Boston Real Estate and Venue Data

Visualization, Clustering, and Modeling

By Theo Pacun
IBM Coursera Capstone
August 5th, 2020

Introduction

Background

The Greater Boston Metro area is one of the most expensive places to live in the United States. As the tenth largest metropolitan statistical area in the US, Boston is home to a continuously growing population of approximately 4.8 million people.[1] This growth is driven by both general global urbanization trends as well as a thriving economy — particularly in the fields of education, technology, and biotech. Demand for housing has exploded over the past decade and is also aided by the annual turnover in student populations. All of these factors contribute to a high median home value of approximately $630,000.[2]

Given the high cost to enter the Boston real estate market, most people — whether professional real estate investors or individuals looking to diversify their investment holdings or purchase a home — want to know as much as possible before they purchase. Trying to find a deal in the city is difficult, but can be aided by looking in areas that are cheaper but also have the desired surrounding amenities and venues.

Problem

This project aims to discover more information about Bostonian zip codes and the most popular venues within each. Additionally it will look at the possibility of segmenting, classifying, and modeling the median prices of the neighborhood.

Audience

Since many individuals and institutions are interested in the prices of real estate, the audience for this survey is fairly broad. This includes not only companies and investors, but also homeowners and renters. Generally, anybody who is interested in what drives changes in the real estate market beyond general larger demographic shifts will find some value in the results.

[1] Metropolitan and Micropolitan Statistical Areas Totals: 2010-2019

[2] Boston MA Home Prices & Home Values

Methodology

Data Collection and Cleaning

Zillow Home Value Index[3]

This is not the full dataset of all Zillow listings, but an aggregated month-by-month compilation of them dating back to 1994. There are many different segments available, including smaller areas such as neighborhoods. However, Zillow unfortunately does not provide updated information on the geographic boundaries it uses for each neighborhood. I was able to find an older version of their neighborhood boundaries, however the coverage was extremely sporadic with many uncovered areas in the middle of the metropolitan area.[4] Furthermore, from personal knowledge of Boston neighborhoods, several were irregularly shaped, noncontinuous, and/or only encompassed an extremely small area of one square block — not something I would refer to as a neighborhood.

All of this meant that for the purposes of this project, I selected the dataset segmented by zip codes, as boundary information was much more readily available. This dataset contains data for all zip codes in the United States and does have a "Metro" feature which includes "Boston-Cambridge-Newton." However, upon further inspection, it also contained more distant towns such as Plymouth[5] which I was uninterested in. I narrowed the dataset down by selecting for the cities of Boston, Cambridge, Somerville, and Brookline, as well as the most recent real estate prices which were from May 31st, 2020.

Google Geocode API[6]

Since the Foursquare API is only capable of providing location data, I used the Google Geocode API for a few queries. The first was to obtain the center of Boston, as a centerpoint for the map visualizations I created. I also used the Geocode API to obtain and add the latitude and longitude of the center of each zip code to the existing dataframe.

Massachusetts GeoJson[7]

This was used as the parameters for the zip codes of Boston. It wasn't in the cleanest format, however — the zip codes in the geojson file itself were encoded as integers. It also contained every zip code in Massachusetts[8] which had more boundaries than I wanted. For

[3] Zillow Housing Data

[4] Zillow - US Neighborhoods - OpenDataSoft

[5] According to Google Maps, Plymouth is approximately 40 miles from the center of Boston.

[6] Overview | Google Geocoding API

[7] OpenDataDE Github | State Zip Codes

[8] Except for those without geographical boundaries.

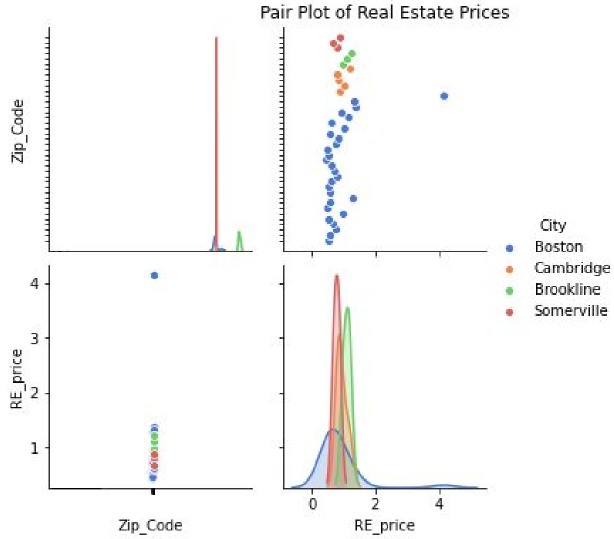
easier manipulation, I read the file into a GeoPandas DataFrame, cleaned the zip codes and selected those which matched the zip codes in the real estate data.

Foursquare API[9]

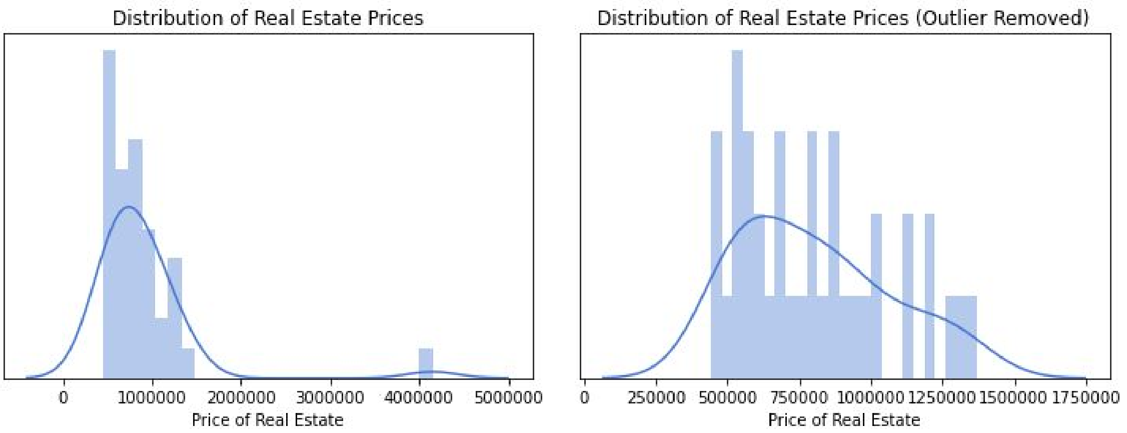
I made calls to the Foursquare API to generate lists of up to the 100 most frequent venues within 750 meters of the center of each zip code. I then grouped the information by each zip code and selected for the category of each venue. This allowed me to compute the frequency of 294 unique venue categories for each individual zip code.

Exploratory Data Analysis

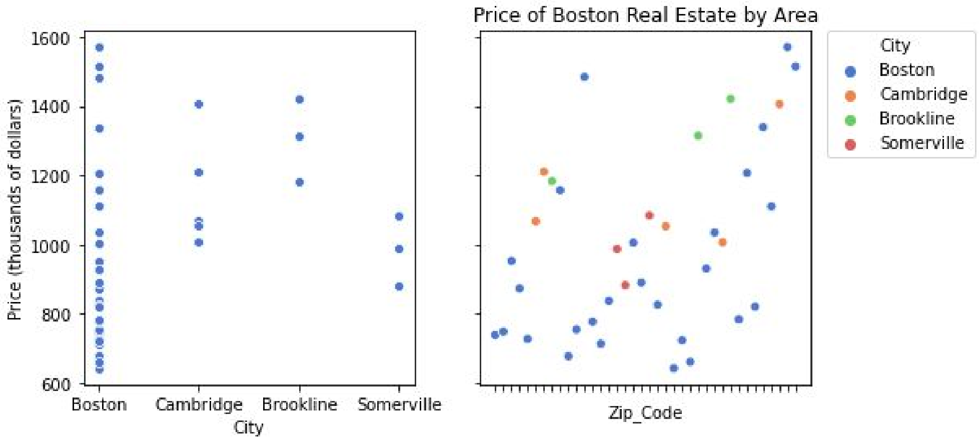
After getting all of the information required, the first thing I did was check the overall distribution of the dataset. As seen below, there was one clear real estate price outlier, with the rest falling within a fairly regular distribution. Given the extreme nature of the data point, I elected to simply drop the entry. The initial pairplots and the distribution of the target feature can be seen both before and after the deletion.

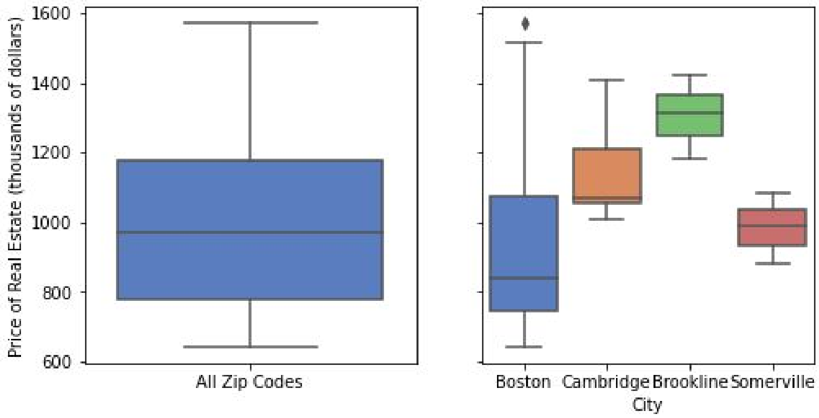


[9] Foursquare Developer Info

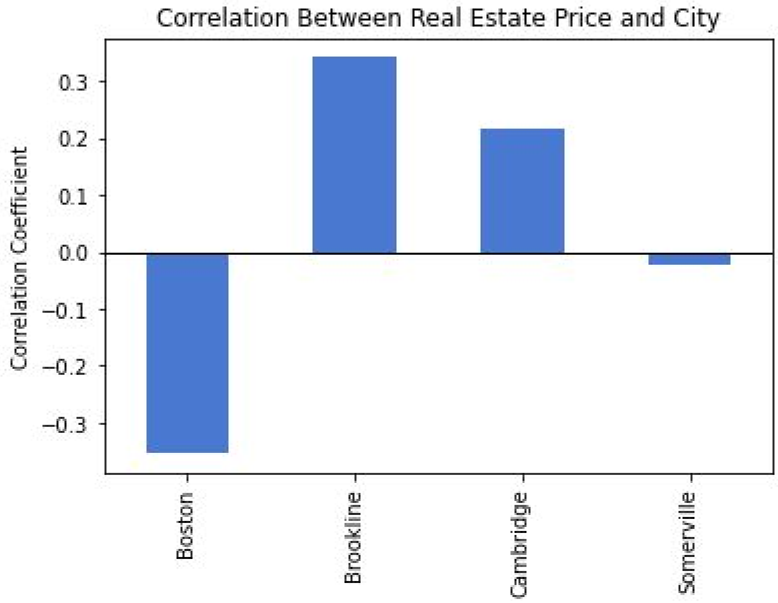


As I decided earlier, the area I examined was the metropolitan Boston area, which is composed of four cities: Boston, Cambridge, Somerville, and Brookline. As shown by the scatter and box plots below, there seemed to be some sort of a relationship between the two.

