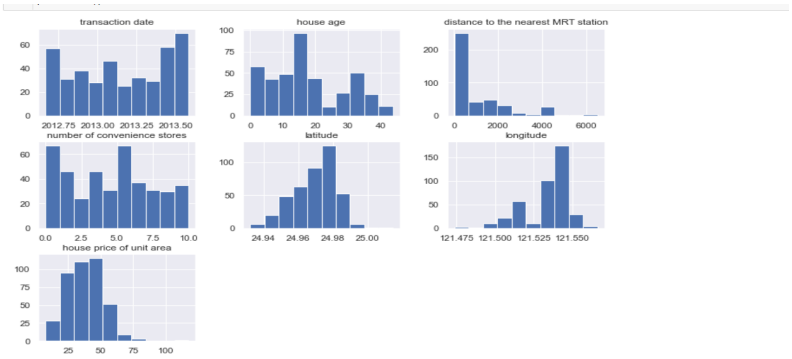
ASSIGNMENT2

ML LAB - 02

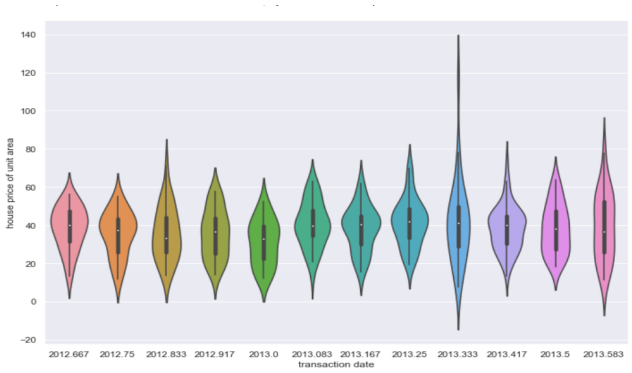
SREELAKSHMI CV
21BDA39

1. **Document 5-6 key insights from EDA and support each point with a visualization.**

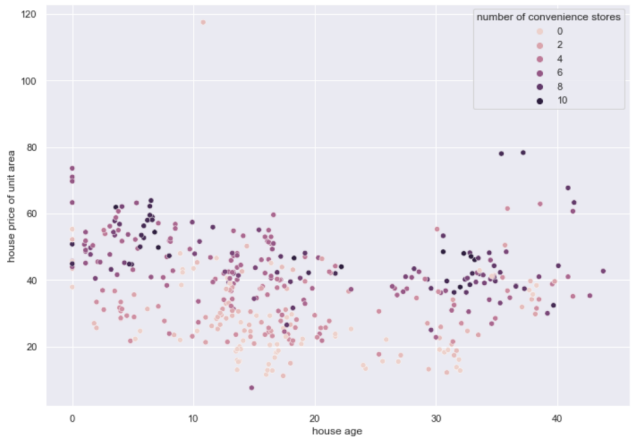
NORMALLY DISTRIBUTED VALUES



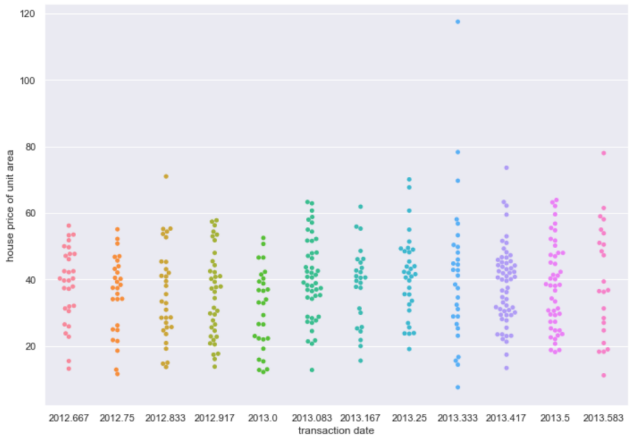
HOUSE PRICE IS HIGH IN TRANSACTION DATE 2013



PRICE OF THE HOUSE DEPEND ON AGE OF THE HOUSE



HOUSE PRICE IS HIGH IN TRANSACTION DATE 2013



2. Answer the following questions:

a) What are the assumptions of linear regression?

The Four Assumptions of Linear Regression is:

Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.

Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

Homoscedasticity: The residuals have constant variance at every level of x.

Normality: The residuals of the model are normally distributed.

b) How can we evaluate a Regression model? Define each metric and Its interpretation.

There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)

3. Mean Absolute Error(MAE)

R Square/Adjusted R Square

R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

R Square is calculated by the sum of squared of prediction error divided by the total sum of the square which replaces the calculated prediction with mean. R Square value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value.

Mean Square Error(MSE)/Root Mean Square Error(RMSE)

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result but it gives you

a real number to compare against other model results and help you select the best regression model.

Mean Absolute Error(MAE)

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

c) **Can R squared be negative**

It is possible to get a negative R-square for equations that do not contain a constant term. Because R-square is defined as the proportion of variance explained by the fit, if the fit is actually worse than just fitting a horizontal line then r square is negative.

d) **What is dummy variable trap.**

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. When we use *one-hot encoding* for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other

dummy variables. Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables for regression models leads to a ***dummy variable trap***. So, the regression models should be designed to exclude one dummy variable.

e) Is One Hot Encoding different from Dummy Variables?

One hot encoding can make values more than 2 while labeling like [1,0,0], [0,1,0], while dummy variables are only 0 and 1.

f) How is polynomial regression different from linear regression?

Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables