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	Senior	zatimur@bu.edu
Akshad Ramnath	Senior	shadr@bu.edu
Baicheng Fang	First Year MS	bcfang@bu.edu

Introduction

The aim of this project is to support BU's Sustainability department in their efforts to transform the university's planning, operations, and culture towards a sustainable and equitable future. Specifically, the project seeks to investigate the correlation between weather conditions and waste production and storage. To accomplish this, we will analyze detailed data sets from various sources, including monitors, overall data, temperature data, and waste generation spreadsheets from Casella.

Our analysis of these datasets will provide valuable insights to BU Sustainability, enabling them to identify potential areas for waste storage improvement in the event of adverse weather conditions. Through a comprehensive analysis of the data provided by both the Sustainability Department and third-party vendors, the project aims to deepen our understanding of how external factors impact the waste collection process.

Our ultimate goal is to offer insights that will help to enhance the waste collection process and support BU's sustainability objectives.

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)

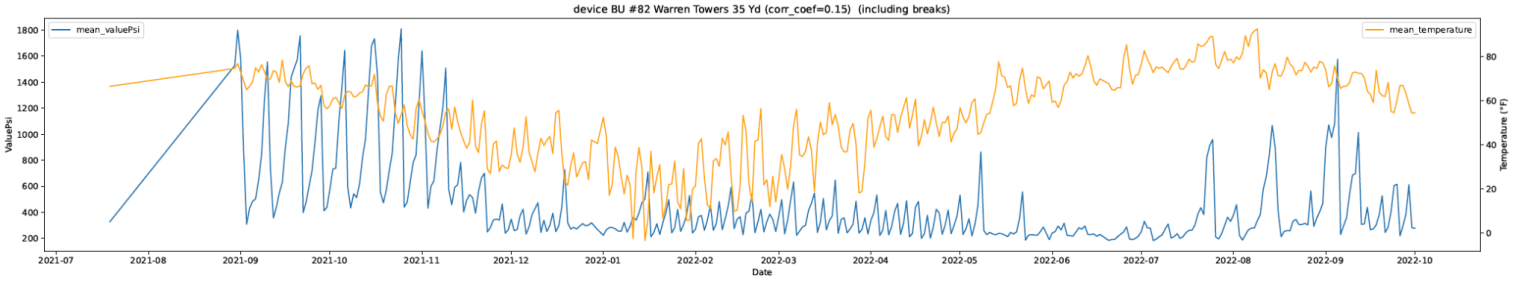
Assumption

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*.

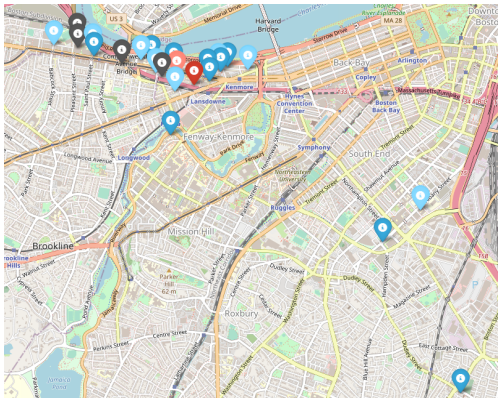
* Correlation coefficient is a statistical measurement that describes the degree of relationship between two variables. (For example the psi value and temperature) and the coefficients are ranged from +1 to -1. If the correlation coefficient is positive, it means that there is a direct relationship between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

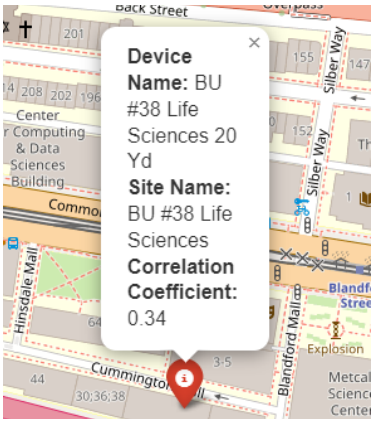
Correlation Coefficients Between Mean ValuePsi and Fahrenheit (All Dates)

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 a coefficient above 0.1 and the blue color represents the negative relationship with a 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:

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 the psi value. Apart from Life Science and Warren Towers, which show a positive correlation, other residents and academic buildings are showing negative correlations. In general, the colder the weather, the more wastes are being generated at the locations, assuming the higher the psi value, the more wastes are generated.

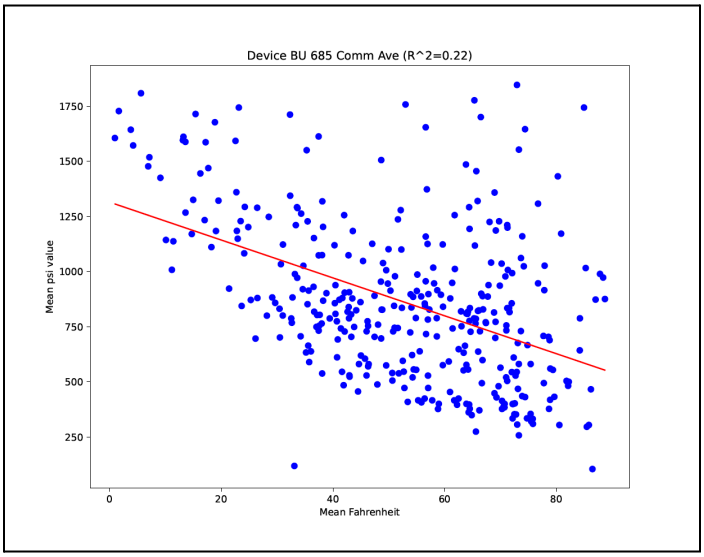
Further Analysis - Linear Regression

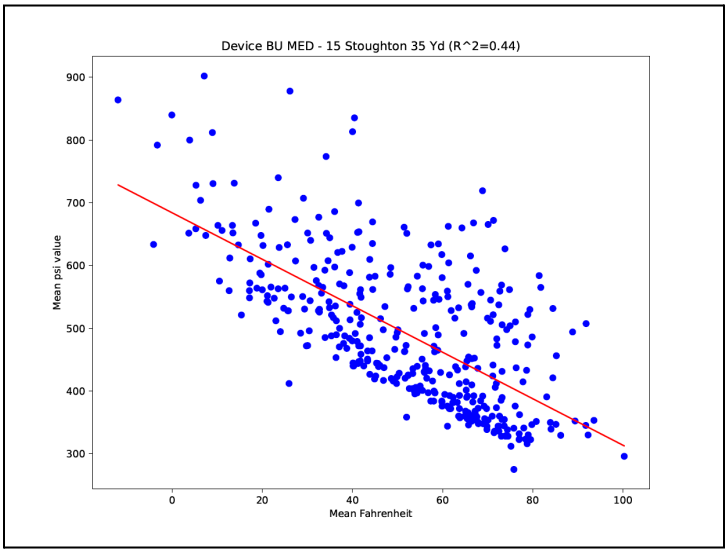
Linear regression provides more information than just the correlation coefficient. While the correlation coefficient measures the strength and direction of the linear relationship between two variables, linear regression can help predict the values of one variable based on the values of another variable.

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:





(for complete graphs, check here)

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.

