Ver. 11.6.2020 Author: Niko Juntunen

Finding the best location for a new vegetarian restaurant in city of Helsinki, Finland

Introduction / Business Problem

In recent years there has been huge grow in popularity of vegetarian diet in Finland and especially in the capital, Helsinki. Many new vegetarian restaurants have been launched during 2010s in addition to punch of veg restaurants that has been around for years. Coming to 2020s the popularity for healthy and environmental friendly restaurant options is growing faster than ever before. Due to this I have choose to find out what location would be the best location for new vegetarian restaurant in Helsinki.

To solve the problem we can easily utilize Foursquare location data of Helsinki city areas neighborhoods and cluster the given venues into desirable segments. Our target audience is any restaurant business owner who is about to launch a new vegetarian eatery in the near future in Helsinki. We will be provide an analysis of the current locations of vegetarian restaurant and answer the question where the competition is favorable for new business.

To answer the main question we will have to find out where the most of Helsinkis current vegetarian restaurants are located. As the city is rather small we will avoid the most popular area given that the competition is there already too high. Instead we will utilize the base information of the citys structure. We know that the "heart" of Helsinki is around the main railway station and we don't want to go too far from that point when creating new business. We will find the optimum location for a new restaurant near that area keeping mind that we should avoid the most competitive area.

We will also consider as a optimum location a place that has many restaurants and other venues but does not have yet any vegetarian ones. In addition there has been recently huge debate about too many new shopping centers in Helsinki area. Causing the situation where the restaurant and other business owners are complaining that there is not anymore enough customers for their business. This goes especially to the new shopping centers. So we will utilize that information when choosing the best location. Even so we will carefully study also those locations and their surroundings.

I don't have any background in running restaurant business so it is presumable that as the project proceeds there will be some additional information I will utilize finding the best location.

Data

In the project I will mainly utilize location information that can be found from Wikipedia and Foursquare. From Wikipedia I will get the data of subdivision of Helsinki and with that I will create dataframe including all the neigborhoods and their subareas called quarters in Finland. In addition I will combine to the dataframe the latitude and longitude information of each location. That I will achieve by utilizing Geopy.

From Foursquare I will request the venues of all the locations I have in the dataframe. After having the venues listed I will start analysing the information by visualizing the venues over a map and creating new dataframes including the favorable features(many venues densely etc.). The outcome will be my data for analysing the current restaurant locations in Helsinki and that I will cluster into veg- and non-veg restaurants and after that I will see if there is need to further clustering for veg restaurants only.

In addition we will determine the main railway station (latitude: 60.1698, longitude: 24.9382) to be the "heart" of Helsinki and because of that it also to be the base location for our study.

Methodology

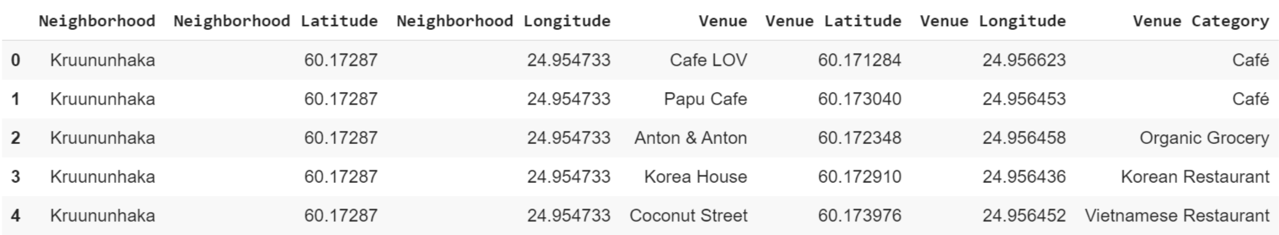
Creating the dataframe and cleaning the data

In order to get data in form of dataframe we are utilising Pandas read funktion. After that we start to clean the data. As Finland is bilingual country the data contains also Swedish names for locations. As we don't need them we drop them from the dataframe. We also create a for -loop so we can manipulate neighorhoods to be in desirible form for us. The neighborhoods that has quarters we drop, but first we make the quartes of them to be the neigborhoods. In our case this method is more efficient and accurate than use the actual neighborhoods which only a few contains quarters and rest are just actual neighborhoods on their own.

