Geospatial Operational Insights for National Library Board (NLB)

Interim Progress Report

ANLY482 – Analytics Practicum AY16/17 Term 1

Prepared by:

Team Qui Vivra Verra

LIU Bowei

PONG Chong Xin

TEO Hui Min

Supervised by:

Prof KAM Tin Seong

Associate Professor of Information Systems

Table of Contents

1. Project Recap

1.1 Project Objectives

The main aim of the project is to provide NLB with valuable operational insights by developing a geospatial dashboard contained in a web-application, which determines the following when an existing library is relocated/removed or when a new library is added:

a. Demand capture area of libraries
b. Predicted patronage levels of libraries
c. Associated operational-related variables e.g. subzones served, distance to the nearest transport network (MRT station and/or bus stop)

To ensure the continued sustainability of the web-application, end-users will be able to upload files of the following format to update the model:

a. .csv format
b. .xlsx format

1.2 Project Problem Statement

There have been renovations and relocation of existing libraries, and new libraries have been commissioned (e.g. library@orchard) to keep up with the times. These constant changes prompt for a reliable system to measure the effectiveness of past policies, as well as an accurate predictive model to conduct what-if analyses for future plans. A user-friendly system which displays geospatial information that can provide operational insights would thus be valuable to the NLB.

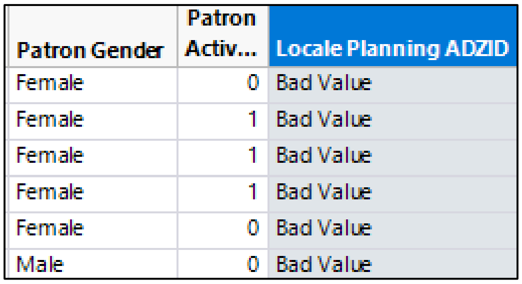
2. Data Exploration & Pre-processing

2.1 Summary of Anomalies & Errors

The team has performed exploratory data analysis and highlighted some anomalies as listed:

- Anomaly 1

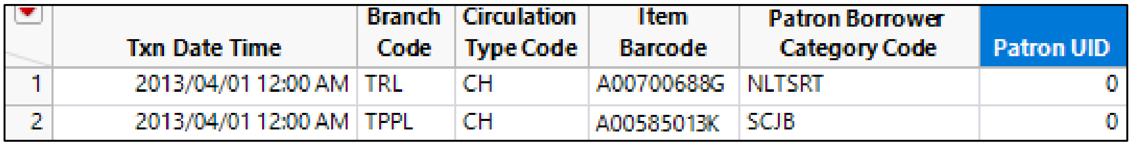
In the 2013 Patron Dataset, 2.534% of all records have *Locale Planning ADZID* set to *"Bad Value"* and *"Missing Value"*. In the 2013 Patron Dataset, 2.876% of all records have *Locale Planning ADZID* set to *"Bad Value"* and *"Missing Value"*. A snapshot of the anomaly is as shown below.



Patron Gender	Patron Activ...	Locale Planning ADZID
Female	0	Bad Value
Female	1	Bad Value
Female	1	Bad Value
Female	1	Bad Value
Female	0	Bad Value
Male	0	Bad Value

- Anomaly 2

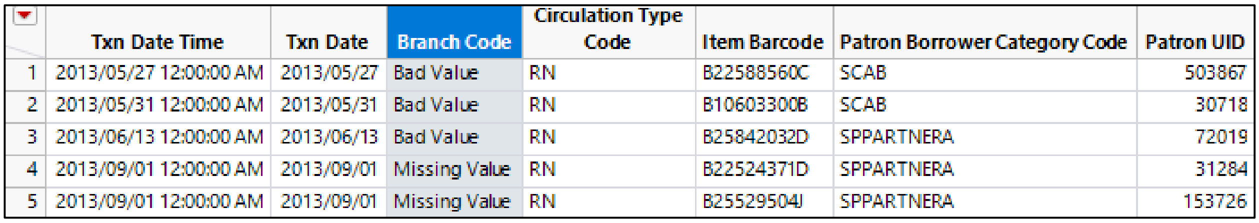
In the 2013 Transaction Dataset, 19,828 records have *Patron_UID* set to *'0'*. In the 2014 Transaction Dataset, 641 records have *Patron_UID* set to *'0'*. A snapshot of the anomaly is as shown below. A snapshot of the anomaly is as shown below.



	Txn Date Time	Branch Code	Circulation Type Code	Item Barcode	Patron Borrower Category Code	Patron UID
1	2013/04/01 12:00 AM	TRL	CH	A00700688G	NLTSRT	0
2	2013/04/01 12:00 AM	TPPL	CH	A00585013K	SCJB	0

- Anomaly 3

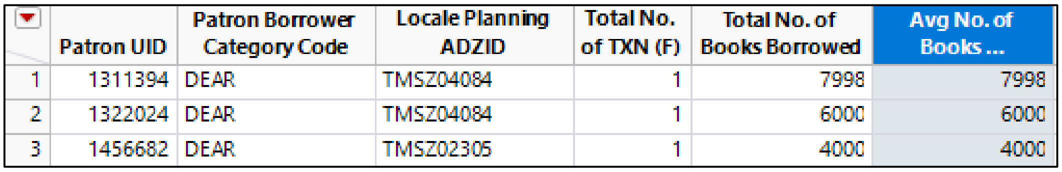
In the 2013 Transaction Dataset, 3 records have *Branch Code* set to *"Bad Value"* and 3,065 records set to *"Missing Value"*. In the 2014 Transaction Dataset, 9,467 records have *Branch Code* set to *"Bad Value"* and 1,600 records set to *"Missing Value"*. A snapshot of the anomaly is as shown below.



	Txn Date Time	Txn Date	Branch Code	Circulation Type Code	Item Barcode	Patron Borrower Category Code	Patron UID
1	2013/05/27 12:00:00 AM	2013/05/27	Bad Value	RN	B22588560C	SCAB	503867
2	2013/05/31 12:00:00 AM	2013/05/31	Bad Value	RN	B10603300B	SCAB	30718
3	2013/06/13 12:00:00 AM	2013/06/13	Bad Value	RN	B25842032D	SPPARTNERA	72019
4	2013/09/01 12:00:00 AM	2013/09/01	Missing Value	RN	B22524371D	SPPARTNERA	31284
5	2013/09/01 12:00:00 AM	2013/09/01	Missing Value	RN	B25529504J	SPPARTNERA	153726

- Anomaly 4

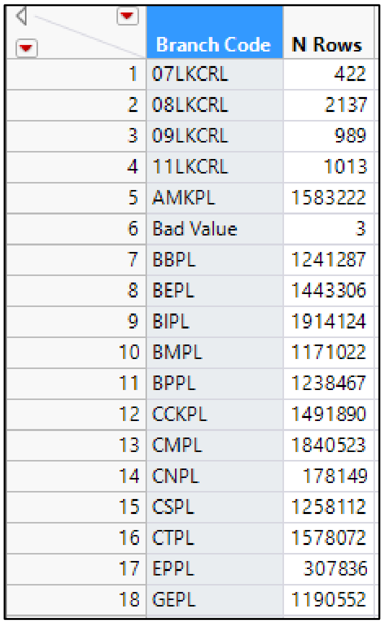
In the 2013 Transaction Dataset, there are 893 patrons with *Avg. No. of Books Borrowed* (aggregated value) exceeding 32. In the 2014 Transaction Dataset, there are 779 patrons with *Avg. No. of Books Borrowed* exceeding 32. Furthermore, from the records with *Avg. No. of Books Borrowed* exceeding 32, we find that most of the records have attribute *Patron Borrower Category Code* set to *"DEAR"*. These records also have unrealistic values in attribute *Patron Birthyear* i.e. *"1900", "2015, and "2016".* These records have *Patron Citizenship, Patron Race, Patron Gender* set to *"Others"*. A snapshot of the anomaly is as shown below.



	Patron UID	Patron Borrower Category Code	Locale Planning ADZID	Total No. of TXN (F)	Total No. of Books Borrowed	Avg No. of Books ...
1	1311394	DEAR	TMSZ04084	1	7998	7998
2	1322024	DEAR	TMSZ04084	1	6000	6000
3	1456682	DEAR	TMSZ02305	1	4000	4000

- Anomaly 5

In the 2013 and 2014 Transaction Datasets, there are records with *Branch Codes* that do not exist in *Collection_Dataset_FY13* and *FY14*, e.g. *'07LKCRL', 08LKCRL'*. A snapshot of the anomaly is as shown below.



	Branch Code	N Rows
1	07LKCRL	422
2	08LKCRL	2137
3	09LKCRL	989
4	11LKCRL	1013
5	AMKPL	1583222
6	Bad Value	3
7	BBPL	1241287
8	BEPL	1443306
9	BIPL	1914124
10	BMPL	1171022
11	BPPL	1238467
12	CCKPL	1491890
13	CMPL	1840523
14	CNPL	178149
15	CSPL	1258112
16	CTPL	1578072
17	EPPL	307836
18	GEPL	1190552

During Sponsor Meeting 01 with the NLB held on 06 October 2016, the team has consulted the NLB Analytics Team regarding the above-mentioned anomalies and noted the following:

- *Patron Borrower Category Code "DEAR"*, is not a unique patron per se, but refers to institutional partnership programmes. NLB suggested to remove all records with *Patron Borrower Category Code* set to *"DEAR"* from further analysis.
- Patrons with *Birthyears '1900', '2015' and '2016'* are due to values being set to the year that the institutional programme was set up.
- NLB suggested to exclude the *Branch Codes* that are not listed in the Collection Dataset.
- NLB is agreeable with removing all records that contain the anomalies as described above.

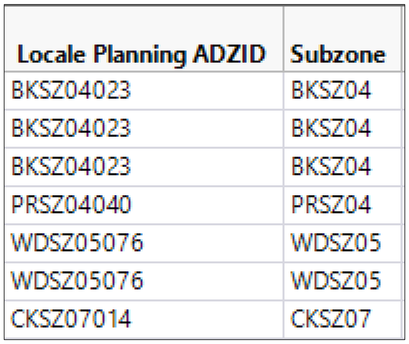
2.2 Data Pre-processing

2.2.1 Planning Area Reconciliation

In the Patron Datasets, the residential areas of the patrons are given as *Local Planning ADZID,* which is equivalent to the subzones where they live in. However, the team find that it may be hard to identify the residential areas with subzones as there are simply too many subzones. Hence the team has decided to match the subzones to URA Planning Areas instead.

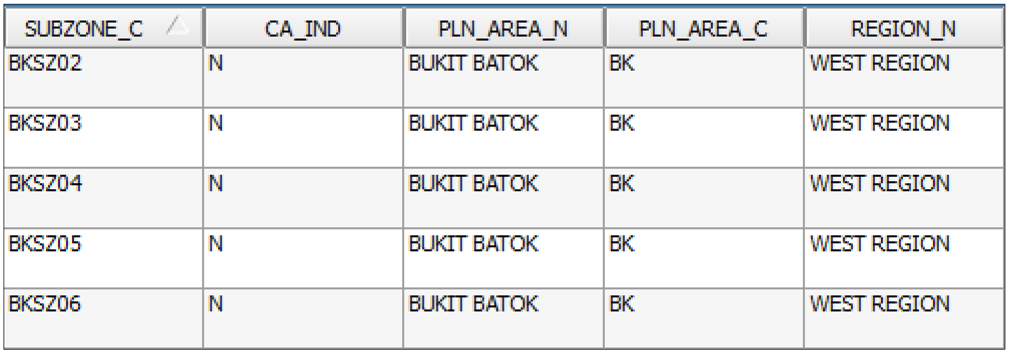
The matching of Subzones to Planning Area was performed in QGIS and the procedure is as follows:

1. Extracted the first 6 characters (column *‘Subzone’*) of *Locale Planning ADZID* from the *Patron Dataset* using JMP Pro.



Locale Planning ADZID	Subzone
BKSZ04023	BKSZ04
BKSZ04023	BKSZ04
BKSZ04023	BKSZ04
PRSZ04040	PRSZ04
WDSZ05076	WDSZ05
WDSZ05076	WDSZ05
CKSZ07014	CKSZ07

2. The *Patron Dataset* is then joined with the *MP14_SUBZONE* shapefile which the team has obtained from data.gov.sg, matching the extracted characters from *Patron Dataset* with the variable *‘SUBZONE_C’* found in the *MP14_SUBZONE* shape file to get the Planning Area for each Subzone. The team has also obtained the region (column *REGION_N*) in which the Subzones fall in, which can be used to display the data in a higher level of detail.



SUBZONE_C	CA_IND	PLN_AREA_N	PLN_AREA_C	REGION_N
BKSZ02	N	BUKIT BATOK	BK	WEST REGION
BKSZ03	N	BUKIT BATOK	BK	WEST REGION
BKSZ04	N	BUKIT BATOK	BK	WEST REGION
BKSZ05	N	BUKIT BATOK	BK	WEST REGION
BKSZ06	N	BUKIT BATOK	BK	WEST REGION

The team also needed the centroid of each Planning Area to be used for initial visualisation in Tableau, thus the team has made use of the ‘Polygon Centroid’ function available in QGIS to obtain the centroid and coordinates (Latitude and Longitude) of each Planning Area.

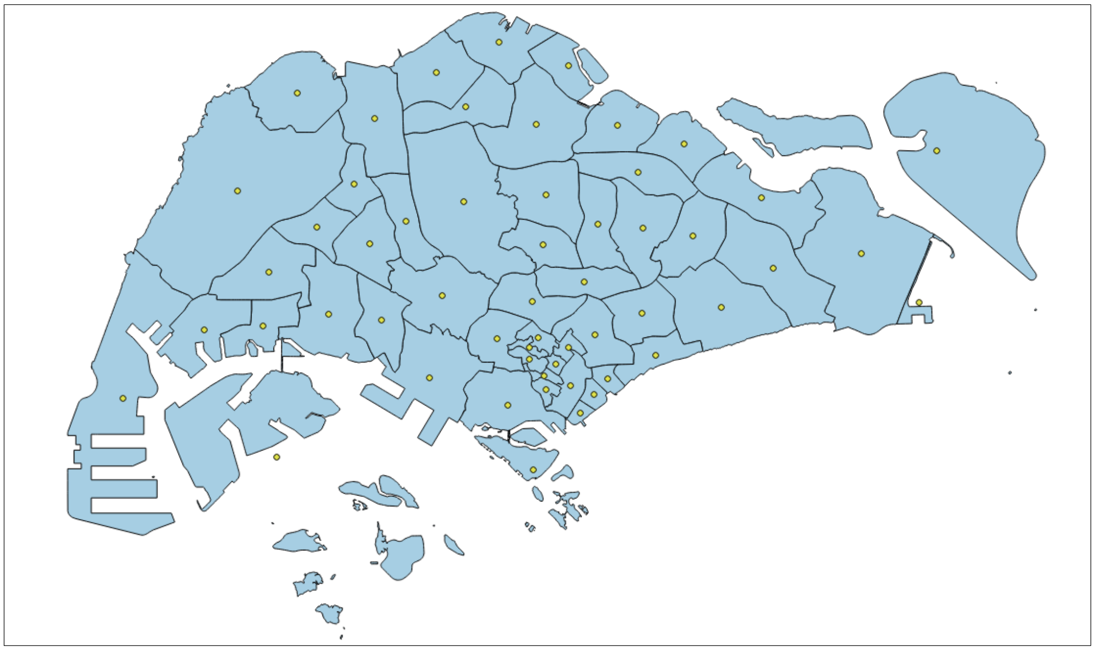
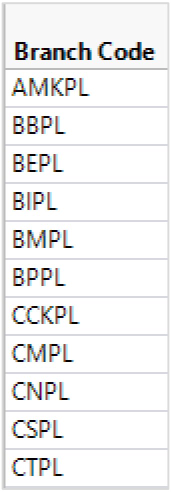


Figure 2.1 The centroid of each Planning Area is represented by the yellow circle.

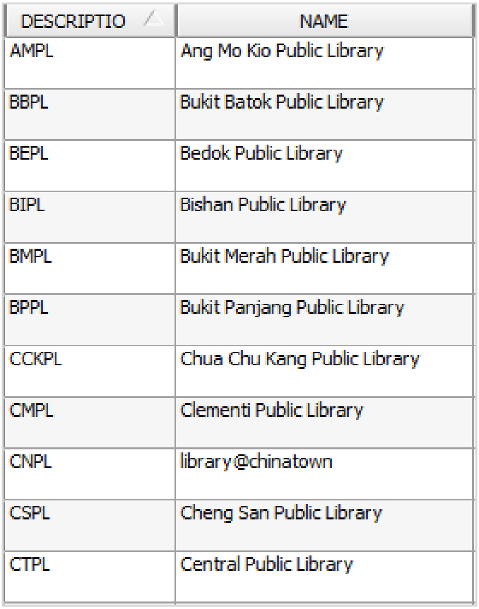
2.2.2 Library Branch Code Reconciliation

From the *Collection Dataset*, the only identifier for the libraries is the attribute *Branch Code*. However, it is hard to identify a library based on the *Branch Code*. Hence, the *Branch Code* is then matched to the *‘DESCRIPTIO’* variable found in the attribute table of the *LIBRARIES* shapefile which the team has obtained from data.gov.sg to get the library names. Furthermore, as the team wanted to perform a geospatial visualisation, we needed to get the coordinates of the libraries which was also obtained from performing the join.



Branch Code
AMKPL
BBPL
BEPL
BIPL
BMPL
BPPL
CCKPL
CMPL
CNPL
CSPL
CTPL

Figure 2.2 Branch Code of libraries in Collection Dataset



DESCRIPTIO	NAME
AMPL	Ang Mo Kio Public Library
BBPL	Bukit Batok Public Library
BEPL	Bedok Public Library
BIPL	Bishan Public Library
BMPL	Bukit Merah Public Library
BPPL	Bukit Panjang Public Library
CCKPL	Chua Chu Kang Public Library
CMPL	Clementi Public Library
CNPL	library@chinatown
CSPL	Cheng San Public Library
CTPL	Central Public Library

Figure 2.3 To match 'Branch Code' to 'DESCRIPTIO' in 'LIBRARIES' shapefile

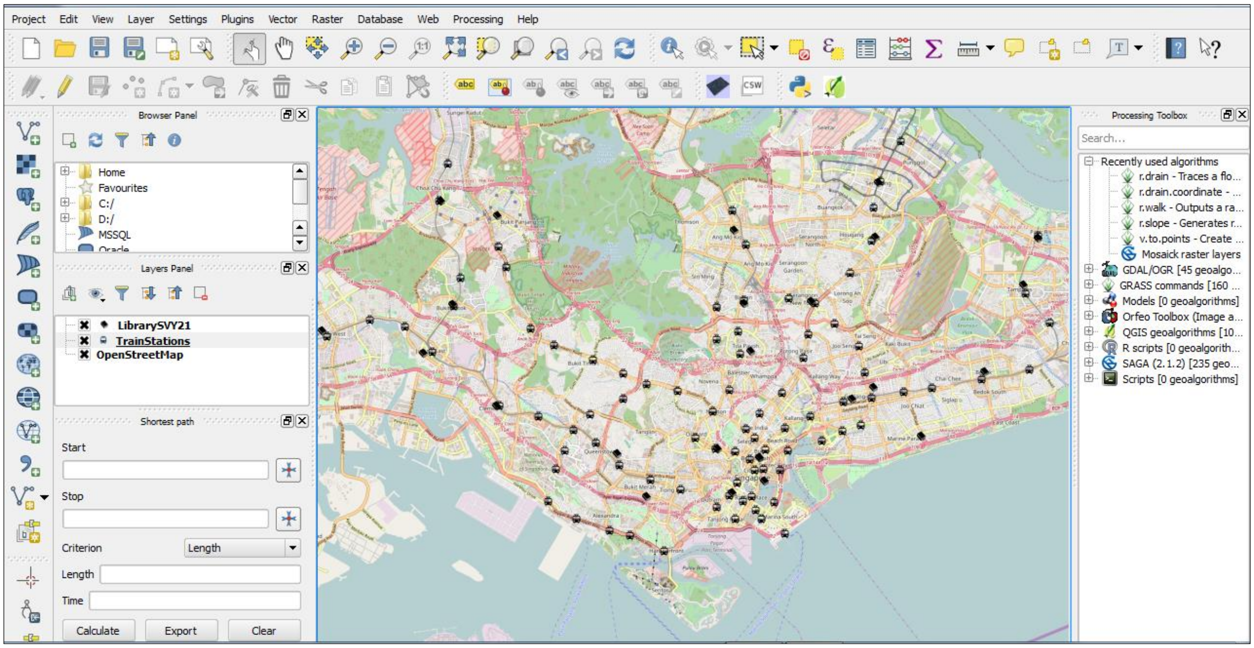
2.3 External Data Sources

The team has sourced for external data, which will be used in conjunction with the provided data to build up the geospatial model. Listed in the following sections are the details of the type of external data derived.

