Destination USA: Where to live as a new immigrant in USA

By. Ernest K. Kwegyir-Afful

1.0 INTRODUCTION/BUSINESS PROBLEM

The United States of America is a country established by an immigrant majority. Every year a number of immigrants become new U.S citizens[1]. For example, in the 2018 fiscal year 163,000 immigrants became new citizens[2]. The U.S also runs a Diversity lottery visa program that officially brings immigrants into the country to fill needed specific job categories where skills sets are lacking in the current population. The diversity visa program makes available 50,000 immigrant visas every year across the globe[3]. To diversify the immigrant population, this program selects applicants from countries with low numbers of immigrants in the previous five years. Some of these immigrants migrate to the U.S without having relatives that they can associate with. The primary source of acclimating to the new environment is to find communities that have immigrants of a similar background or at the very least find a diverse community that is most likely to be tolerant of new immigrants.

The goal of this project is to develop an analysis that compares neighborhoods in different U.S cities to identify those with a cultural composition that will be attractive to a specific group of immigrants with a similar cultural background. New immigrants may then use this to assist in their choice of where to live as they move into a new life.

2.0 DATA

We make the assumption that neighborhoods with specific cultural diversity will have a lot more cuisine catered to that cultural background. Additionally, we will expect specialty grocery shops that cater to these specific cultural backgrounds. The more culturally diverse a neighborhood, the more diverse restaurants and grocery shops there will be. Neighborhoods with a predominant cultural background will feature cuisines and grocery from said background. The data to be used for this project will be Foursquare location data to identify clusters of restaurants and grocery shops in specific neighborhoods. For the purposes of this project we will use neighborhoods in Des Moines, Iowa, Baltimore, Maryland and Washington DC. We will select neighborhoods from these regions by pulling data from the following websites:

Des Moines, Iowa: https://www.areavibes.com/des+moines-ia/neighborhoods/

Baltimore, Maryland: https://en.wikipedia.org/wiki/List_of_Baltimore_neighborhoods

Washington, DC: https://en.wikipedia.org/wiki/Neighborhoods_in_Washington,_D.C.

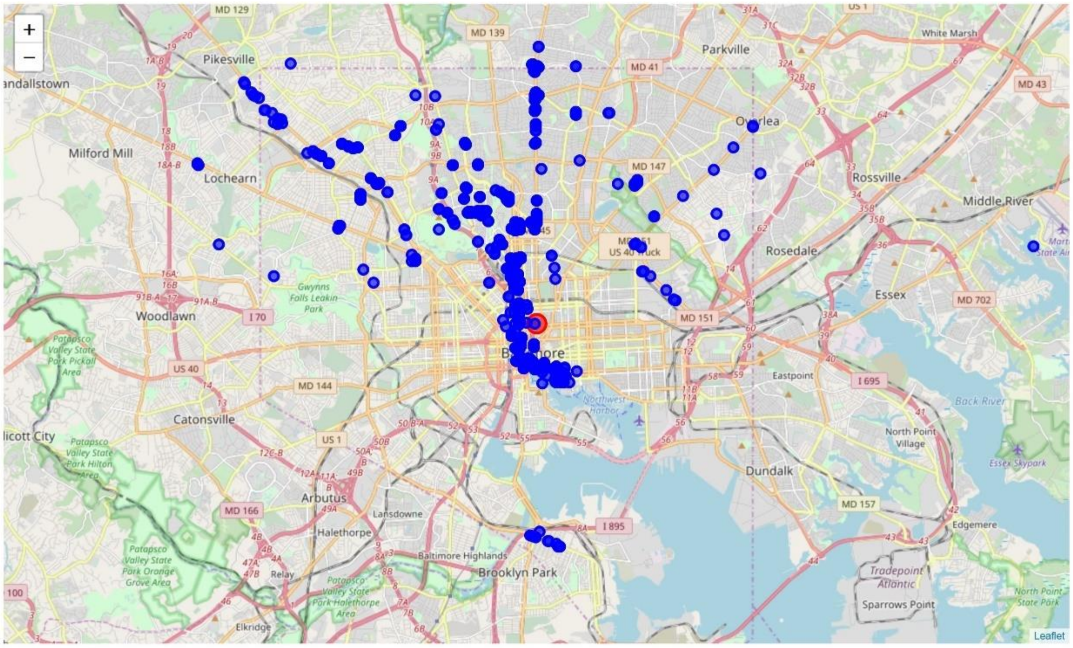
To make the data more manageable, we will select a maximum 100 neighborhoods from each of these cities to compare. We will use the python geocoder to request the geographic coordinates for these neighborhoods. It is our experience from the previous assignment that sometimes the geocoder does not return any results for some locations. We use an API call to Foursquare to return venues that are located within these neighborhoods. We then filter the results to retain only venues for restaurants, grocery shops and other eateries (see Methodology).

From these websites, we created a CSV file that contained the neighborhoods from the different cities. Baltimore has 308 neighborhoods, Des Moines has 53 neighborhoods and Washington DC has 138 neighborhoods. For Baltimore and Washington, DC we selected 100 neighborhoods for our analysis. Since Des Moines has only 53 neighborhoods, we used all 53 neighborhoods. The geocoder returned coordinates for 81 Baltimore neighborhoods, 24 neighborhoods for Des Moines and 98 neighborhoods for DC.

