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

**Michael Alizzi**
**25 March 2025**

---

## Sampling Method

Melbourne housing census data on sale price and the number of rooms was unattainable or impractical due to the following reasons:

- Some house sales are private or aren't publicly listed on real estate associations.
- House prices in Melbourne are volatile, fluctuating from a mean of $908,800 in June 2024 to $874,200 in December 2024 (Australian Bureau of Statistics 2025). Therefore, an estimate from a sample may yield more accurate results (Sharpe, Norean R., De Veaux, Richard D., Velleman, Paul F., Bock, David E., 2015, p. 275).
- The errors associated with attempting to collect Melbourne housing census data (Sharpe, Norean R., De Veaux, Richard D., Velleman, Paul F., Bock, David E., 2015, p. 275).

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 have been chosen at random with each data point having an equal chance of selection (Blair E. and Blair J., 2015, p. 63).

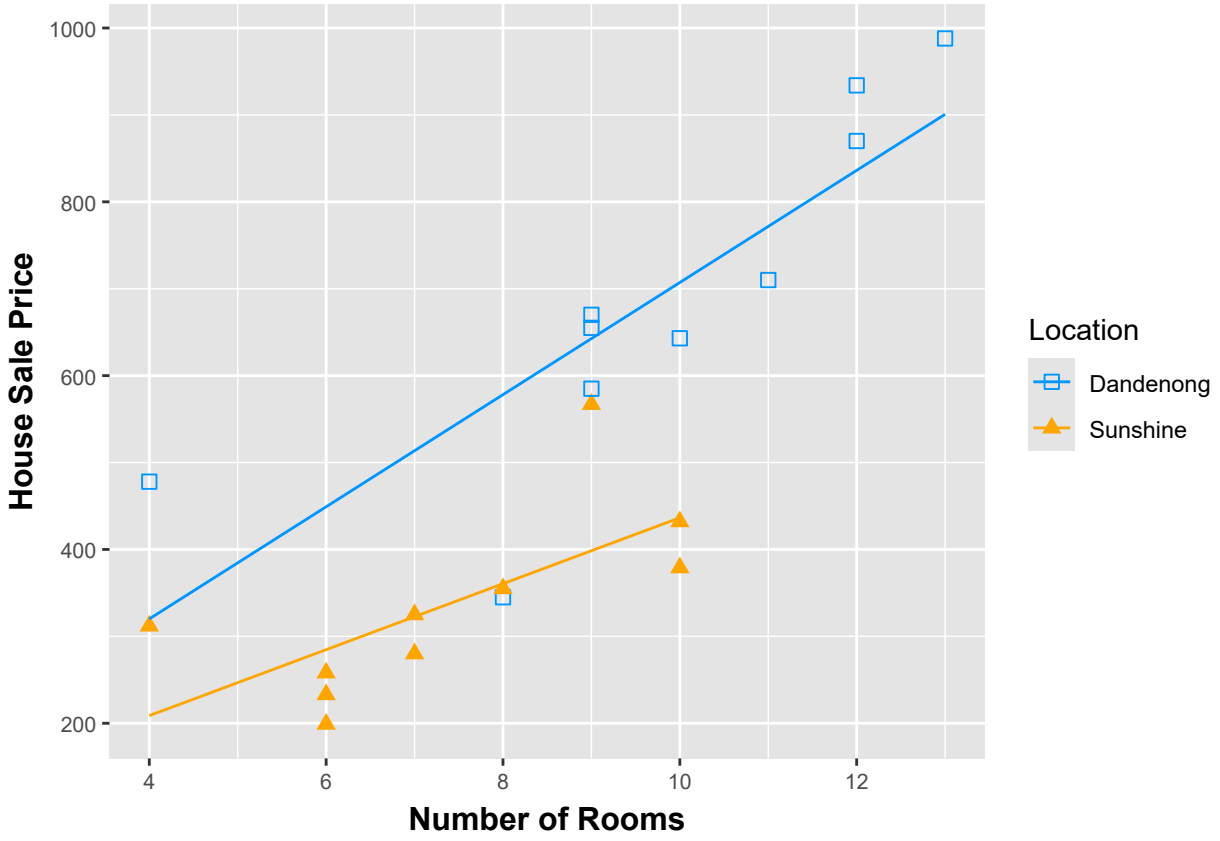
## The Relationship Between Rooms & Price

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 size 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

Two multiple linear regression models were estimated using Ordinary Least Squares (OLS), one with no transformation and one with a log transformation of the dependent variable. A log transformation was chosen as the dependent variable, selling price, is unlikely to be negative. These two regression models will be compared to determine which is the most appropriate.

