Deciphering the Californian Dreamscape: Data-Driven Insights into Housing Market Dynamics

Andrew Boyle, Eric Sun, Feiyang Jiang, Jerry Wu, Yaotian Wu
Mathematics, University of California, San Diego

2024.3.16

Abstract

Housing prices may vary depending on various factors such as its location and the median income of households. Analyzing patterns between housing prices and various other factors can lead to more investment opportunities. We used a dataset containing the 1990 California Census Data, featuring 20640 observations across block groups and 10 different metrics to explore correlations and patterns between housing prices and other variables, aiding investment decisions, real estate development strategies, resource allocation, and market analysis. Utilizing careful preprocessing, exploratory data analysis, and modeling techniques like clustering, linear regression, and logistic regression, we can explore some insights into California's housing market dynamics, guiding both investors and policymakers to predict housing prices for informed decision-making.

1 Introduction

The California housing market presents a fascinating landscape shaped by a myriad of socioeconomic, geographical, and demographic factors. Understanding the intricate dynamics governing housing prices in California is imperative for various stakeholders, including real estate developers, investors, policymakers, and urban planners. In this comprehensive project report, we embark on an in-depth exploration of California's housing market, employing a multifaceted analytical approach to unravel the underlying patterns and determinants influencing housing prices.

Through a combination of hypothesis testing, exploratory data analysis (EDA), clustering analysis, and predictive modeling techniques, we aim to dissect the complex interactions between various variables and median housing prices across different regions of California. Our endeavor begins with formulating hypotheses and conducting hypothesis tests to examine the relationships between key factors such as geographical location, ocean proximity, median housing age, and median household income with housing prices. Leveraging advanced statistical methodologies, we delve into the nuances of California's housing market, seeking to uncover hidden insights and actionable intelligence.

As we delve deeper into the analysis, we employ clustering algorithms to segment California's regions based on similar housing market dynamics, providing spatial insights into the distribution of housing prices and market trends. Furthermore, our predictive modeling efforts utilize linear regression and logistic regression techniques to forecast housing prices and classify regions based on price ranges, respectively. Through meticulous data preprocessing, feature engineering, and model evaluation, we strive to develop robust predictive models that capture the essence of California's housing market dynamics.

Ultimately, this project report aims to offer valuable insights and practical recommendations for stakeholders navigating the California housing market landscape. By leveraging data-driven analytics and sophisticated modeling techniques, we endeavor to provide a holistic understanding of the factors driving housing prices, empowering stakeholders to make informed decisions and navigate the complexities of the ever-evolving housing market in California.

2 Data Description

Repository website https://www.kaggle.com/datasets/shibumohapatra/house-price

Brief Information The problem statement for our project revolves around understanding the factors influencing housing prices in California. The data used for this project will be acquired from a Kaggle dataset that summarizes the 1990 California Census Data from the US Census Bureau. This dataset contains 20640 observations along with 10 different types of metrics for each block group of houses in California: longitude, latitude, median age of the house in the block, total number of rooms (excluding bedrooms) in all houses in the block, total number of bedrooms in all houses in the block, total population in the block, median of the total household income of all the houses in the block, proximity to the ocean, and median of household prices of all houses in the block.

In our heatmap below, we found the most significant finding to be the positive correlation between median income and median house value at 0.69, indicating that higher income is associated with higher house values. We also found that the location of the houses affected the house prices as the latitude of houses had a negative correlation with house values, meaning that houses that were further south had higher median house values. In addition, our map generated by the folium library in Python provides evidence that the coastal zones of California especially the Bay Area and the Southern Coastline of Los Angeles had red markers which indicated that the median house values of $500,000 or higher. As we look more inland, the markers become blue and green which indicate more moderate housing values that are often below $300,000.

Example rows of dataset:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_proximity	median_house_value
0	-122.23	37.88	41	880	129.0	322	126	8.3252	NEAR BAY	452600
1	-122.22	37.86	21	7099	1106.0	2401	1138	8.3014	NEAR BAY	358500
2	-122.24	37.85	52	1467	190.0	496	177	7.2574	NEAR BAY	352100
3	-122.25	37.85	52	1274	235.0	558	219	5.6431	NEAR BAY	341300
4	-122.25	37.85	52	1627	280.0	565	259	3.8462	NEAR BAY	342200

