

Comparison of Beaches

Analysıs and clusterıng of beaches

usıng

data scıence tools and technıques

GOKHAN EKICI | IBM DATA SCIENCE CAPSTONE | 06/07/2020

# Definition of The Problem

When you move to a new country or when you would like to visit a new country for sun, sea, and beaches, you have many websites that provide information about your options. Just like moving to a new neighborhood you want to learn the details, but some details are confusing and some of the content you must read biased. When you search for “best beaches in US” on Google for example, you receive over 400.000 results. And although you know Google sorts them in a relevant order, when you check those results you are confused, you see that the top lists do not overlap, and you might get confused.

On the other hand, when you are deciding to buy a tool, a furniture a holiday package you mostly rely on people with similar tastes like you. But if you are new to a country, if you will visit it for the first time then the chances that you will find someone to understand what you want and respond to it with a good recommendation gets lower.

Today as we have so much information why would not I make my own decision, based on available data out there? At least a similarity analysis, a clustering algorithm might help us all choose places to see given the similarity between the places we like our favorite beach and the surrounding venues and neighborhood. This could be a self-service model which could lead to an affiliate model for tourism agencies or could be directly used by tourism agencies.

A choice of a beach holiday for me has 3 main elements.

1. The beach, its rating and the neighborhood restaurants and cafes and bars etc.
2. The climate
3. The accommodation choices

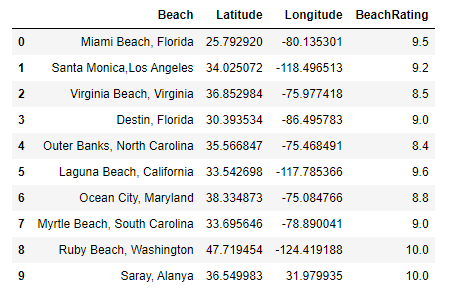
During this analysis I focused on the first 2 as most beautiful beaches already have many accommodation choices and there is plenty of reliable price and feature data on booking.com, hotels.com and Airbnb that can easily be searched and compared. I believe the hard part is finding the right beach.

# Data

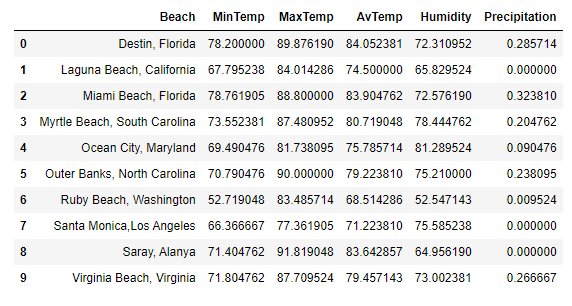
I used 3 main data sets available online, the ones available for free for the sake of this Capstone Project.

1. A list of best beaches in US derived from USNEWS Travel site, Tripadvisor and Foursquare, Tripadvisor, Google ratings for these beaches.
   1. The beach list and subjective was prepared
   2. Foursquare, Tripadvisor, Google ratings for these beaches were collected.
   3. This data was manually and added to a csv file using Python.

Here is a list of the beaches analyzed:

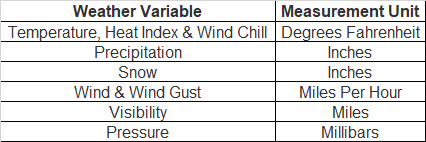


1. Climate data was derived from **visualcrossing.com.** It required a membership. Thefree membership was limited, and I used it according to its limits so that I have at least 3 years data for the 10 beaches analyzed. I only used the weather data between 5th of July and 11th of July 2017-2019 for a 1 week holiday in 2020. The data used for this analysis are the most effective on human comfort, especially while considering a beach holiday in summer. A aggregated sample of the data is:



For statistically better data this could be enlarged to 5-7 years of averages for July. The API returns these values for HISTORY: Address, Date time, Minimum Temperature, Maximum Temperature, Temperature, Dew Point, Relative Humidity, Heat Index, Wind Speed, Wind Gust, Wind Direction, Wind Chill, Precipitation, Precipitation Cover, Snow Depth, Visibility, Cloud Cover, Sea Level Pressure, Weather Type, Latitude, Longitude, Resolved Address, Name, Info, Conditions.

