# Kent State Ambassador Crawford College of Business and Entrepreneurship

# Fall 2023
# BUSINESS ANALYTICS
# (BA-64036-002)

# Project Title: Zillow Home Price Prediction

## Group Participants:

| Name | Contribution Summary |
|---|---|
| Harini Padmaja Solleti (811288992) | Model Building, Model Performance, Report |
| Jaswanth Budigi (811304063) | Predictions and Results, Report |
| Likitha Sree Yarabarla(811289017) | Data Cleaning, Data Exploration,Report |
| Nanaji Chalamalashetty (811295529) | Documentation and presentation |

## PROJECT GOAL:

The project aims to develop predictive models for estimating house prices using Zillow's dataset, using regression techniques and decision tree models.
The focus is on identifying key features for accurate prediction. The project also focuses on OverallQual classification, using machine learning algorithms to categorize properties based on features.
The goal is to improve understanding of factors influencing house prices and create robust predictive models. he project aims to achieve accurate and interpretable results, contributing to a better understanding of the quality assessment process.

## OVERVIEW DATA

Information about various houses like,
These data seem to contain information related to various characteristics concerning houses, like area, features, and selling price. In the real estate industry, the numerical and category data relating to residential properties are included on these columns for analytical and modelling purposes.

| | |
|---|---|
| LotArea: | The area of the lot in square feet. |
| OverallQual: | Overall material and finish quality of the house. |
| YearBuilt: | Year when the house was built. |
| YearRemodAdd: | Year when the house was last remodeled or underwent renovation. |
| BsmtFinSF1: | Finished square feet of the basement area. |
| FullBath: | Number of full bathrooms. |
| HalfBath: | Number of half bathrooms. |
| BedroomAbvGr: | Number of bedrooms above the basement level. |
| TotRmsAbvGrd: | Total rooms above ground (excluding bathrooms). |
| Fireplaces: | Number of fireplaces. |
| GarageArea: | Size of the garage in square feet. |
| YrSold: | Year when the property was sold. |
| SalePrice: | Sale price of the property. |

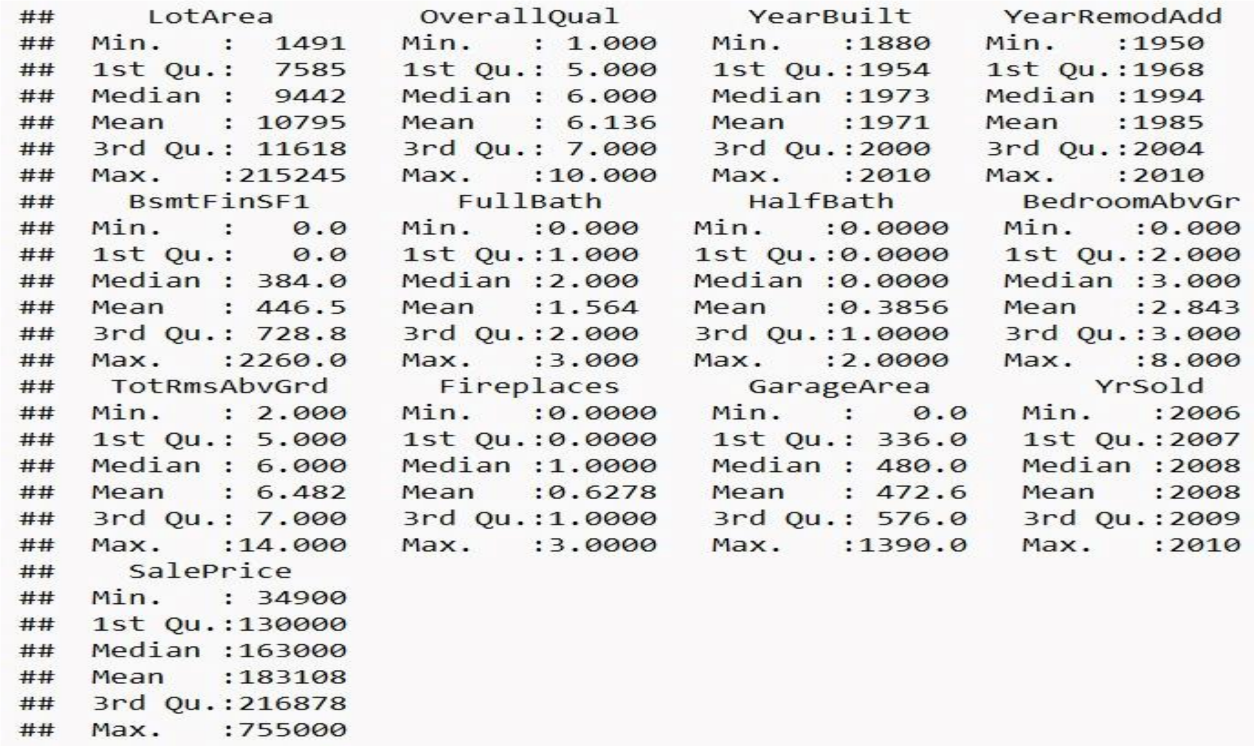
These data seem to contain information related to various characteristics concerning houses, like area, features, and selling price. In the real estate industry, the numerical and category data relating to residential properties are included on these columns for analytical and modelling purposes.

