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

Problem Statement

Investing in rental properties can be a difficult process, but is also can 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, 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 majority 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 there. The remaining 4% was lost due to insufficient tax data and 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 data set that spanned from 2011-10-31 to 2019-05-30, amounting to ~470,000 rows.

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

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 (Value Penguin)
   a. Average monthly insurance amounts by state

Wrangling/Cleaning Process

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

Processing

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	Flexible
=	Gross Rent		
-	Management Fees	8% of monthly rent	Flexible
-	Operating Expenses	- $250/month (Capital Expenditures = $200, Maintenance = $50) - Adjusted by value (* 250)	Flexible monthly amount
-	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 Median Value / Statewide Average) * Statewide average	Update Insurance data as needed
=	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	
-	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		

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

where:

$C_t =$ Net cash inflow during the period t

$C_0 =$ Total initial investment costs

$r =$ The discount rate

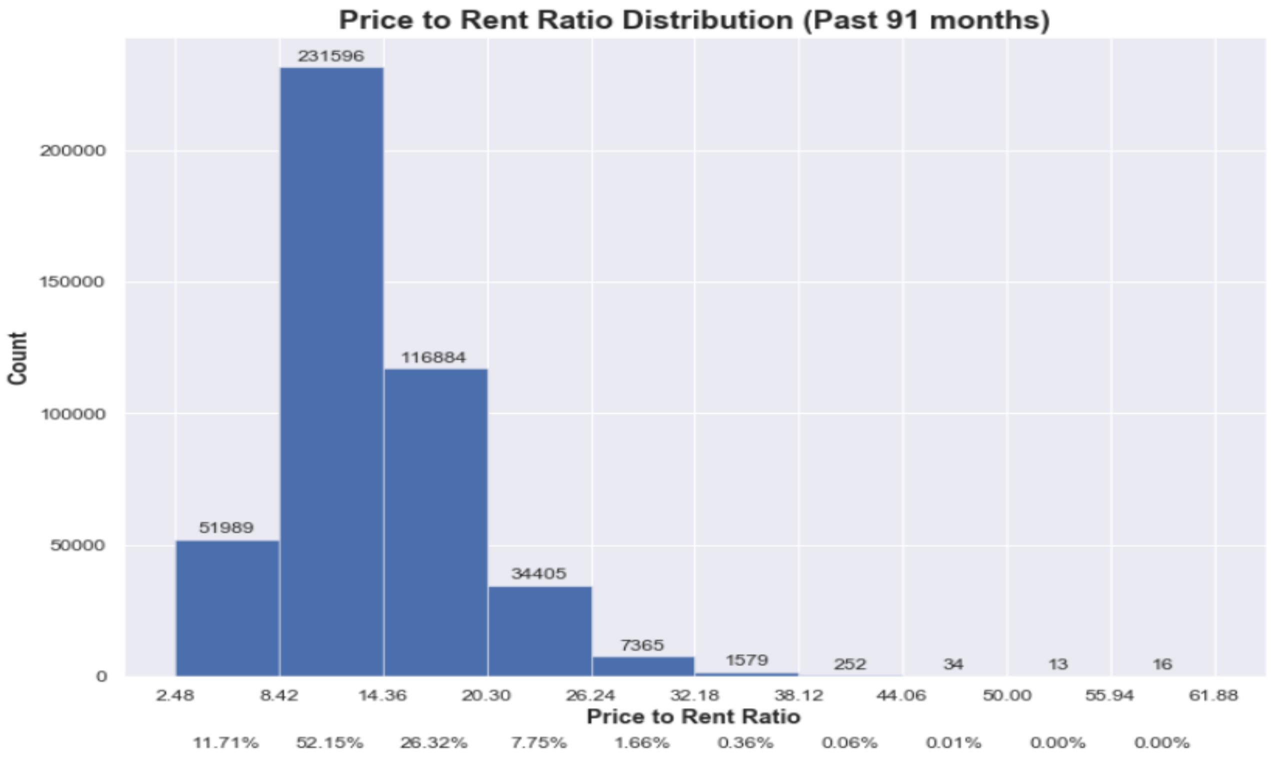
$t =$ The number of time periods

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

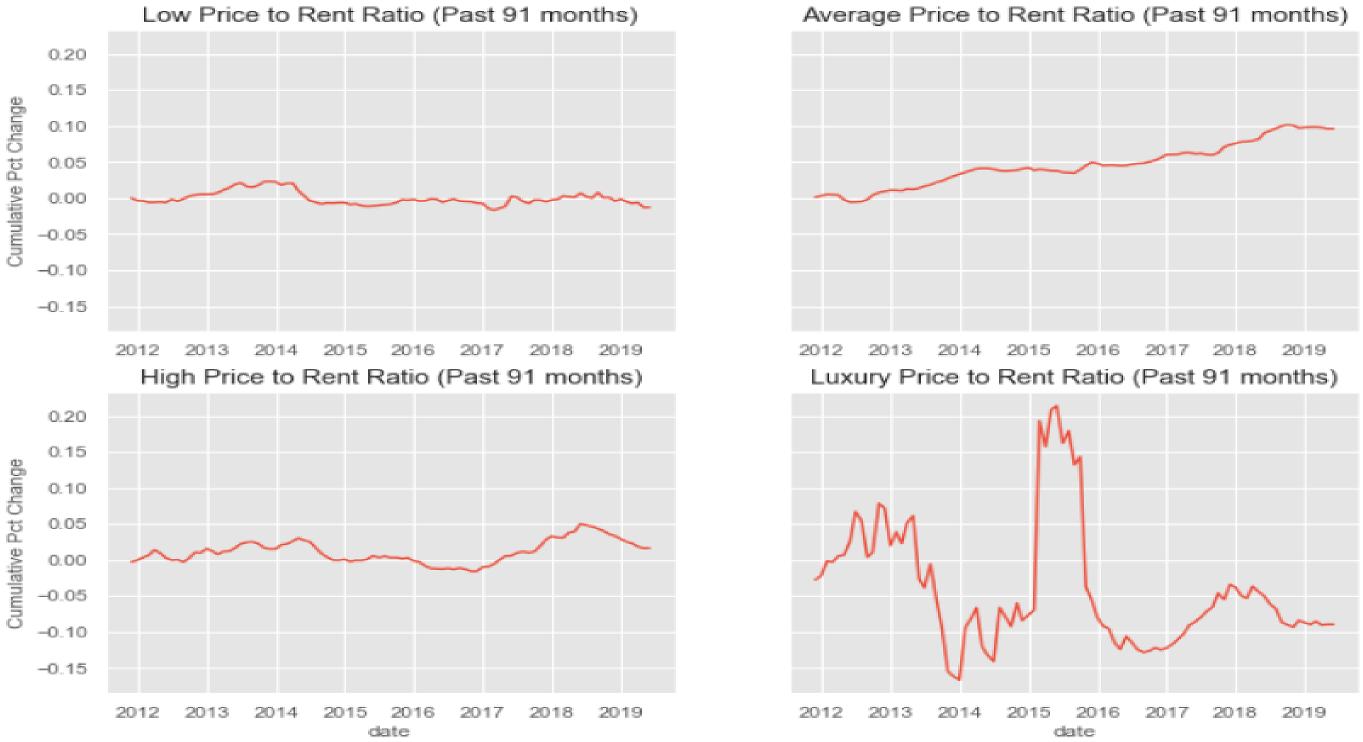
We consider monthly net cash flows on rent and final after-tax sale proceeds (assuming a sale is desired). A custom function was built to provide flexibility with the IRR calculation.

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 will 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.


