Assignment

Data Understanding Competition

REPORT

INDEX

Contents

Team:

Vidhan Dholakia (20151326)

Software Used & Data sets:

I are using Python for the coding purpose.
Gmap: Python graph plotting libraries to plot geo locations.
For the data sets, I have the Transit Data of Kingston City of the year 2017.

Instructions on how to Download & Install the software and the data sets:

How to Download Python & Install Python Map Plotting Libraries (Matplotlib, Pylab)?

Step 1: = Go to https://www.python.org/downloads/

Step 2: = Select my version and directly download as per the requirement and system adaptability.

Step 3: = The installation process is quite simple and straightforward. Just need to follow the obvious steps. I have used PyCharm IDE for execution of Python.

Installing MATPOTLIB & PYLAB: =

Step 1: = Go to https://matplotlib.org/3.1.1/users/installing.html

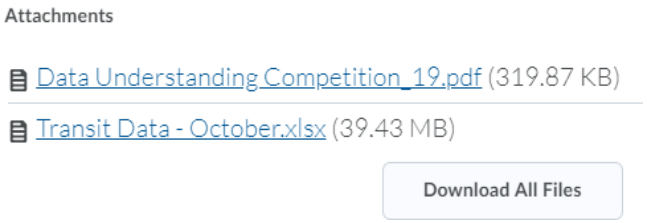
Step2: = Matplotlib and its dependencies are available as wheel packages for macOS, Windows and Linux distributions.

```
python -m pip install -U pip
python -m pip install -U matplotlib
```

How to Download the Data Set?

I have been given the data of people riding the bus in Kingston City in October 2017.

Each data point includes date/time, the type of boarding, bus and route number, and latitude/longitude where the boarding was recorded (note that around 8% of these Ire on buses with malfunctioning GPS antennae, and have the boarding coordinates recorded as 0,0). One note – I will see references to Routes 8 and 13 in the data. These routes do not have public timetables; these are extra buses to help service peak demand in certain areas of the City (primarily, the Union Street corridor betIen St. Lawrence College and Downtown).



Analysis Process Used:

I have the data about Kingston City's Transit from the month of October in the year 2017. The data is about the people riding the bus.

Initially, I had the total data set of 704540 items! This data set is not yet useful as it is not clean and not fit for the data modeling processes.

Date	Time	Class	Operation	Bus	Route	Latitude	Longitude
10/1/2017 0:00	1/1/1900 1:43	QUEENS	Exact Fare	620	17	44.22786	-76.4969383
10/1/2017 0:00	1/1/1900 1:43	QUEENS	Exact Fare	620	17	44.22786	-76.4969383
10/1/2017 0:00	1/1/1900 1:58	ADULT	Pass (Multi-ride card)	620	17	44.232035	-76.4913967
10/1/2017 0:00	1/1/1900 1:58	QUEENS	Exact Fare	620	17	44.232035	-76.4913967
10/1/2017 0:00	1/1/1900 1:58	QUEENS	Exact Fare	620	17	44.232035	-76.4913967
10/1/2017 0:00	1/1/1900 1:58	QUEENS	Exact Fare	620	17	44.232035	-76.4913967
10/1/2017 0:00	1/1/1900 5:47	QUEENS	Exact Fare	1687	7	44.26129667	-76.5073133
10/1/2017 0:00	1/1/1900 0:11	QUEENS	Exact Fare	620	17	44.22783667	-76.4970267
10/1/2017 0:00	1/1/1900 0:13	QUEENS	Exact Fare	620	17	44.22731667	-76.5001917
10/1/2017 0:00	1/1/1900 0:15	QUEENS	Exact Fare	620	17	44.223395	-76.5137567
10/1/2017 0:00	1/1/1900 0:23	QUEENS	Exact Fare	620	17	44.22776	-76.49559
10/1/2017 0:00	1/1/1900 0:27	ADULT	Pass (Multi-ride card)	620	17	44.23204833	-76.4913983
10/1/2017 0:00	1/1/1900 0:40	QUEENS	Exact Fare	620	17	44.22782	-76.4970317
10/1/2017 0:00	1/1/1900 0:40	QUEENS	Exact Fare	620	17	44.22782	-76.4970317
10/1/2017 0:00	1/1/1900 0:42	QUEENS	Exact Fare	620	17	44.22735167	-76.4998983
10/1/2017 0:00	1/1/1900 1:02	QUEENS	Exact Fare	620	17	44.23262	-76.5078217
10/1/2017 0:00	1/1/1900 1:13	QUEENS	Exact Fare	620	17	44.22732333	-76.4998467
10/1/2017 0:00	1/1/1900 1:13	QUEENS	Exact Fare	620	17	44.22732333	-76.4998467
10/1/2017 0:00	1/1/1900 1:13	QUEENS	Exact Fare	620	17	44.22732333	-76.4998467
10/1/2017 0:00	1/1/1900 1:22	QUEENS	Exact Fare	620	17	44.22367167	-76.499225
10/1/2017 0:00	1/1/1900 1:43	QUEENS	Exact Fare	620	17	44.22786	-76.4969383
10/1/2017 0:00	1/1/1900 1:43	QUEENS	Exact Fare	620	17	44.22786	-76.4969383

a. Data Preparation & Data Cleaning: =

I have reduced the data from the original data set by removing unwanted items. Firstly, I have sorted the data as per the time, bus number as Ill as the route number. By doing this, I have reduced my data set from 704540 to just 14086!

To avoid the redundancy, I will consider data of whole Iek (7 days). I will try to avoid the duplication of similar routes or coinciding routes whenever possible.

