CS 418

Intro to Data Science

Prof. Brian Ziebart

Rising Cost of Homes and Unaffordability

Big Data Dunces Team

github.com/uic-ds-fall2025/class-project-big-data-dunces

David Yanjaa Adnan Sufyan Philip

TABLE OF CONTENTS

01. Scope

02. Datasets

03. Exploratory Data Analysis

04. Hypothesis Visualization

05. Machine Learning

Big Data Dunces

01.

Scope

01

Scope

Why owning a home in the United States has become more unattainable?

We chose this problem because of its growing impact on our generation, and so that we could analyze it. Home prices have increased much faster than incomes.

Gen Z and Millennials are currently affected the most by changes in housing prices as a growing number of people cannot afford to buy homes

How has housing affordability in the United States has evolved over time?

Big Data Dunces

02.

Datasets

Overview of Datasets

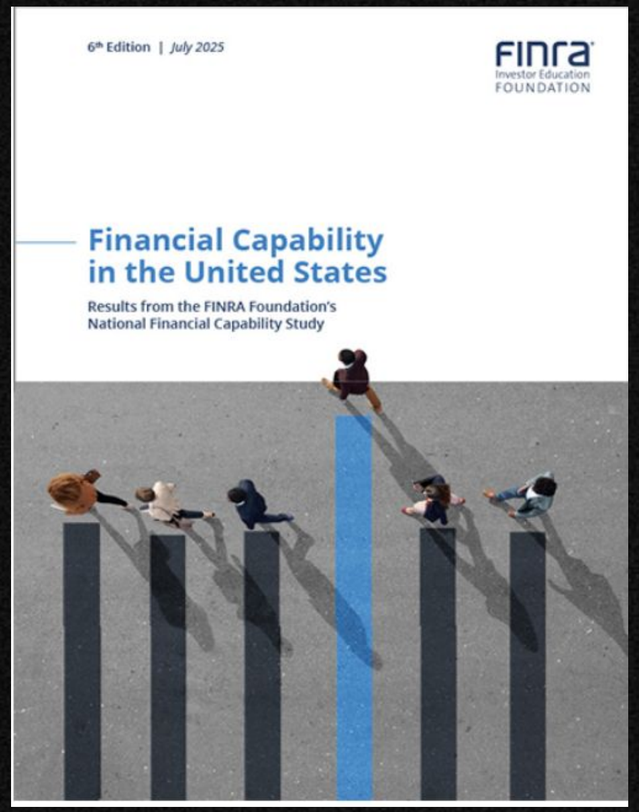
- **Money Management and Financial Literacy**
- **2024 National Financial Capability Study**
- Regression Dataset for Household Income Analysis
- New York Housing Market
- California Housing Prices
- Median Household Income
- Median House Price

1) Money Management and Financial Literacy

- From the District of Columbia
- Created April 30, 2024
- Last updated April 23, 2025
- About services that help with money management, financial planning, and insurance education

2) 2024 National Financial Capability Study

- From the FINRA Foundation
  - Financial Industry Regulatory Authority
- Started in 2009, every 3 years
- Last updated June 30, 2025
- Designed to understand and measure a connected set of perceptions, attitudes, experiences and behaviors to provide a comprehensive view of financial capability across the United States



3) Regression Dataset for Household Income Analysis

- From Kaggle
- Last updated June 4, 2024
- This synthetic dataset simulates various demographic and socioeconomic factors that influence annual household income.

4) New York Housing Market Analysis

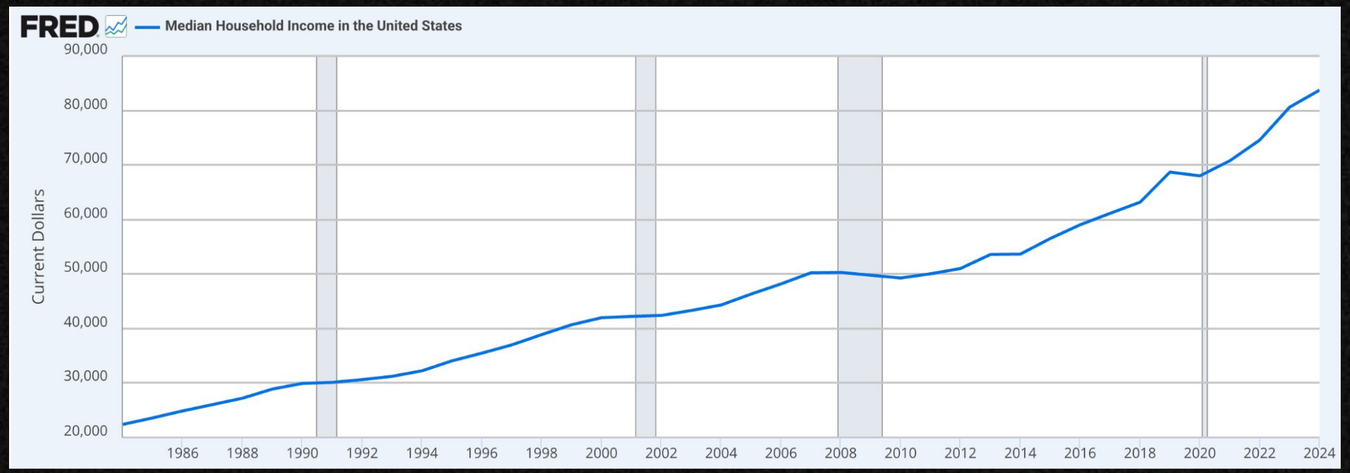
- From Kaggle
- Last updated 2 years ago
- Contains prices of New York houses, providing valuable insights into the real estate market in the region

5) California Housing Prices

- From Kaggle but originally from “Hands-On Machine learning with Scikit-Learn and TensorFlow” by Aurélien Géron
- Median house prices for California districts

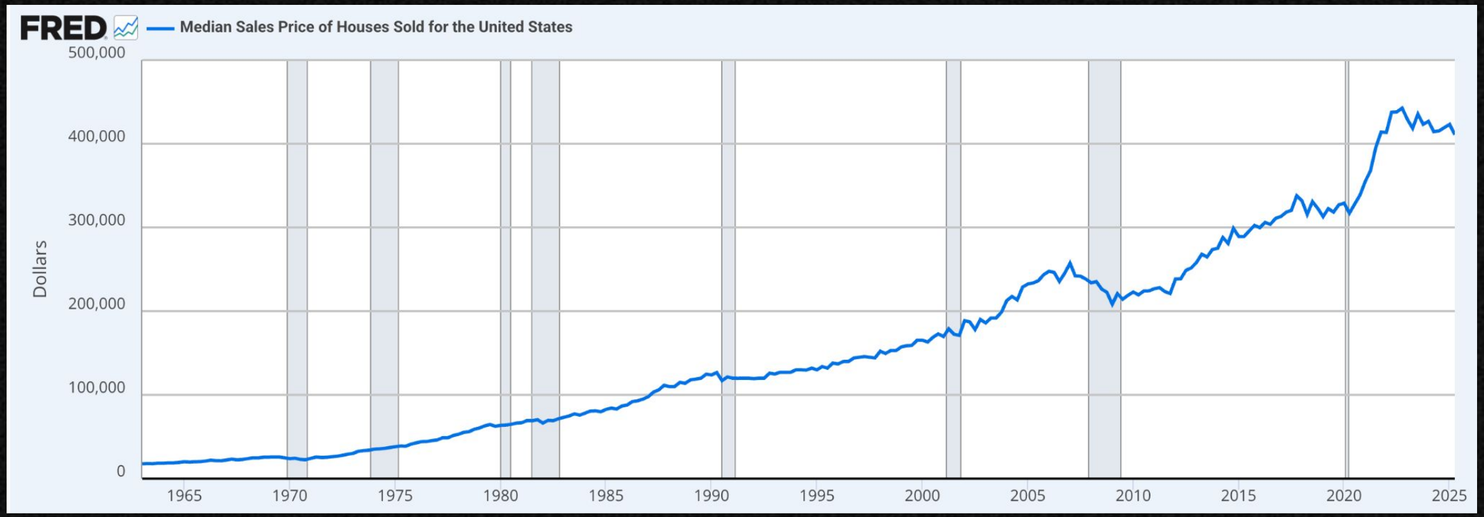
6) Median Household Income in the U.S.

- From Federal Reserve Bank of St. Louis
  - Sourced from U.S. Census Bureau
- Last updated Sep 9, 2025
- Estimation of median incomes using Pareto's linear interpolation to estimate all medians
  - A method to estimate the median or other values by looking at how much of the population falls below specific threshold, providing a more complete picture from **limited data**



7) Median House Price in the U.S.

- From U.S. Census Bureau and U.S. Department of Housing and Urban Development
- Last updated Jul 24, 2025
- Median sales price of houses sold for the United States

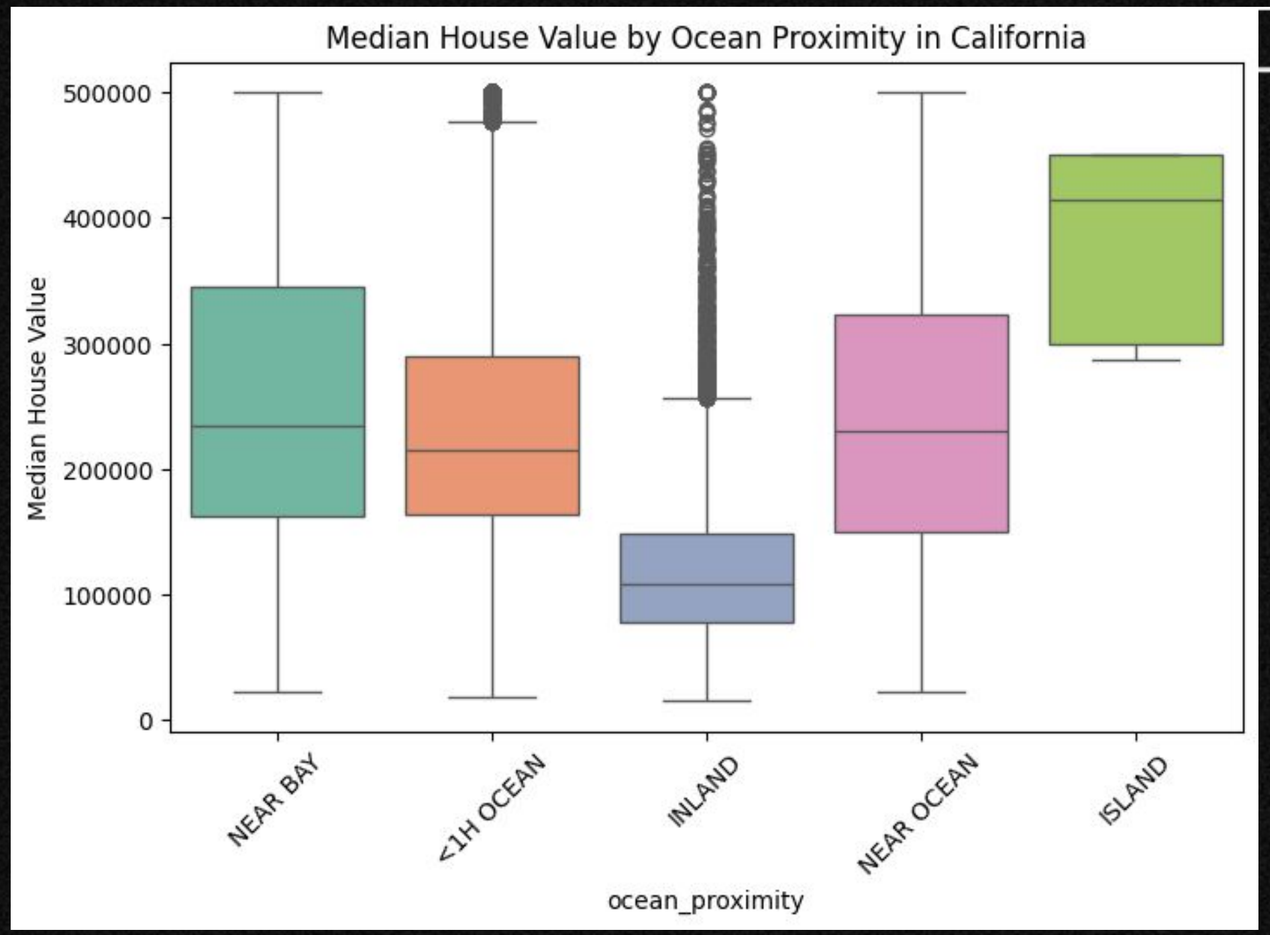


EDA

03.

