Best Locations for a Gourmet Coffee Shop in Apex, NC, USA

Jason Wise

January 2021

1.0 Introduction

1.1 Background & Problem

My hometown of Apex is a growing suburb located outside of Raleigh, NC. Its population is around 60,000. While it has many of the amenities that are attractive for suburban residents, it seems to be lacking in access to high quality coffee shops. There are plenty of places to grab an average beverage, but few establishments that cater to gourmet coffee connoisseurs.

The market for coffee drinkers is strong. Approximately 63% of Americans are coffee drinkers. Of those consumers, around 61% consume "gourmet" coffee[1]. Additionally, around 36% of the population consumes coffee brewed outside of the home[2]. In general, these statistics have an upward trend[3].

1.2 Target Audience

For this project, I am working with a fictional entrepreneur that wishes to open a Gourmet Coffee Shop in Apex, NC. While they have strong knowledge of the coffee shop market and some hunches about the best locale, they wish to base their business decision on more than just intuition. They have hired me to perform a data based analysis to determine a list of the top 5 locations for a Gourmet Coffee Shop in Apex.

References:

1. https://www.washingtonpost.com/news/voraciously/wp/2019/03/28/americas-growing-affection-for-gourmet-coffee-and-other-takeaways-from-a-new-national-survey/
2. https://dailycoffeenews.com/2018/03/21/current-coffee-consumer-trends-inside-the-ncas-2018-report/
3. https://arctoscoffee.com/coffee-trends-show-increase-in-gourmet-and-youth-consumption/

2.0 Data

Based on my analysis, I have chosen to acquire neighborhood and related demographic data for around 70 neighborhoods in Apex, NC that will be used as features within my model. My project plan is to:

- Define the problem and the data required for a solution
- Acquire the needed data and perform analysis
- Explore the makeup of Apex, NC neighborhoods
- Analyze each Neighborhood
- Cluster Neighborhoods using k-means clustering
- Examine Clusters for client recommendations

The following data sources will be used to perform a data based analysis to determine the best locations for a Gourmet Coffee Shop located in Apex, NC.

List of Apex, NC Subdivisions and Related Property Sales Figures
http://triangleareaarealty.com/apex-nc-subdivisions/
I will scrape this website to pull the approximately 70 different neighborhoods and their associated average sales price of homes. This data will be used to segment Apex neighborhoods according to subdivision. Additionally, average home sales prices by neighborhood will represent a demographic feature that will contribute to our analysis

GPS Coordinates for Apex, NC Subdivisions - Google Maps Apex, NC
https://www.google.com/maps/place/Apex,+NC/@35.7275871,-78.8999849,13z/data=!3m1!4b1!4m5!3m4!1s0x89ac92a3c19280d1:0x85cd817e17e28015!8m2!3d35.732652!4d-78.8502856
I was not able to readily find this data in a form that could be imported. I decided to manually compile the latitude and longitude values for all Apex neighborhoods via Google Maps and add them to a CSV file that will be imported to a dataframe and merged to the list of subdivisions. The GPS coordinates will allow for data visualization and k-means clustering according to related features.