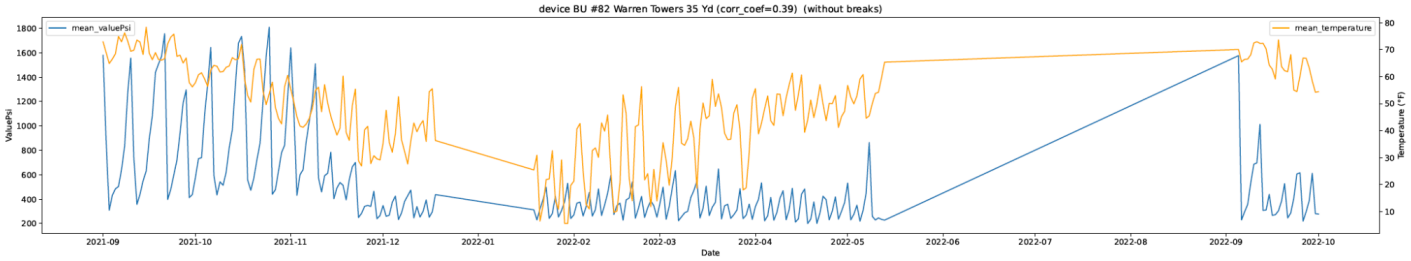
Extension idea

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.

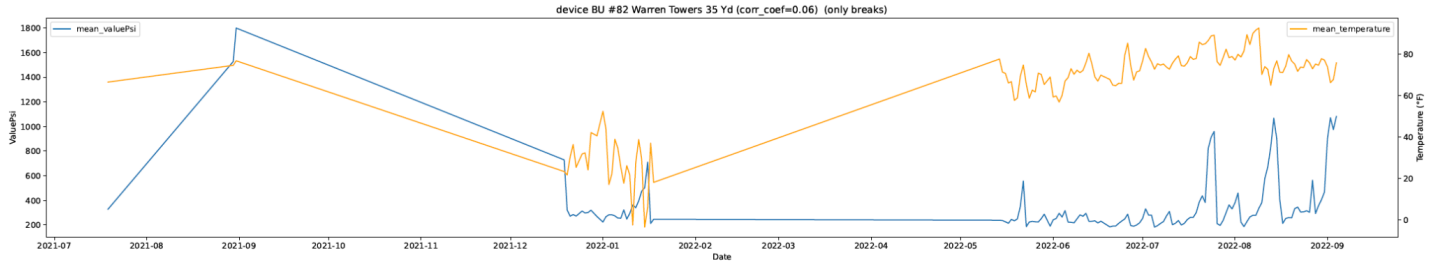
Here I separated the original data with all dates into two sets: **Without Breaks** and **Only Breaks**. Based on the data we have, and by referencing the BU Calendar, I have separated out the following dates as breaks:

- 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 temperature versus Psi value at Warren Towers' data without breaks and only breaks.



(sample line graph of data without break, for complete graph, check here)



(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 and the amount of waste generated is not large enough to have a correlation with the temperature.

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.

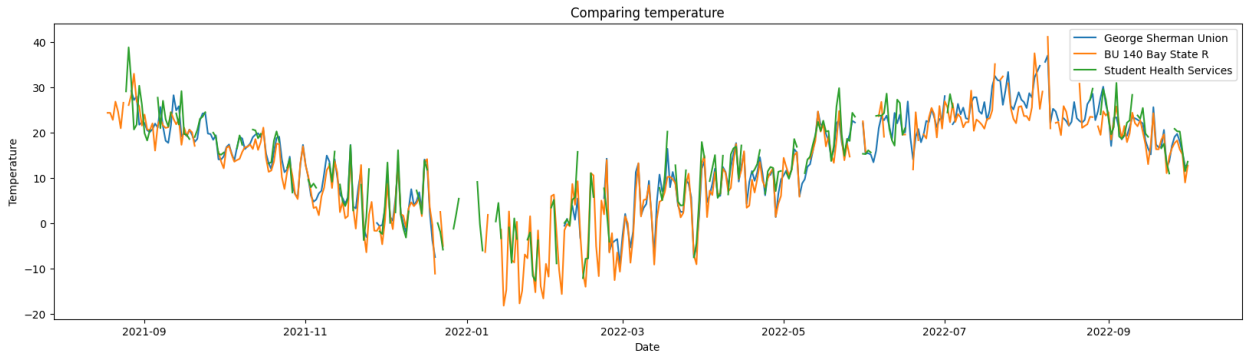
Topic: Analysis of metadata - Timur Zhunussov

Intro

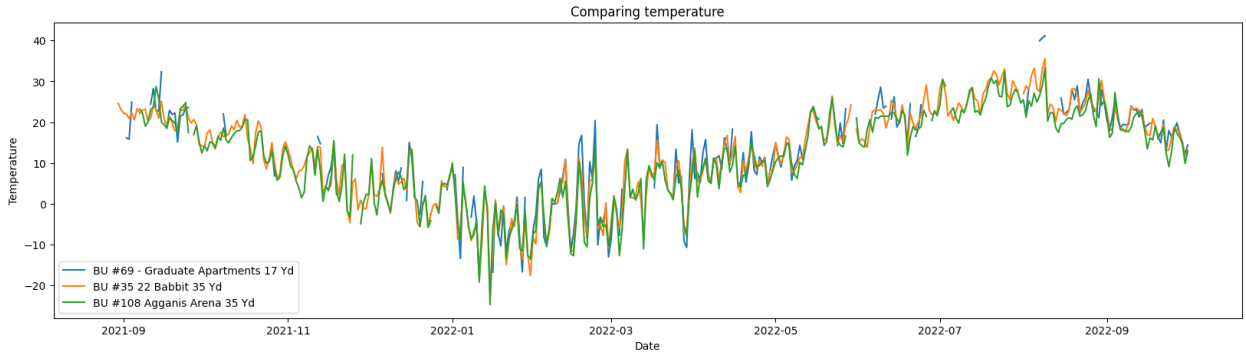
My team has been working on correlating temperature and compressor PSI values. To further our analysis, I decided to investigate additional data, such as temperature readings from different locations, the number of compactions, and the difference between compactions and waste pickups. However, we found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

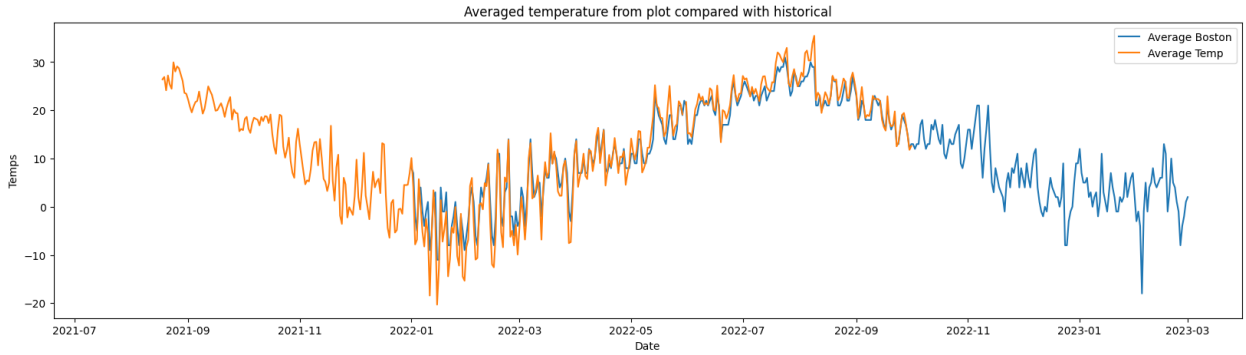
To fill the gaps, I have attempted to combine data from different locations with one another. In the example below, we compare temperature readings at GSU, 140 Bay State, and SHS:



In this plot, we can observe the temperature readings for Grad Apartments, 22 Babit, and Agganis Arena:

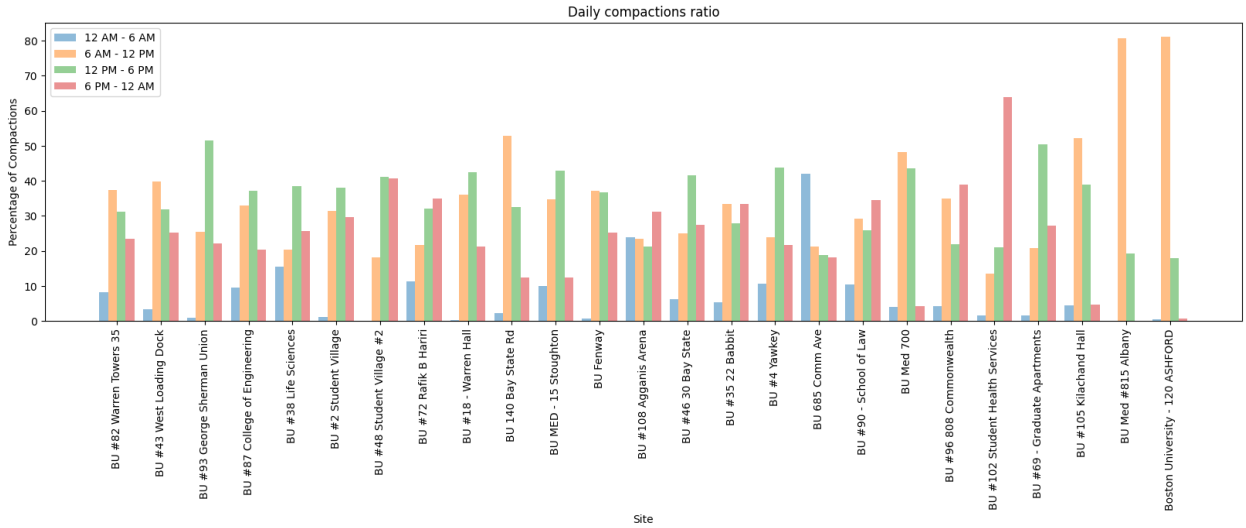


By merging the temperature data from Grad Apartments, 22 Babit, and Agganis Arena and others we can calculate the average temperature and compare it with historical data from the National Centers for Environmental Information Climate Data Online (NCEI CDO):



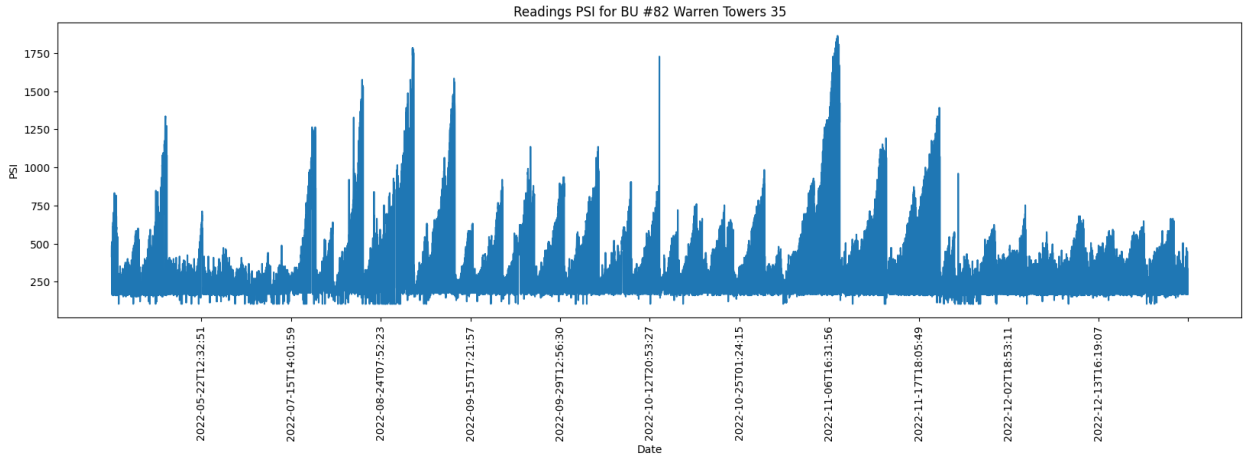
We can observe that the data is nearly identical.

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:

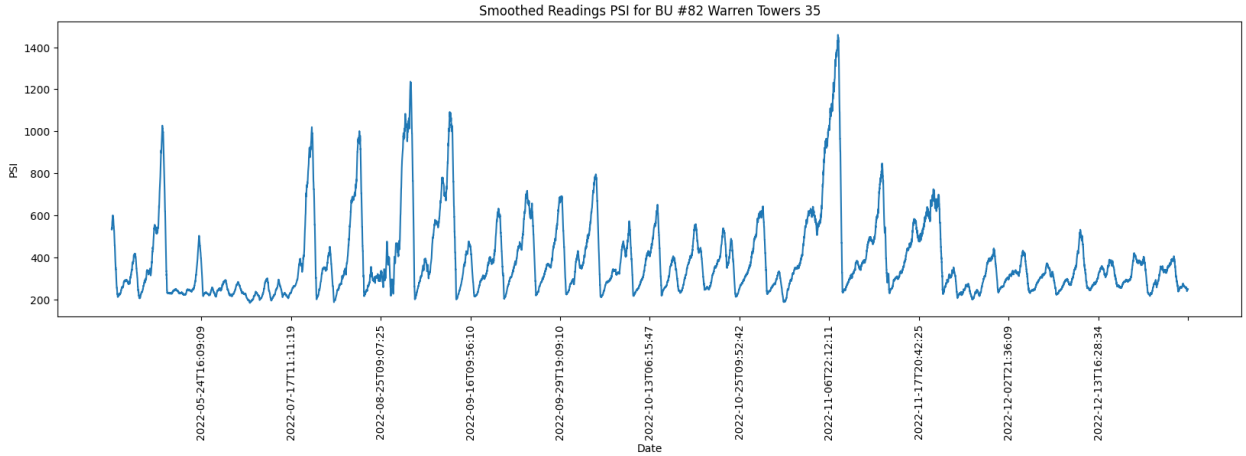


Due to the large number of compactions, we may obtain a substantial amount of data in a single day, making it challenging to interpret the information. As we can see in this example, the PSI

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

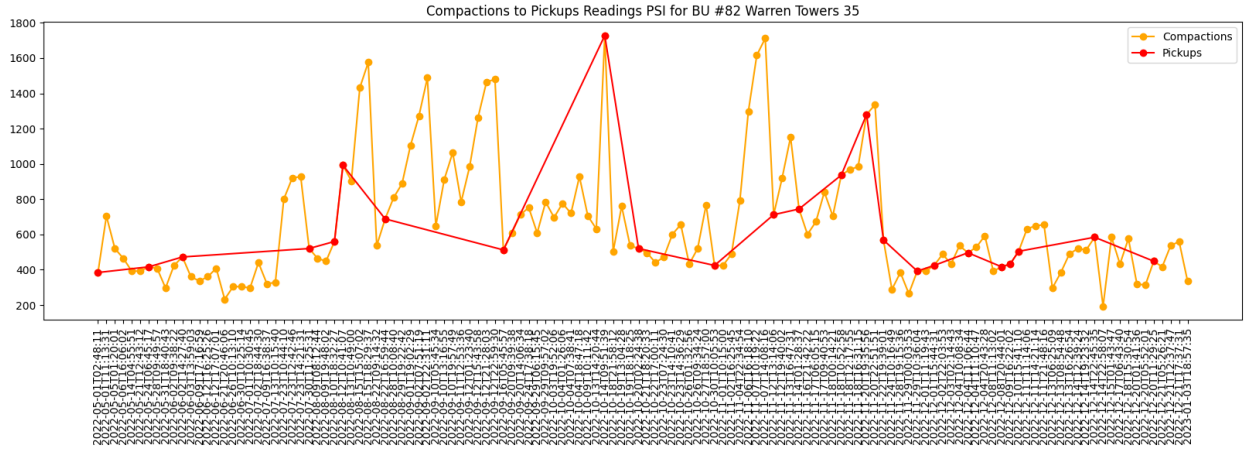


To address this issue, I decided to use the find_peaks method and apply a smoothing function over 50 neighboring data points. This approach made the data more readable and easier to interpret.