b. Feature Transformation & Feature Selection: =

I have use Sine and Cosine functions to get the location and to plot the points on the map. I have also added the location (coordinates) of the park nearby the bus route.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Date	Time	Bus	Route	Latitude	Longitude	Park_Location_X	Park_Location_Y	Park_Location_Z	Park_Latitude	Park_Longitude	Park_Location_X	Park_Location_Y	Park_Location_Z	Stop
2	10/3/2017	########	315	1	44.27231	-76.47819	0.20360329	0.200801936	0.711987897	44.2357422	-76.48823639	0.180283963	0.174265484	0.71900634	44
3	10/3/2017	########	315	1	44.26812	-76.47913167	0.200929204	0.197791742	0.712648811	44.25874718	-76.47917541	0.194520307	0.191466165	0.712679498	44
4	10/3/2017	########	315	1	44.26812	-76.47913167	0.200929204	0.197791742	0.712648811	44.35880621	-76.33462815	0.222150547	0.292976106	0.604201426	44
5	10/3/2017	########	315	1	44.26812	-76.47913167	0.200929204	0.197791742	0.712648811	44.22014021	-76.58364491	0.184231692	0.146864978	0.781945044	44
6	10/3/2017	########	315	1	44.26812	-76.47913167	0.200929204	0.197791742	0.712648811	44.26418821	-76.57979643	0.216846731	0.174233957	0.779540319	44
7	10/3/2017	########	315	1	44.263865	-76.47865667	0.197925482	0.195020126	0.712315509	44.21764566	-76.58937053	0.183164677	0.144306966	0.785501257	44
8	10/3/2017	########	315	1	44.263865	-76.47865667	0.197925482	0.195020126	0.712315509	44.33941667	-76.40043712	0.229075806	0.264061194	0.655294381	44
9	10/3/2017	########	315	1	44.26265	-76.48195833	0.197734065	0.193549008	0.714628928	44.25575982	-76.46739706	0.190226552	0.191703081	0.704367862	44
10	10/3/2017	########	315	1	44.26451667	-76.49050833	0.200673887	0.193094748	0.720583491	44.23901908	-76.49620004	0.183963357	0.175009778	0.724518247	4
11	10/3/2017	########	315	1	44.26451667	-76.49050833	0.200673887	0.193094748	0.720583491	44.23762274	-76.48012209	0.180162301	0.176998106	0.713343264	4
12	10/3/2017	########	315	1	44.26269167	-76.49091167	0.199487888	0.191798645	0.720863091	44.23472624	-76.51078904	0.183445457	0.169488125	0.734496339	44
13	10/3/2017	########	315	1	44.26269167	-76.49091167	0.199487888	0.191798645	0.720863091	44.26035067	-76.45585847	0.191077041	0.197056535	0.696130658	44
14	10/3/2017	########	315	1	44.26269167	-76.49091167	0.199487888	0.191798645	0.720863091	44.2639023	-76.37520845	0.176751642	0.214443737	0.636030799	44
15	10/3/2017	########	315	1	44.26338	-76.49818333	0.201357387	0.190797891	0.725883817	44.26053724	-76.4840922	0.196692262	0.191709115	0.716119947	44
16	10/3/2017	########	315	1	44.26171333	-76.49312333	0.199232358	0.190706805	0.722394183	44.22115572	-76.53910406	0.178253116	0.15557131	0.753414284	44
17	10/3/2017	########	315	1	44.26160167	-76.48933833	0.198431849	0.191385017	0.719771758	44.31029623	-76.52285326	0.239237706	0.215744229	0.742629622	44
18	10/3/2017	########	315	1	44.26160167	-76.48933833	0.198431849	0.191385017	0.719771758	44.25065016	-76.57706756	0.206236285	0.166636746	0.777828185	44
19	10/3/2017	########	315	1	44.26054667	-76.48896333	0.197630287	0.190755025	0.719511378	44.77435502	-76.72071477	0.611995643	0.363484308	0.859785727	44
20	10/3/2017	########	315	1	44.26080833	-76.48623	0.197288698	0.191469722	0.717610445	44.23210834	-76.57240282	0.19156861	0.156268031	0.774888052	44
21	10/3/2017	########	315	1	44.26103667	-76.48172167	0.196580331	0.192510813	0.714463358	44.25636791	-76.52477905	0.20134325	0.180869205	0.74391795	44
22	10/3/2017	########	315	1	44.25633333	-76.483165	0.193620595	0.189065654	0.715472475	44.23559626	-76.52899934	0.187130312	0.166679882	0.746731614	44
23	10/3/2017	########	315	1	44.25633333	-76.483165	0.193620595	0.189065654	0.715472475	44.2489621	-76.47027515	0.1861496	0.186517744	0.706407911	4
24	10/3/2017	########	315	1	44.25633333	-76.483165	0.193620595	0.189065654	0.715472475	44.22597444	-76.61141491	0.192765392	0.145100368	0.798951687	44
25	10/3/2017	########	315	1	44.25633333	-76.483165	0.193620595	0.189065654	0.715472475	44.22562277	-76.48372656	0.17247443	0.168227852	0.715864691	44

Apart from reducing the data, I have added some relevant features like location (co- ordinates) of the park in terms of Latitude and Longitude. This is done to find the correlation betIen the park and the bus route. By doing this, I can check the park ways which are nearby to the bus routes. This is just done so that I can predict if the bus route is accessible to everyone in the city.

Charts showing interesting Trends and Patterns:

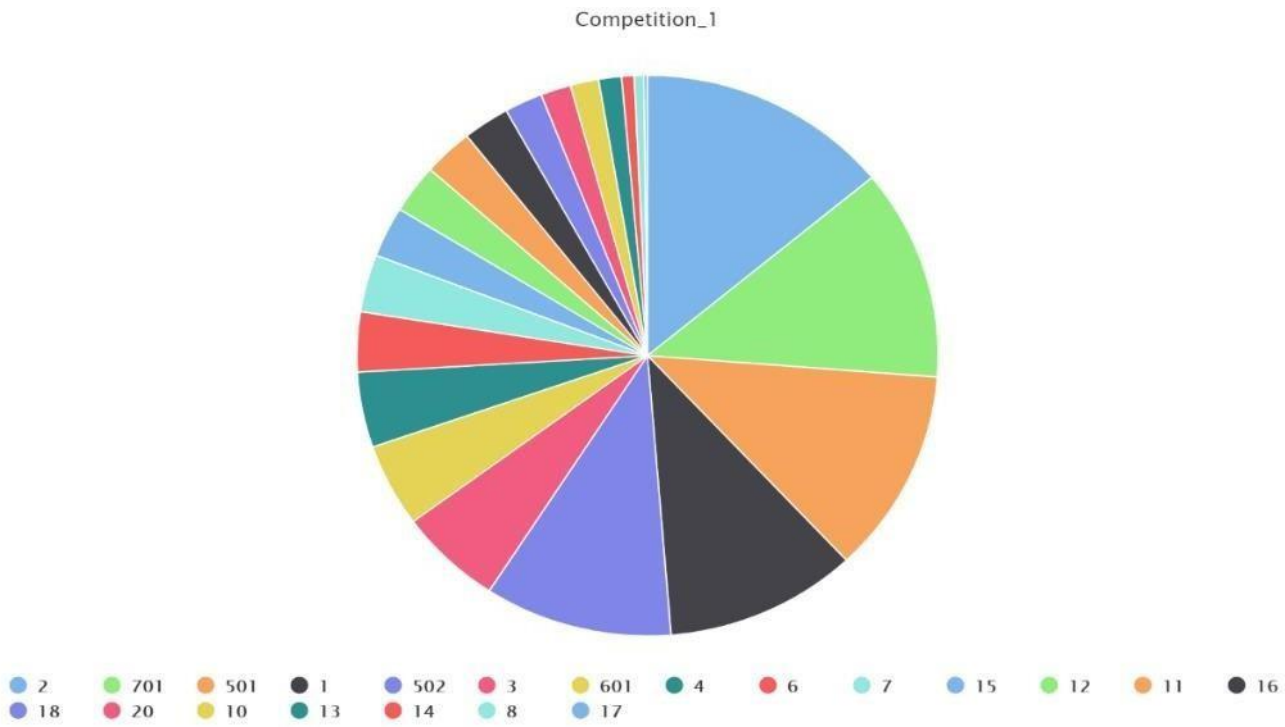


Diagram 1: Pie Chart Representation.

The above chart presents the classification of all the buses. I can deduce the conclusion that route number 701, 2, and 15 constitute a superior part of all the 21 buses. Approximately 14.28% of the transit is done from these 3 buses.

