Project Report:
Analysis of Housing Values in California

1. Goal, Hypothesis, Business Question
2. Data Set And Variables
3. Data Pipeline
4. Visualizing Data for Insights
5. Fitting the Model
6. Summary of Metrics
7. Conclusion/Recommendation

BY ISABELITA SERVANDO, WINTER QUARTER 2023

GOAL OF THIS ANALYSIS

To determine the factors that drive median house values in California

HYPOTHESIS

As the saying goes, “Location, location, location.”

In California (or anywhere else for that matter), location – or in the case of this analysis, latitude, longitude and proximity to the ocean – is believed to be the strongest driver of median house values.

WHAT’S THE BUSINESS QUESTION?

What factors determine median house values in California?

Is location truly the strongest predictor of house values?
If so, should real estate companies prioritize selling houses in prime locations to ensure faster turnaround and maximum profits?

THE DATA SET

housing.csv, from the California Housing data set

This data set first appeared in a 1997 paper titled Sparse Spatial Autoregressions by Pace, R. Kelley and Ronald Barry, published in the Statistics and Probability Letters journal.

The researchers built it using the 1990 California census data.

It contains one row per census block group*, i.e. each row/observation pertains to a group of houses representing medians for groups of houses in close proximity to each other.

The data set has 20640 observations, 10 variables.

*A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people).

THE VARIABLES

1. Median_house value is the target or response variable, what we're trying to predict

   These are the predictor variables, what we will test if they can predict or determine median house value.

2. longitude
3. latitude
4. housing_median_age
5. total_rooms
6. total_bedrooms
7. population
8. households
9. median_income
10. ocean_proximity

All variables are numerical, except for ocean_proximity, which is a character that, as the name implies, indicates a house's proximity to the ocean.

DATA PIPELINE

A series of transformations, aka munging, cleaning or wrangling are needed to prepare the data for visualization and machine learning.

The following transformations were done for this data set:

- Transformed the variable ocean_proximity from character to factor to produce these levels: "<1H OCEAN" "INLAND" "ISLAND" "NEAR BAY" "NEAR OCEAN"
- Split the ocean_proximity variable into a number of binary categorical variables consisting of 1s and 0s.
- Ran a correlation of all numerical variables, including the target variable (median_house_value) to determine their relationship (how strongly they affect each other, whether positively or negatively)
- Imputed NAs in total_bedrooms with the median value for total_bedrooms
- Created two new variables: mean_bedrooms and mean_rooms (derived from: total bedrooms and total rooms by number of households
- Removed the variables total_bedrooms and total_rooms after creating the above variables
- Scaled each numerical variable except for median_house_value (as this is theresponse variable), and the binary categorical variables.
- Created a new data frame named cleaned_housing with the following variables.
    - "NEAR BAY"
    - "<1H OCEAN"
    - "INLAND"
    - "NEAR OCEAN"
    - "ISLAND"
    - "longitude"
    - "latitude"
    - "housing_median_age"
    - "population"
    - "households"
    - "median_income"
    - "mean_bedrooms"
    - "mean_rooms"
    - "median_house_value"
- Divided the data set into training and test sets

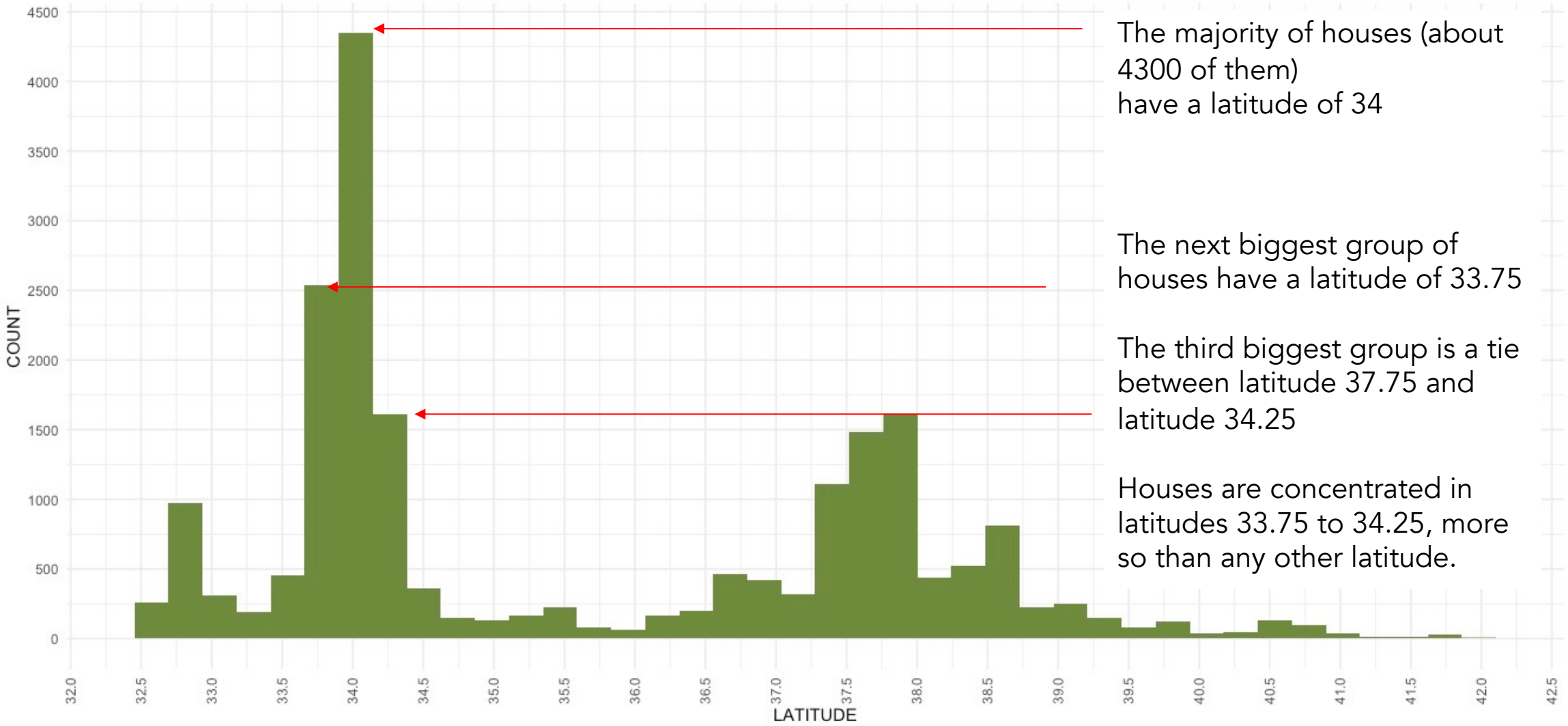
- Training set for regression
- Test set for prediction

VISUALIZING DATA FOR INSIGHTS

- HISTOGRAMS FOR ALL NUMERICAL VARIABLES
- BOXPLOTS FOR ALL NUMERICAL VARIABLES
- BOXPLOTS FOR HOUSING MEDIAN AGE X OCEAN PROXIMITY,

MEDIAN INCOME X OCEAN PROXIMITY, MEDIAN HOUSE VALUE X OCEAN PROXIMITY

HISTOGRAM: LATITUDE



The majority of houses (about 4300 of them)
have a latitude of 34

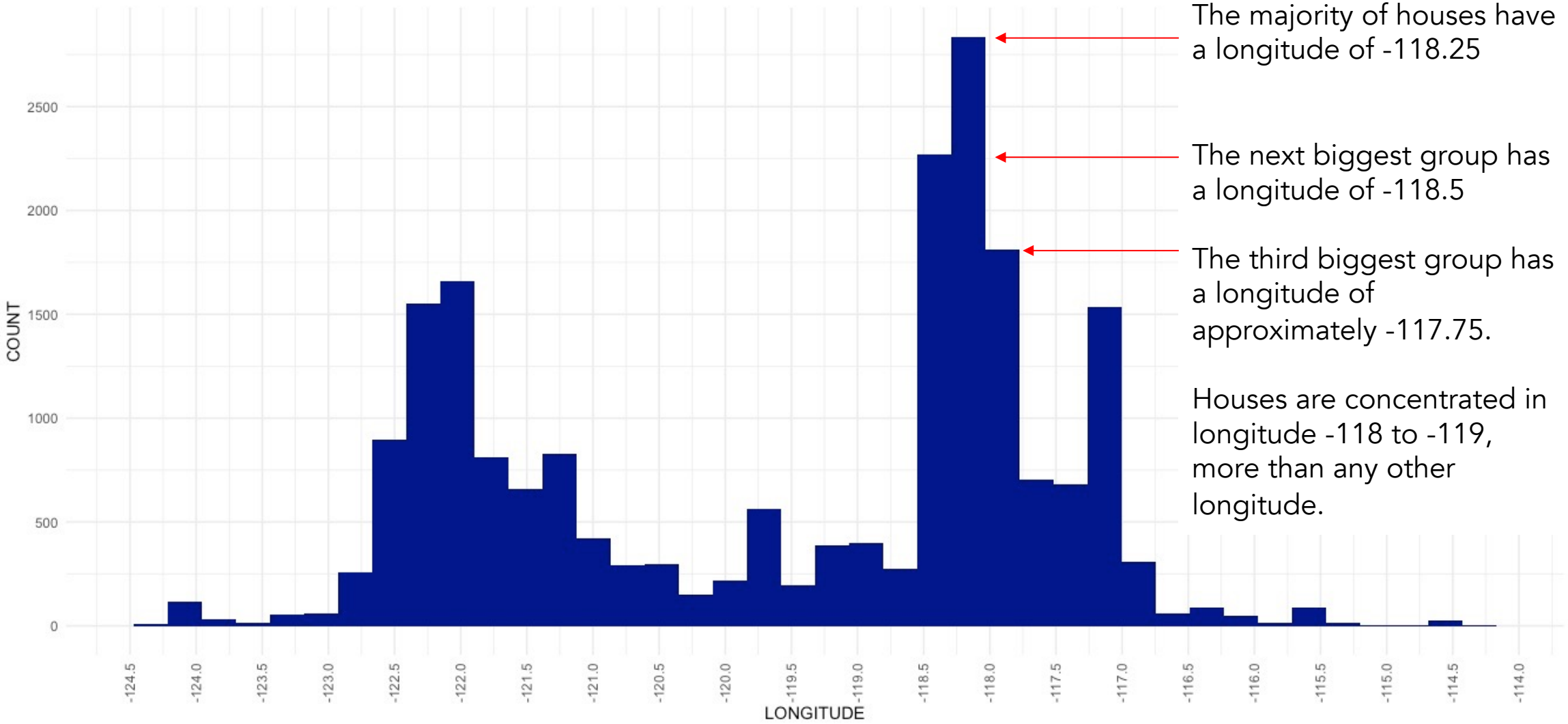
The next biggest group of houses have a latitude of 33.75

The third biggest group is a tie between latitude 37.75 and latitude 34.25

Houses are concentrated in latitudes 33.75 to 34.25, more so than any other latitude.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

HISTOGRAM: LONGITUDE



The majority of houses have a longitude of -118.25

The next biggest group has a longitude of -118.5

The third biggest group has a longitude of approximately -117.75.

Houses are concentrated in longitude -118 to -119, more than any other longitude.

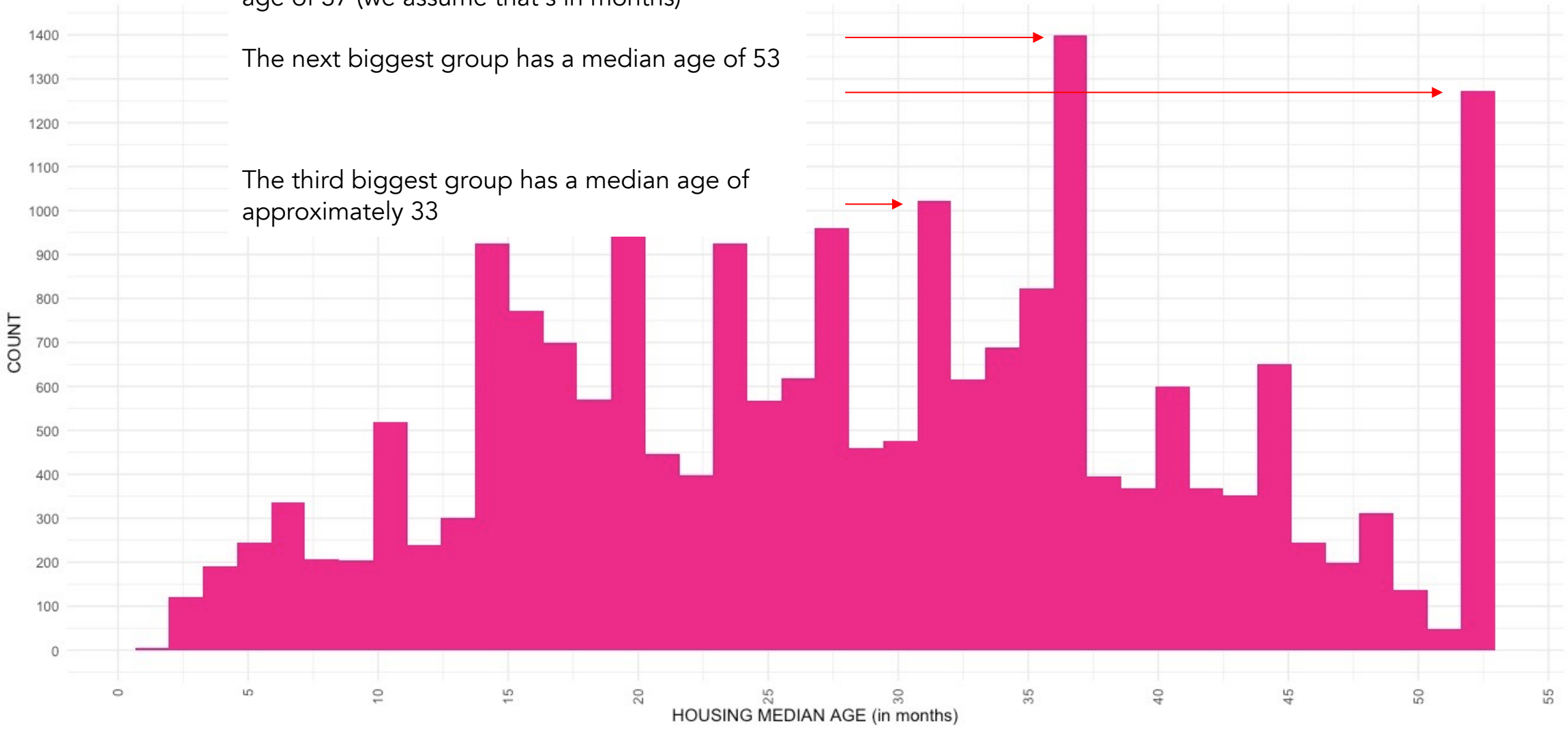
NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

