RALEIGH NEIGHBORHOODS ANALYSIS

IBM Coursera Data Science Specialization

Report by:
Sakshi Jain

1. Introduction

1.1 Background

Raleigh is one of the three hub cities comprising the Research Triangle Park (RTP) which is the largest research park in the United States. RTP is thus named for the three hub cities of Raleigh, Durham, and Chapel Hill, or more properly, for the three major research universities in those three cities (North Carolina State University, Duke University, and the University of North Carolina at Chapel Hill respectively). The park is home to more than 300 companies employing 55,000 workers and an additional 10,000 contractors. Raleigh soared to third place in a study, regarding the best American cities to work in technology, conducted by financial news and information site SmartAsset. Given its increasing reputation as a high-tech hub, it is becoming an attractive destination for investments, also because Raleigh's business costs are 14 percent below the national average, per Moody's Analytics. The downtown of Raleigh has evolved dramatically and its transformation into lively centers with high-rises and new restaurants provides great market opportunity for hotel development.

1.2 Problem

Raleigh is the capital city of North Carolina and consists of many neighborhoods. This project is aimed at determining the neighborhoods which would be optimal choices for the location for opening a new hotel based on their proximity to other venues of interest such as restaurants, attractions, etc. and where there are none or a smaller number of already existing hotels.

1.3 Interest

Location is a critical consideration for business owners whenever choosing to open a new hotel. This analysis would be of interest to stakeholders/business owners looking to open a new hotel in the RTP area. It could also be of interest to people looking for investment in the area or hotel business.

2. Data acquisition and cleaning

2.1 Data sources

The main data required for the project were extracted as follows:

- Wikipedia - The list of neighborhoods for Raleigh were extracted from the Wikipedia page(https://en.wikipedia.org/wiki/Raleigh,_North_Carolina_neighborhoods).
- Python Geocoder Nominatim - The location coordinates (latitude, longitude) for the neighborhoods obtained using the Python geocoder package.
- latlong.net - The missing coordinates looked up using this online geographic tool.
- Foursquare API (Search) - Search and extract the already existing hotels in the neighborhoods.
- Foursquare API (Explore) – Explore the neighborhoods.

2.2 Using Data to solve the Problem

The basic idea behind analyzing the neighborhoods in Raleigh is to find a location which hasn't been saturated yet with already existing hotels so that competition is comparatively less and the probability of the business thriving is more. Raleigh is gaining popularity as a technology hub and is one of the fastest growing cities, hence the demand is going to be increasing in future. As the area remains a premier destination for business and tourism, it makes sense to consider location based on proximity to businesses and industry, colleges, hospitals, attractions, services and entertainment as these are the important generators of room demand. Hence the neighborhood with higher number of venues and having less than three already existing hotels will be important consideration in finalizing the location. This analysis can be done using the data by collecting information about the neighborhoods and checking their corresponding counts.

2.3 Data cleaning

Data scraped from Wikipedia page contained only the name of neighborhoods and regions. The names of the neighborhoods and regions extracted were thus stored in excel. While fetching the location coordinates for the neighborhoods using the geocoder, there were a few missing values encountered. The coordinates were then manually obtained using the online geographic tool(latlong.net). Still, the location coordinates for two neighborhoods couldn't be obtained which upon further online research revealed that the areas are non-commercial ones and hence, could be left out.
Next, while searching for hotels in the area and fetching the data using Foursquare API, there were a few duplicate entries obtained since the neighborhoods radius overlapped. The duplicate entries were removed using the hotel id as the unique key and using the distance parameter to assign it to the neighborhood whose center was the closest. Similarly, when exploring for venues in the neighborhoods, again there were duplicate entries obtained and thus were dropped using the venue id as the key and using the distance parameter to assign it to the neighborhood whose center was the closest.

Figure 1: Raleigh Hotels

	Id	Neighbourhood	Name	Address	Latitude	Longitude	Distance	Category
0	5b338e832a7ab6002c75d7d0	Anderson Heights	Comfort Inn Raleigh Midtown	[1001 Wake Towne Dr, Raleigh, NC 27609, United...	35.824470	-78.624070	1439	Hotel
1	56e73487498e53fd654e55ef	Crabtree Valley	courtyard	[Raleigh, NC, United States]	35.834861	-78.673948	631	Hotel
2	5367bfb7498ec156685fe45b	Crabtree Valley	Hilton Garden Inn Raleigh /Crabtree Valley	[3912 Arrow Dr, Raleigh, NC 27612, United States]	35.835510	-78.673219	658	Hotel
3	4bc37c7adce4eee189a0719d	Crabtree Valley	Courtyard Raleigh Crabtree Valley	[3908 Arrow Drive, Raleigh, NC 27612, United S...	35.835099	-78.673796	630	Hotel
4	4bd2330f462cb7134fe1db07	Glenwood South	Days Inn by Wyndham Raleigh Downtown	[300 North Dawson Street (at W Lane St), Ralei...	35.784282	-78.642445	461	Hotel

Figure 2: Raleigh Venues

	Id	Neighbourhood	Venue	Venue Latitude	Venue Longitude	Distance	Venue Category
0	4d45a17014aa8cfa5e4e743d	Anderson Heights	Fallon Park	35.815168	-78.637047	307	Park
1	517abc13e4b03fca3ead4a0c	Anderson Heights	Crabtree Creek Trailhead	35.821286	-78.634901	461	Trail
2	4e00dda262e12fb08938acb2	Drewry Hills	Crabtree Creek Trail Entrance @ Marlowe	35.823670	-78.639982	532	Trail
3	4ea82f090aaf6e058655d91f	Anderson Heights	Calibre Chase Gym	35.820212	-78.630409	715	Gym
4	4c7175cf1f58199c331d407c	Hi-Mount	Kiwanis Park	35.814533	-78.630892	705	Park

While grouping the list of hotels using the neighborhoods to get the counts for the number of hotels in the area, neighborhoods having no hotels in the corresponding area were assigned zero. But the similar assignment was not done while retrieving the venue counts for the neighborhoods because the neighborhoods having no venues will not be ideal for a hotel location. The counts were then merged in the total_counts dataframe. There weren't any outliers in our data as we will be analyzing the neighborhoods and hence weren't taken into consideration.

