K Nearest Neighbor Clustering to Identify Developing Neighborhoods in Baltimore, MD

Andrew Timmons

Coursera Data Science Capstone Project

Goals

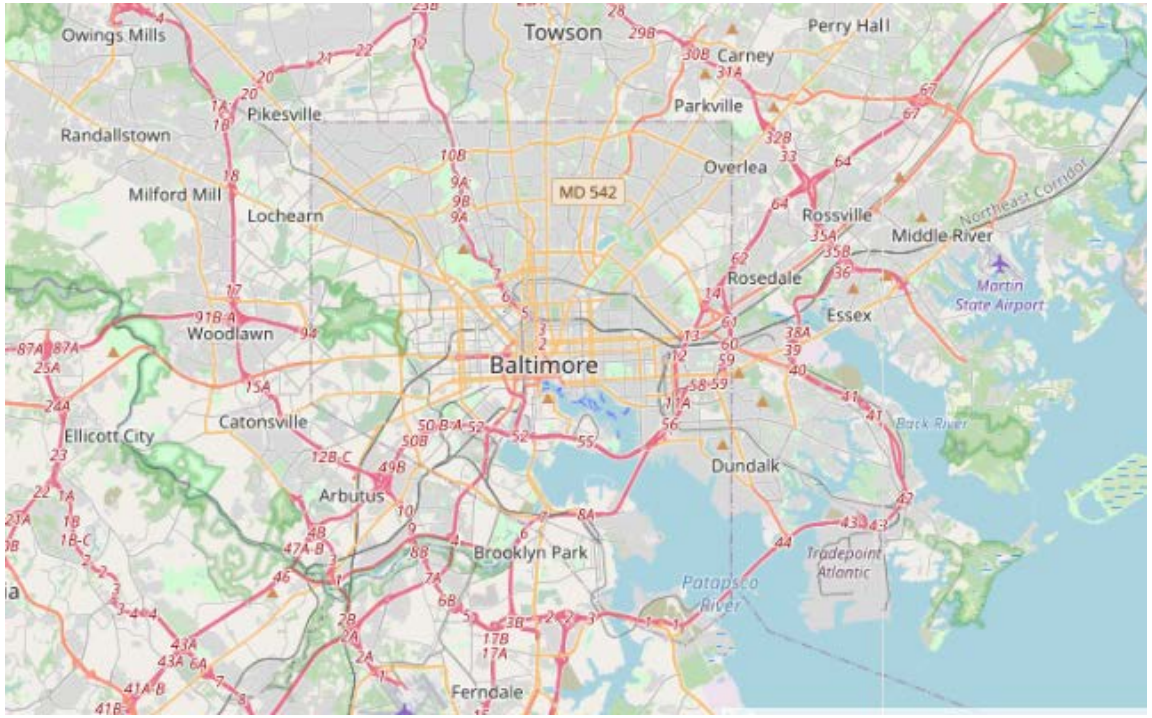
- I recently completed my PhD at Johns Hopkins University and have secured a research position in Baltimore
- As I will be in the area for the foreseeable future, I want to buy a home instead of renting.
- I want to determine neighborhoods that are candidates for further gentrification, because I want my home value to grow over time.

Approach

- I plan to use K Nearest Neighbor (KNN) algorithms to cluster neighborhoods in Baltimore according to their crime rates, income levels, and venue distributions
- Using KNN clustering, I will identify less expensive neighborhoods that have similar characteristics to expensive neighborhoods.
- I reason that A) my home value will increase as less expensive neighborhoods gentrify, and B) I will enjoy the lifestyle of more expensive neighborhoods without paying as much for my home.

Background

- Few American cities exemplify the pervasive issue of 'urban blight' to the extent of Baltimore City, Maryland
- As a result of public and private investments, certain neighborhoods have begun experiencing a revitalization of neighborhood character and local business.



