BU Sustainability: Understanding How Weather Impacts Waste

Spring 2023 CS506 Data Science

Team-1 Deliverable 1

Required Tasks:

- Collect and pre-process a preliminary batch of data
- Perform a preliminary analysis of the data
- Answer 1-2 key questions
- 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!)
- Submit the Weekly Scrum report to the gradescope and upload to google drive.

Overview:

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 performed their share of the data analysis.
- 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.
- Each section has its own insights of the dataset. Through the course of the work put towards this deliverable we gained a substantially better understanding of the datasets and scope of the project
- We also communicated with the PM and client to clarify certain questions relating to the integrity/correctness of the data and general question regarding the data/columns/project scope.
- The scrum report is also attached with the Gradescope submission

Name: Junyi Zhu

Github branch [team-1-JunyiZ]

Exploring the relationship between value_psi and weather-related features from the perspective of each device

dev	date	psi	celsius	AWND	PRCP	SNOW	TAVG	TMAX	TMIN	WDF2	WSF2	index	absC
39569	2021-08-25	1120	22.70	5.82	0.0	0.0	84	94	76	240	12.1	0	0.70
39569	2021-08-25	1024	22.70	5.82	0.0	0.0	84	94	76	240	12.1	1	0.70
39569	2021-08-25	944	22.70	5.82	0.0	0.0	84	94	76	240	12.1	2	0.70
39569	2021-08-25	872	22.70	5.82	0.0	0.0	84	94	76	240	12.1	3	0.70
39569	2021-08-25	808	25.74	5.82	0.0	0.0	84	94	76	240	12.1	4	3.74

Beside the data provided by our client, which is mainly about temperature, we collected data from [NOAA database] to provide extra information about the weather, including min and max temperature.

Feature	Description
celsius	Temperature by celsius
AWND	Average wind speed
PRCP	Precipitation
SNOW	Snowfall
TAVG	Average Temperature
TMAX	Maximum temperature
TMIN	Minimum temperature
WDF2	Direction of fastest 2-minute wind
WSF2	Fastest 2-minute wind speed
absC	abs(Temperature-22℃)

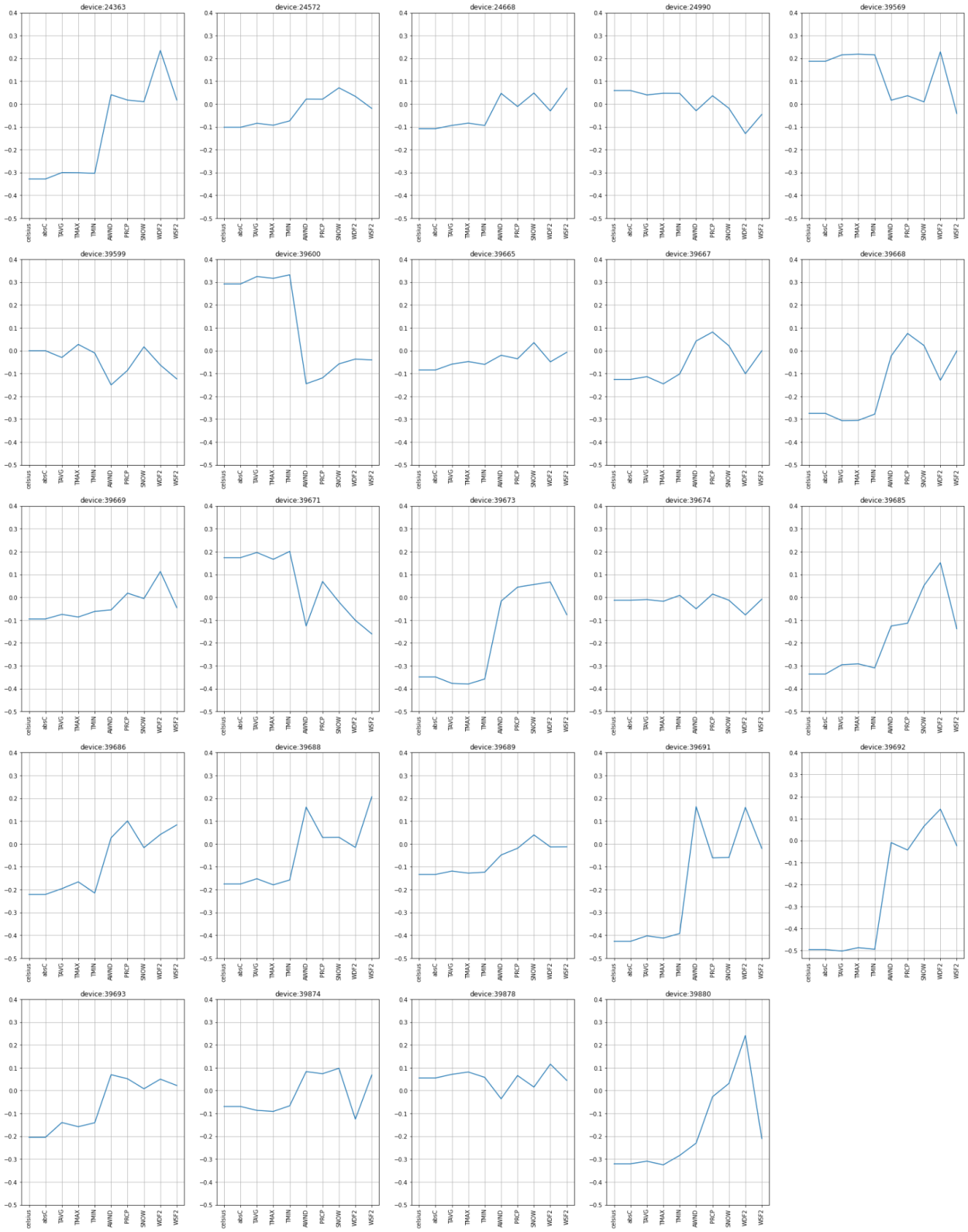
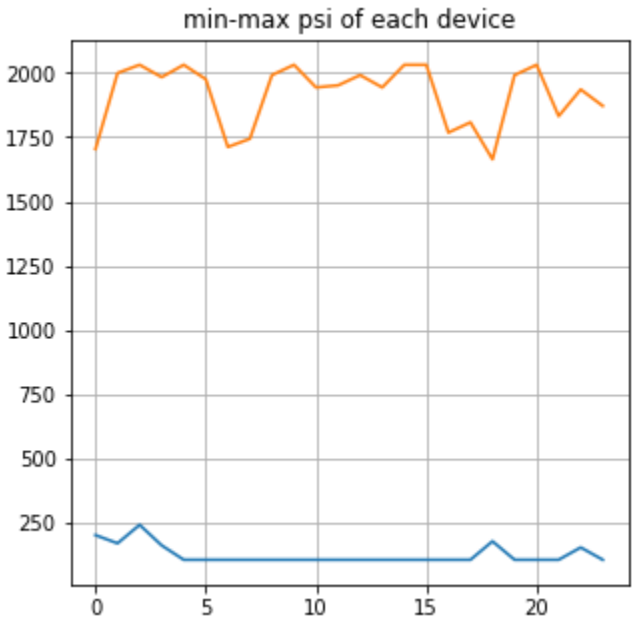


