Visualization Term Project

20141087 Ryeongyun Kim 20161206 Jaewoong Lee

November 25, 2019

Contents

List of Tables

List of Figures

1 Introduction

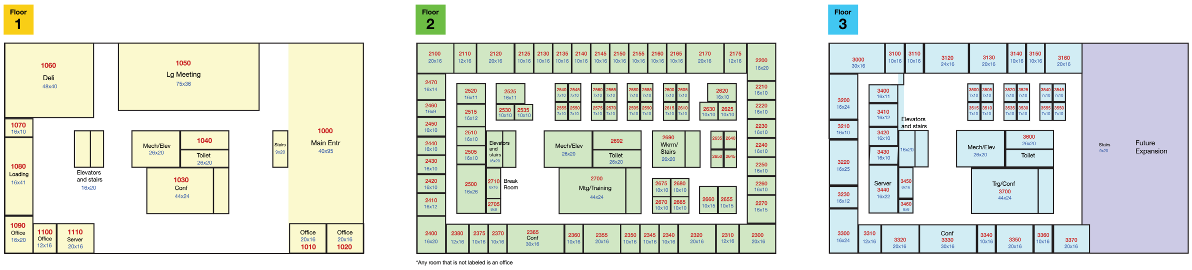
In this term project, we have to answer several question with virtual building data.

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

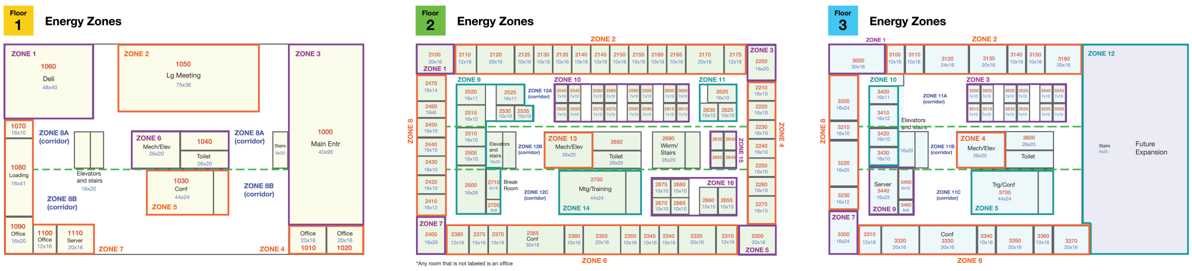


(a) First Floor (b) Second Floor (c) Third Floor

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.

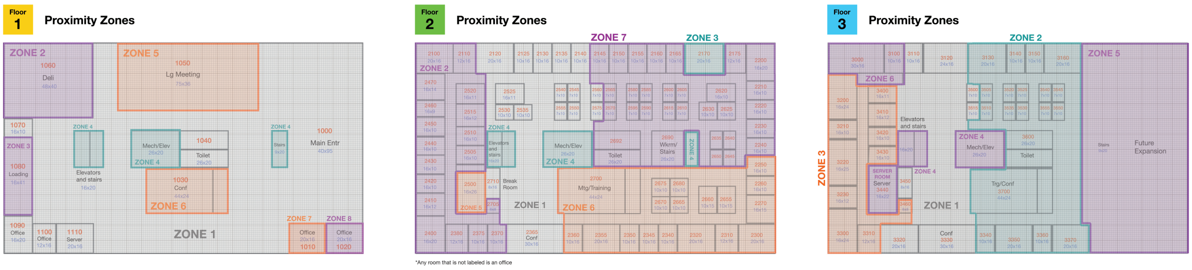
3 Methods

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].

3.2 TSNE

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

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.

