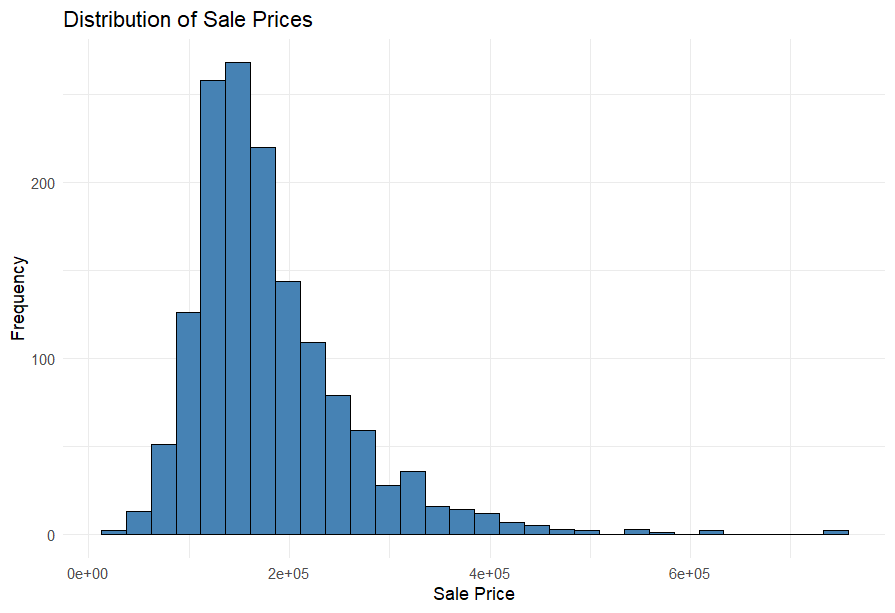
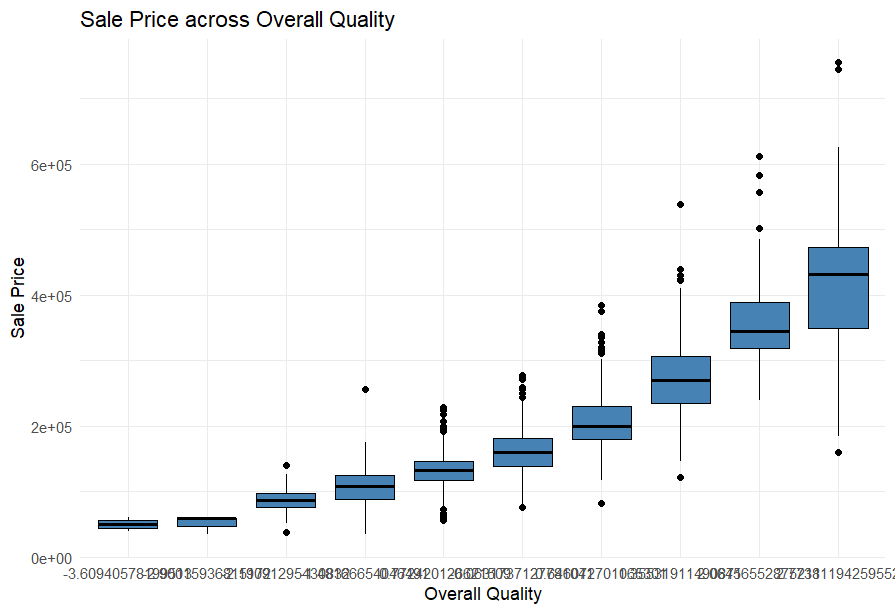
