Visualization Term Project

20141087 Ryeongyun Kim 20161206 Jaewoong Lee

December 4, 2019

Contents

List of Tables

List of Figures

1 Introduction

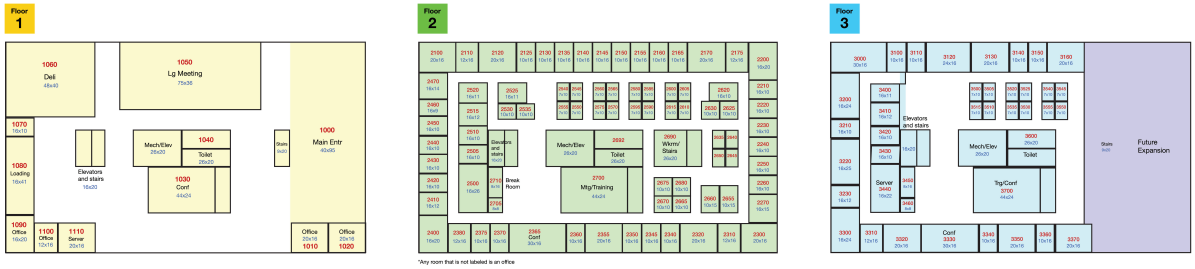
In this term project, we have to answer several question with virtual building data.

2 Materials

2.1 Building Layout

To analyzing movement data, we should find corresponding coordinate with zone data. To find matching coordinate, we calculate the approximate center of all zones, and consider the approximate center coordinate as representative of its zone.

2.1.1 Main Layout

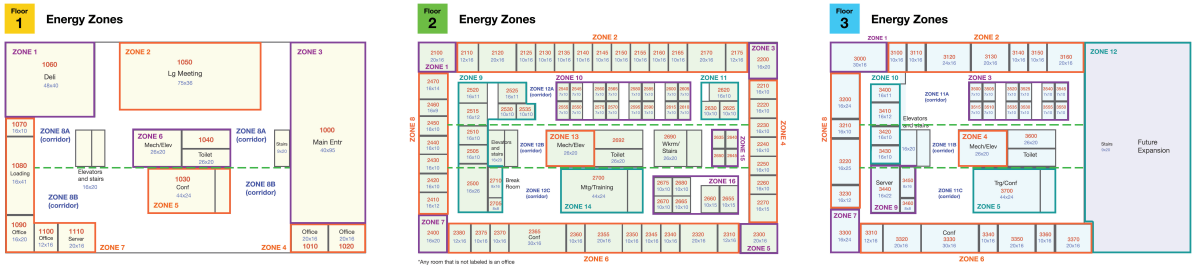


(a) First Floor (b) Second Floor (c) Third Floor

Figure 1: Main Layout of the building

The main layout of this building is as figure 1.

2.1.2 Energy Zone Layout



(a) First Floor (b) Second Floor (c) Third Floor

Figure 2: Energy Zone of the Building

The energy zone of this building is as figure 2.

2.1.3 Prox Zone Layout

The prox zone of this building is as figure 3.

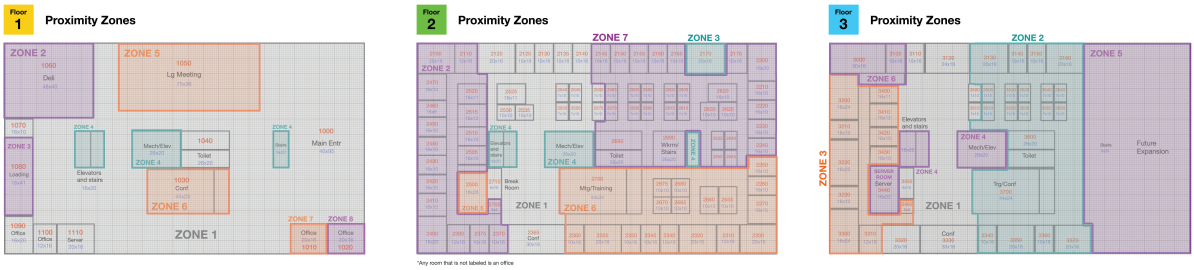
3 Methods

3.1 Python Packages

To analyze data, we used Python programming language. Also, we adopt many Python modules as hereinafter.

3.1.1 Scikit-learn: Machine Learning in Python

Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems [1].



(a) First Floor (b) Second Floor (c) Third Floor

Figure 3: Prox zone of the Building

3.1.2 Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms [2].

3.1.3 Pandas

Pandas is a Python library of rich data structures and tools for working with structured data sets common to statistic, finance, social sciences, and many other fields [3].

3.1.4 SciPy

SciPy is a Python-based ecosystem of open-source software for mathematics, science, and engineering [4].

3.1.5 Pillow

Pillow is the Python Imaging Library. [5]

3.2 JavaScript

For implementing interactive plots in web, we used JavaScript for showing plots.

3.2.1 Plotly

Plotly is a high-level, declarative charting library. plotly.js ships with over 40 chart types, including 3D charts, statistical graphs, and SVG maps. [6]

3.3 TSNE

T-distributed Stochastic Neighbor Embedding (TSNE) is a machine learning algorithm for visualization high-dimensional data in a low-dimensional space [7].

3.4 Standardization

Note that, in this analysis, all values are standardized. In other words, all values are adjusted for the mean value is zero, and standard deviation is 1. If all values in one columns are same, then the column will be discarded.

4 Results

4.1 What are the typical patterns in the prox card data? What does a typical day look like for GAStech employees?

4.1.1 General Information of prox Data

First of all, we drew the distribution of movement distance as figure 4. Also, the basic statistics values, such as minimum, maximum, and average, of movement distance is in table 1.

Furthermore, the extreme value of moving information is in tables 2 and 3.

