Seeing into the Future: Using Time Series Forecasting to Predict Housing Trends

María José del Granado (Mary Jo) ‘24 Data Science Major Capstone
Advisor: Eni Mustafaraj

Research Question

The ever fluctuating housing market is a topic of tremendous financial significance. Accurately predicting future home prices holds immense value for both buyers and sellers. However, predicting housing prices is a notoriously complicated task. For this reason I will investigate the following question: **To what extent can past price data alone be used to predict future trends of average home prices in the Boston metropolitan area?**

Data Collection

Scraping

I scraped data on houses in the recently sold section off of Zillow. I focused on houses within the Boston Metropolitan Area— homes that were within Middlesex, Essex, Norfolk or Suffolk county. The full dataset has 130,536 rows.

Data

I picked both starting and final periods that maximized the number of rows within my dataset: Jan 2021 to November 2023 (See Figures 1-2). After dropping homes that weren’t within this range of dates (2857) and homes with missing dates (1018), my dataset was reduced to 126,661 homes.

Splitting Data into Training and Testing

I labeled data up until June 2023 as training data, and the last six periods as testing data.

Figure 1: House Count by Year-Month

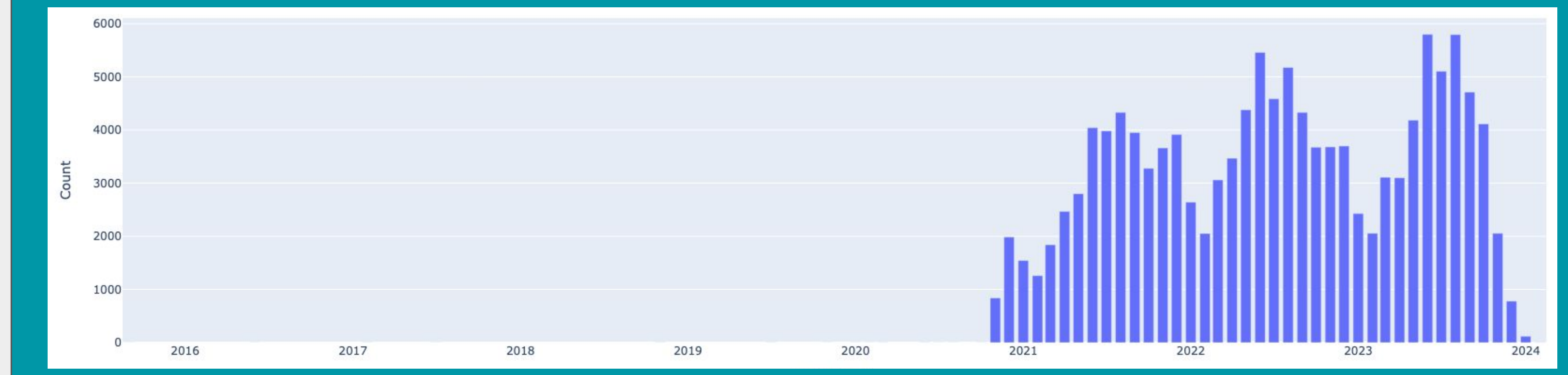

Figure 2: Earliest Dates at City Level

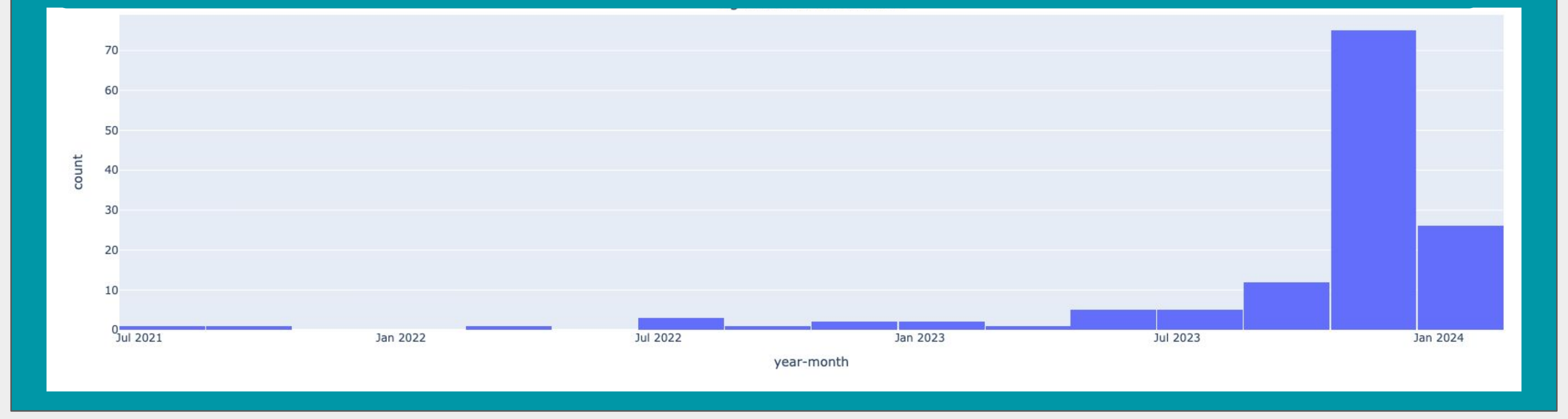

Figure 3: Average Sale Prices of Homes in Boston Metro Area at City Level

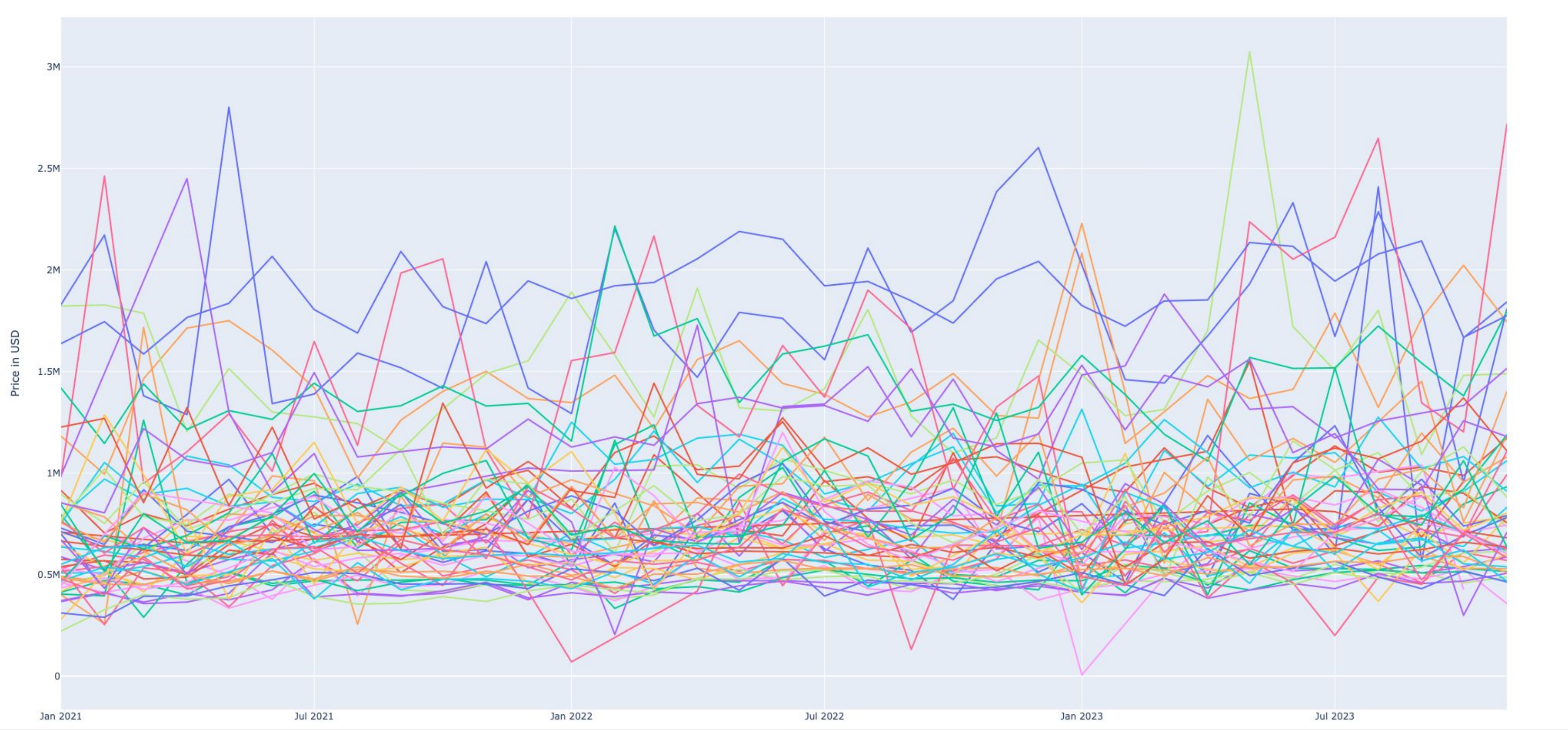

Figure 4: Average Sale Prices of Homes in Boston Metro Area with Model Predictions

