MSCI718 2023W Individual Assignment 3 -Multiple Linear Regression

Data description

The Melbourne housing dataset contains 13580 observations across 21 variables. The focus of study is on, "**Which feature outperforms the other in investment analysis if we prioritize "Type" and "Region name"?**". This research is significant because it helps in the identification and prioritization of critical decision variables in future investments. As a result, a subset of five features with a strong correlation, and that may have the greatest impact on home prices are selected are shown below:

- **Price** (Price of property in dollars): (interval, continuous) output variable, **Bathroom** (Number of Bathrooms): (ordinal or continuous, coded as integers), **Room** (Number of Rooms): (ordinal or continuous, coded as integers), **YearBuilt** (Year of Built of property): (interval, integers), **Distance** (Distance from CBD): (interval, integers), **Landsize** (Land size of property): (interval, integers).
- **Type** (Type of Real Estate): (categorical, ordinal) h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse, **Regionname** (Name of Region): (categorical, ordinal) General Region (Eastern Metropolitan, Eastern Victoria, Northern Metropolitan, Northern Victoria, South-Eastern Metropolitan, Southern Metropolitan, Western Metropolitan, Western Victoria)

Visual inspection of the Boxplot and inspection of the minimum and maximum datapoints revealed some outliers. There were some abnormal values at YearBuilt=1196 and Landsize=37000 that were influencing the model, so outliers and abnormal values (shown in Appendix 1 and 2) that are outside 2 times standard deviation of the mean are removed, and some null values are omitted, resulting in 8205 observations in the final dataset.

Planning

In this analysis, three models are constructed to evaluate the significance of the "Region name" and "Type" features. Certain assumptions should be checked in order to have a more reliable regression model that generalizes well:

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 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*)

Table 1: Model Assumptions Test

Test/Model	Without Type and Regionname	Model with Type	Model with Regionname
Multiple R-squared	0.4986	0.5202	0.5963
Durbin-Watson Test	1.4774, p=0	1.369, p=0	1.6834, p=0
Max Cook's Distance	0.0097	0.0115	0.0159
Max VIF	1.784	1.498	1.364
Min 1/VIF	0.5605	0.6674	0.7331
Residual Outliers	361	343	330

The Durbin-Watson test for independent errors was not significant at the 5% level of significance. We assume that the data was sampled independently because **d is close to 2 (indicating no autocorrelation)**. Figures 1 and 4 show that **neither model has a linear relationship in residuals**. The **QQ-plots** in Figures 2 and 5, show that **residuals are not normal** and Figure 3 and Figure 6 explains both **models are not Homoscedastic**. However, it can be noted that model with Region name is more ideal than model with Type feature.
The VIF test for influential points reveals **no influential points, no collinearity** in the data (as max VIF is less than 10 and min 1/VIF is greater than 0.2), and **no outliers influencing model** results (Cook's distance is well below the chosen cutoff value of 1). As a result, we conclude that there is no compelling case.
In addition, only 343 residuals (model with Type) and 330 residuals (model with Region name) are greater than or less than 1.96 standard deviations. We **do not consider any of these as outliers** because they represent approximately 4% of the observations (8205). (till 5% can be considered)

Residuals plot of Model with Type category lm(Price ~ Rooms+Bathroom,+YearBuilt+Landsize+Distance+Type)

