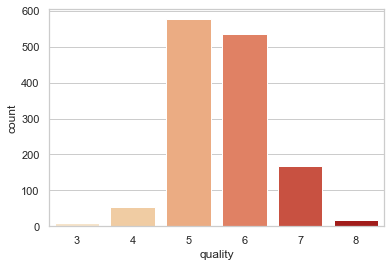
**MACHINE LEARNING LAB**

**ASSIGNMENT – 02(B)**

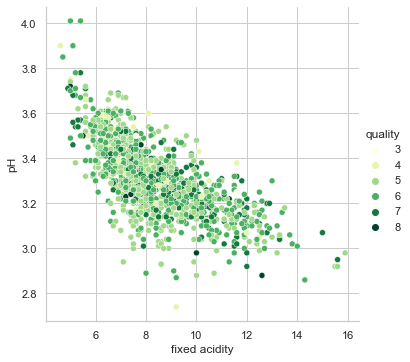
**21BDA11**

**1st May 2022**

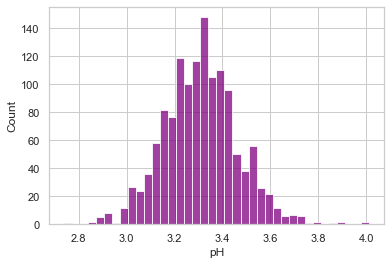
1. Document 5-6 key insights from EDA and support each insight with a visualization.
2. Most of the wine tasted (~560 of them) had a quality indicator of 5. This is followed closely by the quality indicator of 6. There were very few wines with quality indicator of 3.



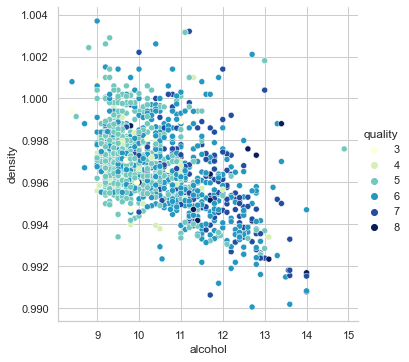
1. As the acidity increases, the pH value decreases.



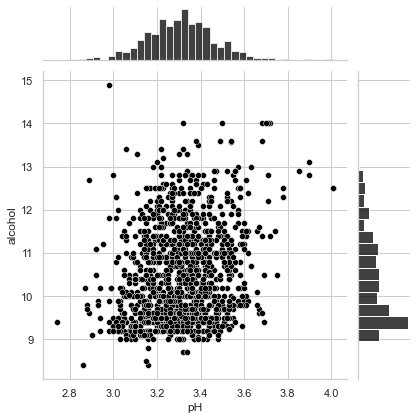
1. Most of the wines tasted had a pH value of 3.3



1. As the alcohol content increases, the density of the wine decreases. However, the quality of the wine increases.



1. With the increase in alcohol content, there is an increase in the pH levels.



1. Answer the following questions:
2. What are the assumptions of linear regression?

Linear regression is an analysis that assesses whether one or more predictor variables can explain the variation in a dependent variable.

There are certain assumptions that are made before one begins linear regression. They include:

1. Linear relationship

There must be a linear relationship between the predictor and the dependent variables.

1. Multivariate normality

The variables must have a normal distribution. This is checked using Q-Q plot.

1. No or little multicollinearity

The independent variables must not be highly correlated to each other.

1. Homoscedasticity

The residuals need to be evenly distributed for every variable.

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

A regression model provides a function that describes the relationship between the independent variable(s) and the dependent variable.

There are various ways in which one can evaluate a regression model.

The most common or the most naïve way of evaluating a model is to use the R-square value. Other than this, there are many more ways in which one can evaluate a model. They are called the metrics. Some of them include:

1. Co-efficient of Determination (R2)

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.

The value of R2 tells how much percentage of variation in the dependent variable can be explained by the independent of variables. An R2 of more than 0.7 (or 70%) is considered to be a high effect size.

1. Mean Squared Error (MSE)

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 an absolute number on how much your predicted results deviate from the actual number. One cannot interpret many insights from one single result but it gives a real number to compare against other model results and help select the best regression model. Lower the value of MSE, the better.

1. Mean Absolute Error (MAE)

Mean Absolute Error 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. A perfect mean absolute error value is 0.0, which means that all predictions matched the expected values exactly.

