House Price Prediction Project Final Report

Stefan Jafry
stefanj@my.yorku.ca

Executive Summary:

This project aimed to develop a machine learning model to forecast U.S. housing prices using the Zillow Home Value Index (ZHVI) dataset. After a systematic data cleaning process including the removal of incomplete records and transformation from wide to long format we engineered time-based features (Year, Month, Quarter) and lag variables to support autoregressive modeling. Exploratory Data Analysis (EDA) revealed clear regional clusters and macroeconomic effects, such as the 2008 housing crash and the COVID-era price boom.

Three models were evaluated: **LightGBM**, **XGBoost**, and **SARIMAX**. LightGBM achieved the best overall performance ($R^2 = 0.9339$, MAE < $10K), outperforming XGBoost slightly in terms of stability and generalization. SARIMAX, applied only to top regions, performed poorly due to limited features and failed to generalize. K-Fold cross-validation confirmed high model stability ($R^2 > 0.999$), though slight overfitting was observed.

To address heteroskedasticity and non-normal residuals in high-priced regions, models were segmented by price tier (Low, Mid, High). This tiered approach significantly improved residual balance and reduced prediction errors. A forecasting loop was implemented to generate six-month-ahead predictions for each region, with LightGBM and XGBoost yielding near-identical results in Low and Mid tiers. However, High-tier markets exhibited large prediction divergences due to price volatility and limited feature granularity.

Overall, the models show high reliability for low and mid-tier housing segments. Future improvements could include incorporating property-level features (e.g., bedrooms, square footage) to enhance high-tier predictions.

Data Cleaning:

To ensure the Zillow Home Value Index (ZHVI) dataset was suitable for a forecasting model, we undertook a systematic data cleaning and reshaping process. This involved addressing missing values, normalizing the data set's structure, and engineering temporal features to support time-series modeling.

Step 1: Basic Cleaning

We began by removing rows that lacked RegionName or StateName. These fields are essential for regional analysis, and their absence would hinder interpretation and grouping. Rows missing this metadata were dropped from the dataset. Furthermore the dataset included monthly home price valued across multiple years (2000-2025), organized in wide format with one column per month. These monthly prices often contained missing values. To handle this we applied both a forward-fill and backward-fill methods to impute missing prices based on available neighboring months. This ensures complete data for each month with region name and state name.

Step 2: Reshaping

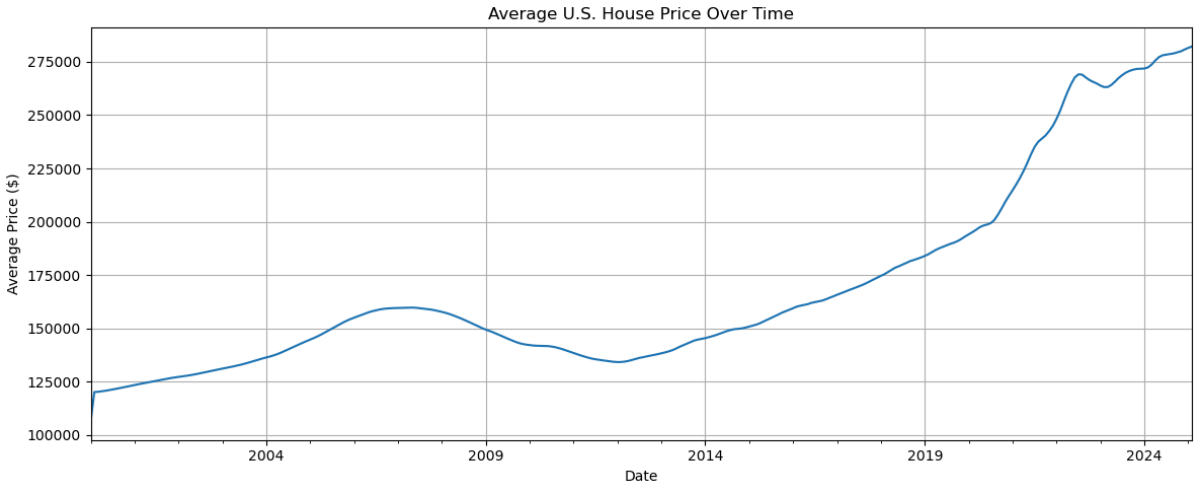
The dataset was reshaped from wide format (one row per region, many columns for dates) to long format using a melt operation. Each row in the transformed dataset corresponds to a unique combination of region and date, with a single Price value. The reasoning behind this decision is that long-form structures are more compatible with machine learning pipelines and visualization tools. Additionally columns previously containing string-formatted dates were converted into proper datetime objects. This ensures compatibility with Python's time-based operations and allows for accurate sorting, filtering, and aggregation. Any null values in the price column were removed to ensure clean and complete data.