Baltimore, MD

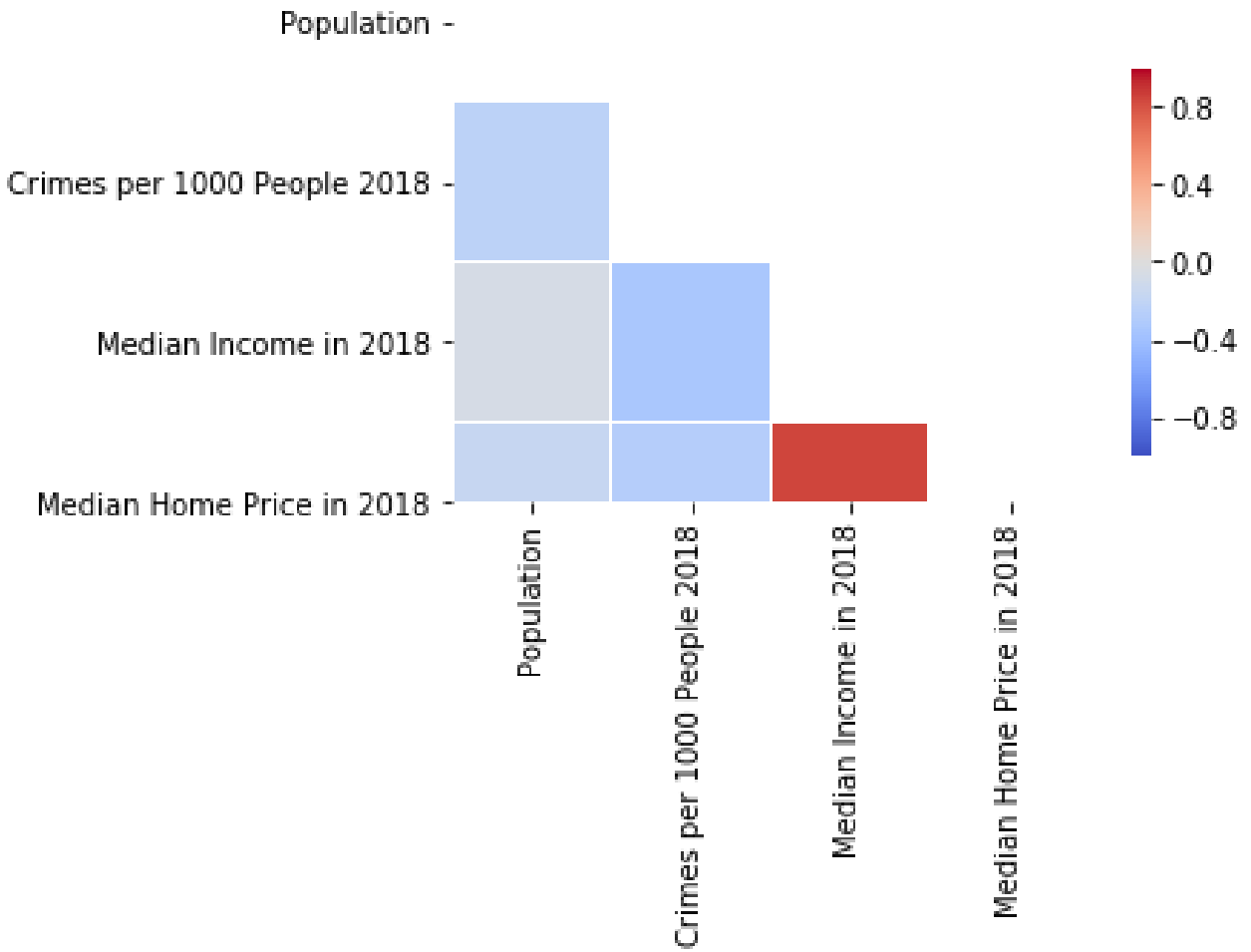
Approach

- Using K Nearest Neighbor clustering approaches, I will identify neighborhoods with low median property values in Baltimore that show similarity to more expensive neighborhoods. Clustring will take into account a variety of different venues as taken from FourSquare

- I will use multiple datasets as input to the clustering algorithm
  1. Demographic data on neighborhood population (Source: US Census)
  2. Population-weighted violent crime data (Source: Baltimore City OpenData)
  3. Venue data for neighborhoods (Source: FourSquare API)

Preliminary Findings

- Certain types of demographic data are (unsurprisingly) well correlated
- There is a strong correlation between neighborhood incomes and home prices



