Report

House Price Prediction

REPORT

HOUSE PRICE PREDICTIONS

Course Name :- ADVANCED PROGRAMMING LABORATORY

Couse Code :- 3CP05

Students ID & Name :- 18CP066 - Patel Ankur N
19CP311 - Parmar Vivek M

Abstract

Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Here we aim is to predict price according to attributes like no of bedrooms, no of bathrooms, area of living room, area of living room, area of lot, area of above, area of basement, year of built, year of renovation, area, city, state.

We are also going to analyse price as per the individual attributes. And we will see how graph varies as per these attributes.

Table of Content

Chapter 1 – Introduction

The starting step in exploring any big amount of data is Exploratory Data Analysis (EDA). It starts with a basic understanding of data and the framing of which factors to focus on. This is done with a broad view of the patterns and quantitative techniques which give a basic understanding of what the dataset depicts. Our quest for clues that help in framing the future steps and the question that arise from the data set is what truly drives the Exploratory Data Analysis, as there are no standard set of rules which tell the user how to approach data. That aside EDA gives life to a set of statistical methods which help to define the purpose of the data. Almost all EDA practices are graphical in nature with a few quantitative techniques. The motive for the high dependence on graphics is that by its very nature the main role of EDA is to liberally investigate, and graphs gives the examiners unparalleled chances to do as such, tempting the information to uncover its basic mysteries, and being constantly prepared to increase some new, regularly unsuspected, understanding into the information. In blend with the characteristic example acknowledgment abilities that we as a whole have, design gives, obviously, unparalleled energy to do this.

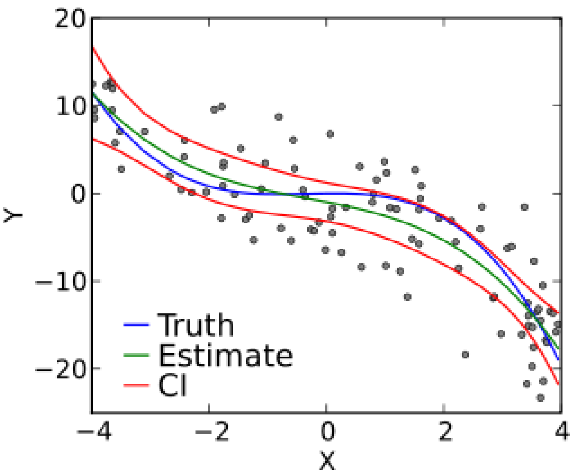
Chapter 2 – Methods and Methodologies

Linear Regression:

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called **multiple linear regression**. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Standard linear regression models with standard estimation techniques make a number of assumptions about the predictor variables, the response variables and their relationship. Numerous extensions have been developed that allow each of these assumptions to be relaxed (i.e. reduced to a weaker form), and in some cases eliminated entirely. Generally, these extensions make the estimation procedure more complex and time-consuming, and may also require more data in order to produce an equally precise model.



Example of a cubic polynomial regression, which is a type of linear regression.

The following are the major assumptions made by standard linear regression models with standard estimation techniques (e.g. ordinary least squares):

- **Weak exogeneity**. This essentially means that the predictor variables *x* can be treated as fixed values, rather than random variables. This means, for example, that the predictor variables are assumed to be error-free—that is, not contaminated with measurement errors. Although this assumption is not realistic in many settings, dropping it leads to significantly more difficult errors-in-variables models.
- **Linearity**. This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables. Note that this assumption is much less restrictive than it may at first seem. Because the predictor variables are treated as fixed values (see above), linearity is really only a restriction on the parameters. The predictor variables themselves can be arbitrarily transformed, and in fact multiple copies of the same underlying predictor variable can be added, each one transformed differently. This technique is used, for example, in polynomial regression, which uses linear regression to fit the response variable as an arbitrary polynomial function (up to a given rank) of a predictor variable. With this much flexibility, models such as polynomial regression often have "too much power", in that they tend to overfit the data. As a result, some kind of regularization must typically be used to prevent unreasonable solutions coming out of the estimation process. Common examples are ridge regression and lasso regression. Bayesian linear regression can also be used, which by its nature is more or less immune to the problem of overfitting. (In fact, ridge regression and lasso regression can both be viewed as special cases of Bayesian linear regression, with particular types of prior distributions placed on the regression coefficients.)
- **Constant variance** (a.k.a. **homoscedasticity**). This means that different values of the response variable have the same variance in their errors, regardless of the values of the predictor variables. In practice this assumption is invalid (i.e. the errors are heteroscedastic) if the response variable can vary over a wide scale. In order to check for heterogeneous error variance, or when a pattern of residuals violates model assumptions of homoscedasticity (error is equally variable around the 'best-

fitting line' for all points of x), it is prudent to look for a "fanning effect" between residual error and predicted values. This is to say there will be a systematic change in the absolute or squared residuals when plotted against the predictive variables. Errors will not be evenly distributed across the regression line. Heteroscedasticity will result in the averaging over of distinguishable variances around the points to get a single variance that is inaccurately representing all the variances of the line. In effect, residuals appear clustered and spread apart on their predicted plots for larger and smaller values for points along the linear regression line, and the mean squared error for the model will be wrong. Typically, for example, a response variable whose mean is large will have a greater variance than one whose mean is small. For example, a given person whose income is predicted to be $100,000 may easily have an actual income of $80,000 or $120,000 (a standard deviation of around $20,000), while another person with a predicted income of $10,000 is unlikely to have the same $20,000 standard deviation, which would imply their actual income would vary anywhere between -$10,000 and $30,000. (In fact, as this shows, in many cases—often the same cases where the assumption of normally distributed errors fails—the variance or standard deviation should be predicted to be proportional to the mean, rather than constant.) Simple linear regression estimation methods give less precise parameter estimates and misleading inferential quantities such as standard errors when substantial heteroscedasticity is present. However, various estimation techniques (e.g. weighted least squares and heteroscedasticity-consistent standard errors) can handle heteroscedasticity in a quite general way. Bayesian linear regression techniques can also be used when the variance is assumed to be a function of the mean. It is also possible in some cases to fix the problem by applying a transformation to the response variable (e.g. fit the logarithm of the response variable using a linear regression model, which implies that the response variable has a log-normal distribution rather than a normal distribution).

- **Independence** of errors. This assumes that the errors of the response variables are uncorrelated with each other. (Actual statistical independence is a stronger condition than mere lack of correlation and is often not needed, although it can be exploited if it is known to hold.) Some methods (e.g. generalized least squares) are capable of handling correlated errors, although they typically require significantly more data unless some sort of regularization is used to bias the model towards assuming uncorrelated errors. Bayesian linear regression is a general way of handling this issue.

- **Lack of perfect multicollinearity** in the predictors. For standard least squares estimation methods, the design matrix X must have full column rank p; otherwise, we have a condition known as perfect multicollinearity in the predictor variables. This can be triggered by having two or more perfectly correlated predictor variables (e.g. if the same predictor variable is mistakenly given twice, either without transforming one of the copies or by transforming one of the copies linearly). It can also happen if there is too little data available compared to the number of parameters to be estimated (e.g. fewer data points than regression coefficients). In the case of perfect multicollinearity, the parameter vector $\boldsymbol{\beta}$ will be non-identifiable—it has no unique solution. At most we will be able to identify some of the parameters, i.e. narrow down its value to some linear subspace of $\mathbf{R}^p$. See partial least squares regression. Methods for fitting linear models with multicollinearity have been developed; some require additional assumptions such as "effect sparsity"—that a large fraction of the effects are exactly zero. Note that the more computationally expensive iterated algorithms for parameter estimation, such as those used in generalized linear models, do not suffer from this problem.