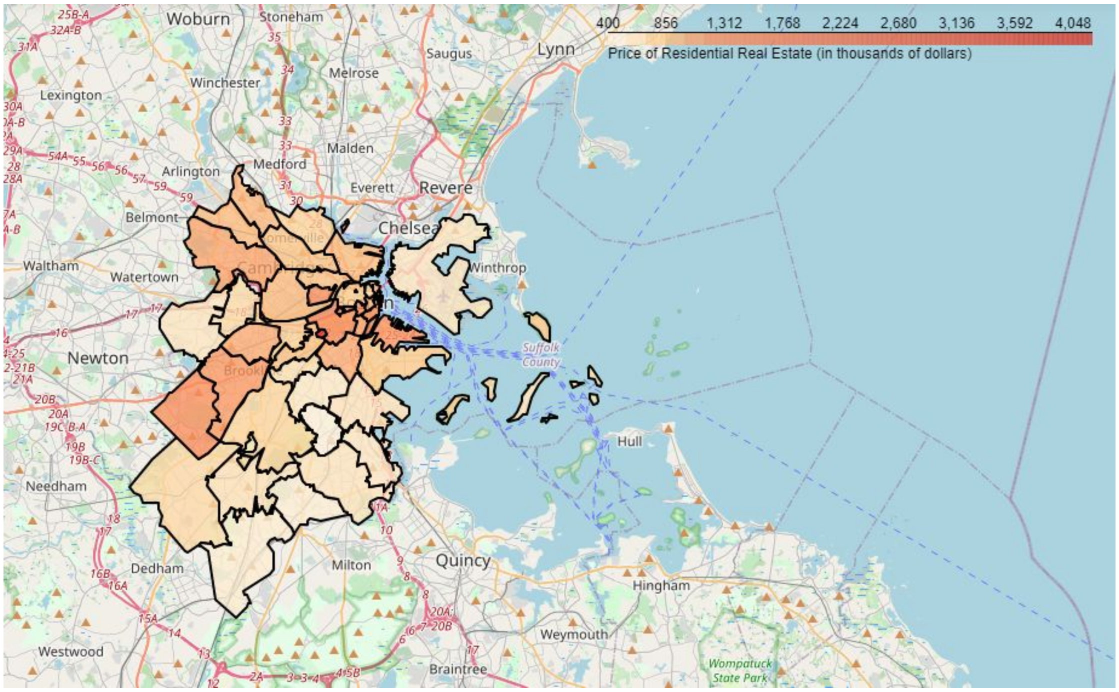




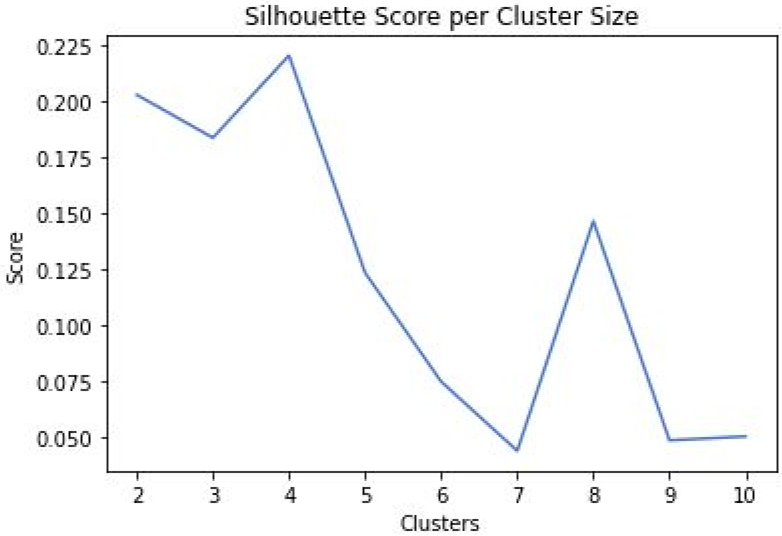
I was interested in seeing if there was any sort of relationship between that and the target feature. The visual results were encouraging, so I calculated the correlation coefficients between each city and the target and decided to include them in any possible modeling. Both the city segmented visualizations and the correlation coefficient calculation are below.



