House Prices:

Data Cleaning, Data Visualization, Data Preprocessing, Dimensionality Reduction & Feature Engineering

Aymen Rumi

```r
# Importing data

housing <- read.csv('train.csv', sep=',')

housing[c("Id")]<-NULL
attach(housing)
```

Data Investigation

```r
# Data dimensions

dim(housing)
```

```
[1] 1460    80
```

```r
# Data types

types<-as.data.frame(sapply(housing,class))
types%>%group_by(sapply(housing, class))%>%summarise(count=n())
```

```
`summarise()` ungrouping output (override with `.groups` argument)

# A tibble: 2 x 2
  `sapply(housing, class)` count
  <chr>                    <int>
1 factor                      43
2 integer                     37
```

```r
# missing data amount & percentages

missing<-housing%>%is.na()%>%colSums()
missing_percent<-housing%>%is.na()%>%colSums()/dim(housing)[1]


# vVariables with missing data

missing_percent[missing>0]
```

1

```
 LotFrontage         Alley    MasVnrType    MasVnrArea      BsmtQual      BsmtCond
0.1773972603 0.9376712329 0.0054794521 0.0054794521 0.0253424658 0.0253424658
BsmtExposure BsmtFinType1 BsmtFinType2    Electrical   FireplaceQu    GarageType
0.0260273973 0.0253424658 0.0260273973 0.0006849315 0.4726027397 0.0554794521
  GarageYrBlt GarageFinish    GarageQual    GarageCond        PoolQC          Fence
0.0554794521 0.0554794521 0.0554794521 0.0554794521 0.9952054795 0.8075342466
  MiscFeature
0.9630136986
```

```
length(missing_percent[missing>0])
```

```
[1] 19
```

```
# variables with missing data above 45%
```

```
missing_percent[missing_percent>0 & missing_percent>0.45]
```

```
      Alley FireplaceQu        PoolQC         Fence MiscFeature
  0.9376712    0.4726027    0.9952055    0.8075342    0.9630137
```

```
length(missing_percent[missing_percent>0 & missing_percent>0.45])
```

```
[1] 5
```

Observations

Our dataframe contains 80 variables, thus 1 response variable & 79 predictor variable; Of the 79 predictor variables, 43 are categorical & 36 are numerical; Of these 79 variables 19 contain missing values of which 5 have more than 45% missing

Data Cleaning

We will visualize and handle missing data; we will vizualize missing data patterns, identify sources for missing data, fill them as either factors levels or values through imputations.

Visualizing Missing Data
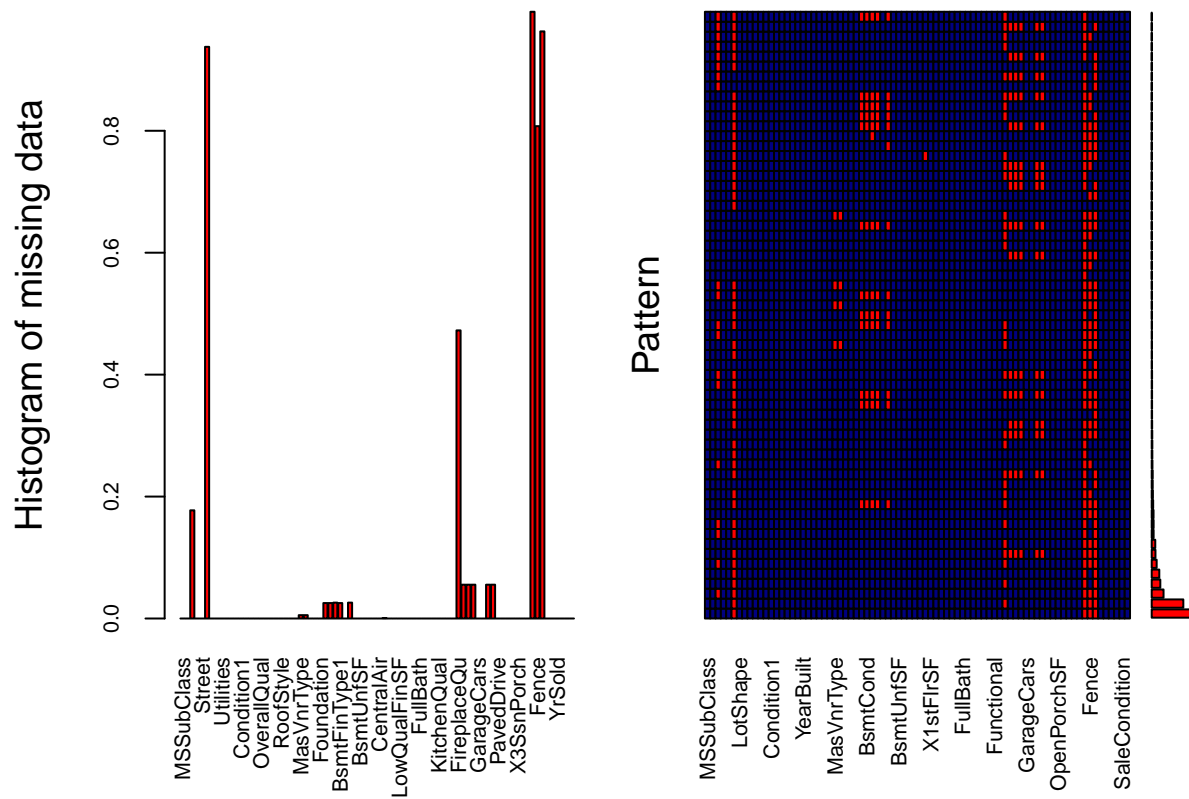
```
# Visualizing missing data
```

```
aggr(housing, col=c('navyblue','red'),
     numbers=TRUE, labels=names(housing), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Patte
```

```
Warning in plot.aggr(res, ...): not enough vertical space to display frequencies
(too many combinations)
```
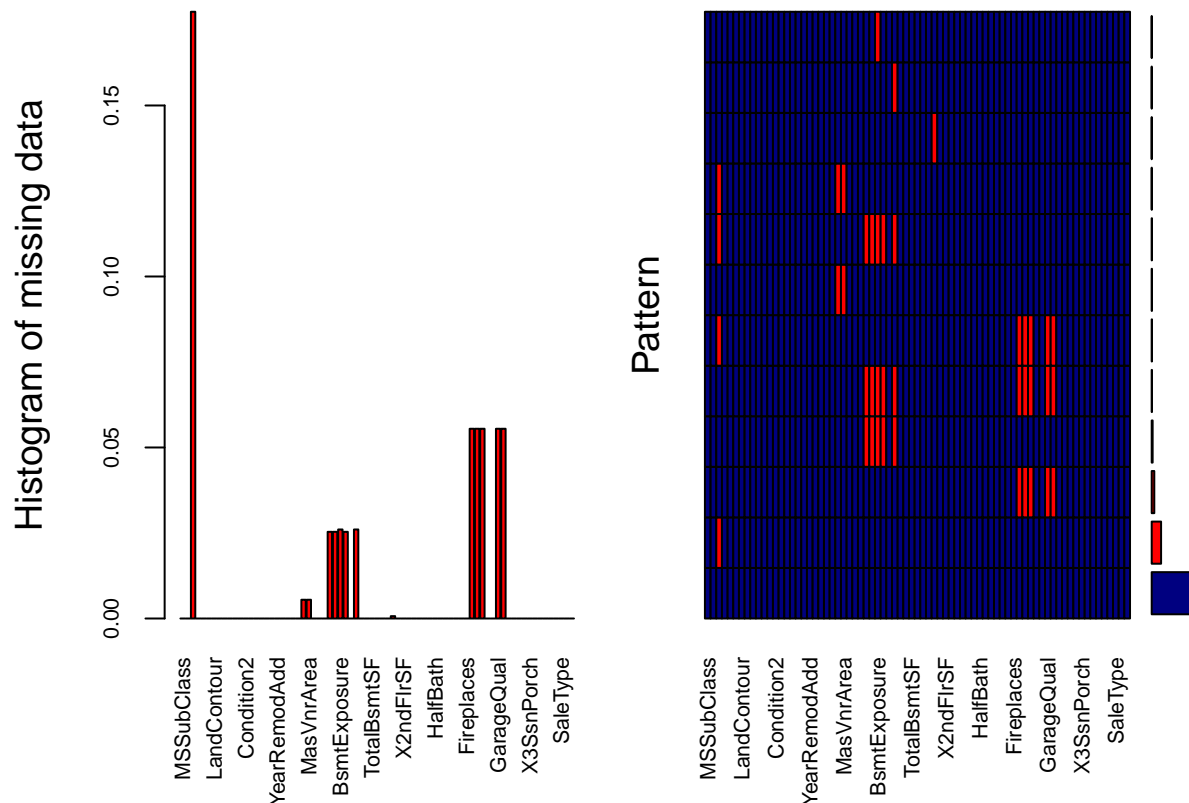
```
# Removing rows with more than 45% missing data

housing[c("Alley","FireplaceQu","PoolQC","Fence","MiscFeature")]<-NULL

aggr(housing, col=c('navyblue','red'),
     numbers=TRUE, labels=names(housing), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Patter
```

```
Warning in plot.aggr(res, ...): not enough horizontal space to display
frequencies
```

3

Many of the data seem to be missing in patterns, this can indicate that the data is not missing but part of a factor level that was not accounted for seperately

