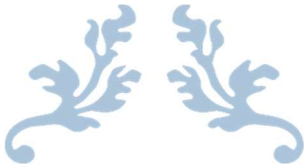


BATTLE OF NEIGHBORHOOD

Capstone Project Report v1.0

DECEMBER 25, 2019

PAUL CHELLADURAI R

Title	Version	Date
Battle of Neighborhood – Capstone Project Report	1.0	25-Dec-2019

Table of Contents

Title	Version	Date
Battle of Neighborhood – Capstone Project Report	1.0	25-Dec-2019

Purpose:

The purpose of this report is to provide the overview of the problem that we are trying to address through this final capstone project, data we are using to analyze and the methodology we follow to uncover the facts. This report also has the final conclusion from the author based on the insights derived from the data.

Problem Background:

Being the land of opportunities, United States of America becomes the world's most welcoming place for work for the tech geeks around the world. People from different culture, ethnics move into USA for work, studies, business etc. The immigration of people goes up and up every year. People who migrate gets professionally settled over a period of time would like move up further to get settled in USA. The general problem for such people is to find a city/state which is financially affordable and with a better living standard. The better place to live is defined by the availability of residents with the same culture background, crime rate, Cost of living index, Poverty level, employment level, per capita income and the ongoing rental rate in the location. So, developing a model to cluster the locations based on these attributes will help the people to select the location.

Description and Usage of data:

As part of this capstone project assignment, I am taking different cities in Montgomery County in Pennsylvania and trying to segment the cities by weighing them in terms of crime rate, rental cost, Cost of living index, Poverty level, Employment rate and per capita income. I am considering one another data to choose the better city is the availability of restaurants with different cuisines. For example, if a location is having multiple Indian Restaurants, then that neighborhood must have more Indian community. It is assumed that the restaurant's owner must have done the analysis on population with different ethnicity before to open the restaurants. These datas are used to analyze and build a model to cluster the cities to decide the better livelihood.

- Crime Rate: Needless to say. Cities with less crime rate will be a better place to live.
- Cost of Living Index: A measure of expenses in the neighborhood
- Poverty Level: This must be direct proportional to Crime Rate. More in Poverty lead to more in crime and turns up less likelihood place to live.
- Employment Level, Per Capita Income: Indirect proportional to crime rate and contributes significantly to decide the better neighborhood
- Ongoing rental rate/Real Estate cost: This is an important factor while choosing a place to live.

Methodology Highlights:

In this analysis, we shall take different cities in Montgomery county in Pennsylvania and will estimate their goodness as livelihood.

1. Import the Montgomery county map

2. Import the different cities in the county from https://data.montgomerycountymd.gov/

3. Mark those cities in the Montgomery map

4. Fetch the availabilities of different restaurants using foursquare API

5. Analyze different cuisines available in different cities to decide different community presence

6. Finally cluster the cities based on the cost of living index, crime rate, employment rate, poverty rate and rental rate

Lets import the geo locations of different cities of Montgomery county into a dataframe and mark these cities in the geospatial map of Montgomery county.

