Foundations of Data Science (H323 Online) - Assessment Report
Melbourne Housing Price Data - Statistical Analysis

Rushmila Islam

2023-06-05

Introduction

The value of a house is influenced by various factors, encompassing location, supply and demand dynamics, economic outlook, interest rates, size, and demographics. This report aims to statistically analyse the relationship between house prices and family house size in greater Melbourne area. Using basic statistical approaches using R (R Core Team [2017]) we investigate the relationship between the size of a family house measured by the number of rooms and selling prices in eastern and western regions and investigate the influence of location on a house's price.

House	Selling_Price	Location	Number_of_Rooms
1	345	0	8
2	655	0	9
3	325	1	7
4	478	0	4
5	432	1	10
6	233	1	6
7	567	1	9
8	988	0	13
9	199	1	6
10	934	0	12
11	258	1	6
12	379	1	10
13	355	1	8
14	643	0	10
15	710	0	11
16	585	0	9
17	312	1	4
18	870	0	12
19	670	0	9
20	280	1	7

Exploratory Data Analysis using Summary Statistics

In this section, we explore the initial dataset using simple summary statistics in R. At first, we drop the House column as it only represents the row id and does not have any significance value. Next, we rename and reorder the columns for easier analysis. We also create two separate datasets for Melbourne East and Melbourne West for later analysis and explore detail summary statistics such as mean, standard deviation (sd), standard error (se), median, trimmed mean, median absolute deviation (mad), minimum (min), maximum (max), range, skewness (skew), kurtosis, and interquartile range (IQR).

Summary statistics for greater Melbourne area:

	Location	Rooms	Price
vars	1.00	2.00	3.00
n	20.00	20.00	20.00
mean	0.50	8.50	510.90
sd	0.51	2.54	239.50
median	0.50	9.00	455.00
trimmed	0.50	8.56	491.50
mad	0.74	2.97	269.09
min	0.00	4.00	199.00
max	1.00	13.00	988.00
range	1.00	9.00	789.00
skew	0.00	-0.10	0.52
kurtosis	-2.10	-0.96	-1.02
se	0.11	0.57	53.55
IQR	1.00	3.25	337.00

Summary statistics for Melbourne East and Melbourne West, respectively:

	Rooms (East)	Price (East)	Rooms (West)	Price(West)
vars	1.00	2.00	1.00	2.00
n	10.00	10.00	10.00	10.00
mean	9.70	687.80	7.30	334.00
sd	2.58	199.98	1.95	107.52
median	9.50	662.50	7.00	318.50
trimmed	10.00	693.12	7.38	321.75
mad	2.22	194.22	1.48	89.70
min	4.00	345.00	4.00	199.00
max	13.00	988.00	10.00	567.00
range	9.00	643.00	6.00	368.00
skew	-0.75	-0.03	0.03	0.77
kurtosis	-0.21	-1.20	-1.31	-0.36
se	0.82	63.24	0.62	34.00
IQR	2.75	230.50	2.75	109.50

