Carnegie Mellon University

Can median house value be predicted based on population characteristics?

Arnab Dey Sarkar, Emmy Moore, Iana Iacob, Oladayo Oladeji, Tyler Jaffe, Sarah Pitell

Introduction

The value of a house can be defined as the current worth or market value of that household, i.e., the monetary value of a house to prospective buyers. There are certain variables that tend to influence the median household value. This project seeks to determine what, if any, relationships exist between the variables of our dataset and median household values. We will also determine if median household value can be effectively predicted with such variables.

Methodology

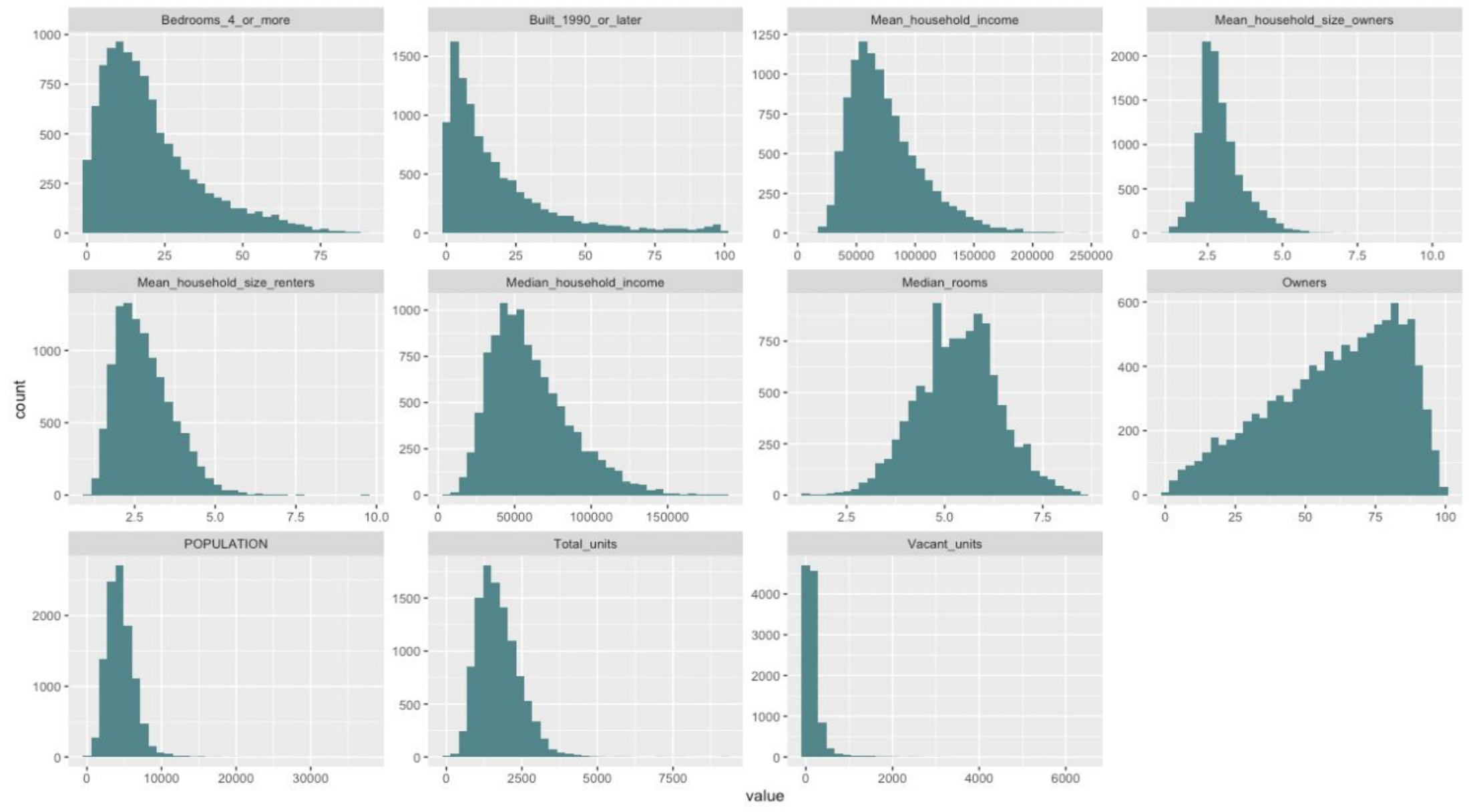
- We utilized two linear regression approaches and two classification techniques through a cross-validation procedure and Akaike Information Criterion for best-subset-selection
  - Analysis methods: Linear Regression, Linear Regression with Best-Subset-Selection, Decision Tree Classification, Random Forest
- The cross validation procedure was a random sampling of the entire data set where 70% of the data was used for training the model and 30% was used for validation of the model
- The quality of the model was determined by the overall mean squared error (MSE) and R-squared value

Data Processing

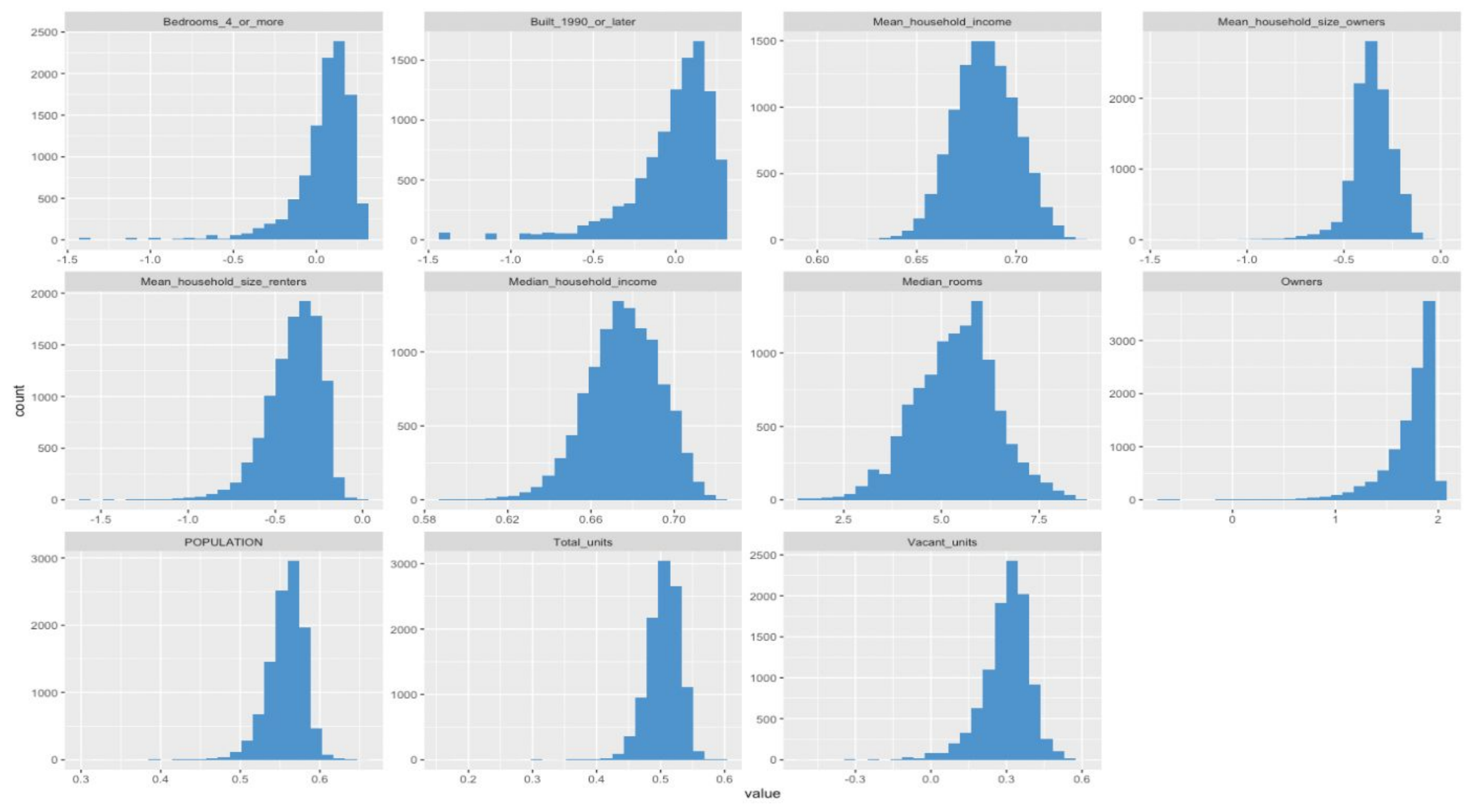
This dataset contains housing market characteristic information in 14 different variables for 10,605 houses. The data includes a response variable, median household value, and 13 predictor variables: location (latitude and longitude markers), population size, total number of housing units, total number of vacant units, median number of rooms per unit, average number of people in rented homes, average number of people in owned homes, percentage of units that are owned (as opposed to rented), median household income, mean household income, percentage of units built after 1989, and percentage of units with four or more bedrooms.