Fig. Pearson correlation coefficient between psi and each feature

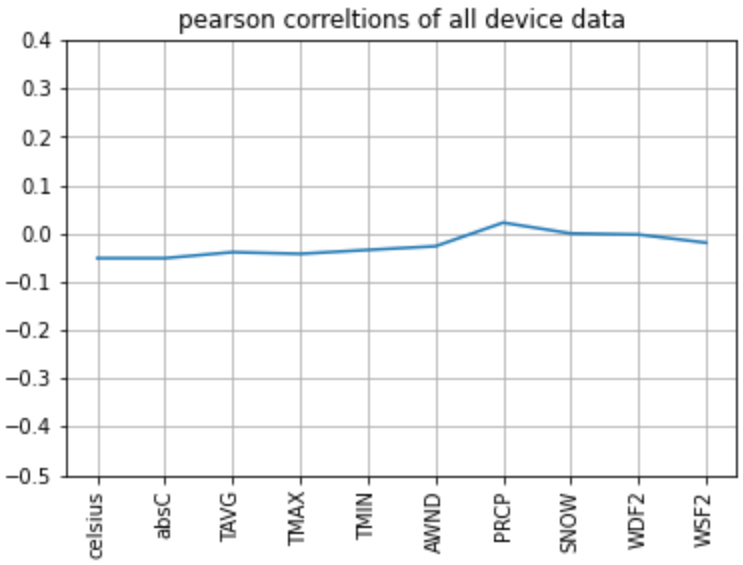
Among the features, temperature-related features, including min, max, mean, abs(temp-22℃), and sample temperature, have the strongest correlation with psi.

Observation A: Some devices do not correlate significantly with these features, even temperature

Observation B: Though temperature shows a strong correlation, being negative or positive depends on the specific device.



When it comes to min/max-psi, it seems there's no significant difference among the devices, so I treat each device the same when I combine the data of each device to analyze the impact of weather on the quantity of waste from the aspect of psi (the figure below).



The impact at the macro level is much smaller than at the micro(device) level, as we analyzed above, due to the different influences at the micro level.
Based on the chart, we may say there is a minor impact that when it’s colder and it’s rainy/snowy, there is more waste generated.

Insights:

a) Further research might emphasize the temperature features
b) It is unlikely to get a linear law that applies to all devices, alternatively, we divide by device/location(indoor or outdoor).
c) To study the impact of weather on garbage, we need to confront the enormous “noisy” impact of people's daily routines, such as work and rest.
d) By intuition and results, the weather has a more pronounced effect on distribution than quantity. For example, if it's raining or cold one day, people's life is likely to go on as usual, and there is not so much waste generated due to the weather. However, in such weather, people are more reluctant to go out, there the distribution is affected.

Name: Zhengyu Liu
Preliminary Analysis on Sites Report Dataset
GitHub branch: team-1-Zhengyu_Liu

We also performed a preliminary analysis on the dataset sites_report_2023-02-28_0820, given by our client.

	Category	Site Name	Timezone	GPS Latitude	GPS Longitude	Address Line 1	Address Line 2	City	Zip/Postal Code	State	Country
0	NaN	BU #102 Student Health Services	America/New_York	42.35	-71.11	881 Commonwealth Avenue	NaN	Boston	2215.0	MA	US
1	NaN	BU #105 Kilachand Hall	America/New_York	42.35	-71.10	91 Bay State Rd	NaN	Boston	2215.0	MA	US
2	NaN	BU #108 Agganis Arena	America/New_York	42.35	-71.12	925 Commonwealth Avenue	NaN	Boston	2215.0	MA	US
3	NaN	BU #18 - Warren Hall	America/New_York	42.35	-71.10	14 Buswell St	NaN	Boston	2215.0	MA	US
4	NaN	BU #2 Student Village	America/New_York	42.35	-71.12	10 Buick St	NaN	Boston	2215.0	MA	US

Fig. An Overview of sites_report Dataset

This dataset contains information about the GPS Latitude and Longitude of each device across the whole BU campus. After performing a series of actions to fetch the necessary information based on the given dataset, we get a neater version:

	GPS Latitude	GPS Longitude
0	42.35	-71.11
1	42.35	-71.10
2	42.35	-71.12
3	42.35	-71.10
4	42.35	-71.12

Fig. An Overview on Location Information of Each Device

We are interested in the correlation between the location of each device, the amount of waste of them, and the weather. To further analyze this, we plotted a heatmap based on the location of each device on a Boston base map with default location of (42.3505° , 71.1054°).

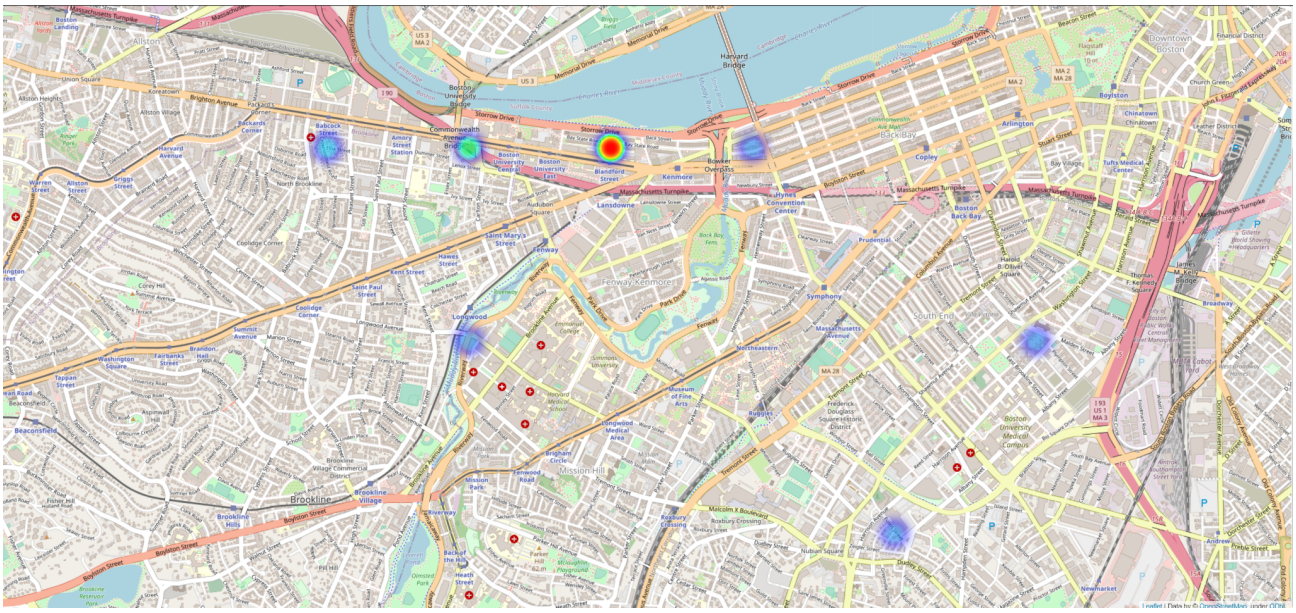


Fig. Device Location Heatmap

The heat map above contains the distribution of the devices based on their location. We obtained the following observations: waste collection devices are more grouped at the main campus, which indicates more wastes are produced at those locations. Nevertheless, the figure doesn't show a direct correlation between the weather and the amount of waste produced. In the next stage, we will make adjustments to the dataset such as merging the weather information to this dataset and then plotting the heatmap again to show a more accurate result.

Possible **Key questions** to be answered:

1. Does weather affect waste?
   Yes.
2. Which features dominate the influence?
   Temperature-related features including real-time temperature and max/min/average temperature of the day.
3. In which ways does weather affect waste?
   Quantity(based on tons produced) and distribution(by location of devices).
4. How does weather affect those aspects(e.g., distribution)?
   Quantity:
   TBA

   Distribution:

It depends on the specific device. Some can be irrelated to temperature, some can be positively related, and some can be negatively related to.

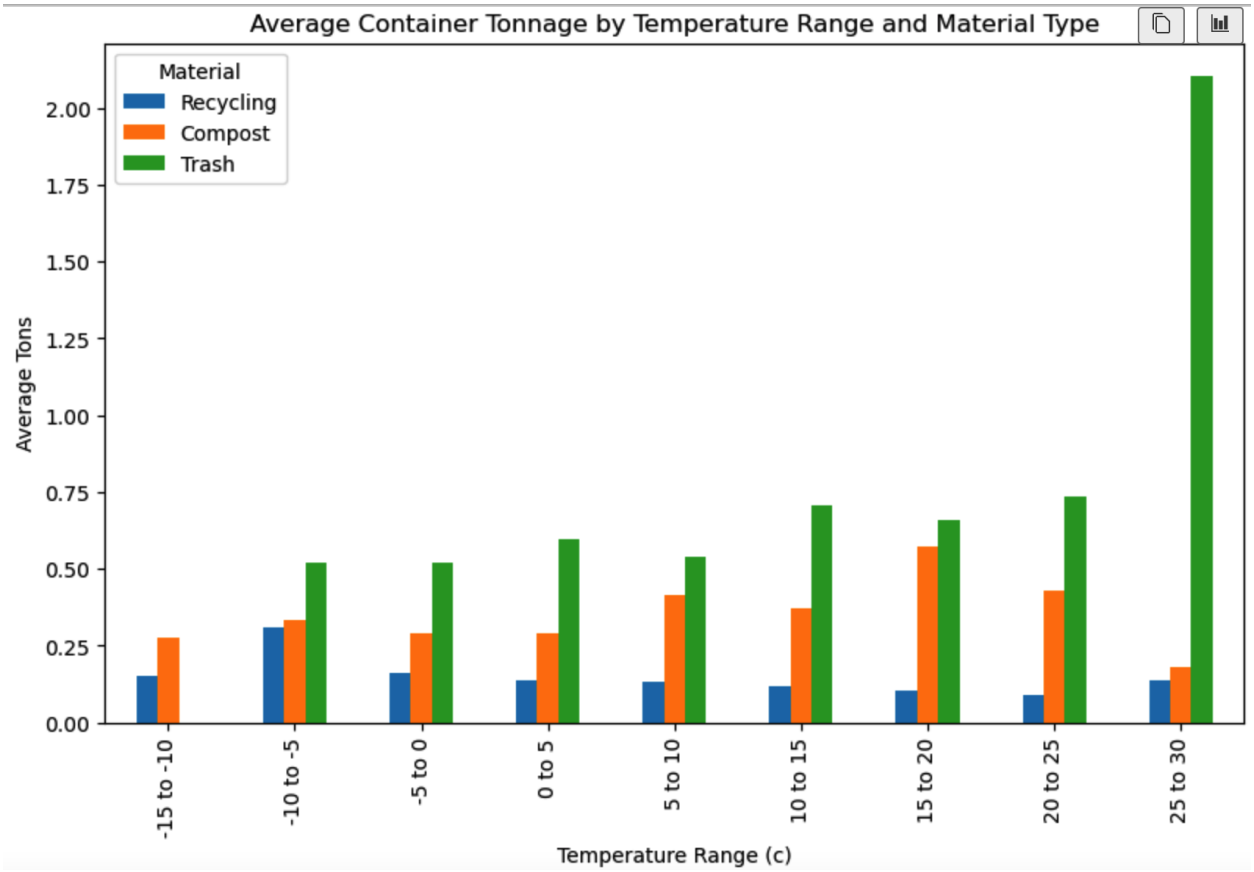
If considering types of waste, the above research can be further developed into sub-tasks like quantity-recycle.

