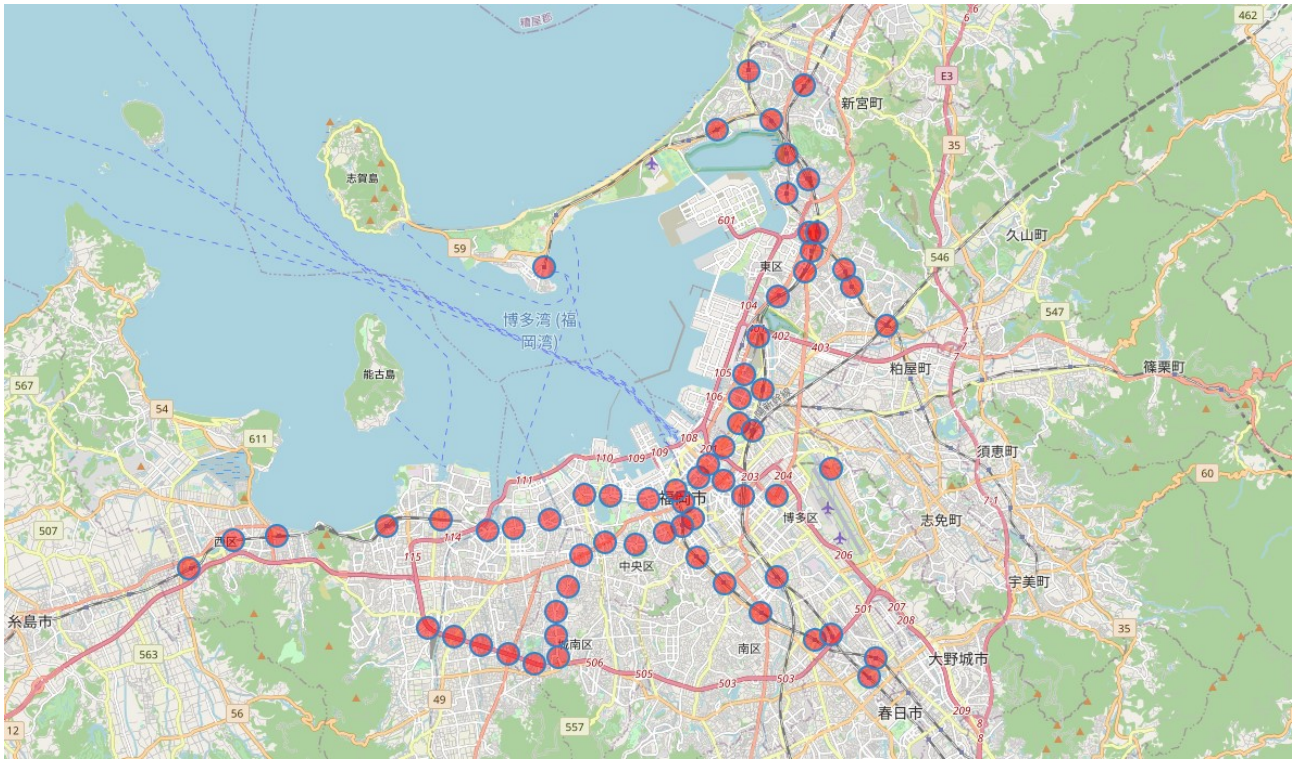
IBM Data Science - Capstone Project

The Battle of Neighbourhoods

Find a good place for a new coffee shop in Fukuoka City using Data Science methodologies



Abstract

This project is the capstone assignment for the IBM Data Science Professional Certification on Coursera. [1] The goal of the assignment was to come up with an idea to leverage the Foursquare API location data to explore or compare neighborhoods or a city of choice and develop a report outlining the business problem, dataset, data analysis methodology, results, observations, and conclusions. As the city of choice I picked the city of Fukuoka, that is the city of my current residence. The selected business problem will be about finding a nice location for a new coffee shop in the city. This analysis is not requested by any particular business customer.

