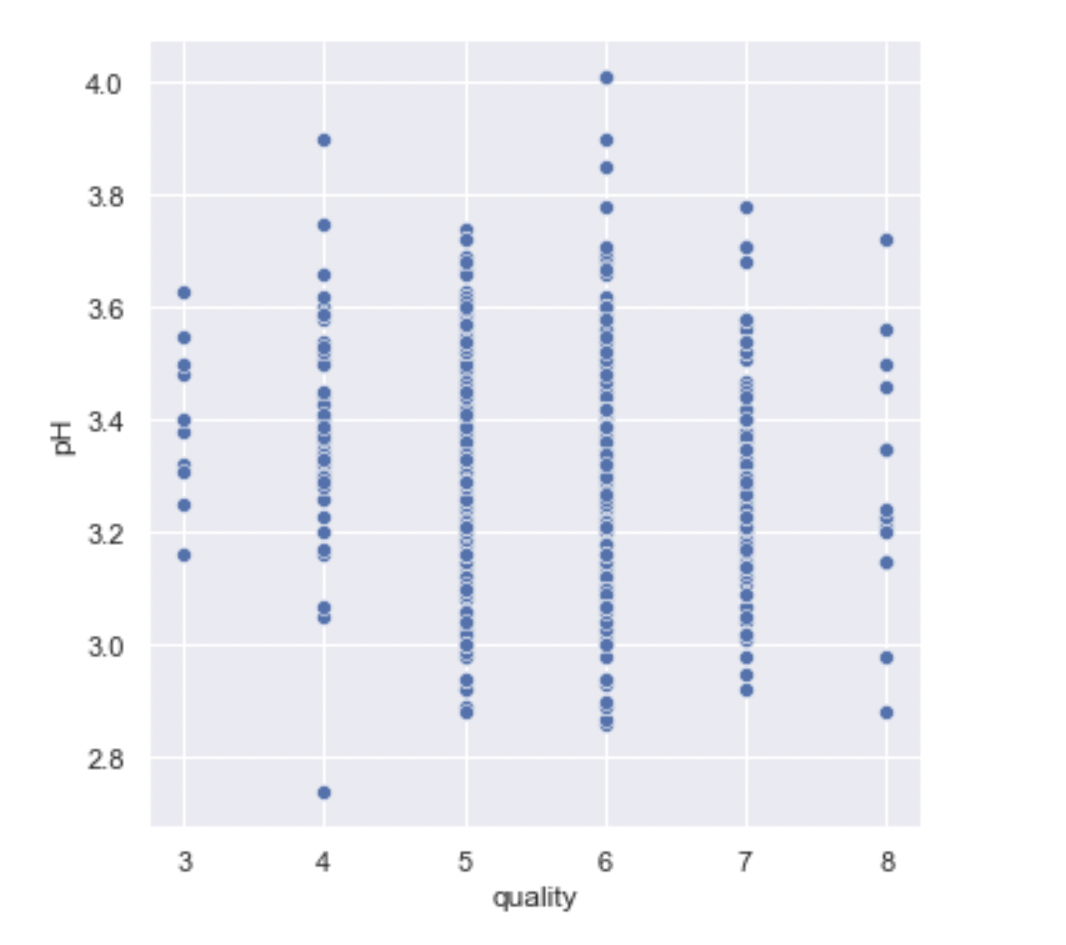
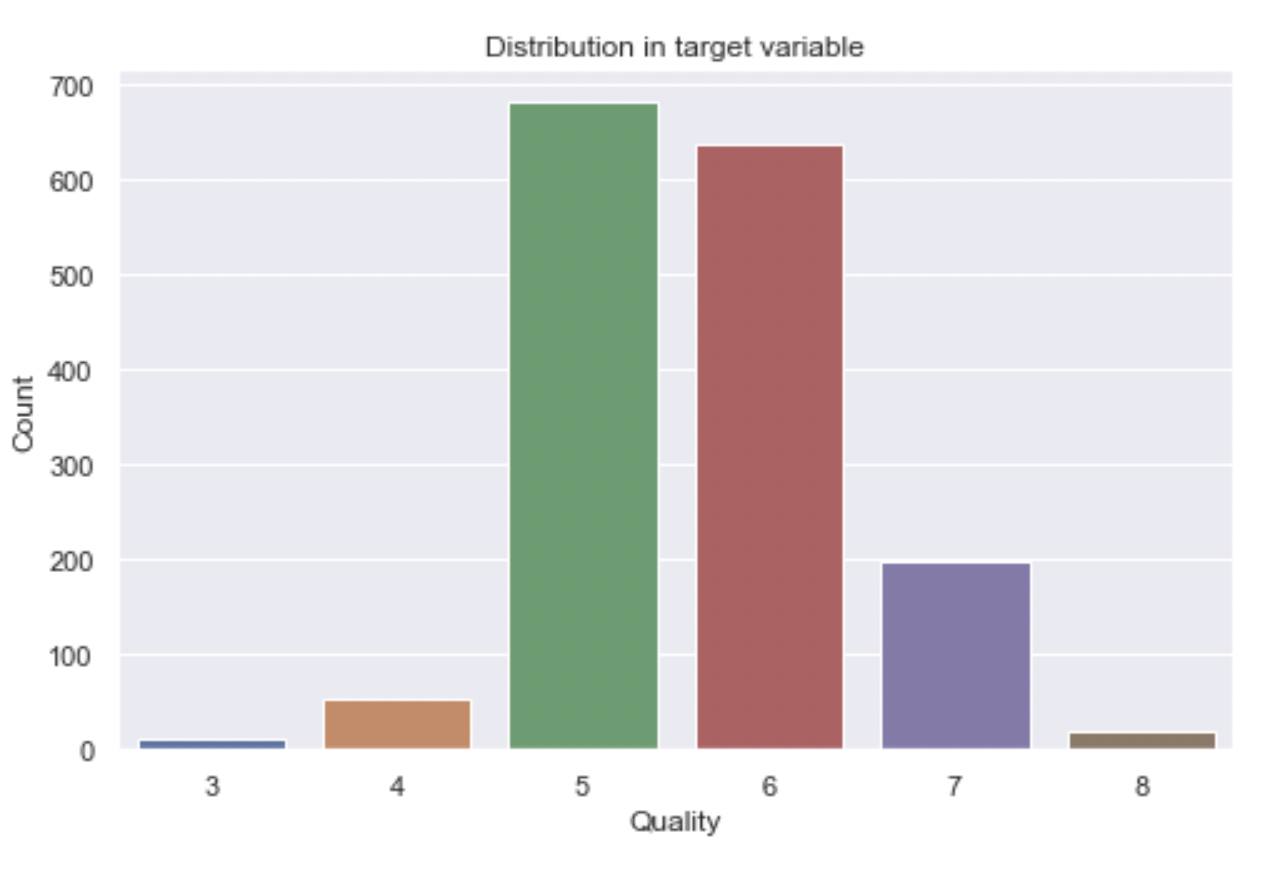
**ASSIGNMENT 02**

