IBM Data Science Professional Certificate
Capstone Project

Opening a restaurant in Zadar, Croatia

Data Analysis Approach

Matija Sorić

July, 2020

Contents

List of tables....2
List of figures....2
Summary....3
1. Introduction (Business Problem)....4
2. Methodology....5
3. Data Preparation....5
4. Descriptive analysis....8
4.1 K-Means....10
5. Results....11
6. Discussion....12
7. Conclusion....13

List of tables

Table 1. Merged data - neighborhoods + venues....7
Table 2. Neighborhood venues....8
Table 3. Neighborhoods with the most 'food' venues....9
Table 4. Assigned cluster labels....11

List of figures

Figure 1. Dataset creation....6
Figure 2. Geographical map showing Zadar neighborhoods....6
Figure 3. Venue frequency by neighborhood....9
Figure 4. Visualizing ‘food’ venue frequency by neighborhood....10
Figure 5. Generated clusters....11

Summary

This paper is written as a result of a Capstrone Project for the IBM Data Science Professional Certificate.

The analysis which the paper, and the project itself is focused to, is the tourism sector, namely, opening a restaurant in the city of Zadar in Croatia. As Zadar is a popular tourist destination, it always seeks and strives to attract more tourists, offering them its sunny beaches, beautiful sunsets and of course, local mediterranean cuisine, specific for this part of Croatia, and world, in general. The aim of this project is to find the most suitable place for the new restaurant, taking into consideration competition, in form of other restaurants.

The first part of the paper will present neighborhoods in the city of Zadar, extracting their location for further analysis, creating a *Folium* map of the neighborhoods for the easier representation, as well as showing the venues of the each neighborhood.

After that, each neighborhood will be analyzed and the results presented in the form of the heatmaps, showing neighborhoods where the venues of interest are situated.
In the end, *K means* clustering will be implemented, to cluster the neighborhoods with more than 5 venues, for the potential client who opens a restaurant to see which neighborhood is the most optimal as a new restaurant location. As a most optimal neigborhood is chosen a Jazine-Stanovi-Višnjik neighborhoods triangle, as a great leverage between tourist circulation and location.

The results found correspond to the actual situation in the city of Zadar, and can be further used to plan the city development.

1. Introduction (Business Problem)

Zadar is one of the most popular places in Croatia for tourism, each year having tens of thousands of tourists from all over the world. These tourists often seek places with genuine mediterrranean cuisine, which can usually be found in the old city core. However, as the tourists from some countries could not afford the high prices of restaurants in the old city core, this presents the opportunity for the new restaurants, which, if opened on the right location, can present a real success in terms of tourism business.

The investors who would be interested in this analysis are future restaurants owners who are seeking the best place to open a restaurant in terms of profit return.

2. Methodology

The metodology is described into three parts:

- Data preparation – obtaining the dataset and preparing it for analysis
- Descriptive analysis – used in order to better understand the data and to enhance the decision making process
- K-means – understanding the *machine learning* (ML) algorithm used for clustering

3. Data Preparation

Data consists of 25 geolocations (neighborhoods), with their corresponding longitude and latitude values. As the data for this city was not available in an organized format, manual extraction of the 'neighborhoods' was done, having their longitude and latitude taken from a webpage *Hoodmaps*[1]. Result of the extraction and dataset creation (excel file) is depicted on figure 1.

[1] https://hoodmaps.com/zadar-neighborhood-map

A	B	C
Neighborhood	Latitude	Longitude
Brodarica	44,12361111	15,22611111
Voštarnica	44,11694444	15,23583333
Peninsula	44,11444444	15,22555556
Plovanija	44,12611111	15,24833333
Špada	44,12888889	15,24000000
Skročini	44,13277778	15,24472222
Bokanjac	44,14416667	15,24500000
Bili Brig	44,11972222	15,26083333
Crvene kuće	44,11222222	15,26111111
Bulevar	44,11055556	15,24666667
Stanovi	44,11500000	15,24583333
Arbanasi	44,10222222	15,24138889
Jazine	44,11111111	15,23444444
Borik	44,13333333	15,21638889
Puntamika	44,13138889	15,20638889
Mocire	44,13083333	15,22444444
Petrići	44,12750000	15,22861111
Belafuža	44,13222222	15,23472222
Maslina	44,12444444	15,23722222
Smiljevac	44,10972222	15,25250000
Ričina	44,10500000	15,25277778
Sinjoretovo	44,10722222	15,26888889
Gazenica	44,09722222	15,27333333
Višnjik	44,12166667	15,24416667
Diklo	44,13833333	15,21833333

