

# Visualization Term Project

20141087 Ryeongyang Kim      20161206 Jaewoong Lee

December 1, 2019

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Materials</b>	<b>4</b>
2.1	Building Layout . . . . .	4
2.1.1	Main Layout . . . . .	4
2.1.2	Energy Zone Layout . . . . .	4
2.1.3	Prox Zone Layout . . . . .	4
<b>3</b>	<b>Methods</b>	<b>4</b>
3.1	Python Packages . . . . .	4
3.1.1	Scikit-learn: Machine Learning in Python . . . . .	4
3.1.2	Matplotlib . . . . .	5
3.1.3	Pandas . . . . .	5
3.1.4	SciPy . . . . .	5
3.2	TSNE . . . . .	5
3.3	Standardization . . . . .	5
<b>4</b>	<b>Results</b>	<b>5</b>
4.1	Question 1 . . . . .	5
4.1.1	General Information of prox Data . . . . .	5
4.1.2	Workflow . . . . .	6
4.1.3	Movement Direction and Distance . . . . .	6
4.1.4	Department Distribution . . . . .	6
4.1.5	Typical Patterns in prox Data . . . . .	8
4.1.6	Typical Day Look for GAStech employees . . . . .	8
4.2	Question 2 . . . . .	8
4.2.1	General Information of General Building Data . . . . .	8
4.2.2	Workflow . . . . .	10
4.2.3	Correlation within General Building Data . . . . .	10
4.2.4	Plots of General Building Data . . . . .	10
4.2.5	Interesting Patterns . . . . .	10
4.3	Question 3 . . . . .	12
4.3.1	General Information of Hazium Concentration . . . . .	12
4.3.2	Workflow . . . . .	12
4.3.3	Find Abnormality in General Building Data . . . . .	12
4.3.4	Abnormality of Hazium Data . . . . .	12
4.3.5	Abnormality of General Building Data and Hazium Data . . . . .	13
4.3.6	Score of Classification of Abnormality . . . . .	13
4.3.7	Danger for Building Operation . . . . .	13
4.4	Question 4 . . . . .	13
4.4.1	General Information of Moving Average for General Building Data . . . . .	13
4.4.2	Workflow . . . . .	13
4.4.3	Frequency of prox Data . . . . .	13
4.4.4	Correlation between General Building Data and prox Data . . . . .	13
4.4.5	Cause and Effect for the Correlation . . . . .	13
<b>5</b>	<b>Discussion</b>	<b>13</b>

## List of Tables

1	Basic Statistics Data within Movement Distance . . . . .	5
2	Minimum Moving Employees . . . . .	5
3	Maximum Moving Employees . . . . .	6
4	Department Information by Quartiles . . . . .	9
5	Basic Statistics of R-Values . . . . .	10
6	Minimum Values of R-Values . . . . .	11
7	Maximum Values of R-Values . . . . .	11

## List of Figures

1	Main Layout of the building . . . . .	4
2	Energy Zone of the Building . . . . .	4
3	Prox zone of the Building . . . . .	5
4	Distribution of Movement Distance . . . . .	6
5	Workflow for Question 1 . . . . .	6
6	Movement Direction/Distance on First Floor . . . . .	7
7	Movement Direction/Distance on Second Floor . . . . .	7
8	Movement Direction/Distance on Third Floor . . . . .	8
9	Pie Plot of Department Distribution by Quartiles . . . . .	8
10	TSNE for General Building Data . . . . .	9
11	Workflow for Question 2 . . . . .	9
12	Correlation Heatmap within General Building Data . . . . .	10
13	R-value Distribution within General Building Data . . . . .	10
14	Plot of the Lowest R-values . . . . .	11
15	Plots of General Building Data . . . . .	11
16	Hazium Data from Different Data Sources . . . . .	12
17	Workflow for Question 3 . . . . .	12
18	Abnormality in General Building Data by Timeline . . . . .	12
19	Abnormality in Hazium Data by Timeline . . . . .	13
20	Abnormality in General Building Data and Hazium Data . . . . .	13
21	Workflow for Question 4 . . . . .	13