NAME-MAHIKA JADHAV REG NO-21BDA53

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



The above graph tells us that the points are clustered more around quality 5 and 6.The pH level is scattered more around quality 8.There are many Ph values ranging from 3.0 to 3.6.



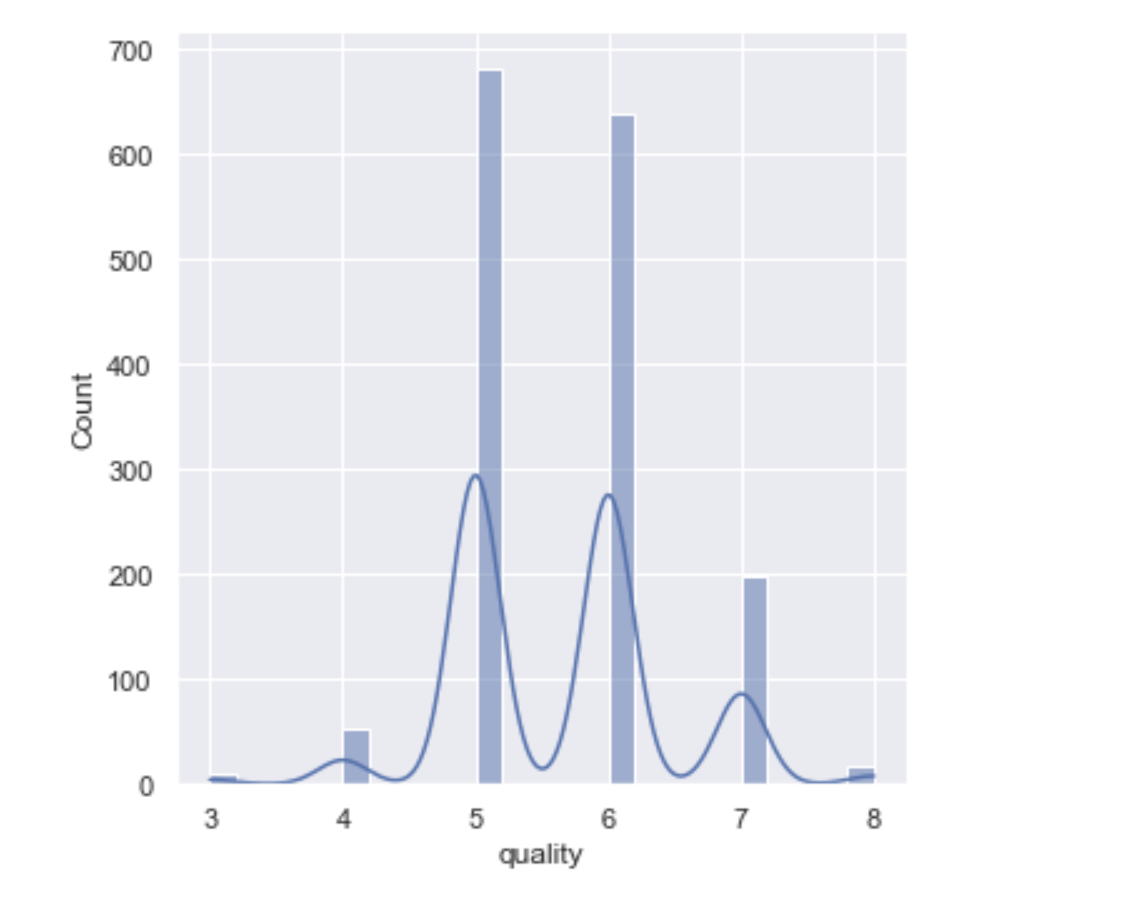
The above graph tells us that the quality of wine is rated 5 and 6 most number of

