

**Log Book Entry 2**

**Analysis of Relationships between House Characteristics and Sale Price using Ames Housing Data**

Joseph Deery

40078793

Contents

[1.0 Introduction 2](#_Toc56499222)

[2.0 Measures of Association 2](#_Toc56499223)

[3.0 Regression Analysis 3](#_Toc56499224)

[4.0 Summary of Insights 9](#_Toc56499225)

[5.0 Reflective Commentary 10](#_Toc56499228)

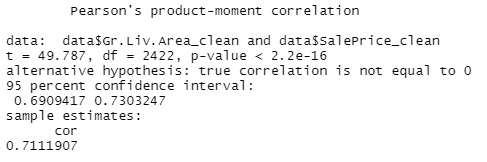
[Appendix 1: R Code Used 11](#_Toc56499229)

# Introduction

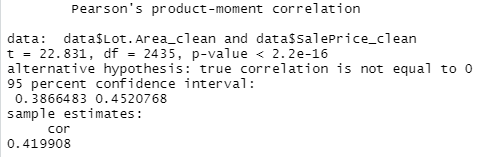
This Log Book entry builds on the foundations laid in Log Book 1. Having successfully addressed quality issues, provided descriptive statistics and visualisations, this entry will focus on using inference and prediction through association and regression to provide a more in-depth examination of the factors that influenced the sale price of a house in the Ames area in 2010. Logically, and based on the explorations carried out in Log Book 1, some variables appear to have more of an influence than others on sale price. 5 key independent variables that will be used in measuring association are Ground Floor Living Area, Total Rooms, Lot Area, Overall Quality and Number of Bedrooms. These variables will also form the baseline for a regression model, which will be checked for both predicted accuracy and assumption before being interpreted in section 4.0.

# Measures of Association

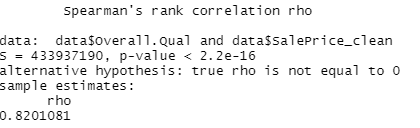
**Figure 1**: Correlation between Ground Floor living area and Sale Price



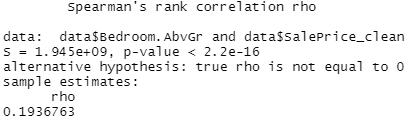
**Figure 2**: Correlation between Lot Area and Sale Price



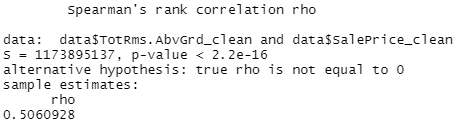
**Figure 3**: Correlation between Overall Quality Rating and Sale Price