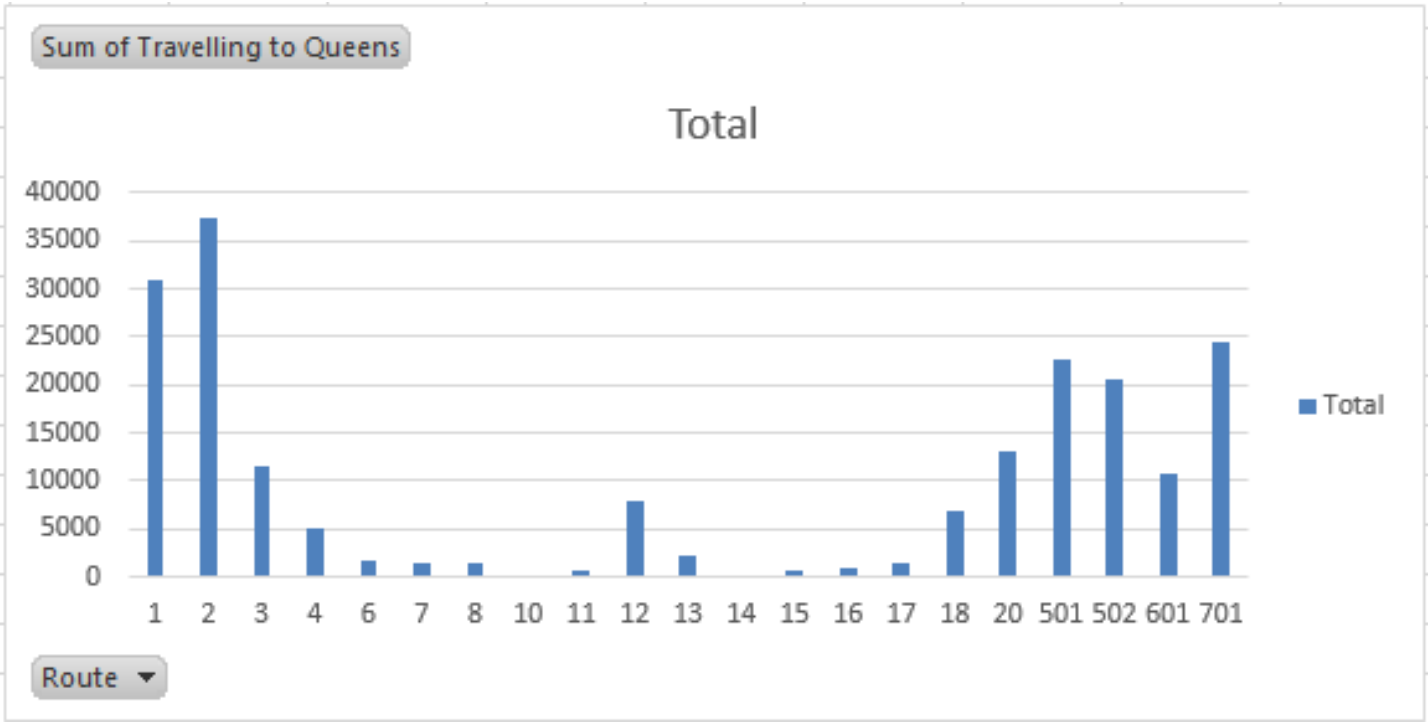


Diagram 2: Total Sum of Travelling to Queen's University

This is another representation of the total sum of people travelling to Queen's University. As I can see, maximum number of people travelling to Queen's prefer bus route number 2 which consists of nearly 90% of the people.

Row Labels	Count of Date
Sunday	49073
Monday	103352
Tuesday	127622
Wednesday	100935
Thursday	105790
Friday	101845
Saturday	60169
Grand Total	**648786**

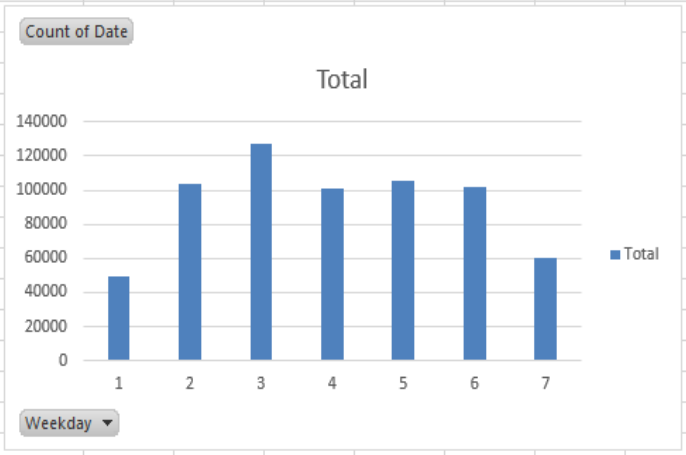


Diagram 3: Day wise Comparison

From this diagram, I can conclude that people tend to travel more on the Iek days as compared to the Iekends.

A majority of the people travel during Monday, and the counts goes on until the Iekend.

Count of Bus	Column Labels							
Row Labels	1	2	3	4	5	6	7	Grand Total
1	6414	11719	13171	10335	11010	10258	6179	69086
2	4373	12721	15218	15782	15216	13257	8886	85453
3	2633	3913	8284	6130	5951	6508	3674	37093
4	1785	4833	5450	4004	4825	4252	4561	29710
6		2237	3163	2319	2474	1473	1657	13323
7	2462	2956	4397	3279	3536	3035	2633	22298
8		358	589	487	365	486		2285
10	797	1803	2120	1733	1697	1857	968	10975
11	1216	2812	3381	2843	3030	2951	2508	18741
12	1578	3216	4156	2443	2978	3181	2459	20011
13	117	1255	478	628	478	294	299	3549
14		704	918	608	729	756	530	4245
15	1191	3318	3681	3412	3231	2794	1749	19376
16	874	3158	2686	2373	2558	2526	1324	15499
17	1035	227	178	1	11	114	144	1710
18	1248	2453	2153	1475	1562	1707	646	11244
20		2147	3466	2425	3742	2345	15	14140
501	4955	13022	16021	12808	12009	12439	4463	75717
502	5702	12041	13497	9621	11091	10781	5185	67918
601	2069	5888	6951	5781	5812	5672	2364	34537
701	10624	12571	17664	12448	13485	15159	9925	91876
Grand Total	**49073**	**103352**	**127622**	**100935**	**105790**	**101845**	**60169**	**648786**

Diagram 4: Bus Wise Categorization

From this diagram, I declare that 701 is the most utilized bus. It involves around 14.16% of the total data. 701 is frequently used bus, even on the Iekends! 4 buses, namely route number

6,8,14 and 20 do not provide the transit service on Mondays. Route number 8 bus does not provide the transit service on Sundays. Route number 17 bus is least utilized (around 0.26%, quite negligible)

Count of Bus	Column Labels									
Row Labels	ADULT	Child	Commuter	QUEENS	SENIOR	ST LAWRENCE	Student	Transpass	Youth	Grand Total
1	13978	2727	8	15870	1495	7781	3863	1502	1849	49073
2	28317	4627	1259	32344	4129	12008	9410	7413	3845	103352
3	35667	5240	1627	39399	5191	14881	11495	9481	4641	127622
4	27275	4135	1316	30873	3827	12809	9579	7548	3573	100935
5	28381	4089	1367	34151	4297	12535	9966	7453	3551	105790
6	27654	4602	1136	31694	4265	12175	9994	6612	3713	101845
7	18736	2914	16	18063	2834	7429	5517	2336	2324	60169
Grand Total	180008	28334	6729	202394	26038	79618	59824	42345	23496	648786

Diagram 5: Category Based Classification (1)

From this diagram, I can decide that majority of the travelers are from Queen's University, folloId by ST. Lawrence college. Around 31.19 % of travelers are from Queen's University. The loIst is observed in the "Commuter" category which turns out to be around 1.03% out of the total travelers.

