California Housing Price Analysis

Jing (Joyce) Chen
Haomin Mai
Euijun(Jun) Kim
Casey Truong
Sungwon (Alex) Lee
Celeste Vargas

STATS 140XP Final Project
University of California, Los Angeles
Professor. Vivian Lew
November 27, 2022

Contents

1 Abstract

It is known that California is the most populous state and the largest state economy in the United States, with one of the highest average housing prices at $683,000, compared to the national average of $450,600. This can make buying a home in California a slow, and sometimes unattainable, process. However, to help current and the next generation of home buyers, we wanted to find the most important factors affecting average housing prices that buyers should take into consideration. To employ our analysis, we performed hypothesis testing, multiple linear regression, cross-validation, Lasso and Ridge regression, XGBoost, and a Random Forest. Lastly, in this analysis, we used the California Housing Prices Dataset from Kaggle.

2 Introduction

Home prices in California steadily rose 63.1% in the span of five years, and around 20% in 2020, with one of the factors pointed to the increase in home prices as the imbalance in the supply and demand in the housing market. In this paper, however, we want to explore additional factors that can also drive up housing prices. Some real state companies indicate that location, interest rates, home size, and location affect the average home value. However, to dive further, we utilized the California Housing Prices Dataset from Kaggle to predict the variables affecting the median house value.

2.1 Data Description

The original data set consisted of a total of 10 features and 20,640 observations.

The following are the variables:

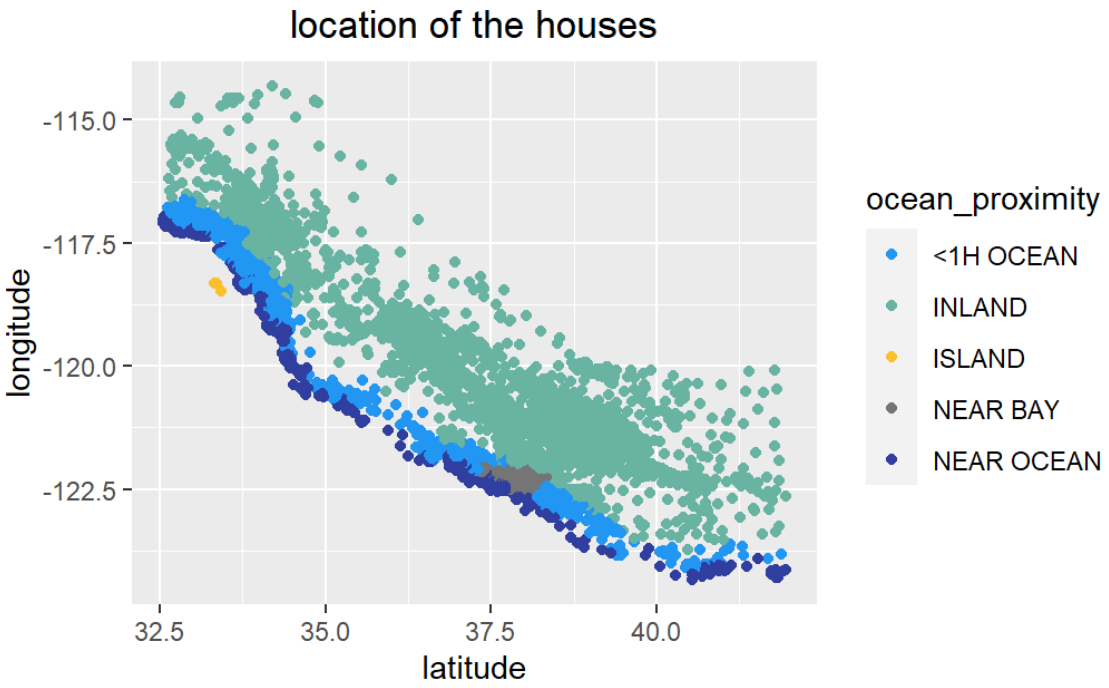
1. longitude: A measure of how far west a house is; a higher value is farther west
2. latitude: A measure of how far north a house is; a higher value is farther north
3. housingMedianAge: Median age of a house within a block; a lower number is a newer building
4. totalRooms: Total number of rooms within a block
5. totalBedrooms: Total number of bedrooms within a block
6. population: Total number of people residing within a block
7. households: Total number of households, a group of people residing within a home unit, for a block
8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. medianHouseValue: Median house value for households within a block (measured in US Dollars)
10. oceanProximity: Location of the house w.r.t ocean/sea

2.2 Data Cleaning

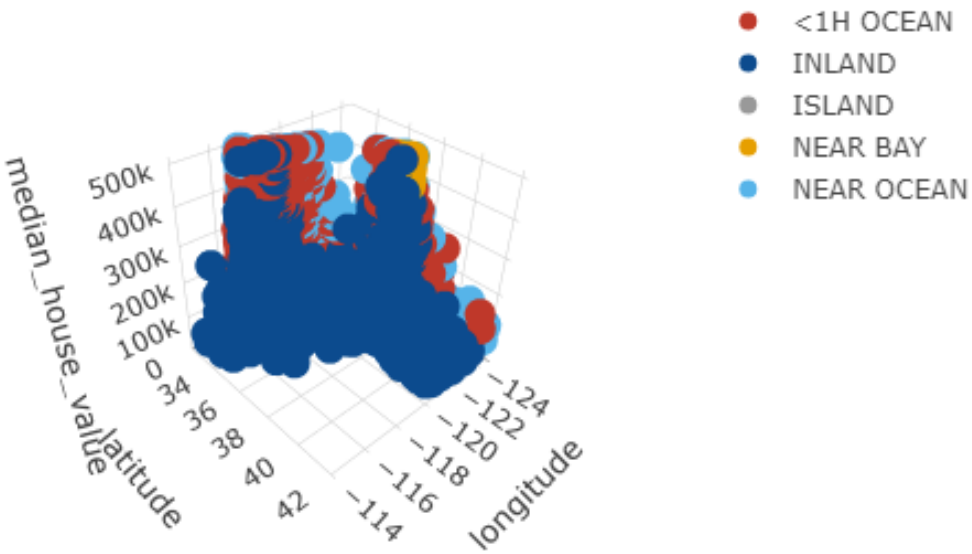
To clean our data, we decided to drop any missing values. By doing this, we lost around 207 observations, amounting to 1% of our total data observations. Additionally, we also encoded our Ocean Proximity variable into numerical to be able to use for modeling and was treated as a factor. Lastly, we also standardized our variables so that our measurements were not constricted within a block of homes, but rather per observation, and removed the columns longitude and latitude because they did not serve any meaningful analysis.