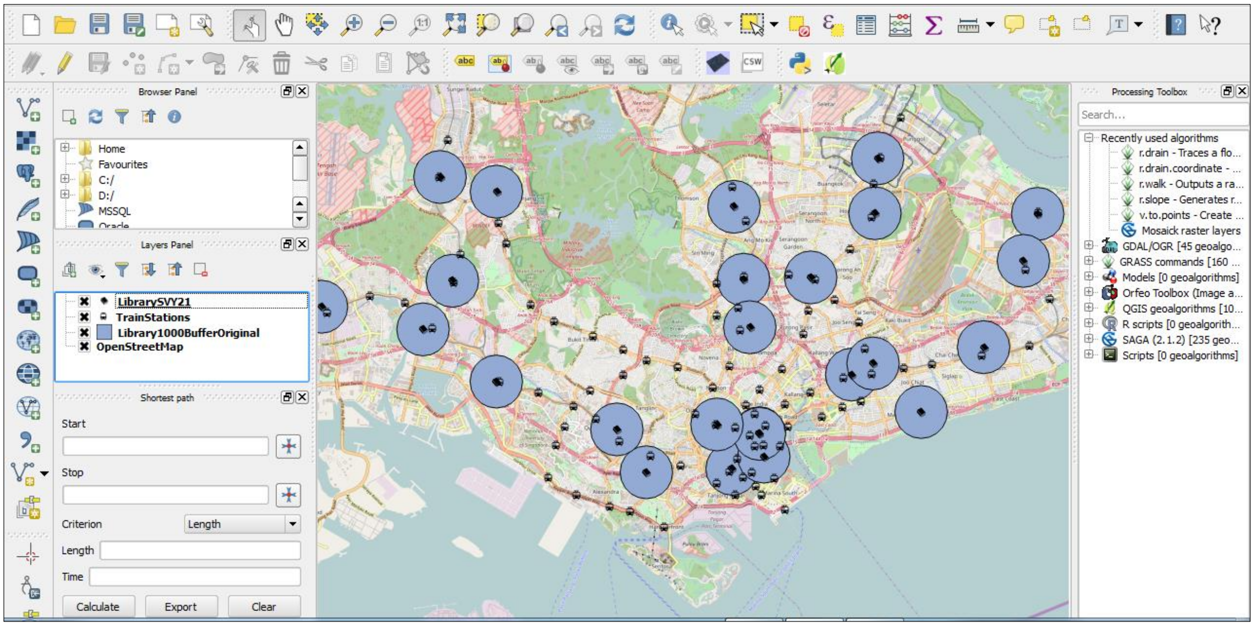
2.3.1 MRT Stations

The team felt that the libraries' proximity to MRT stations is possibly a factor explains the difference in patronage flow to a library. For example, an individual may prefer to patronise a particular library due to the presence of a MRT station near the library. Hence, libraries with a greater number of MRT stations located nearby would draw more patrons.

Therefore, we decided to compile information on the locations of the various MRT stations along Singapore. Fortunately, this information is readily available at data.gov.sg, in SHP format. After downloading this dataset, we visualized the distribution of MRT stations along with the library branches in QGIS.



Using QGIS, we then created buffer region with a radius of 1 kilometre around the library branches, to simulate the area within walking distance to the branches.

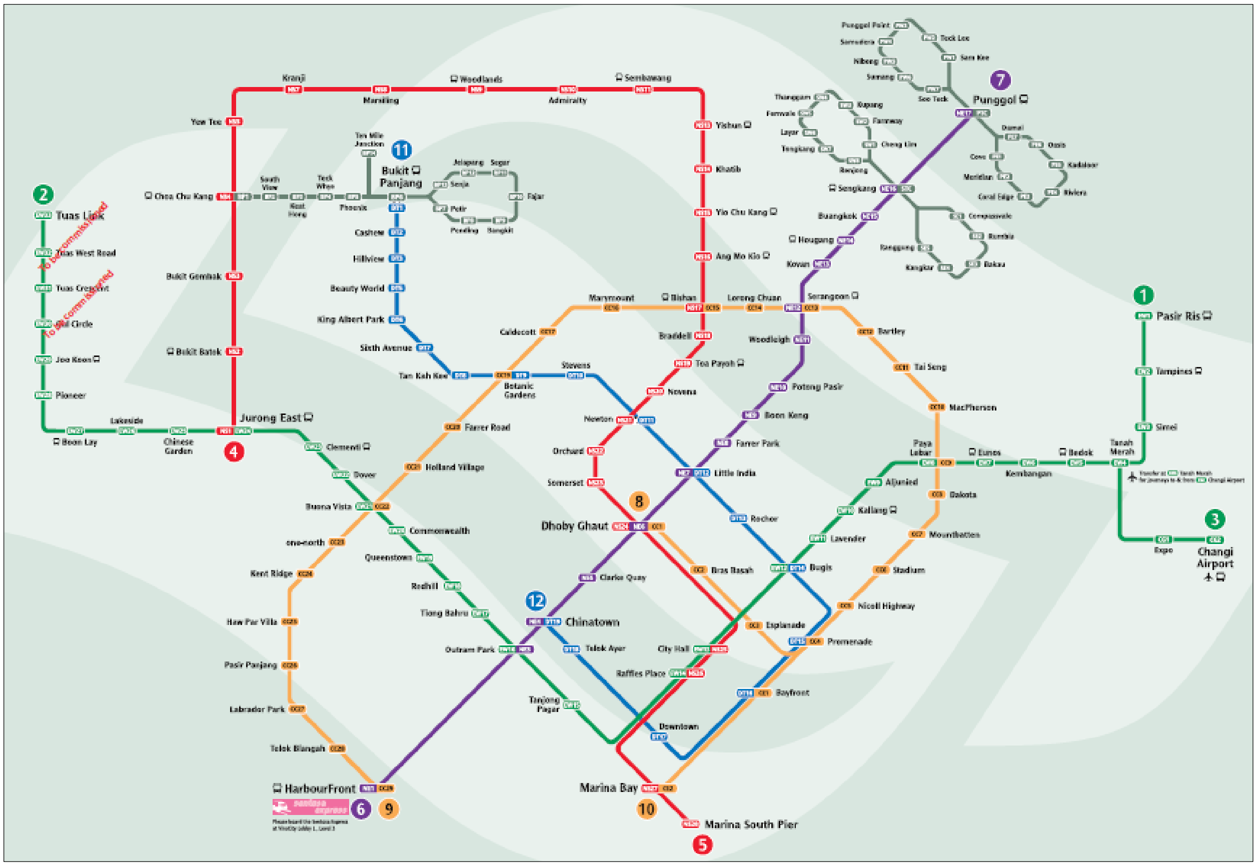


Criterion Length

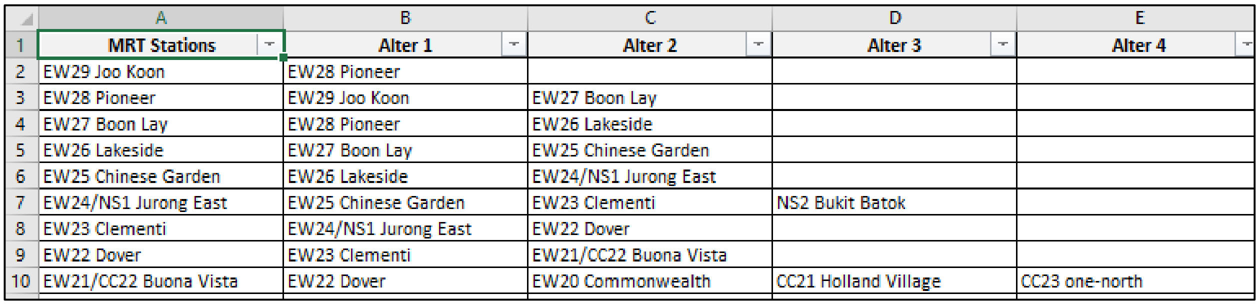
We then counted the number of MRT stations within this specified area, and exported the derived data in the form of a .csv file to be used in Tableau.

To add another dimension in the data, the team have researched on using centrality measures of the MRT stations within the MRT transport network, instead of simply using the absolute number of MRT stations within 1 kilometre of the libraries.

First, we derived the MRT network map from the SMRT official webpage:

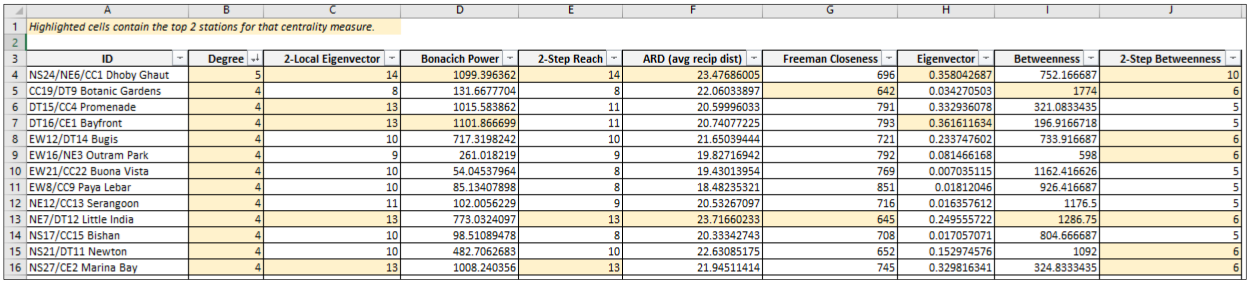


Next, we manually created a matrix, in which each row shows a unique MRT stations (ego) and the other MRT stations it is adjacent to (alters). A snapshot of the matrix is as shown below:



	A	B	C	D	E
1	MRT Stations	Alter 1	Alter 2	Alter 3	Alter 4
2	EW29 Joo Koon	EW28 Pioneer			
3	EW28 Pioneer	EW29 Joo Koon	EW27 Boon Lay		
4	EW27 Boon Lay	EW28 Pioneer	EW26 Lakeside		
5	EW26 Lakeside	EW27 Boon Lay	EW25 Chinese Garden		
6	EW25 Chinese Garden	EW26 Lakeside	EW24/NS1 Jurong East		
7	EW24/NS1 Jurong East	EW25 Chinese Garden	EW23 Clementi	NS2 Bukit Batok	
8	EW23 Clementi	EW24/NS1 Jurong East	EW22 Dover		
9	EW22 Dover	EW23 Clementi	EW21/CC22 Buona Vista		
10	EW21/CC22 Buona Vista	EW22 Dover	EW20 Commonwealth	CC21 Holland Village	CC23 one-north

Using UCINET, an open-source software package for social network analysis, we calculated the centralities of each MRT station in the overall network (excluding LRT stations), using different formulations of centralities, such as degree centrality, eigenvector centrality, betweenness centrality and closeness centrality. A screenshot is provided below:



Highlighted cells contain the top 2 stations for that centrality measure.

ID	Degree	2-Local Eigenvector	Bonacich Power	2-Step Reach	ARD (avg recip dist)	Freeman Closeness	Eigenvector	Betweenness	2-Step Betweenness
NS24/NE6/CC1 Dhoby Ghaut	5	14	1099.396362	14	23.47686005	696	0.358042687	752.166687	10
CC19/DT9 Botanic Gardens	4	8	131.6677704	8	22.06033897	642	0.034270503	1774	6
DT15/CC4 Promenade	4	13	1015.583862	11	20.59996033	791	0.332936078	321.0833435	5
DT16/CE1 Bayfront	4	13	1101.866699	11	20.74077225	793	0.361611634	196.9166718	5
EW12/DT14 Bugis	4	10	717.3198242	10	21.65039444	721	0.233747602	733.916687	6
EW16/NE3 Outram Park	4	9	261.018219	9	19.82716942	792	0.081466168	598	6
EW21/CC22 Buona Vista	4	10	54.04537964	8	19.43013954	769	0.007035115	1162.416626	5
EW8/CC9 Paya Lebar	4	10	85.13407898	8	18.48235321	851	0.01812046	926.416687	5
NE12/CC13 Serangoon	4	11	102.0056229	9	20.53267097	716	0.016357612	1176.5	5
NE7/DT12 Little India	4	13	773.0324097	13	23.71660233	645	0.249555722	1286.75	6
NS17/CC15 Bishan	4	10	98.51089478	8	20.33342743	708	0.017057071	804.666687	5
NS21/DT11 Newton	4	10	482.7062683	10	22.63085175	652	0.152974576	1092	6
NS27/CE2 Marina Bay	4	13	1008.240356	13	21.94511414	745	0.329816341	324.8333435	6

A high centrality measure for a MRT station suggests that it holds a greater importance within the network, and hence a library located to a MRT station with a high centrality score may attract more patrons than one located to a MRT station with a low centrality score. We will be testing the statistical significance of the different centralities and to prevent multicollinearity in the regression model that will be built, we may be only using one type of centrality that best fits the data.

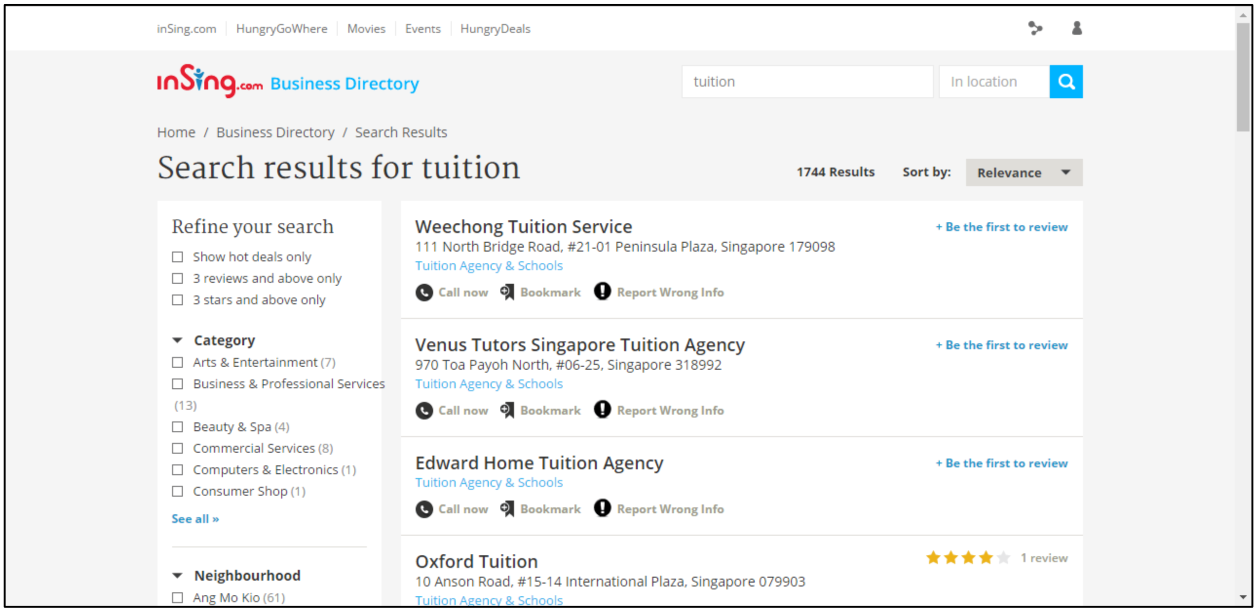
2.3.2 Shopping Malls

The team felt that proximity to shopping malls would be another factor that explains difference in libraries' patronage flow. Visitors to malls might patronise the libraries near the malls due to convenience. To collect geographical data on shopping malls in Singapore, we downloaded the shopping mall data set from data.gov.sg, similar to the process we used to obtain the MRT stations data. The process used to derive the .csv file for visualizing Tableau is similar to the process rendered for the MRT stations' data set.

2.3.3 Tuition Centres

The presence of tuition centres near the libraries may also explain the differences in patronage levels for different libraries (due to differences in number of nearby tuition centres). Parents dropping off their children at children centres may arrive early and decide to let their children spend some time at a nearby library. However, it is difficult to obtain geographical data set for tuition centres in Singapore as the information is not readily available on data.gov.sg. Instead, we have decided to use inSing, a website that functions like a phone book for businesses. Information available on the site includes addresses of businesses.