## DESCRIPTIVE STATISTICS:

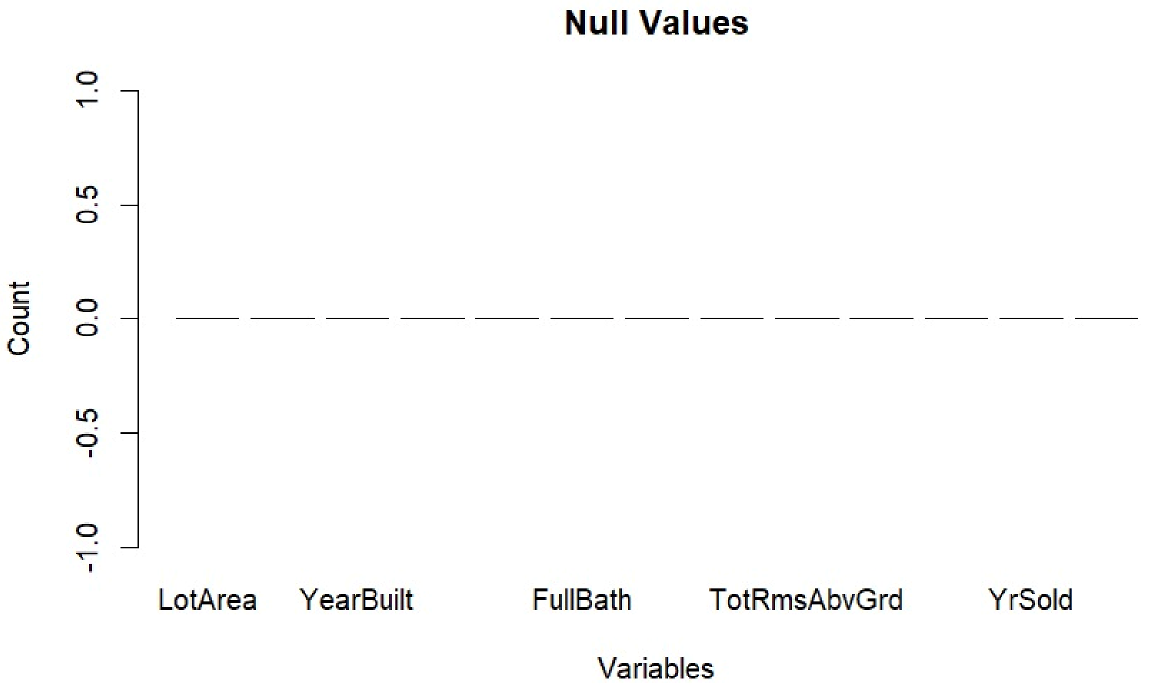
Continous columns has All 13 continous variables.
There are no missing values in the dataset.

