Exploring Boston Real Estate Prices

Boston Overview

- 10th largest metropollitan area in the U.S.
- 4.8 million population
- Aprox. 90 square miles
- Major economic sectors: education, technology, biotech

Housing Overview

- Median home value of $630,000
- Aprox. 148% appreciation since 2000
- Many looking to buy real estate, either for purposes of home or investment earnings

Data Acquisition and Cleaning

Zillow Home Value Index

- Aggregated dataset of houses segmented by zip code

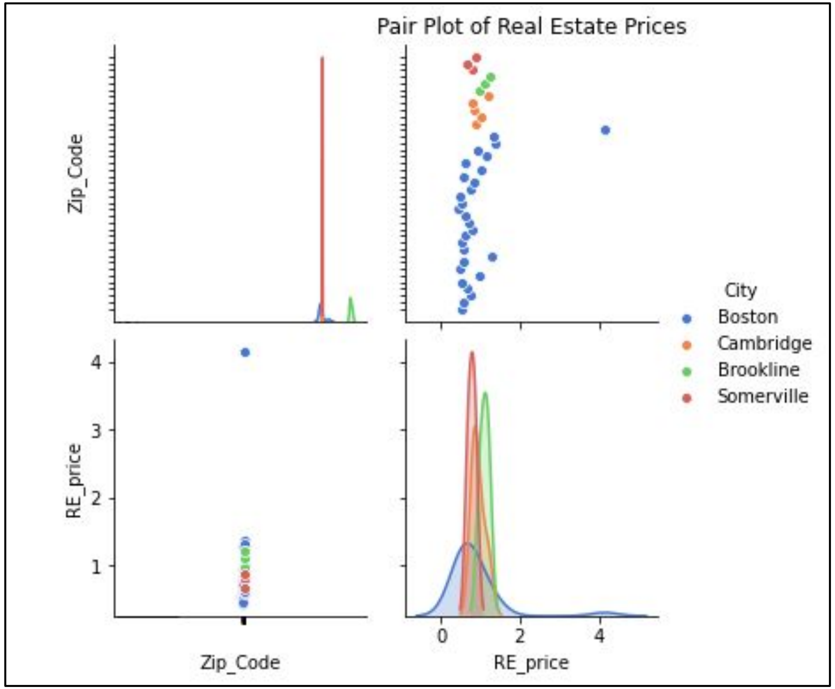
Foursquare Venues API

- Used to gather venue data for individual zip codes

Google Geocode API

- Generated latitude and longitude coordinates of zip code

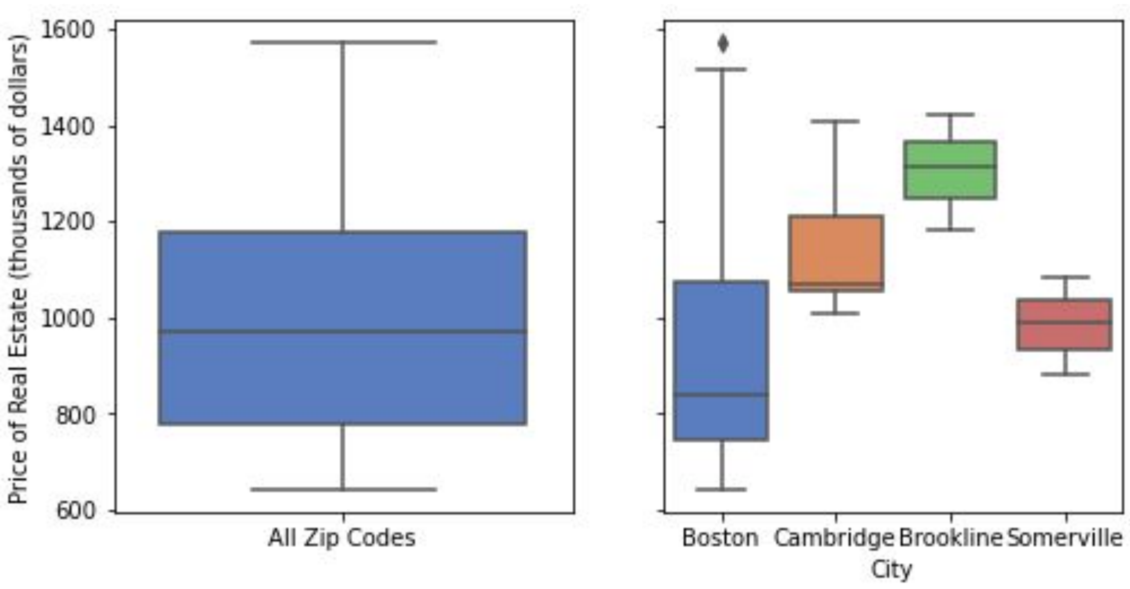
Data Acquisition and Cleaning



Variables

- One clear outlier was identified (zip code 02199)
- Leaves with 38 individual zip codes

