Individual Assignment 4

Ali Rafieepouralavialavijeh, 20900871

Problem statement:

Dear manager, in this analysis, the dataset that I am focusing on has 13580 observations, my focus would be on "***If we were to prioritize between Type and Suburb features in our investment analysis, which one outperforms the other?***". This analysis is significant because it helps identify and prioritize essential decision factors in future investments. For this purpose, I have selected a subset of features with significant correlation with the houses' price (-*suburb*, -*Room*, -*Type*, -*bathroom*, and -*YearBuilt*) I have done a data cleaning process to identify outliers and Null values in the dataset. As a result, I noticed an abnormal data entry for YearBuilt=1196, so this data entry is omitted throughout this analysis. Besides, Null values are omitted, which results in 8204 observations.

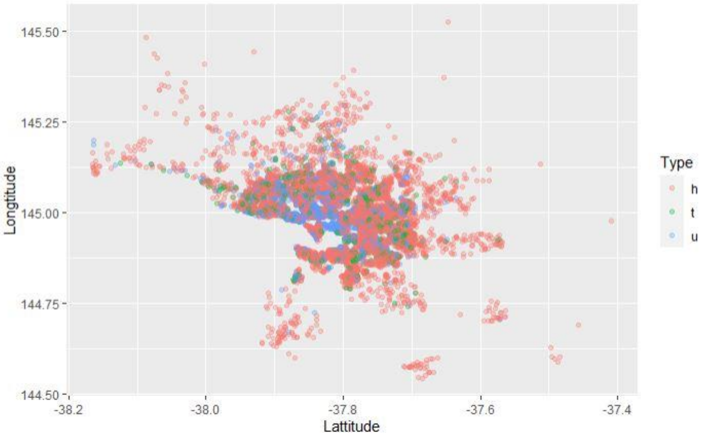


Figure 1 Distribution of houses based on the geographical position and the type of the house

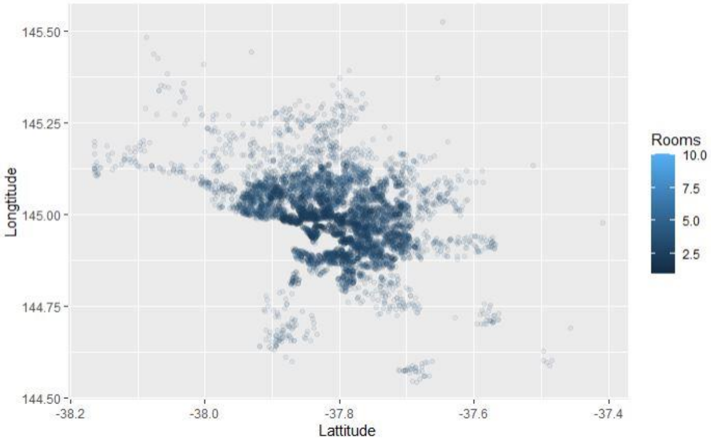


Figure 2 Distribution of houses based on the geographical position and the number of rooms in each house

Planning:

In this analysis, three models are built to study the significance of either "Suburb" and "Type" features. For having a more reliable regression model which generalizes well, certain assumptions should be checked:

1. All predictor variables must be quantitative or categorical, and the outcome must be quantitative, continuous, and unbounded *(all predictors are either quantitative or categorical, and the outcome is quantitative, continuous, and can be considered unbounded).*
2. The variance should be non-zero *(can be easily verified, the variance is non-zero)*
3. No perfect multicollinearity, predictor variables should not correlate highly *(visual inspection shows no perfect multicollinearity, but this is verified using VIF test)*
4. Predictors should be uncorrelated with external variables *(while there could be many factors involved in prices, I can assume that this assumption holds)*
5. The residuals should be normal, homoscedastic, and independent *(analysis will follow)*

The Durbin-Watson test results for a 5% significant level shows enough evidence to reject the hypothesis of independence of the residuals. Figure 3 shows a linear relationship in residuals. Figure 4, along with the Shapiro-Wilk test of normality, shows that data is not normal. VIF test for influential points shows no influential points, and no outliers are affecting the model results (Cook's distance (Figure 5) was a maximum of 0.15, far below the chosen cutoff value of 1. We thus conclude that there is no compelling case. In addition, only 364 of observations are beyond 1.96 confidence interval which indicates that the model has no outliers). [For complete results please refer to the Appendix]

Analysis:

The linear regression model built to analyze the significance of either “Type” or “Suburb” shows that all predictor variables influence the price at the 5% level of significance.

