Battle of the Neighborhoods - Coursera Capstone Project

Patrick Haney

May 17, 2020

1. **Introduction**

**1.1 Background**

ACME Realty, a nation-wide realty group, is interested in expanding their services to include Downsizing Assistance. This new service will work to match adults who are 50+ years old with living spaces that are in very accessible neighborhoods that have services and amenities such as hospitals / medical practices, walking trails, parks, and grocery stores. ACME Realty would like to pilot their service in Phoenix, Arizona, to gauge if it is a viable idea that can be expanded to other cities. They want to build a deep understanding of what the neighborhoods in Phoenix, Arizona offer so that they can find the optimal neighborhoods for their clients to buy new living spaces in. ACME Realty has enlisted your services to help them find the best neighborhoods in Phoenix for them to target during the pilot test.

* 1. **Problem**

ACME Realty wants to have data-driven profiles of neighborhoods in Phoenix to understand which ones are most suitable and least suitable for their new Downsizing Assistance service. They also want to know which neighborhoods are similar to each other, so that if one neighborhood doesn’t have anything on the market, they can recommend other suitable neighborhoods that have similar characteristics.

**1.3 Interest**

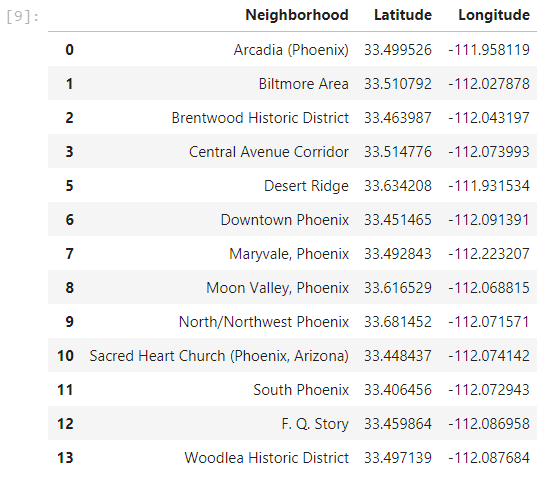
The audience of this project is ACME Realty’s Director of Market Strategy and their Chief Marketing Officer. Once the analysis has been completed, they would like to see the results in a PowerPoint presentation, with clearly represented findings and suggestions for next steps.

1. **Data**

**2.1 Data Sources**

The data for this project was derived from Geopy, a popular Python Geography library, along with the Foursquare API's exploration endpoint, a list of Phoenix neighborhoods from [Wikipedia](https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Phoenix,_Arizona), and a few coordinate pairs from Google Maps to fill in gaps in Geopy’s database. This data consists of the coordinates and names of all of Phoenix, Arizona's neighborhoods; the coordinates of the center of the city itself; and lists containing details on up to 100 venues per neighborhood.

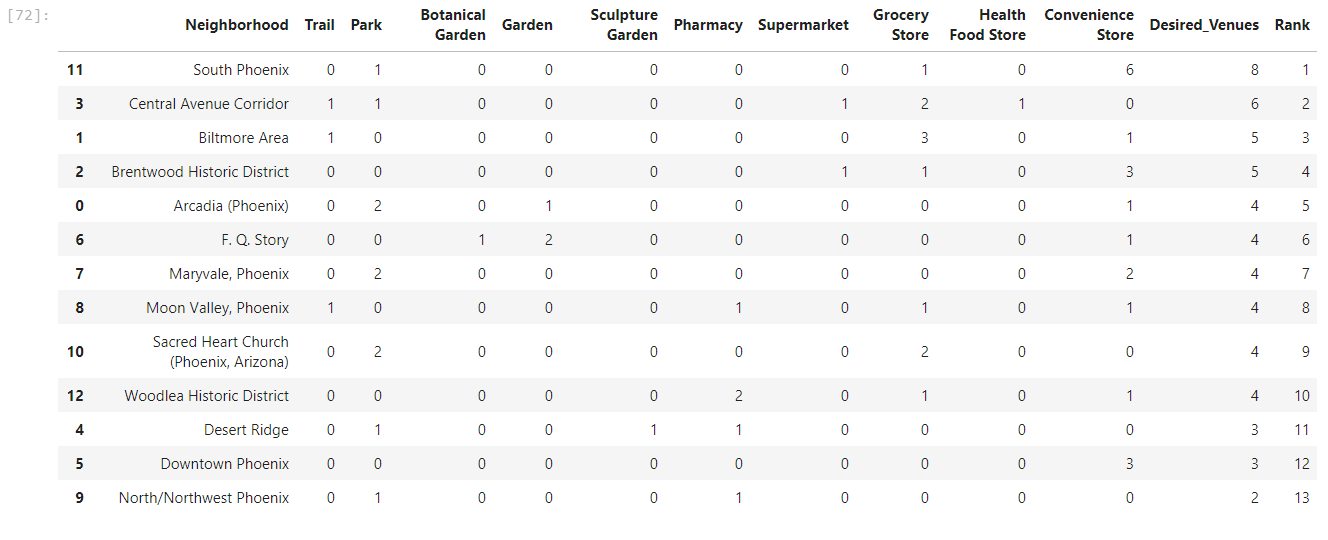
Example of Geopy/Google Maps derived neighborhood coordinates used in analysis:



Each venue list will contain the name, coordinates, and venue category for each venue in the list.

These data will be merged together in a dataframe, and the venue categories will then be one-hot encoded, averaged to show what categories are most prevalent, and merged in as well. The resulting dataframe will be used for k-means clustering with the scikit-learn library, which will help with identifying the neighborhoods that will be more suitable for the 50+ aged community.

Example of Foursquare Venue data used in analysis:

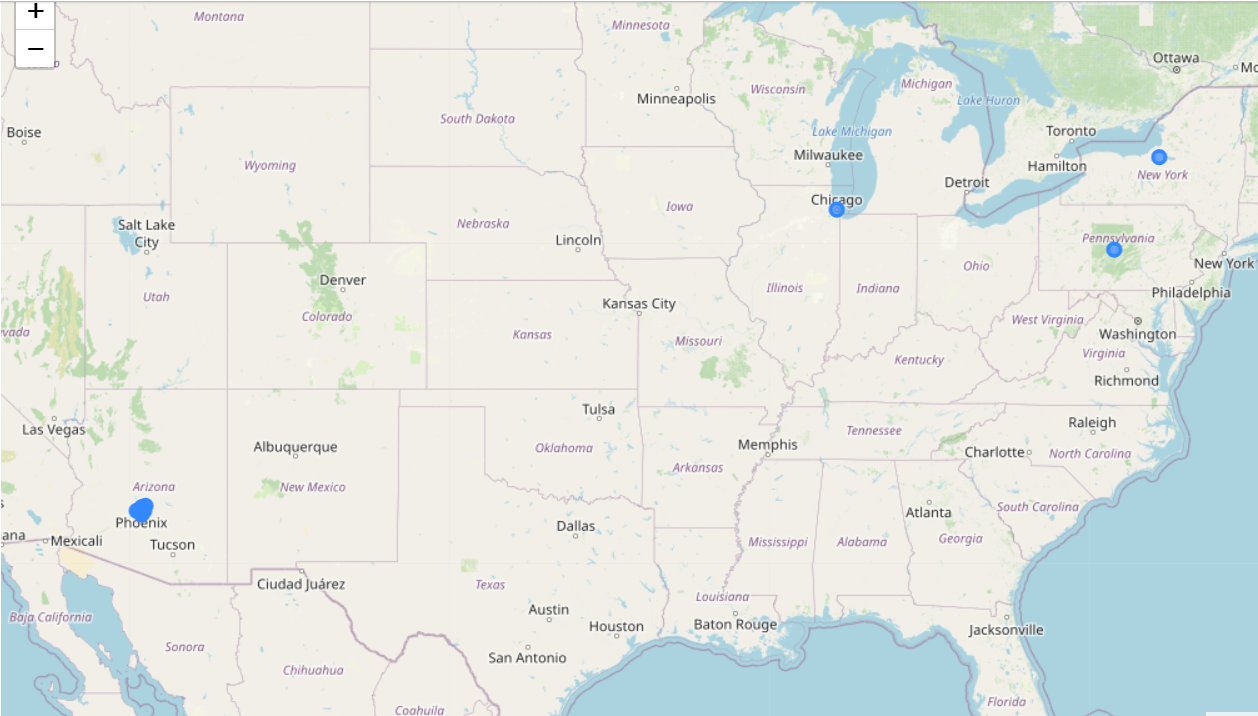


Once the data has been clustered, each of the clusters will be analyzed and the overarching themes of the neighborhoods will be described.