HISTOGRAM: HOUSING MEDIAN AGE

The majority of houses has a housing median age of 37 (we assume that's in months)

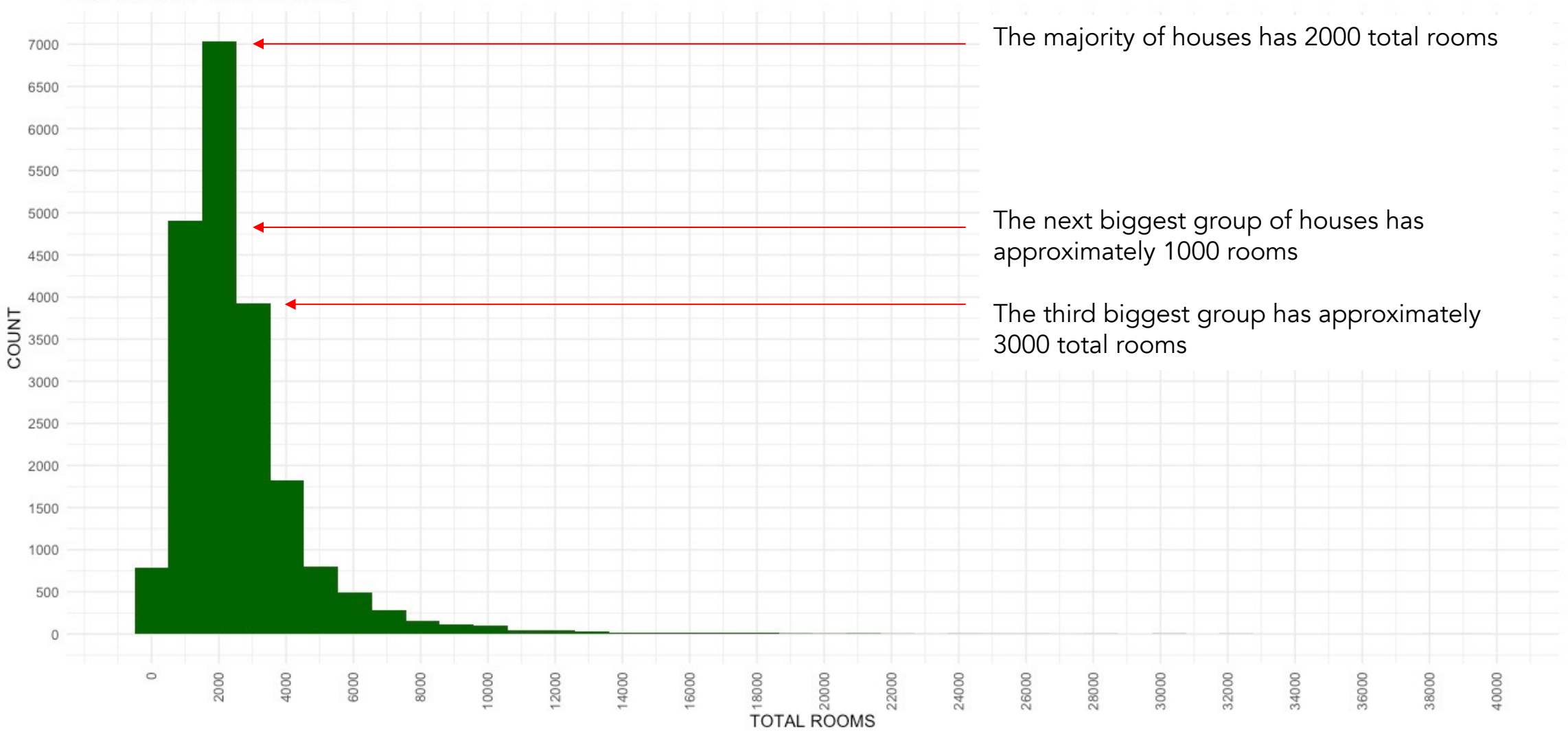
The next biggest group has a median age of 53

The third biggest group has a median age of approximately 33



NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

HISTOGRAM: TOTAL ROOMS



The majority of houses has 2000 total rooms

The next biggest group of houses has approximately 1000 rooms

The third biggest group has approximately 3000 total rooms

NOTE:
These values don't represent individual houses, but rather groups of houses in close proximity to each other.
That's why total rooms are in the thousands, instead of the usual 2-3 or 4-5.

HISTOGRAM: TOTAL BEDROOMS

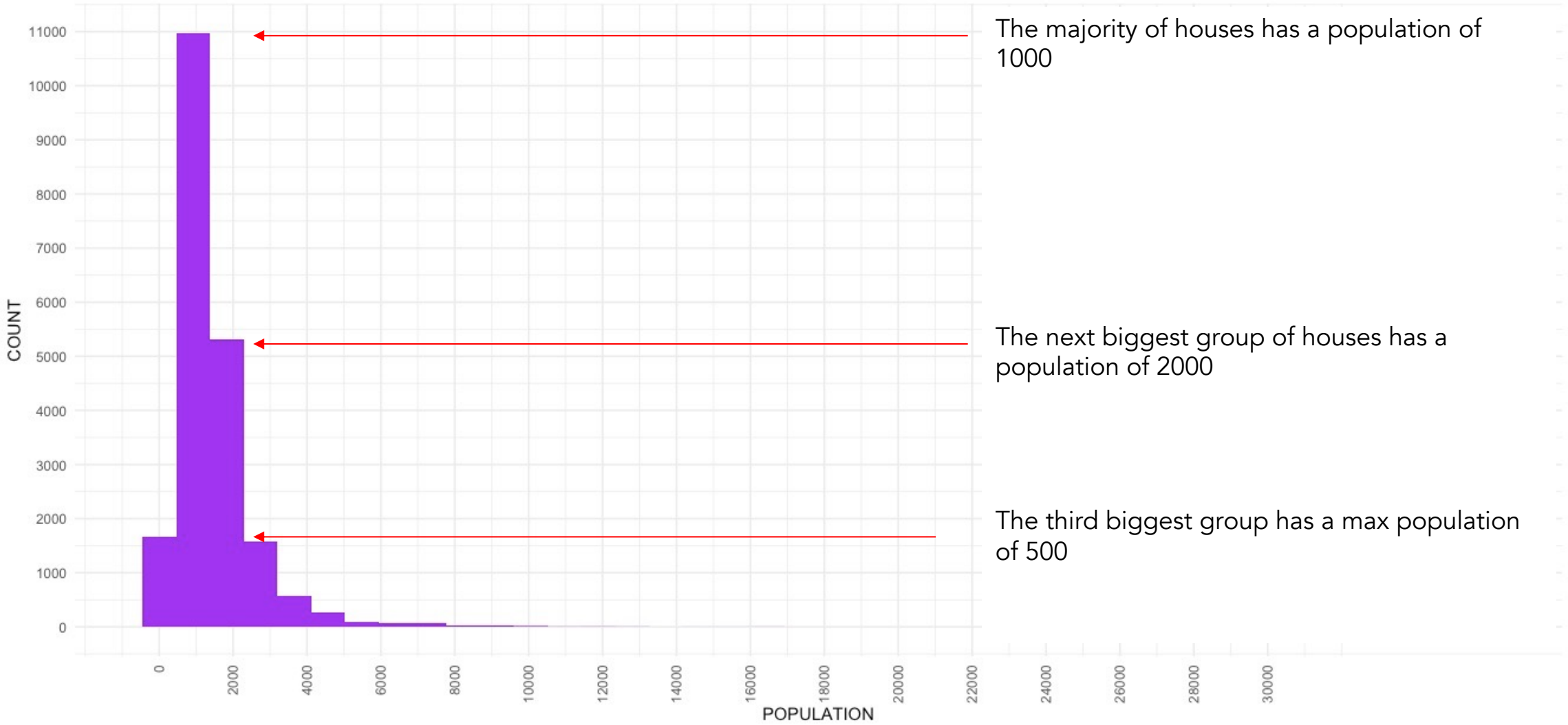
The majority of houses have approximately 300 total bedrooms

The next biggest number of houses have 500 total bedrooms

The third biggest number of houses have approximately 150 total bedrooms.

NOTE:
These values don't represent individual houses, but rather groups of houses in close proximity to each other.
That's why total bedrooms are in the hundreds, instead of the usual 2-3 bedrooms.

HISTOGRAM: POPULATION



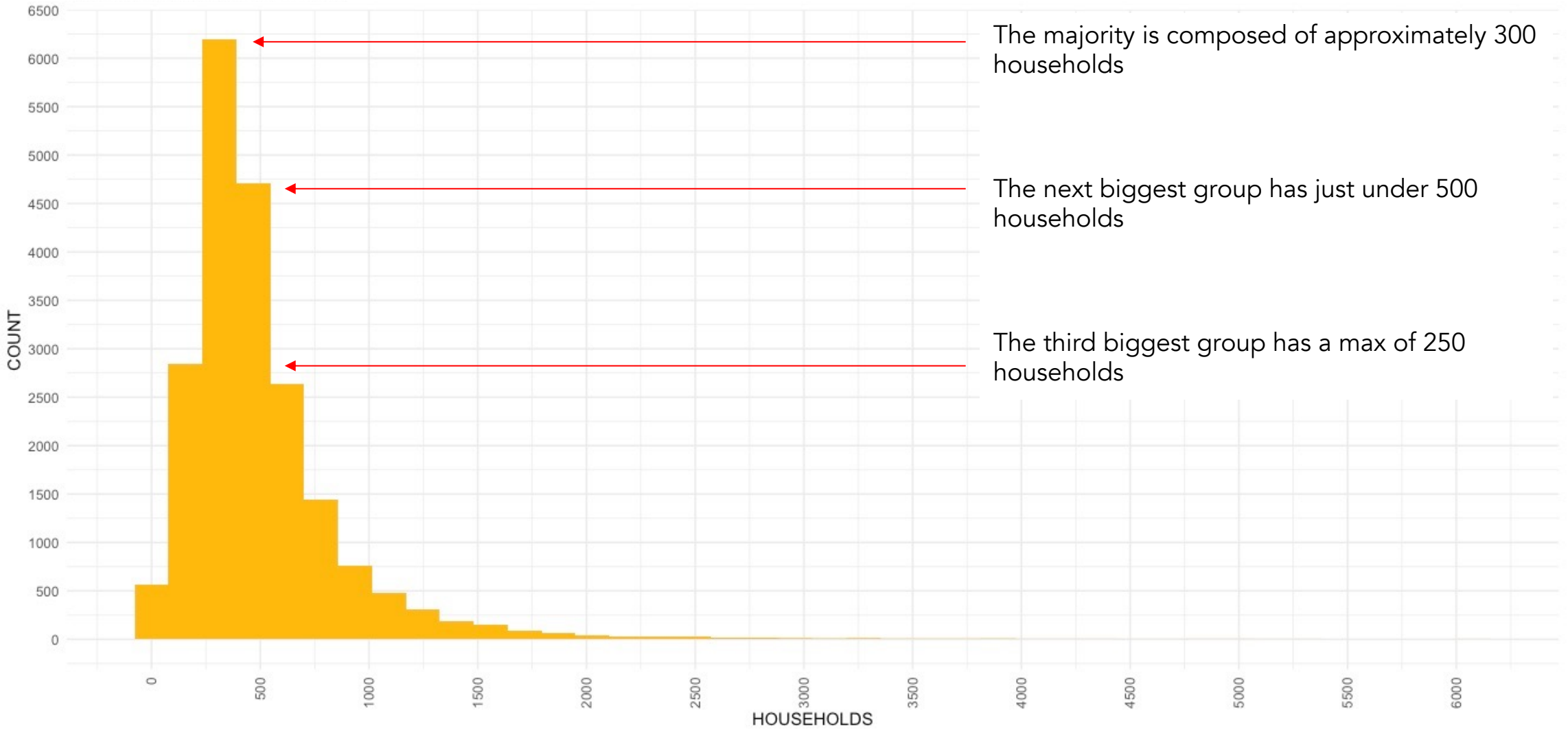
The majority of houses has a population of 1000

The next biggest group of houses has a population of 2000

The third biggest group has a max population of 500

NOTE:
These values don't represent individual houses, but rather groups of houses in close proximity to each other.

