Applied data science capstone

Using foursquare api to predict the best location for a new Japanese restaurant in Groningen

By: Alex de Vries

Background

This report will discuss the possibilities of setting up a new Japanese restaurant in the Dutch city of Groningen. This is a student city of around 300.000 people. Because of the university there are students from all over the world. This is not completely represented by the variety of food places in the city. There are two Japanese resaurants that I personaly know of (before starting this analysis). We will find out how many there already are and what the best location would be if someone would want to start a new Japanese restaurant in Groningen. The insights resulting from this analysis could be used for a future business plan.

Problem definition

The goal of this analysis is in the end give advise on what the best location for a new Japanese restaurant will be. Because of the limited knowledge of Dutch people regarding Japanese food, any places related to Japanese food will be considered in this analysis. We might even consider Korean food places as any new Japanese place would likely not want to be close to Korean places as it would detract customers. Chinese restaurants are well known and distinquisable for Dutch people so we can leave those out of the analysis.

Data overview

This notebook is highly inspired by the template given in the course. I will keep the idea of clustering the city by area and then plot heatmap to find better area.

I will change some data:

Country: The Netherlands

City: Groningen

Goal: Open a Japanese restaurant/little shop

So, I will cross data from working days, and localisations.

The following API will be used:

Foursquare API: to find restaurant/venues

Google API: reverse geolocalisation

Methodology

In the Methodology section the data is imported, organised and analysed. The longitude and latitude of Groningen is aquired from Google Maps. The radius value is set to 600 based on testing to cover the relevant area. Tese values including the Foursquare secret ID and version are combined in a Foursquare API call.

Note that the details of the code can be found in the actual report. This presentation will focus on the data manipulation and conclusions.

Methodology

The first 5 values of the dataset after manipulation are shown below. It contains 5 column values and 100 rows in total.

```
grunn_venues.head()
```

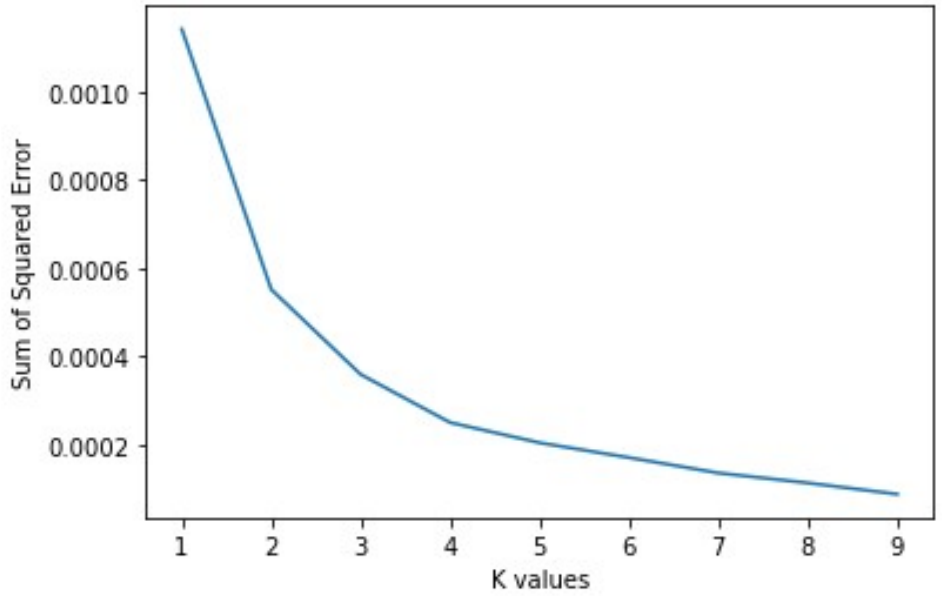
Out[5]:

	name	categories	crossStreet	lat	lng
0	Pure Groningen	Frozen Yogurt Shop	Koude Gat	53.217708	6.566397
1	Der Witz	Bar	NaN	53.218259	6.567054
2	Mr. Mofongo	Restaurant	Academieplein	53.219369	6.564100
3	Gustatio	Italian Restaurant	NaN	53.218200	6.568896
4	Grote Markt	Plaza	NaN	53.218754	6.567711

Results

Before doing k means analysis we need to find the optimal value for K. See the following Figure. We choose K = 4.

```
Out[24]: [<matplotlib.lines.Line2D at 0x7fd3f1cd9ac8>]
```



Results

The following image shows the K-means grouping with K = 4.

