**The Battle of Neighborhoods**

Conor Zimmerman

***1. Introduction / Business Problem***

The problem this report aims to solve consists of the following question, "Which neighborhoods in and around Pittsburgh could be considered suitable locations to host a new grocery store." To solve this problem, we will focus on factors such as the variety of venues within Pittsburgh's neighborhoods and their locations throughout the area. We will then group the neighborhoods based on these factors and make observations. The target audience of this report includes the stakeholders of a national grocery chain.

***2. Data***

For this project, we will be using Allegheny County Zip Codes data in conjunction with FourSquare location data. The Allegheny Zip Code dataset demarcates the zip code boundaries that lie within Allegheny County and provides zip codes, neighborhoods, and geospatial coordinates. The FourSquare Places API will be used to gather venue locations found within a 500m radius of a given neighborhood. Data will be organized into data frames and projected on to an interactive map for visualization.

The Allegheny Zip Codes dataset is accessed via Pennsylvania Spatial Data Access (PASDA), an official public access open geospatial data portal.

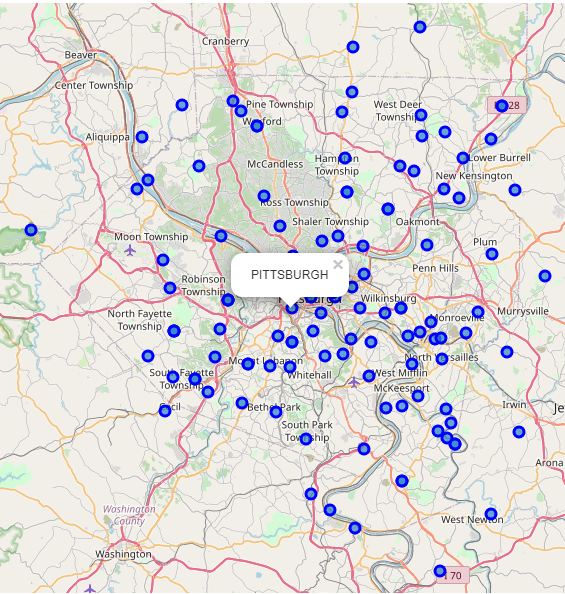
***2.1 Data Cleansing***

Upon loading in our data, we can see the variety of fields that are initially included (F2.1). Selecting the features that we want to work, we create our ‘neighborhoods’ data frame (F2.2). Next, we remove any rows where geospatial data is empty or unavailable to avoid any issues with mapping our neighborhoods. After removing these rows, we are left with a data frame containing 120 zip codes and 73 unique neighborhoods.

***3. Methodology***

For our analysis, we primarily utilize the *pandas, geopy, folium,* and *sklearn* libraries in Python to organize and map our data. Additionally, we install *numpy,* *requests*, *json*, and *matplotlib*.

***3.1 Mapping***

To begin, we utilize *geopy* and *folium* to find the coordinates of Pittsburgh, PA and map our neighborhoods (F3.1). We can now see how the neighborhoods in our data frame are spread throughout Pittsburgh and the surrounding metropolitan area.

F3.1

***3.2 FourSquare API***

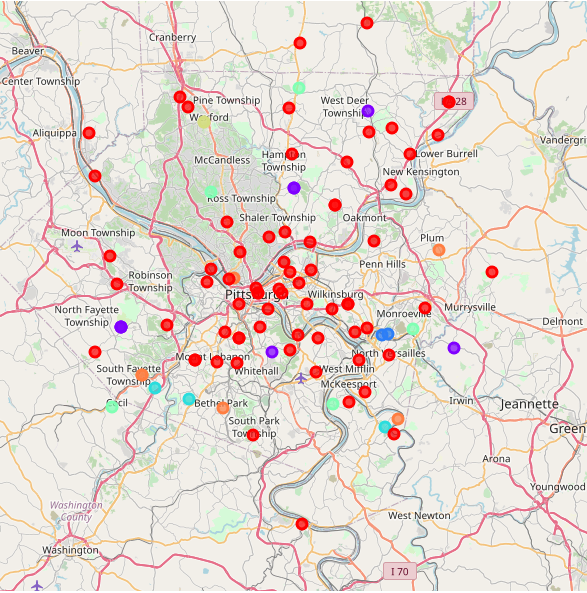
Selecting our first neighborhood, we can now make a request using the FourSquare API to get a list of venues within a 500m radius. Passing our FourSquare credentials and placing the results into a *pandas* data frame, we are able to see that 39 venues were returned, including the venue category and spatial coordinates (F3.2). Next, we make a similar request, this time for all neighborhoods. Placing the results of this call into a *pandas* data frame, and joining it with our ‘neighborhoods’ data frame, we can now see each venue that falls within a 500m radius of each neighborhood, along with the venue category and venue coordinates. (F3.3).

***3.3 Venue Frequency***

With our neighborhoods and venues data now organized together, we can find how often each venue category occurs within each neighborhood (i.e., venue frequency). Using venue frequency, we can find the most common occurring venue categories for each neighborhood. Placing this data into a new data frame, we can see the top 10 most common venues for each neighborhood (F3.4).

