California Housing Price Analysis

Group 12 Project Presentation - STA9750
Alaa El Hajjar, Momo Ogawa, Momtahin Masud

12/10/2024

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, 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.

01

Data Exploration and Cleaning

Data Understanding & Cleaning

```
'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" ...
```

longitude: 0 **latitude:** 0 **housing_median_age:** 0 **total_rooms:** 0 **total_bedrooms:** 207 **population:** 0 **households:** 0
median_income: 0 **median_house_value:** 0 **ocean_proximity:** 0

Data Understanding & Cleaning

```
> 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
```

Creating New Variables and dropping null values

```r
# 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
```

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

Reason:

The creation of new variables through combining and transforming existing variables has given us extra room to come up with more ideas for us to run a thorough analysis and come up with interesting findings and insights.

Data Summary

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	-122.23	37.88	41	880	129	322	126	8.3252
2	-122.22	37.86	21	7099	1106	2401	1138	8.3014
3	-122.24	37.85	52	1467	190	496	177	7.2574
4	-122.25	37.85	52	1274	235	558	219	5.6431
5	-122.25	37.85	52	1627	280	565	259	3.8462
6	-122.25	37.85	52	919	213	413	193	4.0368

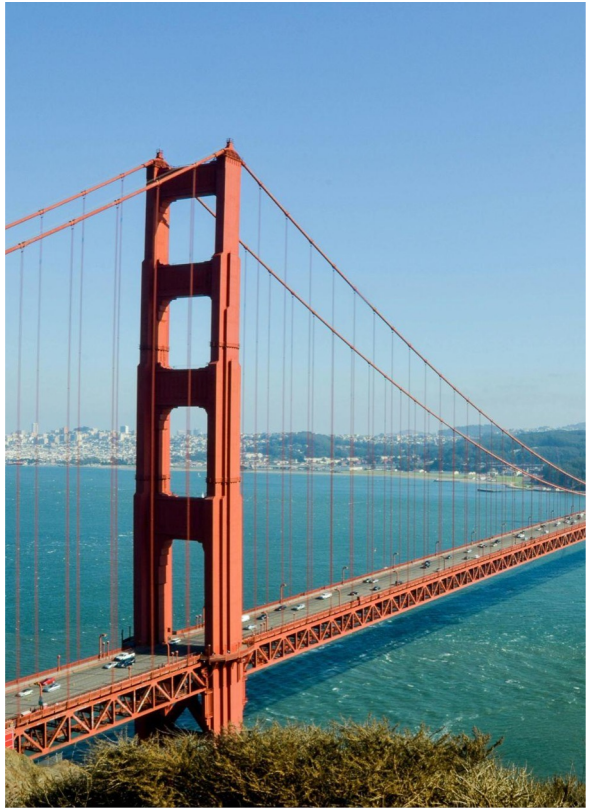
bedroom_room_ratio	rooms_per_household	income_rooms_ratio	ocean_proximity	median_house_value
<dbl>	<dbl>	<dbl>	<chr>	<dbl>
0.1465909	6.984127	1.1920173	NEAR BAY	452600
0.1557966	6.238137	1.3307499	NEAR BAY	358500
0.1295160	8.288136	0.8756372	NEAR BAY	352100
0.1844584	5.817352	0.9700462	NEAR BAY	341300
0.1720959	6.281853	0.6122715	NEAR BAY	342200
0.2317737	4.761658	0.8477719	NEAR BAY	269700

