Determining Retail Categories to Open in the Church Hill Neighborhood of Richmond, VA

Mitch Phillips

June 10, 2021

1. Introduction

1.1. Background

Across the country the retail landscape is constantly changing. Factors like consumer trends, neighborhood revitalization, and even demographic migration inform the success of different types of businesses in a particular neighborhood. I want to determine under-represented retail categories which would potentially succeed in a changing neighborhood. There are several key players in retail that would benefit from this insight. Property managers like to know what types of businesses would be wise for renting their space. Entrepreneurs with an eye on a particular neighborhood may want to know what retail void could be filled. Existing business owners seeking to expand their offerings may look to these insights as well. I attempt to determine the under-represented retail categories for Church Hill, a neighborhood in Richmond, VA. I have chosen this neighborhood for several reasons. First, I am familiar with the neighborhood and I want to have an intuition for the validity of the results. I have also chosen it because it is in the early stages of revitalization and it sits right in the middle of the state surrounded by several similar cities.

1.2. Problem

I cannot assume the current retail categories present in Church Hill represent the entirety of local customers' wants. Often it is the case that consumers do not know what they want. The problem property managers, entrepreneurs, and existing business owners face is looking beyond their neighborhood for deeper insight into what may be successful. This problem is commonly addressed in a limited way by exchanging information with people in nearby cities with similar markets. This process is cumbersome because it is difficult to network beyond one's city and it is flawed because it relies too much on human perspective which is prone to bias.

1.3. Plan

To address the challenges of looking beyond Church Hill, I will compile and analyze retail category and location data from neighborhoods in nearby cities. Based on the category data, I will use clustering to determine which neighborhoods may have similar retail markets to Church Hill. Then I will generate a ranked list of retail categories present in those similar, nearby neighborhoods which are under-represented from Church Hill. It is important to note I make the assumption consumers in similar, nearby neighborhoods will behave similarly. Based on this assumption, the ranked list will show which retail categories have a high likelihood of success in Church Hill.

1.4. City Choices

I have decided to look at all the neighborhoods of Baltimore, Charlotte, Charlottesville, Norfolk, Raleigh, Richmond, Virginia Beach, and Washington, DC. These cities are are all near Richmond where Richmond is in the center. I think choosing nearby cities is a good way to capture neighborhoods with similar demographics and retail markets. Of course there will certainly be useful neighborhoods with retail categories similar to Church Hill in cities not captured here. This may be considered in further work.

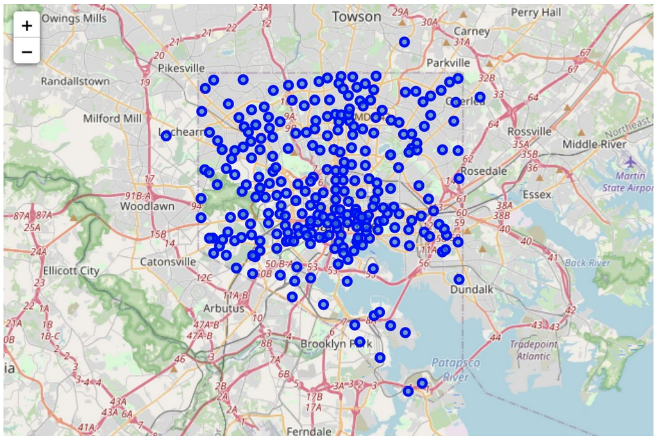
2. Data Collection

2.1. Data Source Choices

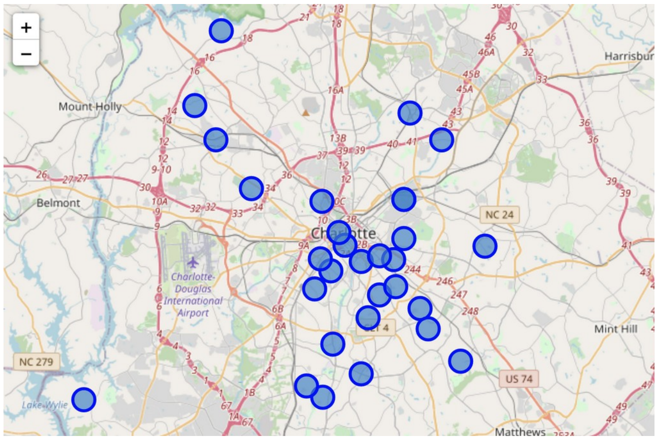
I chose to use Google's geocoder because unlike some alternatives, it recognizes neighborhood names and returns requests quickly. It could be cost prohibitive if I were making a larger volume of requests but I will not exceed the free limit with this project. I used the Foursquare API introduced by the IBM data science course to retrieve retail venue information. This service allows me to send GPS coordinates and a radius to receive a list of retail venues in that circular area along with category information about each venue.

2.2. Geographic Data Collection from Google Geocoder

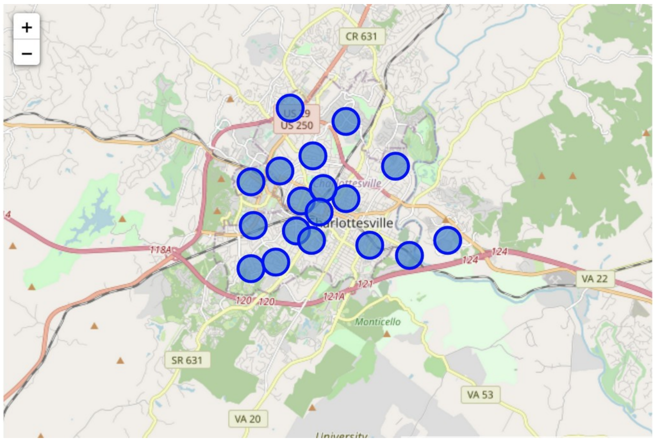
I collected GPS coordinates of every neighborhood in each city and then chosen a neighborhood radius for each city.