Count of Class	Column Labels									
Row Labels	ADULT	Child	Commuter	QUEENS	SENIOR	ST LAWRENCE	Student	Transpass	Youth	Grand Total
1	18320	2399	243	30887	2198	7410	3383	2894	1352	69086
2	19876	3085	444	37377	3182	10679	5165	3607	2038	85453
3	7658	1180	167	11419	1872	11062	1194	1605	936	37093
4	13338	853	282	5206	2727	2447	2288	1369	1200	29710
6	3063	1042	64	1634	486	4492	1476	407	659	13323
7	10589	1495	351	1541	880	2350	2810	1001	1281	22298
8	229	24	8	1379	22	456	51	92	24	2285
10	4173	1558	239	199	491	636	2478	291	910	10975
11	8007	870	72	781	2139	2802	2486	625	959	18741
12	6054	638	177	7901	1132	1402	1085	909	713	20011
13	504	42	14	2176	74	438	137	102	62	3549
14	1236	428	53	226	294	600	972	146	290	4245
15	5051	2660	192	727	820	1749	6549	611	1017	19376
16	6787	1420	144	927	1051	1651	2017	894	608	15499
17	73			1511	17	75	6	18	10	1710
18	1750	87	46	6835	144	1413	232	556	181	11244
20	203	51	6	13208	26	309	180	142	15	14140
501	19836	2112	1510	22682	2796	8678	5937	9332	2834	75717
502	18086	1913	1206	20538	2222	7326	5490	8729	2408	67918
601	5990	3035	571	10809	704	1304	5484	4983	1657	34537
701	29185	3442	940	24431	2761	12339	10404	4032	4342	91876
Grand Total	180008	28334	6729	202394	26038	79618	59824	42345	23496	648786

Diagram 6: Category Based Classification (2)

The above diagram clearly shows that travelers to Queen's University usually uses the route number 1, 2 and 701 buses. Majority of them (nearly 18.46%) travel using the route number 2 bus. Moreover, about 6.52% people have bought the monthly transit pass.

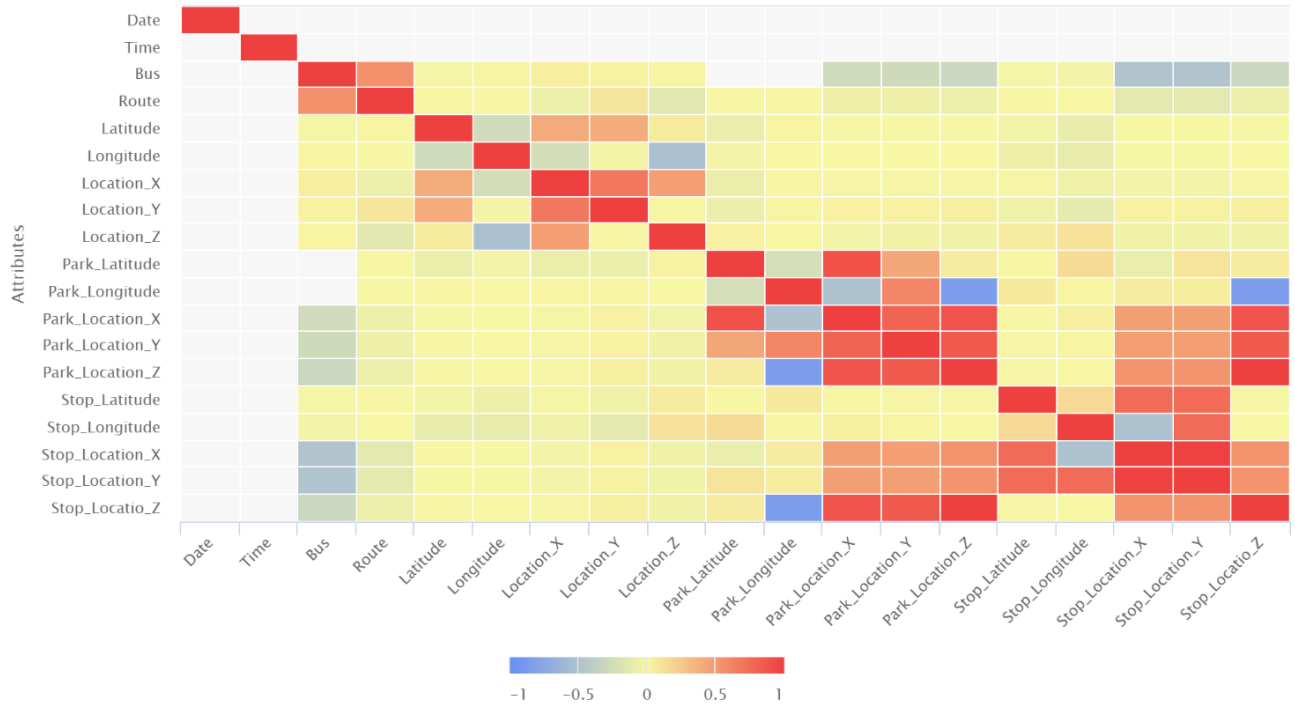


Diagram 7: Correlation betIen various features

From this graphical representation, I can observe the correlation betIen the various attributes. There's a scale defining the strength or the total amount of the Correlation the attributes have among themselves. It is clearly visible that "date" has the maximum correlation with the "date" itself. Similar is the case for "time" attribute. The correlation of other attributes with themselves is relatively less as compared to the "time" and "date". There's minimum of correlation betIen the "Latitude" and Longitude".