Exploratory Data Analysis

Given the size of the dataset and the goal of finding significant predictors for the response variable, the predictor variables had to be investigated for potential transformation and outliers had to be identified and removed appropriately. The sections below show the original set of response variable data (latitude and longitude were excluded here due to the data type) as histograms.



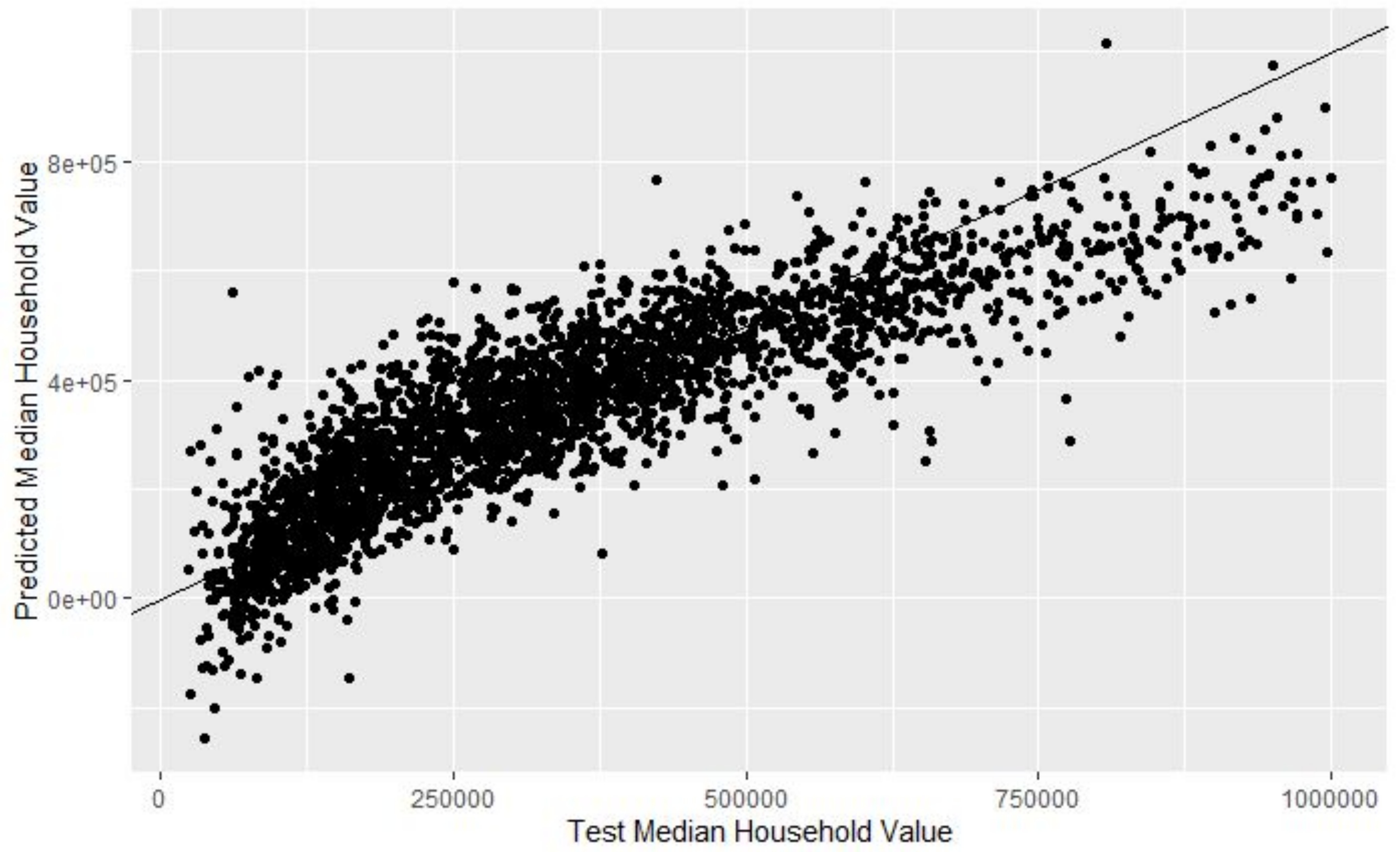
The histograms above show several left- and right-skewed variables. The variables that may benefit from a logarithmic transformation based on these are: `Bedrooms_4_or_more`, `Built_1990_or_later`, `Mean_household_income`, `Mean_household_size_owners`, `Mean_household_size_renters`, `Median_household_income`, `Owners`, `Population`, `Total_units`, and `Vacant_units`.



