

Mod 4 project Executive Summary

Andrew Hotchkiss

Assignment

- What are the top 5 best zip codes to invest in?
 - Use forecasted data (i.e. not data from the original data set) for your recommendations
- Define “Best”
 - Lowest risk investment → which zip codes are the safest investment where I have the least chance of losing my money?
- Ada County (Boise), Idaho selected as best county in the country to invest in 2021 for the lowest risk, 5-year investment

Why Boise, ID?

HOUSING MARKET OVERVIEW BOISE, IDAHO



OSEMN process overview

Obtain

Scrub

Explore


Model

iNterpret

Obtain/Scrub

- Zillow data overview
 - 14,723 zip codes with 265 months of data from 1996-04 to 2018-04 → **sampled monthly**
 - Wide format
- Pre-processing steps
 - Drop unnecessary columns → anything but County == 'Ada'
 - Convert from Wide to Long format for visualization and modeling
 - Cut data from before 2013 → only need data from 5-years ago (from 2018 back) to make forecasts 5-years into future

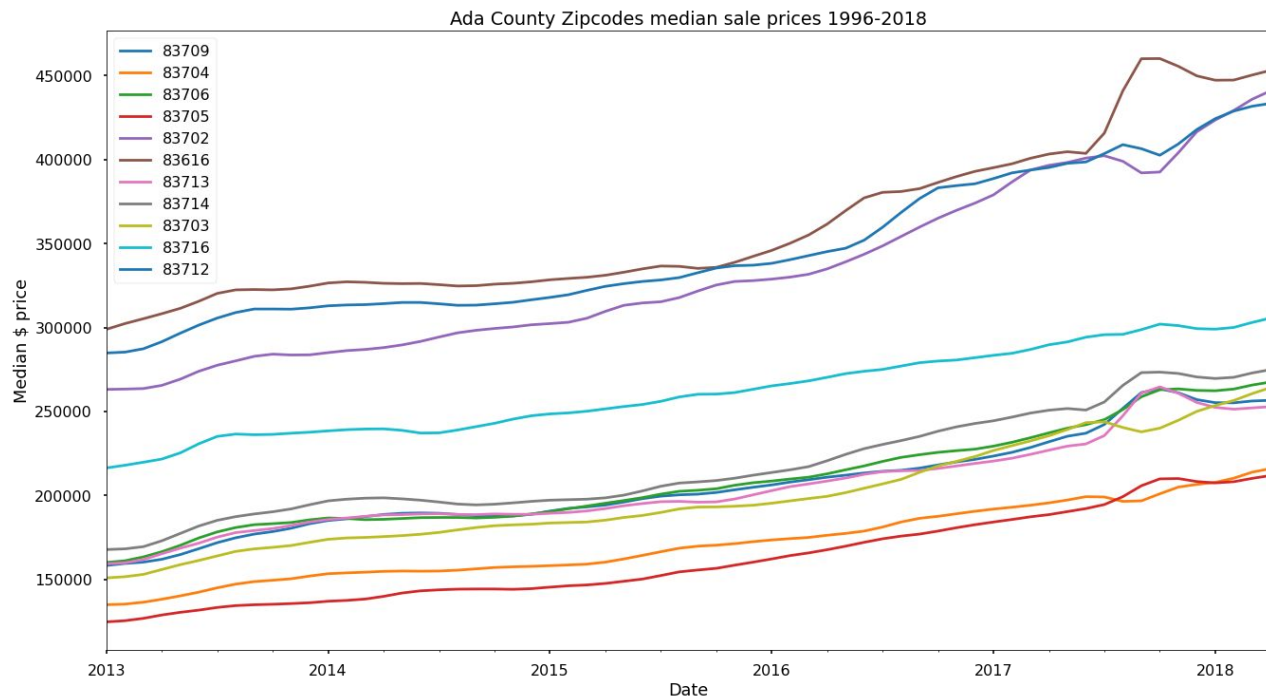
Wide-to-Long format



ID	Product1	Product2	Product3	Product4
1	1	NA	1	1
2	1	1	NA	1
3	1	1	NA	NA
4	1	1	1	1

ID	Product	value
1	Product1	1
1	Product3	1
1	Product4	1
2	Product1	1
2	Product2	1
2	Product4	1
3	Product1	1
3	Product2	1
4	Product1	1
4	Product2	1
4	Product3	1
4	Product4	1