Median House Value by Ocean Proximity in California

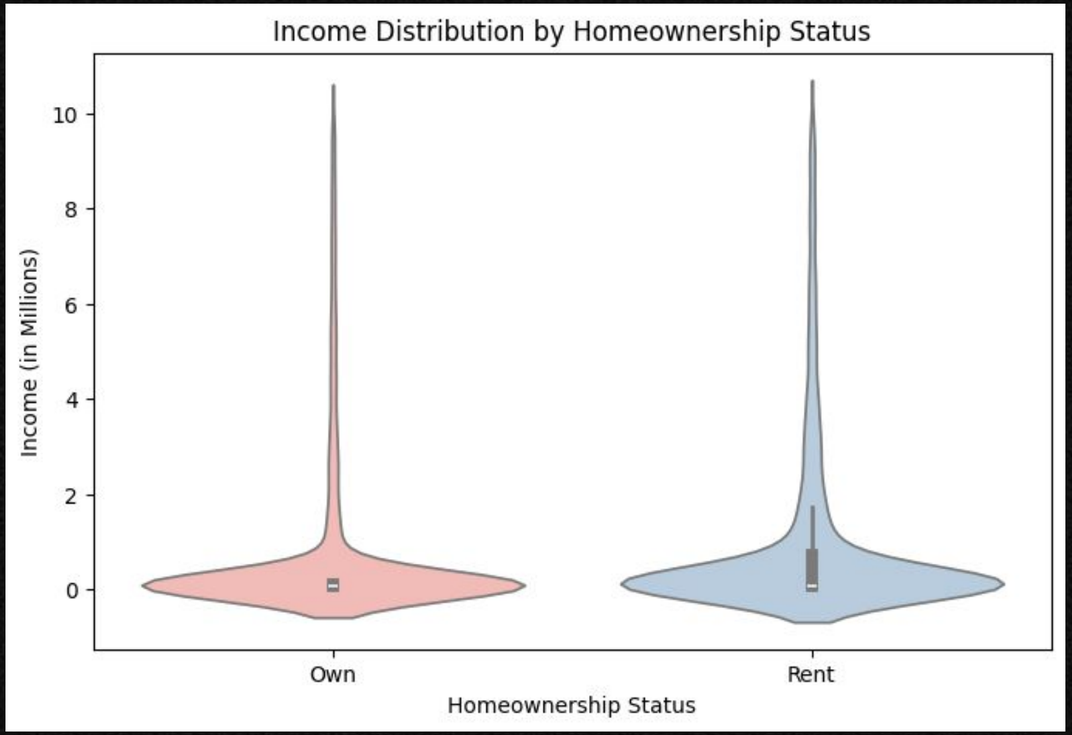
Inland areas show lower, more compact price distributions, highlighting a strong coastal premium in California housing



