Linear regression, feature engineering, AIC process

Project of supervised learning

Statistical Learning Course Projects

By Wang Jian(Steven)

Course:Data science for economics (classe lm-data) (code: B79, class: LM-DATA - Data science for economics) Year 2

Matriculation number:14326A

Contents

Abstract

This report presents a comprehensive analysis of supervised learning techniques applied to the Melbourne housing market dataset[1]. The study focuses on linear regression models, first, through feature engineering strategies and manual selection of feature variables to define and analyze the model, and then through model optimization using the Akaike Information Criterion (AIC) process to select model automatically. By examining various models, this project aims to enhance predictive accuracy and understanding of the factors influencing property prices[2].

Statement of the Problem/Goal and Dataset Description

The primary goal of this analysis is to develop a predictive model for property prices in the Melbourne housing market using supervised learning techniques. The dataset, sourced from Kaggle, includes features like suburb, number of rooms, property type, distance from the city center, and building area. The challenge involves identifying significant predictors of price and optimizing model performance through feature engineering and selection.

Potential problem of the models designed or derived during the analytical process

A. Based on previous understanding:

1. The property price changes over time, even through short time span. All analysis does not take this into account.
2. Factors that influence the price may not exhausted by the dataset. Below are some objective factors. An example would be, taking Milan as an example, property prices are different for two properties very close but one is inside of Milan but the other is slightly outside of Milan. Or, like the insurance company may give different zones of a city a risk score, which may impact the car insurance price and also hint the block wise difference(even people's preferences on different blocks). In addition, although property age has been taken into account, the maintenance fee is not available and might impact the price.
3. Factors that influence the price subjectively, may base on the agent as the seller, or buyer's judgment, where in both cases, price would deviate from intrinsic values.

B. Based on the trial and test of different steps of the analysis

1. Scaling data may not always produce more significant statistical results or more robust models. In the meantime, interpretability may also be impacted.
2. Dummy variables are sometimes difficult to define. If the factor variable has more than 10 values, if some factors are grouped as others, the results may not be easily interpreted.

Findings/Key Points

Exclusion of outliers and log transformation of variables significantly improve model performance.

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

[2] as per building area value, since overall property value is not better measurement than per square meter price

Stepwise regression, guided by AIC, effectively identifies the most relevant features, balancing model complexity and goodness of fit.

Models incorporating geographical location (distance from the center) and property characteristics (such as type and size) provide better predictions of housing prices.

The application of log transformations to price-related variables helps in normalizing data distribution, enhancing the model's predictive accuracy.

In general, automatic feature selection process is more efficient than manual selection process, but there is still room for models produced from such processes to improve by manual intervention, either resolving violations to model assumptions or improve the explainability of the models.

Analysis and Commentary

The analysis began with data cleaning and preparation, followed by the development of multiple linear regression models. Techniques such as outlier removal, log transformation, and feature scaling were employed to refine the models. The AIC process was utilized for systematic feature selection, leading to the identification of models that offer a good trade-off between complexity and predictive performance.

Theoretical Background

The study relies on linear regression theory, which assumes a linear relationship between the dependent and independent variables. Feature engineering techniques like log transformation and scaling were used to meet model assumptions and improve interpretability. The AIC method, a cornerstone of the analysis, helps in selecting a model that adequately balances model fit and complexity.

Conclusions

The analysis demonstrates the effectiveness of combining feature engineering with AIC-guided stepwise regression[3] in improving the predictive performance of linear regression models[4]. Key findings include the importance of treating outliers, the value of log transformation, and the significance of geographical and property-specific variables in predicting housing prices. Detailed results are in the following appendixes and the R markdown files.

It is worth noting that either real estate manager who sells properties or buyers of properties could potentially benefit from the analytical results, to improve business strategy, segregate and

[3] In the AIC process part, there could be further explored, with classification dummy inclusion, distance variable inclusion, to conduct multiple tests of AIC result and additional models.

[4] Although in both stepwise processes, the final models both have two variables with GVIF value higher than 5(some higher than 10), adjusted GVIF of all variables are not concerning. However, there still exist obvious multicollinearity issue and the models could potentially be further improved manually, since the AIC method stopped at the presented states.

focus on market sections[5], or utilize the information to better decide weather a purchase decision is statistically sound or requires further consideration[6].

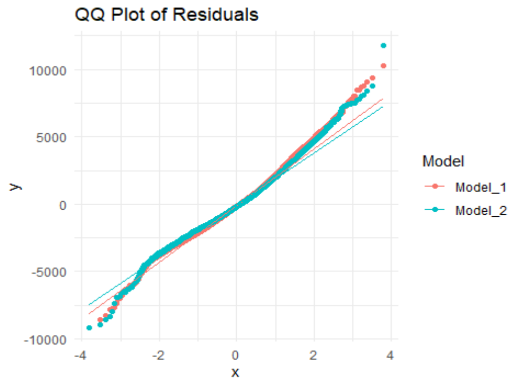


Figure 1 CLASSIFICATION AND REGRESSION WITH EXCLUSION OF OUTLIER

Model 1Geo KNN Dummy[7]; Model 2 Distance[8]

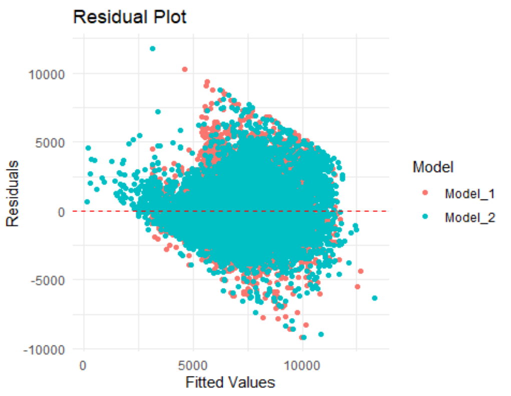


Figure 2 CLASSIFICATION AND REGRESSION WITH EXCLUSION OF OUTLIER

Model 1Geo KNN Dummy; Model 2 Distance

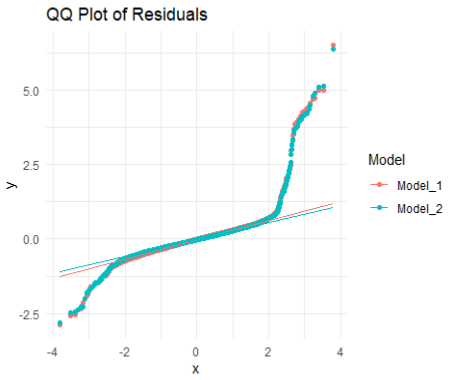


Figure 3 CLASSIFICATION AND REGRESSION WITH LOG

Model 1Geo KNN Dummy[9]; Model 2 Distance[10]

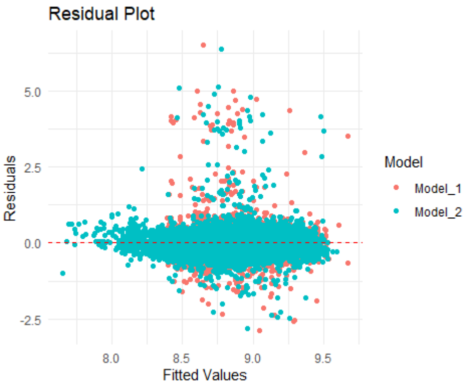


Figure 4 CLASSIFICATION AND REGRESSION WITH LOG

Model 1Geo KNN Dummy; Model 2 Distance

[5] Further sub setting dataset to derive additional models, or using models with classification to interpret sectional market characteristics. As for detailed market subsection analysis, unsupervised learning part, clustering, would be an alternative.

[6] LogPricePerBuildingArea ~ Type + Distance + Bedroom2 + Bathroom + Car + Landsize + BuildingArea + CouncilArea + YearsAfterBuilt
By choosing the optimal model, type, distance, bedroom2, bathroom, car, landsize, buildingarea, councilarea, should all be taken into consideration from both buyer and seller side.

[7] lm(Priceperbuildingarea ~ ., data = trainData_trimmed_1)

[8] lm(Priceperbuildingarea ~ ., data = trainData_trimmed_2)

[9] lm(formula = LogPriceperbuildingarea ~ ., data = trainData_trimmed_1)
trainData_trimmed_1 <- subset(trainData_1, select = c(Class, YearsSinceBuilt, LogPriceperbuildingarea))

[10] lm(LogPriceperbuildingarea ~ ., data = trainData_trimmed_2)
trainData_trimmed_2=subset(trainData_2, select = c(distance_from_center, YearsSinceBuilt, LogPriceperbuildingarea))

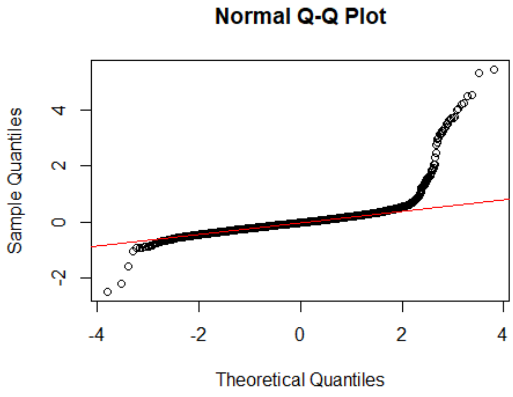


Figure 5 Stepwise Regression (Best Var Comb Log)[11]

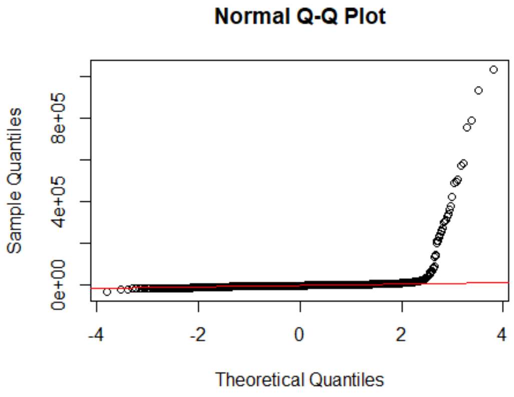


Figure 7 Stepwise Regression (Best Var)[12]

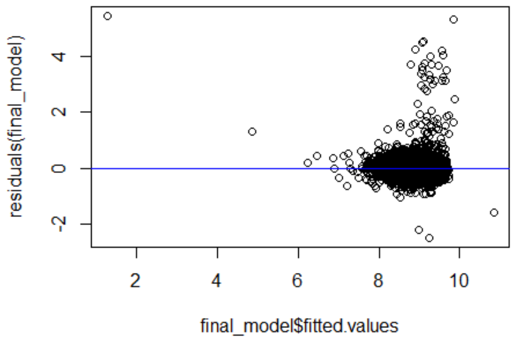


Figure 6 Stepwise Regression (Best Var Comb Log)

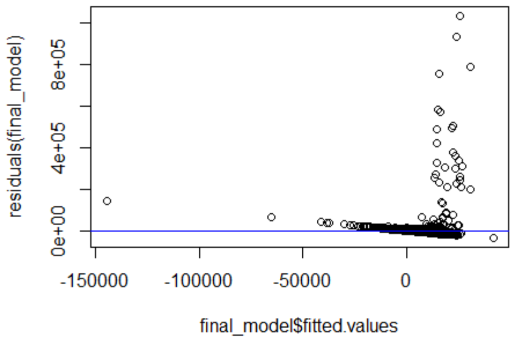


Figure 8 Stepwise Regression (Best Var)

[11] final_model <- lm(LogPricePerBuildingArea ~ Type + Distance + Bedroom2 + Bathroom + Car + Landsize + BuildingArea + CouncilArea + YearsAfterBuilt, data = train_data)

[12] final_model <- lm(PricePerBuildingArea ~ Rooms + Type + Distance + Bedroom2 + Bathroom + Car + BuildingArea + YearsAfterBuilt, data = train_data)

Appendix

The appendix would contain brief processing steps and all R code used in the project, detailing each step of the data cleaning, preparation, clustering, and analysis process. This includes code for data manipulation, outlier detection, clustering, and result visualization.

Appendix 1, comparison table

The appendix contains all R code used in the project, detailing data cleaning, model building, feature engineering, and model selection processes, thereby providing a comprehensive guide to the analytical approach adopted in this study.

