Milestone Report – Prospecting the U.S. Rental Market (Mining for Cash Flow)

Problem Statement

Investing in rental properties can be a difficult process, but it also can be rewarding, offering investors the potential for long-term income followed by the crescendo of a lump sum sale (if you so choose). Successful rental property investors do their homework ahead of time by doing the necessary math to better understand what returns they can expect for a given property in a given region, but many investors are unable to compare returns across vast regions or have little flexibility in their assumptions; paving the wave for costly outcomes.

Solution

A scalable rental property model that can look at vast amounts of data across neighborhoods solves these problems. An optimal model would offer everyday property investors institutional quality analysis. The model would give the user the ability to iterate through various scenarios, change assumptions, and compare returns across broad geographies. My capstone project aims to solve this problem by leveraging my domain knowledge and growing capabilities within the data science discipline.

The Data

The data wrangling process was fairly straightforward overall, which was more than offset by the processing requirements. In the following summary, I list the data sources and primary fields used:

1) Zillow Data
    a. Estimated median home values (per sq. ft and median value)
    b. Estimated rental income (per sq. ft. and median value)
    c. Price-to-rent ratios
    d. Interest Rates

Wrangling/Cleaning Process

The data was relatively clean and was acquired across six csv files that can be downloaded directly from Zillow's data page. All csv files were loaded into Python via a for loop where I modified column names, performed moderate string formatting, and created a list of dataframes that were subsequently merged. The loop structure will allow for efficient loading of additional csv files in the future. The one exception was the interest rate data, which was loaded separately given its different date frequency.

I lost about 46% of the original dataset post-formatting, which sounds alarmingly high, but the majority of the loss can be explained. The median value data pertained to all regions and since the analysis is focusing on rental markets, a sizeable proportion of neighborhoods don't have an active rental market. Secondly, the rental value data began in October 2011 versus 1996 for the median value data, so a large portion was lost historically. The remaining 4% was lost due to insufficient tax data as well as a lack of a full 91 month dataset for a given neighborhood - a percentage I felt acceptable to avoid introducing bias into the dataset. The majority of the nan's were handled in the aforementioned filtering with the exception of some interest rate data having an occasional day gap where I filled using ffill() and bfill(). The wrangling exercise was a bit tricky,

but I managed to match all counties with a robust monthly data set that spanned from October 2011 to present.

2) Property tax data (source: U.S. Census Bureau, NAHB, American Community Survey)
    a. Property tax data for all counties within all 50 states (as of 2014).

Wrangling/Cleaning Process

The county real estate tax data was in xls format with 50 sheets (one for each state labeled using state abbreviation). I created a list of unique states and looped through that list loading each sheet as a separate dataframe in a list. I then concatenated those dataframes and merged with the rental data on state abbreviations and county name.

3) Insurance data (source: Value Penguin)
    a. Average monthly insurance amounts by state (as of 2019).

Wrangling/Cleaning Process

The insurance data was copied into a csv file and loaded into Python. Little cleaning was required.

<u>Processing</u>

After all the data was wrangled and cleaned, each dataset was merged into one large dataset, which was then ready for the processing stage. Processing for this project requires a large time investment due to the domain specific knowledge required to get accurate calculations given inputs. Let's breakdown the domain specific processing components to better understand the process:

Net Cash Flow			
Component		**Assumption**	**Notes**
	Gross Rent	Zillow Estimate	
-	**Vacancy (months)**	One month each year	Flexible
=	**Gross Rent**		
-	**Management Fees**	8% of monthly rent	Flexible
-	**Operating Expenses**	- $250/month (Repairs = $200, Maintenance = $50) - Adjusted by value (* 250)	Flexible
-	**Taxes**	(Zillow Estimate * Multiplier @ 80%) * (County Avg/12)	Flexible multiplier
-	**Interest Payment**	- Zillow estimate on given date = purchase price - 30 year interest rates (equal to monthly average) - 20% down payment	- Function creates amortization table for each region given purchase date - Flexible down payment amount
-	**Insurance**	(Region Med. Value / Statewide Avg. Value) * Statewide average home insurance cost	Update Insurance data each year
=	**Net Cash Flow**		