Preliminary Findings

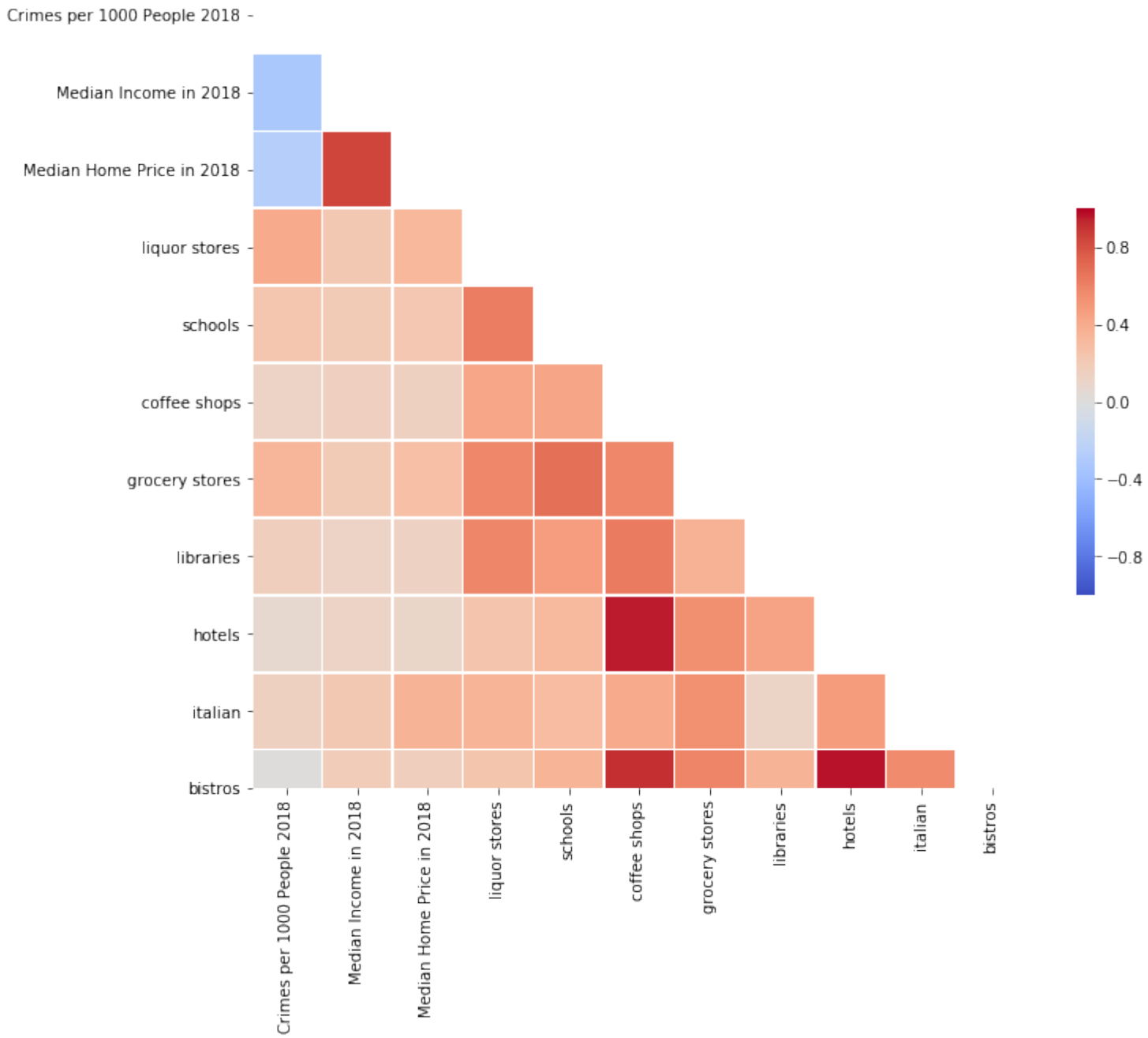
- Given the correlation between income and home prices, if I used KNN clustering only on demographic data, my clusters of neighborhood home prices would be almost entirely based on income, thereby not saving me any money
- I will next collect venue data from FourSquare to allow more nuanced comparisons between neighborhoods

Venue Data

- I used reverse geocoding to determine the latitude and longitude for each neighborhood
- For each neighborhood's latitude and longitude, I used an 800 meter radius (~10 American city blocks) as a search area for specific types of venues.
- The specific venues analyzed were chosen to represent meaningful neighborhood amenities
  - *I used this approach to minimize the number of API calls I had to make to FourSquare. There are hundreds of categories, and 55 neighborhoods being analyzed. This analysis can definitely be extended by adding in additional venue data*

