MSCI 718 – INDIVIDUAL ASSIGNMENT 3

Data Summary and Cleaning

The dataset used is Melbourne Housing dataset which has various information such as the Price in Australian Dollars, Suburb, Address, Number of rooms and bathrooms, size od the building, the regon distance from Melbourne CBD etc. This dataset contains 13580 objects of 21 variables and the variables are either continuous or categorical. In this report I will go through the regression analysis I did to find the **Price** of the house(in AUD) with respect to the following variables: **the type of the building**(Categorical; Type; h: house/villas/cottages, u: unit/duplex, t: townhouse)**, number of Rooms**(Categorical; Rooms; min: 1, max: 10)**, number of Bathrooms**(Categorical; Bathrooms; min:0, max:8)**, Area covered by the building**(BuildingArea; min:0, max: 44515) **, Distance from Melbourne CBD**(Distance; min:0, max:48.10)**, Year the building was built**(YearBuilt; min: 1196, max: 2018)**, the Region**(Eastern Metropolitan, Eastern Victoria, Northern Metropolitan, Northern Victoria, South-Eastern Metropolitan, Southern Metropolitan, Western Metropolitan, Western Victoria).

As part of data cleaning, missing data was removed and plotted boxplots to check outliers and removed them using IQR method.

Planning

The regression analysis is to **predict the prices of house**(dependent variable) by the **type**, the **number of bedrooms and bathrooms**, **area covered by the building, distance from Melbourne CBD**, the **region** and **built year**(the independent variables). In order to draw conclusions from the regression analysis the several assumptions were checked.

- All the chosen predictor variables are either quantitative or categorical and the output variable is quantitative, continuous and unbounded.
- Our data clearly has non-zero variance.
- The VIF and tolerance statistics were used to access collinearity. The largest VIF is 2.9 which is much lesser than 10; the average VIF is 1.72 which is close to 1. The lowest tolerance(1/VIF) is 0.346, greater than 0.1(which would indicate serious problem) and 0.2(which would indicate potential problem). Thus, concluded that there is no collinearity in my dataset.
- The predictors are uncorrelated with external variables.
- The Durbin-Watson test for independent errors was not significant at the 5% level of significance (d=1.94, p=0). As d is very close to 2 (which would indicate no autocorrelation detected), we do not reject the null hypothesis that the errors are independent, and continue with the assumption of independence met.
- Homoscedasticity and linearity were tested by plotting the model. From the graph, the model is almost homoscedastic and normal.

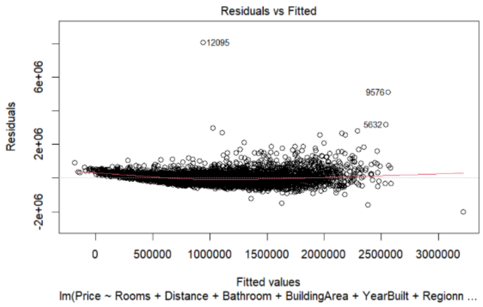


Fig 1: Plot of Residual vs Fitted

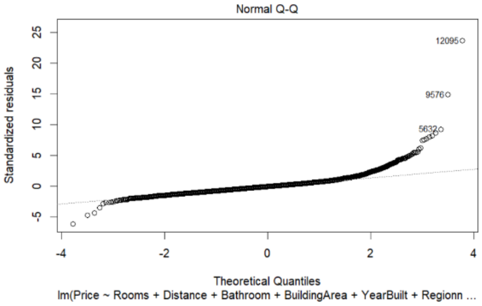


Fig 2: QQ plot of Standardized residuals

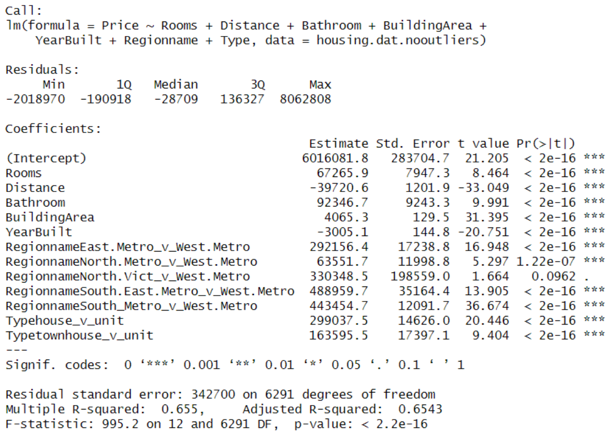
- After assessing the assumptions, I checked for outliers in the model and found 293 residuals are above or below 1.96 standard deviations. This represents 4.8% of the observations; less than 5%. Therefore, these observations are not considered as outliers and continued with all the observations included in the model.
- To investigate influential points, I calculated Cook's distance on the model. The maximum Cook's distance is 0.39, far below the cut-off 1. Thus, concluded there are no influential cases.

Before doing the regression analysis, I created dummy variables for the categorical variables Type and Regionname.

