**House Prices - Advanced Regression Techniques**

**Introduction**

In this study, prompted by our client, Century 21 Ames, we were tasked with first developing a model to predict the Sale Price of houses with three Neighborhoods based on the Square Footage. Next we were tasked with trying to develop a Simple Linear Regression model, and two Multiple Regression Models using any of the variables.

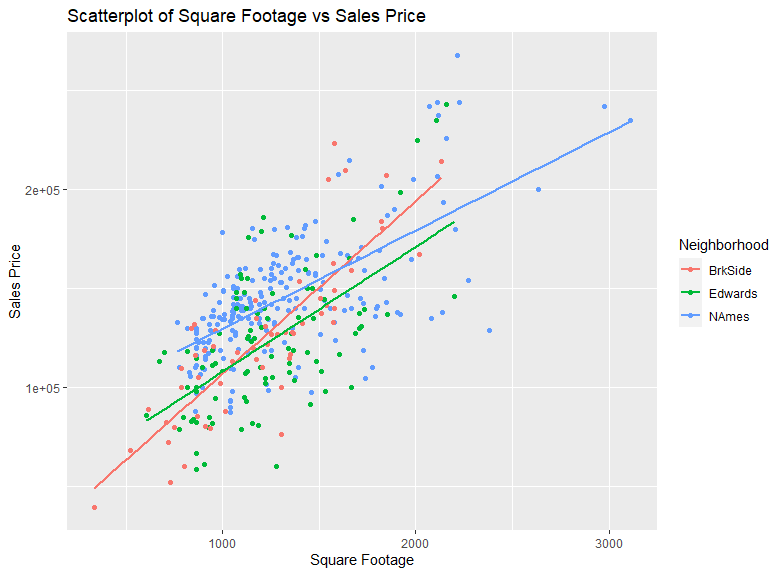
**Data Description**

The data that is being analyzed was obtained from the website Kaggle, they obtained it from this data from Century 21 Ames. This data contains 79 different explanatory variables to try and predict 1 dependent variable, Sale Price. There are 1460 different homes that are being observed in this data set. In our analysis of the data some of the key variables we looked at to explain the Sale price were the Square footage, Neighborhood, Full Bath, and Overall Quality.

**Analysis Question 1:**

**Restatement of Problem**

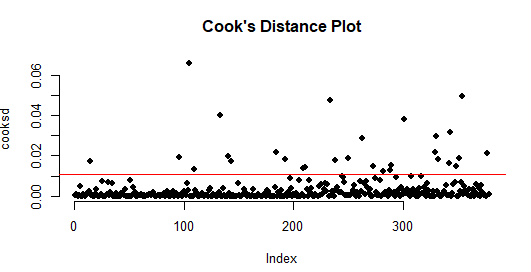
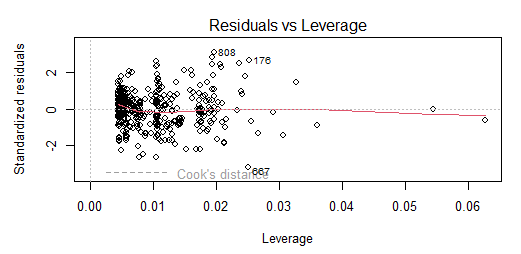
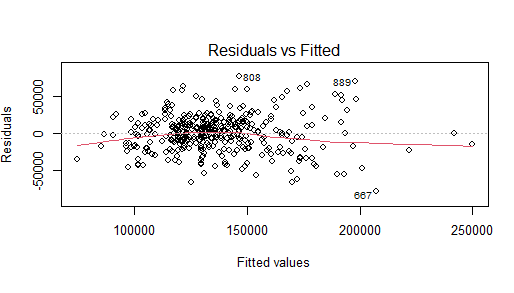
The first problem we were tasked to look at and estimate a model between Sale Price and Square footage with respect to three Neighborhoods. The three Neighborhoods in question were NAmes, Edwards and BrkSide.



We built a model to estimate the Sale Price for each of the three Neighborhoods. Each off the neighborhoods has their own slopes:

BrkSide: SalePrice = 19971.514 + 87.163\*GrLIvArea

Edwards: SalePrice = 88353.1 + 29.751\*GrLIvArea

 NAmes: SalePrice = 74676.4 + 54.316\*GrLIvArea

|  |  |
| --- | --- |
|  |  |

**Assumptions**

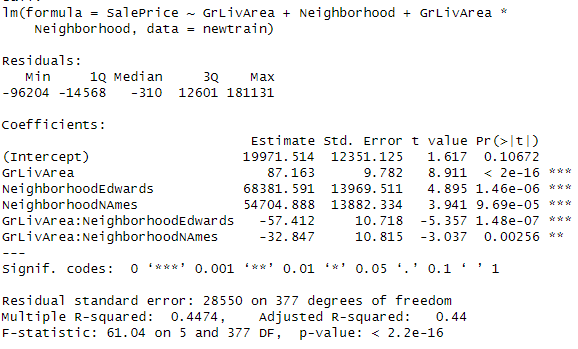
The plots bove are the Residual plot, Cook’s D plot, and Leverage plot that were created after the elimination of a few outliers. The outliers were taken out of the data set because they were high priced homes with low square footage for the neighborhoods. These houses were ID # 524, 1299, 643, 725 and 1424. The residual plot shows that there is a fairly random distribution of points along with an equal amount positive and negative. For the Cook’s D plot it shows that there are no outliers within the data set and that there aren't any influential points. The Leverage plot shows that there are no points with significant leverage.

**Comparing Competing Models**

Adj R2 = 0.44

**Internal CV Press**

* BrkSide = 0.1871246213
* Edwards = 0.2393758557
* NAmes = 0.2363974899

**Parameters** 

**Interpretation**

Based on the BrkSide: SalePrice = 19971.514 + 87.163\*GrLIvArea equation for every 100 square feet increase the Sale Price for BrkSide Neighborhood is estimated to increase $8716.3. Based on the Edwards: SalePrice = 88353.1 + 29.751\*GrLIvArea equation for every 100 square feet increase the estimated Sale Price will increase by $2975.1 in the Sale Price. Based on the NAmes: SalePrice = 74676.4 + 54.316\*GrLIvAre equation for every 100 square feet increase the estimated Sale Price will increase by $5431.6.

**Confidence Intervals**

For the BrkSide Neighborhood we are 95% confident that for every 100 square feet increase the Sale Price is estimated to increase between $6792.85 and $10639.66. For the Edwards Neighborhood we are 95% confident that for every 100 square feet increase the Sale Price is estimated to increase between $-1055.76 and $7005.82. For the NAmes Neighborhood we are 95% confident that for every 100 square feet increase the Sale Price is estimated to increase between $1381.58 and $9481.59.

**Conclusion**

In conclusion, based on the Models for every 100 Square feet increase the BrkSide Neighborhood Sale Price is estimated to increase by $8716.3, the Edwards Neighborhood Sale Price is estimated to increase by $2975.1, and the NAmes Neighborhood Sale Price is estimated to increase by $5431.6. Each neighborhood was determined to have their own independent slopes as depicted in the graph above, and thus their own equation.

**R Shiny: Price v. Living Area Chart**

<https://nolandulude.shinyapps.io/StatsProject/>

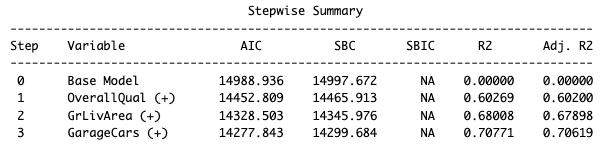
**Analysis Question 2:**

**Restatement of Problem**

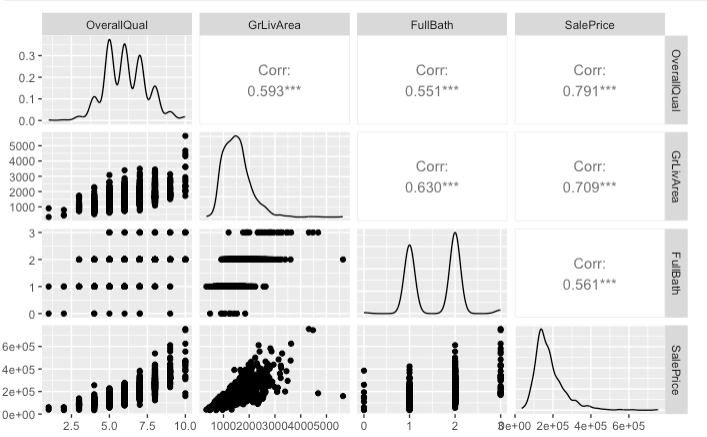
In this problem, we were asked to build the most predictive models for sales prices of homes in all of Ames, Iowa: one being a simple linear regression model in which we pick the explanatory variable, a multiple linear regression model (SalePrice~GrLivArea + FullBath), and another multiple linear regression model where we select the explanatory variables. We were to generate an adjusted R2, CV Press and Kaggle Score for each of these models and clearly describe which model is the best in predicting future sale prices of homes in Ames, Iowa.

**Candidate Models:**

With the same data used to solve Analysis Problem 1, we sought to identify the best variables to predict sale price of Ames homes. After dropping variables not suited for Linear Regression Models, we put together our models.

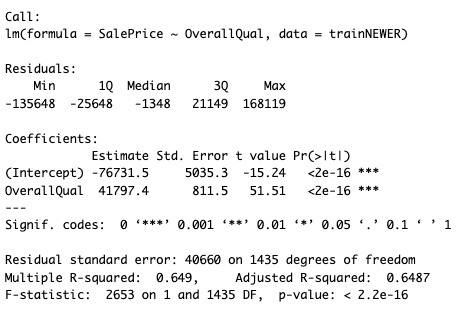


With the stepwise model selection, we found the two best fit variables to be OverallQual and GrLivArea. These two, along with FullBath, will be used in the Linear regression model and multiple linear regression models we intend to conduct. Before running those models, we first want to check assumptions and assess the normality of the data.



There is visual evidence of a relationship between overall qual and GrLivArea, OverallQual and FullBath, OverallQual and SalePrice, GrLivArea and FullBath, FullBath and SalePrice. There appears to be a linear relationship between GrLivArea and SalePrice with outliers (those we struck out in the first problem).

**SLR (SalePrice ~ OverallQual)**

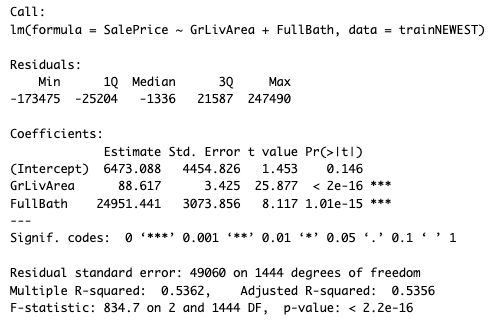


With the model **SalePrice = -76731.5 + 41797.4OverallQual**, there is a significant relationship between the overall quality of the home and the sale price of the homes in Ames, as evidenced by the p-value of the OverallQual variable (<0.0000000000000002) and the overall p-value of <0.00000000000000022.

Checking assumptions for this model, we found that the data was adequately scattered in the residual plot so there was little evidence of variance , the data was reasonably fitted to the line in the qq-plot, suggesting normal distribution, and through the process of deleting high leverage outliers, the Cook’s d plot showed a good distribution for us to move forward with the model.

|  |  |
| --- | --- |
|  |  |

**MLR 1: Model Provided by Century21 Ames (SalePrice ~ GrLivArea + FullBath)**

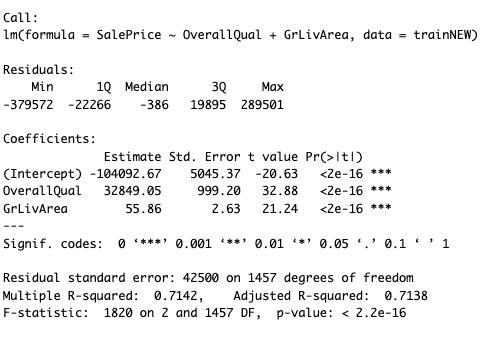


With the model **SalePrice = 6473.1 + 88.62GrLivArea + 24951.44FullBath**, there is a statistically significant relationship between the gross living area and number of full baths on the sale price of the homes in Ames, as evidenced by their respective small p-values (<0.0000000000000002, .00000000000000101) and the overall p-value of <0.00000000000000022.

Checking assumptions for this model, we found that the data was adequately scattered in the residual plot so there was little evidence of variance , the data was reasonably fitted to the line in the qq-plot, suggesting normal distribution, and through the process of deleting high leverage outliers, the Cook’s d plot showed a good distribution for us to move forward with the model

|  |  |
| --- | --- |
|  |  |

**MLR 2 (SalePrice ~ OverallQual + GrLivArea)**



With the model **SalePrice = -104092.67 + 32849.1OverallQual + 55.86GrLivArea**, there is a statistically significant relationship between the overall quality of the homes and the gross living area of the homes in predicting the sale price of the homes in Ames, as evidenced by their respective small p-values (<0.0000000000000002, <0.0000000000000002) and the overall p-value of <0.00000000000000022.

