Title

Yiannis Hadjiyianni

March 20, 2021

1. Introduction

1.1 Background

Cyprus, a Greek island located in the Eastern part of the Mediterranian sea, has seen a great rise in demand for technology jobs in the city of Limassol. This resulted in a great deal of people moving from the island's capital city, Nicosia, to Limassol in search of a job. This means that a lot of people must leave their homes and move to different environments. Helping people find a place to live in Limassol where the environment is similar to where they lived in Nicosia is beneficial for both the moving workers and their employees because moving to a similar urban environment will take a smaller toll on their productivity.

1.2 Data

Data that might contribute to determining similar neighborhoods between Nicosia and Limassol might include their locations and nearby venues.

1.3 Interest

As stated before, people moving to Limassol for work would be interested in finding the best areas to move to based on their current living location as well as their employees that would benefit from their employees' well-being and increased productivity.

2. Data Acquisition and Cleaning

2.1 Data sources

Both districts of the cities' of Nicosia and of Limassol are divided into municipalities which are going to be used as the locations of interest that a person might move to. To get the municipalities of each cities' we are going to use an excel file provided by the Postal Service of the Republic of Cyprus, which we can find on the Service's website. We are also going to need the venues of each municipality along with each venue's category. This data will be acquired with the use of Foursquare's API. In order to use the API effectively we are also going to need the coordinates (latitude and longitude) of each municipality. We are going to get this data using the Nominatim API.

2.2 Data Cleaning and Extraction

Firstly, all the data from the excel file provided by the Postal Service was read and any unnecessary, missing or duplicate data was removed. What we were left with was all the municipalities in the districts of Nicosia and Limassol.

Secondly, changes were made to our clean municipalities dataset in order to face with the following problems:

- Municipalities whose names were of the following form: "X of Nicosia/Limassol" (where X is the municipality's name). These municipalities had their name changed to X.
- Municipalities whose names were in Greek (e.g. "Lefkosia" instead of Nicosia). These municipalities had their name changed to its English version.
- Municipalities who were in too rural of an area were removed.

The above stated problems needed to be dealt with in order for the Nominatim API to be used effectively.

Thirdly, after each municipality's coordinates were acquired using Nominatim API, most common venues in each municipality (explained in **2.3**) were found using Foursquare API.

At the end all acquired data was combined into one table.

2.3 Venue Selection

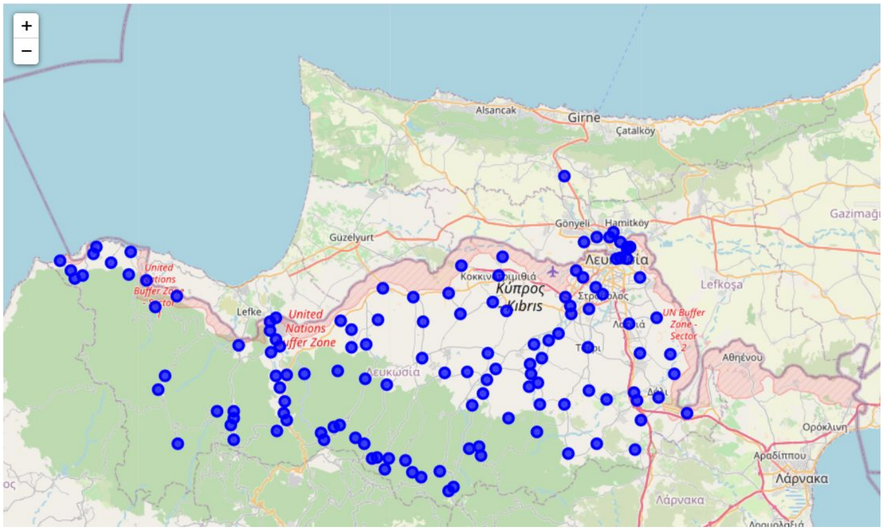
For each municipality a sample of 100 venues was acquired from which the 6 most commonly appeared categories would represent the municipality's "taste". The sample consisted of 100 venues because that was considered as a logical number of venues to be in a 2km radius of the municipality's center. Also, a number of different numbers of most commonly appeared categories were tested and 6 was considered the best. The reason that this occurs is because any number above 6 will result in missing values in our dataset and any number below would provide less of the available data.

