Capstone Project

Clustering mountains and cities in Greece

Introduction/Business Problem

Greece attracts a lot of tourists for hiking cause of the number and beauty of her nature.

The aim of this project is to recommend city destinations for mountain lovers in Greece.

The first task would be to segment the cities based on their geographic distance (on km) from every mountain and cluster them based on that and as second step I will make another segmentation based on the similarity of the cities.

After that I will merge them to filter the cities that are both similar and close to same montains.

Finally I will find the closest 3 mountains , based on their distance in km for every city in the final merged cluster.

I will make use of our data science tools to analyse data and focus on the relationship of cities and mountains in Greece.

Data

We will need two types of data.

- Wikipedia data for mountain informations like height , regional unit and coordinates.
- Wikipedia data for cities informations like name , coordinates
- Foursquare data about venues on every city in Greece.

I scrape the wikipedia data from https://en.wikipedia.org/wiki/List_of_mountains_in_Greece .

I used **Scrapy** an open source and collaborative framework for extracting the data from websites
I end up with a dataset that contains the **peak** , the **height** , the **mountain range** , the **coordinates** and the **regional unit** the mountain belongs.

I follow a similar approach for the **region** / **cities** . I got the **name** and the **coordinates** of the cities which is usefull to use this dataset in collaboration with the previous.
Finally I will use foursquare data about the near cities from each mountain to find venues in a given radius, to separate them from each other. I will take the top 10 venues of every city. We will see at the end that cities having some specific similarities like islands cluster together.

Methodology

First I read our two csv files that we create in the previous step

I plot my data to understand the geographical distribution.

Here we have with red dots our cities and with blue the mountains

In this project as I mentioned previously we gonna cluster our data with k means algorithm based on two criteria. First based on the distance of every mountain with every city in kilometer using distance formulas and then based on city venues similarities that i took from foursquare API.

For distance based clustering I calculate the distance between the mountains and cities passing the coordinates to Haversine formula https://en.wikipedia.org/wiki/Haversine_formula . I run a loop and parse mountain location one by one.

Methodology (2)

I continue with Foursquare API to explore the cities in Greece . Because I explore whole cities I put as radius limit 15 kilometer .

After the call of the API I have my new Dataframe with the top 100 venues of every city.

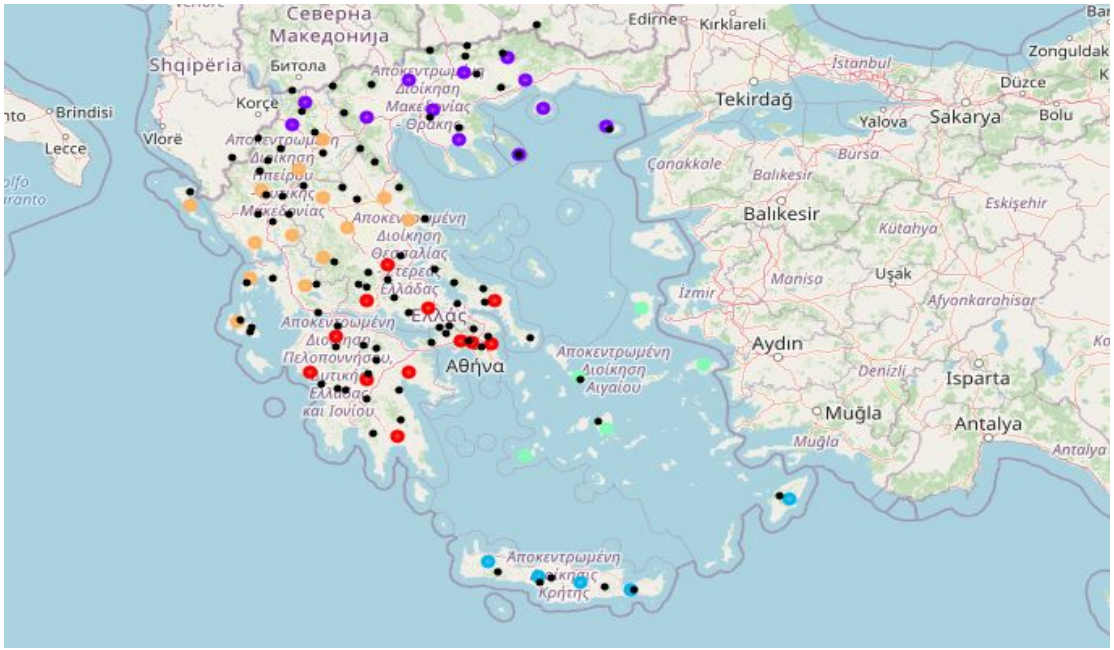
I group them based on the city name , analyse every city by making every venue type as new column and take the mean to execute my second clustering.

After the execution of this step we have 10 sub-Dataframe , 5 from distance based and 5 from similarity based and I need to merge them to end up with cities that have both venue similarity and geographical similarity.
I create two list that contain the 10 Dataframes. This step help me with the Dataframes processing that occur with iterating .

As last step remain to find the 3 mountains that are the closest to every city in the same Dataframe. I create a row that contains the mean from the distance of all mountains of all cities and take the 3 minimum values. I then pass it to list
.
With this list i filtered the Dataframes take the final ones that contain the 3 closest mountains on every team of cities.

