Problem Statement:

When people look to start a family, buying a house always come as a priority although it is a huge investment for everyone. This report aims to predict the price of the houses in Melbourne by using Rooms, Bathroom, Distance from Central Business District (CBD), Land size of the house and Building area of the house.

Data:

The dataset used in this analysis is the subset of the original data retrieved from Kaggle and was scraped by Tony Pino from Domain's website. The original data contains 13,580 records with 21 variables. It contains more information about houses sold from January 2016 to the end of 2018. Also, there are several attributes of the houses in Melbourne along with their prices. The dataset provided is a fabricated dataset by using real data after a number of other analyses for the population distributions of the variables included in the dataset. For the purpose of this report, the data is filtered to 586 records with 21 variables. Below are the variables, which we will analyse in this report:

Rooms: Number of rooms
Bathroom: Number of bathrooms in each house
Distance: Distance of each house from Melbourne CBD (in Km)
Price: Price of house that was sold
Landsize: Land Size
BuildingArea: Building Size

Planning:

Both the models are built using Multiple Linear Regression and will be compared through ANOVA Test. All the missing values have been removed from the dataset. Prices for the regions "Western Metropolitan" and "Eastern Metropolitan" will be predicted. Land sizes less than 1,000 with no more than 5 bedrooms and bathrooms will be considered, so the dataset tidy up accordingly. There are 586 data points that will be considered for modelling.

Assumptions:

- The predictor variables are quantitative and the outcome variable is quantitative, continuous, and unbounded for both the models.
- Through the variances value, I can clearly say that the data meets the assumption of non-zero variances for both the models.
- Multicollinearity can be verified through VIF Factor. The largest VIFs are 1.18 and 1.51 which are less than 10; the average vifs are 1.12 and 1.39, close to 1. The lowest tolerances (1/VIF) are 0.84 and 0.65. Thus, we can say there is no collinearity in our data.
- Through the "Residual Vs Fitted" plots shown below in the Appendix (***Figure 01***) and (***Figure 03***) for both the model, the data meets the Assumptions of Linearity. Also, the residuals are equally spread above and below the regression line in "Residual Vs Fitted" plots, so both the models meet the Assumption of homoscedasticity.
- Through the visual inspection of the QQ plots shown in Appendix (***Figure 02***) and (***Figure 04***), it shows the residuals for both the models follow a normal pattern.
- From the output of Durbin-Watson test it is clear that that the both the models do meet the assumption of independent errors as the p value is greater than alpha value (0.05) value and DW test statistic is close to 2 for both the Models. Please refer to the Appendix (***Figure 10***)

Outliers:

The 23 residuals are above or below 1.96 standard deviations. As this represents 3.92% of the observations, I do not consider any of these observations as outliers and continued with all 586 observations included in the model.

Influential Cases:

Cook's distance was a maximum of 0.3384975 and 0.1415757 for Model 1 and Model 2 respectively, far below the value of 1.

