Queen Mary
University of London

DATA ANALYTICS

GROUP ASSESSMENT

GROUP 4

SUBMITTED BY:

- MUHAMMAD ZAIN HASHMI
- SNEHAL SRIVASTAVA
- WEIDONG LIANG

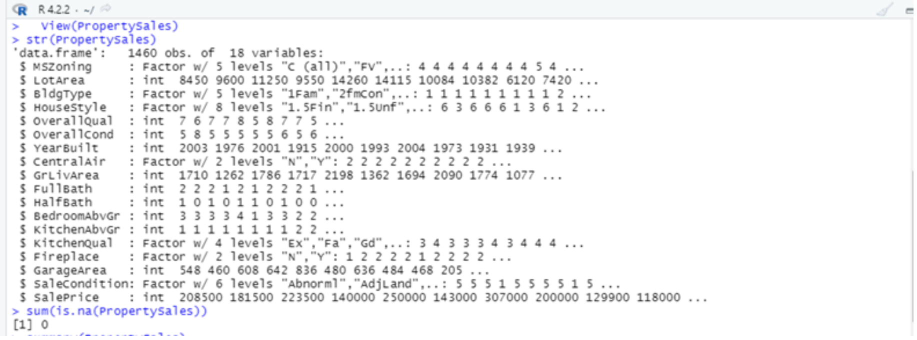
Abstract

The study aims to develop statistical analysis with programme R to identify the key characteristics of a historical dataset of property, to develop a regression model with one or more variables for predicting the Sales Price and development of a classification model for predicting the installation of Fireplace in the property.

1. Exploratory Data Analysis

1.1 Data Import, Cleaning and Processing

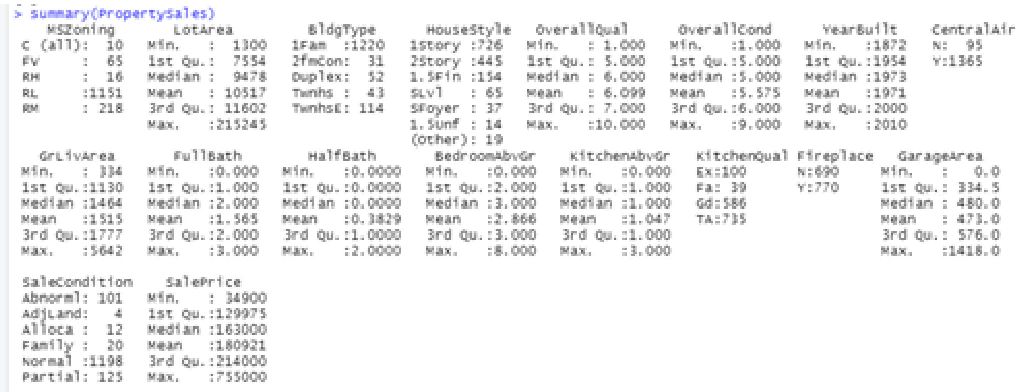
The property sales dataset contained 1460 observations and 18 variables, out of which 11 variables were numerical and 7 variables were qualitative character data. The following information in Figure A shows that the dataset was imported into R with the name of ‘PropertySales’ and further investigation was done to find the missing values and key characteristics.



```
>   View(PropertySales)
> str(PropertySales)
'data.frame':   1460 obs. of  18 variables:
 $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
 $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
 $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
 $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
 $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
 $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
 $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
 $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
 $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
 $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
 $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
 $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
 $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
 $ Fireplace    : Factor w/ 2 levels "N","Y": 1 2 2 2 2 1 2 2 2 2 ...
 $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
 $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...
 $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
> sum(is.na(PropertySales))
[1] 0
```

Figure A

With reference to Figure B, overall 1460 houses were sold at an average price of US$180,921 during Year 1872 to Year 2010. Most of the property sold was in the residential low density of MS Zoning. 94% of the properties had central air conditioning installed, 53% had a fireplace. On an average, each property had approximately 2.9 bedrooms above ground, 1.57 full bathrooms, 0.38 half bathroom, and 1.05 kitchens above ground. Out of all the property sold, approximately 473 properties had garage area.



```
> summary(PropertySales)
    MSZoning        LotArea          BldgType      HouseStyle    OverallQual      OverallCond      YearBuilt    CentralAir
 C (all):  10   Min.   :  1300   1Fam  :1220   1Story :726   Min.   : 1.000   Min.   :1.000   Min.   :1872   N:  95    
 FV     :  65   1st Qu.:  7554   2fmCon:  31   2Story :445   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   Y:1365    
 RH     :  16   Median :  9478   Duplex:  52   1.5Fin :154   Median : 6.000   Median :5.000   Median :1973             
 RL     :1151   Mean   : 10517   Twnhs :  43   SLvl   : 65   Mean   : 6.099   Mean   :5.575   Mean   :1971             
 RM     : 218   3rd Qu.: 11602   TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000             
                Max.   :215245                 1.5Unf : 14   Max.   :10.000   Max.   :9.000   Max.   :2010             
                                               (Other): 19                                                            
   GrLivArea       FullBath        HalfBath        BedroomAbvGr    KitchenAbvGr    KitchenQual Fireplace   GarageArea    
 Min.   : 334   Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100      N:690     Min.   :   0.0  
 1st Qu.:1130   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39      Y:770     1st Qu.: 334.5  
 Median :1464   Median :2.000   Median :0.0000   Median :3.000   Median :1.000   Gd:586                Median : 480.0  
 Mean   :1515   Mean   :1.565   Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735                Mean   : 473.0  
 3rd Qu.:1777   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000                         3rd Qu.: 576.0  
 Max.   :5642   Max.   :3.000   Max.   :2.0000   Max.   :8.000   Max.   :3.000                         Max.   :1418.0  

  SaleCondition    SalePrice     
 Abnorml: 101   Min.   : 34900  
 AdjLand:   4   1st Qu.:129975  
 Alloca :  12   Median :163000  
 Family :  20   Mean   :180921  
 Normal :1198   3rd Qu.:214000  
 Partial: 125   Max.   :755000  
```