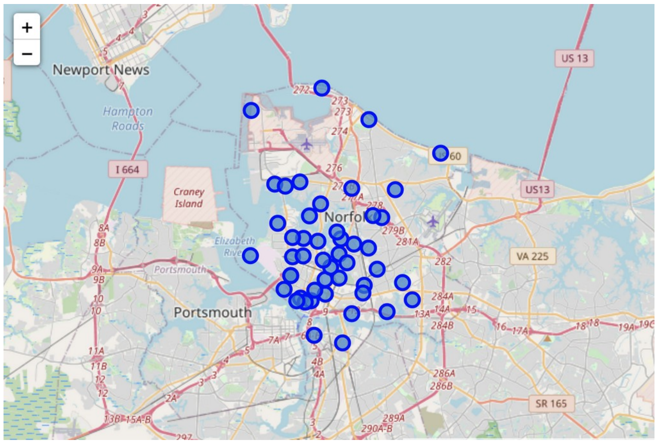
Baltimore, Maryland



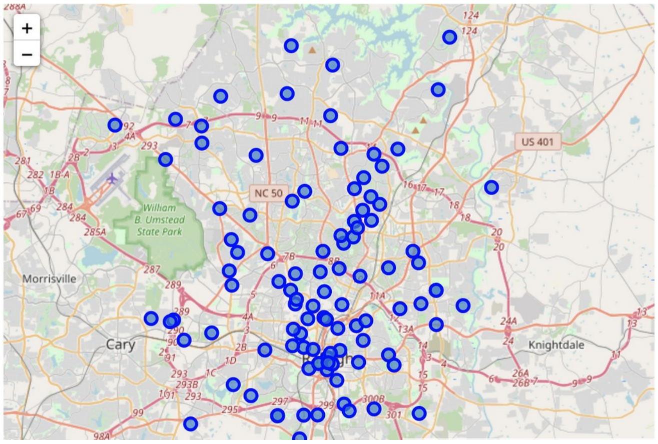
Charlotte, North Carolina



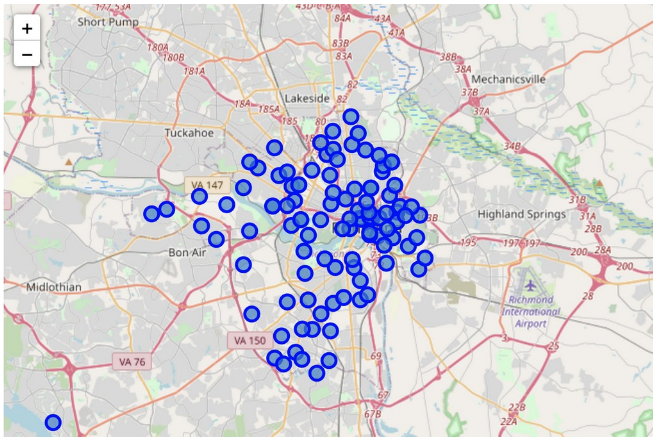
Charlottesville, Virginia



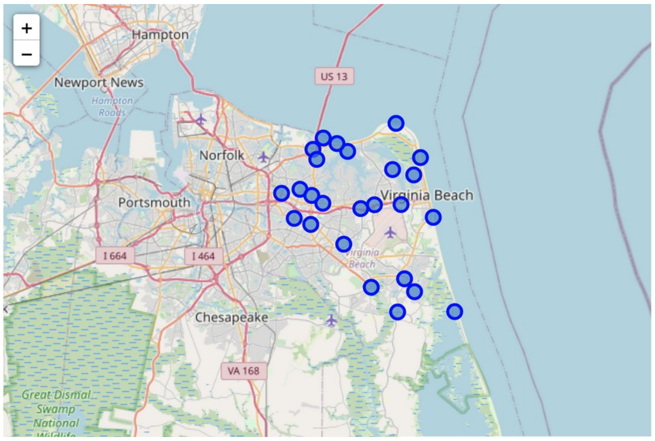
Norfolk, Virginia



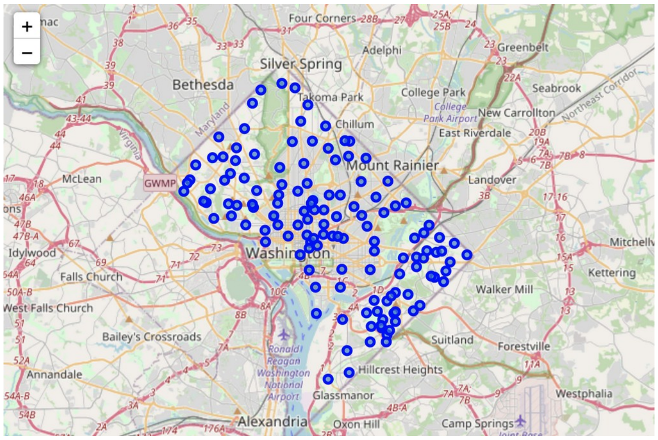
Raleigh, North Carolina



Richmond, Virginia



Virginia Beach, Virginia

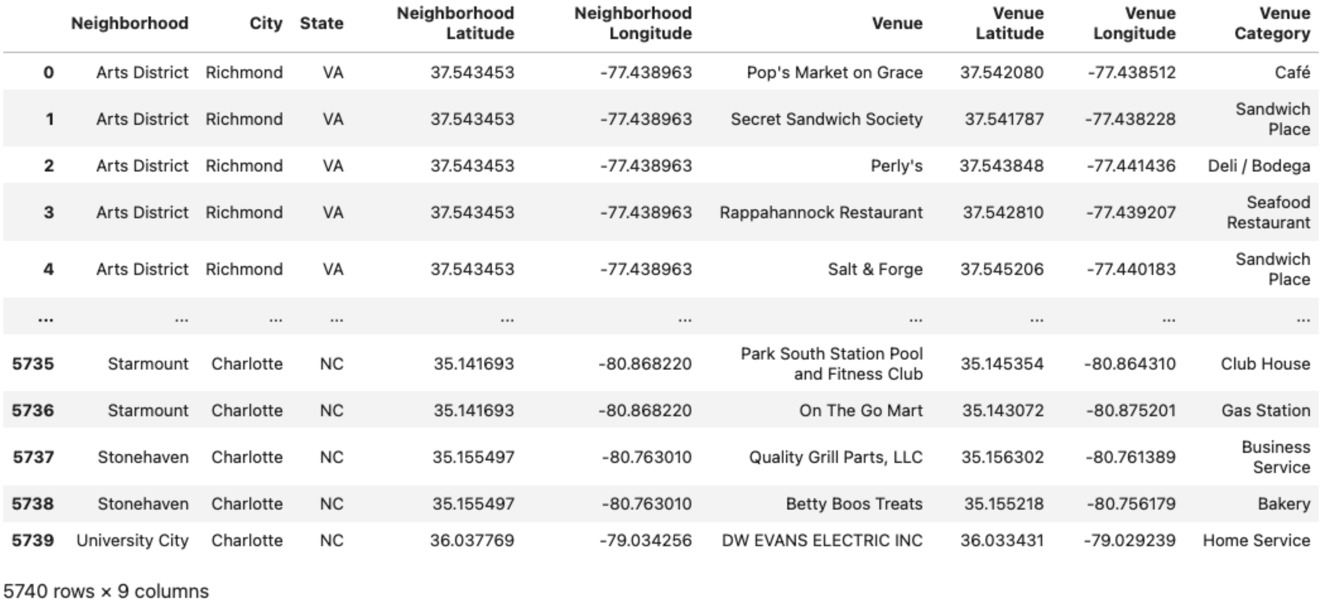


Washington, DC

There 747 neighborhoods on record throughout the eight cities.