3. Exploratory Analysis

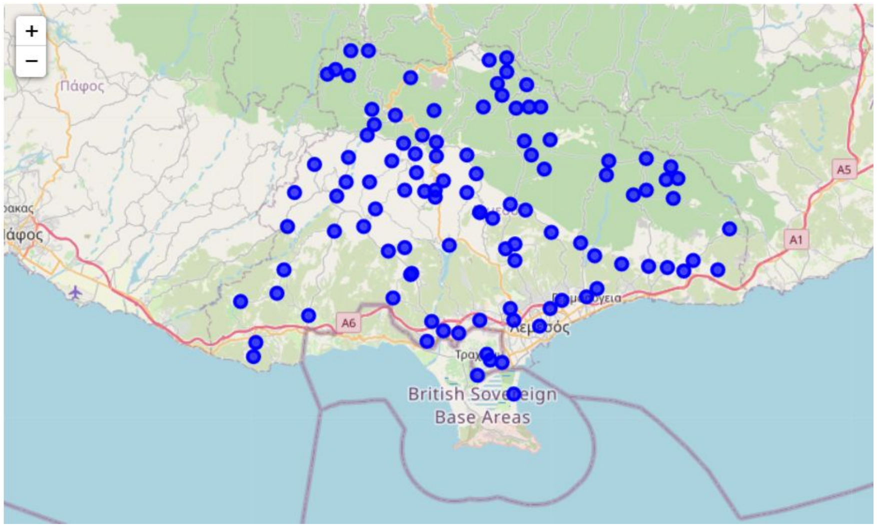
3.1 Locations of Municipalities

Once coordinates of all Municipalities were acquired using Nominatim API the following maps were produced to visualize and explore municipalities' locations.

For Nicosia:



For Limassol:



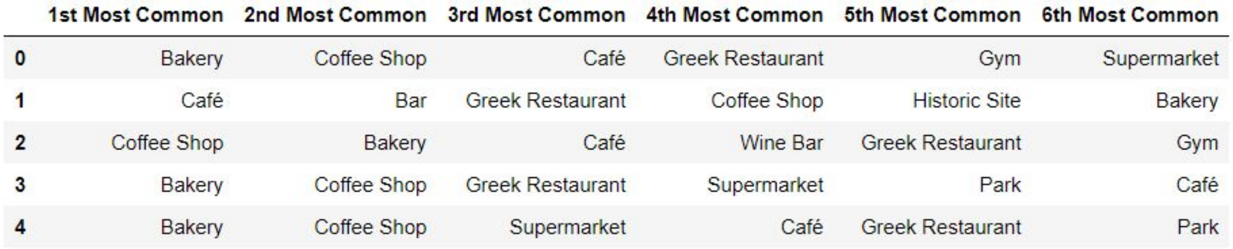
Consider the following:

- Locations of interest are spread all across each district, thus better representing a wide range of areas.
- Both rural, coastal and urban areas are represented.
- Subcategories of environments are also represented (e.g. areas near bodies of freshwater, areas near main highway roads etc.)

3.2 Venues of Municipalities

Using Forsquare API we acquire a sample of 100 venues per municipality (size of sample explained in **2.3**). These venues will be used as the variable of similarity between locations of interest, but using a list of 100 venues per municipality wouldn't be efficient or effective at all. Instead we decided to use, as a determining factor, each venue's category. Thus, we categorized each venue and got the top 6 most common venues per municipality (6 most common explained in **2.3**)

The following is a sample of the data generated:



	1st Most Common	2nd Most Common	3rd Most Common	4th Most Common	5th Most Common	6th Most Common
0	Bakery	Coffee Shop	Café	Greek Restaurant	Gym	Supermarket
1	Café	Bar	Greek Restaurant	Coffee Shop	Historic Site	Bakery
2	Coffee Shop	Bakery	Café	Wine Bar	Greek Restaurant	Gym
3	Bakery	Coffee Shop	Greek Restaurant	Supermarket	Park	Café
4	Bakery	Coffee Shop	Supermarket	Café	Greek Restaurant	Park

4. Predictive Modeling

4.1 Types of models chosen

The approach to facing the problem of comparing the different locations of interest (municipalities of Nicosia vs municipalities of Limassol) was decided to be based on two types of models.

Initially, a clustering model would be used on the data regarding locations of interest in Nicosia, as according to the stated problem, people are moving from Nicosia to Limassol and not in both ways (i.e. from Limassol to Nicosia). The clustering model would help in grouping similar locations of interest in Nicosia, thus creating "categories" of municipalities. Based on those categories decisions could be made about what are the best locations to move to in Limassol.

After clustering Nicosia's locations of interest, a classifying model would be used on the data regarding locations of interest in Limassol, which will help classify each location to the respective cluster of municipalities in Nicosia. Thus, for each cluster of locations in Nicosia we will have a respective cluster of similar locations in Limassol.

