USA Real Estate Market Analysis

Team Members:

Abhiram Ravipati (A20539084)

Sumanth Kalyan Bandigupthapu (A20544342)

Abstract:

The research aimed to analyze the real estate market trends and factors affecting property prices using a dataset of real estate listings. The study focused on exploring patterns in property prices across different cities and states, identifying key factors influencing price variations, and developing predictive models to forecast property prices. The research findings provide valuable insights into the dynamics of the real estate market, aiding stakeholders in making informed decisions.

Overview:

The problem statement addressed in this research is understanding the determinants of property prices in the real estate market. With the increasing complexity of the real estate landscape, it's crucial to analyze various factors influencing property prices to provide accurate insights for buyers, sellers, and investors. Relevant literature was reviewed to understand existing research on real estate market analysis, including factors such as location, property characteristics, and market demand.

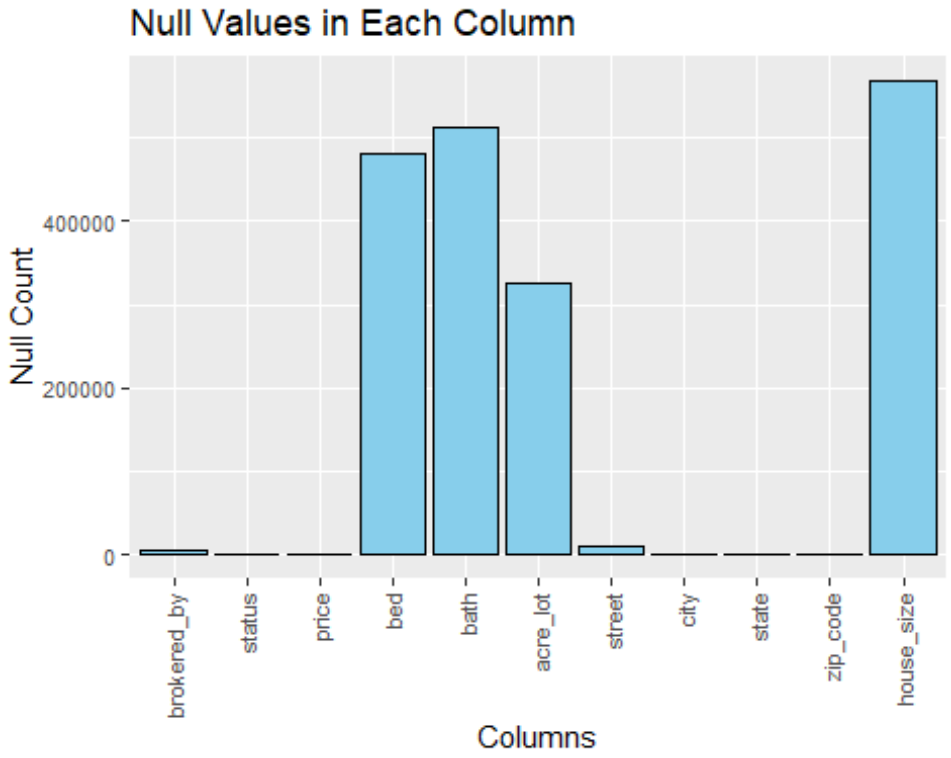
The proposed methodology involved exploratory data analysis (EDA) to uncover patterns and trends in the dataset, followed by feature engineering to identify significant predictors of property prices. Regression analysis and machine learning techniques were employed to develop predictive models for estimating property prices. The research methodology aimed to provide a comprehensive understanding of the real estate market dynamics and develop robust models for price prediction.

Methodology:

Data Cleaning and Processing:

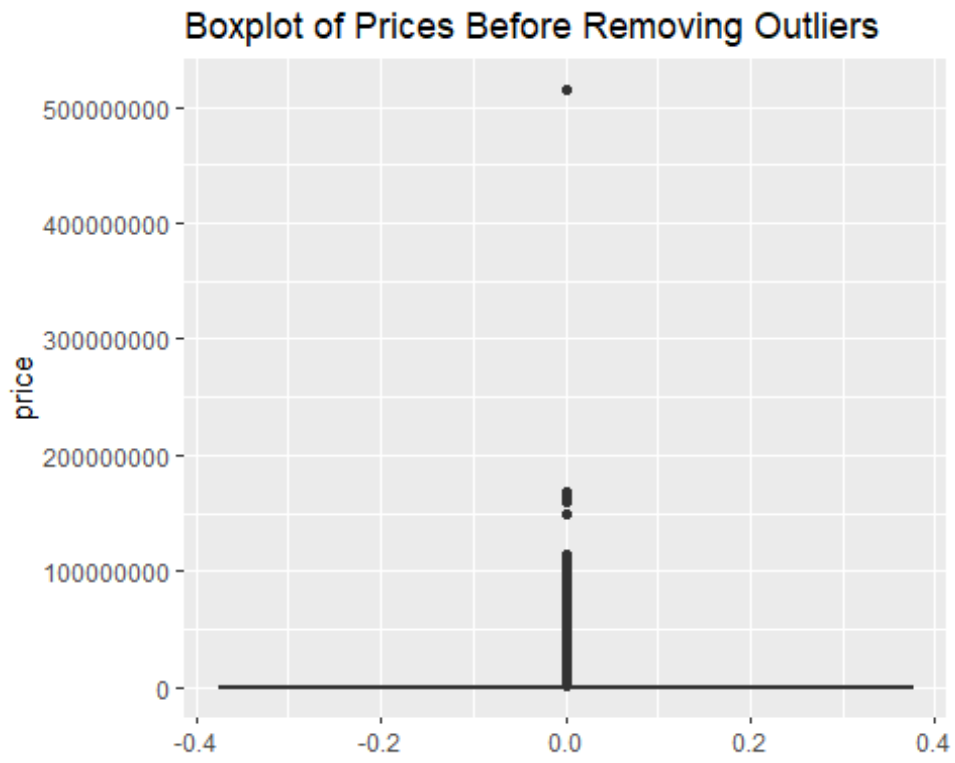
In the initial stages of data processing, the dataset underwent several steps to ensure its quality and suitability for analysis. One crucial step involved the removal of the "prev_sold_date" column due to its high proportion of null values. Retaining columns with significant missing data could compromise the integrity and reliability of subsequent analyses. By dropping this column, we aimed to enhance the quality of our dataset and facilitate more robust analysis.

Following the removal of the "prev_sold_date" column, the dataset underwent further cleaning procedures to address missing values. Null values were systematically identified and subsequently removed from the dataset. This step was essential to maintain data consistency and accuracy throughout the analysis process. Removing null values ensured that the subsequent analyses were conducted on complete and reliable data, minimizing the potential for biased results or inaccurate conclusions.



