Recommendations of new restaurant locations – Neighborhood comparison between Pittsburgh and Cleveland

Mei-Yu Wang

May, 29, 2020

1. Introduction:

1.1 Background:

These two cities: **Pittsburgh, PA** and **Cleveland, OH**, are known rivalry cities due to their football teams. They are only 2 hours driving distance away from each other and are similar in population (Pittsburgh: 305,012; Cleveland: 388,812). However, their restaurant landscapes are different, even for those fast-food restaurant chains.

1.2 Problem:

In this study, I am interested in categorizing neighborhoods in Pittsburgh and Cleveland using census data and venue data from Foursquare.com to make recommendations to fast-food restaurant chain stores when they want to open new locations in another city. For example. I am using four fast-food restaurant chains: **Shake Shack, BIBIBOP Asian Grill, Potbelly Sandwich Shop,** and **Krispy Krunchy Chicken**, for which they only exist in Cleveland and other city, as examples to generated recommended neighborhood in Pittsburgh to open new stores. This methodology can also be applied to any other cities or making recommendations to other types of business.

2. Data acquisition and cleaning

The dataset consists of three parts: **census data, geometric data** (neighborhood json files), and **venue data**. In the following I will explain the data acquiring and data cleaning processes for these three types of data:

2.1 Data acquisition and cleaning:

1. **.json files:**

   Zillow provides neighborhood json files for every cities in the U.S. They are available at "opendatasoft" (https://data.opendatasoft.com/). They provide the boundary of each neighborhood and also the geometric center of the neighborhood. Please note that the neighborhood number and names may not consistent with the census data described below. I utilized information from Wikipedia and other online information to match the neighborhood in each data. So the final number of neighborhoods and their names in my data may not be the same as those shown in the original Zillow json files.

2. **Census data:**

U.S. census data can be acquired from the United States Census Bureau website : www.census.gov. However, we are interested in census data presented by city neighborhoods, which are people usually refer to when they search for certain venues. Here we search online to find the most recently available Pittsburgh & Cleveland neighborhood census data. Since they may not be in csv form and may not be all up-to-date (we choose year 2018 to be the data collection time), we also explain how we apply data mining and cleaning procedures.

1) Pittsburgh data:

   The main content of Pittsburgh neighborhood census data at year 2018 is from University of Pittsburgh Center for Social & Urban Research website, where they provide census data from 2018-2014:

   https://ucsur.pitt.edu/files/census/ACS_Pgh_Profile_of_Change_2009-2013_v_2014-2018_Tables.pdf

   The data is stored in pdf form. In order to scratch the table from the pdf, we first convert the pdf file into excel using Adobe Acrobat pro and then read in the excel file using Pandas in Python. The median household income information (2016) comes from the website:

   http://www.city-data.com/nbmaps/neigh-Pittsburgh-Pennsylvania.html

   I used the read_html function from Pandas to read in the income table, then a cumulative inflation rate of 4.62% from year 2016 to 2018 to correct income values.

2) Cleveland data:

   The main content of Cleveland neighborhood census data at year 2014 comes from the Center for Community Solutions website:

   https://www.communitysolutions.com/resources/community-fact-sheets/cleveland-neighborhoods-and-wards/

   The data is again stored in pdf form. I first convert the pdf file into excel using Adobe Acrobat pro and then read in the excel file using Pandas in Python.

   To extrapolate the data to year 2018, I used the total population change in Cleveland from 2014 to 2018 acquired from the “World Population Review”:

   https://worldpopulationreview.com/us-cities/cleveland-population/

to correct the total population in each neighborhood. Here I assumed that the population ratio in each neighborhood doesn't change from year 2014 to 2018, so are the percentage values in different race, age, and education groups. The median household income values are corrected using a cumulative inflation rate of 6.07% from year 2014 to 2018.

The final data used in my analysis consist of **88** neighborhoods in Pittsburgh and **36** neighborhoods in Cleveland, for which I had went through the process of matching the information in the Zillow json files and those census data. In some cases there are ""missing data," and they are replaced by the average value of each columns in each city. In total, **124** neighborhood areas are included in this analysis.

In figure 1. I show the distribution of median household income in each neighborhood in Pittsburgh (right panel) and Cleveland (left panel). Please note that both "total population" and "income" columns have been normalized using the "min-max normalization" across these two cities. It is interesting to note that Pittsburgh has much higher average income than Cleveland.

