Housing Prices

Which features have influenced California's housing prices in 1990?

Youssef Abdelwahab, Sara Dutton, Golin Chen, Danae McCulloch

Introduction

Dataset: California Housing Data (1990)

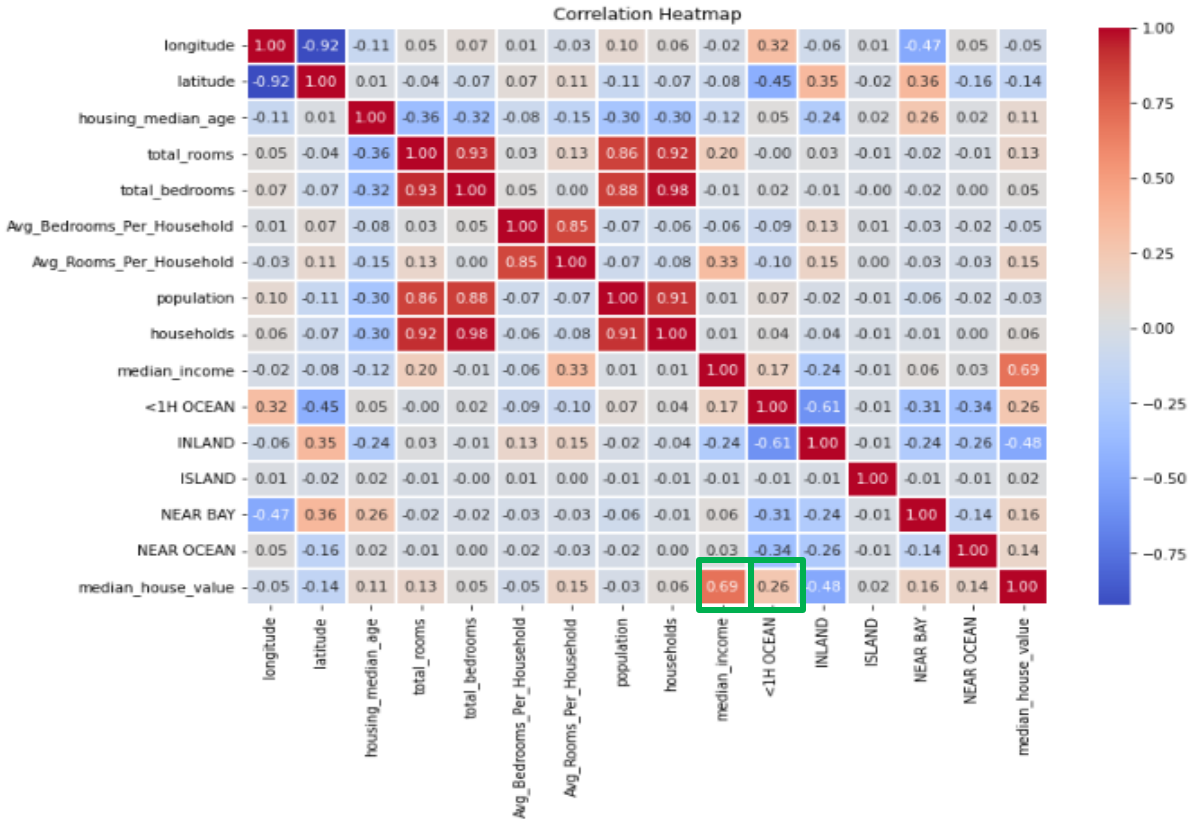
- Features are **by block**

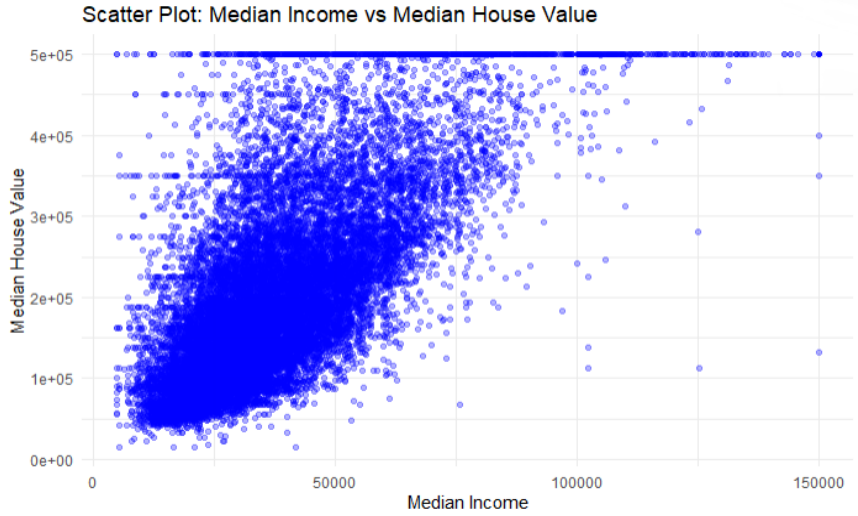
COLUMNS

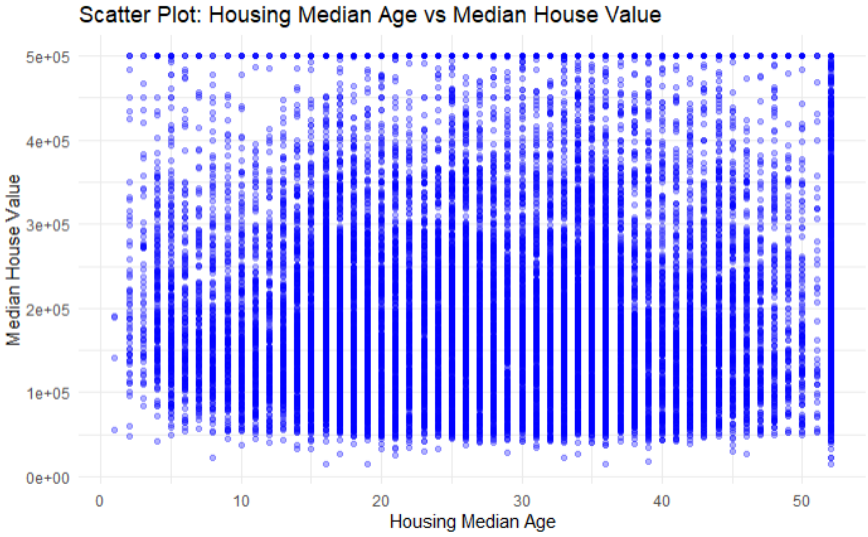
- Longitude
- Latitude
- Housing Median Age
- Total Rooms
- Total Bedrooms
- Avg. Bedrooms Per House
- Avg. Rooms Per House
- Population
- Household
- Median Income
- Ocean Proximity
- Median House Value

- Data Cleaning
  - Method na.omit() to remove missing values
  - Total of 20,433 rows
  - Dummy Variables to distinguish Ocean Proximity
  - Calculated the Avg. Bedrooms Per House and Avg. Rooms Per House column

Which feature has the strongest correlation? What is the weakest?

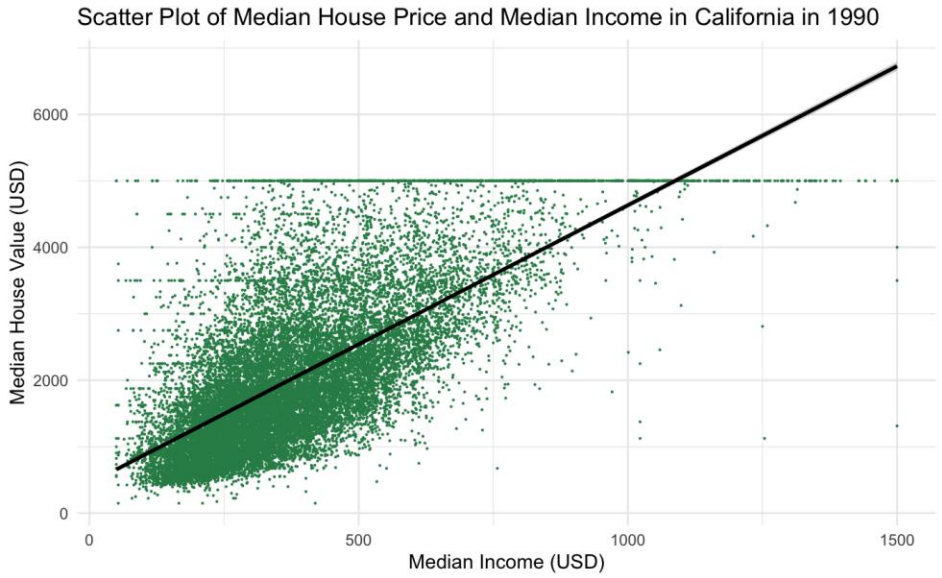






Does **higher income** have a higher median house value?

H0: $\widehat{\beta 1}$ = 0 (There is no positive relationship between median income and median house value)
HA: $\widehat{\beta 1} > 0$ (There is a positive relationship between median income and median house value)




```
Call:
lm(formula = normalized_median_house_value ~ normalized_median_income,
    data = housing_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5411.7  -558.6  -169.6   369.0  4341.8

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               449.06369   13.29965   33.77   <2e-16 ***
normalized_median_income    4.18371    0.03084  135.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 837.4 on 20431 degrees of freedom
Multiple R-squared:  0.4738,    Adjusted R-squared:  0.4738
F-statistic: 1.84e+04 on 1 and 20431 DF,  p-value: < 2.2e-16
```

- Our regression model can be defined as: **median house value = $44,900 + $418(median income)**
- Statistically Significant **(p-value < 0.05)** with **correlation, r = 0.688**
- We have enough evidence to **reject the null in favor of the alternative**

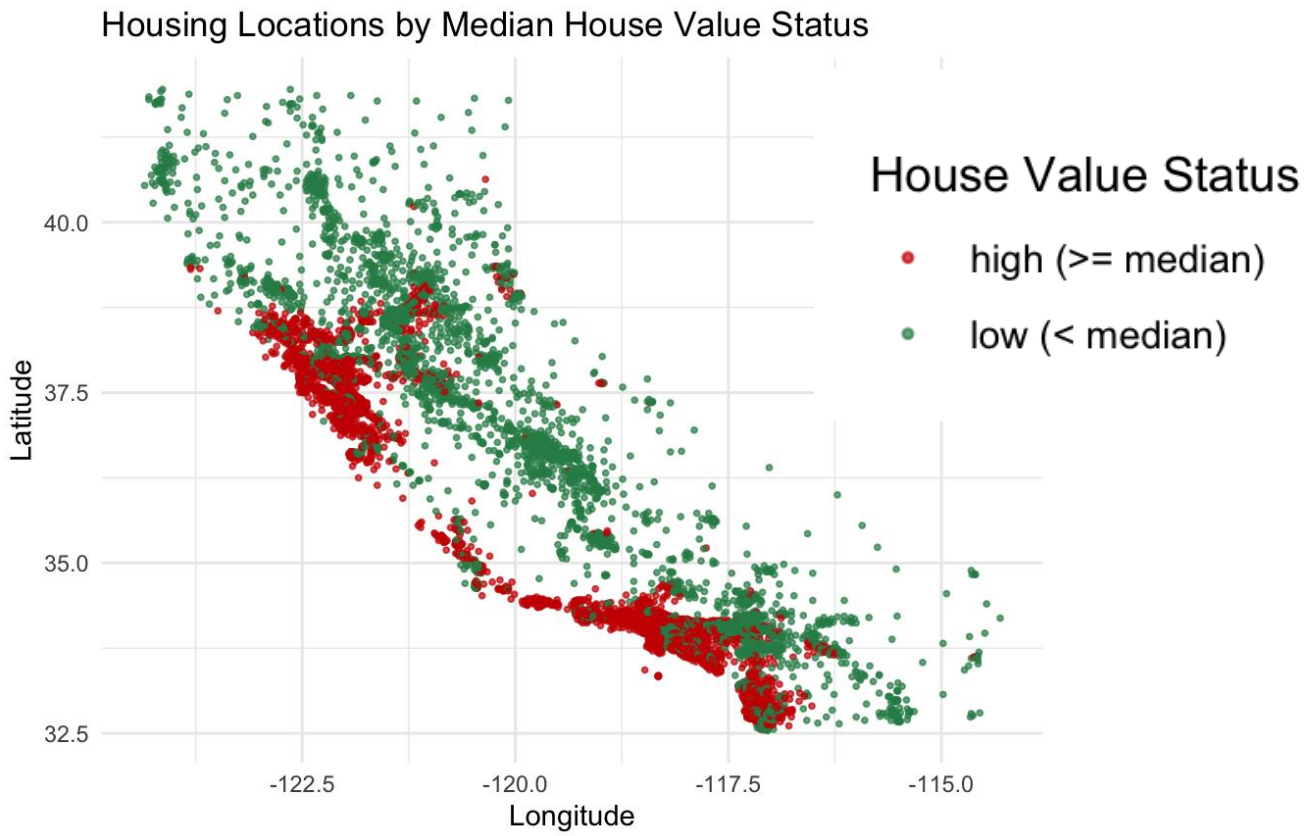
*The median income and median house value were normalized by dividing the values by 100

How does median house values vary by **geography** in California?

- Split (median) house value into two classes based on median:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14999	119500	179700	206864	264700	500001

- Clear pattern separating location of house with high vs low median house value
- Suggests something about geography is influencing house value
  - Looks coastal
  - Let's investigate **ocean proximity**…



Does **ocean proximity** affect the housing price?

Our dataset had 5 categories (i):

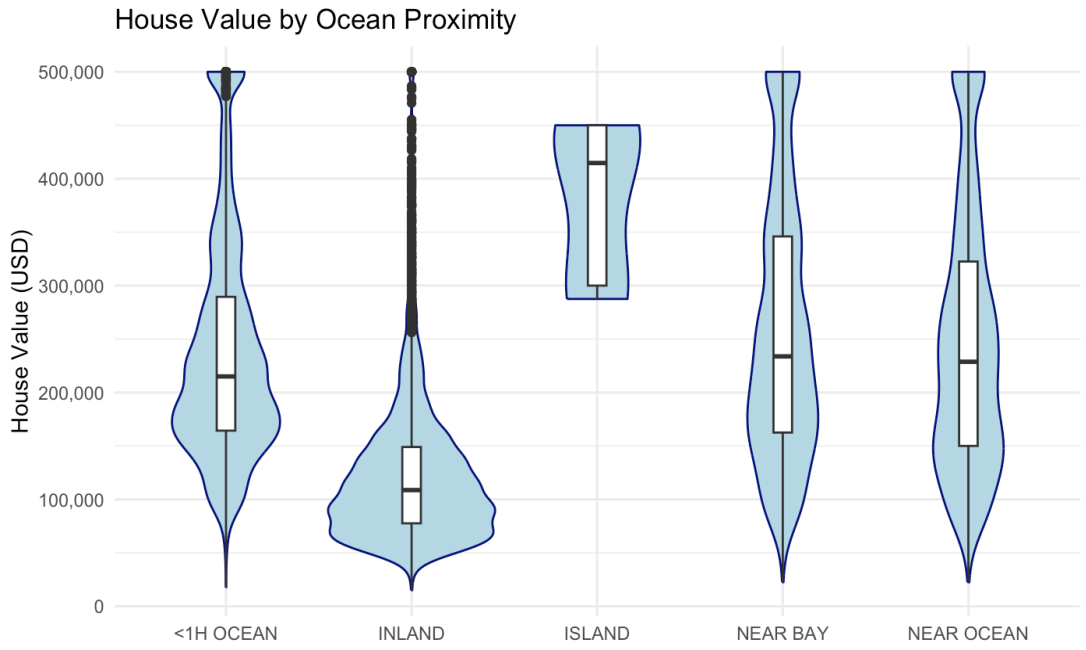
1. <1 Hour Ocean
2. Inland
3. Island
4. Near Bay
5. Near Ocean