Name: Abdelazim Lokma
Preliminary Analysis of Container Tonnage Upon Disposal:
Github branch: team-1-abdelazim link

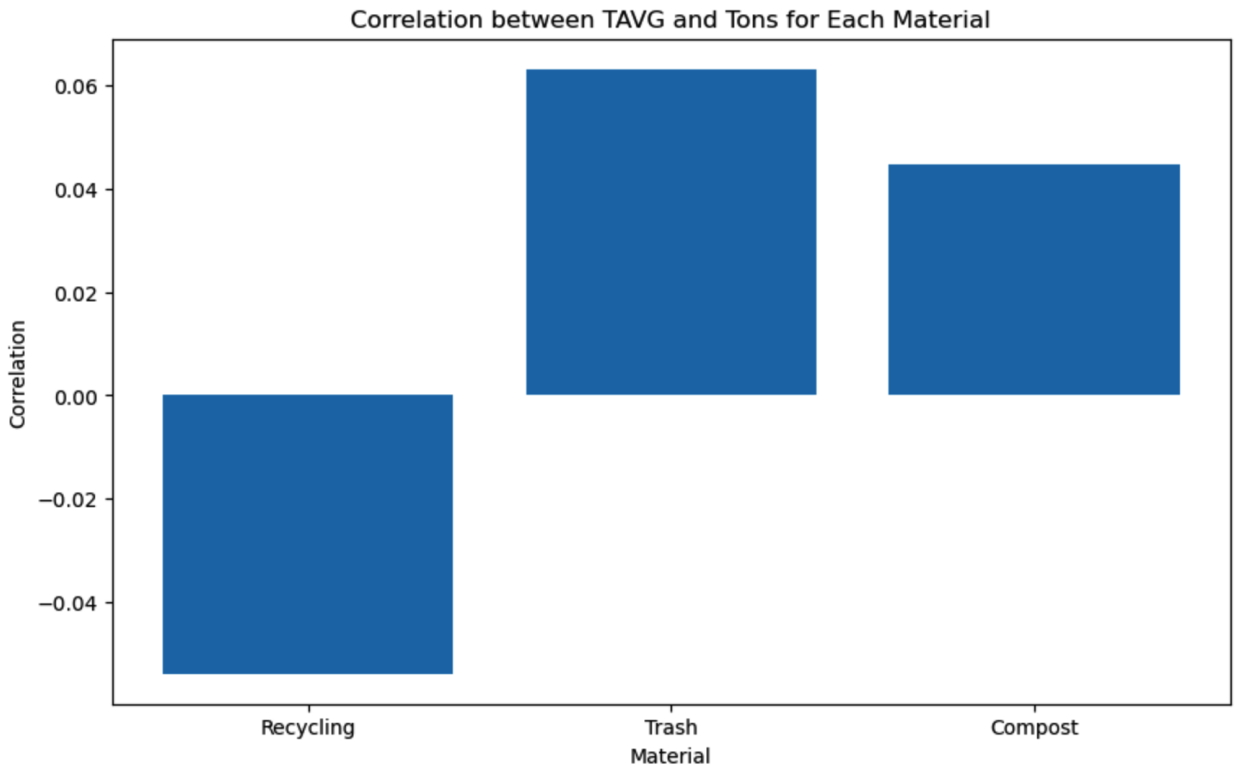
This section aims to analyze the “BU Daily Weights FY22.xlsx” document, which contains information on the tonnage of all compacting containers on campus just before they were emptied by the waste vendor Cacella. We first had to merge temperature data from the NOAA database(www.ncei.noaa.gov) with the data provided in the excel sheet in order to determine the Average Daily temperature on the days in which the containers were emptied.

	Date	AWND	TAVG	TMAX	TMIN	WDF2	WSF2	Customer Key	Location	Address	Material	Tons
0	7/1/2021	4.0	25.4	29.4	20.0	220.0	7.6	31903.0	BU #72 - Rafik B Hariri	853 BEACON ST	Recycling	0.0300
1	7/2/2021	6.3	18.3	20.6	15.0	50.0	9.8	31950.0	BU #89 - Brownstones	91 BAY STATE RD	Recycling	0.0800
2	7/6/2021	6.6	24.9	33.3	20.6	270.0	12.5	31769.0	BU #112B - 949 Comm Ave	36 Cummington Mall	Recycling	0.5670
3	7/6/2021	6.6	24.9	33.3	20.6	270.0	12.5	31946.0	BU #99 - College of Fine Arts	120 ASHFORD STREET	Trash	0.0470
4	7/6/2021	6.6	24.9	33.3	20.6	270.0	12.5	32111.0	BU MED- 815 Albany	815 Albany Street	Compost	0.0725
...	...	...	...	...	...	...	...	...	...	...	...	...
17798	6/30/2022	4.6	22.9	28.3	19.4	320.0	8.9	32111.0	BU #29 - Harriet Richards House	640 Commonwealth Ave	Recycling	0.0000
17799	6/30/2022	4.6	22.9	28.3	19.4	320.0	8.9	32450.0	BU #37 - African Studies	940 Commonwealth Ave	Recycling	0.0250
17800	6/30/2022	4.6	22.9	28.3	19.4	320.0	8.9	32493.0	BU #43 - West Loading Dock	275 Babcock	Recycling	0.0580
17801	6/30/2022	4.6	22.9	28.3	19.4	320.0	8.9	32494.0	BU #78 - 660 Beacon	100 BAY STATE RD	Recycling	0.0555
17802	6/30/2022	4.6	22.9	28.3	19.4	320.0	8.9	33165.0	BU MED- 635 Albany Street	232 baystate Rd	Recycling	0.0195

After preparing the data and plotting onto a scatter plot, where each datapoint was clustered by Material type, we noticed a possible trend between the tonnage and the temperature. To further explore this, we graphed the average weight in tons of each type of container relative to the changing daily average temperature.



The grouped bar chart above shows that, on average, the tonnage of Trash containers seems to increase with temperature. For the recycling containers, it seems to decrease in most cases. However, for Compost, it is difficult to tell. We may need to find the Pearson correlation coefficient for each container type in order to make a final observation.

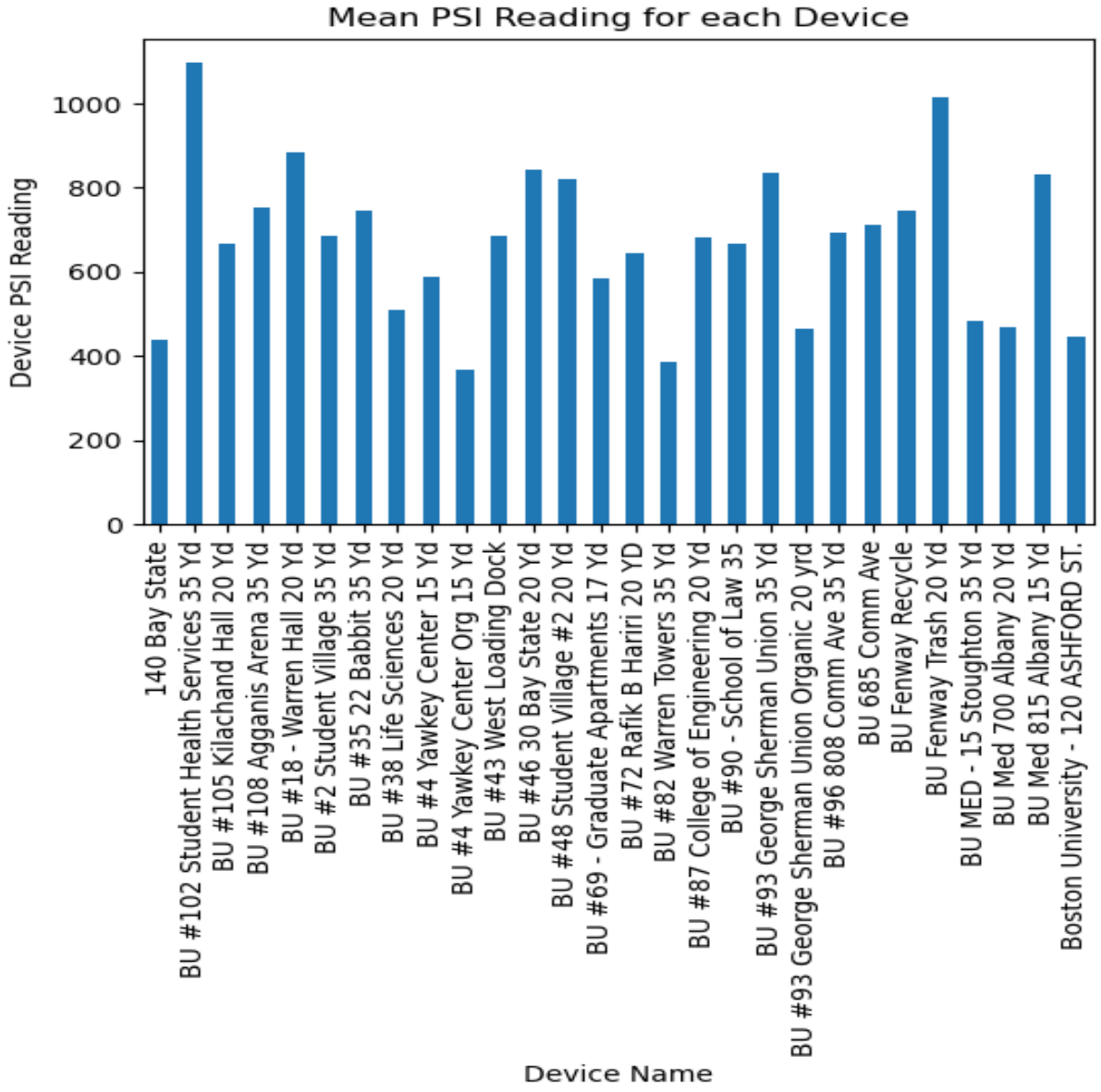


Interestingly, the bar chart above reveals more detail about recycling containers, it appears recycling tonnage goes down as temperatures increase. Whereas there seems to be a minor positive correlation between Trash/Compost tonnage with Average daily temperature.

Name: **Karan Vombatkere**
GitHub branch: **team-1-kv**
https://github.com/BU-Spark/ds-bu-sustainability-waste/tree/team-1-kv/spring23-team-1
*Code/Analysis is in **DataAnalysis_KV.ipynb** notebook.*

