**Predictive analytics**

**Discussing & predicting housing prices by using linear regression and decision tree modelling using R.**

|  |  |
| --- | --- |
| PART-A | FACTORS TO BE CONSIDERED WHILE EVALUATING AND PREDICTING THE REAL ESTATE PRICES. |
| PART B | PREROCESSING AND CLEANING THE DATA BY REMOVING NULL VALUES AND CONVERTING CATEGORICAL VARIABLES INTO NUMERICALS. |
| PART C | DEVELOPING AND EVALUATING LINEAR REGRESSION MODEL AND DECISION TREE MODELS FOR OUR HOUSING DATA SET. |

PART –A

**1-**

When buying a house there are many factors that can significantly affect the prices of the house.

As a normal buyer there are a few indicators of the property value that we should consider when planning to buy a house. We should first look around for the comp prices in the subdivision that we are planning to buy a house in, a comp is the house which is in the same subdivision as yours and have similar features. Another major thing is the location of the house, the more closer it is to a train or tram station higher will be its demand in the market. Another factor to consider is the square footage of the house or in simple words you can say the livable area in the house also, we should look for number of bedrooms and toilets in the house, the higher these numbers are the higher will be the price for a house. We should also look for the age of the house which is a crucial factor in determining the prices a newer house tend to have higher prices than a house which is has got more age due the fact that the newer houses do not require any significant repairs and upgrades which saves the buyer a major chunk of money. There is a long list of factors that should be considered but lastly I will discuss a couple of them which is economic indicators which is the presence of good employment options which in turn gives people purchasing power and vice versa.

**2-**

When considering above factors it is crucial to have access to the relevant information most of which is not available easily. To get data on the above factors we can approach the real estate agency that is involved in managing the lot/area you want to buy a house in. Also we can go to local councils to ask for age related documents of the house. Another source is the satellite images or online maps that pretty much show you the vicinity of the house, availability of train stations, schools, shopping centers etc. in the locality which affects the house prices significantly. Also to get data on interest rates we can access the news or economic journals and to predict if the government has any plans in the future to hike the interest rates. Also, comp data can also be collected form real estate agency websites like Zillow.com, realtor.com etc. Above sources can help gather data on the housing market in any locality. But the project which are under development, it becomes hard to get data for such projects and investing such projects require professional expertise in real estate field. So, gathering data on such houses is a bit difficult and requires a professional firm on the other hand for existing houses data is widely available in formats like, jpg, text, excel etc. Although in order to draw insights on a broader level this data need to be organized, validated and cleaned before it becomes fro decision making.

**3-**

Significant variables when building predictive model for housing prices can vary from model to model and developer to developer but the widely used variables for housing prices prediction are number of bedrooms, number of bathrooms, lot area, year built, basement, totalBSF, kitchen quality, Garage type, year sold and sold price.

References:

https://www.opendoor.com/w/blog/factors-that-influence-home-value

Author : Joe Gomez

2- <https://www.mashvisor.com/blog/factors-that-affect-property-value>

3-https://www.investopedia.com/best-real-estate-websites-5069964

Author: Bryan Carmody

**PART- B**

**1A-**

|  |  |  |  |
| --- | --- | --- | --- |
| **CATEGORICAL** | | **NUMERICAL** | |
| **Nominal** | **Ordinal** | **Discrete** | **Continuous** |
| LotConfig. | .LotShape  OverallQuality | TotalRmsAbvGrd | TotalBSF |
| DwellClass.  Utilities. | YearBuilt  LotConfig | PoolArea | LowQualFinSF |
| CentralAir.  Slope. | ExteriorCondition  BasementCondition | FullBath  GarageCars | LivingArea |
| PavedDrive.  LandContour. | KitchenQuality  OverallCondition | HalfBath  YrSold | SalePrice |
| GarageType. |  | BedroomAbvGr  Fireplaces | Id  Open porchSF |
|  |  | KitchenAbvGr  MoSold |  |

**1B-**

Stats summary

LotArea

Min. Median Mean Max. SD Skewness

1300 9478 10521 215245 10000.46

OverallQuality

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 5.000 6.000 6.103 7.000 10.000

OverallCOndition

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.000 5.000 5.000 5.576 6.000 9.000

TotalBSF

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0 799 999 1065 1304 6110 49

LowQualitySF

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 0.000 0.000 5.869 0.000 572.000

LivingArea

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

334 1131 1467 1517 1780 5642 22

FullBath

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 1.000 2.000 1.566 2.000 3.000

HalfBath

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 0.000 0.0000 0.3831 1.0000 2.0000

BedRoomABvGr

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 2.000 3.000 2.869 3.000 8.000

KitchenAbvGr

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 1.000 1.000 1.047 1.000 3.000

TotalRmsAbvGrd)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 5.00 6.00 6.52 7.00 14.00

Fireplaces)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.0000 1.0000 0.6142 1.0000 3.0000

GarageCars)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 1.000 2.000 1.771 2.000 4.000

PoolArea)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 0.00 0.00 2.77 0.00 738.00

OpenPorchSF)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 0.00 25.00 46.37 68.00 547.00

MoSold)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 5.000 6.000 6.319 8.000 12.000

YrSold)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2006 2007 2008 2008 2009 2010

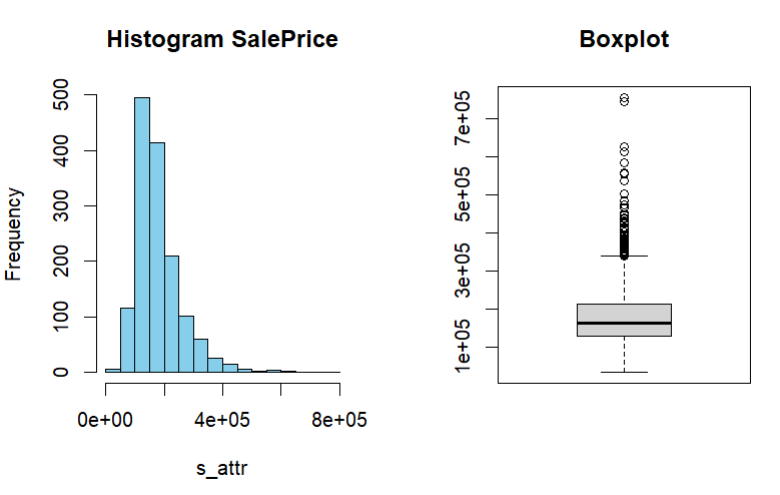
SalePrice)

