Improving Home Price Prediction Accuracy

Oscar DeHamer

Heidi Lin

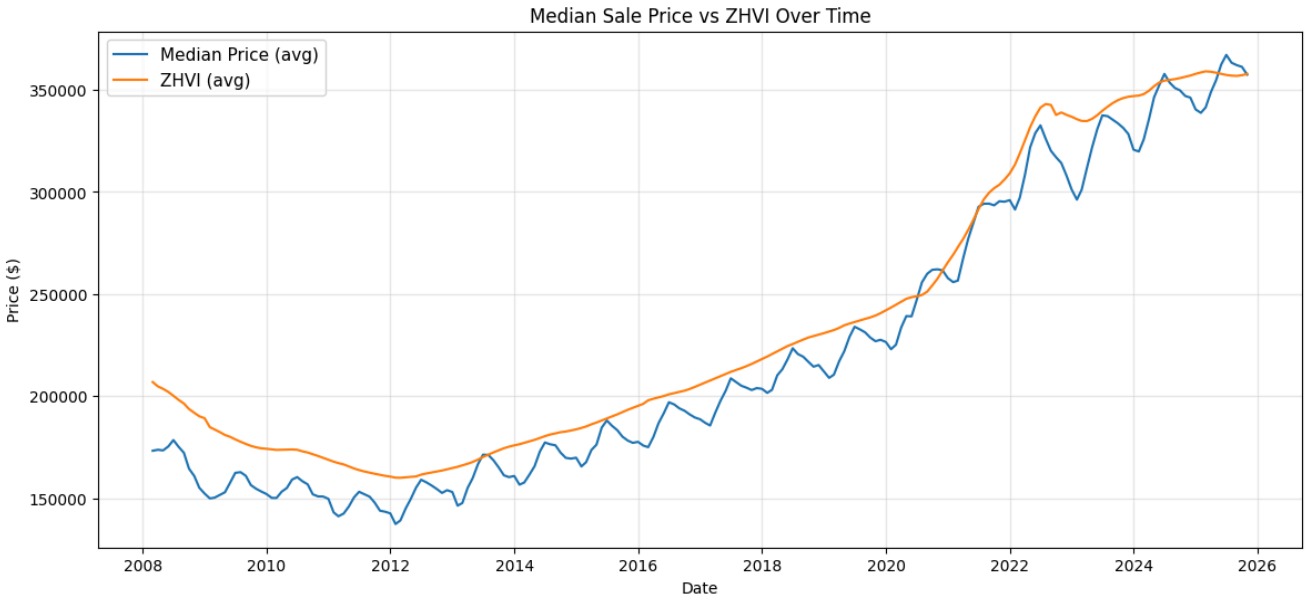
Andre Lasola

The goal of our project is to determine whether there are inefficiencies or biases in the most popular home pricing models and to explore whether we can leverage data from multiple sources to build a model that more accurately predicts a house's value.

The datasets we are using come from Zillow and Redfin. We are using the Zillow Home Value Index (ZHVI) dataset, which is a measure of the typical home value in a given region. This dataset is separated by metro areas, with a column containing their estimate for the price of the typical house in that region for every month since January 2000. The unit of observation in this data set is the typical home value estimate for each region over time. We are also using Zillow's median sale price dataset. This dataset is laid out similarly to the Zillow Home Value Index, with each row being a region, and there being columns for every month since January 2008, containing the median sale price during that month. The unit of observation for this data set is the median sale price every month. We did have to do some data cleaning with these two datasets. Since the column names were dates, we had to turn those into datetime objects and melt them to turn them into rows instead of columns to better graph them. Redfin's version of the ZHVI is the Redfin Home Price Index (RHPI). This dataset is laid out differently from the Zillow dataset, though. Instead of showing the price like the Zillow Home Value Index does, it shows the

month-over-month change in their models' pricing for each region, so it will take some cleaning to convert it to something that can be used alongside the Zillow datasets.