***3.4 Clustering***

Next, we can begin grouping our neighborhoods into clusters based on their top 10 most common venues. To do this, we will use k-means clustering to sort our neighborhoods into 7 distinct clusters. By studying the top 10 most common venues, the k-means clustering method will assign each neighborhood a cluster label between 0 and 6.

***3.5 Mapping Clusters***

With our neighborhoods now assigned to their clusters, we can construct a new map to visualize how the clusters fall throughout Pittsburgh and the metropolitan area (f3.5).

Here, we can see how each cluster differs in size and location.

F3.5

Legend: Cluster 0

Cluster 1

Cluster 2

Cluster 3

Cluster 4

Cluster 5

Cluster 6

***4. Results***

Finally, we can now analyze our neighborhood clusters to determine which neighborhoods could utilize a new grocery store location.

Immediately, we can see that Cluster 0 is our largest and most broad ranging cluster, containing 70 of our 96 clustered neighborhoods. Covering many of the more confined areas of the map with many neighborhoods near one another, including all downtown Pittsburgh, this cluster was likely formed due to the assigned 500m radius. As the distance between neighborhoods becomes smaller, the search radius will begin to overlap, resulting in several areas sharing the same venue results. Because of this added complexity, we will ignore Cluster 0 for now and focus our analyses on the remaining clusters.

Next, Cluster 1 is our second largest cluster, containing 8 of our 96 clustered neighborhoods. Appearing in distinct locations around the Pittsburgh area, Cluster 1’s neighborhoods share common venues such as Flower Shops, Farms, and Farmer’s Markets, with the most common venue type being Construction and Landscaping venues. Additional common venues in these neighborhoods include Electronics Stores, Restaurants, and Fields.

Thirdly, Cluster 2 is one of our smallest clusters, containing only 2 neighborhoods. Looking at the map, we can see that these neighborhoods fall right next to each other. As a result, the venues found in this cluster are likely the same, just shared between the 2 neighborhoods. Looking at the map further, we can discover that these neighborhoods are located near the Good Shephard Catholic Cemetery, the Restland Memorial Park, and the Monroeville Landfill.

Moving on, Cluster 3 is located on the southern edge of the Pittsburgh area. Looking further, we can discover that that the neighborhoods of this cluster are each located near Country Clubs: the St. Clair Country Club, and the Youghiogheny Country Club. Thus, the cluster’s most common venue of Golf Course is explained.

Cluster 4 contains 5 neighborhoods which can be found in distinct locations throughout the mapped area. These neighborhoods share common venues such as Restaurants, Yoga Studios, Farms and Farmer’s Markets.

Cluster 5 is tied for our smallest cluster, also containing just 2 neighborhoods. Looking at the map, we can see that these neighborhoods are also located beside each other, practically overlapping. Like Cluster 2, the common venues between these neighborhoods are identical, including a Recreation Center as the most common.

Lastly, Cluster 6 contains 6 neighborhoods which can be found in distinct locations on the southern and eastern edges of the Pittsburgh area. Common venues within these neighborhoods include Baseball Fields, Home Services, Parks, Event Spaces, and Restaurants.

***5. Discussion.***

With our results now gathered and analyzed, let us discuss the outcome and determine which of these clusters may include suitable neighborhoods.

To begin, we must first address the complexities of our study. The dataset used for our analysis, sourced from Pennsylvania Spatial Data Access (PASDA), was a subset of a larger dataset designed for mapping the Zip Code boundaries of Pittsburgh and the surrounding area. When searching for relevant data that could be used to answer our question, "Which neighborhoods would be suitable to host a new grocery store location?" this dataset was chosen due to it conveniently including both zip codes, which could be used for identification, and geospatial coordinates, used for mapping. However, the ‘NAME’ field of the data is troublesome, due to 'PITTSBURGH' frequently being listed as the neighborhood, as opposed to the proper Township or Borough. Fortunately, this problem did not directly affect the results of our study and could be solved in future analyses by acquiring additional data on the Townships and Boroughs of the area.

Next, let us re-acknowledge Cluster 0. Containing more than 70% of our clustered neighborhoods, Cluster 0 shows little trends amongst most common venues upon initial review. Due to the proximity of neighborhoods found within downtown Pittsburgh and nearby areas, the size and variety of this cluster makes analysis difficult. Solutions to this issue may include restricting our analysis to include only these downtown neighborhoods, altering our search radius to a more confined area, adding additional clusters when conducting k-means, or attempting a different method of clustering altogether.

Despite these complexities, the clusters provided by our study still provide interesting results. Particularly, Clusters 1, 4, and 6 each provide insights into several neighborhoods in the Pittsburgh metropolitan area. Beginning with Cluster 1, we see that these neighborhoods share Construction and Landscaping venues as their most common venue, suggesting ongoing industry and development within these areas. Additionally, these neighborhoods host Fields, Farms, and Farmer's Markets, which suggest more suburban/rural environments. With the implication of ongoing development going on in these areas, a new grocery store location could help centralize the services provided by the existing Farms and Farmer’s Markets.

