# Analyzing Neighborhoods of Philadelphia For Starting A New Restaurant

## Introduction

### Background

The city of Philadelphia, or Philly, is the largest city in the U.S. state of Pennsylvania and the sixth-most populous U.S. city, with an estimated population of more than 1.5 million in 2019. Since 1854, the city has had the same geographic boundaries as Philadelphia County, the most-populous county in Pennsylvania and the urban core of the eighth-largest U.S. metropolitan statistical area, with over 6 million residents as of 2017 in over 100 neighborhoods. Philadelphia is also the economic and cultural center of the greater Delaware Valley along the lower Delaware and Schuylkill rivers within the Northeast megalopolis. The Delaware Valley's population of 7.2 million makes it the eighth-largest combined statistical area in the United States. In this project, I will explore the neighborhoods of Philadelphia and make a prediction on where the best location might be for starting a new restaurant in the area.

### Business Problem

A client who is interested in investing in a restaurant in Philadelphia has reached out to us. They would like us to study the market and suggest them a location in one of the neighborhoods which would be in best interest of the business. Our main objectives of this project are to extract and analyze the data from multiple sources about various neighborhoods of Philadelphia using various data science techniques and suggest our client with the most appropriate location for their restaurant. Although this project is designed for finding the best location for a new restaurant, people who would like to start their business other than restaurant in the Philadelphia area maybe also find this project interesting.

## Data

The data used in this project comes from multiple sources including:

- List of Philadelphia neighborhoods
- Geographical coordinates of the neighborhoods of Philadelphia.
- Venue data for each neighborhood.

### Neighborhood Data

This data was extracted from Category : Neighborhoods in Philadelphia Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Philadelphia) using web scraping with BeautifulSoup library in Python. This will give us a detailed list of neighborhoods' names in Philadelphia.

Geographical Coordinates

Next, the geographical coordinates of the center of each neighborhood were obtained using geocoder in Python. Geographical coordinates are necessary for locating neighborhoods in the map of Philadelphia and inquiring venue data from FourSquare. After using geocoder, we added two columns to our dataframe with latitudes and longitude information of each neighborhood as shown below:

```
In [11]: philly_data.head()
```

Out[11]:

| | Neighborhoods | Latitude | Longitude |
|---|---|---|---|
| 0 | Academy Gardens | 40.06178 | -74.99628 |
| 1 | Allegheny West | 40.00361 | -75.17716 |
| 2 | Andorra | 40.07261 | -75.23129 |
| 3 | Angora | 39.94399 | -75.23803 |
| 4 | Ashton-Woodenbridge | 40.07178 | -75.02295 |

### Venue data from FourSquare API

Finally, we obtained venue data using FourSquare API. This venue data was used to study the venues in Philadelphia. By analyzing venues categories and find out the top 10 most popular venues for each neighborhood as shown below, we will divide all neighborhoods into multiple clusters and draw the conclusion and decide what are most appropriate neighborhood to start a new restaurant for our client.

| | Neighborhood | 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 | Academy Gardens | Golf Course | Liquor Store | Gym / Fitness Center | Zoo Exhibit | Farm | Electronics Store | Entertainment Service | Ethiopian Restaurant | Event Space | Exhibit |
| 1 | Allegheny West | Intersection | Grocery Store | Pizza Place | Fast Food Restaurant | Farm | Entertainment Service | Ethiopian Restaurant | Event Space | Exhibit | Falafel Restaurant |
| 2 | Andorra | American Restaurant | Playground | Zoo Exhibit | Fast Food Restaurant | Entertainment Service | Ethiopian Restaurant | Event Space | Exhibit | Falafel Restaurant | Farm |
| 3 | Angora | Chinese Restaurant | Park | Grocery Store | Shopping Plaza | Donut Shop | American Restaurant | Discount Store | Road | Supermarket | Light Rail Station |
| 4 | Ashton-Woodenbridge | Liquor Store | Business Service | Miscellaneous Shop | Zoo Exhibit | Farmers Market | Ethiopian Restaurant | Event Space | Exhibit | Falafel Restaurant | Farm |

## Methodology

### Feature Extraction

Feature extraction was carried out through One Hot Encoding as shown below. In this method, each feature is a category that belongs to a venue which is then converted into binary, this means that 1 means this category is found in the neighborhood and 0 means the opposite. Then, all the venues are grouped by the neighborhoods and the mean was computed at the same time. This will give us a neighborhood for each row and each column will contain the frequency of occurrence of that category of a venue.