Analysis

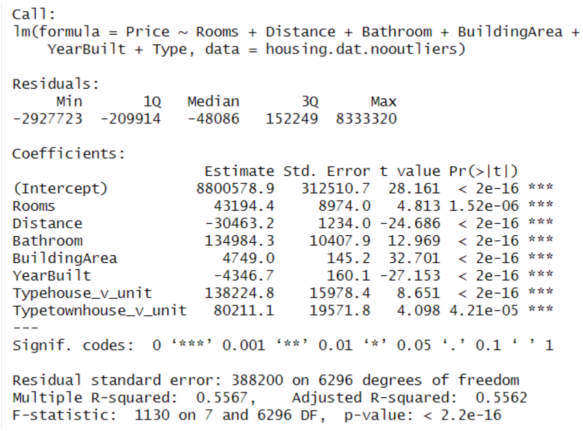
Model 1: Multiple linear regression is conducted on the dataset and the model is created. From the model, it can be concluded that all the seven predictor variables have an influence on the price of the house. **65.5% variance** in the outcome variable, Price of house is accounted by this model.

Model 2: Another model is created by removing the variable Regionname and multiple linear regression analysis is done. It accounts only **55.67% variance** in the outcome variable, Price of the house.



```
Call:
lm(formula = Price ~ Rooms + Distance + Bathroom + BuildingArea +
    YearBuilt + Regionname + Type, data = housing.dat.nooutliers)

Residuals:
     Min       1Q   Median       3Q      Max
-2018970  -190918   -28709   136327  8062808

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                           6016081.8   283704.7  21.205  < 2e-16 ***
Rooms                                   67265.9     7947.3   8.464  < 2e-16 ***
Distance                               -39720.6     1201.9 -33.049  < 2e-16 ***
Bathroom                                92346.7     9243.3   9.991  < 2e-16 ***
BuildingArea                             4065.3      129.5  31.395  < 2e-16 ***
YearBuilt                               -3005.1      144.8 -20.751  < 2e-16 ***
RegionnameEast.Metro_v_West.Metro      292156.4    17238.8  16.948  < 2e-16 ***
RegionnameNorth.Metro_v_West.Metro      63551.7    11998.8   5.297 1.22e-07 ***
RegionnameNorth.Vict_v_West.Metro      330348.5   198559.0   1.664   0.0962 .
RegionnameSouth.East.Metro_v_West.Metro 488959.7   35164.4  13.905  < 2e-16 ***
RegionnameSouth_Metro_v_West.Metro     443454.7    12091.7  36.674  < 2e-16 ***
Typehouse_v_unit                       299037.5    14626.0  20.446  < 2e-16 ***
Typetownhouse_v_unit                   163595.5    17397.1   9.404  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 342700 on 6291 degrees of freedom
Multiple R-squared:  0.655,	Adjusted R-squared:  0.6543
F-statistic: 995.2 on 12 and 6291 DF,  p-value: < 2.2e-16
```

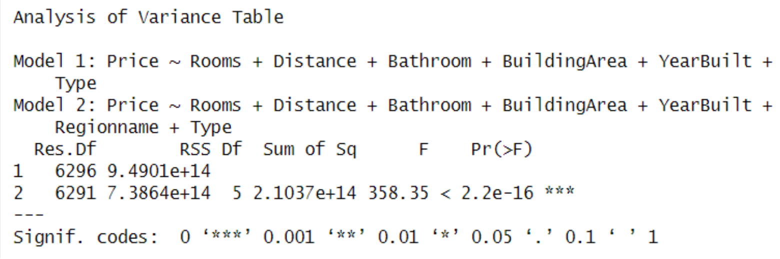
Fig 3: Summary of Model 1



```
Call:
lm(formula = Price ~ Rooms + Distance + Bathroom + BuildingArea +
    YearBuilt + Type, data = housing.dat.nooutliers)

Residuals:
     Min       1Q   Median       3Q      Max
-2927723  -209914   -48086   152249  8333320

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          8800578.9   312510.7  28.161  < 2e-16 ***
Rooms                  43194.4     8974.0   4.813 1.52e-06 ***
Distance              -30463.2     1234.0 -24.686  < 2e-16 ***
Bathroom              134984.3    10407.9  12.969  < 2e-16 ***
BuildingArea            4749.0      145.2  32.701  < 2e-16 ***
YearBuilt              -4346.7      160.1 -27.153  < 2e-16 ***
Typehouse_v_unit      138224.8    15978.4   8.651  < 2e-16 ***
Typetownhouse_v_unit   80211.1    19571.8   4.098 4.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 388200 on 6296 degrees of freedom
Multiple R-squared:  0.5567,	Adjusted R-squared:  0.5562
F-statistic:  1130 on 7 and 6296 DF,  p-value: < 2.2e-16
```

