**(B) In a word file, 1. Document 5-6 key insights from EDA and support each point with a visualization.**

1. Higher the residual sugar content in the wine, higher the density. As the density increases, the sugar content also tends to increase.

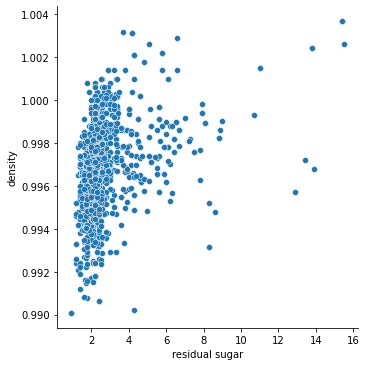


Figure 1: Relational Plot

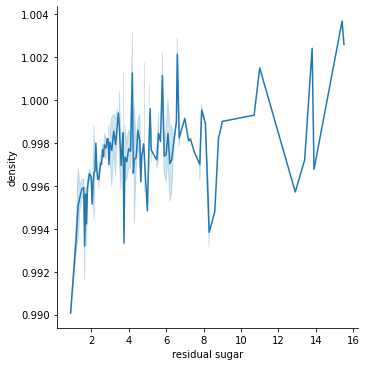


Figure 2: Relational plot with kind=line

1. Higher the alcohol content, higher the quality of the wine.

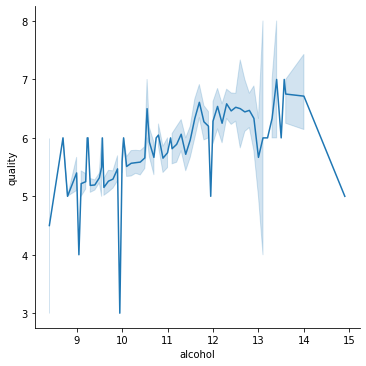


Figure 3: Relational Plot

1. Lower the density, greater the pH level

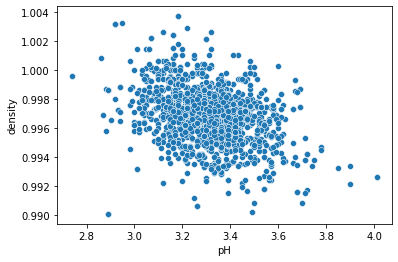


Figure 4: Scatterplot

1. As the fixed acidity of the wine increases, pH level decreases

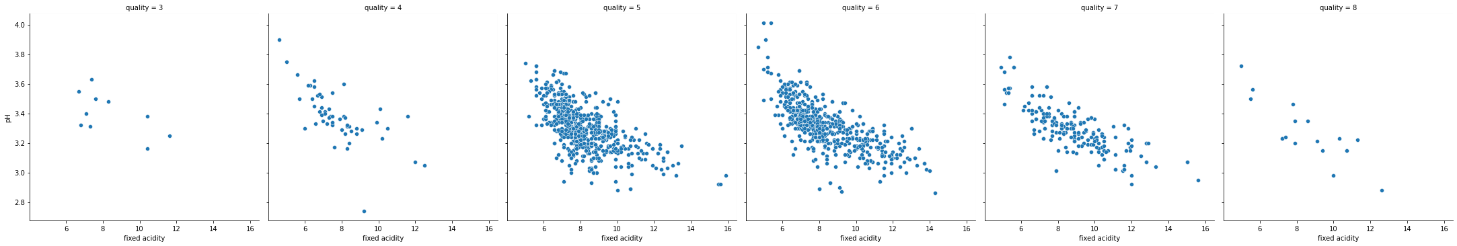


Figure 5: Relational Plot

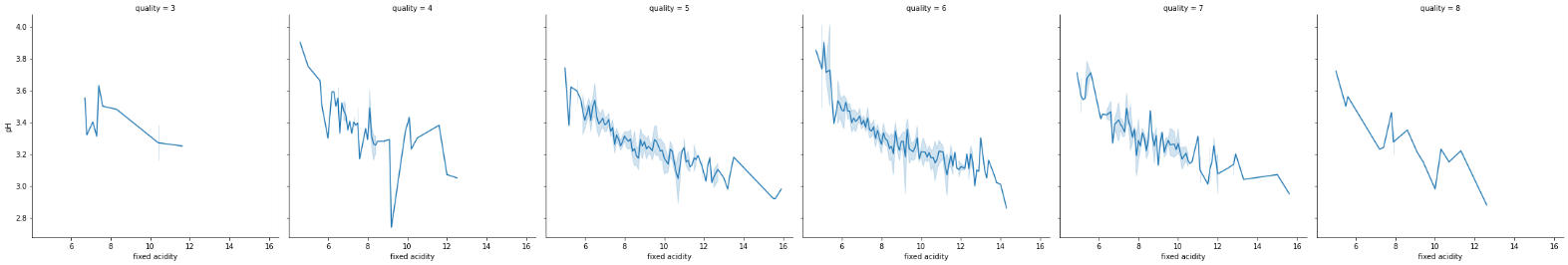


Figure 6: Relational Plot with kind=line

1. There are many overlapping points in the range 9 to 12, i.e., many wines have alcohol content in this range and there are some outliers present as well for the alcohol content

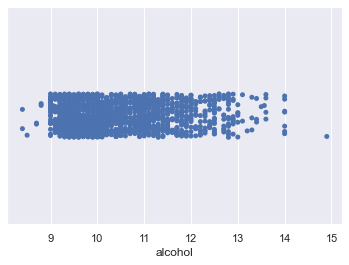


Figure 7: Strip Plot

1. Maximum values for the sulphate content present in the wine, range between 0.25 to 1.25, i.e., most wines have sulphate content in this range. There are present some extreme values for the sulphate content as well (can be treated as outliers)

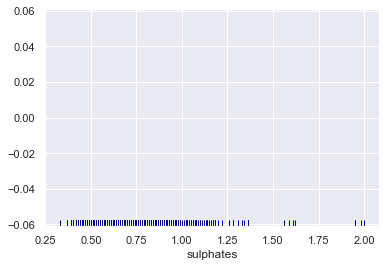


Figure 8: Rug plot

**2. Answer the following questions:**

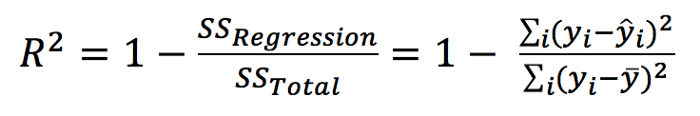