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.



Finally, I explored four different models to assess any correlation in prediction, similar to what we did in class.
Choosing the right model depends on the specific problem we are solving, as well as the quality of your data, the complexity of the model, and other factors. So I chose these models:

1. Linear Regression: This is a simple model that tries to fit a linear relationship between the input features and the target variable. It works well when the relationship between the features and the target is linear, and when there are few outliers or noise in the data.
2. Support Vector Regression: This is a more powerful model that tries to find the best "hyperplane" to separate the input features and the target variable. It works well when the relationship between the features and the target is nonlinear, and when there are many outliers or noise in the data.
3. Decision Trees: This is a model that tries to create a tree-like structure to represent the relationship between the input features and the target variable. It works well when there are many possible combinations of features that can predict the target, and when there are non-linear relationships between the features and the target. However Decision trees are often prone to overfitting and may not generalize well to new data like in our case.
4. MLPRegressor: This is a neural network model that tries to find the best combination of weights and biases to map the input features to the target variable. It works well when there are many possible combinations of features that can predict the target, and when there are complex, non-linear relationships between the features and the target.

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) and Machine learning neural network models to evaluate their performance and identify any predictive correlation between them.

	Linear regression		MLPRegressor		Support Vector Regressor model		Decision Tree Regressor	
Location	Mean Squared Error:	R2 Score:	Mean Squared Error:	R2 Score:	Mean Squared Error:	R2 Score:	Mean Squared Error:	R2 Score:
BU #82 Warren Towers 35	67377.7331180751	0.055698641246183	54892.2549607304	0.230683067748191	74100.1375851558	-0.038516099716274	69109.9194139194	0.031421987861094
BU 685 Comm Ave	100976.516048402	0.233205310386646	97371.9624526949	0.260577541711109	101193.176907185	0.231560032826443	86001.527638191	0.346922262004064
BU #87 College of Engineering	78854.7925924103	0.047336397963115	69616.1355393785	0.158950822613468	81045.0889649751	0.020874903826654	3293.83729662078	0.960206364123717
BU #108 Agganis Arena	130728.798081008	0.123984088834434	123883.17486113	0.169856727078542	137521.0957206	0.078468786215586	26532.1149773071	0.822207843886144
BU #102 Student Health Services	229846.249010561	0.338282067108051	211503.488257438	0.391090123716932	330488.630850318	0.048536773639000	172392.421052632	0.503689283613482
BU #90 - School of Law	88341.6306385578	0.071955582900738	82999.7520248526	0.128072960273523	88639.8807315468	0.068822413049819	74976.9928400955	0.212353461067177
BU #35 22 Babbit	95327.0473616633	0.064126292799813	91998.7649525261	0.096801720006017	99834.8969907191	0.019870459218398	9284.53731343284	0.908849014045354
BU #72 Rafik B Hariri	39883.9454972524	0.043540856127409	35891.4361309839	0.139285448162005	41236.2034608535	0.011112281721453	24249.7746144721	0.418464788834417
BU #48 Student Village #2	148228.055191439	0.055669837237208	128194.038916591	0.183302395224865	162215.439853557	-0.03344088621892	11504.5005045409	0.926707215986025
BU MED - 15 Stoughton	11507.8978574742	0.135986960441621	11088.094137792	0.16750582621127	12187.3657197957	0.084972509298802	4812.63977746871	0.638666976894513
BU #38 Life Sciences	95935.7873106331	0.061528726695001	80725.6392303962	0.210318739641874	100754.916724546	0.014386626295567	34917.7662337662	0.658424437251302
BU #96 808 Commonwealth	79336.2321458321	0.058535393187572	73313.2892303404	0.130008255339181	82645.2417529984	0.019268146124174	66867.7293233083	0.206496214994304
BU 140 Bay State Rd	115795.258312892	0.065322370620836	99981.5419673159	0.192967726060334	141152.428404311	-0.139355955367	46770.0100502513	0.622481241833876

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

Topic: Multi-linear Analysis - Baicheng Fang

Intro

Our team has conducted numerous analyses exploring the relationship between temperature and waste production, with the aim of fitting an appropriate linear model. Despite our efforts, the results derived from the linear model have been less than satisfactory. In an attempt to enhance the performance and predictive capabilities of our model, we have decided to shift our focus towards a multiple linear regression approach. By incorporating an additional significant variable, namely pressure, we hope to develop a more robust and accurate model that better captures the underlying relationships among these factors.

Add a new feature

Thanks to Zeqi for his diligent efforts in pre-processing an ample amount of pressure data and merging it according to the respective sites. The dataset now contains average, maximum, and minimum pressure values for each location. Although our aim is to add only one extra feature to avoid overcomplicating the model, it is essential to determine which pressure metric—average, maximum, or minimum—holds the greatest significance. To accomplish this, I conducted a preliminary analysis using Spearman's rank correlation coefficient as a non-parametric measure of association. The outcomes of this analysis are depicted in Figure 3-1, which provides

valuable insight into the relative importance of each pressure variable, allowing us to select the most pertinent feature for our refined model.