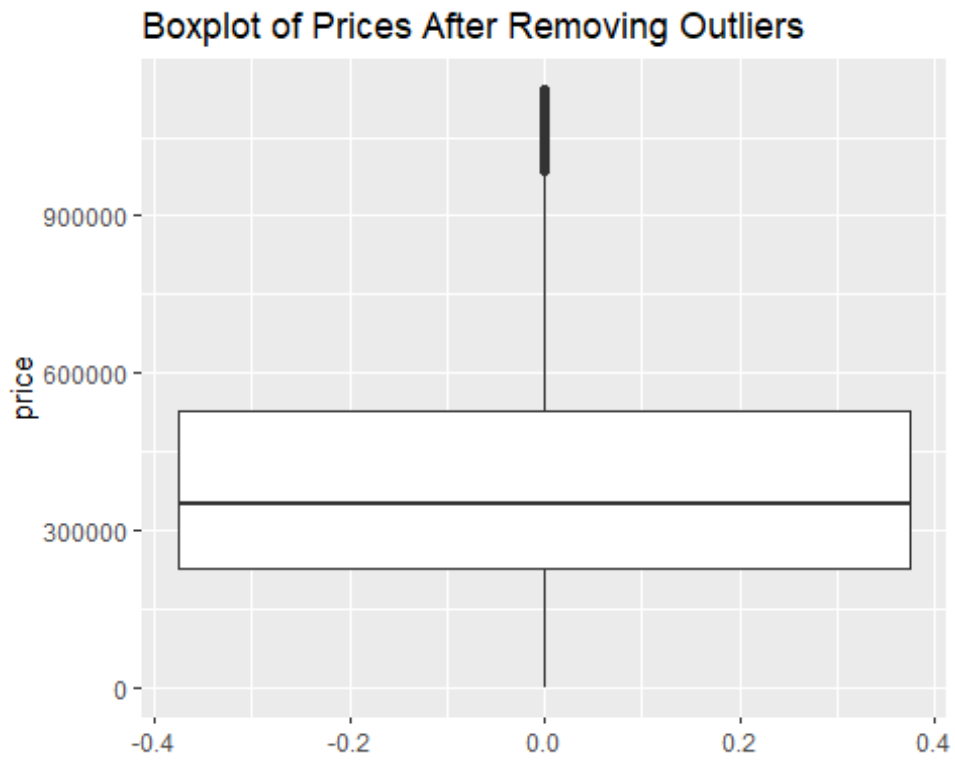
The bar chart above displays the number of null values in each column of a dataset. Columns such as "bed", "bath", "acre_lot", and "house_size" have significant null counts, indicating that these attributes are frequently missing or unrecorded. Specifically, the "house_size" column contains the highest count of null values, while columns like "status", "price", "brokered_by", "city", and "state" have minimal or no null values. This visualization helps identify which data columns require cleaning or imputation to maintain data integrity.

Additionally, outlier cleansing was performed to identify and address any extreme or anomalous values within the dataset. Outliers have the potential to distort statistical analyses and modeling outcomes, leading to erroneous conclusions. Through outlier cleansing techniques, such as identifying values beyond a certain threshold or using statistical methods like z-scores, we aimed to mitigate the influence of outliers on our analysis and ensure the robustness of our findings.



This boxplot illustrates the distribution of house prices before removing outliers. It shows a highly skewed distribution, with several extreme values far beyond the main concentration of data.

The majority of the data points lie within a narrow range near the lower end of the price scale, indicated by the thick line and small box (the interquartile range). The upper whisker stretches upwards, highlighting the presence of numerous outliers. These outliers represent significantly high prices, which distort the central tendency and spread of the data. Removing these outliers can result in a clearer and more accurate understanding of the general pricing trends.

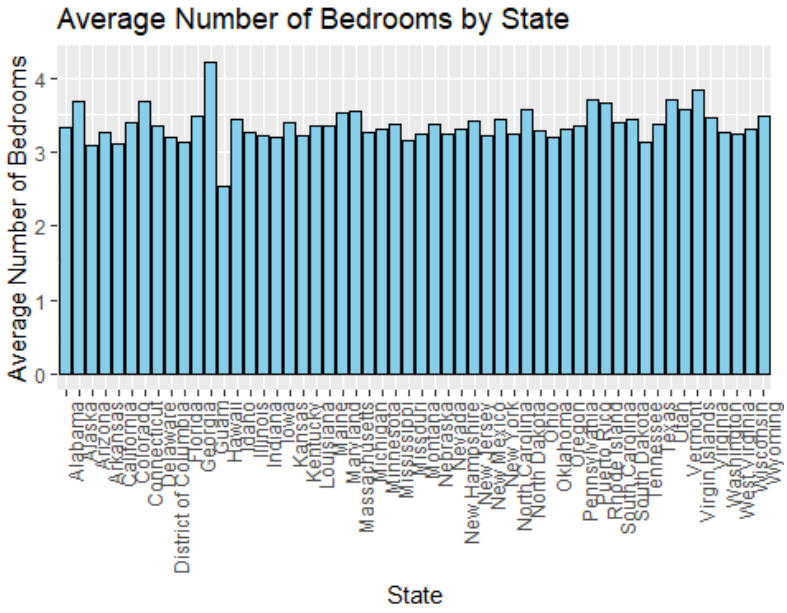


This boxplot displays the distribution of house prices after removing outliers, offering a more focused view of the central tendency and spread of prices. The box represents the interquartile range (IQR), where the middle 50% of the data resides, with the thick line indicating the median. The whiskers extend to the smallest and largest values within 1.5 times the IQR, showing a more compact data range compared to the previous plot.

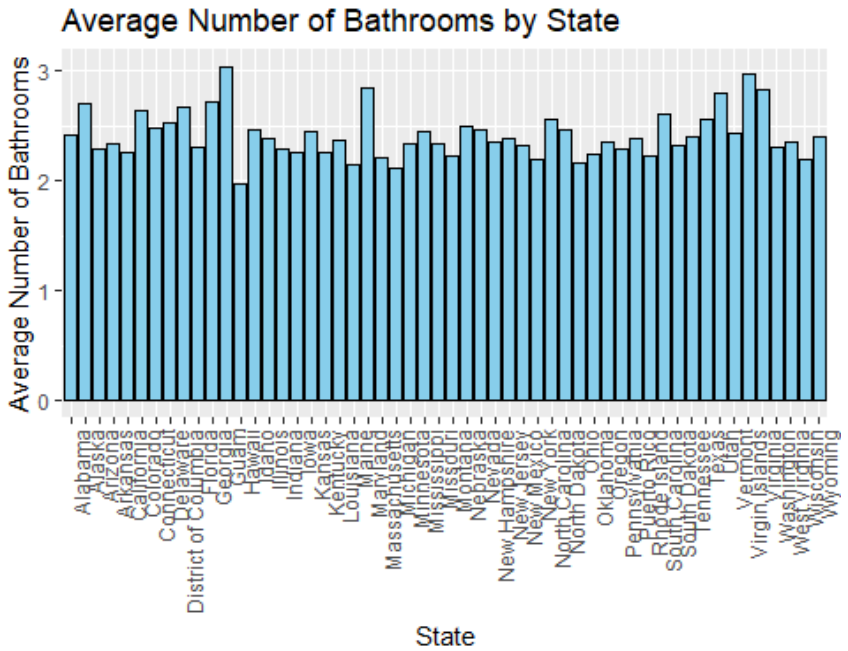
Without extreme outliers, the data appears more symmetric and easier to analyze. This approach can help identify trends and make accurate comparisons without the influence of extreme values.

Overall, the data processing and cleaning procedures employed in this study were essential for preparing the dataset for rigorous analysis. By systematically addressing missing values and outliers, we enhanced the quality and reliability of the data, laying the groundwork for insightful and accurate analyses of real estate market trends and factors affecting property prices.