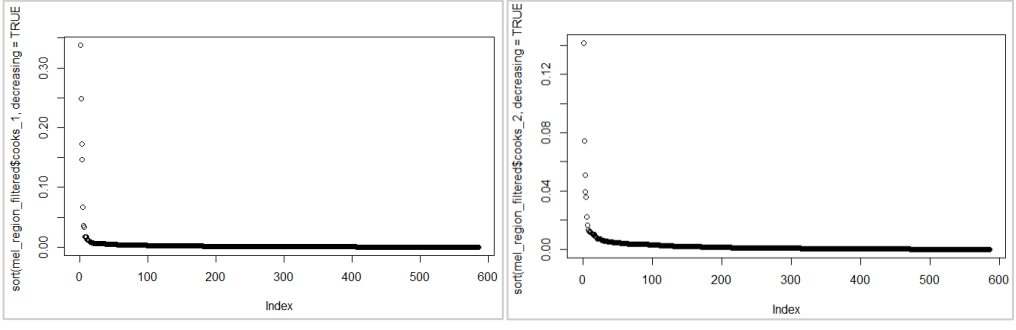

Figure 05: Cook's distance graph for Model-1 and Model-2 respectively

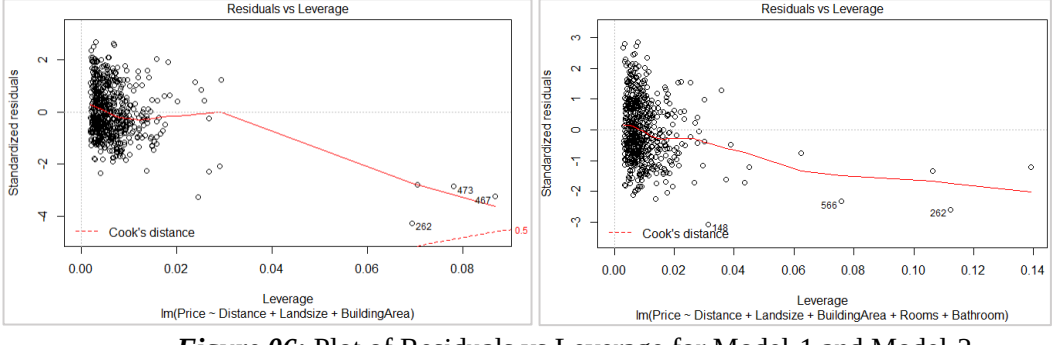

Figure 06: Plot of Residuals vs Leverage for Model-1 and Model-2

There are no values fall outside the bands; therefore, no evidence of influential cases.

Analysis of Model-1

All 3 predictor variables have an influence on Price of the House at the 5% level of significance. Since R^2 value is 0.27, we can say 27% change in the price of the house can be predicted by Distance form CBD, Land Size and Building Area. Please refer to the Appendix (***Figure 07***)

Analysis of Model-2

All 5 predictor variables have an influence on Price of the House at the 5% level of significance. Since R^2 value is 0.38, we can say 38% change in the price of the house can be predicted by Distance form CBD, Land Size Building Area, Rooms and Bathrooms. Please refer to the Appendix (***Figure 08***)

Model Comparison

The ANOVA test result shows significant result ($F(582,582)=51.43$, $p < 2.2e-16$) at 5% Level of significance. This means adding the 2 variables (Rooms and Bathroom) to the model made it a better model as compared to Model-1. Please refer to the Appendix (***Figure 09***)

Conclusion

Model-2 is a better model as compared to Model-1. Hence, we say that Rooms and Bathrooms does add significance to the Model-1. Distance from city center, Rooms and Bathrooms are the three most important factors in determining the Price of the house. Price of the house tends to decreases by $64,775.79 for every 1 km from Melbourne city center.

Appendix:

Bibliographic Reference

Hardin, J. (2021) *Example R code / analysis for housing data* Retrieved from https://pages.pomona.edu/~jsh04747/courses/math58/Final_examp.html

Srun, K. (October 24, 2019) *RPubs – Housing price in Melbourne* Retrieved from https://rpubs.com/Vatey/542526

Hasnain, A. (2021) *RPubs – Melbourne Housing Price* Retrieved from https://rpubs.com/The_Analyst/431304