3 Exploratory Data Analysis

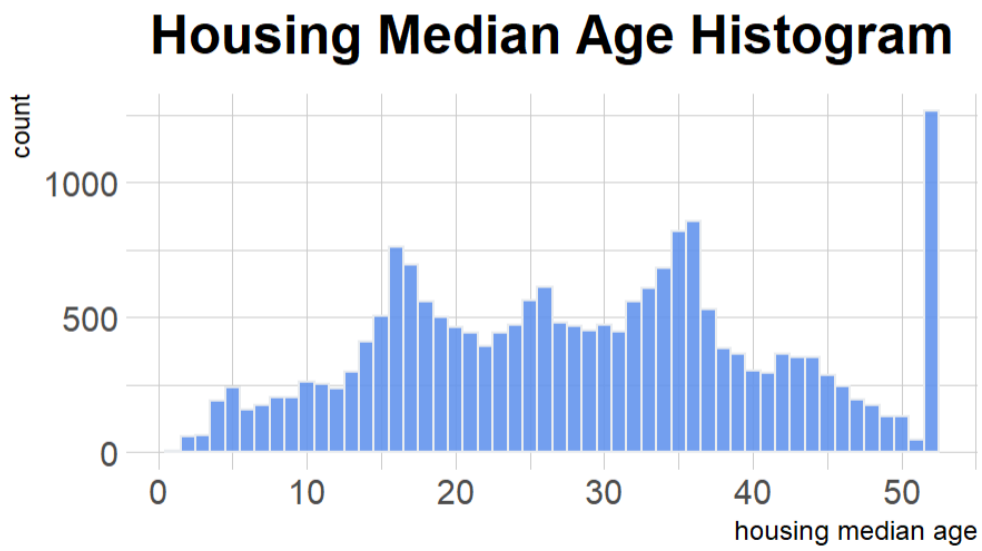
To begin our Exploratory Data Analysis, we were interested in observing the number of homes in the data by ocean proximity. As we can see in the figure down below, most of the homes in the data were inland, followed by homes less than one hour away from the ocean. A takeaway from this analysis is that perhaps homes near the ocean are more coveted because of limited availability, and encourages us to take note of ocean proximity in our future modeling.



To continue our analysis, we also observed the relationship between ocean proximity and our predictor median housing value. From the figure below, we can see that home values do tend to increase as they get closer to the ocean; however, there are still some homes that are inland where their values are also high.



Lastly, we explored the Housing Median Age variable. From our histogram, we are able to see that the distribution of this variable is bimodal, with most homes being around 20 years old, 35 years old and over 50 years old.

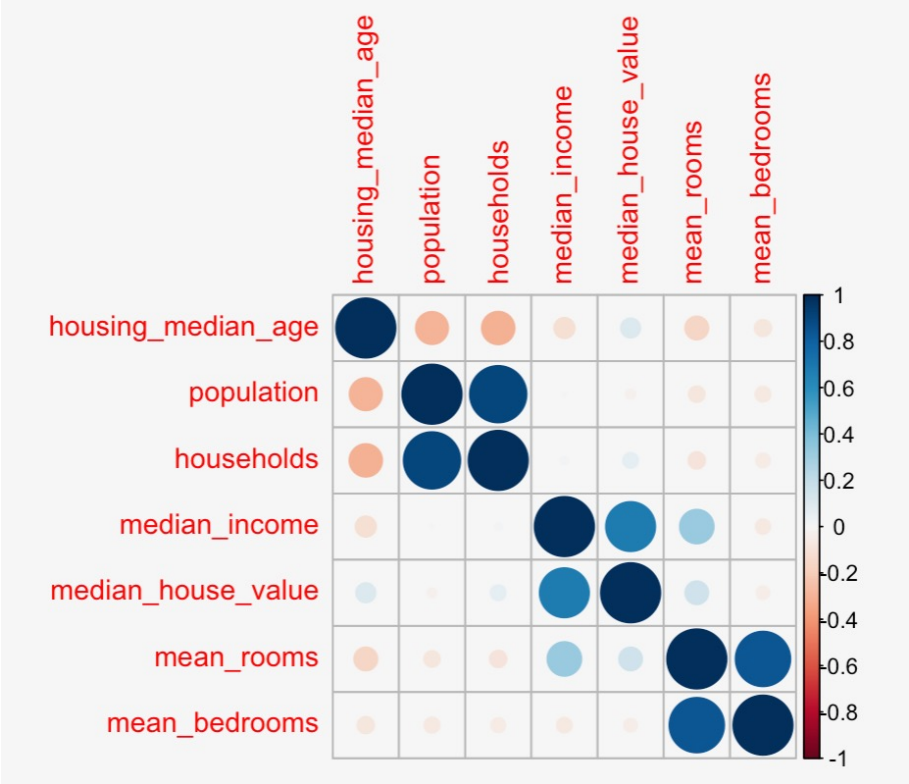


4 Feature Selection

For our feature selection, we observed the multicollinearity between all of our variables. We also employed stepwise and subset methods; however, we decided to use the variables obtained from multicollinearity because it reduced the number of columns needed the most.

From the figure below, we are able to observe that variables such as population, households, median income, median house value, mean bedrooms and median rooms were the variables with the most multicollinearity between each other.

