Greek Travel

A REPORT ON THE GREEK ISLANDS’ POPULARITY

Zacharias Detorakis | The Battle of Neighborhoods (Week 2) | March 25, 2019

INTRODUCTION

I work at a tourist office that provides travel packages to people who want to visit Greece for the summer. Lately we've experienced a higher demand for people wanting to visit the Greek islands following Greece's economic crash and the drop in the prices of accommodations. To address this need more effectively the travel agency has decided to revisit the way the packages to the islands have been designed and to create custom versions based on the customer's needs. These packages would include multiple islands per package that the client can visit during their stay in Greece.

Greece has a great number of islands cited as between 166 and 227. However, for the purposes of this analysis the focus is on the major 40 islands that are the main touristic destinations and can be accessed via boat and/or plane. Smaller islands would require special arrangements for travel and are therefore not in scope of the analysis.

The problem we are facing is to try and identify groups of islands that are either:

- **very touristic** and therefore appeal to a younger demographic looking for fun and partying or
- **less touristic** and appeal to an older demographic looking for relaxation.

The question we need to answer is *how can we find groups of islands that are similar in terms of their popularity with the tourists*

DATA

In order to find the answer to the question of clustering the islands by popularity we first need to define the universe of the data (i.e. which islands to select). A list of the islands that we can use can be found online on the Wikipedia page:

https://simple.wikipedia.org/wiki/List_of_Greek_islands

This is a list of the 40 islands that we consider for their touristic popularity. As mentioned, the agency has decided to limit the analysis to the islands that are easily accessibly which is why smaller islands are not included.

Following is the sample of the data available for the islands

Rank by size	Island name	Area (km²)
1	Crete	8.336
2	Euboea	3.655
3	Lesbos	1.630
4	Rhodes	1.398

Table 1: Sample data for the islands from the Wikipedia page

Next we will be looking at the Foursquare database to identify the venues on these islands. We will be doing that based on the island name for two reasons:

- The latitude and longitude are not readily available. This obstacle could be overridden as we found those data points on a separate dataset that we could merge
- The islands are of different sizes and in close proximity to one another therefore defining a radius around a central location would yield incorrect results. Often times the coordinates of an island are those of its main town ('Chora') and therefore are located near the port and not in the centre of the islands

We will set the limit to a very high value in order to include all the venues on the island rather than only the top x. We will review the request responses for any inconsistencies or bad data and cleanse the dataset. Finally, we will extract the features we think are needed for the clustering of the islands and perform the ML algorithm to group them together. These steps are referenced better in the following methodology section.

METHODOLOGY

Step 1: Get the data from the Wikipedia page and identify a normalization factor

As shown in Table 1 there the islands in scope differ a lot in size from 8k (km²) to 64 (km²). Therefore, it decided to calculate a normalization factor in order to take into account the difference in size. This is based on the assumption that islands of a bigger size are expected to have more venues. A sample of the normalization factor is shown in the following table

Rank by size	Island name	Area (km^2)	NormFactor
1	Crete	8.336	1.000000
2	Euboea	3.655	0.438460
3	Lesbos	1.630	0.195537
4	Rhodes	1.398	0.167706
…	…	…	…
39	Kasos	66	0.007917
40	Alonnisos	64	0.007678

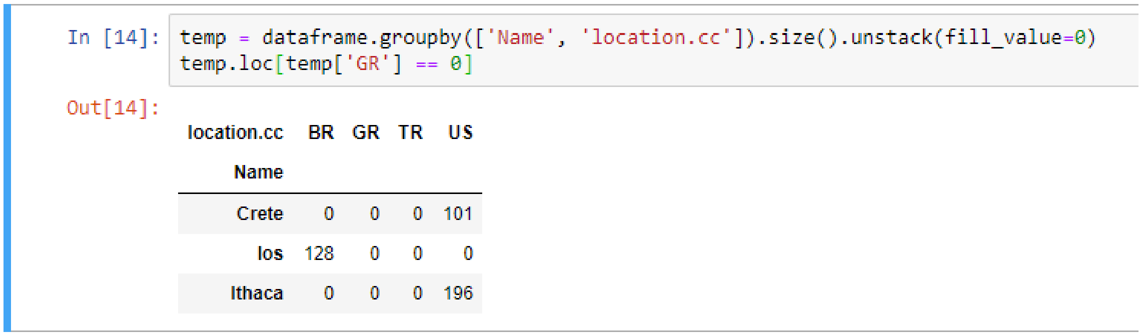
Table 2: Sample data for the islands with the normalization factor