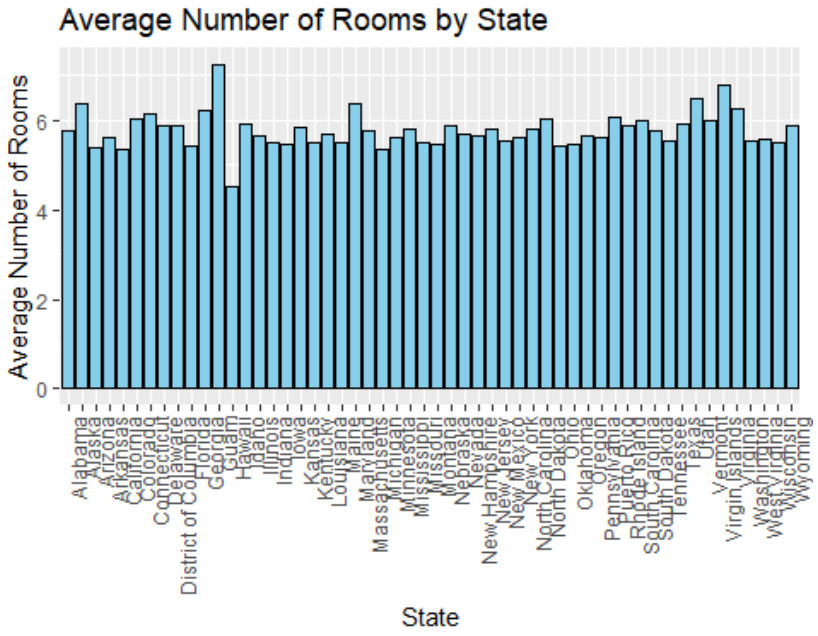
Exploratory Data Analysis (EDA):



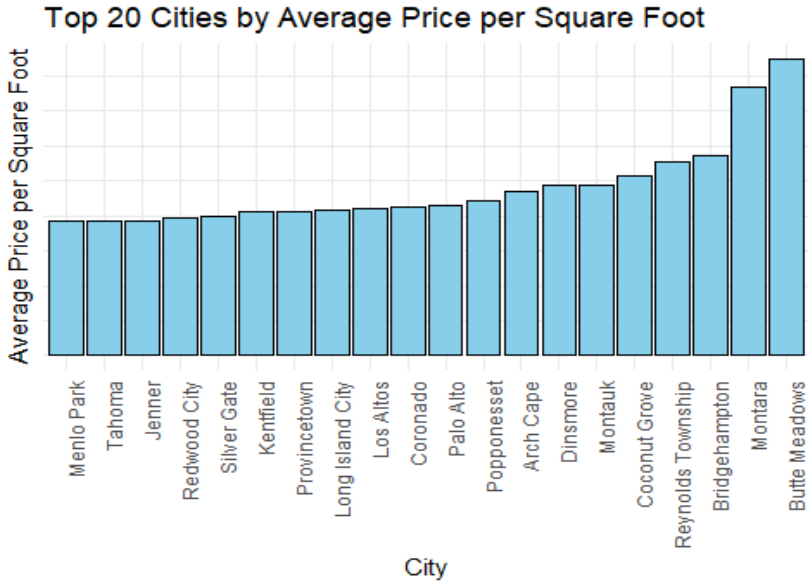
This bar chart illustrates the average number of bedrooms by state in the United States. Most states have an average of around 3 bedrooms, with some states having slightly higher or lower averages. States like Alabama and District of Columbia stand out with higher averages, while others like Mississippi and Arkansas are on the lower end.



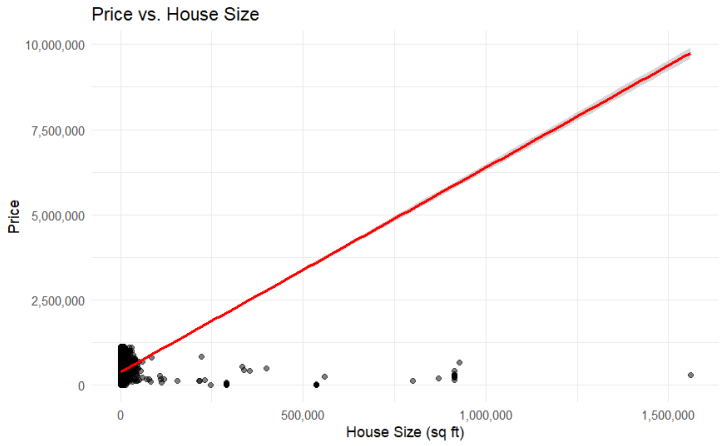
This bar chart represents the average number of bathrooms by state across the United States. Most states have an average of around 2 to 3 bathrooms, with some showing slight variations. States like Alabama and Florida tend to have a higher average, suggesting larger or more luxurious homes, while others, such as Mississippi and Arkansas, may have a slightly lower average.



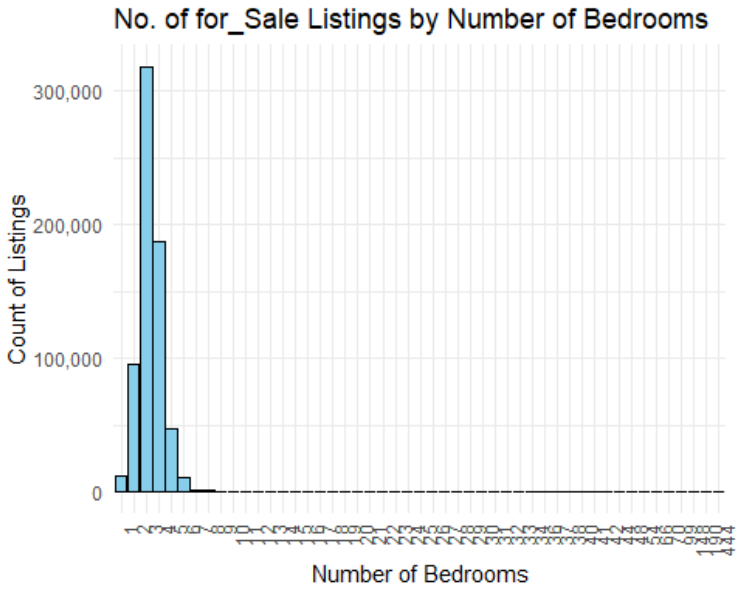
This bar chart shows the average number of rooms by state in the United States. Rooms indicate combination of bedrooms and bathrooms. Most states have an average of about 5 to 6 rooms, with some variations. States like Alabama and Florida have higher averages, while others like Arkansas and Mississippi have lower averages. This pattern can reflect different housing trends across the country, with some states tending to have larger homes.



In this bar chart, you can see which cities top the list in terms of real estate prices. Menlo Park, Tahoma, and Jenner start at the lower end of the spectrum, indicating they're among the more affordable cities within the top 20. As you move along, prices increase, with Montara and Butte Meadows standing out as the most expensive cities. The variance in price across these cities suggests differences in factors like location desirability, local amenities, or housing demand.

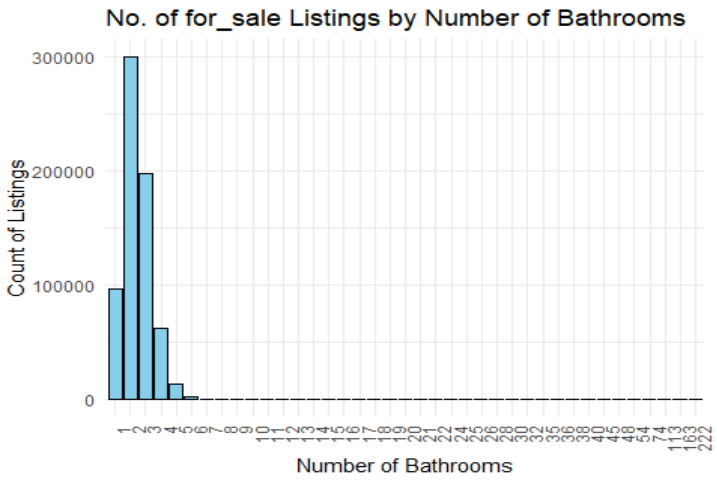


This scatter plot illustrates the relationship between house size (measured in square feet) and house price. The red line is a trend line indicating the general pattern or direction of the relationship. The plot suggests a positive correlation between house size and price. As the size of the house increases, the price tends to rise as well, with the trend line following an upward slope. However, you can also see a lot of variability in the smaller house sizes, with some points clustered near the origin, indicating that smaller houses can still command a range of prices.

