Data 602 Group Project: Housing Prices in California

Danae McCulloch, Sara Dutton, Youssef Abdelwahab, Golin Chen

2024-10-10

Introduction

Housing is a necessity; therefore it is important to analyze which socio-economic features impact housing prices. Our group decided to investigate which features will impact the housing market value the most and which has the least impact. The research question that we will be answering is *"Which features have influenced California's housing prices in 1990?"*.

Our dataset, "California Housing Data (1990)", was derived from Kaggle. This dataset includes columns analyzing specific house features such as total bedrooms, total rooms, ocean proximity, and the median house value to name a few. Note that the data represents the median housing value based by blocks within the California area and not by individual households; however, we will solve this issue in the data cleaning and data exploration phase of the project.

To ensure accuracy in our analysis, we performed data cleaning by removing rows with missing values using `na.omit()`, dropping around 200 rows of 20,640. We looked over all 200 rows of missing value and found that there is no significant information we can conclude from the rows with the missing values, even if we had filled in with the mean or median value. Additionally, to provide more meaningful context of these features, we calculated the average number of rooms and bedrooms per household by dividing total rooms and total bedrooms by number of households (as our dataset had a column that showed the total population within the block).

The dataset was further altered to investigate the relationship between ocean proximity and the median house value. The ocean proximity of the houses falls under five categories: Near Ocean, Near Bay, More than 1 Hour (<1 Hour Ocean), Inland, and Island. Our group incorporated dummy variables to assign the houses to their respective proximity. The houses were assigned to a '1' if they were found within the proximity and were assigned to a '0' in the other 4 categories to indicate that they were not within that proximity.

In this report, we will analyze the following features separately: the household income, ocean proximity, average number of rooms, as well as analyzing the proportions of both lower and higher incomes. These features will be compared against the housing value by using methods such as linear regression, data visualization, hypothesis testing, and confidence intervals.

Hypothesis Test: Does higher income have a higher median house value?

In this Exploratory Data Analysis (EDA), we will test the relationship between the median income and the median house value of a neighborhood by asking "how does median income of neighborhood influence the median house value of neighborhood in California in 1990", and see if there is sufficient evidence to suggest a positive relationship between these two features.

We define our hypotheses as follows:

$H_0 : \widehat{\beta}_1 = 0$ (There is no positive relationship between median income and median house value

$H_A : \widehat{\beta}_1 > 0$ There is a positive relationship between median income and median house value

Analysis:

Linear Regression Model: Median Income vs. Median House Value

```
## linear regression model

housing_data <- read_csv("housing_data_cleaned_oceanencoded.csv")
```

```
## New names:
## Rows: 20433 Columns: 19
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): ocean_proximity, Lat-Long dbl (17): ...1, longitude, latitude,
## housing_median_age, total_rooms, total_...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
housing_data <- housing_data

summary(housing_data$median_income) # print summary stats
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4999   25637   35365   38712   47440  150001
```

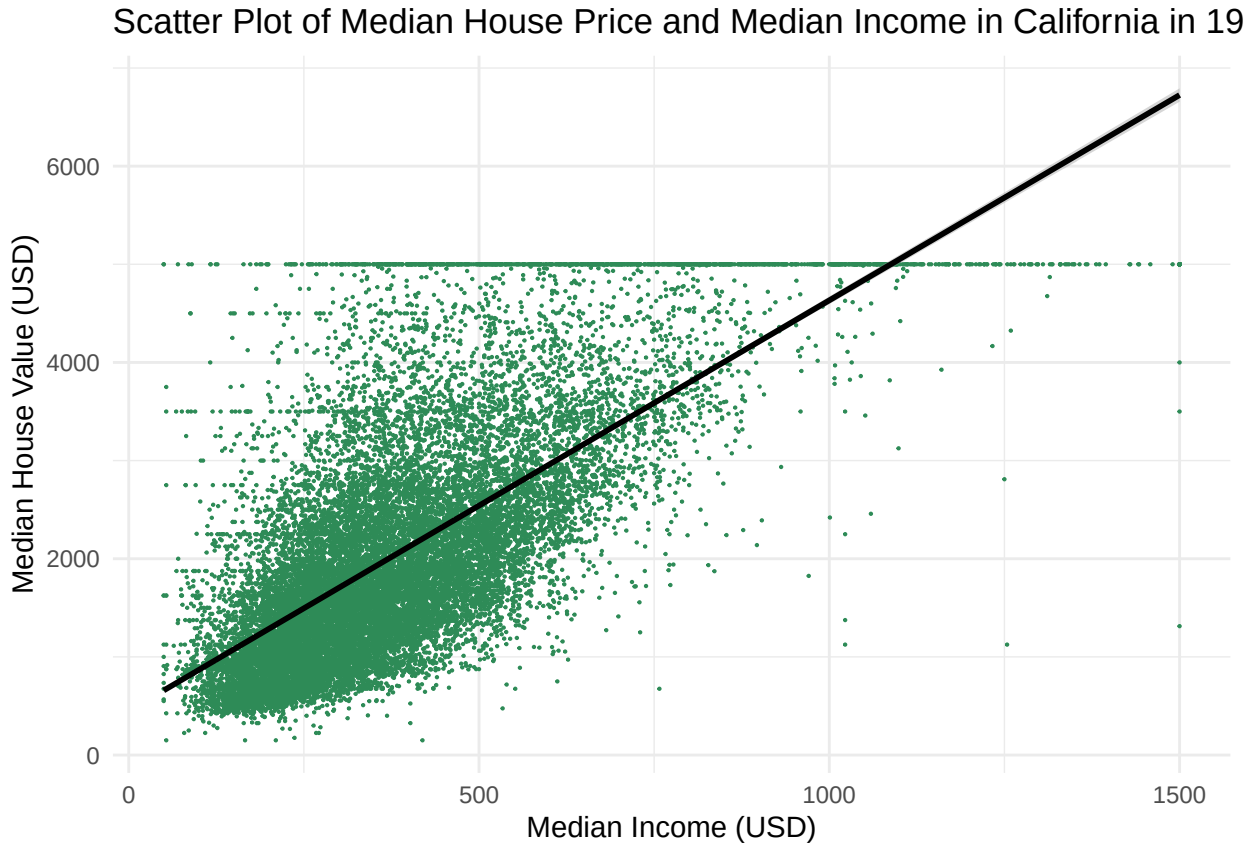
```
summary(housing_data$median_house_value) # print summary stats
```

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

```
# normalize prices by 100-fold
housing_data$normalized_median_income <- housing_data$median_income / 100
housing_data$normalized_median_house_value <- housing_data$median_house_value / 100

# plot linear regression model
price_vs_income <- ggplot(housing_data, aes(x = normalized_median_income, y = normalized_median_house_va
  geom_point(color = "seagreen", size = 0.1) + stat_smooth(method = "lm", formula = y ~ x, geom = "smoo
  labs(title = "Scatter Plot of Median House Price and Median Income in California in 1990",
       x = "Median Income (USD)",
       y = "Median House Value (USD)") +
  theme_minimal()

print(price_vs_income)
```




```
# create regression model
price_vs_income_reg <- lm(normalized_median_house_value ~ normalized_median_income, data = housing_data)
summary(price_vs_income_reg)
```

```
##
## 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
```

```r
# p value - for B_1 > 0
1 - pt(q=135.64, df=20431)
```

```
## [1] 0
```

```r
# test for correlation
cor(housing_data$normalized_median_income, housing_data$normalized_median_house_value)
```

```
## [1] 0.6883555
```

```r
## predictions
# predict median house value if income ranges from 50,000 to 1,000,000, increasing by 50,000
x <- data.frame(normalized_median_income = seq(50000,1000000,by=50000))

predictions <- data.frame(predict(price_vs_income_reg, x, interval = "prediction", level =0.95))

predicted_value <- cbind(x, predictions)
```

Linear Regression Model - Conclusions

Model Evaluation

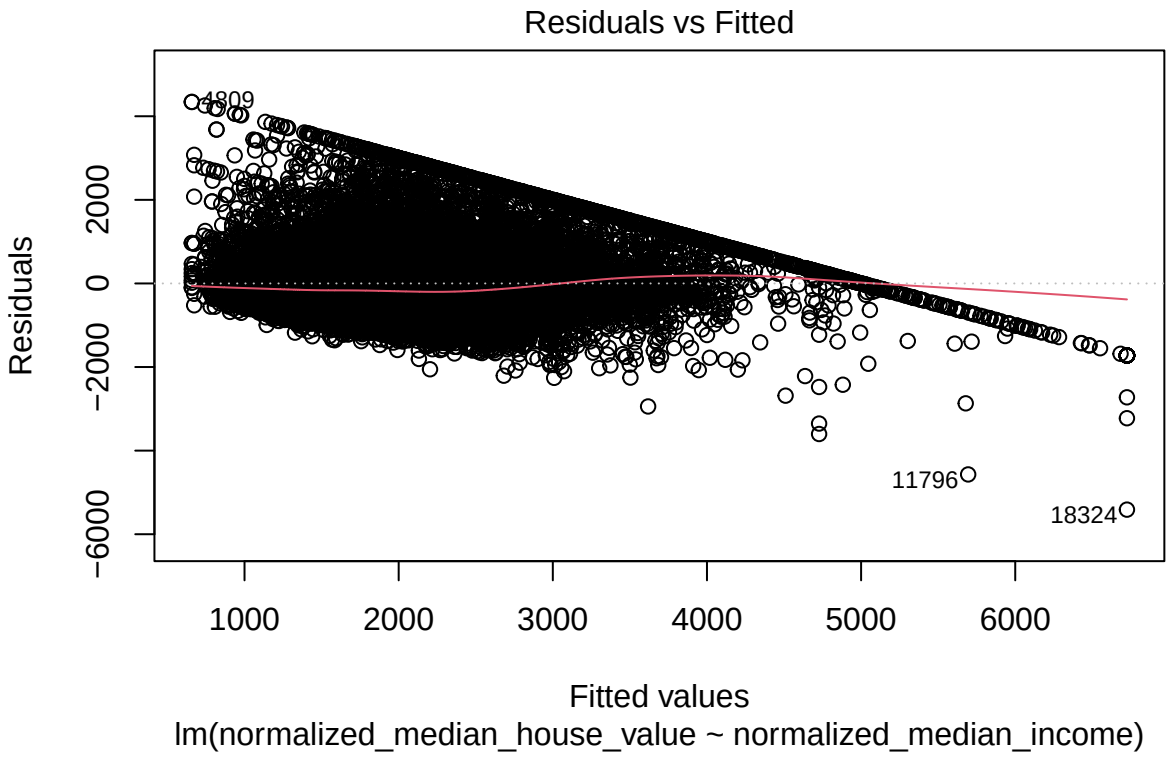
Our model is defined as:

$$\text{median\_house\_value} = 449.1 + 4.184 \times \text{median\_income}$$

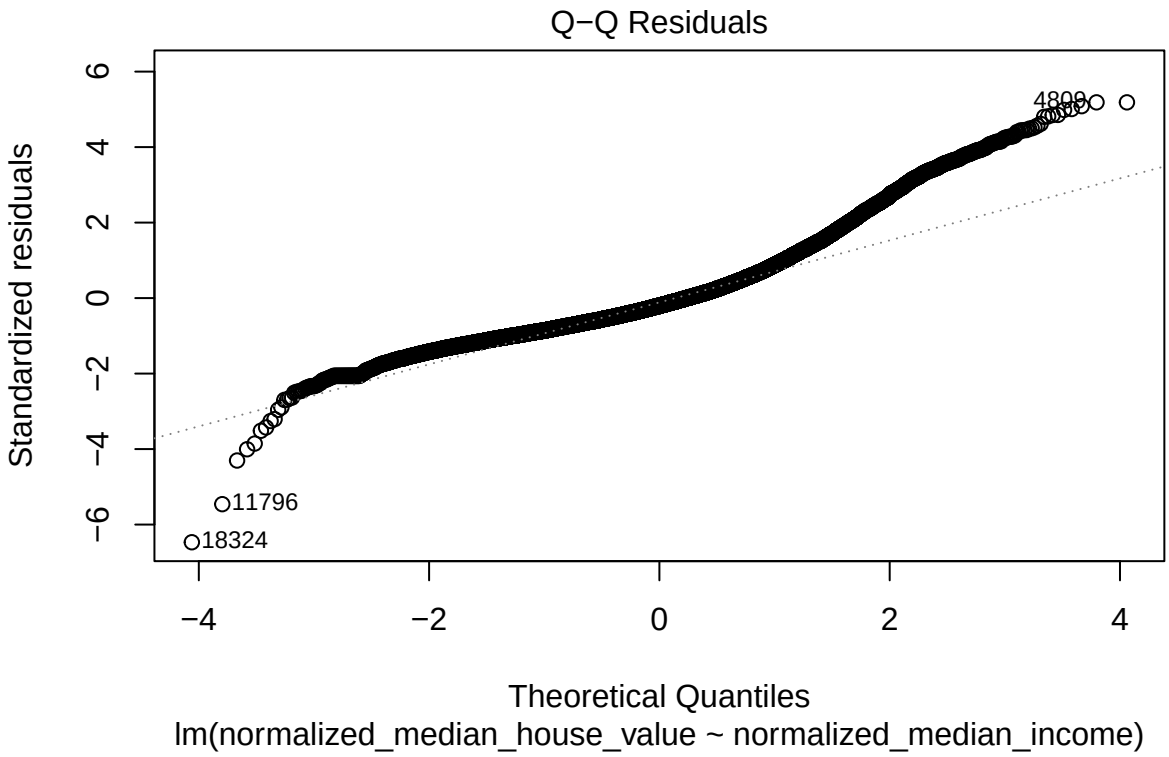
This means that for every one unit increase in median income, the median house value increases by approximately $418.4. When the median income is 0 the median house value is 449.1. The p-value is < 0.05 meaning we have strong evidence reject the null, that there is no positive relationship between median income and median house value, in favour of the alternative. This model has a R-squared 0.4738 indicates that approximately 47.4% of the variability in median house value can be attributed to median income. While this is a moderate level of explanatory power, it also suggests that other factors likely contribute to house value.The correlation between median income and median house value is r = 0.688 indicating positive correlation that aligns with the model. Therefore, we conclude there is significant evidence that there is a positive relationship between median income and median house value.

Testing Assumptions

```r
# test for assumptions
plot(lm(normalized_median_house_value ~ normalized_median_income, data = housing_data), which = 1)
```




```
plot(lm(normalized_median_house_value ~ normalized_median_income, data = housing_data), which = 2)
```



1. Test for normal distributions - The QQ plot follows the linear line closely before tapering off likely due to some more extreme values. It appears to start as a roughly a normal distribution but the ends indicate there may not be a normal distribution.