Author: E. Malamura
Publication date: July 21, 2021

Table of Content

1. Introduction

Fukuoka is Japan's fifth largest city and the capital of Fukuoka prefecture situated on the northern side of Kyushu island. According to the preliminary data of the 2020 national census, the city experienced a five-year population growth of 4.9%, the fastest rate among the ordinance-designated cities in Japan. Fukuoka's population is also relatively youthful, with the percentage of 20 to 30 year olds considerably higher than the Japanese average. [2]

Along with the population growth, in recent years there is a trend of growing popularity of cafes and coffee culture. There are many new venues opened in recent years, as in busy districts near big city transport hub stations, as well as more coffee venues are being opened in quieter residential neighbourhoods too.

The goal of this analysis is to identify a good location for opening a new coffee shop in the city, such that: a) the area has a decent foot traffic, b) the area is popular among people, who are most likely to be active customers of a coffee shop. In this analysis we will focus mainly on the young population of Fukuoka City, trying to detect locations of their interest and gathering, c) the location has few coffee shops opened yet, that would give some competitive advantage to a new venue.

The results of this analysis will be useful for entrepreneurs, who are planning to open a coffee shop venue in the city and thus looking for a promising location.

2. Data

2.1 Data description

Instead of analysis of districts and neighbourhoods of the city by their names and official borders, we will use train stations as centers of neighbourhoods. That sounds reasonable for Japanese cities, due to high levels of development and popularity of public transport. A subway or a city railroad station is always a hub of foot traffic and a center of a neighbourhood.

In Fukuoka, we will analyse stations of 3 main city railways: Fukuoka City Subway (35 stations), Japan Railways (18 stations within the city) and Nishi-Nippon Railroad (16 stations within the city). The list of names of all stations (in Japanese) of 3 operators are found and scraped from a website of local 'yellow pages'. [3]

Every train station has a page in wikipedia that includes information about their locations, including geographical coordinates. Links to those pages were assembled using the names of stations in Japanese, then the geographical coordinates of each station were scraped from corresponding wikipedia pages. [4]

The load (foot traffic) of every station is obtained from another source of regional statistical data in CSV format, and then merged with the scraped data by train station names. [5]

At last, but not least, we used the Foursquare API to gather information about venues in the vicinity (500m radius) of each train station [6]. We were interested in 3 large categories of venues, which are:

- Coffee shops;
- Colleges and universities;
- Arts and entertainment venues

The first category will give us information about possible competitors in the vicinity of a station. The other two categories may give us insights about places of gathering of young audience, that is the primary target group of customers, as it is stated in Introduction.

2.2 Data acquisition and preparation

The code for the data scraping and cleaning process can be obtained from the github repository. [0]

After cleaning data from three different sources, we compiled them into one dataframe of venues with their geographical coordinates and categories, related neighbourhood stations and coordinates of stations. The resulting data frame can be obtained from github in csv table format. Summary of data is in Table 1.

Table 1. Summary of the main dataset of venues.

#	Column	Count Values	Type	Sample
0	Station	1404	object	Tenjin
1	Station Latitude	1404	float64	33.591489
2	Station Longitude	1404	float64	130.399358
3	Daily Passengers	1404	int64	148957
4	Venue	1404	object	Tenjin Style
5	Venue Latitude	1404	float64	33.593734
6	Venue Longitude	1404	float64	130.394685
7	Venue Category	1404	object	Cafes
8	Venue Subcategory	1404	object	Coffee Shop

3. Methodology

In this next section we will start with an exploratory analysis intending to understand underlying features of the data. We will proceed with an analysis of locations from the point of view of each of the chosen criteria (foot traffic, popularity among the target customers and density of existing coffee shop venues in the area). Then finally we will apply a machine learning technique (clustering) to detect the most promising groups of locations for opening a new coffee shop business.

3.1 Exploratory analysis

We started with an exploratory analysis of the collected data. We have collected data of geographical locations of 1404 venues of 3 major categories (coffee shops, entertainment venues and educational institutions), which are situated in the vicinity of 62 train stations. Note that the default Foursquare API does not allow us to extract more than 100 venues by one request, thereby data about a number of coffee shops near some busiest stations is not precisely correct. However, we assume that it does not affect the results of analysis and conclusions.

