Behavior Prediction and Safety Control of Clients Based On WiFi Data

Team Member:

Name: Ruichao Ma (Team Leader)
Address: 3150 Wilshire Blvd, Los Angeles, 90010
Email: ruichaom@usc.edu

Name: Chuqi Liu
Address: 158 N Mariposa Ave, Los Angeles 90004
Email: chuqiliu@usc.edu

Name: Jianing Chen
Address: 158 N Mariposa Ave, Los Angeles 90004
Email: jchen801@usc.edu

Project Information

Project Title: Behavior Prediction and Safety Control of Clients Based On WiFi Data

Date Started: 2021/09/16

Date Completed: 2021/12/10

Project Sponsor/Champion: Kiana Analytics

Executive Summary: Six Sigma Project

- **Problem Statement:** We are aiming to provide a safe and comfortable working environment for clients, so we first built a time series model to capture the trend and patterns of WiFi connection frequency. The IT department can try to maintain network stability in peak time. Then we used a visualization tool (Tableau) to compare people gathering areas before and after COVID-19, and established a group detection system. For the Security department, they can track clients' movement and provide a COVID-19 alert if one of the clients got infected.

- **Project Scope:**

  Objective:

  Provide a safe and comfortable working environment for clients.

  Deliverables:

  1. Final presentation
  2. Final report
  3. Tableau Dashboard

  Cost:

  Since the dataset is very large, we are concerned that computation time would be costly, and some cluster algorithms will be affected by a very large dataset.

  Out of scope:

  1. Detect whether the COVID-19 pandemic will affect clients' attendance and behaviors by comparing data in 2019 and 2020
  2. Develop a group detection system used for alerting clients if the same group client get infected by COVID-19
  3. Computational cost might be large

- **Major project phase (DMAIC) milestones and key learning:**

Phase	Due	Details
Define	From 09/16 to 10/14	● Collecting ideas and potential topics of the Kiana dataset. ● Finish 2-3 pages report on big data, implement project charter based on the presentation feedback and explore the whole dataset.
Measure	From 10/15 to 11/03	● Cleaning and processing the data. ● Using clustering algorithms and silhouette scores to make clusters for clients, and make a weekly time-series pattern on the dataset. ● Report on the project processing and part of the analysis.
Analysis	From 11/04 to 11/19	● Identify Potential Causes, determine root causes using Pareto Chart and Fishbone Diagram ● Analyze the root causes.
Improve	From 11/20 to 11/30	● Summarization of predictive models for tracing clients' movement, improve the model analysis.
Control	From 11/31 to 12/03	● Fix any technical problems that occurred, improve prediction accuracy and lower complexity, etc

Figure 1: DMAIC milestones

We utilized the knowledge gained from the class in a real business case, and we learned how to manage a project and build an integrated system. Besides, we learned how to communicate among our group members, and have a better understanding of business processes.

- **Test Scenario**

1. If one person gets infected, but there are no severe and moderate contacts, but huge amounts of mild contacts, should the Security department announce all mild contacts?
2. If there are too many severe contacts, should the Security department announce all severe contacts to work from home?
3. What happens if there are huge groups of people gathering together?
4. What happens if there is a drastic increase in WiFi connection during weekends, how should the IT department deal with it?
5. How should the IT department deal with the WiFi sensors during peak WiFi connection hours?
6. If more after COVID-19 dataset is given, the updates of the comparison of before and after pandemic will have any changes, and how will the security department deal with this change?

- **Conclusions**
  - ❖ After COVID-19, clients distribution patterns in the UK offices become sparser and the quantity of clients decreases drastically.
  - ❖ In group detection, around the end of April, the “severe” alert notice reached a peak. After May, the “severe” notice decrease drastically
  - ❖ For Time Series, clients have a periodical working schedule pattern weekly. Interestingly, Wednesday and Thursday are the busiest workdays every week.

- **Project recommendations**

  Non-technical part:
  - ❖ Pay more attention to VOC and understand what customers really need; Make sure our project goals must have practical uses and satisfy their needs
  - ❖ Making milestone charts to make our project work structured; making sure what we need to finish in each milestone specifically and timely.

- ❖ Pay attention to stakeholders' and professors' feedback for weekly presentations and make appropriate adjustments if our project deviates.

Technical part:

- ❖ Select appropriate amounts of data since the dataset is a little bit overwhelming and we do not want to run out of my GPUs.
- ❖ Using appropriate machine learning models to make predictions and systems based on our project goals.
- ❖ Doing some basic data processing and cleaning steps to avoid any missing data, imbalance ratios which have negative influences for us to do data analysis

- **Actions Taken**
  1. Doing the VOC survey to understand what customers really need.
  2. Making milestones charts to specify every task we need to finish for our project ultimate goal
  3. Write a brief summary of TA's suggestions, Stakeholders' feedback to our project, and try to improve it.
  4. Evaluate our models frequently on the different testing datasets to enhance and stabilize our accuracy and recall.

- **Benefits:**
  - ❖ Process capability: Security guards and IT workers can work more efficiently
  - ❖ Financial benefits: Lower the cost of hiring employees if they work more efficiently, such as security guards. Stabilized network will make clients work more efficiently and provide benefits to the company.
  - ❖ Other benefits: Protect the clients' health and maintain a safe working environment

Lean Six Sigma Project

DEFINE PHASE

Cost of Poor Quality Statement

Our macro goal is to let clients work in a secure atmosphere. Specifically, we mainly use the time series SARIMA model to predict WiFi connection frequency for each floor. And this is designed to deliver to the IT department to help clients work in a stable WiFi environment. Besides, after the pandemic starts, we observe that few people are in the building. We hope that the IT department can close some WiFi sensors or hence WiFi speed based on our dashboard. We also design two systems for the Security department. They correspondingly are “Clustering comparison before and after COVID-19” and “COVID-19 infected group detection”. By doing these tasks, we hope the department can pay more attention to clients’ safety since we notice Kiana has no specific protection rules for clients so we really hope we help that.

Customer Satisfaction (Voice of the Customer)

The customers we are facing are the both Security department and IT department. The poor quality can affect both the Security department and IT department to make decisions to maintain a good working environment, indeed to affect UK building employees’ moods negatively and may even potentially affect their working efficiency. To avoid these from happening, learning what’s customer needs is very important. We can hear their voices by doing some sample surveys, which are objectives for us to implement in our project. And we believe our actions can only be meaningful if we truly understand what they want. Besides, we can define customer needs and requirements based on the Voice of the Customer. Critical-to-quality characteristics for providing a safe and comfortable working environment need to consider both physical and model accuracy. Specifically, physical refers to dashboard design and how easy to use it. Model accuracy refers to how our models perform.

Tools Application

- Create a Project Charter that defines the business case, problem statement, goal statement, and scope for the project. We noticed that, other than business case, goal, and scope, constraints and project assumptions are also very important. Given the limited features of the dataset, we have to make clear assumptions before we start data analysis.

Project Charter

Analysis:	Analysis on Wifi Data for clients connecting wifi and then speculate their working time range and frequency
Conclusion:	Give approximate clients' generalized working time range and frequency and we can give clients safer working environments based on their working schedule
Project Title:	Behavior Prediction of Clients Based on Wifi Data
Team Name:	Client Observers

Business Case	**Problem/Opportunity Statement**
We're given more than one hundred thousand clients' datasets related to a specific time in connecting wifi, their location, longitude and latitude information. Then we need to analyze clients' past working time schedule pattern based on the dataset of the time period of connecting wifi in one day and predict future pattern base on clients' history. By checking overall clients working hour distribution patterns, we will enhance safety and security, wher many clients work in that building.	1. There is a very small difference between each record of longitude and latitude information, which might be hard for visualization. Also, it may be hard to distinguish individuals' belonging and may require a very precise classification system. 2. The local time column records hours, minutes, and seconds, the time range covered in the whole dataset is merely from 2019-2020, so this will be too small if we do a time series analysis.
Goal Statement	**Scope**
1. From clients' perspective: to analyze clients' working time schedule pattern based on dataset of a time period of connecting wifi in one day and predict future behavior schedule based on history. 2. From companies' perspective: based on clients' working time schedule pattern and time series analysis on the working hours to adjust companies rules to provide clients with safe and comfortable working environment.	**Objective**: analysis client behaviors, for both individual and company **Deliverables**: report and presentations to demonstrate our results **Dataset**: all clients' addresses and time of using wifi from 2019 to 2020 in Kiana **Functions**: analyze clients behaviors pattern in the past and then predict future behaviors **Tasks**: perform data cleaning, paralle l processing, model training, hyperparameter tuning, summary **Costs**: the dataset is massive, and we're concerned that computation time would be costly **Out of Scope**: Detect whether covid-19 issues will affect clients attendence and behaviors by comparing data in 2019 and 2020

Team Members		**Project Timeline**	
Key Stakeholders	Member Names	Key Milestone	Target Date
Team Leader:	Ruichao Ma	Start Date:	09/16/2021
Team Members:	Ruichao Ma	Data Cleaning	09/30/2021
	Chuqi Liu	Clustering, Time Series & Midterm:	10/21/2021
	Jianing Chen	Behavior Model & Second Report	11/4/2021
		Final Report:	12/2/2021

Project Budget:	N/A	**Project Resources:**	

Constraints, Assumptions, Risks and Dependencies	
Constraints	1. Features to analyze behaviors are limited 2. Computation time will be costly due to the massive dataset 3. The difference individuals' latitude and longitude are small
Assumptions	1. Clients mac address directly refer to the individual client. 2. All longitude and latitude data refers to a building belongs to a large chemical and pharmaceutical company on the ground floor to the 2nd floor.
Risks & Dependencies	1. We are not sure whether individuals or several people can use one mac address. 2. It is hard to determine the location of people in the building, but the device is in airplane mode. 3: If sensors detect devices, it's hard to distinguish whether devices in unused status or they are used by some real clients

Figure 2: The Project Charter

- Determine the Voice of Customer (VOC). The important aspect of VOC is to capture the feedback from customers so that we can improve our product. For our project, the customers are both Security department and IT department people, suppliers include our team, Client Observer, and Kiana

company. Our team will use the wifi data provided by the Kinaa company to create 3 dashboards for both the Security department and IT department people in the UK company to use.

