IE 7280 STATISTICAL METHODS IN ENGINEERING

FINAL PROJECT

Contents:

BY :

ADITYA CHANDRASEKARAN

INTRODUCTION AND PROBLEM STATEMENT

We are given a dataset consisting of house details in Ames, Iowa . Our problem statement is to build a predictive model that can be used to guide our customers on where, when and how to build a property with minimum expense to maximize profit during resale

DESCRIPTION OF THE DATA

The dataset consists of the various features of the houses in AMES IOWA and the sales spread 2006-2010

The dataset consists of 3000 transactions that has happened at Ames and 80 features of the houses that has been sold. The columns consist of 36 numerical features and 43 categorical features . Few important features that has been collected are mainly Type of dwelling involved in the sales, building types, Zones of the household , Neighborhood, Basement Conditions, Garage maintenance, year the house was build, Year of the house sold etc. These features must be investigated individually and in-collaboration with other features to find out their influence on the marketplace and buying patterns so that it can help us in building a recommendation system for our customers based on their budget

CODE IMPLEMENTATION STEPS

Step 1 :

We Examine the data for Null or Missing values and found quite a few columns with them. We decided to keep a threshold of 40 % and any column that has more than 40% of their values as null was eliminated. We found that 5 columns had more than 40% Null or missing values and the 5 columns were removed from the data

For the columns which had less than 40 % missing values we chose to substitute the values with the mode of the column for categorical and mean for numerical.

	NullCount <int>	PctNull <dbl>
PoolQC	2909	1.00
MiscFeature	2814	0.96
Alley	2721	0.93
Fence	2348	0.80
FireplaceQu	1420	0.49
LotFrontage	486	0.17
GarageYrBlt	159	0.05
GarageFinish	159	0.05
GarageQual	159	0.05
GarageCond	159	0.05

We also found the variance of the columns and decided to remove columns which don't have much variance as they would not contribute much to our model and would not be of any significance in our analysis.

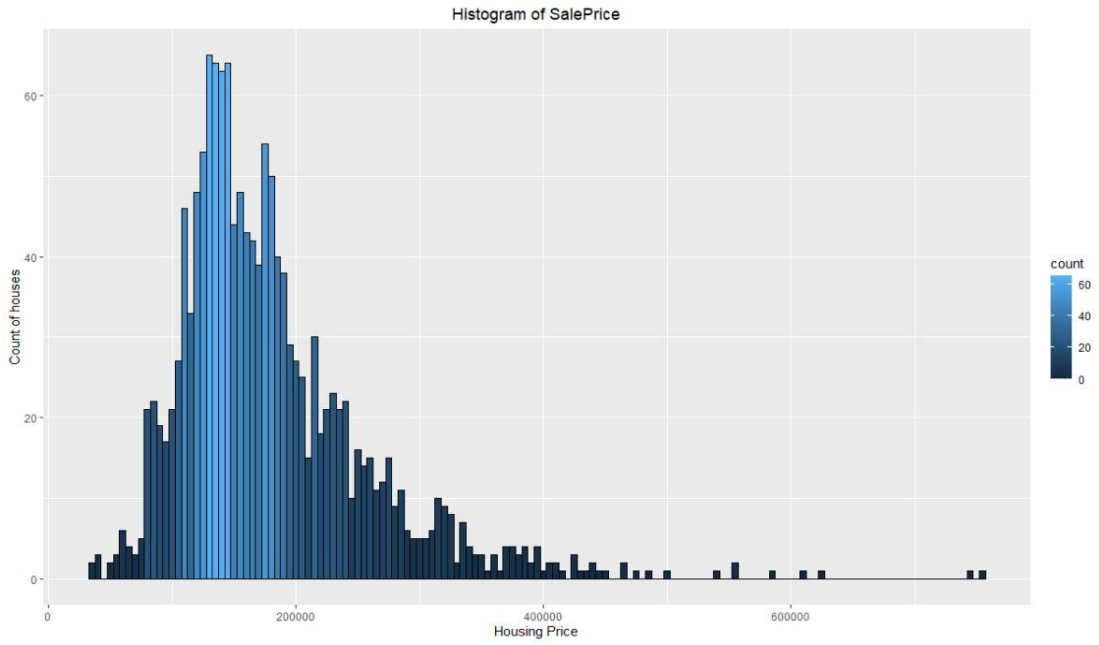
Step 3:

Exploratory Data Analysis (EDA)

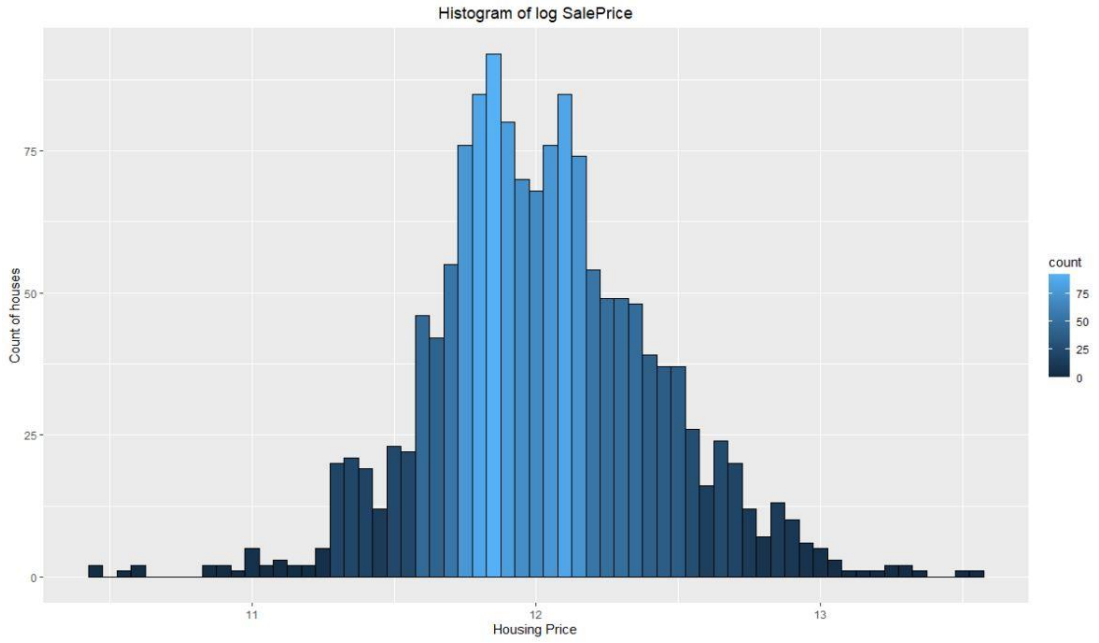
Numerical features

We visualized the target variable so that we can get a better sense of the data's distribution. We found that the target variable was heavily right skewed. Our model assumes that the data is normally distributed. So, we decided to log transform the data so that we make the data normally distributed.

