DSCI-D 590 Time Series Analysis

Final Project Part 2

LE Lee | Andrew Mouat | Nicholas Perry

Project Description

Objective:

The problem that our team elected to solve is the lack of tools available to potential home buyers, or sellers, for accurate and timely reporting of price signals. It is important to have readily available analysis and forecasting on real estate pricing as this will allow users to buy, or sell, the right properties at the right time to maximize their investment. Therefore, the primary objective of this project is to create a web-based application that allows users to generate price predictions for the current value of a home and value forecasts for residential properties within an input, Indiana-based zip code. Furthermore, users will have the ability to enter various parameters, such as number of bedrooms, number of bathrooms, square footage, etc. to refine the tool's accuracy. In addition to historical pricing data and user input, our final model aims to incorporate other external, influencing factors including, but not limited to, mortgage interest rates and total home inventory. This tool will allow prospective buyers/investors the opportunity to independently compare the current listed value of a property to the predicted and forecasted value from our model to make informed decisions. Additionally, the tool can be used by current homeowners and real estate agents to determine a listing price for their property that is reasonable and competitive with the rest of the local market.

Usefulness:

The housing market is noticeably susceptible to changes over time. These fluctuations have a vast impact on local economies, personal finances, and general public perceptions of the economy as a whole. Forecasting real estate prices can help stakeholders (buyers, sellers, investors, real estate agents, etc.) in the decision-making process of purchasing and selling properties. Our web tool will show the viability of investments for future profit. It will also help to identify undervalued homes, which could be exploited for profit through real estate arbitrage. Finally, it will also assist users by properly valuing their current primary residence and their potential next home. While Zillow has an estimated home value, or "Zestimate," to help users value properties, it does not have them for every property. The estimates are also tied to individual properties and so they do not allow a user to get an idea of the accepted price range for homes in an area with certain parameters. Zillow also does not provide future forecasts of prices based on time series principles (seasonality, trends, etc.).

Dataset Generation

In order to generate a dataset that satisfactorily met the needs of our project, our team developed a custom web-scraping script in Python that pulls relevant information from Zillow for homes in Indiana. The script generated two primary files that our team utilized as input for the development of our Time Series Analysis Framework. At the time of this report, our dataset is comprised of 3718 properties. However, as our team progresses with the project, this number will increase as the scraping script has more time to run.

1. Home Information File - This file contains pertinent input information for each home that can be used to filter the data points used for analysis based on user

input. Currently, the web-scraping tool pulls the number of bedrooms, number of bathrooms and number of square feet as our team collectively decided these factors to be a reasonable starting point for our tool. As we continue progressing through the project, we may elect to add additional features to improve the forecasting accuracy of our tool. The table below provides an example of the file contents for reference.

Table 1 *Example of Web Scraping Output for Home Information*

Address	Beds	Baths	Sqft
1012 W Water St, Berne, IN 46711	3	2	1410
1015 W Main St, Berne, IN 46711	4	3	2174
1017 Dearborn St, Berne, IN 46711	4	2	2157
1017 Rose Ln, Berne, IN 46711	3	2	1700
1017 W Water St, Berne, IN 46711	3	3	2065
1018 Rose Ln, Berne, IN 46711	4	2	1428
1018 W Clark St, Berne, IN 46711	3	2	1624
102 W 300 S, Berne, IN 46711	4	3.5	2308
1020 Dearborn St, Berne, IN 46711	3	2	1479
1020 W Main St, Berne, IN 46711	1	1	728
1021 E 550 S, Berne, IN 46711	3	2.5	1456
1022 E 550 S, Berne, IN 46711	4	1.5	2060
1023 W Main St, Berne, IN 46711	3	1.5	1377
1023 W Water St, Berne, IN 46711	3	1.5	1676
1024 Rose Ln, Berne, IN 46711	4	3.5	3258

2. Home Sale History File - This file contains any listed sale history for each property (when available). This information is critical in developing the temporal element to our dataset as it gives us several data points for each applicable property throughout time. Then, by combining several like properties (based on the information we have - number of beds, baths, zip code, etc.), our team can develop a time series with an adequate number of data points. The table below provides an example of the file contents for reference.

Table 2 *Example of Web Scraping Output for Home Sales History*

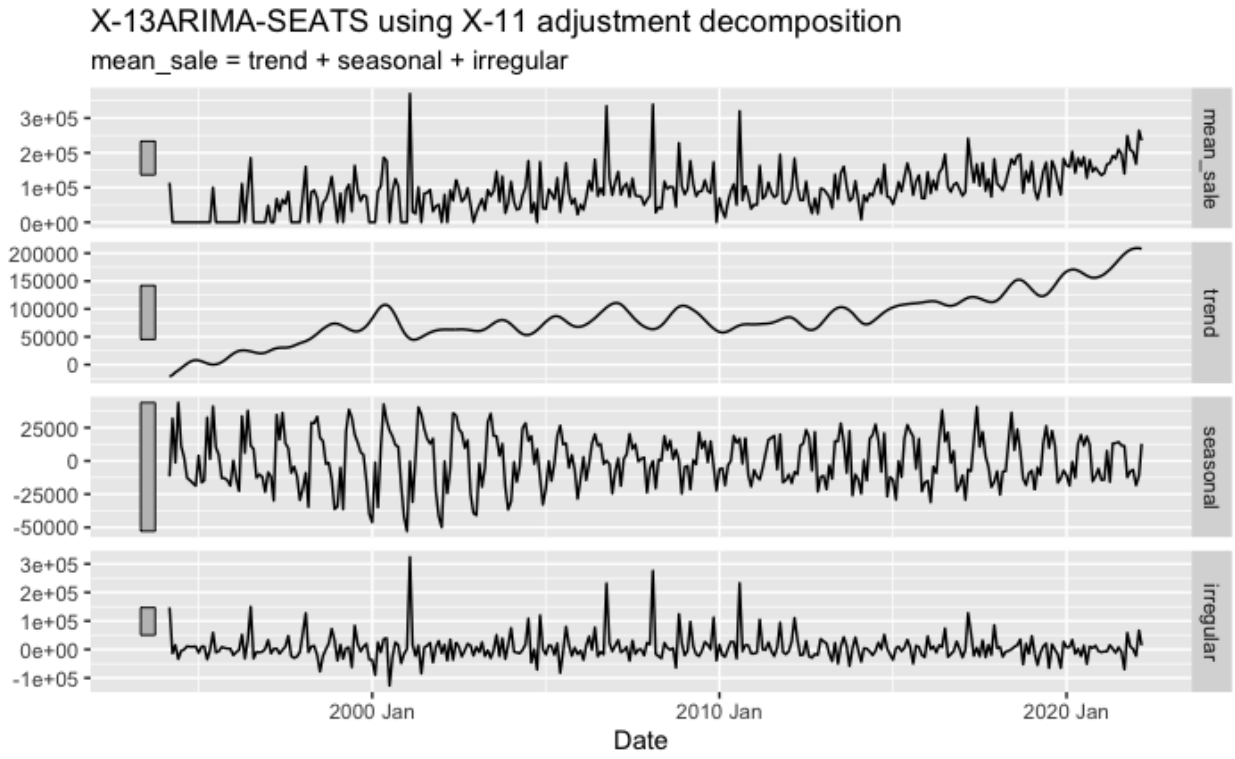
Address	Date	Event	Price
1024 W Water St, Berne, IN 46711	3/5/2018	Listed for sale	$123,000 (0%)
1024 W Water St, Berne, IN 46711	3/6/2018	Listing removed	$123,000 (0%)
1024 W Water St, Berne, IN 46711	5/19/2018	Sold	$126,000 (+2.4%)
1027 W 700 S, Berne, IN 46711	12/1/2021	Listed for sale	$142,000 (0%)
1027 W 700 S, Berne, IN 46711	12/11/2021	Pending sale	$142,000 (0%)

