Term Project Summary

Teresa Ferrill

Bellevue University

Abstract

My term project was an analysis of red wine to determine if there are some physicochemical properties in red wine that affect quality. Instead of determining how the chemicals come together in those wines identified as high in quality, I decided to determine if the variance in quality ranking can be explained by the input variables. The null hypothesis is that none of the variance in the quality ranking is explained by the input variables. Repeatedly, my analysis supported the null hypothesis. The outcome of my exploratory data analysis was that, while there are differences in the chemical makeup of various wine quality groups, the actual quality of wine is more subjective than objective.

*Keywords:* analysis, physicochemical, wine, quality

Term Project Summary

My term project was an analysis of red wine to determine if there are some physicochemical properties in the wine that affect quality. I used a red wine dataset from University of California Irvine’s Machine Learning Repository (UCI, 2009). The dataset is based on wines produced at Vinho Verde, a region in the north of Portugal. The dataset contains eleven variables plus an additional attribute to record quality, which is a sensory score between zero and ten.

Instead of determining how the chemicals come together in those wines identified as high in quality, I decided to determine if the variance in quality ranking can be explained by the input, or explanatory, variables. The null hypothesis is that none of the variance in the quality ranking is explained by the input variables. The alternate hypothesis is that the input variables contribution to the variance in the quality ranking is significantly different from zero. My analysis sought to either prove or disprove the null hypothesis.

My analysis included computation and comparison of correlation and covariance along with boxplots, scatter plots, and histograms. I targeted six input variables which included fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, and free sulfur dioxide. Review of the boxplots revealed that just about all variables had outliers, some with a substantial number. For some of my analysis, I grouped the dataset into three groups based on the quality then focused on those with the highest quality and the lowest quality.

Repeatedly, my analysis supported the null hypothesis. The outcome of my exploratory data analysis was that, while there are differences in the chemical makeup of various wine quality groups, the actual quality of wine is more subjective than objective. I used the fixed acidity chemical variable in the majority of my analysis. I chose this variable due to the fact it had a more normal distribution than most of the others. In the end, I believe my analysis would have been more complete had I repeated some of my analysis steps with other chemicals. I may have missed the opportunity to challenge the alternative hypothesis.

I believe the variable volatile acidity may have been helpful in the analysis. It may be that some of the assumptions I made were incorrect due to the fact I used fixed acidity by itself instead of doing additional analysis with other variables.

**Conclusion**

I believe the most challenging part of this analysis was interpreting the results of each analysis and having to arrive at a conclusion. For the most part, I still find it difficult interrupting the whole story the descriptive characteristics present. I need to become better at understanding how these values come together and the information they provide.

**References**

UCI. (2009, October 7). *Wine Quality Data Set*. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/wine+quality