Table 2. Selected types of venues in the vicinity of train station (* Note: 100 is not a true maximum number of venues, but the upper limit of results returned by Foursquare API [6])

#	Venue category	Color on map	Max	Min	Mean	Total
0	Coffee shops	blue	100 *	0	13	809
1	Arts & Entertainment	yellow	56	0	6	368
2	College & University	red	28	0	4	227

Another important part of the collected data is the information about average daily traffic of passengers. In Fukuoka city, the average daily passengers flow at a train station is 28724 people. Yet, there is a huge gap between the stations with minimum and maximum passenger flow.

The busiest station of Fukuoka city is Hakata with a daily traffic of 461079 people. Hakata is the large hub of Japan Railways trains between cities, Fukuoka's bullet trains (shinkansen) and a large subway station. The daily flow of passengers at Hakata is 3 times more intense than the flow of the second busiest station - Tenjin (148957 people daily).

The smallest city train station is Tonoharu station of the Nishi-Nippon Railroad line with just 1029 daily passengers.

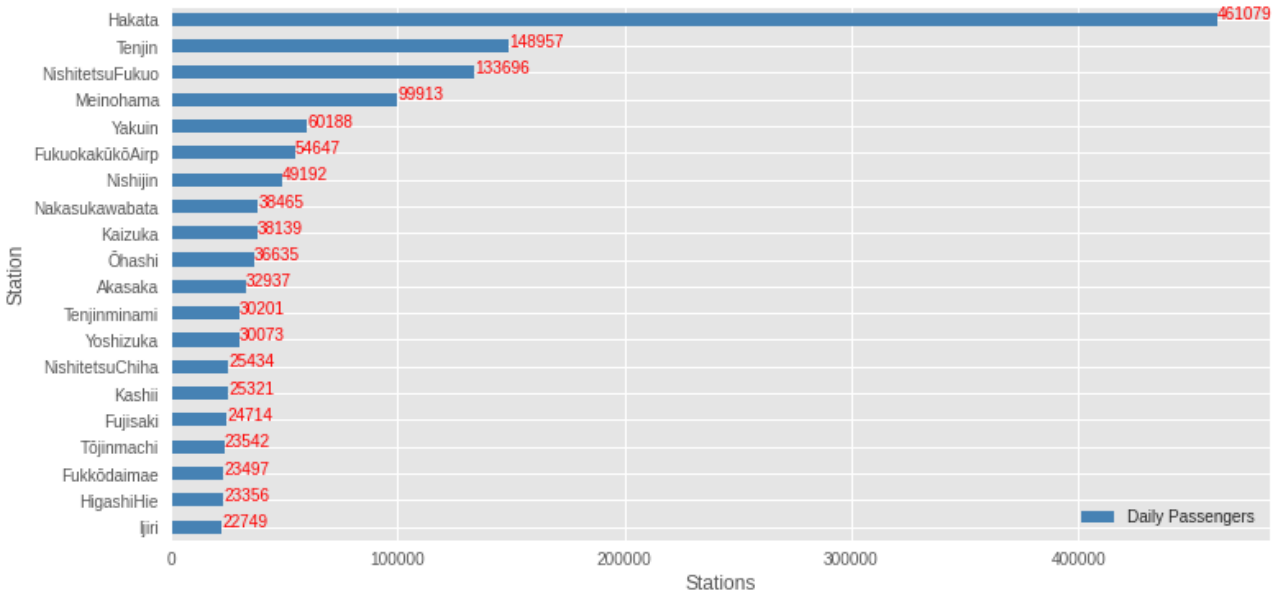


Figure 1. Top 20 stations in Fukuoka city by daily passenger traffic.

Finally, our dataset includes coordinates of stations and nearby venues. Their distribution across the map of the city is visualized on the map using Folium in Figure 0.

