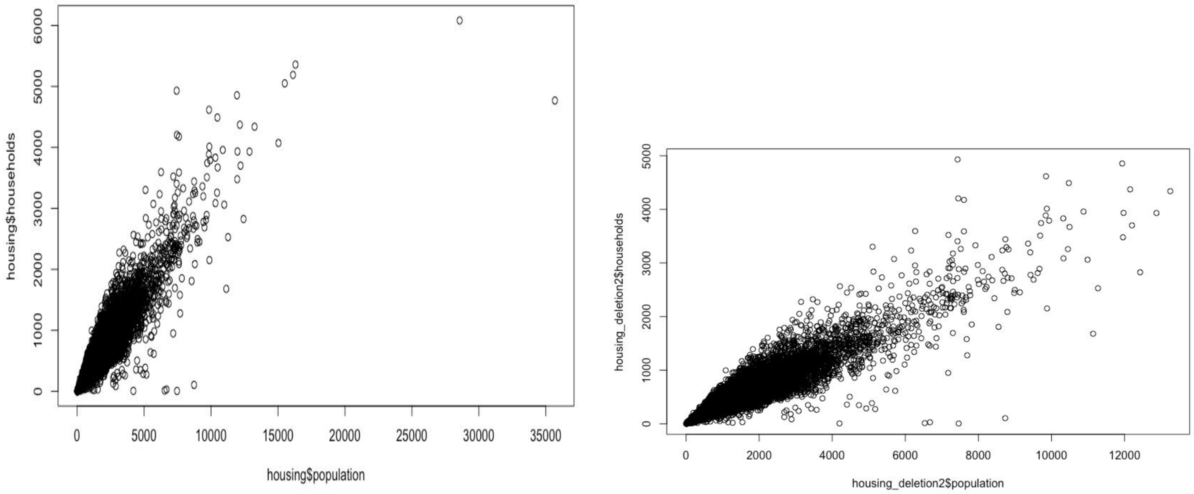
Executive summary

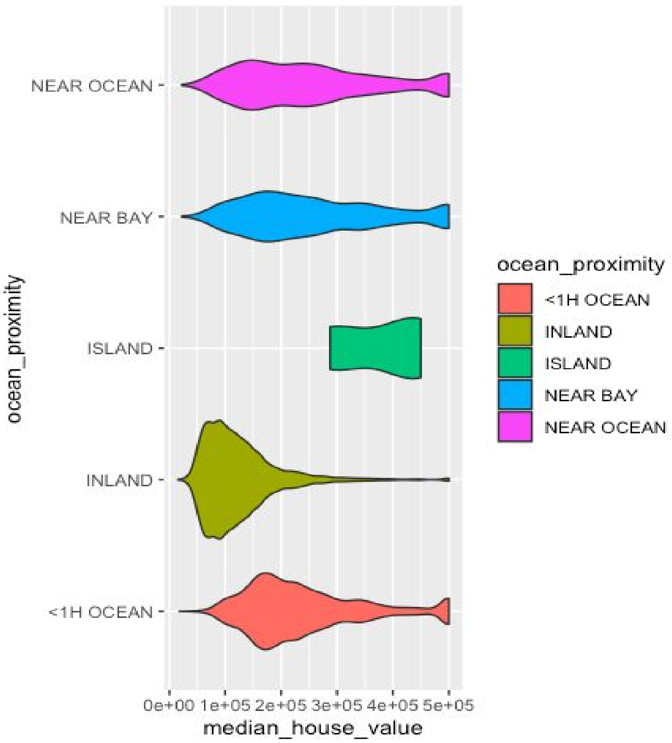
The objective of this report is to describe how various housing factors affect housing prices in California. The data was collected from the 1990 California census report and contains nine numerical variables and one categorical variable. The numerical variables include longitude and latitude values which measure the distance west and north, respectively. The remaining variables are measured as within a block of the longitude and latitude coordinates and those are as follows: housing median age, total rooms, total bedrooms, population, households (defined as a group of people residing within a home unit), median income (measured in tens of thousands of USD), and median house value (measured in USD). The categorical variable of "Ocean Proximity" contains five categories: <1H Ocean, Inland, Near Ocean, Near Bay, Island. The overall structure of the dataset includes 20,640 observations and 10 variables.

