device Readings file by device_id:** First, the different device readings files were concatenated for the different date ranges available for each device. The device_id was extracted from the file name and all readings data, keyed by timestamp for the different devices were appended into a single dataframe. The following figure shows the schema of the concatenated readings data.

	timestamp	valuePsi	celsius	fahrenheit	device_id	serial_no	date
0	2022-07-01 04:35:52	256	20.20	68.360	39671	31180	2022-07-01
1	2022-07-01 04:36:39	200	20.20	68.360	39671	31180	2022-07-01
2	2022-07-01 04:37:11	240	20.20	68.360	39671	31180	2022-07-01
3	2022-07-01 06:42:44	200	20.20	68.360	39671	31180	2022-07-01
4	2022-07-01 08:04:31	280	22.52	72.536	39671	31180	2022-07-01
5	2022-07-01 08:04:43	184	22.52	72.536	39671	31180	2022-07-01

2. **Create Device_ID : Serial No, and Serial No: Site Mappings:** The Events data file was used to extract a device_id to serial number mapping as well as a Serial No: Site mapping. These mappings helped join data between different files, since different files used different ID values.
3. **Add Site Name to Readings Data:** Using the mappings created in the previous step, the *unique* Site Name corresponding to the device id/serial number for that reading was joined onto the dataframe.
4. **Aggregate data to Daily ranges:** Since the original raw data provided was at an extremely detailed timestamp level, in order to analyze further at a suitable

level of granularity, the data was grouped by the multi-index: [Site, Date] and then mean aggregated. The result was a single data point with the mean values corresponding to each Site-Date combination in the dataset, for each of the sites, and for the entire timeline of data provided in the raw readings files. The following screenshot shows the format of the dataset, with the data aggregated to the Site-Date level for each site.

	device_id	serial_no	date	valuePsi	celsius	fahrenheit	site_name
0	39671	31180	2022-07-01	253.333333	24.822500	76.680500	BU #82 Warren Towers 35
1	39671	31180	2022-07-02	331.200000	28.178000	82.720400	BU #82 Warren Towers 35
2	39671	31180	2022-07-03	280.400000	26.006000	78.810800	BU #82 Warren Towers 35
3	39671	31180	2022-07-04	279.272727	24.530909	76.155636	BU #82 Warren Towers 35
4	39671	31180	2022-07-05	182.800000	22.599000	72.678200	BU #82 Warren Towers 35
...	...	...	...	...	...	...	...
7161	39880	31186	2022-09-27	502.000000	19.840000	67.712000	BU #69 - Graduate Apartments
7162	39880	31186	2022-09-28	525.333333	17.156667	62.882000	BU #69 - Graduate Apartments
7163	39880	31186	2022-09-29	552.000000	15.150000	59.270000	BU #69 - Graduate Apartments
7164	39880	31186	2022-09-30	590.000000	12.020000	53.636000	BU #69 - Graduate Apartments
7165	39880	31186	2022-10-01	590.000000	14.400000	57.920000	BU #69 - Graduate Apartments

5. **Site Key Extraction:** A custom method was written to generate a unique key from the string name for each site. For example the key ‘bu#82’ corresponded to the unique Warren Towers site. The goal of creating a unique site key was to facilitate an easy joining process further down the line when attempting to merge the data with the Daily Weights data provided.
6. **Preprocess Daily Weights Data and PSI data:** The daily weights data was preprocessed to ensure the site name corresponded to the same site keys outlined in the previous step. Similar steps were performed on the PSI values dataset, however due to a date range mismatch, most of the PSI data provided could not be used.
7. **Use SQL Inner Join to merge:** The Daily weights dataset was successfully merged with the concatenated device readings data using the join key [Date, Site Key]. Some readings data that was outside the date range available in Daily weights was dropped. The following screenshots show the final dataset

that was created by merging the different sources, and the number of reading data points and date ranges available.

- **Date ranges and counts of readings available by Site**

	site_name	min	max
0	BU #102 Student Health Services	2021-09-02	2022-06-16
1	BU #105 Kilachand Hall	2021-08-24	2022-05-24
2	BU #2 Student Village	2021-09-03	2022-06-30
3	BU #35 22 Babbit	2021-08-31	2022-06-30
4	BU #38 Life Sciences	2021-08-31	2022-06-28
5	BU #4 Yawkey	2021-09-10	2022-06-30
6	BU #43 West Loading Dock	2021-08-24	2022-06-30
7	BU #46 30 Bay State	2021-08-31	2022-06-30
8	BU #48 Student Village #2	2021-10-01	2022-06-30
9	BU #69 - Graduate Apartments	2021-08-31	2022-06-24
10	BU #72 Rafik B Hariri	2021-09-07	2022-06-30
11	BU #82 Warren Towers 35	2021-09-02	2022-06-30
12	BU #87 College of Engineering	2021-08-27	2022-06-24
13	BU #90 - School of Law	2021-08-20	2022-06-27
14	BU #93 George Sherman Union	2021-08-27	2022-06-30
15	BU #96 808 Commonwealth	2021-08-27	2022-06-24
16	BU 140 Bay State Rd	2021-08-20	2022-06-24
17	BU MED - 15 Stoughton	2021-09-07	2022-06-30
18	BU Med #815 Albany	2021-12-10	2022-06-30
19	BU Med 700	2021-09-17	2022-06-30

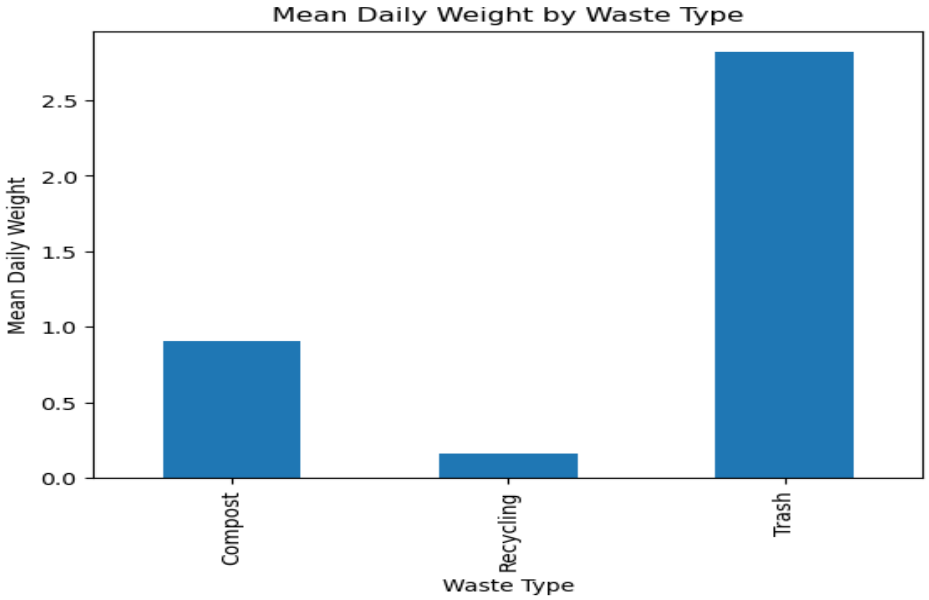
```
BU #35 22 Babbit                396
BU #82 Warren Towers 35         394
BU #2 Student Village           392
BU #90 - School of Law          388
BU #87 College of Engineering   385
BU #46 30 Bay State             382
BU MED - 15 Stoughton           379
BU #43 West Loading Dock        377
BU #38 Life Sciences            377
BU #72 Rafik B Hariri           372
BU #48 Student Village #2       361
BU 140 Bay State Rd             360
BU #4 Yawkey                    359
BU 685 Comm Ave                 356
BU #93 George Sherman Union     336
BU #96 808 Commonwealth         335
BU Med 700                      279
BU #69 - Graduate Apartments    270
BU #105 Kilachand Hall          240
BU #102 Student Health Services 236
BU Med #815 Albany              192
Name: Site, dtype: int64
```

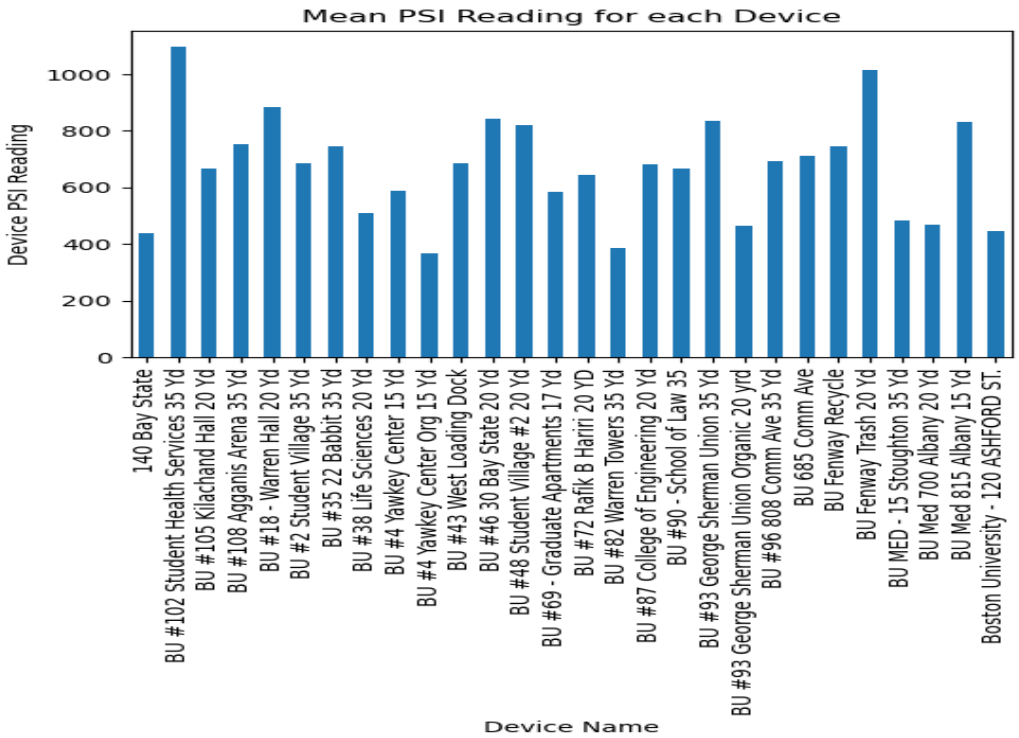
- **Schema of merged data available at Site-Date granularity.**
- **Features:** Date, PSI, Temperature, Site Info, Material, Waste Tonnage

