Alaa El Hajjar, Momo Ogawa, Momtahin Masud
Professor Chung Eun Lee
STA9750/OPR9750 - Software Tools for Data Analysis
December 16, 2024

Final Project
California Housing Price Analysis

1. Introduction

The housing market in California is a dynamic and vital part of the state's economy, influenced by a variety of social, economic, and demographic factors. This analysis aims to address a real-world problem by examining the factors that influence housing prices in California, with a particular focus on predicting the *median_house_value* using a robust data-driven approach.

To achieve this, we began by using the California housing dataset from Kaggle (https://www.kaggle.com/datasets/camnugent/california-housing-prices). The data underwent comprehensive cleaning, including the removal of rows with missing values and the creation of new variables—such as the bedroom-to-room ratio, rooms-per-household, and income-to-rooms ratio—to enhance analytical insights. Subsequent analyses included association analysis to explore the relationships between key variables and regression analysis to build predictive models for housing prices.
The findings of this study are intended to provide actionable insights for stakeholders in the housing industry by shedding light on the key factors that drive housing prices in California.

2. Data Cleaning

We used the command **"str(df)"** to display the structure of our data frame. We can see below, the dataset has 20640 rows and 10 columns, with 9 columns being numerical and 1 column being categorical.

```
'data.frame':   20640 obs. of  10 variables:
 $ longitude         : num  -122 -122 -122 -122 -122 ...
 $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
 $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
 $ total_rooms       : num  880 7099 1467 1274 1627 ...
 $ total_bedrooms    : num  129 1106 190 235 280 ...
 $ population        : num  322 2401 496 558 565 ...
 $ households        : num  126 1138 177 219 259 ...
 $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
 $ median_house_value: num  452600 358500 352100 341300 342200 ...
 $ ocean_proximity   : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
```

We ran the command **"summary(df)"** to find the 5 number summary for each variable and spot any missing values. The column *total_bedroom* has 207 missing values as shown below:

```
> summary(df)
   longitude         latitude     housing_median_age  total_rooms    total_bedrooms     population      households
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2   Min.   :   1.0   Min.   :    3   Min.   :   1.0
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448   1st Qu.: 296.0   1st Qu.:  787   1st Qu.: 280.0
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127   Median : 435.0   Median : 1166   Median : 409.0
 Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636   Mean   : 537.9   Mean   : 1425   Mean   : 499.5
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148   3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320   Max.   :6445.0   Max.   :35682   Max.   :6082.0
                                                                    NA's   :207

 median_income     median_house_value ocean_proximity
 Min.   : 0.4999   Min.   : 14999     Length:20640
 1st Qu.: 2.5634   1st Qu.:119600     Class :character
 Median : 3.5348   Median :179700     Mode  :character
 Mean   : 3.8707   Mean   :206856
 3rd Qu.: 4.7432   3rd Qu.:264725
 Max.   :15.0001   Max.   :500001
```

In our next step we dropped the rows with null values using the code **"df[!is.na(df$total_bedrooms),]"** and we assigned it to a new variable called "cleaned_df"

```
# Drop rows with missing values in 'total_bedrooms'
cleaned_df <- df[!is.na(df$total_bedrooms), ]
```

We then decided to create 3 new variables on the cleansed data frame with no missing value through combining existing variables in the dataset to enhance our insight and provide us with more in depth and interesting findings about the data. The variables we created were: ***Bedroom-to-room ratio***, which gives us a sense of housing quality. ***Rooms-per-household***, reflecting housing density. ***Income-to-rooms ratio***, which links economic status to housing availability.

```
# Create new variables based on the cleaned data
cleaned_df$bedroom_room_ratio <- cleaned_df$total_bedrooms / cleaned_df$total_rooms
cleaned_df$rooms_per_household <- cleaned_df$total_rooms / cleaned_df$households
cleaned_df$income_rooms_ratio <- cleaned_df$median_income / cleaned_df$rooms_per_household
```

In our next step, we decided to assign a clean dataset (cleaned_df) to our final dataset called "housing". This attaches the 3 new variables that we just created to our final dataset. The code is shown below:

```
housing <- cleaned_df[, c("longitude", "latitude", "housing_median_age", "total_rooms",
                          "total_bedrooms", "population", "households", "median_income",
                          "bedroom_room_ratio", "rooms_per_household", "income_rooms_ratio",
                          "ocean_proximity", "median_house_value")]
```

