**PREDICTIVE MODELLING FOR RESIDENTIAL PROPERTY PRICES.**

**BY**

**IDOWU DANIEL**

**Executive Summary**

This study aims to understand the relationships between variables in the dataset and relationships between the variable and the target variable. The study also seeks to predict the house price from the other variable in the dataset and see how they perform. The best three predictors are ocean proximity inland, median income and population. After conducting the exploratory data analysis on each variable in the dataset, I discovered that the median age of people purchasing a house was around 29years old, the median rooms for each block is around 2,127, also the median number of bedrooms in each block is about 438. The median population size of people in each block is around 1,166 and the median households in each block is 409. The median income of each household in a block is around $35,000 and the median price for houses in each block is around $179,700. The reason for using median is to solve the issue of outliers, using the mean value will only calculate the average of each variable without considering the outliers. I used a bar chart to check the relationship between the ocean proximity and the house values. From the data I saw that people who bought houses near the ocean spend more on houses and this is because the Pacific Ocean is closer to California. After this I found out that people who bought houses near bays and on islands spend about the same amount on house prices. People who buy houses inlands spend the least amount of money on houses. This makes sense because wealthy and rich people are those who afford houses on islands and beaches or bays. I used the correlation plot to check for correlation between variables in the data. There is a strong relationship or correlation of about 0.93 between the number of bedrooms and the total number of rooms. This suggests that as the number of rooms increases, the number of bedrooms also tends to increase. Similarly, there is a correlation of around 0.86 between the population and the total number of rooms. This indicates that areas with more rooms tends to have more populations. Additionally, there is a robust correlation of about 0.92 between the number of households and the total number of rooms. This suggests that areas with more rooms tends to have more households. Furthermore, a noteworthy correlation of approximately about 0.87 exists between the total number of bedrooms and the population. This implies that areas with larger population tends to have more bedrooms. A striking correlation of about 0.97 is observed between the total number of bedrooms and the number of households. This indicates a strong relationship between these two variables, suggesting that areas with more bedrooms also tend to have more households. Moreover, there is a correlation of about 0.91 between the population and number of households. This suggests that areas with larger population tend to have more households. Lastly, there is a moderate correlation of about 0.69 between the median house value and the median income. This indicates a relationship between these variables, suggesting that areas with higher median incomes tend to have higher median house values. After building the dataset into train and test set, train set was used to build the model and the test set was used to test and predict from the model. The mean square error is about 4617386073 while the root mean squared error (RMSE) is about 67951.35. This means that the predicted results were a little closer to the actual results and the number of errors gotten was not much but the model can still do better. The R-Squared shows that about 62% of the independent variable is explained by the dependent variable. This means that the model performed about 62% which is fairly ok.­­­

Contents

[1. Introduction 1](#_Toc160158339)

[1.1. Problem Domain 1](#_Toc160158340)

[1.2. Statistical Question 1](#_Toc160158341)

[2. Methodology 1](#_Toc160158342)

[2.1. Define the objectives 1](#_Toc160158343)

[2.2. Data Collection 2](#_Toc160158344)

[2.3. Data Cleaning 2](#_Toc160158345)

[2.4. Exploratory Data Analysis 2](#_Toc160158346)

[2.5. Data Preprocessing 2](#_Toc160158347)

[2.6. Model Selection 2](#_Toc160158348)

[2.7. Model Training 3](#_Toc160158349)

[2.8. Model Evaluation 3](#_Toc160158350)

[3. Results and Discussion 3](#_Toc160158351)

[3.1. Data Collection 3](#_Toc160158352)

[3.2. Data Cleaning 4](#_Toc160158353)

[3.3. Exploratory Data Analysis 4](#_Toc160158354)

[3.3.1.1. Visualization of the variables in the data 5](#_Toc160158355)

[3.3.1.2. Relationship Between the dependent and independent variable 6](#_Toc160158356)

[3.3.1.3. Relationships between variables using correlation plot 7](#_Toc160158357)

[3.4. Data Preprocessing 8](#_Toc160158358)

[3.5. Model Selection 8](#_Toc160158359)

[3.6. Model Training 9](#_Toc160158360)

[3.7. Model Evaluation 10](#_Toc160158361)

[4. Conclusion and Future work 11](#_Toc160158362)

# Introduction

The housing market is a dynamic and complex system influenced by various factors that affect property prices. Understanding these factors and their impact on housing prices is crucial for homeowners, real estate agents, investors, and policymakers alike. Predictive modelling offers a powerful tool to analyse historical data, identify trends, and forecast future price movements in the residential property market. In this study, we aim to explore the intricate relationship between different variables and housing prices, leveraging advanced statistical techniques to develop accurate predictive models. By examining factors such as location, property characteristics, economic indicators, and demographic trends, we seek to provide valuable insights into the dynamics of the housing market and enable stakeholders to make informed decisions. Through our analysis, we strive to contribute to a deeper understanding of the housing market landscape and provide practical solutions for price prediction and trend analysis in residential real estate.