Next we request the latitude and longitude information with Geopys Nominatim. Then we add the given values for the existing dataframe of location. This we achieve with a simple for -loop. To make this method working we just make sure that the order of the origal dataframe and lati/longitude list maintains mutable during the operation.

Then we request the location data from Foursquare. To see everything works we first request data just for one location. We limit the results to 150 and in radius of 500 meters. Inspecting the result and we are find it to be correct. Then we request location data for all our locations. For the process we are using getNearbyVenues -function which was provided in the course materials(IBMs Applied data science). This will get the venues from Foursquare and compine them to our dataframe.

During the process important observation is that we can easily run out of our maximum amount of daily requests from Foursquare as we are using free account. Problem occurs when we need to run our code multible times during the same day due to the fixing the code or becouse of a crashing platform/kernel. This issue we fix by saving the Foursquare results in a csv -file. Then we can load that any time we like without consuming our daily requests.

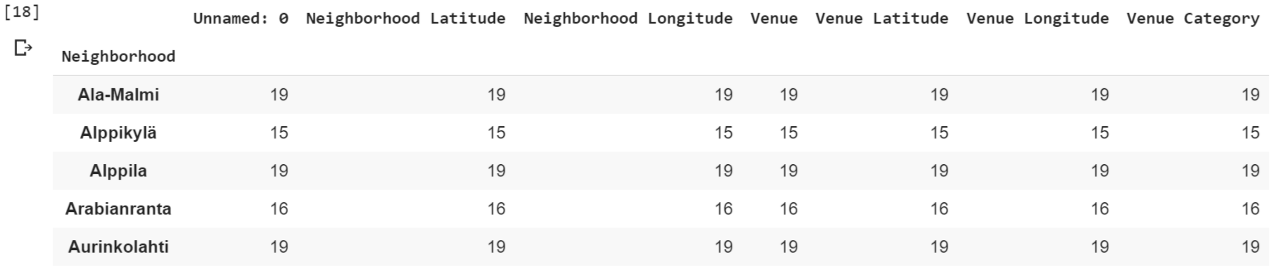


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Kruununhaka	60.17287	24.954733	Cafe LOV	60.171284	24.956623	Café
1	Kruununhaka	60.17287	24.954733	Papu Cafe	60.173040	24.956453	Café
2	Kruununhaka	60.17287	24.954733	Anton & Anton	60.172348	24.956458	Organic Grocery
3	Kruununhaka	60.17287	24.954733	Korea House	60.172910	24.956436	Korean Restaurant
4	Kruununhaka	60.17287	24.954733	Coconut Street	60.173976	24.956452	Vietnamese Restaurant

SCREENSHOT - Our newly created and cleaned dataframe with location, coordinates and venues compined.

Analysing the data

The venues and categories



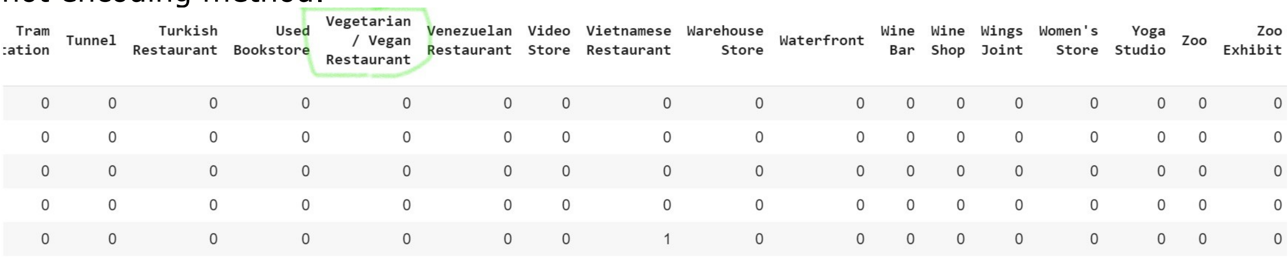
[18]

