CISC 3225: Final Project
Analysis of USA Real Estate Data
Name: Said Lfagrouche
Date: 05/22/2024

Introduction:

Kaggle Dataset: https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data

- **About the Dataset**
  This report is based on an analysis of the USA Real Estate dataset, which includes comprehensive real estate listings across various states and zip codes in the United States. The data is sourced from Realtor.com, a leading real estate listing website in the United States with over 100 million monthly active users as of 2024. The dataset was used under the terms that it is intended for educational purposes and contains information collected for analyzing trends in the housing market.

- **Dataset Description**
  The dataset comprises a single CSV file, "realtor-data.csv," which contains 2,226,382 entries spread across 10 key columns:

  - Brokered by: Categorically encoded agency or broker.
  - Status: Housing status (ready for sale or ready to build).
  - Price: Listing price or recently sold price.
  - Bed: Number of bedrooms.
  - Bath: Number of bathrooms.
  - Acre Lot: Property or land size in acres.
  - Street: Categorically encoded street address.
  - City: City name.
  - State: State name.
  - Zip Code: Postal code of the area.
  - House Size: Living space in square feet.
  - Prev Sold Date: Date when the property was last sold.
  - For privacy considerations, the dataset encodes broker and street addresses categorically. The "Acre Lot" and "House Size" columns specify the extent of the property and the living area, respectively.

- **Purpose of the Analysis**
  The primary objective of this analysis is to uncover insights into the U.S. housing market through a series of high-level analytical approaches, including statistical tests and data modeling. The questions guiding this analysis are:
    - Is it possible to forecast housing prices utilizing the characteristics detailed in the dataset?
    - What is the relationship between housing prices and location-specific attributes?
    - How do housing prices vary across different locations within the USA?
    - To what extent do features like the number of bedrooms and bathrooms influence housing prices? Also, are there underlying patterns that are not immediately apparent?

Exploratory analysis:

Observations from Descriptive Statistics of Real Estate Data:

- 1. The 'price' variable shows a significant range, with a maximum of over 2 billion, which suggests potential outliers or special property listings.
- The 'bed' (bedrooms) and 'bath' (bathrooms) columns have unusually high max values (473 for bedrooms and 830 for bathrooms), indicating possible data entry errors or extreme cases.
- The 'lot_size_acres' has a maximum of 100,000 acres, which is extraordinarily large for residential properties, pointing towards potential inclusion of commercial or agricultural land, or errors.
- The 'living_area_sq_ft' also indicates extreme values with a maximum of over 1 billion square feet, highly unlikely for individual real estate properties and likely indicating data errors.
- The standard deviation in 'price' and 'living_area_sq_ft' is considerably large, highlighting substantial variability in the dataset, which could affect modeling accuracy if not addressed (e.g., through log transformation or removing outliers).
- The minimum values for 'price' and 'living_area_sq_ft' are 0, which may require further investigation to determine if these represent missing or placeholder values that need to be handled.

	price	bed	bath	lot_size_acres	zip_code	living_area_sq_ft
count	1.495891e+06	1.495891e+06	1.495891e+06	1.495891e+06	1.495891e+06	1.495891e+06
mean	5.562242e+05	3.252950e+00	2.407734e+00	1.591930e+01	4.700219e+04	2.774207e+03
std	2.545138e+06	1.479286e+00	1.630392e+00	7.766317e+02	2.888466e+04	8.507990e+05
min	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	4.000000e+00
25%	1.595000e+05	3.000000e+00	2.000000e+00	1.800000e-01	2.248500e+04	1.515000e+03
50%	3.250000e+05	3.000000e+00	2.000000e+00	3.100000e-01	3.813500e+04	1.803000e+03
75%	5.650000e+05	4.000000e+00	3.000000e+00	9.500000e-01	7.564700e+04	2.175000e+03
max	2.147484e+09	4.730000e+02	8.300000e+02	1.000000e+05	9.999900e+04	1.040400e+09

Analysis of Correlation Coefficients:

- Price and Bed: A correlation coefficient of 0.107 suggests a weak positive relationship. More bedrooms tend to increase property prices, but the effect is not very strong.

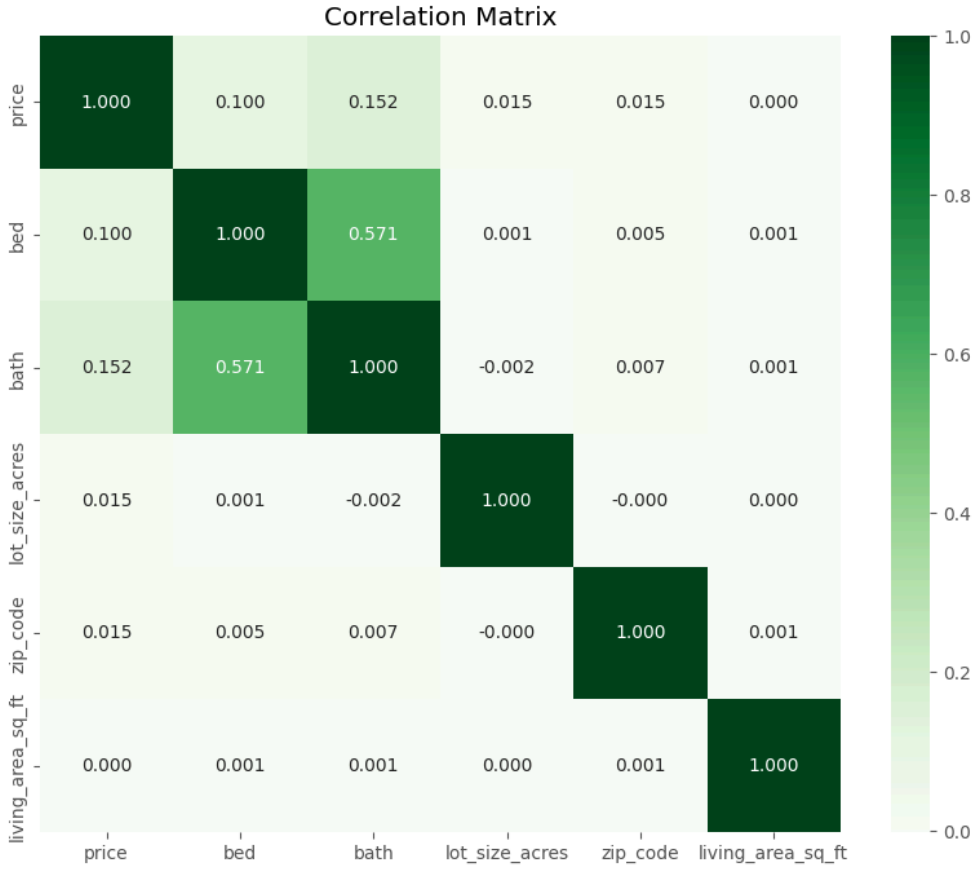
- Price and Bath: The correlation of 0.166 with bathrooms is slightly stronger than with bedrooms, indicating that properties with more bathrooms are likely to be priced higher.

- Price and Lot Size Acres: The very weak correlation of 0.014 suggests almost no linear relationship between the size of the lot and the price.

- Price and Living Area: Surprisingly, there is almost no correlation (0.000246) between the living area square footage and the price. This is unusual as larger living areas typically command higher prices.

- Bed and Bath: There is a moderate positive correlation of 0.581 between the number of bedrooms and bathrooms, which makes sense as larger homes typically have more of both.

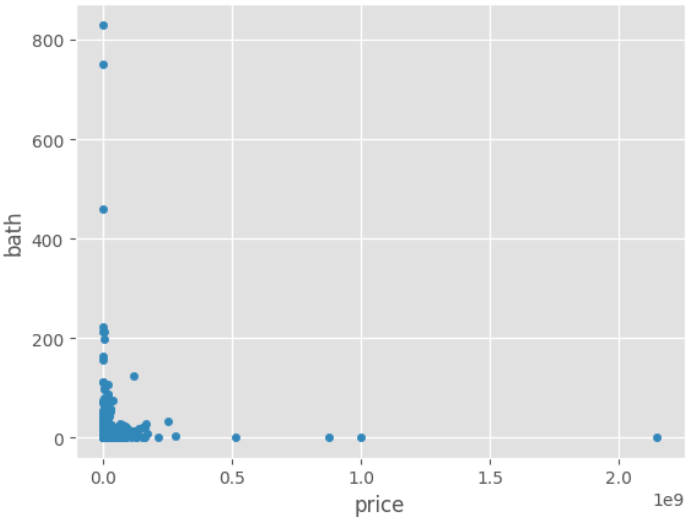
- Lot Size, Zip Code, and Living Area:
  These variables show very low to negligible correlations with other features. Lot size acres and living area sq ft, particularly, show no meaningful correlations with price or other variables, which is atypical in real estate analysis.

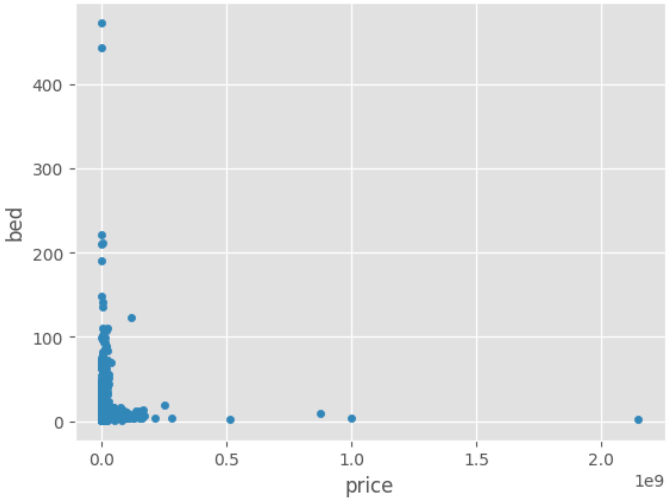


Some interesting relationship visualization:

- Displays a weak positive relationship where properties with more bathrooms tend to be priced higher, correlation coefficient = 0.166.

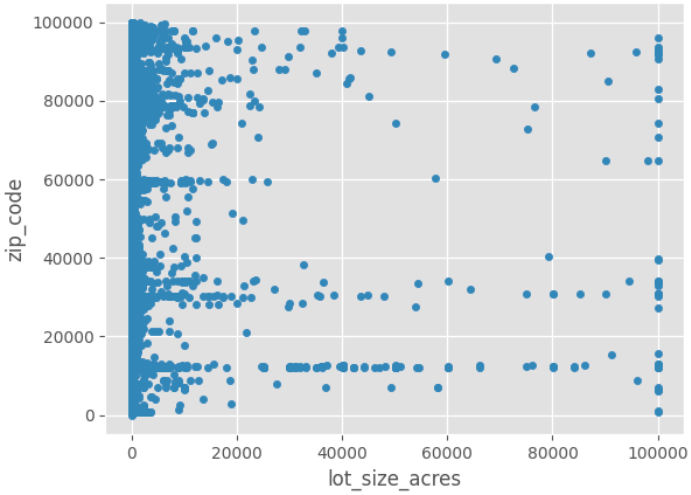
- Shows a weak positive relationship, suggesting that an increase in bedrooms can slightly increase property prices, correlation coefficient = 0.107.

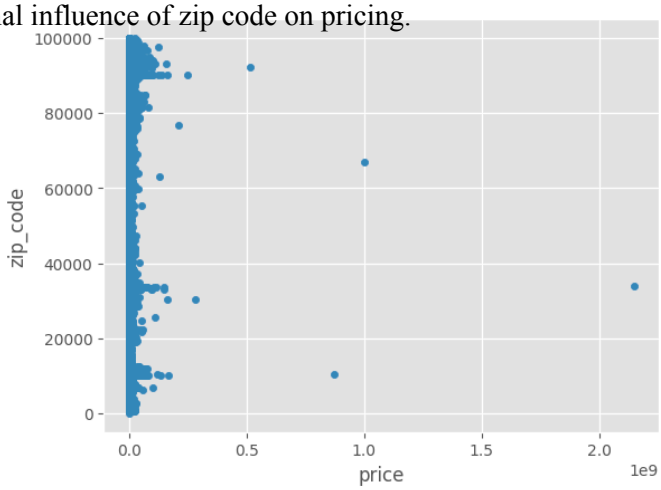




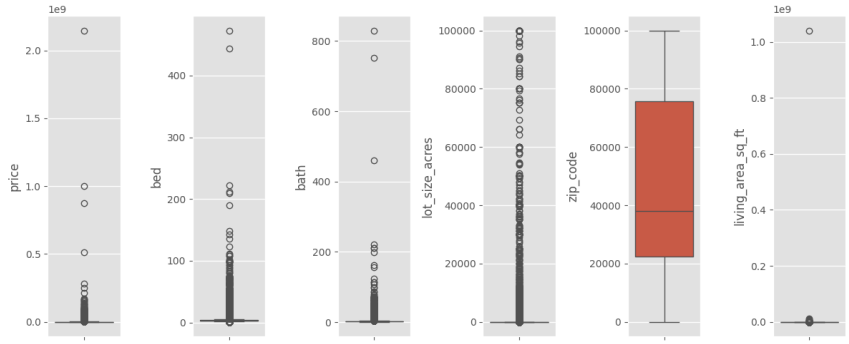
- The plot shows negligible correlation between lot size and zip code, correlation coefficient = -0.0015. Indicates that lot size doesn't vary systematically across zip codes.

- There appears to be a very weak relationship between the price and the zip code, correlation coefficient = 0.028. This indicates minimal influence of zip code on pricing.

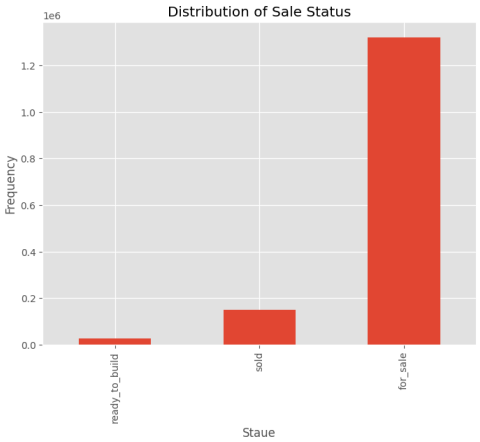




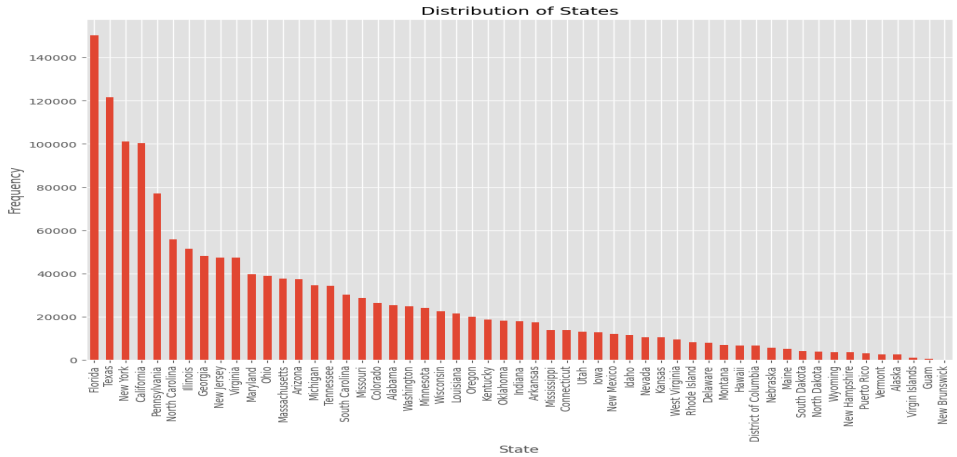
- Visualizations such as boxplots could effectively illustrate the extreme variability in price and living area, highlighting the long tails that suggest the presence of luxury or outlier properties.
  Distribution of Sale Status:



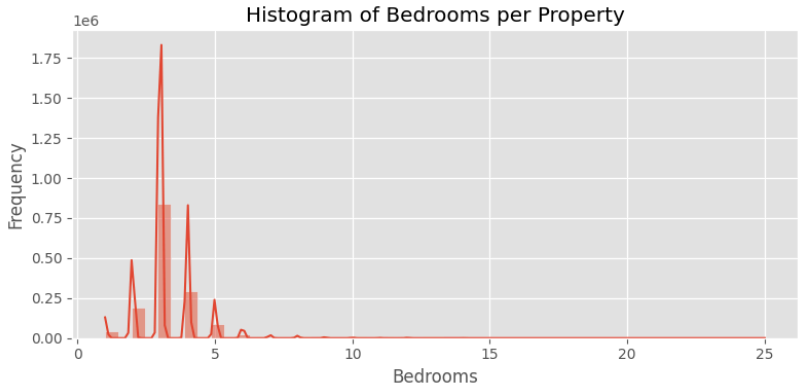
- The dataset includes a significant majority of properties listed for sale (1,321,436), with far fewer properties sold (149,453) or ready to build (25,002). This distribution points to a high inventory of available properties relative to those that are under construction or have recently changed ownership.



- The distribution of real estate listings is highly concentrated in states like Florida, Texas, and New York, which are known for their large populations and dynamic real estate markets. In contrast, states such as North Dakota and Wyoming show significantly fewer listings, reflecting their lower population densities and real estate activity.



- The majority of properties listed have between 2 to 4 bedrooms, aligning with the typical family unit's needs. Properties with a high number of bedrooms (over 20) are less common and might represent outliers or specialized real estate such as luxury estates or commercial lodging facilities.



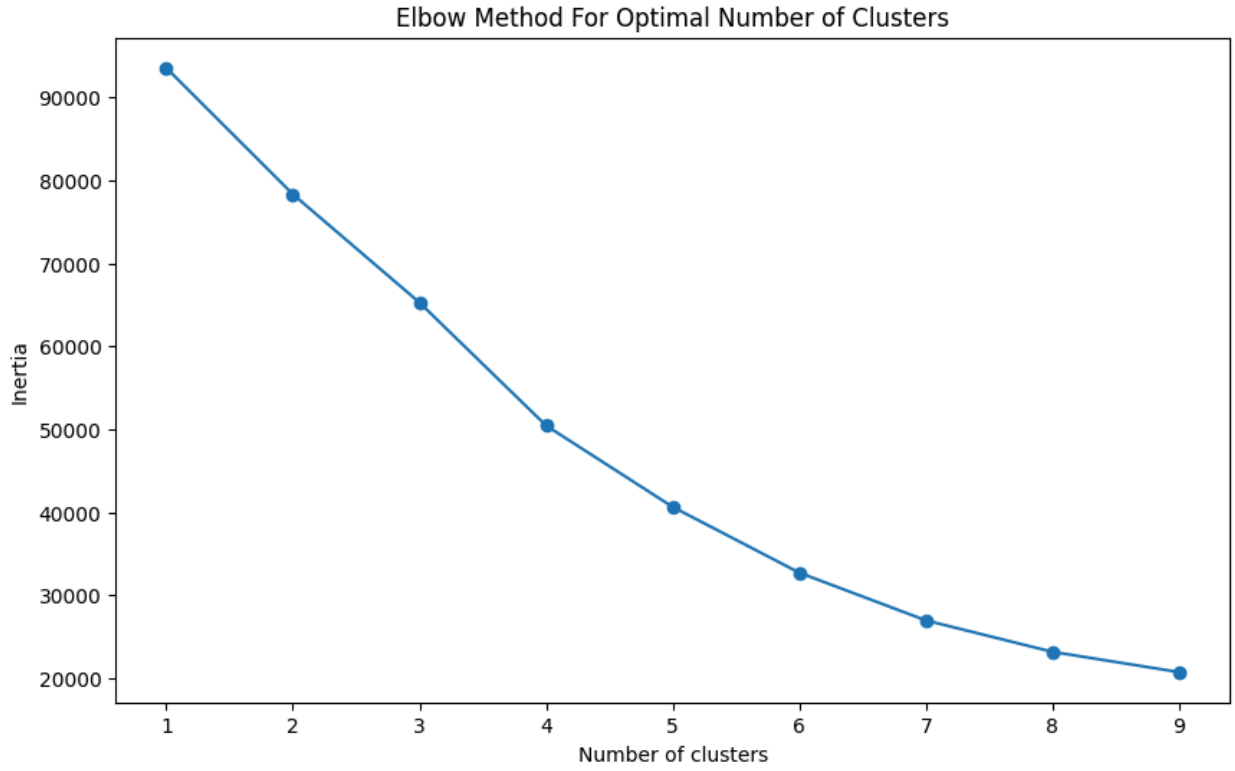
Conclusion:
The exploratory analysis highlights several key aspects of the USA real estate market, including high variability in property features and prices, the presence of outliers, and weak correlations between traditionally predictive features and property prices. These findings underscore the complexity of the real estate market and the need for careful data handling and robust analytical techniques to derive accurate insights. This analysis sets the stage for more detailed statistical tests and predictive modeling to better understand market dynamics and aid stakeholders in making informed decisions.

High-level analysis:

1. Clustering Analysis:

Objective: Identify distinct clusters within the housing market based on price, number of bedrooms, number of bathrooms, property size, and house size.

Cluster	Price	Beds	Baths	Acre Lot	Zip Code	House Size (sq. ft.)	Status
0	$438,249.88	3.67	2.54	5.65	6933.96	2131.80	For Sale
1	$7,061,831.00	9.16	9.07	16.08	2763.71	8185.26	For Sale
2	$704,111.82	3.70	2.55	5.61	2193.83	2268.30	For Sale
3	$585,000.00	4.00	4.00	100000.00	926.00	3300.00	For Sale
4	$8,250,000.00	5.00	6.00	33.29	775.00	1450112.00	For Sale



Insights:

- Cluster 0:
  - Houses in this cluster are mid-range in price and size.
  - They tend to have fewer acres compared to other clusters.
  - All houses in this cluster are for sale.

- Cluster 1:
  - Houses here are the highest in price among all clusters.
  - They have a significantly larger number of bedrooms and bathrooms.
  - All houses in this cluster are for sale, indicating luxury properties.

- Cluster 2:
  - Houses in this cluster are higher in price with moderate house sizes.
  - The lot size is similar to Cluster 0, indicating suburban properties.
  - All houses in this cluster are for sale.

- Cluster 3:
  - This cluster contains houses with very large lot sizes and moderate house sizes.
  - The price is mid-range, similar to Cluster 0 and 2.
  - All houses in this cluster are for sale, indicating unique property types with large lots.
- Cluster 4:
  - Houses in this cluster are very high in price, with a moderate number of bedrooms and bathrooms.
  - They have the largest house sizes among all clusters.
  - All houses in this cluster are for sale, suggesting they might cater to ultra-luxury buyers.

Answering the Objective:

- Can we identify distinct clusters of housing markets based on housing price, number of beds and baths, property size, and house size?

