Clarence Lin, David Moran

Github: https://github.com/1Clarence3/AppliedML-Final-Project/tree/main

State the problem:
Which socioeconomic & housing-market factors most strongly influence annual county-level housing price growth, and how do those drivers vary across Census regions.

Audience:
Intended for young adults (20 - 30) looking to buy houses in the near future. Our potential value for this project would be to hopefully help these people reduce time researching about what factors would most influence housing prices as well as which factors are important for which regions. We would also then recommend counties that are undervalued vs. counties that are overvalued, which provides the audience a sense of where we should buy our next house.

Data Introduction & Processing:
We approached our problem by assuming housing prices will be driven by supply and demand. On the demand side, we looked at common socioeconomic factors we thought would influence housing prices. To do so, we looked at datasets from the United States Census Bureau (https://data.census.gov). We obtained 4 datasets corresponding to population, income, employment, and education features for US counties from the above link. In terms of the time range, we obtained a decade of data from 2012 to 2022 (excluding 2020 due to COVID-19) and the number of counties was ~450. Below is a bulleted list of features we originally had:

- Population features (raw population counts): total, male, female, multiple ages (20-24, 25-34, 35-44, 45-54, 55-59, 60-64), median age, multiple races (hispanics, whites, african americans, asians, american indians), total housing units
- Income features: median household income, total number of households, median incomes for each race, median income for 4 age groups (15-24, 25-44, 45-64, 65+), median income of family and nonfamily households
- Education features (% of county population): % with bachelor's degree or higher for age groups (18-24, 25-34, 35-44, 45-64, 65+)
- Employment features: labor force total population, % unemployed

On the other hand, for the supply side, we found a dataset relating to permits (https://hudgis-hud.opendata.arcgis.com/datasets/HUD%3A%3Aresidential-construction-permits-by-county/about). Specifically, we used single-family permits, which refer to the number of building permits issued for new single-family homes (as opposed to multi-family residences, like apartments). Lastly, we obtained the median housing price dataset by county on the Census Bureau website. An alternative to using the housing prices from the census was to use monthly average housing prices from Zillow. Although this dataset would be more granular (monthly price vs. annual price), we wanted temporal consistency with the other features, which were all annual averages. Also, the values fluctuate too much and were inconsistent in terms of having missing months or counties for certain years.

With this, we moved forward to download these datasets, which had counties as rows and the features as columns. We merged all the features together via an inner join on GEO_ID, which is a unique identifier

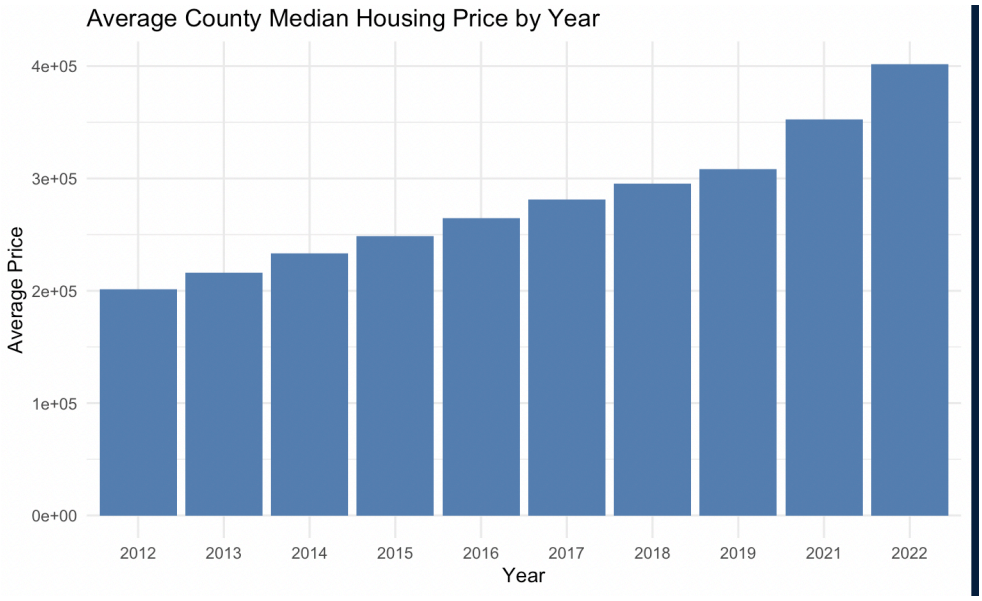
for each US county. Lastly, we separated the datasets into 1 csv file per year (2012.csv, 2013.csv, …, 2022.csv).

