**Coursera Capstone Project**

Choose an area to live in Philadelphia by exploring all its neighborhoods

By Linh Tran



**Introduction:**

My sister just landed her first full time job in Philadelphia, one of the biggest cities in Pennsylvania. She has never been to Philadelphia before so she is quite new to this area. She wants to find housing in Philadelphia area but does not know much about the city and housing price for each neighborhood in Philly. She wants to live in a nice neighborhood, preferably around city center, and where there’s high population and affordable housing price. She can afford higher rent since she care more about the neighborhood itself. I will help her explore every neighborhood in Philly so that she can have a broad over view of each neighborhood here and make her own decision on where to live. Thus, this project is about clustering neighborhoods in Philadelphia so that we can get a general view of what each neighborhood is known for while consider other demographic data for each neighborhood such as: population density, housing units, average home value, average household income

**Business Problem:** Where to live in Philadelphia?

**Goals/Objectives:** The objective of this capstone project is to analyze Philadelphia’s neighborhoods dataset and select the best location for a housing. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: where to live in Philadelphia. In order to answer this, we need to understand the characteristics of each neighborhood.

**Stakeholders**

This project is helpful to anybody who wants to know more about Philadelphia, not limited to my sister only. Since many people might want to understand in and out about each neighborhood in Philadelphia, this project will help them a lot as it will segment all neighborhoods in the city and show the most frequent venue categories, population density, home value (similar to housing price), housing units and average household income in each neighborhood.

**Data:**

**Sources of data and how it is used:**

Philadelphia’s neighborhood dataset including names of boroughs and their corresponding neighborhoods. The table that contain needed data is located at the bottom of this wiki website:

<https://en.wikipedia.org/wiki/Callowhill,_Philadelphia>. This data defines the scope of this project which is confined to the city of Philadelphia.

Philadelphia’s demographic data: population density, home value, housing units and average household income for each neighborhood.

Coordinates (latitude, longitude) of each neighborhood. This data is needed to generate maps, get demographic data and get venues data using Foursquare APIs.

Foursquare APIs: that contains data about all venues data around Philadelphia

**Methods to extract Data:**

Philadelphia’s neighborhood dataset on Wiki: use Beautiful Soup package to scrape data from Wiki website.

Coordinates for each neighborhood: use geocoder package to get coordinates(latitude, longitude). In order to get coordinates for each neighborhood, address, in other words, name of neighborhood is needed

Philadelphia’s demographic data: Use uszipcode package and SearchEngine library to get demographic data (population density, average household income) for each **zipcode**. 1 zipcode can contain multiple neighborhoods; thus, we will define the boundaries of each neighborhood inside zipcode and collect the most relevant data.

Foursquare APIs to get venues data: create API url request and get request to get all venues data.

**Methodology**

1. **Get neighborhoods, coordinates and demographic data and transform into dataframe**

Firstly, we need to get the list of neighborhoods in the city of Philadelphia. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Callowhill,\_Philadelphia.).We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API as well as get demographic data. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. Next, we will use the latitude and longitude of each neighborhood to get demographic data for each neighborhood. By using uszipcode package with SearchEngine library applying on the previous coordinates data, we can gather demographic data for each **zipcode**, but not for each neighborhood. Thus, we need to find the boundary of each neighborhood within each zipcode to get the most accurate data to analyze. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Philadelphia.

1. **Explore each neighborhood using Foursquare APIs**

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 10 clusters. The results will allow us to identify popular venue categories within each cluster and we can combine with the demographic data to get a better view of each neighborhood cluster. For example, for each cluster, we can look further at the average household income and average home value to identify if the neighborhood cluster is affluent neighborhoods, or we can look more at population and housing units data to estimate the crowdedness of each cluster/neighborhood.

**Results:**

1. **Cluster Modeling**

Scikit-learn’s K-Means clustering was used to determine similar neighborhoods based on music venue percentage. The image below shows the data being scaled and the K-Means model being created:

A screenshot of a social media post

Description automatically generated

1. **Cluster Visualization**

A picture containing text, map

Description automatically generated

1. **Cluster Evaluation:**

Using K-means algorithm, we cluster the neighborhoods into 10 clusters. Each cluster shows a list of neighborhoods with their respective top venue categories and demographic data.

