1. Introduction

1.1 Background and Business Problem

How can a new retail business determine the ideal location to open a store in the city of Atlanta?

Finding the right location for a retail store is crucial to the long-term growth and sustainability of the business. In a world where physical clothing stores are facing a decline due to the rapid adoption of online shopping by younger generations, it is even more important to choose a suitable site for opening a retail store.

In order to enjoy healthy profits, retailers have to find the best possible way to attract crowds. Factors such as proximity to densely populated areas, access to parking lots and major roads as well as the general nature of the area play an important part in choosing the most favorable location for maximizing traffic to the store.

The city of Atlanta provides a host of promising locations containing a variety of venues which can complement a retail store by attracting crowds. Venue characteristics such as type of establishment, neighborhood population and position within the city can be leveraged to decide the ideal location.

The *target audience* for this analysis would be the owner of the business or the people in charge of marketing and research within the company as they can use the insights and recommendations to make the best decision regarding the location of their new retail store.

1.2 Data Description

Atlanta neighborhood information was retrieved from *Wikipedia* which also consisted of the *population* and *neighborhood planning unit (NPU)* for each instance [1]. There are a total of **161 neighborhoods** in the city which are separated into **25 NPUs**.

Coordinates (latitude and longitude) of each neighborhood were retrieved using the Python library called *geopy*. In some cases, the coordinates were not available in which case they had to be manually searched on *Google*.

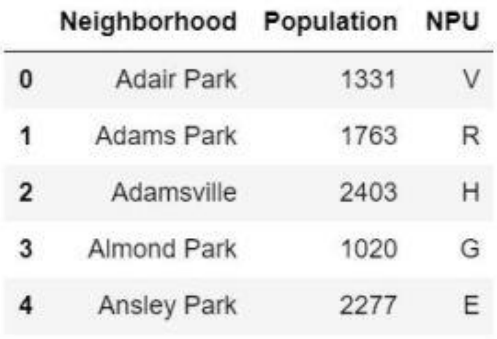
The *Foursquare* location app was used to search and retrieve venues across neighborhoods in the city of Atlanta using latitude and longitude data which provides important information that can be used to determine the most promising places to open the retail store. In Python, the app can be utilized by calling the Foursquare API using custom user credentials. The data was retrieved in the form of a *JSON* file which contained the coordinates, venue type, venue name, distance from neighborhood and other information. All this data was combined into a single dataset that was used for analysis.

2. Methodology

The neighborhoods data was saved as a *.csv* file and imported into the Python *Jupyter Notebook* using the *Pandas* library in a *dataframe* format.

The dataset consists of 3 columns:

- **Neighborhood**: Lists all the neighborhoods in the city
- **Population**: Lists the population in each of the neighborhoods
- **NPU**: This is known as the *Neighborhood Planning Unit*. Each NPU consists of several neighborhoods.



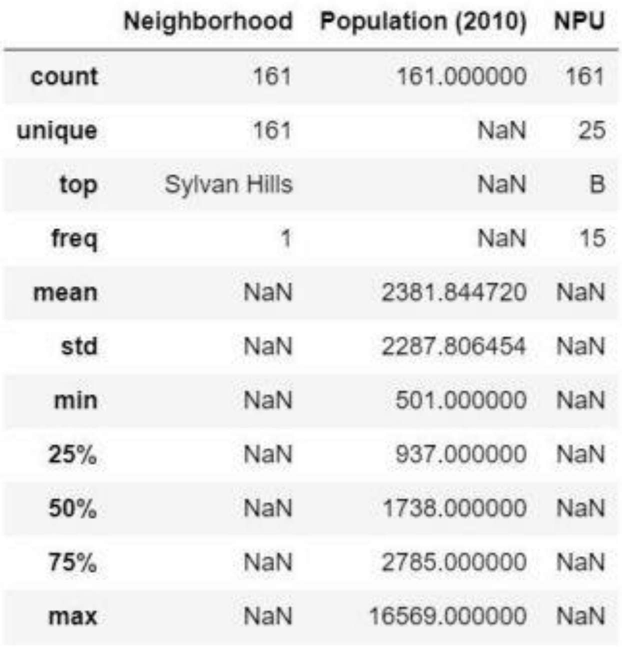
	Neighborhood	Population	NPU
0	Adair Park	1331	V
1	Adams Park	1763	R
2	Adamsville	2403	H
3	Almond Park	1020	G
4	Ansley Park	2277	E

Fig 1. Atlanta neighborhood population and NPU table

Above, the first five instances of the dataset are displayed in a dataframe structure. Both *Neighborhood* and *NPU* columns are of type *Object* whereas the *Population* attribute is of type *Integer*.

The summary statistics of the dataset is displayed in the table below. From these stats, we can see that the mean population of a neighborhood is **2381.84472** and the largest population is **16569**.

Another thing to note is that NPU *B* contains the most number of neighborhoods with a total of **15**. The *unique* stat confirms that the dataset in fact 25 unique NPUs. Note that the Population column name was changed in *Fig 1* from the one in *Fig 2*.



	Neighborhood	Population (2010)	NPU
count	161	161.000000	161
unique	161	NaN	25
top	Sylvan Hills	NaN	B
freq	1	NaN	15
mean	NaN	2381.844720	NaN
std	NaN	2287.806454	NaN
min	NaN	501.000000	NaN
25%	NaN	937.000000	NaN
50%	NaN	1738.000000	NaN
75%	NaN	2785.000000	NaN
max	NaN	16569.000000	NaN

Fig 2. Summary statistics of the neighborhoods dataset

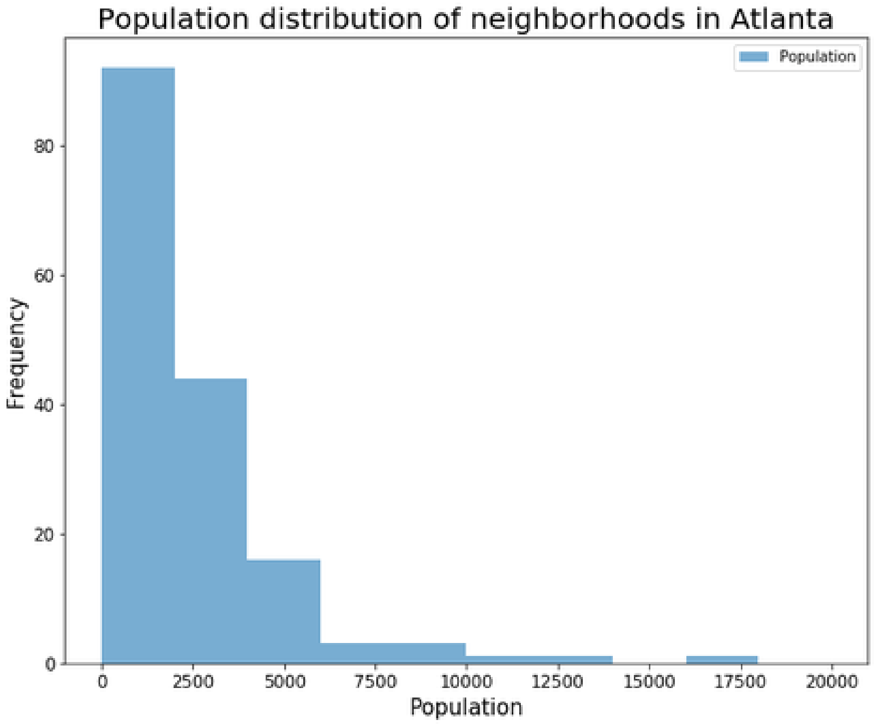


Fig 3. Histogram displaying the population distribution of the neighborhoods

From the distribution above, it appears that most of the neighborhoods have a population of under **2000**. In fact, the distribution is highly **right skewed**. From the summary stats table we saw earlier, the **upper-quartile for population is 2785** which means that **75% of the neighborhoods have a population under 2785**.

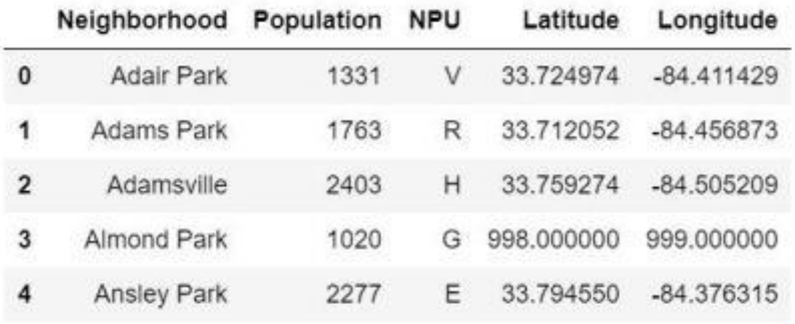
This number falls on the extreme left of the histogram, which is one way to get a measure of the skewness. For a perfectly normal distribution, the median would lie somewhere around the center of the graph.

2.1 Retrieving coordinates for each neighborhood

In order to use the Foursquare API, the latitude and longitude of each neighborhood need to be determined, which is retrieved using Python's *geopy* library.

The coordinates are retrieved by looping through the neighborhoods list and calling the location function for each neighborhood, which are appended together into a list.

After combining the coordinates data to our main dataset, it was discovered that some of the values were not able to be retrieved using the library. For example, coordinates for *Almond Park* were not available, for which an *except* object was included which would assign dummy values *(998,999)* for such neighborhoods as shown in *Fig 4*.



	Neighborhood	Population	NPU	Latitude	Longitude
0	Adair Park	1331	V	33.724974	-84.411429
1	Adams Park	1763	R	33.712052	-84.456873
2	Adamsville	2403	H	33.759274	-84.505209
3	Almond Park	1020	G	998.000000	999.000000
4	Ansley Park	2277	E	33.794550	-84.376315

Fig 4. The original dataset with the retrieved coordinates

On a closer look, there turned out to be **38 neighborhoods** with missing coordinates. To reduce time and effort, these missing coordinates were manually looked up on *Google* and substituted into the dataset.