Data Summary

```
  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
```

02

Association Analysis



Exploring Variable Associations

Identify top correlations with **median_house_value (y)**

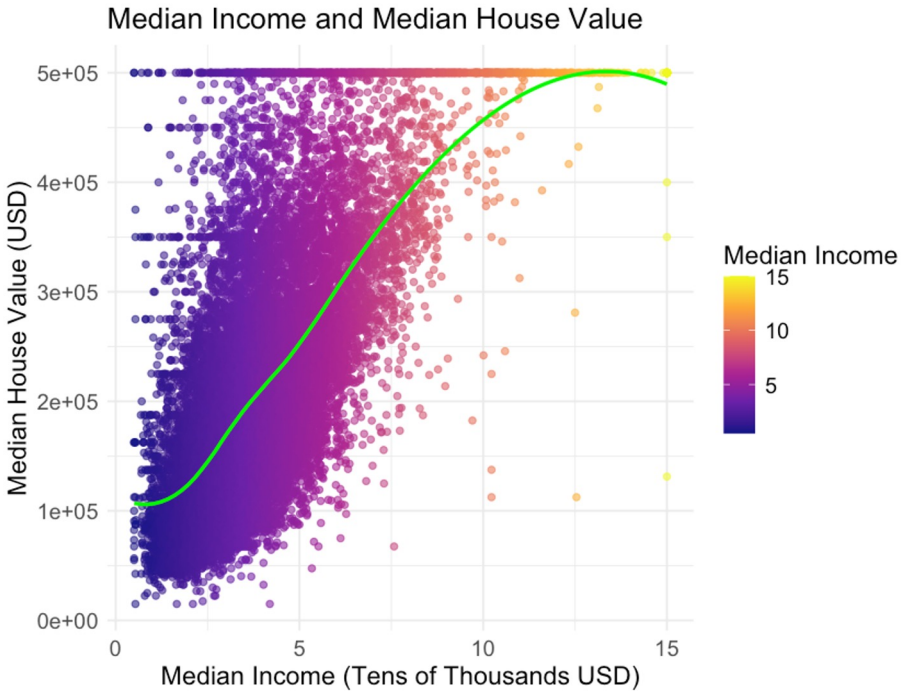
var1	var2	correlation	pval
<chr>	<chr>	<dbl>	<dbl>
median_income	median_house_value	0.68835548	0.000000e+00
income_rooms_ratio	median_house_value	0.66497497	0.000000e+00
bedroom_room_ratio	median_house_value	-0.25588015	8.160022e-303
rooms_per_household	median_house_value	0.15134408	5.822073e-105
latitude	median_house_value	-0.14463821	6.132893e-96
total_rooms	median_house_value	0.13329413	1.221172e-81
housing_median_age	median_house_value	0.10643205	1.496134e-52
households	median_house_value	0.06489355	1.611514e-20
total_bedrooms	median_house_value	0.04968618	1.191968e-12
longitude	median_house_value	-0.04539822	8.450466e-11
population	median_house_value	-0.02529973	2.982633e-04

Highlights:
Median Income (r = 0.688, strongest),
Population (r = -0.025, weakest).

Next Step:
Select the top 5 variables for visualization and further analysis.

1. Median Income and Median House Value

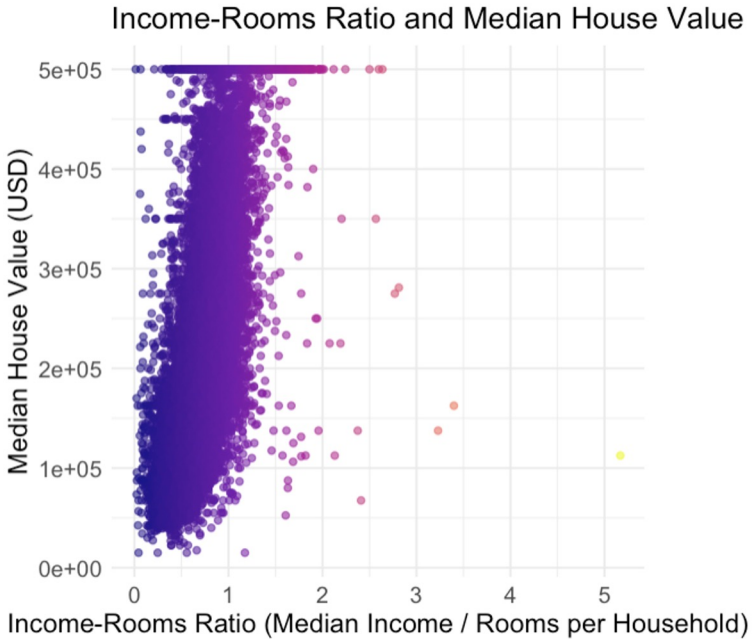
Conclusive Results: Statistically Significant




```
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
```

