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