Capstone Two - Final Project Report

Problem Statement

Analyze the American housing market dataset to understand trends, relationships, and factors influencing property prices and to tell a compelling data story.

Approach

The analysis involved loading and exploring the dataset, performing exploratory data analysis using visualizations (histograms, scatter plots, box plots, geographical scatter plot, bar plot) to identify trends and correlations between various features and property prices, and summarizing the key findings and insights.

Findings

- The dataset contains information on housing properties across different states and zip codes, including price, living space, number of beds and baths, location (latitude, longitude), and demographic information (zip code population, zip code density, median household income).
- The distribution of property prices is heavily right-skewed.
- Living space and the number of beds and baths show positive relationships with property price.
- Median household income is positively correlated with property price.
- Zip code density has a weak positive correlation with price.
- There is no strong linear correlation between zip code population and price.
- Property prices exhibit geographical clustering, with higher prices observed in certain regions.
- The dataset's coverage varies by state, with California and Texas having the most properties.

Insights

- Living space, median household income, and geographical location are significant factors influencing property prices.
- The number of beds and baths are also important determinants of price.
- The skewed price distribution highlights the presence of a luxury market segment.
- The uneven state-wise data distribution should be considered when interpreting results.

Hypotheses

1. Zip codes with a higher median household income have a significantly higher average property price compared to zip codes with a lower median household income.

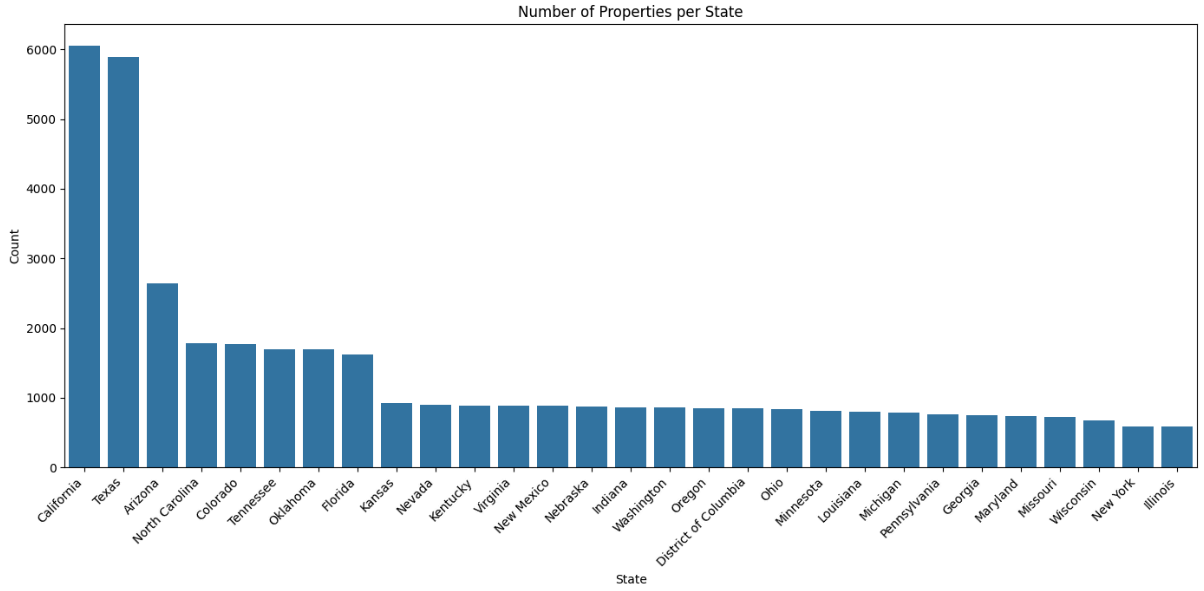
2. There is a strong positive linear correlation between the living space of a property and its price, independent of its geographical location.