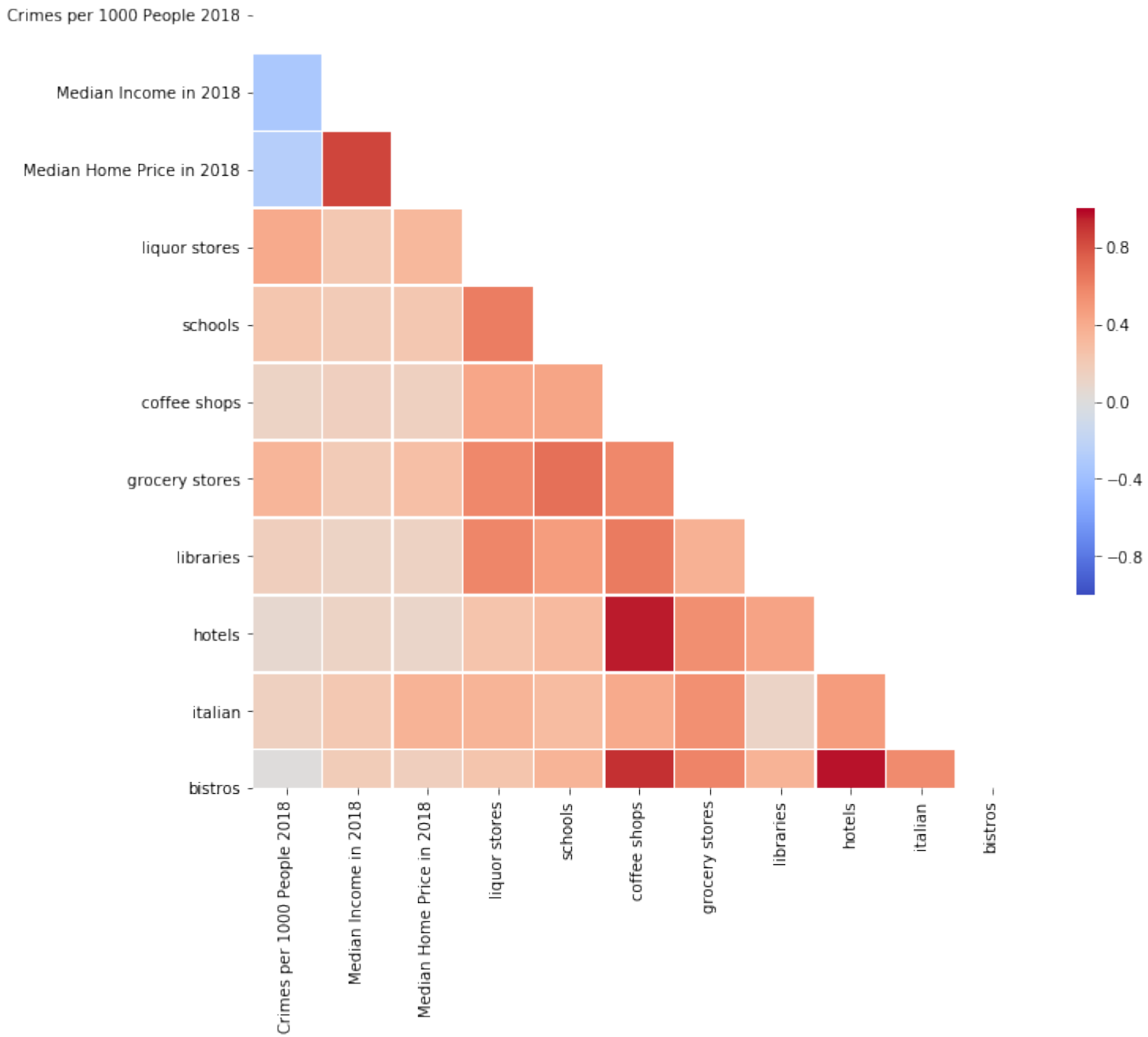
Venue Data

- Venues Analyzed:
  - Shopping
    - Liquor Stores
    - Grocery Stores
  - Educational
    - Schools
    - Libraries
  - Services
    - Hotels
  - Dining
    - Coffee Shops
    - Italian Restaurants
    - Bistros