In our final step of data cleaning, we decided to run the command **"summary(housing)"** on our cleansed and final dataset to see the 5 number summary along with the 3 new variables we just created that is ready to go through association and regression analysis.

```
   longitude         latitude     housing_median_age  total_rooms
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1450
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127
 Mean   :-119.6   Mean   :35.63   Mean   :28.63      Mean   : 2636
 3rd Qu.:-118.0   3rd Qu.:37.72   3rd Qu.:37.00      3rd Qu.: 3143
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320
 total_bedrooms     population      households     median_income
 Min.   :   1.0   Min.   :    3   Min.   :   1.0   Min.   : 0.4999
 1st Qu.: 296.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5637
 Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5365
 Mean   : 537.9   Mean   : 1425   Mean   : 499.4   Mean   : 3.8712
 3rd Qu.: 647.0   3rd Qu.: 1722   3rd Qu.: 604.0   3rd Qu.: 4.7440
 Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001
 bedroom_room_ratio rooms_per_household income_rooms_ratio ocean_proximity
 Min.   :0.1000     Min.   :  0.8461    Min.   :0.01321    Length:20433
 1st Qu.:0.1754     1st Qu.:  4.4414    1st Qu.:0.54267    Class :character
 Median :0.2032     Median :  5.2308    Median :0.70775    Mode  :character
 Mean   :0.2130     Mean   :  5.4313    Mean   :0.71540
 3rd Qu.:0.2398     3rd Qu.:  6.0524    3rd Qu.:0.86165
 Max.   :1.0000     Max.   :141.9091    Max.   :5.16803
 median_house_value
 Min.   : 14999
 1st Qu.:119500
 Median :179700
 Mean   :206864
 3rd Qu.:264700
 Max.   :500001
```

3. Association Analysis

This section explores associations between *median_house_value* (y-variable) and explanatory variables to guide the construction of a robust regression model. Visualizations and statistical tests reveal patterns to enhance predictive accuracy.

3-1. Exploring Variable Associations (without Rooms per Household)

First, we analyzed correlations to identify numerical variables most closely associated with *median_house_value* (fig. 1) and selected the top 5 for further analysis.

var1	var2	correlation	pval
<chr>	<chr>	<dbl>	<dbl>
median_income	median_house_value	0.68835548	0.000000e+00
income_rooms_ratio	median_house_value	0.66497497	0.000000e+00
bedroom_room_ratio	median_house_value	-0.25588015	8.160022e-303
rooms_per_household	median_house_value	0.15134408	5.822073e-105
latitude	median_house_value	-0.14463821	6.132893e-96

fig.1

We visualized and tested (permutation number: 500) associations between the selected variables and housing prices. (Due to page limits, results for *rooms_per_household* are in the code and slides. A p-value of 0 (<0.05) shows a significant and conclusive relationship in Spearman's test).

3-2. Median Income and Median House Value

Areas with higher income levels tend to have higher housing prices. Since the point cloud is heteroscedastic and there are outliers, we used Spearman's test (fig. 2, 3). A p-value of essentially 0, below 0.05, indicates the relationship is statistically significant and conclusive (fig. 4).

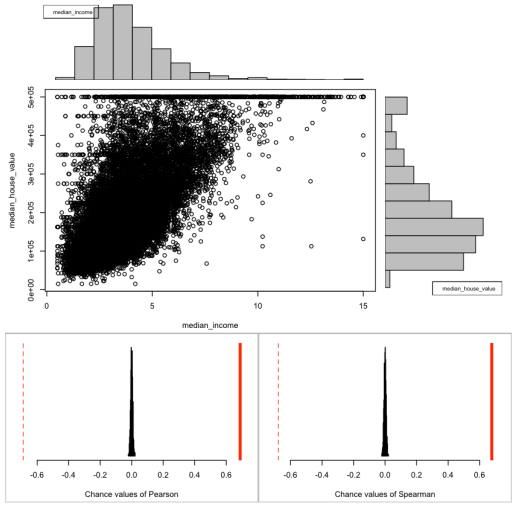

fig.2

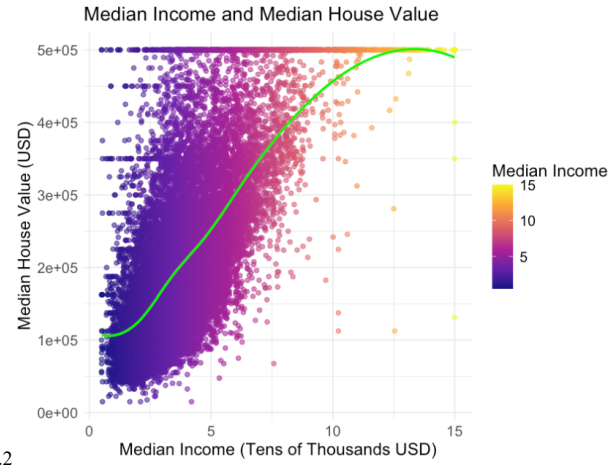

fig.3

```
Association between median_income (numerical) and  median_house_value (numerical)
 using 20433 complete cases
Permutation procedure:
                                Value Estimated p-value
Pearson's r                0.6883555                 0
Spearman's rank correlation 0.6771076                0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
```

fig.4

3-3. Income-Rooms Ratio and Median House Value

Areas with a higher Income-Rooms Ratio—indicating higher income or fewer rooms per household—tend to have higher house values. Due to heteroscedasticity and outliers, we used Spearman's test (fig. 5, 6). A p-value of essentially 0, below 0.05, confirms the relationship is statistically significant and conclusive (fig. 7).