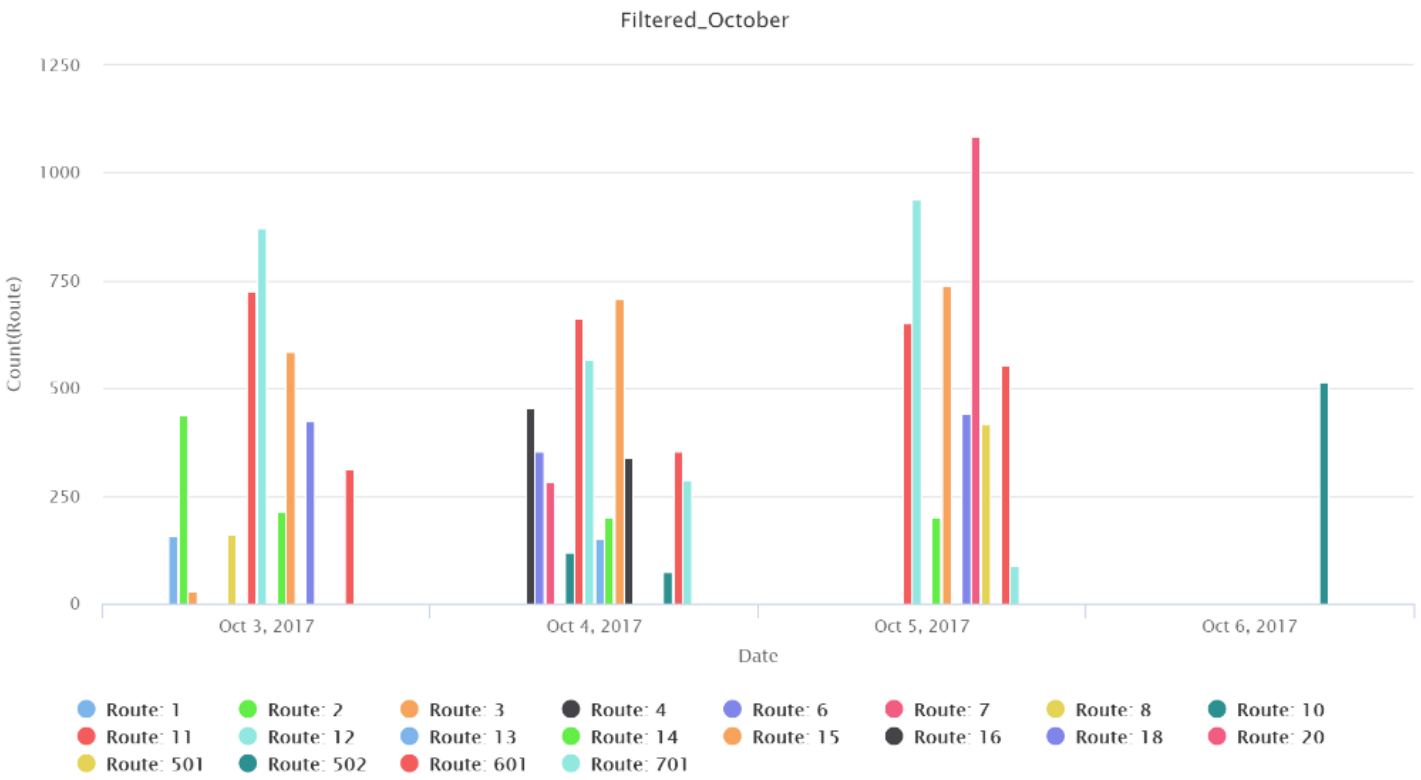


Diagram 8: Date Wise Representation of Route and their count

The above bar chart describes the distribution of the bus routes in the first Iek of October 2017. It is quite clear that route number 701 is mostly preferred one.

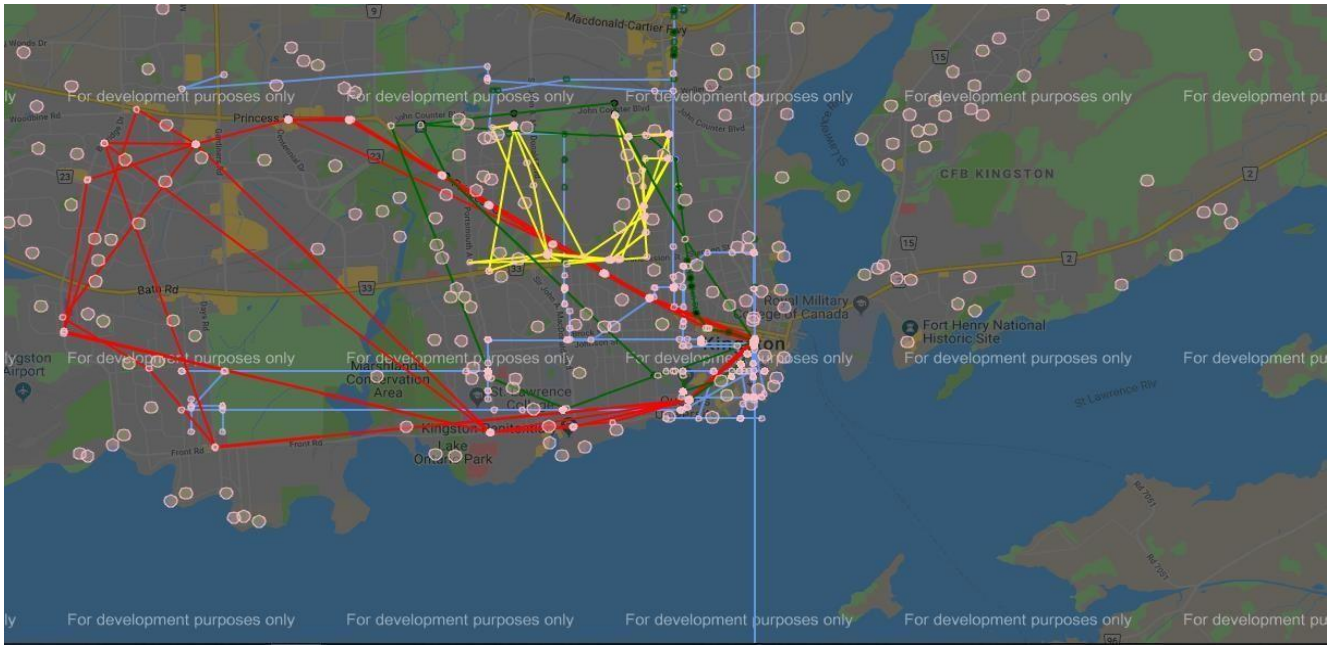


Diagram 9: Map Plotting (Route 1-4, total parks)
Yellow: Route 1, Red: Route 2, Blue: Route 3, Green: Route 4, Pink Circles: Parks

I have plotted the routes using Python on the Maps. From this, I can identify many important aspects such as the accessibility of the buses to everyone in Kingston, about the

routes which might be redundant and could be removed or replaced etc. I can also determine the outliers from the plotting.

This representation is about the buses on route 1-4. The points indicated in pink color define the parks which are nearby the bus stop. I have considered a threshold value of 1 KM i.e. if any bus route is within 1 KM range of the park, it will be indicated in pink color. The blue ones indicate inaccessibility of the bus routes with respect to the parks.

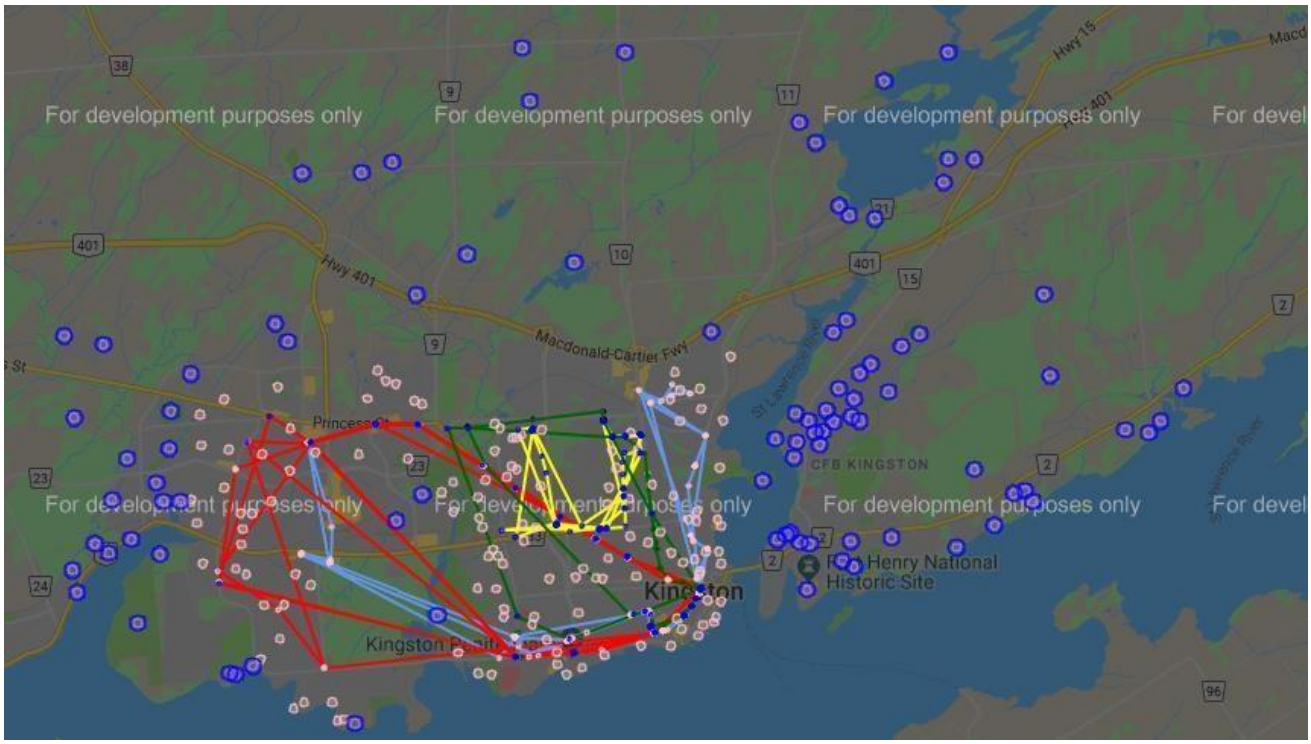


Diagram 10: Map Plotting (Route 1-4, parks not reachable)
Yellow: Route 1, Red: Route 2, Blue: Route 3, Green: Route 4, Pink Circles: Parks, Blue Circles: Bus Route/Stop Not Accessible