# 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

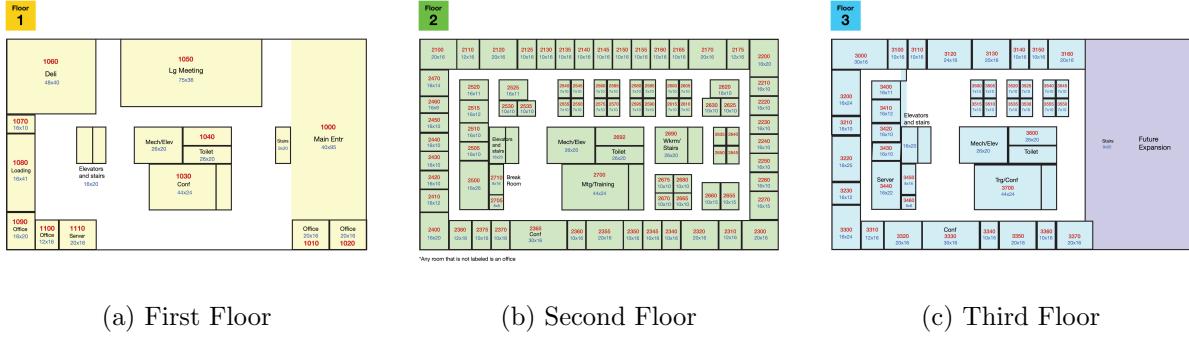


Figure 1: Main Layout of the building

The main layout of this building is as figure 1.

#### 2.1.2 Energy Zone Layout

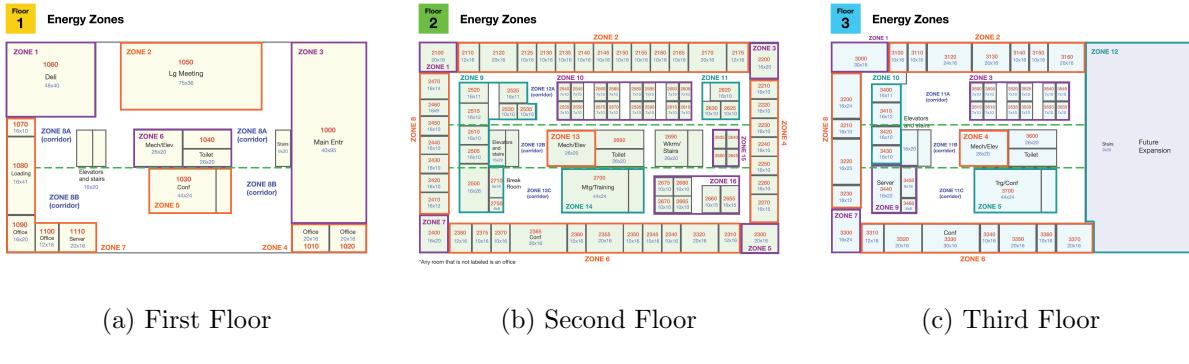


Figure 2: Energy Zone of the Building

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].