2. Income-Rooms Ratio and Median House Value

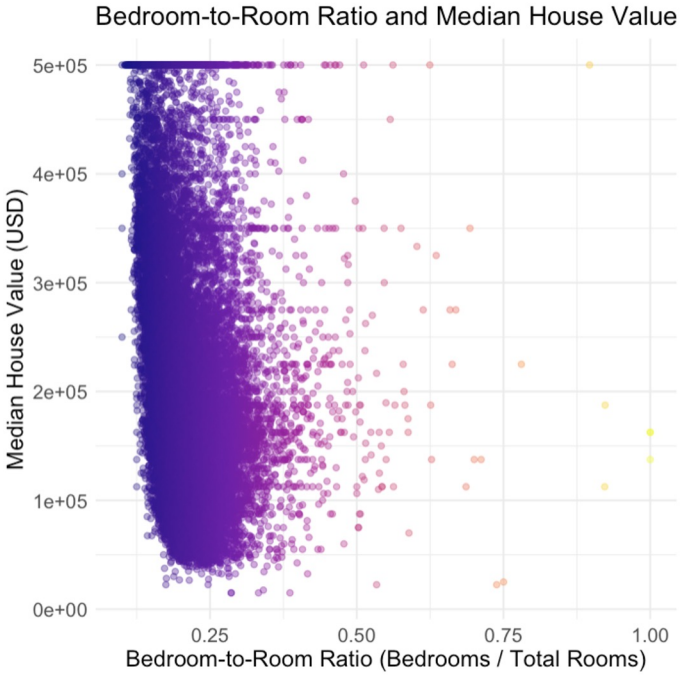
Conclusive Results: Statistically Significant




```
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
```

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

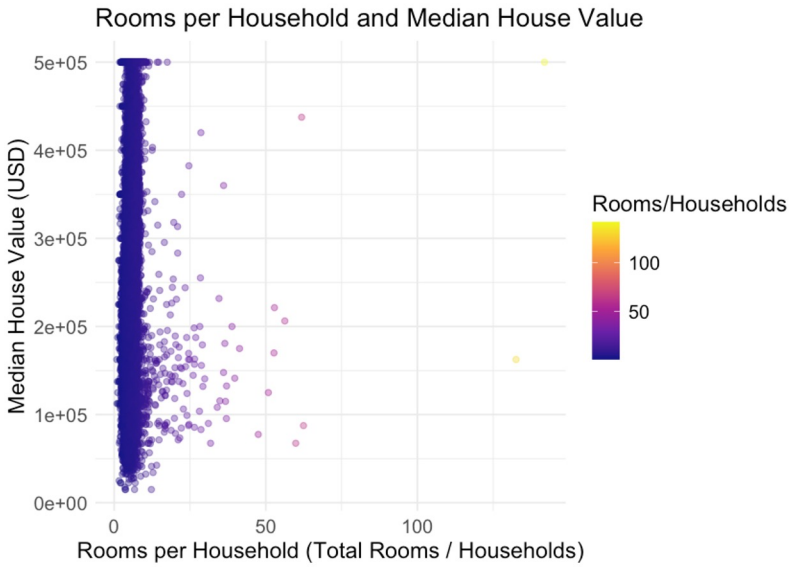
Conclusive Results: Statistically Significant