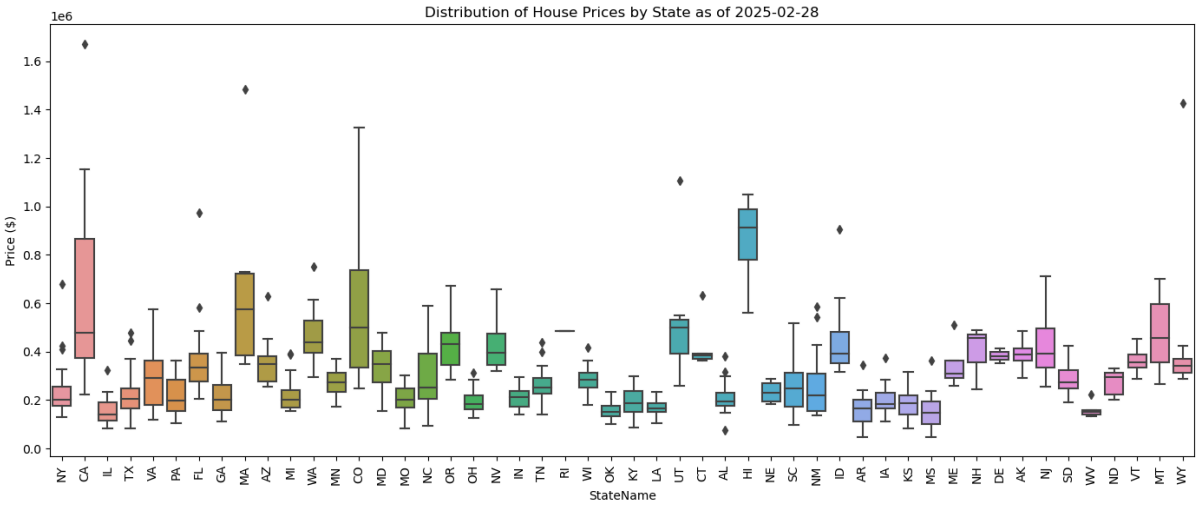
Step 3: Feature Engineering

To enrich the dataset we added, Year, Month and Quarter (Assigned each observation to a calendar quarter (e.g., 2020Q1). These features will help detect seasonal patterns, trend analysis and enhance predictive model performance.

EDA (Exploratory Data Analysis):

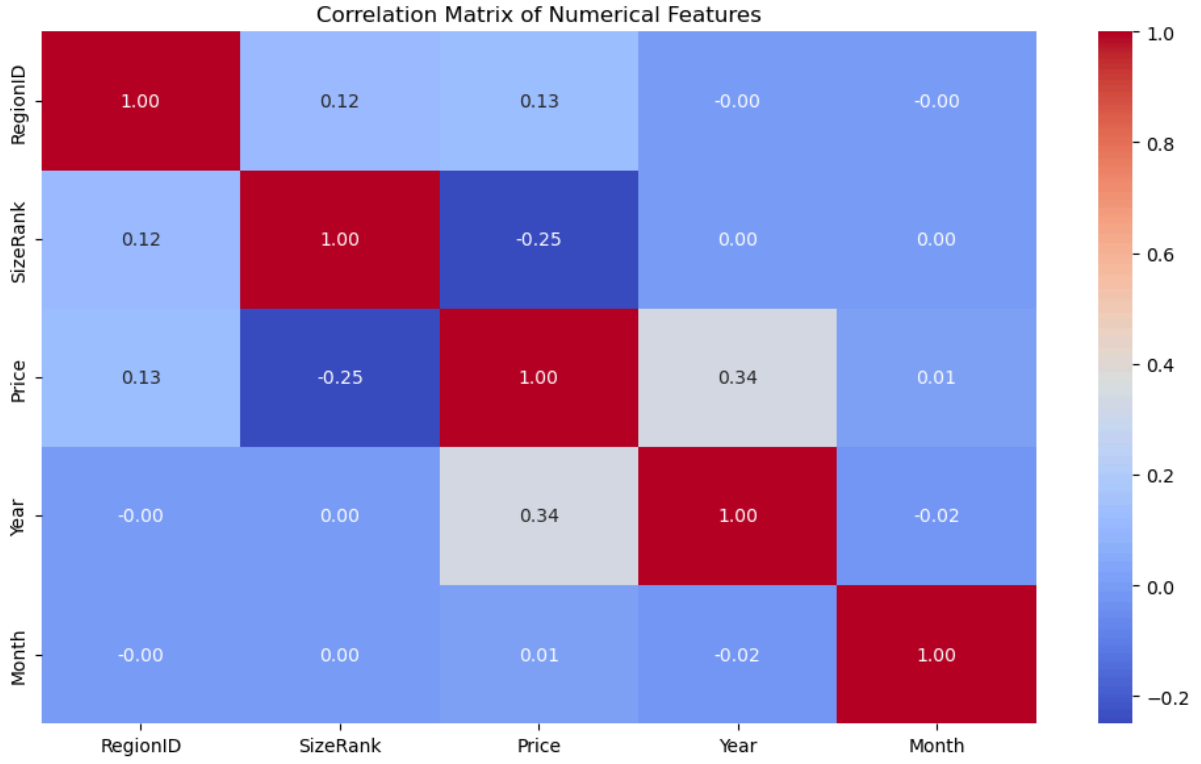
EDA is Crucial step in data science for a multitude of reasons. It can help understand price levels, variability, trends, Identify regional clusters, outliers, reveal seasonality, trends, cycles, Detect errors, anomalies, inconsistent records and guide feature engineering, just to name a few.