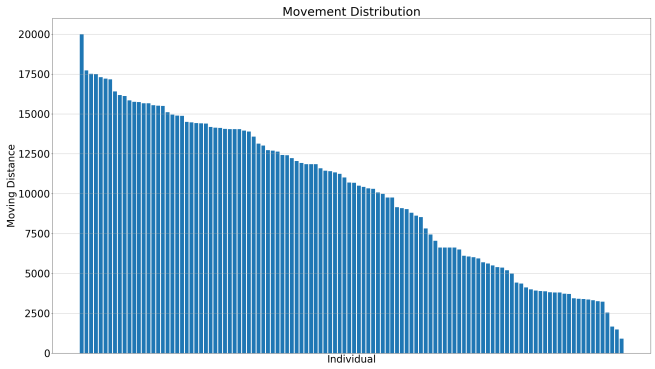


Figure 4: Distribution of Movement Distance

Table 1: Basic Statistics Data within Movement Distance

Item	Minimum	Maximum	Mean	q1	Median	q3	Standard Deviation
Value	902.44	19999.38	10083.95	5642.54	10688.57	14134.16	4750.46

Table 2: Minimum Moving Employees

Moving Distance	ID
902.4411338142776	earpa
1482.4411338142788	vawelon
1667.820095117141	jfrost
2550.198060545332	ibarranco
3233.4833039729083	cstaley

Table 3: Maximum Moving Employees

Moving Distance	ID
19999.386059326014	chawelon
17719.3756100877	hmies
17507.957735800737	eminto
17478.788651971503	monda
17302.863257165674	ldedos

4.1.2 Workflow

With the general information of prox data, we have decided our workflow for question 1 as figure 5.

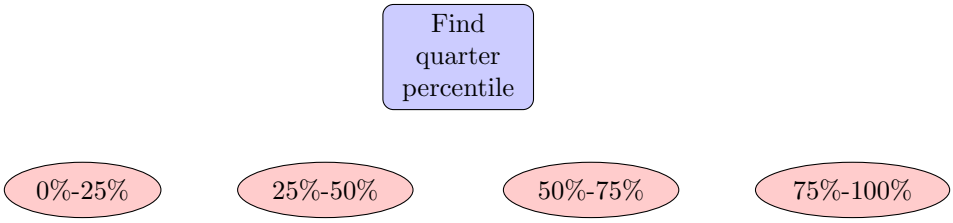
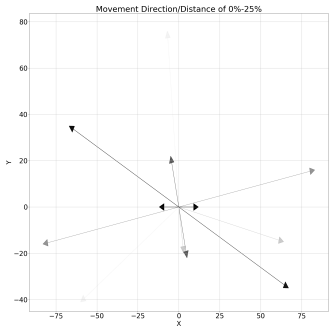


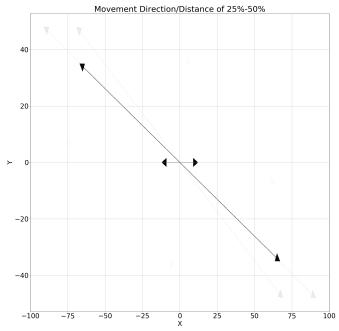
Figure 5: Workflow for Question 1

4.1.3 Movement Direction and Distance

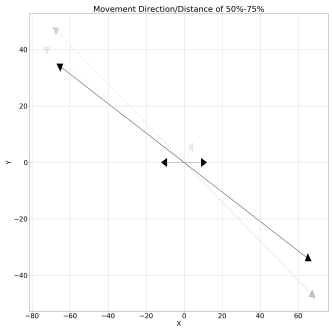
We drew the plot about movement direction and distance with each sub-group as figures 6, 7, and 8. Note that the darkness of arrow is proportioned with number of duplicates.



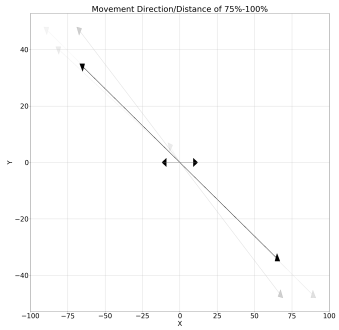
(a) 0%-25%



(b) 25%-50%



(c) 50%-75%



(d) 75%-100%

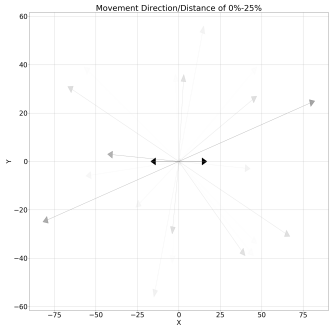
Figure 6: Movement Direction/Distance on First Floor

The movement direction and distance on the first floor is shows as figure 6. In figure 6-(b, c, d), you can see two arrows: one is left-upward arrow, the other is right-downward arrow.

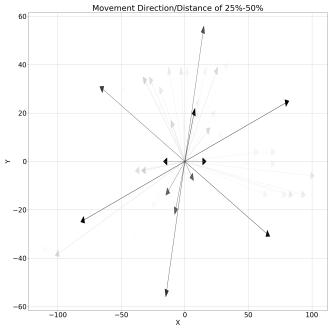
4.1.4 Department Distribution

There are seven departments in the data as followings: (in alphabetical)

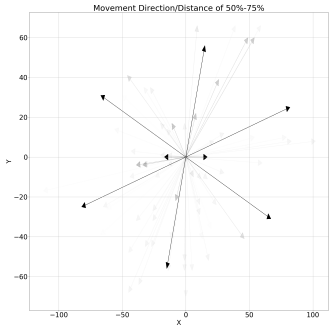
1. Administration
2. Engineering
3. Executive
4. Facilities



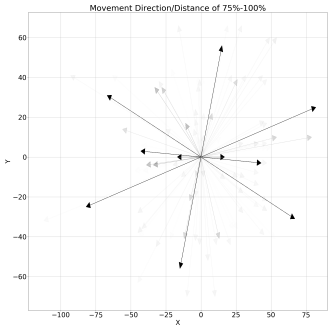
(a) 0%-25%



(b) 25%-50%

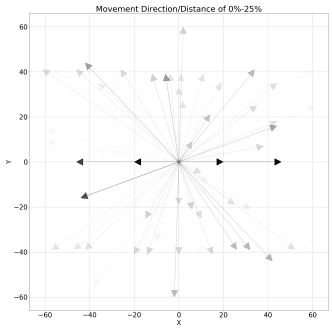


(c) 50%-75%

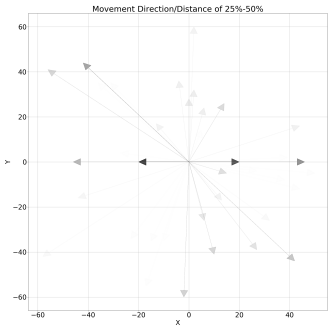


(d) 75%-100%

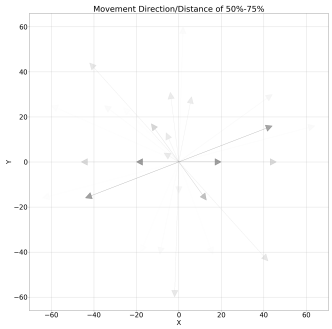
Figure 7: Movement Direction/Distance on Second Floor



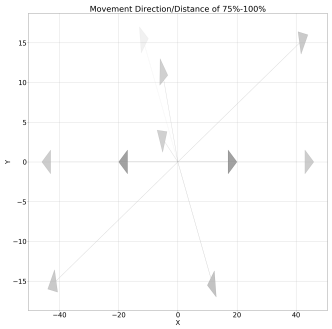
(a) 0%-25%



(b) 25%-50%



(c) 50%-75%



(d) 75%-100%

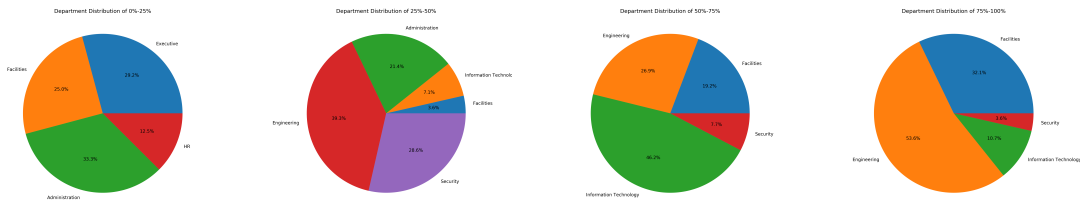
Figure 8: Movement Direction/Distance on Third Floor

5. HR
6. Information Technology
7. Security

Table 4 shows the distribution of department by quartiles. Also, with the data in table 4, we drew the four pie plots as 9.

Table 4: Department Information by Quartiles

Quartiles	Department	Counts
q1 (Minimum 25%)	Administration	8
	Executive	7
	Facilities	6
	HR	3
q2	Engineering	11
	Security	8
	Administration	6
	Information Technology	2
	Facilities	1
q3	Information Technology	12
	Engineering	7
	Facilities	5
	Security	2
q4 (Maximum 25%)	Engineering	15
	Facilities	9
	Information Technology	3
	Security	1

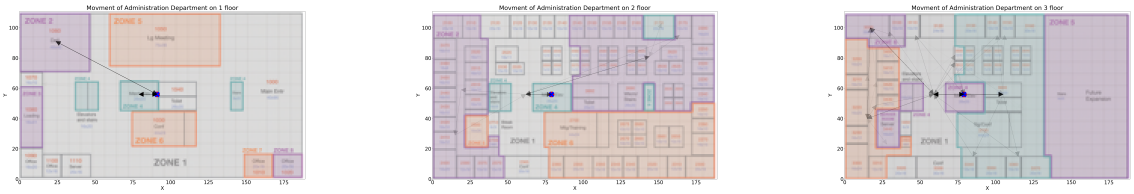


(a) q1 (b) q2 (c) q3 (d) q4

Figure 9: Pie Plot of Department Distribution by Quartiles

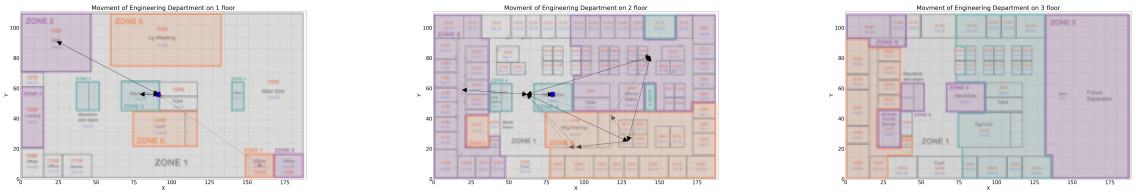
4.1.5 Typical Day Look for GAStech employees

First of all, when you look at prox card data by floor, there is a hub which almost everyone goes through. This pattern is most common on the second floor, with prox zone 1 going through the center the most. One can assume that this is because it is just in front of the elevator's entrance and is a passageway linking each office. Also, this is an area that must be passed if you use an elevator when you go to wok or go home.



(a) First Floor (b) Second Floor (c) Third Floor

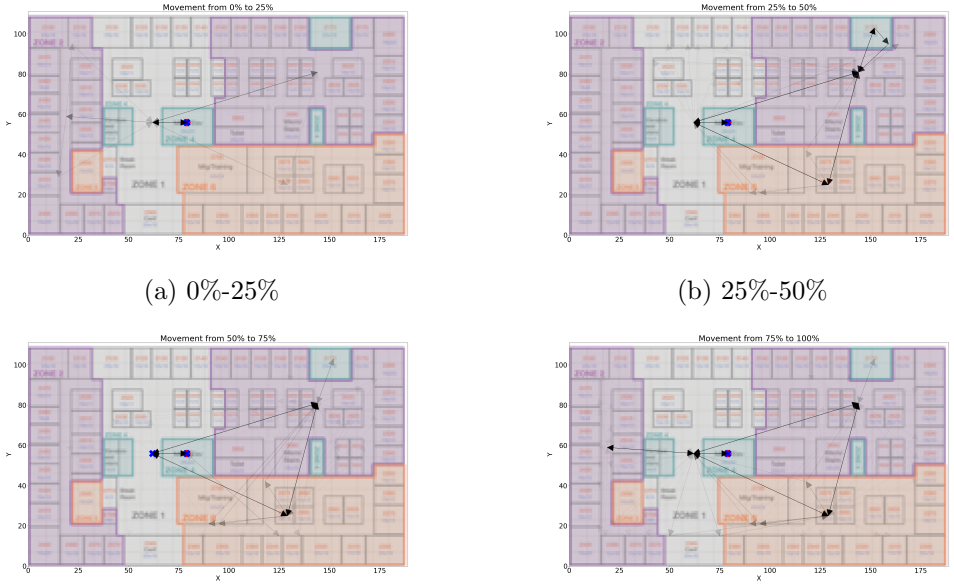
Figure 10: Movement within Administration Department



(a) First Floor (b) Second Floor (c) Third Floor

Figure 11: Movement within Engineering Department

Secondly, there are distinction layers of movement by department. In the case of the Administration department, as figure 10, there was the most movement on the third floor and there was clearly less movement on the second floor. According to the employee data, the administration department has an office in each of floor, so it needs further investigation into why this pattern comes out. In the engineering department, as figure 11, there was a lot of movement on the second floor and nobody went on the third floor. In the case of Executive Department, it moved the most on the third floor and hardly moved on the second floor. The other departments did not have much variation in their movement on each floor.