HISTOGRAM: HOUSEHOLDS



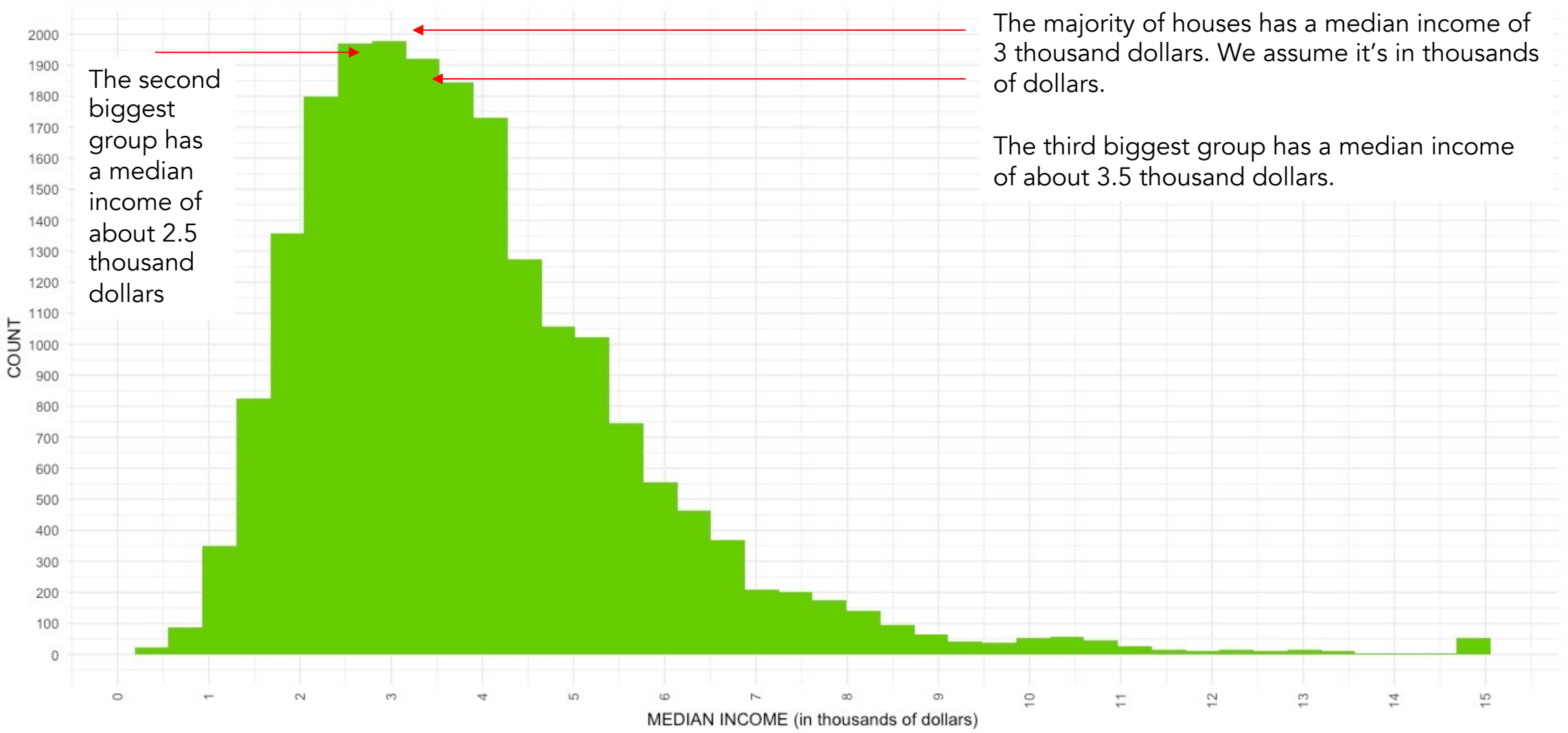
The majority is composed of approximately 300 households

The next biggest group has just under 500 households

The third biggest group has a max of 250 households

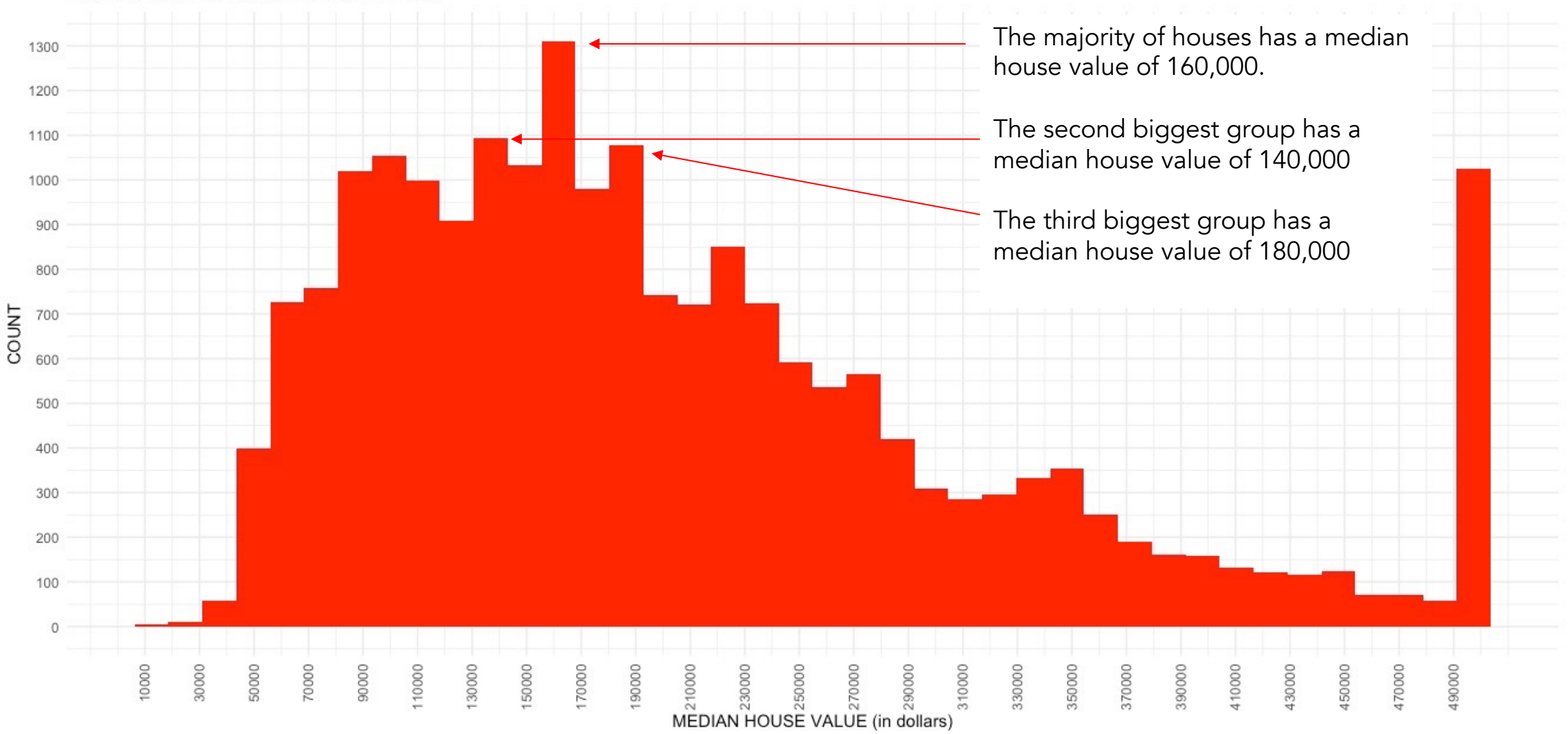
NOTE:
These values don't represent individual houses, but rather groups of houses in close proximity to each other.

HISTOGRAM: MEDIAN INCOME



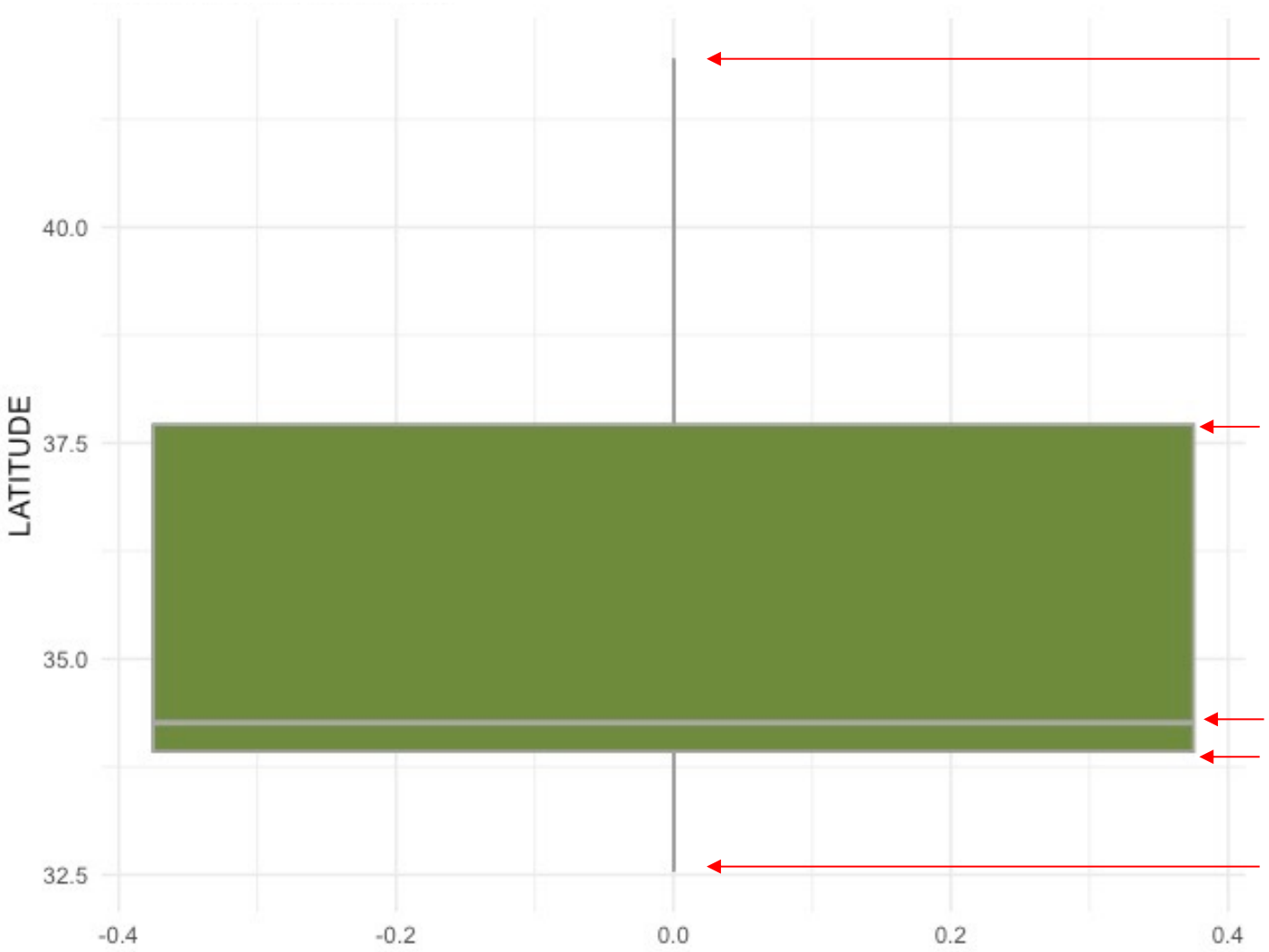
NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

HISTOGRAM: MEDIAN HOUSE VALUE, THE TARGET OR RESPONSE VARIABLE



NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: LATITUDE



Upper whisker: This marks the highest latitude in the dataset.