In [25]:
```
km = KMeans(n_clusters=4)
cluster = km.fit_predict(grunn_df)
dining_df['cluster'] = cluster
dining_df.head()
```

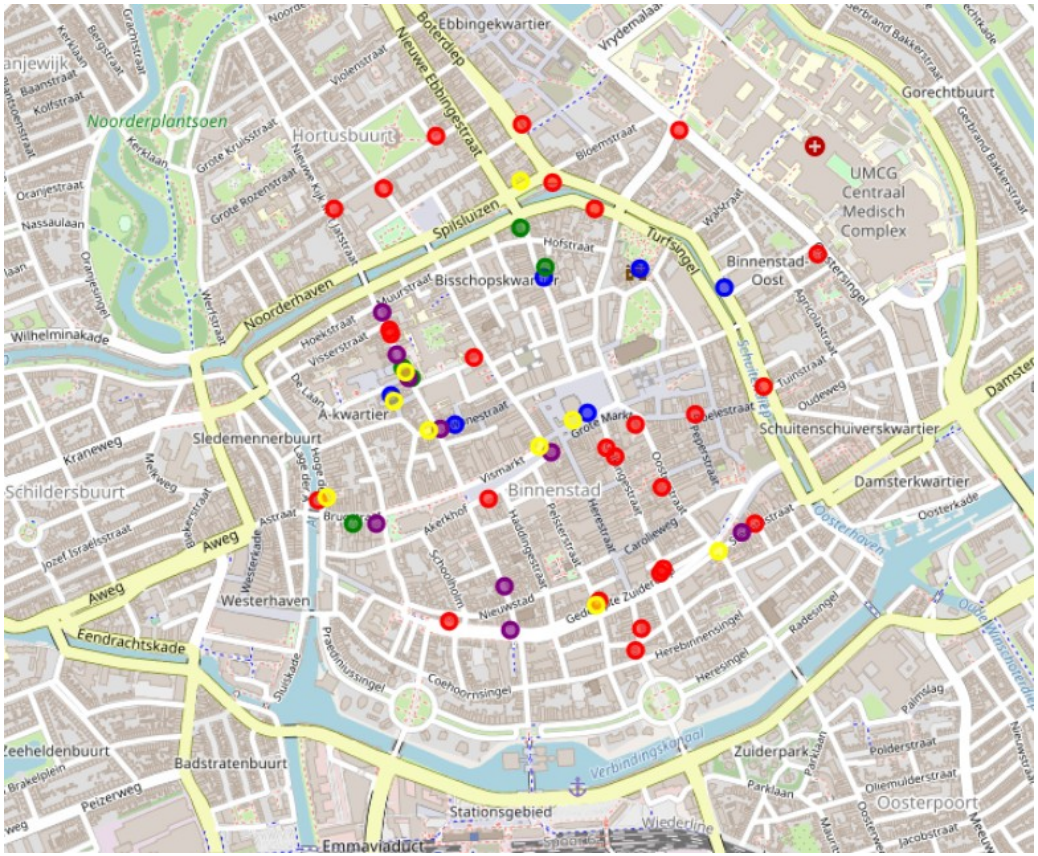
Out[25]:

	name	categories	crossStreet	lat	lng	cluster
2	Mr. Mofongo	Restaurant	Academieplein	53.219369	6.564100	3
3	Gustatio	Italian Restaurant	NaN	53.218200	6.568896	0
5	Cho Fah Eatthai	Thai Restaurant	NaN	53.217775	6.568035	0
20	Imono	Sushi Restaurant	NaN	53.217628	6.568279	0
21	Osteria Da Vinci	Italian Restaurant	NaN	53.221979	6.567679	1