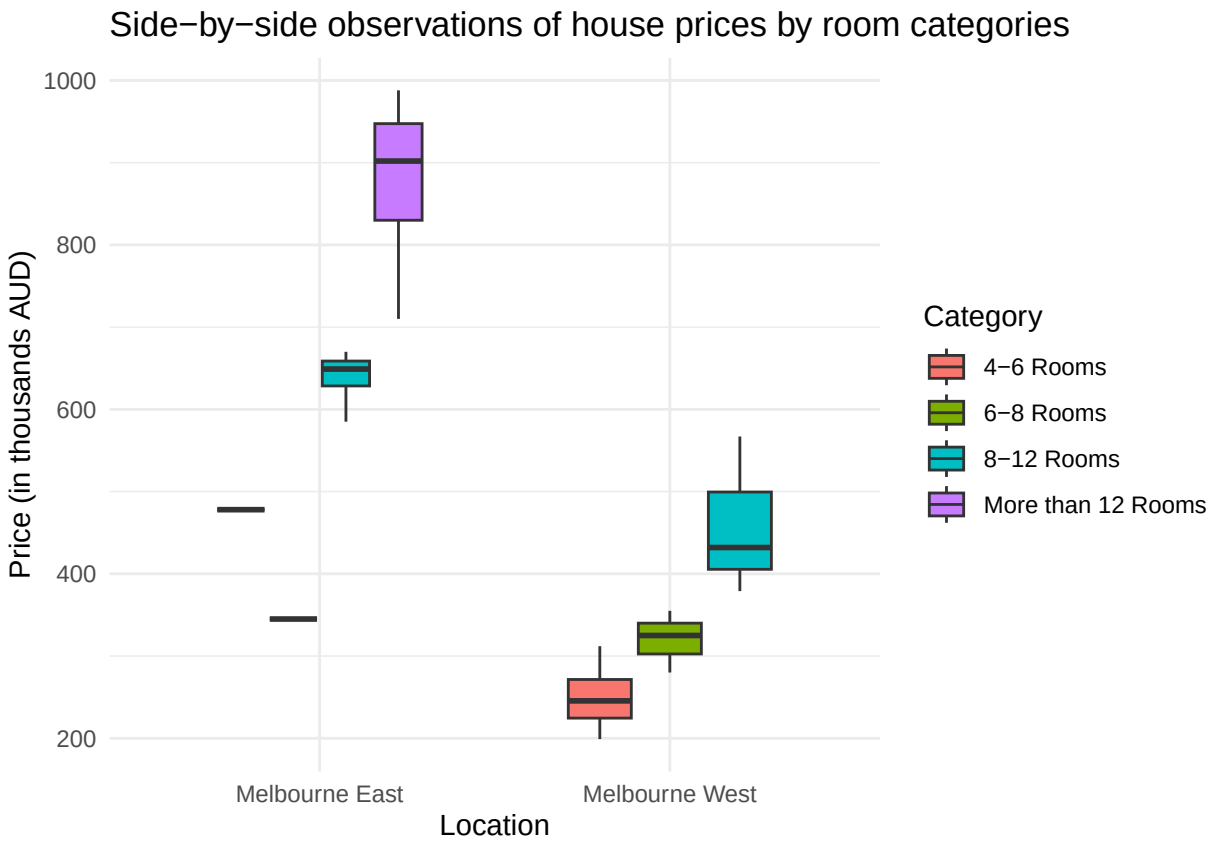
Sampling Method

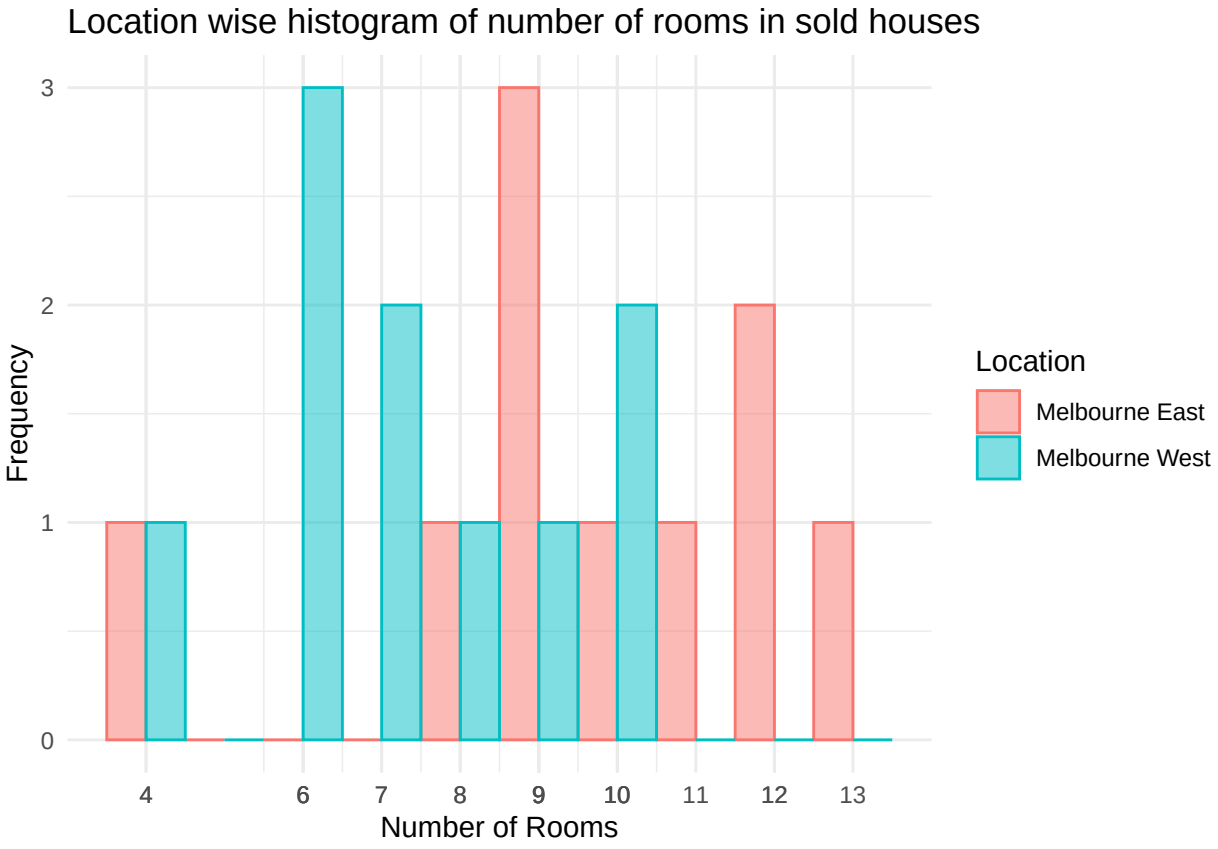
Based on our observations, in the given dataset Sampling is used to select a subset of units e.g., houses in this study from a larger population as it is not possible to obtain the results for the whole population. Here, based on previous studies sampling population is divided into two homogeneous groups of Melbourne East and Melbourne West. These groups are called strata, then random sampling is used to select 10 houses from each stratum to study. This is a stratified random sampling technique with a combination of random sampling. So, we can safely assume it is a multistage sampling technique where several method of sampling techniques are used.