times.

Followed by 7 rating and 3 is the least rated wine quality.

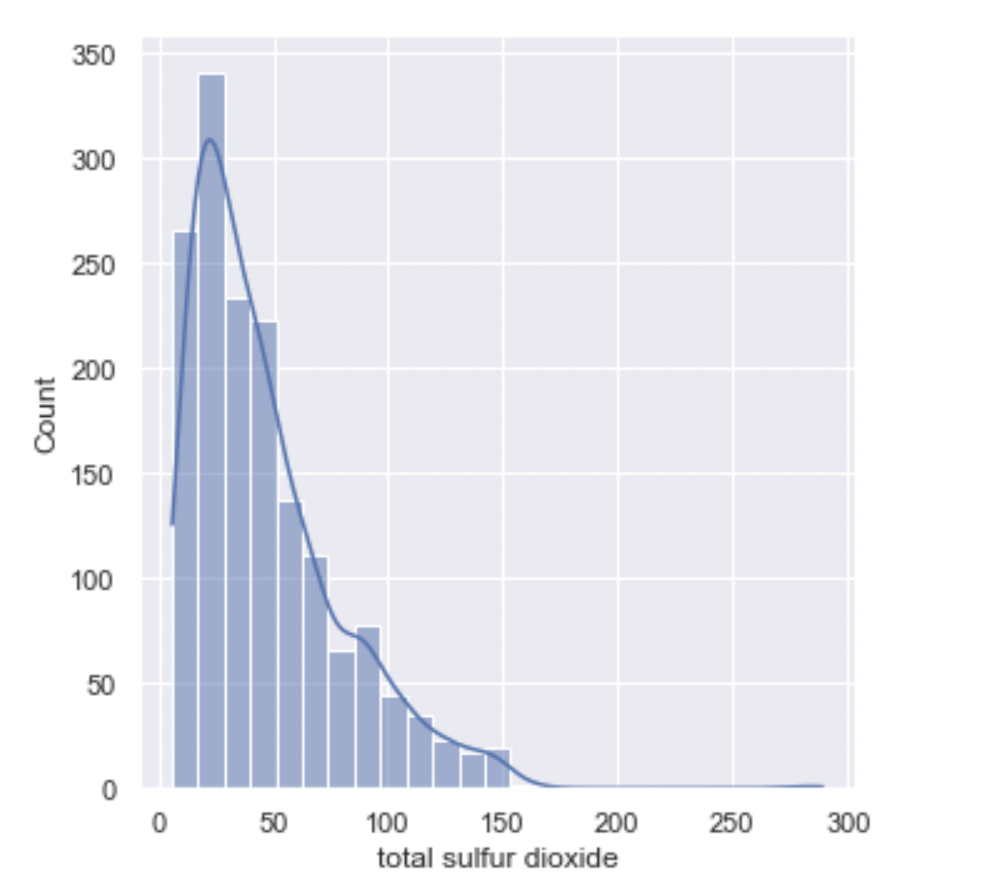


The highest correlation is between fixed acidity and citric acid=0.67.pH and fixed acidity have the least correlation.