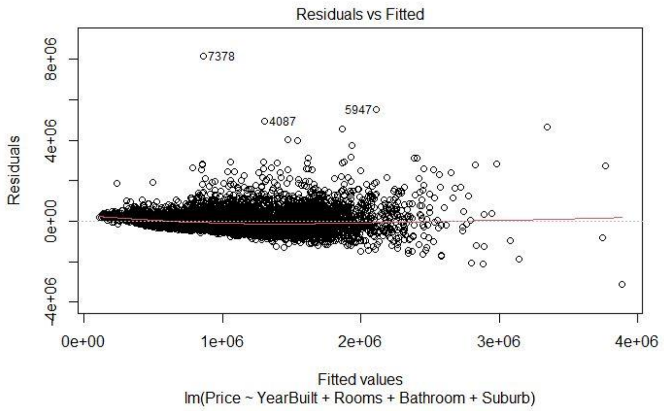


Figure 3 Distribution of residuals vs fitted values

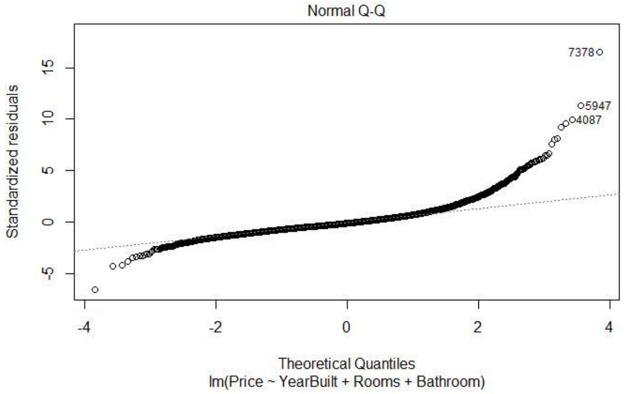


Figure 4 QQ plot, showing normality of the data

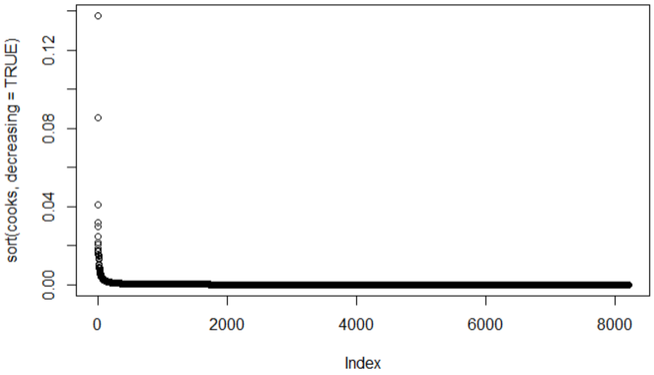


Figure 5 Cook distance plot for finding influential points

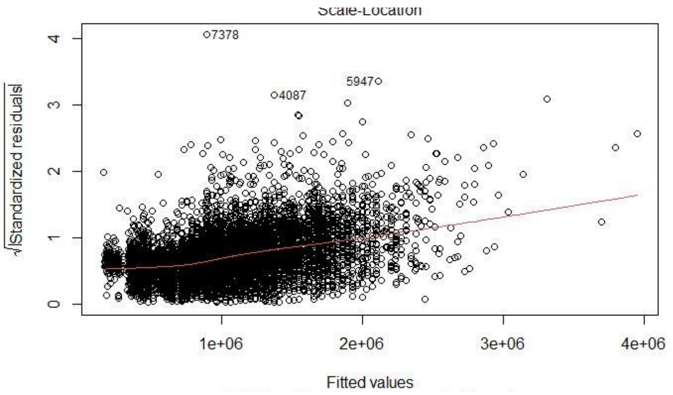


Figure 6 Distribution of standardized residuals in comparison to fitted values

The F-statistic and its corresponding p-value for each of our models give us enough evidence to make sure that our models outperform the simple mean. This basically means, at 5% significance interval I have enough evidence to conclude the models built till now perform better than the simple mean.