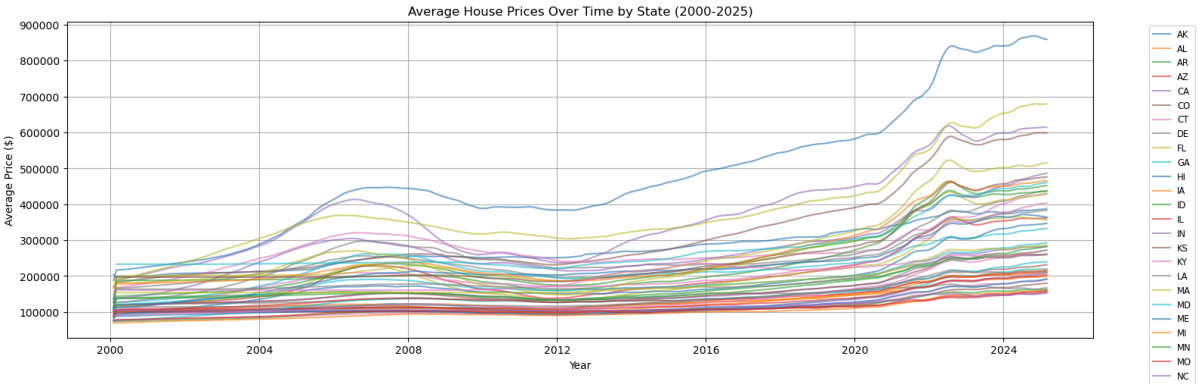




Line Graph: 2000-2007 is the pre-financial crisis growth, showing a steady increase in average house prices, which was driven by housing demand, loose credit, and speculative investments. During 2006-2012, the average house price started to decline, which indicates the collapse of the housing bubble, mortgage defaults, and the overall recession. The price decline is substantial, approximately a 15-25% decline. From 2012 onwards, average housing prices started increasing and eventually exploded. The main drivers are low interest rates, the economic recovery from the financial crises, and urban job growth. The explosive price appreciation has mainly been observed from the late 2020s through 2022 due to the pandemic. Remote work migration, low mortgage rates, and inventory shortages are contributing factors to this explosion.

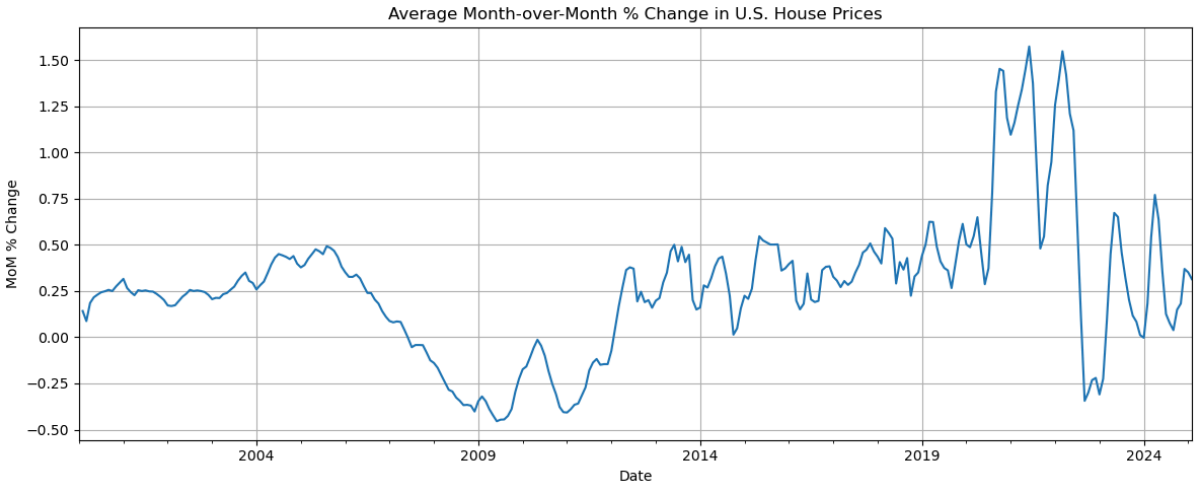
Box Plot: States such as NY, HU, MA, WA, CO, and UT exhibit high prices and high variability, reflecting urban densities, coastal metros, and investment markets (e.g., San Francisco, NYC, and Honolulu). States such as TX, FL, AZ, NC, and VA show moderate prices and spread, which reflects balanced growth states with strong migration trends. States such as WV, MS, AR, ND, SD, and IN have low prices and variability, indicative of stable and affordable rural/inland markets. Finally, something important to note is how even affordable states (e.g., VT, ME, MT) show luxury-region outliers, which May reflect second-home markets or tourism-driven areas.

