

Midterm report

This semester, we have learnt how to use QGIS as the visualization tool and PostgreSQL as database. These tools are powerful in illustrating the geographic information from a visionary perspective. With the practice during the course, I managed to learn how to generate database from various methods; how to export SQL files; how to generate graph in QGIS; and most importantly, how to adjust the graph elements to optimize the reception experience of viewers.

My report is a very simple one. Although I am not able to produce stunning graphs with my humble aesthetics ability, I managed to produce a decent graph with figures on the stations along the train lines.

Given limited data, my initial attempt was to calculate the ratio of empty seats pf each line. First of all, let me explain how the ratio of empty seats bothers me. Let us assume we have a line between a and b. And $a \rightarrow b$ is busy while $b \leftarrow a$ is empty. And if a train goes from a to b, since the seats are all occupied, we can say the energy is well spent because the train fulfills its job as a commuting tool as well as earns profit. Yet on its way back, this train is so empty that there is no economic profit in this run and thus the budget is wasted. I believe the occupation of seats is very important to measure if a line is energy efficient and therefore environmental friendly.

However, there is no way to measure the seat occupation directly from the data. First of all, we only get the information of departure and arrival in each station. Beside, some of the stations are shared by different lines. Therefore, the exact number of seat occupation is impossible to get regarding the missing data. So, I measured the difference between arrival and departure of each station instead.



The picture above is the part of my map1. In which, I found the station with highest influx(arrival-difference) is 池袋.

name	prefecture	city	departure	arrival	difference	total	id	cluster_id
池袋	東京都	豊島区	188757	195534	6777	384291	8132	468
大手町	東京都	千代田区	66508	71358	4850	137866	1455	97
大手町	東京都	千代田区	66508	71358	4850	137866	4055	246
表参道	東京都	港区	40084	44687	4603	84771	8263	468
新宿三丁目	東京都	新宿区	36539	40656	4117	77195	8196	468
銀座	東京都	中央区	49054	51834	2780	100888	4018	246
錦糸町	東京都	墨田区	43967	46383	2416	90350	4049	246
大宮	埼玉県	さいたま市大宮区	61957	64315	2358	126272	6606	382
大宮	埼玉県	さいたま市大宮区	61957	64315	2358	126272	5701	329
新大久保	東京都	新宿区	20165	22517	2352	42682	8205	468
六本木	東京都	港区	43142	45426	2284	88568	3919	246
赤羽	東京都	北区	30580	32832	2252	63412	8264	468
吉祥寺	東京都	武蔵野市	50628	52772	2144	103400	3640	234
飯田橋	東京都	千代田区	53322	55465	2143	108787	3991	246
恵比寿	東京都	渋谷区	63737	65820	2083	129557	8655	489
西新宿	東京都	新宿区	18328	20349	2021	38677	8214	468
三越前	東京都	中央区	27593	29609	2016	57202	4086	246
新横浜	神奈川県	横浜市港北区	35667	37668	2001	73335	1336	92
原宿	東京都	渋谷区	25255	27158	1903	52413	8159	468
北千住	東京都	足立区	51438	53328	1890	104766	4110	246
船橋	千葉県	船橋市	30190	31975	1785	62165	6866	393
蒲田	東京都	大田区	39867	41645	1778	81512	8648	489
下北沢	東京都	世田谷区	24358	25937	1579	50295	8268	468
藤沢	神奈川県	藤沢市	29779	31332	1553	61111	8332	470
御徒町	東京都	台東区	19479	20976	1497	40455	3972	246
馬喰横山	東京都	中央区	7824	9320	1496	17144	4002	246
横浜	神奈川県	横浜市西区	153757	155206	1449	308963	1267	92
日比谷	東京都	千代田区	27705	29100	1395	56805	4023	246
川崎	神奈川県	川崎市川崎区	59741	61092	1351	120833	8763	489

While the station with highest outflux is 渋谷.

name	prefecture	city	departure	arrival	difference	total	id	cluster_id
渋谷	東京都	渋谷区	208836	204199	-4637	413035	8127	468
西武新宿	東京都	新宿区	18996	15160	-3836	34156	8203	468
秋葉原	東京都	千代田区	81136	77720	-3416	158856	3962	246
新橋	東京都	港区	91888	89285	-2603	181173	3946	246
神田	東京都	千代田区	40261	37915	-2346	78176	59	3
神田	東京都	千代田区	40261	37915	-2346	78176	3933	246
東京	東京都	千代田区	130868	128559	-2309	259427	4034	246
赤坂見附	東京都	港区	25211	23444	-1767	48655	3956	246
品川	東京都	港区	83314	81715	-1599	165029	8679	489
有楽町	東京都	千代田区	54125	52670	-1455	106795	4094	246
後楽園	東京都	文京区	16121	14679	-1442	30800	3947	246
九段下	東京都	千代田区	22342	20901	-1441	43243	3963	246
浜松町	東京都	港区	30042	28615	-1427	58657	4117	246
京成上野	東京都	台東区	5313	3937	-1376	9250	4039	246
水道橋	東京都	千代田区	25251	24016	-1235	49267	4076	246
日本橋	東京都	中央区	34656	33434	-1222	68090	7004	402
日本橋	東京都	中央区	34656	33434	-1222	68090	4021	246
田町	東京都	港区	41913	40703	-1210	82616	3890	243
田町	東京都	港区	41913	40703	-1210	82616	3924	246
新宿	東京都	新宿区	244909	243718	-1191	488627	8201	468
駒沢大学	東京都	世田谷区	14259	13095	-1164	27354	8738	489
宮前平	神奈川県	川崎市宮前区	8243	7231	-1012	15474	1278	92
要町	東京都	豊島区	8098	7108	-990	15206	8171	468
三鷹	東京都	三鷹市	28128	27140	-988	55268	3613	234
月島	東京都	中央区	14183	13200	-983	27383	4087	246
池尻大橋	東京都	世田谷区	12515	11543	-972	24058	8754	489
神保町	東京都	千代田区	27668	26709	-959	54377	3965	246
早稲田	東京都	新宿区	12505	11569	-936	24074	8137	468
不動前	東京都	品川区	7461	6572	-889	14033	8731	489

