THE RELATIONSHIP BETWEEN THE SIZE OF A FAMILY HOUSE AND SELLING PRICE OF THE HOUSE IN THE GREATER MELBOURNE AREA

Introduction

This is a report addressing a commercial consulting project to study the relationship between the size of a family house (measured by the number of rooms) and selling price of the house in the greater Melbourne area. The research methods, statistics models and framework used in this project will be discussed in the following section and findings will be presented in accordance with the attached Appendix. All tables, graphical demonstrations and calculations were also presented in the Appendix.

Methods

To investigate the relationship, a dataset consisting of selling prices of 20 houses with varying number of rooms was established. Then several linear Regression models were adopted to test whether the number of rooms of a house would impact its selling price or whether there were other contributing factors. The hypothesis tests utilised in this report included T-tests - to test significance of the independent variables; and F-test - to test the joint effects to the prediction.

Findings

The housing data were collected by choosing two representative suburbs (strata) - Dandenong and Sunshine West, and randomly selecting 10 properties from each of the two suburbs. This method was named Stratified Random Sampling (Sharpe et al., 2015, p. 278). One of the major benefits of Stratified Random Sampling is that it reduces sampling variability by having restrictions between strata and assisted in creating samples that had similar components as the population to alleviate sampling biases.

A single factor linear regression analysis was conducted as per graph 1-1 of the Appendix. The upward-sloping line of best fit indicated positive relationship between number of rooms and house selling price. However, in the sample, houses in Dandenong with four rooms and nine rooms were more expensive than their counterparts in Sunshine West respectively. After differentiating houses between the two suburbs using different colour and shape as shown in graph 1-1, one could see that the majority of houses in Dandenong were plotted on the top half of the graph, indicating generally higher prices than houses in Sunshine West. This could be further illustrated in graph 1-2 of the Appendix using a boxplot.

To test whether the location, together with number of rooms, affects house selling prices, a multiple regression analysis was conducted with the location and number of rooms being the independent variables plotted against the actual selling prices. Reading from the constructed summary table for the model in the Appendix, a regression equation could be deduced as follows:

$$\textit{Predicted price} = 155.29 - 222.04 \times \textit{location parameter} + 54.9 \times \textit{number of rooms}$$

In the above equation, the slope -222.04 indicated that, when sampling a property in Sunshine West (parameter 1), the predicted house price would reduce by 222.04 (in thousands of dollars). The slope 54.90 indicated that, with one additional room, the predicted house price

would increase by 54.90 (in thousands of dollars). For example, a property with 9 rooms in Dandenong, as calculated in the Appendix, would have a predicted house price of $649,371.

In addition to this, a F-test that test whether a jointly significant relationship between selling price and the two independent variables exists was also conducted. The 42.95 F-statistic and the 2.257 × 10!" p-value for this multiple regression model indicated that at the 5% level of significance, the null hypothesis should be rejected and the alternative hypothesis - a jointly-significant relationship between selling price and the two independent variables does exist, should be accepted (p-value < 0.05) (James et al., 2013, p. 68).

Lastly, the summary table listed the respective p-values from the t-test conducted for both of the independent variables. At the 5% level of significant, the p-value of 5.74×10^{-4} for the variable 'location' and 7.58×10^{-5} for the variable 'number of rooms' both indicated that, the alternative hypothesis that each independent variable had contributed to the predicted selling price, should be accepted (p-values for both were smaller than 0.05).

To further deduce whether there was interaction effect between the two independent variables so that the effect of one of the variable would towards the dependent variable is dependent on the other independent, a joint term representing the multiplication of the two independent variables could be added to construct a new multiple regression model. Refer to the Appendix for the summary of the new model. Based on the coefficients of the model, the equation for the new model could be interpreted as follows:

$$Predicted\ price$$

$$= 61.997 - 5.011 \times location\ parameter + 64.516 \times number\ of\ rooms - 26.569 \times location\ parameter \times number\ of\ rooms$$

After also conducting a t-test to the new model, the p-value for the joint term was 0.23962. This was insignificant at the 5% level of significance. Consequently, a null hypothesis that, the effects of number of rooms in a house on the selling price is not dependent on the location of the house could not be rejected (Jaccard, J and Turrisi, R, 2003). The result of this implied that the new model with a joint term was not an appropriate model to be adopted in this scenario, due to the lack of evidence that the interaction effect between the two independent variables exists.

To further validate the appropriateness of the original multiple regression without the joint term, a Residual Plot and a Residuals Normal Quantile-Quantile (Q-Q) Plot were conducted as per graphs 1-3 and 1-4 in the Appendix. By observing the data in the Residual Plots, no obvious patterns or heteroskedasticity could be concluded and the distributions of residuals were plotted approximately evenly above and below the 0 reference line, indicating a fitted regression line (Sharpe et al., 2015, p. 590). What's more, residual plots in the Residuals Normal Q-Q Plot were distributed approximately around the diagonal line, indicating the near normality of the residuals' distributions (Sharpe et al., 2015, p. 591). The above attributes, together with the multiple r-squared statistic of 0.835 made this a relatively accurate regression model to use to interpret the relationship of the subjects.

Conclusion

To summarise, the regression model with number of rooms and location as independent variables was an appropriate model to be adopted to predict the house selling price. Apart from implying a positive correlation between selling price and the number of rooms, the

model also concluded that the location of the house was a contributor to the selling price (houses in Dandenong were priced at premiums over houses in Sunshine West).

Reference List

Jaccard, J & Turrisi, R, 2003, 'Interaction effects in multiple regression', Quantitative applications in the social sciences, 2nd edn, SAGE Publications, Inc., Thousand Oaks, California, viewed 6 June 2020,
< https://dx-doi-org.wwwproxy1.library.unsw.edu.au/10.4135/9781412984522.n2>
James, G, Hastie, T, Tibshirani, R & Witten, D, 2013, An Introduction to Statistical Learning : with Applications in R, Springer, New York.
Sharpe, NR, De, VRD, & Velleman, P 2015, Business Statistics, Global Edition, Pearson Education Limited, Harlow, United Kingdom.