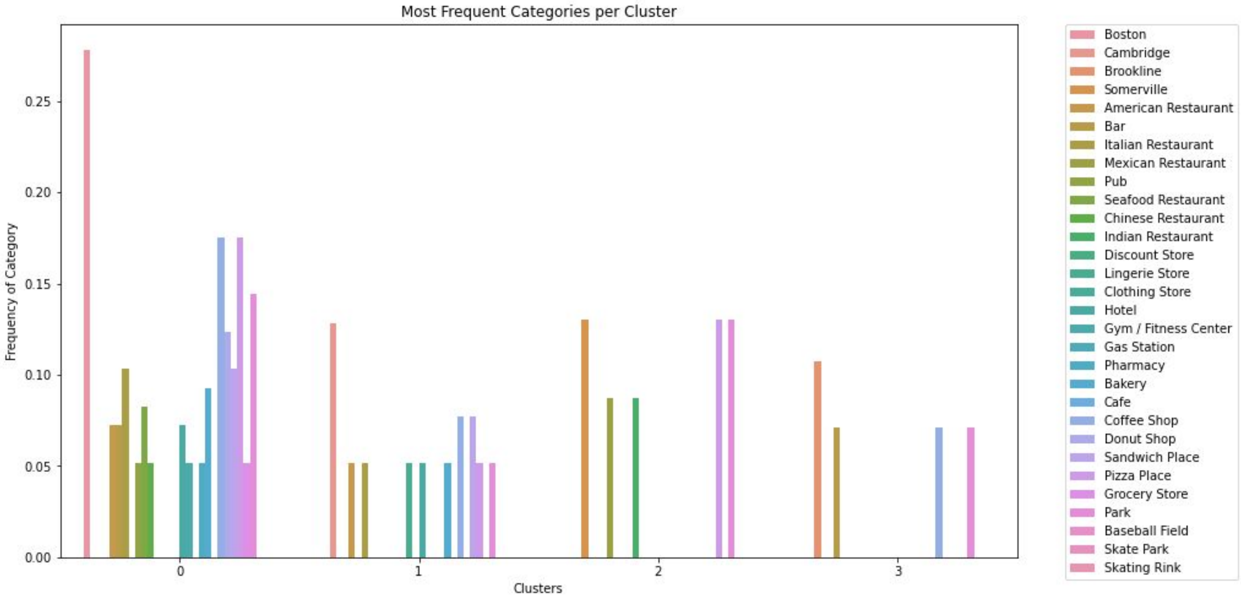
Moving forward, I mapped the real estate value by zip code out onto a choropleth map, with darker reds being higher values and lighter yellows being lower ones. It appears that most of the higher priced real estate is closer to the center of the city.



Using the frequency of venue data gathered by Foursquare, I then clustered each zip code according to both the (one hot encoded) city feature as well as the venue frequency. I used the unsupervised clustering algorithm K-Means. After iterating through up to 10 clusters and calculating the silhouette score for each, it was clear that four was the correct number of clusters.



Basic distribution information about the target feature of each cluster showed that this might have been an effective model, as shown by the boxplot below. Since the clusters represented the possibility of a viable model, I grouped each according to the frequency of venue and city, then graphed the results below.



Given this, I have made the following inferences about each zip code based upon the frequency of venue type in each zip code:

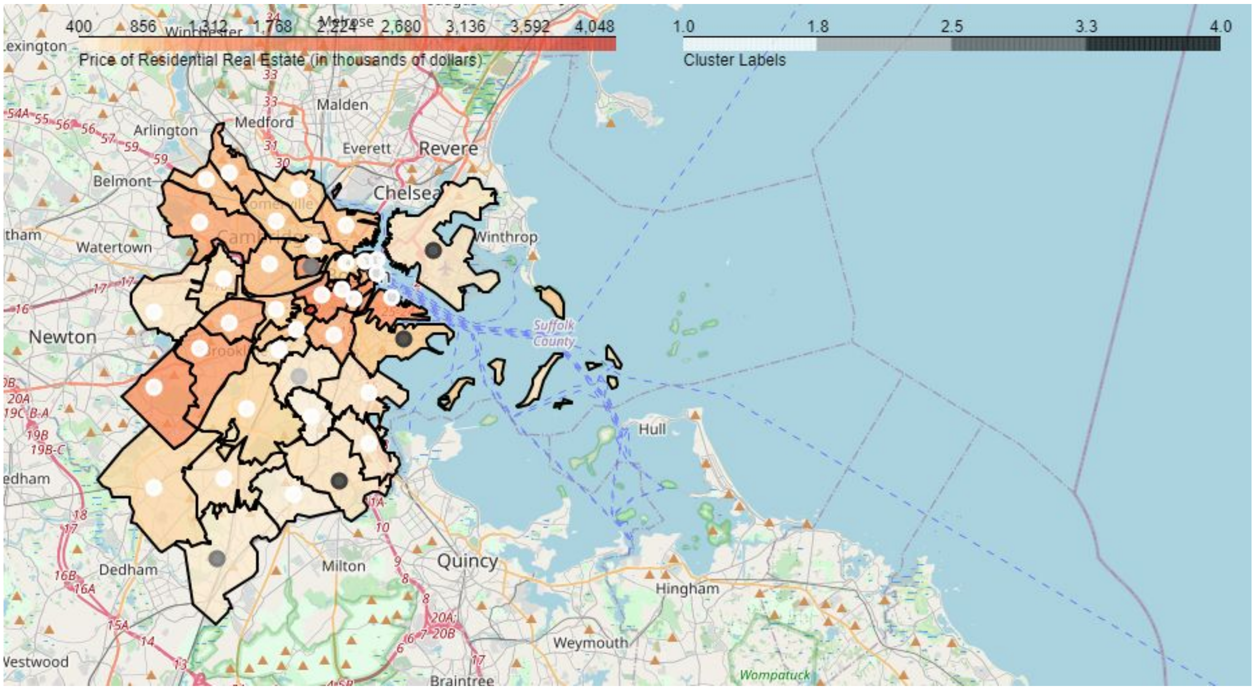
Cluster 1: high frequency of venues, city of Boston, quick service restaurants, retail locations, public service venues