Figure 1. Dataset creation

Also, data (neighborhoods) can be showed geographically, plotted on the map using Folium. Result of plotting is visible on image below, showing 25 neighborhoods:

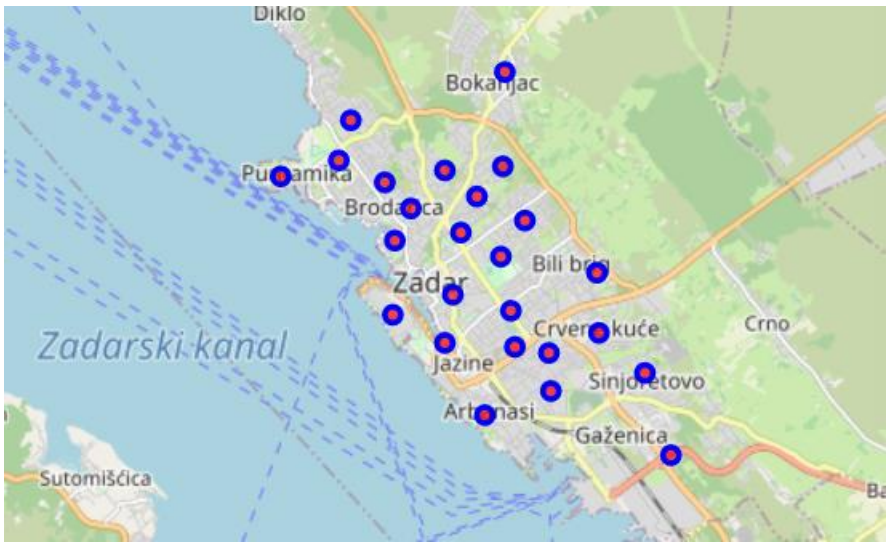

Figure 2. Geographical map showing Zadar neighborhoods

This data was used in conjuction with the Foursquare API, to get the necessary venues in order to perform further analysis. These venues were combined with the above neighborhood data, and the result is depicted in a table below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Main Venue Category	Venue Category
0	Brodarica	44.123611	15.226111	Slatka tajna	44.122089	15.227802	food	Donut Shop
1	Brodarica	44.123611	15.226111	Maraska Park	44.122748	15.224458	parks_outdoors	Garden
2	Brodarica	44.123611	15.226111	Palacinka bar	44.122658	15.224491	food	Restaurant
3	Brodarica	44.123611	15.226111	Djiga	44.122731	15.224572	parks_outdoors	Harbor / Marina
4	Brodarica	44.123611	15.226111	Marex	44.122187	15.223801	education	Cafeteria
5	Brodarica	44.123611	15.226111	Sfinga Park	44.126215	15.227031	parks_outdoors	Park
6	Brodarica	44.123611	15.226111	Restaurant Lungo Mare	44.125226	15.225334	food	Mediterranean Restaurant
7	Voštarnica	44.116944	15.235833	Fast food Papica	44.115216	15.232901	food	Fast Food Restaurant
8	Voštarnica	44.116944	15.235833	Mala Posta	44.116898	15.232496	travel	Bus Station
9	Voštarnica	44.116944	15.235833	da vinci	44.114765	15.233130	food	Café
10	Voštarnica	44.116944	15.235833	Garage Center	44.114113	15.234721	shops	Auto Garage

Table 1. Merged data - neighborhoods + venues

Now, the data has the shape of 144 rows and 8 columns, having columns ‘Venue’, ‘Venue latitude’, ‘Venue longitude’, ‘Main Venue Category’ and ‘Venue Category’, attaached to the initial data.

4. Descriptive analysis

After this step, all the components to perform analysis were collected, having necessary neighborhoods, plus the venues for each eneighborhood. Next step was to group these venues by neighborhood, to see the neighborhoods with the most venues.

As said in the Introduction part of this paper, as expected, the most venues are exactly at the city core, or ‘Peninsula’, as shown on table below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Counts
0	Arbanasi	44.102222	15.241389	4
1	Belafuža	44.132222	15.234722	3
2	Bili Brig	44.119722	15.260833	2
3	Bokanjac	44.144167	15.245000	1
4	Borik	44.133333	15.216389	8
5	Brodarica	44.123611	15.226111	7
6	Bulevar	44.110556	15.246667	1
7	Crvene kuće	44.112222	15.261111	1
8	Diklo	44.138333	15.218333	6
9	Gazenica	44.097222	15.273333	4
10	Jazine	44.111111	15.234444	8
11	Maslina	44.124444	15.237222	5
12	Mocire	44.130833	15.224444	2
13	Peninsula	44.114444	15.225556	47
14	Petrići	44.127500	15.228611	5
15	Plovanija	44.126111	15.248333	1
16	Puntamika	44.131389	15.206389	5
17	Ričina	44.105000	15.252778	6