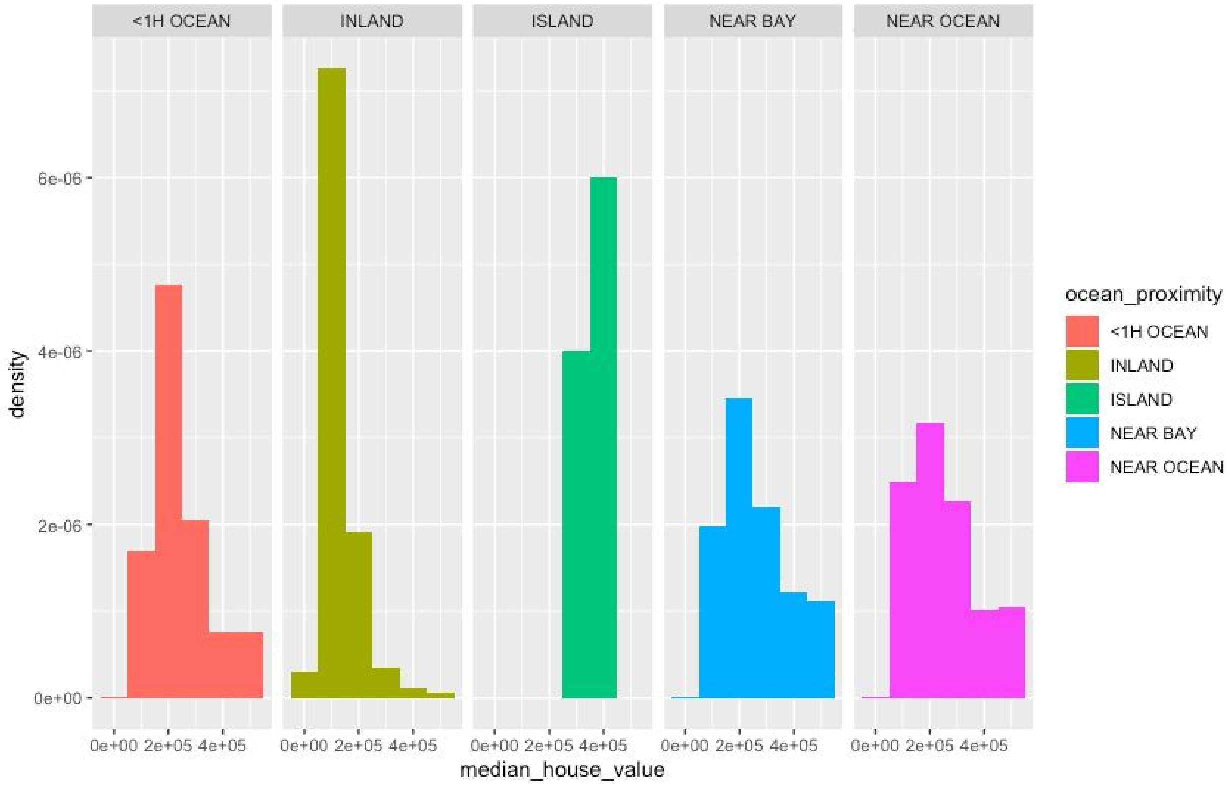
Cleaning/Preprocessing/Exploratory Analysis Visuals:

After conducting several preprocessing analyses and generating several plots I deleted variables that indicated outliers. These included populations greater than 5,500, households greater than 2,000, median house values greater than $500,000, median income less than $10,000, mean rooms greater than 11, mean bedrooms greater than 2.5 and less than 0.5. This is a new variable I created by dividing total rooms/bedrooms by households. Since the households variable describes a home unit, we are essentially dividing total rooms and total bedrooms by number of housing units. I also used a deletion when necessary command to remove the NA outputs from the total bedrooms variable. Then, I created a new binary variable called "location" from the ocean proximity categorical variable and removed the ocean proximity column from the cleaned data. Finally, I partitioned the data into training and testing groups. The figures below provide a visual representation of key variables of interest and their distribution:



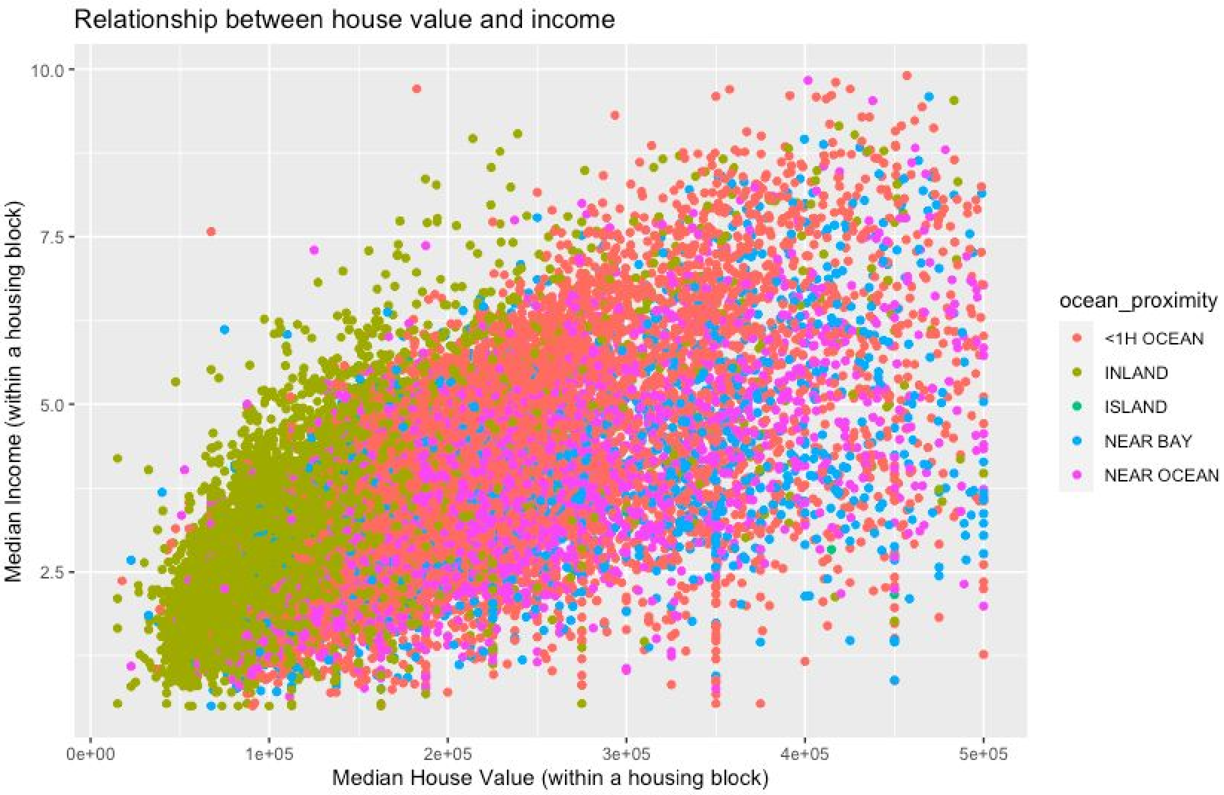
Example of before and after deletion of population variable


NEAR OCEAN
NEAR BAY
ISLAND
INLAND
<1H OCEAN
ocean_proximity
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
median_house_value
ocean_proximity
<1H OCEAN
INLAND
ISLAND
NEAR BAY
NEAR OCEAN

