SIM - Project

Ignacio Lloret and Iñigo Pikabea

2023-10-09

## Data Reading

df <- read.csv("C:/MDS/SIM/Project/data/train.csv")  
# df <- read.csv("C:/Users/inigo/Documents/UPC/SIM/SIM-project/data/train.csv")

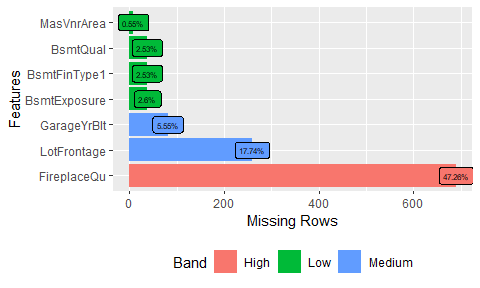
## Keep columns

numeric\_columns <- c("LotFrontage", "LotArea", "YearBuilt", "YearRemodAdd", "MasVnrArea", "BsmtFinSF1", "BsmtFinSF2", "BsmtUnfSF", "TotalBsmtSF", "X1stFlrSF", "X2ndFlrSF", "LowQualFinSF", "GrLivArea", "BsmtFullBath", "BsmtHalfBath", "FullBath", "HalfBath", "BedroomAbvGr", "KitchenAbvGr", "TotRmsAbvGrd", "Fireplaces", "GarageYrBlt", "GarageCars", "GarageArea", "WoodDeckSF", "OpenPorchSF", "EnclosedPorch", "X3SsnPorch", "ScreenPorch", "PoolArea", "MiscVal", "MoSold", "YrSold", "OverallCond","OverallQual","SalePrice")  
  
cat\_keep <- c("FireplaceQu","KitchenQual","BsmtFinType1","BsmtExposure","BsmtQual","Foundation","Neighborhood","LotShape","MSSubClass","Exterior1st","Exterior2nd", "SaleCondition")  
  
df1 <- df %>% select(all\_of(numeric\_columns), all\_of(cat\_keep))

# Exploratory Data Analysis

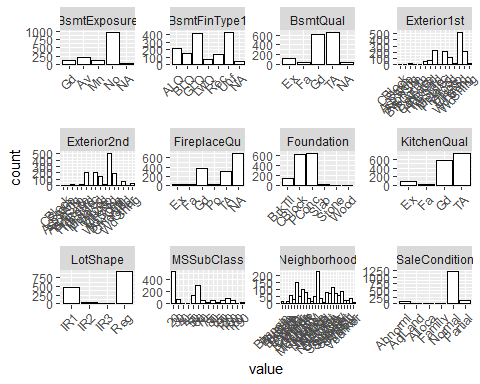
## Missing data

plot\_missing(df1, missing\_only = TRUE, group = list("Low" = 0.05, "Medium"=0.25, "High"=0.5, "Very High" =1), geom\_label\_args = list("size" = 2))



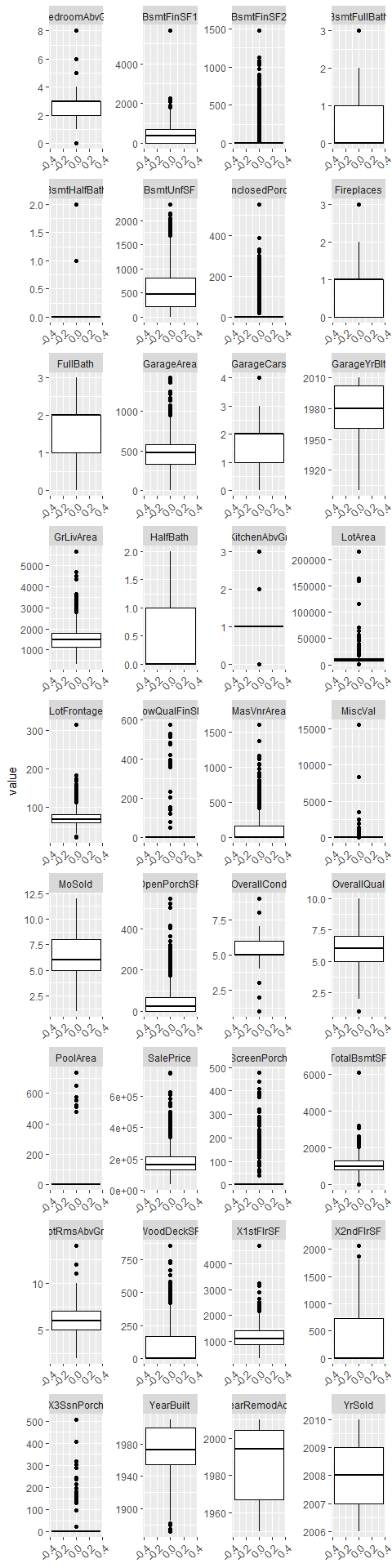
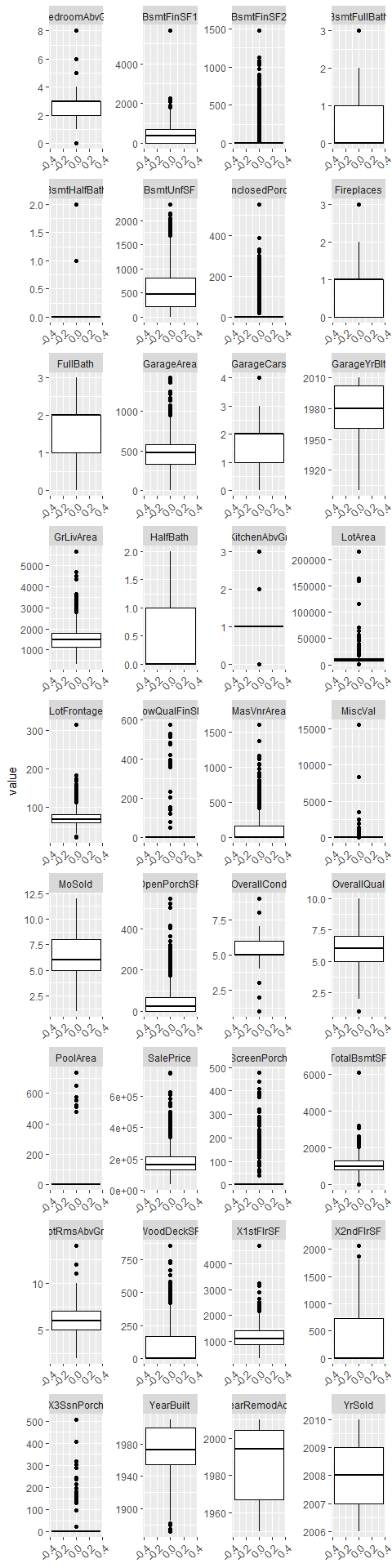
## Categorical values

df1[cat\_keep] <- lapply(df1[cat\_keep], as.factor) Create Factors  
df1[numeric\_columns] <- lapply(df1[numeric\_columns], as.numeric)   
  
  
p1 <- df1 %>%   
 select(all\_of(cat\_keep)) %>%  
 pivot\_longer(cols=everything()) %>%  
 ggplot(data=.) +  
 geom\_bar(aes(x=value), col="black", fill="white") +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 facet\_wrap(~name, scales="free", ncol=4)  
p1



## Numerical Data

p2 <- df1 %>%   
 select(all\_of(numeric\_columns)) %>%  
 pivot\_longer(cols=everything()) %>%  
 ggplot(data=.) +  
 geom\_boxplot(aes(y=value), col="black", fill="white") +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 facet\_wrap(~name, scales="free", ncol=4)  
p2

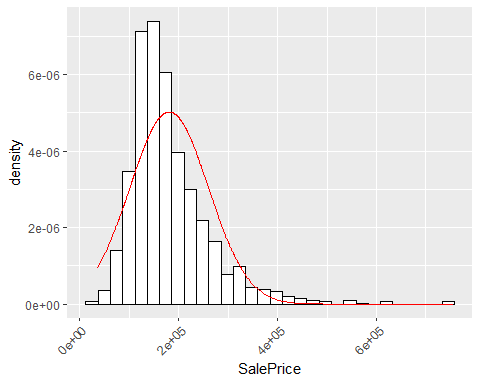


## Normality Test

Based on the analysis performed, we concluded that the distribution of the SalePrice variable in the dataset does not follow a normal distribution.

library(stats)  
# Calculate mean and standard deviation  
mean <- mean(df1$SalePrice)  
sd <- sd(df1$SalePrice)  
  
# Create a variable of sequence from minimum to maximum with 0.01 increments  
x <- seq(min(df1$SalePrice), max(df1$SalePrice), length = 100)  
  
# Add a 'Density' column to the data  
data <- df1 %>%  
 mutate(Density = dnorm(SalePrice, mean = mean, sd = sd))  
  
# Generating the histogram  
ggplot(data, aes(x = SalePrice)) +  
 geom\_histogram(aes(y = ..density..), bins = 30, colour = 'black', fill = 'white') +  
 geom\_line(aes(y = Density), colour = 'red') +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
 ℹ Please use `after\_stat(density)` instead.  
 This warning is displayed once every 8 hours.  
 Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
 generated.



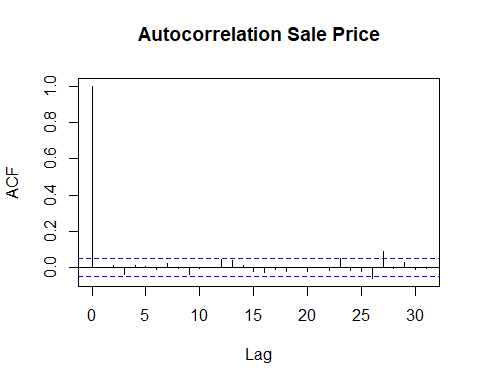
shapiro.test(df1$SalePrice) It is not normal

Shapiro-Wilk normality test  
   
 data: df1$SalePrice  
 W = 0.86967, p-value < 2.2e-16

## Test serial correlation

There is no autocorrelation in the data as we see in the plot and in the Durbin Watson test. As the p-value is greater than .05.

acf(df1$SalePrice, main="Autocorrelation Sale Price")



dwtest(SalePrice ~1, data=df1)

Durbin-Watson test  
   
 data: SalePrice ~ 1  
 DW = 2.0026, p-value = 0.5199  
 alternative hypothesis: true autocorrelation is greater than 0

## Univariate Outliers transformation

The transformations we will do to the variables are the following. For the numeric values that have mild and extreme boundaries in 0, as they have near zero variability we will convert them to categorical. For the ones that show extreme values we will created a logarithmic transformation and for those which have little extreme outliers we will keep them as they are but we will be careful to not influence the models with these observations. So for the first case we will transform BsmtFinSF2, EnclosedPorch, LowQualFinSF, MiscVal, PoolArea, ScreenPorch, MasVnrArea. For the second case we will create log variables for: LotArea, LotFrontage, TotalBsmtSF, we will sum 1 in order to avoid the log(0). And we will keep them as they are the rest of them. These transformations will be held in the steps of transformation and categorization.