To get a list of tuition centres, we used the keyword *'tuition'* and performed a search on the inSing website, as shown in the screenshot below.



inSing.com | HungryGoWhere | Movies | Events | HungryDeals
inSing.com Business Directory
tuition
In location

Home / Business Directory / Search Results

Search results for tuition

1744 Results Sort by: Relevance

Refine your search

□ Show hot deals only
□ 3 reviews and above only
□ 3 stars and above only

Category

□ Arts & Entertainment (7)
□ Business & Professional Services (13)
□ Beauty & Spa (4)
□ Commercial Services (8)
□ Computers & Electronics (1)
□ Consumer Shop (1)

See all »

Neighbourhood

□ Ang Mo Kio (61)

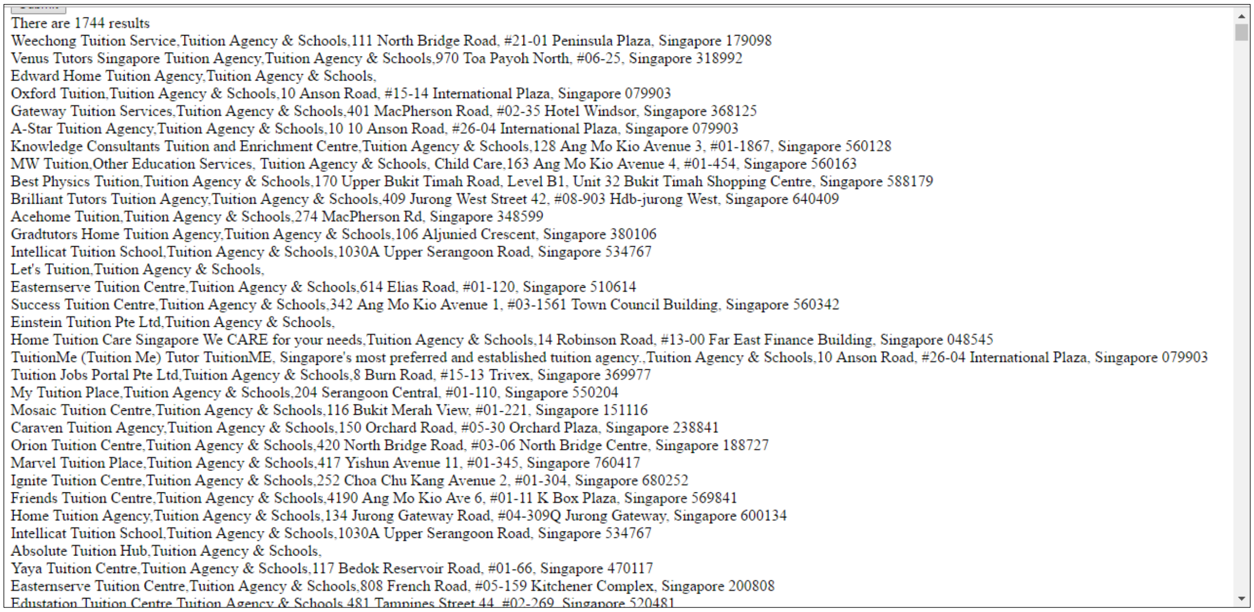
Weechong Tuition Service
111 North Bridge Road, #21-01 Peninsula Plaza, Singapore 179098
Tuition Agency & Schools
Call now Bookmark Report Wrong Info + Be the first to review

Venus Tutors Singapore Tuition Agency
970 Toa Payoh North, #06-25, Singapore 318992
Tuition Agency & Schools
Call now Bookmark Report Wrong Info + Be the first to review

Edward Home Tuition Agency
Tuition Agency & Schools
Call now Bookmark Report Wrong Info + Be the first to review

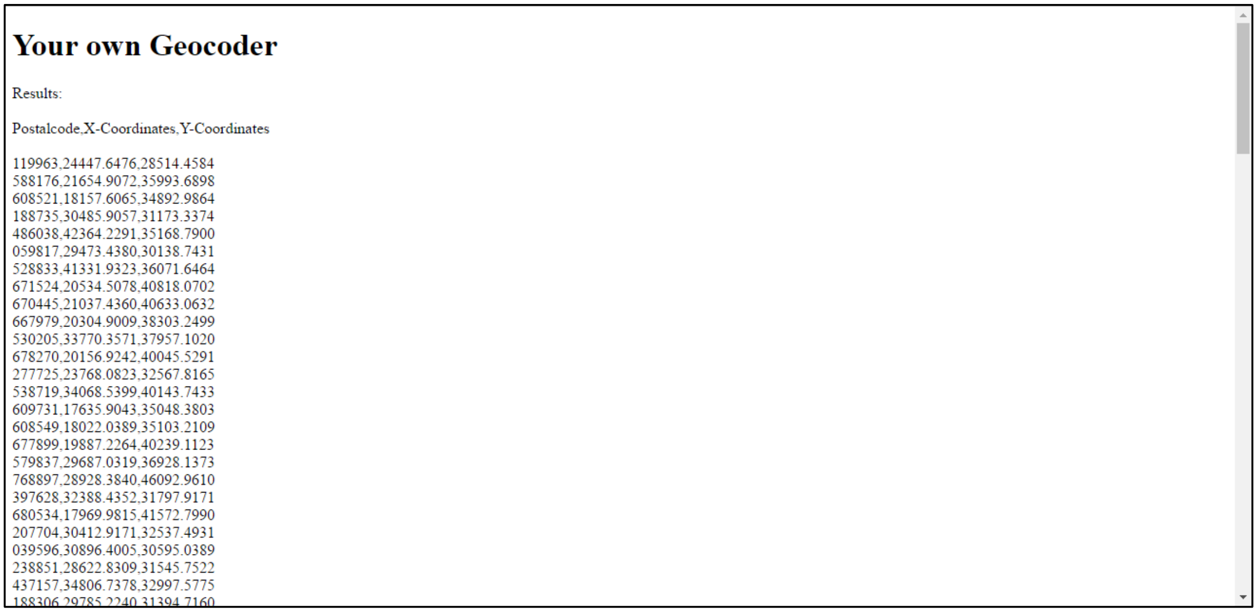
Oxford Tuition
10 Anson Road, #15-14 International Plaza, Singapore 079903
Tuition Agency & Schools
★★★★☆ 1 review

However, due to the large number of results (1,744), copying and pasting the information manually will take too long, as inSing shows information for 10 tuitions centres at a time. Instead, we created a list of links to inSing, with each link giving us the information about 10 tuition centres, by changing the page number in the link. For example, the link http://search.insing.com/s/tuition?page=1 lists first 10 tuition centres in the results, while the link http://search.insing.com/s/tuition?page=10 lists the 100th to 110th tuition centres. The full list of 175 links were inserted to a self-coded html parser, to extract the information needed. The of the process is shown below, and specifically the last column contains the postal code of the tuition centre.



```
There are 1744 results
Weechong Tuition Service,Tuition Agency & Schools,111 North Bridge Road, #21-01 Peninsula Plaza, Singapore 179098
Venus Tutors Singapore Tuition Agency,Tuition Agency & Schools,970 Toa Payoh North, #06-25, Singapore 318992
Edward Home Tuition Agency,Tuition Agency & Schools,
Oxford Tuition,Tuition Agency & Schools,10 Anson Road, #15-14 International Plaza, Singapore 079903
Gateway Tuition Services,Tuition Agency & Schools,401 MacPherson Road, #02-35 Hotel Windsor, Singapore 368125
A-Star Tuition Agency,Tuition Agency & Schools,10 10 Anson Road, #26-04 International Plaza, Singapore 079903
Knowledge Consultants Tuition and Enrichment Centre,Tuition Agency & Schools,128 Ang Mo Kio Avenue 3, #01-1867, Singapore 560128
MW Tuition,Other Education Services, Tuition Agency & Schools, Child Care,163 Ang Mo Kio Avenue 4, #01-454, Singapore 560163
Best Physics Tuition,Tuition Agency & Schools,170 Upper Bukit Timah Road, Level B1, Unit 32 Bukit Timah Shopping Centre, Singapore 588179
Brilliant Tutors Tuition Agency,Tuition Agency & Schools,409 Jurong West Street 42, #08-903 Hdb-jurong West, Singapore 640409
Acehome Tuition,Tuition Agency & Schools,274 MacPherson Rd, Singapore 348599
Gradtutors Home Tuition Agency,Tuition Agency & Schools,106 Aljunied Crescent, Singapore 380106
Intellicat Tuition School,Tuition Agency & Schools,1030A Upper Serangoon Road, Singapore 534767
Let's Tuition,Tuition Agency & Schools,
Easternserve Tuition Centre,Tuition Agency & Schools,614 Elias Road, #01-120, Singapore 510614
Success Tuition Centre,Tuition Agency & Schools,342 Ang Mo Kio Avenue 1, #03-1561 Town Council Building, Singapore 560342
Einstein Tuition Pte Ltd,Tuition Agency & Schools,
Home Tuition Care Singapore We CARE for your needs,Tuition Agency & Schools,14 Robinson Road, #13-00 Far East Finance Building, Singapore 048545
TuitionMe (Tuition Me) Tutor TuitionME, Singapore's most preferred and established tuition agency.,Tuition Agency & Schools,10 Anson Road, #26-04 International Plaza, Singapore 079903
Tuition Jobs Portal Pte Ltd,Tuition Agency & Schools,8 Burn Road, #15-13 Trivex, Singapore 369977
My Tuition Place,Tuition Agency & Schools,204 Serangoon Central, #01-110, Singapore 550204
Mosaic Tuition Centre,Tuition Agency & Schools,116 Bukit Merah View, #01-221, Singapore 151116
Caraven Tuition Agency,Tuition Agency & Schools,150 Orchard Road, #05-30 Orchard Plaza, Singapore 238841
Orion Tuition Centre,Tuition Agency & Schools,420 North Bridge Road, #03-06 North Bridge Centre, Singapore 188727
Marvel Tuition Place,Tuition Agency & Schools,417 Yishun Avenue 11, #01-345, Singapore 760417
Ignite Tuition Centre,Tuition Agency & Schools,252 Choa Chu Kang Avenue 2, #01-304, Singapore 680252
Friends Tuition Centre,Tuition Agency & Schools,4190 Ang Mo Kio Ave 6, #01-11 K Box Plaza, Singapore 569841
Home Tuition Agency,Tuition Agency & Schools,134 Jurong Gateway Road, #04-309Q Jurong Gateway, Singapore 600134
Intellicat Tuition School,Tuition Agency & Schools,1030A Upper Serangoon Road, Singapore 534767
Absolute Tuition Hub,Tuition Agency & Schools,
Yaya Tuition Centre,Tuition Agency & Schools,117 Bedok Reservoir Road, #01-66, Singapore 470117
Easternserve Tuition Centre,Tuition Agency & Schools,808 French Road, #05-159 Kitchener Complex, Singapore 200808
Edustation Tuition Centre,Tuition Agency & Schools,481 Tampines Street 44, #02-269, Singapore 520481
```

Next, the list of postal codes generated by the html parser were used as inputs in a geocoder app (as recommended for use by Prof Kam) to derive the corresponding coordinates.

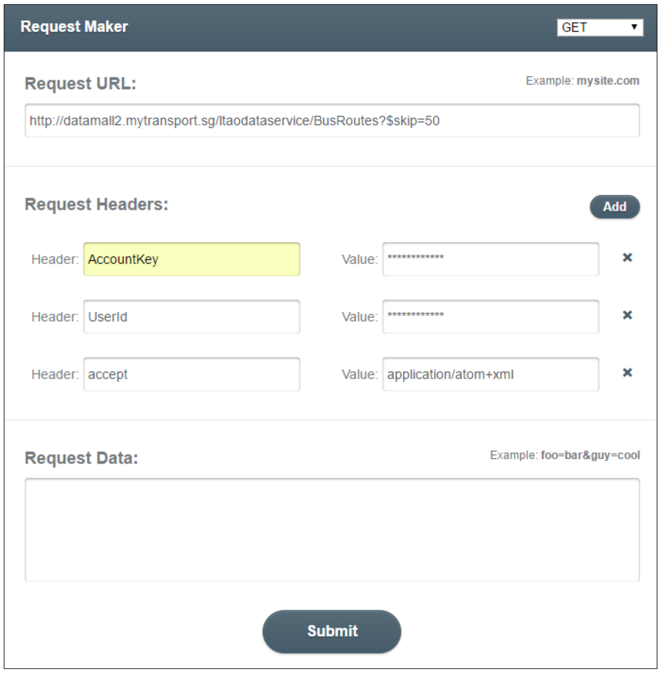


```
Your own Geocoder

Results:

Postalcode,X-Coordinates,Y-Coordinates

119963,24447.6476,28514.4584
588176,21654.9072,35993.6898
608521,18157.6065,34892.9864
188735,30485.9057,31173.3374
486038,42364.2291,35168.7900
059817,29473.4380,30138.7431
528833,41331.9323,36071.6464
671524,20534.5078,40818.0702
670445,21037.4360,40633.0632
667979,20304.9009,38303.2499
530205,33770.3571,37957.1020
678270,20156.9242,40045.5291
277725,23768.0823,32567.8165
538719,34068.5399,40143.7433
609731,17635.9043,35048.3803
608549,18022.0389,35103.2109
677899,19887.2264,40239.1123
579837,29687.0319,36928.1373
768897,28928.3840,46092.9610
397628,32388.4352,31797.9171
680534,17969.9815,41572.7990
207704,30412.9171,32537.4931
039596,30896.4005,30595.0389
238851,28622.8309,31545.7522
437157,34806.7378,32997.5775
188306,29785.2240,31394.7160
```

These coordinates are finally read into QGIS to give a layer showing the geographical location of tuition centres in Singapore. The process used to derive the .csv file for visualizing Tableau is the same as the one used for the two data sets mentioned above.

2.3.4 Bus Stops

Instead of counting the number of bus stops within a buffer, the team would like to get the number of bus services instead as it better show how convenient people could get to the libraries. The bus stop data was retrieved from LTA's DataMall@MyTransport API.



GET

Request URL: http://datamall2.mytransport.sg/ltaodataservice/BusRoutes?$skip=50

Request Headers:

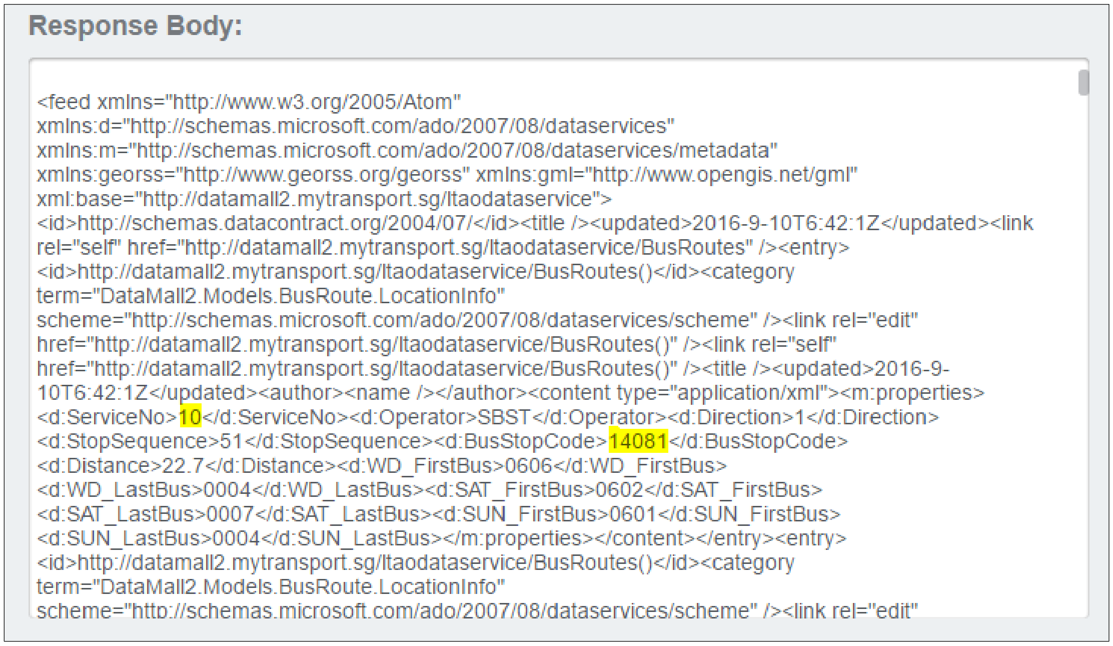
Header: AccountKey Value: *************

Header: UserId Value: *************

Header: accept Value: application/atom+xml

Request Data:

The API calls returned information of every bus service and the bus stops that it will stop by, which could be aggregated to get the number of bus service at a bus stop. The data also provided the coordinates of the bus stops which would be useful to visualise the location of bus stops.



Response Body:

```
<feed xmlns="http://www.w3.org/2005/Atom"
xmlns:d="http://schemas.microsoft.com/ado/2007/08/dataservices"
xmlns:m="http://schemas.microsoft.com/ado/2007/08/dataservices/metadata"
xmlns:georss="http://www.georss.org/georss" xmlns:gml="http://www.opengis.net/gml"
xml:base="http://datamall2.mytransport.sg/ltaodataservice">
<id>http://schemas.datacontract.org/2004/07/</id><title /><updated>2016-9-10T6:42:1Z</updated><link
rel="self" href="http://datamall2.mytransport.sg/ltaodataservice/BusRoutes" /><entry>
<id>http://datamall2.mytransport.sg/ltaodataservice/BusRoutes()</id><category
term="DataMall2.Models.BusRoute.LocationInfo"
scheme="http://schemas.microsoft.com/ado/2007/08/dataservices/scheme" /><link rel="edit"
href="http://datamall2.mytransport.sg/ltaodataservice/BusRoutes()" /><link rel="self"
href="http://datamall2.mytransport.sg/ltaodataservice/BusRoutes()" /><title /><updated>2016-9-
10T6:42:1Z</updated><author><name /></author><content type="application/xml"><m:properties>
<d:ServiceNo>10</d:ServiceNo><d:Operator>SBST</d:Operator><d:Direction>1</d:Direction>
<d:StopSequence>51</d:StopSequence><d:BusStopCode>14081</d:BusStopCode>
<d:Distance>22.7</d:Distance><d:WD_FirstBus>0606</d:WD_FirstBus>
<d:WD_LastBus>0004</d:WD_LastBus><d:SAT_FirstBus>0602</d:SAT_FirstBus>
<d:SAT_LastBus>0007</d:SAT_LastBus><d:SUN_FirstBus>0601</d:SUN_FirstBus>
<d:SUN_LastBus>0004</d:SUN_LastBus></m:properties></content></entry><entry>
<id>http://datamall2.mytransport.sg/ltaodataservice/BusRoutes()</id><category
term="DataMall2.Models.BusRoute.LocationInfo"
scheme="http://schemas.microsoft.com/ado/2007/08/dataservices/scheme" /><link rel="edit"
```

Figure 2.4 The Bus service number and Bus stop code returned from the API call are as highlighted.

2.4 Cluster Analysis

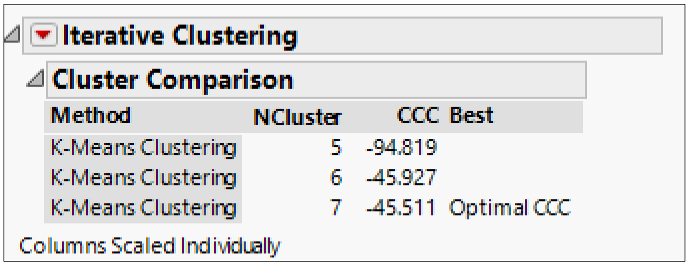
After removing the anomalies mentioned in section 2.1, the team decided to apply the method of cluster analysis to identify possible segmentations in the library patrons, based on differences in their borrowing patterns.

The variables that are used for the cluster analysis are:

a. End Date – Last TXN Date (Recency)
b. Total No. of TXN in the FY (Frequency)
c. Avg. No. of Books Borrowed Per TXN (Monetary)

These variables are adapted from the concept of RFM Analysis, which is popular in marketing analysis. By looking at the how recent was the patron’s last transaction (recency), how many times did they visit the library in a year (frequency), and the average number of books borrowed per transaction (monetary), we can potentially divide patrons into distinct groups based on differences in their borrowing patterns.