1. Can R squared be negative?

While most commonly, the lowest R2 is 0.0, sometimes it can go below 0 and can hence be negative. If the regression line is worse than using the mean value, the r-squared value that is calculated will be negative. A negative r-squared value can also occur when there is no constant present in the equation.

1. What is dummy variable trap?

Dummy variables are used mainly for categorical variables. Using dummy variables, one can give values to the categorical variables and hence makes it easier to analyse.

A dummy variable trap occurs when variables have a high correlation. This nullifies one of the assumptions of linear regression thereby defeating the purpose of the linear regression model. Regression models must exclude one dummy variable in order to avoid the dummy variable trap.

1. Is One Hot Encoding different from Dummy Variables?

Yes, one hot encoding is different from dummy variable encoding.

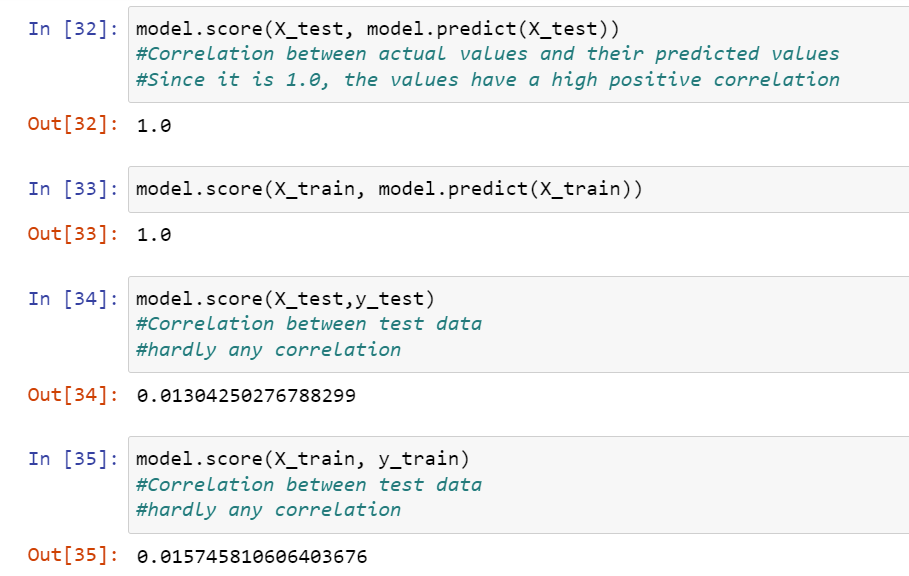
One hot encoding is a way of encoding categorical variables. It uses binary representation. If there are k number of distinct values for a categorical variable, OHE creates k number of dummy variables. Wherever the category is present, it marks as one, else it is zero. Due to the number of k dummy variables created, there is duplication.

In dummy variable encoding, if there are k number of distinct values for a categorical variable, there are k-1 number of dummy variables created. This does away with the duplication.

1. How is polynomial regression different from linear regression?

Polynomial regression is a specific sub-case of linear regression. The inverse does not hold true. Linear regression establishes the relationship between two or more variables by trying to fit them onto a straight line. Polynomial regression establishes the relationship between the independent and the dependent variables as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E (y |x). Polynomial Regression provides the best approximation of the relationship between the dependent and independent variable.

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



1. model.score(X\_test, model.predict(X\_test))

This command computes the correlation between the actual values of the variable which is contained in X\_test and the predicted values of that variable which is contained in X\_test. This is done on the testing or validating data. Here, the value is 1.0 meaning that the values and the predicted values are highly positively correlated or are the same.

1. model.score(X\_train, model.predict(X\_train))

This command computes the correlation between the actual values of the variable which is contained in X\_train and the predicted values of that variable which is contained in X\_train. This is done on the training data. Here, the value is 1.0 meaning that the values and the predicted values are highly positively correlated or are the same.

1. model.score(X\_test,y\_test)

This command computes the correlation between the testing data of variable X and variable Y. Here, the value returned is 0.01304 which indicates that there is a very minimal positive correlation between X and Y.

1. model.score(X\_train, y\_train)

This command computes the correlation between the training data of variable X and variable Y. Here, the value returned is 0.01547 which indicates that there is a very minimal positive correlation between X and Y.