Analyzing Neighborhoods of Philadelphia For Starting A New Restaurant

TY

12/12/2020

Introduction

- Philadelphia is one of the largest cities in the U.S.
- 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.
- The population comprises of people of various ethnicities from all over the world.

Business Problem

- One of our clients would like to start a new restaurant in Philadelphia.
- We are asked to give suggestion on the location for the business based on data science and machine learning techniques.

Data

The data used in this project includes:

- List of neighborhood scraped from Wikipedia page: https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Philadelphia)
- Geographical coordinates of each neighborhood using Geocoder.
- Venue data of each neighborhood using FourSquare API.

Methodology

- Data scraping using BeautifulSoup library

Scrape the neighborhoods list from wikipedia

In [2]:
```
source = requests.get('https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Philadelphia').text
soup = BeautifulSoup(source, 'lxml')
mwcg = soup.find_all(class_ = "mw-category-group")
```

Save the data in a csv file

In [3]:
```
csv_file = open('philly.csv', 'w')
csv_writer = csv.writer(csv_file)
csv_writer.writerow(['Neighbourhood'])
length = len(mwcg) # Gets the length of number of `mw-category-groups` present

for i in range(1, length):  # Gets all the neighbourhoods
    lists = mwcg [i].find_all('a')
    for list in lists:
        nbd = list.get('title') # Gets the title of the neighbourhood
        csv_writer.writerow([nbd]) # Writes the name of the neighbourhood in the csv file

csv_file.close()
```

import the data from the csv file

In [4]:
```
philly_raw = pd.read_csv('philly.csv',encoding='cp1252')
philly_raw.head()
```

Out[4]:

	Neighbourhood
0	Category:Bridesburg-Kensington-Richmond, Phila...
1	Category:Bridesburg, Philadelphia
2	Category:Callowhill, Philadelphia
3	Category:Chestnut Hill, Philadelphia
4	Category:Chinatown, Philadelphia

Methodology

- Feature Extraction: One Hot Encoding

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

# add neighborhood column back to dataframe
philly_onehot['Neighborhood'] = 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()
```

	Neighborhood	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	Resta
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	

Methodology

- Geographic Visualization: Folium

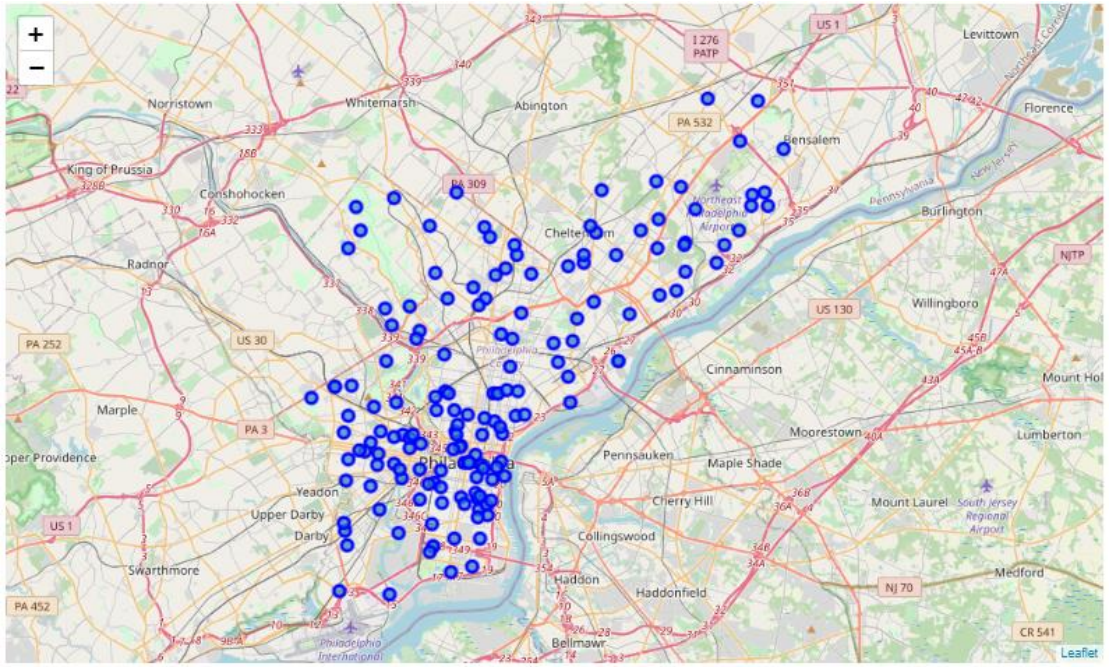
In [14]:
```
#define latitude and longitude of Philadelphia
latitude = 39.952583
longitude = -75.165222
print('The geograpical coordinate of Philadelphia are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Philadelphia are 39.952583, -75.165222.
```

In [15]:
```
# plot the map around Philadelphia and label all neighborhoods as dots on the map using the geographical information
man_map = folium.Map(location=[latitude, longitude], zoom_start=11)