After-Tax Sale Proceeds			
Component		Assumption	Notes
	Sale Price	Zillow Estimate	Specify sale date in function
-	Sales Commission	5% of sales price	Flexible
=	Gross Sales Proceeds		
-	Mortgage Balance	30 year mortgage rates	Need to implement different tenors
-	Capital Gain Taxes	((Gross Sale Proceeds - (Purchase Price - Cumulative Depreciation)) + (Depreciation Recapture Rate * Cumulative Depreciation)) * Capital Gain Tax Rate	Depreciation Recapture Rate, Capital Gains Tax Rate, and Depreciation time (default = 27.5 years) are all flexible
=	After Tax Proceeds		

Lastly, a return metric such as internal rate of return (IRR) may be calculated to compare returns across regions.

$$\text{IRR} = \text{NPV} = \sum_{t=1}^{T} \frac{C_t}{(1+r)^t} - C_0 = 0$$

where:

C_t = Net cash inflow during the period t

C_0 = Total initial investment costs

r = The discount rate

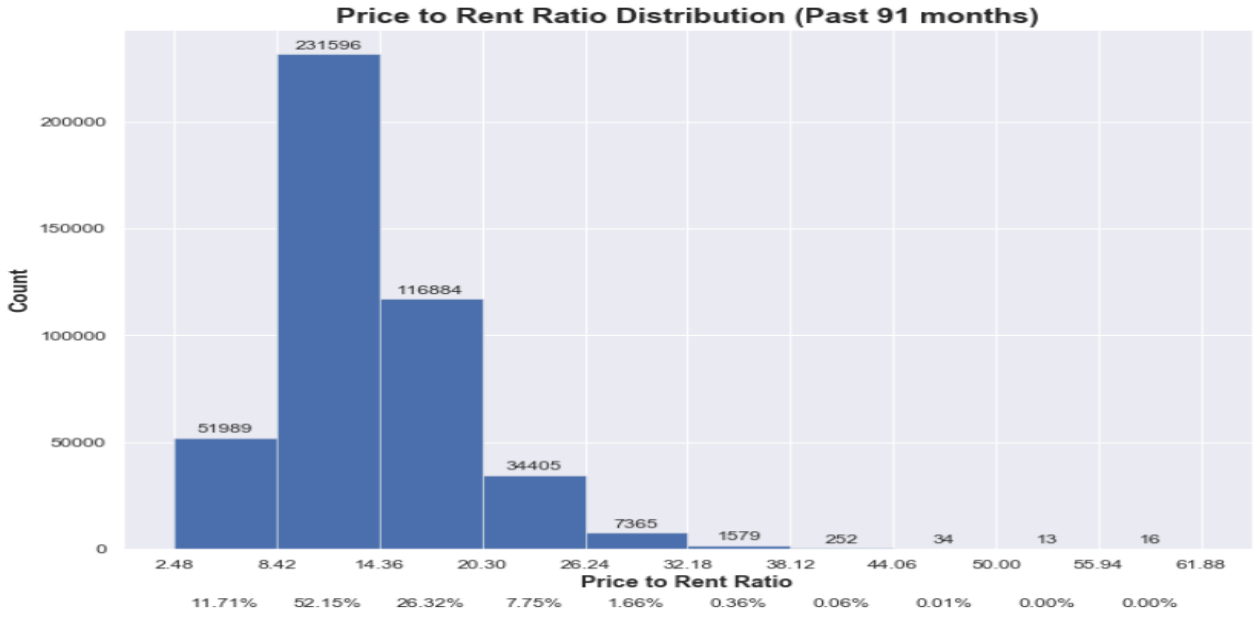
t = The number of time periods

IRR is commonly used for corporate project comparisons and real estate properties.