Each of the 900 rows contains entire data.
The dataset has 11,700 observations.

**DATASET FOR QUARTILE DISTRIBUTION:**



```
##     LotArea        OverallQual       YearBuilt     YearRemodAdd
##  Min.   :  1491   Min.   : 1.000   Min.   :1880   Min.   :1950
##  1st Qu.:  7585   1st Qu.: 5.000   1st Qu.:1954   1st Qu.:1968
##  Median :  9442   Median : 6.000   Median :1973   Median :1994
##  Mean   : 10795   Mean   : 6.136   Mean   :1971   Mean   :1985
##  3rd Qu.: 11618   3rd Qu.: 7.000   3rd Qu.:2000   3rd Qu.:2004
##  Max.   :215245   Max.   :10.000   Max.   :2010   Max.   :2010
##    BsmtFinSF1        FullBath        HalfBath       BedroomAbvGr
##  Min.   :   0.0   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:   0.0   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
##  Median : 384.0   Median :2.000   Median :0.0000   Median :3.000
##  Mean   : 446.5   Mean   :1.564   Mean   :0.3856   Mean   :2.843
##  3rd Qu.: 728.8   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2260.0   Max.   :3.000   Max.   :2.0000   Max.   :8.000
##   TotRmsAbvGrd      Fireplaces       GarageArea         YrSold
##  Min.   : 2.000   Min.   :0.0000   Min.   :   0.0   Min.   :2006
##  1st Qu.: 5.000   1st Qu.:0.0000   1st Qu.: 336.0   1st Qu.:2007
##  Median : 6.000   Median :1.0000   Median : 480.0   Median :2008
##  Mean   : 6.482   Mean   :0.6278   Mean   : 472.6   Mean   :2008
##  3rd Qu.: 7.000   3rd Qu.:1.0000   3rd Qu.: 576.0   3rd Qu.:2009
##  Max.   :14.000   Max.   :3.0000   Max.   :1390.0   Max.   :2010
##    SalePrice
##  Min.   : 34900
##  1st Qu.:130000
##  Median :163000
##  Mean   :183108
##  3rd Qu.:216878
##  Max.   :755000
```



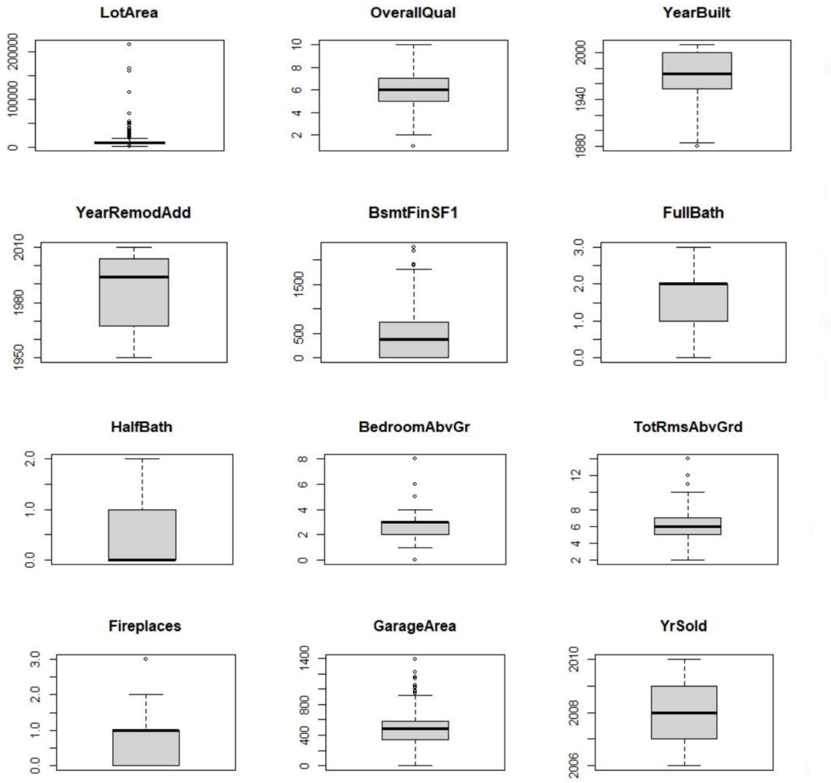
## OUTLIER DETECTION:

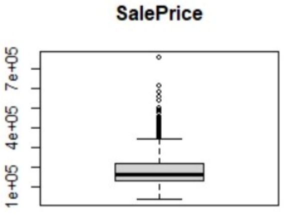
Detecting and investigating outliers that may bias analytical or predictive models. The winsorize method detects and removes outliers. Following are the results after deleting the outliers.

## VISUAL REPRESENTATION:

### Box plot

The code below generates boxplot for several numerical variables within the dataset HP_train. Once setting up an array of numbers under various column titles, its configuration of a 2-way 3-cell is for plotting the layout. A second for loop traverses through all the variables contained in the list and tries plotting a boxplot on each of them onto the appropriate grid. If the variable exists in {HP_train}, then the code produces a box plot as the main title; otherwise, it shows a message stating that the variable is absent and hence cannot be used in the dataset. This coding seeks to demonstrate their distinctive distribution values, such as any probable outlier, as well as deeper research and understanding of the data's properties.

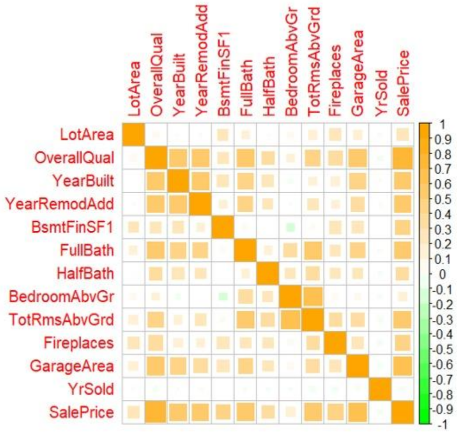


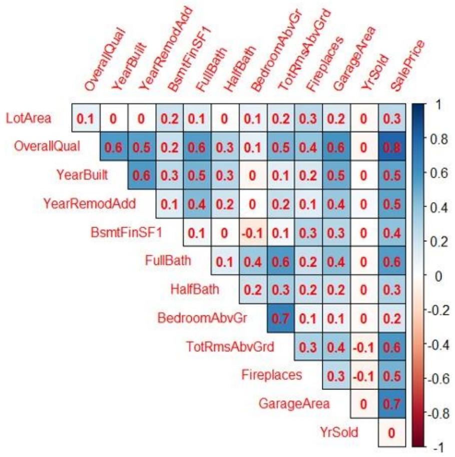


## DATA ANALYSIS:

**Correlation:**

A correlation matrix depicts the relationships between the variables in your data collection.The integers in this matrix range from -1 to +1, with 0 representing no link, -1 representing a perfect negative relationship, and the other representing a perfect positive association.



