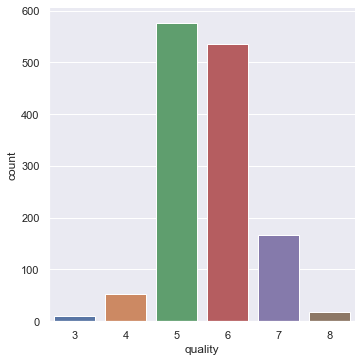
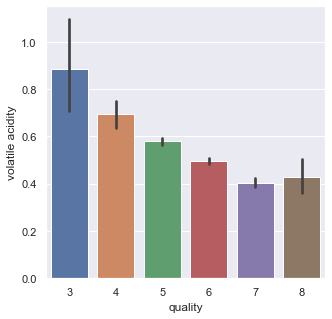
**(B)**

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

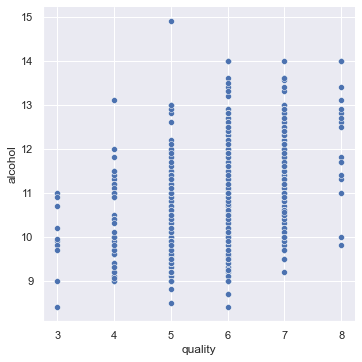
* Mainly the quality of wine is 5 followed by 6 according to our data.



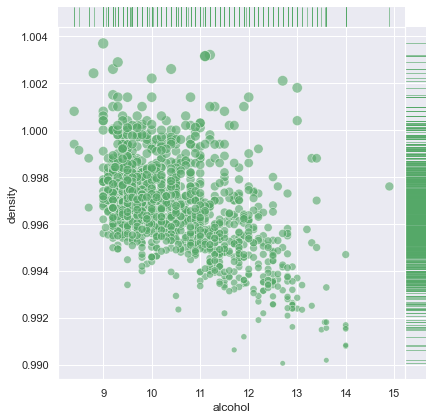
* To prepare good wine, quantity of volatile acidity should be less as quality and volatile acidity is inversely related.

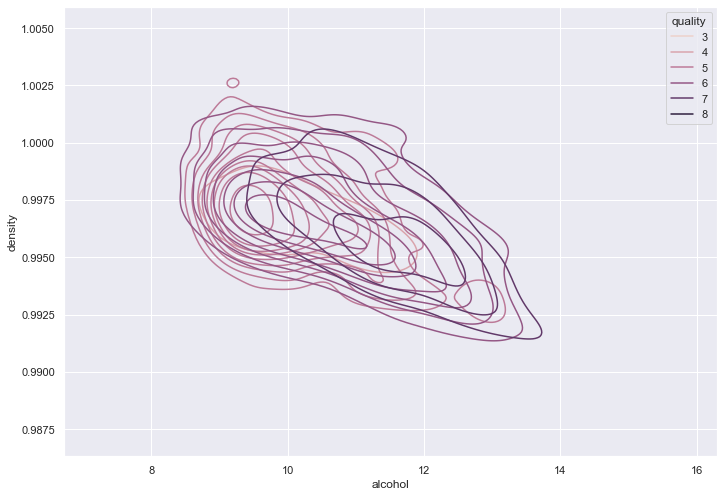


* Increase in the alcohol qty, increases the quality of the wine

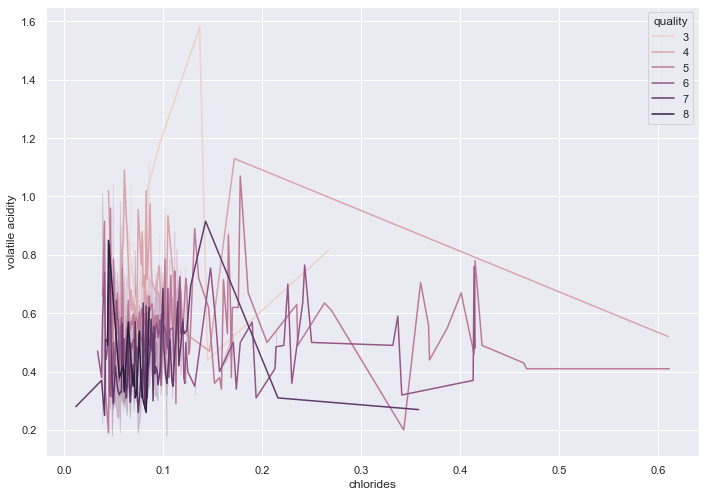


* If the density of the wine is more it means it has less alcohol which leads to low quality





* Chlorides and Volatile Acidity both are inversely related to the quality of the wine. So in order to have make good quality wine both the things should be in less amount.



**2. Answer the following questions:**

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

There are four assumptions associated with a linear regression model:

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
5. **Autocorrelation**: One of the analytical assumptions of linear regression is that the given dataset should not be autocorrelated.

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

The evaluation metrics are:

* **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. 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**.

* **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 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.

* **RMSE**

Root Mean Square Error (RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.

* **RMSLE**

Taking the log of the RMSE metric slows down the scale of error. The metric is very helpful when you are developing a model without calling the inputs. In that case, the output will vary on a large scale.

To control this situation of RMSE we take the log of calculated RMSE error and resultant we get as RMSLE.

* **R squared**

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.

* **Adjusted R Squares**

R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration of overfitting problem. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalize additional independent variables added to the model and adjust the metric to prevent overfitting issues.

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

Note that **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.

**v. Is One Hot Encoding different from Dummy Variables?**

**One-hot encoding**

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**

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.

**vi. 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** we add some polynomial terms to linear regression to convert it into Polynomial regression.

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

The coefficient of determination is the square of the correlation(r), if R2 is equal to 1, then the dependent variable can be predicted from the independent variable without any error.

The predicted values of X\_test are very close to the real values of X\_test, that’s why we got value 1 for coefficient of determination.

**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.**

Other options for Sweetviz are:

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

Out of these dtale seemed to be more user-friendly and easier to understand. It works more or less similar to tableau.