Cluster 2: medium frequency of venues, city of Cambridge, quick service restaurants, retail locations

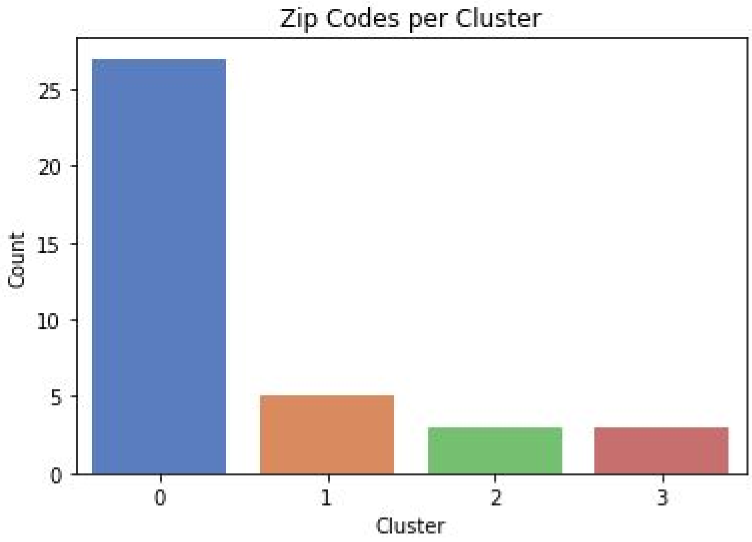
Cluster 3: low frequency of venues, city of Somerville, quick service restaurants, retail locations

Cluster 4: very low frequency of venues, city of Brookline

I then mapped out each cluster onto the choropleth map generated earlier to see if any discernible insight was possible.



As described above, it appears that the K-Means clustering algorithm predominantly used the city feature. However, you can see that the discrepancy in the number of zip codes in each cluster is very large and the distribution uneven so the relevance of the clustering may be statistically unimportant.



Predictive Modeling and Results

Linear Regression

I made a few attempts with linear regression — one with all 294 unique venue frequencies and one with a set of eight features selected for largest correlations (both positive and negative). Below are the means of cross validated scores with five folds for each model.As you can see by the scores in the table below, both of the models performed extremely poorly, and did not warrant further exploration.[10]

	Full Features	Highly Correlated Features
Full List of Cross Validation Scores	0.5137, 0.7326, -0.687, 0.0461, -0.1249	0.6358, 0.6618, -0.7135, 0.7464, -0.1177
Mean Score	0.0961	0.2426

[10] I did also explore using Lasso regression to handle the large number of features relative to the sample size of data. However, though it did perform better than the standard multiple linear regression above, it did not warrant full inclusion in this report.