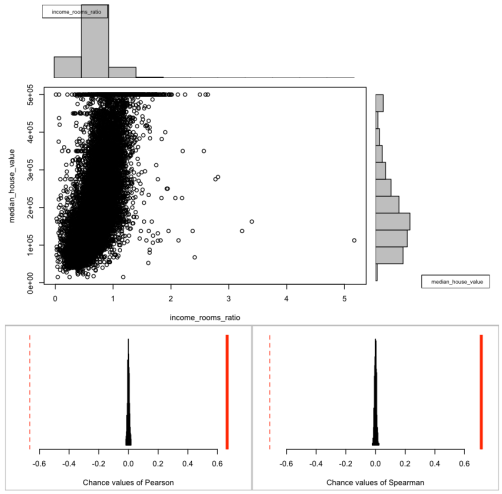

fig.5

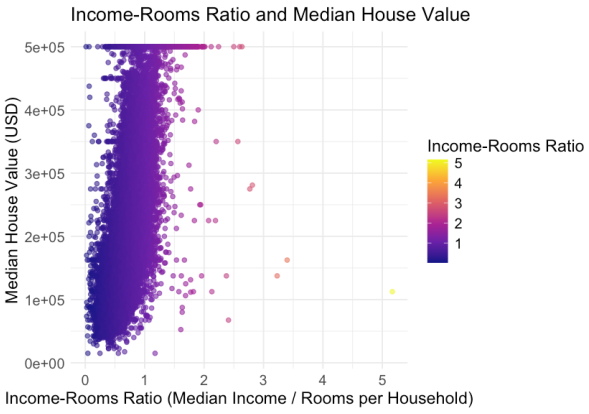

fig.6

```
Association between income_rooms_ratio (numerical) and  median_house_value (numerical)
 using 20433 complete cases
Permutation procedure:
                                 Value Estimated p-value
Pearson's r                   0.664975                 0
Spearman's rank correlation 0.712601                   0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
```

fig.7

3-4. Ratio of Bedrooms to Total Rooms and Median House Value

Homes with fewer bedrooms relative to total rooms tend to have higher values (fig. 9). Due to heteroscedasticity and outliers, we used Spearman's test (fig. 8). A p-value of essentially 0, below 0.05, confirms the relationship is statistically significant and conclusive (fig. 10).

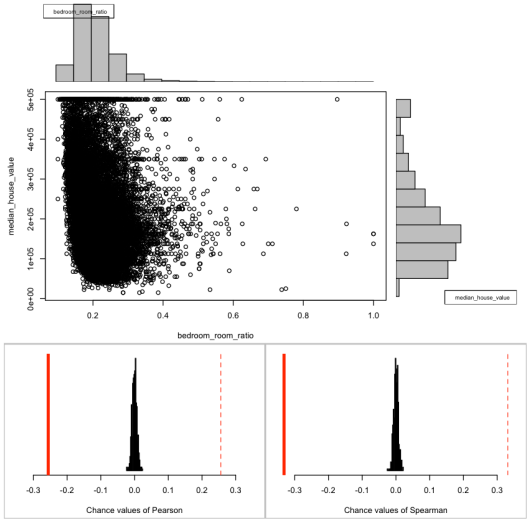

fig.8

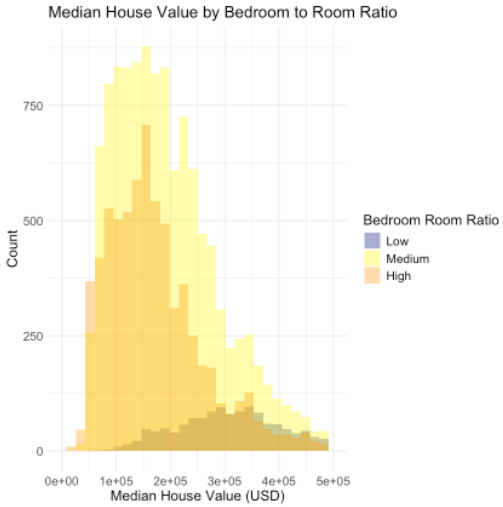

fig.9

```
Association between bedroom_room_ratio (numerical) and  median_house_value (numerical)
 using 20433 complete cases
Permutation procedure:
                                 Value Estimated p-value
Pearson's r                  -0.2558801                 0
Spearman's rank correlation -0.3316925                  0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
```

fig.10

3-5. Latitude and Median House Value

Higher housing values are observed in central and some coastal areas (fig. 12). To avoid multicollinearity, we conducted an association test using only latitude, excluding longitude (fig. 11).

Due to heteroscedasticity and outliers, we used Spearman's test. A p-value of essentially 0, below 0.05, confirms the relationship is statistically significant and conclusive (fig. 13).

