*David Dancis*

*IBM Data Science Capstone Project*

**Exploring Philadelphia Neighborhoods with Machine Learning**

***Leveraging geolocation data and k-means clustering to identify the best neighborhoods to open an Italian restaurant in Philadelphia***

### Background

Location is a critical factor for any business, and it’s especially important in the restaurant industry which typically operates on thin margins, relies heavily on foot traffic, and is heavily impacted by the level of competition in the area. While COVID-19 pandemic has taken a toll on the industry, the vaccine rollout has provided good reason to believe that it will rebound over the next year, and that the fundamentals of the industry should return remain relatively constant.

This analysis examines the restaurant distribution in Philadelphia neighborhoods, and focuses on identifying the best potential locations to open an Italian restaurant in the city. Throughout our examination, we seek to address several key questions:

* What neighborhoods have the highest concentration of Italian restaurants?
* Are there neighborhoods with similar characteristics without a strong Italian restaurant presence that are likely to be amenable to a new restaurant opening in the area?

We will begin by examining the types of restaurants currently in each neighborhood, and then utilize k-means clustering to identify areas that present the best opportunity to open a new Italian restaurant.

### Stakeholders

This analysis will be of interest to the following groups of people:

* Data science students interested in learning how geolocation data and machine learning can be leveraged to solve a practical business problem.
* Philadelphia residents seeking to learn more about business and restaurant distribution in their city.
* Business owners seeking to identify the best potential locations to open an Italian restaurant Philadelphia.

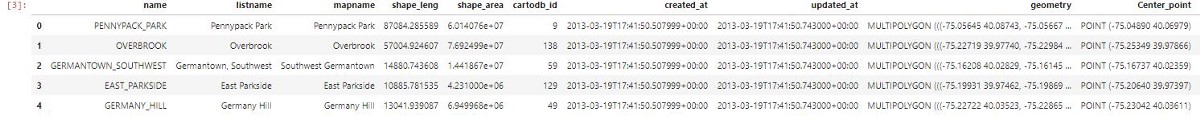
### Data

This analysis relies on the following datasets:

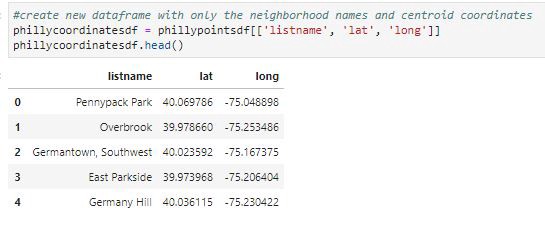
* [OpenDataPhilly](https://www.opendataphilly.org/) to find geographical coordinates for Philadelphia neighborhoods.
* [Foursquare’s](https://foursquare.com/) venue dataset accessed via API, to identify area restaurants in each neighborhood.

### Methodology

We will begin by importing geographical data from OpenDataPhilly to determine the coordinates for each neighborhood. As neighborhood boundaries are complex, the coordinates are provided in a MultiPolygon geometry via a GeoJSON file. To simplify our analysis, we will use the GeoPandas library calculate the centroid for each neighborhood, and will then use those coordinates as the basis for the rest of our analysis.

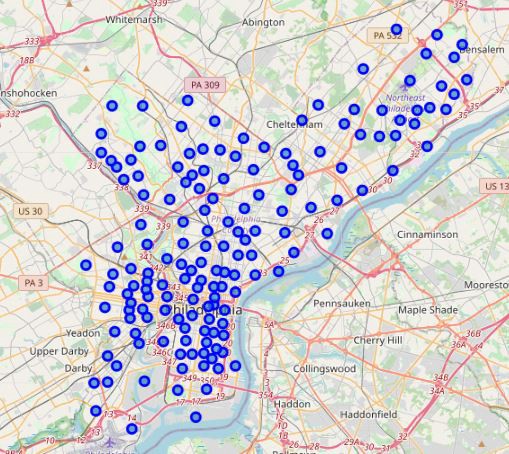
*Philadelphia neighborhood coordinates, geometry, and centroids*

Next, we will create a new dataframe in pandas to extract only the neighborhood name, and latitude and longitude coordinates for each centroid.



*Philadelphia neighborhood coordinates*

With our neighborhood coordinates established, we will leverage Folium’s mapping library to create a map of all Philadelphia neighborhoods. Each neighborhood centroid is marked by a blue dot in the map below.