device_id	serial_no	date	valuePsi	celsius	fahrenheit	site_name	site_key	Material	Tons
24668	30932	2021-10-01	394.000000	15.765000	60.377000	BU #48 Student Village #2	bu#48	Recycling	0.0565
24668	30932	2021-10-02	431.636364	15.744545	60.340182	BU #48 Student Village #2	bu#48	Recycling	0.2830
24668	30932	2021-10-04	576.666667	14.909167	58.836500	BU #48 Student Village #2	bu#48	Recycling	0.0920
24668	30932	2021-10-05	726.000000	14.385000	57.893000	BU #48 Student Village #2	bu#48	Recycling	0.0920
24668	30932	2021-10-07	866.222222	18.500000	65.300000	BU #48 Student Village #2	bu#48	Recycling	0.0920
...	...	...	...	...	...	...	...	...	...
39880	31186	2022-06-07	492.000000	21.960000	71.528000	BU #69 - Graduate Apartments	bu#69	Recycling	0.0250
39880	31186	2022-06-10	684.000000	23.440000	74.192000	BU #69 - Graduate Apartments	bu#69	Recycling	0.0250
39880	31186	2022-06-14	796.000000	26.220000	79.196000	BU #69 - Graduate Apartments	bu#69	Recycling	0.0300
39880	31186	2022-06-16	438.000000	19.980000	67.964000	BU #69 - Graduate Apartments	bu#69	Trash	3.9700
39880	31186	2022-06-24	360.000000	24.230000	75.614000	BU #69 - Graduate Apartments	bu#69	Recycling	0.0000

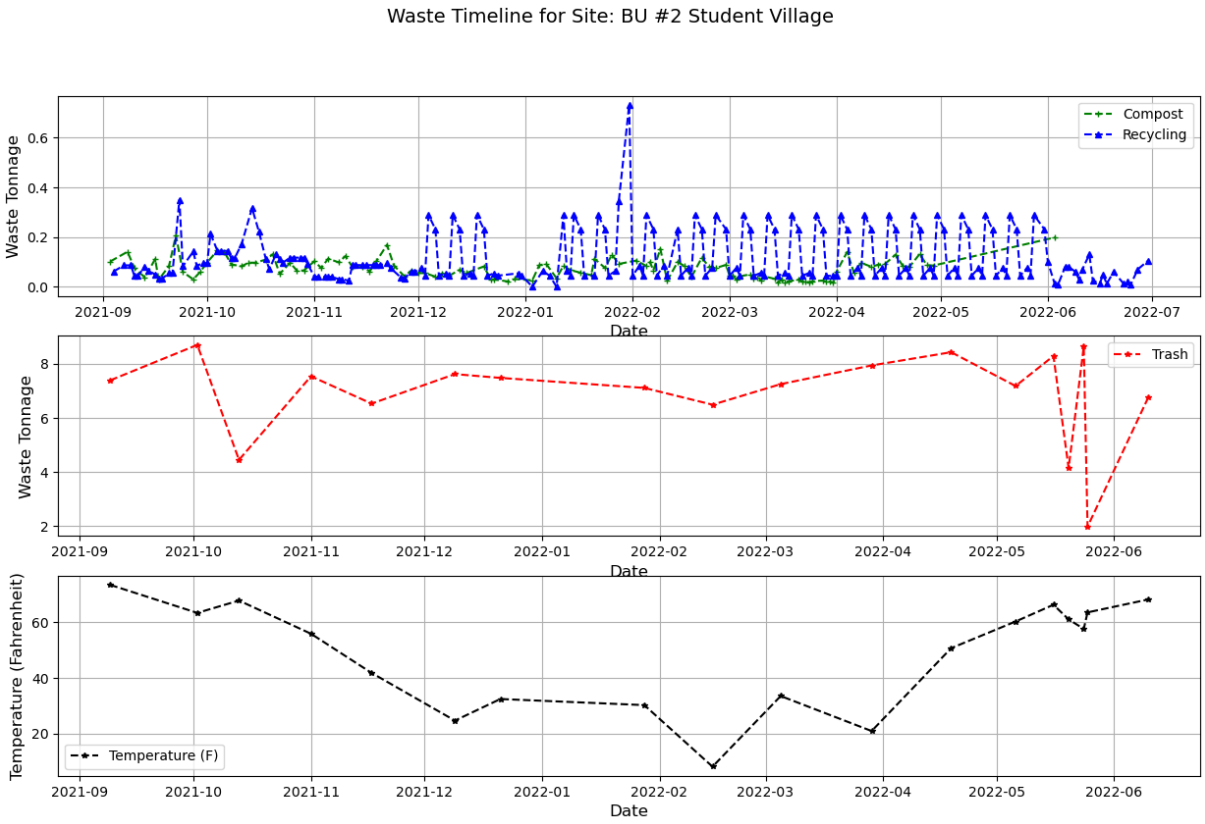
Preliminary Analysis overview (KV)

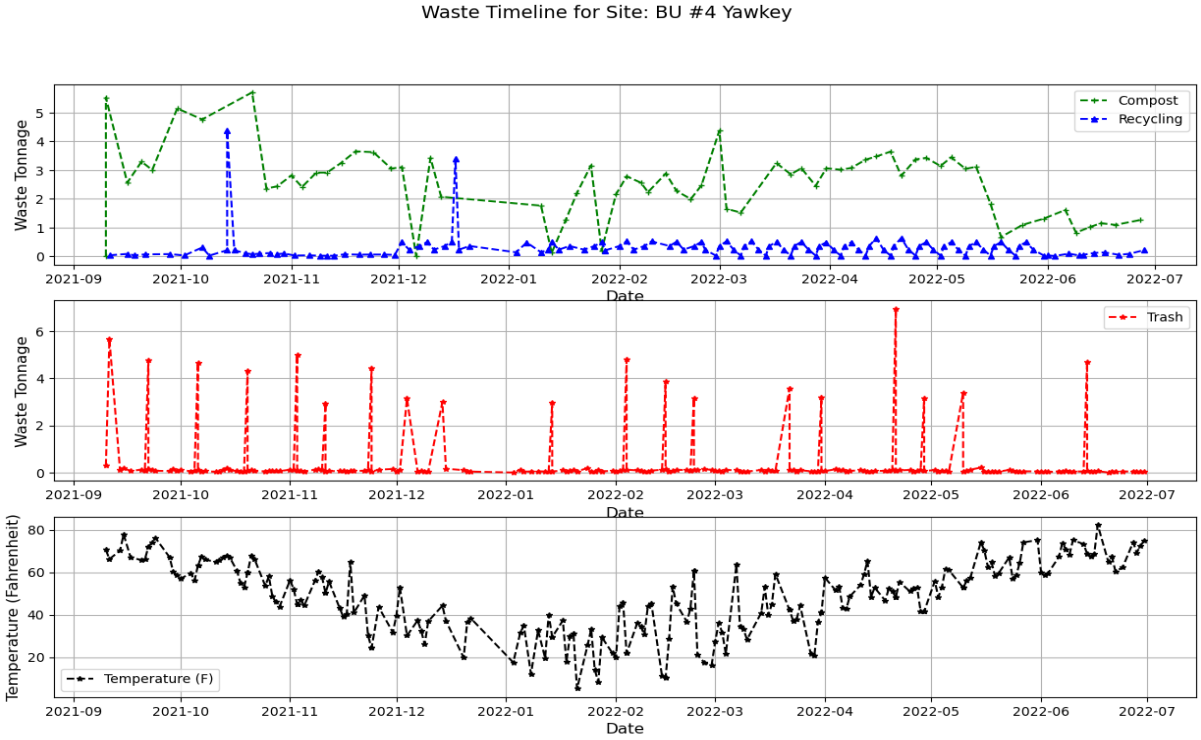
- **For each device, explore different patterns between waste generated, temperature and psi.** This was done at both the aggregate level and daily timeline level. Some of the plots below show the highlights of this preliminary analysis that motivated future work.