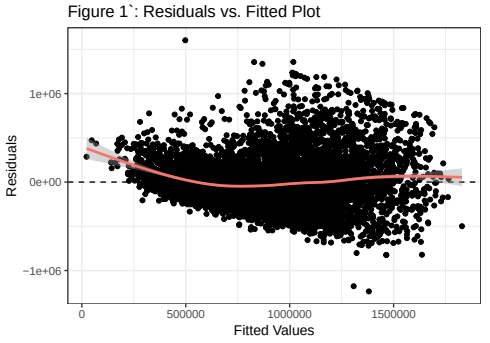


Figure 1`: Residuals vs. Fitted Plot

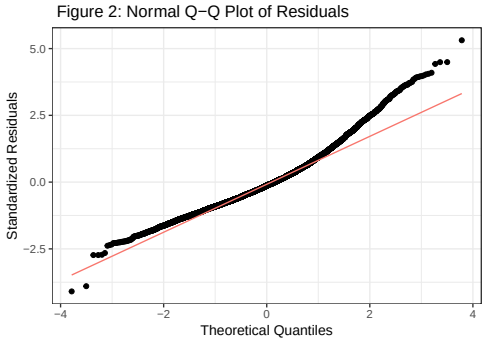


Figure 2: Normal Q−Q Plot of Residuals

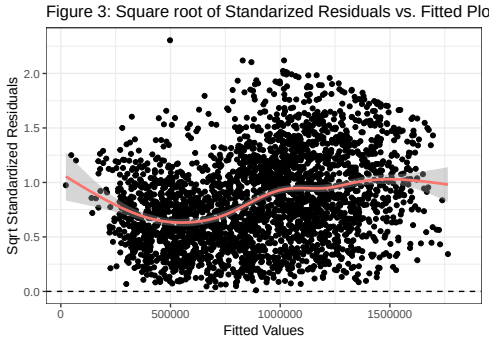


Figure 3: Square root of Standarized Residuals vs. Fitted Plo

Residuals plot of Model with Region name category lm(Price ~ Rooms+Bathroom,+YearBuilt+Landsize+Distance+Regionname)

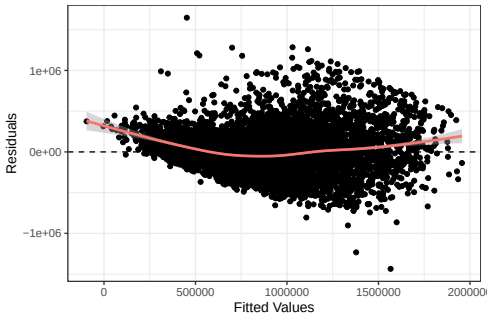


Figure 4: Residuals vs. Fitted Plot

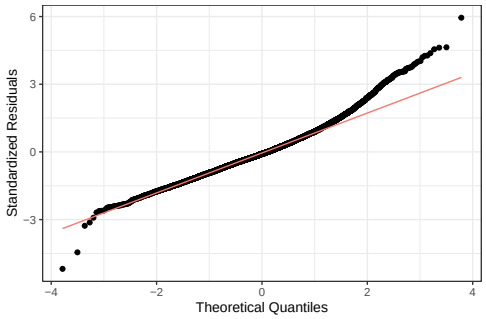


Figure 5: Normal Q−Q Plot of Residuals

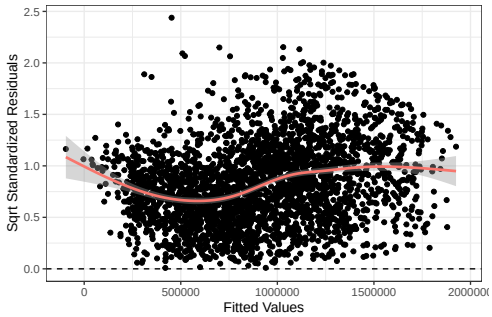


Figure 6: Square root of Standarized Residuals vs. Fitted Plo

Analysis

To Compare whether two models are significantly different or not, ANOVA test is performed and the results shows the model with "Type" outperforms the model without it and the model with the "Regionname" feature outperforms the model without it. The F-statistic and related p-value for each of our models provide sufficient evidence that our models outperform the simple mean. This effectively means that at the 5% significance level, I have enough evidence to conclude that the models created thus far outperform the simple mean. When the two models "with Type feature and without Region name" and "with Region name feature and without Type" are compared, the results show that the **model with Region name (Res Df = 6486) outperforms the model with Type feature (Res Df = 6483) with F = 407.3, p<0.01.**

Table 2: Coefficient of Model with Region name

Feature	2.5%	Mean	97.5%	t-value	p-value
(Intercept)	8039464.84	8495355.52	8951246.12	36.530	<2e-16
Rooms	243993.59	255231.04	266468.48	44.524	<2e-16
YearBuilt	-4419.81	-4187.48	-3955.15	-35.33	<2e-16
Bathroom	180897.89	198120.57	215343.25	22.551	<2e-16
landsize	116.47	138.86	161.24	12.16	<2e-16
Distance	-30856.73	-28973.54	-27090.35	-30.16	<2e-16
Region name_Northern Metropolitan	-212939.82	-186626.91	-160313.25	-13.904	<2e-16
Region name_Northern Victoria	-329854.80	-14851.92	300150.96	-0.092	0.926
Region name_South-Eastern Metropolitan	67489.14	125167.24	182845.33	4.254	2.13e-05
Region name_Southern Metropolitan	54635.45	80542.99	106450.52	6.094	1.16e-09
Region name_Western Metropolitan	-266519.06	-239912.97	-213306.88	-17.677	<2e-16

