**Assignment 1**

**Introduction:**

Economy is such a muddle subject and its branches are so dense to crack. Digitalization of data is doing its job but segregating and learning from the patterns can solve a lot of difficulties. Just like housing prices. The machine learning model will predict the price of a house. The prediction of the house prices based on several factors affecting it will be considered. Here the goal is to use a generalize linear model (GLM) that can take all the aspects into consideration and give a range with high accuracy.

The objective of this assignment is as follows,

* To perform data visualization techniques to understand the insight of the data.
* This leading us to perform the machine learning model more smoothly.
* Apply various R tools to get a visual understanding of the data
* Clean it to make it ready to apply machine learning models thus predicting the housing prices

**Problem Statement:**

Consider a real estate company that has a dataset containing the prices of properties in a region. It wishes to use the data to optimise the sale prices of the properties based on important factors such as lot size, bedrooms, stories, etc.

Essentially, the company wants,

* To identify the variables affecting house prices, e.g. lot size(area), number of rooms, bathrooms, etc.
* To create a linear model that quantitatively relates house prices with variables such as the number of rooms, area, number of bathrooms, etc.
* To know the accuracy of the model, i.e. how well these variables can predict house prices.

**Data Acquisition:**

The dataset is provided in the assignment name “Housing”. This data set about the house for a particular region.

The dataset contains 546 observations

(housing samples) and 13 attributes related to housing.

**Data Dictionary:**

1. No. - Series number

2. price- The price of the house in us dollars.

3. Lotsize - Area of the house in SQFT(Square Feet)

4. bedrooms - Number of bedrooms

5. bathrms - Number of bathrooms

6. stories - Number of storage units

Categorical variables with values Yes or no

7. Driveway – Drive-way

8. Recroom – Recreational Room

9. Fullbase – Full basement

10. gashw – Gas water heater

11. airco – Air conditioner

12. garagepl – Garage place(Unique values are 0 and 1), and

13. prefarea – Area

**Data Overview:**

Required packages have been imported before loading the dataset into R.

* Head of the dataset with columns and first few rows.

**A screenshot of a cell phone

Description automatically generated**

**Exploratory Data Analysis(EDA):**

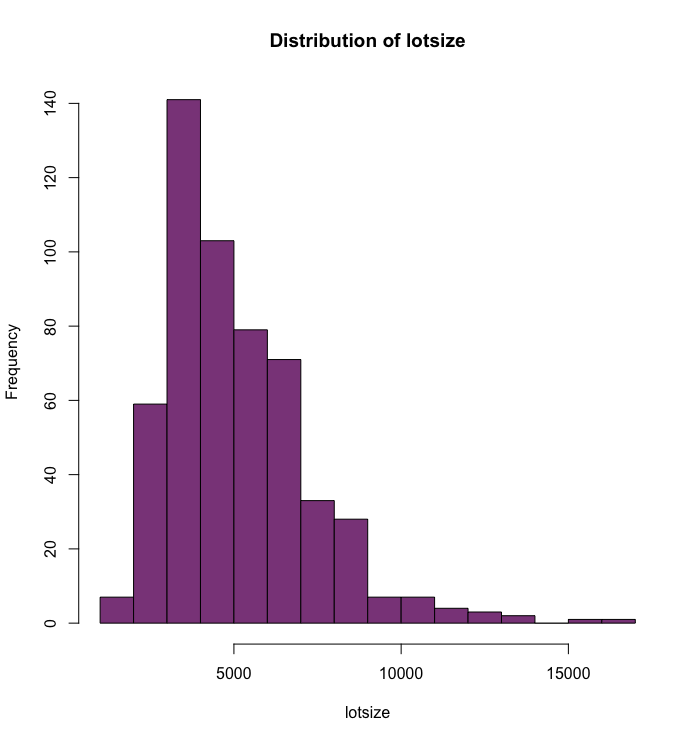
The most important step is to understand the data and identify if there is some obvious multicollinearity present. Here’s where we will also identify if predictors have a strong association with the outcome variable.