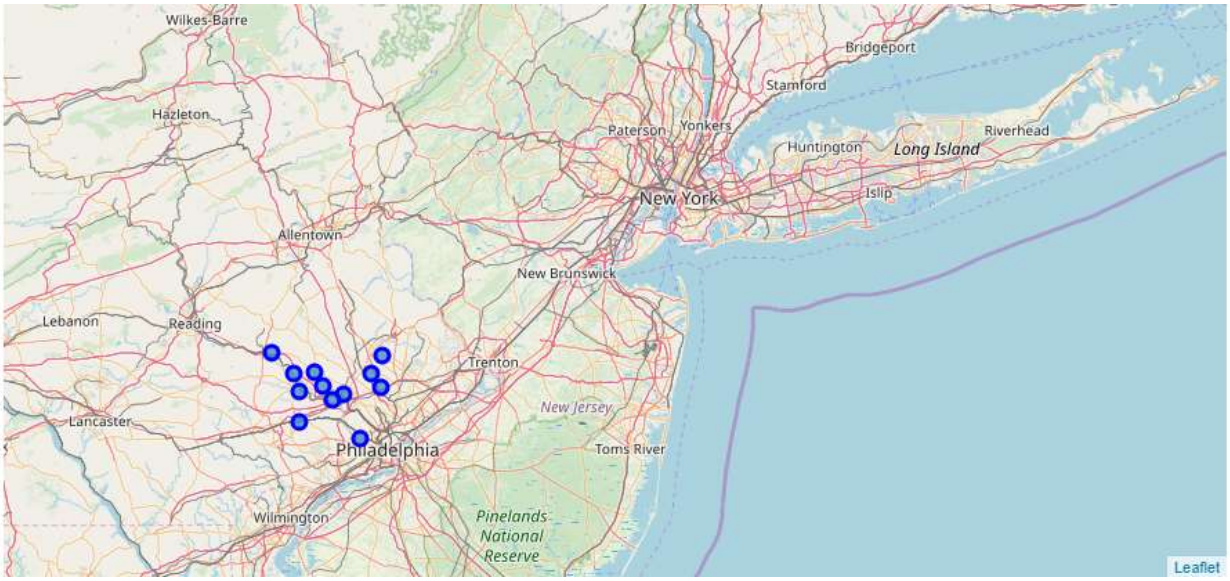
Refer the details code in the notebook:

https://github.com/paulchelladurai/Coursera_Capstone/blob/master/capstone%20project%20final%20(1).ipynb

```
#Import the geo locations of different cities in Montgomery County in the state of Pennsylvania and display it
Montgomerydf=pd.read_csv('pennsylvania.csv')
Montgomerydf
```

	Latitude	Longitude	State	County	City	Name	RegionID
0	40.248003	-75.626843	PA	Montgomery	Pottstown	Washington/Rosedale	761971
1	40.184300	-75.538000	PA	Montgomery	Royersford	East End South	761171
2	40.184365	-75.226319	PA	Montgomery	Lower Gwynedd	Spring House	16311
3	40.143300	-75.422800	PA	Montgomery	Lower Providence	Downtown North	761169
4	40.185700	-75.451600	PA	Montgomery	Collegeville	Beech/Wilson	761168
5	40.130400	-75.514900	PA	Montgomery	Phoenixville	Downtown South	761170
6	40.121500	-75.339900	PA	Montgomery	Norristown	North End	761970
7	40.101300	-75.383600	PA	Montgomery	King Of Prussia	Dresher	38234
8	40.139832	-75.188891	PA	Montgomery	Upper Dublin	Fort Washington	24770
9	39.985409	-75.272983	PA	Montgomery	Wynnewood	Penn Wynne	6460
10	40.036200	-75.513800	PA	Montgomery	Malvern	West End	275966

Geo spatial map:



Let's then fetch all the restaurants available in these cities of Montgomery county using Foursquare API and move them to a dataframe. First few rows look as shown below.

```
Restaurantdf= Montgomery_venues[Montgomery_venues['Venue Category'].str.contains('Restaurant')]
print('There are {} venus for the food (holding Restaurants in their name)'.format(Restaurantdf.shape[0]))
Restaurantdf.head(10)
```

```
There are 224 venus for the food (holding Restaurants in their name)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Pottstown	40.248003	-75.626843	McDonald's	40.242988	-75.619734	Fast Food Restaurant
3	Pottstown	40.248003	-75.626843	McDonald's	40.252671	-75.659760	Fast Food Restaurant
4	Pottstown	40.248003	-75.626843	Wendy's	40.253671	-75.660442	Fast Food Restaurant
5	Pottstown	40.248003	-75.626843	Arby's	40.254643	-75.661653	Fast Food Restaurant
6	Pottstown	40.248003	-75.626843	Burger King	40.265440	-75.628389	Fast Food Restaurant
11	Pottstown	40.248003	-75.626843	TGI Fridays	40.234440	-75.661947	American Restaurant
12	Pottstown	40.248003	-75.626843	Chili's Grill & Bar	40.265175	-75.650780	Tex-Mex Restaurant
13	Pottstown	40.248003	-75.626843	Boston Market	40.254724	-75.662670	American Restaurant
15	Pottstown	40.248003	-75.626843	Friendly's	40.254824	-75.659765	Restaurant
16	Pottstown	40.248003	-75.626843	Red Lobster	40.254113	-75.660487	Seafood Restaurant

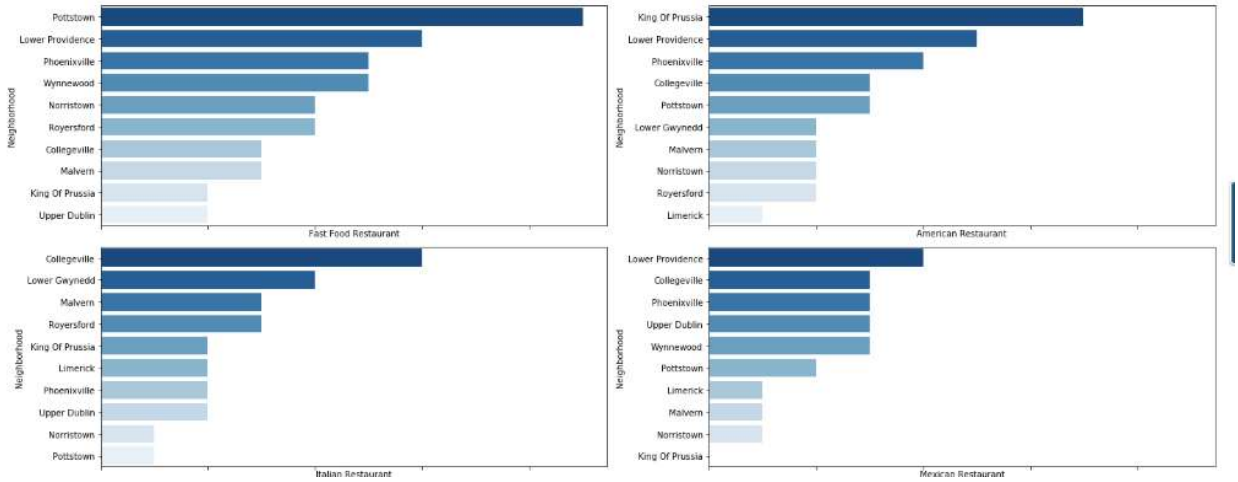