**2.2 Data Cleaning**

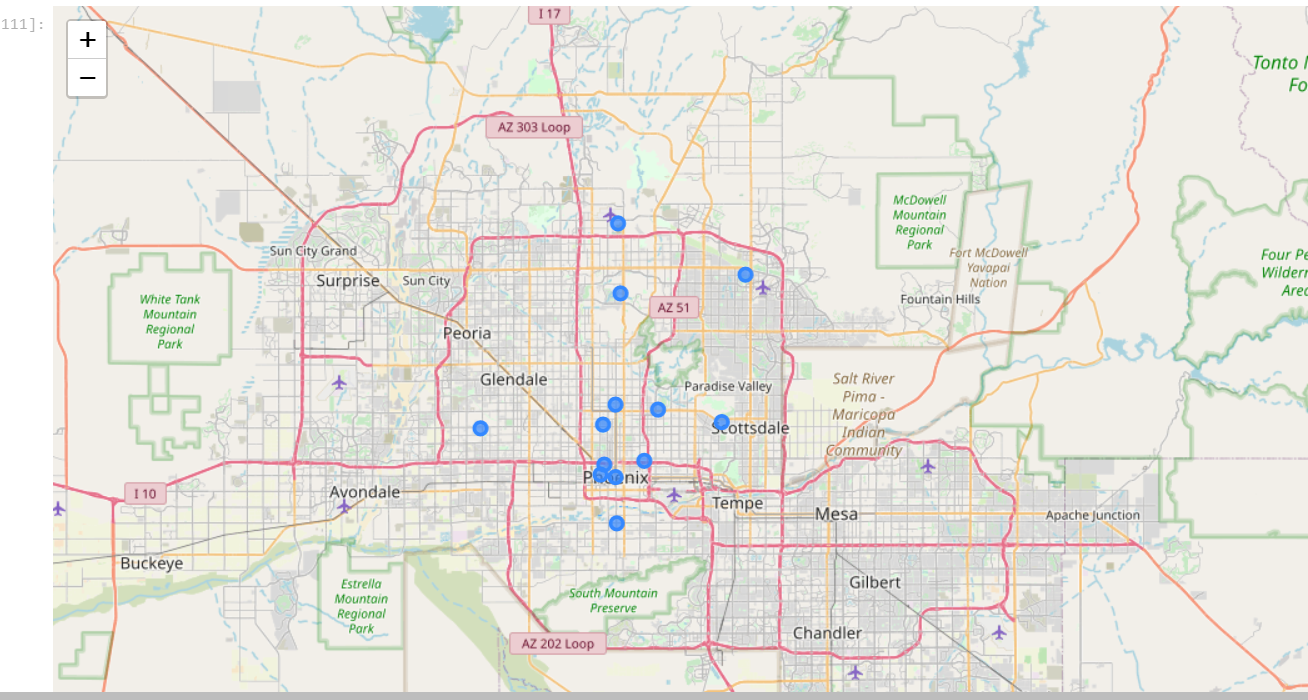
The neighborhood data gathered from Wikipedia was difficult to use. Initially there were 15 neighborhoods listed by Wikipedia. When plugged in to Geopy, Central Avenue Corridor, Golden Gate Barrio, and F.Q. Story Historical District neighborhoods did not return results. It turned out that Central Avenue Corridor was not in Geopy’s database, so I had to gather the coordinates from Google Maps. Upon further research, I determined that Golden Gate Barrio is no longer a neighborhood, and that Wikipedia had included both current and past neighborhoods. F.Q. Story Historical District was not recognized by Geopy, but F.Q. Story was.

After fixing gathering coordinates for these neighborhoods and removing the defunct neighborhood, I plotted them on a map. That made it apparent that three neighborhoods in the data set were on the other side of the country.



The API had misinterpreted the queries for Chinatown, Downtown, and North/Northwest Phoenix. Upon further research, I found that Chinatown was another defunct neighborhood and omitted it from the analysis. Downtown and North/Northwest Phoenix’s coordinates were gathered from Google Maps. With the new coordinates entered and the data clean, it was time to start the analysis.