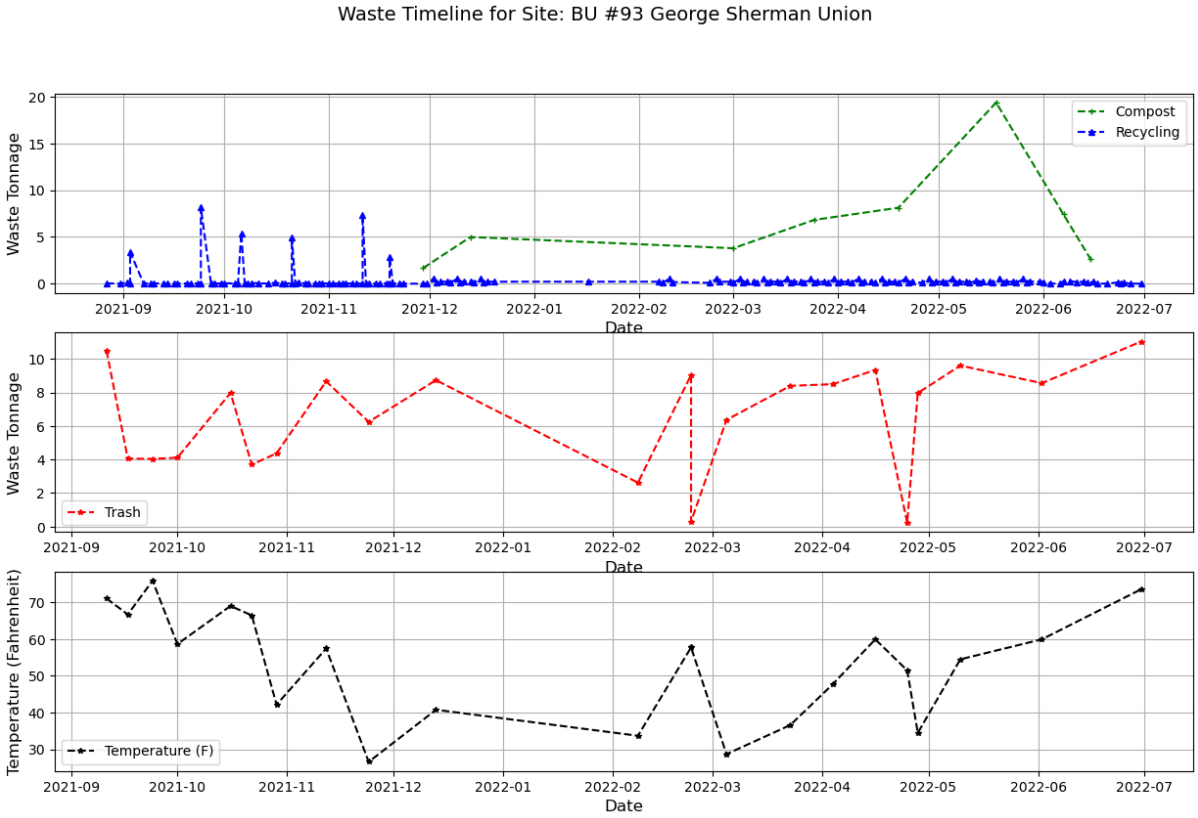




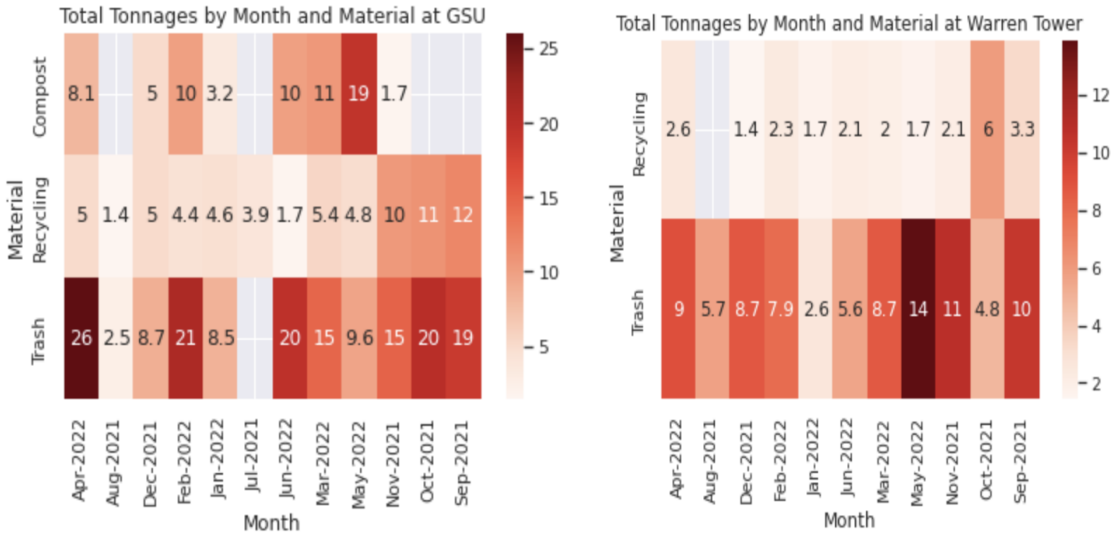
- **For each Site, plot timelines by waste type, and temperature.** Examine Pearson correlations between Temperature, Waste Tonnage and PSI values.



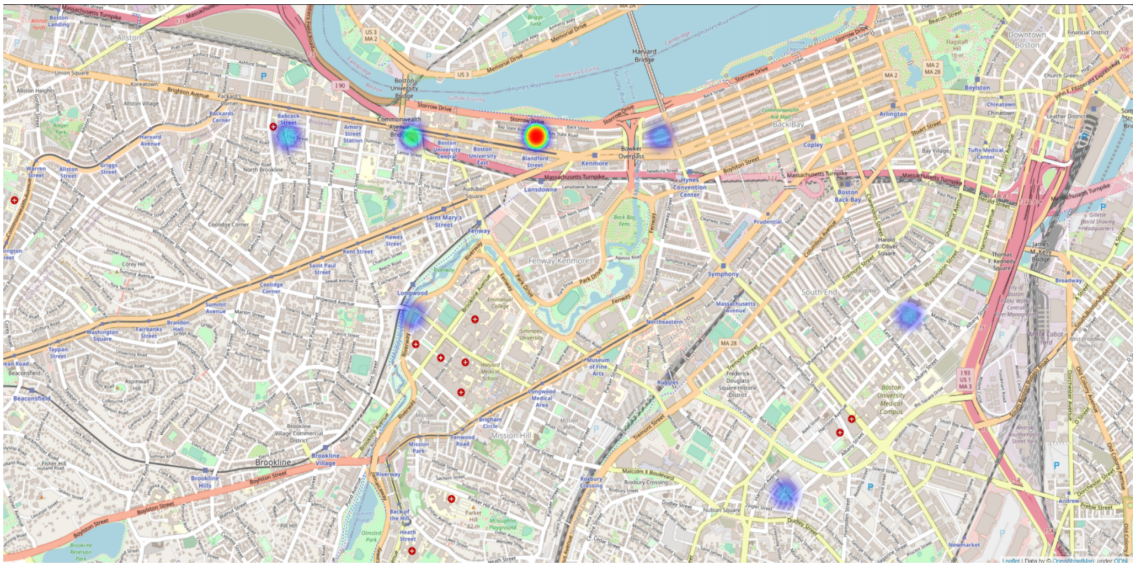




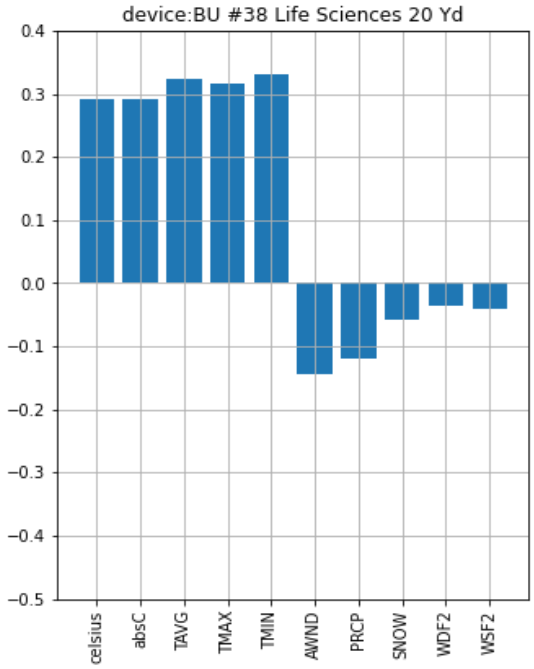
- **Tonnage changes at popular sites - focus on the locations that have the most waste, and also explore the seasonality of waste generation at these sites**

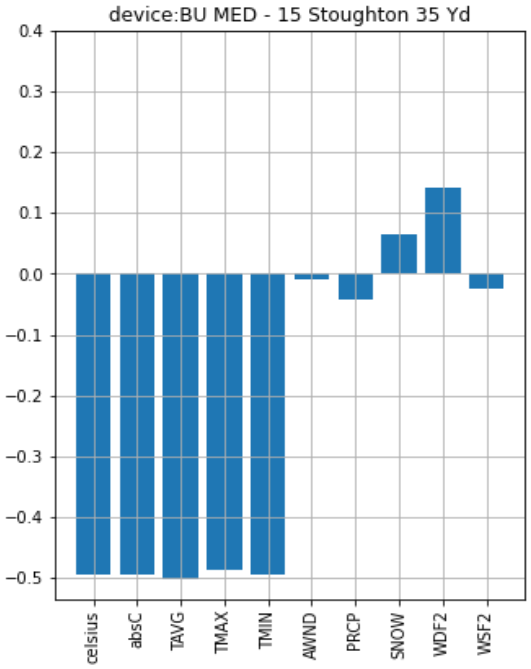


- **Heatmap for each location. Shows Comm Ave produced the most trash (Charles River). There's more traffic on the BU Main Campus compared to other campuses**

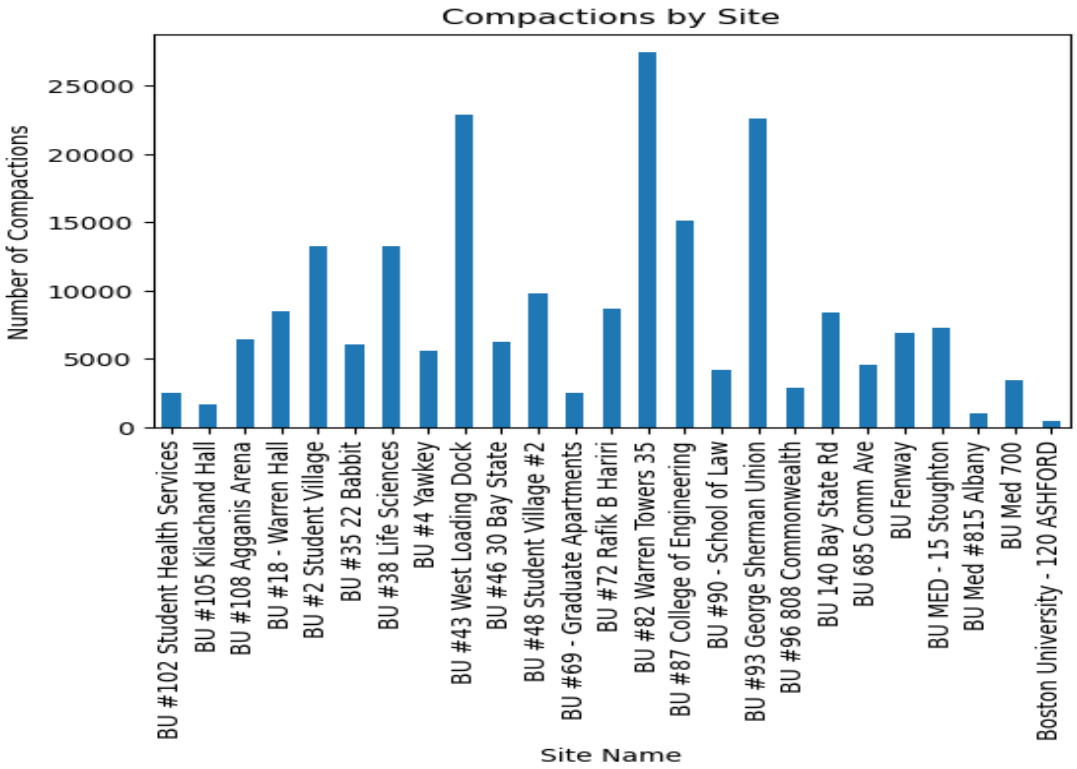


- **Correlation between temperature and tonnage of Trash containers. Collected extra data from NOAA database and performed analysis**