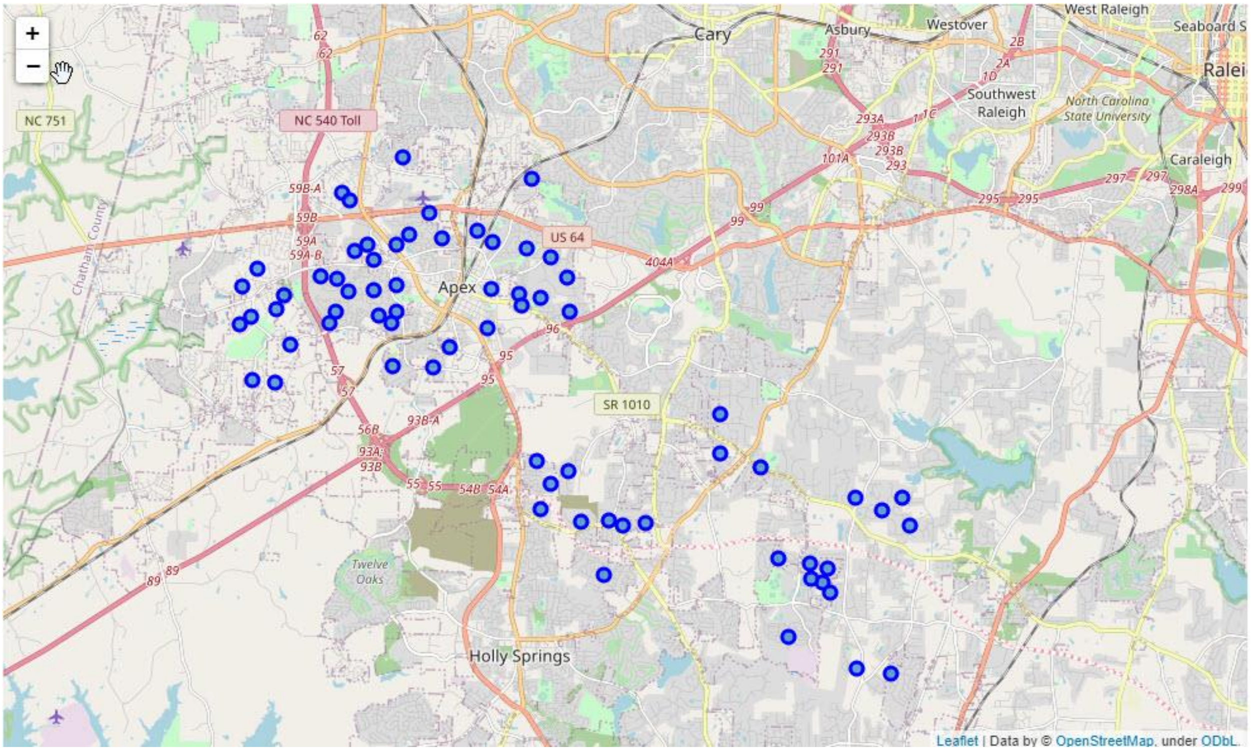
Subdivision Demographic Information
http://www.city-data.com/nbmaps/neigh-Apex-North-Carolina.html
Also for this category, I was not able to readily find this data in a form that could be imported. I decided to manually compile Median Household Income, Population and Median Resident Age and store them to the CSV file of Apex neighborhoods that will be imported into the notebook. While this data is visualized and made available on the City-Data website, it is sourced from the United States Census data. This demographic data will be aligned to the list of neighborhoods and used as features in our model to determine the best location for a coffee shop.

Foursquare Venues and Categories
I will utilize the Foursquare Places API to conduct a comparison of neighborhoods that may support a coffee shop. I will look at complimentary businesses in the area as well as competing businesses.

3.0 Methodology

3.1 Map Neighborhoods to Visualize with Folium

After acquisition of data, data cleansing and formatting. I proceeded with a map of all neighborhoods using the Folium library. The allowed for a visualization of the distribution of all Apex neighborhoods. From this distribution you can see how they lie in comparison to major venues, attractions and highways.



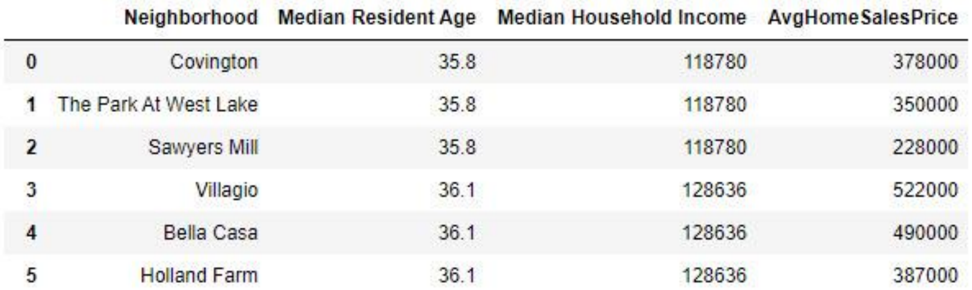
3.2 Filter the Dataset to our Base Demographic Market

An examination of the demographic features showed that population would not be useful due to differences between the geographic shape of US Census Data versus the shapes of each individual neighborhood. Therefore, this feature was dropped.

Further analysis of the dataset revealed that several neighborhoods were clearly outside of the target market of our client. To sharpen our approach, we filtered out several neighborhoods based on our client's base market demographics being:

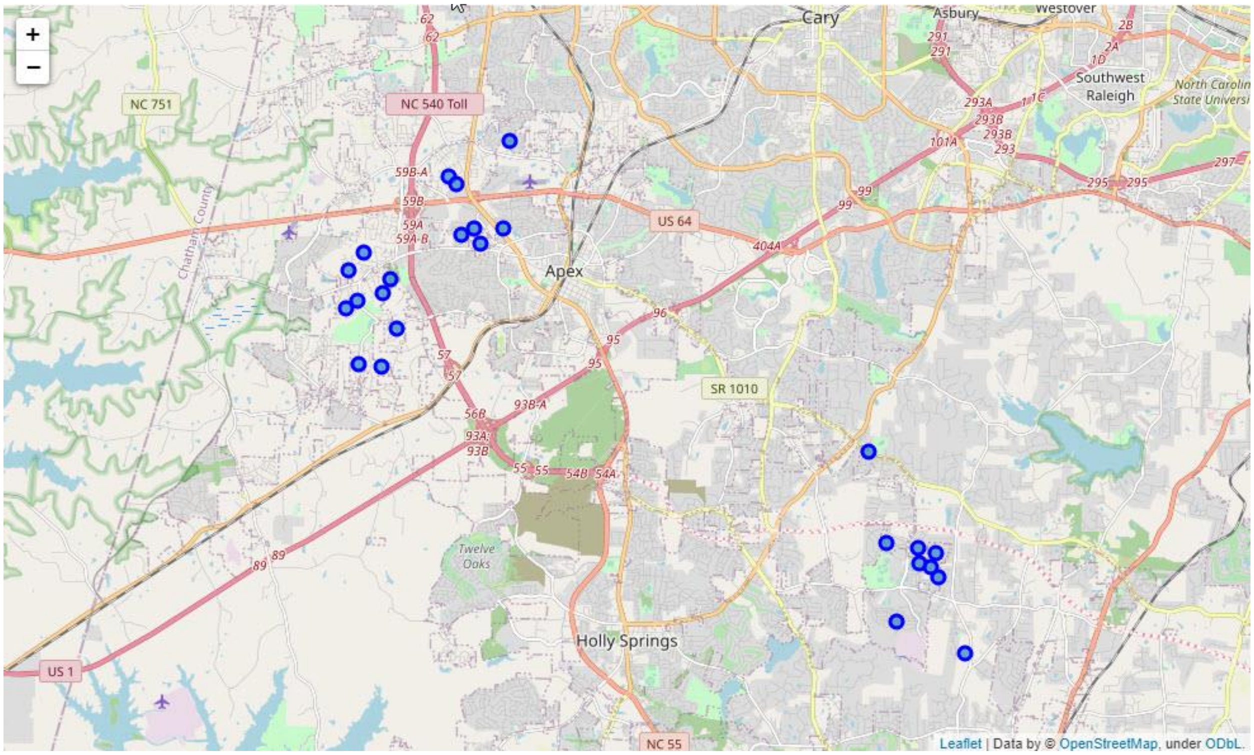
1. Age – Median Resident Age < 40 Years
2. Income – Median Household Income Above the Mean for our Population (All Apex Neighborhoods)

Filtering out neighborhoods reduced our list to 23 base target market representative neighborhoods.



	Neighborhood	Median Resident Age	Median Household Income	AvgHomeSalesPrice
0	Covington	35.8	118780	378000
1	The Park At West Lake	35.8	118780	350000
2	Sawyers Mill	35.8	118780	228000
3	Villagio	36.1	128636	522000
4	Bella Casa	36.1	128636	490000
5	Holland Farm	36.1	128636	387000

We then visualized again with Folium to see our new distribution of neighborhoods within our base target market.



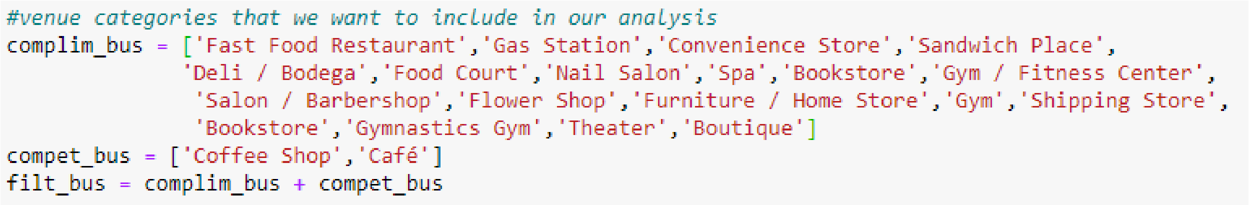
3.3 Get Venue Information from Foursquare

Our next step was to gather local venue information via the Foursquare Places API. This API returns JSON data which is then transformed to a dataframe for our consumption.Our initial run of all venues returned 1,014 venues using a limit of 100 venues per neighborhood and a radius of 1.7 miles.

A review of venues and their respective categories told us that there was a lot of data here that was not useful to us. Rather than include features in our model that are not important to us, we then decided to limit our features to the following:

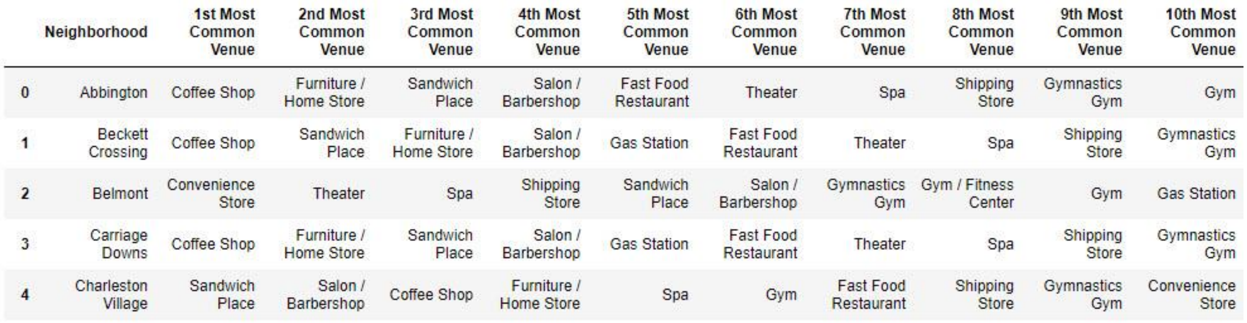
1. Complimentary Businesses – businesses that are deemed favorable by the client and the coffee industry in terms of location for a Gourmet Coffee Shop. It is important to measure the number of these businesses to determine the strength of a potential location.
2. Direct Competing Businesses – cafes and coffee shops that represent direct competition to the client. The client desires to locate their coffee shop in an area with a relatively lower concentration of coffee shops. This will allow us to determine market saturation.

The following lists were used to pare down our list of venues/features for analysis. Filtering pared down our dataset to 209 venues.



```
#venue categories that we want to include in our analysis
complim_bus = ['Fast Food Restaurant','Gas Station','Convenience Store','Sandwich Place',
               'Deli / Bodega','Food Court','Nail Salon','Spa','Bookstore','Gym / Fitness Center',
                'Salon / Barbershop','Flower Shop','Furniture / Home Store','Gym','Shipping Store',
                'Bookstore','Gymnastics Gym','Theater','Boutique']
compet_bus = ['Coffee Shop','Café']
filt_bus = complim_bus + compet_bus
```

Through our Foursquare data, we were able to capture the 10 most common filtered venues by neighborhood

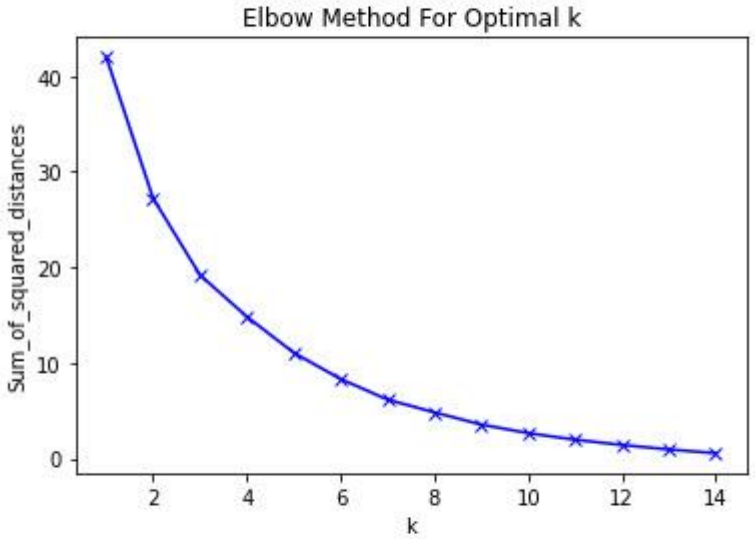


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abbington	Coffee Shop	Furniture / Home Store	Sandwich Place	Salon / Barbershop	Fast Food Restaurant	Theater	Spa	Shipping Store	Gymnastics Gym	Gym
1	Beckett Crossing	Coffee Shop	Sandwich Place	Furniture / Home Store	Salon / Barbershop	Gas Station	Fast Food Restaurant	Theater	Spa	Shipping Store	Gymnastics Gym
2	Belmont	Convenience Store	Theater	Spa	Shipping Store	Sandwich Place	Salon / Barbershop	Gymnastics Gym	Gym / Fitness Center	Gym	Gas Station
3	Carriage Downs	Coffee Shop	Furniture / Home Store	Sandwich Place	Salon / Barbershop	Gas Station	Fast Food Restaurant	Theater	Spa	Shipping Store	Gymnastics Gym
4	Charleston Village	Sandwich Place	Salon / Barbershop	Coffee Shop	Furniture / Home Store	Spa	Gym	Fast Food Restaurant	Shipping Store	Gymnastics Gym	Convenience Store