Replacing NA with "None"

```r
# Adding levels to NA that should be category "None"

levels(housing$GarageType)<-c(levels(housing$GarageType),"None")
levels(housing$GarageFinish)<-c(levels(housing$GarageFinish),"None")
levels(housing$GarageQual)<-c(levels(housing$GarageQual),"None")
levels(housing$GarageCond)<-c(levels(housing$GarageCond),"None")


levels(housing$BsmtQual)<-c(levels(housing$BsmtQual),"None")
levels(housing$BsmtCond)<-c(levels(housing$BsmtCond),"None")
levels(housing$BsmtExposure)<-c(levels(housing$BsmtExposure),"None")
levels(housing$BsmtFinType1)<-c(levels(housing$BsmtFinType1),"None")
levels(housing$BsmtFinType2)<-c(levels(housing$BsmtFinType2),"None")

# Filling NA values with "None"
```

```r
aggr(housing, col=c('navyblue','red'),
     numbers=TRUE, labels=names(housing), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Patter
```

```
Warning in plot.aggr(res, ...): not enough horizontal space to display
frequencies
```
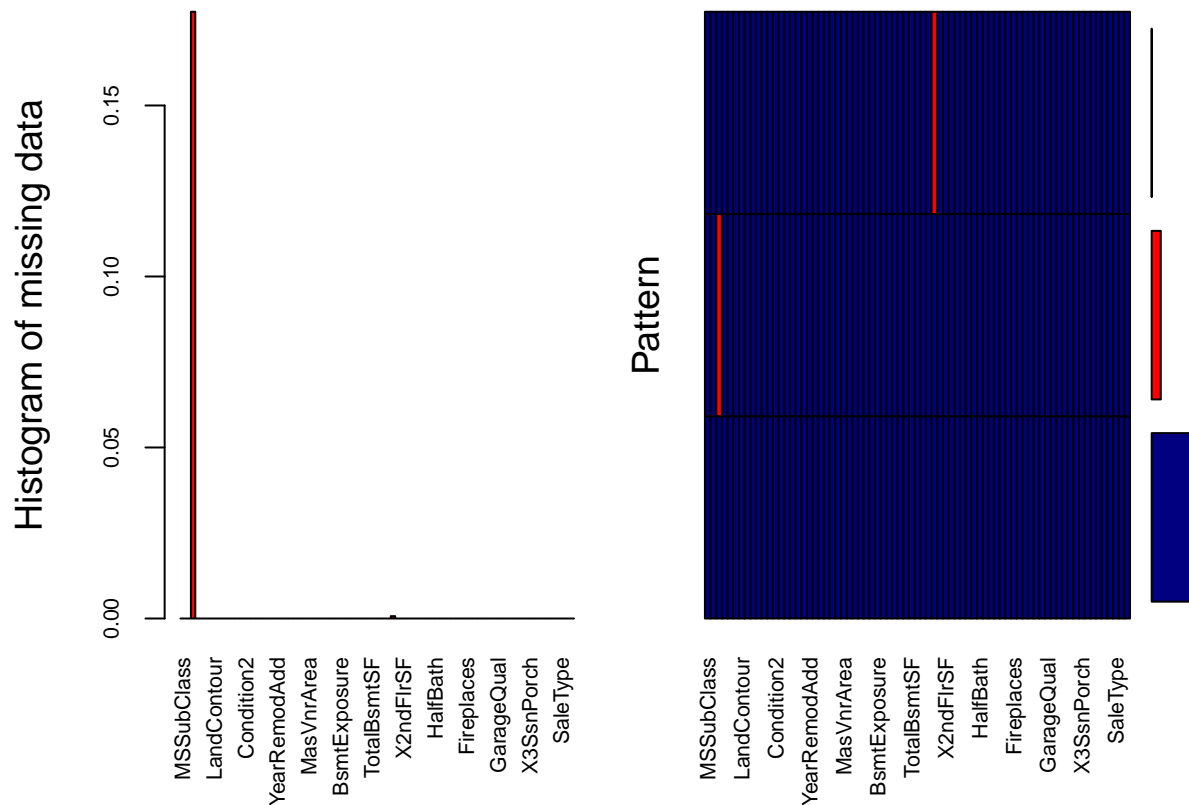
```
missing<-housing%>%is.na()%>%colSums()
missing[missing>0]
```

```
LotFrontage   Electrical
        259            1
```

Data Imputation

The remainder of the missing values need to be imputed, we have a numerical variable in which we will use other numerical variables to perform a mutliple linear regression model based imutation, the other categorical variable will be imputed with hot deck imputation

```
# Hot Deck Imputation

housing<-hotdeck(housing,variable=c("Electrical"))
housing$Electrical_imp<-NULL
```

```
# Model Based Multiple Regression Imputation

regression_imputation<-housing[c("LotFrontage","LotArea","X1stFlrSF","X2ndFlrSF","OverallCond","OverallC
regression_imputation<-regression_imputation%>%mutate(imputed=is.na(LotFrontage))

ggplot(data=regression_imputation,aes(x=LotFrontage,y=LotArea))+geom_point(colour="black")
```

```
Warning: Removed 259 rows containing missing values (geom_point).
```
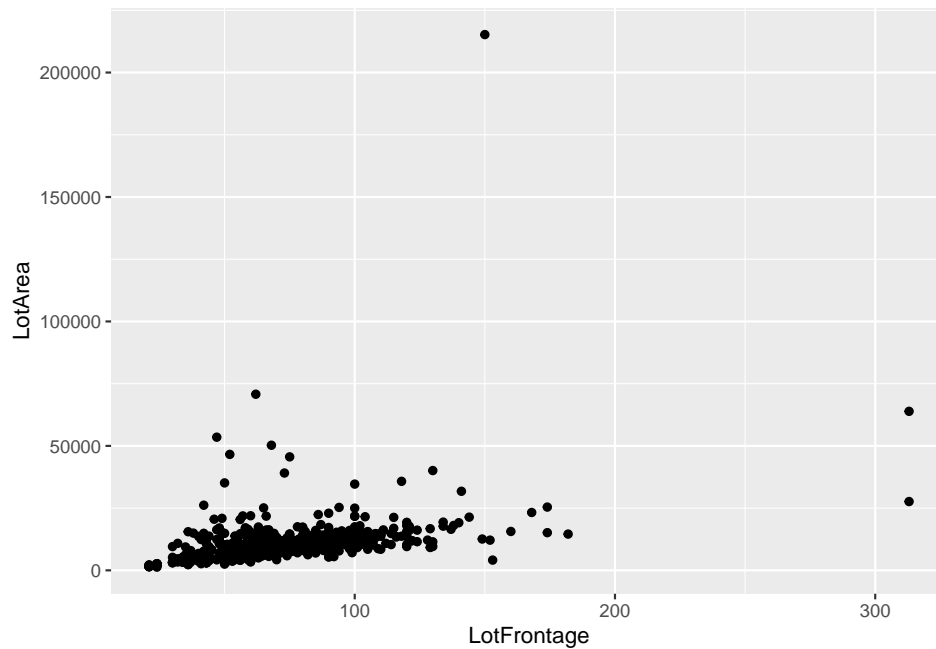
```
regression_imputation<-impute_lm(regression_imputation,LotFrontage~LotArea+X1stFlrSF+X2ndFlrSF+OverallC
```

Evaluating Imputation

We will asses the quality of our imputed data with a scatterplot

```
# Evaluating our imputation
ggplot(data=regression_imputation,aes(x=LotFrontage,y=LotArea,colour=imputed))+geom_point()
```

6
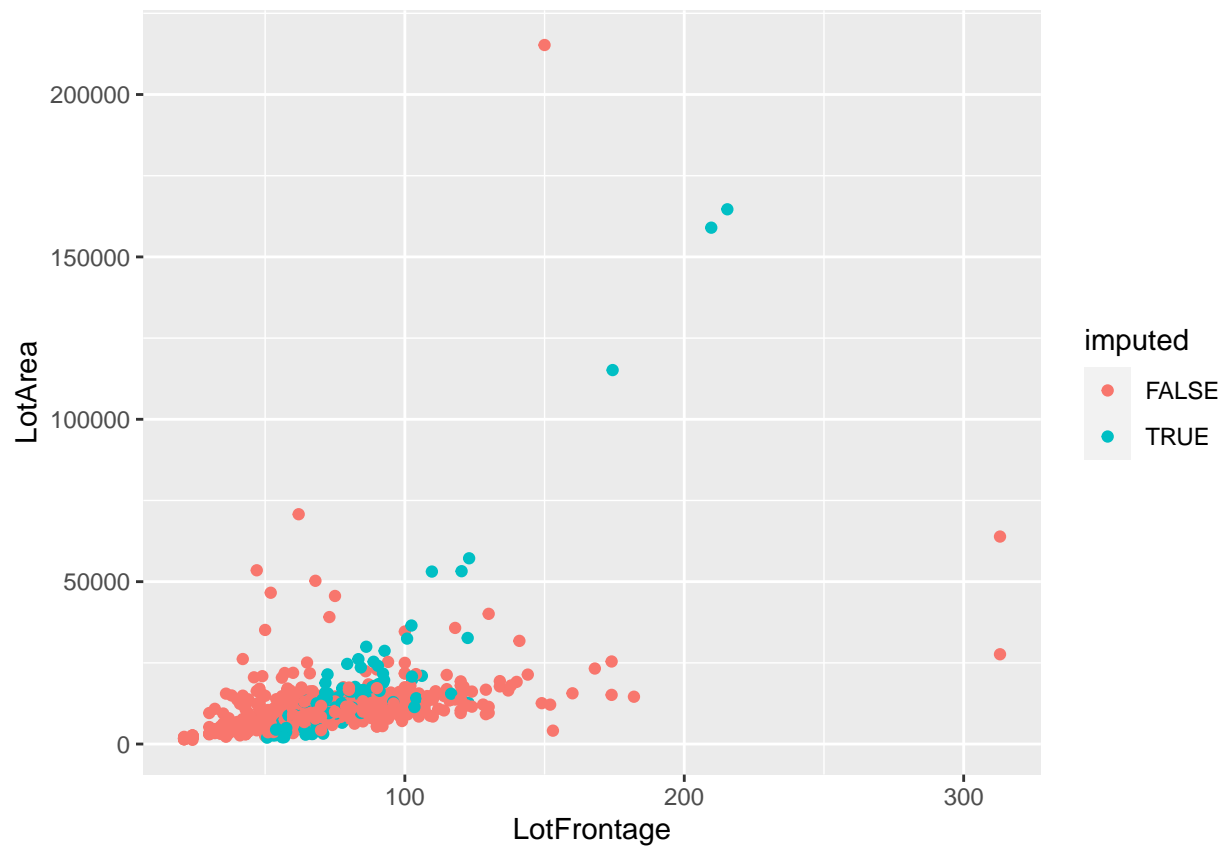
```
housing$LotFrontage<-regression_imputation$LotFrontage
```