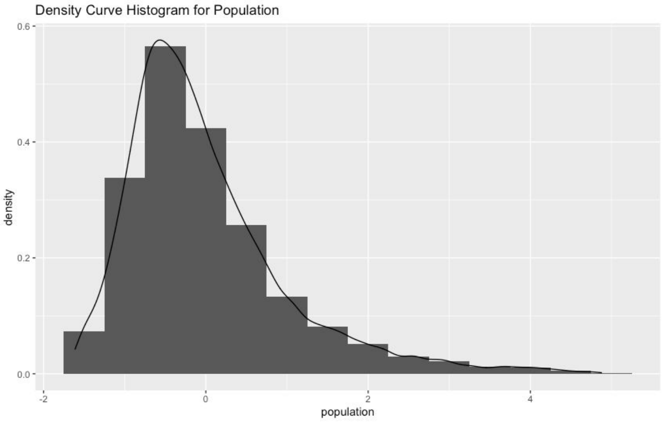

<1H OCEAN
INLAND
ISLAND
NEAR BAY
NEAR OCEAN
density
6e-06
4e-06
2e-06
0e+00
0e+00
2e+05
4e+05
median_house_value
ocean_proximity
<1H OCEAN
INLAND
ISLAND
NEAR BAY
NEAR OCEAN

Exploratory Analysis Visuals:

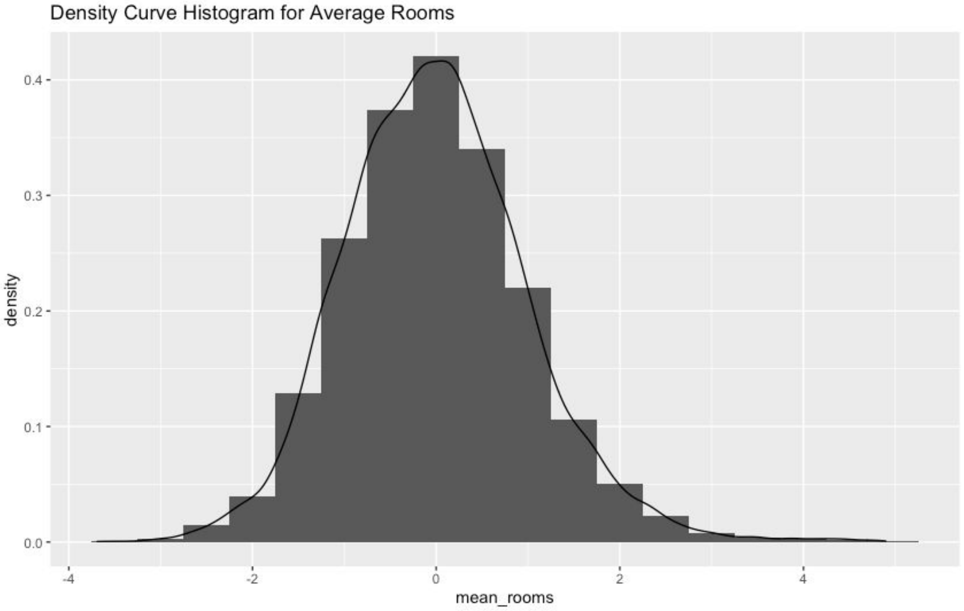
Plot a: Ggplot of median house value and median income for a housing block colored by ocean proximity

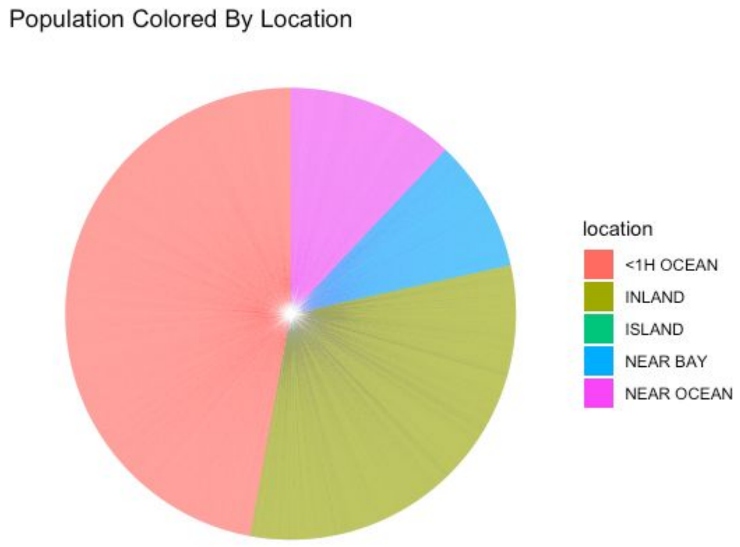


Plot b: Density curve histogram for population variable. Skewed indicating there are high density housing blocks in this dataset