  Yes, distinct clusters can be identified in the housing market based on the mentioned features. The analysis reveals five clusters with unique characteristics:
  - Cluster 0: Represents mid-range homes currently on the market.
  - Cluster 1: Represents high-priced luxury homes currently on the market.
  - Cluster 2: Shows higher-priced homes with moderate house sizes, currently on the market.
  - Cluster 3: Comprises homes with very large lot sizes and moderate house sizes, currently on the market.
  - Cluster 4: Includes ultra-luxury homes with very high prices and the largest house sizes, currently on the market.

2. Chi-square test

Objective: Is there a significant association between the housing status (status) and the state (state)?

Analysis of Results

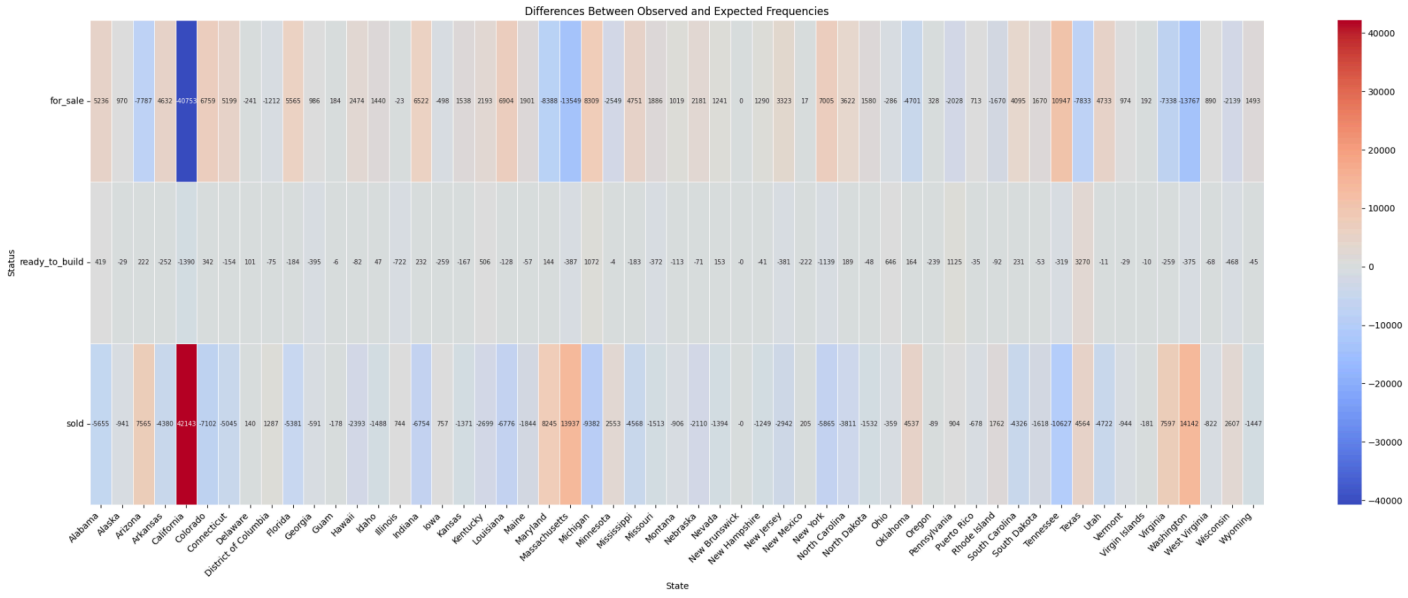
Chi-Square Test Outputs

- **Chi-Square Statistic**: 213189.42
- **P-Value**: 0.0
- **Degrees of Freedom**: 108

Interpretation

1. **Chi-Square Statistic**: A high value of 213189.42 indicates a substantial difference between the observed and expected frequencies, suggesting a strong association between the variables.
2. **P-Value**: A p-value of 0.0 indicates the observed association is highly statistically significant, reinforcing the rejection of the null hypothesis.
3. **Degrees of Freedom**: 108, reflecting the complexity of the association being tested.
4. **Expected Frequencies**: Represent the counts we would expect if there were no association between status and state.

Based on the Chi-Square test results, there is a highly significant association between the housing status (status) and the state (state). This means that the distribution of housing status (for sale, ready to build, sold) varies significantly by state.



Detailed Analysis of Selected States

I conducted a detailed Chi-Square test with a few selected states to understand the association better. Here are the results:

California:

- **For Sale**: Observed (101,034) significantly lower than expected (141,786.92).
- **Ready to Build**: Observed (1,168) significantly lower than expected (2,557.63).
- **Sold**: Observed (125,013) significantly higher than expected (82,870.45).
- **Conclusion**: High market turnover with fewer new listings and constructions.

New York

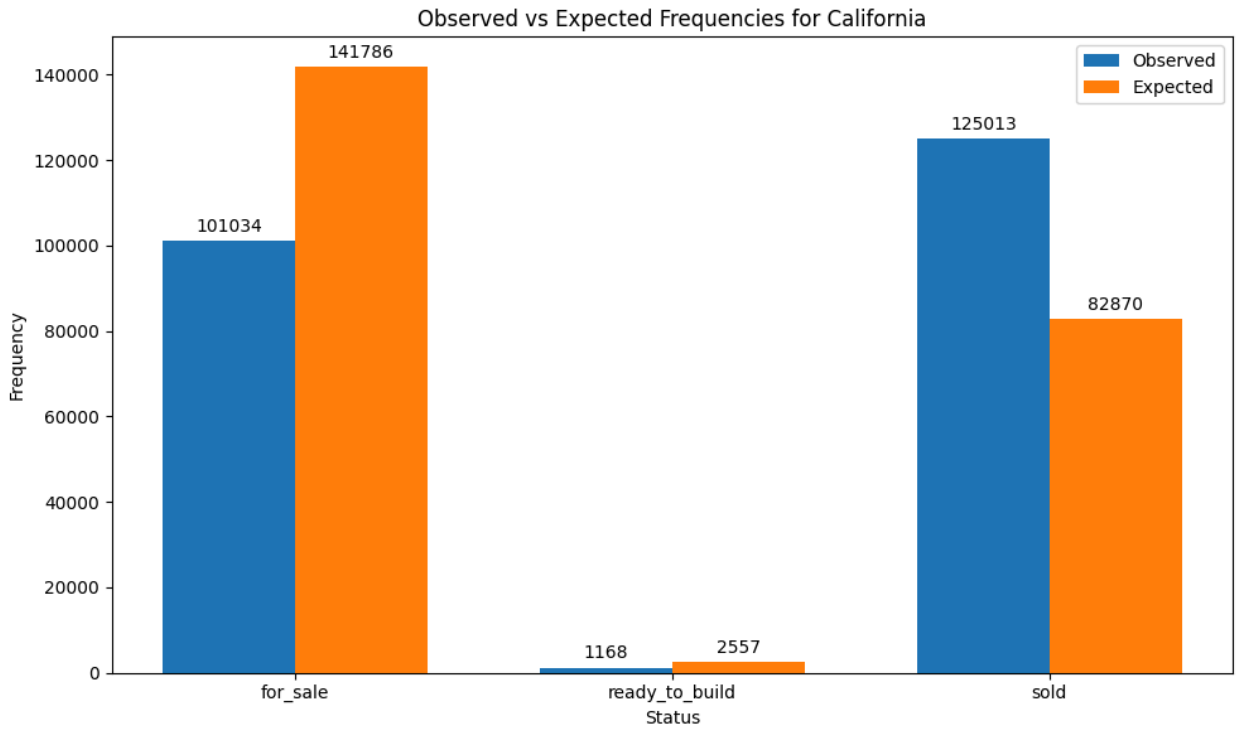
- **For Sale**: Observed (71,378) higher than expected (64,373.38).
- **Ready to Build**: Observed (22) much lower than expected (1,161.20).
- **Sold**: Observed (31,759) lower than expected (37,624.42).
- **Conclusion**: Stagnant market with few new developments and lower turnover.