Checking assumptions for this model, we found that the data was adequately scattered in the residual plot so there was little evidence of variance, the data was reasonably fitted to the line in the qq-plot, suggesting normal distribution, and through the process of deleting high leverage outliers, the Cook’s d plot showed a good distribution for us to move forward with the model.

|  |  |
| --- | --- |
|  |  |

**Comparing Competing Models**

Comparing the three models using test data, our multiple linear regression model is best fit, evidenced by the higher r-squared, adjusted r-squared, and small mean CV press value. You will note that the MLR we created has a slightly higher mean CV press in comparison to Century21 Ames’, suggesting a slightly better predictive performance, however, with much higher correlation of variables in our MLR, we decided that the latter was best fit in predicting the sale prices using Test data.

| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| --- | --- | --- | --- |
| Simple Linear Regression | 0.6487 | 0.002726416 | 0.48351 |
| MLR\_C21\_2 | 0.5356 | 0.0009129459 | 0.61812 |
| Multiple Linear Regression | 0.7138 | 0.001248771 | 0.28542 |

**Conclusion**

Through simple and multiple linear regression analysis, we found that the explanatory variables that best predicted the sale price of homes in Ames, IA were GrLivArea and OverallQual. These variables were used in the multiple linear regression model and proved to explain the variation in sale price in homes. We suggest Century21 Ames utilize this model to predict sale prices while conducting business in the Ames, IA area.

**Appendix**

1. **R Code**

**Analysis Problem # 1**

library(tidyverse)

library(car)

train <- read.csv(choose.files())

head(train)

#Question 1

#Selecting for NAmes, Edwards, and BrkSide

match1 <- grepl("NAmes", train$Neighborhood)

match2 <- grepl("Edwards", train$Neighborhood)

match3 <- grepl("BrkSide", train$Neighborhood)

new1 <- train[match1, ]

new2 <- train[match2, ]

new3 <- train[match3, ]

#New data set

newtrain <- rbind(new1, new2, new3)

#Graphing each Neighborhood alone and together

newtrain %>% filter(Neighborhood == "NAmes") %>% ggplot(aes(x = GrLivArea, y = SalePrice, colour = Neighborhood)) + geom\_point() + labs(x = "Square Footage", y = "Sales Price") + ggtitle("Scatterplot of Square Footage vs Sales Price for the NAmes Neighborhood")

newtrain %>% filter(Neighborhood == "Edwards") %>% ggplot(aes(x = GrLivArea, y = SalePrice, colour = Neighborhood)) + geom\_point() + labs(x = "Square Footage", y = "Sales Price") + ggtitle("Scatterplot of Square Footage vs Sales Price for the Edwards Neighborhood")

newtrain %>% filter(Neighborhood == "BrkSide") %>% ggplot(aes(x = GrLivArea, y = SalePrice, colour = Neighborhood)) + geom\_point() + labs(x = "Square Footage", y = "Sales Price") + ggtitle("Scatterplot of Square Footage vs Sales Price for the BrkSide Neighborhood")

newtrain %>% ggplot(aes(x = GrLivArea, y = SalePrice, colour = Neighborhood))+geom\_point() + geom\_smooth(method = "lm", se = FALSE) + labs(x = "Square Footage", y = "Sales Price") + ggtitle("Scatterplot of Square Footage vs Sales Price")

#After looking at the residual plot and Cooks plot there appears to be a couple outliers.

which(newtrain$Id == 1299)

id1 <- newtrain[313, c("GrLivArea", "SalePrice", "Neighborhood")]

which(newtrain$Id == 524)

id2 <- newtrain[258, c("GrLivArea", "SalePrice", "Neighborhood")]

summary(newtrain$SalePrice)

summary(newtrain$GrLivArea)

fit = lm(SalePrice~GrLivArea + Neighborhood, data = newtrain)

fit\_summary <- summary(fit)

#Visualize LRM

par(mfrow = c(2, 2))

plot(fit)

#Calculate Cook's distances

cooksd <- cooks.distance(fit)

#Plot Cook's distances

plot(cooksd, pch = 19, frame = FALSE, main = "Cook's Distance Plot")

abline(h = 4/length(cooksd), col = "red") # Add a horizontal line at Cook's distance = 4/n

#Set a threshold for identifying outliers.

threshold <- 16/length(cooksd)

#Identify outliers based on Cook's distance exceeding the threshold

outliers <- which(cooksd > threshold)

#Print the indices of outliers

print(outliers)

print(train[c(524, 1299, 643, 725, 1424), ])

#Outliers were removed

newer\_train<- newtrain[-c(313, 258, 99, 275, 322), ]

#After removal of Outliers

cooksd <- cooks.distance(fit)

plot(cooksd, pch = 19, frame = FALSE, main = "Cook's Distance Plot")

abline(h = 4/length(cooksd), col = "red") # Add a horizontal line at Cook's distance = 4/n