It is interesting to see that some clusters are very small, sometimes only holding a single neighborhood, and appear to have identified a niche venue category. Examples of this is cluster 3: This cluster only has one neighborhood named “Andorra” with top venue category “Playground”.

A screenshot of a cell phone

Description automatically generated

Other clusters are very large and appear to be grouping neighborhoods with assortments of restaurants, coffeeshops, flea markets, etc.

**Overview of each cluster:**

Cluster 1: South Philadelphia area – Restaurant/Bar/Intersection

A screenshot of a cell phone

Description automatically generated

Cluster 2: Multiple neighborhoods around the city – Coffeeshop, Bakery, Hotel, Park, Bar, etc.

A screenshot of a cell phone

Description automatically generated

Cluster 3: Andorra – Playground, Exhibit, Eastern European Restaurant

A screenshot of a cell phone

Description automatically generated

Cluster 4: Morrell Park and Crestmont Farms – Café, Exhibit, Ethiopian Restaurant

A screenshot of a cell phone

Description automatically generated

Cluster 5: Passyunk Square and Wsest Passyunk - Flea Market

A screenshot of a cell phone

Description automatically generated

Cluster 6: Multiple neighborhoods around the city – Restaurants, Bars, Pubs, Coffeeshops, Store, etc.

A screenshot of a cell phone

Description automatically generated

Cluster 7: Academy Gardens and Ashton Wooden Bridge - Garden, Exhibit, Ethiopian Restaurant

A screenshot of a cell phone

Description automatically generated

Cluster 8: Several neighborhoods (5) scattered in the city – Intersection, Park, Fast food, Gym

A screenshot of a cell phone

Description automatically generated

Cluster 9: East Oak Lane – Lake, Exhibit, Electronic Store

A screenshot of a cell phone

Description automatically generated

Cluster 10: Hog Island – Airport, Exhibit, Field

A screenshot of a cell phone

Description automatically generated

**Discussions:**

As seen in the above clusters, cluster 2 and cluster 6 seem to be ideal clusters to start digging more into because these two clusters consist of many neighborhoods located in the central of Philadelphia with diverse venue categories such as coffeeshops, stores, bakeries, park, restaurants, etc. As my sister like a populated, high-class neighborhood, I sort each neighborhood with highest population and household income and finally come up with this list of neighborhoods that she might be interested in:

|  |  |  |  |
| --- | --- | --- | --- |
| **Neighborhoods** | **Average Income** | **Population density**  **(number of people/mi2)** | **Top Venue Categories** |
| 30th Street Station | 63709 | 34284 | Sandwich, Food truck, Café, Pub, Train Station |
| Rittenhouse Square/ Filter Square | 52888 | 27533 | Café, American and Italian restaurant, Hotel, yoga, Hotel |
| Callowhill/ Belmont Village | 44402 | 27085 | Beer, Pub, Chinese restaurant, Bakery |
| Bella Vista/ Italian Market | 60400 | 25741 | Mexican, Italian, Vietnamese, French restaurant, Pizza place and Coffeeshop, Bakery, Pharmacy |

**Conclusions:**

Machine learning and clustering algorithms can be applied to multi-dimensional datasets to find similarities and patterns in the data. Clusters of neighborhoods can be generated using high-quality venue location data, and for instance, in this project I am using Foursquare API data. There is a preface on high-quality because analysis models are only as good as the input into them (garbage in, garbage out). Luckily, Foursquare offers a robust ‘Places API’ service that, although (as we have seen) not perfect (nothing is), can be leverages in similar studies and model-making.

In addition to the algorithms and data science techniques introduced in the course, in order to finish the project, I needed to do a lot of side research and utilize many different python packages to solve the problem or get the data that I want. It’s definitely fun to apply all the knowledge I learned throughout the course and research ability into this project and finally come up with the final deliverable that might be impactful for some people. I’m looking forward to do more data science projects so as to sharpen my skills and get more experience in data analysis.