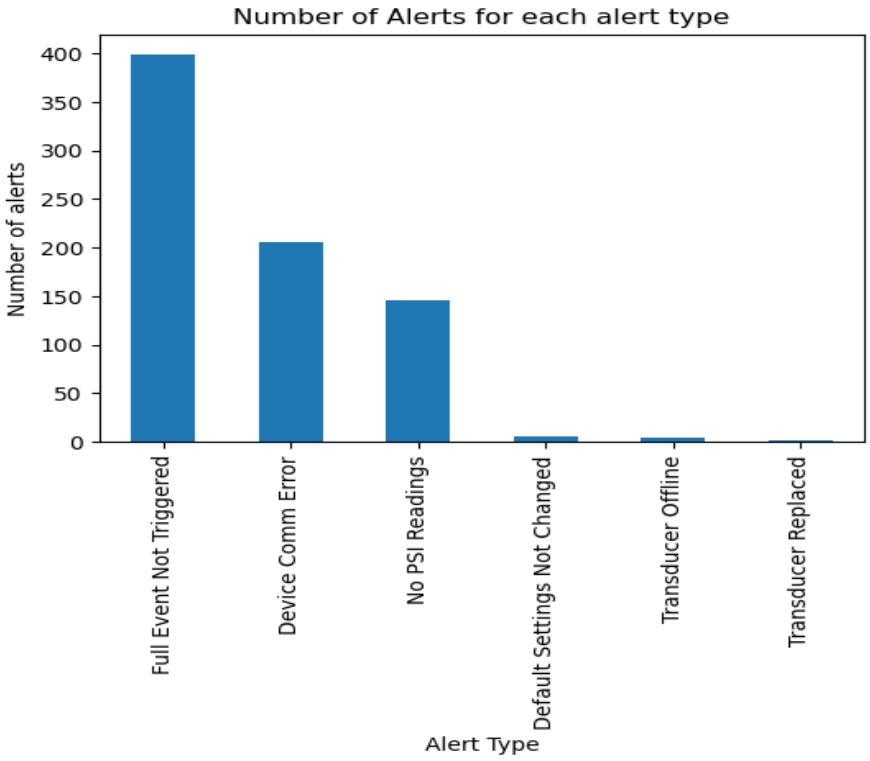




- **Compactions by site - Warren towers and GSU should be considered separately**



- **Analysis of Alert types**

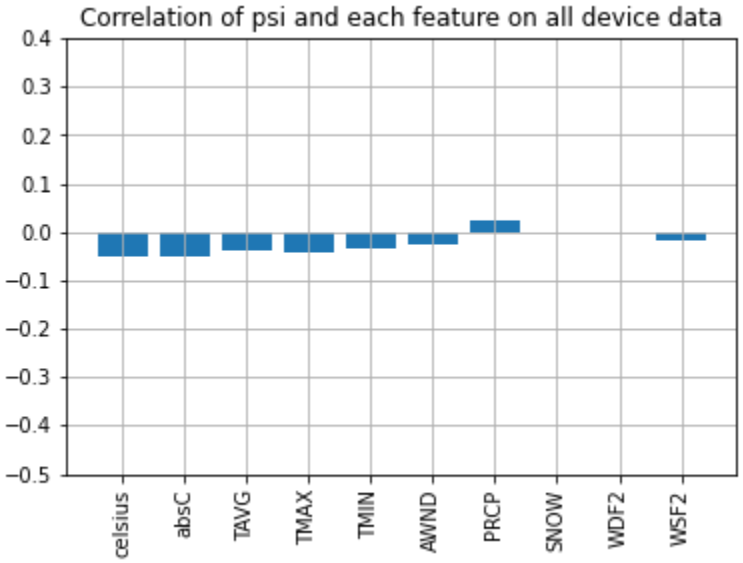


Does temperature impact waste generation?

(J.Z)

We collect data that includes temperature(in celsius), wind speed, rain/snow fall and calculate the Pearson correlation coefficient with PSI value.

The impact of temperature on waste generation is very weak on the macro level, as showed in the fig below that the absolute correlation for every feature is lower than 0.1.The result shows temperature features have the most significant influence on waste generation.



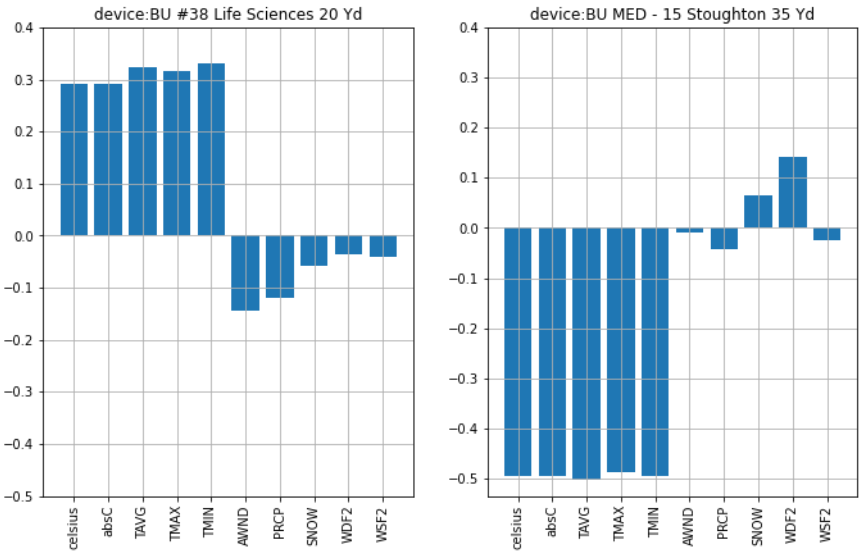
However, the pattern of each device(site) is different from each other(see (b)), so at least from this perspective we can say temperature impacts waste generation.

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

- Location:

(J.Z)

We analyzed a list of locations that generate waste the most by the correlation of features with PSI. The correlation to PSI figs (below) of Life Sciences and BU MED show completely different patterns. Life Sciences shows a positive correlation while BU MED shows a negative correlation, indicating that waste generation increases in Life Sciences and decreases in BU MED as temperature increases.



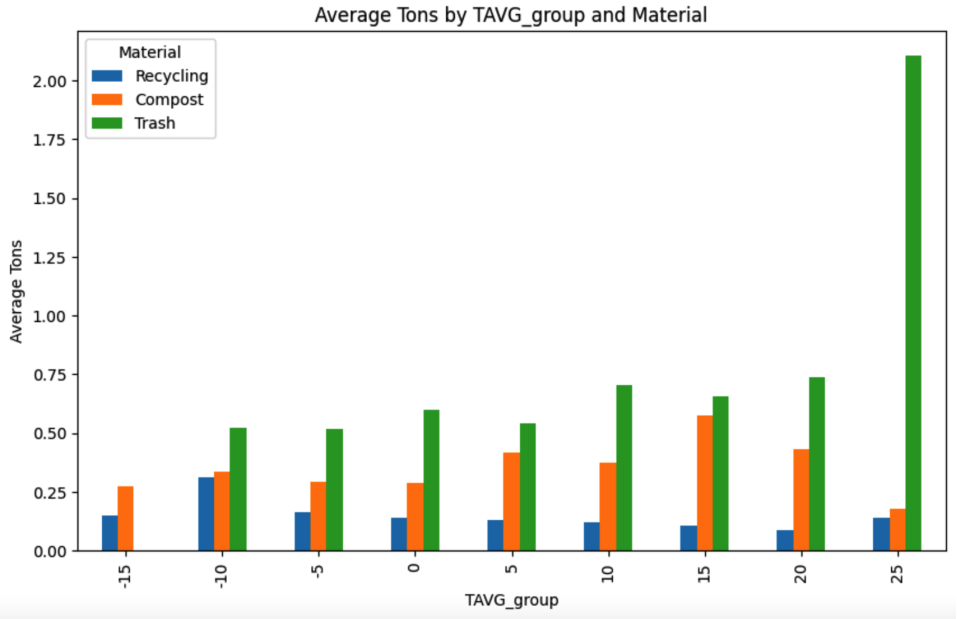
- Waste Type

(A.L):

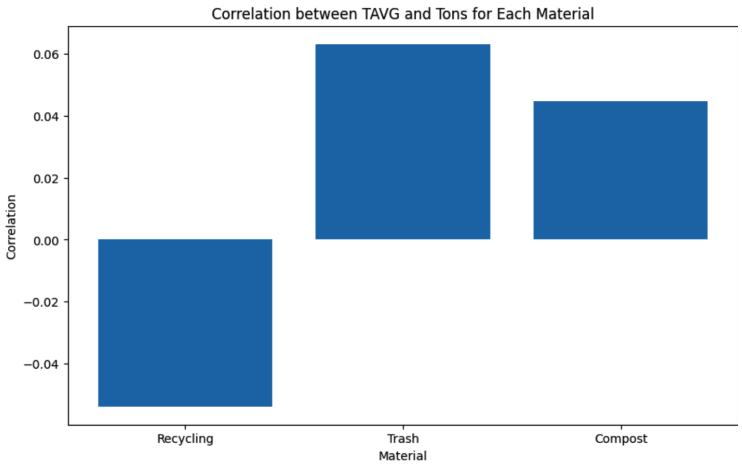
The “BU-Daily-Weights” document contains information on all waste containers across campus, and the tonnage of waste they contained when they were being emptied. I prepared this document by merging it with NOAA weather information on the days each container was emptied. To get a basic understanding of how tonnage relates to weather. I grouped each entry of the document into a temperature range in which the trash was emptied, and plotted the average tonnage by waste type collected in each temperature range.

The resulting graph above indicated to me a potential relationship between increasing temperatures and waste generation, so I decided to explore this relationship further.

In order to determine how waste output changes with time, I decided to explore
The graph below displays the Pearson Correlation coefficient between average daily temperature and the generation of Recycling, Trash, and Compost respectively. This coefficient is a number that measures how strong the relationship is between two variables (in our case, temperature and the three different waste types generated and collected).