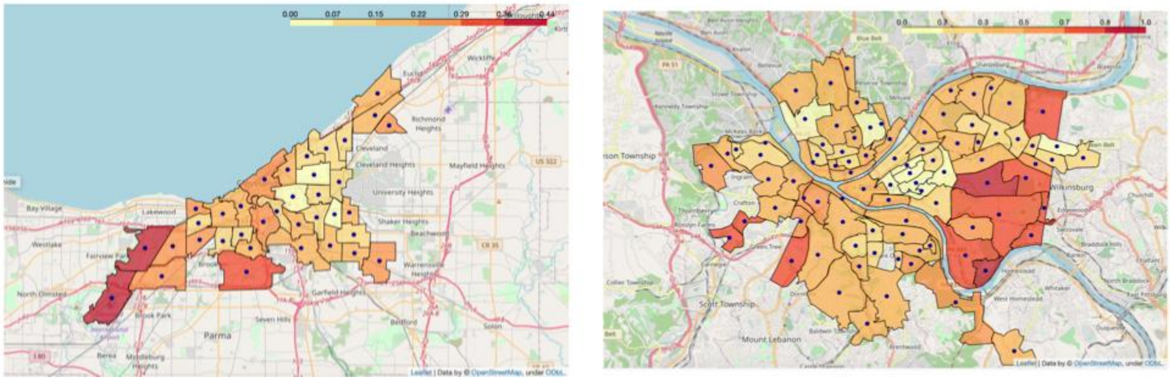


Fig. 1: distribution of median household income in each neighborhood in Pittsburgh (right panel) and Cleveland (left panel).

3. Venue data:

The venue data consists of two parts:

- Venues in each neighborhood.
- Information of the target restaurant chains in this study: Shake Shack, BIBIBOP Asian Grill, Potbelly Sandwich Shop, and Krispy Krunchy Chicken

The methods and resources to acquire both data are similar: using Foursquare data (foursquare.com) to link the geometric information to each venue data. However, the detailed procedure to acquire those data are slightly different. Here I elaborate on the details of those processes:

1) Venues in each neighborhood:

Since I had the boundaries of each neighborhood provided by the json files, I searched for all the venues lying within each neighborhood. Please note that although the foursquare sandbox account usage limits the max venue number of each query to be 100, by scanning small overlapping regions within a neighborhood and removing duplicated data, I can make sure to acquire a complete list of all venues within a neighborhood. To make sure a certain venue is located within a neighborhood, I used the "shapely" modules (pypi.org/project/Shapely/) to do so. In figure 2. I show the spatial distribution of all the venues in red points queried from Foursqure.com for Pittsburgh.

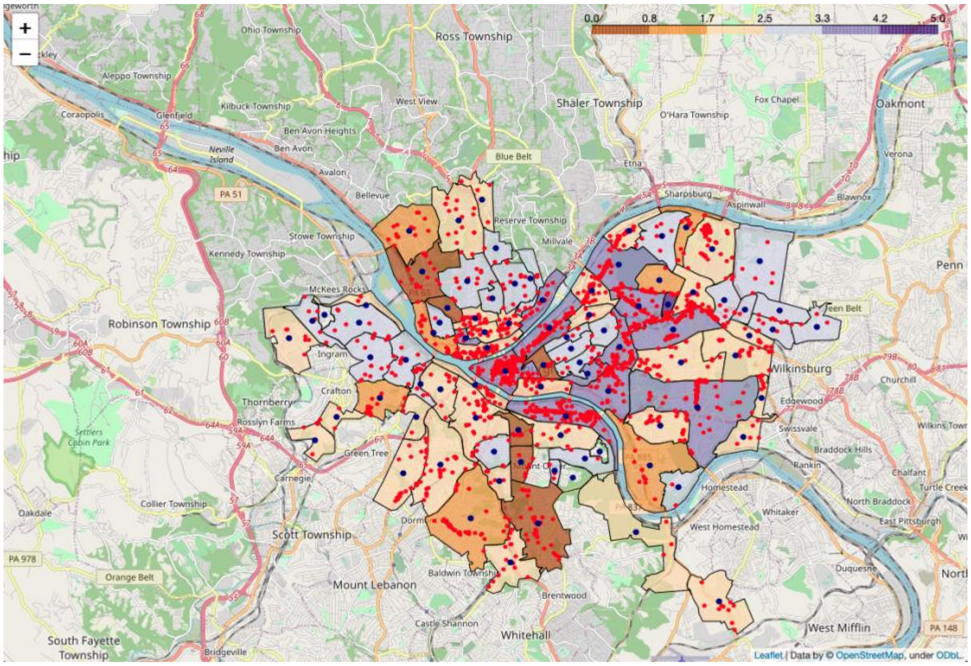
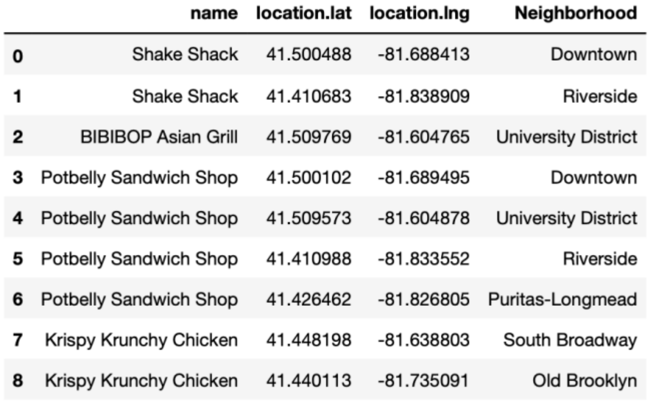