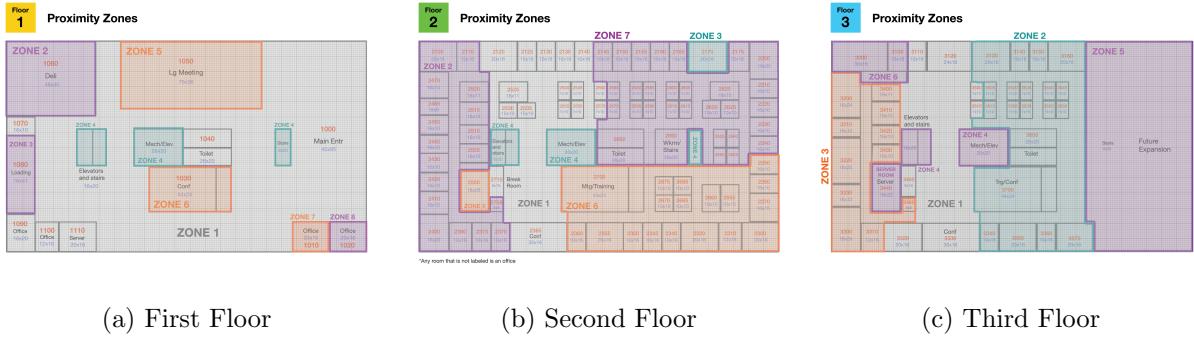


Figure 3: Prox zone of the Building

### 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].

### 3.2 TSNE

T-distributed Stochastic Neighbor Embedding (TSNE) is a machine learning algorithm for visualization high-dimensional data in a low-dimensional space [5].

### 3.3 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.

Table 1: Basic Statistics Data within Movement Distance

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

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

Table 2: Minimum Moving Employees

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

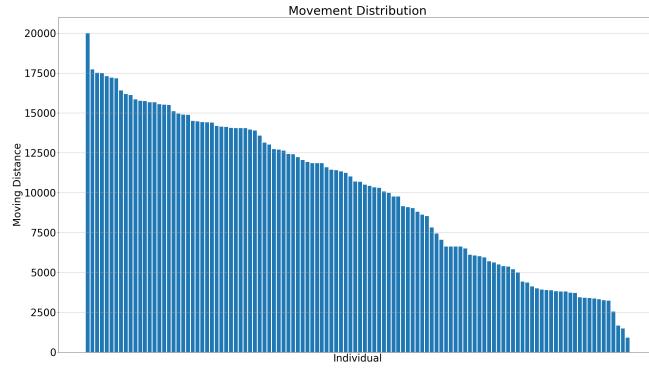


Figure 4: Distribution of Movement Distance

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.

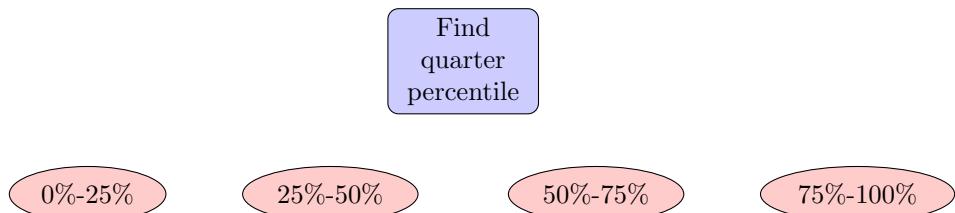


Figure 5: Workflow for Question 1

#### 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.

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
5. HR
6. Information Technology
7. Security