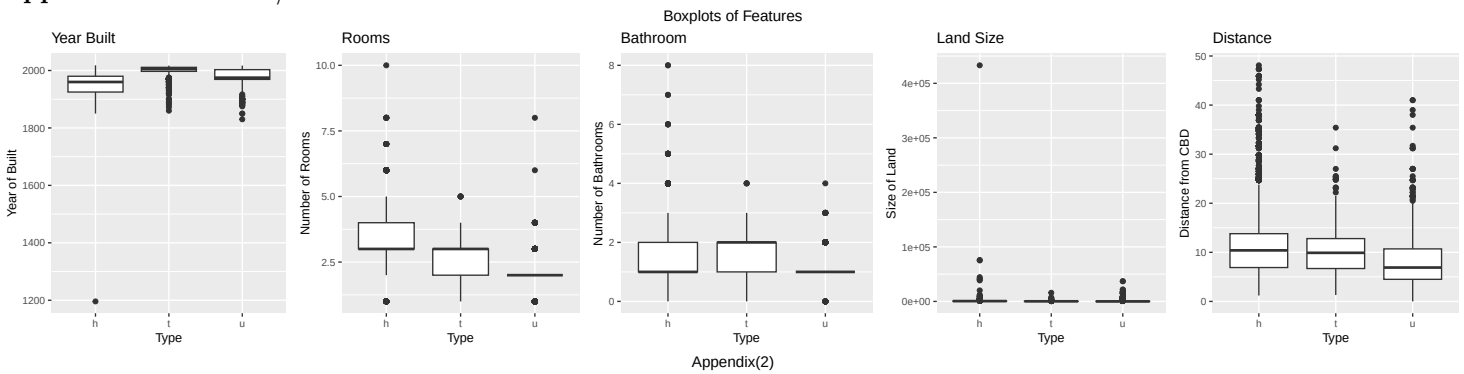
Conclusion

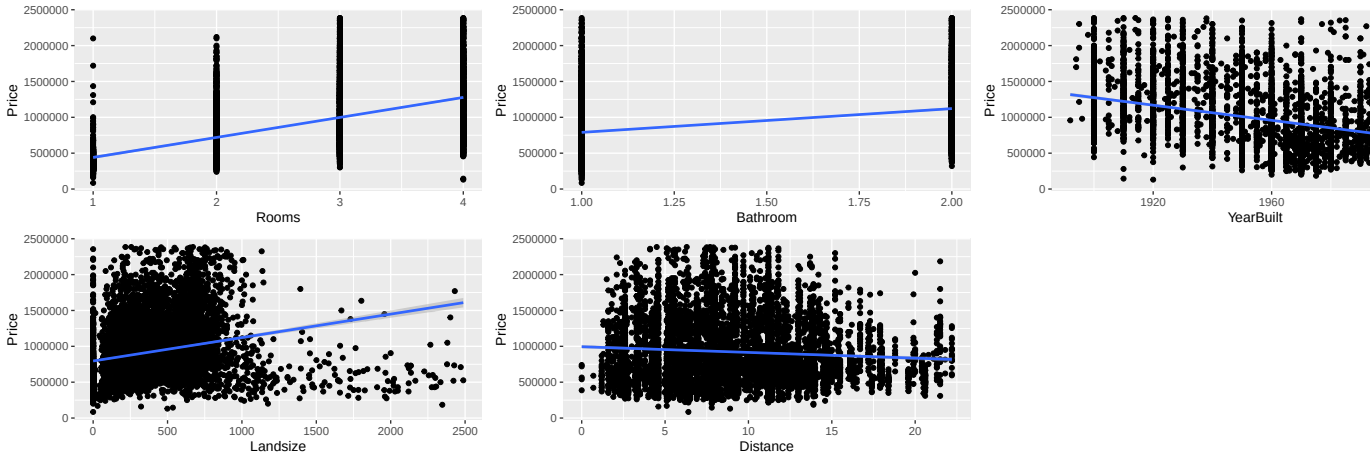
The outcome of this analysis's investigation is that **Region name is a better predictor variable than Type** as such; at a 5% level of significance, their impacts on price are marginally significant different. Additional research shows that including both features improves the model's performance. Please refer to the Appendix 6 for managerial thoughts regarding this model. We infer from this model that **59% of the variation in house prices can be accounted for by the number of rooms, year of built, the number of bathrooms, the land size, the distance from CBD, and the name of the region of the property**. Additionally, we can forecast how these factors will affect future house prices. For instance, **if nothing else changes, adding a bathroom to a property will raise the price by 180897.89 to 215343.25 dollars**.
The evaluated models do not generalize very well because some of their assumptions violated, despite the fact that improving their quality is not a primary objective of this analysis. As a result, they cannot be trusted to predict prices accurately, necessitating the use of potentially more complex methods.