2. Test for linear relationship - The plot of residuals vs fitted roughly follows a straight line, with no distinct pattern observed on the red line. However, the residual tapper of towards the end suggesting their may be another pattern.

These plots indicate there may be some other pattern and that there is not a strong evidence they pass these assumptions. Based on this, despite having a moderate R squared in the model model these plots indicate that a linear regression might not be the best model to explain this relationship.

Exploratory Data Analysis: How does House Value by Geography?

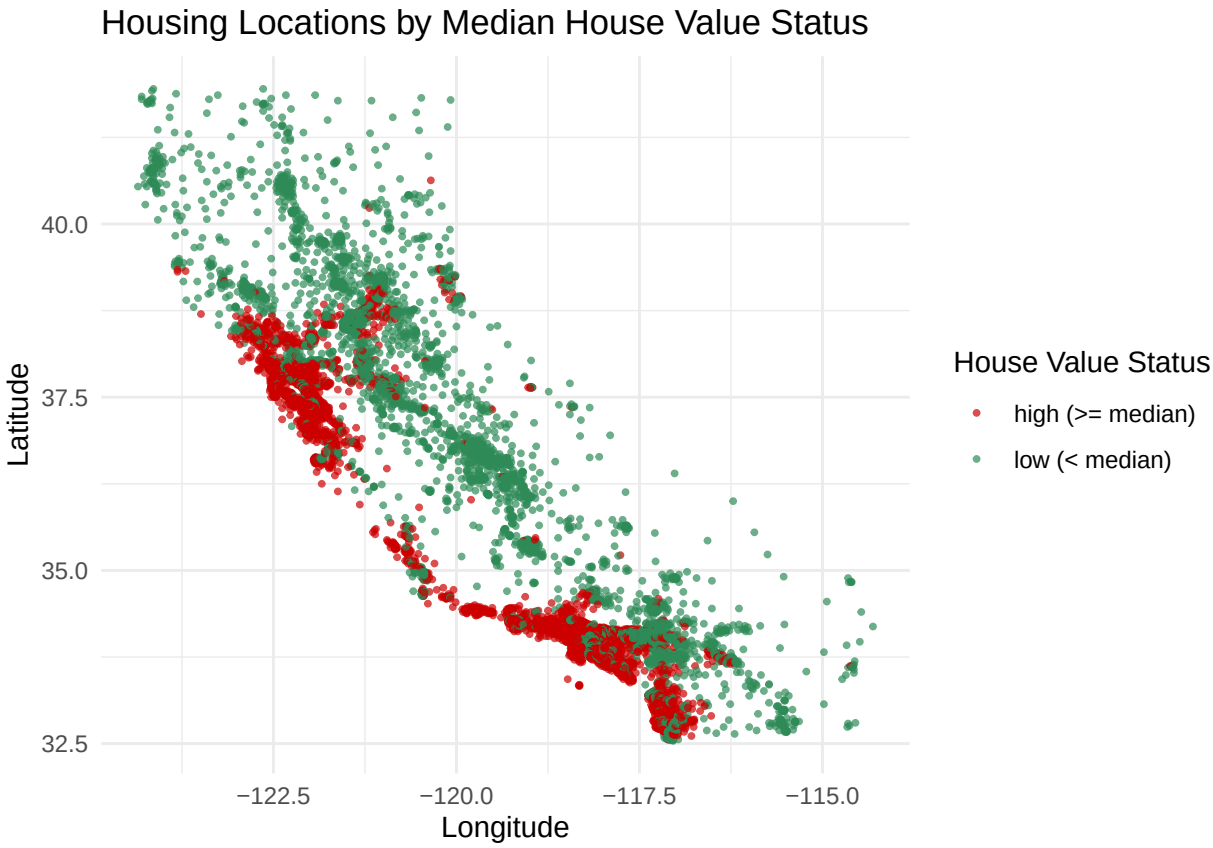
EDA

```
# categorize income status
housing_data$income_status <- ifelse(housing_data$median_income >= 35365, "high", "low") # based on med
# categorize house value status
housing_data$house_status <- ifelse(housing_data$median_house_value >= 179700, "high (>= median)", "low

value_geog <- ggplot(housing_data, aes(x = longitude, y = latitude, color = house_status)) +
  geom_point(alpha = 0.7, size = 0.7) + # Adjust alpha for transparency
  labs(title = "Housing Locations by Median House Value Status",
       x = "Longitude",
```

```
       y = "Latitude",
       color = "House Value Status") +
  scale_color_manual(values = c("high (>= median)" = "red3", "low (< median)" = "seagreen")) + # Custom
  theme_minimal()

print(value_geog)
```



Key Takeaways

The analysis reveals a clear pattern that distinguishes the locations of houses with high median house values from those with low values (split by median). This suggests that geographic factors are influencing house prices. Given the observed trends, we hypothesize ocean proximity has an impact, and will furhter investigate the impact of ocean proximity on housing values.

Analysis: Does ocean proximity affect the housing price?

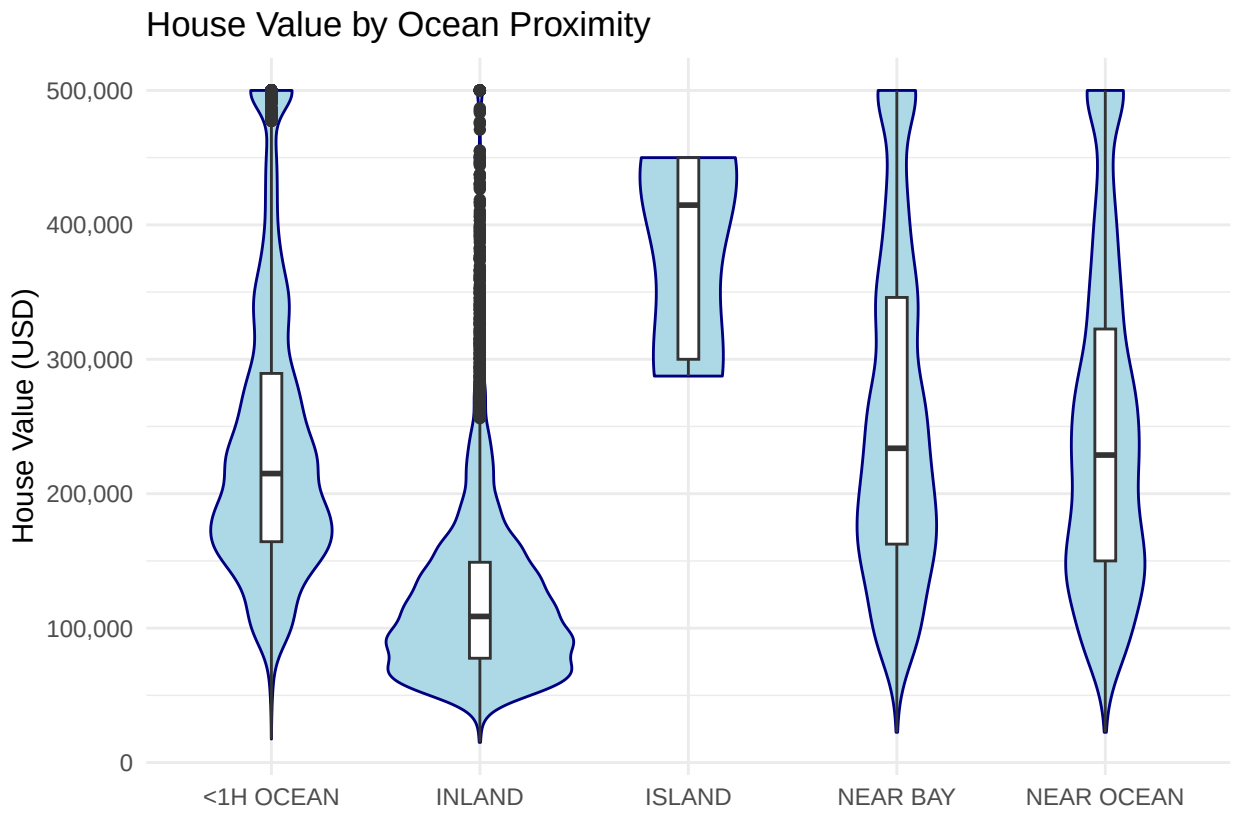
In this section, we will be investigating if ocean proximity affects the median house value. We have five categories to investigate: <1 Hour Ocean (within one hour of the ocean), Inland, Island, Near Bay, and Near Ocean. Note that in the introdcution we mentioned that we used dummy variables for this analysis which allowed us to perform a linear regression.

```
housing_data_cleaned_oceanencoded_in_ <- read_csv("housing_data_cleaned_oceanencoded.csv")
```

```
## New names:
```

```
## Rows: 20433 Columns: 19
## -- Column specification
## ------------------------------------------------------- Delimiter: "," chr
## (2): ocean_proximity, Lat-Long dbl (17): ...1, longitude, latitude,
## housing_median_age, total_rooms, total_...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
ggplot(housing_data_cleaned_oceanencoded_in_, aes(x = ocean_proximity, y = median_house_value)) +
  geom_violin(fill = "lightblue", color = "navy") +
  geom_boxplot(width=0.1) +
  labs(title = "House Value by Ocean Proximity",
       x = "",
       y = "House Value (USD)") +
  scale_y_continuous(labels = comma) +
  theme_minimal()
```



Description of Violin Plot: In this violin plot we can see that Island houses have resulted the highest median income out of all of the categories.This is mostly due to the fact that purchasing Island housing is more competitive due to land space and ocean proximity/views. The median housing value for Island houses is found around 420,000 USD. This is followed by <1 Hour Ocean, Near Bay, Near Ocean and finally, Inland with the lowest calculated median house value. Interesting enough, Near Bay and Near Ocean were very familiar in their shape and median house value (~230,000 USD). Not surprisingly, Inland houses were valued at the lowest with a median at ~115,000 USD. This is most likely due to having less competition and a less desirable location as we are further away from the ocean.

```r
housing_data_cleaned_oceanencoded_in_$ocean_proximity <- as.factor(housing_data_cleaned_oceanencoded_in

ocean_model <- lm(median_house_value ~ ocean_proximity, data = housing_data_cleaned_oceanencoded_in_)
summary(ocean_model)
```

```
##
## 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
```

Ocean Proximity: Hypothesis Testing $H0 = \beta i = 0$ Ocean Proximity has no effect on the median house value. $HA = \beta i > 0$ Ocean Proximity does have an effect on the median house value.

Note that βi (the regression coefficient) represents the five following categories: $i =$ <1 Hour Ocean, Near Bay, Near Ocean, Island, and Inland.

Summary of Linear Regression between Ocean Proximity and Median House Value To quickly summarize, the regression coefficients represent the following categories: $\beta_0 =$ <1 HOUR OCEAN, $\beta_1 =$ INLAND, $\beta_2 =$ ISLAND, $\beta_3 =$ NEAR BAY, and $\beta_4 =$ NEAR OCEAN

The overall p-value of $< 2.2e-16$ concludes that there is a significant relationship between ocean proximity and median house value as it falls below the significance level ($\alpha < 0.05$). Our R-squared was valued at 0.238 which indicates that 23.8% variability in the median house value is explained by ocean proximity. Therefore, we reject the null hypothesis in favor of the alternative which states ocean proximity is statistically significant.

It is important to note that our ocean proximity '<1 Hour Ocean' is our category baseline, $\beta 0$. The intercept is valued at 240,268 USD and this value is compared against the remaining four categories. For example, if we were to compare <1 Hour to Inland, this model is suggesting that Inland houses are 115,371 USD cheaper (as represented by the negative sign). In contrast, Island homes are shown to be 140,172 USD more than the <1 Hour homes. The line that represents the expected housing price is the following equation: $E[MedianHouseValue] = \beta 0 + \beta 1 INLAND + \beta 2 ISLAND + \beta 3 NEARBAY + \beta 4 NEAROCEAN$. β_0 is the intercept (240,268) and it also represents the value of the housing price when all variables are equaled to zero.

Analysis: Comparing the proportions of Lower Income vs. High Income and their purchasing behaviour.

For this we decided to make the median of the median_income data and anything above will be consider high income and the rest will be consider as low income. (note the reason we decided to use the median is because mean could be skew by the top 1% individual who earns significantly more than the average individual)

Then graph the information we obtain using a box and scatterplot in order to extract valuable insight.

```
median_income_value <- median(housing_data$median_income, na.rm = TRUE)

housing_data <- housing_data %>%
  mutate(income_category = ifelse(median_income < median_income_value, "Lower Income", "Higher Income"))

head(housing_data)
```

```
## # A tibble: 6 x 24
##    ...1 longitude latitude housing_median_age total_rooms total_bedrooms
##   <dbl>     <dbl>    <dbl>              <dbl>       <dbl>          <dbl>
## 1     0     -122.     37.9                 41         880            129
## 2     1     -122.     37.9                 21        7099           1106
## 3     2     -122.     37.8                 52        1467            190
## 4     3     -122.     37.8                 52        1274            235
## 5     4     -122.     37.8                 52        1627            280
## 6     5     -122.     37.8                 52         919            213
## # i 18 more variables: Avg_Bedrooms_Per_Household <dbl>,
## #   Avg_Rooms_Per_Household <dbl>, population <dbl>, households <dbl>,
## #   median_income <dbl>, ocean_proximity <chr>, median_house_value <dbl>,
## #   `<1H OCEAN` <dbl>, INLAND <dbl>, ISLAND <dbl>, `NEAR BAY` <dbl>,
## #   `NEAR OCEAN` <dbl>, `Lat-Long` <chr>, normalized_median_income <dbl>,
## #   normalized_median_house_value <dbl>, income_status <chr>,
## #   house_status <chr>, income_category <chr>
```

```
ggplot(housing_data, aes(x = income_category, y = median_house_value, fill = income_category)) +
  geom_boxplot() +
  labs(title = "Income Level (Lower vs Higher) vs Median House Value",
       x = "Income Category", y = "Median House Value") +
  theme_minimal() +
  scale_fill_manual(values = c("Lower Income" = "lightgreen", "Higher Income" = "orange")) +
  theme(legend.position = "none")
```