Below is the first correlation plot:



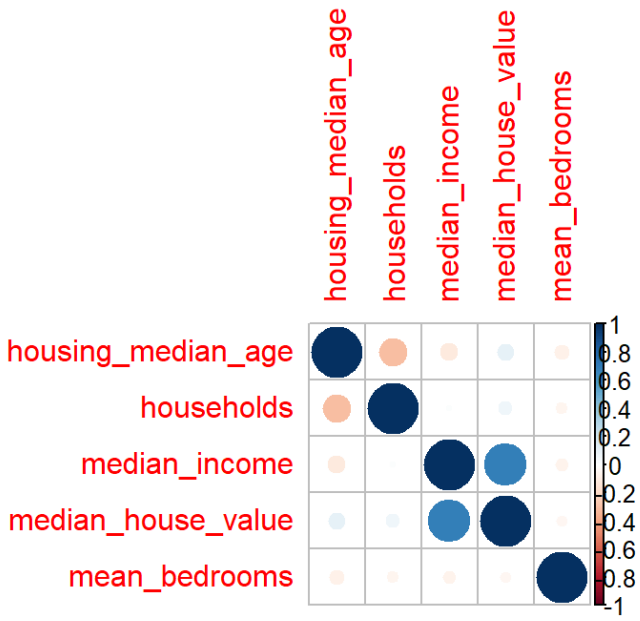
We decided to keep the median house value because it is our predictor and median income because it was correlated with our predictor.

Moreover, we also decided to remove mean rooms because we observed in future analysis that the mean number of bedrooms was a more accurate predictor of median housing prices.

To continue, we decided to remove population because the column households accounts for population and also because it is correlated to mean rooms. Once we removed these variables, we are left with no high multicollinearity.

As we can notice, we did not include ocean proximity in our correlation analysis but will be included during our modeling process because our exploratory data analysis suggested we should further explore that variable's effect on median housing prices.

Below is the second correlation plot:


housing_median_age
households
median_income
median_house_value
mean_bedrooms
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
-1

5 Hypothesis Testing

There are various hypothesis that can be considered with our data. We decided to come up with three hypotheses regarding the factors of effect on median housing value.

1. A home's proximity to the ocean does not affects median housing value.
2. Housing median age affects median housing value.
3. Population increases median housing value.

5.1 First Hypothesis

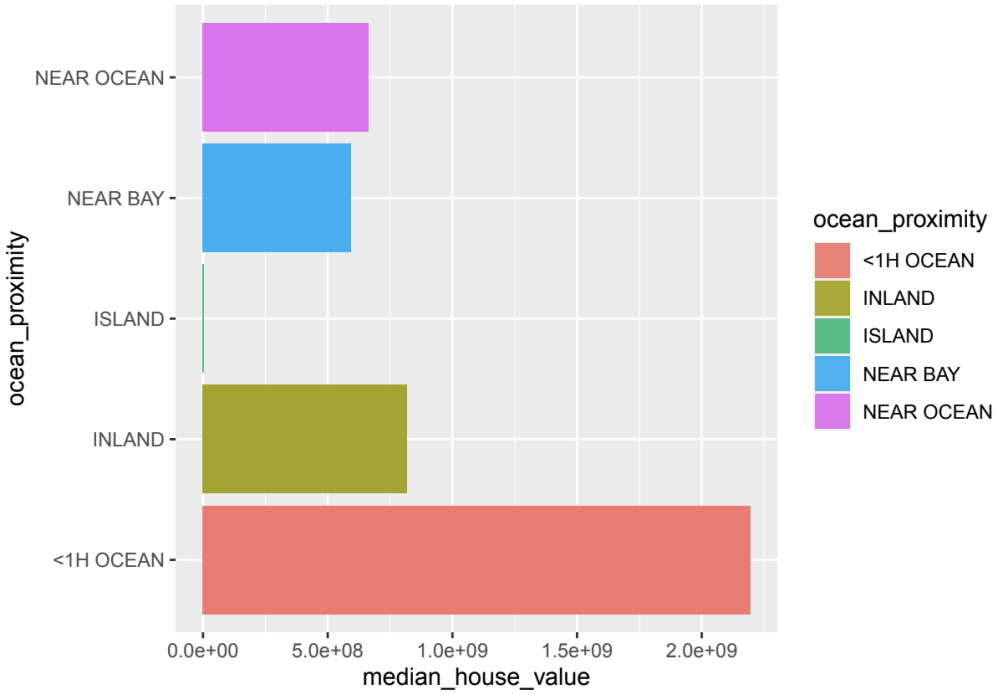
The first hypothesis we made was the proximity to the ocean for the house does not affects median housing value (Null hypothesis).

```
summary(lm(median_house_value ~ factor(ocean_proximity), new_df))
```

```
##
## Call:
## lm(formula = median_house_value ~ factor(ocean_proximity), data = new_df)
##
## 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 ***
## factor(ocean_proximity)2   -115371       1639 -70.372  < 2e-16 ***
## factor(ocean_proximity)3    140172      45083   3.109  0.00188 **
## factor(ocean_proximity)4     19011       2366   8.035 9.88e-16 ***
## factor(ocean_proximity)5      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
```

By looking at the summary table where we factorized ocean proximity, It is clearly labeled that p-value is less than 0.05, which is our significance level α. Thus, we do have enough convincing evidence to reject null hypothesis and accept alternative hypothesis - "A home's proximity to the ocean does affects median housing value."

Also, if we visualize the relationship from the plot, we can show as following:



This shows that the proximity to the ocean for the house which is less than 1 hour has significantly higher median house value compare to other located houses.