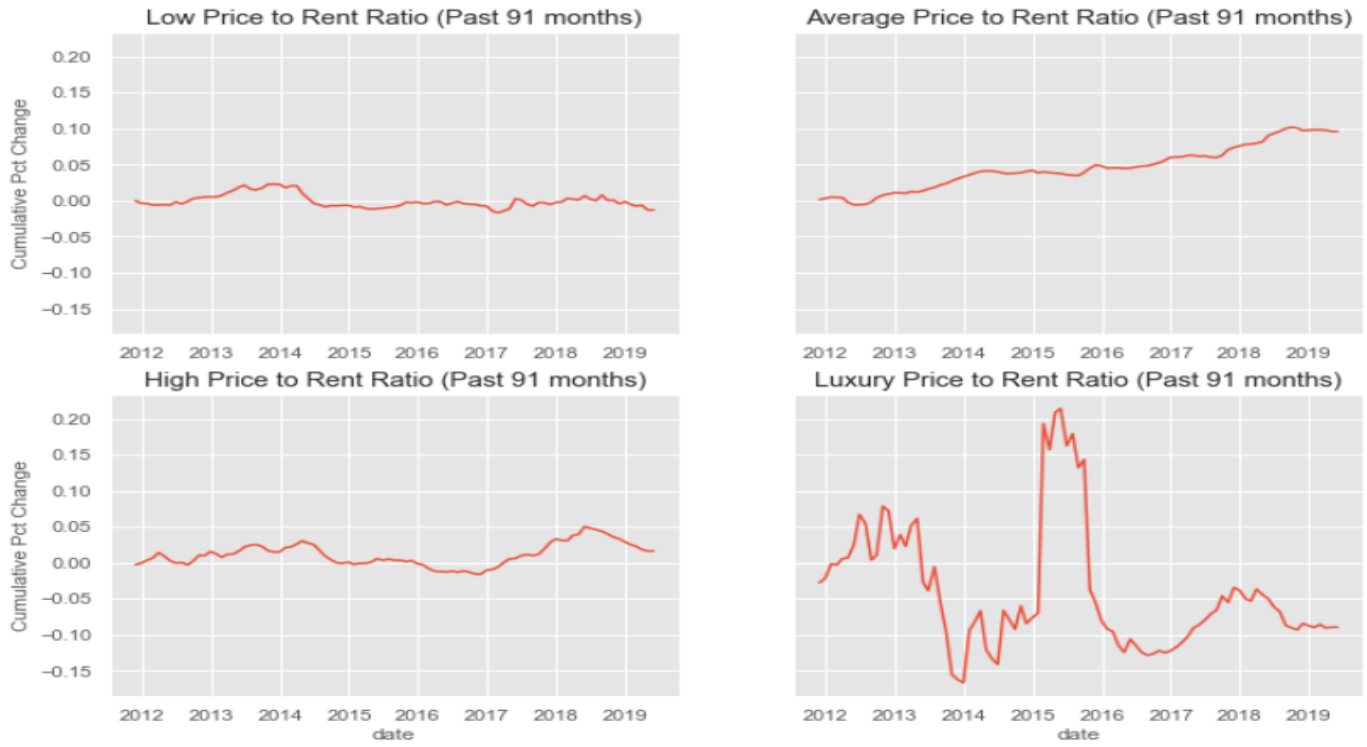
We consider monthly net cash flows on rent and final after-tax sale proceeds (assuming a sale is desired). A custom function was built to provide flexibility with the IRR calculation.

Exploratory Data Analysis

The first area of focus for exploratory data analysis was on the price-to-rent ratio. This data point is provided by Zillow Data, which I've found to provide clean time-series data. The price-to-rent ratio provides a normalized way to compare properties across different geographies, which could explain a large portion of expected returns before expenses are considered. I looked at this ratio through various lenses that provided initial insights that will help aid in building a robust rental property model. We began by looking at the distribution of price-to-rent ratios across all regions that had 91 months of data.