- Create key milestones of the project so we can follow the steps and keep track of project deadlines. The milestones not only incorporate the details of the project phrase but also include each homework and presentation for this course.

Key Milestone	**Due**	**Details**
Brainstorm	09/16	Collecting ideas and potential topics of the Kiana dataset.
First report and data exploration	09/23	Finish 2-3 pages report on big data, implement project charter based on the presentation feedback and explore the whole dataset.
Data cleaning	10/15	Cleaning and processing the data.
Clustering and time series	10/30	Using clustering algorithms and silhouette scores to make clusters for clients, and make a weekly time-series pattern on the dataset.
Monthly report	11/04	Report on the project processing and part of the analysis.
Midterm presentation	11/12	Team presentation on the monthly report, and stakeholders.
Predictive model	11/19	Summarization of predictive models for tracing clients' movement, assist for the later analysis.
Implement the model	11/26	Fix any technical problems that occurred, improve prediction accuracy and lower

		complexity, etc
Team presentation	12/03	Team presentation on stakeholder summit.
Final report	12/10	The comprehensive report with analysis, final project closeout.

Figure 3: Key Milestones of Project

- Create a SIPOC diagram that lists supplier, input, process, output, and customer. SIPOC is a tool that summarizes every important flow in table form. It is used to define a business process from beginning to end before work begins. The key learning point using SIPOC is to have a general understanding of projects.

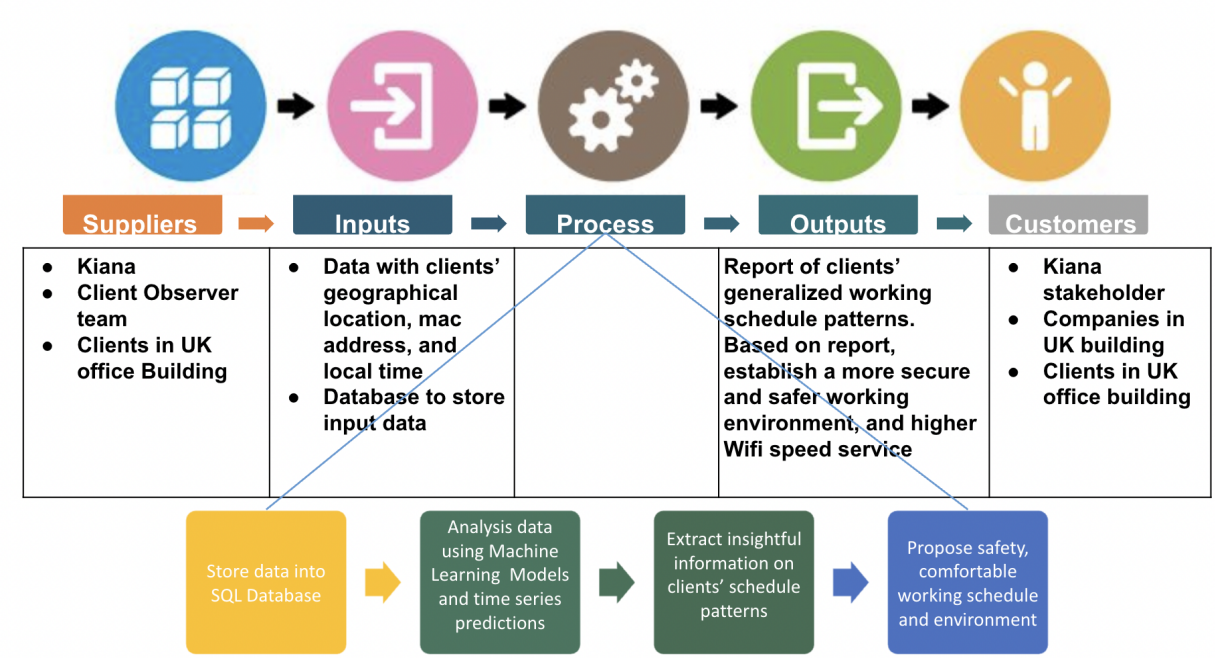


Figure 4: SIPOC Diagram

MEASURE PHASE

Process Mapping/Process Visualization

In processing mapping, we first built a high-level process map which is the same as SIPOC. In this part, we described the major tasks and listed suppliers and customers. Also, we clearly claimed what the process required input, process, and output. Then we built a basic flow map (common process mapping) to describe the whole process of this project, each step we did and plan to do. Besides, we plot each step's relationship including loops and decision points. Then we built a detailed mapping. In this map, we described each step in detail and value-add status such as how to store the data, what steps we did in data cleaning and what kind of models we are planning to build, etc. Finally, we finished with a functional mapping in which we described each step in a bounded process and separated it into different functional areas, such as which department is responsible for each step and phase. All of the process maps are shown in the figures below.