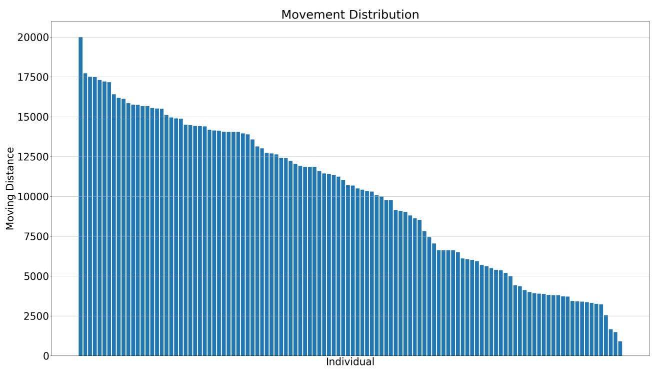


Figure 4: Distribution of Movement Distance

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

4.1.2 Workflow

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

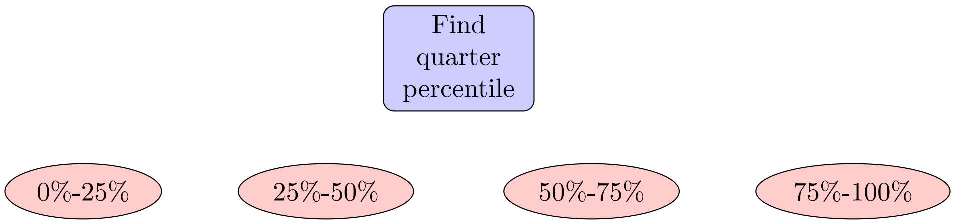
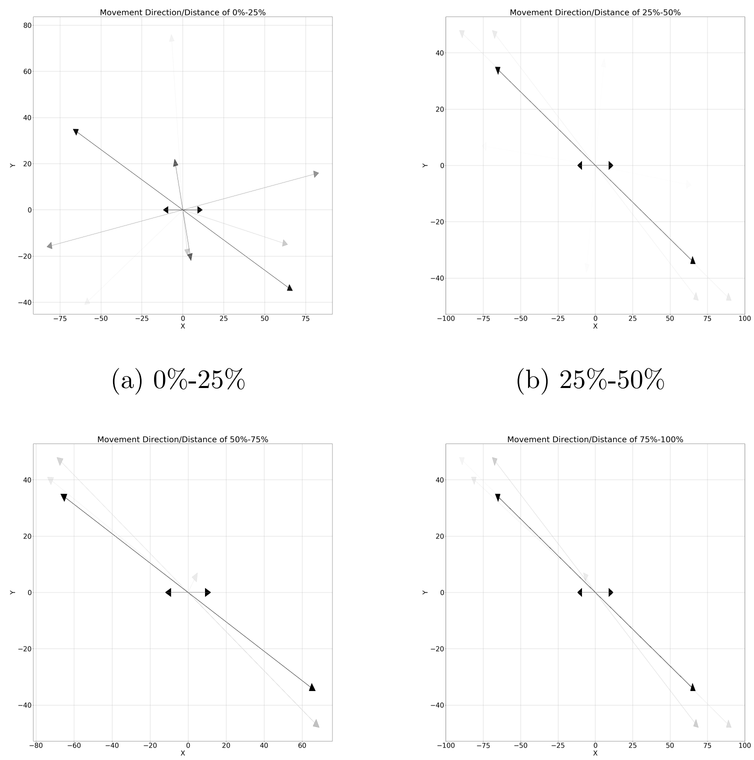


Figure 5: Workflow for Question 1

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%

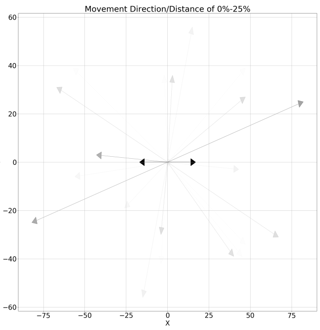
(b) 25%-50%

(c) 50%-75%

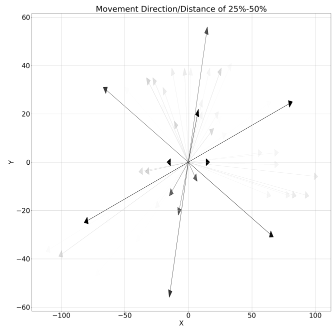
(d) 75%-100%

Figure 6: Movement Direction/Distance on First Floor

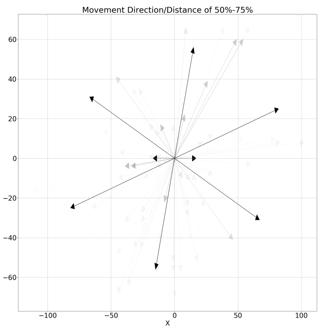
The movement direction and distance on the first floor is shows as figure 6. In figure 6-(b, c, d), you can see two arrows: one is left-upward arrow, the other is right-downward arrow.



