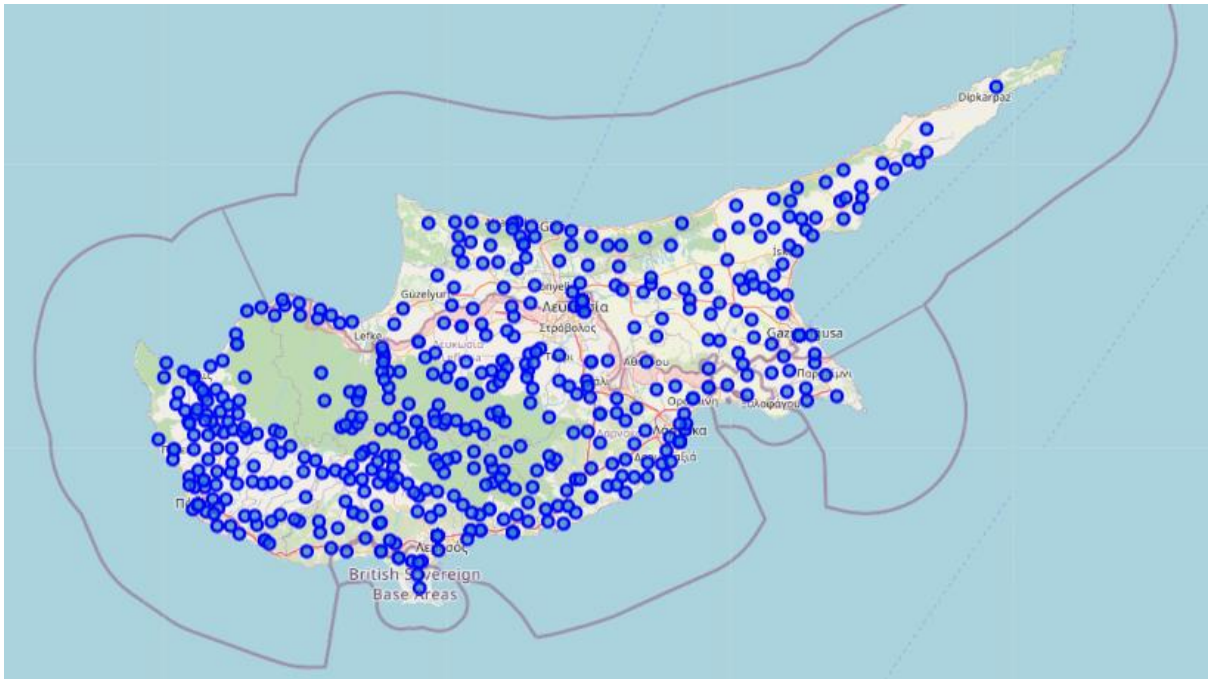
Coursera IBM data science capstone project:

Determine the best venues by area in Cyprus



Andreas Antoniou

https://github.com/AndreasAntoniou

Contents

1. Introduction

Cyprus is one of the 27 members of the European Union. The island currently has around 1 million population, that is distributed in 4 main cities and many villages, with individuals having minimum knowledge of the internet. In addition, due to the old infrastructure used by the country, it is very difficult to acquire relevant data for a business problem and further the accuracy of the data might be questioned. With the two aforementioned issues, the Cypriot community struggles to assess issues using analytics and statistical models.

With the use of recent software, such as Google maps APIs, 2GIS and Foursquare, it is more convenient to generate specific geographical information that will provide the opportunity for further analysis using analytical tools and software.

To that instance, this project will aim to gather all the areas in Cyprus, including main cities, as well as villages, and deploy a visualization of the island, providing crucial information such as the most visited venues of the whole island.

2. Aim and objectives

The aim of this project is to generate data through the utilization of methods and software that will be used to provide analytical insights for geolocation-based data.

