Identifying the Undervalued "Fixer Upper" Properties in Staten Island

Dan Fisher
August 20, 2020

1. Introduction

1.1 Background

New York City is one of the most expensive real estate markets in the world. Divided into five boroughs, the variety of housing styles and prices are as diverse and wide ranging as the different neighborhoods that make up one of the world's most wellknown cities. As such, many individuals are shut out of the New York housing market. Because of the nature of the New York real estate market, it is often quite difficult to gauge the true value of home relative to the asking price on the real estate listing. More specifically, sometimes homes are listed as "fixer-uppers", which many would conclude the price has been lowered to allow for future repairs and upgrades, however it take a highly trained and experience real estate professional and potentially home builder or developer to understand how much a home is actually worth, or would be worth if a certain amount of money was put into it. That is what this project aims to help with – finding undervalued homes or home with potential with investment in repairs.

1.2 Problem

The specific output of this project will be an algorithm which identifies potentially undervalued properties on Staten Island – one of the five New York Boroughs which is known to have a diverse mix of real estate, including traditional single-family homes, condos, multi-family units. The different neighborhoods in Staten Island are also quite diverse in that some are more traditionally residential, while others are more suburban and mixed use. Aside from a general property evaluation, the output of this exercise will be directly applied to a number of current listing marked 'fixer uppers' and we will use it to gauge if they are actually being discounted or if they are already priced as if they are ready for the market.

1.3. Interest

The output of this algorithm could be used by a number of individuals ranging from a potential home buyer to a real-estate investor or developer. For example, as an investor I could use this program to identify current listings that have an asking price lower than the "comps" in the area, and as such, may be undervalued and could be a good investment.

2. Data and Cleaning

2.1 Data Sources

The first is the active real estate data with is currently listed on Redfin.com. From Redfin I will be able to query all current active listings and a number of their attributes, such as price, bedrooms, neighborhood, HOA, etc.

The second data source will be the foursquare API for the general exploration of the neighborhoods. Using the potential housing available I will explore the different neighborhoods and ultimately choose the best houses to make offers on.

2.2 Cleaning the Data

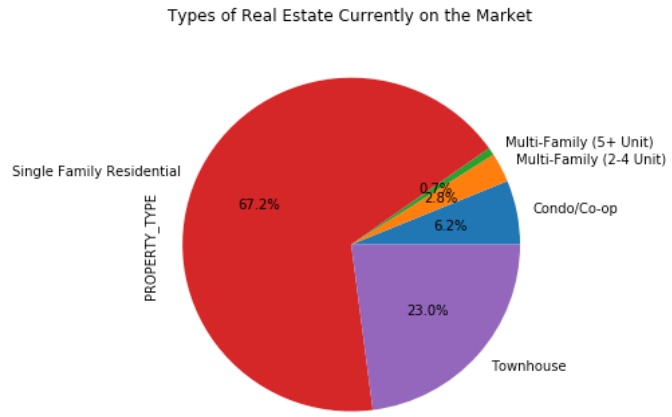
Real Estate data. The data coming from the Redfin database in generally pretty clean and usable, however it does require further modification in order to fit the needs of this particular analysis. For example, much of the data is fed in from the listing agents, so data may be missing, the neighborhoods may have typos, contain abbreviations, or be called by an alternative name. As such, I spent a significant amount of time correcting these issues and standardizing many of the property' attributes.

Neighborhood data. Using foursquare API we will examine the different neighborhoods on Staten Island in which we have our real estate listings. We will cluster these neighborhoods based on the different characteristics of the mix of venues. These clusters will then be appended to the real estate listings.

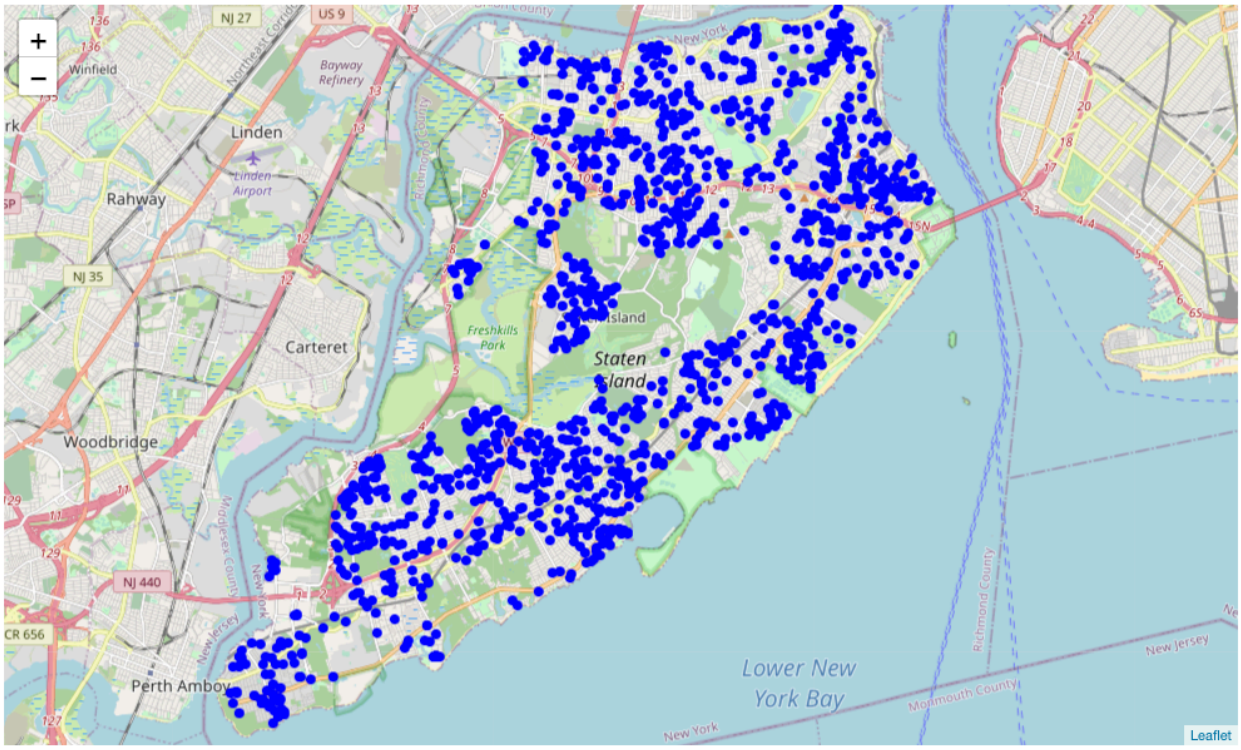
3. Methodology

Gaining a baseline understanding of the pricing of Staten Island Homes

In order perform this study I used housing market data from Redfin and a number of python libraries such as pandas for general data analysis, matplotlib for visualizations, sklearn for machine learning and the folium library to visualize the areas of Staten Island in which these potential homes are located. The primary data points for mapping were the latitude and longitude of each residence. Much of the housing specific data is from the real estate listings, and a number of variables were ultimately created in order to perform this analysis.
The following visualizations demonstrate the variety of housing currently on the market. The majority are single family homes, however there are a significant proportion of townhomes and finally a small percentage of condos and multi-family properties.

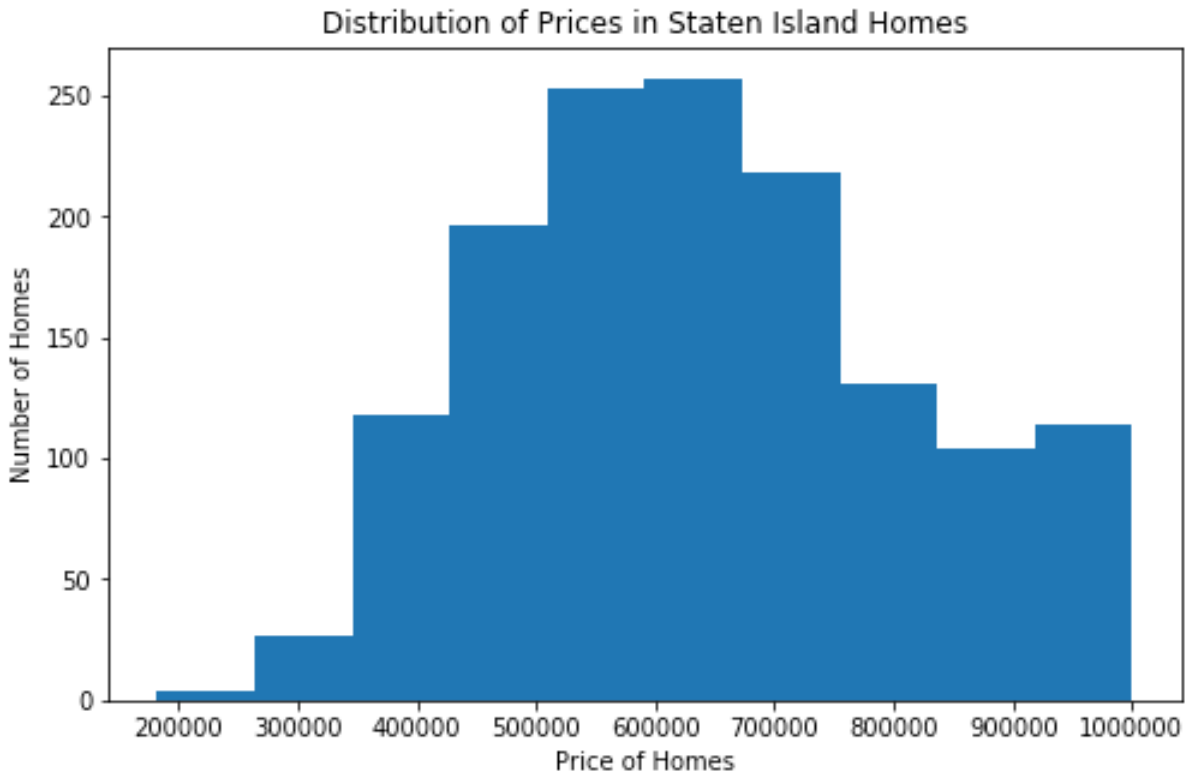


The map below shows the location of the different listing across Staten Island. There are currently 1755 listings. As one would expect in Staten Island, the price range for houses is quite broad. Currently on the market there are home costing as much as 2 million dollars.



In order to focus the inquiry and help control for many of the potential outliers, the data set was narrowed to only single-family homes and townhouses listed for less than 1 million. This decreased our overall housing file to 1400 listings. As a potential real estate investor these are the most common types of homes I am likely to consider buying and they are same type as the 21 of the 23 “fixer upper” homes, thus should give us a decent basis for our initial modeling.

The histogram below shows the distribution of prices for the homes in the data set. As you can see, we do have a general bell curve, but not an exact normal distribution.



Off of the basic real estate listings we have a number of variables that one would think would have a correlation with the asking price. We can see that PRICE is most strongly correlated with SQUARE FEET, followed by Bath, and Beds – the year and lot size show little correlation and will be left out of this initial analysis.

LOT_SIZE	0.02953
YEAR_BUILT	0.10821
BEDS	0.51565
BATHS	0.52282
SQUARE_FEET	0.68259
PRICE	1.00000

Using these three variables I have built a basic linear model. The results are:

```
lm.intercept_
```

```
array([209806.22179413])
```

```
lm.coef_
```

```
array([[33594.71024177, 38172.17146878,   132.14890801]])
```

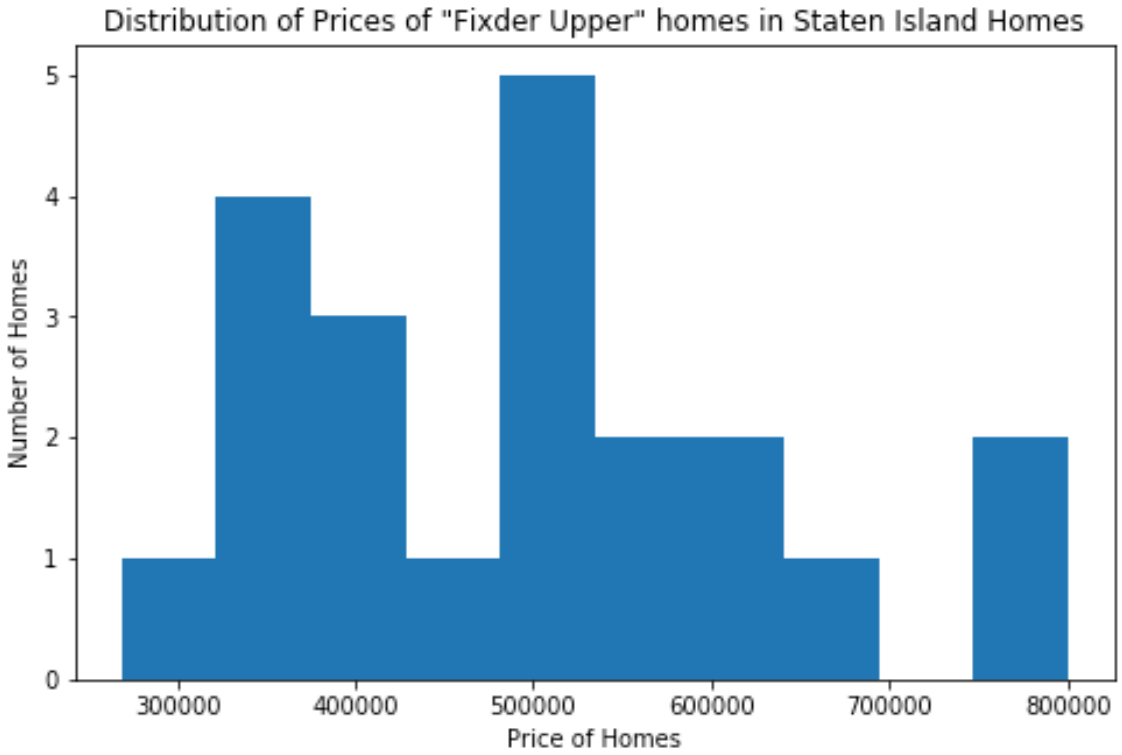
```
lm.score(Z,P)
```

```
0.5370579468171552
```

Or simply speaking – starting at 209806.22, we can add 33594 for each bedroom, 38172 for each bath, and 132 for each square foot of space.

Also, these three simple independent variables explain 53.7% of the variance in the housing price. This is actually quite good considering all of the other attributes that influence the price of a home not currently part of the model.

Looking at the current price distributions we do not have a clean a histogram, mostly due to the difference in records – there are only 21 "fixer uppers" to evaluate.



Despite the low number for comparison we can still perform an initial evaluation of the discounted prices. By applying the regression formula to our fixer upper houses, we can derive what they "should" cost if fixed up. Using this simple method, we can see a calculated price that is on average 65k less than projected for a fixer upper home. This does tell us these homes are generally discounted.

	PRICE	projected_price	price_diff
count	21.00000	21.00000	21.00000
mean	499071.42857	564717.33257	-65645.90400
std	143316.67085	88575.73164	119800.43788
min	269000.00000	438859.19164	-257636.73782
25%	399000.00000	502932.54908	-174115.87956
50%	489000.00000	572500.64991	-53040.24393
75%	579000.00000	592151.92494	-14684.36118
max	800000.00000	825378.14909	161040.80836

