Stamford, CT Neighborhood Analysis

A Look at What Neighborhoods in Stamford Have to Offer Prospective Residents and Businesses

Created by

Jonathan Lowthert

Introduction / Business Problem

Before we get into the data, we begin with a little background on the problem being faced, and how we aim to solve it.

Background

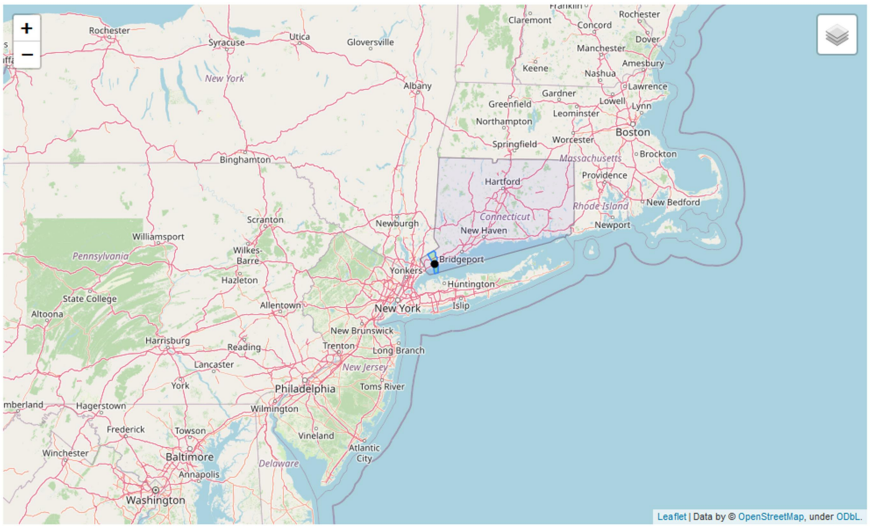
Professional life begins for many in a big city, which offers myriad job opportunities as well as numerous social activities. After several years of living in the big city, many people tire of the fast pace and limited breathing room, and opt to move to the suburbs. Sometimes this move is precipitated by changes in family, such as having young children.

A popular misconception is that the suburbs are completely opposite of the big city, consisting of small towns with large houses and no social amenities. But, in many areas this is not true — the suburbs contain many larger towns and smaller cities that offer a great combination of city life and non-city life.

New York City is not only the paragon of big cities, but its surroundings provide great examples of suburban cities and towns. The tri-state area of New York, New Jersey, and Connecticut provide numerous areas in which to live that echo many of the amenities that New York City offers. Many people live in these areas outside of the city and continue to work in New York City, commuting every day to work in the city. And if a person living in one of those towns just has to have something that can only be found in NYC, then it is a short car or train ride away.

Stamford, Connecticut

The city of Stamford is located in the state of Connecticut, and is just 45 minutes away from New York City by train. Stamford is a small city, with a population of between 120,000 and 130,000 people. Stamford is also a varied city, containing a sparsely-populated northern district, a metropolitan downtown area, and numerous residential areas. The downtown area can feel much like Manhattan, with its numerous corporate offices, especially those from the financial industry.



Problem

Since Stamford is such a varied city, it can be difficult for somebody moving from another area to understand the neighborhoods and what they each offer. Stamford, like most suburban cities, is not so much of a destination city, so it is unlikely that most people moving into it will have had much experience with its neighborhoods and amenities.

The problem that is to be solved for many people is where within Stamford, Connecticut to move to.

Audience

The audience for this report is people who have a vested interest in understanding the neighborhoods in Stamford, CT. Many of these people will be involved in moving to Stamford and needing to find housing. Other members of the audience may want to understand the neighborhoods to make business decisions.