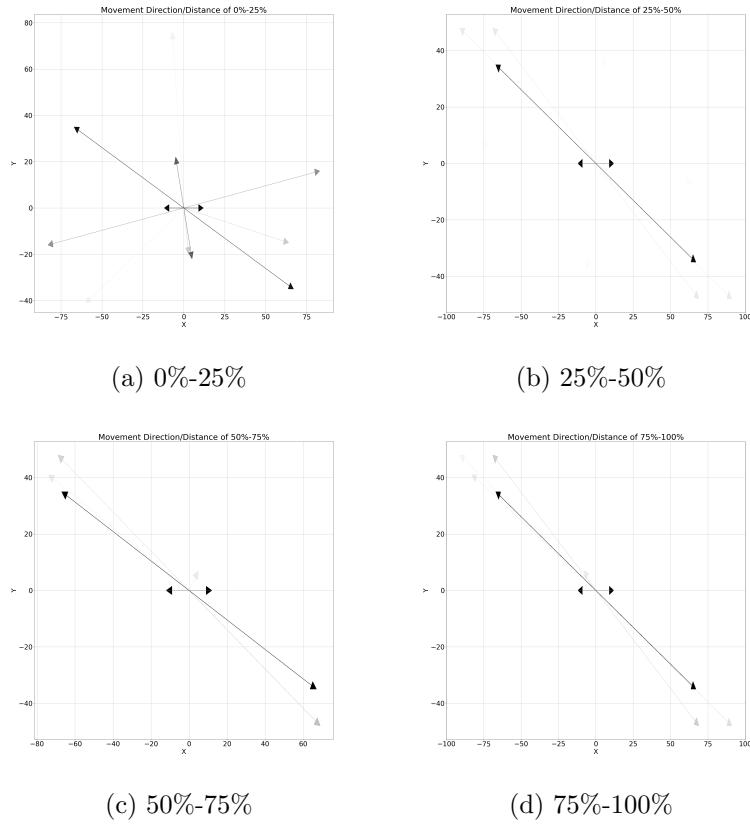


Figure 6: Movement Direction/Distance on First Floor

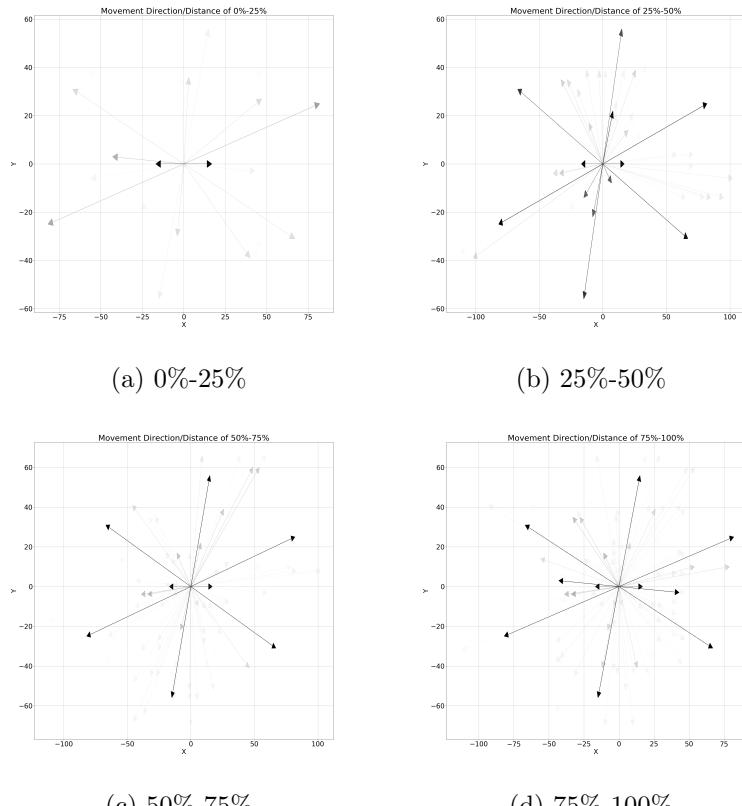


Figure 7: Movement Direction/Distance on Second Floor

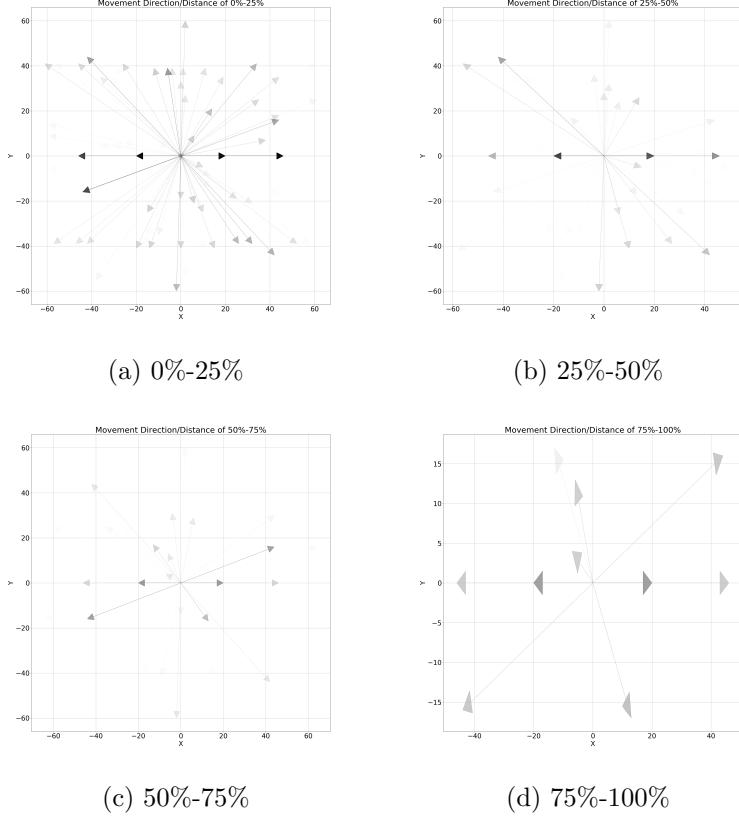


Figure 8: Movement Direction/Distance on Third Floor

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

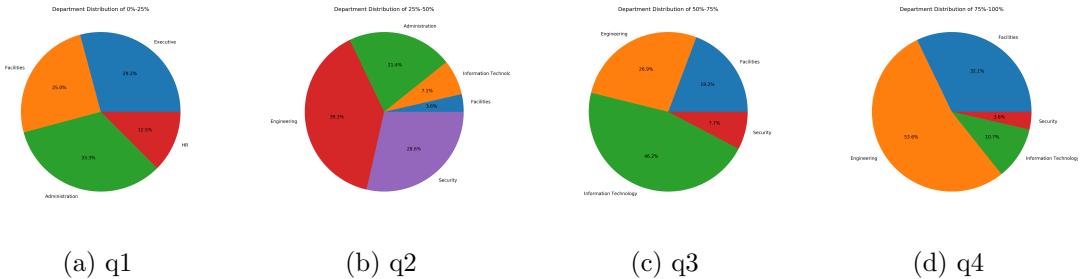


Figure 9: Pie Plot of Department Distribution by Quartiles

#### 4.1.5 Typical Patterns in prox Data

#### 4.1.6 Typical Day Look for GASTech employees

**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 10.

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
q3	Facilities	1
	Information Technology	12
	Engineering	7
	Facilities	5
q4 (Maximum 25%)	Security	2
	Engineering	15
	Facilities	9
	Information Technology	3
	Security	1

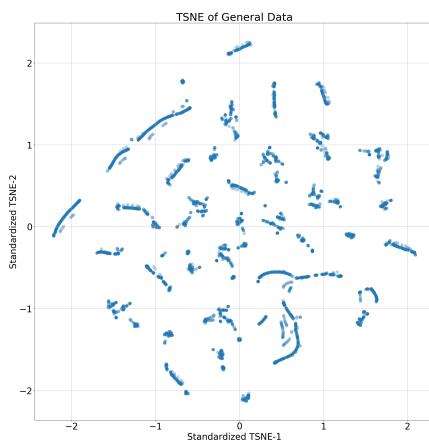


Figure 10: TSNE for General Building Data

Figure 11: Workflow for Question 2

#### 4.2.2 Workflow

#### 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 12. Moreover, the R-value distribution with the data which are used in figure 12 is as shown as figure 13.