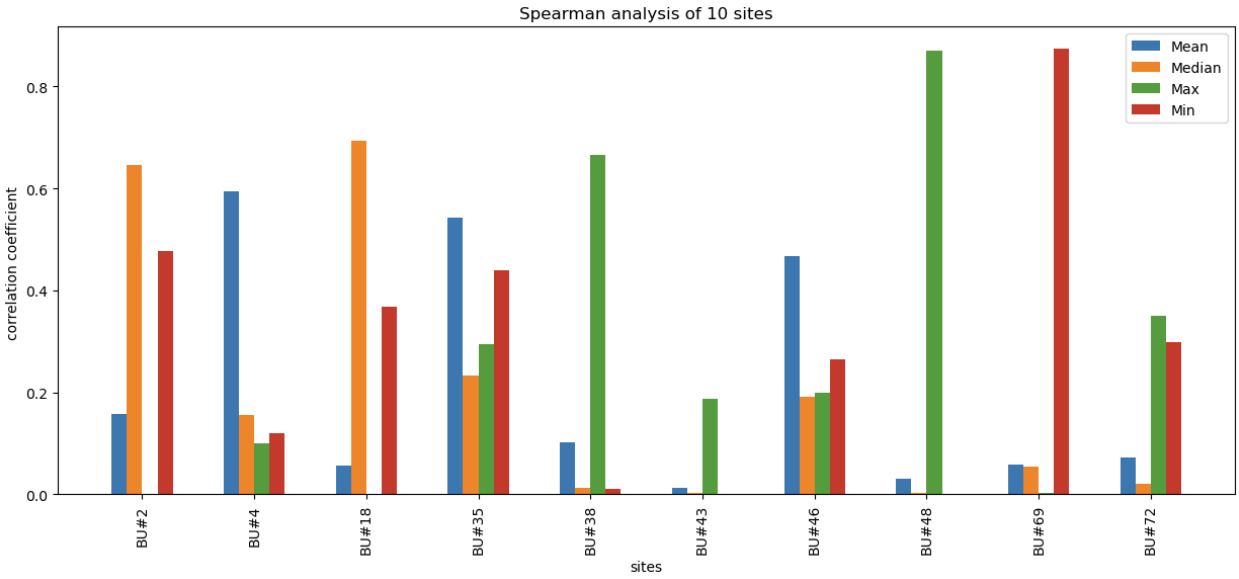


Figure 3-1

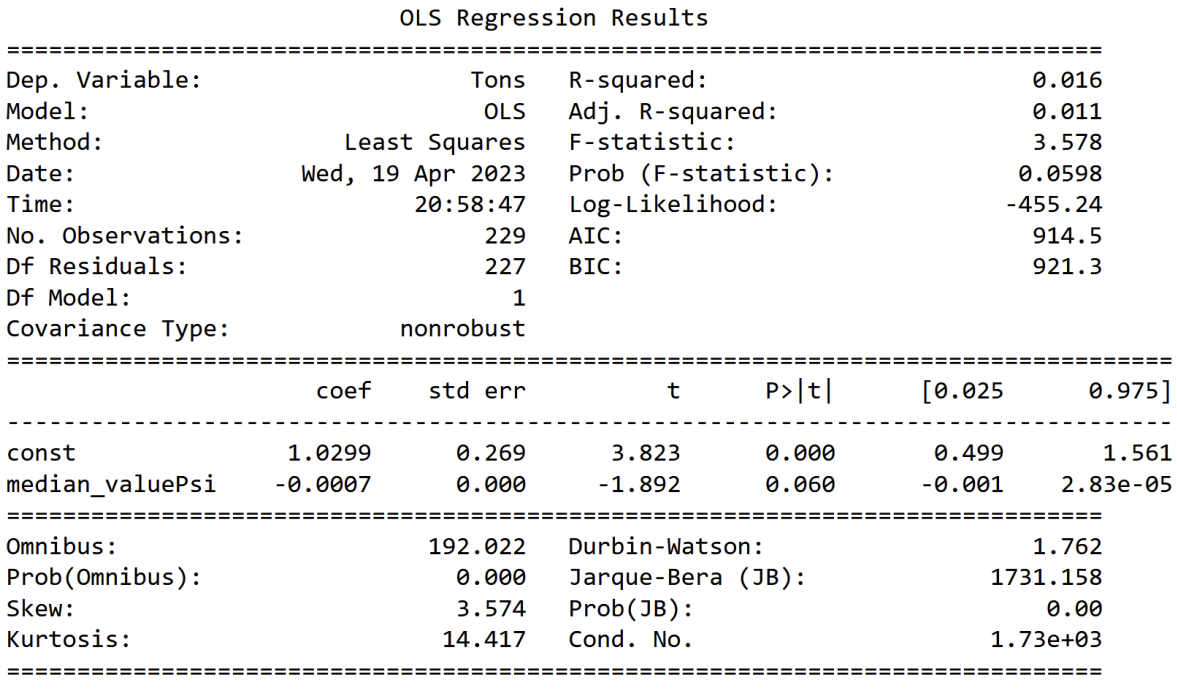
The plot reveals that the key factors influencing the relationship between temperature and waste production vary across different sites. In order to generate a model that is as accurate and precise as possible, I have taken a tailored approach by fitting each site's data with the most relevant vital factors. This customized method allows us to account for the unique characteristics and driving forces at each location, ultimately leading to a more robust and reliable model that better captures the intricacies of the underlying relationships.

OLS Analysis

Ordinary Least Squares (OLS) is a popular statistical technique employed to estimate the coefficients of a linear regression model. The primary goal of OLS is to identify the best-fitting line through the data points, which is achieved by minimizing the sum of the squared differences (also known as residuals) between the observed values of the dependent variable and the values predicted by the linear regression model.

By minimizing the sum of squared residuals, OLS seeks to reduce the overall discrepancy between the actual data points and the model's predictions, resulting in a more accurate and reliable representation of the relationships between the independent and dependent variables. This method is widely utilized in various fields for its simplicity and effectiveness in uncovering linear relationships and generating meaningful insights from data.