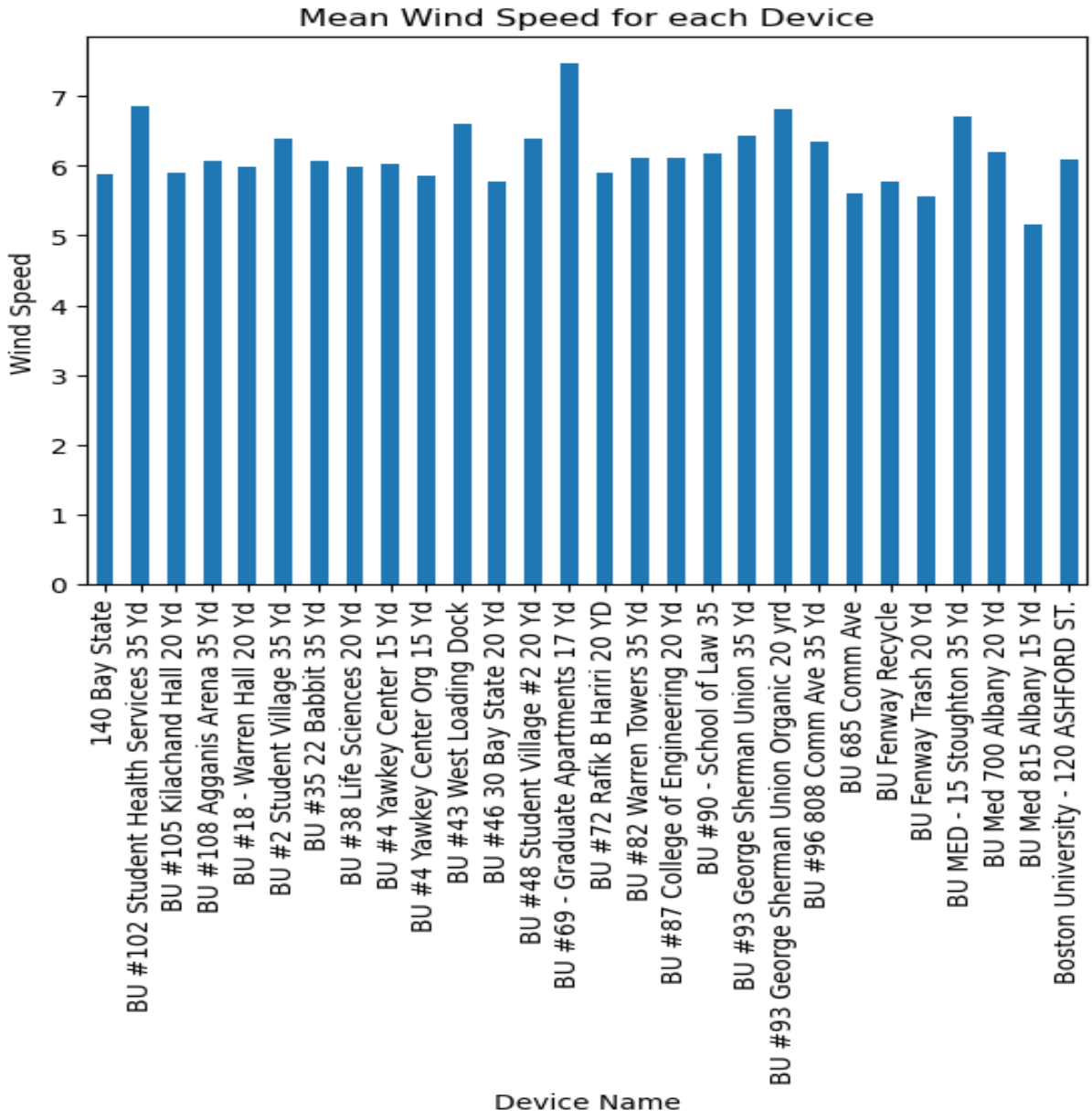
Preliminary Analysis

PSI Readings - *PSI_Readings_with_Weather_2023-01-04_1026*



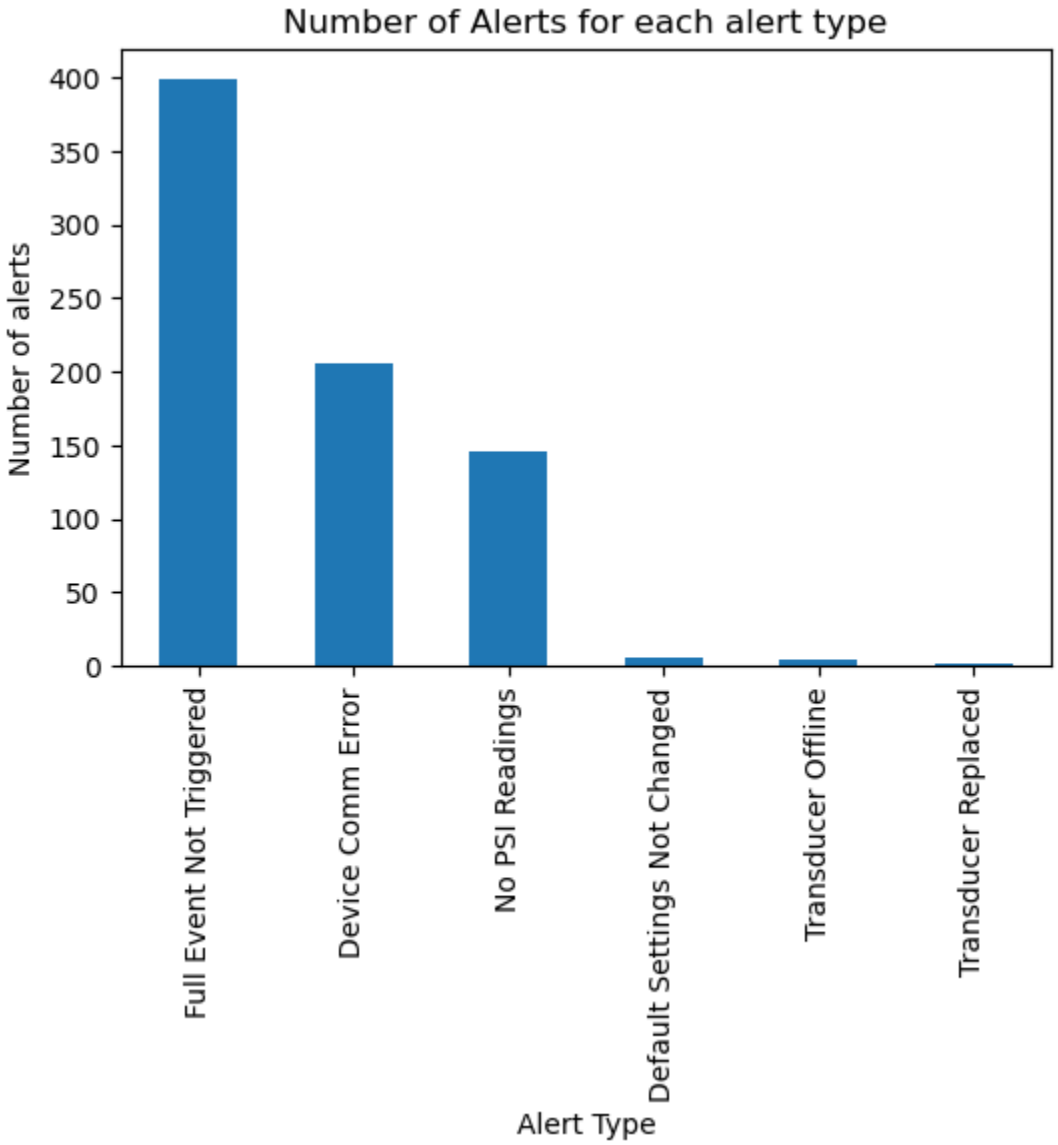
Insights: We observe that the mean PSI readings greatly varies by device. This could mean that some devices are likely being allowed to fill up more before compaction than

others. It could also be related to the location/weather of the device, since the PSI reading gives an indication of how full the compactor is.



