Week 4 Notes

Interesting Notes from the Dataset Research Paper: ***Modeling wine preferences by data mining from physiochemical properties***

p. 1: Three regression techniques were performed using simultaneous variable and model selection.

p. 2: Wine certification is assessed by physiochemical and sensory tests. Routine physiochemical tests generally include pH, alcohol, and density. Sensory tests rely on experts. Taste is the least understood of the human senses, making wine classification a difficult task.

p. 3: When continuous data is modeled, linear/multiple regression (MR) is the classic approach. Neural networks (NN) and support vector machines (SVM) have gained popularity, due to greater flexibility and nonlinear learning capabilities; often attain high predictive performances.

When data mining (DM) is used, variable and model selection are key. Models that are too complex may overfit the data and models too simplistic offer limited learning capabilities.

There is great capability for data mining to be used to predict quality based on physiochemical properties---that research is scarce/uses small datasets.

p. 4: Authors argued using a sensory taste panel is difficult and used a neural network (NN) that was fed with data from an electronic tongue.

p. 5: Highlights the main contributions of the work.

p. 6: Materials and methods explanation: samples collected between May 2004 and February 2007. Sample evaluated by a minimum of three sensory assesors.

p. 7: The scale that was used ranged from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations.

Data mining approach with mention that regression was used, preserving the order of the preferences. Also describes their confusion matrix.

p. 8: A "more robust" estimation is the k-fold cross-validation, where data is divided into k-partitions of equal size.

p. 11: They indicated that R was the exclusive model used for all experiments reported, specifically the RMiner library package.

p. 17-18: Reiterates the importance of the work for the wine industry, both for certification and under Portugese law (requiring sensory tests to be completed by human testors).

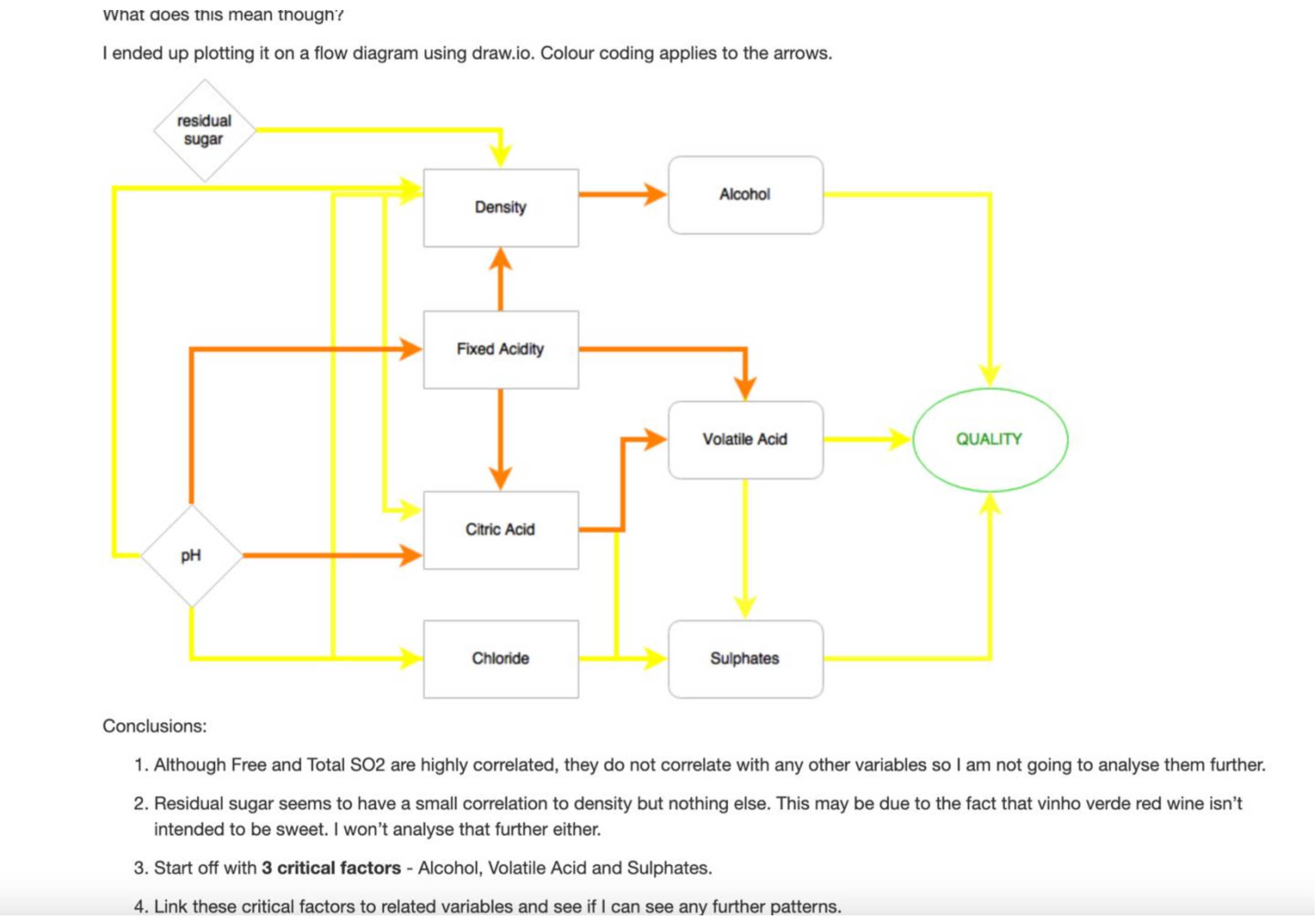
p. 23-29: Helpful graphs/tables to show their data analysis.

Interesting Notes from *Red Wine Vino Verde Analysis*:

This data analysis doesn't have regressions but contains some Bivariate and Multivariate Analysis. At the end of the project when questions for possible exploration are proposed, the use of k-means:

> "A possible next step could be to apply machine learning: Can I predict top quality wines based on the links I've discovered? What does KMeansBest say about my features? Perhaps I could create some new composite features? PCA of alcohol/density springs to mind."

What I really liked from this data project was the flow chart that was depicted as part of the explanation to one of the analyses sets:



What does this mean though?

I ended up plotting it on a flow diagram using draw.io. Colour coding applies to the arrows.

Conclusions:

1. Although Free and Total SO2 are highly correlated, they do not correlate with any other variables so I am not going to analyse them further.
2. Residual sugar seems to have a small correlation to density but nothing else. This may be due to the fact that vinho verde red wine isn't intended to be sweet. I won't analyse that further either.
3. Start off with **3 critical factors** - Alcohol, Volatile Acid and Sulphates.
4. Link these critical factors to related variables and see if I can see any further patterns.

Other work I’m looking at:

RED AND WHITE WINE DATA ANALYSIS PREDICT QUALITY OF WINE
(LASSO Regression and Linear Model Discussion in Chapter 5)

https://github.com/YangDS/Vinho-Verde-Wine-Quality
Stepwise regression is used in this code

Project description: “This projects demos how to use R for analyzing the wine quality data (https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/) using linear regression, logistic regression, random forest, and other clustering algorithms.”