#Graphing each Neighborhood alone and together

newer\_train %>% filter(Neighborhood == "NAmes") %>% ggplot(aes(x = GrLivArea, y = SalePrice, colour = Neighborhood)) + geom\_point() + labs(x = "Square Footage", y = "Sales Price") + ggtitle("Scatterplot of Square Footage vs Sales Price for the NAmes Neighborhood")

newer\_train %>% filter(Neighborhood == "Edwards") %>% ggplot(aes(x = GrLivArea, y = SalePrice, colour = Neighborhood)) + geom\_point() + labs(x = "Square Footage", y = "Sales Price") + ggtitle("Scatterplot of Square Footage vs Sales Price for the Edwards Neighborhood")

newer\_train %>% filter(Neighborhood == "BrkSide") %>% ggplot(aes(x = GrLivArea, y = SalePrice, colour = Neighborhood)) + geom\_point() + labs(x = "Square Footage", y = "Sales Price") + ggtitle("Scatterplot of Square Footage vs Sales Price for the BrkSide Neighborhood")

newer\_train %>% ggplot(aes(x = GrLivArea, y = SalePrice, colour = Neighborhood))+geom\_point() + geom\_smooth(method = "lm", se = FALSE) + labs(x = "Square Footage", y = "Sales Price") + ggtitle("Scatterplot of Square Footage vs Sales Price")

#Here is the Model and with the ADJR^2, and internal CV Press, along with confidence intervals.

fit1 = lm(SalePrice~GrLivArea + Neighborhood+GrLivArea\*Neighborhood, data = newtrain)

fit\_summary1 <- summary(fit1)

adj\_r\_squared <- fit\_summary1$adj.r.squared

internal\_cv\_press <- fit\_summary1$cov.unscaled

conf\_intervals <- confint(fit1)

par(mfrow = c(2, 2))

plot(fit1)

#Here is the code of the R Shiny app

library(shiny)

library(ggplot2)

train <- read.csv(choose.files())

#Selecting for NAmes, Edwards, and BrkSide

match1 <- grepl("NAmes", train$Neighborhood)

match2 <- grepl("Edwards", train$Neighborhood)

match3 <- grepl("BrkSide", train$Neighborhood)

new1 <- train[match1, ]

new2 <- train[match2, ]

new3 <- train[match3, ]

#New data set

newtrain <- rbind(new1, new2, new3)

ui <- fluidPage(

titlePanel("House Price vs. Square Footage"),

sidebarLayout(

sidebarPanel(

selectInput("neighborhood",

"Choose a Neighborhood:",

choices = c("NAmes", "Edwards", "BrkSide"),

selected = "NAmes")

),

mainPanel(

plotOutput("scatterplot"),

plotOutput("combined\_plot")

)

)

)

# Server Logic

server <- function(input, output) {

output$scatterplot <- renderPlot({

neighborhood\_data <- subset(newtrain, Neighborhood == input$neighborhood)

ggplot(neighborhood\_data, aes(x = GrLivArea, y = SalePrice)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE) + # Add linear trend line

labs(title = paste("House Price vs. Square Footage in", input$neighborhood),

x = "Square Footage",

y = "Price")

})

output$combined\_plot <- renderPlot({

ggplot(newtrain, aes(x = GrLivArea, y = SalePrice, color = Neighborhood)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE) + # Add linear trend line

labs(title = "House Price vs. Square Footage (Combined)",

x = "Square Footage",

y = "Price",

color = "Neighborhood")

})

}

# Run the App

shinyApp(ui = ui, server = server)

Code for Analysis Problem #2:

library(tidyverse)

library(ggplot2)

library(scales)

library(pwr)

library(agricolae)

install.packages("huxtable")

library(huxtable)

install.packages("lawstat")

library(lawstat)

library(lsmeans)

library(dplyr)

library(WDI)

library(investr)

library(multcomp)

library(pairwiseCI)

install.packages("DescTools")

library(DescTools)

install.packages("GGally")

library(GGally)

install.packages("olsrr")

library(olsrr)

library(tidyverse)

library(car)

```

In order to create our simple linear regression, we will use an automatic variable selection technique.

Select explanatory variable for determining sales prices of homes in Ames

```{r}

#Review data

head(train)

summary(train)

#Create tentative linear regression model to plug into Backward, Forward, and Stepwise Selection Models

# Load the necessary libraries

library(MASS) # For stepAIC function

# Start with an empty model

best\_model <- lm(SalePrice ~ 1, data = train)

# Number of predictors in the dataset

num\_predictors <- ncol(train) - 1 # Excluding the target variable 'SalePrice'

# Initialize variables to store the best predictor and its associated AIC

best\_predictor <- NULL

best\_AIC <- Inf

# Forward selection loop

for (predictor in names(train)[-which(names(train) == "SalePrice")]) {

# Construct formula for current predictor

formula\_str <- paste("SalePrice ~", predictor)