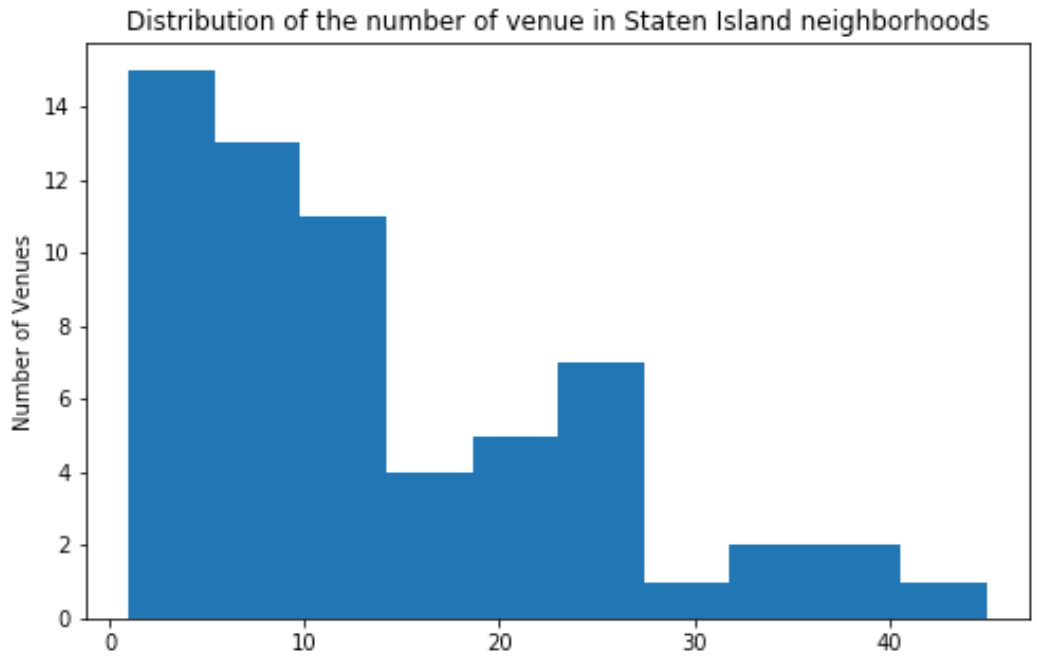
In fact, only 4 homes have a discounted price above what their calculated normal price should be. Again, these three variables only explain slightly more than half of the variance in price, so we must add additional variables to the data set and perform a more formal multivariant regression model with machine learning.

Examining the Neighborhoods and developing additional independent (contributing) variables for our model)

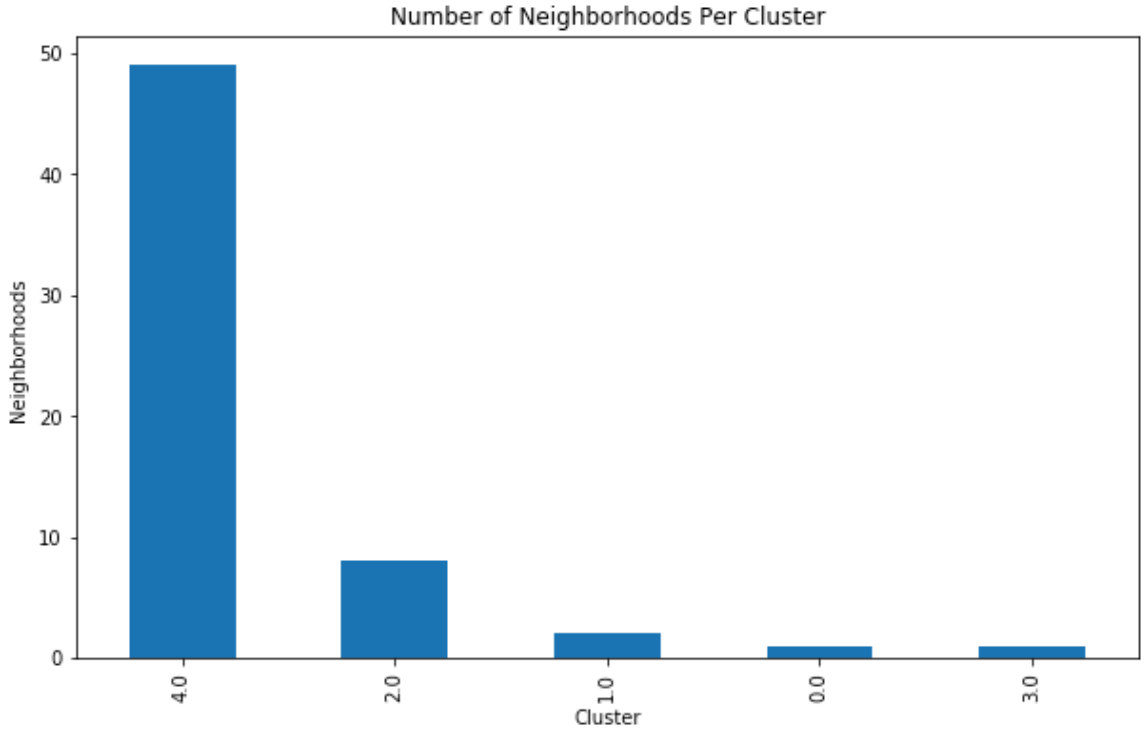
As we all know, the same home exact home in a sought-after area will cost significantly less in a less desirable area. Using the Foursquare API, I will examine the different neighborhoods these homes are located and see if we can identify what areas are more desirable based on the types of venues in the surrounding areas.