Figure 1: A glimpse of dataset

3 Literature Review

Before the analysis, we looked into some previous research and works that have been done on this California Housing Dataset. There are other studies on housing prices, with many focusing on the influence of single specific features on prices within particular cities. For instance, Conroy and Milosch conducted a study in 2009 on housing prices in San Diego county and found the relationship between ocean proximity and housing price to be positively linearly associated. Others have studied and proposed theories on features affecting housing prices through analyzing patterns in other parts of the world. For instance, Selim used hedonic pricing and machine

learning techniques like neural networks to propose theory on the positive correlation between number of rooms and housing price in 2001, particularly on housing prices in Turkey. In our study, we are going to build upon those insights by taking possible features that might affect housing prices as mentioned in theories proposed in many of those studies, and conduct analysis on how well those features work in housing patterns in California in a more holistic view. Particularly, we review the impact of different features on general housing price in California first using hypothesis tests with multivariate linear regression, then gradually dive in to examine how well those theories generalize to smaller regional subgroups in California through techniques such as clustering and regressions. This comprehensive analysis gives a more nuanced understanding of the dynamics of housing prices across various regions in California.

Descriptive Analysis The research explores housing trends in specific California districts by describing various features of the dataset, such as median house value, median income, median age, total rooms, total bedrooms, population, households, latitude, longitude, and distances to coast and major cities.

Exploratory Data Analysis (EDA) The research employs EDA techniques to understand the structure of the data and uncover patterns and trends within it. This includes visualizations such as scatter plots, histograms, pair plots, and 3D scatter plots to explore relationships between different variables.

Correlation Analysis The research calculates correlation coefficients between different numeric variables to understand the relationships between them. For example, it examines the correlation between median house value and other variables such as median income, median age, and geographic coordinates.

Visualization Various visualization techniques, such as histograms, count plots, pair plots, and 3D scatter plots, are used to visually explore the data and relationships between different variables.

Geospatial Analysis The research utilizes geographic coordinates (latitude and longitude) to explore spatial patterns and relationships between housing variables across different cities in California.

Feature Comparison Scatter plots are used to compare median house value with other numeric features, providing insights into how these features are related to housing prices.

Data Cleaning and Handling Missing Data The research visualizes missing data using a heatmap to identify any missing values in the dataset, which is an essential step in data preprocessing.

Exploratory Data Analysis for California Housing Here is a example that have done an quite interesting EDA on California Housing data: Exploratory Data Analysis for California Housing — Kaggle.

4 Data Imputation

In addressing the inherent complexities of data preparation for the California housing market analysis, a meticulous approach was adopted to manage the missing values within the dataset, particularly for the **total_bedrooms** feature. Acknowledging the criticality of this variable in

housing studies, and to preserve the statistical validity of our analysis, a methodologically robust strategy of predictive imputation was deployed, leveraging linear regression techniques over simplistic mean or median imputations. This decision was predicated on the reasonable hypothesis that the number of bedrooms in a dwelling is predictably influenced by the total number of rooms and the count of households, thus warranting their use as independent variables in our regression model.

The operationalization of this imputation involved curating a training dataset devoid of missing values, which was imperative for establishing a well-fitted linear regression model. Subsequent to the training phase, the model exhibited its utility by accurately estimating the missing **total_bedrooms** data, effectively filling the gaps in our dataset. The imputation's efficacy was validated through a comprehensive post-imputation analysis, which confirmed the eradication of missing values from the **total_bedrooms** variable.

Employing this advanced imputation method not only enriched the dataset for heightened analytical accuracy but also fortified the foundational data upon which the housing market dynamics could be scrutinized. This rigorous treatment of missing data underscores our commitment to methodological excellence, ensuring that the inferences drawn from our study are both reliable and insightful, thereby contributing a layer of scientific rigor to the exploration of the determinants of housing prices in California.

5 Data Preprocessing

Before analysis, data preprocessing steps like imputation, encoding, and scaling were meticulously applied to enhance model performance. Specifically, missing values in the **total_bedrooms** column were addressed through imputation, leveraging linear regression with correlated columns to predict and fill in gaps, ensuring a complete dataset without discarding valuable information. Categorical variables, notably "ocean proximity," underwent one-hot encoding, converting them into a set of binary variables that maintain the original information while making it digestible for machine learning algorithms. Furthermore, to mitigate the influence of outliers and prevent features with larger scales from overshadowing others, scaling methods, particularly standardization, were employed across the dataset. This comprehensive preprocessing ensures a level playing field for all features, facilitating more accurate and insightful model analysis.

6 Exploratory Data Analysis

Folium Geospatial Plot From the coastal regions to the inland areas, there is a clear gradation in property values. The coastal zones, particularly in the Bay Area and along the southern coastline near Los Angeles, are marked predominantly with red markers, indicating median house values of $500,000 or higher. This reflects the well-known high cost of living in these regions. As the visualization moves inland, the prevalence of red markers diminishes, giving way to blue and green markers, signifying more moderate housing values, often below $300,000, which is characteristic of California's inland real estate market. The Central Valley shows a mixed pattern, with a greater number of affordable housing options, as denoted by the green markers. It is also evident that the density of markers correlates with population density, with urban centers showing a higher concentration of markers and thus a broader range of housing values. This map serves as a graphical illustration of the diverse housing market across California, with a stark contrast between the affluent coastal cities and the more modestly priced inland regions.

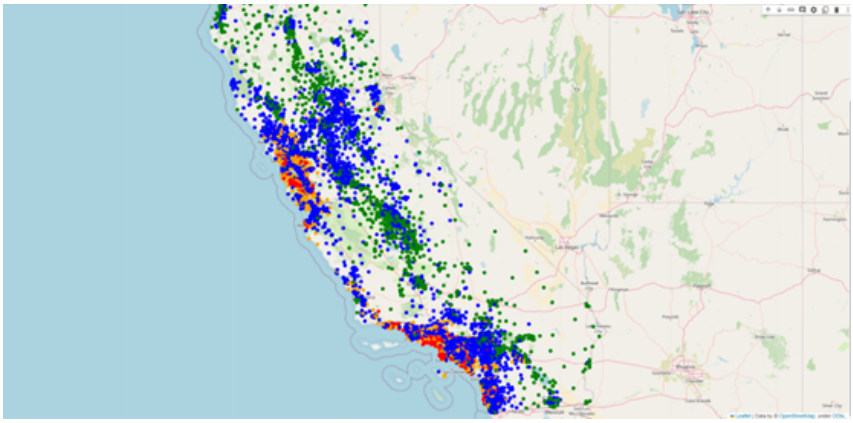

Figure 2: Choropleth for California Housing

Correlation Heatmap The most significant finding is the positive correlation between median income and median house value (0.69), indicating that higher income is associated with higher house values. Latitude shows a slight negative correlation with house value, suggesting that house values decrease as one moves north. Other factors such as the number of rooms or households show strong interrelations but do not significantly correlate with house values, highlighting the complex nature of real estate pricing.

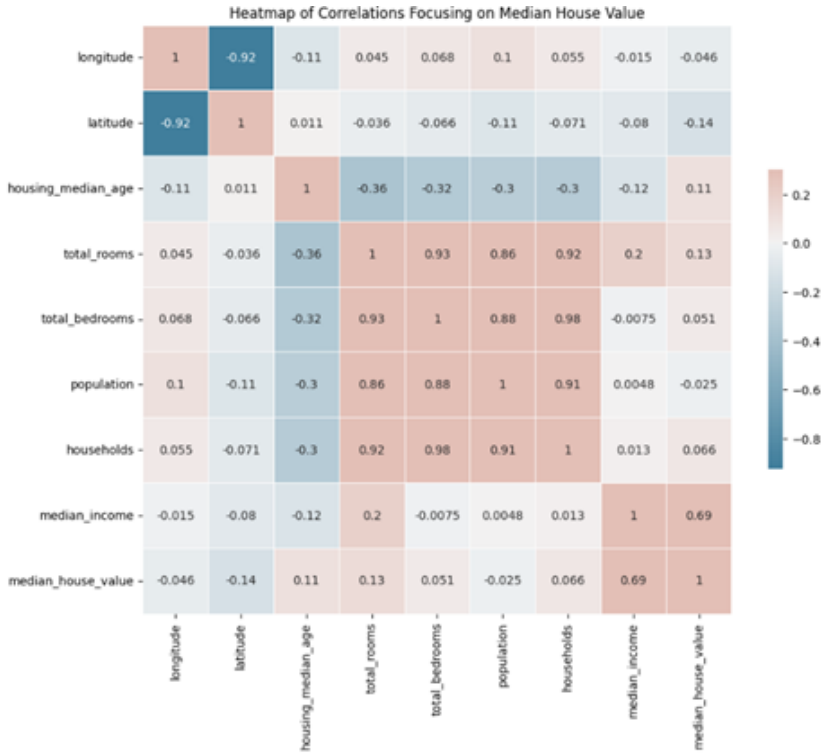

Figure 3: Feature Heatmap

Violin Plot The plots show variation in age distribution, with 'NEAR BAY' and 'NEAR OCEAN' having wider spreads, indicating a diverse range of housing ages. In contrast, 'ISLAND' has a narrow plot, suggesting less variation in housing ages. The box plots within each violin reveal the medians and interquartile ranges. The removal of graph spines creates a clean visual

presentation.