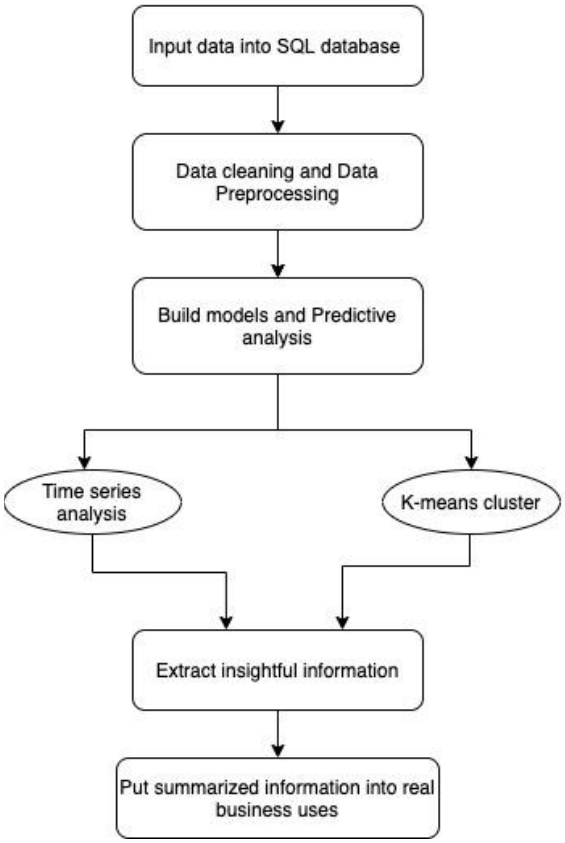


Figure 5: Common Process Mapping

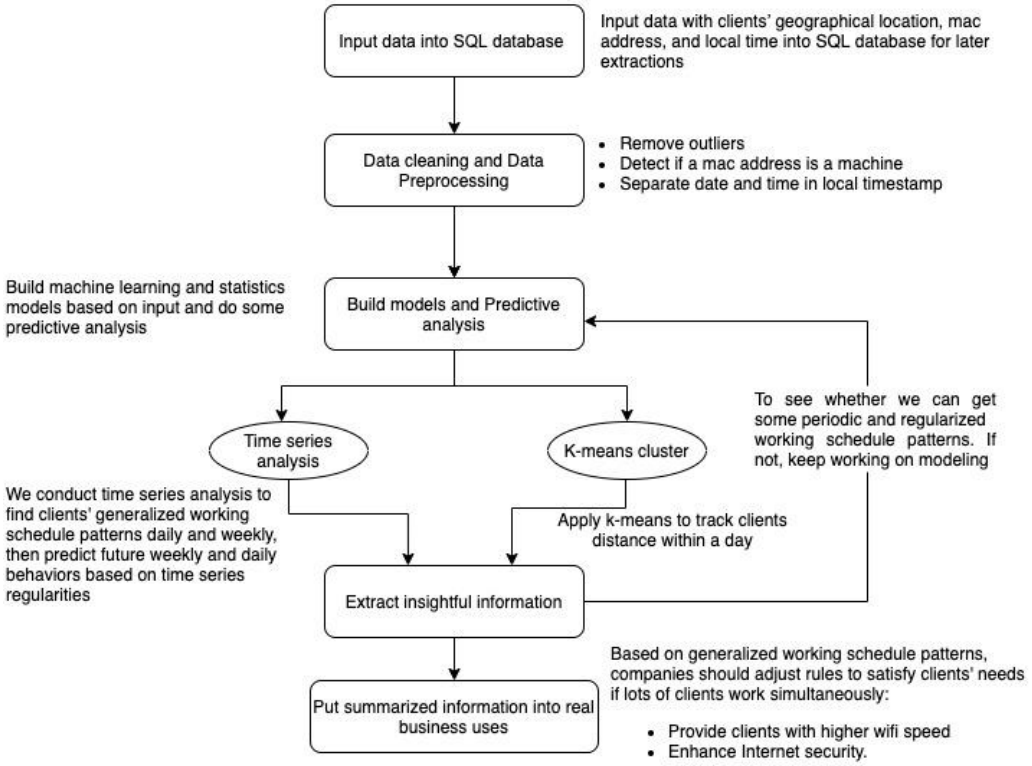


Figure 6: Detailed Process Mapping

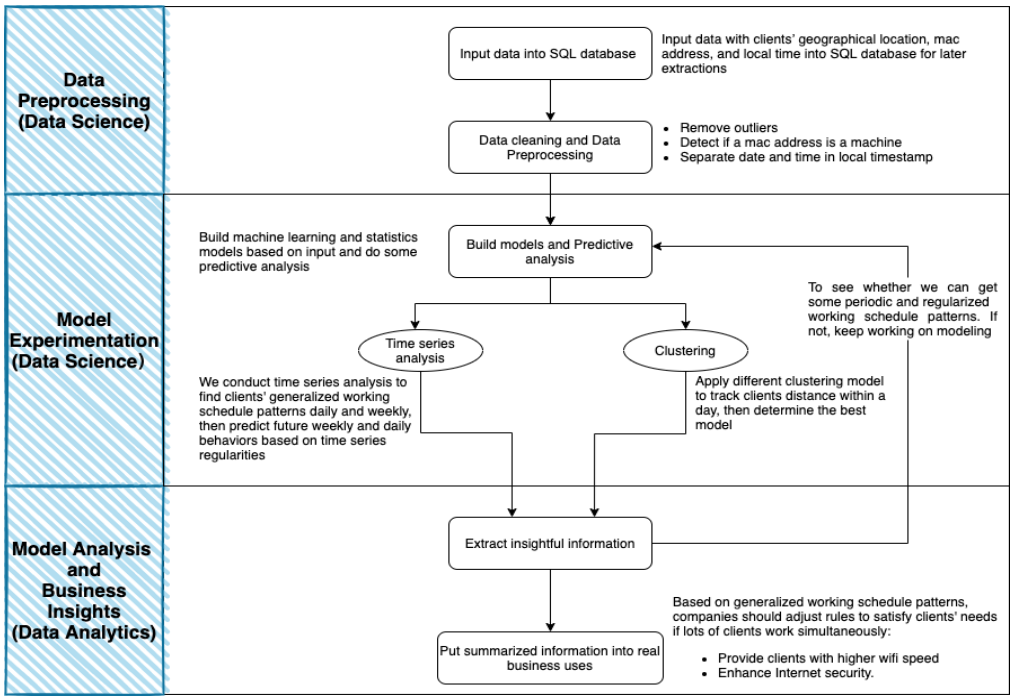


Figure 7: Functional Process Mapping

The Vital Few

Pareto's chart is the best method we employed to select the "vital few" input variables (X's) affecting the output (Y). Since by using this visualization tool, we can truly know what elements take around 80 percentages of effects for our whole project. We define Y as an uncomfortable working environment. Furthermore, we have a few significant factors of X's:

1. Low WiFi speed
2. Unstable internet connection
3. Insufficient security guard
4. Device malfunction
5. Outdated device
6. Loss of clients' privacy disclosure

Data Collection Planning and Execution

In the data collection part, we are given 88 separate text files which contain clients' mac addresses and their location information. But data amounts are too overwhelming, each txt file contains 3 million data. If we combine all datasets one, GPU will be overloaded and computation speed will be really slow. So we use stratified sampling by date, each date we only select the fifth percentile of data. Also, we present a normal working time range from 6:00 AM to 10:00 PM. Moreover, we discovered that some mac addresses appear in very high frequency on some days, even repetition occurrences can be as high as 8000 times per day. So, we will treat this as an autonomous device and delete it from our dataset since it cannot represent real clients' information. We also preset the threshold for repetition occurrences as 3000. If a mac address appears more than 3000 times, we just remove this mac address. Moreover, we separate data and time into two features since we will use them separately in our future model. The original data has both data and time in one column, so we felt it is not convenient to use if we want to only select certain days or only want to look at a few hours. Finally, we applied many different data transformation method, and finished with a box cox

transformation to transform all non-normal dependent variables into normal shapes and make it easier for us to analyze in the future.

	Site	Level	ClientMacAddr	lat	lng	date	time
0	UK Office	1st Floor	9c:8c:6e:46:1c:5e	51.4604	-0.933048	2020-09-01	23:58:49
1	UK Office	1st Floor	9c:8c:6e:46:0b:7c	51.4608	-0.932288	2020-09-01	23:58:45
2	UK Office	1st Floor	9c:8c:6e:46:1c:5e	51.4604	-0.933048	2020-09-01	23:58:38
3	UK Office	1st Floor	9c:8c:6e:46:0b:7c	51.4608	-0.932292	2020-09-01	23:58:35
4	UK Office	1st Floor	9c:8c:6e:46:0b:7c	51.4608	-0.932292	2020-09-01	23:58:29
...	...	...	...	...	...	...	...
3733370	UK Office	1st Floor	c4:6e:1f:1a:68:a1	51.4607	-0.932252	2020-03-21	23:26:33
3733371	UK Office	1st Floor	a4:e9:75:77:7e:c3	51.4605	-0.93235	2020-03-21	23:26:28
3733372	UK Office	1st Floor	9c:8c:6e:46:0b:7c	51.4608	-0.932317	2020-03-21	23:26:27
3733373	UK Office	1st Floor	c4:6e:1f:1a:68:a1	51.4607	-0.932252	2020-03-21	23:26:27
3733374	UK Office	1st Floor	9c:8c:6e:46:1c:5e	51.4604	-0.933043	2020-03-21	23:26:27

3733375 rows × 7 columns

Figure 8: Data After Pre-processing

	date	count	box_cox_count
0	2019-08-12	66579	8.511450
1	2019-08-13	1813781	10.254144
2	2019-08-14	1685426	10.218500
3	2019-08-15	1587230	10.189249
4	2019-08-16	1010854	9.966545
...	...	...	...
360	2020-08-28	222808	9.182205
361	2020-08-29	172802	9.044431
362	2020-08-30	169946	9.035335
363	2020-08-31	163341	9.013670
364	2020-09-01	536612	9.645327

365 rows × 3 columns

Figure 9: Box-Cox Transformation of WiFi Connection Count

Measurement System Analysis

For the time series analysis, we use MSE as our evaluation metrics. We fine-tuned our model hyperparameters and used MSE to check the distance between predicted and real data points. If MSE is too large, we will try the other hyperparameters.

For the clustering part, we use silhouette scores and elbow curves as our evaluation metrics for detecting the optimal number of clusters in K-Means. We detect every silhouette score for each number of clustering: K. And we found that when K=3, the silhouette score is highest. Moreover, the inertia value decreases most drastically when K increases from 1 to 3. So we judge the optimal number of clusters should equal to 3. In the visualization plot, we see it makes sense when K is equal to 3 because inter-cluster distances exist between groups. We mainly detect recall, precision, and F1 score of clustering results in the group detection system. We previously set up a testing dataset in a CSV file. After DBSCAN shows some results, we will save the output and compare it to our testing dataset directly and calculate the corresponding F1 score, precision, and recall. Fortunately, all these values are around 0.92, proving DBSCAN is very useful for designing group detection systems.

Tools Application

For the time series analysis, we use MSE as our evaluation metrics. MSE tells us how close predicted points are close to real data points. It does this by taking the distances from the points to the regression. As a result, MSE will be influenced by the range of data points. Therefore, we also incorporate residual analysis after we build the model. The residual analysis plays an important role in validating our models. Suppose the error term in the regression model satisfies the assumptions, such as it having a constant mean, constant variance, or standard deviation, and the auto-covariance should not depend on time. In that case, the model is considered valid.

We mainly used packages in python to detect precision, recall, and F1 score as our evaluation metric defined for measurement system analysis in clustering. Since all evaluation metrics are numerical values, it is easy to detect whether we are satisfied with our models by a numerical value. We also gain lots of key learnings arising from the application of these tools. In this process, we learned precision is a measure of correctly identifying positive cases from all predicted positive cases. Also, recall is a measure of correctly identified positive cases from all the actual positive cases. Accuracy is the measure of all correctly identified cases. So far, using any of these evaluation metrics solely cannot get satisfactory results, the best method we think so far is to use F1-score. This is the harmonic mean of precision and recall that give a better measure of incorrectly classified cases than accuracy metrics. Below is the attached formula:

$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Figure 10: F1-score Formula

Moreover, if imbalance cases exist in the dataset, F1 can also deal with it perfectly.

ANALYSIS PHASE

What charts will help you analyze the data collected from Measure?

Among all the charts we learned this semester so far, fishbone helps us a lot in identifying cause and effect in business problems. Since we want to design a cluster map to track clients' locations, we first need to discern the fact that clients' physical movements and locations are difficult to track. For instance, for different groups of clients, their working schedules may vary. Also, their experiences, specialties, and department belongings are different. Only by understanding these difficulties, we can come up with a solution to avoid all these difficulties, even if it is a struggle to find. Also, since our clustering group detection system is designed for the security department. We need to consider some realistic causes. Like we

cannot expect a company to hire 1000 guards to ensure safety for every client since it is very costly. To summarize, in the “people” part of the fishbone diagram, we set some realistic expectations: companies may think it is unnecessary to hire so many guards. So many guards in a company are hard to divide labor and financially costly, so on and so forth. By only considering all these realistic factors, we can avoid letting our solutions be too fancy. To sum up, we use the fishbone diagram tool to keep the team focused on the causes of the problem, rather than symptoms. Also, fishbone provides us a chance to brainstorm all possible causes of the problem. Ask “Why did this happen?” As each idea is given, the facilitator writes the causal factors as branches from appropriate categories and causes can be written in several places if they relate to several categories.

Pareto charts are also very important for us to recognize real important causes for the project. Since there are 80/20 rules. Only a few factors can largely affect our outputs and all remaining factors play trivial roles. Specifically, we designed a Pareto chart to reflect in which parts clients are most unsatisfactory for companies. Finally, we found that “insufficient security guard” and “low WiFi speed” are two top reasons. Other unsatisfactory factors like “device malfunction, dissatisfied salaries” exist but are trivial. Thus, our project goals mainly want to deal with the top two factors clients are most unsatisfactory about. In some sense, the Pareto chart gives us clear direction for our project to strive for.

