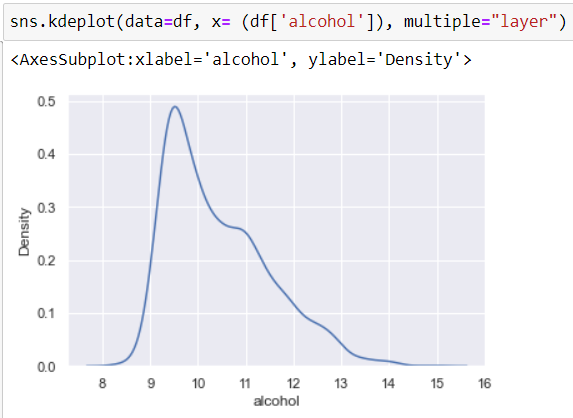
**ML Lab 2**

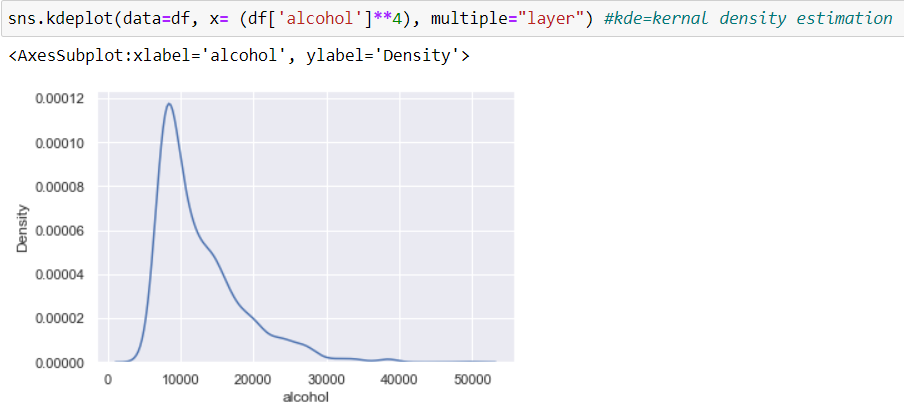
**Alexei Beuno. H**

**21BDA01**

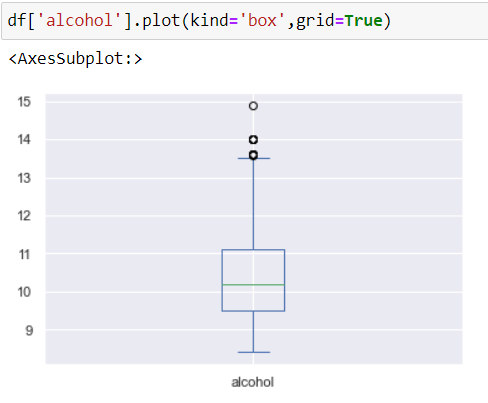
1. Document 5-6 key insights from EDA and support each point with a visualization.

1.



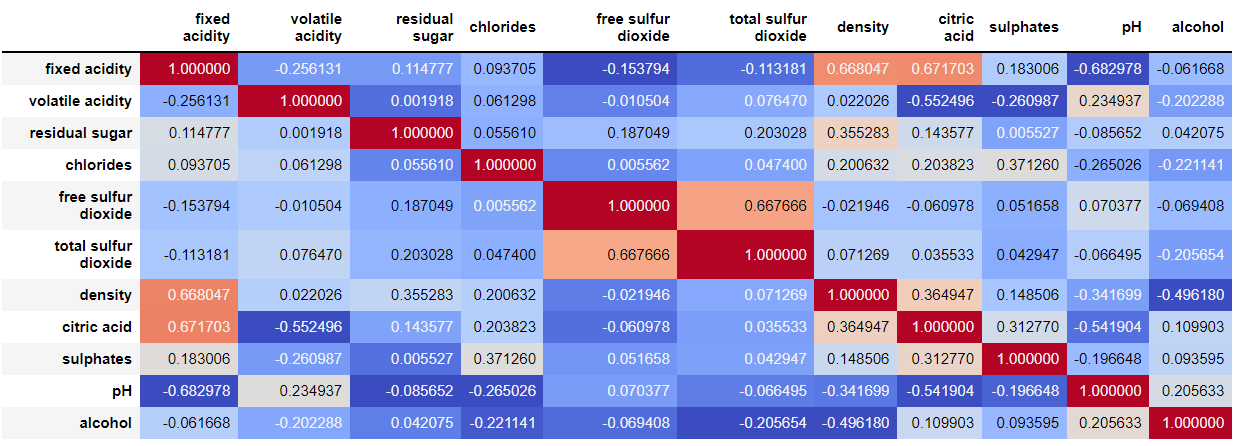
  
The target distribution was not normal and hence changed it to x^4 to make it normal.

1. To check for outliers



Alcohol has 3 outliers and the median alcohol concentration is around 10.2

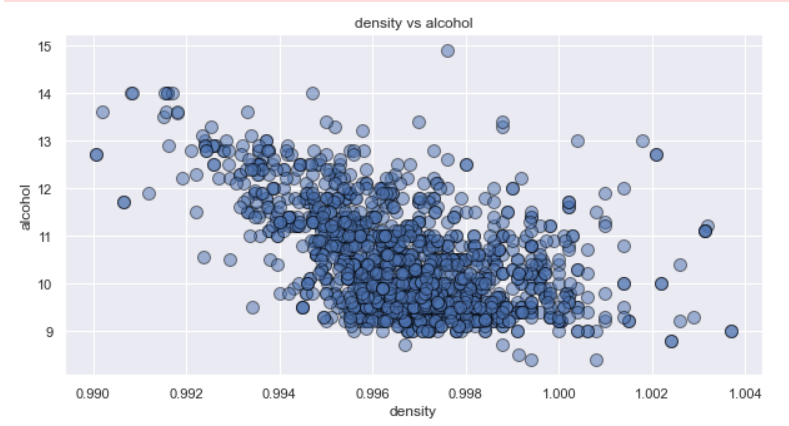
1. Correlation matrix for continuous columns



Fixed acidity and sulfur dioxide needs to be removed.

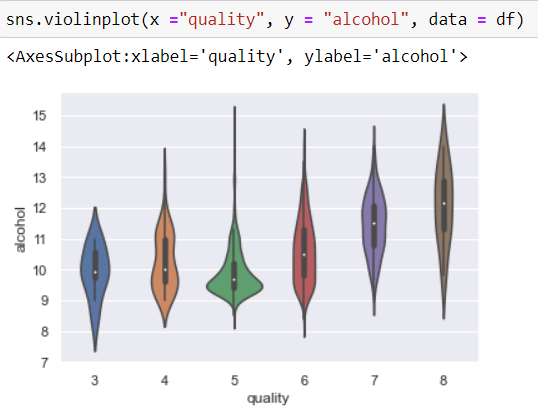
Scale of total sulfur is higher than others which may cause trouble.

4.



Lower the density of wine, higher is the alcohol concentration.

5. Quality vs Alcohol



Qualities 5 and 8 have higher alcohol concentration.

2. Answer the following questions

1. What are the assumptions of linear regression?

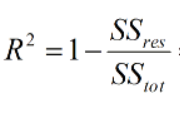
* Linear relationship: A linear relationship is said to exist between the dependent and the independent variables.
* No autocorrelation or independence: The residual terms are independent of each other.
* No multicollinearity: The independent variables should not be correlated.
* Homoscedasticity: The residuals must have constant variance.
* The error terms should be normally distributed.

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

* Coefficient of determination(r2) : It's a statistical measure that shows how much the independent variables explain the variation of dependent variables.
* Mean Absolute Error(MAE): Measure of absolute errors. It is calculated taking the absolute difference between the predicted and actual values.
* Mean Squared Error(MSE): Average of the squares of the errors. It is always non-negative, values closer to zero indicate best fit.
* Root Mean Squared Error(RMSE): Taking the square root of MSE, will give us the RMSE. A value of 0, will indicate a perfect fit.

1. Can R squared be negative?

R2 can be negative, indicating that the model is poorly fit.



Looking at this equation, if the SSres/SStot is more than 1, then the R2 is likely to be negative.

1. What is a dummy variable trap?

When the number of dummy variables created is equal to the number of values the categorical value can take on. This leads to multicollinearity, which causes incorrect calculations of regression coefficients and p-values.

1. Is One Hot Encoding different from Dummy Variables?

Dummy variable is when the category is converted to one column and the value 0 or 1 indicates if the category is present for each record.

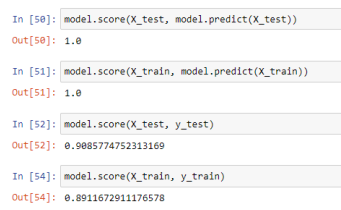
One Hot Encoding is similar to dummy variables, but one column is dropped as its value can be derived from other columns. This is to prevent multicollinearity and dummy variable traps.

Reference: [reference request - One hot encoding vs dummy variables best practices for explainable AI (XAI) - Artificial Intelligence Stack Exchange](https://ai.stackexchange.com/questions/26747/one-hot-encoding-vs-dummy-variables-best-practices-for-explainable-ai-xai)

1. How is polynomial regression different from linear regression?

Linear regression estimates the linear dependence of the independent and dependent variable whereas polynomial regression is a form of linear regression which estimates the non linear dependence of the variables.

1. Interpret the screenshot.



The R2 value of 1st and 2nd is 1 because we are determining how much variation of x\_test and x\_train is explained by x\_test and x\_train respectively. Since the prediction is from the same variables, R2 value is likely to be 1.

Whereas in the 3rd and 4th line, we see that x\_train and x\_test explains the variability of y\_train and y\_test respectively, therefore, it is not equal to 1.

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

There are three different types of libraries that perform automated EDA. They are,

Sweetviz, Pandas Profiling and autoviz

I personally liked the autoviz library, because it gives more information compared to the other libraries.

Code reference: [Automated EDA using pandas profiling,sweetviz,autoviz | by Guhanesvar | Analytics Vidhya | Medium](https://medium.com/analytics-vidhya/automated-eda-using-pandas-profiling-sweetviz-autoviz-4f15c4031a12#:~:text=EDA%20can%20be%20automated%20using%20a%20Python%20library,taken%20a%20dataset%20on%20heart%20failure%20from%20Kaggle.)