1. House price distribution using histogram



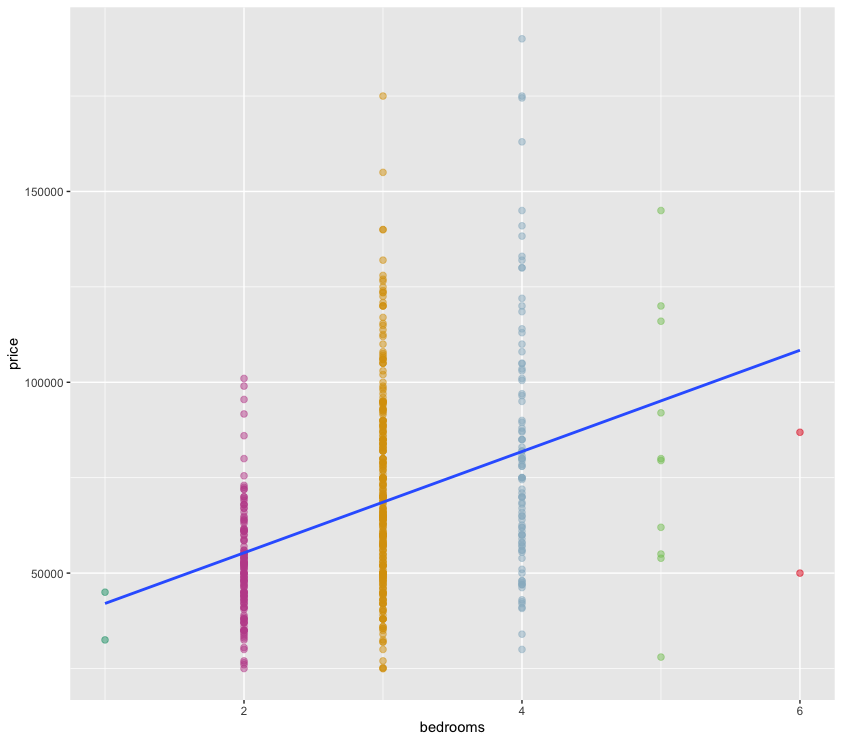
* Distribution of house prices is right-skewed. Most house prices are denser around 50k to 100k.

1. Lotsize distribution using histogram



* Most of the houses are of size 3000sqft to 7000sqft.

1. Price Vs Bedroom



* The above graph shows the number of bedroom count Vs price. Most of the houses have two to four bedrooms.
* The blue lines show a positive correlation between the number of bedrooms and the price.