Fig 4: Summary of Model 2

The two models are compared using ANOVA and from the result we can say that **Model 1 significantly improved the fit of the model to the data** compared to Model 2, F(5, 6921) = 358.35, p<2.2e-16



```
Analysis of Variance Table

Model 1: Price ~ Rooms + Distance + Bathroom + BuildingArea + YearBuilt +
    Type
Model 2: Price ~ Rooms + Distance + Bathroom + BuildingArea + YearBuilt +
    Regionname + Type
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1   6296 9.4901e+14
2   6291 7.3864e+14  5 2.1037e+14 358.35 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 5: Analysis of variance for Model 1 and Model 2

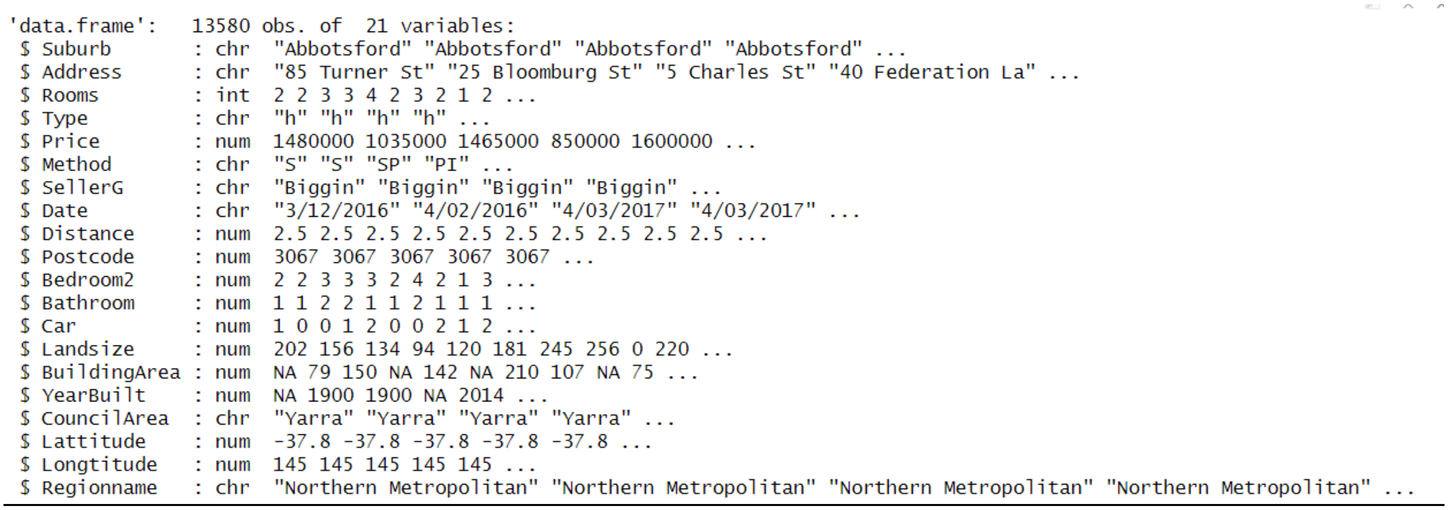
Conclusion

2 regression models were created to predict the price of houses in Melbourne city within 23.35km radius. Model 1 had 7 variables and Model 2 had 6 variables. To compare the models, anova() function was used and from the result it is clear that Model 1 is better.