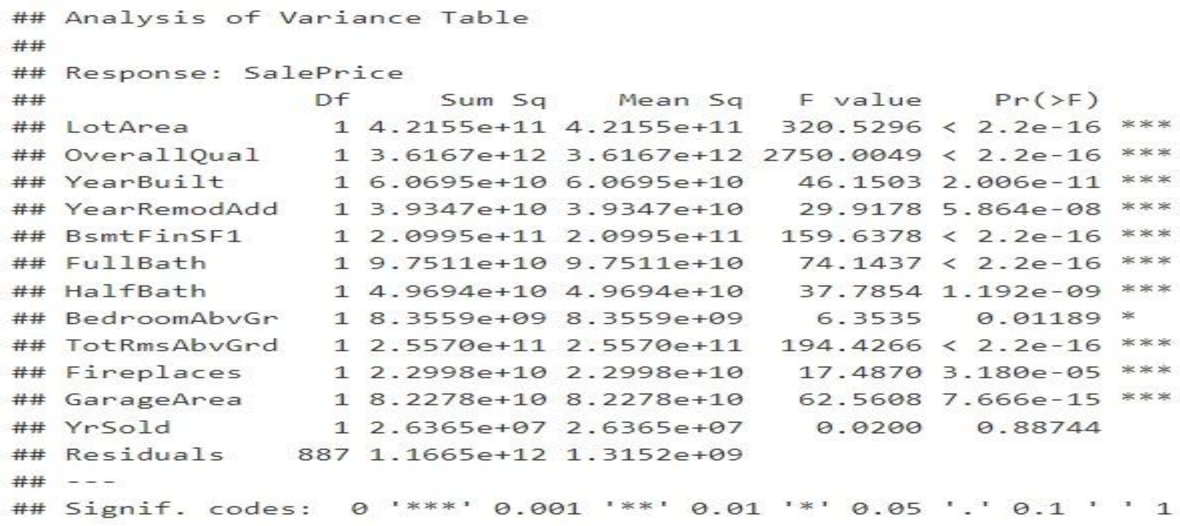


| | OverallQual | YearBuilt | YearRemodAdd | BsmtFinSF1 | FullBath | HalfBath | BedroomAbvGr | TotRmsAbvGrd | Fireplaces | GarageArea | YrSold | SalePrice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LotArea | 0.1 | 0 | 0 | 0.2 | 0.1 | 0 | 0.1 | 0.2 | 0.3 | 0.2 | 0 | 0.3 |
| OverallQual | | 0.6 | 0.5 | 0.2 | 0.6 | 0.3 | 0.1 | 0.5 | 0.4 | 0.6 | 0 | 0.8 |
| YearBuilt | | | 0.6 | 0.3 | 0.5 | 0.3 | 0 | 0.1 | 0.2 | 0.5 | 0 | 0.5 |
| YearRemodAdd | | | | 0.1 | 0.4 | 0.2 | 0 | 0.2 | 0.1 | 0.4 | 0 | 0.5 |
| BsmtFinSF1 | | | | | 0.1 | 0 | -0.1 | 0.1 | 0.3 | 0.3 | 0 | 0.4 |
| FullBath | | | | | | 0.1 | 0.4 | 0.6 | 0.2 | 0.4 | 0 | 0.6 |
| HalfBath | | | | | | | 0.2 | 0.3 | 0.2 | 0.2 | 0 | 0.3 |
| BedroomAbvGr | | | | | | | | 0.7 | 0.1 | 0.1 | 0 | 0.2 |
| TotRmsAbvGrd | | | | | | | | | 0.3 | 0.4 | -0.1 | 0.6 |
| Fireplaces | | | | | | | | | | 0.3 | -0.1 | 0.5 |
| GarageArea | | | | | | | | | | | 0 | 0.7 |
| YrSold | | | | | | | | | | | | 0 |

```
anova_model<- aov(SalePrice~.,data = HP_train)
anova_result<- anova(anova_model)
print(anova_result)
```