Altered the dataset by incorporating "Dummy Variables"

- 1 is "Yes"
- 0 is "No"

median_income	ocean_proximity	median_house_value	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
83252	NEAR BAY	452600	0	0	0	1	0
83014	NEAR BAY	358500	0	0	0	1	0
72574	NEAR BAY	352100	0	0	0	1	0
56431	NEAR BAY	341300	0	0	0	1	0
38462	NEAR BAY	342200	0	0	0	1	0

- *H0: βi = 0: There is no relationship between ocean proximity and median house value*
- *HA: βi > 0: There is a positive relationship between ocean proximity and median house value*

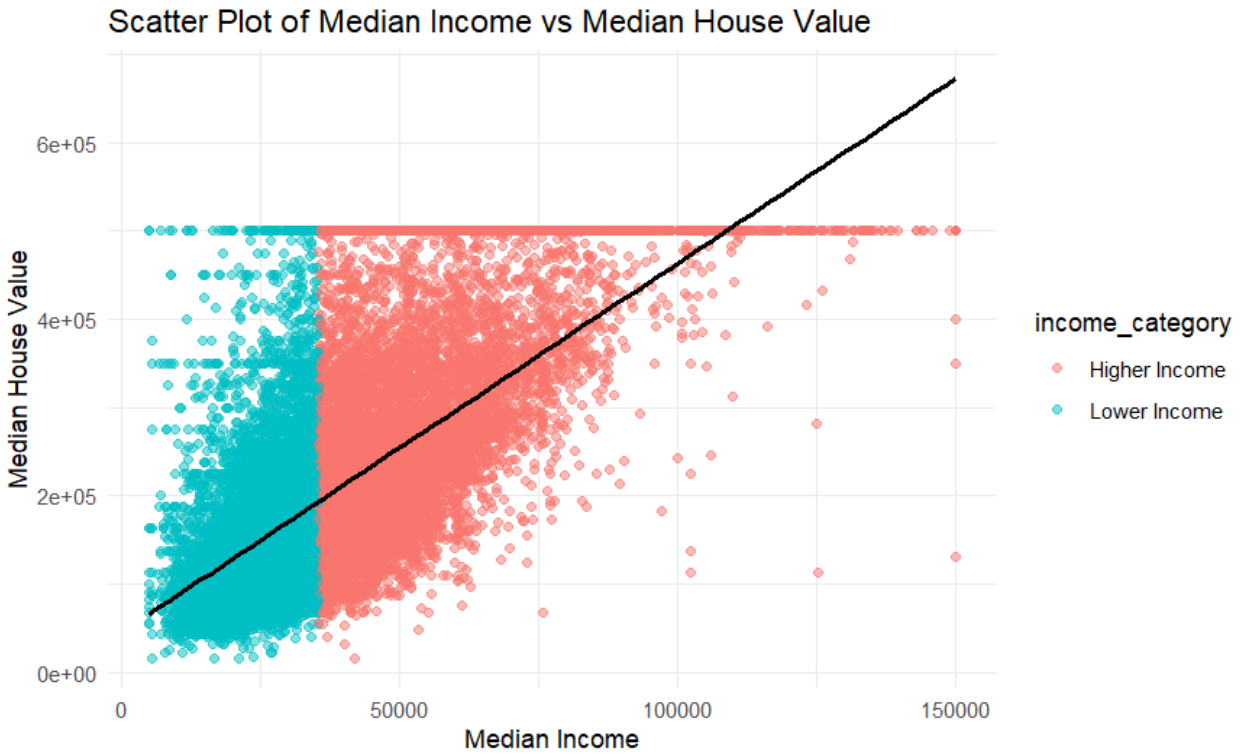



```
Call:
lm(formula = median_house_value ~ ocean_proximity, data = housing_data_cleaned_oceanencoded_in_)

Residuals:
    Min      1Q  Median      3Q     Max
-236779  -66268  -20897   42332  375104

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                240268       1060 226.602  < 2e-16 ***
ocean_proximityINLAND      -115371       1639 -70.372  < 2e-16 ***
ocean_proximityISLAND       140172      45082   3.109  0.00188 **
ocean_proximityNEAR BAY      19011       2366   8.035 9.88e-16 ***
ocean_proximityNEAR OCEAN     8774       2234   3.928 8.58e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 100800 on 20428 degrees of freedom
Multiple R-squared:  0.238,    Adjusted R-squared:  0.2378
F-statistic:  1595 on 4 and 20428 DF,  p-value: < 2.2e-16
```

i
- Category Baseline: <1 Hour (β_0)
- β_1 = Inland
- β_2 = Island
- β_3 = Near Bay
- β_4 = Near Ocean

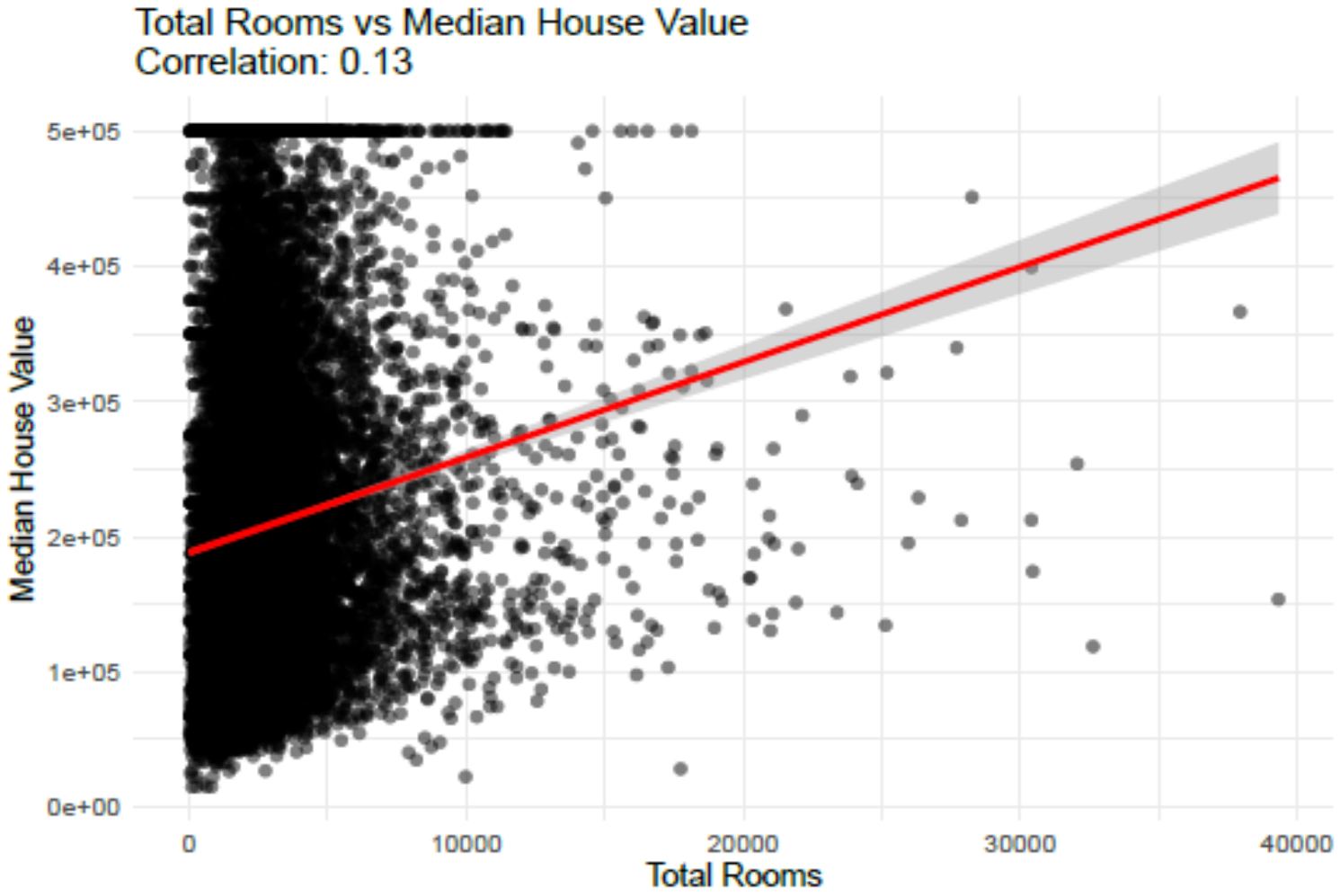
- E[House Price] = (β_0) + β_1 Inland + β_2 Island + β_3 Near Bay + β_4 Near Ocean
- Statistically Significant ($\alpha < 0.05$)
- Reject H0 in favor of HA

Proportions of Lower Income vs. High Income



Since we realized the correlation between income and house value is the strongest, we decided to take the median income level and split it into high vs. low income to see if higher income households purchase more expensive homes.