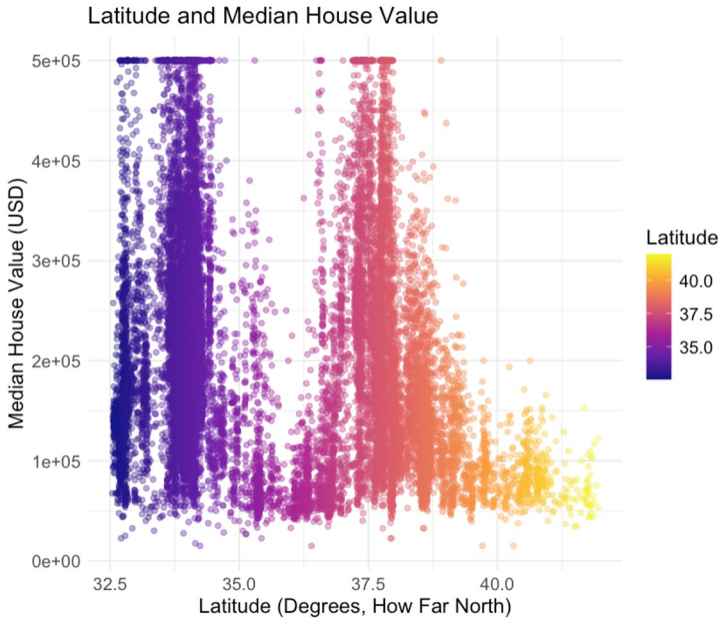

```
Permutation procedure:
                                  Value Estimated p-value
Pearson's r                  -0.2558801                 0
Spearman's rank correlation -0.3316925                 0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between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

4. Rooms per Household, Latitude, and Median House Value

Conclusive Results: Statistically Significant



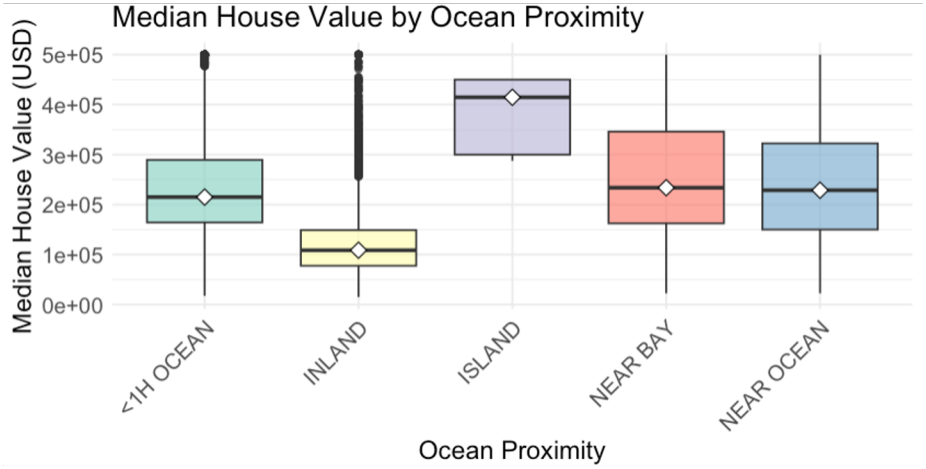

```
                                Value Estimated p-value
Pearson's r                 0.1513441                 0
Spearman's rank correlation 0.2633387                 0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
```

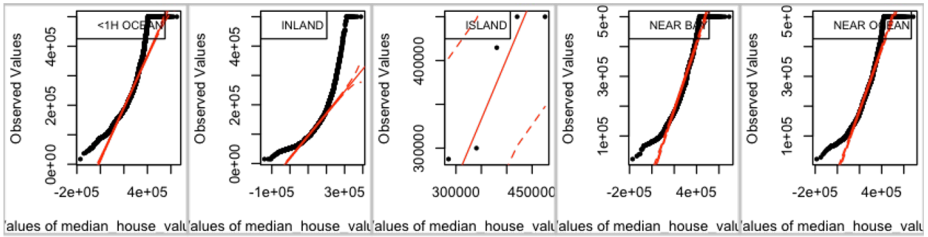



