* Descriptive statistics – in ‘Section Summary’ you wrote two times: The data type of the 'quality' column is integer.
* In plot ‘"Number of Wines in Each Quality Category" I would consider to not include yticks because you annotate bars so thanks to removing numbers on y axis you can minimize data-ink ratio
* Hypothesis tests: you wrote nice functions for calculating two sample t-test and two tailed t-test but I would consider improving this section with either using: <https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ttest_ind.html> <https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html> or putting your own functions in a separate file (named utils or something like that and import them in exactly the same way as you import functions from scikit-learn) or at least putting them in one cell at the beginning of the analysis, right below IMPORTS. Also I would consider to use type hints 😊 <https://docs.python.org/3/library/typing.html>
* I think you should consider to be careful with strong conclusions with section summaries in hypothesis testing, because for instance in graph: ("Effect of Fixed Acidity on Quality") you can see that mean fixed acidity in wines with the score 3 is higher than with those with 4 score, and similar situation occurs with wines with score 7 and 8. However, you are plotting means and in that particular case, the number of wines with those scores differs significantly. You only have 10 wines with score 3 and 53 with score 4, also number of wines with score 8 is much lower than those wines with 7. And my point is that maybe if you would have comparable number of samples for those wines, means would change their values, and the trend would be more visible. We don’t know that, and most of all, we won’t know that, but should I really trust the statement of ‘fixed acidity decrease from quality 7 to quality 8’? I’m not sure about it.
* Correlation between features – I don’t know why you plotted all correlation separately for wines with different quality score? How would you explain it?
* Data Scaling should be conducted after splitting the dataset. This is important because you have to keep in mind that we always treat test data as ‘never seen before’. It means that you can’t scale all the data at once, because you practically don’t have access to test set. Thus, you perform fit\_transform on train set and just transform on test set. ‘fit’ method from scaler learns how to scale data based on training data and keeps in memory parameters of this. Then ‘transform’ apply this parameters on a data. So we want to learn from training and only apply this transformation on test set. You can read about the concept of data leakage (what happens in that particular case) in this article: <https://machinelearningmastery.com/data-leakage-machine-learning/>
* Before performing logistic regression on each feature from a dataset separately I would comment why you do it. I was truly surprise with that attitude (but I think it might make sense as long as you have only few features) to perform this task.
* Section summary that followed logistic regression in that section in my opinion is a little bit too shallow. First of all, I would order features in descending order of accuracy score starting from alcohol content. It helps to distinguish which features might have the strongest predictive power. Additionally I would comment on chlorides, density, fixed acidity, total sulfur dioxide and free sulfur dioxide that the values of accuracy scores are only slightly higher than 0,5 and that means that they predict only slightly better that the random choice. Is it a good indicator of good/bad quality? I would argue 😊
* Maybe you could also draw graphs with quality on X axis and each feature on Y axis and those graphs would somehow give you an idea if you can expect any predictive power of those features (<https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290> Assumption 2 ‘visual check’)?
* For alcohol content prediction I would also go for checking linear regression assumptions first. [Assumptions of Linear Regression | Towards Data Science](https://towardsdatascience.com/assumptions-of-linear-regression-fdb71ebeaa8b).