New York: Property Size vs Price

Location and neighborhood effects influence prices more than just size

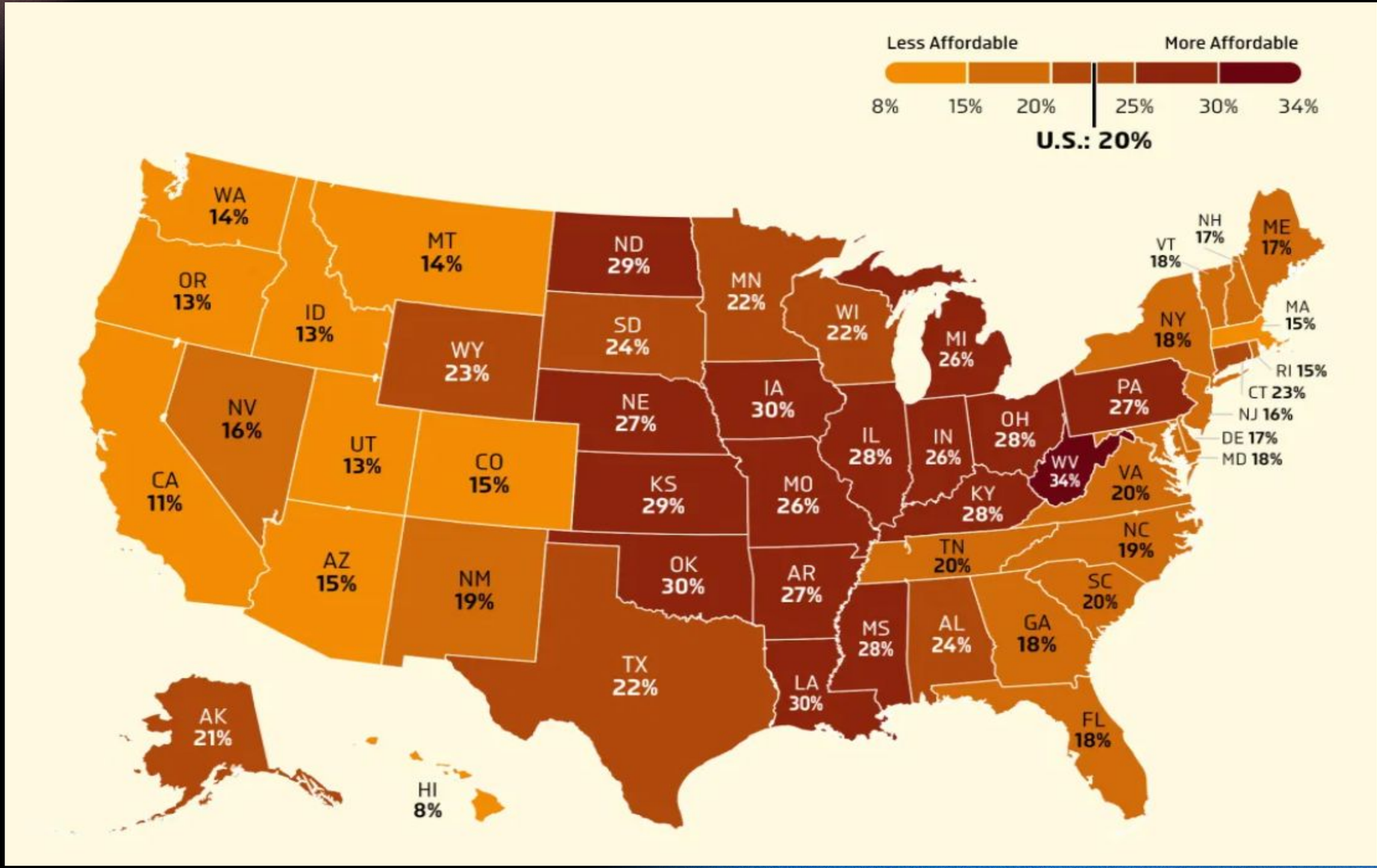
Income Distribution by Homeownership Status



Implies that income alone doesn't fully determine homeownership. Other factors like debt, credit, or savings likely play important roles