Appendix 1 Summary statistic of Dataset

```
##      Price             Type               Rooms          Bathroom
##  Min.   :  85000   Length:13580       Min.   : 1.000   Min.   :0.000
##  1st Qu.: 650000   Class :character   1st Qu.: 2.000   1st Qu.:1.000
##  Median : 903000   Mode  :character   Median : 3.000   Median :1.000
##  Mean   :1075684                      Mean   : 2.938   Mean   :1.534
##  3rd Qu.:1330000                      3rd Qu.: 3.000   3rd Qu.:2.000
##  Max.   :9000000                      Max.   :10.000   Max.   :8.000
##
##   Regionname          YearBuilt       Landsize          Distance
##  Length:13580       Min.   :1196   Min.   :     0.0   Min.   : 0.00
##  Class :character   1st Qu.:1940   1st Qu.:   177.0   1st Qu.: 6.10
##  Mode  :character   Median :1970   Median :   440.0   Median : 9.20
##                     Mean   :1965   Mean   :   558.4   Mean   :10.14
##                     3rd Qu.:1999   3rd Qu.:   651.0   3rd Qu.:13.00
##                     Max.   :2018   Max.   :433014.0   Max.   :48.10
##                     NA's   :5375
```

Appendix 2 Outliers/Abrnormal values in Dataset





Appendix 3

Appendix 4 Anova results:

1. Anova on with and without Type

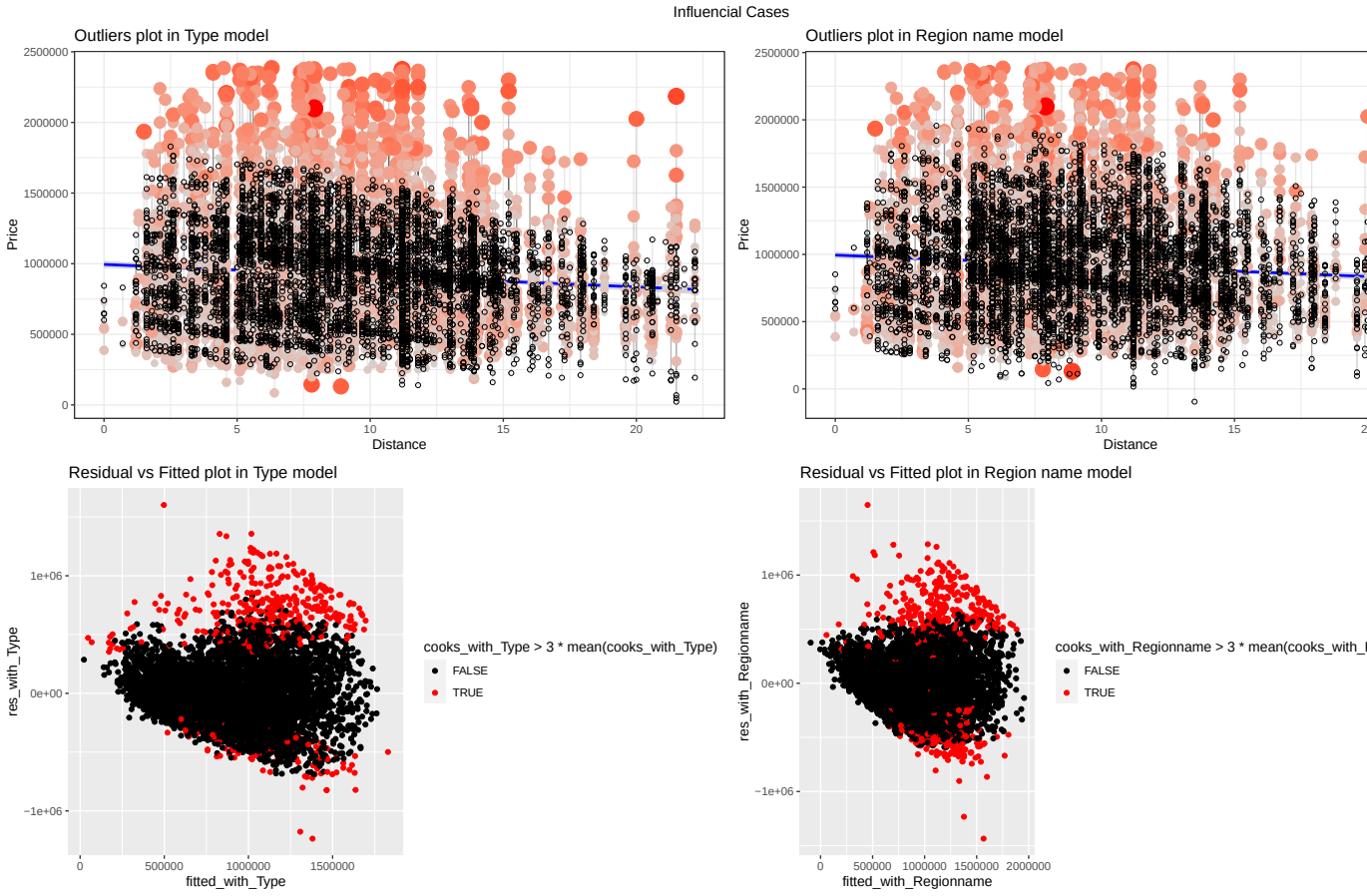
```
## Analysis of Variance Table
##
## Model 1: Price ~ Rooms + YearBuilt + Bathroom + Landsize + Distance
## Model 2: Price ~ Rooms + YearBuilt + Bathroom + Type + Landsize + Distance
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   6488 6.1892e+14
## 2   6486 5.9219e+14  2 2.6725e+13 146.35 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Anova on with and without Region name