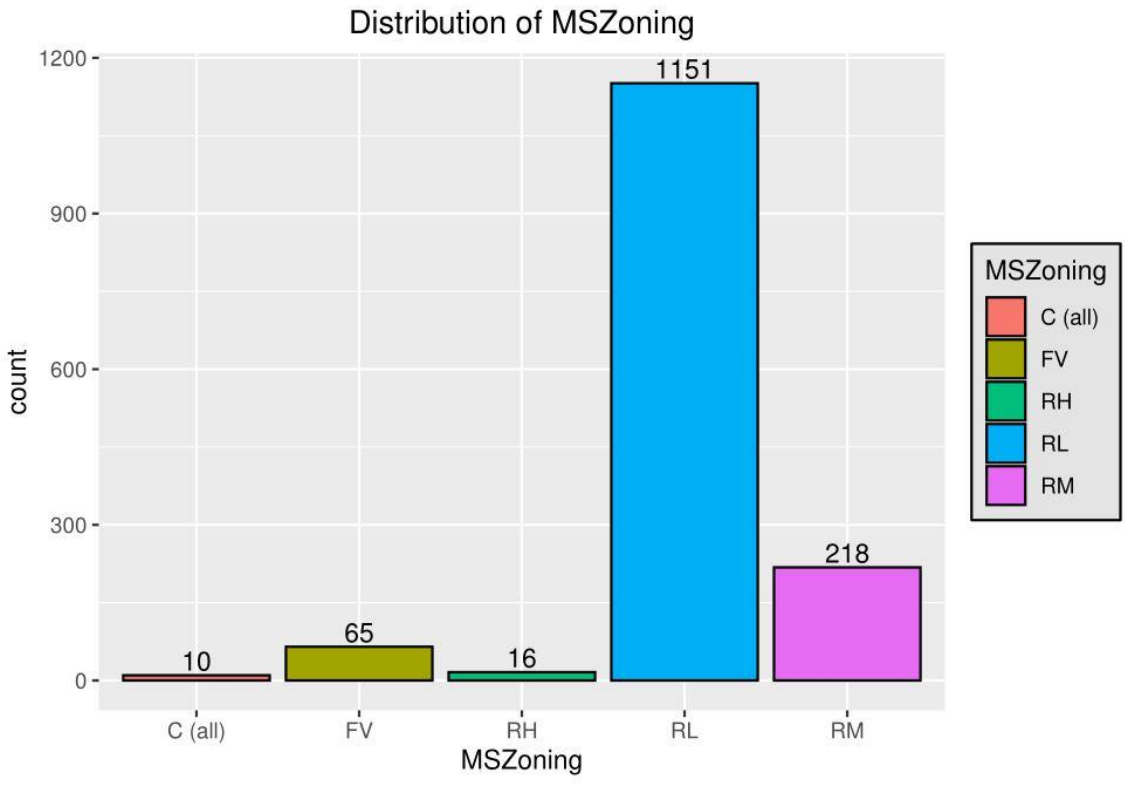
Before transform:



After transformation:

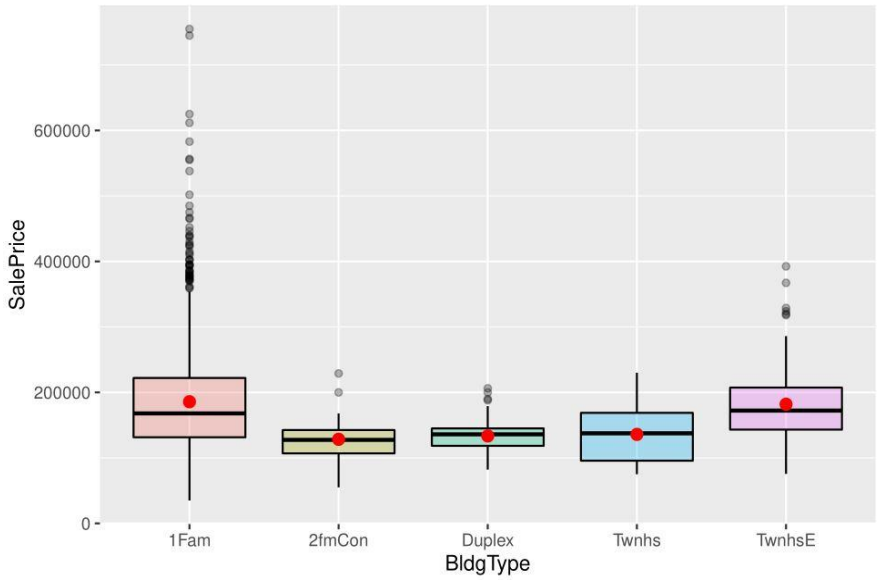


We then plot the Distribution of MSzoning to get an idea which type of house sold the most. We found RL houses sold the most which tells us that people prefer lower residential type houses.