The top 10 restaurants from all these cities and their data description is as looks below:

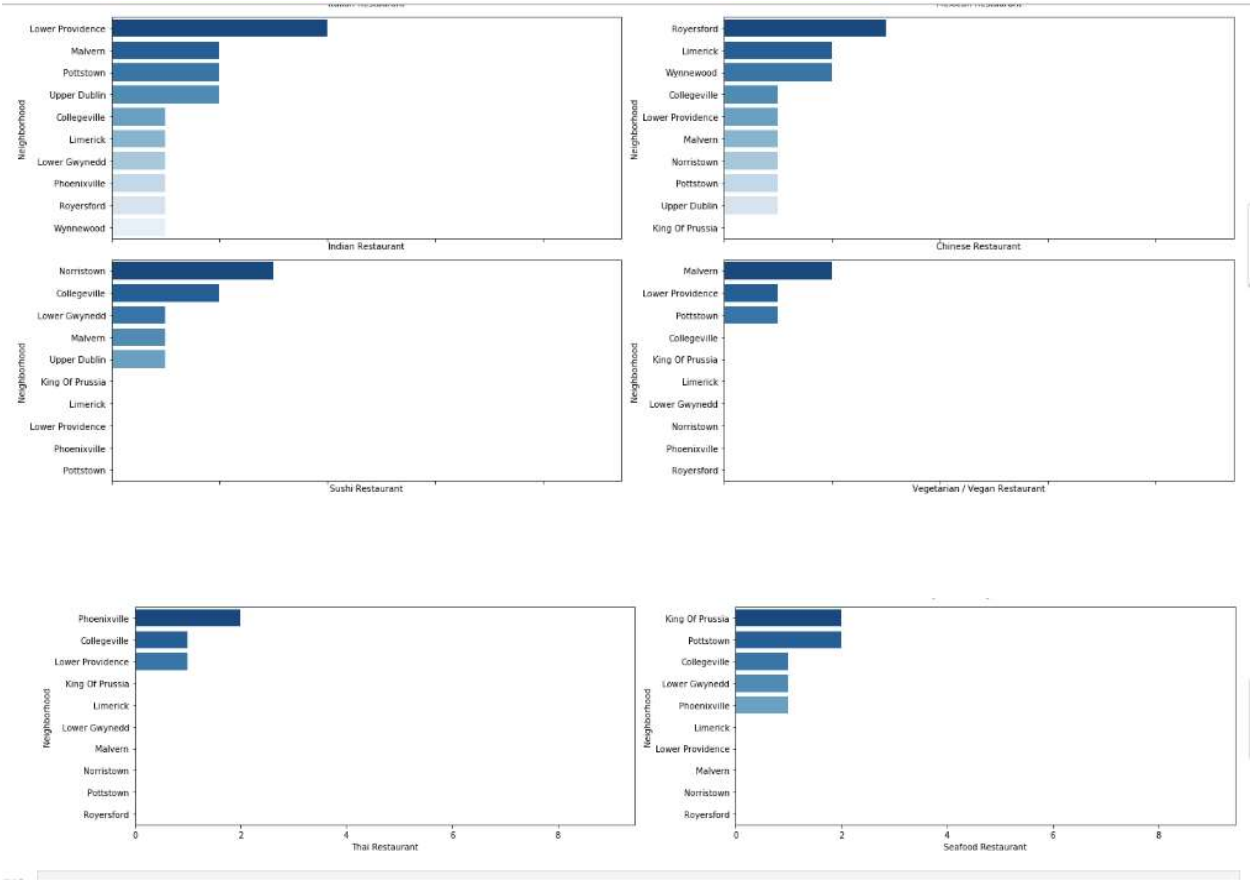
```
venue_counts_described = venue_counts.describe().transpose()
venue_top10 = venue_counts_described.sort_values('max', ascending=False)[0:10]
venue_top10
```

	count	mean	std	min	25%	50%	75%	max
Fast Food Restaurant	12.0	3.750000	2.301185	1.0	2.00	3.5	5.00	9.0
American Restaurant	12.0	2.750000	1.815339	1.0	1.75	2.0	3.25	7.0
Italian Restaurant	12.0	2.250000	1.602555	0.0	1.00	2.0	3.00	6.0
Mexican Restaurant	12.0	1.750000	1.422226	0.0	0.75	1.5	3.00	4.0
Indian Restaurant	12.0	1.333333	1.073087	0.0	1.00	1.0	2.00	4.0
Chinese Restaurant	12.0	1.083333	0.900337	0.0	0.75	1.0	1.25	3.0
Sushi Restaurant	12.0	0.666667	0.984732	0.0	0.00	0.0	1.00	3.0
Vegetarian / Vegan Restaurant	12.0	0.333333	0.651339	0.0	0.00	0.0	0.25	2.0
Thai Restaurant	12.0	0.333333	0.651339	0.0	0.00	0.0	0.25	2.0
Seafood Restaurant	12.0	0.583333	0.792961	0.0	0.00	0.0	1.00	2.0

Based on our analysis, Fast Food restaurant are the most followed by American Restaurants. Over all the top 10 food categories gives us an idea that Montgomery county has the good mix of food ethnicity that includes American. Italian, Mexican, and other Asian countries

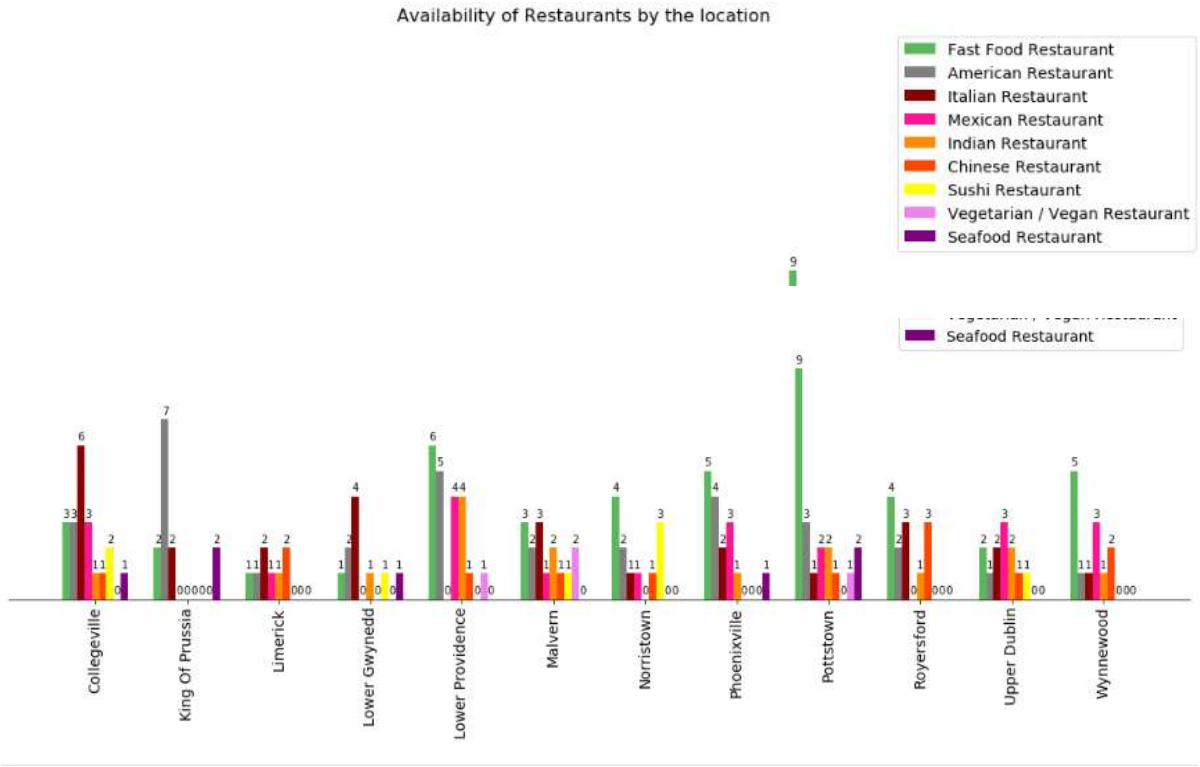
Now I'm trying to put the availability of different cuisines in the different locations...Let's see how the Horizontal bar chart comes in and then let's explore it in group chart