1. Locate available data that can be used as a basis for geolocation
2. Using APIs such as Google Maps API, 2GIS API or Foursquare, to generate more specific data such as longitude and latitude of each village and other areas in Cyprus.
3. Develop visualization of geographical data that will be used as complementary findings to stakeholders for better understanding
4. Identify patterns in customer preferences and provide suggestions to potential stakeholders regarding the findings

3. Methodology

For the development of this project, the following data are required:

- Basic geolocation data for Cyprus (Areas, Regions)
- Specific geolocation
- Acquire statistical data regarding customer preferences

As already mentioned, due to the old infrastructure of Cyprus and the Cypriot network, including the old systems used by the government and the limited knowledge of the population living in rural areas, there are limited data available for geolocation. To tackle this issue 4-steps will be followed to acquire and generate geolocation data. Firstly, data from the Cypriot post office, regarding postcodes and areas will be downloaded. The list is comprehensive and mentions most of the areas (neighbourhoods) in Cyprus, postcodes and their districts (region). Despite the comprehensive list, geo-coordinates – latitude and longitude – are missing from the

dataset. Since it is not possible to generate a geolocation map without these data, an algorithm should be used to acquire geo-coordinates from the data acquired through the Cypriot post office. One way to do this, is to use the google maps reverse geolocation API, that returns latitude and longitude based on the postcode and area name provided. To acquire access to the Google API, it is required to register for free to receive an API key. The usage of the API is free as long as the calls are made within a certain limit. While the API is relatively precise, on certain occasions there are areas that are not recognised, while on others the geo-coordinates returned are not representative to the actual areas located in Cyprus. Consequently, after using the reverse geolocation API, it is crucial to clean the dataset from missing values and outliers. All observations that have null longitude and/or latitude, will be removed from the dataset by using a simple python function. Further, investigation regarding incorrect coordinates will be first observed through geographical visualization of the observations and the Cypriot map using Folium. After cleansing the dataset, Foursquare API is used to acquire nearby venues in each of the areas of Cyprus. To acquire information from the Foursquare API, free registration along with an account ID and API are required. Using Foursquare, the names and unique categories that exist in each area are acquired. Further, after acquiring the venue data, k-means clustering is applied to the dataset to observe the difference cluster and identify patterns within the data. After the clustering is performed, a visualization is developed using the results of the cluster as well as a map of Cyprus. The observations are coloured based on the cluster that they belong in. Further analysis is performed using frequency of the 1st most common type of venue in each area. Lastly, for visualization purposes and to provide a more robust demonstration of the results, the WorldCloud library is used that reveals the

3.1 Tools and data used in methodology

The most specific tools and their URL that are used in methodology are mentioned below.

1. **Basic geolocation data for Cyprus**

   Source: https://www.cypruspost.post/uploads/postal-codes/dd934fd71a.xlsx

   The file has the Greek and English translations of Area/Municipalities, Districts and Postcodes.

2. **Specific geolocation (Postcodes, Longitude, Latitude)**

   Source:

   https://developers.google.com/maps/documentation/javascript/examples/geocoding-reverse

   This data will be generated using the reverse geocoding algorithm of the Google API. Specifically, Postcodes will be set as an input parameter and Longitude with Latitude will be the return parameter. Afterwards the two datasets will be merged.

3. **Statistical data regarding customer preferences and spending (if available)**

   Source:

   https://www.mof.gov.cy/mof/cystat/statistics.nsf/publications_en/publications_en?OpenForm&OpenView&RestrictToCategory=14&SrcTp=1&Category=0&Subject=1&SubSubject=4&subsubtext=4&

4. **Data regarding venues location and preference**

   Source:

https://developer.foursquare.com/

Using the Foursquare places API, generate a list of places in each area of Cyprus

3.2 General libraries that are used to develop the project

Pandas: For creating and manipulating Dataframes.

Folium: Python visualization library would be used to visualize the neighbourhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

XML: To separate data from presentation and XML stores data in plain text format.

Geocoder: To retrieve Location Data.

Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.

Google API: To acquire geolocation data

Numpy: Handle data in a vectorized manner

Json_normalize: transform json file to pandas Dataframe

Wordcloud: package to develop a visualization based on the frequency of words

4.0 Results

4.1 Observations used

The total areas in Cyprus acquired through the post office were 729. After using the Google API for reverse geolocation, 666 observations returned a longitude and latitude, while 62 returned 0 (Null) values. The 62 observations with null values were removed, while the rest were visualized geographically on a map. As observed in figure 1 below, there were certain observations that returned an inaccurate geo-location. The red circle below shows the island of Cyprus, indicating that the rest of the observations are considered outliers.