As an exploratory starting point, according to foursquare, Staten Island consists of 63 neighborhoods – all of which have been matched to the location listed by Redfin.

According to Foursquare, there are over 800 venues across 180 categories on Staten Island, most neighborhoods have relatively few as Staten Island is know to be one of the more residential boroughs of NYC, however there are neighborhoods that clearly have much more going on.

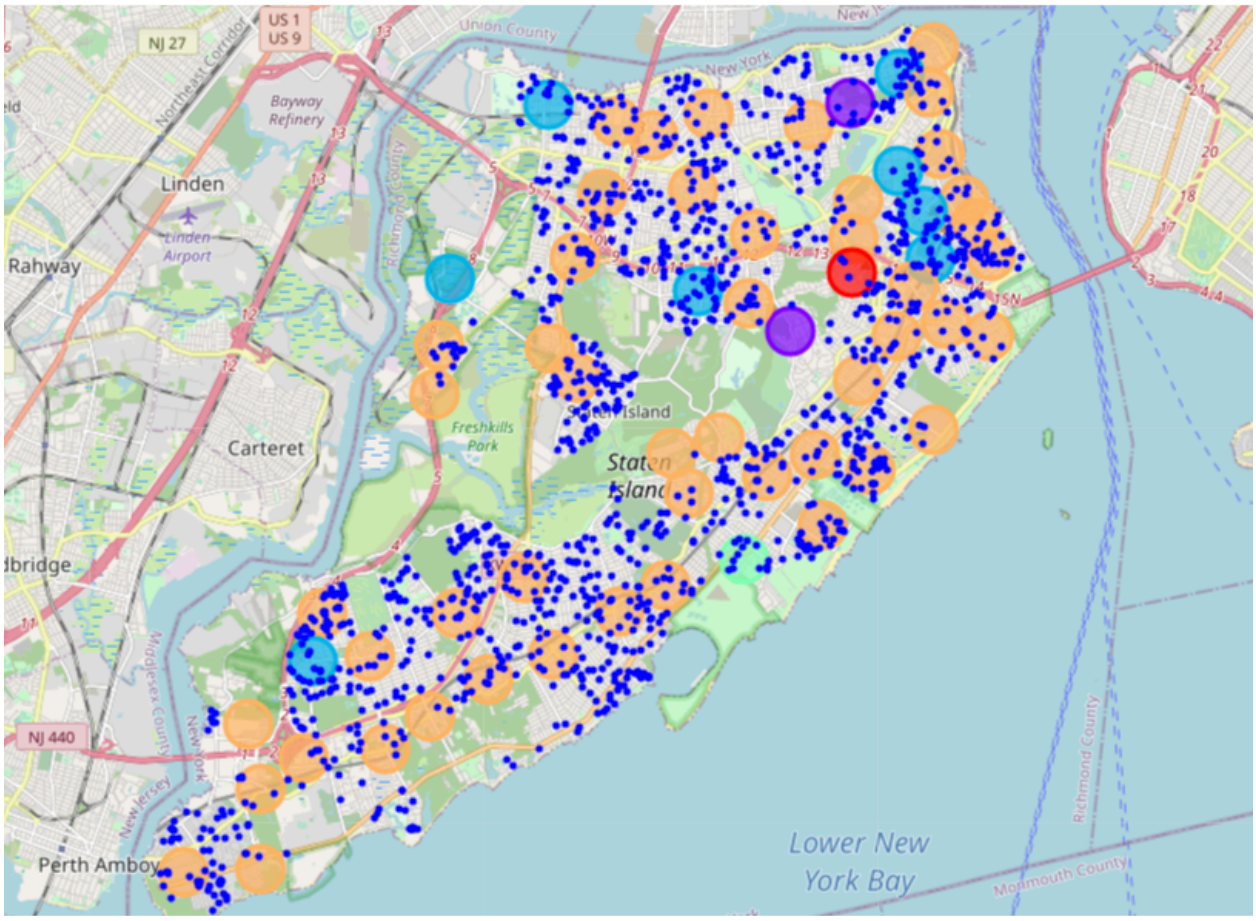


After running the initial cluster evaluation for the Staten Island Neighborhoods it appears 5 clusters are the optimal number for our evaluation



After re-running the analysis we now have our five clusters.

The following map shows the cluster breakdown, as well as the current listings.



Looking at the outputs of these clusters specifically focusing on cluster 4 – the most common cluster, we see that it is what I would consider a typical residential area with a variety of venue options. We often see restaurants, grocery stores, retail, and many of the typical venues we see in suburban settings in many US cities.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	St. George	Clothing Store	Italian Restaurant	Bar	Sporting Good	Pharmacy	Tapas Restaurant	Farmers Market	Burger Joint	Snack Place	Scenic

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