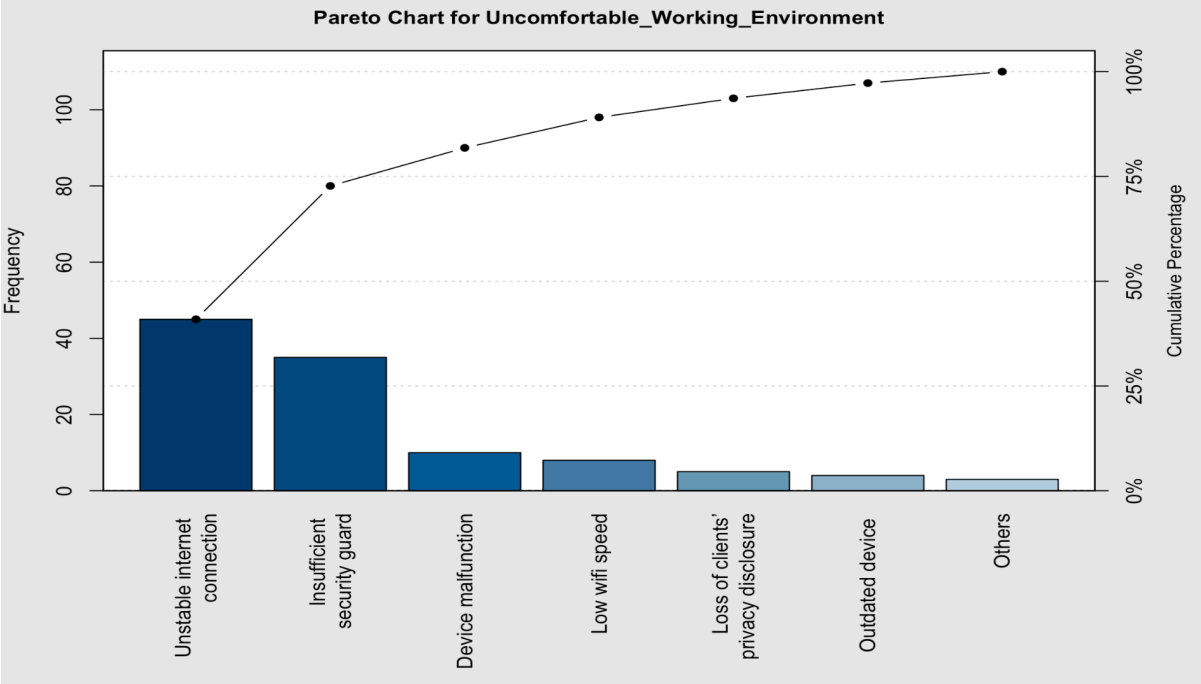


Figure 11: Pareto Chart

The 5 whys diagram is also an important chart in the analysis phase. When we come up with doubt or question, it is easy to think too superficially and ignore the true reason behind this. For instance, in our project, when we come up with a question: “what causes insufficient security guards for the company?” Initially, we thought the company probably had budget deficits. But when we truly deploy the 5 why diagram, we found the truth may not be like this. The superficial reason is that “the security department does not arrange appropriately for clients' working schedules.” And the deep reason is that “some clients work with no regular patterns”. If we want to extract some regularities from these, we need to use machine learning models to do that. Thus, 5 why diagrams give us a chance to think deeply about the root cause of problems. And we believe that only if we address the root cause, our project deliverables can be really meaningful and make some impact.

In summary, we mainly use fishbone diagrams, 5 WHY diagrams, and Pareto charts. By using Pareto Charts, we can gain a deeper understanding of what

elements will affect our project implementation most significantly and what matters least, and then we can focus on those “significant” elements to produce higher efficiency. From the fishbone diagram, we can visualize the potential cause of a problem. From the 5 WHY diagram, we can delve into one problem statement deeply and successively to know the root reason behind this. Those all are key learning we learned from the process mapping activity.

List your Y = f(x) formula

Y equals the outcome for the potential solutions, so we will define it as providing a safe and comfortable working environment for clients. X refers to the inputs necessary to achieve this solution, so it comes from Measurement, for example, the difficulty to discern clients’ physical movement and location; X also can come from People, which means the security guards do not arrange appropriately by clients working schedules; X also can be Machine, which can be both device malfunction and internet speed may unstable and slow. In addition, X also includes Method and it is no reasonable way to detect clients’ privacy disclosure amounts. Lastly, X can also be Data, because features are limited and geographical locations between rows are tiny, and database storage is a difficult issue since data is overwhelming.

SWOT

- Strengths: All dependent variables in Y = f(x) are easy to detect and measure theoretically. With more and more known variables, Y is easier to quantify and the value of Y can be more accurate. Moreover, our project output and dependent variable can form a linear regression relationship. All dependent variable effects can be added together to cause effects on Y.

- Weakness: The Kiana company provided the dataset is too overwhelming and we cannot gain comprehensive statistical views for all dependent variables. We only select subsets of representative data to use in our project. Also, how to decide representative data and implementation is very time-consuming because the data amount is too large. Also, useful features

are limited since we are only given mac address and correspondingly latitude and longitude information so not too much useful information can be extracted.

- Opportunities: If the geographical map for the UK office can be more specific, including stairs location and obstacles in building, we can delve deeper insights into specific distances between clients given latitude and longitude information. Moreover, if more features for clients information can be given, like weather, mac address devices type, etc. We can analyze for more useful information.

- Threats: Although relevant features for clients' information are easy to find, merging all datasets into one is a difficult task. We cannot ensure all clients have information on additional variables. Even if we use data imputation techniques to fill out these missing values, deciding to use which values to fill out is difficult because it is very subjective. Using majority vote, or median to fill out all missing values is not a good idea since not all data are representative since some mac addresses can appear many many times.

What are the Root Causes?

The root causes are the factors that caused a nonconformance and should be dealt with and eliminated through process improvement. In our project, we will consider the root causes come from 5 aspects which are Measurement, People, Machine, Method, and Data respectively. Those are all the factors that can cause an uncomfortable working environment. Each of the aspects has one or two detailed causes and we explored 5 whys for each cause. The detailed fishbone diagram is shown below.

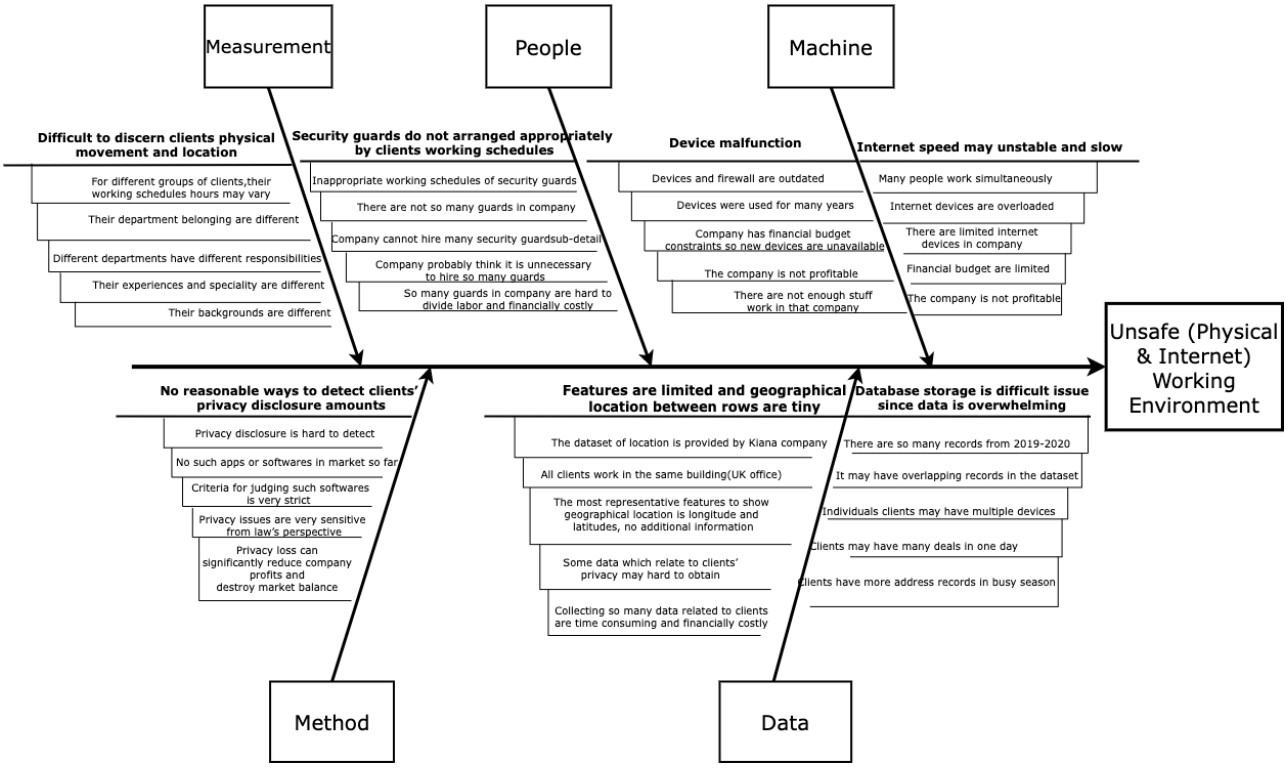


Figure 12: Fishbone diagram and 5 whys

Are there any Correlations?

For the Machine, Measurement and Method causes, since it is hard to discern whether devices are used by corresponding users, it will increase the difficulty of Measurement. Also, if the device malfunctions such that devices were used for many years and firewalls are outdated, the devices cannot have a good performance to protect the privacy of users.

For Measurement and Data causes, since it is difficult to discern clients' physical movement and location, so our dataset features are limited. Then, we delve into the deeper reason for the difficulty of discernment, and we found that it probably relates to the Machine part, since some devices are malfunctioning and outdated, probably the tracking system, or something like that is not "smart" and accurate enough.

Also, for People and Measurement parts, it is reasonable to speculate that since clients' locations are difficult to discern, then the security department does not have a clear picture about clients' working schedule patterns, so they have difficulty appropriately arranging security guards for clients.

Also, there is a correlation between Machine and Data parts. Since some machines are outdated and malfunctioning, we can expect the computer system does not have identification to different mac addresses and speculate their belongings. So the data amount is too overwhelming because machines just simply transfer all data into text files but have no capability to recognize and delete some repetition occurrences for mac addresses. So it directly causes difficulty to store data into the database.

What are the sources of variation?

Date and Time selection: Clients' locations may change at different dates and times with no regularity. Since clients are not robots, predicting their movement in the next few days on a very precise scale is almost impossible. So, sometimes it may cause misclassification in our machine learning algorithms. For instance, not all clients' locations can be classified into the correct clustering group perfectly.

Moreover, there exists a class imbalance problem. Clients' data on the 3rd floor are very minor in quantity so it may cause some difficulties to improve accuracy on the group detection system and time series prediction on the 3rd floor.

List your Potential Solutions

Here are potential solutions to the root cause:

1. Measurement and Data: Trying to find distribution plots for daily clients' locations and delete redundant ones to reduce data amounts. Also, trying to find more datasets which contain more features for clients that are useful for projects.

2. People and Method: Designing more precise classification and tracing systems for clients and trying to come up with features thoroughly that are irrelevant to clients' privacy but useful to trace their location.
3. Machine: Trying to expand the company's budget and replace old computers or outdated devices with a new one.

IMPROVE PHASE

Alternative Solutions Considered

1. Provide more comprehensive data with more features, such as IP address, IP name, etc, so the time series analysis can be more accurate. Put it into other words, we can use a more complicated model structure, such as deep learning models. For example, feedforward neural networks that can extract different features and learn complicated non-linear relations by adding multiple hidden layers. Or, using LSTM, which provides some short-term memory so it is good at extracting patterns over long sequences.
2. If we can access the detailed description of the UK office, such as the structure of the building (rooms, offices, etc.) and WiFi sensor locations, we may give a more insightful analysis on how to maintain the WiFi connection. Based on the sensor location, the IT department can try to close a few sensors if there are very little people in the building to save some energy.
3. Trying to find more datasets that contain relevant features for clients which are useful for both predicting clients' working patterns and clients group clustering
4. Pay more attention to date selection and delete mac addresses that has high repetitions occurrences, we can decrease data amounts and avoid high computation cost by using this way
5. Trying to explore more fast-speed machine learning algorithms and experiment it in our project to see whether it is suitable.

Recommended Solution(s)

All alternative solutions listed above are recommended solutions. Admittedly, implementing them one by one is somewhat difficult. For instance, we have concerns about how to expand our dataset with more useful features, since finding one that fits our original dataset perfectly is hard. Also, some machine learning algorithms to make clusters are not applicable to our project since they are not suitable for analyzing geolocation data,(Hierarchical clustering, Mean-Shifting-Algorithm, etc) like longitude and latitudes. We will consider these concerns fully and try our best to come up with “perfect” solutions which can avoid all these concerns.

How will this be piloted?

For clustering parts, we make a group detection system on several separate dates to evaluate its effectiveness since using all dataset is too overloaded and will make the model not very interpretable. We mainly select after COVID-19 era date, with 2 weeks as one interval. We select :04-01, 04-15, 04-29, 05-13, 05-27 as five “experimental” day. We apply a group detection system to these five days and detect “moderate” “mild” “severe” alerts. We found dates 04-01, 04-15, 04-29. Severe alert notice takes most percentage. It makes sense since dispersion of COVID-19 in England reaches a peak in April. In 05-13 and 05-27, the percentage of moderate and mild alert increase and percentage of severe alert decrease proportionally. This is also reasonable because due to political policy, more and more people choose to work at home. So the population in the UK office decreased drastically. We applied the same experiment to a client's distribution map by K-Means.

Create a project plan using WBS

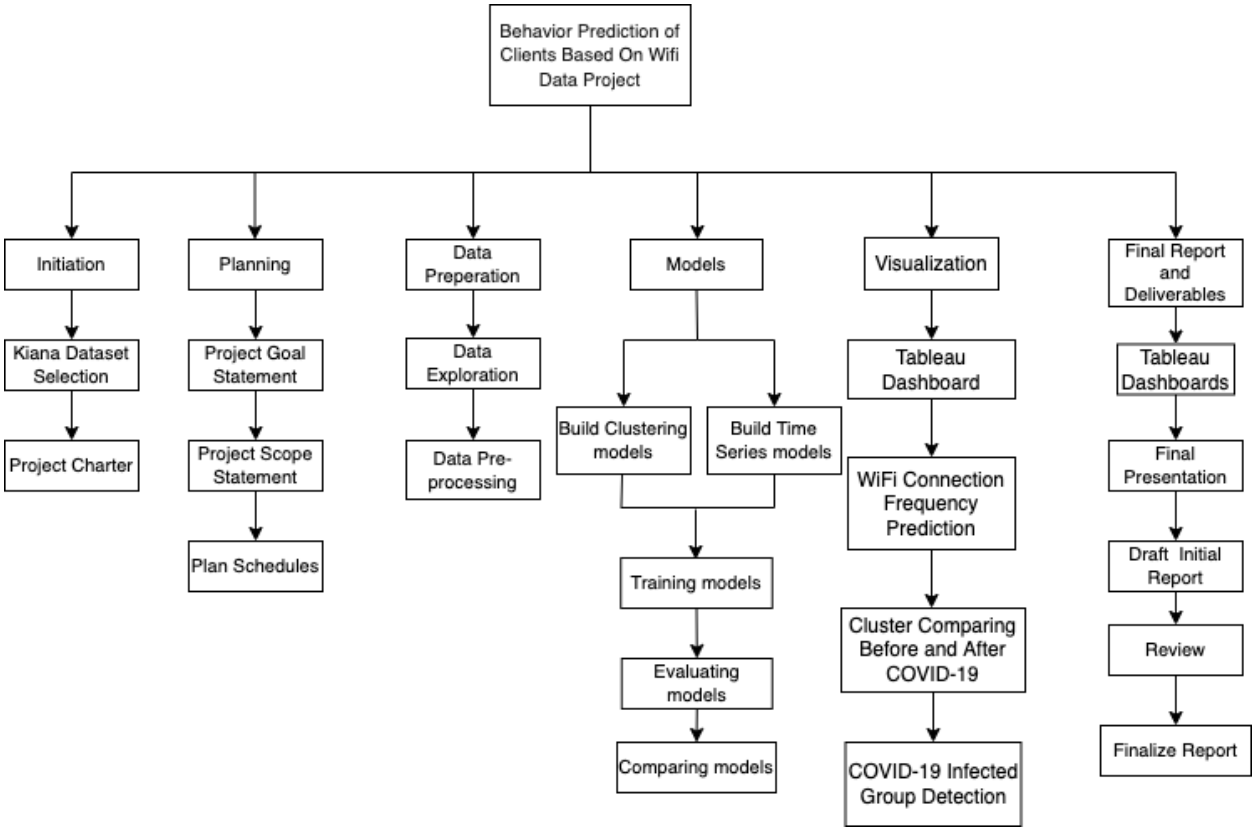


Figure 13: Work Breakdown Structure (WBS)

The figure above shows our Work Breakdown Structure, we totally have six phases from the very beginning of initiating this project to the final deliverables. We followed the time schedule as planned. In different phrases, we updated some detailed points such as instead of only observing clients' working patterns and tracking their movements, we built a COVID-19 infected group detection system as we mentioned in our out-of-scope part, and we finally decided to visualize our output by using Tableau Dashboards.

CONTROL PHASE

During the control phase, we implemented our day-to-day process management. First, instead of finishing our initial objectives, we decided to add a new COVID-19 infected group system for users to make this project more widely used for the UK company. Also, at the initiation phase of this project, we did not plan to have a UI or Dashboard for users, instead, our first plan was to plot all of the graphs and compare results for our findings. But soon, we found it is very inconvenient because we need to compare lots of visualization plots and get results, which is very time consuming. We want to come up with more straightforward ways to compare all graphs with filters. We kept finding a better final deliverables format for users to clearly visualize our results and can compare by themselves. Finally, we decided to visualize our results by using Tableau Dashboards, and designed an integrated and interactive user interface. We generate all new data from the original one after data cleaning and preprocessing. In this process, we met some challenges. Like tableau cannot separate different groups by using different colors. By solving this, we mainly adjust groups' transparency in color to make it visualizable for different groups. Also, we notice that tableau is easy to overload with large datasets. So we use a stratified sampling method for each day and only extract ten percentages of all data to fit in a tableau.

For the follow-up updates, we have uploaded all model code and processed datasets to GitHub at https://github.com/juliachenc/dsci560, and the dashboard can be assessed with a Tableau account at Tableau Dashboard. In the future, we will add a streaming dataset into our project after closure. Streaming dataset is real-time and we believe it is useful to visualize all data on time. Then, in the future, group detection systems can play a more important role because they can send alert notices on time and prevent more people being infected by COVID-19.

RESULT AND SYSTEM IMPLEMENTATION

Machine Learning Approaches

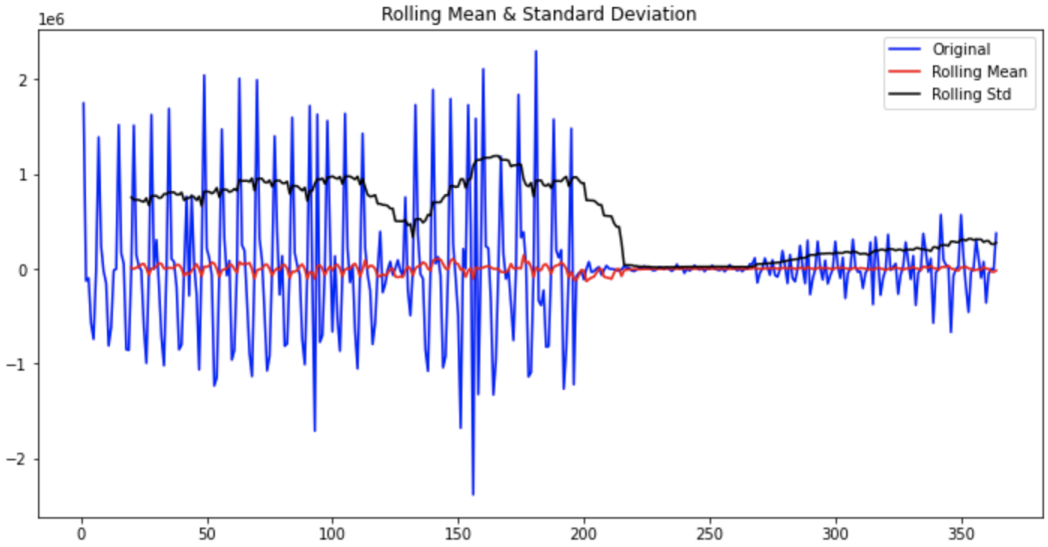
- For the **time series analysis** part, we, first of all, count the number of Mac Addresses by date using all data that Kiana provided. Time series analysis is a statistics tool to extract the previous observation taken at specific time intervals. It also enables model trends and patterns which can support the prediction of the value at a future point. In our case, we used the time series model to understand past behavior and plan for future WiFi frequency connections.