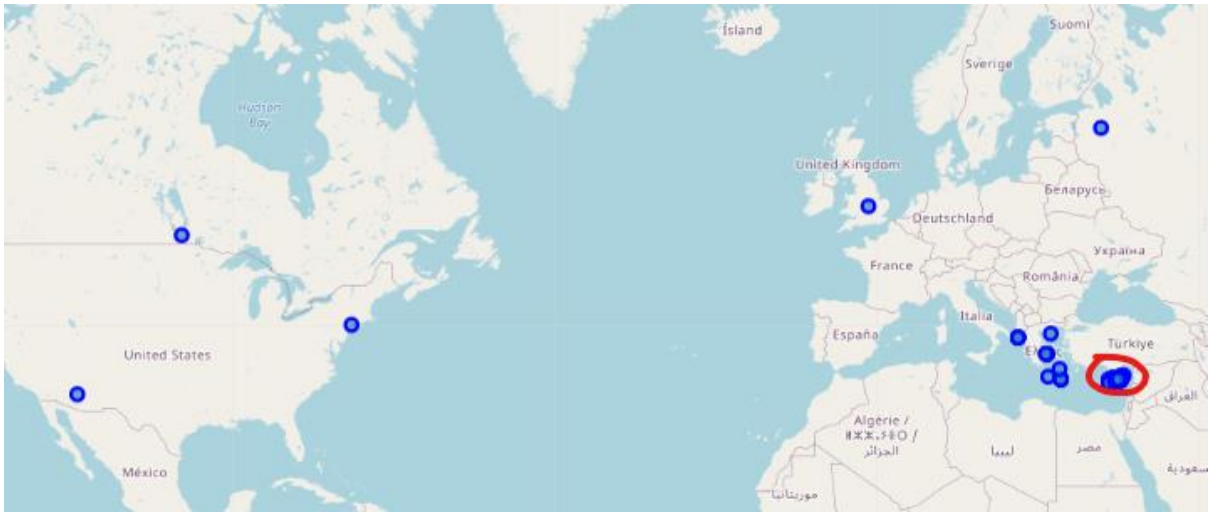


Figure 1 Visualization of observations on a geographical map using Folium

To solve this issue, summary statistics and histogram visualizations were used.

	Lat	Long
Count	666	666
Mean	35	32
Std	1.39	8.82
Min	33.41	-111.831
25%	34.82	32.76
50%	34.99	33.15
75%	35.19	33.42
max	58.52	34.40

The table above provides summary statistics information regarding Lat and Long. It is observed that the distance between the 75% to maximum of Lat and 25% to minimum of Long are way higher than

the standard deviation. Therefore, it can be assumed that the outliers exist in the two aforementioned distances. After removing 2 standard deviations from the mean (median was considered more accurate to be used in this case), most of the outliers were removed. Figure 2 below shows the outliers that were removed.

Figure 2 shows the visualization of the outliers on a map

While this method removed most of the outliers, some of the outliers still existed in the dataset. Since these outliers were close to the long/lat of Cyprus and could not be removed by basic statistical methods, through the visualization, the longitude and latitude of the furthest observations within Cyprus were gathered and anything further than the geo-coordinates of those observations were removed. These method was very effective since all of the outliers were removed providing a map with observations only on the island of Cyprus.