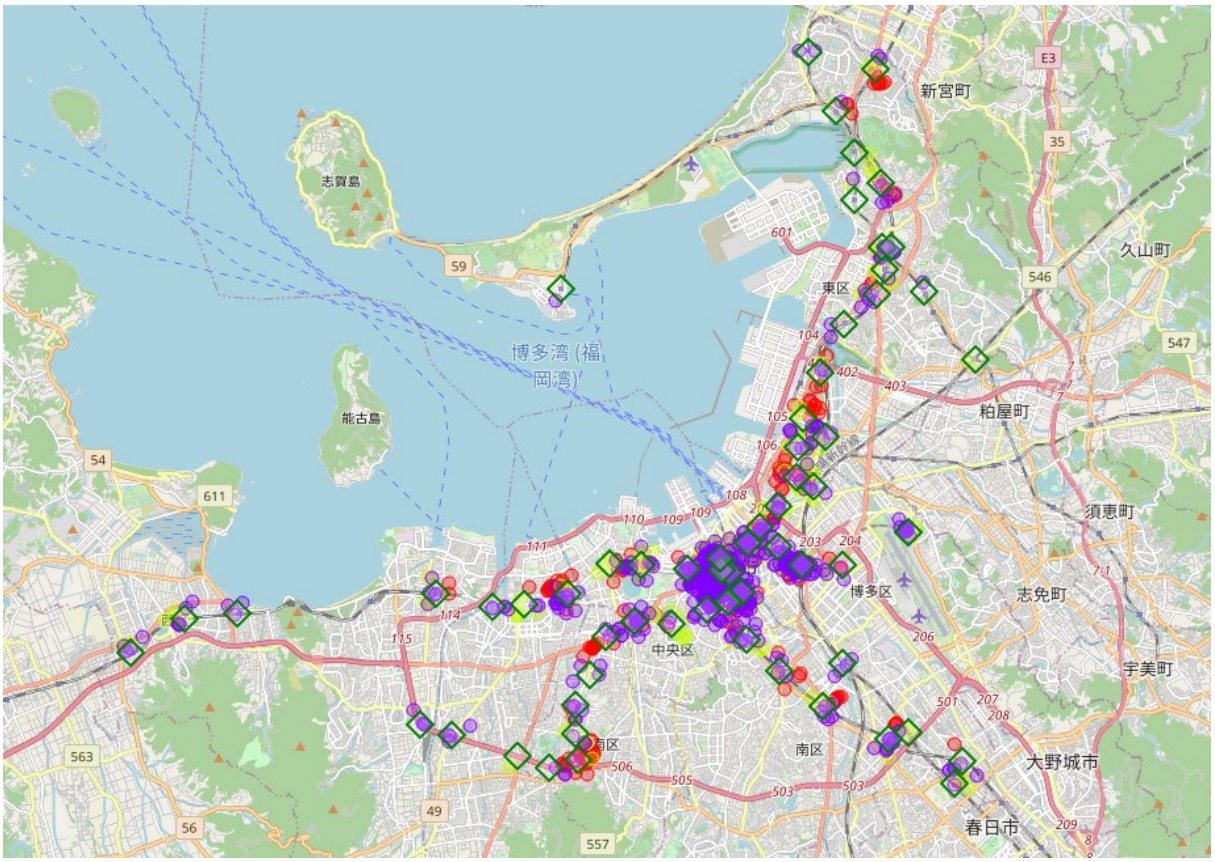


Figure 2. Distribution of coffee shops (blue), entertainment venues (yellow) and educational institutions (red) located in the vicinity of train stations (green square marks) in Fukuoka City.

3.2 Clustering

In this section we will use scikit-learn's k-means algorithm to divide all Fukuoka's neighbourhoods into clusters by their similarity from the point of view of the number of different types of venues around the stations and daily flow of passengers. To do that, first we did one-hot encoding for the Venue Category column and calculated a sum of venue types per station name. The resulting table of stats for 62 stations is summarized in Table 3.

Table 3. Data grouped by station names after one-hot encoding of venue categories.

	Station	Cafes	Education	Entertainment	Daily Passengers
0	Akasaka	54	4	16	32937
1	Befu	5	12	1	11453
2	Chayama	4	1	1	4986
3	ChiyokenchōguchiChiyoPrefecturalOffice	6	3	3	9423
4	Doi	0	0	1	2963
...	...	...	...	...	...
57	Yakuin	41	5	4	60188
58	YakuinodoriZooBotanicalGarden	28	2	2	6258
59	Yoshizuka	4	3	5	30073
60	Zasshonokuma	4	2	3	15047
61	Ōhashi	6	10	4	36635

62 rows × 5 columns

After that, we used the MinMaxScaler algorithm to normalize the numerical data before clustering.