Results

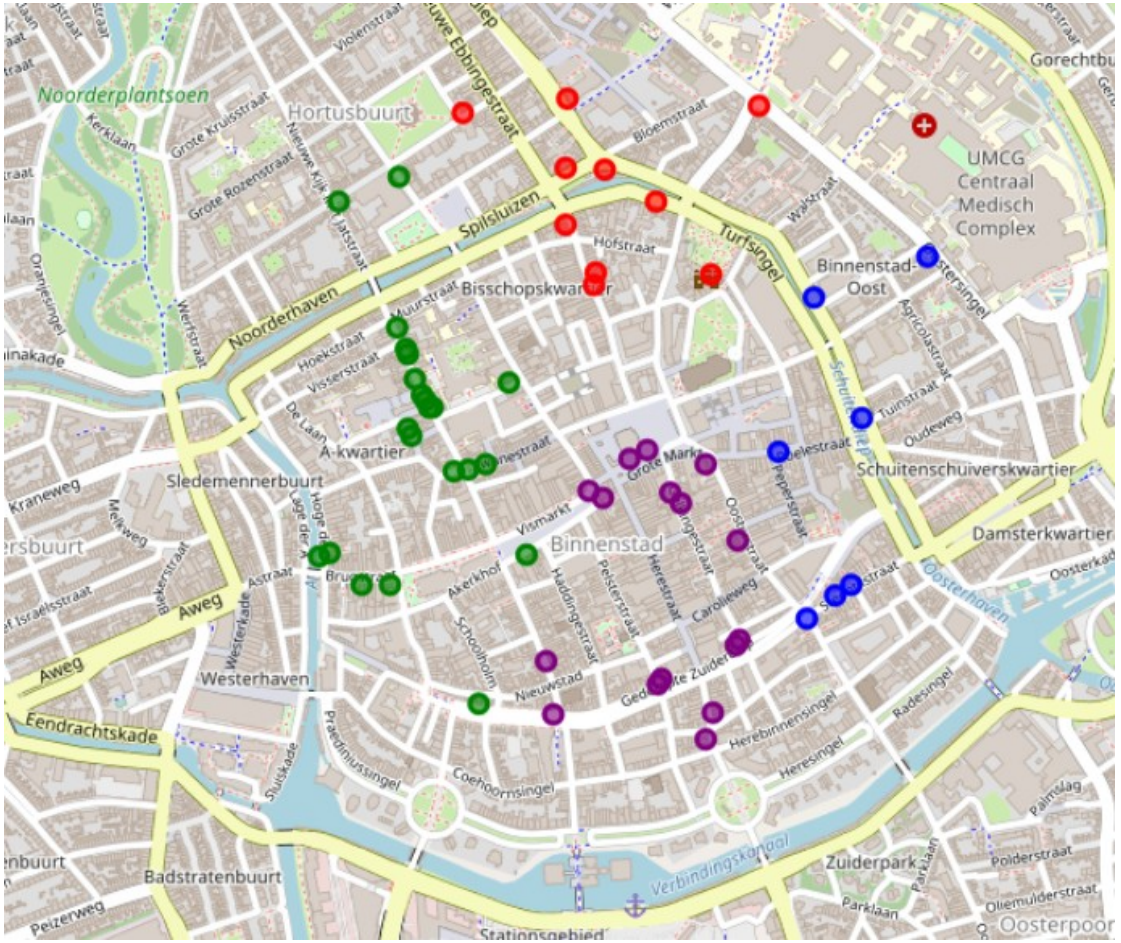
Plotting all food venues on the Groningen map shows the following image.

Important to note is we cannot see a correlation between venue type and location. Next is to label based on K-means grouping.



Results

Grouping with K-means shows 4 clear groups. We can now check if in each of these 4 groups a specific food venue type occurs more often than the others.



Conclusions

We find looking at the previous two Foursquare plots that firstly the different types of food venues are distributed across the whole city. Secondly we find that food venues can be grouped into 4 sections. These groups do seem to line up with the four sections in which the city itself has been sectioned off in. It is yet too early to make conlusions on the data. The last step is to see if the four groups show a pattern of containing more or less of a particular type of venue.

Discussion

The first thing to note is that the analysis is done purely on Foursquare data. To be more confident in our analysis and conclusions we could crossrefference our findings with a different dataset.

Based on the final code in the previous chapter we can show the final distribution of various food venues for each section in the city. It can be seen that there is not an equal distribution of each food venue in each section.

Discussion

There could be arguments to be made for building a new restaurant in either cluster 3 and cluster 4. The reason is that cluster 3 is the most popular area for restaurants, namely 66.67% of all food venues are restuarants. This might be a saturated market and cluster 4 is not yet. This could indicate that cluster 4 is the better area for a new restaurant.

The following table shows for each cluster the distribution of types of food venues:

[29]:

	Cluster	Restaurants	Cafes	Places	Shops	Bars
0	One	10	1	1	3	3
1	Two	6	2	3	0	1
2	Three	4	1	0	1	0
3	Four	6	2	3	4	4

Percentage of food venues that are restaurants for each cluster:

- Cluster 1: 55%
- Cluster 2: 50%
- Cluster 3: 66.67%
- Cluster 4: 31.57%

Conclusion

The goal of this analysis was to find out the best location for a new restaurant. Initially there was an interest to focus on asian restaurants to decide on the best location of a new Japanese restaurant. It turns out that there are hardly any asian restaurants in the city of Groningen to begin with. Because of this the scope of the analysis was increased to focus on all restaurants and other food related venues.

Currently an argument could be made to build a new restaurant in two different sections of the city. Because the city is easily walkable, visitors could access eaither areas without any difficulties. The east section of the city contains the highest concentration of restaurants which indicates that it is a popular area. Since there are hardly any asian restaurants yet it would stand out and attract visitors. The west section of the city contains the lowest concentration of restaurants. Only 33.57% of all food venues are restaurants. This could indicate a less saturated market.

Advise for future analysis: It is possible to evaluate foot traffic for the two identified areas to differentiate between them. A second analysis is to review the resaurants ratings or time they have been in business. A last analysis that could be done to decide which area is more ready for a new restaurant is evaluating rent prices for the building locations.