Investigating the Relationship between House Size (Number of Rooms) and Selling Price in Two Melbourne Suburbs

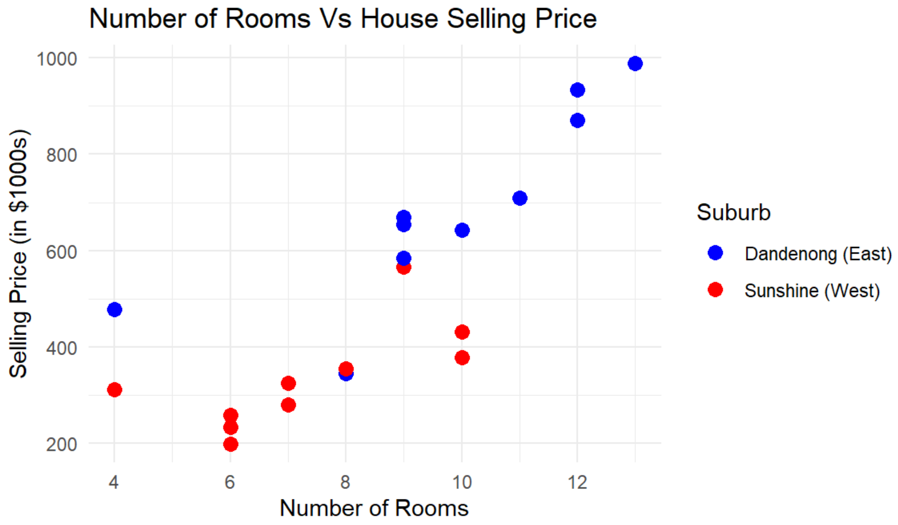
Author	Anirban Chakrabarty
z-ID	z5626947
Course	ZZSC9001 - Foundations of Data Science (H524 Online)
Tutorial Time	10am Au EST
Tutor	Dr Valentyn Panchenko
Date	9/23/2024

1. Method of Sampling

The sampling method used here is ***stratified random sampling***. The population is divided into two distinct strata based on location: East (Dandenong) and West (Sunshine). A random sample of 10 houses was selected from each suburb. Stratified sampling ensures that both regions are equally represented, which is crucial given previous research suggesting differences between the East and West sides of Melbourne. The assumption is that the selected houses are representative of the general population of homes in their respective suburbs.

2. Data Display Method

A scatter plot is the most appropriate method to visualize the relationship between house size (number of rooms) and selling price, with different colors to differentiate between the two locations (East and West). This allows us to observe any potential correlation between the two variables and any differences between the two suburbs.



It shows the relationship between continuous dependent variable (selling price) and independent variable (number of rooms), while also revealing location-based differences. The R code that generates the above scatter plot is attached in the Appendix I.

3. Multiple Regression Equation

The multiple regression model to be estimated is:

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

Where:

- Y is the selling price (in thousands of dollars),
- X_1 is the number of rooms,
- X_2 is the location (0 for Dandenong, 1 for Sunshine),
- β_0, β_1 and β_2 are the regression coefficients,
- ϵ is the error term.

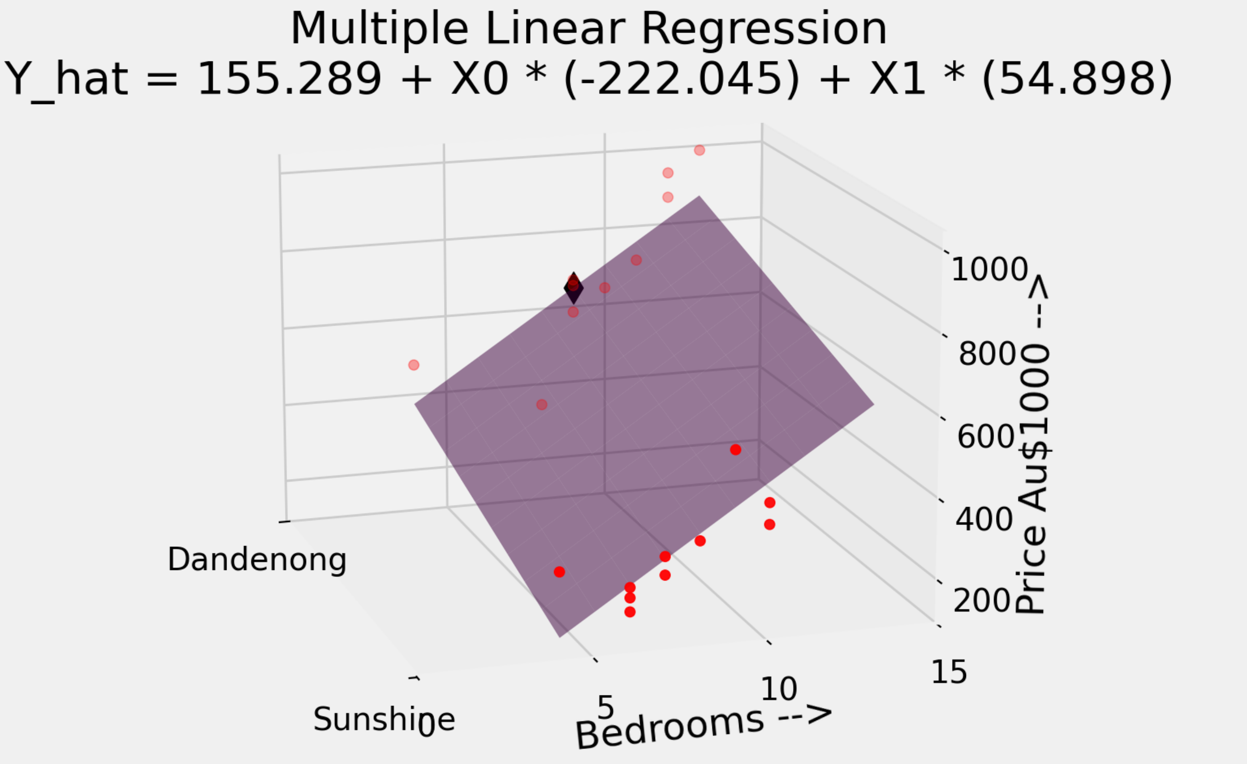
The slope β_1 indicates how much the selling price changes for each additional room, while β_2 represents the price difference attributed to being in Sunshine as opposed to Dandenong.

Using the R code in Appendix I, we find the following:

- Regression coefficient of intercept β_0 = 155.28854
- For number of rooms, the positive coefficient or slope (54.8981) suggests that, on average, each additional room increases the selling price by approximately $54,898.10, holding the location constant.
- For location, β_2 = the negative coefficient (-222.0446) suggests that, on average, houses in Sunshine (marked 1) sell for approximately $222,044.60 less than those in Dandenong (marked 0), holding the number of rooms constant.
- Therefore, the final equation for the prediction model becomes:

  ŷ = 155.28854 + 54.89809 * rooms - 222.04459 * location
- Therefore, the selling price for a house with 9 rooms in Melbourne's east (location = 0) has been predicted to be **$649,371.30**.

The following is the 3D representation of Multiple Linear Regression showing House Prices depending on Location and Number of Bedrooms. Python code in Appendix II.



4. F-Test for Joint Significance

The F-test evaluates whether there is a significant relationship between the selling price and the two independent variables (number of rooms and location). We test the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = 0$$

At a 5% level of significance, we assess whether the F-statistic exceeds the critical value, which would suggest a joint significant relationship. If the p-value is less than 0.05, we reject the null hypothesis and conclude that both variables together significantly affect the selling price. Using the R code in Appendix I, we find the following:

The F-statistic for the model is approximately 42.95, and the p-value is 2.26×10^{-7}

Since the p-value is significantly lower than the 5% significance level (0.05), we reject the null hypothesis and can conclude that there is a jointly significant relationship between the selling price and the two independent variables (number of rooms and location).

5. Significance of Each Independent Variable

For individual variables, we conduct t-tests to determine whether each independent variable significantly affects the selling price at the 5% significance level.

- Null hypothesis for X_1: $\beta_1 = 0$ (Number of rooms has no effect).
- Null hypothesis for X_2: $\beta_2 = 0$ (Location has no effect).

The t-test results and corresponding p-values for both X_1 (number of rooms) and X_2 (location) will reveal whether each variable is significant on its own. If the p-values are below 0.05, we reject the null hypotheses.

From the regression analysis summary (R code in Appendix I) we get the p-values as follows:

1. p-value for location is 0.00574
2. p-value for the number of Rooms is 7.58×10^{-5}

Therefore, as both p-values are less than 0.05, we reject the null hypotheses and conclude that both location and number of rooms significantly affect the selling price of houses at the 5% level of significance

6. Joint Term $X_1 X_2$

We introduce a joint term $X_1 X_2$ to examine whether the relationship between the number of rooms and selling price differs between the two suburbs. The regression equation now becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \epsilon$$

From the R code in Appendix I, the values of the coefficients of the new equation are re-estimated as:

$$Y = 61.997 - 5.011\, X_1 + 185.099\, X_2 - 26.569\, X_1 X_2 + \epsilon$$

The **p-value for the interaction term $X_1 X_2$ is = 0.239620** which is greater than the 0.05 significance level, indicating that the interaction between location and number of rooms does not have a statistically significant effect on the selling price at the 5% level of significance.

Hence, the relationship between the number of rooms and the selling price is consistent across both suburbs and does not significantly affect the selling price of houses.

7. Most Appropriate Model

The following could be concluded, based on the results of the regression analysis, including the significance tests for the individual coefficients and the interaction term. If the interaction term is not significant, the simpler model without the interaction must be preferred.

Y = 155.28854 + 54.89809 * rooms - 222.04459 * location + ϵ

Appendix

Following code files are separately submitted.

Appendix I

R code to generate:

1. Scatter Plot between house size (number of rooms) and selling price.
2. Regression model summary, prediction, F-test, p-values and interaction terms.

01RegressionR.R

Appendix II

Python code for 3D Multiple Linear Regression.

02MultiLinReg3DVisPy.py