House Prices Prediction

Introduction

Whether or not to buy a house can be one of the most important decisions for people. Predicting house value can guide people when they are making such an important decision so that they can plan their finances more efficiently and appropriately. Meanwhile, people may be indecisive among several houses due to their budgets. Obtaining prices of house features through regression coefficients can assist people in quantitatively weighing various house features. Our prediction and analysis would assist people in rationally choosing a desirable and valuable house.

Problem Statement

This project plans to solve the problem of predicting housing prices in Melbourne and identifying the important variables for predicting prices. The only dependent variable is house prices. Eventually, the predicted results are compared with the actual prices. The project will use 3 performance metrics, i.e., adjusted R square, mean absolute percentage error, and AIC, to measure the models' performance.

Data Description and Preliminary Analysis

The data is collected from the Melbourne Housing Market dataset on Kaggle[1]. The dataset contains 21 columns and 34858 rows. Each row represents a transaction record in the housing market, and out of 21 columns, one column stands for the transaction price and the other 20 are all characteristics of the transaction or the underlying property. A detailed description of all columns can be found in the appendix. Of these feature columns, 8 columns are qualitative, including Method, Type, etc. All categories for the qualitative variables are also included in the appendix. 12 columns are quantitative, including Distance, Bathroom, etc. The dataset contains many missing values and requires cleaning. After removing all the missing values, 8895 rows are left in the dataset.

To gain some basic understanding of the dataset, the distribution of price is generated as shown in the bottom left:

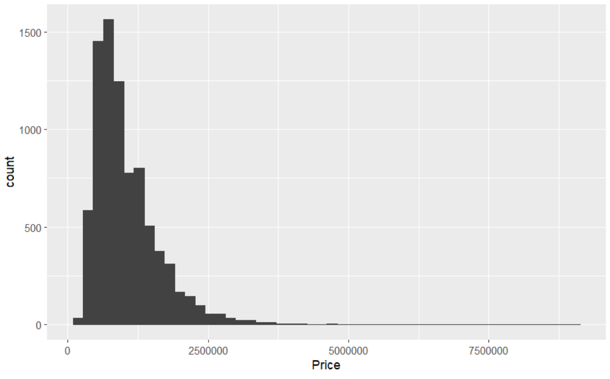


Figure 1: Distribution of Price before removing outliers

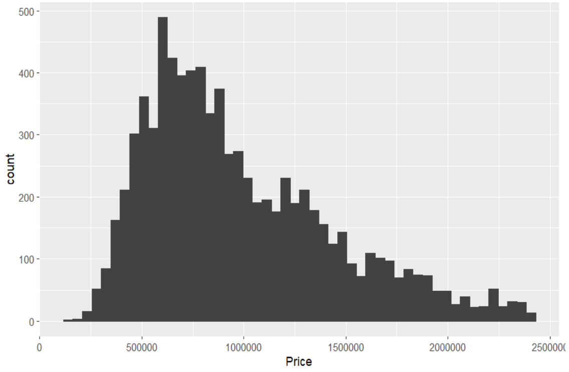
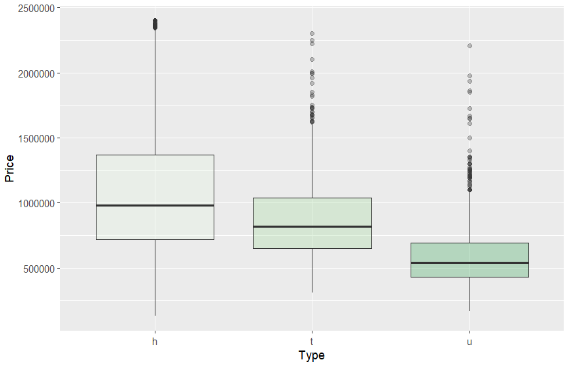


Figure 2: Distribution of Price after removing outliers

The distribution of price is highly right-skewed, while most data points are concentrated between 100000 and 1500000. It indicates that there are many outliers in Price, which is also the case for many quantitative variables in the dataset, such as BuildingArea, Landsize, etc. Therefore, outliers are removed in Price, BuildingArea, Landsize, and YearBuilt. After the cleaning process, the dataset

[1] https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market

contains 8043 records. The distribution of price after cleaning is shown in the top right, which is roughly normal and can be used in further analysis.



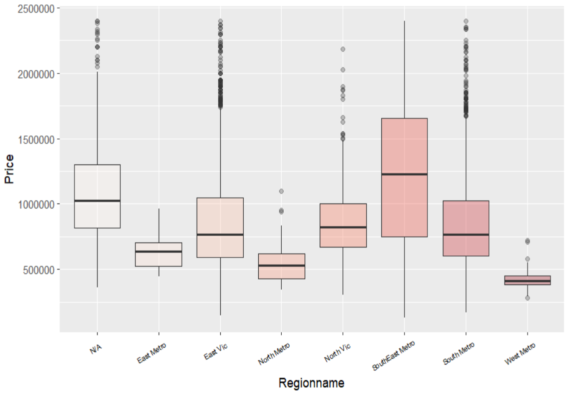


Figure 3: Boxplot for Price by Type

Figure 4: Boxplot for Price by Regionname

Two box plots are generated to investigate the relationship between price and qualitative variables, Type and Regionname correspondingly. The box plot in the top left shows that the property with type u has a notably lower median than property h, which is intuitive because property with type house is usually more expensive. The box plot between price and Regionname in the top right indicates that the housing price has different distribution across different regions, especially in Southern Metropolitan, which is significantly higher than in other regions. The above analysis concludes that some of the qualitative variables are correlated with the price, therefore, should be included in the regression model.

Next, scatter plots are generated between some quantitative variables and price:

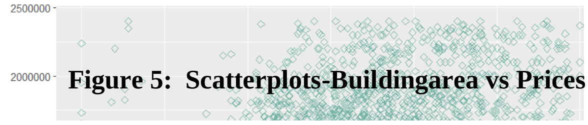


Figure 5: Scatterplots-Buildingarea vs Prices

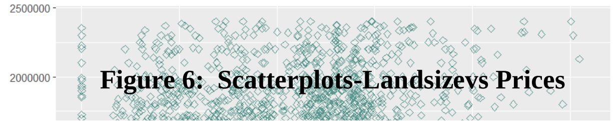


Figure 6: Scatterplots-Landsizevs Prices

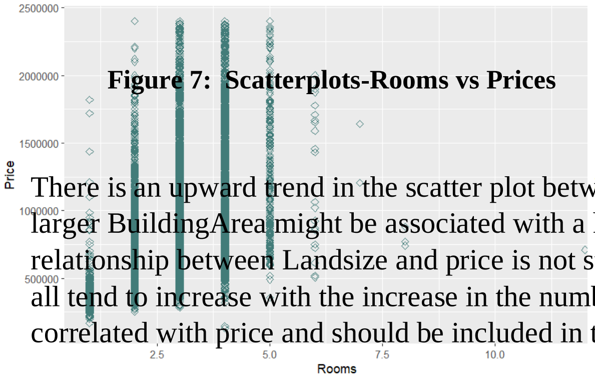


Figure 7: Scatterplots-Rooms vs Prices

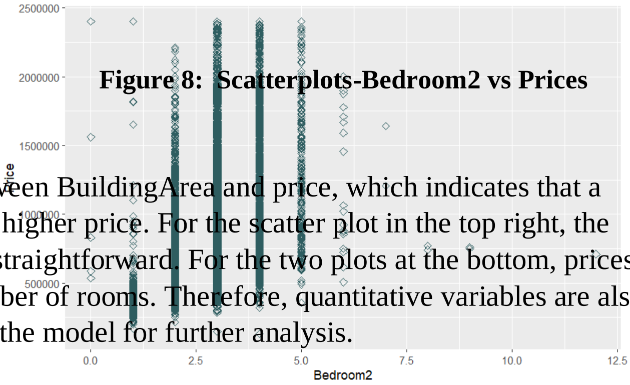


Figure 8: Scatterplots-Bedroom2 vs Prices

There is an upward trend in the scatter plot between BuildingArea and price, which indicates that a larger BuildingArea might be associated with a higher price. For the scatter plot in the top right, the relationship between Landsize and price is not straightforward. For the two plots at the bottom, prices all tend to increase with the increase in the number of rooms. Therefore, quantitative variables are also correlated with price and should be included in the model for further analysis.

Finally, the correlation plot is generated as follows:

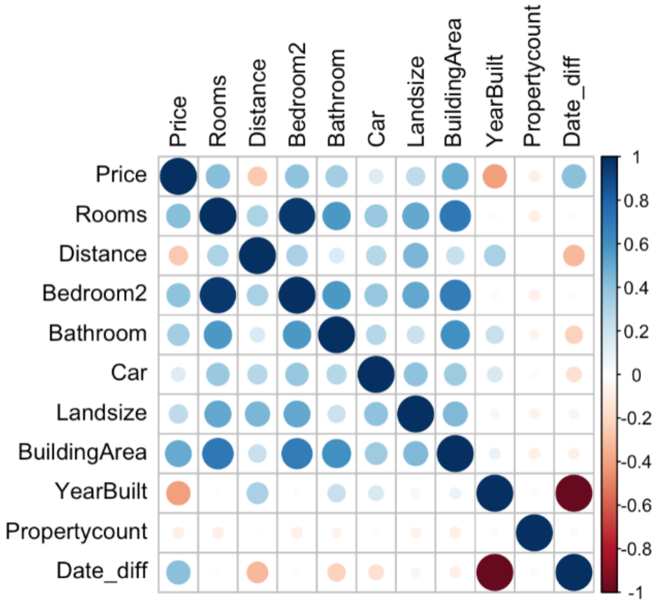


Figure 9: Correlation Matrix

The correlation between Bedroom2 and Rooms is high, and 96% of these two values are the same. Therefore we decide to keep only the variable Rooms. The correlation between YearBuilt and Date_diff is also highly negative because Date_diff is generated by using the transaction date minus the YearBuilt. We only keep one of these two variables in any of our models.