Using JMP Pro, we used the cluster analysis module and applied the Johnson Transformation and standardization functions, then performed k-means clustering on the data. Results of the cluster analysis is as displayed below:

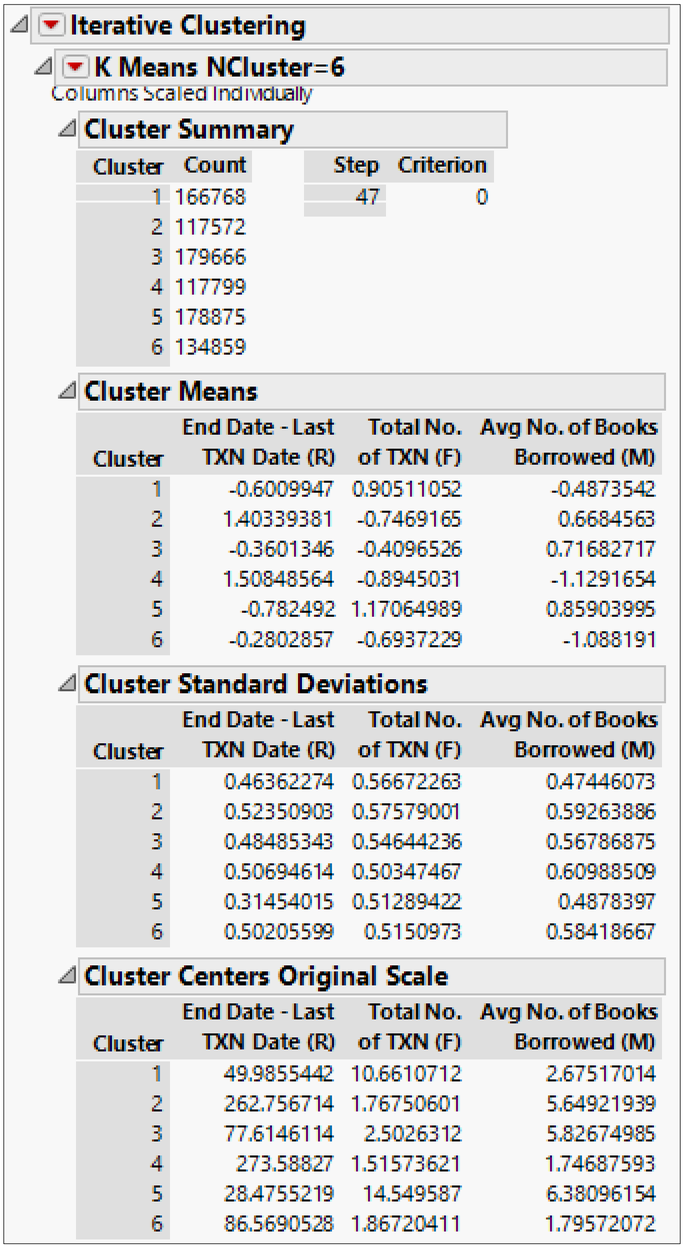


Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	5	-94.819	
K-Means Clustering	6	-45.927	
K-Means Clustering	7	-45.511	Optimal CCC

Columns Scaled Individually



Iterative Clustering

K Means NCluster=6

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	166768
2	117572
3	179666
4	117799
5	178875
6	134859

Step	Criterion
47	0

Cluster Means

Cluster	End Date - Last TXN Date (R)	Total No. of TXN (F)	Avg No. of Books Borrowed (M)
1	-0.6009947	0.90511052	-0.4873542
2	1.40339381	-0.7469165	0.6684563
3	-0.3601346	-0.4096526	0.71682717
4	1.50848564	-0.8945031	-1.1291654
5	-0.782492	1.17064989	0.85903995
6	-0.2802857	-0.6937229	-1.088191

Cluster Standard Deviations

Cluster	End Date - Last TXN Date (R)	Total No. of TXN (F)	Avg No. of Books Borrowed (M)
1	0.46362274	0.56672263	0.47446073
2	0.52350903	0.57579001	0.59263886
3	0.48485343	0.54644236	0.56786875
4	0.50694614	0.50347467	0.60988509
5	0.31454015	0.51289422	0.4878397
6	0.50205599	0.5150973	0.58418667

Cluster Centers Original Scale

Cluster	End Date - Last TXN Date (R)	Total No. of TXN (F)	Avg No. of Books Borrowed (M)
1	49.9855442	10.6610712	2.67517014
2	262.756714	1.76750601	5.64921939
3	77.6146114	2.5026312	5.82674985
4	273.58827	1.51573621	1.74687593
5	28.4755219	14.549587	6.38096154
6	86.5690528	1.86720411	1.79572072

From the results, even though k=7 gives us the optimal Cubic Clustering Criterion (CCC), our team decided to go with k=6, as the difference in CCC is not much (-45.511 vs -45.927), and a smaller no. of clusters may allow us to make more intuitive interpretation of the clustering results. A further discussion of the clustering results can be found in 3.3. RFM Analysis.

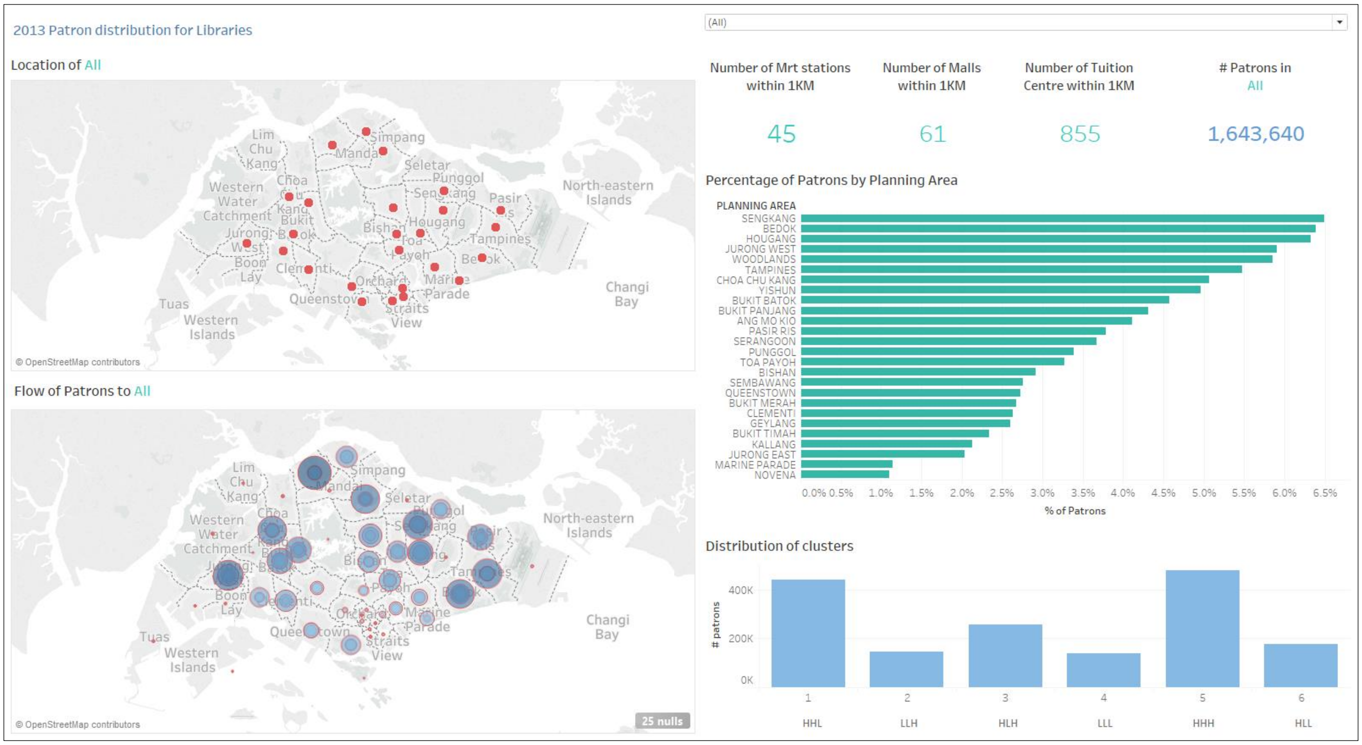
3. Initial Visualization & Findings

Listed in the following sections are the initial visualization & findings presented to the NLB Analytics Team in Sponsor Meeting 01.

3.1 Patron Flow at Library Level

3.1.1 Features

The dashboard visualisation below allows users to understand the flow of patrons to each of the libraries using the FY2013 datasets. The following section will explain the different features in the dashboard.



2013 Patron distribution for Libraries

(All)

Location of All

Flow of Patrons to All

Number of Mrt stations within 1KM: 45

Number of Malls within 1KM: 61

Number of Tuition Centre within 1KM: 855

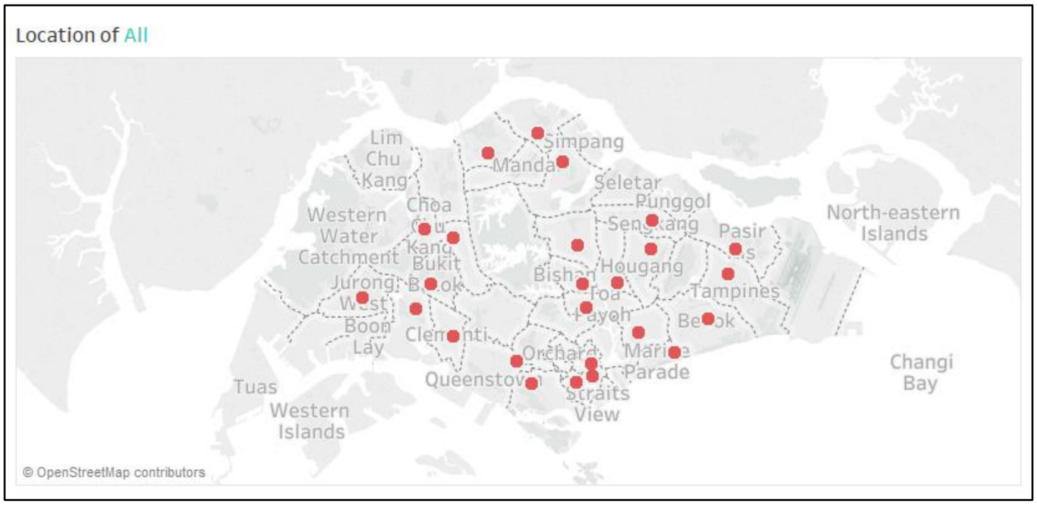
Patrons in All: 1,643,640

Percentage of Patrons by Planning Area

Distribution of clusters

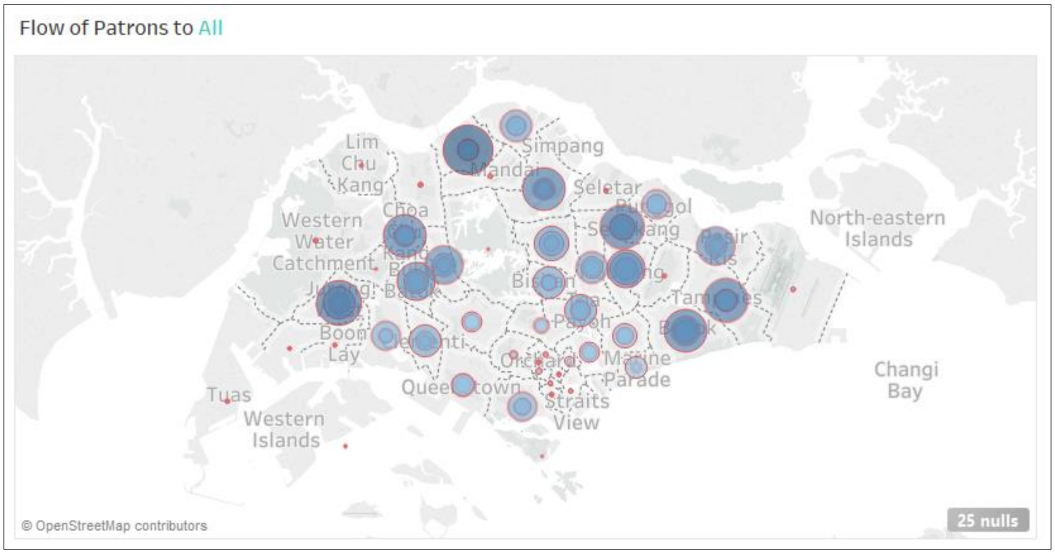
3.1.1.1 Location of Library

The map visualises the location of the selected library.



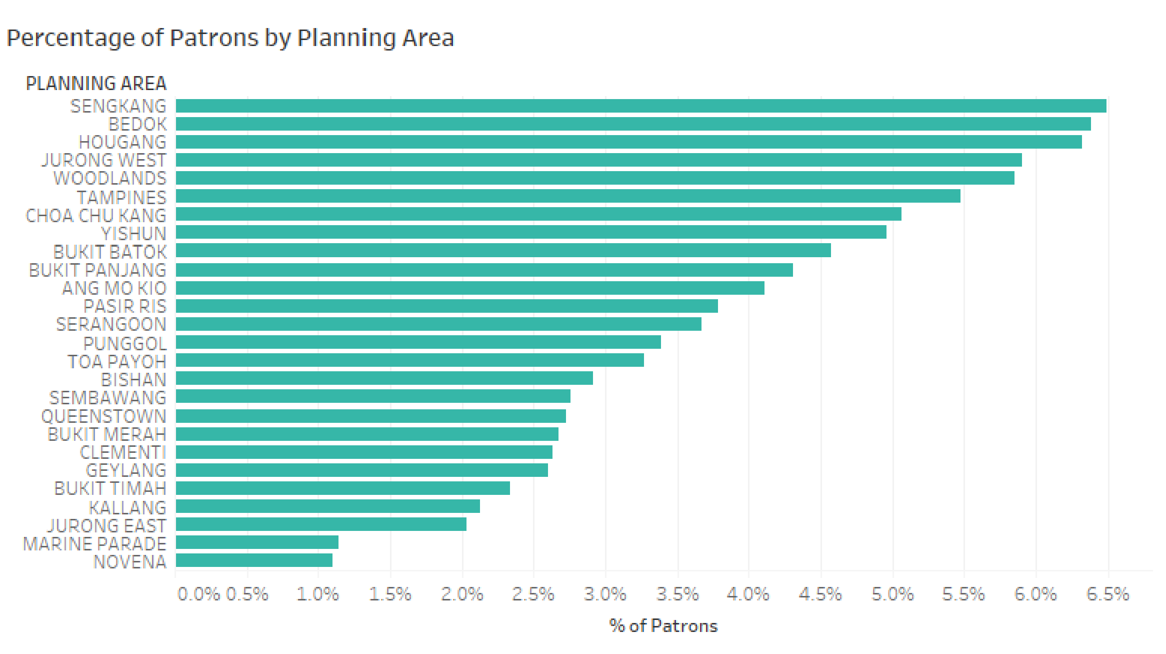
3.1.1.2 Geographical distribution of patrons

The map visualises the flow of patrons (by Planning Area) to the selected library.



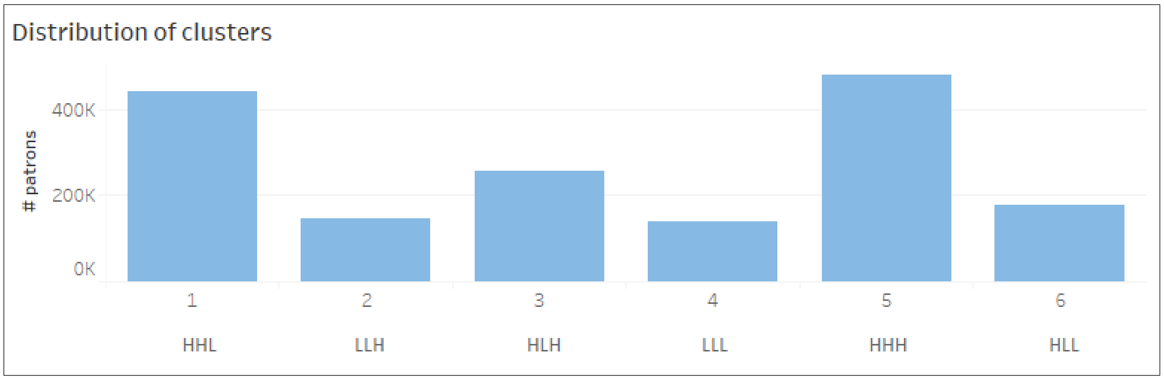
3.1.1.3 Distribution of patrons

The bar chart is sorted in descending order to identify the top Planning Areas with higher number of patrons for the selected library.



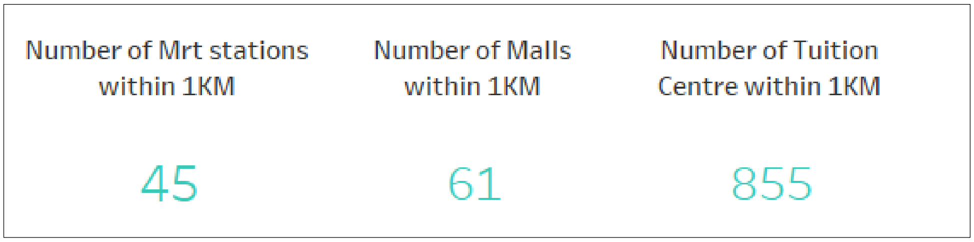
3.1.1.4 Distribution of patrons by Clusters

The distribution of patrons by clusters (from RFM Analysis in Section 3.2) for selected library is as shown.



3.1.1.5 Number of nearby amenities from Library

The number of nearby amenities (MRT stations, Malls and Tuition Centres) within 1 km from the selected library.



Number of Mrt stations within 1KM

45

Number of Malls within 1KM

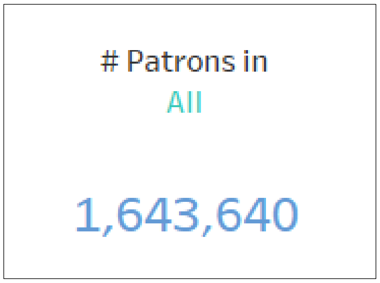
61

Number of Tuition Centre within 1KM

855

3.1.1.6 Number of patrons in Library

The number of unique patrons visiting the selected library.



Patrons in All

1,643,640

3.1.2 Findings

3.1.2.1 Community Library vs Regional Library

Through the initial visualisations, the team has discovered some patterns in the *Patron Dataset* provided. The team has observed different patterns of distribution of patrons for community libraries and regional libraries, where majority of the patrons from the community libraries tend to come from only the Planning Area closest to the library, whereas majority of the patrons from the regional libraries are dispersed across more Planning Areas nearer to the library.

For example, comparing the patron distribution for Jurong West Library (community library) and Jurong Regional Library (regional library), 65.91% of the patrons at Jurong West Library (Figure 3.1) are from Jurong West, where the Planning Area is closest to the library.