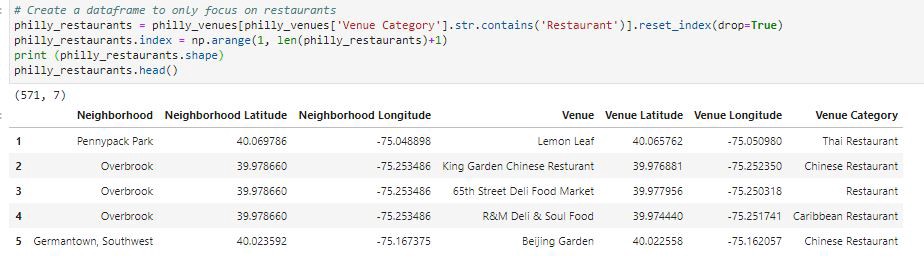


*Philadelphia neighborhood map using Folium’s Python library*

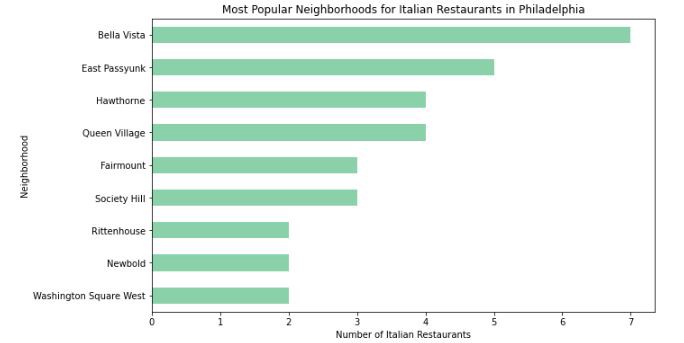
Now that we have our neighborhoods mapped and the dataframe has been prepared, it is time to begin exploring the venues in each neighborhood. For this portion of our analysis, we utilized Foursquare’s API to identify the top 100 venues within 500 meters of each neighborhood centroid.

In our initial approach, we leveraged k-means clustering to examine all venues in each neighborhood, regardless of the business type to try and identify meaningful clusters. However, we found that this approach did not provide any useful information, as nearly 90% of our neighborhoods fell within a single cluster, regardless of the k selected. To hone our analysis, we’ve reduced our dataset to only focus on the *restaurants* within each neighborhood, and removed all other business venue types.

We expect this approach to be more instructive than examining all venue types, since our objective is to open a new restaurant, and our results can be skewed by confounding factors that are not applicable to achieving this objective if we expand the dataset to include other venue types. To support this approach, we have created a new pandas dataframe that only focuses on restaurants in Philadelphia.

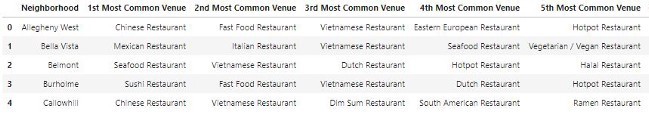
*Dataframe consisting entirely of Philadelphia restaurants, excluding all other venue types*

Next, we will use one hot encoding to get a count of the venue types in each neighborhood. As our objective is to open an Italian restaurant, we will plot the top 10 neighborhoods in Philadelphia with the highest concentration of Italian restaurants.

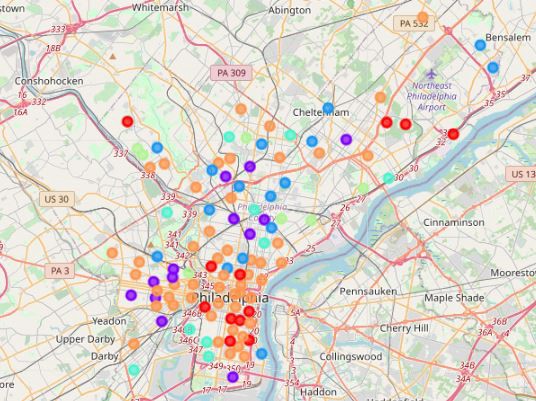
*Italian restaurant concentration by Philadelphia neighborhood*

This chart provides some useful information around Philadelphia neighborhoods where the highest number of Italian restaurants currently reside. Unsurprisingly, these areas are predominantly located in South Philadelphia where there is a higher concentration of Italian neighborhoods. However, if we are to open a new restaurant, it does not behoove us to do so in an area where demand may already be saturated by already-established restaurants. Instead, we will continue our analysis by leveraging k-means clustering to identify neighborhoods that may have similar restaurant compositions as those with high Italian restaurant concentrations, but do not currently have many Italian restaurants in the area.