Data Visualisation

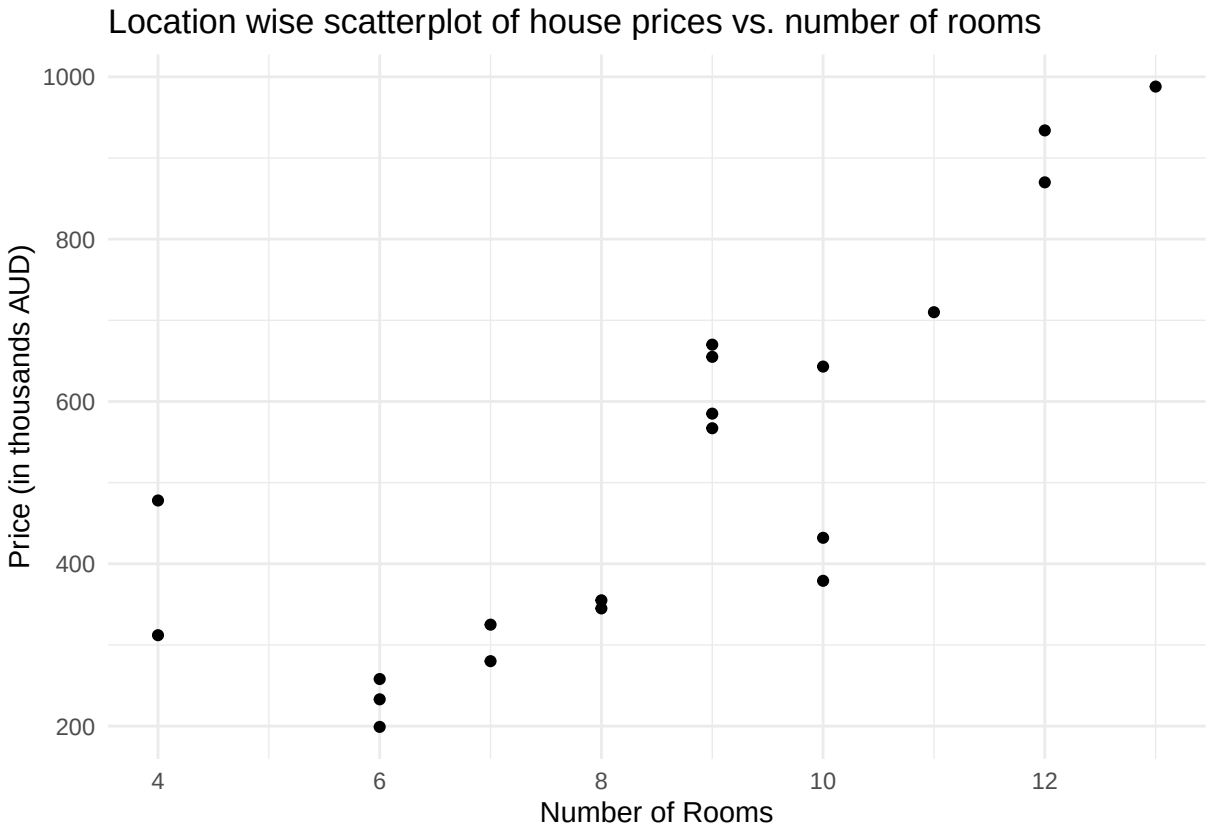
We first observe the Melbourne East and Melbourne West datasets in side-by-side box plots grouped by location which gives us visual understanding of how the prices are varied between these two greater Melbourne areas. From the box plots its very clear that the sold house prices in Melbourne East are higher for all room categories comparing to Melbourne West houses. Next,