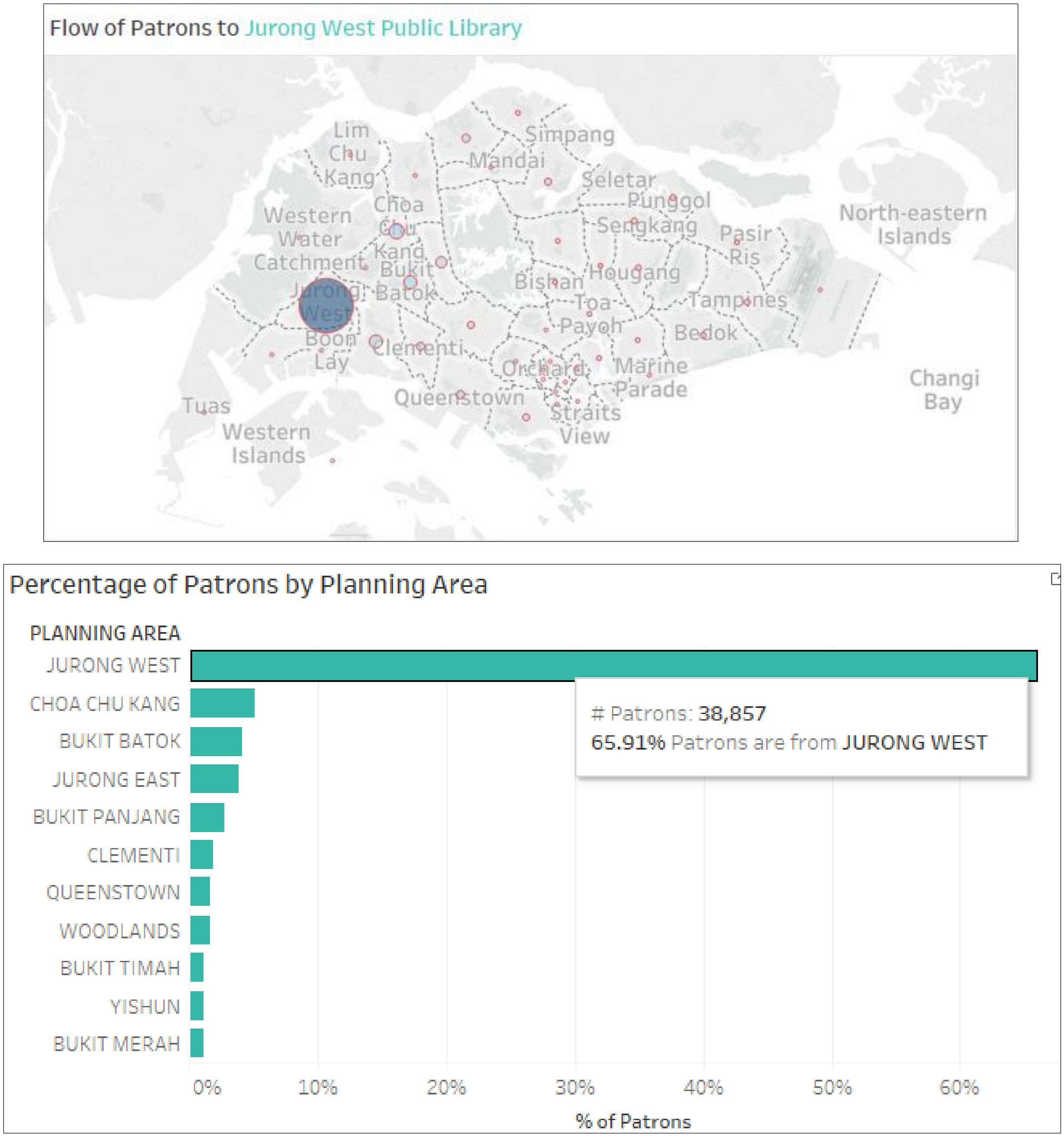
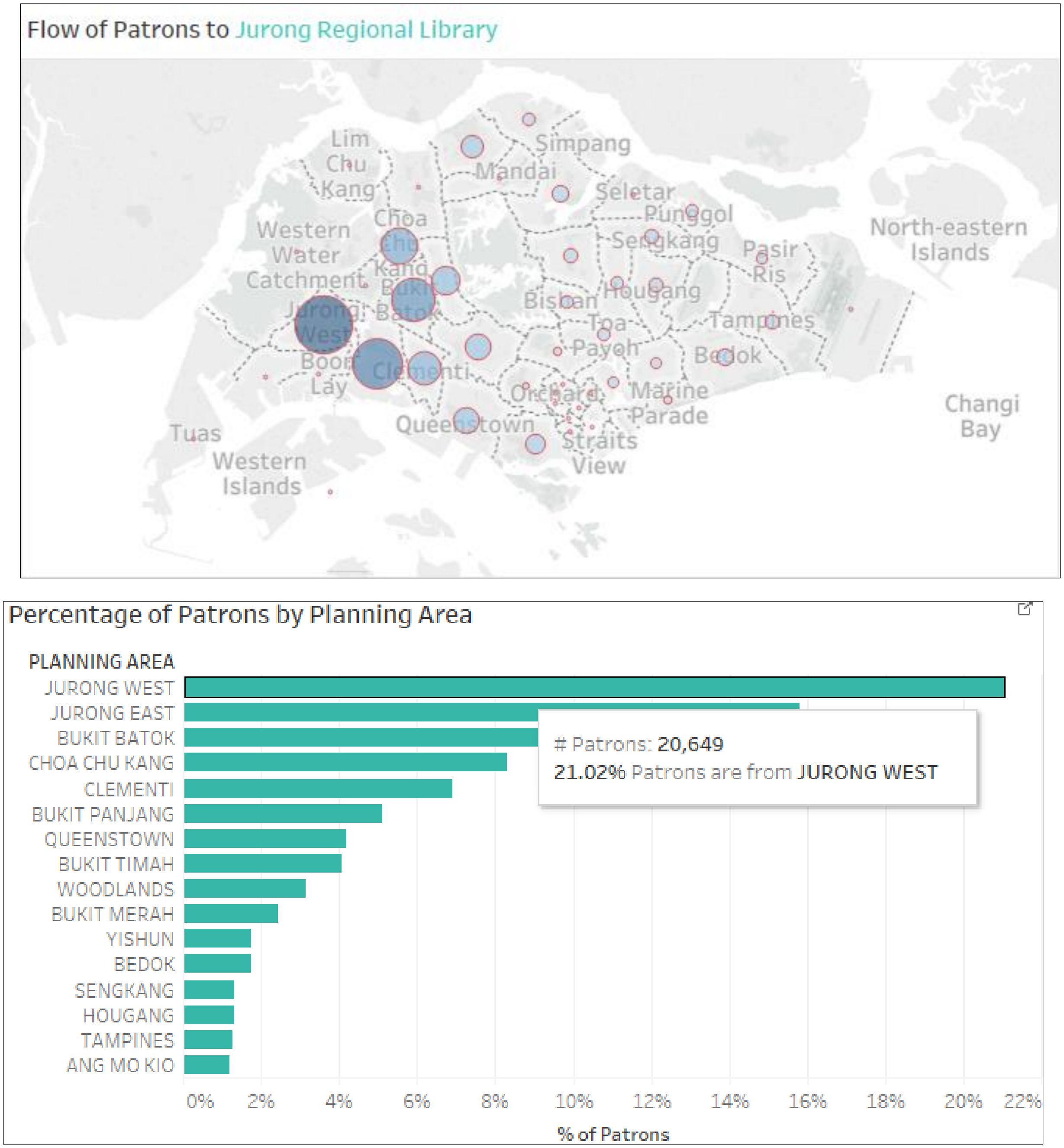


Figure 3.1 The distribution of patrons at Jurong West Library.

On the other hand, majority of the patrons to Jurong Regional Library are dispersed across more Planning Areas (Figure 3.2) such as Jurong West, Jurong East and Bukit Batok which are located nearer to the library.


Flow of Patrons to Jurong Regional Library

Percentage of Patrons by Planning Area

Figure 3.2 The distribution of patrons at Jurong Regional Library.

Regional library tends to have a larger collection of materials and a greater floor space as compared to the community library, which may help explaining why there are more patrons coming from different areas of Singapore. However, both community and regional library are usually more localised, thus drawing patrons who are living nearby.

3.1.2.1 Libraries located in the Central Region

Unlike the trend observed in the previous section where patrons tend to live near to the libraries, the team has discovered that libraries such as Central Public Library, library@chinatown, and library@esplanade that are located in the central region (Figure 3.3) tended to draw in patrons from many different areas across Singapore.

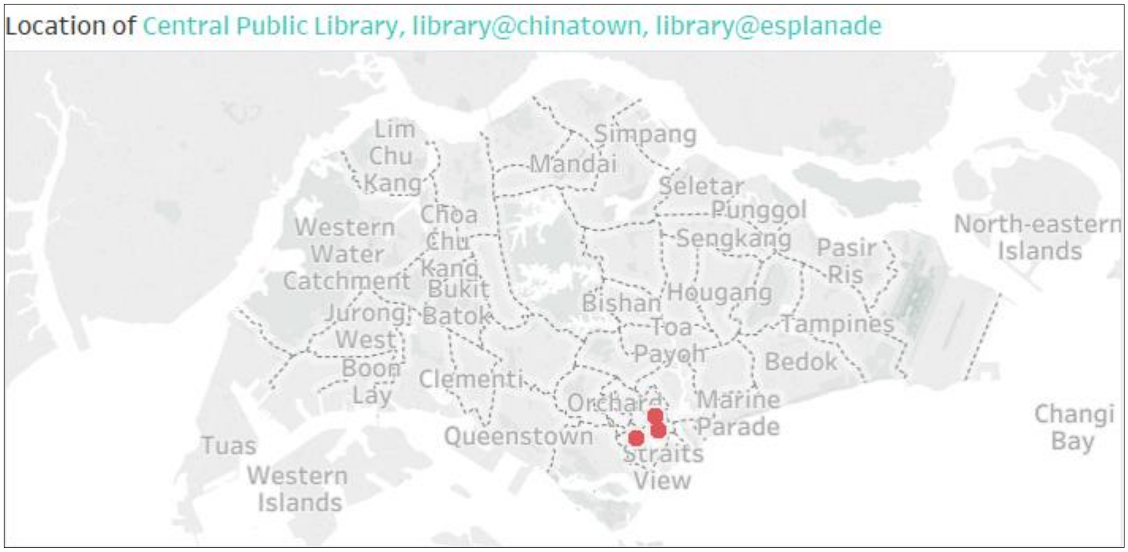


Figure 3.3 Location of Central Public Library, library@chinatown, and library@esplanade.

The trend is apparent in all 3 libraries where the distributions of patrons are quite even distributed across Singapore (*Figure 3.4, 3.5 and 3.6*) and this could be due to the accessibility to the libraries and the amenities around the libraries.

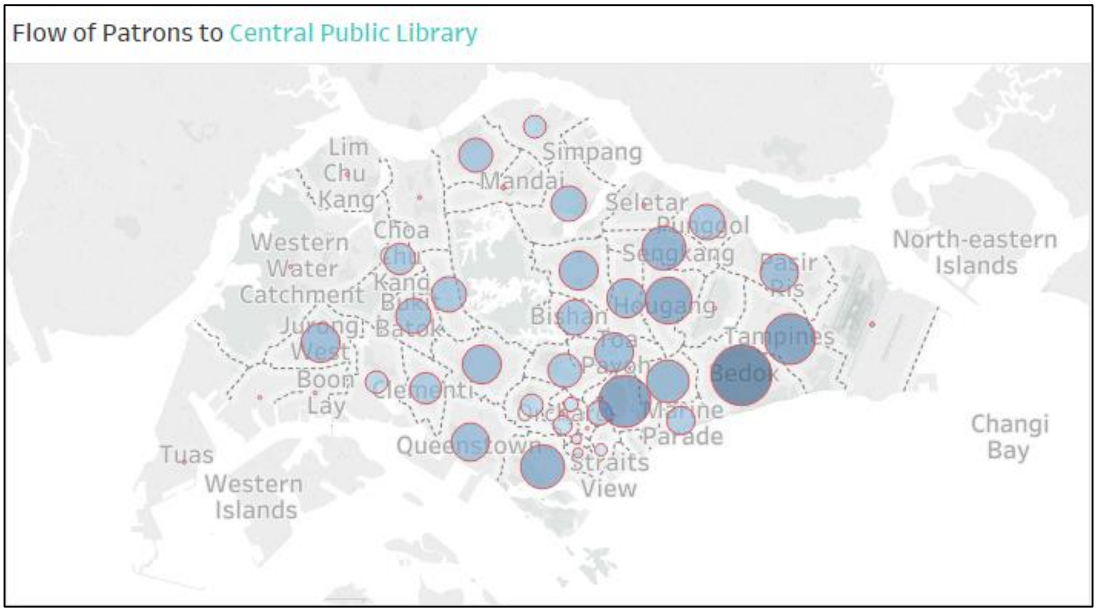


Figure 3.4 Distribution of patrons at Central Public Library. Patrons are distributed evenly across Singapore.

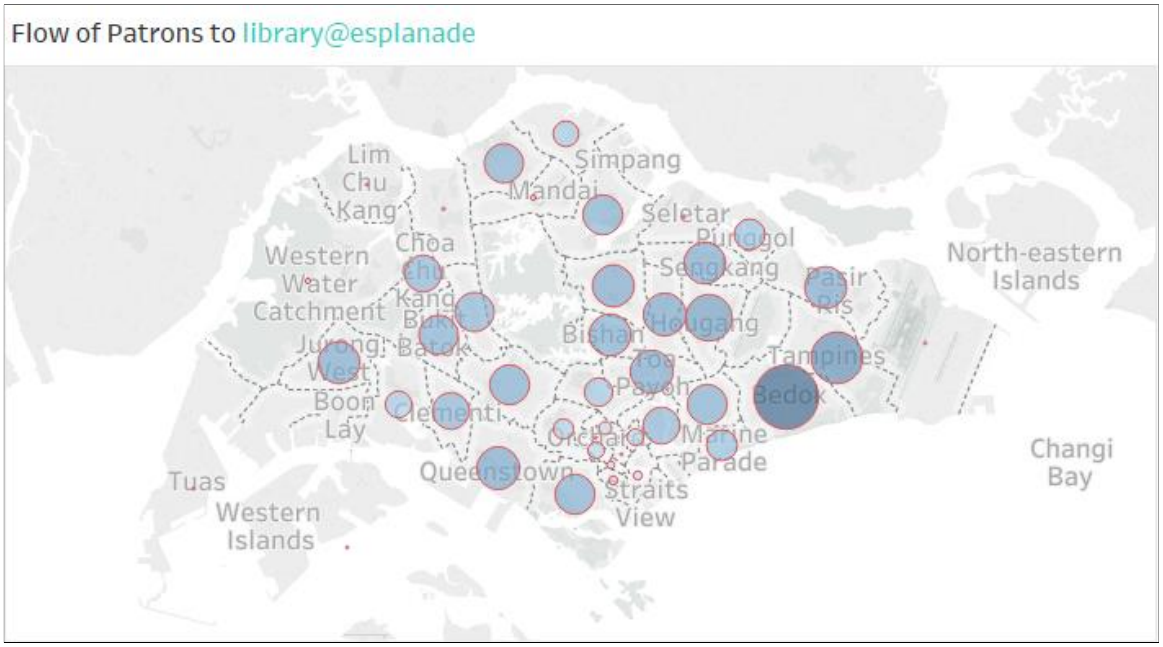


Figure 3.5 Distribution of patrons at library@esplanade. Patrons are distributed evenly across Singapore.

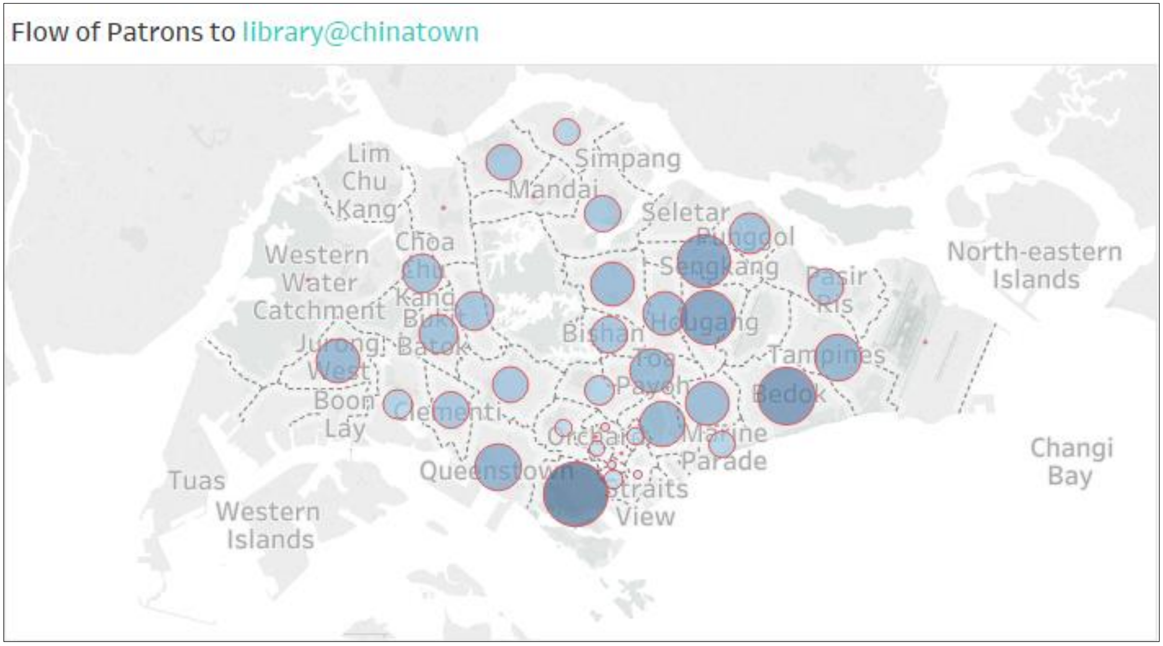
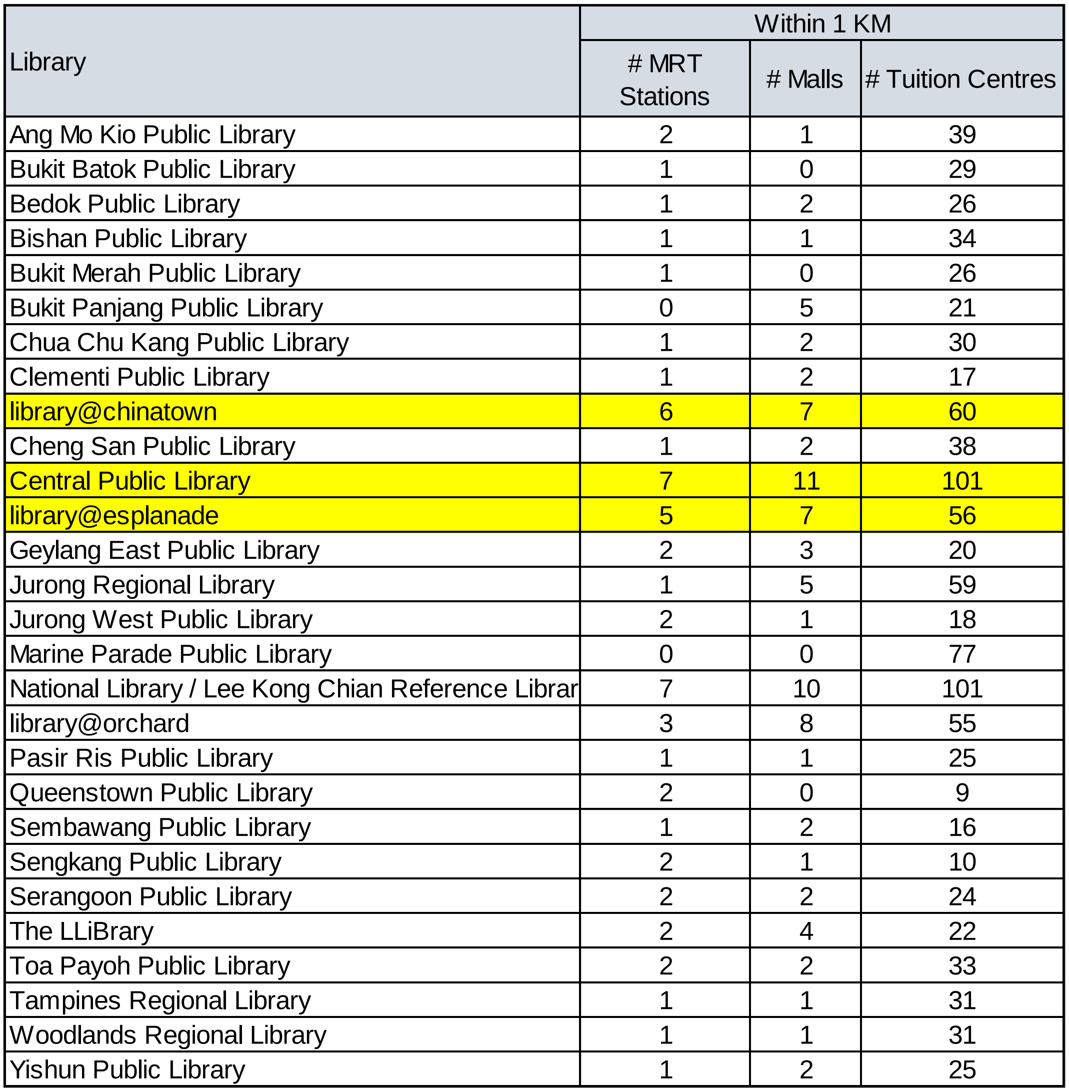


Figure 3.6 Distribution of patrons at library@chinatown. Patrons are distributed evenly across Singapore.

