LINKING VENUES TO THE MEDIAN INCOMES OF A LOCATION

Jyotishka Misra

July 2020

Table of Contents

Introduction

Modern retail transactions are widely dependent on the acceptance of credit cards. Banks issuing the credit cards earn revenue both from any annual fees on the cards paid by the customer and a portion of each transaction value paid by the retailer. Therefore, a bank will earn more revenue if customers accrue a high value in transactions (either more purchases, or more expensive purchases, or both) and are able to pay off any balance on their cards on time. It is important for the bank to find such potential customers and offer their credit services to them. One way that banks do this is by directly sending mail to addresses with a description of their services and how they can be availed.

In this direct mail advertising, businesses send physical mail to every address in a particular region. However, the conversion rate of a person receiving mail into a paying customer is very low. Therefore, a business will prefer getting the best possible customers for the funds being spent in the effort. In order to attract better customers, it is beneficial to target those with more disposable income to spend. As we cannot accurately predict the exact budgets of each and every household in a region, we can use the median household income as a proxy for this measure. It is important to use the median rather than the mean household income as it is more representative of the entire population given that it is not influenced by the outliers that may exist in the region.

However, the problem with this approach is that it automatically excludes any region for which data is not readily available. This can mean new communities and regions that have not been surveyed for a long period of time will likely be not advertised to. In addition, it is quite expensive to survey entire regions for data about their incomes. This results in banks missing out on a considerable amount of revenue and large portions of the population remaining underserved in terms of credit services. To overcome this hurdle, we can link higher median incomes to the types of businesses that open and operate in a given region. This allows us to make approximations about the expected median income of a given region without having to rely on expensive surveys.

Data

Phoenix is not only one of the largest cities in the U.S. but also one of its fastest growing cities. This will serve as the city studied to try and link median household incomes to a specific cluster of business types. To achieve this end, all of Maricopa county (within which the city of Phoenix resides) will be used for the project. The data thus required will be gathered as follows:

1. A list of locations within Maricopa County will be used from the website of the newspaper Phoenix New Times.
2. The forward geocoding for each of those locations will then be done using LocationIQ and the geocoder library in python.
3. The latitude and longitude from the forward geocoding will then be used with the uszipcode python library to ascertain which U.S. zip code that particular location falls in and from that will be used to return the median income of that location.
4. The Foursquare API will then be used to return the venues for each of the locations, their respective latitudes and longitudes, and the venue category of which they are a part.

Methodology

The web-scraped data was used to with forward geocoding to create a dataset of their respective geographical coordinates and median incomes of the zipcodes within which the coordinates fall. The data from the Forusquare API was used to find and assign venues and their venue type to a given location. Therefore, the characteristic venues of a given location can be ascertained. The radius for the search for venues was done in a manner so that it was equal to roughly half the distance of the two distinct locations. This was important to ensure that the same venue was not assigned to two different locations which would make the two different locations seem similar to one another even if they really were not.

One-Hot encoding was done for each venue type within every location. This allowed the calculation of the mean venue type of a given location. Carrying out the K-means clustering with K=5, allowed us to group together locations with similar venue types and calculate the average median income of that specific cluster.

Results

The K-Means clustering yields us the following map:

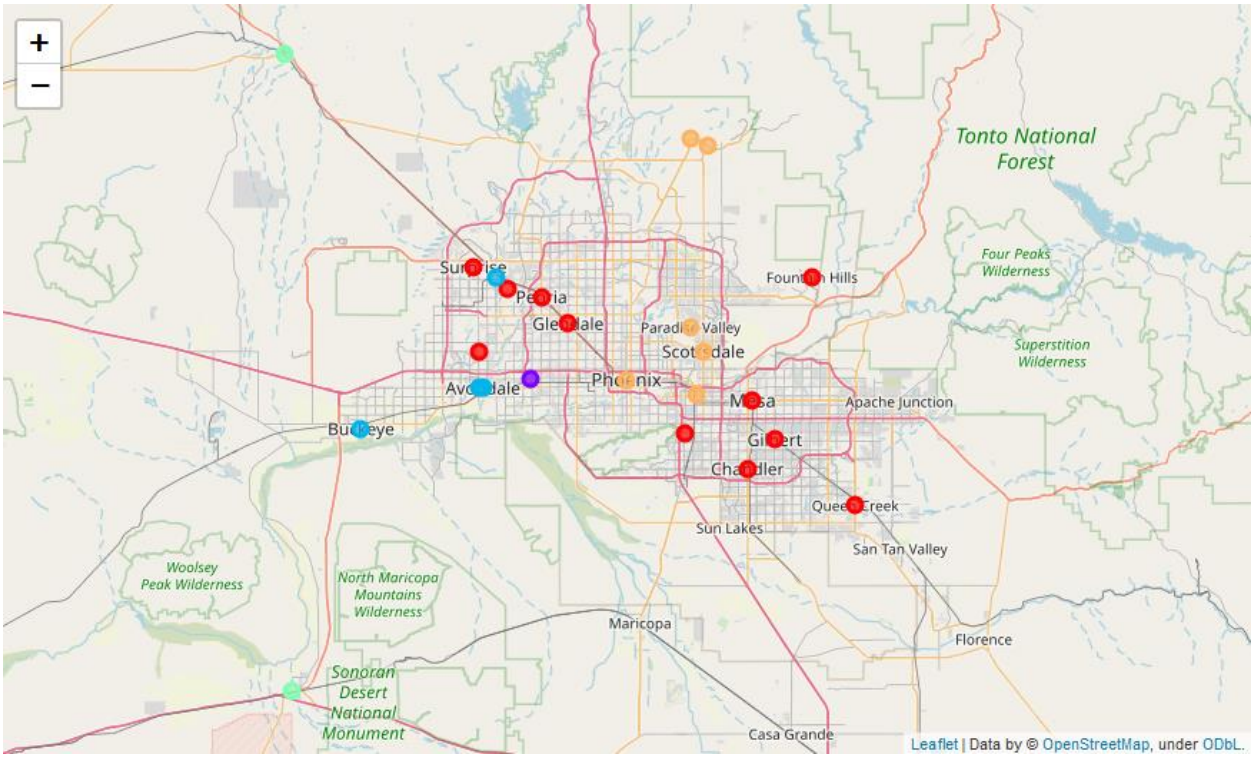

Figure 1: Map with all Clusters visualized

Each of the different clusters is represented by a different colour. Looking at them individually, we can see:

Cluster 1

This is the largest cluster. It has an average median income of $55,462.02 and contains the locations Youngtown, Surprise, Guadalupe, Fountain Hills, Mesa, Gilbert, Queen Creek, Peoria, Litchfield Park, Chandler, and Glendale.

	Place	Latitude	Longitude	Median_Income	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
0	Youngtown	33.593730	-112.303326	33884	0	Convenience Store	Fast Food Restaurant
4	Surprise	33.629227	-112.368019	70302	0	Sandwich Place	Fast Food Restaurant
5	Guadalupe	33.363125	-111.962533	53022	0	Fast Food Restaurant	Sandwich Place
9	Fountain Hills	33.611711	-111.717361	73608	0	Pizza Place	Italian Restaurant
10	Mesa	33.415112	-111.831479	36586	0	Mexican Restaurant	Convenience Store
12	Gilbert	33.352763	-111.789037	75365	0	Mexican Restaurant	Sandwich Place
13	Queen Creek	33.248386	-111.634158	73367	0	Mexican Restaurant	Pizza Place
16	Peoria	33.580612	-112.237294	45886	0	Fast Food Restaurant	Convenience Store
17	Litchfield Park	33.493380	-112.358124	64383	0	Coffee Shop	Grocery Store
18	Chandler	33.306160	-111.841250	56414	0	Mexican Restaurant	Convenience Store
20	Glendale	33.538686	-112.185994	27267	0	Convenience Store	Mexican Restaurant