Feature Engineering:
Given that the nature of housing prices and real estate is very time sensitive, we wanted to engineer some dynamic features that weren't static raw population counts or percentages. We decided on engineering the annual median housing price growth as a percentage, which would become the main target variable for all models later on. One reason we chose housing price growth instead of raw housing price itself is that percent growth shows how fast prices are rising, which is useful for our audience in estimating future affordability. Also, this allows users to make meaningful comparisons across regions with different price baselines (comparing median prices across 2 counties could be misleading if one county has a higher average salary, lower tax rate, etc.). We can then recommend our young buyers areas where prices could accelerate rapidly vs. areas with slower growth.

Below are a few other features we engineered:

- Population growth %, Income growth %, housing growth %, labor force growth %: All growth rate features were computed with the growth from the previous year to the current year
- income_per_permit: The ratio of median income to permit count could reflect the balance between income (demand) and housing supply (permits issued)
- pop_per_housing: This ratio of total population to total housing units indicates the pressure on housing supply. A higher ratio means more people per available unit, which increases competition (higher demand)

With a full set of different types of features ranging from raw counts, percentages, growth rates, and ratios, let's get a sense of the current housing price trend. We did EDA by looking at the average county-level housing price for each year just to validate our problem at the beginning. We also plotted the average annual growth rate (% change in housing price relative to the prior year) of county housing prices over time to better visualize the jump in prices:

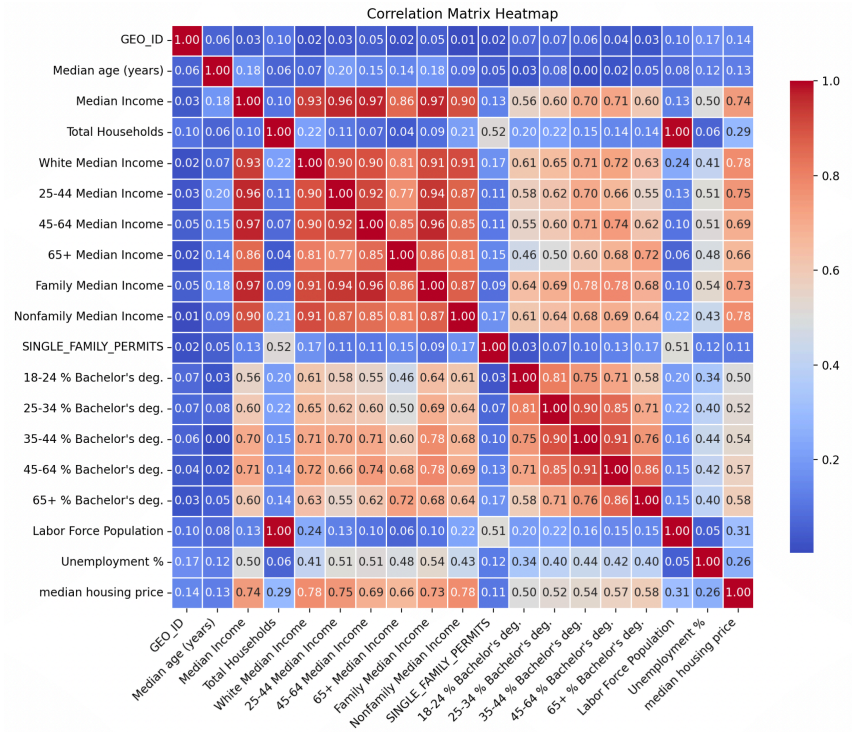


year <fctr>	avg_price_growth <dbl>
2012	*NA*
2013	2.090652
2014	4.051098
2015	6.010329
2016	4.971150
2017	5.312599
2018	5.752227
2019	4.317050
2021	15.974445
2022	13.844155

We chose bar graphs instead of line plots because you can't assume a perfect linear relationship in price between one year and the next. We only have discrete points per year. As expected, housing prices are rising fast and after covid, there was a sharp jump in price percentage change. For the second bar graph, 2012 is NA because we have no data prior to 2012.