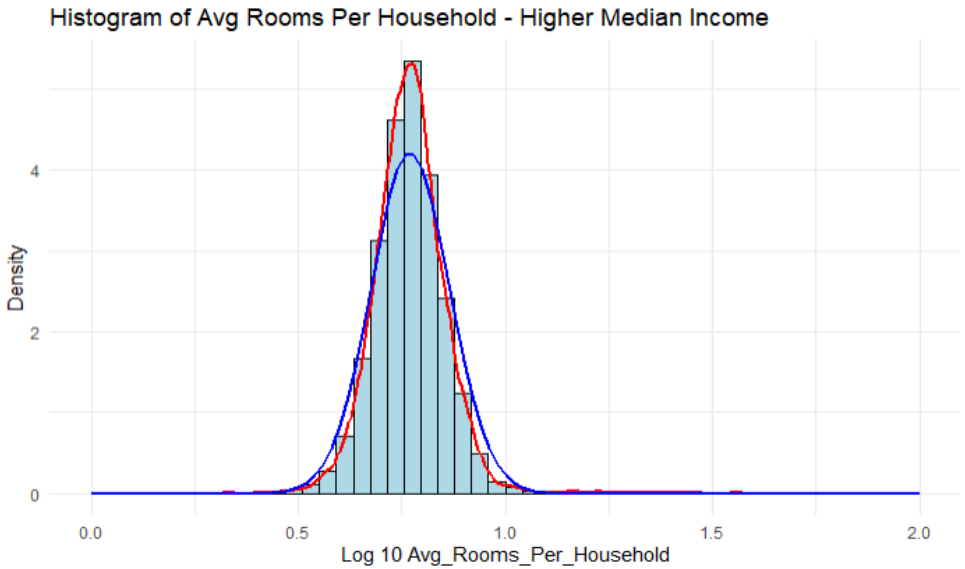
Total Rooms vs Median House Value



- A positive correlation suggests that more rooms in a block increase house values
- Linear regression will indicate how much `median house value` changes with each additional room

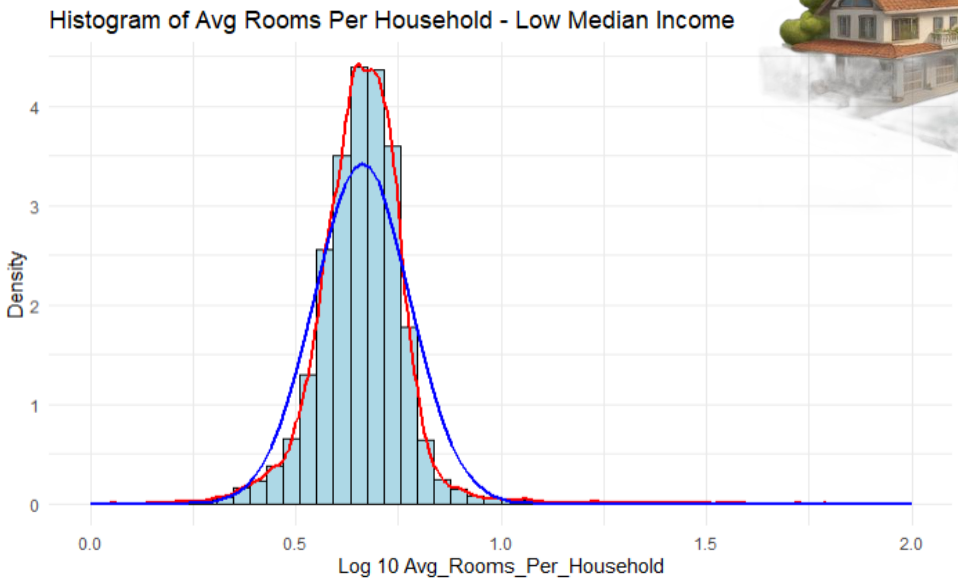
```
Residuals:
    Min      1Q  Median      3Q     Max
-311460  -86505  -26706   55721  311644

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.883e+05  1.254e+03  150.13   <2e-16 ***
total_rooms 7.041e+00  3.663e-01   19.22   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 114400 on 20431 degrees of freedom
Multiple R-squared:  0.01777,   Adjusted R-squared:  0.01772
F-statistic: 369.6 on 1 and 20431 DF,  p-value: < 2.2e-16
```

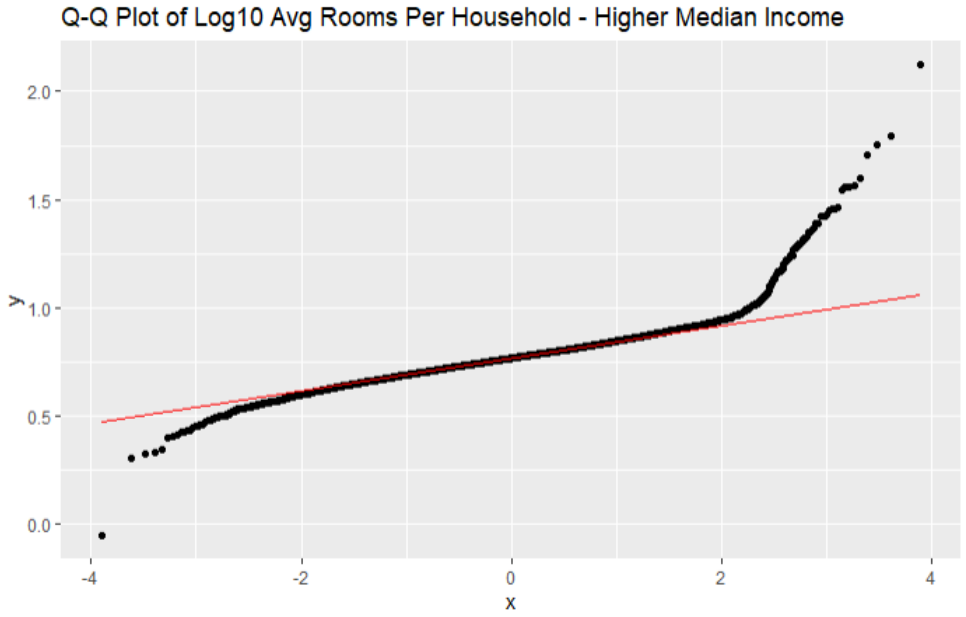
Does higher median income lead to larger houses (i.e., greater number of rooms) ?