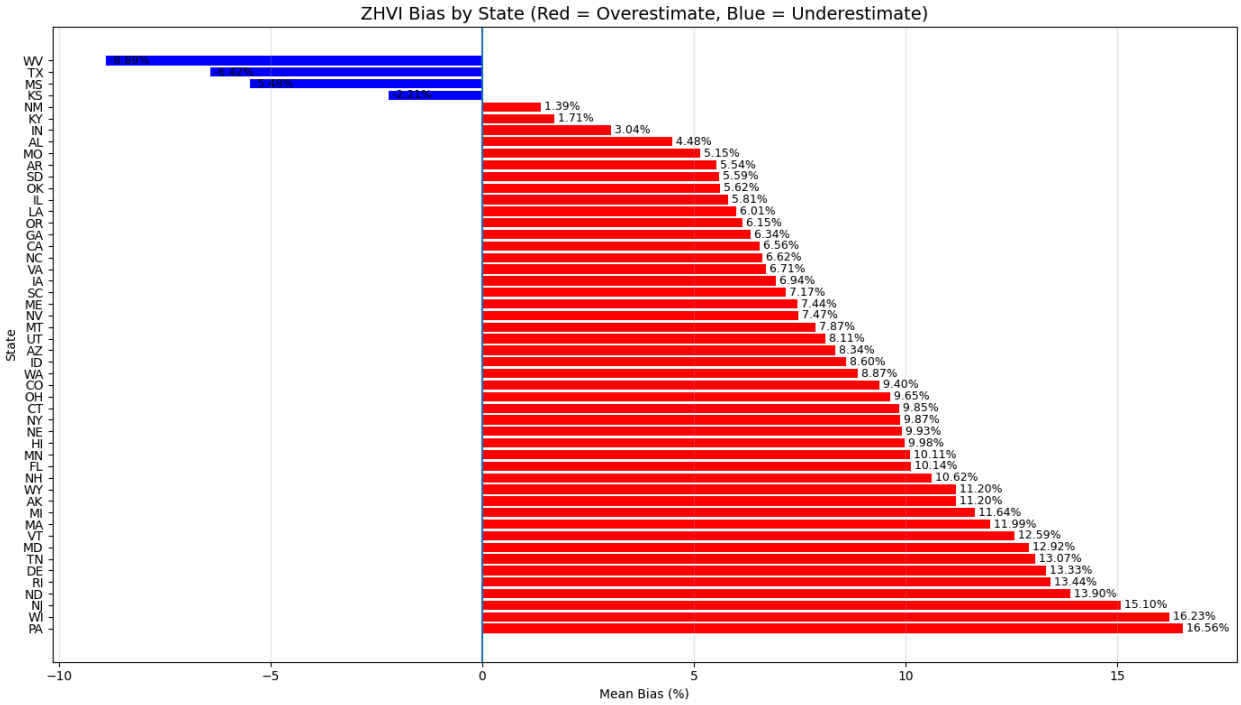
We want to find how well of a job the current models for predicting home prices do, so first, we decided to compare the median sale price with Zillow's Home Value Index over time. We did this for every region and date in the datasets, and we were able to create this graph to show the median sale price vs ZHVI over time.



We noticed that the ZHVI is almost always greater than the median sale price, so there might be some bias in Zillow's estimates and possible room for improvement for our model. We decided to find the mean bias of ZHVI by comparing it to the median sale price at each point. We did this by using the formula $Bias = \frac{ZHVI - Median\ Sale\ Price}{Median\ Sale\ Price} \times 100\%$, and then finding the average of all the biases. Doing so, we were able to find that on average, the ZHVI is 7.76%