2.3. Retail Venue Data Collection from Foursquare

Next I made a collection of retail venues found within each neighborhood.



	Neighborhood	City	State	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arts District	Richmond	VA	37.543453	-77.438963	Pop's Market on Grace	37.542080	-77.438512	Café
1	Arts District	Richmond	VA	37.543453	-77.438963	Secret Sandwich Society	37.541787	-77.438228	Sandwich Place
2	Arts District	Richmond	VA	37.543453	-77.438963	Perly's	37.543848	-77.441436	Deli / Bodega
3	Arts District	Richmond	VA	37.543453	-77.438963	Rappahannock Restaurant	37.542810	-77.439207	Seafood Restaurant
4	Arts District	Richmond	VA	37.543453	-77.438963	Salt & Forge	37.545206	-77.440183	Sandwich Place
...	...	...	...	...	...	...	...	...	...
5735	Starmount	Charlotte	NC	35.141693	-80.868220	Park South Station Pool and Fitness Club	35.145354	-80.864310	Club House
5736	Starmount	Charlotte	NC	35.141693	-80.868220	On The Go Mart	35.143072	-80.875201	Gas Station
5737	Stonehaven	Charlotte	NC	35.155497	-80.763010	Quality Grill Parts, LLC	35.156302	-80.761389	Business Service
5738	Stonehaven	Charlotte	NC	35.155497	-80.763010	Betty Boos Treats	35.155218	-80.756179	Bakery
5739	University City	Charlotte	NC	36.037769	-79.034256	DW EVANS ELECTRIC INC	36.033431	-79.029239	Home Service

5740 rows × 9 columns

I collected data on a total of 5740 retail venues from the eight cities and 747 neighborhoods.

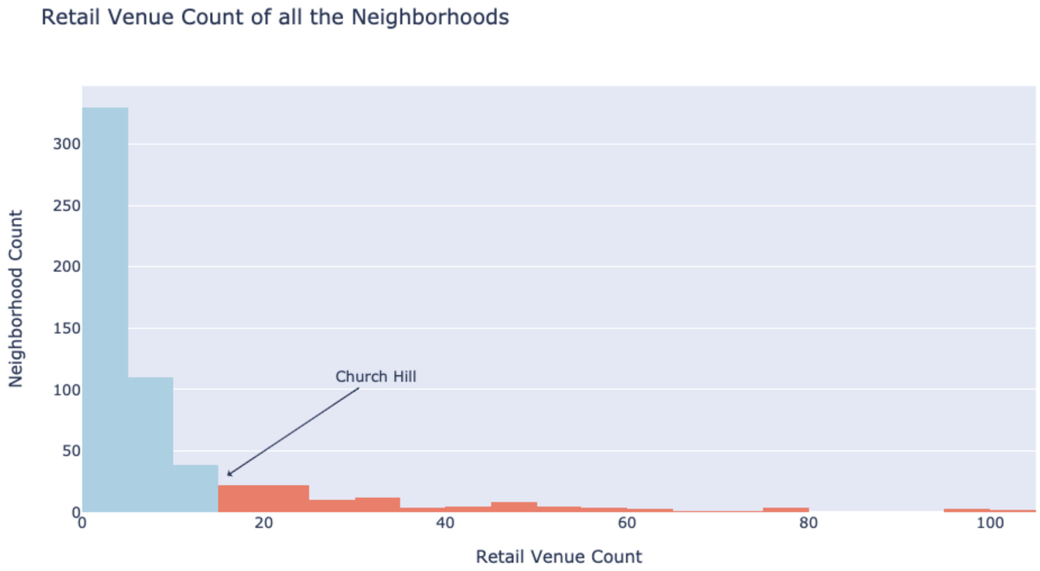
3. Data Cleaning and Preparation

3.1. Remove Unwanted Foursquare Venue Entries

There are some venues collected from Foursquare with category listed as neighborhood, housing development, beach, lake, river, or building. I got rid of each of these venues because they do not represent any kind of relevant retail venue. This removed 57 unwanted venues leaving 5683 remaining venues.

3.2. Remove Neighborhoods with Minimal Retail

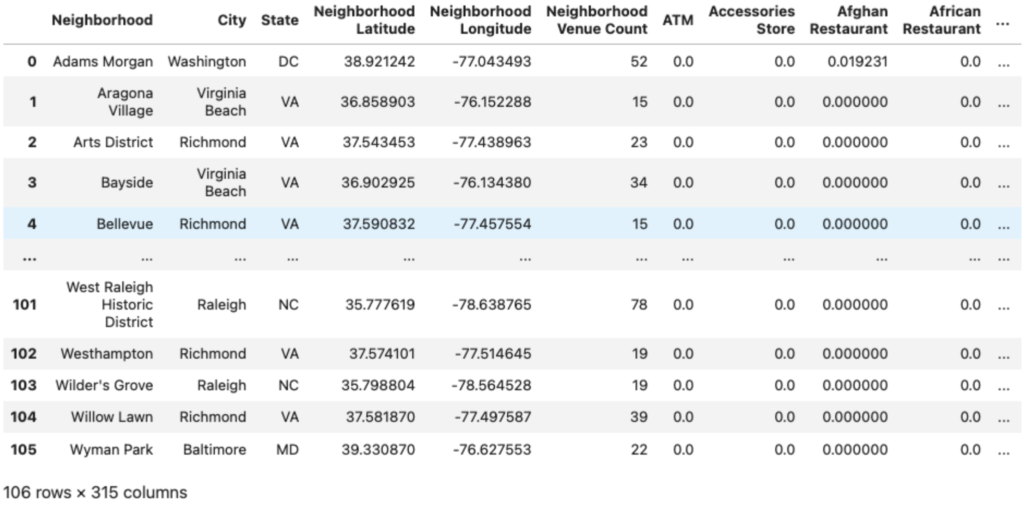
When I counted how many retail venues are in each neighborhood I found that only 584 of the 747 neighborhoods have at least one retail location on record. I also saw that Church Hill has 15 retail locations on record. Here I plot a histogram of the venue counts of the 587 remaining neighborhoods.