In [21]:
```
# one hot encoding
philly_onehot = pd.get_dummies(philly_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
philly_onehot['Neighbourhood'] = philly_venues['Neighbourhood']

# move neighborhood column to the first column
fixed_columns = [philly_onehot.columns[-1]] + philly_onehot.columns[:-1].values.tolist()
philly_onehot = philly_onehot[fixed_columns]
philly_onehot.head()
```

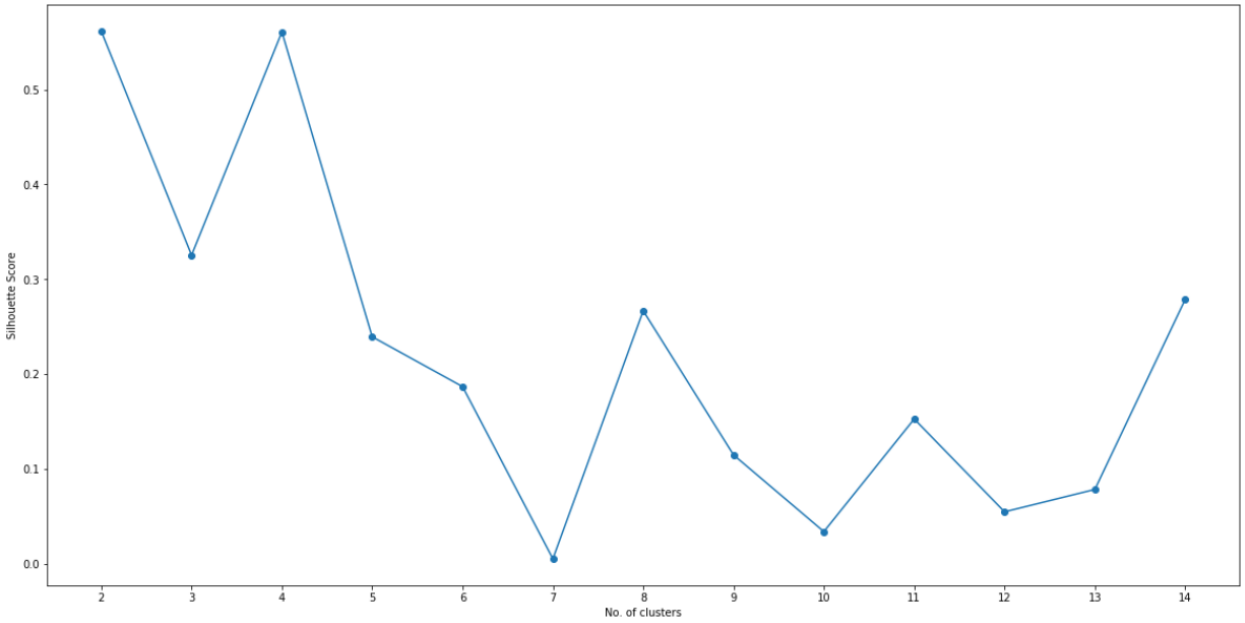
Out[21]:

| | Neighbourhood | ATM | Adult Boutique | Advertising Agency | African Restaurant | Airport Terminal | American Restaurant | Antique Shop | Arcade | Argentinian Restaurant | Art Gallery | Art Museum | Arts & Crafts Store | Arts & Entertainment | Rest |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Academy Gardens | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 1 | Academy Gardens | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 2 | Academy Gardens | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 3 | Academy Gardens | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 4 | Allegheny West | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |

## K-means clustering

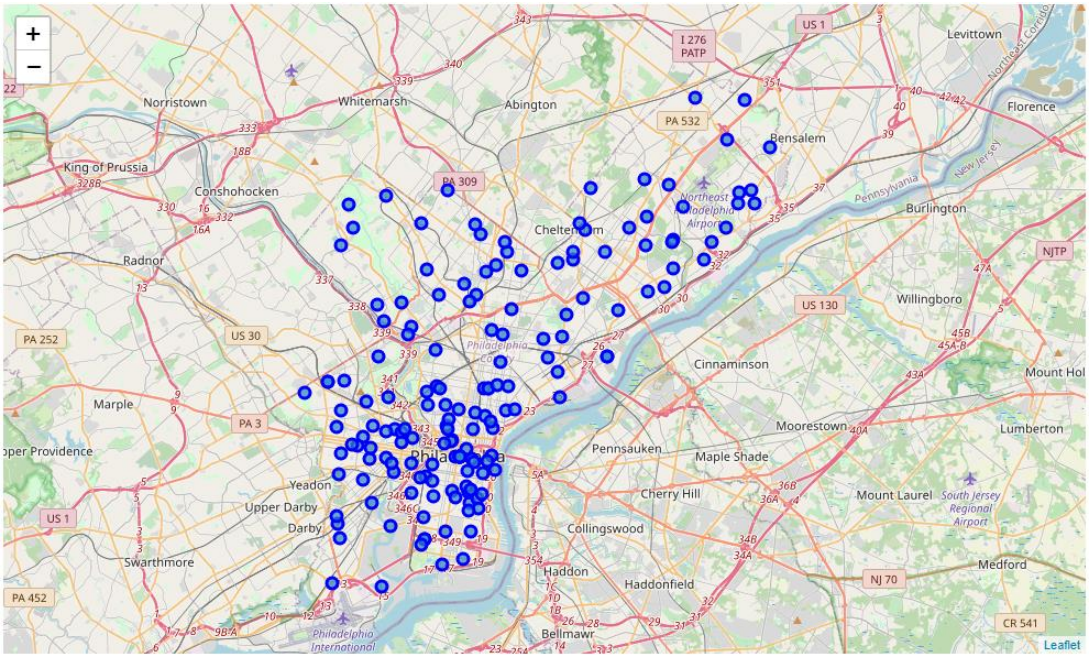
Unsupervised learning was carried out in order to discover the similarities between neighborhoods. K-means clustering was implemented due to its simplicity and its similarity approach to find patterns. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the multi-dimensional features. It is necessary for this algorithm to have a prior idea about the number of clusters since it is considered an input of this algorithm. For this reason, the elbow method is implemented. A plot that compares error vs number of clusters is done and the optimum value of k is selected as shown below. Then, the optimum value of cluster number was defined and the data is divided into the predefined number of clusters. Finally, further analysis of each cluster is done to investigate similarities within each cluster.

```
# find the best k value for K-means clusering
max_range = 15 #Max range 15 (number of clusters)

from sklearn.metrics import silhouette_samples, silhouette_score


indices = []
scores = []

for philly_clusters in range(2, max_range) :

    # Run k-means clustering
    philly_gc = philly_grouped_clustering
    kmeans = KMeans(n_clusters = philly_clusters, init = 'k-means++', random_state = 0).fit_predict(philly_gc)

    # Gets the score for the clustering operation performed
    score = silhouette_score(philly_gc, kmeans)

    # Appending the index and score to the respective lists
    indices.append(philly_clusters)
    scores.append(score)
```




**Data Visualization**

Various plotting techniques we used as well in order to visualize the data. Visualizing data often gives a clear understanding of the data as it is easier to spot patterns in a visualized data as compares to quantitative data. A couple of libraries were used in this project to visualize data and help analysis. Matplotlib library was used to compares error vs number of clusters to select the best value of k as shown above and Folium library was exploited to plot the geographical map of Philadelphia for demonstration, as shown below, and clustering.