Figure 2: Cluster 1 Locations

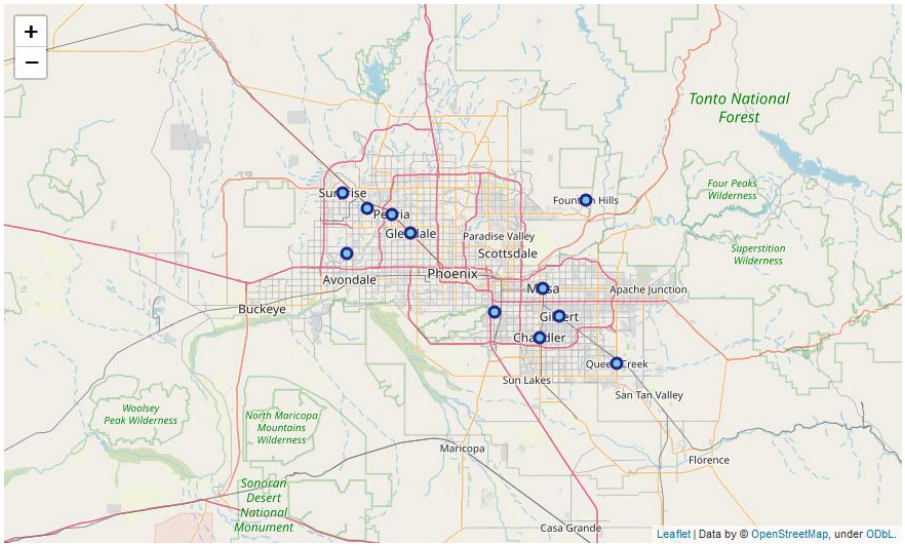

Figure 3: Cluster 1 Map

Cluster 2

This cluster has an average median income of $50,066.02 and contains the single location of Tolleson.

	Place	Latitude	Longitude	Median_Income	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
3	Tolleson	33.45005	-112.259309	50066	1	Convenience Store	Mexican Restaurant

Figure 4: Cluster 2 Location(s)

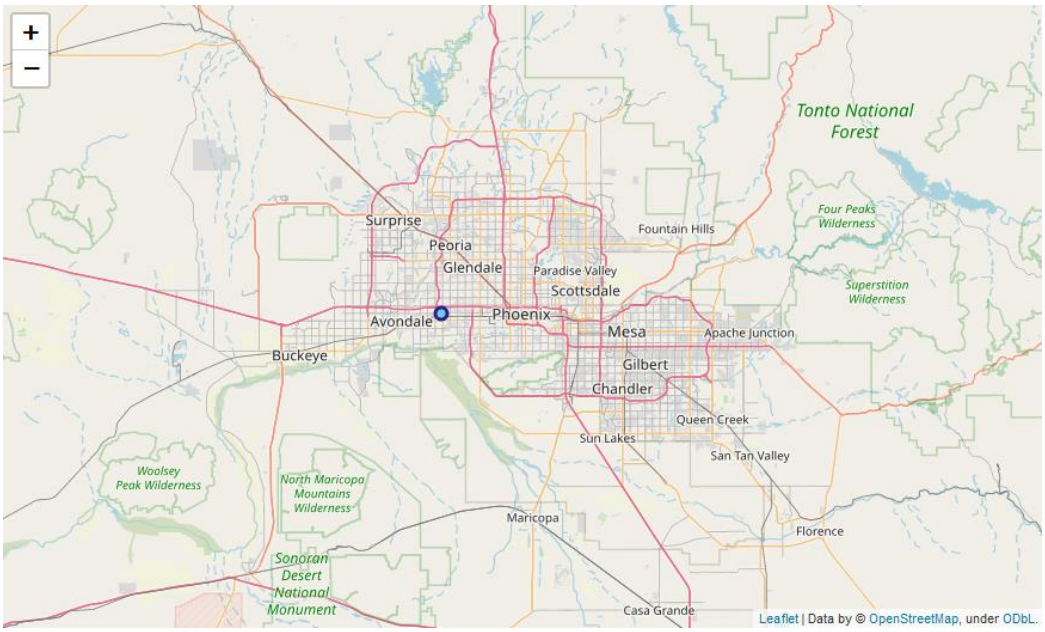

Figure 5: Cluster 2 Map

Cluster 3

This cluster has an average median income of $51,348.02 and contains the locations El Mirage, Avondale, Buckeye, and Goodyear.