When the neighborhoods were plotted on a map using the *Folium* library, coordinates for **5 neighborhoods** appeared to be wrongly retrieved as they are clearly far from Atlanta or Georgia on the map, shown in *Fig 5*.

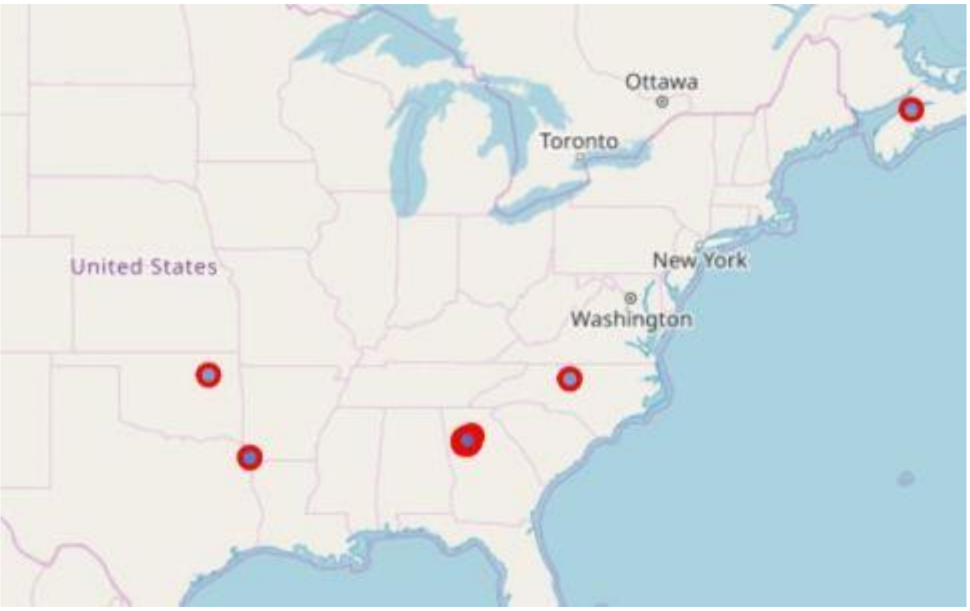
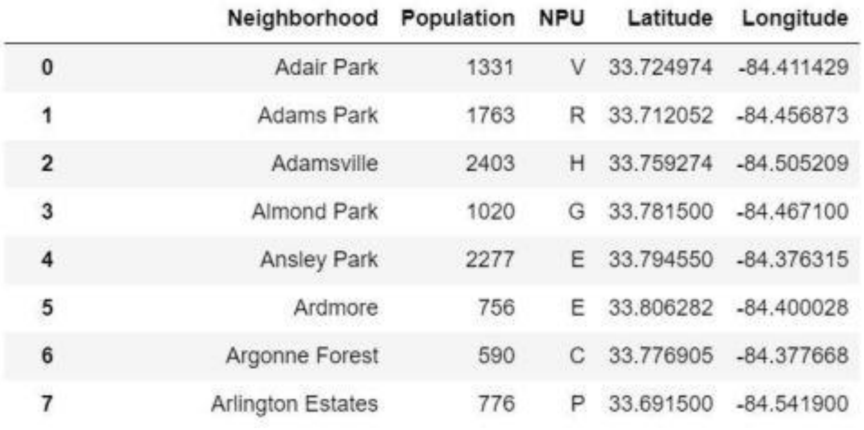


Fig 5. Folium map showing the initial location of the neighborhoods

To fix this, these outlying neighborhoods were looked up manually on *Google* as well. Once all the missing coordinates were retrieved, they were put into a *list* of *tuples* which allowed for easy extraction using a f*or loop*.

By looping through these tuples, the dummy values in the dataframe were replaced by the actual coordinates, whereas the 5 misplaced coordinates were appended at the end of the dataset.

The neighborhoods known as *Pamond Park* and *Kings Forest* were neither found using the *geopy* library nor the internet. Therefore, these neighborhoods were removed from the data for the sake of simplicity. Following a re-indexing, the resulting dataframe contained **159 neighborhoods**.



	Neighborhood	Population	NPU	Latitude	Longitude
0	Adair Park	1331	V	33.724974	-84.411429
1	Adams Park	1763	R	33.712052	-84.456873
2	Adamsville	2403	H	33.759274	-84.505209
3	Almond Park	1020	G	33.781500	-84.467100
4	Ansley Park	2277	E	33.794550	-84.376315
5	Ardmore	756	E	33.806282	-84.400028
6	Argonne Forest	590	C	33.776905	-84.377668
7	Arlington Estates	776	P	33.691500	-84.541900

Fig 6. Dataset following the retrieval and manipulation of the neighborhood coordinates

A snippet of the dataset following this exercise is shown in *Fig 6*. This organized dataset contains the initial neighborhood data needed to explore venues using the neighborhood name, latitude and longitude.

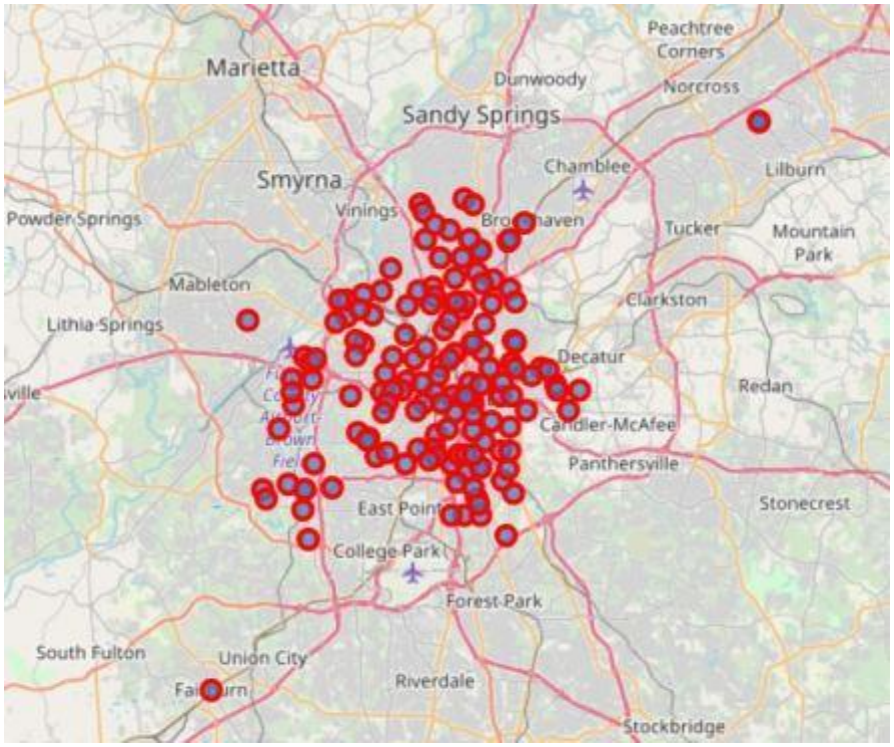


Fig 7. Map of Atlanta displaying all the 159 neighborhoods

The map in *Fig 7* shows all 159 neighborhoods scattered densely across Atlanta. The dataset was further narrowed down by filtering out neighborhoods that lie outside the **I-285 circle**. This is the interstate that circles the city and can be seen in *Fig 7* enclosing most of the neighborhoods.

Since the aim was to find areas of maximum traffic, it is more likely that areas within the circle are more densely population than the ones outside of it.

In order to visualize the density of the various neighborhoods, the dataset was split into high population count neighborhoods and low ones. This was achieved by sorting the dataset by population and separating the dataset into the top and bottom halves.

From the map in *Fig 8*, no clear distinction was observed between the densely populated areas and the less dense ones. However, the dataset could still be reduced to include points inside the *I-285 highway circle*.

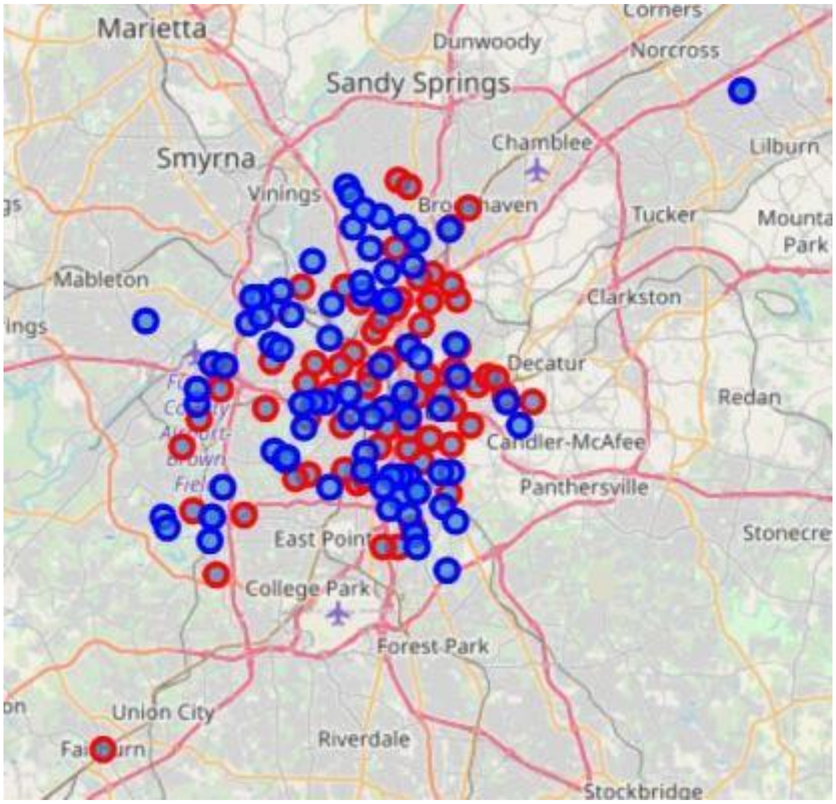
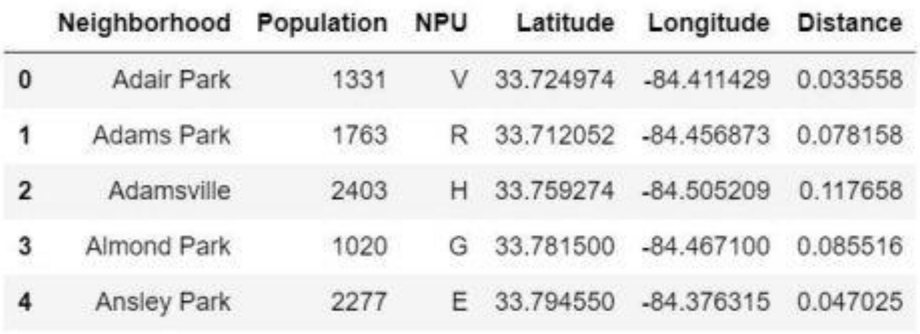


Fig 8. Neighborhood distribution in Atlanta based on low and high population

2.2 Finding neighborhoods that are inside the I-285 circle

For this, the distance from the center of Atlanta *(33.7490, -84.388)* to one of the coordinates on the circumference of the I-285 circle *(33.793697, -84.4878)* was first determined. Then the points that lie outside of this radius were removed.