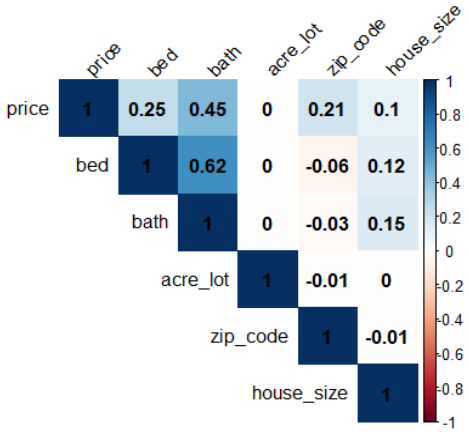


This histogram represents the distribution of for_sale real estate listings by the number of bedrooms. The histogram has a clear peak, indicating that most listings have a typical number of

bedrooms, probably around 3 to 4. As the number of bedrooms increases, the count of listings drops sharply, suggesting that larger homes with more bedrooms are less common in the market. This trend is not surprising, as most households prefer a moderate number of bedrooms, while homes with more bedrooms are generally larger and rarer. The rapid decline in listing counts as bedrooms increase implies that homes with a high bedroom count are either not frequently built or are less likely to be put on the market.



This histogram displays the distribution of for_sale real estate listings by the number of bathrooms. The x-axis represents the number of bathrooms, while the y-axis shows the count of listings that have that specific number of bathrooms. The chart has a significant peak, indicating that the majority of listings have between 2 and 3 bathrooms. As the number of bathrooms increases, the count of listings decreases steeply, suggesting that properties with a higher number of bathrooms are less frequent in the real estate market. This distribution reflects common housing patterns, with most households having a typical number of bathrooms for daily use. Properties with a larger count of bathrooms tend to be larger homes or luxury listings, which are not as common as regular family homes.



This correlation matrix plot shows the relationships between various real estate features. The matrix helps understand how different features are correlated with each other, with positive values indicating a positive correlation, and negative values indicating a negative correlation.

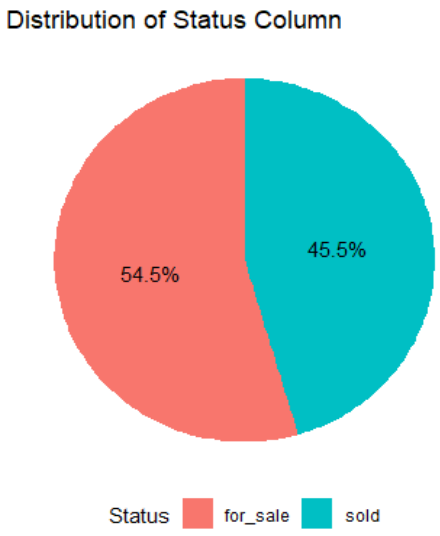
The key relationships to note are:

Price and Other Features: Price has the highest positive correlation with the number of bathrooms (0.45), suggesting that homes with more bathrooms tend to be more expensive. There's also a moderate correlation between price and the number of bedrooms (0.25), indicating that larger homes with more bedrooms tend to have higher prices.

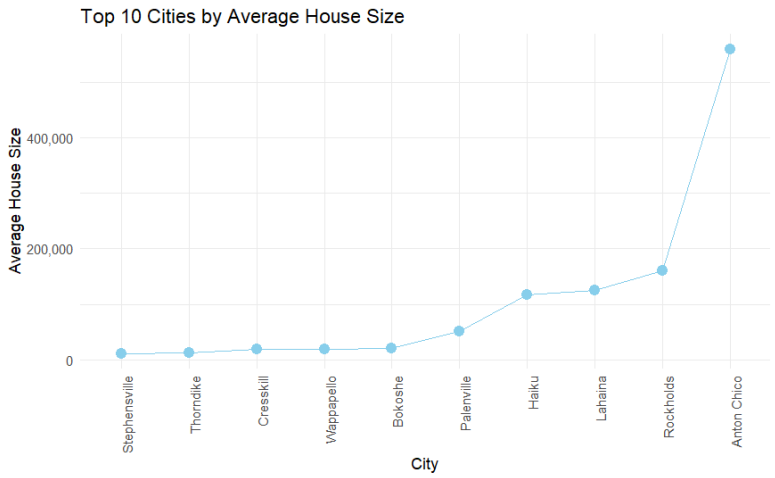
Bedrooms and Bathrooms: The correlation between the number of bedrooms and bathrooms is quite high (0.62). This relationship indicates that as the number of bedrooms increases, the number of bathrooms also tends to increase.

House Size Correlation: House size has the lowest correlation with other features, suggesting it doesn't strongly depend on them.

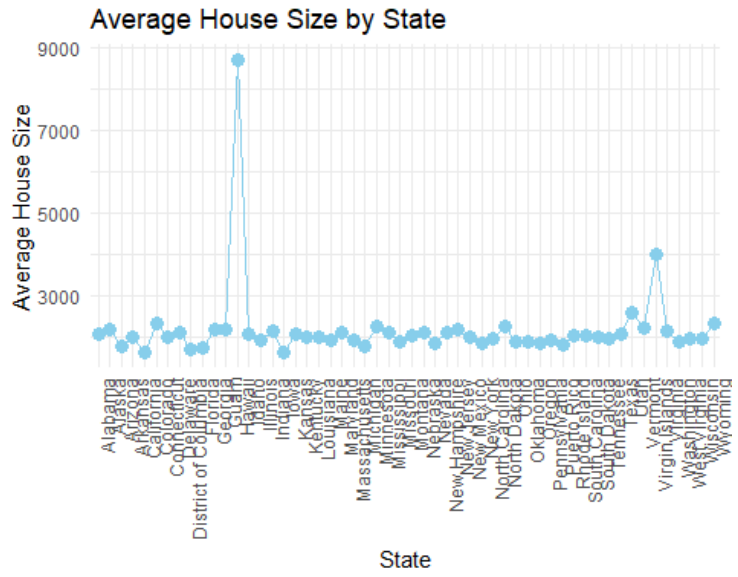
Other Features: The correlations between features like zip code and acre lot with other attributes are generally lower, indicating less dependence.



This pie chart shows the distribution of the "status" column in a dataset, with two categories: "for_sale" and "sold." The chart divides the circle into segments to represent the proportion of each category. In this case, the segment for "for_sale" makes up 54.5% of the pie, indicating that more than half of the data consists of properties currently on the market. The "sold" segment accounts for 45.5%, representing the proportion of properties that have already been sold.