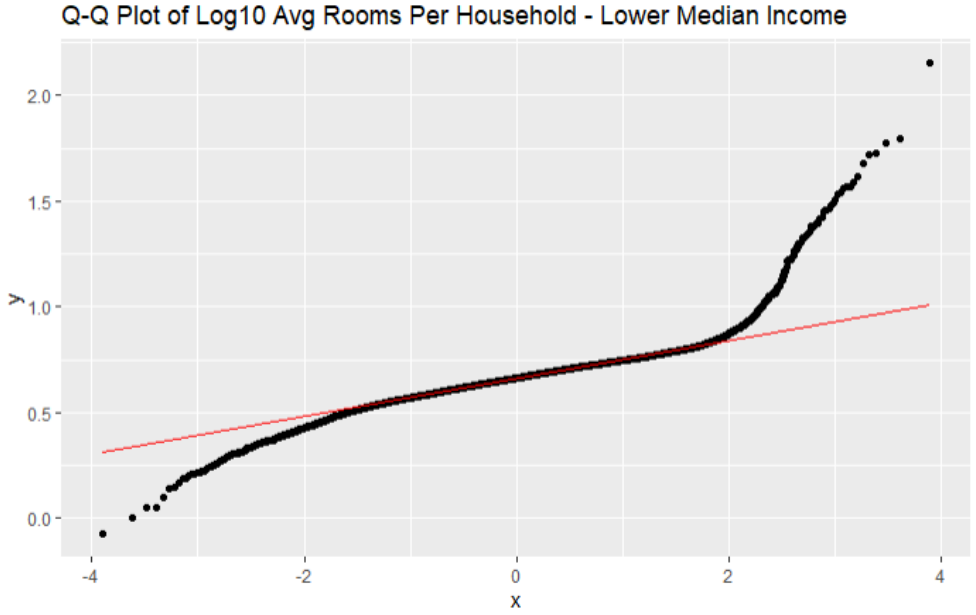
Blue: Normal Distribution

Red: Observed Distribution





- Heavier tails due to extreme outliers
- Reduces the relatability of inferences using parametric approaches using the assumption of normality