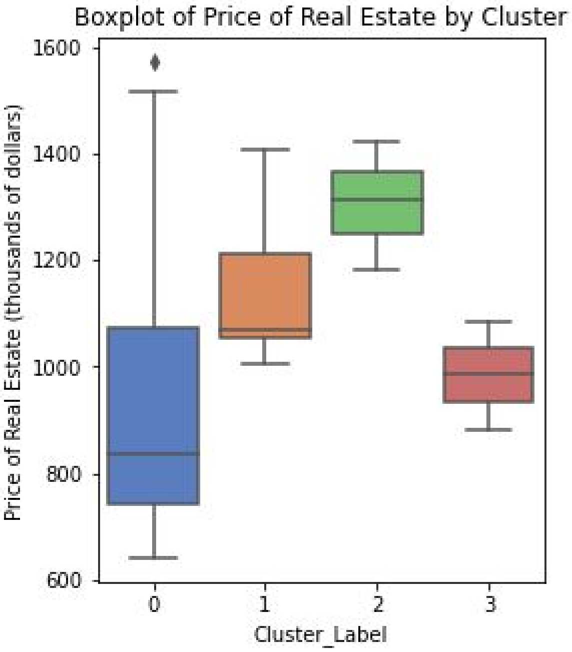
Random Forest Regression

I also made a few attempts with random forest regression which definitely performed better if not with statistically actionable modeling abilities. I attempted the random forest regression technique with both the full featured set of 294 as well as the same highly correlated eight features from above. The best result — again, still a very poor one — came from the model trained on the complete feature set.

	Full Features	Highly Correlated Features
Full List of Cross Validation Scores	.0.0628, 0.0051, -0.4316, 0.0766, -1.8256	0.029, 0.0969, -0.3505, 0.5107, -0.6033
Mean Score	-0.4226	-0.0634

K-Means Classification

Lastly, I tried segmenting the neighborhoods into different clusters using the K-Means algorithm. In order to do this, I selected the top ten most frequent venue categories in each zip code and iterated through fitting the model with different numbers of clusters. As shown earlier, I used the silhouette score which showed the optimal number of clusters was three. Using these clusters as the feature selection, I then looked at whether or not any of the clusters could be significantly correlated with higher or lower real estate prices.



Though this didn’t appear to have an obvious visual difference, it did warrant statistical investigation. Accordingly, I ran a one-way ANOVA test, of which the results are in the figure below.

F Statistic	P-Value
2.584	.0693

Discussion

While there are some interesting inferences to be made from this data and the ways the K-Means clustering algorithm mainly grouped each zip code, much of it is unactionable. I tried three different types of modeling on the datasets: KMeans classification, linear regression, and random forest regression. Unfortunately, all three ran into problems. Both of the regression models simply could not account for a meaningful amount of variance in the samples, whether the features were complete or explicitly selected. The K-means classification model appeared to possibly show encouraging results with an F statistic of around 2.584; however, the p-value of .0693 showed the sample size is simply far too small to prove that the score is not a random chance.

The fact of the matter is there are only 39 zip codes in Boston and that is simply too small of a sample size to effectively train and test a model and receive results with any sort of statistical significance. Even though there were 300 unique venue categories this is still not enough to create reliable models with that sample size — and in fact is somewhat detrimental to them.

Looking forward, in order to gain a better understanding of the effects of venues on real estate price, a better dataset is needed. The most obvious solution is a more exact real estate dataset. If for example, an API could be provided which returned the average price of real estate in a given radius at a given location, then it would be possible to develop a much better model. Indeed Zillow used to provide this but no longer supports such queries. Even if this wasn’t possible, a better neighborhood segmented dataset could be provided.

Conclusion

In this study, I analyzed the relationship between venues and zip code real estate prices. I visualized each region’s real estate price and pulled the top 100 most frequented venues from each zip code. I then used this information to attempt to model real estate prices using both supervised and unsupervised machine learning algorithms. Unfortunately, due to the shortcomings of the data and the scope of this assignment, no statistically significant models were able to be developed. Gaining access to a more comprehensive data source than the

Foursquare API or Zillow’s publicly accessed data would be the first step in expanding further upon this report