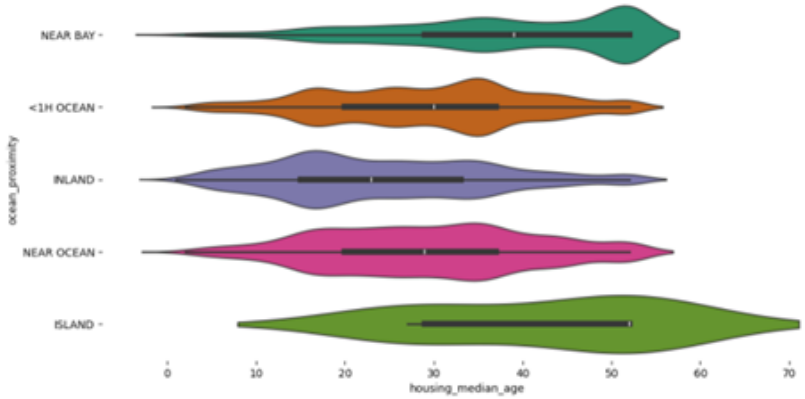

Figure 4: Violin Plot on Ocean Proximity & Median Age

Horizontal Barplot The bar chart illustrates the count of housing units by their proximity to the ocean. 'INLAND' areas have the highest number of units, significantly more than other categories, while 'ISLAND' areas have the fewest. 'NEAR OCEAN' and '<1H OCEAN' also have a substantial number of housing units, with '<1H OCEAN' being the second most populated category. The 'NEAR BAY' area has a moderate count. The horizontal orientation of the bars and the palette choice enhance readability, and the absence of 'top' and 'right' spines offers a cleaner look to the presentation.

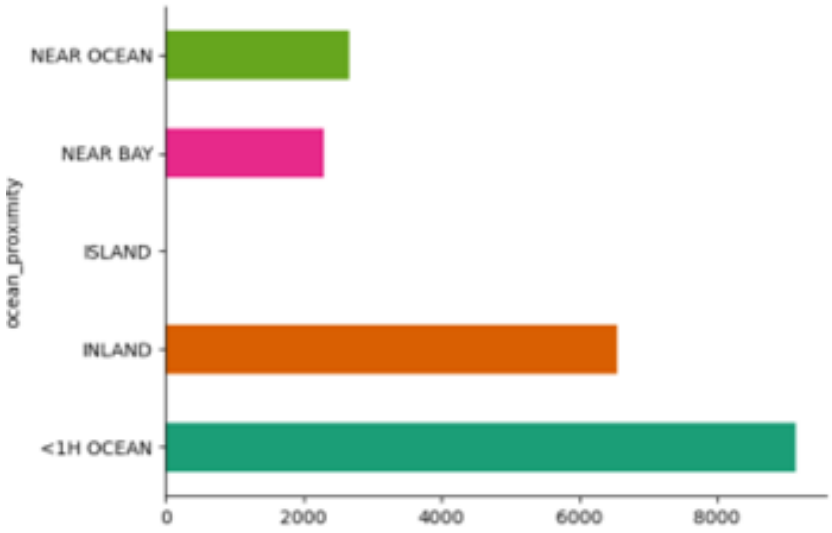

Figure 5: Bar Plot on categories of Ocean Proximity

7 Model Implementation

7.1 Hypothesis Testing

We wanted to see if Longitude, Latitude, Ocean Proximity, and median housing age affects the median housing value. We conducted a mixture of multiple and single linear regressions and performed five hypothesis tests.

- Longitude, Latitude vs. Median House Value
- House located on Island vs. Median House Value
- House located near the ocean or near the bay vs. Median House Value
- House located inland or less than 1 hour away from the ocean vs. Median House Value
- Housing Median Age vs. Median House Value

7.1.1 Hypothesis Test To See If Longitude And Latitude Affects Median House Value

- Null Hypothesis (H_0): Longitude and latitude will not affect the **median_house_value**.
- Alternative Hypothesis (H_1): Longitude and latitude will affect **median_house_value**.

```
==============================================================================
Dep. Variable:     median_house_value   R-squared:                       0.242
Model:                            OLS   Adj. R-squared:                  0.242
Method:                 Least Squares   F-statistic:                     3302.
Date:                Sat, 09 Mar 2024   Prob (F-statistic):               0.00
Time:                        21:40:07   Log-Likelihood:                -24642.
No. Observations:               20640   AIC:                         5.285e+04
Df Residuals:                   20637   BIC:                         5.287e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -52.3105      0.711    -73.530      0.000    -53.705    -50.916
longitude     -0.6171      0.008    -77.702      0.000     -0.633    -0.602
latitude      -0.6027      0.007    -80.908      0.000     -0.617    -0.588
==============================================================================
Omnibus:                     2567.179   Durbin-Watson:                   0.406
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3642.564
Skew:                           0.973   Prob(JB):                         0.00
Kurtosis:                       3.671   Cond. No.                     1.47e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.47e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The coefficients of longitude and latitude are -0.6171 and -0.6027 respectively, meaning that longitude and latitude both negatively influences the median house price. The p-value is also 0.000 for both, meaning that they are both statistically significant at the 5% level. Therefore, we reject the null hypothesis and conclude that longitude and latitude will affect the median house value.

7.1.2 Hypothesis Test To See If Houses Located on an Island Affects Median House Value

- Null Hypothesis (H_0): Houses located on an island will not affect its **median_house_value**.
- Alternative Hypothesis (H_1): Houses located on an island will affect **median_house_value**.

