Analysis and Recommendation for SU REIT Investment Opportunities

By: Nate Hoffelmeyer
August 24, 2018

I. Introduction

I.I Background

Investing is a data driven field. Countless simulations and models are run to determine where capital must flow, all to boost returns for investors depending on future funds for college tuition, vacations, retirement, or even for instituions who are aiming to fund pensions, etc.

Real Estate is one very popular asset, and is often regarded as one of the fastest ways to build wealth in America. Real Estate Investment Trusts (REIT) are companies who generally own, operate, or finance income generating real estate[i]. In particular, REITs might be interested in multi-tenant complexes for rental purposes, resorts & hotels, or in some case single family homes that can be easily rented, or flipped for profit soon.

I.II Data and Problem Specification

This paper will look at a time series of housing data from Zillow[iiiii]. The objective is to provide an analysis and recommendation for three zip codes, such that the Syracuse University REIT can decide where to invest next. The data, described below, was distilled to a working set, with NaN (empty or missing) values removed.

The initially cleaned data set contained 131,500 records of average home price by zip code across the United States. This data spanned from April of 1996 through June of 2018. The formal analysis selected the state of Connecticut to narrow the problem, based on research from realtor.com[iv]. In the realtor.com research, they identified the areas with the highest price increases going back as far as 2008.

Using this as a spring board for analysis, the paper aims to guide SU REIT on the top three areas to invest, and provide a prediction of what the area median average price will be 3 months into the future.

	RegionID	RegionName	SizeRank	1996-04	1996-05	1996-06	1996-07	1996-08	1996-09	1996-10	...	
count	13150.000000	13150.000000	13150.000000	1.315000e+04	1.315000e+04	1.315000e+04	1.315000e+04	1.315000e+04	1.315000e+04	1.315000e+04	...	1.
mean	80965.863650	47473.394525	7139.665932	1.204022e+05	1.205257e+05	1.206477e+05	1.207666e+05	1.208972e+05	1.210484e+05	1.212458e+05	...	2.
std	33354.707112	29862.588767	4353.439216	8.693276e+04	8.709021e+04	8.724795e+04	8.741076e+04	8.759811e+04	8.782421e+04	8.810979e+04	...	3.
min	58196.000000	1001.000000	1.000000	1.140000e+04	1.150000e+04	1.160000e+04	1.180000e+04	1.180000e+04	1.200000e+04	1.210000e+04	...	1.
25%	66626.250000	20755.250000	3371.250000	7.050000e+04	7.070000e+04	7.090000e+04	7.100000e+04	7.110000e+04	7.130000e+04	7.150000e+04	...	1.
50%	77520.500000	45119.000000	6868.500000	1.014000e+05	1.015000e+05	1.015000e+05	1.016000e+05	1.017000e+05	1.018000e+05	1.018000e+05	...	2.
75%	91030.500000	75237.750000	10818.750000	1.450000e+05	1.449000e+05	1.451000e+05	1.452000e+05	1.452000e+05	1.453750e+05	1.456000e+05	...	3.
max	753844.000000	99901.000000	15245.000000	3.676700e+06	3.704200e+06	3.729600e+06	3.754600e+06	3.781800e+06	3.813500e+06	3.849600e+06	...	7.

II. Obtain

Data was obtained from a files link (see endnotes) provided by Professor Jon Fox. The initial data set contained the RegionID, RegionName, City, State, Metro, CountyName, SizeRank, and average price for the zip code (RegionName) dating from 1996-04 through 2018-06.

	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1996-05	1996-06	...	2017-09	2017-10	2017-11	2017-12	2018-01	
1334	60390	6606	Bridgeport	CT	Stamford	Fairfield	1335	87700.0	87700.0	87600.0	...	179300	181000	182600	184200	185800	18
3795	60388	6604	Bridgeport	CT	Stamford	Fairfield	3796	98100.0	97900.0	97600.0	...	206600	211500	215400	216600	214400	21
4743	60393	6610	Bridgeport	CT	Stamford	Fairfield	4744	79700.0	79700.0	79700.0	...	157000	158300	159900	161800	164500	16