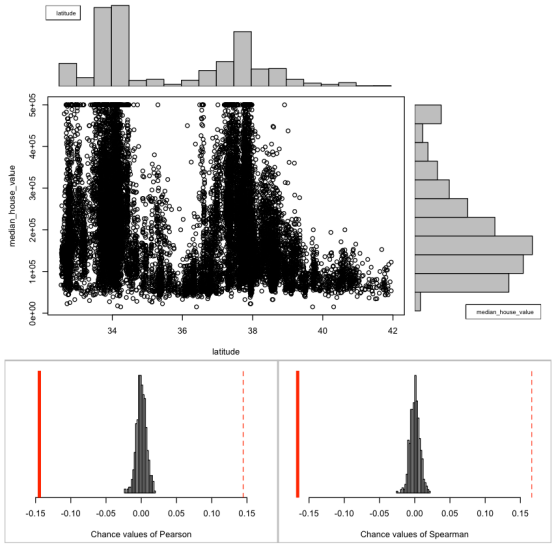

fig.11

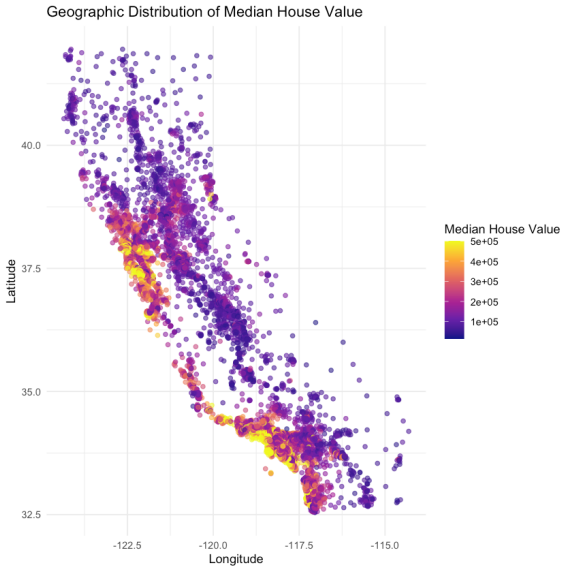

fig.12

```
Association between latitude (numerical) and  median_house_value (numerical)
 using 20433 complete cases
Permutation procedure:
                                Value Estimated p-value
Pearson's r                -0.1446382                 0
Spearman's rank correlation -0.1661199                 0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
```

fig.13

3-6: Ocean Proximity and Median House Value

We visualized and tested the categorical variable *ocean_proximity*. Housing prices differ significantly by proximity to the ocean, with island properties having the highest median prices (fig. 14). Due to skewed distributions and outliers, we used a median test (fig. 15). A p-value of 0, below 0.05, confirms the relationship is statistically significant and conclusive (fig. 16).

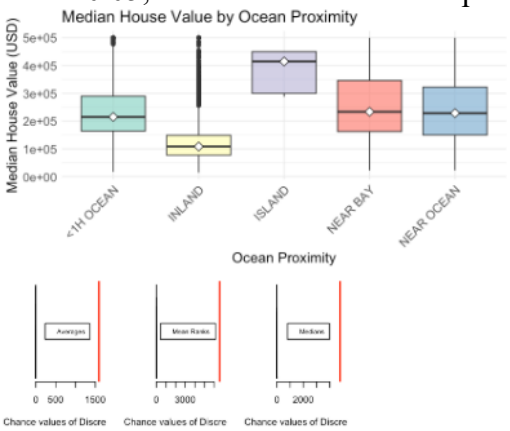

fig.14

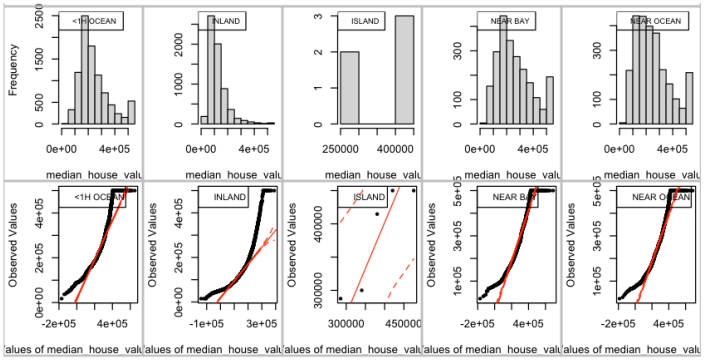

fig.15

```
Association between ocean_proximity (categorical) and  median_house_value (numerical)
 using 20433 complete cases

Sample Sizesx
 <1H OCEAN      INLAND      ISLAND    NEAR BAY NEAR OCEAN
      9034        6496           5        2270        2628

Permutation procedure:
                  <1H OCEAN INLAND ISLAND NEAR BAY NEAR OCEAN Discrepancy
Averages (ANOVA)     240268 124897 380440   259279     249042        1595
Mean Ranks (Kruskal)  10280   9950  11606     9902      10930        6565
Medians              215000 108700 414700   233800     228750        4788
```

```
                    Estimated p-value
Averages (ANOVA)                    0
Mean Ranks (Kruskal)                0
Medians                             0
With 500 permutations, we are 95% confident that
 the p-value of ANOVA (means) is between 0 and 0.007
 the p-value of Kruskal-Wallis (ranks) is between 0 and 0.007
 the p-value of median test is between 0 and 0.007
```

fig.16

3-7. Handling Multicollinearity

Before proceeding to regression modeling, we addressed multicollinearity. To avoid redundancy, we will select only one variable from strongly associated pairs (fig. 17, 18). This approach will help keep the model strong and reliable.

