1. Introduction

The dataset house data was analysed and used to forecast house prices. The primary objective of statistical analysis is to forecast the price of a house based on the information provided. The dataset is evaluated and summarised using R programming. Overall Condition and sale price are the two main variables of focus because they are the two significant variables that can be used to predict price. It is possible to interpret the significance of each variable after careful analysis. Any interdependence between these variables is also considered. House prices can be predicted with additional training and testing of the dataset. Any additional methodology that can be used is also investigated.

Any missing data or irregularities in the data are being investigated. The houses are divided into three categories based on their condition: poor, average, and good. A regression model is going to be created that can fit and predict the house price. A group of observations is chosen after predicting house prices using two different methods and models. These observations are then chosen to estimate the fitting test errors. This is accomplished by employing resampling methods and applying them to both prediction methods. It can be used to compare the reduction of error in task performance. This analysis will aid in the generation of accurate price evaluation and prediction for real estate companies. This can save a lot of time and effort when it comes to achieving proper negotiation in the field of estate agencies. The implementation of machine learning into the housing market reduces manpower investment as well.

1. a Preliminary Analysis

The data set contains 1460 details of various houses with unique Ids. These particulars are described by 51 variables. The variables are a mix of categorical and numerical values. The variable type 'character' appears in 28 columns of categorical variables and 23 columns of numerical values. OverallCond is used to predict the overall condition of houses and classify them as Poor, Average, or Good. The prediction of house prices is made using various relevant variables. These feature selections are included in the report's appendix.

Following an examination of the dataset's status, it is clear that the variable 'Id' is the only one with uniqueness in the entire dataset. This leads to the conclusion that there is a relationship between the variables. These are dealt with separately in a subsequent process in this report.