From the model, it can be concluded that all the seven predictor variables have an influence on the price of the house at the 5%level of significance: number of rooms(t(6291) = 8.46, p<2e-16), number of bathrooms(t(6291) = 9.99, p<2e-16), distance from Melbourne CBD(t(6291) = -33.049, p<2e-16), area covered by building(t(6291) = 31.39, p<2e-16), built year(t(6291) = -20.75, p<2e-16), Eastern Metropolitan Region (t(6291) = 16.95, p<2e-16) , North Metropolitan Region(t(6291)=5.3,p=1.22e-07), South Eastern Metropolitan Region (t(6291)=13.905,p<2e-16), Southern metropolitan Region (t(6291) = 36.67, p<2e-16), Type “h” buildings (t(6291) = 20.446, p<2e-16), Type “t” buildings (t(6291) = 9.4, p<2e-16). The intercept is significantly different from 0(t(6291) = 21.2, p<2e-16). **65.5% variance** in the outcome variable, Price of house is accounted by this model. The adjusted R^2(0.6543) value is very similar to the observed R^2, indicating that the cross validity of this model is good.

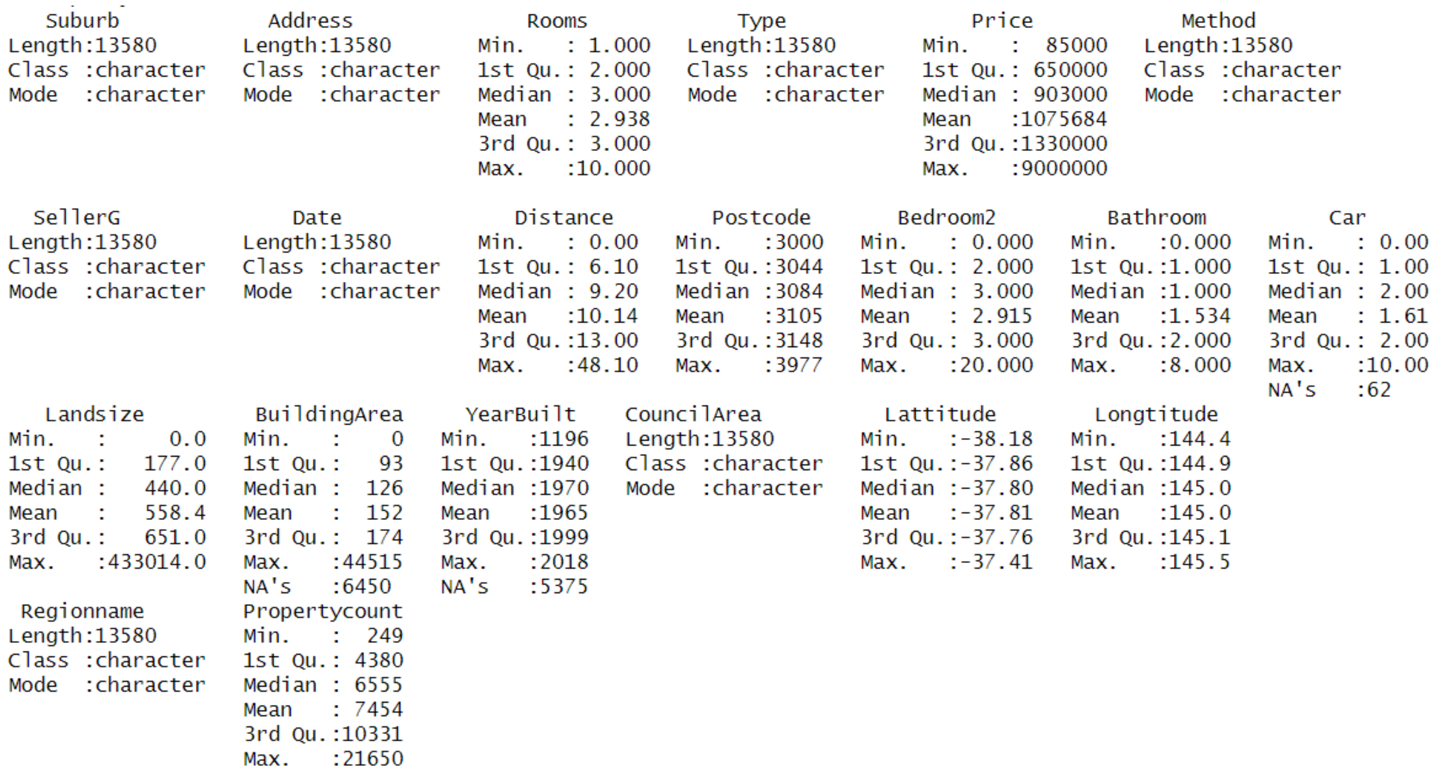
Appendix

Structure of Melbourne Housing dataset



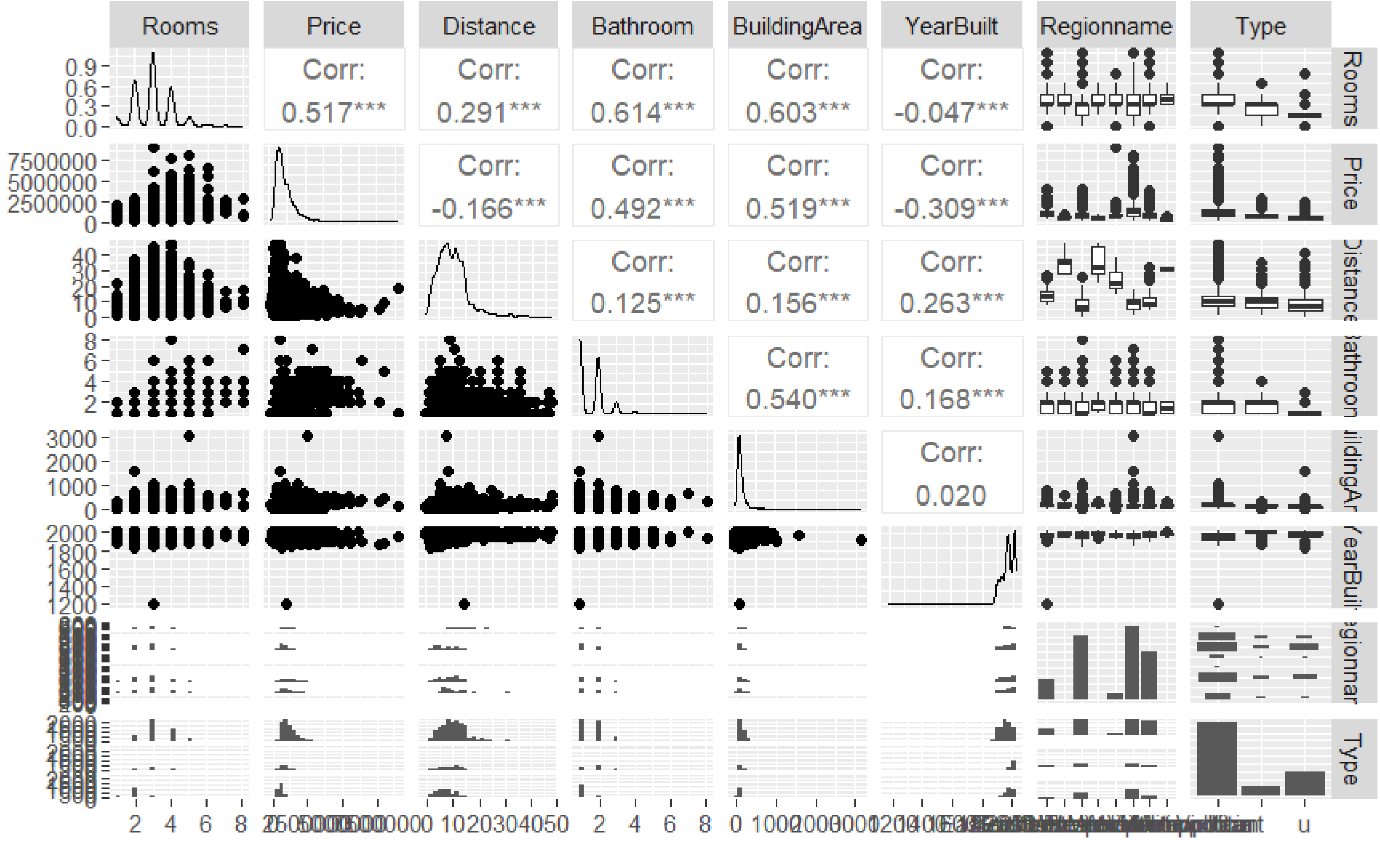
```
'data.frame':	13580 obs. of  21 variables:
 $ Suburb      : chr  "Abbotsford" "Abbotsford" "Abbotsford" "Abbotsford" ...
 $ Address     : chr  "85 Turner St" "25 Bloomburg St" "5 Charles St" "40 Federation La" ...