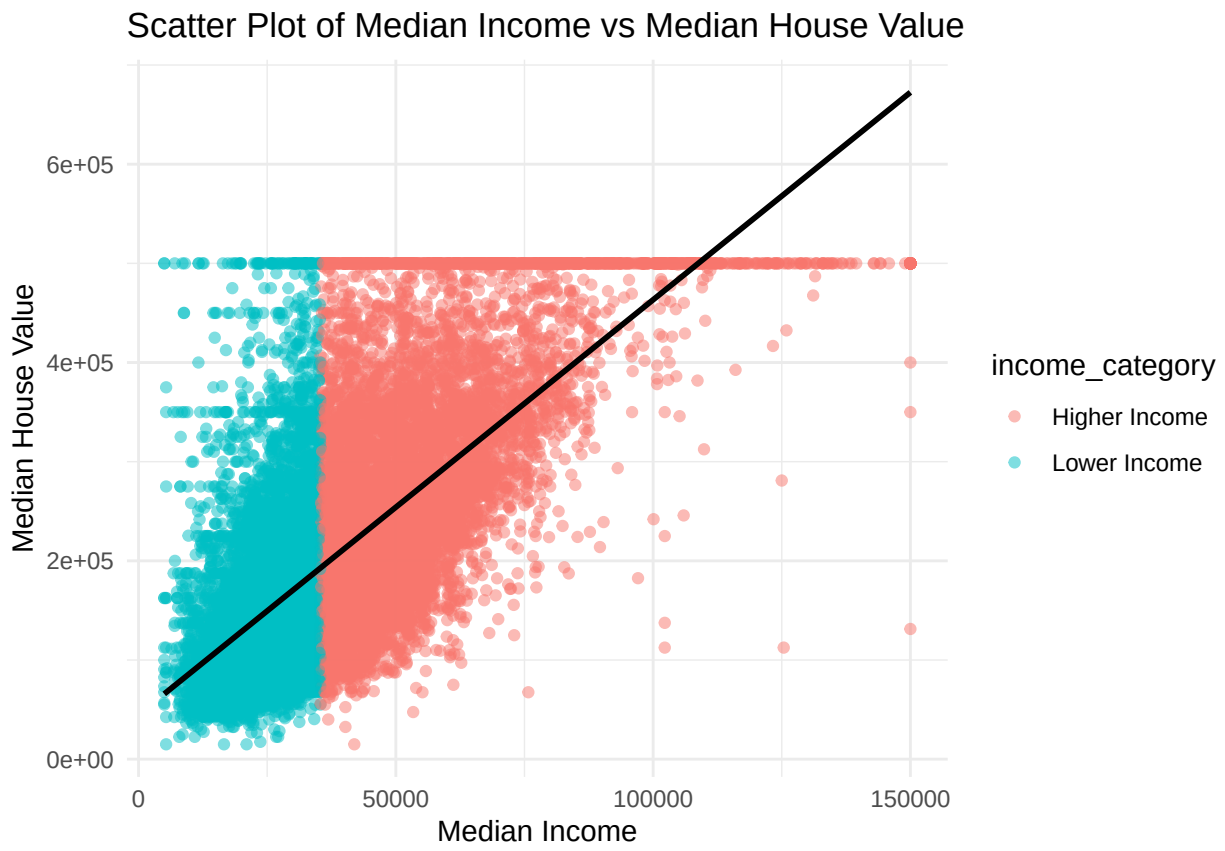

```
avg_house_value_by_income <- housing_data %>%
  group_by(income_category) %>%
  summarise(avg_median_house_value = mean(median_house_value, na.rm = TRUE))

# Preview the summarized data
print(avg_house_value_by_income)
```

```
## # A tibble: 2 x 2
##   income_category avg_median_house_value
##   <chr>                            <dbl>
## 1 Higher Income                  266368.
## 2 Lower Income                   147355.
```

```
ggplot(housing_data, aes(x = median_income, y = median_house_value, color = income_category)) +
  geom_point(alpha = 0.5) +  # Scatter points
  geom_smooth(method = "lm", se = FALSE, , color = "black") +  # Add a trend line for each income categ
  labs(title = "Scatter Plot of Median Income vs Median House Value",
       x = "Median Income", y = "Median House Value") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



To test the hypothesis that higher-income individuals buy more expensive houses compared to lower-income individuals, we can perform a hypothesis test and calculate a 95% confidence interval for the difference in mean house prices between the two groups.

$$H_0 : \mu_{high} = \mu_{low} \text{ vs. } H_a : \mu_{high} > \mu_{low}$$

Our null hypothesis is if There is no difference in the median house values between lower-income and higher-income households and our alternative hypothesis is that higher-income households tend to buy more expensive houses.

```
lower_income_group <- housing_data %>% filter(income_category == "Lower Income")
higher_income_group <- housing_data %>% filter(income_category == "Higher Income")

t_test_result <- t.test(higher_income_group$median_house_value, lower_income_group$median_house_value, a

print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  higher_income_group$median_house_value and lower_income_group$median_house_value
## t = 85.993, df = 18303, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  116736.6      Inf
## sample estimates:
## mean of x mean of y
##  266368.1  147354.9
```

the p-value is below 0.05, this means we can conclude that higher-income individuals are likely to buy more expensive houses.

Analysis - Does the supply of rooms in a block bring down the house value?

For this question, I decided to look at the correlation and the R-squared between the total rooms and the median house value to see if the linear regression show a positive relationship between the two variables.

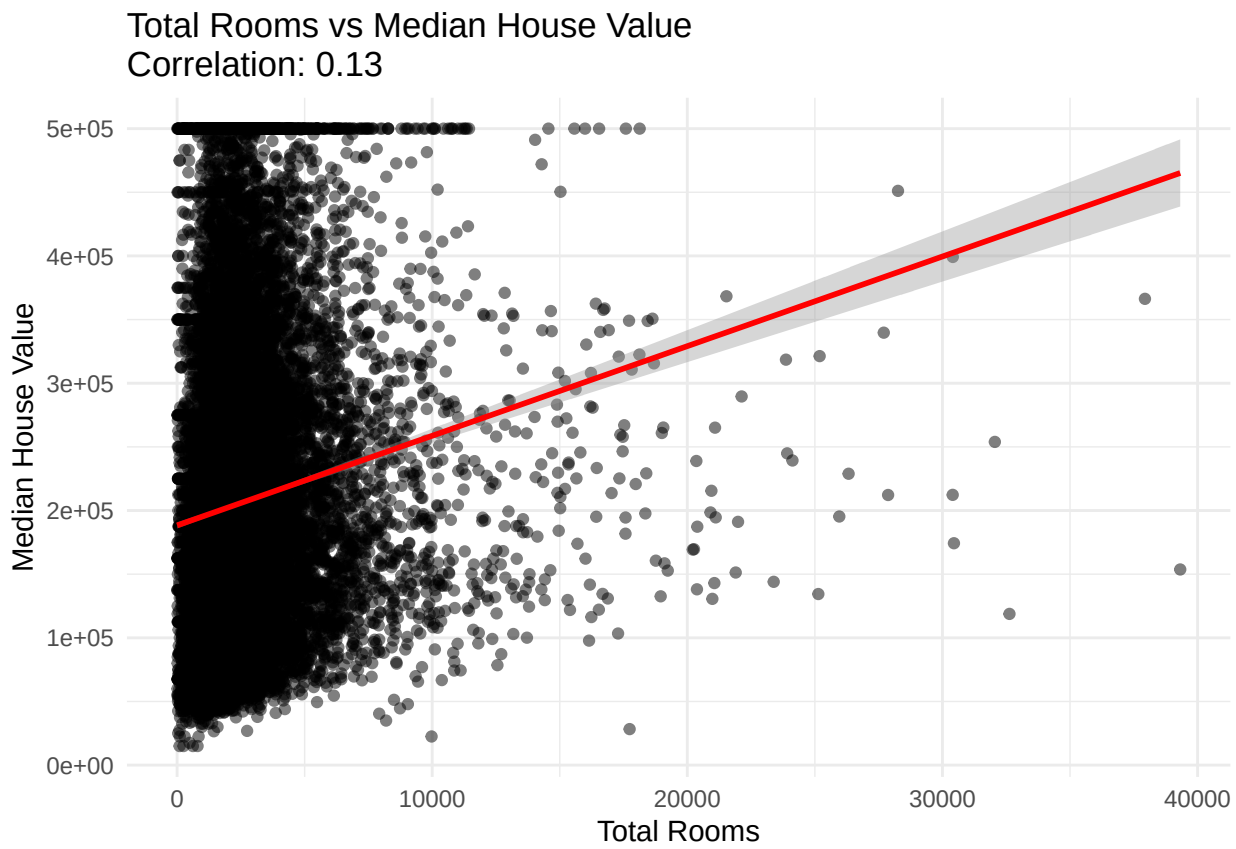
```
correlation <- cor(housing_data$total_rooms, housing_data$median_house_value, use = "complete.obs")
print(paste("Correlation:", round(correlation, 2)))
```

```
## [1] "Correlation: 0.13"
```

We can see the correlation is relatively low at 0.13 let's see what may cost this issue.

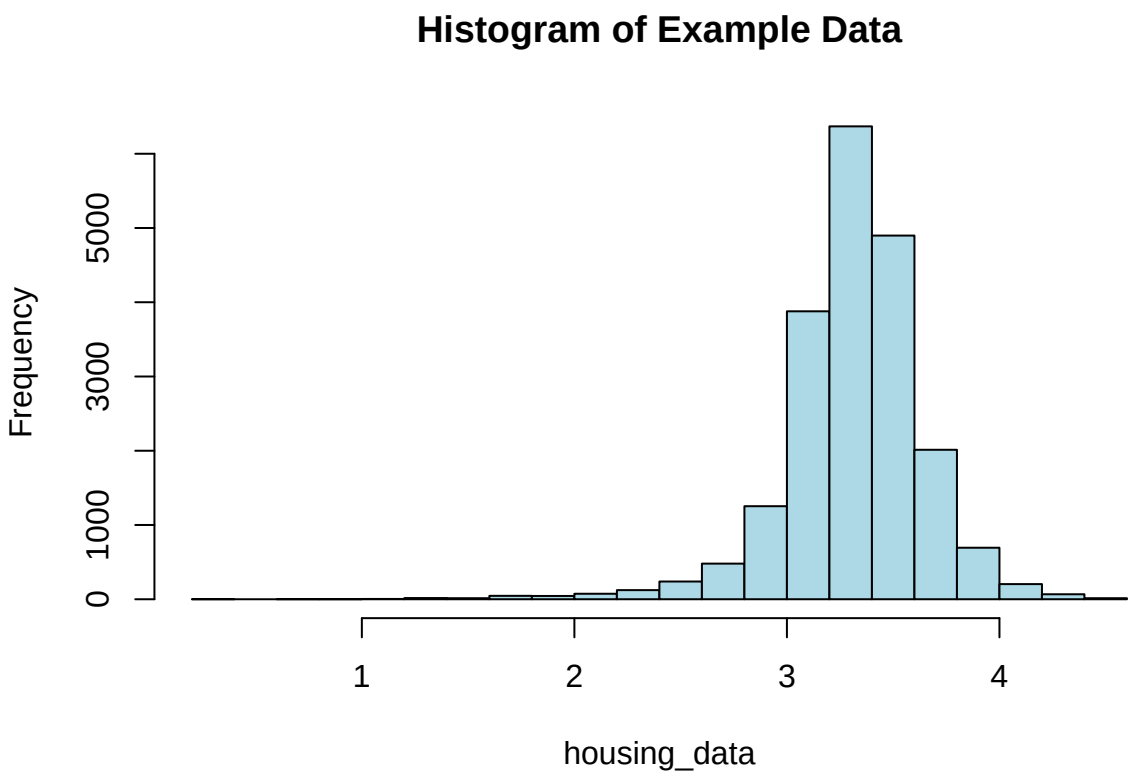
```
ggplot(housing_data, aes(x = total_rooms, y = median_house_value)) +
  geom_point(alpha = 0.5) +  # Scatter points
  geom_smooth(method = "lm", color = "red") +  # Add regression line
  labs(title = paste("Total Rooms vs Median House Value\nCorrelation:", round(correlation, 2)),
       x = "Total Rooms", y = "Median House Value") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



We can see from this scatter plot that the general trends of the graph isn't clear and well defined.

```r
hist(log10(housing_data$total_rooms),
     main = "Histogram of Example Data",
     xlab = "housing_data",
     ylab = "Frequency",
     col = "lightblue",
     border = "black")
```



The histogram for the total_rooms we can see the distribution shifted to the left following a binomial distribution

```
model <- lm(median_house_value ~ total_rooms, data = housing_data)

summary(model)
```

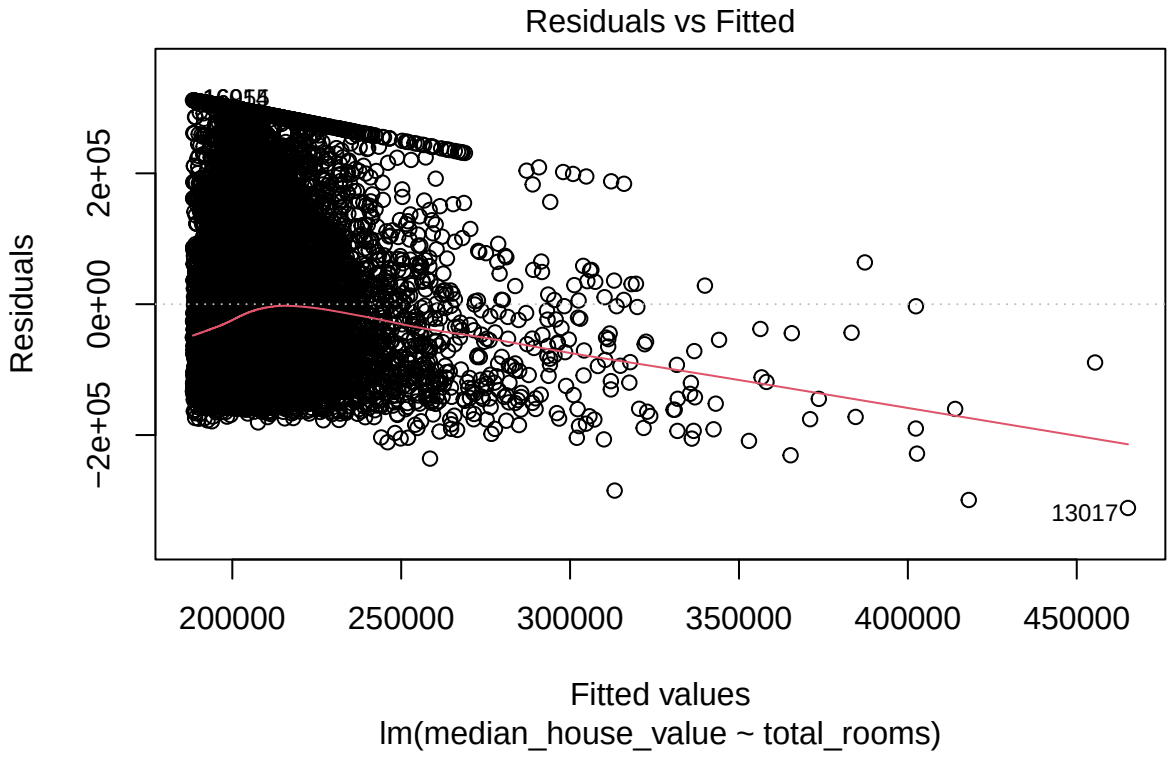
```
##
## Call:
## lm(formula = median_house_value ~ total_rooms, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -311460  -86505  -26706   55721  311644
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.883e+05  1.254e+03  150.13   <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
```

```r
confint(model, level = 0.95)
```