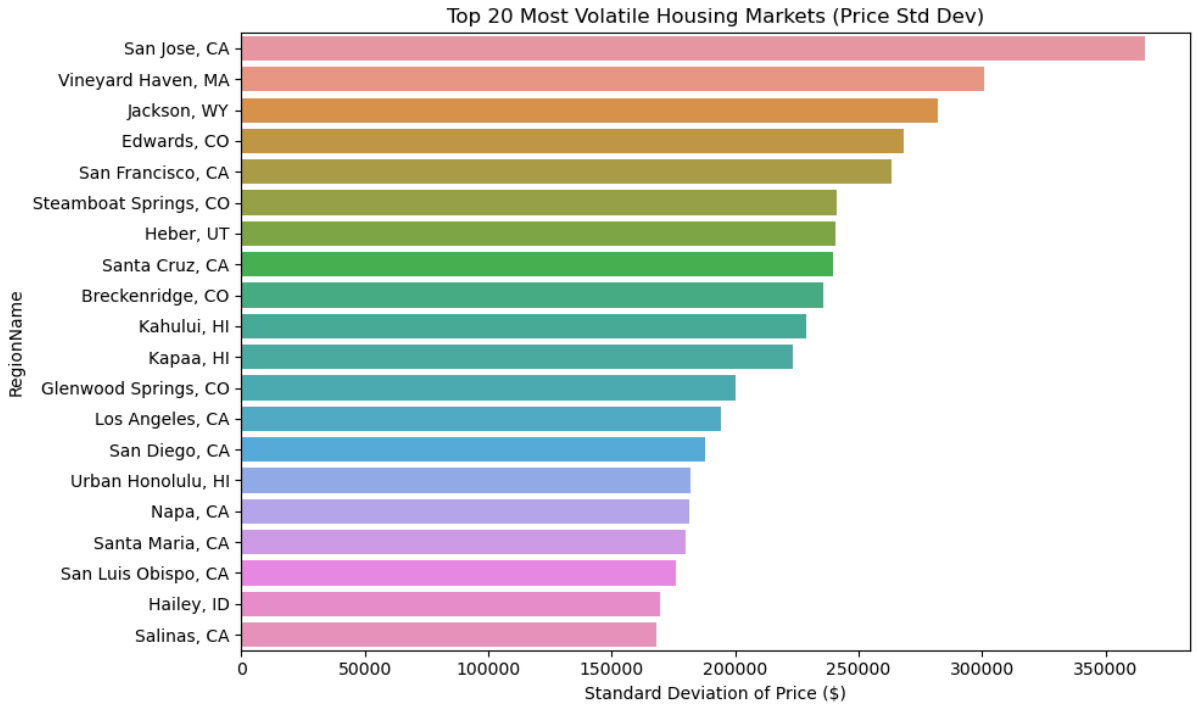


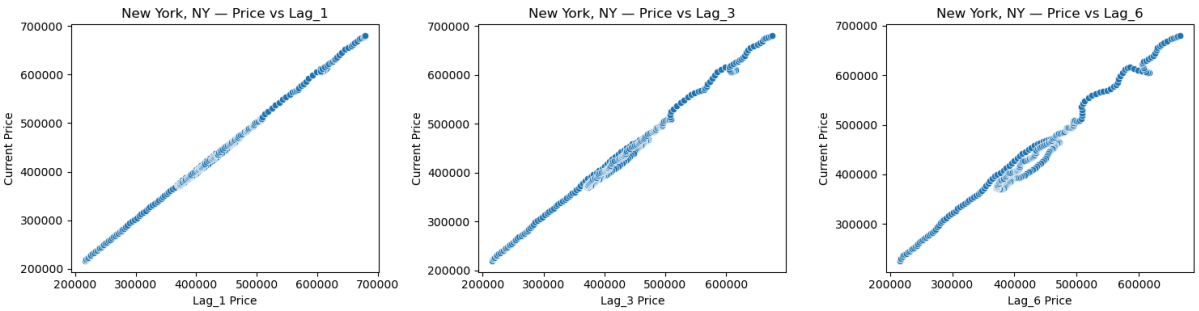


Line Plot: States like CA, HI, MA, DC, and WA will exhibit the highest price levels by 2025. CA (California) is a strong outlier, with prices nearing or exceeding 900k. States such as *WV*, *MS*, *ND*, *IN*, *OH* and *AR* remaining under 900K. States like WV, MS, ND, IN, OH, and AR remain under 200K even in 2025. This also shows modest growth, lower volatility, and possibly less speculative activity. All states follow the general trend of pricing decline from 2007-2012, followed by a recovery from 2012-2020, then the pandemic housing boom.

Heat Map: Price and year exhibit a moderate positive correlation (+0.34), reflecting the long-term appreciation trend in housing markets. As time progresses, house prices increase.
Price and size rank exhibit a moderate negative correlation. A lower size rank means a larger or more prominent job market. Thus, prices tend to be higher in major markets (e.g., NYC, SF, LA), which aligns with economic intuition. Everything else exhibits low or no correlation between variables. Even price and month demonstrate how seasonality is a non-factor in this dataset. Because nothing shows a high correlation when creating our forecasting models we do not need to consider multicollinearity.



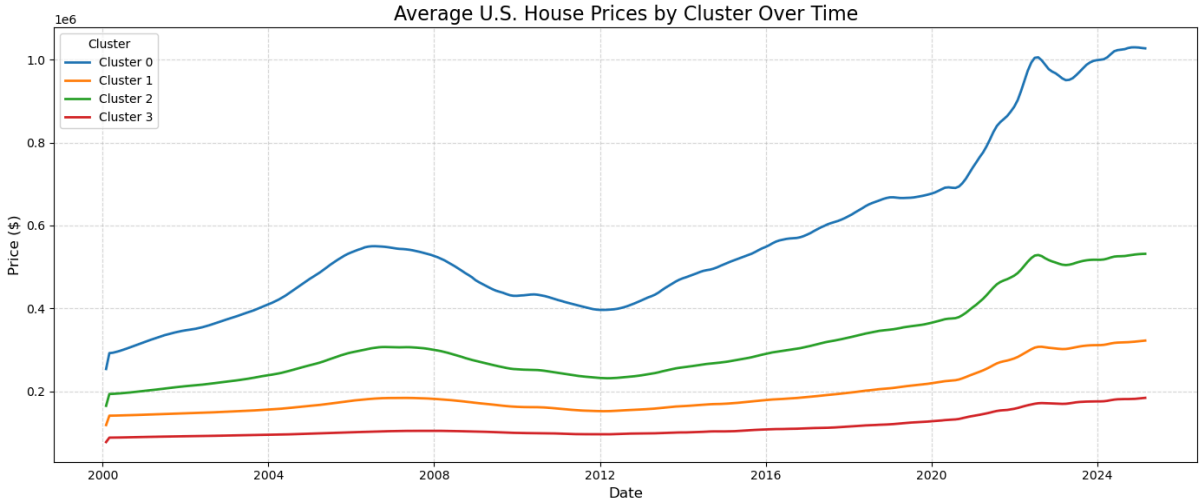




Overall, month-by-month insight shows similar trends as other line graphs. However, between 2022 and 2023, there was a sharp decline and rebound, which reflects the Federal Reserve rate hikes, Sudden shifts in affordability and buyer sentiment, and the Market reacting to inflation and macro policy.