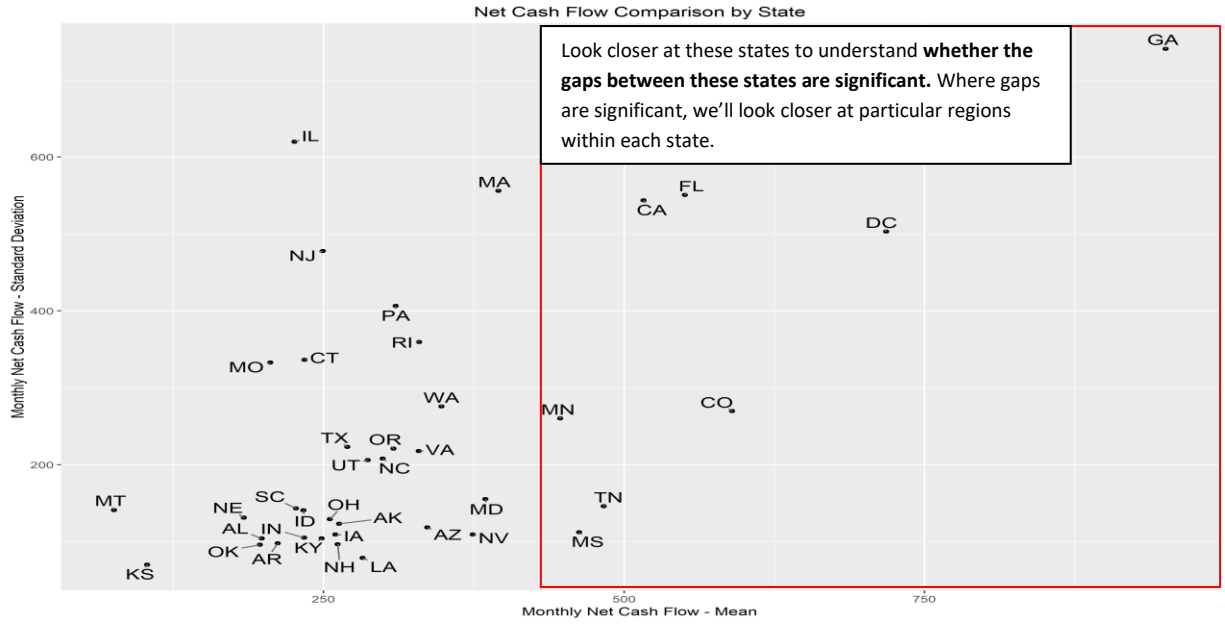


I then partitioned the ratio across four different categories; Low (0-8.42), Average (8.42-20.3), High (20.3-32.18), and Luxury (>32.18) to observe growth trends over time in each category.



It may be tempting to target the lowest price-to-rent ratios exclusively, but the data suggests that you could miss out on price appreciation over time as compared to average price-to-rent ratios. High and Luxury categories appear to provide mixed growth, but are unlikely to provide desirable cash flows over time given the high purchase prices and ceilings on rent that would be required to achieve desirable returns where no buyers exist (unless used as a vacation rental).

Our initial EDA steps gave us some initial clues on where we might focus our attention, but the underlying goal is to understand the return drivers within regions and make them comparable. In order to do this, we must look deeper at our processed data point, which boils down to net cash flow. Below we show estimated mean monthly net cash flows (x-axis) versus one standard deviation differences within each state (46 total) to see if we can further focus our analysis.