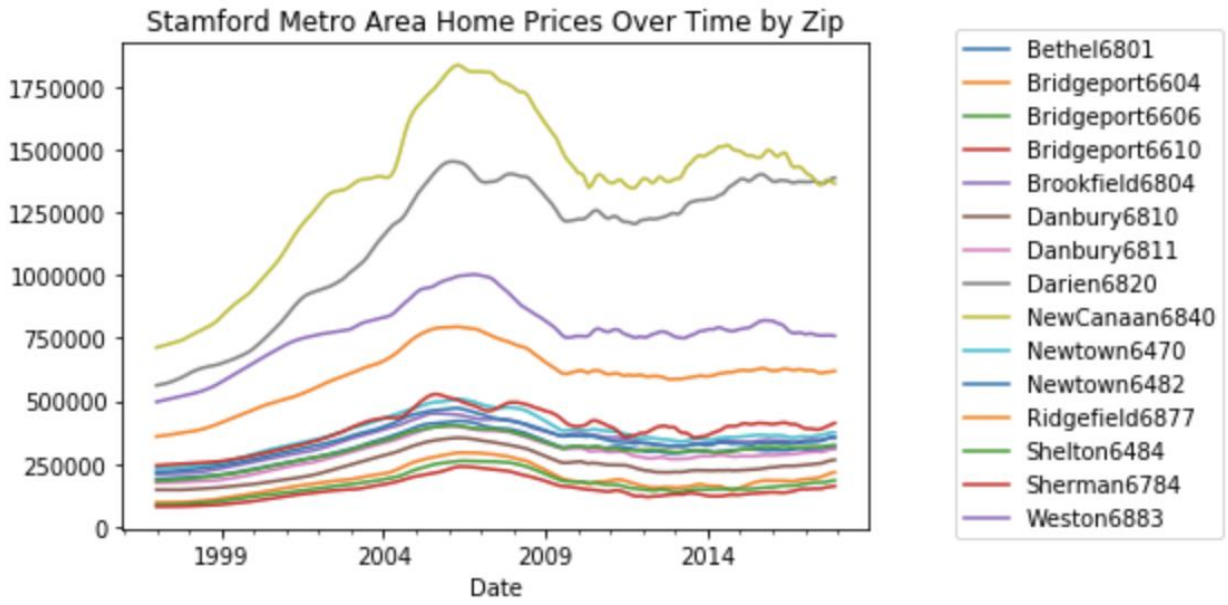
No additional data was collected for the purposes of this analysis. The full data set can be found in the files link in the endnotes.

III. Explore and Scrub

After obtaining the data described previously, exploration and scrubbing commenced using statistical analysis tools and libraries from the programming language Python. The first task was to filter down the data set to the top area as mentioned in the realtor.com study: Bridgeport CT.

Bridgeport was identified as a part of the Stamford metro area. Accordingly, knowing that this area specifically has seen high returns (Bridgeport, that is) but given the research around the Bridgeport area having already been made public ,and consequently perhaps having missed some opportunity to obtain max ROI (return on investment) in just Bridgeport alone; the analysis was broadened to the Stamford metro area. There, the analysis looks to see if there are 3 zip codes throughout the entire metro area that might outperform, and provide the SU REIT with high ROI.

Further refining the analysis, the data was filtered to the Stamford area from 1997 through 2018, spliced and melted to produce a time series, and then charted to show how home prices have risen over time:



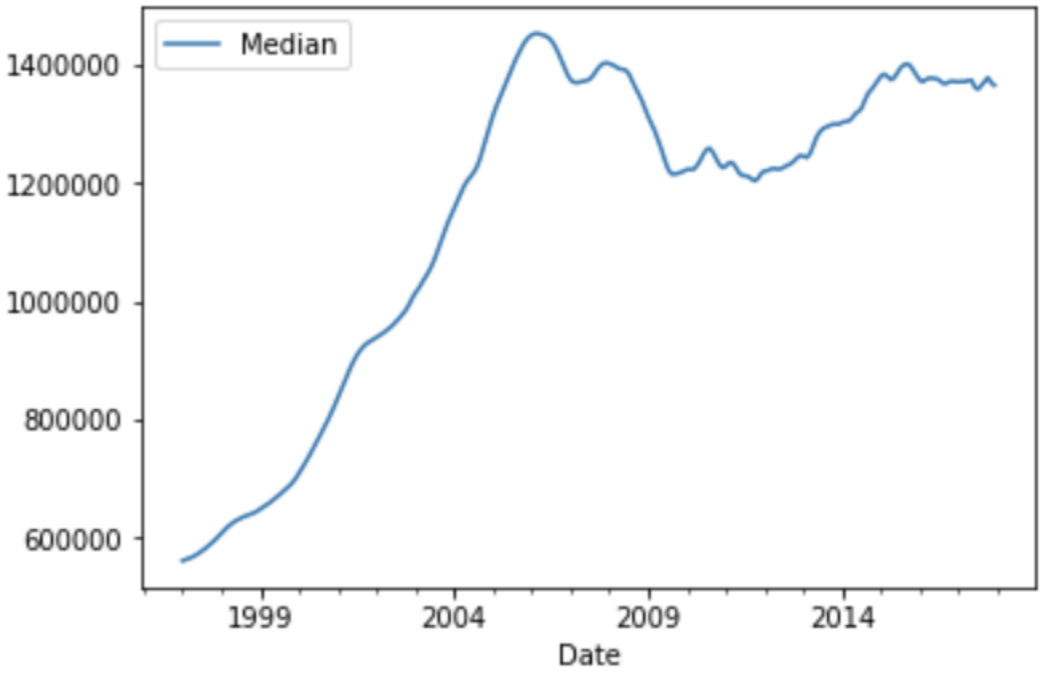
The plot indicates there are clearly four highly increasing in price areas over the last 20 years of analysis (1997-2017: New Canaan, CT 06840, Darien, CT 06820, and Brookfield, CT 06804, plus Bridgeport, CT 06604 look to have about the same increase over the period relative to where they started, with Brookfield maybe slightly higher. This is interesting, because the realtor article indicated highest profitability in the Bridgeport areas, and yet our data shows the top 3 growth areas don't include Bridgeport, assuming Brookfield is indeed a bit higher in the growth realm as mentioned above.