5.2 Second Hypothesis

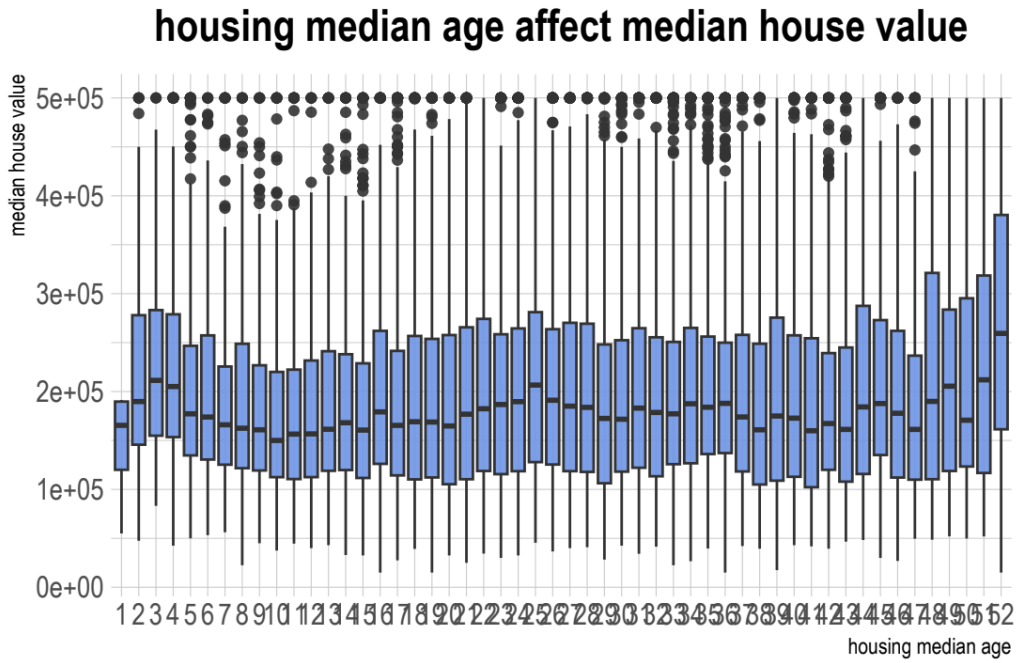
The second hypothesis we made was the housing median age does not affects median housing value (Null hypothesis).

```
summary(lm(median_house_value ~ housing_median_age, new_df))
```

```
##
## Call:
## lm(formula = median_house_value ~ housing_median_age, data = new_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -214665  -85114  -25771   58290  319123
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        178926.58    1994.76    89.7   <2e-16 ***
## housing_median_age    975.72      63.77    15.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114800 on 20431 degrees of freedom
## Multiple R-squared:  0.01133,    Adjusted R-squared:  0.01128
## F-statistic: 234.1 on 1 and 20431 DF,  p-value: < 2.2e-16
```

By looking at the summary table of the linear regression model, where response variable (y) is set to median house value and predictor (βs) includes only intercept and housing median age. It is clearly labeled that p-value is less than 0.05, which is our significance level α. Thus, we do have enough convincing evidence to reject null hypothesis and accept alternative hypothesis - "housing median age does affects median housing value."

If we visualize the relationship between the response variable and predictor, we can show as following:



By observing the plot above, we can assume that housing median age greater than 51 would worth more value as their maximum house value are higher than all other housing age.

5.3 Third Hypothesis

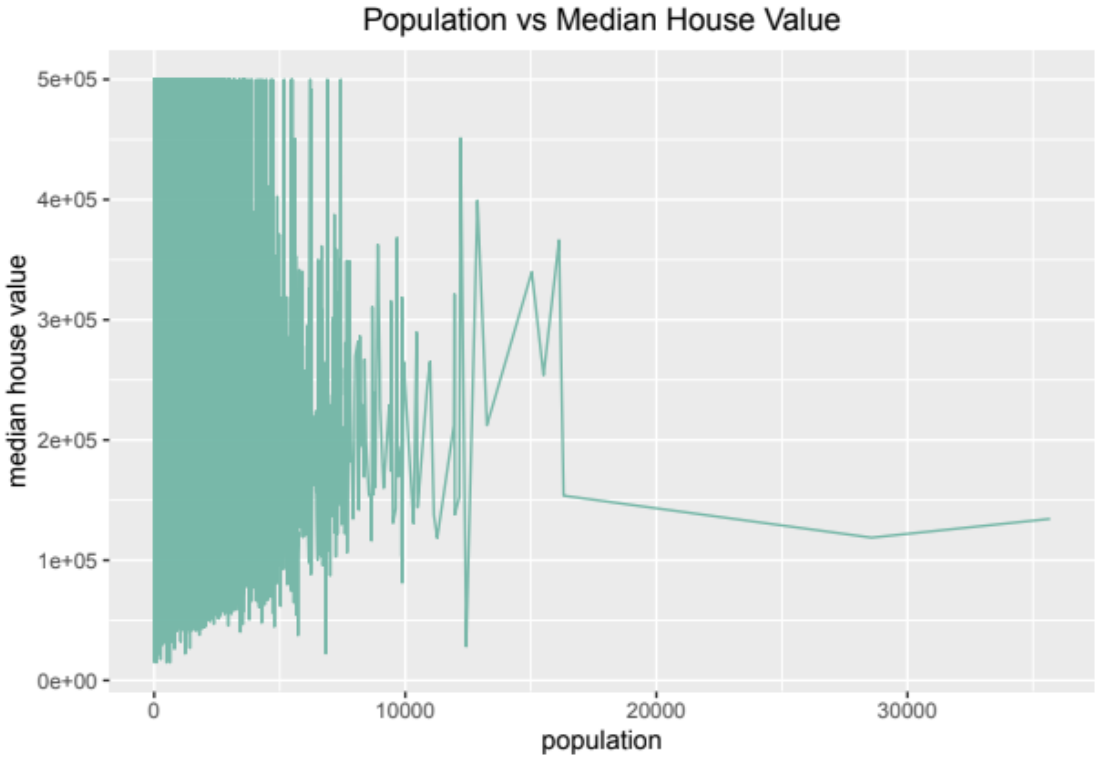
The third hypothesis we made was the population does not increases median housing value (Null hypothesis).

```
t.test(data$population, data$median_house_value, alternative = c('greater'))
```

```
##
##  Welch Two Sample t-test
##
## data:  data$population and data$median_house_value
## t = -255.75, df = 20643, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -206751.6       Inf
## sample estimates:
```

By conducting the two sample t-test between the two variables, population and median house value, we observe that the p-value is 1, which is greater than our significance level $\alpha = 0.05$. Thus, we do not have enough convincing evidence to reject null hypothesis - "population does not increases median housing value."