Analysis

In exploratory data analysis, we found the missing data problem, outlier problem, and multicollinearity problem between variable Room and variable Bedroom2 (correlation is 0.96). These three problems will significantly impact multiple linear regression fit if we do nothing. Considering we have lots of data, we decided to drop rows that contain the missing value. Because we are interested in predicting the housing price, not the outliers, we decided to drop the outliers. In definition, variable Rooms is the number of rooms. Moreover, variable Bedroom2 is the Scraped number of Bedrooms from a different source. We found that the values of these two variables are highly overlapping. As a result, we decided to drop the Bedroom2 variable. (We ran the models without doing these data preprocess steps. We can see the model performance improved after doing that.)

After the data preprocessing, we started to build our first model. To keep it simple, we first tried the model 1 of removing categorical variables with many unique values. This model ended up having 8 numeric variables, 3 categorical variables, and 22 coefficients. The adjusted R square is 0.7063 in the training dataset. The F statistic is 538.2 on 21 and 4670 degrees of freedom with a p-value less than 2.2e-16. This means that at least one variable in our model is statistically significant. In residual analysis, we found that the residuals show a larger variance as the fitted values increase (We fixed this issue in the later model). The points at the high end deviate from the reference line.

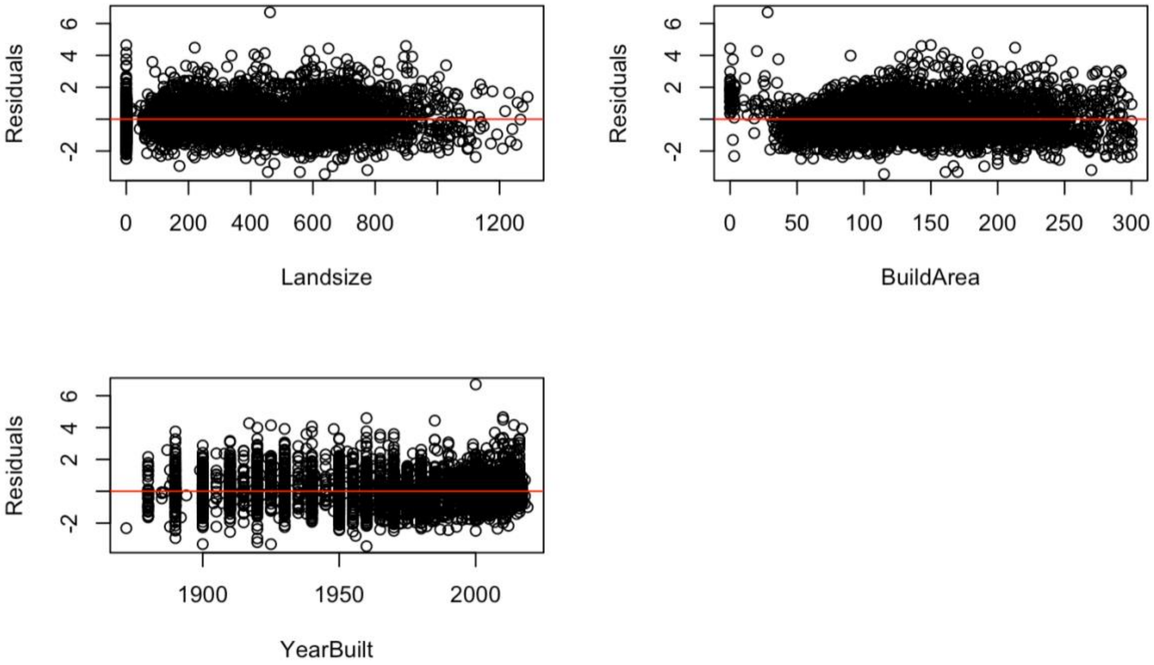


Figure 10: Model 1 Residual Analysis

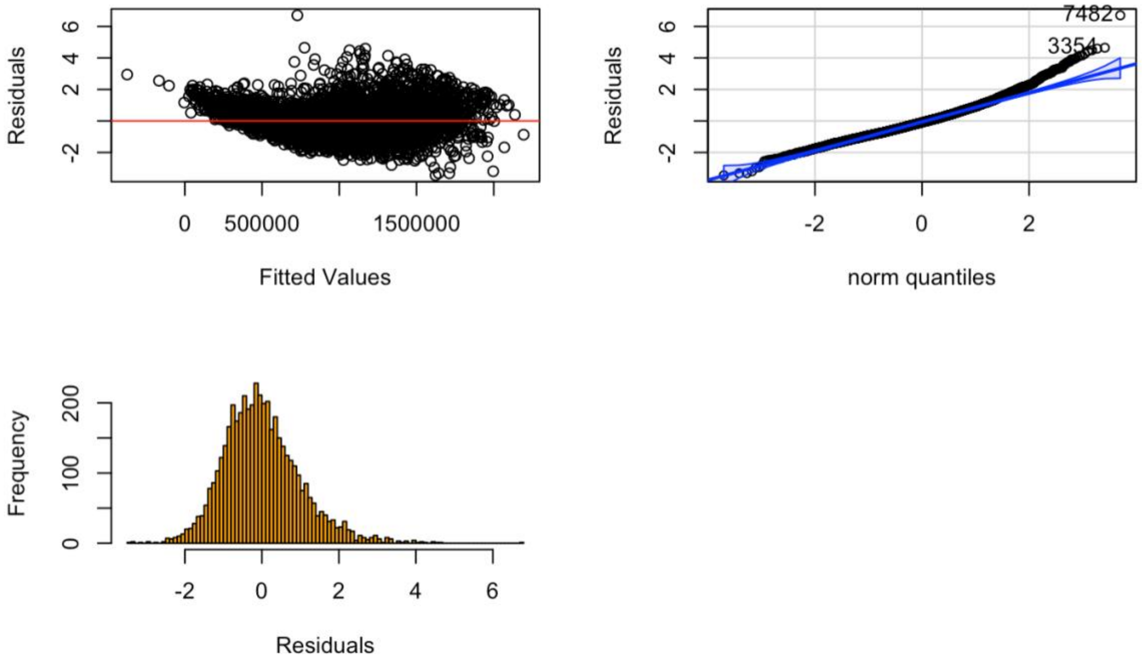


Figure 11: Model 1 Assumptions Checking

In model 2, we used a mean encoding technique to deal with the categorical variables with many unique values and added a new variable, "Date_diff". In mean encoding, each distinct value of categorical value is replaced with the average value of the target variable we are trying to predict. In our case, the target variable is housing price. Compared with the dummy variable encoding, mean encoding does not add too much sparseness to the data. In model 1, we excluded the variable "Date" which is the date the property is sold. However, we thought the newness of the house would affect the price of the house at the moment of selling it. We can get this information by subtracting the year the house was built from the date the house was sold. We named this information the new variable "Date_diff". To avoid the multicollinearity problem, we dropped the variable "YearBuilt" and included the variable "Date_diff" in model 2. After feature transformation and adding the new feature, the adjusted R square of model 2 jumped to 0.7747 in the training dataset. The F statistic is 649 on 24 and 4499 degrees of freedom with a p-value less than 2.2e-16. In residual analysis, the fan shape in the residuals vs. fitted value plot does not improve, and as shown in the Q-Q plot, the target variable is skewed to the right. Therefore, we need to transform it to become normally distributed.