```
                                Value Estimated p-value
Pearson's r                 -0.1446382                 0
Spearman's rank correlation -0.1661199                 0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
```

5: Ocean Proximity and Median House Value

Conclusive Results: Statistically Significant





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

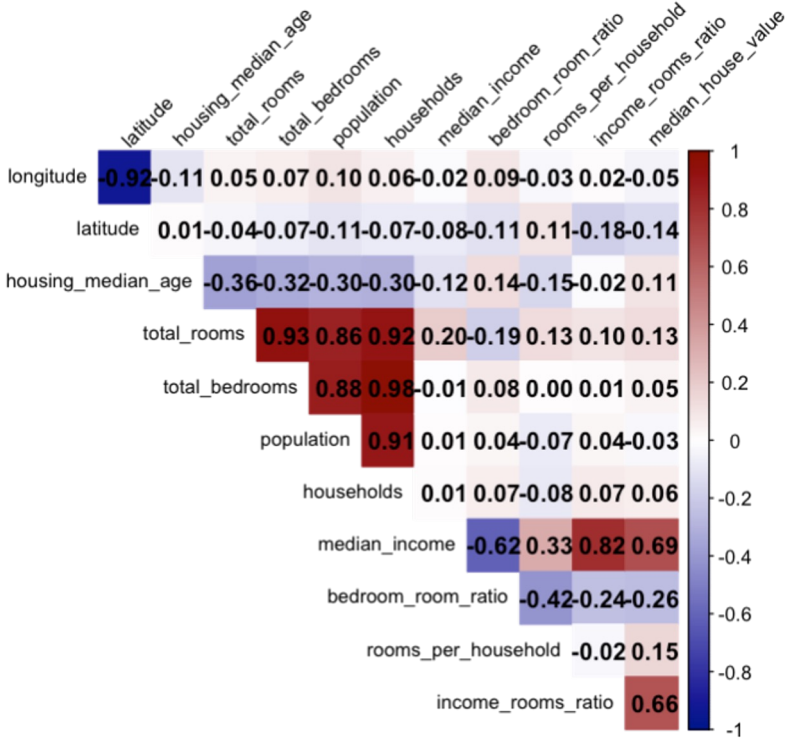
	Estimated p-value
Averages (ANOVA)	0
Mean Ranks (Kruskal)	0
Medians	0

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

Handling Multicollinearity

Identify highly correlated variable pairs (|r| >= 0.7) to avoid multicollinearity

	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



03

Regression Model

Baseline Regression Model- All independent Variables included

Code

```
lm(formula = median_house_value ~ longitude + latitude +
housing_median_age +
    total_rooms + total_bedrooms + population + households +
    median_income + bedroom_room_ratio + rooms_per_household +
    income_rooms_ratio + ocean_proximity, data = housing)

confint(model1,level=0.95)
summary(model1)
```

Output

```
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
```

Multicollinearity using VIF (Variance Inflation Factor)

	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

Package installation:

install.packages("car")

library(car)

Calculate vif of baseline model (model1):

vif(model1)

Model 2-Removing income_rooms_ratio

Code

```
lm(formula = median_house_value ~ longitude + latitude +
housing_median_age +
    total_rooms + total_bedrooms + population + households +
    median_income + bedroom_room_ratio + rooms_per_household +
    ocean_proximity, data = housing)

confint(model2,level=0.95)
summary(model2)
```

Output

```
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

Multicollinearity using VIF on Model 2

	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

Calculate vif of model 2:
vif(model2)

Model 3-Removing longitude

Code

```
lm(formula = median_house_value ~ latitude +
housing_median_age + total_rooms + total_bedrooms + population
+ households +
    median_income + bedroom_room_ratio + rooms_per_household +
    ocean_proximity, data = housing)

confint(model3 , level=0.95)
summary(model3)
```

Output

```
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

Multicollinearity using VIF on Model 3

	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

Calculate vif of model 3:
vif(model3)

Model 4-Removing household

Code

```
lm(formula = median_house_value ~ latitude +
housing_median_age + total_rooms + total_bedrooms + population
+
    median_income + bedroom_room_ratio + rooms_per_household +
    ocean_proximity, data = housing)

confint(model4 , level=0.95)
summary(model4)
```

Output

```
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

Multicollinearity using VIF on Model 4

	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

Calculate vif of model 4:
vif(model4)

Model 5-Removing total_bedrooms and keeping total_rooms

Code

```
lm(formula = median_house_value ~ latitude +
housing_median_age + total_rooms + population +
    median_income + bedroom_room_ratio + rooms_per_household +
    ocean_proximity, data = housing)

confint(model5 , level=0.95)
summary(model5)
```

Output

```
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

Multicollinearity using VIF on Model 5

	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

Calculate vif of model 5:
vif(model5)

Model 6-Removing total_rooms and keeping total_bedrooms

Code

```
lm(formula = median_house_value ~ latitude +
housing_median_age + total_bedrooms + population +
    median_income + bedroom_room_ratio + rooms_per_household +
    ocean_proximity, data = housing)

confint(model6 , level=0.95)
summary(model6)
```

Output

```
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

Multicollinearity using VIF on Model 6

	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

Calculate vif of model 6:
vif(model6)

Model 6- Best Fit

- Model 6 had the best F-statistics = 3274 "highest value" among all the models with a p-value of < 2.2e-16.