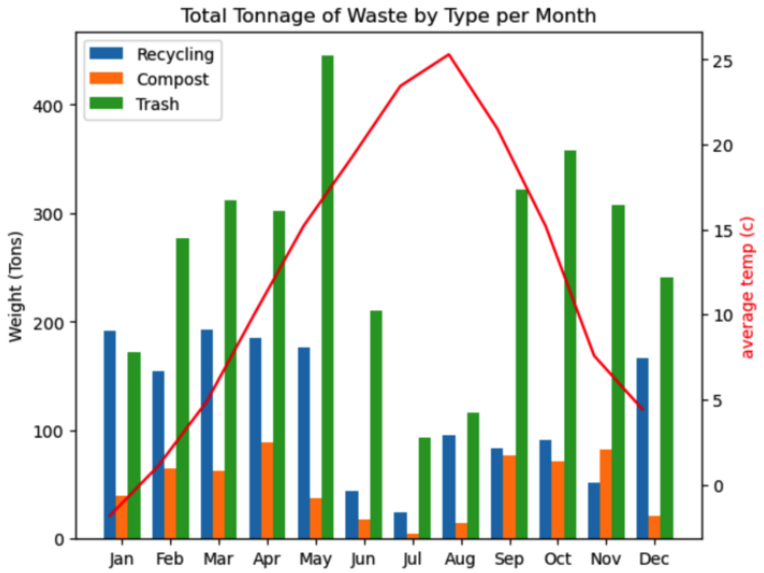


The results of the graph below shows a weak correlation between changes in temperature and changes in the generation of different waste types. Note that correlation does not necessarily equal causation, and that there are likely other factors that may drive waste output.



To get a better understanding of this correlation, we can graph the total waste output, per waste type (in Tons) across 12 months of monitoring. Also displayed in the graph below is a line representing the daily average temperature (in Celcius). As can be seen in the bar chart, the weak positive correlation between Trash/Compost

and temperature is caused by the increased output of waste during the Spring Semester (Jan - May), which is when temperatures are rising. The correlations are prevented from being higher due to the Fall semester & Summer period, during which we see either falling temperatures (from Sep - Dec) or decreased waste generation (Jun - Aug).



Overall, it seems to me that there is a weak correlation between temperature and waste generation. But correlation between two variables does not always mean causation between them. Taking a closer look at the graph might reveal a connection between campus activities and waste generation. As it seems clear that during the Summer break (June to August) the waste generated is significantly reduced as compared to the Spring and Fall semesters. I think it would be interesting to explore this in the extension questions of our report.

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

(J.Z)

It can be used as a feature of a predictive model, but far from deciding the result. There are many factors that make it unwise to predict the total waste generation of the area:

- The total waste generation is dominated by some large contributors like BU MED

- The date range of the data of different sites do not match
- We do not know the exact waste generation in a specific day, instead, we have to simulate in this way: we divide the tonnage data by the interval to simulate each day's generation.

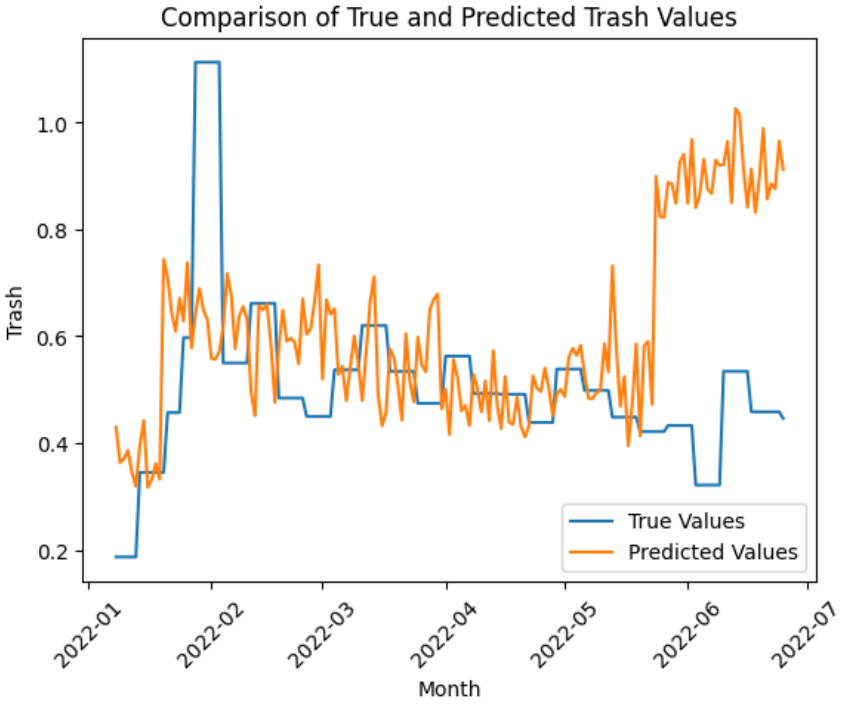
Therefore, we decide to predict a site's waste generation instead of the BU area.

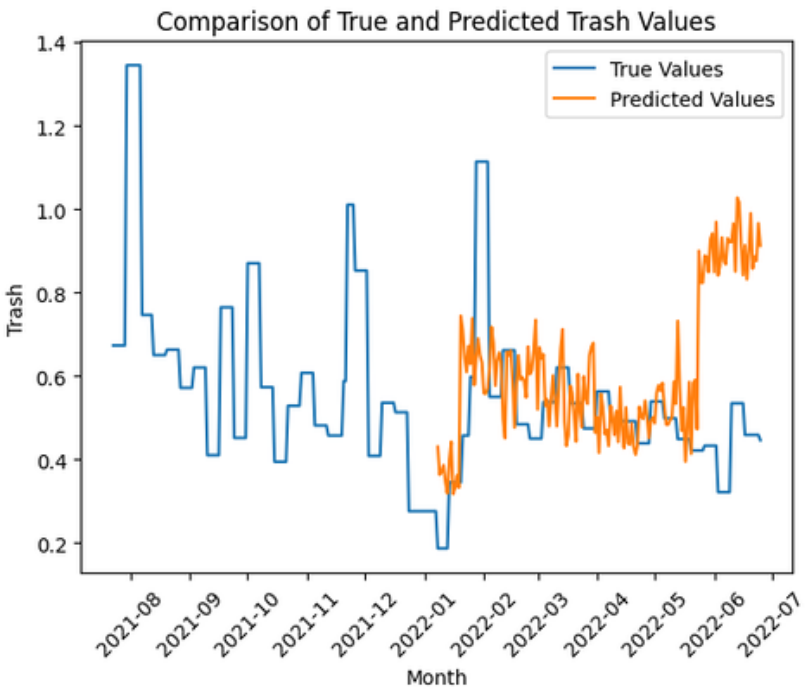
Here is a predictive model that we made on **BU MED- 15 Stoughton** (This is an example site, transferable to other sites), which uses LightGBM.
Features used to train the model:

- Date
- Semester
- Weather

Training method:
(The following figs show comparison of predicted trash generation and true trash generation in tons).

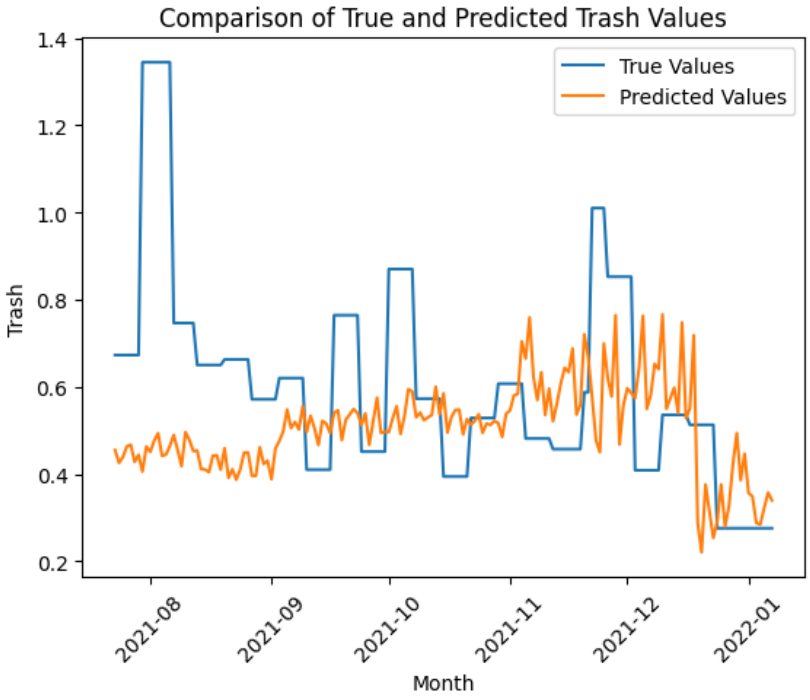
- Use the first 50% of the data for training and the last 50% for testing

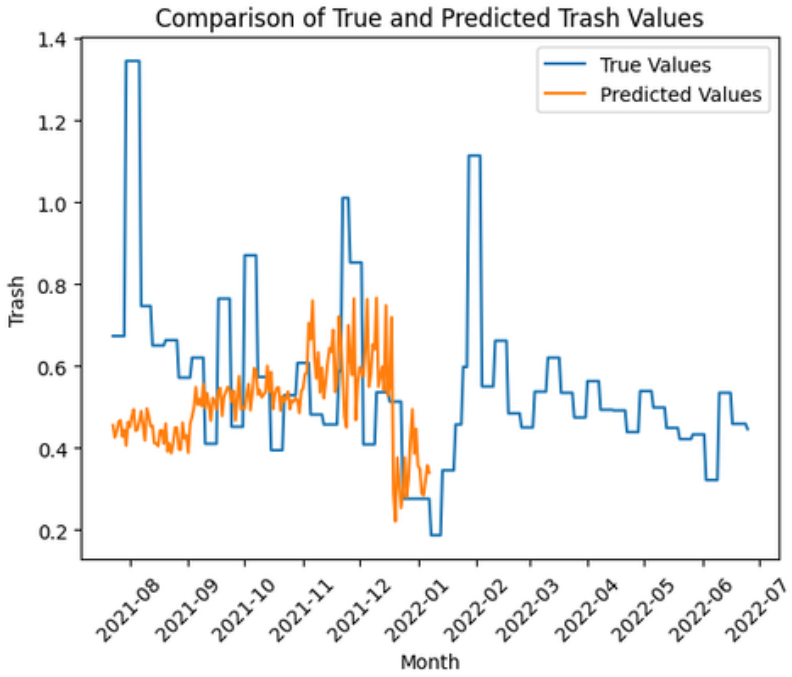




Because we use the first 50% data (2021) for training and the last 50% data (2022) for testing, it result in a problem that the model learns too well about the 2021 summer so that the model believes the 2022 summer will be the same.
However, instead of a model issue, this reveals a dataset problem. Our dataset is too small and full of coincidences such as festivals so that the model cannot learn too well and predict very precisely.