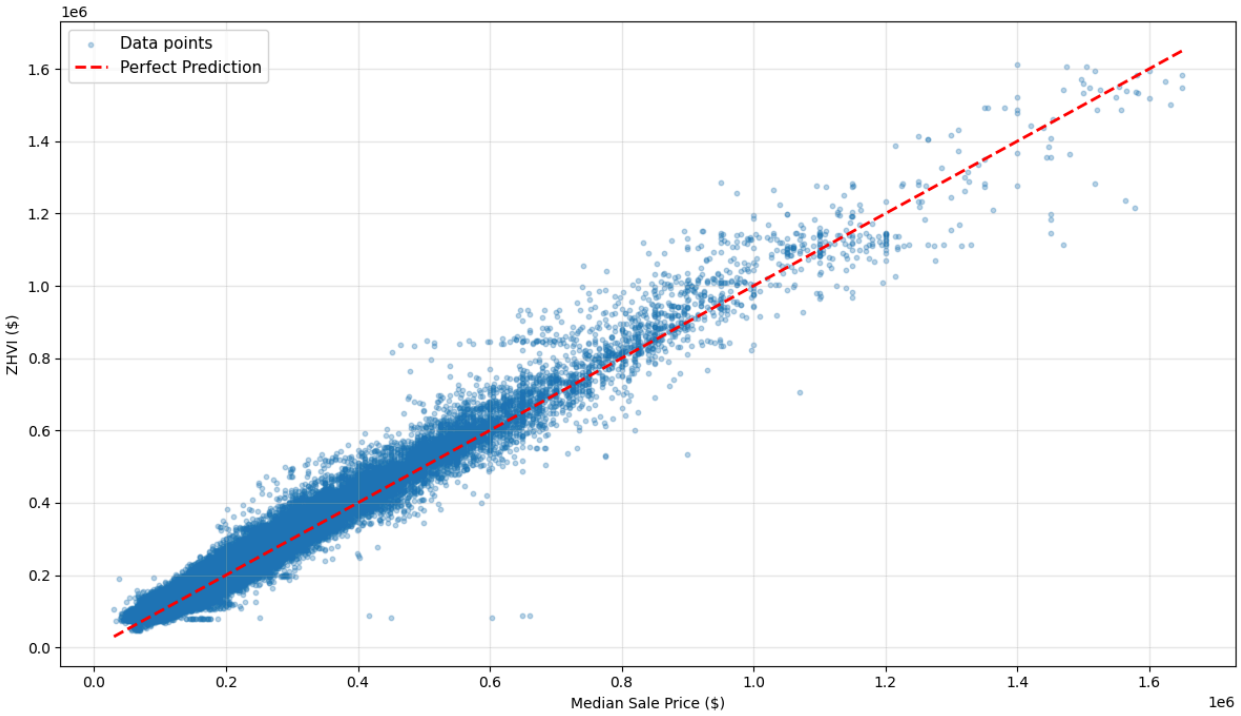
higher than the median sale price. This means that if a region has a median sale price of $400,000, ZHVI would, on average, value it around $431,000.

Next, we decided to find out if there is any pattern to the bias by state, so we sorted all the data and found the mean bias in each state, giving us this graph.



From this, we can see that the ZHVI model has vastly different results depending on the states, but in all but four states, it overestimates how much homes are worth.

We also decided to find the Root Mean Squared Error and the Mean Absolute Error of the ZHVI compared to the median sale price. Comparing the two on a scatter plot, you get:



You can see the bias in the graph, as there are more data points above the perfect prediction line than below it. We found that the Root Mean Squared Error is $29,971.41 and the Mean Absolute Error is $21,358.23, so hopefully our model will be able to beat that.

So far, we have performed several analyses to understand how well Zillow's Home Value Index aligns with actual market data. To make the data easier to work with, we melted the datasets from wide to long formats, which allowed us to graph and analyze trends more efficiently. Our first analysis involved comparing the median sale price with the Zillow Home Value Index over time for each region. By plotting these two measures, we observed that the ZHVI was almost always greater than the median sale price, suggesting a consistent overestimation of home values by Zillow's model. To quantify this, we calculated the bias