					s Shop						Lookout
2	Stapleton	Pizza Place	Discount Store	Bank	Bar	Sandwich Place	Park	Cosmetics Shop	Residential Building (Apartment / Condo)	Restaurant	Coffee Shop
3	Rosebank	Italian Restaurant	Pizza Place	Grocery Store	Breakfast Spot	Cajun / Creole Restaurant	Sandwich Place	Restaurant	Ice Cream Shop	Discount Store	Pharmacy
4	West Brighton	Coffee Shop	Breakfast Spot	Italian Restaurant	Pharmacy	Bar	Bank	Music Store	Cosmetics Shop	Sandwich Place	Food & Drink Shop

Since we now have both the neighborhood, cluster, and real estate listings combined into one data set we are able to perform a more complex regression analysis.

I created categorical variables base off of the house having a garage or a pool and the neighborhood and clusters in which the listing is located. Because there are 63 distinct neighborhoods, many of which are very similar, having all of them in the model would lower the predictability. So, I removed the neighborhoods that do not appear to have much variance in pricing.

By adding in these categorical data points and using machine learning we have increased our R2 to .65 – an increase of 12ppts.

The new regression formula =
"projected_price" =239324.35233967BEDS'*
39891.2083+'BATHS'*30049.2911+'SQUARE_FEET'*119.361383+'Pool_Y'*13364.6119+'Garage_Y'*30764.3129+'Neighborhood_Graniteville'*-114952.955+'Neighborhood_Port

