Recap of our question

The purpose of this analysis has been to look at historical data on housing costs to create a 3 year forecast of future prices in five midwestern US states (Minnesota, Wisconsin, Illinois, Indiana, and Michigan). Our goal behind this analysis is to identify which state will likely have the most affordable housing costs for myself and my fiancee in the timeframe we expect to be looking to purchase in. Housing costs have continued to rise over the past several years, so while cost will not be our sole deciding factor, it will influence our search strategy and how we evaluate other criteria such as amenities, proximity to family and our hobbies.

Analysis Strategy

We sourced our data from Zillow.com, a leading real estate marketplace platform in the United States. Zillow has a dataset which contains 24 years' worth of property value data, captured monthly, on hundreds of nationwide properties. From there, we dropped all properties that were not in our target states, which left us with approximately 30 properties per state.

During the course of our EDA, we made a surprising discovery that Illinois, despite having the third largest city in the United States (Chicago), had the lowest average property value among our target states. Because of this we conducted a brief test to ensure that there was no bias we referred to as a 'metro area' bias, in which we have artificially high average property values in some states owing to an overrepresentation of properties in major cities, as opposed to more rural homes which are expected to have lower values. Fortunately, there does not seem to be such a bias; while there were in fact no properties in Chicago for Illinois, no other states had more than a single property in their respective major metro areas.

Our strategy for determining expenses is to investigate each state for outliers, then use all data for each state to train a model. We will then take the result of that model, analyze it by itself against the other forecasts, then combine it with the average historical data for each state to get a full picture of both historical trends and the anticipated 3-year forecast to determine the most inexpensive state in the future to shop in.

Findings

Our findings were surprising fairly early on. We first began by visualizing the average property values for each of our target states from 1/2000 through 3/2024 (Figure 1), discovering that Illinois had the lowest average home value throughout the entire timespan, and by a fair margin. As discussed above, we conducted a brief follow up visualization to discover whether that result was due to a 'metro bias' of varying proportions of properties in urban vs rural areas (Figure 2), but that did not appear to be the case. Finally, we investigated to see whether there were any outliers to make the mean an unreliable metric by plotting the average price per state against all properties in that state (See appendix A), which did not appear to be the case.