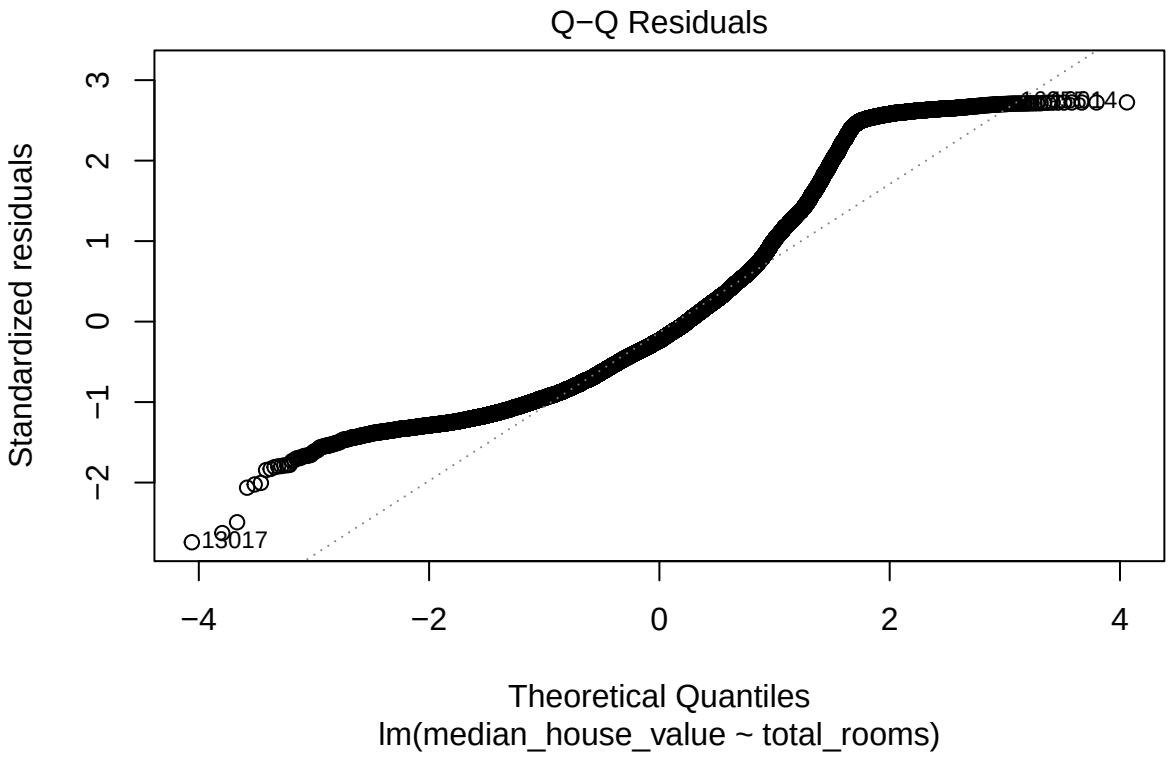
```
##                     2.5 %       97.5 %
## (Intercept) 1.858419e+05 1.907587e+05
## total_rooms 6.323278e+00 7.759102e+00
```

A positive correlation suggests that more rooms in a block increase house values.However, the low R-squared value suggests that the independent variables in the regression model are not effectively explaining the variation in the dependent variable.

```r
plot(lm(median_house_value ~ total_rooms,data = housing_data), which = 1)
```

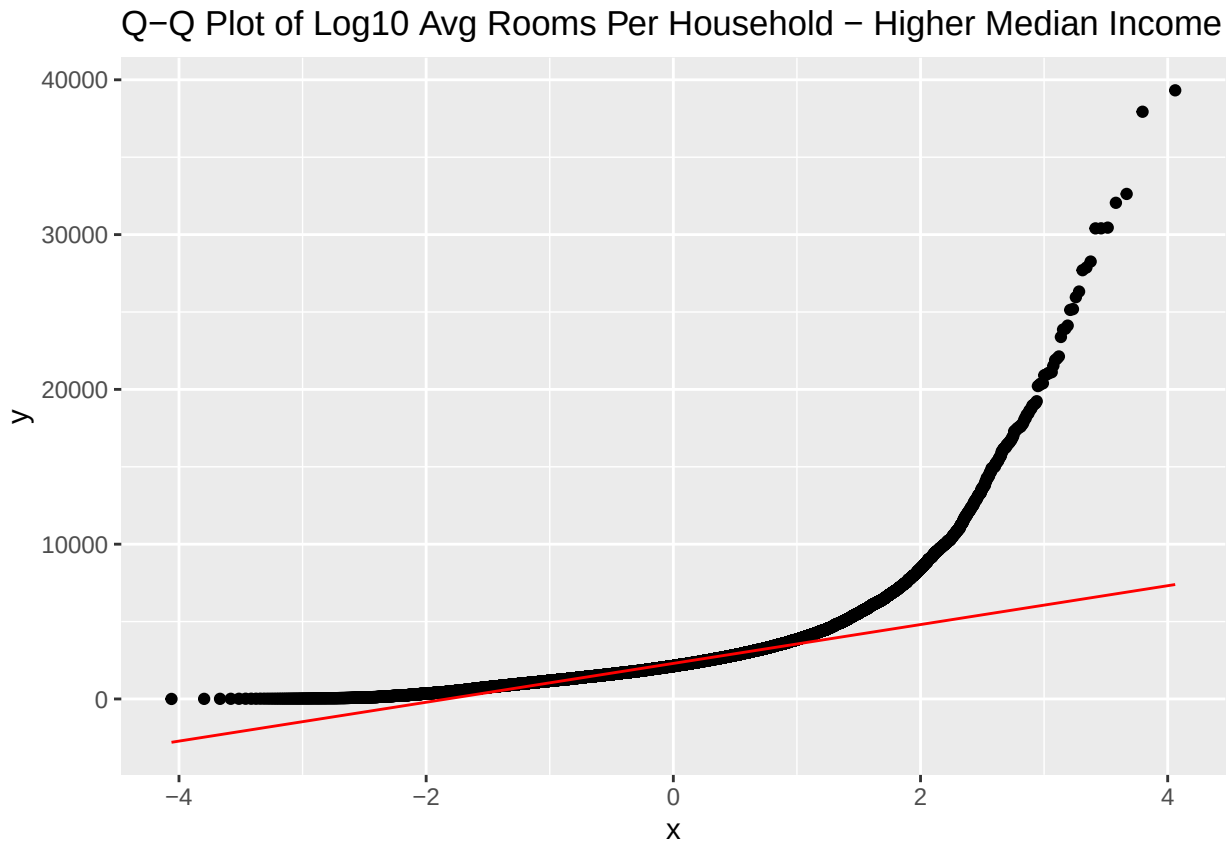



```r
plot(lm(median_house_value ~ total_rooms,data = housing_data), which = 2)
```




```
ggplot(housing_data, aes(sample = housing_data$total_rooms)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  ggtitle("Q-Q Plot of Log10 Avg Rooms Per Household - Higher Median Income")
```

```
## Warning: Use of ‘housing_data$total_rooms‘ is discouraged.
## i Use ‘total_rooms‘ instead.
## Use of ‘housing_data$total_rooms‘ is discouraged.
## i Use ‘total_rooms‘ instead.
```



In fact when we looked at the Residuals values, we can see that it is showing a down shift comparing to the correlation graph, the Q-Q graph it shows a lot of deviation from being normally distributed which means the data is not normally distributed. We can conclude that is it not reliable to predict the median_house_value in the 1990's California area with only using the total_rooms as the independent variables.

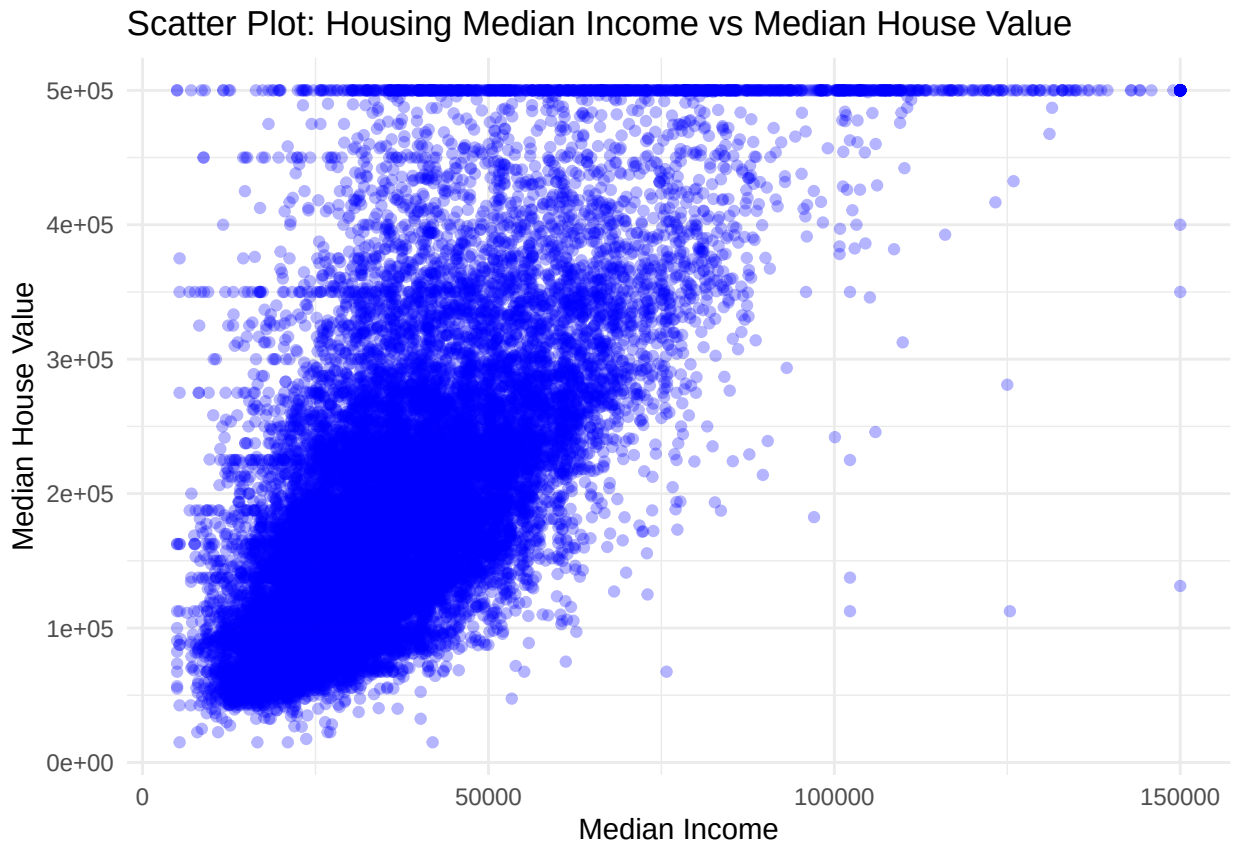
Analysis: Does higher median income lead to larger houses (i.e., greater number of rooms) ?

```
# reading dataset
sf90_housing_data <- read.csv("housing_data_cleaned_oceanencoded.csv")
```

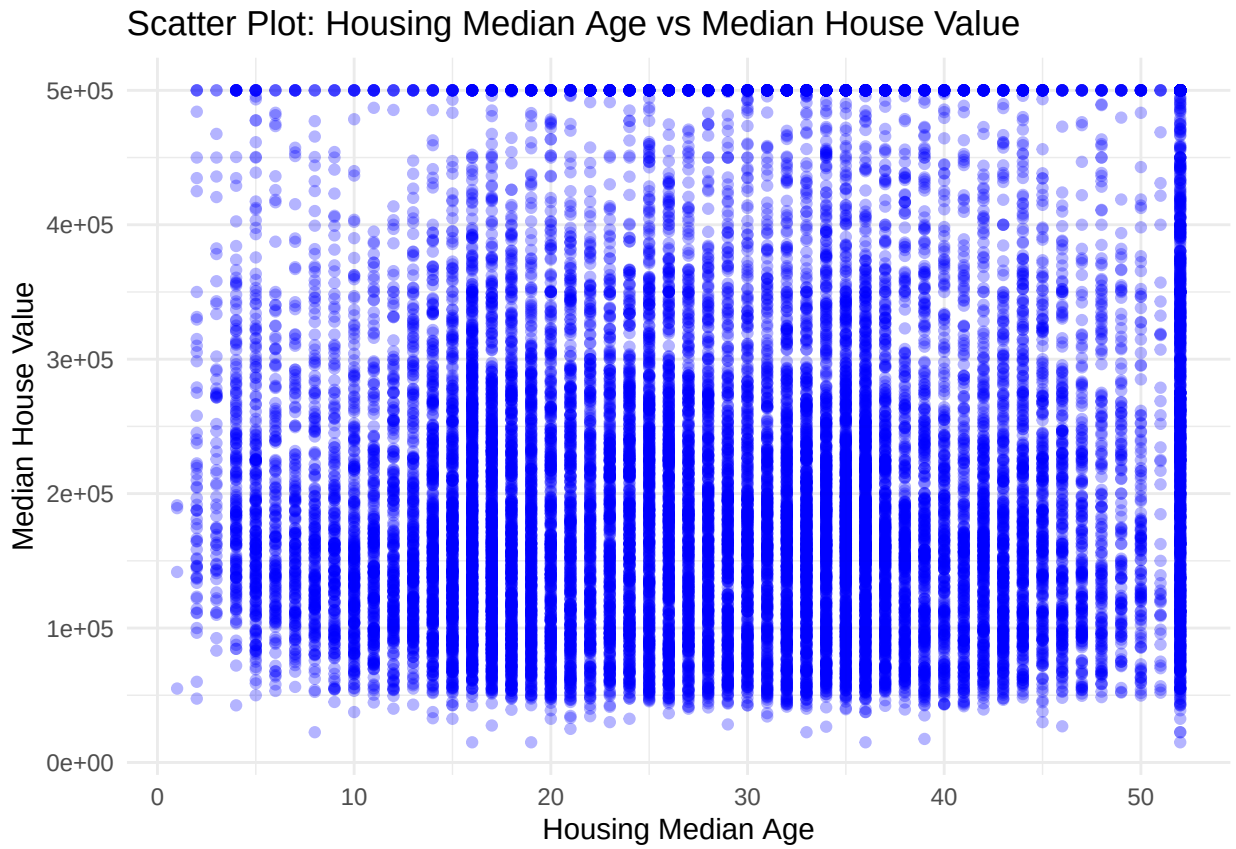
Visualizing the Strongest and Weakest Relationships

We used scatter plots to visualize the weakest and strongest relationships to the median house value

```
ggplot(sf90_housing_data, aes(x = median_income, y = median_house_value)) +
  geom_point(color = 'blue', alpha = 0.3) +
  ggtitle("Scatter Plot: Housing Median Income vs Median House Value") +
  xlab("Median Income") +
  ylab("Median House Value") +
  theme_minimal()
```




```
ggplot(sf90_housing_data, aes(x = housing_median_age, y = median_house_value)) +
  geom_point(color = 'blue', alpha = 0.3) +
  ggtitle("Scatter Plot: Housing Median Age vs Median House Value") +
  xlab("Housing Median Age") +
  ylab("Median House Value") +
  theme_minimal()
```



Analysis on Avg. Rooms Per Household

- Since we established that there is a clear positive relationship between median income and median house value in a block. We wanted to explore what causes higher median income blocks to have higher house values. One factor that we wanted to explore is the Avg Rooms per Household. Since we don't have the official size of the house we can reasonably say that a house more rooms will be larger than one with less rooms.