Preparing Features for Models:

Given we have so many features, many of them just from eyeballing are likely correlated. So the next step was to compute a correlation matrix to see which pairs of features were very highly correlated. Note that the image below isn't the full matrix.



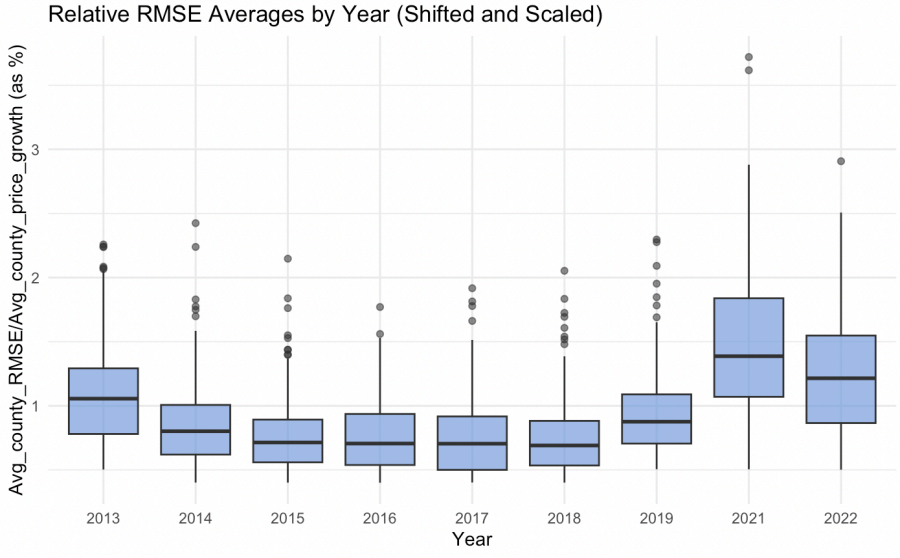
Some features were dropped. For instance, male and female population were dropped (kept total population), many of the race features for income were dropped (kept median income), and the middle age categories were dropped (kept young age and old age). Most of the growth features as well as ratios were kept, as they weren't correlated with many other features. A possible reason for why the code retained a young age category (i.e. 18-24) and an old one (i.e. 65+) could be because this is indicative of characteristics between younger working class and older generations within each county, which probably have a larger difference in values across the features (hence the lower correlation).

We decided to also normalize the features prior to doing any model training to standardize the results and values across our dataframes. We applied a Z-score normalization for each year across the individual columns, or features.

Models:
To begin our first model, we tested using a baseline of linear regression. To simulate this, we used basic assumptions: 1 demand feature (median income) and 1 supply feature (single family permits) to predict annual housing price growth. The reason for choosing population as the demand feature was that we thought higher median income would increase purchasing power, enabling more households to afford homes. So then we have increased demand for housing which drives housing prices up. On the supply side, permits were the only feature we had and we thought using the count would signal housing supply increase or decrease.

Moving onto cross validation, we made sure that if we are predicting price growth for a county in year x, then x is only in the test set, and we train on the other years. In other words, at each fold, the same year can't be in both sets at one time. The reason we allowed the model to train on all other years except x instead of all years less than x was that we assumed housing price growth is often driven by structural factors (e.g., income levels, housing supply) that vary more by location compared to small year-to-year changes (we will see this in map later on). This suggests limited temporal correlation. Another reason is that our objective isn't to deploy a real time model for future predictions, but rather to explore feature importance and regional variation. So we thought pooling all the years helps improve statistical power.