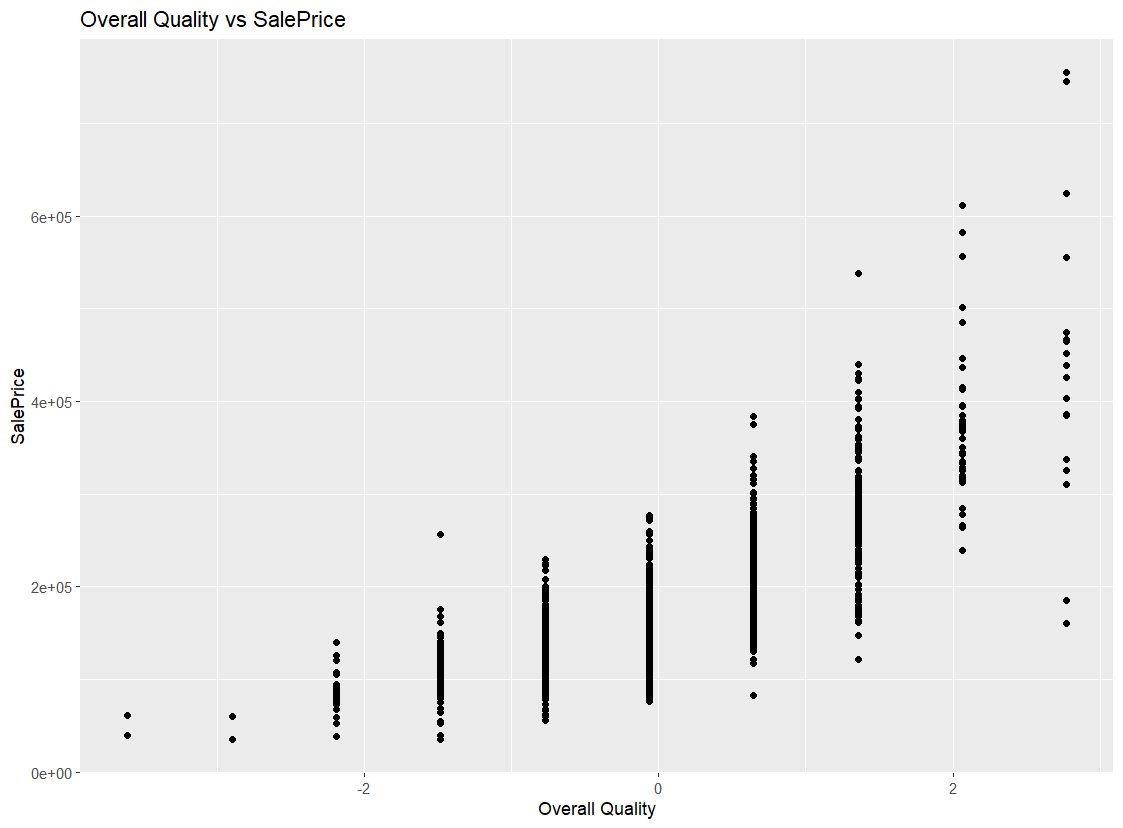