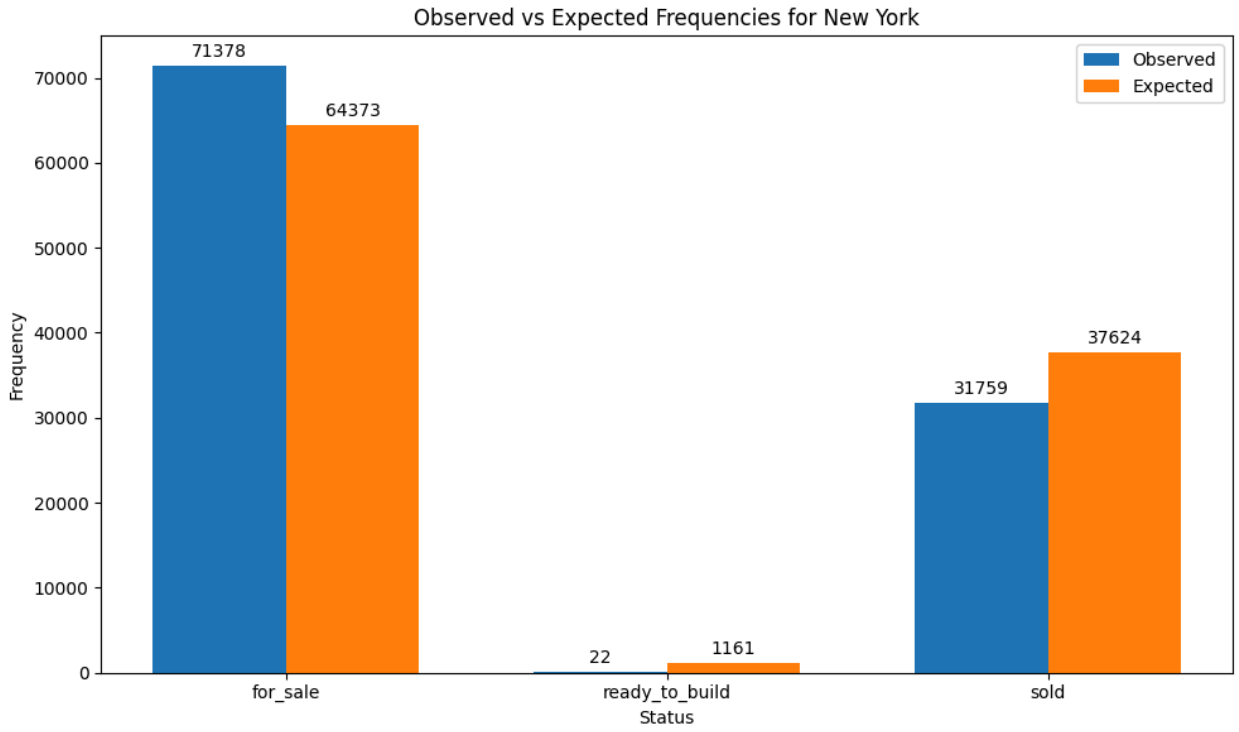
Texas

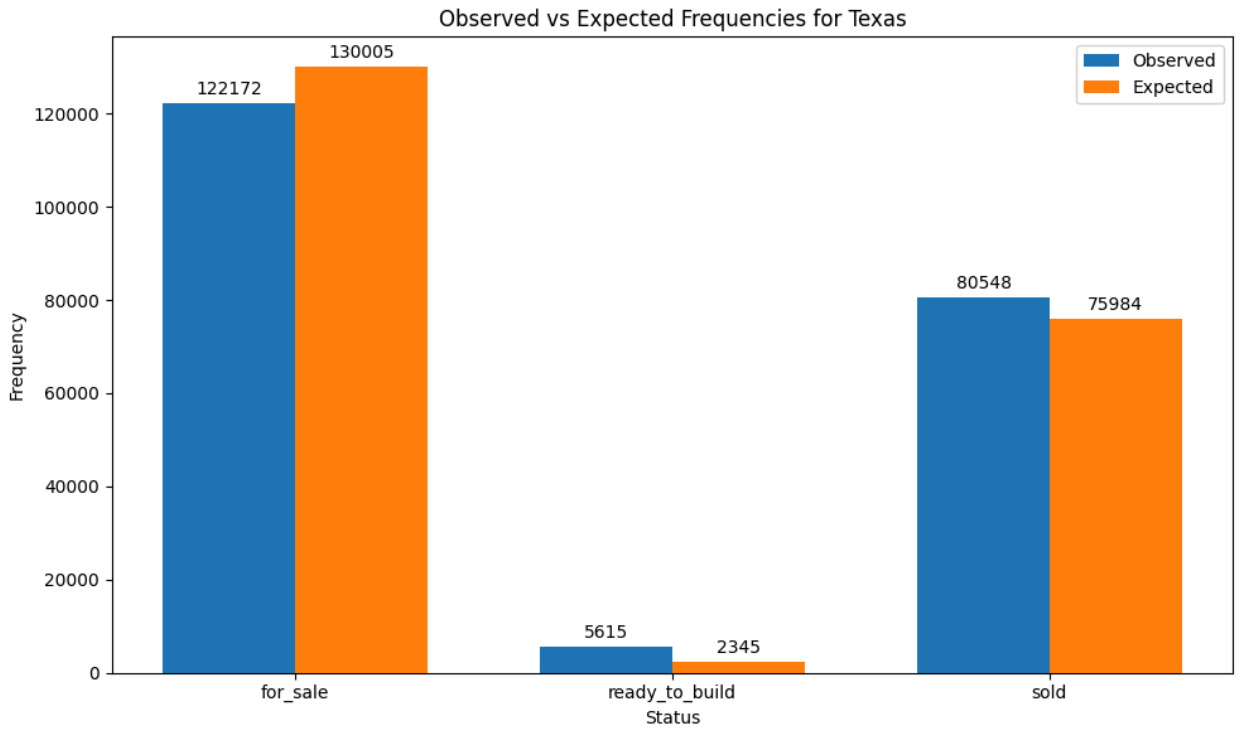
- **For Sale**: Observed (122,172) slightly lower than expected (130,005.40).
- **Ready to Build**: Observed (5,615) significantly higher than expected (2,345.11).
- **Sold**: Observed (80,548) higher than expected (75,984.49).
- **Conclusion**: Dynamic market with substantial new construction activity and high turnover.

The Chi-Square test reveals a significant association between housing status and state. Key insights include:

- **California**: High market turnover with fewer new listings and constructions.
- **New York**: Stagnant market with fewer new developments and lower turnover.
- **Texas**: Dynamic market with substantial new construction activity and high turnover.


Observed vs Expected Frequencies for California
Observed
Expected
Frequency
Status
for_sale
ready_to_build
sold
101034
141786
1168
2557
125013
82870
0
20000
40000
60000
80000
100000
120000
140000


Observed vs Expected Frequencies for New York
Observed
Expected
Frequency
Status
for_sale
ready_to_build
sold
71378
64373
22
1161
31759
37624
0
10000
20000
30000
40000
50000
60000
70000



3. Chi-square test

Objective: Is there a significant difference in housing prices among houses that are 'for_sale', 'ready_to_build', and 'sold'?