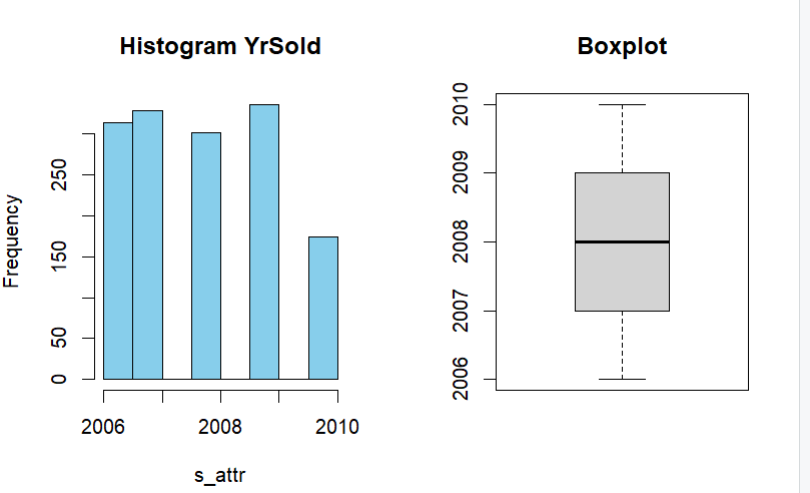
Min. 1st Qu. Median Mean 3rd Qu. Max.

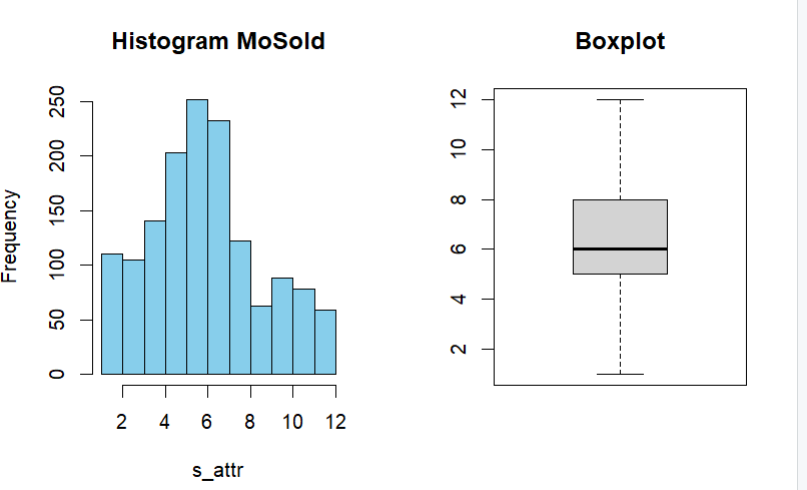
34900 130000 163250 181112 214000 755000