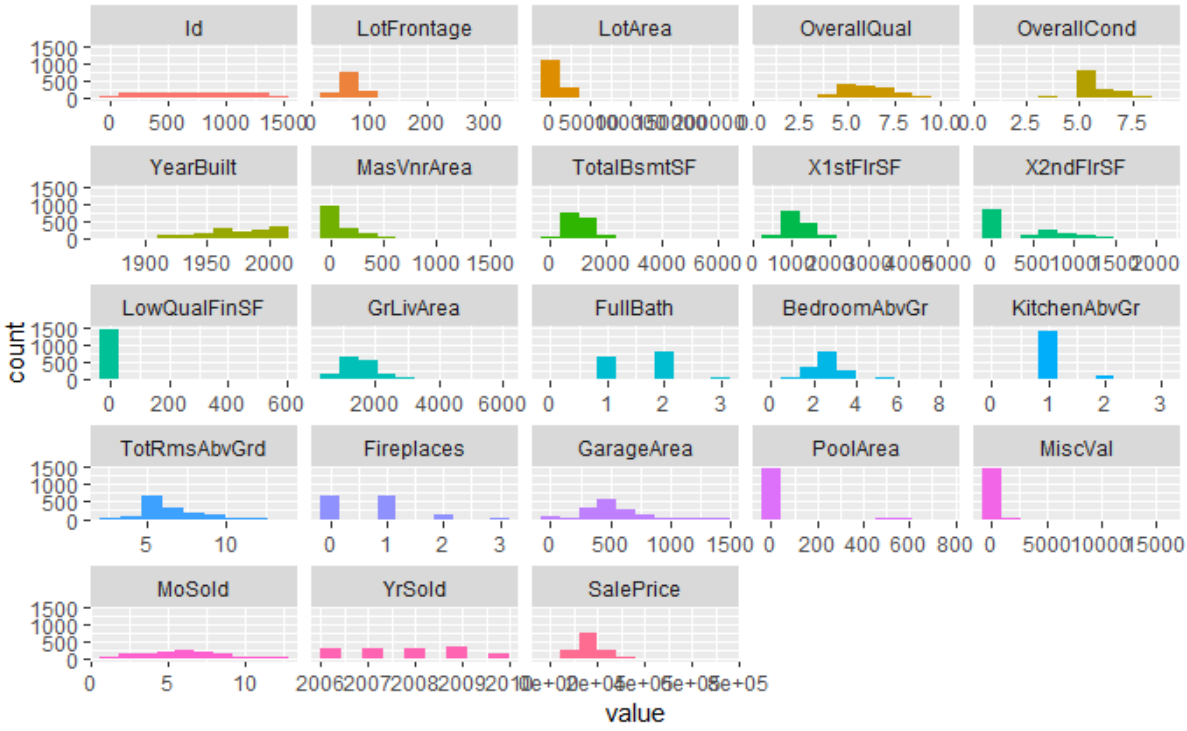

Figure 1: Numerical variable plot

1. b Missing Data Imputation and Encoding

The data set is not completely clean and must be cleaned before it can be processed. After examining the data, it is discovered that there are numerous missing values across multiple columns. Due to the nature of the data, two columns of numerical values are missing, as are eight columns of categorical values. Because these missing values do not belong in the same category, they must be handled separately. There are several approaches to dealing with numerical and categorical values. After carefully scrutinizing the numerical values in LotFrontage and MasVnrArea, it is highly recommended to use MICE imputation so that the imputed values also follow the nature of the data set as in the columns.

We begin looking at categorical values in detail after completing imputation on continuous values. It is discovered that the NA values in various columns correspond to specific categories as described in the description and are not true NA values. There are many methods for dealing with this, and even before that, those NA's should be replaced with some meaningful information so that they don't lose their true meaning of data. As a result, I converted all of those NA values into something useful. MasVnrArea and LotFrontage are two numerical data points imputed by mice. The 8 columns that are replaced and encoded with useful data are MiscFeature, Fence, PoolQC, GarageCond, GarageType, BsmtQual, BsmtCond, Alley.

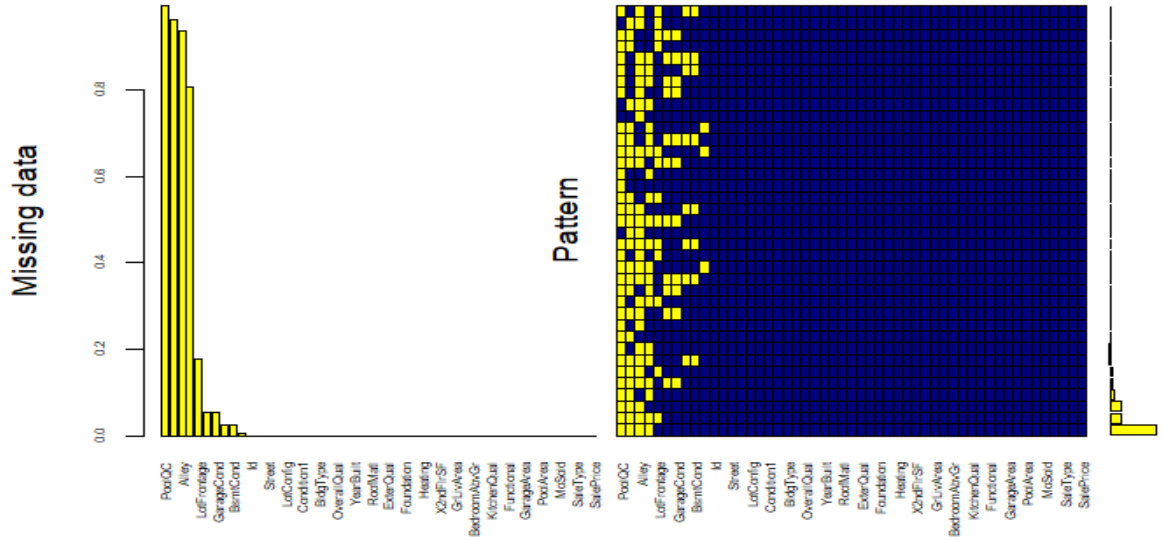

Figure 2: Missing data

For categorical values, missing values are highly identified. The method for encoding these values is included in the Appendix. The percentage of missing values for numerical variables observed is as follows. 7.9% of the data is missing, while 92.1% is present. In order to process the data correctly, these missing data must be treated specifically.