Fig. 2: Spatial distribution of all the venues in red points queried from Foursqure.com for Pittsburgh.

2) Information of the target restaurant chains:

To acquire the location of those restaurants, I searched the restaurant names and find out which Cleveland neighborhood they are in based on their locations. Again, I used the "shapely" modules to do so. In figure 3, I show the data of those restaurants (name, coordinate, neighborhood).



	name	location.lat	location.lng	Neighborhood
0	Shake Shack	41.500488	-81.688413	Downtown
1	Shake Shack	41.410683	-81.838909	Riverside
2	BIBIBOP Asian Grill	41.509769	-81.604765	University District
3	Potbelly Sandwich Shop	41.500102	-81.689495	Downtown
4	Potbelly Sandwich Shop	41.509573	-81.604878	University District
5	Potbelly Sandwich Shop	41.410988	-81.833552	Riverside
6	Potbelly Sandwich Shop	41.426462	-81.826805	Puritas-Longmead
7	Krispy Krunchy Chicken	41.448198	-81.638803	South Broadway
8	Krispy Krunchy Chicken	41.440113	-81.735091	Old Brooklyn

Fig. 3: data of target restaurants (name, coordinate, neighborhood) in Cleveland in this study.

2.2 Feature selection:

For the census data, the following features are selected for each neighborhood at year 2018:

- **Total Population**
- **Income** (median household income)
- **White alone** (in percentage)
- **Black alone** (in percentage)
- **Asian alone** (in percentage)
- **Other races** (in percentage)
- **Under age 18** (in percentage)
- **Age 18-64** (in percentage)
- **Age 65 and over** (in percentage)
- **With a High School diploma or less** (in percentage)
- **With a Bachelor's degree or higher** (in percentage)

For the venue data, the features are derived out of different venue categories using one hot encoding methods. There are originally 380 kinds of venue categories in Pittsburgh and Cleveland. However, there are many venue categories only have a handful of counts across Pittsburgh and Cleveland and will cause overfitting issues for those neighborhoods which have them. Therefore, I drop the venue categories that have counts lower than 20. The resulting number of venue categories is **49.** Combining with the features from the census data, I used a total of **61** features in this analysis.

3. Exploratory Data and Analysis

3.1 Overall behaviors of each city:

These two cities are many differences. In figure 4, I show the mean values of census features among different neighborhood in each city. There are a few noticeable behaviors:

- The median population of Cleveland neighborhoods is much higher than Pittsburgh
- Pittsburgh neighborhoods have higher average income than Cleveland.

- Pittsburgh has higher white population than Cleveland.
- Cleveland has higher African American population.
- There are higher fraction of people in Pittsburgh with higher education background.