Lag features: These plots show how past prices (lags of 1, 3, and 6 months) correlate with current prices, valid for forecasting and autoregressive modeling. Lag_1 (1-month delay): Near-perfect linear correlation implies that house prices are highly persistent, especially short-term. There is still a strong correlation, but it is slightly more spread. Some "banding" in Lag_6 suggests possible regime shifts in price levels.

Bar Chart: San Jose, CA Leads with Exceptionally High Volatility With a standard deviation approaching $370,000, San Jose tops the list. This suggests extreme price swings in its housing market. This aligns with known macroeconomic trends in the Bay Area, where tech-driven booms and busts strongly influence housing demand and prices. Many of the most volatile regions are high-end or seasonal markets, including Vineyard Haven, MA (Martha's Vineyard); Jackson, WY; Breckenridge, CO; Edwards, CO (ski resorts); Hawaiian cities like Kahului, Kapaa, and Urban Honolulu. These regions experience fluctuations in demand due to seasonal tourism, speculative investment patterns, and greater sensitivity to macroeconomic conditions like interest rates and vacation home demand. Lastly, California dominates the volatility spectrum. Over 10 of the 20 cities are in California, including San Francisco, Santa Cruz, LA, San Diego, and Napa.



Cluster 0: CA, CO, CT, FL, HI, ID, MA, NY, UT, WY
Cluster 1: AK, AL, AR, AZ, CA, CO, CT, DE, FL, GA, IA, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, ND, NE, NH, NJ, NM, NV, NY, OH, OR, PA, SC, SD, TN, TX, UT, VA, VT, WA, WI, WV, WY
Cluster 2: AK, AZ, CA, CO, DC, FL, HI, ID, MA, MD, ME, MT, NC, NH, NJ, NM, NV, OR, RI, SC, TX, UT, VA, VT, WA, WY
Cluster 3: AL, AR, AZ, CO, FL, GA, IA, IL, IN, KS, KY, LA, MD, ME, MI, MN, MO, MS, MT, NC, ND, NE, NH, NM, NY, OH, OK, OR, PA, SC, SD, TN, TX, UT, VA, WI, WV

Cluster 0:
Has High-Growth, High-Value, and Coastal/urban markets. States include CA, NY, MA, HI, CT, FL, CO, and UT. Common traits include top-tier metros (e.g., San Francisco, NYC, Boston). Strong appreciation was observed post-2012. Cluster 0 may have markets that are price-inelastic, investor-driven, and reflect high demand and low supply. Furthermore, volatility may be present due to market overheating or policy sensitivity.

Cluster 1:
Has Mixed Growth and Diverse Geography. These States include CA, FL, GA, TX, AK, AL, and AR. Cluster 1 represents a mix of fast-growing secondary markets and stable metros. Cluster 1

exhibits Mid-tier appreciation, potential indicating suburban regions around big cities. These regions are in transition, neither speculative nor stagnant, reflecting suburban sprawl, remote work relocations, and moderate affordability.

Cluster 2:
Represents tourist/Seasonal and Coastal Booms. States include HI, FL, CA, MA, MD, MT, ID, ME. Many states are vacation destinations, resort economies, or remote-friendly areas. They exhibit nonlinear growth patterns, likely with COVID-induced surges. These regions are influenced by tourism, second-home buyers, and telework migration and are highly susceptible to macroeconomic shocks (e.g., interest rate sensitivity).

Cluster 3:
Represents stable/flat Growth and Interior/Rural States. States include AL, IN, KS, LA, MS, MO, OH and OK. Cluster 3 is Dominated by inland, southern states, smaller cities, and rural economies with Price trajectories show low volatility and low appreciation. Additionally, they reflect stable but less dynamic markets and are attractive for affordability but lack investor attention or growth potential.

Machine Learning Models:

We decided to create a LightGBM, XGBoost, and a Sarimax model. Light GBM models help directly support categorical features, have fast training and low memory usage, and capture nonlinear feature interactions. XGBoost is more stable and regularized than some other tree-based methods and often outperforms linear models when feature interactions are complex. Sarimax is designed for explicit time series modeling and helps forecast future values.

```
📊 LightGBM (All Regions) Performance:
MAE:     8,151.45
MedAE:   1,218.31
RMSE:    42,100.84
MAPE:    1.41%
SMAPE:   1.49%
R²:      0.9339

📊 XGBoost (All Regions) Performance:
MAE:     8,732.62
MedAE:   1,134.24
RMSE:    43,319.88
MAPE:    1.49%
SMAPE:   1.55%
R²:      0.9300

📊 SARIMAX (Average over Top Regions):
MAE:     37,788.39
MedAE:   45,648.18
RMSE:    42,295.59
R²:      -1.2158
```

