A Family Business Moving to a New Location

Capstone Project, IBM Data Science Professional Certificate - The Battle of Neighborhoods (Week 2)

Muhammad Faizan Siddiqui
(09-May-2020)

1) Introduction

a) Background

A Huntsville, Alabama based family has been mostly into the health-and-fitness business over the past many decades. Some family members have retail stores selling exercise machines and related equipment; others have gyms and medium-sized swimming pools. Some of the members sell maintenance services for the equipment sold by family-run as well as other retail stores.

Unfortunately, in the past couple of years, all segments of the business have seen a sharp decline in revenue. For them, people are no more investing in health and fitness that seems unlikely but valid for them. While the family is well off and can easily withhold a couple of more years like this but there seems to be no silver lining in the short term.

b) Problem

The family is considering relocating to a different city, most probably in another state, which is lucrative for their family-run business. Moreover, rebuilding the business from scratch is a gigantic task, and in their business, real estate is very important; therefore, they are also looking for the city with relatively requires low investment in real estate. Finally, as the whole family will be relocating along with the business, they would like to remain on the South or East side and their elders are especially concerned about the changes in the weather.

c) Approach

We will address the problem in reverse order as it would simplify the process, reduce processing overhead and help in achieving objective within API limits.

- Create list of states on the South and East side.
- Shortlist those with similar weather conditions using KMeans Clustering.
- Select states from these state with average low-cost of real estate
- Finally, determine the cities that may have more potential for health-and-fitness business

2) Data

a) Requirements and Sources

In line of our approach, following primary data sources will be used to obtain desired information.

- Download census regions data to create list of states on the South and East side
- Web scrape state-wise average weather data.
- Download real estate median home value dataset from Zillow via Opendatasoft: Data Network.
- Use Foursquare API to determine health-and-fitness related venues within the selected cities.
- Download list of cities from United States Cities Database (basic version) from SimpleMaps.

b) Data cleansing and feature selection

First of all we loaded region-wise states list from Cphalpert (GitHub). The data was based on the US census bureau. We discarded the division column as we did not need it. Since, we needed only selected records where the region was either West or South, we filtered rest out, except the record for the Alabama state that we needed for clustering purposes.

Next, we scraped the weather data from the Current Results website. We retrieved data for average temperature, precipitation, snowfall, humidity, and sunshine. We removed the features that were not needed for the clustering.

Then we loaded the real estate median home value dataset from Zillow via Opendatasoft: Data Network. While we only used home value data as it was freely available, similarly we could have used commercial data for their business properties. To minimize memory consumption and processing time we read only the required columns. We then converted the date to state-wise means.

We also loaded city-wise states list with additional information like population density and city ranking information.

Used geolocator API to locate geographical coordinates of the USA.

Foursquare API was used to fetch venues of desired categories required to identify cities favorable for the family business.

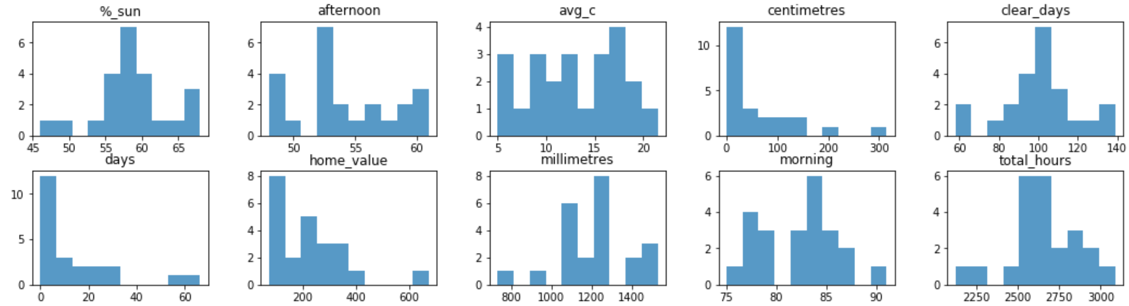
Finally, we also used a geo-json file required for choropleth maps using folium library.

We standardized all column names – removed extra spaces, parenthesis and replaced spaces with underscores. We also removed any non-ASCII characters from all column names and from certain columns to help us in join operations and while persisting the files as CSV.