- Use the last 50% of the data for training and the first 50% for testing



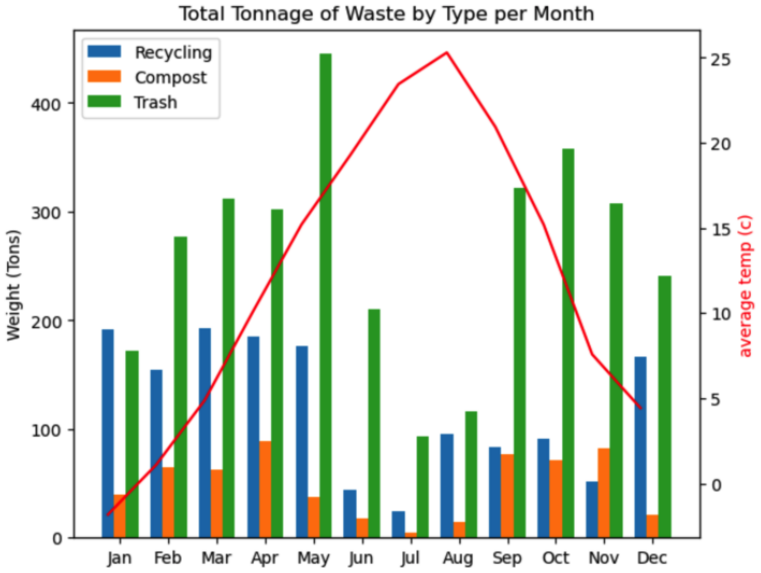


Similar issue also happens here. However, the figs indicate that it is possible to predict on these features about the trash generation, especially the wave-like pattern. Even the 2021 and 2022 data are completely different, the model learns well about the pattern and successfully complies with the general pattern.

Extension Analysis
(Overview of extension ideas and why we chose them - answer 2-3 key questions)

(A.L)
Does University Activity Impact Waste Generation?

Continuing where I left off, I decided to have a closer look at the graph We are looking at analyzing how waste generation changes during the Fall/ Spring Semesters, and how it changes during the Summer Period, when campus activities fall heavily. Based on the graph that we had previously shown, it seems that waste output decreases during the months in which



campus activity decreased. The Fall semester typically ranges from August to September, whilst the Spring semester ranges from January to May. The graph clearly shows a decrease in activity from June to August, which is when the Summer break occurs. This is inline with the fact that most students choose to not remain on campus for the entirety of the Summer break, resulting in decreased activity overall during that time. In addition to that, we can explain some of the outlying results in the graph such as the month of May, which shows a peak Trash output due to events such as Move Out and Commencement. Both of these events must have caused an increase in waste due to students disposing of their furniture or waste, as well as the increased campus activity caused from hosting students and their families during

Commencement. However, we would also like to see how waste generation changes for smaller events and breaks. Ultimately, it seems that when campus activities decrease, waste also decreases.

How do events and campus activity affect the container sites across campus, which sites contribute the most throughout the year?

To be able to visualize the waste generated by each site on a given day or month throughout the year, we plotted a HeatMap that plots a color on the map in relation to the tonnage of waste extracted from containers. These heatmaps contain markers showing all the containers across campus, color coded by the type of waste they store. Hovering over these markers gives you the site names. These heatmaps will be accessible through the links below so that the client can freely interact with and explore them. Hopefully using their experiential knowledge to get a better understanding of how waste generation occurs across campus throughout the year.

Monthly Heat Map Link

Daily Heat Map Link

Here are instructions for how to load these documents into your browser after downloading the files:

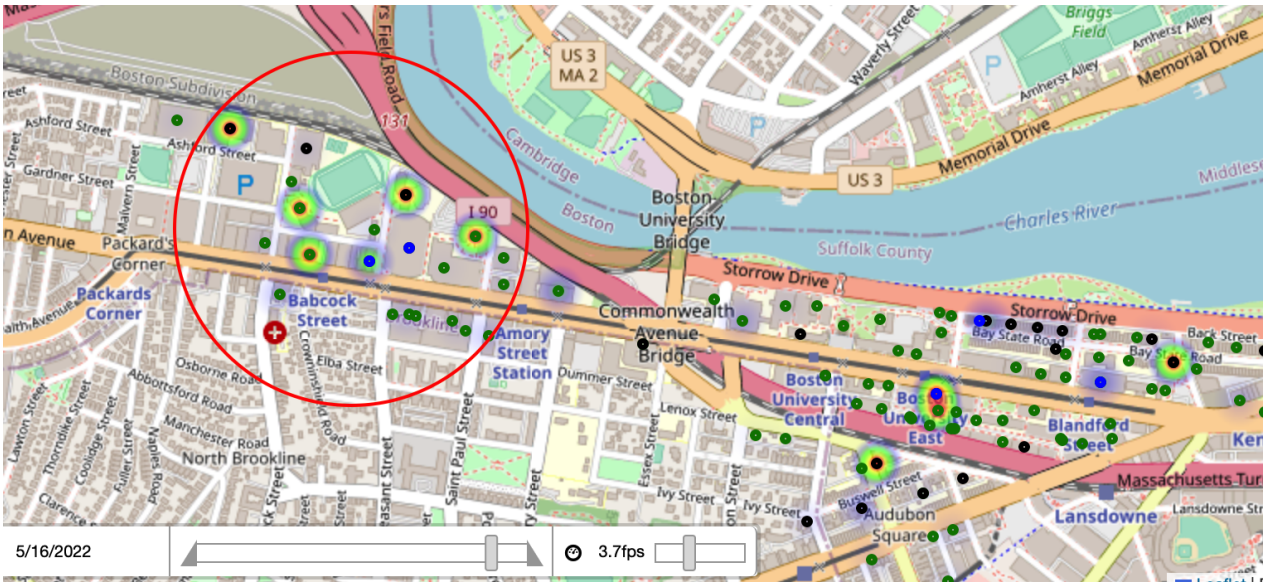
1. Open your web browser (e.g. Google Chrome, Firefox, Safari, etc.)
2. Click on File > Open File (or press Ctrl + O on Windows or Command + O on Mac)
3. Navigate to the location where you saved the HTML file and select it
4. Click Open
5. The HTML file should now open in your browser and you can interact with the heatmap

Let us observe the week in which Commencement occurred to see how the containers across campus are strained. Commencement weekend occurred

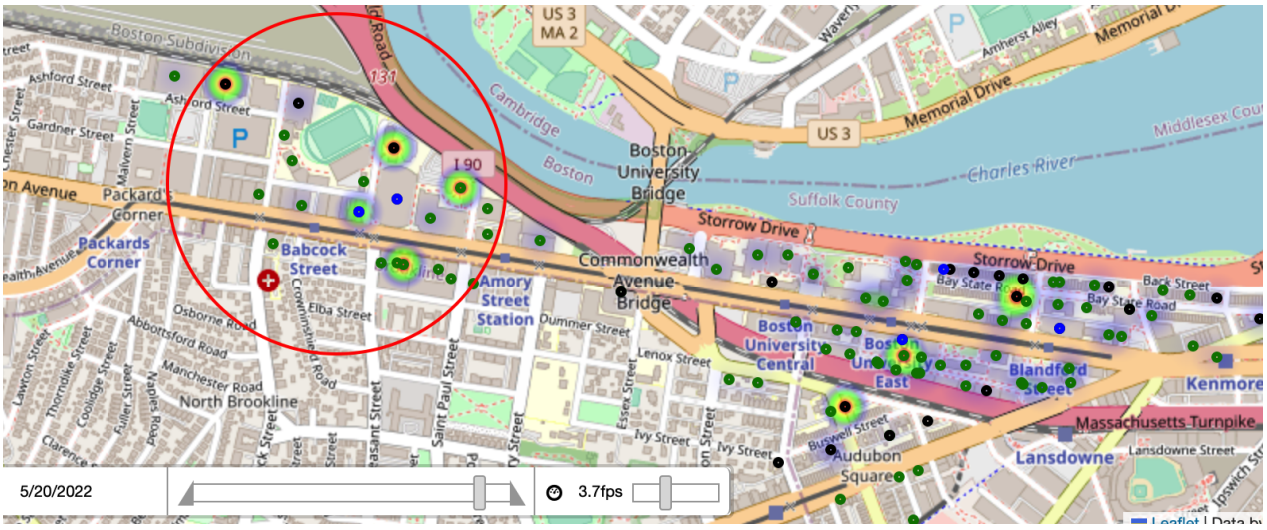
from the 19th to the 22nd of May. We will add a few days to this date range as a margin of error in case the containers were emptied at a delayed rate (at least 2 days according to the client). So we will take a 5 day range from the 14th to the 27th.