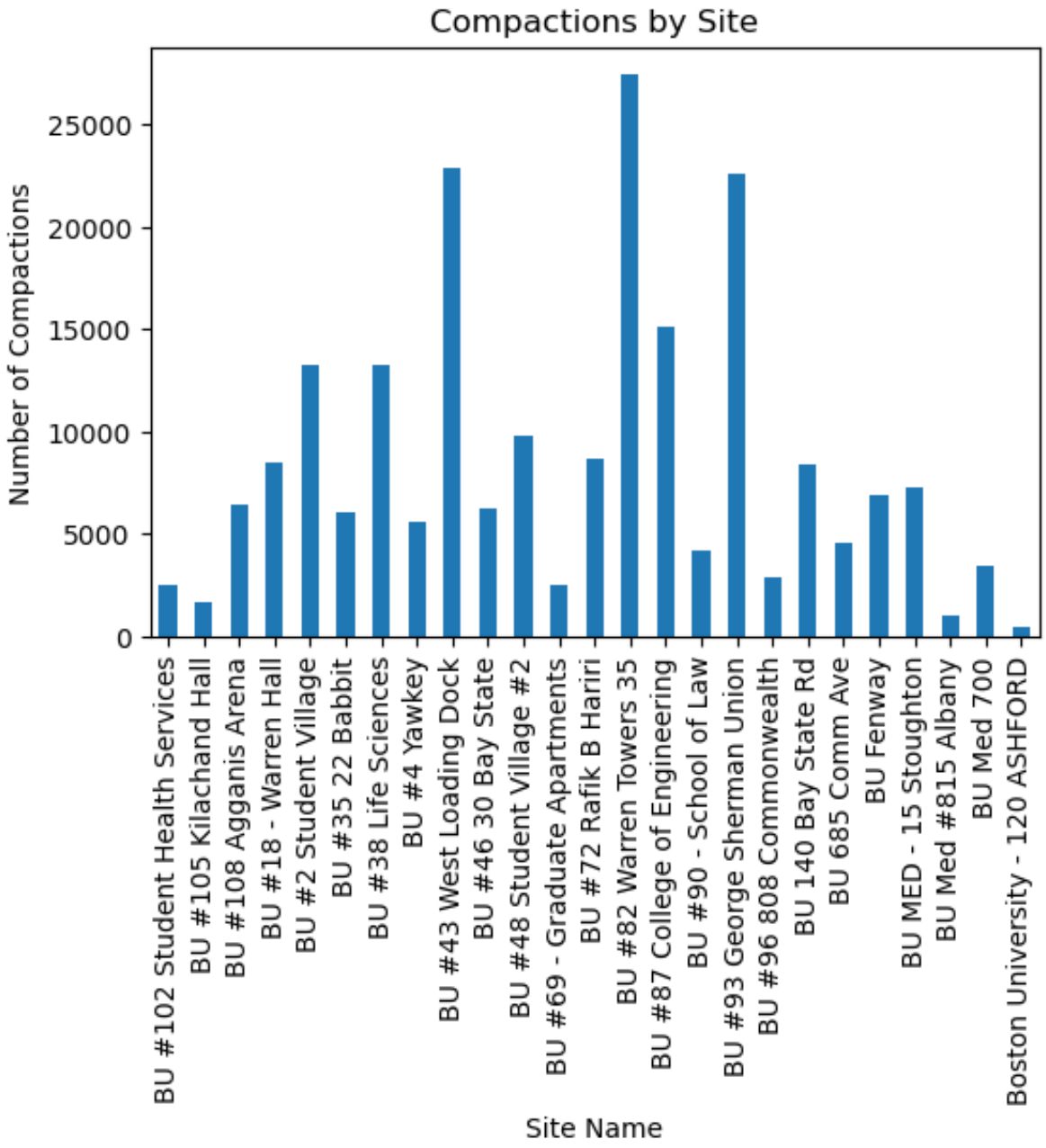
Insights: We observe that the mean wind speed doesn't significantly vary by device. However there are a couple of locations with higher wind speeds. This raises the question of investigating those locations specifically to see if they correlate with greater/lesser amount of waste generation.

Alert Types - *Alert_Flag_History_2023-01-04_1029*



Insights: We observe that the “Full Event not Triggered” is by far the most common alert flag, followed by, “Device Comm Error” and “No PSI Readings”. Since the Alert flag dataset has a wide date range, we can conclude that the Alerts likely don’t have a significant impact on the quality of the rest of the data.
Ideally we could map the alert date and type to the device readings using the date and device ID and then filter out/delete those “bad” readings, but since there are very few alerts, this is not an essential step at this stage.