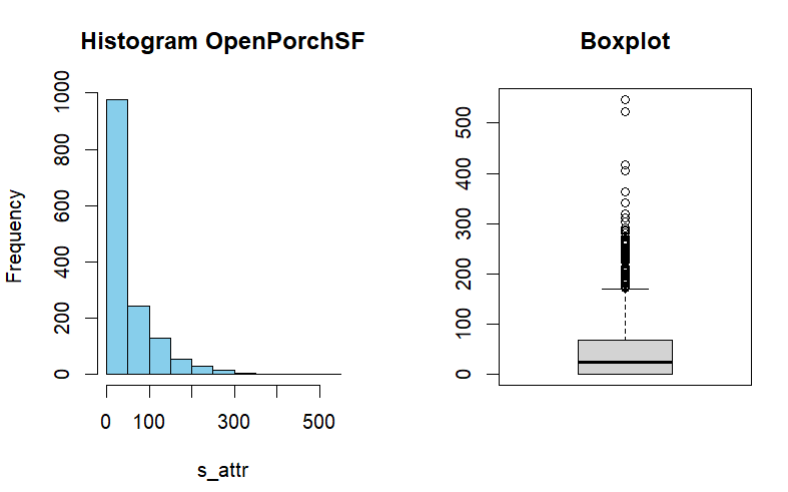
**PARTB**

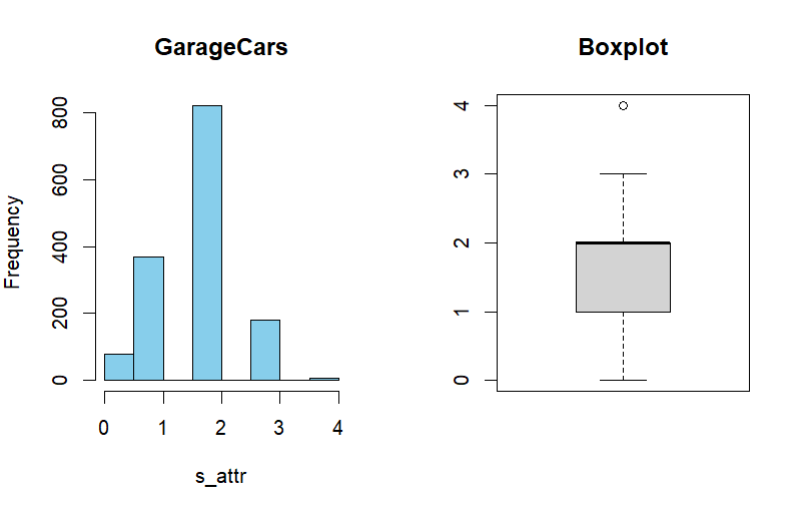
**Q2-(a,b)**

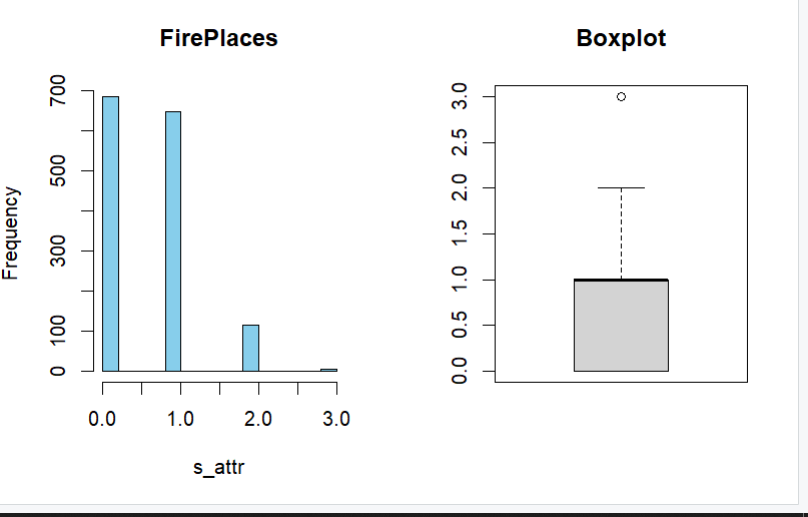


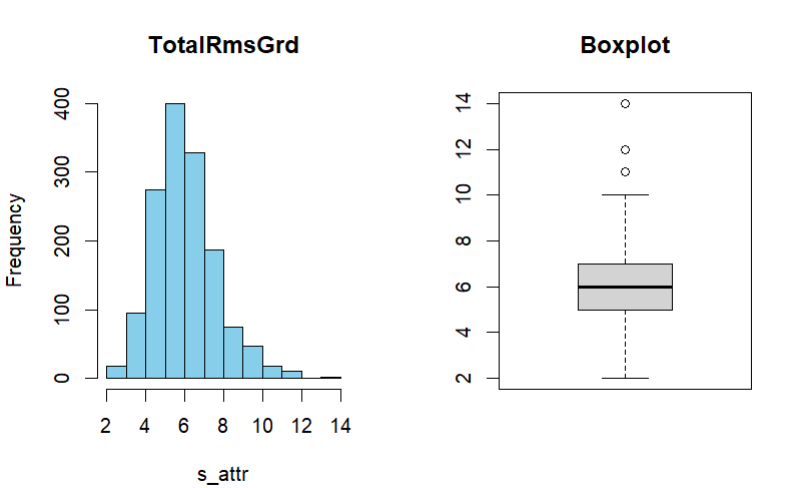


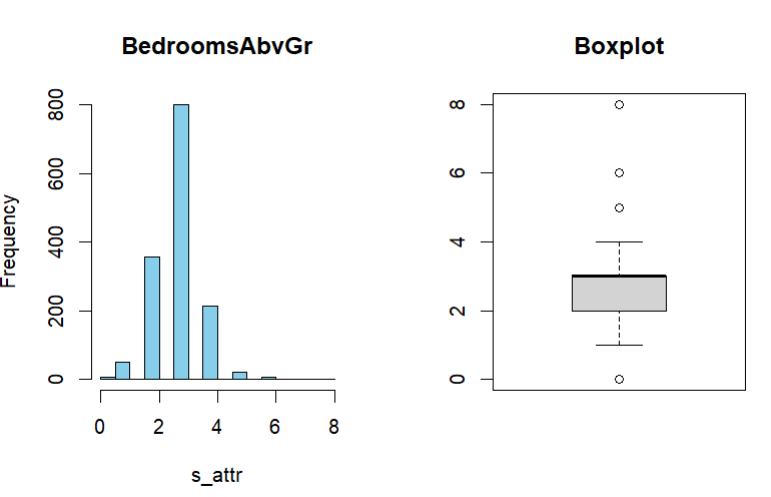


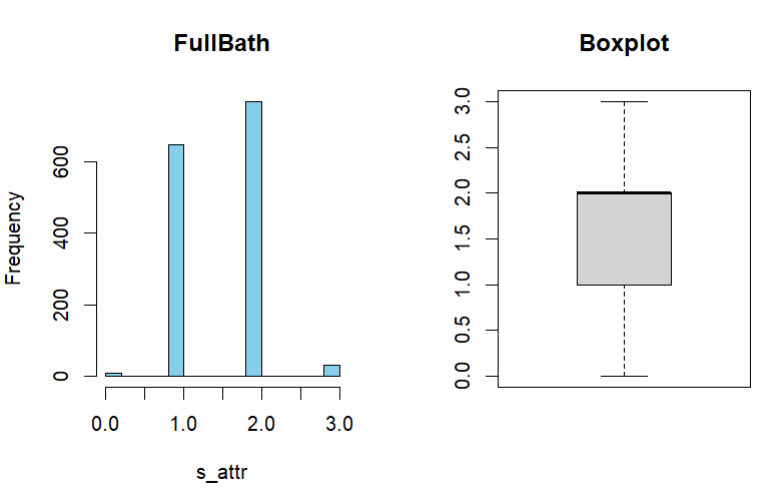


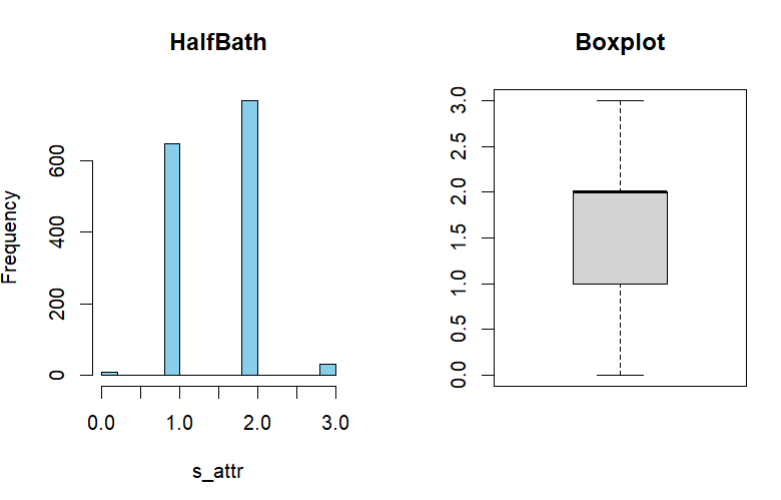


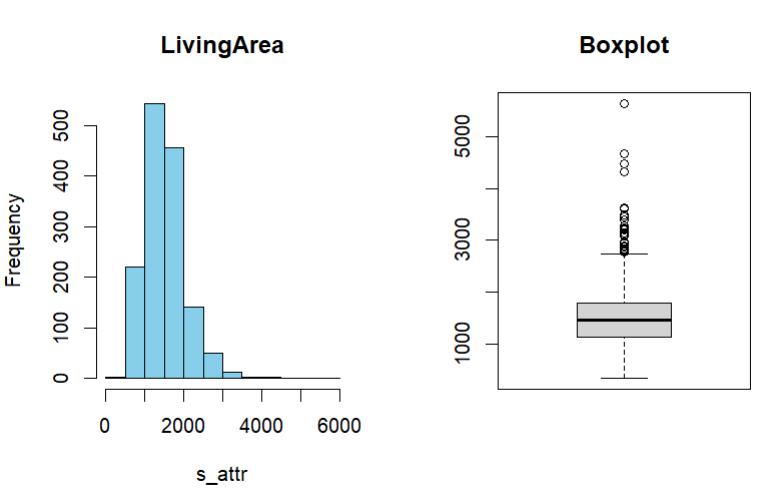


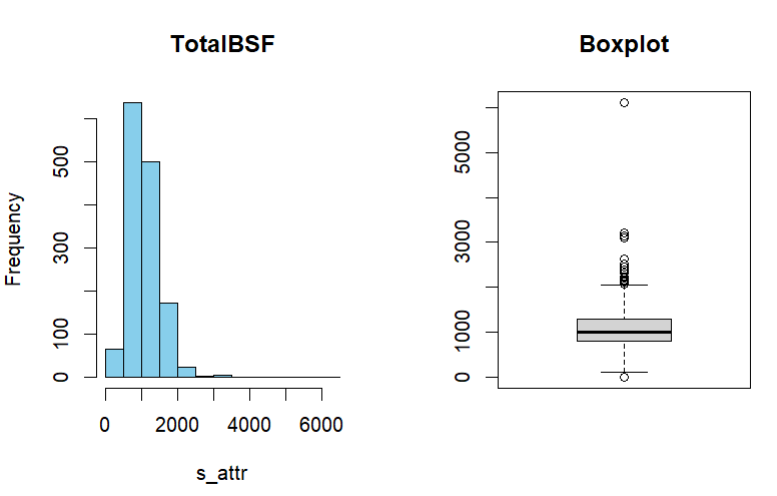


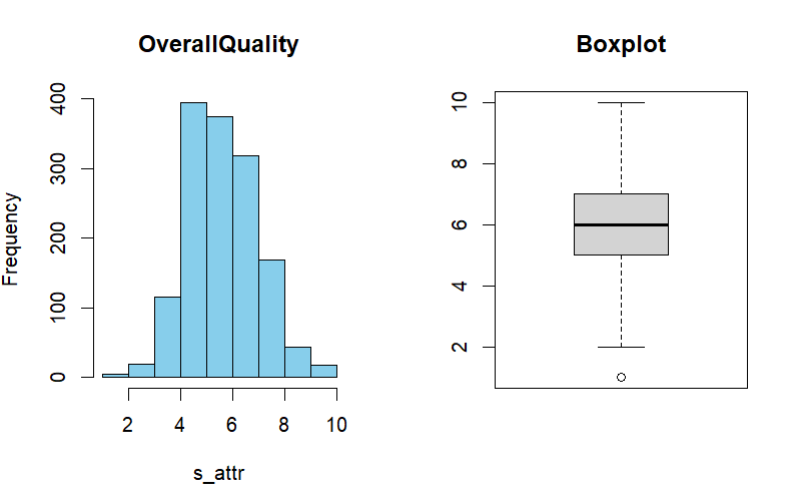


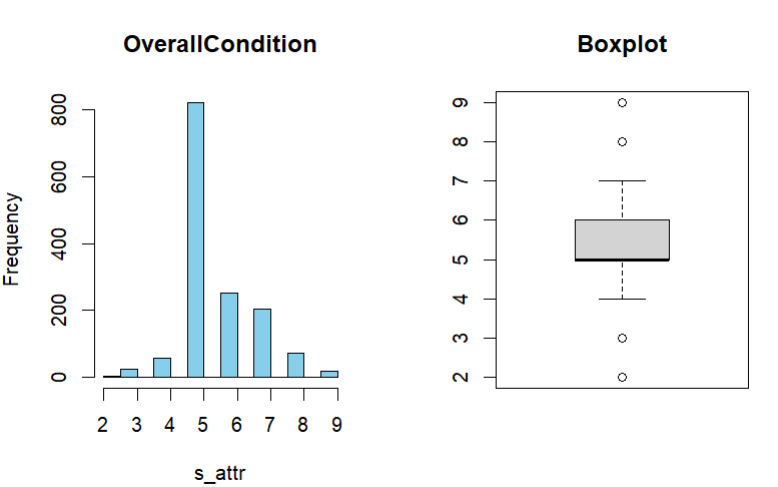












**Q2-C**-

1- The variables which have the highest variability are :-

Lot area, TotalBSF, Livingarea, SalePrice

2- As evident from the histograms and the statistical values observed we can say that following values are highly skewed (left/right):

SalePrice, OpenPorchSF, PoolArea, KitchenAbvGr, LowQualFinSF,

Moderately skewed are:

YearBuilt, TotalRmsAbvGrd, OverallCondition

3-Variables with extreme values are:-

SalePrice, OverallCondition, overallQuality, TotalBSF, LivingArea, HalfBath, FUllBath, RoomsAbvGrd, TotalRoomsGrd, Frieplaces, GarageCars, OpenPorchSF.