Final Result

```
aggr(housing, col=c('navyblue','red'),
     numbers=TRUE, labels=names(housing), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Patter
```
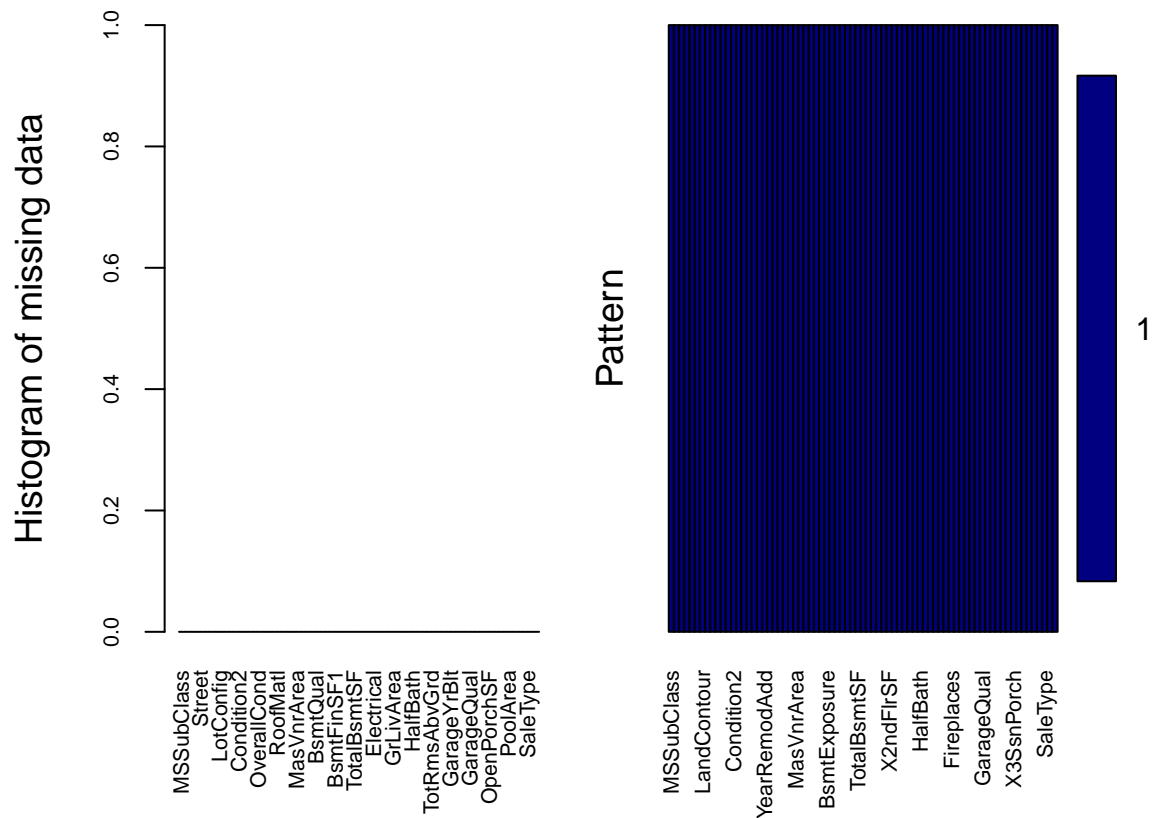
We have cleaned our dataset and it is now ready for analysis, we will start with data vizualizations to understand the nature of our data

Data Visualization

Since we have many categorical variables, we will group our response variable by categorical variable groups to identify & visualize particular patterns and difference in distribution shapes according to these variables

```r
# Distribution of Sales Price

ggplot(housing,aes(x=SalePrice))+geom_density(color="darkblue", fill="lightblue")
```
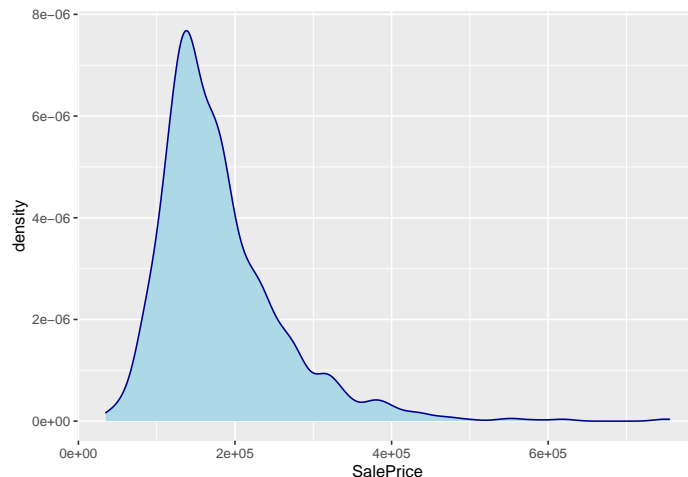
```r
# Distribution of Sales Price for GarageTypes
ggplot(housing, aes(x=SalePrice,fill=GarageType)) +   geom_density()
```
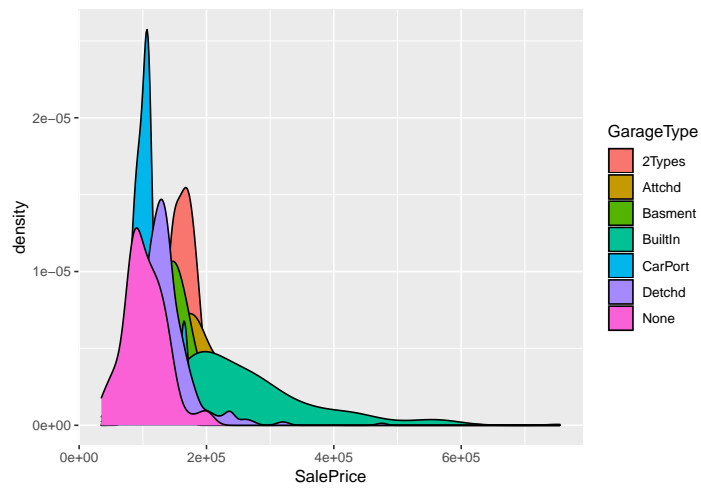


```r
# Distribution of Sales Price for Foundations
ggplot(housing, aes(x=SalePrice,fill=Foundation)) +   geom_density()
```
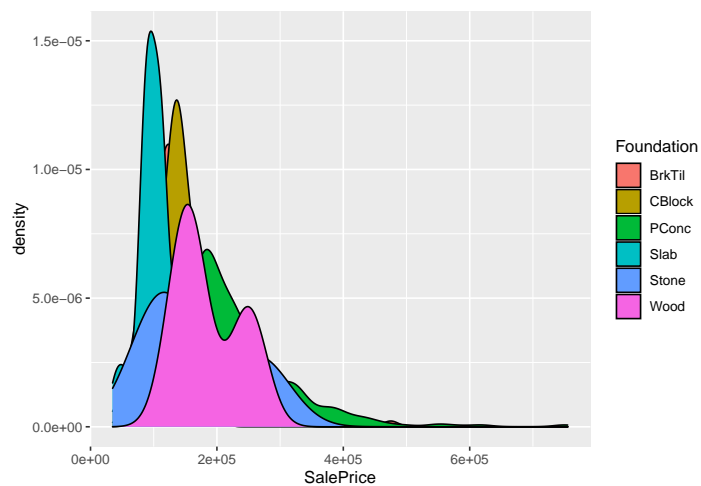


```r
# Distribution of Sales Price for BsmtQuals
ggplot(housing, aes(x=SalePrice,fill=BsmtQual)) +   geom_density()
```
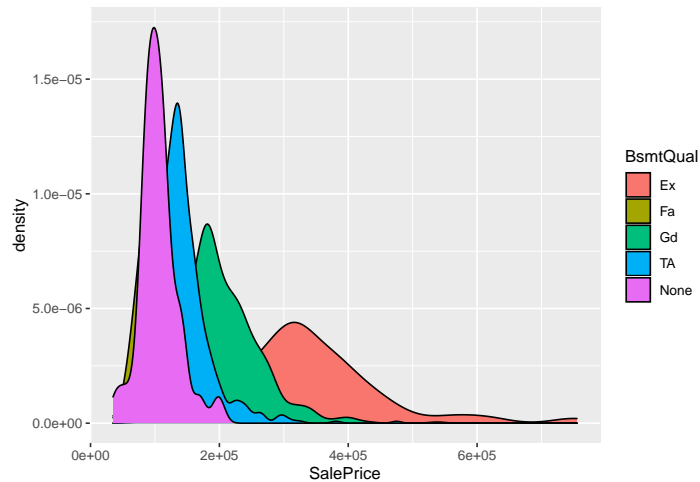
```r
# Distribution of Sales Price for SaleTypes
ggplot(housing, aes(x=SalePrice,fill=SaleType)) +   geom_density()
```
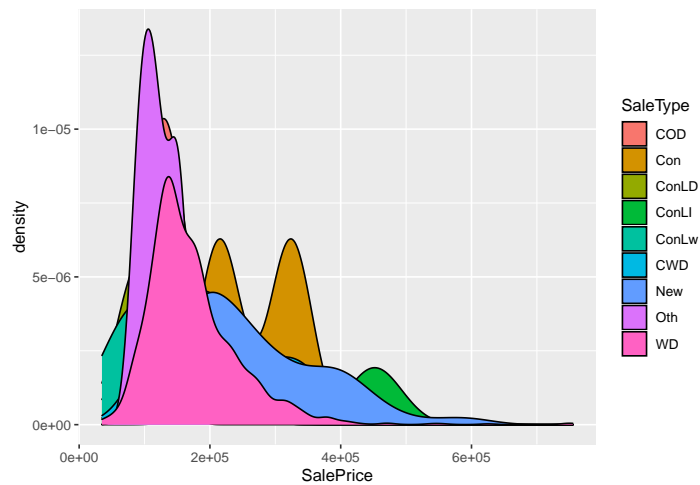

We will now look at patterns and trends in Sales Price with respect to numerical predictors, we will group colors by other numerical or categorial variables to identify & visualize further patterns

```r
# LotFrontage vs SalePrice, colored by Overall Quality

ggplot(data=housing, aes(x = LotFrontage,y=SalePrice,colour=OverallQual)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
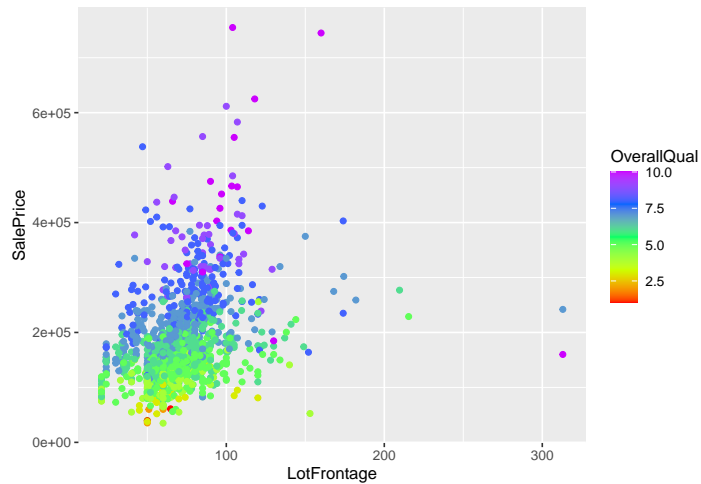
```r
# LotArea vs SalePrice, colored by Overall Quality

ggplot(data=housing, aes(x = LotArea,y=SalePrice,colour=OverallQual)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
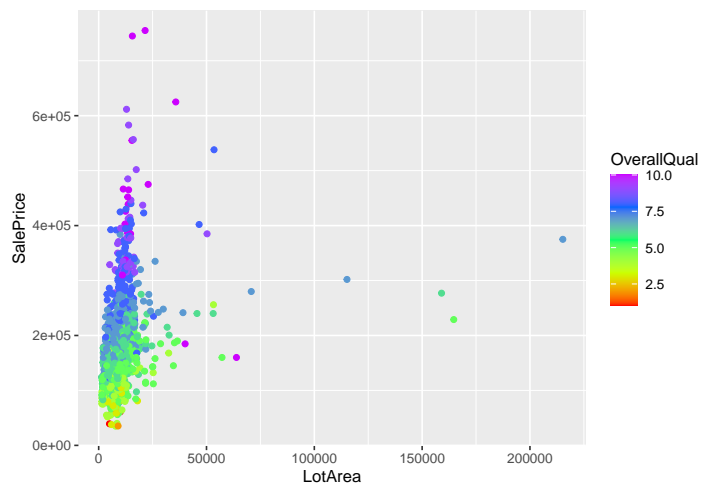


```r
# OverallQual vs SalePrice, colored by Year Built

ggplot(data=housing, aes(x = OverallQual,y=SalePrice,colour=YearBuilt)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
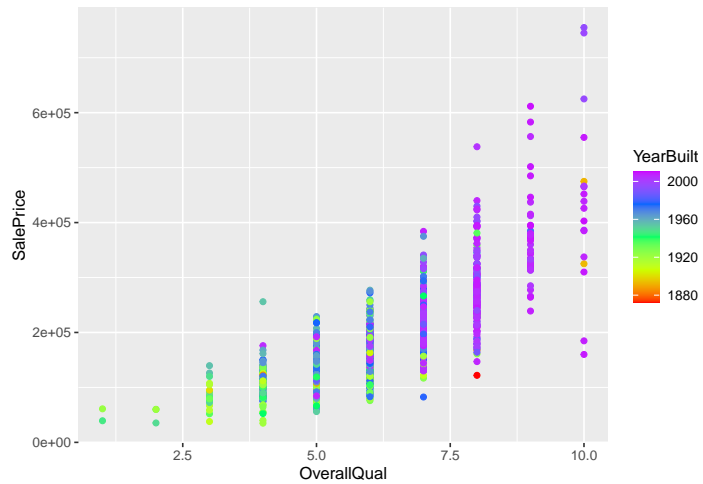
```
# OverallCond vs SalePrice, colored by Year Built

ggplot(data=housing, aes(x = OverallCond,y=SalePrice,colour=YearBuilt)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
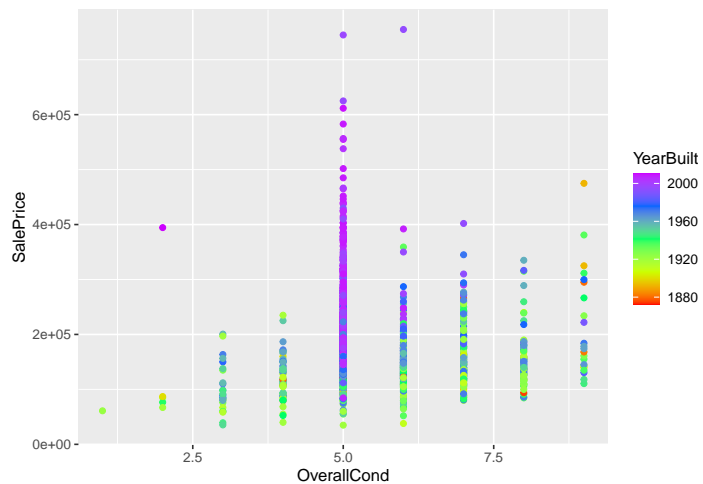


```
# Year Built vs SalePrice, colored by Overall Quality

ggplot(data=housing, aes(x = YearBuilt,y=SalePrice,colour=OverallQual)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
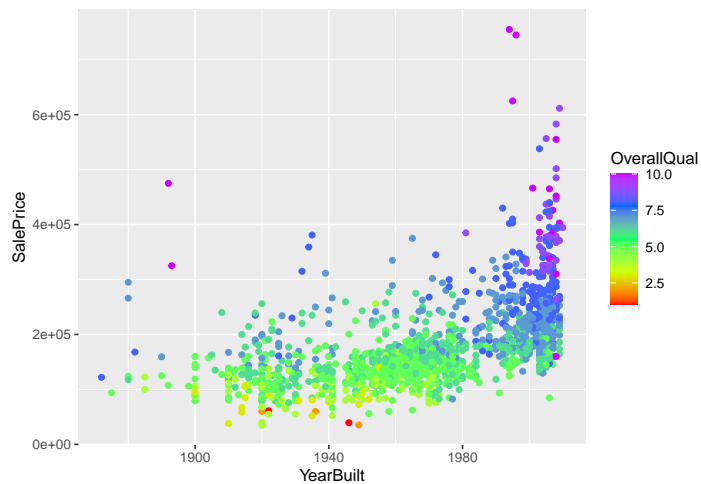
```
# X1stFlrSF vs SalePrice, colored by OverallQual

ggplot(data=housing, aes(x = X1stFlrSF,y=SalePrice,colour=OverallQual)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
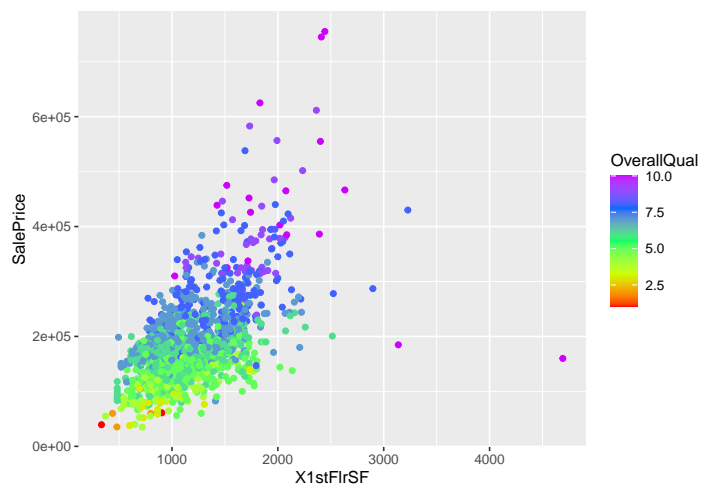


```
# X2ndFlrSF vs SalePrice, colored by Year Built

ggplot(data=housing, aes(x = X2ndFlrSF,y=SalePrice,color=YearBuilt)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
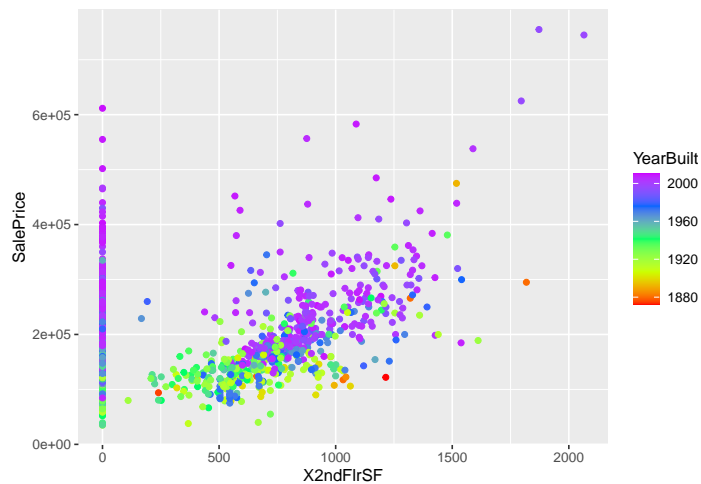
```
# BsmtFinSF1 vs SalePrice, colored by Overall Quality