```
==============================================================================
Dep. Variable:     median_house_value   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     11.32
Date:                Sat, 09 Mar 2024   Prob (F-statistic):           0.000763
Time:                        21:40:07   Log-Likelihood:                -29281.
No. Observations:               20640   AIC:                         5.857e+04
Df Residuals:                   20638   BIC:                         5.858e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0004      0.007     -0.052      0.958      -0.014      0.013
ISLAND         1.5047      0.447      3.365      0.001       0.628      2.381
==============================================================================
Omnibus:                     2434.269   Durbin-Watson:                   0.338
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3386.334
Skew:                           0.978   Prob(JB):                         0.00
Kurtosis:                       3.330   Cond. No.                         64.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

The coefficient of island is +1.5047, meaning that a house located on the island increases the median house price by 1.5047. The p-value is also 0.001, meaning that the island coefficient statistically significant at the 5% level. Therefore, we reject the null hypothesis and conclude that houses located on an island will affect the median house value.

7.1.3 Hypothesis Test To See If Being Near The Bay or Ocean Affects Median House Value

- Null Hypothesis (H_0): Being near the bay or ocean will not affect the **median_house_value**.
- Alternative Hypothesis (H_1): Being near the bay or ocean will affect **median_house_value**.

```
==============================================================================
Dep. Variable:     median_house_value   R-squared:                       0.053
Model:                            OLS   Adj. R-squared:                  0.053
```

```
Method:                 Least Squares   F-statistic:                     577.1
Date:                Sat, 09 Mar 2024   Prob (F-statistic):          1.29e-244
Time:                        21:40:07   Log-Likelihood:                -28725.
No. Observations:               20640   AIC:                         5.746e+04
Df Residuals:                   20637   BIC:                         5.748e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1287      0.008    -16.567      0.000      -0.144    -0.113
NEAR OCEAN     0.4977      0.020     24.381      0.000       0.458     0.538
NEAR BAY       0.5824      0.022     26.753      0.000       0.540     0.625
==============================================================================
Omnibus:                     2436.967   Durbin-Watson:                   0.329
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3384.027
Skew:                           0.968   Prob(JB):                         0.00
Kurtosis:                       3.432   Cond. No.                         3.44
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

The coefficients of 'NEAR OCEAN' and 'NEAR BAY' are +0.4977 and +0.5824 respectively, meaning that 'NEAR OCEAN' and 'NEAR BAY' both positively influence the median house price. The p-value is also 0.000 for both, meaning that they are both statistically significant at the 5% level. Therefore, we reject the null hypothesis and conclude that being near the ocean or being near the bay will affect the median house value.

7.1.4 Hypothesis Test To See If Being Inland or Less Than 1 Hour From Ocean Affects Median House Value

- Null Hypothesis (H_0): Being inland or less than 1 hour from ocean will not affect the **median_house_value**.
- Alternative Hypothesis (H_1): Being inland or less than 1 hour from ocean will affect **median_house_value**.

```
OLS Regression Results
==============================================================================
Dep. Variable:     median_house_value   R-squared:                       0.237
Model:                            OLS   Adj. R-squared:                  0.237
Method:                 Least Squares   F-statistic:                     3212.
Date:                Sat, 09 Mar 2024   Prob (F-statistic):               0.00
Time:                        21:40:07   Log-Likelihood:                -26490.
No. Observations:               20640   AIC:                         5.299e+04
Df Residuals:                   20637   BIC:                         5.301e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
```

```
const          0.4093      0.012     32.984      0.000       0.385       0.434
INLAND        -1.1204      0.016    -68.130      0.000      -1.153      -1.088
<1H OCEAN     -0.1213      0.015     -7.875      0.000      -0.152      -0.091
==============================================================================
Omnibus:                     2659.105   Durbin-Watson:                   0.413
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3832.931
Skew:                           0.987   Prob(JB):                         0.00
Kurtosis:                       3.749   Cond. No.                         4.37
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

The coefficients of **INLAND** and **<1H Ocean** are -1.1204 and -0.1213 respectively, meaning that **INLAND** and **<1H Ocean** both negatively influences the median house price. The p-value is also 0.000 for both, meaning that they are both statistically significant at the 5% level. Therefore, we reject the null hypothesis and conclude that being inland or less than 1 hour from the ocean will affect the median house value.