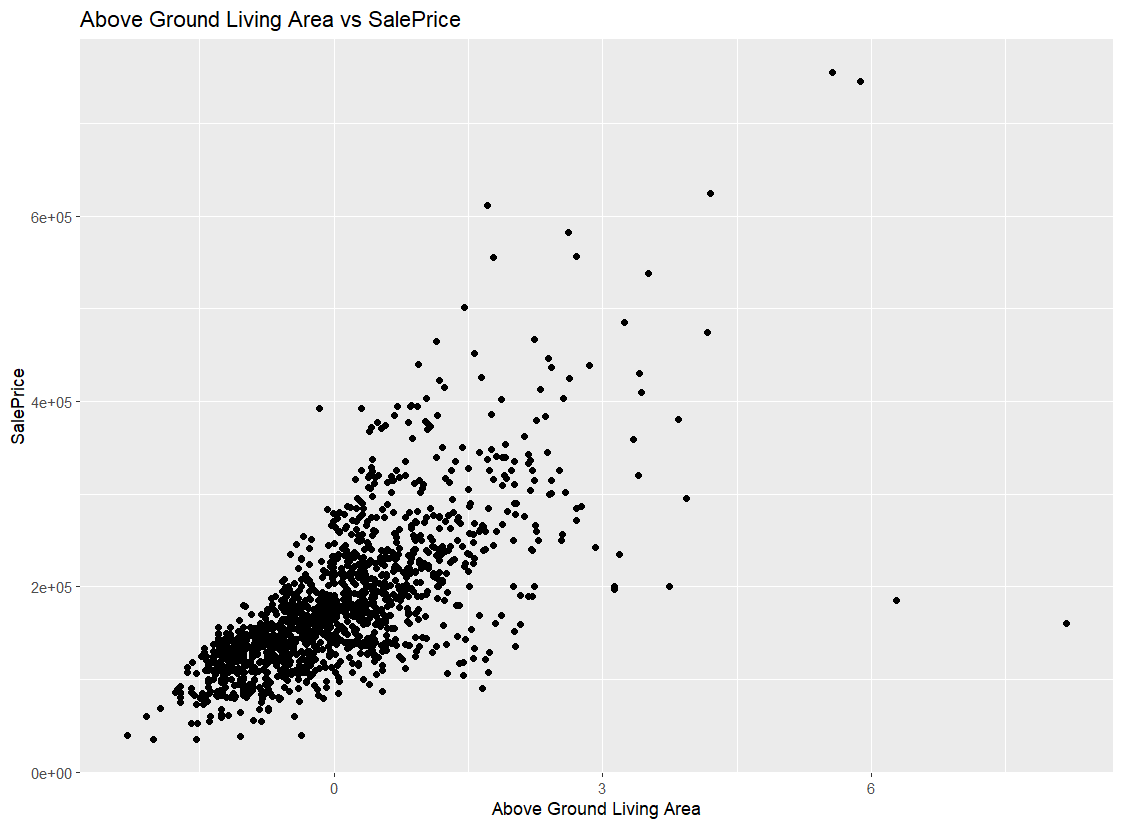
**Predicting Housing Price**

