BU Sustainability: Understanding How Weather Impacts Waste

Spring 2023 CS506 Data Science

Team Members

Name	Year	Email
Zeqi Wang	Senior (BA/MS)	zw100107@bu.edu
Timur Zhunussov		
Akshad Ramnath	Senior	shadr@bu.edu
Baicheng Fang		

Introduction

BU Sustainability supports the transformation of Boston University's planning, operations, and culture toward a sustainable and equitable future.They have pulled multiple spreadsheets with data from their 3rd party vendor Contelligent and our waste vendor Casella. Contelligent provides PSI fullness monitoring of our compactors. The client has provided 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.

The primary motivation for the project is to find a correlation between weather and waste generation at BU so that BU sustainability can potentially improve waste storage location in the case of adverse weather conditions and help take care of our campus better.

Topic: Relationship between Compressor PSI and Temperature - Zeqi Wang

Intro

The research I focused on was finding out if there is a relationship between temperature and the compressor's psi value. The dataset I was using was the 'readings_device.#####.csv'. The time frame of the data record was from 2021-07-01 to 2022-10-02, and 24 compressor machines were located at the BU campus (Charles, Fenway, and MED)

Assumption

The assumption we made here is the higher the psi value, the fuller the trash machine will be, which means there will be more waste generated in that certain machine.

Data Processing & Strategies

The initial data we obtained was very detailed in that it was recorded every few tens of seconds. To make the analysis more convenient, I calculated a summary of the data for each machine for each day, including the mean, median, minimum, and maximum values of PSI and temperature.

For initial visualization, I plotted a line chart of temperature (Fahrenheit) with the Psi value for each machine, and here is an example of Warren Towers:
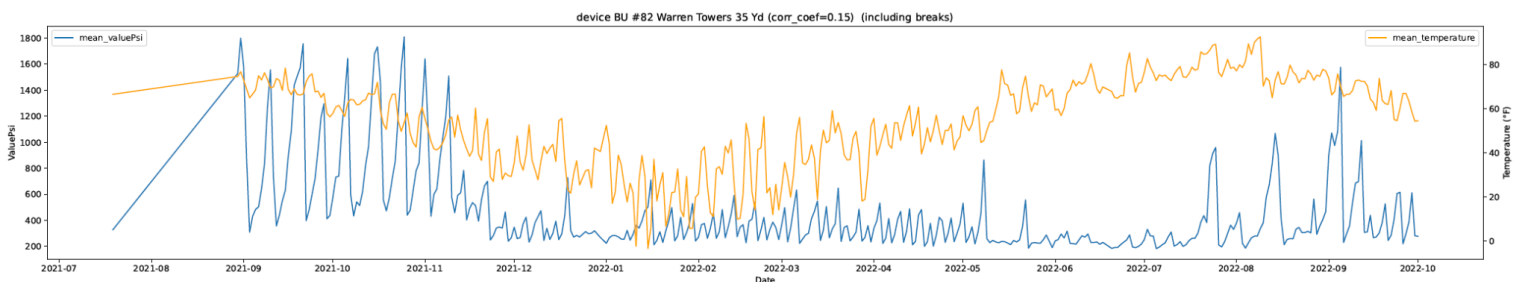
(For complete graphs, check [here])

This initial visualization didn't provide too much information, to find out the relationship, my first strategy is to calculate the correlation coefficient*.

* Correlation coefficient is a statistical measurement that describes the degree of relationship between two variables. (For example the psi value and temperature) and the coefficients are ranged from +1 to -1. If the correlation coefficient is positive, it means that there is a direct