we observe a location wise histogram of number of rooms in the sold houses. From the graph we can understand that many of houses in Melbourne East had higher number of rooms comparing to Melbourne West. Hence, it matches with our previous observations for sold house prices.





Next, we draw a scatter plot which shows the relation of in house prices with number of rooms of a house. We can look at the scatterplot and see that house prices increases when there a more rooms in the house. The larger houses cost more. We see there is an association between two quantitative variables. We also observe from the data the direction of the association is positive and it has a linear form although some data points are bit away from it.



Regression Analysis

In our first regression model we only consider a single variable of Rooms i.e., the number of rooms against Price.

```
##
## Call:
## lm(formula = Price ~ Rooms, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -246.74 -116.45   12.48  109.57  311.62
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -139.87     114.29  -1.224    0.237
## Rooms          76.56      12.91   5.932  1.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.2 on 18 degrees of freedom
## Multiple R-squared:  0.6615, Adjusted R-squared:  0.6427
## F-statistic: 35.18 on 1 and 18 DF,  p-value: 1.297e-05
```

based on the regression chart with only Room variable the simple linear regression equation for this model will be:

$$\hat{Price} = -139.87 + 76.56 \times Rooms$$

From the above regression summary the lower P value indicates that the association we see in the data is unlikely to have occurred by chance. Therefore, we reject the null hypothesis and conclude that there is strong evidence that house price linearly related to number of rooms.

Next, we want to test the null hypothesis that the house price is not related to number of rooms. The slope of this relationship will model the relationship between house price and number of rooms of the house. Our null hypothesis is that the slope of the regression is zero:

H_0: The price of a house is not linearly related to the number of rooms: $\beta_1 = 0$.

H_A: The price of a house is related to the number of rooms: $\beta_1 \neq 0$.

At 5% level of significance (95% confidence interval) for β_1 is $b_1 \pm t^*_{n-2} \times SE(b_1)$

With 18 degree of freedom we found t value from t distribution table 2.1009 and calculate value as $76.56 \pm 2.1009 \times 12.91 = (49.44337, 103.6786)$ at thousands AUD per house.

Next, we calculate the confidence intervals for the coefficients.

```
##                   2.5 %   97.5 %
## (Intercept) -379.97783 100.2412
## Rooms         49.44337 103.6786
```

