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. We 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. We will attempt to determine the under-represented retail categories for Church Hill, a neighborhood in Richmond, VA. This is a good neighborhood because it is in the early stages of revitalization. It’s also in a city that sits right in the middle of the state surrounded by several similar cities.

#### 1.2. Problem

We can’t assume the current retail categories present in Church Hill represent the entirety of local customers’ wants. Often it is the case that consumers don’t 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’s difficult to network beyond one’s city and it’s 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, we will compile and analyze retail category and location data from neighborhoods in nearby cities. Based on the category data, we will use clustering to determine which neighborhoods we may consider to be similar markets to Church Hill. Then we 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 we make the assumption consumers in similar, nearby neighborhoods will behave similarly. Based on this assumption, our ranked list will tell us what retail categories have a high likelihood of success in Church Hill.

#### 1.4. City Choices

I've 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.

#### 1.X Intro from jupyter notebook

The purpose of this project is to determine new retail that should open in the Church Hill neighborhood of Richmond, VA. We will compare Church Hill's retail offerings to those of similar neighborhoods in Richmond and other nearby cities. We can gather retail data on all the neighborhoods of Richmond and seven other nearby cities and make clusters of similar neighborhoods. We can then deduce the retail categories missing from Church Hill but found in high frequency within its cluster. These retail categories would likely do well in Church Hill because they've done well in neighborhoods we've deemed similar.

### **2. Data Collection**

#### 2.1. Data Source Choices

I'm choosing 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 will use the Foursquare API introduced by the IBM data science course to retreive retail venue information. This service allows me to send a gps coordinate and radius and recieve a list of retail venues in that circular area along with category information about each venue. It would be more ideal if I could recieve a batch of retail venues based on their exact neighborhood membership because a circle can only be a rough approximation for any geogaphic region. However this is the best method I can find at present.

#### 2.2. Geographic Data Collection from Google Geocoder

|  |  |
| --- | --- |
| Baltimore, Maryland | Charlotte, North Carolina |
| Charlottesville, Virginia | Norfolk, Virginia |
| Raleigh, North Carolina | Richmond, Virginia |
| Virginia Beach, Virginia | Washington, DC |

#### 2.3. Retail Venue Data Collection from Foursquare

Now we collect all retail venues found within a radius of each neighborhood collected above. I'm going to chose an appropriate radius for each city by roughly estimating based on the map views above. Notice that more dense cities will have a smaller neighborhood radius.

We've collected data on a total of 5740 retail venues from eight cities and 747 neighborhoods.

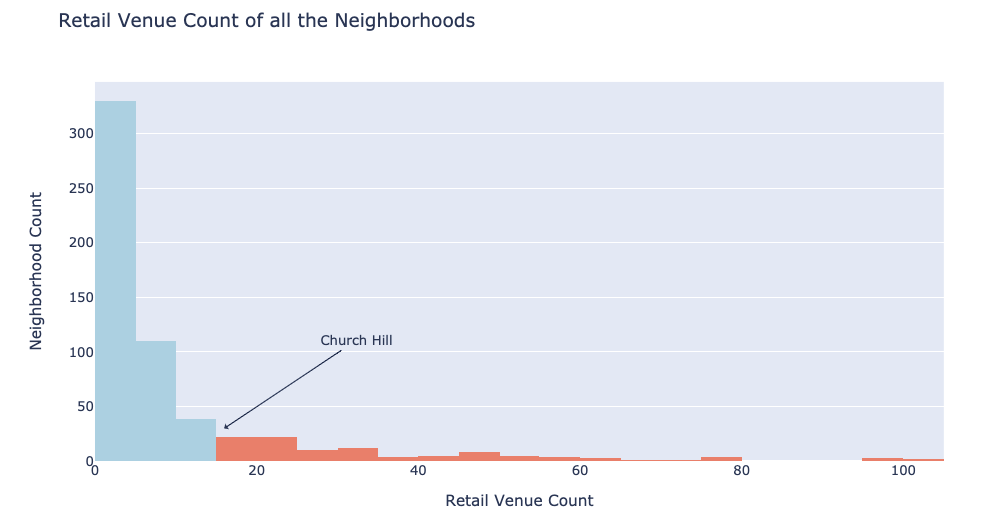
### 3. Data Cleaning and Preparation

#### **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. We're going to get rid of each of these venues because they do not represent any kind of relevant retail venue. This removed 57 unwanted venues with 5683 remaining.

#### **3.2.** Count the Number of Retail Venues in Each Neighborhood

There may be some neighborhoods with vastly different numbers of retail venues. Let's count them for each neighorhood and look at the counts in a plot.   
We see that only 584 of the 747 neighborhoods have at least one retail location on record. We also see that Church Hill has 15 retail locations on record. Let's make a histogram of the venue counts of the 587 remaining neighborhoods.