Plot c: ggplot of mean rooms: Fail to reject null hypothesis, data is normally distributed (supported by jarque bera test)



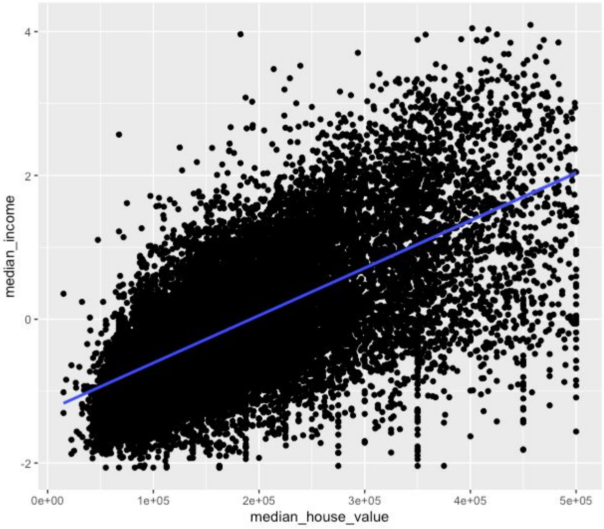


Predictive Modelling

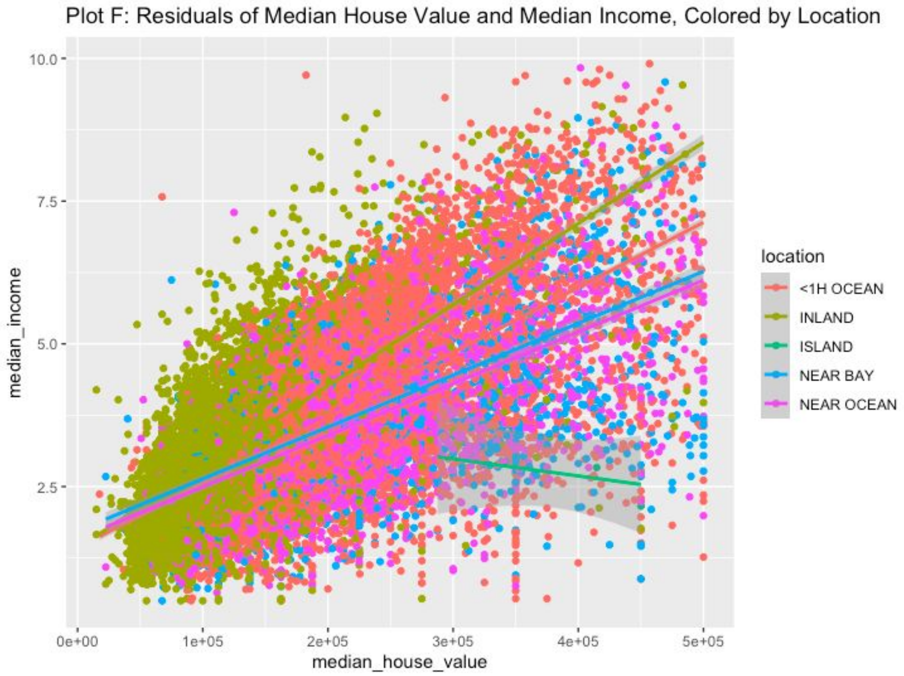
With linear regression model, I mainly wanted to test the significance of all variables to see if there are any that were not significant to our analysis From linear regression model LR3, found that homes that were inland, on an island, and near the bay were highly significant predictors of house value at the 1% level. Interestingly enough, homes that were near the ocean however were only significant at the 10% level in predicting house value. This could be due to a variety of factors such as wide variety in housing types by the ocean (older homes compared to luxury vacation homes intended for rentals). Also found that median home values for houses inland is $101,536.90 less than the average, $157,191.20 more than average for houses on an island, $13,295.30 more than average for houses near a bay, and $3,646.20 more than average for homes near the ocean. Below is an outline of other significant variables from linear regression models:

- Mean rooms is highly significant at predicting median house value. As the average number of rooms increases by 1, the median house value increases by $23,169.50.
- Mean bedrooms is highly significant at predicting median house value. As the average number of bedrooms increases by 1, interestingly enough, the median house value decreases by $10,207.10
- Population is **not** significant at predicting median house value. As population of a housing block increases by 1, median house value increases by $416.40
- Number of households is highly significant at predicting median house value with the increase of 1 household in a block, increasing the median house value by $10,628

Also generated visualization of residuals below:



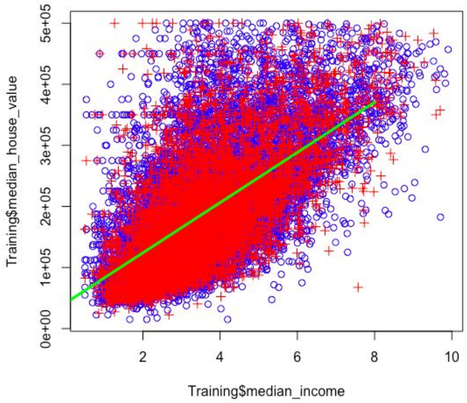
Ggplot of median house value and median income of residuals from linear regression model of the two (plot e)



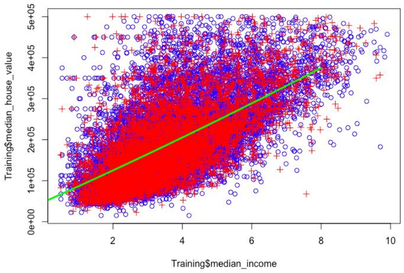
Plot F: Residuals of Median House Value and Median Income, Colored by Location

Modeling building

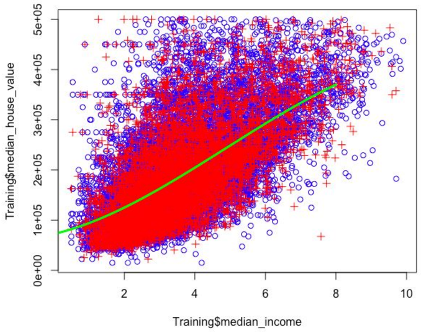
I built four multiple regression models to test which model would be best at predicting median house value: linear regression, quadratic linear regression, cubic linear regression, and logarithmic linear regression. I found that the quadratic model was the best model out of the four as it produced the lowest out of sample error of $20,2591.8 off from predicting median house value from median income. Interestingly enough though, while this model was the best at predicting on the training data there was very little difference between model 1 and model 2 in how well the model fit the data, as indicated by the R-Squared output. The best model that fit the data, with the highest R-Squared value was model 3, but had a higher out of sample error than model 2. Since our objective is to generate a model that best predicts housing value, we would select model 2 as it has the lowest out of sample error. Below are visualizations of the four models against both partitions:



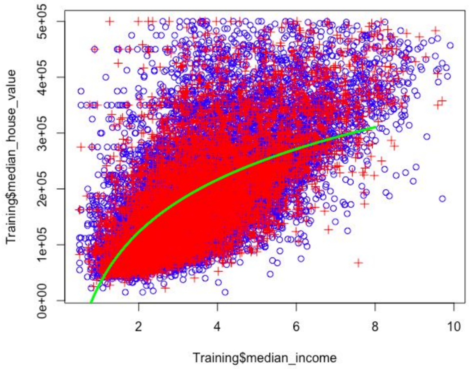
Model 1: Single Linear Regression



Model 2: Quadratic Linear Regression



Model 3: Cubic Linear Regression



Model 4: Logarithmic Linear Regression