The Upside of Upsizing
A Deep Dive into House Size and Price in Melbourne

Michael Alizzi
25 March 2025

Sampling Method

The following multi-stage sampling method was conducted.

1. Stratified Sampling

Stratification is appropriate when strata have different variances for some variable being studied (Blair E. and Blair J., 2015, p. 113). Considering previous studies have identified that there are differences between Melbourne's East and West, this study has assumed that there are differences in the variance of sale price between the two geographical strata.

2. Cluster Sampling

Clustering is appropriate when there are substantial fixed costs associated with each data collection location (Blair E. and Blair J., 2015, p. 129). Considering the constrained budget, this study has assumed that the two clusters chosen for Melbourne's East and West suburbs (Dandenong and Sunshine respectively) are representative of the population of each strata (Sharpe, Norean R., De Veaux, Richard D., Velleman, Paul F., Bock, David E., 2015, p. 278).

3. Simple Random Sampling (SRS)

From the sampling frames of house data available in Dandenong and Sunshine, this study has assumed that the 10 houses selected in each suburb/cluster have been chosen at random with each data point having an equal chance of selection (Blair E. and Blair J., 2015, p. 63).

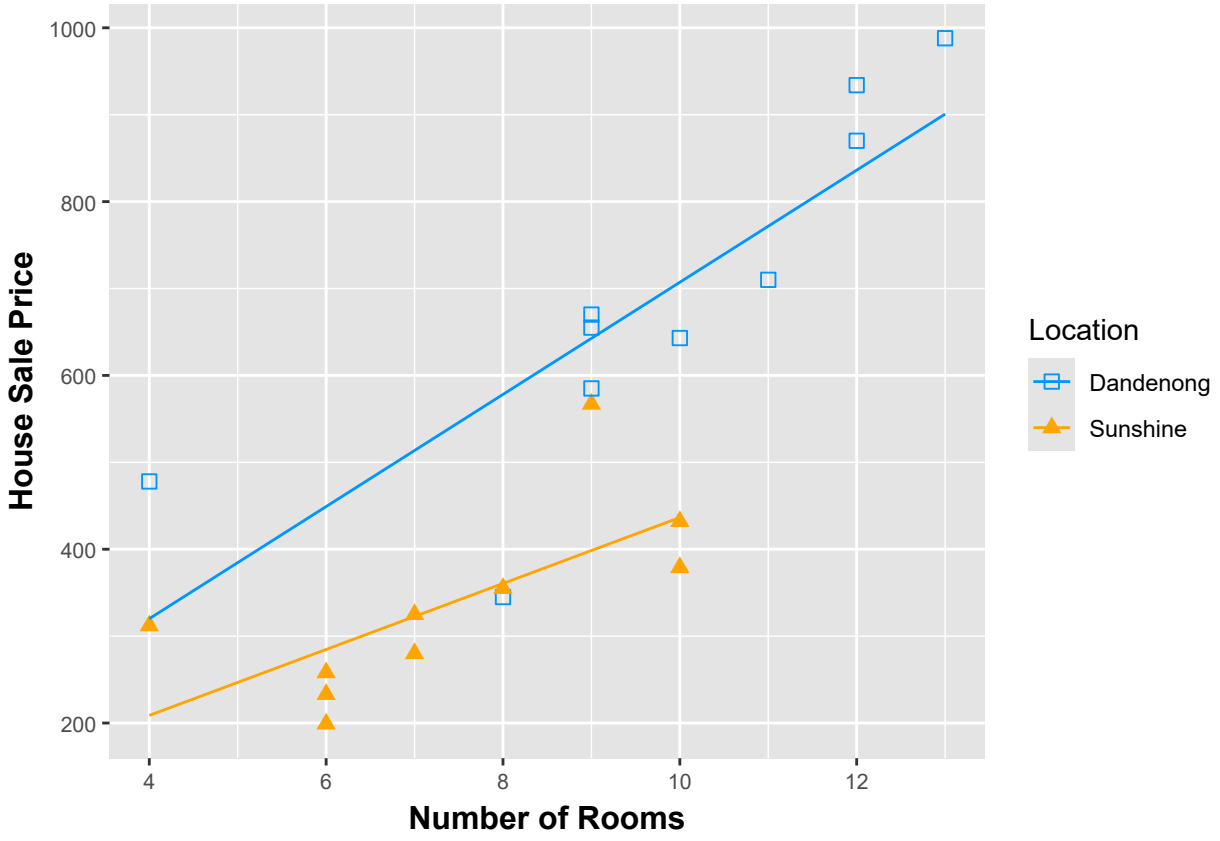
The Relationship Between Rooms & Price

A scatter plot was chosen to identify, at a glance, the correlation and strength of the relationship between sale price and number of rooms. The actual data points in the plot allow clear identification of a generally a positive correlation between number of rooms and sale price (as one would expect), regardless of suburb.

The use of shape and color to distinguish between Dandenong and Sunshine visualises the validity of previous studies that have suggested differences between the two suburbs.

When plotting two lines of best fit with simple linear regression, Dandenong has a steeper slope than its Sunshine counterpart, indicating a stronger relationship and that increases in room number have a larger effect on sale price. For the Sunshine strata, this could mean that simple linear regression is not the most appropriate regression model.

Additionally, houses in Dandenong generally sell for higher than those in Sunshine, as seen in the differences in the actual data points and the difference in intercepts of the lines of best fit.



Initial Estimation of Regression Equation

A multiple linear regression model were estimated using Ordinary Least Squares (OLS).