The next step in the process was to determine the distance between the center of Atlanta to each of the neighborhoods using the *Euclidian* distance.



	Neighborhood	Population	NPU	Latitude	Longitude	Distance
0	Adair Park	1331	V	33.724974	-84.411429	0.033558
1	Adams Park	1763	R	33.712052	-84.456873	0.078158
2	Adamsville	2403	H	33.759274	-84.505209	0.117658
3	Almond Park	1020	G	33.781500	-84.467100	0.085516
4	Ansley Park	2277	E	33.794550	-84.376315	0.047025

Fig 9. Dataset with the *Distance* variable added

Fig 9 shows the dataset with the *Distance* variable added which contains the the distance from each neighborhood to the center of the city.

The general distance from the circle to the center of Atlanta was determined by the *Euclidian* distance as well following which, any neighborhoods with distances in the table in *Fig 9* that were less than this distance were removed from the dataset.

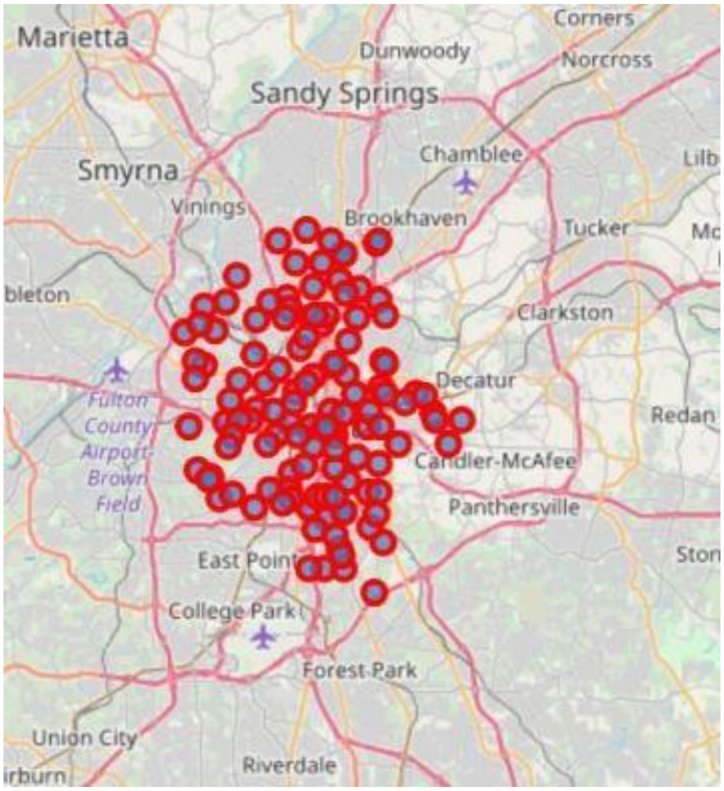


Fig 10. Map of Atlanta only showing neighborhoods inside the I-285 circle

The map in *Fig 10* shows the resulting map containing only the **127** neighborhoods that lie within the I-285 circle.

2.3 Using Foursquare API to retrive venues

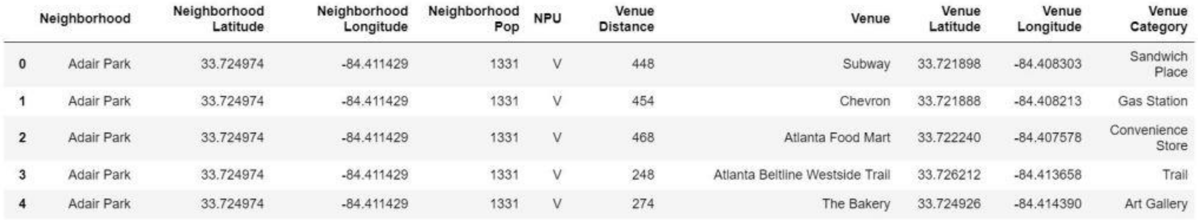
The Foursquare API is used to get all nearby venues for each neighborhood. First, credentials are used to initialize the API and then a *request* is made to the Foursquare website using certain parameters to retrieve venue information.

To discover venues around a certain coordinate, the Foursquare API contains the *explore* endpoint that allows retrieval of venues by supplying the **credentials** as well as the **coordinates**, **radius** (around the coordinate) and **limit** (number of results to retrieve).

Then, the request function is used to retrieve the data from Foursquare using the a *URL* that contains the aforementioned parameters, which in turn returns the information in *JSON* format.

The following data was retrieved for each venue within each neighborhood:

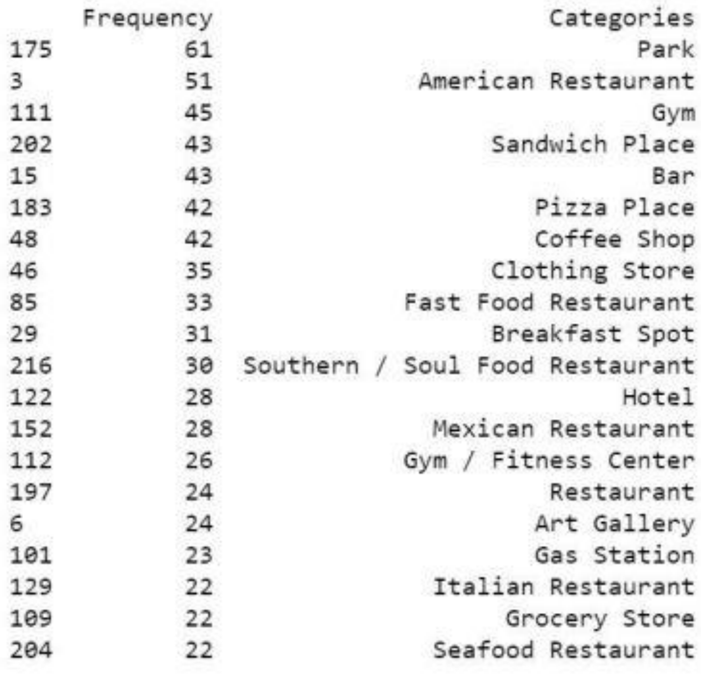
- *Distance from neighborhood*
- *Name*
- *Latitude*
- *Longitude*
- *Category*



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Neighborhood Pop	NPU	Venue Distance	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adair Park	33.724974	-84.411429	1331	V	448	Subway	33.721898	-84.408303	Sandwich Place
1	Adair Park	33.724974	-84.411429	1331	V	454	Chevron	33.721888	-84.408213	Gas Station
2	Adair Park	33.724974	-84.411429	1331	V	468	Atlanta Food Mart	33.722240	-84.407578	Convenience Store
3	Adair Park	33.724974	-84.411429	1331	V	248	Atlanta Beltline Westside Trail	33.726212	-84.413658	Trail
4	Adair Park	33.724974	-84.411429	1331	V	274	The Bakery	33.724926	-84.414390	Art Gallery

Fig 11. Dataset with appended venue information for each neighborhood

Fig 11 shows the updated dataset with the venue information added. There are **262 unique categories** across **1882 venues.**



```
     Frequency                      Categories
175         61                            Park
3           51             American Restaurant
111         45                             Gym
202         43                  Sandwich Place
15          43                             Bar
183         42                     Pizza Place
48          42                     Coffee Shop
46          35                  Clothing Store
85          33            Fast Food Restaurant
29          31                  Breakfast Spot
216         30  Southern / Soul Food Restaurant
122         28                           Hotel
152         28              Mexican Restaurant
112         26             Gym / Fitness Center
197         24                      Restaurant
6           24                     Art Gallery
101         23                     Gas Station
129         22              Italian Restaurant
109         22                   Grocery Store
204         22               Seafood Restaurant
```

Fig 12. Highest occurring venue categories across the 159 neighborhoods

In *Fig 12*, the highest occurring categories in the dataset are shown. **Parks, American restaurants, gyms, sandwich places, bars, pizza places and coffee shops** are the most common types of venues with all of them having a **frequency of more than 40**. These are in fact some of the most popular and regular destinations for people in a city.

Apart from the top 7 categories, some of the other important venue types that complement the location for a retail store are also in the table above, such as **clothing stores**, **hotels**, **gas stations** and **restaurants** (many).

This is good information as it indicates which categories to look for and which categories to potentially avoid when looking for a location to maximize traffic.

It is a trend among businesses to situate their stores close to competitors simply to attract crowds with similar needs. Therefore, one will often see multiple fast food restaurants close to one another and several gas stations in close proximity to each other. This makes other clothing/retail stores, boutique stores, vintage stores and malls important for determining the ideal location for a new retail store.