**Figure 4:** Correlation between number of bedrooms and Sale Price

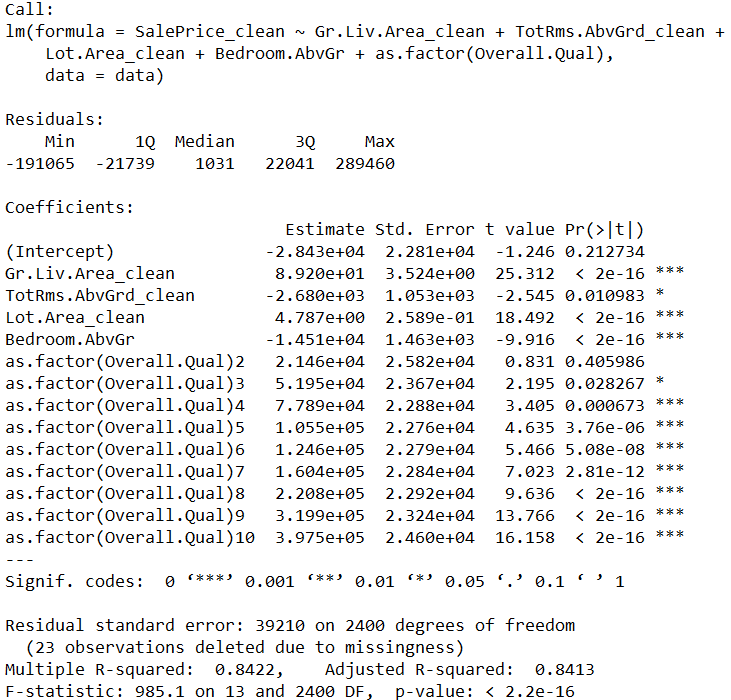


**Figure 5:** Correlation between Total number of rooms and Sale Price

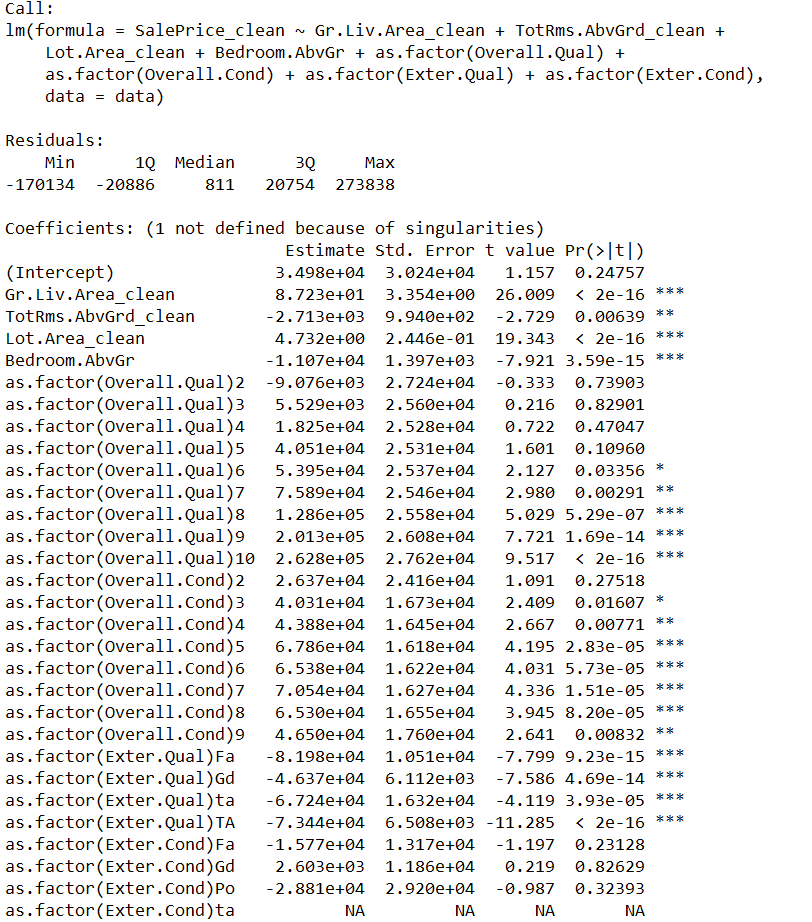


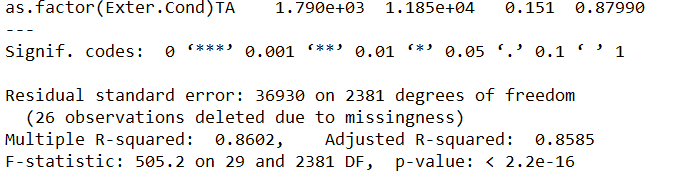
# Regression Analysis

**Figure 6:** Model 1 (baseline model) Summary