	CLASSIFICATION AND REGRESSION WITH EXCLUSION OF OUTLIER		CLASSIFICATION AND REGRESSION WITH LOG			
METRICS / MODELS	**Geo KNN Dummy**	**Distance**	**Geo KNN Dummy**	**Distance**	**Stepwise Regression (Best Var Comb Log)**	**Stepwise Regression (Best Var)**
RESIDUAL STANDARD ERROR	2211 (on 6827 DF)	2070 (on 6829 DF)	0.464 (on 7069 DF)	0.4416 (on 7071 DF)	0.3449 (on 7042 DF)	29880 (on 7065 DF)
MULTIPLE R-SQUARED	0.3621	0.441	0.2201	0.2934	0.5565	0.02958
ADJUSTED R-SQUARED	0.3618	0.4408	0.2197	0.2932	0.5539	0.02834
F-STATISTIC (AND DF)	968.9 (on 4 and 6827 DF)	2693 (on 2 and 6829 DF)	498.7 (on 4 and 7069 DF)	1468 (on 2 and 7071 DF)	215.5 (on 41 and 7042 DF)	23.93 (on 9 and 7065 DF)
P-VALUE	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
FEATURE ENGINEERING, TRANSFORMATION AND SELECTION	Feature engineering: create dummy on geo classification Features: manually selected Outlier: filtered using 1.5*IQR away from Q1 and Q3	Feature engineering: create distance from center Features: manually selected Outlier: filtered using 1.5*IQR away from Q1 and Q3	Feature engineering: • create dummy on geo • Log transform Y label variable classification Features: manually selected Outlier: No filter	Feature engineering: • create distance from center • Log transform Y label variable Features: manually selected Outlier: No filter	Feature engineering: • create distance from center • Log transform Y label variable • Scale feature X variables Features: AIC variable selection Outlier: No filter	Feature engineering: • create distance from center • Scale feature X variables Features: AIC variable selection Outlier: No filter

Appendix 2, Model Processing

Based on the supervised learning techniques from the lectures and the dataset from Kaggle about Melbourne-housing-market[13], the above models or process to identify optimal model have been used.

The base idea is to train a linear regression model with below variations.

1. Identify Y lable as price per buildingarea
2. Classify groups of property and produce a classification dummy variable to perform regression
3. Calculate distance variable to perform regression
4. Treat outliers in quartile elimination to create variation of modeling
5. Treat Y label in log term to create variation of modeling

The above steps have been updated during the experimental process in order to improve the robustness of the model. Then, stepwise AIC methodology has been used to following steps.

1. Stepwise regression process to identify optimal linear regression model
2. Treat Y lable in log term and normal term(scale only the feature variables)

Appendix 3, Comparison of the models:

Typically, the application of stepwise variable selection utilizing the Akaike Information Criterion (AIC) method is considered superior to models constructed with manually selected variables. This is due to its systematic approach in identifying the most statistically significant variables while penalizing model complexity.

Additionally, the exclusion of outliers in linear models often results in an enhancement of the adjusted R-squared value, thereby improving the overall explanatory power and reliability of the model. But, as shown in the result, this strategy can not capture the whole picture well, and, is less efficient than AIC method to improve model efficiency on R square and adjusted R square.

Residual Standard Error: Lower values indicate better fit. The "Stepwise Regression (Best Var Comb Log)" model has the lowest residual standard error among the regression models, suggesting better fit compared to others. The "Distance" approach in the exclusion of outlier context also shows improvement over the "Geo KNN Dummy" in the same context.

Multiple R-squared and Adjusted R-squared: Higher values indicate a model explains more variance of the dependent variable. The "Stepwise Regression (Best Var Comb Log)" model has significantly higher values than others, indicating a superior fit to the data. The "Distance" models generally outperform the "Geo KNN Dummy" models, indicating a better explanatory power.

F-statistic and Degrees of Freedom: Higher F-statistic values suggest a model's explanatory variables collectively significantly predict the outcome variable. The "Distance" model in the

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

exclusion of outlier context and the "Stepwise Regression (Best Var Comb Log)" show strong significance levels, with high F-statistics relative to their degrees of freedom[14].

p-value: All models show significant p-values (< 2.2e-16), indicating the models' predictors are, collectively, significantly different from zero. However, this does not compensate for low R-squared values in some models, such as the "Stepwise Regression (Best Var)".