Where Americans can still Afford Homes

*Annual Income as Percentage of Home Price



Big Data Dunces

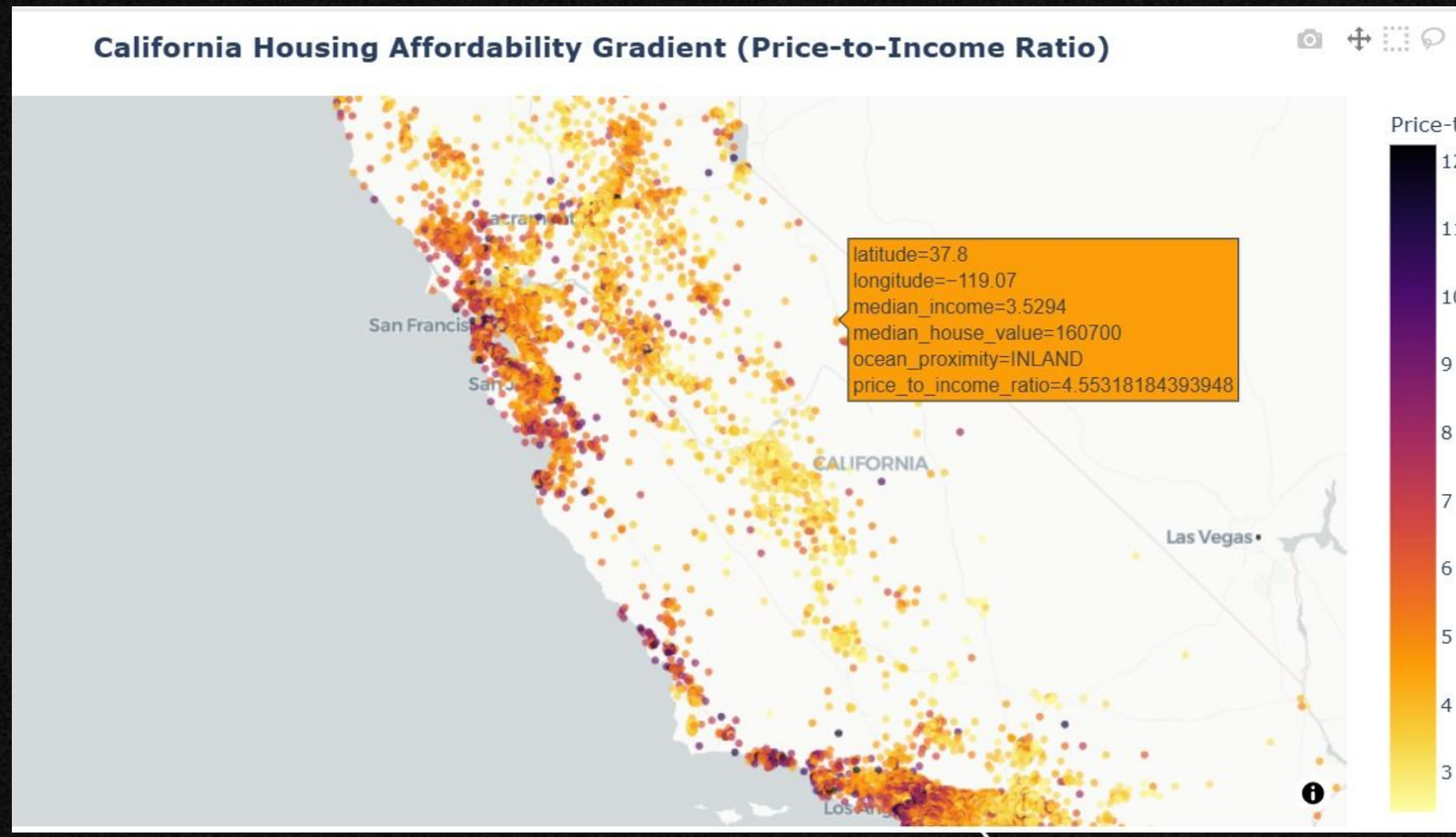
04.

Hypothesis

Hypothesis Claim 1

Coastal regions show higher price-to-income ratios, indicating low affordability.