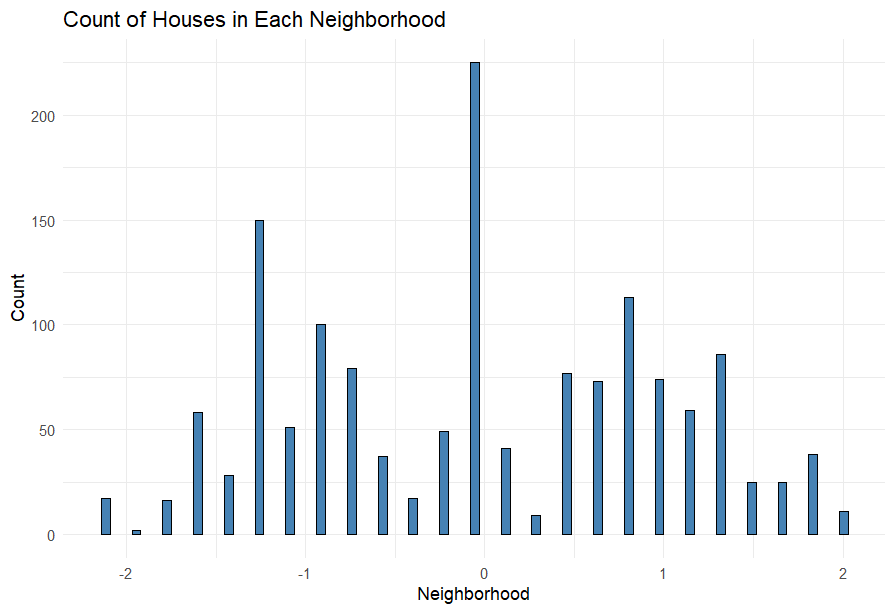
**Data Visualization**

****

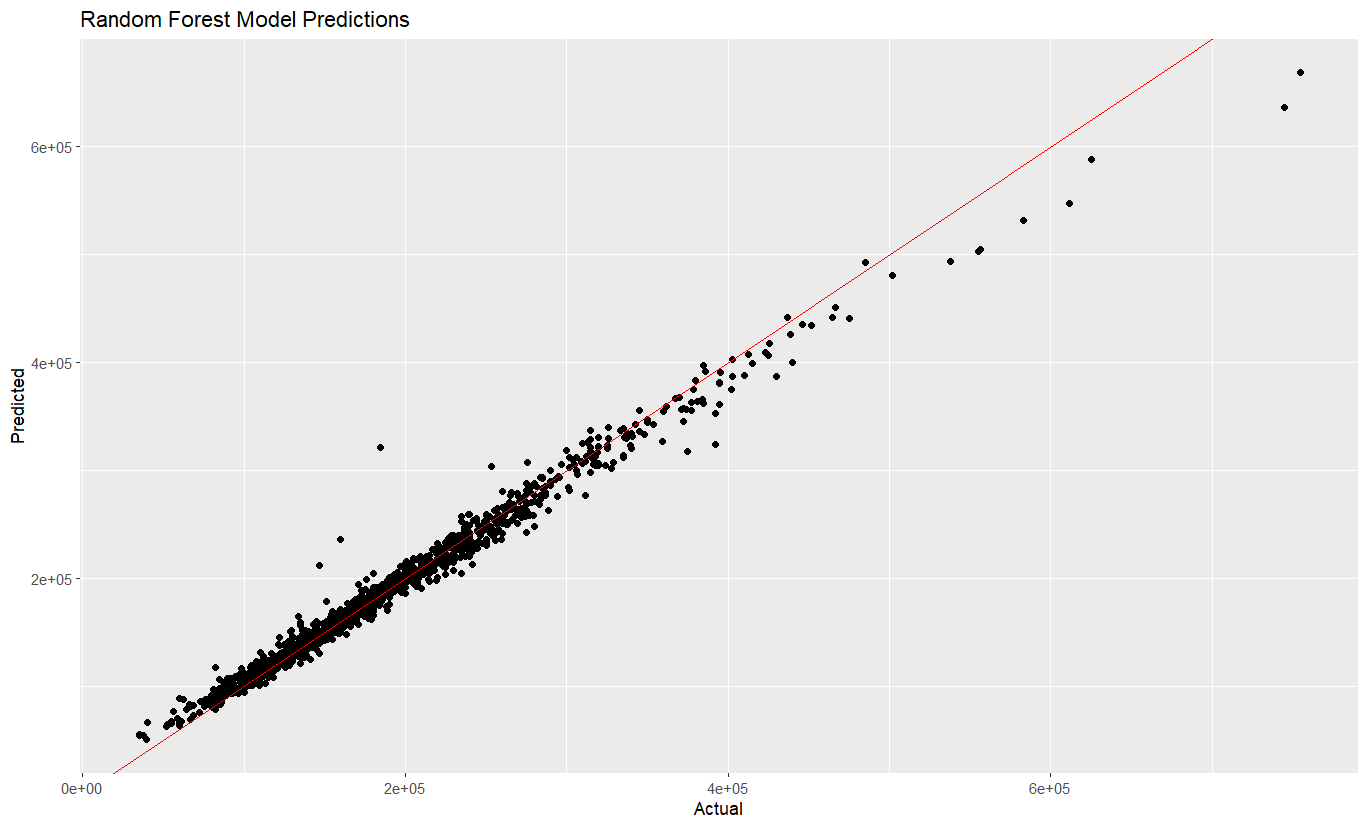
****

****

****

****

**Model Predictions:**

****

**Preprocessing:**

**Remove Null Values**

* The training and testing data have 13965 **null** values in 34 columns, which we replaced with the **median** of each column

**Encoding**

* Convert categorical columns to numeric by Convert it to factors and then to numeric

