FINDING THE 5 BEST ZIP CODES FOR REAL ESTATE INVESTMENT IN THE NYC SUBURBS

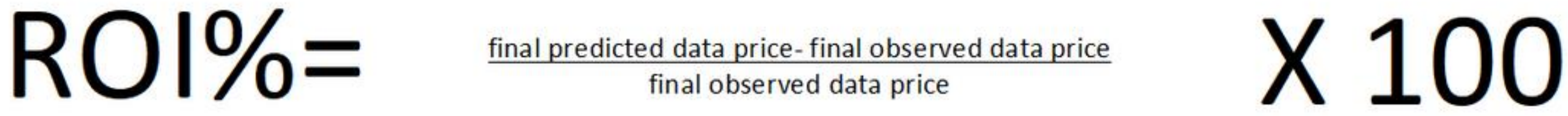
Business Understanding

Thanks for hiring me . It’s May 2018 and you, a Private Equity company, want to invest in NYC suburban real estate.

You have narrowed your focus down to 2 counties in NY: Westchester and Nassau and 2 counties in New Jersey: Bergen and Hudson.

The question you have is: **which zip codes should you invest in**?

Investment Horizon: 5 Years



$$\text{ROI\%} = \frac{\text{final predicted data price- final observed data price}}{\text{final observed data price}} \text{ X } 100$$

Business Understanding

- You are seeking the 5 zip codes with the highest ROI%.

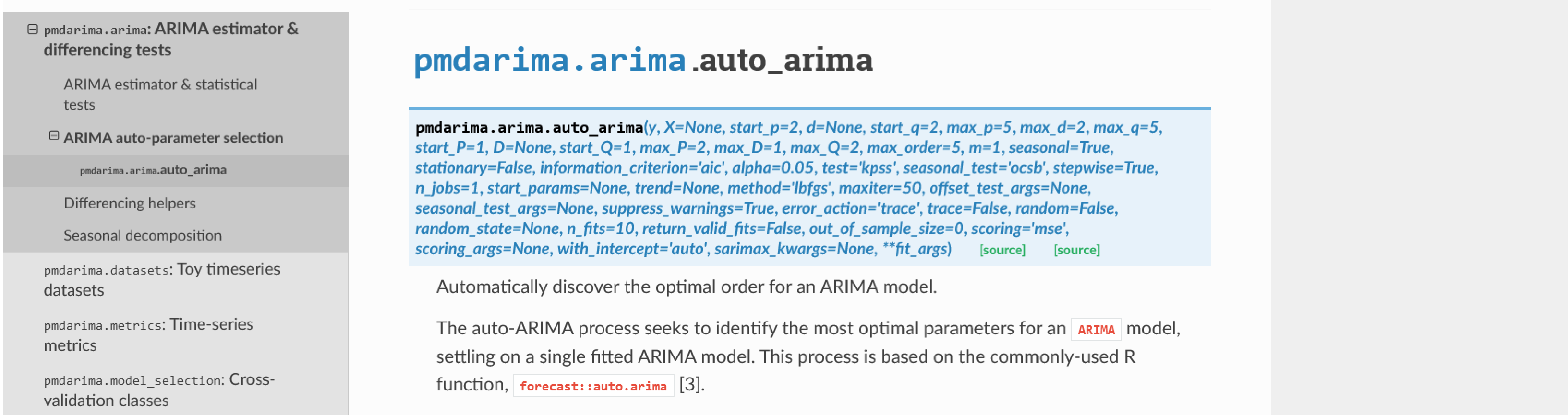
Data Understanding

I used an outstanding dataset from Zillow that contains monthly average sales data from almost every zip code in the United States from 1996-2018.

I narrowed the data down to the 201 zip codes in the 4 chosen counties and used that data to predict which would be the 5 most profitable in the next 5 years.

Modeling

I used an auto.arima model to predict the prices.