7.1.5 Hypothesis Test To See If The Median Housing Age Affects Median House Value

- Null Hypothesis (H_0): The Median Housing Age will not affect the **median_house_value**.
- Alternative Hypothesis (H_1): The Median Housing Age will affect **median_house_value**.

```
==============================================================================
Dep. Variable:     median_house_value   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     230.8
Date:                Sat, 09 Mar 2024   Prob (F-statistic):           2.76e-52
Time:                        21:40:08   Log-Likelihood:                -29171.
No. Observations:               20640   AIC:                         5.835e+04
Df Residuals:                   20638   BIC:                         5.836e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.735e-16    0.007   2.51e-14      1.000      -0.014      0.014
housing_median_age  0.1056      0.007     15.259      0.000       0.092      0.119
==============================================================================
Omnibus:                     2269.585   Durbin-Watson:                   0.325
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3093.615
Skew:                           0.938   Prob(JB):                         0.00
Kurtosis:                       3.281   Cond. No.                         1.00
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

The coefficient of **housing_median_age** is +0.1056, meaning that the median housing age increases the median house price by 0.1056. The p-value is also 0.000, meaning that the island coefficient statistically significant at the 5% level. Therefore, we reject the null hypothesis and conclude that the housing median age will affect the median house value.

7.2 Clustering Model

Firstly, the data was preprocessed to ensure it was suitable for clustering. Since clustering algorithms like K-Means require numerical input, categorical variables were one-hot encoded, and numerical variables were scaled or standardized if necessary. The median_house_value was excluded from clustering since it's the target variable for potential predictive modeling.

The K-Means algorithm was chosen for its efficiency and simplicity. It partitions the data into 'k' clusters, each with a centroid, by minimizing the variance within each cluster. To determine the optimal number of clusters 'k', the Elbow Method was employed. We computed the within-cluster sum of squares (inertia) for a range of cluster numbers and plotted them. The 'elbow' in the plot is where the inertia's rate of decrease sharply changes, suggesting the addition of more clusters does not significantly improve the fit of the model.

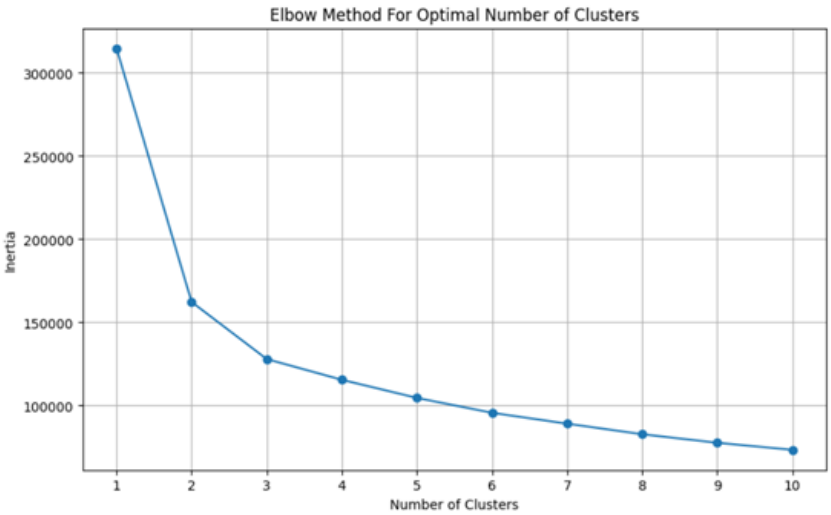

Figure 6: Elbow plot for optimal clusters

In the provided Elbow plot, although there was no sharply defined elbow, a subtle bend after 2 clusters indicated that 3 clusters could be optimal, balancing the model's complexity and the data's intrinsic structure. The choice of the number of clusters can also be informed by domain expertise and additional validation methods like the Silhouette Score.

Figure 7: PCA reduction Method

The final step was to visualize the clusters on a map to interpret the geographic distribution of the clusters. In the map visualization, data points are color-coded based on their cluster assignments.

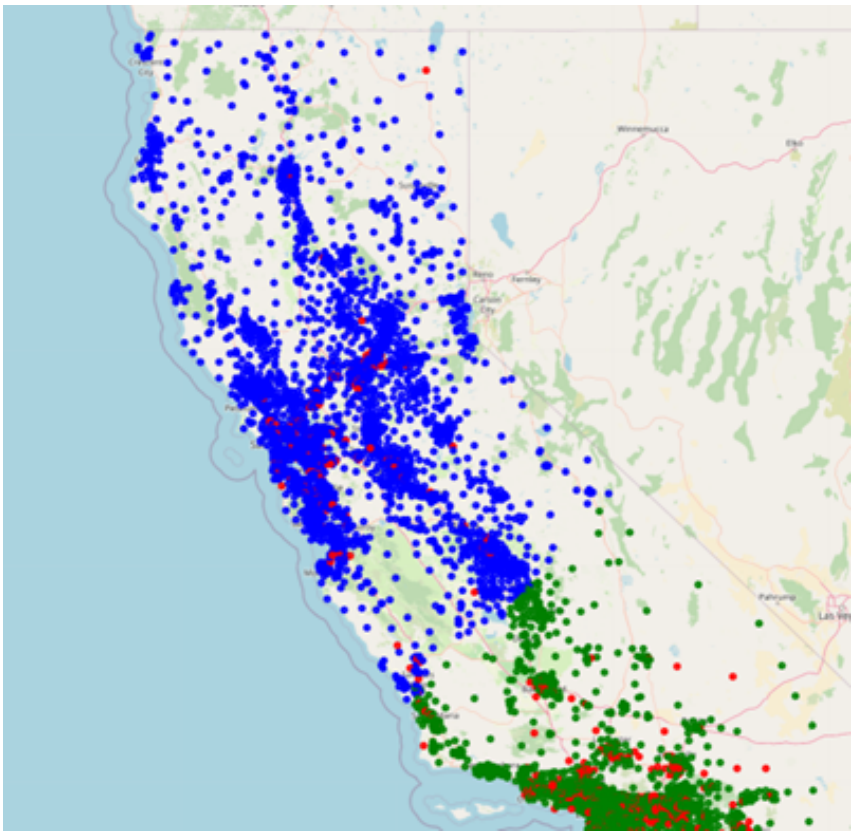

Figure 8: Clustering Map

Blue Cluster (Predominantly Northern Region) **Economic**: This cluster seems to be in less densely populated areas, possibly indicating rural or semi-urban settings. The economic activities here might be less diverse, with a possible focus on agriculture, manufacturing, or resource-based industries. **Social**: Housing in these areas might be more affordable, attracting people who prefer or require larger living spaces, such as families. The population density seems to be lower, which could translate into a different community lifestyle compared to urban centers. **Regional Planning**: The distribution could reflect regional development policies that encourage or restrict growth in certain areas, preserving natural lands or focusing on specific economic zones.

Red Cluster (Central Coastal Region, including Los Angeles) **Economic**: This cluster likely represents a high-density urban area with diverse economic opportunities, higher cost of living, and significant real estate development. It may include business districts, cultural hubs, and a mix of residential types. **Social**: Urban areas often have a higher concentration of services, amenities, and entertainment options, which attract younger populations, professionals, and smaller households that prefer urban living. **Urban Planning**: This cluster's location suggests a strategic focus on coastal development, where urban sprawl meets natural constraints, leading to higher housing densities and vertical development.