Table 2. Neighborhood venues

Each neighborhoods main venue category was then analyzed, however, for this analysis, or category of interest was ‘food’, since restaurants fall into that category. Then neighborhoods were analyzed in a manner to see the frequency of venues in each neighborhood, results of which are shown in a figure below:

```
----Peninsula----
                      venue   freq
0   Neighborhood Latitude  44.11
1  Neighborhood Longitude  15.23
2                    food   0.51
3      arts_entertainment   0.15
4               nightlife   0.13


----Petrići----
                      venue   freq
0   Neighborhood Latitude  44.13
1  Neighborhood Longitude  15.23
2                    food   0.20
3          parks_outdoors   0.20
4                   shops   0.20
```

Figure 3. Venue frequency by neighborhood

For the ‘food’ category to be shown in numbers and proven with the analysis, table below can be useful.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	food
4	Borik	44.133333	15.216389	5
5	Brodarica	44.123611	15.226111	3
13	Peninsula	44.114444	15.225556	24
20	Smiljevac	44.109722	15.252500	3

Table 3. Neighborhoods with the most 'food' venues

As visible in an output above, the old city core, ‘Peninsula’, has the most frequent venue exactly the ‘food’ category, which is researched in this project. As said, probably because the tourist frequency is the highest in that part of the city.

‘Peninsula’ has 24 venues categorized as ‘food’, which is presented on two figures below, where the supremacy of the restaurants in the ‘Peninsula’ neighborhood can be observed.

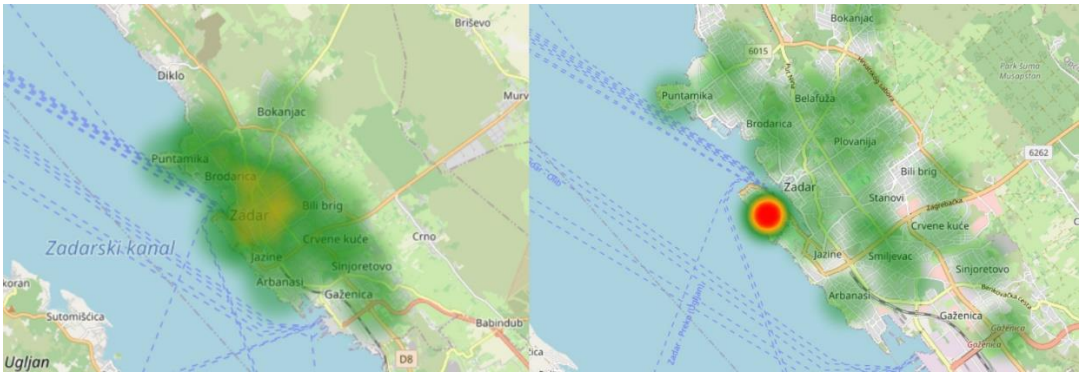

Figure 4. Visualizing 'food' venue frequency by neighborhood

4.1 K-Means

In order to develop further data insights, K-means clustering was used. Goal of K-means is to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid). Let the set of points be $X = x_1, x_2, x_m$ in n dimensional Euclidean space $\mathbb{R}_n$. The goal of the k-mean algorithm is to find, for a predetermined number k >= 2, the optimal k-partition of the set X in k sets, $C_1, C_2, C_3 \ldots C_k$ by measuring the means of the points and aligning the centroid accordingly, until the centroid fixates and no longer moves. At that point, k-means algorithm is finished. In *sci-kit learn* (*sklearn*) library in Python this can be done very easily by calling 'from sklearn.cluster import Kmeans' and using the Kmeans method. The most difficult thing is to find the optimal number of clusters.

In this case, K-means essentially grouped the spots in the city in clusters, depending on the number of restaurants in each neighborhood. Results were the clusters, with each cluster representing certain number of restaurant in that area. K-means algorithm was hereby used for the clustering of the aforementioned data, because it is simple to implement and intuitive to understand. For calculating inter and intra-cluster distances, Euclidian distance was calculated between the points, so that the points could be separated in different clusters. K-means algorithm was developed by by Arthur and Vassilvitskii in 2007.

Data was cleansed beforehand forwarding it to the algorithm, in that sense that only the neighborhoods where there are more than two 'food' categories were included in the K-means. That is because, if the neughborhood has 2 or more food venues on a relatively small

area of one neighborhood, our new restaurant would have more competition, and therefore it woule be less profitable.