Figure 3-2 shows an example of OLS analysis on pressure data vs temperature on BU #2 Student Village.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Tons   R-squared:                       0.016
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     3.578
Date:                Wed, 19 Apr 2023   Prob (F-statistic):             0.0598
Time:                        20:58:47   Log-Likelihood:                -455.24
No. Observations:                 229   AIC:                             914.5
Df Residuals:                     227   BIC:                             921.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              1.0299      0.269      3.823      0.000       0.499       1.561
median_valuePsi   -0.0007      0.000     -1.892      0.060      -0.001    2.83e-05
==============================================================================
Omnibus:                      192.022   Durbin-Watson:                   1.762
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1731.158
Skew:                           3.574   Prob(JB):                         0.00
Kurtosis:                      14.417   Cond. No.                     1.73e+03
==============================================================================
```

Figure 3-2

Among the various details presented in Figure 3-2, we should pay particular attention to the Prob (F-statistic), or the p-value, which is associated with each independent variable. The p-value represents the probability of observing a more extreme test statistic (t-value) if the null hypothesis were true, i.e., if there were no relationship between the independent variable and the dependent variable.

A small p-value (typically less than 0.05) suggests that the corresponding independent variable is significantly related to the dependent variable, thereby indicating that the variable contributes meaningfully to the model. By analyzing the p-values, we can determine the significance of each independent variable in the multilinear regression model and better understand the relationships between the independent variables and the dependent variable. This information is crucial for selecting the most relevant predictors and developing a robust and reliable model.

In this case, we perform a linear regression with OLS on a dataset with one feature, median_valuePsi and one dependent variable mean_fahrenheit. The p-value is 0.0598, which is slightly greater than the threshold value of 0.05, so we may conclude that pressure has correlation, but not very significant, with temperature.

In addition to the OLS regression analysis, I have incorporated the Mean Squared Error (MSE) as another metric to evaluate the performance of our regression models, including linear regression. The MSE quantifies the average squared difference between the observed values of the dependent variable and the predicted values generated by the model. Lower MSE values are indicative of a better model fit, as they suggest that the model's predictions are closer to the actual data points.

By utilizing both the OLS and MSE evaluation metrics, we can obtain a more comprehensive understanding of the models' performance and their ability to accurately capture the relationships between temperature, pressure, and waste production. This dual-metric approach allows us to identify the most robust and reliable models for each site, ultimately enhancing our analysis and interpretation of the underlying data.

Figure 3-3 shows the result of MSE of temperature vs waste production on 10 sites.

Indeed, a lower MSE value signifies a better fit for the linear model, and in this case, site #38 and site #72 demonstrates the most optimal fitting performance. However, it is crucial to recognize that the interpretation of MSE is contingent upon the scale of the dependent variable. As such, comparing MSE values across different models or incorporating additional evaluation metrics can offer a more comprehensive assessment of model performance.

In this context, I have employed the Ordinary Least Squares (OLS) regression analysis mentioned earlier as an alternative metric to further evaluate and compare the fitting quality of the models for each site. By utilizing both MSE and OLS, we can gain a more nuanced understanding of the models' accuracy and effectiveness in capturing the underlying relationships between temperature, pressure, and waste production across various locations.

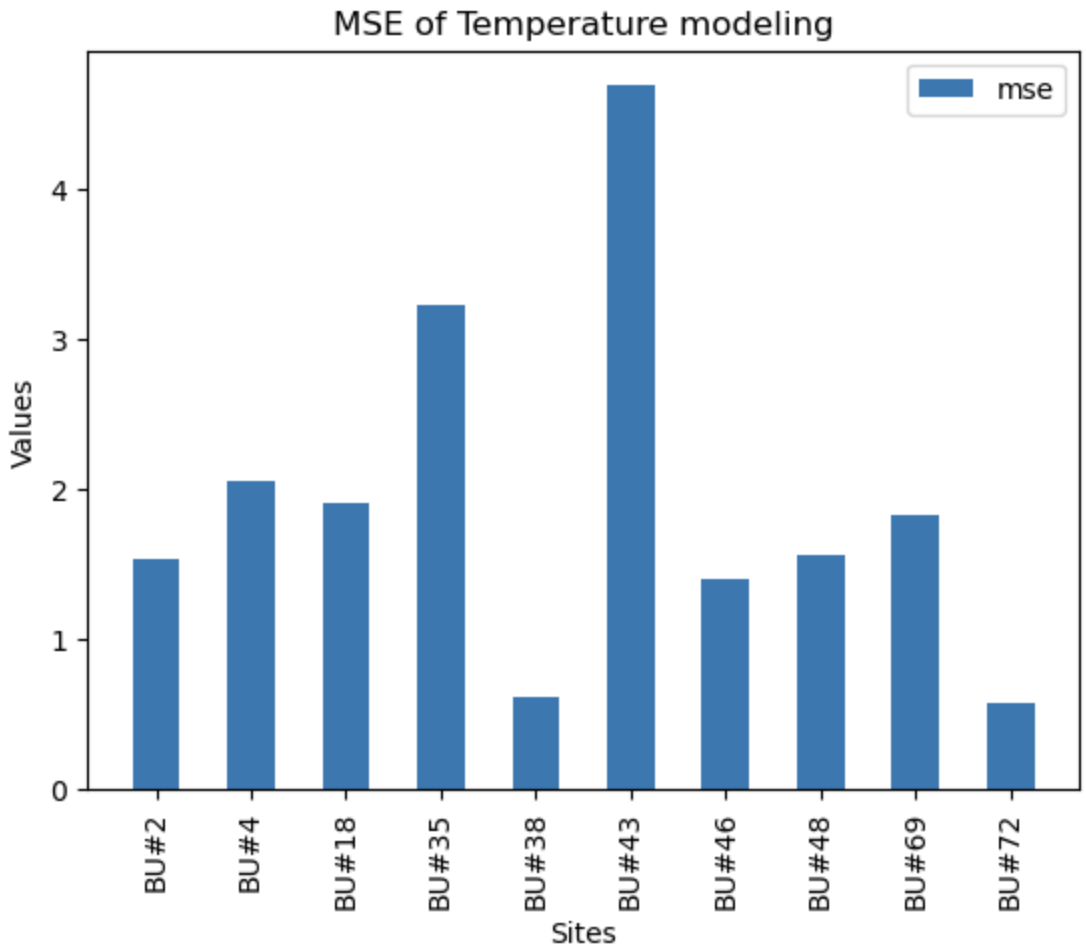


Figure 3-3

Figure 3-4 presents the Ordinary Least Squares (OLS) regression analysis results for 10 distinct sites. In addition to the sites previously identified as having promising fitting performance based on the Mean Squared Error (MSE) analysis, sites #2, #43, and #48 also exhibit strong potential in terms of model fit. The combination of these findings highlights a subset of locations where the customized multiple linear regression models demonstrate enhanced accuracy and predictive capabilities, ultimately contributing to a more effective understanding of the relationships between temperature, pressure, and waste production.

Multi-Linear Regression

Multilinear regression analysis, commonly referred to as multiple linear regression, is a statistical technique that extends simple linear regression to model the relationship between a dependent variable and multiple independent variables. In this method, the dependent variable is assumed to be a linear combination of the independent variables, such as pressure and temperature in this case.

Multiple linear regression aims to create a linear equation that best describes the relationship between the dependent variable and the set of independent variables, allowing for the prediction of the dependent variable based on the values of the independent variables. This approach

enables the analysis of more complex relationships and provides a deeper understanding of the factors influencing the dependent variable, leading to more accurate predictions and insights.

Figure 3-5 shows the comparison of P-values of single feature, i.e., temperature with multi-linear regression, i.e., pressure and temperature.

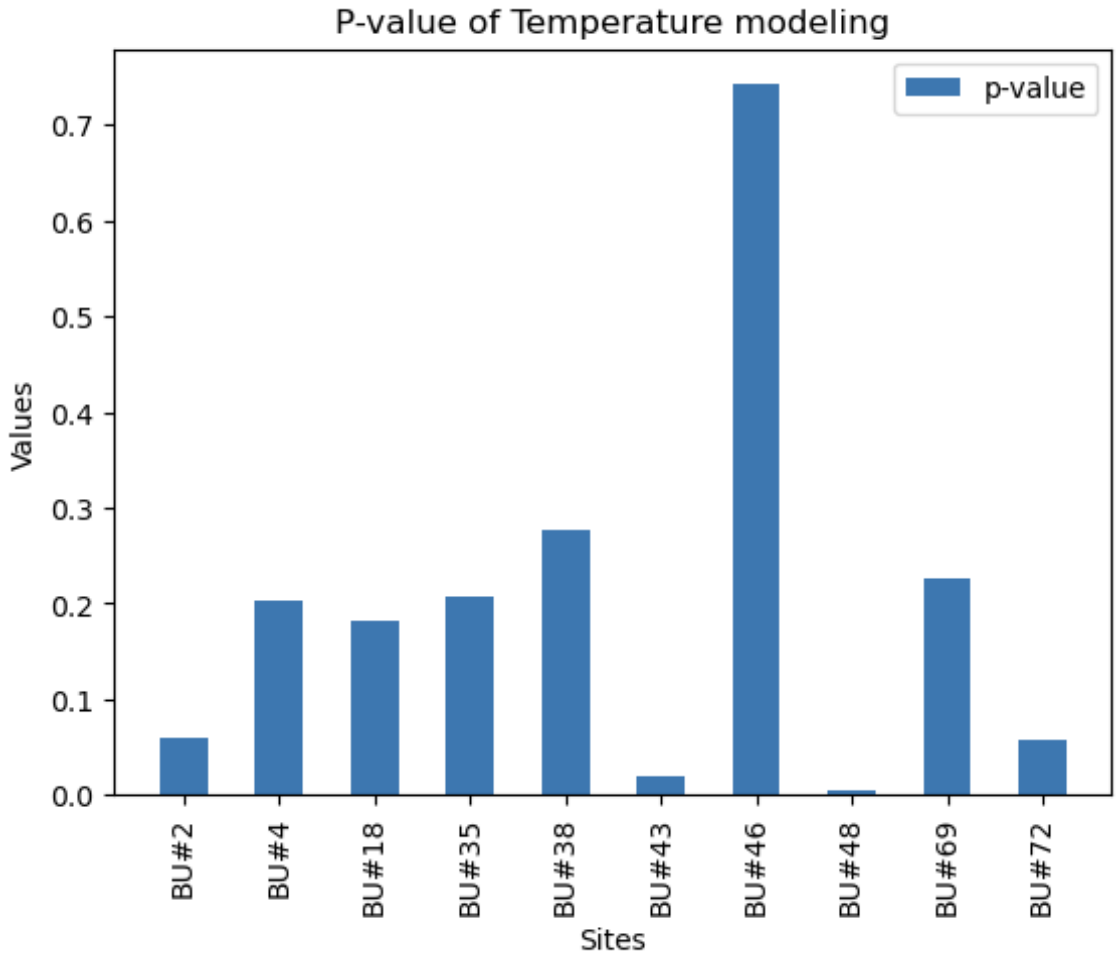


Figure 3-4

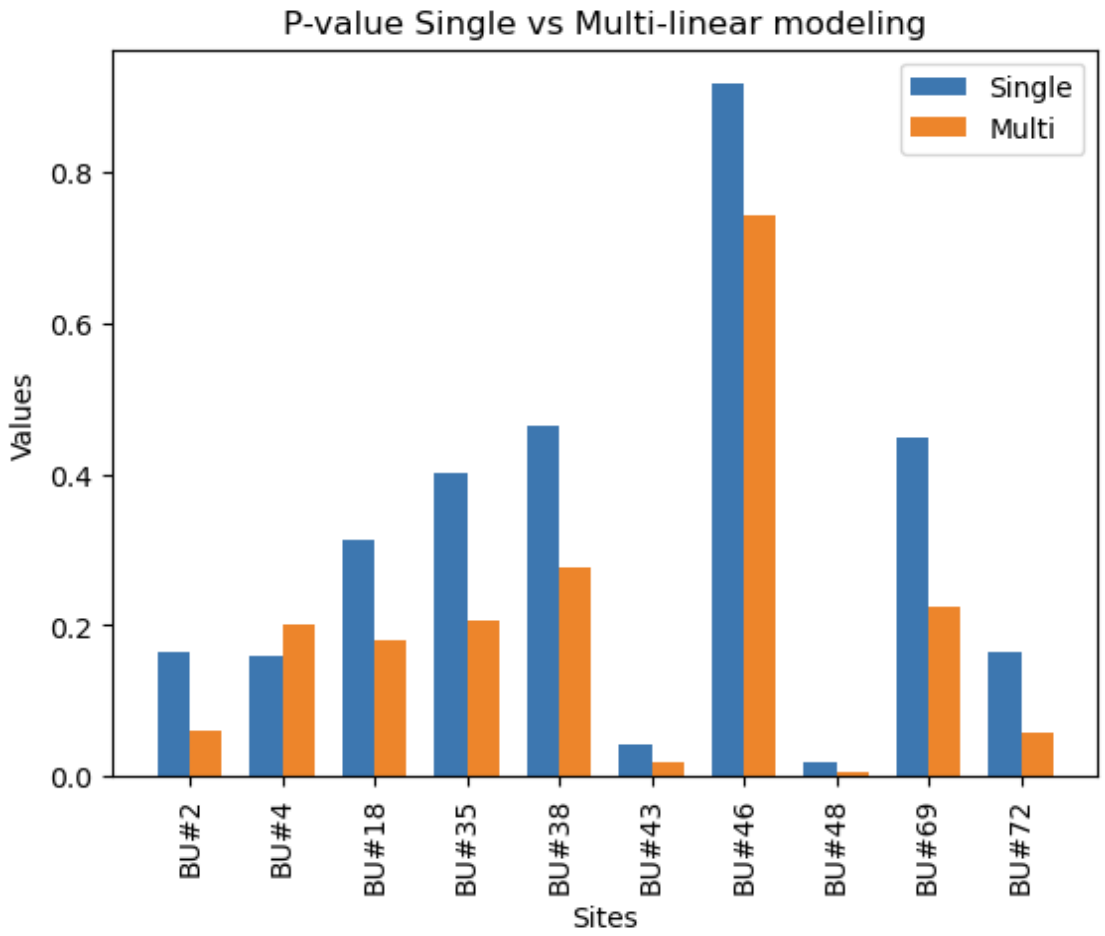


Figure 3-5