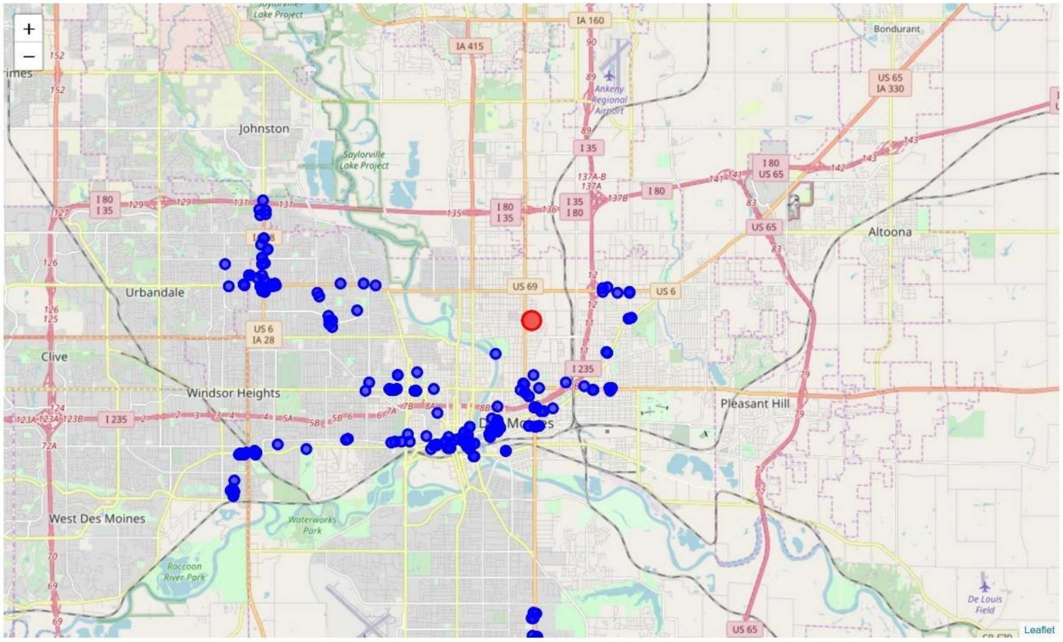
3.0 METHODOLOGY

With the prior knowledge that not all neighborhoods may return a set of coordinates from geocoder, I visualized the datasets that were returned to determine if any neighborhood didn't return a set of coordinates. Viewing up to 20 records at a time to identify records with no coordinates. I then set my code to replace locations coordinates with NA for locations for which there were no coordinates returned. This made it easier to eliminate these locations from further processing. Locations with coordinates were then passed to the Foursquare API. To understand the types of venue categories Foursquare was returning, we once again viewed the different records, viewing up to 20 records at the time. This allowed us to understand what categorization to use for locations that were eateries or grocery shops. We then filtered the data returned from Foursquare based on key words in the venue category that were identified. We used these keywords: Restaurant, Market, Pizza, Food and Sandwich to identify our locations of interest. For our Foursquare API call, we looked within a radius of a 1000 meters from our location coordinates and returned 300 venues. We chose 1000 meters to minimize the chances of overlap that will cause the return of the same venues from adjacent neighborhoods. Prior exploration of our data had shown that some duplicate venues were returned for adjacent neighborhoods. We thus eliminated any duplicates from our data sources. We explored the data to see how many unique categories were returned for each city. For Baltimore, 249 number of unique categories were returned. 190 unique categories and 323 unique categories were returned for Des Moines and Washington DC respectively. We also visualized the venues within each city to understand the spread of the communities (Figure 1 a, b, c)