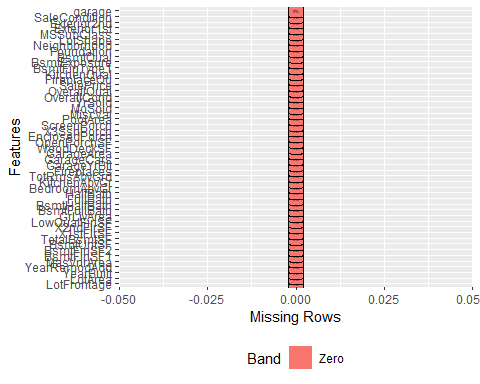
Looking at the data exploration part we can distinguish multiple columns which have outliers, we will plot a geom\_histogram to see them better  
  
num\_outliers <- c("BsmtFinSF1", "BsmtFinSF2", "BsmtHalfBath", "BsmtUnfSF", "EnclosedPorch", "GarageArea", "GrLivArea","LotArea", "LotFrontage", "LowQualFinSF", "MasVnrArea","MiscVal","OpenPorchSF","PoolArea","ScreenPorch","TotalBsmtSF", "WoodDeckSF","X1stFlrSF","X3SsnPorch")  
  
# Create a list to store the plots  
plots <- list()  
  
columna <- "MasVnrArea"  
# Loop through each numeric column  
for(i in 1:length(num\_outliers)) {  
 columna <- num\_outliers[i]  
   
 # Calculate the thresholds  
 q1 <- quantile(df1[columna],0.25, na.rm = TRUE)   
 q3 <- quantile(df1[columna],0.75, na.rm = TRUE)   
 iqr <- q3 - q1  
 mild\_l <- q1 - iqr\*1.5  
 mild\_h <- q3 + iqr\*1.5  
 high\_l <- q1 - iqr\*3  
 high\_h <- q3 + iqr\*3  
   
 # Create the plot  
 p <- ggplot(df1, aes(x=!!sym(columna))) +  
 geom\_histogram(color="black", fill="white", bins=30) +  
 geom\_vline(aes(xintercept=mild\_l), color="blue", linetype="dashed") +  
 geom\_vline(aes(xintercept=mild\_h), color="blue", linetype="dashed") +  
 geom\_vline(aes(xintercept=high\_l), color="red", linetype="dashed") +  
 geom\_vline(aes(xintercept=high\_h), color="red", linetype="dashed") +  
 labs(x = columna, y="Frequency", title = paste("Histogram of", columna))  
 # Add the plot to the list  
 print(p)  
}  
  
  
 Multivariate

## NA imputation

After imputing the data logically, creating the levels NoFirePlace and NoBasement, we perform advanced imputation for imputing numerical values, creating sinthetic values.

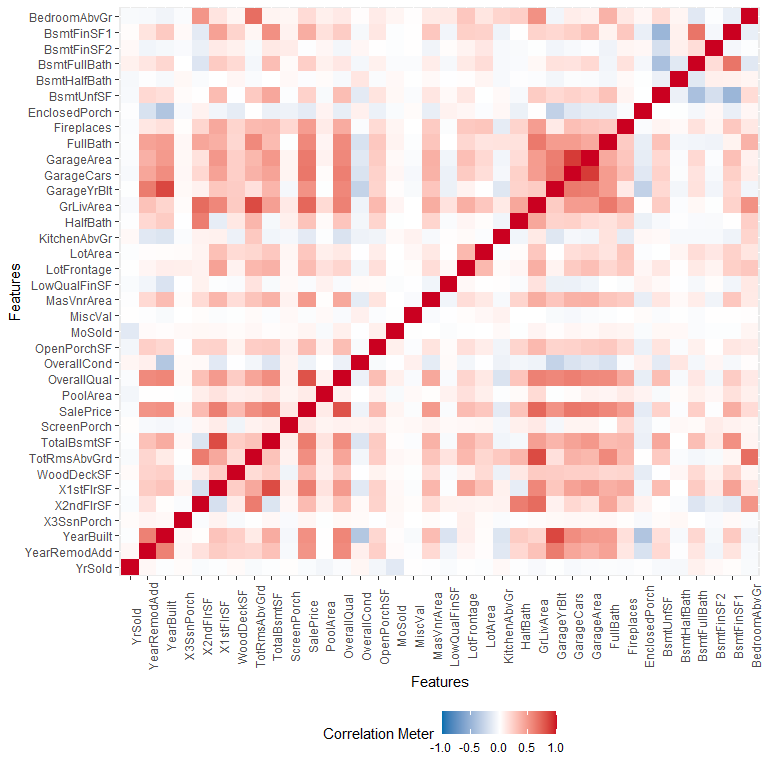
Impute missing data by creating extra modality  
  
df1$FireplaceQu <- ifelse(df1$FireplaceQu %>% is.na, "NoFirePlace", df1$FireplaceQu)  
df1$BsmtExposure <- ifelse(df1$BsmtExposure %>% is.na, "NoBasement", df1$BsmtExposure)  
df1$BsmtFinType1 <- ifelse(df1$BsmtFinType1 %>% is.na, "NoBasement", df1$BsmtFinType1)  
df1$BsmtQual <- ifelse(df1$BsmtQual %>% is.na, "NoBasement", df1$BsmtQual)  
  
df1$garage <- ifelse(df1$GarageYrBlt %>% is.na, "NO", "YES")  
 Impute with synthetic values  
df2 <- mice(df1, method = "cart")

df3 <- complete(df2)  
  
plot\_missing(df3, missing\_only = FALSE, group = list("Zero"=0,"Low" = 0.05, "Medium"=0.25, "High"=0.5, "Very High" =1), geom\_label\_args = list("size" = 1))



## Correlation Matrix

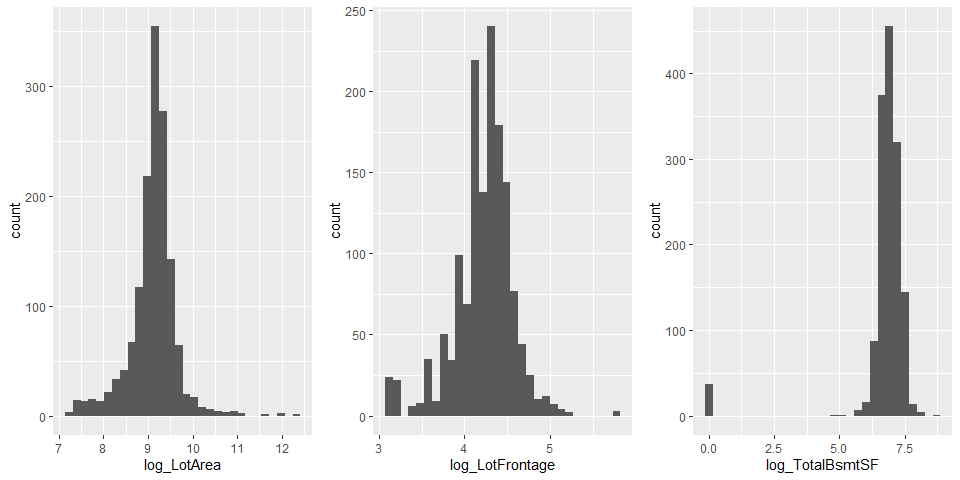
df\_num <- df3[order(numeric\_columns, decreasing = TRUE)]  
  
plot\_correlation(df\_num)



## Numerical transformation

In order to minimize the effect of 0 values in the dataset we will create a dummy variable for No Basement. It will be very useful for modelling as we will use it to create interactions with the numerical features.

df3$log\_LotArea <- log(df3$LotArea + 1)  
df3$log\_LotFrontage <- log(df3$LotFrontage + 1)  
df3$log\_TotalBsmtSF <- log(df3$TotalBsmtSF + 1)  
  
df3$YearRemodAdd <- df3$YearRemodAdd - df3$YearBuilt  
  
g1 <- ggplot(df3) + geom\_histogram(aes(x=log\_LotArea))  
g2 <- ggplot(df3) + geom\_histogram(aes(x=log\_LotFrontage))  
g3 <- ggplot(df3) + geom\_histogram(aes(x=log\_TotalBsmtSF))  
  
grid.arrange(g1,g2,g3, ncol=3)



df3$bsmt <- ifelse(df3$log\_TotalBsmtSF == 0, "NO","YES")

## Categorization transformation

We opted to transform some features into categories because many data points in some variables are zeros, which might be less informative for analysis. Variables like BsmtFinSF2, EnclosedPorch, LowQualFinSF, and others are transformed. If the original value is greater than zero, it’s categorized as “Yes”; otherwise, it’s labeled “Zero” or “No”. This transformation improves interpretability and is beneficial for statistical analyses, particularly when dealing with variables heavily skewed towards zero.

df3$cat\_BsmntFinSF2 <- ifelse(df3$BsmtFinSF2 > 0, "Yes","Zero")  
df3$cat\_EnclosedPorch <- ifelse(df3$EnclosedPorch > 0, "Yes","Zero")  
df3$cat\_LowQualFinSF <- ifelse(df3$LowQualFinSF > 0, "Yes","Zero")  
df3$cat\_MiscVal <- ifelse(df3$MiscVal > 0, "Yes","Zero")  
df3$cat\_PoolArea <- ifelse(df3$PoolArea > 0, "Yes","No")  
df3$cat\_ScreenPorch <- ifelse(df3$ScreenPorch > 0, "Yes","No")  
df3$cat\_X3SsnPorch <- ifelse(df3$X3SsnPorch > 0, "Yes","No")  
df3$remod <- ifelse(df3$YearRemodAdd == 0, "No", "Yes")  
  
df3$SaleCondition <- relevel(x=df3$SaleCondition, ref="Normal")

## Interaction between numerical and categorical

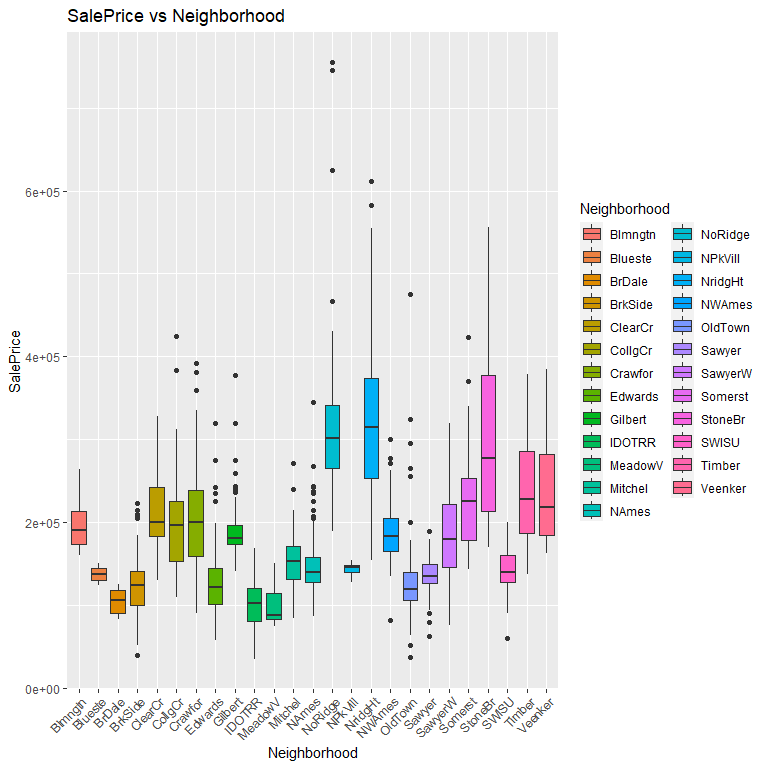
The first graph shows the variation in sale prices of properties across different Neighborhoods. The neighorhood can significantly influence its sale price.This graph helps to understand if certain neighorhoods have higher or lower average prices.