Neighborhood	Unnamed: 0	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Ala-Malmi	19	19	19	19	19	19	19
Alppikylä	15	15	15	15	15	15	15
Alppila	19	19	19	19	19	19	19
Arabianranta	16	16	16	16	16	16	16
Aurinkolahti	19	19	19	19	19	19	19

SCREENSHOT - We start our analyse section by using groupbys count funtion to see how many venues there are in each loacation.

Here we get the first idea of amounts of the venues. We can see that only a punch of locations have a big numbers of venues. We will get into those numbers more deebly in later stage of our study.

In order to inspect what kind of venue categories we have in our locations we use one hot encoding method.



Tram :ation	Tunnel	Turkish Restaurant	Used Bookstore	Vegetarian / Vegan Restaurant	Venezuelan Restaurant	Video Store	Vietnamese Restaurant	Warehouse Store	Waterfront	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	Zoo	Zoo Exhibit
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

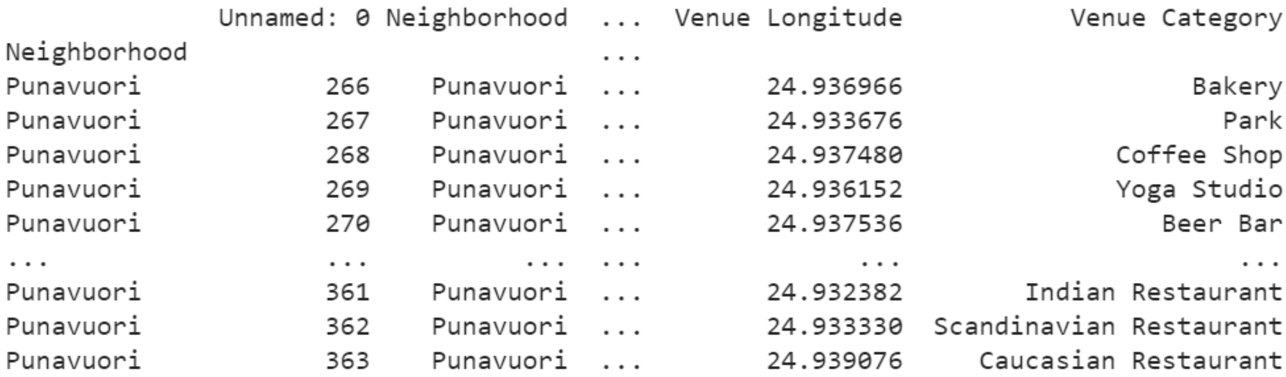
SCREENSHOT When inspecting our one hot encoded dataframe we observe that one category is especially well fitted for our needs.

The venues “hot spot”

Next we will see what location have the most venues. This location will be our “hot spot” of Helsinki venues. We utilize again groupby and this time we use idmax -funciton which reveals us the location. As the result is the name of the location we next use the given name to check how many venues there actually is by using .loc -method for our goupby dataframe. The result is Punavuori with 100 venues.

Our observation at this point is that the excact number of 100 is quite unlikely and points at our fixed limit in Foursquare request which we set originally to be 100. To make sure that the number in Punavuori is really 100 and not more we go back in the code and set limit to 150 and make the Foursquare request again. For a little suprise we find out that the actual number really happens to be excactly 100.