Richmond'*-147583.2+Neighborhood_Mariner's Harbor*-144458.81+'Neighborhood_Elm Park'*-89199.1133+ 'Neighborhood_Midland Beach'*-15599.8027+ 'Neighborhood_Concord'*-76698.2555+'Neighborhood_Arden Heights'*-124651.041+'Neighborhood_West Brighton'*-105796.391+'Neighborhood_Stapleton'*-102029.582+ 'Neighborhood_Park Hill'*-72887.6042+'Neighborhood_New Brighton'*-117147.577+'Neighborhood_Silver Lake'*52561.6692+'Neighborhood_Arrochar'*72053.3744+'Neighborhood_Emerson Hill'*23061.7507+Neighborhood_Annadale'*17664.9116+ 'Neighborhood_Pleasant Plains'*45574.4585+'Neighborhood_Woodrow'*48093.9137+ 'Neighborhood_Tottenville'*11808.784+Neighborhood_Prince's Bay*48905.3072+ Cluster_Labels_1.0'*24788.0281+ 'Cluster_Labels_2.0'*33664.5601+ 'Cluster_Labels_3.0'*43928.2372+ 'Cluster_Labels_4.0'*-1770.13343

4. Results

After apply this formula to the fixer upper data set we get the following price projections and differences from asking price:

	PROPERTY_TYPE	PRICE	projected_price	price_diff
0	Single Family Residential	499000.00000	572256.92806	-73256.92806
1	Townhouse	419000.00000	685156.97491	-266156.97491
2	Single Family Residential	499000.00000	567004.45379	-68004.45379
3	Single Family Residential	440000.00000	508046.34813	-68046.34813
4	Single Family Residential	599900.00000	344155.22610	255744.77390
5	Single Family Residential	349000.00000	482818.10469	-133818.10469
6	Single Family Residential	375000.00000	466048.61750	-91048.61750
7	Single Family Residential	329900.00000	484822.50830	-154922.50830
8	Single Family Residential	488000.00000	544960.28291	-56960.28291
9	Townhouse	549000.00000	546618.81481	2381.18519

	PROPERTY_TYPE	PRICE	projected_price	price_diff
10	Single Family Residential	360000.00000	541685.57339	-181685.57339
11	Townhouse	658800.00000	637906.26179	20893.73821
12	Single Family Residential	420000.00000	478649.80751	-58649.80751
13	Single Family Residential	269000.00000	506638.51791	-237638.51791
14	Single Family Residential	399000.00000	351654.91216	47345.08784
15	Single Family Residential	519000.00000	625247.53766	-106247.53766
16	Single Family Residential	639000.00000	543316.61617	95683.38383
17	Single Family Residential	489000.00000	572368.86961	-83368.86961
18	Single Family Residential	800000.00000	771382.45496	28617.54504
19	Townhouse	799900.00000	727685.25715	72214.74285
20	Single Family Residential	579000.00000	590609.37671	-11609.37671

The output shows that 14 of the 21 Fixer Upper listings are in fact priced lower than what the model predicts they would be under normal conditions. If I were an investor or house flipper, I have a number of houses that seem like they may be great opportunities. For example #13 seems to be priced at nearly 50% discount. Properties 1, 5, and 10 are also siginificantly discounted. These are homes I would look at in order estimate the actual repair work they required.

5. Discussion

This project was not a total success nor a total failure. The model could be significantly improved with more precise contributing variables into the price of the homes. Items such as school districts, crime data, and other data about the property would allow for more accurate projections. However, despite some weaknesses in the model itself, I believe it does present a solid framework for examining properties and their potential. For this exercise I had a very specific output of evaluating the “fixer-upper” properties. However, this model of evaluation

could be shifted to examining the neighborhoods that are trading at a discount. For example, the vast majority of neighborhoods were in cluster 4, however there were many neighborhoods within that cluster who contributed higher or lower increase or decreases to the price. An argument could be made that these neighborhoods were similar enough in their cluster analysis that a potential home buyer would be receiving the same neighborhood benefits in both neighborhoods, and thus should buy the home in the lower price area, all things being equal.

Again, there are a number of improvements that need to be made on the input side of the evaluation, but I do believe it is conceptually sound.

6. Conclusion:

Overall, the fixer upper listing did, on average, show to be discounted from a similar home under normal conditions. Of course, the real question is are they discounted enough? That is not knowable unless you knew what needed to be done to the homes, which is not part of this analysis. This would just be a starting point and a way to narrow the homes being examined if I were an active real estate investor or developer. I do believe it does function in the regard for which it was designed.