There are 478 neighborhoods with less than 15 retail locations. Since I will cluster to consider Church Hill as a growing neighborhood, it would make sense to ignore these 478 neighborhoods with less than 15 retail locations. So I removed 1837 retail venues by removing all the neighborhoods with less than 15 retail venues.

3.3. Create Dummy Feature Columns for Each Retail Category

Then I grouped all the venue data by neighborhood, created a column for every possible retail venue category, and inserted the frequency of occurrence of that category for each neighborhood as a percent. This frequency comes from taking the mean of 0's and 1's where 0 is category not present and 1 is category present. There are 106 neighborhoods and 309 dummy feature columns based on retail venue category. I used the Neighborhood Venue Count column as a feature in the clustering since the number of retail venues in a neighborhood can characterize the neighborhood.

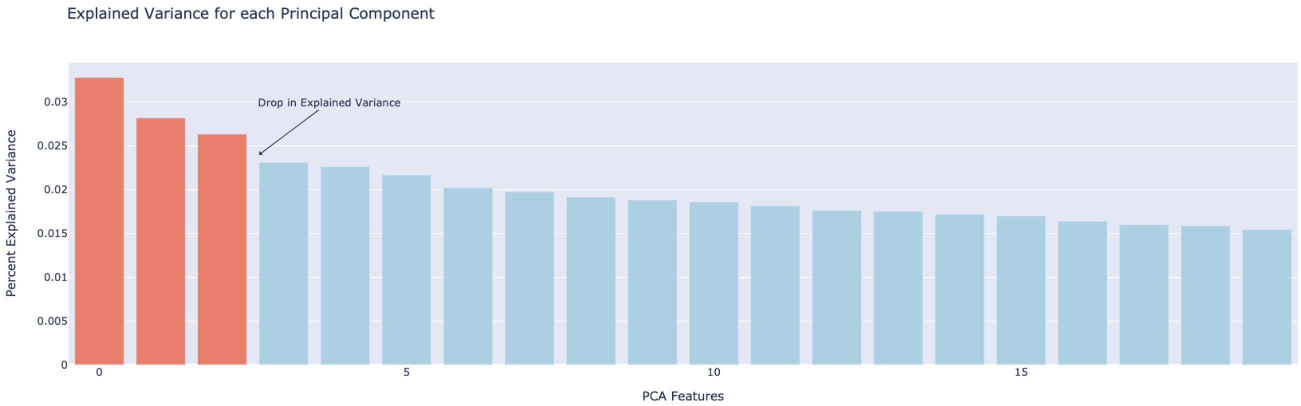


	Neighborhood	City	State	Neighborhood Latitude	Neighborhood Longitude	Neighborhood Venue Count	ATM	Accessories Store	Afghan Restaurant	African Restaurant	...
0	Adams Morgan	Washington	DC	38.921242	-77.043493	52	0.0	0.0	0.019231	0.0	...
1	Aragona Village	Virginia Beach	VA	36.858903	-76.152288	15	0.0	0.0	0.000000	0.0	...
2	Arts District	Richmond	VA	37.543453	-77.438963	23	0.0	0.0	0.000000	0.0	...
3	Bayside	Virginia Beach	VA	36.902925	-76.134380	34	0.0	0.0	0.000000	0.0	...
4	Bellevue	Richmond	VA	37.590832	-77.457554	15	0.0	0.0	0.000000	0.0	...
...	...	...	...	...	...	...	...	...	...	...	...
101	West Raleigh Historic District	Raleigh	NC	35.777619	-78.638765	78	0.0	0.0	0.000000	0.0	...
102	Westhampton	Richmond	VA	37.574101	-77.514645	19	0.0	0.0	0.000000	0.0	...
103	Wilder's Grove	Raleigh	NC	35.798804	-78.564528	19	0.0	0.0	0.000000	0.0	...
104	Willow Lawn	Richmond	VA	37.581870	-77.497587	39	0.0	0.0	0.000000	0.0	...
105	Wyman Park	Baltimore	MD	39.330870	-76.627553	22	0.0	0.0	0.000000	0.0	...

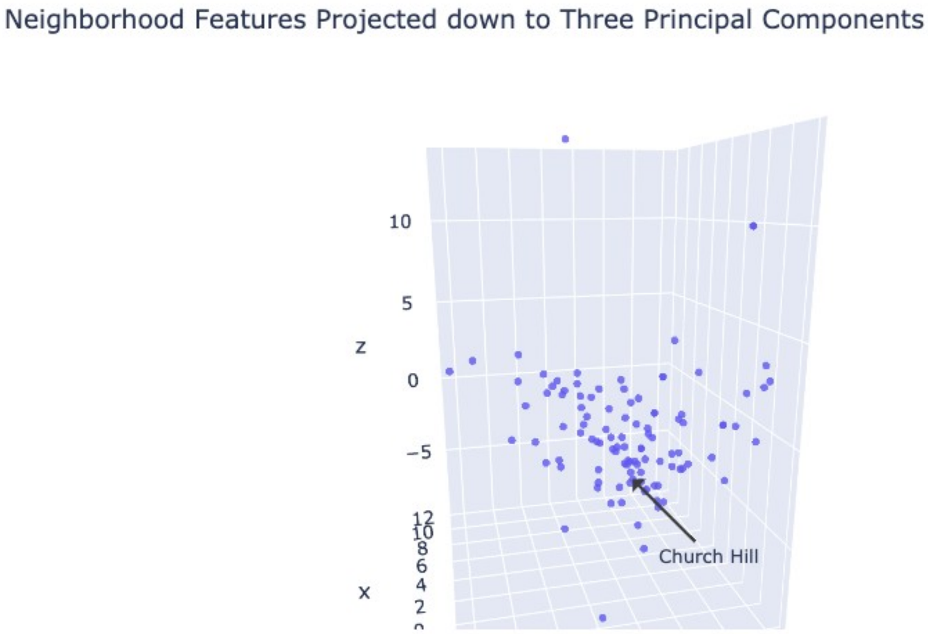
106 rows × 315 columns

3.4. Scale the Feature Columns and Apply PCA Dimensionality Reduction¶

In order to prepare the data for clustering, I processed it in two ways. First I applied a standard scaler to give each feature a mean of 0 and a variance of 1. Next, I applied dimensionality reduction with Principle Component Analysis or PCA. The reason for this is because the data is quite sparse with most entries being zero and with there being more feature columns than sample rows. This sparse data is likely going to introduce a great deal of noise and using PCA will reduce this by projecting the data down to its most important components. I need to decide how many principle components to consider and I can use an explained variance plot.



