

History Flow Analysis in WiFi-based Environments

--EE 5003 project

Fei Yang

(A0169096N)

Department of Electrical & Computer Engineering, NUS

In partial fulfillment of the requirements for the Degree of Master of Engineering

National University of Singapore

12 April 2018

Abstract

In this report it proposes a history flow analysis method and designs a friendly application which can automatically download, process data and show result. By employing Python3.6, MySQL5.7, Tableau, the application shows good performance for NUS WiFi history flow analysis.

This report mainly analyzes the historical flow of NUS buildings from user dwell time and flow rate. It has addressed the download inconvenience from datacommons website and extract problem from a week's compressed file. In order to improve the read and write speed, the application uses MySQL to store data. To improve result presentation, it applies Tableau to display line chart of flow density, pie chart of user type. In addition, it also utilizes the python tkinter GUI to make this application more friendly.

It has provided a detailed method of python, including designing UI, processing csv file, achieving analysis algorithm and operating MySQL. It also demonstrated basic operation of MySQL and Tableau. This detailed process serves as a guide for researchers to continue optimizes our application. Last but not least, it will get better performance with more analysis algorithms and more optimized data structure. And this more accuracy result can be used to estimate building occupancy and utilization to do management for saving energy.

Declaration

I hereby declare that this report is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the report.

This report has also not been submitted for any degree in any university previously.

Fei Yang

13 April 2018

Acknowledgement

In this project I get support from many individuals. Therefore, I would like to take this opportunity to express my profound gratitude to all of them.

First of all, I want to express my special gratitude to my supervisor, Prof Lawrence Wong, Department of Electrical and Computer Engineering, National University of Singapore, for his constant encouragement and great support all the time. I benefit a lot from his excellent guidance and valuable suggestions on my master project.

Secondly, I am very grateful to Huang Jinliang Raymond who has been devoting his time and knowledge to providing me with access to the related resources on my master project.

Thirdly, I also want to express my thanks to Soh Hock Heng. He gave me some useful advice on the implementation of the project and shared his experience of life.

Last but not least, I wish to take this opportunity to thank my parents and friends for their great encouragement, support and comforting help provided throughout the process of the thesis writing.

Table of Content

	Page
Abstract.....	i
Declaration.....	ii
Acknowledgement	iii
Table of Content.....	iv
List of Figures.....	vi
List of Tables	vii
List of Abbreviations	viii
Chapter 1 Introduction.....	1
1.1 Background	1
1.2 Motivation and Challenges.....	1
1.3 Data structure	2
1.4 Tools introduction	5
Chapter 2 Data Preprocess.....	7
2.1 Introduction of UI.....	7
2.2 Download Data.....	8
2.3 Data preprocess	10
2.4 Improvement in user type.....	11
Chapter 3 Data Analysis.....	15
3.1 Generate time for mac address count	15
3.2 Flow Rate	16
3.3 Dwell time	19
3.4 Reduce fluctuation.....	22
Chapter 4 The Results Show	23
4.1 Write data into MySQL	23
4.2 Data visualization	24
4.3 Experimental demonstration.....	29
Chapter 5 Conclusion	34

Reference.....	35
-----------------------	-----------

List of Figures

Figure1. 1 General flow chart	2
Figure1. 2 Python icon	5
Figure1. 3 MySQL icon	6
Figure1. 4 Tableau icon	6
Figure2. 1 Friendly UI	7
Figure2. 2 Data Commons	8
Figure2. 3 Download process.....	9
Figure2. 4 Unification process	11
Figure2. 5 Improve user type based on history	12
Figure2. 6 The pie chart before improvement.....	13
Figure2. 7 The pie chart after improvement	14
Figure3. 1 Flow chart of generating time.....	16
Figure3. 2 Flow chart of calculating rate	18
Figure3. 3 Detail of calculating inflow and outflow	19
Figure3. 4 Flow chart of calculating dwell time	21
Figure3. 5 Dwell time of E1 6floor.....	21
Figure3. 6 Dwell time after improvement.....	22
Figure4. 1 Whole process of reslut show	23
Figure4. 2 Whole process of data write	24
Figure4. 3 Initial page of Tableau	25
Figure4. 4 Tooltip in Tableau	26
Figure4. 5 Line chart of library	26
Figure4. 6 Pie chart of library	27
Figure4. 7 Type line chart of library	28
Figure4. 8 Connect to shared floder.....	29
Figure4. 9 File structure	30
Figure4. 10 Application interface	30
Figure4. 11 Line chart of E1	31
Figure4. 12 Pie chart of E1	32
Figure4. 13 Type line chart of E1	32

List of Tables

Table1. 1 Data structure of localization records	3
Table3. 1 Raw data.....	17
Table3. 2 Modified data for devices count	17
Table3. 3 First located time.....	20

List of Abbreviations

Abbreviation	Description
WLAN	Wireless local area network
GPS	Global positioning system
RSS	Received signal strength
LOS	Line of sight
GUI	Graphical user interface

Chapter 1 Introduction

1.1 Background

There is energy waste in many buildings today. For example, after class there is a few students in the classroom, but the lights and air conditioning are already over-supply, it is possible to save energy by modification if we know the number of students. Manual counting is the traditional way to count the number of people. However, it requires high labor costs and there is a problem that accuracy is decreased under a crowded environment. Therefore, there have proposed a various methods for counting people or estimating the degree of congestion automatically by using cameras, sensors, and devices attached on people such as RFID and smart phones.[1]

On the other hand, with the popularity of smart phones, more and more buildings have WiFi. For example, shopping malls, office buildings, airports, subway stations, and campuses all have WiFi coverage. There are more and more WIFI access point and served for more and more clients in recent years. In NUS campus, it has constructed more than 4,000 access point in campus to provide Internet connection.

With this development of technology, to estimate the level of utilization in a particular environment is available based on WIFI infrastructure. If this can be achieved accurately and cost-effectively, such metrics can be very useful for a variety of buildings to save energy.

1.2 Motivation and Challenges

This project aims to develop a flow analysis system that operates in a WIFI-based environment using access point monitoring number of devices like phone, laptop. we have considered to use the localization information by CISCO system which are on received signal strength indication (RSSI) at devices from an WIFI access point for estimating the number of people in an area. The information on RSSI provided by CISCO WIFI access point will provide the localization of devices by longitude and latitude as well as some extra information like localization time etc. And in this research, we assume that every people take one device that can be localization by access point. We use this existing WIFI networks for estimating the number of people. Therefore, low-cost people counting can be implemented.

In this report, as the first step of our research, we design a friendly UI and propose methods to automatically download and extract csv file from datacommons in chapter 2. Then to preprocess data and analyze data to get user dwell time and flow rate in chapter 3. Finally, in chapter 4, to write data into MySQL and use Tableau connecting MySQL to show the result. You can see the whole process in the following figure:

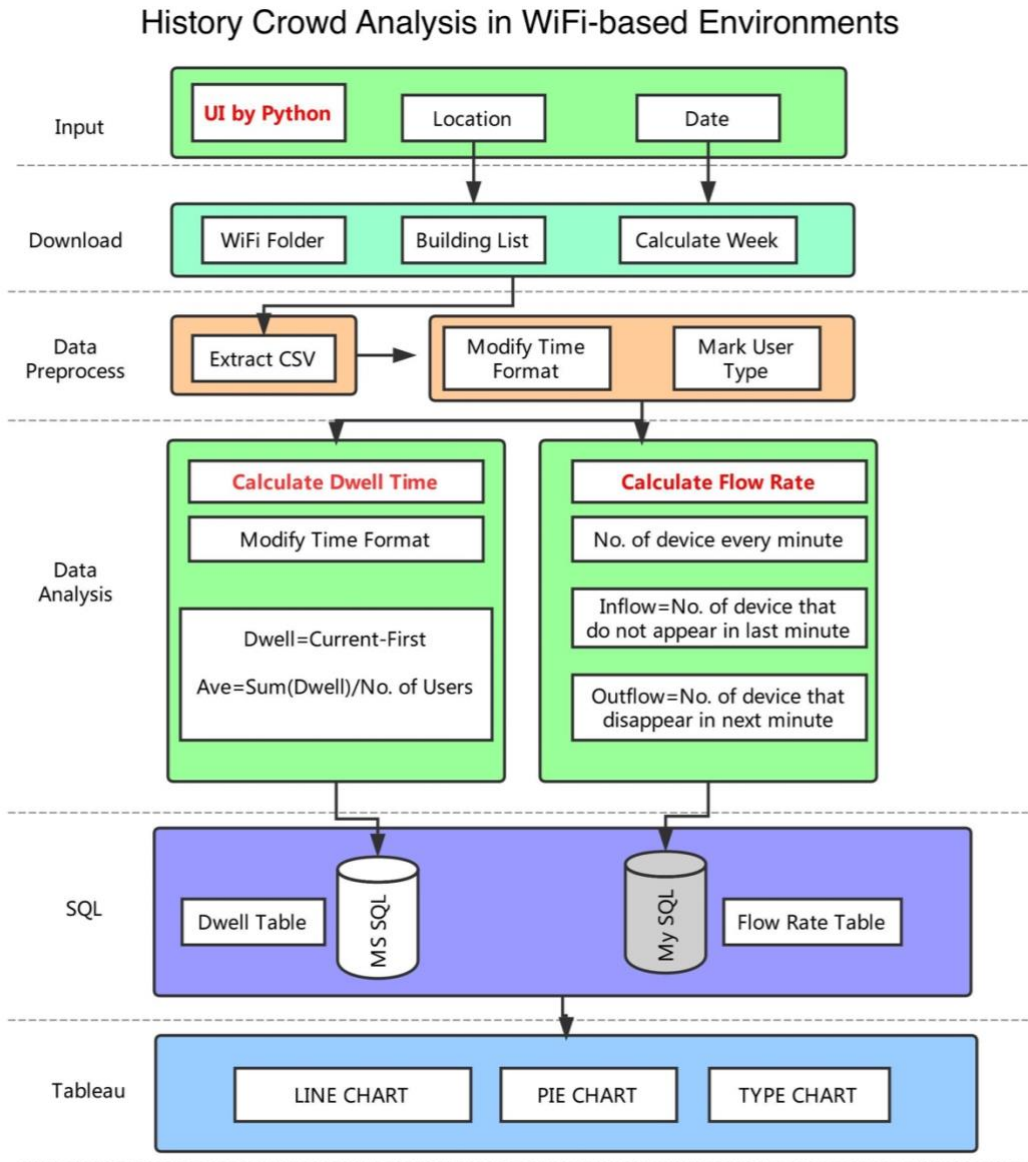


Figure1. 1 General flow chart

1.3 Data structure

A rough introduction of data structure and situations are as following:

- 1) Data structure

For this project we use history data (which means devices records in previous times of the same places) from NUS datacommens CISCO localization system. These data are as following:

time	macaddress_hash	isTracked	confidence	ip	username	ssid	band
20180318-00:00:01	27da2ea94f14725a4b0b138008ad1f27ddd10c1e	TRUE	56	172.17.113.140	ec77bc57401fe8cb856f888e3b0e55a0ffbd85a5	NUS_STU_2-4GHz	UNKNOWN
20180318-00:00:01	e733670253f4a7e5b52541571f6c1bcd3f5f5dc2	TRUE	32	172.17.60.162	ec77bc57401fe8cb856f888e3b0e55a0ffbd85a5	NUS_STU	UNKNOWN
20180318-00:00:01	ce7678fdcafa2c7a7a4b6220e4b4849e7dc5ffb5	TRUE	488				UNKNOWN
20180318-00:05:01	a4fc1587135608cc7d25668377619ff61ee97600	TRUE	448				UNKNOWN
20180318-00:05:01	07d8bc41615aafba81b4725f44b297b1dec409d1	TRUE	160				UNKNOWN
20180318-00:05:01	c2896ed29c4ec8f0a031a620b48dfdee1c5687c	TRUE	128	172.17.115.125	d74840836783d8c46a77409d19727f98689cdd11	NUS_STU_2-4GHz	UNKNOWN
20180318-00:05:01	0d26cb7fd9f29017830ef3e855abd7c8c80e1b82	TRUE	424				UNKNOWN
20180318-00:05:01	218c75c810b61775914a0c3cd6d1f920c298709f	TRUE	440				UNKNOWN
20180318-00:05:01	1439d192f1d6c0058b82f417fd95bc3be11f8e7d	TRUE	480				UNKNOWN
20180318-00:05:01	6ac8610f7f813646cc210bac99198641470ef46	TRUE	448				UNKNOWN
20180318-00:05:01	959c858dc33fd0fab368f5e9b4e8dc5e819705	TRUE	448				UNKNOWN
20180318-00:05:01	f81c802899c719fe48adf35674134df9be7d5945	TRUE	480				UNKNOWN
20180318-00:05:01	8ee62d306bd258fde30a74738097e4b9e95f6399	TRUE	504				UNKNOWN
20180318-00:05:01	796bbaac48fb74ef947733c6bc4d2bcbf1206266	TRUE	56		7af94e72e5a5b5641ac065fe859a2e54b5066959	NUS_STU	UNKNOWN

apMacAddress	isGuest	dot11Status	mapX	mapY	currentServerTime	firstLocatedTime	lastLocatedTime	latitude	longitude
92d95d5ae40de34b5327b848202098fa92575f4f	FALSE	ASSOCIATED	199.16	292.99	2018-03-18T00:00:03.221+0800	2018-03-17T23:43:34.276+0800	2018-03-17T23:43:34.277+0800	1.29844511	103.771288
92d95d5ae40de34b5327b848202098fa92575f4f	FALSE	ASSOCIATED	205.63	287.66	2018-03-18T00:00:03.221+0800	2018-03-17T22:32:44.042+0800	2018-03-17T23:55:21.948+0800	1.29845972	103.771306
	FALSE	PROBING	125.75	155.05	2018-03-18T00:00:03.221+0800	2018-03-17T21:33:57.439+0800	2018-03-17T23:51:46.874+0800	1.29882312	103.771087
	FALSE	PROBING	121.44	148.63	2018-03-18T00:05:09.127+0800	2018-03-18T00:03:16.509+0800	2018-03-18T00:03:16.511+0800	1.2988407	103.771075
	FALSE	PROBING	216.46	292.15	2018-03-18T00:05:09.127+0800	2018-03-17T21:33:23.070+0800	2018-03-18T00:03:14.812+0800	1.29844743	103.771336
803acd1979f3fb6cec9dec510491520a3cc21e4e	FALSE	ASSOCIATED	218.02	304.31	2018-03-18T00:05:09.127+0800	2018-03-17T22:24:43.301+0800	2018-03-18T00:05:07.061+0800	1.29841409	103.77134
	FALSE	PROBING	109.24	140.89	2018-03-18T00:05:09.127+0800	2018-03-17T23:56:22.370+0800	2018-03-17T23:56:22.372+0800	1.29886192	103.771042
	FALSE	PROBING	185.2	266.04	2018-03-18T00:05:09.127+0800	2018-03-18T00:04:36.902+0800	2018-03-18T00:04:36.902+0800	1.29851897	103.77125
	FALSE	PROBING	125.7	154.95	2018-03-18T00:05:09.127+0800	2018-03-17T23:48:37.861+0800	2018-03-17T23:48:37.861+0800	1.29882338	103.771087
	FALSE	PROBING	116.43	144.63	2018-03-18T00:05:09.127+0800	2018-03-18T00:00:17.563+0800	2018-03-18T00:00:17.564+0800	1.29885168	103.771061
	FALSE	PROBING	114.7	145.61	2018-03-18T00:05:09.127+0800	2018-03-17T23:52:33.086+0800	2018-03-18T00:03:10.477+0800	1.298849	103.771057
	FALSE	PROBING	125.39	148.54	2018-03-18T00:05:09.127+0800	2018-03-17T23:27:29.306+0800	2018-03-18T00:03:20.529+0800	1.29884095	103.771086
	FALSE	PROBING	130.36	163.74	2018-03-18T00:05:09.127+0800	2018-03-18T00:00:47.714+0800	2018-03-18T00:00:47.714+0800	1.29879931	103.771099
92d95d5ae40de34b5327b848202098fa92575f4f	FALSE	ASSOCIATED	206.66	279.86	2018-03-18T00:05:09.127+0800	2018-03-18T00:04:36.904+0800	2018-03-18T00:04:59.021+0800	1.29848109	103.771309

Table1. 1 Data structure of localization records

First, we need to get knowledge of this data columns meaning. Refer to the CISCO Community the columns meaning is as following:

1. The first column ‘time’ is the time that the server records the data. This column depends on the server time stamp and isn’t totally same as current time. That means when a device is localized by CISCO WIFI system, the information of this device may cannot be updated immediately in system space ;
2. The second column ‘macaddress_hash’ is the hashed mac address of device by SHA encryption. For privacy of devices owner, we use SHA encryption to encode the MacAddress of device which is impossible to decode directly;
3. The third column ‘isTracked’ is the state whether the device can be tracked by server;
4. The fourth column ‘confidence’ is generate by CISCO system. With every calculated location (say x1, y1), a confidence factor (err_ft) is returned. This is a floating point scalar used to calculate 95% confidence square. The device is estimated to be inside the square centered at (x1, y1) with sides 2 x (err_ft) with 95% confidence;
5. The fifth column ‘ip’ is the IP4 address of devices;

6. The sixth column 'username' is the account number of device;
7. The seventh column 'ssid' is the identity of device, containing: NUS, NUS_2-4GHz, NUS_STU, NUS_STU_2-4GHz, NUSOPEN, NUS_Guest, eduroam and blank. The blank means CISCO system fails to identity the network by too short communication time;
8. The eighth column 'band' is used in CISCO system now;
9. The ninth column 'apMacAddress' is the access point that localized the device. Its value is the nearest AP mac address from device and hashed with SHA encryption same as device mac address;
10. The tenth column 'isGuest' is whether its identity is guest;
11. The 11st column 'dot11Status' indicates connection status between NUS WIFI and devices. It contains 3 states: UNKNOWN, PROBING, which means the device can't communicate by NUS WIFI, but it is connected to NUS WIFI. ASSOCIATED, which means the device can communicate by NUS WiFi. Only in ASSOCIATED status, can CISCO system get ip, username, ssid, apMacAddress of devices;
12. The 12st 'mapX' is the offset of image in longitude in inches, which is calculated by device longitude and image longitude;
13. The 13st 'mapY' is the offset of image in latitude in inches, which is calculated by device latitude and image latitude;
14. The 14st column is the 'currentServerTime' which shows the time when the record package update to server;
15. The 15st column is the 'firstLocatedTime' which shows the time when this record generated or when the device is first associated with CISCO WIFI system. It will be regenerated if a device lose connection with WIFI for one hours;
16. The 16st column is the 'lastLocatedTime' which shows the time when the record last updated. The column will change in several situations as following:
 - (a) The device move beyond the threshold of WIFI system;
 - (b) The device move from the zone of an AP to another;

- (c) The localization record interval between `currentServerTime` and `lastLocatedTime` time exceeds 15 minutes;
- 17. The 17st is the localization data in latitude;
- 18. The 18st is the localization data in longitude.

1.4 Tools introduction

1) Python

For programming beginner, Python is a good choice because of its friendly, concise, easy-to-learn features. Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. The Python language has a rich library and powerful features for processing data.[2]

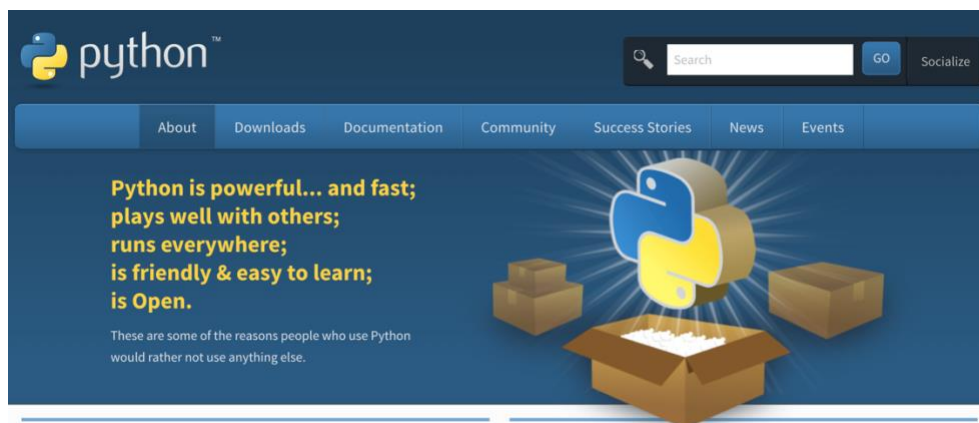


Figure1. 2 Python icon

2) MySQL

In the past, MySQL has become the most popular open source database because of its high performance, low cost, and high reliability. Therefore, it has been widely used in small and medium-sized websites on the Internet. As MySQL continues to mature, it is gradually used for more large-scale websites and applications.

In this project, at the beginning we use csv file to store data which capture from NUS WIFI. But when using Tableau to read csv file, we find the speed is too low to tolerate. So we use MySQL to store our data, and Tableau shows great performance

cooperating with MySQL.



Figure1. 3 MySQL icon

3) Tableau

Tableau Software is a software company that produces interactive data visualization products that focus on business intelligence.[3] Students can apply for free education version for one year. It is an efficient tool for data visualization, we plot line chart and pie chart in this project. Besides, Tableau provides some useful filters and tooltips, which give us great help in select datas.



Figure1. 4 Tableau icon

Chapter 2 Data Preprocess

Before we learn how to apply some algorithms to solve problem above, we need to download the data from datacommons and preprocess data. In this chapter, we introduction the UI and how to automatically download and uncompromising zip file. Finally we present the detail how to deal with the raw data.

2.1 Introduction of UI

As an engineer, when we are designing an application, we must consider that non-technical people can easily use this application. Only presenting a simple and easy-to-understand interface to users, the UI needs to encapsulate the technical details. For this project, we design a friendly UI based on python tkinter just to show users the time and location, which users want to check.

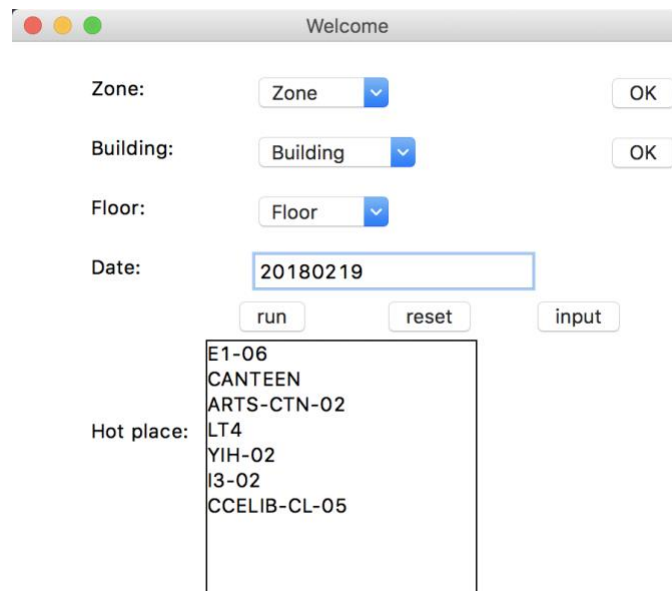


Figure2. 1 Friendly UI

From the above picture, we can see Zone, Building, Floor and Date. For selecting location, we divide NUS campus to several zones, every zone has a wide range of buildings with several floors. When users chose the zones and click ok button on the right side, then to select which building they want to look and click ok button on the right side. Finally deciding the floor and inputting the date. Then clicking the run button, our application will work and you may wait for one minute to get the result. Besides, there is a area named hot place, like canteen, library and LT.

If users want to look information of some hot places, they can select one area and click the input button to run.

2.2 Download Data

We have introduced the NUS WIFI data structure in the chapter 1.3 and then we will talk about how to get these data from databases. There are two ways to download data from university database. The first one is the most convenient way to download from datacommons website. Firstly, going to the website, <https://datacommons.nus.edu.sg/> default and using student account to login. After login successful, you will see the following picture:

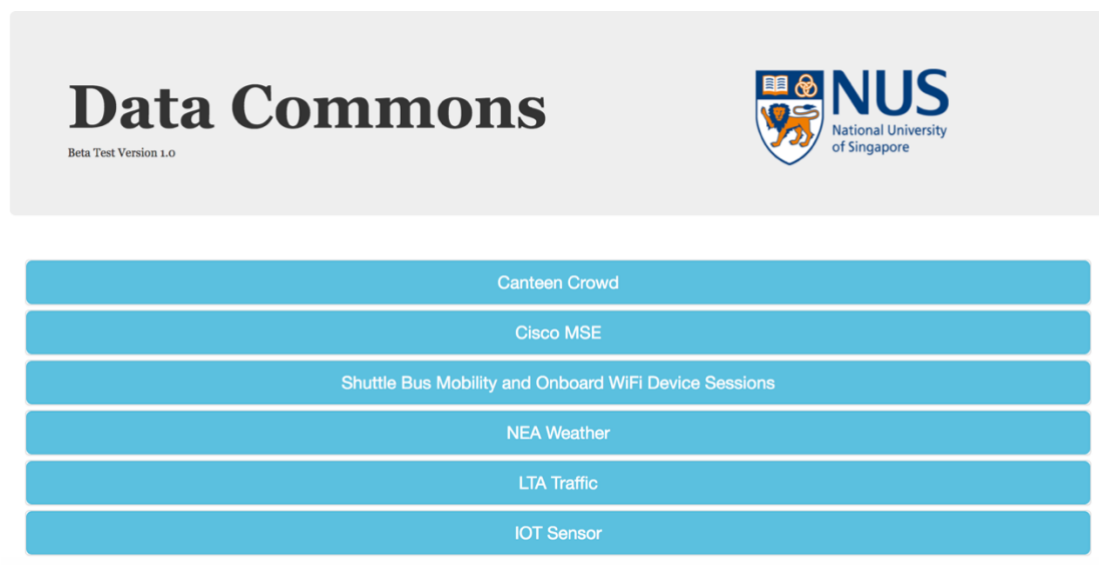


Figure2. 2 Data Commons

The button CISCO MSE is the source of our project data. After clicking the button, you can choose the location and the time period. Finally you download the zip file of data which includes the whole week of that date because our CISCO system processes the data once a week. One week data is compressed into one zip file. Although this method is universal, this method is not feasible for people without an NUS account. And every time the site needs to log in, it is also a waste of time.

So we have another advanced method to get data from data commons. In order to access the database faster, we use the api that staff provides to download the data from share folder. First, you need to connect NUS WiFi to access the shared folder. Then you have to locate the specified building folder and select the date to download the file. As mentioned above, data is packed once every week, so you also need to

convert the date to the corresponding week.

As we all know, a big feature of Python is that it has a lot of packages, so we can call packages that are already mature enough to download and convert dates. The download process looks like this:

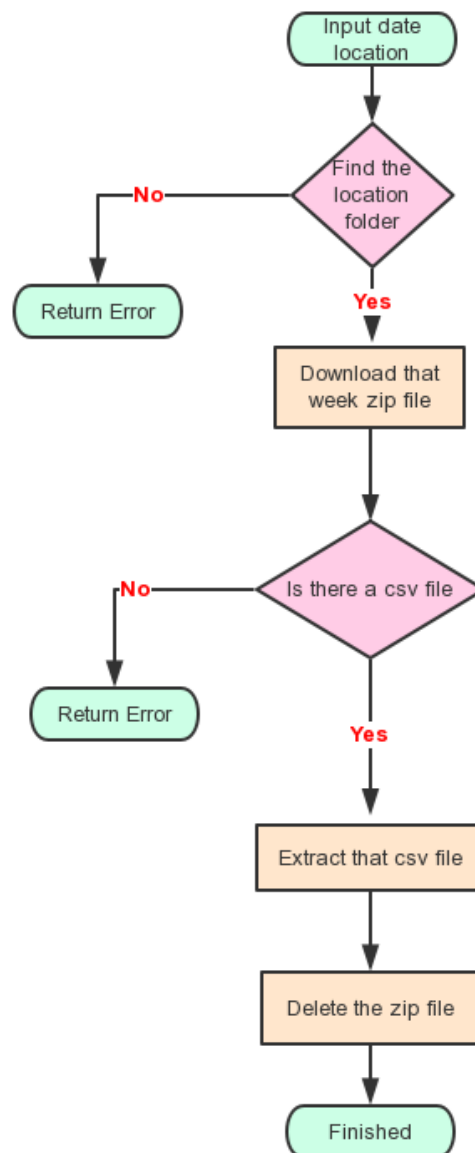


Figure2. 3 Download process

We use python datetime library to convert the date to the corresponding week and the variable workdir is the path of the NUS WIFI share folder. The variable filelist is the building list in NUS campus, you can see we compare the subpath with the filelist, if they are matched, program returns the complete path of that building . If we cannot

find the corresponding folder, results returns error and shows “no this location”.

After downloading the zip file successfully, we need to extract the csv file from the one week compressed packet. After downloading the data, we output “download success”. As for uncompressing, we call the python library zipfile to extract the date_csv, which is the name of the file we get in figure 2.3. Finally, after unzip operation, we delete redundant files and output the “unzip success”. At this point, we have obtained the data we want. The next step is how to preprocess the data.

2.3 Data preprocess

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. [4]

In our project, we need to process the time(current time and first locate time) of the data so that the time format meet to MySQL requirement. In addition, we need to annotate the user type(staff, student, other) according to ssid. Since the CISCO system records once every minute, its seconds per minute are random. In order to unify and calculate the dwell time, we change the seconds of recording time to 01. Because MySQL time format YYYY-MM-DD HH:MM:SS, our program will modify the time format. NUS ssid has 3 types: student using NUS_STU or NUS_STU_2-4GHz, staff using NUS or NUS_2-4GHz, guest using NUSOPEN or NUS_Guest. Specially, if the device is probing rather than associated, we cannot get its ssid. For probing devices, their ssid is blank. So our program need to mark these user type according to the ssid name. The unification process is in following figure:

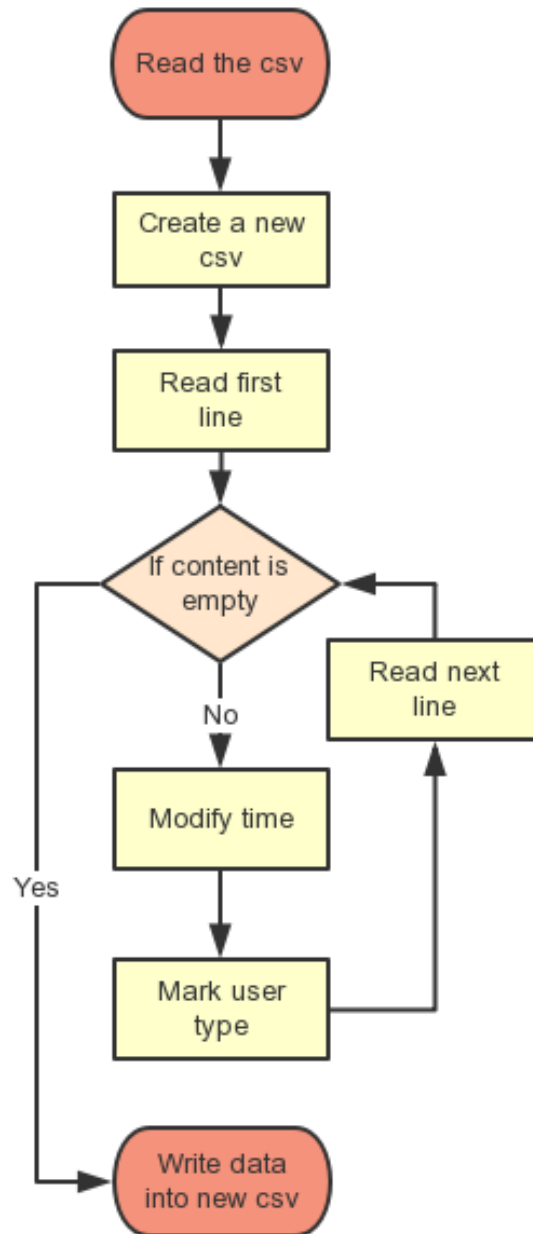


Figure2. 4 Unification process

2.4 Improvement in user type

There is also room for improvement on the user type. For some devices, the CISCO system deletes the device's record from the cache because it left the network for too long or the sleep time exceeded 15 minutes. When the device becomes probing, CISCO cannot identify its ssid. But in fact, based on past records, we can identify the user type of the device, as long as the device was previously successfully identified. In this way, we reduce the error for the user type. Please refer to the figure below for the improvement process:

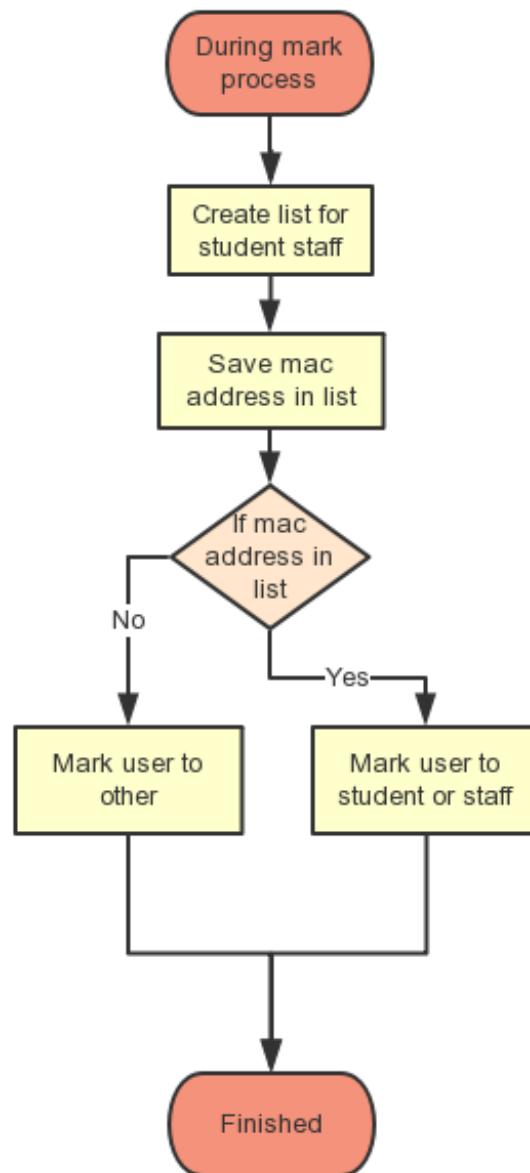
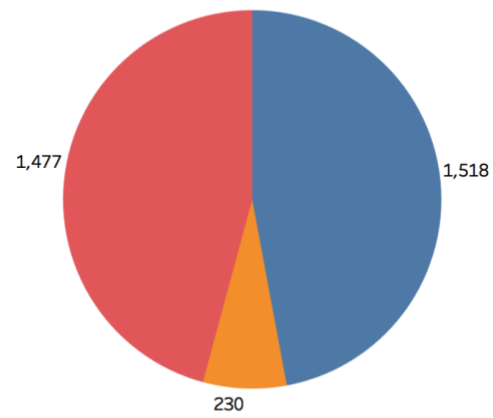


Figure2. 5 Improve user type based on history

We tested and compared the results before and after improvement, as shown in the following figures :

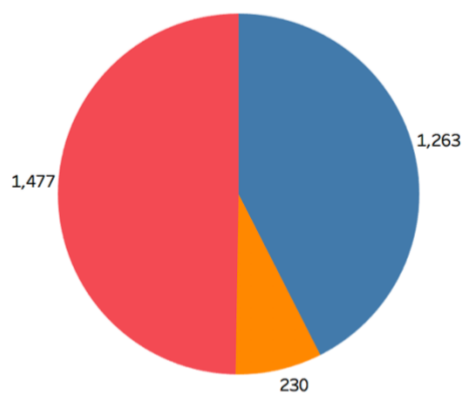
PIE CHART



User Type	Other	Staff	Student
-----------	-------	-------	---------

Figure2. 6 The pie chart before improvement

PIE CHART



User Type	Other	Staff	Student
-----------	-------	-------	---------

Figure2. 7 The pie chart after improvement

We can see from the above two figures, after improvement the Other type number decreases from 1518 to 1263. Because some probing records have been modified to student or staff based on history. As for why other type is reduced, the other two categories do not increase. This is because our statistics are based on the number of unique mac addresses instead of record's number.

So far, our data preprocessing has been completed. We mainly deal with the time format and user type. The next step is to perform data analysis.

Chapter 3 Data Analysis

Data analysis refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements.[6] In this chapter, we mainly show our data analysis based on flow rate and dwell time, which can reflect buildings utilization and occupation.

3.1 Generate time for mac address count

Because the CISCO system records once every minute for connected devices, there are multiple records in one minute (as shown in the table 3.1). For calculation, the form we want to get should be like table 3.2. To adjust the records of multiple devices at the same time to one minute containing multiple devices, we need to calibrate the time.

Initially we count all the time of the day and then delete the duplicated time, but there is a problem: that is, due to the occasional failure of the system, some records will be lost, leading to time discontinuities. In order to solve this problem, we create a function to generate all the time records of a certain day just like the normal system. The code that produces the time series is shown in the figure below:

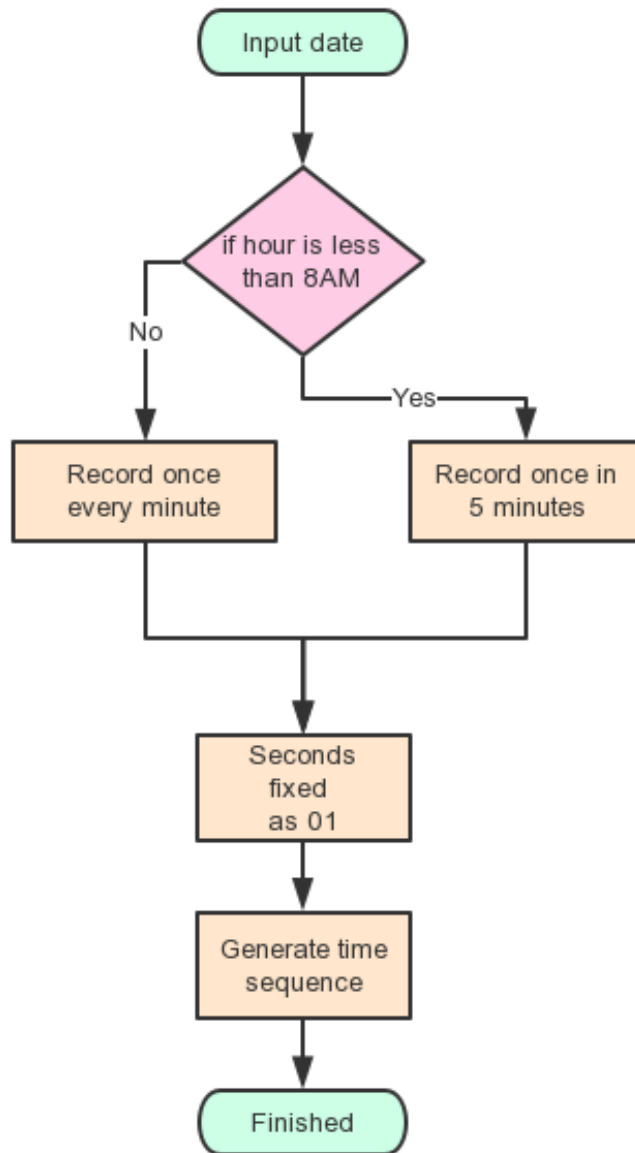


Figure3. 1 Flow chart of generating time

3.2 Flow Rate

Since we have the mac address of the devices in our data, we can calculate the inflow and out flow based on the change in the mac address per minute recorded by the system. Firstly, we should modify our data format and count how many devices every minute. The following two tables show the raw data and processed data:

time	macaddress_hash
4/3/18 00:00	e6b2f09ef58f0fba25fb6da81c62ed27c3778a64
4/3/18 00:00	5a0dbb50bcb193794c031186d4248d9fee213960
4/3/18 00:05	5a0dbb50bcb193794c031186d4248d9fee213960
4/3/18 00:05	6dbbc27a3d8e4d560e444b12e62be1de59060627
4/3/18 00:10	437432a17277c08eb72a671a68ca8fb2569fa15a
4/3/18 00:10	6dbbc27a3d8e4d560e444b12e62be1de59060627
4/3/18 00:15	fde07c6bb366d4b8ea0398719281bef758f2e40a
4/3/18 00:15	437432a17277c08eb72a671a68ca8fb2569fa15a
4/3/18 00:15	6dbbc27a3d8e4d560e444b12e62be1de59060627
4/3/18 00:20	fde07c6bb366d4b8ea0398719281bef758f2e40a
4/3/18 00:20	392051312477dda9a5e35c58344b767f5bc7acc5
4/3/18 00:20	437432a17277c08eb72a671a68ca8fb2569fa15a
4/3/18 00:20	6dbbc27a3d8e4d560e444b12e62be1de59060627
4/3/18 00:25	37bbe42a8999891419c1ed5431439811bdc84ddf
4/3/18 00:25	437432a17277c08eb72a671a68ca8fb2569fa15a

Table3. 1 Raw data

time	No.						
4/3/18 00:00	2	e6b2f09ef58f0fba2	5a0dbb50bcb193794c031186d4248d9fee213960				
4/3/18 00:05	2	5a0dbb50bcb19379	6dbbc27a3d8e4d560e444b12e62be1de59060627				
4/3/18 00:10	2	437432a17277c08e	6dbbc27a3d8e4d560e444b12e62be1de59060627				
4/3/18 00:15	3	fde07c6bb366d4b8	437432a17277c08eb72a671a68ca8fb2569fa15a	6dbbc27a3d8e4d560e444b12e62be1de59060627			
4/3/18 00:20	4	fde07c6bb366d4b8	392051312477dda9a5e35c58344b767f5bc7acc5	437432a17277c08eb72a671a68ca8fb2569fa15a	6dbbc27a3d8e4d560e444b12e62be1de59060627		
4/3/18 00:25	3	37bbe42a89998914	437432a17277c08eb72a671a68ca8fb2569fa15a	6dbbc27a3d8e4d560e444b12e62be1de59060627			
4/3/18 00:30	3	0268ed606086ddb1	437432a17277c08eb72a671a68ca8fb2569fa15a	6dbbc27a3d8e4d560e444b12e62be1de59060627			
4/3/18 00:35	4	fde07c6bb366d4b8	1bb589e4b8	437432a17277c08eb72a671a68ca8fb2569fa15a	6dbbc27a3d8e4d560e444b12e62be1de59060627		
4/3/18 00:40	2	437432a17277c08e	6dbbc27a3d8e4d560e444b12e62be1de59060627				
4/3/18 00:45	3	35b8534965798cfd	437432a17277c08eb72a671a68ca8fb2569fa15a	6dbbc27a3d8e4d560e444b12e62be1de59060627			
4/3/18 00:50	3	35b8534965798cfd	437432a17277c08eb72a671a68ca8fb2569fa15a	6dbbc27a3d8e4d560e444b12e62be1de59060627			
4/3/18 00:55	2	fde07c6bb366d4b8	6dbbc27a3d8e4d560e444b12e62be1de59060627				
4/3/18 01:00	2	981e4a3e81404a27	6dbbc27a3d8e4d560e444b12e62be1de59060627				
4/3/18 01:05	2	fde07c6bb366d4b8	6dbbc27a3d8e4d560e444b12e62be1de59060627				
4/3/18 01:10	3	fde07c6bb366d4b8	30872062db	981e4a3e81404a276278da80d308dc9a74288e10			
4/3/18 01:15	2	37bbe42a89998914	981e4a3e81404a276278da80d308dc9a74288e10				
4/3/18 01:20	1	6dbbc27a3d8e4d560e444b12e62be1de59060627					

Table3. 2 Modified data for devices count

Because the CISCO system's record is based on every minute, we calculate inflow and outflow rate also based on every minute. As is known to all, the media access control address (MAC address) of the device is a unique identifier assigned to the network interface controller for communication at the data link layer of the network segment. The MAC address is used as the network address for most IEEE 802 network technologies, including Ethernet and Wi-Fi. Therefore, in order to calculate the inflow and outflow rate, we only need to compare the mac address before and after two

minutes (our default initial value is null). The entire calculation process is shown in the figure below:

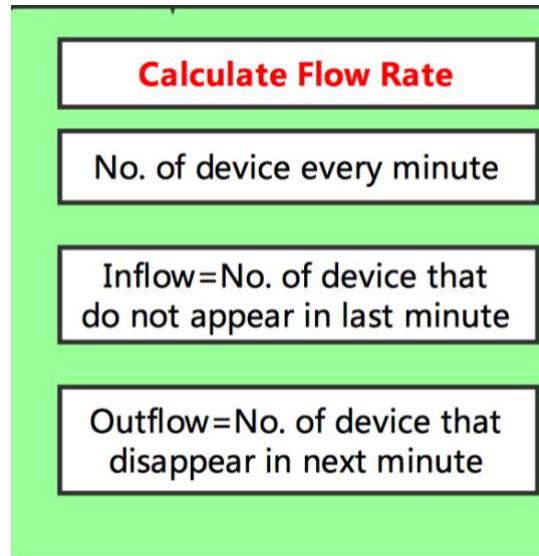


Figure3. 2 Flow chart of calculating rate

The specific idea we already have, the next step is to implement these functions in python. First of all, we want to convert the original data to better compare the mac address, and then compare the mac address between the two records before and after, the new mac address number is inflow, the number of leaving the mac address is outflow. Specific detail process see the figure below:

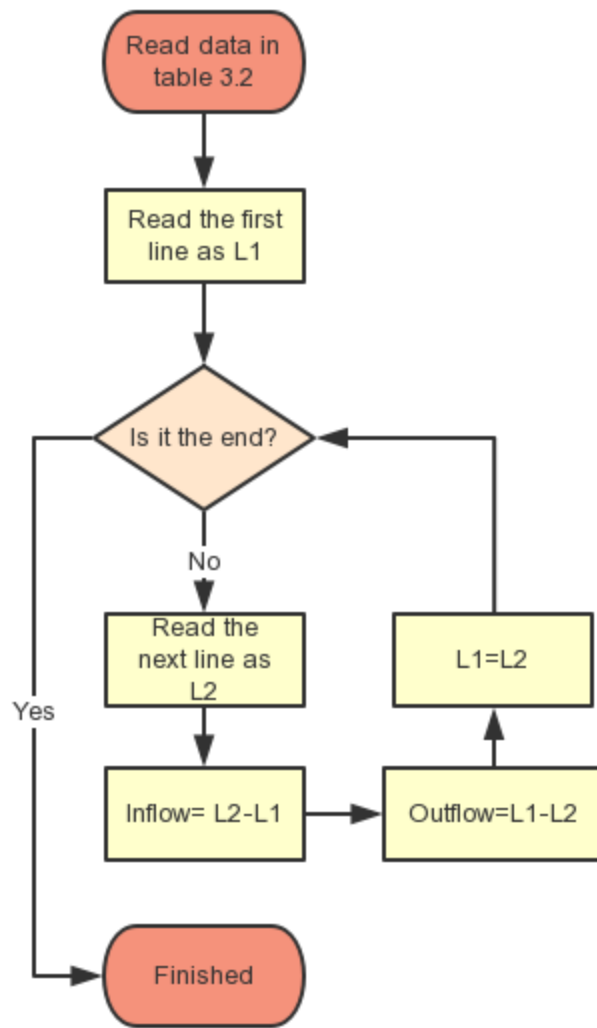


Figure3. 3 Detail of calculating inflow and outflow

From the above figure we can see that `time_mac` is a function that adjusts the data format. `L1` is the previous minute's mac address list, and `L2` is the mac address list for the current minute. `Come` is the variable that counts inflow, and `leave` is the variable that counts outflow. We have got inflow and outflow, the next step is to calculate the dwell time.

3.3 Dwell time

Firstly, there is a question what is dwell time? Dwell time is the time for each device to stay in this environment. The second question, why do we calculate the dwell time? Because the dwell time can reflect the pattern of the device, we can know what kind of behavior patterns people in the environment have, and thus we calculate the utilization and occupancy of the environment.

CISCO's system will automatically record the time for each device when it connects to NUS WIFI from the first time, so we can calculate the time the device stays in that environment. The raw data is shown in the following table:

currentServerTime	firstLocatedTime	lastLocatedTime
2018-03-04T00:00:15.	3/3/18 23:59	2018-03-04T00:00:14.
2018-03-04T00:00:15.	3/3/18 23:58	2018-03-03T23:59:16.
2018-03-04T00:05:11.	3/3/18 23:58	2018-03-03T23:59:16.
2018-03-04T00:05:11.	3/3/18 20:49	2018-03-04T00:05:06.
2018-03-04T00:10:11.	4/3/18 00:03	2018-03-04T00:08:51.
2018-03-04T00:10:11.	3/3/18 20:49	2018-03-04T00:09:15.
2018-03-04T00:15:11.	3/3/18 23:46	2018-03-04T00:14:00.
2018-03-04T00:15:11.	4/3/18 00:03	2018-03-04T00:08:51.
2018-03-04T00:15:11.	3/3/18 20:49	2018-03-04T00:14:30.
2018-03-04T00:20:10.	3/3/18 23:46	2018-03-04T00:14:00.
2018-03-04T00:20:10.	4/3/18 00:18	2018-03-04T00:18:32.
2018-03-04T00:20:10.	4/3/18 00:03	2018-03-04T00:08:51.
2018-03-04T00:20:10.	3/3/18 20:49	2018-03-04T00:19:44.
2018-03-04T00:25:11.	4/3/18 00:23	2018-03-04T00:23:19.

Table3. 3 First located time

We only need to subtract the first located time from the current time to calculate the dwell time. It should be noted that the first located time will be updated more than 15 minutes after the device leaves the network. Therefore, the dwell time we get is more accurate, even if there is a short time when the device is offline. The process for calculating the dwell time is shown in the figure below:

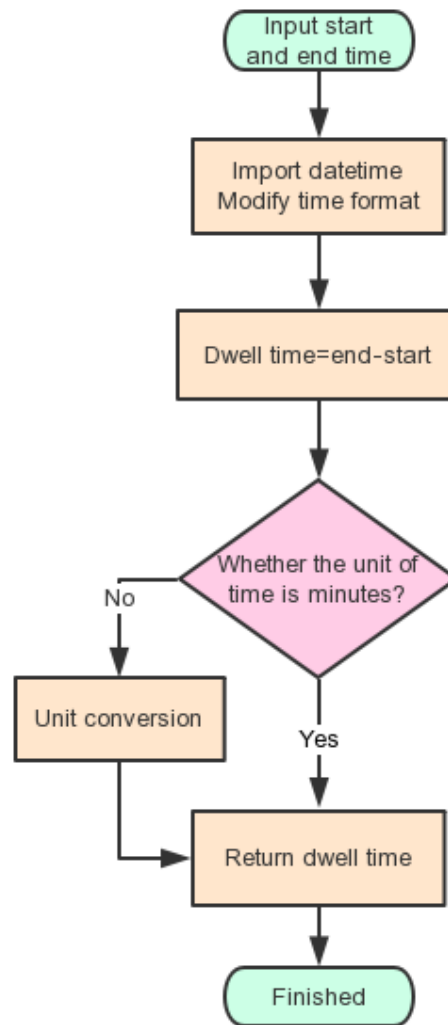


Figure3. 4 Flow chart of calculating dwell time

As you can see from the above figure, we need to adjust the format of time for calculation. Then we call python datetime library to calculate the time difference, and finally convert the result's units to minutes.

Now we just need to average the dwell time. Because tableau has the function of averaging, we don't need to write our own function. The following figure shows the dwell time for a day on E1 6floor:



Figure3. 5 Dwell time of E1 6floor

The above figure shows the dwell time changes over time in a day. We can see that

the dwell time increases with the progress of a lesson. After the class ended, the dwell time quickly declined again.

3.4 Reduce fluctuation

This line is not smooth in figure 3.5. There are many fluctuations, what caused this? After observing the raw data, we found that these fluctuations were caused by fixed devices. At the beginning of the day, the dwell time of the fixed devices is still very small. Over time, their dwell time will increase. When we average dwell time, the weights of the fixed devices will be very large, resulting in huge results. The effect of pulling high averages caused a fluctuation.

We assume that the activities on the NUS campus generally take place at 7am-10pm for a total of 15 hours or 900 minutes. If a device's dwell time in a certain area is greater than 900 minutes a day, then we can think of this device as a fixed device.

The next step is to filter out the fixed devices. At first, we tried to use python to delete the devices whose dwell time was greater than 900 minutes while processing the csv files, but we found that this method is not very efficient. In order to improve efficiency, we use MySQL to operate, and it only takes one command to complete, and the time is much faster. The result of removing fixed devices is shown in the figure below:



Figure3. 6 Dwell time after improvement

As can be seen from the above figure, after the filtration, the trend of the line has not changed, but the line is more smooth and there is no fluctuation.

Chapter 4 The Results Show

We have analyzed and calculated the data. In this chapter we will present the data visualization results at tableau. Data visualization is mainly aimed at the use of graphical means to communicate and communicate information clearly and effectively. [7]

However, this does not mean that the data visualization must be boring because of its functional purpose, or it is extremely complicated to look colorful. In order to effectively convey the concept of ideas, aesthetic forms and functions need to go hand in hand to convey deep insights into rather sparse and complex data sets by intuitively conveying key aspects and features.

However, before the data visualization, we need to write our data into the mysql database. This is because we found through testing that if we read the csv file directly with tableau, the data processing speed is very slow. When we use tableau to connect mysql, the data processing speed is greatly improved. There is a flow chart as following:

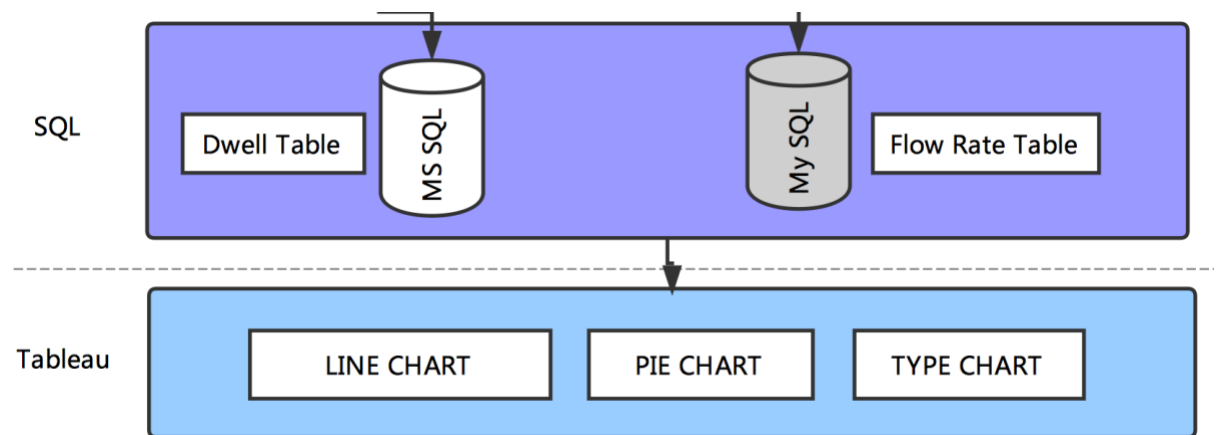


Figure4. 1 Whole process of reslut show

4.1 Write data into MySQL

First we need to create a table in MySQL, define the required attributes, and then use python to read the contents of the csv file and write it to the table. Finally removing the fixed devices in the database. The whole process is shown below:

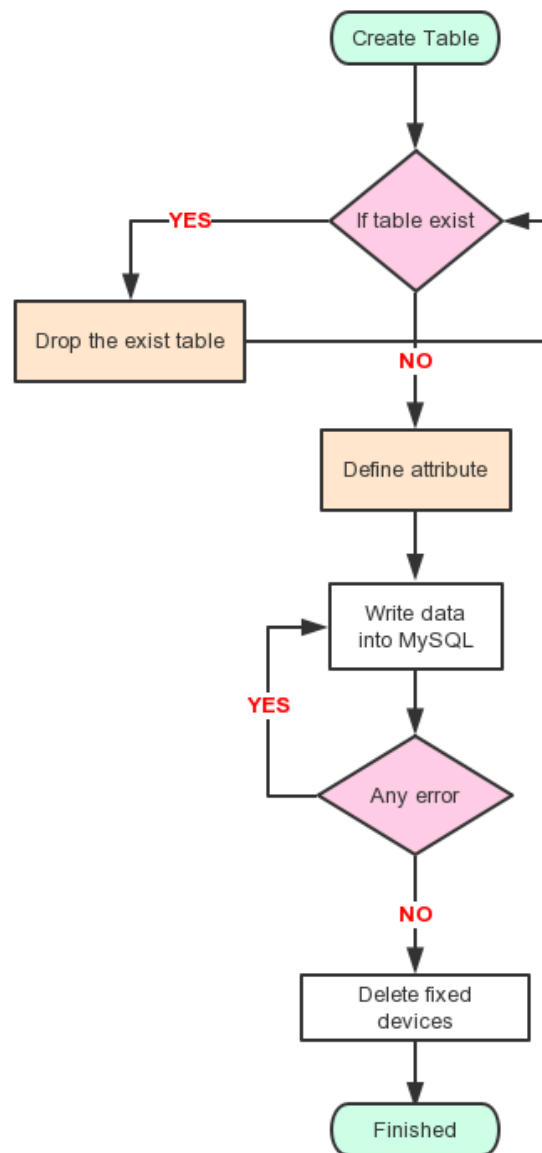


Figure4. 2 Whole process of data write

We implement these functions on Python based on the above flow chart. When finishing writing, our application will output “success write MySQL” .2

After we write the data to MySQL, the final step is to use Tableau to connect to the database for data visualization.

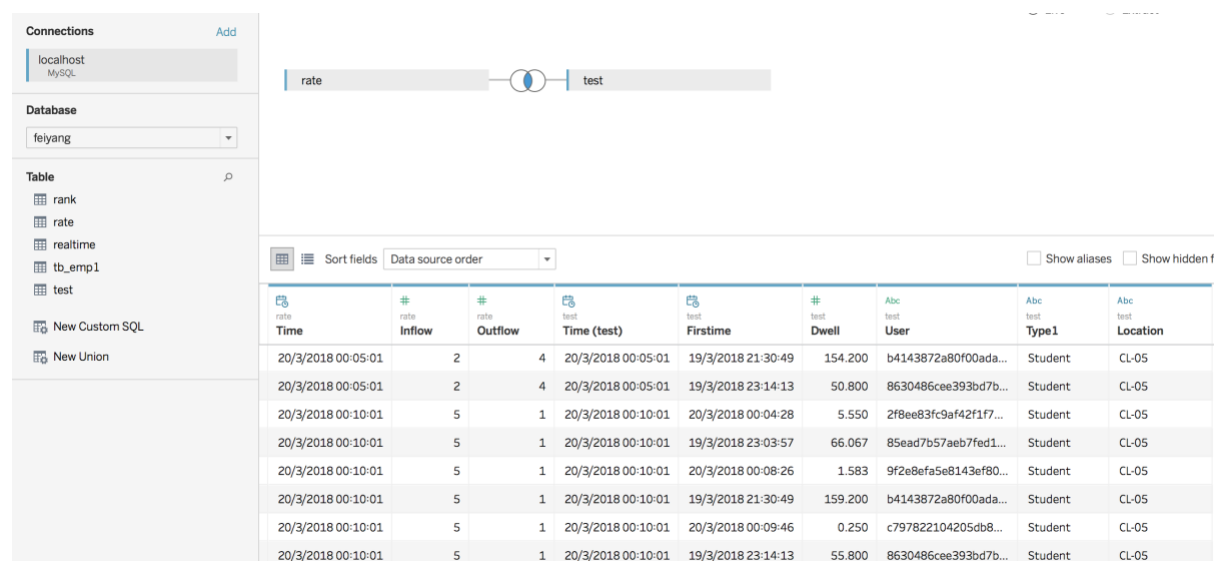
4.2 Data visualization

First of all, why we need data visualization? Because we use the amount of information we obtain visually, it is far more than other senses. People's analysis of image information is more efficient than text, and when the human brain parses the image, it is a complex process. Text parsing is a systematized and systematic process

resulting from acquired education. Therefore, it is difficult to ensure that text can express its information accurately when expressing information. But the image just makes up for its shortcomings. A lot of research has shown that, in the task of comprehension and learning, graphic texts can help readers better understand what they want to learn. Images are easier to understand, more interesting, and easier to remember.[5]

Before we begin, driver is required for MySQL. This connector requires a driver to talk to the database. You might already have the required driver installed on your computer. If the driver is not installed on your computer, when you try to connect, Tableau displays an error message with a link to the [Driver Download](#) page where you can find driver links and installation instructions.

After installing the driver, we can run our tableau and connect with MySQL. The first step is to enter the database password and successfully connect to the database. Selecting the required data in the data source and adjust the data type. For example, the time should be date&time instead of the string type. This modification is for us to display the image correctly. You can see the example below:



The screenshot shows the Tableau interface. On the left, the 'Connections' pane shows 'localhost MySQL' connected. The 'Database' dropdown is set to 'feiyang'. The 'Table' list includes 'rank', 'rate', 'realtime', 'tb_emp1', 'test', 'New Custom SQL', and 'New Union'. The main view displays a table with the following data:

rate Time	# rate Inflow	# rate Outflow	test Time (test)	test Firsttime	# test Dwell	Abc test User	Abc test Type1	Abc test Location
20/3/2018 00:05:01	2	4	20/3/2018 00:05:01	19/3/2018 21:30:49	154.200	b4143872a80f00ada...	Student	CL-05
20/3/2018 00:05:01	2	4	20/3/2018 00:05:01	19/3/2018 23:14:13	50.800	8630486cee393bd7b...	Student	CL-05
20/3/2018 00:10:01	5	1	20/3/2018 00:10:01	20/3/2018 00:04:28	5.550	2f8ee83fc9af42f1f7...	Student	CL-05
20/3/2018 00:10:01	5	1	20/3/2018 00:10:01	19/3/2018 23:03:57	66.067	85ead7b57aeb7fed1...	Student	CL-05
20/3/2018 00:10:01	5	1	20/3/2018 00:10:01	20/3/2018 00:08:26	1.583	9f2e8efa5e8143ef80...	Student	CL-05
20/3/2018 00:10:01	5	1	20/3/2018 00:10:01	19/3/2018 21:30:49	159.200	b4143872a80f00ada...	Student	CL-05
20/3/2018 00:10:01	5	1	20/3/2018 00:10:01	20/3/2018 00:09:46	0.250	c797822104205db8...	Student	CL-05
20/3/2018 00:10:01	5	1	20/3/2018 00:10:01	19/3/2018 23:14:13	55.800	8630486cee393bd7b...	Student	CL-05

Figure4. 3 Initial page of Tableau

Then we create a new worksheet named line chart, selecting dimensions and measures that we want to present and dragging them to the corresponding position. To make the graphics more obvious, we can select the colour of each line in the marks. At the same time, we can also drag the dimensions and measures that need to be filtered into the

filters. Here we mainly filter the time and user type. Because most of the data in the night is of no use to us, user type we mainly focus on students. In order to make the interface friendlier, we can modify the description about dimensions and measures on tooltip. As shown below:

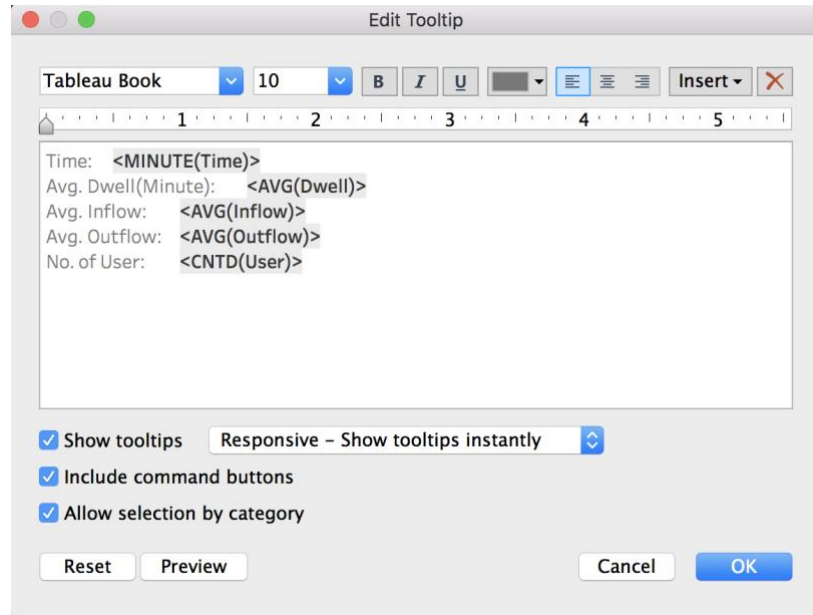


Figure4. 4 Tooltip in Tableau

The following figure shows flow information of the 5th floor in the Central Library on March 20, 2018:

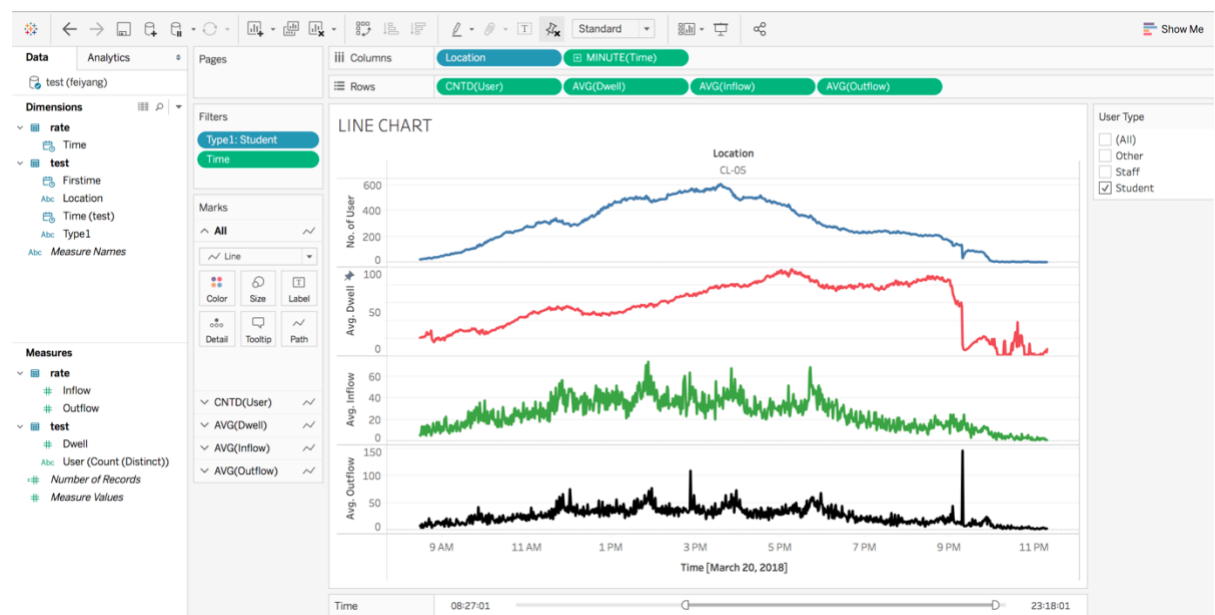


Figure4. 5 Line chart of library

As you can see from the above chart, we only select students and the time interval is from 8am to 12pm. The blue line represents the overall flow of people, the red line represents the dwell time (in minutes), the green line represents the inflow rate, and the black line represents the outflow rate. From this figure we can probably understand the library's flow density and usage on this day.