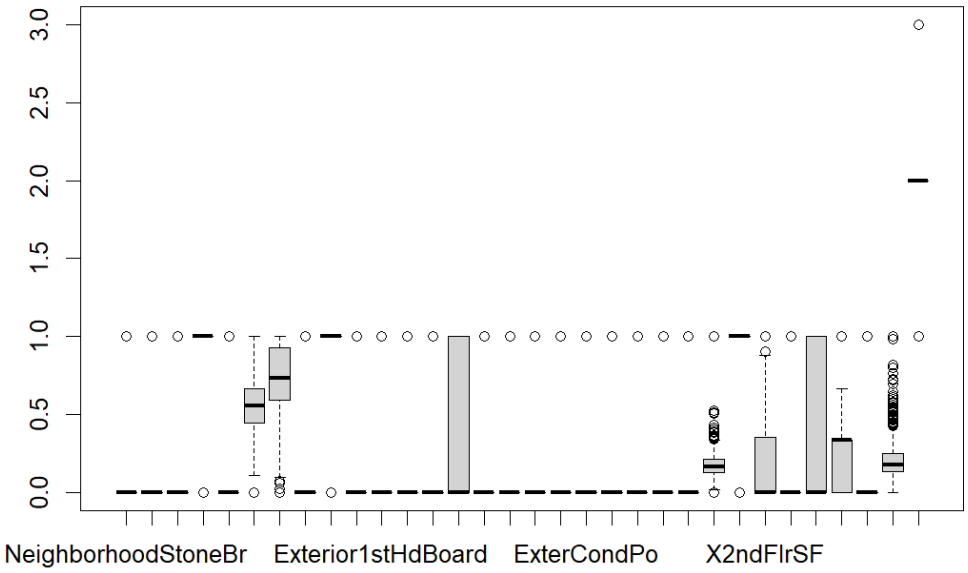

Figure 3: Visualizing Outliers

Once completed, those categorical values must be converted into numerical values in order to proceed. There are various encoding schemes currently available; however, we are proceeding with one hot encoding in which each category in each column is replaced with a new column in order to preserve the entire information. The number of columns was originally 51, but after encoding, it was increased to 207. To address the first question, which is the classification of houses based on their overall quality, the response variable that categorises house condition into poor, average, and good is restructured. Because the response variable continuous values ask for the question, we must convert them to categorical values. It is classified as poor if the value ranges from zero to three, average if the value ranges from four to six, and good if the value ranges from seven to ten.

The data cleaning is nearly complete, and the data is now ready to be used to build our model. It is discovered that the number of features in 207 is excessive, and before proceeding further, it is necessary to reduce the number of features based on their significance level and importance in predicting house quality. There are various methods for removing significant features. A few of them are developing a correlation matrix, as given in figure 6 to visually aid us in evaluating and identifying important features. Another measure would be to use feature selection, either backward or forward, to extract meaningful features.

We are continuing with a backward feature selection method, and after using it, the number of features produced has increased significantly 207, and we attempted to fit the features in a linear regression so that we could evaluate the status of the features and the values in a fitted line. After evaluating the model, it is discovered that the data sets still contain a large number of features that do not contribute significantly to predicting the response variable.

The feature set was narrowed down further with 21 features based on the significance level of various features. Finally, we will use 21 highly significant variables to predict house quality. Data are scattered in a different ways after plotting the statistics of data distribution within the data set. As a result, streamlining this data is critical, and it was decided to perform additional normalization with min Max scaling. There are various methods for standardizing data. we can perform normalization and scale for the same.