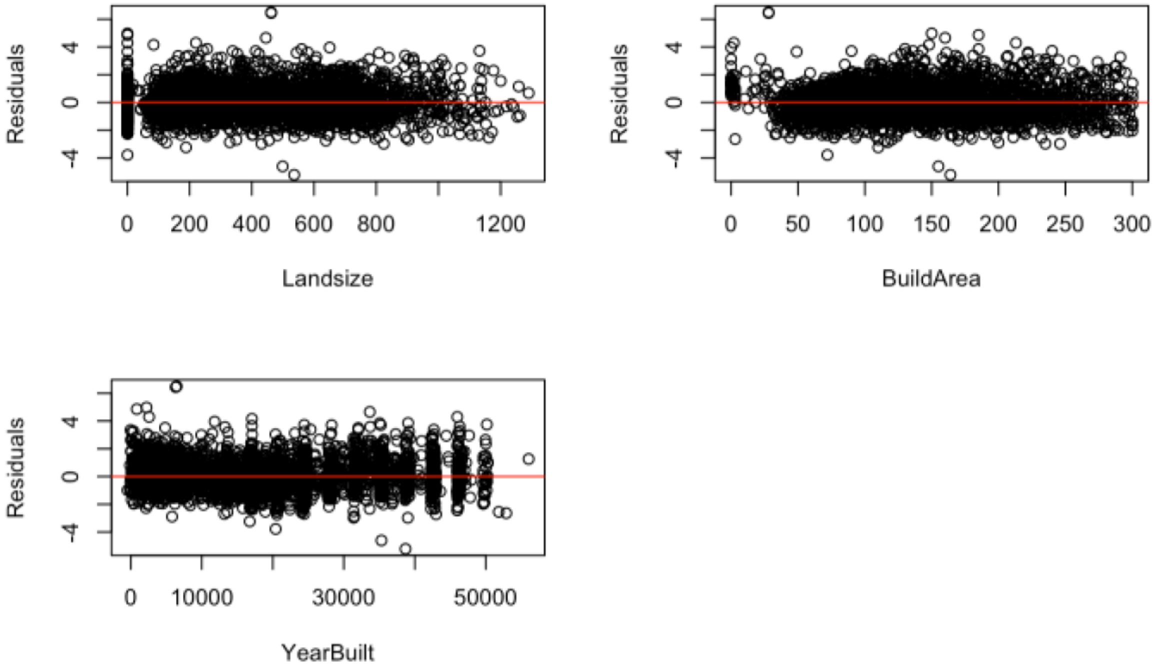


Figure 12: Model 2 Residual Analysis

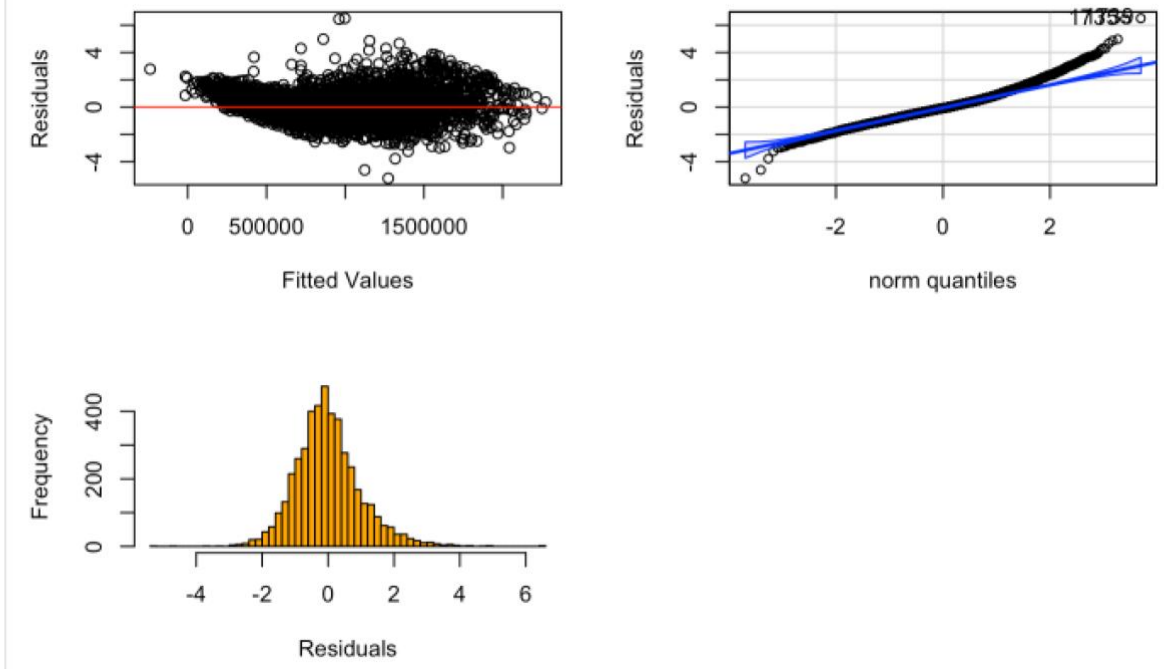


Figure 13: Model 2 Assumptions Checking

In the residual analysis of model 2, we suspected the constant-variance assumption had been violated. To solve this problem, we did a log transformation of the targeted variable y in model 3. The adjusted

R squared increases to 0.7942. In addition, we can see that the residuals are uniformly scattered after log transformation, and the Q-Q plot is closer to the reference line.

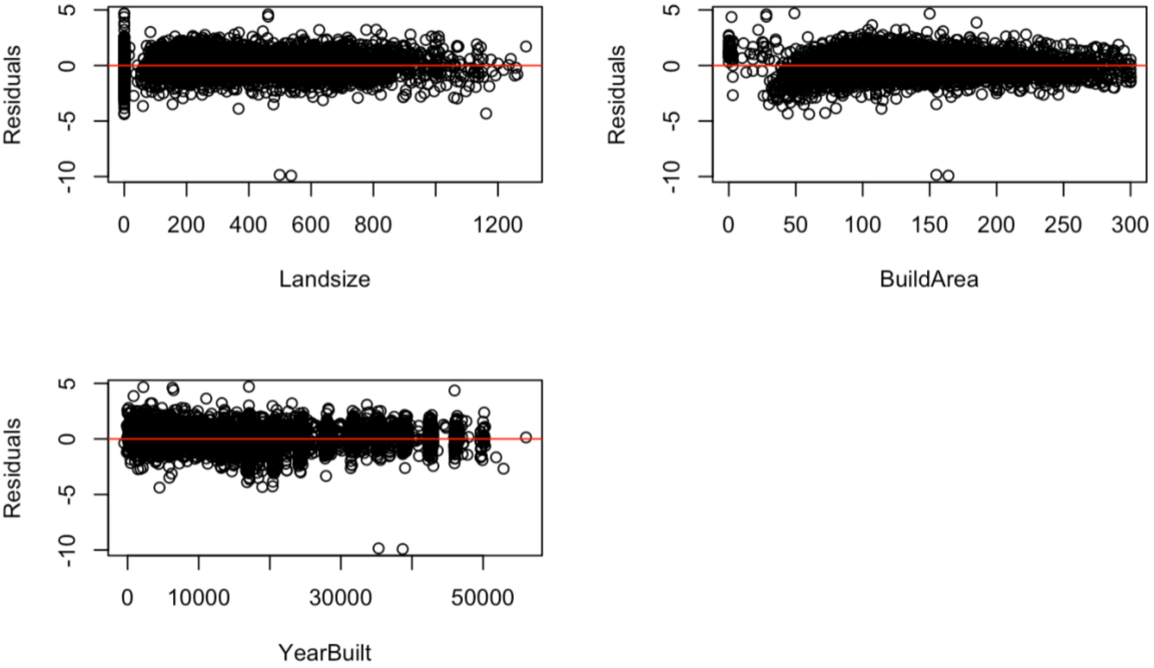


Figure 14: Model 3 Residual Analysis

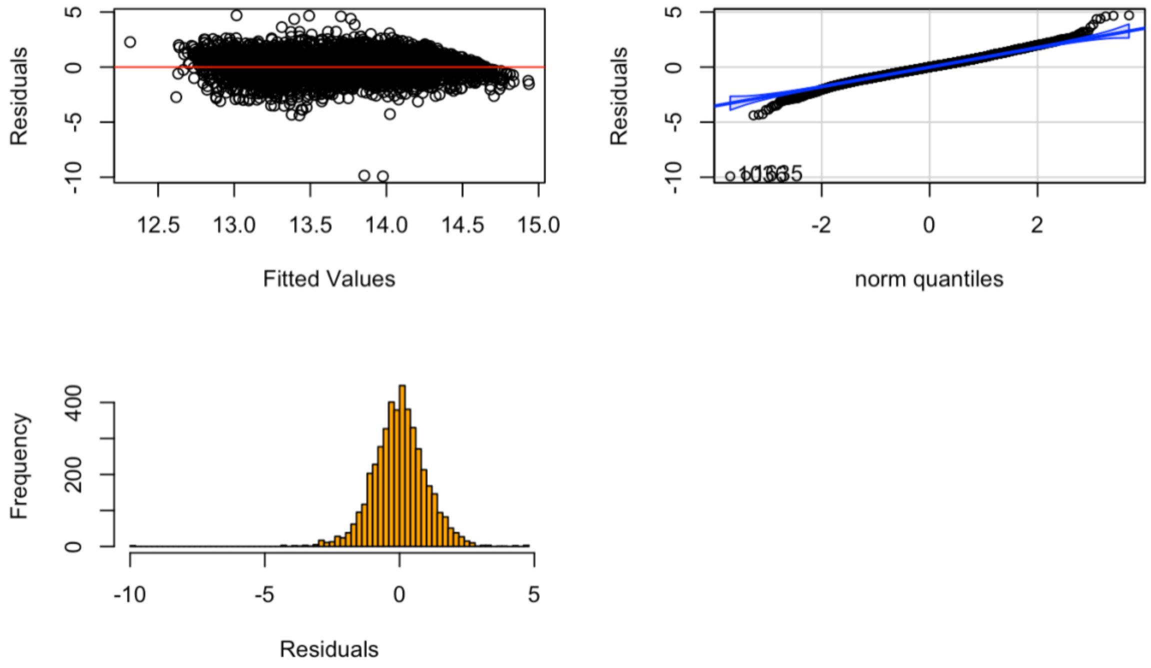


Figure 15: Model 3 Assumptions Checking

To get a better interpretation of our model and further improve our model prediction accuracy, we applied LASSO regression to help us make the variable selection. We used 10-fold cross-validation to help us choose the best penalty parameter **λ**, which is 0.0009319311. From the LASSO estimation of regression coefficients, we can see that none of the coefficients is 0, so it means LASSO doesn’t remove any variables.

As professor Goldsman suggested during our presentation, we included time and inflation indicators such as Year_of_Sale, Month_of_Sale, Season_of_Sale, and CPI. To remain our model at effective rank, we kept Season_of_Sale and CPI in Model 4. Compared with model 3, model 4 has a slightly lower adjusted R square and MAPE while much higher AIC. In addition, we also applied LASSO regression to make a variable selection, and it forced the coefficient of variable "CPI" to equal 0. The regression coefficients obtained from LASSO are less efficient than those abstained from Ordinary Least Squares. Therefore, we rerun model 4 after removing the variable "CPI". Unfortunately, its performance was still not as good as model 3 on the validation data set. What's more, as shown in the Appendix - Model 4 regression summary, none of Season_of_Sale and CPI is significant at a 5% significance level. Thus, our group ended up using model 3.