In summary, while all models show statistical significance, the "Stepwise Regression (Best Var Comb Log)" stands out for its lower residual standard error and higher R-squared values, suggesting it provides the best fit and explanatory power among the models compared.

classification-and-regression-with-exclusion-of-outlier

The comparison of metrics for Model 1 and Model 2 is as follows:

Model 1 Metrics:

R-squared: 0.3621. This value indicates that approximately 36.21% of the variance in the dependent variable is explained by the model.

Adjusted R-squared: 0.3618. This is a modification of the R-squared that adjusts for the number of predictors in the model, providing a more accurate measure in the context of multiple variables.

Residual Standard Error: 2211. This value represents the standard deviation of the residuals, indicating the average distance that the observed values fall from the regression line.

F-statistic: 968.9. This value tests the overall significance of the regression model. A higher F-statistic indicates a more significant predictive capability of the model variables.

```
## Call:
## lm(formula = Priceperbuildingarea ~ ., data = trainData_trimmed_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9142.2 -1537.9  -169.2  1298.5 10299.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5354.0584    69.6539  76.867   <2e-16 ***
## Class2            26.4614    72.5400   0.365    0.715
## Class3           917.8225    79.4400  11.554   <2e-16 ***
## Class4         -2329.5063   114.2920 -20.382   <2e-16 ***
## YearsSinceBuilt   38.3039     0.7665  49.972   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2211 on 6827 degrees of freedom
```

[14] Compared with other models, this model has the lowest F value. But a VIF test has been done to ensure multicollinearity should not be a significant concern.

```
## Multiple R-squared:  0.3621, Adjusted R-squared:  0.3618
## F-statistic: 968.9 on 4 and 6827 DF,  p-value: < 2.2e-16
```

Model 2 Metrics:

R-squared: 0.441. This indicates that about 44.1% of the variance in the dependent variable is explained by the model, which is higher than that of Model 1.

Adjusted R-squared: 0.4408. Similar to Model 1, this adjusts for the number of predictors but is higher than Model 1, suggesting a better fit given the number of variables.

Residual Standard Error: 2070. This is lower than in Model 1, suggesting that Model 2's predictions are, on average, closer to the actual values.

F-statistic: 2693. This significantly higher value compared to Model 1 suggests that the variables in Model 2 have a stronger combined effect on predicting the dependent variable.

```
## Call:
## lm(formula = Priceperbuildingarea ~ ., data = trainData_trimmed_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9173.5 -1389.2  -185.5  1214.2 11771.7
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7594.6911    67.6008  112.35   <2e-16 ***
## distance_from_center -146.4118     3.3025  -44.33   <2e-16 ***
## YearsSinceBuilt        31.2631     0.7414   42.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2070 on 6829 degrees of freedom
## Multiple R-squared:  0.441,  Adjusted R-squared:  0.4408
## F-statistic:  2693 on 2 and 6829 DF,  p-value: < 2.2e-16
```

From these metrics, it can be concluded that Model 2 generally provides a better fit to the data than Model 1, as indicated by its higher R-squared and Adjusted R-squared values, and lower Residual Standard Error. The higher F-statistic in Model 2 also indicates a stronger overall significance of the model.

classification-and-regression-with-log

Model 1 Metrics:

R-squared: 0.2201, indicating that approximately 22.01% of the variance in LogPriceperbuildingarea is explained by this model.

Adjusted R-squared: 0.2197, slightly lower than the R-squared, adjusted for the number of predictors in the model.

Residual Standard Error (RSE): 0.464, representing the average distance that the observed values deviate from the regression line.