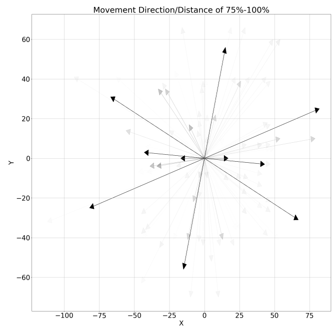
(a) 0%-25%



(b) 25%-50%

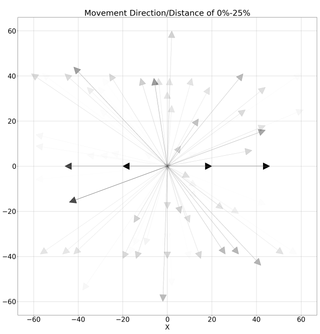


(c) 50%-75%

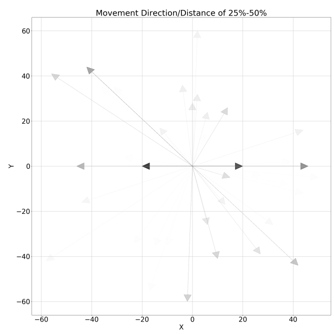


(d) 75%-100%

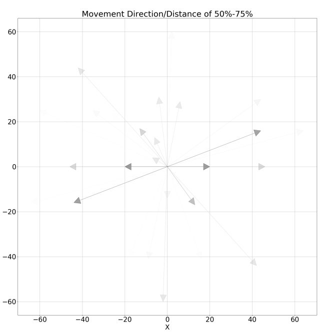
Figure 7: Movement Direction/Distance on Second Floor



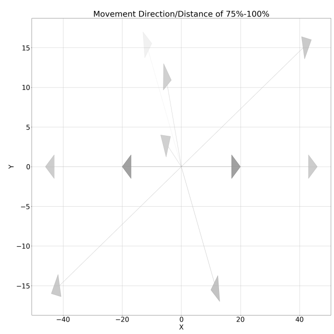
(a) 0%-25%



(b) 25%-50%



(c) 50%-75%



(d) 75%-100%

Figure 8: Movement Direction/Distance on Third Floor

4.1.4 Department Distribution

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

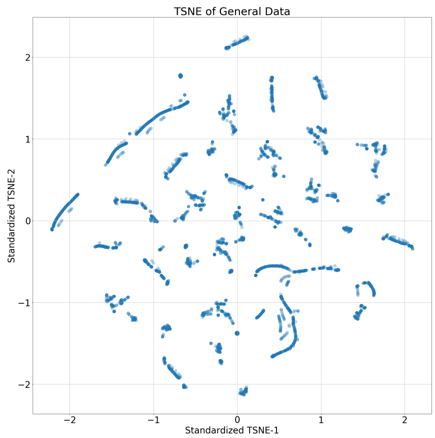


Figure 9: TSNE for General Building Data

4.2.2 Workflow

Figure 10: 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 11. Moreover, the R-value distribution with the data which are used in figure 11 is as shown as figure 12.

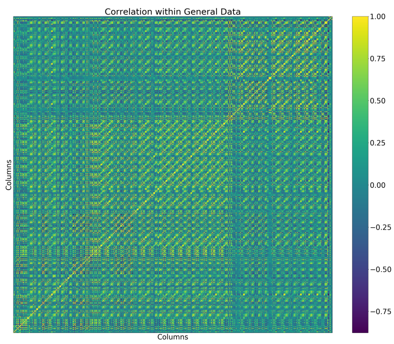


Figure 11: Correlation Heatmap within General Building Data

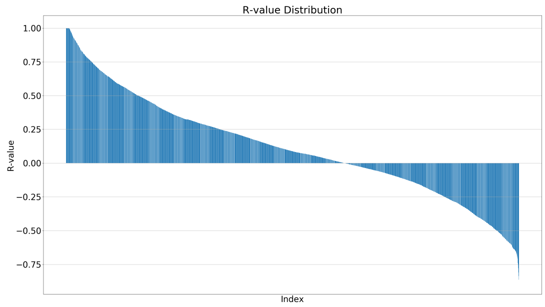


Figure 12: R-value Distribution within General Building Data

The basic statistics of these R-values are in table 2.

Table 2: 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	0.36

4.2.4 Plots of General Building Data

4.2.5 Interesting Patterns

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.

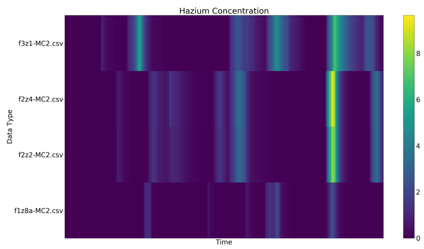


Figure 13: Hazium Data from Different Data Sources

In the figure 13, we can see Hazium concentration of many sources.

4.3.2 Workflow

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].

Moreover, with the data in figure **??**, we can display the timeline of abnormality as figure 15.

In the figure 15-(a), we can know that which algorithm consider specific time as abnormal events (yellow marked is abnormal); and, in the figure 15-(b), we can realize that how many algorithms consider specific time as abnormal events.

Figure 14: Workflow for Question 3

(a) (b)

Figure 15: Abnormality by Timeline

4.3.4 Abnormality of Hazium Data

4.3.5 Correlation between General Building Data and Hazium Data

4.3.6 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 16: 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.