1027 W 700 S, Berne, IN 46711	1/7/2022	Sold	$135,000 (-4.9%)
1034 W 700 S, Berne, IN 46711	7/6/2020	Listed for sale	$137,500 (0%)
1034 W 700 S, Berne, IN 46711	7/8/2020	Pending sale	$137,500 (0%)
1034 W 700 S, Berne, IN 46711	7/27/2020	Listing removed	$137,500 (0%)
1034 W 700 S, Berne, IN 46711	8/19/2020	Sold	$130,000 (-5.5%)

Time Series Analysis Framework

I. Time Series Decomposition

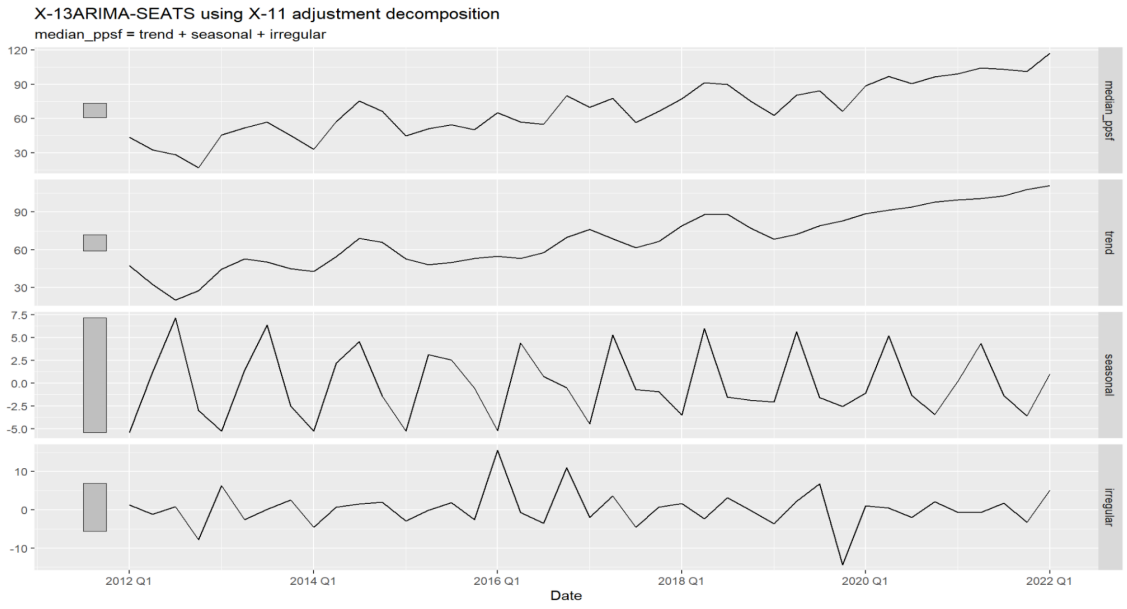
Shown below is an X-11 Method Decomposition for the time series of mean home sale prices in Indiana (based on the data generated to date). Using the X-11 decomposition method, the seasonal component is allowed to vary over time. This method is also highly robust to outliers and level shifts, making it preferred to alternative methods such as classical or STL decomposition for our application.



Analyzing the above output, it is clear that our data set exhibits an upward trend with time and has strong seasonality, as one would expect. It is worth noting that our team implemented an inflation adjustment to the data prior to performing the decomposition.

The above decomposition was performed on all properties in our dataset. However, the dataset can be quickly filtered based on user input criteria. For example, say a user was interested

in a 3 bedroom property in the zip code 46711. The overall data set could be filtered and the decomposition reconducted, producing the plot below. Note that for the below plot, our team has utilized the median price per square foot instead of the sale price as means of normalization and also grouped data by quarter to help alleviate issues resulting from filtering the dataset to a smaller number of points. Again, we see an upward trend with strong seasonality.



II. Time Series Visualizations

This section provides several types of visualizations for historical home sale prices. The time series utilized are monthly sale price of all homes (mean and median), monthly sale price of homes in different regions (mean and median), and monthly sale price of homes of different square footage (mean and median). We have determined that the distribution of mean and median sale prices are approximately equivalent through the comparison of their respective time series.

Figure 1 demonstrates the monthly mean sale price for homes in Indiana. Figure 2 demonstrates the monthly median sale price for homes in Indiana.