```
## Analysis of Variance Table
##
## Model 1: Price ~ Rooms + YearBuilt + Bathroom + Landsize + Distance
## Model 2: Price ~ Rooms + YearBuilt + Bathroom + Regionname + Landsize +
##     Distance
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   6488 6.1892e+14
## 2   6483 4.9828e+14  5 1.2064e+14 313.92 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Anova on Type and Region name models

```
## Analysis of Variance Table
##
## Model 1: Price ~ Rooms + YearBuilt + Bathroom + Type + Landsize + Distance
## Model 2: Price ~ Rooms + YearBuilt + Bathroom + Regionname + Landsize +
##     Distance
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1   6486 5.9219e+14
## 2   6483 4.9828e+14  3 9.3914e+13 407.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



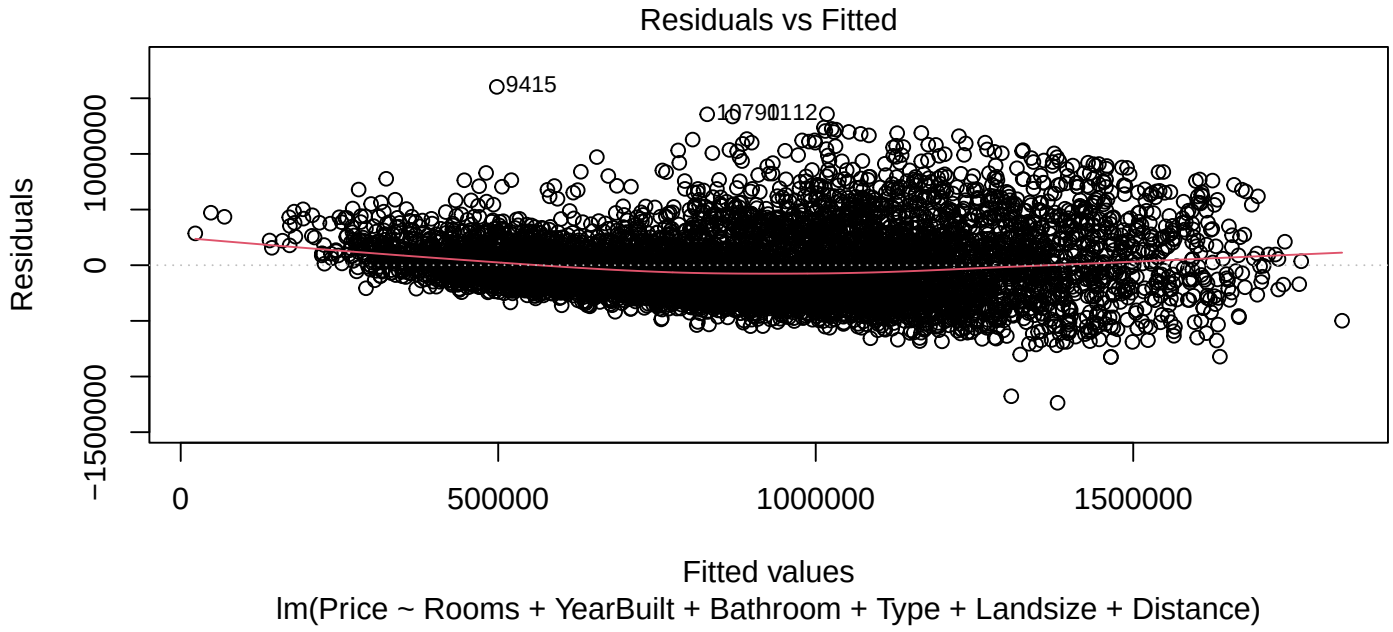
Appendix 5

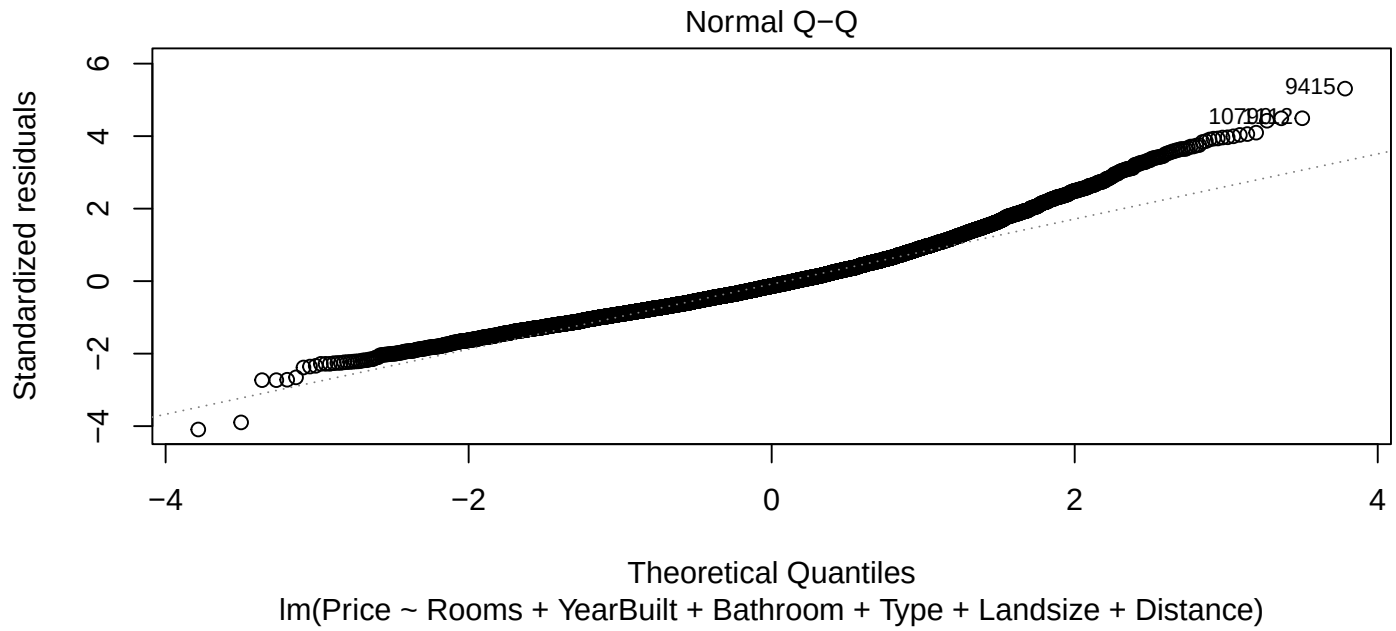
Appendix(5)

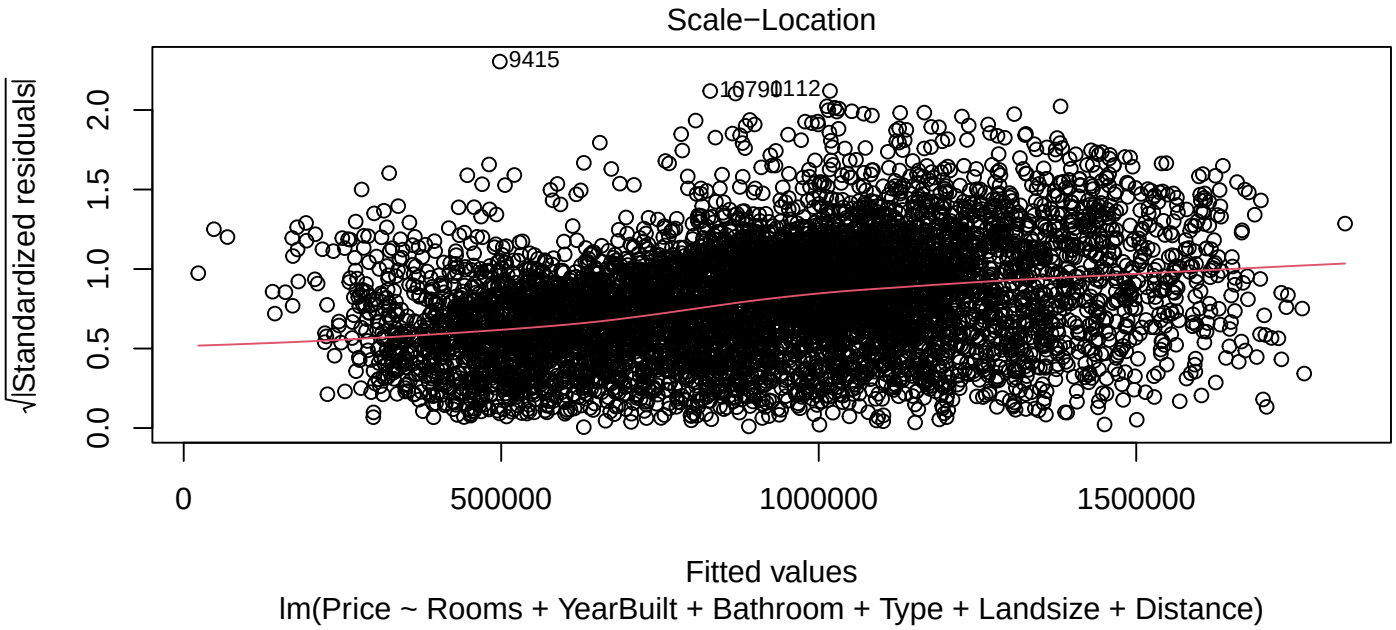
Appendix 6 Variance Explained by taking both Type an Region name categories is around 65% higher than taking into account individual categorical features, so its better to consider both Type and Region name for Price prediction. Although, the model still violates the assumption of Linearity, Homoscedasticity and Normality.

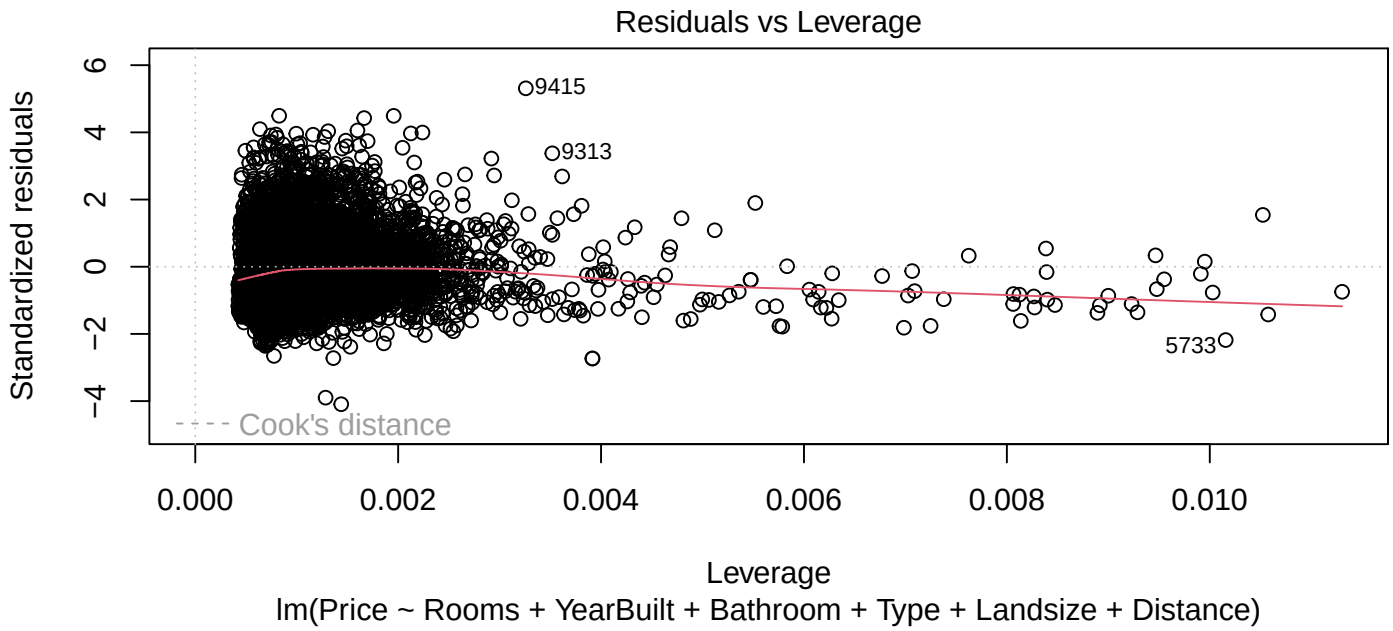
```
## Warning: contrasts dropped from factor Regionname due to missing levels
```