relationship between the two variables - as one increases, so does the other. If it is negative, it means that there is an inverse relationship - as one increases, the other decreases.

```
+----------------------------------------------------------------------------------+
|         Correlation Coefficients Between Mean ValuePsi and Fahrenheit (All Dates) |
+------------------------------------+---------------+-------+---------+-------+-------+
|              Device                | Building Type | MeanF | MedianF |  MaxF | MinF  |
+------------------------------------+---------------+-------+---------+-------+-------+
|        BU #35 22 Babbit 35 Yd      |    Academic   | -0.02 |  -0.02  | -0.00 | -0.01 |
|      BU #38 Life Sciences 20 Yd    |    Academic   |  0.34 |   0.32  |  0.34 |  0.34 |
|      BU #72 Rafik B Hariri 20 YD   |    Academic   | -0.22 |  -0.22  | -0.19 | -0.22 |
|  BU #87 College of Engineering 20 Yd |  Academic   | -0.19 |  -0.20  | -0.17 | -0.19 |
|      BU #90 - School of Law 35     |    Academic   | -0.23 |  -0.23  | -0.20 | -0.24 |
|      BU #96 808 Comm Ave 35 Yd     |    Academic   | -0.05 |  -0.07  | -0.05 | -0.04 |
|          BU 685 Comm Ave           |    Academic   | -0.47 |  -0.47  | -0.44 | -0.47 |
|       BU Fenway Trash 20 Yd        |    Academic   | -0.39 |  -0.39  | -0.38 | -0.38 |
|     BU MED - 15 Stoughton 35 Yd    |    Academic   | -0.66 |  -0.66  | -0.61 | -0.68 |
|     BU Med 700 Albany 20 Yd        |    Academic   | -0.17 |  -0.18  | -0.18 | -0.17 |
| BU #102 Student Health Services 35 Yd | Medical    | -0.20 |  -0.20  | -0.19 | -0.21 |
|      BU #43 West Loading Dock      |     Other     | -0.16 |  -0.19  | -0.12 | -0.17 |
|          140 Bay State            |    Resident   | -0.13 |  -0.13  | -0.12 | -0.11 |
|    BU #105 Kilachand Hall 20 Yd    |    Resident   | -0.24 |  -0.24  | -0.25 | -0.23 |
|     BU #18 - Warren Hall 20 Yd     |    Resident   | -0.13 |  -0.13  | -0.14 | -0.13 |
|    BU #2 Student Village 35 Yd     |    Resident   | -0.12 |  -0.13  | -0.11 | -0.11 |
|    BU #46 30 Bay State 20 Yd       |    Resident   | -0.15 |  -0.16  | -0.13 | -0.14 |
|   BU #48 Student Village #2 20 Yd  |    Resident   | -0.08 |  -0.09  | -0.06 | -0.10 |
| BU #69 - Graduate Apartments 17 Yd |    Resident   | -0.33 |  -0.34  | -0.32 | -0.33 |
|    BU #82 Warren Towers 35 Yd      |    Resident   |  0.15 |   0.14  |  0.13 |  0.15 |
|     BU Med 815 Albany 15 Yd        |    Resident   | -0.20 |  -0.21  | -0.16 | -0.22 |
|     BU #4 Yawkey Center 15 Yd      |    Service    | -0.36 |  -0.35  | -0.35 | -0.35 |
| BU #93 George Sherman Union 35 Yd  |    Service    | -0.20 |  -0.19  | -0.16 | -0.20 |
|     BU #108 Agganis Arena 35 Yd    |     Sport     |  0.03 |   0.03  |  0.05 |  0.02 |
+------------------------------------+---------------+-------+---------+-------+-------+
```

The table above shows the correlation coefficient of Mean ValuePsi and Fahrenheit for each location's machine. The red color represents the positive relationship with coefficient above 0.1 and the blue color represents the negative relationship with coefficient below -0.1. There is no significant difference in different types of Fahrenheit values, and for further analysis, I will be using Mean Fahrenheit only.

For visualization purposes, I plotted them on a map with colors based on their coefficient:
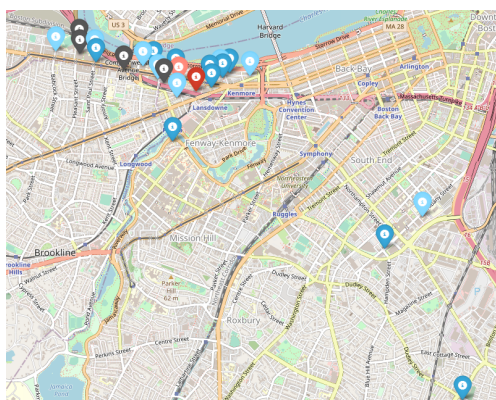
(link to the complete map all dates)
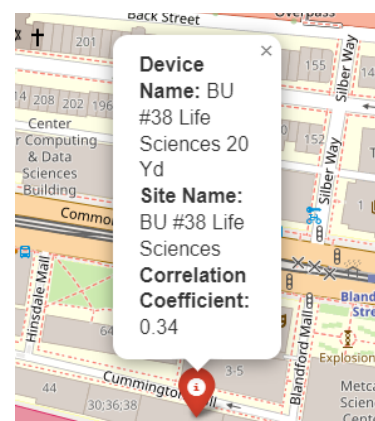
(sample information of Life Science Building)

The map above visually represents the relationship between the psi value and temperature for each location. For a more detailed description, I separated the strength of correlation by the darkness of color:

 Strong Positive   Weak Positive   Strong Negative   Weak Negative   Almost No Relation

The Strong relationships are for those locations with an absolute value of the Correlation Coefficient larger and equal to 0.2. The Weak relationships are for those locations with an absolute value of the Correlation Coefficient smaller than 0.2 but greater than 0.1. The gray color pin indicates we cannot find out an obvious relationship between temperature and psi value.

By looking at the data, we can see that most of the Academic and Residential buildings show some level of relationship between temperature and waste generation. Apart from Life Science and Warren Towers, which show a positive correlation, other residents and academic buildings are showing negative correlations.


Further Analysis - Linear Regression

Linear regression provides more information than just the correlation coefficient. While the correlation coefficient measures the strength and direction of the linear relationship between two variables, linear regression can help predict the values of one variable based on the values of another variable.

Additionally, linear regression provides other important metrics such as R-squared (coefficient of determination), which measures the proportion of the variance in the dependent variable that is explained by the independent variables. This gives a more complete understanding of the relationship between the variables and the predictive power of the model.

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. In other words, it's a measure of how well the regression line fits the data. R-squared values range from 0 to 1, with higher values indicating that more of the variance in the dependent variable is explained by the independent variable(s) in the model. A value of 1 indicates a perfect fit, where all of the variance in the dependent variable is explained by the independent variable(s), while a value of 0 indicates no relationship between the variables.