Figure 1

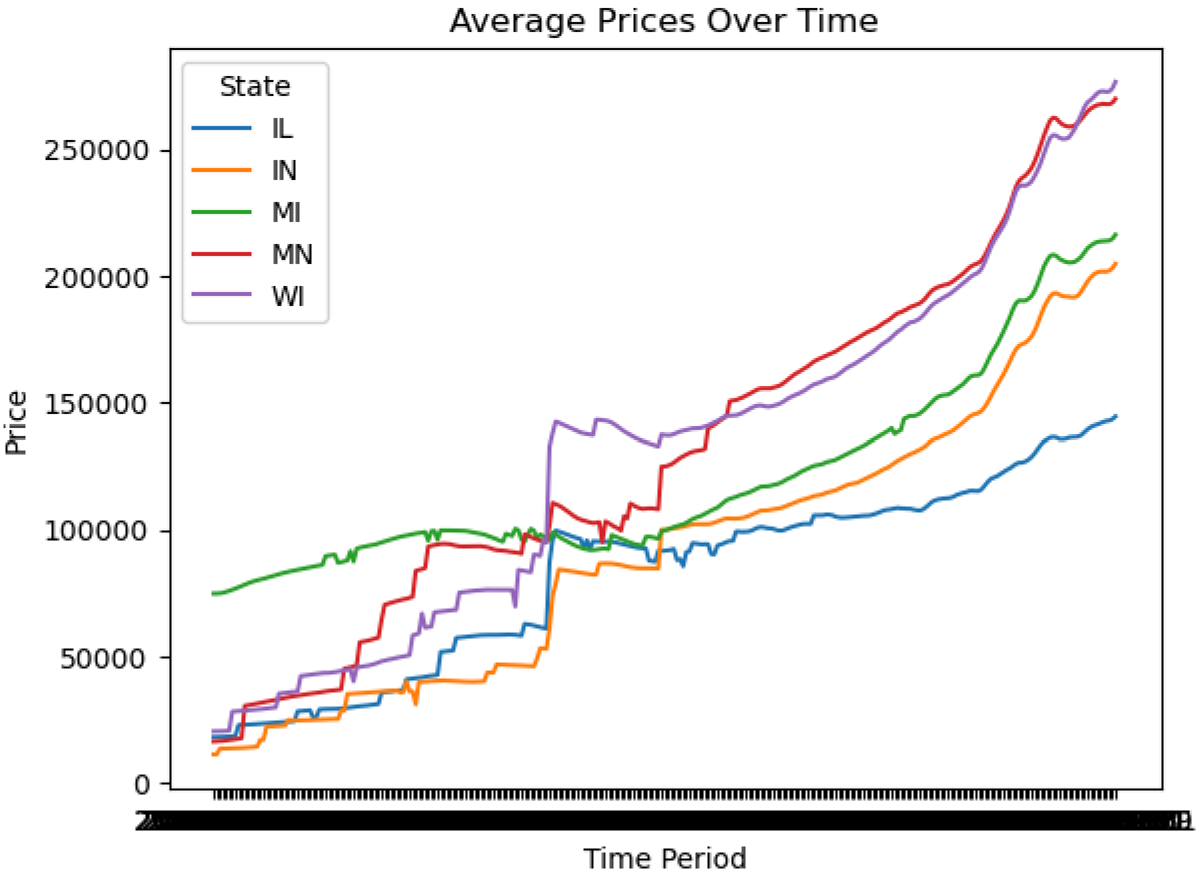
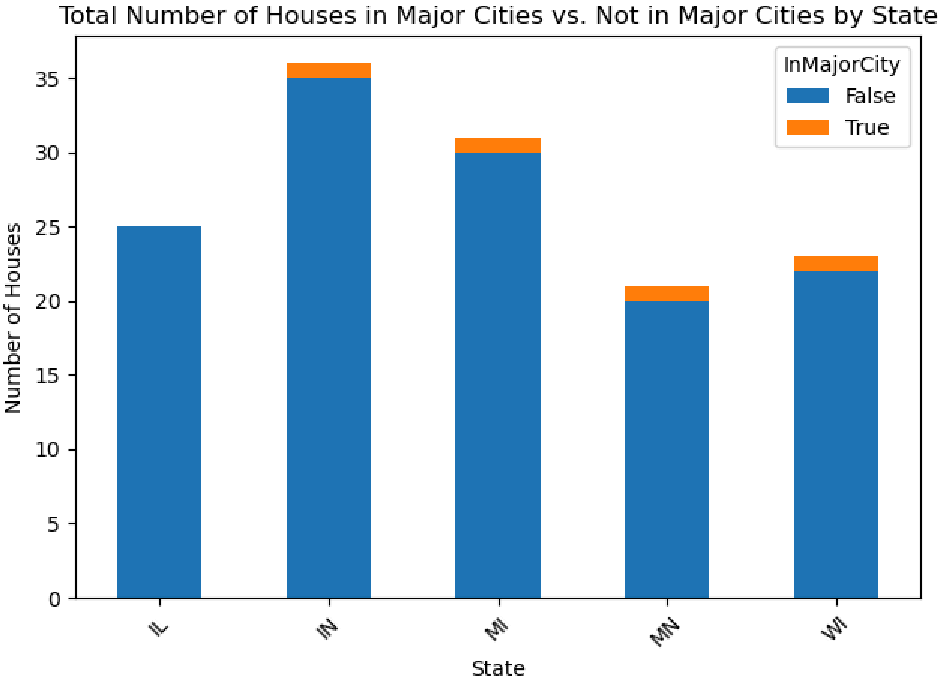


Figure 2



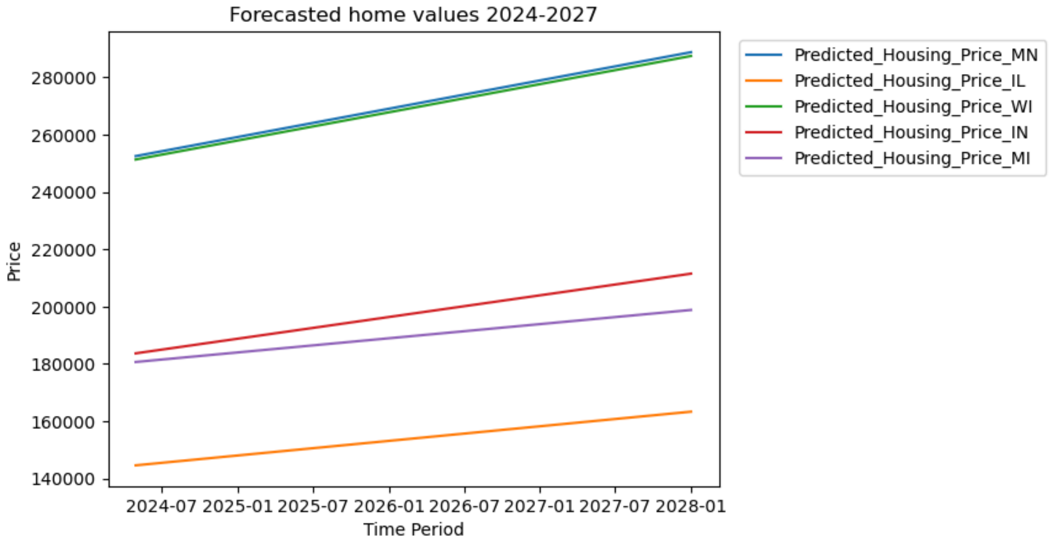
Fortunately, given the fact that all our data is of the same scale, and that we are analyzing our sole categorical variable (each state) as its own dataframe, we didn't need any real data preprocessing aside from splitting the single dataframe we'd been working from into 5 dataframes, divided by state. Our next and greatest challenge then became model selection.