Figure 3: Total Counts

	Borough	Neighbourhood	Latitude	Longitude	Hotel_Count	Venue_Count
0	Beltline	Anderson Heights	35.817863	-78.637782	4.0	3
1	Beltline	Avent West	35.778662	-78.716634	0.0	18
2	Beltline	Belvidere Park	35.798775	-78.619352	0.0	11
3	Beltline	Battery Heights	35.777058	-78.617563	0.0	12
4	Beltline	Bloomsbury	35.808897	-78.648599	0.0	1
5	Beltline	Boylan Heights	35.774159	-78.652102	1.0	18

3. Methodology

The project is designed to find optimal neighborhoods for the location of a new hotel in Raleigh, North Carolina. The neighborhoods should be such that there are a smaller number of already existing hotels so that the competition is not high but have a fair number of things to do in the vicinity. In the first step which included the data collection and gathering phase, we collected the data about the already existing hotels in the neighborhoods and their counts. We then also collected data about the number of nearby venues as well as their type in those neighborhoods. In the second step which will include the data analysis part, first we will examine if there is any correlation in the number of hotels and number of venues in the neighborhoods. We will also identify the frequency of occurrence of various categories of venues existing in the neighborhoods.
In the next step, we will cluster the neighborhoods based on the type of venues as well as the counts of hotels and nearby venues and then find the cluster having higher number of venue categories and counts which are generally of interest to visitors like restaurants, stores, attractions, etc. We will visualize the resulting categories/counts on a bar chart to identify the promising cluster of neighborhoods. In the final step, from the identified cluster we will extract the list of neighborhoods that meet our criteria of three or less than three hotels and more than 19 venues in the vicinity. We will present the map of identified neighborhoods to visually see if the neighborhoods are in diverse locations. These neighborhood centers could be the starting point for further exploration and finding the best possible location.

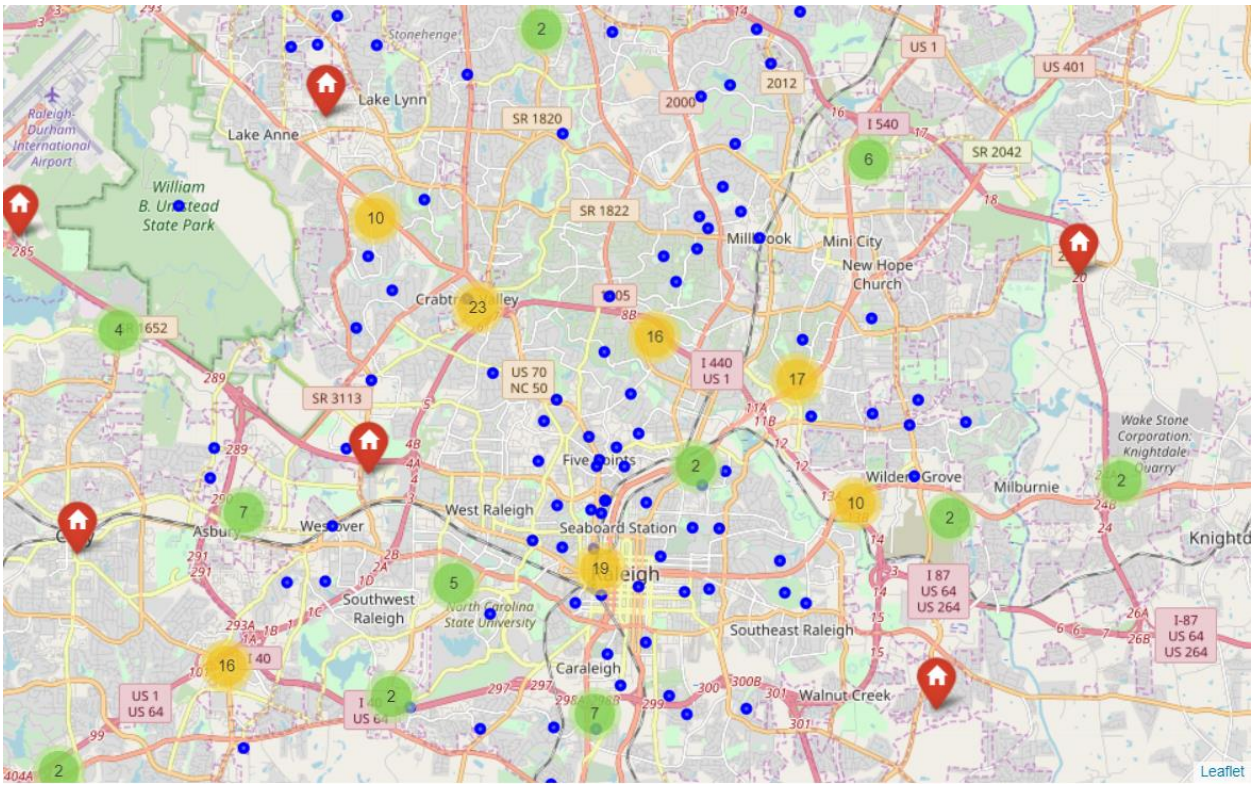
3.1 Exploratory Data Analysis

3.1.1 Distribution and Number of Hotels in Neighborhoods

The City of Raleigh, also known as the City of Oaks, is the capital of North Carolina, the seat of Wake County, and one of the most economically diverse cities in the state. The Raleigh hotel market is divided

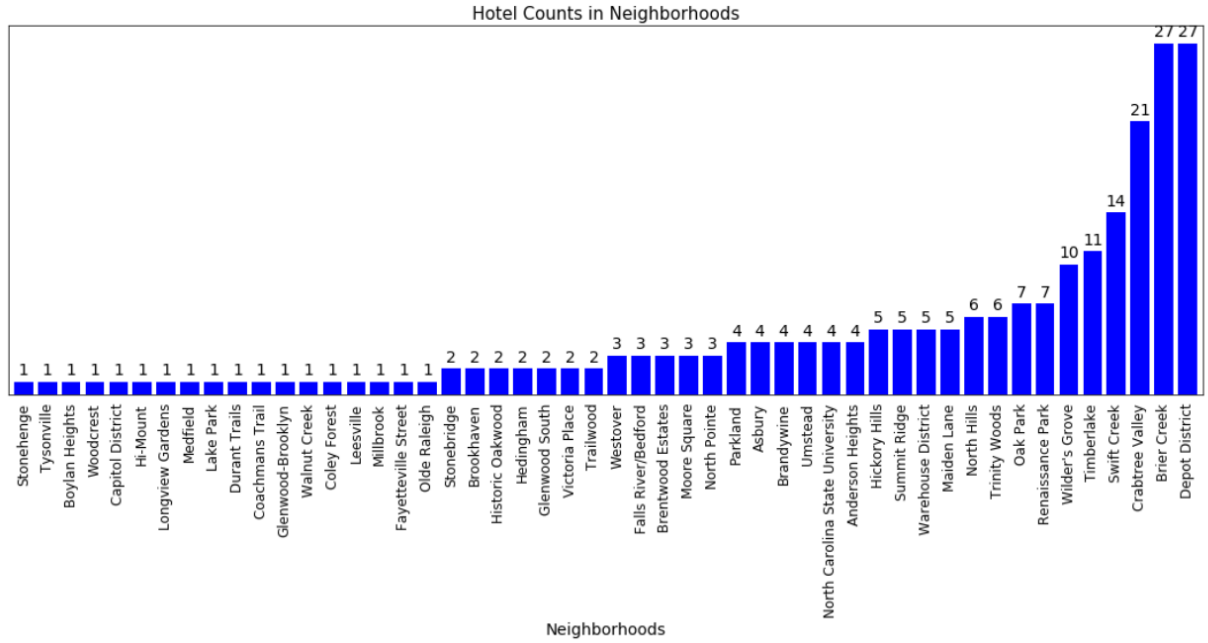
into several different submarkets that have individual demand generators and guest profiles. Hotel demand in each individual submarket is largely driven based on the proximate location to a variety of economic and leisure demand generators.
Let's visualize how the hotels are distributed currently in the region.