After performing Linear Regression on each location's data, we found some locations that indicate a strong relationship between temperature and Psi, as shown in the graph below:
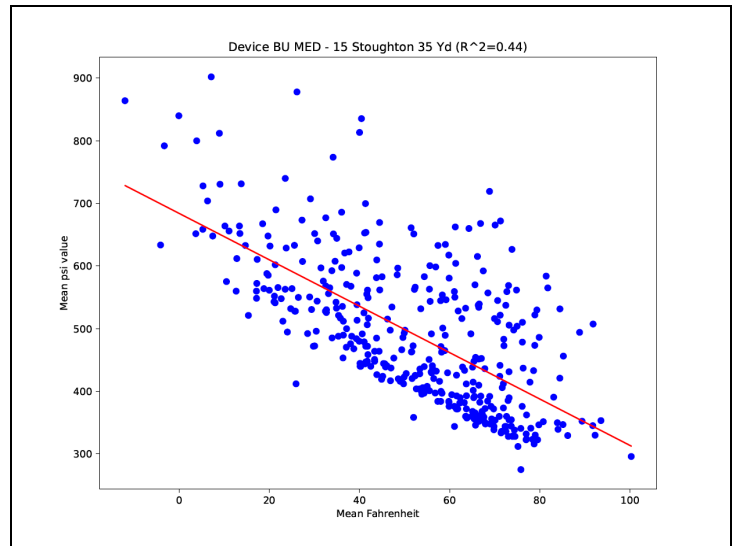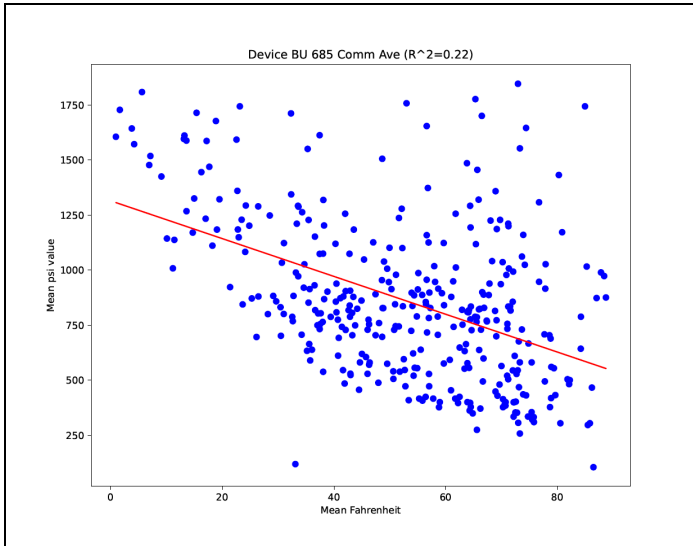
After using the linear regression, we are able to find out the relationship with higher reliability since there are more locations identified as 'no obvious relationship'. The above two locations are the ones that show a strong correlation.

Extension idea

During the research, I realized that temperature data alone may not be enough to accurately predict waste generation as there may be other factors at play, such as occupancy rates, types of buildings, etc. Our team tried to find out the population of each location but we didn't have enough sources to look for that information. Therefore, one of the possible strategies is to analyze the data based on BU's opening and closing dates, since there is a big population difference between school days and breaks.
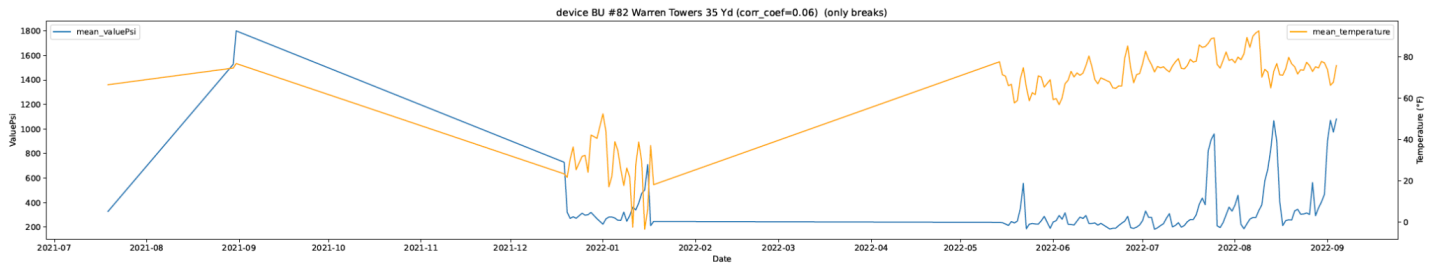
Here I separated the original data with all dates into two sets: **Without Breaks** and **Only Breaks**. Based on the data we have, and by referencing the BU Calendar, I have separated out the following dates as breaks:

- 2021 Summer Break: May 23 - September 1
- 2021 Winter Break: December 19 - January 19
- 2022 Summer Break: May 14 - September 5