	var1	var2	correlation	pval
	<chr>	<chr>	<dbl>	<dbl>
1	total_bedrooms	households	0.9797283	0
2	total_rooms	total_bedrooms	0.9303795	0
3	longitude	latitude	-0.9246161	0
4	total_rooms	households	0.9189915	0
5	population	households	0.9071859	0
6	total_bedrooms	population	0.8777467	0
7	total_rooms	population	0.8572813	0
8	median_income	income_rooms_ratio	0.8195145	0

fig.17

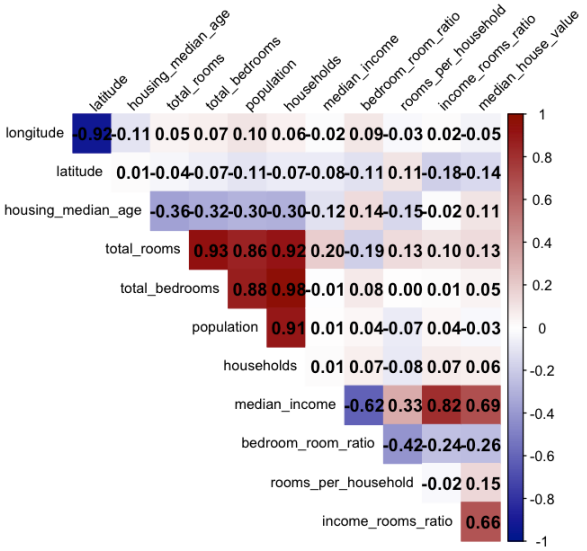

fig.18

4. Regression Model

In order to predict *median_house_value*, multiple linear regression analyses will be performed. Model metrics will be analyzed in order to find the best fit model.

4-1. Baseline Regression Model (all Independent Variables Included)

```
Residuals:
   Min     1Q Median     3Q    Max
-575550 -41675 -10166  28347 817928

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.457e+06  8.817e+04 -27.871  < 2e-16 ***
longitude               -2.800e+04  1.020e+03 -27.445  < 2e-16 ***
latitude                -2.650e+04  1.008e+03 -26.294  < 2e-16 ***
housing_median_age       1.085e+03  4.342e+01  25.001  < 2e-16 ***
total_rooms              1.781e+00  9.431e-01   1.888  0.05898 .
total_bedrooms           1.458e+01  7.985e+00   1.826  0.06786 .
population              -4.032e+01  1.071e+00 -37.629  < 2e-16 ***
households               1.060e+02  8.562e+00  12.377  < 2e-16 ***
median_income            4.338e+04  7.883e+02  55.025  < 2e-16 ***
bedroom_room_ratio       2.928e+05  1.502e+04  19.491  < 2e-16 ***
rooms_per_household      2.477e+03  2.600e+02   9.529  < 2e-16 ***
income_rooms_ratio      -1.474e+04  4.913e+03  -2.999  0.00271 **
ocean_proximityINLAND   -3.503e+04  1.745e+03 -20.073  < 2e-16 ***
ocean_proximityISLAND    1.454e+05  3.041e+04   4.782 1.75e-06 ***
ocean_proximityNEAR BAY -4.304e+03  1.893e+03  -2.274  0.02298 *
ocean_proximityNEAR OCEAN 3.852e+03  1.555e+03   2.476  0.01329 *

Residual standard error: 67900 on 20417 degrees of freedom
Multiple R-squared:  0.6542,    Adjusted R-squared:
0.654
F-statistic:  2575 on 15 and 20417 DF,  p-value: < 2.2e-16
```

fig.19

Fig.19 shows the output of model1, baseline model, where F-statistics= 2575, RSE=67900, Adjusted R-square= 0.654 and p-value < 2.2e-16. P-value <0 showing significance. Also ***, **,* shows the order of significance for the independent variables in the baseline model. '***' has the highest significance and the lack of '*' shows no significance for that independent variable.

4-1-1. Multicollinearity using VIF

To enhance and find the best linear regression model to predict *median_house_value,* we will be checking for multicollinearity and addressing it using VIF. VIF or Variation Inflation Factor measures the degree of multicollinearity in a regression analysis. A higher VIF indicates stronger multicollinearity. In order to calculate VIF, "car" package was installed.

After running VIF on our baseline model, fig.20 shows that *longitude, latitude, total_rooms, total_bedrooms, households* have an GVIF value>10, therefore indicating severe multicollinearity. Furthermore, *income_rooms_rate* and *median_income*, showing high multicollinearity. To see also in detail what variables are highly correlated to each other, we can refer to fig.17**.**

	GVIF	Df	GVIF^(1/(2*Df))
longitude	18.508830	1	4.302189
latitude	20.548136	1	4.533005
housing_median_age	1.324424	1	1.150836
total_rooms	18.821448	1	4.338369
total_bedrooms	50.162354	1	7.082539
population	6.533316	1	2.556035
households	47.479365	1	6.890527
median_income	9.933538	1	3.151752
bedroom_room_ratio	3.362344	1	1.833670
rooms_per_household	1.846450	1	1.358841
income_rooms_ratio	6.981029	1	2.642164
ocean_proximity	4.219021	4	1.197158

fig.20

4-2. Model 2 (income_rooms_ratio Excluded)

For the next model since *median_income* & *income_rooms_ratio* are highly correlated (fig. 17), one of the variables will be chosen. Since *median_income* explains slightly more variance (correlation= 0.688) than *income_rooms_ratio* (correlation=0.665), *median_income* will be prioritized, and *income_rooms_ratio* will be dropped.