The number of clusters k for the k-means algorithm was decided by running k-means multiple times with different k, and measuring k-means performance with the Silhouette Score Elbow tool as it is illustrated in Figure 0. In result we decided that the optimal number of clusters for our analysis is 5.

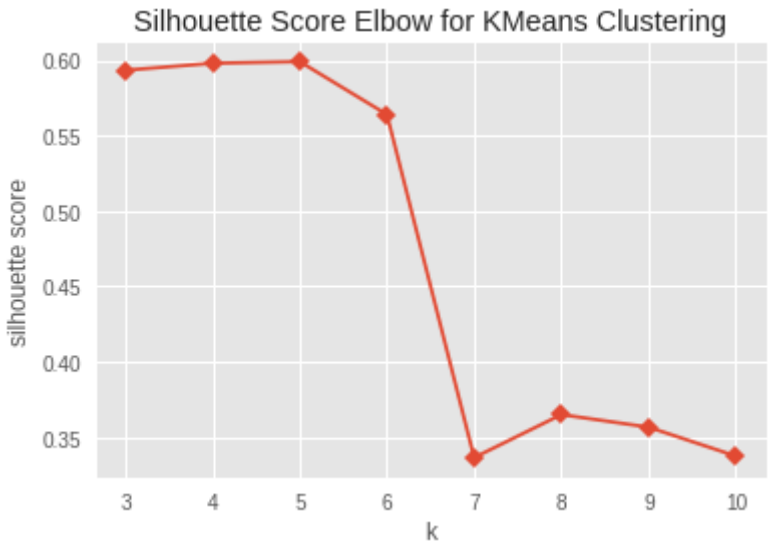


Figure 3. Selection an optimal number of clusters for k-means algorithm with silhouette score elbow (optimal n_clusters = 5).

The following Tables 0-0 show samples from the

Table 4. Cluster of 46 stations in residential neighbourhoods

	Station	Station Latitude	Station Longitude	Cafes	Education	Entertainment	Daily Passengers	Cluster
23	Meinohama	33.583703	130.325236	5	2	4	99913	0
48	FukuokakūkōAirport	33.597078	130.448542	13	0	5	54647	0
42	Kaizuka	33.632175	130.425578	2	3	1	38139	0
29	Yoshizuka	33.607161	130.423906	4	3	5	30073	0
31	NishitetsuChihayaChihaya	33.649258	130.440164	6	1	3	25434	0
...	...	...	...	...	...	...	...	...
50	Doi	33.634994	130.466097	0	0	1	2963	0
27	KashiiMiyamae	33.654658	130.442525	2	2	4	2418	0
58	Kashiijingū	33.649753	130.452631	1	0	0	2033	0
60	Saitozaki	33.650236	130.357825	1	0	0	1808	0
53	Tōnoharu	33.680050	130.434506	0	0	1	1092	0

46 rows × 8 columns

Table 5. Hakata station cluster

	Station	Station Latitude	Station Longitude	Cafes	Education	Entertainment	Daily Passengers	Cluster
7	Hakata	33.59	130.420611	71	9	17	461079	1

Table 0. Cluster of 3 central stations of Tenjin area

	Station	Station Latitude	Station Longitude	Cafes	Education	Entertainment	Daily Passengers	Cluster
3	Tenjin	33.591489	130.399358	100	5	56	148957	2
2	NishitetsuFukuokaTenjin	33.588778	130.399889	100	6	47	133696	2
1	Tenjinminami	33.588197	130.401728	100	5	39	30201	2

Table 6. Cluster of 5 stations near the central Tenjin area

	Station	Station Latitude	Station Longitude	Cafes	Education	Entertainment	Daily Passengers	Cluster
8	Yakuin	33.581925	130.401683	41	5	4	60188	3
0	Nakasukawabata	33.594983	130.406603	28	2	25	38465	3
4	Akasaka	33.589117	130.390850	54	4	16	32937	3
5	Watanabedori	33.583969	130.404844	35	5	7	7966	3
10	YakuinodoriZooBotanicalGarden	33.580406	130.396158	28	2	2	6258	3

Table 7. Cluster of 7 stations with more educational institutions in vicinity.