Statistical Analysis

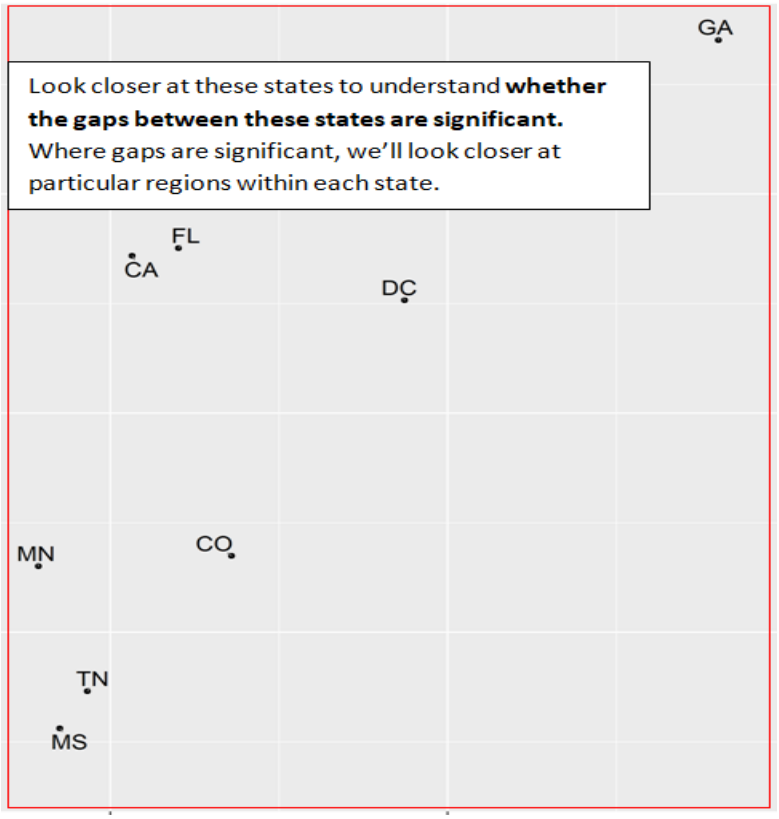
Based on our exploratory data analysis process, we gained the following knowledge that we'd like to apply to our next steps in statistical analysis. The following filters were applied to isolate the most attractive areas given our investment goals:

1) Top states for average monthly net cash flow: CA, CO, DC, FL, GA, MN, MS, and TN.
2) Focus on properties with price-to-rent-ratios below 20.3.
3) Remove regions with average monthly net cash flows less than $100/month. This is the minimum margin of error for net cash flow.
4) Remove regions with median values above $300,000 (latest). I assumed that properties above this price range become cash flow prohibitive.

It was important to use these insights to filter the dataset to more digestible pieces and focus on differences amongst regions within the most attractive markets that meet typical cash flow property objectives. Based on the filters above, we narrowed the focus from 4,303 regions to 385. The next step was to then exhaustively compare each region and observe differences in z-scores.

```
Significant of total combinations: 21 of 28
```

	feature_combo	z_score	p_value	reject_null
16	(DC, MS)	41.974147	0.0000	Yes
17	(DC, TN)	38.154405	0.0000	Yes
1	(CA, DC)	35.220434	0.0000	Yes
15	(DC, MN)	29.599508	0.0000	Yes
13	(DC, FL)	20.126311	0.0000	Yes
23	(GA, MS)	14.617400	0.0000	Yes
3	(CA, GA)	11.075889	0.0000	Yes
22	(GA, MN)	11.071221	0.0000	Yes
24	(GA, TN)	10.897475	0.0000	Yes
11	(CO, MS)	8.919605	0.0000	Yes
7	(CO, DC)	8.475178	0.0000	Yes
18	(FL, GA)	7.996370	0.0000	Yes
14	(DC, GA)	7.840824	0.0000	Yes
27	(MS, TN)	7.346001	0.0000	Yes
10	(CO, MN)	6.506391	0.0000	Yes
0	(CA, CO)	6.173050	0.0000	Yes
5	(CA, MS)	5.966631	0.0000	Yes
12	(CO, TN)	5.901506	0.0000	Yes
20	(FL, MS)	5.575971	0.0000	Yes
8	(CO, FL)	4.477124	0.0000	Yes
25	(MN, MS)	3.854824	0.0002	Yes



Next Steps

Given we've filtered our search results down to eight states, the logical next steps would be to begin examining and comparing the regions ("neighborhoods") within each state. Neighborhoods is the most detailed level of data we have, which will allow us to examine the myriad of assumptions we chose during the processing stage as well as give us some confidence around how useful this model is. We'll begin by looking at Georgia, which looked like the most attractive state post-filter.