Clean neighborhood coordinates



Methodology

To answer the question of what neighborhoods in Phoenix, Arizona are most suitable for the 50+ audience, I went back to what ACME Realty had initially specified in the prompt they had provided. They were particularly interested in the existence of convenient access to hospitals / medical practices, walking trails, parks, and grocery stores within the neighborhoods. Since all of these neighborhoods have those features within driving distance, the analysis focused on what the neighborhoods had within a 15 minute walking distance.

The analysis’ key focuses were to rank each neighborhood based on the concentration of these desirable services; to build a data-driven understanding of what each of the neighborhoods had to offer; and to find out which neighborhoods were similar to one another, so that ACME Realty could recommend similar neighborhoods to homebuyers if they are unable to find a home in a neighborhood they liked.

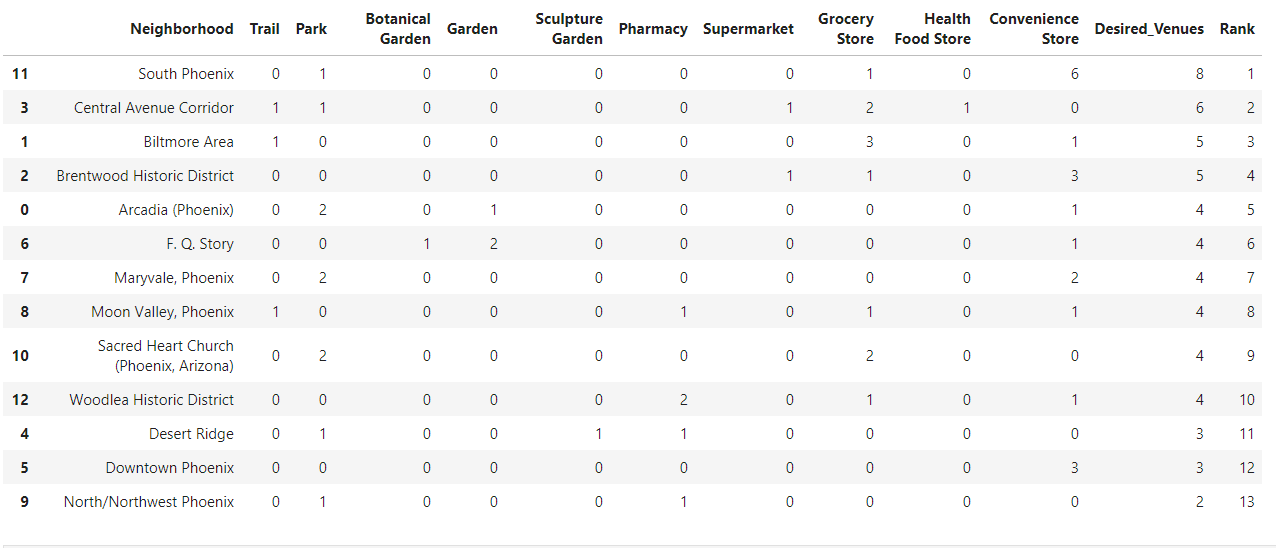
To rank the neighborhoods, I went into Foursquare’s [API Documentation](https://developer.foursquare.com/docs/build-with-foursquare/categories/) and found all venue categories that pertained to healthcare, walking trails/parks, and grocery stores. Then, utilizing a dataframe that had each neighborhood and one-hot encoded records for each venue that was within walking distance, I decided that the sum of the occurrences of venues that fell within the categories that ACME Realty was interested in would serve as a good metric for gauging the accessibility of the neighborhood. I summed up all of the relevant categories, and then calculated a grand total for each neighborhood. The neighborhoods were then ranked, with the highest grand total having the best rank.

After ranking the neighborhoods, I proceeded to use the dataframe with all of the one-hot encoded records to develop a description of each neighborhood based on its offerings. This involved averaging the occurrences of venue categories to calculate the density at which venues of certain types appear in the neighborhood. The top 10 most frequently appearing venue categories for each neighborhood were found using this method. This top 10 list is open to a degree of subjective interpretation by the analyst, the realtor, and the homebuyer, but serves as a data-driven resource to help develop a profile for the neighborhood.