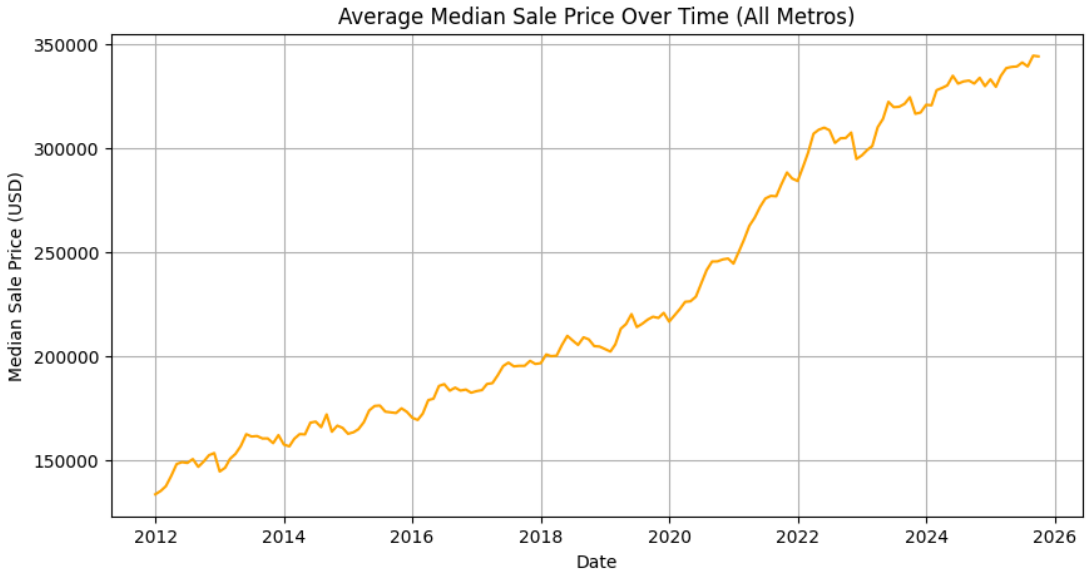
between ZHVI and the median sale price and found that, on average, the ZHVI is 7.76% higher than the actual median sale price. Meaning a home that sold for $400,000 would typically be estimated as being worth $431,000 by Zillow.

Next, we examined whether this bias varied by state. We grouped the data by state and calculated the mean bias for each one. By visualizing these results, we discovered that while the degree of bias differed across states, Zillow's model overestimated home values in nearly every state, suggesting Zillow's model may systematically overvalue homes.

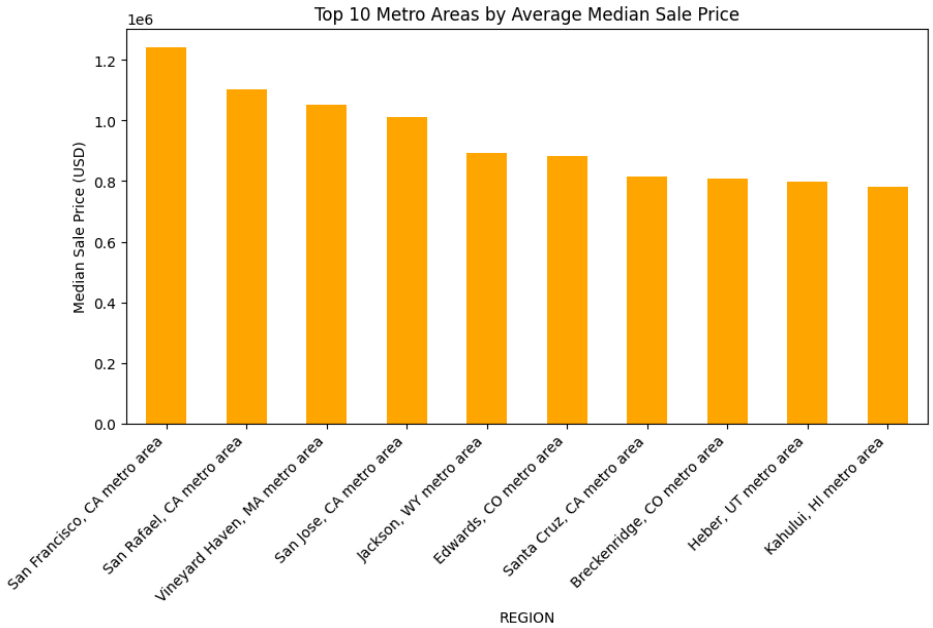
Our implicit objective function so far has been to minimize the prediction bias between model estimates (ZHVI) and actual observed market prices. Using the formula $Bias = \frac{ZHVI - Median\ Sale\ Price}{Median\ Sale\ Price} \times 100\%$, we're able to measure the relative over- or underestimation of ZHVI compared to real sale data.