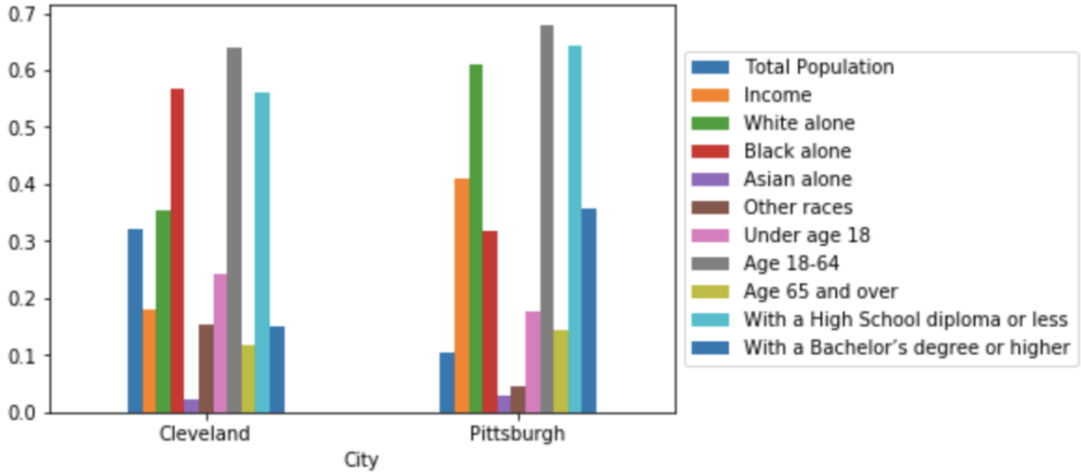
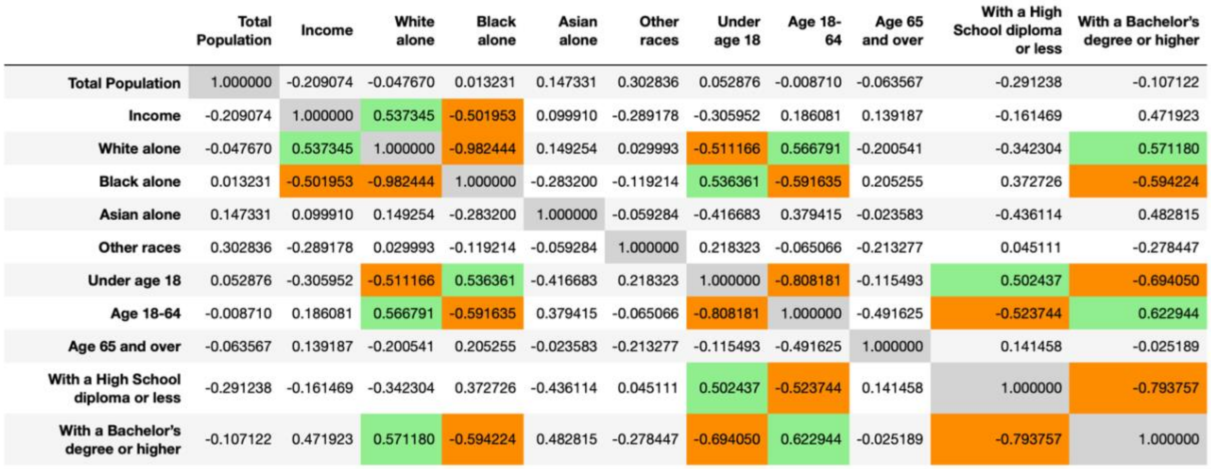


Fig. 4: Mean values of features from census data among different neighborhood in each city.

3.2 Relationship between features in the census data:

There are a few features that are either correlated or anti-correlated. In figure 5 I show the Pearson correlation coefficient values between those features. Those marked by green show high correlation with value > 0.5, and those marked by orange show high anti-correlation. Please note that those with high or low values may not always mean true correlations since many features here are correlated (e.g., "With a high School diploma or less" will be anti-correlated with "With a Bachelor's degree or higher"). Nevertheless, here are several noticeable behaviors among those features:



	Total Population	Income	White alone	Black alone	Asian alone	Other races	Under age 18	Age 18-64	Age 65 and over	With a High School diploma or less	With a Bachelor's degree or higher
Total Population	1.000000	-0.209074	-0.047670	0.013231	0.147331	0.302836	0.052876	-0.008710	-0.063567	-0.291238	-0.107122
Income	-0.209074	1.000000	0.537345	-0.501953	0.099910	-0.289178	-0.305952	0.186081	0.139187	-0.161469	0.471923
White alone	-0.047670	0.537345	1.000000	-0.982444	0.149254	0.029993	-0.511166	0.566791	-0.200541	-0.342304	0.571180
Black alone	0.013231	-0.501953	-0.982444	1.000000	-0.283200	-0.119214	0.536361	-0.591635	0.205255	0.372726	-0.594224
Asian alone	0.147331	0.099910	0.149254	-0.283200	1.000000	-0.059284	-0.416683	0.379415	-0.023583	-0.436114	0.482815
Other races	0.302836	-0.289178	0.029993	-0.119214	-0.059284	1.000000	0.218323	-0.065066	-0.213277	0.045111	-0.278447
Under age 18	0.052876	-0.305952	-0.511166	0.536361	-0.416683	0.218323	1.000000	-0.808181	-0.115493	0.502437	-0.694050
Age 18-64	-0.008710	0.186081	0.566791	-0.591635	0.379415	-0.065066	-0.808181	1.000000	-0.491625	-0.523744	0.622944
Age 65 and over	-0.063567	0.139187	-0.200541	0.205255	-0.023583	-0.213277	-0.115493	-0.491625	1.000000	0.141458	-0.025189
With a High School diploma or less	-0.291238	-0.161469	-0.342304	0.372726	-0.436114	0.045111	0.502437	-0.523744	0.141458	1.000000	-0.793757
With a Bachelor's degree or higher	-0.107122	0.471923	0.571180	-0.594224	0.482815	-0.278447	-0.694050	0.622944	-0.025189	-0.793757	1.000000