# Fit a model with the current predictor

model <- lm(formula\_str, data = train)

# Compute AIC for the current model

model\_AIC <- AIC(model)

# Update the best predictor if current AIC is lower

if (model\_AIC < best\_AIC) {

best\_AIC <- model\_AIC

best\_predictor <- predictor

}

}

# Display the best predictor found

print(best\_predictor)

```

Using Forward Step Selection, we found that the best predictor variable in the 79 variable dataset is PoolQC. We will use this variable in the simple linear regression model.

#Selection for top variables

```{r}

#TRANSFORM CAT VARS TO FACTORS AND TRY STEPWISE

# Identify variables with categorical data types

categorical\_vars <- sapply(train, function(x) is.factor(x) || is.character(x))

# List variables with categorical data types

cat\_vars\_names <- names(categorical\_vars)[categorical\_vars]

cat\_vars\_names

# Convert variables with categorical data types to factors

train[, cat\_vars\_names] <- lapply(train[, cat\_vars\_names], as.factor)

fit <- lm(SalePrice ~ ., data = train)

result <- ols\_step\_both\_p(fit, penter = 0.01, prem = 0.05, details = FALSE)

# Filter the output of str() to only display factor variables

str(train[, sapply(train, is.factor)])

# Check levels of each factor variable

lapply(train[, cat\_vars\_names], function(x) levels(x))

```

#In the process of running the forward step, we found that the PoolQC, Fence, MiscFeature, Alley, and Utilities variables had a number of NA values that were not suitable to be run in a linear regression model. We dropped those variables from the dataset and ran stepwise step again.

```{r}

trainNEW = subset(train, select=-c(PoolQC, Fence, MiscFeature, Alley, Utilities))

fit <- lm(SalePrice ~ ., data = trainNEW)

```

#Now we will try stepwise selection.

```{r}

# Stepwise

# Perform stepwise selection with different p-values for entering and exiting variables

result <- ols\_step\_both\_p(fit, penter = 0.01, prem = 0.05, details = FALSE)

print(result)

```

With the stepwise model selection, we found the two best fit variables to be OverallQual and GrLivArea. These two, along with FullBath, will be used in the Linear regression model and multiple linear regression models we intend to conduct. Before running those models, we first want to check assumptions and assess the normality of the data.

```{r}

# Create scatterplot matrix

ggpairs(trainNEW[, c("OverallQual", "GrLivArea", "FullBath", "SalePrice")])

```

#There is visual evidence of a relationship between overall qual and GrLivArea, OverallQual and FullBath, OverallQual and SalePrice, GrLivArea and FullBath, FullBath and SalePrice. There appears to be a linear relationship between GrLivArea and SalePrice with outliers (those we struck out in the first problem).

#We will assess outliers in the models we create below.

```{r}

#Simple Linear Regression

SLR = lm(SalePrice ~ OverallQual, data = trainNEW)

summary(SLR)

#Visualize SLR

plot(SLR)

#Calculate Cook's distances

cooksd <- cooks.distance(SLR)

#Plot Cook's distances

plot(cooksd, pch = 19, frame = FALSE, main = "Cook's Distance Plot")

abline(h = 4/length(cooksd), col = "red") # Add a horizontal line at Cook's distance = 4/n

```

#The qqplot and cook's d plots show evidence of outliers. In an effort to give Century21 Ames the best results informed by a best fit model, we will remove these outliers.

```{r}

#Set a threshold for identifying outliers.

threshold <- 16/length(cooksd)

#Identify outliers based on Cook's distance exceeding the threshold

outliers <- which(cooksd > threshold)

#Print the indices of outliers

print(outliers)

```

#There are 20 outliers in the dataset with leverage that impact the our model's ability to predict saleprice. The outliers stand out as anomalies that should be selected and deleted.

```{r}

print(trainNEW[c(179,186,350,376,441,458,474,497,524,528,534,592,692,770,799,804,899,1047,1170,1183,1244,1299,1374), ])

trainNEWER<- trainNEW[-c(179,186,350,376,441,458,474,497,524,528,534,592,692,770,799,804,899,1047,1170,1183,1244,1299,1374), ]

#Simple linear regression

SLR\_2 = lm(SalePrice ~ OverallQual, data = trainNEWER)

summary(SLR\_2)

#Calculate Cook's distances

cooksd <- cooks.distance(SLR\_2)

#Plot Cook's distances

plot(cooksd, pch = 19, frame = FALSE, main = "Cook's Distance Plot")

abline(h = 4/length(cooksd), col = "red") # Add a horizontal line at Cook's distance = 4/n

```

#The residuals are better fit on the qqplot and the cook's d plot shows differences in residuals on a much smaller scale. Outliers have been sufficiently eliminated. Checking assumptions for this model, we found that the data was adequately scattered in the residual plot so there was little evidence of variance , the data was reasonably fitted to the line in the qq-plot, suggesting normal distribution, and through the process of deleting high leverage outliers, the Cook’s d plot showed a good distribution for us to move forward with the model.

#Now we will visualize and assess the simple linear regression.