The goal of this problem is to categorize the entire data set based on the conditions given. This task necessitates the use of predictive classification analysis. There are various methods for

predicting the response variable based on a set of dependent features such as sales price, condition, quality, and so on. Neha chose decision tree-based classification and support vector machine-based classification methods from a list of classification algorithms. To compare the performance of those models, it is necessary to train and test the model before applying those methods.

We kept the ratio of 8:2 to divide the entire data set into train and test sets so that 80 percent of the dataset is classified as a training set and the remaining 20 percent is classified as test a set. Now the model will be built using the training data, and the model will be tested using the data set. Finally, the accuracy of these two methods was compared using various methods. There are visible outliers in the data, which can be eliminated with proper feature selection. It is also normalized for ease of examination.

2. Discussion of Classification Methods

Predictive classification refers to the classification of house conditions based on given features, in which prediction and classification are performed using an existing data set and trained model. Two approaches have been proposed to accomplish this.

1. Multinomial Logistic Regression
2. Support Vector Machine Regression
3. Random Forest Regression

There are numerous methods available, but we will only use those discussed during the class lectures. The multinomial regression model is a subset of logistic regression in which the response variable belongs to a different category, hence the name multivariate classification. The response variable is divided into three categories and has 21 independent features.

2. a Analysis of Classification Methods

The support vector machine model regression analysis is also one of the best-performing models, with classification performed effectively using a smooth margin. Table 1 provides an analysis of both models.

	Multinomial Logistic Regression	Support Vector Machin	Random Forest
Accuracy	0.7869	0.8041	0.8247
P Value	0.3668	0.1452	0.02628

Table 1: Analysis of Classification Methods

Based on accuracy and p-value, we can gain a better understanding of both models' performance. Random Forest outperformed both SVM and multinomial regression in terms of performance, with an accuracy of 82.47 whereas SVM and Multinomial had 80.41, and 78.69 percent respectively. Because both p values are (p>.1), it is proof that the system and assumptions are true and it works well with the given feature; thus, the null hypothesis could not be rejected, and the model works well for classification. The analysis also shows that the majority of the houses fall into the average category in terms of quality.

3. House Price Prediction & Test Error Estimation

3.a. House Price Prediction

3.a (i) Support Vector Regression Model (SVM)

Support Vector Machine with the linear kernel is the best regression method to predict the house price because there are 51 features and many training examples in the given house data set. SVM regression is based on geometric principles of data rather than its statistical principles. The SVM approach provides the best margin in the hyperplane where the data points are from different classes. It is one of the best methods to treat outliers in the data, it produces minimal error rate in the estimation. It performs well in unstructured data like text, in the house data set there are more categorical values. It reduces the chance of overfitting the data and produces higher accuracy in prediction.