Conclusions and Recommendations

The following table provides each model's performance on the validated set. R-square indicates how much the independent variables explain the variation of a dependent variable. Mean Absolute Percentage Error (MAPE) is the mean or average of the absolute percentage errors of the predictions[2]. We use MAPE here because MAPE is easy to read, and values in ys and y hats are large. MSE is scale-dependent, whereas MAPE is not. When comparing accuracy with different scales due to the model with log y transformation, it is better to use MAPE. AIC is an estimator of prediction error and, thereby, the relative quality of statistical models.

By comparing all 4 models, it is clear that Model 3 gives the best prediction on house price because it has relatively high variability within the target variables, low mean absolute percentage error, and low AIC. We also applied a time series regression model and added it onto the model 3. Our finding is that adding seasonality and CPI (Consumer Price Index) doesn’t help significantly improve the actual model performance. Therefore, we decided to still keep model 3 as our final model. To test lag-1 autocorrelation, we use the Durbin Waston test and it indicates there is no autocorrelation for the model being added with seasonality and CPI with D Statistics close to 2.

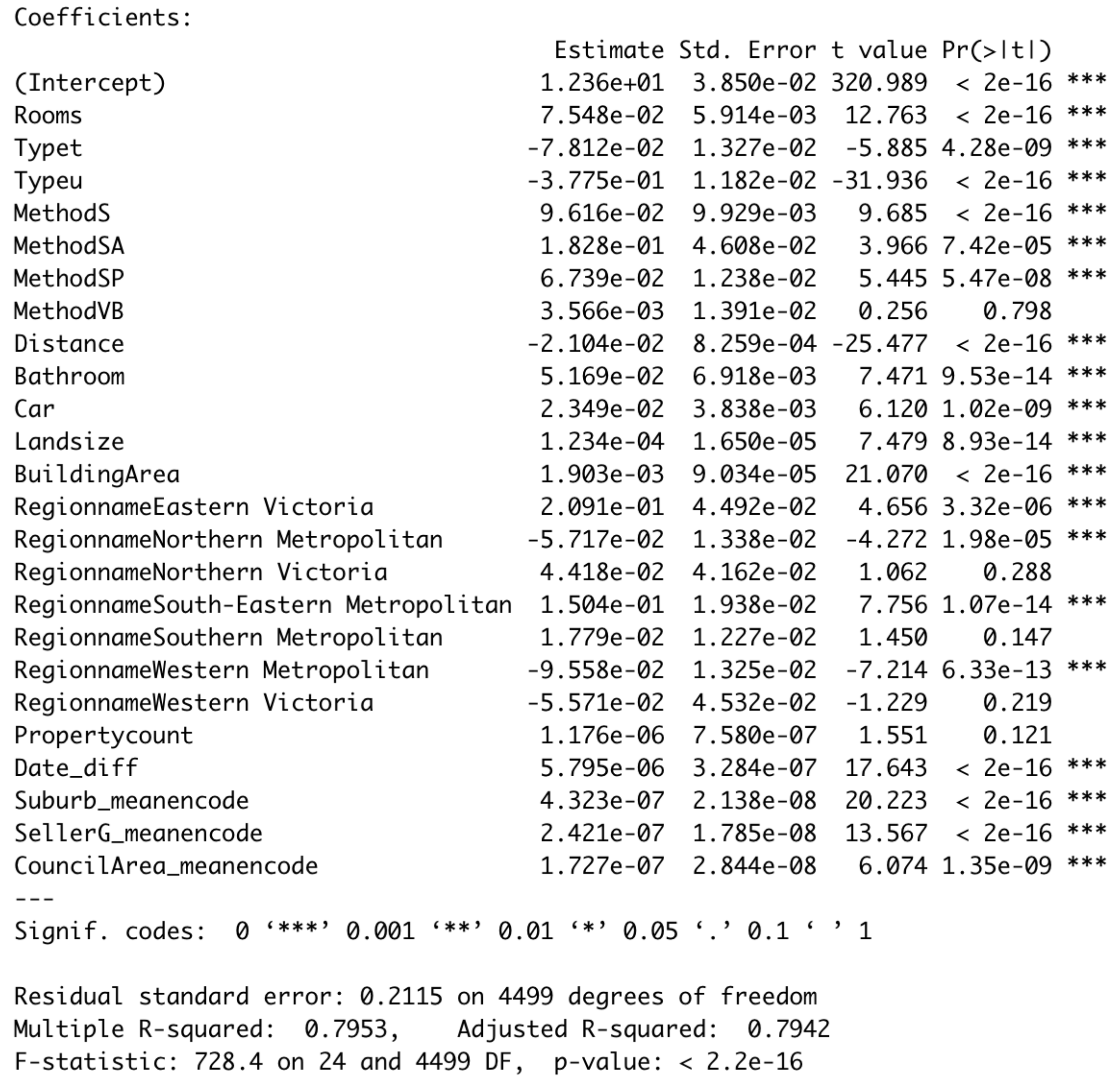
Table 1: Comparing 4 Models’ Performance in Validated Set

Validation Set	Adjusted R-squared	MAPE	AIC
Model 1	0.687	0.230	129982.2
Model 2	0.766	0.194	124030.6
Model 3 (Final model)	0.761	0.168	122548.8

[2] (2000). MEAN ABSOLUTE PERCENTAGE ERROR (MAPE) . In: Swamidass, P.M. (eds) Encyclopedia of Production and Manufacturing Management. Springer, Boston, MA . https://doi.org/10.1007/1-4020-0612-8_580

Model 4	0.764	0.165	126009.6

The regression summary shows that all coefficient estimates except property count and three other categorical dummy variables are statistically significant. One example of interpretation for a coefficient of our model with log y transformation is that as the number of rooms increases by 1, the housing price increases by 7.193%, holding all other factors constant. Having high regression coefficients, the Number of Rooms, Type of Housing, Sales Method, Bathroom, Building Size, Region, Property Age, Suburb, Seller Agent, and Council Area are important variables.



```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          1.236e+01  3.850e-02 320.989  < 2e-16 ***
Rooms                                7.548e-02  5.914e-03  12.763  < 2e-16 ***
Typet                               -7.812e-02  1.327e-02  -5.885 4.28e-09 ***
Typeu                               -3.775e-01  1.182e-02 -31.936  < 2e-16 ***
MethodS                              9.616e-02  9.929e-03   9.685  < 2e-16 ***
MethodSA                             1.828e-01  4.608e-02   3.966 7.42e-05 ***
MethodSP                             6.739e-02  1.238e-02   5.445 5.47e-08 ***
MethodVB                             3.566e-03  1.391e-02   0.256    0.798    
Distance                            -2.104e-02  8.259e-04 -25.477  < 2e-16 ***
Bathroom                             5.169e-02  6.918e-03   7.471 9.53e-14 ***
Car                                  2.349e-02  3.838e-03   6.120 1.02e-09 ***
Landsize                             1.234e-04  1.650e-05   7.479 8.93e-14 ***
BuildingArea                         1.903e-03  9.034e-05  21.070  < 2e-16 ***
RegionnameEastern Victoria           2.091e-01  4.492e-02   4.656 3.32e-06 ***
RegionnameNorthern Metropolitan     -5.717e-02  1.338e-02  -4.272 1.98e-05 ***
RegionnameNorthern Victoria          4.418e-02  4.162e-02   1.062    0.288    
RegionnameSouth-Eastern Metropolitan 1.504e-01  1.938e-02   7.756 1.07e-14 ***
RegionnameSouthern Metropolitan      1.779e-02  1.227e-02   1.450    0.147    
RegionnameWestern Metropolitan      -9.558e-02  1.325e-02  -7.214 6.33e-13 ***
RegionnameWestern Victoria          -5.571e-02  4.532e-02  -1.229    0.219    
Propertycount                        1.176e-06  7.580e-07   1.551    0.121    
Date_diff                            5.795e-06  3.284e-07  17.643  < 2e-16 ***
Suburb_meanencode                    4.323e-07  2.138e-08  20.223  < 2e-16 ***
SellerG_meanencode                   2.421e-07  1.785e-08  13.567  < 2e-16 ***
CouncilArea_meanencode               1.727e-07  2.844e-08   6.074 1.35e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2115 on 4499 degrees of freedom
Multiple R-squared:  0.7953,	Adjusted R-squared:  0.7942 
F-statistic: 728.4 on 24 and 4499 DF,  p-value: < 2.2e-16
```

Figure 16: Summary Information for Model 3

Linearity assumption, constant variance assumption, and normality assumption hold for model 3. The Q-Q plot has a good tail, and residuals in the histogram look normally distributed. The correlation matrix and linearity assumption plots from the appendix show that the factors and price have a linear relationship. Improves the residual vs. Fitted from the funeral shape to the random patterns.