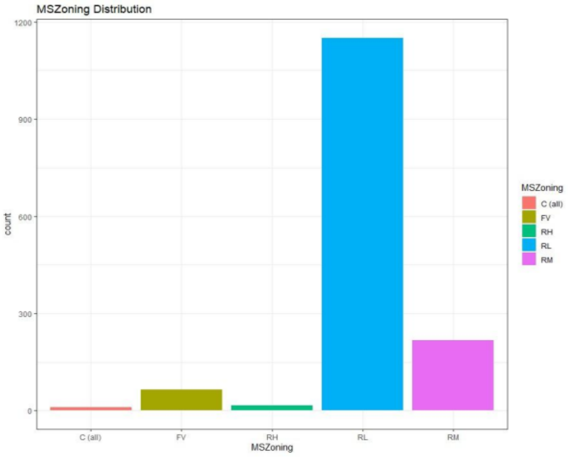
Figure B

1.2 Key Characteristics of the Dataset

Exploratory data analysis was performed on the given dataset to identify the key characteristics of the data. The following graphs represent our findings:

Graph 1.1 was plotted to find the MSZoning distribution of the sale. On the x and y axis we have MSZoning and count of properties, respectively. From this graph, maximum sales have taken place in the RL (Residential Low Density) Zone. It refers to the residential areas which have low density of population and low-rise properties.

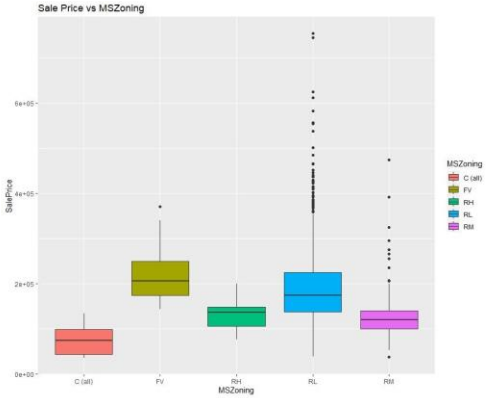
```
>ggplot(PropertySales,aes(x=MSZoning,fill=MSZoning))+theme_bw()+geom_bar()+ylab("count")+ggtitle("MSZoning Distribution")
```



Graph 1.1

Graph 1.2 explores the relationship between MSZoning and Sales Price of the properties, with the former on x-axis and latter on the y-axis. Looking at this boxplot, we can see that the customers are willing to pay more for properties in floating village (FV) zone. The outliers of the RL box plot indicate that the customers are willing to pay much higher than the median price for a property in RL zone, and hence indicates that RL is more valued.

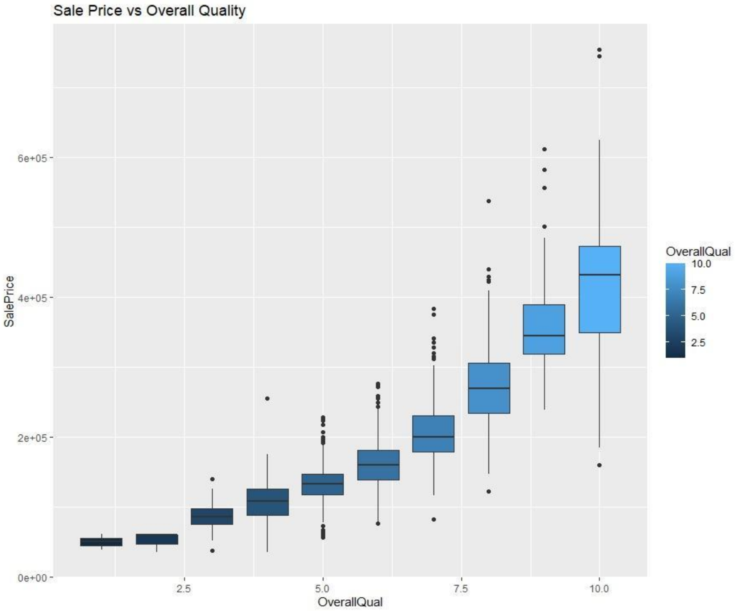
```
>ggplot(PropertySales,aes(x=MSZoning,y=SalePrice,group=MSZoning,fill=MSZoning))+geom_boxplot(alpha=1.5)+ggtitle("Sale Price vs MSZoning")
```



Graph 1.2

The relationship between Sales Price and the Overall Quality of the property is depicted through Graph 1.3. We can very clearly see the positive trend in this graph between the two variables, i.e., with higher overall quality of property, the customers are willing to pay higher prices. We can also see the increased variability in the sales prices quoted by the customer with an increase in the overall quality.

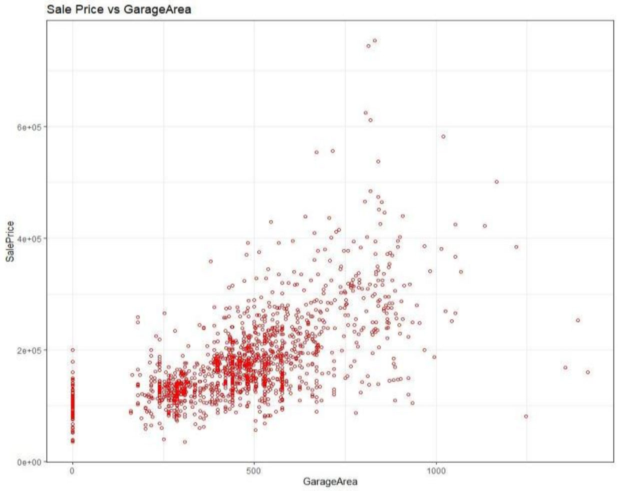
```
>ggplot(PropertySales,aes(x=OverallQual,y=SalePrice,group=OverallQual,fill=OverallQual))+geom_boxpl
ot(alpha=1.5)+ggtitle("Sale Price vs Overall Quality")
```



Graph 1.3

Graph 1.4 shows the impact of garage area of a property in terms of its sales price. Garage area (GA) in square feet is plotted on the x-axis whereas sales price is plotted on the y-axis. We can observe from this scatter plot that customers are willing to buy a property even with no garage area (0 ft^2). Furthermore, there is a concentration of observations in the lower price range for a GA of approximately 0-675 ft^2. This indicates that the customers hardly value the GA within this range. When the GA increases beyond this point, the variability for sales price increases exponentially, with respect to GA.