If we visualize the relationship between the two variables, we can show as following:



By observing the plot above, we can see that population increase does not affect the median housing value since median housing value converges to around 150,000 as the population increases far beyond.

6 Models

Before fitting the model with each analyses, we first transformed our predictor variables in order to enhance the performance of our prediction model as following:

```
new_df <-
  transform(
    df,
    housing_median_age = df$housing_median_age,
    mean_rooms = (df$mean_rooms)^-0.3005615,
    mean_bedrooms = (df$mean_bedrooms)^-1.627276,
    population = (df$population)^0.2358979,
    households = (df$households)^0.2453846,
    median_income = (df$median_income)^0.09188503
  )
library(psych)
```

We created 5 different models for predicting median house value, using:

1. Multiple Linear Regression (MLR)
2. Linear Regression using Cross-Validation
3. LASSO and Ridge Regression
4. XGBoost
5. Random Forest

6.1 Multiple Linear Regression (MLR)

We found that there is a high correlation between mean_rooms predictor and mean_bedroom predictor, so we decided to analyze with only mean_bedrooms predictor in our model. We also tried multiple transformations regarding response variable to enhance our prediction, and we decided to use the log transformation for the multiple linear regression model.

Excluding mean_rooms predictor, we got the following prediction model

```
Call:
lm(formula = log(median_house_value) ~ ., data = new_data3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.35927 -0.22325 -0.02435  0.19513  2.66924

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.6821536  0.0617547  59.625  < 2e-16 ***
housing_median_age  0.0041223  0.0002163  19.061  < 2e-16 ***
households          0.0587831  0.0033778  17.403  < 2e-16 ***
median_income       7.0318857  0.0529567 132.786  < 2e-16 ***
mean_bedrooms      -0.2185191  0.0147347 -14.830  < 2e-16 ***
less_than_1H_OCEAN  0.4766533  0.0059581  80.001  < 2e-16 ***
ISLAND              1.1482510  0.1527109   7.519 5.74e-14 ***
NEAR_BAY            0.5079986  0.0090255  56.285  < 2e-16 ***
NEAR_OCEAN          0.5188346  0.0080971  64.077  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.341 on 20424 degrees of freedom
Multiple R-squared:  0.6412,    Adjusted R-squared:  0.6411
F-statistic:  4563 on 8 and 20424 DF,  p-value: < 2.2e-16
```

Where our R^2 value was around 0.64. Thus, we concluded that the R^2 value for the multiple linear regression model was around 0.64.

6.2 Linear Regression Using Cross-Validation

For this model, we found the prediction model as well as the relative variable importance for the response variable.

```
set.seed(1000)
data_with_rooms <- new_data2
data_with_bedrooms <- new_data3
# CV_data <- data_with_rooms
CV_data <- data_with_bedrooms
library(caret)
```

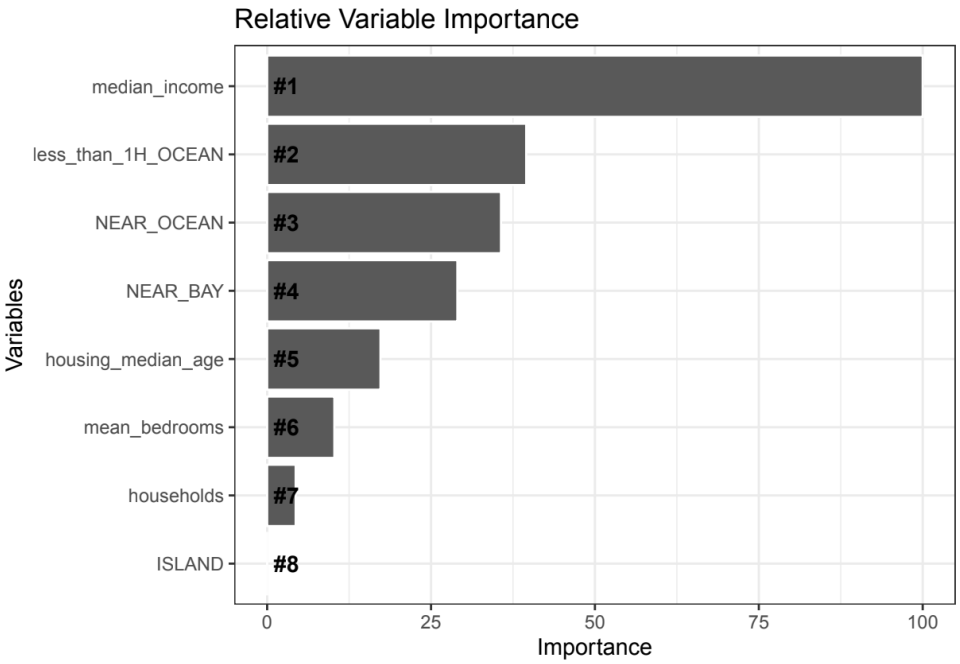
```
formula = median_house_value ~ .
fitControl <- trainControl(method="cv",number = 5)
HousingDataModel = train(formula, data = CV_data,
                  method = "lm",trControl = fitControl,metric="RMSE")
importance = varImp(HousingDataModel)


PlotImportance = function(importance)
{
  varImportance <- data.frame(Variables = row.names(importance[[1]]),
                              Importance = round(importance[[1]]$Overall,2))

  rankImportance <- varImportance %>%
    mutate(Rank = paste0('#',dense_rank(desc(Importance))))

  rankImportancefull = rankImportance
```

```
  ggplot(rankImportance, aes(x = reorder(Variables, Importance),
                             y = Importance)) +
    geom_bar(stat='identity',colour="white") +
    geom_text(aes(x = Variables, y = 1, label = Rank),
              hjust=0, vjust=.5, size = 4, colour = 'black',
              fontface = 'bold') +
    labs(x = 'Variables', title = 'Relative Variable Importance') +
    coord_flip() +
    theme_bw()
}
PlotImportance(importance)
```