The following graphs are the line chart of Warren Towers' data without breaks and only breaks.
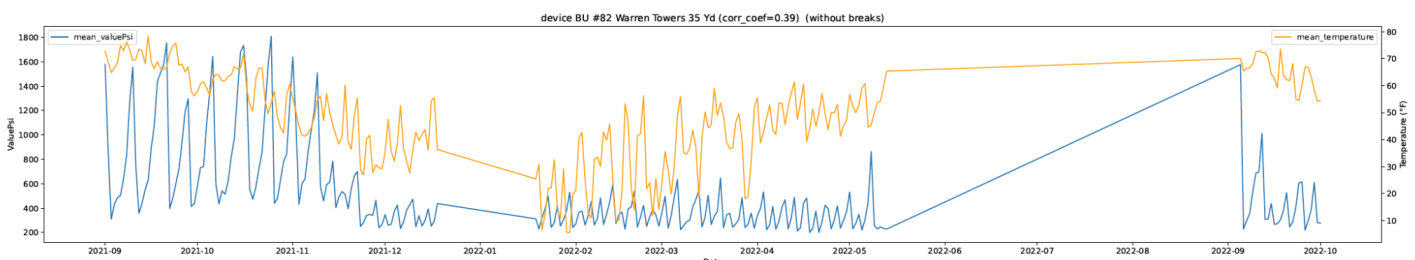
(sample line graph of data with only break, for complete graph, check [here](#))

Note: The straight stroke between empty values is not being calculated during the following analysis.

I also calculated the correlation coefficient for these two data sets.

Correlation Coefficients Between Mean ValuePsi and Fahrenheit (Without Breaks)					
Device	Building Type	MeanF	MedianF	MaxF	MinF
BU #35 22 Babbit 35 Yd	Academic	0.13	0.13	0.14	0.12
BU #38 Life Sciences 20 Yd	Academic	0.29	0.24	0.29	0.30
BU #72 Rafik B Hariri 20 YD	Academic	0.05	0.04	0.08	0.06
BU #87 College of Engineering 20 Yd	Academic	−0.11	−0.12	−0.09	−0.08
BU #90 − School of Law 35	Academic	−0.12	−0.13	−0.10	−0.12
BU #96 808 Comm Ave 35 Yd	Academic	−0.12	−0.14	−0.10	−0.10
BU 685 Comm Ave	Academic	−0.26	−0.26	−0.25	−0.26
BU Fenway Trash 20 Yd	Academic	−0.19	−0.18	−0.20	−0.16
BU MED − 15 Stoughton 35 Yd	Academic	−0.62	−0.63	−0.57	−0.64
BU Med 700 Albany 20 Yd	Academic	0.01	−0.00	−0.02	0.02
BU #102 Student Health Services 35 Yd	Medical	−0.41	−0.40	−0.41	−0.41
BU #43 West Loading Dock	Other	−0.03	−0.05	−0.03	−0.02
140 Bay State	Resident	−0.00	−0.00	−0.01	0.03
BU #105 Kilachand Hall 20 Yd	Resident	−0.18	−0.17	−0.19	−0.17
BU #18 − Warren Hall 20 Yd	Resident	−0.07	−0.07	−0.09	−0.05
BU #2 Student Village 35 Yd	Resident	−0.08	−0.08	−0.11	−0.05
BU #46 30 Bay State 20 Yd	Resident	−0.08	−0.10	−0.06	−0.07
BU #48 Student Village #2 20 Yd	Resident	−0.15	−0.16	−0.15	−0.16
BU #69 − Graduate Apartments 17 Yd	Resident	−0.26	−0.27	−0.25	−0.26
BU #82 Warren Towers 35 Yd	Resident	0.39	0.38	0.37	0.41
BU Med 815 Albany 15 Yd	Resident	−0.26	−0.27	−0.18	−0.29
BU #4 Yawkey Center 15 Yd	Service	−0.29	−0.27	−0.30	−0.26
BU #93 George Sherman Union 35 Yd	Service	−0.16	−0.16	−0.13	−0.15
BU #108 Agganis Arena 35 Yd	Sport	0.02	0.02	0.03	0.03