Our first step is two seprate the two populations. We will do so based on median income. Values higher than the median will be considered higher median income and those equal to or less will be considered lower median income

```
library(dplyr)

#filtering dataset based on income
median_sample_income <- median(sf90_housing_data$median_income)

higher_median_income <- filter(sf90_housing_data, median_income > median_sample_income)

lower_median_income <- filter(sf90_housing_data, median_income <= median_sample_income)
```

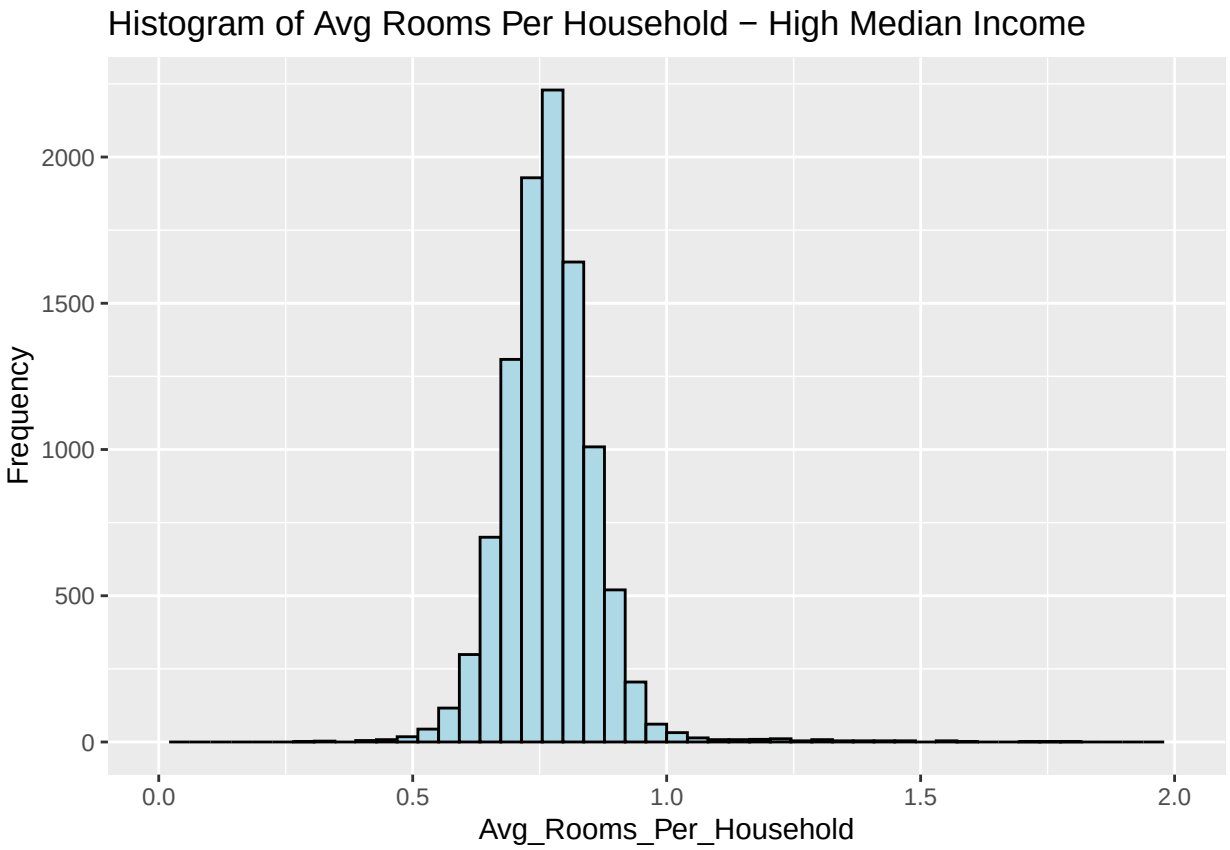
EDA

Before we get to hypothesis testing we want to explore our two populations. We will conduct EDA methods learned in class 4 to gain a clearer picture of our data points.

```r
# Histogrom of avg rooms per households in high income blocks
ggplot(higher_median_income, aes(x = log10(Avg_Rooms_Per_Household))) +
  geom_histogram(bins = 50, fill = "lightblue", color = "black") +
  xlim(0, 2) +
  ggtitle("Histogram of Avg Rooms Per Household - High Median Income") +
  xlab("Avg_Rooms_Per_Household") +
  ylab("Frequency")
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

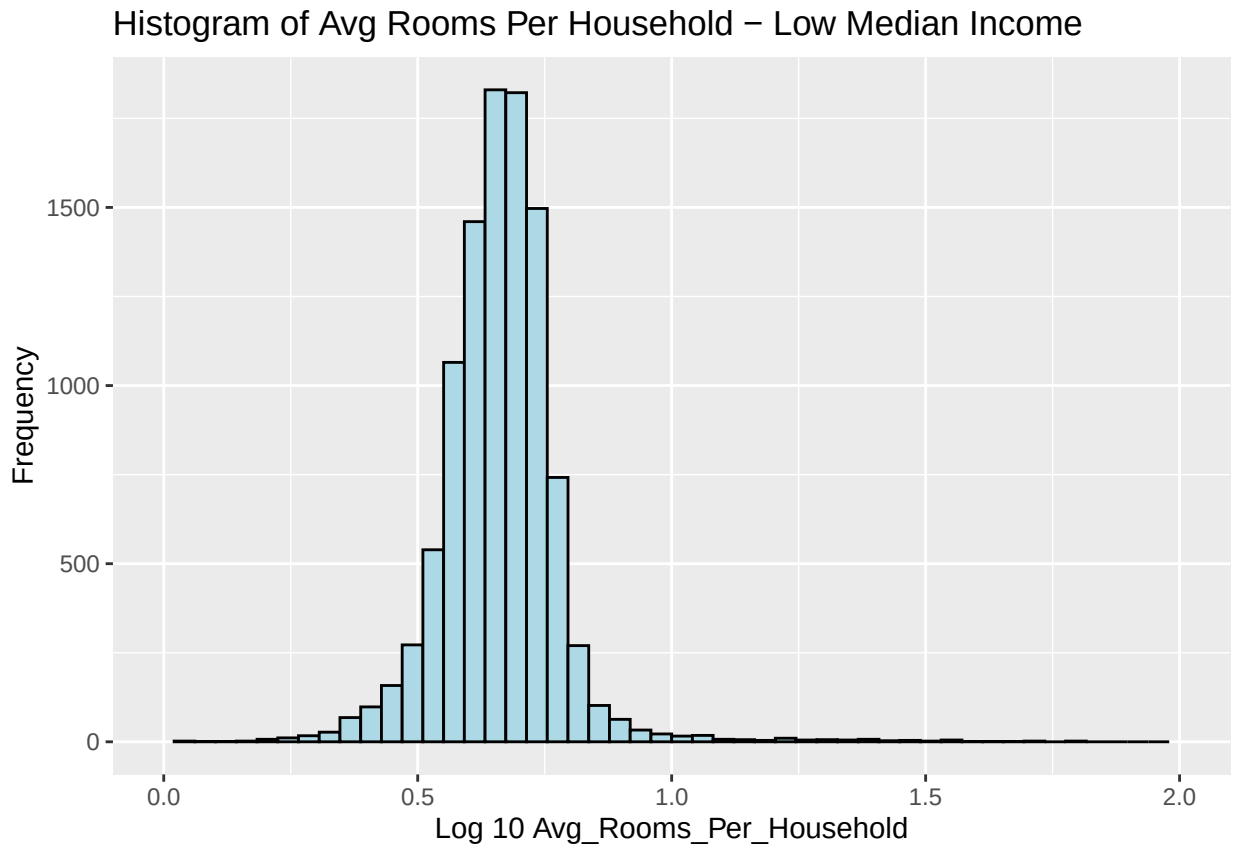
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```




```r
# histogram of avg rooms per households in lower median income blocks
ggplot(lower_median_income, aes(x = log10(Avg_Rooms_Per_Household))) +
  geom_histogram(bins = 50, fill = "lightblue", color = "black") +
  xlim(0, 2) +
  ggtitle("Histogram of Avg Rooms Per Household - Low Median Income") +
  xlab("Log 10 Avg_Rooms_Per_Household") +
  ylab("Frequency")
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

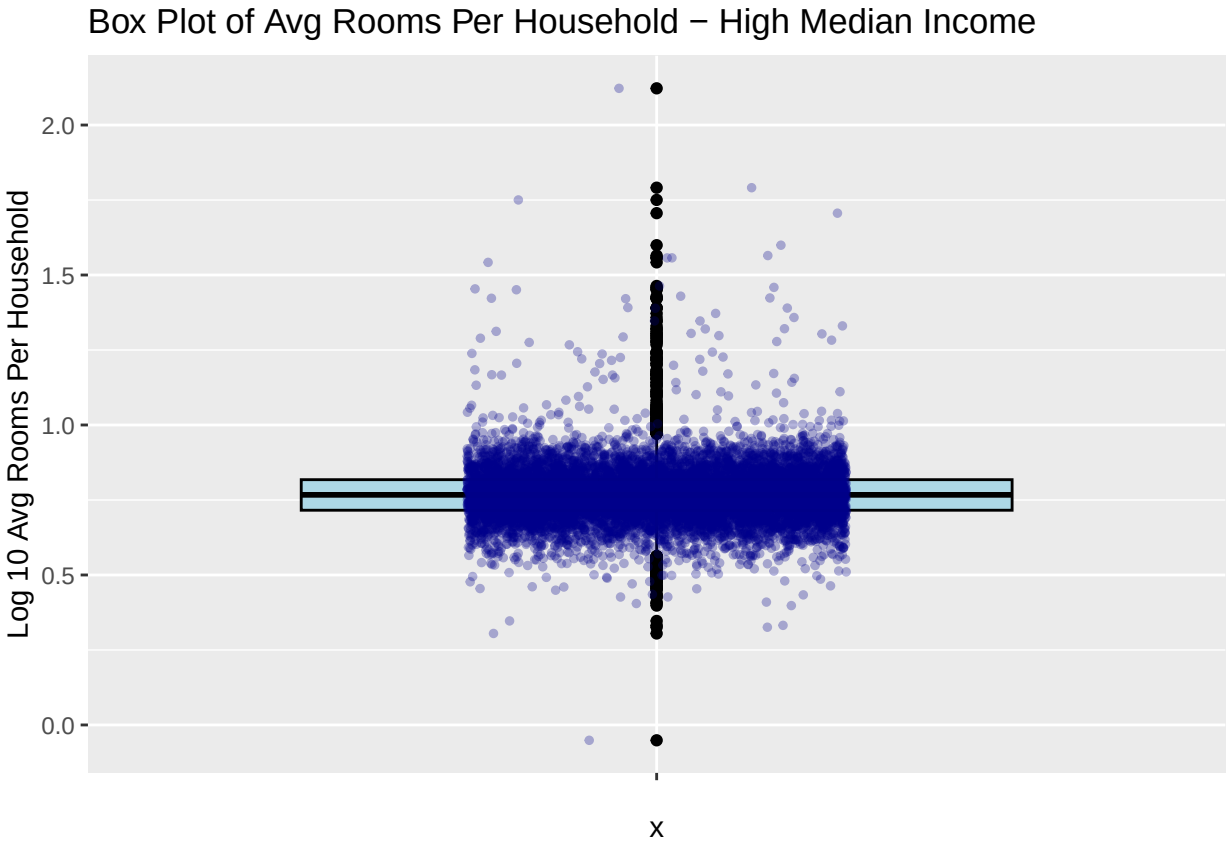
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



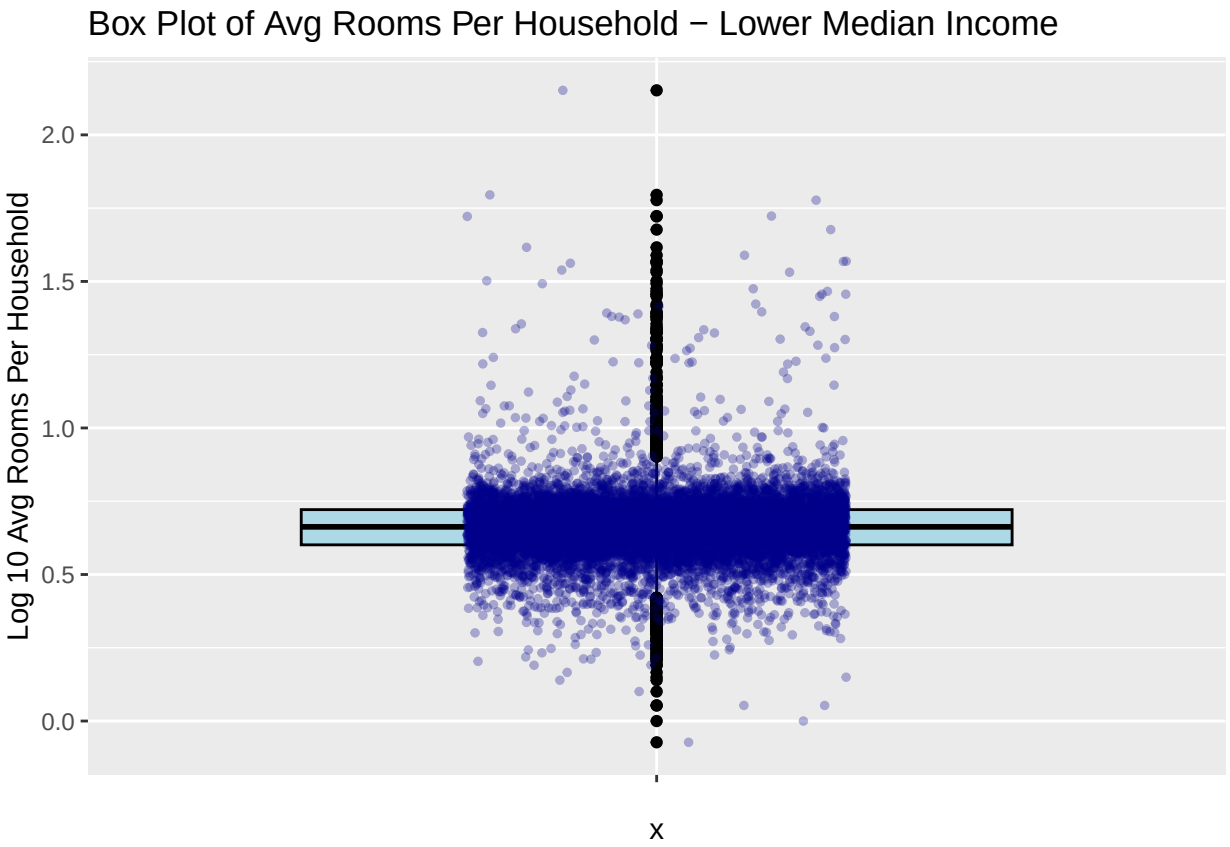
We opted to the normalize the Avg Rooms Per Household values using a log10 transformation because the data was substantially rightly skewed. The positive skew is also observable for both populations after normalization. That indicates that there outliers influencing the data set. Regarding the peak of the histograms we see that the average rooms per households in the higher median income population is centered around $10^{0.75}$ while the peak of lower median income households is centered slightly lower between $10^{0.60}$ *and* $10^{0.70}$

We want to have a clearer look on the extend of the outliers in both datasets so we will use box plots with jitters effect to visualize where the data points lie within their observed ranges.

```
ggplot(higher_median_income, aes(x="",y = log10(Avg_Rooms_Per_Household))) +
  geom_boxplot(fill = "lightblue", color = "black") +
  geom_jitter(width = 0.2, size=1, alpha = 0.3, color = "darkblue") +
  ggtitle("Box Plot of Avg Rooms Per Household - High Median Income") +
  ylab("Log 10 Avg Rooms Per Household")
```




```
ggplot(lower_median_income, aes(x="",y = log10(Avg_Rooms_Per_Household))) +
  geom_boxplot(fill = "lightblue", color = "black") +
  geom_jitter(width = 0.2, size=1, alpha = 0.3, color = "darkblue") +
  ggtitle("Box Plot of Avg Rooms Per Household - Lower Median Income") +
  ylab("Log 10 Avg Rooms Per Household")
```




```
summary(higher_median_income$Avg_Rooms_Per_Household)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.8889   5.1978   5.8500   6.0604   6.5722 132.5333
```

```
summary(lower_median_income$Avg_Rooms_Per_Household)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.8461   3.9916   4.5978   4.8023   5.2648 141.9091
```

From this plot, we can truly see the extent of the outliers. We see median average rooms per households in the high median income dataset to be 5.85 while the median for its counterpart is at 4.578. With regards to the outliers we see max and min values in the ranges of [0.8 to 141] average rooms per households which indicates that certain blocks consists of different types of houses. We suspect that areas located centrally in downtown San Francisco might see more apartments than detached homes which contribute to this.

Checking The Normality of the dataset

In order to use the appropriate method to conduct our hypothesis test we have to visualize our population distribution and see if it approximates the normal distribution or not.

```r
mean_value <- mean(log10(higher_median_income$Avg_Rooms_Per_Household), na.rm = TRUE)
sd_value <- sd(log10(higher_median_income$Avg_Rooms_Per_Household), na.rm = TRUE)


ggplot(higher_median_income, aes(x = log10(Avg_Rooms_Per_Household))) +
  geom_histogram(aes(y = ..density..), bins = 50, fill = "lightblue", color = "black") +
  geom_density(color = "red", size = 1) +
  stat_function(fun = dnorm, args = list(mean = mean_value, sd = sd_value), color = "blue", size = 1) +
  xlim(0, 2) +
  ggtitle("Histogram of Avg Rooms Per Household - Higher Median Income") +
  xlab("Log 10 Avg_Rooms_Per_Household") +
  ylab("Density") +
  theme_minimal()
```

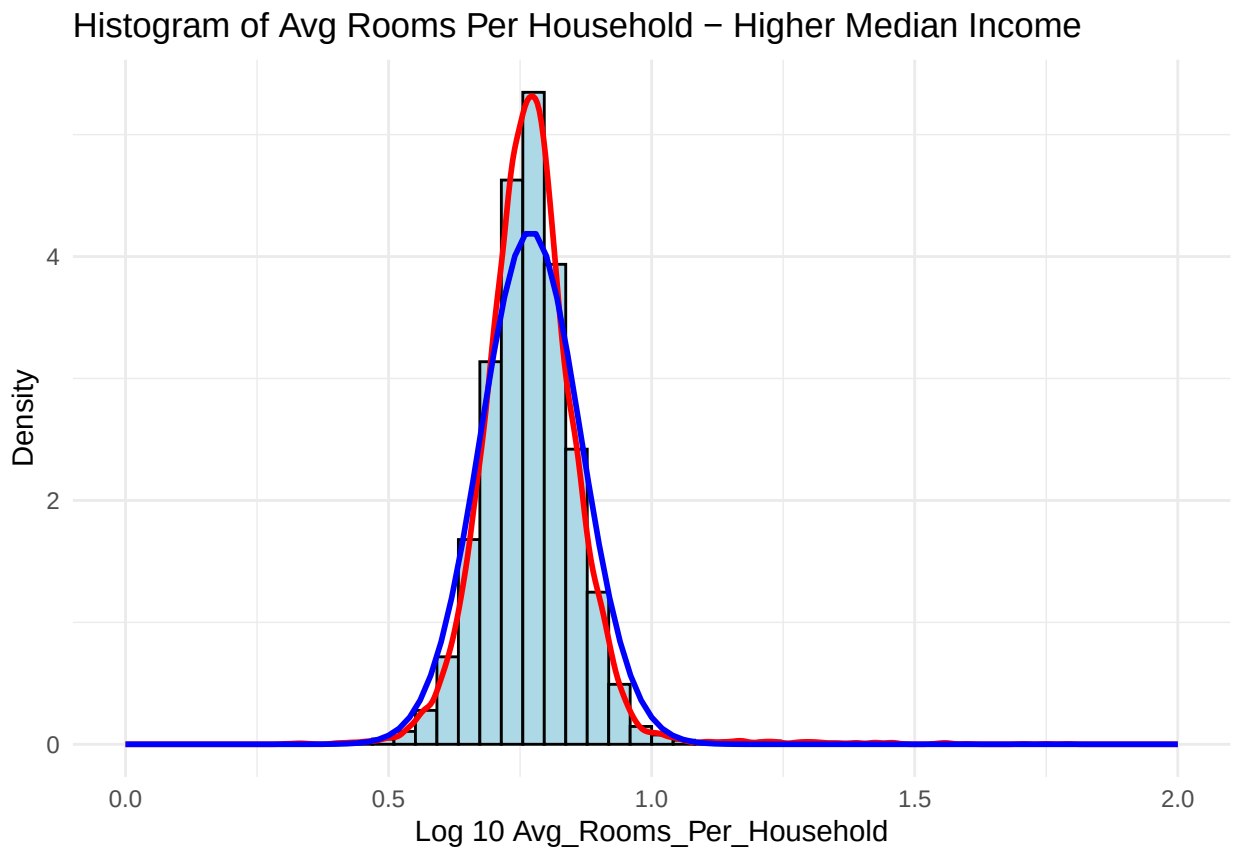
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning: Removed 2 rows containing non-finite outside the scale range
## (‘stat_bin()‘).

## Warning: Removed 2 rows containing non-finite outside the scale range
## (‘stat_density()‘).

## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_bar()‘).
```



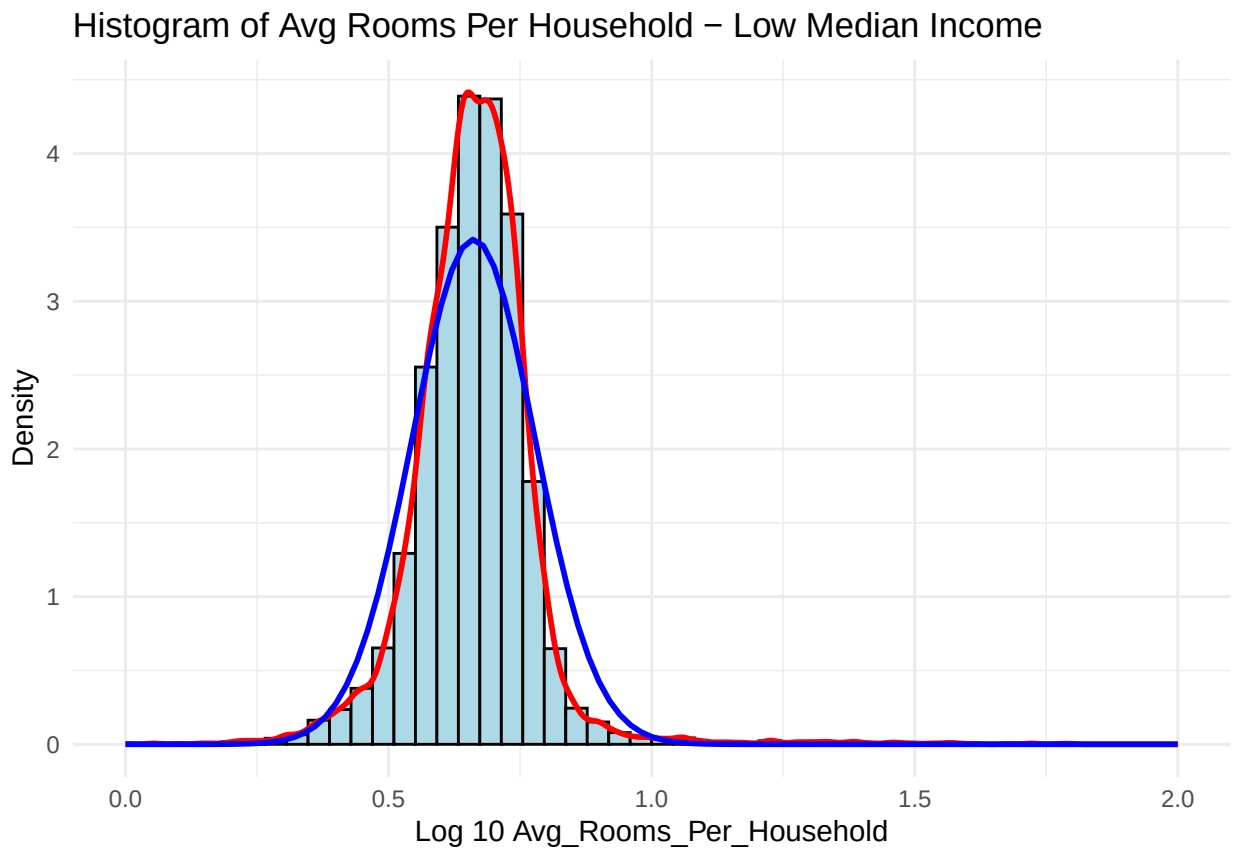

```
mean_value <- mean(log10(lower_median_income$Avg_Rooms_Per_Household), na.rm = TRUE)
sd_value <- sd(log10(lower_median_income$Avg_Rooms_Per_Household), na.rm = TRUE)