Green Cluster (Southern Region, including San Diego) **Economic**: This cluster may represent a mix of suburban and urban areas, potentially with a combination of residential neighborhoods, commercial zones, and possibly military bases given its proximity to San Diego. There could be a blend of employment opportunities from tech to border-related industries. **Social**: Suburban areas often offer a balance between access to urban jobs and amenities with the benefits of suburban living, such as more space and perceived safety. There might be a diverse population with a variety of household sizes and types. **Regional Planning**: These areas could be influenced by regional planning that promotes balanced development, accommodating population growth while preserving quality of life. It might also reflect the influence of cross-border economic dynamics with Mexico.

7.3 Linear Model on Price

In our analysis of California housing prices, we employed a linear regression model to delve into the factors influencing housing prices across various regions. Prior to constructing the model, we meticulously prepared our dataset by normalizing the data using the StandardScaler method. This normalization ensured that all features were on the same scale, preventing any one feature from dominating the model due to its larger magnitude. Additionally, we applied Principal Component Analysis (PCA) to reduce the dimensionality of the dataset while retaining essential information. This step was crucial in managing computational complexity and enhancing model interpretability.

By calculating the correlation matrix, we gained insights into which features exhibited stronger associations with housing prices. Leveraging this understanding, we performed feature selection to focus on the most influential variables for our predictive model. Features with low correlation to housing prices were dropped, allowing us to streamline our analysis and concentrate on the most relevant factors. With our refined dataset, we proceeded to build the linear regression model. This model aimed to predict housing prices based on the selected features. We trained the model using the training data and evaluated its performance on the test data using metrics such as Root Mean Squared Error (RMSE) and R-squared score. The RMSE provided us with a measure of the average difference between predicted and actual housing prices, while the R-squared score indicated the proportion of variance in housing prices explained by the model.

Upon fitting the model, we conducted a thorough analysis of the coefficients. Each coefficient represented the strength and direction of the relationship between a particular feature and housing prices. Positive coefficients indicated a positive influence on housing prices, while negative coefficients signified a negative influence. This analysis allowed us to discern the relative importance of each feature in determining housing prices across California.Our linear regression model provided valuable insights into the complex dynamics of housing prices in California. By uncovering the factors driving these prices, our analysis can inform real estate investment strategies, policy-making decisions, and urban planning initiatives.

7.4 Logistic classification

With some basic patterns in California housing uncovered through previous models, the results can be used to conduct a closer examination on which features might contribute more to the variation in housing price. Logistic classification is a common type of classification model and could help with identifying the influence of different features in houses of different price ranges in California, and the resulting model could serve as a price predictor that helps industry in making construction decisions. The median housing price is divided into 4 smaller groups based on their distribution, then the coefficient and odds for each feature on each of the smaller groups are examined, resulting in the conclusion that geographical location in terms of longitude and

latitude are the most determinant factors influencing the regional housing price, followed closely by regional median income, and lastly the proximity to ocean and the age of houses. Other features like housing size in terms of rooms and regional population density does not have as strong an impact on California housing prices as others.