All 3 libraries have a considerably high number of amenities as compared to the other libraries as shown below:

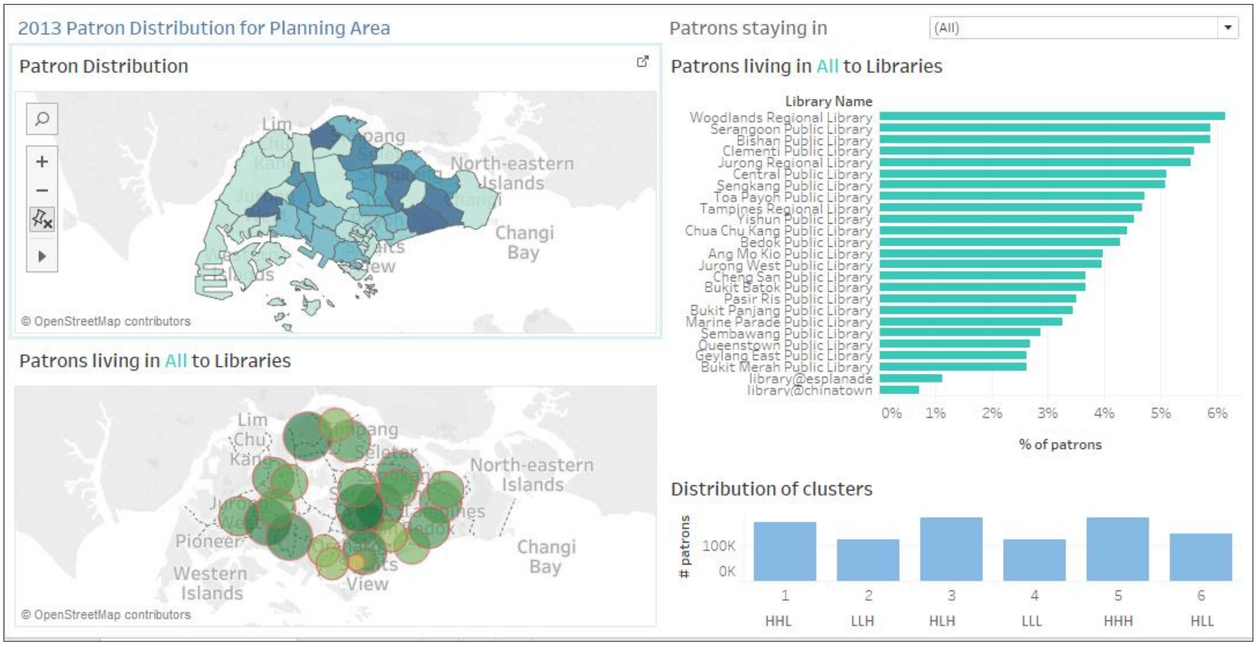


Library	Within 1 KM		
	# MRT Stations	# Malls	# Tuition Centres
Ang Mo Kio Public Library	2	1	39
Bukit Batok Public Library	1	0	29
Bedok Public Library	1	2	26
Bishan Public Library	1	1	34
Bukit Merah Public Library	1	0	26
Bukit Panjang Public Library	0	5	21
Chua Chu Kang Public Library	1	2	30
Clementi Public Library	1	2	17
library@chinatown	6	7	60
Cheng San Public Library	1	2	38
Central Public Library	7	11	101
library@esplanade	5	7	56
Geylang East Public Library	2	3	20
Jurong Regional Library	1	5	59
Jurong West Public Library	2	1	18
Marine Parade Public Library	0	0	77
National Library / Lee Kong Chian Reference Librar	7	10	101
library@orchard	3	8	55
Pasir Ris Public Library	1	1	25
Queenstown Public Library	2	0	9
Sembawang Public Library	1	2	16
Sengkang Public Library	2	1	10
Serangoon Public Library	2	2	24
The LLiBrary	2	4	22
Toa Payoh Public Library	2	2	33
Tampines Regional Library	1	1	31
Woodlands Regional Library	1	1	31
Yishun Public Library	1	2	25

3.2 Patron Flow at Planning Area Level

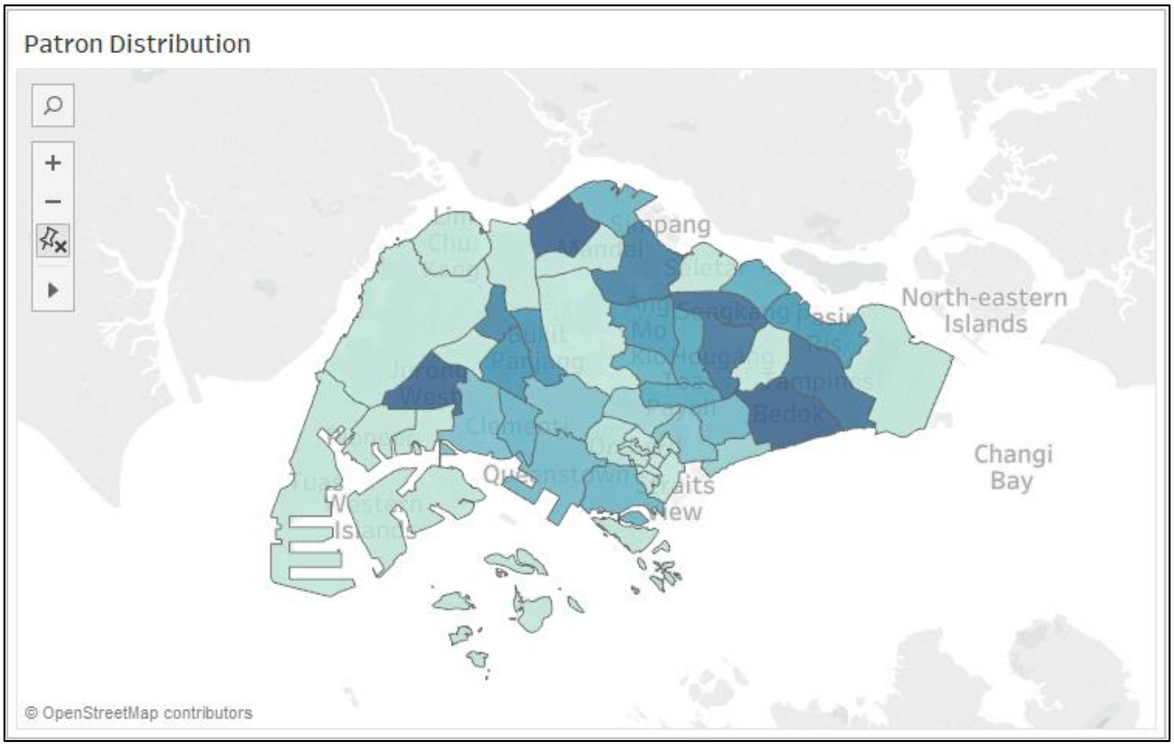
3.2.1 Features

The dashboard visualisation below allows users to understand the flow of patrons from each planning area using the FY2013 dataset. The following section will explain the different features in the dashboard.



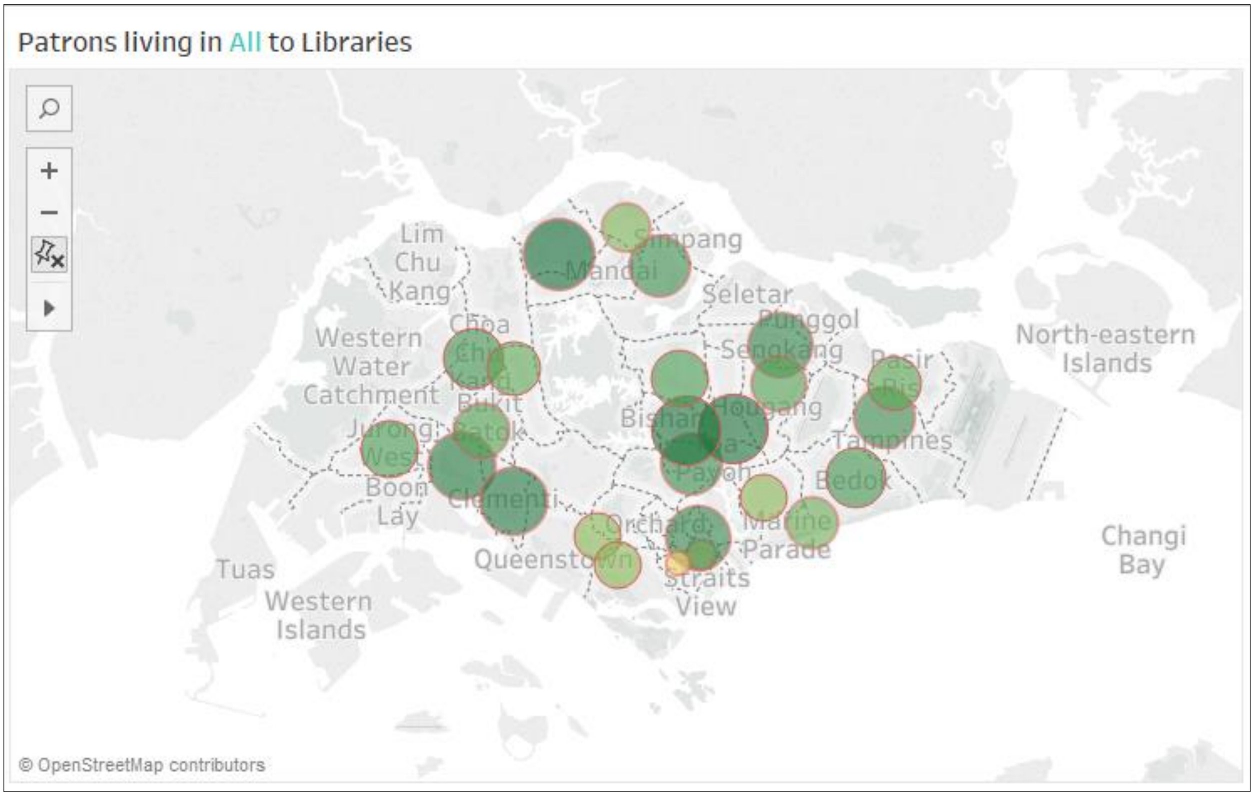
3.2.1.1 Location of Library

The map visualises the patron distribution across the country in terms of planning areas. Each polygon on the map is a planning area of the country.



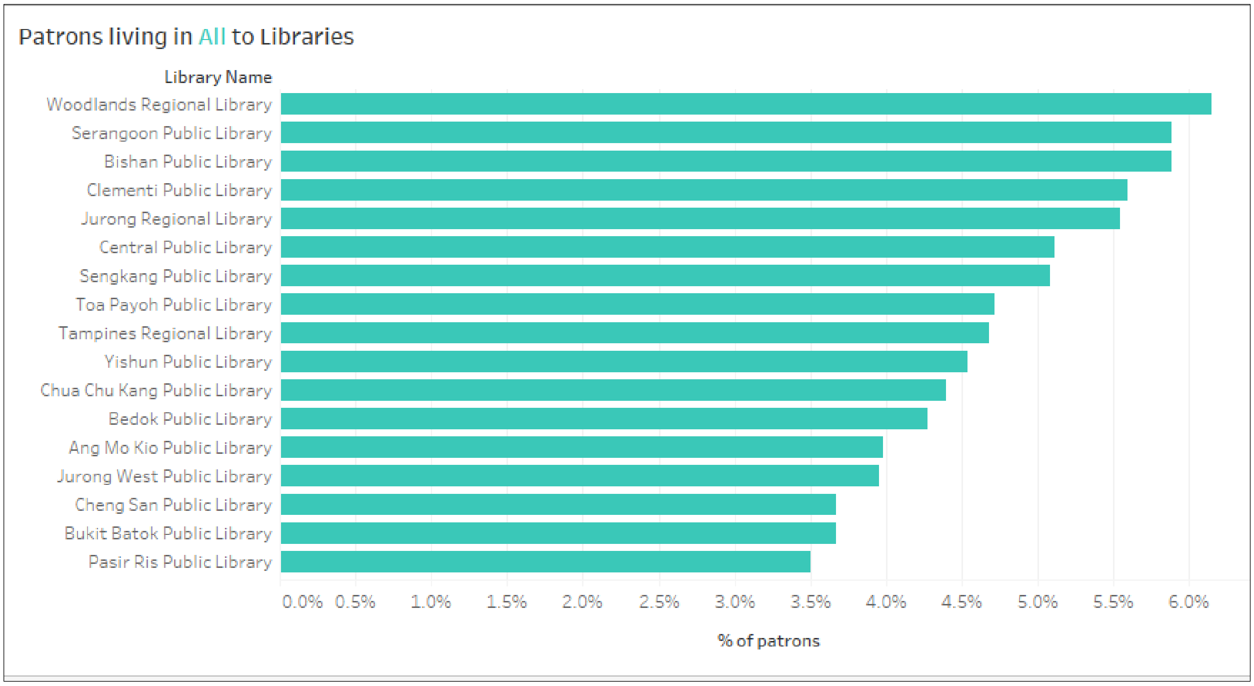
3.2.1.2 Geographical distribution of patrons

The map visualises the flow of patrons from a planning area to all libraries.



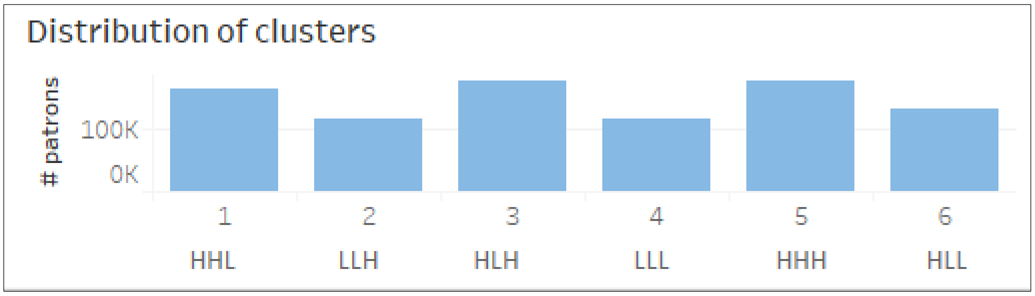
3.2.1.3 Distribution of patrons

The bar chart is sorted in descending order to identify the top Libraries with higher number of patrons from the selected region.



3.2.1.4 Distribution of patrons by Clusters

The distribution of patrons by clusters (from RFM Analysis in Section 3.2) for selected planning region.

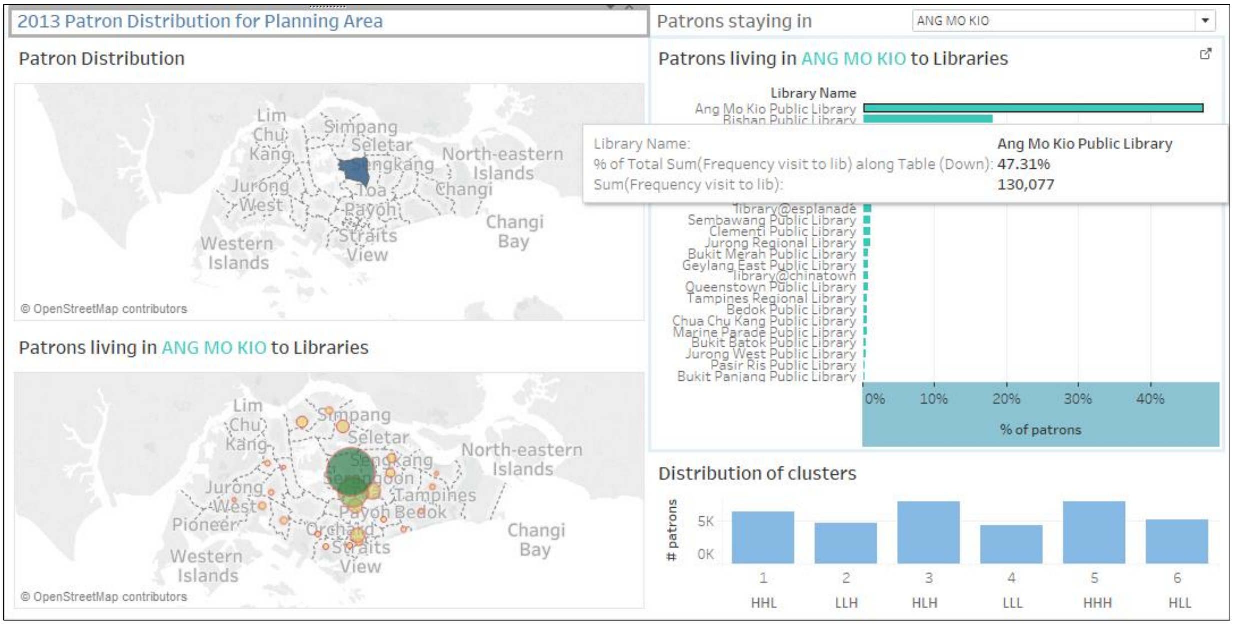


3.2.2 Findings

3.2.2.1 Proximity to library

Through the initial visualisations, the team has discovered some patterns in the *Patron Dataset* provided. There are different patterns of distribution of patrons for each planning area, where the majority of the patrons for each planning area tends to go to libraries near to their planning area.