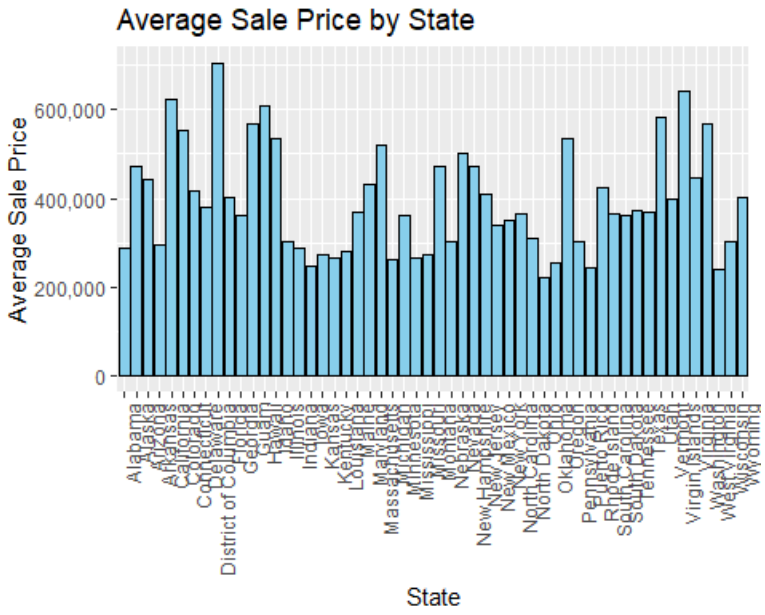


This line plot shows the top 10 cities by average house size. The cities are arranged from smallest to largest in terms of average house size. The plot demonstrates a sharp increase in average house size as you move from left to right. The first few cities, like Stephenville and Thorndike, have relatively smaller average house sizes, indicating smaller or more compact homes. However, there's a significant jump as you reach the last few cities on the right. Rockholds and Anton Chico show much larger average house sizes, suggesting that these cities have significantly larger homes or properties.

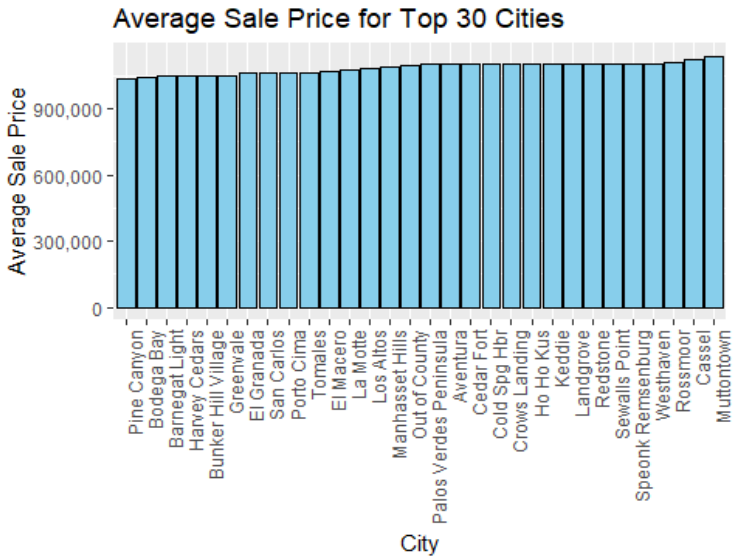


This line plot represents the average house size by state. The x-axis displays the names of all 50 states, while the y-axis indicates the average size of houses in square feet. Each point on the plot represents a state, showing the general distribution of average house sizes across the United States. A notable feature of this plot is the clear spike in average house size for one state, likely due to a few very large properties. The majority of states cluster along the lower end, with average house sizes ranging from about 1,000 to 3,000 square feet, suggesting that most homes

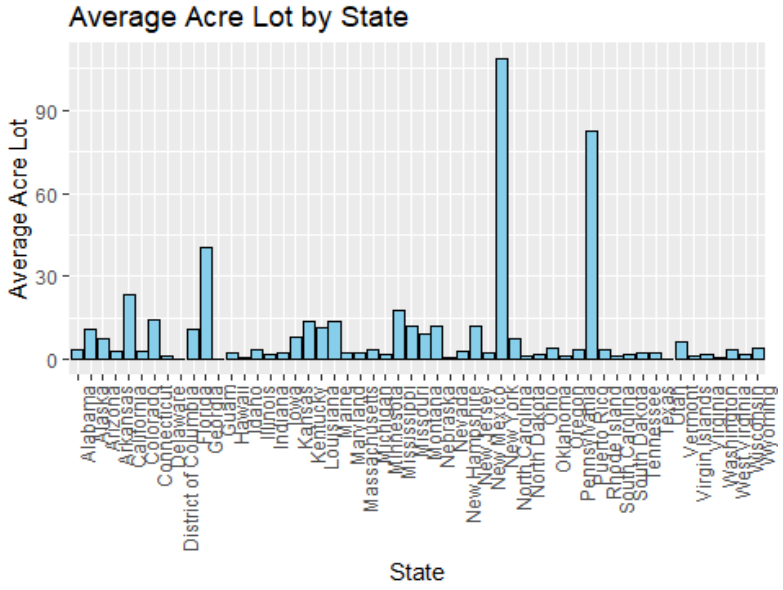
in these states fall within a typical range. This spike indicates an unusually high average house size, which can be due to a unique housing market in that state.



This bar chart shows the average sale price of homes by state in the United States. Each bar represents a state, with the height indicating the average sale price of homes in that state. The chart provides a quick comparison of real estate pricing trends across different states. The chart reveals some notable variations in average sale prices. States like California and New York, typically known for their high real estate costs, have bars reaching the upper range, indicating higher average sale prices. On the other hand, states such as Mississippi and West Virginia show lower average sale prices, suggesting a more affordable housing market. The overall pattern indicates a diverse range of average sale prices across the states. This distribution reflects factors like regional cost of living, housing market dynamics, and state-specific economic conditions.

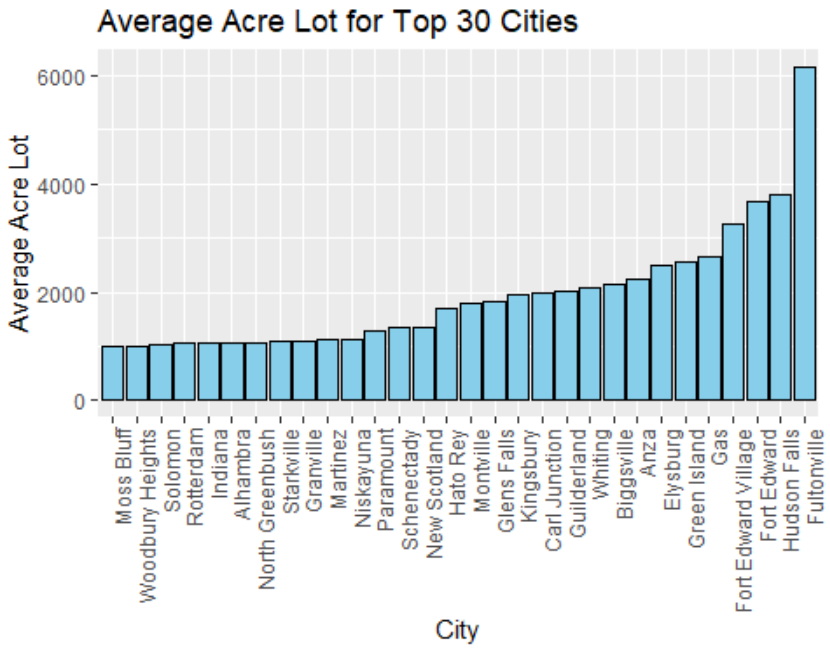


This bar chart displays the average sale price for the top 30 cities. The cities are arranged from left to right in order of increasing average sale price. The bars show a gradual rise in average sale prices across these top 30 cities, indicating that the cities on the right tend to have higher real estate values compared to those on the left. In this chart, we can see a general pattern of increasing prices from Pine Canyon to Muttontown, suggesting that cities on the left have more affordable real estate, while those on the right represent higher-end markets. The spread of prices reflects the diversity in real estate values, influenced by factors such as location, demand, and housing market conditions.