Figure 4: Distribution of hotels



As we see from the map there are a few regions which are not saturated with hotels yet and could be of interest.
Let's also visualize the top 50 neighborhoods with the highest number of hotel counts.

Figure 5: Top 50 Neighborhoods with highest Hotel Counts

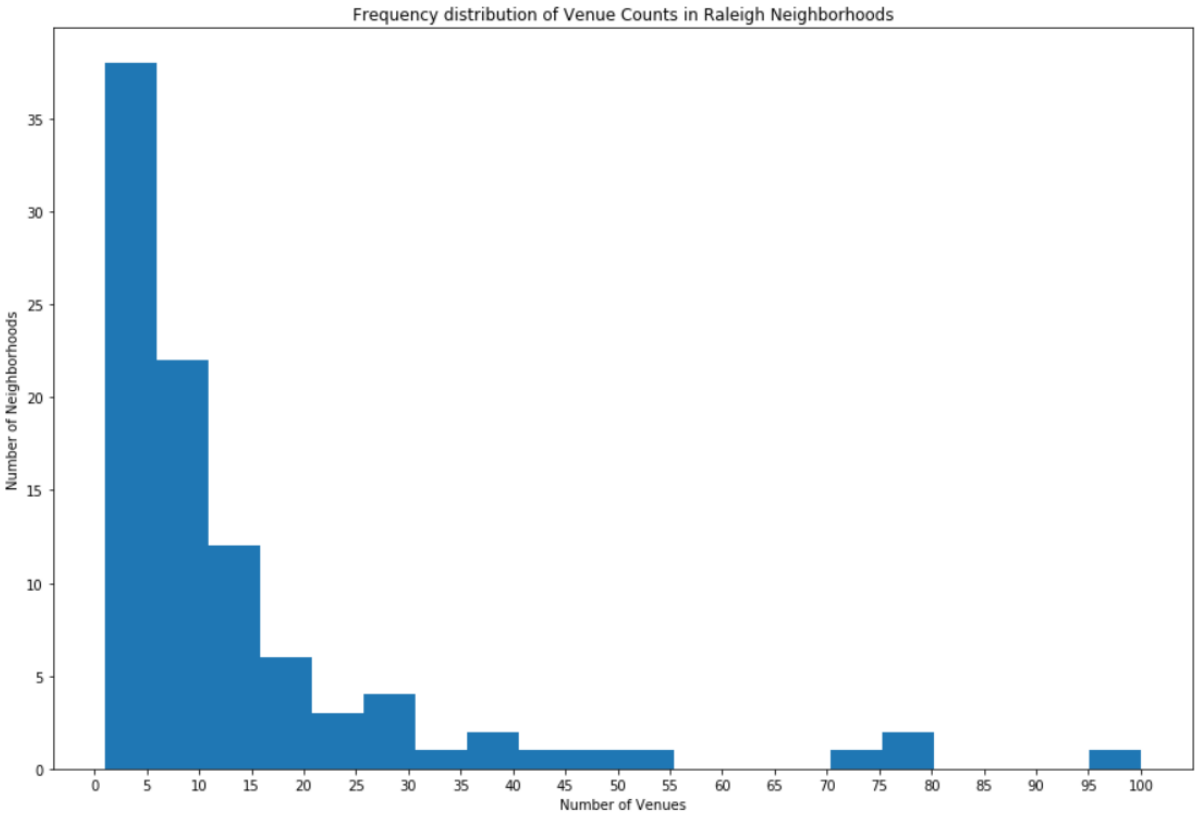


Depot District has the highest number of hotels followed by Brier Creek and Crabtree Valley.

3.1.2 Distribution and Number of Venues in Neighborhoods

Next, let's see the distribution of count of venues among neighborhoods, i.e., what is the common number of venues present in Raleigh neighborhoods.

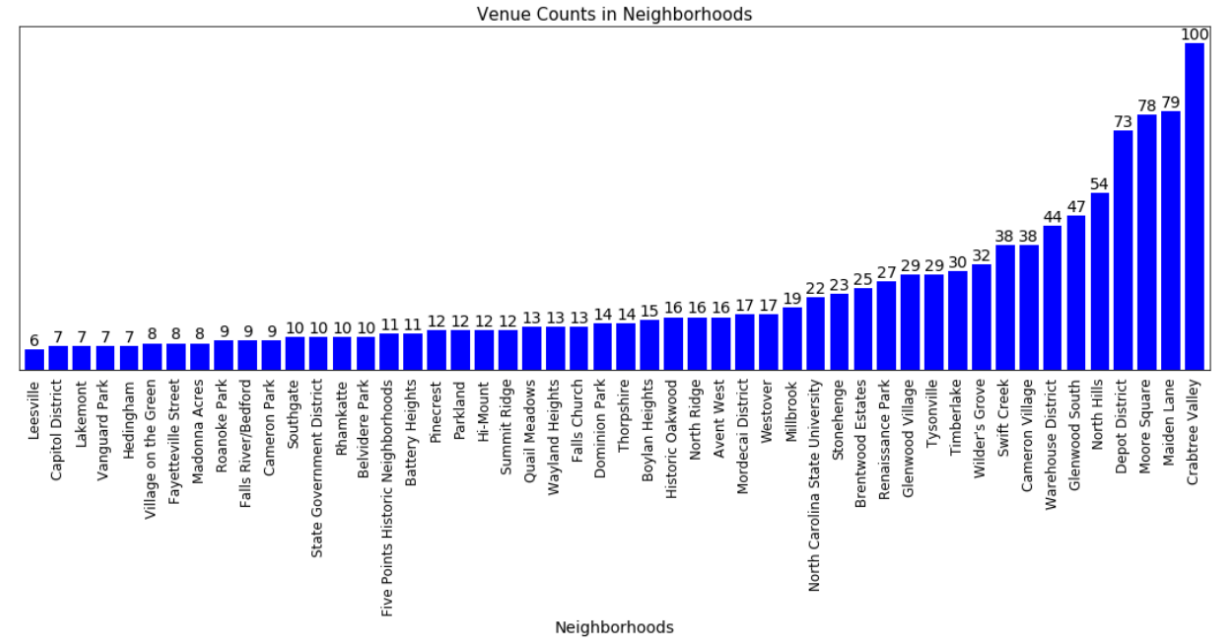
Figure 6: Distribution of Venue Counts



Looks like most of the neighborhood centers have a low concentration of venues and it will be difficult to find appropriate neighborhoods that have not been already saturated with hotels.