Figure 1 *Monthly Mean Sale Price*

Figure 2 *Monthly Median Sale Price*

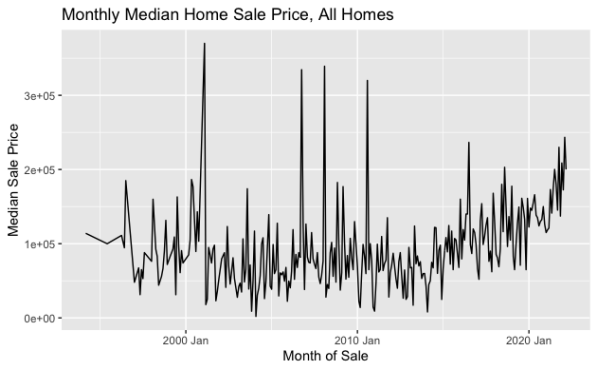
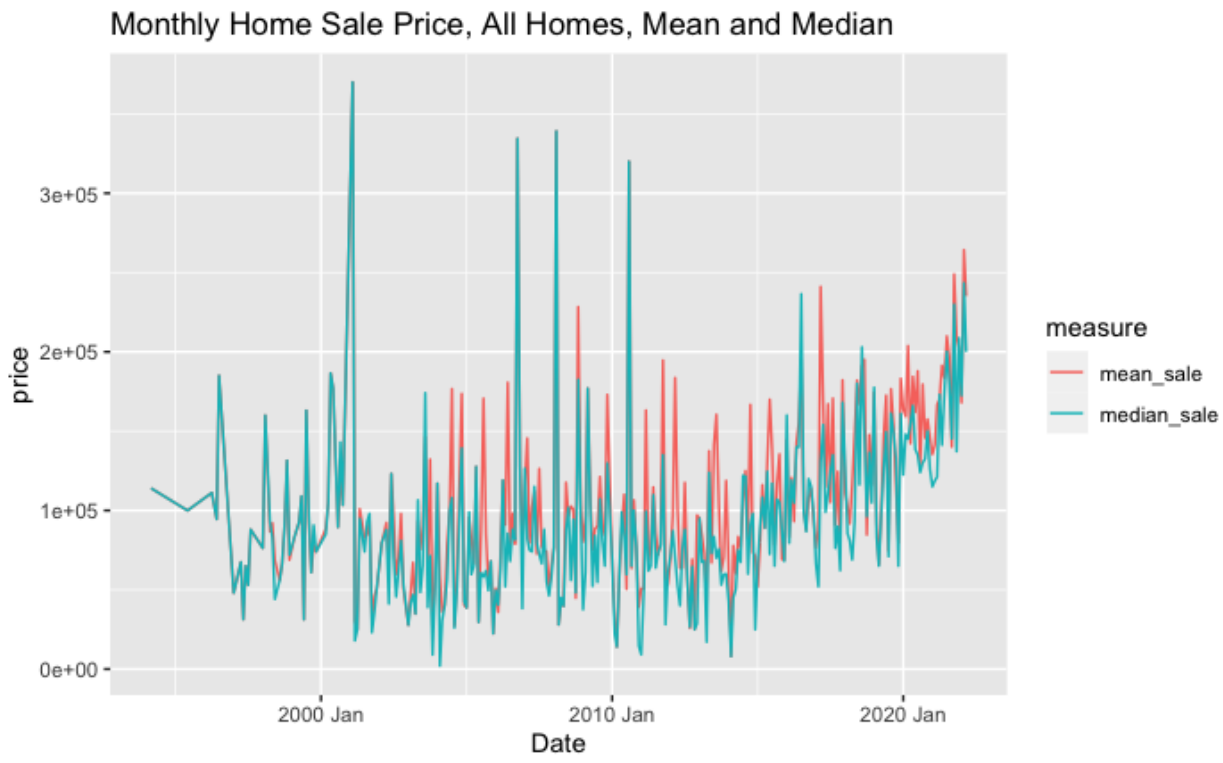


Figure 3 provides a plot in which the monthly mean and median sale price of homes can be compared. It can be observed that the differences in the time series for the mean and median monthly sale price are very minimal, as they are overall quite similar in shape; the mean appears to be higher on average than the median. As the size of our dataset increases, our team will further evaluate if we should utilize mean or median price for modeling purposes. The median price, in theory, should be more robust to outliers, making it more favorable initially.

Figure 3 *Monthly Sale Price, Mean and Median*



Figures 4 and 5 respectively demonstrate the monthly mean and median sale price of homes in different regions of the state of Indiana (C = Central, N = Northern and S - Southern). Figures 6 and 7 demonstrate the same plots, but faceted by region for viewability.. It can be seen that the Central Region has a tendency to have a higher monthly mean and median sale price, followed by the Northern Region. There appears to be a lack of data for the Southern Region, which will be resolved as our dataset continues to grow.