The normalization factor is calculated by dividing the size of each island in square km by the size of the biggest one. In this case the largest island in Greece is Crete. If we inspect the tail of that normalization table looks like the factor becomes to small so we will need to make sure that the normalization of the features later on is not heavily impacted by this.

Step two: Request Foursquare data

As discussed the request of the venues from the Foursquare API is not done via latitude and longitude but by name.

By inspecting the data retrieved from the API it was identified that for a few islands the country is not GR therefore Foursquare API has not returned the correct results for the venues. For those islands we either need to find a different name with which they are represented in the Foursquare API or exclude them from our analysis.



```
In [14]: temp = dataframe.groupby(['Name', 'location.cc']).size().unstack(fill_value=0)
         temp.loc[temp['GR'] == 0]
```

```
Out[14]:
         location.cc   BR  GR  TR   US
                Name
               Crete    0   0   0  101
                 Ios  128   0   0    0
              Ithaca    0   0   0  196
```

After a few trials with the API we've found the following replacements are necessary to rectify the data:

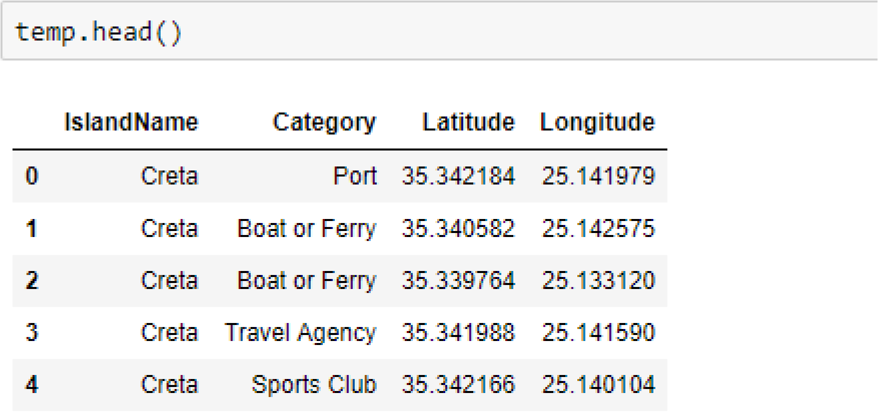
- Crete --> Creta
- Ithaca -->Ithaka
- Ios --> Íos

We've found the new values by manually inserting the latitude and longitude in the Foursquare API and observe some of the value returned. Therefore, we amended the values (i.e. names) of the islands and submitted the requests once again.

After replacing the values as mentioned above we call the Foursquare API once again. Following that we reviewed the results once again we found that there are is only one result outside of Greece.

We then selected only the category column and the columns that have the island name and the location country, latitude and longitude. The country is selected because there is one row that has a country other than GR which needs to be removed.

After dropping the non-relevant rows the dataset looks as follows:



```
temp.head()
```

	IslandName	Category	Latitude	Longitude
0	Creta	Port	35.342184	25.141979
1	Creta	Boat or Ferry	35.340582	25.142575
2	Creta	Boat or Ferry	35.339764	25.133120
3	Creta	Travel Agency	35.341988	25.141590
4	Creta	Sports Club	35.342166	25.140104

With the category in place we review the unique values with the ultimate goal to create our feature list for each island which will be whether they have port/airport, the number of hotels and the number of bars/coffee shops and restaurants they have.