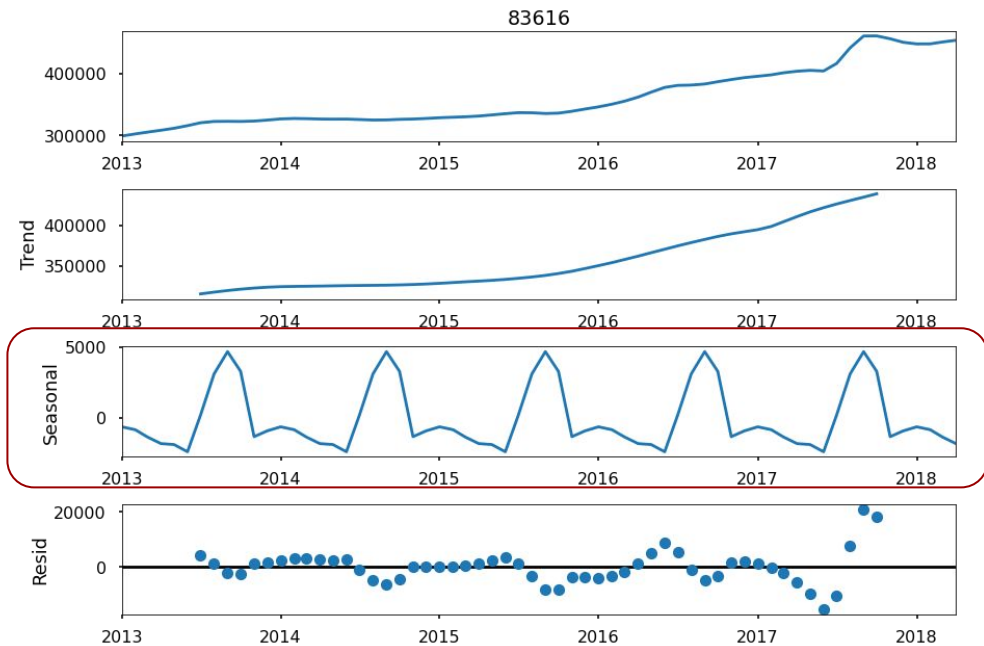
Explore - Non-stationary analysis



Strong positive trend in all zip codes

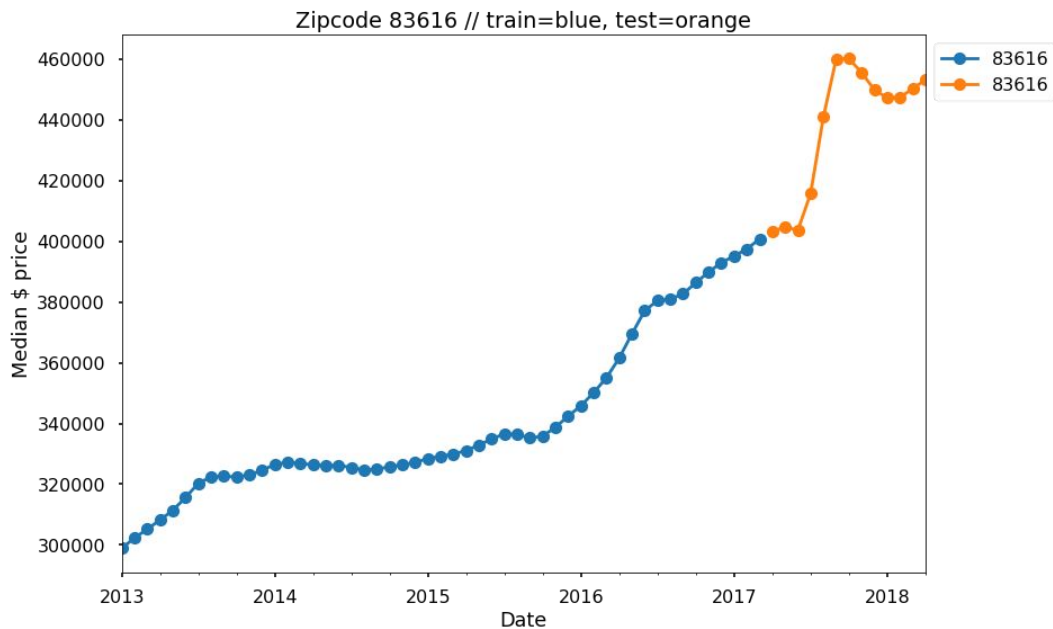
Explore - Non-stationary analysis (cont.)