Let's have a look at the top 50 neighborhoods with the highest number of venue counts.

Figure 7: Top 50 Neighborhoods with highest Venue Counts

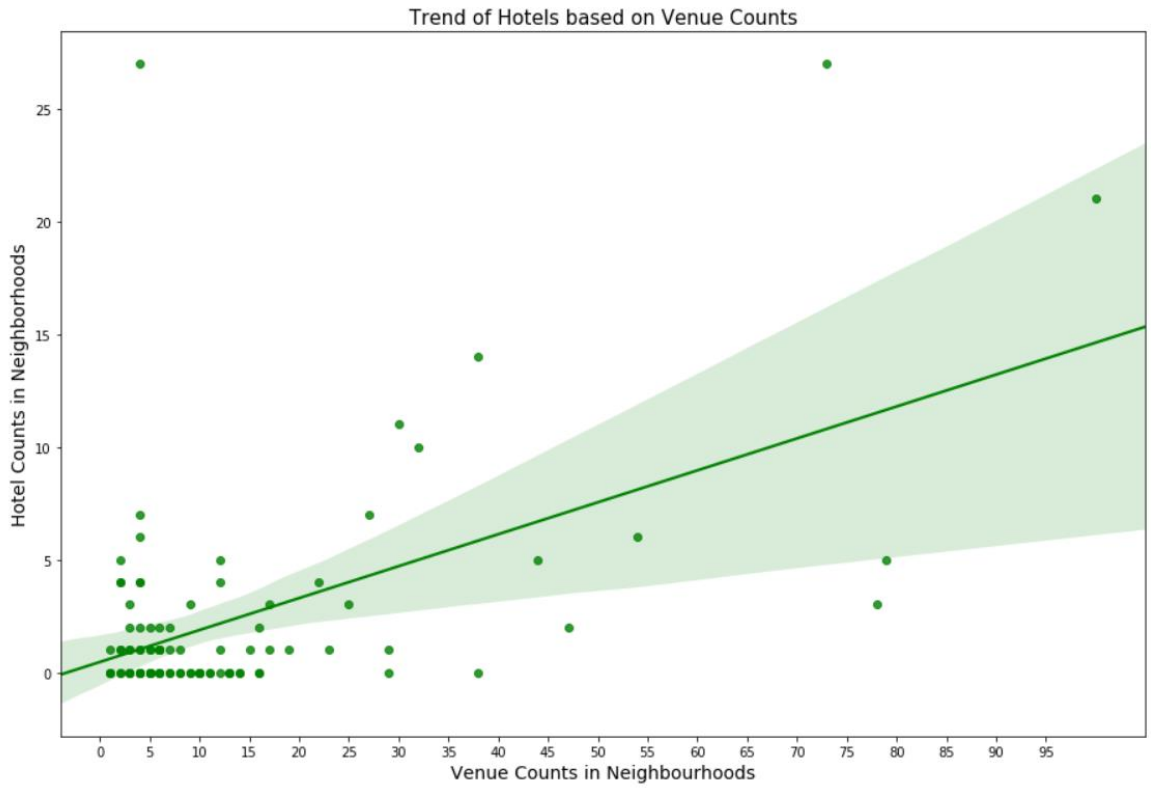


We see that the Crabtree Valley has the highest number of venues followed by Maiden Lane, Moore Square and Depot District. Both the Crabtree Valley and the Depot District have a higher concentration of hotels as well as nearby venues. Does that mean the hotel count is related to venues count?

3.1.3 Relationship between Hotel Counts and Venues Count in Neighborhoods

Let's see if there is any relation between the hotels count and the venues count in the neighborhoods:

Figure 8: Relationship between Hotel and Venue counts

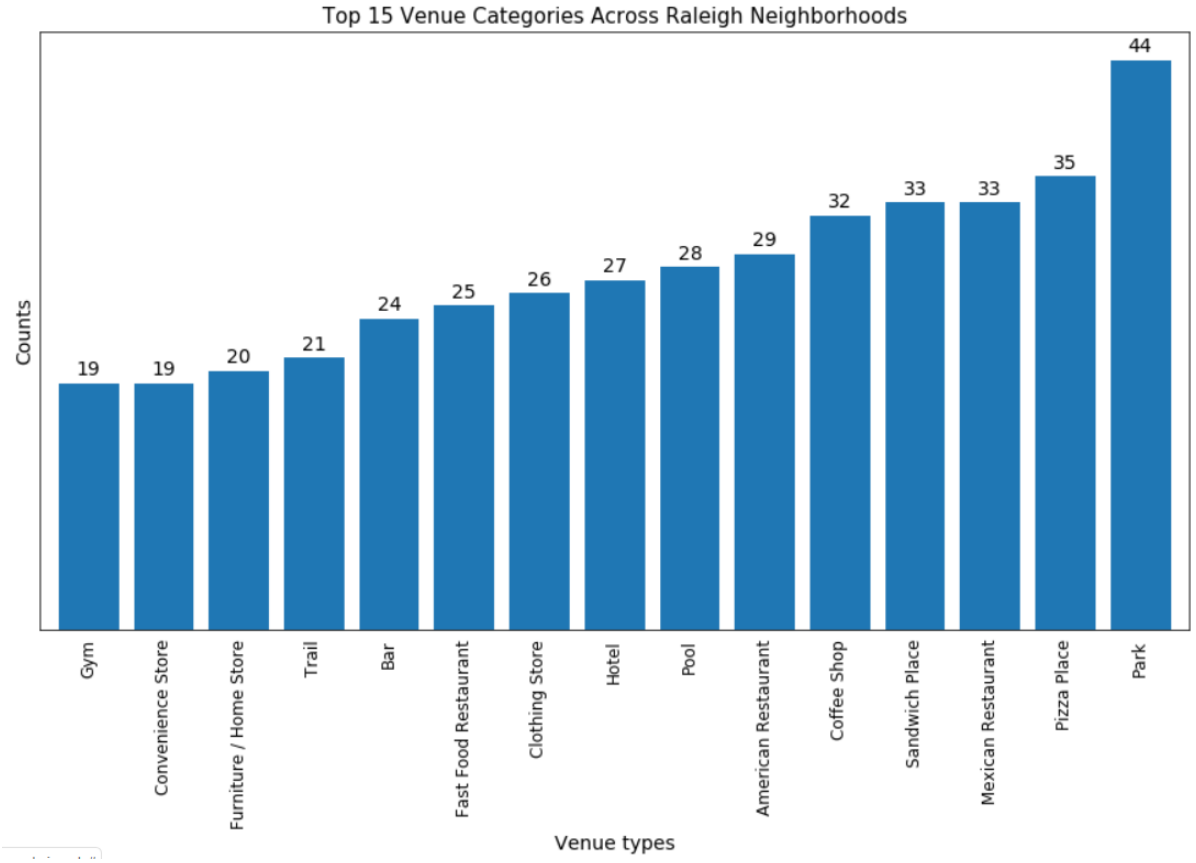


As we see from the plot above, with an increase in the number of venues in neighborhoods, the number of hotels tend to increase. Also, there appears to be a few neighborhoods where the number of hotels is less although the number of nearby venues is more. These are the neighborhoods we are interested in. From the plot, we can estimate that neighborhoods having venue counts more than or equal to 20 and having number of hotels less than 4 will be a good fit.

Now, since the number of venues could be made up of any type of venues, we will also need to group the neighborhoods based on the type of venues so as to get a better idea of regions of interest to the visitors.

3.1.3 Grouping the Neighborhoods by the type of venues

If we sum up the number of venues obtained from neighborhoods according to their categories, we get the most common types of venues in Raleigh neighborhoods:



This is interesting as there are not just restaurants or cafes nearby but also quite a lot of places like, park, trails, etc. for leisure activities. These are just the total counts and maybe biased towards neighborhoods having too many venues. Since different submarkets in Raleigh have individual demand generators and guest profiles, we need to find the neighborhoods based on the type of venue categories existing in them and thus we need to get the proportion being contributed by each category towards the neighborhood.

Figure 9: Mean of Frequency of Occurrence of each Category

	Neighbourhood	ATM	Accessories Store	Adult Boutique	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	...	Video Store	Vietnamese Restaurant	Warehouse Store	Weight Loss Center
0	Anderson Heights	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000
1	Asbury	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000
2	Avent West	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	...	0.055556	0.000000	0.0	0.000000
3	Battery Heights	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000
4	Belvidere Park	0.090909	0.00	0.000000	0.090909	0.000000	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000
5	Biltmore Hills	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000
6	Bloomsbury	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000
7	Boylan Heights	0.000000	0.00	0.000000	0.055556	0.055556	0.00	0.000000	0.111111	0.000000	...	0.000000	0.000000	0.0	0.000000
8	Brandywine	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000

Now, if we have a look at the ten most common venue in each neighborhood:

Figure 10: Most Common Venues

	Neighbourhood	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	Anderson Heights	Gym	Park	Trail	Dry Cleaner	Electronics Store	Event Service	Event Space	Eye Doctor	Farmers Market	Fast Food Restaurant
1	Asbury	Thrift / Vintage Store	Convenience Store	Baseball Field	Taco Place	Yoga Studio	Financial or Legal Service	Eye Doctor	Farmers Market	Fast Food Restaurant	Filipino Restaurant
2	Avent West	Park	Lawyer	Cosmetics Shop	Grocery Store	Pizza Place	Convenience Store	Discount Store	Sandwich Place	Chinese Restaurant	Shopping Plaza
3	Battery Heights	Park	Fried Chicken Joint	Caribbean Restaurant	Hardware Store	Department Store	Southern / Soul Food Restaurant	Discount Store	Automotive Shop	Scenic Lookout	Event Space
4	Belvidere Park	Coffee Shop	ATM	Home Service	Gas Station	Park	Seafood Restaurant	Gym / Fitness Center	Pawn Shop	Hot Dog Joint	American Restaurant

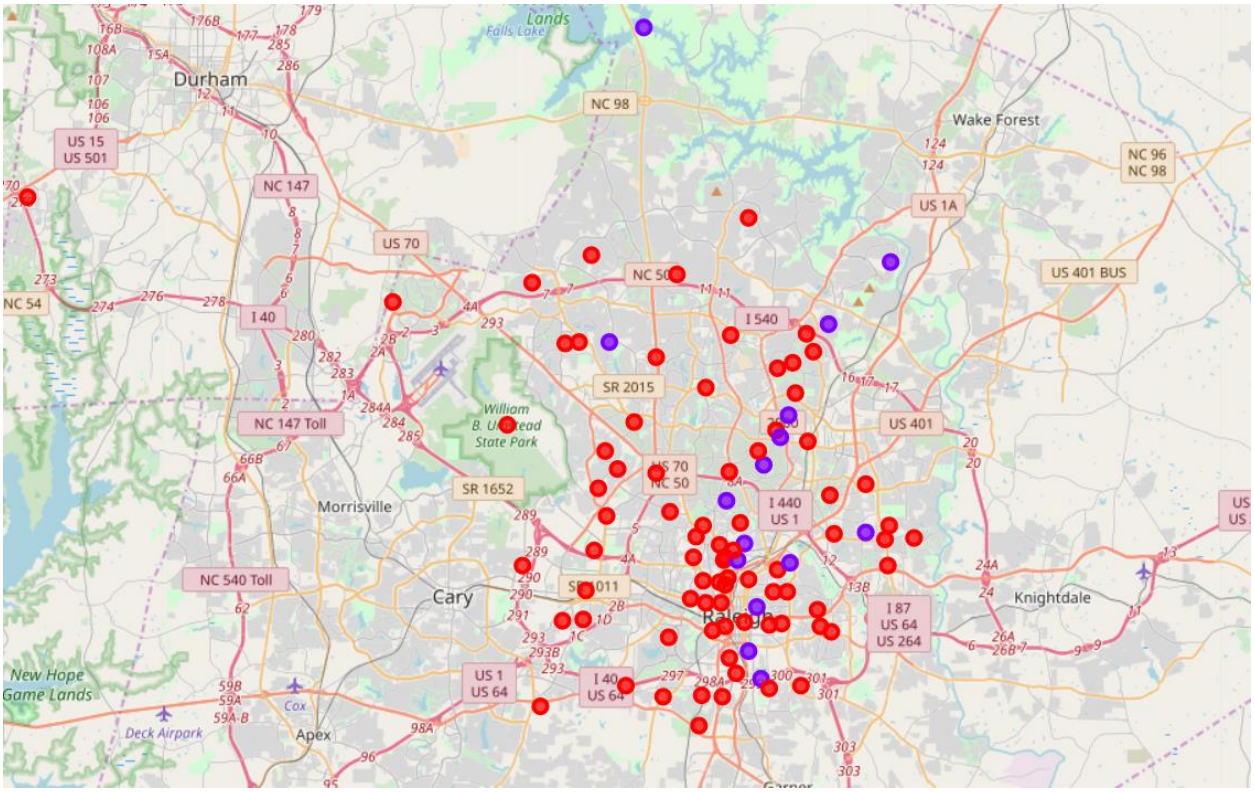
Although we can see the most common venues of each neighborhood from the above dataframe, we cannot exactly infer which neighborhood will be the best fit just by looking at it. We need to further group them based on the type of venues most common among them. Hence, we will be using the K-means clustering algorithm to find the groupings in our data.
Since we also need to know the number of existing hotels and venues to decide on the optimal location for the hotel, we need to include those parameters as well while clustering the data for analysis.