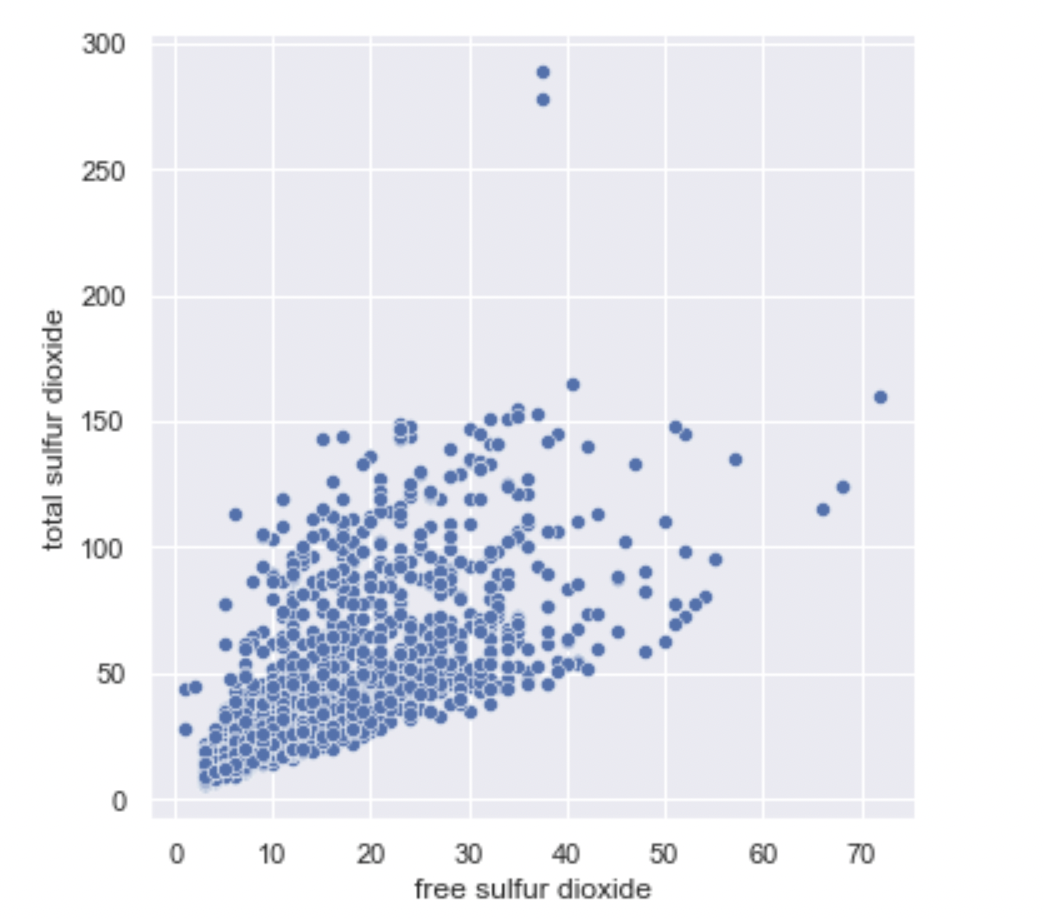


The above graph tells us that the quality of wine is rated 5 and 6 most number of

times.



Total sulphur dioxide lies the most between 0 and 50.



As total sulfur dioxide increases free sulfur dioxide also increases. free sulfur dioxide has clustered points around 5 to 30.

2.Answer the following questions:

What are the assumptions of linear regression?

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

ii. How can we evaluate a Regression model? Define each metric and its interpretation.

**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.

**Mean Square Error(MSE)/Root Mean Square Error(RMSE) :** While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

**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.

iii. 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.

iv. 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.**

v. Is One Hot Encoding different from Dummy Variables?

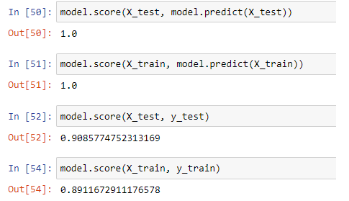
If there are n values in a category variable. It is converted into n variables using one-hot encoding, and n-1 variables using dummy encoding. If we have k categorical variables with n values each. Hot encoding produces kn variables, whereas dummy encoding produces kn-k variables.

vi. 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.

vii. Interpret the screenshot below from the notebook we discussed in class today:



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 perform 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.

viii. Bonus: We saw Sweetviz as an Automated EDA option. What are the other options? Try a few of them and share which one did you find the best.

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.