3.4 Conduct Cluster Analysis using k-means via Scikit Learn

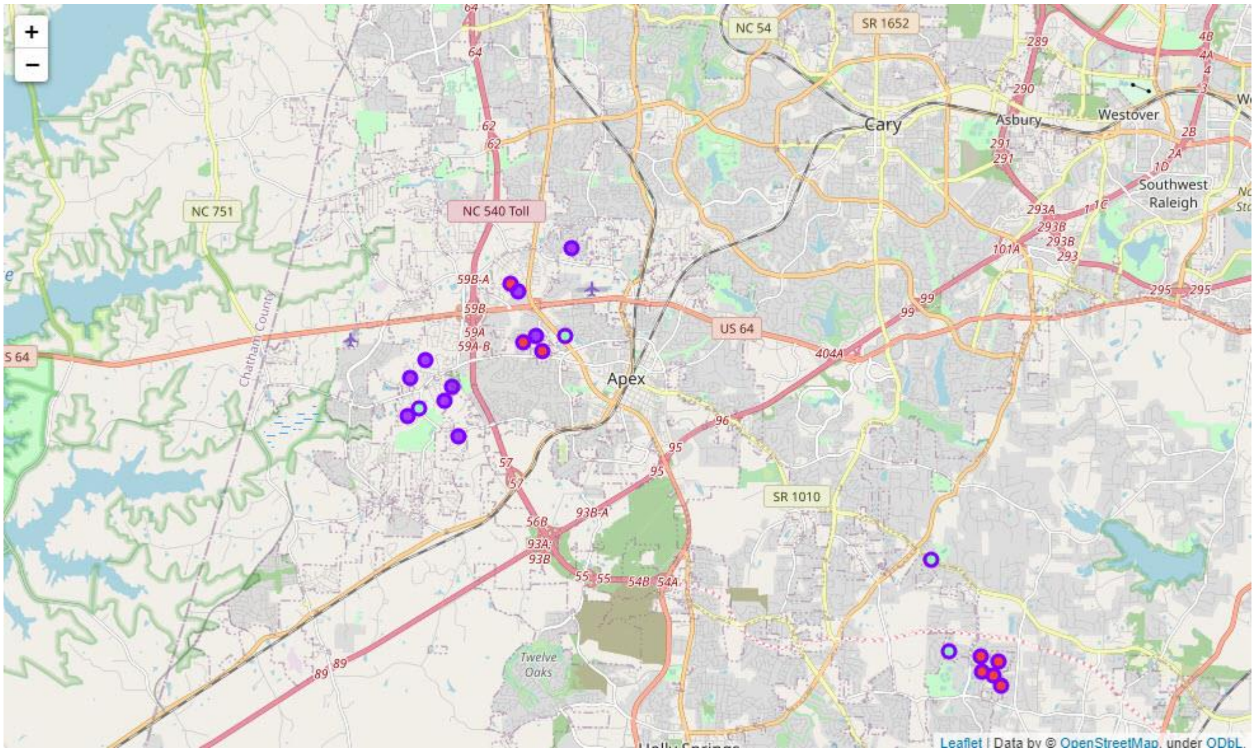
The next step is to segment our data into clusters that represent similar customers. The venue data contained features that were categorical in nature. These are not well suited to analysis, especially clustering. As such, we used the get_dummies function to conduct one hot encoding and break out our categorical data into new columns of 1s and 0s representative of a venue being present or not present near a neighborhood.

The one hot data was combined with our demographics features and then normalized using the MinMaxScaler function from the Scikit Learn library. Normalized data was then plotted for Optimal k as follows:

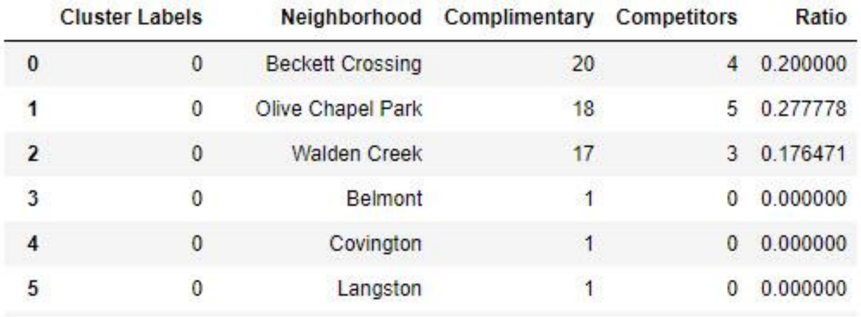


The fact that our graph is more round than sharp is likely due to filtering of data outside of our base demographic. Therefore, we have data that is more similar to each other. We have decided to use a k value of 3 to break our data into three separate segments of customers.

Again, the data was visualized on a map with Folium where you can clearly see the distribution of the three segments.