Upon examining Figure 3-5, we can confidently deduce that incorporating pressure as an additional independent variable results in a substantially reduced P-value. This decrease in P-value is a strong indicator that the extended linear model, which now includes both pressure and temperature as predictors, has a higher likelihood of effectively accounting for the variability observed in the dependent variable. Consequently, the inclusion of pressure as a new feature enhances the overall explanatory power and robustness of the linear model.

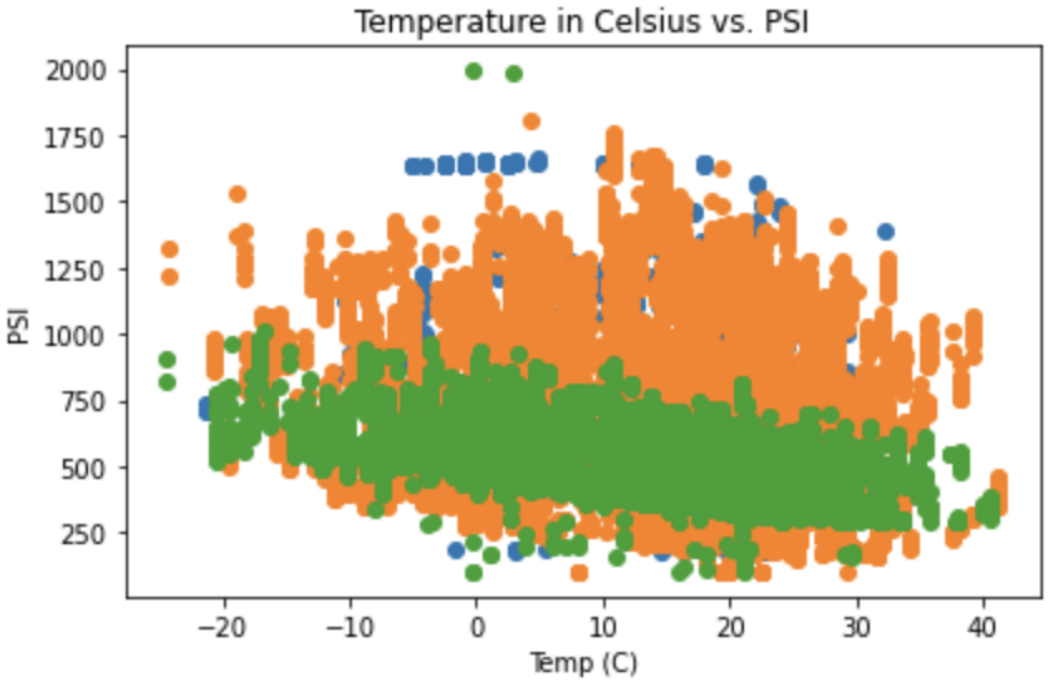
Topic: Analysis - Akshad Ramnath

Intro

Just like my team members, I tried to perform some linear regression to find some correlations between temperature and psi values. With little success in finding correlations, I looked to do some non-linear regression using a polynomial regression model. I then proceeded to try to find outliers using z-score and excluding them from the data to see if there were any correlations.

Preliminary analysis

I created a single data frame containing all the data from the data files, drop any rows with missing values, and plotted the scatter plot for the combined data.



As Zeqi did earlier, I decided to find the correlation coefficient as well to find patterns.
The correlation coefficient is a statistical measure that tells you how strong the relationship is between two variables. It ranges from -1 to 1, with values closer to -1 or 1 indicating a stronger relationship, and values closer to 0 indicating a weaker relationship. To calculate the correlation coefficient between temperature and PSI values in the data, I used the pandas corr() function.
Correlation Coefficient: -0.16769074792721786
The correlation coefficient being closer to 0 indicates a weak relationship between temperature and psi values.

Regression Analysis

I also decided to perform OLS (ordinary least squares) which is a statistical method used to analyze the relationship between a dependent variable and one or more independent variables. The goal of OLS regression is to find the line (or hyperplane) that best fits the data, in the sense that it minimizes the sum of the squared differences between the predicted values and the actual values. To do this I used the statsmodels library in Python. This will output a summary of the regression results, including the regression coefficients and other statistical measures.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                valuePsi   R-squared:                       0.028
Model:                            OLS   Adj. R-squared:                  0.028
Method:                 Least Squares   F-statistic:                     537.1
Date:                Mon, 24 Apr 2023   Prob (F-statistic):          3.75e-117
Time:                        14:10:31   Log-Likelihood:            -1.3097e+05
No. Observations:               18564   AIC:                         2.619e+05
Df Residuals:                   18562   BIC:                         2.620e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        716.1158      3.105    230.638      0.000     710.030     722.202
celsius       -4.2883      0.185    -23.175      0.000      -4.651      -3.926
==============================================================================
Omnibus:                     1972.507   Durbin-Watson:                   0.095
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2669.155
Skew:                           0.880   Prob(JB):                         0.00
Kurtosis:                       3.597   Cond. No.                         25.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The regression results will tell you the strength and direction of the relationship between temperature and PSI values in your data. In particular, you can look at the sign of the regression coefficient to determine whether the relationship is positive or negative. A positive coefficient indicates a positive relationship (as temperature increases, so does PSI), while a negative coefficient indicates a negative relationship (as temperature increases, PSI decreases). Since the coefficient is a positive value, there is a positive relationship.

Exploring Non-Linear Relationships

While a simple linear regression assumes a linear relationship between temperature and PSI values, there may be more complex, non-linear relationships at play. Thus, I tried fitting a polynomial regression model to see if it provides a better fit to the data. This would allow for a more complex, non-linear relationship between temperature and PSI values. To do this, I used the numpy and sklearn libraries.

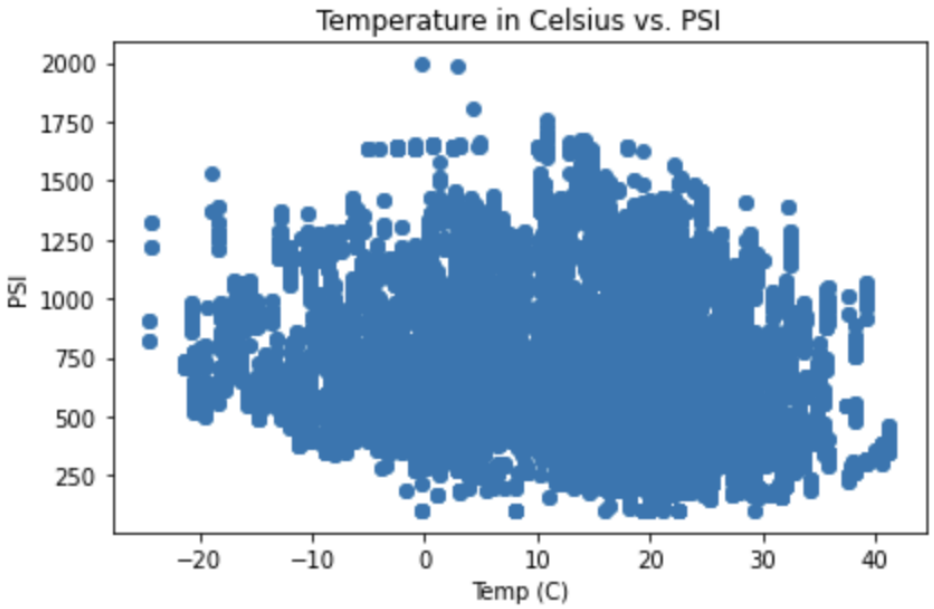
```
                            OLS Regression Results
==============================================================================
Dep. Variable:               valuePsi   R-squared:                       0.030
Model:                           OLS   Adj. R-squared:                  0.030
Method:                Least Squares   F-statistic:                     289.3
Date:               Mon, 24 Apr 2023   Prob (F-statistic):          1.95e-124
Time:                       14:13:09   Log-Likelihood:            -1.3095e+05
No. Observations:              18564   AIC:                         2.619e+05
Df Residuals:                  18561   BIC:                         2.619e+05
Df Model:                          2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        717.4413      3.109    230.787      0.000     711.348     723.535
x1            -2.5544      0.330     -7.748      0.000      -3.201      -1.908
x2            -0.0821      0.013     -6.351      0.000      -0.107      -0.057
==============================================================================
Omnibus:                     1924.759   Durbin-Watson:                   0.095
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2584.763
Skew:                           0.868   Prob(JB):                         0.00
Kurtosis:                       3.573   Cond. No.                         605.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The R-squared value: This measures the proportion of variation in the dependent variable (PSI) that is explained by the independent variable(s) (temperature). An R-squared value of 1 indicates a perfect fit, while a value of 0 indicates no relationship between the variables. In general, a higher R-squared value indicates a better fit of the model to the data. Since the R-squared value is close to 0, there is no relation between the variables. The results show an R-squared value of 0.03 which indicated a weak relation between the variables.