  By looking at the plot of all WiFi connection frequencies, we observe that the connection of WiFi frequency appears to be a strong seasonal trend. Therefore, instead of ARIMA, we use SARIMA to capture the overall pattern. In addition, there are a few important components of the SARIMA model: $SARIMA(p,\ d,\ q)(P,\ D,\ Q)_S$

  1. Autoregressive terms (AR: p, P), associate with the past observations
  2. Number of differentiation(D, d): stabilize the mean of a time series model
  3. Moving average terms(MA: q, Q): associate with current value against past observed values
  4. Seasonal order (S): the overall periodicity pattern

  There is another important aspect of the SARIMA model, which is stationarity. The stationary is the statistical properties of a time series that do not change over Time. If a time series has a particular behavior over a time interval, then there's a high probability that over a different interval, it will have the same behavior. We applied the Dicky-Fuller Test to check if data points are stationary. Below is the example of Dicky-Fuller Test, where rolling mean and rolling standard deviation does not change based on the time.

Also, the test gives a very small p-value, so we can conclude that all observations are stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                -6.812447e+00
p-value                        2.099518e-09
#Lags Used                     1.300000e+01
Number of Observations Used    3.500000e+02
Critical Value (1%)           -3.449173e+00
Critical Value (5%)           -2.869833e+00
Critical Value (10%)          -2.571188e+00
dtype: float64
```

Figure 14: Dicky-Fuller Test and its statistical results

We performed the Dicky-Fuller Test on each floor-based data. If the data points are not stationary, we took the differentiation. Below are the predictions of each floor.

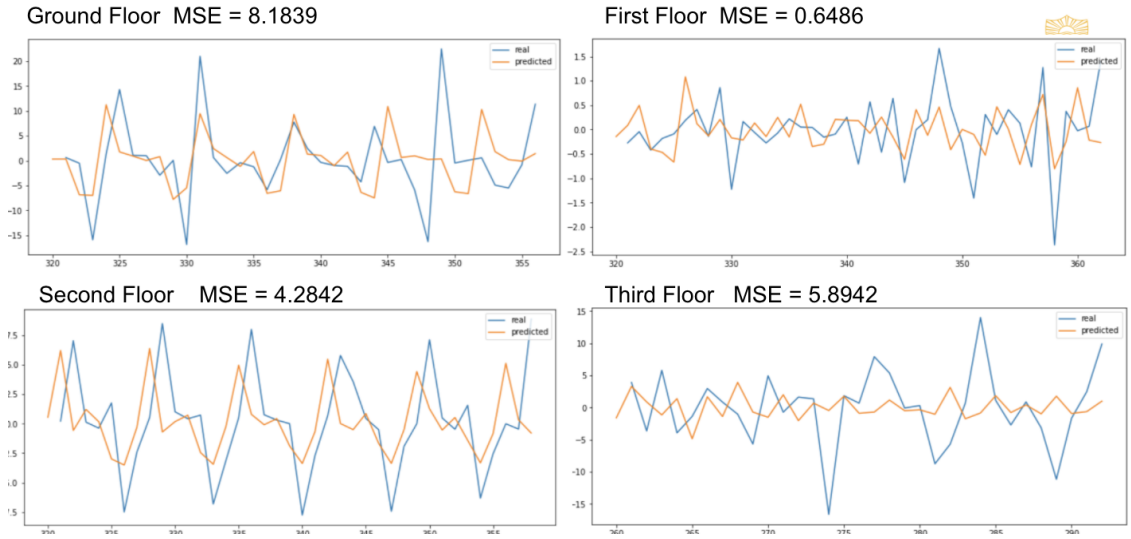


Figure 15: True Value and Prediction Comparison of each floor

We have tried turning hyperparameters for our model, is this the final outcome for each floor prediction. The y-axis is transformed WiFi connection frequency, the x-axis is the date. We change data back to the original scale in the dashboard. We can notice that some floors have larger MSE but look good. For example, the ground floor and second floor. It is because the MSE will be affected by the y-axis. We can see that the gourd floor has a larger range, but the first floor has a smaller field.

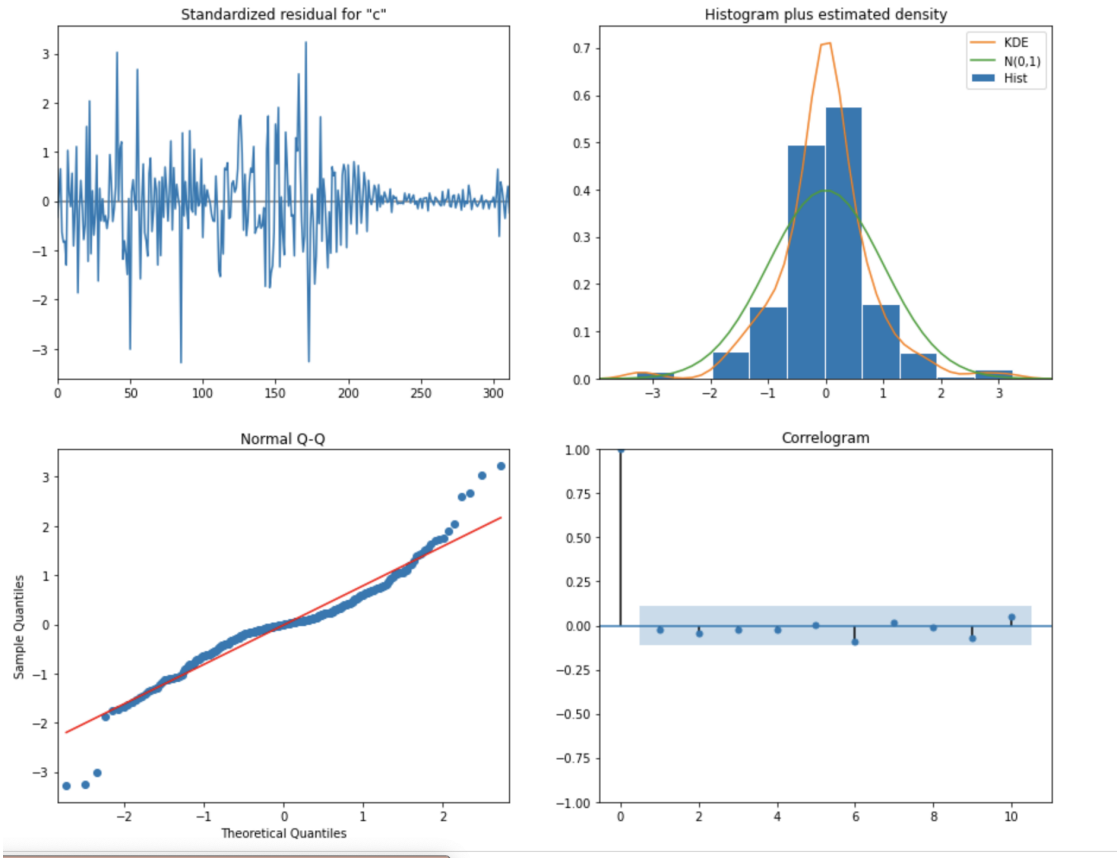


Figure 16: Model Validness Check

So in order to check if our models are valid. We went ahead to check the residual plots. We want it to have a constant mean, constant variance, or standard deviation. Auto-covariance should not depend on time. In the plots, the residuals seem to be normally distributed around 0 — which is the condition that we need. The mean and sd are also normally distributed. From the normal Q-Q plot, it does not severely tail off. We apply this model diagnosis to all floors and make sure all of them are valid.

- For clustering parts, we mainly design two models by using **K-means clustering** and DBSCAN algorithms correspondingly. The first model we design is a “clients’ distribution map in the UK office” by using k-means since it is computationally fast even with a large dataset. Moreover, it is