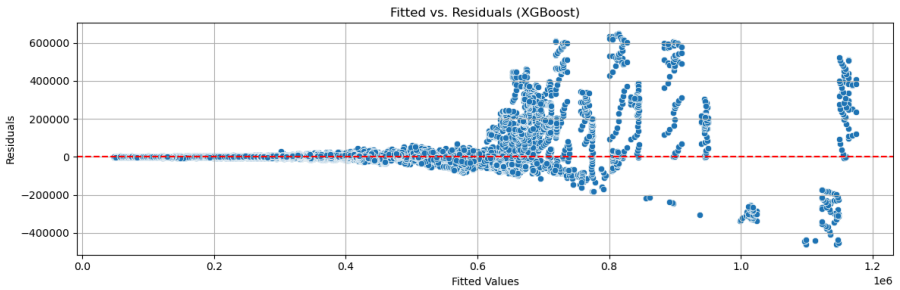
So far, LightGBM has the best overall performance in almost all metrics: lowest MAE, RMSE, MAPE, SMAPE, and highest R^2. $R^2 = 0.9339$ Explains ~93% of the variance in housing prices across all regions. MAE under $10K means it's strong even at the regional scale. It is slightly more accurate than XGBoost but with better stability (lower RMSE and SMAPE).
XGBoost also performed very well. $R^2 = 0.93$ shows near-identical variance capture. MAE is only ~$580 higher than LightGBM. There was slightly more error volatility (higher RMSE and SMAPE).
Sarimax performed terribly. MAE is 4.5x worse than LightGBM. $R^2 = -1.22$ indicates it's performing worse than a horizontal mean line. This is because Sarimax only applied to the top 5 regions, which have high prices and cities with large volatility. It will not generalize well, especially considering that our dataset does not contain in-depth columns to increase model accuracy. Therefore, we will only consider the first two models moving forward.

The next step is to perform K-fold cross validation. K-fold validation is a technique used to assess how well a model generalises to a dataset, helping reduce variance and bias in model performance estimates. For our use in 5-fold validation, the data is split into 5 parts and iteratively trained and validated. For each iteration it uses k-1 for training and the last fold for testing. It rotates the test fold into the training data each time. Lastly it averages all the evaluation metrics giving a more accurate performance evaluation.

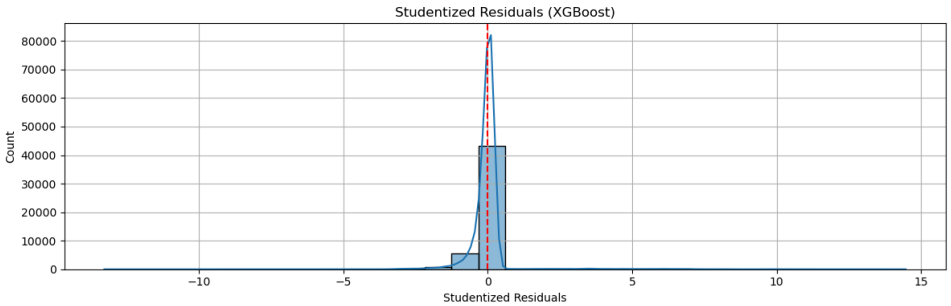
	Model	Metric	Mean	Std Dev
0	LightGBM	MAE	1032.394613	10.768528
1	LightGBM	MedAE	462.976666	2.856250
2	LightGBM	RMSE	2861.181333	85.038169
3	LightGBM	MAPE	0.498142	0.003646
4	LightGBM	SMAPE	0.497572	0.003589
5	LightGBM	R^2	0.999370	0.000034
6	XGBoost	MAE	886.614937	8.585176
7	XGBoost	MedAE	432.259482	7.389027
8	XGBoost	RMSE	2222.967931	69.818206
9	XGBoost	MAPE	0.447291	0.005514
10	XGBoost	SMAPE	0.447126	0.005473
11	XGBoost	R^2	0.999619	0.000028

In this scenario, XG boost performed better. For both models, the MAE and RMSE improvements are significant, with MAPE/SMAPE under 0.5%, which is incredibly accurate. Standard deviations are very low, suggesting high model stability across folds. For both models, the R^2 is >0.999, which could imply overfitting.

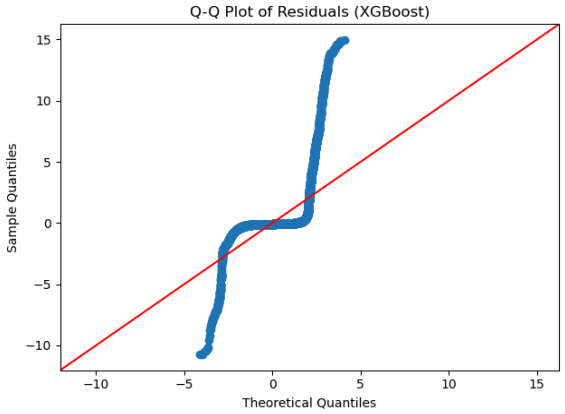
Residual Analysis:



Fitted Values Vs. Residuals: A clear funnel pattern exists where the residual variance increases with higher fitted values. This means that model performance worsens in the later years. This suggests heteroscedasticity.



Student Residuals: It is a longer right-skewed distribution, which indicates overpredictions. It is still very compact, but mitigating this factor may help model performance.