Lower Providence
Malvern
Pottstown
Upper Dublin
Collegeville
Limerick
Lower Gwynedd
Phoenixville
Royersford
Wynnewood
Neighborhood
Indian Restaurant
Royersford
Limerick
Wynnewood
Collegeville
Lower Providence
Malvern
Norristown
Pottstown
Upper Dublin
King Of Prussia
Neighborhood
Chinese Restaurant
Norristown
Collegeville
Lower Gwynedd
Malvern
Upper Dublin
King Of Prussia
Limerick
Lower Providence
Phoenixville
Pottstown
Neighborhood
Sushi Restaurant
Malvern
Lower Providence
Pottstown
Collegeville
King Of Prussia
Limerick
Lower Gwynedd
Norristown
Phoenixville
Royersford
Neighborhood
Vegetarian / Vegan Restaurant
Phoenixville
Collegeville
Lower Providence
King Of Prussia
Limerick
Lower Gwynedd
Malvern
Norristown
Pottstown
Royersford
Neighborhood
0
2
4
6
8
Thai Restaurant
King Of Prussia
Pottstown
Collegeville
Lower Gwynedd
Phoenixville
Limerick
Lower Providence
Malvern
Norristown
Royersford
Neighborhood
0
2
4
6
8
Seafood Restaurant

Title	Version	Date
Battle of Neighborhood – Capstone Project Report	1.0	25-Dec-2019

Group Chart:



Fast food restaurants, American Restaurants, Italian and Mexican Restaurants are common across locations. Indian, Chines Restaurants are more in some locations.

Clustering and Segmenting the Cities:

Now lets cluster and segment the Cities by Cost of Living, Crime Rate, employment rate, Poverty rate and Rental rate into 2 groups.