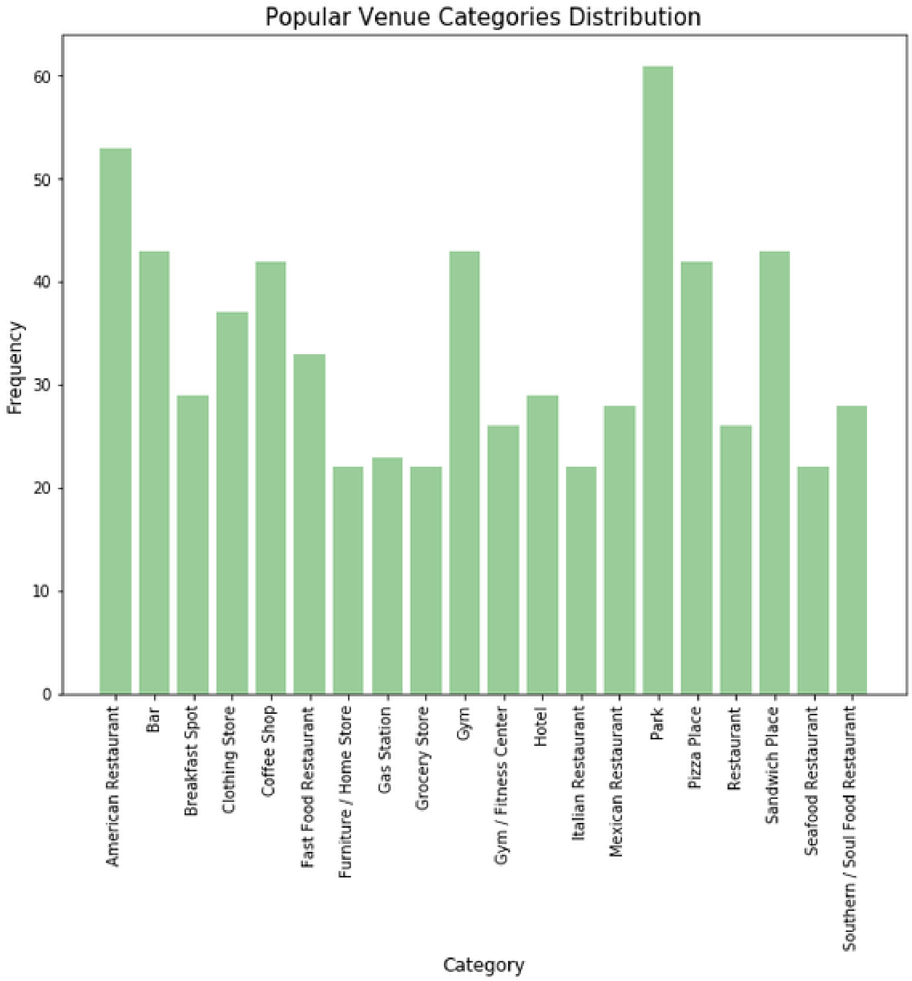


Fig 13. Venue category histogram showing the highest occurring categories

The top categories are visualized in *Fig 13* using a histogram which provides a better idea of the most popular venue types.

Next, the neighborhoods are grouped by venues and the total number of venues contained in each neighborhood is displayed in *Fig 14*. Note that these are just the top 30 neighborhoods.

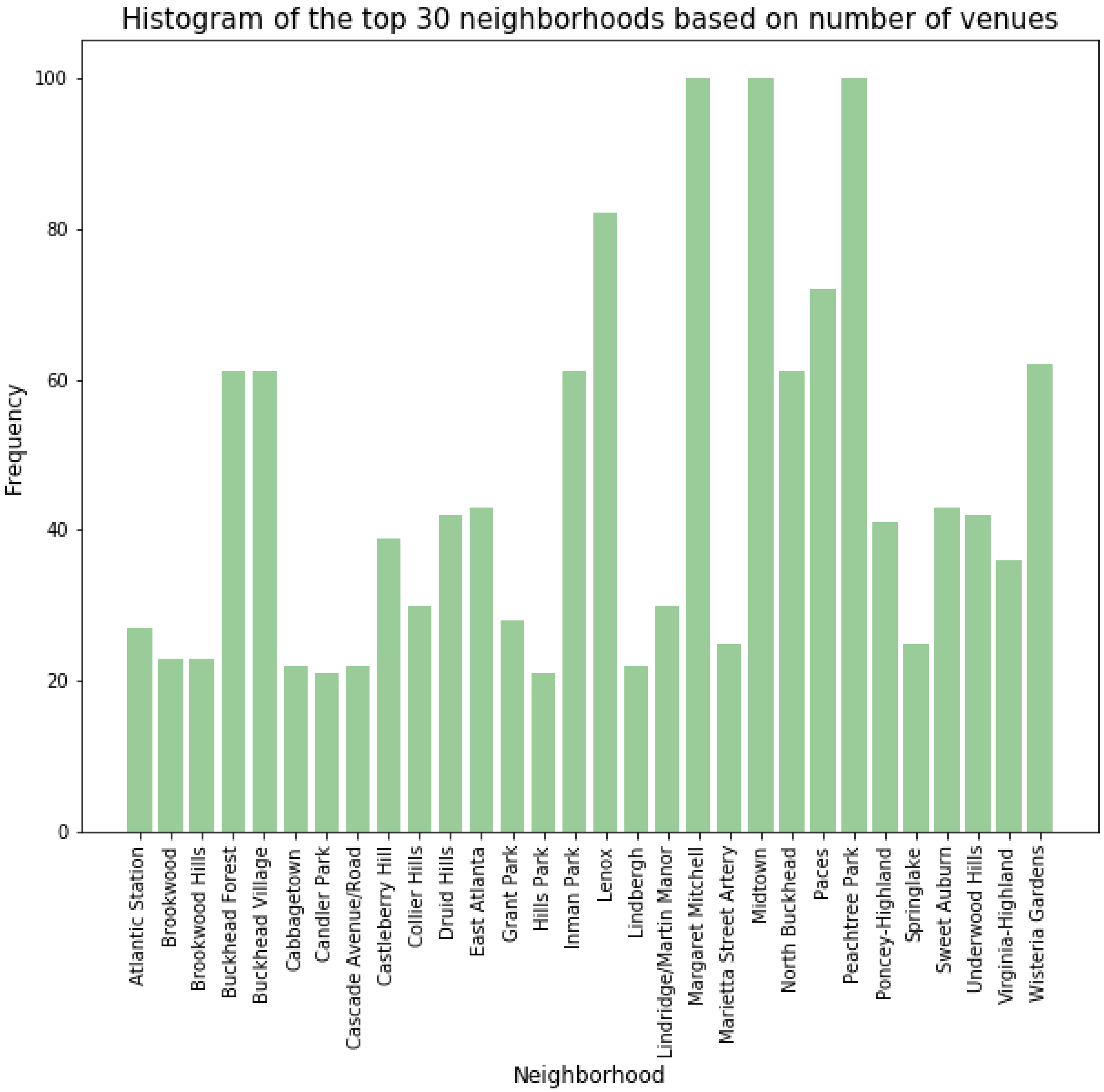


Fig 14. Histogram showing the distribution of venues across the top 30 neighborhoods based on total number of venues

Out of the remaining neighborhoods, only the top 30 by total number of venues are selected for further analysis as they consist of at least 20 venues. This is done under the assumption that neighborhoods with more venues are likely to have a **higher flow of traffic** on average.

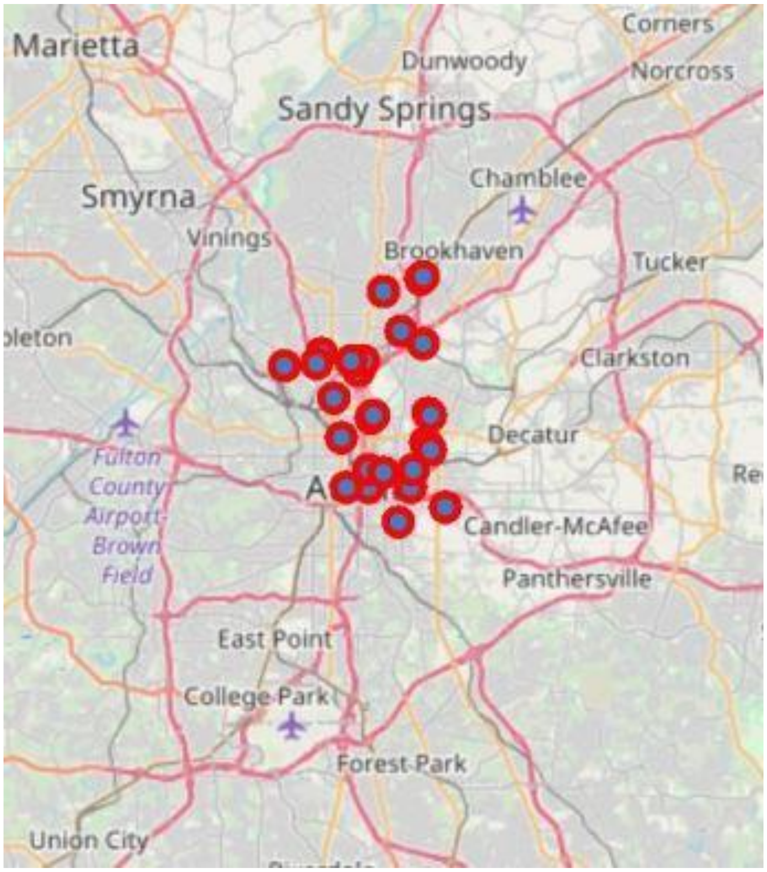


Fig 15. Top 30 neighborhoods in Atlanta based on number of venues

Fig 15 shows the remaining 30 neighborhoods on the Atlanta map and all of the points are concentrated at the **center of the I-285 circle**. This makes sense as the more dense neighborhoods are expected to be located at the heart of the city.

2.4 Feature engineering

From the **1365 different venues** in the reduced dataset, the neighborhoods need to be clustered by preferably distinguishing between the high traffic and low traffic venues. The venue categories are used to achieve this.