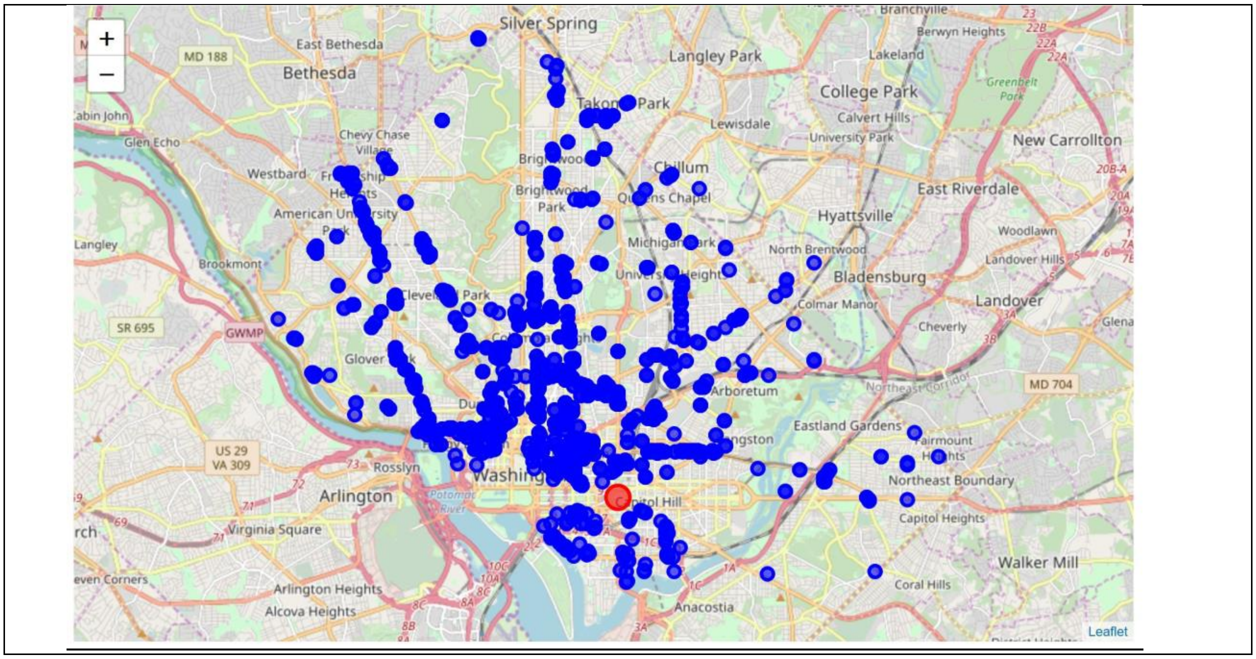
A. Map of Baltimore venues



B. Map of Des Moines, Iowa venues



C. Map of Washington DC venues



We used k-means clustering to generate clusters of the different neighborhoods based on the venues returned by Foursquare. The clusters were thus based on restaurants or eateries in these neighborhoods. K-means provides a relatively simple algorithm to cluster data in an unsupervised manner and is one of the more popular unsupervised machine learning algorithms. Prior to clustering dummy variables were created for all the venue categories returned for each city. The k-means algorithm requires us to predetermine the number of clusters in the data. To determine the optimum k clusters, we used the "elbow" method[4]. The Elbow method is an approach that computes the sum of squared distances (SSD) from each point to its assigned cluster centroid. Since the K-means algorithm seeks to reduce the SSD within each cluster, as we increase the number of clusters, the SSD also decreases till a point (i.e. cluster number) where the rate of decrease of the SSD is minimal compared to the increase in cluster number. When we plot the number of clusters (k) against the SSD, we can find the point where this occurs. This is the elbow, this is then chosen as the optimum number of clusters. In our case, for most of the cities we chose, there wasn't a clear elbow, rather there was a relatively smooth curve. In such cases we also employed the use of silhouette analysis to identify the optimum k. Silhouette analysis is used to study the separation distance between resulting clusters from a K-means clustering analysis[5]. The silhouette plot displays a measure of how close each point is to points in the neighboring clusters and provides a way to assess the number of clusters visually. Silhouette coefficients range from -1 to 1. Coefficients of +1 indicate that the sample is far away from the neighboring clusters whereas a value of 0 indicates the sample is close to a decision boundary. Negative value indicate that the sample may have been assigned to the wrong cluster. A third metric that can be used identify the optimum number of clusters is the Calinski-Harabaz method[6,7]. This Calinski Harabaz score computes the ratio between within-cluster dispersion and the between-cluster dispersion. The "Elbow" within this plot also denotes the optimum K.

We also explored using Density-Based Clustering (DBSCAN) to assess if it will overcome some of the challenges observed with using K-means clustering. This was because we hypothesized that some clusters may be shaped arbitrarily and that different cultures maybe spread across a variety of neighborhoods.

Density based clustering is able to locate regions of high density separated from each other by regions of low density. Density is defined as the number of data points within a specified radius. However, our analysis of this indicated that the clusters generated by DBSCAN did not answer our question better than those generated by K-means. We will discuss this further in the results using Baltimore as our example.

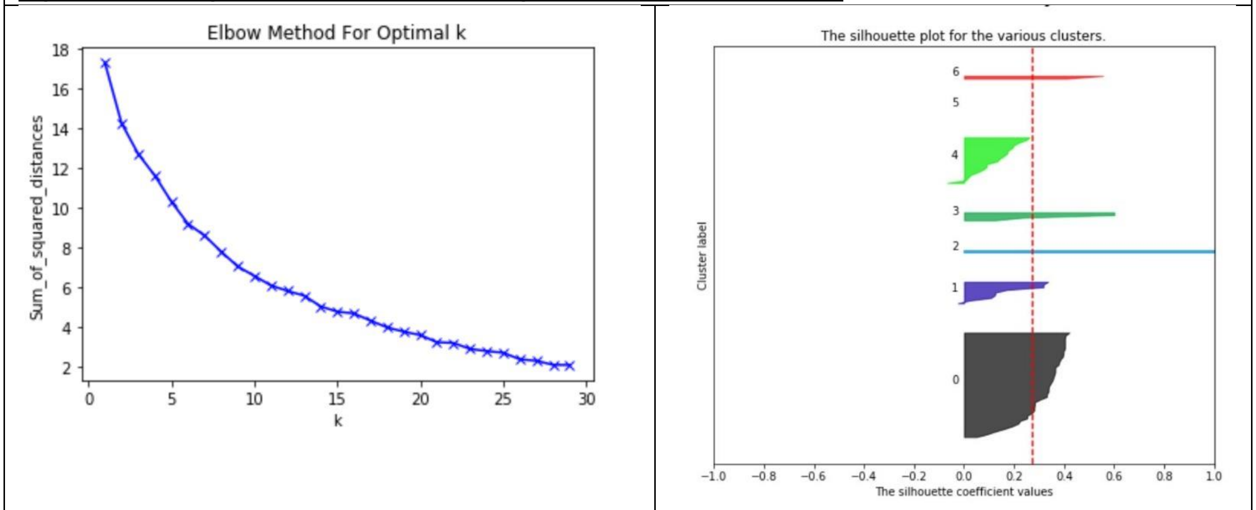
After clustering, we explored the clusters to identify the categories restaurants that were most represented in different neighborhoods. We used the top 10 most venue in each neighborhood to identify the predominant cultural make-up of the community.

4.0 RESULTS

4.1 Baltimore Analysis

We selected 100 neighborhoods from Baltimore and used the python geocoder to return Longitude and Latitude for these neighborhoods. Out of the 100 neighborhoods, coordinates were returned for 81 neighborhoods. We proceeded with the analysis for these 81 neighborhoods. Of the 81 neighborhoods, Foursquare returned 3056 venues. We then filtered this data to obtain only locations for restaurants, grocery shops and general places where people go to buy a meal. This resulted in 875 locations with 247 unique venue categories. We then used K-means to cluster these locations. K-means clustering algorithm requires that we pre-specify the number of clusters we desire for the data. This requires that we explore the data and metrics that may suggest the optimum number of clusters for each data set. To do that we employed one or more of three methods, Elbow, Silhouette or Calinski Harabaz score (see methods for descriptions). To understand the potential number of clusters within the Baltimore data we used the Elbow and Silhouette methods.

Fig. 2 – Selecting K for K-,means clustering for Baltimore venue data



a. Elbow plot to determine K for Baltimore data.

b. Silhouette plot to determine K for Baltimore data.