Output

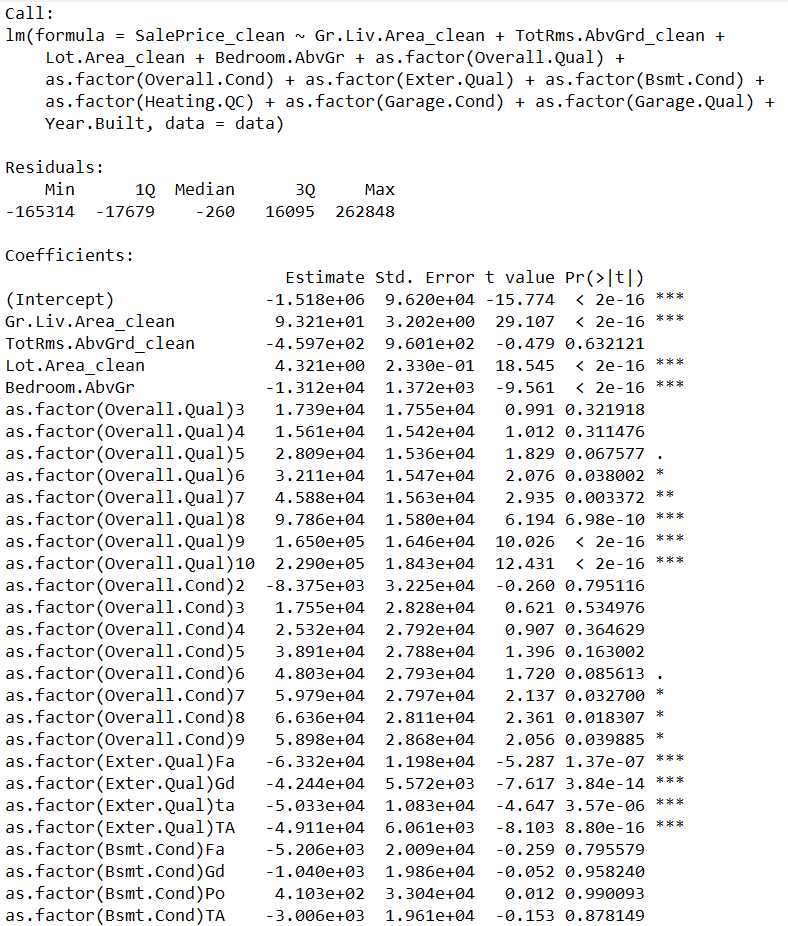
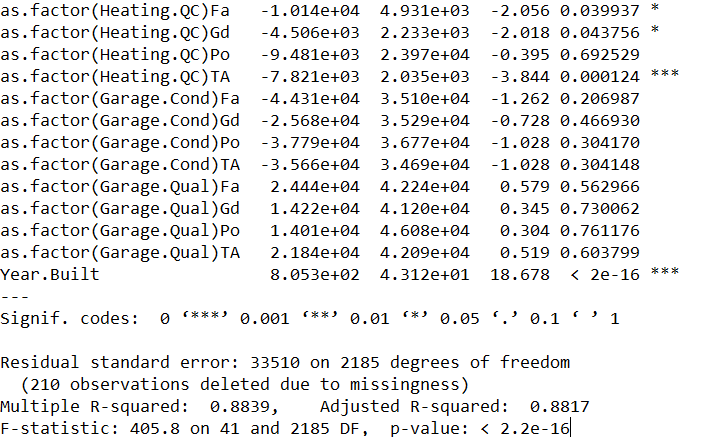


**Figure 7:** Model 2 Summary Output





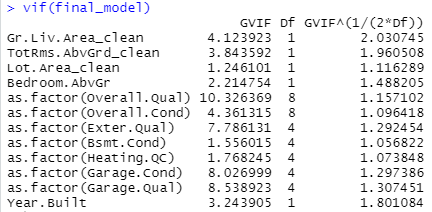
**Figure 8**: Model 3 (Final model) Summary Output



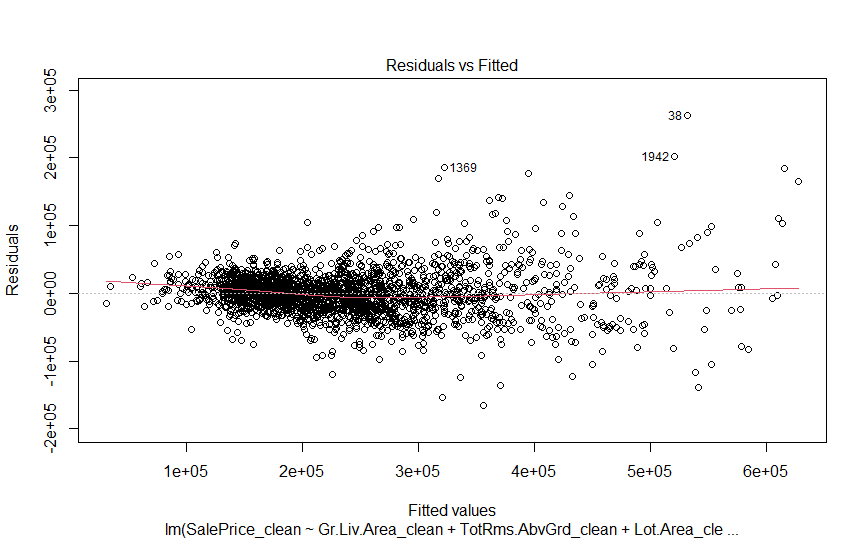
**Figure 9:** Final Model Prediction Accuracy Evaluation