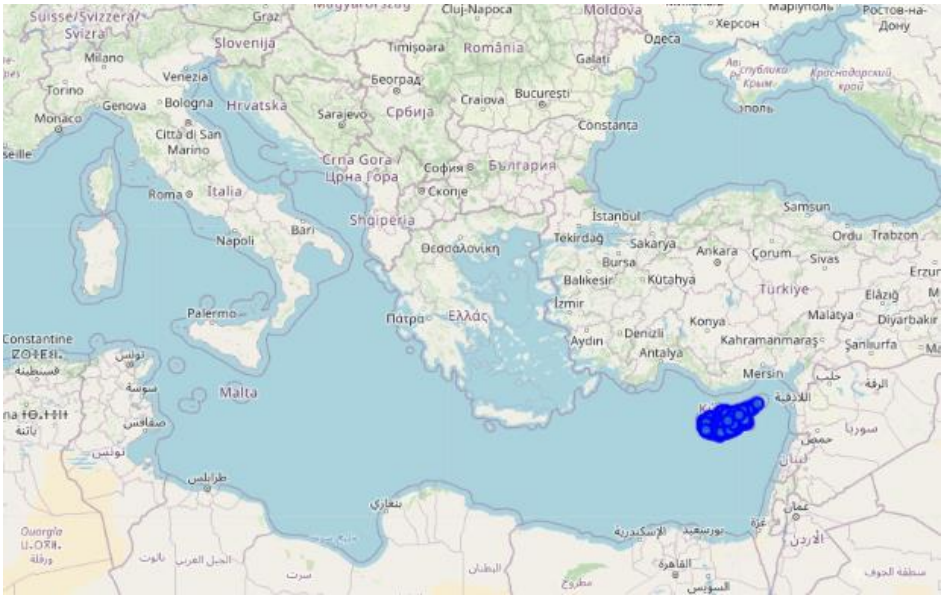


Figure 3 shows a visualization of the cleansed dataset

After removing all of the outliers, 657 areas remained, and 9 were removed.

4.2 Venus results

To retrieve the venues for each area, the Foursquare API was used. Foursquare API is a city guide and an application that provides information regarding venues, including name, reviews and type of venue. Foursquare API was used to generate venues from each area in the dataframe. For the 657 areas, a total of 2290 venues were retrieved that belonged in a total of 244 unique categories.

4.3 Cluster results

Results regarding this analysis were crucial. Firstly, cluster analysis revealed the 1st most common venue for each area and their results were documented as shown below.

Clusters	Area and frequency
1	Café: 11
2	Coffee shop: 37/ Leather goods store: 29 / Sports club: 29 / Mountain: 14
3	Kafenio: 56
4	Beer Store: 15
5	Greek restaurant: 17

While the 1st most common venue of the three first clusters are similar, certain assumptions can be made based on the results. In fact, Café, Coffee Shop and Kafenio can be considered as similar shops, although *"Kafenio"* are venues that will most likely include traditional Cypriot coffee, while *"Café"* are venues that are considered to be a place that will also attract tourists and do not necessarily adopt a traditional kind of coffee. Further, *"coffee shops"* are venues that are likely to be a combination of "*Kafenio"* and "*Café"* and include both traditional as well as exotic coffees. The differences between these categories can be based on the difference of the areas. For example, traditional coffee venues, *"Kafenio"*, are most likely to exist in rural areas where the population is closest to the traditions whereas *"Coffee shops"* could appear in tourist areas. Further analysis regarding the frequency of clusters 1 -3 will be observed through the visualization.

Regarding Cluster 4, it can be observed that *"Beer stores"* have the highest frequency, while in cluster 5, *"Greek restaurant"* have the highest frequency with 17 occurrences.

4.4 Wordcloud Visualization

Using wordcloud library, a visualization was developed regarding the frequency of restaurants in Cyprus. As observed in figure 4 below, words *"Kafeneio"*, *"Greek Restaurant"*, *"Coffee shop"*, *"Restaurant"*, *"Good Store"*, are the words that have the highest frequencies. Further understanding of the frequency can be observed using a folium visualization map with the clusters above.

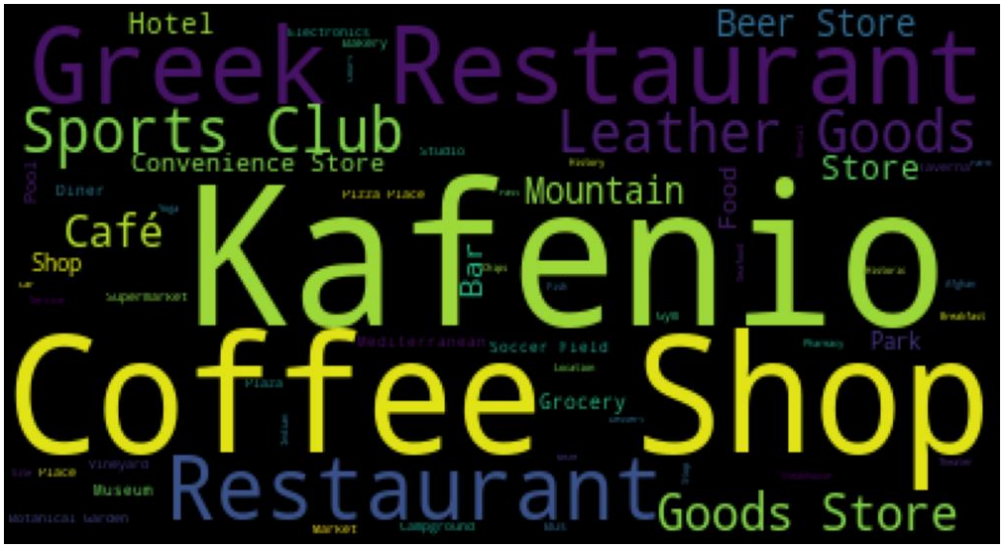


Figure 4 shows a visualization of word frequencies of venues in Cyprus using wordcloud