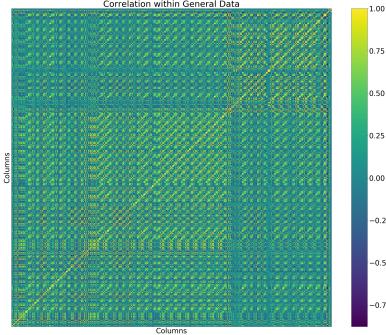


Figure 12: Correlation Heatmap within General Building Data

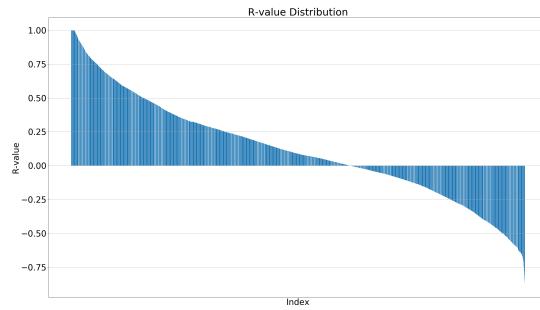


Figure 13: R-value Distribution within General Building Data

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

Table 5: Basic Statistics of R-Values

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

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 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 14. In the figure 14, the R-values are about 0.88, so the R-squared values ( $R^2$ ) will be about 0.77. Therefore, in figure 14, we can argue that they have negative correlation, but it is not *strong* negative correlation.

#### 4.2.4 Plots of General Building Data

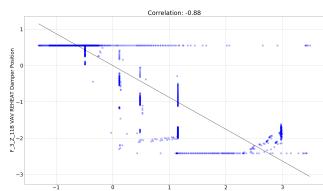
In the figure 15, there are two plots: Figure 15-(a) is stacked plot about all columns, and figure 15-(b) has basic statistics plot of general building data.

Table 6: Minimum Values of R-Values

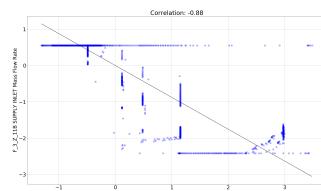
Column1	Column2	R-Value
F_3_Z_11C: Thermostat Temp	F_3_Z_11B VAV REHEAT Damper Position	-0.877076318455652
F_3_Z_11C: Thermostat Temp	F_3_Z_11B SUPPLY INLET Mass Flow Rate	-0.8770746421685551
Supply Side Inlet Temperature	F_3_Z_6: Equipment Power	-0.8764301266869379
Supply Side Inlet Temperature	F_3_Z_6: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

Table 7: Maximum Values of R-Values

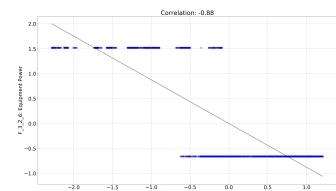
Column1	Column2	R-Value
Water Heater Tank Temperature	Supply Side Outlet Temperature	1.0
F_3_Z_7: Thermostat Heating Setpoint	F_3_Z_3: Thermostat Heating Setpoint	1.0
F_3_Z_7: Thermostat Heating Setpoint	F_3_Z_2: Thermostat Heating Setpoint	1.0
F_3_Z_7: Thermostat Heating Setpoint	F_3_Z_11B: Thermostat 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...)



(a) First



(b) Second

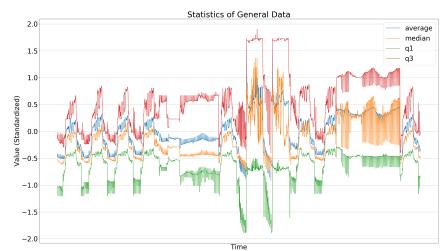


(c) Third

Figure 14: Plot of the Lowest R-values



(a) Stacked



(b) Statistics

Figure 15: Plots of General Building Data

#### 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)

### 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.

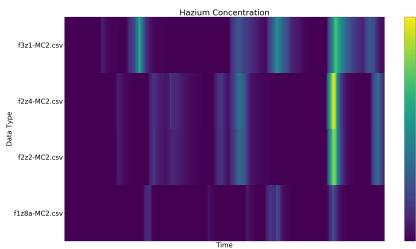


Figure 16: Hazium Data from Different Data Sources

In the figure 16, we can see Hazium concentration of many sources.