1. **People relocating to Stamford**
   People who are relocating to Stamford care where they want to live — they want to have amenities nearby that they are interested in, whether they are social, shopping, or educational.
2. **Real estate agents**
   Real estate agents need to know what each neighborhood offers in order to find buyers the best home when they are looking, and to provide sellers with a best sale price.
3. **Corporations investigating moving to Stamford**
   Companies moving into Stamford need to know about neighborhoods both for their prospective office site, as well as for the good of their employees.
4. **Businesses investigating opening a store in Stamford**
   Businesses that might be looking to open a physical store in Stamford need to know what other types of businesses are in prospective neighborhoods, as they want to be able to drive shoppers/clients to their new storefront, but without having an overabundance of competition.

Goal

The goal of this project is to investigate the neighborhoods in Stamford, Connecticut. The first step will be to determine where the various neighborhoods are located. The second step will be to understand the characteristics of each neighborhood.

Data

The dataset that we will be using is from FourSquare. FourSquare data contains information on numerous points of interest in the city under investigation, with categories such as restaurants, parks, and schools. The terms “point of interest” and “POI” will be used interchangeably throughout this report.

Many of the previous projects in this course have used neighborhoods as they are defined by external datasets. These external datasets have defined neighborhoods by such arbitrary measures as postal codes and historical/political districts. It does not seem that this always represents a good measure of what constitutes a neighborhood.

I will be using the FourSquare data itself to define neighborhoods based on where points of interest are located geographically. I will be using clustering algorithms as part of this process, and will specifically be using the latitude and longitude fields from the FourSquare data.

Once the neighborhoods have been defined, I will dig deeper into the characteristics of each point of interest, and attempt to understand what major amenities define each neighborhood.

More specifics of the data are covered in the Methodology section, when the data is actually collected, explored, and then analyzed.

Libraries

To implement the investigation of the neighborhoods in Stamford, CT, the following Python libraries and web services will be used:

- Numpy and Pandas (for data analysis)
- GeoPy (for location services)
- Requests (to request data from web services, such as FourSquare)
- Matplotlib (for plotting)
- Sklearn (for machine learning, such as clustering algorithms)
- Folium (to create maps)
- Shapely (to deal with mapping geometry)
- FourSquare (to provide data on points of interest)

Methodology

In this section, I will discuss the methods used to collect, process, and analyze the data that was used in this project. For more details, one can view the Jupyter Notebook that was used.

A limited subset of data science tools were used in this project. No Inferential Statistical Testing was done. The primary machine learning technique that was used was k-Means Clustering, which is an unsupervised segmentation algorithm that was used to determine where POIs were grouped together geographically.

Exploratory Data Analysis

Before beginning to collect data, I queried from FourSquare for a subset of the city of Stamford. This allowed me to investigate the data to see what the data schema looked like, as well as to judge what portions of the data would be relevant to the goal of this project.

The FourSquare API provides a call to explore venues and allows one to pass in category ID(s) to query, and defines an extensive hierarchy of categories. I chose to use the top-level categories to see what kind of data was available. I then looked at the results of this subset query to decide what data would be relevant to meeting the goal of this project. For some categories the results just were not relevant, and for others they were flooded with results that hid the data that I thought would be relevant.

For example, in the Arts & Entertainment category there was an over-abundance of entries for statues that were part of a 2012 street show; somebody involved in the show must have checked in all of the statues and they remain in the FourSquare database. Not only do these entries flood the data making it look like there are a lot of Art Gallery entries, but these statues are not even displayed anymore. So, it is desirable to remove these entries.

Another example pertains to Gyms, many of which were labelled as belonging to apartment complexes. These are not publically available facilities, so they exaggerate the number of gyms that are available in a neighborhood; it is also desirable to remove these entries. There are other recreation entries that are not flooded that can provide more meaningful data, such as those in the categories of Yoga Studio and Gym Pool.

In an attempt to make the data more relevant, I chose to selectively add or subtract sub-categories from the FourSquare top-level categories to make each top-level category more relevant. The following table outlines what categories I used and why.