We can see from the results that there seems to be an increase in the intensity of the colors of the sites around Nickerson Field (where Main commencement occurs)

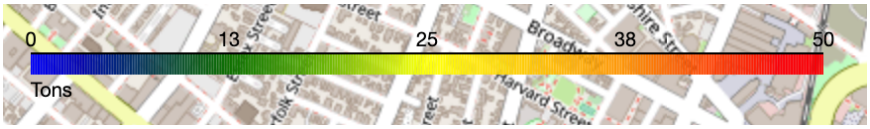
May 16th, 2022



May 20th, 2022



June 9th, 2022



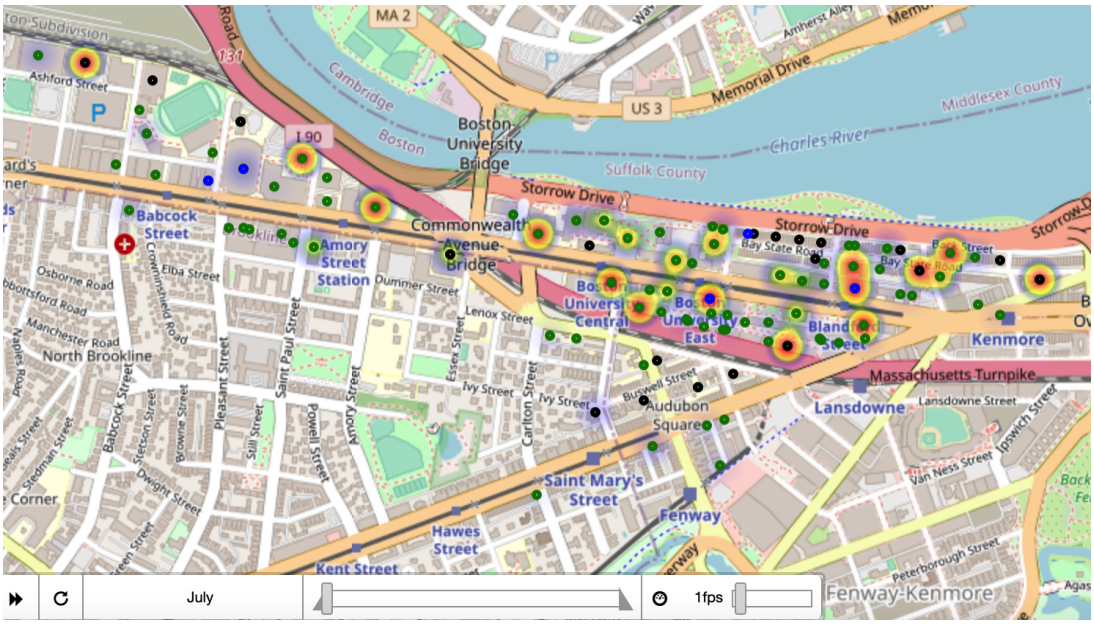
As you can see from the snapshots above, Many of the sites that are located around Nickerson Field show a very high tonnage range (from 38 to 40 tons). Indicating that Commencement may indeed have an impact on the waste generated, specifically in the sites surrounding the field. However the differences are not clear enough to conclude this without further analysis.

It is also worth noting that Sites such as the ones located at Warren Towers and Bay State show very high levels of waste due to the fact that they contain large food courts that feed many students on a regular basis.

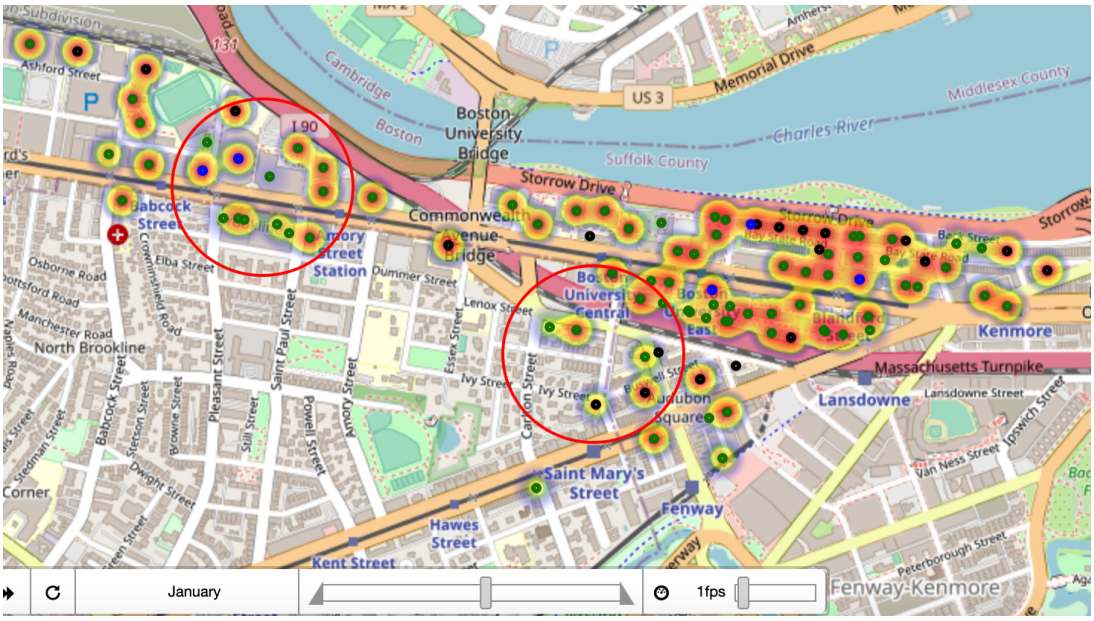
In conclusion, it seems as though in this current representation. It is not possible to use the heatmap in order to determine day to day changes in waste output as not all containers are emptied on the same interval. It would be more ideal to use a

monthly rate in order to see overall changes in waste generation across campus. Let's look at the month of May in comparison to the month of January.

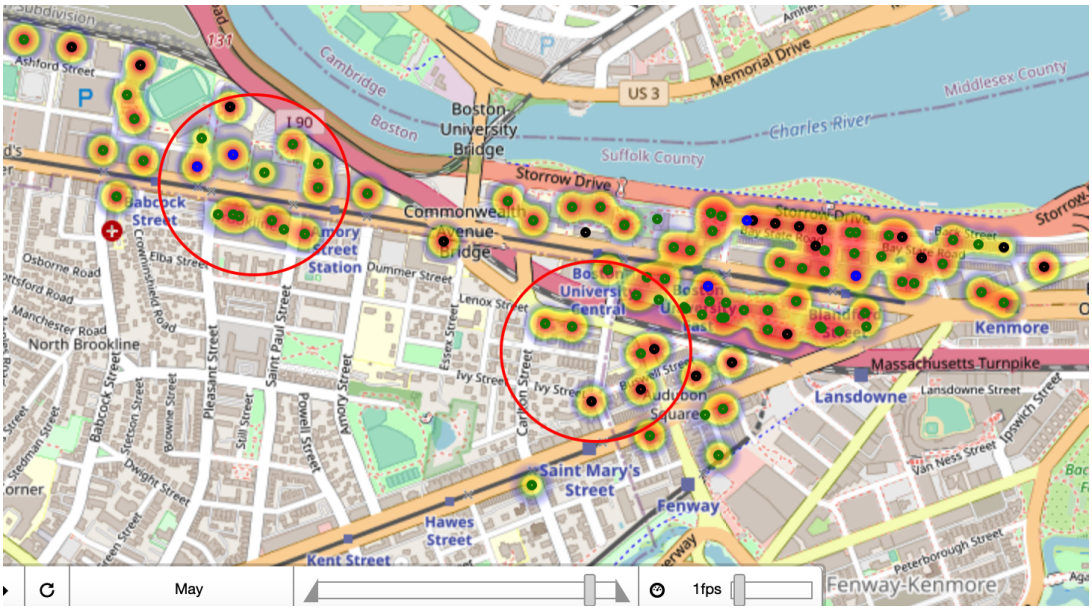
July 2021 (Summer Break)



January 2022 (1st month of Spring 2022)

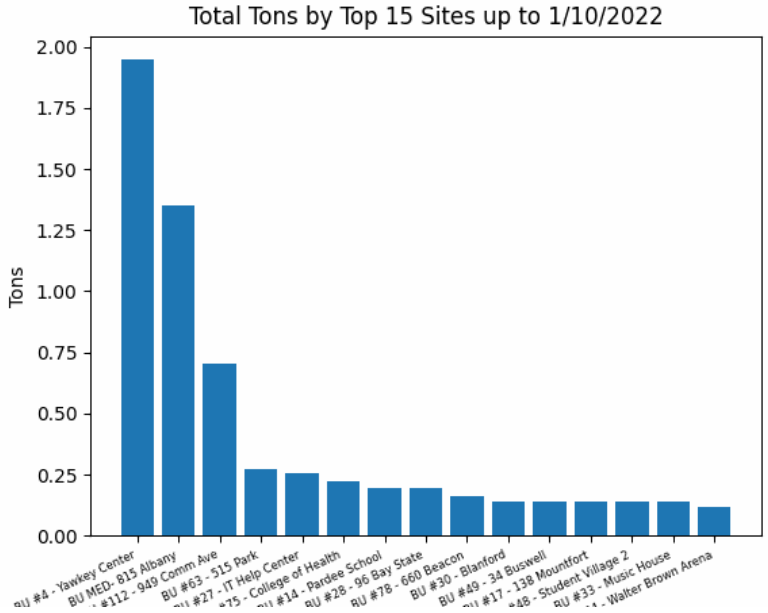


May 2022 (Last month of Spring 2022)



Though the differences are minute, we can see that some sites across campus show a much lower output in January compared to May, which is inline with the results of our first graph. Hower, during the summer break (especially in July) we can see how much of a difference there is in monthly waste output. With only sites such as Warren Towers, Bay State, GSU, College of Fine Arts, College of Engineering, Photonics, Graduate Apts. and others contributing to the majority of the waste.

To help visualize how the top contributing sites change over time, I've gone ahead and created a dynamic graph that shows which sites contribute the most over the year:



It appears that Warren Towers, Yawkey Center, George Sherman Union and West Loading Dock seem to the the top contributors throughout the year, momentarily exchanging places with other sites such as MED-15 Stougton, 120 Ashford and Student Village 2. (Since the deliverable is likely to be submitted as a PDF, please find the animated gif HERE)