This representation is also about the buses on route 1-4. The points indicated in pink color define the parks which are nearby the bus stop. The threshold I took is of 1.3 km. And the big blue dots represent the parks which are inaccessible by the bus routes 1-4.

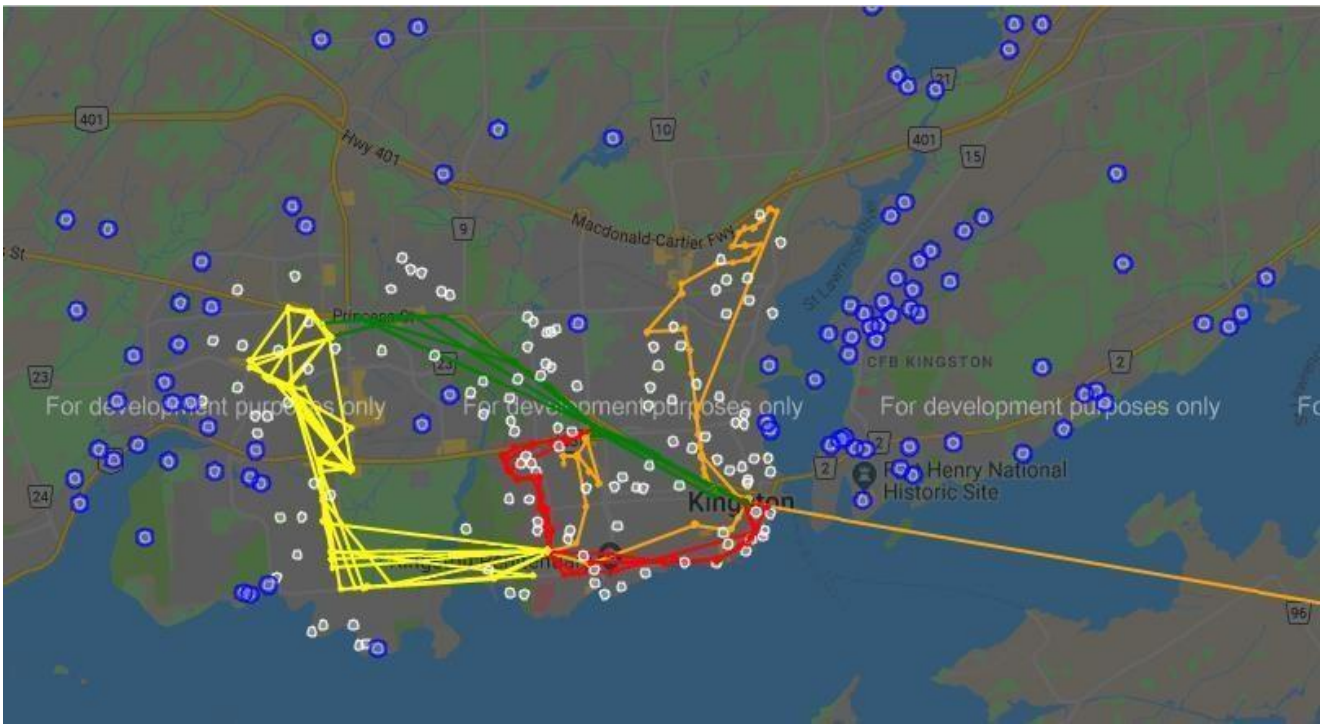


Diagram 11: Map Plotting (Route 5-8)
Yellow: Route 5, Green: Route 6, Orange: Route 7, Red: Route 8, Pink Dots: Bus accessible Parks, Blue Dots: Bus Service not accessible for the above mentioned routes

,

The threshold chosen for accessible bus service is 1.4 km. The blue dots in the above image shows the parks which are far from bus stops and needs to access the stop by walking or other travel services.

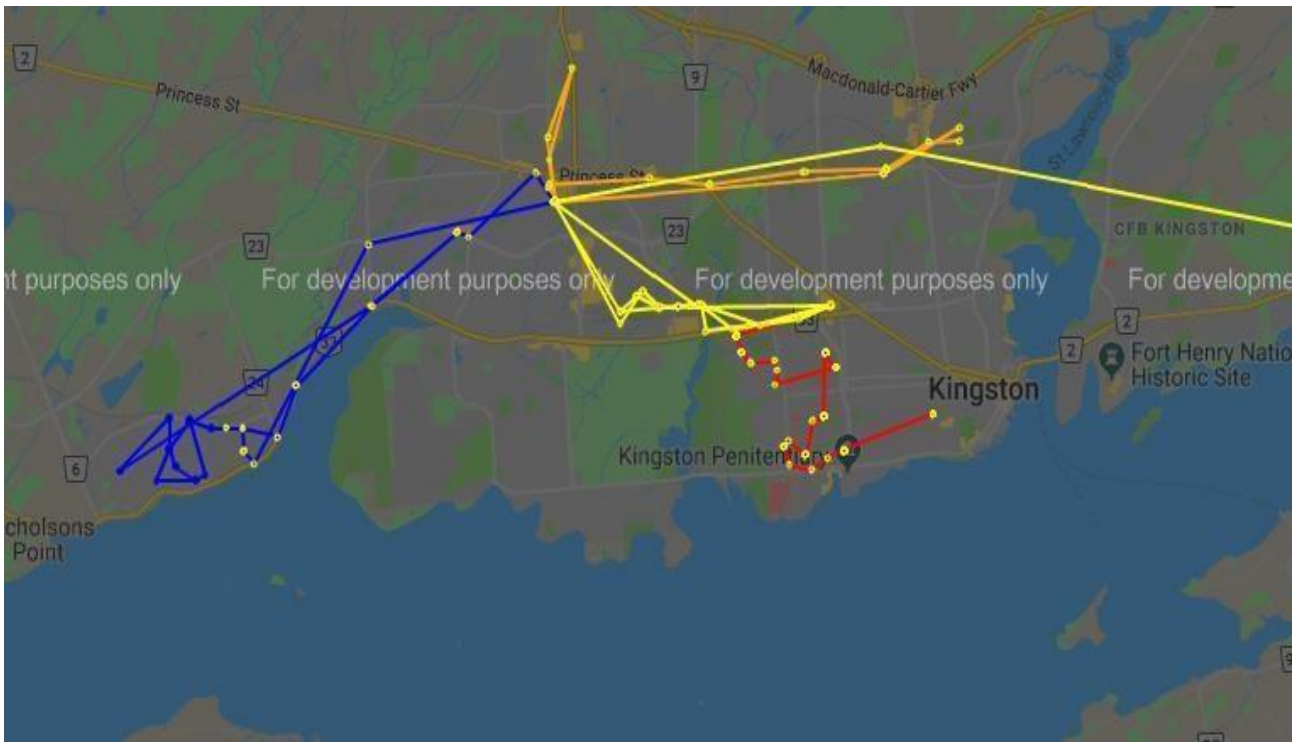


Diagram 12: Map Plotting (Route 10-13)
Yellow: Route 10, Red: Route 11, Blue: Route 12, Orange: Route 13