## Problem Domain

The problem domain for house price prediction revolves around the challenge of accurately estimating the market value of residential properties. This task is of significant interest to various stakeholders, including homebuyers, sellers, real estate agents, financial institutions, and policymakers. The primary objective is to develop predictive models that can effectively forecast housing prices based on relevant variables and market dynamics. This study will use linear regression model to carry out the predictive analysis.

## Statistical Question

These are some of the statistical questions that needs to be answered over the course of this research:

1. What is the relationship between housing prices and various independent variables such as location, square footage, number of bedrooms, number of bathrooms, and amenities?
2. Can the model accurately predict housing prices based on the selected independent variables, and what is the level of confidence or uncertainty associated with these predictions?

# Methodology

The following statistical techniques are used to analyse the dataset and extract meaningful information to reach the desired objective.

## Define the objectives

Clearly defining the objective and goals of a project is the first step in research analysis. The objectives and questions to be answered has been done in the introductory section and this shows the problems I aim to answer and solve with this data.

## Data Collection

The dataset used for this project is gotten from Kaggle. Kaggle is a popular online platform for data science and machine learning enthusiasts, researchers, and professionals. It offers a wide range of datasets, competitions, courses, and collaborative features to support data-driven projects and facilitate learning and knowledge sharing in the data science community. The link to the dataset [California House Price (kaggle.com)](https://www.kaggle.com/datasets/shibumohapatra/house-price/data).

## Data Cleaning

I cleaned the data to address issues such as missing values, outliers, duplicate records and inconsistences. I used the **any(is.na())** function to check for null values in the dataset and it returned True, meaning there are null values in my dataset. I used **colSums(is.na())** to check for the columns with null values and discovered it was the total bedroom variable with 207 null values. I used a for loop to input the mean values of the total bedroom variable to replace the null values. After I removed all the null values, I dropped the longitude and latitude values because I don’t need them for my analysis.

## Exploratory Data Analysis

I Explored the data using descriptive statistics and visualization techniques to gain insights into its distribution, patterns, and relationships. EDA helps identify trends, outliers, and potential relationships between variables. I used the structure and summary function to check the structure of the data and to also check the summary and perform statistical analysis such as mean, median, minimum, maximum among others on the data.

I used a visualization package in R called ggplot2 to visualize each variable in my dataset to see and understand the distribution of variables in the dataset. I used ggplot2 with histogram to see the distribution of each variable against the count so as to know and drive insights. I also used scatterplot to show the relationships between each the dependent variable (median house values) against each independent variable in the dataset.

## Data Preprocessing

It preprocesses the data to prepare it for analysis. This may involve standardizing or normalizing numeric features, encoding categorical variables, and scaling the data to ensure all variables are on a similar scale. I pre-processed my data by encoding the categorical variables in my dataset. The ocean proximity variable has four (4) and I used the dummy package to encode the categorical variables so as to be able to use it when building the model.

## Model Selection

Choose appropriate models or algorithms based on the nature of the data and the objectives of the analysis. Consider factors such as the type of problem (classification, regression, clustering), the size of the dataset, and the interpretability of the model. For this study I chose to use the linear regression model because this is a regression problem and want to see how the model performs.

## Model Training

I used the caTools package to split my dataset into train and test data, with the split ratio of 0.8, which means that the training set takes 80% of the data and testing set took about 20% of the data.

## Model Evaluation

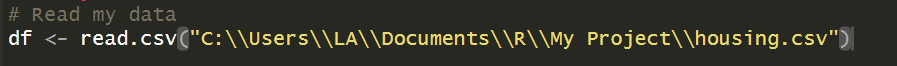
Evaluate the performance of the trained models using appropriate metrics and techniques. This step helps assess how well the models generalize to new data and whether they meet the objectives of the analysis. After I created the model using the training set, I checked the summary of the model to see the results, I evaluated the coefficients, the standard errors, t-values, p-values, residuals, the adjusted R values, the level of freedom amongst others. Then I calculated the R2 to see if the models if the models fit and how the independent variables performed against the dependent variables.

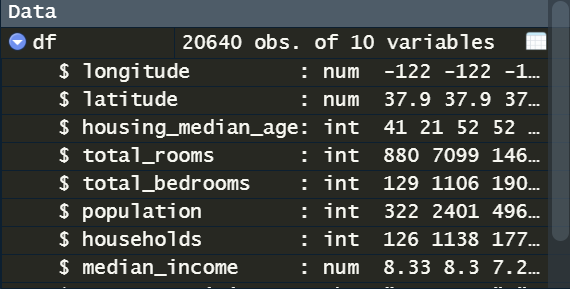
# Results and Discussion

This section discusses the results gotten from the experiments carried out and the results will be discussed to see the insights and knowledge gained from this research.

## Data Collection

The data is collected from Kaggle as explained in the methodology section. The dataset was read into R studio Using read.csv () function. This function reads the csv file into R to perform analysis on them. The dataset has 10 variables with 20,640 observations. Then I also did a quick summary and checked the structure of my data.