After carefully reviewing the different categories in the Foursquare data we decided that the following belong to accommodation and food entertainment:

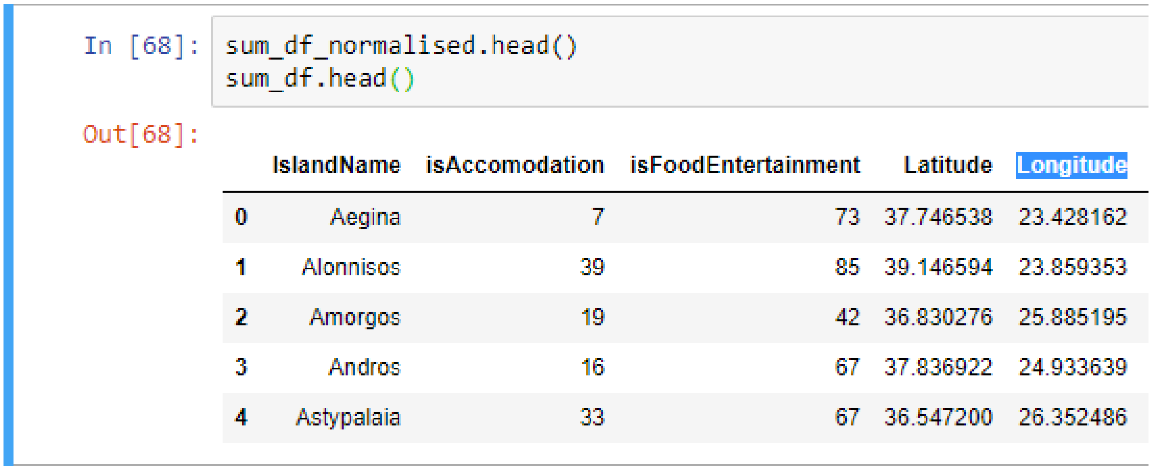
- Accomodation: 'Hotel','Bed & Breakfast','Vacation Rental','Resort','Residential Building (Apartment / Condo)','Spa','Hotel Bar','Hostel','Motel'
- Food Entertainment: 'Greek Restaurant','CafÃ©','Bar','Bakery','Cocktail Bar','Dessert Shop','Coffee Shop','Nightclub','Souvlaki Shop','Pizza Place','Taverna','Fast Food Restaurant','Restaurant','Grocery Store','Snack Place','Seafood Restaurant','Italian Restaurant','Kafenio','Mediterranean Restaurant','Meze Restaurant','Fish Taverna','Burger Joint','Breakfast Spot','Sandwich Place','Candy Store','Creperie','Food & Drink Shop','Frozen Yogurt Shop','Ouzeri','Bistro','Deli / Bodega','Internet Cafe','Food','Modern Greek Restaurant','Bougatsa Shop','Donut Shop','Pub','Steakhouse','Grilled Meat Restaurant','Cafeteria','Chinese Restaurant','Cupcake Shop','Falafel Restaurant','Juice Bar','Cretan Restaurant','Diner','Sushi Restaurant','Karaoke Bar','Fish & Chips Shop','Paella Restaurant','Piano Bar','Turkish Restaurant'

Finally, we grouped the rows by island creating the following features

Feature	**Description**	**Aggregation Method**
isAccomodation	This feature is a measure of how many hotels, bed&breakfast etc are available for accommodation on the island	sum
isFoodEntertainment	This feature is a measure of how many restaurants, coffee shops, bars, etc. are available on the island	sum
Latitude	The mean of the coordinates of the venues on the island. This was included in order to group together islands that are similar but also in close proximity so that one can easily 'hop' from one to the other as part of the same vacation package	mean
Longitude	Same as above	mean