Figure 4 *Monthly Mean Sale Price by Region*

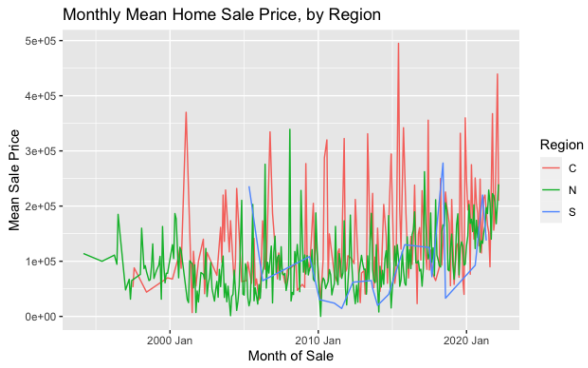


Figure 5 *Monthly Median Sale Price by Region*

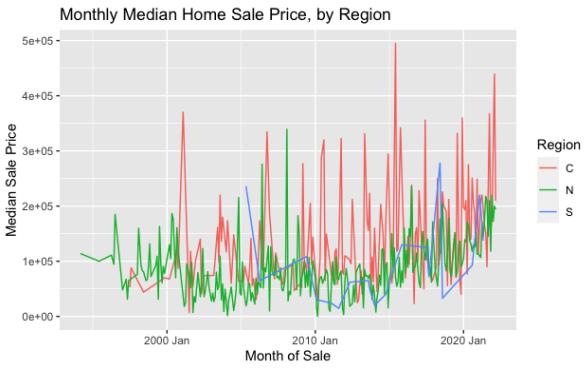


Figure 6 *Monthly Mean Sale Price by Region*

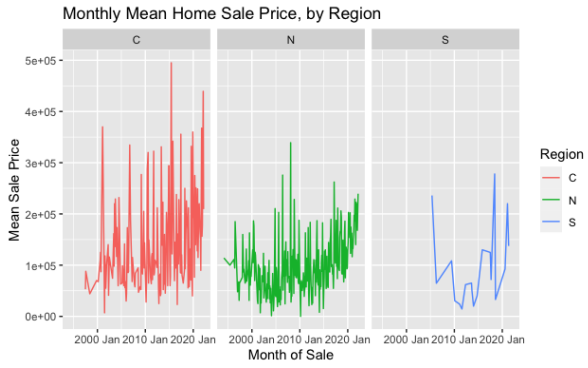


Figure 7 *Monthly Median Sale Price by Region*

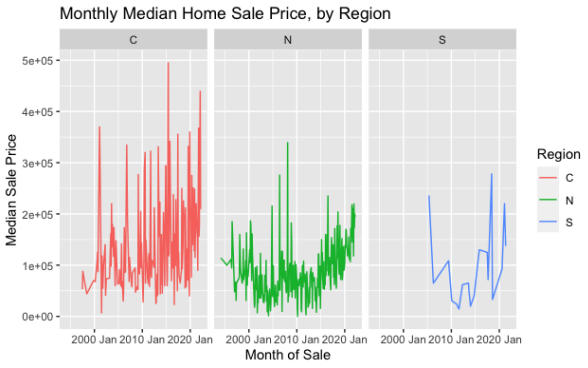
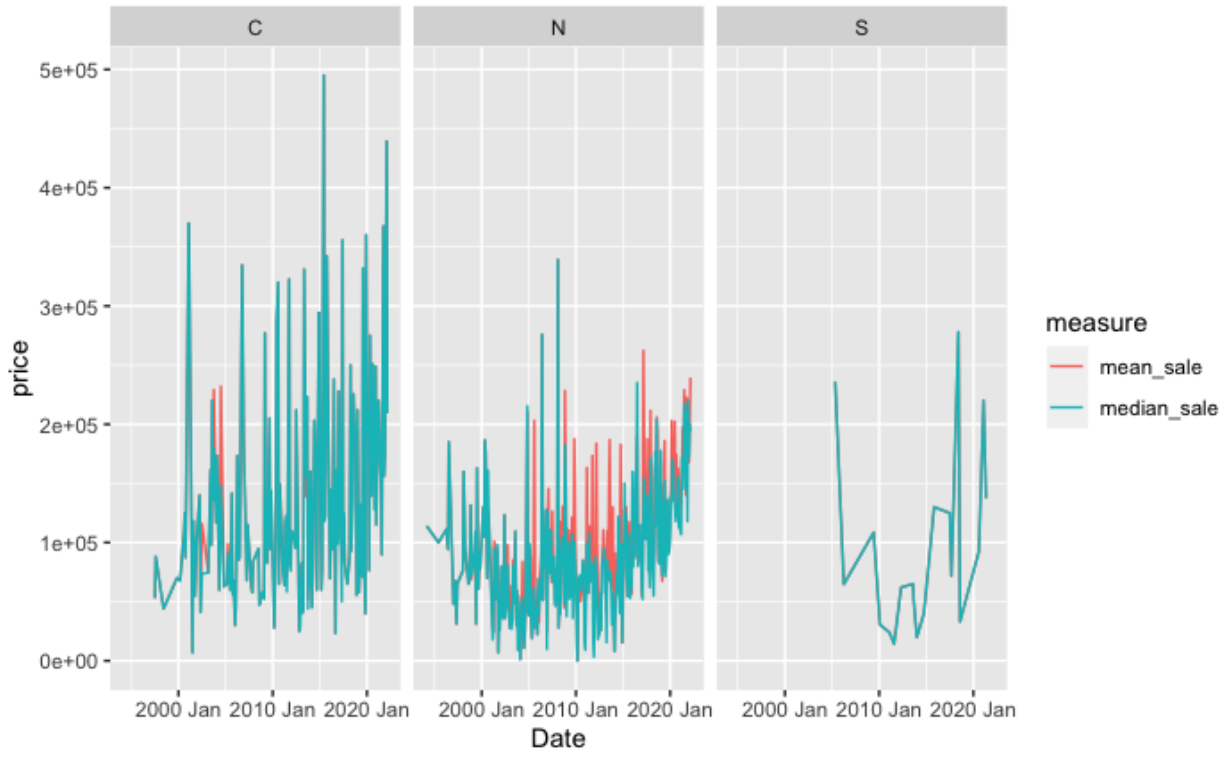


Figure 8 again demonstrates the monthly sale price by region such that comparison between the mean and median measures in the time series is allowed. It may be observed that there is not much variation between the two measures.

Figure 8 *Monthly Sale Price by Region, Mean and Median*



Figures 9 and 10 respectively demonstrate the monthly mean sale price of homes in different regions, faceted by their square footage attribute. It can be seen from these visualizations that both the mean and median sale price of a home in all three regions appears to have a direct relationship with the square footage interval in which the house lies, and all homes appear to be trending upward in price over time. This fact is what drove our team to generate the "price per square foot" feature to simplify the workflow.

Figure 9 *Monthly Mean Sale Price by Region & Square Footage*

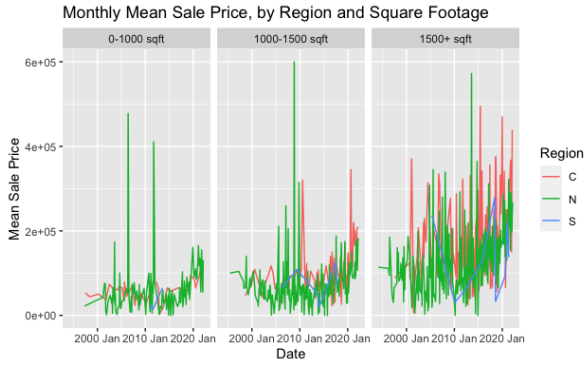
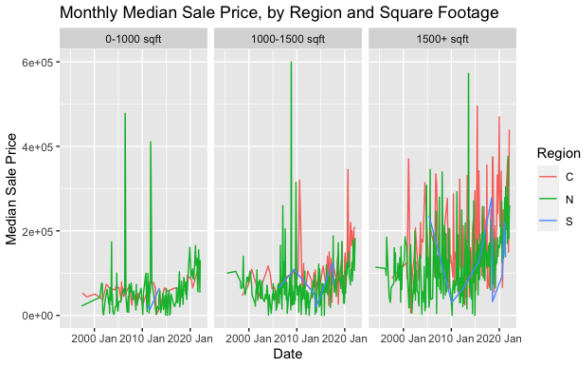


Figure 10 *Monthly Median Sale Price by Region & Square Footage*



Figures 11 and 12 respectively demonstrate the monthly mean and median sale price of all homes in Indiana depending on their square footage. Again, it can again be observed that there appears to be a direct relationship on average between the square footage of a home and its sale price throughout the time interval contained in the dataset.