```
# Modules
import matplotlib.pyplot as plt
from matplotlib.image import imread
from sklearn.datasets.samples_generator import (make_blobs,
                                                make_circles,
                                                make_moons)
from sklearn.cluster import KMeans, SpectralClustering
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_samples, silhouette_score

%matplotlib inline
sns.set_context('notebook')
plt.style.use('fivethirtyeight')
from warnings import filterwarnings
filterwarnings('ignore')

pdf=pd.read_csv('pennsylvania_Metadata.csv')

Citydf=pdf['City']
Montgomerypdf=pdf.drop('City',axis=1)

from sklearn.cluster import KMeans
# Standardize the data
X_std = StandardScaler().fit_transform(Montgomerypdf)

# Run local implementation of kmeans
km = KMeans(init="k-means++",n_clusters=2, n_init=50).fit(X_std)
print('The labels out of Kmeans: ', km.labels_)
centroids=km.cluster_centers_
print('The centroids that come out of clustering: ',centroids)

#Insert the cluster label in the parent dataframe
try:
    pdf.drop('Cluster Labels', axis=1)
except:
    pdf.insert(0, 'Cluster Labels', km.labels_)
# Merge the dataframes that holdes the clustered data and the geospatial data of cities in the montgomery county
finalpdf=pd.merge(pdf, Montgomerydf, on='City')
print('\n\n\nFinal dataframe with the clustered data (refer cluster lable column) and Geospatial data of cities looks as follows \n')
finalpdf
```

The cluster outputs attributes are as follows:

```
The labels out of Kmeans:  [1 1 0 0 0 0 1 0 0 0 0 1]
The centroids that come out of clustering:  [[ 0.57767656  0.14431655 -0.55347007  0.57942639 -0.585853    0.43813729]
 [-1.15535313 -0.2886331   1.10694013 -1.15885278  1.171706   -0.87627458]]
```

The Geospatial map with the clustered cities marked:

Lets see the cities in Cluster – 0

```
cluster_0 = finalpdf.loc[finalpdf['Cluster Labels'] == 0, finalpdf.columns[1:12]]
cluster_0
```

	City	Income per capita	Unemployment rate	Poverty level	Cost of living index	Crime per 100k peopl	Mean rental rate	Latitude	Longitude	State	County
2	Lower Gwynedd	35755	0.05	0.10	111	832	1350	40.184365	-75.226319	PA	Montgomery
3	Lower Providence	43387	0.02	0.05	131	688	1200	40.143300	-75.422800	PA	Montgomery
4	Collegeville	33510	0.03	0.02	124	966	1300	40.185700	-75.451600	PA	Montgomery
5	Phoenixville	32881	0.05	0.09	113	1317	1300	40.130400	-75.514900	PA	Montgomery
7	King Of Prussia	44934	0.04	0.07	124	800	1600	40.101300	-75.383600	PA	Montgomery
8	Upper Dublin	45745	0.03	0.01	135	638	1400	40.139832	-75.188891	PA	Montgomery
9	Wynnewood	54087	0.03	0.03	137	645	1400	39.985409	-75.272983	PA	Montgomery
10	Malvern	48086	3.50	0.11	132	684	1700	40.036200	-75.513800	PA	Montgomery

Lets see the cities in Cluster – 1