By leveraging the one hot encoding performed earlier, we can count the top 10 most common types of restaurants in each neighborhood. We will use this dataset as the basis for our clustering algorithm.

*Most common restaurant types in each Philadelphia neighborhood*

Using the elbow method, we determined that this data was best broken apart into 6 clusters. We then utilized scikit-learn to cluster our data with k=6, and plotted these neighborhood clusters using Folium in the map below.



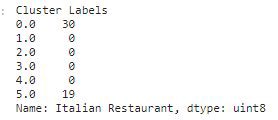
*Philadelphia restaurant clusters*

Next, we will examine a breakdown of restaurant types by cluster, as this will be instructive in looking for the similarities that helped shape each cluster.



*Cluster breakdown by restaurant type. (Figure shows only first 5 restaurant types in dataframe for easier viewing)*

As we are interested in opening an Italian restaurant, we will take a closer look at the clusters for these types of establishments. Here, we see that all Italian restaurants in Philadelphia are located in clusters 0 and 5.



There are 12 neighborhoods in cluster 0, and we see that Italian restaurants are among the top 3 most common restaurant types for all of them. This cluster is clearly shaped by the popularity of Italian restaurants in these neighborhoods, but it also leaves us very little opportunity to open a new restaurant as these areas are already saturated by Italian restaurants.



*The top 3 most common venues in cluster 0 show the popularity of Italian restaurants in these neighborhoods*

Therefore, we move on to cluster 5. This is a much larger cluster, as it contains 50 neighborhoods, and has a much more diverse set of restaurant venues. Ideally, we should seek to open our restaurant in a neighborhood that is not already saturated by Italian restaurants. However, we may want there to at least be some presence of Italian restaurants in the area to prove that there is a demand for this type of restaurant in that neighborhood. To do this, we will filter our dataframe to only show neighborhoods are between the 6th and 10th most common venues. This will show that there is some demand for the service in the area, while helping avoid stiff competition where Italian restaurants are among the top 5 most common venues.

*Recommendations for Philadelphia neighborhoods to open an Italian restaurant*

### Results

After filtering the data, we see that Passyunk Square, Manayunk, Old Kensington, Chinatown, Lower Moyamensing, Spruce Hill, and University City appear to be the best areas to open an Italian restaurant. If we eliminate Chinatown from this list, we are left with 6 neighborhoods that are best suited for our objective of opening an Italian restaurant.

Our analysis has revealed that cluster 5 appears to be the best location to open an Italian restaurant, and within this cluster, we identified 6 neighborhoods that appear to be the best candidate to open our business. The k-means clustering algorithm has shown us that these neighborhoods have similar features as those with a heavy presence of Italian restaurants, but by filtering our data to show only those where Italian restaurants are between the 6th and 10th most common venues, we have identified the neighborhoods where there is a small presence of Italian restaurants to prove the concept, but not so many that we expect too much competition.

### Discussion

While our analysis is generally instructive toward the appropriate areas to open an Italian restaurant in Philadelphia, there are limitations to our approach.

First of all, early in our analysis we calculated the centroid for each neighborhood and then conducted a search for venues within 500 meters of that point. This is an imperfect approach, as not every venue in each neighborhood will be within the given radius for that particular centroid. Some of the larger neighborhoods are likely to have venues that fall outside that range which will result in missed data that could potentially skew our results.

Secondly, the number of venues returned for each neighborhood had a limit of 100. Certainly there are neighborhoods with well over 100 venues, but we limited this number to simplify our analysis and for easier computing.

Finally, there are some limitations to our dataset. The number of restaurants across all neighborhoods appear to be smaller than expected, even with the limitations listed above, which could mean the Foursquare API is not up to date or is missing restaurants and venues, which would of course impact our results.

Even with these limitations, our analysis is still directionally informative in identifying similar areas within Philadelphia as it pertains to restaurant makeup, and that information can still be leveraged to help us achieve our objective of identifying the best neighborhoods to open an Italian restaurant in the city.

### Conclusion

Throughout our analysis, we utilized datasets from OpenDataPhilly and Foursquare, and used a number of python libraries including pandas, scikit-learn, Folium, and NumPy. By leveraging geolocation data and unsupervised machine learning, we examined 158 Philadelphia neighborhoods, and narrowed it down to 6 potential areas that are best suited to open an Italian restaurant.