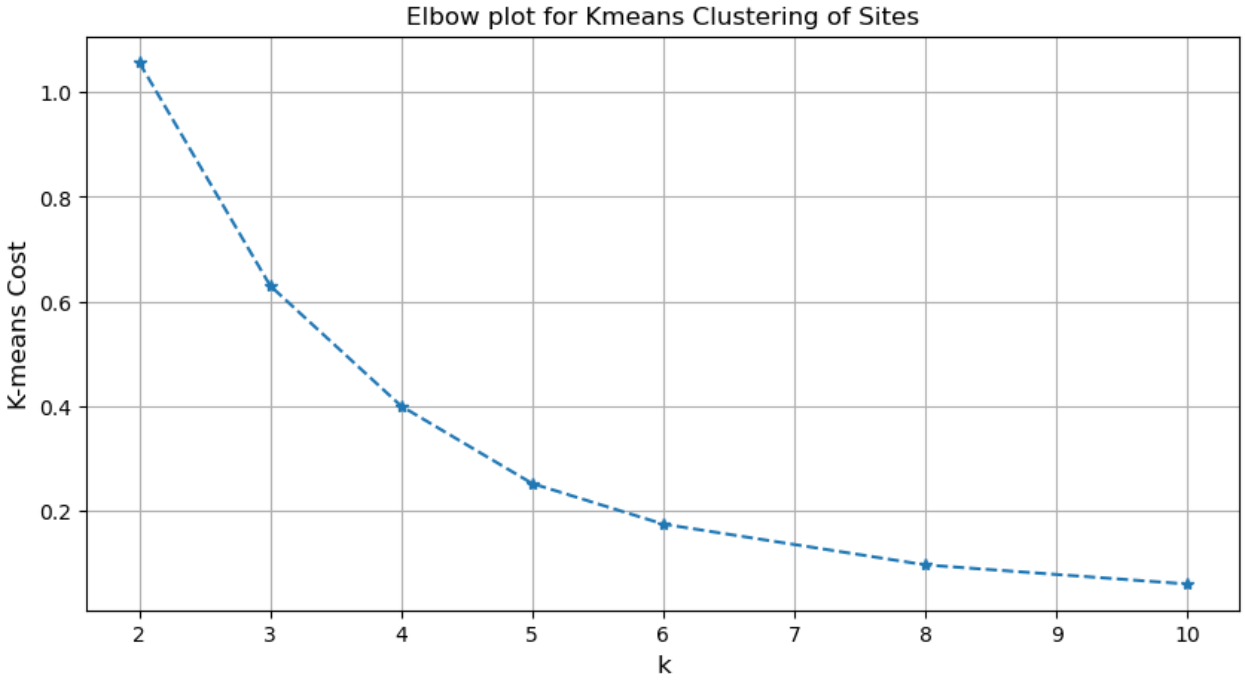
Ultimately, we can conclude that from these visualizations, campus activity seems to be the main driver for the waste generation across B.U's campus, with most sites showing very little outputs aside from the main sites that contain dorm food courts or other popular locations that might remain active during the summer. Events were harder to track, with a small change being noted for events such as Commencement. In the future, we will try to find better ways to further explore the effects of large events such as Commencement on the waste container network across campus.

Clustering Analysis of Different Sites by Waste Tonnage, Temperature (KV)

The main idea behind this analysis was motivated by the heat maps and that different sites appear to have different waste generation patterns. The hypothesis was that we could perhaps group or cluster these sites into a few distinct clusters such that we could then analyze their waste generation patterns in a more comprehensive manner, specifically for that group or cluster.

Elbow plot of KMeans Clustering

Using the elbow plot below it was determined that k=4 is the optimal clustering to cluster sites based on the temperature, waste tonnage and PSI.

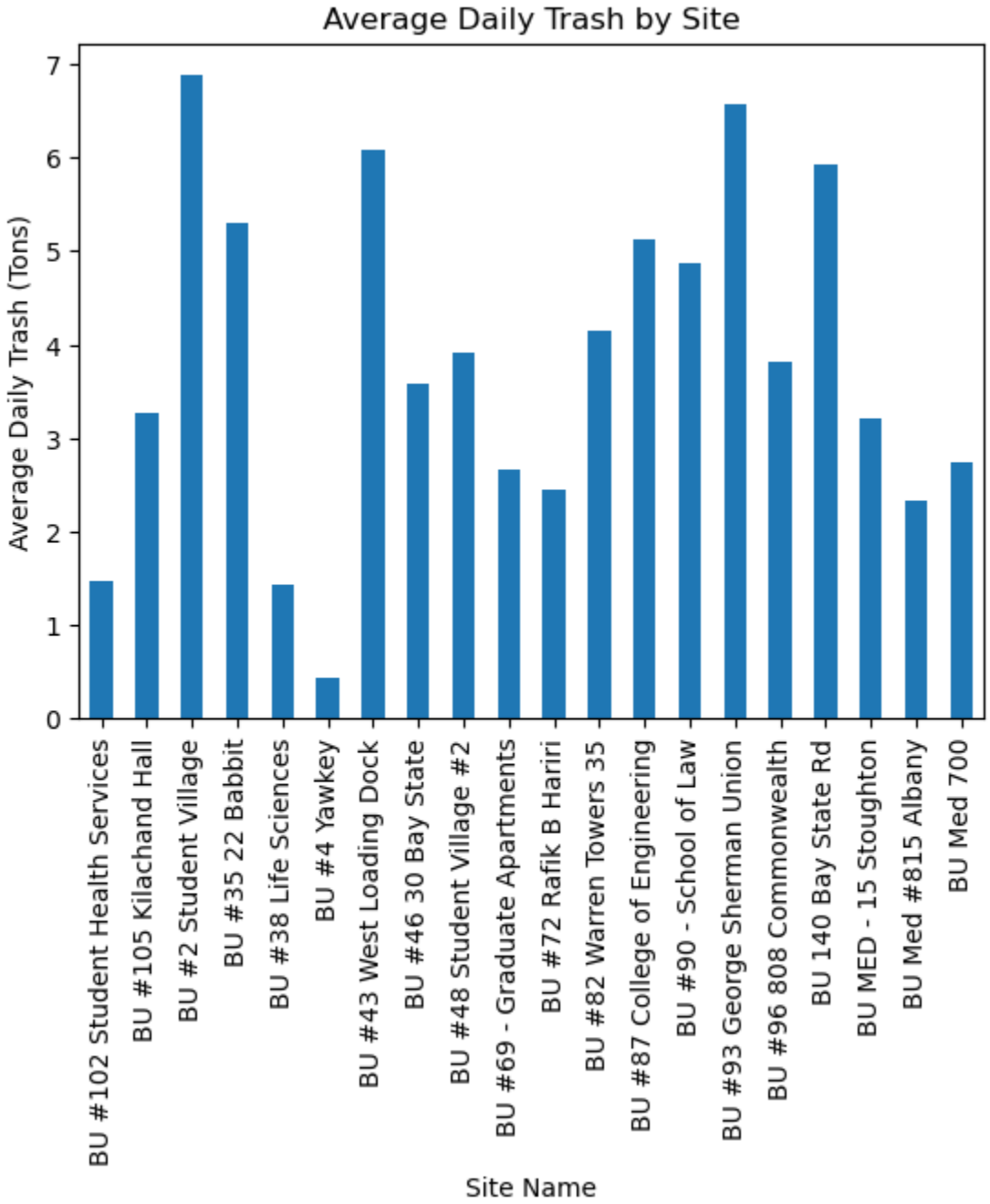


Visualize Clustering on Map

Analysis of Trash by Material Type (KV)

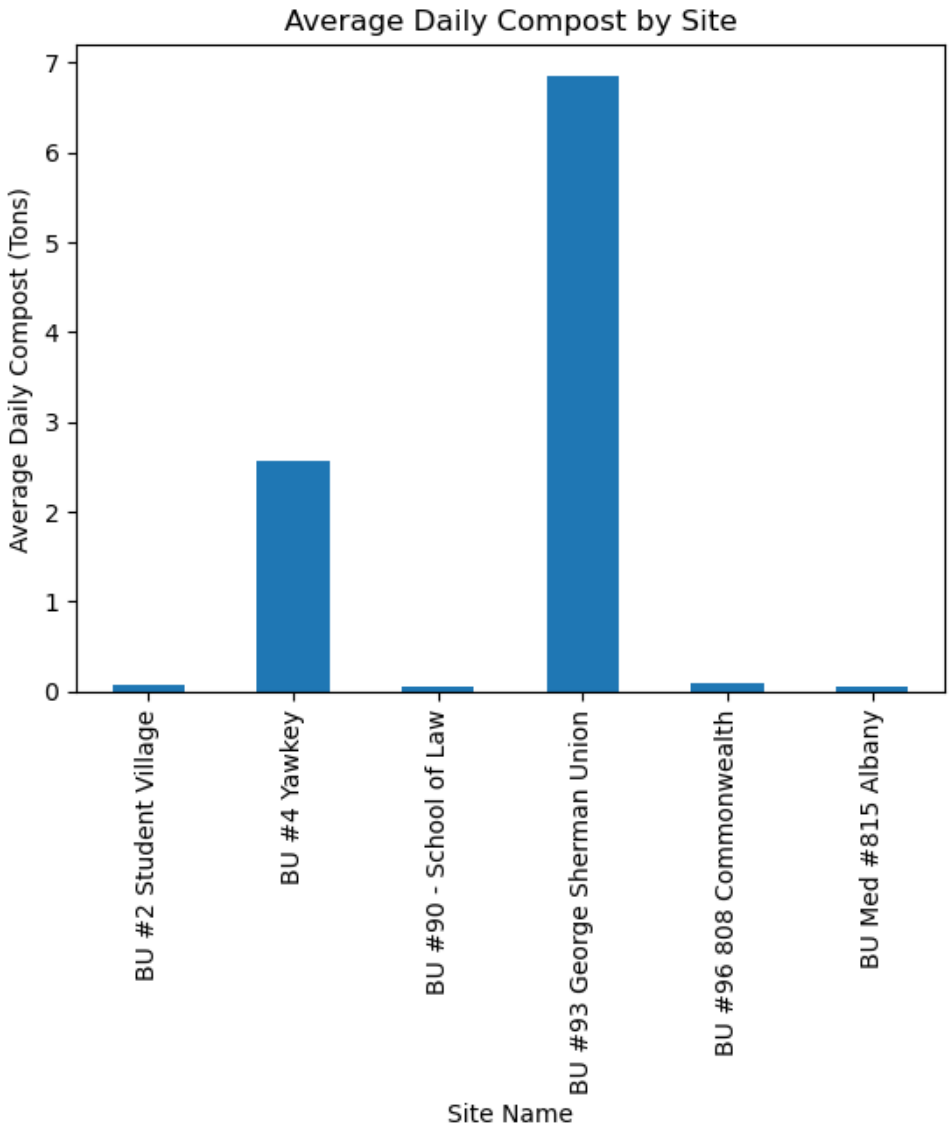
Trash Generation Daily Average by Site

We observe that Student Village, West Loading Dock, GSU and Bay State road generate the highest daily average of trash across all sites. The Life sciences and Yawkey locations have the lowest trash daily average



Compost Generation Daily Average by Site

We observed that GSU and Yawkey generate the overwhelming majority of compost waste across all 5 sites. Their waste accounts for over 99% of all compost generated.



Recycling Daily Average by Site.

We observe that GSU and Yawkey create a disproportionate amount of daily recycling waste tonnage. Additionally, Rafik B Hariri, Student Village and Graduate Apartments also have a high amount of recycling waste, as compared to the average recycling amount across all sites.