

# Project Proposal

Bob Ding, Lynn Fan, Alice Jiang

11/08/2018

```
library("cowplot")

## Loading required package: ggplot2
##
## Attaching package: 'cowplot'
## The following object is masked from 'package:ggplot2':
##
##      ggsave
library("tibble")
library("dplyr")

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##      filter, lag
## The following objects are masked from 'package:base':
##
##      intersect, setdiff, setequal, union
library("ggplot2")
library("broom")
library("knitr")
```

## Introduction

### Project Goal: Explanation

To identify variables that are important in explaining variation in the response.

We are interested in researching **what factors contribute to the quality of wine** for different types of red vinho verde from Portugal. This data set was used to predict quality of wine for future wine certification, complementary to human wine tasters, in the paper we cited (*Modeling wine preferences by data mining from physicochemical properties*). We believe that this dataset can also be used to analyze what chemical factors are attributable to the final rating of wine (measured by the variable quality). If we can understand how chemical factors affect the wine quality, it may shed light on future R&D directions for chemical methods that could improve/preserve wine quality.

In the beginning stage of the modeling process, we will **try to include all variables** that might be related to the response variable (quality), even if they are not statistically significant. By doing so, we can maximize the fitness of this model. We would also like to **mean-center** all numerical explanatory variables to obtain more meaningful interpretation of each chemical factor by its statistical coefficient. At the end, we will produce a regression model that best explains how different chemical compositions of the wine (red vinho verde) determines the variation in wine quality.

## Data

```
# input data source
data <- read.csv("./redwine_quality.csv")
```

```
# overview of the dataset
glimpse(data)
```

```
## Observations: 1,599
## Variables: 12
## $ fixed.acidity      <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, ...
## $ volatile.acidity  <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, ...
## $ citric.acid        <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, ...
## $ residual.sugar     <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2...
## $ chlorides          <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, ...
## $ free.sulfur.dioxide <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15...
## $ total.sulfur.dioxide <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, ...
## $ density            <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0...
## $ pH                 <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, ...
## $ sulphates          <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, ...
## $ alcohol            <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, ...
## $ quality            <int> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, ...
```

There are 1,599 observations and 12 variables in this dataset.

```
# summary statistics of each variable
summary(data)
```