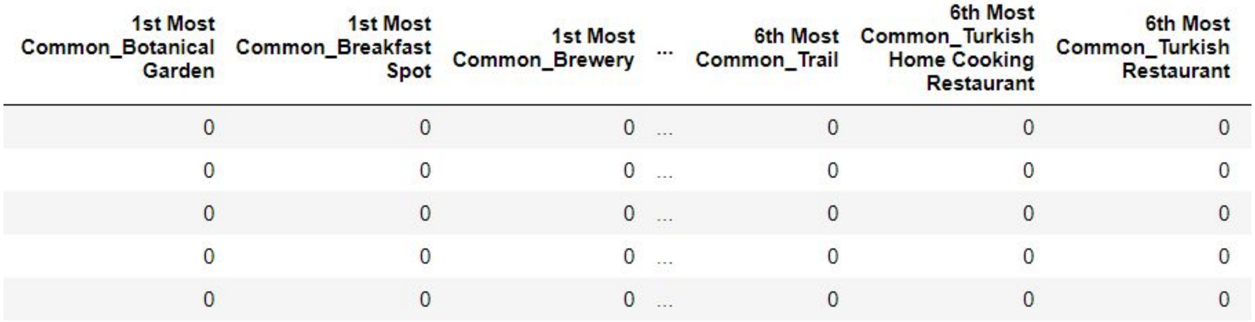
4.2 Preparing the Data for use in Models

In order to prepare our data for training the types of models specified we needed to make changes when it comes to the categorical data.

The categorical data that is going to be used, this being the most common types of venues, must be turned into numerical data that will allow the models to find insights and make predictions.

To achieve this a table with the most common categories of venues per location was one-hot encoded.

The following is an extract of that table:

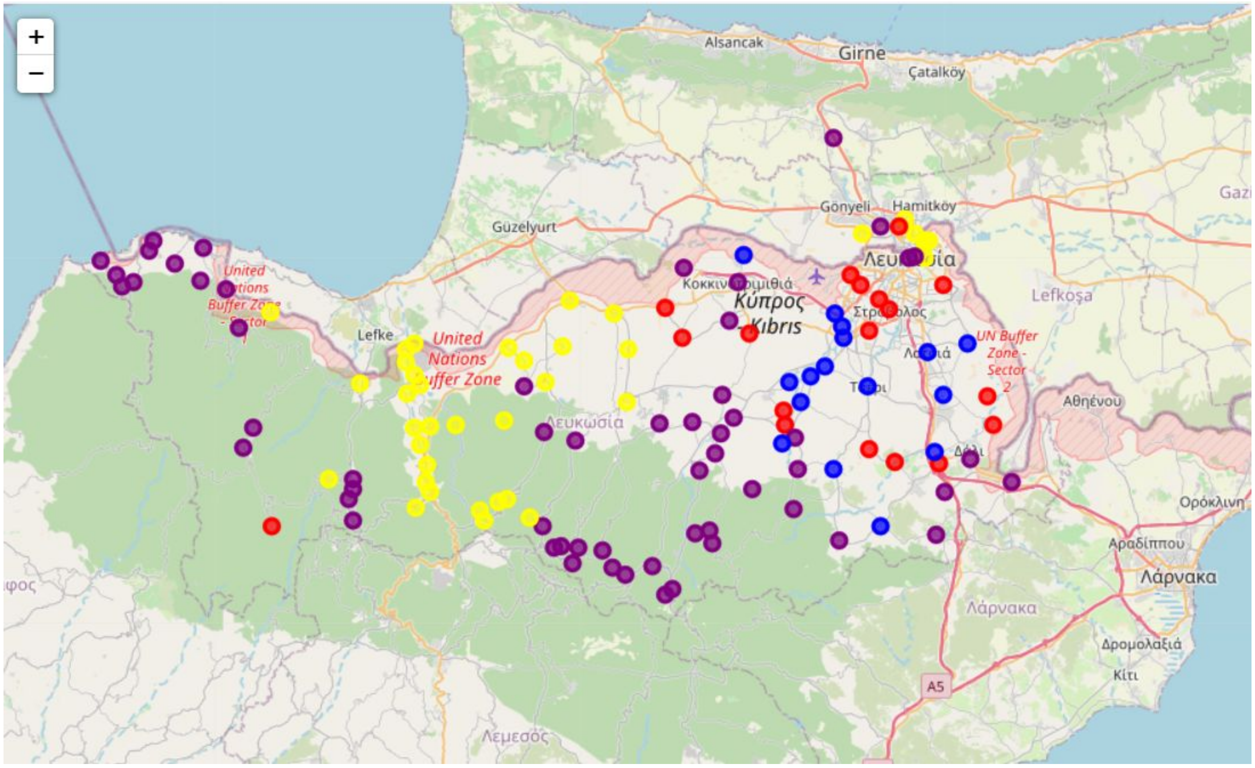


1st Most Common_Botanical Garden	1st Most Common_Breakfast Spot	1st Most Common_Brewery	...	6th Most Common_Trail	6th Most Common_Turkish Home Cooking Restaurant	6th Most Common_Turkish Restaurant
0	0	0	...	0	0	0
0	0	0	...	0	0	0
0	0	0	...	0	0	0
0	0	0	...	0	0	0
0	0	0	...	0	0	0

4.3 Clustering Algorithm

When it comes to the clustering part of the process stated in **4.1** the K-Means Clustering Algorithm was picked as the best one to be used. The reason for that is that based on the nature of the data available and the information needed the K-Means Clustering Algorithm was the most appropriate. Other algorithms that were considered but not picked due to their incompatibility with the problem include Density-Based Clustering Algorithms and Hierarchical Clustering Algorithms.