(a) 0%-25% (b) 25%-50%

(c) 50%-75% (d) 75%-100%

Figure 12: Movement Distribution by Percentile on Second Floor

Third, we looked at the pattern differences amongst people with many movements and people with few movements. According to figure 12, only those whose movement is between 25% and 75% did not move to zone 2 on second floor. What is really interesting about this pattern is that people on the second floor in zone 2 have a lot of movement, or very little movement. If you can see the role of each zone, you can see why this pattern appears.

Fourth, people with 0 % to 25 % movements are more likely to move on the third floor than other groups, indicating that people working on the third floor are less likely to move on the third floor than people on other floor. We think taht it will be because the department on the third floor is the execute department.

Finally, we identify the pattern by focusing on *Deli*. 23 of people were found not to use *Deli* at all, and more than half of those who did not use *Deli* were found to be security departments or facility departments. Also, only one person in the execute department and only one person in the administration department do not use *Deli*.

4.2 Describe up to five of the most interesting patterns that appear in the building data. Describe what is notable about the pattern and explain its possible significance.

4.2.1 General Information of General Building Data

First of section 4.2, we should find about the distribution of general building data. With TSNE technique, we can draw the TSNE plot of general building data as figure 13.

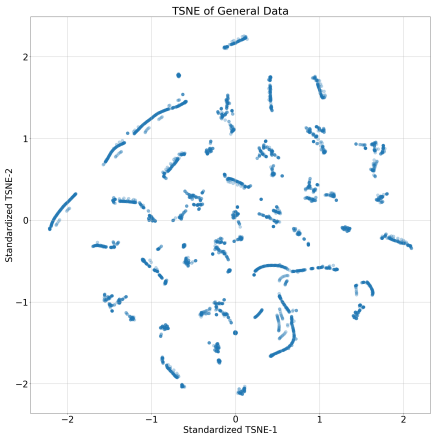


Figure 13: TSNE for General Building Data

4.2.2 Workflow

Figure 14: Workflow for Question 2

4.2.3 Correlation within General Building Data

We made the correlation heatmap within the general building data to find two columns which have strong positive or negative correlation. The correlation heatmap is as figure 15. Moreover, the R-value distribution with the data which are used in figure 15 is as shown as figure 16.

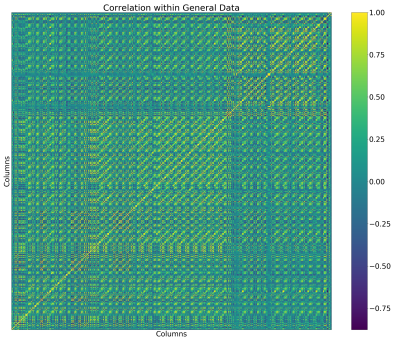


Figure 15: Correlation Heatmap within General Building Data

The basic statistics of these R-values are in table 5.

Furthermore, the extreme values of R-values are in tables 6 and 7.

As table 6, no combination of columns make R-value to -1; but, there are many combinations of columns make R-value 1. In other words, many columns have strong positive correlation with others rather than negative

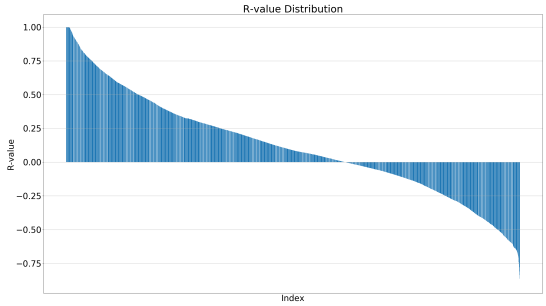


Figure 16: R-value Distribution within General Building Data

Table 5: Basic Statistics of R-Values

Item	Minimum	Maximum	Mean	q1	Median	q3	Standard Deviation
Value	-0.88	1.0	0.11	-0.12	0.08	0.34	0.37

Table 6: Minimum Values of R-Values

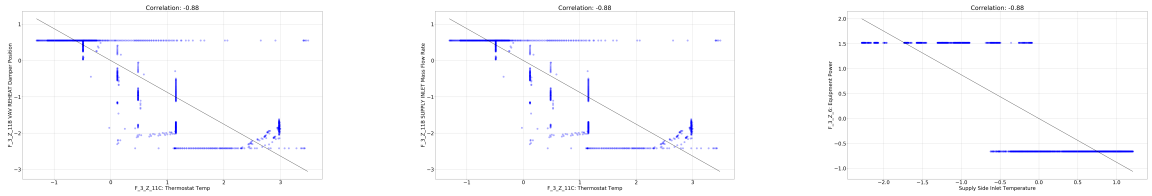
Column1	Column2	R-Value
F_3_Z_11C: Thermostat Temp	F_3_Z_11B VAV REHEAT Damper Position	-0.877076318455652
F_3_Z_11C: Thermostat Temp	F_3_Z_11B SUPPLY INLET Mass Flow Rate	-0.8770746421685551
Supply Side Inlet Temperature	F_3_Z_6: Equipment Power	-0.8764301266869379
Supply Side Inlet Temperature	F_3_Z_6: Lights Power	-0.8764301266869379
Supply Side Inlet Temperature	F_1_Z_4: Lights Power	-0.8743564507417133

