U.S.A. Real Estate Analysis

Introduction:
This case study looked at real estate across the U.S.A., specifically across the different regions and investigating whether certain variables influenced the market.

Goal:
Analyze the United States real estate market to see what factors or variables influence sales the most.

Hypothesis:
As house size increases, price increases.

Steps and Skills:

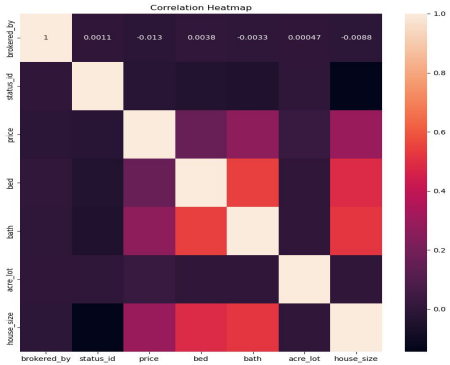
- Data Cleaning
- Exploratory Analysis
- Linear Regression
- Scatterplots
- Clusterplot Analysis
- Pairplots

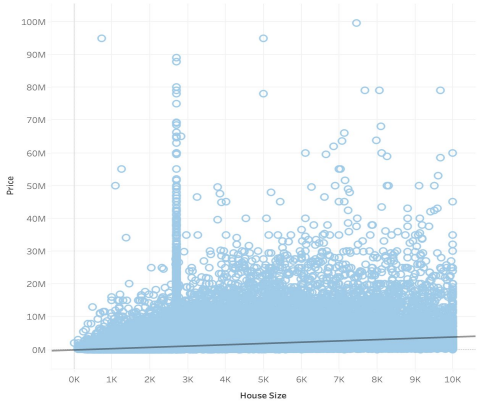
Tools Used:

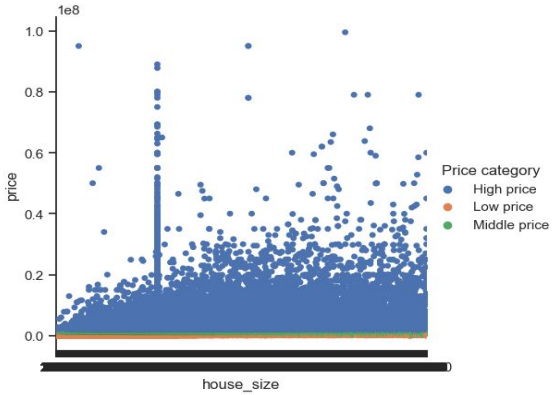
U.S.A. Real Estate Analysis

Exploratory Analysis:

By using this correlation heatmap (below), I was able to determine the strongest positive correlations. Since bed, bath, and house size all seem to be obvious in their correlation, it was determined to further explore the relationship between house size and price (noted by the lighter purple color).







***HYPOTHESIS:**

As the house size increases, so does the price.

To test this hypothesis, I conducted a linear regression.
There are many points that fall outside of the regression line. This isn't enough to draw a significant conclusion. I will try another approach -> Clusters

In order to prove the hypothesis besides only using linear regression, I conducted a cluster analysis. This categorical plot groups data into 'clusters' in order to compare each group to uncover patterns.

-Cluster 0 represents low priced homes.
-Cluster 1 represents medium to high priced homes
-Cluster 2 represents very high priced homes

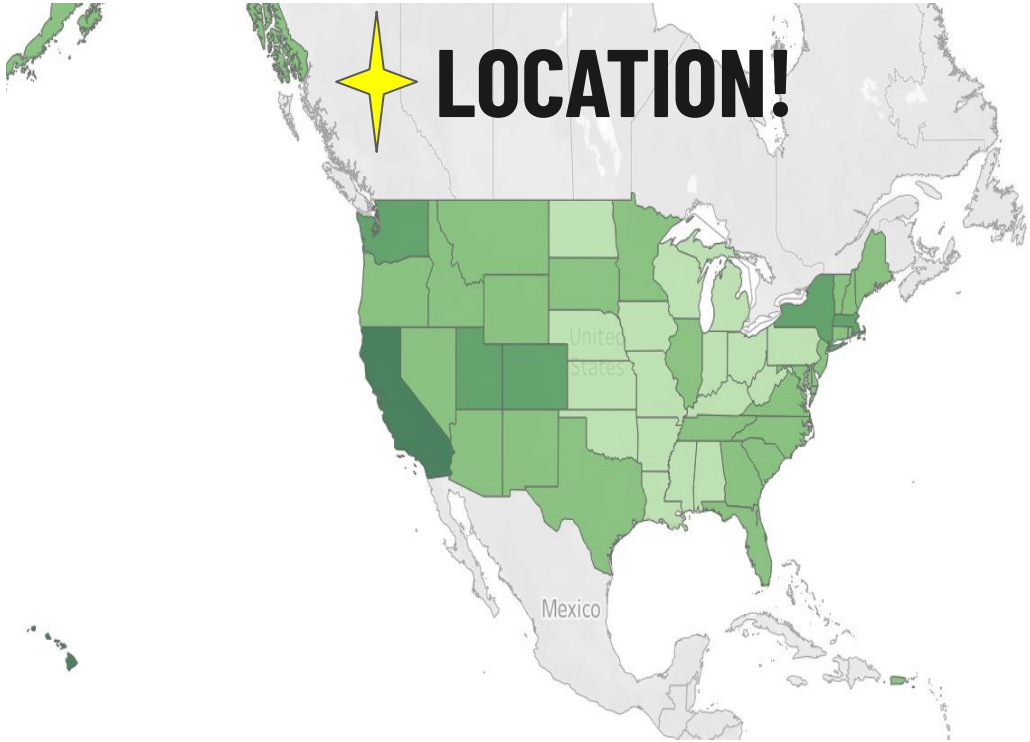
This plot shows that cluster 1 (med-to-high priced homes) have outliers for very high priced homes while the other two clusters stay within a certain range.