Residuals:

Min	1Q	Median	3Q	Max
-569697	-41662	-10261	28478	815798

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.440e+06	8.799e+04	-27.727	< 2e-16	***
longitude	-2.778e+04	1.018e+03	-27.297	< 2e-16	***
latitude	-2.628e+04	1.005e+03	-26.137	< 2e-16	***
housing_median_age	1.086e+03	4.343e+01	25.003	< 2e-16	***
total_rooms	2.188e+00	9.335e-01	2.344	0.01908	*
total_bedrooms	1.851e+01	7.878e+00	2.349	0.01881	*
population	-4.011e+01	1.069e+00	-37.506	< 2e-16	***
households	9.882e+01	8.225e+00	12.015	< 2e-16	***
median_income	4.129e+04	3.706e+02	111.426	< 2e-16	***
bedroom_room_ratio	2.742e+05	1.368e+04	20.044	< 2e-16	***
rooms_per_household	2.693e+03	2.499e+02	10.776	< 2e-16	***
ocean_proximityINLAND	-3.461e+04	1.740e+03	-19.891	< 2e-16	***
ocean_proximityISLAND	1.473e+05	3.041e+04	4.844	1.28e-06	***
ocean_proximityNEAR BAY	-4.213e+03	1.893e+03	-2.226	0.02603	*
ocean_proximityNEAR OCEAN	4.113e+03	1.553e+03	2.648	0.00811	**

Residual standard error: 67920 on 20418 degrees of freedom
Multiple R-squared: 0.6541, Adjusted R-squared: 0.6538
F-statistic: 2758 on 14 and 20418 DF, p-value: < 2.2e-16

fig.21

The output, fig.21, shows F-statistics= 2758, RSE=67920, Adjusted R-square= 0.6538 and p-value < 2.2e-16. P-value <0 showing significance. F-statistics has increased from model 1.

	GVIF	Df	GVIF^(1/(2*Df))
longitude	18.419721	1	4.291820
latitude	20.432726	1	4.520257
housing_median_age	1.324416	1	1.150833
total_rooms	18.431606	1	4.293205
total_bedrooms	48.812569	1	6.986599
population	6.505323	1	2.550553
households	43.797607	1	6.617976
median_income	2.194074	1	1.481241
bedroom_room_ratio	2.786144	1	1.669175
rooms_per_household	1.705330	1	1.305883
ocean_proximity	4.178679	4	1.195721

fig.22

As seen in fig.22, Model 2 shows improvement, as *median_income* is no longer correlated to other variables in the model, GVIF<5.

4-3. Model 3 (longitude Excluded)

For the next model since *longitude* & *latitude* are highly correlated (fig. 17), one of the variables will be chosen. Since *latitude* explains slightly more variance (correlation: -0.145) than *longitude* (correlation: -0.045), *latitude* will be prioritized, and longitude will be dropped.

```
Residuals:
    Min      1Q  Median      3Q     Max
-577891  -41999  -10139   28433  778473

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -56029.299  10998.147  -5.094 3.53e-07 ***
latitude                   111.353    281.014   0.396   0.6919
housing_median_age        1196.329     44.018  27.178  < 2e-16 ***
total_rooms                  1.479      0.950   1.557   0.1196
total_bedrooms               6.740      8.008   0.842   0.4000
population                 -39.078      1.088 -35.915  < 2e-16 ***
households                 114.089      8.354  13.656  < 2e-16 ***
median_income            42834.246    372.835 114.888  < 2e-16 ***
bedroom_room_ratio      273499.434  13925.267  19.641  < 2e-16 ***
rooms_per_household       1797.797    252.212   7.128 1.05e-12 ***
ocean_proximityINLAND    -63815.079   1396.737 -45.689  < 2e-16 ***
ocean_proximityISLAND    168958.035  30950.764   5.459 4.85e-08 ***
ocean_proximityNEAR BAY     4569.961   1898.915   2.407   0.0161 *
ocean_proximityNEAR OCEAN  13536.685   1541.813   8.780  < 2e-16 ***

Residual standard error: 69140 on 20419 degrees of freedom
Multiple R-squared:  0.6414,    Adjusted R-squared:
0.6412
F-statistic:  2810 on 13 and 20419 DF,  p-value: < 2.2e-16
```

fig.23

Fig. 23, shows an improved F-statistic (higher F-statistic than model 2), a bit lower Adjusted r-square (0.6412), and p-value<0, showing significance.

	GVIF	Df	GVIF^(1/(2*Df))
latitude	1.540278	1	1.241079
housing_median_age	1.312903	1	1.145820
total_rooms	18.417317	1	4.291540
total_bedrooms	48.666376	1	6.976129
population	6.497189	1	2.548958
households	43.595135	1	6.602661
median_income	2.142962	1	1.463886
bedroom_room_ratio	2.786135	1	1.669172
rooms_per_household	1.675964	1	1.294590
ocean_proximity	2.059009	4	1.094479

fig.24

As seen in fig.24, Model 3 showed improvement, as latitude is no longer correlated to other variables in the model, GVIF<5.