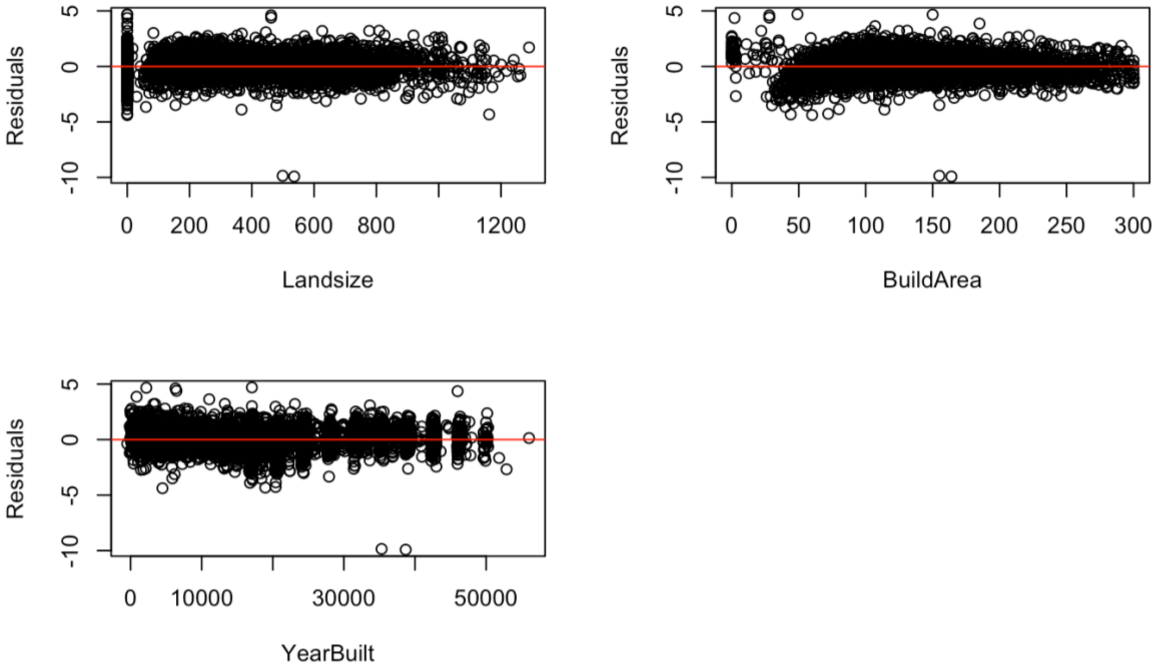


Figure 17: Model 3 Assumptions Checking

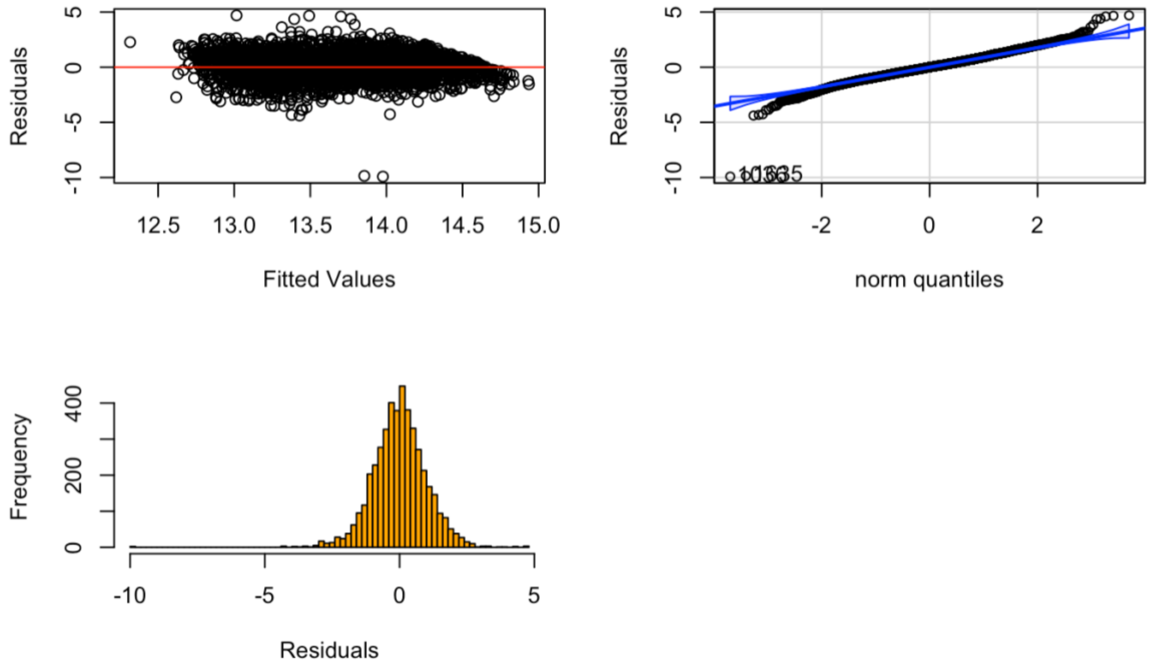


Figure 18: Model 3 Assumptions Checking

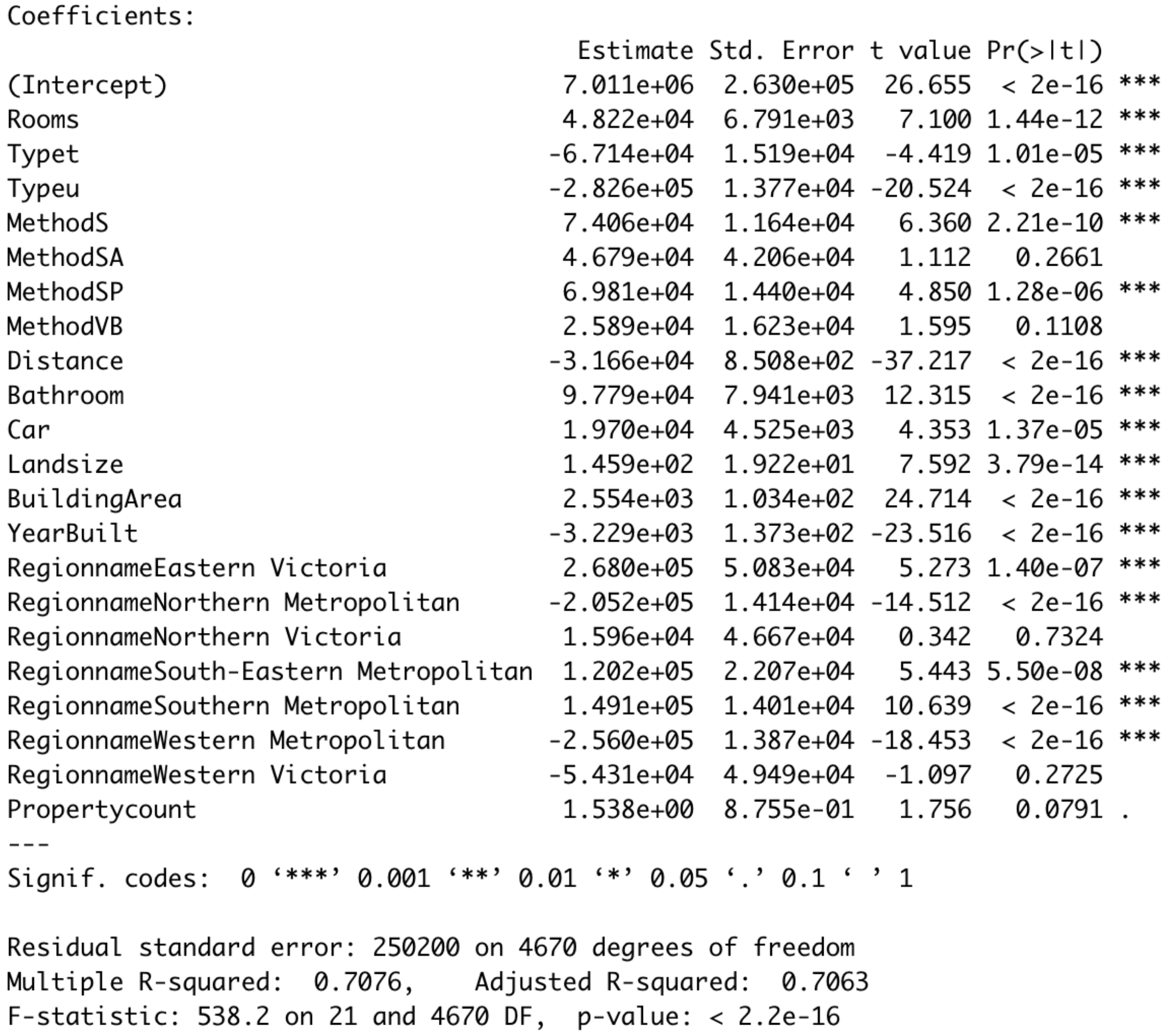
The coefficient of determination, R square, for the Model performance on the test set is 0.74, and the MAPE is 17%.

Last but not least, we find that price ranges typically between 500k and 1 million. Buyers can purchase units or duplexes in West Metro to save more, and sellers can increase their revenue by selling a house, cottage, villa, or terrace in Southeast Metro. Moreover, the variable building area is statistically significant and positively correlated to the revenue variable. We can look up the group we want based on the mean value of each group of the encoded variables. By comparing the absolute value of coefficients of lasso regression, the type of housing, the number of bathrooms, and the number of rooms are the three most important factors that predict the price value.

We applied linear models with feature engineering in this project. We performed nonlinear transformations but still failed to capture unobservable nonlinear relationships. To capture those nonlinear relationships, some future actions we can take is to try nonlinear models such as random forest and light GBM method.