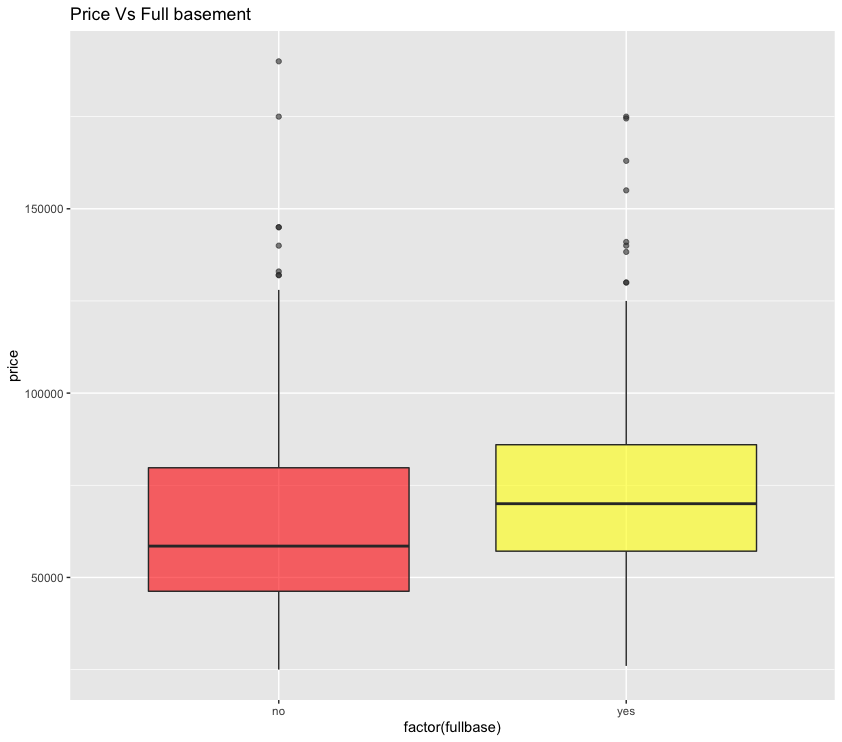
1. Bathroom count

A picture containing drawing

Description automatically generated

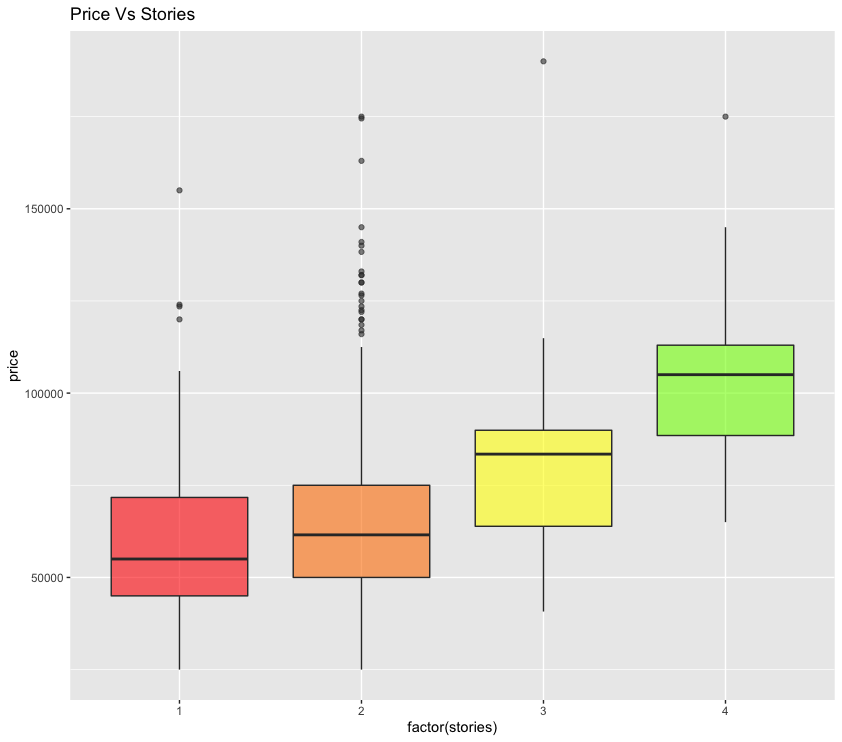
* Houses which are for sell out of those 400 of them have at most one bathroom. Very few Houses has two or more bathrooms

1. Price Vs Full basement



* The average house prices are around 70k when the full basement is available otherwise the mean price would be around 55k.
* The maximum house price is scattered abound 80k to 85k no matter basement is available or not.
* Box plot also helps to identify the outliers in the dataset.

1. Price Vs Stories



* The x-axis shows the number of stories or the floor of the house. It is apparent that as the number of stories is increasing the price is also increasing meaning that there is a positive correlation between price and stories.
* The average price for a one-story house is approximately 55k whereas four stories house has an average price above 100k which shows a huge difference in mean price.