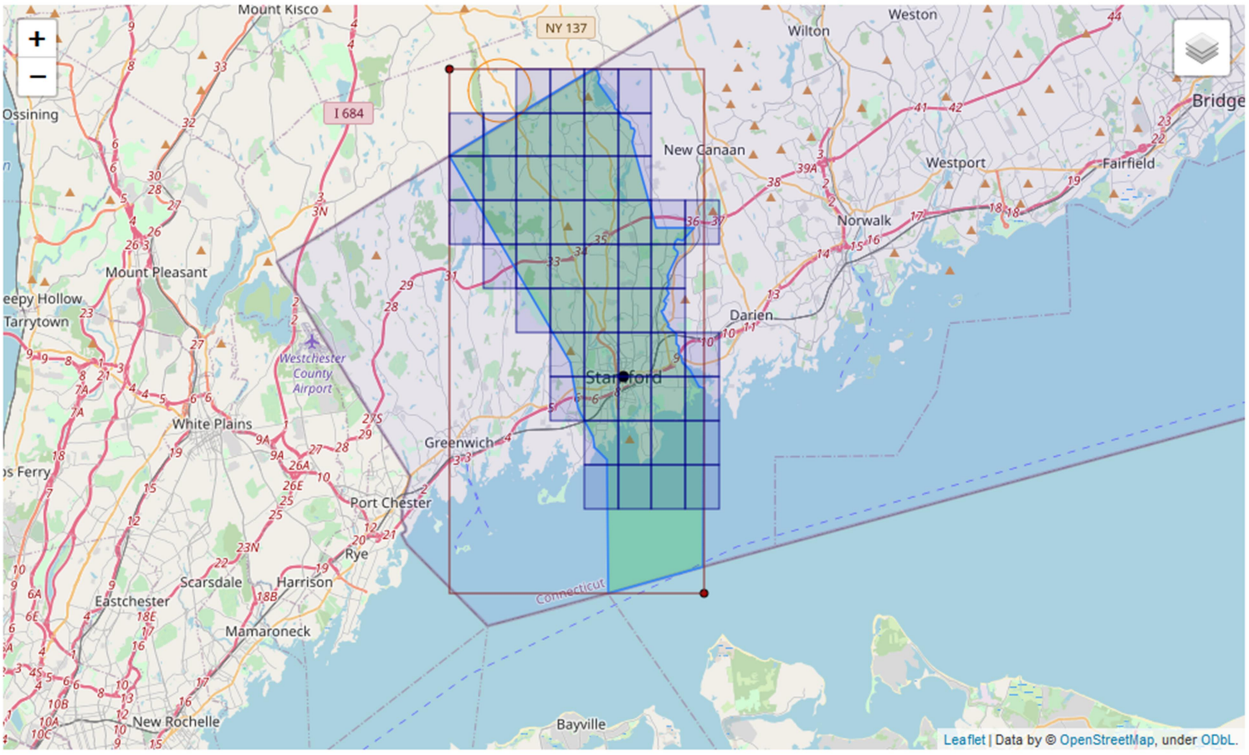
Category	**Queried**	**Removed**
Arts & Entertainment	All	Art Gallery (too many POI entries from a street show in 2012)
College & University	Community College, Trade School, University	-
Food	All	-
Nightlife Spot	All	-
Outdoors & Recreation	All	Removed Gym, Gym / Fitness Center, Roof Deck (many of which were not stand-alone, but provided by an apartment building; Summer Camp (not a permanent POI)
Other Places	Library, Post Office, Social Club	-
Shop & Service	All	Lawyer (not relevant)
Travel & Transport	Many Airport, Train, and Bus categories	Bus Stop (too many)
Residence	None	-

In order to implement the above selective querying and dropping of data, I created a helper class and instantiated an instance of the class that described which categories to query and which to delete; see the '`top_level_categories`' object in the Python code.

Data Collection

The next step in this project was to gather the data necessary. The data was to come from FourSquare, which lets one query for POI information in a given radius around a central latitude-longitude location. For each query, FourSquare provides the most popular POIs in the region, and has some internal limit on the number of results returned. So, if one were to query FourSquare with a location in the center of Stamford and a large radius, then only the POIs that were deemed the most popular would appear in the results, and the majority of POIs in the city would not be counted at all.