**PART B**

**Q3**

**a**- Using the R function colSums(is.na(df)) we found out that two columns have null values namely:

LivingArea 22

TotalBSF 49

GarageType 78

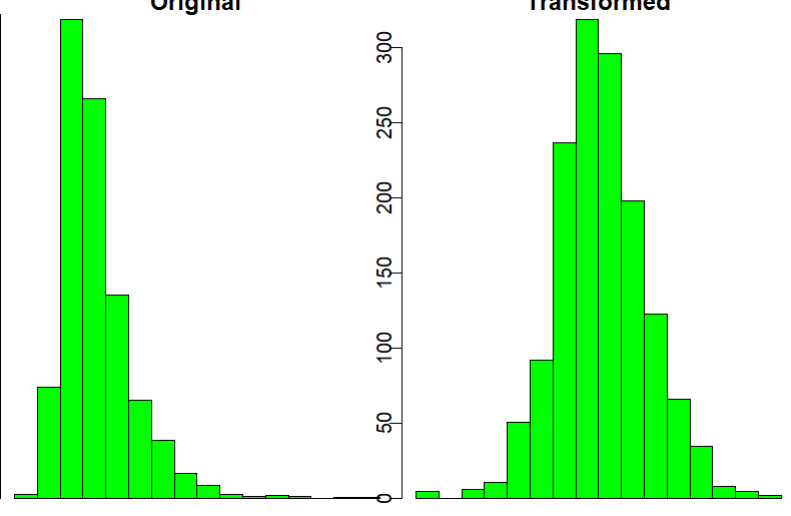
**b-** Three methods that we have learnt for dealing with the missing values in the dataset are:

1- We can replace the missing values with mean and mode for categorical variables.

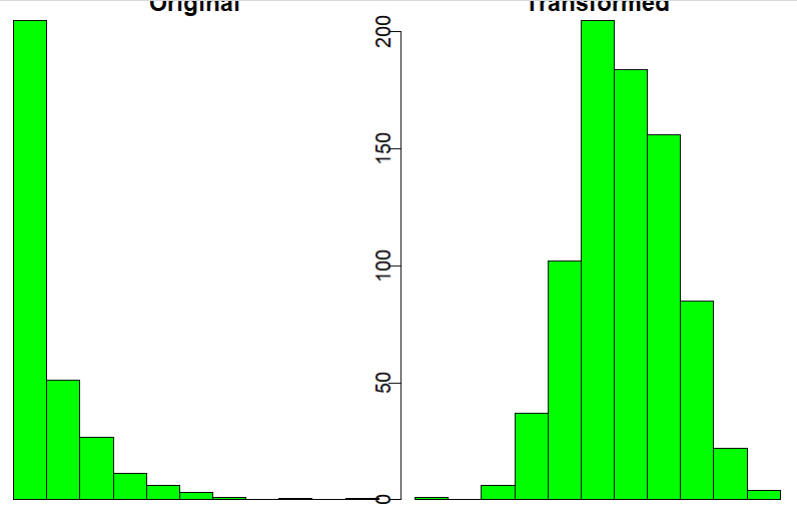
2- Other way is to delete the record or that row.

3- We can replace missing values with a specific value like 0.

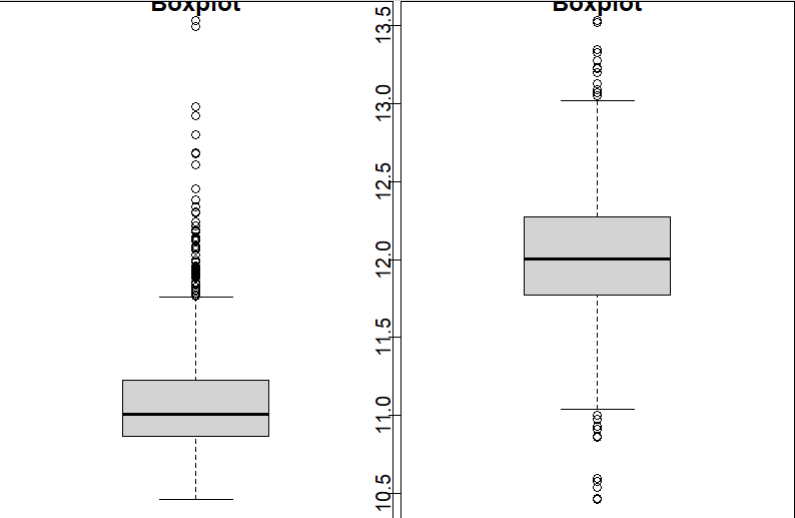
**e-** Logarithmic transformation for saleprice original vs. transformed

****

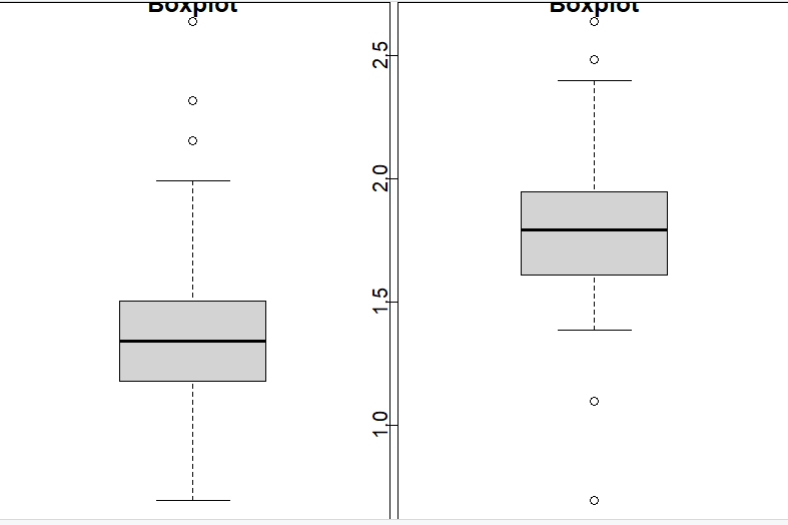
Logarithmic transformation for openporchsf original vs. transformed.



Before and after transformation boxplots of Saleprice



Before and after transformation boxplots of TotalRmsAbvGrd.



**Q4**

**a-** The methods for transforming categorical variables is to convert them into numerical variables, one hot encoding or by removing the data if feasible.

**b-**

#transforming categorical(nominal) into numeric

#by one hot encoding

slope\_tr1<- as.numeric(df$Slope == "Gtl")

slope\_tr2<- as.numeric(df$Slope == "Mod")

slope\_tr3<- as.numeric(df$Slope == "Sev")

util\_tr1 <- as.numeric(df$Utilities == "AllPub")

util\_tr2 <- as.numeric(df$Utilities == "NoSewr")

util\_tr3 <- as.numeric(df$Utilities == "NoSewa")

util\_tr4 <- as.numeric(df$Utilities == "ELO")

LotConfig\_tr1 <- as.numeric(df$LotConfig == "Inside")

LotConfig\_tr2 <- as.numeric(df$LotConfig == "Corner")

LotConfig\_tr3 <- as.numeric(df$LotConfig == "CulDSac")

LotConfig\_tr4 <- as.numeric(df$LotConfig == "FR2")

LotConfig\_tr5 <- as.numeric(df$LotConfig == "FR3")

DwellClass\_tr1 <- as.numeric(df$DwellClass == "1Fam")

DwellClass\_tr2 <- as.numeric(df$DwellClass == "2FmCon")

DwellClass\_tr3 <- as.numeric(df$DwellClass == "Duplx")

DwellClass\_tr4 <- as.numeric(df$DwellClass == "TwnhsE")

DwellClass\_tr5 <- as.numeric(df$DwellClass == "TwnhsI")

Centralair\_tr1 <- as.numeric(df$CentralAir == "Y")

Centralair\_tr2 <- as.numeric(df$CentralAir == "N")

LandContour\_tr1 <- as.numeric(df$LandContour == "Lvl")

LandContour\_tr2 <- as.numeric(df$LandContour == "Low")

LandContour\_tr3 <- as.numeric(df$LandContour == "Bnk")

LandContour\_tr4 <- as.numeric(df$LandContour == "HLS")

GarageType\_tr1 <- as.numeric(df$GarageType == "Attchd")

GarageType\_tr2 <- as.numeric(df$GarageType == "Detchd")

GarageType\_tr3 <- as.numeric(df$GarageType == "BuiltIn")

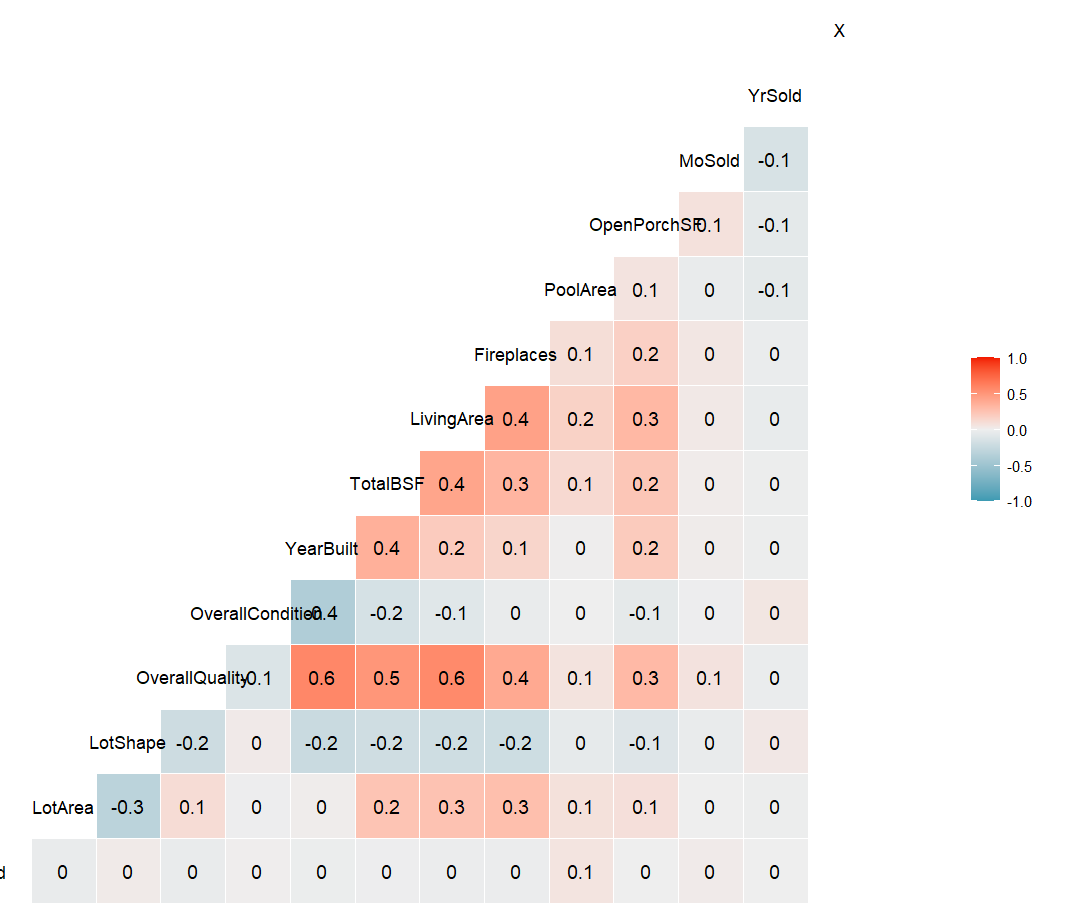
GarageType\_tr4 <- as.numeric(df$GarageType == "0")

