California Housing Prices
Exploratory Analysis

Michael Zelaya, mzelaya@bellarmine.edu
Radjamin Hukom, rhukom@bellarmine.edu

I. INTRODUCTION

Our data set is about California housing prices from the 1990 California census. We found this data set on Kaggle (https://www.kaggle.com/datasets/camnugent/california-housing-prices). We chose this data set because we thought it looked interesting and wanted to explore the housing prices based on income in California even further.

II. DATA SET DESCRIPTION

Narrative summary of the data set: This data set contains 20,641 samples with ten columns, with most data types being quantitative and just one column being categorical.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Longitude	Ratio and Quantitative	0%
Latitude	Ratio and Quantitative	0%
Housing Median Age	Ratio and Quantitative	0%
Total Rooms	Ratio and Quantitative	0%
Total Bedrooms	Ratio and Quantitative	0%
Population	Ratio and Quantitative	0%
Households	Ratio and Quantitative	0%
Median Income	Ratio and Quantitative	0%
Median House Value	Ratio and Quantitative	0%
Ocean Proximity	Ordinal and Categorical	0%

III. Data Set Summary Statistics

- Longitude: A measure of how far west a house is; a higher value is farther west.
- Latitude: A measure of how far north a home is; a higher value is farther north.
- Housing_median_age: The median age of a house within a block; a lower number is a newer building.
- Total_rooms: Total number of rooms within a block.
- Total_bedrooms: Total number of bedrooms within a block.
- Population: Total number of people residing within a block.
- Households: Total number of families, a group living within a home unit for a block.
- Median_income: Income for households within a block of houses (measured in tens of thousands of US Dollars).
- Median_house_value: Median house value for homes within a block (measured in US Dollars).
- Ocean_proximity: Location of the house w.r.t ocean/sea.

Table 2: Summary Statistics for California Housing Prices

Variable Name	*Count*	*Mean*	*Min*	*25^{th}*	*50^{th}*	*75^{th}*	*Max*
Longitude	*20,640*	*-119.6*	*-124.3*	*-121.8*	*-118.5*	*-118.0*	*-114.3*
latitude	*20,640*	*35.63*	*32.54*	*33.93*	*34.26*	*37.71*	*41.95*
Housing Median Age	*20,640*	*28.64*	*1.00*	*18.00*	*29.00*	*37.00*	*52.00*
Total Rooms	*20,640*	*2636*	*2*	*1448*	*2127*	*3148*	*39,320*
Total Bedrooms	*20,640*	*536.8*	*1.0*	*297.0*	*435.0*	*643.2*	*6445.0*
Population	*20,640*	*1425*	*3*	*787*	*1166*	*1725*	*35682*
Households	*20,640*	*499.5*	*1.0*	*280.0*	*409.0*	*605.0*	*6082.0*
Median Income	*20,640*	*3.8707*	*0.4999*	*2.5634*	*3.5348*	*4.7432*	*15.0001*

Median House Value	*20,640*	*206,856*	*14999*	*119600*	*179700*	*264725*	*500,001*

There should be a table for **EACH** categorical variable.

Table 3: Proportions for California Housing Prices near ocean_proximity

Ocean Proximity

Category	*Frequency*	*Proportion (%)*
<1H Ocean	*9136*	*9136/20,640 = 44.26%*
Inland	*6551*	*6551/20,640=31.74%*
Island	*5*	*5/20,640= .02%*
Near Bay	*2290*	*2290/20,640=11.09%*
Near Ocean	*2658*	*2658/20,640=12.88%*

IV. EDA FINDINGS

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalances, etc., that you notice.