```
cluster_1 = finalpdf.loc[finalpdf['Cluster Labels'] == 1, finalpdf.columns[1:12]]
cluster_1
```

	City	Income per capita	Unemployment rate	Poverty level	Cost of living index	Crime per 100k peopl	Mean rental rate	Latitude	Longitude	State	County
0	Pottstown	23346	0.06	0.22	94	4325	1050	40.248003	-75.626843	PA	Montgomery
1	Royersford	28169	0.06	0.13	110	2184	1300	40.184300	-75.538000	PA	Montgomery
6	Norristown	21986	0.07	0.22	103	2214	900	40.121500	-75.339900	PA	Montgomery
11	Limerick	24380	0.03	0.11	96	1917	1300	40.238400	-75.184329	PA	Montgomery

The final dataframe with the cluster labels and geospatial data

]:

	Cluster Labels	City	Income per capita	Unemployment rate	Poverty level	Cost of living index	Crime per 100k peopl	Mean rental rate	Latitude	Longitude	State	County	Name	RegionID
0	1	Pottstown	23346	0.06	0.22	94	4325	1050	40.248003	-75.626843	PA	Montgomery	Washington/Rosedale	761971
1	1	Royersford	28169	0.06	0.13	110	2184	1300	40.184300	-75.538000	PA	Montgomery	East End South	761171
2	0	Lower Gwynedd	35755	0.05	0.10	111	832	1350	40.184365	-75.226319	PA	Montgomery	Spring House	16311
3	0	Lower Providence	43387	0.02	0.05	131	688	1200	40.143300	-75.422800	PA	Montgomery	Downtown North	761169
4	0	Collegeville	33510	0.03	0.02	124	966	1300	40.185700	-75.451600	PA	Montgomery	Beech/Wilson	761168
5	0	Phoenixville	32881	0.05	0.09	113	1317	1300	40.130400	-75.514900	PA	Montgomery	Downtown South	761170
6	1	Norristown	21986	0.07	0.22	103	2214	900	40.121500	-75.339900	PA	Montgomery	North End	761970
7	0	King Of Prussia	44934	0.04	0.07	124	800	1600	40.101300	-75.383600	PA	Montgomery	Dresher	38234
8	0	Upper Dublin	45745	0.03	0.01	135	638	1400	40.139832	-75.188891	PA	Montgomery	Fort Washington	24770
9	0	Wynnewood	54087	0.03	0.03	137	645	1400	39.985409	-75.272983	PA	Montgomery	Penn Wynne	6460
10	0	Malvern	48086	3.50	0.11	132	684	1700	40.036200	-75.513800	PA	Montgomery	West End	275966
11	1	Limerick	24380	0.03	0.11	96	1917	1300	40.238400	-75.184329	PA	Montgomery	Maple Glen	19189

Discussion and insights of Analysis:

By looking at the group chart by restaurants, we can conclude that Fast food restaurants, American Restaurants, Italian and Mexican Restaurants are wide spread across all the cities which is as expected. The Indian Restaurants are more in Lower Providence city and Sushi Restaurants are more in Collegeville.

The cities Pottstown, Royersford, Norristown and Limerick are clustered into one group. The other cities are clustered into group 0. Again, we can observe that the cities in Cluster-1 has more crime rate and less per capita income. we can see the uniqueness. On the other end, the cities in Cluster-0 has less crime rate and the other factors are good compared to cities in cluster-1.

Conclusion:

Lets assume a Indian looking for the better city to settle, he or she will prefer to choose Lower providence since it is in cluster-0 with less crime rate and other better factors. The city Lower Providence also has more Indian Restaurants and as per our initial assumption, more Indian Restaurant means more Indians in the city.

In case if a real estate house looks for a new location to start a housing project, it will choose a city from Cluster-0.

The same analysis can further be enhanced to find better counties in the state or the better state in the country by following the same approach.

End User/Beneficiary:

This analysis will be beneficial for anyone who wish to choose a location to buy house/get settled. This model/analysis can also be used by real estate business house' to understand the customer preferences on locations.

Data References:

1. https://data.opendatasoft.com/
2. https://foursquare.com/
3. https://data.montgomerycountymd.gov/
4. Google Geo Locator