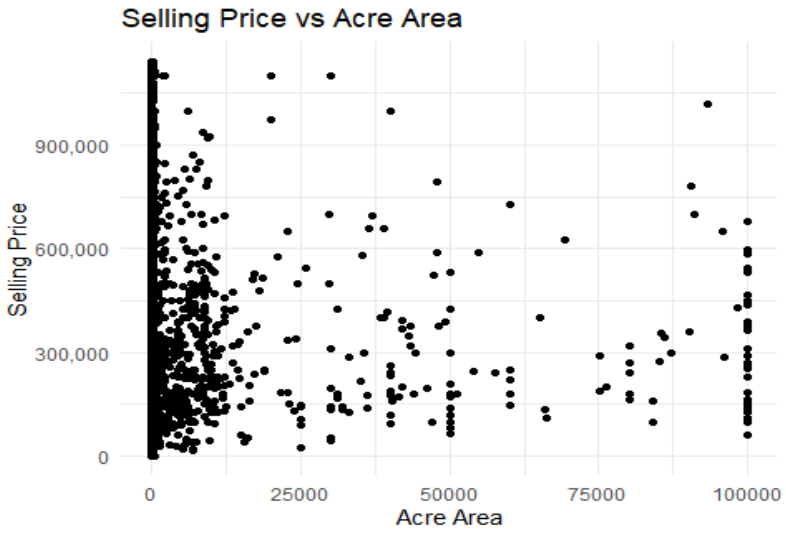


This bar chart represents the average size of acre lots by state. In this chart, the average lot size varies significantly across states. Some states, like Vermont and North Dakota, have much larger

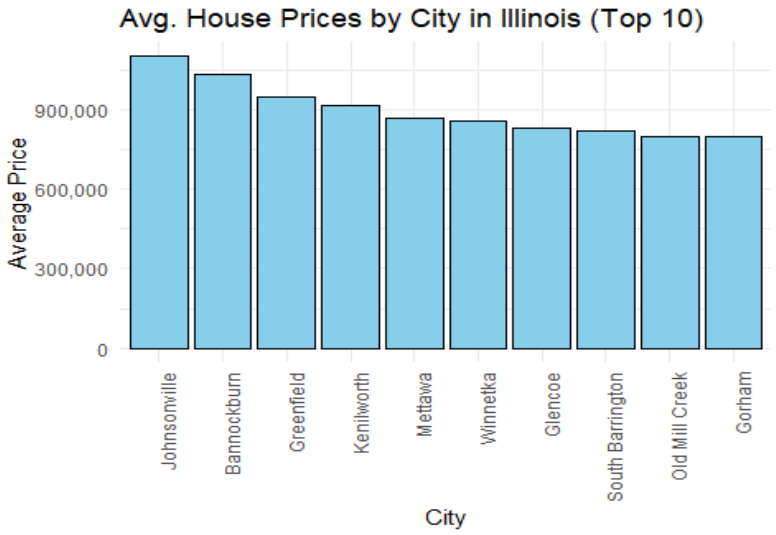
average acre lots, suggesting that properties in these regions are on more expansive land. This could be due to lower population density or a different approach to property development. Most states cluster around the lower end of the scale, with smaller average acre lots. This distribution might reflect denser urban environments, smaller property sizes, or varying land use patterns. The high variability in average acre lot sizes among states could be due to regional differences in housing markets, zoning regulations, or geographic factors. The chart provides an interesting overview of how property sizes differ across the United States and can guide further analysis into why these differences exist.



This bar chart represents the average acre lot for the top 30 cities. The chart demonstrates a clear progression from left to right, with Moss Bluff, Woodbury Heights, and Solomon having the smallest average acre lot sizes among the top 30 cities. As you move to the right, the average lot size increases significantly, with Hudson Falls and Fultonville having the largest average acre lots. The spread between the smallest and largest acre lots is considerable, indicating that certain cities have much larger properties on average. This disparity could be due to various factors, such as zoning regulations, population density, or geographic characteristics.

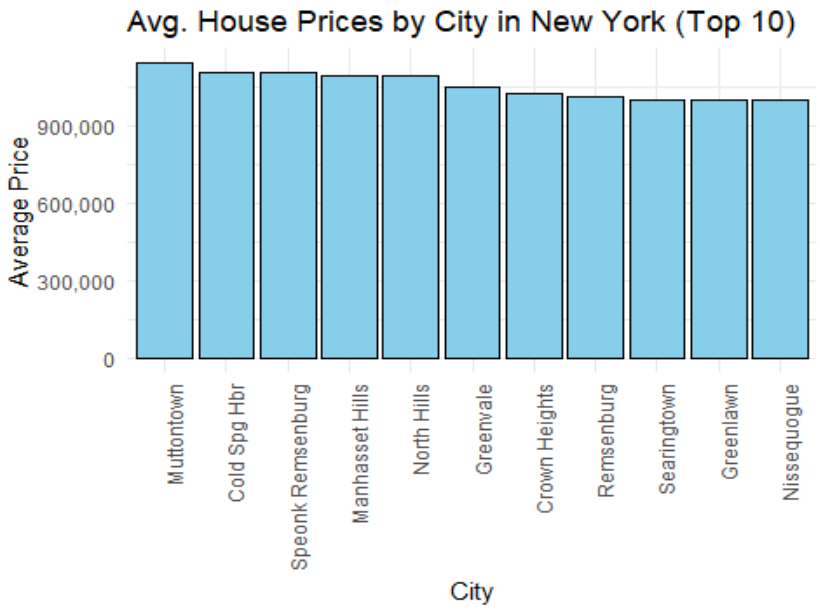


This scatter plot shows the relationship between selling price and acre area. The plot has a dense concentration of points on the left side, where the acre area is smaller. This indicates that many properties are on smaller plots of land, with prices clustering in the lower to mid-range. As acre area increases, the scatter of points becomes more spread out, suggesting a broader range of selling prices for larger properties. There's a general upward trend, indicating that as acre area increases, selling prices can also increase, but the correlation is not strong. This suggests that while larger properties tend to have higher selling prices, many other factors can influence the price.