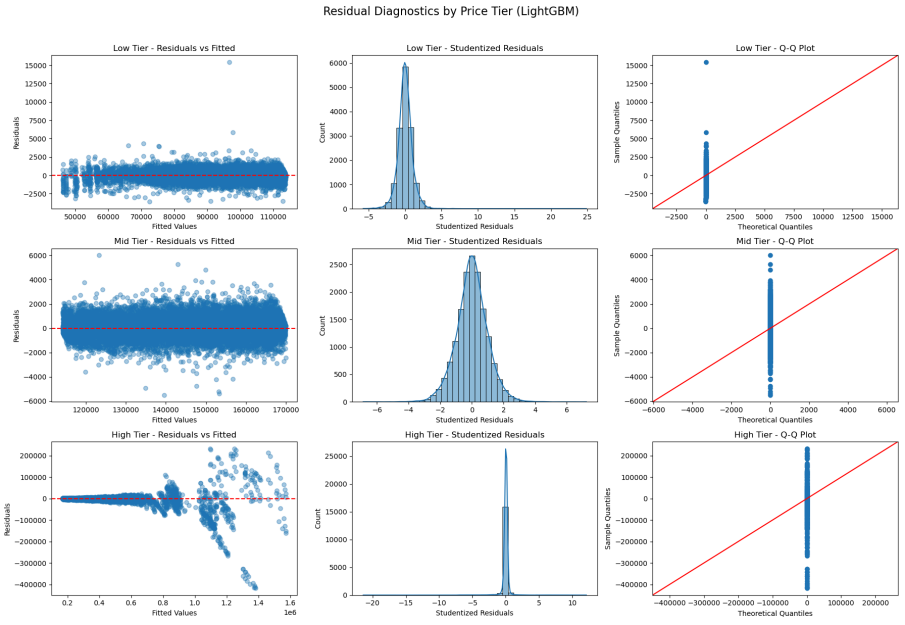
Q-Q plot: There is an apparent deviation at both tails of the Q-Q plot. The S shape indicates something called fat-tailed residuals. This is when extreme under and over-predictions exist more than expected.

(LightGBM has identical residuals so analysis is same as XGBoost)

Creation Of better Model:

We will split the machine learning model into low, mid, and high-tier models to fix the high levels of heteroscedasticity and S curve Q-Q plots. This model fixes the 'one size fits all' mistake the previous models were making.

	Tier	Model	MAE	MedAE	RMSE	MAPE	SMAPE	R2
0	Low	LightGBM	446.156922	336.459471	620.577190	0.499008	0.498723	0.998109
1	Low	XGBoost	401.506389	300.873400	564.930152	0.447968	0.447799	0.998433
2	Mid	LightGBM	624.794413	490.780941	820.587643	0.437362	0.437621	0.997269
3	Mid	XGBoost	552.991275	434.961809	730.698272	0.387882	0.388039	0.997834
4	High	LightGBM	4856.133531	1670.406761	19490.526161	0.956785	0.943829	0.988020
5	High	XGBoost	4915.139244	1565.346838	19165.706205	0.934996	0.923022	0.988416



Because both machine learning models are incredibly similar, we will aggregate rational analysis as graphs are identical.

Low Tier: The residuals are balanced and randomly scattered across the horizontal red line for our residuals vs fitted values, indicating that the linearity assumptions are met. A constant spread across the range of fitted values suggests that homoscedasticity is satisfied. There are no outliers in the scatter plot. Our studentized residuals are slightly right-skewed but primarily symmetric. There are no strong signs of non-normality indicating an accurate model. Due to splitting the dataset into tiers, the Q-Q plot will be vertical (because splitting into tiers results in narrowing the range of Price, Training models on more homogeneous data, and Making residuals much more tightly clustered). However, this is of no concern, as outliers, heteroscedasticity, and non-normality do not exist in the low tier.

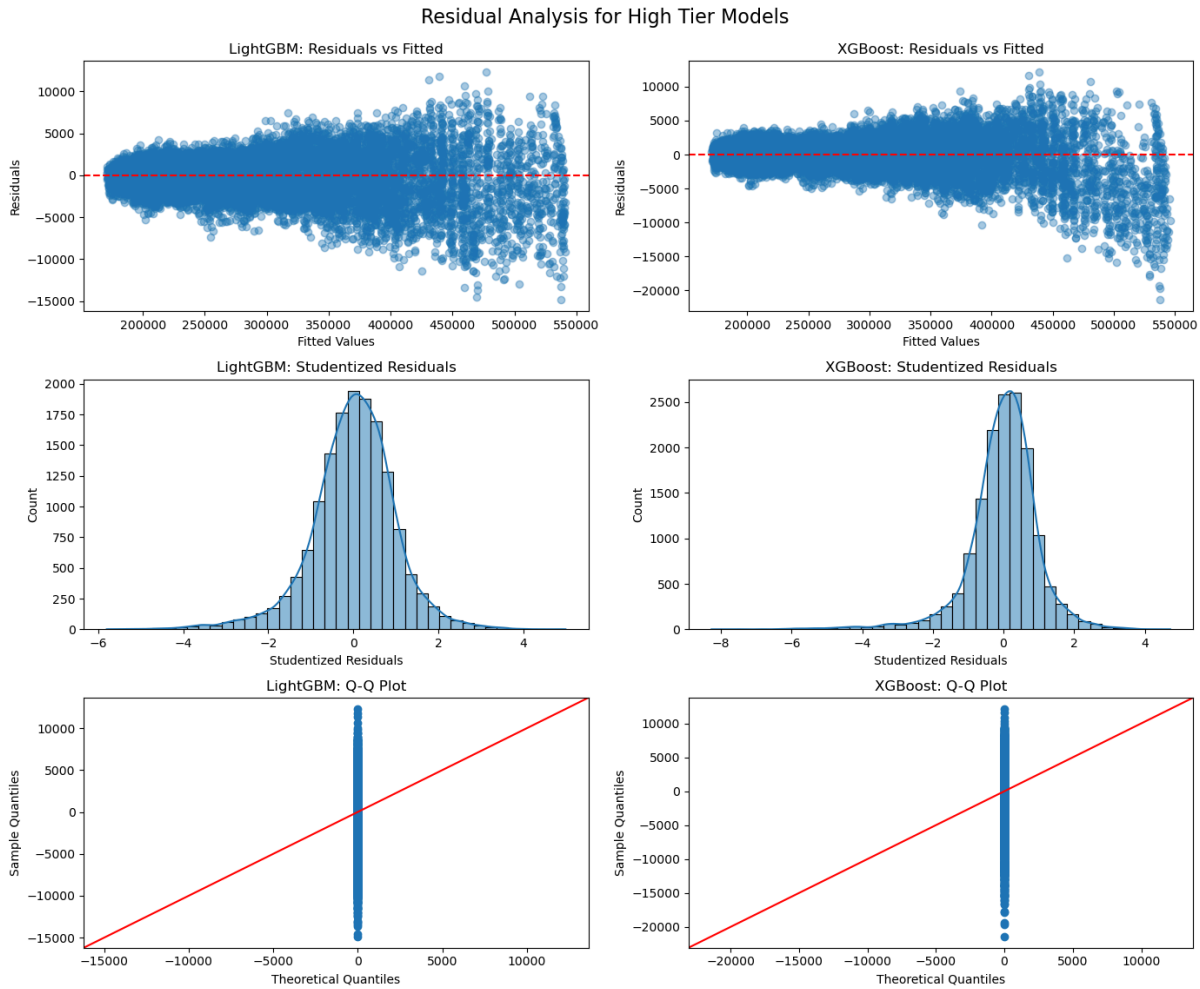
Mid Tier: For our residuals vs. fitted values, the residuals are balanced and randomly scattered across the horizontal red line, indicating that the linearity assumptions are met. There are more outliers than the low tier, but not enough to skew the data. The studentized residuals have a Clean bell curve shape, indicating good symmetry. It is the same story as the Q-Q plot.