There is a bit of a drop in variance after the first three components which means I can plot the first three components and see a good amount of the overall variance. I can plot the first three principle components in 3D to get a good idea of the spread of the neighborhoods.


Neighborhood Features Projected down to Three Principal Components

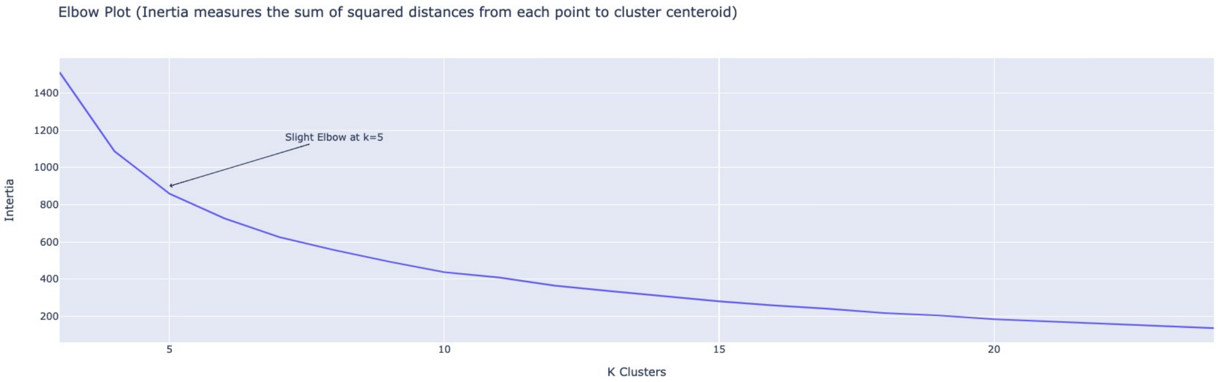
4. Train the K-Means Clustering Model

4.1. Model Choice

Out of the many clustering algorithms available, I chose to use the K-Means algorithm because it is efficient and allows control over the number of clusters produced. I will use K-Means clustering on the first three principle components of the data.

4.2. Determine an Appropriate Number of Clusters, k

I applied the k-means clustering algorithm to the 106 neighborhoods with 15 or more retail venues over a range of cluster counts k from 3 to 25 to determine the best choice. With each model I recorded the inertia or the sum of the squared distances from each point to it's centroid, the center of its cluster. Minimizing this inertia measurement is one way to indicate a good clustering, however inertia will always continue to decrease as the cluster count increases. Therefore, I looked for an elbow point in the inertia plot below which would indicate a point where increasing the cluster count had diminished returns.

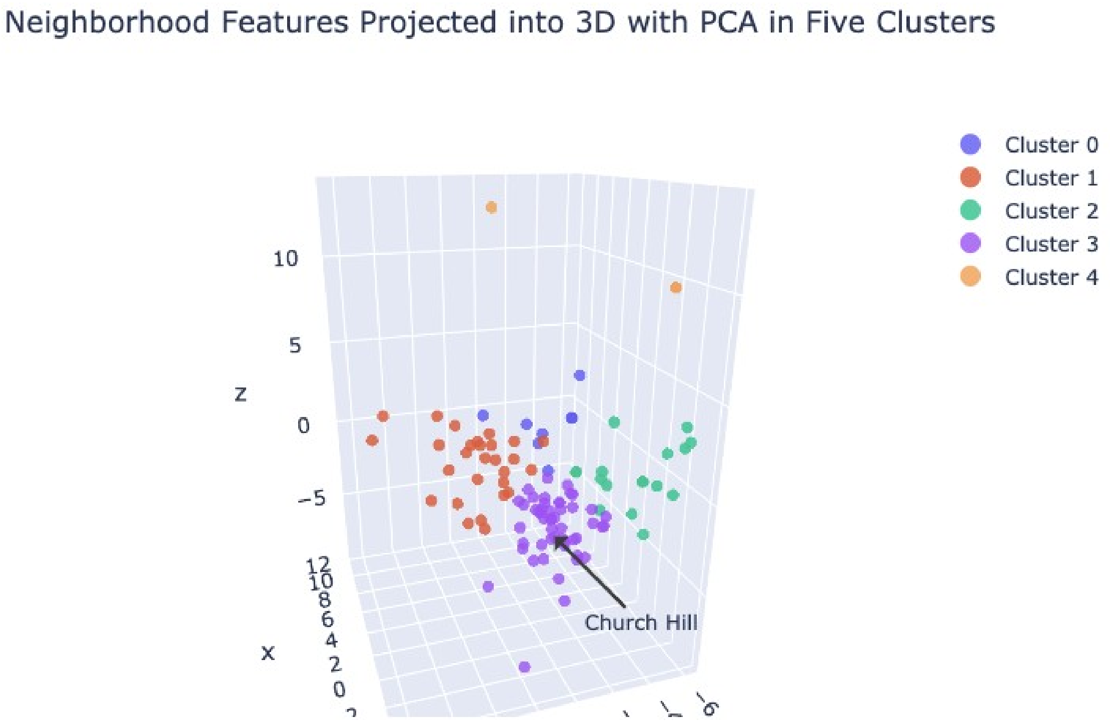

Elbow Plot (Inertia measures the sum of squared distances from each point to cluster centeroid)

There was a slight elbow around 4-6 clusters so any one of those values would be a good choice. There is no significant elbow represented here which may indicate the model does not have clearly defined clusters. I could see in the 3D plot earlier that this is true. I went with k=5 meaning the algorithm will group the 106 neighborhoods into five clusters.