Using the K-Means Clustering Algorithm on the data considering locations in Nicosia produced the following results (visualised on a map format):



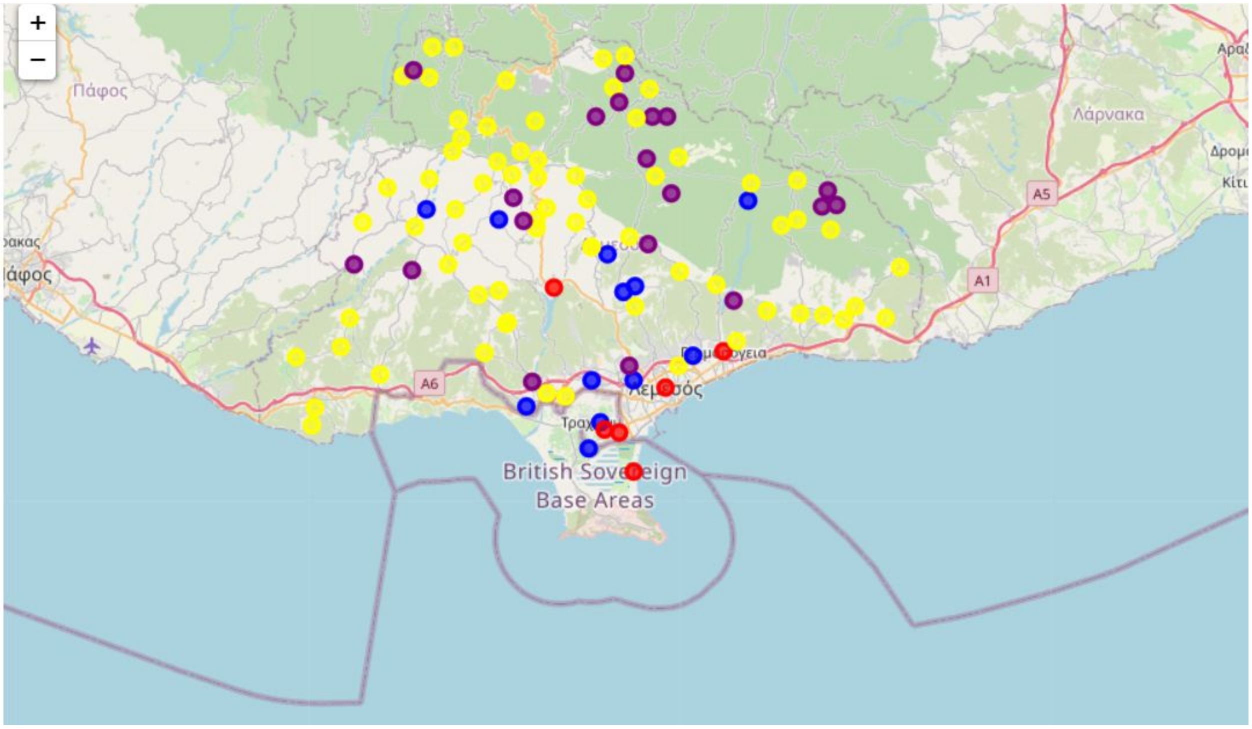
Consider the following:

- **Purple Locations**: tend to be on more mountainous land.
- **Yellow Locations**: tend to be following highway roads.
- **Blue and Red Locations**: both tend to be in more urban areas and seem to be subclusters of the same parent cluster, their main difference being that red points tend to be closer to bodies of freshwater.

4.4 Classification Algorithm

When it comes to the classification part of the process stated in **4.1** the K-Nearest Neighbor Classification Algorithm was picked as the best one to be used. Other Algorithms that were considered include Logistic Regression, Decision Trees and Support Vector Machine but were deemed incompatible with the data's nature and the approach we decided to take on solving the problem.

Using the K-Nearest Neighbor Classification Algorithm on the data considering locations in Limassol produced the following results (visualised in a map format):



5. Conclusions

We have clustered our data on Nicosia municipalities using a k-means clustering algorithm, and then the municipalities of Limassol were classified to each cluster by using a k-nearest neighbor algorithm.

Our final results represent which municipalities in Limassol are better to move to for someone living in the District of Nicosia, based on in what municipality of Nicosia they live.