John Manzo

DSC530-301T

August 14, 2021

**Course Project Summary Paper**

**Statistical/Hypothetical Question**

Research question: Does higher alcohol volume improve the quality of red wine?

Null Hypothesis: Alcohol volume does not have an effect on red wine quality.

Alternative Hypothesis: Alcohol volume does have an effect on red wine quality.

**Outcome of your EDA**

The overall outcome is I was able to reject the null hypothesis and answer the research question in the affirmative. The hypothesis testing allowed me to reject null (there is an effect), and the test statistics, comparative PMF plot, and regression analysis allowed me to answer the research question yes (i.e., the correlation is positive).

Moreover, variable-specific histograms, box plots, and summary statistics helped rule out some variables and gave me the confidence to include outliers. There are three reason I kept outliers in: 1) how few there were for the explanatory and dependent variables, 2) domain knowledge of red wine leads one to expect very few outstanding and very few terrible vintages, and 3) confidence in the data source that there were not misreported values. In short, I felt the small number if outliers included were important to the analysis, given the subject, and were all correctly recorded values.

**What do you feel was missed during the analysis?**

In an effort to utilize my six variables in one way or another, I feel I missed some opportunities to focus a little more directly on the research question with a couple of the tasks. I also made a point to not use the course text author’s custom classes and functions for the project, so I may have missed out of some better (or easier) ways of accomplishing the requirements. Although, that also allowed me to learn more about various Python libraries and code techniques.

**Were there any variables you felt could have helped in the analysis?**

I feel the selling price, grape/wine type (versus just red), and vintage year would have helped a great deal. I believe it is a fair assumption, until demonstrated otherwise, that at least one of those three variables correlates quite well with red wine quality. “Quality” in this dataset, by the way, is a 1-10 score given by three different wine assessors in blind taste tests.

**Were there any assumptions made you felt were incorrect?**

I went back and forth for weeks on whether I should exclude all outliers. I believe I made the right decision to keep them in the dataset for the reasons previously discussed, but I am certain others would think differently.

**What challenges did you face, what did you not fully understand?**

Perhaps the biggest challenge I faced was replicating (or attempting to replicate) the functionality of some of the course text author’s routines, such as CDF generation/manipulation. Weak knowledge areas tended to be in the areas of analytical distribution model selection, regression model evaluation statistics (beyond the basics), and deep residual analysis.