```{r}

#Summarize Simple linear model

summary(SLR\_2)

#Visualize LRM

plot(SLR\_2)

#Find Adjusted R-Squared Value

SLR\_2\_fit\_summary <- summary(SLR\_2)

SLR\_2\_adj\_r\_squared <- SLR\_2\_fit\_summary$adj.r.squared

print(SLR\_2\_adj\_r\_squared)

#Find CV Press

SLR\_2\_internal\_cv\_press <- SLR\_2\_fit\_summary$cov.unscaled

mean(SLR\_2\_internal\_cv\_press)

print(mean(SLR\_2\_internal\_cv\_press))

```

#Now we will run a Multiple Linear Regression with the variables you all provided wherein GrLivArea + FullBath predict SalePrice.

```{r}

#Multiple Linear Regression

MLR\_C21 = lm(SalePrice ~ GrLivArea + FullBath, data = trainNEW)

summary(MLR\_C21)

#Visualize multiple linear regression

plot(MLR\_C21)

#Calculate Cook's distances

cooksd <- cooks.distance(MLR\_C21)

#Plot Cook's distances

plot(cooksd, pch = 19, frame = FALSE, main = "Cook's Distance Plot")

abline(h = 4/length(cooksd), col = "red") # Add a horizontal line at Cook's distance = 4/n

```

#There are clearly some outliers in this plot.

```{r}

#Set a threshold for identifying outliers.

threshold <-32/length(cooksd)

#Identify outliers based on Cook's distance exceeding the threshold

outliers <- which(cooksd > threshold)

#Print the indices of outliers

print(outliers)

```

```{r}

print(trainNEW[c(54,441,524,636,665,692,770,804,899,1047,1170,1183,1299), ])

trainNEWEST<- trainNEW[-c(54,441,524,636,665,692,770,804,899,1047,1170,1183,1299), ]

#Simple linear regression

MLR\_C21\_2 = lm(SalePrice ~ GrLivArea + FullBath, data = trainNEWEST)

summary(MLR\_C21\_2)

plot(MLR\_C21\_2)

#Calculate Cook's distances

cooksd <- cooks.distance(MLR\_C21\_2)

#Plot Cook's distances

plot(cooksd, pch = 19, frame = FALSE, main = "Cook's Distance Plot")

abline(h = 4/length(cooksd), col = "red") # Add a horizontal line at Cook's distance = 4/n

```

#The residuals are better fit on the qqplot and the cook's d plot shows differences in residuals on a much smaller scale. We also got rid of the outlier with high leverage. Outliers have been sufficiently eliminated. Checking assumptions for this model, we found that the data was adequately scattered in the residual plot so there was little evidence of variance , the data was reasonably fitted to the line in the qq-plot, suggesting normal distribution, and through the process of deleting high leverage outliers, the Cook’s d plot showed a good distribution for us to move forward with the model.

#Now we will visualize and assess the simple linear regression.

```{r}

#Summarize Simple linear model

summary(MLR\_C21\_2)

#Visualize LRM

plot(MLR\_C21\_2)

#Find Adjusted R-Squared Value

MLR\_C21\_2\_fit\_summary <- summary(MLR\_C21\_2)

MLR\_C21\_2\_adj\_r\_squared <- MLR\_C21\_2\_fit\_summary$adj.r.squared

print(MLR\_C21\_2\_adj\_r\_squared)

#Find CV Press

MLR\_C21\_2\_internal\_cv\_press <- MLR\_C21\_2\_fit\_summary$cov.unscaled

mean(MLR\_C21\_2\_internal\_cv\_press)

print(mean(MLR\_C21\_2\_internal\_cv\_press))

```

#Judging from the parameter estimate table, there is overwhelming evidence to suggest that the combination of GrLivArea and FullBath are statistically significant in predicting SalePrice.

#Our team was able to develop a model that similarly predicts sale price.

```{r}

MLR = lm(SalePrice ~ OverallQual + GrLivArea, data = trainNEW)

summary(MLR)

#Visualize multiple linear regression

plot(MLR)

#Calculate Cook's distances

cooksd <- cooks.distance(MLR)

#Plot Cook's distances

plot(cooksd, pch = 19, frame = FALSE, main = "Cook's Distance Plot")

abline(h = 4/length(cooksd), col = "red") # Add a horizontal line at Cook's distance = 4/n

```

#There are clearly some outliers in this plot. Let's identify and remove them if necessary.