Upper quartile: 75% of the latitude values is below 37.75

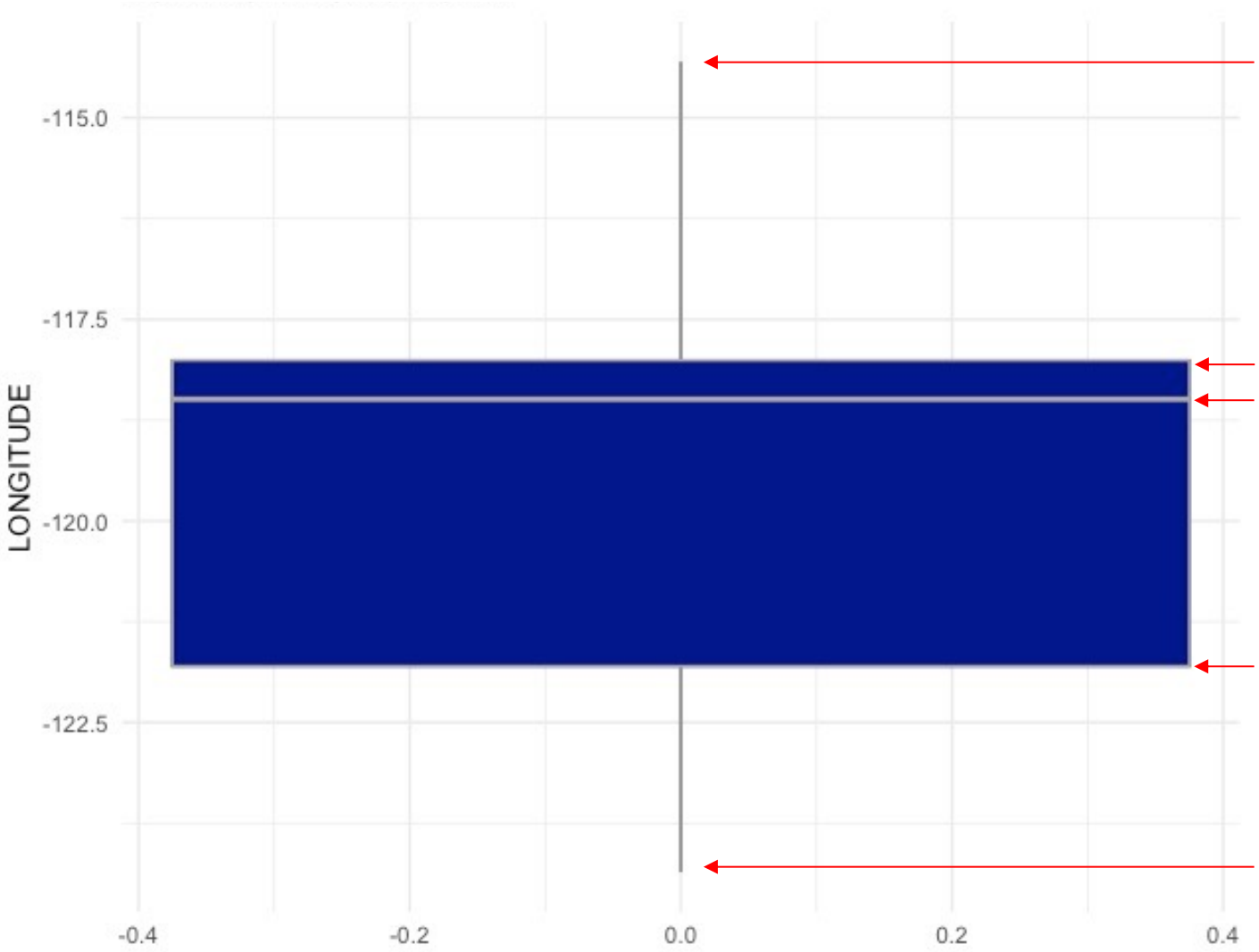
The **median** latitude is about 34. Half the latitudes in the data set is below this median, while the other half is above it.
The median is not in the middle of the boxplot, because it's positively skewed: lower latitudes outnumber the higher ones.

Lower quartile: 25% of the data points for latitude is below this value.

Lower whisker: This marks the lowest latitude in the dataset.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: LONGITUDE



Upper whisker: This marks the highest longitude in the dataset.

Upper quartile: 75% of the longitude values is below -118.

The **median** longitude is about -118.75 . Half the longitudes are below the median, while the other half is above it.

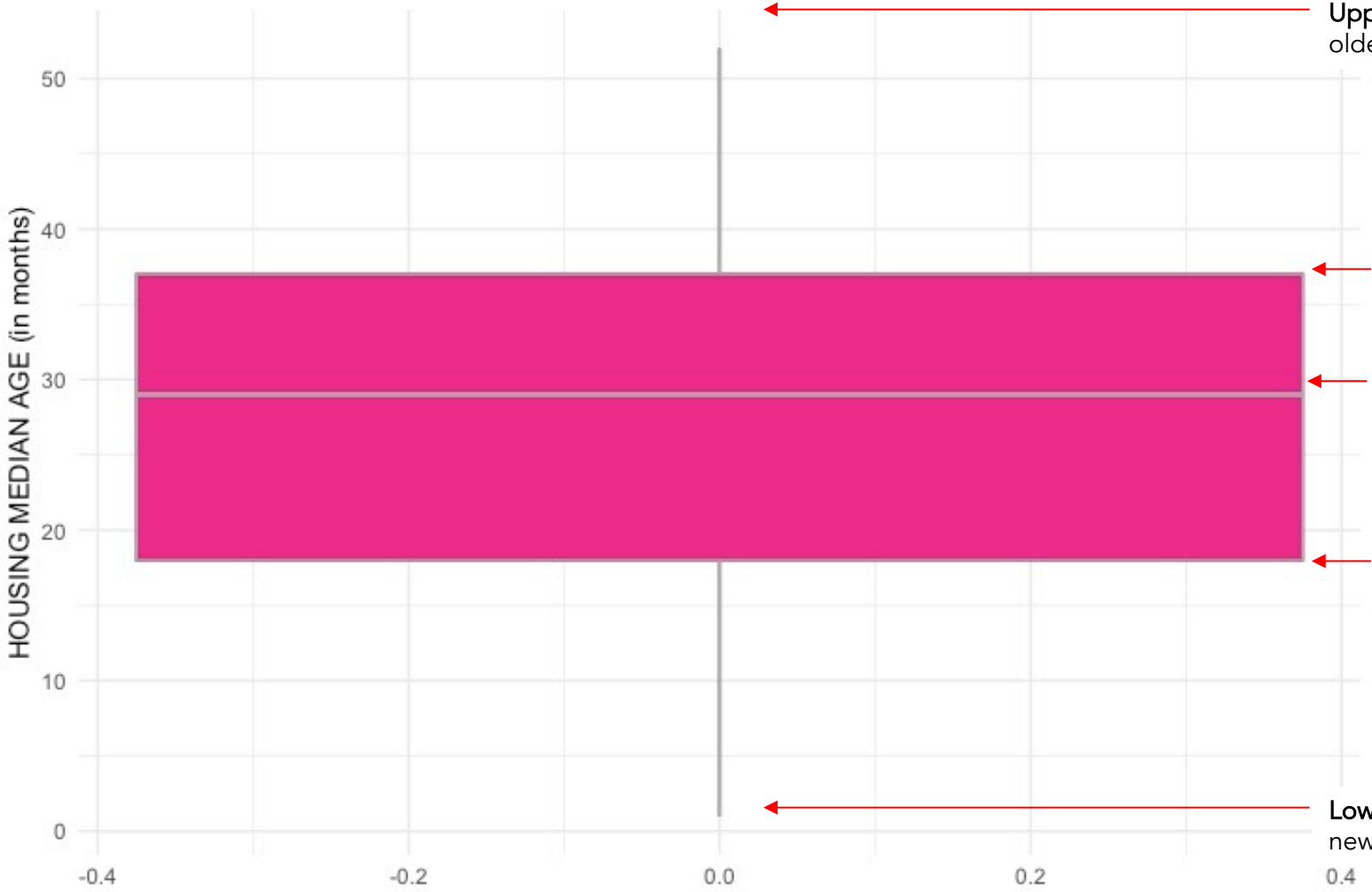
The median is not in the middle of the boxplot, however, which means it's negatively skewed: there are more higher longitudes than lower ones.

Lower quartile: 25% of the longitudes is below this value.

Lower whisker: This marks the lowest longitude in the dataset.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: HOUSING MEDIAN AGE



Upper whisker: This marks the highest median age, i.e. the oldest of the housing groups.

Upper quartile: 75% of the median ages is below 37 (we assume that's in months).

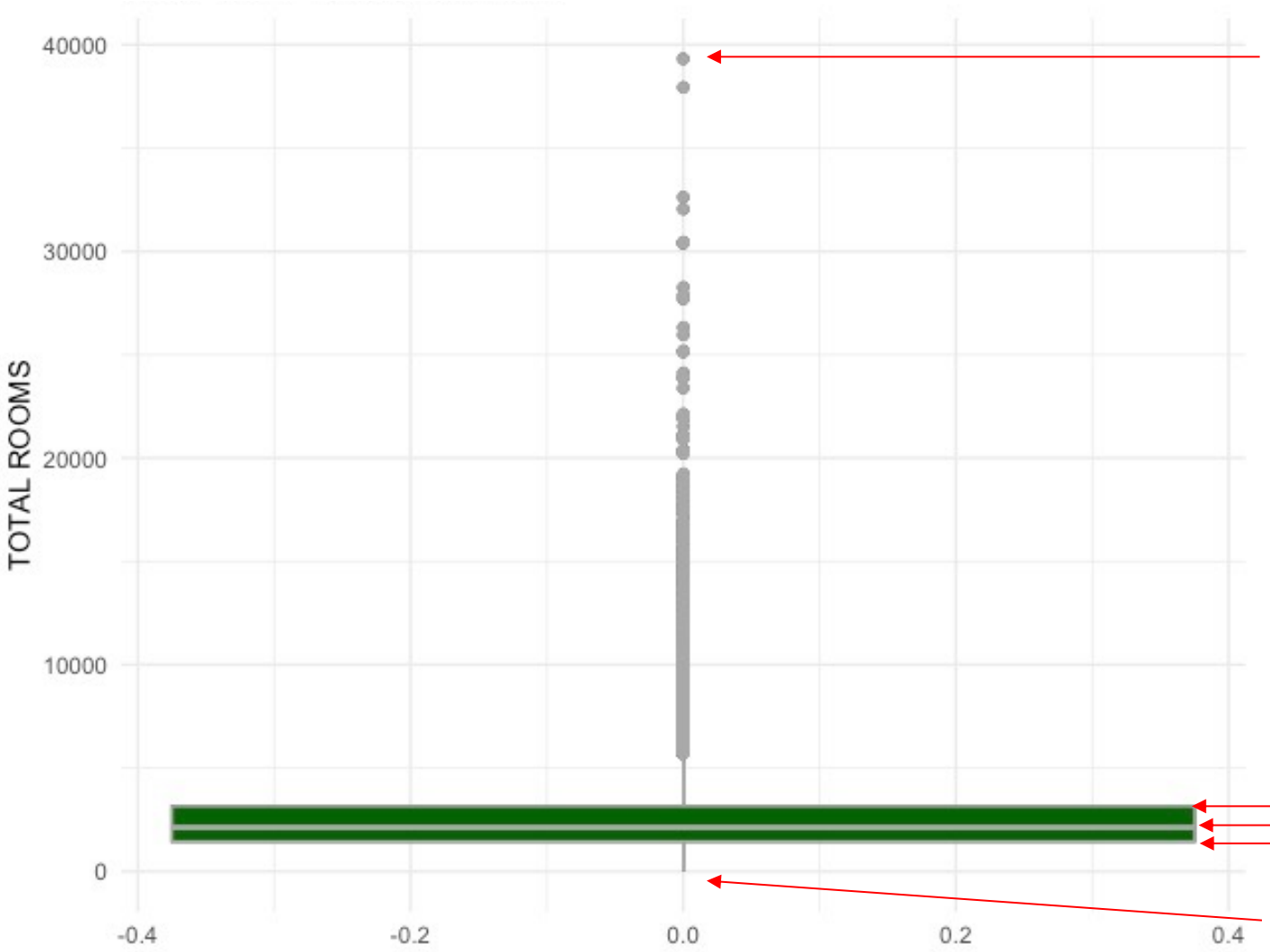
The **median** of all median ages. Half the median ages is below this line, while the other half is above it.
The median is not in the middle of the boxplot, however, which means it's negatively skewed: there are more higher median ages than lower ones. Simply put, there are a lot more older houses than new ones.

Lower quartile: 25% of the housing median age values is below this line.

Lower whisker: This marks the lowest median age, i.e. the newest housing group, approximately 1 month old.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: TOTAL ROOMS



Outliers: All the dots represent outliers, observations/data points with total rooms that are significantly higher than 75% of the dataset.

Upper quartile: 75% of the total rooms is below this value.

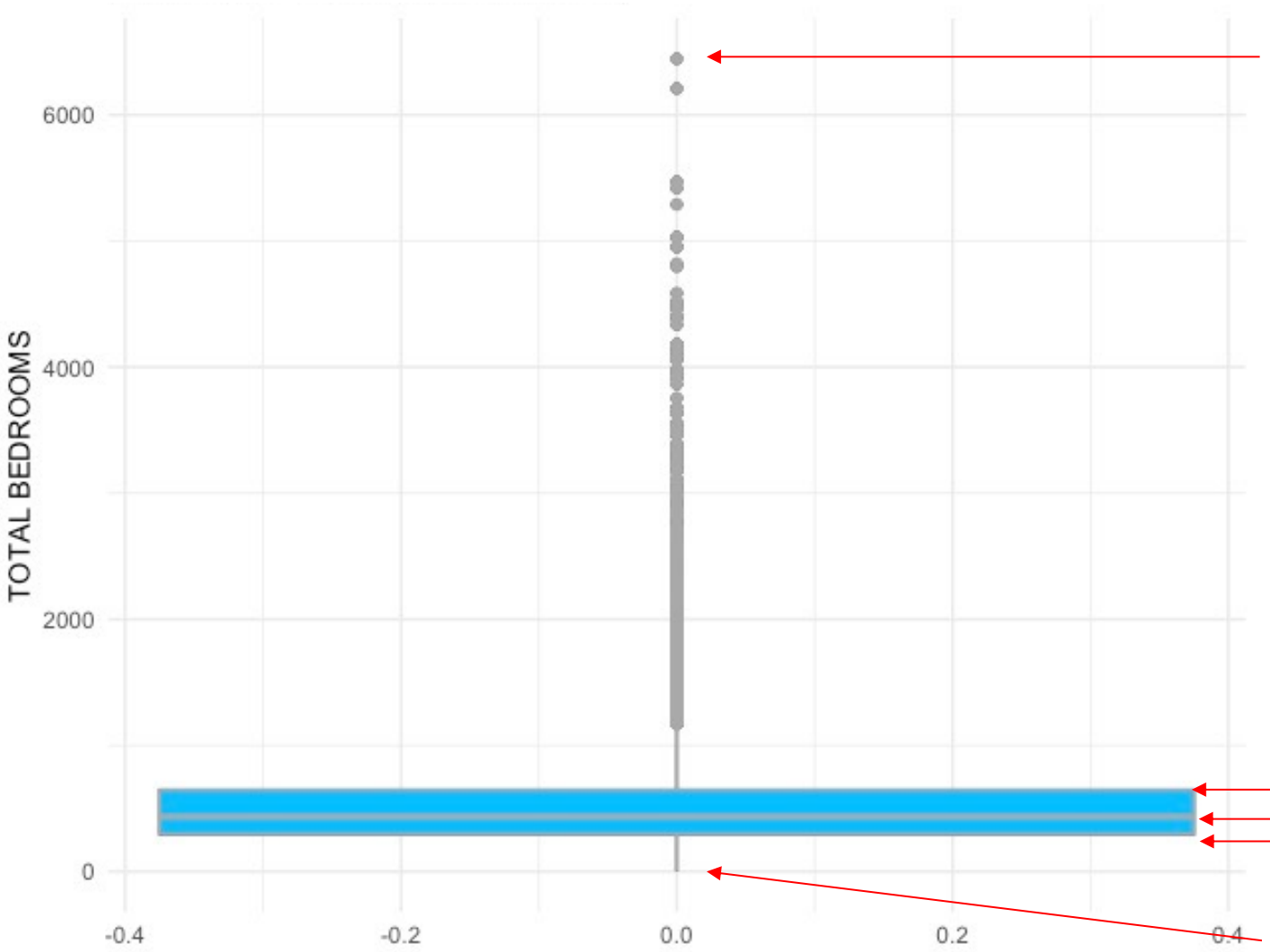
The **median** of the total rooms. Half the total rooms is below this line, while the other half is above it.
The median appears in the middle of the boxplot, which means the number of total rooms is more or less evenly distributed.