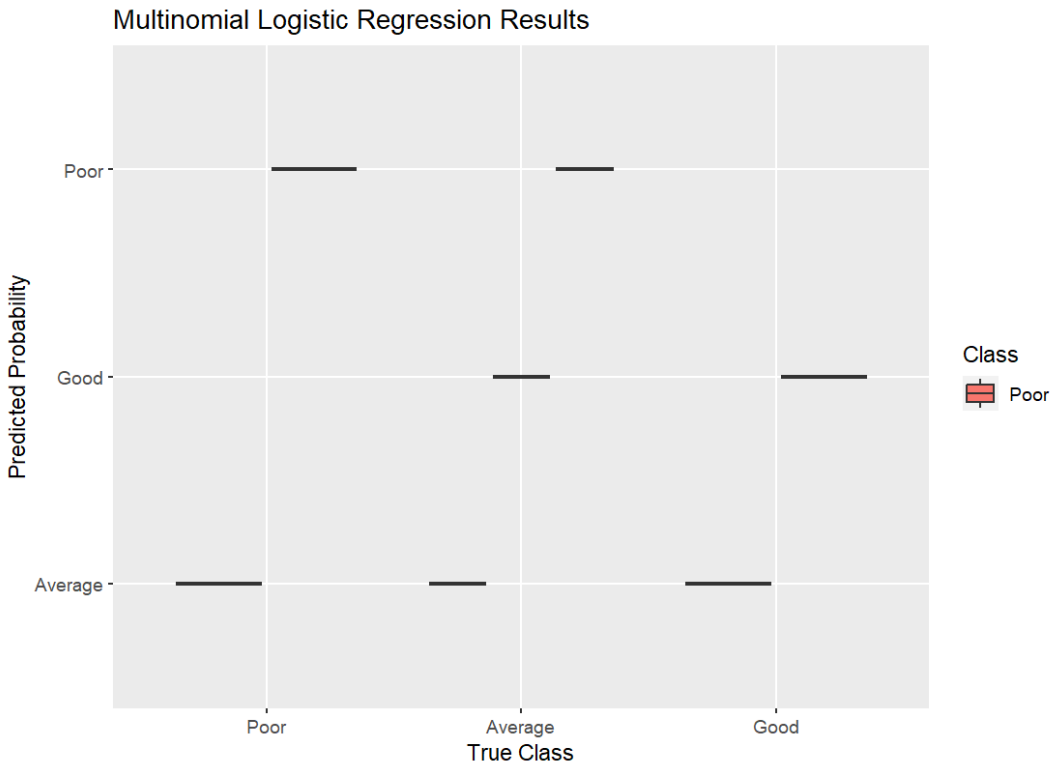

Figure 4: Fitted Multinomial Logistic Regression

SVM Prediction on Trained Data set

For SVM model implementation, after preprocessing the dataset, the house data set is divided into two sets, a trained set, and a test set. The trained model is created by applying the SVM regression approach to the trained data set. After fitting the model, the sale price of the house is predicted on the test data.

SVM Model Evaluation

The trained SVM model is evaluated using the root mean square error (RMSE) value and R^2 value. It shows how well the model predicts the data or fits the data. SVM has got an R^2 value of **0.8323** and a root mean square error of **0.039655**. The RMSE value close to zero indicates the perfect model and the R^2 value of 0.83 indicate 83% of the variation in the response variable. Moreover, the R^2 value greater than 0.7 indicates a great effect on the dependent variables. In the given graph due to the high R^2 value (0.83), most of the data points are close to the regression line. The red points indicate the predicted values fit in the model and the black points indicate the actual or observed values in the given data set.