RegionName	mean	std	count	samp_var
Bolton	938.902	220.5	92	528.479
Georgetown	506.435	43.6713	92	20.7302
Midwest Cascade	742.965	103.58	92	116.618
Princeton Lakes	744.201	82.2381	92	73.512

I decided to select Bolton as the first neighborhood to examine closer via looking at individual properties on Zillow. After some time spent, Bolton does appear to be quite attractive with opportunities abound. Many homes are priced in the mid 200's with homes across the street that are well above $400K. Additionally, numerous townhome and new builds are being developed nearby. The model focused us immediately on an attractive area, which bodes well given the initial pass. Closer examination will be required and after looking at some expense metrics, I'd suspect that the mean monthly net cash flows are overstated, but not egregiously. Specifically, the size adjustment we make for median homes in regions relative to the state average appears to be aggressive. Overall, a promising first look at this basic model.

Machine Learning Application

For simplicity and to gain broader understanding, we began by looking at the most attractive states and then examined specific neighborhoods within those states. The shortcoming of this approach is we ignore potentially attractive neighborhoods that get lost in the state average calculations. Wouldn't it make more sense to just look at the 4,000+ neighborhoods regardless of state? This is where machine learning applications can help as we look at the data in more abstract granularity.

Since we aren't predicting prices (Zillow has already used ML to estimate rental and home values) and we aren't predicting labels, the insights gleaned from this data will come from unsupervised learning techniques. By pointing to specific neighborhood groupings using a **clustering model,** we can identify neighborhoods which offer similar opportunities as well as identify varying investor strategies. The goal would be first to create the optimal number of clusters, then begin to develop custom labels for the clusters that best associate the group with the investment strategy (i.e. vacation rentals, college rentals, flipping neighborhoods, etc.). Furthermore, we'd be able to see holistically what areas can offer as potential returns for varying strategies.

Overall, this model offers immense value for focusing an investor's attention to specific areas (as well as allowing them to compare them) in a more systematic manner. Due to the many factors that determine the actual transaction of buying a property, the finer details are outside the scope of the data we have for this model. Good old 'boots to the ground' work is needed to finalize the purchase, but this model can confidently allow you to focus on ideal neighborhoods for your investment strategy.

Clustering Model

The ideal use of this clustering model would be for property types, meaning we could run this process iteratively for Single Family Homes (SFH), Multi-Family Homes (MFH), Condos/Townhomes, and Commercial properties. The focus of this project is on Single Family homes as a proof of concept. The following filters were applied to create a realistic subset of neighborhoods for SFH investment:

Initial Filters

1) Home values less than or equal to $500,000 (-1,283 neighborhoods)

Note: These filters reduce our neighborhood count from 4,303 to 3,202.

Features

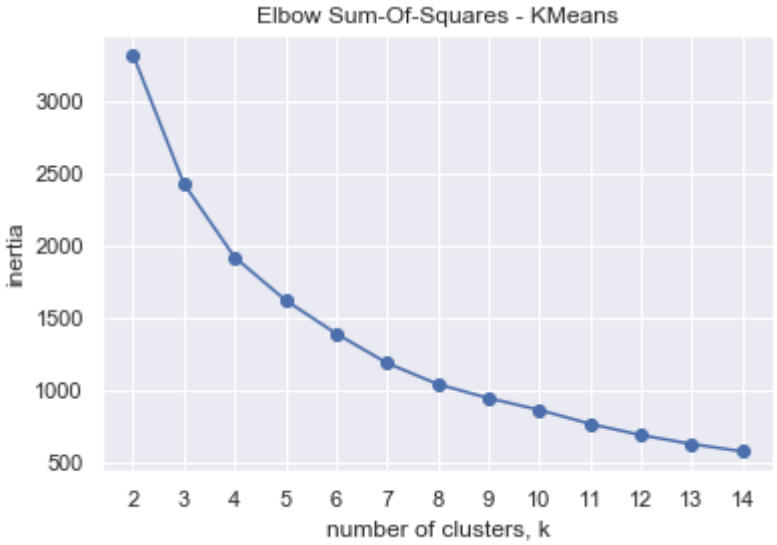
Features
1) Net Cash Flow
2) Neighborhood Zillow Home Value Estimate - Single Family Residence

Note: Given each feature has different scales, I first scaled all features using StandardScaler.