As we can see from the Table, after deleting the break dates from the original data, we are able to see there is one more Academic building (BU #35) showing a positive correlation while previously we only has BU #38 Life Science Building. And for Warren Towers, the original all-date data set only indicates a correlation of 0.15, and now we have 0.39. This means that the school break is also an important factor in affecting waste generation.

```
+-------------------------------------------------------------------------------------------+
|              Correlation Coefficients Between Mean ValuePsi and Fahrenheit (Only Breaks)   |
+----------------------------------------+---------------+-------+---------+-------+-------+
|                 Device                 | Building Type | MeanF | MedianF |  MaxF | MinF  |
+----------------------------------------+---------------+-------+---------+-------+-------+
|           BU #35 22 Babbit 35 Yd       |    Academic   | -0.22 |  -0.23  | -0.19 | -0.23 |
|         BU #38 Life Sciences 20 Yd     |    Academic   |  0.28 |   0.29  |  0.28 |  0.26 |
|         BU #72 Rafik B Hariri 20 YD    |    Academic   | -0.55 |  -0.55  | -0.52 | -0.55 |
|    BU #87 College of Engineering 20 Yd |    Academic   | -0.19 |  -0.18  | -0.15 | -0.23 |
|          BU #90 - School of Law 35     |    Academic   | -0.30 |  -0.30  | -0.28 | -0.35 |
|          BU #96 808 Comm Ave 35 Yd     |    Academic   | -0.06 |  -0.06  | -0.05 | -0.05 |
|              BU 685 Comm Ave           |    Academic   | -0.71 |  -0.71  | -0.64 | -0.72 |
|           BU Fenway Trash 20 Yd        |    Academic   | -0.58 |  -0.58  | -0.53 | -0.59 |
|         BU MED - 15 Stoughton 35 Yd    |    Academic   | -0.65 |  -0.64  | -0.59 | -0.68 |
|          BU Med 700 Albany 20 Yd       |    Academic   | -0.29 |  -0.29  | -0.27 | -0.30 |
| BU #102 Student Health Services 35 Yd  |    Medical    | -0.02 |  -0.03  |  0.01 | -0.06 |
|         BU #43 West Loading Dock       |     Other     | -0.08 |  -0.11  | -0.01 | -0.13 |
|              140 Bay State             |    Resident   | -0.04 |  -0.03  | -0.05 | -0.02 |
|        BU #105 Kilachand Hall 20 Yd    |    Resident   | -0.41 |  -0.40  | -0.42 | -0.39 |
|        BU #18 - Warren Hall 20 Yd      |    Resident   | -0.09 |  -0.08  | -0.08 | -0.10 |
|        BU #2 Student Village 35 Yd     |    Resident   | -0.01 |  -0.04  |  0.03 | -0.02 |
|         BU #46 30 Bay State 20 Yd      |    Resident   | -0.20 |  -0.20  | -0.17 | -0.20 |
|      BU #48 Student Village #2 20 Yd   |    Resident   |  0.12 |   0.11  |  0.15 |  0.11 |
|    BU #69 - Graduate Apartments 17 Yd  |    Resident   | -0.43 |  -0.43  | -0.41 | -0.45 |
|         BU #82 Warren Towers 35 Yd     |    Resident   |  0.06 |   0.07  |  0.06 |  0.04 |
|         BU Med 815 Albany 15 Yd        |    Resident   | -0.04 |  -0.05  | -0.02 | -0.05 |
|         BU #4 Yawkey Center 15 Yd      |    Service    | -0.29 |  -0.29  | -0.26 | -0.30 |
|    BU #93 George Sherman Union 35 Yd   |    Service    | -0.04 |  -0.05  |  0.00 | -0.08 |
|         BU #108 Agganis Arena 35 Yd    |     Sport     |  0.19 |   0.20  |  0.21 |  0.16 |
+----------------------------------------+---------------+-------+---------+-------+-------+
```

When we only look at the data during breaks, we can find out some resident buildings like BU #48 student village, and BU# 69 Graduate Apartments show a higher relationship than during school days, and for Warren Towers, there is no relationship to be found. We can easily understand the reason by assuming students are leaving Warren Towers during breaks.

There are also Maps and Linear Regression Graphs for these two datasets, to make the page short, I provide the link below:
- Correlation Map for all locations with all days
- Correlation Map for all locations without breaks
- Correlation Map for all locations with only breaks
- Linear Regression graphs for all locations with all days
- Linear Regression graphs for all locations without breaks
- Linear Regression graphs for all locations with only breaks

Conclusion

Does temperature impact waste generation (in terms of Psi)?
The temperature may not be the reason or the cause for waste generation, but during the analysis, we can find out that temperature and waste generation do have some correlations.

Can we use temperature as a predictor of waste generation and service level requirements?
We can definitely use the temperature as a reference for predicting waste generation. However, this has to be done by cases, we need to consider whether we want to predict the waste

generation during breaks or during school opening days. Once we decide which dates to predict, we need to focus on different locations. In general, we can find that during school opening days, there are stronger relationships being found in Academic Areas, and during school breaks, there are stronger relationships being found in Residential Areas. However, we cannot guarantee that we can make precise predictions on all locations. By looking at the higher absolute value of correlation coefficients and higher value of R-squared value of Linear Regression, we can make reliable predictions in locations like **BU 685 Comm Ave**, **BU MED**, and **Warren Towers**. For a complete prediction graph for all locations, please check the above Correlation Coefficient tables & Maps and Linear Regression Graphs.

Topic: Analysis of metadata - Timur Zhunussov

Intro

My team has been working on correlating temperature and compressor PSI values. To further our analysis, I decided to investigate additional data, such as temperature readings from different locations, the number of compactions, and the difference between compactions and waste pickups. However, we found that there was a low correlation between temperature and PSI values, so I explored different models to predict temperature.

Assumption

Upon an initial analysis of the data, I discovered that some inconsistencies exist. For instance, certain locations lack temperature readings for specific time periods. To address these gaps, we need to determine an appropriate method for filling them. Given that all the locations are within Boston, we can reasonably assume that there won't be significant differences in temperature readings between these sites.