The second graph shows the variation in sale prices of properties across different municipal zoning classifications (MSZoning), which may include residential, commercial, agricultural, etc. The zoning of a property can significantly influence its sale price due to factors such as proximity to amenities, population density, and municipal regulations. This graph helps to understand if certain zones have higher or lower average prices.

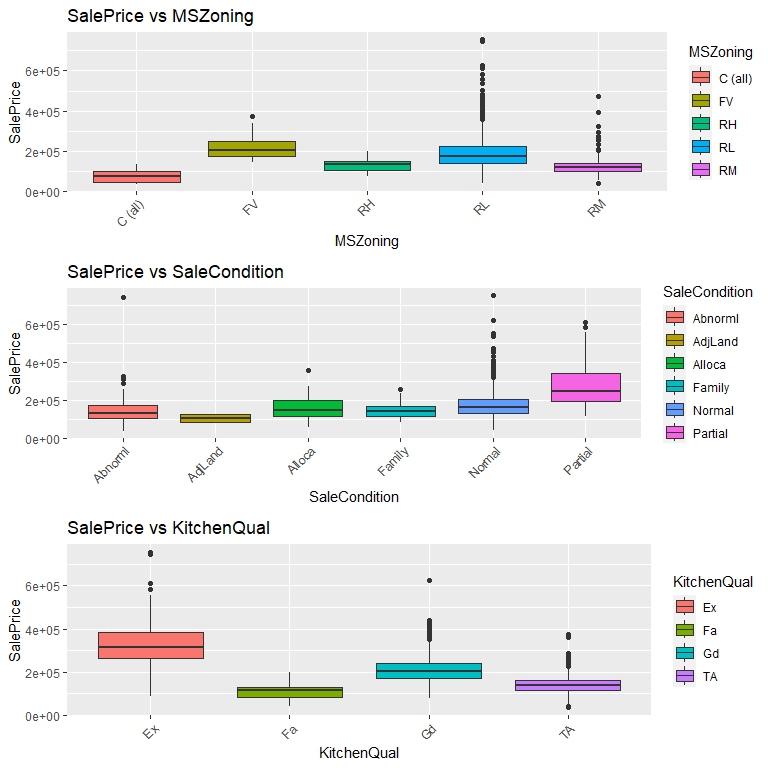
The second graph compares the sale price with the sale condition (SaleCondition), which includes categories like normal, foreclosure, and urgency sale. The condition of the sale can reflect specific circumstances that affect the price, like urgent sales that might result in lower prices or normal condition sales that might reflect the market value. This analysis helps to identify if certain sale conditions are associated with significant price variations.

In the last graph, it’s shown how the sale price varies with the quality of the kitchen (KitchenQual), which can be excellent, good, average, etc. The quality of a kitchen is a significant factor in property valuation. A high-quality kitchen can increase a property’s value, while a low-quality one can decrease.

g1 <- ggplot(data=df3) +  
 geom\_boxplot(aes(y=SalePrice, x = Neighborhood, fill=Neighborhood)) +  
 labs(title="SalePrice vs Neighborhood") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
g2 <- ggplot(df, aes(x=MSZoning, y=SalePrice, fill=MSZoning)) +  
 geom\_boxplot() +  
 labs(title="SalePrice vs MSZoning") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
g3 <- ggplot(df, aes(x=SaleCondition, y=SalePrice, fill=SaleCondition)) +  
 geom\_boxplot() +  
 labs(title="SalePrice vs SaleCondition") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
g4 <- ggplot(df, aes(x=KitchenQual, y=SalePrice, fill=KitchenQual)) +  
 geom\_boxplot() +  
 labs(title="SalePrice vs KitchenQual") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
g1



grid.arrange(g2,g3,g4)



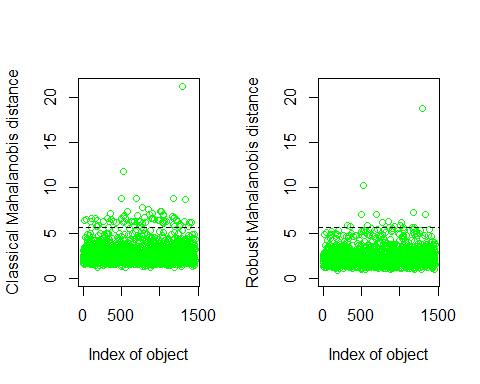
## Multivariate outliers

In our analysis with the Moutlier function, we faced issues due to the poor distribution of some numeric variables, which caused computational errors. To resolve this, we excluded variables with fewer than 50 unique values. This step was necessary to ensure the effectiveness of the outlier detection process, as variables with limited unique values can hinder the accurate identification of outliers. Additionally, we identified and removed multivariate outliers (moutliers) from our training dataset. These outliers are observations that significantly deviate from the norm across several dimensions. Removing these moutliers is crucial for enhancing model accuracy, as they can skew results and lead to misleading interpretations.

set.seed(1)  
numeric\_columns <- c("log\_LotFrontage", "log\_LotArea", "MasVnrArea", "BsmtFinSF1", "BsmtUnfSF", "log\_TotalBsmtSF", "X1stFlrSF", "X2ndFlrSF", "GrLivArea", "BsmtFullBath", "BsmtHalfBath", "FullBath", "HalfBath", "BedroomAbvGr", "KitchenAbvGr", "TotRmsAbvGrd", "Fireplaces", "GarageYrBlt", "GarageCars", "GarageArea", "WoodDeckSF", "OpenPorchSF", "MoSold", "YrSold", "OverallCond","OverallQual","SalePrice")  
  
  
df\_clean = df3[complete.cases(df3[, numeric\_columns]), numeric\_columns]  
  
sum(is.na(df\_clean))

[1] 0

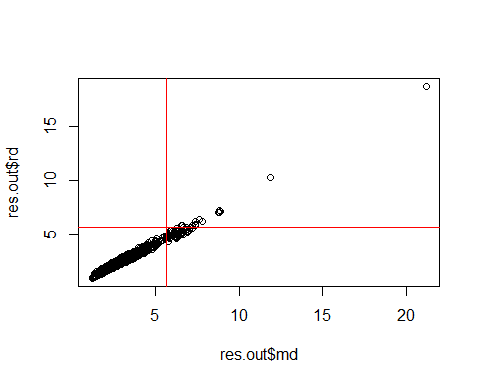
data\_numeric <- df\_clean[, sapply(df\_clean, is.numeric)]  
  
threshold <- 50  
  
filtered\_numeric\_vars <- sapply(data\_numeric, function(x) length(unique(x)) > threshold)  
final\_numeric\_data <- data\_numeric[, filtered\_numeric\_vars]  
  
res.out = Moutlier(final\_numeric\_data[, !(names(final\_numeric\_data) %in% c("LotFrontage", "log\_LotArea", "TotalBsmtSF", "YearBuilt", "YearRemodAdd", "MasVnrArea", "BsmtFinSF2", "X2ndFlrSF", "WoodDeckSF", "EnclosedPorch", "ScreenPorch"))], quantile = 0.9995, col="green")



outlier\_index <- which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))  
length(outlier\_index)

[1] 17

par(mfrow=c(1,1))  
plot( res.out$md, res.out$rd )  
abline(h=res.out$cutoff, col="red")  
abline(v=res.out$cutoff, col="red")



df3 <- df3[-outlier\_index,]

## Model

First, we will add all the variables that are do not have correlation between each other. We have iterated to keep those that are not multicollinear with each other by adding and removing. But keeping always the ones that we preprocessed.

df\_model <- df3 %>% select(c("log\_LotFrontage","log\_LotArea","YearBuilt","YearRemodAdd","FullBath","garage","GarageArea","bsmt","log\_TotalBsmtSF","cat\_PoolArea","YrSold","MoSold","remod","GarageCars","GarageYrBlt","BedroomAbvGr","KitchenAbvGr","TotRmsAbvGrd","Neighborhood","GrLivArea","MSSubClass","X1stFlrSF","X2ndFlrSF","OverallCond","OverallQual","FireplaceQu","cat\_MiscVal","OpenPorchSF", "cat\_EnclosedPorch","cat\_LowQualFinSF","cat\_X3SsnPorch","ScreenPorch","Foundation","Exterior1st","Exterior2nd","SaleCondition","SalePrice"))  
  
  
attach(df\_model)  
mod1\_raw <- lm(SalePrice ~ log\_LotFrontage + log\_LotArea + X1stFlrSF + X2ndFlrSF + YearBuilt + YearRemodAdd + FullBath + GarageArea + GarageCars + GarageYrBlt:garage + BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd + Neighborhood + MSSubClass + log\_TotalBsmtSF + cat\_PoolArea + YrSold:MoSold + OverallCond + OverallQual + FireplaceQu + cat\_MiscVal + OpenPorchSF + ScreenPorch + Foundation +SaleCondition, data=df\_model)  
summary(mod1\_raw)

Call:  
 lm(formula = SalePrice ~ log\_LotFrontage + log\_LotArea + X1stFlrSF +   
 X2ndFlrSF + YearBuilt + YearRemodAdd + FullBath + GarageArea +   
 GarageCars + GarageYrBlt:garage + BedroomAbvGr + KitchenAbvGr +   
 TotRmsAbvGrd + Neighborhood + MSSubClass + log\_TotalBsmtSF +   
 cat\_PoolArea + YrSold:MoSold + OverallCond + OverallQual +   
 FireplaceQu + cat\_MiscVal + OpenPorchSF + ScreenPorch + Foundation +   
 SaleCondition, data = df\_model)  
   
 Residuals:  
 Min 1Q Median 3Q Max   
 -156169 -13653 315 12070 200829   
   
 Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
 (Intercept) -1.167e+06 1.757e+05 -6.642 4.44e-11 \*\*\*  
 log\_LotFrontage -8.227e+03 3.410e+03 -2.413 0.015969 \*   
 log\_LotArea 2.020e+04 2.542e+03 7.947 3.98e-15 \*\*\*  
 X1stFlrSF 7.607e+01 4.375e+00 17.388 < 2e-16 \*\*\*  
 X2ndFlrSF 5.398e+01 5.014e+00 10.766 < 2e-16 \*\*\*  
 YearBuilt 5.549e+02 8.859e+01 6.263 5.04e-10 \*\*\*  
 YearRemodAdd 3.990e+01 5.426e+01 0.735 0.462187   
 FullBath -2.312e+03 2.102e+03 -1.100 0.271603   
 GarageArea 9.685e+00 8.333e+00 1.162 0.245327   
 GarageCars 8.719e+03 2.397e+03 3.637 0.000286 \*\*\*  
 BedroomAbvGr -1.088e+04 1.403e+03 -7.758 1.68e-14 \*\*\*  
 KitchenAbvGr -2.252e+04 5.755e+03 -3.914 9.52e-05 \*\*\*  
 TotRmsAbvGrd 3.543e+03 9.974e+02 3.552 0.000395 \*\*\*  
 NeighborhoodBlueste 2.238e+04 2.036e+04 1.099 0.271851   
 NeighborhoodBrDale 1.922e+04 1.114e+04 1.724 0.084904 .   
 NeighborhoodBrkSide 8.918e+03 9.024e+03 0.988 0.323231   
 NeighborhoodClearCr 3.923e+03 9.432e+03 0.416 0.677526   
 NeighborhoodCollgCr -4.686e+03 7.780e+03 -0.602 0.547088   
 NeighborhoodCrawfor 2.278e+04 8.778e+03 2.596 0.009543 \*\*   
 NeighborhoodEdwards -5.183e+03 8.383e+03 -0.618 0.536474   
 NeighborhoodGilbert -1.145e+04 8.183e+03 -1.400 0.161824   
 NeighborhoodIDOTRR -1.652e+03 9.476e+03 -0.174 0.861666   
 NeighborhoodMeadowV 1.738e+04 1.118e+04 1.555 0.120065   
 NeighborhoodMitchel -8.492e+03 8.585e+03 -0.989 0.322745   
 NeighborhoodNAmes -4.733e+03 8.126e+03 -0.582 0.560354   
 NeighborhoodNoRidge 2.734e+04 8.825e+03 3.098 0.001991 \*\*   
 NeighborhoodNPkVill 1.753e+04 1.158e+04 1.514 0.130380   
 NeighborhoodNridgHt 4.052e+04 7.736e+03 5.238 1.88e-07 \*\*\*  
 NeighborhoodNWAmes -1.651e+04 8.288e+03 -1.992 0.046574 \*   
 NeighborhoodOldTown -4.975e+03 8.757e+03 -0.568 0.570064   
 NeighborhoodSawyer -4.759e+03 8.480e+03 -0.561 0.574739   
 NeighborhoodSawyerW -5.926e+03 8.221e+03 -0.721 0.471157   
 NeighborhoodSomerst 6.694e+03 8.015e+03 0.835 0.403732   
 NeighborhoodStoneBr 5.316e+04 8.700e+03 6.110 1.29e-09 \*\*\*  
 NeighborhoodSWISU 5.330e+03 9.948e+03 0.536 0.592208   
 NeighborhoodTimber 1.213e+03 8.741e+03 0.139 0.889665   
 NeighborhoodVeenker 1.672e+04 1.087e+04 1.538 0.124333   
 MSSubClass30 2.563e+02 4.860e+03 0.053 0.957940   
 MSSubClass40 1.720e+03 1.345e+04 0.128 0.898305   
 MSSubClass45 -6.090e+03 8.319e+03 -0.732 0.464273   
 MSSubClass50 -5.730e+03 4.212e+03 -1.360 0.173951   
 MSSubClass60 -2.501e+03 4.521e+03 -0.553 0.580228   
 MSSubClass70 -7.554e+03 6.108e+03 -1.237 0.216367   
 MSSubClass75 -7.603e+03 8.860e+03 -0.858 0.390941   
 MSSubClass80 -1.363e+03 3.817e+03 -0.357 0.721051   
 MSSubClass85 4.090e+03 6.217e+03 0.658 0.510733   
 MSSubClass90 -3.477e+03 6.202e+03 -0.561 0.575180   
 MSSubClass120 -2.167e+04 4.432e+03 -4.891 1.12e-06 \*\*\*  
 MSSubClass160 -2.392e+04 6.794e+03 -3.521 0.000445 \*\*\*  
 MSSubClass180 -1.248e+04 1.105e+04 -1.129 0.259073   
 MSSubClass190 1.484e+02 6.701e+03 0.022 0.982330   
 log\_TotalBsmtSF 2.856e+03 1.104e+03 2.586 0.009804 \*\*   
 cat\_PoolAreaYes 9.717e+03 1.228e+04 0.791 0.428895   
 OverallCond 6.531e+03 8.257e+02 7.910 5.26e-15 \*\*\*  
 OverallQual 1.168e+04 9.940e+02 11.755 < 2e-16 \*\*\*  
 FireplaceQu2 -1.422e+04 7.622e+03 -1.865 0.062364 .   
 FireplaceQu3 -1.529e+04 5.979e+03 -2.558 0.010641 \*   
 FireplaceQu4 -1.541e+04 8.540e+03 -1.804 0.071408 .   
 FireplaceQu5 -1.887e+04 6.185e+03 -3.051 0.002327 \*\*   
 FireplaceQuNoFirePlace -1.706e+04 6.222e+03 -2.742 0.006195 \*\*   
 cat\_MiscValZero 3.212e+03 3.847e+03 0.835 0.403917   
 OpenPorchSF 3.441e+01 1.296e+01 2.655 0.008025 \*\*   
 ScreenPorch 5.266e+01 1.319e+01 3.994 6.85e-05 \*\*\*  
 FoundationCBlock 3.231e+03 3.289e+03 0.983 0.326018   
 FoundationPConc 7.042e+03 3.634e+03 1.938 0.052840 .   
 FoundationSlab 1.734e+04 9.973e+03 1.739 0.082265 .   
 FoundationStone 1.029e+04 1.112e+04 0.925 0.354924   
 FoundationWood -1.660e+04 1.595e+04 -1.041 0.298011   
 SaleConditionAbnorml -9.446e+03 2.830e+03 -3.338 0.000868 \*\*\*  
 SaleConditionAdjLand 9.308e+03 1.395e+04 0.667 0.504813   
 SaleConditionAlloca -6.013e+03 8.195e+03 -0.734 0.463217   
 SaleConditionFamily -6.772e+03 5.999e+03 -1.129 0.259201   
 SaleConditionPartial 1.630e+04 2.960e+03 5.506 4.39e-08 \*\*\*  
 GarageYrBlt:garageNO -4.447e+01 6.030e+01 -0.737 0.460977   
 GarageYrBlt:garageYES -5.338e+01 6.028e+01 -0.885 0.376066   
 YrSold:MoSold -1.823e-01 1.305e-01 -1.396 0.162845   
 ---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   
 Residual standard error: 25950 on 1367 degrees of freedom  
 Multiple R-squared: 0.8887, Adjusted R-squared: 0.8826   
 F-statistic: 145.5 on 75 and 1367 DF, p-value: < 2.2e-16

vif(mod1\_raw) %>% round(2) %>% t

log\_LotFrontage log\_LotArea X1stFlrSF X2ndFlrSF YearBuilt  
 GVIF^(1/(2\*Df)) 1.72 1.90 2.33 3.14 3.91  
 YearRemodAdd FullBath GarageArea GarageCars BedroomAbvGr  
 GVIF^(1/(2\*Df)) 1.94 1.68 2.56 2.61 1.67  
 KitchenAbvGr TotRmsAbvGrd Neighborhood MSSubClass  
 GVIF^(1/(2\*Df)) 1.84 2.34 1.14 1.31  
 log\_TotalBsmtSF cat\_PoolArea OverallCond OverallQual  
 GVIF^(1/(2\*Df)) 1.77 1.06 1.34 1.98  
 FireplaceQu cat\_MiscVal OpenPorchSF ScreenPorch Foundation  
 GVIF^(1/(2\*Df)) 1.12 1.04 1.16 1.04 1.34  
 SaleCondition GarageYrBlt:garage YrSold:MoSold  
 GVIF^(1/(2\*Df)) 1.08 1.75 1.03

## Pruned model

Deleting not significant variables   
df\_model <- df3 %>% select( c("log\_LotFrontage","log\_LotArea","YearBuilt","YearRemodAdd","FullBath","garage","GarageArea","bsmt","log\_TotalBsmtSF","cat\_PoolArea","YrSold","MoSold","remod","GarageCars","GarageYrBlt","BedroomAbvGr","KitchenAbvGr","TotRmsAbvGrd","Neighborhood","GrLivArea","MSSubClass","X1stFlrSF","X2ndFlrSF","OverallCond","OverallQual","FireplaceQu","cat\_MiscVal","OpenPorchSF", "cat\_EnclosedPorch","cat\_LowQualFinSF","cat\_X3SsnPorch","ScreenPorch","Foundation","Exterior1st","Exterior2nd","SaleCondition","SalePrice"  
 )  
)  
attach(df\_model)  
mod1 <- lm(SalePrice ~ log\_LotFrontage + log\_LotArea + X1stFlrSF + X2ndFlrSF + YearBuilt + GarageCars + BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd + Neighborhood + MSSubClass + log\_TotalBsmtSF + SaleCondition + OverallCond + OverallQual + FireplaceQu + ScreenPorch, data=df\_model)  
summary(mod1)

Call:  
 lm(formula = SalePrice ~ log\_LotFrontage + log\_LotArea + X1stFlrSF +   
 X2ndFlrSF + YearBuilt + GarageCars + BedroomAbvGr + KitchenAbvGr +   
 TotRmsAbvGrd + Neighborhood + MSSubClass + log\_TotalBsmtSF +   
 SaleCondition + OverallCond + OverallQual + FireplaceQu +   
 ScreenPorch, data = df\_model)  
   