Lower quartile: 25% of the total rooms is below this value

Lower whisker: The lowest number of total rooms. Zero rooms may indicate a warehouse or studio.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: TOTAL BEDROOMS



Outliers: All the dots represent outliers, observations/data points with total bedrooms that are significantly higher than 75% of the dataset.

Upper quartile: 75% of the total bedrooms are below this value.

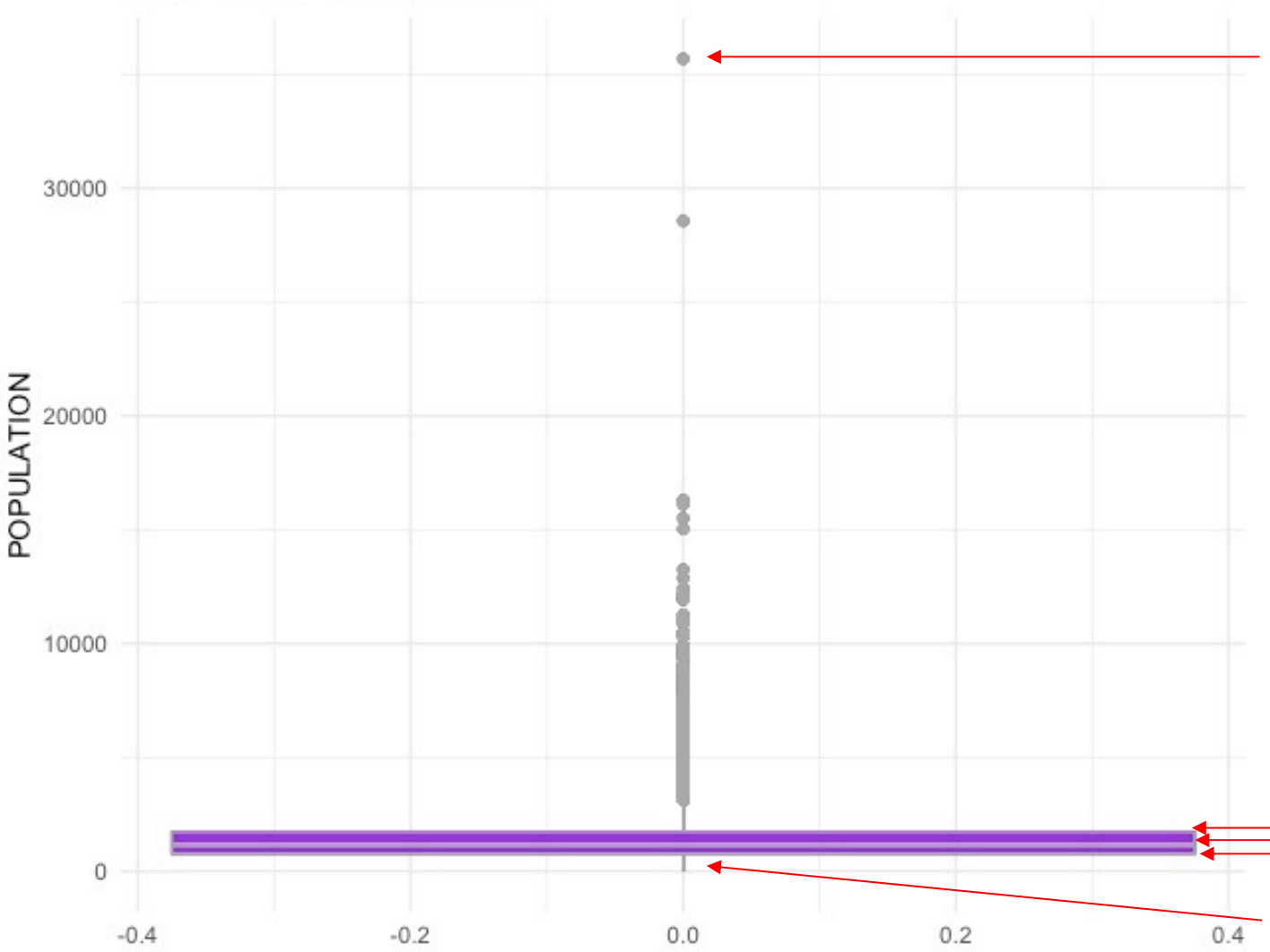
The **median** of the total bedrooms. Half the total bedrooms is below this line, while the other half is above it.
The median appears in the middle of the boxplot, which means the number of total bedrooms is more or less evenly distributed.

Lower quartile: 25% of the total bedrooms is below this value

Lower whisker: The lowest number of total bedrooms. Zero bedrooms may indicate a studio, warehouse, office or any type of commercia/non-residential space.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: POPULATION



Outliers: All the dots represent outliers, observations/data points with a population that are significantly higher than 75% of the dataset.

Upper quartile: 75% of the population is below this line.

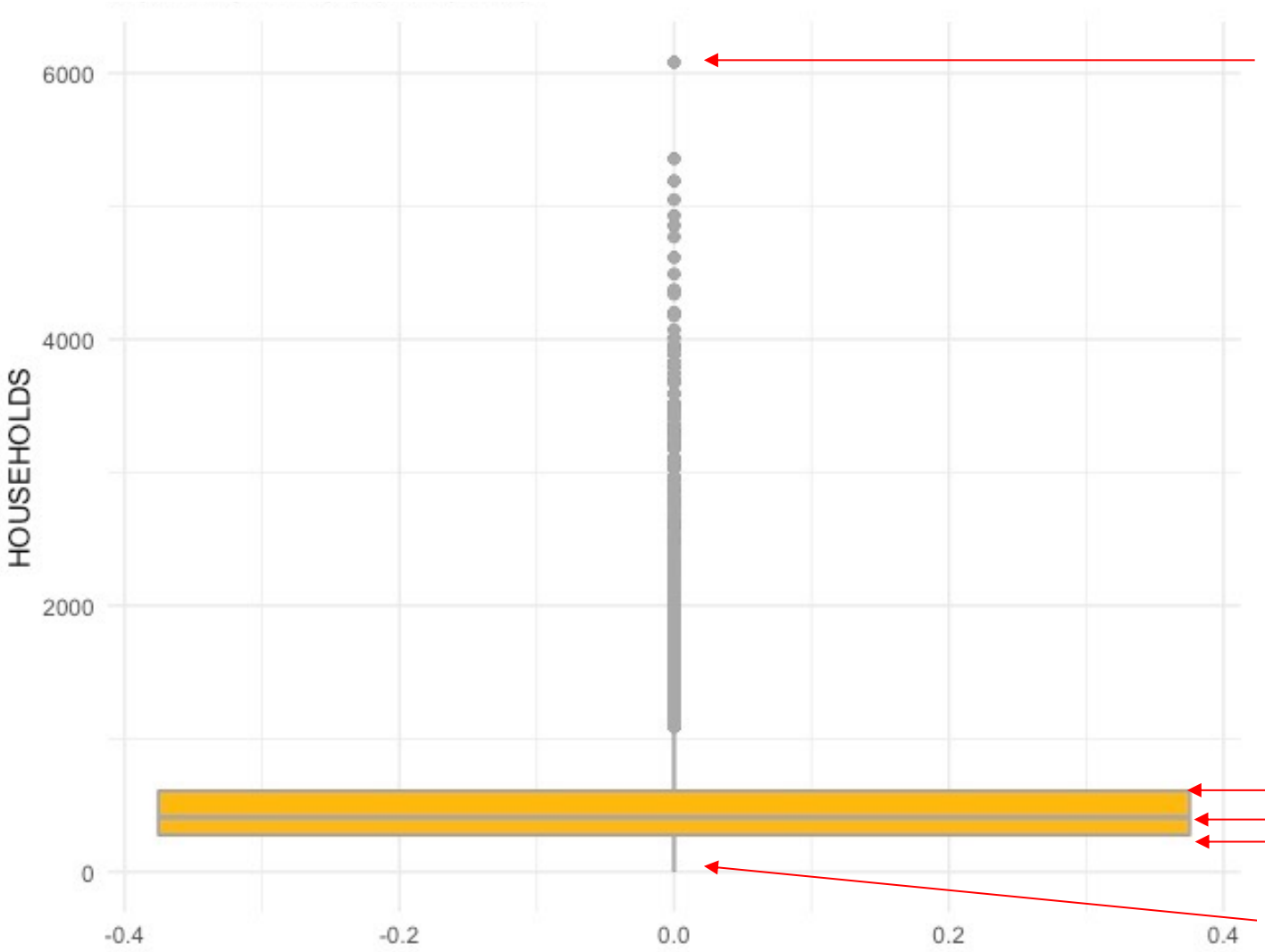
The **median** of population. Half population is below this line, while the other half is above it.
The median appears in the middle of the boxplot, which means the population is more or less evenly distributed.

Lower quartile: 25% of the population is below this value

Lower whisker: The lowest population. Zero population may indicate a new development, i.e. no residents yet.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: HOUSEHOLDS



Outliers: All the dots represent outliers, observations/data points with households that are significantly higher than 75% of the dataset.

Upper quartile: 75% of the households is below this value.

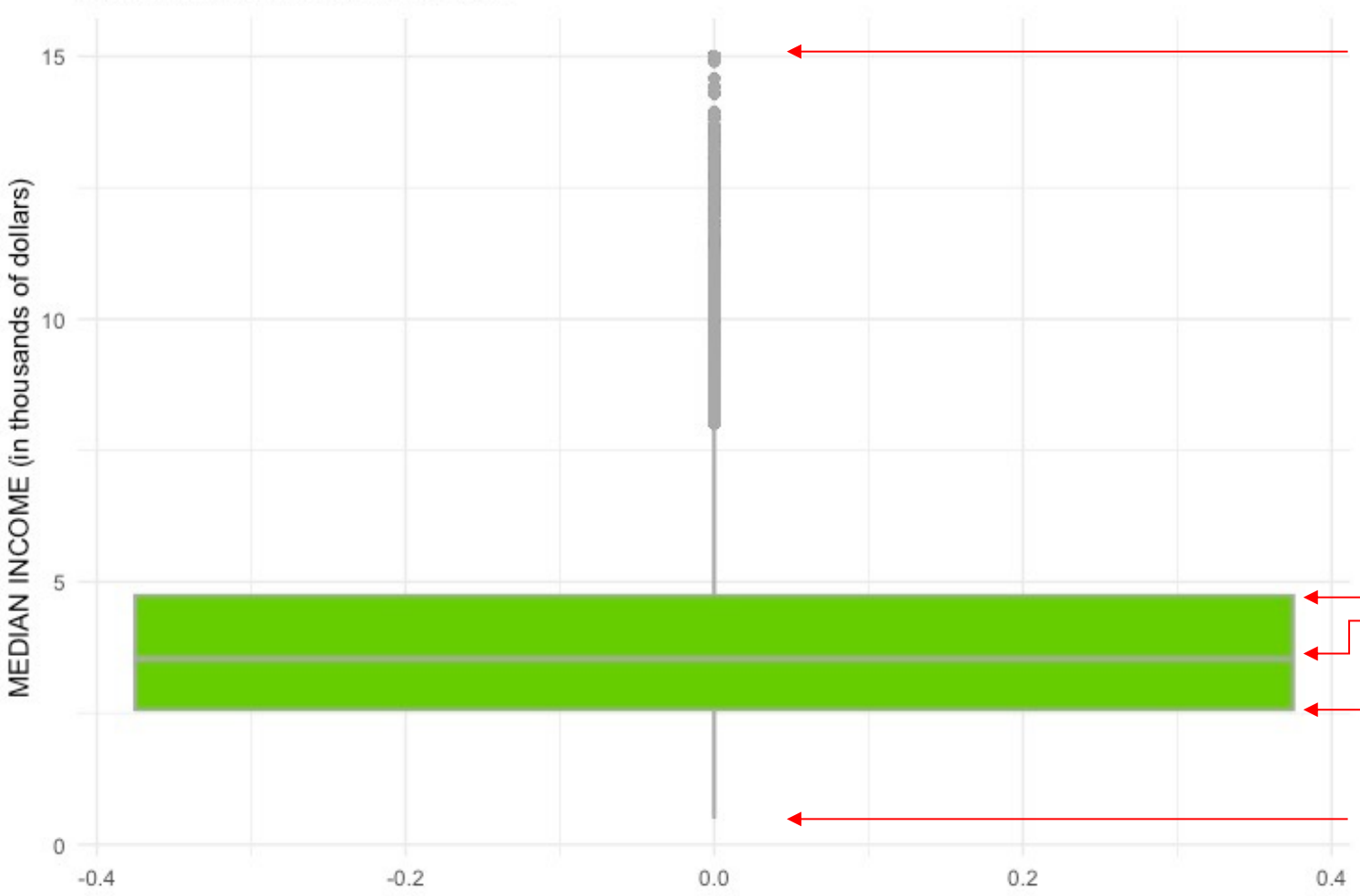
The **median** of household values. Half the households is below this line, while the other half is above it.
The median appears in the middle of the boxplot, which means the number of households is more or less evenly distributed.