Lastly, scikit-learn’s k-means algorithm was utilized to cluster similar neighborhoods together. I tested cluster counts of 3 – 8 and found k=5 to be the best. When there were only k=3, almost all of the neighborhoods ended up in clusters 0 and 1, and their top 10 venues had very few similarities. Conversely, with k=8, the clusters became way too specific, with half of the clusters only having 1 neighborhood in them. With k=5, the clusters were sensitive enough to find similarities and differences between neighborhoods, but not enough to make every neighborhood its own cluster. After the right amount of clusters was found, I looked for characteristics that made the clusters unique and then created descriptions, which are shared in the Results section. I then merged the top 10 venue categories and the cluster labels with the venue ranking and created a final, consolidated dataframe that could be supplied to ACME Realty as a deliverable that provides data-driven profiles of each neighborhood.

Results

All the neighborhoods had some of ACME Realty’s desired venue types within walking distance, but there was a stark difference in the amount. South Phoenix, the top ranked neighborhood, had 8 desirable venues within walking distance, which was 4x as many as North/Northwest Phoenix, the lowest ranked neighborhood.



Although many of the neighborhoods had similar offerings, it was useful to find out the top 10 most common venue types, as it helped to describe each of the neighborhoods in a data-driven fashion.



The top 10 venue categories were used not only to describe the neighborhoods at the individual level, but also to find thematic commonalities between neighborhoods that were in the same cluster. Below are descriptions of the 5 clusters that were generated through the k-means algorithm.

In Cluster 0 (Brentwood Historic District and Woodlea Historic District), there is an emphasis on Mexican Food, Cafes, and Gay Bars.

Cluster 1 (Central Avenue Corridor, Biltmore Area, F. Q. Story, Moon Valley, Phoenix, Sacred Heart Church (Phoenix, Arizona), Desert Ridge, and Downtown Phoenix) is the largest cluster by far. Its neighborhoods mostly have Pizza Places, American Restaurants, and Coffee Shops.

Cluster 2 (South Phoenix and Maryvale) has Mexican Food, Convenience Stores, and Fried Chicken Joints in common.

Cluster 3 (Arcadia (Phoenix)) only had one neighborhood. This neighborhood was very hotel/resort focused, but had parks as well.

Cluster 4: North/Northwest Phoenix This cluster also had one neighborhood. It had a hardware store, shipping store, and Construct & Landscaping services in its top 10.

Discussion

It was unexpected that Downtown Phoenix would be one of the lowest ranking neighborhoods in terms of accessibility. Out of the venue categories that ACME Realty was interested in, it only had convenience stores within walking distance, which means that the 50+ year old homebuyer would need to drive to fulfill most of their needs. Surprisingly, neighborhoods on the edges of Phoenix’s city boundaries like Maryvale, Arcadia, and Moon Valley, had better accessibility ranks.

Overall, the clusters were useful in identifying neighborhoods that were similar to one another. This data could be utilized by the realtor to come up with suitable descriptions for what makes the neighborhoods unique. An example of what the realtor could say would be, “North/Northwest Phoenix (Cluster 4) is unique because it has a hardware store, shipping store, and construction & landscaping services within walking distance. This area is less accessible than most in terms of grocery, but it has a convenient pharmacy and park, and is overall a better choice for do-it-yourselfers.”

Conclusion

In conclusion, ACME Realty has been provided with data that will help them build a data-driven understanding of each neighborhood, along with the common characteristics that it shares with other similar neighborhoods. Every homebuyer’s needs and preferences are unique, so this data will be especially useful in helping to guide them along their home buying journey. The accessible desirable venues rank, combined with a table showing the top 10 venue categories for each neighborhood, and the list of cluster descriptions, should equip ACME and its real estate agents with the insights that they need to match their clients up with the right home in the right neighborhood.

Sources

List of neighborhoods in Phoenix, AZ: <https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Phoenix,_Arizona>

Foursquare API Venue Categories

<https://developer.foursquare.com/docs/build-with-foursquare/categories>

Foursquare API

<https://developer.foursquare.com/>

Google Maps (for Central Avenue Corridor, Downtown, and North/Northwest Phoenix’s coordinates)

<https://www.google.com/maps/place/Phoenix,+AZ/@33.6050991,-112.4052392,10z/data=!4m5!3m4!1s0x872b12ed50a179cb:0x8c69c7f8354a1bac!8m2!3d33.4483771!4d-112.0740373>

Geopy (Phoenix neighborhood coordinates)

<https://anaconda.org/conda-forge/geopy>

Scikit-Learn (K-means clustering algorithm)

<https://anaconda.org/anaconda/scikit-learn>

Folium (Maps)

<https://anaconda.org/conda-forge/folium>