⊟ pmdarima.arima: ARIMA estimator & differencing tests
ARIMA estimator & statistical tests
⊟ ARIMA auto-parameter selection
pmdarima.arima.auto_arima
Differencing helpers
Seasonal decomposition
pmdarima.datasets: Toy timeseries datasets
pmdarima.metrics: Time-series metrics
pmdarima.model_selection: Cross-validation classes

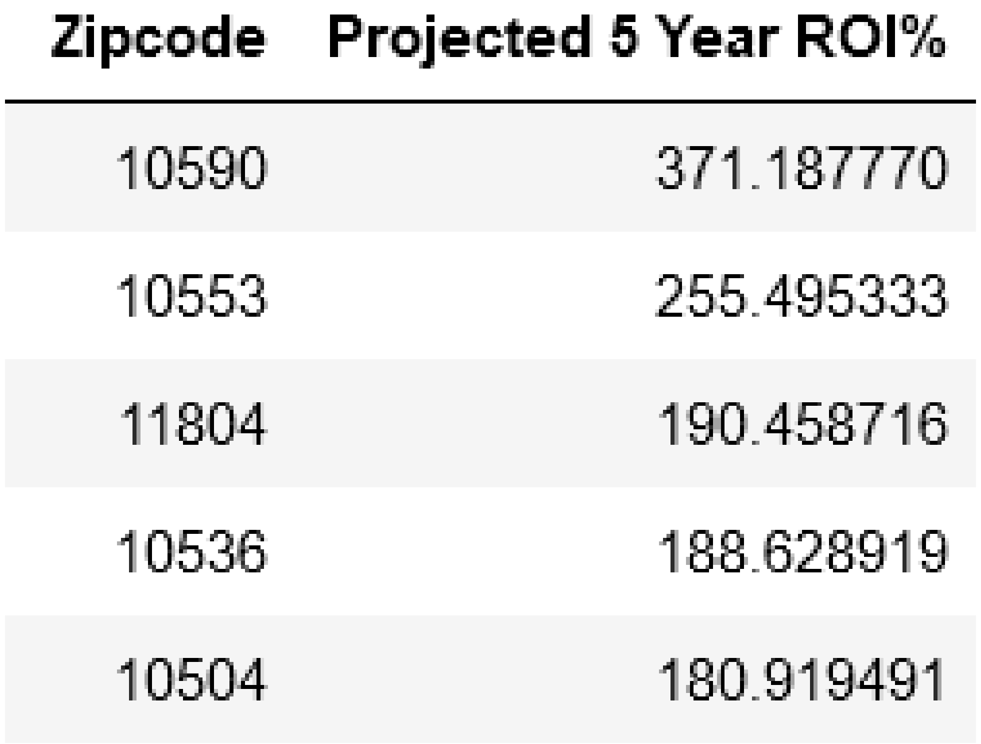
pmdarima.arima.auto_arima

```
pmdarima.arima.auto_arima(y, X=None, start_p=2, d=None, start_q=2, max_p=5, max_d=2, max_q=5,
start_P=1, D=None, start_Q=1, max_P=2, max_D=1, max_Q=2, max_order=5, m=1, seasonal=True,
stationary=False, information_criterion='aic', alpha=0.05, test='kpss', seasonal_test='ocsb', stepwise=True,
n_jobs=1, start_params=None, trend=None, method='lbfgs', maxiter=50, offset_test_args=None,
seasonal_test_args=None, suppress_warnings=True, error_action='trace', trace=False, random=False,
random_state=None, n_fits=10, return_valid_fits=False, out_of_sample_size=0, scoring='mse',
scoring_args=None, with_intercept='auto', sarimax_kwargs=None, **fit_args)    [source]    [source]
```