Beyond these assumptions, several other statistical properties of the data strongly influence the performance of different estimation methods:

- The statistical relationship between the error terms and the regressors plays an important role in determining whether an estimation procedure has desirable sampling properties such as being unbiased and consistent.
- The arrangement, or probability distribution of the predictor variables $\mathbf{x}$ has a major influence on the precision of estimates of $\boldsymbol{\beta}$. Sampling and design of experiments are highly developed subfields of statistics that provide guidance for collecting data in such a way to achieve a precise estimate of $\boldsymbol{\beta}$

Chapter 3 - Implementation Results and Discussion

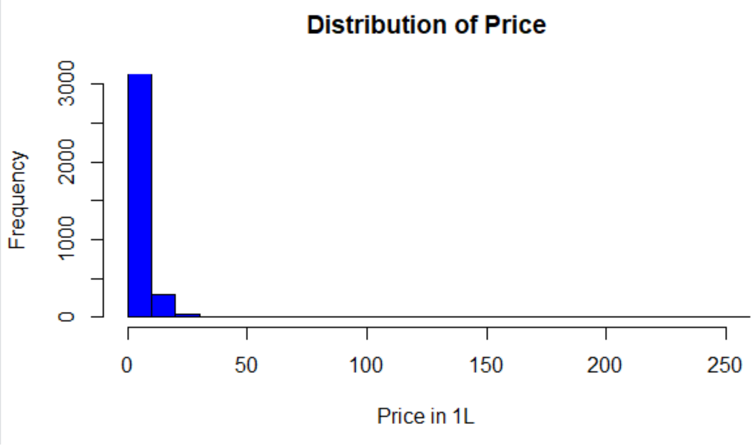
Our dataset has attributes no of bedrooms, no of bathrooms, area of living room, area of living room, area of lot, area of above, area of basement, year of built, year of renovation, area, city, state and also a price. First of all, we have trained the model with the available data set. And then after test is being conducted. In between different graphs are generated.

Result and Discussion :

Firstly we plot some attributes and discuss its behaviour,

- Distribution of Price

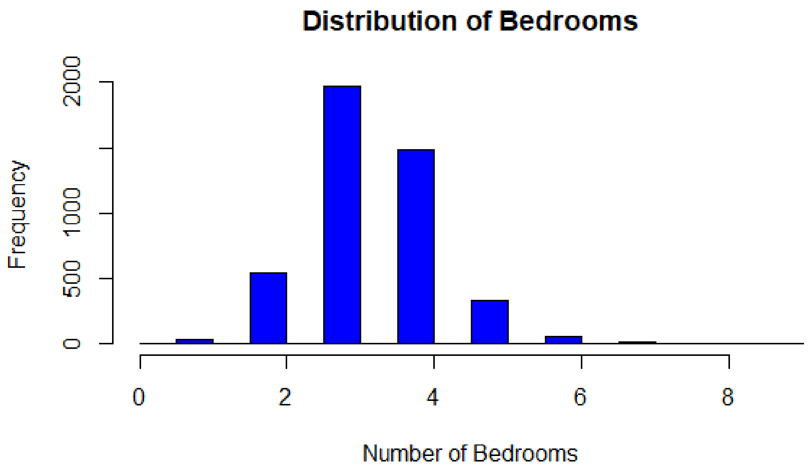
```
#price distribution
hist(priceIN1L,
    main = "Distribution of Price",
    xlab = 'Price in 1L',
    ylab = 'Frequency',
    col = 'blue',
    breaks = 20,
    xlim = c(0,250),
    ylim = c(0,3000)
)
```



Discussion :- From the above graph, we visualize that price is less than 50 lakhs

