Final Draft

Project Title: California Housing Price Trends and Predictive Modeling
Course: MATH 120
Author: Elliot Wiley

1. Introduction

This project analyzes long-term housing price trends in California using Zillow Home Value Index (ZHVI) data. The goal is to understand how median home values have evolved over time in major California cities and to evaluate how well different modeling approaches capture these trends. Housing prices are a critical economic indicator in California, influencing affordability, migration patterns, and regional inequality. The questions I aim to answer are:

1. How have median home values changed over time in major California cities over the past decade?
2. Which California cities have experienced the fastest and slowest housing price growth?
3. Did the COVID-19 pandemic correspond with noticeable changes in housing price trends?
4. How well do different regression-based models describe and predict long-term housing price trends?

The dataset comes from the Zillow Home Value Index, which provides monthly median home values for cities across the United States. This project focuses on California cities only. The analysis combines time-series visualization, rolling-average/smoothing, linear regression, and polynomial regression to compare model performance and interpret housing market behavior.

2. Methods

The Zillow dataset was loaded directly from Zillow’s public research website to ensure reproducibility. The data was filtered to include only California cities, and a subset of major cities was selected for analysis. Each row in the dataset represents a city, while columns represent monthly median home values, creating a wide time-series structure. Date columns were converted to datetime objects to allow plots to display actual calendar years. For regression modeling, time was represented numerically as a sequential index. Cities with excessive missing values were removed, and small gaps were handled using forward-filling to preserve continuity. Several analytical methods were applied. Time-series line plots were created to visualize long-term housing trends within and across cities. A 12-month rolling average was computed for each city to smooth short-term fluctuations and better show long term trends. Linear regression models were fit separately for each city to establish baseline growth rates however since the linear versions were not super accurate I moved to try some polynomial regression (quadratic) was also applied to capture non-linear growth patterns, particularly during periods of rapid price acceleration. Model performance was evaluated using visual comparisons of predicted versus actual values and by calculating mean absolute error (MAE) for each model and city.

3. Results

Time-series plots show a consistent increase in median home values across all selected California cities. Coastal metropolitan areas such as San Francisco, San Jose, and Los Angeles display higher overall prices and steeper growth than inland cities like Fresno and Sacramento. The rolling-average smoothing reveals a clear acceleration in housing prices beginning around 2020, corresponding with the COVID-19 pandemic. While short-term fluctuations are present, the smoothed trends confirm sustained growth across regions. Linear regression models capture the general upward trend but struggle during periods of rapid change. Polynomial regression improves fit in many cities by accounting for non-linear growth, especially in the post-2020 period. In most cases, polynomial models achieve lower MAE values than linear models, indicating improved predictive accuracy. Despite improvements, both regression approaches show limitations when applied to volatile housing markets, particularly when prices change abruptly due to external economic factors. One key thing to look out for is the large dip in San Jose before 2024, this dip doesn’t really have an explanation even after research.

4. Discussion

The results demonstrate that California housing prices have increased quite a lot over the past decade, with notable differences between coastal and inland regions. Rolling averages proved effective for identifying long-term trends, while polnomial regression offered a meaningful improvement over linear models by capturing accelerating growth. However, the models remain simplified representations of a complex system. Housing prices are influenced by interest rates, employment trends, migration, and policy decisions, none of which are directly included in this analysis. As a result, regression models are best interpreted as descriptive rather than fully predictive. From an ethical and social perspective, rising housing prices raise concerns about affordability and displacement. Understanding these trends can inform discussions about housing policy and urban planning. With more time, this project could be extended by incorporating economic indicators, regional comparisons beyond California, or more advanced time-series models.

5. Conclusion

This project analyzed California housing price trends using Zillow data and multiple modeling approaches. Time-series visualization and rolling averages revealed consistent long-term growth, while polynomial regression improved upon linear models by capturing non-linear price behavior. Overall, the analysis highlights both the value and the limitations of simple models when studying complex economic systems.