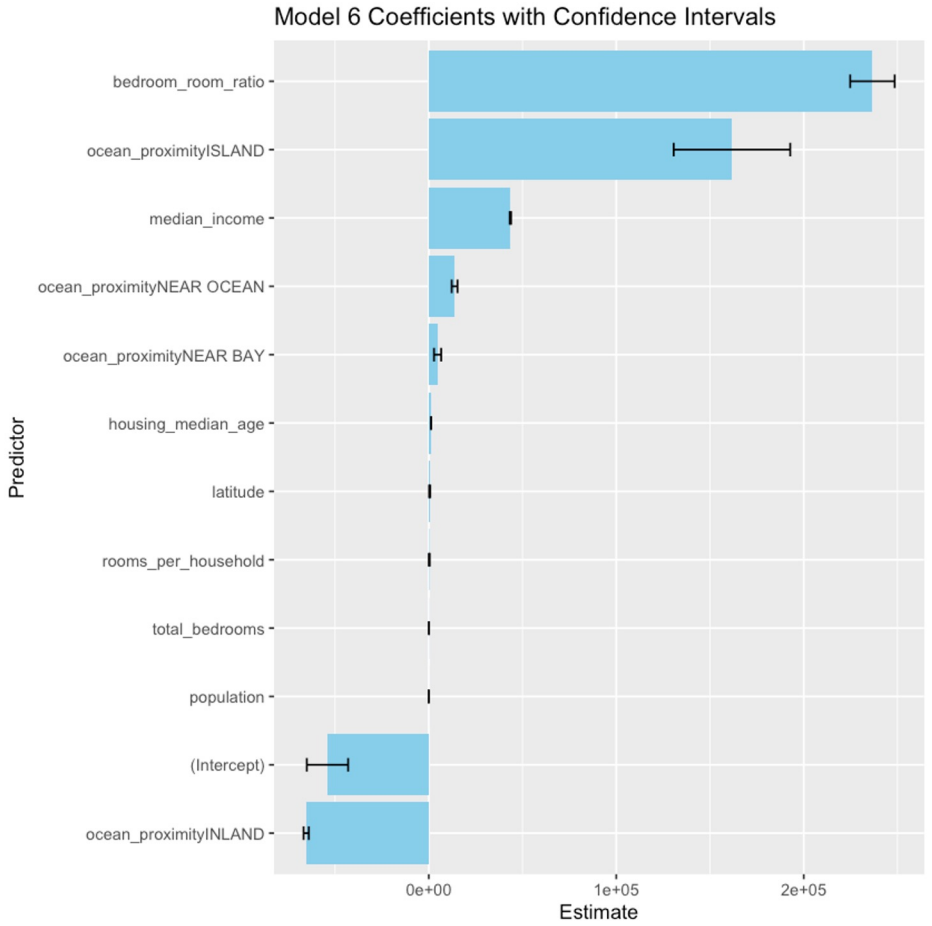
- Model is overall meaningful and statistically significant.

- Even though Adjusted r-square is 0.638, which is slightly lower than other models, model 6 takes into account multicollinearity which avoids potential instability in the coefficients and that is essential. Therefore, we believe model 6 is the best fit for predicting median_house_value.

Conclusion

Conclusion

- **Bedroom-to-room ratio** is the strongest predictor, followed by **proximity to the ocean** and **median income**.
- The model highlights **key drivers**, offering actionable insights for **stakeholders**.



Thank you!

Feel free to ask any questions.

Group 12 Project Presentation - STA9750

Alaa El Hajjar, Momo Ogawa, Momtahin Masud