```{r}

#Set a threshold for identifying outliers.

threshold <-32/length(cooksd)

#Identify outliers based on Cook's distance exceeding the threshold

outliers <- which(cooksd > threshold)

#Print the indices of outliers

print(outliers)

```

#The outliers were homes with unusual living area sizes which stood out from the homes in Ames, IA. Remove the outliers and reassess model fit.

```{r}

print(trainNEW[c(179,441,524,692,770,804,899,1047,1170,1183,1299), ])

trainNEW\_MLR<- trainNEW[-c(179,441,524,692,770,804,899,1047,1170,1183,1299), ]

#Simple linear regression

MLR\_2 = lm(SalePrice ~ OverallQual + GrLivArea, data = trainNEW\_MLR)

summary(MLR\_2)

plot(MLR\_2)

#Calculate Cook's distances

cooksd <- cooks.distance(MLR\_2)

#Plot Cook's distances

plot(cooksd, pch = 19, frame = FALSE, main = "Cook's Distance Plot")

abline(h = 4/length(cooksd), col = "red") # Add a horizontal line at Cook's distance = 4/n

```

#The residuals are better fit on the qqplot and the cook's d plot shows differences in residuals on a much smaller scale. We also got rid of an outlier with relatively high leverage. Outliers have been sufficiently eliminated. Checking assumptions for this model, we found that the data was adequately scattered in the residual plot so there was little evidence of variance , the data was reasonably fitted to the line in the qq-plot, suggesting normal distribution, and through the process of deleting high leverage outliers, the Cook’s d plot showed a good distribution for us to move forward with the model.

#Now we will visualize and assess the multiple linear regression.

```{r}

#Summarize Simple linear model

summary(MLR\_2)

#Visualize LRM

plot(MLR\_2)

#Find Adjusted R-Squared Value

MLR\_2\_fit\_summary <- summary(MLR\_2)

MLR\_2\_adj\_r\_squared <- MLR\_2\_fit\_summary$adj.r.squared

print(MLR\_2\_adj\_r\_squared)

#Find CV Press

MLR\_2\_internal\_cv\_press <- MLR\_2\_fit\_summary$cov.unscaled

mean(MLR\_2\_internal\_cv\_press)

print(mean(MLR\_2\_internal\_cv\_press))

```

#Judging from the parameter estimate table, there is overwhelming evidence to suggest that the combination of OverallQual and GrLivArea are statistically significant in predicting SalePrice.

#We decided to test the models used in this exercise using Kaggle's test data.

```{r}

#Read in data

test<- read.csv(choose.files())

#Check data

head(test)

#Test Simple Linear Regression

predictions\_SLR <- predict(SLR\_2, newdata = test)

head(predictions\_SLR)

test$SalePrice\_Predicted <- predictions\_SLR

SLR\_test <- test[c("Id", "SalePrice\_Predicted")]

write.csv(SLR\_test, "SLR\_predictions.csv", row.names = FALSE)

#Test Multiple Linear Regression provided by Century21 Ames

predictions\_MLR\_C21\_2 <- predict(MLR\_C21\_2, newdata = test)

head(predictions\_MLR\_C21\_2)

test$SalePrice\_Predicted2 <- predictions\_MLR\_C21\_2

MLR\_test <- test[c("Id", "SalePrice\_Predicted2")]

write.csv(MLR\_test, "MLR\_C21\_2\_predictions.csv", row.names = FALSE)

#Test Multiple Linear Regression

predictions\_MLR\_2 <- predict(MLR\_2, newdata = test)

head(predictions\_MLR\_2)

test$SalePrice\_Predicted3 <- predictions\_MLR\_2

MLR\_2\_test <- test[c("Id", "SalePrice\_Predicted3")]

write.csv(MLR\_2\_test, "MLR\_2\_predictions.csv", row.names = FALSE)

```

#Comparing the three models using test data, our multiple linear regression model is best fit, evidenced by the higher r-squared, adjusted r-squared, and small mean CV press value. You will note that the MLR we created has a slightly higher mean CV press in comparison to Century21 Ames’, suggesting a slightly better predictive performance, however, with much higher correlation of variables in our MLR, we decided that the latter was best fit in predicting the sale prices using Test data.

1. **Further Information on the Researchers**

For further information on the researchers, we invite you to visit our github pages which document our bios and previous work in the field of data science.

Nolan Dulude: [NolanDulude.github.io](http://nolandulude.github.io)

Kenya Roy: [KenyaRoy.github.io](http://kenyaroy.github.io)