5. Analysis of Results

5.1. Observe the Clusters

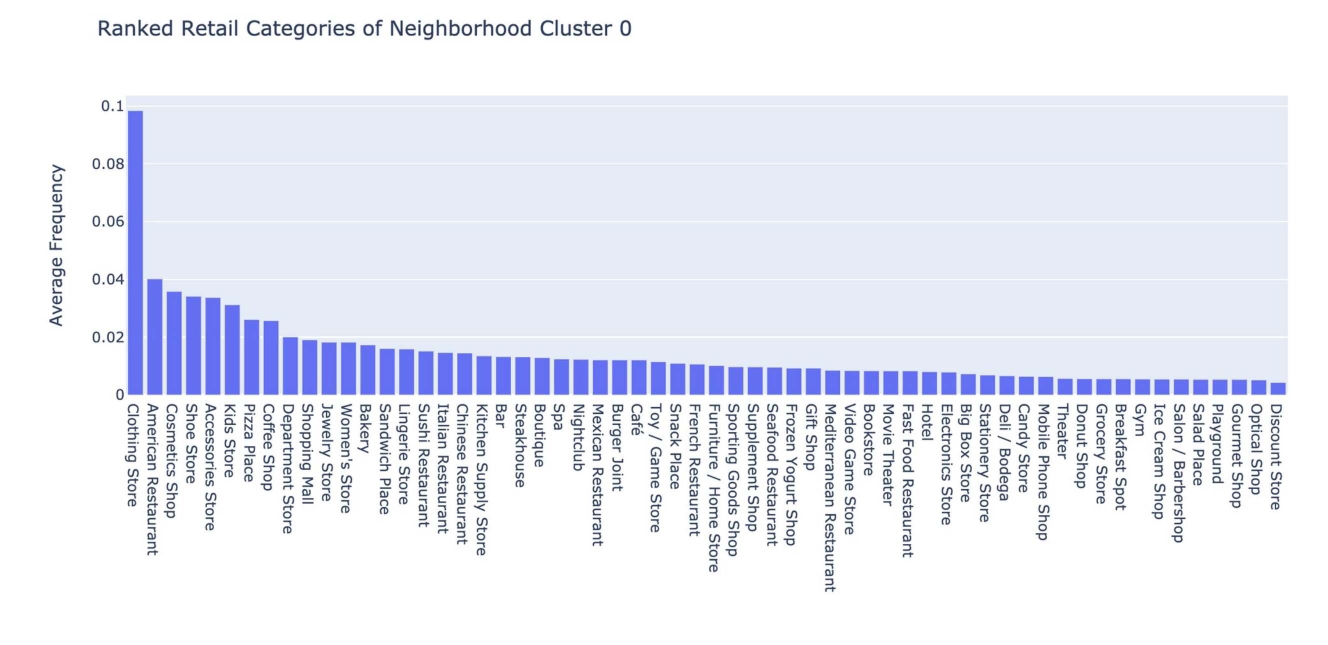
After fitting the K-Means clustering model with k=5 clusters, I plotted the data with the results of the clustering represented in color.

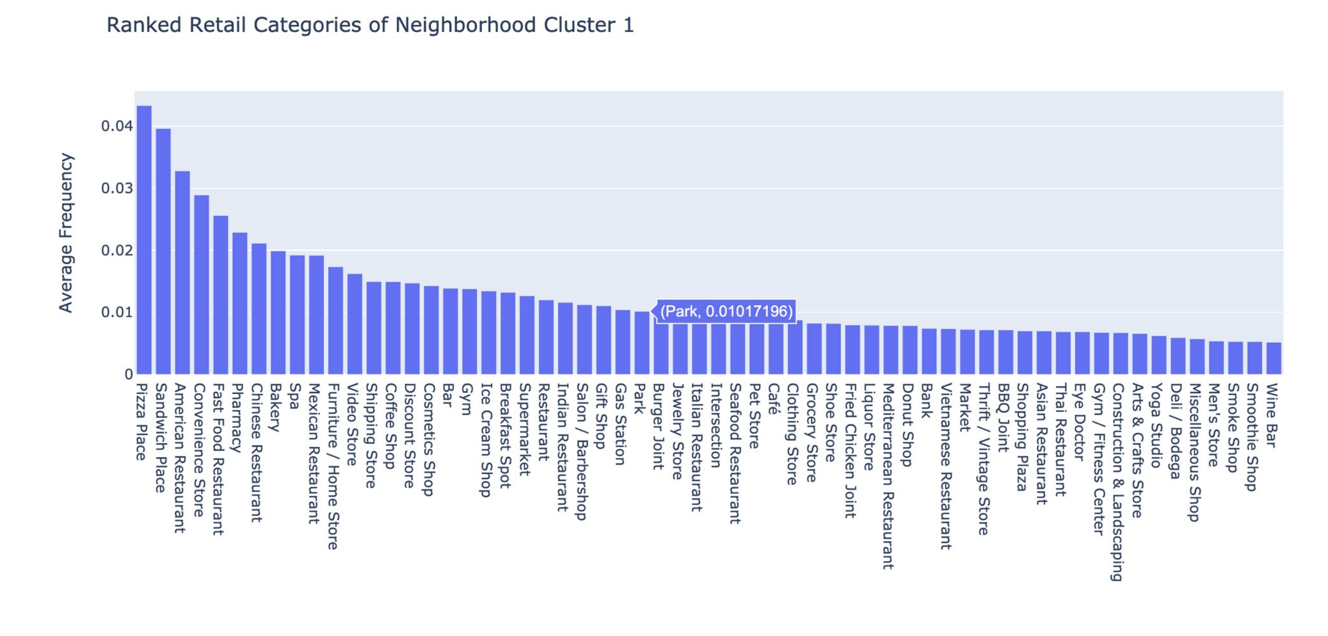


Right away I saw that some clusters are larger than others. For instance, Cluster 0 and 4 seem to consist of outlier neighborhoods. Church Hill was placed in Cluster 3 which appears to be the largest and most central cluster. This is ideal because I wanted a large cluster of similar neighborhoods to improve the characterization of the cluster and thus the confidence in the results.