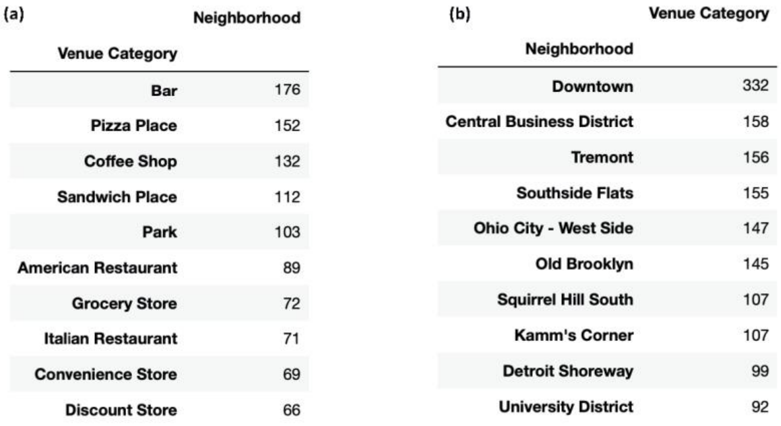
Fig. 5: Pearson correlation coefficient values between features from census data.

- Higher white population is mildly correlated with higher income

- Higher black population is mildly anti-correlated with higher income
- The areas with higher black population tend to have higher fraction of youth (< age 18)
- The areas with higher black population tend to have lower fraction of people in the age of 18- 64.
- Areas with higher population of higher education background tend to be with higher white population.

3.3 Venue data:

Here I show the top 10 most common venues in Pittsburgh and Cleveland in figure 6. Bar is the most common venues; Pizza Place is the second; and Coffee shop is the third. Since there are 124 neighborhoods in this analysis, on average each neighborhood has about one bar, Pizza Place, and Coffee shop.



(a)

Venue Category	Neighborhood
Bar	176
Pizza Place	152
Coffee Shop	132
Sandwich Place	112
Park	103
American Restaurant	89
Grocery Store	72
Italian Restaurant	71
Convenience Store	69
Discount Store	66

(b)

Neighborhood	Venue Category
Downtown	332
Central Business District	158
Tremont	156
Southside Flats	155
Ohio City - West Side	147
Old Brooklyn	145
Squirrel Hill South	107
Kamm's Corner	107
Detroit Shoreway	99
University District	92

Fig. 6: Top 10 most common venues in Pittsburgh and Cleveland.

As shown in the right panel of figure 8, the neighborhood which has the most venue categories is Downtown in Cleveland, which has 332 different kind of venues. The second one is Central Business District in Pittsburgh (downtown), which has 158 different venues.

4. Neighborhood Clustering with K-means clustering algorithm

To discover similar neighborhood, I applied the K-means clustering method to the data. The goal is to select similar neighborhood and make recommendations when restaurants/stores want to open new locations in Pittsburgh or Cleveland.

To determine the number of clusters, I used the "elbow method" to figure out if there is an optimized k value for this problem. However, as shown in figure 7, there is no obvious elbow as value k increases. I chose k **= 6** as I would hope that each cluster can on average contain ~20 neighborhoods. Another reason, which is explained below, is to minimize the number of clusters with only one neighborhood.

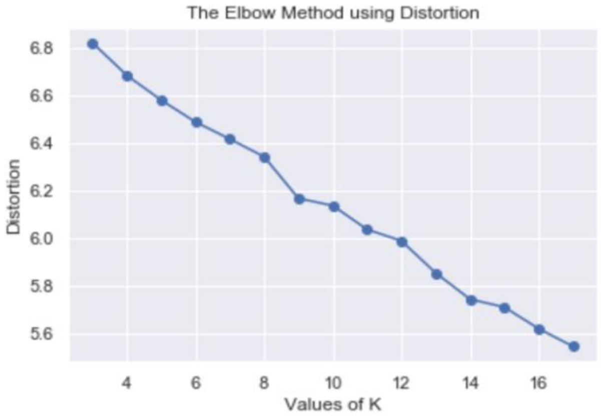


Fig. 7: Distortion as a function of cluster number k.