	Station	Station Latitude	Station Longitude	Cafes	Education	Entertainment	Daily Passengers	Cluster
17	Nishijin	33.583803	130.359742	15	22	8	49192	4
20	Ōhashi	33.559203	130.426344	6	10	4	36635	4
21	Fukkōdaimae	33.698442	130.440117	2	9	3	23497	4
36	Fukudaimae	33.547469	130.362228	3	28	2	14098	4
11	MaidashikyūdaibyōinmaeMaidashiKyushuUniversity...	33.609214	130.419317	4	12	5	12551	4
18	Befu	33.574169	130.369628	5	12	1	11453	4
34	HakozakikyūdaimaeHakozakiKyushuUniversity	33.622111	130.421039	4	10	6	6546	4

4. Results

Table 8. Summary of clusters of the neighbourhoods

Cluster Id	Color on map	N Stations	Mean Cafe	Mean Edu.	Mean Enter.	Mean Pass.
0	Blue	46	4.63	1.76	2.74	15372.78
1	Sky blue	1	74.00	9.00	17.00	461079.00
2	Green	3	100.00 *	5.33	47.33	104284.67
3	Orange	5	37.20	3.60	10.80	29162.80
4	Red	7	5.57	14.71	4.14	21966.00

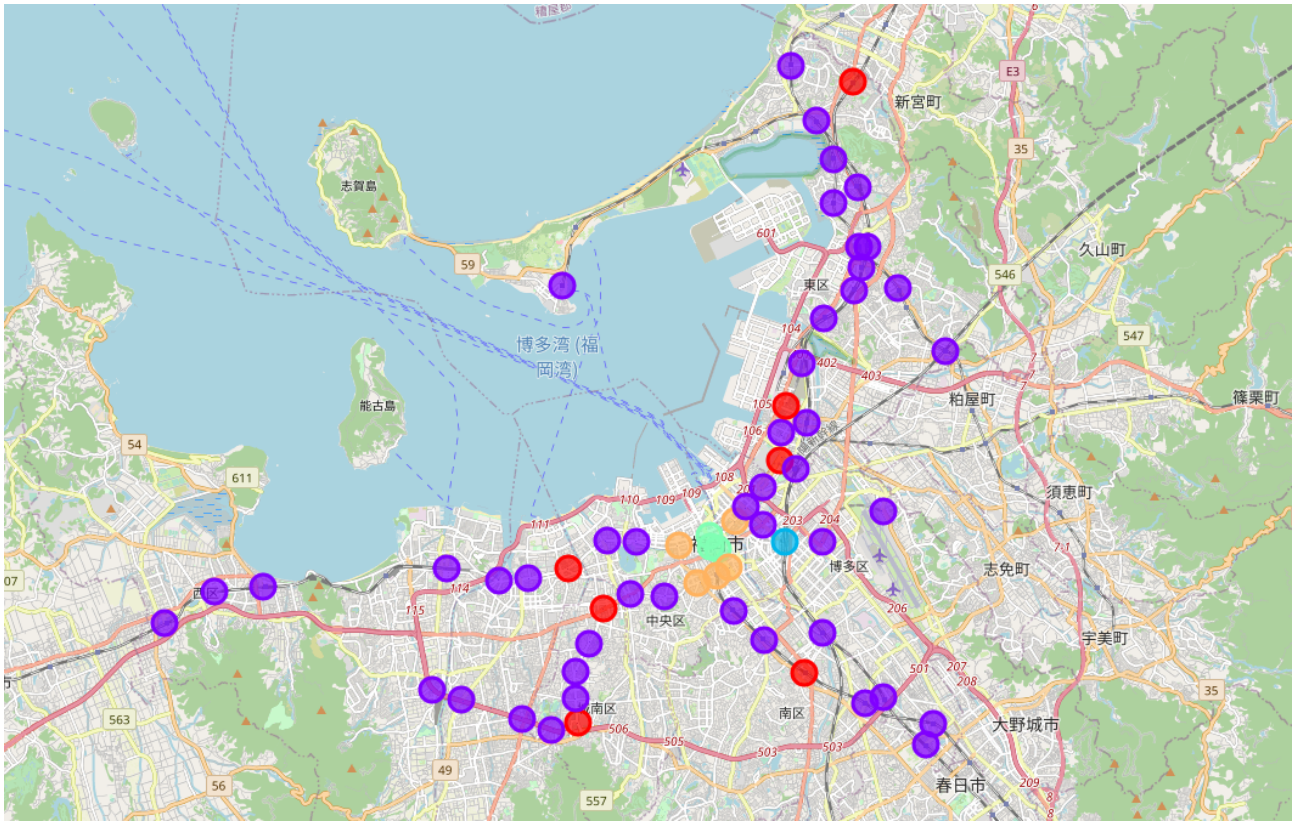


Figure 4. Map of station clusters

4.1 Cluster 0 (dark blue)

The biggest cluster includes 46 mostly remote stations.

The daily foot traffic at stations of this cluster varies a lot from 99913 daily passengers at Meinohama station (that is the 4th biggest station by passengers flow in Fukuoka city) to 1092 daily passengers at the smallest station of Tonoharu. Average foot traffic at this cluster is 15373 passengers.

The common feature of stations in this cluster is a small number of all types of venues in the vicinity. Average number of cafes near stations in these neighbourhoods is 5.

We may conclude that this cluster represents mostly quiet residential neighbourhoods.

4.2 Cluster 1 (sky blue)

One of the clusters includes just one Hakata station.

Hakata station is the busiest in the city, that significant feature is the foot traffic of 461079 passengers daily.