**I. What are the assumptions of linear regression?**

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**

1. **R Square/Adjusted R Square**

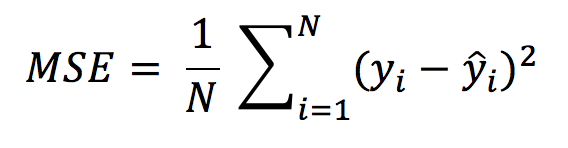
R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.



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.

1. **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.

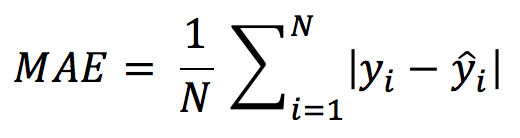


Mean Square Error formula

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.

1. **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.



Mean Absolute Error formula

Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. **MSE gives larger penalization to big prediction error by square it while MAE treats all errors the same**.

**III. Can R squared be negative?**

Yes. If the regression line is worse than using the mean value, the calculated r squared value will be 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?**

* 1. **How is polynomial regression different from linear regression?**

**Simple Linear Regression** establishes the relationship between two variables using a straight line. It attempts to draw a line that comes closest to the data by finding the slope and intercept which define the line and minimize regression errors. Simple linear regression has only one x and one y variable.

**Polynomial Regression**is a one of the types of linear regression in which the relationship between the independent variable x and dependent variable y is modelled 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.