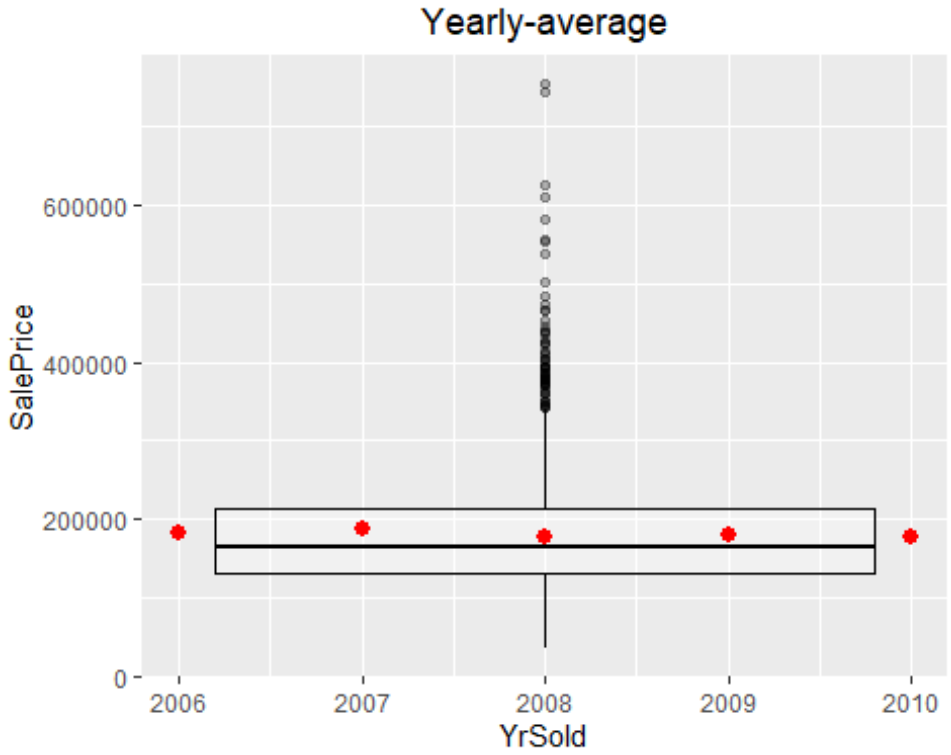


We have plotted the average sales price of each type of house we found that single family homes had an average selling price higher than that of the others.

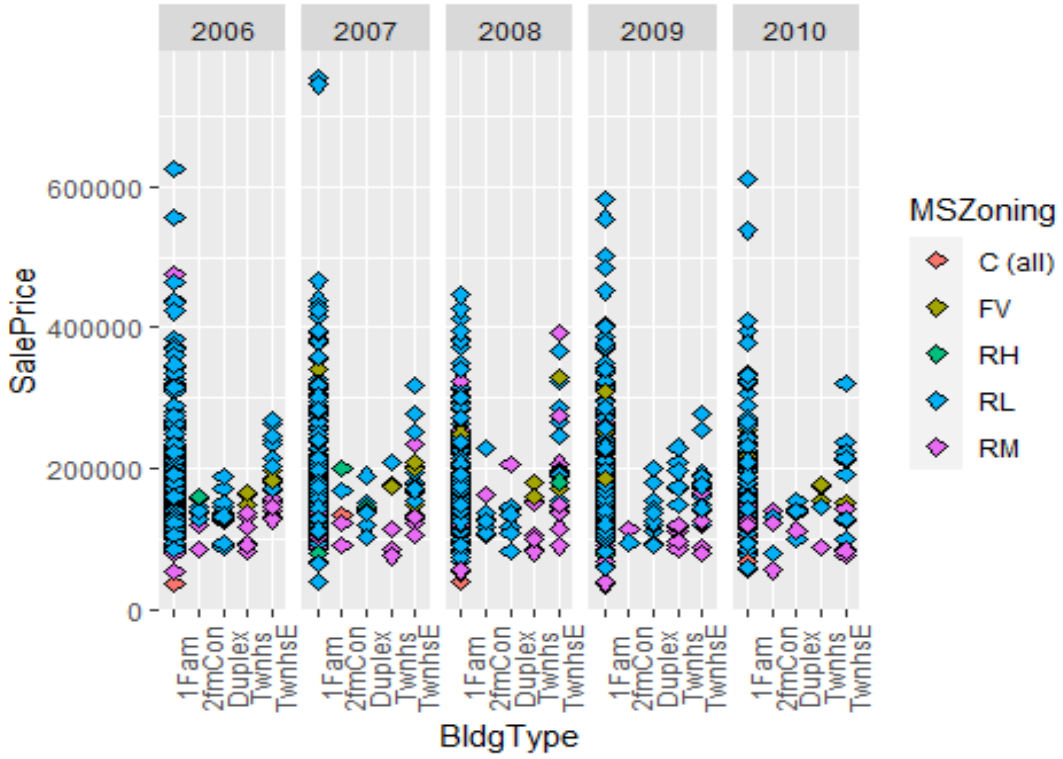
Figure 4 Boxplot of SalePrice by MSZoning



Let's suppose a client approaches us and tells us that he has around 400000 dollars to invest and asks us which house would be a good investment. We see that a single family house has the highest sales price and we would most likely recommend the client to go with a single family house.