In figure 8 I show the fraction of neighborhood in each cluster in Pittsburgh (left pie chart) and in Cleveland (right pie chart). Pittsburgh has at least one neighborhood in each cluster, and Cleveland neighborhoods only occupy 4 different clusters. I found that as I increase k, the number of clusters with only one neighborhood increases while the type of clusters in Cleveland doesn't increase significantly.

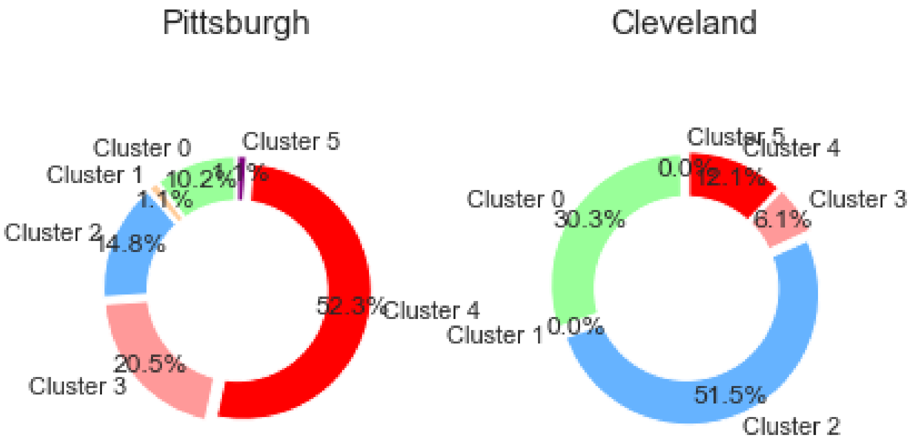


Fig. 8: Fraction of neighborhood in each cluster in Pittsburgh (left pie chart) and in Cleveland (right pie chart).

5. Results:

5.1 Cluster type spatial distribution:

In figure 9 I show the distribution of cluster types in Cleveland (left panel) and Pittsburgh neighborhood (right panel). Cluster type 3 (light purple) consists of both downtown area in both cities.

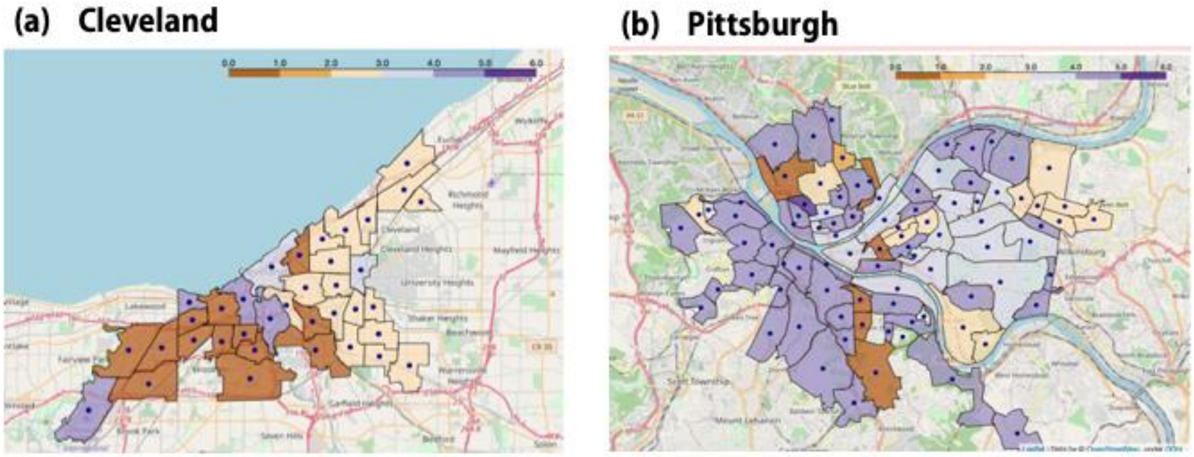


Fig. 9: The distribution of cluster types in Cleveland (left panel) and Pittsburgh neighborhood (right panel).

5.2 Cluster Characteristics:

I calculated the mean feature values in each cluster and listed those that are the highest or the lowest among the clusters.