```
##
## Call:
## lm(formula = Price ~ Rooms + YearBuilt + Bathroom + Regionname +
##     Type + Landsize + Distance, data = selected_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1414657  -170481   -25581   141710  1292649
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        5636669.82  247329.00  22.790  < 2e-16 ***
## Rooms                               177346.93    5885.16  30.135  < 2e-16 ***
## YearBuilt                            -2686.76     126.66 -21.212  < 2e-16 ***
## Bathroom                            173491.70    8257.23  21.011  < 2e-16 ***
## RegionnameNorthern Metropolitan    -215110.56   12536.29 -17.159  < 2e-16 ***
## RegionnameNorthern Victoria         -30203.79  149663.69  -0.202     0.84
## RegionnameSouth-Eastern Metropolitan 145979.55   27411.42   5.326 1.04e-07 ***
## RegionnameSouthern Metropolitan     109729.93   12343.88   8.889  < 2e-16 ***
## RegionnameWestern Metropolitan     -270628.92   12684.31 -21.336  < 2e-16 ***
## Typet_H                             326605.30   10481.13  31.161  < 2e-16 ***
## Typet_T                             215157.54   12603.76  17.071  < 2e-16 ***
## Landsize                                99.49      10.86   9.163  < 2e-16 ***
## Distance                               -33854.16     910.07 -37.200  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 258200 on 6481 degrees of freedom
## Multiple R-squared:  0.6499, Adjusted R-squared:  0.6493
## F-statistic:  1003 on 12 and 6481 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ Rooms + YearBuilt + Bathroom + Type + Landsize + Distance)


Scale−Location
√|Standardized residuals|
Fitted values
lm(Price ~ Rooms + YearBuilt + Bathroom + Type + Landsize + Distance)


Residuals vs Leverage
Standardized residuals
6
4
2
0
−2
−4
9415
9313
5733
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
Leverage
lm(Price ~ Rooms + YearBuilt + Bathroom + Type + Landsize + Distance)