ANOVA Test Results:

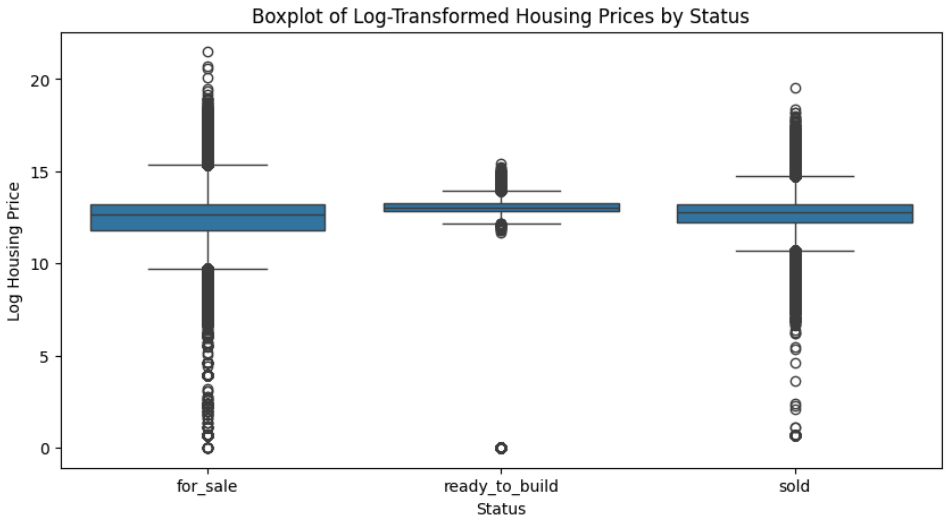
- F-statistic: 227.7586956494143
- P-value: 1.2467513401901643e-99

Given the extremely small p-value, which is much less than the significance level of 0.05, the result is statistically significant. Therefore, we reject the null hypothesis and conclude that there is a significant difference in housing prices among the 'for_sale', 'ready_to_build', and 'sold' groups.

Boxplot:

The boxplot of log-transformed housing prices shows the following:

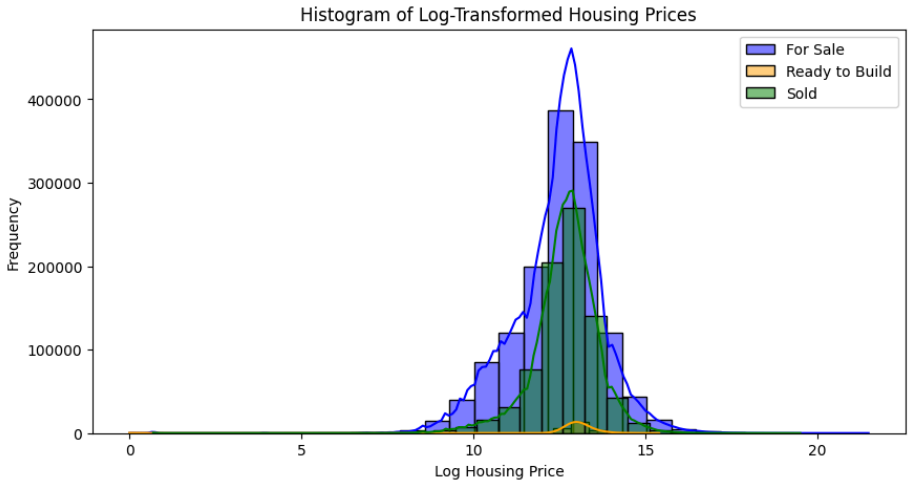
- Houses that are "for_sale" have a wider range of prices and more outliers compared to "ready_to_build" and "sold" groups.
- The median price for "for_sale" houses is slightly higher than that for "sold" houses, with "ready_to_build" houses showing the lowest median price.
- The presence of outliers in all three categories indicates variability in housing prices, especially for the "for_sale" category.



Histogram:

The histogram of log-transformed housing prices indicates:

- The distribution of housing prices for "for_sale" houses is broader, showing a higher frequency of higher prices compared to "ready_to_build" and "sold" houses.
- "Ready_to_build" houses have a narrower distribution, concentrated around a specific price range, indicating less variability in their prices.
- "Sold" houses also show a wide distribution, but with fewer extreme high prices compared to "for_sale" houses.



Density Plot:

The density plot of log-transformed housing prices reveals:

- A sharp peak for "ready_to_build" houses, indicating that most of these houses are priced within a narrow range.
- "For_sale" houses have a broader peak, suggesting more variability in pricing.

- "Sold" houses have a distribution similar to "for_sale" houses but with a slightly lower peak, indicating less concentration around a specific price point.

Insights and Observations

1. **Significant Price Differences**: The ANOVA test confirms that there is a statistically significant difference in housing prices among the three status groups. This suggests that the market values houses differently based on their status.
2. **Price Variability**: The "for_sale" category exhibits the highest variability in prices, with a broader range and more outliers. This could be due to the diverse nature of houses on the market, ranging from affordable to luxury properties.
3. **Narrow Price Range for "Ready to Build"**: Houses categorized as "ready_to_build" have a more concentrated price range. This might indicate that these properties are more standardized or have less variation in features and sizes compared to "for_sale" houses.
4. **Market Trends**: The "sold" houses' price distribution suggests that the final selling prices of houses tend to be lower and less variable than the listing prices ("for_sale"). This could reflect market negotiations and adjustments based on buyer demand and market conditions.
5. **Strategic Insights**: For real estate professionals, understanding these price dynamics can help in pricing strategies, marketing efforts, and targeting potential buyers. Sellers of "for_sale" houses might need to consider competitive pricing and highlighting unique features to stand out in a variable market.

4. T-test

Objective: Is there a significant difference in housing prices between houses that are 'for_sale' and houses that are 'ready_to_build'?

T-Test Results:

- **T-statistic**: 2.2982664540388513
- **P-value**: 0.02154677378302286

The p-value obtained from the t-test is 0.0215, which is less than the common significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a statistically significant difference in housing prices between "for_sale" and "ready_to_build".

Implications:

- The data provides evidence that housing prices differ significantly between the two statuses.
- This indicates that the market values houses listed as "ready_to_build" differently compared to those listed as "for_sale".

5. Linear Regression

Objective: Relationship Between Features and Price

1. Relationship Between Features and Price

The coefficients of the linear regression model provide insight into how each feature affects housing prices:

- **bath**:Each additional bathroom increases the price by approximately $382600.
- **bed**: Each additional bedroom decreases the price by approximately $74,397.
- **acre_lot**: Each additional acre has a negligible positive impact on the price.
- **house_size**: Each additional square foot increases the price by approximately $15.96.
- **zip_code**: Changes in zip code (treated as a categorical variable) have a very small positive impact on the price.

2. Model Performance

The performance metrics indicate how well the model fits the data:

- **R-squared**: The low R-squared values (0.15 for training and 0.19 for testing) indicate that the model does not explain much of the variance in housing prices. This suggests that either the relationships between the features and price are not linear or that additional relevant features are needed.

- **Mean Squared Error (MSE)**: The high MSE values indicate that there is a significant difference between the predicted prices and actual prices, further confirming that the model does not fit the data well.