For the next models and in order to choose which variables to include, we will also be referring to fig.17. GVIF for *total_rooms*, *total_bedrooms* and *households* show severe association among themselves. To choose which variables to keep, different variations will be analyzed in the next section and model performance metrics will be taken into account.

4-4. Model 4 (households Excluded)

In this model, *households* will be removed first, as it has a high GVIF.

```
Residuals:
    Min      1Q  Median      3Q     Max
-584510  -42191  -10101   28324  665408

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -5.488e+04  1.105e+04  -4.968 6.84e-07 ***
latitude                   4.603e+02  2.811e+02   1.638   0.1015
housing_median_age         1.212e+03  4.420e+01  27.413  < 2e-16 ***
total_rooms                9.530e-01  9.535e-01   1.000   0.3176
total_bedrooms             9.522e+01  4.728e+00  20.141  < 2e-16 ***
population                -3.226e+01  9.711e-01 -33.219  < 2e-16 ***
median_income              4.344e+04  3.719e+02 116.813  < 2e-16 ***
bedroom_room_ratio         2.437e+05  1.382e+04  17.640  < 2e-16 ***
rooms_per_household        2.402e+02  2.260e+02   1.063   0.2877
ocean_proximityINLAND     -6.550e+04  1.398e+03 -46.863  < 2e-16 ***
ocean_proximityISLAND      1.614e+05  3.109e+04   5.193 2.09e-07 ***
ocean_proximityNEAR BAY    4.638e+03  1.908e+03   2.431   0.0151 *
ocean_proximityNEAR OCEAN  1.370e+04  1.549e+03   8.848  < 2e-16 ***

Residual standard error: 69460 on 20420 degrees of freedom
Multiple R-squared:  0.6382,    Adjusted R-squared:
0.638
F-statistic:  3001 on 12 and 20420 DF,  p-value: < 2.2e-16
```

fig.25

Fig. 25 shows Improved F-statistics (3001), a slightly lower Adjusted r-square, and p-value<0

	GVIF	Df	GVIF^(1/(2*Df))
latitude	1.527540	1	1.235937
housing_median_age	1.312039	1	1.145443
total_rooms	18.387082	1	4.288016
total_bedrooms	16.809261	1	4.099910
population	5.128618	1	2.264645
median_income	2.112696	1	1.453511
bedroom_room_ratio	2.717783	1	1.648570
rooms_per_household	1.333222	1	1.154653
ocean_proximity	2.037529	4	1.093045

fig.26

As seen in fig.26, Model 4 showed improvement, as *total_bedroom* GVIF decreased significantly. *Household* will be removed.

4-5. Model 5 (total_bedrooms removed and total_rooms kept)

```
Residuals:
    Min      1Q  Median      3Q     Max
-563363  -42862  -10705   28989  576112

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -7.576e+04  1.111e+04  -6.821 9.31e-12 ***
latitude                  4.510e+02  2.839e+02   1.589  0.11214
housing_median_age        1.174e+03  4.460e+01  26.334  < 2e-16 ***
total_rooms               1.720e+01  5.130e-01  33.539  < 2e-16 ***
population               -2.841e+01  9.615e-01 -29.549  < 2e-16 ***
median_income             4.214e+04  3.698e+02 113.941  < 2e-16 ***
bedroom_room_ratio        3.875e+05  1.194e+04  32.444  < 2e-16 ***
rooms_per_household       2.291e+02  2.282e+02   1.004  0.31536
ocean_proximityINLAND    -6.710e+04  1.409e+03 -47.623  < 2e-16 ***
ocean_proximityISLAND     1.604e+05  3.139e+04   5.110 3.25e-07 ***
ocean_proximityNEAR BAY   5.859e+03  1.925e+03   3.043  0.00234 **
ocean_proximityNEAR OCEAN 1.380e+04  1.564e+03   8.822  < 2e-16 ***
---
```

Residual standard error: 70140 on 20421 degrees of freedom
Multiple R-squared: 0.631, Adjusted R-squared: 0.6308
F-statistic: 3174 on 11 and 20421 DF, p-value: < 2.2e-16

fig.27

Fig. 27 shows Improved F-statistics (3174), a slightly lower Adjusted r-square, and p-value<0

	GVIF	Df	GVIF^(1/(2*Df))
latitude	1.527536	1	1.235935
housing_median_age	1.309742	1	1.144440
total_rooms	5.218778	1	2.284465
population	4.930298	1	2.220427
median_income	2.048751	1	1.431346
bedroom_room_ratio	1.991875	1	1.411338
rooms_per_household	1.333215	1	1.154649
ocean_proximity	2.023799	4	1.092121

fig.28

As seen in fig.28, *total_rooms* GVIF decreased significantly, but still slightly above 5 so in the next model total_rooms will be removed and *total_bedrooms* will be kept.