**ANOVA MODEL:**

ANOVA table displaying statistical significance for different predictors relative to the SalePrice response variable. Degrees of freedom, sum of squares, mean square, F -value and P -value are used to test for each predictor’s significance. The relationships between predictors such as OverallQual, LotArea, BsmtFinSF1, TotRmsAbvGrd etc., are said to have significant effects on SalePrice as they yield lower p-values or higher F-values that indicate strong relationships with SalePrice. Therefore, a variable like YrSold does not prove meaningful to predict SalePrice on the contrary. The following table explains how each independent variable influences SalePrice allowing to understand its relevance for this dataset set.



```
## Analysis of Variance Table
##
## Response: SalePrice
##               Df     Sum Sq    Mean Sq   F value    Pr(>F)    
## LotArea        1 4.2155e+11 4.2155e+11  320.5296 < 2.2e-16 ***
## OverallQual    1 3.6167e+12 3.6167e+12 2750.0049 < 2.2e-16 ***
## YearBuilt      1 6.0695e+10 6.0695e+10   46.1503 2.006e-11 ***
## YearRemodAdd   1 3.9347e+10 3.9347e+10   29.9178 5.864e-08 ***
## BsmtFinSF1     1 2.0995e+11 2.0995e+11  159.6378 < 2.2e-16 ***
## FullBath       1 9.7511e+10 9.7511e+10   74.1437 < 2.2e-16 ***
## HalfBath       1 4.9694e+10 4.9694e+10   37.7854 1.192e-09 ***
## BedroomAbvGr   1 8.3559e+09 8.3559e+09    6.3535   0.01189 *  
## TotRmsAbvGrd   1 2.5570e+11 2.5570e+11  194.4266 < 2.2e-16 ***
## Fireplaces     1 2.2998e+10 2.2998e+10   17.4870 3.180e-05 ***
## GarageArea     1 8.2278e+10 8.2278e+10   62.5608 7.666e-15 ***
## YrSold         1 2.6365e+07 2.6365e+07    0.0200   0.88744    
## Residuals    887 1.1665e+12 1.3152e+09                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## DETAILS OF YOUR MODELING STRATEGY

Based on the p-values and correlation plots, BedroomAbvGr and YrSold have no meaningful link with the target variable, SalePrice. Their impact on SalePrice appears to be minor or non-existent. As a result, these factors are removed from further research or modelling due to their insignificant impact on the target.Therefore the selected variablesfor the analysis are

1.LotArea 2.OverallQual 3.YearBuilt 4.YearRemodAdd 5.BsmtFinSF1 6.FullBath 7.HalfBath 8.TotRmsAbvGrd 9.Fireplaces10 GarageArea

## Estimation of the model's performance

The model perfome both HP_train and BA_pred_test capabilities.