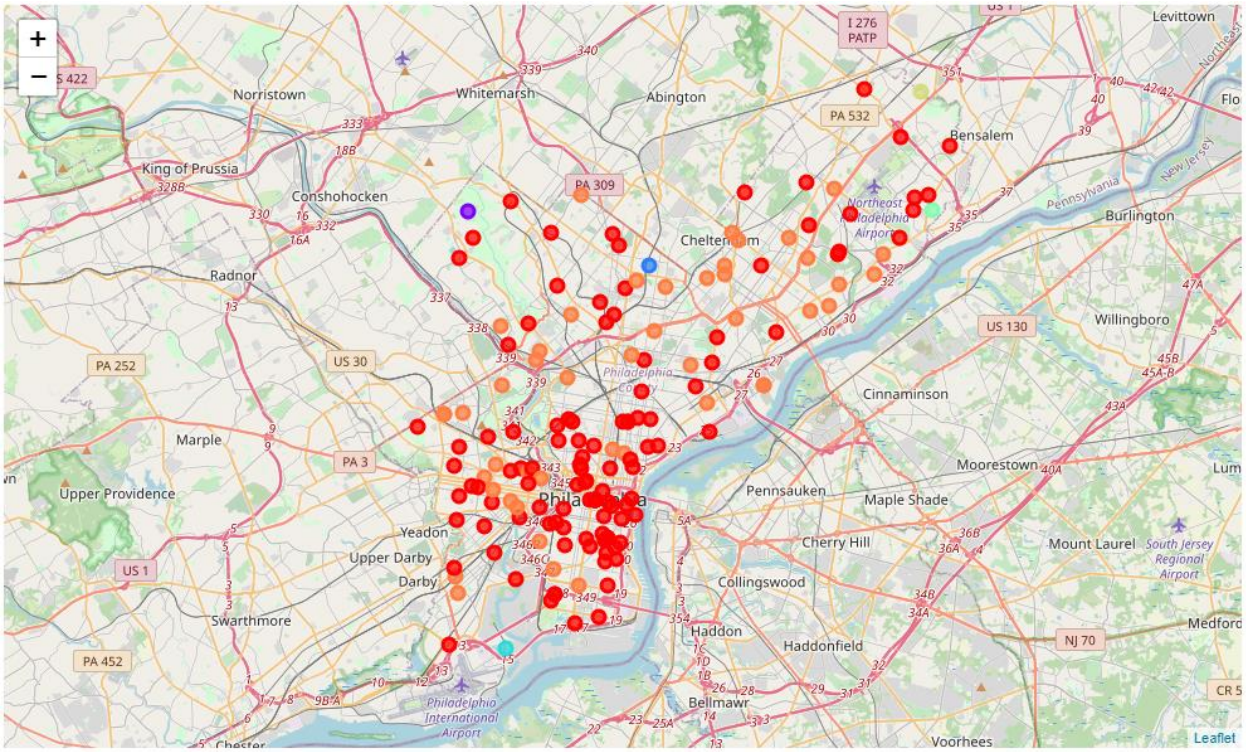


## Results

The above mentioned, k-means clustering method was applied to the dataframe of neighborhoods of Philadelphia. As mentioned earlier the number of clusters that was derived from elbow method was 7. The code as well as plotting of clusters can be seen below:

```
# use the optimum k value from the previous step
opt_value = 7
philly_clusters = opt_value

# Run k-means clustering
philly_gc = philly_grouped_clustering
kmeans = KMeans(n_clusters = philly_clusters, init = 'k-means++', random_state = 0).fit(philly_gc)
```

```
# create map
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=11)

# set color scheme for the clusters
x = np.arange(opt_value)
ys = [i + x + (i*x)**2 for i in range(opt_value)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(philly_merged['Latitude'], philly_merged['Longitude'], philly_merged['Neighborhood'], philly_m
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[int(cluster)-1],
        fill=True,
        fill_color=rainbow[int(cluster)-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```



After visualizing the clusters, the individual clusters were studied and some important conclusions were derived. The neighborhood that had the greatest number of restaurants was cluster number 0 which is indicated by the red dots on the map.

## Discussion

After running the k-means clustering algorithm, all neighborhoods in Philadelphia were group into 7 groups (cluster number 0-6). Among all clusters, neighborhoods in cluster number 0 seems to be the best location for starting a new restaurant as a lot of the neighborhoods in cluster numbers 0 have their most common venue as restaurant. Depending on the cuisine type of the new restaurant, specific neighborhood within in cluster number 0 can be recommended. For example, Bella Vista, East Passyunk Crossing and Southwark could be the potential neighborhoods if our client would like to open a Mexican restaurant. After finding the perfect location for their restaurant, our client can proceed and make a decision depending on other factors like availability and legal requirements that are out of scope of this project.

## Conclusion

To sum up, we extracted and combined data from Wikipedia, Geocoder and FourSquare API to study over 100 neighborhoods in Philadelphia in the United States. Using k-means clustering algorithm, we grouped all neighborhood into 7 clusters. By analyzing detailed information within each cluster, we decided that neighborhoods in cluster number 0 is the most appropriate locations to start a new restaurant.