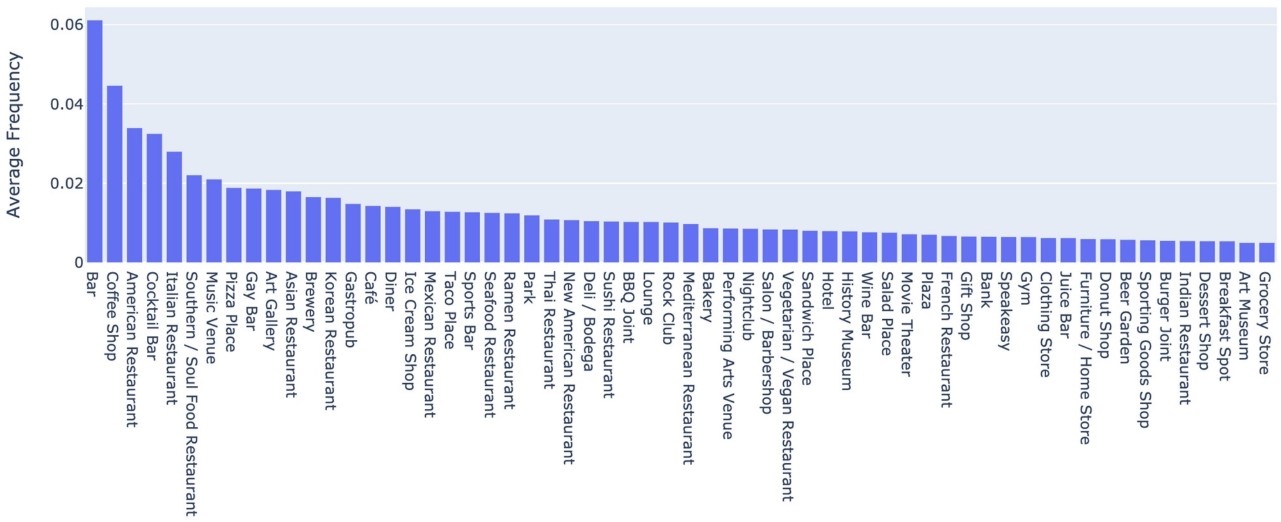
5.2. Describe each Neighborhood Cluster with Bar Plots

I then plotted each retail venue category average frequency and ranked them left to right for each neighborhood cluster. First I looked at the four clusters that don't contain Church Hill.

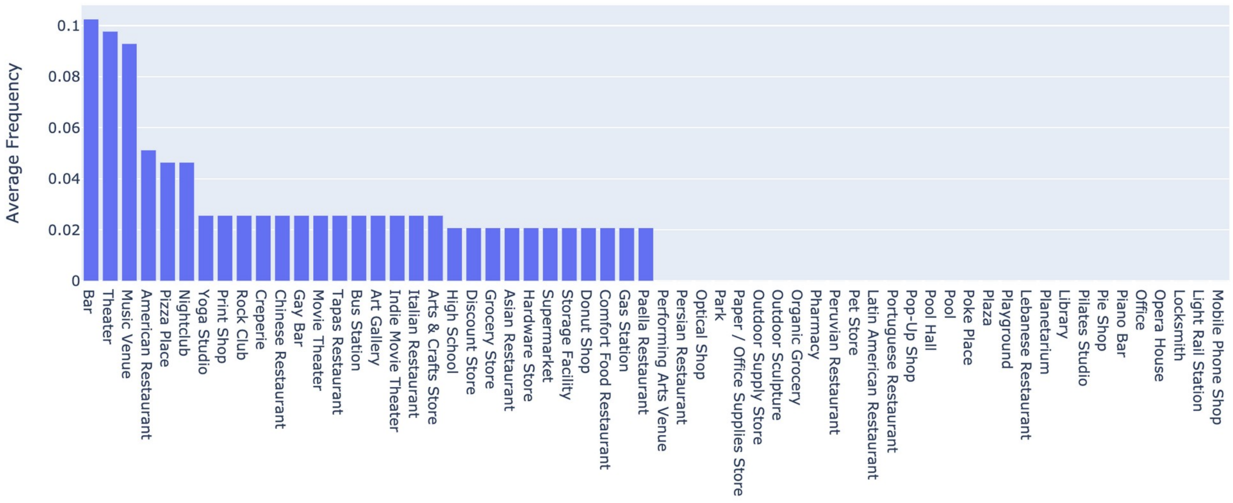

Ranked Retail Categories of Neighborhood Cluster 0


Ranked Retail Categories of Neighborhood Cluster 1

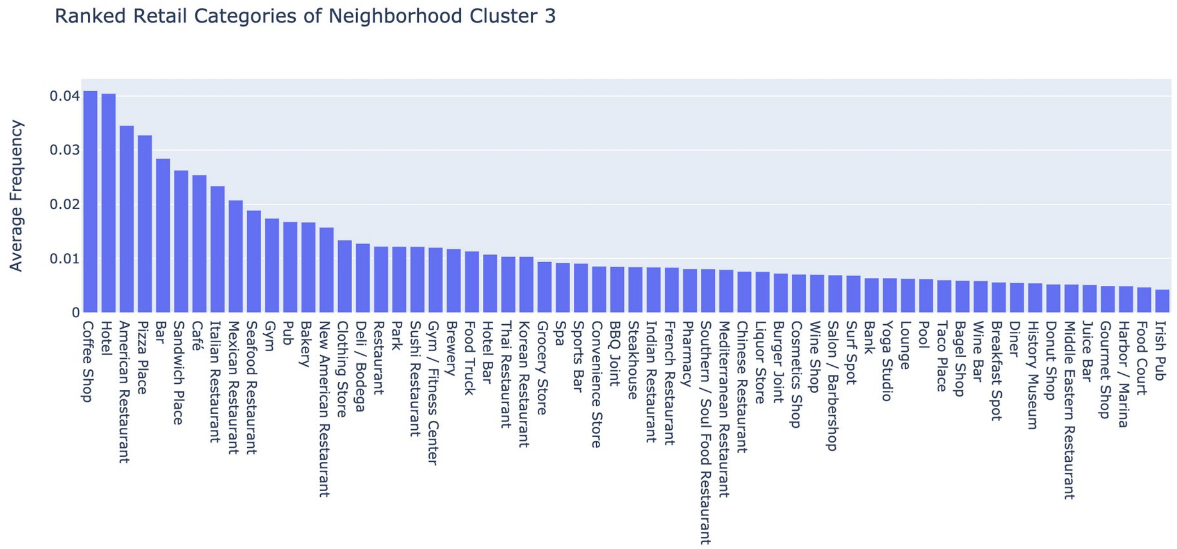
Ranked Retail Categories of Neighborhood Cluster 2