Figure 11 *Monthly Mean Sale Price by Square Footage*

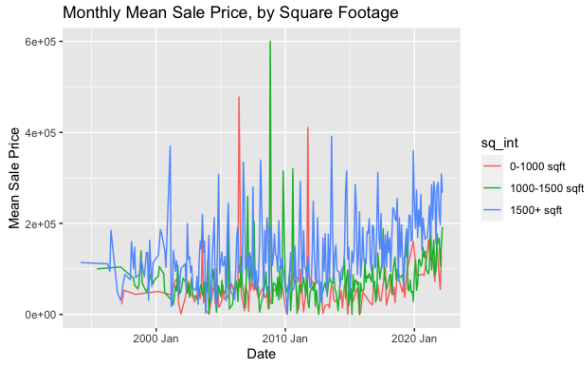
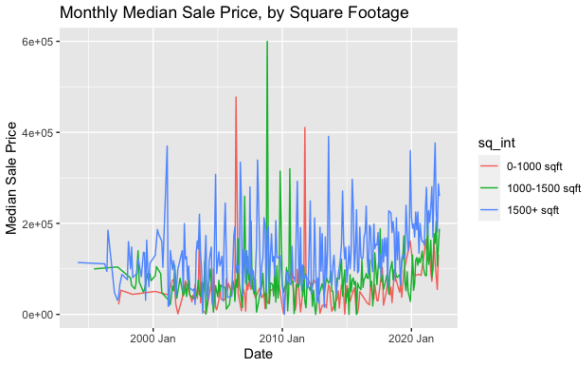
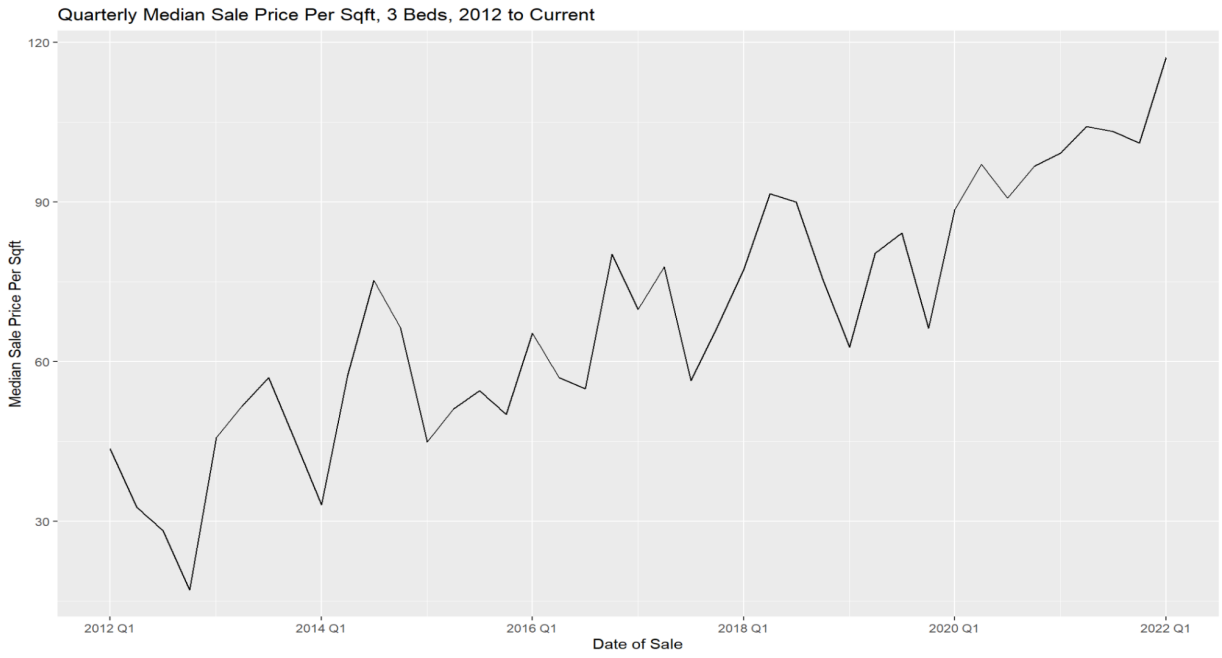


Figure 12 *Monthly Median Sale Price by Square Footage*



As was done with the decomposition, the above visualizations can be re-created for a subset of the overall dataset. For example, Figure 13 below shows the Median Sale Price (normalized by square footage) for 3 bedroom homes. This workflow will be critical as we progress with the project and accept user input. The visualization generation will be dynamic based on the input values and should also allow for the user to perform comparisons between different sets of inputs, allowing them to become more informed on how property features or location impact the time series.

Figure 13 *Quarterly Median Sale Price Per Square Foot, 3 Bedrooms, 2012 to Current*



III. Description of Time series

As is evident by the decomposition, our time series appears to demonstrate both trend and seasonality, even when subset. By definition, a stationary time series is one whose properties do not depend on the time at which the series is observed. As such our time series which has both