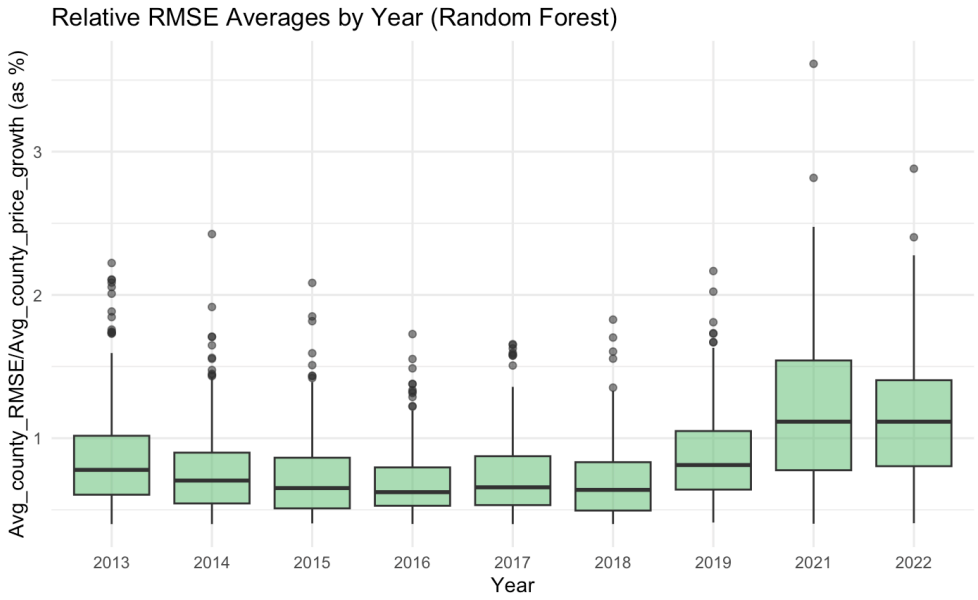
For reporting performance, the county-level average RMSE was used for each year. However, this value itself does not hold much meaning unless we compare it to the actual price growth for the county. Thus, we divided the county-level average RMSE by its price growth to obtain essentially relative RMSE ratio:



test_year <dbl>	rel_rmse <dbl>
2013	1.0665021
2014	0.8451080
2015	0.7666225
2016	0.7662723
2017	0.7446014
2018	0.7526866
2019	0.9318709
2021	1.4678833
2022	1.2462800

Since RMSE represents the error on predictions of price growth, we interpret this relative RMSE as follows: For a given year, if the ratio of RMSE to the average county price growth = 1, the model is as good as guessing the mean. If the ratio is < 1, then it did relatively well for that year and on the flip side, relatively poor if the ratio > 1. The units, thus, can be thought of like a percent error. From the graph and values, we were able to predict the growth somewhat well between 2014 - 2018 while the tail-end years had a worse performance. A possible explanation is that for the post-COVID-19 years, the data was more volatile as we saw earlier there were large jumps in prices. Also, the housing market shifts could then be a result of a lot more factors (record-low interest rates, migration patterns, supply-demand imbalances, etc.) that were harder to capture with traditional predictors like income or permits, leading to higher prediction errors. On the other hand, the middle years were more stable. Overall, it seems median income and permit counts are somewhat respectable features.

Moving on from the baseline, we used a random forest model to again predict the annual housing price growth rate (target variable). We chose this model because based on the linear model above as well as the complex nature of real estate, we thought there could be many complex and nonlinear relationships which random forests can handle. Again, we trained using the same cross validation technique (dropping current year per fold) and used RMSE as our validation metric. Here were the results:



test_year <dbl>	rel_rmse <dbl>
2013	0.8582983
2014	0.7647664
2015	0.7210225
2016	0.6911789
2017	0.7338787
2018	0.6927318
2019	0.8723713
2021	1.1739409
2022	1.1386458

As we can see, the results were slightly better compared to the linear regression model in terms of the median relative RSMEs, but not by much. Again, a similar pattern emerges where the latter years had a lower performance compared to the middle years from 2014-2018. Given the similar performances between models, this helps answer our problem statement in that most of our selected features earlier do not influence price growth much.

