**Capstone Project**

**Finding a Philadelphia Neighborhood Most Similar to Almagro, Madrid**

**Introduction**

The project objective is to find a neighborhood in Philadelphia, PA, USA, that is most similar to Almagro, Madrid, but the program is applicable for finding a neighborhood in any city that is similar to some other target neighborhood. The algorithm could be used by people who are planning to move to a new city and want to find a neighborhood that is most similar to a known neighborhood that they find desirable. The algorithm also incorporates real estate price data so that the user can compare similarity between neighborhoods and median housing prices to find a most optimal match for his/her particular situation.

**Data Sources**

### The following data sources will be utilized in the algorithm:

### GPS data from Nominatim: [www.openstreetmap.org](http://www.openstreetmap.org/)

### The Foursquare Places API for exploring neighborhood venues <https://developer.foursquare.com/>

### A list of Philadelphia neighborhoods and housing price statistics published by Philadelphia magazine at <https://www.phillymag.com/property/house-prices-philadelphia-suburbs/#philly>

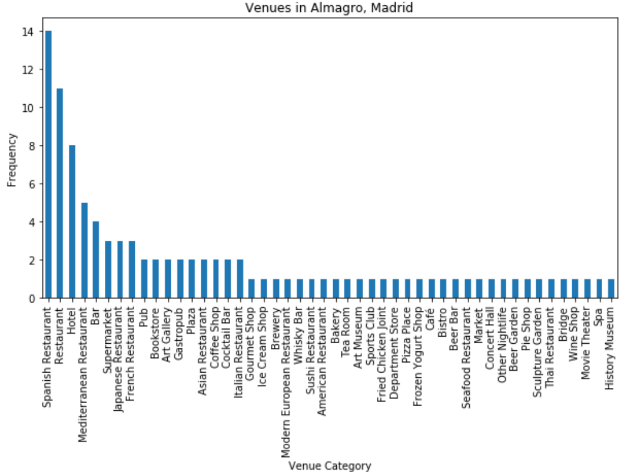
**Methodology**

The target neighborhood, Almagro, Madrid will be characterized through an analysis of nearby venues utilizing the Foursquare Places API. Philadelphia neighborhoods will then be characterized in the same manner, and k-means clustering will be utilized to group Philadelphia neighborhoods by similarity in venue types. Almagro, Madrid will be included in the clustering analysis to identify the cluster of Philadelphia neighborhoods that is most similar to it.

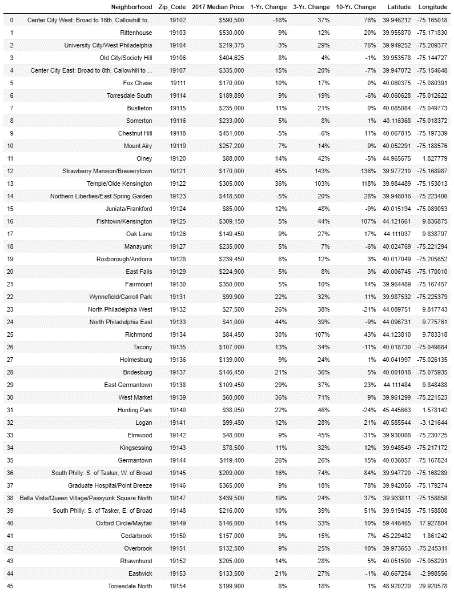
To further identify the Philadelphia neighborhoods that are most similar to Almagro, exact Euclidian distance from Almagro will be calculated in the multidimensional venue space. The neighborhoods will then be plotted in a scatter plot against median housing prices to provide an assessment of affordability along with similarity to Almagro. The ideal neighborhood will be similar to Almagro, but also affordable for the user of the algorithm.

**Analysis, Results, and Discussion**

The Foursquare Places API was utilized to characterize the target neighborhood, Almagro, Madrid by its categories and quantities of venues. As can be seen in the following bar chart, the Almagro neighborhood has a high concentration and wide variety of restaurants. Spanish and Mediterranean restaurants are most common. There are also a good number of bars/pubs and hotels, and some art galleries.



A list of Philadelphia neighborhoods and housing price data was scraped from an article published by Philadelphia magazine at <https://www.phillymag.com/property/house-prices-philadelphia-suburbs/#philly>. After reformatting the column data, the neighborhood names were fed into the Nominatim database to collect GPS coordinates. The resulting dataframe is shown below.