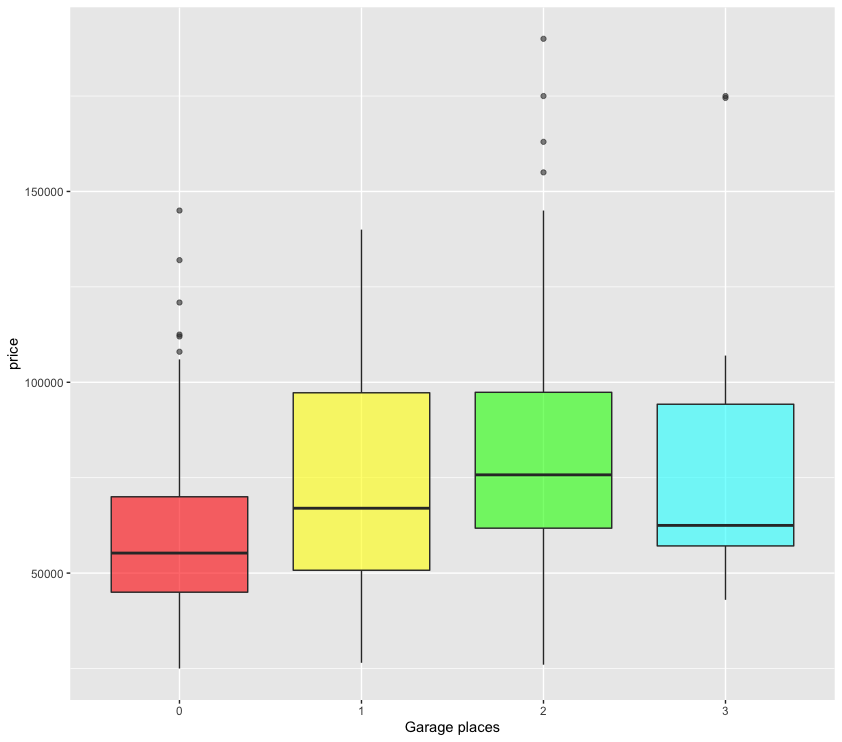
1. Number of houses by stories

A picture containing drawing

Description automatically generated

* More than 200 houses have one to two stores house whereas only 45 houses have three to four floors.

1. Price Vs Garage place



* This box plot shows that price does change according to the number of garage places. In general, houses which have one to two garage place can have average price around 60k to 70k whereas no garage space and more than two garage space has lower price values. That is approximately 60k or less.

1. Price distribution by Square feet(SQFT)

A screenshot of a cell phone

Description automatically generated

* This scatter plot shows the distribution of price per square feet.
* Price per square feet is mostly scattered around $7 to $20.

**Data Preparation:**

The first step is to check for null values in the dataset.



* As we can see that the sum of null values is zero meaning there is no null values in our dataset.

The dataset has many columns with values as 'Yes' or 'No'(categorical values). In order to fit a regression line, we would need numerical values and not string. Hence, we need to convert them to 1s and 0s, where 1 is a 'Yes' and 0 is a 'No'.

A picture containing knife

Description automatically generated

* In order to convert string to number factor() function is used. It stores the categorical values as a vector of integers in the range [1... k], (where k is the number of unique values in the nominal variable) here 0 and 1 are nominal values and an internal vector of character strings (the original values) like yes and No, mapped to these integers.

