Wine Analysis: DSCI 100 Project Proposal

Daniel Alimohd, Alexandria Ahluwalia

Introduction

What makes a good wine? What separates an okay glass of wine from a great glass of wine? Many will answer these questions with a certain growing region, winery, or grape variety. But we would like to take this analysis further and see what top chemical factors are responsible for differentiating the quality of wine. We will perform our analysis by using the wine quality data set. This data set contains observations from a variety of Red and White Vinho Verde (wine only from a special region in Portugal). Each separate wine observation contains its perceived “quality” from wine experts and information on the content of alcohol, sugar, acidity, density, citric acid, sulfur dioxide, pH, and sulphates. The quality in the data set is measured on a scale from 0 to 10, but only quality ratings from 3 to 9 appear in the data set. Through the analysis of this information, we hope to find trends for which chemical factors have the most significant impact on wine quality and what makes both the red and white Vinho Verde wines good.

Methods

This dataset is in a downloadable .csv format and is composed of two separate .csv files for both red and white wine. This dataset uses semicolons as the delimiters. It will be downloaded locally from the machine learning repository at UC Irving into our data folder and then read into R using the read_delim function as part of the tidyverse library.

```
red_wine <- read_delim("data/winequality-red.csv",delim=";")
white_wine <- read_delim("data/winequality-white.csv",delim=";")
```

The wine csv files provided are almost good enough to start analysis. However our column names currently use spaces in the csv file. We will change the column names to use underscores using gsub() from base R, so it be easier to plot and perform other operations on the data such as filtering out all the wines with a quality 9 score and summarizing the data in a table.

```
# replacing whitespace in column names to underscores
names(red_wine) <- gsub(" ", "_", names(red_wine))
names(white_wine) <- gsub(" ", "_", names(white_wine))
```

Visually, we will use a bar chart to explore the distribution of the quality rankings as the data seems to have a lot of 5’s and 6’s. To then analyze the data, we will visualize our data (both red and white wine separately) through use of scatter plots. We will explore on the variable on the y-axis and the quality on the x-axis. Since the quality scores are all integers, we will make use of the jitter geom to avoid overplotting on each quality line, making the charts easier to read. We will also make use of transparency to discern dense areas more easily.

After plotting, we will see trends within the data and can eliminate the chemical variables that have no relationship. We will then analyze the variables that have the most prominent trends and find which has the most significant impact on the quality of the wine which we will do through regression. We plan to take the top 4 variables with the strongest/most dominant trends to then do the regression with only those.

```
# Scatterplot example
variable_plot <- red_wine %>%
   ggplot(aes(x = quality, y =  variable)) +
       geom_point(position = "jitter", alpha = 0.3) +
       xlab("Quality") +
       ylab("Variable (unit)") +
       ggtitle("Variable vs Quality")
```

Will we conduct the regression using the caret package with 4 variables to see which trend is the strongest and affects the quality the most. We will use the variables first individually and then all 4 together.

Expected outcomes and significance

We expect to find the variables that are the most significant in affecting the quality for both the red and white wines. These findings could help wine producers tweak their production methods to create better wine and as a result have a positive outcome with their business. The outcome of this analysis can also be significant as the effects of climate change could impact the growth of the grapes that compose the wine in that region and as a result, impact their chemical composition. Lastly, it will help us know which Vinho Verde wines to buy to ensure good quality and good times.