Nana Noda-Morgan

**DSC530 – Exploratory Data Analysis Final Term Project**

The main hypothesis of my analysis was that wines with higher alcohol content tend to have better quality ratings. Using the UCI Machine Learning Repository, I found a dataset on wine quality that was perfect for this purpose. I went through the various steps of EDA on the different properties of both red and white wines, including alcohol content, residuals sugar, acidity, pH, and sulphates. The goal was to identify variables that were predictors of the quality ratings.

The analysis showed us that alcohol content and sulphates were the most significant predictors for wine quality among the variables we examined. Specifically, alcohol content had a linear relationship with quality, which supported my original hypothesis that wines with higher alcohol content tend to receive higher quality ratings. The distribution analysis, as well as the scatter plots for alcohol vs. quality also showed us a strong positive trend. Other variables like residual sugar showed weaker relationships.

The main thing that stood out to me when working through this analysis was that the variables presented in the dataset were limited to numerical properties and were missing descriptors you would typically expect when referring to wines, such as aroma and flavor profiles. In addition, it was lacking in variables such as grape variety, production region, or anything related to the fermentation process. I believe including such variables would make the analysis more robust as it pertains to quality.

One of the primary assumptions in this analysis was that the numerical properties in the dataset, independent of the other descriptive characteristics mentioned earlier, could serve as strong predictors of wine quality. But it is important to keep in mind that wine quality is quite subjective and influenced by factors that are not measurable. Additionally, I assumed that the relationship between variables (such as alcohol content and quality) was mostly linear, which may not fully capture non-linear patterns between variables. Assuming a linear relationship between alcohol and quality may have oversimplified a potentially more complex relationship. Because the dataset likely does not capture all factors that could influence quality, these assumptions could affect the model’s accuracy.

My main challenge was deciding how to handle outliers and, in the process, researching wines to gain a little more domain knowledge so I can properly decide whether or not some values represented true extremes or were just errors. I can really see here why working with a subject matter expert would be extremely helpful as a data scientist.

Finally, the EDA and regression analysis I was able to perform provided valuable insights, but incorporating additional descriptive variables could improve the model. Moreover, this analysis highlighted the importance of a comprehensive approach when studying multifaceted data like wine quality.