4-6. Model 6 (total_bedrooms kept and total_rooms removed)

```
Residuals:
    Min      1Q  Median      3Q     Max
-585806  -42146  -10081   28360  660014

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -5.404e+04  1.102e+04  -4.906 9.39e-07 ***
latitude                  4.648e+02  2.811e+02   1.654   0.0982 .
housing_median_age        1.210e+03  4.416e+01  27.396  < 2e-16 ***
total_bedrooms            9.922e+01  2.519e+00  39.393  < 2e-16 ***
population               -3.197e+01  9.287e-01 -34.428  < 2e-16 ***
median_income             4.352e+04  3.622e+02 120.150  < 2e-16 ***
bedroom_room_ratio        2.366e+05  1.187e+04  19.944  < 2e-16 ***
rooms_per_household       2.660e+02  2.245e+02   1.185   0.2361
ocean_proximityINLAND    -6.538e+04  1.393e+03 -46.943  < 2e-16 ***
ocean_proximityISLAND     1.617e+05  3.108e+04   5.202 1.99e-07 ***
ocean_proximityNEAR BAY   4.689e+03  1.907e+03   2.459   0.0139 *
ocean_proximityNEAR OCEAN 1.377e+04  1.547e+03   8.896  < 2e-16 ***
```

Residual standard error: 69460 on 20421 degrees of freedom
Multiple R-squared: 0.6382, Adjusted R-squared: 0.638
F-statistic: 3274 on 11 and 20421 DF, p-value: < 2.2e-16

fig.29

Fig. 29 shows that the removal of *total_rooms* and keeping *total_bedrooms*, improved F-statistics, higher Adjusted r-square and p-value<0

	GVIF	Df	GVIF^(1/(2*Df))
latitude	1.527147	1	1.235778
housing_median_age	1.309706	1	1.144424
total_bedrooms	4.770948	1	2.184250
population	4.691119	1	2.165899
median_income	2.004716	1	1.415880
bedroom_room_ratio	2.004574	1	1.415830
rooms_per_household	1.315892	1	1.147123
ocean_proximity	2.022476	4	1.092032

fig.30

Fig.30 shows that *total_bedrooms* GVIF is under 5. Thus, removing *total_rooms* and keeping *total_bedrooms* is better.

4-7. Model 6 as a best Fit Model

Model 6 has the best F-statistics = 3274, which is the highest value among all the other models, and a p-value of < 2.2e-16 meaning that the model is overall meaningful and statistically significant. Even though Model 6 has an adjusted r-square =0.638, which is slightly lower than other models and an RSE= 69460, which is higher than the first 3 models. Model 6 takes into account multicollinearity which avoids potential instability in the coefficients which is essential. Therefore, we believe model 6 is the best fit for predicting *median_house_value*.

5. Other Techniques

1. Multicollinearity Analysis: Pairwise correlations were visualized using corrplot(), and multicollinearity was assessed with vif() (car package).

2. Log Transformation: Log transformations were tested on *median_house_value* and key predictors to address non-linear associations. However, since log transformations did not result in clear linear relationships, the original values were retained.

3. Regression Model Summarization: The broom package was used to tidy and visualize regression outputs, making coefficients and confidence intervals easier to interpret.

4. Advanced Visualizations: Scatter plots and histograms created with ggplot2 effectively summarized associations and transformation impacts.

6. Conclusion

Through association analysis and regression modeling, we identified key drivers of housing prices in California. The Ratio of Bedrooms to Total Rooms emerged as the strongest predictor, followed by proximity to the ocean and median income (fig. 31). Addressing multicollinearity improved the model's reliability. For further precision, coefficients and confidence intervals can provide additional insights into each variable's impact, enhancing the model's interpretability. These findings offer valuable guidance for stakeholders making data-driven decisions.

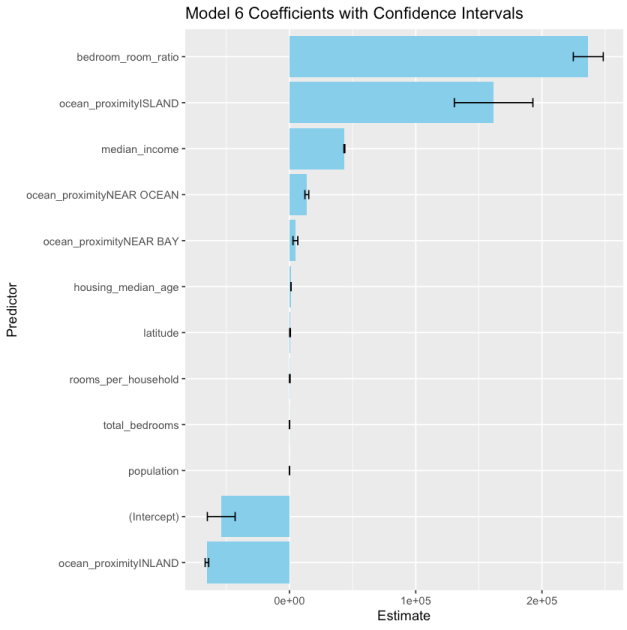

fig.31

7. Appendix I

Alaa El Hajjar: Regression Model, Other Techniques
Momo Ogawa: Association Analysis, Other Techniques, Conclusion
Momtahin Masud: Introduction, Data Cleaning

8. Appendix II (Separate Files)

Real data (housing.csv), R codes (STA9750-Project.r), and presentation slides (STA9750-Project-Presentation-Slides.pdf).