#### 4.3.2 Workflow

Figure 17: Workflow for Question 3

#### 4.3.3 Find 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* [6], *OneClassSVM*, *IsolationForest* [7, 8], and *LocalOutlierFactor* [9].

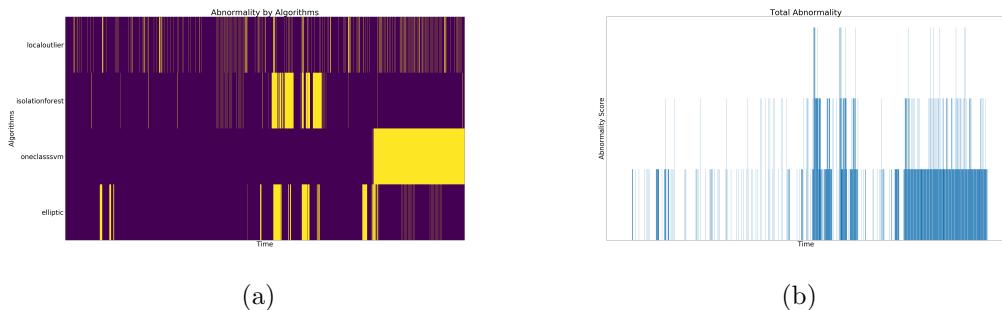


Figure 18: Abnormality in General Building Data by Timeline

Moreover, we can display the timeline of abnormality as figure 18. In the figure 18-(a), we can know that which algorithm consider specific time as abnormal events (yellow marked is abnormal); and, in the figure 18-(b), we can realize that how many algorithms consider specific time as abnormal events.

#### 4.3.4 Abnormality of Hazium Data

As figure 18, we draw figure 19 with four algorithms which mentioned hereinbefore.

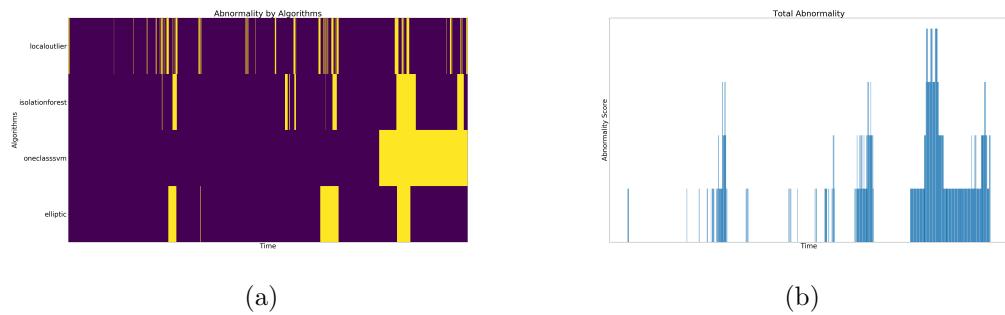


Figure 19: Abnormality in Hazium Data by Timeline

#### 4.3.5 Abnormality of General Building Data and Hazium Data

In figure 20, you can see the abnormality both in general building data and Hazium data. Green means single abnormality; and yellow means abnormal in both.

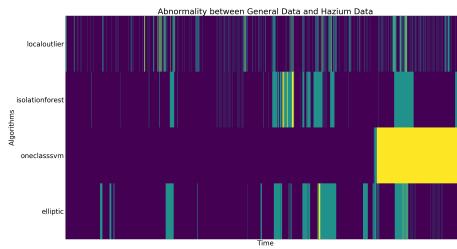


Figure 20: Abnormality in General Building Data and Hazium Data

#### 4.3.6 Score of Classification of Abnormality

#### 4.3.7 Danger for Building Operation

**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 General Information of Moving Average for General Building Data

#### 4.4.2 Workflow

Figure 21: Workflow for Question 4

#### 4.4.3 Frequency of prox Data

#### 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] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.