****

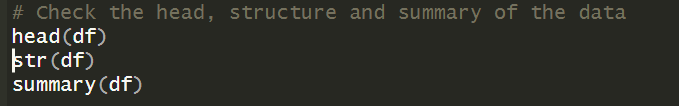
****

Figure 1: Showing a picture of my dataset.

## Data Cleaning

After collecting the data and checking the data to see the variables and what it constitutes, I checked to see if there are no null values in the data and if there were any, I will use the median of the column or variable where there are null values to replace any null value found. To do this I used the following to check for null values.

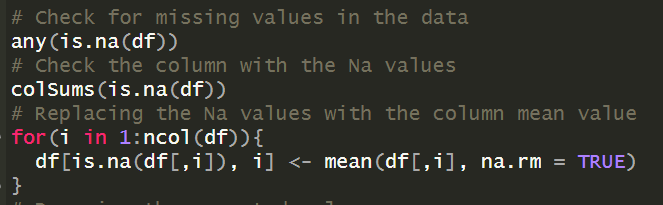


Figure 2: Checking for null values and replacing the null values.

After this, I dropped the columns I was not going to use and those columns are the longitude and the latitude column.

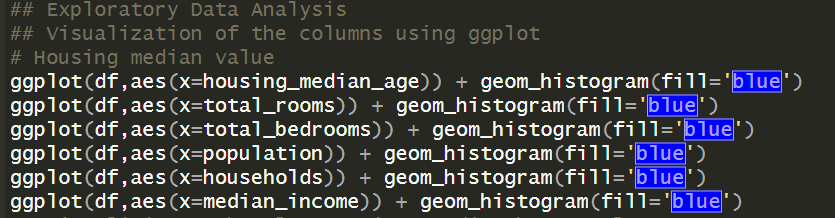


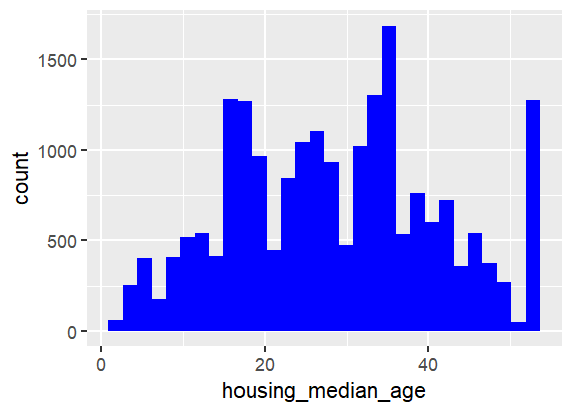
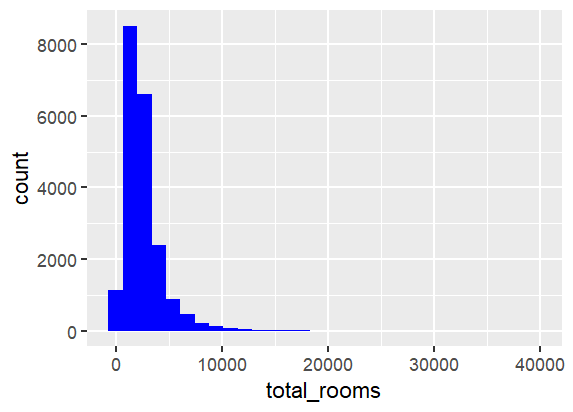
Figure 3: Dropping the longitude and latitude column.

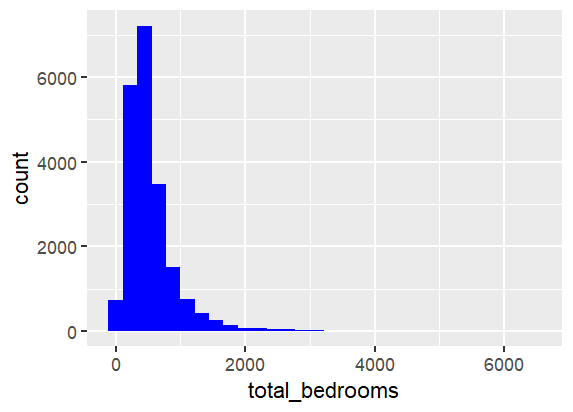
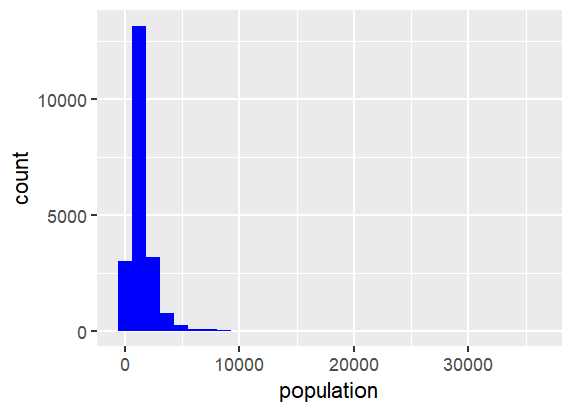
## Exploratory Data Analysis

I used the EDA to visualize each variable in the dataset and relationships between the independent variables and the target variables. With this, I was able to gain insights to see relationships within our dataset and the best tool for this is to use the ggplot2 package.