ggplot(data=housing, aes(x = BsmtFinSF1,y=SalePrice,colour=OverallQual)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
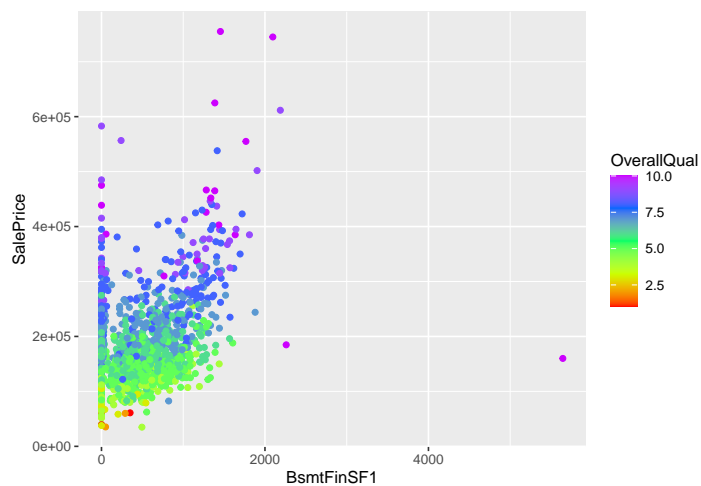


```
# TotalBsmtSF vs SalePrice, colored by Overall Quality

ggplot(data=housing, aes(x = TotalBsmtSF,y=SalePrice,colour=OverallQual)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
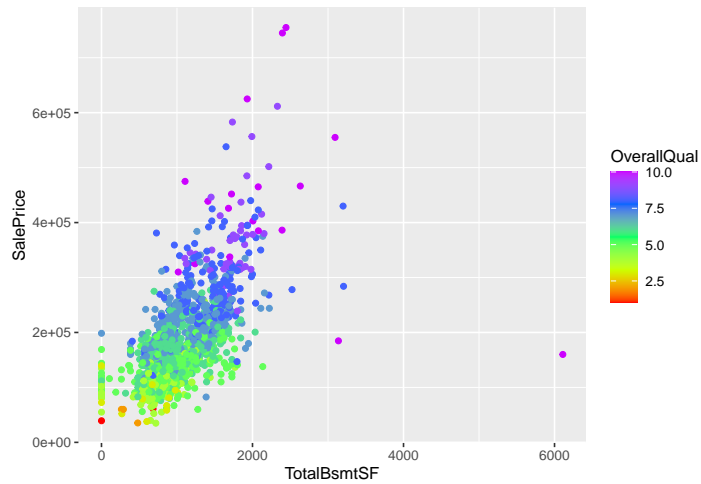
```
# GrLivArea vs SalePrice, colored by Overall Quality

ggplot(data=housing, aes(x = GrLivArea,y=SalePrice,colour=OverallQual)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
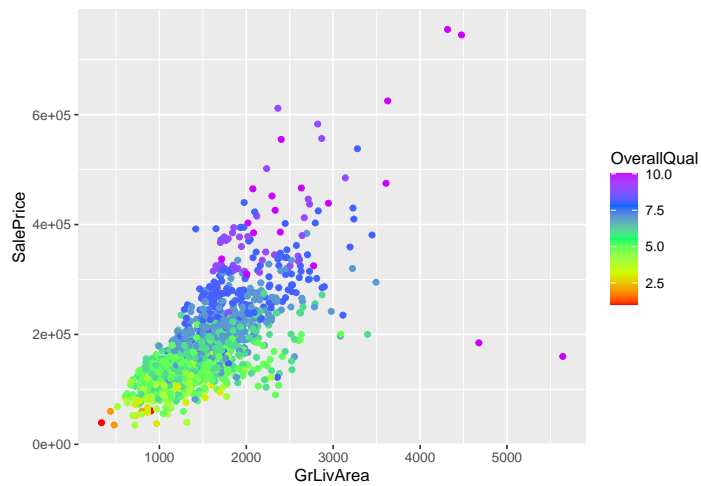


```
# GrLivArea vs SalePrice, colored by Foundation

ggplot(data=housing, aes(x = GrLivArea,y=SalePrice,colour=Foundation)) +

    geom_point()
```
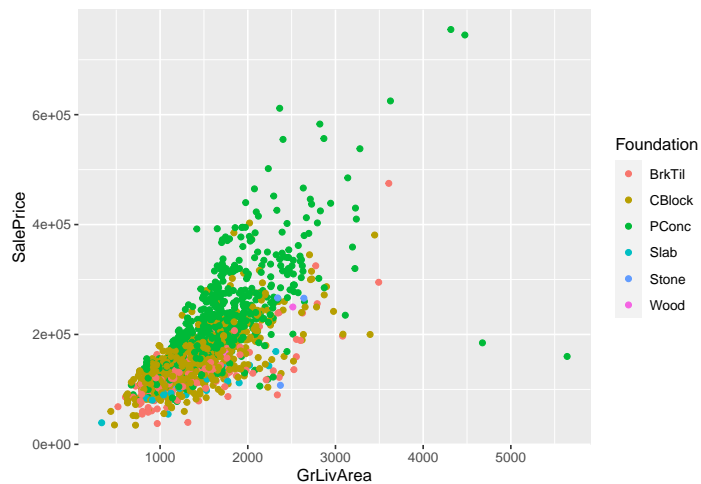
```
# GrLivArea vs SalePrice, colored by GarageType

ggplot(data=housing, aes(x = GrLivArea,y=SalePrice,colour=GarageType)) +

    geom_point()
```
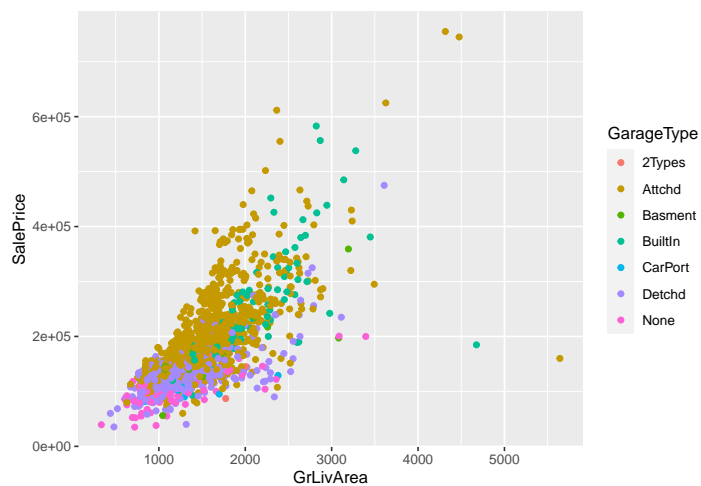


```
# GrLivArea vs SalePrice, colored by GarageCars

ggplot(data=housing, aes(x = GrLivArea,y=SalePrice,colour=GarageCars)) +

    geom_point()+scale_color_gradientn(colours = rainbow(5))
```
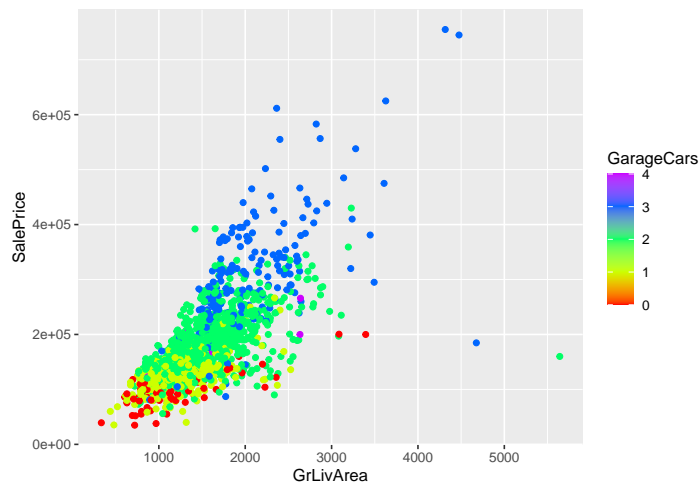
Observations

Now that we have visualized our data & identified predictor variables that show particular trends and patterns with our response variable, we will select features and preprocess our data for building predictive models

Dimensionality Reduction

We will reduce the dimension of our dataset and visualize the results to spot trends and patterns. 3 classes will be created

```
encoder2 <- onehot(housing)
```

```
Warning: Variables excluded for having levels > max_levels:
NeighborhoodVariables excluded for having levels > max_levels:
Exterior1stVariables excluded for having levels > max_levels: Exterior2nd
```

```r
preprocessed2 <- as.data.frame(predict(encoder2, housing))

classifier= function(x){
    if(x<=150000){
        return ("low")
    }
    else if (x>150000 & x<250000){
        return ("med")
    }
    else{
        return ("high")
    }
}

preprocessed2<-preprocessed2%>%mutate(class=sapply(SalePrice,classifier))
df<-preprocessed2[1:223]
pca_res<-prcomp(df,scale.=TRUE)
autoplot(pca_res,data=preprocessed2,colour="class",main = "PCA-2D Dimensionality Reduction: Scaled")
```

```
Warning: `select_()` is deprecated as of dplyr 0.7.0.
Please use `select()` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_warnings()` to see where this warning was generated.
```