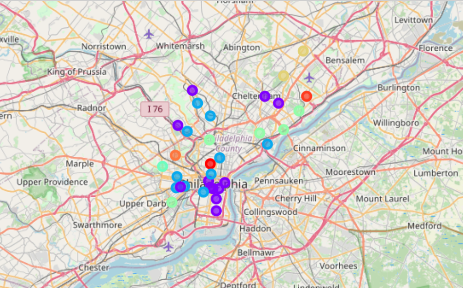
The GPS data were used to create a map of the Philadelphia neighborhoods to visualize their locations with respect to each other

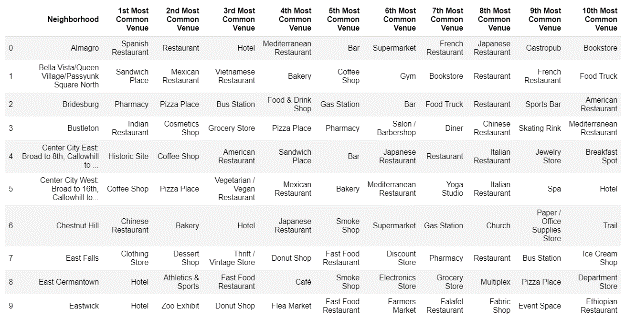


The GPS data were then run through the Foursquare places API to generate a list of venues in each neighborhood. The resulting json file was converted to a dataframe for subsequent one-hot encoding and analysis by k-means clustering. Prior to clustering, the venue data for Almagro, Madrid was added to the Philadelphia dataframe to identify the cluster into which it would fall. This would identify the cluster of most similar neighborhoods to Almagro.

K-means clustering showed that Almagro fell into cluster number 1, along with Center City West, Rittenhouse Square, Old City/Society Hill, and nine other Philadelphia neighborhoods, which are visualized by the purple markers in the following map. These neighborhoods had a high number and wide variety of restaurants, as well as hotels.

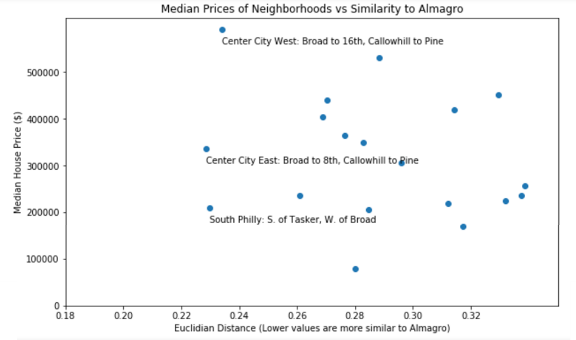
Two other clusters were identified in the Philadelphia neighborhoods. More casual venues like pizza resaurants and sandwich places were more popular in Cluster 4, and Cluster 8 included neighborhoods with fewer sit-down restaurants and more fast food venues.





Euclidian distance from Almagro was then calculated in the multidimensional venue space to identify the neighborhoods that were most similar to Almagro. Finally, the housing price statistics were incorporated in the analysis to identify neighborhoods that might be more affordable for the user, yet still relatively similar to the target neighborhood.

When applied for finding a neighborhood in the city of Philadelphia, Pennsylvania, USA that is most similar to Almagro in Madrid, Spain, three neighborhoods are identified as promising candidates: (1) Center City East: Broad to 8th, Callowhill to Pine, (2) South Philly: S. of Tasker, W. of Broad, and (3) Center City West: Broad to 16th, Callowhill to Pine. Of these neighborhoods, South Philly is clearly the most affordable, with a median house price of $209,000. These results are visualized in the following scatter plot of median housing prices versus similarity to Almagro.



**Conclusions**

An effective algorithm has been developed for identifying neighborhoods in a selected city that are most similar to some desirable neighborhood in another city. The algorithm utilizes k-means clustering to find groups of similar and dissimilar neighborhoods, and then identifies the most similar neighborhoods by calculating Euclidian distance from the target neighborhood in the multidimensional space. Finally, the housing price statistics are incorporated in the analysis to identify neighborhoods that might be more affordable for the user, yet still relatively similar to the target neighborhood.

When applied for finding a neighborhood in the city of Philadelphia, Pennsylvania, USA that is most similar to Almagro in Madrid, Spain, three neighborhoods are identified as promising candidates:

* Center City East: Broad to 8th, Callowhill to Pine
* South Philly: S. of Tasker, W. of Broad
* Center City West: Broad to 16th, Callowhill to Pine.

Of these neighborhoods, South Philly is clearly the most affordable, with a median house price of $209,000.