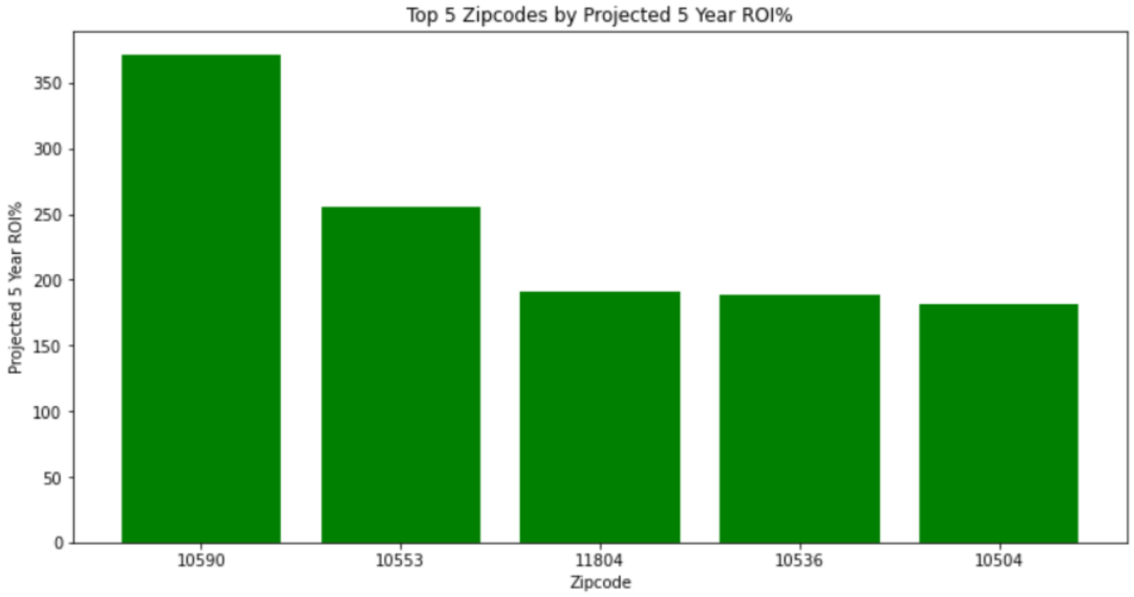
Automatically discover the optimal order for an ARIMA model.

The auto-ARIMA process seeks to identify the most optimal parameters for an `ARIMA` model, settling on a single fitted ARIMA model. This process is based on the commonly-used R function, `forecast::auto.arima` [3].