We continue by inspecting Punavuori area more deeply. From our original dataframe with venues we print full results of Punavuori with .loc -method. This gives us good idea of what kind of area the given location is. We can see that it is clearly a place where people go eat and drink. Location contains very high density of restaurants but not any vegetarian restaurant. That is important notice for us and we will keep that in mind for further analysing.



```
             Unnamed: 0 Neighborhood  ...  Venue Longitude           Venue Category
Neighborhood                          ...                                          
Punavuori           266    Punavuori  ...        24.936966                   Bakery
Punavuori           267    Punavuori  ...        24.933676                     Park
Punavuori           268    Punavuori  ...        24.937480              Coffee Shop
Punavuori           269    Punavuori  ...        24.936152              Yoga Studio
Punavuori           270    Punavuori  ...        24.937536                 Beer Bar
...                 ...          ...  ...              ...                      ...
Punavuori           361    Punavuori  ...        24.932382        Indian Restaurant
Punavuori           362    Punavuori  ...        24.933330  Scandinavian Restaurant
Punavuori           363    Punavuori  ...        24.939076     Caucasian Restaurant
```

SCREENSHOT Shortened version of our print out inlcuding all the venues in Punavuori.

Venue frequensies

The next step is to see what kind of venues we have in each location. This we can achieve by cheking the frequensies of the venue categories. We utilize our previously made groupby dataframe and loop it to get the results. We limit the result to 5 categories and with .sort function we get them in right order to form top5 lists.

We found out that Kaartinkaupunki and Munkkiniemi, has vegeratian restaurants among their top5 venues. With that information we can decide that Kaartinkaupunki and Munkkiniemi has already too much of competition.

Next we will remove all the other venues but vegetarian / vegan restaurant and see the frequency for them only. We achieve this by the same method we previously used, in this time we just make a new dataframes of gropby dataframe and use the category: “Vegetarian / vegan Restaurant.

By inspecting the frequesies we will find out that only a few places has vegetarian restaurants in Foursquares database. This is possibly because of unpopularity of the service inFinland where most people use Google nowadays. With a quick Google Maps search we can see that there are more vegetarian restaurants in Helsinki locations that Foursquare claims. Anyway as we are tide to Foursquare in this project we will continue.

The good news in the data is that it reveals some facts that are well known in Helsinki area: by looking the data we can see that vegetarian restaurant "hot spots" in Helsinki are Harju, Torkkelinmäki and Linjat, these areas being the well known for veg options. In addition data reveals three not so well known locations: Kaartinkaupunki, Ala-Malmi, Itä-Pasila and Munkkiniemi. Now we will check how many venues each of these location has, so we can see if any of them is our intrest for the best location.

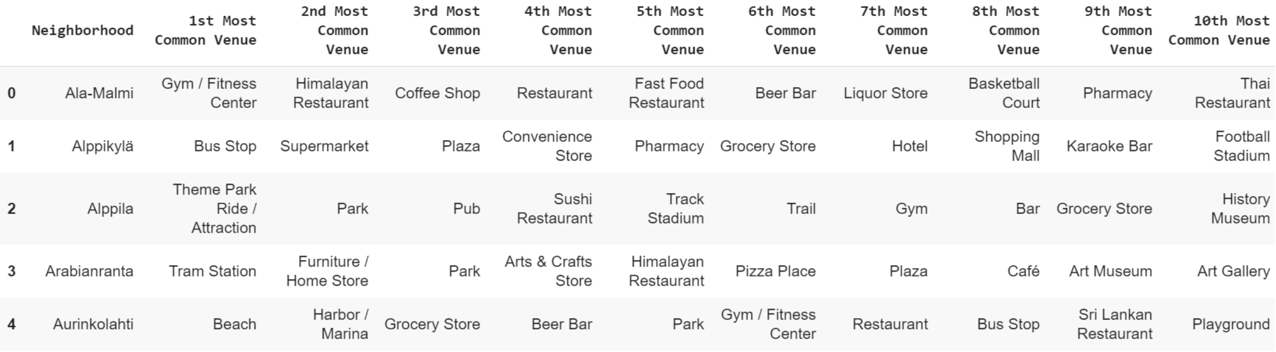
Venues in numbers, clustered and visualized

```
Venues by location: Harju: 59 Torkkelinmäki: 88 Linjat: 69 Kaartinkaupunki:
56 Ala-Malmi: 19 Itä-Pasila: 25 Munkkiniemi: 17
```

SCREENSHOT Our print out reveals the numbers of venues in desired locations.