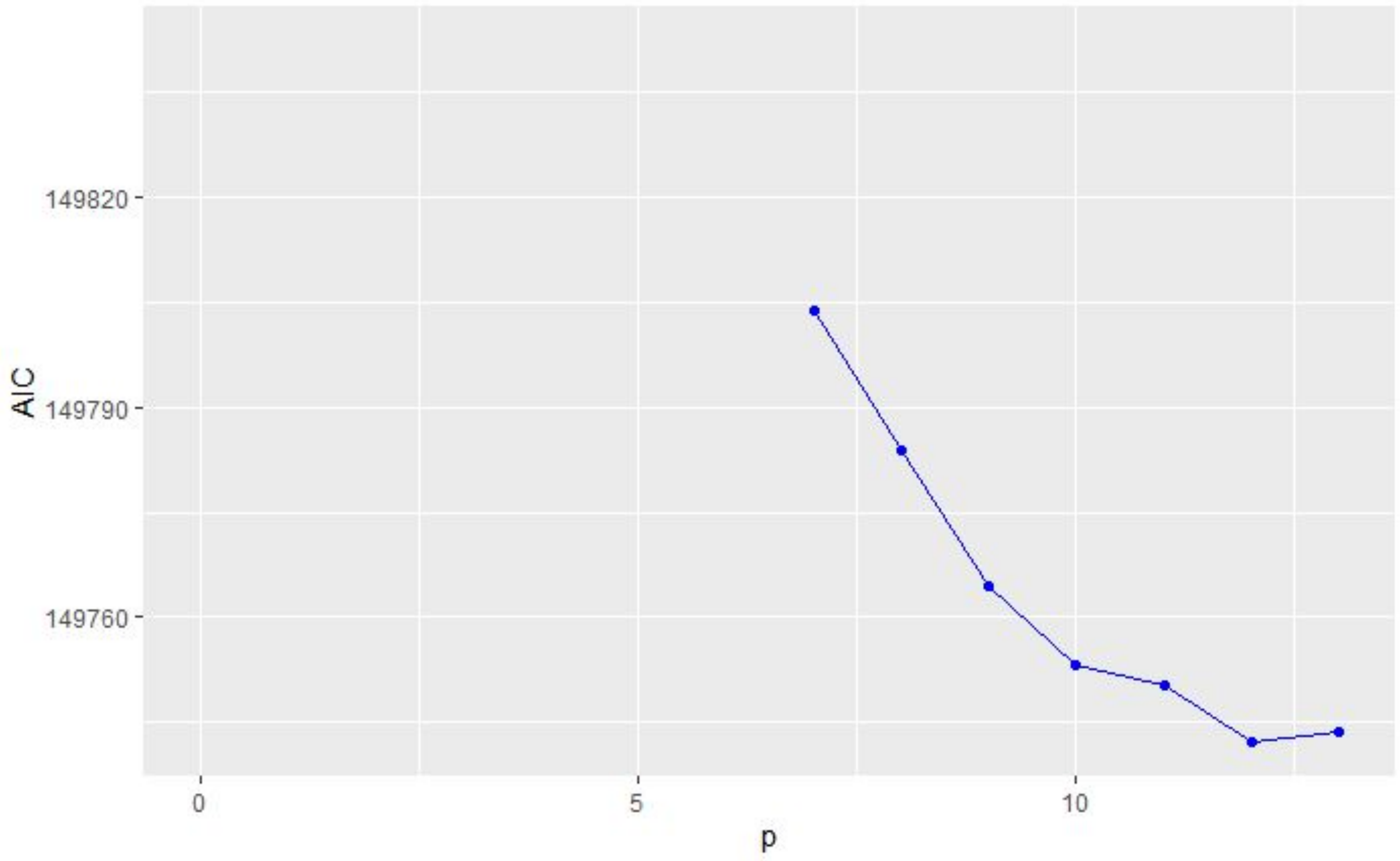
After transforming and removing outliers, the total sample size was reduced to 9,259. Histograms were not created for latitude and longitude, as these are dependent on one another to provide location data.

Analysis

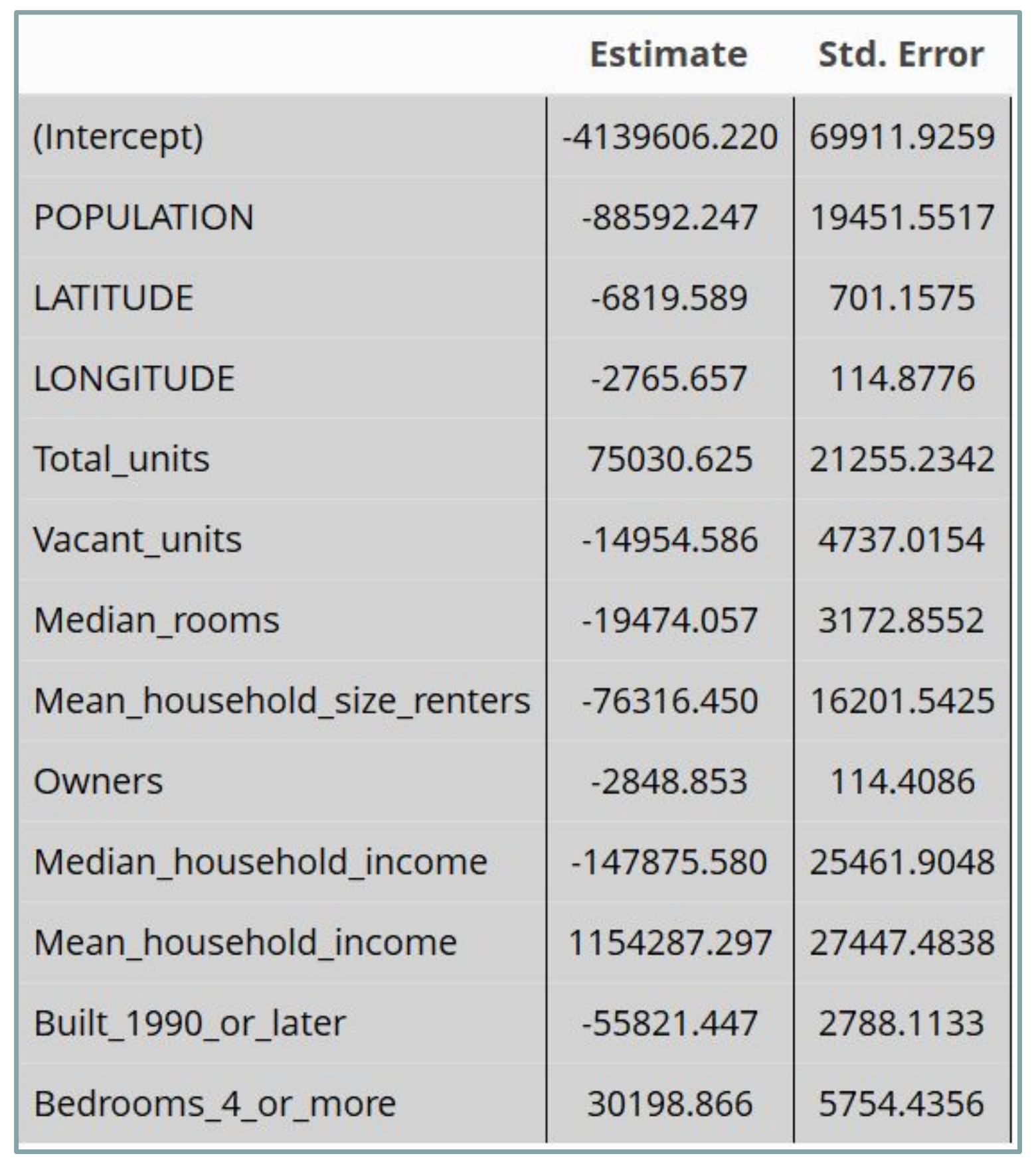
Linear Regression & Best Subset Selection



According to the results, linear regression achieved a mean squared error of $10.5B and an adjusted R-squared value of 0.76. This signifies that the linear regression model is not recommended for prediction of the median household value in this dataset.



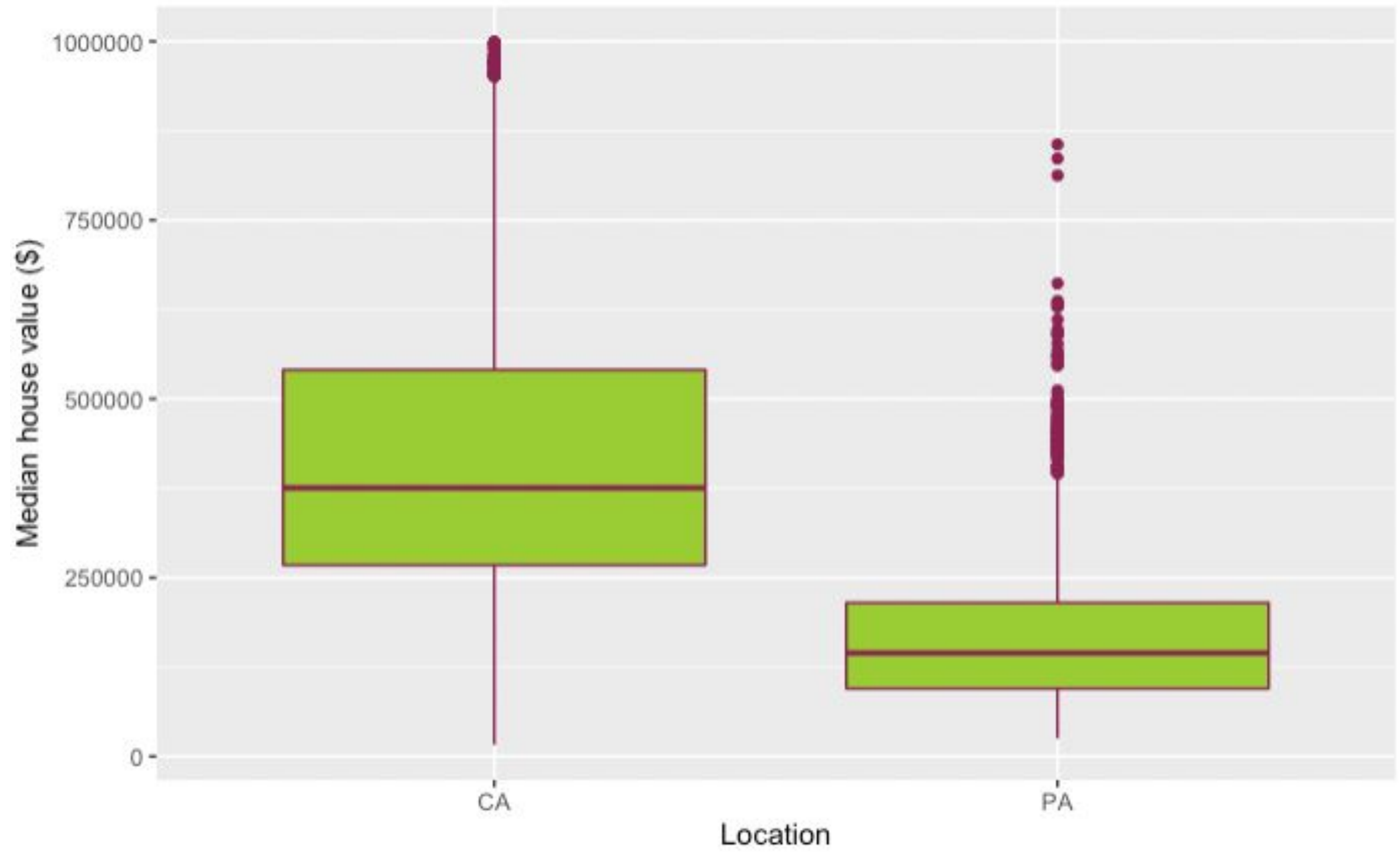
A best-subset-selection analysis was conducted to help improve the linear regression approximation, but achieved the same R-squared value (0.76) by removing `Mean_household_size_owners` as an uninformative predictor.



	Estimate	Std. Error
(Intercept)	-4139606.220	69911.9259
POPULATION	-88592.247	19451.5517
LATITUDE	-6819.589	701.1575
LONGITUDE	-2765.657	114.8776
Total_units	75030.625	21255.2342
Vacant_units	-14954.586	4737.0154
Median_rooms	-19474.057	3172.8552
Mean_household_size_renters	-76316.450	16201.5425
Owners	-2848.853	114.4086
Median_household_income	-147875.580	25461.9048
Mean_household_income	1154287.297	27447.4838
Built_1990_or_later	-55821.447	2788.1133
Bedrooms_4_or_more	30198.866	5754.4356

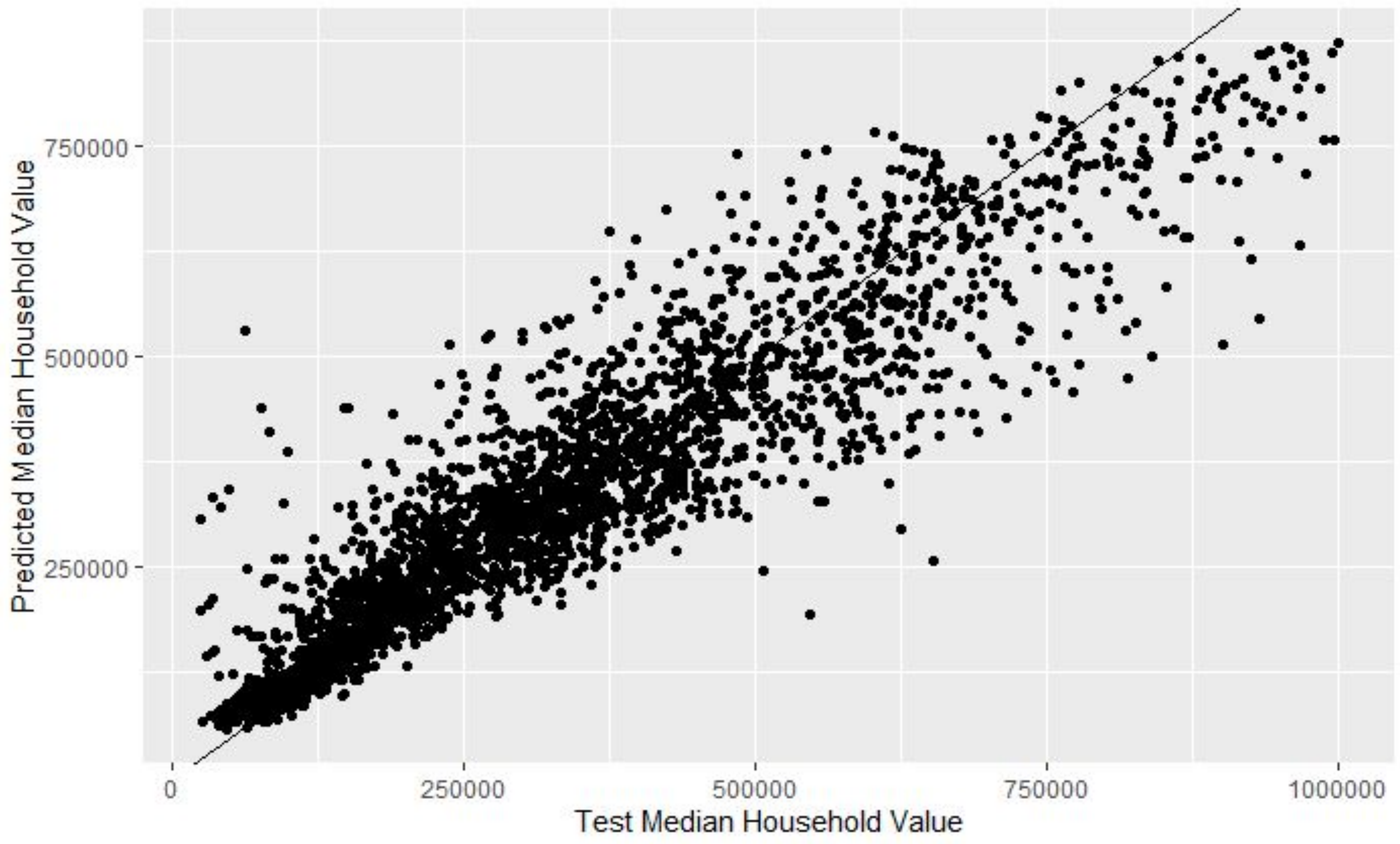
Best-subset-selection analysis coefficient estimates and standard errors for each informative predictor.

Location Data



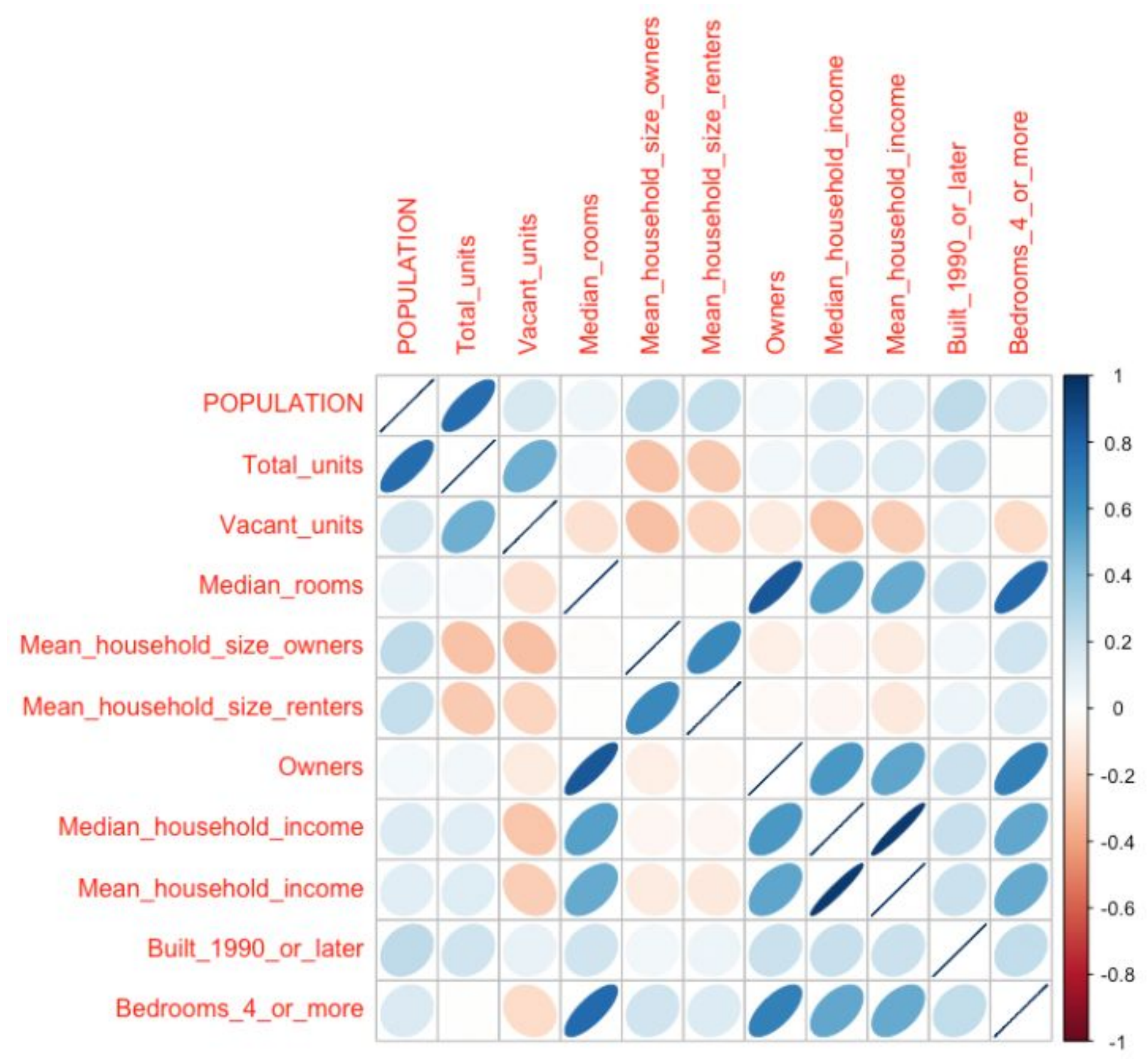
When latitude and longitude data were combined, they showed that this dataset is made up of houses located in either California or Pennsylvania. When comparing on a state-by-state basis, houses in California are generally more expensive than houses in Pennsylvania. There is a large range of median house value for both states, as cost of living is different in different parts of a state as well as different parts of a country.

Random Forest



The random forest model was also used to predict the median house value for the locales. The model generated has an R-squared value of 0.85, which is significantly better than the linear regression model. The variable importance sorting was used to sort out the most important predictor variables. According to this model, the most important predictors are `Mean_household_income`, `Median_household_income`, `Latitude`, and `Longitude`.

Multicollinearity



This dataset exhibits strong multicollinearity between predictor variables. Most correlations are positive, with the exception of `Vacant_units`, which has a slightly negative correlation with most other predictors.

Conclusion

Our analysis shows that people with higher income more commonly afford four or more bedroom apartments. The dataset includes data for two U.S. states, California and Pennsylvania; houses in CA tend to have a higher median value than those in PA. This dataset has high multicollinearity in the predictor variables. The MSE of the random forest model is lower than that of the linear model, which leads us to believe that random forest is the most suitable model for prediction with this dataset. This model also suggests that `Mean_household_income`, `Median_household_income`, `Latitude`, and `Longitude` are the primary predictors for median house value.