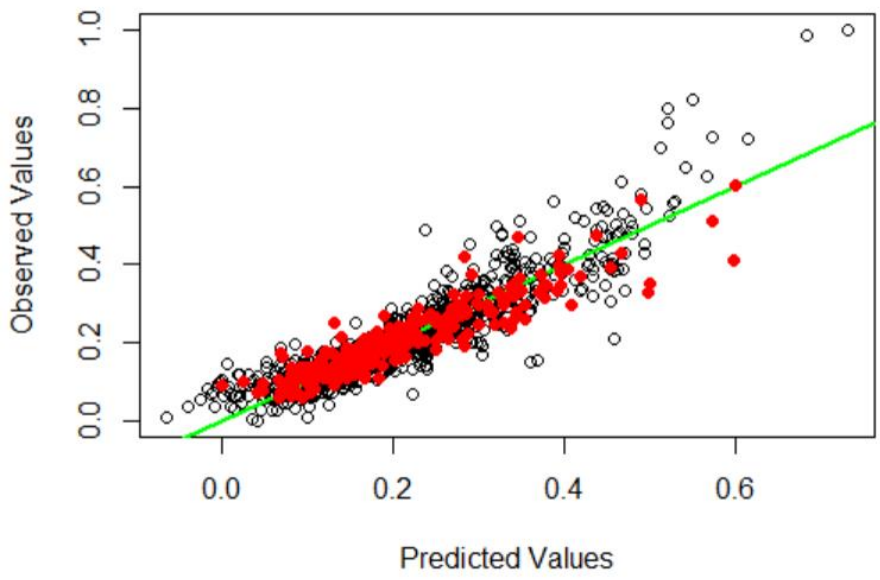

Fig 5: SVM Actual Vs Predicted Sales Price

3.a (ii) Random Forest Regression for House Price Prediction

Random Forest is a supervised learning algorithm used for both classification and prediction. It combines multiple learning algorithm outputs to predict more accurately. This is more accurate on a large data set. It performs well in data sets with both categorical and continuous values. In addition to this, it handles missing values in the data set effectively and reduces the overfitting of data.

Prediction on the Dataset using Random Forest Approach

For implementing the random forest approach the data set is divided into two sets test set and a trained set. The trained model is built on the trained set and the model is applied to the test set to predict the sales price of the house. Fig 3.2 shows the plot between actual and predicted values. The blue points show the predicted values fit near the regression line and the black points indicate the actual or observed values.

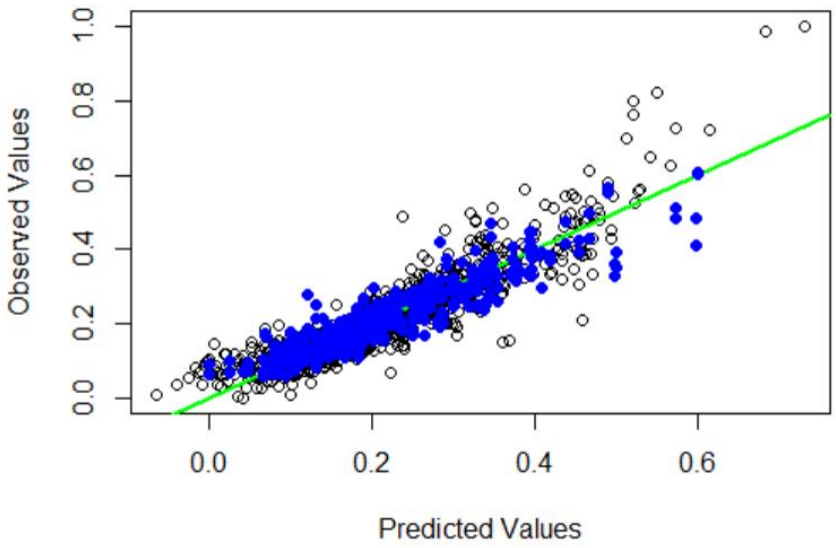

Fig.6 Random Forest Actual Vs Predicted Sales Price

Random Forest Model Evaluation

The trained model is evaluated based on the root mean square error (RMSE) value, the value obtained for the model is **0.0367,** the smaller the value higher the predicted accuracy. It has always been in the range of 0 to infinity.

3.a (iii) House Price Prediction Model Analysis

Algorithm	RMSE
Support Vector Machine	**0.039655**
Random Forest	**0.0367**

Table 2

Random Forest and SVM approaches were used to build the model for predicting the house price. While comparing fig. 3.1 and 3.2, random forest approaches best fit the model near the regression line. In addition to this, from the above-mentioned table, the root means square error value of the random forest approach is less compared to the SVM method. As a result, the random forest approach is selected as the best method to fit the model for predicting the sales price of the house.

3. b. Re-Sampling Methods for Estimating the Test Error

The resampling technique used in data analysis involves repeatedly selecting samples from a dataset to estimate the parameter population. It also tests the significance of results and helps in

the validation of the model by calculating accuracy and error. Different resampling methods include cross-validation, bootstrap, and permutation testing.

3.b (i) Bootstrap resampling

Bootstrap resampling involves drawing samples with replacements from the original dataset, and a new model is trained on each sample. This method helps to estimate the uncertainty and variability of the model's parameters.