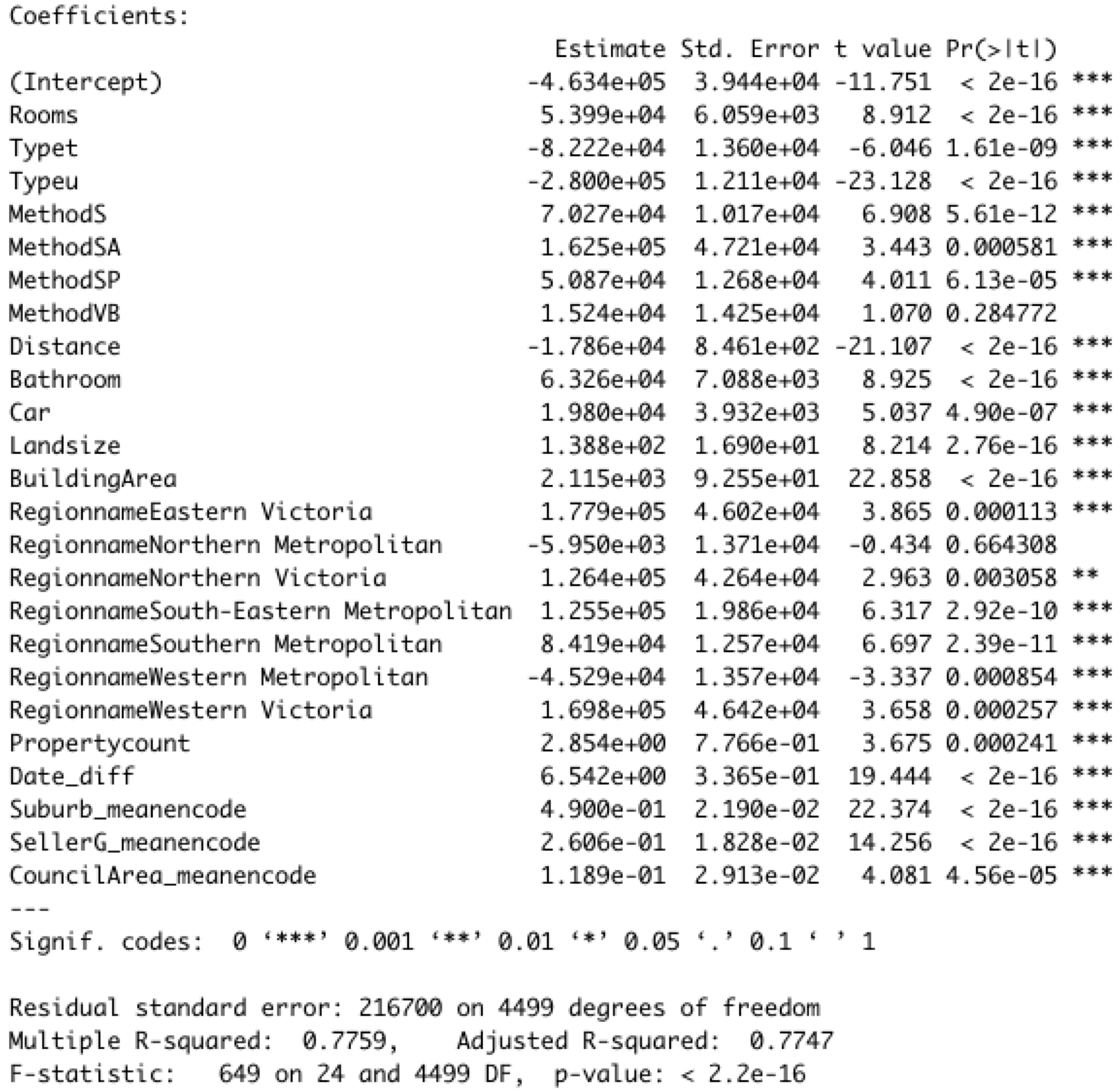
Appendix

Model 1: price ~ categories



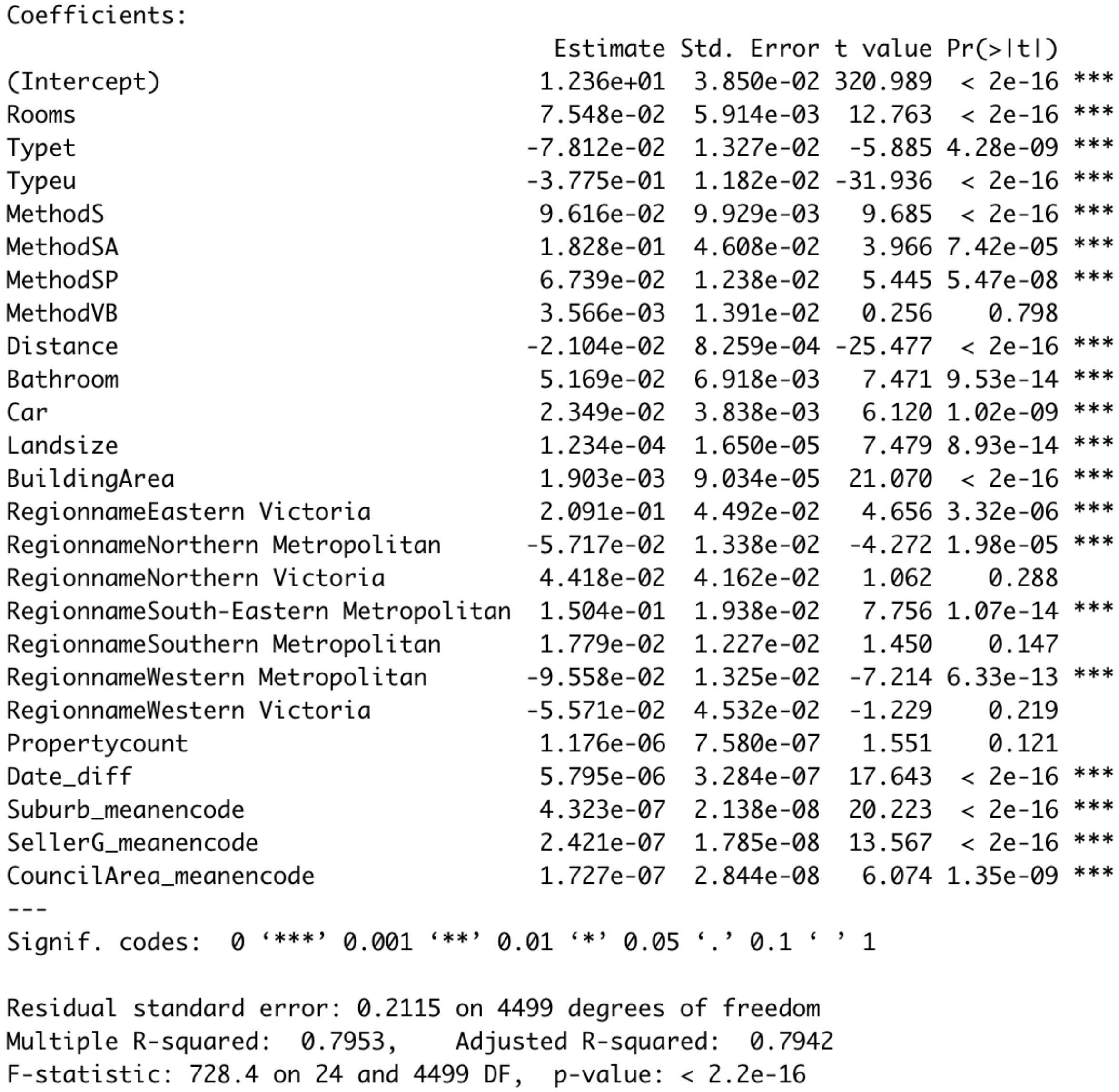
```
Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           7.011e+06  2.630e+05  26.655  < 2e-16 ***
Rooms                                 4.822e+04  6.791e+03   7.100 1.44e-12 ***
Typet                                -6.714e+04  1.519e+04  -4.419 1.01e-05 ***
Typeu                                -2.826e+05  1.377e+04 -20.524  < 2e-16 ***
MethodS                               7.406e+04  1.164e+04   6.360 2.21e-10 ***
MethodSA                              4.679e+04  4.206e+04   1.112   0.2661    
MethodSP                              6.981e+04  1.440e+04   4.850 1.28e-06 ***
MethodVB                              2.589e+04  1.623e+04   1.595   0.1108    
Distance                             -3.166e+04  8.508e+02 -37.217  < 2e-16 ***
Bathroom                              9.779e+04  7.941e+03  12.315  < 2e-16 ***
Car                                   1.970e+04  4.525e+03   4.353 1.37e-05 ***
Landsize                              1.459e+02  1.922e+01   7.592 3.79e-14 ***
BuildingArea                          2.554e+03  1.034e+02  24.714  < 2e-16 ***
YearBuilt                            -3.229e+03  1.373e+02 -23.516  < 2e-16 ***
RegionnameEastern Victoria            2.680e+05  5.083e+04   5.273 1.40e-07 ***
RegionnameNorthern Metropolitan      -2.052e+05  1.414e+04 -14.512  < 2e-16 ***
RegionnameNorthern Victoria           1.596e+04  4.667e+04   0.342   0.7324    
RegionnameSouth-Eastern Metropolitan  1.202e+05  2.207e+04   5.443 5.50e-08 ***
RegionnameSouthern Metropolitan       1.491e+05  1.401e+04  10.639  < 2e-16 ***
RegionnameWestern Metropolitan       -2.560e+05  1.387e+04 -18.453  < 2e-16 ***
RegionnameWestern Victoria           -5.431e+04  4.949e+04  -1.097   0.2725    
Propertycount                         1.538e+00  8.755e-01   1.756   0.0791 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 250200 on 4670 degrees of freedom
Multiple R-squared:  0.7076,    Adjusted R-squared:  0.7063 
F-statistic: 538.2 on 21 and 4670 DF,  p-value: < 2.2e-16
```