### Visualization of the variables in the data







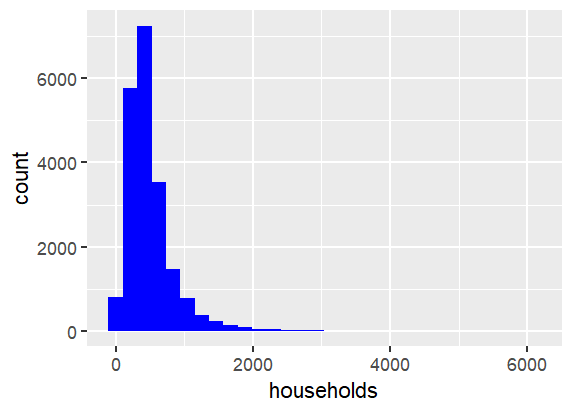
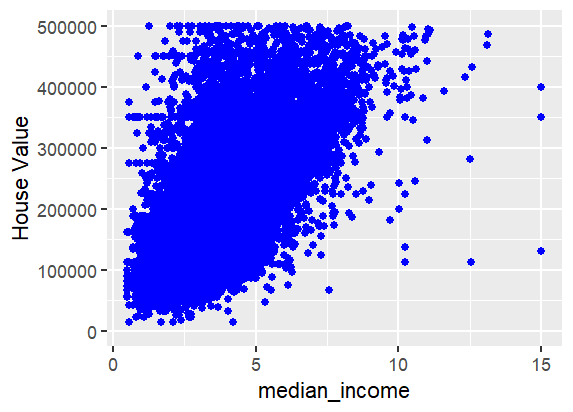


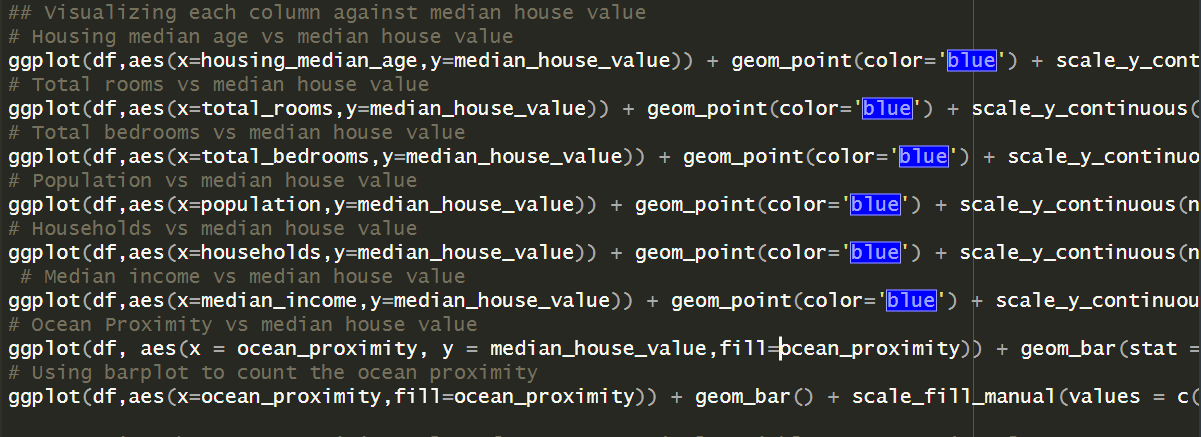
Figure 4: Showing the code to generate the plot and each plot generated.

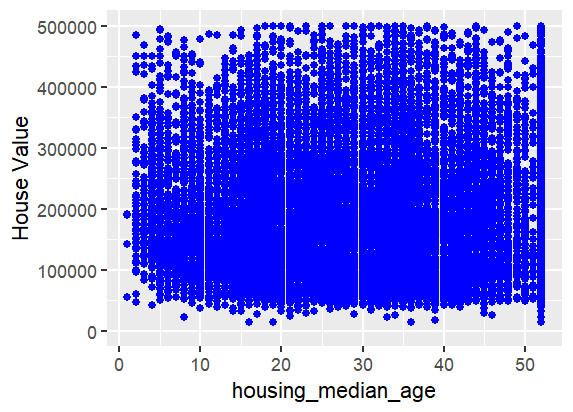
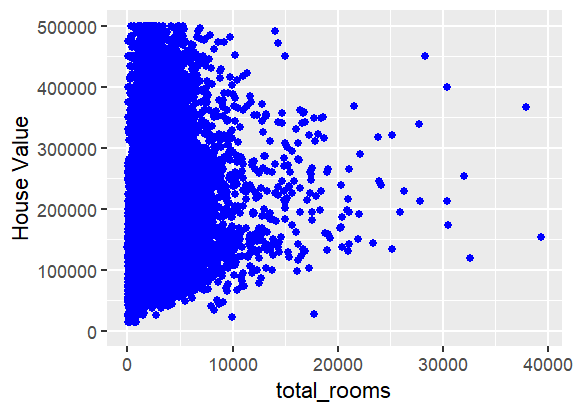
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary**  **statistics** | **Housing**  **Median**  **Age** | **Total Rooms** | **Total Bedrooms** | **Population** | **Households** | **Median Income** | **Median House Value** |
| **Min** | 1.00 | 2 | 1.0 | 3 | 1.0 | 0.4999 | 14999 |
| **1st Quarter** | 18.00 | 1448 | 297.0 | 787 | 280.0 | 2.5634 | 119600 |
| **Median** | 29.00 | 2127 | 438.0 | 1166 | 409.0 | 3.5348 | 179700 |
| **Mean** | 28.64 | 2636 | 537.9 | 1425 | 499.5 | 3.8707 | 206856 |
| **3rd Quarter** | 37.00 | 3148 | 643.2 | 1725 | 605.0 | 4.7432 | 264725 |
| **Max** | 52.00 | 39320 | 6445.0 | 35682 | 6082.0 | 15.0001 | 500001 |

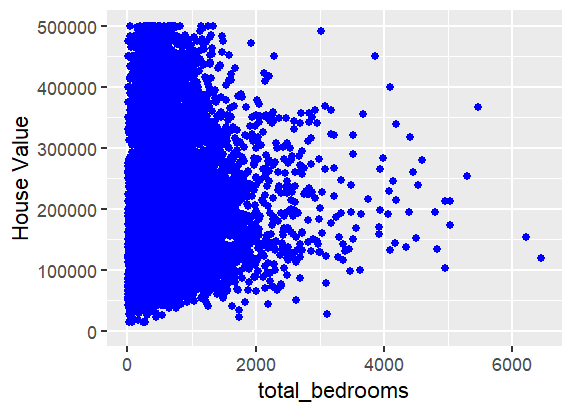
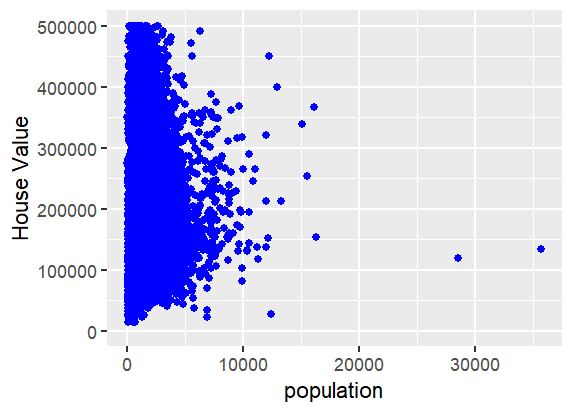
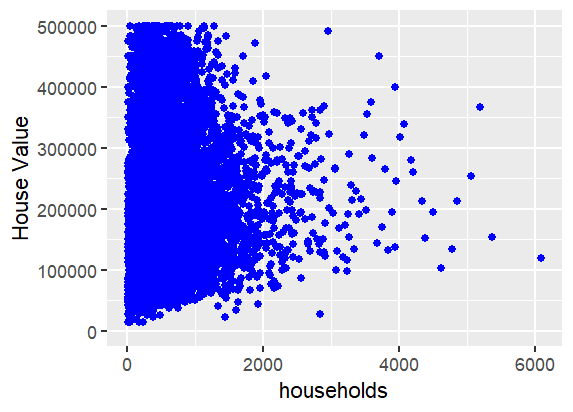
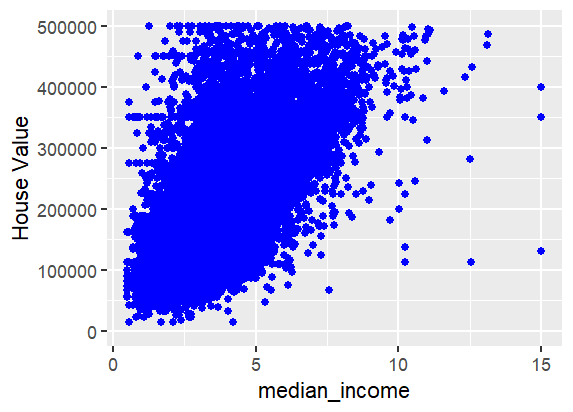
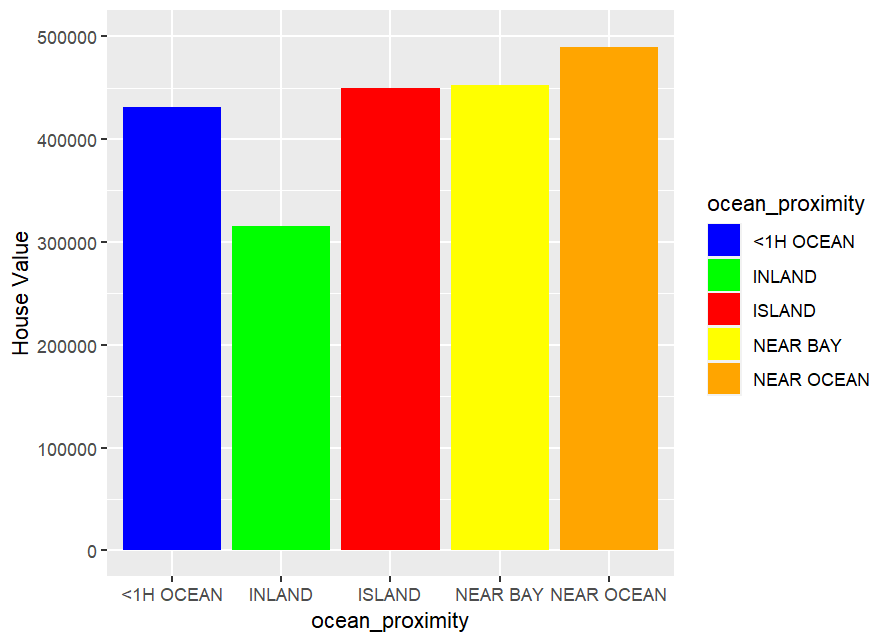
**Table 1**: Showing a summary statistic of each variable in the dataset.