There are 478 neighborhoods with less than 15 retail locations. Since we are clustering to consider Church Hill as a growing neighborhood, it would make sense to ignore these 478 neighborhoods with less than 15 retail locations.

#### **3.3.** Remove Neighborhoods with Minimal Retail

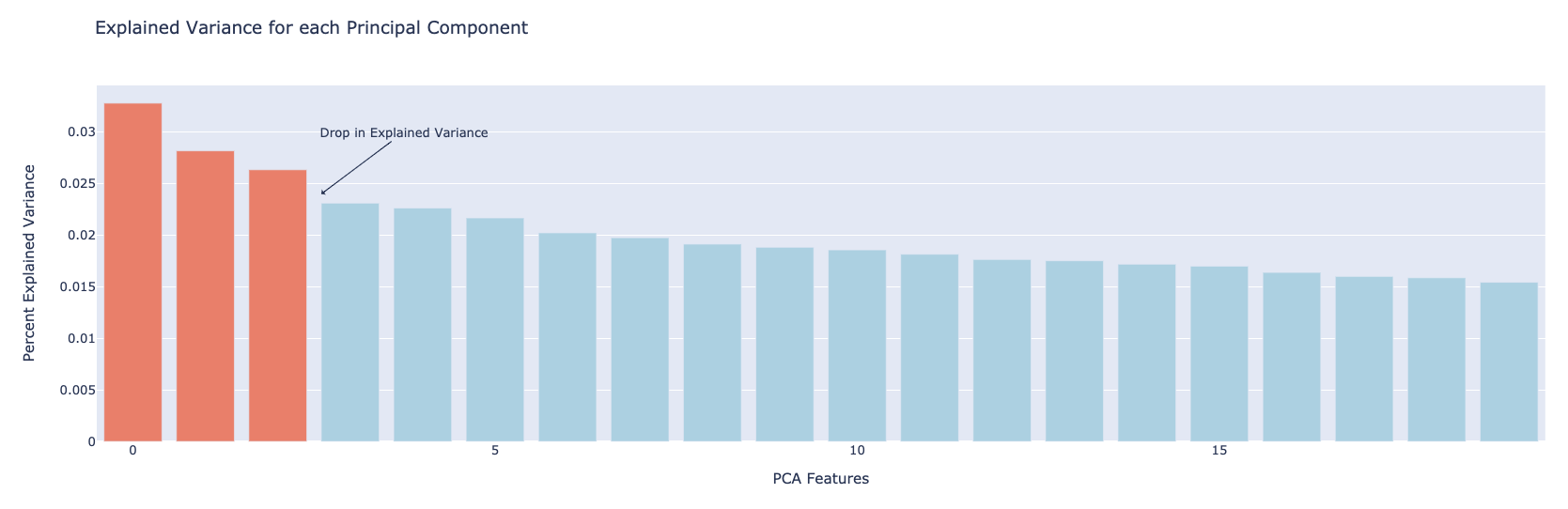
With 3846 retail venues remaining, we've removed 1837 retail venues by removing all the neighborhoods with less than 15 retail venues.

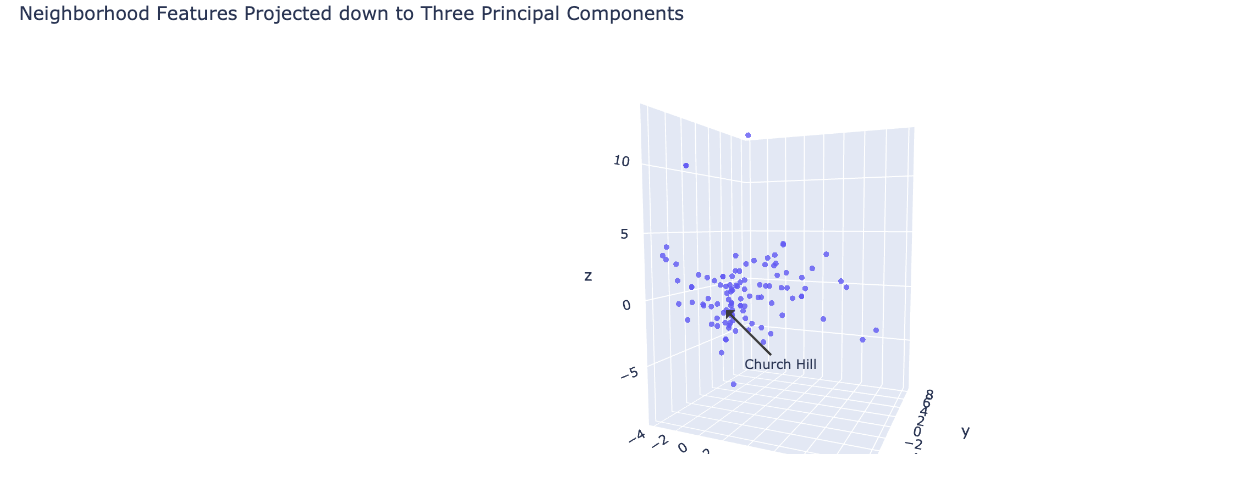
#### **3.4.** Create Dummy Feature Columns for Each Retail Category

Now we will group all our venue data by neighborhood, create a column for every possible retail venue category, and insert the frequency of occurance 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 not present and 1 is present. We have 106 neighborhoods and 309 dummy feature columns based on retail venue category. We will also use the Neighborhood Venue Count column as a feature in our clustering since the number of retail venues in a neighborhood can characterize the neighborhood.

#### **3.5. Scale the Feature Columns and Apply PCA Dimensionality Reduction**[**¶**](_self)

In order to prepare our data for k-means clustering, we are going to process it in two ways. First we will apply a standard scaler to give each feature a mean of 0 and a variance of 1. Next, we will apply dimensionality reduction with Principle Component Analysis or PCA. The reason for this is because our 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 our data down to its most important components.

We have a bit of a drop in variance after the first three components which means we we can plot the first three components and see a good amount of the overall variance. This data in three dimensions is what we will cluster with the k-means algorithm. As stated above, we've reduced the dimensionality of our data to reduce noise caused by the sparsity of dummy features.



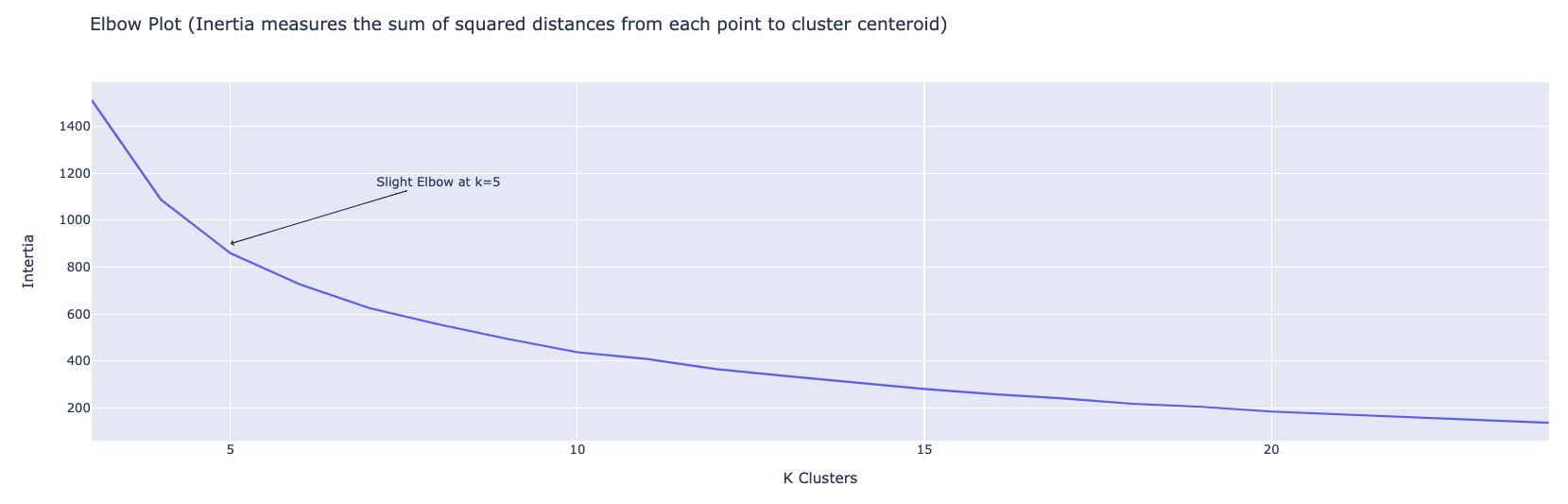
### 4. Train the K-Means Clustering Model

#### **4.1. Model Choice**

### Out of the many clustering algorithms available, I've chosen to use the K-Means algorithm because it's efficient and allows us to have control over the number of clusters produced. To miminize noise we will use PCA dimensionality reduction. I may consider redoing this project with a DBSCAN algorithm later to see if that produces a more useful clustering.