**Linear Equation**

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

**Linear Interpretation**

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

Holding disturbance ($\varepsilon$) 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:

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

$$649.3713$$

$$Or$$

$$\$649371.3$$

**Log-Linear Equation**

$$\log(\text{SalePrice}) = 5.5 + (0.1 \cdot \text{NumberOfRooms}) - (0.48 \cdot \text{Location}) + \varepsilon$$

**Log-Linear Interpretation**

Holding disturbance ($\varepsilon$) and location constant, for every additional room in a house, sale price increases approximately on average by 10%.

Holding disturbance ($\varepsilon$) and the number of rooms constant, when location equals 1 (or alternatively, the house is in Sunshine), sale price decreases approximately on average by 48%.

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

$$\log(\widehat{\text{SalePrice}}) = 5.5 + (0.1 \cdot 9) - (0.48 \cdot 0) =$$

$$6.4197$$

$$Or$$

The expected value of sale price would be:

$$\widehat{\text{SalePrice}} = exp(5.5 + (0.1 \cdot 9) - (0.48 \cdot 0)) =$$

$$\$613835.8$$

The approximate percent increase in sale price from eight to nine bedrooms in Dandenong (Melbourne's East) can be calculated as:

$$\log(\widehat{\text{SalePrice}}) = 5.5 + (0.1 \cdot 9) - (0.48 \cdot 0)$$

$$-$$

$$\log(\widehat{\text{SalePrice}}) = 5.5 + (0.1 \cdot 8) - (0.48 \cdot 0)$$

$$\approx$$

$$10.2\%$$

## Significance of Independent Variables on Sale Price

### Linear Model Inference Table (Table 1)

```
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
```

### Log-Linear Model Inference Table (Table 2)

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

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.47412 -0.08715  0.02902  0.07265  0.39911

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)    5.50113    0.22340  24.624 0.00000000000000976 ***
NumberofRooms  0.10207    0.02196   4.648            0.000230 ***
Location      -0.47848    0.10891  -4.393            0.000397 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2131 on 17 degrees of freedom
Multiple R-squared:  0.8233,    Adjusted R-squared:  0.8025
F-statistic: 39.61 on 2 and 17 DF,  p-value: 0.0000003991
```

#### Joint Significance

An F-test was conducted for both models to determine whether the two independent variables, the number of rooms and location, have 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 $\beta_{\text{Location}} \neq 0$

For either model, the probability (or p-value) of $\beta_{\text{Number of Rooms}} = 0$ and $\beta_{\text{Location}} = 0$ 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 using either of these models with $\beta_{\text{Number of Rooms}}$ and $\beta_{\text{Location}}$ will have better predictive power than if the intercepts were the only parameter.

#### Individual Significance

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

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

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

For either model, the probability (or p-value) of $\beta_{\text{Number of Rooms}} = 0$ 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.

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

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

For either model, the probability (or p-value) of $\beta_{\text{Location}} = 0$ 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.

## Impact of an Interaction Term

Add a joint term, X1X2 into the regression model, and estimate the regression equation again. Report the p-value for the joint term X1X2. What does this imply?

## Impact of an Interaction Term

Add a joint term, X1X2 into the regression model, and estimate the regression equation again. Report the p-value for the joint term X1X2. What does this imply?

## Conclusion

On the basis of these results, indicate the most appropriate regression model for this set of data.

## Appendix

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

Australian Bureau of Statistics Dec-quarter-2024, Total Value of Dwellings, ABS, viewed 30 March 2025, https://www.abs.gov.au/statistics/economy/price-indexes-and-inflation/total-value-dwellings/latest-release.

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.

The formatting of this markdown pdf document has been created with the use of AI tools (OpenAI. 2025).