GarageType\_tr5 <- as.numeric(df$GarageType == "Basment")

GarageType\_tr6 <- as.numeric(df$GarageType == "CarPort")

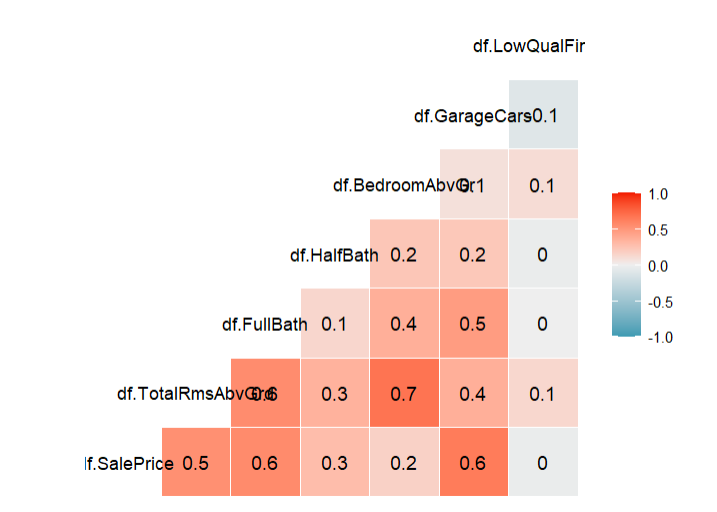
GarageType\_tr7 <- as.numeric(df$GarageType == "2Types")

**5-a**



**b-** I removed TotalRmsAbvGrd, FullBath, Halfbath, BedRoomAbvGr, Garagecars, HalfBath, LowQualFinSF because they had coeff of correlation greatehr than 0.5 which is quite high.

Also the same is evident from the plot obtained.

**c- **

Above correlation graph shows the distribution of selected variables against the target variable.

GarageCars and BedroomAbvGr show a moderate correlation with the target variable SalePrice.

**PART C:**

**1a-** I made a linear model using the less correlated variables which I found out using the correlation plot for making the model.

LivingArea, OpenPorchSF, FullBath, PoolArea

And the results are as follows:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10402.1556 5803.290105 1.792458 7.337305e-02

LivingArea 79.0523 4.009499 19.716251 5.764328e-73

OpenPorchSF 134.6629 28.663362 4.698085 3.007427e-06

FullBath 30743.7705 4024.909014 7.638377 5.292239e-14

PoolArea -113.8413 39.027248 -2.916970 3.616492e-03

Above is the summary of the model.

“R Squared: 0.537342921291359", this R squared suggest a moderate fit.

**1b-**

**1c- Model formula**

Ist model-

y ~ 10402.16 + 79.05 \* LivingArea + 134.66 \* OpenPorchSF + 30743.77 \*

FullBath + -113.84 \* PoolArea

2nd model

**C-2-a-**

library(rpart.plot)

library(GGally)

setwd("C:/Users/ASUS/Desktop/La\_trobe/Assignment and tutorials/Predic\_analy")

df <- read.csv("C:/Users/ASUS/Desktop/La\_trobe/Assignment and tutorials/Predic\_analy/Assignment1/HousingValuationTest-V2.csv")

df <- replace(df, is.na(df), 0)

colSums(is.na(df))

df\_D <- data.frame(df$SalePrice, df$LotArea, df$LotShape, df$Utilities, df$OverallQuality, df$LivingArea)

View(df\_D)

smp\_size <- floor(2/3 \* nrow(df\_D))

nrow(df\_D)

set.seed(2)

df\_D.selected <- df\_D[sample(nrow(df\_D)), ]

sample(5)

SalePrice.train <- df\_D.selected[1:smp\_size, ]

SalePrice.test <- df\_D.selected[(smp\_size+1):nrow(df\_D.selected), ]

#Specifying target and input variables

formula = SalePrice.train$df.SalePrice ~.

SalePrice.train

dtree <- rpart(formula, data=SalePrice.train, method="anova")

#Checking the feature importance

dtree$variable.importance

#Visualizing the decision tree

rpart.plot(dtree, type = 4, fallen.leaves = FALSE)

**C-2-B**

**#**Model assessment and making new trees using pruning.

print(dtree)

#Making Predictions and Assessment

#Predictions

predicted.SP1 <- predict(dtree, SalePrice.test)

#Model assessment

error <- SalePrice.test$df.SalePrice - predicted.SP1

rmse <- sqrt(mean(error^2))

print(paste("Root Mean Square Error: ", rmse))

printcp(dtree)

dtree$cptable[which.min(dtree$cptable[,"xerror"]),"CP"]

pruned\_dtree <- prune(dtree, cp = 0.04)

#decision tree with cp value 0.04

rpart.plot(pruned\_dtree, type = 4, fallen.leaves = FALSE)

predicted\_pruned.df.SalePrice <- predict(pruned\_dtree, SalePrice.test)

error\_new <- SalePrice.test$df.SalePrice-predicted\_pruned.df.SalePrice

rmse\_new <- sqrt(mean(error\_new^2))

print(paste("New Root Mean Square Error: ", rmse\_new))

#decision tree with cp value 0.022

pruned\_dtree <- prune(dtree, cp = 0.022)

rpart.plot(pruned\_dtree, type = 4, fallen.leaves = FALSE)