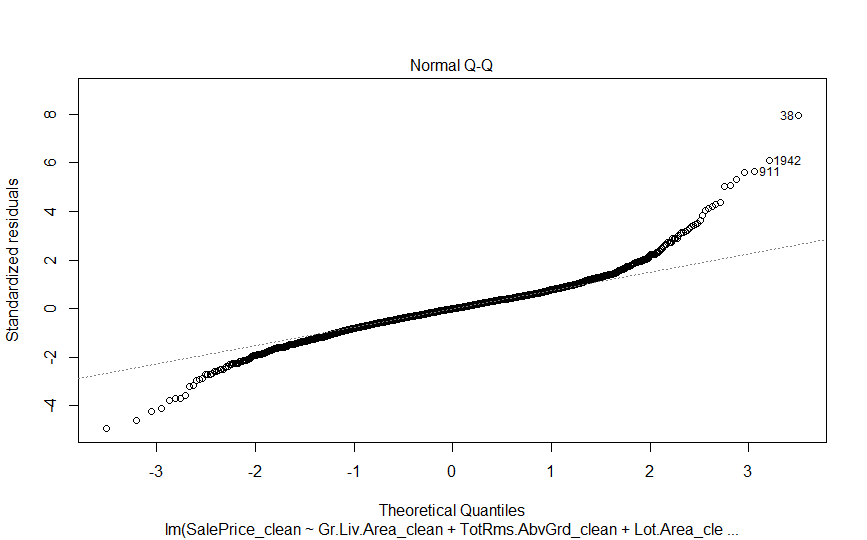


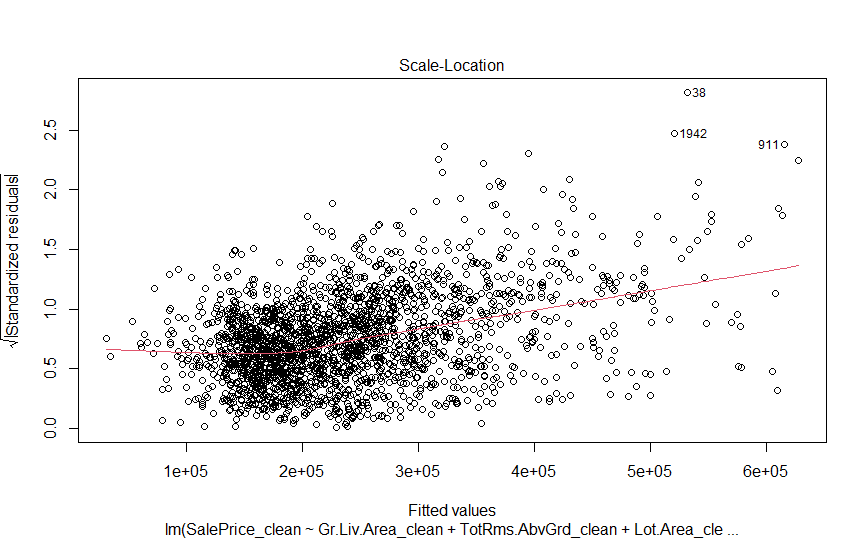
**Figure 10**: Variance Inflation Factor