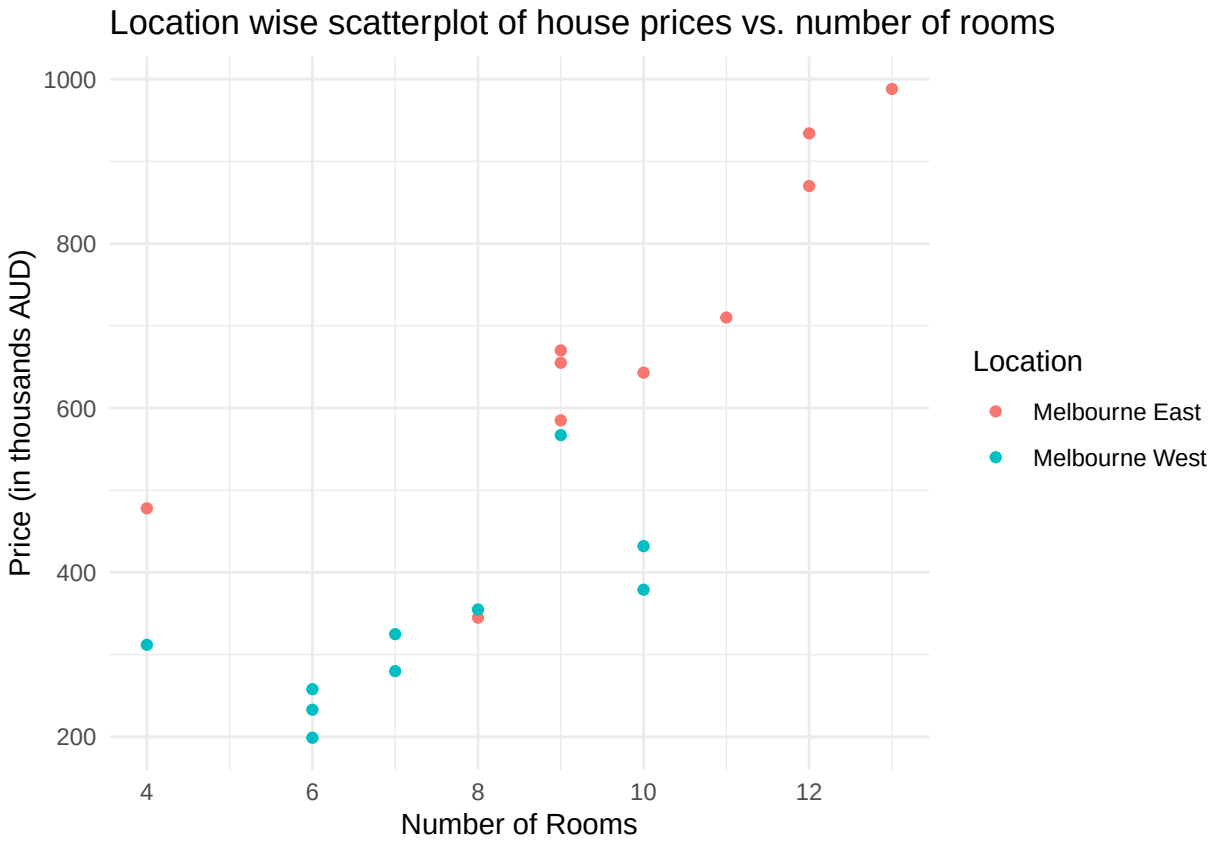
The regression analysis shows that on average house price is higher by AUD 76.56 thousands for every additional room. We are 95% confident that the actual price is higher by between AUD 49.44 and 103.68 thousands for each additional room.

Next, we come up with a multiple regression model where we include Location as a variable.

```
##
## Call:
## lm(formula = Price ~ Rooms + Location, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -249.47  -52.99  -11.03   67.73  159.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   155.29     107.88   1.439 0.168191
## Rooms          54.90      10.60   5.177 7.58e-05 ***
## Location     -222.04      52.59  -4.222 0.000574 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.9 on 17 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.8153
## F-statistic: 42.95 on 2 and 17 DF,  p-value: 2.257e-07
```

The previous model that only consider Rooms has R^2 value of 66.15% but this model considers 83.48% of the variability. We can understand that the new variable location contributes to good prediction of the house. Collecting the coefficients of the multiple regression of Price on Rooms

and Location from the above output we can get the mutiple linear regression equation of this model.



In the above scatterplot we plot the house selling data divided into two groups based on location. We create a new variable here that indicates the location of the house, giving it value 0 for east and 1 for west.

Below equation defines the multiple linear regression model with two variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

β_0 = Y Intercept

β_1 = Slope of Y with variable X_1 when X_2 is constant

β_2 = Slope of Y with variable X_2 when X_1 is constant

ϵ = Random error

Multiple linear regression equation to estimate Price with two independent variable Rooms and Location:

$$\hat{Price} = 155 + 54.9 \times Rooms - 222.04 \times Location$$

The slope of Rooms with Price (54.90) indicates that, for a increase of room, average house Price is increased by AUD 54.90 thousands dollar. The slope of the Location with Price (-222.04) indicates that, for a given location, average house Price is estimated to decrease by AUD 222.04 thousands dollar.

Looking at the regression coefficients in this model we can state that there is a constant effect of number of rooms on house price, for each additional room there is an increase of price, we estimate that on average house price increases by AUD 54.90 thousands.

The effect of location on house price, saying the house in East Melbourne compared a house in West Melbourne. Thus, considering Rooms as constant, we estimate that the average price of a house is AUD 222.04 thousands less for western suburbs than a house in eastern suburb. The t statistics representing the slopes for number of Rooms and Location are 5.177 and -4.222. The corresponding p-values are very low as 0.0000758 and 0.000574. So, both Rooms and Location are statistically significant in determining house prices with the 0.05 level of significance.