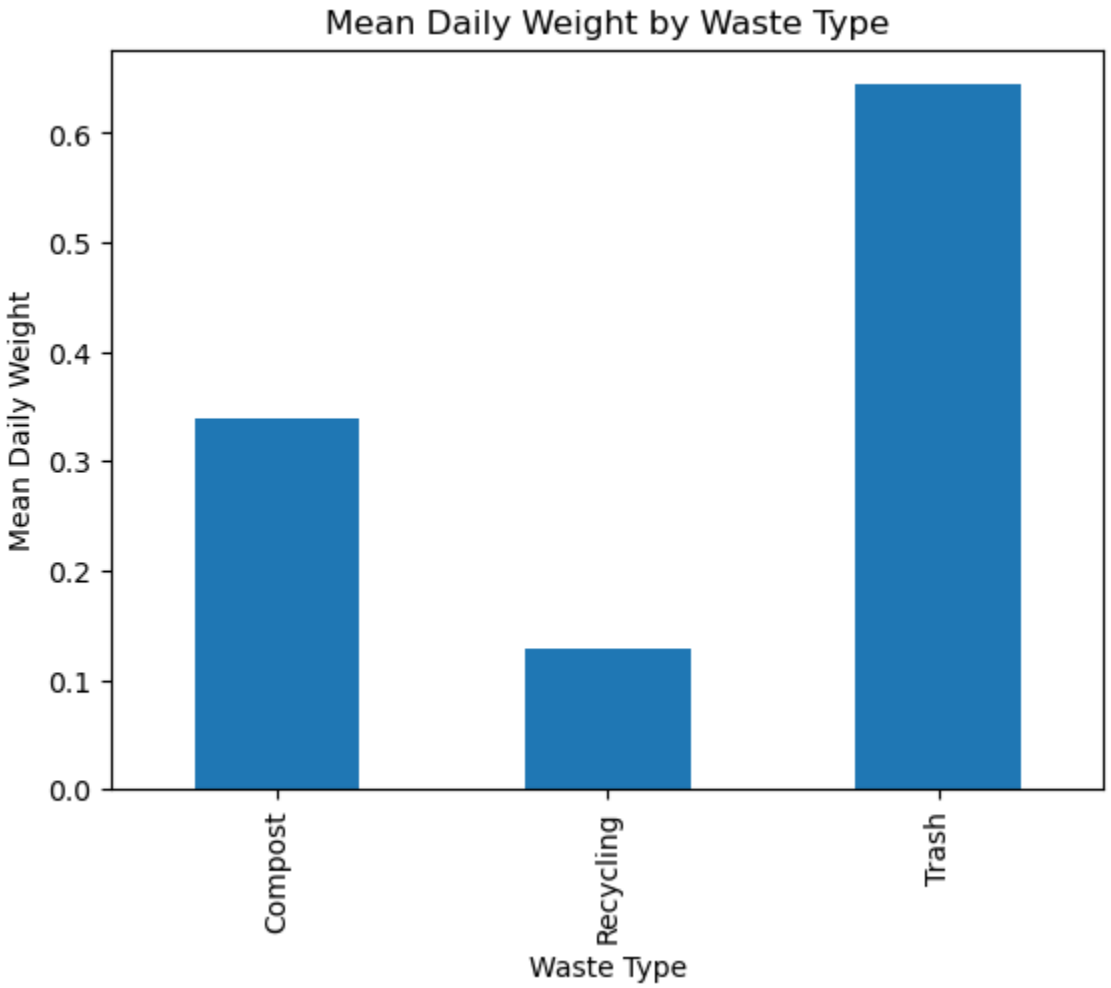
Device Compactions - *Device_Compaction_Frequency_2023-01-04_1029*



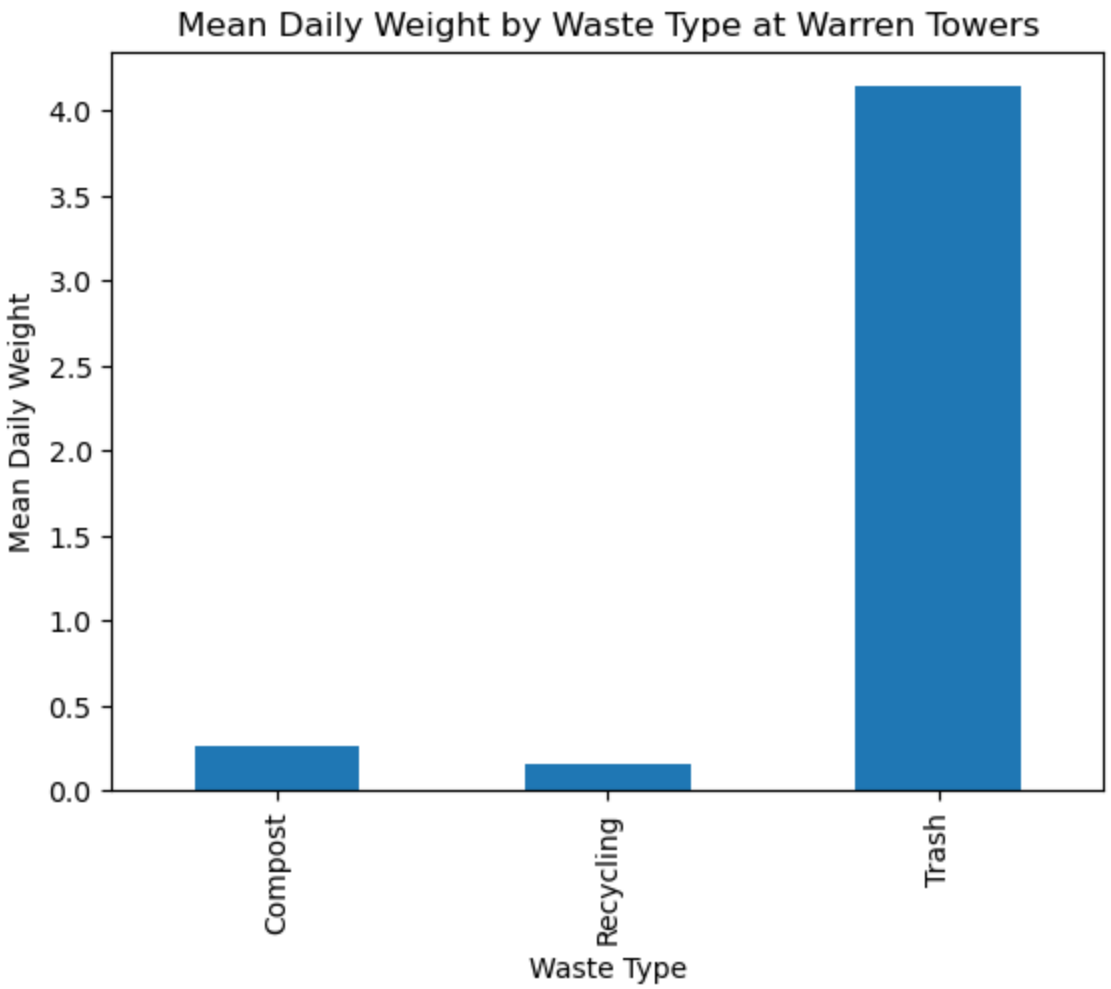
Insights: We observe that the Device compaction frequency varies greatly by location. Of particular interest are the Warren Towers and GSU sites, which have the highest number of compactions. This should validate that these 2 locations have the greatest amount of waste generation - to be investigated further.

This insight also hints that it would probably make sense to first cluster the Devices by location/amount of waste generated and then apply ML models specifically tailored to each cluster.

Daily Weight waste Types



Insights: It is easy to see that trash is the most common type of waste, followed by compost and recycling.



Insights: Warren Towers produces a lot of daily trash, on average more than 6x the daily mean across all sites. This indicates that there are outliers in the data set that need to be dealt with separately.

TO-DOs Going forward

- Need to integrate datasets into one comprehensive dataframe. Use Device ID as join key
- Cluster waste devices by type/location/amounts
- Analyze each cluster separately to answer the key questions regarding weather impact.