Having explored our data briefly to determine the three areas with the highest previous growth in home prices, the exploration now moves to modeling the time series and producing a forecast of prices for these areas.

I.V. Model and Interpret

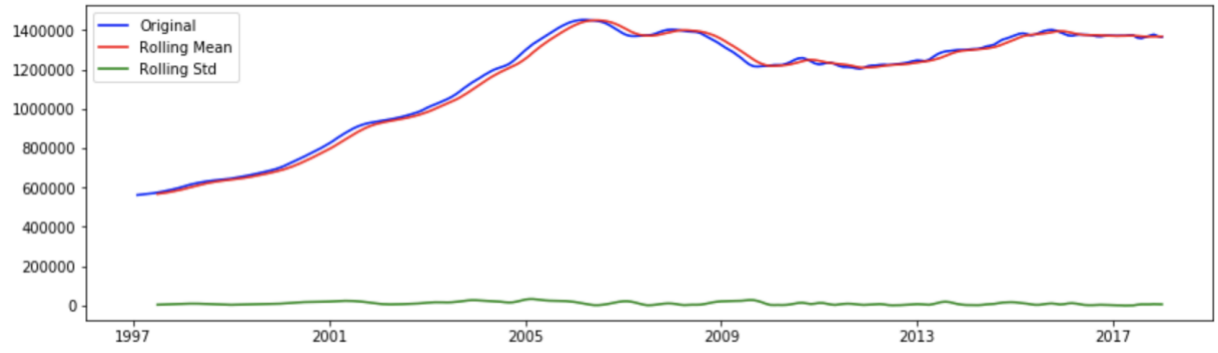
To begin modeling time series, the first thing to look at is whether our data has trend, is seasonal, or if it is stationary. This will help to choose the appropriate model.

Having added a median column to the data set such that the analysis aims to predict / forecast median home prices across the top three zip codes identified previously, a basic plot shows there is unquestionable trend in the data.



To be extra sure, analysis of rolling mean can commence; as can analysis through ADF (augmented dickie fuller testing). If the ADF results show a significant p-value (lower than 0.05), and stationarity in the data consequent of that, then the analysis is able to reject the null hypothesis that there is no stationarity – and can commence normal analysis.