Each resampling method has its own advantages and limitations. From the above methods, the cross-validation method has the best performance by preventing overfitting as well. In most, the predictive research cross-validation and bootstrap have the best performance and that is why it is being used in this report.

3.b (ii) Cross-Validation

Cross-Validation methods are of different types such as Leave – one – out, k-fold, and stratified cross-validation.

Leave-One-Out method uses each observation as a test sample and all others as a train. In this method, it is really expensive to deal with a very large data set in case of computation. One of the main advantages of this method is that it provides unbiased estimates for the model. The sale price is used as a test variable and all other variables are trained. These samples are then used to predict the Sale Price. It can be observed that the accuracy rate has increased from to.

Since our objective is to predict the price of the house, we use the Sale price variable. The change in accuracy is reflected here as well. The Bootstrap and Leave-One-Out cross-validation resampling methods are performed, and the prediction is made in two different prediction methods. After performing the above-mentioned resampling tasks, it can be inferred that the accuracy of the model and its reliability has improved.

	Bootstrap		**Leave One Out**	
Method	SVM	Random Forest	SVM	Random Forest
RMSE	0.0577	0.0459	0.0541	0.0412
R Squared	0.7537	0.8259	0.7604	0.8475

Table: 3 Performance Evaluation Using Different Sampling Methods

The result shows that the leave-one-out cross-validation resampling has the lowest error rate in both random forest and SVM methods. The random forest approach has an RMSE value of **0.0412** and SVM has **0.0541** in leave-one-out cross-validation. According to the test error rate, the random forest approach has better performance when compared with other methods. Each method implemented to predict the house price is deployed in resampling, these methods show a significant increase in the root means square error value on resampling compared with the value obtained in the above section.

4. Research Question

People buy property or houses based on their several requirements; the most important feature is road access to the property. This section investigates the type of road access customers are looking for when buying a house and classifies customers based on which type of road access they selected.

Decision Tree classification algorithm used to classify the road access type; people most likely to buy a house when they consider. The model is built on the trained data and predictions made on the test data. The trained model is evaluated in terms of accuracy. The model has got an accuracy of 97.88.

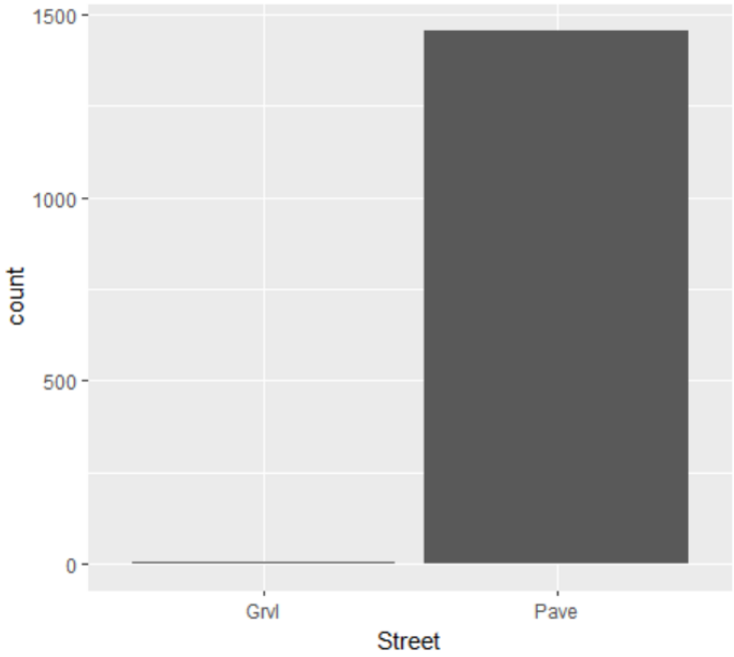

Figure 7: Prediction Based on Street