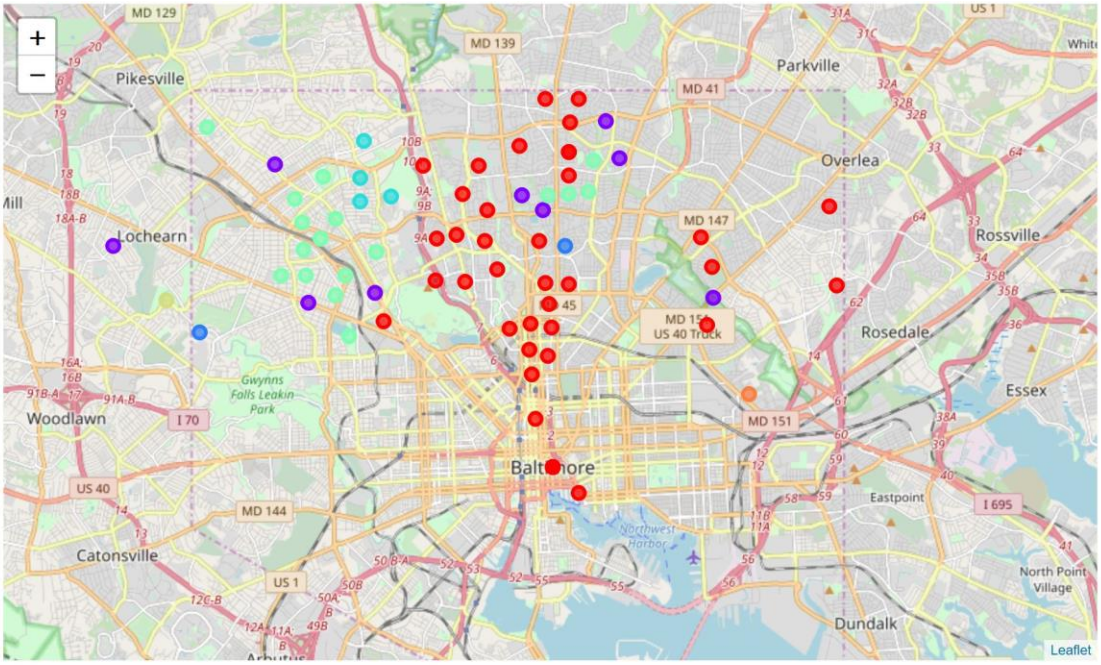
The elbow method for determining K, didn't provide a distinct elbow in the graph for which we could choose the K from. While the change in SSD for subsequent Ks decreased after 7 clusters, there wasn't a distinct elbow that could be determined from the graph. We then did a silhouette analysis. Table 1 shows the scores for the Silhouette analysis for 15 clusters. Cluster 7 gives the highest coefficient indicating that the best separation of clusters occurs when we have 7 clusters.

Table 1 – Coefficient Scores for Silhouette analysis of Baltimore venue data

No. of Clusters (K)	Score
2	0.198496
3	0.218676
4	0.241563
5	0.218919
6	0.222749
7	0.274201
8	0.185495
9	0.180894
10	0.20338
11	0.199456
12	0.184712
13	0.215188
14	0.212285
15	0.208928

Figure 3 shows the 7 clusters for Baltimore displayed on a geographic map of Baltimore.