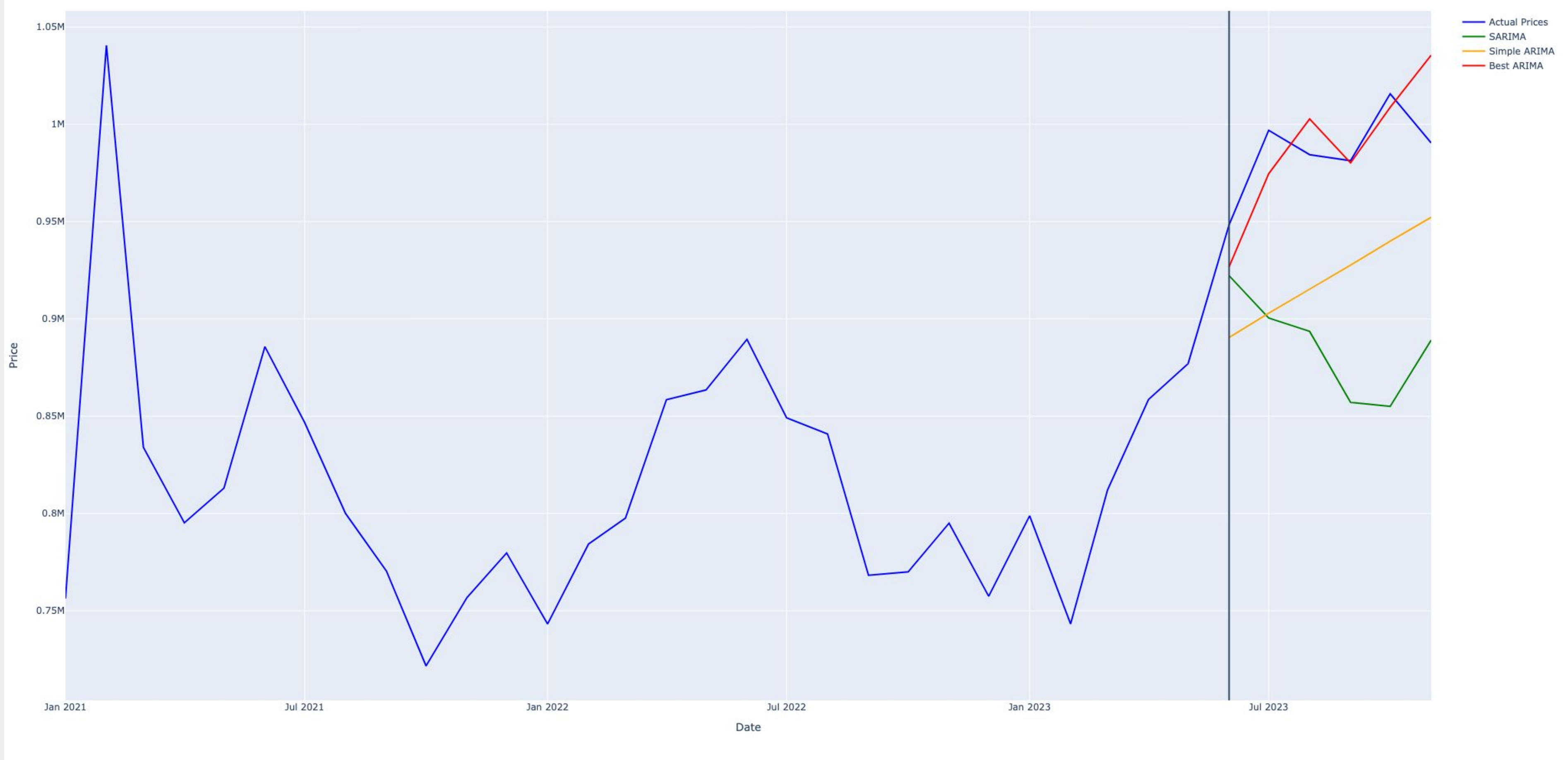

Table 1. RMSE by Model

Model	Training RMSE	Testing RMSE
Baseline	260,798.75	71,310.26
ARI(1,2)	210,595.41	67,136.47
ARIMA(5,2,5)	95,282.21	23,735.44
SARIMA(0,2,0)(0,1,1,12)	241,461.62	107,792.43

Model & Analysis

Baseline Model

Predicts a future month’s prices to be same as previous month. While the baseline models offers a simple starting point, it’s limited both by its poor performance and it’s inability to make long-term predictions.

Autoregressive and ARIMA Models

- AR - Autoregressive (Number of lagged observations)
- I - Integration (Order of differencing)
- MA - Moving Average (Size of moving average window)

To assess the suitability of fitting an autoregressive model, I conducted both Ljung-Box and Augmented Dickey-Fuller (ADF) tests on my data. The ADF test confirmed the presence of non-stationarity, rendering a simple autoregressive model inadequate. So, I fit an ARI(1,2) model to account for the non-stationarity. The integration order (d) was set to 2, as this was the first differencing step that visually indicated stationarity in the data.

Optimal ARIMA Model

To optimize the hyperparameters of a secondary ARIMA model, I employed a grid search. This search exhaustively evaluated all possible combinations of parameters (p, d, and q) up to order 6. The model with the lowest Akaike Information Criterion (AIC) value was then selected. The best model performance I achieved was with the optimized ARIMA model.

SARIMA Model

A seasonal decomposition revealed the presence of seasonal trends, prompting the use of a SARIMA model. I employed a package that uses numerical approximation to best estimate optimal SARIMA hyperparameters by comparing the AIC of the models.

Discussion and Conclusion

Predicting the average sale price of a home, merely on historical data isn’t sufficient to get accurate predictions. After all, there are several different factors that play a role in determining home sale prices. However, my models are limited by the fact that the data that I collected doesn’t go that far back in time. It could very well be that better performance could be achieved, if I had more data.

Acknowledgements

Special thanks to everyone in the ECON/CS 350 Independent Study, Professor Kyung Park and Professor Eni Mustafaraj.