As we remember Helsinki venue hot spot is the area of Punavuori that has 100 venues. Compared to that only Torkkelinmäki and Linjat (of high frequensy veg restaurant locations) stands up with venues of 88 and 69. We will keep those locations in mind for further analyse. Next we will see what kind of venues each location has, cluster and visualize them.

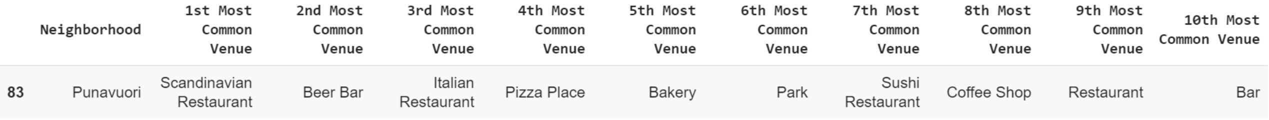
For a first step of clustering the location by venues we first create a new dataframe that shows us the top10 most common venue categories for each location. We utilize return_most_common_venues -function and for-loops that were introduced in the course materials (IBMs Applied data science) .



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Ala-Malmi	Gym / Fitness Center	Himalayan Restaurant	Coffee Shop	Restaurant	Fast Food Restaurant	Beer Bar	Liquor Store	Basketball Court	Pharmacy	Thai Restaurant
1	Alppikylä	Bus Stop	Supermarket	Plaza	Convenience Store	Pharmacy	Grocery Store	Hotel	Shopping Mall	Karaoke Bar	Football Stadium
2	Alppila	Theme Park Ride / Attraction	Park	Pub	Sushi Restaurant	Track Stadium	Trail	Gym	Bar	Grocery Store	History Museum
3	Arabianranta	Tram Station	Furniture / Home Store	Park	Arts & Crafts Store	Himalayan Restaurant	Pizza Place	Plaza	Café	Art Museum	Art Gallery
4	Aurinkolahti	Beach	Harbor / Marina	Grocery Store	Beer Bar	Park	Gym / Fitness Center	Restaurant	Bus Stop	Sri Lankan Restaurant	Playground

SCREENSHOT The new dataframe with most common venues gives us even more accurate information of our locations.

Having sorted the venue information, we will check once again how does the venue hot spot, Punavuori looks like.