From Table 1 and Figure 4 above, this shows the statistics of the variables in the dataset. The housing median age in the data is 29 years old and this means that for a block the average age of people that buys a house is 29 years old. The median total rooms in a block is 2,127 and the median total bedrooms in a block is 438. The median population in a block is 1,166 and the median household in a block is 409. The median income in each block is around $35,000 and the median house value in each block is $179,700.

### Relationship Between the dependent and independent variable







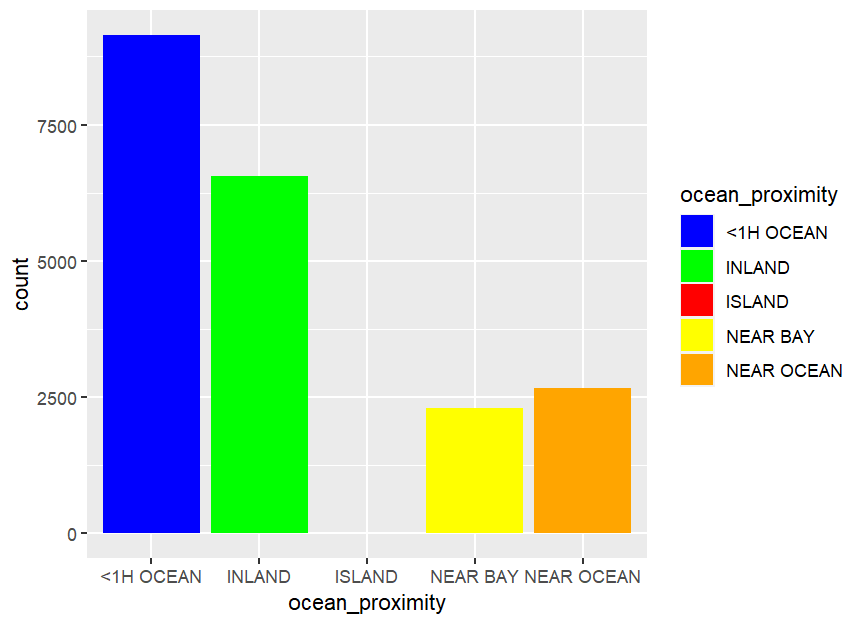
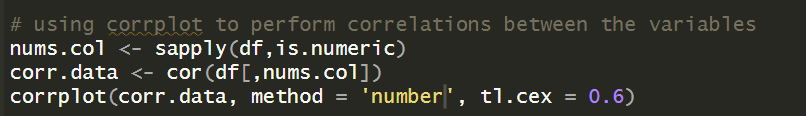


Figure 5: Showing a scatter plot for each variable against the target variable and the code to generate it.

### Relationships between variables using correlation plot

The correlation matrix of the variables under study is presented below:



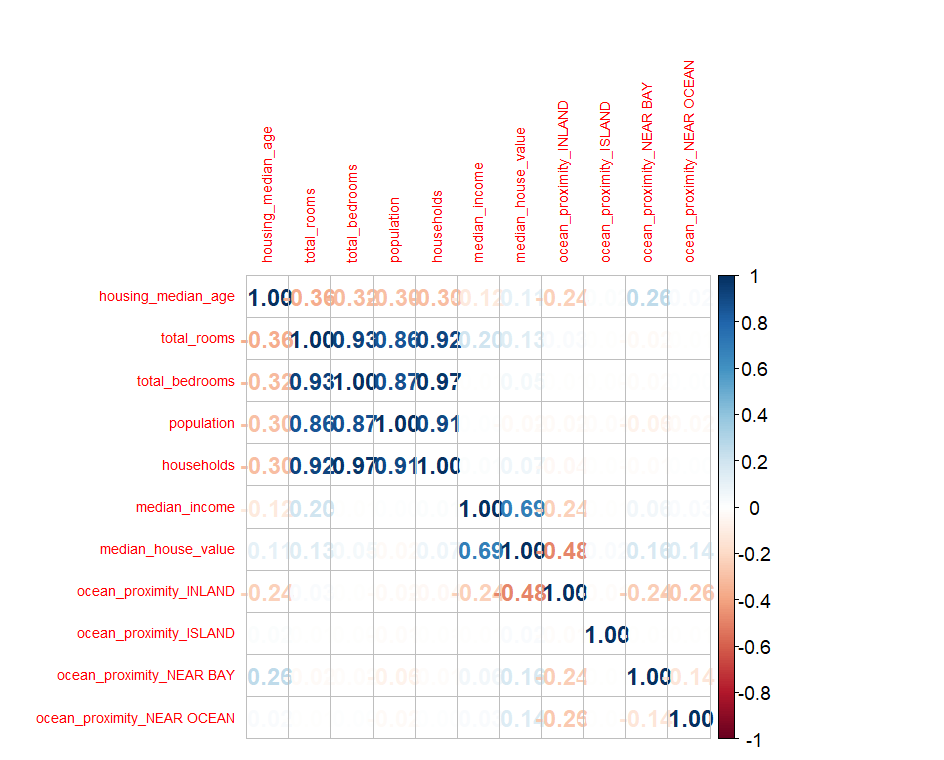


Figure 6: Showing a correlation plot among variables and the code to generate it.

From Figure 6 above, there are significant correlations between several variables. Firstly, there is a robust correlation of approximately 0.93 between total bedrooms and total rooms. Similarly, a correlation of around 0.86 exists between population and total rooms. Moreover, a strong correlation of about 0.92 is observed between households and total rooms. Additionally, there is a noteworthy correlation of approximately 0.87 between total bedrooms and population, and a striking correlation of 0.97 between total bedrooms and households. Furthermore, a correlation of approximately 0.91 is noted between population and households. Lastly, there is a moderate correlation of about 0.69 between median house value and median income.

## Data Preprocessing

I pre-processed my data by encoding the categorical variables in my dataset. The ocean proximity variable has four (4) and I used the dummy package to encode the categorical variables so as to be able to use it when building the model.

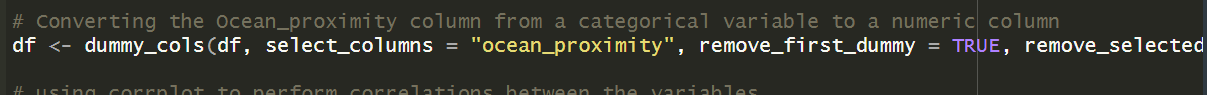


Figure 7: Showing a code that encodes the categorical variable.

## Model Selection

I chose to use linear regression because this is a regression problem and I want to see how the algorithm performed with the dataset provided.

## Model Training

I used the caTools to split my data into 80% for training set and 30% for test set. I used all the variables against the median house value which is the independent variable and I passed in the train set as the data.

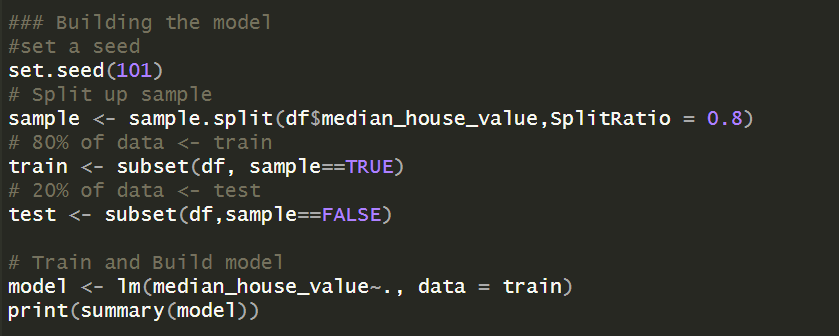


Figure 8: Showing the model building.

After calling the summary on the model created, the results gotten are shown below:

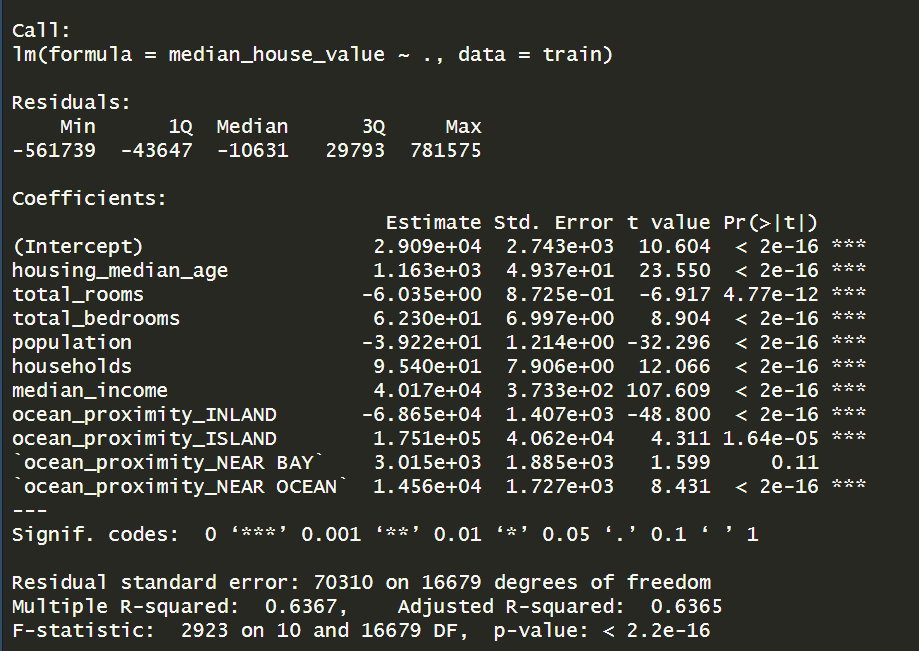


Figure 9: Showing the summary of the model.