Using that information, a price range predictor is also built using a logistic regression model with ridge regularization to handle multicollinearity and prevent overfitting, and trained with Limited-memory Broyden–Fletcher–Goldfarb–Shanno optimization on the dataset. With testing on the unseen subset of the dataset, the multi-class predictor achieved an overall accuracy of over 73%, and has best performance in detecting regions that would likely have housing prices between 300,000 to 500,000 with an 86% accuracy. The logistic regression model, with its interpretable coefficients, serves as a robust tool for predicting price ranges, aiding industry stakeholders in making informed decisions.

7.5 Model Evaluation

In our model evaluation process, we rigorously tested both linear and logistic regression models, employing specific metrics tailored to each model's nature to ensure a thorough analysis of their performance. For the linear regression model, we measured the goodness of fit using the R^2 metric, applying k-fold cross-validation to ensure reliability and mitigate overfitting risks. This method provided a comprehensive overview of how well the linear model could predict continuous outcomes, such as housing prices, based on its ability to capture the variance in the dataset. On the other hand, the logistic regression model, designed for classification tasks, was evaluated using accuracy to measure its success rate in categorizing homes into distinct price brackets. To deepen our understanding of its performance, we also incorporated the AUC (Area Under the Curve) and ROC (Receiver Operating Characteristic) curve analysis. These tools offered insights into the logistic model's capability to distinguish between different classification outcomes—true positives, true negatives, false positives, and false negatives—thereby assessing its predictive precision and robustness. An AUC score of 0.88 was particularly telling, signifying a high degree of accuracy in classifying housing prices, which not only confirmed the model's effectiveness but also established a benchmark for future model comparisons within the domain of housing data analysis. This multifaceted approach to model evaluation allowed us to validate the models' predictive power and ensure their reliability in real-world applications, enhancing confidence in their deployment for housing price predictions.

8 Conclusion

Our comprehensive investigation into the California housing market employed a multifaceted approach to decode the intricacies influencing housing prices, offering critical insights for stakeholders and policymakers. By conducting an exploratory data analysis (EDA), we illuminated the stark geographical disparities in median housing prices, pinpointing the Bay Area and the southern coastal regions near Los Angeles as hotspots for elevated housing costs. This geographic analysis was complemented by examining the age distribution of houses in relation to their proximity to the ocean, revealing that homes categorized as 'NEAR BAY' and 'NEAR OCEAN' displayed a broader range of ages when compared to those on 'ISLAND', indicating a diverse architectural timeline across locales. Our data preprocessing regimen included one-hot encoding for categorical variables and the application of standardization and normalization across all data features to refine our models' accuracy. Through hypothesis testing, we established a significant impact of factors like ocean proximity and household income on housing prices, further supported by a clustering analysis that identified four distinct groups for in-depth analysis. The deployment of regression models shed light on the nuanced relationship between

housing features and their market prices; the linear regression model underscored the positive influence of income and the detrimental impact of inland locations on prices, reflected by an RMSE of 0.64 and an R^2 value of 0.57. The logistic regression model's prowess in categorizing homes into four price brackets, with an average accuracy of 73%, emphasizes the pivotal roles of geographic location, income, and property age in the housing valuation process. These revelations not only guide real estate development and investment decisions but also pave the way for future inquiries into advanced feature engineering and modeling techniques to capture the housing market's complexity more effectively.

References

Hasan Selim, "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network," *Expert Systems with Applications*, vol. 36, no. 2, 2009.

Marlon G. Boarnet and Saksith Chalermpong, "New highways, house prices, and urban development: A case study of toll roads in Orange County, CA," *Housing Policy Debate*, vol. 12, no. 3, 2001.