implemented easily with great interpretability. The only thing we need to notice is that K-means require us to manually set up a cluster number. But it is not very hard to address. We use the elbow curve and silhouette score as our evaluation metrics for the optimal number of clustering. On the ground floor, first floor, and second floor. The optimal number of clustering is always equal to 3. We compared several weeks of data before and after COVID-19 and we found that clients' distribution is sparser after COVID-19. It is easy to understand because, after the COVID-19 era, more and more people choose to work at home.

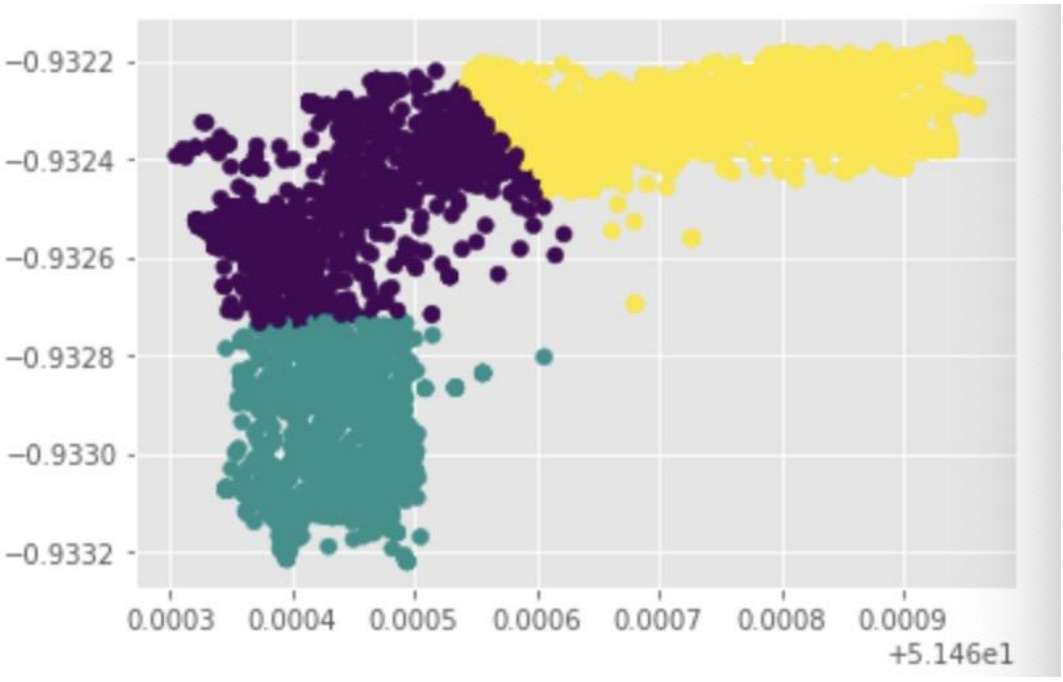


Figure 17: K-means Clustering Plot Before COVID-19

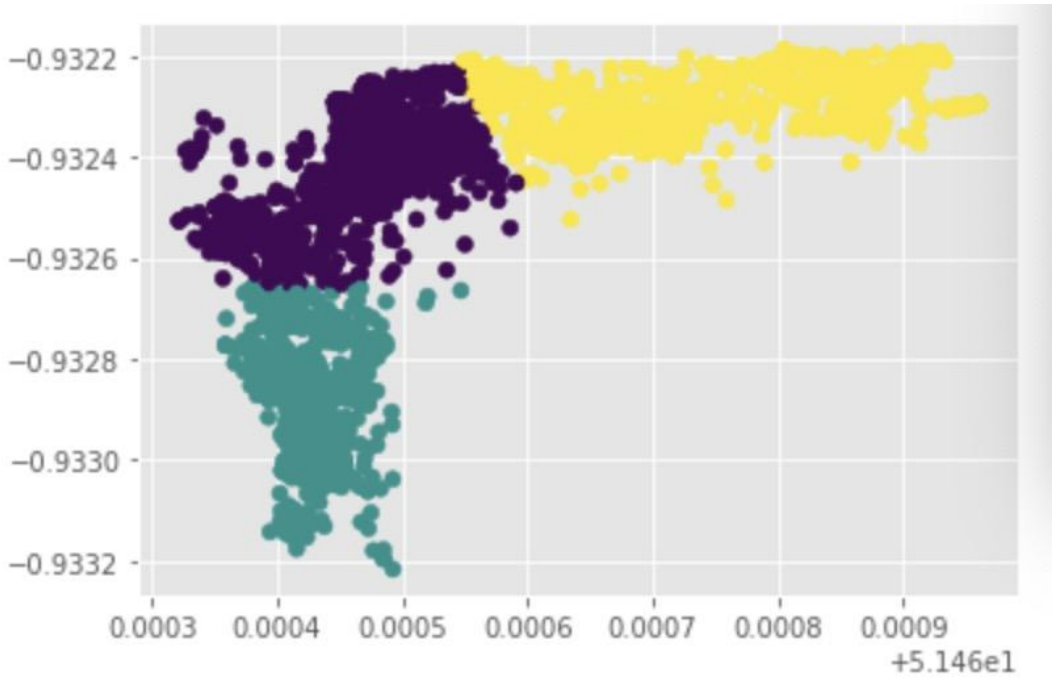


Figure 18: K-means Clustering Plot After COVID-19

- We also design a group detection system by using **DBSCAN algorithms**. This system aims to provide alert systems for clients to keep safe if other clients get infected by COVID. DBSCAN works on the assumption that clusters are dense regions in space separated by regions of lower density. In this algorithm, we don't need to set up clustering numbers manually. Instead, we set safe distances for clients and can be regarded as epsilon in the DBSCAN parameter. We also set 4 for “min points” since it will not incur too many clusters as a result. Specifically, for groups with a safe distance of fewer than 5 meters, if one client gets infected, all other members will receive “severe” warning alerts, if safe distances are between 5-10 meters, only “moderate” alerts will be sent. if safe distances are between 10-20 meters, “mild” alerts will be sent. If safe distances are larger than 20 meters, no alert will be sent. Also, we set one hour as one interval so the circumstance below can be avoided: an infected client stands in one location at 9:00 am. Then, in the afternoon, other clients standing in the same location will not receive alert warnings since a long time has passed by. We

found that between mid and end of April, “severe” alert notice takes up most percentage. That is easy to understand because COVID dispersion around that time is most severe in England. After May, “Severe” notice quantities decrease drastically and mild and moderate alerts increase proportionally. That is also reasonable because probably due to political policies and personal safety, more and more people choose to work from home.

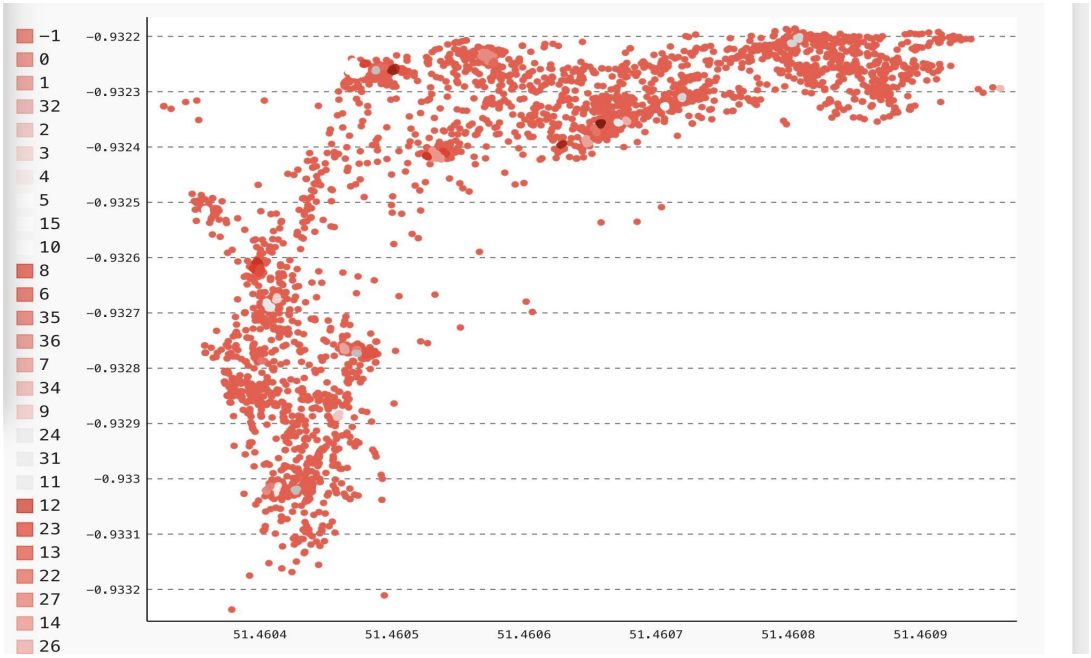


Figure 19: DBSCAN Clustering Plot for Infected Group Detection System

System Implementation

We utilized Tableau to create dashboards for visualizing the results. There are mainly 3 dashboards and we will introduce all of them respectively.

1. Dashboard 1: WiFi connection frequency for each floor, users can explore each floor's true frequency with its prediction using the sidebar filter. The x-axis is the date and y-axis is the connection frequency.

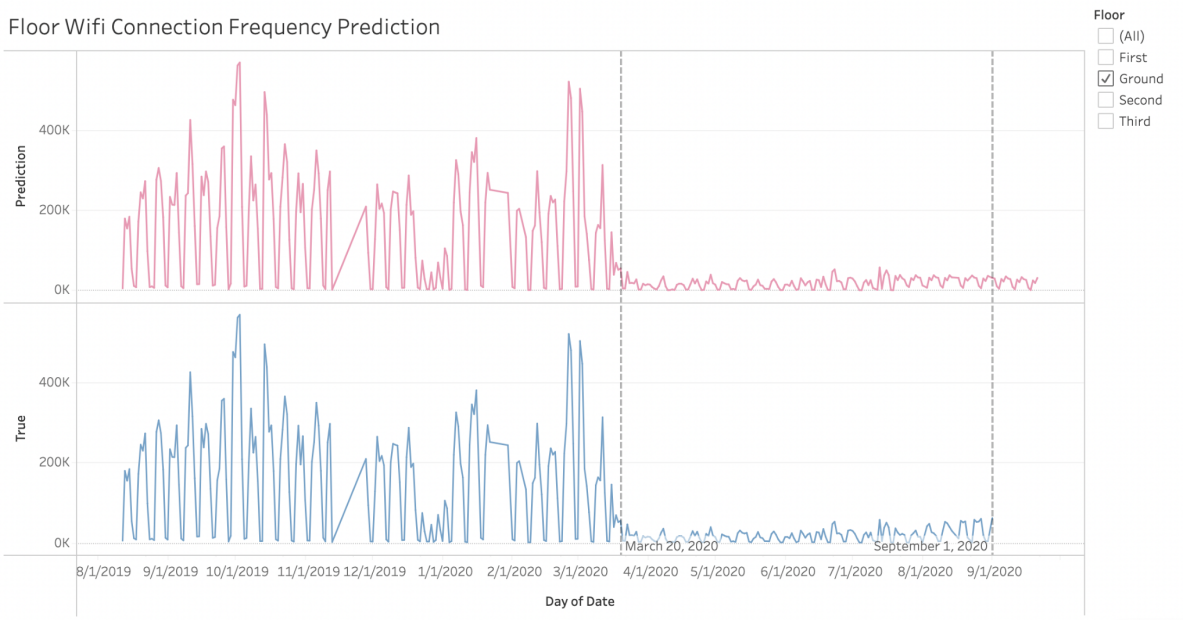


Figure 20: WiFi Connection Frequency Dashboard

2. Dashboard 2: Clustering comparison before and after COVID-19, users can choose one day before and one day after COVID-19 pandemic with a specific location and time ranges by sidebar. To compare by controlling variables, users can choose to select the same workday (such as both Wednesdays) before and after COVID-19 and compare how the pandemic affected the clients' gathering areas within this building on a specific floor.