4.5 Visualization of clusters

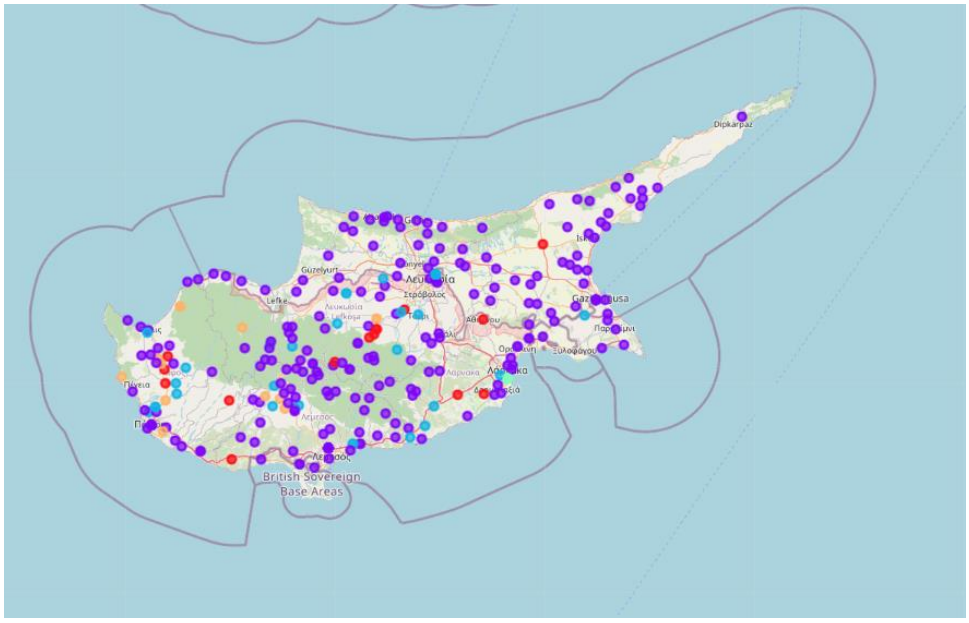


Figure 5 shows a visualization of the clusters of venues in Cyprus

With the visualization of the map, expert judgment is used to determine the venues and the reason of their existence in certain areas. As observed in figure 5, Cluster 2, with coffee shops, appears in all areas of Cyprus irrespectively. Cluster 1 (red) with the venue category *"Café"* appears mostly in tourist areas that are located out of urban cities. Cluster 2 (Teal) with the most frequent venue category *"Kafeneio"* are located in rural areas, since they are and have been a traditional place for individuals to enjoy Cypriot coffee. Aside from coffee, Greek restaurants are also very popular in Cyprus, with 17 areas having the category of venue as more frequent. Greek restaurants appear to have a higher frequency in rural areas.

Coffee shops are attractive to all segments and can generate profit in all of the areas. Venues with the category of Café

Cluster	Colour
1	Red
2	Purple
3	Teal
4	Light Green/Blue
5	Yellow

5.0 Recommendations and limitations

While this report provides crucial information regarding the frequency of venues in Cyprus and in specific areas, there are other factors that should be considered before opening a venue. Specifically, consumer expenditure, venue rating and review, income of residents and rent are some of the variables that should be estimated to provide further information regarding a possible investment.

6.0 Conclusion

In conclusion, coffees are a very popular destination for Cypriots. It is observed that coffee shops are frequent in various locations, including tourists, urban and rural areas, allowing the investor to also consider other variables before establishing such a venue. In addition, for variances of venues that sell coffees, such as “Kafenio” and “Café”, it is possible for the investor to target segment that love traditional or segments that love international tastes, respectively. Aside coffee shops, an investor could also consider Greek restaurants as they are popular, although only exist in rural areas. Importantly, Greek restaurants are a place of preference for both tourists and local population, allowing the venue to target a variety of segmentations.