 Residuals:  
 Min 1Q Median 3Q Max   
 -154719 -13592 74 12516 200220   
   
 Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
 (Intercept) -1.255e+06 1.419e+05 -8.842 < 2e-16 \*\*\*  
 log\_LotFrontage -7.149e+03 3.383e+03 -2.113 0.034784 \*   
 log\_LotArea 1.972e+04 2.534e+03 7.782 1.39e-14 \*\*\*  
 X1stFlrSF 7.917e+01 4.079e+00 19.410 < 2e-16 \*\*\*  
 X2ndFlrSF 5.493e+01 4.872e+00 11.275 < 2e-16 \*\*\*  
 YearBuilt 5.543e+02 6.882e+01 8.054 1.71e-15 \*\*\*  
 GarageCars 6.564e+03 1.368e+03 4.796 1.79e-06 \*\*\*  
 BedroomAbvGr -1.142e+04 1.383e+03 -8.256 3.48e-16 \*\*\*  
 KitchenAbvGr -2.269e+04 5.729e+03 -3.961 7.85e-05 \*\*\*  
 TotRmsAbvGrd 3.709e+03 9.970e+02 3.720 0.000207 \*\*\*  
 NeighborhoodBlueste 2.259e+04 2.048e+04 1.103 0.270085   
 NeighborhoodBrDale 1.886e+04 1.103e+04 1.711 0.087376 .   
 NeighborhoodBrkSide 1.060e+04 9.002e+03 1.177 0.239392   
 NeighborhoodClearCr 5.145e+03 9.445e+03 0.545 0.585978   
 NeighborhoodCollgCr -3.098e+03 7.790e+03 -0.398 0.690957   
 NeighborhoodCrawfor 2.178e+04 8.781e+03 2.481 0.013231 \*   
 NeighborhoodEdwards -3.101e+03 8.393e+03 -0.369 0.711820   
 NeighborhoodGilbert -1.213e+04 8.202e+03 -1.479 0.139274   
 NeighborhoodIDOTRR 1.441e+03 9.440e+03 0.153 0.878733   
 NeighborhoodMeadowV 2.095e+04 1.110e+04 1.886 0.059444 .   
 NeighborhoodMitchel -6.931e+03 8.597e+03 -0.806 0.420315   
 NeighborhoodNAmes -4.969e+03 8.065e+03 -0.616 0.537929   
 NeighborhoodNoRidge 3.033e+04 8.840e+03 3.431 0.000619 \*\*\*  
 NeighborhoodNPkVill 1.712e+04 1.149e+04 1.490 0.136415   
 NeighborhoodNridgHt 4.364e+04 7.727e+03 5.648 1.97e-08 \*\*\*  
 NeighborhoodNWAmes -1.698e+04 8.266e+03 -2.055 0.040091 \*   
 NeighborhoodOldTown -3.652e+03 8.742e+03 -0.418 0.676236   
 NeighborhoodSawyer -5.322e+03 8.470e+03 -0.628 0.529879   
 NeighborhoodSawyerW -5.283e+03 8.247e+03 -0.641 0.521907   
 NeighborhoodSomerst 1.078e+04 7.935e+03 1.359 0.174360   
 NeighborhoodStoneBr 5.419e+04 8.741e+03 6.200 7.45e-10 \*\*\*  
 NeighborhoodSWISU 7.231e+03 9.921e+03 0.729 0.466222   
 NeighborhoodTimber 1.933e+03 8.790e+03 0.220 0.825972   
 NeighborhoodVeenker 1.846e+04 1.091e+04 1.692 0.090879 .   
 MSSubClass30 9.802e+02 4.705e+03 0.208 0.834999   
 MSSubClass40 -2.777e+03 1.348e+04 -0.206 0.836805   
 MSSubClass45 -4.971e+03 8.285e+03 -0.600 0.548580   
 MSSubClass50 -7.764e+03 4.167e+03 -1.863 0.062677 .   
 MSSubClass60 -1.647e+03 4.514e+03 -0.365 0.715279   
 MSSubClass70 -7.165e+03 6.071e+03 -1.180 0.238142   
 MSSubClass75 -4.326e+03 8.823e+03 -0.490 0.623974   
 MSSubClass80 -2.360e+03 3.825e+03 -0.617 0.537465   
 MSSubClass85 4.264e+03 6.273e+03 0.680 0.496775   
 MSSubClass90 -3.020e+03 6.101e+03 -0.495 0.620671   
 MSSubClass120 -2.209e+04 4.405e+03 -5.014 6.02e-07 \*\*\*  
 MSSubClass160 -2.608e+04 6.800e+03 -3.836 0.000131 \*\*\*  
 MSSubClass180 -1.191e+04 1.112e+04 -1.072 0.284108   
 MSSubClass190 2.679e+03 6.724e+03 0.398 0.690372   
 log\_TotalBsmtSF 1.708e+03 7.286e+02 2.344 0.019207 \*   
 SaleConditionAbnorml -8.672e+03 2.844e+03 -3.049 0.002340 \*\*   
 SaleConditionAdjLand 1.531e+04 1.386e+04 1.105 0.269357   
 SaleConditionAlloca -4.834e+03 8.132e+03 -0.595 0.552270   
 SaleConditionFamily -7.674e+03 6.052e+03 -1.268 0.205020   
 SaleConditionPartial 1.722e+04 2.959e+03 5.820 7.32e-09 \*\*\*  
 OverallCond 6.625e+03 7.378e+02 8.979 < 2e-16 \*\*\*  
 OverallQual 1.183e+04 9.879e+02 11.970 < 2e-16 \*\*\*  
 FireplaceQu2 -1.571e+04 7.658e+03 -2.051 0.040428 \*   
 FireplaceQu3 -1.658e+04 5.998e+03 -2.765 0.005769 \*\*   
 FireplaceQu4 -1.522e+04 8.577e+03 -1.774 0.076225 .   
 FireplaceQu5 -1.986e+04 6.180e+03 -3.215 0.001336 \*\*   
 FireplaceQuNoFirePlace -1.813e+04 6.227e+03 -2.911 0.003660 \*\*   
 ScreenPorch 5.167e+01 1.327e+01 3.892 0.000104 \*\*\*  
 ---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   
 Residual standard error: 26220 on 1381 degrees of freedom  
 Multiple R-squared: 0.8852, Adjusted R-squared: 0.8801   
 F-statistic: 174.5 on 61 and 1381 DF, p-value: < 2.2e-16

vif(mod1) %>% round(2) %>% t

log\_LotFrontage log\_LotArea X1stFlrSF X2ndFlrSF YearBuilt  
 GVIF^(1/(2\*Df)) 1.69 1.88 2.15 3.02 3.01  
 GarageCars BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Neighborhood  
 GVIF^(1/(2\*Df)) 1.47 1.63 1.81 2.32 1.12  
 MSSubClass log\_TotalBsmtSF SaleCondition OverallCond  
 GVIF^(1/(2\*Df)) 1.29 1.15 1.07 1.19  
 OverallQual FireplaceQu ScreenPorch  
 GVIF^(1/(2\*Df)) 1.95 1.11 1.04

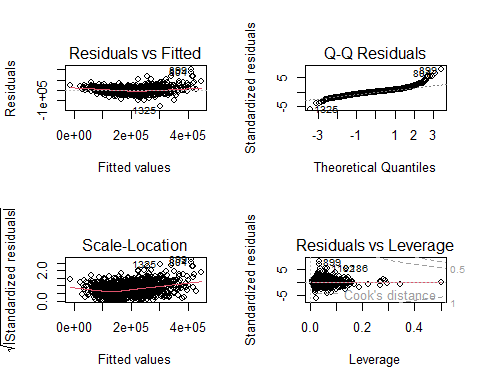
# Stepwise   
mod\_step <- step(mod1\_raw, trace = FALSE)  
summary(mod\_step)

Call:  
 lm(formula = SalePrice ~ log\_LotFrontage + log\_LotArea + X1stFlrSF +   
 X2ndFlrSF + YearBuilt + GarageCars + BedroomAbvGr + KitchenAbvGr +   
 TotRmsAbvGrd + Neighborhood + MSSubClass + log\_TotalBsmtSF +   
 OverallCond + OverallQual + FireplaceQu + OpenPorchSF + ScreenPorch +   
 SaleCondition + GarageYrBlt:garage + YrSold:MoSold, data = df\_model)  
   
 Residuals:  
 Min 1Q Median 3Q Max   
 -156109 -13643 379 12442 200563   
   
 Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
 (Intercept) -1.184e+06 1.521e+05 -7.783 1.38e-14 \*\*\*  
 log\_LotFrontage -8.493e+03 3.366e+03 -2.523 0.011742 \*   
 log\_LotArea 1.979e+04 2.517e+03 7.862 7.57e-15 \*\*\*  
 X1stFlrSF 7.709e+01 4.069e+00 18.946 < 2e-16 \*\*\*  
 X2ndFlrSF 5.390e+01 4.837e+00 11.145 < 2e-16 \*\*\*  
 YearBuilt 5.449e+02 7.362e+01 7.402 2.33e-13 \*\*\*  
 GarageCars 1.086e+04 1.719e+03 6.319 3.54e-10 \*\*\*  
 BedroomAbvGr -1.116e+04 1.375e+03 -8.117 1.05e-15 \*\*\*  
 KitchenAbvGr -2.218e+04 5.684e+03 -3.902 0.000100 \*\*\*  
 TotRmsAbvGrd 3.395e+03 9.933e+02 3.418 0.000649 \*\*\*  
 NeighborhoodBlueste 2.131e+04 2.030e+04 1.050 0.293951   
 NeighborhoodBrDale 1.884e+04 1.097e+04 1.717 0.086182 .   
 NeighborhoodBrkSide 9.184e+03 8.982e+03 1.023 0.306714   
 NeighborhoodClearCr 4.536e+03 9.414e+03 0.482 0.629973   
 NeighborhoodCollgCr -3.227e+03 7.757e+03 -0.416 0.677449   
 NeighborhoodCrawfor 2.326e+04 8.763e+03 2.654 0.008041 \*\*   
 NeighborhoodEdwards -4.335e+03 8.364e+03 -0.518 0.604301   
 NeighborhoodGilbert -1.121e+04 8.153e+03 -1.375 0.169210   
 NeighborhoodIDOTRR 4.969e+02 9.418e+03 0.053 0.957931   
 NeighborhoodMeadowV 1.863e+04 1.105e+04 1.686 0.092110 .   
 NeighborhoodMitchel -8.530e+03 8.562e+03 -0.996 0.319270   
 NeighborhoodNAmes -4.131e+03 8.057e+03 -0.513 0.608225   
 NeighborhoodNoRidge 2.949e+04 8.772e+03 3.362 0.000795 \*\*\*  
 NeighborhoodNPkVill 1.499e+04 1.143e+04 1.311 0.189949   
 NeighborhoodNridgHt 4.193e+04 7.689e+03 5.453 5.86e-08 \*\*\*  
 NeighborhoodNWAmes -1.739e+04 8.237e+03 -2.111 0.034924 \*   
 NeighborhoodOldTown -4.493e+03 8.738e+03 -0.514 0.607216   
 NeighborhoodSawyer -4.200e+03 8.437e+03 -0.498 0.618718   
 NeighborhoodSawyerW -5.097e+03 8.214e+03 -0.621 0.535013   
 NeighborhoodSomerst 8.063e+03 7.971e+03 1.012 0.311902   
 NeighborhoodStoneBr 5.410e+04 8.680e+03 6.233 6.07e-10 \*\*\*  
 NeighborhoodSWISU 6.091e+03 9.897e+03 0.615 0.538396   
 NeighborhoodTimber 8.560e+02 8.730e+03 0.098 0.921905   
 NeighborhoodVeenker 1.769e+04 1.084e+04 1.633 0.102729   
 MSSubClass30 -1.059e+03 4.688e+03 -0.226 0.821339   
 MSSubClass40 6.338e+02 1.337e+04 0.047 0.962202   
 MSSubClass45 -7.420e+03 8.247e+03 -0.900 0.368420   
 MSSubClass50 -6.718e+03 4.136e+03 -1.624 0.104560   
 MSSubClass60 -2.439e+03 4.503e+03 -0.542 0.588090   
 MSSubClass70 -7.057e+03 6.017e+03 -1.173 0.241039   
 MSSubClass75 -7.282e+03 8.778e+03 -0.830 0.406939   
 MSSubClass80 -1.177e+03 3.798e+03 -0.310 0.756683   
 MSSubClass85 3.904e+03 6.216e+03 0.628 0.530129   
 MSSubClass90 -4.901e+03 6.092e+03 -0.805 0.421176   
 MSSubClass120 -2.222e+04 4.391e+03 -5.059 4.78e-07 \*\*\*  
 MSSubClass160 -2.544e+04 6.765e+03 -3.761 0.000176 \*\*\*  
 MSSubClass180 -1.475e+04 1.103e+04 -1.336 0.181633   
 MSSubClass190 -4.274e+02 6.697e+03 -0.064 0.949130   
 log\_TotalBsmtSF 1.783e+03 7.222e+02 2.469 0.013676 \*   
 OverallCond 6.638e+03 7.332e+02 9.054 < 2e-16 \*\*\*  
 OverallQual 1.170e+04 9.807e+02 11.926 < 2e-16 \*\*\*  
 FireplaceQu2 -1.594e+04 7.594e+03 -2.099 0.035966 \*   
 FireplaceQu3 -1.671e+04 5.946e+03 -2.811 0.005015 \*\*   
 FireplaceQu4 -1.651e+04 8.523e+03 -1.937 0.052968 .   
 FireplaceQu5 -2.092e+04 6.132e+03 -3.412 0.000664 \*\*\*  
 FireplaceQuNoFirePlace -1.875e+04 6.175e+03 -3.037 0.002437 \*\*   
 OpenPorchSF 3.762e+01 1.293e+01 2.909 0.003684 \*\*   
 ScreenPorch 5.388e+01 1.316e+01 4.095 4.47e-05 \*\*\*  
 SaleConditionAbnorml -9.331e+03 2.826e+03 -3.302 0.000985 \*\*\*  
 SaleConditionAdjLand 8.558e+03 1.380e+04 0.620 0.535133   
 SaleConditionAlloca -4.820e+03 8.061e+03 -0.598 0.550002   
 SaleConditionFamily -6.360e+03 6.001e+03 -1.060 0.289452   
 SaleConditionPartial 1.662e+04 2.951e+03 5.631 2.17e-08 \*\*\*  
 GarageYrBlt:garageNO -1.619e+01 5.627e+01 -0.288 0.773563   
 GarageYrBlt:garageYES -2.524e+01 5.622e+01 -0.449 0.653565   
 YrSold:MoSold -2.135e-01 1.302e-01 -1.640 0.101328   
 ---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   
 Residual standard error: 25980 on 1377 degrees of freedom  
 Multiple R-squared: 0.8876, Adjusted R-squared: 0.8823   
 F-statistic: 167.3 on 65 and 1377 DF, p-value: < 2.2e-16

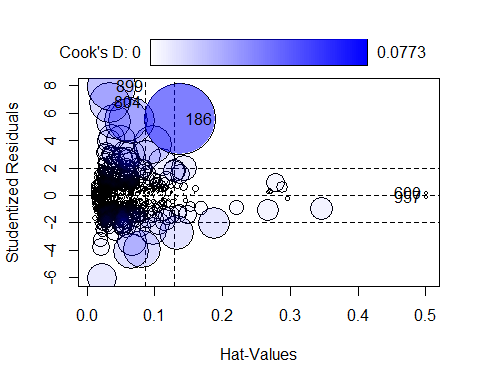
So in the end little difference from removing not significant variables with the stepwise procedure. We will keep mod1 as the final one.

## Validation

par(mfrow=c(2,2))  
plot(mod1)

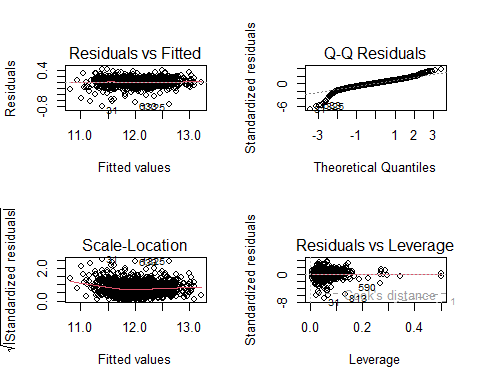


We can see that the extreme values are not being predicted very well so we will apply a log transformation in order to improve the performance.   
par(mfrow=c(1,1))  
influencePlot(mod1)

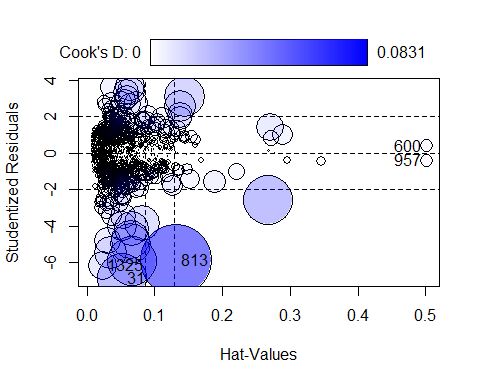


StudRes Hat CookD  
 186 5.5190153 0.13846887 0.0773117892  
 600 0.1235395 0.50085246 0.0002471785  
 804 6.8005397 0.03305736 0.0246922886  
 899 7.9489956 0.03593819 0.0363542665  
 957 -0.1235395 0.50085246 0.0002471785

mod2 <- lm(log(SalePrice) ~ log\_LotFrontage + log\_LotArea + X1stFlrSF + X2ndFlrSF + YearBuilt + GarageCars + BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd + Neighborhood + MSSubClass + log\_TotalBsmtSF + OverallCond + OverallQual + FireplaceQu + ScreenPorch + SaleCondition, data=df\_model)  
  
par(mfrow=c(2,2))  
plot(mod2)



par(mfrow=c(1,1))  
influencePlot(mod2)



StudRes Hat CookD  
 31 -6.8728339 0.05186097 0.040322419  
 600 0.4108596 0.50085246 0.002733617  
 813 -5.8864221 0.13213140 0.083063049  
 957 -0.4108596 0.50085246 0.002733617  
 1325 -6.1619063 0.02226965 0.013584999

Delete non significant   
summary(mod2)

Call:  
 lm(formula = log(SalePrice) ~ log\_LotFrontage + log\_LotArea +   
 X1stFlrSF + X2ndFlrSF + YearBuilt + GarageCars + BedroomAbvGr +   
 KitchenAbvGr + TotRmsAbvGrd + Neighborhood + MSSubClass +   
 log\_TotalBsmtSF + OverallCond + OverallQual + FireplaceQu +   
 ScreenPorch + SaleCondition, data = df\_model)  
   
 Residuals:  
 Min 1Q Median 3Q Max   
 -0.79217 -0.06052 0.00323 0.07072 0.42598   
   
 Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
 (Intercept) 2.923e+00 6.513e-01 4.487 7.81e-06 \*\*\*  
 log\_LotFrontage -1.031e-02 1.552e-02 -0.664 0.506670   
 log\_LotArea 8.987e-02 1.163e-02 7.727 2.10e-14 \*\*\*  
 X1stFlrSF 3.719e-04 1.872e-05 19.869 < 2e-16 \*\*\*  
 X2ndFlrSF 2.519e-04 2.236e-05 11.266 < 2e-16 \*\*\*  
 YearBuilt 3.562e-03 3.158e-04 11.279 < 2e-16 \*\*\*  
 GarageCars 4.825e-02 6.280e-03 7.683 2.92e-14 \*\*\*  
 BedroomAbvGr -2.296e-02 6.345e-03 -3.620 0.000306 \*\*\*  
 KitchenAbvGr -7.863e-02 2.629e-02 -2.991 0.002828 \*\*   
 TotRmsAbvGrd 6.451e-03 4.575e-03 1.410 0.158753   
 NeighborhoodBlueste 4.650e-02 9.396e-02 0.495 0.620767   
 NeighborhoodBrDale -1.002e-02 5.060e-02 -0.198 0.843093   
 NeighborhoodBrkSide 3.799e-02 4.131e-02 0.920 0.357957   
 NeighborhoodClearCr 8.601e-02 4.334e-02 1.985 0.047391 \*   
 NeighborhoodCollgCr 2.579e-02 3.574e-02 0.721 0.470747   
 NeighborhoodCrawfor 1.324e-01 4.029e-02 3.285 0.001046 \*\*   
 NeighborhoodEdwards -2.357e-02 3.851e-02 -0.612 0.540689   
 NeighborhoodGilbert -6.643e-03 3.764e-02 -0.177 0.859920   
 NeighborhoodIDOTRR -7.340e-02 4.332e-02 -1.694 0.090403 .   
 NeighborhoodMeadowV -6.194e-02 5.095e-02 -1.216 0.224257   
 NeighborhoodMitchel -1.002e-02 3.945e-02 -0.254 0.799538   
 NeighborhoodNAmes 2.541e-03 3.701e-02 0.069 0.945272   
 NeighborhoodNoRidge 8.630e-02 4.056e-02 2.128 0.033551 \*   
 NeighborhoodNPkVill 3.767e-02 5.271e-02 0.715 0.474950   
 NeighborhoodNridgHt 1.219e-01 3.545e-02 3.437 0.000606 \*\*\*  
 NeighborhoodNWAmes -4.600e-02 3.793e-02 -1.213 0.225387   
 NeighborhoodOldTown -3.463e-02 4.012e-02 -0.863 0.388160   
 NeighborhoodSawyer -1.656e-02 3.886e-02 -0.426 0.670101   
 NeighborhoodSawyerW 1.911e-03 3.784e-02 0.051 0.959731   
 NeighborhoodSomerst 9.232e-02 3.641e-02 2.535 0.011340 \*   
 NeighborhoodStoneBr 1.667e-01 4.011e-02 4.156 3.43e-05 \*\*\*  
 NeighborhoodSWISU 5.727e-02 4.552e-02 1.258 0.208581   
 NeighborhoodTimber 4.001e-02 4.034e-02 0.992 0.321361   
 NeighborhoodVeenker 8.046e-02 5.008e-02 1.607 0.108340   
 MSSubClass30 -6.282e-02 2.159e-02 -2.909 0.003680 \*\*   
 MSSubClass40 -3.029e-02 6.185e-02 -0.490 0.624433   
 MSSubClass45 -5.505e-02 3.802e-02 -1.448 0.147883   
 MSSubClass50 -8.295e-03 1.912e-02 -0.434 0.664510   
 MSSubClass60 9.427e-03 2.071e-02 0.455 0.649129   
 MSSubClass70 2.210e-02 2.786e-02 0.793 0.427831   
 MSSubClass75 3.780e-02 4.049e-02 0.934 0.350620   
 MSSubClass80 6.724e-03 1.755e-02 0.383 0.701721   
 MSSubClass85 3.114e-02 2.878e-02 1.082 0.279517   
 MSSubClass90 1.286e-03 2.800e-02 0.046 0.963361   
 MSSubClass120 -1.178e-02 2.021e-02 -0.583 0.560112   
 MSSubClass160 -5.658e-02 3.120e-02 -1.813 0.069993 .   
 MSSubClass180 -1.301e-02 5.102e-02 -0.255 0.798737   
 MSSubClass190 3.021e-02 3.085e-02 0.979 0.327701   
 log\_TotalBsmtSF 2.568e-02 3.343e-03 7.682 2.95e-14 \*\*\*  
 OverallCond 5.438e-02 3.386e-03 16.063 < 2e-16 \*\*\*  
 OverallQual 5.978e-02 4.533e-03 13.187 < 2e-16 \*\*\*  
 FireplaceQu2 -2.131e-02 3.514e-02 -0.606 0.544293   
 FireplaceQu3 -3.253e-02 2.752e-02 -1.182 0.237387   
 FireplaceQu4 -5.987e-02 3.936e-02 -1.521 0.128461   
 FireplaceQu5 -4.168e-02 2.836e-02 -1.470 0.141840   
 FireplaceQuNoFirePlace -6.565e-02 2.857e-02 -2.297 0.021744 \*   
 ScreenPorch 2.321e-04 6.091e-05 3.811 0.000145 \*\*\*  
 SaleConditionAbnorml -7.647e-02 1.305e-02 -5.859 5.81e-09 \*\*\*  
 SaleConditionAdjLand 4.085e-02 6.359e-02 0.642 0.520740   
 SaleConditionAlloca -4.150e-02 3.731e-02 -1.112 0.266223   
 SaleConditionFamily -5.917e-02 2.777e-02 -2.131 0.033288 \*   
 SaleConditionPartial 3.612e-02 1.358e-02 2.660 0.007898 \*\*   
 ---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   
 Residual standard error: 0.1203 on 1381 degrees of freedom  
 Multiple R-squared: 0.91, Adjusted R-squared: 0.906   
 F-statistic: 228.9 on 61 and 1381 DF, p-value: < 2.2e-16

