**Recommendation Engine**

**Philadelphia Neighborhoods for NYC Commuters**

Kelly Justice

May 2020

**Introduction**

The New York City real estate is one of the most expensive real estate markets in the U.S. and around the globe. Due to the high costs of living, many people that work in NYC would prefer to commute and own a home outside of the city. However, these same people also want the benefits that come along with big city life.

Philadelphia has much of the same benefits of NYC with a diverse community, major league sports, large population, similar climate, large cultural events and locations, and a significant number of highly-rated restaurants and other establishments. In addition, commuters can get trains directly from Philadelphia to downtown NYC in about an hour to an hour and half. Yet, property prices in Philadelphia per square foot are significantly lower than those in NYC (by a factor of 4).

Therefore, over the past several years, Philadelphia has seen an influx of people from NYC. Especially people with families that are not looking to spend their entire salary on a small place in the city and want some of the amenities that big cities have to offer.

**Problem Statement :** How does someone from NYC know which neighborhoods are similar to the ones they like in NYC? By creating a clustering model of both NYC and Philly neighborhoods, we should be able to provide them with like neighborhoods. This should help to narrow or simplify their search for a place in Philadelphia that is similar to those they know in NYC in terms of demographics and the variety and quantify of establishments within major categories like restaurants, bars, religious, financials, and others.

**Data**

The data required to create this analysis is drawn from the following sources:

* Opendatasoft API : provides the list of neighborhoods, geographic coordinates, and demographic data from Zillow
* Google Maps API : provides postal codes for the neighborhoods based on the latitude and longitude
* Foursquare API : provides the venue data based on latitudes and longitudes of each neighborhood

The data is brought into 3 separate DataFrames, transformed, and linked and then combined at various steps in the analysis by neighborhood. The DataFrames are listed below:

* Neighborhoods : Contains a list of the neighborhoods for both cities with latitude, longitude, and postal code
* Demographics : Contains the list of neighborhoods with Median Age, Income per Household, and racial diversity data
* Places : Contains a list the neighborhoods and the places, categories, major categories, and quantity

The common key to all of the tables is the name of the neighborhood which links the tables together

**Methodology & Analysis Section**

The neighborhoods recommendation engine uses a k means clustering model to find similar neighborhoods based upon various demographic and venue data.

The first step was to connect to external APIs to get a list of neighborhoods in the 2 cities under review (New York and Philadelphia) including a list of latitudes and longitudes so that we could map the neighborhoods and determine a list of places/venues within a walking distance of the neighborhood center. The venue data was summarized by neighborhood to get counts (totals) for each major category of venue.

The second step was to connect the postal codes tied to the neighborhoods so that we could create a link to demographic data as this data was only available at the postal code level. Demographic data include income, household, and racial diversity data.

The data was collected through various APIs (listed in the Data section) and loaded into 3 separate data frames for review.

Finally, the data was consolidated into a final table and transformed to normalize the values for processing by the kmeans clustering model.

**Analysis**

In order to simplify the analysis, the categories supplied by Foursquare were transformed from over 400 to subset of 15 major categories. A review of the instances of those categories can be seen in the chart below.

A close up of text on a white background

Description automatically generated

The overwhelming majority of the locations data provided by Foursquare are for restaurants followed by shopping places. Noticeably not as present are locations for education (schools), financial (banks, check cashing, etc), lodging (hotels, motels, etc.) and religious organizations. Knowing that these are large metropolitan areas, it is somewhat surprising to see so few of these. This could be a result of the source of the Foursquare data.

In reviewing the demographics data, there were few surprises in that the quantities and ranges of the data appear to be in line with expectations for those areas. Histograms on this data can be seen below.

A picture containing screenshot, cabinet

Description automatically generated

**Results**

The results of running the kMeans clustering model were a group of 6 neighborhood clusters aligned across the demographics + places data. The model performed with an accuracy score (inertia score) of 24.6.

A summary table of the results by clusters is below. The clusters clearly show grouping along the various features especially around Median Household Income, racial diversity, and locations within walking distance.

A screenshot of a cell phone

Description automatically generated

The neighborhood clusters were also mapped to show geo location for both cities

**Cluster Legend : 0 1 2 3 4 5**

A close up of a map

Description automatically generated

**Discussion Section**

The results of the analysis show a similarities in the types of most neighborhoods of Philadelphia and New York that should allow the user to leverage the model to narrow down the selection process when looking to relocate between the two cities.

Below is a list of the clusters with some of the main differences highlighted.

**Cluster 0** : **2nd highest Median Age, 2nd highest average household income**, below average household size, average house value, and **the least racially diverse cluster with 77% white**, below average number of locations within walking distance

**Cluster 1** : **2nd lowest Median age**, average income levels, household size, and home values, **racially diverse**, average number of places within walking distance

**Cluster 2** : Average age, **2nd lowest average income, average home size, low median home value, low racial diversity , heavily African American, lowest total number of locations within walking distance**

**Cluster 3** : Average age, **highest avg income, lowest household size, highest median home value, second least racially diverse but 16% Asian population, highest (4X others) average locations especially bars, restaurants, etc.**

**Cluster 4** : **Youngest Median Age, lowest household income, very diverse population, largest household size**, average number of locations

**Cluster 5 :** **Oldest Median age**, average household income, **high household size and home value, largest Asian population percentage but racially diverse**, second highest number of places within walking distance especially restaurants. **Especially of note in this cluster is that they are all NYC neighborhoods with no matching Philadelphia neighborhoods.**

**Conclusion Section**

The purpose of this analysis was to build a model that could recommend neighborhoods for relocation to people in NYC looking to relocate to Philadelphia and commute to NYC. The premise of the model would be to find similar neighborhoods within the 2 cities based on demographic data and places of interest. This recommendation engine would allow the users to find neighborhoods in the Philadelphia area that were similar in makeup to those that they were familiar with in New York and help to speed the research process.

The model shows that it is possible to create a model based on these parameters. Similar features were found in online models by real estate agents but not ones where the customer could select a neighborhood in their city that they were familiar with and then have the system provide similar neighborhoods in other cities.

This model could be refined further to allow the customer to weight the importance of the factors making the model more flexible and make the decision and research process simpler for the end users.