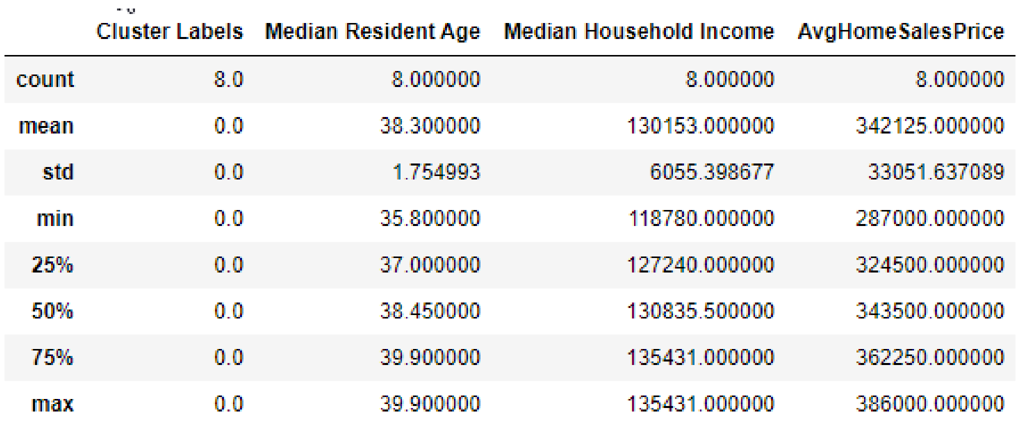
At this point, it is important to understand the differences in each segment to see the meaning behind the data. We took the clustered data by neighborhood and summed up the number of complimentary and competitive businesses. Additionally, measured the ratio of Competitors to Complimentary businesses by neighborhood. These values were added to our dataset to further our analysis of each cluster.



	Cluster Labels	Neighborhood	Complimentary	Competitors	Ratio
0	0	Beckett Crossing	20	4	0.200000
1	0	Olive Chapel Park	18	5	0.277778
2	0	Walden Creek	17	3	0.176471
3	0	Belmont	1	0	0.000000
4	0	Covington	1	0	0.000000
5	0	Langston	1	0	0.000000

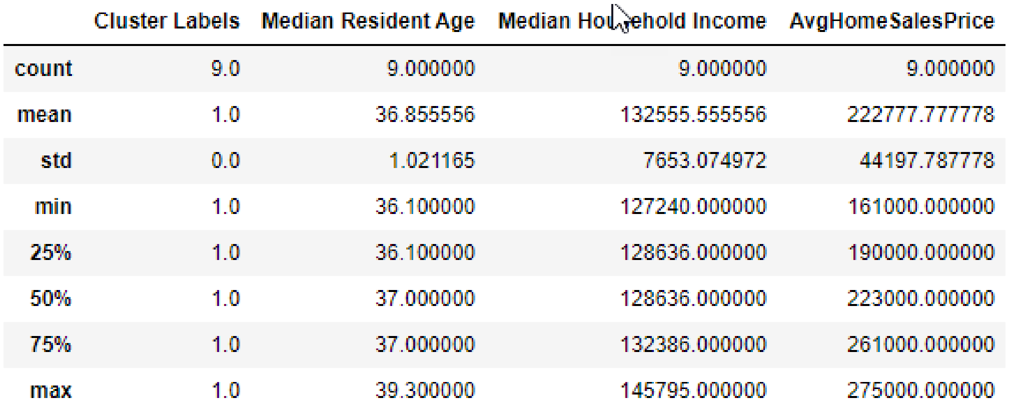
Our next approach was to take the cluster labeled data and slice so that we can do some data analysis and determine labels for our three segments. We added demographic data back in and ran descriptive statistics on our segments.

From cluster 0, we see that this is our Oldest Aged group, Lowest Income Group and Mid Home Values group.



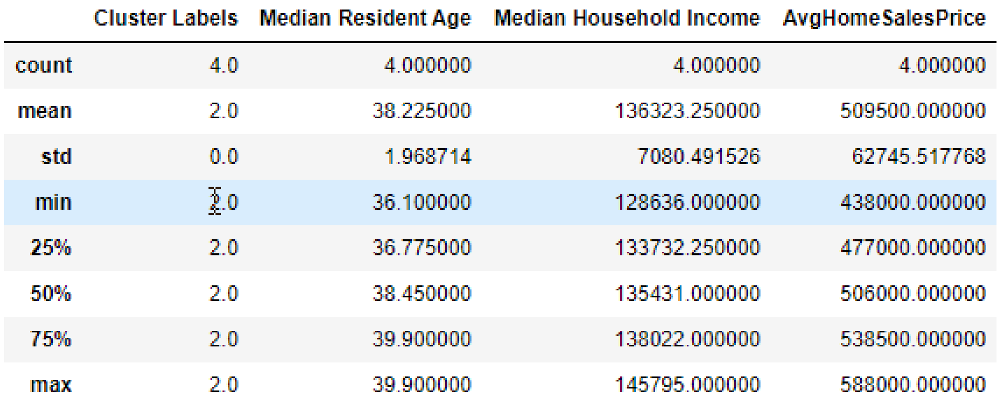
	Cluster Labels	Median Resident Age	Median Household Income	AvgHomeSalesPrice
count	8.0	8.000000	8.000000	8.000000
mean	0.0	38.300000	130153.000000	342125.000000
std	0.0	1.754993	6055.398677	33051.637089
min	0.0	35.800000	118780.000000	287000.000000
25%	0.0	37.000000	127240.000000	324500.000000
50%	0.0	38.450000	130835.500000	343500.000000
75%	0.0	39.900000	135431.000000	362250.000000
max	0.0	39.900000	135431.000000	386000.000000

From cluster 1, we see that this is our Youngest Aged group, Mid Income Group and Lowest Home Values group.