Table 7: Maximum Values of R-Values

Column1	Column2	R-Value
Water Heater Tank Temperature	Supply Side Outlet Temperature	1.0
F_3_Z_7: Thermostat Heating Setpoint	F_3_Z_3: Thermostat Heating Setpoint	1.0
F_3_Z_7: Thermostat Heating Setpoint	F_3_Z_2: Thermostat Heating Setpoint	1.0
F_3_Z_7: Thermostat Heating Setpoint	F_3_Z_11B: Thermostat Heating Setpoint	1.0
F_3_Z_7: Thermostat Heating Setpoint	F_3_Z_11A: Thermostat Heating Setpoint	1.0
F_3_Z_7: Thermostat Heating Setpoint	F_3_Z_10: Thermostat Heating Setpoint	1.0
F_3_Z_7: Thermostat Cooling Setpoint	F_3_Z_3: Thermostat Cooling Setpoint	1.0
F_3_Z_7: Thermostat Cooling Setpoint	F_3_Z_2: Thermostat Cooling Setpoint	1.0
F_3_Z_7: Thermostat Cooling Setpoint	F_3_Z_11B: Thermostat Cooling Setpoint	1.0
F_3_Z_7: Thermostat Cooling Setpoint	F_3_Z_11A: Thermostat Cooling Setpoint	1.0
F_3_Z_7: Thermostat Cooling Setpoint	F_3_Z_10: Thermostat Cooling Setpoint	1.0
F_3_Z_6: Thermostat Heating Setpoint	F_3_Z_5: Thermostat Heating Setpoint	1.0
F_3_Z_6: Thermostat Cooling Setpoint	F_3_Z_5: Thermostat Cooling Setpoint	1.0
F_3_Z_6: Lights Power	F_3_Z_6: Equipment Power	1.0
F_3_Z_5: Lights Power	F_3_Z_5: Equipment Power	1.0
F_3_Z_3: Thermostat Heating Setpoint	F_3_Z_2: Thermostat Heating Setpoint	1.0
F_3_Z_3: Thermostat Heating Setpoint	F_3_Z_11B: Thermostat Heating Setpoint	1.0
F_3_Z_3: Thermostat Heating Setpoint	F_3_Z_11A: Thermostat Heating Setpoint	1.0
F_3_Z_3: Thermostat Heating Setpoint	F_3_Z_10: Thermostat Heating Setpoint	1.0
F_3_Z_3: Thermostat Cooling Setpoint	F_3_Z_2: Thermostat Cooling Setpoint	1.0
F_3_Z_3: Thermostat Cooling Setpoint	F_3_Z_11B: Thermostat Cooling Setpoint	1.0
F_3_Z_3: Thermostat Cooling Setpoint	F_3_Z_11A: Thermostat Cooling Setpoint	1.0
F_3_Z_3: Thermostat Cooling Setpoint	F_3_Z_10: Thermostat Cooling Setpoint	1.0
F_3_Z_3: Lights Power	F_3_Z_3: Equipment Power	1.0
(omitted...)	(omitted...)	(omitted...)

correlation. However, in table 7, most of combination are (Thermostat Heating Setpoint), (Thermostat Cooling Setpoint), or (Lights Power & Equipment power). Also, not all Thermostat Setpoints in same floor have strong correlation; but, Thermostat Setpoints in different floor do not have strong correlation.

Furthermore, we draw plot between three general building data which have the lowest R-values as figure 17. In the figure 17, the R-values are about 0.88, so the R-squared values (R^2) will be about 0.77. Therefore, in figure 17, we can argue that they have negative correlation, but it is not *strong* negative correlation.

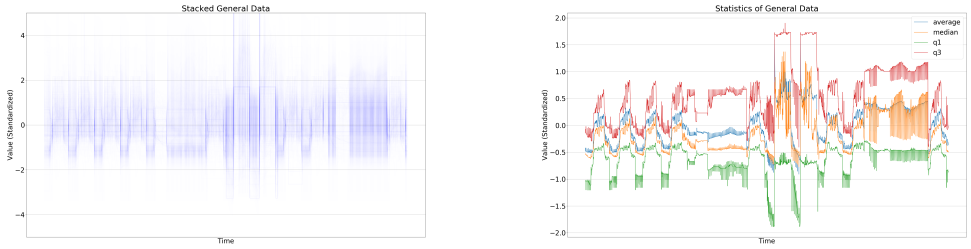


(a) First (b) Second (c) Third

Figure 17: Plots of the Lowest R-values

4.2.4 Plots of General Building Data

In the figure 18, there are two plots: Figure 18-(a) is stacked plot about all columns, and figure 18-(b) has basic statistics plot of general building data. In figure 18-(b), we can see that the pattern of general building data is changing by quarters: in first-quarters, the cyclic pattern is shown; in second-quarters, rapidly increasing can be detected; in third-quarters, many general building data jump and hold a while; and, in fourth-quarters, general building data are suddenly decreasing.



(a) Stacked (b) Statistics

Figure 18: Plots of General Building Data

In the first-quarter, we find that the general building data make a cycle every *288 indices* as figure 19. The general building data are reported on every 5 minutes, so the general building data make a cycle every *1440 minutes* or every *24 hours* or every *1 day*.

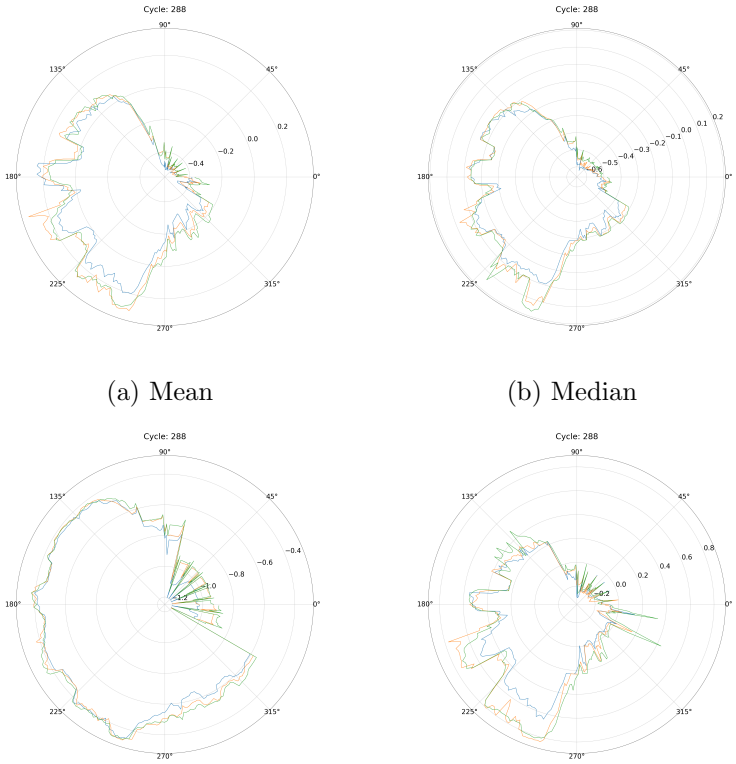
In the second-quarter, we calculated the peak in every columns in general building data. With figure 20-(a), the distribution of peak width is shown; and, with figure 20-(b), you can know that where is peak in general building data. Also, table 8 displays basic statistics value of peak width. With the data in figure 20 and table 8, we can know that the general building data are suddenly increasing about *15 indices*; in other words, the general building data has steeps every *75 minutes*.