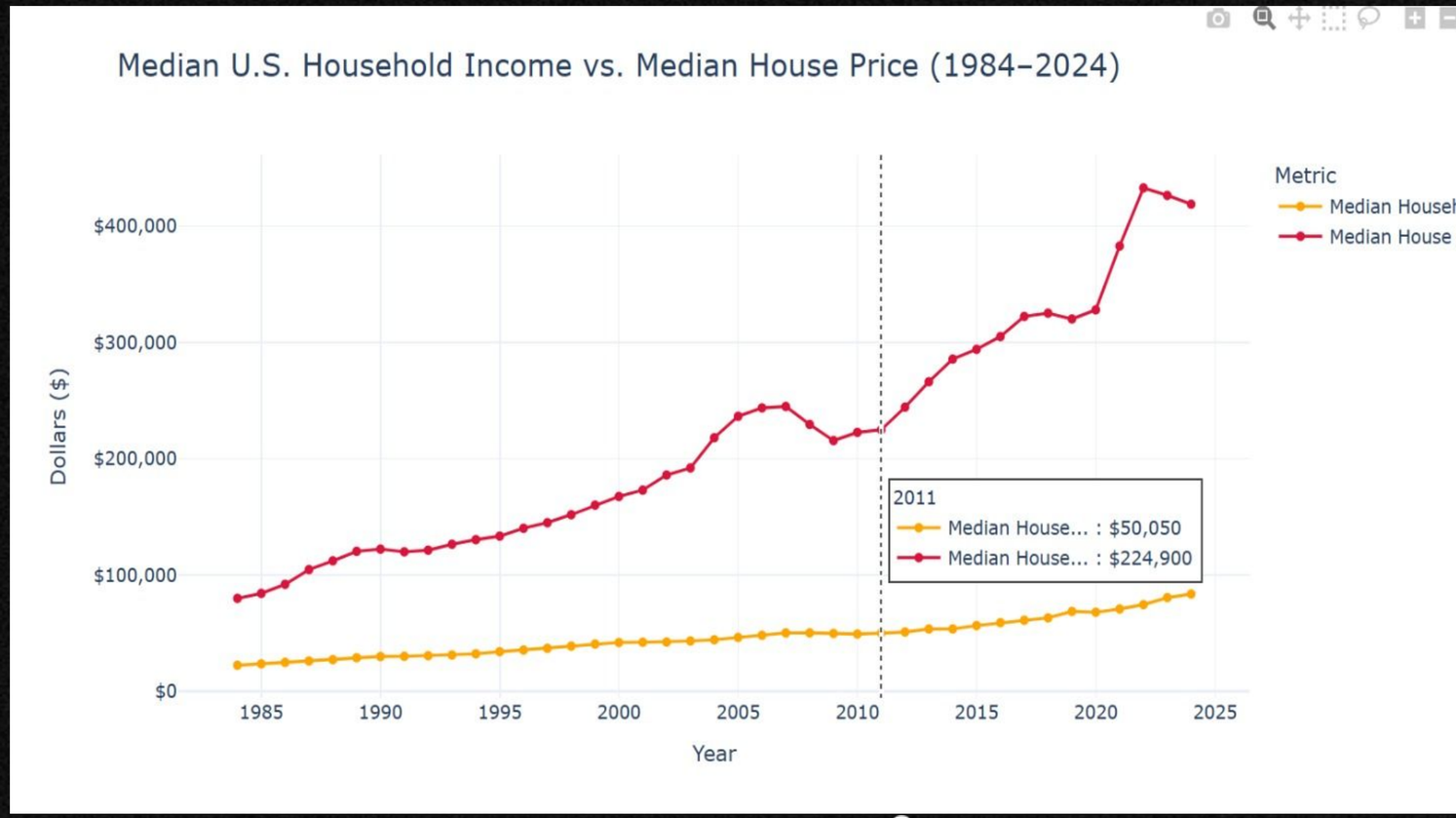
The closer a home is to California's coast, the higher its price-to-income ratio is; indicating that homes near coastal areas have become increasingly unaffordable relative to all income levels.



Hypothesis Claim 2

Home prices have grown nearly 4 times faster than incomes

Median household income in the United States has not kept pace with rising median home prices over the past 40 years, making homeownership increasingly unaffordable for the average household



Big Data Dunces

05.

ML

Accuracy and Baseline

- Logistic regression with class weighting achieved 63.2% accuracy, improving over the 59.6% majority baseline.
- Class performance is asymmetric: the model identifies owners well (precision = 0.67, recall = 0.77), but under-detects renters (precision = 0.55, recall = 0.42)
- This indicates useful signal in the features, but also a bias toward the majority class.
- Adding engineered income features did not increase accuracy, suggesting that affordability is influenced by unobserved variables such as regional housing costs or credit constraints.

Conclusion

The **hardest part** of the project so far has been working with multiple datasets that differ in scale and structure. Aligning income data, housing markets, and demographic attributes required careful cleaning and interpretation.

Our results **successfully answered** our research questions: Identified generational ownership gaps. Quantified long-term affordability decline. Linked education, occupation, and marital status to financial outcomes.

Homeownership has become **less attainable** for the average American. Wage stagnation and **rising property costs** are widening the gap. Data-driven insight like this helps policymakers target affordability reforms.

These insights emphasize the growing affordability crisis and **regional economic disparities** across the U.S

Big Data Dunces

THANK YOU!

David Yanjaa Adnan Sufyan Philip