Given the nature of our data, I worked through various types of models, including Linear Regression, Decision Tree Regression and Random Forest Regression. My metric in evaluating these models was Root Mean Squared Error. Decision Tree Regression and Random Forest produced very different results (32666.25 RMSE for Decision Tree, 22572.99 for Random Forest), a simple Linear Regression model proved the most effective (16199.03 RMSE).

When it came time for prediction, rather than randomizing the data with a train/test split (this would not make sense for time series data), we allowed the model to evaluate all available data in order, while making predictions for 45 future months (in this case 01/2000 through 03/2024 were given, 04/2024 through 12/2027 were predicted). One key detail to understand is that the forecasted data will not appear at all like historical data when visualized - without much more robust data to influence month to month predictions, the linear regression model will produce a straight line of values, predicted at a slope it believes most accurately predicts future values given historical values.

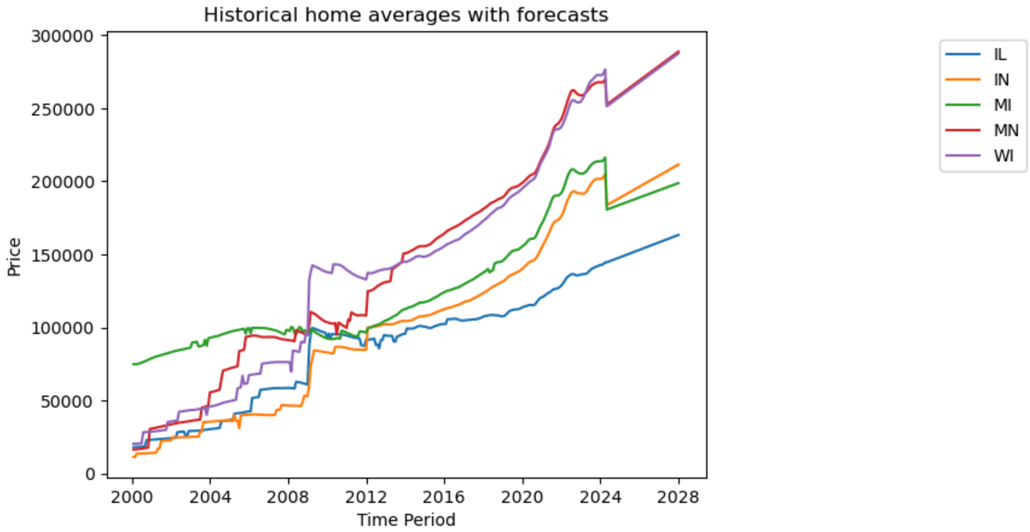
A brief glance at the forecasted future values (Figure 3) provides trends we may have expected, but also some unexpected insights. Illinois remains the most inexpensive state forecasted and Minnesota remains the most expensive, but Minnesota and Wisconsin are forecasted to grow so similarly that on a house by house basis, the two markets are likely to be functionally indistinguishable in price over the next few years. Another interesting feature we see is that while Indiana and Michigan start fairly close to one another in value, Indiana properties are forecasted to increase in value at a significantly greater pace than those in Michigan. This is surprising, given that although Michigan has two of the most dangerous cities in the United States (Detroit and Flint), there was only one property in this dataset between them, and Michigan is a state of much greater natural beauty (considering lakefront property, national parks and tourism).