Hakata also has many cafe venues in the vicinity, thereby we can conclude that opening another coffee shop business here would be challenging from a competition perspective.

4.2 Cluster 2 (green)

The next cluster is the cluster of the central area of Tenjin. This cluster includes neighbourhoods of 3 stations

Specific features of this cluster are the maximum number of cafes and a large number of entertainment venues. There is high foot traffic too.

The area is obviously the most popular in the city and promising for entrepreneurs, who are ready for the challenges of business rivalry.

4.3 Cluster 3 (orange)

A noticeable cluster is a cluster of stations surrounding the central area. This cluster includes 5 stations, half of which has above average foot traffic and above average number of cafes and entertainment venues.

Yet, the number of cafes and entertainment venues are prominently smaller than those for neighbourhoods of the Tenjin area.

Considering the geographic location of these stations, we conclude that some of these stations may be centers of actively growing neighbourhoods, which are merging into the central area.

4.4 Cluster 4 (red)

The last cluster of 7 stations obviously represents neighbourhoods of educational institutions.

These neighbourhoods will be the areas of interest for entrepreneurs, whose target customers are students.

The number of coffee shops and entertainment venues is below average (13) in these areas, except for the neighbourhood of Nishijin (15) and significantly smaller than in the areas of the sky blue, green and orange clusters of central areas.

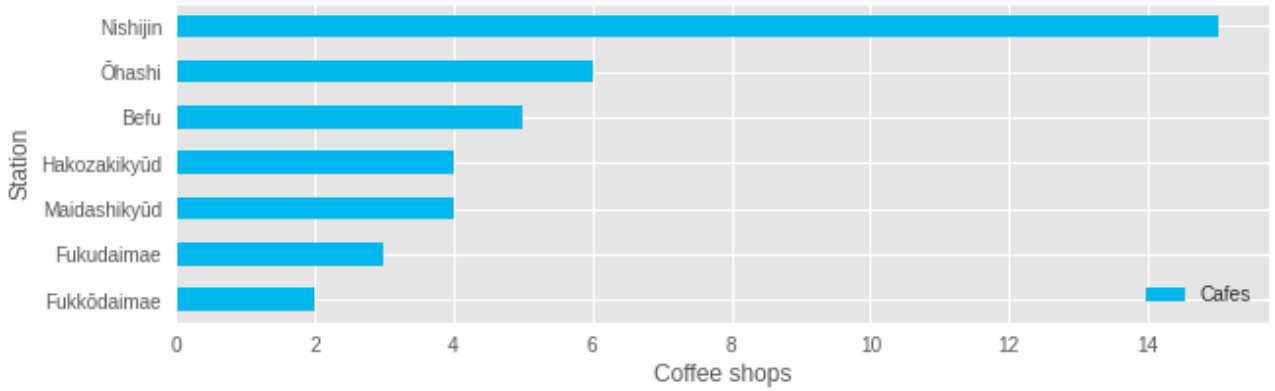


Figure 5. Number of coffee shops near station of the cluster of educational institutions