3.2 K-Means Clustering

Based on the type of venues in the neighborhoods, the number of already existing hotels and nearby venues, we have clustered the neighborhoods. Only the neighborhoods having venue counts greater than zero have been used for clustering since if there are no venues nearby, it doesn't make sense to choose that neighborhood and we also cannot obtain the categories of venues present in those neighborhoods. Such neighborhoods are basically just outliers and will not be used in analysis.

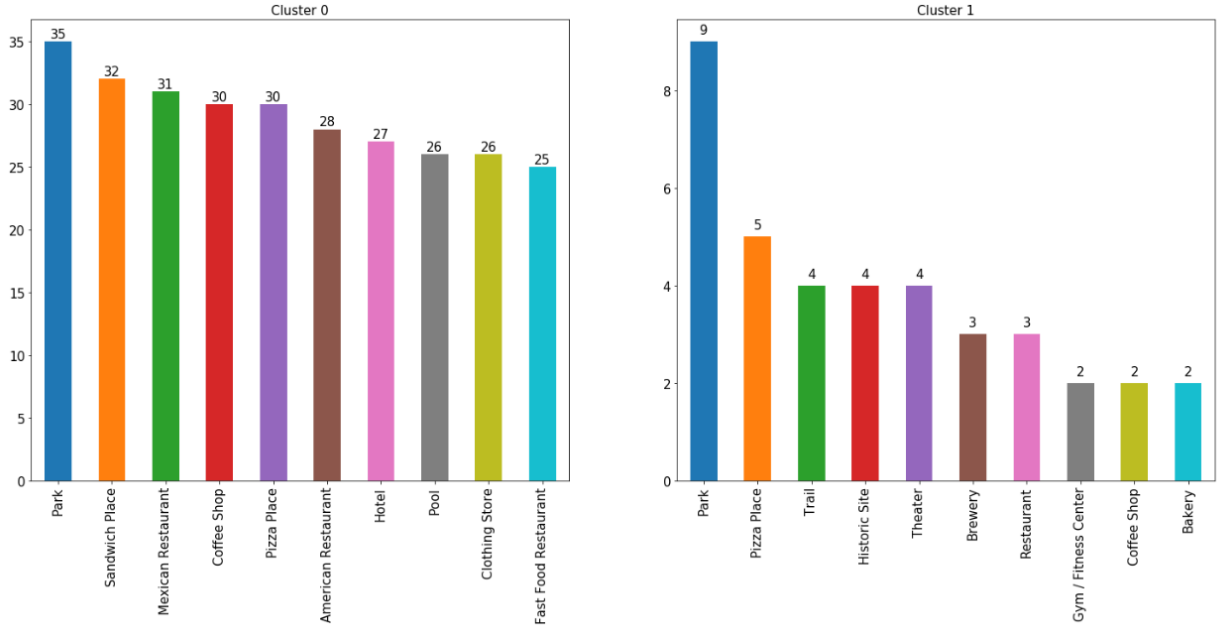
Let's visualize the created clusters on map.

Figure 11: Neighborhood Clusters



To examine the clusters further, we identified the top ten venue categories comprising each cluster.

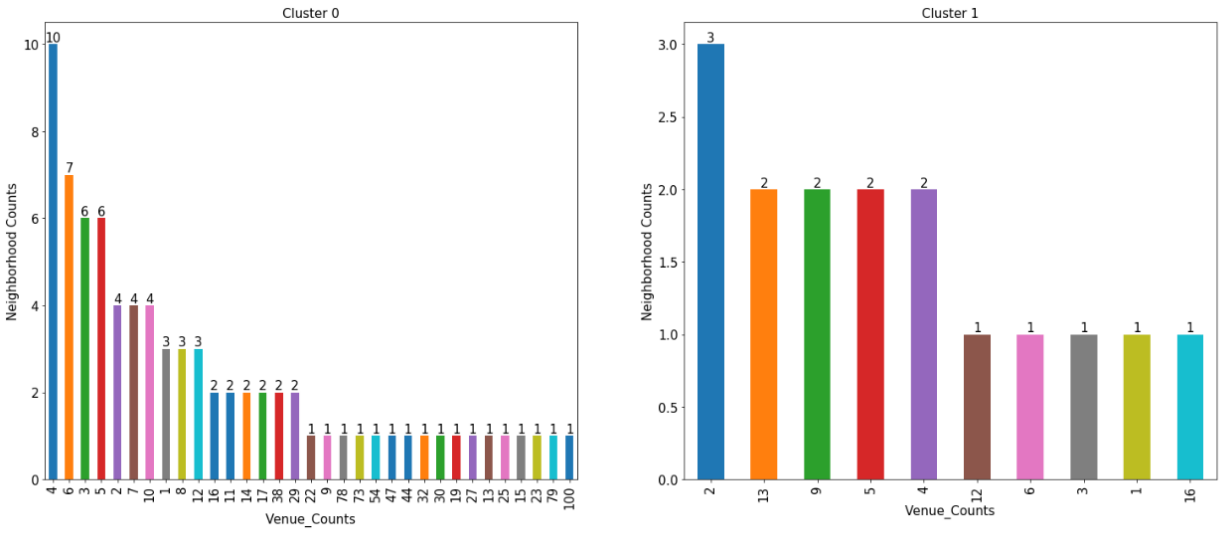
Figure 12: Top 10 Venue Categories in Each Cluster



From the plots above, we see that among the top ten categories in each of the clusters, the venue categories which may be of interest to visitors like restaurants, shopping stores, etc. are higher in cluster 0.

Let's also visualize the venue counts along with the corresponding count of neighborhoods with those venue counts among clusters.

Figure 13: Frequency of Venue Counts in each Cluster



From the plots above, we also see that the neighborhoods with higher venue counts (more than or equal to 20) are also present in cluster 0.

