In Search of Lower Cost of Living

Jonas Andersson

Coursera IBM Data Science Capstone

Feb 2021

Introduction

Everybody has a dream place where they want to live. But as housing prices have skyrocketed during the past years, the area where one ideally would want to live, may not be in range when it comes to affordability. In this project, I try to mitigate some of that issue by providing a framework where a person who has seen their dream area but think it's too expensive, can see what other areas that are similar, but cheaper.

Data

As data I use income data from St. Louis Fed (income data from St. Louis Fed and lat/long data from simplemaps.com. I then use the lat/long data per city to retrieve the 50 most popular spots at that specific location via the Foursquare explore service.

For the purpose of this lab i limit i limit my selection to the west coast of the US (as defined by west of lat -117), with my favorite area set to Seattle. I then focus on the top three cities in each region that has a population over 10k. I crudely proxy the cost of living with the median household income, and assume it's strongly correlated.

Methodology

I first downloaded a file from St Louis Fed which I then uploaded to my Github account for further retrieval.

I used Folium map library to visualize the maps

I used the Foursquare API with limit of 1000m, but as the API limits to 50 results at the time, it’s a moot point really. I chose to sort the results based on popularity to mitigate the lack of number of venues.

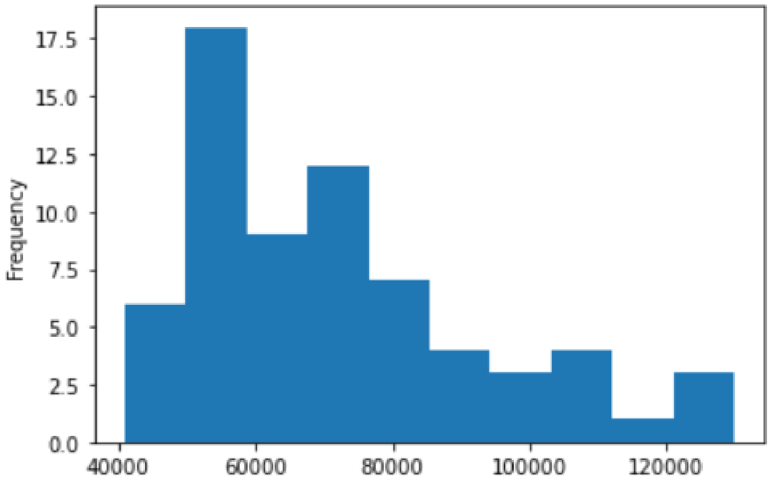
Knowing that many cities in the same country look the same I opted to further differentiate the different cities using TF-IDF. TD-IDF is a method that focuses on how many occurrences of an item in all cities compared to the items in that specific city.

After that I normalize the data using Standard Scaler from the SkLearn library.

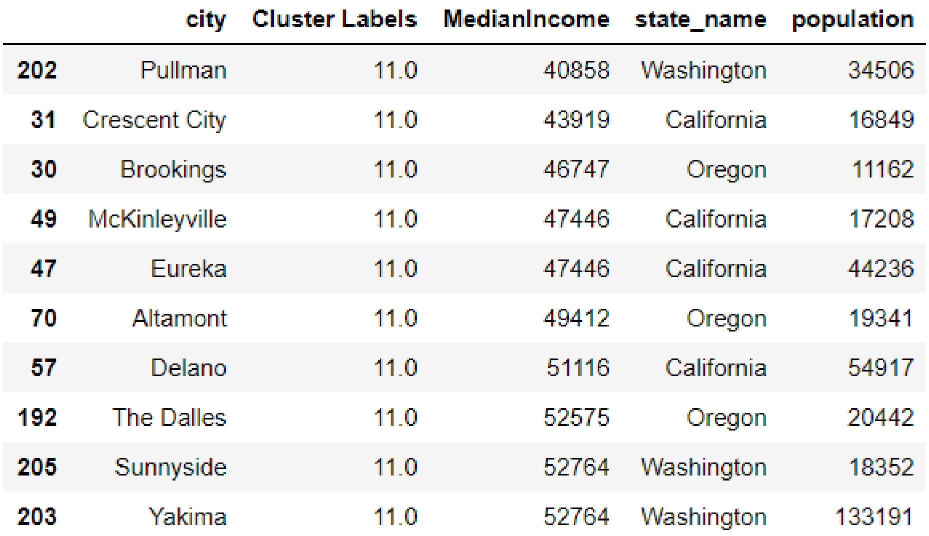
Then I proceed to cluster the cities using Spectral Clustering, with 14 clusters. Since there are over 300 features (venues, restaurants, etc) the Spectral Clustering has an edge as there is a dimensionality reduction feature in there.

Results

After clustering we can see that Seattle is in a group with 66 other cities, so there will be lots to chose from. After sorting the data on Median Income, and whereas Seattle has a median income of just north of 100k, the below histogram shows that most cities in the cluster offer the same mix of venues for a cheaper cost of living.



The top 10 cities from the same cluster when it comes to affordability is shown below.



	city	Cluster Labels	MedianIncome	state_name	population
202	Pullman	11.0	40858	Washington	34506
31	Crescent City	11.0	43919	California	16849
30	Brookings	11.0	46747	Oregon	11162
49	McKinleyville	11.0	47446	California	17208
47	Eureka	11.0	47446	California	44236
70	Altamont	11.0	49412	Oregon	19341
57	Delano	11.0	51116	California	54917
192	The Dalles	11.0	52575	Oregon	20442
205	Sunnyside	11.0	52764	Washington	18352
203	Yakima	11.0	52764	Washington	133191

Discussion

As can be seen above, the recommendation for anyone wanting to live in Seattle could consider these similar cities instead. Of course, this report only looks at the venues and the service offerings in the cities, and does not care about scenery or any environmental factors. Another limiting factor in this analysis is the Foursquare limitation of how many venues that can be retrieved from the API for the same place. Furthermore, it would be interesting to look at environmental variables, like distance to bodies of water, mountains and similar to see if this would yield different results. Unfortunately, this has been deemed out of scope for this report.

Conclusion

In conclusion, I believe there is some truth to the findings. Looking at perhaps ***Eureka*** from the above select cities, it may be a viable alternative to Seattle as it is very close to the Pacific Ocean as well.