F-statistic: 498.7, suggesting the model's explanatory variables significantly explain the variation in LogPriceperbuildingarea.

```
##                                                                       Call:
## lm(formula = LogPriceperbuildingarea ~ ., data = trainData_trimmed_1)
##
##                                                                  Residuals:
##          Min             1Q     Median             3Q                  Max
##     -2.8777     -0.2358     -0.0054           0.1996               6.5075
##
##                                                               Coefficients:
##                               Estimate  Std.  Error  t  value  Pr(>|t|)
##  (Intercept)           8.3889617    0.0141821    591.52      <2e-16  ***
## Class2                   0.3830909    0.0164309     23.32      <2e-16  ***
## Class3                   0.1903375    0.0154050     12.36      <2e-16  ***
## Class4                   0.2301495    0.0193324     11.90      <2e-16  ***
##  YearsSinceBuilt  0.0054086     0.0001555       34.77       <2e-16  ***
##                                                                         ---
## Signif.  codes:   0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
##
##  Residual   standard   error:   0.464   on   7069   degrees   of   freedom
##   Multiple   R-squared:     0.2201,   Adjusted   R-squared:     0.2197
## F-statistic: 498.7 on 4 and 7069 DF,  p-value: < 2.2e-16
```

Model 2 Metrics:

R-squared: 0.2934, showing an improvement over Model 1, with approximately 29.34% of the variance in LogPriceperbuildingarea explained by the model.

Adjusted R-squared: 0.2932, closely aligns with the R-squared, indicating a good fit for the number of predictors.

Residual Standard Error (RSE): 0.442, lower than Model 1, suggesting that predictions are closer to the actual values on average.

F-statistic: 1468, significantly higher than in Model 1, demonstrating a stronger overall significance of the regression model.

```
## Call:
## lm(formula = LogPriceperbuildingarea ~ ., data = trainData_trimmed_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7966 -0.2100 -0.0236  0.1743  6.3809
##
```

```
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           8.9899464  0.0144129  623.74   <2e-16 ***
## distance_from_center -0.0259987  0.0007086  -36.69   <2e-16 ***
## YearsSinceBuilt       0.0039851  0.0001545   25.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4416 on 7071 degrees of freedom
## Multiple R-squared:  0.2934, Adjusted R-squared:  0.2932 
## F-statistic:  1468 on 2 and 7071 DF,  p-value: < 2.2e-16
```

Comparative Analysis:

Predictive Power: Model 2 exhibits a higher R-squared and Adjusted R-squared value than Model 1, implying it better accounts for the variance in LogPriceperbuildingarea.

Precision of Predictions: The lower Residual Standard Error in Model 2 indicates it makes more accurate predictions compared to Model 1.

Model Significance: The substantially higher F-statistic in Model 2 indicates a more statistically significant relationship between the predictors and the dependent variable than in Model 1.

In summary, Model 2 provides a better fit for the data, explaining a higher percentage of the variance in LogPriceperbuildingarea, and offers more precise predictions than Model 1. This suggests that the variables used in Model 2 have a stronger and more significant impact on predicting LogPriceperbuildingarea.

stepwise-regression-process-to-find-best-var-comb-log

The summary of the models and the AIC (Akaike Information Criterion) process is as follows:

1. **Data Cleaning**: The training data is cleaned to remove any rows where the target variable, `LogPricePerBuildingArea`, contains NA, NaN, or Inf values. This ensures that the model is trained on valid, non-missing data.

2. **Model Training - Stepwise Regression**: A stepwise regression model is trained using the 'backward' direction. The initial model includes predictors such as Rooms, Type, Distance, Bedroom2, Bathroom, Car, Landsize, BuildingArea, CouncilArea, Propertycount, and YearsAfterBuilt. The AIC of this initial model is -15035.86.

3. **AIC Improvement Process**: The stepwise process evaluates the removal or retention of variables based on AIC values. For example, removing 'Rooms' improves the AIC to -15038, and removing 'Propertycount' further improves it to -15040. This process identifies the most statistically significant variables while aiming for a model with the lowest possible AIC, indicating a better fit.

4. **Final Model Selection**: The final model includes variables: Type, Distance, Bedroom2, Bathroom, Car, Landsize, BuildingArea, CouncilArea, and YearsAfterBuilt. The AIC after the final

step is -15039.78, indicating this model has a better balance of complexity and fit compared to the initial model.