Figure 3

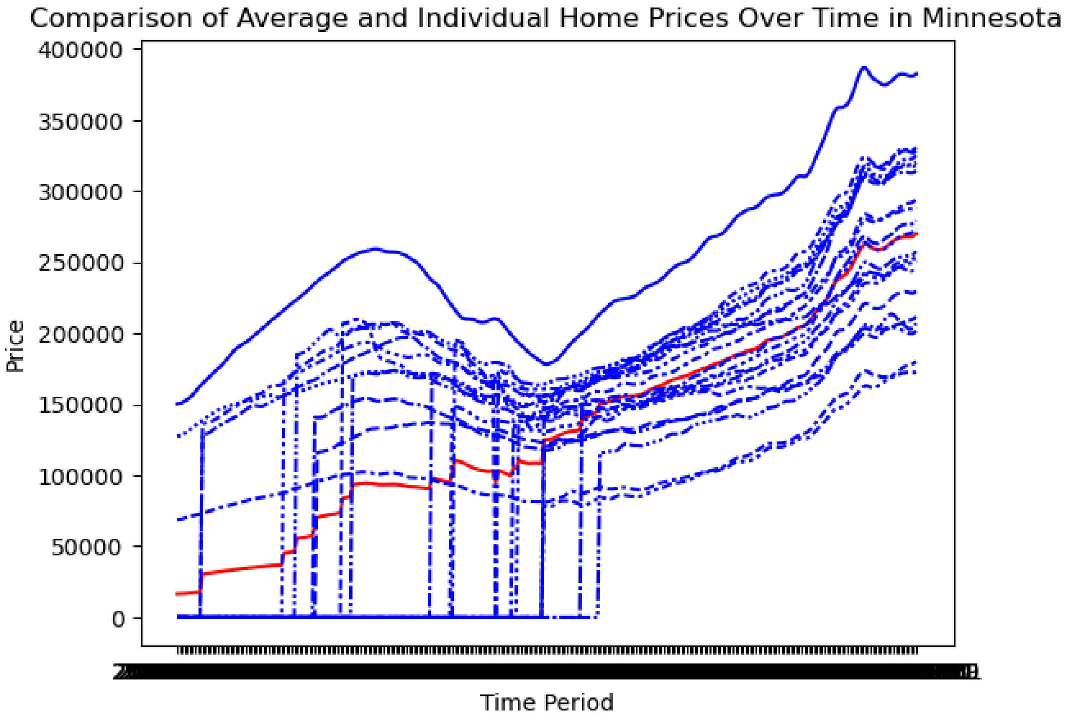


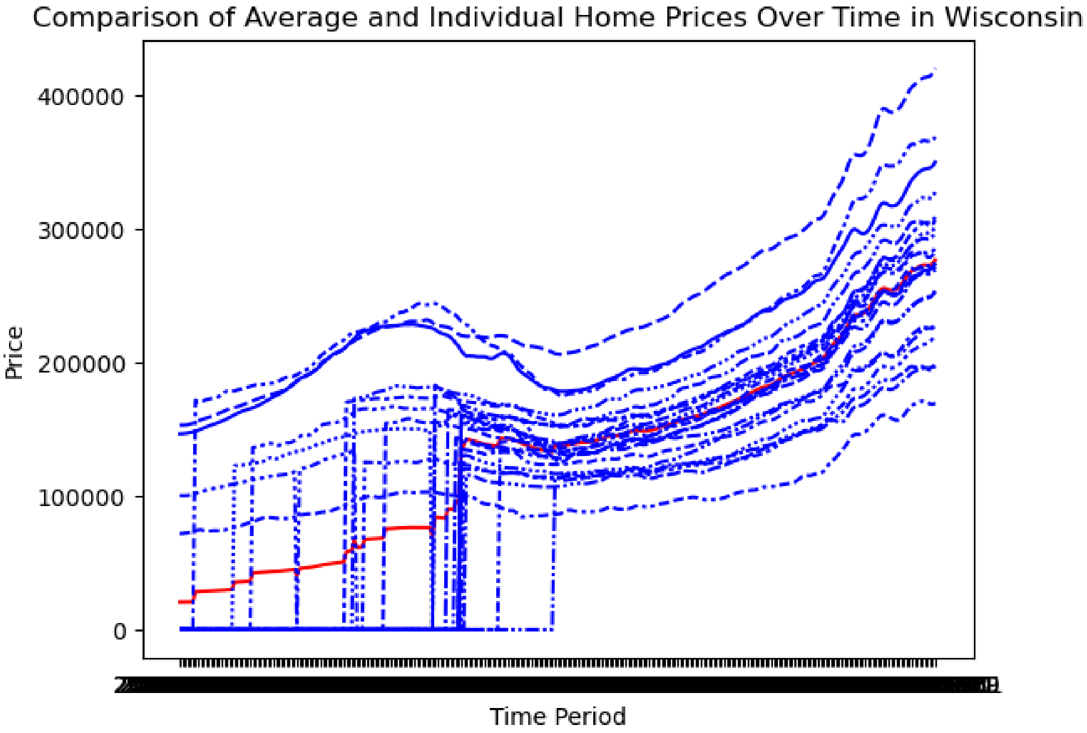
Finally, once we look at the historical data alongside the forecasts (Figure 4), we see that the forecasts largely reflect what we may have expected to see given the historical data by itself, with the exception of an unexpected jump of Indiana home values over Michigan. Further, we see that the divide which seems to have been forming among the three groups of states (MN and WI, IN and MI, and IL) after 2012 appears likely to continue to widen over the coming few years. It may be an interesting future project to look into the roots of what happened to cause the differences in values over the past decade, as well as to look at cost of living per state to see if home prices are an accurate benchmark of overall costs to live in a state, or if some states with higher property values are actually cheaper to live in than others with the proper considerations.