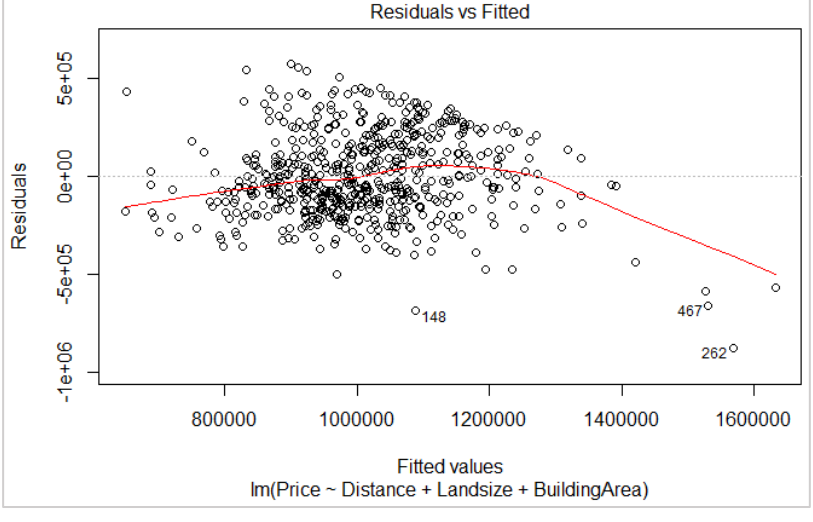

Figure 01: Plot of Residuals vs Fitted

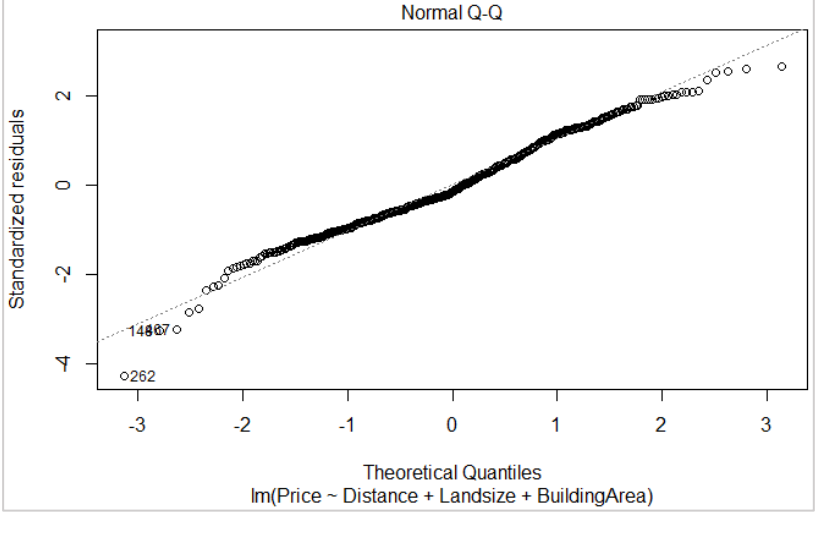

Figure 02: Normal Q-Q plot of Model-1

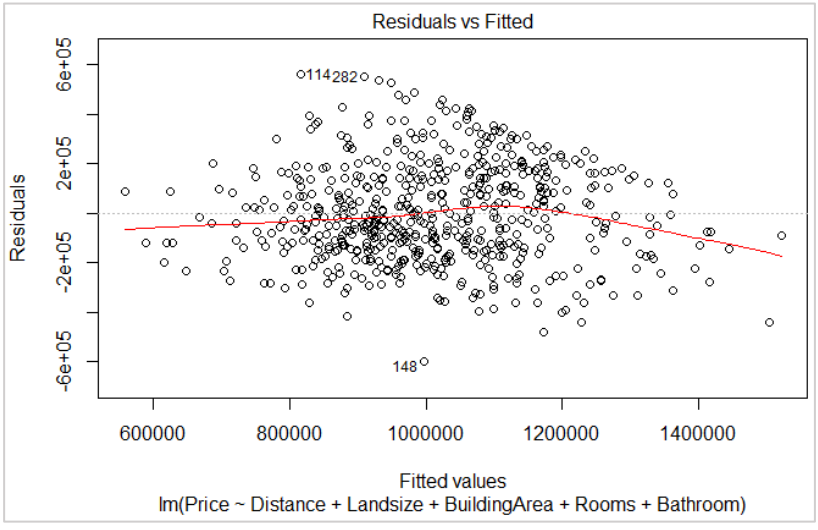

Figure 03: Plot of Residuals vs Fitted for Model-2

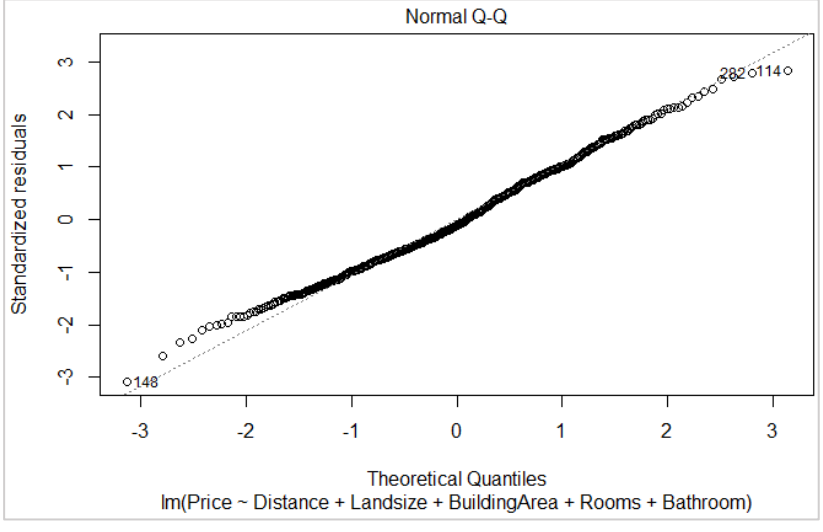

Figure 04: Normal Q-Q plot of Model-2

```
Call:
lm(formula = Price ~ Distance + Landsize + BuildingArea, data = mel_region_filtered)

Residuals:
    Min      1Q  Median      3Q     Max
-878739 -146833  -30424  151984  569241

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1090640.73   54994.05  19.832   <2e-16 ***
Distance      -63003.06    7413.64  -8.498   <2e-16 ***
Landsize         525.95      57.41   9.161   <2e-16 ***
BuildingArea    1319.34     149.59   8.820   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 213900 on 582 degrees of freedom
Multiple R-squared:  0.2728,    Adjusted R-squared:  0.269
F-statistic: 72.76 on 3 and 582 DF,  p-value: < 2.2e-16
```

Figure 07: Summary of Model 1

```
Call:
lm(formula = Price ~ Distance + Landsize + BuildingArea + Rooms +
    Bathroom, data = mel_region_filtered)

Residuals:
    Min      1Q  Median      3Q     Max
-598216 -139860  -20059  140311  559554

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  847400.72   56941.87  14.882  < 2e-16 ***