- Distribution of Bedrooms

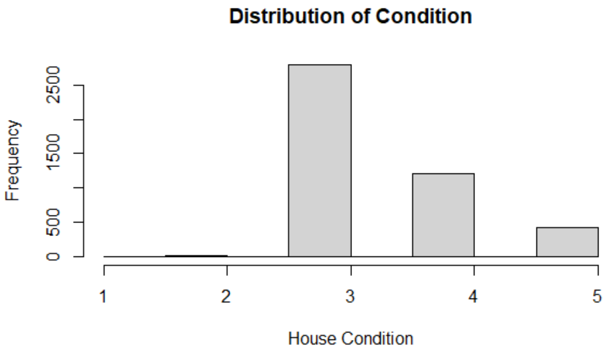
```
#bedroom distribution
hist(train$bedrooms,
    main = "Distribution of Bedrooms",
    xlab = 'Number of Bedrooms',
    ylab = 'Frequency',
    col = 'blue'
)
```



Discussion :- From the above graph, we visualize that most of the houses have 3 to 4 bedrooms

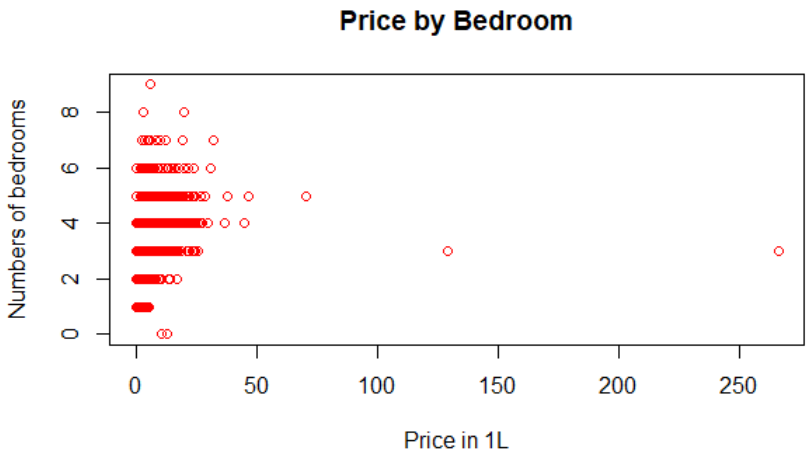
- Distributions of Conditions

```
#condition distribution

hist(train$condition,

    main = "Distribution of Condition",

    xlab = 'House Condition',

    ylab = 'Frequency',

)
```



- Price by Bedrooms

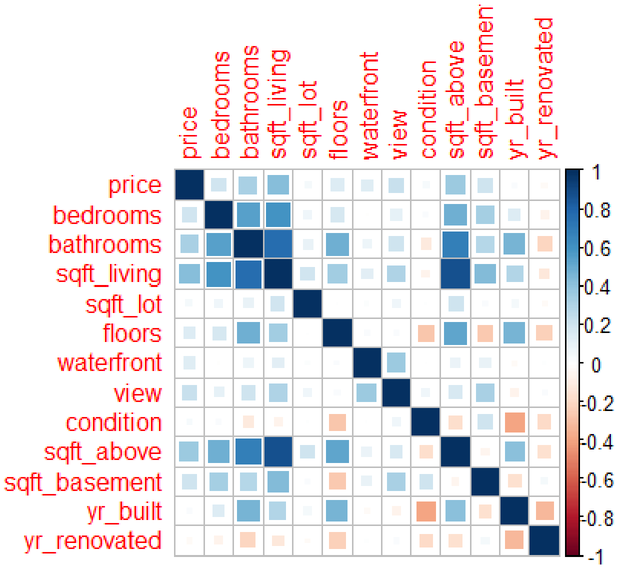
```
plot(priceIN1L,
    train$bedrooms,
    main = "Price by Bedroom",
    xlab = 'Price in 1L',
    ylab = 'Numbers of bedrooms',
    col = 'red')
```



<u>Discussion</u> :- From the above graph, we visualize that price is dependent on bedrooms