Future features to consider adding/testing, but not implemented: Home Building Permits/Starts, Crime Data, School Ratings, Population Growth, and Unemployment Data.

Optimizing clusters and Choosing Model Type

I tested a series of models to see which produced the best average silhouette scores given a parameter or feature. I didn't run a hierarchical model due to the sheer number of neighborhoods (over 3,000), but it's interesting to look at for individual states. Additionally, I ran an elbow sum-of-squares graph for further clarity for K-Means. The results are summarized below:



Model	Metrics Tested	Silhouette Average	Notes
K-Means	n_clusters = 2-10	0.41 (2 clusters), 0.38 (3 clusters), 0.36 (4 clusters), 0.35 (5 clusters)	Weak Structure
Affinity Propagation	damping = 0.5 - .95	0.77 (0.90)	Good Structure
Spectural Clustering	n_clusters = 2-10	0.26 (2 clusters), 0.24 (3 clusters), 0.24 (4 clusters), 0.24 (5 clusters)	Poor Structure
Agglomerative Clustering	n_clusters = 2-10	0.41 (2 clusters), 0.38 (3 clusters) 0.28 (4 clusters), 0.28 (5 clusters)	Weak Structure

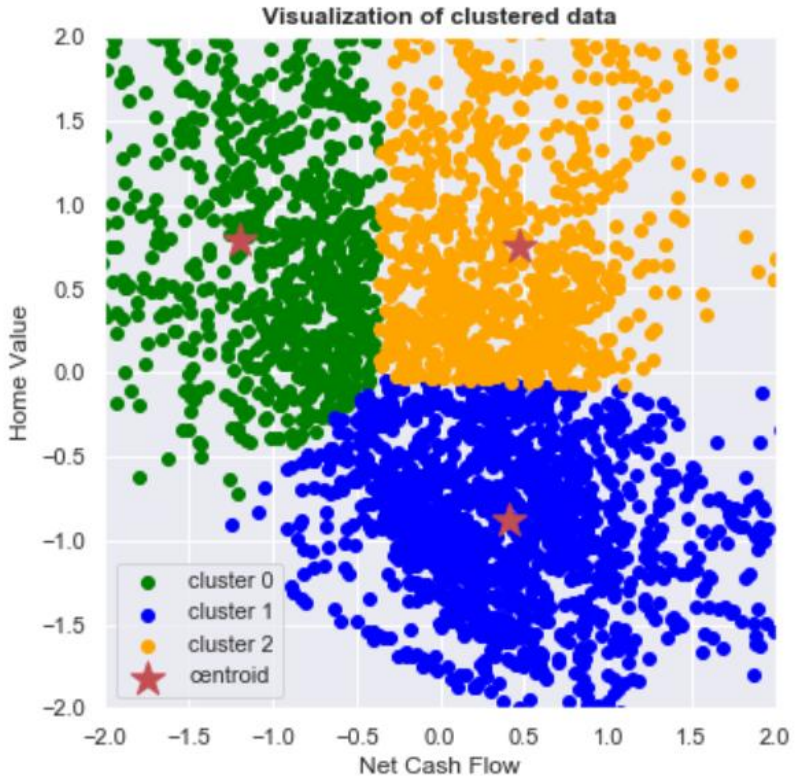
Model specifications

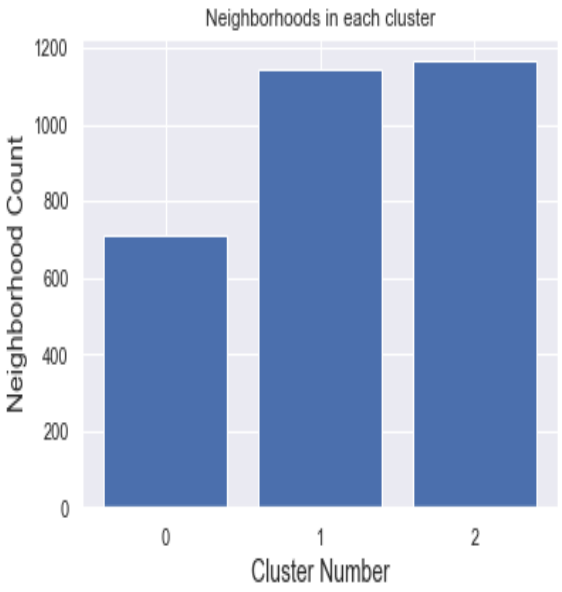
Given the results I made the following choices:

Model Type = K-Means. Provided highest silhouette scores of all models tested.

Clusters = 4. Silhouette score isn't much lower than two clusters and gives more granularity.
Features = 2. Tested varying combinations of features.

Cluster visualization





Application

We now have a useful process for combining features to provide useful groupings of neighborhoods for which to focus our analysis. Additionally, we can constantly refine or add our features as we get more data or notice patterns in our search process. Here's a short example to show how this model could be effectively used to narrow a rental property search:

1) Subset top neighborhoods for estimated net cash flow in MN (market I have domain knowledge in) to see if any cluster patterns can be observed. Compare this top 10 to the top 10 across all states.

RegionName	State	net_cf	cluster
East Isles	MN	1242.0	0
Whittier	MN	727.0	1
Southeast Como	MN	589.0	2
Hawthorne	MN	456.0	1
Mckinley	MN	439.0	1
Willard Hay	MN	417.0	1
Marcy Holmes	MN	411.0	2
Folwell	MN	409.0	1
East Phillips	MN	405.0	1
Webber-Camden	MN	392.0	1

RegionName	State	net_cf	cluster
East Isles	MN	1242.0	0