	Cluster Labels	Median Resident Age	Median Household Income	AvgHomeSalesPrice
count	9.0	9.000000	9.000000	9.000000
mean	1.0	36.855556	132555.555556	222777.777778
std	0.0	1.021165	7653.074972	44197.787778
min	1.0	36.100000	127240.000000	161000.000000
25%	1.0	36.100000	128636.000000	190000.000000
50%	1.0	37.000000	128636.000000	223000.000000
75%	1.0	37.000000	132386.000000	261000.000000
max	1.0	39.300000	145795.000000	275000.000000

Cluster 2, our third and final segment represents our Mid Age Group, Highest Income Group and Highest Home Values.



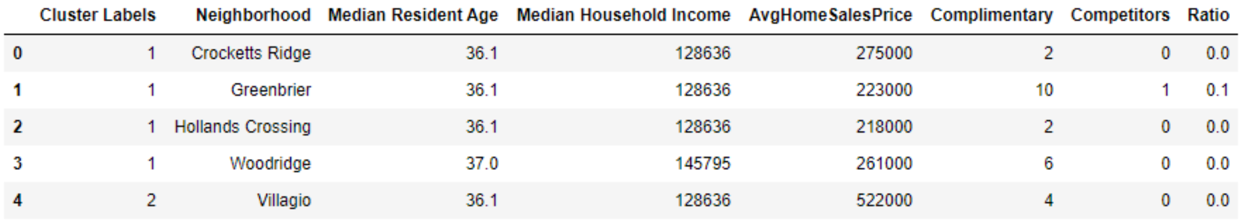
	Cluster Labels	Median Resident Age	Median Household Income	AvgHomeSalesPrice
count	4.0	4.000000	4.000000	4.000000
mean	2.0	38.225000	136323.250000	509500.000000
std	0.0	1.968714	7080.491526	62745.517768
min	2.0	36.100000	128636.000000	438000.000000
25%	2.0	36.775000	133732.250000	477000.000000
50%	2.0	38.450000	135431.000000	506000.000000
75%	2.0	39.900000	138022.000000	538500.000000
max	2.0	39.900000	145795.000000	588000.000000

3.5 Filter Cluster Data for Finely Targeted Demographics

Now that we have an understanding of the makeup of our different potential customer segments, we need to further refine our scope to pinpoint our exact targeted customer demographic segment.

We merged the demographic data with the complimentary vs. competitor data to produce a final dataframe for refinement. After some analysis, we further filtered our data from the exact customer demographic desired by the client. This resulted in the following changes:

- Dropped Cluster 0 -Oldest Age Group, Lowest Income Group, Mid Home Values. While a good market, this least represented our exact targeted demographic out of the three clusters.
- Dropped Neighborhoods with a relatively higher ratio of competitors to complimentary businesses (greater than 10%). This represents our measure of market saturation.
- Dropped the Jamison Park neighborhood due to its outer suburban location –our client is only interested in more centrally-located neighborhoods.
- These changes resulted in reducing our list to the Top 5 Neighborhoods shown as follows:



	Cluster Labels	Neighborhood	Median Resident Age	Median Household Income	AvgHomeSalesPrice	Complimentary	Competitors	Ratio
0	1	Crocketts Ridge	36.1	128636	275000	2	0	0.0
1	1	Greenbrier	36.1	128636	223000	10	1	0.1
2	1	Hollands Crossing	36.1	128636	218000	2	0	0.0
3	1	Woodridge	37.0	145795	261000	6	0	0.0
4	2	Villagio	36.1	128636	522000	4	0	0.0