To collect as much data as possible, I chose to break the city of Stamford up into a grid of boxes, with each box being 2000 meters per side. I then iterated over all of the squares and called the FourSquare API with the center location of each square and a far smaller radius.



The map above shows how I divided the city of Stamford up into a grid. The green polygon defines the boundaries of the city. The red rectangle is a bounding rectangle that includes the entire city and was used to define the left and top edges of the grid pattern. The blue boxes are grid elements that intersect with the city polygon, and which we will query for data; no data is required from boxes that do not lie in the city. Finally, the orange circle represents the radius required to query from FourSquare that will contain all POIs in the grid box. The Shapely library was a great help in dealing with these geographic shapes.

For each box that intersected with Stamford, the following steps were taken:

- Query FourSquare for POIs in each desired category, and remove POIs from sub-categories that are not desired

- Remove entries that are not located in Stamford (i.e. are in neighboring cities/towns and show up giving that a box/circle do not line up on city boundaries)

The data fields that were collected for each box were a subset of the data fields that FourSquare provides. The following DataFrame snippet shows what data fields were relevant for our data project.

```
df[0:3]
```

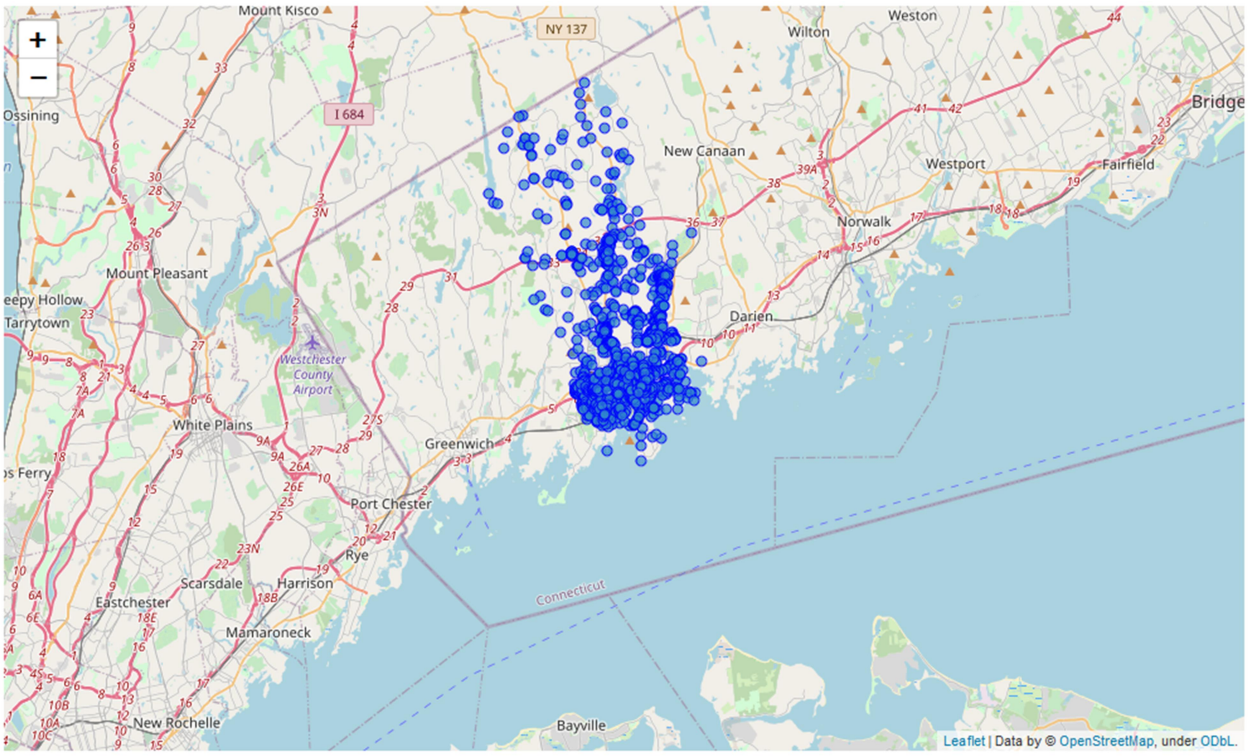
	name	address	city	state	latitude	longitude	zipcode	venue_id	category_name	category_id	top_catego
0	Rockrimmon Country Club	2949 Long Ridge Rd	Stamford	CT	41.160919	-73.594320	06903	4c30a3ed3896e21e09cbe590	Golf Course	4bf58dd8d48988d1e6941735	Outdoors Recreati
1	Optimum WiFi Hotspot	2949 Long Ridge Rd	Stamford	CT	41.160381	-73.596077	06903	5993321edd12f839eb08a7e0	Business Service	5453de49498eade8af355881	Shop Servi
2	Dorothy Heroy Park	1-99 Riding Stable Trail	Stamford	CT	41.174521	-73.560152	06903	4d98abac0caaa1431450b2b3	Park	4bf58dd8d48988d163941735	Outdoors Recreati