Further Questions

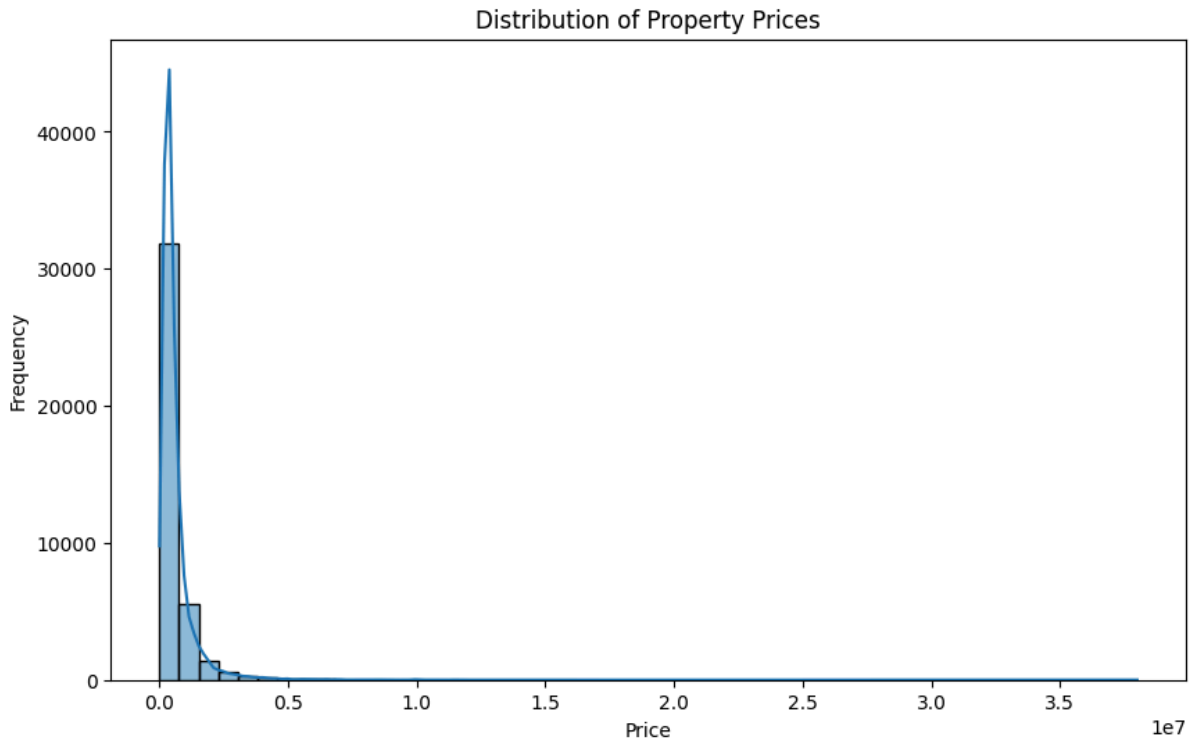
1. What is the average price per square foot in different states or major metropolitan areas within the dataset?
2. Are there specific characteristics (e.g., number of beds/baths, living space, income) that define high-value properties compared to average-value properties?
3. How does the age of the property, if available, correlate with its price, and does this relationship vary by location?