- Correlation Plot

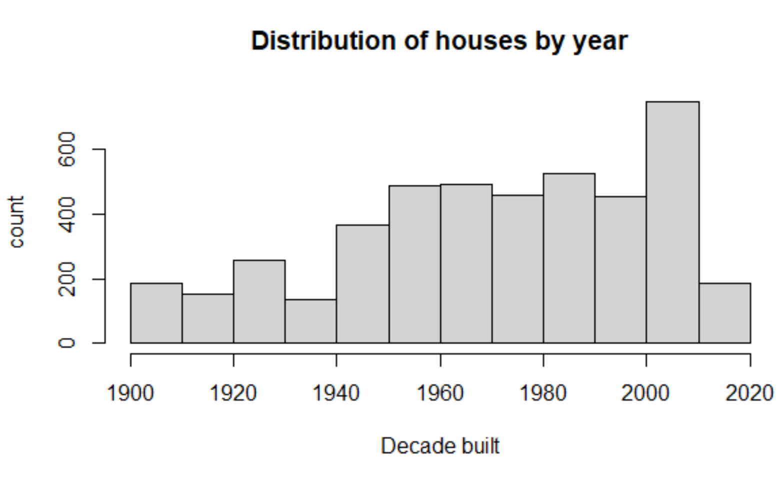
```
correlation <- cor(na.omit(trainNum[,-1]))
corrplot(correlation, method="square")
```



<u>Discussion</u> :- From the above graph, we visualize relationship between two attributes

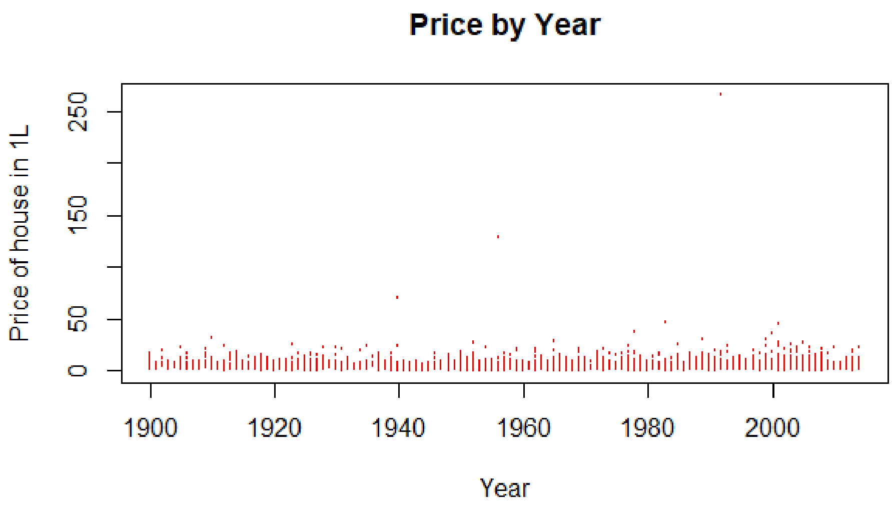
- Distribution of houses by year of build

```
#distribution by year
hist(priceByDecade$decade,
     main = 'Distribution of houses by year',
     xlab = 'Decade built',
     ylab = 'count')
```



- Price by Year

```
#Scatter plot of price by year
plot(
  priceIN1L ~ train$yr_built,
  cex = .2,
  main = 'Price by Year',
  xlab = 'Year',
  ylab = 'Price of house in 1L',
  col = 'red'
)
```



Discussion :- From the above graph, we visualize that recently build houses have more price compared to older houses.

Conclusion

Using Multiple Regression we create a model for various attributes and successfully generated predicted data in file “PredictedData.csv”. We can plot an attributes for visualization and also able to get variations on price versus different attributes.

References

https://r4ds.had.co.nz/exploratory-data-analysis.html

https://www.kaggle.com/

https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php#:~:text=Multiple%20regression%20is%20an%20extension,%2C%20target%20or%20criterion%20variable).

https://machinelearningmastery.com/machine-learning-in-r-step-by-step/