Some of the **high traffic** venues that we are interested in are:

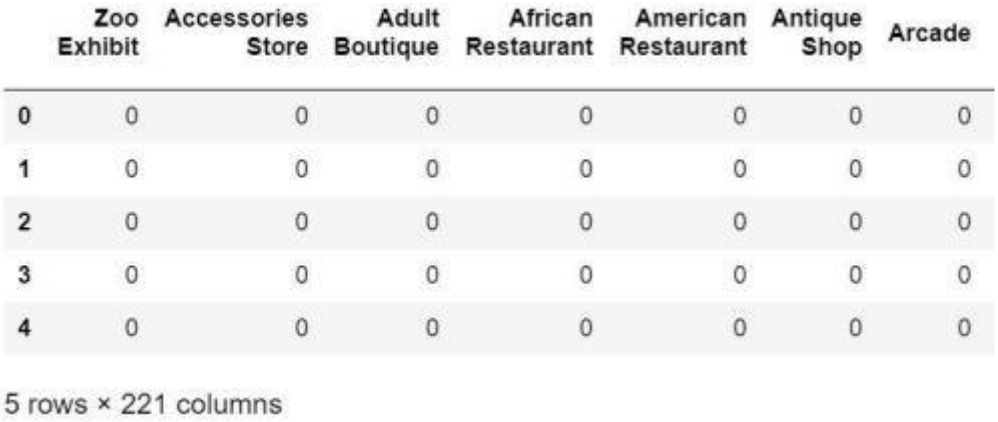
- *Restaurants*
- *Coffee shops*
- *Bars*
- *Parks*
- *Malls*
- *Hotels*

Secondly we also want to distinguish areas with venues that **complement retail stores** such as:

- *Gas stations*
- *Parking lots*
- *Other clothing stores*

First, the different venue categories are analyzed by transforming them into a *sparse matrix* format. A **sparse matrix** is a matrix where each unique category (in this case) is set as a separate *feature*. If the category exists for that instance (neighborhood) in the dataframe, the entry is a *1* otherwise it's a *0*.

For example, if there are **8 unique categories**, the sparse matrix would contain **8 columns (for each category)** and **rows equal to the number of observations (neighborhoods)** in the original dataset.



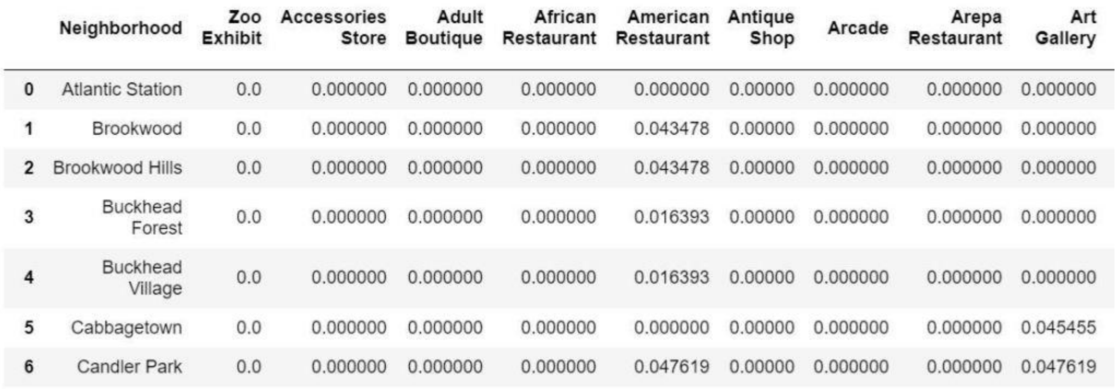
	Zoo Exhibit	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Antique Shop	Arcade
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0

5 rows × 221 columns

Fig 16. Sparse matrix from the venue categories

The sparse matrix is easily created in Python using the *get_dummies* function from the *Pandas* library. Fig 16 shows a snippet of the resulting dataframe which shows just 5 rows and 7 columns. The complete dataframe consists of **1365 rows (total number of venues)** and **221 columns (categories).**

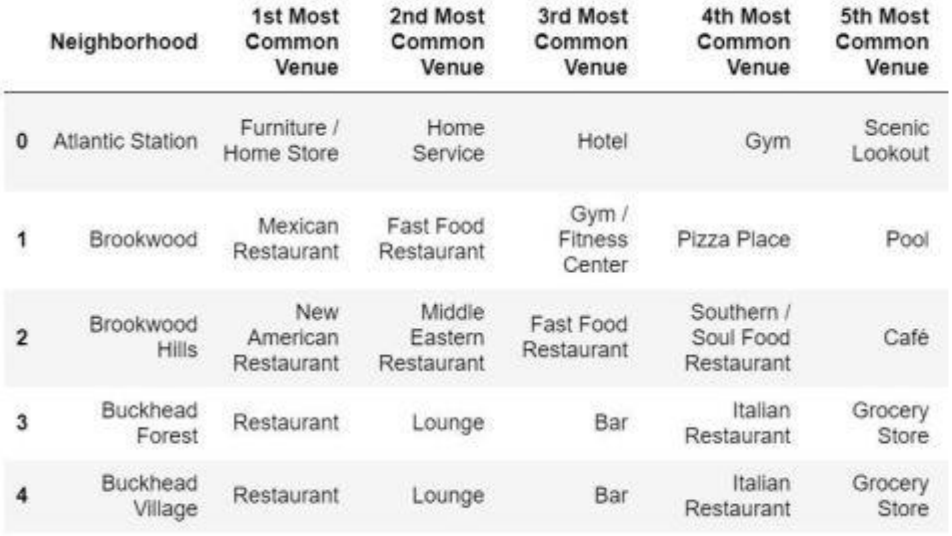
In *Fig 17*, the occurrence of each category is shown across all neighborhoods. This manipulation is important for the upcoming **clustering** exercise. The above dataframe is generated by grouping the data shown in *Fig 6* by neighborhood and applying the *mean* function to generate the **relative frequency** of each category for each neighborhood.



	Neighborhood	Zoo Exhibit	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Art Gallery
0	Atlantic Station	0.0	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000
1	Brookwood	0.0	0.000000	0.000000	0.000000	0.043478	0.00000	0.000000	0.000000	0.000000
2	Brookwood Hills	0.0	0.000000	0.000000	0.000000	0.043478	0.00000	0.000000	0.000000	0.000000
3	Buckhead Forest	0.0	0.000000	0.000000	0.000000	0.016393	0.00000	0.000000	0.000000	0.000000
4	Buckhead Village	0.0	0.000000	0.000000	0.000000	0.016393	0.00000	0.000000	0.000000	0.000000
5	Cabbagetown	0.0	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.045455
6	Candler Park	0.0	0.000000	0.000000	0.000000	0.047619	0.00000	0.000000	0.000000	0.047619

Fig 17. Relative frequency of each category for each neighborhood

To get an idea of the most common venues in each neighborhood, the highest category values from the dataset in *Fig 17* are picked out. **Higher the average value of that category in a neighborhood, more common the category.**



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Atlantic Station	Furniture / Home Store	Home Service	Hotel	Gym	Scenic Lookout
1	Brookwood	Mexican Restaurant	Fast Food Restaurant	Gym / Fitness Center	Pizza Place	Pool
2	Brookwood Hills	New American Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Southern / Soul Food Restaurant	Café
3	Buckhead Forest	Restaurant	Lounge	Bar	Italian Restaurant	Grocery Store
4	Buckhead Village	Restaurant	Lounge	Bar	Italian Restaurant	Grocery Store

Fig 18. Displaying the most common venue categories for each neighborhood

The table in *Fig 18* shows a snippet of the most commonly occurring venue categories across each neighborhood. The complete shape of the dataset is **30 by 10** which represents the **number of neighborhoods (rows)** and **most common venues (columns)**.

2.5 Clustering

K-means clustering was used to group similar neighborhoods together based on the venue categories in each neighborhood. The *unsupervised machine learning* technique was used on the dataframe from *Fig 17*.

The **number of clusters** chosen for the algorithm was **5** and it was implemented using the *KMeans* machine learning function from the *sklearn* library in Python. **Five labels** are generated by the algorithm **representing each cluster**. These labels are then added to the dataset in *Fig 18* for distinguishing similar neighborhoods.

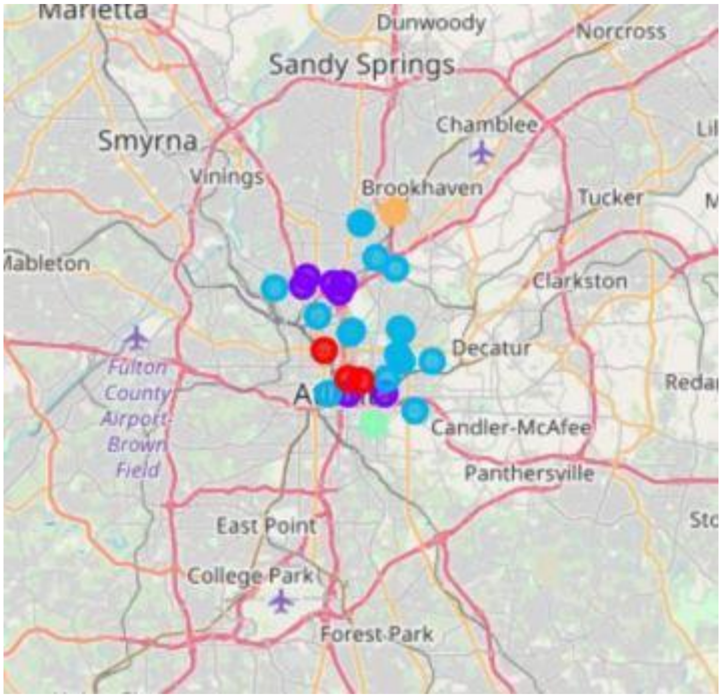
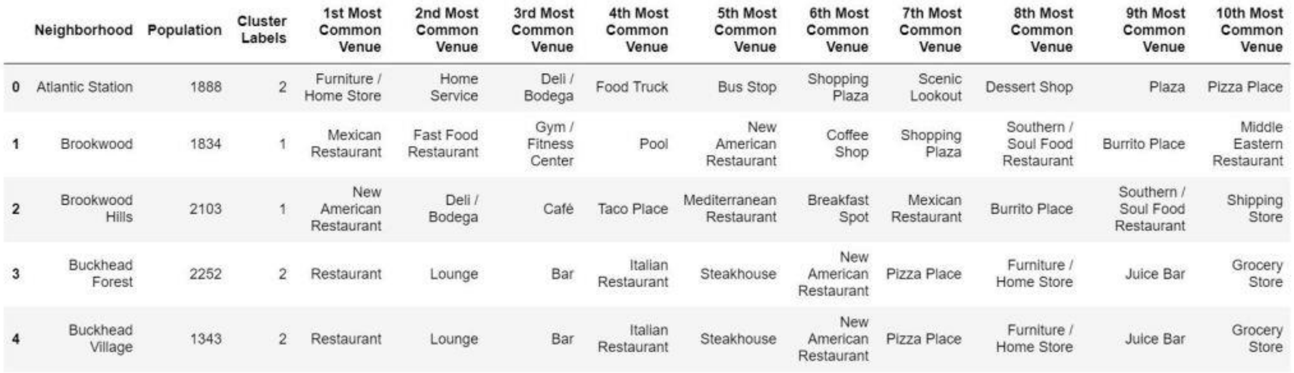


Fig 19. Map of Atlanta showing the 5 clusters of neighborhoods

From the map in *Fig 19*, there appears to be a somewhat distinguishable pattern between the clusters. Noticeably, the **purple**, **blue** and **red** clusters are the largest and are localized, to a great extent, around the same area. There appears to be **3 purple neighborhoods** that are away from the main cluster. Also for the blue neighborhoods, the data points are spread across a relatively larger area.

The **light green** and **orange** clusters appear to be relatively small and are located at the periphery of all the clusters.



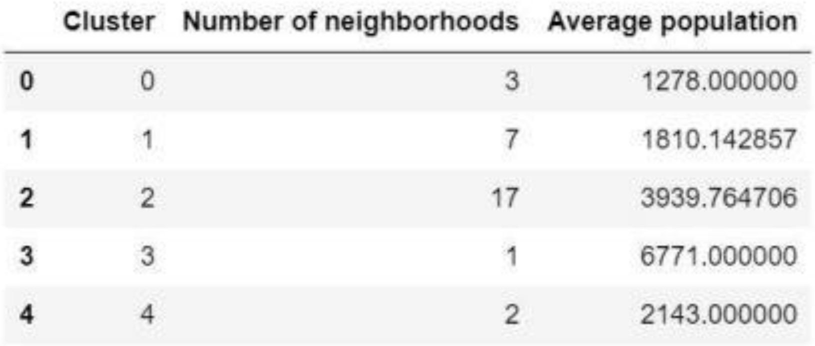
	Neighborhood	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Atlantic Station	1888	2	Furniture / Home Store	Home Service	Deli / Bodega	Food Truck	Bus Stop	Shopping Plaza	Scenic Lookout	Dessert Shop	Plaza	Pizza Place
1	Brookwood	1834	1	Mexican Restaurant	Fast Food Restaurant	Gym / Fitness Center	Pool	New American Restaurant	Coffee Shop	Shopping Plaza	Southern / Soul Food Restaurant	Burrito Place	Middle Eastern Restaurant
2	Brookwood Hills	2103	1	New American Restaurant	Deli / Bodega	Café	Taco Place	Mediterranean Restaurant	Breakfast Spot	Mexican Restaurant	Burrito Place	Southern / Soul Food Restaurant	Shipping Store
3	Buckhead Forest	2252	2	Restaurant	Lounge	Bar	Italian Restaurant	Steakhouse	New American Restaurant	Pizza Place	Furniture / Home Store	Juice Bar	Grocery Store
4	Buckhead Village	1343	2	Restaurant	Lounge	Bar	Italian Restaurant	Steakhouse	New American Restaurant	Pizza Place	Furniture / Home Store	Juice Bar	Grocery Store

Fig 20. Dataset showing the most common venue categories, population and cluster label for each neighborhood

Fig 20 shows the final dataset before results of the clustering algorithm are analyzed. This data shows the **top 10 most common venue categories**, **population** and **cluster label** for each neighborhood.

2.6 Analyzing the clusters

The table in *Fig 21* shows the summary statistics for the cluster data containing the **cluster label**, **number of neighborhoods**and **average population**.



	Cluster	Number of neighborhoods	Average population
0	0	3	1278.000000
1	1	7	1810.142857
2	2	17	3939.764706
3	3	1	6771.000000
4	4	2	2143.000000

Fig 21. Summary statistics for the neighborhood clusters displaying total number of venues and average population

From this table, it is clear that *cluster 2* contains **majority of the neighborhoods (17 out of 30)** and is also one of the **largest clusters by average population (3939.7647)**. Since *cluster 3* contains just 1 neighborhood, the **misleading** high population of 6771 can be ignored due to the lack of a large enough sample size.

In terms of population the clusters can be segmented accordingly:

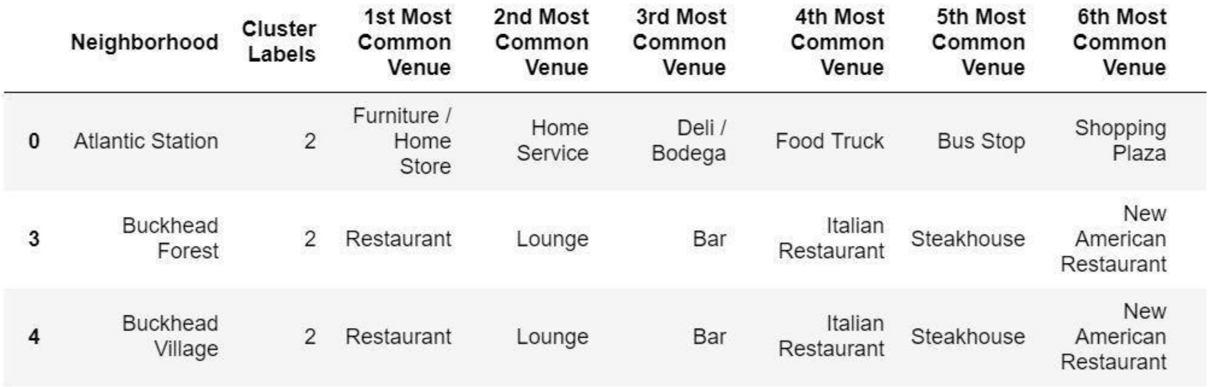
- Cluster 0: *Low population*
- Cluster 1: *Below average population*
- Cluster 2: *Large population*
- Cluster 3: *Extremely large population*
- Cluster 4: *Above average population*

Based on number of neighborhoods, the clusters can be segmented as follows:

- Clusters 0, 3 and 4: *Low number of neighborhoods*
- Cluster 1: *Average number of neighborhoods*
- Cluster 2: *Large number of neighborhoods*

2.7 Examining Cluster 2

Based on the analysis so far, *cluster 2* is the most favorable group to analyze further due to a large population and the highest number neighborhoods.



	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Atlantic Station	2	Furniture / Home Store	Home Service	Deli / Bodega	Food Truck	Bus Stop	Shopping Plaza
3	Buckhead Forest	2	Restaurant	Lounge	Bar	Italian Restaurant	Steakhouse	New American Restaurant
4	Buckhead Village	2	Restaurant	Lounge	Bar	Italian Restaurant	Steakhouse	New American Restaurant

Fig 22. Snippet of cluster 2 dataset

Fig 22 shows a snippet of the cluster 2 dataset. In *Fig 23* the **top 20 most commonly occurring categories** in the cluster are visualized. The most common categories include *pizza place*, *coffee shop* and *bar*.

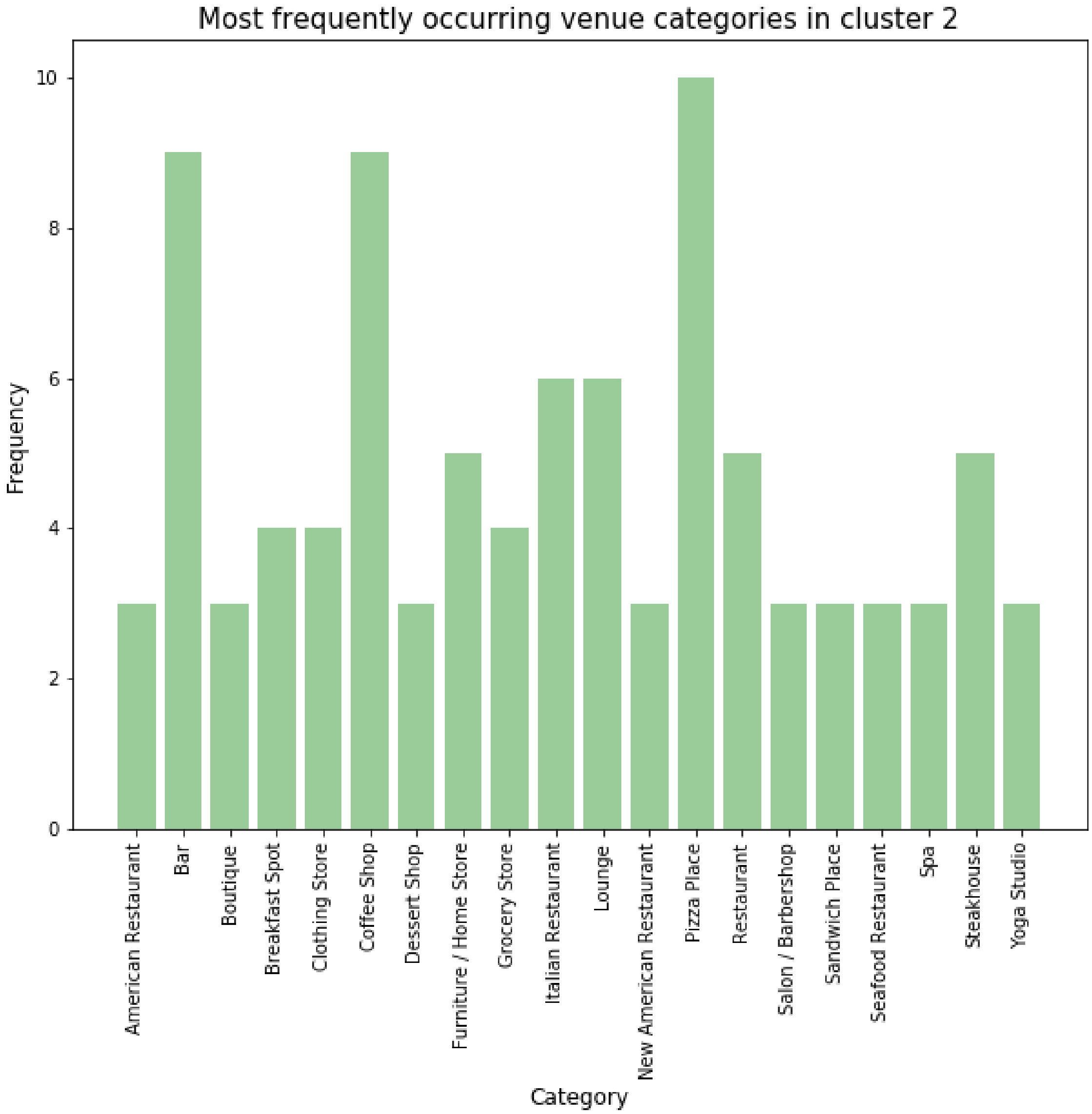
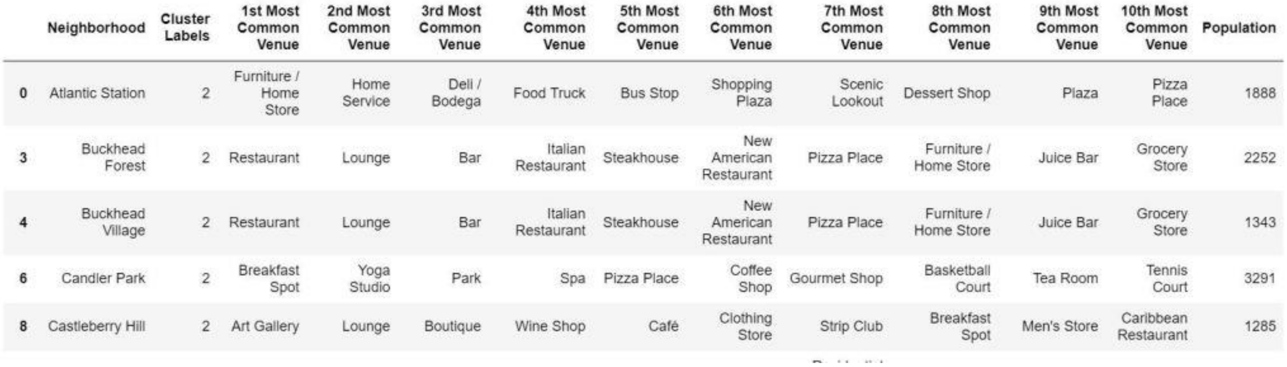


Fig 23. Histogram showing the top 20 venue categories in cluster 2

Cluster 2 is further reduced to just seven neighborhoods which were selected as the ones with **population above 4000**. Since the goal is to find areas of maximum traffic, looking at higher population neighborhoods will ensure greater flow of traffic in these areas.

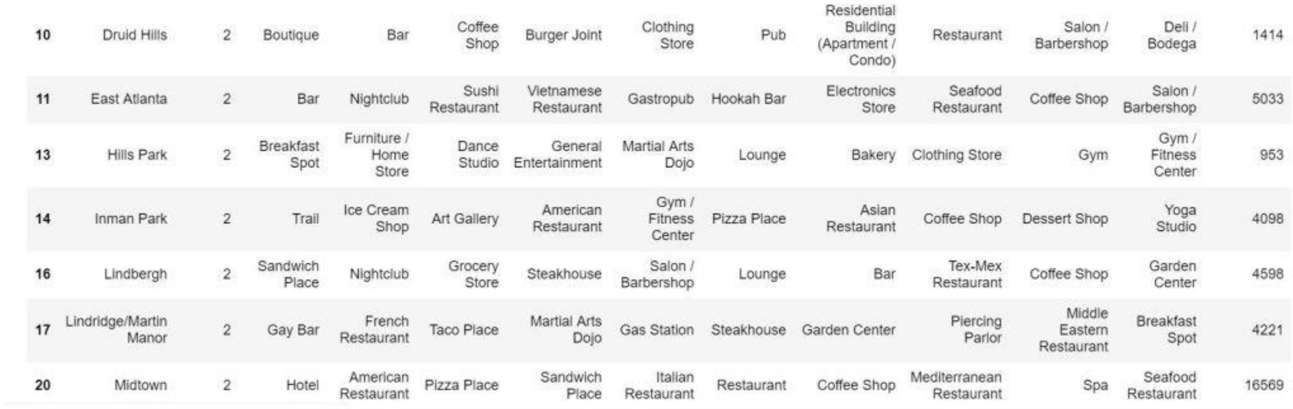
3. Results

Fig 24, *Fig 25* and *Fig 26* together show the entire cluster 2 dataset which contains the most common venue categories, population and cluster labels for each neighborhood.



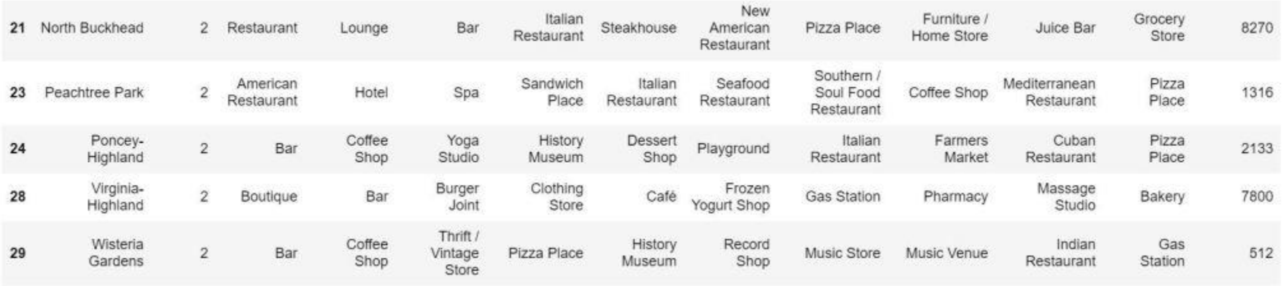
	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Population
0	Atlantic Station	2	Furniture / Home Store	Home Service	Deli / Bodega	Food Truck	Bus Stop	Shopping Plaza	Scenic Lookout	Dessert Shop	Plaza	Pizza Place	1888
3	Buckhead Forest	2	Restaurant	Lounge	Bar	Italian Restaurant	Steakhouse	New American Restaurant	Pizza Place	Furniture / Home Store	Juice Bar	Grocery Store	2252
4	Buckhead Village	2	Restaurant	Lounge	Bar	Italian Restaurant	Steakhouse	New American Restaurant	Pizza Place	Furniture / Home Store	Juice Bar	Grocery Store	1343
6	Candler Park	2	Breakfast Spot	Yoga Studio	Park	Spa	Pizza Place	Coffee Shop	Gourmet Shop	Basketball Court	Tea Room	Tennis Court	3291
8	Castleberry Hill	2	Art Gallery	Lounge	Boutique	Wine Shop	Café	Clothing Store	Strip Club	Breakfast Spot	Men's Store	Caribbean Restaurant	1285

Fig 24. First part of cluster 2 dataset



10	Druid Hills	2	Boutique	Bar	Coffee Shop	Burger Joint	Clothing Store	Pub	Residential Building (Apartment / Condo)	Restaurant	Salon / Barbershop	Deli / Bodega	1414
11	East Atlanta	2	Bar	Nightclub	Sushi Restaurant	Vietnamese Restaurant	Gastropub	Hookah Bar	Electronics Store	Seafood Restaurant	Coffee Shop	Salon / Barbershop	5033
13	Hills Park	2	Breakfast Spot	Furniture / Home Store	Dance Studio	General Entertainment	Martial Arts Dojo	Lounge	Bakery	Clothing Store	Gym	Gym / Fitness Center	953
14	Inman Park	2	Trail	Ice Cream Shop	Art Gallery	American Restaurant	Gym / Fitness Center	Pizza Place	Asian Restaurant	Coffee Shop	Dessert Shop	Yoga Studio	4098
16	Lindbergh	2	Sandwich Place	Nightclub	Grocery Store	Steakhouse	Salon / Barbershop	Lounge	Bar	Tex-Mex Restaurant	Coffee Shop	Garden Center	4598
17	Lindridge/Martin Manor	2	Gay Bar	French Restaurant	Taco Place	Martial Arts Dojo	Gas Station	Steakhouse	Garden Center	Piercing Parlor	Middle Eastern Restaurant	Breakfast Spot	4221
20	Midtown	2	Hotel	American Restaurant	Pizza Place	Sandwich Place	Italian Restaurant	Restaurant	Coffee Shop	Mediterranean Restaurant	Spa	Seafood Restaurant	16569