Lower quartile: 25% of the households is below this value

Lower whisker: The lowest number of households. Zero households may indicate no residents yet; it's a new development.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: MEDIAN INCOME



Outliers: All the dots represent outliers, observations/data points with median incomes that are significantly higher than 75% of the dataset.

Upper quartile: 75% of the median incomes is below 4.8 (we assume that's in thousands of dollars per month)

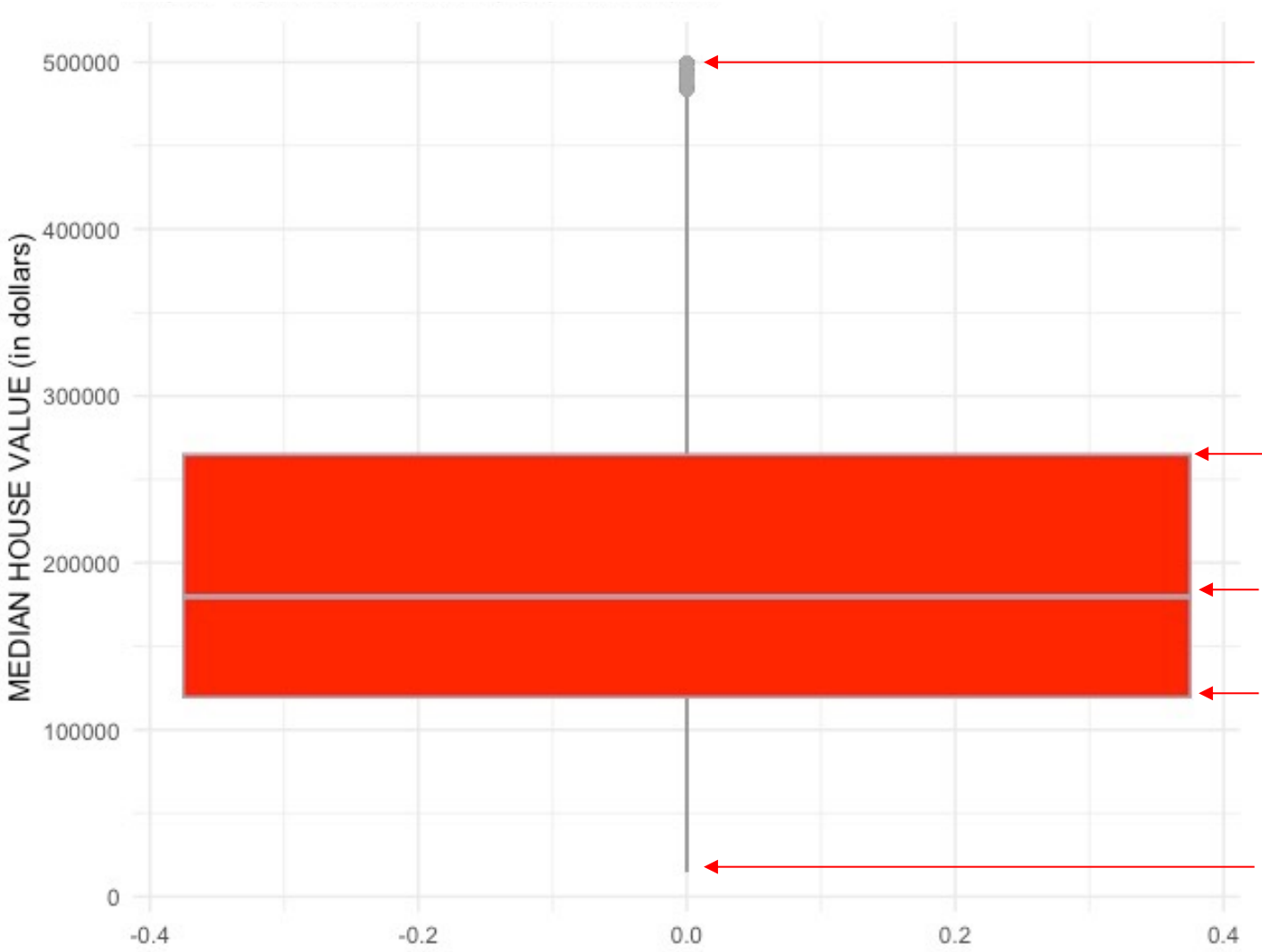
The **median** of median incomes. Half the median incomes is below this line, while the other half is above it.
The median line is in the middle of the boxplot, which means median income is more or less evenly or symmetrically distributed.

Lower quartile: 25% of the median incomes is below this value.

Lower whisker: This marks the lowest median income, approximately 900 or 1000

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: MEDIAN HOUSE VALUE



Outliers: All the dots represent outliers, observations/data points with median house values that are significantly higher than 75% of the dataset.

Upper quartile:
75% of median house values is below 250K.

The **median** of the boxplot is about 180K . Half the median house values are below this value, while the other half is above it.

The median is not in the middle of the boxplot; it's positively skewed: there are lower median house values than high ones.

Lower quartile: 25% of median house values is below this line.

Lower whisker: This marks the lowest median house value, at around 12K.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: HOUSING MEDIAN AGE BY OCEAN PROXIMITY

The median ages of **<1H OCEAN and INLAND** houses are somewhat evenly spread out because their boxplot medians are close to the middle, if not exactly in the middle, of the boxplot. Both groups have upper whiskers, indicating houses with the highest median age of 50 (we assume that's in months), compared to the majority which only ranges from 20 to 38 months. Their lower whiskers indicate the lowest median age, at less than 1-2 months.

The median ages for **NEAR BAY** houses are also somewhat evenly spread out because the boxplot median is also close to the middle, if not exactly in the middle.
This boxplot only has lower whiskers - indicating houses whose median ages are much, much lower, with the lowest at just 1-2 months. There is no upper whisker: there are no houses whose median age is higher than 52 months.

The ages of houses on the ISLAND and NEAR BAY are in the same range.

For **ISLAND** houses, the median is not easily visible because it overlaps with the upper quartile. This means there are lot of identical high values for median age – indicating many equally old houses. Though the short lower whisker indicates newer houses, whose median age is as low as 28 months. There is no upper whisker: there are no houses whose median age is higher than 52 months.

The median age for the **NEAR OCEAN** group appears evenly spread out/symmetrically distributed, because the boxplot median is right in the middle.
Upper whisker indicates houses that are much, much older, while the lower whisker indicates much, much newer houses.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

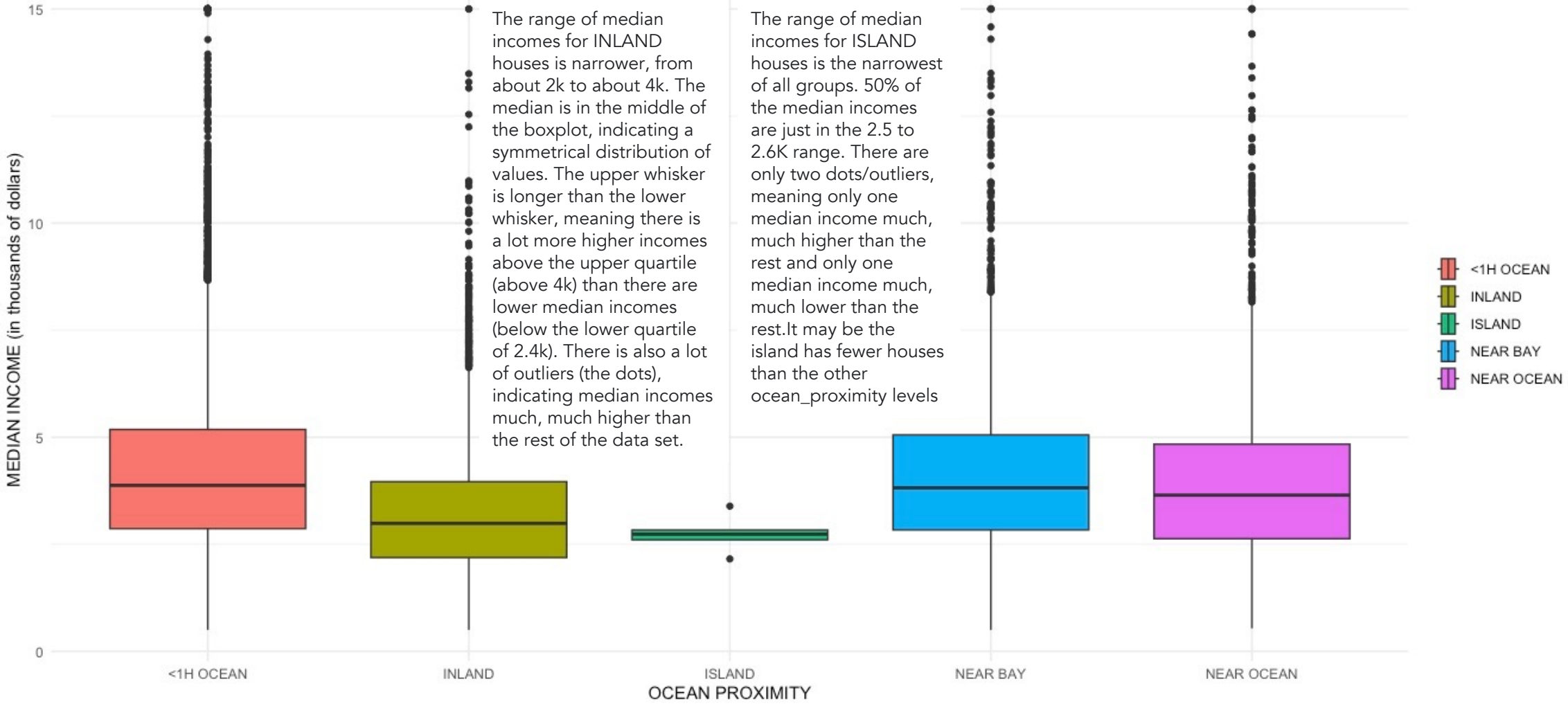
BOXPLOT: MEDIAN INCOME BY OCEAN PROXIMITY

The range of median incomes for <1H OCEAN is pretty wide. 50% of the median incomes are in the 3 to just above 5 (we assume that's in thousands per month) range, with the median at about 4. However, lower whiskers indicate much lower median incomes at under 1, while upper whiskers and outliers (the dots) indicate much, much higher median incomes (up to 15k).

The boxplots for NEAR BAY and NEAR OCEAN are almost identical: the range of median incomes in these two groups are almost similar. Of course, NEAR BAY does have a slightly longer boxplot, indicating a slightly wider range of incomes than NEAR OCEAN. The median for both boxplots are almost the same, too. Both also have similar outliers -- median incomes that are much, much higher than the rest of the data set. Their lower whiskers also look similar, indicating their lowest median incomes are also somewhat the same.

The range of median incomes for INLAND houses is narrower, from about 2k to about 4k. The median is in the middle of the boxplot, indicating a symmetrical distribution of values. The upper whisker is longer than the lower whisker, meaning there is a lot more higher incomes above the upper quartile (above 4k) than there are lower median incomes (below the lower quartile of 2.4k). There is also a lot of outliers (the dots), indicating median incomes much, much higher than the rest of the data set.