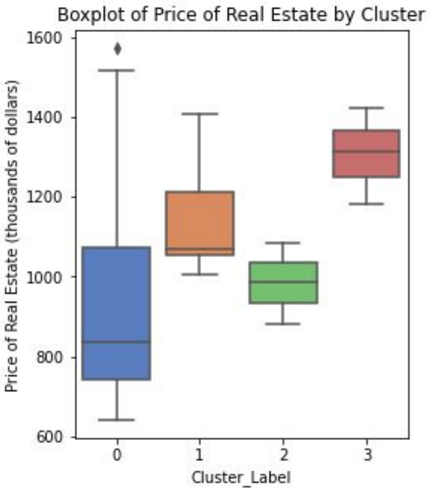
Feature Selection

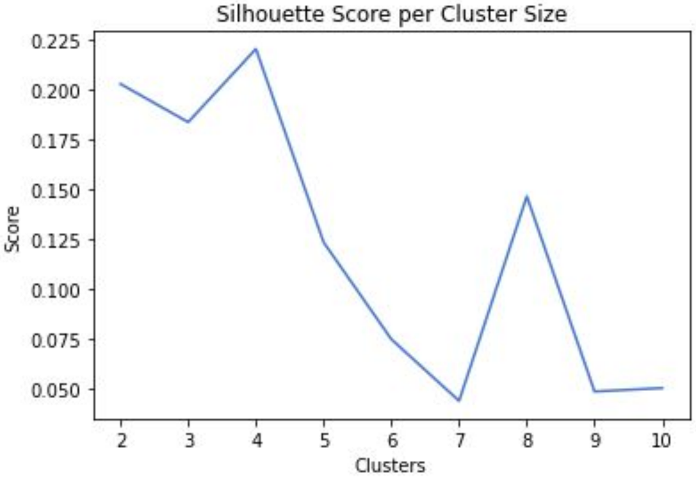


City

- Included in the features due to apparent correlation
- Somerville was least correlated