Yearly average graph is a box plot of average sale price of houses in each year and we can find huge pruchases (outliers) that has been made for the year 2008



Through above graph we understood that 1-Fam (Single Family) houses has been an high seller each year and their predominantly bought in Low-Residential Zones

Then we also performed EDA on specific columns like utilities has only 1 value for NoSeWa and the rest of the values are all AllPub. We can drop this feature from our dataset as the house with 'NoSeWa' will not help with any predictive modelling.

Step 4 :

Feature Engineering

From the data we observed that the GarageType Column has several classifications. However, the classification can be grouped and made into 2 major classifications. We can group attchd, builtin category as a single group called 'attached' and the rest of the classifications under 'Not-attached'. This will give the model a better performance and reduce overfitting , as well as capture the essential information of the data.

```
completeData$GarageAttchd <- ifelse(completeData$GarageType == "Attchd" |
completeData$GarageType == "BuiltIn", 1, 0)
completeData$GarageAttchd <- as.integer(completeData$GarageAttchd)
completeData$GarageDetchd <- (completeData$GarageType == 'Detchd') * 1
```

We do the same thing for other columns with similar cases such as Electrical and BSmtgrade.

In some cases, we change the factors to numeric values.

```
#Changing year to use them in the model
completeData$yrs_since_built <- 2010 - completeData$YearBuilt
completeData$yrs_since_sold <- 2010 - completeData$YrSold
completeData$yrs_since_remod <- completeData$YrSold - completeData$YearRemodAdd

#creating grades
completeData$BsmtGrade <- ((completeData$BsmtQual)*1) * ((completeData$BsmtCond)*1)
completeData$GarageGrade <- completeData$GarageQual * completeData$GarageCond
completeData$TotalArea <- completeData$GrLivArea + completeData$TotalBsmtSF
completeData$MSSubClass<-completeData$MSSubClass*1
```

The dataset that we have is till 2010 and sales made from 2006 to 2010 so we use 2010 as the present year and subtract with the respective year the house was built. We stored it as 'year since built' column. We assume that this column will have a negative correlation as the resale value goes down (after discounting inflation) with the number of years since it's built increases.

There are many features describing basement in the dataset, so we perform feature engineering and collaborate all the other features into a single column called as basement grade. This basement grade will give an overall idea of the basement and its maintenance condition. Similarly, we do the same steps for garage as well.