Venue Data

- Certain venues appear to correlate well with each other (i.e. hotels and coffee shops)

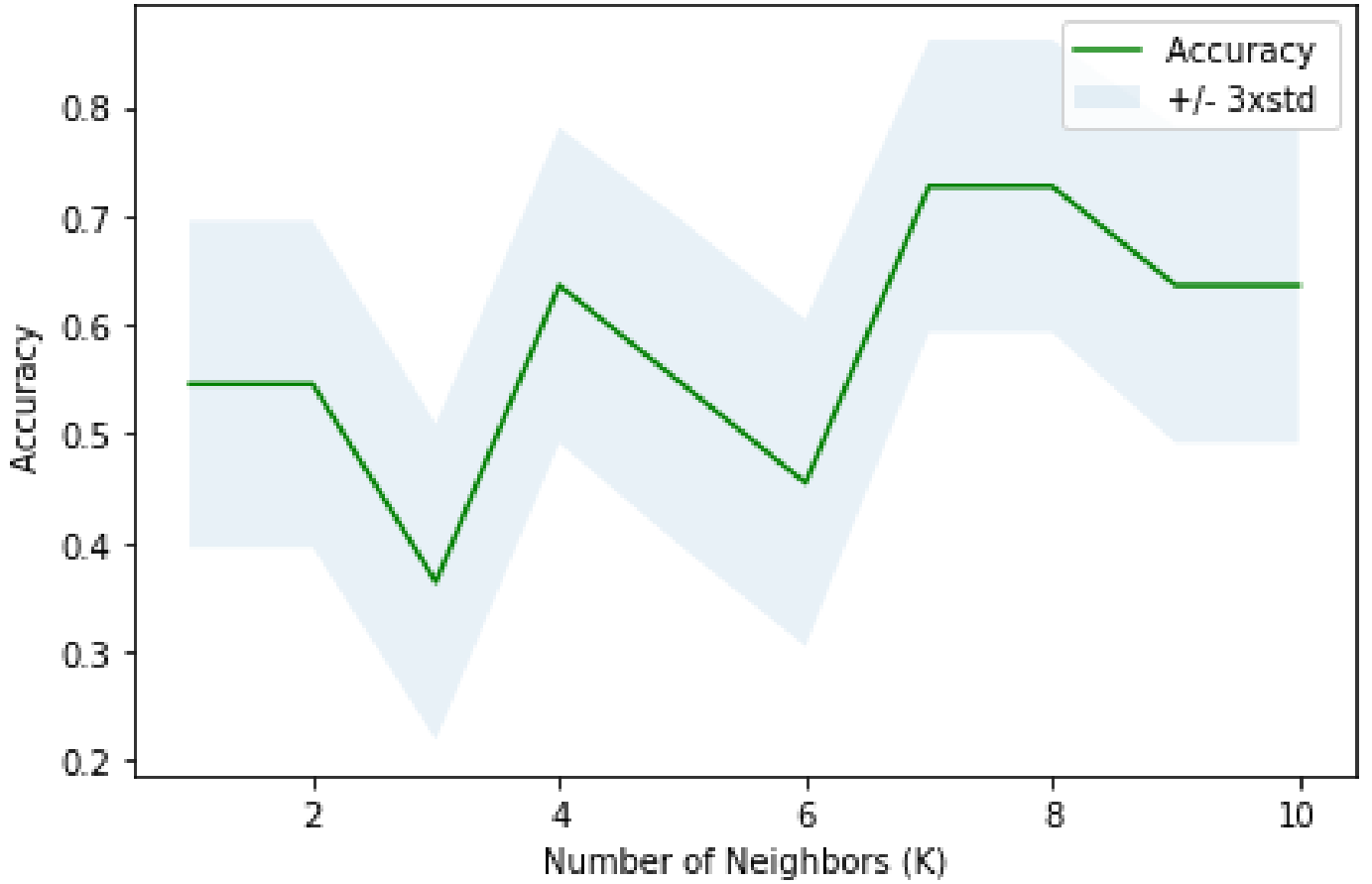


Building the Clustering Algorithm

- K Nearest Neighbor clustering was performed in python using the SciKit learn featureset.
- Clustering was used to predict the category of median home price
  - All neighborhoods in Baltimore were assigned to 5 evenly-spaced ‘buckets’ according to median home price (i.e. bucket 5 has the top 20% of home values)
- Neighborhoods (55 in total) were split into testing (20%) and training (80%) sets.
- Several iterations of the KNN instance were tried to identify the optimal K value for the clustering algorithm