4.0 Results

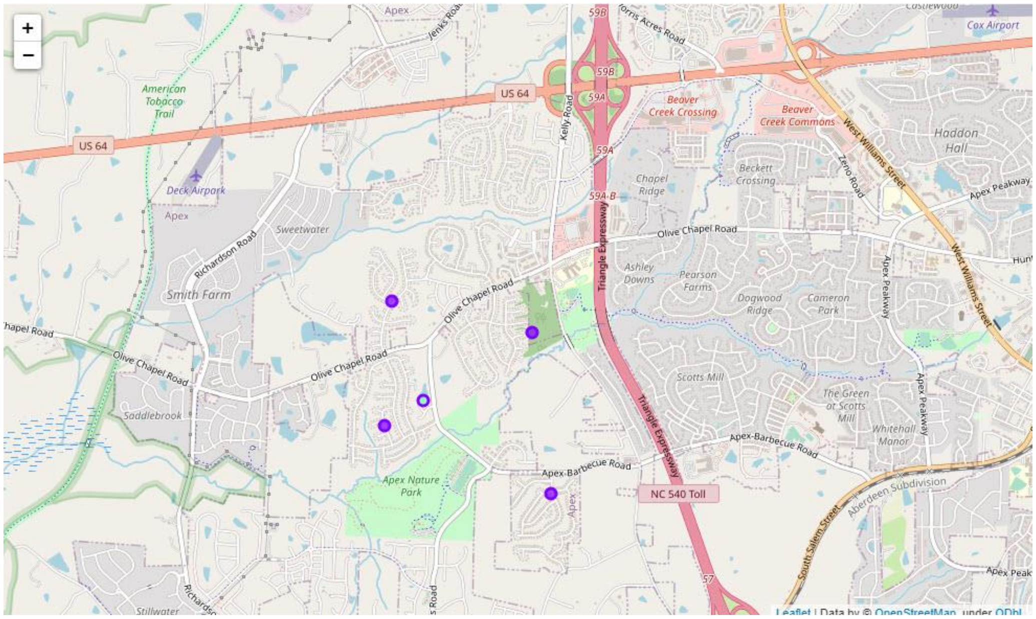
4.1 Targeting the Best Locations

Through my analysis I was able to gather, refine, segment and pinpoint our dataset to be highly targeted with results representative of the client’s ideal customer and ideal business location:

- Demographics - Young - Age Group – Under 40
- Demographics - Relatively Affluent – Median Income Above the Average of our Targeted Population
- Location - Complimentary Businesses within the targeted radius of 1.7 miles
- Location - Low Concentration of Direct Competitors within the targeted radius of 1.7 miles

4.2 Top 5 Neighborhoods

The Top 5 Neighborhoods are visualized as follows with the respective clusters visualized. As you can see, in addition to having favorable demographics, being near complimentary businesses, and havinglow competitor saturation; they are adjacent to major roads and interstates,making for customer ease of access.



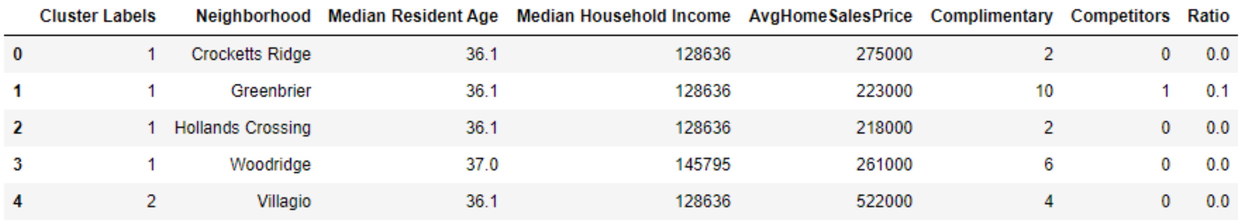
5.0 Discussion

5.1 Observations

- Apex, NC offers a solid demographic based that aligns with the client's target market
- The analysis presented some limitations in the form of a limited data set with a reduced number of features. This meant that our dataset was more closely aligned and resulted in a more "rounded" optimal k analysis.
- Demographic data was roughly aligned to neighborhoods given differing shapes of the geographic borders of census vs subdivision data. This resulted in data results being averaged over adjacent areas.
- With additional resources, this study could be improved through the use of more finely targeted demographic data, as well as qualitative data about competitors and complimentary businesses.

5.2 Recommendations

- I recommend the client target a retail location that is located within 1.7 miles of one the Top 5 Neighborhoods
- Additionally, the target location should be adjacent to one or more Complimentary Businessesto promote increased and dedicated customer traffic.
- This analysis can be further improved through a qualitative study of adjacent complimentary businesses and a study of direct competitors in the area. This qualitative analysis will help to refine the business plan and further differentiate the business opportunity.



	Cluster Labels	Neighborhood	Median Resident Age	Median Household Income	AvgHomeSalesPrice	Complimentary	Competitors	Ratio
0	1	Crocketts Ridge	36.1	128636	275000	2	0	0.0
1	1	Greenbrier	36.1	128636	223000	10	1	0.1
2	1	Hollands Crossing	36.1	128636	218000	2	0	0.0
3	1	Woodridge	37.0	145795	261000	6	0	0.0
4	2	Villagio	36.1	128636	522000	4	0	0.0

6.0 CONCLUSION

- The town of Apex, NC represents a thriving and growing population of currently around 60,000 residents. As a suburb of tech savvy Raleigh, it is by nature, well suited for young, educated and growing families.
- The town demographics are well aligned to the client’s target market, with many opportunities to start a business in a location with strong potential and an optimistic growth outlook.
- While the study methodology can be improved and expanded upon with a qualitative analysis, the results provide a useful tool for finalizing a business location with a Gourmet Coffee Shop retail location within the town of Apex, NC.