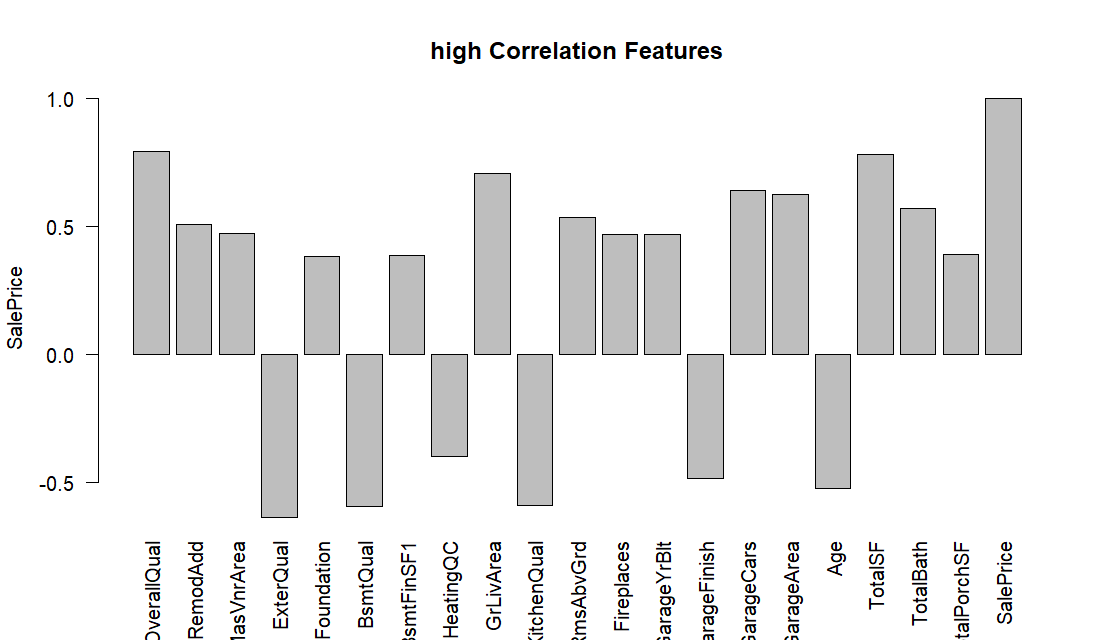
**Creating new Features**

1. We create a new feature called 'Age' by subtracting the 'YearBuilt' from the 'YrSold'.
2. We create a new feature called 'TotalSF'(Total square footage) by adding the '1stFlrSF'(First Floor square feet), '2ndFlrSF'(First Floor square feet), and 'TotalBsmtSF' (Total square feet of basement area) together.
3. We create a new feature called 'TotalBath' by adding the 'FullBath' and 'HalfBath' together.
4. We create a new feature called 'TotalPorchSF' (Total porch area) by adding the 'WoodDeckSF' , 'OpenPorchSF' , 'EnclosedPorch' , '3SsnPorch'(Three season porch area in square feet), and 'ScreenPorch' together

**Data Normalization:** Using **Scale** Built in Function

**Feature Selection**

* Extract the features that have high correlation with the Sale Price column which are 20 Classes



* Apply Different Models and select the best accuracy which is **Random Forest Algorithm**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Testing error | RMSE  Without log | RMSE  With log |
| Logistic Regression | 0.17 | 34112.006 | 0.157 |
| Decision tree | 2.15 | 36381.967 | 0.198 |
| SVM | 1.66 | 34956.789 | 0.164 |
| Random Forest | 0.15 | 10916.417 | 0.055 |

**The impact of different features on the detect news data**

|  |  |  |
| --- | --- | --- |
| Abs Correlation | #of Features | Testing error |
| >=0.35 | 20 | 0.1507 |
| >=0.3 | 22 | 0.1522 |
| >=0.1 | 43 | 0.1535 |
| >=0.5 | 12 | 0.1555 |
| >=0.7 | 3 | 0.1855 |