 $ Rooms       : int  2 2 3 3 4 2 3 2 1 2 ...
 $ Type        : chr  "h" "h" "h" "h" ...
 $ Price       : num  1480000 1035000 1465000 850000 1600000 ...
 $ Method      : chr  "S" "S" "SP" "PI" ...
 $ SellerG     : chr  "Biggin" "Biggin" "Biggin" "Biggin" ...
 $ Date        : chr  "3/12/2016" "4/02/2016" "4/03/2017" "4/03/2017" ...
 $ Distance    : num  2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 ...
 $ Postcode    : num  3067 3067 3067 3067 3067 ...
 $ Bedroom2    : num  2 2 3 3 3 2 4 2 1 3 ...
 $ Bathroom    : num  1 1 2 2 1 1 2 1 1 1 ...
 $ Car         : num  1 0 0 1 2 0 0 2 1 2 ...
 $ Landsize    : num  202 156 134 94 120 181 245 256 0 220 ...
 $ BuildingArea : num  NA 79 150 NA 142 NA 210 107 NA 75 ...
 $ YearBuilt   : num  NA 1900 1900 NA 2014 ...
 $ CouncilArea : chr  "Yarra" "Yarra" "Yarra" "Yarra" ...
 $ Lattitude   : num  -37.8 -37.8 -37.8 -37.8 -37.8 ...
 $ Longtitude  : num  145 145 145 145 145 ...
 $ Regionname  : chr  "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" ...
```

Summary of Melbourne Housing dataset

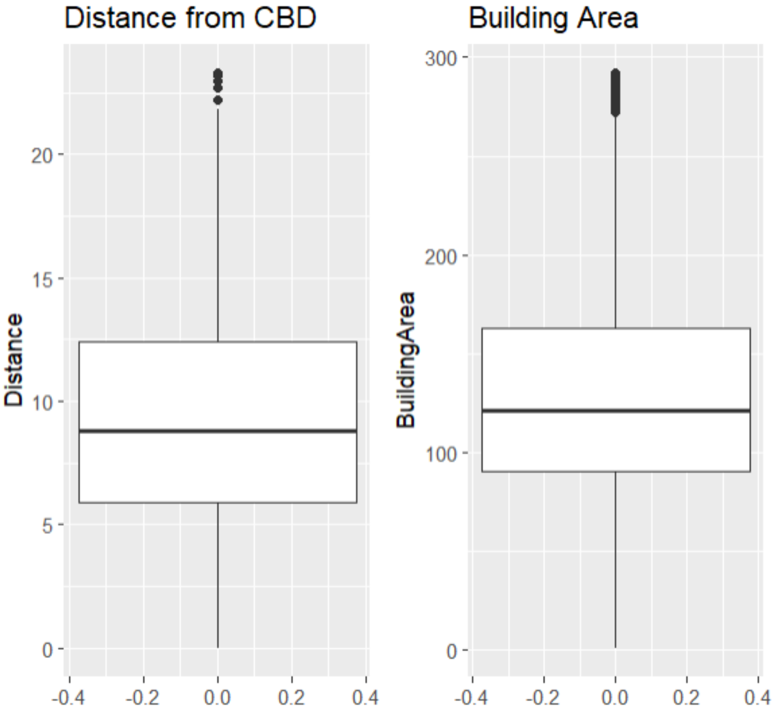