K-Means Clustering

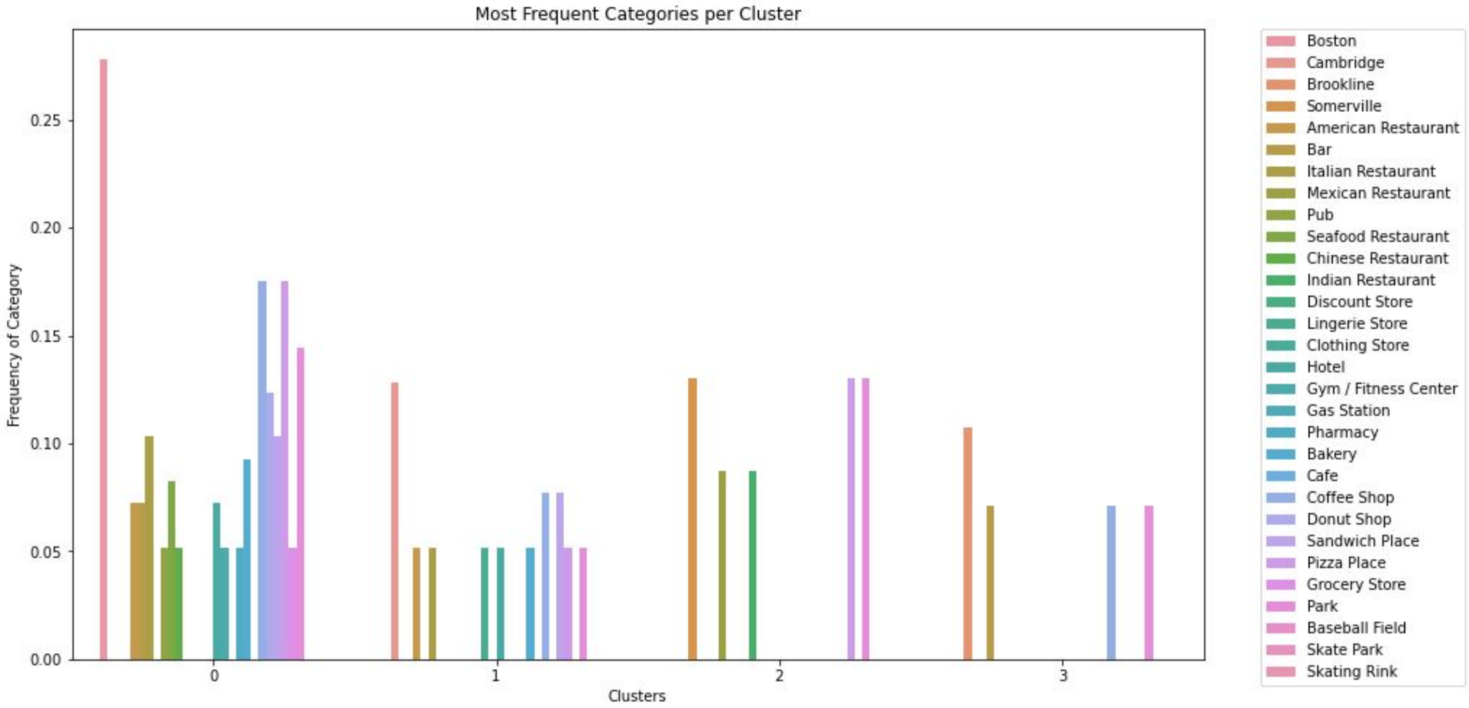

Boxplot of Price of Real Estate by Cluster


Silhouette Score per Cluster Size

Number of Clusters

- According to silhouette score, four clusters was selected
- Possible correlation was seen in cluster divisions

K-Means Clustering



K-Means Clustering

Cluster 1

- High frequency, Boston, QSRs, retail, public service

Cluster 2

- Medium frequency, Cambridge, QSRs, retail

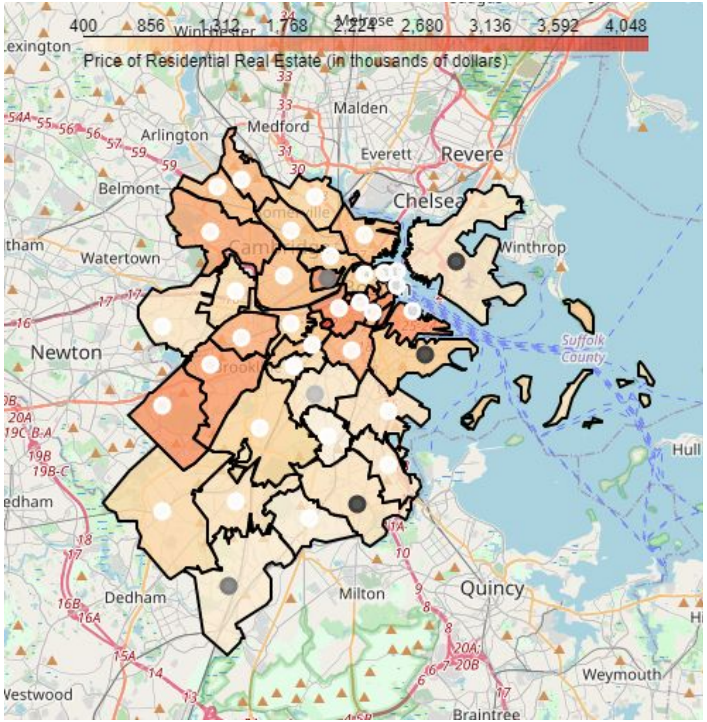
Cluster 3

- Low frequency, Somerville, QSRs, retail

Cluster 4

- Very low frequency, Brookline

K-Means Clustering



Price of Real Estate vs Clusters

- Possible relationship between cluster label and real estate price observed
- Primarily one cluster (white dots)

Statistical Testing and Modeling

Linear Regression

- Tested using five-fold cross validation with two feature sets: full dataset and highly correlated eight feature set
- Full Feature Mean Score: 0 .09
- Selected Mean Score: 0.242
- No statistical significance

Random Forest Regression

- Tested using five-fold cross validation with two feature sets: full dataset and highly correlated eight feature set
- Full Feature Mean Score: -0.42
- Selected Mean Score: -0.06
- No statistical significance

K-Means Clusters

- Tested using one-way ANOVA test with p-value <= 0.05
- F Statistic: 2.584
- P-Value: .069
- No statistical significance

Conclusion and Future Direction

Modeling

- Attempts were unsuccessful due to small size of dataset. Not enough to successfully complete cross validation.

Possible Improvements

- Need more exact information, such as pinpointed real estate prices
- Comparison to larger sets of real estate price data, such as in other comparable cities
- Better venue data - Foursquare can't provide all venues within a set geographic boundary