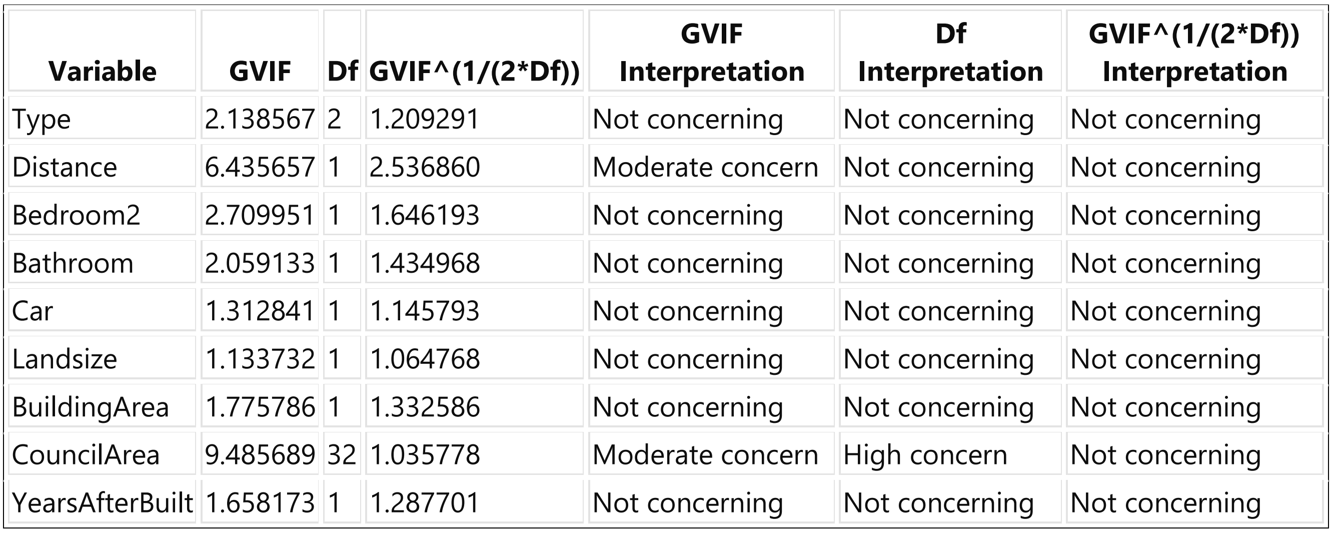
5. **Model Evaluation**: The model is evaluated using RMSE (Root Mean Square Error) on the test data. However, an issue is noted as the RMSE is reported as Inf, indicating a problem with the prediction process or the test data.

6. **AIC Visualization**: The AIC values across different model iterations are plotted against the number of variables. Initially, the model starts with 12 variables and an AIC of -15035.86. After the first step of removal, it goes down to 11 variables with an AIC of -15038, and then to 10 variables with an AIC of -15039.78, showcasing the stepwise optimization process.

7. **Final Model Summary**: The final regression model details are provided, including coefficients for each predictor. For instance, 'BuildingArea' has a significant negative coefficient of -0.246039, indicating its strong inverse relationship with 'LogPricePerBuildingArea'. Various 'CouncilArea' coefficients show the diverse impact of different regions on property prices. The model also confirms significant predictors like 'Distance' and 'Type', with respective coefficients of -0.199143 for each unit increase in distance and adjustments for different types of properties.

```
lm(formula = LogPricePerBuildingArea ~ Type + Distance + Bedroom2 +
##     Bathroom + Car + Landsize + BuildingArea + CouncilArea +
##     YearsAfterBuilt, data = train_data)
## Residual standard error: 0.3449 on 7042 degrees of freedom
## Multiple R-squared:  0.5565, Adjusted R-squared:  0.5539
## F-statistic: 215.5 on 41 and 7042 DF,  p-value: < 2.2e-16
```

Additional check on VIF:



Variable	GVIF	Df	GVIF^(1/(2*Df))	GVIF Interpretation	Df Interpretation	GVIF^(1/(2*Df)) Interpretation
Type	2.138567	2	1.209291	Not concerning	Not concerning	Not concerning
Distance	6.435657	1	2.536860	Moderate concern	Not concerning	Not concerning
Bedroom2	2.709951	1	1.646193	Not concerning	Not concerning	Not concerning
Bathroom	2.059133	1	1.434968	Not concerning	Not concerning	Not concerning
Car	1.312841	1	1.145793	Not concerning	Not concerning	Not concerning
Landsize	1.133732	1	1.064768	Not concerning	Not concerning	Not concerning
BuildingArea	1.775786	1	1.332586	Not concerning	Not concerning	Not concerning
CouncilArea	9.485689	32	1.035778	Moderate concern	High concern	Not concerning
YearsAfterBuilt	1.658173	1	1.287701	Not concerning	Not concerning	Not concerning

stepwise-regression-process-to-find-best-var-comb

In the second project, the process of training a stepwise regression model for predicting PricePerBuildingArea is outlined, along with the steps involved in model selection based on the Akaike Information Criterion (AIC):

Initial Model: The starting AIC is 146189.8 with the full model including predictors: Suburb, Rooms, Type, Distance, Bedroom2, Bathroom, Car, Landsize, BuildingArea, CouncilArea, Propertycount, Date, and YearsAfterBuilt.

Model Refinement Steps:

The first step removes 'Propertycount', maintaining an AIC of 146189.8, indicating no improvement from the removal.

Subsequent steps involve evaluating the removal of different variables while monitoring the AIC for the best model fit. Notably, 'Suburb' is removed leading to a significant drop in AIC to 145831.9, suggesting a better model fit without this variable.

The removal of 'Date' further reduces the AIC to 145830.6, indicating another improvement in model performance.

This iterative process continues, with variables like 'Landsize' being removed to achieve a lower AIC of 145826.7, suggesting incremental improvements in model simplicity and fit.

Final Model Selection: The concluding step of the regression ends with the variables: Rooms, Type, Distance, Bedroom2, Bathroom, Car, BuildingArea, and YearsAfterBuilt, achieving an AIC of 145826.7, reflecting the most balanced model in terms of complexity and goodness of fit from the stepwise process.

Model Evaluation: The final model's performance is evaluated using the RMSE on the test data, where an issue is identified as the RMSE is reported as 'Inf', indicating a possible anomaly or extreme values in predictions or the test dataset.

Visualization of AIC Values: The AIC values are plotted against the number of variables through different model iterations, starting from 13 variables (AIC = 146191.6) and ending with 7 variables (AIC = 145826.6). This graphically illustrates the model selection process, highlighting the trade-off between model complexity and fit.

Final Model Summary: The regression summary for the final model reveals coefficients for each predictor, indicating their impact on PricePerBuildingArea. For instance, 'BuildingArea' has a significant negative impact (coefficient = -4803.9), while 'Type' (specific categories like 't') shows a positive association with the target variable. The model explains a small portion of the variance in the data, as indicated by a Multiple R-squared value of 0.02958.

```
## lm(formula = PricePerBuildingArea ~ Rooms + Type + Distance +
##     Bedroom2 + Bathroom + Car + BuildingArea + YearsAfterBuilt,
##     data = train_data)
## Residual standard error: 29880 on 7065 degrees of freedom
## Multiple R-squared:  0.02958,    Adjusted R-squared:  0.02834
## F-statistic: 23.93 on 9 and 7065 DF,  p-value: < 2.2e-16
```

Additional check on VIF:

Variable	GVIF	Df	GVIF^(1/(2*Df))	GVIF Interpretation	Df Interpretation	GVIF^(1/(2*Df)) Interpretation
Rooms	16.533716	1	4.066167	High concern	Not concerning	Not concerning

Variable	GVIF	Df	GVIF^(1/(2*Df))	GVIF Interpretation	Df Interpretation	GVIF^(1/(2*Df)) Interpretation
Type	1.912470	2	1.175977	Not concerning	Not concerning	Not concerning
Distance	1.354880	1	1.163993	Not concerning	Not concerning	Not concerning
Bedroom2	16.261664	1	4.032575	High concern	Not concerning	Not concerning
Bathroom	1.980821	1	1.407416	Not concerning	Not concerning	Not concerning
Car	1.268326	1	1.126200	Not concerning	Not concerning	Not concerning
BuildingArea	1.774341	1	1.332044	Not concerning	Not concerning	Not concerning
YearsAfterBuilt	1.496941	1	1.223495	Not concerning	Not concerning	Not concerning

Appendix 4, R code as attached