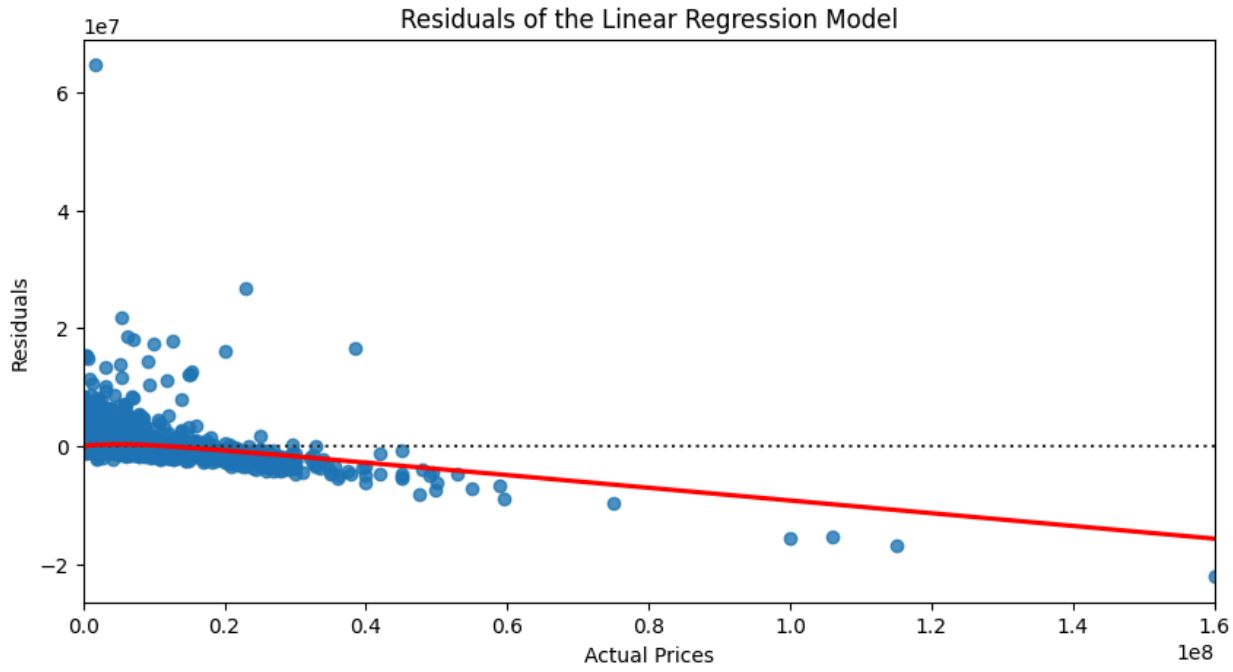
3. Evaluation Metrics Interpretation

- **Training Metrics**: The high MSE and low R-squared value on the training set suggest that the model does not fit the training data well.
- **Test Metrics**: The slightly better performance on the test set suggests that the model has not overfitted to the training data, but the overall performance is still poor, indicating that the features used may not be the best predictors of housing prices or that the relationships are not linear.

4. Visualization Insights

The residuals plot shows the difference between the actual and predicted prices:

- The residuals are not randomly distributed, and there is a clear pattern.
- This pattern indicates that the model may be missing some key features or that the relationship between the features and the target variable is not purely linear.



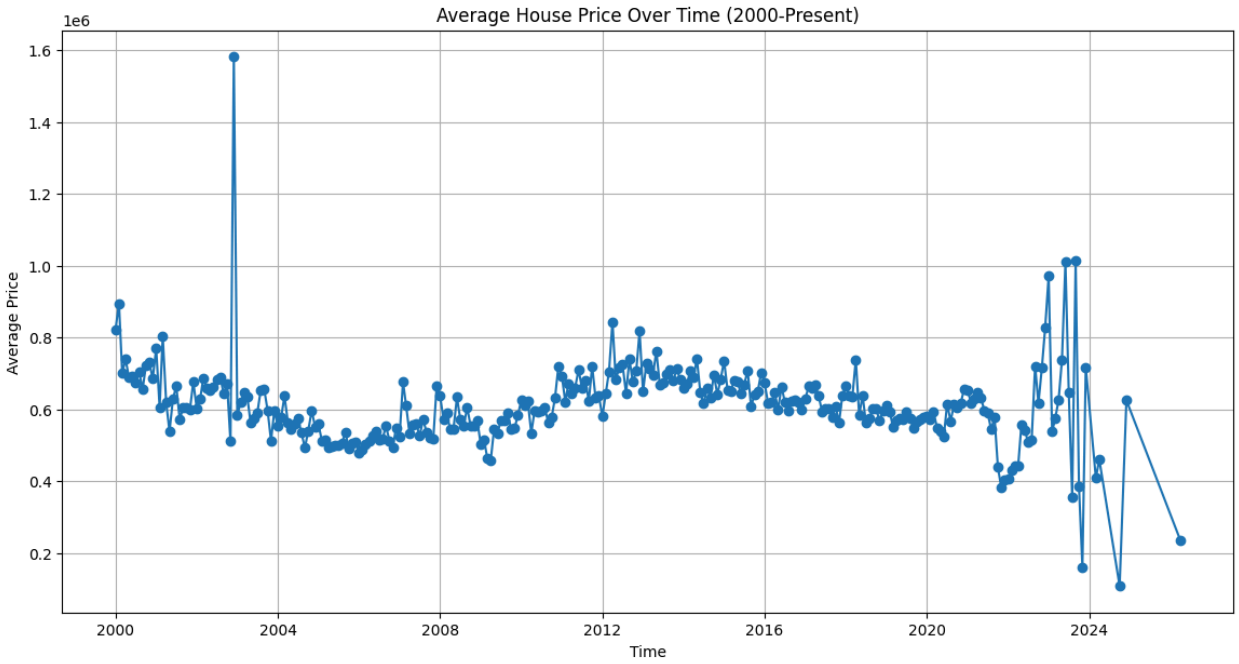
The scatter plot of actual vs predicted prices:

- Most of the points are clustered at the lower end of the price spectrum, indicating that the model struggles to predict higher-priced houses.
- The diagonal red line represents perfect predictions. The fact that many points deviate significantly from this line further indicates that the model's predictions are not very accurate.

6. Time Series Analysis

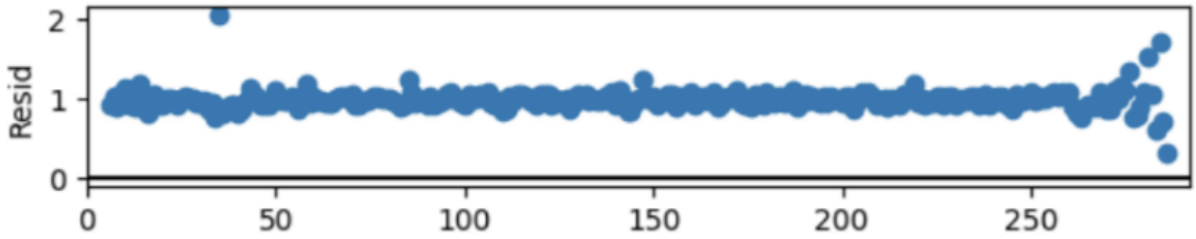
1. What trends can be observed in house prices over time?

- **Initial Spike**: There is an initial spike around 2000-2004. This could be due to outlier transactions or market conditions that need further investigation.
- **General Trend**: Post-2004, house prices show a general decrease until around 2012.
- **Recovery and Stabilization**: From 2012 to 2020, there is a gradual increase and stabilization in house prices.
- **Recent Volatility**: There is noticeable volatility and some spikes in prices around 2022-2024, likely influenced by recent market conditions, economic factors, or specific high-value transactions.



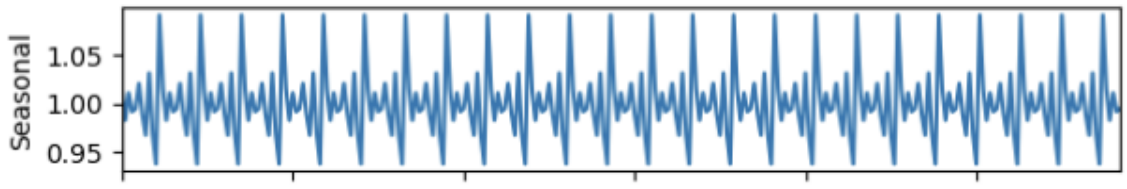
2. Are there any significant outliers in historical house prices?