In order to more clearly see the proportion of flow, we use the pie chart to show the ratio between student, staff and other. The effect chart is as follows:

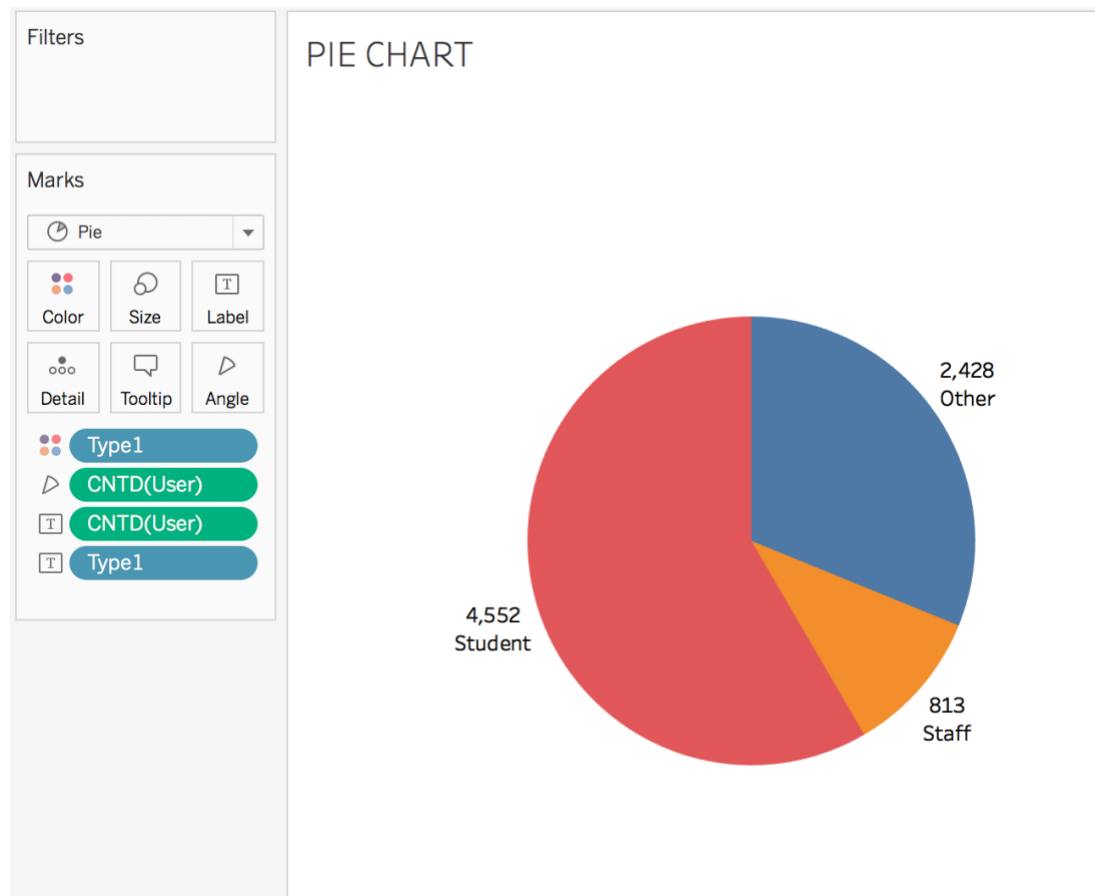


Figure4. 6 Pie chart of library

Looked at the map above, there may be a question why the staff so much? Because we count devices rather than number of people, it is possible that many of the library's public devices are logged in with the staff's ID.

Finally, we will present the type line chart, in order to clearly see the student, staff and other changes over time in one day. The result figure is as follows:

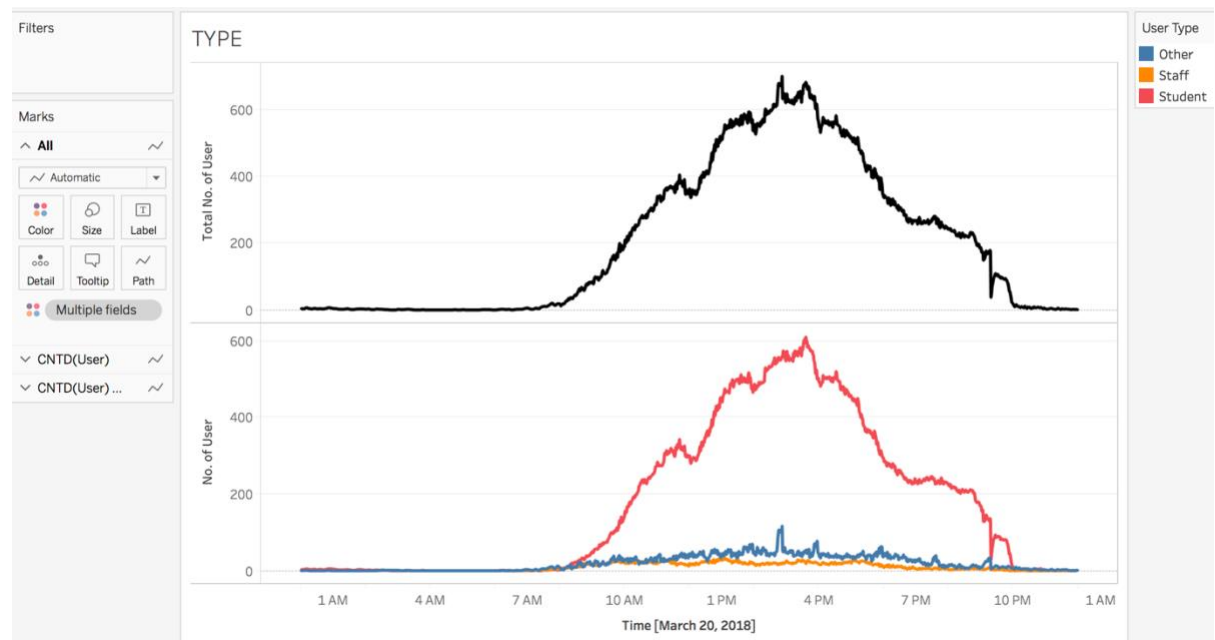


Figure4. 7 Type line chart of library

From the above figure, we can see that the number was 0 at night, and the number started from 8am and gradually increased, and it decreased slightly at 12pm because of lunch. Then the number peaked at 3pm, then the number gradually decreased until the library closures at 10pm when number was 0.

4.3 Experimental demonstration

To better understand this project, we will operate from beginning to end.

1. Connect to a shared folder

The prerequisite is that in the NUS Wi-Fi intranet, use the NUS account to log in to the address `smb://fs9.nus.edu.sg/DATACOMM`. The login interface is as follows:

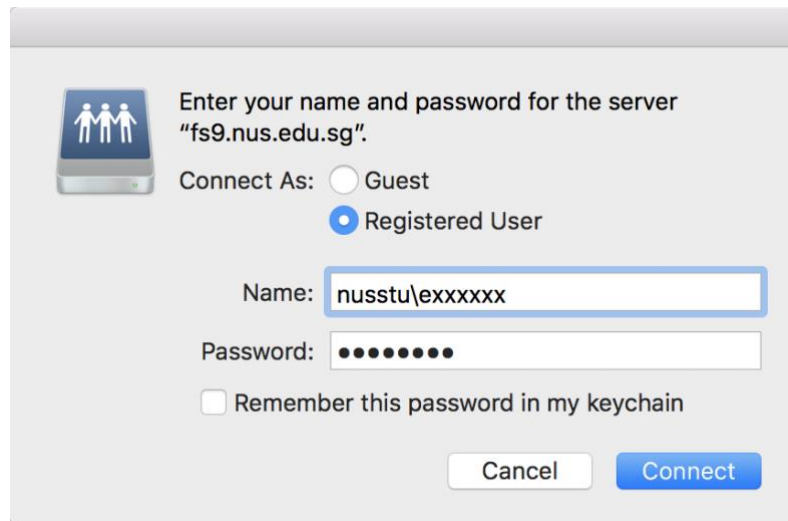


Figure4. 8 Connect to shared floder

After logging in successfully, we can see a folder named nuswifi. Next step is to run our application.

2. Run the application based on python

I have upload my project to my GitHub, so you can click [this](#) to download. When the decompression is completed, you will find a folder named mse sys. This folder is our project, its file structure as shown below:

File Structure

```
| - LICENSE
| - README.md
| - docs
|   |-- feiyang.twb
|   |-- System Architecture Design.pdf
|   |-- quickstart.md
| - main.py
|   |-- __init__.py
|   |-- mysqlwrite.py
|   |-- -- difftime.py
|   |-- copy.py
|   |-- uniform.py
|   |-- main_rate.py
|   |-- --time_mac.py
|   |-- -- -- generate_time.py
| - requirements.txt
```

Figure4. 9 File structure

Just open main.py and we can start our application. The initialization interface is as follows:

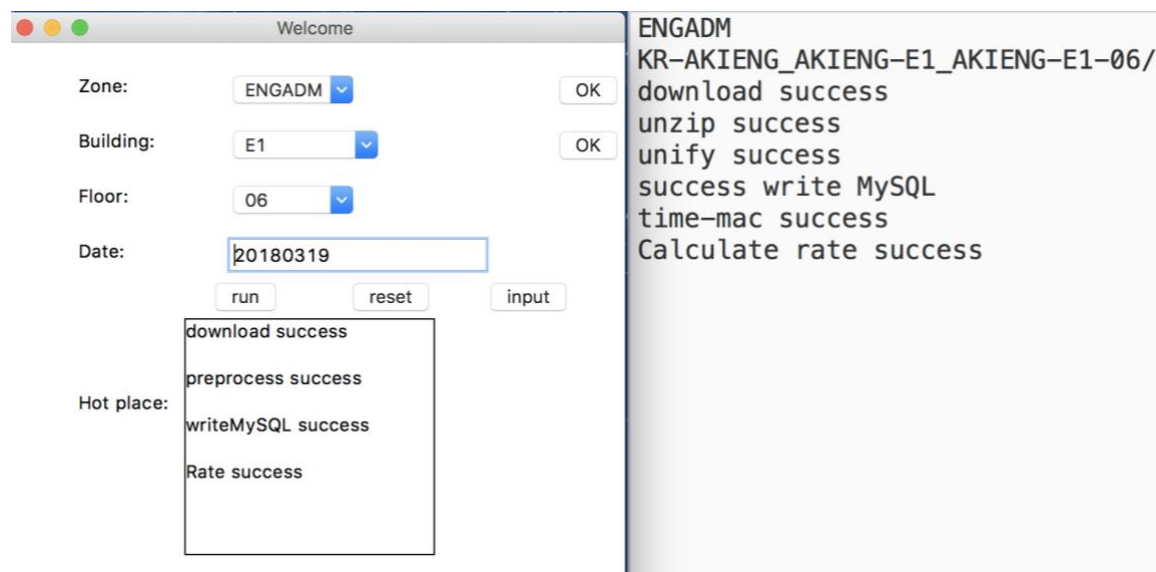


Figure4. 10 Application interface

As shown above, the left part is our UI and the right part is our python result output window. In this UI, after selecting location and date, click on the run button to run our application. Or if we find the place we want to view in the hot place, click the mouse to select it, enter the date, finally click the input button to run. After finished, we can see success output in our window.