One example of this trend is the behaviour of patrons living in the Ang Mo Kio planning area. From the visualization, we can see that the majority (47.31%) of those living in Ang Mo Kio tended to go to the Ang Mo Kio public library.



2013 Patron Distribution for Planning Area

Patrons staying in: ANG MO KIO

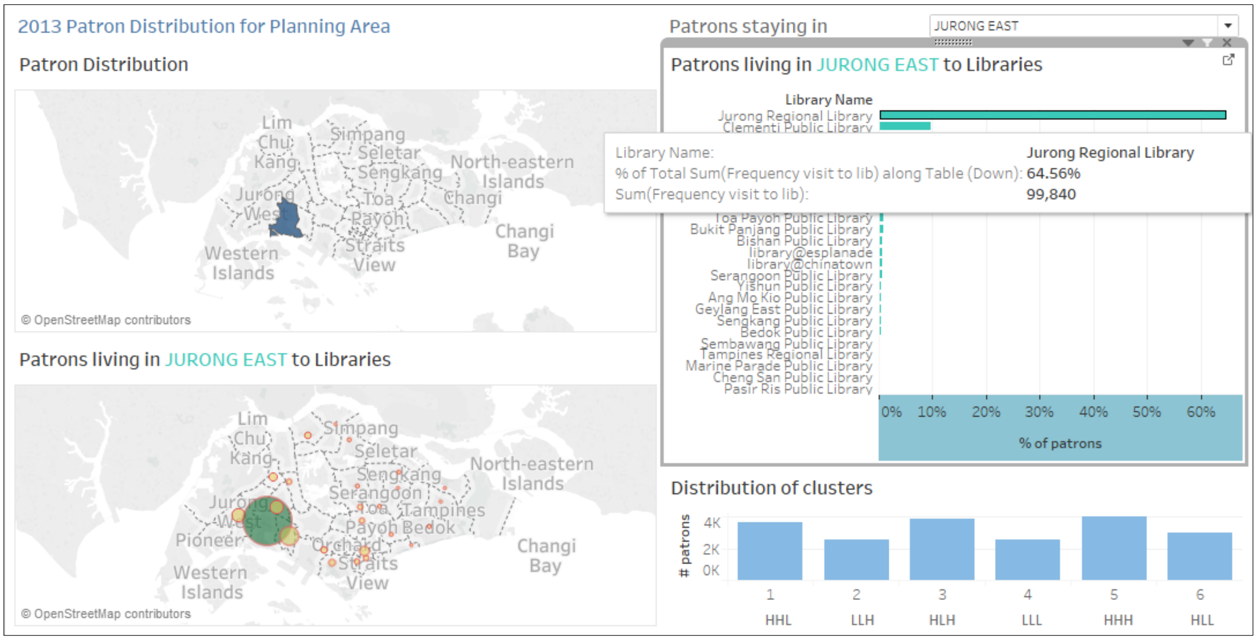
Patron Distribution

Patrons living in ANG MO KIO to Libraries

Library Name: Ang Mo Kio Public Library
% of Total Sum(Frequency visit to lib) along Table (Down): 47.31%
Sum(Frequency visit to lib): 130,077

Distribution of clusters

This trend remains the same for both public libraries (as shown above) and regional libraries (as shown below).



2013 Patron Distribution for Planning Area

Patrons staying in: JURONG EAST

Patron Distribution

Patrons living in JURONG EAST to Libraries

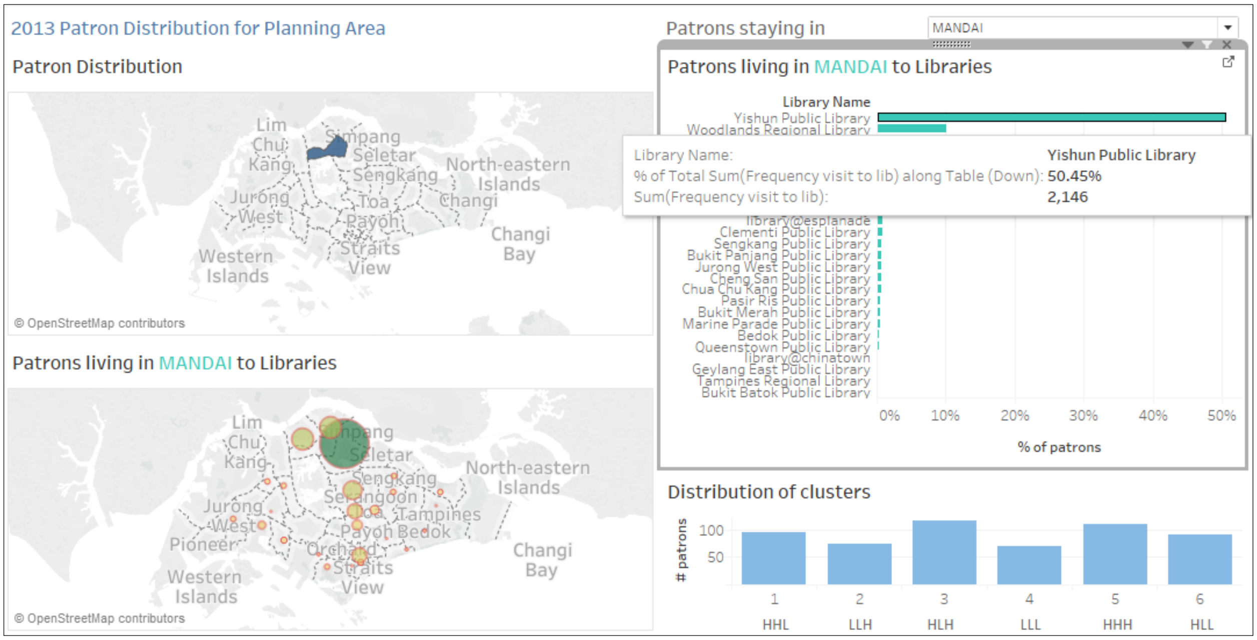
Library Name: Jurong Regional Library
% of Total Sum(Frequency visit to lib) along Table (Down): 64.56%
Sum(Frequency visit to lib): 99,840

Distribution of clusters

Therefore, the proximity of libraries to the planning area may be a significant factor to explain the patronage level to the library from the planning area.

3.2.2.2 Ease of travel to a library

Although we have established that proximity to a planning area is a key factor for determining patron traffic to a library, our visualization has uncovered another key factor, besides proximity, which is ease of travel from the planning area to the library. This trend is uncovered when the team attempted to visualize the patron flow for those living in Mandai planning area. The visualization is shown below.



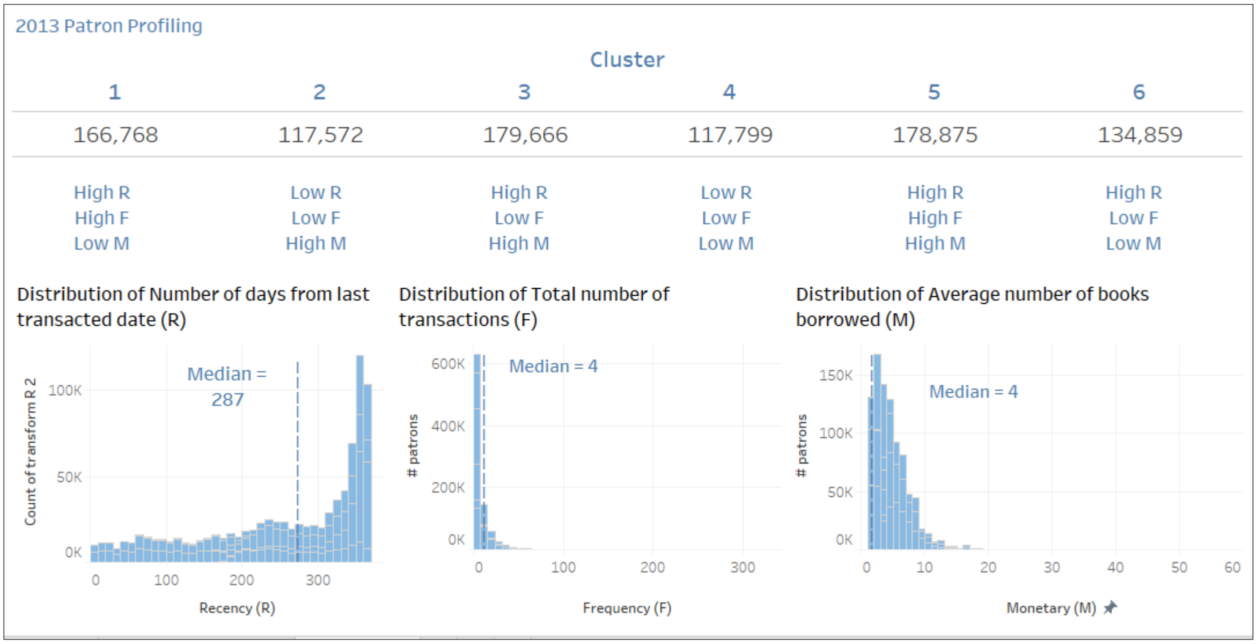
Library Name: Yishun Public Library
% of Total Sum(Frequency visit to lib) along Table (Down): 50.45%
Sum(Frequency visit to lib): 2,146

As there are no libraries in the Mandai planning area, patrons living in that area would have to travel to one of the three libraries in the surrounding planning areas instead. These libraries are Yishun Public Library, Woodlands Regional Library, and Sembawang Public Library. As Mandai is located such that it is equidistance to each of the 3 libraries, we expected that traffic to all three libraries would be approximately the same. Contrary to expectations, we see that that Yishun Public Library had the largest portion of Mandai patrons at 50.45%, nearly five times that of Woodlands Regional Library (10.04%) or Sembawang Public Library (9.87%). This can be attributed to the fact that there is a direct bus route to Yishun through Mandai. Therefore, we can conclude that ease of travel to the library may be an important factor in explaining the difference in patronage levels between libraries.

3.3 RFM Analysis

3.3.1 Features

After conducting clustering on the patron data set using the three attributes, Recency, Frequency, and Monetary (denoted as R, F, and M respectively), we have obtained 6 clusters, which we have visualized using Tableau. The following sections will explain the different features in the dashboard.

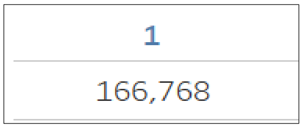


2013 Patron Profiling

Cluster					
1	2	3	4	5	6
166,768	117,572	179,666	117,799	178,875	134,859
High R High F Low M	Low R Low F High M	High R Low F High M	Low R Low F Low M	High R High F High M	High R Low F Low M

3.3.1.1 Patrons in cluster

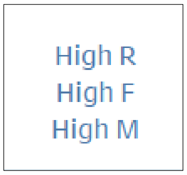
The numbers below the cluster number shows the number of patrons in that cluster.



1
166,768

3.3.1.2 Characteristics of cluster

The text below the number of patrons in the cluster shows the characteristics of that cluster.

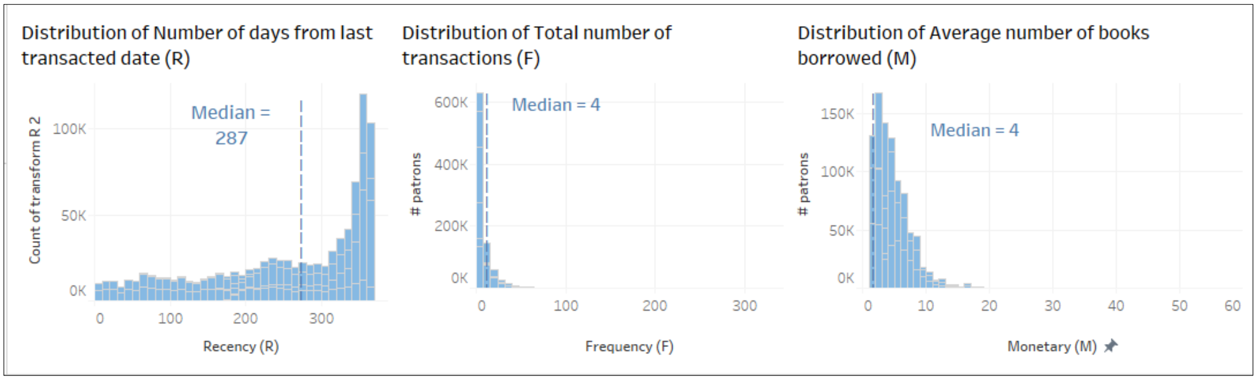


High R
High F
High M

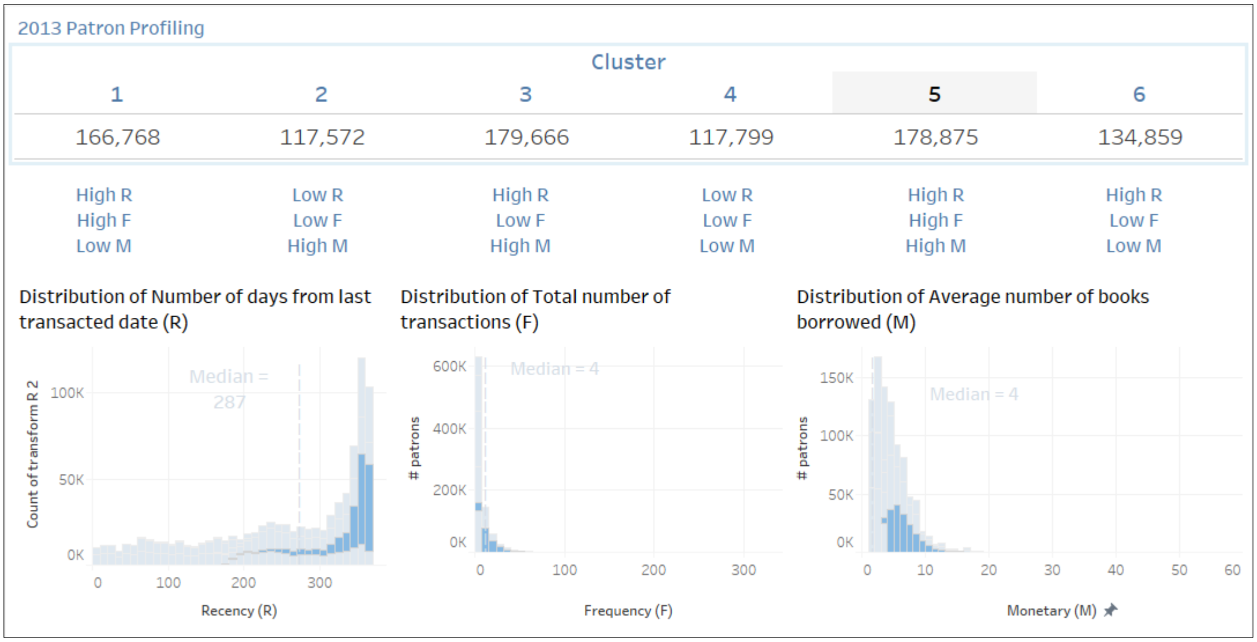
In this case, it shows that this cluster is characterized by patrons who borrowed books recently, at a frequent rate, with a large number of books borrowed at one time.

3.3.1.3 Distribution of patrons in Clusters

The distribution of patrons based on each attribute of R, F and M.



When selecting a cluster, the corresponding region in the bar charts will be selected to show the position of the patrons in the cluster.



2013 Patron Profiling

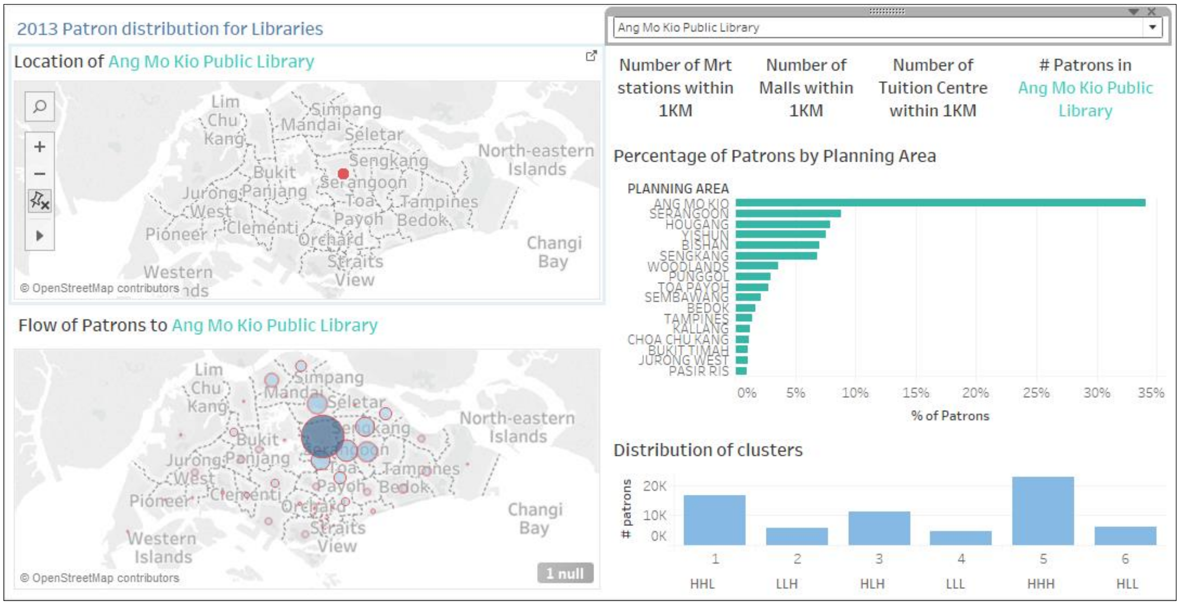
Cluster					
1	2	3	4	5	6
166,768	117,572	179,666	117,799	178,875	134,859
High R High F Low M	Low R Low F High M	High R Low F High M	Low R Low F Low M	High R High F High M	High R Low F Low M

In this case, cluster 5 is selected. The highlighted portion shows that a large number of patrons in cluster 5 have R, F and M values that are above the median.

3.2.2 Findings

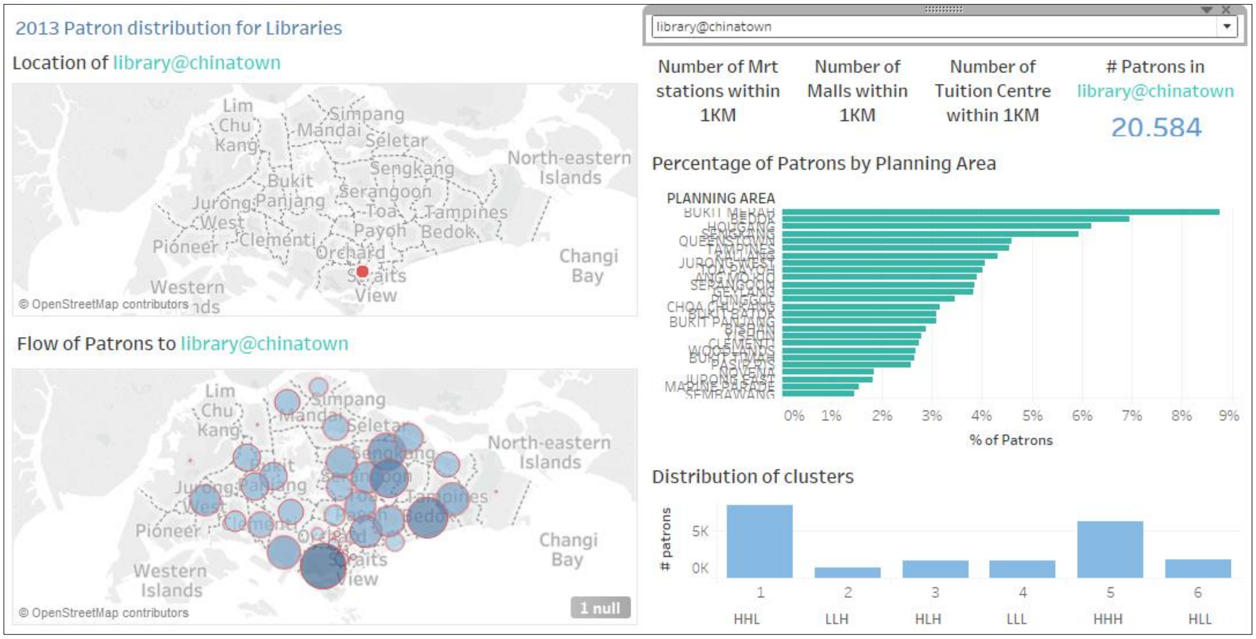
3.2.2.1 Clusters for Patrons in Library

The cluster distribution for patrons in a library is generally similar to the one shown below for Ang Mo Kio Public Library.