ggplot(lower_median_income, aes(x = log10(Avg_Rooms_Per_Household))) +
  geom_histogram(aes(y = ..density..), bins = 50, fill = "lightblue", color = "black") +
  geom_density(color = "red", size = 1) +
  stat_function(fun = dnorm, args = list(mean = mean_value, sd = sd_value), color = "blue", size = 1) +
  xlim(0, 2) +
  ggtitle("Histogram of Avg Rooms Per Household - Low Median Income") +
  xlab("Log 10 Avg_Rooms_Per_Household") +
  ylab("Density") +
  theme_minimal()
```

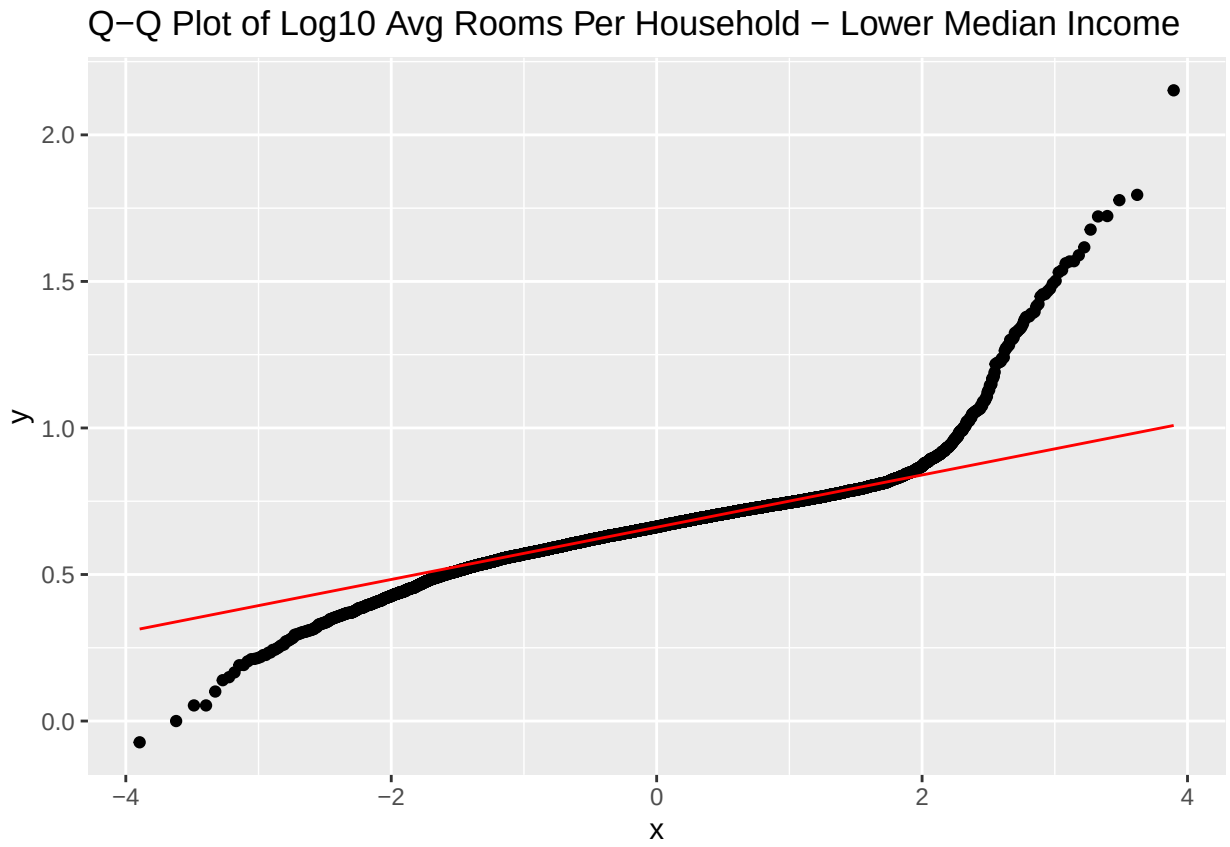
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```

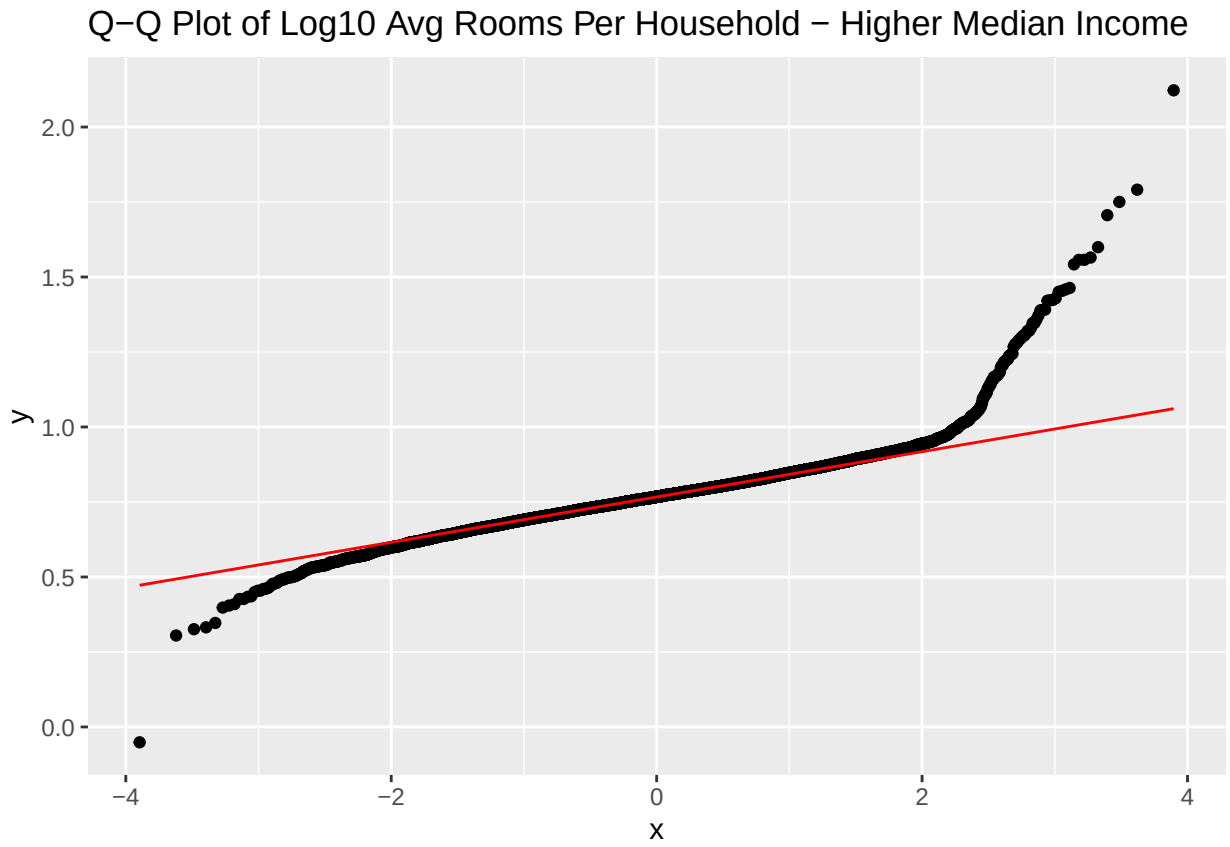
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```




```
ggplot(lower_median_income, aes(sample = log10(Avg_Rooms_Per_Household))) +
  stat_qq() +
  stat_qq_line(color = "red") +
  ggtitle("Q-Q Plot of Log10 Avg Rooms Per Household - Lower Median Income")
```




```
ggplot(higher_median_income, aes(sample = log10(Avg_Rooms_Per_Household))) +
  stat_qq() +
  stat_qq_line(color = "red") +
  ggtitle("Q-Q Plot of Log10 Avg Rooms Per Household - Higher Median Income")
```



Based on the density histograms and the qq plots we observed what we noticed earlier in our EDA. The observed distribution deviate extremely from normality in the tail portions. If we base our hypothesis testing on the assumption of normality given the density plots above then our inferences might produce unreliable/inaccurate results. Having a large portion of the data set influenced by the tails than in a normal distribution would cause our confidence intervals to be narrower understating the likelihood of extreme values which is quite substantial in our data set. This would cause the confidence intervals to not properly account for the actual variability in the data. Therefore using the bootstap approach will likley reduce more accurate results.

Hypthoesis Testing

We want to test to see if there is statistical evidence to higher median income blocks having more avg rooms per household on average than lower median income blocks. We can state our hypothesis test as

μ_1 : Population Mean of Average Rooms per Household in Higher Income Blocksμ_2 : Population Mean of Average Rooms p

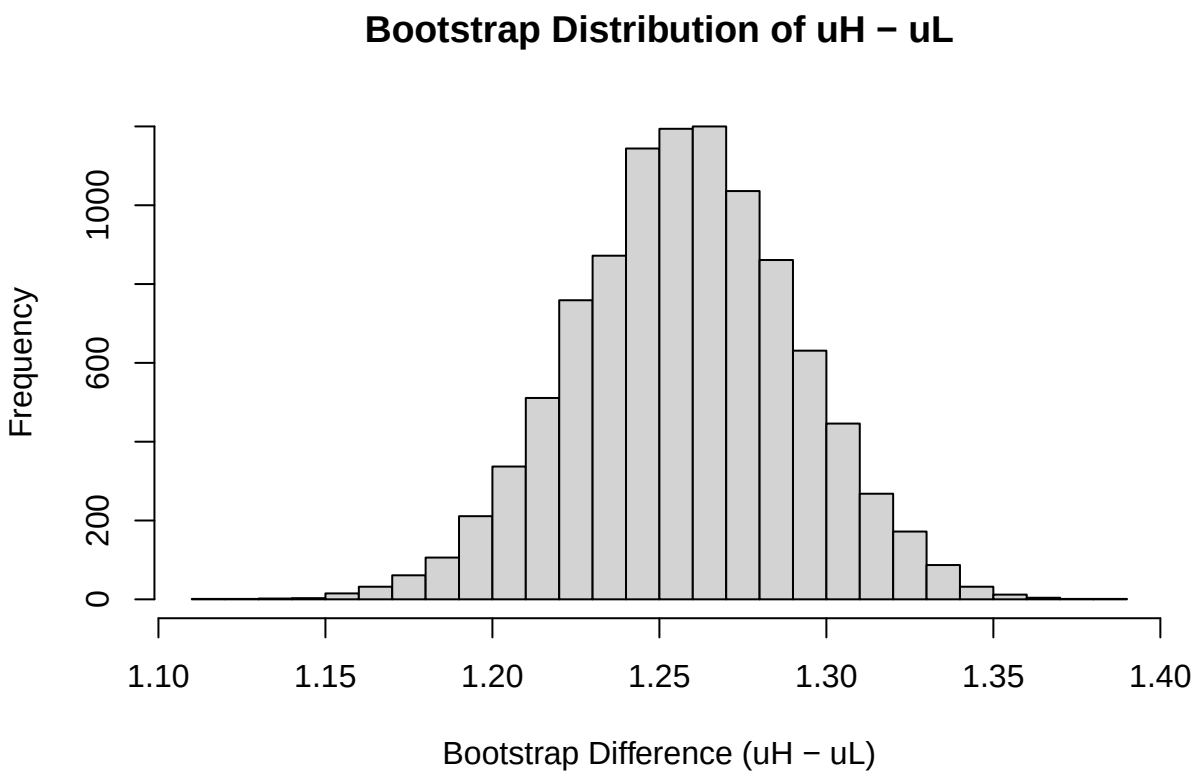
```
bootstap <- 10000

M_high_income = do(bootstap) *mean(sample(higher_median_income$Avg_Rooms_Per_Household, replace = TRUE))

M_low_income = do(bootstap)*mean(sample(lower_median_income$Avg_Rooms_Per_Household, replace = TRUE))

bootstap_diff <- (M_high_income$mean - M_low_income$mean)
```

```r
hist(bootstap_diff, breaks = 30, main="Bootstrap Distribution of uH - uL",
     xlab="Bootstrap Difference (uH - uL)")
```



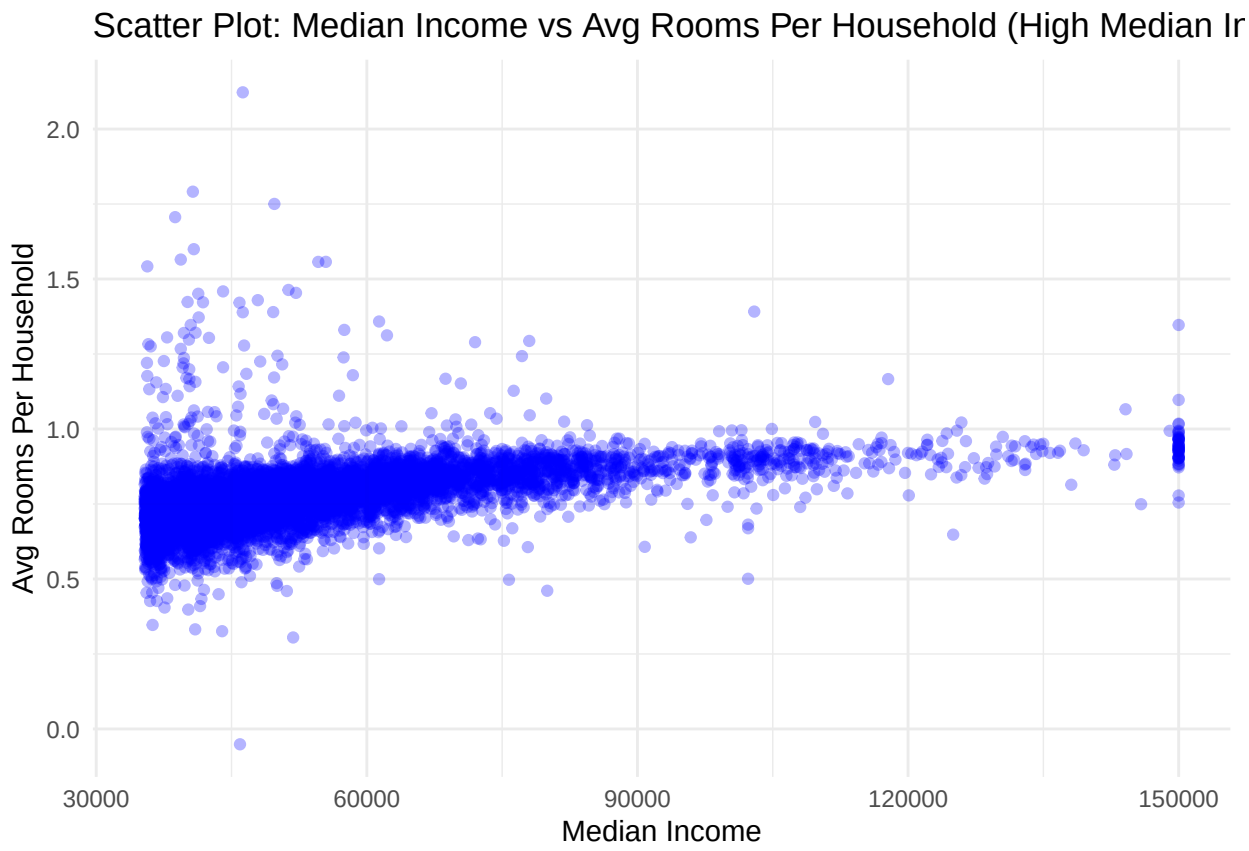

```r
quantile(bootstap_diff , c(.025, .975))
```

```
##     2.5%    97.5%
## 1.191841 1.323473
```

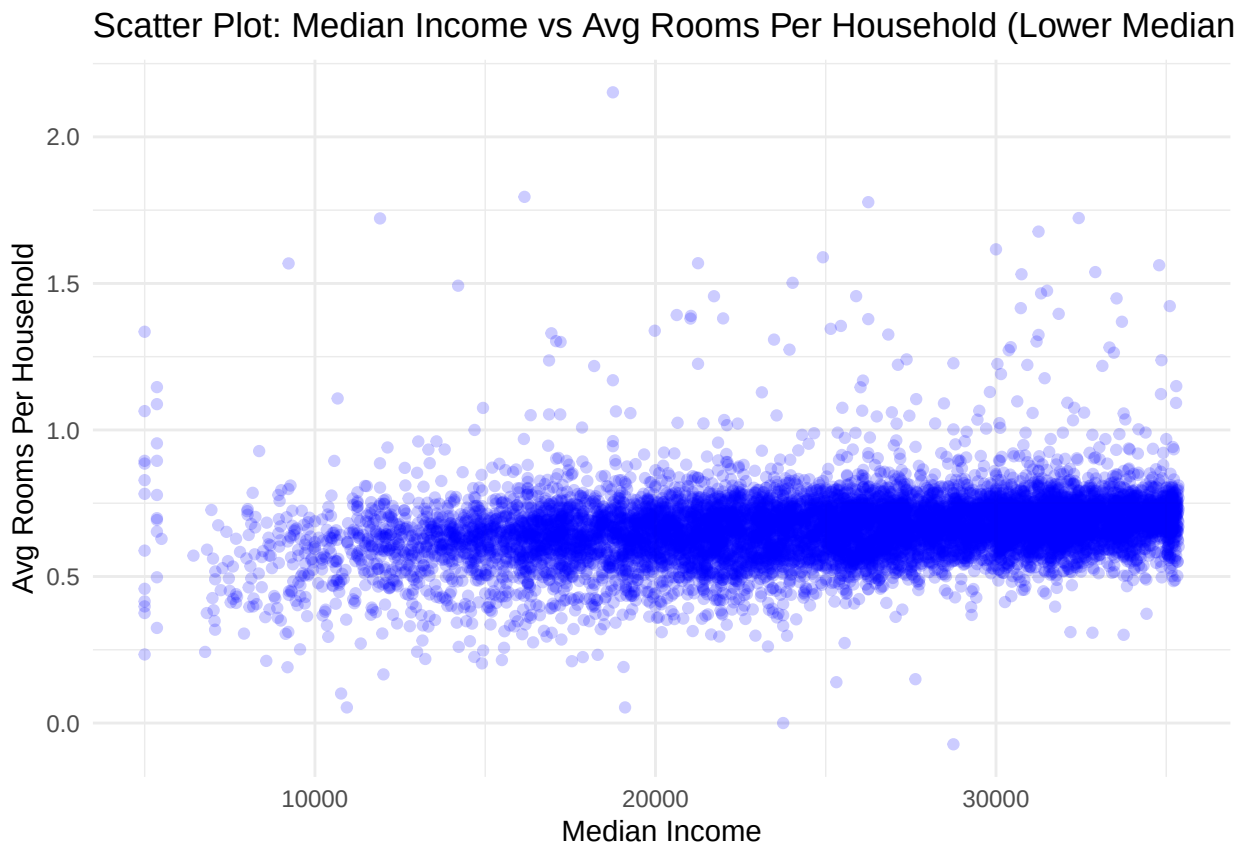
Sufficient evidence to reject the null hypotheses. We are 95% confident that the mean avg rooms per household in higher income blocks will be 1.191 to 1.323 rooms greater than the mean avg rooms per households in lower income blocks

Linear Regression