Now that the models are built, an ANOVA test can compare them with one another. The results for comparing the model without the “Type” feature show that at a 5% significance level, we have enough evidence that our model with “Type” does better than without it. The comparison results without the “Suburb” feature show that likewise, there is enough evidence that the model with the “Suburb” feature does better than without it. However, when the two models “with Type feature and without suburb” and “with Suburb feature and without Type” are compared together, the results indicate that they are not significantly different from one another.

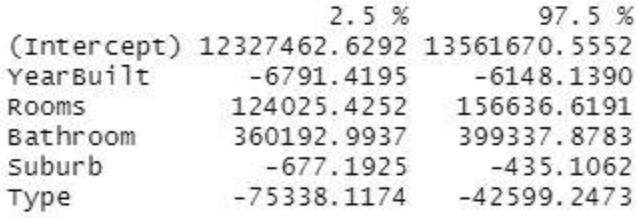
Conclusion:

The answer to this analysis's research question is that no one of these features without the other one does better and as such, their effects on price are similar at a 5% significance level. Further analysis shows that considering both features improve the model performance. Please see managerial insights about this model in appendix.

In terms of the generalizability of the models that are evaluated (whose quality is not a direct goal of this analysis), they do not generalize very well because some of their assumptions are violated. As such, they cannot be relied on for accurate predictions of prices and possibly more complex methods should be applied.

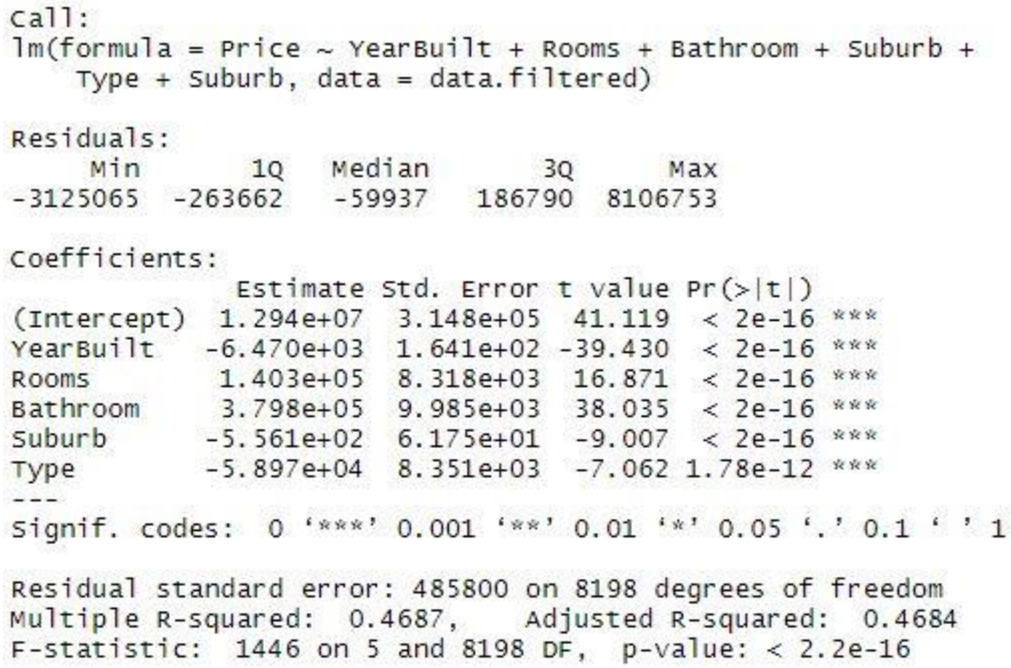
<u>Appendix</u>:

Managerial insights



```
                    2.5 %        97.5 %
(Intercept) 12327462.6292 13561670.5552
YearBuilt      -6791.4195    -6148.1390
Rooms         124025.4252   156636.6191
Bathroom      360192.9937   399337.8783
Suburb          -677.1925     -435.1062
Type          -75338.1174   -42599.2473
```