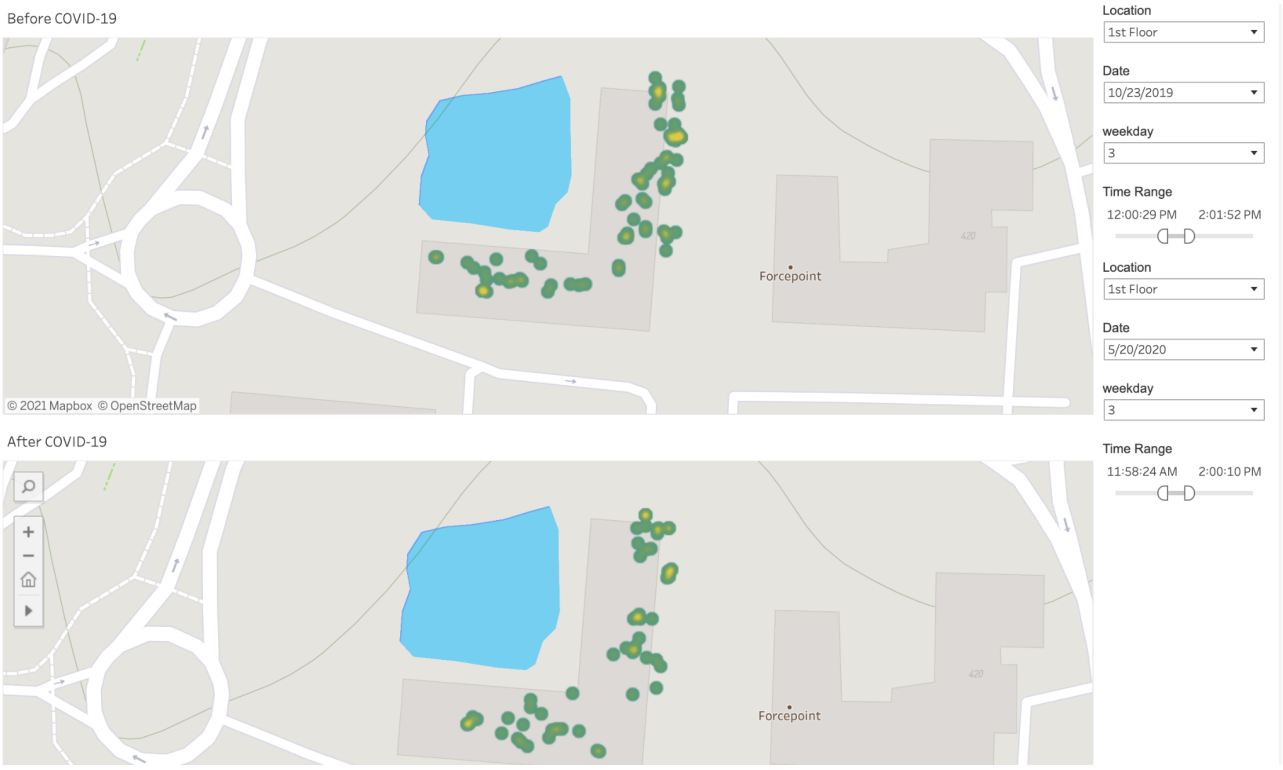
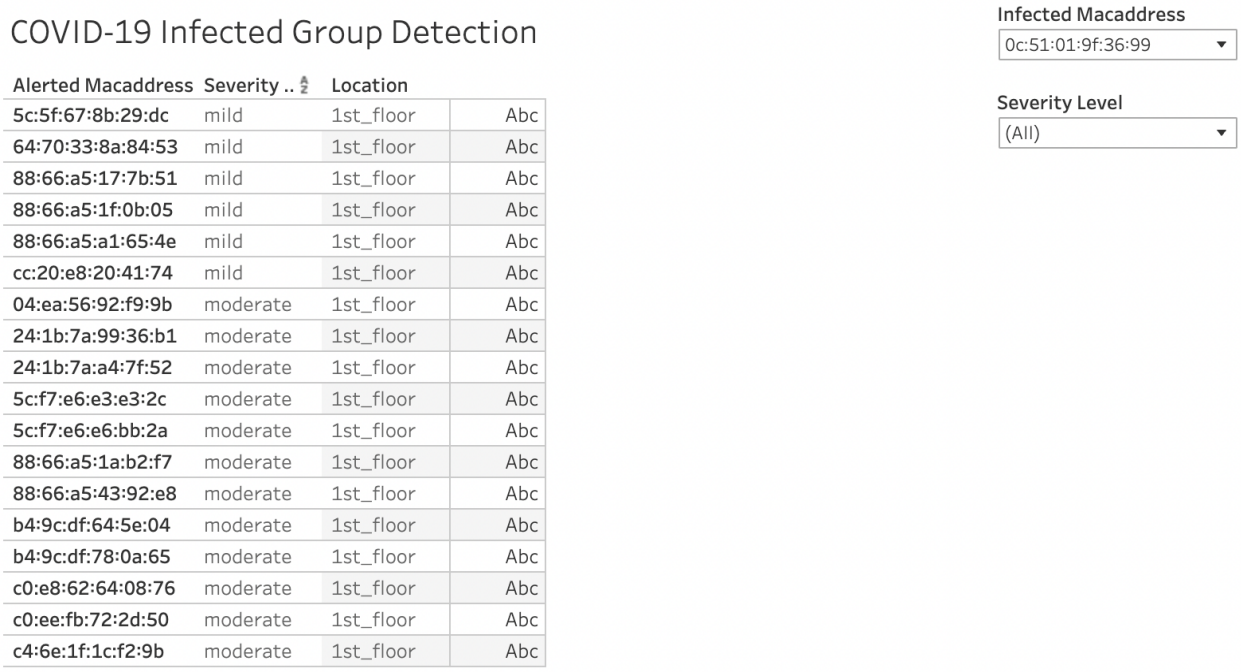


Figure 21: Clustering Comparison Before and After COVID-19 Dashboard

3. Dashboard 3: COVID-19 infected group detection. Users can choose a specific infected mac address to see if other mac addresses may be infected, and listed in a table by severity level. The severity level is defined by the distance of other mac addresses contacting the infected one. We set a range of different contact distances and separate in three severity levels which are mild, moderate and severe respectively. Users also can choose to view the results by filtering by severity levels using the sidebar.



Alerted Macaddress	Severity ..	Location	
5c:5f:67:8b:29:dc	mild	1st_floor	Abc
64:70:33:8a:84:53	mild	1st_floor	Abc
88:66:a5:17:7b:51	mild	1st_floor	Abc
88:66:a5:1f:0b:05	mild	1st_floor	Abc
88:66:a5:a1:65:4e	mild	1st_floor	Abc
cc:20:e8:20:41:74	mild	1st_floor	Abc
04:ea:56:92:f9:9b	moderate	1st_floor	Abc
24:1b:7a:99:36:b1	moderate	1st_floor	Abc
24:1b:7a:a4:7f:52	moderate	1st_floor	Abc
5c:f7:e6:e3:e3:2c	moderate	1st_floor	Abc
5c:f7:e6:e6:bb:2a	moderate	1st_floor	Abc
88:66:a5:1a:b2:f7	moderate	1st_floor	Abc
88:66:a5:43:92:e8	moderate	1st_floor	Abc
b4:9c:df:64:5e:04	moderate	1st_floor	Abc
b4:9c:df:78:0a:65	moderate	1st_floor	Abc
c0:e8:62:64:08:76	moderate	1st_floor	Abc
c0:ee:fb:72:2d:50	moderate	1st_floor	Abc
c4:6e:1f:1c:f2:9b	moderate	1st_floor	Abc

Figure 22: COVID-19 Infected Group Detection Dashboard

Demo

Please watch our demo at https://youtu.be/IVmLc4feiew. The first one is floor-based WiFi connection prediction. The above one is a prediction, and we extended our model to predict the future 19 days so the IT department can plan ahead based on the dashboard. The goal of this dashboard is for the IT department to check WiFi connection and provide a stable network during peak time. The second dashboard is the gathering areas comparison before and after COVID-19. We will see fewer people gathering after the pandemic starts, we will expect to have more after covid data, but due to the long pandemic season, we can have different COVID-19 stage comparisons. This dashboard aims to help the security department work more efficiently, which means the security guard can check the gathering of people and should pay more attention to this area rather than walking around randomly. The last dashboard alerts the clients if one of them gets infected. Also, users can filter by their severity levels, and the security department will be responsible for sending the alert to keep it safe and confidential and protect clients' privacy.

Reference

Roy Gelbard, Orit Goldman, and Israel Spiegler. 2007. Investigating diversity of clustering methods: An empirical comparison. Data Knowl. Eng. 63, 1 (October, 2007), 155–166. DOI:https://doi.org/10.1016/j.datak.2007.01.002

Zhe Wang, Tianzhen Hong, Mary Ann Piette, and Marco Pritoni. Inferring occupant counts from wi-fi data in buildings through machine learning. Building and Environment, 158:281–294, 2019.

Mu Mu. Wifi-based crowd monitoring and workspace planning for COVID-19 recovery. CoRR, abs/2007.12250,2020 https://arxiv.org/abs/2007.12250