```r
# Scatter plot for Median Income vs Avg Rooms Per Household in higher median income dataset
ggplot(higher_median_income, aes(x = median_income, y = log10(Avg_Rooms_Per_Household))) +
  geom_point(color = "blue", alpha = 0.3) +  # Add scatter points with transparency
  ggtitle("Scatter Plot: Median Income vs Avg Rooms Per Household (High Median Income)") +
  xlab("Median Income") +
  ylab("Avg Rooms Per Household") +
  theme_minimal()
```




```r
# Scatter plot for Median Income vs Avg Rooms Per Household in higher median income dataset
ggplot(lower_median_income, aes(x = median_income, y = log10(Avg_Rooms_Per_Household))) +
  geom_point(color = "blue", alpha = 0.2) +  # Add scatter points with transparency
  ggtitle("Scatter Plot: Median Income vs Avg Rooms Per Household (Lower Median Income)") +
  xlab("Median Income") +
  ylab("Avg Rooms Per Household") +
  theme_minimal()
```



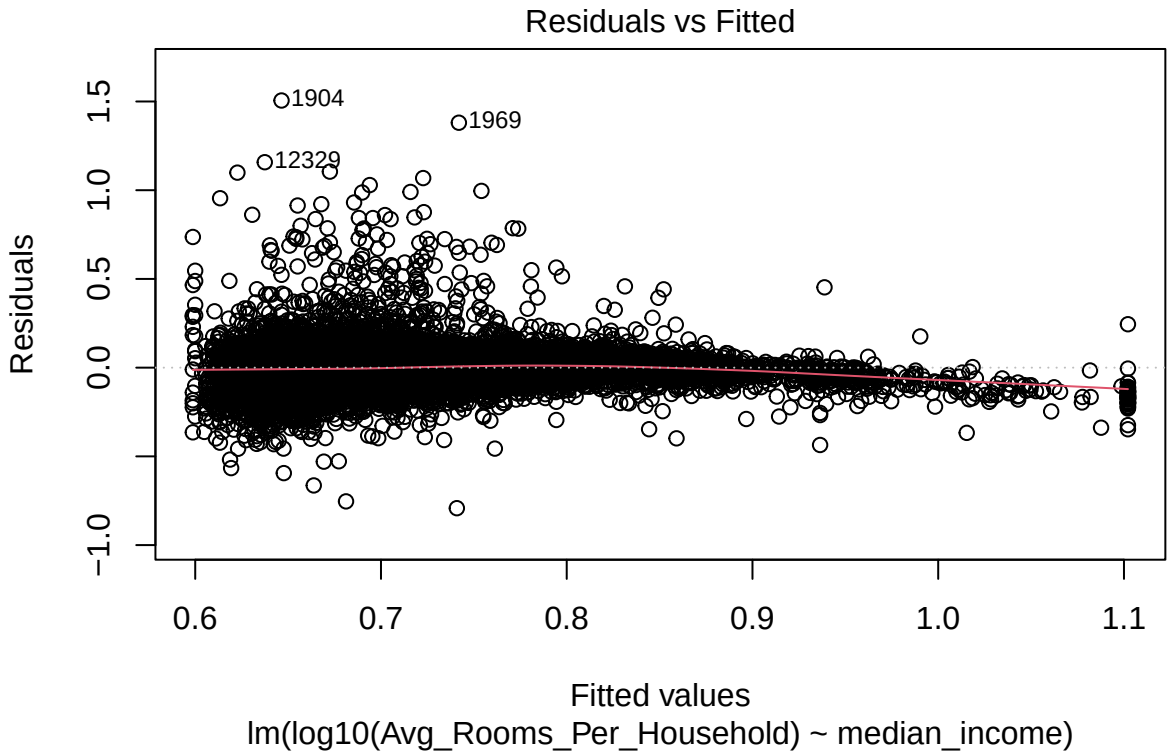

```
room_num_reg <- lm(log10(Avg_Rooms_Per_Household) ~ median_income , data = sf90_housing_data )
summary(room_num_reg)
```

```
##
## Call:
## lm(formula = log10(Avg_Rooms_Per_Household) ~ median_income,
##     data = sf90_housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79200 -0.04846  0.00230  0.04592  1.50555
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.814e-01  1.580e-03  368.05   <2e-16 ***
## median_income 3.472e-06  3.663e-08   94.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09945 on 20431 degrees of freedom
## Multiple R-squared:  0.3054, Adjusted R-squared:  0.3054
## F-statistic:  8982 on 1 and 20431 DF,  p-value: < 2.2e-16
```

The median income coefficient has a significant p-value. With every $1 increase in median income , average rooms per household in a block with increase by 10^3.472e-06 rooms.

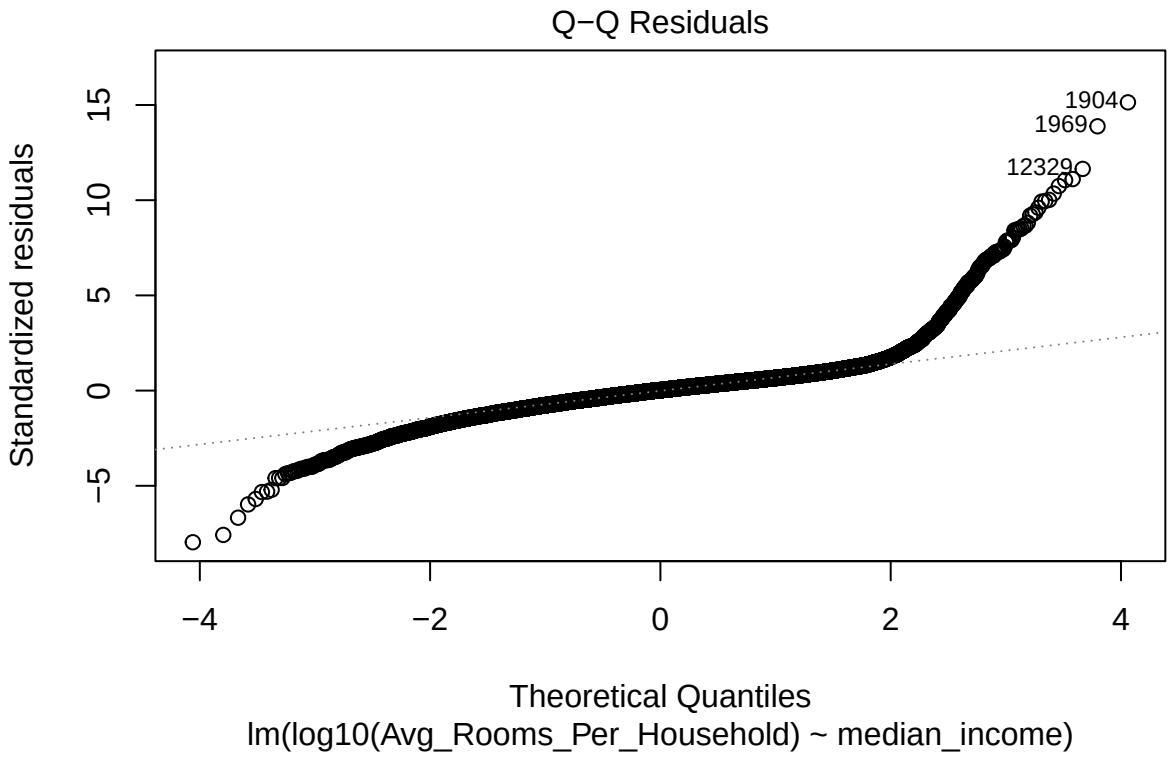
P values indicate that a positive relationship between median income and avg rooms per household is statistically significant. Our R squared value is quite surprising, it indicates that 30% of the variability of the log transformed avg rooms per household is explained by median income. In order to understand more we need to explore our residuals

```
plot(room_num_reg, which = 1)
```



Looking at the residuals we see that a good portion of them are clustered around 0. We do however see the curved line at the higher portion of the fitted patterns, indicating that the errors are not independent. This allows us to conclude that a linear fit might not be the best one.

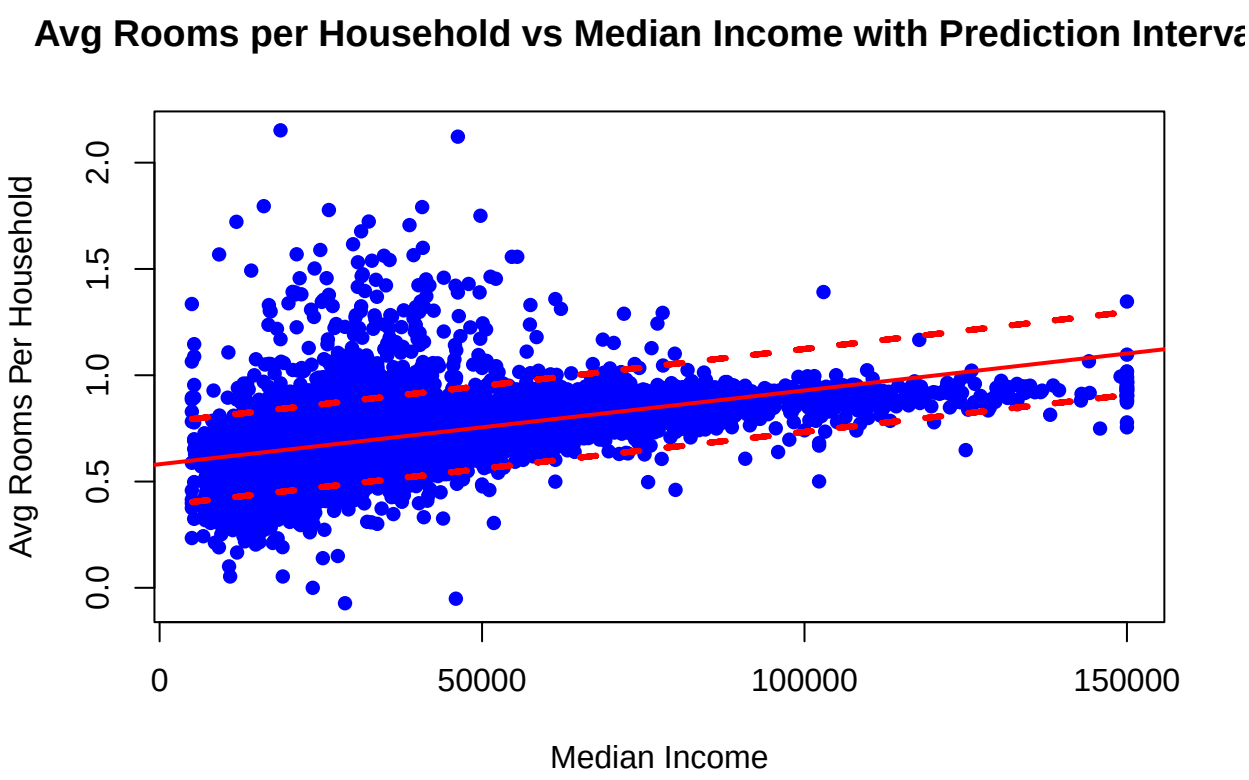
```
plot(room_num_reg, which = 2)
```



Looking at the desnity plot of the residuals, it deviates from normality on the tails, which is another violation of the conditions of the linear model.

```
plot(sf90_housing_data$median_income, log10(sf90_housing_data$Avg_Rooms_Per_Household),
     main = "Avg Rooms per Household vs Median Income with Prediction Intervals",
     xlab = "Median Income", ylab = "Avg Rooms Per Household",
     pch = 16, col = "blue", xlim = range(sf90_housing_data$median_income))
abline(lm(log10(Avg_Rooms_Per_Household) ~ median_income , data = sf90_housing_data),col="red",lwd=2)

x <- data.frame(median_income =seq(min(sf90_housing_data$median_income)
                                   , max(sf90_housing_data$median_income), by = 0.1))
predictions_intervals <- data.frame(predict(room_num_reg, x
                                            , interval="prediction", level=0.95))
lines(x$median_income,predictions_intervals[,2],lwd=3,col="red",lty=2)
lines(x$median_income,predictions_intervals[,3],lwd=3,col="red",lty=2)
```



Conclusion and Future Steps

We found statistical evidence to conclude that median income can explain 48% of the variability in median house values. The relationship between those two variables is positive, indicating a $1 increase in median income will be followed with a $4.1 rise in median house value. In addition, a positive relationship between the baseline proximity to coastlines '<1 Hour' and median house values, and a negative one vice versa. We also found statistical evidence that blocks with higher median incomes will on average have between 1.91 to 1.32 average rooms per household more than blocks with lower incomes.

However, it is important to note some limitations within this dataset and the implications for future directions of this analysis. One important limitation of this dataset was that it does not differentiate between housing types, such as detached homes versus apartments. The lack of this distinction can lead to a skewed interpretation of the relationship between average rooms per household. Additionally, we wanted to investigate the impact of family size but were limited as we would have to infer family size from the ratio of population to the number of households. This would be based on assumptions that overlook diverse living situations and may provide misleading results about the relationship. Knowing additional features such as housing type, family size, and house size (in ft2) would help make for a more comprehensive analysis of features impacting house value.

In conclusion, our finding highlights the significant influence of some features, such as income and distance to ocean, on housing values in California while other features have limited influence. This distinction suggests it is more important to focus on key features when analyzing housing values.

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.