To fill the gaps, I have attempted to combine data from different locations with one another. In the example below, we compare temperature readings at GSU, 140 Bay State, and SHS:
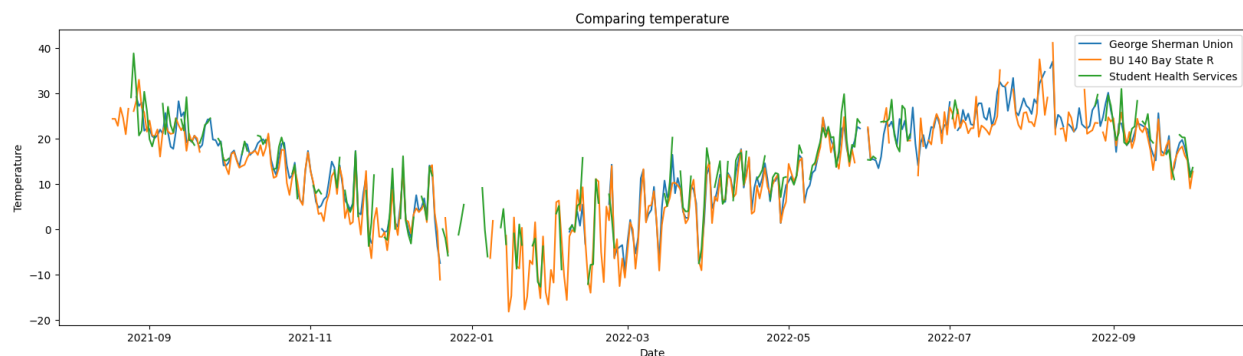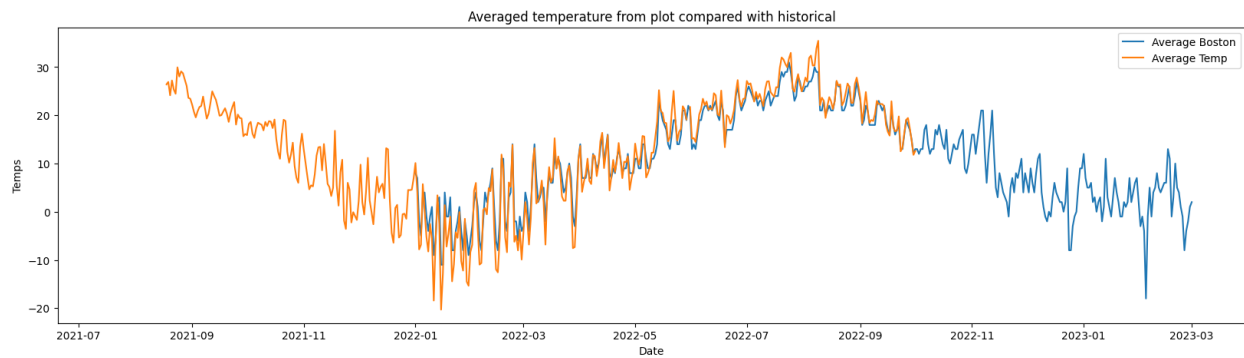
In this plot, we can observe the temperature readings for Grad Apartments, 22 Babit, and Agganis Arena:
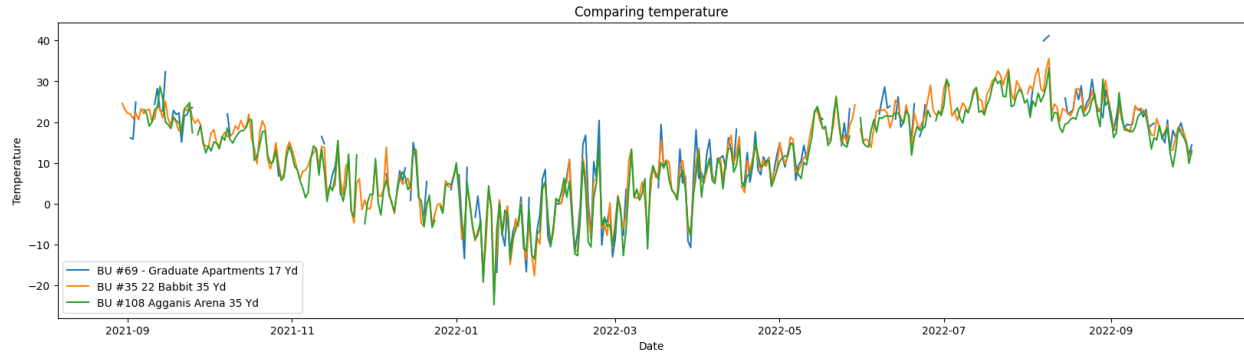
By merging the temperature data from Grad Apartments, 22 Babit, and Agganis Arena and others we can calculate the average temperature and compare it with historical data from the National Centers for Environmental Information Climate Data Online (NCEI CDO):


We can observe that the data is nearly identical.