Equation

$$\text{SalePrice} = 155.29 + (54.9 \times \text{NumberOfRooms}) - (222.04 \times \text{Location}) + \varepsilon$$

Interpretation

Holding disturbance (ε) and location constant, for every additional room in a house, sale price on average increases by $54900.

Holding disturbance (ε) and the number of rooms constant, when location equals 1 (or alternatively, the house is in Sunshine), sale price on average decreases by $222040.

For example, a nine room house in Dandenong (Melbourne's East) will have an expected value of sale price as:

$$\widehat{\text{SalePrice}} = 155.29 + (54.9 \times 9) - (222.04 \times 0) =$$

$$649.3713$$

$$Or$$

$$\$649371.3$$

Significance of Independent Variables on Sale Price

Inferential Statistics

```
Call:
lm(formula = SellingPrice ~ NumberofRooms + Location, data = h_data)

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
NumberofRooms    54.90      10.60   5.177 0.0000758 ***
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: 0.0000002257
```

Joint Significance

An F-test was conducted to determine whether the two independent variables, the number of rooms and location, had a jointly significant relationship on sale price.

H_0: $\beta_{\text{Number of Rooms}} = 0$ and $\beta_{\text{Location}} = 0$

H_1: $\beta_{\text{Number of Rooms}} \neq 0$ or $\beta_{\text{Location}} \neq 0$

Assuming $\beta_{\text{Number of Rooms}} = 0$ and $\beta_{\text{Location}} = 0$ is true, the probability (or p-value) of obtaining the observed data is very small, well below the 5% level of significance The null hypothesis is therefore rejected and it can be concluded with 95% confidence that there is a jointly significant effect of the number of rooms and location on sale price.

This also implies that including $\beta_{\text{Number of Rooms}}$ and β_{Location} will have better predictive power than if the intercept was the only parameter.

Individual Significance

Two t-tests were conducted to determine whether each independent variable, the number of rooms and location, individually had a significant effect on sale price.

H_0: $\beta_{\text{Number of Rooms}} = 0$

H_1: $\beta_{\text{Number of Rooms}} \neq 0$

Assuming $\beta_{\text{Number of Rooms}} = 0$ is true, the probability (or p-value) of obtaining the observed data is very small, well below the 5% level of significance. The null hypothesis is therefore rejected and it can be concluded with 95% confidence that the number of rooms has an individually significant effect on sale price.

H_0: $\beta_{\text{Location}} = 0$

H_1: $\beta_{\text{Location}} \neq 0$

Assuming $\beta_{\text{Location}} = 0$ is true, the probability (or p-value) of obtaining the observed data is very small, well below the 5% level of significance. The null hypothesis is therefore rejected and it can be concluded with 95% confidence that location has an individually significant effect on sale price.

Interaction Term Impact

When re-estimating the regression equation with an interaction term, the location variable was removed to avoid multicollinearity.

Equation with Interaction Term

$$\text{SalePrice} = 59.49 + (64.76 \times \text{NumberOfRooms}) - (27.13 \times \text{NumberOfRooms} \times \text{Location}) + \varepsilon$$

Inferential Statistics with Interaction Term

```
Call:
lm(formula = SellingPrice ~ NumberofRooms + NumberRooms_Location,
    data = joint_h_data)

Residuals:
     Min       1Q   Median       3Q      Max
-232.559  -56.883   -4.613   46.716  168.887

Coefficients:
                     Estimate Std. Error t value   Pr(>|t|)
(Intercept)            59.489     89.788   0.663    0.51650
NumberofRooms          64.759      9.240   7.008 0.00000211 ***
NumberRooms_Location  -27.134      5.911  -4.590    0.00026 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 98.43 on 17 degrees of freedom
Multiple R-squared:  0.8489,    Adjusted R-squared:  0.8311
F-statistic: 47.74 on 2 and 17 DF,  p-value: 0.0000001058
```

Interaction Term Significance

Another t-test was conducted, this time to determine whether the multiplicative interaction term (number of rooms by location), had a significant effect on sale price.

H_0: $\beta_{\text{Number of Rooms} \times \text{Location}} = 0$

H_1: $\beta_{\text{Number of Rooms} \times \text{Location}} \neq 0$

Assuming $\beta_{\text{Number of Rooms} \times \text{Location}} = 0$ is true, the probability (or p-value) of obtaining the observed data is very small, well below even the 1% level of significance. The null hypothesis is therefore rejected and it can be concluded with 99% confidence that the interaction term has an individually significant effect on sale price.

Conclusion

Although both regression models are relatively appropriate, the regression with the interaction term displays a notable improvement. Its p-value when testing for joint significance is approximately half of that of the initial regression model. Adjusted r-squared is also around 1%-2% higher, with the variation in independent variables explaining more of the variation in sale price.

As an additional note, the AIC and BIC for the model with the interaction term (245.08 & 249.06 respectively) are smaller than the no interaction term counterpart (246.86 & 250.85 respectively), albeit by a minor margin, displaying minor improvements in model fit.

Appendix

OpenAI. 2025, ChatGPT [Large language model], Retrieved March 25, 2025, from https://chatgpt.com/.

Sharpe, Norean R., De Veaux, Richard D., Velleman, Paul F., Bock, David E., 2015, Business Statistics, Pearson Education Limited, eBook, accessed 30 March 2025 from ProQuest Online Database.

Blair E. and Blair J., 2015, Applied Survey Sampling, SAGE Publications Inc, eBook, accessed 31 March 2025 from Sage Online Database.

Acknowledgement of AI

I would like to acknowledge the assistance provided by ChatGBT (OpenAI. 2025) which offered formatting suggestions of this markdown pdf document.

Some examples of prompts I used include:

- Can I reference the output of code in latex?
- How to change the font in a r markdown document