Low Price to Rent Ratio (Past 91 months)

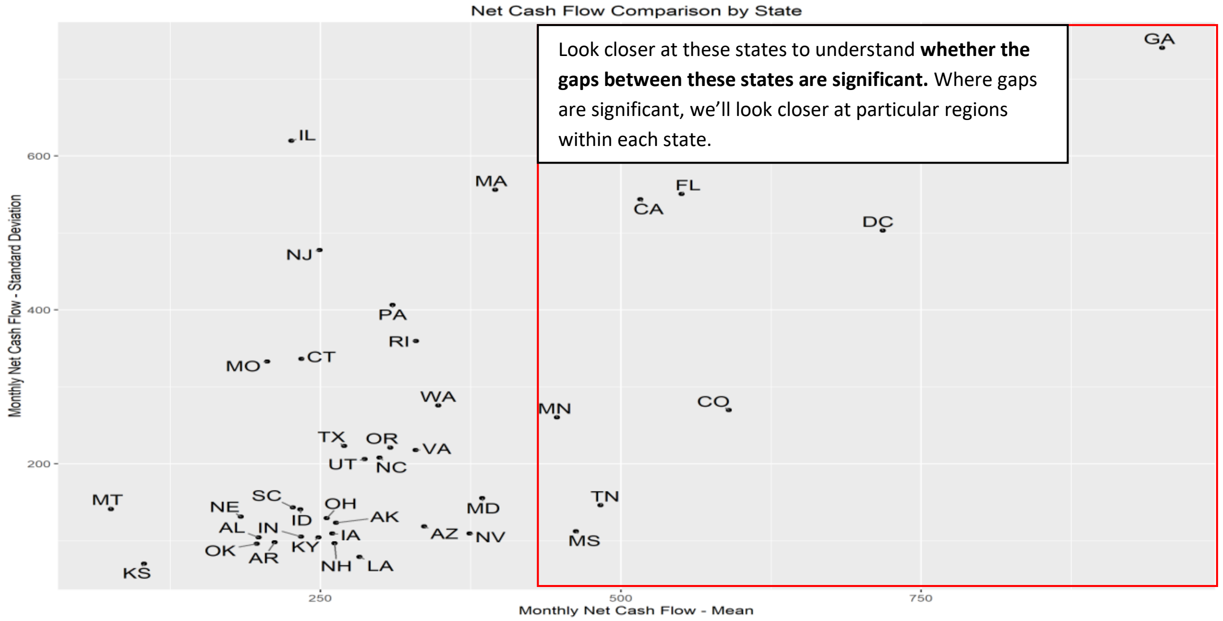
Average Price to Rent Ratio (Past 91 months)

High Price to Rent Ratio (Past 91 months)

Luxury Price to Rent Ratio (Past 91 months)

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.


Net Cash Flow Comparison by State

Look closer at these states to understand **whether the gaps between these states are significant.** Where gaps are significant, we'll look closer at particular regions within each state.

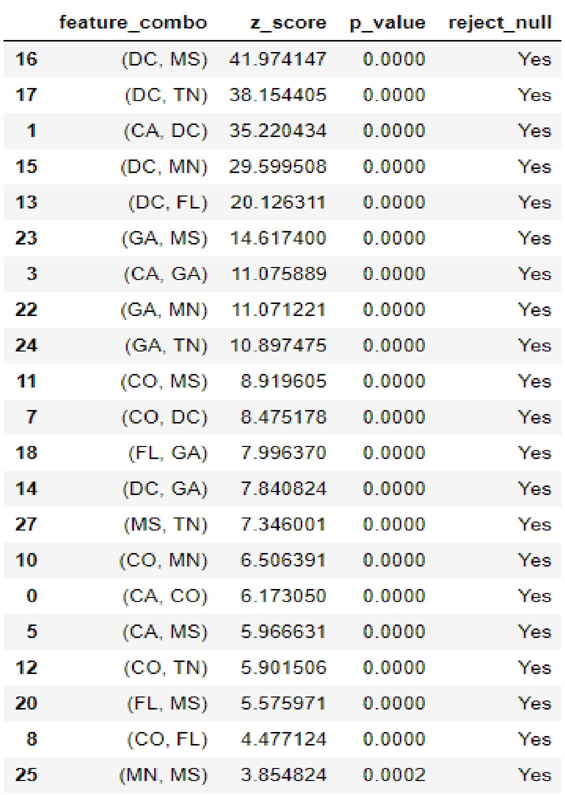
Statistical Analysis

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:

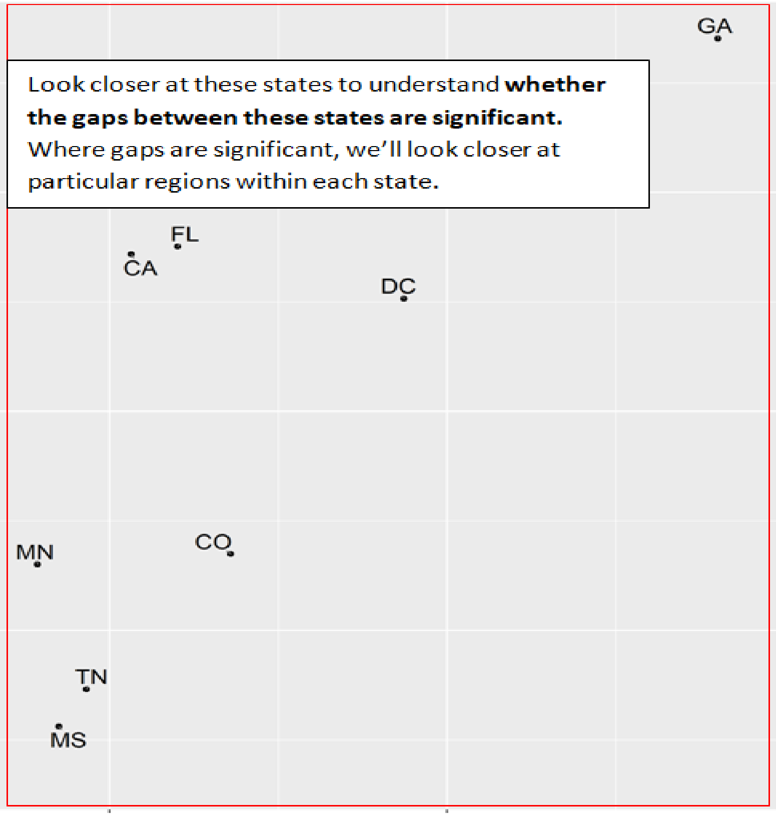
1) Top states for average monthly net cash flow: CA, CO, DC, FL, GA, MN, MS, and TN.
2) Focus on properties with price-to-rent-ratios below 20.3.
3) Remove regions with average monthly net cash flows less than $100/month. This is the minimum margin of error for net cash flow.
4) Remove regions with median values above $300,000 (latest). The client specified they could afford to put $60,000 down and 20% down is a common requirement for rental properties.

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 the client's needs. 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
```



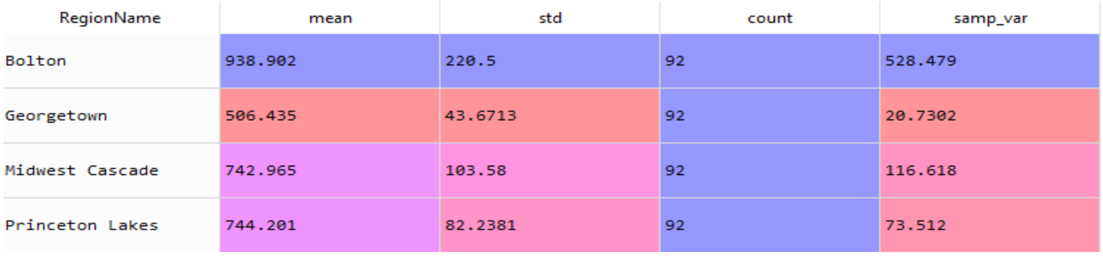
	feature_combo	z_score	p_value	reject_null
16	(DC, MS)	41.974147	0.0000	Yes
17	(DC, TN)	38.154405	0.0000	Yes
1	(CA, DC)	35.220434	0.0000	Yes
15	(DC, MN)	29.599508	0.0000	Yes
13	(DC, FL)	20.126311	0.0000	Yes
23	(GA, MS)	14.617400	0.0000	Yes
3	(CA, GA)	11.075889	0.0000	Yes
22	(GA, MN)	11.071221	0.0000	Yes
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



Look closer at these states to understand **whether the gaps between these states are significant.** Where gaps are significant, we'll look closer at particular regions within each state.

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 in 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 the model.