Here we can observe that the relative variable importance respect to our response variable is ranked as above plot,

1. median_income
2. less_than_1H_OCEAN
3. Near_ OCEAN
4. NEAR_BAY
5. housing_median_age
6. mean_bedrooms
7. households
8. ISLAND

```
HousingDataModel
```

```
## Linear Regression
##
## 20433 samples
##     8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 16346, 16346, 16347, 16346, 16347
## Resampling results:
##
```

```
##   RMSE      Rsquared   MAE
##   75711.91  0.5698454  56810.93
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Here we found the R^2 value as 0.57, which is much lower than our transformed multiple linear regression model in previous section.

6.3 LASSO and Ridge Regression

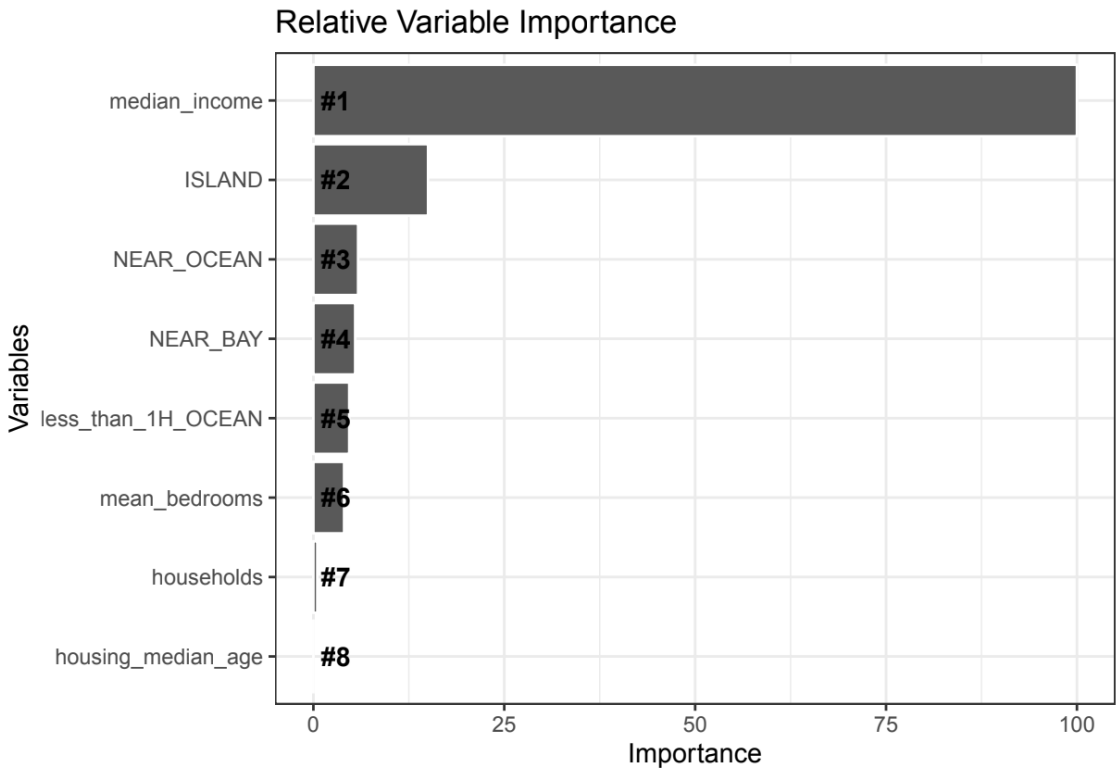
For this model, we found the best prediction model among multiple models that was created by the regression based on the alpha and lambda values. We also found the relative variable importance for the response variable.

```
set.seed(1001)
data_with_rooms <- new_data2
data_with_bedrooms <- new_data3
# RR_data <- data_with_rooms
RR_data <- data_with_bedrooms
library(caret)
formula2 = median_house_value ~ .
fitControl2 <- trainControl(method="cv",number = 5)
HousingDataModel2 = train(formula, data = RR_data,
                  method = "glmnet",trControl = fitControl2,metric="RMSE")
importance = varImp(HousingDataModel2)

PlotImportance = function(importance)
{
  varImportance <- data.frame(Variables = row.names(importance[[1]]),
                              Importance = round(importance[[1]]$Overall,2))
  rankImportance <- varImportance %>%
    mutate(Rank = paste0('#',dense_rank(desc(Importance))))
  rankImportancefull = rankImportance

  ggplot(rankImportance, aes(x = reorder(Variables, Importance),
                             y = Importance)) +
    geom_bar(stat='identity',colour="white") +
    geom_text(aes(x = Variables, y = 1, label = Rank),
              hjust=0, vjust=.5, size = 4, colour = 'black',
              fontface = 'bold') +
    labs(x = 'Variables', title = 'Relative Variable Importance') +
    coord_flip() +
    theme_bw()
}

PlotImportance(importance)
```



Here we can observe that the relative variable importance respect to our response variable is ranked as above plot,

1. median_income
2. ISLAND
3. Near_ OCEAN
4. NEAR_BAY
5. less_than_1H_OCEAN
6. mean_bedrooms
7. households
8. housing_median_age