Results— 5 Highest ROI% Zip Codes

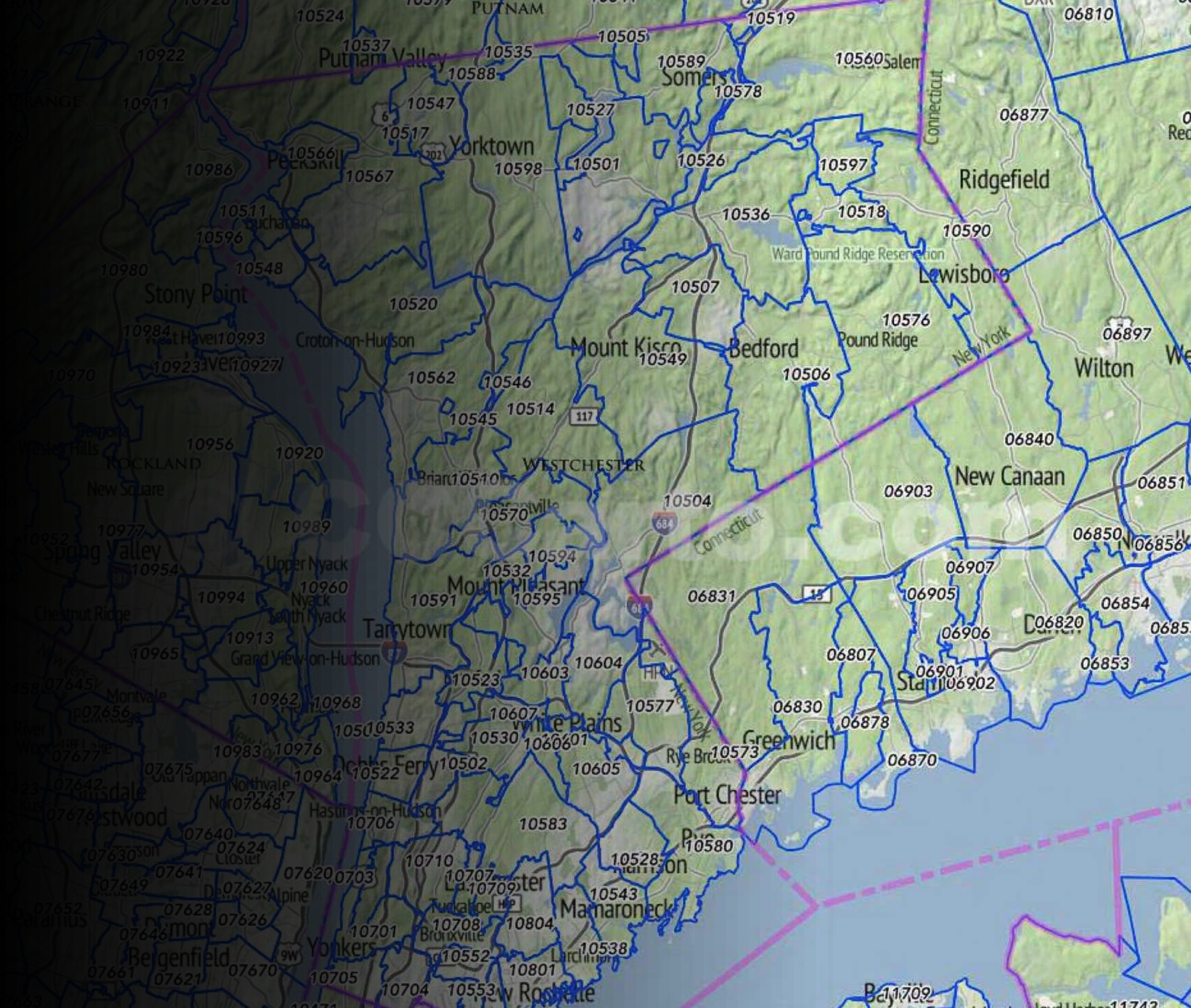


Zipcode	Projected 5 Year ROI%
10590	371.187770
10553	255.495333
11804	190.458716
10536	188.628919
10504	180.919491



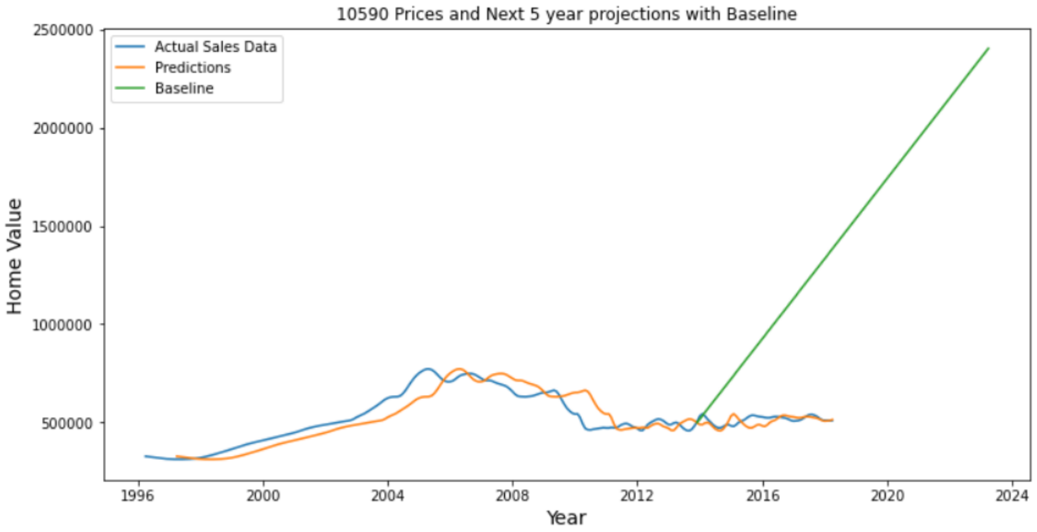
Results

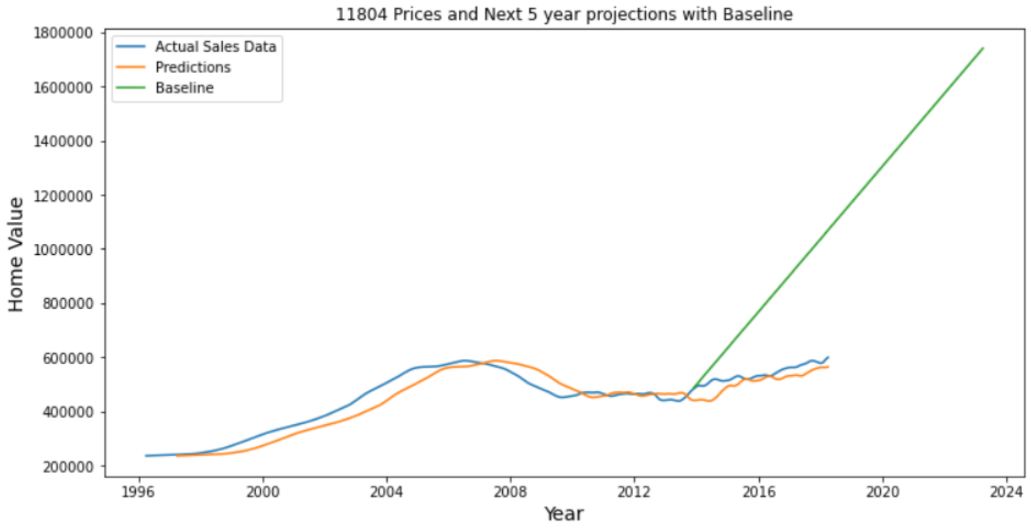
- 4 of the zip codes are in Westchester County.
- All of the zip codes, except 10553, are considered high income areas.
- 10590, 10536, and 10504 are all in Upper Westchester County and are close to each other.



