US Home Price Prediction Report

Problem Statement

This project focused on developing a predictive model that forecasts home prices by location over the next five years using historical real estate data sourced from Zillow. I aimed to forecast the prices for cities, states, and the US as a whole to give buyers a better idea of how the value of homes will change over the next five years. Accurate predictions will assist buyers, sellers, and investors in making informed decisions regarding property transactions and investments. The challenge lies in effectively analyzing and interpreting the large dataset to derive meaningful insights and create a reliable predictive model.

Data Wrangling

The initial step involved collecting extensive historical housing data from Zillow dating back to 2008. The data included only single family homes and included features: Population Rank, Location, Number of Sales, Mean Sales Price (for each location per month), Home Value, and Market Heat Index score. To prepare the data for analysis and modeling, I performed steps to clean, transform, and engineer the features.

Data Extraction

I retrieved the data in CSV format from the Zillow website. I brought in data from 4 different CSVs: Sales, Mean Sales Price, Home Value, and Market Heat Index.

- Sales: Number of home sales per month in each location
- Mean Sales Price: Average sale price per month of homes in each location
- Home Value: Zillow estimated average home value per month
- Market Heat Index: captures dynamic of supply and deman in the market. Higher number is associated with a seller's market.

Each of the CSVs also included data around where the population ranked among US cities and location data. The data ranged from Feb 2008 to January 2025. Market Heat Index only had data from 2018-2025 as it was a newer measurement for Zillow.

Data Cleaning

The data came in with a lot of missing values, specifically in the home value and mean sales columns. For these values I used forward fill to impute the missing values as this would be a good way to keep the linear consistency with the data. To prevent skewing the data, I limited the forward fill method to 6 months and dropped the remaining values.

Other cleanup items involved removing the US row and removing irrelevant columns. I needed to remove the US row as this was a mean of all other cities that would have skewed the data. The mean of our data would also be easy to attain when needed. I removed the columns 'RegionID' and 'RegionType' as these provided no value to our data. 'RegionType' was the same for all states so it was irrelevant.

Data Transformation

Each of the CSVs from Zillow was formatted that each date was a column. To fix this, I transposed the dataframes and created a date column so the dataframes would be easier to read vertically. This would also make combining the data frames easier.

After each of the dataframes from the four CSVs was transposed correctly I combined the dataframes into one for evaluation. This created one data fram with our target variable (mean sales price), and all of our features to make for easier analysis and model training.

I also created separate month and year columns to make it easier to analyze our features in EDA.

Feature Engineering

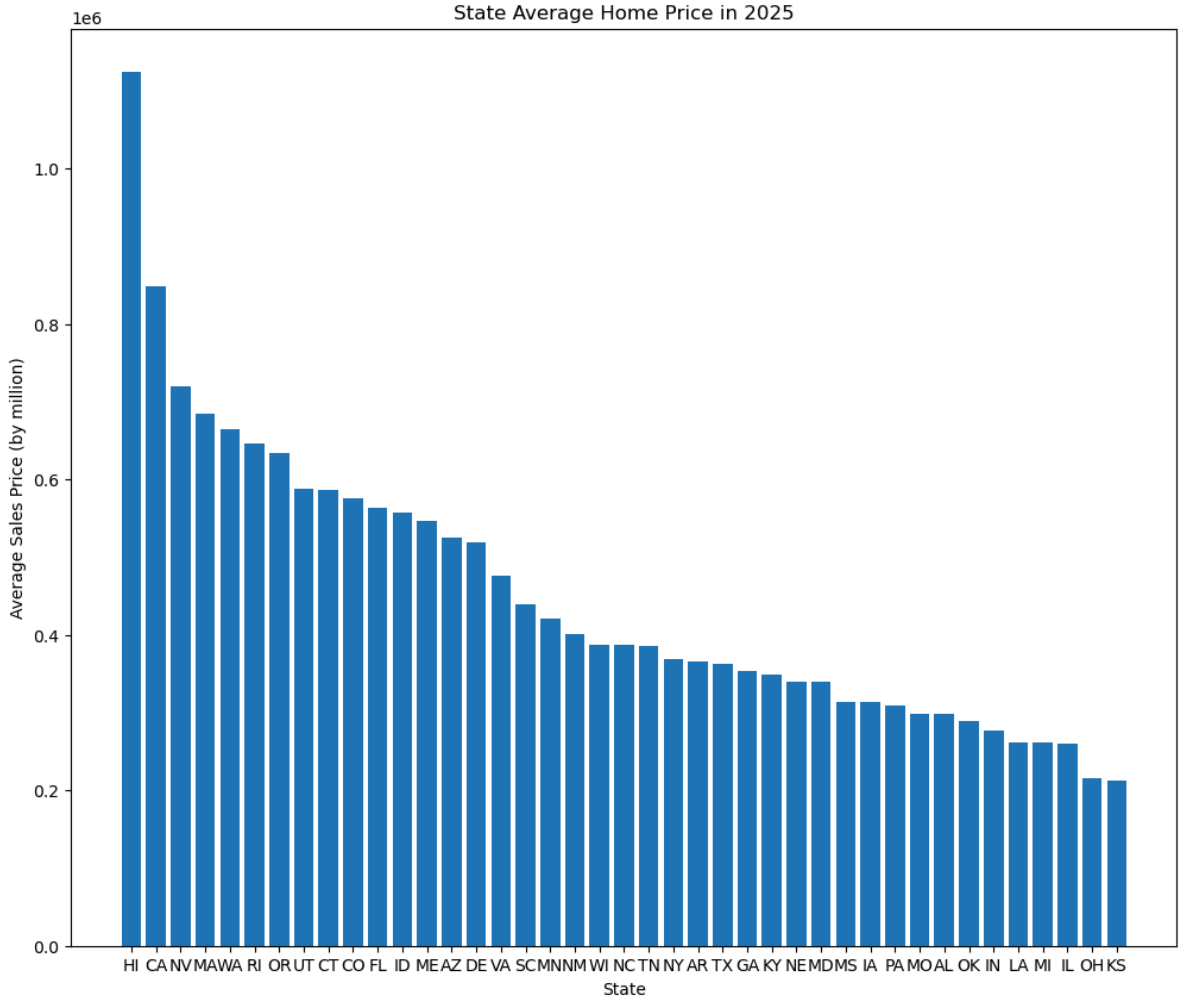
There was not much work needed in feature engineering. One feature I created was value surplus. This took the difference of the sales price and home value. I thought this would help see how often the average sales price would go over/under home value.

Another thing I noticed was that some cities only had a few years worth of data. When making predictions, it would be tough to have a model accurately predict without the proper historical data for each city. I removed 9 cities (of 149) that had less than 5 years of data and kept the rest.

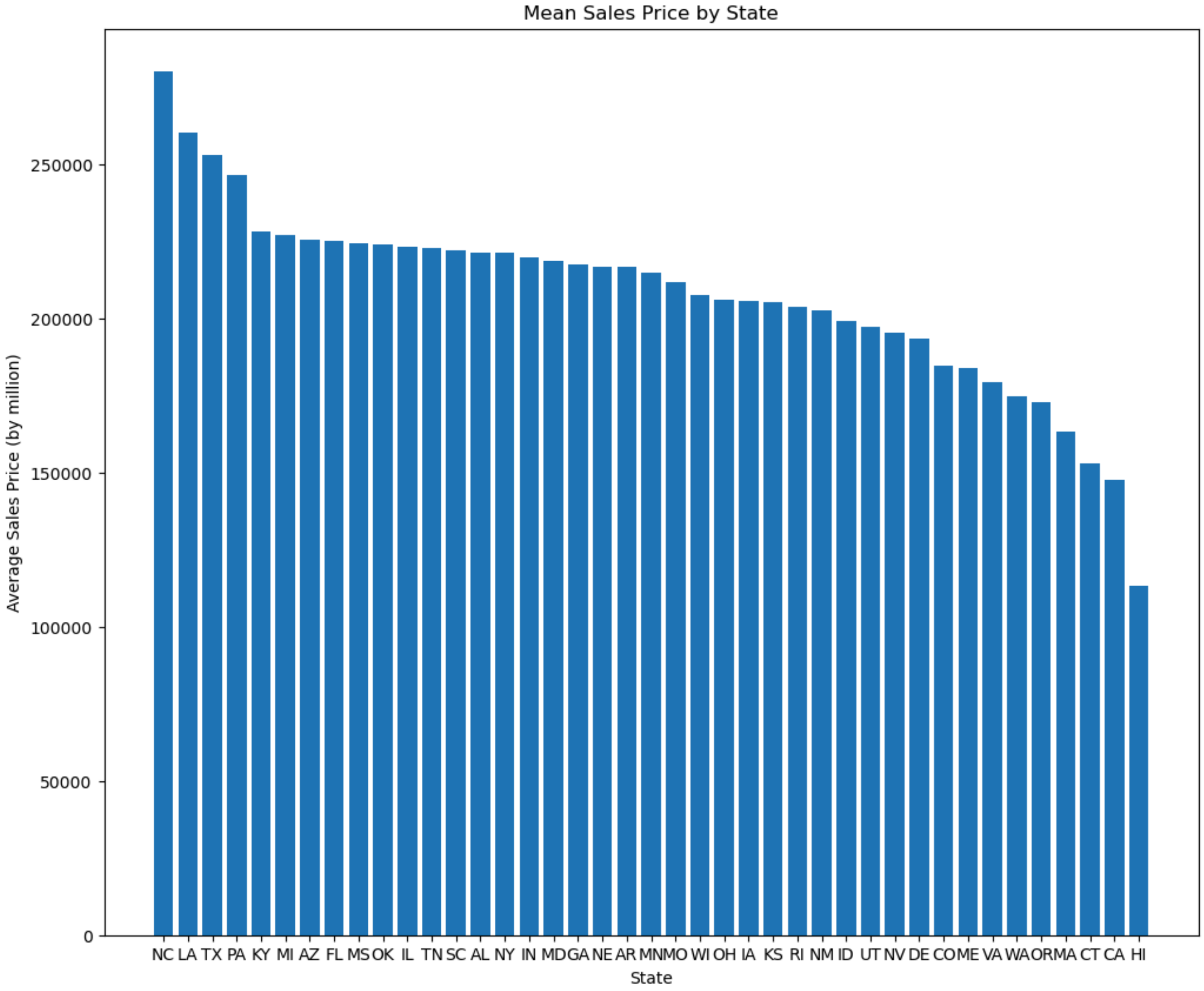
Exploratory Data Analysis (EDA)

State and Year-over-Year breakdown

To start exploring the data, I wanted to review the breakdown by state and over time. First, we'll look at a scope of the current prices across each state. Here we can see that most states are currently sitting between $300,000 to $600,000. Hawaii came in the highest at $1,123,507 and is the only state with over $1,000,000 average home price in 2025. California is not far behind at $800,000. Ohio and Kansas are the only states currently under $250,000. This could signify an opportunity to buy in cheap and see an increase.