All three clusters share a large amount of homes at around 2700 square feet (as noted previously), meaning this is a common house size across all price ranges.

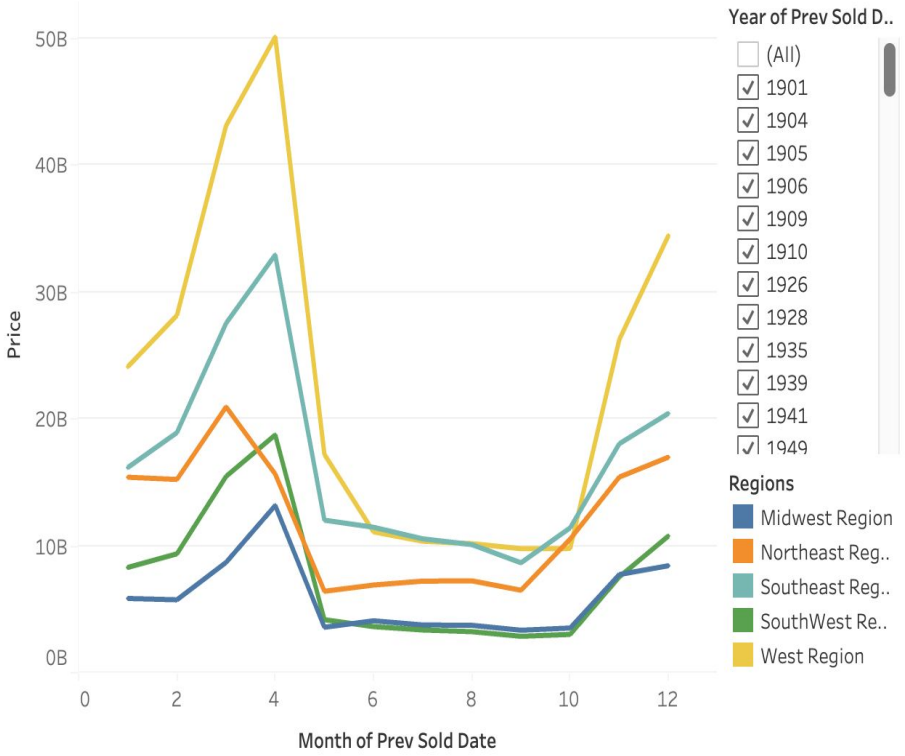
U.S.A. Real Estate Analysis

Geography plays an important role on the price. Western and Northeast Regions, indicated by the darkest green color, demonstrate the highest revenue generated from real estate within the U.S.A. These top states include Hawaii, California, Montana, Colorado, Utah, New York, & Massachesetts.

Lighter green states (midwest) show the least revenue generating states.



U.S.A. Real Estate Analysis



Seasonality (1901-2023)

Stays the same across all regions

Most homes are sold in March and April. This is the same across all regions and across time. Revenue is represented highest in total for West Regions, then Southeast, Northeast, Southwest, and Midwest marked as producing the lowest revenue. This is parallel to cost of living parameters.

Trends remain the same.

April is the hot month for selling/buying.

Note: 2024 was left out as homes without a previous sold date were marked for today's date (raising the month to current month of august 2024). By removing this year, we have eliminated the risk of tainting data to show more sales where they have not occurred.

U.S.A. Real Estate Analysis

Insights:

- PRICES and LOCATION:
  - Highest real estate prices exist within the west and northeast regions:
  - CA, HI, MT, CO, VT, NY, MA
  - Lowest real estate prices are in the midwest
- PRICES and HOUSE SIZE:
  - Strongest correlations exist between house size and price
  - The bigger the house size, the higher the price is, generally.
  - 2700 square feet tends to be a house size that is standard across ALL PRICES!
- SEASONALITY:
  - Most homes are sold in April (and March).

Recommendations:

- There was a large amount of data where previous sold dates were blank. It was assumed that these were newer builds without a previously sold date. These homes were left out of the seasonality. Next steps would be to clarify this.
- Acre lot sizes and house sizes had a HUGE variance in size, along with price range. It would have been insightful to have more details with these homes/listings to understand the context of these ranges. Especially acre lots and house sizes marked as '0'. This does not make sense as it is impossible to have a house or lot size equal to 0. Next steps would be to clarify the number markers, determine if these were missing values filled in, and eliminate or impute values.

Thank you for your time and consideration.

nancykolaski@gmail.com

github.com/Nancy-Kolaski