Distance     -64775.79    6852.01  -9.454  < 2e-16 ***
Landsize        468.30      55.44   8.447 2.41e-16 ***
BuildingArea    412.31     165.88   2.486   0.0132 *
Rooms         92728.50   13705.98   6.766 3.25e-11 ***
Bathroom      82713.17   16441.79   5.031 6.53e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 197500 on 580 degrees of freedom
Multiple R-squared:  0.3823,    Adjusted R-squared:  0.377
F-statistic:  71.8 on 5 and 580 DF,  p-value: < 2.2e-16
```

Figure 08: Summary of Model 2

```
Analysis of Variance Table

Model 1: Price ~ Distance + Landsize + BuildingArea
Model 2: Price ~ Distance + Landsize + BuildingArea + Rooms + Bathroom
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    582 2.6633e+13
2    580 2.2621e+13  2 4.0119e+12 51.433 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 09: ANOVA test results

```
## Durbin-Watson Test - Model_1

```{r echo=FALSE}
library(lmtest)
dwtest(mel_model1)
```


        Durbin-Watson test

data: mel_model1
DW = 1.9186, p-value = 0.1613
alternative hypothesis: true autocorrelation is greater than 0

Durbin-Watson Test - Model_2

```{r echo=FALSE}
library(lmtest)
dwtest(mel_model2)
```


        Durbin-Watson test

data: mel_model2
DW = 1.8953, p-value = 0.102
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 10: Results of Durbin-Watson Test for Model-1 and Model-2