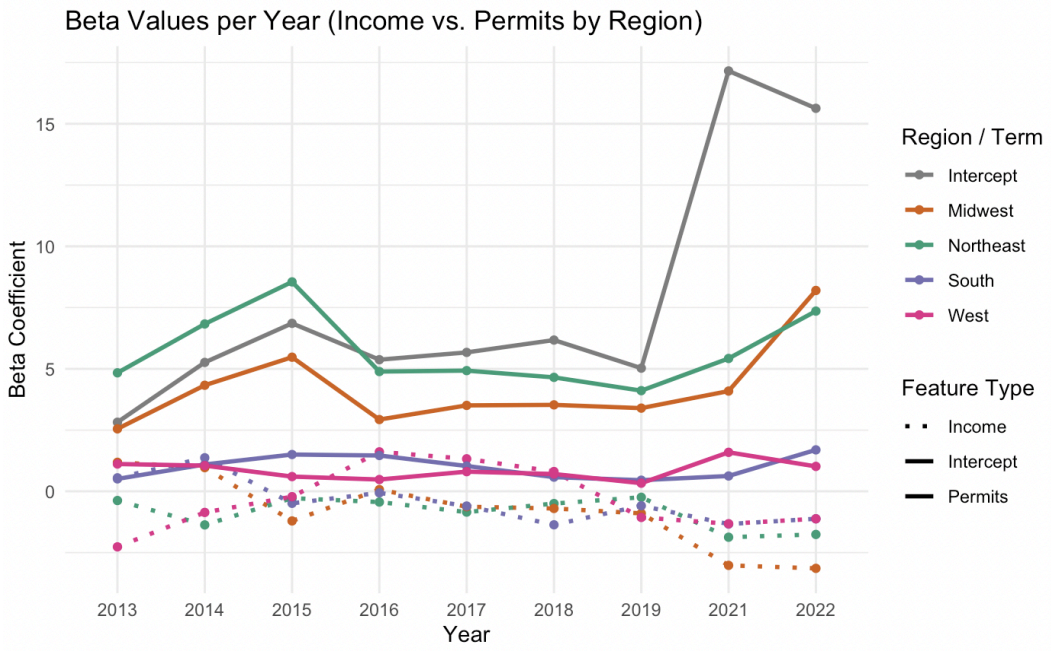
Feature Importance by Region:
Now extending off of these results, it would be helpful to our audience by addressing our question at the very beginning of how top predictors vary across U.S. regions. To do this, let's go back to our linear

regression approach given its comparable results to a more computationally expensive Random Forest regression model. One way to see how our 2 features vary in importance across different regions is to interact our feature columns with each of 4 region columns. Then, we would learn the beta or slope values for each of these interactions per year, giving us the relative importance of each feature for each region.

To give a more detailed explanation, we were interested in dividing the counties into Midwest, West, Northeast, and Southern regions. Thus, we made four additional columns in our dataframe (is_midwest, is_west, is_northeast, is_south) representing if a county was part of that region (1 for yes, 0 for no). This essentially made four one-hot encoding vectors. In terms of how we divided the counties, each county has a unique GEO_ID, and the last 5 digits of this number is known as the FIPS code, which contains state and county info. We utilized the U.S. Census' mapping of FIPS codes to the four regions to separate the counties. Form there, we interacted our median income and permits columns with each of these region columns, which makes our linear formulas as follows:

$$Y \text{ (annual price growth \%)} = \text{income}_t * \text{is\_west}_t * \beta_{\text{income x is\_west}} + \text{(income terms for other 3 regions)} + \text{permits}_t * \text{is\_west}_t * \beta_{\text{permits x is\_west}} + \text{(permit terms for the other 3 regions)}$$

So in total we are learning 8 beta values for 8 terms (4 for income, 4 for permits). The subscript "t" is just the current year and Y is our target variable. Running our linear regression model for each year resulted in the following beta values over time color coded by region (dotted lines are permit features, solid lines is the income feature):