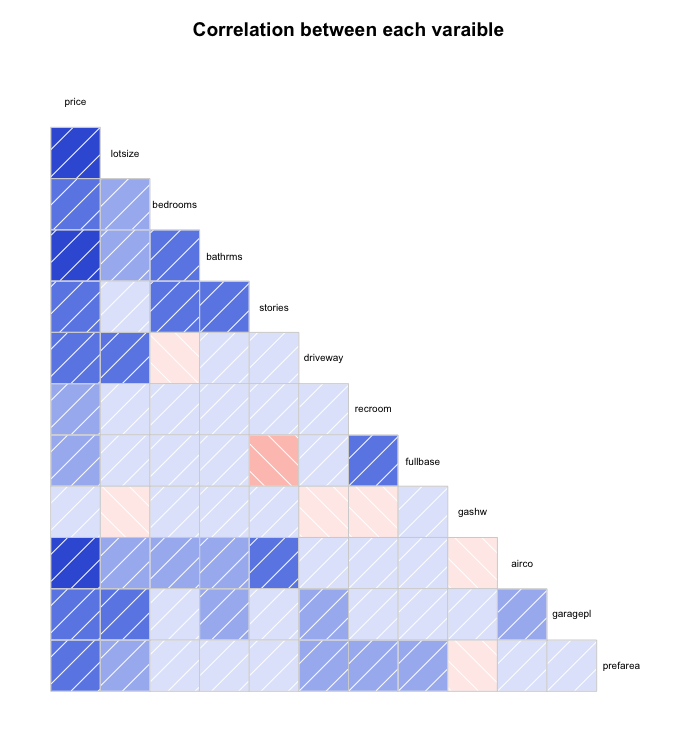
A screenshot of a social media post

Description automatically generated

* First line is used to convert the string data type to numerical after categorical string values to numerical, stored in the
* Here in the second line of code I have dropped the first column name No. which just shows the sequence of the numbers which is not important for further analysis.

Before moving further let’s consider the correlation plot which allows highlighting the variable that is most (positively or negatively) correlated.

As we are going to predict the price of the house knowing the correlation between each variable is very important. This will help the model to predict the price with higher accuracy.

****

* Darker the grid higher the correlation.
* The positive correlation has been observed between -
  + price and lotsize, bedrooms, bathrms, stories, driveway, airco, garagepl, prefarea
* The negative correlation is in a lighter shade of pink.
* Few of them are,
  + stories and fullbase,
  + lotsize, driveway, recroom and gashw.

**Models Implemented:**

In this step, we will decide which model will be appropriate for the given dataset. As discussed earlier the aim is to predict the bidding price for the house and the price is a continuous variable that has numeric values. Considering that independent variables(X) would be all the features except the price. The dependent variable(Y) will be price.

As the dependent variable is continuous we will implement ***multiple linear models*** that fall under the category of generalize linear model(GLM). Multiple linear regression is an extension of simple linear regression. It is used when we want to predict the value of the dependent variable(y) based on the two or more independent variables.

In the previous step as we have analysed the correlation between each variable, I would like to generate two models for the same dataset.

* In the first model, we will consider all the independent variables and predict the bidding price for the house.
* For the second model, we will consider the variables which have a higher correlation with the price and predict the bidding price for a given input.

Here we are going to implement two models because we want to know what is the difference between the accuracy of the model and how the features affect the model. Also wanted to predict the price of all the houses and for the single input values. A detailed explanation will be provided with each step.

**Model 1:**

The first step in the model implementation is to split the dataset to eliminate the bias to training data in machine learning algorithms.

* Split the dataset into train and test sets with ratio 70/30.

A screenshot of a cell phone

Description automatically generated

* As we can see the structure of the train data which is the same as our dataset. In order to check the number of rows in train and test split we can use nrow() which gives the number rows.
* Trian set has 381 rows and the test set has 165 it is obvious that the data is split into a 70:30 ratio.

A picture containing knife

Description automatically generated

* The basic function for fitting a linear model is *lm(formula, data)*. In the parameter where the ~ separates the response variable on the left(price) from the predictor variables on the right(independent variables), dot represents all the features are considered as independent variables except the price.
* The output of the predicted price is stored in the “result” variable. Using View() function we can view the predicted price for the dataset.



* Predicted price is in the range of 31k to 152k.

A close up of text on a white background

Description automatically generated

* This code helps to analyse each column with important statistics along with lower and upper-end points of all confidence intervals for all parameters with lower\_ci and upper\_ci.