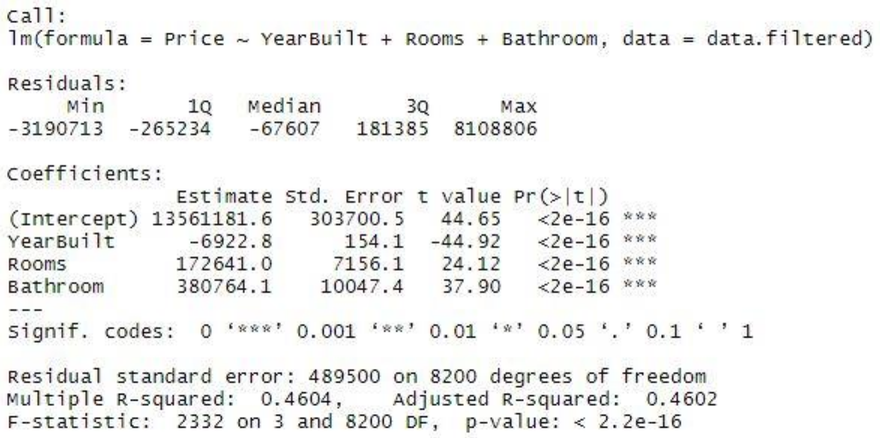
The confidence interval of the proposed model in conclusion is depicted above.



```
Call:
lm(formula = Price ~ YearBuilt + Rooms + Bathroom + Suburb + 
    Type + Suburb, data = data.filtered)

Residuals:
     Min       1Q   Median       3Q      Max 
-3125065  -263662   -59937   186790  8106753 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.294e+07  3.148e+05  41.119  < 2e-16 ***
YearBuilt   -6.470e+03  1.641e+02 -39.430  < 2e-16 ***
Rooms        1.403e+05  8.318e+03  16.871  < 2e-16 ***
Bathroom     3.798e+05  9.985e+03  38.035  < 2e-16 ***
Suburb      -5.561e+02  6.175e+01  -9.007  < 2e-16 ***
Type        -5.897e+04  8.351e+03  -7.062 1.78e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 485800 on 8198 degrees of freedom
Multiple R-squared:  0.4687,	Adjusted R-squared:  0.4684 
F-statistic:  1446 on 5 and 8198 DF,  p-value: < 2.2e-16
```

It can be shown that every increase in the building's age affects the price negatively for -6470 dollars. Every single additional room adds 1403 dollars to the value of the house. Every extra bathroom in the house adds 3798 dollars to the price. So, if we were to invest in increasing our profits from our house's price, bathrooms would be a better choice than the rooms. Investing in short periods is also encouraged if the money-time value is not considered because annually, the prices face reduction.

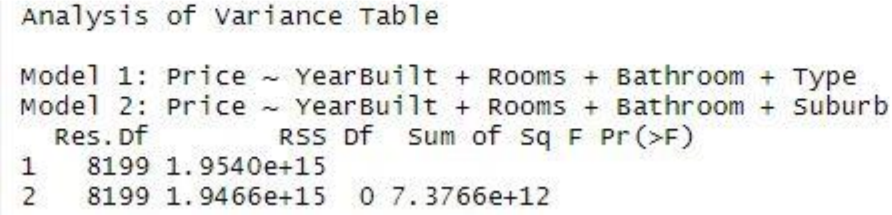
Here is an example of two model comparisons done in this analysis:



```
Call:
lm(formula = Price ~ YearBuilt + Rooms + Bathroom, data = data.filtered)

Residuals:
     Min       1Q   Median       3Q      Max 
-3190713  -265234   -67607   181385  8108806 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13561181.6   303700.5   44.65   <2e-16 ***
YearBuilt      -6922.8      154.1  -44.92   <2e-16 ***
Rooms         172641.0     7156.1   24.12   <2e-16 ***
Bathroom      380764.1    10047.4   37.90   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 489500 on 8200 degrees of freedom
Multiple R-squared:  0.4604,	Adjusted R-squared:  0.4602 
F-statistic:  2332 on 3 and 8200 DF,  p-value: < 2.2e-16
```

All features have a significant effect on the price given their p-values.