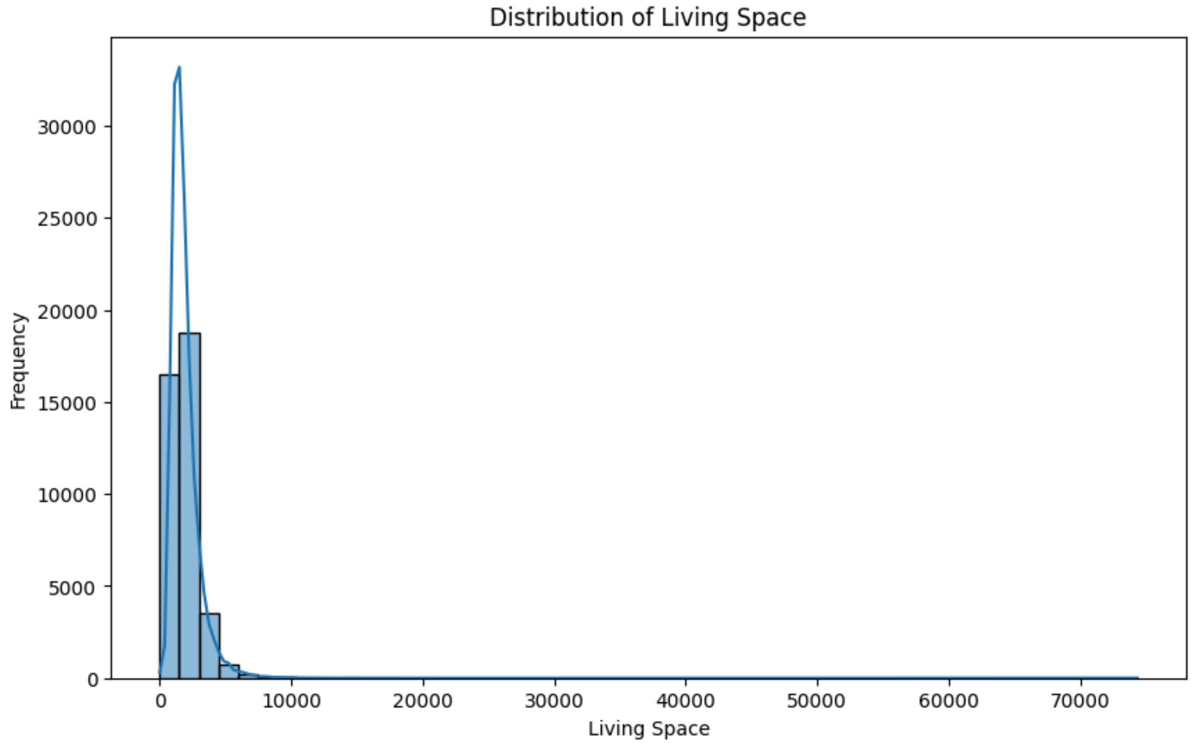
Visualizations

1. Number of properties per state