Table 3: List of features

Following is a sample of how the dataset looks like



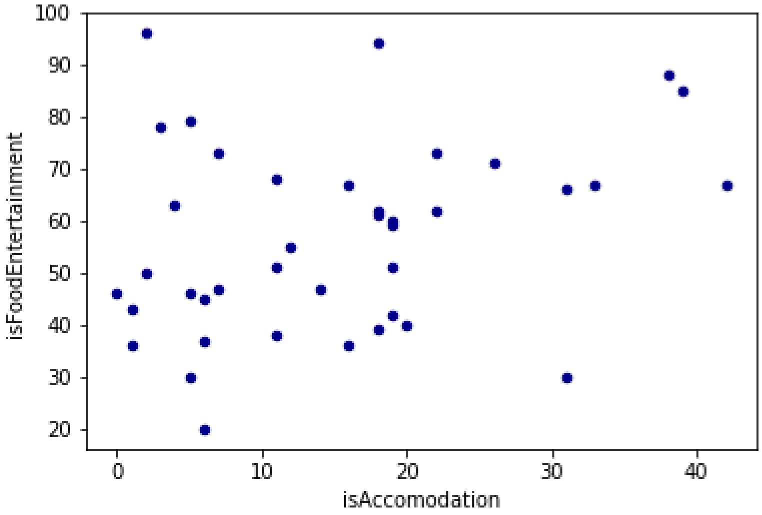
```
sum_df_normalised.head()
sum_df.head()
```

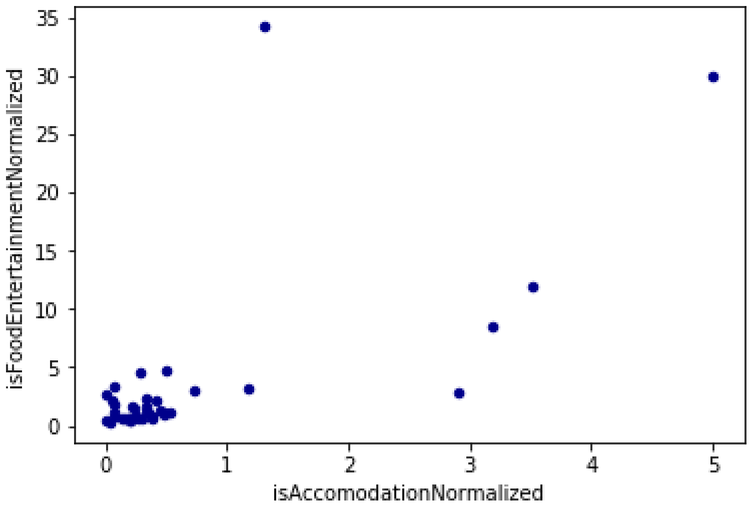
	IslandName	isAccomodation	isFoodEntertainment	Latitude	Longitude
0	Aegina	7	73	37.746538	23.428162
1	Alonnisos	39	85	39.146594	23.859353
2	Amorgos	19	42	36.830276	25.885195
3	Andros	16	67	37.836922	24.933639
4	Astypalaia	33	67	36.547200	26.352486

Step 3: Overview of the data

We then decided to view the data on a 2 by 2 plane base on the two features of accommodation and entertainment both in their normal as well as in their normalized versions

```
Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x2519ebc04e0>
```





It is evident from the figures above that the normalization has distorted the universe making the size a prominent factor and grouping all the instances in the bottom left corner. A logical explanation would be that regardless of the size of the island the venues are all concentrated is small areas in the island therefore the size is not as important. We therefore decided to not considered the normalized versions of the features for the clustering.