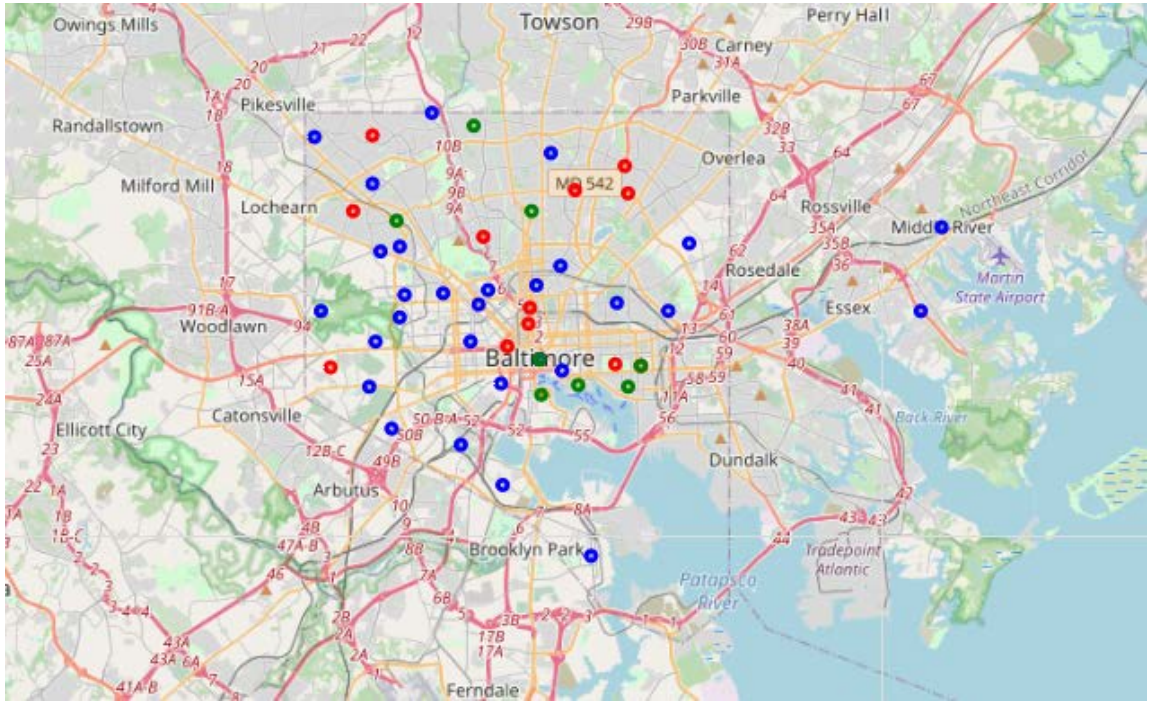
Building the Clustering Algorithm

- A K value of 7 gives the highest accuracy score (~0.73) when used on the test set