mod3 <- lm(log(SalePrice) ~ log\_LotArea + X1stFlrSF + X2ndFlrSF + YearBuilt + GarageCars + BedroomAbvGr + KitchenAbvGr + Neighborhood + MSSubClass + log\_TotalBsmtSF + OverallCond + OverallQual + FireplaceQu + ScreenPorch + SaleCondition, data=df\_model)

## Interpretation of the model

smod <- summary(mod3)  
vif(mod2) %>% round(2) %>% t

log\_LotFrontage log\_LotArea X1stFlrSF X2ndFlrSF YearBuilt  
 GVIF^(1/(2\*Df)) 1.69 1.88 2.15 3.02 3.01  
 GarageCars BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Neighborhood  
 GVIF^(1/(2\*Df)) 1.47 1.63 1.81 2.32 1.12  
 MSSubClass log\_TotalBsmtSF OverallCond OverallQual FireplaceQu  
5.00  
 GVIF^(1/(2\*Df)) 1.29 1.15 1.19 1.95 1.11  
 ScreenPorch SaleCondition  
 GVIF^(1/(2\*Df)) 1.04 1.07

mod2$coefficients %>% round(5)

(Intercept) log\_LotFrontage log\_LotArea   
 2.92273 -0.01031 0.08987   
 X1stFlrSF X2ndFlrSF YearBuilt   
 0.00037 0.00025 0.00356   
 GarageCars BedroomAbvGr KitchenAbvGr   
 0.04825 -0.02296 -0.07863   
 TotRmsAbvGrd NeighborhoodBlueste NeighborhoodBrDale   
 0.00645 0.04650 -0.01002   
 NeighborhoodBrkSide NeighborhoodClearCr NeighborhoodCollgCr   
 0.03799 0.08601 0.02579   
 NeighborhoodCrawfor NeighborhoodEdwards NeighborhoodGilbert   
 0.13236 -0.02357 -0.00664   
 NeighborhoodIDOTRR NeighborhoodMeadowV NeighborhoodMitchel   
 -0.07340 -0.06194 -0.01002   
 NeighborhoodNAmes NeighborhoodNoRidge NeighborhoodNPkVill   
 0.00254 0.08630 0.03767   
 NeighborhoodNridgHt NeighborhoodNWAmes NeighborhoodOldTown   
 0.12185 -0.04600 -0.03463   
 NeighborhoodSawyer NeighborhoodSawyerW NeighborhoodSomerst   
 -0.01656 0.00191 0.09232   
 NeighborhoodStoneBr NeighborhoodSWISU NeighborhoodTimber   
 0.16671 0.05727 0.04001   
 NeighborhoodVeenker MSSubClass30 MSSubClass40   
 0.08046 -0.06282 -0.03029   
 MSSubClass45 MSSubClass50 MSSubClass60   
 -0.05505 -0.00830 0.00943   
 MSSubClass70 MSSubClass75 MSSubClass80   
 0.02210 0.03780 0.00672   
 MSSubClass85 MSSubClass90 MSSubClass120   
 0.03114 0.00129 -0.01178   
 MSSubClass160 MSSubClass180 MSSubClass190   
 -0.05658 -0.01301 0.03021   
 log\_TotalBsmtSF OverallCond OverallQual   
 0.02568 0.05438 0.05978   
 FireplaceQu2 FireplaceQu3 FireplaceQu4   
 -0.02131 -0.03253 -0.05987   
 FireplaceQu5 FireplaceQuNoFirePlace ScreenPorch   
 -0.04168 -0.06565 0.00023   
 SaleConditionAbnorml SaleConditionAdjLand SaleConditionAlloca   
 -0.07647 0.04085 -0.04150   
 SaleConditionFamily SaleConditionPartial   
 -0.05917 0.03612

r2 <- smod$r.squared %>% round(2)  
adj.r2 <- smod$adj.r.squared %>% round(2)  
smod

Call:  
 lm(formula = log(SalePrice) ~ log\_LotArea + X1stFlrSF + X2ndFlrSF +   
 YearBuilt + GarageCars + BedroomAbvGr + KitchenAbvGr + Neighborhood +   
 MSSubClass + log\_TotalBsmtSF + OverallCond + OverallQual +   
 FireplaceQu + ScreenPorch + SaleCondition, data = df\_model)  
   
 Residuals:  
 Min 1Q Median 3Q Max   
 -0.79380 -0.06143 0.00311 0.06767 0.43180   
   
 Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
 (Intercept) 2.901e+00 6.491e-01 4.469 8.51e-06 \*\*\*  
 log\_LotArea 8.742e-02 1.119e-02 7.813 1.09e-14 \*\*\*  
 X1stFlrSF 3.836e-04 1.632e-05 23.503 < 2e-16 \*\*\*  
 X2ndFlrSF 2.643e-04 2.058e-05 12.846 < 2e-16 \*\*\*  
 YearBuilt 3.566e-03 3.157e-04 11.297 < 2e-16 \*\*\*  
 GarageCars 4.827e-02 6.269e-03 7.700 2.59e-14 \*\*\*  
 BedroomAbvGr -1.932e-02 5.712e-03 -3.382 0.000739 \*\*\*  
 KitchenAbvGr -7.077e-02 2.558e-02 -2.766 0.005743 \*\*   
 NeighborhoodBlueste 4.279e-02 9.383e-02 0.456 0.648405   
 NeighborhoodBrDale -7.762e-03 5.033e-02 -0.154 0.877470   
 NeighborhoodBrkSide 3.044e-02 4.073e-02 0.747 0.455000   
 NeighborhoodClearCr 7.992e-02 4.292e-02 1.862 0.062779 .   
 NeighborhoodCollgCr 1.990e-02 3.520e-02 0.565 0.571928   
 NeighborhoodCrawfor 1.247e-01 3.988e-02 3.126 0.001808 \*\*   
 NeighborhoodEdwards -2.937e-02 3.812e-02 -0.771 0.441084   
 NeighborhoodGilbert -1.159e-02 3.725e-02 -0.311 0.755736   
 NeighborhoodIDOTRR -7.835e-02 4.305e-02 -1.820 0.069010 .   
 NeighborhoodMeadowV -6.604e-02 5.075e-02 -1.301 0.193349   
 NeighborhoodMitchel -1.664e-02 3.886e-02 -0.428 0.668569   
 NeighborhoodNAmes -4.204e-03 3.659e-02 -0.115 0.908533   
 NeighborhoodNoRidge 7.674e-02 3.984e-02 1.926 0.054287 .   
 NeighborhoodNPkVill 3.715e-02 5.259e-02 0.706 0.480041   
 NeighborhoodNridgHt 1.179e-01 3.532e-02 3.337 0.000868 \*\*\*  
 NeighborhoodNWAmes -5.208e-02 3.757e-02 -1.386 0.165870   
 NeighborhoodOldTown -4.147e-02 3.957e-02 -1.048 0.294877   
 NeighborhoodSawyer -2.262e-02 3.846e-02 -0.588 0.556533   
 NeighborhoodSawyerW -4.695e-03 3.743e-02 -0.125 0.900188   
 NeighborhoodSomerst 8.629e-02 3.594e-02 2.401 0.016486 \*   
 NeighborhoodStoneBr 1.629e-01 3.978e-02 4.095 4.46e-05 \*\*\*  
 NeighborhoodSWISU 5.045e-02 4.500e-02 1.121 0.262398   
 NeighborhoodTimber 3.534e-02 3.986e-02 0.886 0.375503   
 NeighborhoodVeenker 7.544e-02 4.972e-02 1.517 0.129416   
 MSSubClass30 -6.025e-02 2.151e-02 -2.801 0.005169 \*\*   
 MSSubClass40 -3.085e-02 6.171e-02 -0.500 0.617237   
 MSSubClass45 -5.244e-02 3.794e-02 -1.382 0.167110   
 MSSubClass50 -5.978e-03 1.905e-02 -0.314 0.753715   
 MSSubClass60 1.064e-02 2.068e-02 0.514 0.606988   
 MSSubClass70 2.584e-02 2.775e-02 0.931 0.351837   
 MSSubClass75 4.269e-02 4.033e-02 1.059 0.290006   
 MSSubClass80 6.976e-03 1.754e-02 0.398 0.690846   
 MSSubClass85 3.352e-02 2.875e-02 1.166 0.243784   
 MSSubClass90 6.316e-04 2.793e-02 0.023 0.981959   
 MSSubClass120 -1.129e-02 1.907e-02 -0.592 0.554120   
 MSSubClass160 -5.545e-02 2.976e-02 -1.863 0.062611 .   
 MSSubClass180 -1.014e-02 4.999e-02 -0.203 0.839197   
 MSSubClass190 3.062e-02 3.078e-02 0.995 0.320091   
 log\_TotalBsmtSF 2.562e-02 3.342e-03 7.666 3.32e-14 \*\*\*  
 OverallCond 5.460e-02 3.379e-03 16.159 < 2e-16 \*\*\*  
 OverallQual 5.988e-02 4.529e-03 13.220 < 2e-16 \*\*\*  
 FireplaceQu2 -2.221e-02 3.512e-02 -0.632 0.527300   
 FireplaceQu3 -3.089e-02 2.750e-02 -1.123 0.261430   
 FireplaceQu4 -5.940e-02 3.933e-02 -1.510 0.131183   
 FireplaceQu5 -4.109e-02 2.835e-02 -1.449 0.147488   
 FireplaceQuNoFirePlace -6.477e-02 2.856e-02 -2.268 0.023488 \*   
 ScreenPorch 2.320e-04 6.090e-05 3.810 0.000145 \*\*\*  
 SaleConditionAbnorml -7.665e-02 1.305e-02 -5.873 5.36e-09 \*\*\*  
 SaleConditionAdjLand 4.427e-02 6.355e-02 0.697 0.486189   
 SaleConditionAlloca -4.091e-02 3.732e-02 -1.096 0.273143   
 SaleConditionFamily -5.834e-02 2.776e-02 -2.101 0.035798 \*   
 SaleConditionPartial 3.682e-02 1.356e-02 2.717 0.006679 \*\*   
 ---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   
 Residual standard error: 0.1204 on 1383 degrees of freedom  
 Multiple R-squared: 0.9098, Adjusted R-squared: 0.906   
 F-statistic: 236.5 on 59 and 1383 DF, p-value: < 2.2e-16

As we can see in the summary of model we selected all variables that are significant in order to predict the Sale Price. We have an overall R-squared of 0.91 and an Adjusted R-squared of 0.91 and the p-value of the model is << 0.05, so we can say it is significant and the variables can predict in some way the response variable.