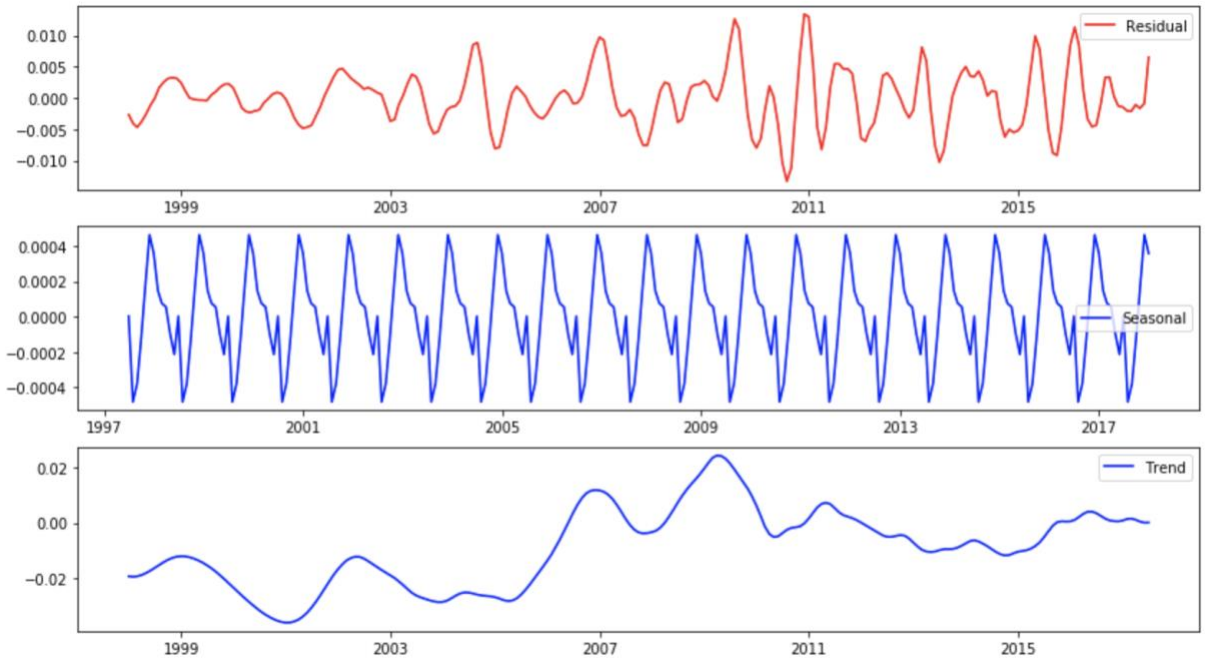
```
Perfom Dickey fuller test
Test Statistic                  -2.199638
p-value                          0.206377
#Lags Used                      15.000000
Number of Observations Used    236.000000
Critical Value (1%)             -3.458366
Critical Value (5%)             -2.873866
Critical Value (10%)            -2.573339
dtype: float64
```



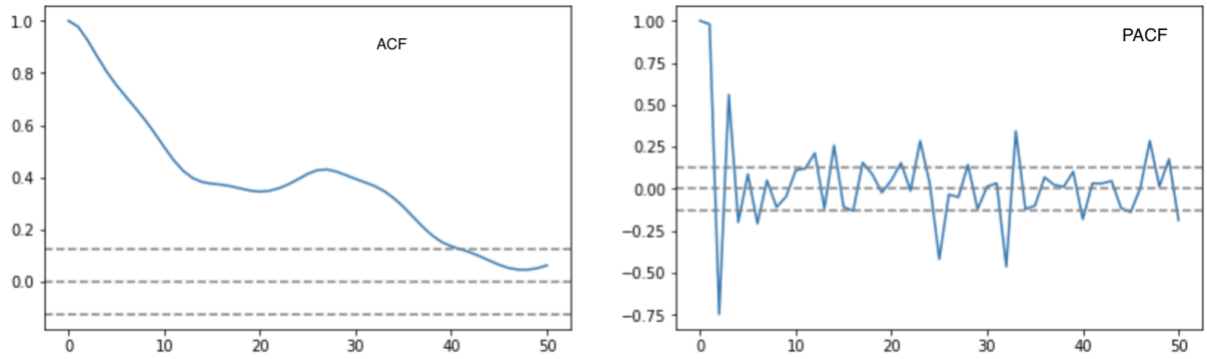
The ADF above shows a very high p-value, and consequently the data is not stationary (the null cannot be rejected here). Further, analysis of the rolling mean shows that there is significant

trend over time, with further analysis of the rolling standard deviation indicating some, though not major, fluctuations: another telltale sign of trend and seasonality in data.

The above is important, because in order to predict trend without overfitting or error, the series must be differenced.



After rigorous testing and analysis, the data being taken the logarithm of, the moving average for, differenced, charted; ARIMA modeling is decided as the method of choice. ACF and PACF charts are drawn to show where the model needs to be set, though later analysis rendered obsolete.



The dotted lines show the confidence interval, and can be used to guide where the model values are set.

The data is final split into training and testing, run with set p values (see accompanying python notebook), and the model summary concludes the data has a lower AIC than other iterations. The Akaike Information Coefficient is a measure of error, and the lower the value the better the model.