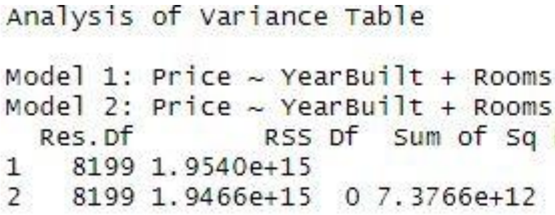
ANOVA test done to compare the significance of “Type” and “Suburb” features:



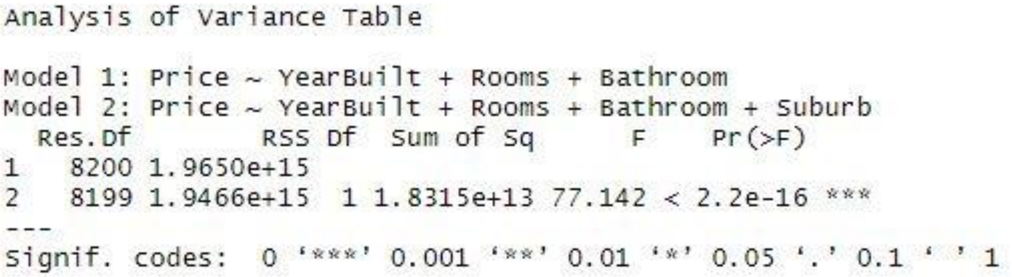
```
Analysis of Variance Table

Model 1: Price ~ YearBuilt + Rooms + Bathroom + Type
Model 2: Price ~ YearBuilt + Rooms + Bathroom + Suburb
  Res.Df        RSS Df  Sum of Sq F Pr(>F)
1   8199 1.9540e+15
2   8199 1.9466e+15  0 7.3766e+12
```

It shows that the difference between the two models are insignificant.

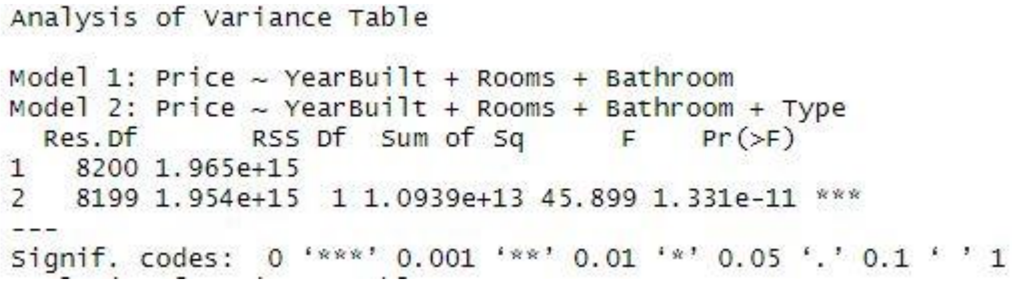
Complete comparison of the models:



```
Analysis of Variance Table

Model 1: Price ~ YearBuilt + Rooms + Bathroom + Type
Model 2: Price ~ YearBuilt + Rooms + Bathroom + Suburb
  Res.Df        RSS Df  Sum of Sq F Pr(>F)
1   8199 1.9540e+15
2   8199 1.9466e+15  0 7.3766e+12
```



```
Analysis of Variance Table

Model 1: Price ~ YearBuilt + Rooms + Bathroom
Model 2: Price ~ YearBuilt + Rooms + Bathroom + Suburb
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1   8200 1.9650e+15
2   8199 1.9466e+15  1 1.8315e+13 77.142 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
Analysis of Variance Table

Model 1: Price ~ YearBuilt + Rooms + Bathroom
Model 2: Price ~ YearBuilt + Rooms + Bathroom + Type
  Res.Df       RSS Df  Sum of Sq      F    Pr(>F)
1   8200 1.965e+15
2   8199 1.954e+15  1 1.0939e+13 45.899 1.331e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```