Project Scope

**Description**

A wine merchant wishes to expand their knowledge and offerings in terms of wine selection. The client requires information on success of various wines by region, country, price, and variety. They want to target future inventory with rating, price, and popularity in mind. In addition, the client would like to know if any variables predict wine rating. Data has been scraped from the accredited Wine Enthusiast Magazine. Various methods of data analysis will be used to gather insights on wine reviews and ratings.

**Deliverables**

- A dashboard or dashboards including the most insightful visualizations to tell the story of the varying wine performances.

- Executive summary report examining text data from wine descriptions specific to top wine varieties and regions using text data cleaning and pre-processing, tokenization, topic modeling, sentiment analysis, and correlation analysis.

- Executive summary containing predictive analysis and findings for variables that determine wine rating.

- Write-up containing final recommendations and conclusions.

**Milestones**

**Nov 4:** Determine and justify analytical methods to be used throughout the project.

**Nov 8:** Implement data hygiene strategies to clean and wrangle data. This includes transforming the data, removing missing data, selecting pertinent variables, etc.

**Nov 11:** Use descriptive analytics strategies and visualizations to gain an in-depth understanding of the dataset.

**Nov 15:** Develop sampling plan.

**Nov 18:** Implement sampling and cross-validation in preparation for predictive modeling.

**Dec 2:** Conduct predictive and prescriptive modeling to determine variables that potentially determine wine points (ratings) and to simulate future variable scenarios.

**Dec 8:** Submit final deliverables.

**Assumptions**

- Cross-validation will involve a 70/30 train/test split of the data.

- K-fold validation will have 10 folds.

- Random number seed 1234 for predictive models.

- Use a tune grid for model tuning.

- Confidence interval used will be 0.95.

**Constraints/Limitations**

- The dataset is fairly large. This may cause time and performance issues when running queries in SQL and scripts in R. This will be especially burdensome when attempting to tokenize the text data.

- Software options available are limited to R, SQL, Excel, and Tableau.

- There are hundreds of varieties and regions in the dataset. In order to effectively implement statistical modeling, these variables must be converted to factors in R. Sampling will have to be used to reduce the amount into a workable quantity.