Moving on to Cluster 4, we see that these neighborhoods are located on the outskirts of the mapped area. Sharing a variety of Restaurants, Shops, and Stores amongst their most common venues, this cluster implies a suburban area with a busy economic base. In contrast to Cluster 1, grocery-style locations such as farms and farmer's markets are less common in this cluster, suggesting the opportunity to provide a new service to the area.

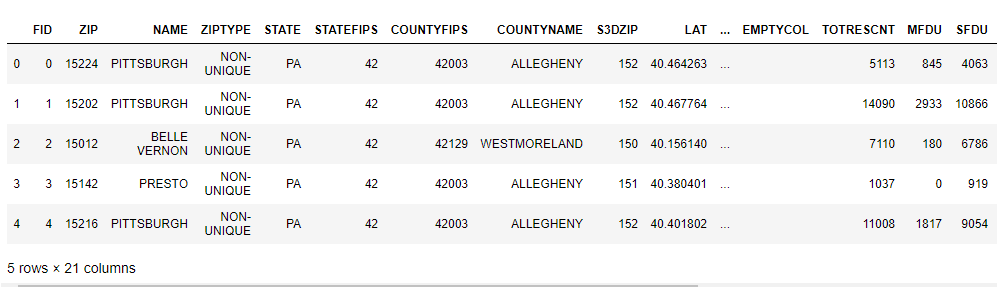
Finally, Cluster 6 shows an interesting most common venue of Baseball Fields. Consisting of suburban areas such as South Fayette, Bethel Park, and Plum, the neighborhoods of this cluster include popular school districts of the Pittsburgh area, suggesting a youthful and family-based population. Within an environment such as this, a new grocery store location could see high levels of activity.

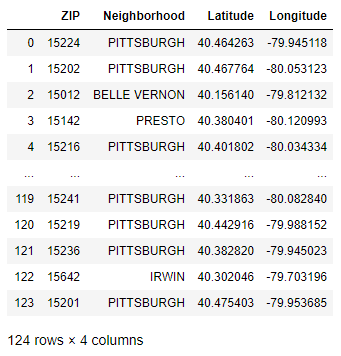
***6. Conclusion***

In conclusion, we have successfully narrowed down our search for which neighborhoods could be suitable to host a new grocery store location. By focusing efforts on Clusters 1, 4, and 6, we have found 19 neighborhoods that could likely support and/or utilize a new grocery store, reducing our initial dataset by 80% (F6.1).

Furthermore, a more complex study into the neighborhoods of Cluster 0 and the downtown Pittsburgh area allows for additional insight. By continuing this study with demographic and economic data on Pittsburgh's neighborhoods, along with data on the locations of pre-existing grocery store locations, we can likely find a more concrete answer to our question, discovering even more specific locations as well as likelihood for success.

***A. Figures***

(F2.1) pittsburgh\_nbh.head()

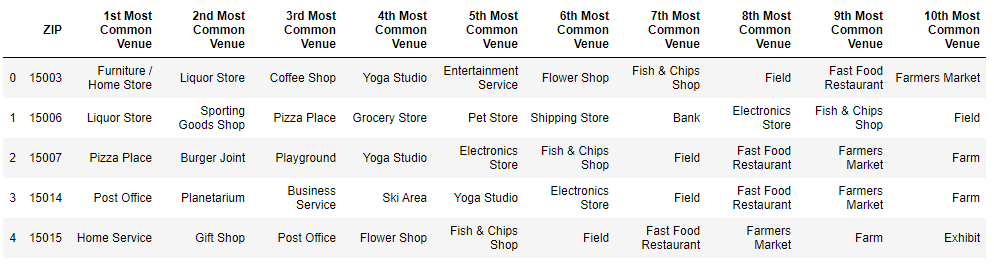
(F2.2) neighborhoods

(F3.2) nearby\_venues.head()

(F3.3) pittsburgh\_venues



(F3.4) neighborhoods\_venues\_sorted.head()



(F6.1)

|  |  |  |
| --- | --- | --- |
| ZIP | Neighborhood | Cluster |
| 15276 | PITTSBURGH | 1 |
| 15126 | IMPERIAL | 1 |
| 15075 | RURAL RIDGE | 1 |
| 15116 | GLENSHAW | 1 |
| 15231 | PITTSBURGH | 1 |
| 15071 | OAKDALE | 1 |
| 15085 | TRAFFORD | 1 |
| 15227 | PITTSBURGH | 1 |
| 15140 | PITCAIRN | 4 |
| 15045 | GLASSPORT | 4 |
| 15237 | PITTSBURGH | 4 |
| 15007 | BAKERSTOWN | 4 |
| 15321 | CECIL | 4 |
| 15057 | MC DONALD | 6 |
| 15239 | PITTSBURGH | 6 |
| 15064 | MORGAN | 6 |
| 15290 | PITTSBURGH | 6 |
| 15047 | GREENOCK | 6 |
| 15102 | BETHEL PARK | 6 |