Individual Assignment 4, Dion Chang - 20812576

1) Problem statement and data used: The objectives with the Melbourne housing data set were to determine the important factors that determine house price as well as create and compare two multiple linear regression models that predict house price. Specifically, there is a budget of 1.2 million dollars and the suburbs of interest were Brunswick, Carlton, South Yarra, and Fitzroy. The filtered building size was greater than 31 m^2 to less than 150 m^2. The type of residential of interest was unit.

2) Planning: Filters were applied as per the conditions listed in **1)**. Missing values in building area and car spaces have also been removed. After going through with the data filtering and cleansing, the dataframe would have 97 observations. This dataframe would then be used to create two different models (models 1 and 2). Model 1 analyzes the influence of suburb, number of car spots, and building area size have on unit prices. In model 2, more variables were added to analyze the influence of unit prices: suburb, number of car spots, building area sizes, number of rooms, bedrooms, and bathrooms. Since the Suburb predictor variable was the only categorical (the rest are quantitative), dummy variables for all suburbs of interest: Brunswick, Carlton, South Yarra, and Fitzroy were created for the models.

Several assumptions need to be checked prior to analysis: normality, linearity, homoscedascity, and multicolinearity. Residual tests (residuals vs. fitted, Q-Q plot, Durbin-Watson, Cook's Distance) and VIF were used to perform assumption tests.

3) Analysis: A summary of Model 1 is presented in Table 1. 2.5% and 97.5% are the 95% confidence interval ranges of the coefficients for each variable. Brunswick, number of car spaces, and building area size are the three predictor variables that have an influence on unit prices at the 5% level of significance as per their p values. The R^2 is .68, and adjusted R^2 is .66.

Table 1. Multilinear regression of model 1.

Variables	Coefficient	p-value	2.5%	97.5%
(Intercept)	13952	.80	-97228.39	125132.28
brunswick compared to fitzroy	-78326	.03	-148878.05	-7774.42
carlton compared to fitzroy	-4262	.93	-97643.33	89119.12
south yarra compared to fitzroy	-7398	.82	-72619.03	57823.02
car spots	102757	< .001	48998.41	156516.38
building area size	7331	< .001	6059.37	8601.82

A summary of Model 2 is presented in Table 2. The factors that had the most impact on unit prices were determined to be Brunswick, number of car spaces, and building area size. This was the same as model 1. The R^2 however is .71, and adjusted R^2 is .69 which are higher than model 1.

Table 2. Multilinear regression of model 2.

Variables	Coefficient	p-value	2.5%	97.5%
(Intercept)	-46628.4	.44	-166802.39	73545.61
brunswick compared to fitzroy	-88435.1	.01	-159098.55	-17771.67
carlton compared to fitzroy	-36165.2	.44	-129208.77	56878.32
south yarra compared to fitzroy	-10575.4	.74	-73469.73	52318.94