	Place	Latitude	Longitude	Median_Income	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
2	El Mirage	33.613034	-112.324487	47237	2	Mexican Restaurant	Convenience Store
6	Avondale	33.435598	-112.349602	44658	2	Mexican Restaurant	Rental Car Location
7	Buckeye	33.370275	-112.583867	68839	2	Mexican Restaurant	Pizza Place
11	Goodyear	33.435367	-112.357601	44658	2	Mexican Restaurant	Fast Food Restaurant

Figure 6: Cluster 3 Locations

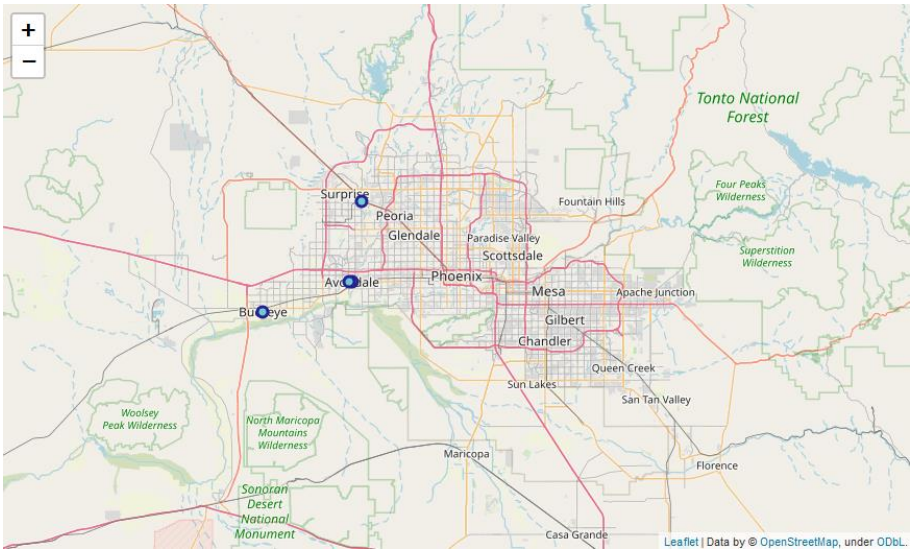

Figure 7: Cluster 3 Map

Cluster 4

This cluster has an average median income of $36,308.02 and contains the locations Gila Bend, and Wickenburg.

	Place	Latitude	Longitude	Median_Income	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
1	Gila Bend	32.947827	-112.716824	30242	3	Fast Food Restaurant	Mexican Restaurant
15	Wickenburg	33.968096	-112.730135	42375	3	Fast Food Restaurant	Mexican Restaurant

Figure 8: Cluster 4 Locations

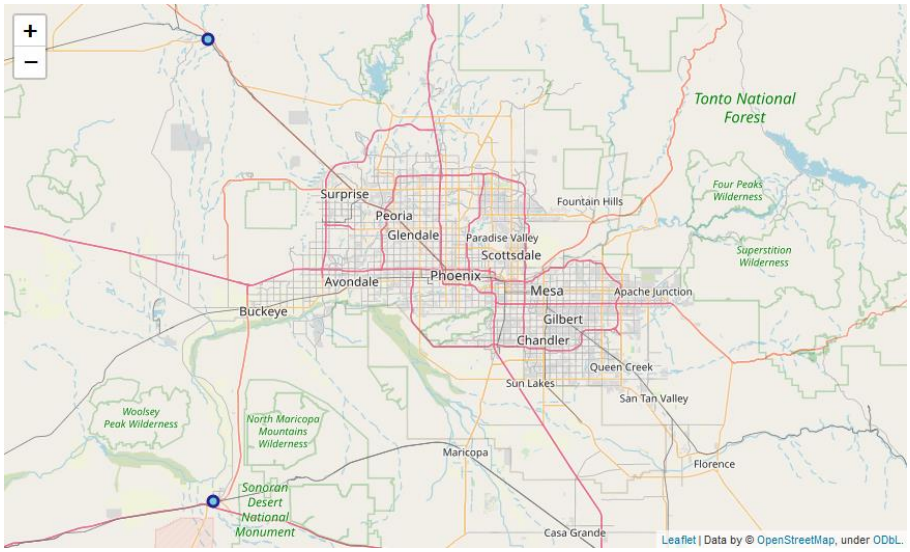

Figure 9: Cluster 4 Map

Cluster 5

This cluster has an average median income of $67,360.02 and contains the locations Carefree, Cave Creek, Paradise Valley, Tempe, Scottsdale, and Phoenix.