Since Cluster 0 meets our criteria with venue counts more than or equal to 20 and also has a higher variety of venue categories of interest, we shall proceed with it and further examine the neighborhoods in this cluster.

If we make a word cloud of neighborhoods in cluster 0 to visualize the neighborhoods with maximum venues:

Figure 14: Neighborhoods in Cluster 0

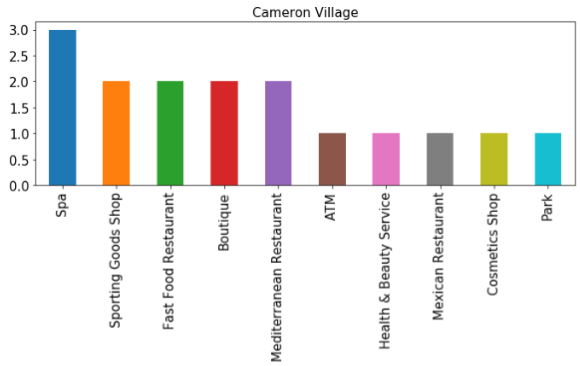
We are able to identify few neighborhoods which seem to have good number of venues.
If we extract them from the cluster based on our criteria of already existing hotels less than or equal to three and nearby venues more than 19, we get our final list of neighborhoods:

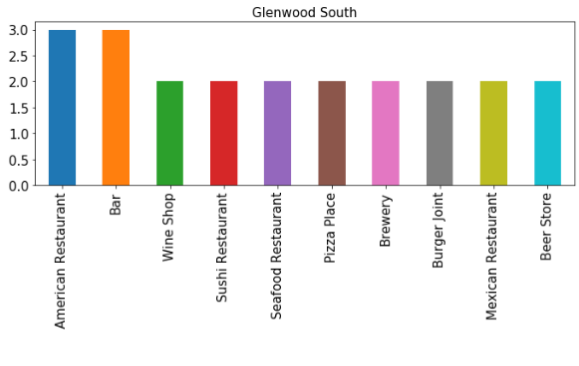
Figure 15: Final List of Neighborhoods

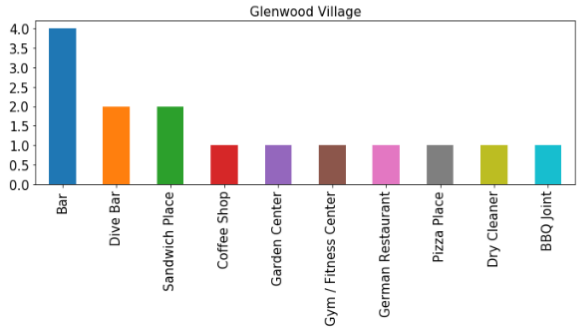
	Cluster_Labels	Region	Neighbourhood	Latitude	Longitude	Hotel_Count	Venue_Count
7	0	Beltline	Cameron Village	35.794548	-78.656789	0.0	38
17	0	Beltline	Glenwood South	35.785436	-78.647354	2.0	47
18	0	Beltline	Glenwood Village	35.792926	-78.645399	0.0	29
25	0	Beltline	Moore Square	35.777559	-78.635786	3.0	78
39	0	North Raleigh	Brentwood Estates	35.829181	-78.593451	3.0	25
63	0	North Raleigh	Stonehenge	35.884593	-78.679962	1.0	23
78	0	West and Southwest Raleigh	Tysonville	35.820706	-78.704727	1.0	29

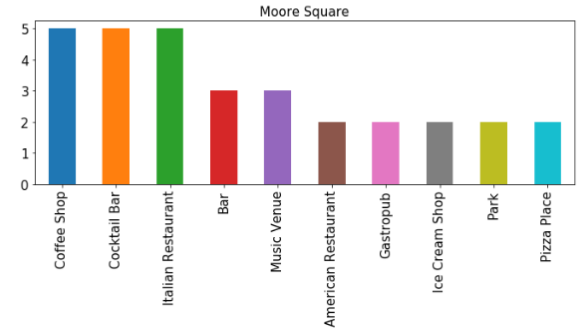
Let’s see the top 10 venue categories in these neighborhoods:

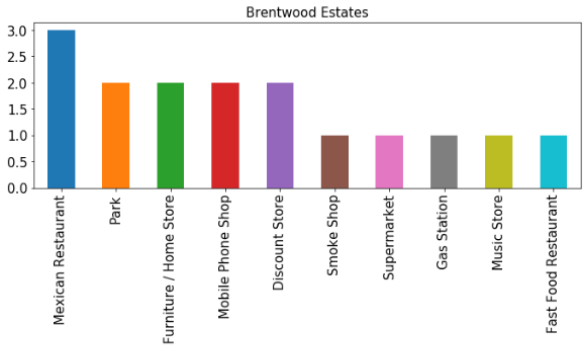
Figure 16: Top 10 Venue Categories in Final list of Neighborhoods