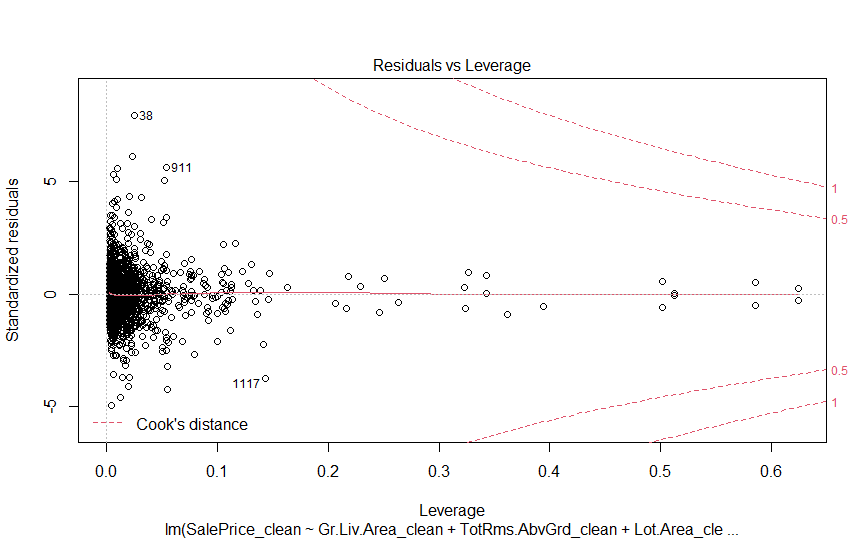


**Figure 11:** Residual Plots





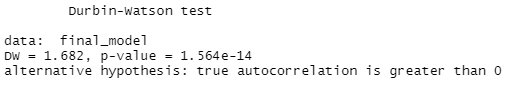




**Figure 12**: Cook’s Distance Output

****

**Figure 13**: Durbin-Watson Test Output



# Summary of Insights

Several insights can be gained from interpretation of both the measures of association and regression model in Section 3.0. Briefly, the measures of association as shown in Figures 1 to 5 show a positive relationship between each independent variable and the sale price, but the key insights are gained from interpreting the linear regression model itself. Figure 8 shows the output from the regression model. The coefficients give a deeper understanding of the significance of each independent variables’ relationship with sale price. As the estimates show, as ground floor living area increases by one square foot, the predicted sale price increases by approximately $93. Comparatively, an increase in lot area by one unit increases the sale price by just $4.32. Bigger houses as you’d expect command a higher price, and both variables show a strong positive relationship, but it is notable that living space is seemingly much more desirable than having a large garden for example, based on this model. This could be trait for the area itself, for example land could be cheaper and building materials more expensive, but more data is needed to determine exactly why this is the case. Nevertheless, both relationships are statistically significant at the 0.001 level. Interestingly, there is a significant negative relationship between number of bedrooms and predicted sale-price at the 0.001 level, with a 1 unit increase in bedroom count equating to a decrease of around $13000 in predicted sale price. Further analysis showed three- and four-bedroom houses commanded the highest overall average price in the area, while six-bedroom houses were some of the lowest. This could be explained by the fact a lower number of 5/6 bedroom houses were sold, either due to scarcity of this trait in the area or difficulty selling them, but more data is needed to determine this. Each exterior quality rating showed a significant negative relationship at the 0.001 level. More data again is needed to determine exactly why this is the case. Houses in the high-quality range (8-10), showed a significant positive relationship with sale price. For example, an increase in quality rating from 8 to 9 showed a predicted difference in sale price of $201,300, significant at the 0.001 level. The year the house was built also appears to have a significant positive relationship at the 0.001 level, with predicted sale price increasing by $805.30 by each yearly increase. This could be due to the quality of building materials improving with time, therefore reflecting in the value of the house. The adjusted r-squared of the model shows that 88.17% of the variability in sale price, is accounted for by the combined independent variables included in the model, which appears to be a good fit for the data. The f-statistic also shows the overall model is statistically significant at less than 0.05, rejecting the null hypothesis. Figure 9 shows the predictive accuracy of the model based on the test data, showing a Root Mean Squared Error (RMSE) of $32,135. This is a reasonably accurate model however could be improved with the collection of more data and inclusion of relevant independent variables. It is also relative to the average sale price in the area, which is $233,234 Figures 10 to 13 show the assumption checks carried out. Figure 10 indicates an issue with multicollinearity in this model, with overall quality , garage quality, garage condition and exterior quality showing very high VIF of between 7 and 11. The first 2 plots shown in figure 11 show homoscedasticity in the model, with a “fanning out” of the data showing a violation of the assumption that residuals should be normally distributed. There appears to be no violation of Cooks’ Distance (plotted in figure 11 with numerical output in figure 12) with none of the observations above 1 and therefore none exerting high influence over the model. The Durbin Watson test is also showing no violations/uncorrelated errors, with an output of 1.682 shown in Figure 13 as non-significant.