PCA−2D Dimensionality Reduction: Scaled
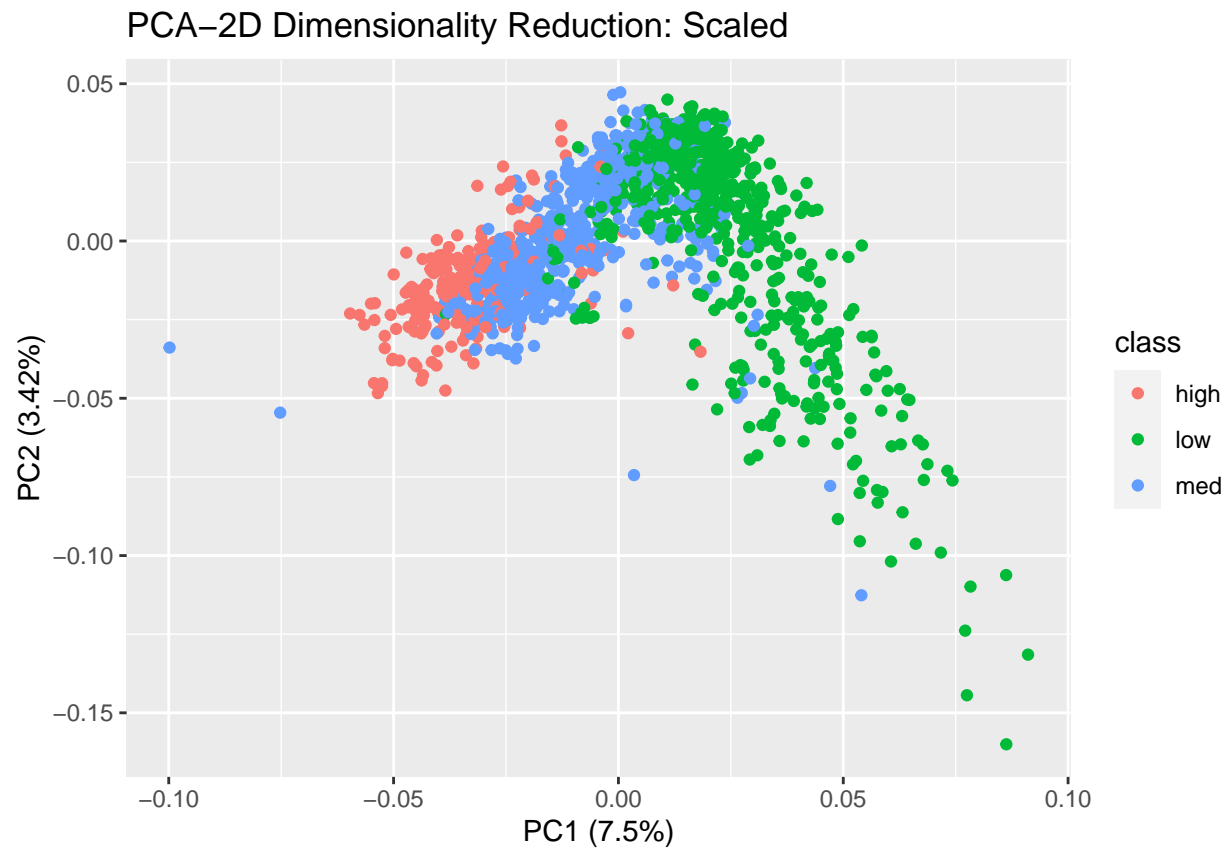

```
pca_res<-prcomp(df)
autoplot(pca_res,data=preprocessed2,colour="class",main = "PCA-2D Dimensionality Reduction: Non Scaled")
```

PCA−2D Dimensionality Reduction: Non Scaled

Data Preprocessing & Feature Engineering

Since our dataset contains an abundance of categorical data, which are inadequeate for many models we will be building ie Multiple Linear Regression, Deep Neural Networks. We will encode these categorical variables with One Hot Encoding

```r
# encoding categorical variables in our dataset

encoder <- onehot(housing)
```

```
Warning: Variables excluded for having levels > max_levels:
NeighborhoodVariables excluded for having levels > max_levels:
Exterior1stVariables excluded for having levels > max_levels: Exterior2nd
```

```r
preprocessed <- as.data.frame(predict(encoder, housing))
```