```
HousingDataModel2
```

```
## glmnet
##
## 20433 samples
##     8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 16346, 16346, 16345, 16348, 16347
## Resampling results across tuning parameters:
##
##   alpha  lambda      RMSE      Rsquared   MAE
##   0.10     152.2623  75695.23  0.5701589  56805.33
##   0.10    1522.6228  75709.68  0.5701273  56811.24
##   0.10   15226.2285  76977.94  0.5669549  57916.28
##   0.55     152.2623  75695.53  0.5701503  56805.35
##   0.55    1522.6228  75764.38  0.5697732  56848.51
##   0.55   15226.2285  80557.45  0.5361319  61082.62
##   1.00     152.2623  75695.69  0.5701466  56805.44
##   1.00    1522.6228  75866.04  0.5689975  56924.94
##   1.00   15226.2285  86152.49  0.4662830  66117.65
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 0.1 and lambda = 152.2623.
```

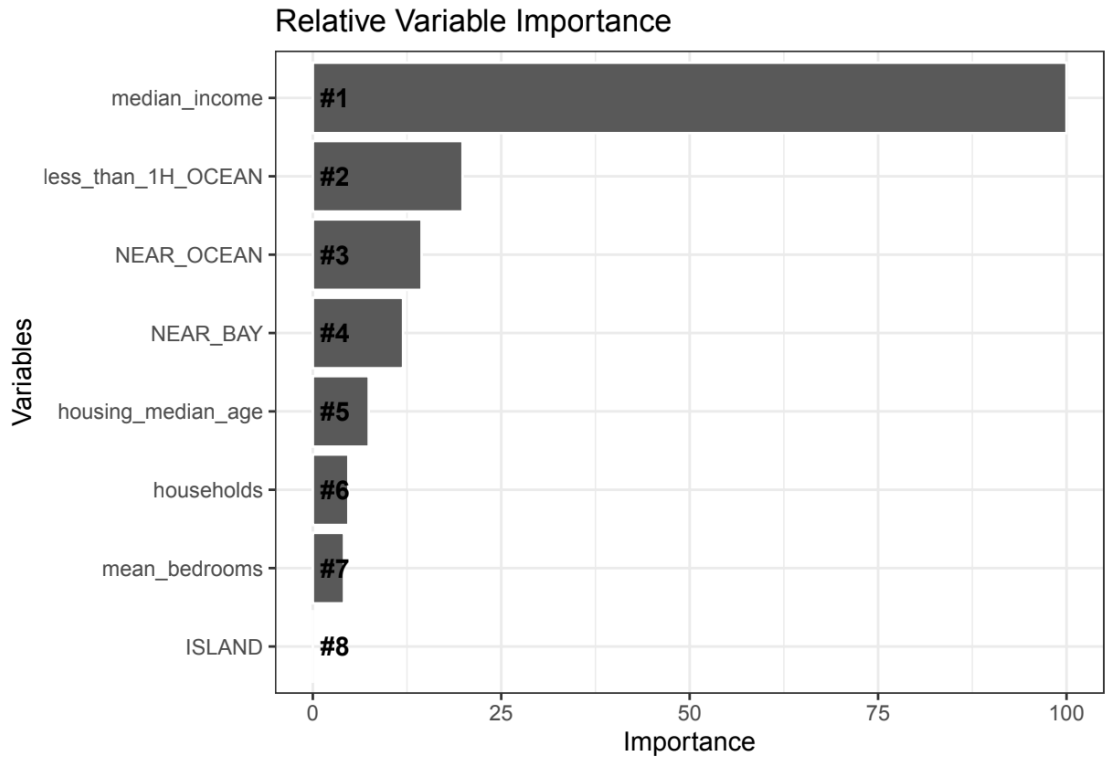
Here we can observe that there are three alpha values which has three lambda, RMSE, Rsquared, and MAE for each alpha. The final model that was chosen based on 9 different models created was where alpha = 0.1 and lambda = 152.2623. If we check the R^2 value of final model, it is shown as $R^2 = 0.57$ which is much lower than the transformed multiple linear regression model.

6.4 XGBoost

We took the same step as we were creating the prediction model with Cross-Validation except using XGBoost package - where XGBoost means Extreme Gradient Boosting.

```
set.seed(1000)
data_with_rooms <- new_data2
data_with_bedrooms <- new_data3
# XG_data <- data_with_rooms
XG_data <- data_with_bedrooms

xgbGrid <- expand.grid(nrounds = 500,
                       max_depth = 4,
                       eta = .05,
                       gamma = 0,
                       colsample_bytree = .5,
                       min_child_weight = 1,
                       subsample = 1)
formula = log(median_house_value) ~ .
fitControl <- trainControl(method="cv",number = 5)
HousingDataModelXGB = train(formula, data = XG_data,
                    method = "xgbTree",trControl = fitControl,
                    tuneGrid = xgbGrid,na.action = na.pass,metric="RMSE")
importance = varImp(HousingDataModelXGB)
PlotImportance(importance)
```



Here we can observe that the relative variable importance respect to our response variable is ranked as above plot,

1. median_income
2. less_than_1H_OCEAN
3. Near_ OCEAN
4. NEAR_BAY
5. housing_median_age
6. households

7. mean_bedrooms
8. ISLAND

```
HousingDataModelXGB
```

```
## eXtreme Gradient Boosting
##
## 20433 samples
##     8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 16346, 16346, 16347, 16346, 16347
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.3171485  0.6895498  0.2392251
##
## Tuning parameter 'nrounds' was held constant at a value of 500
## Tuning
##  held constant at a value of 1
## Tuning parameter 'subsample' was held
##  constant at a value of 1
```

We found the R^2 value as 0.69, which is so far the best R^2 value we found among all models created.

6.5 Random Forest

Using tidyverse and tidymodels packages, we performed the random forest prediction model as following, and instead of using mean_room variable, we used mean_bedroom variable for the prediction. We did not plot the relative importance variable, but we created a tibble using randomForestExplainer package as following:

```
measure_importance(final_rf2$fit) %>%
  as_tibble() %>%
  mutate(imp=node_purity_increase*100/max(node_purity_increase)) %>%
  arrange(-imp) %>%
  select(variable, imp) # Rank of important predictors
```

variable <fctr>	imp <dbl>
median_income	100.0000000
less_than_1H_OCEAN	33.4455998
NEAR_OCEAN	13.9732253
NEAR_BAY	10.6414993
housing_median_age	9.8548049
households	7.8801260
mean_bedrooms	7.1469972
ISLAND	0.2163063

Here we can observe that the relative variable importance respect to our response variable is ranked as above plot,

1. median_income
2. less_than_1H_OCEAN
3. Near_ OCEAN
4. NEAR_BAY
5. housing_median_age
6. households
7. mean_bedrooms
8. ISLAND