Table 8: Basic Statistics Data with the peak in the Second-quarter

Item	Minimum	Maximum	Mean	q1	Median	q3	Standard Deviation
Value	1.0	808.45	15.98	1.51	3.71	12.0	55.36

As figure 19, we drew the cyclic plots of the third-quarters of general building data as figure 21. As the first-quarters, the period is *288 indices*, *24 hours*, or *1 day*. Therefore, there are two days which has increased general building data.

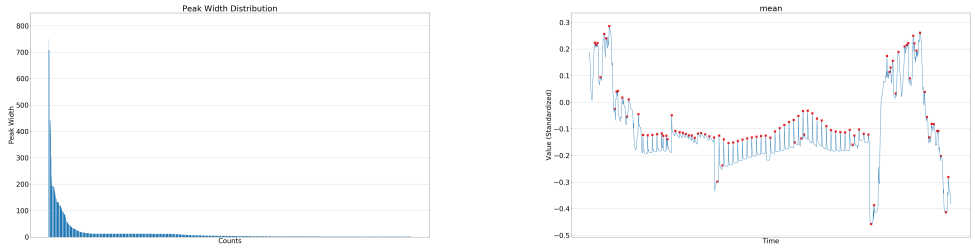
In the fourth-quarter, we calculated the under-peak in every columns in general building data. As figure 20, with figure 22-(a), the distribution of under-peak width is displayed; and, with figure 22-(b), you can know that where is under-peak in general building data. Moreover, table 9 shows basic statistics value of under-peak width. Amongst the data in figure 22 and table 9, we can know that the general building data are rapidly decreasing about *16 indices*; in other words, the general building data has under-steeps every *80 minutes*.



(a) Mean (b) Median

(c) q1 (d) q3

Figure 19: Cyclic Plots on First-quarters

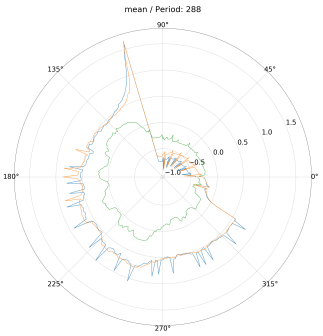


(a) Peak Width Distribution (b) Peak in Mean Column

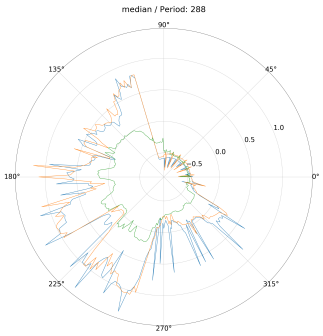
Figure 20: Peak in the Second-quarter

Table 9: Basic Statistics Data with the under-peak in the Fourth-quarter

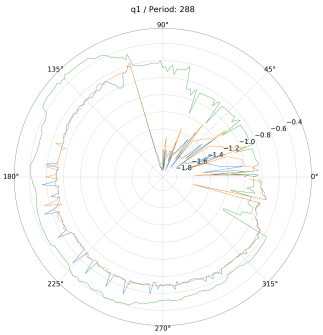
Item	Minimum	Maximum	Mean	q1	Median	q3	Standard Deviation
Value	1.0	850.54	16.36	1.11	2.77	11.68	69.86



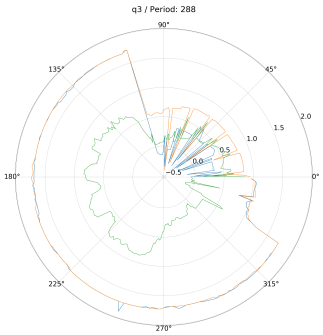
(a) Mean



(b) Median

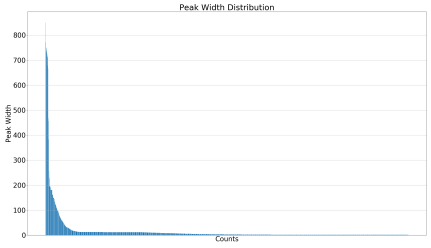


(c) q1

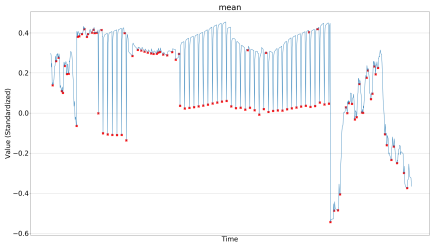


(d) q3

Figure 21: Cyclic Plots on Third-quarters



(a) Under-peak Width Distribution



(b) Under-peak in Mean Column

Figure 22: Under-peak in the Fourth-quarter

4.2.5 Interesting Patterns

According to these fact, we can know followings:

1. *Thermostat Setpoint* is controlled by floor. Someone can adjust specific zone in floor, but the default value is sets on floor. (Table 7)
2. Some zone only have *lights* for power consumption. (Table 7)
3. No *strong* negative correlation exist. (Figure 17)
4. In the first-quarter, the general building data make a cycle everyday. (Figure 19)
5. In the second-quarter, the general building data are suddenly increasing about every *75 minutes*. (Table 8 and Figure 20)
6. There are two days which are increasing whole usage of the general building data. (Figure 21)
7. In the fourth-quarter, the general building data are dramatically decreasing about every *80 minutes*. (Table 9 and Figure 22)

4.3 Describe up to five notable anomalies or unusual events you see in the data. Prioritize those issue that are most likely to represent a danger or a serious issue for building operations.