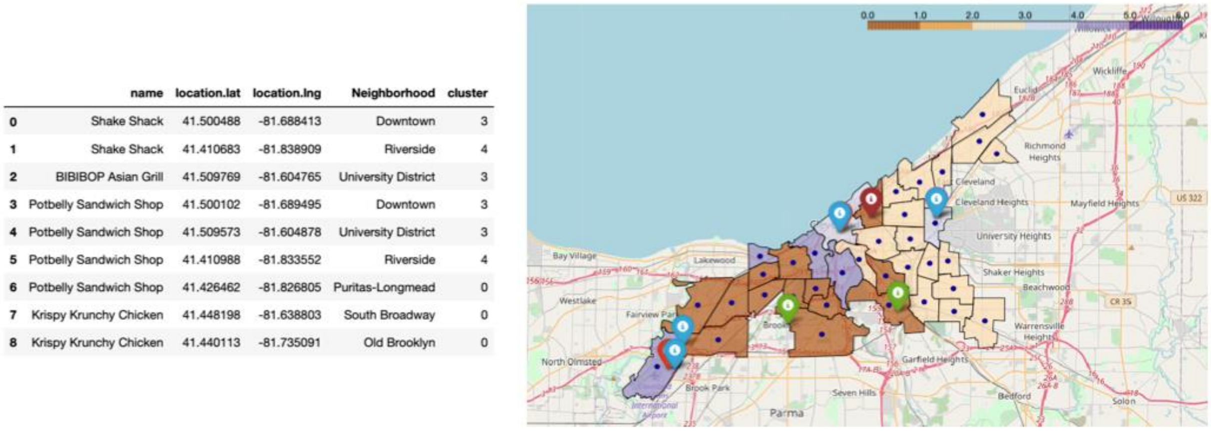
Cluster type	Features with highest average values	Features with the lowest average values
0	'Age 65 and over', 'BBQ Joint', 'Bank', 'Bar', 'Discount Store', 'Food', 'Gym / Fitness Center', 'Harbor / Marina', 'Seafood Restaurant'	'Income', 'Zoo Exhibit'
1	'Asian alone', 'Age 18-64', 'With a Bachelor’s degree or higher', 'American Restaurant', 'Art Gallery' 'Bakery', 'Brewery' 'Burger Joint', 'Café', 'Coffee Shop', 'Food Truck', 'Italian Restaurant', 'Mexican Restaurant', 'Museum', 'New American Restaurant', 'Pub', 'Theater'	'Black alone', 'Under age 18', 'With a High School diploma or less'
2	'Income', 'White alone', 'Baseball Field', 'Furniture / Home Store', 'Hotel', 'Ice Cream Shop', 'Light Rail Station', 'Park', 'Playground', 'Thrift / Vintage Store'	'Harbor / Marina', 'Museum', 'New American Restaurant', 'Theater'
3	'Black alone', 'Under age 18', 'With a High School diploma or less', 'Fast Food Restaurant', 'Intersection'	'White alone', 'Asian alone', 'Other races', 'Age 18-64', 'Age 65 and over', 'With a Bachelor’s degree or higher', 'American Restaurant', 'Art Gallery', 'BBQ Joint', 'Bakery', 'Bank','Bar', 'Baseball Field', 'Brewery', 'Burger Joint', 'Bus Station', 'Café' 'Chinese Restaurant', 'Clothing Store', 'Coffee Shop', 'Convenience Store', 'Deli / Bodega' 'Diner', 'Discount Store', 'Dive Bar', 'Food' 'Food Truck', 'Furniture / Home Store', 'Gas Station', 'Grocery Store', 'Gym / Fitness Center', 'Hotel', 'Ice Cream Shop', 'Italian Restaurant', 'Light Rail Station', 'Liquor Store', 'Mexican Restaurant', 'Park', 'Pharmacy', 'Pizza Place', 'Playground', 'Pub', 'Restaurant', 'Sandwich Place', 'Seafood Restaurant', 'Thrift / Vintage Store'
4	'Bus Station', 'Pharmacy', 'Restaurant'	'Total Population', 'Fast Food Restaurant', 'Intersection'
5	'Total Population', 'Other races', 'Chinese Restaurant', 'Clothing Store', 'Convenience Store', 'Deli / Bodega', 'Diner', 'Dive Bar', 'Gas Station', 'Grocery Store', 'Gym', 'Liquor Store', 'Pizza Place', 'Sandwich Place', 'Zoo Exhibit'	None

According to the features that are highest among other clusters, I categorize those clusters to the following:

- **Cluster 0**: *Elder people with low income*:
  Those areas tend to have more **discount store** and **seafood restaurants**.
- **Cluster 1**: *Working-age population with high education background:*
  Those areas have more **coffee shops, American restaurants, burger joints**, and **pubs**.
- **Cluster 2**: *White population with high income*:
  Those areas tend to have more **parks**.
- **Cluster 3***: Black population with low education background and more youth*:
  Those areas tend to have more **fast food restaurant** and less other kind of restaurants.
- **Cluster 4:** *Low population areas:*
  They may be near **bus stations**.
- **Cluster 5:** *High population residential area:*
  Those areas tend to have more **grocery store, diner, pizza and sandwich places**, and **Chinese restaurants**.

5.3 Recommendations on new restaurant locations based on clustering results:

In figure 10, I show the data of those restaurants (name, coordinate, neighborhood). In the right panel, their locations are marked by red (Shake Shack), pink (BIBIBOP Asian Grill), blue (Potbelly), and green (Krispy Krunchy Chicken) markers.



	name	location.lat	location.lng	Neighborhood	cluster
0	Shake Shack	41.500488	-81.688413	Downtown	3
1	Shake Shack	41.410683	-81.838909	Riverside	4
2	BIBIBOP Asian Grill	41.509769	-81.604765	University District	3
3	Potbelly Sandwich Shop	41.500102	-81.689495	Downtown	3
4	Potbelly Sandwich Shop	41.509573	-81.604878	University District	3
5	Potbelly Sandwich Shop	41.410988	-81.833552	Riverside	4
6	Potbelly Sandwich Shop	41.426462	-81.826805	Puritas-Longmead	0
7	Krispy Krunchy Chicken	41.448198	-81.638803	South Broadway	0
8	Krispy Krunchy Chicken	41.440113	-81.735091	Old Brooklyn	0

Fig. 10: Left panel: data of those restaurants (name, coordinate, neighborhood). Right panel: their locations marked by red (Shake Shack), pink (BIBIBOP Asian Grill), blue (Potbelly), and green (Krispy Krunchy Chicken) markers.

Here I discuss each restaurant one-by-one:

- **Shake Shack:**
  It opens three locations in Cleveland, and two of them lie in our neighborhood data area. Their neighborhoods are in cluster 3 and 4. A few examples of Pittsburgh neighborhood in

cluster 3 are: **Bloomfield, Shadyside, Squirrel Hill North, Squirrel Hill South,** and **Strip District.** A few examples of Pittsburgh neighborhood in cluster 4 are: **Beechview, Garfield, Highland Park, Lower Lawrenceville,** and **Swisshelm Park.**

- **BIBIBOP Asian Grill:**
  It has one location lie in our neighborhood data, although it actually has 5 stores in Cleveland. The only store locates in the cluster 3 neighborhood. Therefore, the recommended neighborhoods in Pittsburgh are: **Bloomfield, Shadyside, Squirrel Hill North, Squirrel Hill South,** and **Strip District.**

- **Potbelly Sandwich Shop:**
  It has five stores in Cleveland, and four of them lie inside our neighborhoods. Two of the stores locate in cluster 3 neighborhood, one in cluster 4, and one in cluster 0. For those neighborhoods in cluster 3 & 4, they are listed in the discussion of Shake Shack locations. For cluster 0, some neighborhoods include**: Allentown, Crawford Roberts,** and **Knoxville**.

- **Krispy Krunchy Chicken:**
  There are ten Krispy Krunchy Chicken locations, and two of them lie within our Cleveland neighborhoods. They are all within cluster 0, and some of the Pittsburgh cluster 0 neighborhood are: **Allentown, Crawford Roberts,** and **Knoxville**, as discussed in Potbelly locations.

6. Discussion:

I utilize census data and venue data to find similar neighborhoods in Pittsburgh and Cleveland, which are two mid-west city with similar size. I find that although these two cities have some fundamental differences, such as average income, population, and race percentage, some of the neighborhoods share similar features as I summarize in different characteristics in clusters in section 5.2.

I use this analysis to make recommendations to new restaurant locations at Pittsburgh based on their locations in Cleveland. This analysis can be applied to any other store types, and it can also be used to make recommendations to new locations in Cleveland as well. This analysis provides a framework to compare neighborhoods in cities for restaurant/store owners to gain insight to expand their franchise.

It will be interesting to include more features, such as school districts, store rental price, to improve this analysis. It will also be good to improve the venue categories, for which some of them may be redundant or not relevant to the research targets.