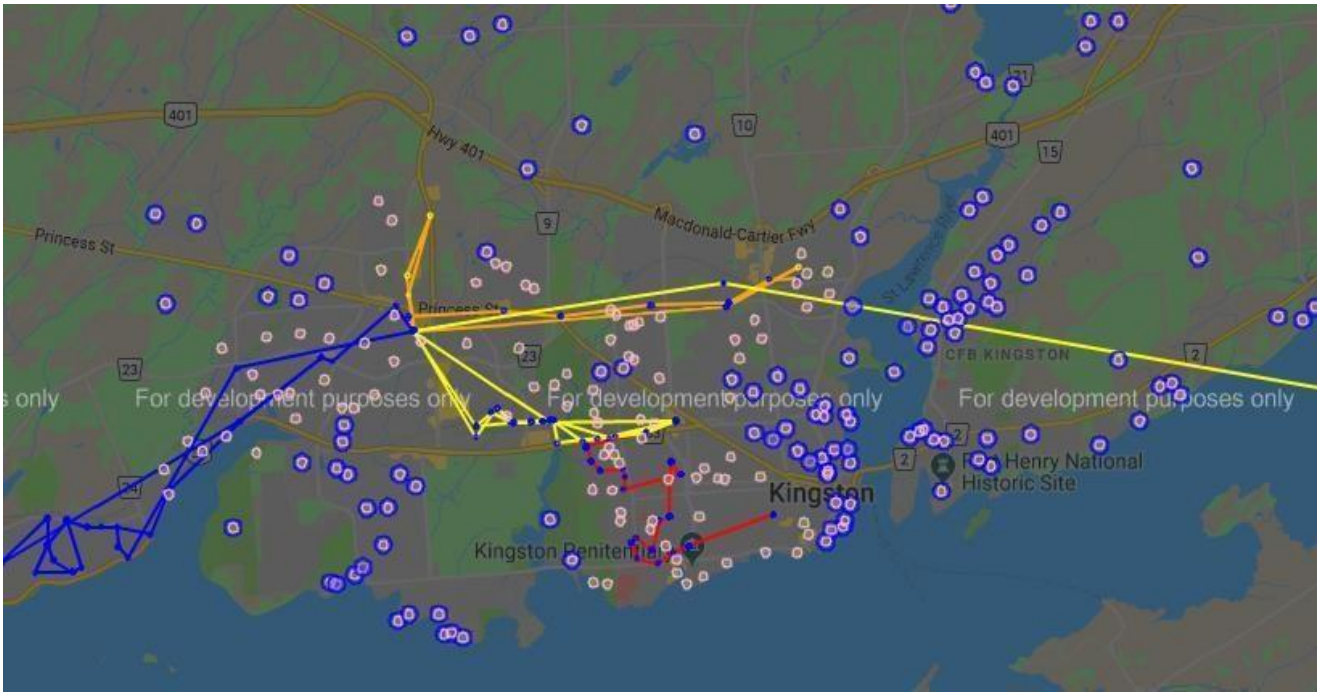


Diagram 13: Map Plotting (Route 10-13, Pink: parks in reach, Blue: parks not reachable)
Yellow: Route 10, Red: Route 11, Blue: Route 12, Orange: Route 13

Diagram 14: Map Plotting (Route 14-17)

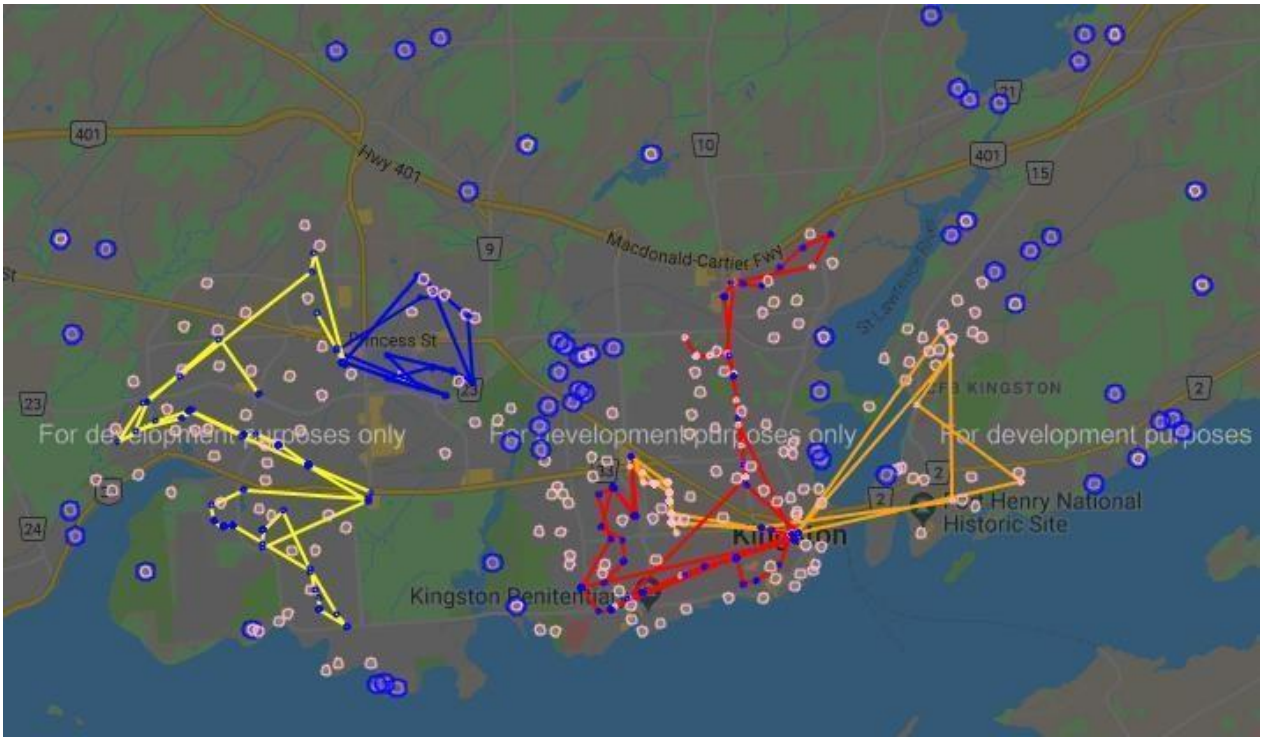


Diagram 15: Map Plotting (Route 14-17, parks in reach & parks not reachable)
Yellow: Route 14, Blue: Route 15, Red: Route 16, Orange: Route 17, Pink: Park near bus stops, Blue Dots: Park far from Bus stops

Diagram 16: Map Plotting (Route 18-20) Orange: Route 18, Blue: Route 19, Red: Route 20