#### **4.2. Determine an Appropriate Number of Clusters, k**

We will apply the k-means clustering algorithm to the 106 neighborhoods with 15 or more retail venues. We'll do so over a range of cluster counts k from 3 to 25 to determine the best choice. With each model we can record the inertia or the sum of the squared distances from each point to it's centroid, the center of its cluster. Minimizing this intertia measurement is one way to indicate a good clustering, however inertia will always continue to decrease as the cluster count increases. Therefore, we will look for an elbow point in the interia plot below which indicates a point where increasing the cluster count has diminishing returns.



It looks like we have 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 our model does not have clearly defined clusters. We could see in our 3D plot earlier that this is true. We will go with k=5 meaning our algorithm will group the 106 neighborhoods into five clusters. Next we can view our clustering and see if k=5 was a good choice.

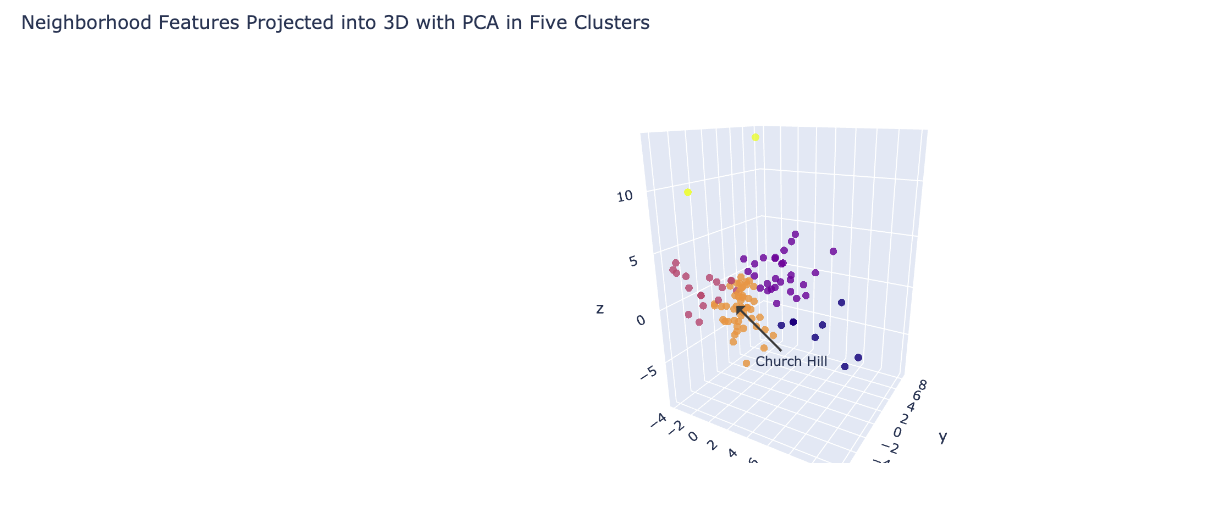
#### **4.3. Fit The K-Means Model**

We've chosen to use five clusters for our k-means model based on the inertia measurements in the elbow plot above.

### **5. Analysis of Result****s**

#### **5.1. Observe the Clusters**

We can plot our data with reduced dimensionality again, this time with the results of our clustering represented in color.



Let's first look at how many neighborhoods were placed into each cluster. Right away we can see that some clusters are larger than others. For instance, Cluster 0 and 4 seem to consist of outlier neighborhoods. Now we should find out what Church Hill's cluster of neighborhoods looks like. First we find out Church Hill's label, and then we pull up all the neighborhoods with this label. Church Hill was put into cluster 3. We expect to see 50 neighborhoods in this cluster, so let's list them now.