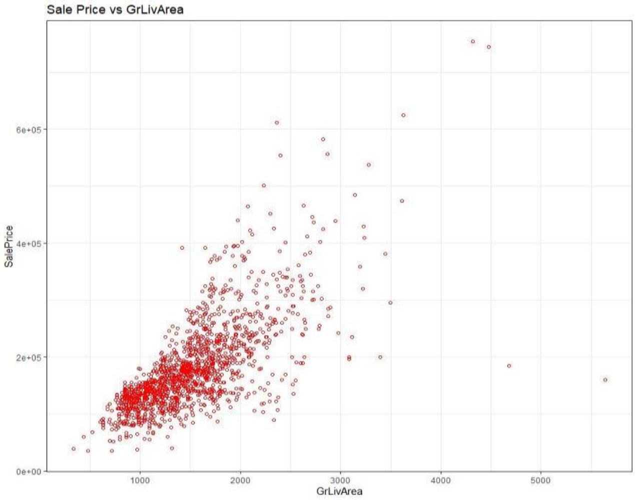
```
>ggplot(PropertySales,aes(x=GarageArea,y=SalePrice))+theme_bw()+geom_point(shape=1,color="Red")
+ggtitle("Sale Price vs GarageArea")
```



Graph 1.4

The relation between Above Ground Living Area in ft^2 and sales price of a property is presented in the form of a scatter plot in Graph 1.5. Sales price is on the y-axis whereas the GrLivArea is on the x-axis. We see a concentration of observations up until about 1800 ft^2 with respect to the sales price. This indicates there is less variability in this range of GrLivArea. Although, we do see more variability of the scatter points with a larger above ground living area, it is not an exponential increase, and the slope of best fit is very steep. This shows that the customers consider a GrLivArea while paying for a property only when it is larger than approximately 1800 ft^2.

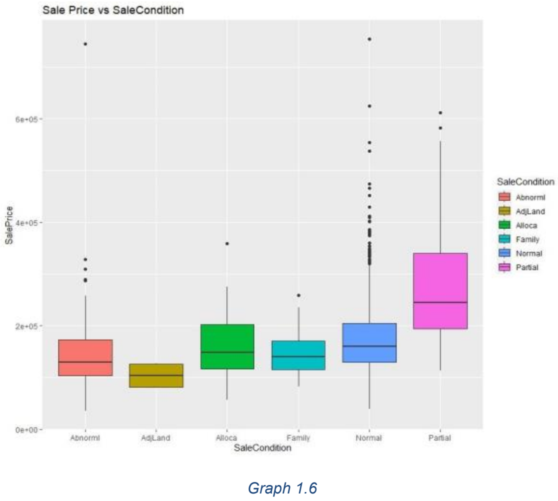
```
>ggplot(PropertySales,aes(x=GrLivArea,y=SalePrice))+theme_bw()+geom_point(shape=1,color="Red")+ggtitle("Sale Price vs GrLivArea")
```



Graph 1.5

Graph 1.6 shows how the condition of sale impacts the sale price. We have SaleCondition on the x-axis and SalePrice on the y-axis of this boxplot. We can infer that customers prefer Partial sale condition over the others, however there is more variability in this sale condition w.r.t. sale price. We also see a lot of outliers in the box plot of Normal sale condition. This indicates that some people are ready to pay more than the sale price included in the 4th quartile for a normal sale condition.

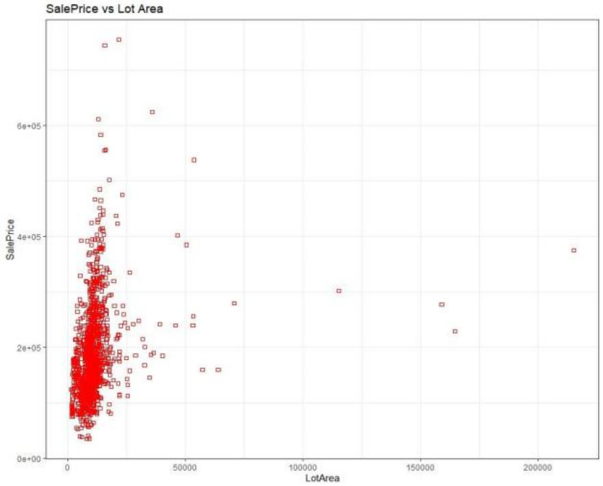
```
>ggplot(PropertySales,aes(x=SaleCondition,y=SalePrice,group=SaleCondition,fill=SaleCondition))+geom_boxplot(alpha=1.5)+ggtitle("Sale Price vs SaleCondition")
```



Graph 1.6

The relationship between Lot Area and Sale Price can be seen in Graph 1.7 through a scatter plot. Very evidently, the customers do not prefer attributes like lot area while considering purchasing a property. Through this scatter plot we see concentration of scatter points towards the y-axis which indicates that irrespective of an increase in the lot area, the customers are willing to pay more for a property based on other attributes.

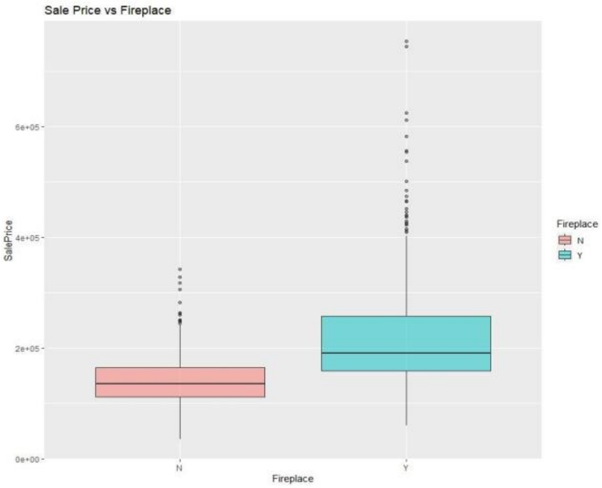
```
>ggplot(PropertySales,aes(x=LotArea,y=SalePrice))+theme_bw()+geom_point(shape=0,color="Red")+ggtitle("SalePrice vs Lot Area")
```



Graph 1.7

Graph 1.8 shows the relation between Sales Price and Fireplace. In this boxplot, x-axis represents whether a property has a fireplace or not (N=No, Y=Yes) and the y-axis represents the sales price of the property. The box plot shows that the median price a consumer is willing to pay for a property with a fireplace is higher than the price they are willing to pay for a property without a fireplace, hence, people prefer properties with a fireplace. Looking at the range of outliers, we can also infer that quite a few people are willing to pay a price well above the 4th quartile for a property with a fireplace. The box plot also shows more variability in the sales price quoted for a property containing a fireplace unlike properties without a fireplace.

```
>ggplot(PropertySales,aes(x=Fireplace,y=SalePrice,group=Fireplace,fill=Fireplace))+geom_boxplot(alpha=0.5)+ggtitle("SalePrice vs Fireplace")
```



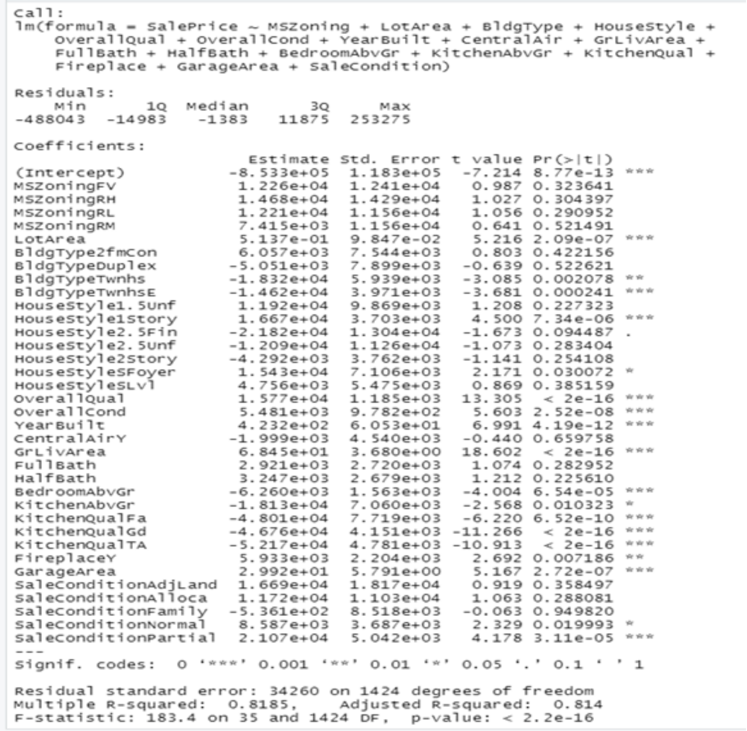
Graph 1.8

2. Sale Price Prediction with Regression Models

To predict the sale price, we performed a multiple linear regression on programme R with all the 17 variables as explanatory variables in the dataset, to check the statistical significance with sale price as a response variable. As can be seen from the corresponding code below, out of 17 only 11 variables had high statistical significance illustrated by high P values, and 6 had low or no, due to which we eliminated the rest from our prediction and further analysis. Overall, regression model shows a good Multiple R squared value with 0.8185.

```
PropertySalesMRegression=lm(SalePrice~
MSZoning+LotArea+BldgType+HouseStyle+OverallQual+OverallCond+YearBuilt+CentralAir+GrLivArea+FullBath+HalfBath+BedroomAbvGr+KitchenAbvGr+KitchenQual+Fireplace+GarageArea+SaleCondition)
```

```
summary(PropertySalesMRegression)
```

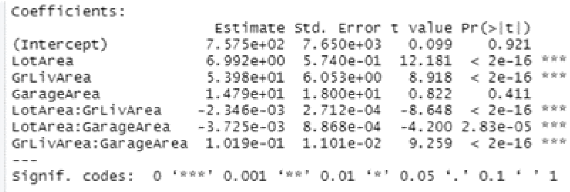


```
Call:
lm(formula = SalePrice ~ MSZoning + LotArea + BldgType + HouseStyle +
    OverallQual + OverallCond + YearBuilt + CentralAir + GrLivArea +
    FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual +
    Fireplace + GarageArea + SaleCondition)

Residuals:
    Min      1Q  Median      3Q     Max
-488043  -14983   -1383   11875  253275

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -8.533e+05  1.183e+05  -7.214 8.77e-13 ***
MSZoningFV            1.226e+04  1.241e+04   0.987 0.323641
MSZoningRH            1.468e+04  1.429e+04   1.027 0.304397
MSZoningRL            1.221e+04  1.156e+04   1.056 0.290952
MSZoningRM            7.415e+03  1.156e+04   0.641 0.521491
LotArea               5.137e-01  9.847e-02   5.216 2.09e-07 ***
BldgType2fmCon        6.057e+03  7.544e+03   0.803 0.422156
BldgTypeDuplex       -5.051e+03  7.899e+03  -0.639 0.522621
BldgTypeTwnhs        -1.832e+04  5.939e+03  -3.085 0.002078 **
BldgTypeTwnhsE       -1.462e+04  3.971e+03  -3.681 0.000241 ***
HouseStyle1.5Unf      1.192e+04  9.869e+03   1.208 0.227323
HouseStyle1Story      1.667e+04  3.703e+03   4.500 7.34e-06 ***
HouseStyle2.5Fin     -2.182e+04  1.304e+04  -1.673 0.094487 .
HouseStyle2.5Unf     -1.209e+04  1.126e+04  -1.073 0.283404
HouseStyle2Story     -4.292e+03  3.762e+03  -1.141 0.254108
HouseStyleSFoyer      1.543e+04  7.106e+03   2.171 0.030072 *
HouseStyleSLvl        4.756e+03  5.475e+03   0.869 0.385159
OverallQual           1.577e+04  1.185e+03  13.305  < 2e-16 ***
OverallCond           5.481e+03  9.782e+02   5.603 2.52e-08 ***
YearBuilt             4.232e+02  6.053e+01   6.991 4.19e-12 ***
CentralAirY          -1.999e+03  4.540e+03  -0.440 0.659758
GrLivArea             6.845e+01  3.680e+00  18.602  < 2e-16 ***
FullBath              2.921e+03  2.720e+03   1.074 0.282952
HalfBath              3.247e+03  2.679e+03   1.212 0.225610
BedroomAbvGr         -6.260e+03  1.563e+03  -4.004 6.54e-05 ***
KitchenAbvGr         -1.813e+04  7.060e+03  -2.568 0.010323 *
KitchenQualFa        -4.801e+04  7.719e+03  -6.220 6.52e-10 ***
KitchenQualGd        -4.676e+04  4.151e+03 -11.266  < 2e-16 ***
KitchenQualTA        -5.217e+04  4.781e+03 -10.913  < 2e-16 ***
FireplaceY            5.933e+03  2.204e+03   2.692 0.007186 **
GarageArea            2.992e+01  5.791e+00   5.167 2.72e-07 ***
SaleConditionAdjLand  1.669e+04  1.817e+04   0.919 0.358497
SaleConditionAlloca   1.172e+04  1.103e+04   1.063 0.288081
SaleConditionFamily  -5.361e+02  8.518e+03  -0.063 0.949820
SaleConditionNormal   8.587e+03  3.687e+03   2.329 0.019993 *
SaleConditionPartial  2.107e+04  5.042e+03   4.178 3.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34260 on 1424 degrees of freedom
Multiple R-squared:  0.8185,	Adjusted R-squared:  0.814
F-statistic: 183.4 on 35 and 1424 DF,  p-value: < 2.2e-16
```

For further analysis to have a good model for the prediction of property sale price, we came to a conclusion that confounding variables should be eliminated from further research in regression model as they will not have a statistical significance with sale price by itself. To eliminate the confounding variables, we worked on the multiple linear regression among all the quantitative variables and concluded that GarageArea was a confounding factor influenced by GrLivArea and Lot Area.

```
PropertySales.CFRegressionA=lm(SalePrice~
LotArea+GrLivArea+GarageArea+LotArea:GrLivArea+LotArea:GarageArea+GrLivArea:GarageArea)
```

```
summary(PropertySales.CFRegressionA)
```

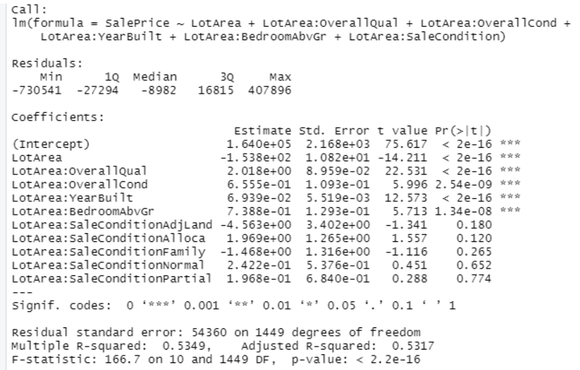


```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.575e+02  7.650e+03   0.099    0.921
LotArea               6.992e+00  5.740e-01  12.181  < 2e-16 ***
GrLivArea             5.398e+01  6.053e+00   8.918  < 2e-16 ***
GarageArea            1.479e+01  1.800e+01   0.822    0.411
LotArea:GrLivArea    -2.346e-03  2.712e-04  -8.648  < 2e-16 ***
LotArea:GarageArea   -3.725e-03  8.868e-04  -4.200 2.83e-05 ***
GrLivArea:GarageArea  1.019e-01  1.101e-02   9.259  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

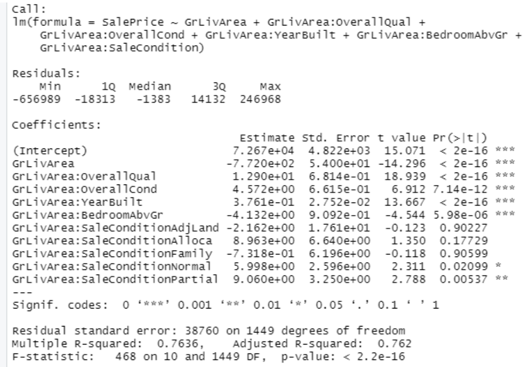
In order to further eliminate the confounding variable, we ran multiple regression on each categorical variable and found out that BldgType, HouseStyle and KitchenQual are confounding and eliminated them from our analysis. Following are the codes we worked on in order to investigate and eliminate the confounding variables.

```
PropertySales.CFRegressionB=lm(SalePrice~ BldgType+BldgType:HouseStyle+BldgType:OverallQual+BldgType:OverallCond+BldgType:YearBuilt+BldgType:BedroomAbvGr+BldgType:KitchenQual+BldgType:SaleCondition)

summary(PropertySales. CFRegressionB)

PropertySales.CFRegressionC=lm(SalePrice~ HouseStyle+HouseStyle:BldgType+HouseStyle:OverallQual+HouseStyle:OverallCond+HouseStyle:YearBuilt+HouseStyle:BedroomAbvGr+HouseStyle:KitchenQual+HouseStyle:SaleCondition)

summary(PropertySales. CFRegressionC)

PropertySales.CFRegressionD=lm(SalePrice~ OverallQual+OverallQual:BldgType+OverallQual:HouseStyle+OverallQual:OverallCond+OverallQual:YearBuilt+OverallQual:BedroomAbvGr+OverallQual:KitchenQual+OverallQual:SaleCondition)

summary(PropertySales. CFRegressionD)

PropertySales.CFRegressionE=lm(SalePrice~ OverallCond+OverallCond:BldgType+OverallCond:HouseStyle+OverallCond:OverallQual+OverallCond:YearBuilt+OverallCond:BedroomAbvGr+OverallCond:KitchenQual+OverallCond:SaleCondition)

summary(PropertySales.CFRegressionE)

PropertySales.CReg4=lm(SalePrice~ YearBuilt+YearBuilt:BldgType+YearBuilt:HouseStyle+YearBuilt:OverallQual+YearBuilt:OverallCond+YearBuilt:BedroomAbvGr+YearBuilt:KitchenQual+YearBuilt:SaleCondition)

summary(PropertySales.CReg4)

PropertySales.CFRegressionF=lm(SalePrice~ BedroomAbvGr+BedroomAbvGr:BldgType+BedroomAbvGr:HouseStyle+BedroomAbvGr:OverallQual+BedroomAbvGr:OverallCond+BedroomAbvGr:YearBuilt:+BedroomAbvGr:KitchenQual+BedroomAbvGr:SaleCondition)

summary(PropertySales.CFRegressionF)

PropertySales.CFRegressionG=lm(SalePrice~ KitchenQual+KitchenQual:BldgType+KitchenQual:HouseStyle+KitchenQual:OverallQual+KitchenQual:OverallCond+KitchenQual:YearBuilt+KitchenQual:BedroomAbvGr+KitchenQual:SaleCondition)

summary(PropertySales.CFRegressionG)

PropertySales.CFRegressionH=lm(SalePrice~ SaleCondition+SaleCondition:BldgType+SaleCondition:HouseStyle+SaleCondition:OverallQual+SaleCondition:OverallCond+SaleCondition:YearBuilt+SaleCondition:BedroomAbvGr+SaleCondition:KitchenQual)

summary(PropertySales. CFRegressionH)
```

After eliminating the investigated confounding variables, we further ran multiple linear regression on the remaining quantitative and qualitative variables in order to form the best regression model for predicting the sale price of property.



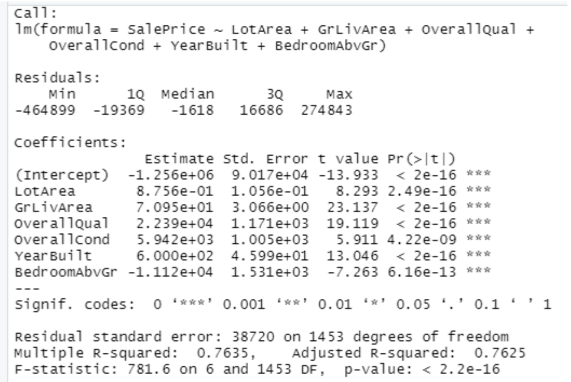
```
Call:
lm(formula = SalePrice ~ LotArea + LotArea:OverallQual + LotArea:OverallCond +
    LotArea:YearBuilt + LotArea:BedroomAbvGr + LotArea:SaleCondition)

Residuals:
    Min      1Q  Median      3Q     Max 
-730541  -27294   -8982   16815  407896 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   1.640e+05  2.168e+03  75.617  < 2e-16 ***
LotArea                      -1.538e+02  1.082e+01 -14.211  < 2e-16 ***
LotArea:OverallQual           2.018e+00  8.959e-02  22.531  < 2e-16 ***
LotArea:OverallCond           6.555e-01  1.093e-01   5.996 2.54e-09 ***
LotArea:YearBuilt             6.939e-02  5.519e-03  12.573  < 2e-16 ***
LotArea:BedroomAbvGr          7.388e-01  1.293e-01   5.713 1.34e-08 ***
LotArea:SaleConditionAdjLand -4.563e+00  3.402e+00  -1.341    0.180    
LotArea:SaleConditionAlloca   1.969e+00  1.265e+00   1.557    0.120    
LotArea:SaleConditionFamily  -1.468e+00  1.316e+00  -1.116    0.265    
LotArea:SaleConditionNormal   2.422e-01  5.376e-01   0.451    0.652    
LotArea:SaleConditionPartial  1.968e-01  6.840e-01   0.288    0.774    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54360 on 1449 degrees of freedom
Multiple R-squared:  0.5349,	Adjusted R-squared:  0.5317 
F-statistic: 166.7 on 10 and 1449 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = SalePrice ~ GrLivArea + GrLivArea:OverallQual +
    GrLivArea:OverallCond + GrLivArea:YearBuilt + GrLivArea:BedroomAbvGr +
    GrLivArea:SaleCondition)

Residuals:
    Min      1Q  Median      3Q     Max 
-656989  -18313   -1383   14132  246968 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     7.267e+04  4.822e+03  15.071  < 2e-16 ***
GrLivArea                      -7.720e+02  5.400e+01 -14.296  < 2e-16 ***
GrLivArea:OverallQual           1.290e+01  6.814e-01  18.939  < 2e-16 ***
GrLivArea:OverallCond           4.572e+00  6.615e-01   6.912 7.14e-12 ***
GrLivArea:YearBuilt             3.761e-01  2.752e-02  13.667  < 2e-16 ***
GrLivArea:BedroomAbvGr         -4.132e+00  9.092e-01  -4.544 5.98e-06 ***
GrLivArea:SaleConditionAdjLand -2.162e+00  1.761e+01  -0.123  0.90227    
GrLivArea:SaleConditionAlloca   8.963e+00  6.640e+00   1.350  0.17729    
GrLivArea:SaleConditionFamily  -7.318e-01  6.196e+00  -0.118  0.90599    
GrLivArea:SaleConditionNormal   5.998e+00  2.596e+00   2.311  0.02099 *  
GrLivArea:SaleConditionPartial  9.060e+00  3.250e+00   2.788  0.00537 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38760 on 1449 degrees of freedom
Multiple R-squared:  0.7636,	Adjusted R-squared:  0.762 
F-statistic:   468 on 10 and 1449 DF,  p-value: < 2.2e-16
```

With reference to above shown codes, it can be depicted that Sale Condition was a confounding variable as it did not show statistical significance with LotArea and GrLivArea, hence we eliminated this factor from our study. The final sale price prediction model consisted of 6 explanatory variables LotArea, GrLivArea, OverallQual, OverallCond, YearBuilt, and BedroomAbvGr.

The following reference shows the multiplier linear regression on the 6 explanatory variables and response variable as sale price. The Multiple R-squared value is 0.7635, which is lower than the first regression we ran in our study but now the variable included are all highly statistically significant.



```
Call:
lm(formula = SalePrice ~ LotArea + GrLivArea + OverallQual +
    OverallCond + YearBuilt + BedroomAbvGr)