```
    Suburb             Address              Rooms             Type               Price             Method         
 Length:13580       Length:13580       Min.   : 1.000   Length:13580       Min.   :  85000   Length:13580      
 Class :character   Class :character   1st Qu.: 2.000   Class :character   1st Qu.: 650000   Class :character  
 Mode  :character   Mode  :character   Median : 3.000   Mode  :character   Median : 903000   Mode  :character  
                                       Mean   : 2.938                      Mean   :1075684                     
                                       3rd Qu.: 3.000                      3rd Qu.:1330000                     
                                       Max.   :10.000                      Max.   :9000000                     

   SellerG              Date              Distance        Postcode       Bedroom2         Bathroom          Car       
 Length:13580       Length:13580       Min.   : 0.00   Min.   :3000   Min.   : 0.000   Min.   :0.000   Min.   : 0.00  
 Class :character   Class :character   1st Qu.: 6.10   1st Qu.:3044   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 1.00  
 Mode  :character   Mode  :character   Median : 9.20   Median :3084   Median : 3.000   Median :1.000   Median : 2.00  
                                       Mean   :10.14   Mean   :3105   Mean   : 2.915   Mean   :1.534   Mean   : 1.61  
                                       3rd Qu.:13.00   3rd Qu.:3148   3rd Qu.: 3.000   3rd Qu.:2.000   3rd Qu.: 2.00  
                                       Max.   :48.10   Max.   :3977   Max.   :20.000   Max.   :8.000   Max.   :10.00  
                                                                                                       NA's   :62     
    Landsize         BuildingArea     YearBuilt    CouncilArea          Lattitude        Longtitude   
 Min.   :     0.0   Min.   :    0   Min.   :1196   Length:13580       Min.   :-38.18   Min.   :144.4  