The range of median incomes for ISLAND houses is the narrowest of all groups. 50% of the median incomes are just in the 2.5 to 2.6K range. There are only two dots/outliers, meaning only one median income much, much higher than the rest and only one median income much, much lower than the rest.It may be the island has fewer houses than the other ocean_proximity levels

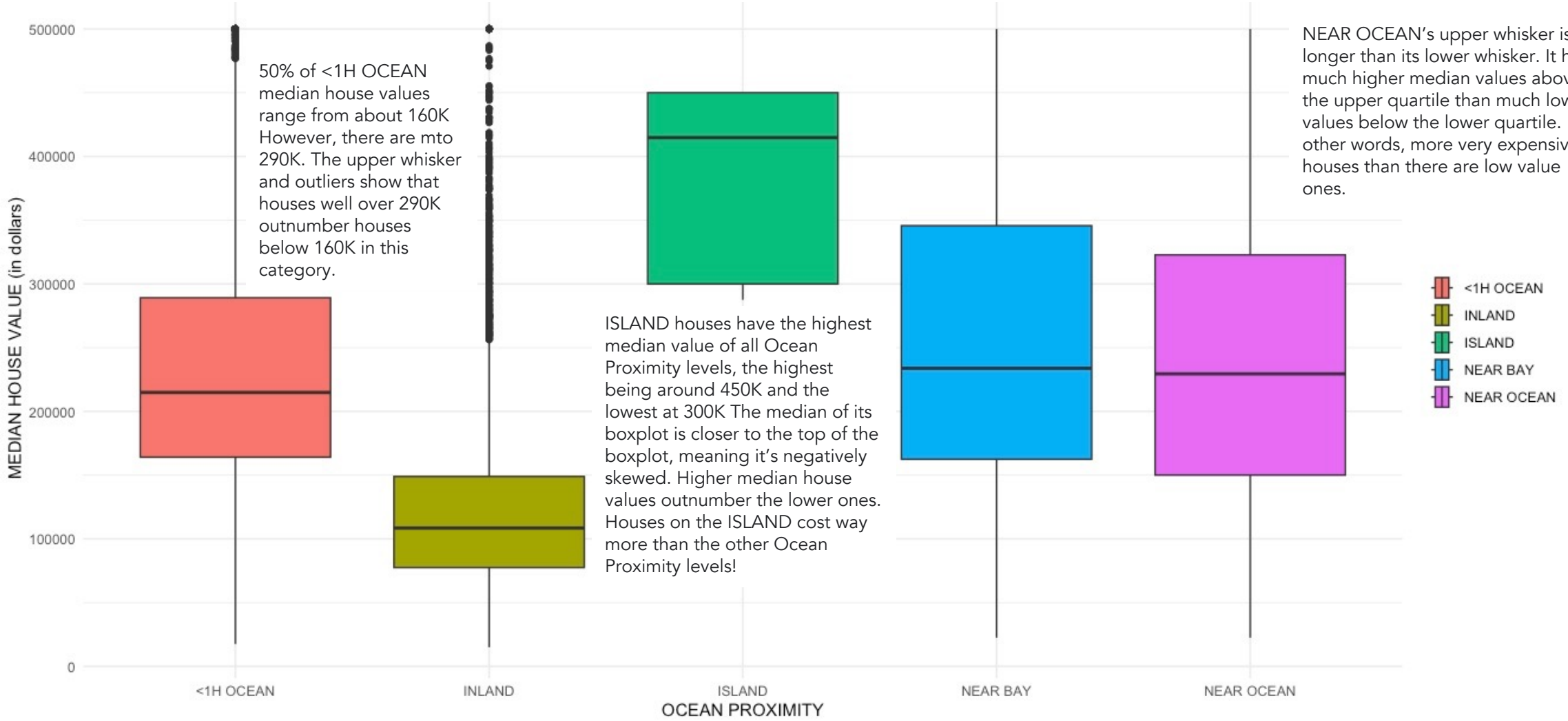


NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

BOXPLOT: MEDIAN HOUSE VALUE BY OCEAN PROXIMITY

50% of median house values INLAND range from just under 80K to about 150K. This would be the most affordable of all the Ocean Proximity levels. However, houses well over 150K (as seen in upper whisker and outliers) outnumber houses below 80K. Best to grab the 80K house as it's quite rare even among INLAND houses.

The median of NEAR BAY and NEAR OCEAN boxplots are the same. This means that 50% of the median house values for both groups is just above 205K (approximately). However, NEAR BAY does have a wider range of median house values than NEAR OCEAN. But with NEAR BAY upper and lower whiskers looking about the same length, this means NEAR BAY has as many much higher median values as there much lower median house values. Its highest median value is 500000, its lowest at about 10000.



50% of <1H OCEAN median house values range from about 160K However, there are mto 290K. The upper whisker and outliers show that houses well over 290K outnumber houses below 160K in this category.

ISLAND houses have the highest median value of all Ocean Proximity levels, the highest being around 450K and the lowest at 300K The median of its boxplot is closer to the top of the boxplot, meaning it's negatively skewed. Higher median house values outnumber the lower ones. Houses on the ISLAND cost way more than the other Ocean Proximity levels!

NEAR OCEAN's upper whisker is longer than its lower whisker. It has much higher median values above the upper quartile than much lower values below the lower quartile. In other words, more very expensive houses than there are low value ones.

NOTE: These values don't represent individual houses, but rather groups of houses in close proximity to each other.

FEATURE SELECTION

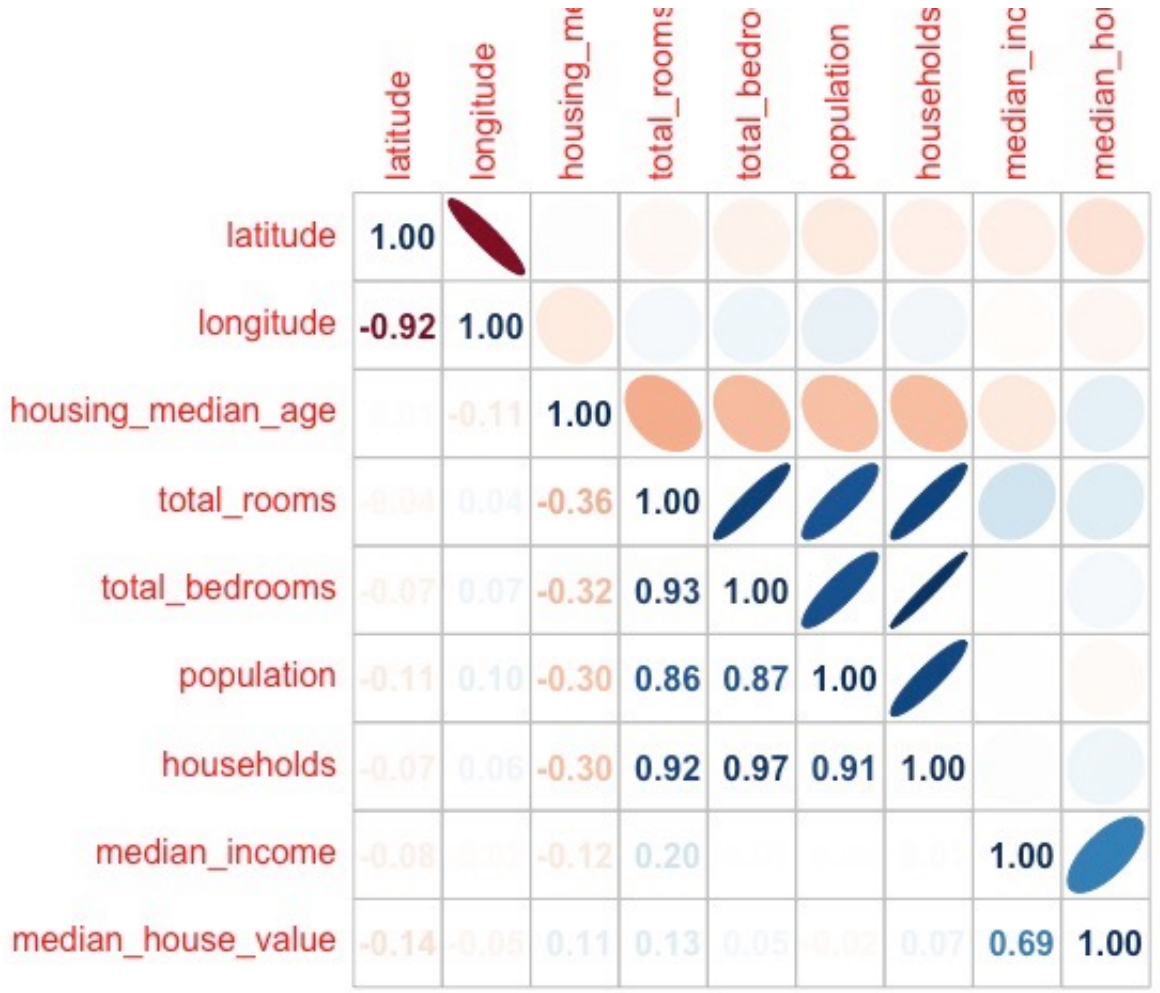
Using correlation
to determine variables for machine learning

Correlation matrix is the most reader-friendly way to demonstrate the correlation

What do the shapes mean?
Narrow and elongated: strong correlations
Rounder, less elongated: weaker associations
Darker colors: stronger correlations vs lighter colors
Positive correlation: slopes from lower left to upper right
Negative if it slopes from the upper left to the lower right.

Therefore:
Very strong negative correlation: latitude to longitude
Very strong positive correlation: total rooms to total bedrooms, population and households
Very strong positive correlation: total bedrooms to total rooms, population and households
Very strong positive correlation: population to total rooms, total bedrooms and households

Strong positive correlation: Median income to median house value
Weak correlation: housing median age to median income
Weak correlation: latitude, longitude, housing median age, total rooms, total bedrooms (even weaker) ,
population (even weaker), households (even weaker) to median house value



FITTING THE MODEL

For this regression, I will use the statistical model random forest. A machine learning algorithm that builds a multitude of decision trees (thus the name forest!), random forest combines the output of these trees into a single result.

As mentioned previously, the target or response variable is median house value.
Below are the predictor variables. Do note that random forest, being random, will not use all variables at the same time. It will instead use one random subset of these variables at a time. This ensures that the decision trees are not correlated.

1. longitude
2. latitude
3. housing_median_age
4. total_rooms
5. total_bedrooms
6. population
7. households
8. median_income
9. ocean_proximity - split into <1H OCEAN, INLAND, ISLAND, NEAR BAY, NEAR OCEAN

PERFORMANCE METRICS FOR THE MODEL

For this regression, the following metrics will be used to determine the performance of the model.

1. **Root Mean Squared Error (RMSE)** is one of the two main performance indicators for a regression model. It measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy).

   The lower the value of the **Root Mean Squared Error**, the better the model is. A perfect model (a hypothetic model that would always predict the exact expected value) would have a **Root Mean Squared Error** value of 0. If you are trying to predict an amount in dollars, then the **Root Mean Squared Error** can be interpreted as the amount of error in dollars.
   Source: https://help.sap.com/docs/SAP_PREDICTIVE_ANALYTICS/41d1a6d4e7574e32b815f1cc87c00f42/5e5198fd4afe4ae5b48fefe0d3161810.html

2. **Rsquared or R^2** tells us **the percentage of variance in the outcome that is explained by the predictor variables** (i.e., the information we do know). A perfect R^2 of 1.00 means that our predictor variables explain 100% of the variance in the outcome we are trying to predict.
   Source: www.causal.app/define/r-squared - :~:text=The closer the r-squared,1 indicates a perfect fit.

   The closer the R^2 value is to 1, **the better the fit**. An R^2 value of 0 indicates that the regression line does not fit the data at all, while an R^2 value of 1 indicates a perfect fit.

   Source: https://www.statisticssolutions.com/r-squared-telling-us-what-we-know-and-what-we-do-not-know/ - :~:text=The R2 tells us,we are trying to predict.

Fitting the model: Random Forest Regression, Round 1

Call:

randomForest(x = train_x, y = train_y, ntree = 500, importance = TRUE, type = regression)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 4

Mean of squared residuals: 2489689423

% Var explained: 81.51

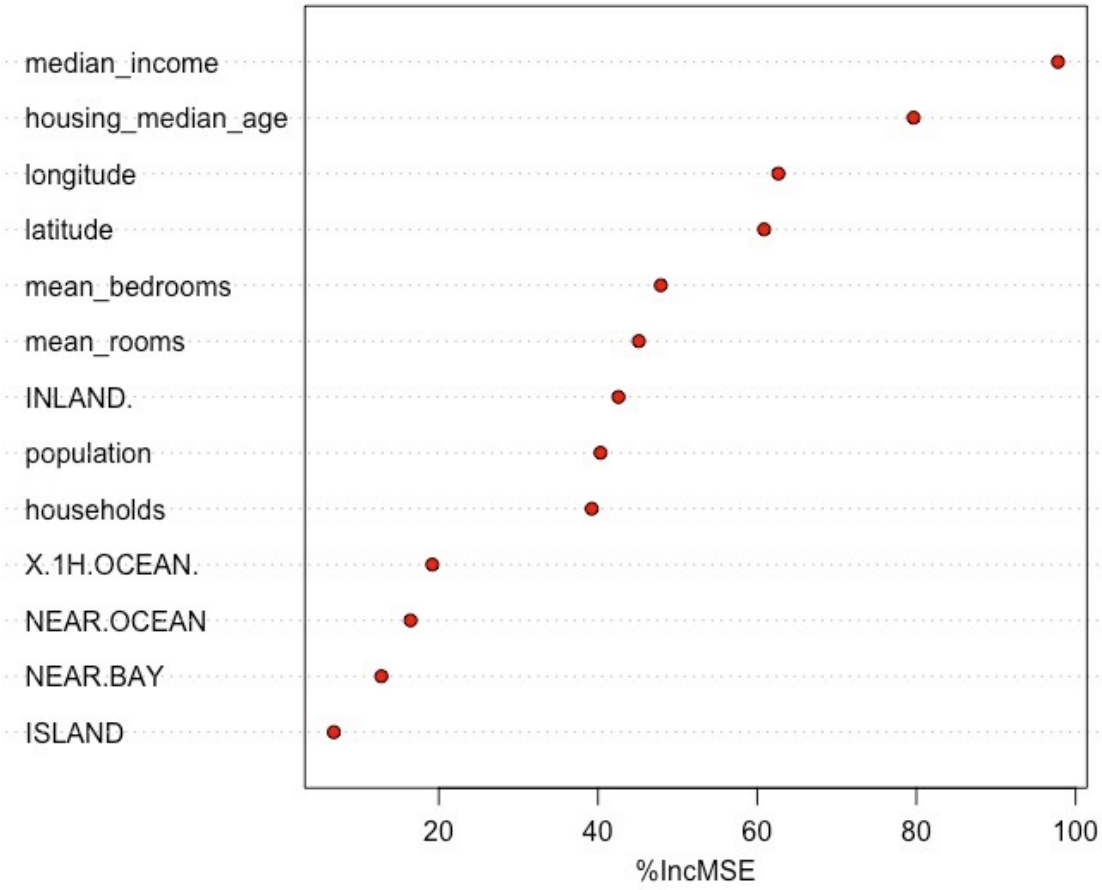
THE METRICS
Train Set RMSE: 49896.79
Test Set RMSE: 48504.14