```
set.seed(1000)
library('tidyverse')
library('tidymodels')

data_with_rooms <- new_data2
data_with_bedrooms <- new_data3

# final_dat <- data_with_rooms
# final_dat$median_house_value = log(final_dat$median_house_value)

final_dat <- data_with_bedrooms
final_dat$median_house_value = log(final_dat$median_house_value)

final_dat -> final_pre
final_pre %>%
  initial_split(prop=0.8) -> final_split

final_split %>% training() %>%
  recipe(median_house_value~.) %>%
  prep() -> final_recipe

final_recipe %>%
  bake(final_split %>% testing()) -> final_testing

final_recipe %>%
  juice() -> final_training

rand_forest(trees=100, mode='regression') %>%
  set_engine('randomForest') %>%
  fit(median_house_value~., data=final_training) -> final_rf
```

```
library(yardstick)
final_rf %>%
  predict(final_testing) %>%
  bind_cols(final_testing) %>%
  metrics(truth=median_house_value, estimate=.pred)

final_pre %>%
  recipe(median_house_value~.) %>%
  prep() -> final_recipe2

final_recipe2 %>%
  bake(final_pre) -> final_testing_pre

final_recipe2 %>%
  juice() -> final_training_pre

rand_forest(trees=100, mode='regression') %>%
  set_engine('randomForest', localImp=TRUE) %>%
  fit(median_house_value~., data=final_training_pre) -> final_rf2

pp <- final_rf2 %>%
  predict(final_testing_pre) %>%
  bind_cols(final_pre)


# scatterplot(pp$median_house_value,pp$.pred)
# summary(lm(median_house_value~.pred,pp))
x <- pp$.pred
y <- pp$median_house_value
plot(x, y, main = "Predicted median_house_price VS Actual median_house_price",
     xlab = "Predicted median_house_price", ylab = "Actual median_house_price",
     pch = 19, frame = FALSE)
abline(lm(y ~ x, data = pp), col = "blue")
abline(a=0, b=1, col= "red")

# install.packages('randomForestExplainer')
library('randomForestExplainer')

measure_importance(final_rf2$fit)

measure_importance(final_rf2$fit) %>%
  as_tibble() %>%
  mutate(imp=node_purity_increase*100/max(node_purity_increase)) %>%
  arrange(-imp) %>%
  select(variable, imp) # Rank of important predictors
```

.metric <chr>	.estimator <chr>	.estimate <dbl>
rmse	standard	0.3427232
rsq	standard	0.6705464
mae	standard	0.2693697

We found that the random forest prediction model has a R^2 value of 0.67 which is similar (but less) than the XGBoost model. It was the second best prediction model we created.

7 Results

Based on 5 prediction models we created using different methods introduced in section 5, the R^2 of each prediction model states as following:

1. Transformed Multiple Linear Regression Model:
   $R^2 = 0.65$
2. Linear Regression Using Cross-Validation model:
   $R^2 = 0.57$
3. LASSO and Ridge Regression model:
   $R^2 = 0.57$
4. XGBoost model:
   $R^2 = 0.69$
5. Random Forest model:
   $R^2 = 0.67$

We have found that the XGBoost prediction model performs the best among all prediction model with the highest R^2 value as 0.69.

8 Conclusion

As far as our conclusion, we rejected two out of the three hypotheses, with ocean proximity and housing median age being significant in their effect to median housing price and not population. These conclusions also agree with our best model XGBoost, where we found that ocean proximity was amongst the the top 4 most important variables, with median income being number 1, and median housing age being number 5.

To briefly mention limitations, some of the limitations we encountered in this experiment were, for instance, that our predictive models did not have an extreme high accuracy; however, this could be due to either over-fitting or the variability in the data itself. For future research, we would like to use a different data set to compare our results. Another interesting future research could also be do a classification model that can detect whether a person will be accepted for a home purchase or not.

A takeaway from these results, is that people that earn more money tend to own property that is more expensive, and this makes sense because they have more freedom in how much they are able to spend on their mortgage. However, with the astronomical home prices in California, it is highlighted that to be able to own a home, then future interested home buyers must ensure that they are earning a high income.

The second feature home buyers should keep in mind is the home's proximity to the ocean. If the home is less than an hour away from the ocean, near the ocean or near the bay, then these factors also affect median housing price.

Lastly, buyers should also keep in mind the age of the home because it also one of the predictors affecting median housing price.

To conclude, we believe that if buyers keep these factors in mind when purchasing a home, they will have better fulfilled expectations on the types of homes that they will be able to ultimately purchase.

References

[1] Biermeier, Deane. "15 States With the Highest Average Home Prices." Forbes Home, 26 Sept. 2022, www.forbes.com/home-improvement/features/states-with-highest-home-prices.
https://www.forbes.com/home-improvement/features/states-with-highest-home-prices/

[2] Desjardins, Jeff. "Animation: The 20 Largest State Economies by GDP in the Last 50 Years." Visual Capitalist, 22 Aug. 2019, www.visualcapitalist.com/animation-the-20-largest-state-economies-by-gdp-in-the-last-50-years.
https://www.visualcapitalist.com/animation-the-20-largest-state-economies-by-gdp-in-the-last-50-years

[3] Gomez, Joe. 8 Critical Factors That Influence a Home's Value | Opendoor. 30 Nov. 2022, www.opendoor.com/articles/factors-that-influence-home-value.
https://www.opendoor.com/articles/factors-that-influence-home-value

[4] Ross, Jenna. "Mapped: The Growth in U.S. House Prices by State." Advisor Channel, 17 Oct. 2022, advisor.visualcapitalist.com/growth-in-u-s-house-prices-by-state.
https://advisor.visualcapitalist.com/growth-in-u-s-house-prices-by-state/