STAT545 Project Proposal

Andrew Cao (hanghec2)

March 2025

1 Background

California's housing market provides an example of complex interactions among geography, socioeconomic forces, and policy challenges. One of the most economically dynamic and demographically diverse regions in the U.S., it has struggled with persistent housing affordability crises due to rapid urbanization, limited land availability and disparate income distribution patterns. Geospatial analysis of housing prices provides critical insight into how physical landscape features such as coastal proximity, terrain constraints and urban clustering intersect with human behavior and policy frameworks to influence real estate dynamics.

The 1990 U.S. Census dataset, which underpins this analysis, captures a pivotal moment in California's demographic history, preceding the tech-driven economic boom of the 2000s. During this period, spatial patterns of housing values began reflecting early signs of the state's modern affordability challenges. Coastal regions, for example, already exhibited premium pricing due to their environmental amenities and economic opportunities, while inland areas housed lower-income populations—a dichotomy that has since intensified. By examining geographic coordinates (longitude/latitude), neighborhood characteristics (e.g., housing age, population density), and economic indicators (median income), this study contextualizes how spatial heterogeneity perpetuates inequities in housing access.

Geospatial methodologies offer a powerful solution to understanding these patterns. While traditional economic models may ignore locational dependencies, spatial regression techniques account for autocorrelation - where housing prices in adjacent regions affect one another- and reveal hidden connections between environmental variables (e.g. ocean proximity) and market behavior. Such analysis not only furthers academic understanding of urban economics but also informs equitable land-use policies in regions like California where climate risks (such as rising sea levels) and gentrification pressures threaten vulnerable communities.

2 Objectives

This project uses geospatial analysis techniques to examine spatial patterns and drivers of housing prices across California using the 1990 U.S. Census dataset. Moran's I will help identify any clustering tendencies in housing values, determining whether high or low-price neighborhoods tend to cluster nonrandomly. Second, hotspot/coldspot analysis will map statistically significant high-value zones (e.g. coastal regions) and low-value clusters (e.g. inland areas), which contextualize spatial inequities. Geographically Weighted Regression (GWR) will model how predictors like median income and ocean proximity influence prices differently across regions, thus accounting for spatial non-stationarity. Kernel Density Estimation (KDE) can also provide an attractive visual display of housing price distribution, overlaid on geographic features like urban centers to reveal latent spatial patterns. These methodologies–applied through R geospatial libraries, such as "`sf`, `tigris`, `ggplot2`, and ect", –will unravel California's housing dynamics by linking concepts of spatial dependence, heterogeneity, and interpolation to real world policy issues related to affordability and equitable development.

3 Data Source

This project employs the 1990 California Housing Dataset, originally sourced from the U.S. Census Bureau [2] and popularized as a pedagogical resource in Aurélien Géron's Hands-On Machine Learning with Scikit-Learn and TensorFlow. The dataset, hosted on Kaggle [1], provides a curated snapshot of California's housing landscape at the census tract level, balancing accessibility and analytical depth.

4 Data Description

This dataset contains **20,640 observations** of **10 variables**, structured as follows:

- `longitude` (float): Geographic longitude coordinate
- `latitude` (float): Geographic latitude coordinate
- `housing_median_age` (float): Median age of houses (years)
- `total_rooms` (int): Total rooms per district
- `total_bedrooms` (float): Total bedrooms per district (contains missing values)
- `population` (int): Total population per district
- `households` (int): Number of households
- `median_income` (float): Median household income (scaled: 1 = $10,000)
- `median_house_value` (int): Median house value (USD)
- `ocean_proximity` (str): Categorical proximity to ocean (5 categories)

Key Notes

- Missing values exist in `total_bedrooms`
- `ocean_proximity` has categorical values: `NEAR BAY`, `<1H OCEAN`, `INLAND`, `NEAR OCEAN`, `ISLAND`
- `median_income` uses a scaled representation (unit = $10,000)

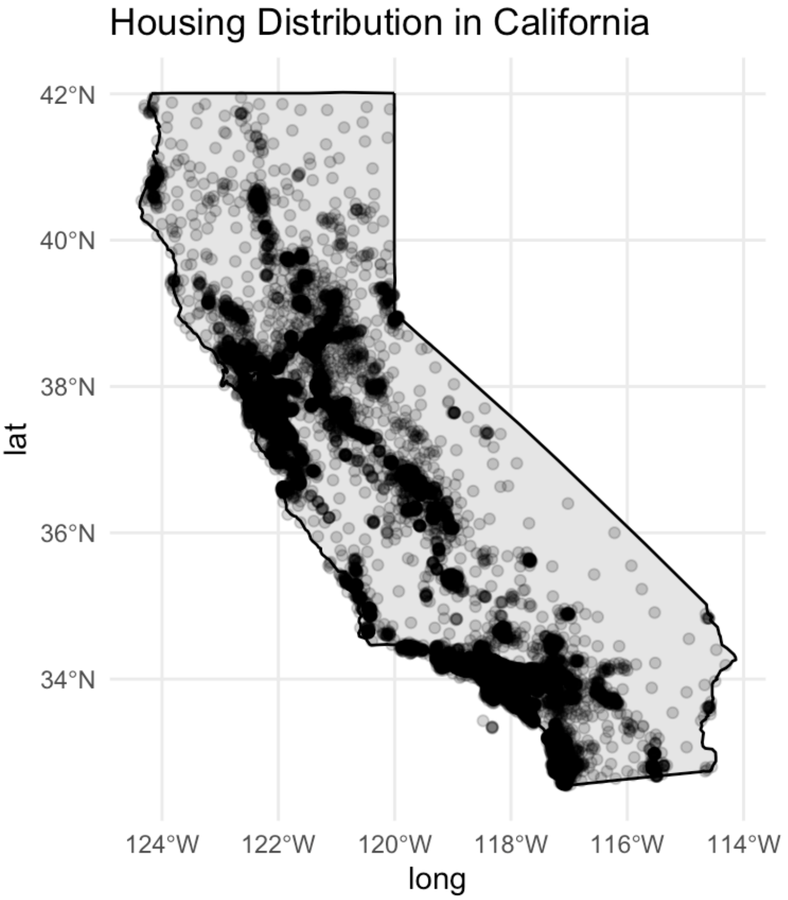

Figure 1: Housing Distribution in California Plot

References

[1] Aurélien Géron. California housing prices. Kaggle Dataset, 2017. Adapted from the 1990 U.S. Census. URL: https://www.kaggle.com/datasets/camnugent/california-housing-prices.

[2] U.S. Census Bureau. 1990 census of population and housing. U.S. Government Publication, 1990. URL: https://www.census.gov/.