The second assumption involved distinguishing between compactions and waste pickups using the provided data, with the goal of identifying any patterns regarding the frequency of pickups at certain locations and times. I began by examining compactions at different locations:
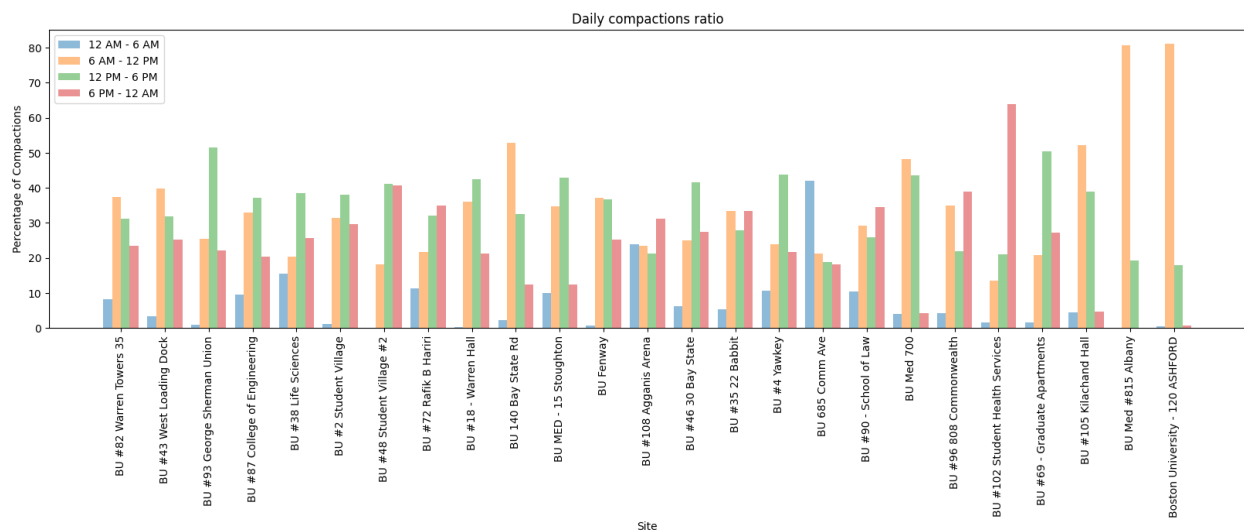

Due to the large number of compactions, we may obtain a substantial amount of data in a single day, making it challenging to interpret the information. As we can see in this example, the PSI

values for Warren Towers are tightly packed together, which complicates our understanding of the data.

Readings PSI for BU #82 Warren Towers 35
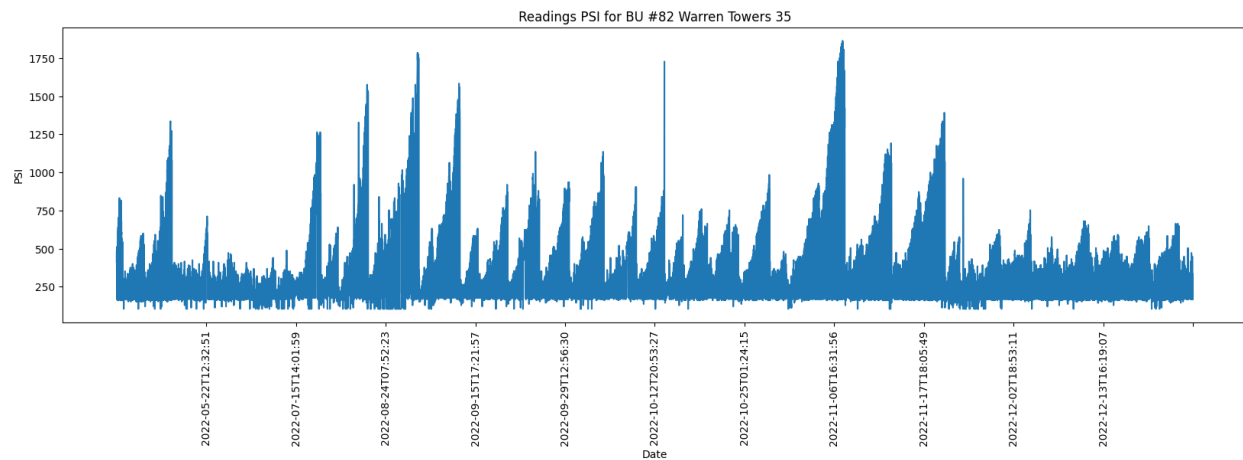
To address this issue, I decided to use the find_peaks method and apply a smoothing function over 50 neighboring data points. This approach made the data more readable and easier to interpret.

Smoothed Readings PSI for BU #82 Warren Towers 35
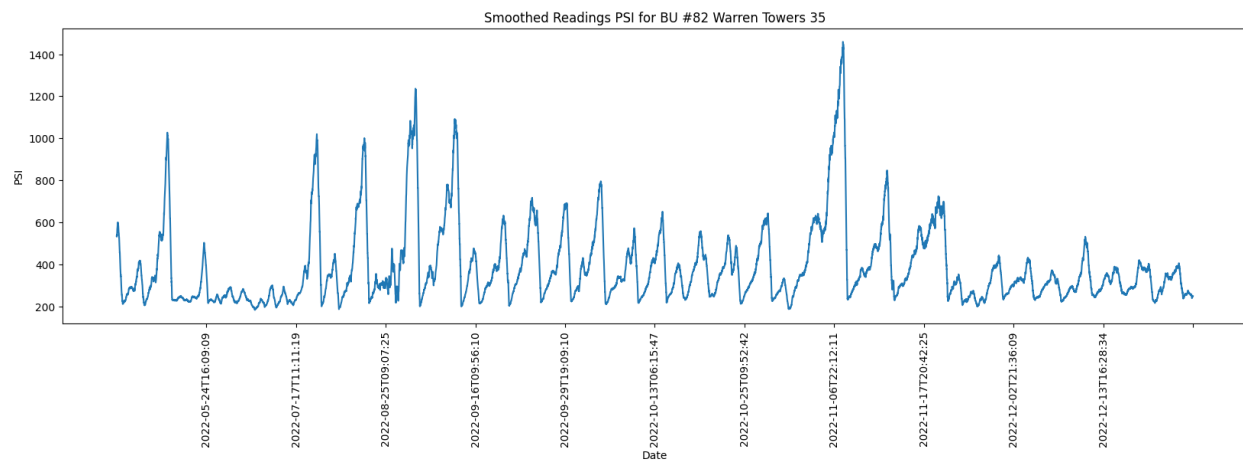
Furthermore, I attempted to count the number of waste pickups from the compaction data by dividing it into two groups: one group with small changes over a given period and another with larger changes. The results showed that there were 130 compactions with PSI values around 30 and 24 pickups with PSI values above 200. This analysis allowed us to gain a better understanding of the distribution of waste pickups and compactions based on PSI values.