Identifying Outliers

I then decided to find and remove outliers to see if the models would provide better results without the outliers. I used the z-score from the scipy library to identify data points that are significantly different from the mean of the data. I printed out any data points that have a z-score greater than 5, indicating that they are significantly different from the rest of the data.
I then proceeded to exclude outliers from my analysis by removing them from the dataset.
In this example, any data points with a z-score greater than 3 are considered outliers.
Now, I can perform analysis on the filtered dataset instead of the original dataset to see if removing outliers had any impact on my results. I plotted the scatterplot and performed linear regression.



OLS Regression Results

Dep. Variable:	valuePsi	R-squared:	0.027
Model:	OLS	Adj. R-squared:	0.027
Method:	Least Squares	F-statistic:	1007.
Date:	Mon, 24 Apr 2023	Prob (F-statistic):	3.78e-218
Time:	14:33:03	Log-Likelihood:	-2.5987e+05
No. Observations:	36960	AIC:	5.197e+05
Df Residuals:	36958	BIC:	5.198e+05
Df Model:	1		
Covariance Type:	nonrobust		

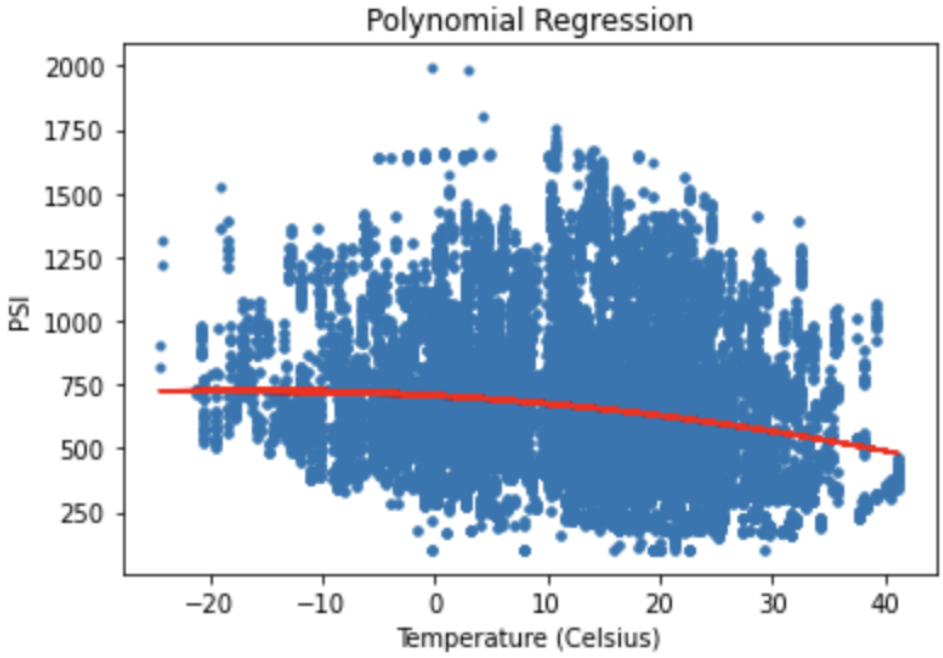
	coef	std err	t	P>|t|	[0.025	0.975]
const	709.0030	2.151	329.661	0.000	704.788	713.218
celsius	-4.0620	0.128	-31.740	0.000	-4.313	-3.811

Omnibus:	3307.423	Durbin-Watson:	0.050
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4259.745
Skew:	0.809	Prob(JB):	0.00
Kurtosis:	3.383	Cond. No.	25.4

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

As you can see, there is not much of a difference between the analysis of filtered data and the unfiltered data since the change in the coefficient and other values are negligible, again

indicating a weak relationship.



On performing polynomial regression, the line does indicate a weak relationship between temperature and psi. As you can see the line starts going down the temperature rises indicating that the trash impact is lower in the warmer climates but, this makes logical sense as the warmer climate would indicate summer temperatures when the campus has the least amount of activity.