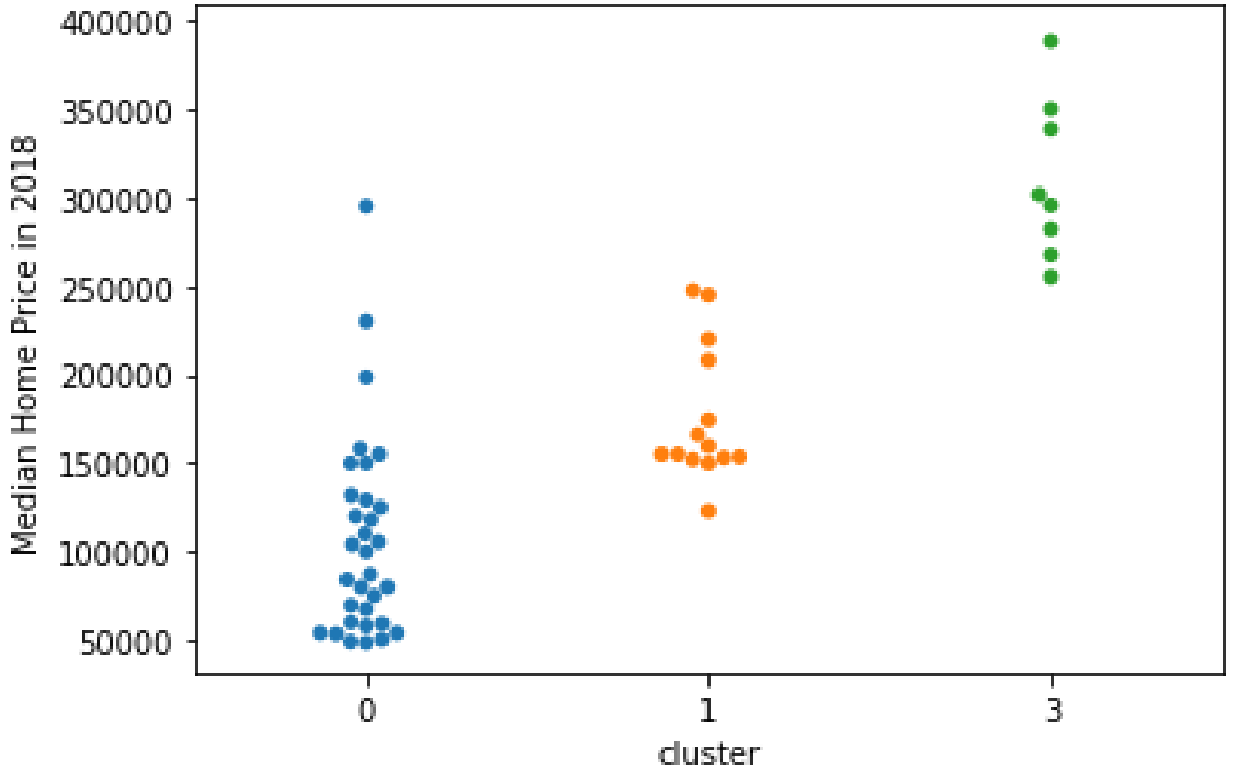
Clustering Neighborhoods in Baltimore

- The clustering algorithm was applied to all neighborhoods in Baltimore
- 3 distinct clusters were identified, which were distributed throughout the city



Baltimore, MD

Cluster Characteristics



- The clusters seem to nicely split the median home prices (the predicted variable in the clustering algorithm)

Cluster Characteristics

- The clusters appear to nicely separate median home prices into low/medium/high prices.
- The cluster with the lowest incomes (cluster 0) appears to have many neighborhoods with very low home prices, and is skewed towards lower home prices.
- Cluster 1 and cluster 2 have a distribution of home prices more centered around an average.
- I believe that the cheaper neighborhoods in cluster 3 represent the most attractive neighborhoods for investment

Cluster Characteristics

	Neighborhood	Population	Crimes per 1000 People 2018	Median Income in 2018	Median Home Price in 2018
37	Mount Washington/Coldspring	5168	5.224458	84848.090909	255000.0
15	Fells Point	9039	19.028654	98981.910064	267500.0
27	Highlandtown	7250	31.034483	88324.005277	282000.0
4	Canton	8100	9.135802	116911.088235	295500.0
29	Inner Harbor/Federal Hill	12855	14.080124	100932.230594	301500.0
38	North Baltimore/Guilford /Homeland	17464	4.466331	98094.852941	338350.0
47	South Baltimore	6406	4.995317	114107.319502	350000.0
21	Greater Roland Park/Poplar Hill	7377	2.575573	117951.409574	388000.0

Avg $268,000

Avg $358,000

Approximately 25% cheaper,
With similar amenities and
Neighborhood characteristics!

- I believe that the cheaper neighborhoods in cluster 3 represent the most attractive neighborhoods for investment

Conclusions

- From this analysis, I would assume that the Mount Washington, Fells Point, and Highlandtown areas are the most attractive neighborhoods in which to purchase a house
- Despite having substantially lower home prices, the Mount Washington, Fells Point, and Highlandtown areas share low crime rates and have similar venue distributions to the more expensive neighborhoods of South Baltimore and Roland Park.

Future Directions

- This analysis can be refined by including more venue data from FourSquare and more demographic data from the Baltimore City opendata project.
- Additionally, repeating this analysis with different machine learning algorithms (such as logistic regression) will allow inspection of factor coefficients. This will permit inference on the relative impact of each factor on a neighborhood’s home prices