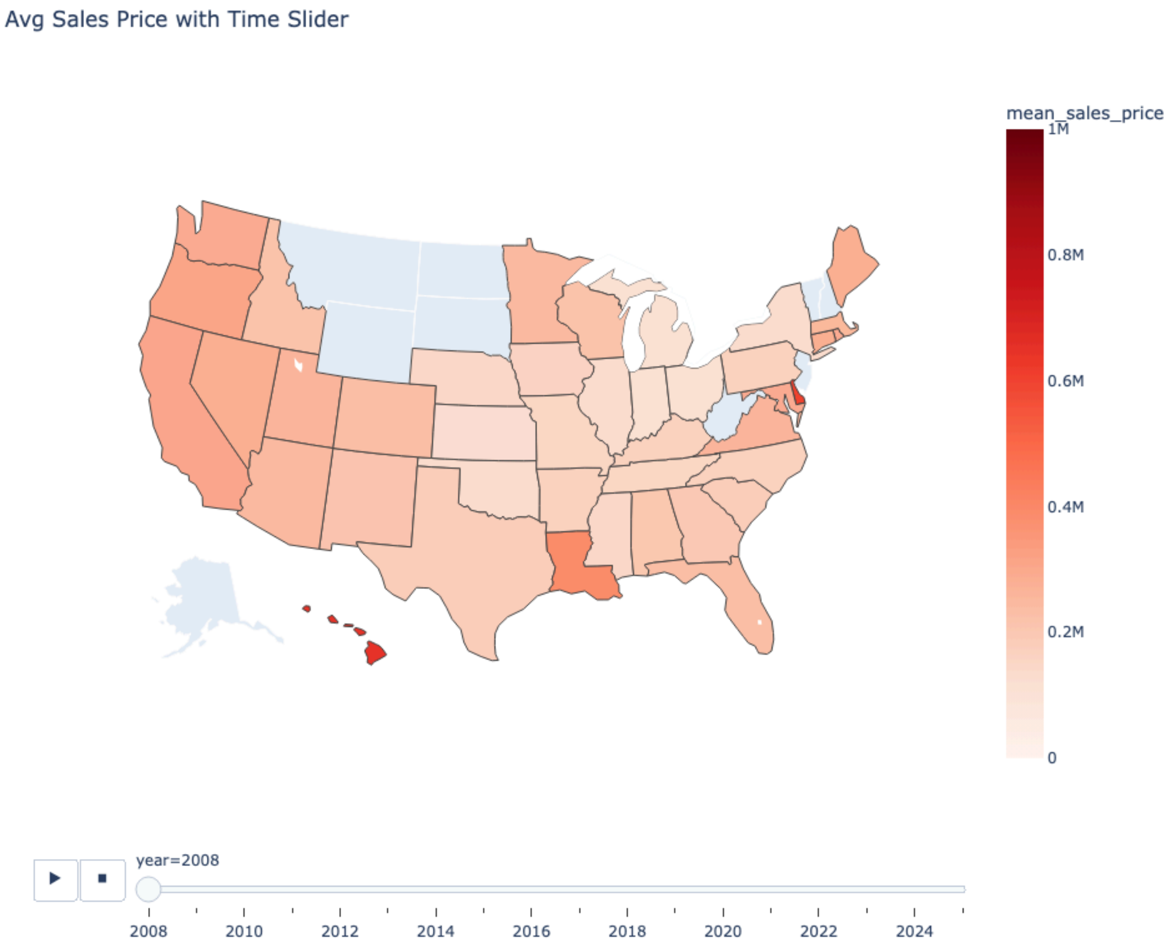


In the following chart, I took the average sales price for each state over the period of 2008 to 2025. We can see here how drastically the real estate landscape has changed over the years. California and Hawaii were at the bottom over the 17 year period but are the highest now in 2025. Other states saw a pretty stark decrease in their rank such as Louisiana and Texas.



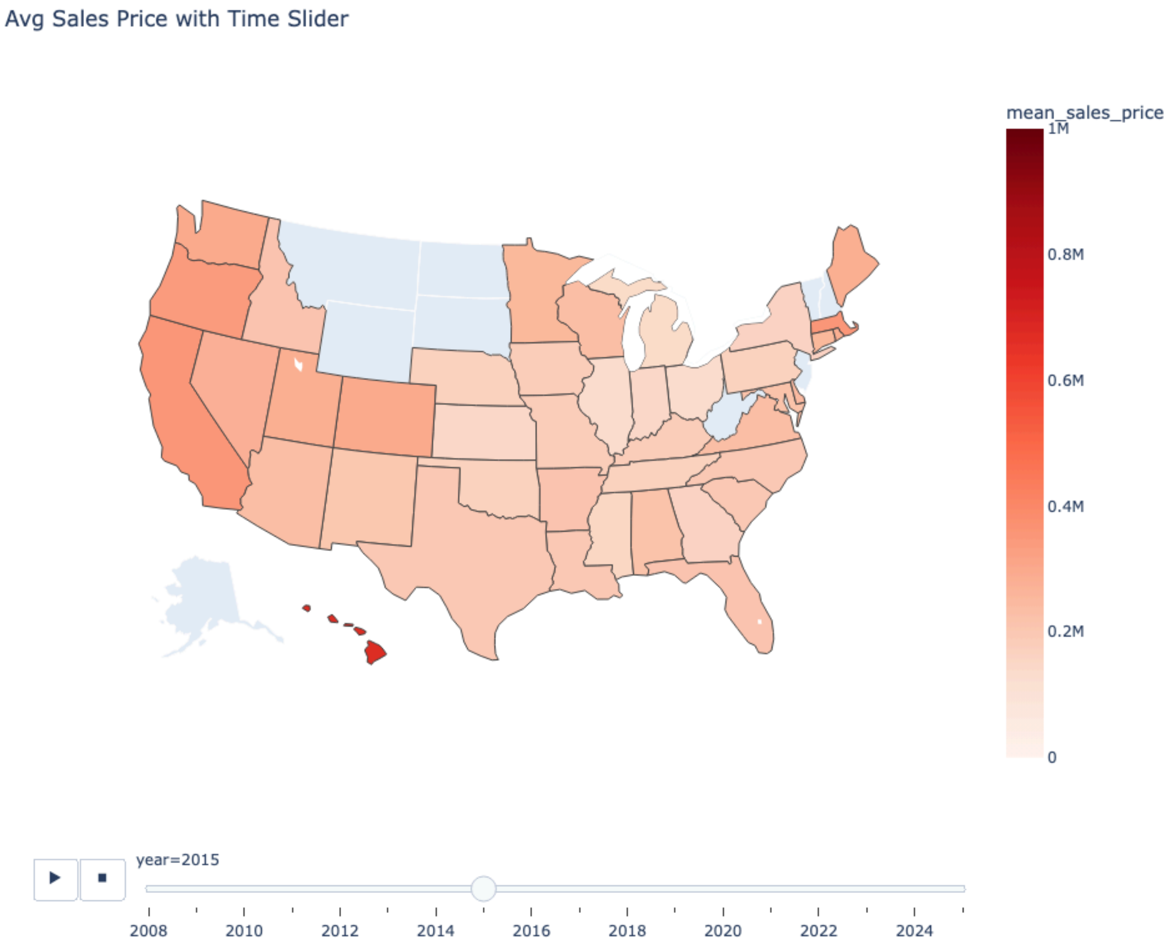
Given the significant change in home prices by state for their average and current prices, I wanted to visualize how it has evolved. This visualization would be easiest with a map.

Since our data started in 2008, we can see how home prices were distributed from the start:

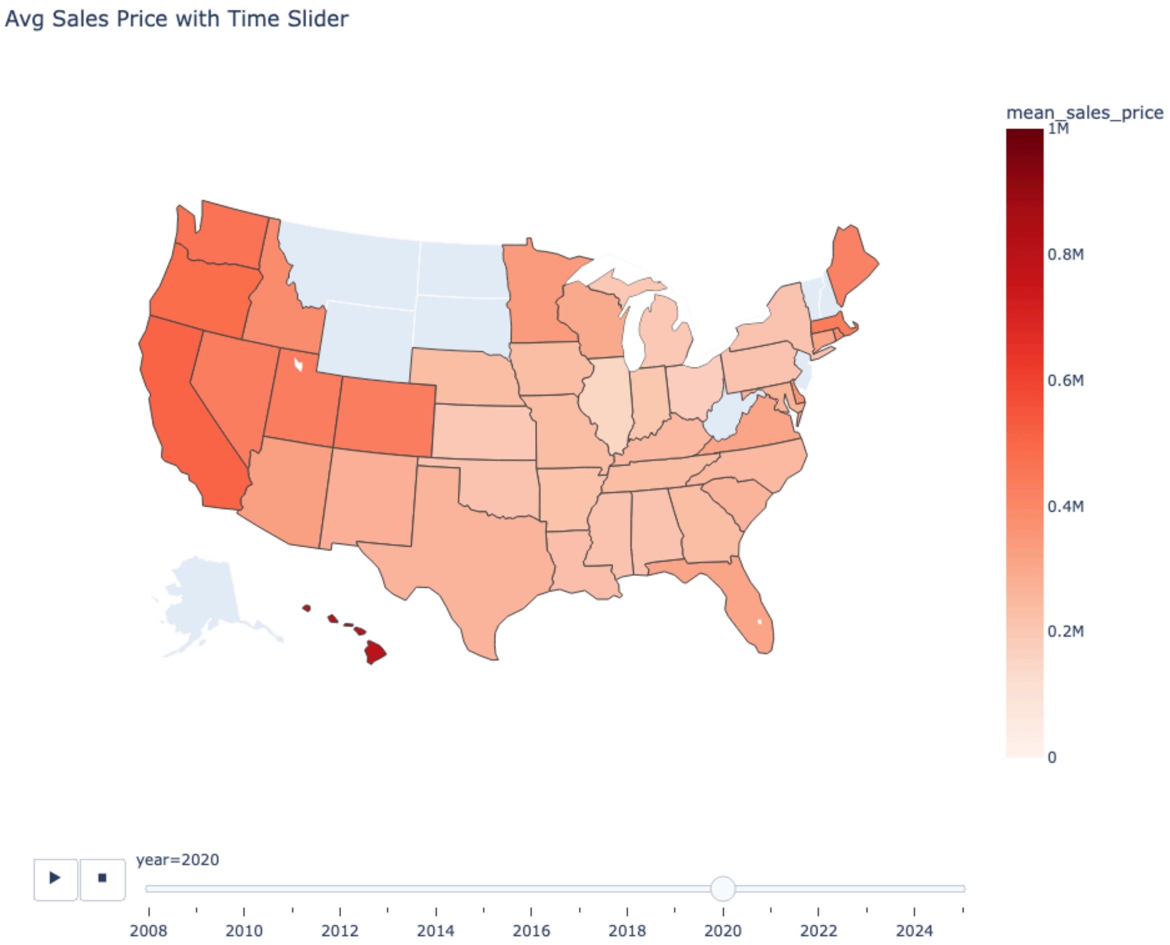


We can see that prices were pretty even across the board and everything was under $400,000 to $500,000. As we move into 2015, the map doesn't change all that drastically. We see slight increases in some west coast states and

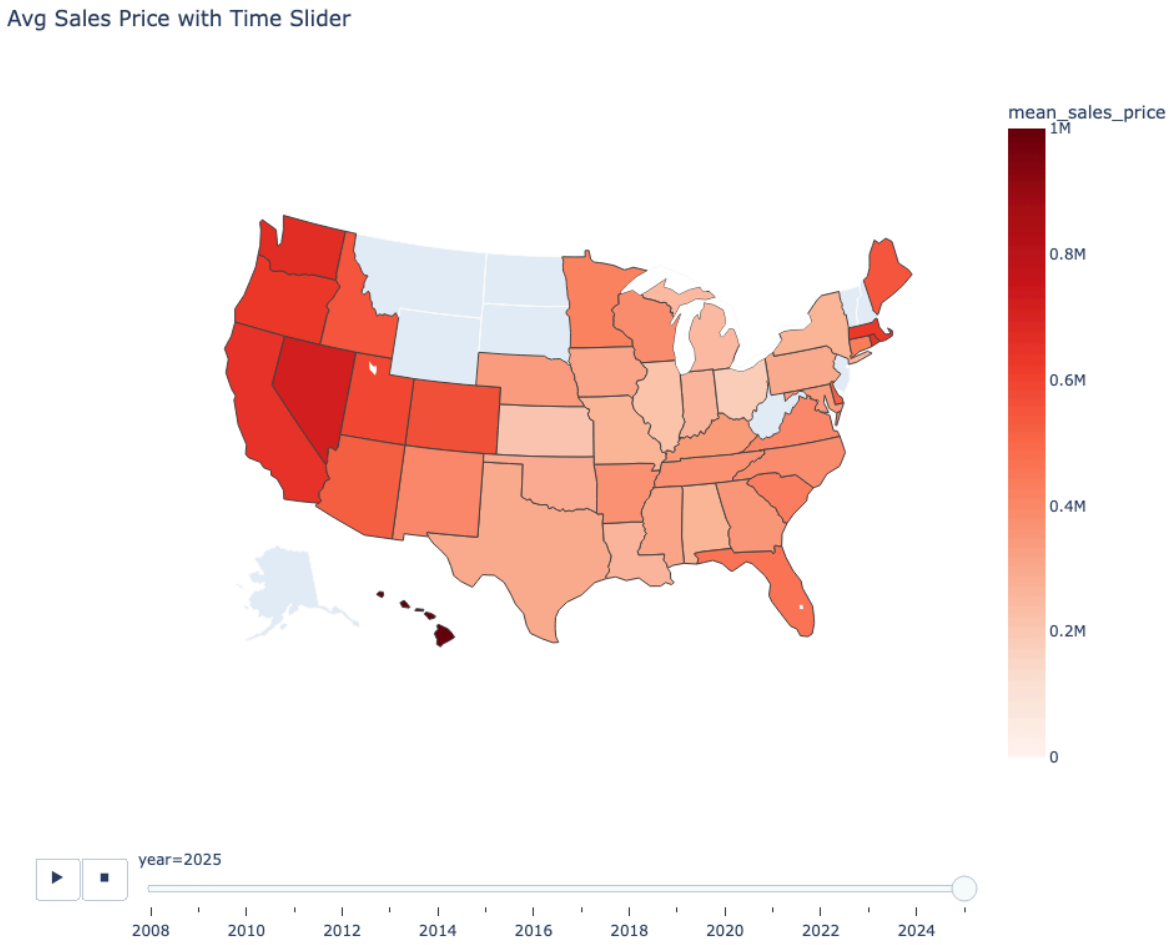
Massachusetts and a decrease in Delaware:



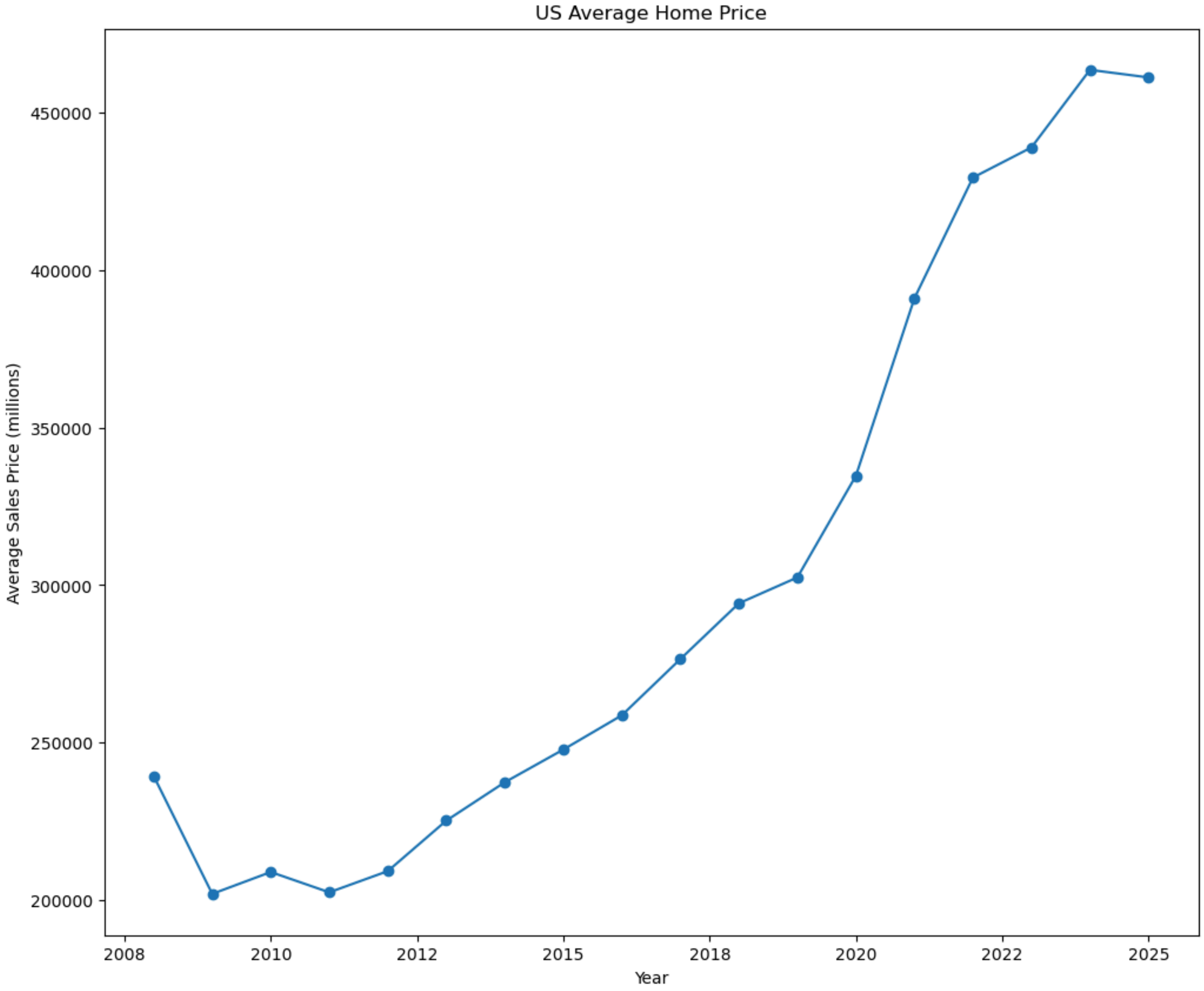
The next plot is 5 years later, in 2020. This is the year of the pandemic hitting. Almost all West Coast states go over $500,000. Other states also have more significant increases than we saw from 2008 to 2015. This is where we really start seeing a large increase in the home prices:



Finally, we move forward to our 2025 plot where the prices increase even more dramatically. All states across the West Coast are bright red, signifying higher prices. States along the East Coast also got significantly redder. Very few states are now under the $400,000 mark:



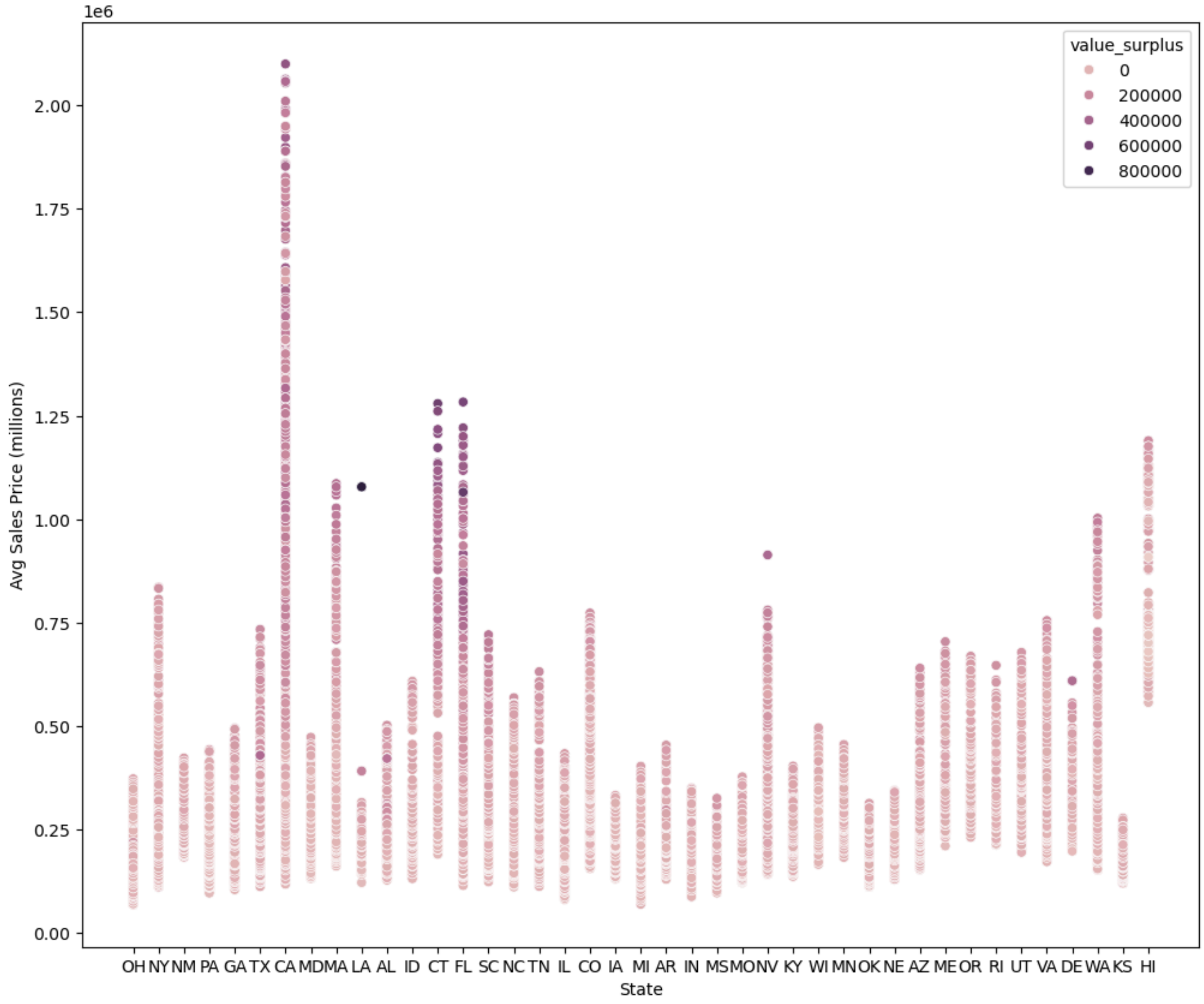
To view how the US average changed over time, I created a line chart to view the change year over year. After the economic issues in 2008, there is a dip in home prices for a few years. It returns to the 2008 prices around 2015 and has a moderate increase from 2015 to 2020. 2020 is when we a steep increase in home prices until about 2022 where it starts leveling off again:



Correlation between variables

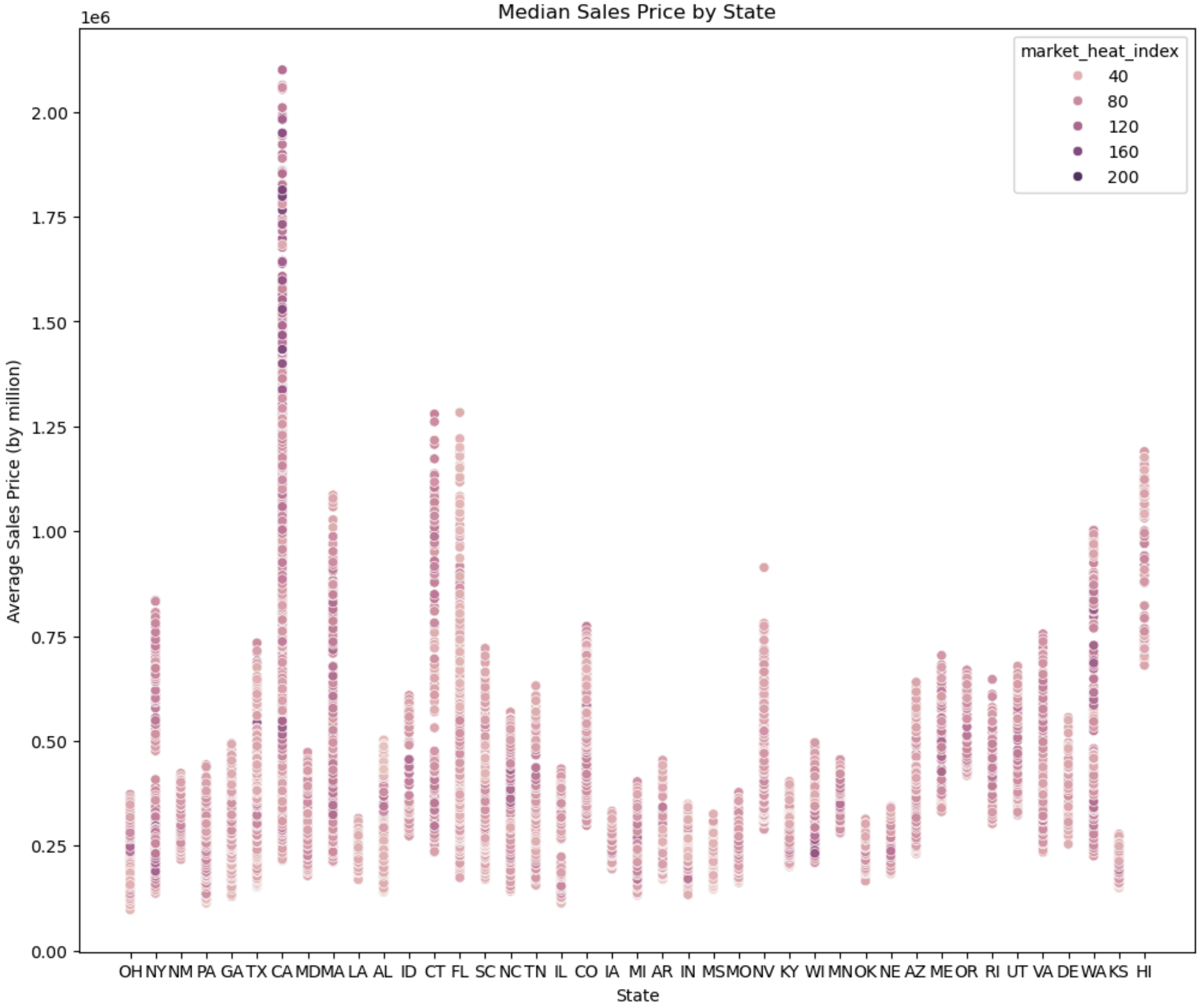
In EDA, one of the things I wanted to find was what correlates most with the sales price. This can tell us more about what influences sales prices in the US. The things that correlated most (outside of home value) were value surplus, market heat index, and population rank, which had the strongest negative correlation.

When visualizing value surplus, we can see that towns with homes that typically sell over value are among the higher prices:



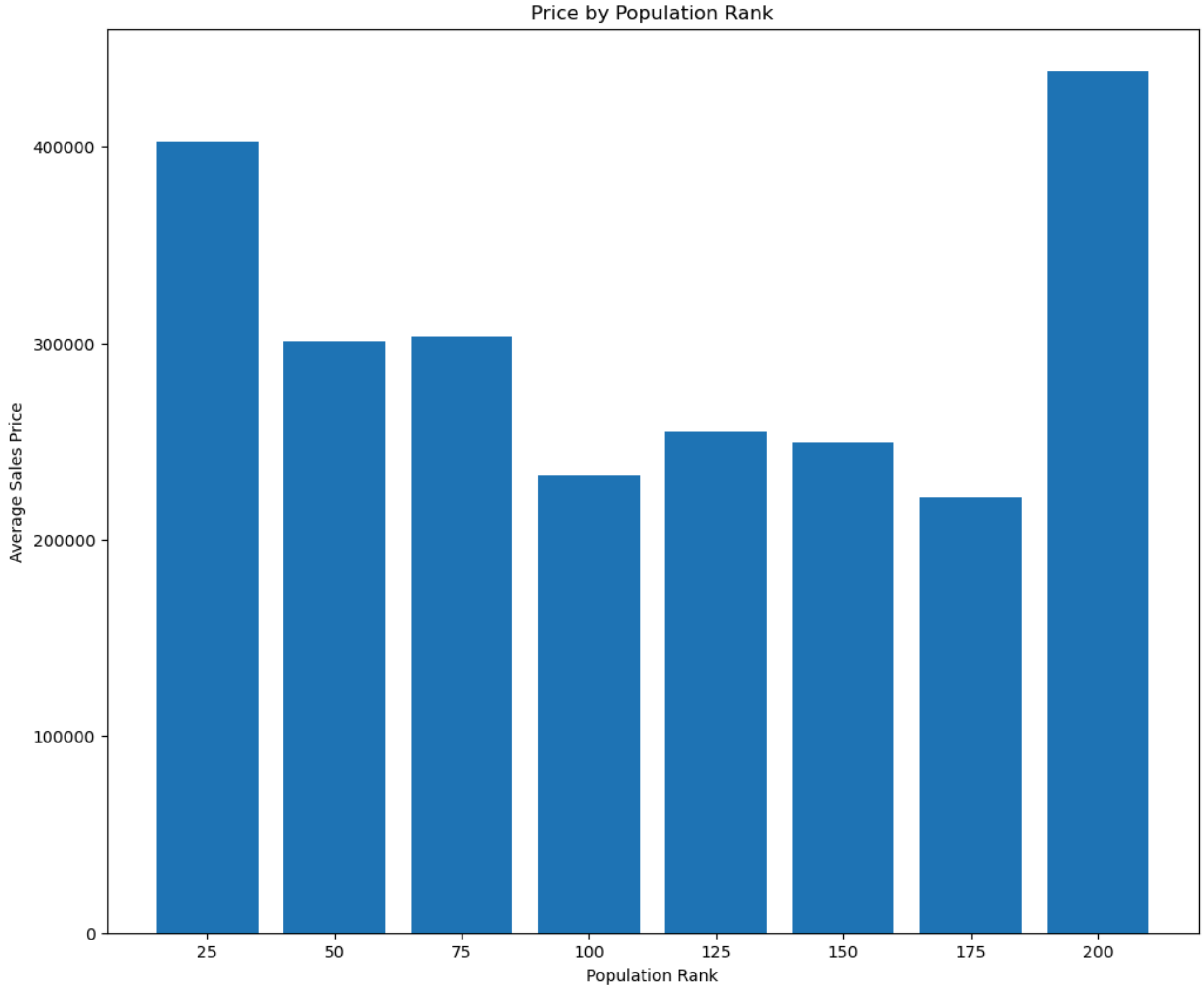
This holds true in every state but we can really see it in Connecticut, Florida, and California.

Market heat index indicates whether it is a buyer’s or seller’s market. Higher numbers indicate it is more of a seller’s market. We can see that a seller’s market typically leads to higher sales prices but the results are more scattered than with value surplus:



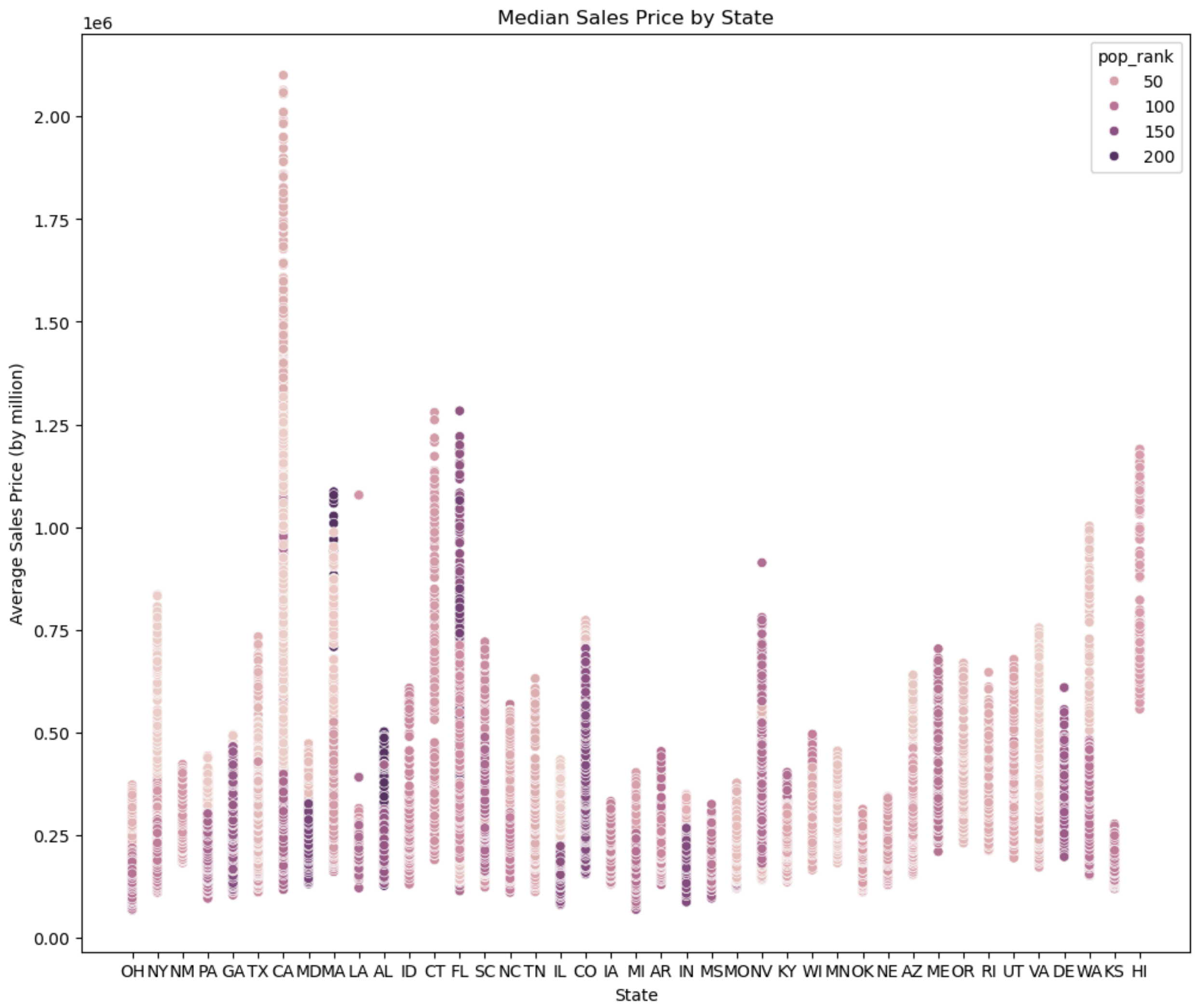
From these two graphs, we can see that value surplus has the stronger positive correlation with sales price.

It was interesting the population rank had the strongest negative correlation with sales price so I was really interested to view the relationship between these variables. You would think that bigger cities usually lead to higher sales price while more remote cities were less expensive. However, we can see here that while bigger cities are higher in price, the smallest cities also have a high sales price:



The home price steadily decreases as we go from the top 25 biggest cities to the top 175 cities. Once we get over 175, the price increases drastically. This could be the result of a few rich smaller cities.

As we did with the other variables, I wanted to view this on a state-by-state basis. We can see that in states like California, New York, and Washington (among others), the more populated cities do lead to higher sales prices. However, in Florida and Massachusetts, the less populated cities lead to higher sales prices. This is probably a result of a few smaller, rich cities in these states may be causing this negative correlation.



Model Selection

I tested a few different regression models on the dataset to determine the one that would be the best fit for predicting the data. The models included linear, ridge and lasso regression, XGBoost regression, and ARIMA.

Before testing and building the models, I engineer the features in preparation for model training. First, grouped the data to train on the total US data by taking the mean for each month from 2008-2025. I also performed scaling on the numerical features and encoding on the categorical variables. I chose to use Robust scaling as the data was pretty strongly skewed during the EDA portion. For the categorical variables (city/state), I used hashing encoder to split the cities into 6 groups and states into 2 groups.

After splitting into the train/test sets, the data was ready to be trained. First I trained on a linear regression model. This strongly overfit our data, returning a 99% accuracy (R-squared score) and a perfect regression line. To answer for this, I moved to Ridge and Lasso regression to regularize the data.

When performing Ridge and Lasso regression, I tested different alphas to see how they would perform. Ridge regression resulted in a R-squared coefficient of 0.888 and a RMSE of about 29000. My only concern was that it fell a little short on the most recent values. Lasso regression performed slightly worse than Ridge regression in terms of RMSE.

To see if I could improve the RMSE, I tried XGBoost regression. This model performed well but after trying different parameters to tune, there wasn’t a good mix for the model that didn’t overfit the data and had a strong RMSE. Most parameters resulted in the regression being capped on each end so it would miss the upper/lower levels of the data. In total, it would overfit the middle values and strongly vary the data closer to the min/max.