I used Minimum Temperature, Maximum Temperature, Temperature (24 Hour Average), Precipitation and Humidity for this beach comparison/clustering analysis. The US units were used for this analysis:



1. Foursquare Venue Data available through Foursquare API.
   1. The venue list for each beach was downloaded in a 2km distance (30 minutes walking distance assumed)
   2. The main category of each venue was identified after downloading by a function that defines each venue’s category’s main category. The categories are:

*'Arts & Entertainment',*

*'Shop & Service',*

*'Travel & Transport',*

*'Food',*

*Nightlife Spot',*

*'Outdoors & Recreation'*

# Methodology

The beach-based rating, important climate information (temperature, humidity etc.), venues and categories of venues in the neighborhood were normalized with min-max normalization method for effective clustering. However, this method is quite prone to the effect of outliers. So, I used the difference between temperature data and Normal Temperature 68 Degrees Fahrenheit (20 Degrees Celsius) and then normalize our temperature data. For Relative Humidity 50 is an acceptable comfort level. ("Winter Indoor Comfort and Relative Humidity", Information please (database), Pearson, 2007)

Precipitation was not be modified prior to min-max normalization.

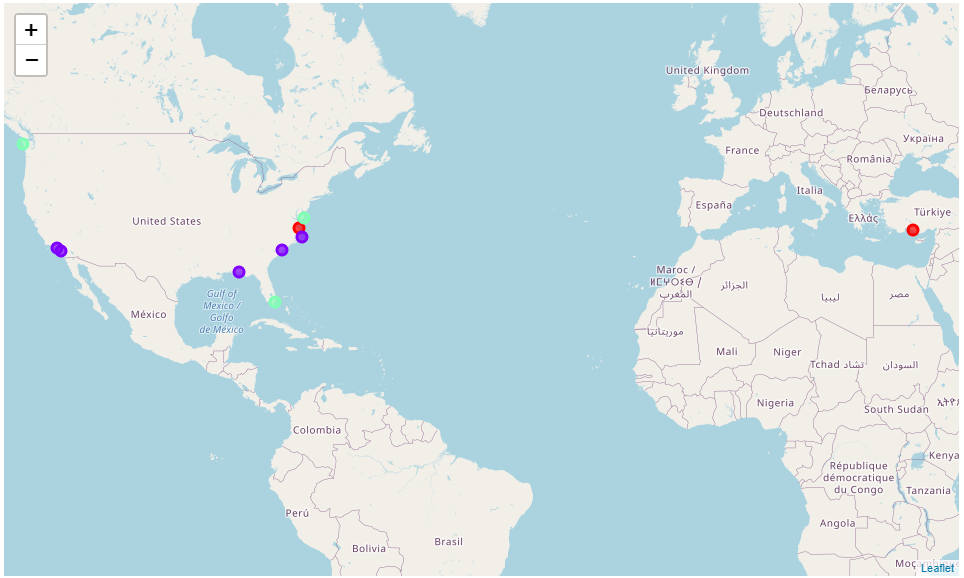
In order to understand the similarities between the shortlist of beaches I have chosen, and our favorite beach (beaches) I added to the list, I used clustering methods K-Means and DBSCAN.

First, I tried K-Means without the climate data, and then I tried K-Means with climate data to see the differences when we run the clustering with climate data. Lastly, I tried DBSCAN and see if it has to offer any insight.