Fig. 3 – Clusters of Baltimore Venues

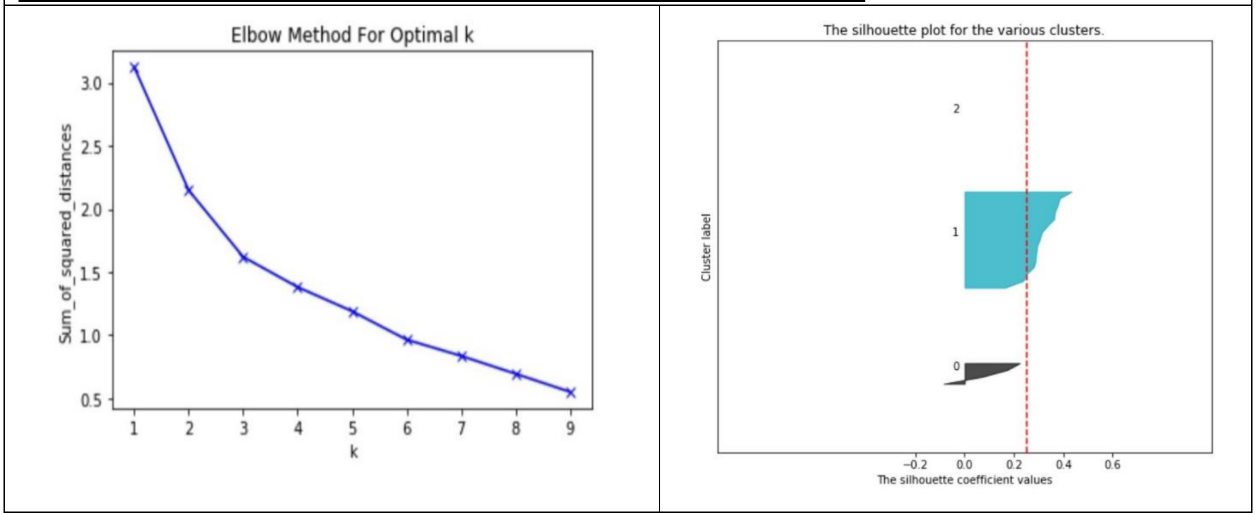


Because the centroids are initialized randomly for K-means, referring to clusters by their cluster number isn't ideal since every run may change the cluster number matching the details for the cluster in terms of cultural diversity. It is however, more appropriate to label the clusters based on their cultural diversity. The first thing one observes from the Baltimore data is that, Baltimore has a very diverse cultural composition. The cluster 0 contains neighborhoods that are very diverse culturally. This includes neighborhoods such as Reisterstown Station, Abell, Barclays, Charles Village, Charles North etc (see python note book - Capstone_Project_Battle_of_the_Neighborhoods_III). This forms the largest cluster and is colored red in the map cluster in Fig 3. The second cluster is a predominantly Asian neighborhood featuring cuisines from China, Vietnam and Japan. This neighborhood also have cuisines from other cultures though not as numerous. These neighborhoods include Central forest park, Glen, Glen Oaks etc. Purnell and Pen Lucy are in the same cluster and they feature predominantly Vietnamese and Japanese cuisine. Cold Spring, Cylburn, Mount Washington and Levindale are predominantly traditional American neighborhoods. Gwynn Oak is a neighborhood that strongly features people of African descent. Cluster 7 in our data was difficult to decipher in terms of cultural background because this cluster features predominantly seafood restaurants. This cluster contain the neighborhoods for Armistead Gardens and Wilson Park. This neighborhood also features other cultures such as Vietnamese, Italian, French and Indian.

4.2 Des Moines Analysis

In the case of Des Moines, there were only 53 neighborhoods identified for the city. Out of the 53 neighborhoods, coordinates were returned for 24 neighborhoods. We proceeded with the analysis for these 24 neighborhoods. Of the 24 neighborhoods, Foursquare returned 905 venues. We then filtered this data to obtain only locations for restaurants, grocery shops and general places where people go to buy a meal. This resulted in 247 locations with 190 unique venue categories. We then used K-means to cluster these locations. We used three methods, Elbow, Silhouette or Calinski Harabaz score (see methods for descriptions) to estimate the optimum K.

Fig. 3 – Selecting K for K-,means clustering for Des Moines venue data



a. Elbow plot to determine K for Des Moines data.

b. Silhouette plot 3 clusters for Des Moines data.

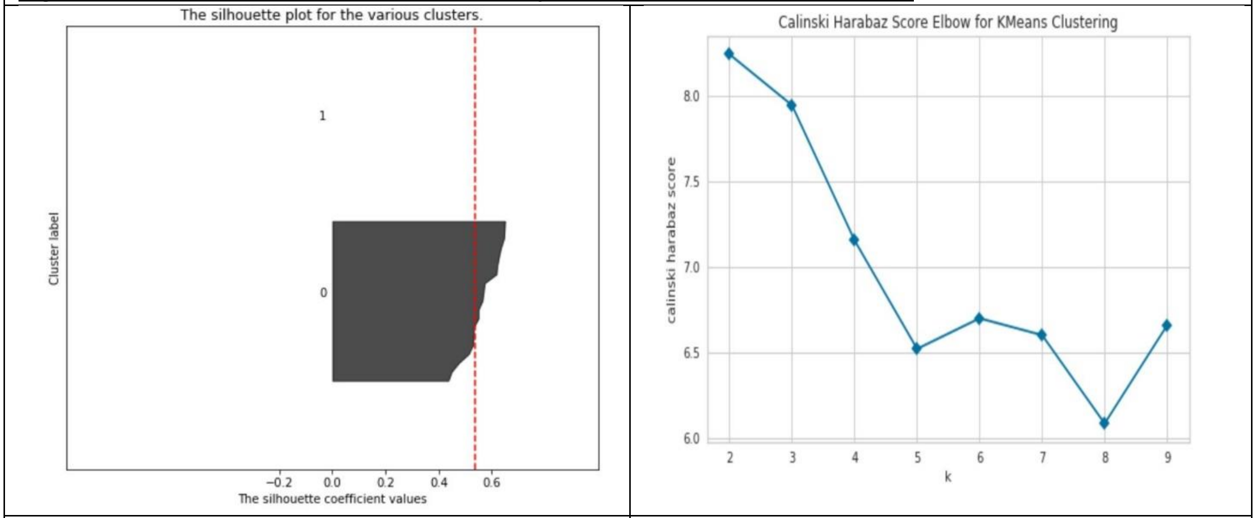
Similar to the Baltimore data, the Elbow plot did not provide us with a clear-cut elbow to determine the optimum number of K. This seem to indicate that the neighborhoods were more uniform. We employed the Silhouette score to determine if that gave us better insights into the potential number of clusters in the neighborhood. The Silhouette scores indicate that we may have 2 or at most three clusters in the Des Moines data. Table II shows us the scores for the Silhouette analysis for Des Moines. When the data is divided into 2 clusters is when we have the most separation with a coefficient of 0.54 while three clusters gave us a coefficient of 0.25.

Table II – <u>Coefficient Scores for Silhouette analysis Des Moines venue data</u>

No. of Clusters (K)	Score
2	0.53544
3	0.252705
4	0.140413
5	0.142071
6	0.132535
7	0.113198
8	0.128956
9	0.143794
10	0.134374

Further analysis with the Calinski-Harabaz analysis didn’t provide any clear insights as to whether two or three clusters was the better analysis for the Des Moines data. Figure 4a and 4b show the Silhouette for 2 clusters and the graph for the Calinski-Harabaz score. The Calinski-Harabaz score doesn’t show a clear elbow, rather, it shows two elbows and an upward turn of the graph.