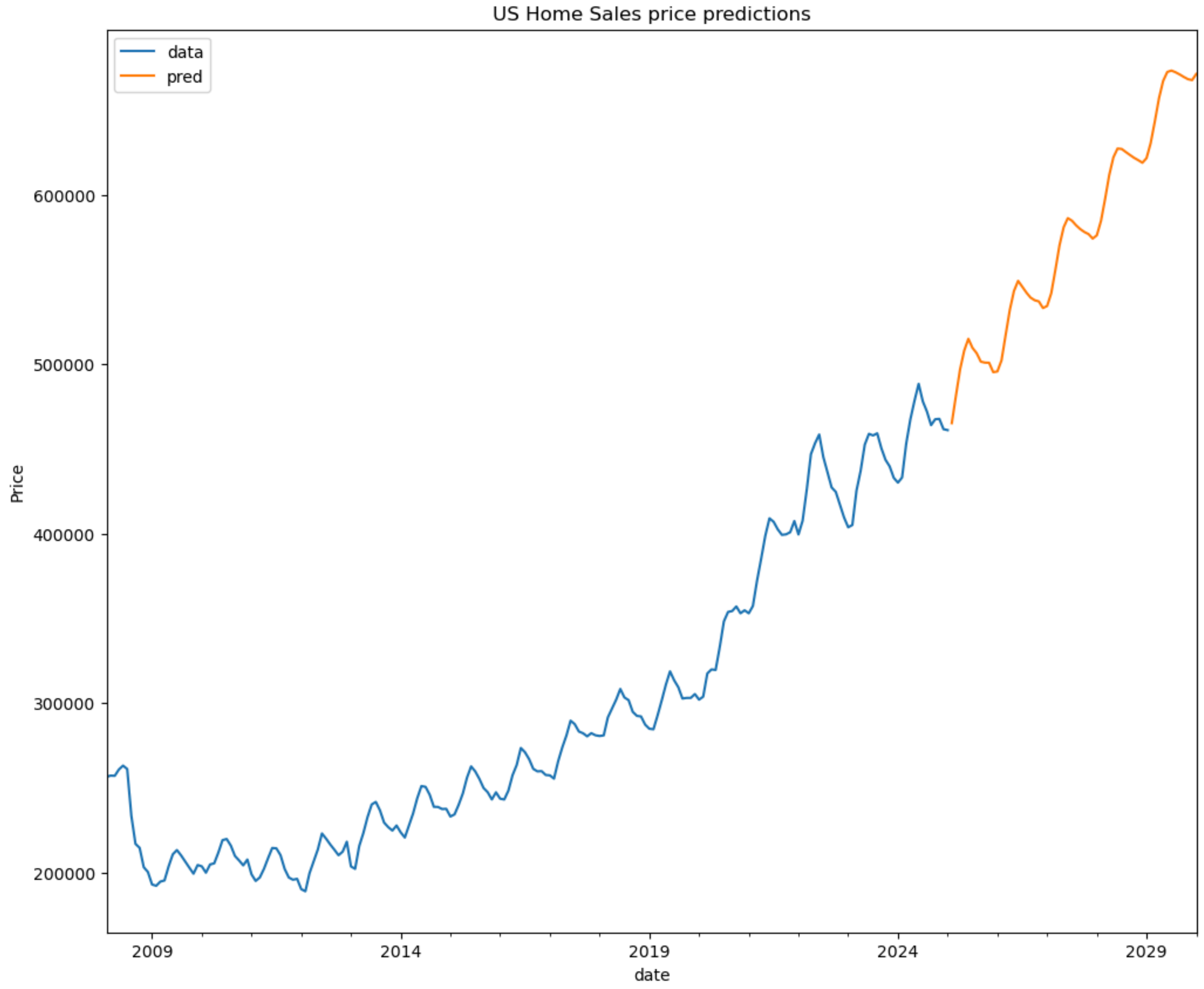
I also tested on the ARIMA model as this is usually a strong model for time-series analysis. After testing different parameters, this model had really strong AIC and BIC values but strongly overfit the data as well.

Comparing these models, the best model to provide generalization and minimize RMSE was Ridge Regression.

Prediction

US Prediction

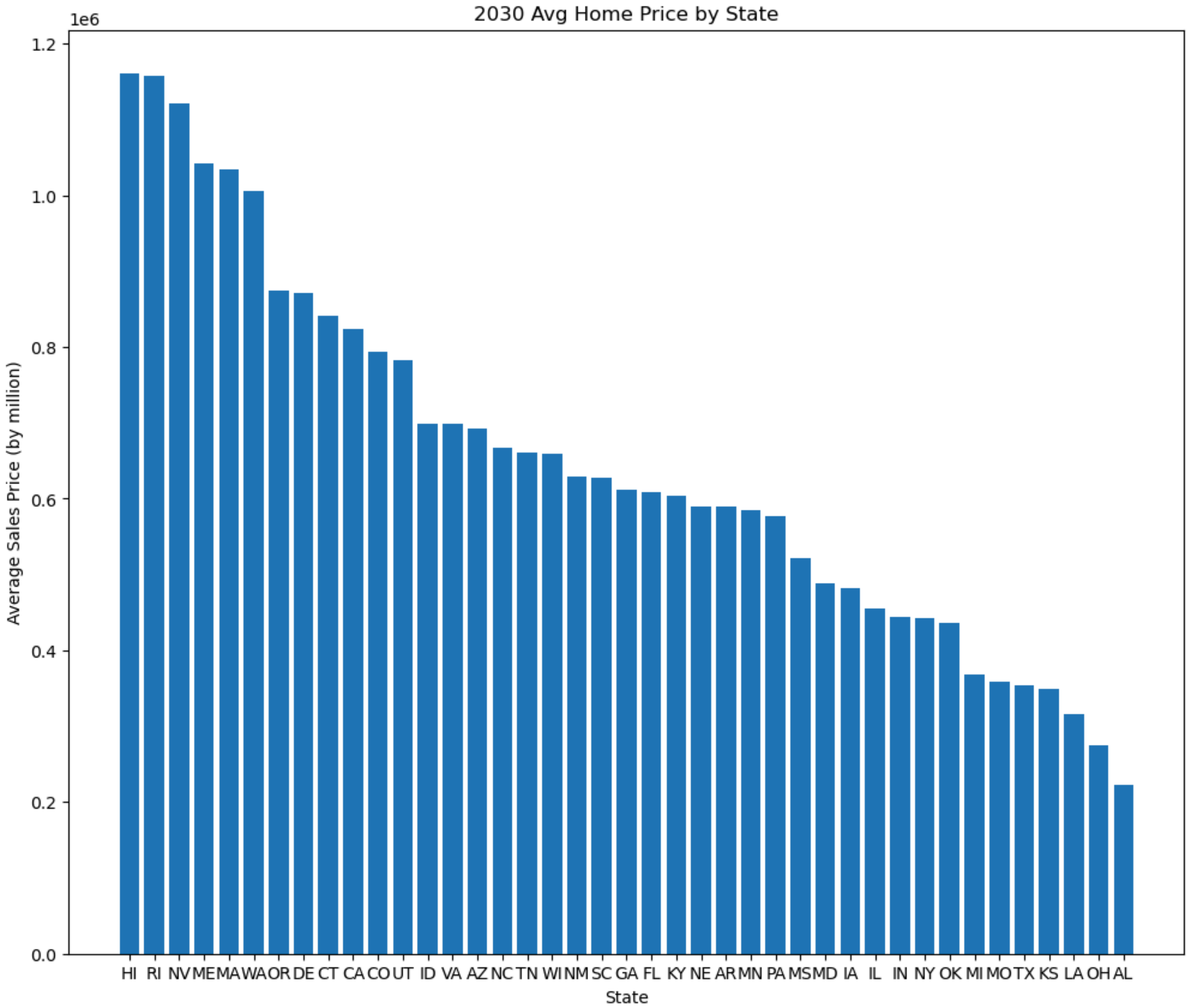
Using the Ridge regression model, I generated predictions for the next 5 years. To start, I generated predictions for the entire country:



In 2030, the average home price in the US will be $671,450. This is an increase of $210,342 over today's average of $461,108.

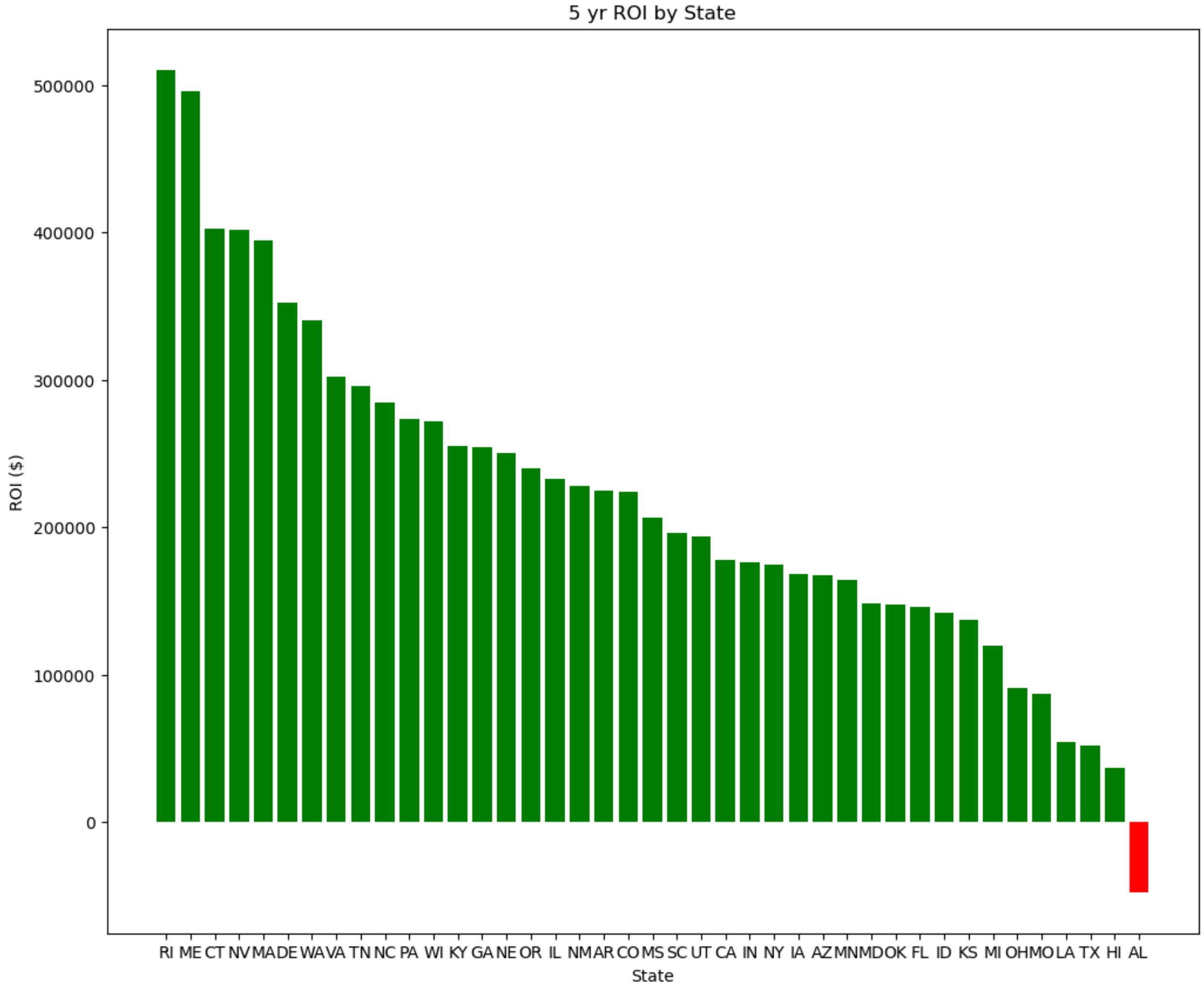
State Predictions

Next, I generated predictions by states to see where home prices would be in each state in 2030:



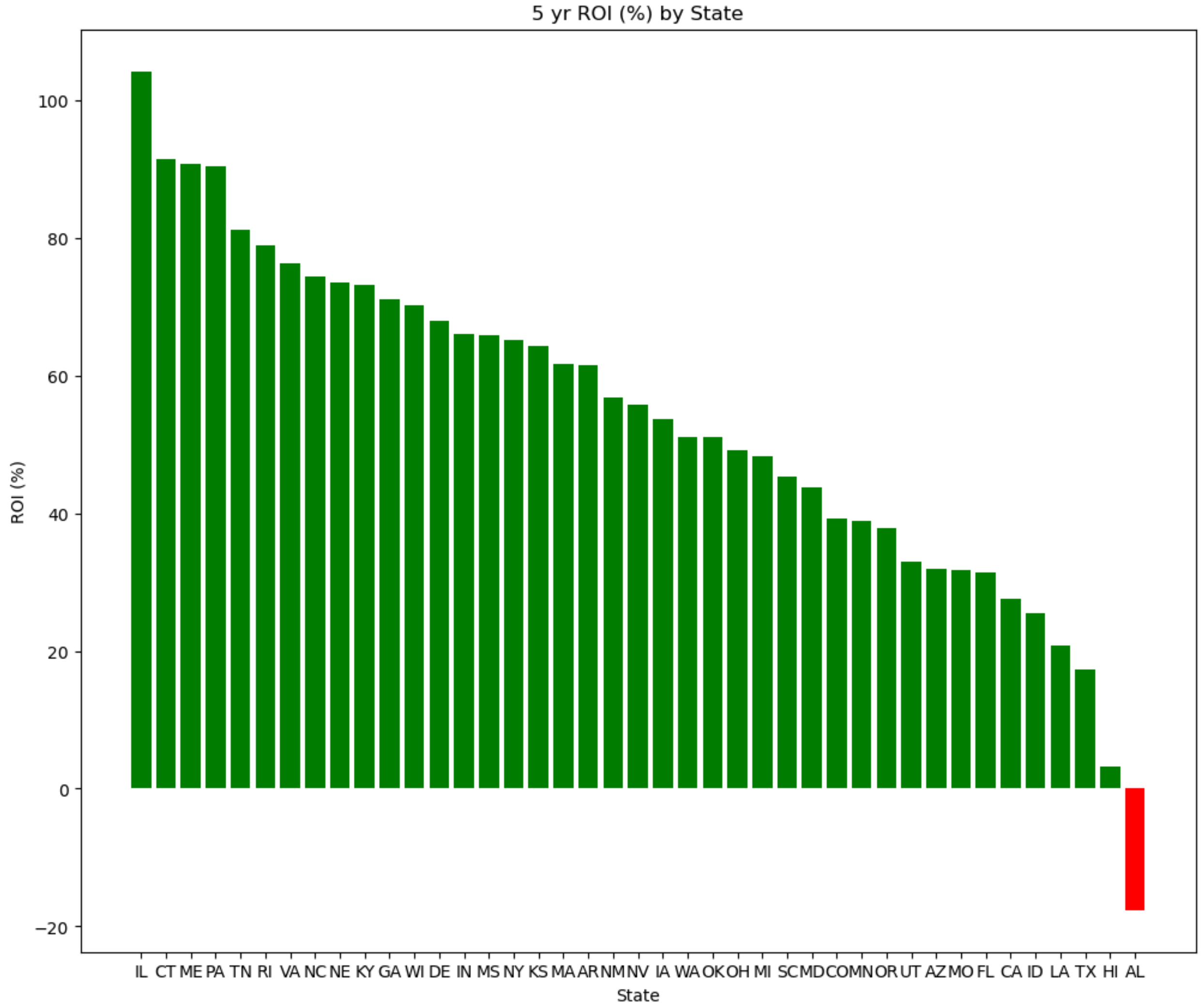
Hawaii has the highest average home price in 2030 at $1,160,246, with Rhode Island and Nevada slightly behind at $1,157,088 and $1,121,203, respectively.

While these states have the highest prices, that doesn't mean that they will have the highest ROI. I examined the difference between the prices in 2030 and 2025 to see which states could have the highest ROI:



Here, Rhode Island leads in the total ROI over 5 years with a total ROI of $510,557. Maine was not far behind with an increase of $496,308. All states saw a positive ROI outside of Alabama, which is negative. Even though Hawaii had the highest price, the total ROI was not very high.

Next, I wanted to see the highest ROI percentage, as this could show opportunities for a strong percentage gain from your initial investment. Here we can see the data tells a slightly different story:

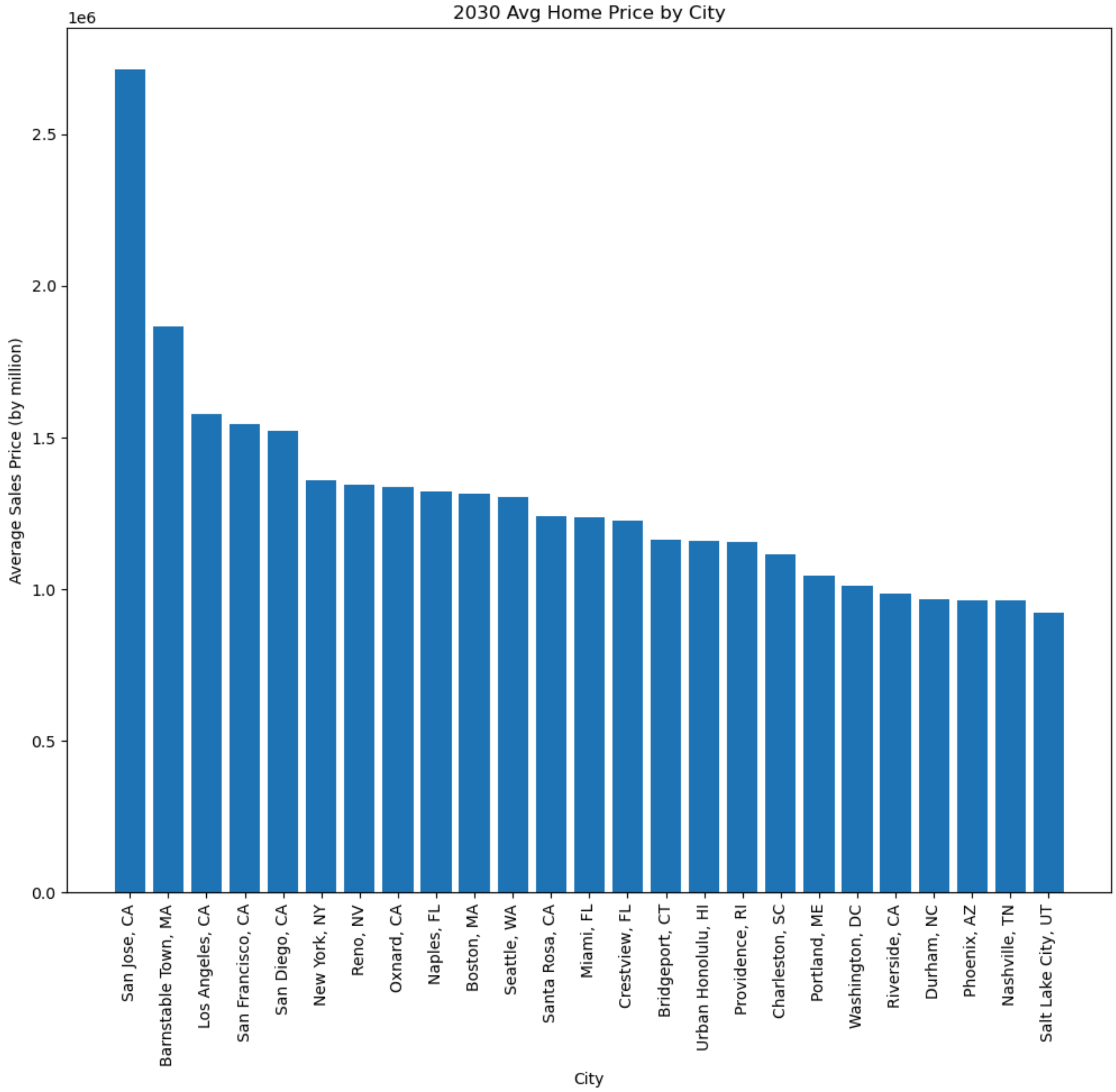


While Illinois was more toward the middle on the predicted prices and overall ROI, it has the highest percentage increase with over 100% ROI. All other states are below 100% ROI but would still be profitable (other than Alabama) over the next 5 years.