Results-Prediction Problems

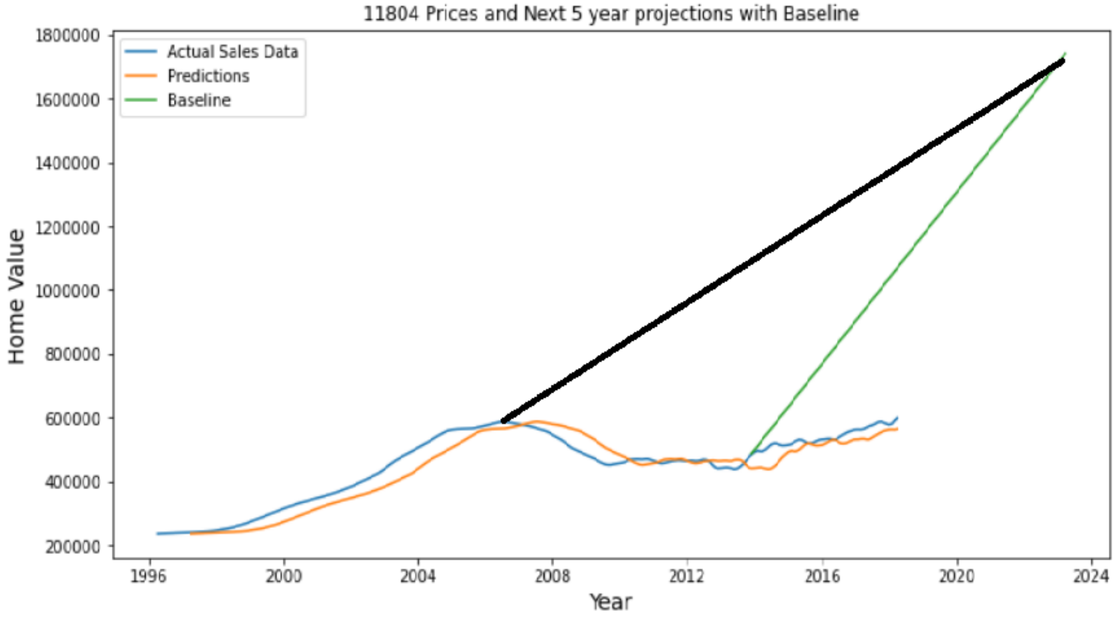
The predictions for 2013-2018 varied greatly from the actual sales data from 2018-2023.





Results-Predictions Problems

- The discrepancy between the predicted and actual prices may be a product of the Great Financial Crash. The model may be expecting the prices to increase after the financial crash. Prices did rebound but the model may be expecting prices to reach the levels they would have had there been no downturn.



Recommendations

- The top 5 zip codes with the highest projected ROI% were 10590, 10553, 11804, 10536, and 10504. Invest in those zip codes.
- The least expensive of the top 5 is 10553 and it has an expected ROI% of 255%. So, if you are targeting homes under $1 million then focus on that zip code.

Next Steps

- More data on Fairfield County, CT would also be very useful to investors and to the model. Unfortunately, the Zillow dataset omitted many zip codes in Fairfield County, CT(parts of which are only a 40-minute train ride to Midtown, Manhattan and contains some of the wealthiest areas in the country) and omitted data many zip codes in the city itself.
- Another ML model may yield more fruitful results. An XGBoost Regressor or a neural network model may yield better and more meaningful results.

Thanks!

Github:https://github.com/icapeli/Phase_4_Time_Series

Image from:https://www.youtube.com/watch?v=1TFzXBLBsAM