 1st Qu.:   177.0   1st Qu.:   93   1st Qu.:1940   Class :character   1st Qu.:-37.86   1st Qu.:144.9  
 Median :   440.0   Median :  126   Median :1970   Mode  :character   Median :-37.80   Median :145.0  
 Mean   :   558.4   Mean   :  152   Mean   :1965                      Mean   :-37.81   Mean   :145.0  
 3rd Qu.:   651.0   3rd Qu.:  174   3rd Qu.:1999                      3rd Qu.:-37.76   3rd Qu.:145.1  
 Max.   :433014.0   Max.   :44515   Max.   :2018                      Max.   :-37.41   Max.   :145.5  
                    NA's   :6450    NA's   :5375                                                      
  Regionname        Propertycount  
 Length:13580       Min.   :  249  
 Class :character   1st Qu.: 4380  
 Mode  :character   Median : 6555  
                    Mean   : 7454  
                    3rd Qu.:10331  
                    Max.   :21650  
```

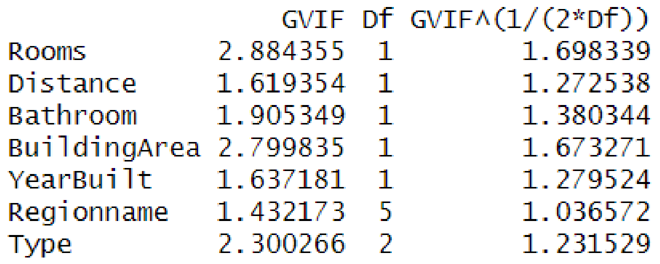
Pairplot of variables under analysis



Boxplots after removing outliers

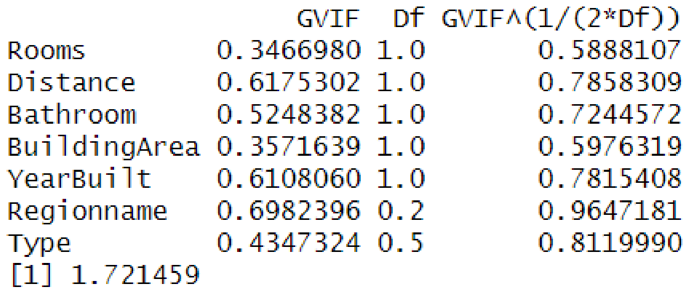


Result of VIF



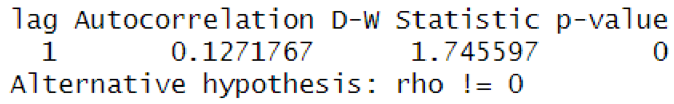
```
                GVIF Df GVIF^(1/(2*Df))
Rooms       2.884355  1        1.698339
Distance    1.619354  1        1.272538
Bathroom    1.905349  1        1.380344
BuildingArea 2.799835  1        1.673271
YearBuilt   1.637181  1        1.279524
Regionname  1.432173  5        1.036572
Type        2.300266  2        1.231529
```

Result of Tolerance and mean(VIF)



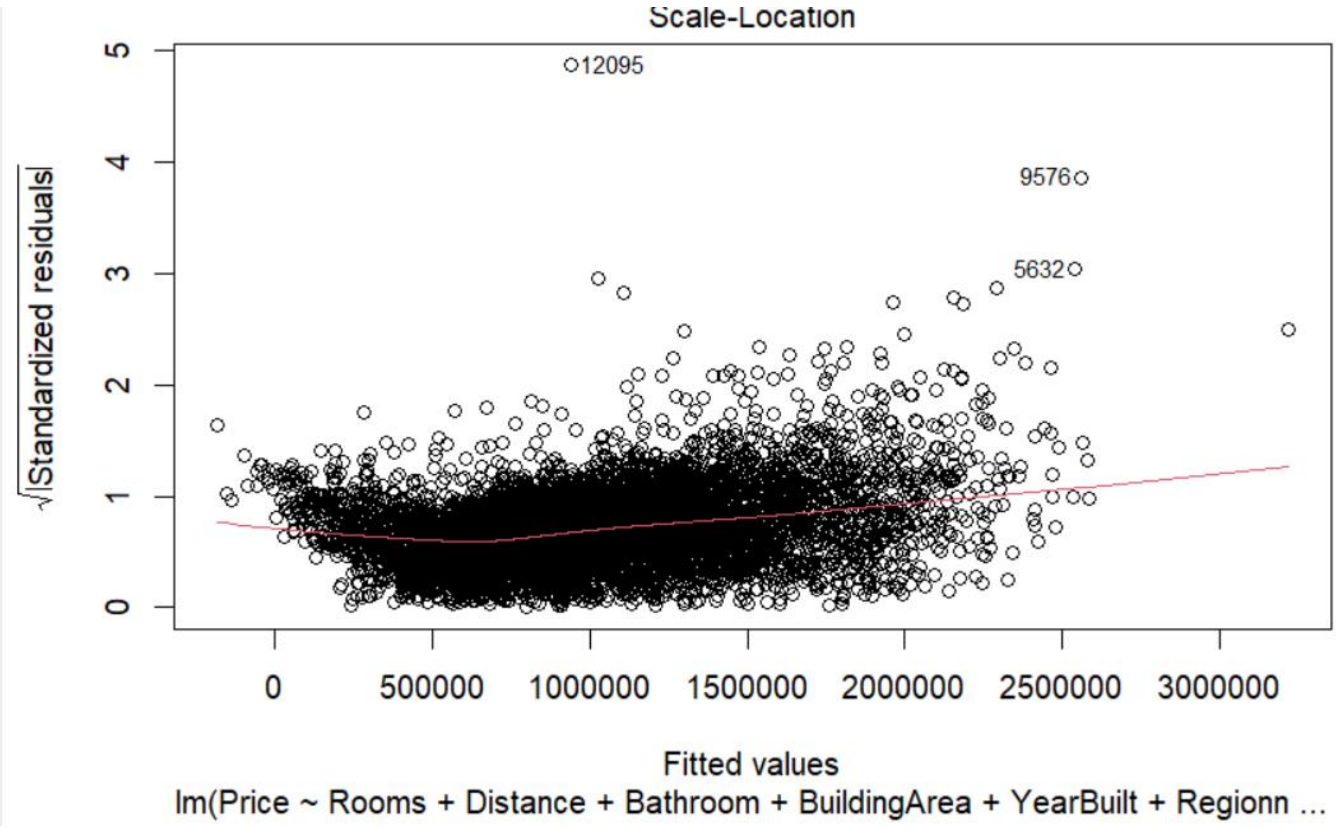
```
                 GVIF  Df GVIF^(1/(2*Df))
Rooms       0.3466980 1.0       0.5888107
Distance    0.6175302 1.0       0.7858309
Bathroom    0.5248382 1.0       0.7244572
BuildingArea 0.3571639 1.0       0.5976319
YearBuilt   0.6108060 1.0       0.7815408
Regionname  0.6982396 0.2       0.9647181
Type        0.4347324 0.5       0.8119990
[1] 1.721459
```

Result of DW Test

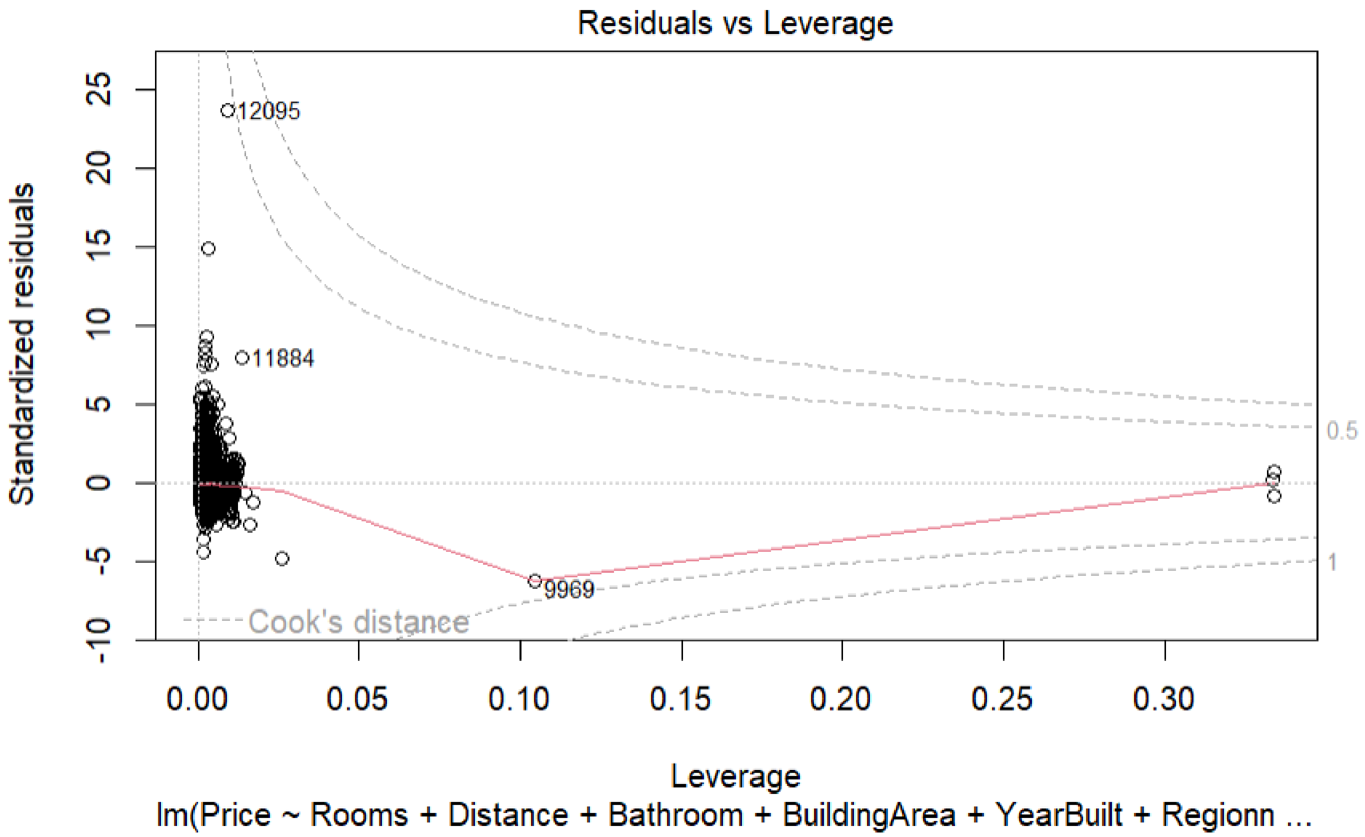


