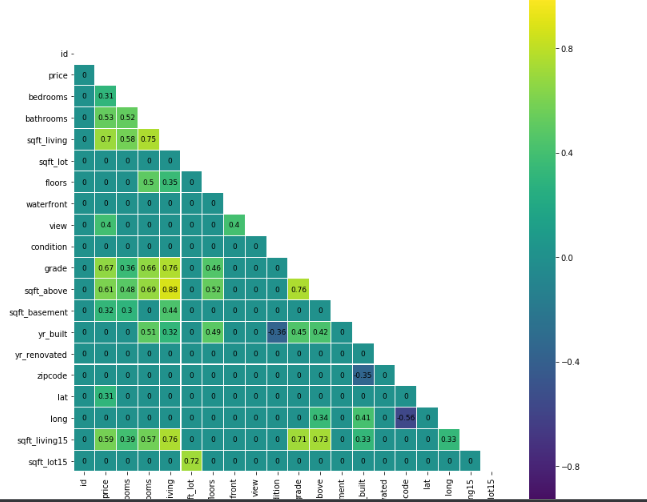
ASSIGNMENT 2

NAME: Yana Kaveramma A A

REGISTER NUMBER: 21BDA32

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

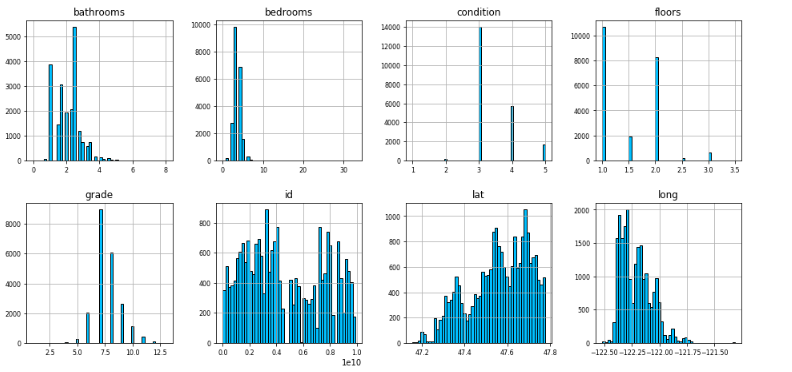
1(a)



The highest correlation is between sqft\_living & sqft\_above.

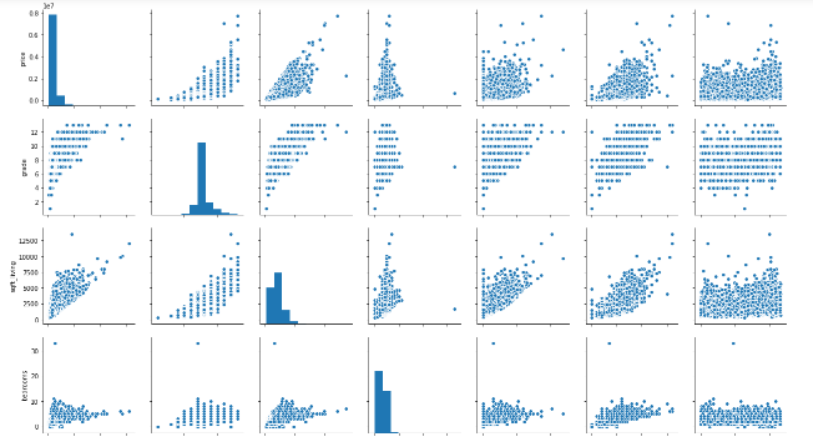
Multicolinearity exists.

1(b)



The histograms show the distribution the entire data. sqrt\_living, sqft\_above, price are left skewed. None of the data are normally distributed.

1(c)



Price vs every other column.

1. Answer the following questions:

2a. What are the assumptions of linear regression?

Linearity: The relationship between X and the mean of Y is linear.

Homoscedasticity: The variance of residual is the same for any value of X.

Independence: Observations are independent of each other.

Normality: For any fixed value of X, Y is normally distributed.

2b. How can we evaluate regression model? Define each metric and its interpretation.

Here are several different metrics used to evaluate regression models. It also depends on which kind of regression model you are using i.e., if it is a linear or non-linear regression, is it logistic regression is it a simple linear/non-linear or a mixed effects regression etc. I would limit my answer to linear and non-linear regression without mixed effects. If you are looking at a particular regression model then some of the following metrics would be useful

1. Metrics like correlation coefficient and coefficient of determination. Root mean square error (RMSE) is a common metric for comparisons
2. goodness of fit plots which include Pred vs. Observed, Pred vs Residuals (which include plots like residuals and weighted residuals). In many instances you may also want to look at Pred vs Independent Variable as well as residuals vs. independent variable.
3. When comparing models you may look for all of the above plus use statistical metrics and tests for selection of a better fit. These depend on whether the two models that are being compared are nested or non-nested. For nested models, comparison can be made with a chi-squared test. We can compare two models (nested or non-nested) using goodness of fit criteria like the Akaike information criteria (AIC) or the Bayesian information criterion (BIC). In general for non-nested models, the lower the AIC value the better is the fit to the data.
4. Other diagnostic plots are also used to evaluate the underlying distribution assumptions (assumptions of normality) which include plots like Q-Q plots.

2c. Can R square 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.

2d. 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.

2e. Is One Hot Encoding different from Dummy Variables?

In one-hot encoding, we create a new set of dummy (binary) variables that is equal to the number of categories (k) in the variable. For example, let’s say we have a categorical variable Color with three categories called “Red”, “Green” and “Blue”, we need to use three dummy variables to encode this variable using one-hot encoding. A dummy (binary) variable just takes the value 0 or 1 to indicate the exclusion or inclusion of a category.

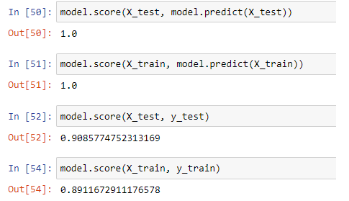
Dummy encoding also uses dummy (binary) variables. Instead of creating a number of dummy variables that is equal to the number of categories (k) in the variable, dummy encoding uses k-1 dummy variables. To encode the same Color variable with three categories using the dummy encoding, we need to use only two dummy variables.

2f. How is polynomial regression different from linear regression?

Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

2g. Interpret the screenshot:



Among the test data, there lies a model accuracy of around 0.9085, which means the model is highly accurate with the given test data. Test data usually accounts for a part of the data taken in order to peform tests and obtain accurate results.

Among the trained data, there lies a model accuracy of 0.8911, which means the model is also highly accurate with the given trained data. Trained data usually accounts for the majority of the data taken also to perform certain tests and obtain results.

Training data is the initial dataset you use to teach a machine learning application to recognize patterns or perform to your criteria, while testing or validation data is used to evaluate your model's accuracy.’

2h**. BONUS:** We saw Sweetviz fas an automated EDA option. What are the other options?

The 4 differed types of automated EDA are:

1. dtale
2. pandas profiling
3. sweetviz
4. autoviz

I also like AutoViz. AutoViz performs automatic visualization of any dataset with just one line of code. AutoViz can find the most important features and plot impactful visualizations only using those automatically selected features. Also, AutoViz is incredibly fast so it creates visualization within seconds.