ARMA Model Results

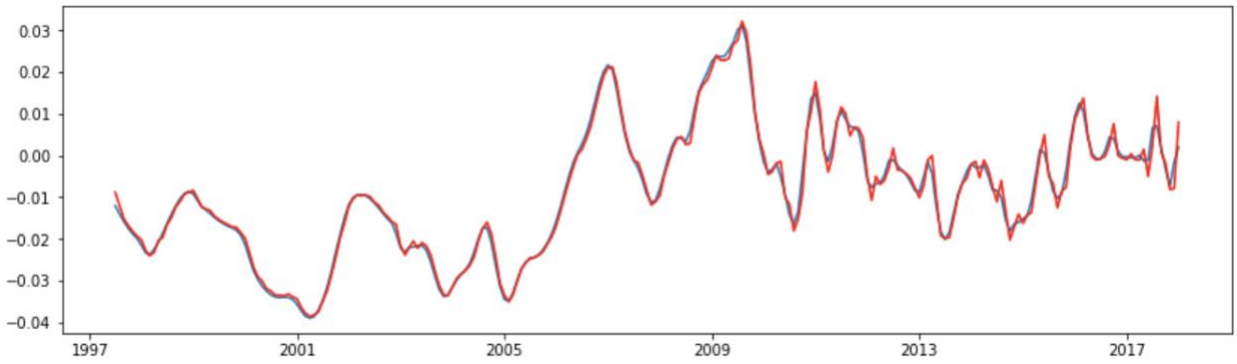
Dep. Variable:	y	No. Observations:	240
Model:	ARMA(7, 4)	Log Likelihood	-2265.903
Method:	css-mle	S.D. of innovations	2902.283
Date:	Thu, 23 Aug 2018	AIC	4557.807
Time:	10:08:28	BIC	4603.055
Sample:	0	HQIC	4576.039

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.95e+04	2.1e+04	1.876	0.062	-1758.043	8.07e+04
ar.L1.y	2.4765	0.072	34.489	0.000	2.336	2.617
ar.L2.y	-1.7145	0.213	-8.060	0.000	-2.131	-1.298
ar.L3.y	-0.3380	0.314	-1.078	0.282	-0.953	0.277
ar.L4.y	0.2825	0.329	0.860	0.391	-0.362	0.927
ar.L5.y	1.1593	0.289	4.010	0.000	0.593	1.726
ar.L6.y	-1.2565	0.207	-6.062	0.000	-1.663	-0.850
ar.L7.y	0.3822	0.073	5.226	0.000	0.239	0.525
ma.L1.y	0.0407	0.025	1.626	0.105	-0.008	0.090
ma.L2.y	-1.0592	0.039	-27.474	0.000	-1.135	-0.984
ma.L3.y	0.0407	0.026	1.554	0.121	-0.011	0.092
ma.L4.y	1.0000	0.014	70.395	0.000	0.972	1.028

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.9106	-0.6927j	1.1441	-0.3965
AR.2	-0.9106	+0.6927j	1.1441	0.3965
AR.3	0.8470	-0.7637j	1.1405	-0.1168
AR.4	0.8470	+0.7637j	1.1405	0.1168
AR.5	1.1500	-0.2366j	1.1741	-0.0323
AR.6	1.1500	+0.2366j	1.1741	0.0323
AR.7	1.1151	-0.0000j	1.1151	-0.0000
MA.1	0.8644	-0.5028j	1.0000	-0.0838
MA.2	0.8644	+0.5028j	1.0000	0.0838
MA.3	-0.8848	-0.4660j	1.0000	-0.4228
MA.4	-0.8848	+0.4660j	1.0000	0.4228

From there, one can predict. To see the value of the predictive model – i.e. a plot of residual erros, we can graph them and look at how well our predictions match the actual data:



The model fits nicely, and so predictions can be made with a reasonable amount of confidence. Accordingly, predictions for the next twelve (12) months (first three months of 2018) following data through 2017 are as follows:

```
Month 1: 1378144.981408
Month 2: 1387929.131385
Month 3: 1396826.606695
Month 4: 1408287.678361
Month 5: 1411452.842236
Month 6: 1407349.488655
Month 7: 1410738.613435
Month 8: 1424312.750661
Month 9: 1439246.398740
Month 10: 1449721.339434
Month 11: 1444359.063108
Month 12: 1439675.540764
```

V. Final Recommendation

Therefore, a recommendation is made for SU REIT to invest in the Stamford metro area, specifically in the zipcodes of New Canaan (06840), Darien (06820), and Brookfield (06840). The REIT can thus expect, with a reasonable level of confidence, median price in that area to increase over the next twelve months (from January 2018 on, in this study) of approximately $62,000 USD, and a 12 month return of about 4.4%.

---

[i] https://www.investopedia.com/terms/r/reit.asp

[ii] files.zillowstatic.com/research/public/Zip/Zip_Zhvi_SingleFamilyResidence.csv

[iii] https://www.zillow.com/research/data/

[iv] https://www.realtor.com/news/trends/americas-profitable-housing-markets/