### 1. REGRESSION MODEL

Intense investigation of the regression model was undertaken, revealing crucial information regarding correlations amongst predictor variables and SalePrice. After closely investigating certain things, some primary discoveries were made. The sale price had a positive relationship with most of the variables such as OverallQual and for every unit rise in these variables, there is an indication of significant rise in sale price. Nevertheless, for example, BedroomAbvGr and YrSold resulted in p-values higher than 5 percent, implying insignificance on SalesPrice. These are in agreement with correlation plots suggesting dropping these variables since they have weak or no impact on the targeted variable.

Result for this Linear Regression R-s:0.8232827
Linear Regression Rmse:28237.95

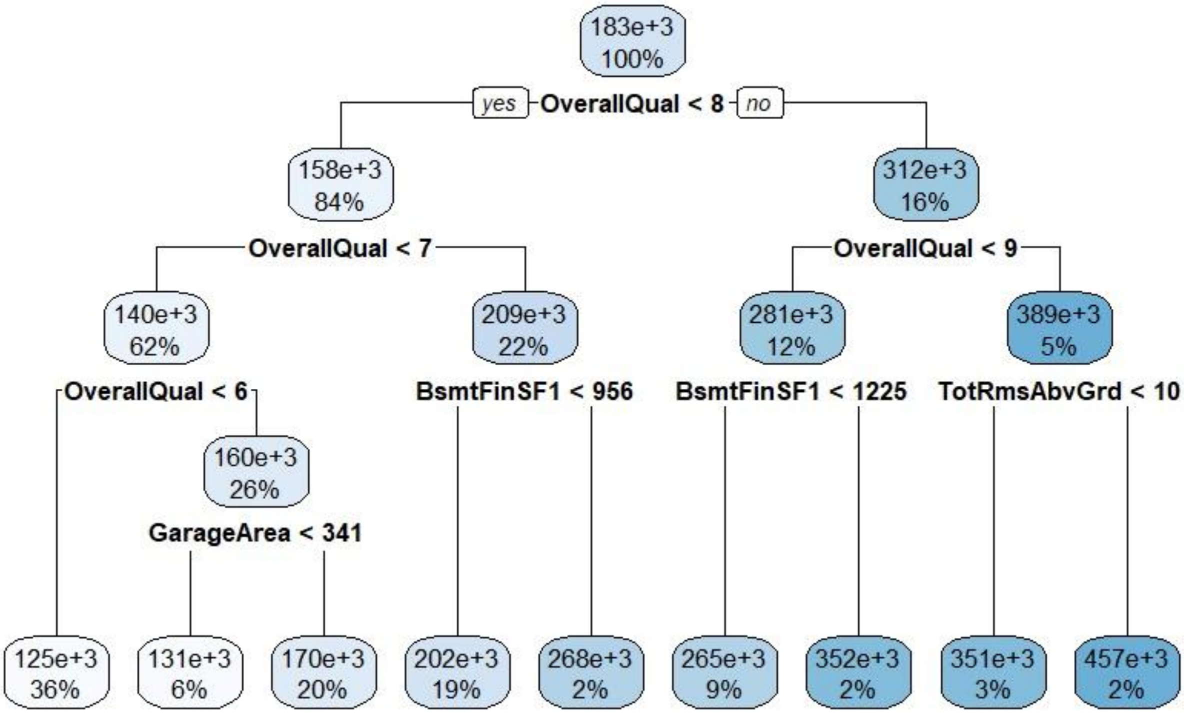
### 2. DECISION TREE

The review of the decision tree model gave important insights concerning influence of the predictors. A tree identified some main variables like OverallQual and GarageArea as prime based on the results found in the regression model, significant predictors of SalePrice include SaleTime, Advertising, and Type. These variables were they are crucial in separating the dataset, thus revealing their significant forecasting ability. Conversely, BedroomAbvGr YrSold and YrandYrSold were found unimportant splitting variables in the tree supporting that they are.

minimal impact on SalePrice. This supported the previous inference from the analysis of the decision tree.

they carried out the regression analysis which emphasized on variables that are significant and can effectively predict SalePrice

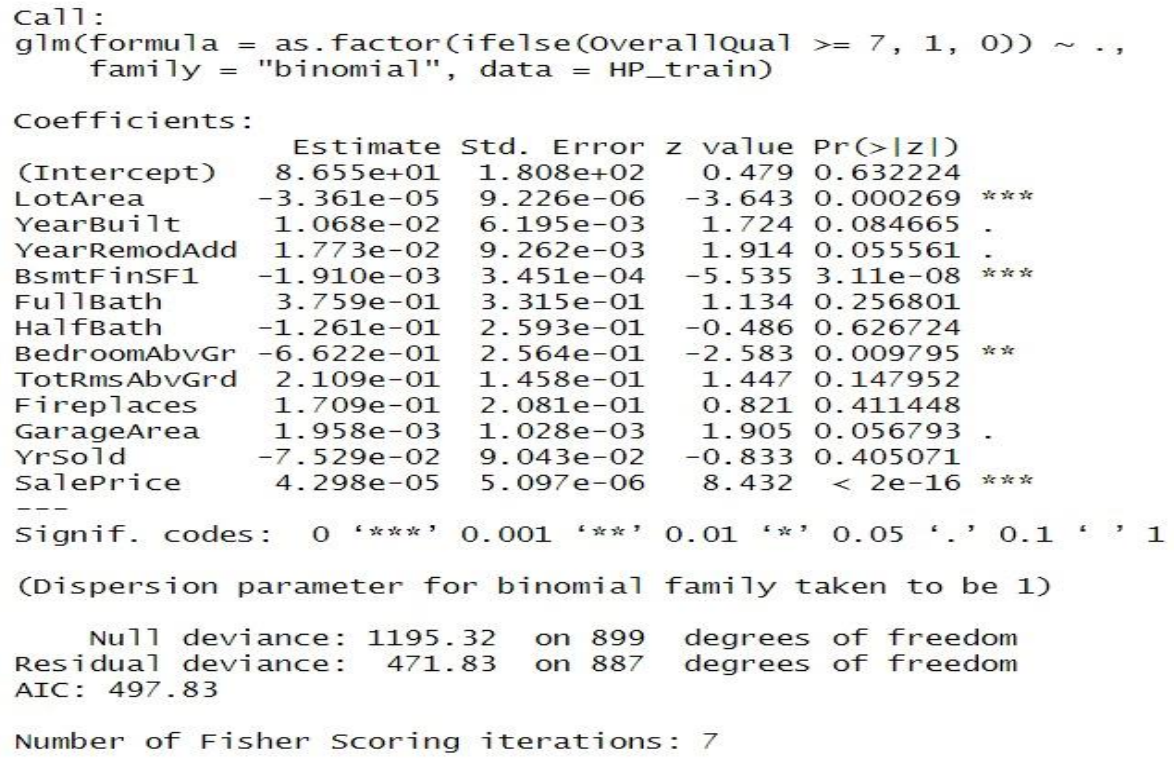
reinforcing the unimportance of BedroomAbvGr, and YrSold in this forecasting framework.



## 3.CLASSIFICATION ANLYSIS

The classification model aimed to predict OverallQual ratings, categorizing them into two classes:
a rating of 7 = 1 and other (lower) ratings = 0. After close examination, the model proved viable strong predictive capacity that can easily separate grades of better or worse quality rating
based on significant features.

To evaluate the classification model's performance, a confusion matrix was built. This matrix provides a detailed breakdown of anticipated versus actual classifications, giving a thorough assessment of the model's accuracy, precision, recall, and overall efficacy in classifying the target variable.
The following metrics are used to evaluate classification model performance.