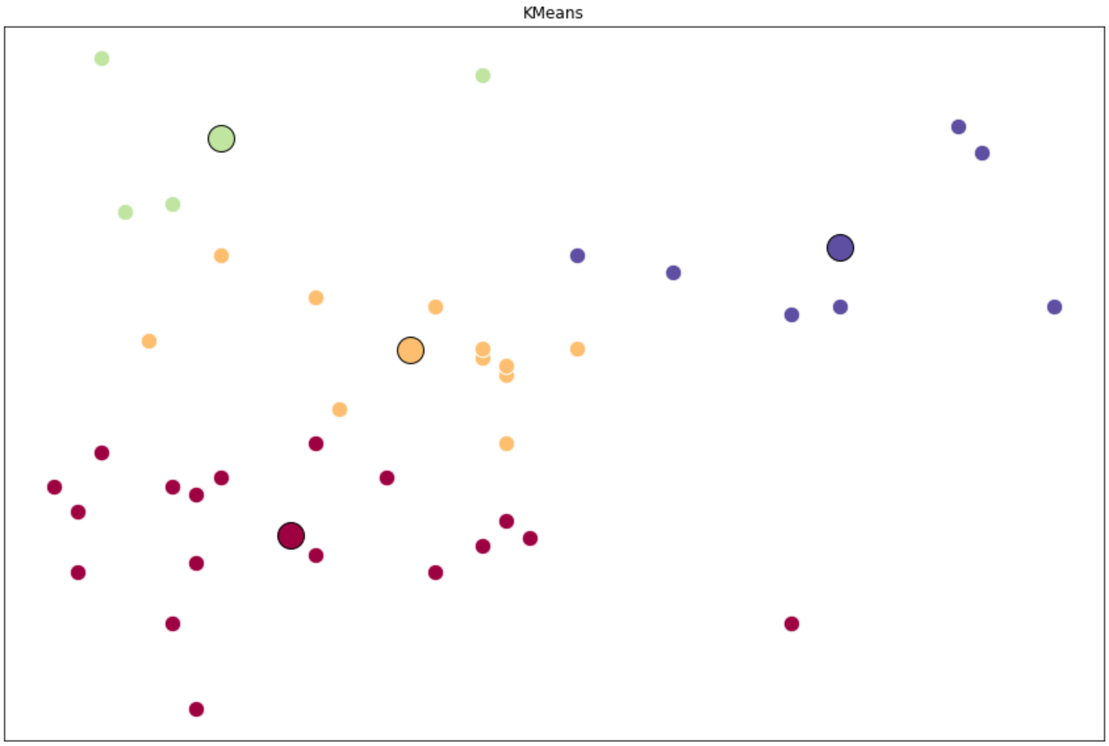
RESULTS

We decided to use a clustering algorithm since the dataset that we have is with unsupervised data. The clustering algorithm we used is the K-means algorithm with 4 clusters. After repeating the algorithm for multiple iterations (i.e. 12) the results are shown below.

IslandName	Cluster
Amorgos	0
Chios	0
Creta	0
Ikaria	0
Karpathos	0
Kasos	0
Kefalonia	0
Kithira	0
Lefkada	0
Limnos	0
Milos	0
Mykonos	0
Naxos	0
Salamis	0
Samos	0
Skopelos	0
Skyros	0
Zakynthos	0
Aegina	1
Andros	1
Corfu	1
Ithaka	1
Kythnos	1
Lesbos	1
Paros	1
Rhodes	1
Samothrace	1
Sifnos	1
Syros	1
Euboea	2
Kalymnos	2
Kos	2
Tinos	2
Alonnisos	3
Astypalaia	3

IslandName	Cluster
Kea	3
Serifos	3
Thasos	3
Thira (Santorini)	3
Ã�os	3

A visual representation of the results based on their key features is also shown below:



DISCUSSION

As illustrated above there is no clear separation between the islands (i.e. the datapoints are not concentrated around the centroids). However, there are clear areas of touristic vs less touristic islands. For instance, it is clear that islands in the first cluster 0 (purple) are less touristic than those in the cluster 3 (blue)

The centroids of the four clusters are listed below:

```
-   [ 9.94444444, 40.16666667, 37.41542023, 24.57764965],
-   [15.        , 61.90909091, 38.05725103, 24.37859527],
-   [ 7.        , 86.75      , 37.46814041, 25.86970018],
-   [33.        , 73.85714286, 37.77395273, 24.92474586]]
```

We could perhaps overlay the result with data for available boats / flights between islands of the same cluster to come up with the packages to offer to the customers

CONCLUSION

The analysis conducted as part of this exercise has revealed groups of clusters that share common characteristics in term of how popular they are with the tourists and how the infrastructure exists to support those tourists.

Based on the results provided by an unsupervised clustering algorithm we've managed to identify islands that can be grouped together to create holiday packages for customers that come to the travel agency.

Further analysis should be conducted to take into account flights and boat trips between these islands in order to better understand how to structure the packages and provide our customers with the best possible service.