A close up of text on a white background

Description automatically generated

* Coefficients are rounded to three decimal points so it would be easy to interpret. As we saw earlier that few variables have a higher coefficient and that impacts a lot on the price prediction.

A screenshot of text

Description automatically generated

* Here comes an important matrix- a summary of model1. These tables also has the error score which helps to identify the accuracy of the model. If the model has a lower error score then the model has higher accuracy.
* Coefficient of determination
  + R-Squared value is a statistical measurement of how close the data are the fitted regression line. It is denoted by R2(R-squared)
  + It is also known as the coefficient of determination.
  + R-square will have a value between 0 and 1.
  + A value of R2 near to 1 indicates that most of the variation of the response data is explained by the different input values, whereas a value of R2 near to 0 indicates that the variation is explained by the different input values.
  + In our case R-square is 0.65 which is acceptable.

Scatter plot to show the fitted model 1 with regression line

A close up of a map

Description automatically generated

* The scatter plot shows the model1 fitted data and it is apparent that the regression line fits almost all the data points.

**Model 2:**

For model 2, we have considered only those variable which has a higher correlation with the dependent variable. Initial steps have been the same as explained in the previous model but the variation will be detailed during implementation.

A screenshot of a social media post

Description automatically generated

* As we can see in the above snippet, we have built a model using the eight most highly correlated variables which are separated by + sign. Data is used from the train set labeled as train2.

A screenshot of text

Description automatically generated

* The summary of model-2 shows the table which has key information. As we have explained all the important matrix previously, so we will focus on the R-square value which is 0.66 relatively moderate on the scale from 0 to 1.

A graph to represent the fitted data using model-2

A close up of a map

Description automatically generated

* As we can see from the above graph the regression line fits almost all the data points and the relation between price and the predicted price with model2 shows a positive relation.

Predicting the price for given eight different features of a house

* To predict the appropriate bidding price for the house based on the features, for instance, a house which is for bidding has a lot size 5500sqft with four bedrooms and two bathrooms as well as one story, driveway, air conditional. Garage space also available with no prefarea.
* The values are passed in predict( ) method.

****

* The predicted value is $88205.19 which approximately $15 per square feet. We can trust this model as the R2 is low around 0.6679 and F-statistic is 97.81.

**Discussion:**

o What makes the problem interesting from the viewpoint of analytics?

* This particular problem is interesting because it is a hands-on case for any data analyst. We get to implement what we learn in the theory and seeking it work and helpful in the market gives satisfaction to the person working on it.

o How did the chosen technique help to illuminate, or solve the problem?

* The chose technique helps to remove the problem by giving us the best bidding price based on the different features of the house. This won’t just help the buyer but also helps the homeowners to keep their house for bidding with appropriate price and that leads to satisfaction for both the parties.

o What analysis do you think should be conducted next?

* Among the recent havoc, a structured and detailed study of COVID-19 patients trauma after being cured. On a scale of 0 to 1, how has the disease changed their physical health?. The study could help us understand their experience and trauma. A clear insight of the same would be resourceful.

**Conclusion:**

We have visualized the data and implemented two different models with almost the same accuracy which shows that even if we consider all the features during modelling only those features give a good impact on the model which has a higher correlation with the dependent variable. In order to predict the bidding price of the house we can use advanced model which may improve r score as it is low.

**Reference:**

1] <https://www.youtube.com/watch?v=CuJc1MFY23k>

2] https://www.statmethods.net/advgraphs/correlograms.html

3]<https://www.rdocumentation.org/packages/moderndive/versions/0.5.0/topics/get_regression_table>

4] Robert Kabacoff. “R in Action, Second Edition: Data analysis and graphics with R”

5]<https://online.stat.psu.edu/stat504/node/216/#:~:text=The%20term%20general%20linear%20model,(with%20fixed%20effects%20only)>

6]https://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713\_randomerror/EP713\_RandomError6.html