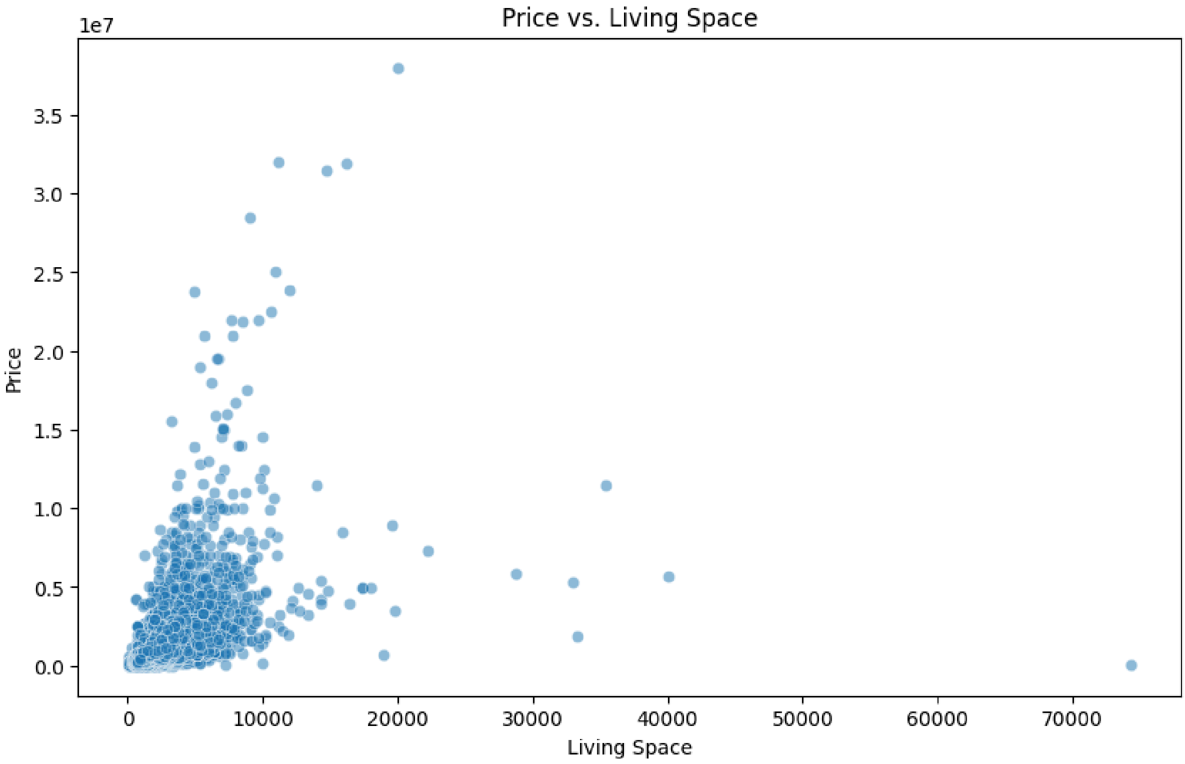


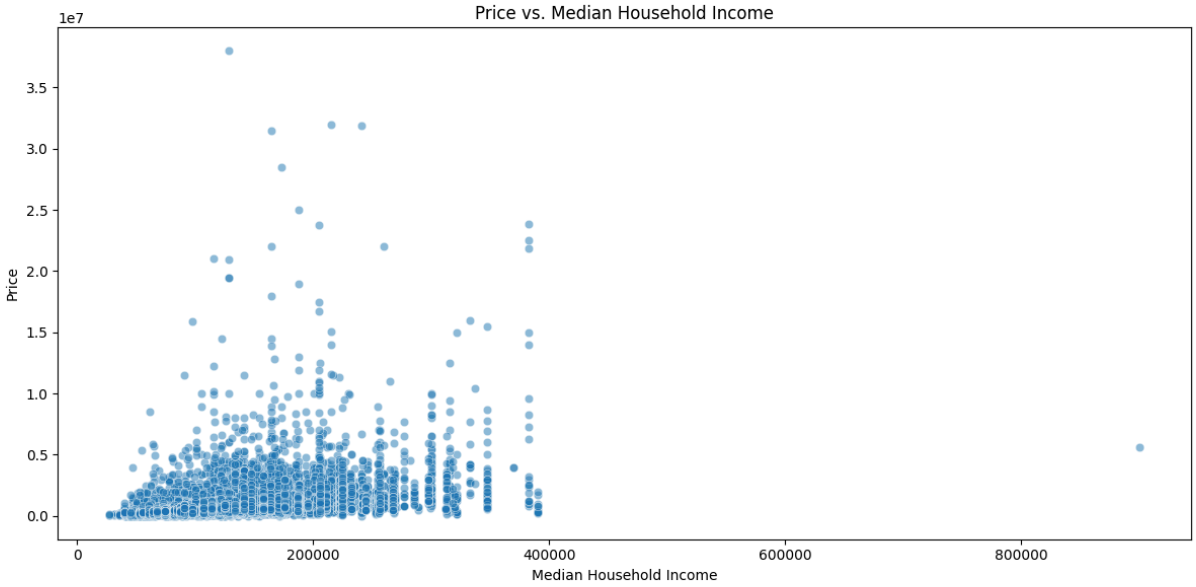
2. histograms for Price and Living Space