Taking a deeper look at each coefficient we can conclude that Lot Area is positively correlated to Price, as well as 1st floor Squared Feet meters and 2nd floor Squared feet meters. Also Year Built is positively correlated with Price, so newer houses are more expensive. The number of cars that the garage can fit is positively correlated, also the bedroom above gradem, and kitchen above grade are positively correlated with Price.

As for the neighorhoods the reference level is Blmngtn that as we can see in the “Interaction between numerical and categorical” section has a mild average and distribution compared to the others. We can see this also in the model as the significant neighorhoods are IDOTRR with an expected lower price over Blmngtn and NridgHt, Somerst, Crawfor and StoneBr with an expected higher price over Blmngtn,

As for MSSubClass the reference level is the level 20 - “1-STORY 1946 & NEWER ALL STYLES”. We have little significant factors with this reference level, we will expect lowe prices for the level 30 - 1-STORY 1945 & OLDER and level 160 - “2-STORY PUD - 1946 & NEWER”.

Moreover, we have negative correlation in the No Fireplaces comparing it with the reference level 1 fireplace. The rest of the levels are not significant compared with the level 1.

Overall Condition and Overall Quality of the house is positively correlated with Price as well as the number of Squared Feet of the basement.

The Screen Porch area is significant and is positively correlated with Price.

The Sale condition reduces the price with the reference level “Normal” when comparing with Abnormal and Family, in contrast increases the price when the Sale Condition is partial.

## Test Preprocessing

#df\_test <- read.csv("C:/Users/inigo/Documents/UPC/SIM/SIM-project/data/test.csv")  
df\_test <- read.csv("C:/MDS/SIM/Project/data/test.csv")  
  
numeric\_columns <- c("LotFrontage", "LotArea", "YearBuilt", "YearRemodAdd", "MasVnrArea", "BsmtFinSF1", "BsmtFinSF2", "BsmtUnfSF", "TotalBsmtSF", "X1stFlrSF", "X2ndFlrSF", "LowQualFinSF", "GrLivArea", "BsmtFullBath", "BsmtHalfBath", "FullBath", "HalfBath", "BedroomAbvGr", "KitchenAbvGr", "TotRmsAbvGrd", "Fireplaces", "GarageYrBlt", "GarageCars", "GarageArea", "WoodDeckSF", "OpenPorchSF", "EnclosedPorch", "X3SsnPorch", "ScreenPorch", "PoolArea", "MiscVal", "MoSold", "YrSold", "OverallCond","OverallQual")  
  
cat\_keep <- c("FireplaceQu","KitchenQual","BsmtFinType1","BsmtExposure","BsmtQual","Foundation","Neighborhood","LotShape","MSSubClass","Exterior1st","Exterior2nd", "SaleCondition")  
  
df1\_test <- df\_test %>% select(all\_of(numeric\_columns), all\_of(cat\_keep))  
  
df1\_test[cat\_keep] <- lapply(df1\_test[cat\_keep], as.factor) Create Factors  
df1\_test[numeric\_columns] <- lapply(df1\_test[numeric\_columns], as.numeric)   
  
num\_outliers <- c("BsmtFinSF1", "BsmtFinSF2", "BsmtHalfBath", "BsmtUnfSF", "EnclosedPorch", "GarageArea", "GrLivArea","LotArea", "LotFrontage", "LowQualFinSF", "MasVnrArea","MiscVal","OpenPorchSF","PoolArea","ScreenPorch","TotalBsmtSF", "WoodDeckSF","X1stFlrSF","X3SsnPorch")  
  
  
 Impute missing data by creating extra modality  
  
df1\_test$FireplaceQu <- ifelse(df1\_test$FireplaceQu %>% is.na, "NoFirePlace", df1\_test$FireplaceQu)  
df1\_test$BsmtExposure <- ifelse(df1\_test$BsmtExposure %>% is.na, "NoBasement", df1\_test$BsmtExposure)  
df1\_test$BsmtFinType1 <- ifelse(df1\_test$BsmtFinType1 %>% is.na, "NoBasement", df1\_test$BsmtFinType1)  
df1\_test$BsmtQual <- ifelse(df1\_test$BsmtQual %>% is.na, "NoBasement", df1\_test$BsmtQual)  
  
df1\_test$garage <- ifelse(df1\_test$GarageYrBlt %>% is.na, "NO", "YES")  
 Impute with synthetic values  
  
df2\_test <- mice(df1\_test, method = "cart", m = 1)

df3\_test <- complete(df2\_test)  
  
df3\_test$log\_LotArea <- log(df3\_test$LotArea + 1)  
df3\_test$log\_LotFrontage <- log(df3\_test$LotFrontage + 1)  
df3\_test$log\_TotalBsmtSF <- log(df3\_test$TotalBsmtSF + 1)  
  
df3\_test$YearRemodAdd <- df3\_test$YearRemodAdd - df3\_test$YearBuilt  
  
df3\_test$bsmt <- ifelse(df3\_test$log\_TotalBsmtSF == 0, "NO","YES")  
  
df3\_test$cat\_BsmntFinSF2 <- ifelse(df3\_test$BsmtFinSF2 > 0, "Yes","Zero")  
df3\_test$cat\_EnclosedPorch <- ifelse(df3\_test$EnclosedPorch > 0, "Yes","Zero")  
df3\_test$cat\_LowQualFinSF <- ifelse(df3\_test$LowQualFinSF > 0, "Yes","Zero")  
df3\_test$cat\_MiscVal <- ifelse(df3\_test$MiscVal > 0, "Yes","Zero")  
df3\_test$cat\_PoolArea <- ifelse(df3\_test$PoolArea > 0, "Yes","No")  
df3\_test$cat\_ScreenPorch <- ifelse(df3\_test$ScreenPorch > 0, "Yes","No")  
df3\_test$cat\_X3SsnPorch <- ifelse(df3\_test$X3SsnPorch > 0, "Yes","No")  
df3\_test$remod <- ifelse(df3\_test$YearRemodAdd == 0, "No", "Yes")

# Predictions of the test data

df\_model\_test <- df3\_test %>% select(  
c("log\_LotFrontage","log\_LotArea","YearBuilt","YearRemodAdd","FullBath","garage","GarageArea","bsmt","log\_TotalBsmtSF","cat\_PoolArea","YrSold","MoSold","remod","GarageCars","GarageYrBlt","BedroomAbvGr","KitchenAbvGr","TotRmsAbvGrd","Neighborhood","GrLivArea","MSSubClass","X1stFlrSF","X2ndFlrSF","OverallCond","OverallQual","FireplaceQu","cat\_MiscVal","OpenPorchSF", "cat\_EnclosedPorch","cat\_LowQualFinSF","cat\_X3SsnPorch","ScreenPorch","Foundation","Exterior1st","Exterior2nd","SaleCondition")  
)  
attach(df\_model\_test)  
  
df\_model\_test <- df\_model\_test[df\_model\_test$MSSubClass %in% (df\_model$MSSubClass %>% unique), ]  
df\_model\_test <- df\_model\_test[df\_model\_test$Neighborhood %in% (df\_model$Neighborhood %>% unique), ]  
  
  
predictions <- predict(mod2, newdata = df\_model\_test, type = "response")  
  
final\_predictions <- exp(predictions)  
  
  
 Assessing forecasting capabilities  
smp <- sample(x=nrow(df\_model),nrow(df\_model)\*0.75, replace=FALSE)  
  
train <- df\_model[smp,]  
test <- df\_model[-smp,]  
  
test <- test[test$MSSubClass %in% (train$MSSubClass %>% unique), ]  
test <- test[test$Neighborhood %in% (train$Neighborhood %>% unique), ]  
  
  
mod2 <- lm(log(SalePrice) ~ log\_LotFrontage + log\_LotArea + X1stFlrSF + X2ndFlrSF + YearBuilt + GarageCars + BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd + Neighborhood + MSSubClass + log\_TotalBsmtSF + OverallCond + OverallQual + FireplaceQu + ScreenPorch, data=train)  
  
predictions <- predict(mod2, test)  
predictions <- exp(predictions)  
  
deviation <- (test$SalePrice - predictions)/test$SalePrice  
deviation\_abs <- (abs(test$SalePrice - predictions)/test$SalePrice)  
mean(deviation\_abs)

[1] 0.0941963

The median deviation is 9.4%   
  
g1 <- ggplot() + geom\_jitter(mapping = aes(x=test$SalePrice, y=deviation, colour=deviation))  
g2 <- ggplot() + geom\_histogram(mapping = aes(x=deviation), fill="white", colour="black")  
g3 <- ggplot() + geom\_boxplot(mapping = aes(y=deviation), fill="cyan", colour="black")  
  
  
  
# Create a data frame for the overlay  
overlay\_data <- data.frame(  
 value = c(predictions, test$SalePrice),  
 group = rep(c("Predictions", "Actual"),   
 each = c(length(predictions), length(test$SalePrice))  
 )  
)

Warning in rep(c("Predictions", "Actual"), each = c(length(predictions), :  
 first element used of 'each' argument

# Create the histogram with ggplot2  
g4 <- ggplot(overlay\_data, aes(x = value, fill = group)) +  
 geom\_histogram(position = "identity", alpha = 0.5, bins = 30) +  
 labs(title = "Histogram Overlay of Predictions and Actual SalePrice",  
 x = "Sale Price",  
 y = "Frequency") +  
 scale\_fill\_manual(values = c("blue", "red"))   
 The distributions look very similar  
  
  
grid.arrange(g1,g2,g3,g4)