# add markers to map
for lat, lng, label in zip(philly_data['Latitude'], philly_data['Longitude'], philly_data['Neighborhoods']):
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(man_map)

man_map
```



Methodology

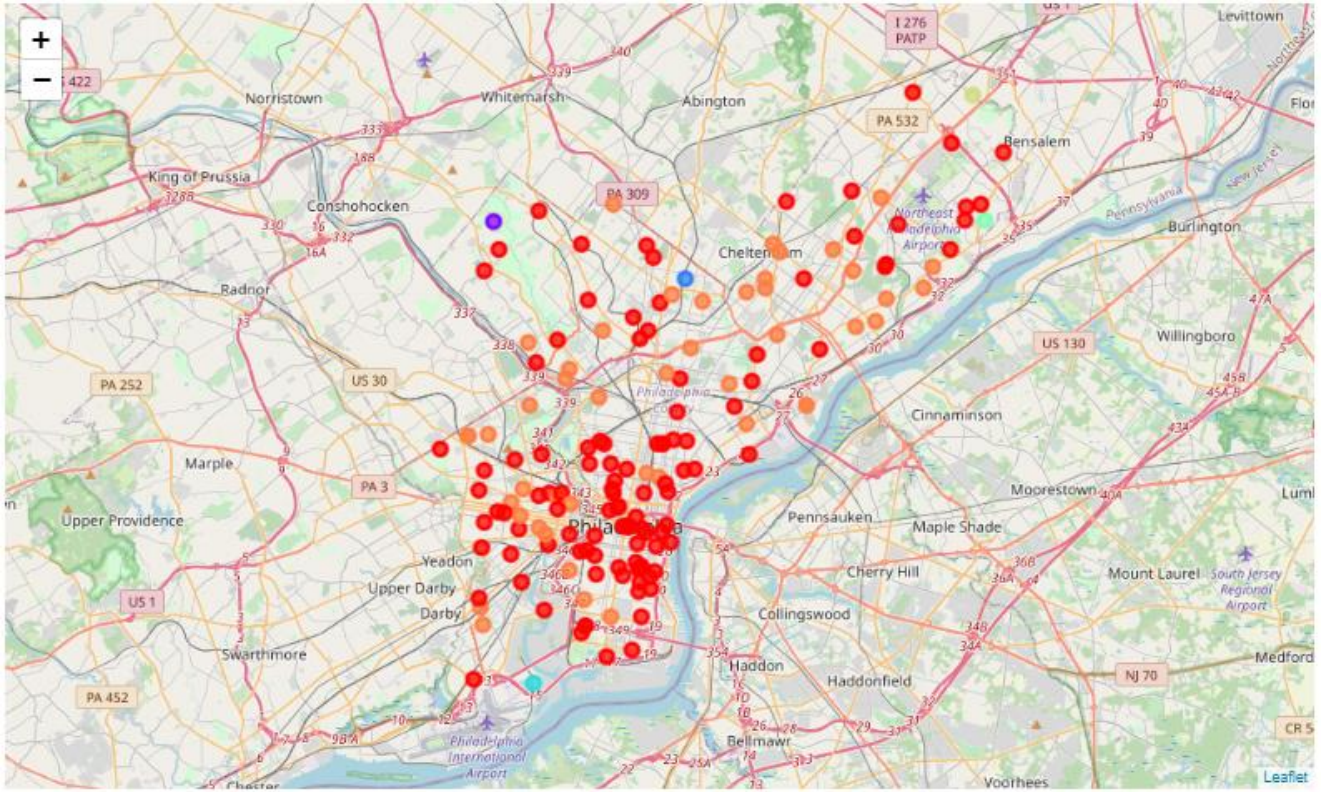
- Unsupervised Learning: K-means Clustering

```
# 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)
```

Results & Discussion

- Based on the results from k-means clustering, the neighborhoods in Philadelphia were divided into 7 clusters:



Results & Discussion

- By looking into detailed venue data within in each cluster, cluster number 0 is the best option for starting a new restaurant.
- Depending on the cuisine style of the restaurant, more specific recommendation on neighborhoods can be given. For example, Bella Vista, East Passyunk Crossing and Southwark could be the potential neighborhoods if our client would like to open a Mexican restaurant.

Conclusion

- 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.
- Based on the cuisine style of the restaurant, specific neighborhoods in cluster number 0 can be selected.