- Seasonal_decompose library from statsmodels
- Explore one zip code (83616)
- Positive trend confirmed, Seasonal component uncovered
- Use **auto_arima** for parameter selection to determine best way to deal with non-stationary features



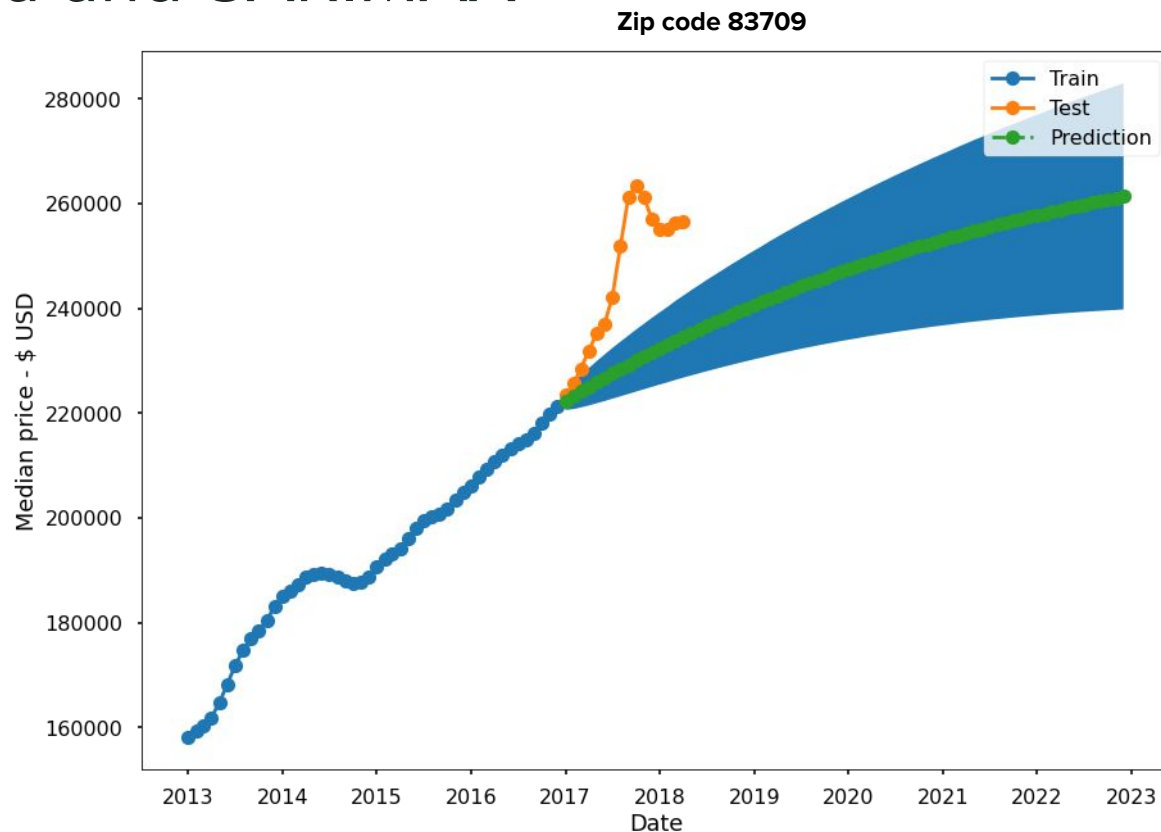
Model - Importance of Train/Test split

- Visualize train/test split with time series
- Models and forecasts are very sensitive to where the split occurs
- Balance between goodness-of-fit and overfitting can change drastically by adjusting train/test split



Model - auto_arima and SARIMAX

- auto_arima for model and parameter selection
- Model with SARIMAX using parameters selected by auto_arima



iNterpret

- Use **lowest relative** Akaike Information Criteria (AIC) to gauge model performance (assess goodness-of-fit)

ZIPCODE: 83705

SARIMAX Results

Dep. Variable:	83705	No. Observations:	48			
Model:	SARIMAX(0, 2, 2)	Log Likelihood	-343.421			
Date:	Sun, 20 Dec 2020	AIC	692.843			
Time:	23:56:18	BIC	698.328			
Sample:	01-01-2013	HQIC	694.898			
	- 12-01-2016					
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
ma.L1	-0.0444	0.085	-0.525	0.600	-0.210	0.121
ma.L2	-0.0441	0.083	-0.533	0.594	-0.206	0.118
sigma2	1.763e+05	4.77e+04	3.695	0.000	8.28e+04	2.7e+05
Ljung-Box (L1) (Q):	1.67	Jarque-Bera (JB):	0.24			
Prob(Q):	0.20	Prob(JB):	0.89			
Heteroskedasticity (H):	0.84	Skew:	-0.00			
Prob(H) (two-sided):	0.75	Kurtosis:	2.65			