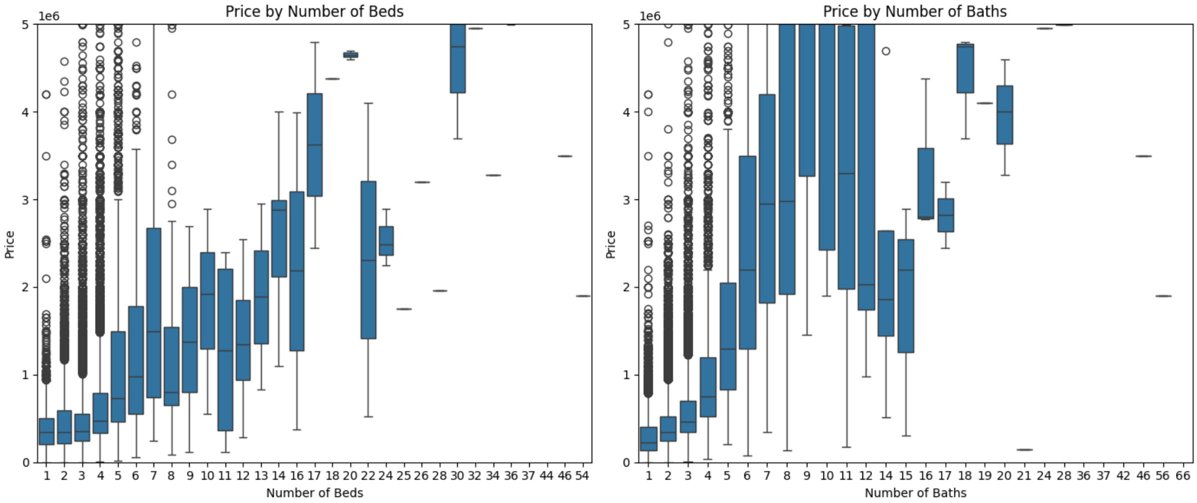


3. Living Space vs. Price

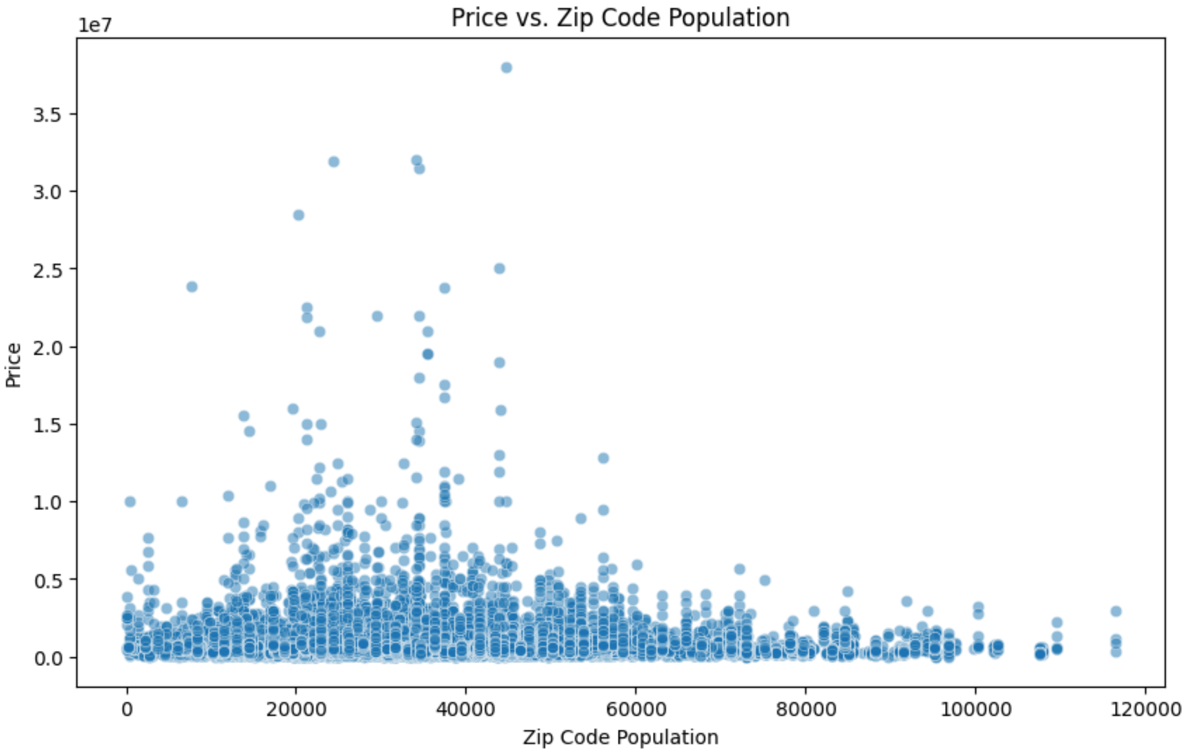




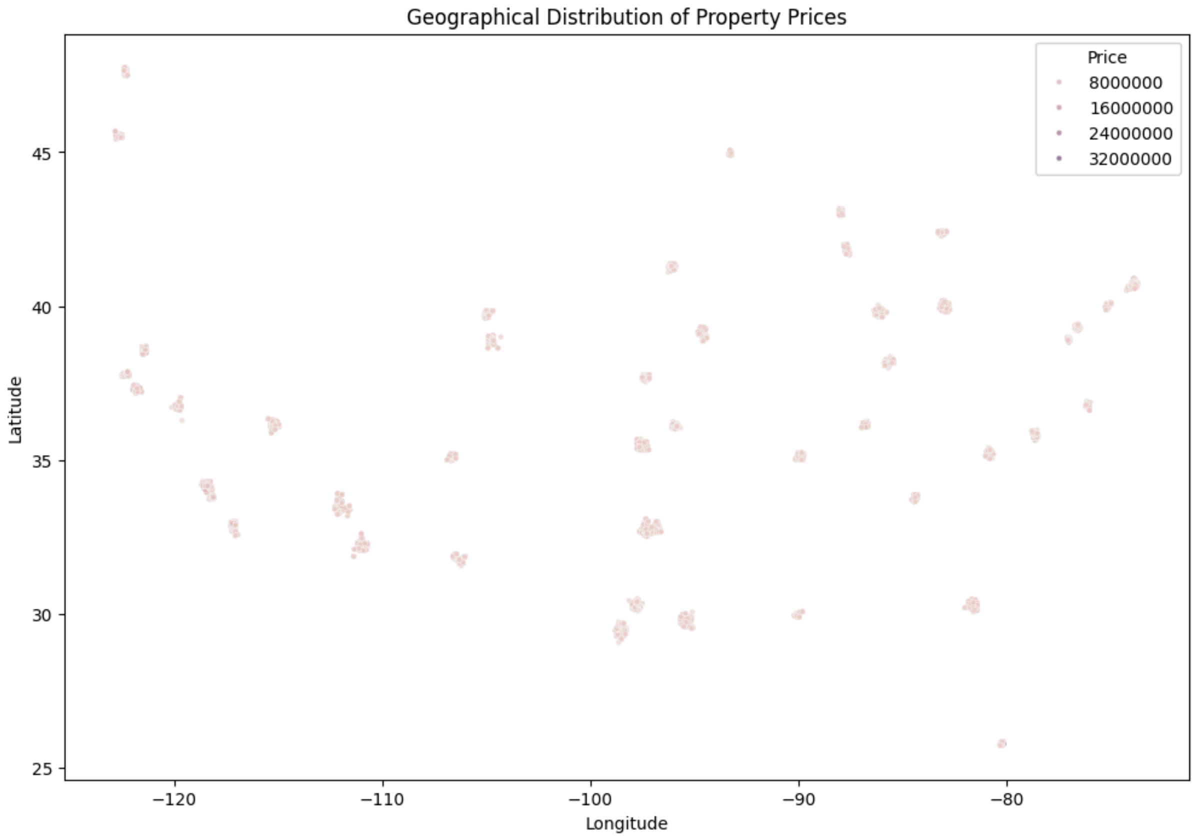
4. box plots for Beds vs. Price and Baths vs. Price



5. Zip Code Population vs. Price



6. geographical scatter plot of Latitude and Longitude colored by Price



Model Metrics

No predictive model was built as part of this exploratory data analysis. The primary focus of this project was to understand the trends, relationships, and patterns within the American housing market dataset through data visualization and interpretation, rather than developing a model for price prediction.

Summary:

Data Analysis Key Findings

- The dataset contains information on housing properties across different states and zip codes, including price, living space, number of beds and baths, location (latitude, longitude), and demographic information.
- The distribution of property prices is heavily right-skewed.
- Living space, number of beds and baths, and median household income show positive relationships with property price.
- There is a weak positive correlation between zip code density and price.
- There is no strong linear correlation between zip code population and price.
- Higher-priced properties tend to cluster in specific geographical locations.
- The dataset's coverage varies by state, with California and Texas having the most properties represented.

Insights or Next Steps

- Living space, median household income, and geographical location are significant factors influencing property prices and could be used in a predictive model.
- Further investigation using statistical methods could quantify the strength of the identified correlations.