Ranked Retail Categories of Neighborhood Cluster 4



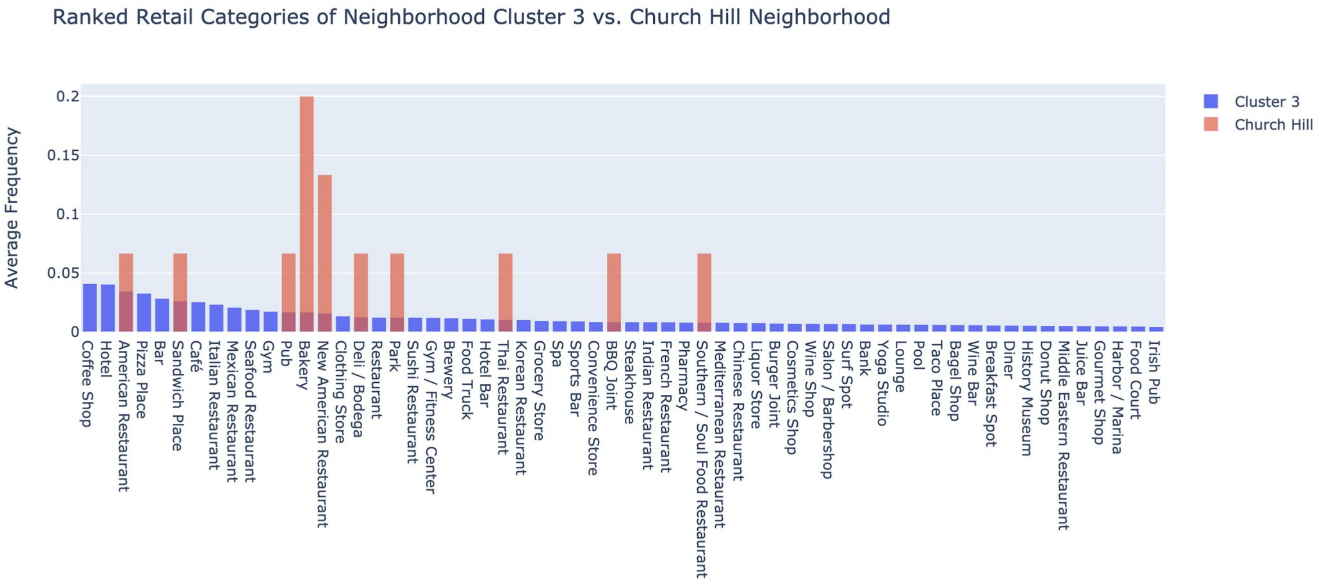
Notice that each cluster characterizes a unique type of neighborhood. I would say cluster 0 looks like a mall where you find shopping. Cluster 1 looks like mostly cheap and quick food places. Cluster 2 seems to be a mix of night life and typical American restaurants. Cluster 4 is small and contains neighborhoods made up of primarily things like clubs, music venues, and theaters. Now I plot the average frequency of retail categories for cluster 3, the cluster containing Church Hill.


Ranked Retail Categories of Neighborhood Cluster 3

This is the cluster I really care about and the one I will be pulling from to determine which retail categories to open in Church Hill. This cluster contains neighborhoods with coffee, hotels, bars, and a variety of restaurants.

5.3. Compare Church Hill Retail to It's Cluster

Now I view cluster 3 again but this time with the retail category frequencies of Church Hill superimposed on top of the averages for the whole cluster. This shows how Church Hill matches up with the averages of its cluster to see what retail is apparently missing from Church Hill.


Ranked Retail Categories of Neighborhood Cluster 3 vs. Church Hill Neighborhood

6. Conclusions

6.1. Results for Church Hill

The purpose of this project was to determine new retail that should open in the Church Hill neighborhood of Richmond, VA. By clustering the neighborhoods of Richmond, VA and seven nearby cities, I was able to come up with a cluster of neighborhoods with similar retail to Church Hill. I could then deduce the retail categories missing from Church Hill yet present in high frequency within its cluster. These retail categories should do well in Church Hill because they've done well in neighborhoods I've deemed similar.

The Church Hill neighborhood would likely do well to add any one of the follow types of retail (just to name the top ten):

1. Coffee shop
2. Hotel
3. Pizza place
4. Bar
5. Cafe
6. Italian Restaurant
7. Seafood Restaurant
8. Gym
9. Clothing store
10. Sushi restaurant

6.2. Validity of the Results

First I will say from personal experience in Church Hill that these feel like the right choices for new retail. Although there are numerous places that sell coffee, there is no official coffee shop or cafe. The neighborhood is missing Italian, seafood, and sushi restaurants and I have a strong feeling based on my familiarity that the people in Church Hill would enjoy these.

There seems to be a small problem with my data not including all the retail venues in Church Hill. For instance, I am a bit confused as to why Pizza place shows up on this list because Church Hill has a popular pizza place called 8-1/2 Pizza. I am suspecting one of two issues: either the radius I created to define Church Hill did not capture this retail venue or it is just not listed on Foursquare yet.

Another problem is the sparsity of the retail category data. Not only are there many zeros throughout the 309 feature columns, but there is also only 106 sample neighborhoods being clustered. While the PCA dimensionality reduction allowed me to reduce noise and focus on components which represented the greatest variance, I have concern that I overcompensated with reduction and neglected some valuable components. If I could reduce data sparsity, I would expect a reduction in noise and a greater consolidation of variance in the first few PCA components, giving a more accurate over-all representation of the data.

Despite these issues, I still have a good amount of confidence in the current results. I have this confidence because other than pizza place, the ten retail categories suggested for Church Hill are indeed missing and seem like popular choices for the neighborhood based on my personal experience here. I have a suspicion that even with noise reduction and improved clustering I would still see a very similar list of retail categories.

6.3. Further Work

PCA was a good way to improve K-Means clustering because I was able to reduce potential noise from sparse data by reducing dimensionality. There are other clustering algorithms which are supposed to do better than K-Means on data with high sparsity and noise, such as DBSCAN or entropy-weighted k-means clustering. It would be interesting to see how either of these algorithms perform compared to the K-Means algorithm.

I could also look to decrease the sparsity of data in numerous ways. First I could try adding more neighborhoods from more cities across the country. By including more sample neighborhoods, the overall sparsity of the data would decrease. Another idea is to add feature columns that would inevitably have less sparsity. This could be median house price, median income, political party affiliation of local representatives, or population density. Any one of these features could likely aid to characterize a retail market. A third way I could decrease sparsity is by reducing the sparsity of the existing dummy category columns by adding more finely tuned category information. I could do this by considering a hierarchy of parent categories and sub-categories defined by a Foursquare API documentation webpage. Some retail venues are only given a broad category such as Restaurant. Then other venues are given a category such as American Restaurant versus New American Restaurant which are considered as different as any other categories, despite being similar or essentially same. If each retail venue were labeled with its assigned category and all of its parent categories, then for example the American Restaurant and the New American Restaurant would now share at least parent category and maybe more. Then the retail venues would be better represented and the data would be less sparse. I would be interested to see what the decrease in sparsity would look like.