```
Call:
glm(formula = as.factor(ifelse(OverallQual >= 7, 1, 0)) ~ .,
    family = "binomial", data = HP_train)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   8.655e+01  1.808e+02   0.479 0.632224
LotArea      -3.361e-05  9.226e-06  -3.643 0.000269 ***
YearBuilt     1.068e-02  6.195e-03   1.724 0.084665 .
YearRemodAdd  1.773e-02  9.262e-03   1.914 0.055561 .
BsmtFinSF1   -1.910e-03  3.451e-04  -5.535 3.11e-08 ***
FullBath      3.759e-01  3.315e-01   1.134 0.256801
HalfBath     -1.261e-01  2.593e-01  -0.486 0.626724
BedroomAbvGr -6.622e-01  2.564e-01  -2.583 0.009795 **
TotRmsAbvGrd  2.109e-01  1.458e-01   1.447 0.147952
Fireplaces    1.709e-01  2.081e-01   0.821 0.411448
GarageArea    1.958e-03  1.028e-03   1.905 0.056793 .
YrSold       -7.529e-02  9.043e-02  -0.833 0.405071
SalePrice     4.298e-05  5.097e-06   8.432  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1195.32  on 899  degrees of freedom
Residual deviance:  471.83  on 887  degrees of freedom
AIC: 497.83

Number of Fisher Scoring iterations: 7
```

## CONCLUSION

Finally, utilising Zillow's dataset, the team successfully constructed predictive models for house price estimation. The models' accuracy was improved by focusing on essential traits revealed through correlation analysis and statistical significance. Variables such as OverallQual, LotArea, and GarageArea were found to be important predictors, while others such as BedroomAbvGr and YrSold had little impact. The regression model performed well, with an R-squared value of 0.823 and an RMSE of 28237.95, indicating a good fit. The decision tree model highlighted the significance of influential variables. Furthermore, the classification study effectively classified OverallQual ratings, demonstrating the models' practical utility. Overall, the study contributes to a better knowledge of the dynamics of the real estate market by providing useful insights into the factors impacting property values.