4.3.1 General Information of Hazium Concentration

In the question 2 or section 4.2, we need to find a danger or a serious issue for building data. Hence, we suppose that a danger will be related with Hazium concentration. In other words, if there are some danger in building operation, then the Hazium data will be increased along.

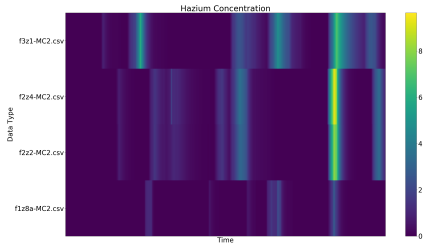


Figure 23: Hazium Data from Different Data Sources

In the figure 23, we can see Hazium concentration of many sources.

4.3.2 Workflow

Figure 24: Workflow for Question 3

4.3.3 Abnormality in General Building Data

To find patterns which appear in the building data, we should find that normality/abnormality in the building data. However, there are over 400 columns in the general building data; therefore, it is almost impossible to find abnormality column-by-column by human. Hence, we used these four algorithms which are included in scikit-learn: *EllipticEnvelope* [8], *OneClassSVM*, *IsolationForest* [9, 10], and *LocalOutlierFactor* [11].

Moreover, we can display the timeline of abnormality as figure 25. In the figure 25-(a), we can know that which algorithm consider specific time as abnormal events (yellow marked is abnormal); and, in the figure 25-(b), we can realize that how many algorithms consider specific time as abnormal events.

(a) (b)

Figure 25: Abnormality in General Building Data by Timeline

4.3.4 Score of Classification of Abnormality

To decide the best algorithm for find abnormality, we use these five classifier algorithms for scoring: *KNeighbor*, *SVC* [12, 13], *DecisionTree* [14, 15, 16], *RandomForest* [16], and *AdaBoost* [17, 18]. The scores are in table 10; the highest score is *0.9938* on *LocalOutlier* algorithm. Therefore, we choose *LocalOutlier* algorithm for finding abnormality. Note that the train data and the test data are randomly selected with $0.8 : 0.2$ ratio, and the seed is specified for repeated result.

Table 10: Scores of Classification of Abnormality

	Elliptic	OneClassSVM	IsolationForest	LocalOutlier
KNeighbor	0.995	0.991	0.929	0.988
SVC	0.994	0.993	0.919	0.985
DecisionTree	0.988	0.989	0.948	0.999
RandomForest	0.993	0.995	0.942	0.999
AdaBoost	0.986	0.985	0.937	0.999
Mean	0.9911	0.9906	0.9351	0.9938

4.3.5 Column for Abnormality

Now, we know that when is abnormal in general building data, and the general building data *per se*. Therefore, we can calculate the difference of mean value for each columns in general building data. In other words, we can measure that which column in general building data most differ between normal and abnormal timing.

The distribution of differences between mean values for whole columns of general building data is figure 26; also, the basic statistics values of differences between mean values are in table 11.

Table 11: Basic Statistics Data with the Differences of Mean Values

Item	Minimum	Maximum	Mean	q1	Median	q3	Standard Deviation
Value	0.00058	1.05075	0.159379	0.05347	0.1097855	0.1756	0.171869

4.3.6 Danger for Building Operation

Figure 27 display two heatmap about R-values: figure 27-(a) is consist with the data which marked as *normal*; and, figure 27-(b) contains the data which marked as *abnormal*. However, as shown as figure 27, the R-values are not significant; in other words, R-squared values are too small to give confidence for correlation. Therefore, we divided the abnormal data with comparing the general building data and Hazium data directly. As we mentioned herein-above, all of values are standardized, so we can compare two values one-by-one.

Figure 28 contains two heatmap plot of divided abnormal data. Figure 28-(a) is drawn with the data which has higher general building data than Hazium data; and also, figure 28-(b) is written with the data which has lower general building data. With comparing figure 27 and figure 28, the R-values in 28 is bigger than the R-values in 28. In other words, in the (General Building Data)–(Hazium Data) plot, the scatter data are aligned near the X-axis and Y-axis.

With the data in figure 28, we can select the column which has the highest sum of R-values. In this data, combination of the column (*F_2_Z_16 SUPPLY INLET Temperature*) and the data set (*f2z2-MC2.csv*) has the highest value. Figure 29 is written with this combination. As we mentioned herein-above, many points are along the X-axis or the Y-axis.

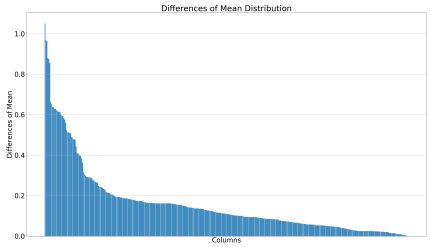
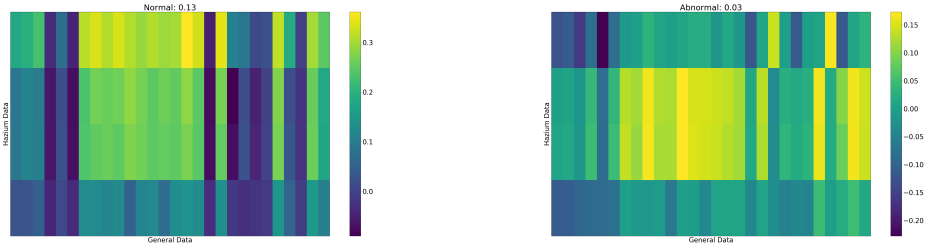
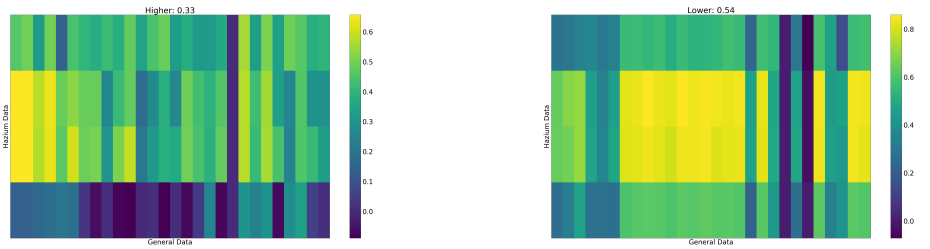


Figure 26: Distribution of Differences between Mean Value



(a) Normal (b) Abnormal

Figure 27: Heatmap of R-values between General Building Data and Hazium Data



(a) Higher (b) Lower

Figure 28: Heatmap of R-values with Higher and Lower General Building Data

∴ Hence, if the data (*F_2_Z_16 SUPPLY INLET Temperature*) is decreasing, then the Hazium concentration of Zone 2 on Floor 2 will be increasing; and *vice versa*. Many other combinations exist; but, we only display this combination which has the highest R-values.