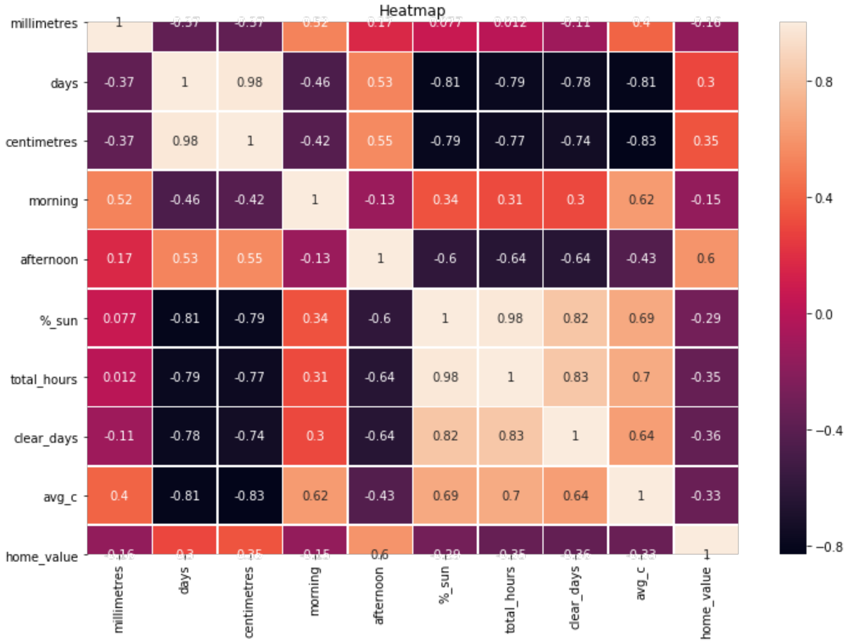
3) Methodology

We have started from creating a list of states on the south and east side. Then we collected average weather data at state level. The next step was to collect the median house value for each state. All of this was merged to create a dataset that was prepared for cluster analysis to identify states like Alabama based on the weather.

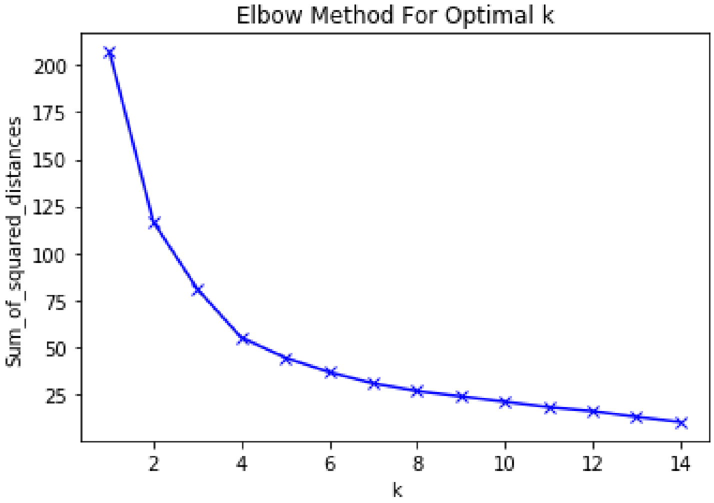
We generated histogram to identify any irrelevant features for clustering. There were outliers within the features, but all features seem to be significant for the job.



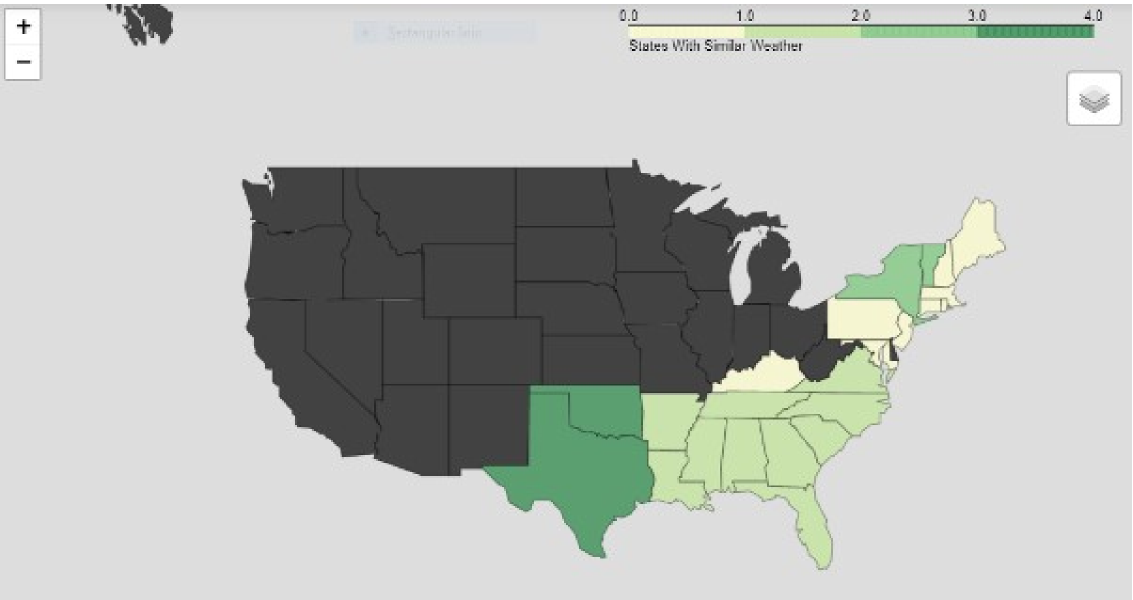
Though, correlation within features is not a major problem for Clustering, let us have a look at our dataset. There is a strong relationship between some attributes e.g., %sun, number of days ice falls and amount of ice etc. but that is not expected to hurt the model.