As data was retrieved for each box, it was added to a main DataFrame that contained the data for all boxes.

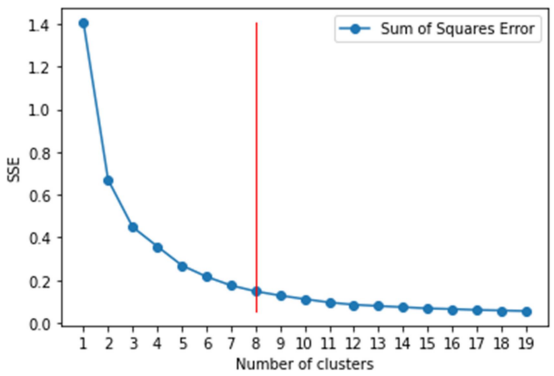
After data for all POIs were collected from FourSquare there were duplicates in the main data set, caused by the fact that the circles required to query with extended beyond the box boundaries and overlapped. Duplicates were eliminated by dropping any excess entries that had the same ‘venue_id’, which is a unique ID provided by FourSquare for each POI.

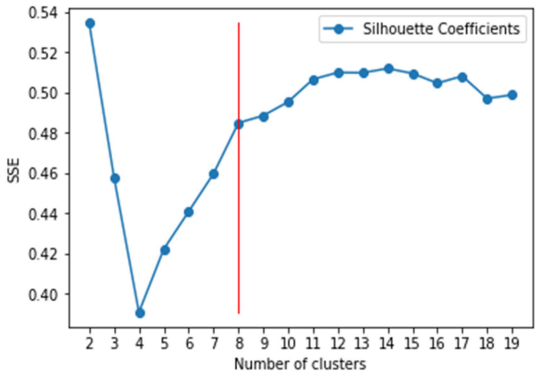
Data Analysis

After collecting the data, it was plotted to get a general idea of what we would be looking at. The following map shows all POIs in Stamford that were collected.



The goal was to segment these POIs into clusters. To do so, I used k-Mean clustering. I first iterated over a number of values of k to find a reasonable number of clusters to use, and created the following graphs.