Heritage Hills	NY	955.0	0
Indian Creek	CO	935.0	2
Brookwood	GA	877.0	2
Meadow Hills	CO	798.0	2
Cadwalader & Hillcrest	NJ	782.0	1
Chambersburg	NJ	777.0	1
Wynnefield	PA	767.0	1
Sable Ridge	CO	759.0	2
Bolton	GA	754.0	2

Observations from top 10 lists

a. Cluster's 0 and 2 contain some net cash flow outliers. Upon closer inspection, all outlier neighborhoods had a common theme; expensive neighborhoods with a large inventory of condos

which carry high HOA fees. HOA fees are not accommodated for in the model as that's a property level data point that cannot be incorporated, but it likely not worth implementing considering the HOA fee will make those neighborhoods a lot less attractive.

b. Cluster 1 appears to be the best cluster to focus on given a spot check on each neighborhood confirmed reasonable accuracy.

2) Subset cluster 1 and filter based on individual preferences (not captured in model) to isolate the most attractive neighborhoods for your given strategy.

Summary

Clearly, the real estate market is a complex market that has many variables embedded. This project began to delve into this market by looking at neighborhoods, but doesn't contain transaction level data – only averaged data for each region. Given this dynamic, this model only points you to potentially attractive neighborhoods, which is a good start considering there are over 4,000 distinct neighborhoods in the Zillow data. Furthermore, additional features would likely improve this model (as mentioned in the features section), but initial results appear to be helpful and neighborhood by neighborhood analysis is beyond the scope of this project. Rather, an investor could use this model to filter opportunities based on individual preferences and focus on individual transactions to complete the process.

Companies who possess the detailed property level data have a distinct advantage as they could take this analysis deeper and select individual properties systematically satisfying a wide range of criteria.

The Good	The Bad
• Provides some clarity amongst 4,691 potential neighborhoods to invest in. • Flexibility of model leaves room for various lenses on data • Cluster 1 provided solid initial leads	• Lack of property-level data is a disadvantage. • Cluster shapes have no separation indicating model could benefit from additional features (crime, school ratings, etc.)