Model 2: price ~ categories + (date_diff + mean_encoding)



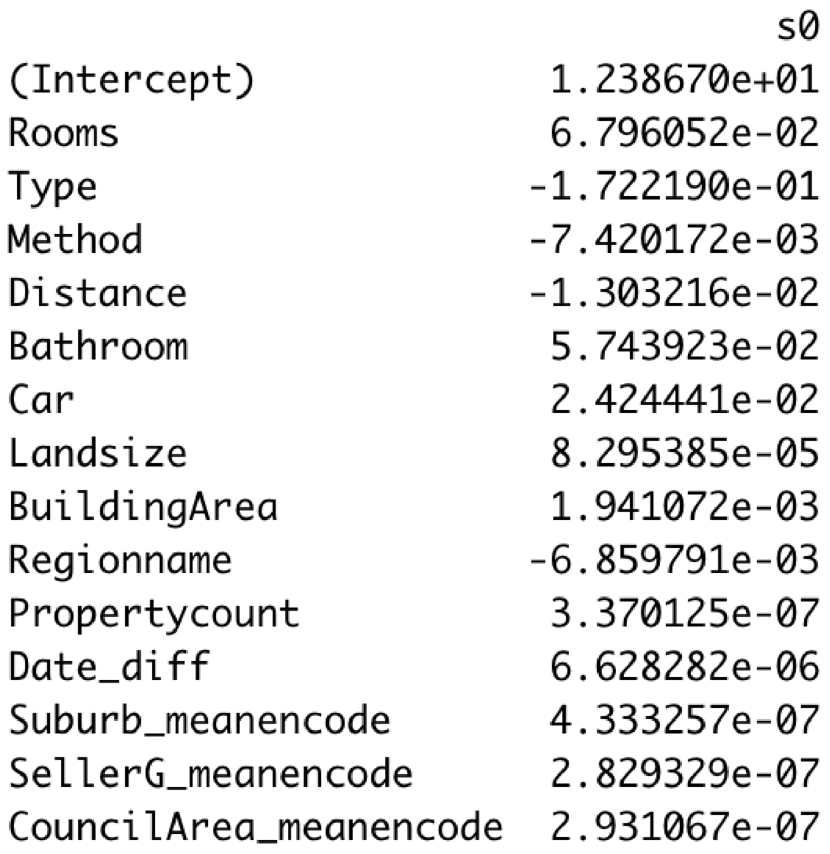
```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         -4.634e+05  3.944e+04 -11.751  < 2e-16 ***
Rooms                                5.399e+04  6.059e+03   8.912  < 2e-16 ***
Typet                               -8.222e+04  1.360e+04  -6.046 1.61e-09 ***
Typeu                               -2.800e+05  1.211e+04 -23.128  < 2e-16 ***
MethodS                              7.027e+04  1.017e+04   6.908 5.61e-12 ***
MethodSA                             1.625e+05  4.721e+04   3.443 0.000581 ***
MethodSP                             5.087e+04  1.268e+04   4.011 6.13e-05 ***
MethodVB                             1.524e+04  1.425e+04   1.070 0.284772    
Distance                            -1.786e+04  8.461e+02 -21.107  < 2e-16 ***
Bathroom                             6.326e+04  7.088e+03   8.925  < 2e-16 ***
Car                                  1.980e+04  3.932e+03   5.037 4.90e-07 ***
Landsize                             1.388e+02  1.690e+01   8.214 2.76e-16 ***
BuildingArea                         2.115e+03  9.255e+01  22.858  < 2e-16 ***
RegionnameEastern Victoria           1.779e+05  4.602e+04   3.865 0.000113 ***
RegionnameNorthern Metropolitan     -5.950e+03  1.371e+04  -0.434 0.664308    
RegionnameNorthern Victoria          1.264e+05  4.264e+04   2.963 0.003058 ** 
RegionnameSouth-Eastern Metropolitan 1.255e+05  1.986e+04   6.317 2.92e-10 ***
RegionnameSouthern Metropolitan      8.419e+04  1.257e+04   6.697 2.39e-11 ***
RegionnameWestern Metropolitan      -4.529e+04  1.357e+04  -3.337 0.000854 ***
RegionnameWestern Victoria           1.698e+05  4.642e+04   3.658 0.000257 ***
Propertycount                        2.854e+00  7.766e-01   3.675 0.000241 ***
Date_diff                            6.542e+00  3.365e-01  19.444  < 2e-16 ***
Suburb_meanencode                    4.900e-01  2.190e-02  22.374  < 2e-16 ***
SellerG_meanencode                   2.606e-01  1.828e-02  14.256  < 2e-16 ***
CouncilArea_meanencode               1.189e-01  2.913e-02   4.081 4.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 216700 on 4499 degrees of freedom
Multiple R-squared:  0.7759,    Adjusted R-squared:  0.7747 
F-statistic:   649 on 24 and 4499 DF,  p-value: < 2.2e-16
```

Model 3: log(price) ~ categories + (date_diff + mean_encoding)



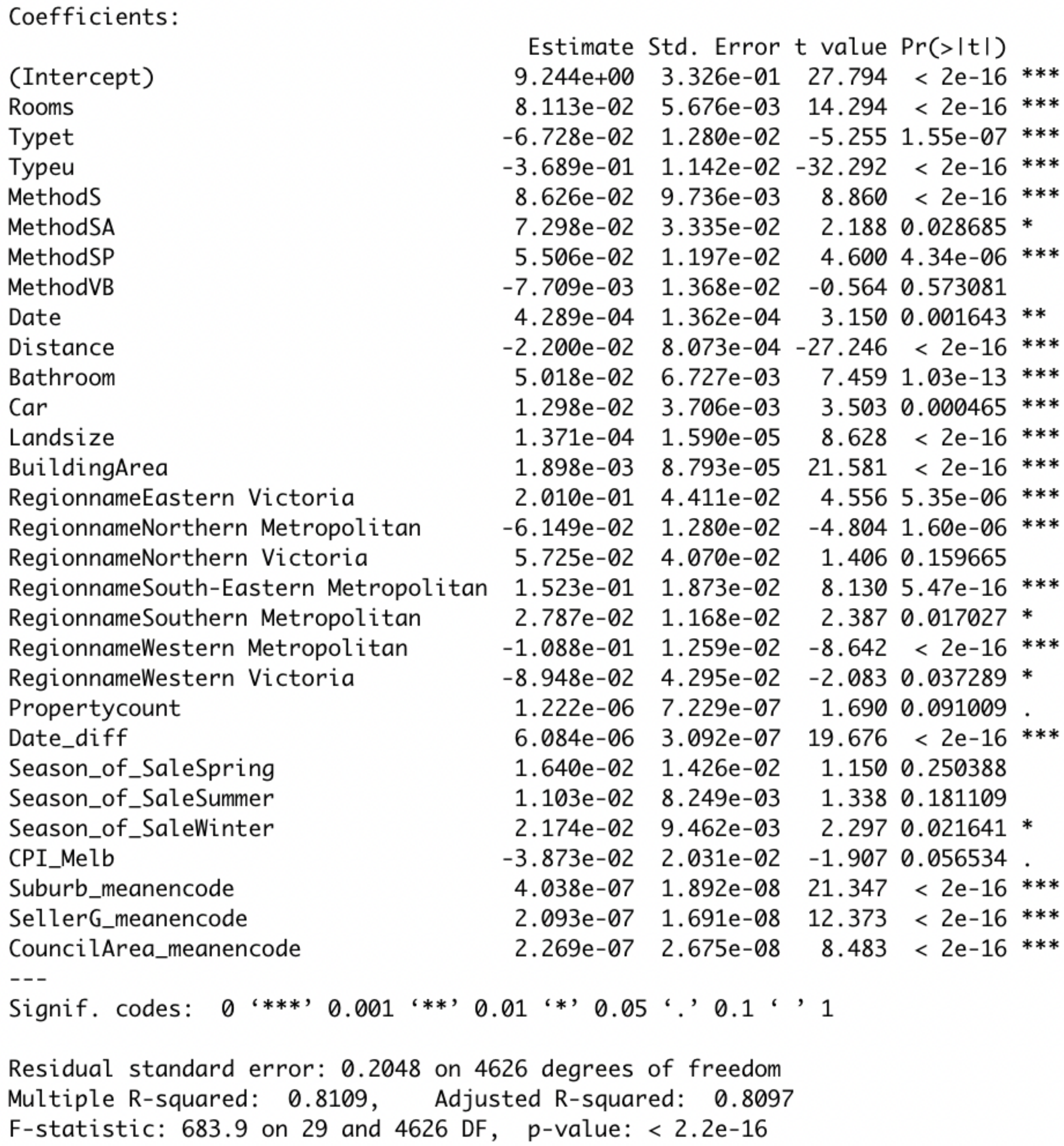
```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          1.236e+01  3.850e-02 320.989  < 2e-16 ***
Rooms                                7.548e-02  5.914e-03  12.763  < 2e-16 ***
Typet                               -7.812e-02  1.327e-02  -5.885 4.28e-09 ***
Typeu                               -3.775e-01  1.182e-02 -31.936  < 2e-16 ***
MethodS                              9.616e-02  9.929e-03   9.685  < 2e-16 ***
MethodSA                             1.828e-01  4.608e-02   3.966 7.42e-05 ***
MethodSP                             6.739e-02  1.238e-02   5.445 5.47e-08 ***
MethodVB                             3.566e-03  1.391e-02   0.256    0.798    
Distance                            -2.104e-02  8.259e-04 -25.477  < 2e-16 ***
Bathroom                             5.169e-02  6.918e-03   7.471 9.53e-14 ***
Car                                  2.349e-02  3.838e-03   6.120 1.02e-09 ***
Landsize                             1.234e-04  1.650e-05   7.479 8.93e-14 ***
BuildingArea                         1.903e-03  9.034e-05  21.070  < 2e-16 ***
RegionnameEastern Victoria           2.091e-01  4.492e-02   4.656 3.32e-06 ***
RegionnameNorthern Metropolitan     -5.717e-02  1.338e-02  -4.272 1.98e-05 ***
RegionnameNorthern Victoria          4.418e-02  4.162e-02   1.062    0.288    
RegionnameSouth-Eastern Metropolitan  1.504e-01  1.938e-02   7.756 1.07e-14 ***
RegionnameSouthern Metropolitan      1.779e-02  1.227e-02   1.450    0.147    
RegionnameWestern Metropolitan      -9.558e-02  1.325e-02  -7.214 6.33e-13 ***
RegionnameWestern Victoria          -5.571e-02  4.532e-02  -1.229    0.219    
Propertycount                        1.176e-06  7.580e-07   1.551    0.121    
Date_diff                            5.795e-06  3.284e-07  17.643  < 2e-16 ***
Suburb_meanencode                    4.323e-07  2.138e-08  20.223  < 2e-16 ***
SellerG_meanencode                   2.421e-07  1.785e-08  13.567  < 2e-16 ***
CouncilArea_meanencode               1.727e-07  2.844e-08   6.074 1.35e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2115 on 4499 degrees of freedom
Multiple R-squared:  0.7953,	Adjusted R-squared:  0.7942 
F-statistic: 728.4 on 24 and 4499 DF,  p-value: < 2.2e-16
```