This model has a R^2 of 83.48% which means this model has explained % of the variation in Price using these two predictors as Rooms and Location. The standard deviation of the residuals is about AUD 102.9 thousands. Using $68-95-99$ Rule we say that most prediction errors will be no larger than AUD 205.8 thousands as we also observe large t-ratios and corresponding small P-values.

Test: Predict the selling price for a house with nine rooms that is located in Melbourne's east

Based on the above model predicting a selling price for a house with nine rooms in Melbourne's east using below equation will cost AUD 649.1 thousands.

$$\hat{Price} = -155 + 54.9 \times 9 - 222.04 \times 0 = 649.1$$

F-test Analysis

As we now have more than one predictor in this model, we test the null hypothesis that all the slope co-efficient are zero:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0. H_A : \text{at least one } \beta \neq 0.$$

To test this hypothesis we do a F-test. F-distribution has two degrees of freedom k and $n-k-1$. Here, k is the number of predictors and n is number of samples. In our study we have predictors $k = 2$ and $n = 20$ houses. Which means the F-value 42.95 has $20 - 2 - 1 = 17$ degrees of freedom. The regression output table shows a very low P-value. The null hypothesis is that all the coefficients are zero. The alternative is that at least one coefficient is no-zero. This test is one sided, bigger F value means smaller P-values. If the null hypothesis were true the F value would have value near 1. Using a 0.05 level of significance, the critical value of the F distribution with 2 and 17 degrees of freedom found from F distribution table is 3.59. The F-statistics here we found in our analysis is 42.95 which is quite large than the critical value in F distribution table, so we can easily reject null hypothesis. In conclusion, we can say that the multiple regression model for predicting house prices with two variables is better than just using the mean.

Alternative Specification and Robustness Check

In the previous regression model, we assumed that the effect of the number of rooms has on the assessed value is independent of the location. In other words, we assumed that the slope of assessed the price of the house in eastern suburb is the same as it is for western suburb. If these two slopes are different, an interaction between price of the apartment and location exists.

Looking in to regression line of two location groups we can observe that in Melbourne east house prices are more than west and the two regression lines have different slopes. To adjust slopes we introduce another variable - the product of Rooms and Location (Rooms * Location).

To evaluate a hypothesis of equal slopes of a Y variable with X, we first define an interaction term that consists of the product of the independent variable number of Rooms and the dummy variable Location. We then test whether this interaction variable makes a significant contribution to a regression model that contains the other X variables. If the interaction is significant, we cannot use the original model for prediction. The regression model for this is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

```
##
## Call:
## lm(formula = Price ~ Rooms + Location + Rooms:Location, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -233.12  -57.50   -5.01   47.19  168.49
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61.997    130.940   0.473 0.642271
## Rooms            64.516     13.087   4.930 0.000151 ***
## Location         -5.011    185.099  -0.027 0.978736
```

```
## Rooms:Location  -26.569     21.752  -1.221 0.239620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101.5 on 16 degrees of freedom
## Multiple R-squared:  0.8489, Adjusted R-squared:  0.8205
## F-statistic: 29.96 on 3 and 16 DF,  p-value: 8.452e-07
```

The equation we derived from the above output is:

$$\hat{Price} = 61.997 + 64.516 \times Rooms - 5.011 \times Location - 26.569 \times Rooms * Location$$

To test for the existence of an interaction, we use the null hypothesis:

$$H_0 : \beta_3 = 0. H_A : \beta_3 \neq 0.$$

From the above analysis, the t statistic for the interaction of Rooms and Location is -1.22 and P-value 0.0239620. Because the P-value $0.0239620 > 0.05$, we do not reject the null hypothesis. Therefore the interaction term does not make any significant contribution to the model. As the model already includes number of Rooms and Location, therefore, the multiple regression model with no interaction terms is the better model.

Conclusion

In conclusion based on the above statistical analysis examines two Melbourne suburbs house prices against location and number of rooms, and although there are other factors involve in house price and the number of data points are very minimal. Therefore, its hard to conclude anything very concrete. But we do understand the house price do vary based on the location and number of rooms. Based on our study we can say the two variable variation works better. In this report and analysis purpose we use [Sharpe et al., 2023] as a key reference to understand the fundamental statistical concepts. We aim to do further analysis in future.

References

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Norean R. Sharpe, Richard D. De Veaux, and Paul F. Velleman. *Business Statistics*. Pearson, 4th edition, 2023.