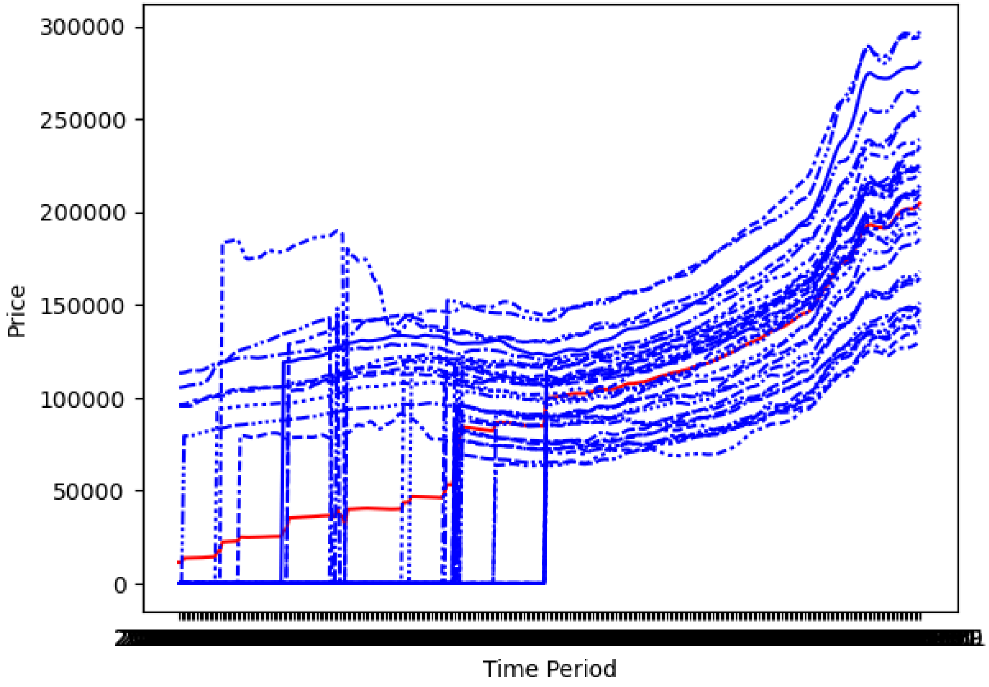
Figure 4



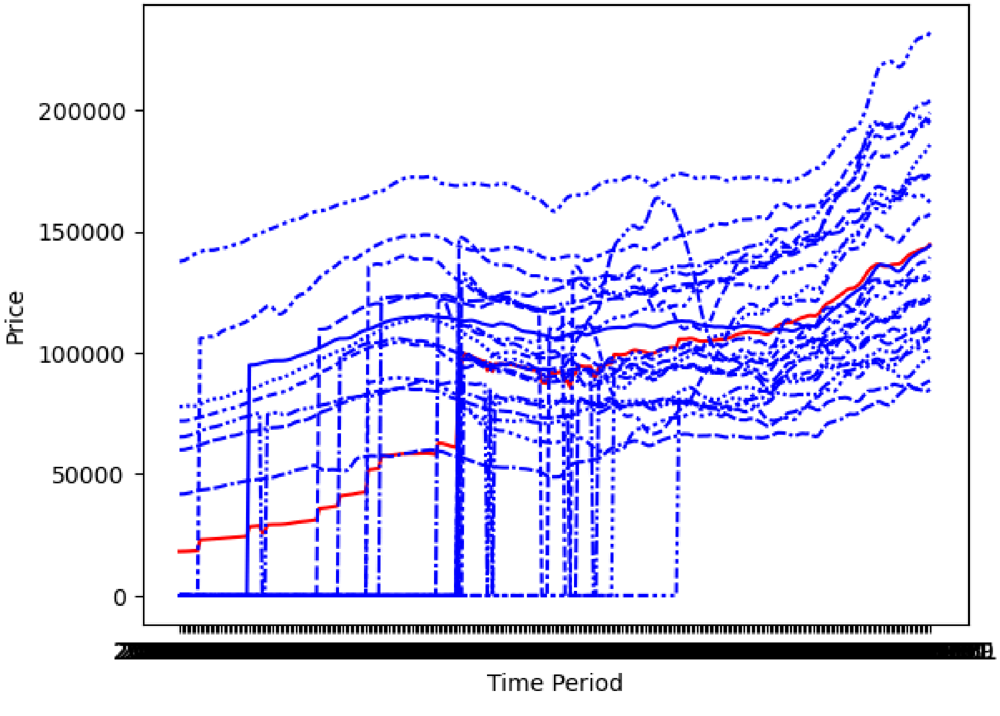
Appendix A


Comparison of Average and Individual Home Prices Over Time in Minnesota

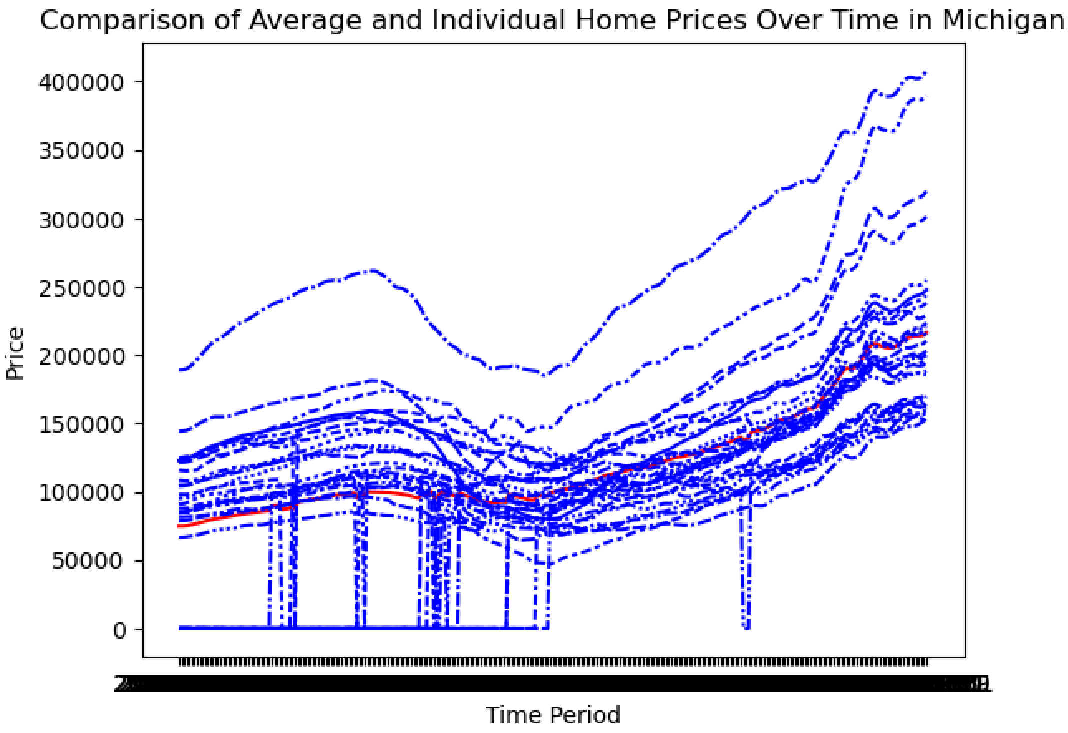

Comparison of Average and Individual Home Prices Over Time in Wisconsin



Comparison of Average and Individual Home Prices Over Time in Indiana



Comparison of Average and Individual Home Prices Over Time in Illinois


Comparison of Average and Individual Home Prices Over Time in Michigan
Price
400000
350000
300000
250000
200000
150000
100000
50000
0
Time Period