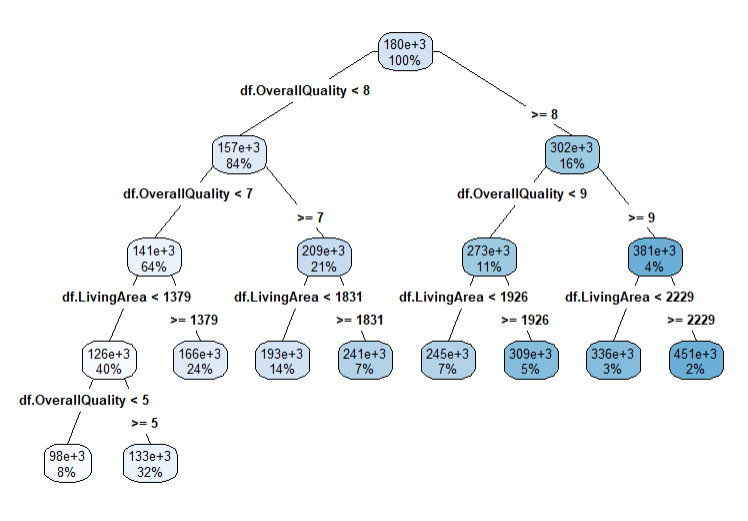
predicted\_pruned.df.SalePrice <- predict(pruned\_dtree, SalePrice.test)

error\_new <- SalePrice.test$df.SalePrice-predicted\_pruned.df.SalePrice

rmse\_new <- sqrt(mean(error\_new^2))

print(paste("New Root Mean Square Error: ", rmse\_new))

**C-2-C**

****

#about the tree

RMSE: 45664.8425330694

The tree splits the decision of SalePrice based on the OverallQuality of the house. As we can see the primary nodes are split based on the same metric and one node has Overallqual < 8 and other child node depicts the values value >=8. Further the split is based on the measure of living area.Also the average house price of the entire data has come out to be roughly “$180k” and increasing as the overallqulty and living area increases and also the color shifts to darker blue shade as a consequence of these values.

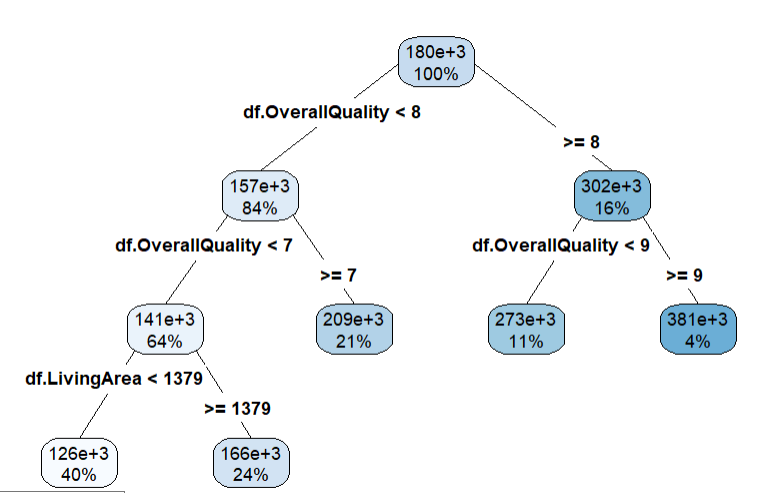
Also the highest price given by the model is for the houses which comprises o only **2% data** for which the prices are around **$451K**. For this segment of results the variables matrix is the highest as we can see the **ovrallqulty is >= 9** and the living area is also the highest which **is >= 2229 sq units**. This leads to the conclusion that the house prices are directly proportional to area and overallqulty out of the chosen variables for model making.

**#Selection of trees.**

On the analysis of first decision tree we found out that out of many chosen variables algorithm found out only LivingArea and OverallQuality to be considered for the model as other variables had not a significant impact on the decision tree their contribution was roughly 1/6th of the most influential variable.

Other decision trees made by pruning are given below.2nd decision tree with cp value 0.04

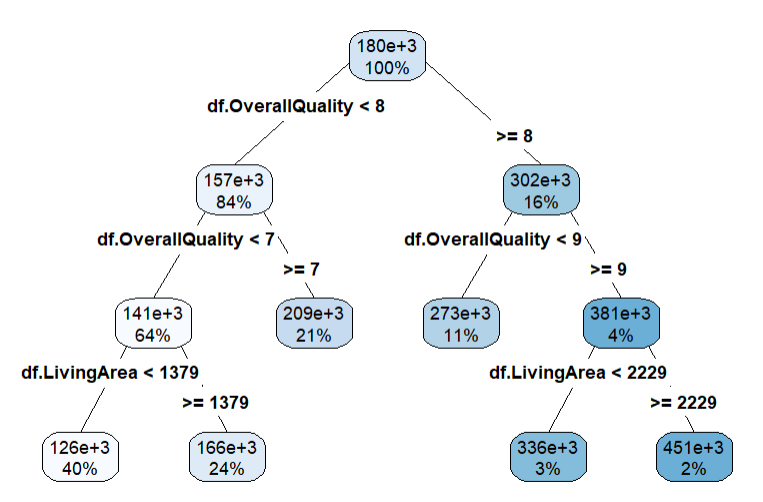
picked up from the summary.



RMSE: 50432.2110472125"

In this tree the average house price for data is also around 180k and 381K for the 4% of the data or 4% houses.

3rd decision tree with cp value 0.022



RMSE: 48966.5657214534"

This tree has the closest results with the first tree we made. The average price for **100%** of our data is roughly **$180k** and the highest price for **2%** of the data is roughly **$451K.** Also by evaluating the parameters we found out that this tree is the best considering the complexity of the tree and effectiveness in the result.

After the analysis we can conclude that the RMSE of all the models is slinging around 48000 +/- 2500, so we can pick up the last model as it has not got a lot of nodes and looks simple for decision making and has the RMSE roughly equal to mean of the three.

**PART-3-A**

We need to build several models as we can see the impact of different variables on the result of the model. Lower the errors higher would be the accuracy of the model. Since it is not feasible to accommodate all the variables in a single model as that might lead to over fitting of the model. So because of these reason it becomes important to consider different parameters to find the best fit.

**3-B**

Decision tree looks to be a better choice for the purpose as it very address our issue by predicting the direct value with higher accuracy. In this type of model we are not required to preprocess the data much which saves time and this model is also easier to understand by the stakeholders and they can very well understand the parameters we have considered without any technical expertise.