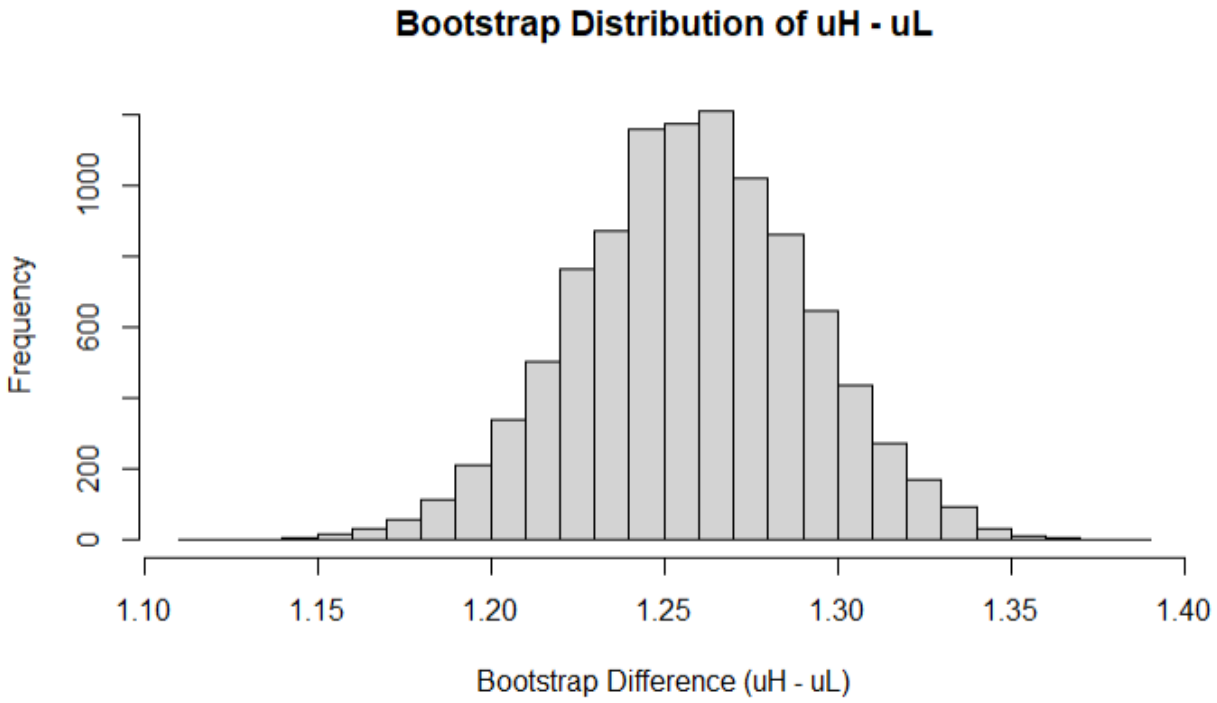


Hypothesis Testing

μ_1 : Population Mean of Average Rooms per Household in Higher Income Blocks

μ_2 : Population Mean of Average Rooms per Household in Lower Income Blocks

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_1 : \mu_1 > \mu_2$$



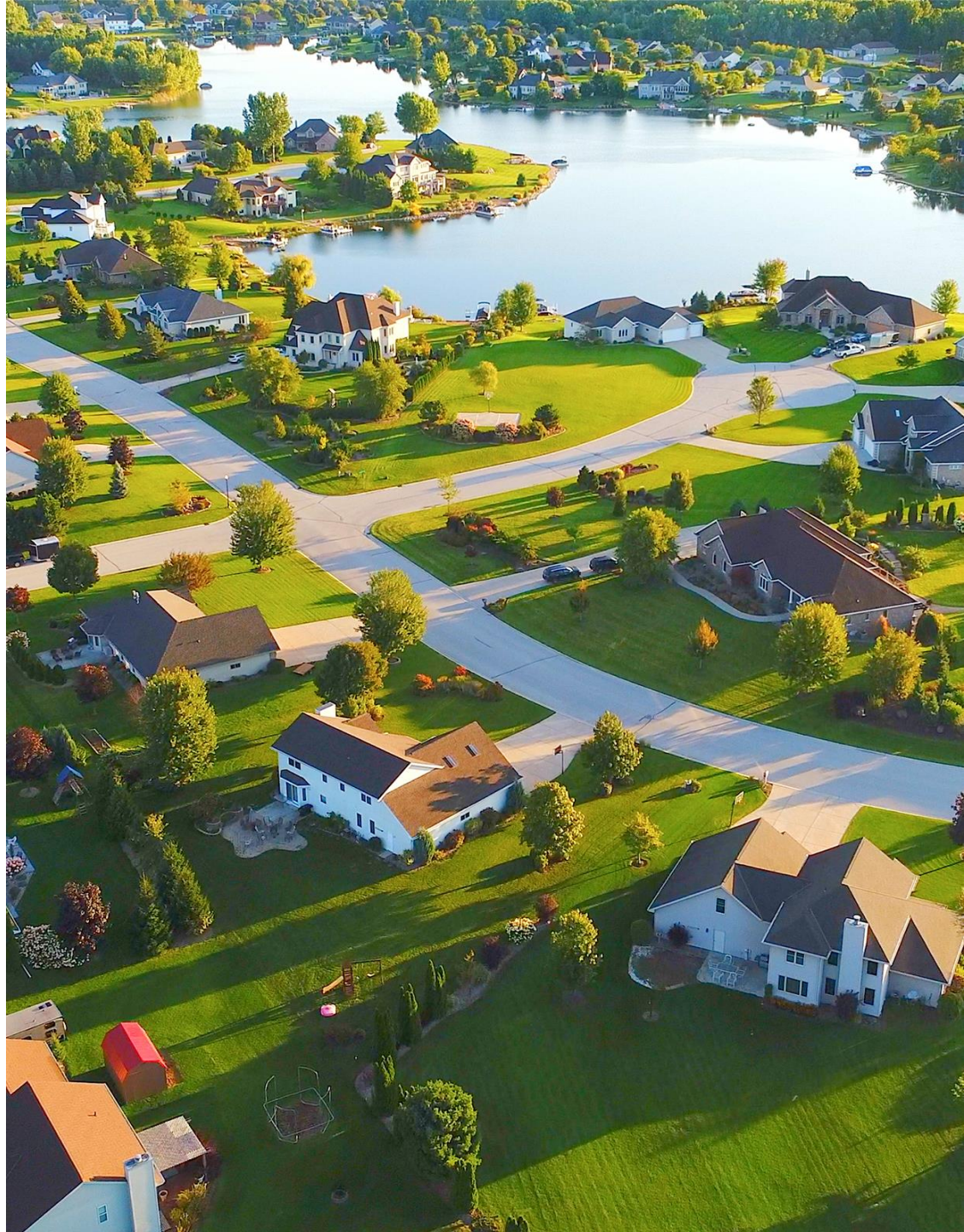
95% Confidence Interval

2.5%	97.5%
1.191945	1.323339

Sufficient evidence to reject the null hypothesis

Conclusions & Findings

- We have significant evidence to support that there is a positive correlation (r = 0.68) between Median Income and Median House Value.
- Ocean Proximity also had statistical significance but not as strong as Median Income.
- Suggestions for future research:
    - Include community/block descriptive variables such as crime rate, walking scores, areas of interest, and property age.
    - Identify the type of household to get a more accurate estimate of its value as apartments might skew the data.
    - Include other house feature variables such as size (square/ft), number of bathrooms and amenities.



References

- *Simple linear regression - one binary categorical independent variable*. Simple Linear Regression - One Binary Categorical Independent Variable | Practical Applications of Statistics in the Social Sciences | University of Southampton. (n.d.). https://www.southampton.ac.uk/passs/confidence_in_the_police/multivariate_analysis/linear_regression.page
- Wang, H. (2018, May 10). *California Housing Data (1990)*. Kaggle. https://www.kaggle.com/datasets/harrywang/housing