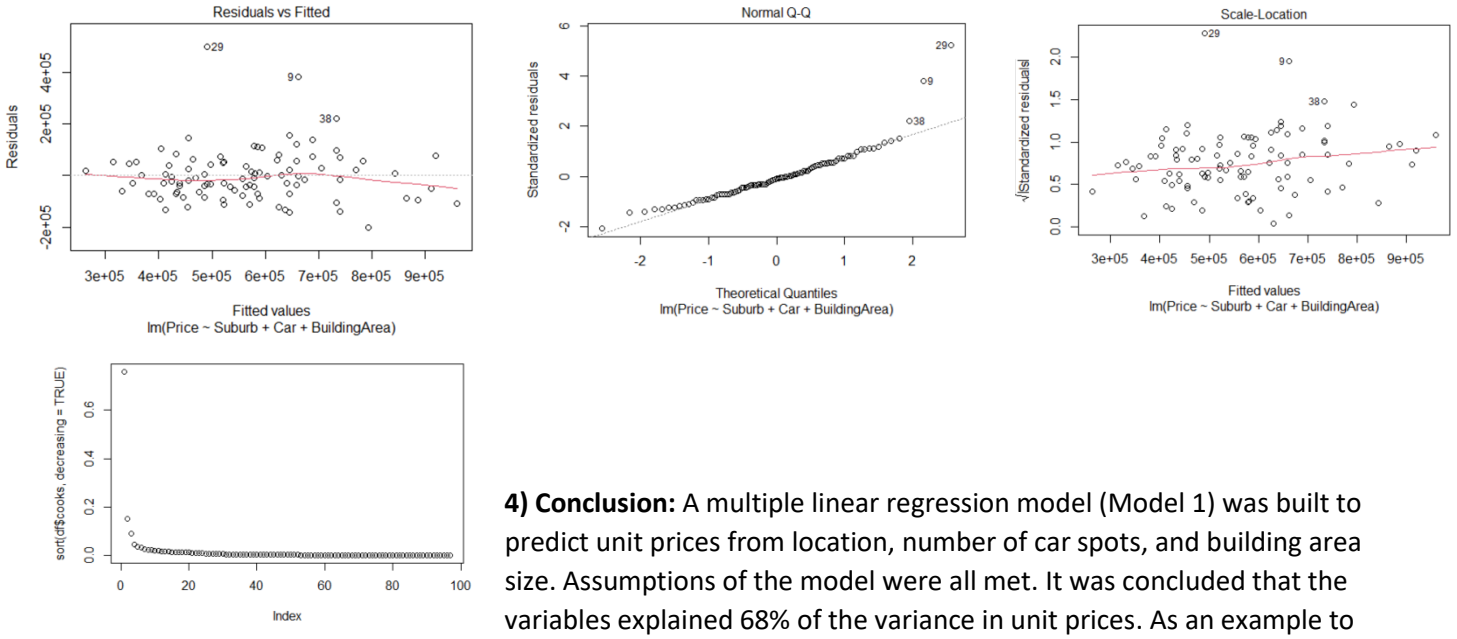
building area size	5455	< .001	3663.55	7246.52
rooms	25668.7	.62	-75802.32	127139.81
bedrooms	43156.7	.35	-47976.5312	134289.853
bathrooms	73210.3	.05	-144.1167	146564.758
car spaces	95477.2	< .001	43760.3165	147194.093

According to the ANOVA test, at 5% level of significance models 1 and 2 are different (p = .01). Since the same variables were determined to be significant in influencing unit price for both models, model 1 will be the final model and assumption tests will be performed only on Model 1 in this report.

The VIF (Variance Inflation Factor) was inspected to check for multicollinearity. The largest VIF was 2.44, less than 10; the average VIF was 1.67, which can be considered close to 1. The lowest tolerance (1/VIF) was .41, which is much greater than 0.2. It can be concluded that there is no collinearity in the data.

Outliers: Four residuals were found to be above or below 1.96 standard deviations. This represents 4% of the observations, which was expected if the residuals are normal (5% of the data expected to be outside the 2 standard deviations). Therefore, these observations are not considered to be outliers and continued with all 97 observations included in Model 1. From the **Durbin-Watson test**, the independent errors were not significant at the 5% level of significance (d = 2.18, p = .49). As d is close to 2, it can be assumed that the residuals or errors are independent. Figure 1 are plots for the assumption tests. Residual vs. fitted, Q-Q, and Scale Location plots showed that the residuals are homoscedastic and linear. Cook's distance was a maximum of .759, below the cutoff value of 1. Therefore, no influential cases.

Figure 1. Residuals vs. Fitted, Q-Q, Scale-Location, and Cook's Distance plots from R for assumption tests for model 1.



4) Conclusion: A multiple linear regression model (Model 1) was built to predict unit prices from location, number of car spots, and building area size. Assumptions of the model were all met. It was concluded that the variables explained 68% of the variance in unit prices. As an example to predict the influence of the variables on house price, increasing the building area by 1 m^2 increases the unit price by \$6059.37 – \$8601.82 (95% of the time).