We will look at particular variables of interest, specifically those that help explain variability in our response variable as indicated by the linear regression function. Once identified we will save a new prepocessed dataset that contain only key properly encoded categorical & key numerical variables variables

```r
summary(lm(data=preprocessed,SalePrice~.))
```

```
Call:
lm(formula = SalePrice ~ ., data = preprocessed)

Residuals:
    Min      1Q  Median      3Q     Max
-205693  -10946     378   10821  205693

Coefficients: (42 not defined because of singularities)
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            -2.228e+04  1.094e+06  -0.020 0.983759
MSSubClass             -1.594e+01  8.572e+01  -0.186 0.852500
`MSZoning=C (all)`     -1.723e+04  9.383e+03  -1.836 0.066598 .
`MSZoning=FV`           9.418e+03  4.182e+03   2.252 0.024503 *
`MSZoning=RH`           4.580e+03  6.902e+03   0.664 0.507128
`MSZoning=RL`           5.702e+03  2.498e+03   2.282 0.022630 *
`MSZoning=RM`                 NA         NA      NA       NA
LotFrontage             6.143e+01  4.684e+01   1.311 0.189931
LotArea                 6.788e-01  1.152e-01   5.892 4.88e-09 ***
`Street=Grvl`          -3.086e+04  1.262e+04  -2.445 0.014637 *
`Street=Pave`                NA         NA      NA       NA
`LotShape=IR1`         -5.098e+02  1.647e+03  -0.310 0.756986
`LotShape=IR2`          3.234e+03  4.550e+03   0.711 0.477398
`LotShape=IR3`         -6.647e+02  9.377e+03  -0.071 0.943494
`LotShape=Reg`               NA         NA      NA       NA
`LandContour=Bnk`      -4.759e+03  3.796e+03  -1.254 0.210242
`LandContour=HLS`       5.373e+03  4.074e+03   1.319 0.187395
`LandContour=Low`      -1.502e+04  5.828e+03  -2.577 0.010074 *
`LandContour=Lvl`            NA         NA      NA       NA
`Utilities=AllPub`      2.891e+04  2.737e+04   1.056 0.291103
`Utilities=NoSeWa`           NA         NA      NA       NA
`LotConfig=Corner`      6.338e+02  1.890e+03   0.335 0.737470
`LotConfig=CulDSac`     9.594e+03  3.123e+03   3.072 0.002169 **
`LotConfig=FR2`        -5.432e+03  3.911e+03  -1.389 0.165158
`LotConfig=FR3`        -6.897e+03  1.321e+04  -0.522 0.601594
`LotConfig=Inside`           NA         NA      NA       NA
`LandSlope=Gtl`         4.675e+04  1.196e+04   3.909 9.77e-05 ***
`LandSlope=Mod`         5.558e+04  1.188e+04   4.680 3.18e-06 ***
`LandSlope=Sev`              NA         NA      NA       NA
`Condition1=Artery`    -7.212e+03  1.298e+04  -0.555 0.578666
`Condition1=Feedr`      2.096e+03  1.270e+04   0.165 0.868906
`Condition1=Norm`       1.186e+04  1.234e+04   0.961 0.336747
`Condition1=PosA`      -4.615e+02  1.560e+04  -0.030 0.976399
`Condition1=PosN`       5.308e+03  1.384e+04   0.383 0.701414
`Condition1=RRAe`      -1.368e+04  1.489e+04  -0.919 0.358334
`Condition1=RRAn`       5.800e+03  1.315e+04   0.441 0.659159
`Condition1=RRNe`      -7.178e+03  2.148e+04  -0.334 0.738271
`Condition1=RRNn`            NA         NA      NA       NA
`Condition2=Artery`     3.787e+03  2.859e+04   0.132 0.894639
`Condition2=Feedr`     -5.450e+03  2.231e+04  -0.244 0.807054
`Condition2=Norm`      -8.248e+03  1.883e+04  -0.438 0.661440
`Condition2=PosA`       1.945e+04  3.929e+04   0.495 0.620724
`Condition2=PosN`      -2.680e+05  2.713e+04  -9.878  < 2e-16 ***
`Condition2=RRAe`      -1.267e+05  4.657e+04  -2.720 0.006623 **
`Condition2=RRAn`      -2.267e+04  3.124e+04  -0.726 0.468099
```

```
`Condition2=RRNn`                  NA        NA        NA        NA
`BldgType=1Fam`            5.762e+03  9.092e+03     0.634  0.526356
`BldgType=2fmCon`        -7.888e+02  7.479e+03    -0.105  0.916023
`BldgType=Duplex`        -1.255e+03  8.738e+03    -0.144  0.885828
`BldgType=Twnhs`         -8.951e+02  4.907e+03    -0.182  0.855285
`BldgType=TwnhsE`                 NA        NA        NA        NA
`HouseStyle=1.5Fin`      -1.391e+03  5.665e+03    -0.245  0.806126
`HouseStyle=1.5Unf`       1.475e+04  9.199e+03     1.603  0.109142
`HouseStyle=1Story`       7.405e+03  6.238e+03     1.187  0.235400
`HouseStyle=2.5Fin`      -2.851e+04  1.407e+04    -2.027  0.042864 *
`HouseStyle=2.5Unf`      -1.408e+04  1.045e+04    -1.348  0.178030
`HouseStyle=2Story`      -7.192e+03  5.436e+03    -1.323  0.186063
`HouseStyle=SFoyer`      -1.169e+03  5.700e+03    -0.205  0.837573
`HouseStyle=SLvl`                 NA        NA        NA        NA
OverallQual               8.308e+03  1.017e+03     8.167  7.54e-16 ***
OverallCond               6.308e+03  9.083e+02     6.945  6.02e-12 ***
YearBuilt                 2.752e+02  7.192e+01     3.826  0.000137 ***
YearRemodAdd              4.867e+01  5.739e+01     0.848  0.396605
`RoofStyle=Flat`         -8.272e+04  3.601e+04    -2.297  0.021775 *
`RoofStyle=Gable`        -7.876e+04  3.093e+04    -2.546  0.011013 *
`RoofStyle=Gambrel`      -7.490e+04  3.204e+04    -2.338  0.019556 *
`RoofStyle=Hip`          -7.863e+04  3.095e+04    -2.540  0.011188 *
`RoofStyle=Mansard`      -7.171e+04  3.089e+04    -2.322  0.020398 *
`RoofStyle=Shed`                 NA        NA        NA        NA
`RoofMatl=ClyTile`       -7.867e+05  3.553e+04   -22.141   < 2e-16 ***
`RoofMatl=CompShg`       -4.492e+04  1.197e+04    -3.753  0.000183 ***
`RoofMatl=Membran`        3.971e+04  3.675e+04     1.080  0.280125
`RoofMatl=Metal`          1.828e+04  3.600e+04     0.508  0.611645
`RoofMatl=Roll`          -6.256e+04  2.903e+04    -2.155  0.031352 *
`RoofMatl=Tar&Grv`       -5.340e+04  2.239e+04    -2.385  0.017232 *
`RoofMatl=WdShake`       -6.396e+04  1.830e+04    -3.495  0.000491 ***
`RoofMatl=WdShngl`               NA        NA        NA        NA
`MasVnrType=BrkCmn`      -1.166e+04  7.504e+03    -1.554  0.120512
`MasVnrType=BrkFace`     -7.004e+03  2.849e+03    -2.458  0.014094 *
`MasVnrType=None`         1.210e+03  3.107e+03     0.389  0.697021
`MasVnrType=Stone`               NA        NA        NA        NA
MasVnrArea                3.159e+01  5.837e+00     5.412  7.44e-08 ***
`ExterQual=Ex`            2.879e+04  5.465e+03     5.267  1.63e-07 ***
`ExterQual=Fa`            1.418e+04  9.239e+03     1.534  0.125159
`ExterQual=Gd`            6.219e+03  2.491e+03     2.496  0.012684 *
`ExterQual=TA`                   NA        NA        NA        NA
`ExterCond=Ex`            1.514e+03  1.828e+04     0.083  0.933987
`ExterCond=Fa`            6.251e+03  5.905e+03     1.058  0.290030
`ExterCond=Gd`           -2.474e+03  2.476e+03    -0.999  0.317991
`ExterCond=Po`            1.657e+04  2.685e+04     0.617  0.537268
`ExterCond=TA`                   NA        NA        NA        NA
`Foundation=BrkTil`       4.048e+04  1.552e+04     2.609  0.009185 **
`Foundation=CBlock`       4.107e+04  1.527e+04     2.690  0.007241 **
`Foundation=PConc`        4.519e+04  1.515e+04     2.982  0.002917 **
`Foundation=Slab`         3.520e+04  1.795e+04     1.961  0.050069 .
`Foundation=Stone`        4.041e+04  1.920e+04     2.105  0.035467 *
`Foundation=Wood`                NA        NA        NA        NA
`BsmtQual=Ex`            -1.994e+04  3.846e+04    -0.519  0.604119
`BsmtQual=Fa`            -3.616e+04  3.833e+04    -0.943  0.345630
```

```
`BsmtQual=Gd`             -3.965e+04  3.825e+04  -1.037 0.300136
`BsmtQual=TA`             -3.881e+04  3.817e+04  -1.017 0.309463
`BsmtQual=None`                  NA         NA      NA       NA
`BsmtCond=Fa`             -3.271e+03  4.459e+03  -0.734 0.463360
`BsmtCond=Gd`             -3.328e+03  3.409e+03  -0.976 0.329160
`BsmtCond=Po`              6.798e+04  3.129e+04   2.172 0.030026 *
`BsmtCond=TA`                    NA         NA      NA       NA
`BsmtCond=None`                  NA         NA      NA       NA
`BsmtExposure=Av`          1.197e+04  2.472e+04   0.484 0.628410
`BsmtExposure=Gd`          2.249e+04  2.480e+04   0.907 0.364735
`BsmtExposure=Mn`          7.246e+03  2.478e+04   0.292 0.770042
`BsmtExposure=No`          6.488e+03  2.466e+04   0.263 0.792530
`BsmtExposure=None`              NA         NA      NA       NA
`BsmtFinType1=ALQ`        -4.428e+03  2.996e+03  -1.478 0.139733
`BsmtFinType1=BLQ`        -1.869e+03  3.269e+03  -0.572 0.567597
`BsmtFinType1=GLQ`         3.100e+03  2.861e+03   1.084 0.278774
`BsmtFinType1=LwQ`        -6.132e+03  3.932e+03  -1.560 0.119117
`BsmtFinType1=Rec`        -3.938e+03  3.298e+03  -1.194 0.232694
`BsmtFinType1=Unf`               NA         NA      NA       NA
`BsmtFinType1=None`              NA         NA      NA       NA
BsmtFinSF1                 4.416e+01  5.431e+00   8.131 1.00e-15 ***
`BsmtFinType2=ALQ`         1.831e+04  2.658e+04   0.689 0.490973
`BsmtFinType2=BLQ`         5.601e+03  2.630e+04   0.213 0.831374
`BsmtFinType2=GLQ`         1.286e+04  2.708e+04   0.475 0.634949
`BsmtFinType2=LwQ`         2.788e+03  2.629e+04   0.106 0.915546
`BsmtFinType2=Rec`         1.062e+04  2.623e+04   0.405 0.685555
`BsmtFinType2=Unf`         1.287e+04  2.619e+04   0.491 0.623329
`BsmtFinType2=None`              NA         NA      NA       NA
BsmtFinSF2                 3.996e+01  9.388e+00   4.257 2.22e-05 ***
BsmtUnfSF                  2.266e+01  4.900e+00   4.625 4.13e-06 ***
TotalBsmtSF                      NA         NA      NA       NA
`Heating=Floor`           -8.115e+03  3.065e+04  -0.265 0.791274
`Heating=GasA`            -1.353e+04  1.470e+04  -0.921 0.357259
`Heating=GasW`            -1.787e+04  1.585e+04  -1.127 0.259764
`Heating=Grav`            -2.184e+04  1.816e+04  -1.203 0.229223
`Heating=OthW`            -4.912e+04  2.420e+04  -2.029 0.042626 *
`Heating=Wall`                   NA         NA      NA       NA
`HeatingQC=Ex`             3.164e+03  2.106e+03   1.502 0.133263
`HeatingQC=Fa`             6.543e+03  4.640e+03   1.410 0.158745
`HeatingQC=Gd`             2.936e+02  2.223e+03   0.132 0.894936
`HeatingQC=Po`             1.725e+03  2.805e+04   0.062 0.950965
`HeatingQC=TA`                   NA         NA      NA       NA