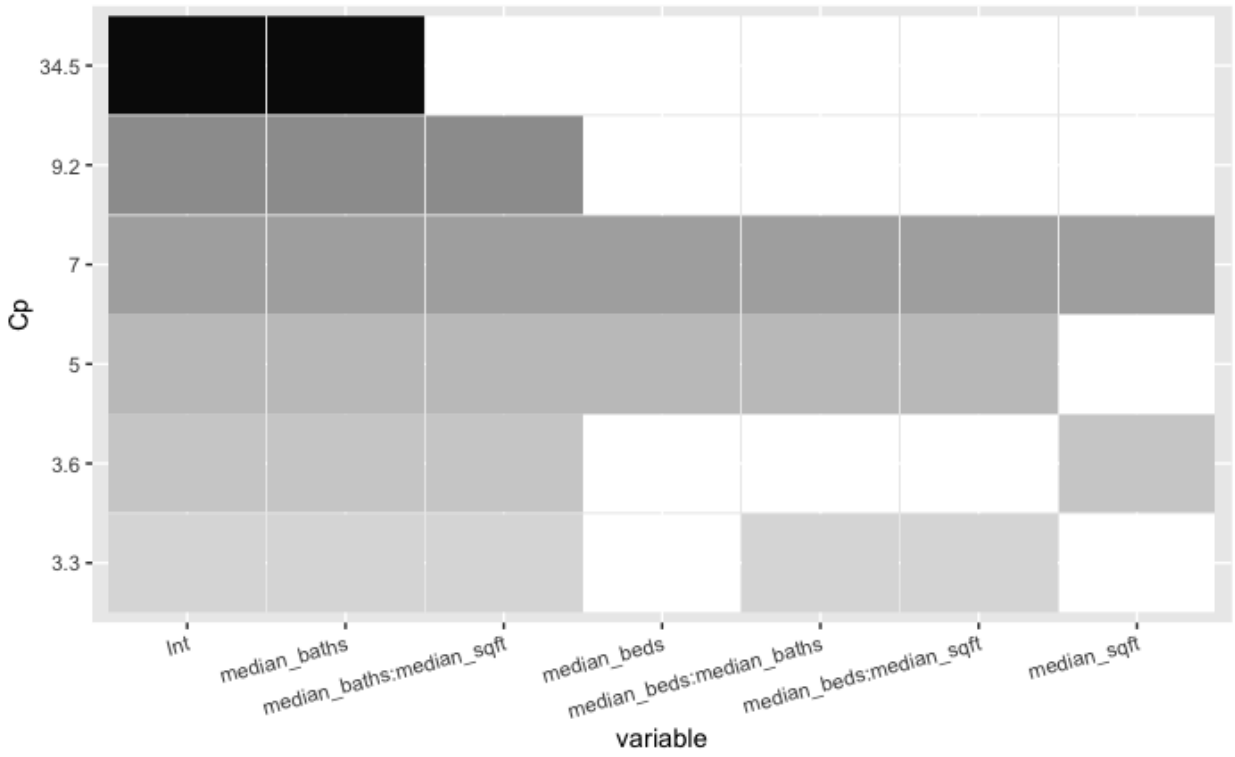
trend and seasonality is not stationary. For purposes of modeling, our team will apply transformations as required to stabilize variance and mean.

IV. Models

Based on the goals our team has for this project, we elected to forgo the implementation of simple models such as Naive or Mean modeling. Therefore, our team focused primarily on developing a workflow for regression and ARIMA models, including ARIMAX or dynamic regression models. As we progress with the project, we look to refine the currently developed models to improve forecasting accuracy. Furthermore, as a clarifying note, all of the models below were developed using the full dataset. In future stages of the project, the dataset will be subsetted based on user input prior to model generation.

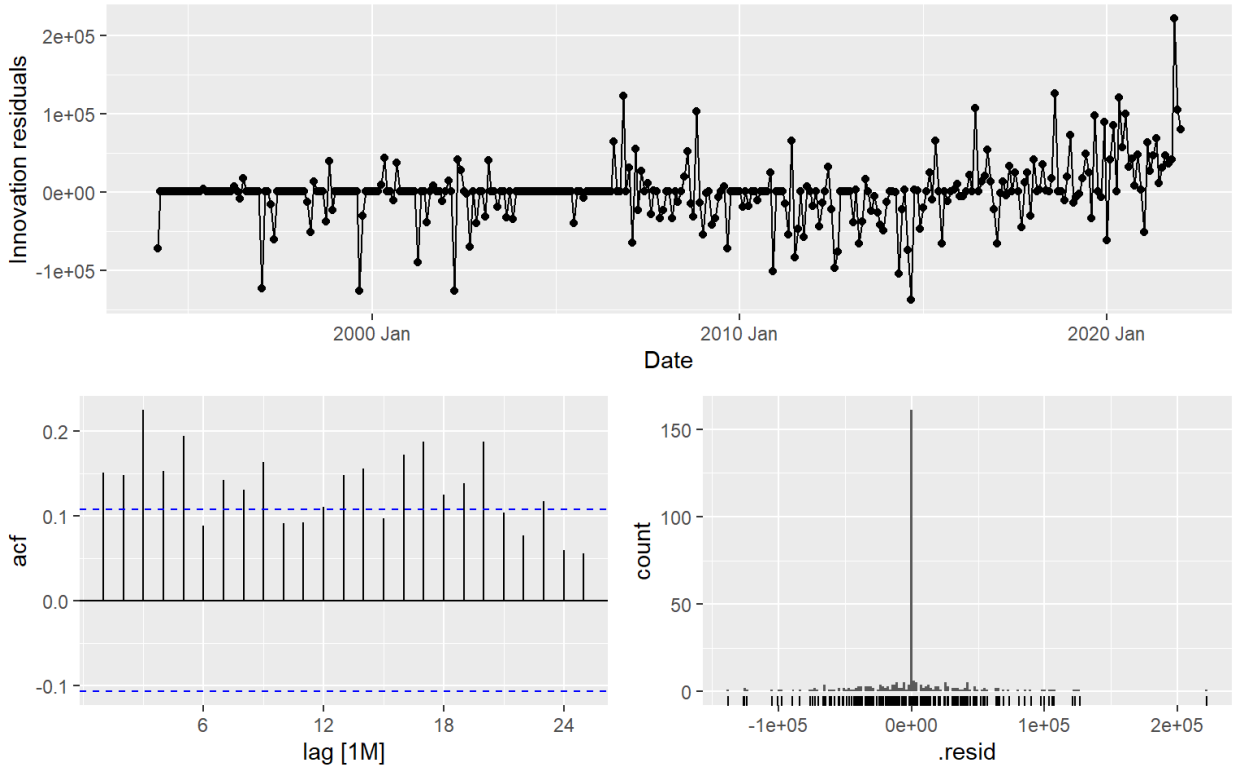
Using a plot of the Mallows's C_p values (Figure 14), instead of individual analysis using least squares regression, we were able to assess the fit of various regression models using the available regressors of median number of beds, median number of bathrooms, median square footage, and all combinations of interactions between those variables on the home price. The model with the lowest C_p values used median number of bathrooms, the interaction variable of median beds and median baths, the interaction variable of medians baths and square footage, and the interaction variable of median beds and square footage to predict home price.