```
 lag Autocorrelation D-W Statistic p-value
   1       0.1271767      1.745597       0
 Alternative hypothesis: rho != 0
```

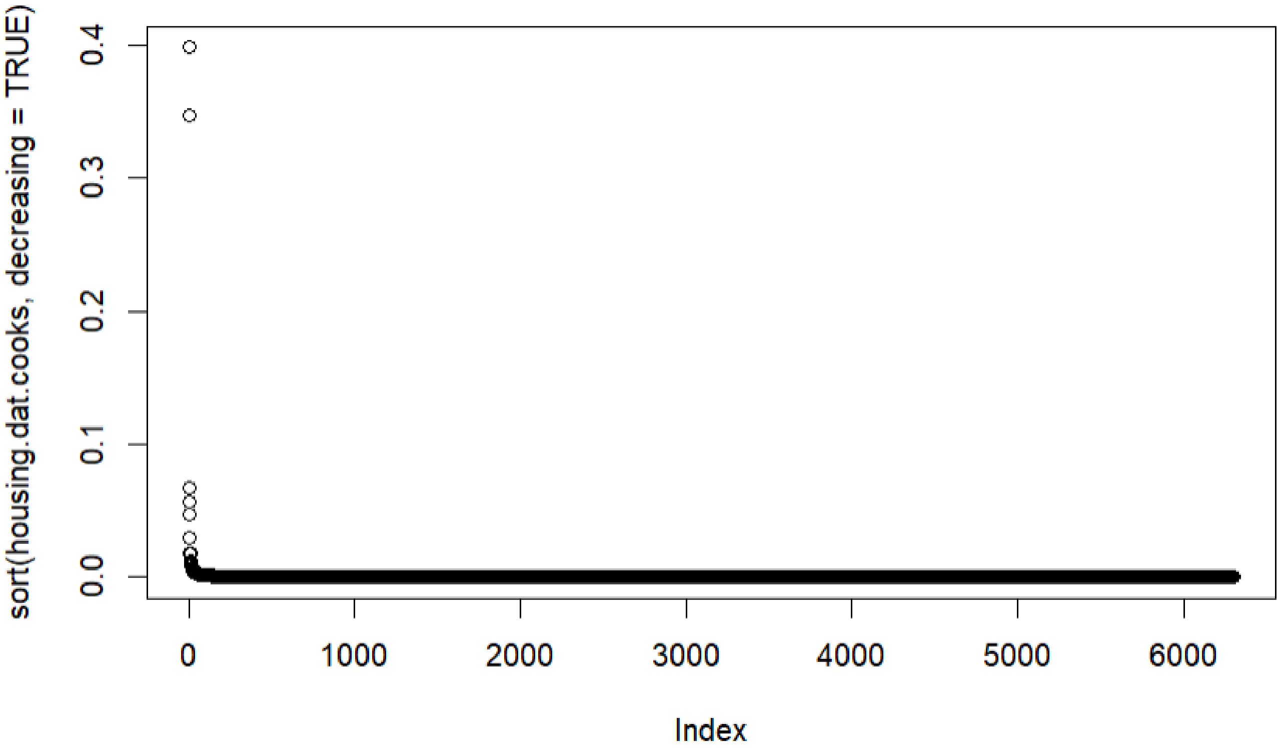
Sqrt(Standardized residuals) vs Fitted values plot



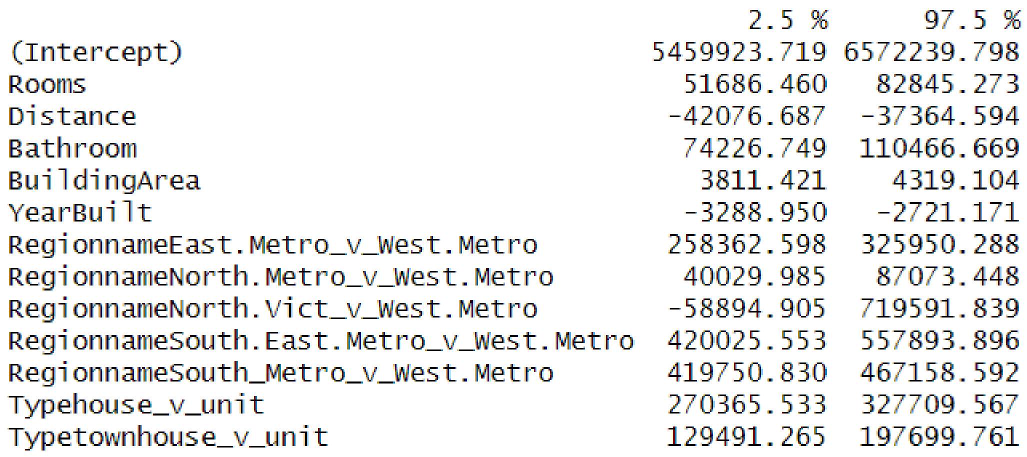
Residuals vs Leverage plot



Cook’s distance plot



Confidence Intervels for the model 1



```
                                           2.5 %      97.5 %
(Intercept)                          5459923.719 6572239.798
Rooms                                  51686.460   82845.273
Distance                              -42076.687  -37364.594
Bathroom                               74226.749  110466.669
BuildingArea                            3811.421    4319.104
YearBuilt                              -3288.950   -2721.171
RegionnameEast.Metro_v_West.Metro     258362.598  325950.288
RegionnameNorth.Metro_v_West.Metro     40029.985   87073.448
RegionnameNorth.Vict_v_West.Metro     -58894.905  719591.839
RegionnameSouth.East.Metro_v_West.Metro 420025.553  557893.896
RegionnameSouth_Metro_v_West.Metro    419750.830  467158.592
Typehouse_v_unit                      270365.533  327709.567
Typetownhouse_v_unit                  129491.265  197699.761
```