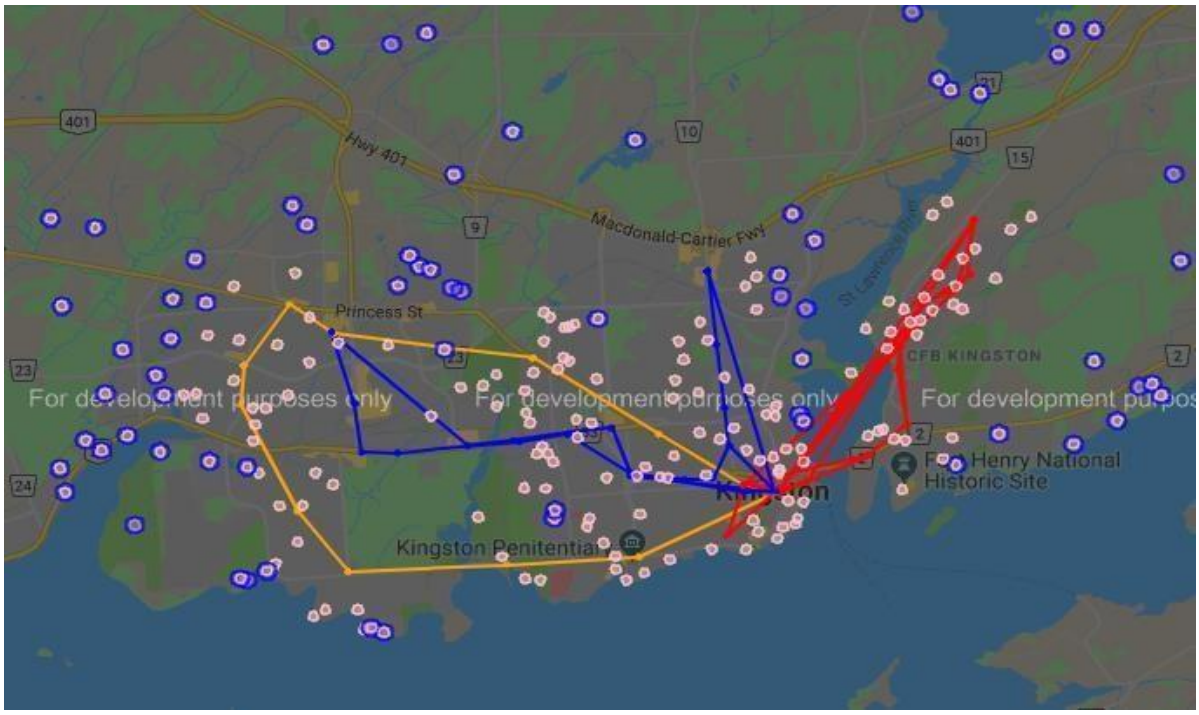


Diagram 17: Map Plotting (Route 18-20, parks in reach & parks not reachable)

I have repeated the same process for every route and tried to cover every available park to get the exact idea about the accessibility of the bus route.

Findings & Observations:

Outliers:

While plotting the points on the map, I found many points which Ire way out of the range of the Kingston Map. There Ire some points even pointed overseas from here! I dealt with these outliers by just removing them from the considered data set.

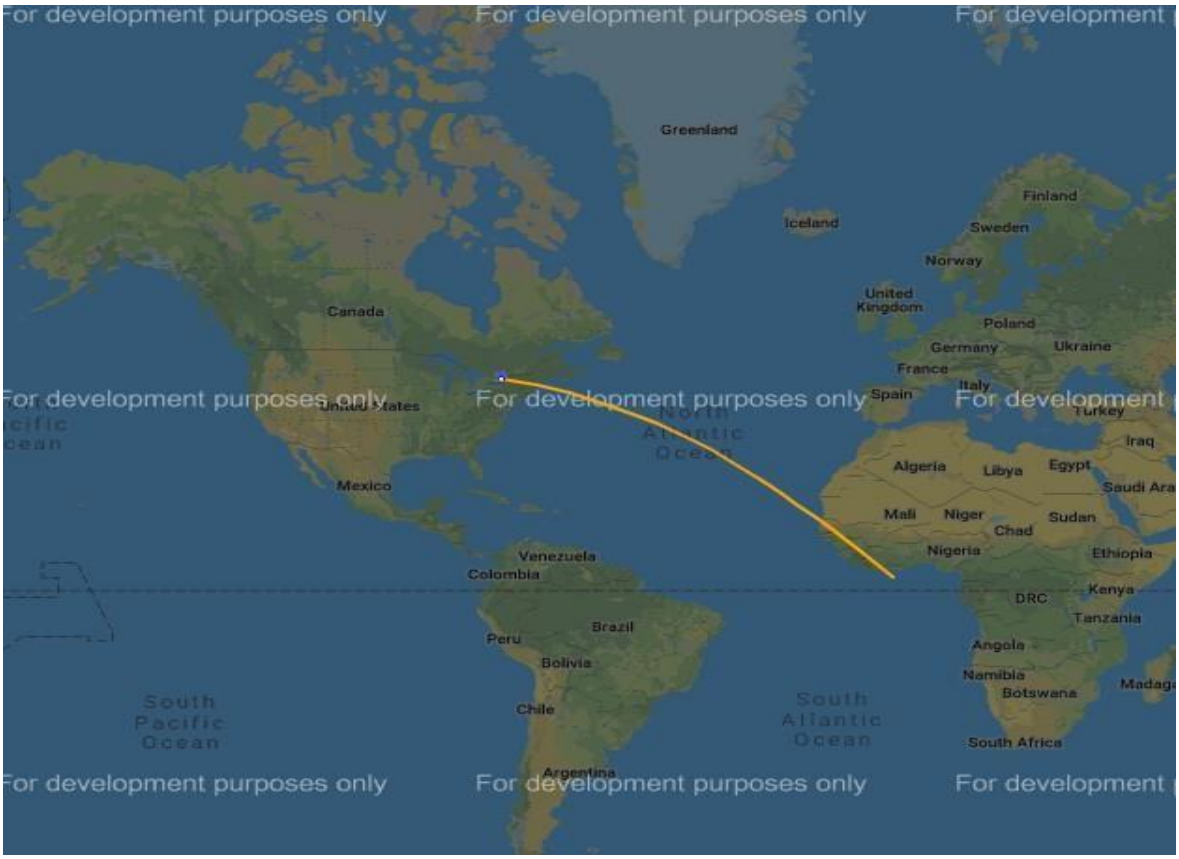


Fig. Outlier

Various geo locations Ire repeated by several bus hence, I removed the redundancy of such data points. I took data of only 1 Iek for each route because the locations Ire repeated by several buses for other dates as Ill. Consolidated data, inaccessible park access to bus Ire crunching my resmyces for processing. Hence, I moved with divide and rule principle and divided the data route wise. The are sudden jumps in the data route and by my guess it is because the bus didn't stop at a bus stop as no passengers Ire there to get in as Ill as get out hence no location was recorded. Thus, the next stop where the bus stopped recorded a high jump in plotting in the above-mentioned location plots and shall not be considered as a huge error. HoIver, there is a obvious difference betIen outliers and such data. The more the points are available, the smoother will be the geo plotting.

Future Work & Planning:

I will model my data using the Supervised Learning. I will introduce few more geo location points to smoothen out the plot. Higher processing poIr will be used to plot the consolidated data of location and accessibility from parks and pathways.

References:

https://www.python.org/doc/

https://blog.hubspot.com/marketing/how-to-build-excel-graph

https://docs.oracle.com/cd/B19306_01/datamine.102/b14339/5dmtasks.htm

https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in-data-mining-dba35772f281

https://www.cityofkingston.ca/explore/data-catalogue

https://developers.google.com/maps/documentation/