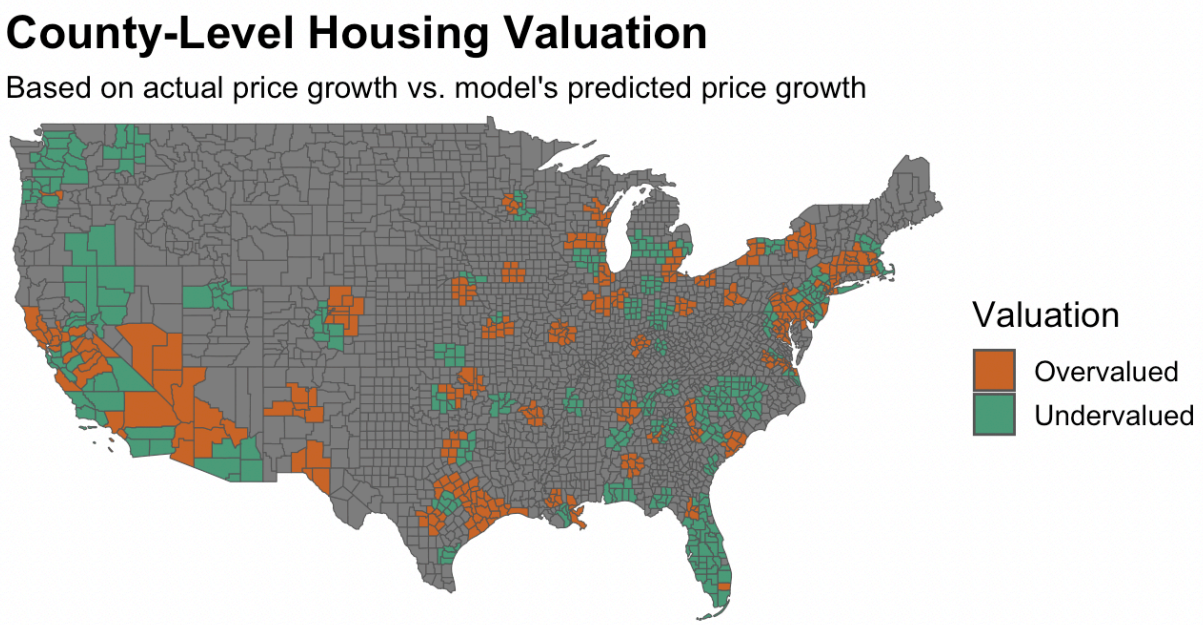


Based on the plot above, let's first understand what the intercept value represents. From a math standpoint, you can think of it as the estimated price growth rate for a hypothetical county with no income or permit info, before considering any region-specific dynamics. This corroborates the average actual county price growth rates we plotted at the beginning. Or in other words, it acts as a normalization anchor or base growth level, against which the region-income and region-permit effects are layered. Now for the

beta values themselves, they should be interpreted relative to each other for a given year *t* in a given region *x*. If we take two such beta, then the ratio $|\beta_{income}/\beta_{permit}| = r$ means median income has $|r|$ times the marginal impact or percentage point increase on predicted price growth compared to permit counts for year *t,* region *x.*

The general trend from the graph is that income is positively correlated with our target variable while permits is negatively correlated (terms above y=0 axis vs. terms below). This intuitively makes sense since higher median income means higher demand and price. For permits, having more of it increases supply and decreases price. Overall, it seems that median income has a stronger influence relative to permits on the price growth over all the years, and especially for the midwest and northeast regions. In fact, these two regions seemed to have the highest beta difference between income and permit, which suggest that one should pay more attention to these two variables if you're buying houses in these counties. In recent times, income seems to have become a slightly more indicative factor for housing price growth. Lastly, for permit features, the betas are relatively stable over time.

Housing Valuation for Counties:
Circling back to our problem statement again, we want to help our audience or prospective homebuyers better understand regional housing markets. So given we have obtained betas for our linear model above, we can feed our model each county's data to get predicted housing price growth in 2022, which we compared against actual observed growth. If the difference of actual growth minus predicted growth > 0, the actual growth is higher than predicted, which means prices grew more than model expectations and the house is worth more. This means the market appears undervalued. On the flip side, if deviation < 0, we said the housing market here was overvalued (price grew slower than expected). For the thresholds of placing this "overvalued" or "undervalued" label, We determined that deviations greater than 0.05 (i.e., 5 percentage points) were undervalued and deviations less than -0.05 were overvalued. Growths within this range were considered fairly valued. The 0.05 threshold was chosen based on the annual growth rates in our first graph being around 5%. This means that ±0.05 from this represents a substantial deviation. In the end, we created a county-level heat map that highlights counties as overvalued or undervalued, or fairly valued.