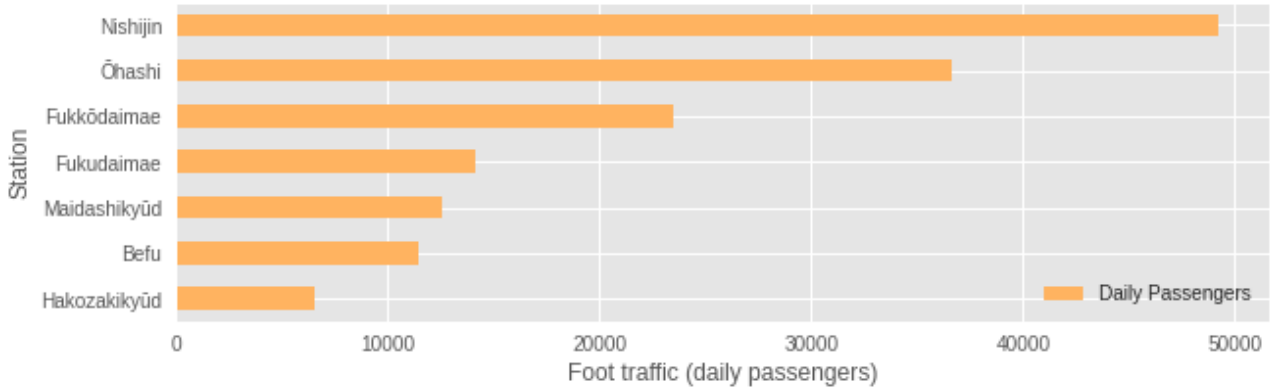


Figure 6. Foot traffic near stations of the cluster of educational institutions.

5. Discussion

As it was stated in the Introduction section, the goal of this analysis was to identify a good location for opening a new coffee shop in the city, such that: a) the area has a decent foot traffic, b) the area is a location of interest and gathering of young people, who are most likely to be active customers of a coffee shop, c) the location has few coffee shops opened yet.

With the consideration of the goal, we would recommend neighbourhoods from two clusters for further investigation, that may include spot inspections or analysis of data from additional sources.

The first recommended cluster is Cluster 3 (orange), due to the geographic location of neighbourhoods of this cluster, we conclude that some of it's stations may be centers of actively growing neighbourhoods, which are merging into the central area that must be popular among young population of Fukuoka city due to a concentration of entertainment venues.

Another recommendation is the Cluster 4 (red) of educational institutions, that includes stations with high foot traffic made by students. In particular, in this cluster we would recommend gathering

more information about Ohashi and Fukkodaimae stations, where the foot traffic is above average in the cluster itself and above average for all stations in the city. At the same time, the number of coffee shop venues is noticeably small at both stations and comparable with the number of venues near small stations in residential neighbourhoods. That would give some competitive advantage to a new venue.

Despite having the recommendations, we acknowledge that there is much room for improvement of this analysis for better conclusions and deeper insights.

One significant gap of this analysis is absence of any information about the difference in cost of rent of premises for business in different neighbourhoods of Fukuoka city.

Another improvement may be done if clustering of areas is performed not only on the higher level of selected types of venues, but with data that includes information about all types of venues too.

6. Conclusion

The purpose of this project was to identify a good location for opening a new coffee shop in Fukuoka city, with respect to criteria identified in the beginning of the project and stated in the Introduction section of this report.

For this report we collected data from 3 different sources and applied different methods of analysis to them, which details are explained in Data and Methodology sections.

In the Results and Discussion sections we provide our recommendations on promising locations and propose a direction for further analysis.

References

1. Coursera IBM Data Science
2. Wiki: Fukuoka City
3. Station names (Mapion)
4. Location data source (Wikipedia)
5. Foot traffic data source (opendata)
6. Foursquare API