3. Check the result in Tableau

First, we need to run Tableau and enter a password to connect to MySQL. Then we create a few worksheets to show the results. The result is shown below:

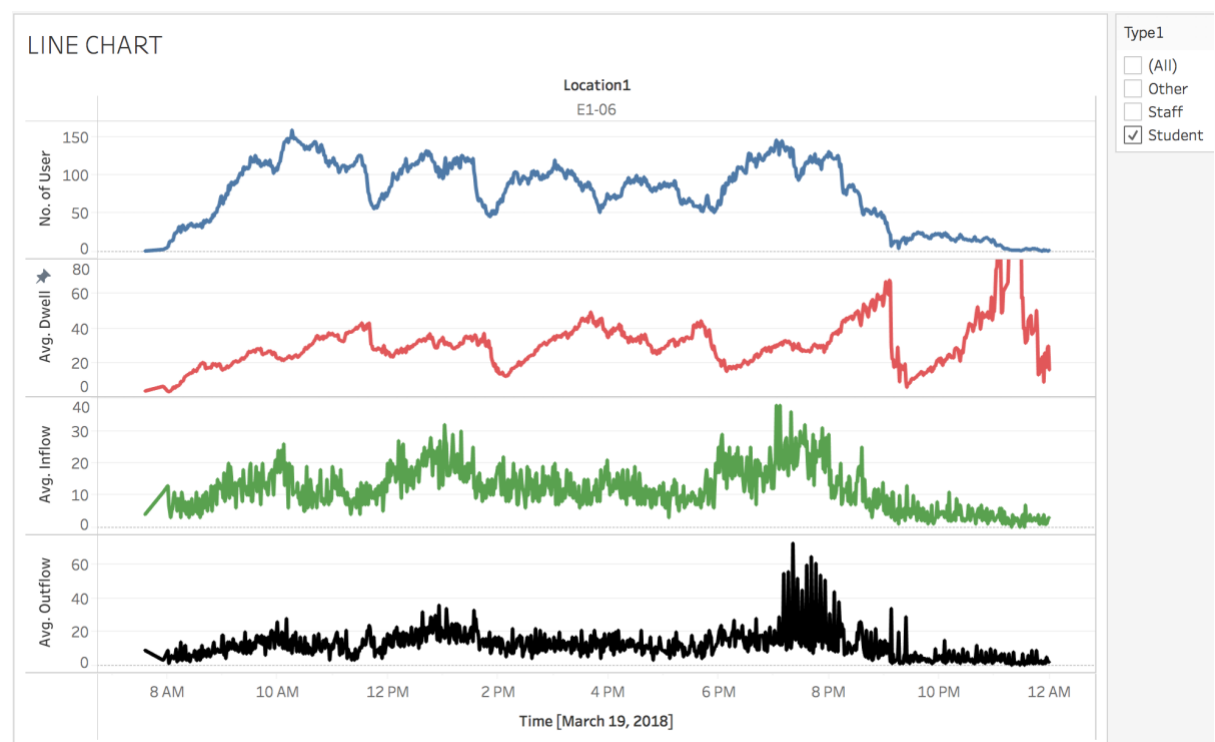


Figure4. 11 Line chart of E1

From above figure, we can find there were three classes in E1 according total number of users and dwell time. The pattern of class is the total number is high and the dwell time is increasing during class time. In this pattern, the three periods are 9AM-11AM, 2PM-4PM, 6PM-9PM.

As for inflow and outflow rate, high rate represents huge crowd mobility. Perhaps there is a question here why the crowd mobility around 8PM is so high? According to our own actual situation, teachers generally take a break at 7.30PM-8.00PM, so crowd mobility during class break is quite high.

In addition to the line graph, we also have a pie chart and type line chart that reflects the proportion between student, staff and other and their detailed changes over time. The following are these figures:

PIE CHART

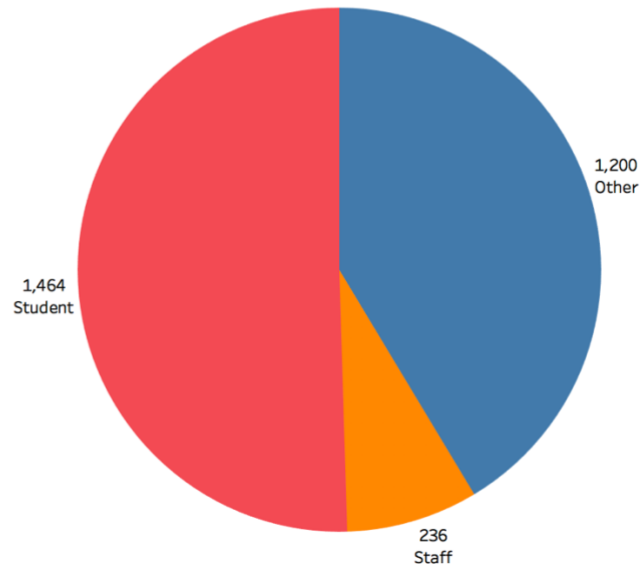


Figure4. 12 Pie chart of E1

TYPE

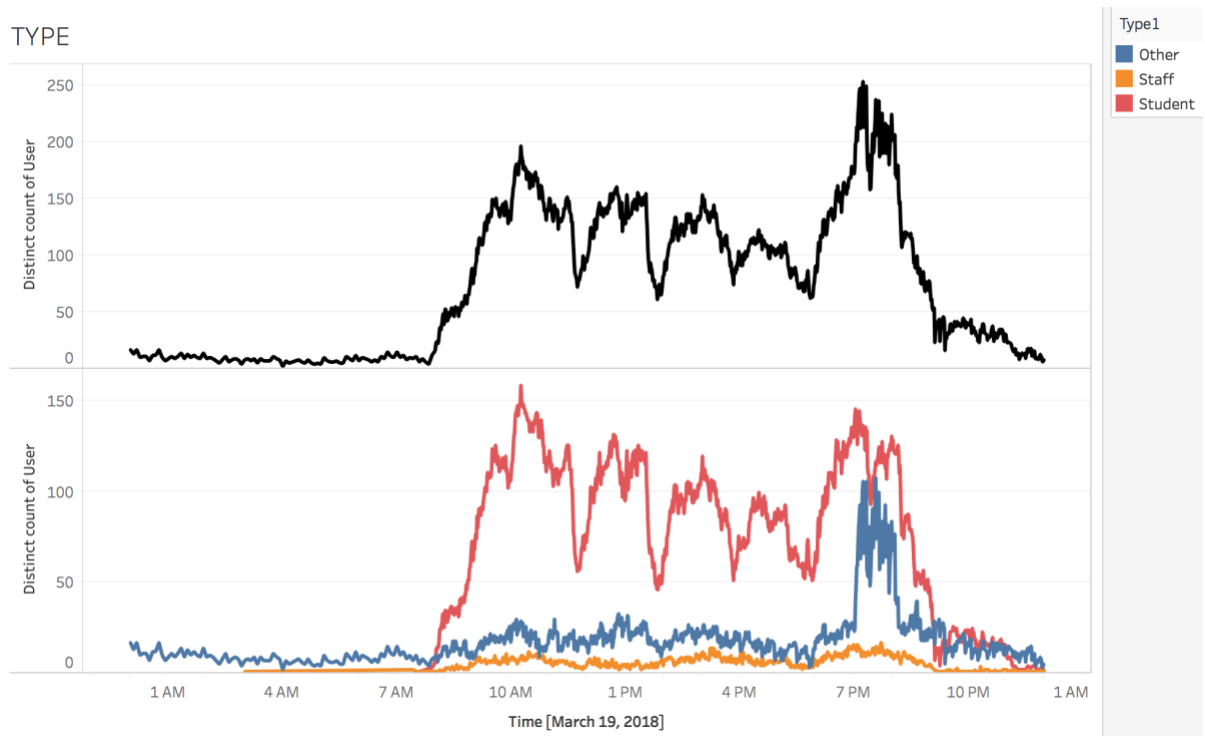


Figure4. 13 Type line chart of E1

From the above two figures we can conclude that students are the main body of the crowd. From the red line, we can see five peaks, which prove there were five classes on this day, each class time is 9 am-11am,12pm-2pm,2pm-4pm,4pm-6pm,6pm-9pm.

Chapter 5 Conclusion

This article designs and implements an application that analyzes NUS WiFi history data, describing the UI design, automatic data download, data preprocessing, data analysis, and data visualization. Finally, this article also demonstrates how to operate this application. The results shown are in good agreement with the actual situation. Therefore, we can use result to analyze the utilization and occupation of building, in order to increase the utilization of building and save the energy of lighting and air conditioning. If you need the source code, you can visit my [github](#) to download this project.

The above research in this paper has achieved some phased results. In order to improve the performance and compatibility of the application, the author believes that further research should be conducted from the following aspects:

1. Use the API to download. The shared folder used in this article has certain limitations. If you leave the NUS network or you do not have an NUS account, you cannot get the data.
2. The NUS WiFi system should use database storage. The current CISCO system uses csv files to store data and it is compressed once a week. The analysis of real-time data and historical data has some trouble due to csv. If database is used, data download and real-time analysis can be greatly improved.
3. Combine with the web. At present, the application is only implemented on the local computer. In order to be available to teachers and students in the future, the application must be deployed on a server and be presented as a web application.

Reference

- [1] Takuya Yoshida, Yoshiaki Taniguchi,” Estimating the number of people using existing WiFi access point in indoor environment”, Advances in Computer Science
- [2] Z. Dobesova, "Programming language Python for data processing," 2011 International Conference on Electrical and Control Engineering, Yichang, 2011, pp. 4866-4869.
- [3] Ramesh Sharda, Dursun Delen, Efraim Turban. Business Intelligence: A Managerial Perspective on Analytics (3rd Edition). ISBN:0133051056 9780133051056
- [4] Hand, D.J. Drug-Safety (2007) 30: 621. <https://doi.org/10.2165/00002018-200730070-00010>
- [5] Hockley, W.E. The picture superiority effect in associative recognition. Memory and Cognition 36 (2009), 1351-1359
- [6] R. Kozik, M. Choraś, D. Puchalski and R. Renk, "Data analysis tool supporting software development process," 2017 IEEE 14th International Scientific Conference on Informatics, Poprad, Slovakia, 2017, pp. 179-184
- [7] Visual Cues: Practical Data Visualization. Peter R. Keller, Mary M. Keller, Scott Markel, A. John Mallinckrodt, and Susan McKay. Computers in Physics 8, 297 (1994); doi: 10.1063/1.4823299