High Tier: When analyzing the high-tier data, specific problems arise. There are strong signs of heteroskedasticity. More specifically, the model is overpricing, and the actual values are hundreds of thousands of dollars below the real value, indicating a tier with high variability in housing prices. There is a Sharp peak at zero but very fat tails for the studentized residuals, indicating significant outliers.

To fix our predictive issues with high-tier values, we will filter out housing prices that should be in the mid-tier model but are being funneled into the high-tier model.

```
📊 LightGBM (High Tier Capped) Performance:
MAE:     1,860.94
MedAE:   1,419.87
RMSE:    2,527.25
MAPE:    0.62%
SMAPE:   0.61%
R²:      0.9992

📊 XGBoost (High Tier Capped) Performance:
MAE:     1,784.27
MedAE:   1,285.07
RMSE:    2,580.48
MAPE:    0.58%
SMAPE:   0.58%
R²:      0.9991
```



Fitted vs. Residuals: Both models show a more even distribution spread for the fitted vs. residuals. Even though the outliers may seem very wide, under or over-predicting by 20,000 is very small considering how expensive houses are in this data bin. The lightGBM rituals are not as tight as the XGBoost. However, the prediction error remains at 10,000-15,000. For XGBoost, the residuals are tighter; however, when the housing prices are more recent, prediction jumps to the 10,000-20,000 range slightly more than LightGBM.

Studentized residuals: Both models follow an ideal bell-shaped distribution. The peak is centered around zero, indicating that the model is unbiased. LightGBM is more unbiased than XGBoost.

Therefore this fix was successful. We only created a 6 month forecast because it takes a very long time for the CSV file to be created.

Creation Of 6 Month Future Prediction

We Created code that automated predictive forecasts from March 2025-August 2025. Using LightGBM and XGBoost 2 separate evaluations were made of each model. The code loops over their price tiers and creates separate modes for each. For each region within a price tier the model generates stepwise forecasts for six months into the future one at a time. For each step it builds a feature row using lagged prices and time-based features, then it predicts the price using both LightGBM and XGBoost. It then appends the prediction to a growing history so that next-step lags are up to date and Stops if any required lag is missing (e.g., short price history). Final predictions are saved as a flat CSV

containing pricetier, regionname,date,LightGBM prediction and XGBoost prediction. (6 month forecast is provided in excel when handed in).

6 Month Prediction Evaluation:

Overall Model Comparison (LightGBM vs XGBoost by Price Tier)

Price Tier	LightGBM Mean	LightGBM Std	XGBoost Mean	XGBoost Std	Mean Error (LGB - XGB)	Std Dev of Error	Mean Absolute Error	Max Absolute Error
High	$337,756.70	$187,804.77	$337,769.34	$190,759.41	-$12.64	$16,516.73	$3,679.63	$305,282.68
Mid	$165,334.30	$10,720.41	$165,450.94	$10,793.86	-$116.63	$396.63	$324.38	$1,908.94
Low	$111,870.72	$7,208.89	$111,926.67	$7,233.49	-$55.95	$273.33	$202.36	$2,489.56

Mean Predictions are nearly identical across both models in all tiers, indicating general agreement in trends, however, absolute error variability is much higher in high-tier markets. Mid and low tier models are extremely consistent, which demonstrates high agreeableness and robustness for most of the data set. The top disagreements in models lie in San Diego CA (high tier, $305,283 error), Santa Maria, CA (high tier, $260,735) and another 4x in San Diego CA (high tier, $295k-$285k).

Overall the most severe mismatches happen in high tier markets, due to greater price volatility and fewer data points for ultra high value homes, for example when viewing housing in San Diego prices range from 500K to 6M. In order for the model to generalize better in high tier markets, more specific regions, amount of bedrooms, bathrooms, and parking would help add predictive power if they existed in the dataset.

In conclusion our model has extremely high agreeableness and robustness for low to mid tier data, however diverges severely in high tier pricing due to a lack of specific data.