Figure 14 *Plot of Mallows' C_p Values*



We see with the interaction variables that the effect of the number of beds changes with the number of bathrooms, for example. The effect of median square footage also changes with the number of bedrooms, as well as the effect of the median number of bathrooms changing with the median square footage. This makes sense intuitively because the larger a house, the less important the number of bedrooms becomes. Similarly, the more bedrooms there are, the more important the number of bathrooms becomes.

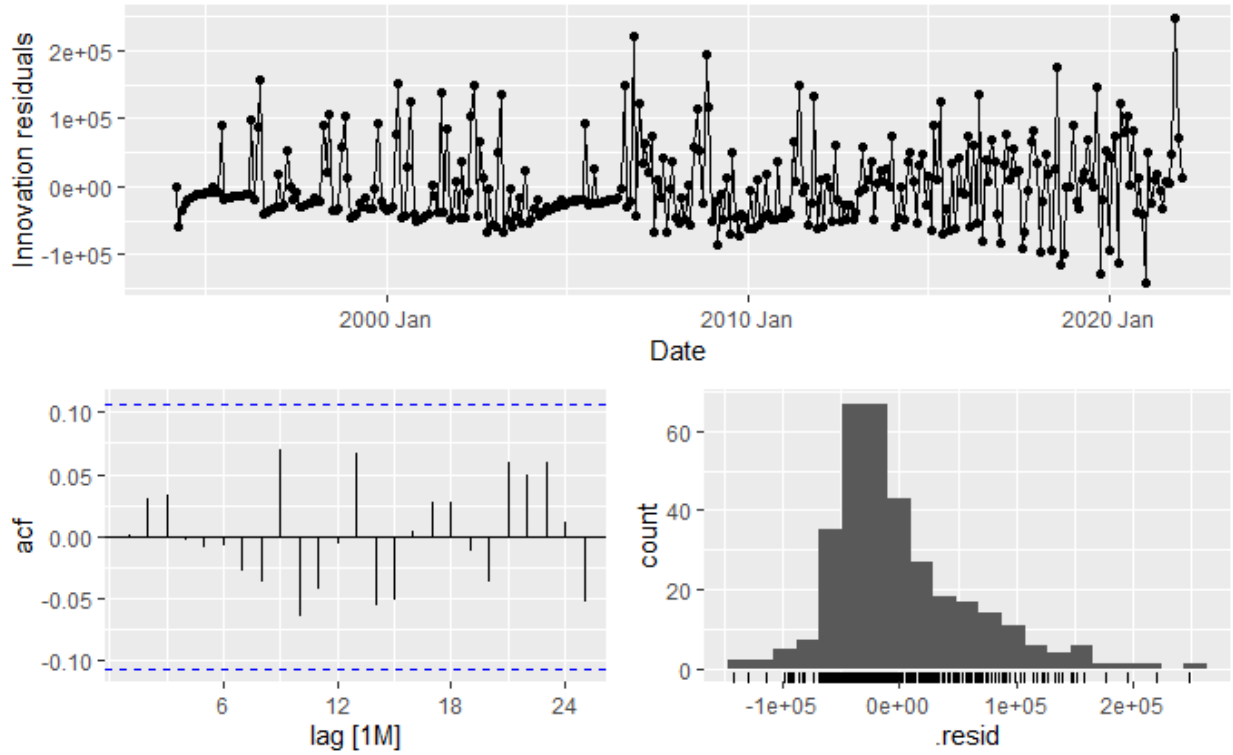
Figure 15 *TSLM Residual Analysis*



The residual plots for the TSLM, seen above in Figure 15, seems to suggest a fairly good model fit. The time plot shows some changing variation over time but is overall not concerning. This heteroscedasticity will potentially make the prediction interval coverage more inaccurate, but the impact is expected to be minimal. The histogram shows that the residuals seem to be approximately normally distributed, which is not required but will improve coverage probability of the prediction intervals. The autocorrelation plot shows several significant spikes, indicating that we should consider alternative or more robust modeling for capturing information left in the residuals.

For the ARIMA model, we see that the AIC value is larger than the TSLM above (8297.54 vs 8017.914, respectively). The best-fitting ARIMA model was determined to be a seasonal ARIMA function with drift. The model definition, $ARIMA(1,1,1)(0,0,1)[12]\ w/\ drift$, shows a nonseasonal first-order autoregressive and moving average component, along with a seasonal moving average piece.

Figure 16 *ARIMA Residual Analysis*



The residual plots of the ARIMA model, seen above in Figure 16, are some cause for concern. The residual histogram shows some fairly significant skew and the time plot of the innovation residuals does not appear to be centered around 0. However, the ACF plot shows no significant autocorrelations. This analysis, combined with the above analysis of the TSLM model drove our team to consider an ARIMAX or dynamic regression model.