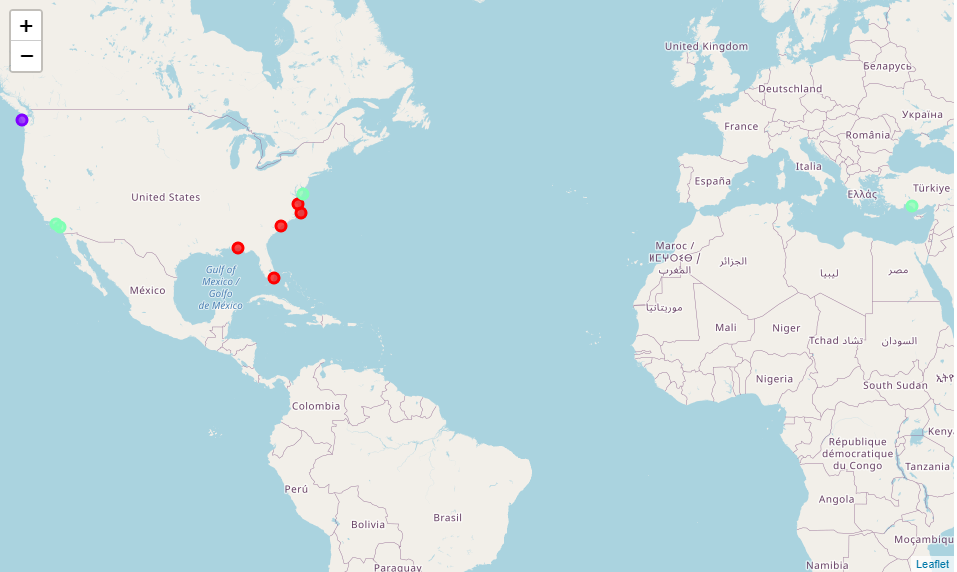
K-Means, although it has model developer’s bias has given some significant results whereas DBSCAN was unable to cluster 8 of 10 beaches in the list, even with different eps values. They were realized as outliers. The diversity of the data might have caused this problem. It would probably give better results with a 2-3 variable clustering analysis.

# Results

K-Means without climate data has created 3 clusters as shown in the map:



It proposed that the most similar beach to Alanya, Turkey was Virginia Beach in Virginia US. As this was with insufficient data, I had to look at the results with climate data by using K-Means. Which has resulted in the following clustering shown on the map:



The resulting comparison set is:



I have found out that California Beaches are more similar to Alanya, given the datasets I used. This is in line with what I have heard of the places from people like me. I have been to Ocean City, Maryland and I had realized how much it is alike with Alanya not only in terms of geographical conditions but also in terms of city facilities, the idea of being able to swim just next to your house etc. So the results here are astonishingly good.

On the other hand, East Coast beaches are clustered together. Except for the Northmost of them which again has a similarity with my favorite beach Kleopatra Beach in Saray, Alanya, Turkey. This tool might help me in my coice of this year's summer vacation. It could be detailed, and the parameters could be changed to see different results.

# Discussion

Other conclusions I could derive from this analysis that are worth discussion are:

1. Climate data is very important in a location comparison. Especially if you are analyzing geographically distinct places.
2. The reviews and ratings on tourism sites are not the only clue you should use. An outlier like the Ruby Beach in Washington for example might be confusing. Although it has good ratings this might be only due to availability, proximity to locals, whereas most people prefer more crowded beaches in the south, mostly California and Florida.
3. DBSCAN did not help where there were outliers to properly cluster and give some significant results. This could be because of too many parameters used for the analysis.
4. Every conclusion is dependent on the data set you have. Reliable data is not easily found, and bias can totally mislead us.
5. It is very important that some organizations publish their data free of charge for at least social use and I think they will be rewarded when someone creates value out of them.

My recommendations for those who would like to work on this and improve it would be

1. Find reliable data sources online. There are lots of free sources, but you should be sure which one come with what kind of bias, and for how much.
2. Do not try to use all data available. Sometimes part of data will represent other parts. E.g climate data had many other fields that were either not relevant for a beach vacation analysis or were already represented by others.
3. Decision for Normalization method is as critical as choosing the machine learning method. E.g climate data should not be directly normalized by min-max etc.

# Conclusion

The model works well to compare beaches and cluster them so that you can choose where you would like to have your next beach vacation in a new country, in a new part of your country etc. The important issue is you should be very much aware of the biased and commercial data out there on the web.

I feel better about my next vacation search process already. I hope this might help others in the future as well.

**Information increases by sharing. So, keep sharing.**