<u>Fig. 4 Silhouette and Calinski-Harabaz analysis for Des Moines venue data</u>

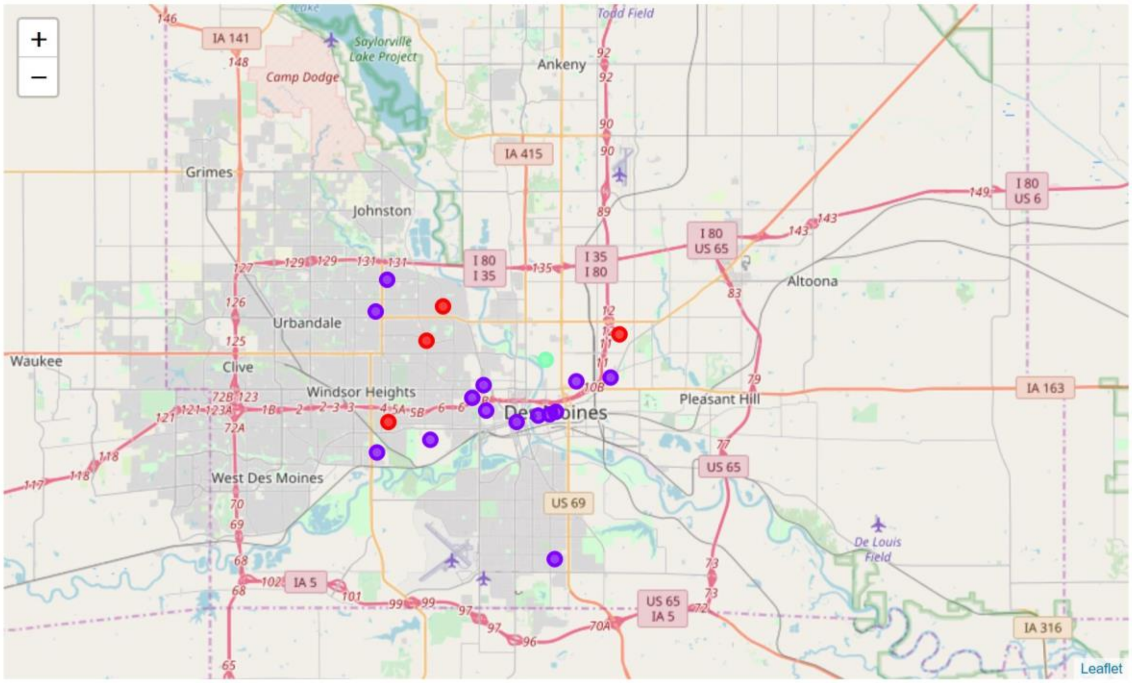


a. Silhouette graph for 2 clusters for Des Moines data

b. Calinski-Harabaz score graph for Des Moines data

Examining the clusters for Des Moines, after we split into 3 clusters indicates that the optimum K will be 2 clusters. The third cluster, Union Park, should be added to the second cluster that has the major migrant communities. The first cluster has a larger density of traditional American cuisines, however, there are also other cultural backgrounds in these neighborhoods. Figure 5 shows the venues plotted on a geographical map of Des Moines. The green dot on the graph indicates the third cluster which is the neighborhood of Union Park. It is easy to see why this should be included in the purple dots (second cluster in this case).

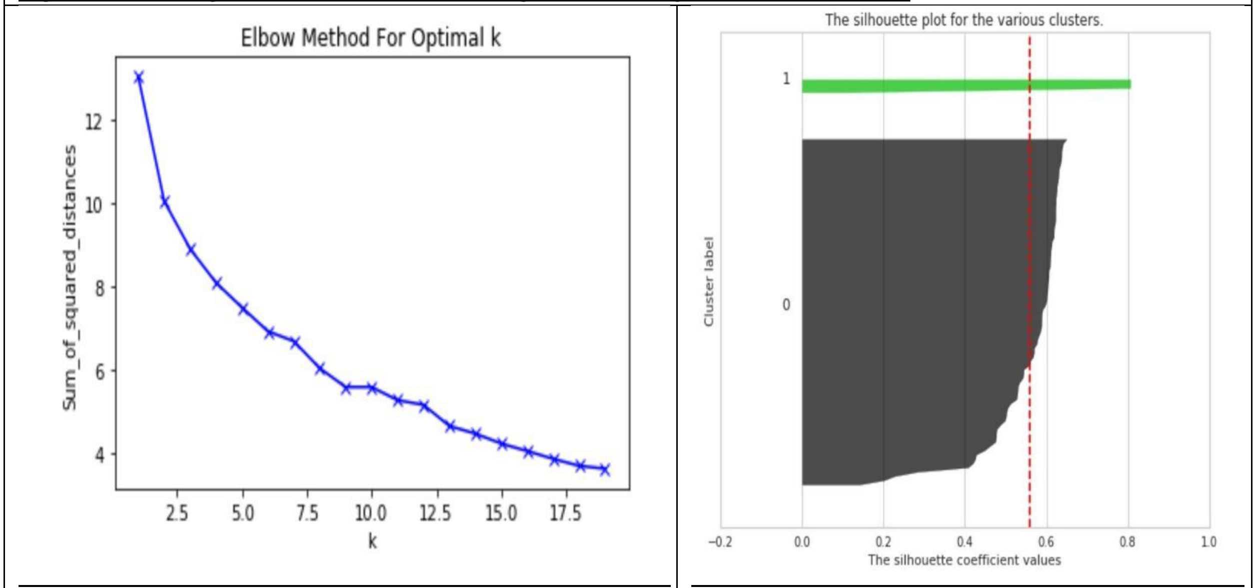
Fig. 5 – Clusters of Des Moines Venues



4.3 Washington DC Analysis

Out of 138 neighborhoods in Washington DC, we selected 100 for our analysis. For the 100 neighborhoods, Python's geocoder returned coordinates for 92. We used Foursquare to extract 5448 venues of which we filtered 1749 as places that fit our criteria for eateries (see methods above). Out of this number there were 323 unique categories. Figure 1c shows the representation of the DC venues superimposed on a DC geographical map. One will observe that the majority of the venues are clustered in the DC downtown area. This similar across all three cities analyzed. We similarly sort to determine the optimum number of clusters before doing the k-means analysis. We used all three methods to determine our optimum K. Similar to the data for Baltimore and Des Moines, the Elbow plot was not a good indicator of the number of clusters in the DC venue data (Fig 6a). The silhouette analysis indicates that a cluster number of 2 with a coefficient of 0.56 may be best for the DC data (Fig 6b and Table III).