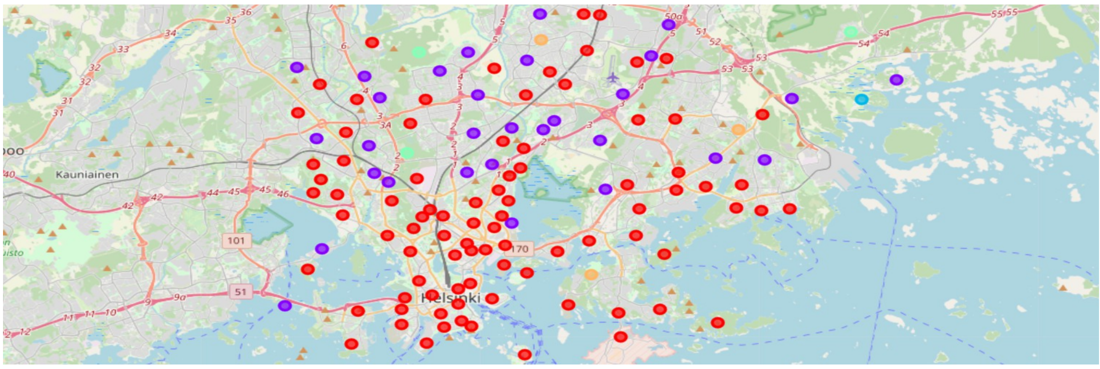


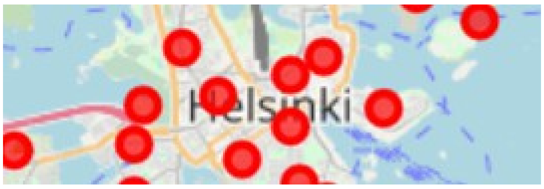
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
83	Punavuori	Scandinavian Restaurant	Beer Bar	Italian Restaurant	Pizza Place	Bakery	Park	Sushi Restaurant	Coffee Shop	Restaurant	Bar

SCREENSHOT Here we can see that Punavuori that has the biggest amout of venues (100) in Helsinki has a very favourable top10 list of venues when looking for an area where people go to eat and drink.

From here we continue by clustering the locations in Helsinki so we can see is there any big differences between the locations near our base location, main railway station. We start by creating a new dataframe that contain a cluster label for each location. We achieve that by utilizing sklearn -librarys KMeans. We set the amount of clusters to be 5 and labels to be 10 as it is our newly made top10 common venues dataframe. By running the Kmeans with our code we will get new dataframe with cluster labels. Now we can utilize that for visualizing the clusters.

By using matloplib and Folium -libraries we visualize the clustered location over Helsinki map. A for -loop will add each clustered location over a map with a color that indicates ist cluster label in range 1-5.





SCREENSHOTS Visualizing the clustered venues we can see that around our base location, main railway station, the locations are very homogeneous.

Results

After clustering and visualizing the venues data we can find that around our base location, main railway station, the locations in matter of venues are very homogeneous. There are variety in the location that are further, among those, we can find locations that we found earlier to have high density of veg restaurants. As they are too far from our base location we wont be analyse them more. The well known "vegetarian hot spots", which our analyse also confirmed, fell into same cluster as all the location around main railway station. We already analysed those locations and find out that most of them have not enough venues to be desireble location for us. We found out that two location, Torkkelinmäki and Linjat has high venues value but as they are already among the location that has high frequency of veg restaurant we wont consider them to be optimium location for us.

After these findings we can reveal the result of our study and state that propability is high for one certain location to be the best area for a new vegetarian restaurant. That is Punavuori which has the highest frequensy of venues and which is close by(in radius of 500 meters) railway station. According to Foursquare database there is not any vegetarian restaurant yet in Punavuori area, but still the area is very popular for restaurants and other venues as our analyse proved.

Discussion

This study was made in part of IBMs Applied data science course and it was requested that study uses Foursquare database. Unfortunately we found found Foursquare database not to be the best one for our case. For further analyse we would use Googles or HappyCow APIs. Due to this observation there is high risk that the result of the study is biased.

Conserning to the viability of Foursquarem database in our case we found out also that for further analyse it would be important to have more in depth information of restaurants. For example in Foursquare database we noticed that there are some ethinic restaurant categories where the restaurant are actually serving only or mostly vegetarian food, and still they are not in vegetarian categories.

In addition one quite important observation was that clustering the locations did not give us so much added value. Anyhow we could achieve more value from clustering by increasing the clusters number. We could also centralise more our inspected area and drop the locations that are far from our base location. But as stated for further analyse we should use Googles or HappyCows APIs.

Conclusion

For the conclusion we can state that dispite of the risk of bias the Foursquare database caused this study gives a good insight at least one of the best locations for new vegetarian restaurant in Helsinki. It proved information that most people in Helsinki area already knows but also revealed added information when considering good locations for veg reastaurants. It also can be said that this study gives good base for further studies and analyses of the given field.