	Place	Latitude	Longitude	Median_Income	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
8	Carefree	33.822261	-111.918203	100338	4	Coffee Shop	American Restaurant
14	Cave Creek	33.833333	-111.950833	88938	4	Coffee Shop	Mexican Restaurant
19	Paradise Valley	33.532429	-111.950512	109185	4	Spa	American Restaurant
21	Tempe	33.425506	-111.940012	30582	4	Coffee Shop	Breakfast Spot
22	Scottsdale	33.494219	-111.926018	49111	4	Coffee Shop	American Restaurant
23	Phoenix	33.448437	-112.074142	26008	4	Coffee Shop	American Restaurant

Figure 10: Cluster 5 Locations

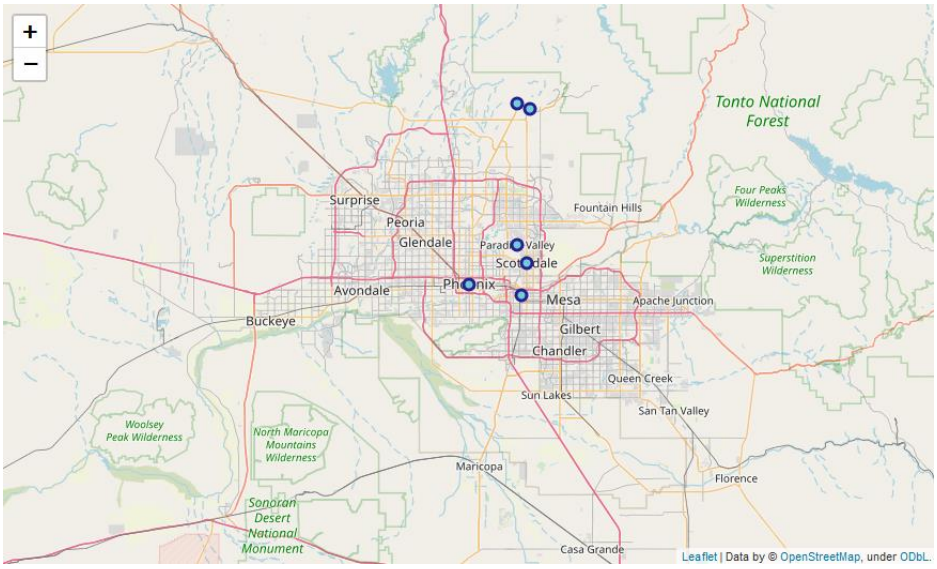

Figure 11: Cluster 5 Map

Discussions

From the clustering, we can see that the cluster with the highest average median income is cluster 5. This region is characterized by an abundance of coffee shops and American Restaurants with some dining establishments and a spa also being present.

The lowest average median income was seen in cluster 4 which was characterized by fast food and Mexican restaurants. The other three clusters had a comparable average median income. In these clusters, the lower end of the average median income was witnessed in clusters having more convenience stores and Mexican restaurants while the higher average median incomes were witnessed in locations with more diversity in venues even though Mexican restaurants remained dominantly present.

Conclusions

As Phoenix is a rapidly growing city, targeted promotions can be applied in regions where we see more coffee shops opening up. Regions in which fast food and Mexican restaurants appear to be shutting down and replaced with coffee shops and American restaurants are likely attracting people with more disposable income and can also be targeted profitably.

Future Directions

The study conducted is fairly localized to Phoenix. Attempting a similar clustering on different cities would be beneficial to gain an understanding the types of venues that are likely to signal that an area has a higher than average median income, and by extension higher disposable income. In addition, converting the incomes of a given city into percentiles would allow us to better compare different cities to each other. This is because we will be removing the regional variations in salaries and allow us to directly compare the impact of the presence or absence of a specific venue. Finally, another method of improving the study would be to factor in the cost of living as it will have a considerable impact on the disposable income of a consumer in a specific area.