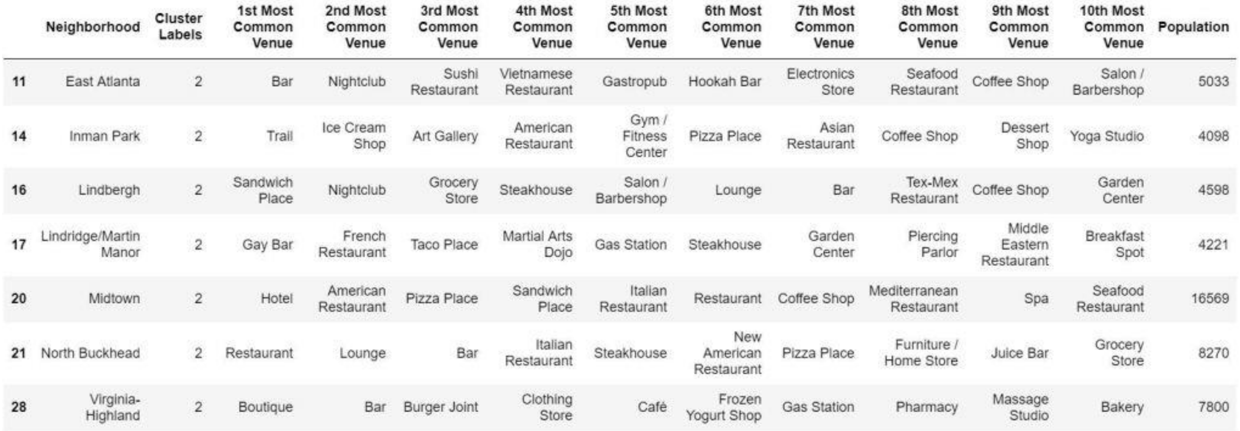
Fig 25. Second part of cluster 2 dataset



21	North Buckhead	2	Restaurant	Lounge	Bar	Italian Restaurant	Steakhouse	New American Restaurant	Pizza Place	Furniture / Home Store	Juice Bar	Grocery Store	8270
23	Peachtree Park	2	American Restaurant	Hotel	Spa	Sandwich Place	Italian Restaurant	Seafood Restaurant	Southern / Soul Food Restaurant	Coffee Shop	Mediterranean Restaurant	Pizza Place	1316
24	Poncey-Highland	2	Bar	Coffee Shop	Yoga Studio	History Museum	Dessert Shop	Playground	Italian Restaurant	Farmers Market	Cuban Restaurant	Pizza Place	2133
28	Virginia-Highland	2	Boutique	Bar	Burger Joint	Clothing Store	Café	Frozen Yogurt Shop	Gas Station	Pharmacy	Massage Studio	Bakery	7800
29	Wisteria Gardens	2	Bar	Coffee Shop	Thrift / Vintage Store	Pizza Place	History Museum	Record Shop	Music Store	Music Venue	Indian Restaurant	Gas Station	512

Fig 26. Third part of cluster 2 dataset

The categories in the above tables are analyzed to determine the ideal neighborhood in which to establish a new retail store. *Fig 27* shows the remaining neighborhoods from *cluster 2* all of which have a population greater than 4000.



	Neighborhood	Cluster Labels	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	Population
11	East Atlanta	2	Bar	Nightclub	Sushi Restaurant	Vietnamese Restaurant	Gastropub	Hookah Bar	Electronics Store	Seafood Restaurant	Coffee Shop	Salon / Barbershop	5033
14	Inman Park	2	Trail	Ice Cream Shop	Art Gallery	American Restaurant	Gym / Fitness Center	Pizza Place	Asian Restaurant	Coffee Shop	Dessert Shop	Yoga Studio	4098
16	Lindbergh	2	Sandwich Place	Nightclub	Grocery Store	Steakhouse	Salon / Barbershop	Lounge	Bar	Tex-Mex Restaurant	Coffee Shop	Garden Center	4598
17	Lindridge/Martin Manor	2	Gay Bar	French Restaurant	Taco Place	Martial Arts Dojo	Gas Station	Steakhouse	Garden Center	Piercing Parlor	Middle Eastern Restaurant	Breakfast Spot	4221
20	Midtown	2	Hotel	American Restaurant	Pizza Place	Sandwich Place	Italian Restaurant	Restaurant	Coffee Shop	Mediterranean Restaurant	Spa	Seafood Restaurant	16569
21	North Buckhead	2	Restaurant	Lounge	Bar	Italian Restaurant	Steakhouse	New American Restaurant	Pizza Place	Furniture / Home Store	Juice Bar	Grocery Store	8270
28	Virginia-Highland	2	Boutique	Bar	Burger Joint	Clothing Store	Café	Frozen Yogurt Shop	Gas Station	Pharmacy	Massage Studio	Bakery	7800

Fig 27. Final cluster showing neighborhoods in cluster 2 that have a population of over 4000

The zoomed-in version of the neighborhood clusters with the neighborhood of *Virginia-Highland* labeled is shown in *Fig 28*.

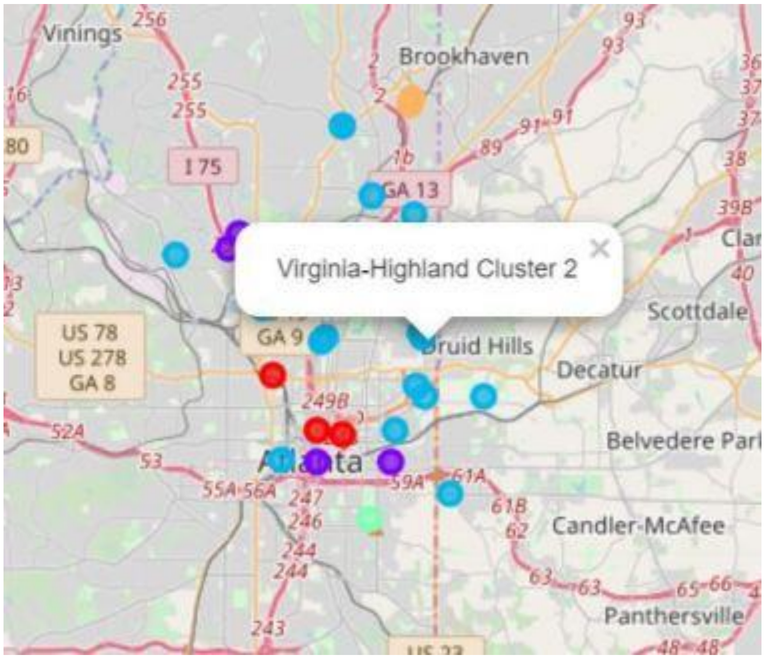


Fig 28. Zoomed in map of neighborhood clusters highlighting Virginia Highland

4. Discussion

In order to narrow down 161 neighborhoods in the city of Atlanta to just one is a difficult task. As seen in this report, several analytic techniques were used to achieve this.

First, the coordinates of each neighborhood were retrieved using the *geopy* library in Python. Then, neighborhoods that were outside the *I-285 highway circle* were filtered out since they are far from the center of the city. This was followed by exploring venues in each neighborhood and generating all the venue categories. Based on the number of venues, the neighborhoods were further reduced by selecting the top 30 neighborhoods which contained the most venues.

After finding the frequency of all categories across each neighborhood, the *KMeans* clustering algorithm was used to **segment the neighborhoods into 5 clusters.** Then, based on population and number of neighborhoods in each group, **cluster 2** was picked as the most important cluster for further analysis.

To summarize, here are the three main aspects that point towards *cluster 2* as being more favorable than the others:

- It is the largest cluster on the map and is mostly centralized to one area. It **can be separated into a somehwat distinguishable area**. Also, the cluster is located right in the heart of Atlanta as seen on the map in *Fig 19*. The neighborhoods are **concentrated at the center of the I-285 circle**.
- Looking at the cluster tables above, *cluster 2* does not only **contain majority of the neighborhoods** but also contains a **good mix of the important venues** that we are interested in, which could help maximize traffic to the new retail store.
- From the population distribution above, *cluster 2* also has a **relatively high average population**. This is considering cluster 3 just has one neighborhood with venues that are not as important, hence we ignore it from the population consideration.

The most common venues for **cluster 2** are **pizza places, coffee shops and bars**. That is followed by **lounges, Italian restaurants, general restaurants, furniture/home stores, steakhouses, clothing stores, breakfast spots and barber shops**. Some of the less common important venues include **clothing stores, music venues, gas stations, music stores, movie theaters, bus stop, boutiques and vintage stores.**

Cluster 2 appears to be promising as it not only contains more venues in general but also contains more “important” venues which we are looking for. Also venues such as gyms, grocery stores and massage studios indicate that there will likely be more regular customers returning to these places.

Most of the above neighborhoods from the final cluster dataset in *Fig 27* contain **restaurants**, **bars** and **coffee shops**, three venues that attract large crowds. However, *Midtown* and *Inman Park* do not appear to have bars as one of the most common venues. So we are left with the remaining 5.

Out of all the remaining neighborhoods, *Virginia-Highland* contains the best mix of venue categories.

5. Conclusion

From the table in *Fig 27*, **Virginia Highlands** appears to be the most promising neighborhood for our new retail store. Here are the key reasons why:

- Virginia Highlands has a *population of 7800* which is substantially higher than the upper-quartile of the population distribution across the neighborhoods. A higher population ensures **higher rate of traffic** in the neighborhood. The presence of restaurants, furniture stores and shop & service venues suggests that this area could have a good amount of residents who are likely to flock to these venues.
- The neighborhood contains a lot of the important venues that we had mentioned earlier – *clothing stores, boutique stores, bars,gas station,cafes* and *restaurants*. This also suggests that tourists could be attracted to the area who are often likely to visit retail stores.
- It is in close proximity to another neighborhood in the same cluster – **Druid Hill**. Although this neighborhood consists of similar venues, this provides customers/tourists with more options. **Poncey-Highland** and **Wisteria Gardens** are also relatively close to Virginia Highlands and are in close proximity to one another as well. They consist of museums, music stores/venues and vintage stores which are not there in either Virginia-Highlands or Druid Hills.

Note that the factors that were considered in this analysis are just a handful of considerations when opening a retail store in a particular city. Factors that were addressed are number of venues in a neighborhood, population of neighborhoods, centrality of neighborhoods, venue categories in a neighborhood and proximity from I-285 circle.

Other important factors could include price of buildings in the area, crime rate in the neighborhood, proximity to parking lots etc. However, using the factors in this report certainly provides a good starting point and solid reference.

6. References

[1] https://en.wikipedia.org/wiki/Table_of_Atlanta_neighborhoods_by_population

[2] Google Search Engine

[3] Foursquare API