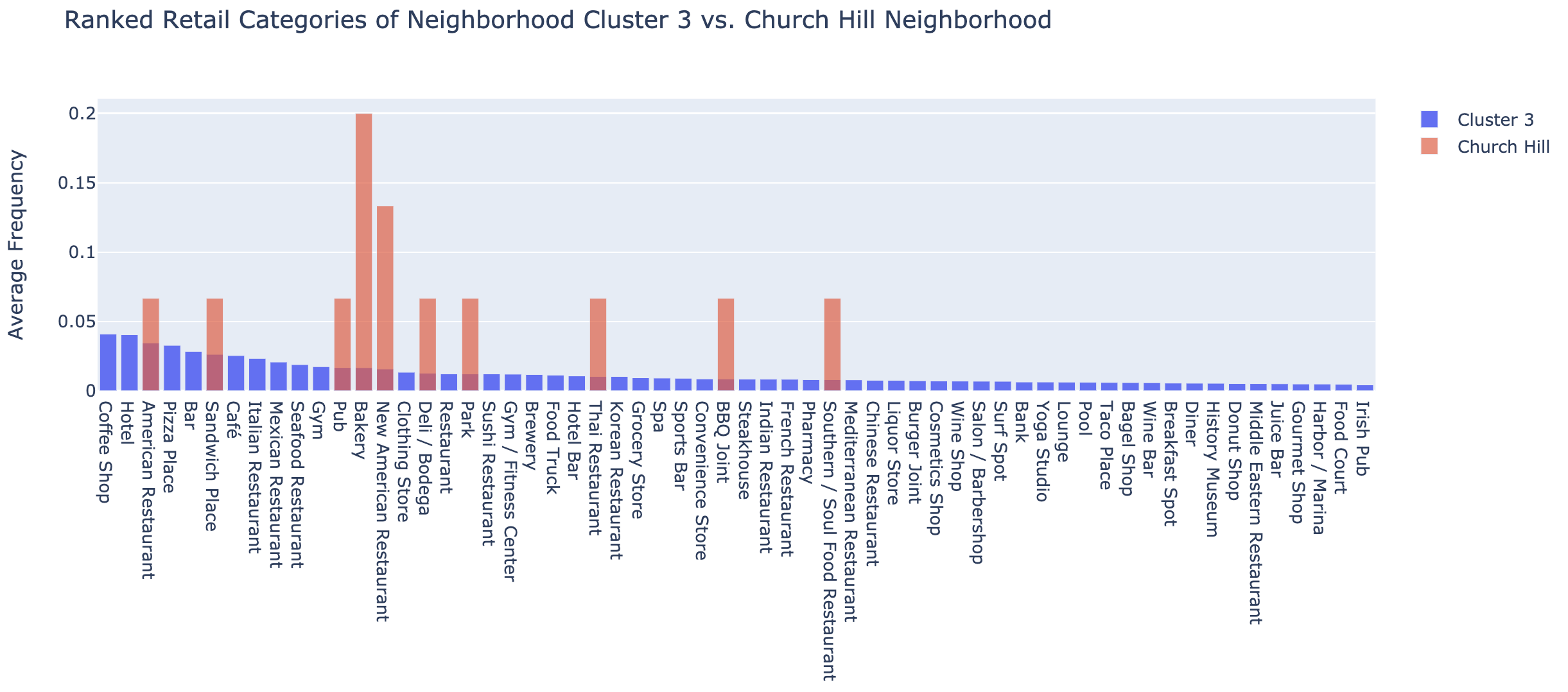
#### 5.2. Describe each Neighborhood Cluster with Bar Plots

Here we can rank each retail venue category left to right and plot average frequency in a bar graph for each neighborhood cluster.

|  |  |
| --- | --- |
|  |  |
|  |  |

#### **5.3. Compare Church Hill Retail to It's Cluster**

Now we will view cluster 3 again but this time with the retail category frequencies of Church Hill superimposed on top. This will allow us to see how Church Hill matches up with the average of its cluster as see what retail is apparently missing from Church Hill.



### **6. Conclusions**

#### **6.1. Results for Church Hill**

The purpose of this project is 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, we are able to come up with a group of neighborhoods with similar retail to Church Hill. We can 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 we've deemed similar.

We gather from the plot above that 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. We are missing Italian, seafood, and sushi restaurants and I have a strong feeling the people in this neighborhood would enjoy those.

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'm suspecting one of two issues: either the radius I created to define Church Hill didn't capture this retail venue or it's just not listed on Foursquare yet.

A bigger problem is the sparsity of the retail category data. Not only are there many zeros throughout the 309 feature columns, but there's also only 106 sample neighborhoods being clustered. While the PCA dimensionality reduction allowed us to reduce noise and focus on components which represented the greatest variance, I have concern that we overcompensated with reduction and neglected some valuable components. If we could reduce data sparsity, we 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, no-brainer choices for the neighborhood based on my personal experience here. I have a suspicion that with noise reduction and improved clustering we would still see a very similar list of retail categories.

#### 6.3. Further Work

PCA was a good way to improve clustering because we were able to reduce potential noise from sparse data by reducing dimensionality. There are other clustering algorithms which are supposed to do better on training sets 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. I could add a wider variety of data to characterize each neighborhood. This data 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. I could also add more neighborhoods from more cities across the United States.

Another interesting way that I could decrease sparsity in the data would not require going beyond Foursquare. Instead, I could improve the quality of the retail venue category labels 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 it's 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. Unfortunately, gathering these parent categories doesn't seem to be built into Foursquare’s API, but this could still be done with a tree data structure and a function on my end.