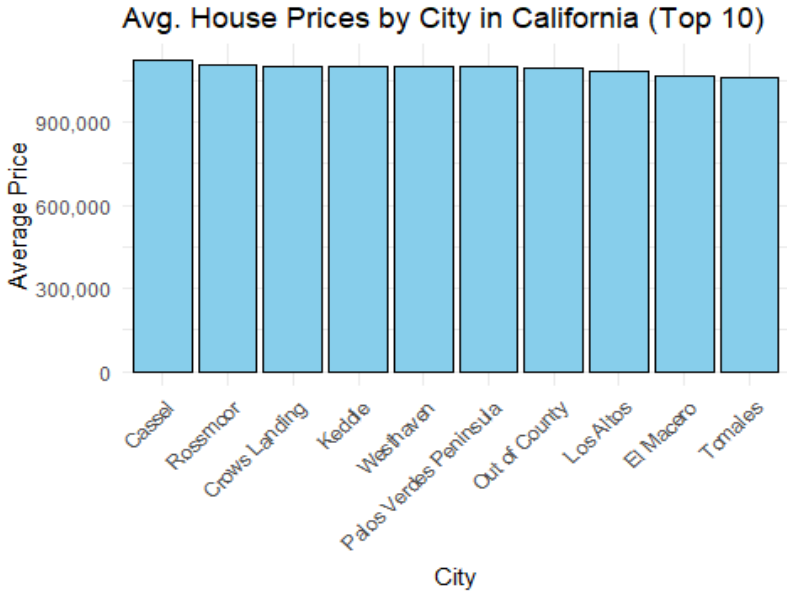


This bar chart displays the average house prices for the top 10 cities in Illinois. From the chart, you can see that Johnsonville has the highest average house price, with a noticeable gap compared to the other cities. Bannockburn and Greenfield follow, with slightly lower average prices. As you move across the chart, the average house prices gradually decrease, with Gorham

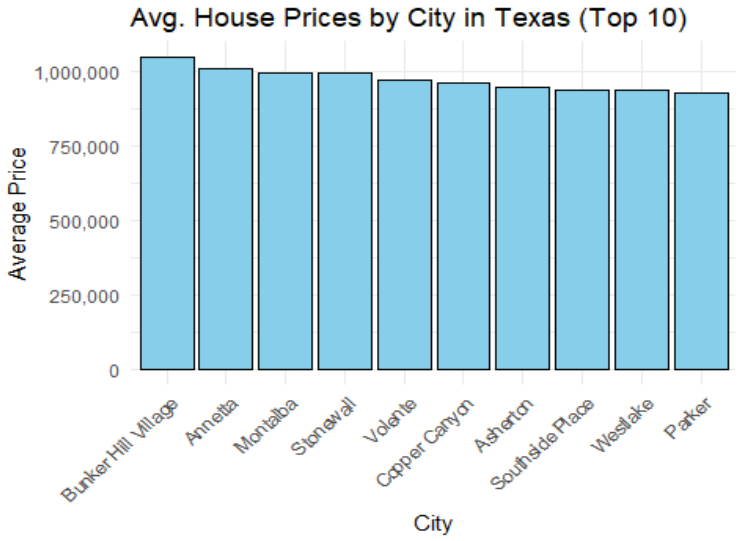
having the lowest average price among the top 10 cities. The bar heights demonstrate the variations in house prices across different cities, suggesting that some areas in Illinois have significantly higher real estate values than others. This pattern could be due to a range of factors, including location desirability, local amenities, or housing market trends.



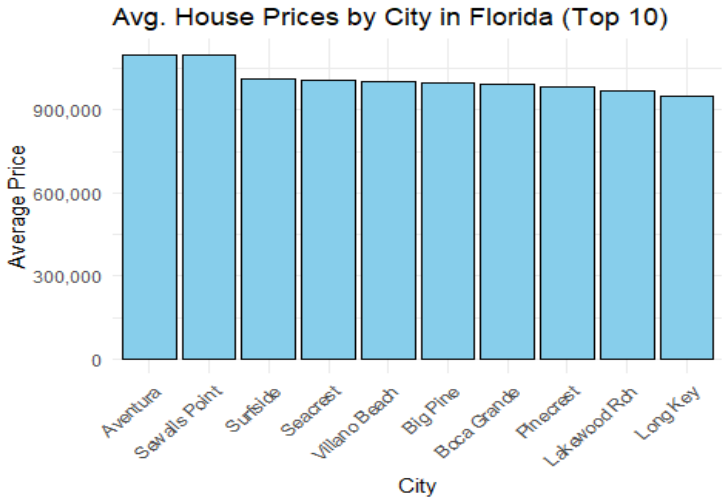
This bar chart illustrates the average house prices in the top 10 cities in New York. In this chart, Muttontown has the highest average house price, followed closely by Cold Spring Harbor and Speonk-Remsenburg. These cities tend to have a higher cost of living or more exclusive real estate markets. Moving from left to right, the average house prices generally decrease, with Nissequogue at the lower end of the scale among the top 10.



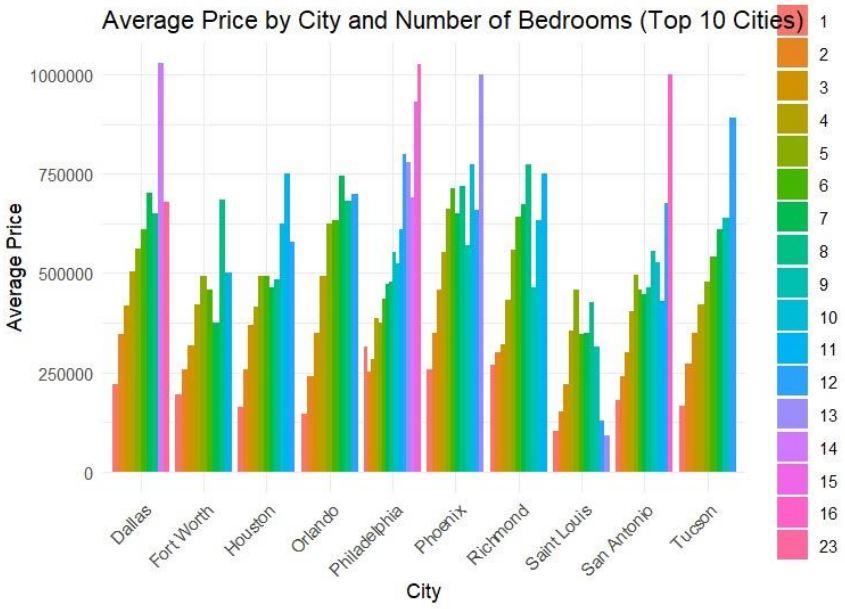
The bar chart titled "Avg. House Prices by City in California (Top 10)" illustrates the average house prices for the top 10 cities in California. Cassel has the highest average house price among the top 10 cities, with other cities like Rossmoor and Crows Landing showing slightly lower average prices. Overall, the average prices are relatively close among these top 10 cities, with Tomales at the lower end of the price range.



This bar chart, titled "Avg. House Prices by City in Texas (Top 10)," displays the average house prices for the top 10 cities in Texas. Bunker Hill Village leads with the highest average house price among the top 10, close to $1,000,000. The other cities follow in a relatively consistent pattern, with the price gradually decreasing as you move from left to right. Annetta and Montaba are close behind Bunker Hill Village, indicating a cluster of cities with high real estate values. Parker, at the lower end of the scale, still has an average price close to $800,000.



The bar chart titled "Avg. House Prices by City in Florida (Top 10)" depicts the average house prices for the top 10 cities in Florida. Aventura leads the list with the highest average house price, closely followed by Sewalls Point and Surfside. The other cities display a similar pattern, with average prices generally decreasing as you move from left to right. Long Key has the lowest average price among the top 10, but it's still quite close to the others, indicating a fairly consistent range of high prices among these cities.

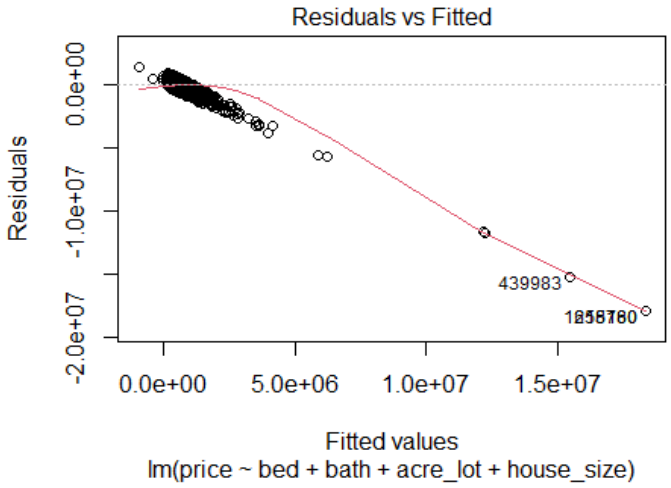


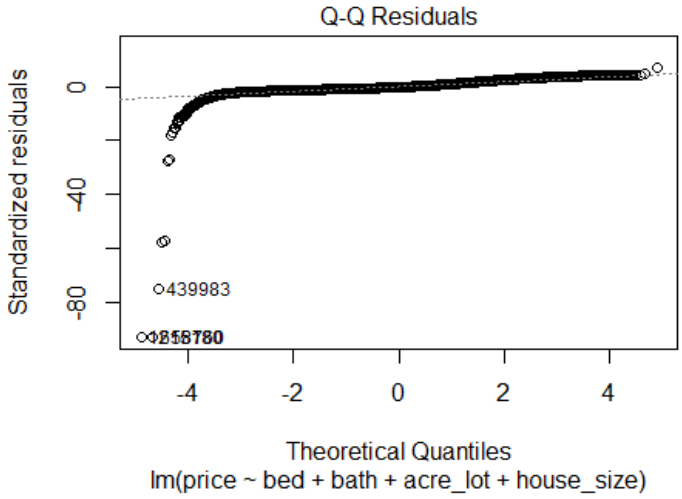
This bar chart, titled "Average Price by City and Number of Bedrooms (Top 10 Cities)," shows the average house price across 10 cities, with each bar indicating the average price by the number of bedrooms. The bars are color-coded to represent the number of bedrooms, with a gradient legend on the right indicating the corresponding number of bedrooms. From the chart, you can see a general trend: as the number of bedrooms increases, the average price tends to rise. Cities like Dallas, Houston, and Tucson display a higher average price for houses with more bedrooms, while cities like Orlando and Richmond show a more gradual increase.

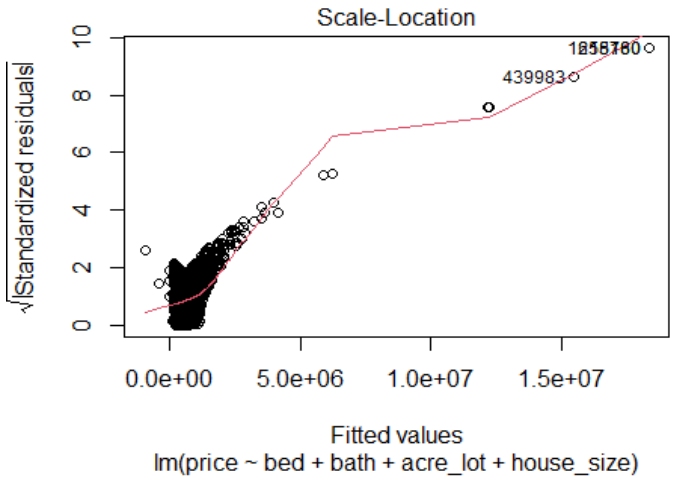
Modelling:

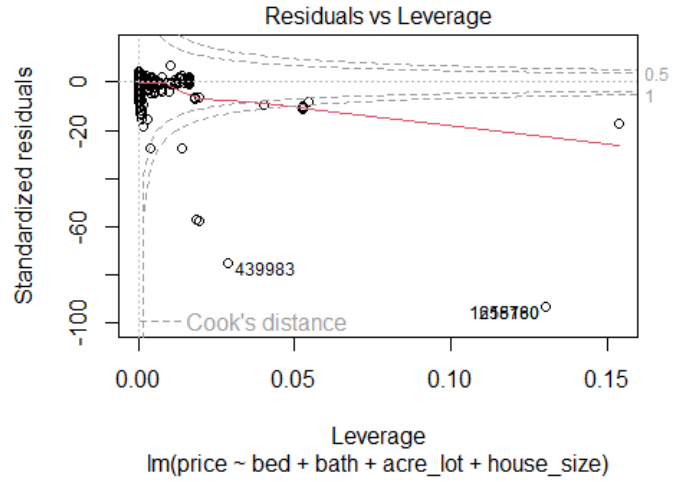
The linear regression model exhibited a notable accuracy of 78.8%, indicating its proficiency in categorizing property prices. With a minimal Mean Squared Error (MSE) of 42.7 billion and a Mean Absolute Error (MAE) of approximately 158,374, the model demonstrates robust performance. Root Mean Squared Error (RMSE) stands at 206,742, reflecting a relatively low level of prediction error. Despite some limitations, such as the model's sensitivity and specificity, the overall balanced accuracy remains satisfactory. This suggests the model's ability to maintain a reasonable balance between true positive and true negative rates. Additionally, the model's Kappa statistic indicates a marginal agreement between observed and predicted classifications.

However, further refinement may be required to enhance the model's predictive capabilities. Overall, the model presents a promising tool for analyzing real estate market trends and making informed decisions.









Conclusion:

The real estate analysis uncovers significant variations in house prices across cities, driven by location and property attributes. Correlations between price, house size, and acre area affirm key determinants of real estate value. Despite achieving 78% accuracy, the model's performance suggests opportunities for enhancement through feature refinement. Market dynamics, as illustrated by the distribution of "for_sale" and "sold" listings, provide valuable context for predicting future trends. Overall, the project offers insights into real estate trends and market dynamics, facilitating informed decision-making in the industry.

Limitations and Future Work:

The analysis has some limitations that we need to address. We're currently relying too much on accuracy alone, which might not give us the full picture, especially considering the diverse nature of real estate data. We should consider using other metrics like precision and recall to get a better understanding, especially when dealing with data that's not evenly distributed. Additionally, our analysis could be affected by incomplete or outdated data, as well as outliers, so we need to make sure we're working with high-quality data.

Our current focus on specific features like house size and acre area might be too narrow. We're missing out on other crucial factors like neighborhood quality and nearby amenities, which can significantly impact property prices. To improve our model, we should expand our dataset to include more diverse features and explore different machine learning methods. It's also important to validate our model and fine-tune its parameters to ensure its accuracy and consistency.

Looking ahead, we should also consider analyzing trends over time. By tracking how real estate trends change over different periods, we can better understand the market dynamics and make more informed investment decisions in the future.

Data Sources:

https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/code?datasetId=3202774

Source Code:

https://github.com/Abhiram1819/CSP_571_DPA_Project.git

References:

[1] Choy, Lennon & Ho, Winky. (2023). The Use of Machine Learning in Real Estate Research. 12.740.10.3390/land12040740.z

[2] Gale, H., Roy, S.S. Optimization of United States Residential Real Estate Investment through Geospatial Analysis and Market Timing. Appl. Spatial Analysis 16, 315–328 (2023).

[3] Kumar, Sailaja & KameshwariSoundarya, & Harshitha, R. & EvangelinGeetha, D. & T.V., Suresh. (2018). Real estate data analysisusing principal component analysis and 'R'. International Journal of Pure and Applied Mathematics. 119. 1535-1541.

[4] Real Estate Data Analysis with Python
https://medium.com/@kingsleyofori/real-estate-data-analysis-with-python-b0004baf9abb

[5] Real Estate Analysis and Modelling using Python
https://medium.com/@filipesampaiocampos/real-estate-data-analysis-and-modeling-using-python-184252d60189