# Reflective Commentary

For this logbook, I adopted a similar approach to the first logbook, by working on each task after the material was covered in class, allowing me to stay on top while also giving me the time to fully comprehend each task. The structure itself made sense and was a nice continuation of the work done in logbook 1. It was good to see the work come together and to learn how and why each component of this logbook was structured and ordered. This is important for me, to have a deeper of the underlying theory to troubleshoot and deliver better quality insights. For example, I encountered an error running the VIF assumption test. Understanding the theory behind multicollinearity allowed me to determine there was perfect collinearity between 2 variables, (basement quality and condition) one of which was removed, and I was able to run the code. I encountered no other real issues technically, which I think is testament to some good progression made on the R programming side (Stack Overflow also helps!). The real learning curve came on the interpretation side, specifically interpreting the linear regression model. Remaining concise while also giving some good critical insight was a challenge while learning the outputs and the underpinning statistical theory behind it will hopefully help me in my career. Understanding the difference between the inferential side and predictive side was a large part of this process which I found insightful too, and seeing how those lines between finding something out(inference) and prediction overlap will be crucial in any future business analytics task, whether its seeking to apply a model to a general population or using that model to predict a specific outcome, be it churn, sales forecasting or other. I feel I have a pretty basic understanding of assumption checks at this stage, so I’m looking forward to developing a deeper understanding of these in future assignments, as well as looking at logistic regression which can be applied to categorical outcomes.

# Appendix 1: R Code Used

**Copy and paste below into R to run code(adjust working directory accordingly, N.B; this is a continuation of code used in Logbook 1)**

#Calculate measures of association

##correlation between ground floor living area and price

cor.test(data$Gr.Liv.Area\_clean,data$SalePrice\_clean, use= "complete.obs", method="pearson")

##correlation between lot area area and price

cor.test(data$Lot.Area\_clean,data$SalePrice\_clean, use= "complete.obs", method="pearson")

##correlation between overall quality and price

cor.test(data$Overall.Qual,data$SalePrice\_clean, use= "complete.obs", method="spearman", exact = FALSE)

##correlation between number of bedrooms and price

cor.test(data$Bedroom.AbvGr,data$SalePrice\_clean, use= "complete.obs", method="spearman",exact=FALSE)

##correlation between total rooms and price

cor.test(data$TotRms.AbvGrd\_clean,data$SalePrice\_clean, use= "complete.obs", method="spearman",exact=FALSE)

#read in test data

install.packages("caret")

library(caret)

test<- read\_excel("ames\_test.xlsx")

#preprocessing test data to match train

install.packages("dplyr")

library(dplyr)

test <- rename(test, SalePrice\_clean = SalePrice)

test <- rename(test, Gr.Liv.Area\_clean = Gr.Liv.Area)

test <- rename(test, Lot.Area\_clean = Lot.Area)

test <- rename(test, TotRms.AbvGrd\_clean = TotRms.AbvGrd)

test$SalePrice\_clean[test$SalePrice\_clean>5000000]<- NA

test$Gr.Liv.Area\_clean[test$Gr.Liv.Area\_clean>3000]<-NA

test$TotRms.AbvGrd\_clean[test$TotRms.AbvGrd\_clean>11] <-NA

test$Lot.Area\_clean[test$Lot.Area\_clean>20000] <-NA

#multiple linear regression models

model1 <-lm(SalePrice\_clean ~ Gr.Liv.Area\_clean + TotRms.AbvGrd\_clean +Lot.Area\_clean + Bedroom.AbvGr + as.factor(Overall.Qual), data=data)

model2 <-lm(SalePrice\_clean ~ Gr.Liv.Area\_clean + TotRms.AbvGrd\_clean + Lot.Area\_clean + Bedroom.AbvGr + as.factor(Overall.Qual) + as.factor(Overall.Cond) + as.factor(Exter.Qual), data=data)

final\_model <-lm(SalePrice\_clean ~ Gr.Liv.Area\_clean + TotRms.AbvGrd\_clean + Lot.Area\_clean + Bedroom.AbvGr + as.factor(Overall.Qual) + as.factor(Overall.Cond) + as.factor(Exter.Qual) + as.factor(Bsmt.Cond)+

as.factor(Heating.QC) + as.factor(Garage.Cond) + as.factor(Garage.Qual) + Year.Built, data=data)

#review models

summary(model1)

summary(model2)

summary(final\_model)

#prediction using the model and test data

multiple\_price\_prediction <- predict(final\_model, newdata = test)

#evaluate prediction accuracy(diff between actual sale price and predicted sale price)

postResample(pred = multiple\_price\_prediction, test$SalePrice\_clean)

##ASSUMPTION CHECKS

#variance inflation factor

install.packages("car")

library(car)

vif(final\_model)

#check residual plots

plot(final\_model)

#cooks distance

c<-cooks.distance(final\_model)

summary(c)

#durbin-watson test

install.packages("lmtest")

library(lmtest)

dwtest(final\_model)