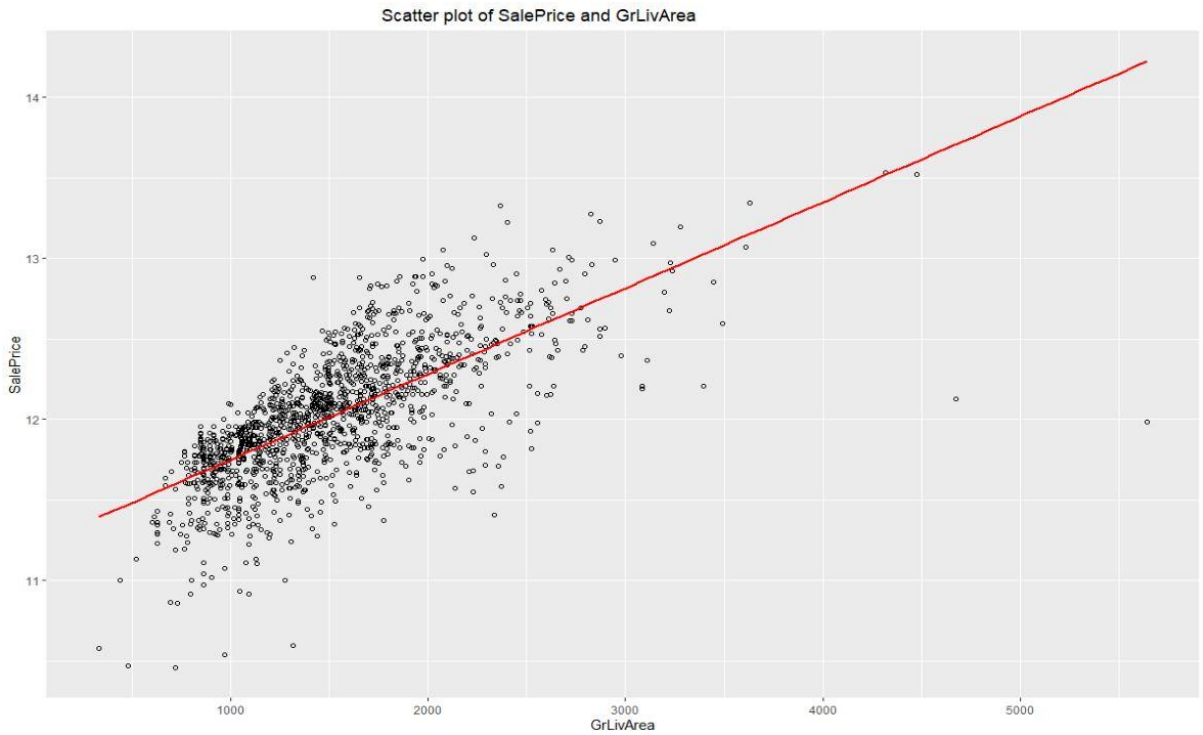
Step 5 Feature Selection :

Remove highly correlated features

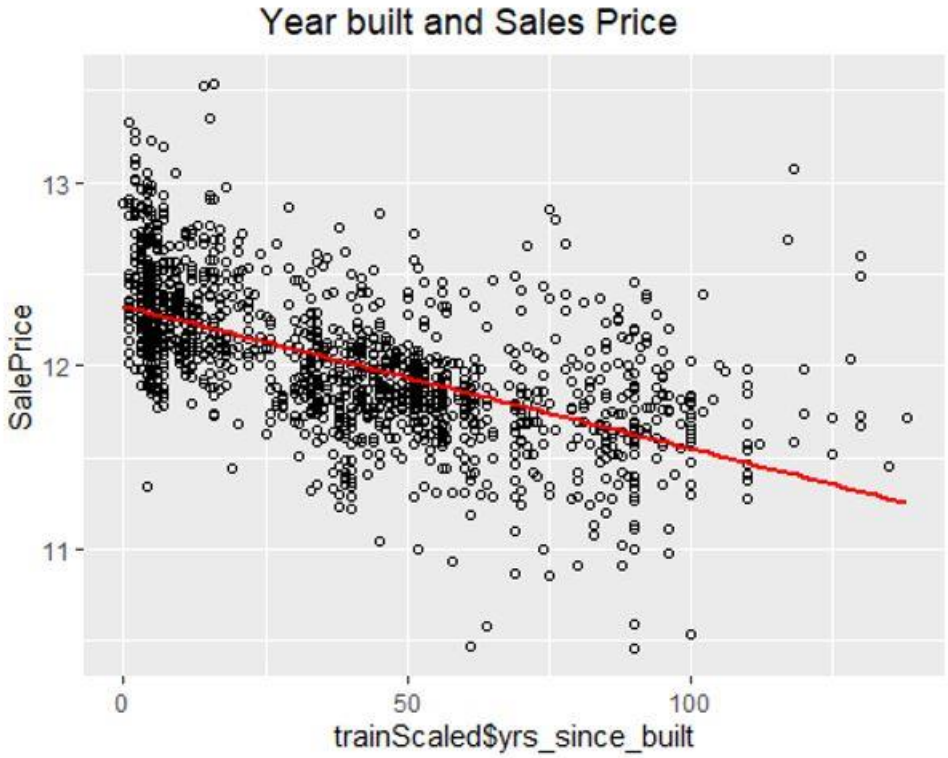
The data contains many features that are either redundant(strongly correlated, linear dependent) or irrelevant, these features can be removed without incurring much loss of information. Multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. Thus, all the features that show more than 70% of correlation are removed. This ensures that our model functions properly.

The number of columns with more than 70% correlated columns is 10. Which were all removed.