Fig. 6 – Selecting K for K-,means clustering for Washington DC venue data



a. Elbow plot for clusters in DC venue data

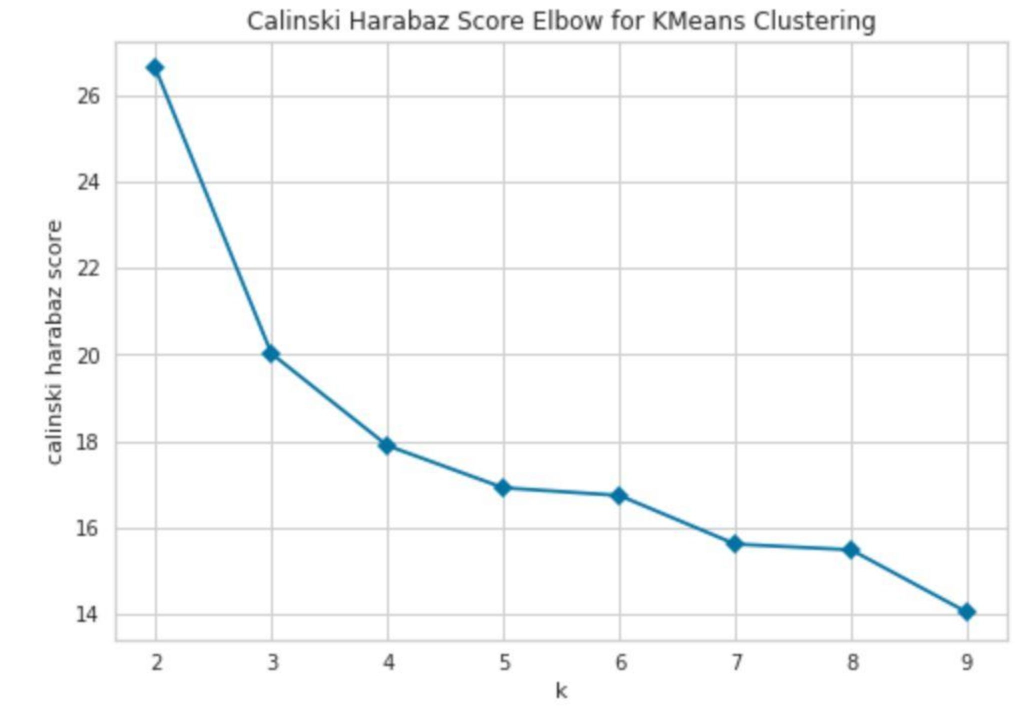
b. Silhouette plot showing 2 clusters for DC data

Table III – Silhouette scores for Washington DC data venue data.

No. of Clusters (K)	Score
2	0.558324
3	0.248493
4	0.269752
5	0.275305
6	0.127688
7	0.092125
8	0.106511
9	0.090174
10	0.091433

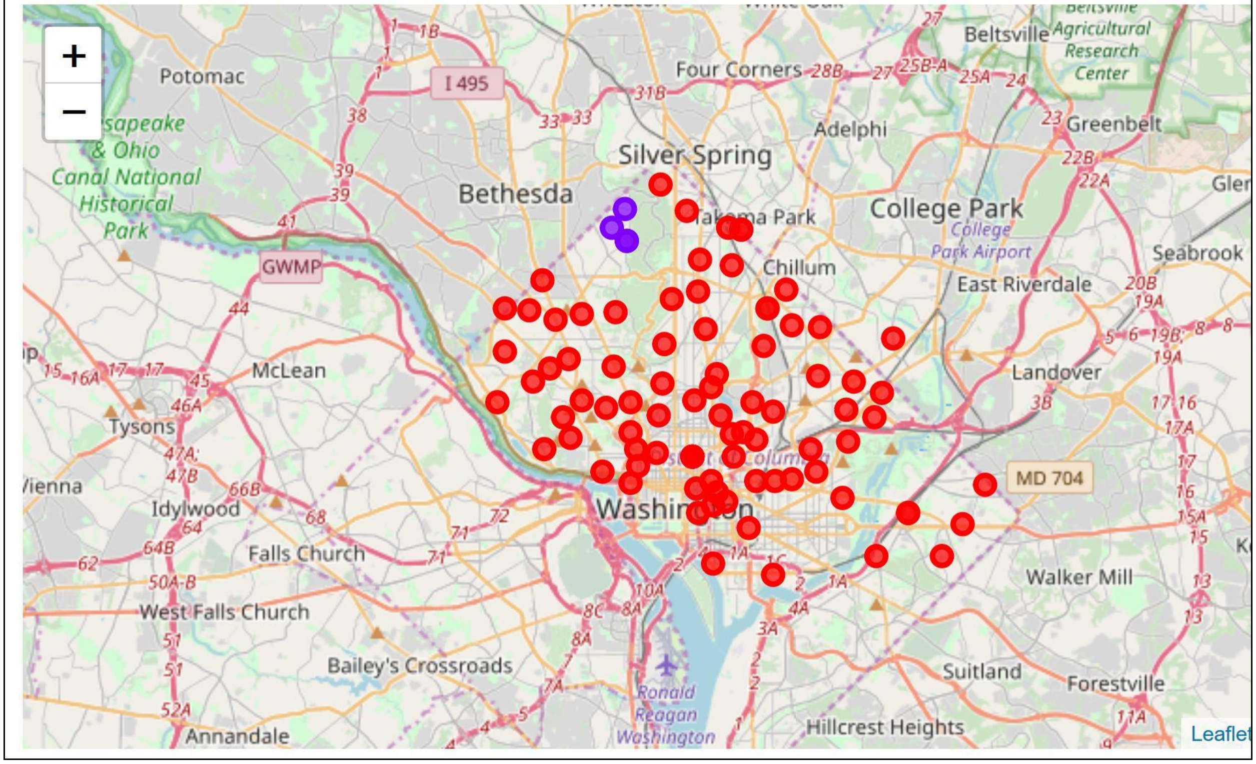
We explored the data for more insights with regards to the optimum number of clusters by doing the Calinski-Harabaz analysis. The Calinski-Harabaz score did not indicate an elbow when plotted against an increasing number of clusters (see Fig. 7).