- **Significant Outliers**: There are a few significant outliers, particularly around the initial spike in the early 2000s and some points towards the recent years.
- **Influence on Analysis**: These outliers can have a substantial influence on the average price trends and might need to be investigated or mitigated to ensure a robust analysis.



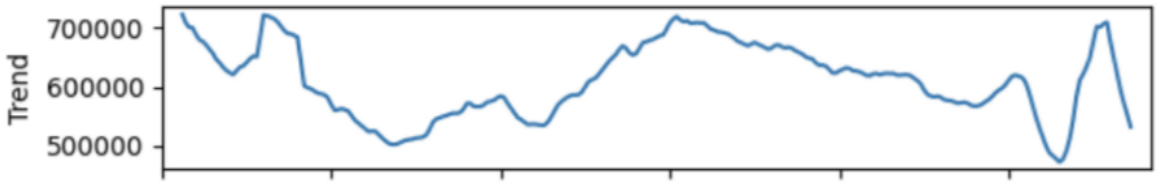
3. What seasonal patterns can be observed in house prices?

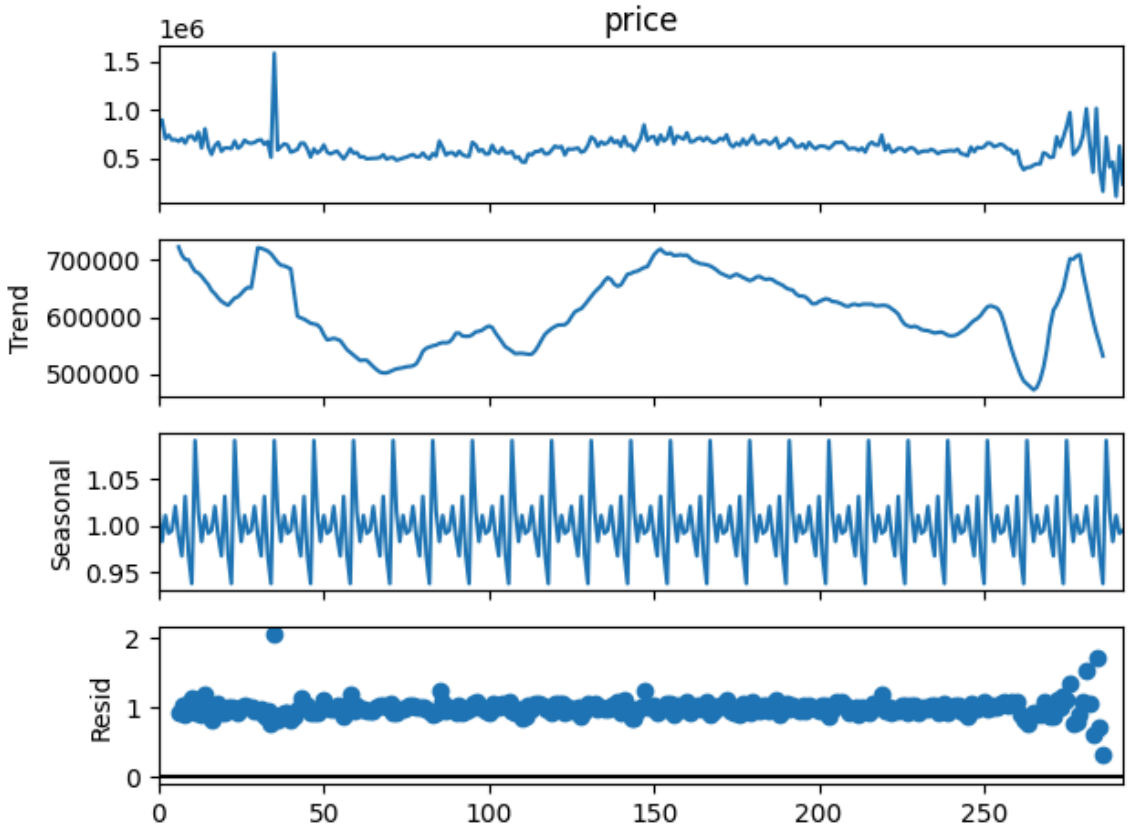
- **Regular Seasonal Pattern**: There is a clear seasonal pattern in house prices, indicating regular fluctuations within each year.
- **Amplitude**: The amplitude of these seasonal fluctuations appears consistent, suggesting regular cyclical variations in house prices, possibly driven by seasonal demand, economic cycles, or other factors.



4. How does the trend component explain the changes in house prices?

- **Initial Decrease**: The trend shows an initial decrease in house prices from 2000 to around 2012.
- **Subsequent Increase**: After 2012, there is a gradual increase in the trend component, indicating a recovery in house prices.
- **Recent Trends**: The trend component shows some volatility in recent years, reflecting current market conditions and economic factors affecting house prices.





Conclusions:

With this project, we analyzed comprehensive real estate listings data to further understand the intricacies that make up the real estate market in the United States. The high-level analysis being done within this project includes, but is not limited to, clustering, Chi-Square tests, ANOVA, T-tests, linear regression, and time series analysis. It gave some peculiar insights from each of these, hence very helpful in understanding different facets of the housing market.

Key Learnings

- Data Variability and Outliers
  - The dataset had a high level of variability, especially in housing prices and property sizes. Extreme values may denote the presence of luxury properties or may result from misspecification of the data. It is really important to treat such types of outliers for models to be estimated properly and analyze better.

- Correlation Analysis:
  For instance, lot size or gross living area correlates only weakly with price; this reflects a need to offer an approach that would actually understand what drives housing prices more insightfully. Moderate correlation of bedrooms and bathrooms is rather typical for the real estate business.

- Geographical Insights:
  The regional diversity of the listing activity of real estate was exposed by remarkable distribution across states. In other words, states like Florida, Texas, and New York register high volumes of listings due to larger and more dynamic real estate markets.

- Cluster Analysis:
  The clustering analysis showed five clearly distinguishable segments within the housing market, all segmented all the way from middle-level homes to ultra-luxury properties. Such segmentation provides insight into value for targeted marketing and investment strategies.

- Association of Housing Status with State:
  That association between housing status and state was significant according to the Chi-Square, suggesting that the differences in the dynamics of the housing market were substantial across regions.

- Price Differentials across Housing Status:
  ANOVA shows that housing prices by status (for sale, ready to build, sold) are statistically different; the T-tests provide a view on market value and buyer preferences.

- Predictive Modeling:

For instance, linear regressions using models showed that some features, such as the number of bedrooms and bathrooms, had a mere slight influence on prices. The overall predictability was moderate, although it suggested other factors were at play.

- Trends over Time
  Time-series analysis has shown the increasing trend in house prices, which take an upward direction over years and periodic fluctuations correlated with economic cycles and changes in policies.

Discussion:

This project further underlined the complexity of the real estate market and showed that it is influenced by many different interrelated factors other than traditional ones like a lot size and living area. Such important correlations with price were established, but effective property value is probably affected by lots of important variables not actively registered in the dataset. All the data are required to be thoroughly taken for a consistent study of real estate information. The above test of Chi-Square and cluster analysis brought out significant regional differences, clearly highlighting that the saying, "real estate is localization," has a very important meaning.

The strategies in the market should embody the particularized nature of the regions in order to adequately respond to the demands of the local market. Both ANOVA and T-tests relay classificatory informative value on how the various housing statuses are valued, with critical importance in informing real estate developers and investors about significance in making decisions about the development and sale of property. This is an important piece of output that comes from the above type of research, in the manner of producing some key insights with regard to the USA real estate market, its applicable trends, associations, and market segments or divisions, very valuable in enabling stakeholders for better-specified markets and sound decisions. This project was able to illustrate the power of data analysis in digging deep into the insights regarding market dynamics and pointing out the areas upon which a line of future research and data collection could be in order to enhance understanding of the real estate markets.