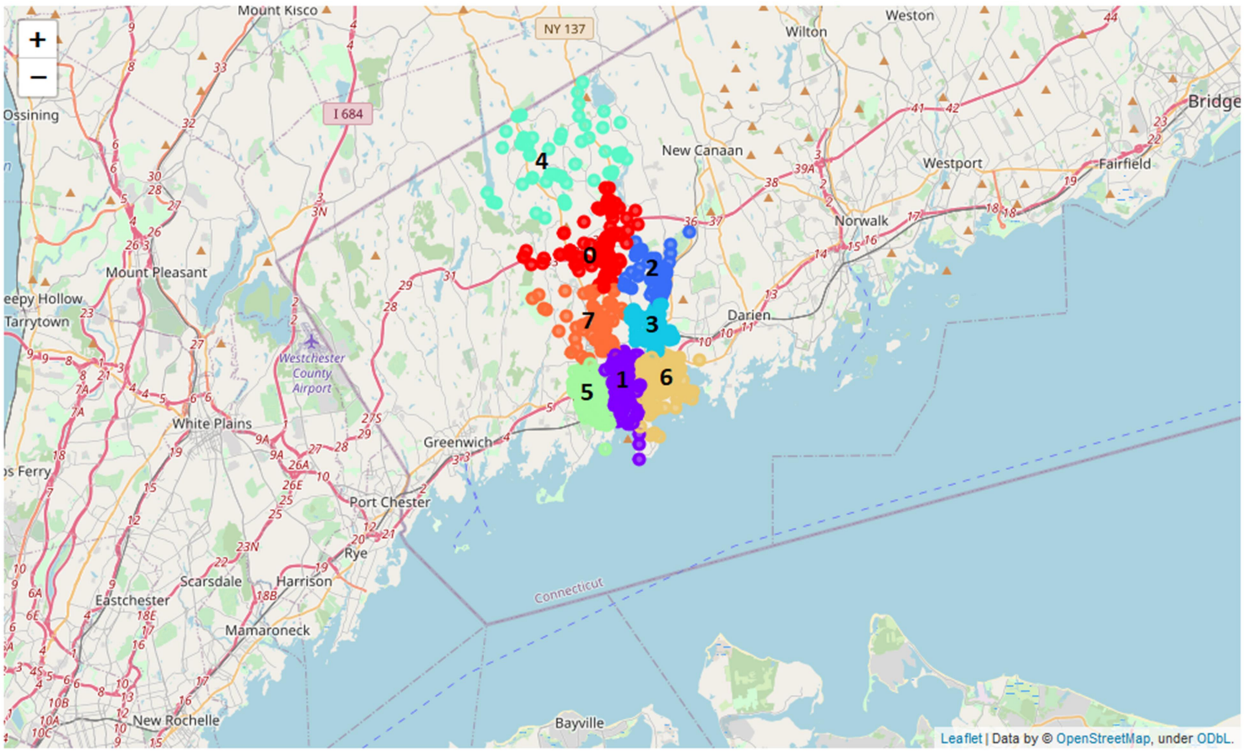


By looking at the graphs of the Sum of Squares Error and Silhouette Coefficients, I chose 8 as the number of clusters since it was far down in the elbow of the SSE graph and had the last large jump up in the Silhouette graph.

Results

After performing the preliminary steps of data collection and analysis, we were able to get the final results—clustered data that describes neighborhoods within Stamford.

The k-Means clustering algorithm was run for a final time on the dataset with a k value of 8. Following this segmentation, I mapped the POIs to see how the data was clustered.