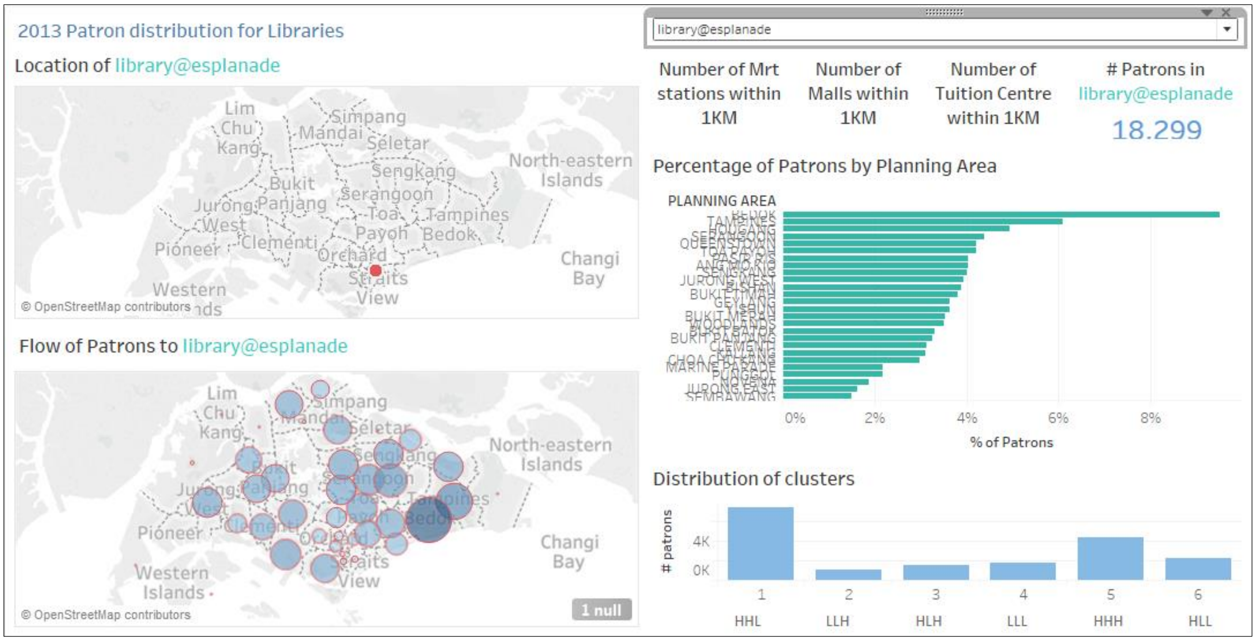


We can see that there are more active patrons (Clusters 1, 3 and 5) than non-active patrons (2, 4, 6). Patrons in the active group are characterized by high Recency, as well as high Frequency or Monetary. They can be considered active because usually, if one borrows frequently, they would need to borrow less books at a time, and vice versa. The non-active groups are characterized by low Recency, or low values for both Frequency and Monetary. This trend persists for both regional and public libraries.

An interesting trend can be observed for the Chinatown library, in that the proportion of patrons in cluster 3 is significantly less than usual.



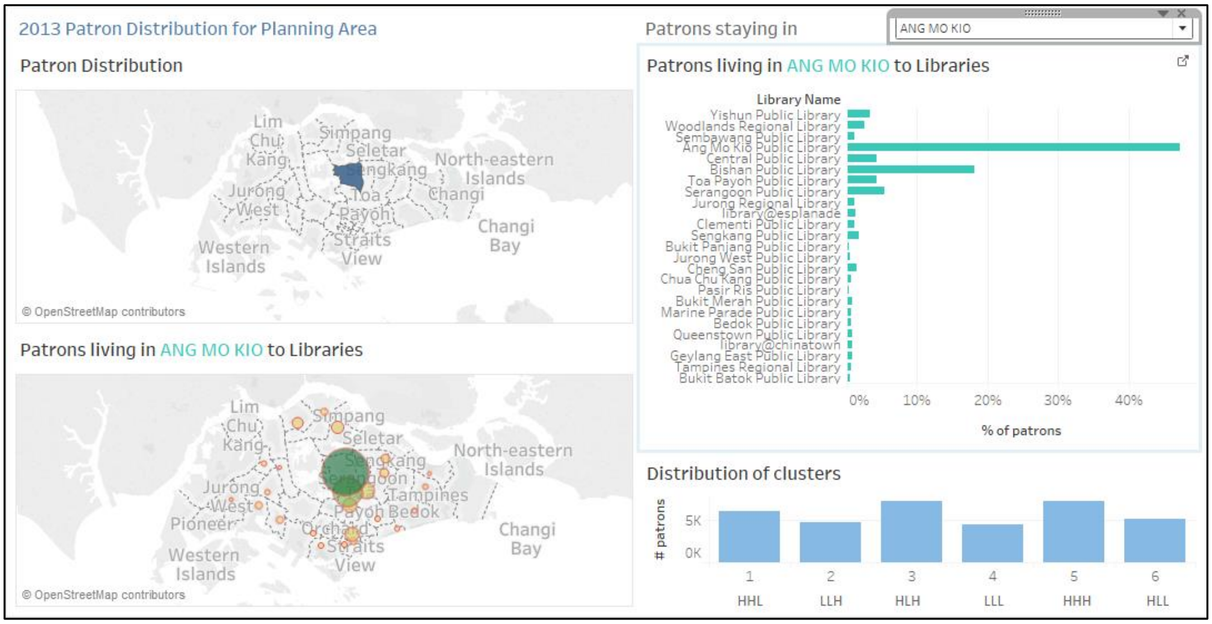
The same trend can be observed for Esplanade Library.



This shows that patrons going to these libraries tended to visit them very often.

3.2.2.2 Clusters for Patrons in Planning Areas

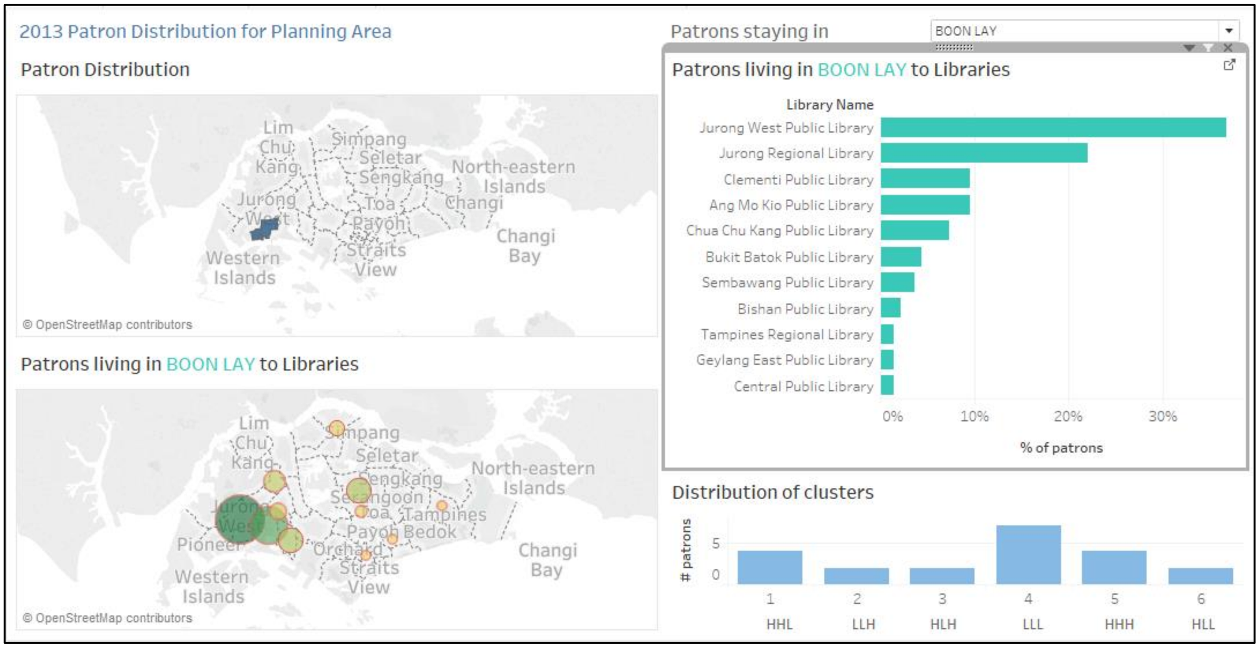
The cluster distribution in planning areas are generally similar to those for libraries, in that clusters 1,2 and 3 are more prominent. This is shown below with Ang Mo Kio planning area as an example.



2013 Patron Distribution for Planning Area

Patrons staying in: ANG MO KIO

However, there are some interesting patterns that can be observed for the few planning areas that deviate from this trend.



2013 Patron Distribution for Planning Area

Patrons staying in: BOON LAY

We can see above that the Boon Lay planning area is dominated by patrons belonging to cluster 4, having lows values in R, F and M. These are patrons that have not borrowed a lot of books, and have not visited the library for a long time. More attention could be paid to these patrons in order to re-engage them.

4. Revision of Methodology

After a discussion with NLB and our supervisor Prof Kam, we have revised the methodology of our project as described in the sections that follow.

4.1 Shiny R

We will now be using only Shiny R, instead of using a combination of Shiny R and Apache Spark as mentioned in our proposal. We have chosen Shiny R mainly due to the statistical function provided by R, which would be useful to our project. R also have many ready-made visualizations which would hopefully save us some time required for coding.

4.1.1 Leaflet

As the project has some geo-spatial nature, we have decided to use the leaflet package available in Shiny R for our visualization and analysis. The visualizations of the Singapore base map, as well as the various layers that will be added in like Libraries, Tuition Centres, Planning Areas, will all be done using leaflet. Users will also be able to select layers and adjust other input variables with this.

4.1.2 D3

For visualizations not provided in R libraries, we intend to make use of D3 to bridge the gap.

5. Revised Scope of Work

Scope of Work	Check
Data Preparation	
Remove anomalies	✓
Document the anomalies	✓
Document the assumptions	✓
Transform skewed data	✓
Match Subzone to URA Planning Area	✓
Aggregate data for visualisation	✓
Match Branch code to Library name	✓
Match libraries to coordinates (Lat Lng)	✓

Initial visualisation with Tableau	
Identify variables for RFM	✓
Perform cluster analysis on patrons based on RFM	✓
Update NLB sponsor on project progress	✓

Dashboard 1 - Distribution of patrons at Library Level	
Discover trends or patterns on the flow of patrons in each library	✓
Visualise the geographical locations of libraries	✓
Visualise the flow of patrons to each library geographically	✓
Visualise the distribution of patrons by clusters	✓
Display the number of nearby amenities for selected library	✓
Display the number of unique patrons in a library	✓
Visualise the number of patrons by planning area in bar chart	✓

Dashboard 2- Distribution of patrons at Planning Area Level	
Discover trends or patterns on the flow of patrons in each Planning Area	✓
Visualise the distribution of patrons with choropleth map	✓
Visualise the flow of patrons in each Planning Area to libraries geographically	✓
Visualise the distribution of patrons in a Planning Area by clusters	✓
Display the number of unique patrons in a Planning Area	✓
Visualise the number of patrons by libraries in bar chart	✓

Dashboard 3 - RFM Analysis	
Display the distribution of R, F and M individually in histogram	✓
Display the characteristics of clusters	✓
Display the number of patrons that falls in each cluster	✓
Indicate the median of R, F and M in the respective histograms	✓
Allow for highlighting of parts of the histograms upon hover of a cluster	✓
Interpret the characteristics of cluster from the visualization	✓

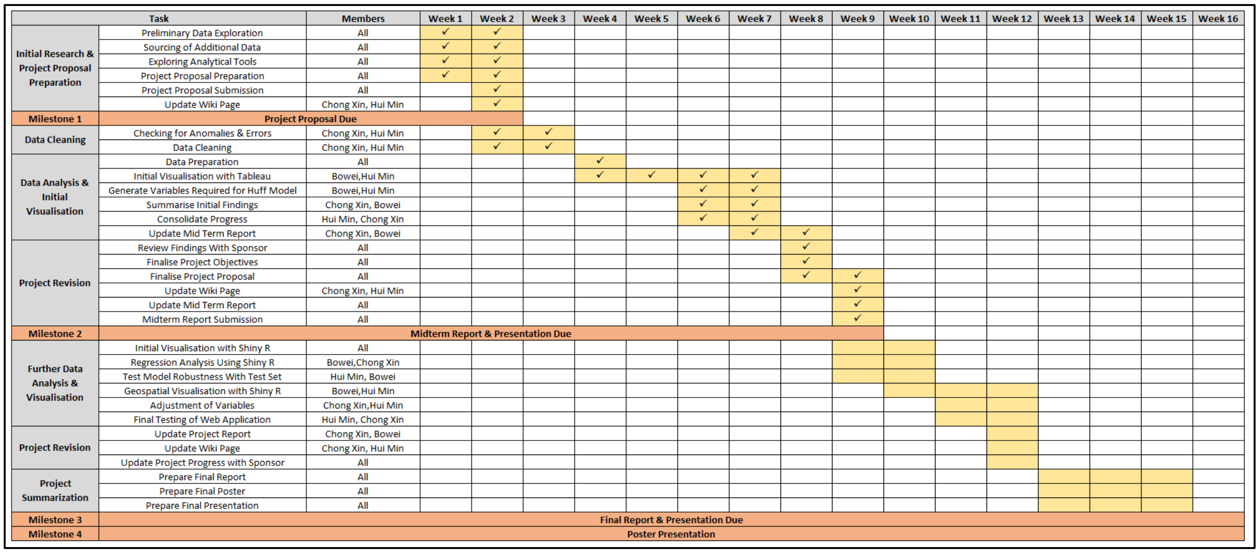
Visualization with Shiny R	
Visualisation from Tableau to be visualised with Shiny R	
Perform regression analysis on the selected variables	
Implement the Huff Model	
Test the Huff Model and adjust accordingly	
Allow the removal of a library	
Allow the adding of a library on a location on the map	
Assess the impact on change in location of library	
Provide interpretation on the impact of change	
Testing of web application 1 (Initial)	
Testing of web application 2	
Testing of web application 3 (Final)	

Project Milestones	
Proposal Report	
Prepare proposal report	✓
Update project wiki	✓
Submission of report	✓

Midterm Report	
Prepare midterm report	✓
Update project wiki	✓
Submission of report	✓

Final Report & Presentation	
Prepare final report	
Prepare final presentation slides	
Update project wiki	
Submission of report	
Submission of poster	

6. Revised Work Plan



Task		Members	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15	Week 16
Initial Research & Project Proposal Preparation	Preliminary Data Exploration	All	✓	✓														
	Sourcing of Additional Data	All	✓	✓														
	Exploring Analytical Tools	All	✓	✓														
	Project Proposal Preparation	All	✓	✓														
	Project Proposal Submission	All		✓														
	Update Wiki Page	Chong Xin, Hui Min		✓														
Milestone 1	Project Proposal Due																	
Data Cleaning	Checking for Anomalies & Errors	Chong Xin, Hui Min		✓	✓													
	Data Cleaning	Chong Xin, Hui Min		✓	✓													
Data Analysis & Initial Visualisation	Data Preparation	All				✓												
	Initial Visualisation with Tableau	Bowei,Hui Min				✓	✓	✓	✓									
	Generate Variables Required for Huff Model	Bowei,Hui Min						✓	✓									
	Summarise Initial Findings	Chong Xin, Bowei						✓	✓									
	Consolidate Progress	Hui Min, Chong Xin						✓	✓									
	Update Mid Term Report	Chong Xin, Bowei							✓	✓								
Project Revision	Review Findings With Sponsor	All								✓								
	Finalise Project Objectives	All								✓								
	Finalise Project Proposal	All								✓	✓							
	Update Wiki Page	Chong Xin, Hui Min									✓							
	Update Mid Term Report	All									✓							
	Midterm Report Submission	All									✓							
Milestone 2	Midterm Report & Presentation Due																	
Further Data Analysis & Visualisation	Initial Visualisation with Shiny R	All									▨	▨						
	Regression Analysis Using Shiny R	Bowei,Chong Xin									▨	▨						
	Test Model Robustness With Test Set	Hui Min, Bowei									▨	▨						
	Geospatial Visualisation with Shiny R	Bowei,Hui Min										▨	▨	▨				
	Adjustment of Variables	Chong Xin,Hui Min											▨	▨				
	Final Testing of Web Application	Hui Min, Chong Xin											▨	▨				
Project Revision	Update Project Report	Chong Xin, Bowei												▨				
	Update Wiki Page	Chong Xin, Hui Min												▨				
	Update Project Progress with Sponsor	All												▨				
Project Summarization	Prepare Final Report	All													▨	▨	▨	
	Prepare Final Poster	All													▨	▨	▨	
	Prepare Final Presentation	All													▨	▨	▨	
Milestone 3	Final Report & Presentation Due																	
Milestone 4	Poster Presentation																	

7. References

Brueckner, J. (2011). *Lectures on Urban Economics*. The MIT Press.

Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology, 92*(5), 1170-1182.

Bonacich, & Lloyd. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks, 23*(3), 191-201.

Dursun, & Caber. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives, 18*, 153-160.

Freeman, L. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215-239.

Huff, D. (1964). Defining and Estimating a Trading Area. *Journal of Marketing, 28*(3), 34-38.

Okabe, Atsuyuki, & Sugihara, Kokichi. (2012). Network Huff Model. In *Statistics in Practice* (pp. 213-230). Chichester, UK: John Wiley & Sons.

SMRT Journeys. (n.d.). Retrieved October 02, 2016, from *http://journey.smrt.com.sg/journey/mrt_network_map/*

Varma, A. (2016, May 15). More primary and secondary school students are getting private tuition years in advance of their grade in school. Retrieved August 25, 2016,

from *http://www.straitstimes.com/lifestyle/more-primary-and-secondary-school-students-are-getting-private-tuition-years-in-advance-of*