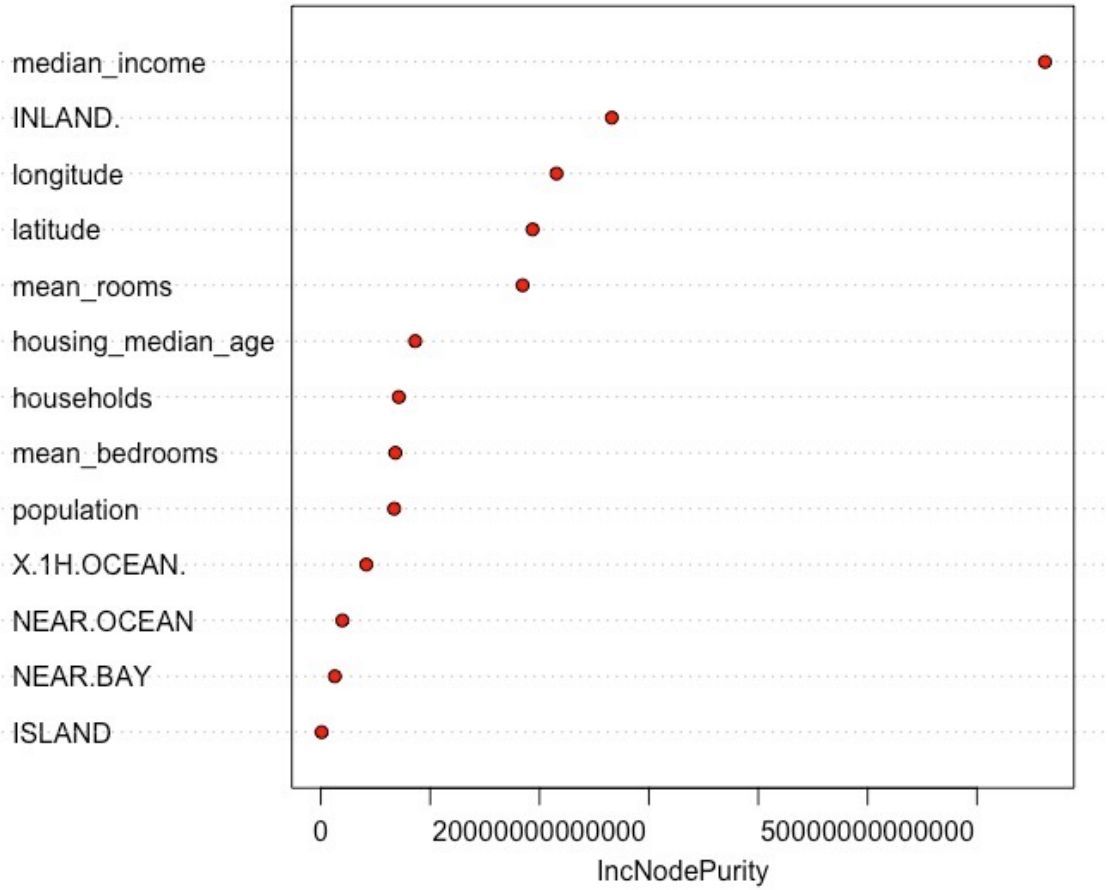
While the RMSE of both train and test sets are high, they are pretty close to each other, give or take a thousand dollars. This means the model doesn't overfit and makes good predictions. Moreover, we can accept such high RMSE, as an RMSE of 48504.14 simply means the prediction is off by $48504.14, an acceptable error vis a vis a house valued at $500,001 (the max median housing value in the dataset)

RSquared: 0.8151249
Closer to 1 than 0, the Rsquared also confirms the model fits the data well.

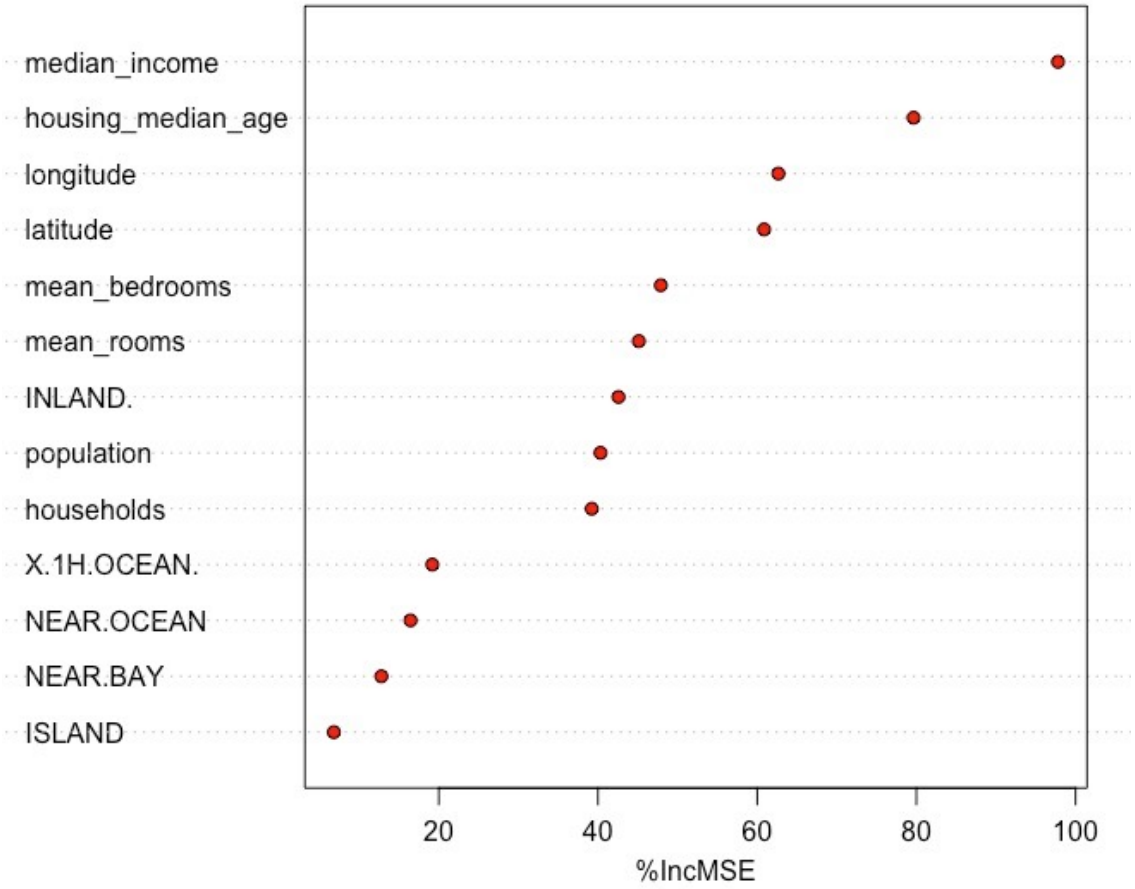
After running random forest, I generated the VARIABLE IMPORTANCE PLOT to determine the most important variables in the regression.

Please see next page for the interpretation





VARIABLE IMPORTANCE PLOT OF RANDOM FOREST REGRESSION



For this report, we will only focus on the leftmost Variable Importance Plot.

What do the plots mean?

- Median income is the most important feature, removing it will mean a 98-99% increase in Mean Squared Error (MSE), which will be the basis of RMSE
- Next is housing median age, removing it will mean about 80% increase in MSE
- Longitude and latitude are 3rd and 4th most important, respectively, removing either will mean about 62-63
- % increase in MSE
- Mean bedrooms and mean rooms are 5th and 6th most important, removing either will mean about 47-48% increase in MSE
- INLAND and population are 7th and 8th most important, removing either will mean about 40-42% increase in MSE
- Householads is 9th most important, removing it will mean about 39% increase in MSE
- The rest are not that important, removing them will mean less than 20% increase in MSE

Fitting the model: Random forest regression, Round 2

This time, without the 4 least important variables, namely: <1H OCEAN ,NEAR OCEAN,NEAR BAY and ISLAND, based on the Variable Importance Plot

Call:

randomForest(x = train_x, y = train_y, ntree = 500, importance = TRUE, type = regression)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 3

Mean of squared residuals: 2527682876

% Var explained: 81.23

THE METRICS

Train Set RMSE 50276.07

Test Set RMSE 48862.85

While the RMSE of both train and test sets are high, they are pretty close to each other, give or take a thousand dollars. This means the model doesn't overfit and makes good predictions. Moreover, we can accept such high RMSE, as an RMSE of 48862.85 simply means the prediction is off by $48862.85, an acceptable error vis a vis a house valued at $500,001 (the max median housing value in the dataset)

RSquared: 0.8123037

Closer to 1 than 0, the Rsquared also confirms the model fits the data well.

Fitting the model: Random forest regression, Round 3

After the 2nd regression, I tuned the model to determine the best m (in this case, 6) or the optimal number of variables to randomly sample. I also determined the optimal number of trees (in this case, 372) that would produce the lowest MSE. These values were used to refine the hyperparameters of this third round, as seen below.

Call:

randomForest(x = train_x, y = train_y, ntree = 372, mtry = best.m, importance = TRUE, type = regression)

Type of random forest: regression

Number of trees: 372

No. of variables tried at each split: 6

Mean of squared residuals: 2463601186

% Var explained: 81.71

THE METRICS

Train Set RMSE 49634.68

Test Set RMSE 48288.36

While the RMSE of both sets are high, they are pretty close to each other, give or take a thousand dollars. This means the model doesn't overfit and makes good predictions. Moreover, we can accept such high RMSE, as an RMSE of 48288.36 simply means the prediction is off by $48288.36, an acceptable error vis a vis a house valued at $500,001 (the max median housing value in the dataset)

RSquared: 0.8170621

Closer to 1 than 0, the Rsquared also confirms the model fits the data well.

Fitting the model: Random forest regression, Round 4

This time, with just the 4 most important variables, namely: latitude, longitude, housing_median_age and median_income, apart from the target variable: median housing value.

Call:

randomForest(x = train_x, y = train_y, ntree = 500, importance = TRUE, type = regression)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 1

Mean of squared residuals: 2676204794

% Var explained: 80.13

THE METRICS

Train Set RMSE 51732.05

Test Set RMSE 49735.01

While the RMSE of both sets are high, they are pretty close to each other, give or take a few thousand dollars. This means the model doesn't overfit and makes good predictions. Moreover, we can accept such high RMSE, as an RMSE of 49735.01 simply means the prediction is off by $49735.01, an acceptable error vis a vis a house valued at $500,001 (the max median housing value in the dataset)

RSquared: 0.8012750

Closer to 1 than 0, the Rsquared also confirms the model fits the data well.

Fitting the model: Random forest regression, Round 5

After tuning the model again with bestm and optimal number of trees.
Using only the 4 most important variables, namely: latitude, longitude, housing_median_age and median_income, apart from the target variable: median housing value.

Call:

randomForest(x = train_x, y = train_y, ntree = 289, mtry = best.m, importance = TRUE, type = regression)

Type of random forest: regression

Number of trees: 289

No. of variables tried at each split: 4

Mean of squared residuals: 2435340839

% Var explained: 81.92

THE METRICS

Train Set RMSE 49349.17
Test Set RMSE 48063.6

While the RMSE of both sets are high, they are pretty close to each other, give or take a few thousand dollars. This means the model doesn't overfit and makes good predictions. Moreover, we can accept such high RMSE, as an RMSE of 49735.01 simply means the prediction is off by $49735.01, an acceptable error vis a vis a house valued at $500,001 (the max median housing value in the dataset)

RSquared: 0.8191606

Closer to 1 than 0, the Rsquared also confirms the model fits the data well.

Summary of Metrics for the different random forest regressions

RANDOM FOREST 1	RANDOM FOREST 2	RANDOM FOREST 3	RANDOM FOREST 4	RANDOM FOREST 5
Train Set RMSE: 49896.79 Test Set RMSE: 48504.14 RSquared 0.8151249	Train Set RMSE 50276.07 Test Set RMSE 48862.85 RSquared 0.8123037	Train Set RMSE 49634.68 Test Set RMSE 48288.36 RSquared 0.8170621	Train Set RMSE 51732.05 Test Set RMSE 49735.01 RSquared 0.8012750	Train Set RMSE 49349.17 Test Set RMSE 48063.6 RSquared 0.8191606

To be sure, I also normalized the RMSE,
using the formula
Normalized RMSE = RMSE / (max value – min value of target variable)

Actual calculation:
NormalizedRMSE <- 49349.17/(500001 - 14999)

Result:
NormalizedRMSE
0.1017504

Based on slide 31,
an RMSE close to 0 indicates the model predicts well.

CONCLUSION:

The variables that drive median housing value are as follows, in order of importance:

- Median income. This may indicate that a more affluent household is able to maintain the house properly, thus raising its value. This may also indicate causation, i.e. the higher the income, the more likely a person is to afford a higher value home.
- Housing median age. The age of a house is also a logical predictor of its value.
- Longitude. This indicates the location of a house, affirming the common belief of “location, location, location”
- Latitude. This also indicates the location of a house, affirming the common belief of “location, location, location”

LEARNINGS/RECOMMENDATIONS:

- Location is not the strongest factor that drives the value of a house, though, of course, it is very much relevant/important.
- Moving forward, the real estate company should also look into the income of the homeowner/potential seller and the home’s age, rather than prioritize selling a home based solely on location.
- For greater accuracy in predictions, it is also recommended that the company source the actual house values, actual incomes, actual house age, rather than run the regression on median values.

THANK YOU!