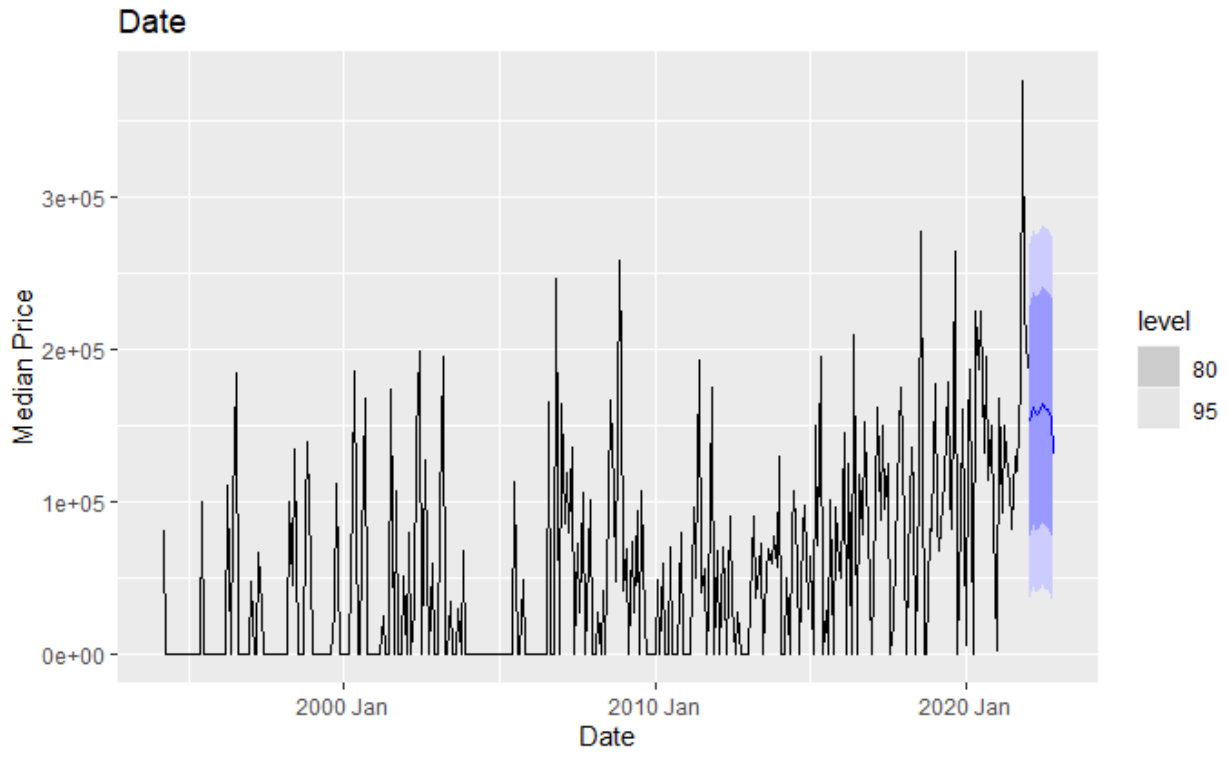
Combining the two approaches above, we can fit an ARIMAX model to see if we can take advantage of both exogenous regressors and ARIMA modeling of the time series component. Fitting an ARIMAX model using the median number of bathrooms as an exogenous regressor gives us a regression with ARIMA(1,1,1)(0,0,1)[12] errors. The AIC value is lower than either the TSLM or the ARIMA models shown above with a value of 7953. Using a Ljung-Box test, we can see that the p-value for the ARIMAX model is 0.2, which means we fail to reject the null hypothesis that the values are autocorrelated. Therefore, based on the improved AIC score, the result on the Ljung-Box test and a check that the residuals had approximately 0 mean, were normally distributed and showed no significant lags on the ACF plot, the ARIMAX model appears suitable for use.

V. Forecasts

With several models developed, our team was able to utilize these models for forecasting. Though the ARIMAX model was identified above as the most suitable model for our application, our team has elected to generate forecasts using both the ARIMA and ARIMAX models, for comparative purposes.

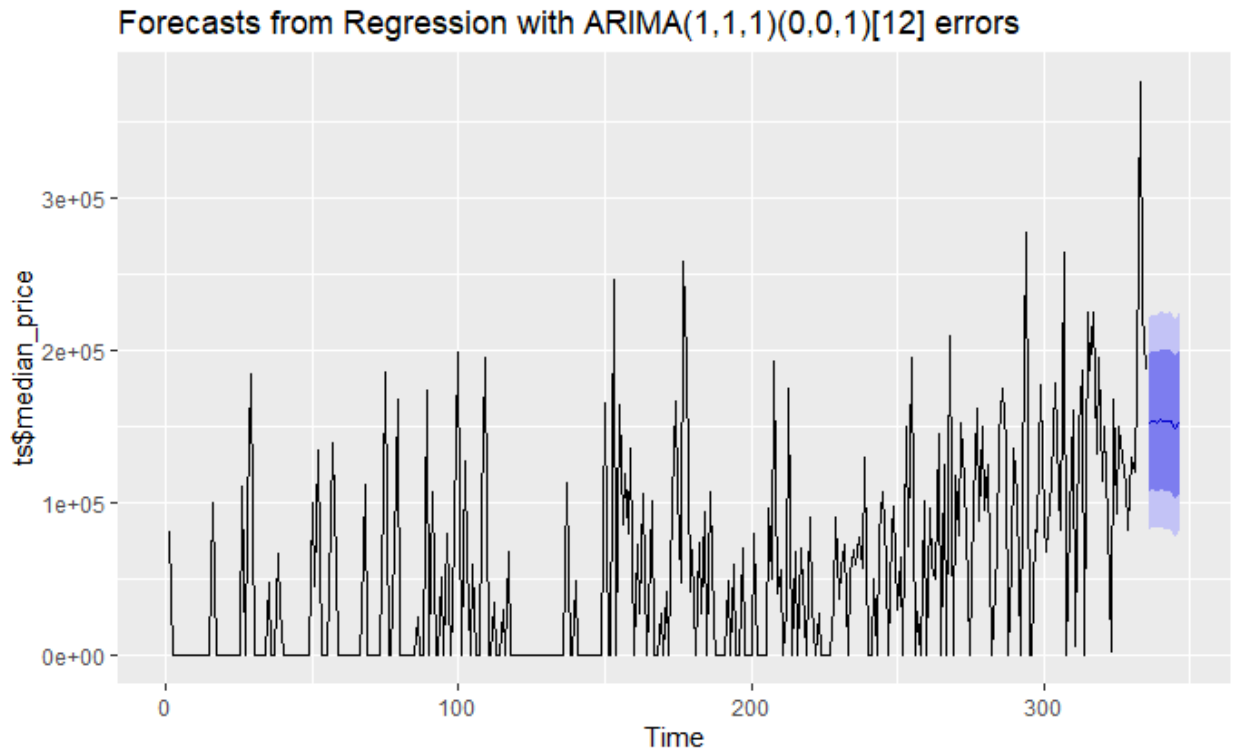
Based on the ARIMA model, we can forecast the expected value for the median price for the next 12 months, as well as bands of 80% and 95% confidence. These forecasts are shown in the figure below.

Figure 17 *ARIMA 12 Month Forecast*



For the ARIMAX model, we obtain the following forecast for the next 12 months, as well as bands of 80% and 95% confidence.

Figure 18 *ARIMAX 12 Month Forecast*



Therefore, forecasting workflows have been successfully developed and will be able to be easily implemented for subsetted data. Additionally, future model improvement will reduce prediction intervals and allow for longer forecast horizons.

Next Steps

Over the upcoming weeks, our team will continue to expand our data set using the web-scraping tool we have developed. In addition, we will make further attempts at improving model accuracy through the addition of new inputs, engineered features and transformations, as necessary. Additionally, our team will begin generation of our web based application and interfacing the application with the workflow outlined above (visualization, modeling and forecasting).