## Model Evaluation

After the model has been trained using the training set, then prediction will be done with the model created and the test data is being used to test the trained model.

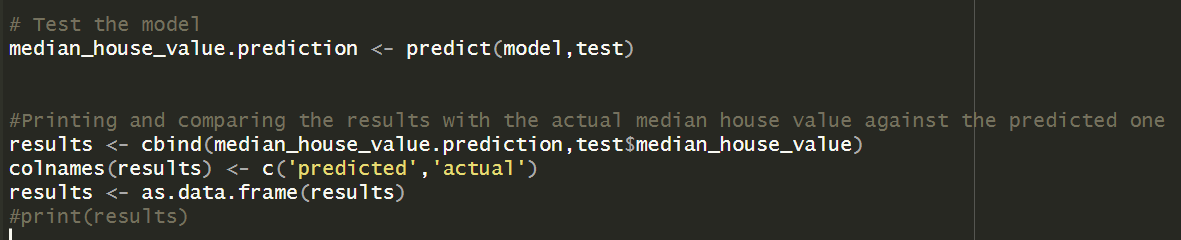


Figure 10: Showing the model being predicted.

From Figure 10 above, after the prediction is done, I used the column bind function to bind together the predicted value with the predictor variable from the test dataset and renamed the column to predicted and actual respectively and passed them as a data frame. Why I did is because with this newly created data frame, I will be able to get the mean squared error and the root mean squared error and likewise get the R squared value.

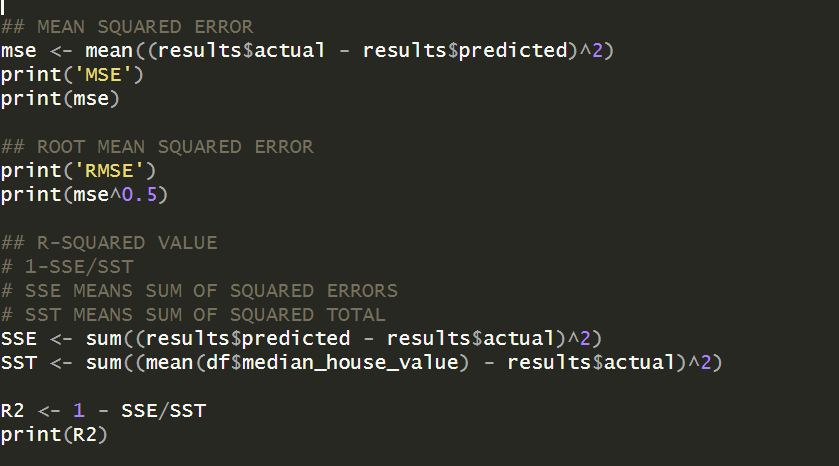
****

Figure 11: Showing the mean squared error, root mean squared error and the R-squared.

The mean squared error (MSE) is 4617386073 while the root mean squared error (RMSE) is 67951.35. what mean squared error does is to calculate the mean of difference between the actual and predicted variable and it is used if the model’s prediction is closer to the actual variable. The closer it is the better. Root Mean Squared Error (RMSE) is a measure of the average deviation between the actual and predicted values in a regression problem. It is calculated by taking the square root of the Mean Squared Error (MSE). RMSE provides a more interpretable measure of error compared to MSE because it is in the same units as the dependent variable. R- squared measures the proportion of the variance in the dependent variable that is explained by the independent variable in a regression model. A higher R2 values indicates that the model fits the data better. The formula to calculate the R2 is **1 – SSE/SST.**

Where **SSE: Sum of squared errors,**

**SST: Sum of squared total.**

The R-squared value for this model is 0.6186742, which means that about 61.87% of the variance in the dependent variable is explained by the independent variables included in the regression model.

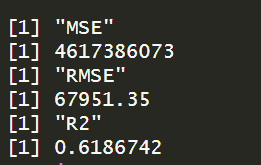


Figure 12: Showing the results from the evaluation.

# Conclusion and Future work

This study answered all the statistical questions raised in this report. I was able to understand and see relationships between different variables in the dataset. I was able generate insights and see the median house value of each house in a block in California cost using the linear regression model. Linear regression model was able to explain about 62% of variables in the dataset which was ok but didn’t perform well. In the future I will try to use other regression model such as random forest, support vector machine, gradient boosting and lasso regression to see which models best fit my dataset.