However, I think it is not reasonable to infer any results from this data. First of all, these two stations has largest passenger flow among all the stations. And the larger the flow is, the more likely to have higher difference. Therefore, I managed to sort the stations by the ratio of influx/outflux to the total number of passengers. And map2 is the result. However, this data is not very convincing since I found some data is abnormal.

name	prefecture	city	departure	arrival	diffrence	total	ratio
白丸	東京都	奥多摩町	16	0	-16	16	1
竹沢	埼玉県	小川町	0	11	11	11	1
笹子	山梨県	大月市	9	0	-9	9	1
大宝	茨城県	下妻市	0	12	12	12	1
東浪見	千葉県	一宮町	17	6	-11	23	0.4782608695652173913
東成田	千葉県	成田市	131	367	236	498	0.47389558232931726908
初狩	山梨県	大月市	19	38	19	57	0.33333333333333333333
合戦場	栃木県	栃木市	22	11	-11	33	0.33333333333333333333
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
武蔵横手	埼玉県	日高市	38	63	25	101	0.24752475247524752475
川崎新町	神奈川県	川崎市川崎区	363	230	-133	593	0.22428330522765598651
相模金子	神奈川県	大井町	62	40	-22	102	0.21568627450980392157
塔ノ沢	神奈川県	箱根町	15	23	8	38	0.21052631578947368421
柴又	東京都	葛飾区	1502	985	-517	2487	0.20788098110172899075
府中競馬正門前	東京都	府中市	47	31	-16	78	0.20512820512820512821
鳥沢	山梨県	大月市	121	81	-40	202	0.1980198019801980198
柳生	埼玉県	加須市	83	56	-27	139	0.19424460431654676259
松田	神奈川県	松田町	51	75	24	126	0.19047619047619047619
水郷	千葉県	香取市	19	13	-6	32	0.1875
葛生	栃木県	佐野市	62	43	-19	105	0.18095238095238095238
柳小路	神奈川県	藤沢市	807	564	-243	1371	0.17724288840262582057
多々良	群馬県	館林市	28	20	-8	48	0.16666666666666666667
田島	栃木県	佐野市	21	15	-6	36	0.16666666666666666667
大平下	栃木県	栃木市	63	46	-17	109	0.15596330275229357798
大田郷	茨城県	筑西市	56	41	-15	97	0.15463917525773195876

I got some stations with result ratio of 1 or 0. I was a little bit confused about the result. And I assume the result is caused by a small amount of data from abnormal stations. So, I normalized the data by lift the bar by 20. And I got the third data.

name	prefecture	city	departure	arrival	diffrence	total	ratio
東成田	千葉県	成田市	131	367	236	498	0.47389558232931726908
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
栄町	千葉県	千葉市中央区	85	142	57	227	0.2511013215859030837
武蔵横手	埼玉県	日高市	38	63	25	101	0.24752475247524752475
松田	神奈川県	松田町	51	75	24	126	0.19047619047619047619
新高島	神奈川県	横浜西区	2046	2741	695	4787	0.1451848750503446835
葭川公園	千葉県	千葉市中央区	319	413	94	732	0.12841530054644808743
神泉	東京都	渋谷区	3218	4108	890	7326	0.12148512148512148512
扇町	神奈川県	川崎市川崎区	76	96	20	172	0.11627906976744186047
扇町	神奈川県	川崎市川崎区	76	96	20	172	0.11627906976744186047
石下	茨城県	常総市	89	111	22	200	0.11
新整備場	東京都	大田区	1239	1544	305	2783	0.10959396334890406037
東武竹沢	埼玉県	小川町	97	120	23	217	0.10599078341013824885
武蔵白石	神奈川県	川崎市川崎区	339	405	66	744	0.08870967741935483871
新芝浦	神奈川県	横浜市鶴見区	175	209	34	384	0.08854166666666666667
馬喰横山	東京都	中央区	7824	9320	1496	17144	0.08726084927671488567
昭和	神奈川県	川崎市川崎区	137	162	25	299	0.08361204013377926421
新千葉	千葉県	千葉市中央区	421	497	76	918	0.08278867102396514161
青堀	千葉県	富津市	174	205	31	379	0.08179419525065963061
小絹	茨城県	つくばみらい市	202	237	35	439	0.07972665148063781321
南新宿	東京都	渋谷区	1140	1335	195	2475	0.07878787878787878788
二重橋前	東京都	千代田区	6500	7595	1095	14095	0.07768712309329549486
京成高砂	東京都	葛飾区	6366	7410	1044	13776	0.07578397212543554007
石園	茨城県	石岡市	510	592	82	1102	0.07441016333938294011
松が谷	東京都	八王子市	530	612	82	1142	0.07180385288966725044
大崎広小路	東京都	品川区	2263	2609	346	4872	0.0710180623973727422

Now the data start to make sense. And it is also shown on map3.

In conclusion, although the desired ratio of empty seats is not acquired. But, I think the difference of influx/outflux can be used as a side-factor indicating the tend of utilization of space of each train. So far, I think in order to infer the empty seats ratio, we might need more data on the extent of crowdedness or more specific data on passenger counts of each line at various points of the time. Also, there should require more advanced statistic tools to make a reasonable prediction.