Figure 7. Calinski-Harabaz Elbow plot for Washington DC venue data



Clustering analysis of the DC data show the largest cluster (Fig. 8) includes people of diverse cultural background. This cluster highlights the diversity of DC neighborhoods. Even in neighborhoods where the most restaurants are traditional American, there is also a sizeable representation of cuisines from other cultures. The second cluster, which is much smaller, includes neighborhoods such as Chevy Chase, Colony Hill, Potomac Heights, Barnaby Woods and Hawthorne. These neighborhoods are distinct in that the are traditionally American but have a good representation of Asian and African immigrants.

Fig. 8 Clusters of Washington DC venues.



5.0 DISCUSSION

This study was designed to provide an analytical process for new immigrants to choose their places of domicile when they first migrate to the United States of America. We purposely included three cities, Baltimore, Maryland, Washington DC and Des Moines, Iowa. This was because we expected some cities to be more diverse than others. We also expected new immigrants will more likely move to the big cities rather than towns that they haven't heard about before. We started this analysis with the assumption that neighborhoods with diverse cultural compositions will have restaurants and grocery shops that cater distinctly to these cultures. We also expected some neighborhoods to have a certain level of diversity. Thus, neighborhoods that generally welcome immigrants will tend to have more immigrant population regardless of their countries of origin/birth. We also expected people from the same cultural background will gravitate towards each other so there will be neighborhoods with a predominance of certain cultural backgrounds.

Exploring the data from all three cities, we realized that Washington DC, has the most diversity, followed by Baltimore and then Des Moines. Based on the number of restaurants returned by Foursquare it becomes obvious that Washington DC and Baltimore are bigger cities and tend to attract people from different cultural backgrounds compared to Des Moines. That Des Moines is a smaller city maybe evident from the number of neighborhoods, however, a smaller number of neighborhoods doesn't necessarily equate to a smaller city. One possibility is to have a smaller number of neighborhoods but have them densely populated or have a much bigger landmass per neighborhood. A good contrast is to compare Baltimore and Washington DC. Whereas Baltimore had more neighborhoods, DC returned the most venues that we were interested in. With Baltimore having more neighborhoods, it was also the one that could be separated into the most clusters (7) even though it had a smaller number of venues than DC. Washington DC had the highest number of venues and upon looking at how dense these venues are clustered within the neighborhoods (see Fig. 1c), it isn't surprising that we had a smaller number of clusters. This is because different types of cuisines appeared clustered within the same neighborhoods underscoring the cultural diversity of DC neighborhoods. Des Moines, Iowa had the least cultural diverse neighborhoods, however there was still a good representation of cuisines from differently cultural backgrounds even though they didn't appear with the same density as Baltimore or DC. We explored using Density Based clustering to assess if it could overcome the shortcomings of K-means. However, we realized that given the density of clustering of cuisines within the same neighborhoods, DBSCAN did not give up a better clustering path than was available with K-means. This is because we could draw an arbitrarily shaped cluster for a particular cuisine that spans over multiple neighborhoods. What became quickly apparent was that we were just slicing up the neighborhoods as though they were different neighborhoods rather than representing them as the truly diverse neighborhoods that they are.

We believe this analysis provides a crude approximation of the cultural diversity of various cities that we looked at. To make this a truly robust analysis and a tool for new migrants to use, we will have to consider including U.S censors' data as that will provide a concrete representation of where migrants actually live. One observation we made was that the majority of the restaurants are clustered in what will be considered the downtown areas of these cities. This is because most people will find it comfortable to travel to these downtown areas for lunches and dinner dates. So, while a neighborhood might have diverse cuisines, the migrants may not actually live in the immediate neighborhood but could be living in the bordering neighborhoods that do not show clusters of the restaurants that we are interested in. Adding the U.S censor's data will address this limitation. Additionally, is important to estimate the cost of living for the different neighborhoods to make this a truly useful tool for new migrants. New migrants typically are not wealthy and may be looking for a place where the cost of living is reasonable. There will be trade-offs between a truly culturally diverse neighborhood and the cost of living in that neighborhood. Applying these new metrics to our cluster analysis will provide a more meaningful tool that new immigrants can use in choosing their first place of domicile on migrating to the United States of America.

6.0 CONCLUSION

In doing this analysis we explored different concepts in unsupervised machine learning including what algorithm to use and how to choose the optimum number of clusters. We can use the analysis as-is to determine which neighborhoods are more likely to be accepting of new immigrants from either similar backgrounds or different backgrounds. However, we will suggest making this analysis more robust and useful by including data from the U.S Census Bureau and also data on the cost of living in these neighborhoods. In order to identify neighborhoods with specific cultural backgrounds, we will suggest filtering the clusters to include only areas with the required background represented in the top 5 venues. Other top N can be considered depending on the results of the top 5 as the less popular areas will indicate less of that cultural background. From our analysis we conclude that Washington DC has the most culturally diverse population, followed by Baltimore and then Des Moines.

7.0 REFERENCES

1. https://en.wikipedia.org/wiki/Immigration_to_the_United_States

2. http://immigrationimpact.com/2018/07/03/citizens-naturalization-ceremonies-fourth-july/

3. https://en.wikipedia.org/wiki/Diversity_Immigrant_Visa

4. https://en.wikipedia.org/wiki/Elbow_method_(clustering)

5. https://en.wikipedia.org/wiki/Silhouette_(clustering)

6. T. Calinski and J. Harabasz, 1974. “A dendrite method for cluster analysis”. Communications in Statistics

7. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabaz_score.html