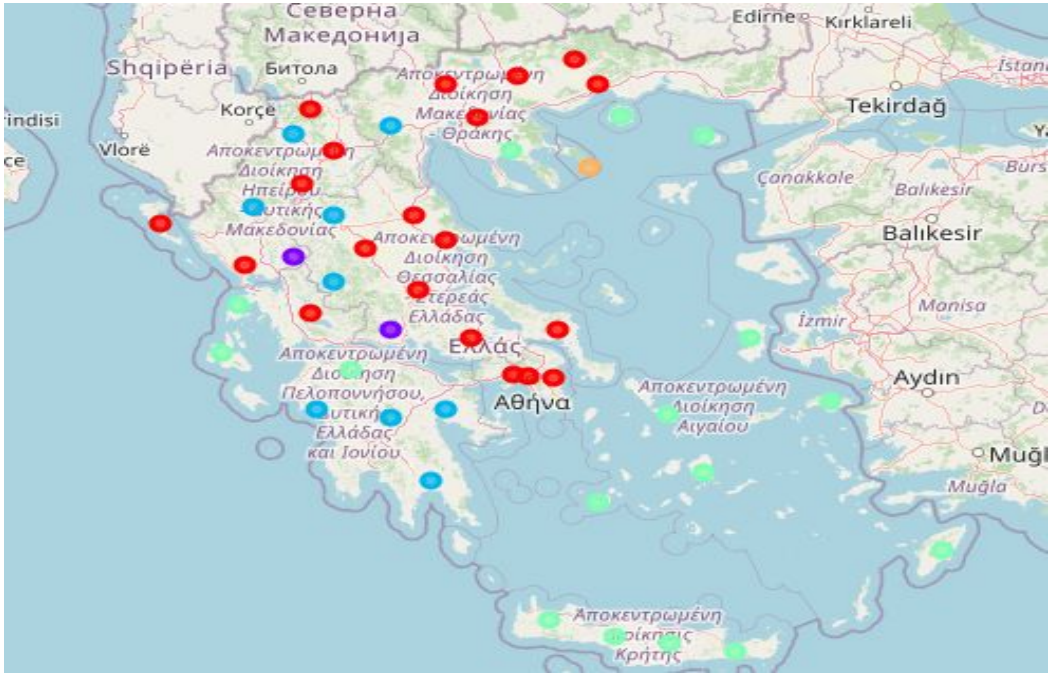
Results

Here is the map of Greece and the clustered cities based on the distance they have from every mountain. With colored dots its the every clustered unit and with black dots the mountains just to have an overview.



Results (2)

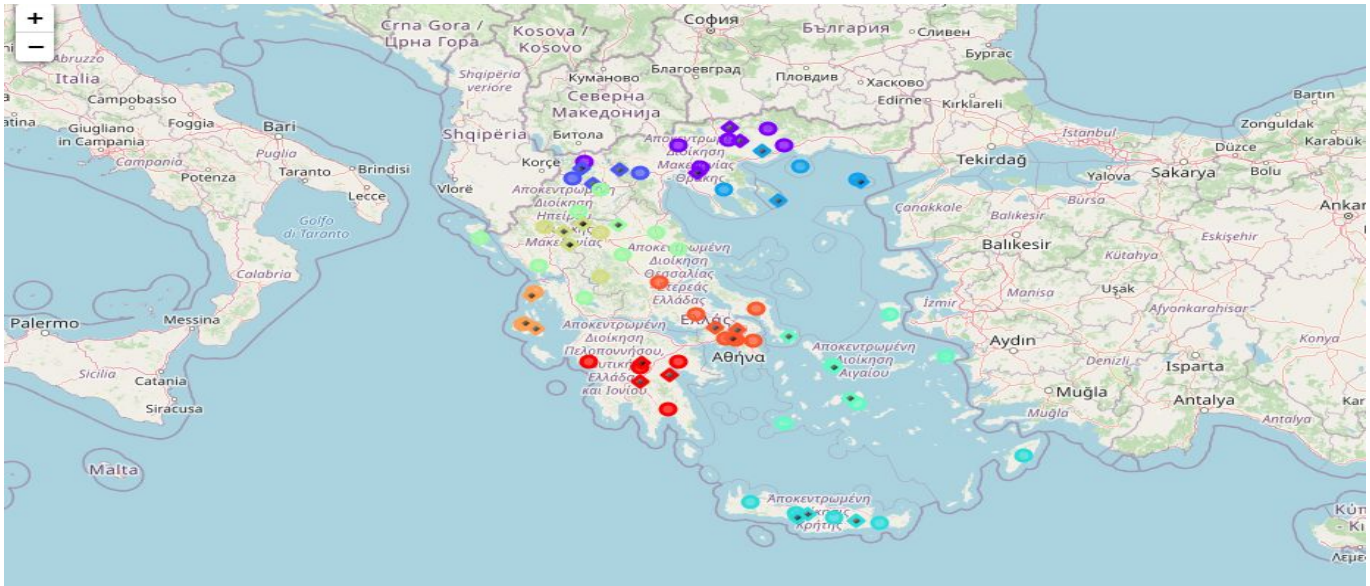
Here is the map of Greece and the clustered cities based on venue similarity.



Results (3)

Finally I reach my goal to have both geographical and also venue similar cities. You will see that the results is much less than the original because there were some cities that didnt meet the two criteria.

With circle I represent the cities and with square the three mountains that is on every area. Every mountain and city that belongs in the final cluster have the same color.



Discussion

We see that places that have some strong characteristics like islands is in the same category because they have some venues in common like beaches and ports.

As further analysis we can use the heights of the mountains to group them in 3 categories like under 1000 meters , between 1000 and 1500 , over 1500 etc and recommend one of each kind based on the same criteria.

Conclusion

In this report I make an explanation of this project , to help understand easier both the code as its functionality.

Its a project to have as goal to make recommendations to the mountain lovers tourists with respect to the cities themselfs.

You can find the whole project here :
https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/0e808e6b-b1e3-41ac-b45d-c79086e6f604/view?access_token=3a759fdb35b1443148092247dd538970da2dad01311520972c9c05f46b496a7c