5. Results

As there were 4 distinct values (number of restaurants) for each neighborhood, K-means algorithm implemented here separated the data in 4 clusters. Neighborhoods were, after running the algorithm, found to have the following cluster labels:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	food	Cluster Labels
4	Borik	44.133333	15.216389	5	2
5	Brodarica	44.123611	15.226111	3	3
8	Diklo	44.138333	15.218333	2	0
10	Jazine	44.111111	15.234444	2	0
13	Peninsula	44.114444	15.225556	24	1
16	Puntamika	44.131389	15.206389	2	0
17	Ričina	44.105000	15.252778	2	0
20	Smiljevac	44.109722	15.252500	3	3
23	Voštarnica	44.116944	15.235833	2	0

Table 4. Assigned cluster labels

After creating cluster labels, data was visualized on Folium map, and the results are the following:

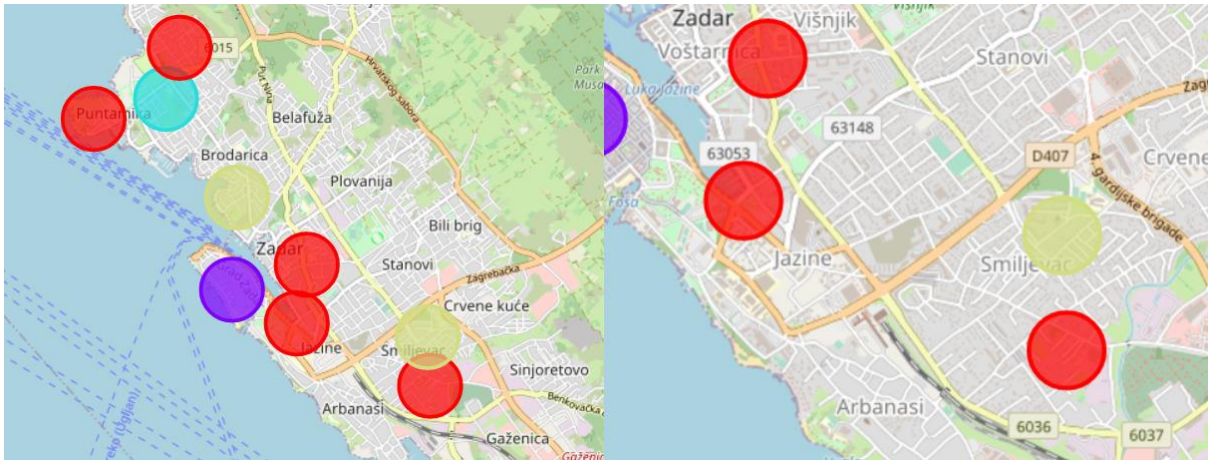

Figure 5. Generated clusters

As there are 4 distinct clusters, each one is representing a different number of restaurants in each neighborhood:

1. Red cluster – 2 food venues
2. Yellow cluster – 3 food venues
3. Light green cluster – 5 food venues
4. Purple cluster – 24 food venues

That being said, the most logical place, taking into consideration the competition, in the form of other restaurants, rent of the location, as well as the tourist circulation in Zadar, is the Stanovi-Jazine-Višnjik-Smiljevac quadrant.

6. Discussion

By this analysis, it was shown that an ML algorithm can be implemented in tourism sector, and help potential restaurant owners. Aside from clustering, machine learning can be implemented in other areas in tourism, such as predicting a number of tourists in a year, taking various features into consideration.

On this example, it is visible that not only data is enough for making a usefuul analysis. Because, by the analysis, it could simply be concluded that a new restaurant should be in the neighborhoods where there arent any restaurants, such as Bokanjac, Bili brig or Gaženica. If that was the case, the analysis would fail miserably, because if it concluded that there should be a restaurant in Gaženica, which is a port, we can deduce that the restaurant opened there would not be very sucessful. So, the matter is delicate, because it is needed to pick a place where the tourist circulation is existent, and it is exactly in the quadrant mentioned in the chapter above, however, we should not choose the neighborhoods where there are no ‘food’ or other attractive tourist venues, because tourists usually do not spend time on that places.

7. Conclusion

As found in the analysis, the best place for a new restaurant in the city of Zadar would be in the Stanovi-Jazine-Višnjik-Smiljevac quadrant.
However, although the most restaurants are situated on ‘Peninsula’, if the client ordering this analysis would be able to obtain the ‘right’ location in the ‘Peninsula’ with an economic (cheap) rent, despite the competition from the other restaurants, it could also be successful.
As there are more and more new places opening everyday in Zadar before the summer season, this analyis could already be outdated, and presumably another (updated) analysis should be conducted, with the exact number of venues for each neughborhood. For the most accurate results, it should be conducted before every summer season, for the employer (in this case, restaurant owner) to be able to generate the most revenue.