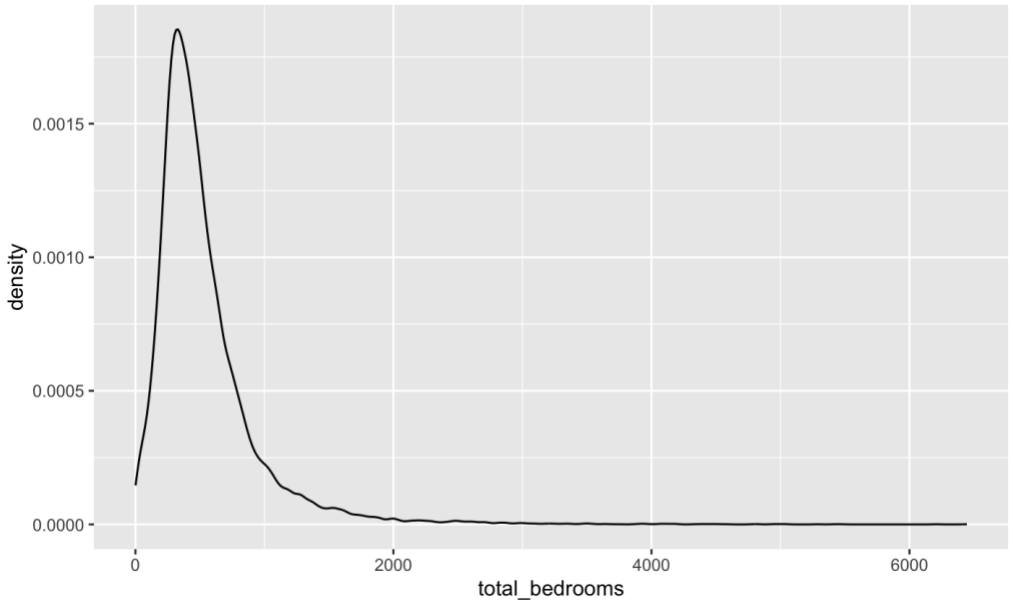

Figure 1: Comparison of total_bedrooms by density

This density curve demonstrates a right-skewed of total bedrooms by population. As you can see, the median of complete bedrooms is approximately 435, which makes the median less than the mean.

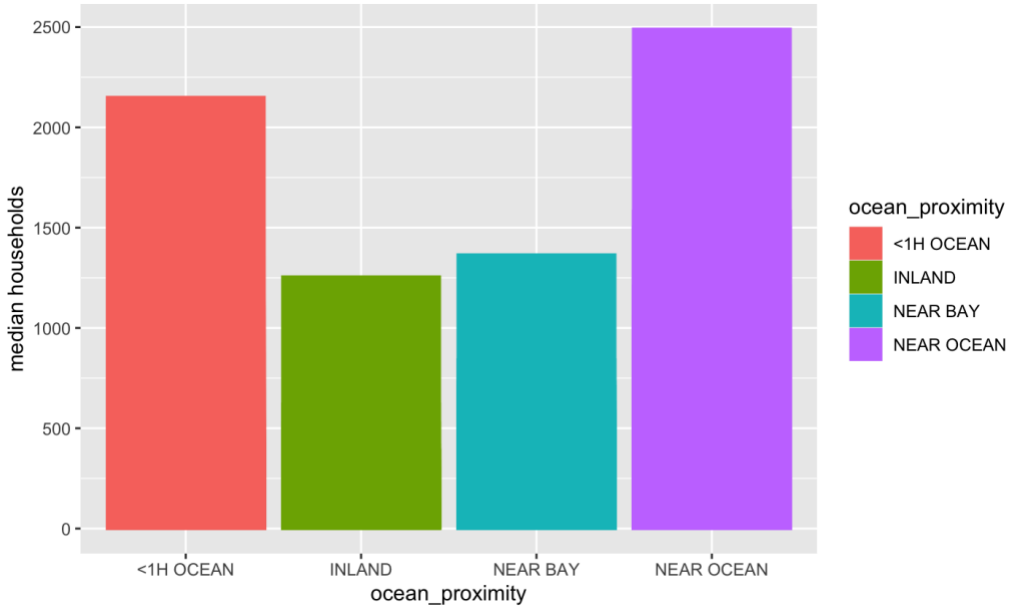

Figure 2: Ocean proximity by median households.

This bar graph compares the ocean proximity by the median of households for the data set. This graph demonstrates how more families live near the ocean than other housing locations on the median.