Then we plot several scatterplots of salesprice with other columns and we include a regression line to see how the trend is.



We see an upward trend in this scatter plot. Hence, we can say that the Living Area and salesprice have a positive correlation



We see a downward trend in the plot with years since built and salesprice. Thus, we can infer that as the years increase the salesprice decreases. We can use this to predict the future price of the house.

From the remaining plots we found that GrLivArea, TotalBsmtSF, TotRmsAbvGrd, and GarageArea are positively correlated with SalePrice, which means with the increase of GrLivArea, TotalBsmtSF, TotRmsAbvGrd and GarageArea, the SalePrice also increases. TotalBsmtSF has more concentrated distribution than others

MODEL FITTING:

After all the preprocessing is done , the data is split into Train and Test and then we scale it.

We first choose a Multiple Linear Regression Model as we have done all the preprocessing required such as :

- linear relationship between the dependent and independent variables
- The independent variables are not highly correlated with each other
- The variance of the residuals is constant
- Independence of observations

Multiple Linear Regression Model is used to identify the strength of the effect that the independent variables have on a dependent variable. Multiple Regression Analysis helps us to understand how much will the dependent variables change when we change the independent variables. Multiple Linear Regression Analysis predicts trends and future values.

Thus, we select a Multiple Linear Regression model for our prediction.

Then we split the train data into Ames_traindata and ames_validationdata. So that we can validate our model on the training set.

Now fit a Multiple Linear Regression Model to predict the sales price given the necessary features.

We evaluate our model and we find that we have a really good R-squared. And a P-value of 0.000000000000022 which is very less but also very good. Thus, the model is significant at the 99% level also.

```
## Multiple R-squared:  0.8872, Adjusted R-squared:  0.8829
## F-statistic: 207.3 on 36 and 949 DF,  p-value: < 0.00000000000000022
```

We are using the RMSE metric to evaluate the performance as it's a numerical value and RMSE would give us the best test results

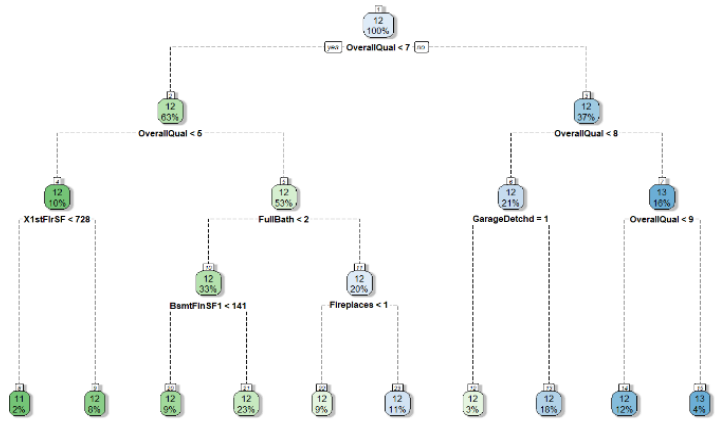
RMSE 0.1703

Which is quite a low value and implies that our model is very accurate.

With this model we predict the salesPrice for our test set.

Problem Solved :

Thus problem we are solving here is that, given a client wants to know the sales price of a house in a particular year with the type of house, garage requirement and any other feature that is available (in our dataset) and we want to suggest him a house as per his requirement and maximize his profit if he decides to sell it in a particular year. Let's assume for the sake of this argument that a client gives the specification which is exactly a entry in our test set then all we have to do is give them the necessary salesprice. We can do this for any custom values the clients may require using our model to predict the sales price.



We have also implemented a decision tree model. So that we can compare the two models.

We follow the implementation in the same way as above for decision model and we found a final RMSE value of 0.2213. This model is also good. And we visualize it to gain an idea at which values the split occurs. We are only implementing this model to get a visual idea of where the spilt occurs.

The Multiple Linear Regression model is much better, and we have solved our business problem using it.

Conclusion

From our above analysis we conclude that we have found a way to provide the clients with the information as to which property to invest in or purchase and when and how much returns can they expect on selling the property. We used a Multiple Linear Regression model with RMSE 0.1703 to predict the sales price value and provide our client with the necessary details.