Model 3(LASSO variable selection): log(price) ~ categories + (date_diff + mean_encoding)



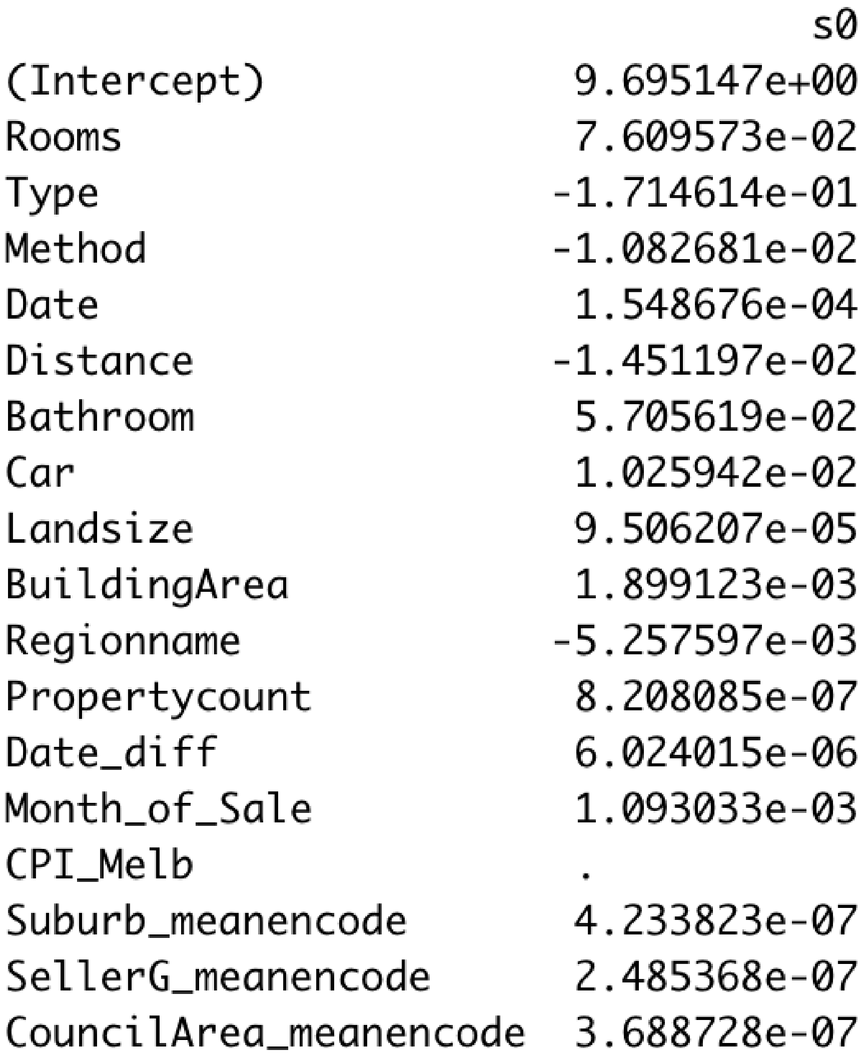
```
                                   s0
(Intercept)             1.238670e+01
Rooms                   6.796052e-02
Type                   -1.722190e-01
Method                 -7.420172e-03
Distance               -1.303216e-02
Bathroom                5.743923e-02
Car                     2.424441e-02
Landsize                8.295385e-05
BuildingArea            1.941072e-03
Regionname             -6.859791e-03
Propertycount           3.370125e-07
Date_diff               6.628282e-06
Suburb_meanencode       4.333257e-07
SellerG_meanencode      2.829329e-07
CouncilArea_meanencode  2.931067e-07
```

Model 4: log(price) ~-categories + (date_diff + mean_encoding) + (season & CPI)



```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          9.244e+00  3.326e-01  27.794  < 2e-16 ***
Rooms                                8.113e-02  5.676e-03  14.294  < 2e-16 ***
Typet                               -6.728e-02  1.280e-02  -5.255 1.55e-07 ***
Typeu                               -3.689e-01  1.142e-02 -32.292  < 2e-16 ***
MethodS                              8.626e-02  9.736e-03   8.860  < 2e-16 ***
MethodSA                             7.298e-02  3.335e-02   2.188 0.028685 *  
MethodSP                             5.506e-02  1.197e-02   4.600 4.34e-06 ***
MethodVB                            -7.709e-03  1.368e-02  -0.564 0.573081    
Date                                 4.289e-04  1.362e-04   3.150 0.001643 ** 
Distance                            -2.200e-02  8.073e-04 -27.246  < 2e-16 ***
Bathroom                             5.018e-02  6.727e-03   7.459 1.03e-13 ***
Car                                  1.298e-02  3.706e-03   3.503 0.000465 ***
Landsize                             1.371e-04  1.590e-05   8.628  < 2e-16 ***
BuildingArea                         1.898e-03  8.793e-05  21.581  < 2e-16 ***
RegionnameEastern Victoria           2.010e-01  4.411e-02   4.556 5.35e-06 ***
RegionnameNorthern Metropolitan     -6.149e-02  1.280e-02  -4.804 1.60e-06 ***
RegionnameNorthern Victoria          5.725e-02  4.070e-02   1.406 0.159665    
RegionnameSouth-Eastern Metropolitan 1.523e-01  1.873e-02   8.130 5.47e-16 ***
RegionnameSouthern Metropolitan      2.787e-02  1.168e-02   2.387 0.017027 *  
RegionnameWestern Metropolitan      -1.088e-01  1.259e-02  -8.642  < 2e-16 ***
RegionnameWestern Victoria          -8.948e-02  4.295e-02  -2.083 0.037289 *  
Propertycount                        1.222e-06  7.229e-07   1.690 0.091009 .  
Date_diff                            6.084e-06  3.092e-07  19.676  < 2e-16 ***
Season_of_SaleSpring                 1.640e-02  1.426e-02   1.150 0.250388    
Season_of_SaleSummer                 1.103e-02  8.249e-03   1.338 0.181109    
Season_of_SaleWinter                 2.174e-02  9.462e-03   2.297 0.021641 *  
CPI_Melb                            -3.873e-02  2.031e-02  -1.907 0.056534 .  
Suburb_meanencode                    4.038e-07  1.892e-08  21.347  < 2e-16 ***
SellerG_meanencode                   2.093e-07  1.691e-08  12.373  < 2e-16 ***
CouncilArea_meanencode               2.269e-07  2.675e-08   8.483  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2048 on 4626 degrees of freedom
Multiple R-squared:  0.8109,	Adjusted R-squared:  0.8097 
F-statistic: 683.9 on 29 and 4626 DF,  p-value: < 2.2e-16
```

Model 4(LASSO variable selection): log(price) ~-categories + (date_diff + mean_encoding) + (season & CPI)



```
                                       s0
(Intercept)                  9.695147e+00
Rooms                        7.609573e-02
Type                        -1.714614e-01
Method                      -1.082681e-02
Date                         1.548676e-04
Distance                    -1.451197e-02
Bathroom                     5.705619e-02
Car                          1.025942e-02
Landsize                     9.506207e-05
BuildingArea                 1.899123e-03
Regionname                  -5.257597e-03
Propertycount                8.208085e-07
Date_diff                    6.024015e-06
Month_of_Sale                1.093033e-03
CPI_Melb                     .
Suburb_meanencode            4.233823e-07
SellerG_meanencode           2.485368e-07
CouncilArea_meanencode       3.688728e-07
```