`CentralAir=N`             2.153e+02  3.977e+03   0.054 0.956843
`CentralAir=Y`                   NA         NA      NA       NA
`Electrical=FuseA`         2.651e+03  3.059e+03   0.867 0.386369
`Electrical=FuseF`         8.794e+02  5.687e+03   0.155 0.877136
`Electrical=FuseP`        -4.452e+03  1.883e+04  -0.236 0.813167
`Electrical=Mix`          -5.714e+04  4.698e+04  -1.216 0.224165
`Electrical=SBrkr`               NA         NA      NA       NA
X1stFlrSF                  5.213e+01  5.658e+00   9.214  < 2e-16 ***
X2ndFlrSF                  7.300e+01  5.549e+00  13.155  < 2e-16 ***
LowQualFinSF               9.303e+00  1.910e+01   0.487 0.626309
GrLivArea                        NA         NA      NA       NA
BsmtFullBath              -8.394e+02  2.060e+03  -0.407 0.683789
```

```
BsmtHalfBath            -3.280e+03  3.118e+03  -1.052 0.293021
FullBath                 1.679e+03  2.222e+03   0.756 0.450073
HalfBath                 7.561e+02  2.153e+03   0.351 0.725574
BedroomAbvGr            -4.660e+03  1.391e+03  -3.351 0.000830 ***
KitchenAbvGr            -1.802e+04  5.836e+03  -3.087 0.002064 **
`KitchenQual=Ex`         2.459e+04  4.023e+03   6.111 1.31e-09 ***
`KitchenQual=Fa`         4.810e+03  5.031e+03   0.956 0.339294
`KitchenQual=Gd`        -1.590e+03  2.224e+03  -0.715 0.474924
`KitchenQual=TA`              NA         NA       NA       NA
TotRmsAbvGrd             1.270e+03  9.638e+02   1.318 0.187816
`Functional=Maj1`       -2.818e+04  7.635e+03  -3.690 0.000233 ***
`Functional=Maj2`       -1.870e+04  1.298e+04  -1.441 0.149714
`Functional=Min1`       -1.111e+04  4.901e+03  -2.267 0.023559 *
`Functional=Min2`       -1.053e+04  4.749e+03  -2.218 0.026745 *
`Functional=Mod`        -2.327e+04  7.516e+03  -3.096 0.002007 **
`Functional=Sev`        -6.907e+04  2.984e+04  -2.315 0.020790 *
`Functional=Typ`             NA         NA       NA       NA
Fireplaces               2.919e+03  1.358e+03   2.150 0.031742 *
`GarageType=2Types`      1.068e+05  1.232e+05   0.867 0.386369
`GarageType=Attchd`      1.223e+05  1.231e+05   0.993 0.320662
`GarageType=Basment`     1.266e+05  1.229e+05   1.031 0.302854
`GarageType=BuiltIn`     1.241e+05  1.232e+05   1.007 0.314089
`GarageType=CarPort`     1.277e+05  1.238e+05   1.032 0.302471
`GarageType=Detchd`      1.270e+05  1.232e+05   1.031 0.302762
`GarageType=None`            NA         NA       NA       NA
GarageYrBlt             -6.868e+01  6.309e+01  -1.089 0.276564
`GarageFinish=Fin`      -5.762e+02  2.442e+03  -0.236 0.813486
`GarageFinish=RFn`      -2.474e+03  2.223e+03  -1.113 0.266000
`GarageFinish=Unf`           NA         NA       NA       NA
`GarageFinish=None`          NA         NA       NA       NA
GarageCars               3.760e+03  2.339e+03   1.608 0.108175
GarageArea               2.382e+01  7.992e+00   2.981 0.002931 **
`GarageQual=Ex`          1.028e+05  3.131e+04   3.282 0.001058 **
`GarageQual=Fa`         -4.535e+03  5.079e+03  -0.893 0.372131
`GarageQual=Gd`          4.747e+03  7.984e+03   0.595 0.552207
`GarageQual=Po`         -1.315e+04  2.542e+04  -0.517 0.604966
`GarageQual=TA`              NA         NA       NA       NA
`GarageQual=None`            NA         NA       NA       NA
`GarageCond=Ex`         -9.866e+04  3.607e+04  -2.735 0.006319 **
`GarageCond=Fa`         -2.237e+03  5.620e+03  -0.398 0.690700
`GarageCond=Gd`         -3.340e+03  9.392e+03  -0.356 0.722188
`GarageCond=Po`          3.590e+02  1.462e+04   0.025 0.980413
`GarageCond=TA`              NA         NA       NA       NA
`GarageCond=None`            NA         NA       NA       NA
`PavedDrive=N`          -1.199e+03  3.569e+03  -0.336 0.736920
`PavedDrive=P`          -4.538e+03  5.098e+03  -0.890 0.373580
`PavedDrive=Y`               NA         NA       NA       NA
WoodDeckSF               1.251e+01  6.056e+00   2.065 0.039118 *
OpenPorchSF             -9.592e+00  1.185e+01  -0.810 0.418245
EnclosedPorch            1.044e+01  1.282e+01   0.814 0.415576
X3SsnPorch               2.954e+01  2.359e+01   1.252 0.210715
ScreenPorch              2.393e+01  1.289e+01   1.857 0.063590 .
PoolArea                 6.898e+01  1.912e+01   3.607 0.000321 ***
MiscVal                 -1.622e-01  1.506e+00  -0.108 0.914256
```

```
MoSold                       -1.934e+02  2.570e+02  -0.753 0.451799
YrSold                       -3.101e+02  5.386e+02  -0.576 0.564924
`SaleType=COD`               -4.864e+03  4.367e+03  -1.114 0.265552
`SaleType=Con`                3.468e+04  1.768e+04   1.961 0.050043 .
`SaleType=ConLD`              8.762e+03  9.314e+03   0.941 0.347019
`SaleType=ConLI`              4.773e+02  1.144e+04   0.042 0.966730
`SaleType=ConLw`              4.801e+03  1.200e+04   0.400 0.689065
`SaleType=CWD`                1.738e+04  1.296e+04   1.340 0.180336
`SaleType=New`                1.680e+04  1.580e+04   1.063 0.287893
`SaleType=Oth`                4.814e+03  1.487e+04   0.324 0.746171
`SaleType=WD`                        NA         NA      NA       NA
`SaleCondition=Abnorml` -5.197e+03  1.573e+04  -0.330 0.741157
`SaleCondition=AdjLand`  7.007e+02  2.120e+04   0.033 0.973634
`SaleCondition=Alloca`   1.296e+03  1.774e+04   0.073 0.941801
`SaleCondition=Family`  -8.208e+03  1.654e+04  -0.496 0.619723
`SaleCondition=Normal`  -1.814e+02  1.557e+04  -0.012 0.990711
`SaleCondition=Partial`              NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24410 on 1278 degrees of freedom
Multiple R-squared:  0.9173,    Adjusted R-squared:  0.9056
F-statistic: 78.35 on 181 and 1278 DF,  p-value: < 2.2e-16
```

```r
ceoffs <- coef(summary(lm(data=preprocessed,SalePrice~.)))
ss_sig <- ceoffs[ceoffs[,"Pr(>|t|)"]<0.1,]

printCoefmat(ss_sig)
```

```
                      Estimate  Std. Error  t value   Pr(>|t|)
`MSZoning=C (all)`  -1.7226e+04  9.3826e+03  -1.8359 0.0665976 .
`MSZoning=FV`        9.4175e+03  4.1822e+03   2.2518 0.0245028 *
`MSZoning=RL`        5.7016e+03  2.4981e+03   2.2824 0.0226296 *
LotArea              6.7882e-01  1.1522e-01   5.8917 4.881e-09 ***
`Street=Grvl`       -3.0858e+04  1.2623e+04  -2.4446 0.0146367 *
`LandContour=Low`   -1.5020e+04  5.8283e+03  -2.5771 0.0100743 *
`LotConfig=CulDSac`  9.5939e+03  3.1227e+03   3.0723 0.0021687 **
`LandSlope=Gtl`      4.6747e+04  1.1960e+04   3.9085 9.773e-05 ***
`LandSlope=Mod`      5.5579e+04  1.1877e+04   4.6796 3.180e-06 ***
`Condition2=PosN`   -2.6800e+05  2.7131e+04  -9.8778 < 2.2e-16 ***
`Condition2=RRAe`   -1.2666e+05  4.6573e+04  -2.7197 0.0066232 **
`HouseStyle=2.5Fin` -2.8512e+04  1.4066e+04  -2.0271 0.0428643 *
OverallQual          8.3085e+03  1.0174e+03   8.1666 7.543e-16 ***
OverallCond          6.3081e+03  9.0833e+02   6.9447 6.021e-12 ***
YearBuilt            2.7516e+02  7.1922e+01   3.8258 0.0001366 ***
`RoofStyle=Flat`    -8.2724e+04  3.6013e+04  -2.2971 0.0217752 *
`RoofStyle=Gable`   -7.8756e+04  3.0933e+04  -2.5460 0.0110130 *
`RoofStyle=Gambrel` -7.4900e+04  3.2040e+04  -2.3377 0.0195558 *
`RoofStyle=Hip`     -7.8633e+04  3.0952e+04  -2.5405 0.0111884 *
`RoofStyle=Mansard` -7.1710e+04  3.0885e+04  -2.3218 0.0203979 *
`RoofMatl=ClyTile`  -7.8667e+05  3.5529e+04 -22.1415 < 2.2e-16 ***
`RoofMatl=CompShg`  -4.4922e+04  1.1971e+04  -3.7527 0.0001828 ***
`RoofMatl=Roll`     -6.2558e+04  2.9030e+04  -2.1550 0.0313519 *
`RoofMatl=Tar&Grv`  -5.3403e+04  2.2393e+04  -2.3848 0.0172323 *
```

```
`RoofMatl=WdShake`     -6.3957e+04  1.8301e+04  -3.4947 0.0004908 ***
`MasVnrType=BrkFace`   -7.0036e+03  2.8490e+03  -2.4582 0.0140938 *
MasVnrArea              3.1590e+01  5.8371e+00   5.4119 7.440e-08 ***
`ExterQual=Ex`          2.8785e+04  5.4655e+03   5.2667 1.628e-07 ***
`ExterQual=Gd`          6.2186e+03  2.4914e+03   2.4960 0.0126845 *
`Foundation=BrkTil`     4.0482e+04  1.5516e+04   2.6091 0.0091851 **
`Foundation=CBlock`     4.1068e+04  1.5268e+04   2.6899 0.0072406 **
`Foundation=PConc`      4.5187e+04  1.5153e+04   2.9821 0.0029169 **
`Foundation=Slab`       3.5205e+04  1.7950e+04   1.9612 0.0500692 .
`Foundation=Stone`      4.0412e+04  1.9196e+04   2.1052 0.0354671 *
`BsmtCond=Po`           6.7978e+04  3.1295e+04   2.1722 0.0300263 *
BsmtFinSF1              4.4161e+01  5.4314e+00   8.1306 1.001e-15 ***
BsmtFinSF2              3.9963e+01  9.3878e+00   4.2569 2.225e-05 ***
BsmtUnfSF               2.2660e+01  4.8998e+00   4.6247 4.132e-06 ***
`Heating=OthW`         -4.9120e+04  2.4204e+04  -2.0294 0.0426260 *
X1stFlrSF               5.2131e+01  5.6579e+00   9.2137 < 2.2e-16 ***
X2ndFlrSF               7.3001e+01  5.5493e+00  13.1550 < 2.2e-16 ***
BedroomAbvGr           -4.6599e+03  1.3908e+03  -3.3505 0.0008301 ***
KitchenAbvGr           -1.8016e+04  5.8358e+03  -3.0872 0.0020642 **
`KitchenQual=Ex`        2.4586e+04  4.0234e+03   6.1107 1.315e-09 ***
`Functional=Maj1`      -2.8177e+04  7.6351e+03  -3.6905 0.0002333 ***
`Functional=Min1`      -1.1110e+04  4.9010e+03  -2.2670 0.0235593 *
`Functional=Min2`      -1.0533e+04  4.7491e+03  -2.2178 0.0267451 *
`Functional=Mod`       -2.3265e+04  7.5157e+03  -3.0956 0.0020069 **
`Functional=Sev`       -6.9071e+04  2.9841e+04  -2.3146 0.0207900 *
Fireplaces              2.9188e+03  1.3576e+03   2.1500 0.0317420 *
GarageArea              2.3820e+01  7.9919e+00   2.9806 0.0029315 **
`GarageQual=Ex`         1.0277e+05  3.1310e+04   3.2822 0.0010580 **
`GarageCond=Ex`        -9.8661e+04  3.6070e+04  -2.7353 0.0063191 **
WoodDeckSF              1.2507e+01  6.0565e+00   2.0651 0.0391179 *
ScreenPorch             2.3934e+01  1.2891e+01   1.8567 0.0635902 .
PoolArea                6.8982e+01  1.9122e+01   3.6074 0.0003212 ***
`SaleType=Con`          3.4683e+04  1.7682e+04   1.9615 0.0500427 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
feature_engineered_preprocessed_data<-preprocessed[c("MSZoning=C (all)","MSZoning=FV","MSZoning=RL","Lo
```

Conclusions

Our dataset is properly processed & ready for modelling