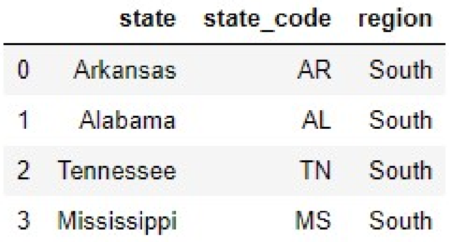
We used KMeans clustering with k-means++ initializer algorithm to optimize the step of picking the clusters’ centroids which increases the likelihood of finding an optimal solution. Further, we used Elbow method to find optimal value for ‘k’ i.e., 4 as follows.



Using the optimal value, we generated a choropleth map using a folium library to visually depict the cluster of states based on the weather data.



Finally selected the caster containing Alabama state and then we selected top 4 states where median house value was the lowest.



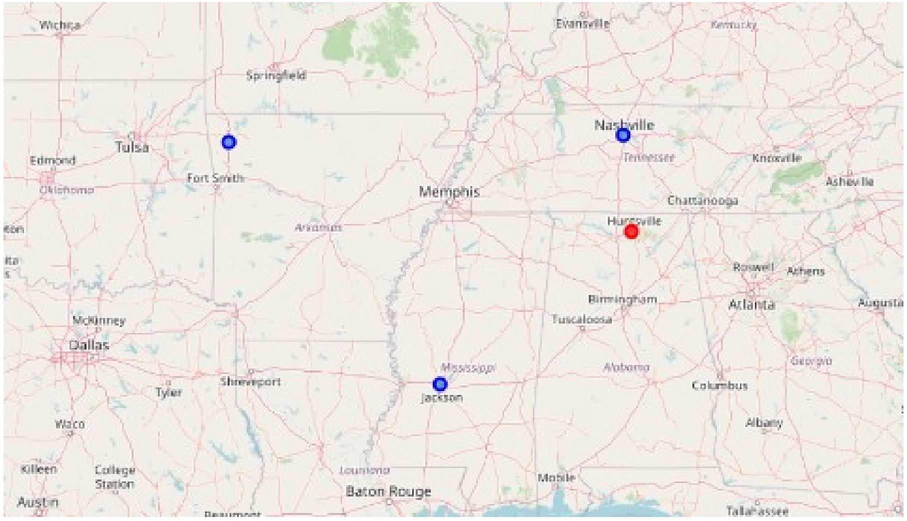
	state	state_code	region
0	Arkansas	AR	South
1	Alabama	AL	South
2	Tennessee	TN	South
3	Mississippi	MS	South

Then be loaded top cities for each shortlisted state filtered around four cities per state based on population density and city ranking.

Finally, we used Foursquare API to retrieve venue information for the categories that either belonged directly to health-and-fitness or sports in general. We wanted to identify cities with maximum number venues in these categories based on the selected keywords.

4) Results

We ended up finding **Nashville, Fayetteville and Jackson** as also shown on the following folium map (in blue) which also shows the original city of current residence, Huntsville (in red), by the family in Alabama.



5) Discussion

First, we made this assumption that the business indeed required to be relocated and there is nothing that can be done to improve its current state where it is currently operating.

Then, we have taken a very simplistic approach to a very complex business problem; and therefore, have not considered various aspects for relocating a business other than the aspects directly related to Data Analytics assignment to minimize the scope of the work.

Next for the purpose of minimizing real estate investment we have only considered house median values as the dataset was freely available; however, for the actual business problem this data cannot be relied upon.

Last but not the least, we have always tried to minimize scope of work by selecting minimum number of states and then again minimum number of cities, and then by avoiding to going down to the borough or neighborhood level to also minimize the use of Foursquare API free version.

In future these shortcomings should be overcome to prepare a commercially viable data analysis solution for the problem of such a scale.

6) Conclusion

The objective of this exercise was to help a poorly performing family business in relocation, given a number of constraints - the family wanted to remain on the east or south side, they were very concerned about the weather conditions and they were looking for low investment in real estate. Finally, the core objective was to find a city where their health-and-fitness business could thrive like good old days.

At the end of the exercise we were able to identify three cities in different states that met the stated requirements.