However it is difficult to be sure that data is correct since it has some fluke in data.

Compactions to Pickups Readings PSI for BU #82 Warren Towers 35
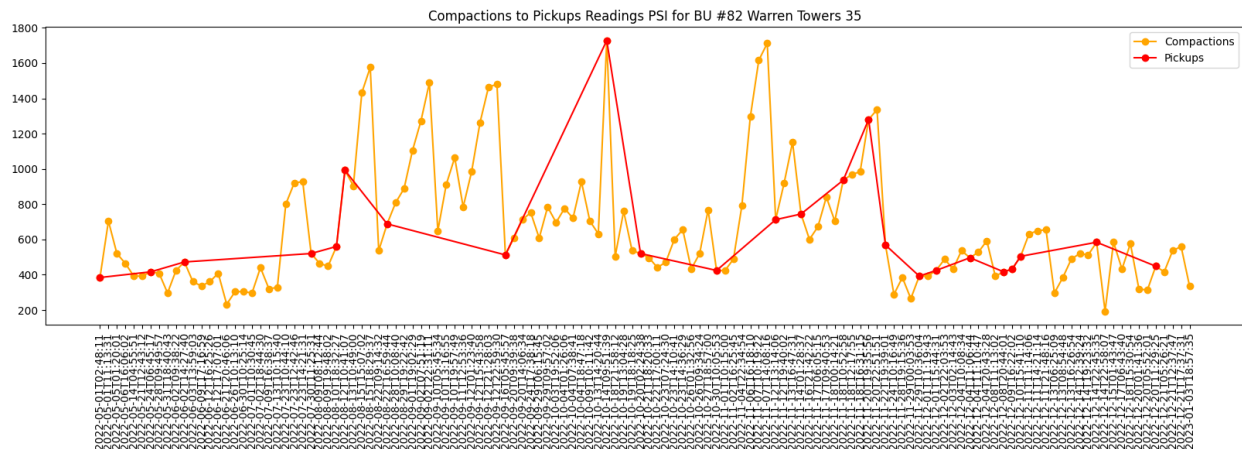
Finally, I explored three different models to assess any correlation in prediction, similar to what we did in class. Although the available data is limited and there aren't many features to build a robust model, I proceeded with splitting the data into training and testing sets using an 80/20 split. Then, I applied Linear Regression, Decision Trees, and Support Vector Machine (SVM) models to evaluate their performance and identify any predictive correlation between them.

	Linear regression		Decision Tree Regressor		Support Vector Regressor model	
Location	Mean Squared Error:	R2 Score:	Mean Squared Error:	R2 Score:	Mean Squared Error:	R2 Score:
BU #82 Warren Towers 35	67377.7331180751	0.055698641246183	69109.9194139194	0.031421987861094	74100.1375851558	-0.038516099716274
BU 685 Comm Ave	100976.516048402	0.233205310386646	86001.527638191	0.346922262004064	101193.176907185	0.231560032826443
BU #87 College of Engineering	78854.7925924103	0.047336397963115	3293.83729662078	0.960206364123717	81045.0889649751	0.020874903826654
BU #108 Agganis Arena	130728.798081008	0.123984088834434	26532.1149773071	0.822207843886144	137521.0957206	0.078468786215586
BU #102 Student Health Services	229846.249010561	0.338282067108051	172392.421052632	0.503689283613482	330488.630850318	0.048536773639000
BU #90 - School of Law	88341.6306385578	0.071955582900738	74976.9928400955	0.212353461067177	88639.8807315468	0.068822413049819
BU #35 22 Babbit	95327.0473616633	0.064126292799813	9284.53731343284	0.908849014045354	99834.8969907191	0.019870459218398
BU #72 Rafik B Hariri	39883.9454972524	0.043540856127409	24249.7746144721	0.418464788834417	41236.2034608535	0.011112281721453
BU #48 Student Village #2	148228.055191439	0.055669837237208	11504.5005045409	0.926707215986025	162215.439853557	-0.033440886218921
BU MED - 15 Stoughton	11507.8978574742	0.135986960441621	4812.63977746871	0.638666976894513	12187.3657197957	0.084972509298802
BU #38 Life Sciences	95935.7873106331	0.061528726695001	34917.7662337662	0.658424437251302	100754.916724546	0.014386626295567
BU #96 808 Commonwealth	79336.2321458321	0.058535393187572	66867.7293233083	0.206496214994304	82645.2417529984	0.019268146124174

From the results, we can observe that Linear Regression and SVM display similar outcomes, indicating low or no correlation in their predictions. However, Decision Trees demonstrate better results, although it's important to consider the possibility of overfitting, which might be causing this improved performance.

Topic: - Baicheng Fang

Topic: - Akshad Ramnath