Elishia Fitzgerald

Western Governors University

D214: Data Analytics Graduate Capstone

Instructor: William Sewell

April 26, 2025

Problem & Hypothesis

This project explores whether there is a statistically significant difference in average home prices between low-priced and high-priced zip codes in Houston, Texas, using a dataset obtained from Kaggle that contains Houston housing data from June 2024.

Research Question- Is there a statistically significant difference in average home prices between low-priced and high-priced zip codes in Houston, TX?

Null hypothesis- There is no significant difference in the average home prices between low-priced and high-priced zip codes in Houston.

Alternate Hypothesis- There is a significant difference in the average home prices between low-priced and high-priced zip codes in Houston.

Data Analysis Process

All data preparation and analysis were performed in Python using Jupyter Notebook, with libraries including pandas, seaborn, matplotlib, and scipy.

The dataset consisted of over 25,900 Houston housing listings, collected from Zillow in June 2024. The original format was a nested JSON file, which was converted into a structured DataFrame using pandas.

The data contained 58 columns, many of which were unnecessary for this analysis. Columns related to listing images, agent contacts, and tour dates were dropped to focus on price, square footage, bedrooms, bathrooms, zip code, and property location. A subset of key numerical columns was retained and renamed for clarity.

To address missing data, I used median imputation grouped by zip code. This ensured that missing value for features like the number of bedrooms, bathrooms, and square footage were filled in using more group-specific data rather than an overall average. The decision to retain high-priced outliers was intentional, as they reflect true values in the housing market and could impact pricing insights.

Zip codes were grouped into low, medium, and high-priced categories using price quartiles. While the original research question compared only low- and high-priced groups, the addition of a medium tier provided a more accurate reflection of the distribution of prices across Houston zip codes.

During data exploration, I used histograms, scatterplots, and a correlation heatmap to assess variable distributions, relationships, and outlier behavior. These visualizations helped support decisions made during the data cleaning and feature selection process.

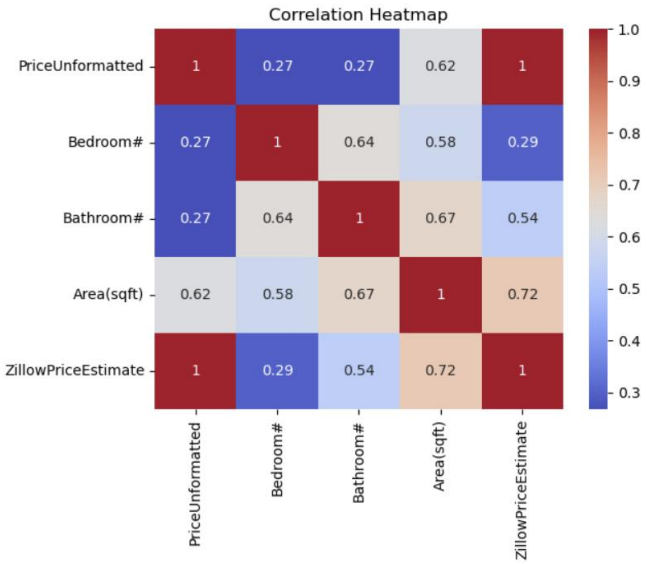

Figure 1: Correlation Heatmap

Findings

After cleaning and preparing the dataset, I ran several statistical tests to answer the research question. First, the Shapiro-Wilk test was performed on the home prices for each zip code group (low, medium, high). The results showed that the price data was not normally distributed ($p < 0.001$ for all groups). Next, Levene's test was conducted, which indicated that variances across the groups were not equal ($p < 0.001$). Because of these violations of assumptions for standard ANOVA, I used Welch's ANOVA, which does not assume equal variances.

Welch's ANOVA results: $F(2, 25945) = 387.14$, $p < 0.001$

This small p-value led to the rejection of the null hypothesis, confirming that there is a statistically significant difference in average home prices across the zip code pricing groups.

To gain a further understanding of which groups were different, I also ran a Tukey's HSD post-hoc test. The results showed that:

- High-priced zip codes were significantly different from medium-priced zip codes
- High-priced zip codes were significantly different from low-priced zip codes
- Medium-priced zip codes were significantly different from low-priced zip codes

These findings further support the conclusion that home prices vary significantly based on zip code economic groups in Houston.

Limitations

One limitation of this project is that the dataset only represents a snippet of Houston housing listings from June 2024. Because it only covers one month, the results do not account for

seasonal trends, economic shifts, or yearly market cycles that could influence home prices. As a result, the findings may not fully reflect long-term pricing patterns across Houston zip codes.

A statistical limitation of this analysis is, although Welch's ANOVA was the correct choice due to the unequal variances detected, it also has its disadvantages. Welch's test identifies that differences exist between groups but does not indicate which specific groups differ. As a result, Tukey's HSD post-hoc test was necessary to fully interpret which groups differed. Additionally, Welch's test can be slightly less powerful than a standard ANOVA when variances are actually equal, but its use was appropriate with the variance violation that was observed in the dataset.

Recommended Actions

Based on the results, it is recommended that real estate professionals, investors, and buyers incorporate a zip code-level pricing analysis when evaluating home values in Houston. Stakeholders could use price grouping strategies such as low, medium, and high to target buying opportunities, predict market behavior, and optimize pricing strategies. Instead of viewing Houston's market as a single market, focusing on specific zip code price tiers will allow for more accurate property valuation and better investment decision-making.

Expected Benefits

Incorporating zip code-level analysis into real estate evaluation offers several key benefits:

- Increased Investment Accuracy: Investors can better identify undervalued zip codes and avoid overpaying in high-priced areas.
- Improved Pricing Strategies: Real estate agents can tailor listing prices more competitively based on zip code pricing tiers, potentially leading to faster sales and higher closing rates.

- Smarter Buying Decisions: Buyers can make more informed choices by understanding how home prices vary across economic zip code groups, improving the chances of securing better deals.

By using pricing insights based on zip code groups, stakeholders could potentially optimize property purchases or sales by 5% to 15% compared to strategies that do not consider localized price trends, depending on market conditions.