ZIPCODE: 83702

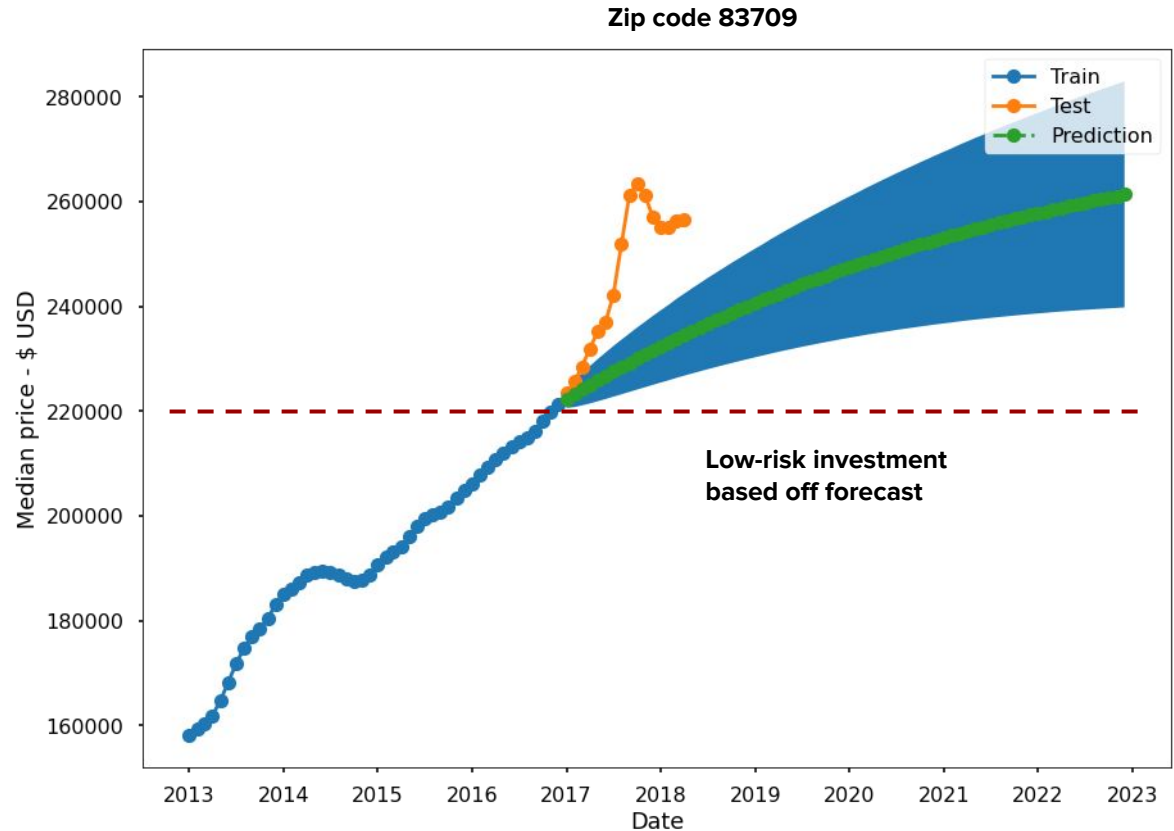
SARIMAX Results

Dep. Variable:	83702	No. Observations:	48			
Model:	SARIMAX(0, 2, 0)	Log Likelihood	-382.589			
Date:	Sun, 20 Dec 2020	AIC	767.178			
Time:	23:56:19	BIC	769.006			
Sample:	01-01-2013	HQIC	767.863			
	- 12-01-2016					
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
sigma2	9.795e+05	2.37e+05	4.127	0.000	5.14e+05	1.44e+06
Ljung-Box (L1) (Q):	8.24	Jarque-Bera (JB):	0.64			
Prob(Q):	0.00	Prob(JB):	0.73			
Heteroskedasticity (H):	0.60	Skew:	-0.17			
Prob(H) (two-sided):	0.33	Kurtosis:	2.53			

Zip code 83705 has a lower AIC [692.843] compared to 83702 [767.178]

iNterpret (cont.)

- **Visualize confidence intervals**
to determine risk of losing money
- If confidence interval drops below current value, negative ROI is more likely



Recommendations

- The top five lowest risk zip codes to invest in based off 5-year forecasts are:
 1. 83706
 2. 83709
 3. 83713
 4. 83704
 5. 83616
- This means that you are least likely to lose your money and most likely to increase your ROI by investing in these zip codes.

Future Work

- Dig deeper into `auto_arima` to see if there are better options for hyper parameter tuning to improve our models' fit.
- Get more recent data (the current data set only includes dates up to April 2018) to be able to forecast more accurately into the future.
- Research other data preprocessing techniques to see if we can improve model performance and forecasts.

Thank you!

Contact info: a.hotchkiss13@gmail.com