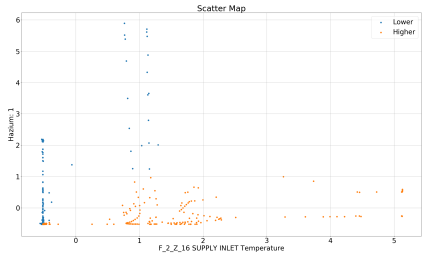


Figure 29: Scatter Plot between General Building Data and Hazium Data

4.4 Describe up to three observed relationships between the proximity card data and building data elements. If you find a causal relationship, describe your discovered cause and effect, the evidence you found the support it, and your level of confidence in your assessment of the relationship.

4.4.1 Frequency of prox Data

The general building data has been reported every five minutes; however, prox data is a kind of movement data, so prox data has no cycle pattern at all. Therefore, we should gather and count prox data for comparing with general building data. In this manner, table 12 and figure 30 have the distribution information about frequence of prox data.

Table 12: Basic Statistics Data with the Frequency of prox Data

Item	Minimum	Maximum	Mean	q1	Median	q3	Standard Deviation
Value	0	141	8.014	0	0	10	15.35

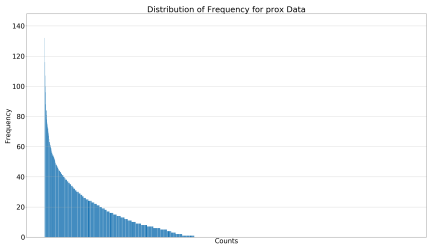


Figure 30: Distribution with Frequency of prox Data

Moreover, with figure 31, there are some peaks in the prox data. Also, the five peaks in the row, we can have reasonable doubt about weekday and weekend.

With the reasonable doubt from figure 31, we can draw the cycle plot with daily as figure 32. With figure 32, we can know what the employees look like such as their attendance time.

4.4.2 Workflow

4.4.3 Abnormality of General Building Data and prox Data

As question 3 or section 4.3, we ran the abnormality finding algorithms of the merged data amongst the general building data and prox frequency data. Figure 34 shows the abnormality of prox data which calculated by four algorithms.

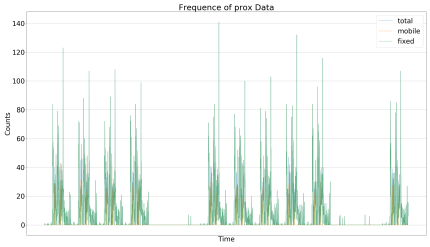
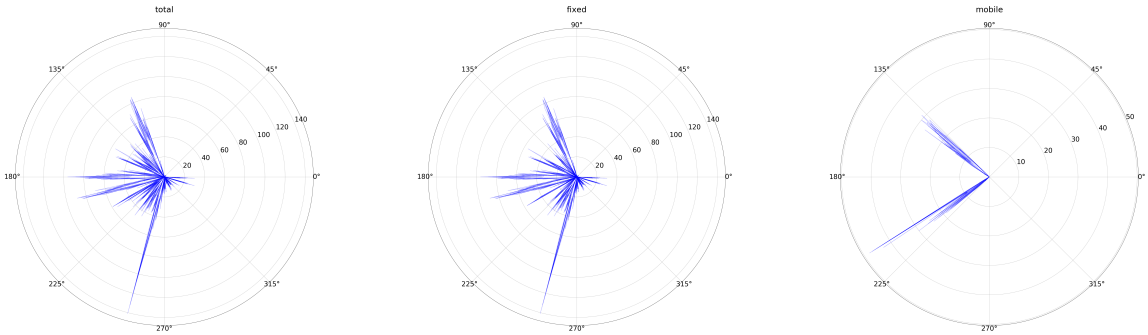


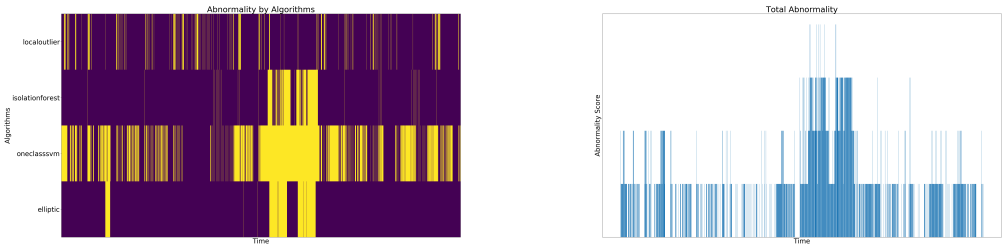
Figure 31: Timeline with Frequency of prox Data



(a) Total data (b) Fixed data (c) Mobile data

Figure 32: Daily Cycle of prox Data

Figure 33: Workflow for Question 4



(a) (b)

Figure 34: Abnormality in prox Data by Timeline

4.4.4 Correlation between General Building Data and prox Data

4.4.5 Cause and Effect for the Correlation

5 Discussion

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[2] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science & engineering*, vol. 9, no. 3, p. 90, 2007.

[3] W. McKinney, "pandas: a foundational python library for data analysis and statistics," *Python for High Performance and Scientific Computing*, vol. 14, 2011.

[4] E. Jones, T. Oliphant, P. Peterson, *et al.*, "Scipy: Open source scientific tools for python," 2001.

[5] A. Clark, "Pillow (pil fork) documentation," 2015.

[6] C. Sievert, C. Parmer, T. Hocking, S. Chamberlain, K. Ram, M. Corvellec, and P. Despouy, "plotly: Create interactive web graphics via 'plotly. js'," *R package version*, vol. 4, no. 1, p. 110, 2017.

[7] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[8] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[9] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, IEEE, 2008.

[10] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, p. 3, 2012.

[11] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, pp. 93–104, ACM, 2000.

[12] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[13] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[14] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[16] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[18] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.