In addition to Zillow's datasets, we explored Redfin's Metro Monthly Housing Market Data (which can be found here), which includes monthly statistics for each metro area from January 2012 to the most recent release. This dataset reports median sale price, number of homes sold, new listings, inventory, and other market activity measures. Compared to Zillow, Redfin's metro datasets were much more complex and harder to interpret, with extra variables and property-type breakdowns that made the raw files confusing to interpret. As a result, we had to perform substantial data cleaning and filtering before we could obtain a usable dataset.

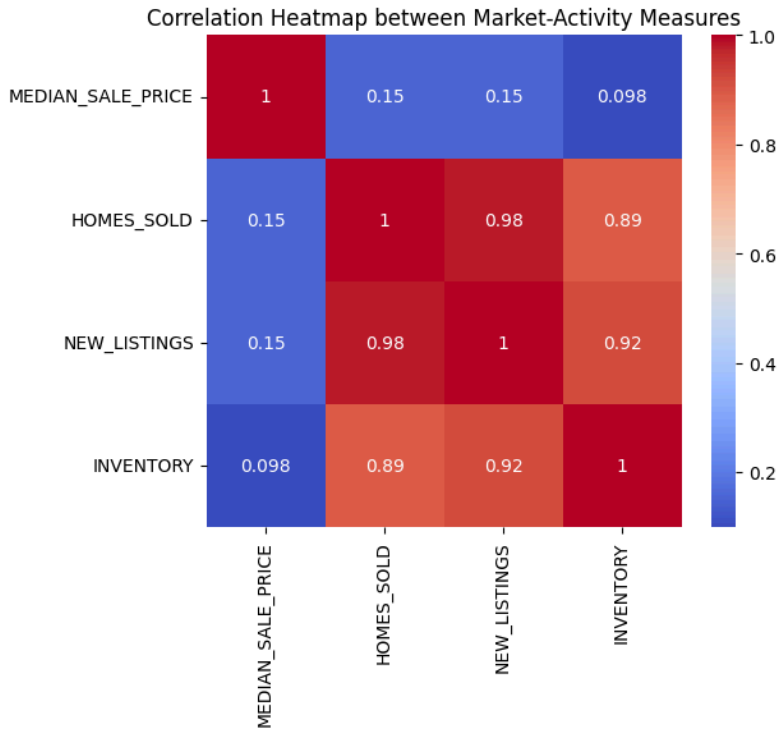
Using this cleaned dataset, we first compute the average median sale price across all metros for each month. This gives us a price trend that confirms a steadily increasing price level that gradually increases in the post-pandemic period.



Next, we calculated the average median sale price for each metro area since 2022 and plotted the ten most expensive metros, which highlights how concentrated the highest price levels are in a small set of markets.



Finally, we examine how Redfin's key variables are related by computing a correlation matrix for median sale price, homes sold, new listings, and inventory.



The resulting heatmap shows that three market-activity measures, namely "homes sold", "new listings", and "inventory", are strongly positively correlated with each other, and each is also moderately correlated with the median sale price.

Some future analysis we have planned is to integrate Redfin's Home Price Index (RHPI) and compare its bias patterns with Zillow's to identify which is more accurate or consistent. We are also going to develop our own predictive model combining Zillow and Redfin's data, and potentially more features, to try to make a better model for predicting home prices. We will explore using Multiple Linear Regression and Random Forests to try to make a better model and evaluate using cross-validation and test-train splits to ensure robustness.

Some difficulties we've run into so far were how to visualize the Zillow data, as it was organized in a way that we have not worked with yet. We had to learn how to convert the data from a wide form to a long form using the melt function, making it easier to visualize. Another issue we ran into is that the Redfin data is much different from the Zillow data that we have. It is already set up in a long form, but instead of having their price estimates for each region during the time period, it has the percent change in their price estimate from last month. This makes it difficult to compare to the median sale price that we already have. We're going to have to find the percent change in price for every month in the median sale price and compare that to the Redfin model to find how well it does on predicting home prices, and how we could use it to create a better model.

We still have a long way to go, but we have made progress and found that there is room for improvement for these models. We hope to have our model that uses the data from both Zillow and Redfin, which can compete with their models of pricing homes by removing the bias that we have found.