Summarising your data analysis procedures in up to 20 slides for a 5-minutes presentation. The slides should include the following contents:

Hi Dr Tran.

This is my presentation for SIT718 Real World Analytics – Assessment 2.

Our first task was to create scatterplots and histograms of the available data.

I paired the scatter plot and histogram of the independent variables together and we can see a positive skew displayed in the histograms of chlorides, sulfur dioxide and alcohol. Citric acid also skews positive though not as strongly as the other three. pH levels appear to be normally distributed. As does the quality though the stepped and clumped accumulation of data makes a visual assessment challenging.

Reviewing the central tendency, correlation and k-s results, only pH has a p-value above 0.05. The p-value for citric acid, chlorides, sulfur dioxide and alcohol fell well short of 0.05 as does the p-value for quality.

This leads to part two of the assessment. The selection of four independent variables and the associated data transformation. I omitted pH levels as its correlation coefficient is closest to zero, leaving us with citric acid, sulfur dioxide, chlorides and alcohol.

Based on their negative correlation, chlorides and sulfur dioxide both required negation. From the p-value, chlorides, quality and alcohol would require a log transformation to make the positive skew more normal. I applied a square root polynomial transformation to sulfur dioxide and alcohol as they are also positively skewed though not as strongly the other three variables.

Feature scaling was also applied to all the variables to get a common interval [0,1].

The fitting function that is the best fitting model for my data was the Power Mean with a p=2. Aside from Spearman Correlation, its measures of accuracy were greater than the alternative models. The Spearman correlation was greater in the Weighted Arithmetic Mean and Power Mean with a p=0.5 however across the other three measures of accuracy they were less.

The Ordered Weighed Average model was excluded as both correlation coefficient values were lower and the orness was above 0.5 suggesting the model favoured higher values – elevating the weighting for alcohol and citric acid in favour of chlorides and sulfur dioxide.

After transforming the predicted value back to the original Y scale, the provided varialbles returned a value of 4. The value is below both the median and mean quality of the sample data.

This result seems reasonable. By only looking at the raw data of the sample:

- the alcohol value is below the minimum value

- the chlorides value is above the maximum value

- citric acid would be the second highest value, exceed only by the highest value of 1 and the next cloest value of 7.9

- sulfur dioxide was around the median

-Discuss the best conditions (in terms of your chosen four variables) under which ahigher quality wine will occur.

Given the weighing of the model, it would suggest the best conditions for a higher quality of wine will occur when the chloride, citric acid and sulfur dioxide values are around or slightly below the mean and alcohol values around or slightly above the mean.

The implication for the model suggest there is a mechanism for determining the quality of a wine prior to it’s release to the market. A predictive model would be valuable tool for wine makers as they have to balance the expectation of their brand with the release of an equivalent quality of product. Knowing in advance where a wine’s quality might fall, a producer could withhold or outsource some wine if it is likely to fall below the quality value they were hoping for.

The limitation of the model used is the number of inputs used to predict the quality value and the nature of the quality value itself.

I am assuming the data used for my model is the same as that used for the article written by Cortez el al in 2009. In that study the authors relied on 11 attributes to determine a wine’s quality. For this model, I have used four of them. This could place greater weight on some variables where otherwise there should not be. In the model developed by the authors, the most important inputs to a red wine were: sulphates, pH levels, total sulfur dioxide and alcohol. Citric acid and chlorides ranked toward the lower end of importance. By contrast my model weights chlorides and alcohol highest followed by citric acid and sulfur dioxide.

The other limitation is on the quality variable. A wine’s quality is determined by a minimum of three sensory assessors (usually blind tasing) and then grading a wine on a range from 0 to 10 – bad to excellent. The final quality score is a median of the evaluations. The conversion of ordinal data from the sensory assessors to discrete data

In addition to the unknown number of sensory assessors providing feedback and the conversion of their ordinal feedback into discrete data, there is a lot of scope of distortion in the quality variable.