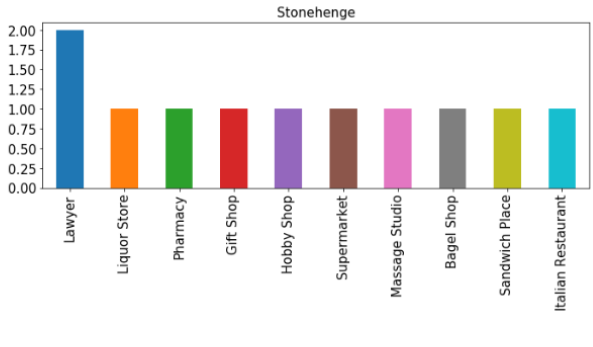


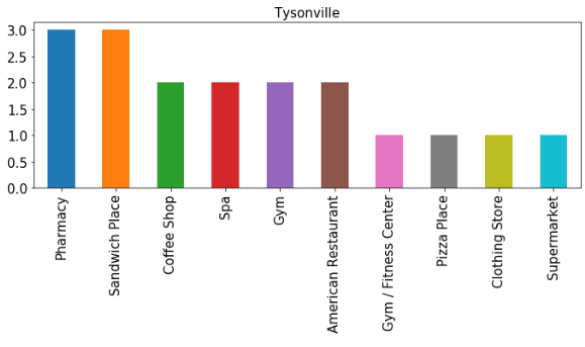








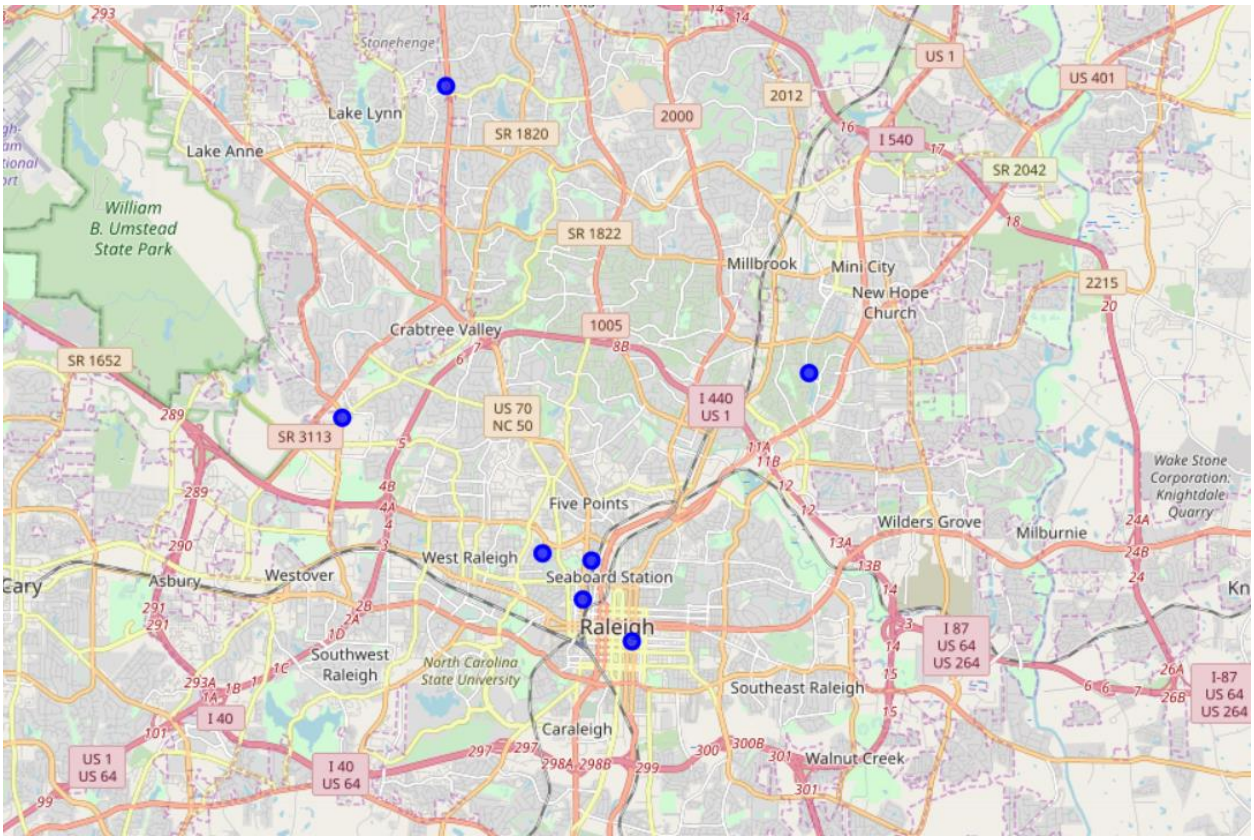




As we can infer from the plots, most of the venues in these neighborhoods are restaurants, bars etc. which are one of the things visitors look for nearby a hotel. Hence, these could be good neighborhoods for the location of hotel.

Let's also visualize these final neighborhoods on a map to see how they are distributed.

Figure 17: Distribution of Obtained Neighborhoods



As we can see from the map, the neighborhoods are well spread out and thus gives a variety of places that could be considered.

This concludes our analysis. We have identified 7 neighborhoods with lower concentration of hotels and higher concentration of venues and they are spread out in three of the four regions in Raleigh, i.e., Beltline, north Raleigh and west-southwest Raleigh thus giving options for the choice of location of hotel near or between downtown, university or research triangle park. These neighborhoods should serve as the reference point for initiating the ground level exploration of potential hotel location.

4. Results

Our analysis shows that although the region appears to be saturated at first sight, there are a lot of potential areas which are still developing and given the booming reputation of city as a hub, could be a good investment choice. Although it seems fit to choose downtown or the research triangle park area as the location but given their popularity, they might already be fully saturated with the hotels. Besides, as

the region keeps growing, the other neighborhoods could be expected to grow and should be taken into consideration. Hence, it becomes necessary to analyze the neighborhoods before reaching to any conclusion. Thus, we extracted the list of all the neighborhoods and explored them in detail further. Upon exploring these neighborhoods, we got the list and counts of hotels and venues in the neighborhoods and we noticed that there is some trend between hotel counts and venue counts in the neighborhood, i.e., the hotel counts tend to increase with the increase in venue counts. Of course, there are other factors on which the hotel count depends like proximity to tourist attractions, businesses, levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of neighborhood etc. but the venue count factor also matters. From that we observed that there are a few neighborhoods which have a good amount of venues and can be good choices for hotel location. On further performing the clustering on neighborhoods venues and hotels data, we got the similar kinds of neighborhoods in separate clusters and examining them based on the categories of venues present and their corresponding counts, we found the cluster which had a higher concentration of venues which may be of interest to visitors like restaurants, amenities, etc. From that cluster, we then extracted the final list of neighborhoods which had a hotel count of less than or equal to three but nearby venues count of more than 19 and visualized them on the map to assure they are well spread out thus giving choices in separate regions.

4. Discussion

This list of neighborhoods and their centers could be used as a starting point for further street-level analysis of the possible locations. This is a potential list and does not imply that these are actually optimal locations for a hotel. There are, of course, other factors like real estate, planned projects, local community laws, hotel sales data of previous quarters, etc. which need to be considered before deciding on the final location. The purpose of this analysis was to only get the possible neighborhood choices for a new hotel and it is possible that regardless of low hotel density, it might not be viable to open a hotel there due to some other reasons. Thus, this recommended list should be considered only as a starting point and if further detailed analysis could be done and more data obtained regarding profits and sales projections, it can lead to getting the best possible location.

5. Conclusion

After all the analysis we did, we identified a few neighborhoods which met our criteria of low hotel density and higher nearby venues density which thus befits our purpose of this project. This analysis can aid stakeholders in narrowing down their search for optimal location for a new hotel in the region. There are a lot of other factors which need to be considered for a hotel like capital, financial projections, zoning and permit issues, etc. before a final decision can be taken but a location choice gives a starting point for all these further steps. This analysis can be further improved if more data related to demand, occupancy, revenue, customer surveys could be obtained for the existing hotels in the region but for now, should be contributory towards exploring few good choices.

References:

- https://www.hvs.com/article/8144-market-pulse-raleigh-nc
- https://en.wikipedia.org/wiki/Raleigh,_North_Carolina_neighborhoods
- https://smartasset.com/mortgage/tech-workers-2019