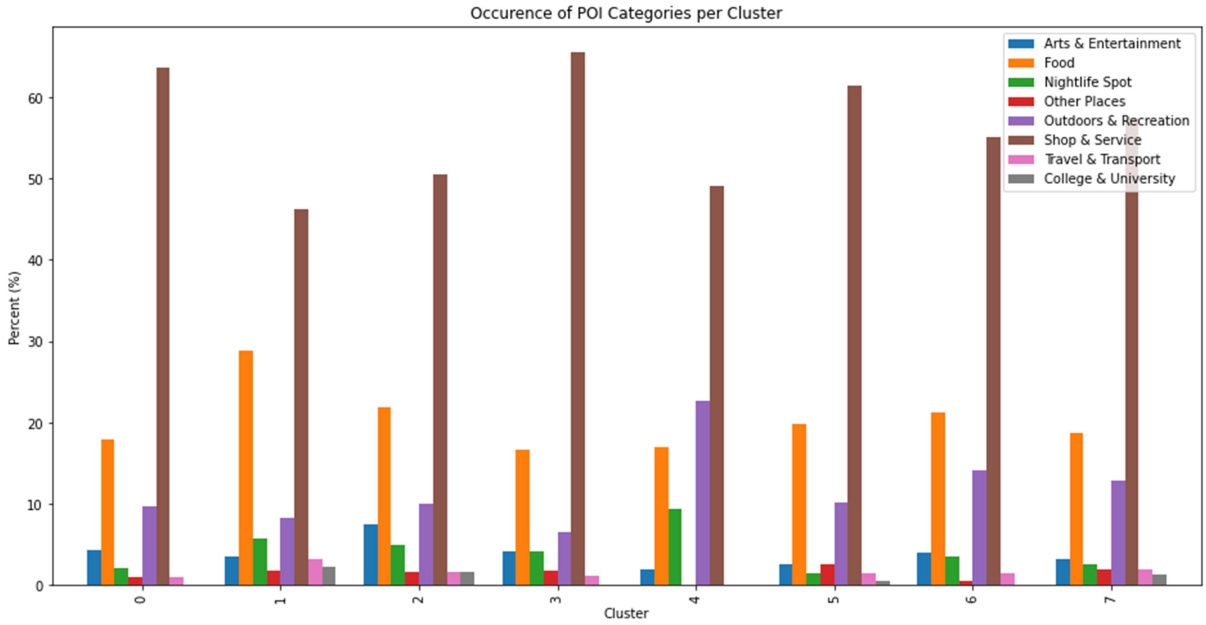


One can see the eight clusters in the map, with three along the coastline, four in the center of the city, and one in the northern Stamford area.

The next step was to see what types of amenities were offered in each cluster, i.e. if each cluster would have some characteristic category types that might be desirable to certain individuals. To do this, I grouped the data by clusters and summed up the entries in each category, getting the following DataFrame.

	Arts & Entertainment	Food	Nightlife Spot	Other Places	Outdoors & Recreation	Shop & Service	Travel & Transport	College & University
0	8.0	33.0	4.0	2.0	18.0	117.0	2.0	0.0
1	14.0	114.0	23.0	7.0	33.0	183.0	13.0	9.0
2	9.0	26.0	6.0	2.0	12.0	60.0	2.0	2.0
3	7.0	28.0	7.0	3.0	11.0	110.0	2.0	0.0
4	1.0	9.0	5.0	0.0	12.0	26.0	0.0	0.0
5	5.0	39.0	3.0	5.0	20.0	121.0	3.0	1.0
6	8.0	42.0	7.0	1.0	28.0	109.0	3.0	0.0
7	5.0	29.0	4.0	3.0	20.0	89.0	3.0	2.0

Although one can start to see some differences between the clusters by looking at the table, I then graphed the data by category percentage cluster for better visualization.



Discussion

The results that were obtained in the previous section paint a picture of what types of amenities each neighborhood/cluster in Stamford provides.

By looking at the above graph, we can decide what types of points of interest each cluster/neighborhood is most represented by.

The "Shop & Service" category is the most prevalent for all clusters, as might be expected because it covers a large range of entry types, from customer-facing shopping to landscaping and accounting services. Because it is so large for all clusters, we will ignore it in defining neighborhood characteristics.

For the most part, the clusters have very similar offerings in terms of points of interest. The following clusters have a slightly more unique profile:

- 1 = lots of food; also has most transportation options
- 2 = well represented by arts & entertainment
- 4 = has more outdoor & recreation activities, lots of nightlife; maybe a younger crowd

One conclusion that we can draw is that all of the neighborhoods offer a good range of points of interest. This means that a person or business moving to Stamford will be able to choose from neighborhoods based on other characteristics, such as by how much housing costs in that neighborhood. And if a person has very specific need or desire, there are a few neighborhoods that might provide an environment that is just a little more to their liking!

Conclusion

The data that was collected, and the results that were achieved, show that almost every neighborhood in Stamford provides a great selection of amenities! If a person is particularly focused on a few specific needs, there may be a specific neighborhood that would be better suited to him or her.

If one were to want a more detailed picture of the amenities available in each neighborhood in Stamford, one could dig even deeper into the data. For example, FourSquare breaks restaurant down into numerous types, as detailed as Cafes, Chinese Restaurant, and Taco Joint. One could focus on different categories to see if the different clusters had differences in the types of food offered, or shopping and services provided.