Much of the counties are greyed out due to either being “fairly valued” as mentioned before, or not being a part of the census data. This was due to many low-populated countries not having enough people to survey for different features, which forced me to exclude the county in our dataset. From the map above and code, We were able to find the top states with the highest percentage of counties being overvalued as well as being undervalued. With that said, we put a threshold that the state needed to have at least a total of 20+ counties. Otherwise, for instance, District of Columbia had 1 county in our data be labelled as overvalued, which gives it a 100% overvalued percentage. The results below were copied over from code results. The reason the tables are manually typed up with values is because the output in my code looked a bit messy at first, so I thought the below was cleaner:

Overvalued States	Overvalued Counties	Total Counties	Overvalued %
Maryland	13	24	54.167
New York	26	62	41.935
California	23	58	39.655

Undervalued States	Undervalued Counties	Total Counties	Undervalued %
Florida	46	67	68.657
New Jersey	13	21	61.905
Washington	20	39	51.282

For overvalued states, Maryland was definitely the surprising state since New York and California are sort of expected to have very expensive houses. As for external validation, the closest related news was a wall-street backed landlord in 2024 that bought 264 homes for $98 million in Nevada (https://horsford.house.gov/media/in-the-news/swapping-homes-like-stocks-wall-street-backed-firm-buys-264-valley-homes-in-a-day). Although Nevada did not show up as the top 3 states in the table above, it was ranked 5th in our results for most undervalued states. So buying in any of the above undervalued states or Nevada seems to show promise.

Alternative approach & Why it was bad:

Initially, we had thought about using K-means clustering instead of doing different forms of regression as our main method of tackling this problem. This would’ve meant clustering the counties by socioeconomic factors and trying to interpret the results, which could possibly indicate if certain counties in an area were experiencing stagnant prices or the possibility of great price increase in the near future. However, we thought many of the clusters would just be based on location, since counties in the same area likely have similar feature values. This wouldn’t help our target audience make a decision as to whether to buy a house there are not. On initial testing, this was one of the testing run results:

cluster	num_counties	majority_state
0	188	Texas

1	91	California
2	3	Arizona
3	24	Florida
4	8	California
5	37	Virginia
6	1	California
7	101	North Carolina

Indeed, a lot of the counties would be clustered based on location. This was likely primarily because from the census, a majority of the counties surveyed were near highly populated cities, with smaller populated states having fewer counties. Thus, the populous regions were overrepresented by populous regions and the clustering would be biased toward creating groups that reflect urban density and regional economic patterns, not necessarily categories like "affordable" vs "unaffordable".

Limitations:

In terms of immediate limitations from the data, having data from more counties, especially more rural places, would open the door for possibly doing clustering and identifying promising small patches of land that could have a steep incline in housing prices in the near future (so buy the dip!). Also, just incorporating the data for another decade (2000 - 2022) would help. Also incorporating features that are indicative of future growth, such as birth and death rate, could be interesting.

Conclusion:

Overall, we wanted to investigate what demand and supply factors most strongly influence annual county-level housing price growth. We also looked at how those factors vary across different regions. To do so, we obtained and engineered several population, income, education, employment, and permit features across ~450 counties from 2012 to 2022 (excluding 2020). After dropping strongly correlated features and normalizing, we trained models. A linear model using median income and permit counts as well as a random forest model with selected features were trained with cross validation to predict the price growth for each year. Based on RMSE box plots, the results were comparable. Thus, we used our linear model to investigate how the aforementioned two features would vary in importance over time based on four regions. Plotting the beta values for these terms revealed income being a stronger predictor relative to permits and these two features being inversely correlated with each other and with the target. Lastly, examining a heatmap of overvalued and undervalued counties based on actual vs. predicted price growth encourages our audience to seek houses in Florida, New Jersey, Washington, or Nevada.