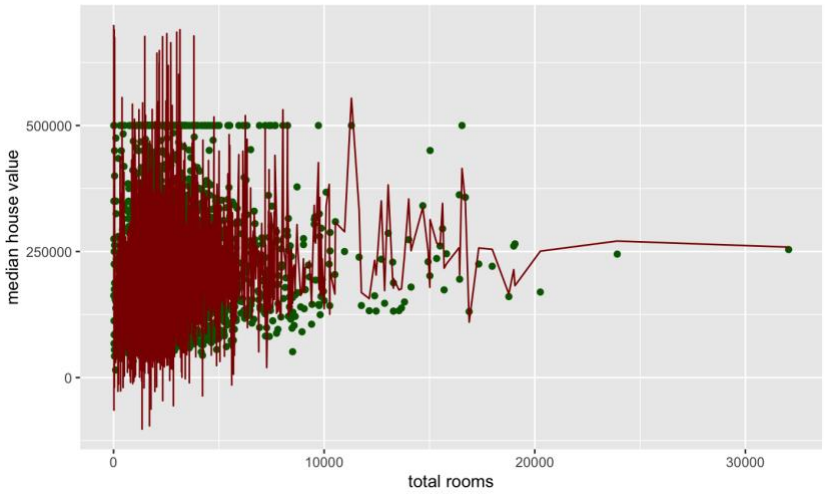

Figure 3: Total rooms by median house value.

After doing the multiple linear regression (MLR) test, the graph demonstrates how median house value fluctuates from 0 to 1,000 total rooms. As the number of rooms increases, the median house value stabilizes.

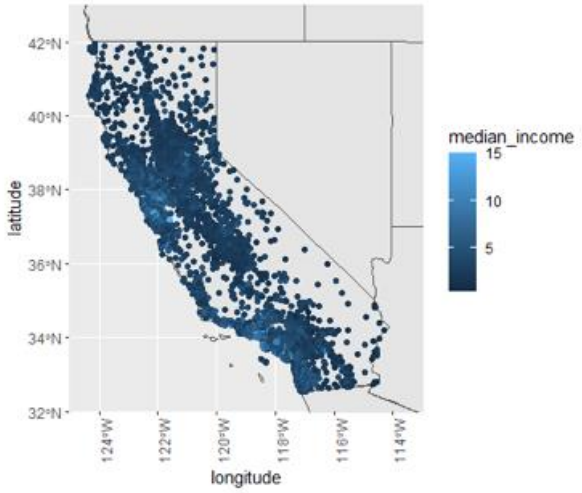

Figure 4: Median income by longitude and latitude (location).

This map demonstrates the median income (measured in tens of thousands of US Dollars) by location in California. As you can see, the median income increases for houses near the ocean. At the same time, lower median income is found more frequently in homes more inland in California.

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.772  -0.583  -0.018   0.540  12.261

Coefficients:
                     Estimate    Std. Error t value             Pr(>|t|)
(Intercept)         5.4203879042  1.1974633372    4.53          0.00000604 ***
longitude           0.0259694817  0.0137570040    1.89               0.059 .
latitude            0.0063615123  0.0131275417    0.48               0.628
housing_median_age -0.0222046951  0.0007709128  -28.80 < 0.0000000000000002 ***
total_rooms         0.0008676146  0.0000121148   71.62 < 0.0000000000000002 ***
total_bedrooms     -0.0042217264  0.0001031980  -40.91 < 0.0000000000000002 ***
population          0.0000850124  0.0000198943    4.27           0.00001938 ***
households         -0.0006103130  0.0001205938   -5.06           0.00000042 ***
median_house_value  0.0000102832  0.0000000952  108.06 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.12 on 16503 degrees of freedom
Multiple R-squared:  0.654,     Adjusted R-squared:  0.654
F-statistic: 3.9e+03 on 8 and 16503 DF,  p-value: <0.0000000000000002
```

This summary analysis demonstrates how all variables except longitude and latitude have a p-value = 0, which we can reject the null hypothesis and enormously significant in affecting California housing prices.

V. REGRESSION ANALYSIS

Describe the regression analysis, including the choice of predictor variables, model assumptions, and the interpretation of regression coefficients.

We observed how the rest of the variables affected median income and household values. We also observed how longitude and latitude affect those two variables. We assumed that the median income and household value would be more significant near the ocean, proven through our models.

VI. CONCLUSIONS AND RECOMMENDATIONS

Provide conclusions based on the EDA and regression analysis results.
Suggest any recommendations or further analysis if applicable.

All in all, after exploring and testing many different types of EDA and regression analysis. We can conclude that median income can correlate based on where someone lives in California. For instance, higher-income lived closer to the ocean, whereas lower-income households lived more inland. In conclusion, we enjoyed exploring this dataset and learning about housing prices in California.