Residuals:
    Min      1Q  Median      3Q     Max 
-464899  -19369   -1618   16686  274843 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.256e+06  9.017e+04 -13.933  < 2e-16 ***
LotArea       8.756e-01  1.056e-01   8.293 2.49e-16 ***
GrLivArea     7.095e+01  3.066e+00  23.137  < 2e-16 ***
OverallQual   2.239e+04  1.171e+03  19.119  < 2e-16 ***
OverallCond   5.942e+03  1.005e+03   5.911 4.22e-09 ***
YearBuilt     6.000e+02  4.599e+01  13.046  < 2e-16 ***
BedroomAbvGr -1.112e+04  1.531e+03  -7.263 6.16e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38720 on 1453 degrees of freedom
Multiple R-squared:  0.7635,	Adjusted R-squared:  0.7625 
F-statistic: 781.6 on 6 and 1453 DF,  p-value: < 2.2e-16
```

When the model was arranged, the dataset was separated into test and train informational indexes. We investigated 40% of total 1460 observations to be in test and 60% of observation in the train set. The trained model was used for prediction of our response variable sale price. The following study is depicted in the figure below.

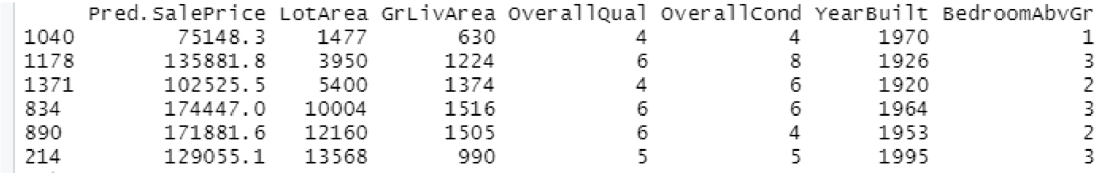
```
testindex=sample(1:1460,584,replace=FALSE)

> PropertySaleTest=PropertySales[testindex,]

> PropertySaleTrain=PropertySales[-testindex,]

> SalePredict_Model=lm(SalePrice ~
LotArea+GrLivArea+OverallQual+OverallCond+YearBuilt+BedroomAbvGr,data=PropertySaleTrain)
```

```
> SalePricePrediction=lm(SalesPredict_Model,PropertySaleTest)
```

```
> SalePrice_Pred=predict(SalesPredict_Model,newdata = PropertySaleTest)
```

```
>Pred_SalePrice=data.frame(Pred.SalePrice=SalePrice_Pred,LotArea=PropertySaleTest$LotArea,GrLivArea=PropertySaleTest$GrLivArea,OverallQual=PropertySaleTest$OverallQual,OverallCond=PropertySaleTest$OverallCond,YearBuilt=PropertySaleTest$YearBuilt,BedroomAbvGr=PropertySaleTest$BedroomAbvGr)
```

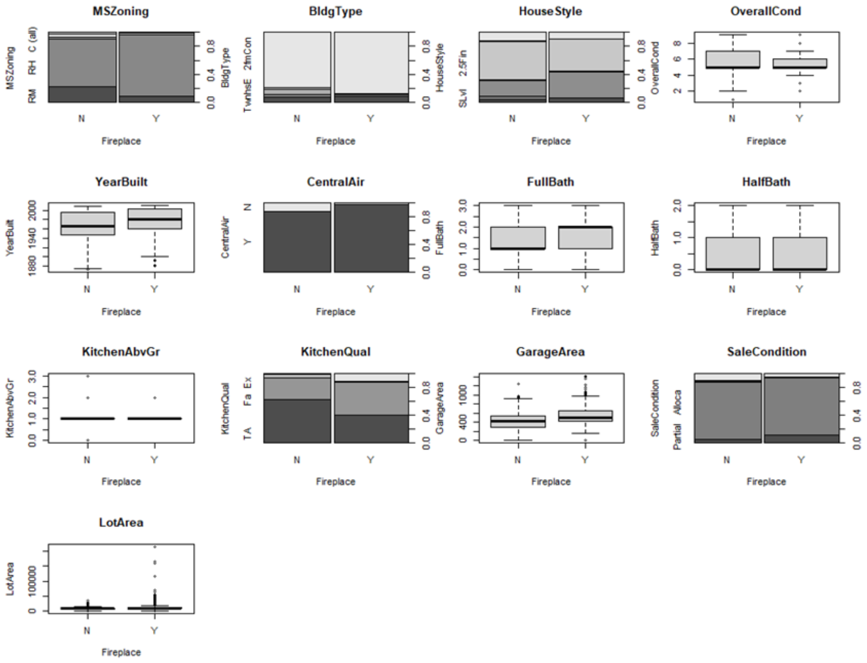
```
> head(Pred_SalePrice)
```



```
     Pred.SalePrice LotArea GrLivArea OverallQual OverallCond YearBuilt BedroomAbvGr
1040        75148.3    1477       630           4           4      1970            1
1178       135881.8    3950      1224           6           8      1926            3
1371       102525.5    5400      1374           4           6      1920            2
834        174447.0   10004      1516           6           6      1964            3
890        171881.6   12160      1505           6           4      1953            2
214        129055.1   13568       990           5           5      1995            3
```

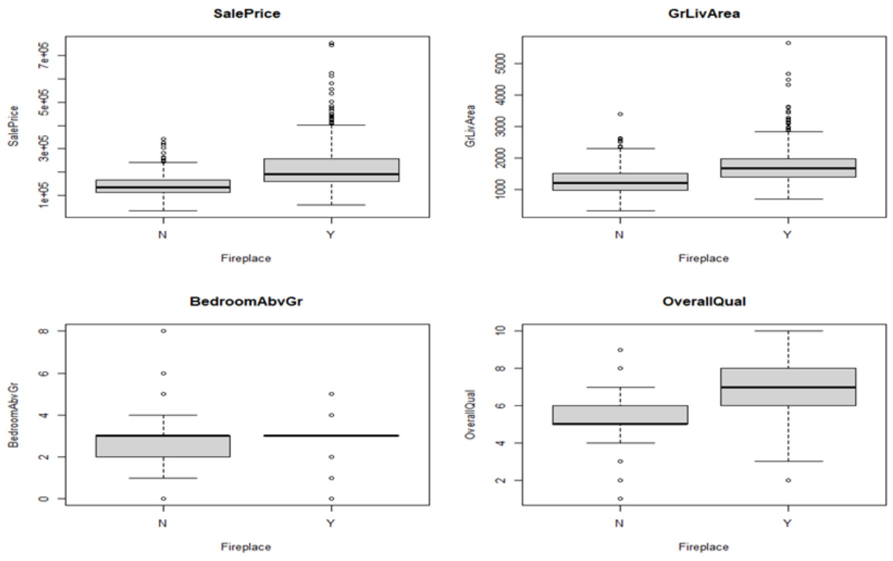
3. Classification Model for Fireplace Prediction

To develop a classification model for the prediction of fireplace in properties, we started off with installing the required Library package and input dataset. We converted the character variables to factor variables that were used to draw box diagrams. We used “as.factor” to convert “OverallQual, MSZoning, BldgType, HouseStyle, CentralAir, KitchenQual, Fireplace, SaleCondition” to factor variables, through which we plotted boxplot to observe correlation and selected high correlation variables as predictors.



Low Correlation Variables

Figure C



High Correlation Variables

Figure D

From Figure D, it is clear that the four variables are significantly correlated with the presence or absence of fireplaces.

At the same time, when OverallQual was converted into a factor variable, it was found that its correlation was different from that of OverallQual in numerical variables. It can be seen in Figure E and the corresponding code below.

As shown:

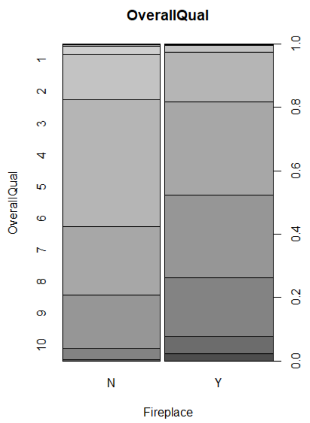
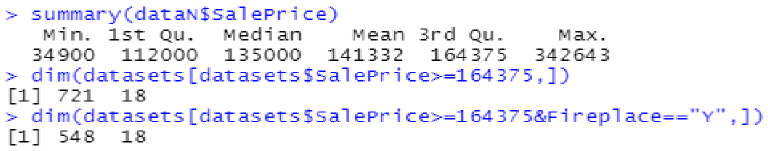
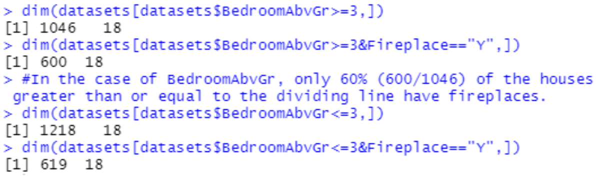


Figure E