In summary, Rhode Island and Maine are the greatest opportunities for total dollar amount gain while Illinois is the best opportunity for percentage gain. for our states,

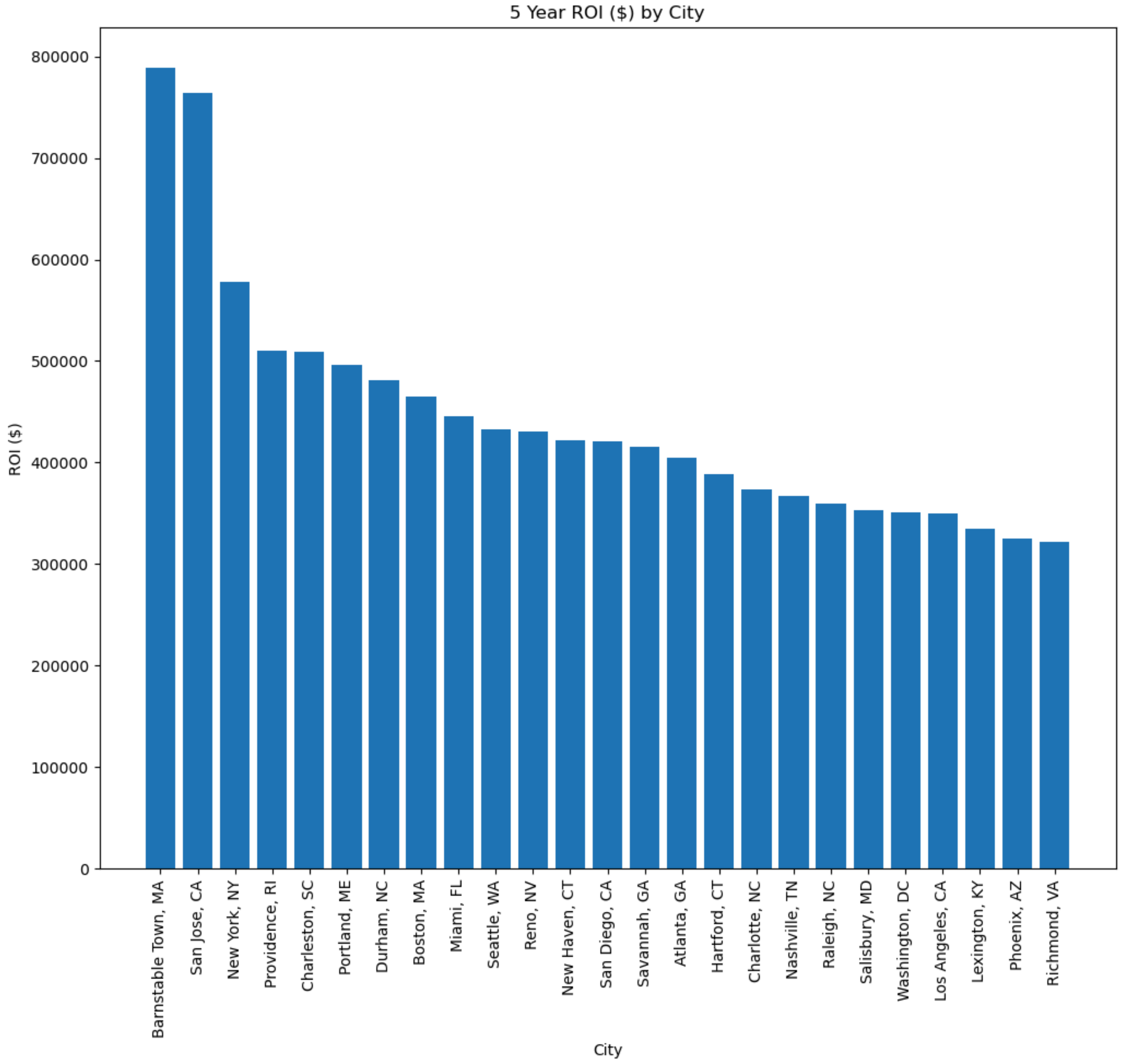
City Predictions

Finally, I used the Ridge Regression model to predict housing prices in cities in the US. As there are over 140 cities in the final dataset, I limited these results to the top 25:



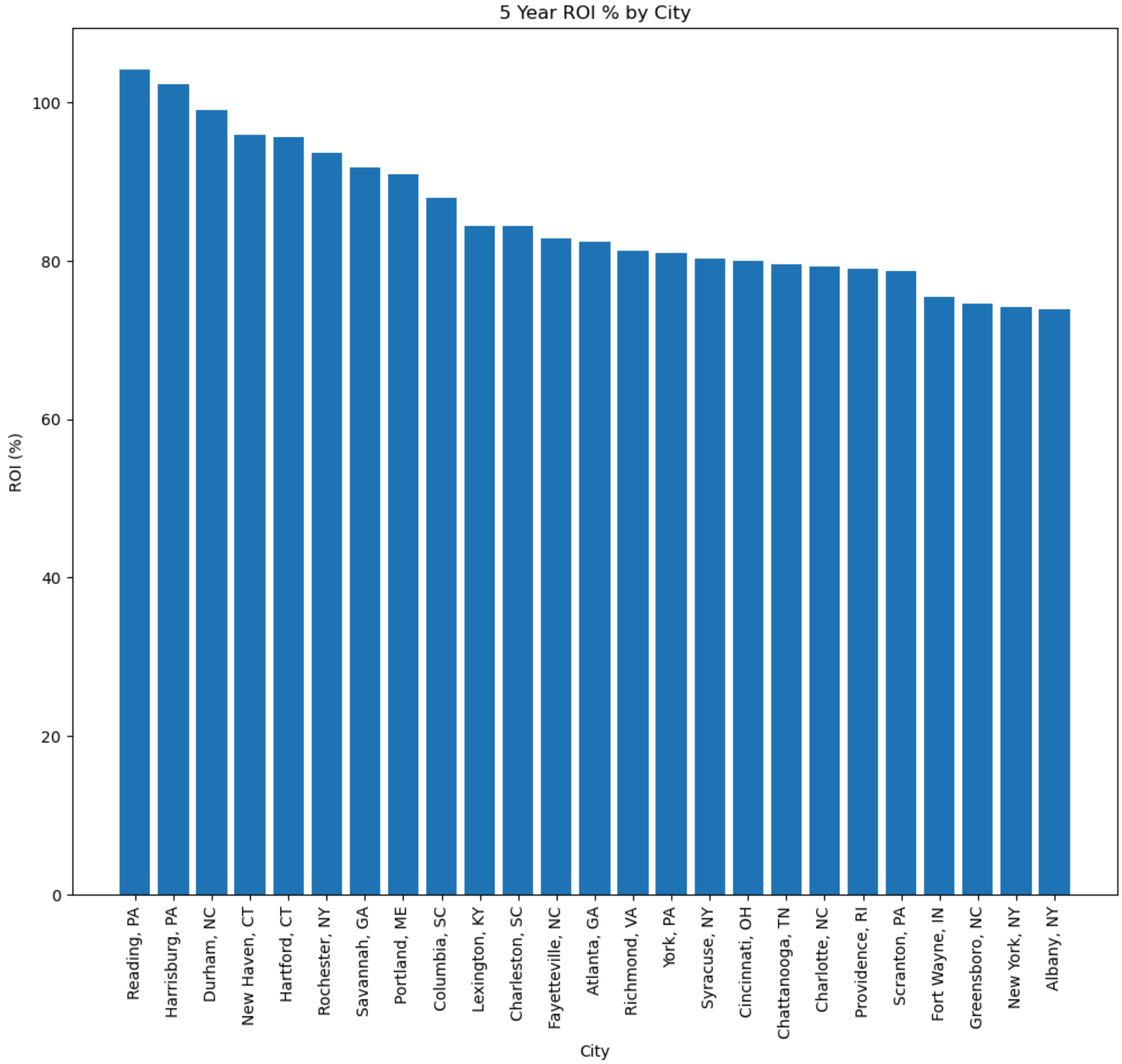
San Jose, CA will have by far the highest home price in 2030 at $2,713,933. All other cities are below $2M on average.

As I did for the states, I broke down the cities based on ROI to see where the top cities would be by total dollar amount and percentage. First, the total ROI:



Barnstable, MA and San Jose, CA will have the highest overall growth in price by a wide margin over other cities. Barnstable will grow by $788,990 and San Jose will increase by $764,034 by 2030.

Finally, I examined the ROI percentage across the cities in our dataset:



Here we see that both Reading, PA and Harrisburg, PA will have over 100% ROI. Reading will go from $298,995 in 2025 to $610408 in 2030. Harrisburg has a very similar increase, going from $303,263 in 2025 to $613,155 in 2030.

In conclusion, Reading and Harrisburg will be the best cities for percentage gains in ROI, and Barnstable and San Jose will be the best for total dollar amount gains in ROI.

Suggestions for the Future

During the project, I thought of a few data points to add that coul would have been helpful for creating a better model and getting more accurate prediction. Economic data such as average income, GDP, and cost of living could help show how that impacts home prices. Demographic trends would also be helpful around age, education, race, etc. Climate data could also be interesting to incorporate to see if weather has any impact on housing prices.

It also would have been helpful to access the Zillow API to streamline data and incorporate more housing data. Characteristics such as bedrooms, bathrooms, home square footage, overall property area, and neighborhood characteristics could help get a more accurate model. Unfortunately my request to access this API was denied for this project

Seeing historical data further back could be interesting for the model as well. The data ranged from Feb 2008-Jan 2025 so I had around 17 years of data. A few key economic events occurred during this time that can skew the data. First the 2008 recession caused housing prices to steeply decrease for the first few years of data. Then COVID hit in 2020 and we’ve seen a huge increase in housing prices over the last 5 years. Getting further historical data could help smooth out these steep trends