```
> summary(dataN$SalePrice)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  34900  112000  135000  141332  164375  342643 
> dim(datasets[datasets$SalePrice>=164375,])
[1] 721  18
> dim(datasets[datasets$SalePrice>=164375&Fireplace=="Y",])
[1] 548  18
```

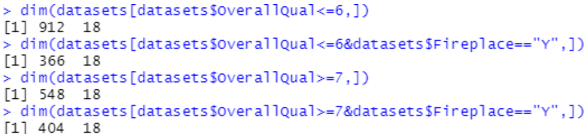
Similarly, the number of houses greater than or equal to the GrLivArea boundary is 653, of which 480 (73%) had fireplaces.



```
> dim(datasets[datasets$BedroomAbvGr>=3,])
[1] 1046   18
> dim(datasets[datasets$BedroomAbvGr>=3&Fireplace=="Y",])
[1] 600  18
> #In the case of BedroomAbvGr, only 60% (600/1046) of the houses
 greater than or equal to the dividing line have fireplaces.
> dim(datasets[datasets$BedroomAbvGr<=3,])
[1] 1218   18
> dim(datasets[datasets$BedroomAbvGr<=3&Fireplace=="Y",])
[1] 619  18
```

In the case of BedroomAbvGr, only 60% (600/1046) of the houses greater than or equal to the dividing line had fireplaces.

Only 50% of the homes below the boundary line have fireplaces. Taken together, the correlation is not so clear.



```
> dim(datasets[datasets$OverallQual<=6,])
[1] 912  18
> dim(datasets[datasets$OverallQual<=6&datasets$Fireplace=="Y",])
[1] 366  18
> dim(datasets[datasets$OverallQual>=7,])
[1] 548  18
> dim(datasets[datasets$OverallQual>=7&datasets$Fireplace=="Y",])
[1] 404  18
```

For OverallQual Houses of larger than 6 are 548, there are 404 (about 80%) have fireplaces

For OverallQual houses of 6 or less there are 912, 366 and less than a third have fireplaces.

In other words, when the OverallQual is greater than or equal to 7, the house will most likely have a fireplace, while when the OverallQual is less than or equal to 6, the house will most likely have no fireplace.

Next, we worked on Split.Train dataset and Test dataset
70% of dataset as Train dataset, 30% as Test dataset.

```
> dim(train)

[1] 1022  18

> dim(test)

[1] 438  18
```

For lda model and logistic regression we ran the following code on R programme.

```
>model.lda=lda(Fireplace~OverallQual+BedroomAbvGr+GrLivArea+SalePrice,train)
```

163TP, 51FP

73FN, 151TN

Test error:28.31%

From the previous illustrations, the correlations between OverallQual and BedroomAbvGr are not very good. Next, we tried different subsets to optimize predictor.

Subsets without overallqual:

```
>model.lda=lda(Fireplace~OverallQual+GrLivArea+SalePrice,train)
```

The confusion matrix:

   1 0

1 161 39

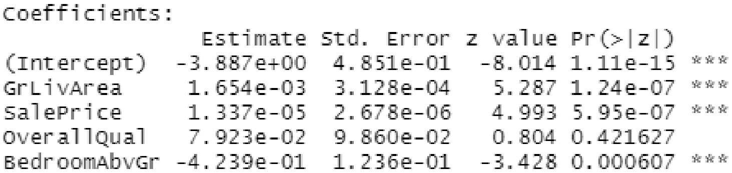
0 75 163

Test error: 26.02%

In summary, without OverallQual, the accuracy of the training set is improved.

Select the subset with higher accuracy after testing different subsets (BedroomAbvGr, GrLivArea, SalePrice)

To verify the appropriateness of this decision, the lda model was replaced with a logistic regression.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -3.887e+00  4.851e-01  -8.014 1.11e-15 ***
GrLivArea     1.654e-03  3.128e-04   5.287 1.24e-07 ***
SalePrice     1.337e-05  2.678e-06   4.993 5.95e-07 ***
OverallQual   7.923e-02  9.860e-02   0.804 0.421627    
BedroomAbvGr -4.239e-01  1.236e-01  -3.428 0.000607 ***
```

The discovery that OverallQual’s P-value is too high further confirms that OverallQual is not suitable for predictive fireplace.

Finally, when (BedroomAbvGr, GrLivArea, SalePrice) was selected as predictor, the prediction error rate was 26.02% as the performance of the model-glm.