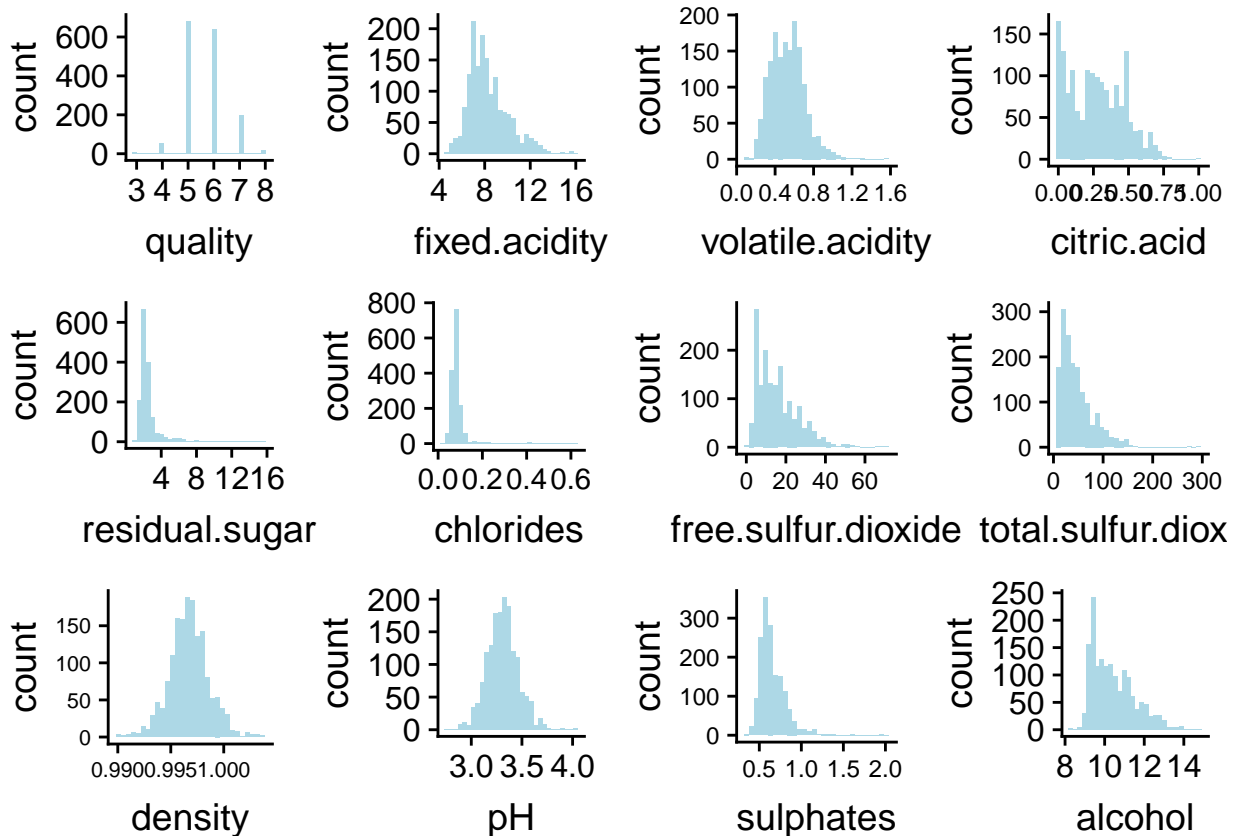
```
## fixed.acidity    volatile.acidity    citric.acid      residual.sugar
## Min.   : 4.60    Min.   :0.1200    Min.   :0.000    Min.   : 0.900
## 1st Qu.: 7.10    1st Qu.:0.3900    1st Qu.:0.090    1st Qu.: 1.900
## Median : 7.90    Median :0.5200    Median :0.260    Median : 2.200
## Mean   : 8.32    Mean   :0.5278    Mean   :0.271    Mean   : 2.539
## 3rd Qu.: 9.20    3rd Qu.:0.6400    3rd Qu.:0.420    3rd Qu.: 2.600
## Max.   :15.90    Max.   :1.5800    Max.   :1.000    Max.   :15.500
## chlorides        free.sulfur.dioxide    total.sulfur.dioxide
## Min.   :0.01200    Min.   : 1.00        Min.   : 6.00
## 1st Qu.:0.07000    1st Qu.: 7.00        1st Qu.: 22.00
## Median :0.07900    Median :14.00        Median : 38.00
## Mean   :0.08747    Mean   :15.87        Mean   : 46.47
## 3rd Qu.:0.09000    3rd Qu.:21.00        3rd Qu.: 62.00
## Max.   :0.61100    Max.   :72.00        Max.   :289.00
## density          pH          sulphates      alcohol
## Min.   :0.9901    Min.   :2.740    Min.   :0.3300    Min.   : 8.40
## 1st Qu.:0.9956    1st Qu.:3.210    1st Qu.:0.5500    1st Qu.: 9.50
## Median :0.9968    Median :3.310    Median :0.6200    Median :10.20
## Mean   :0.9967    Mean   :3.311    Mean   :0.6581    Mean   :10.42
## 3rd Qu.:0.9978    3rd Qu.:3.400    3rd Qu.:0.7300    3rd Qu.:11.10
## Max.   :1.0037    Max.   :4.010    Max.   :2.0000    Max.   :14.90
## quality
## Min.   :3.000
## 1st Qu.:5.000
## Median :6.000
## Mean   :5.636
## 3rd Qu.:6.000
```

```
## Max. :8.000
```

```
p1 <- ggplot(data = data, aes(x = quality) ) + geom_histogram( fill= "lightblue")
p2 <- ggplot(data = data, aes(x = fixed.acidity) ) + geom_histogram(fill= "lightblue")
p3 <- ggplot(data = data, aes(x = volatile.acidity) ) + theme(axis.text=element_text(size=9)) + geom_h
p4 <- ggplot(data = data, aes(x = citric.acid) ) + theme(axis.text=element_text(size=9)) + geom_histo
p5 <- ggplot(data = data, aes(x = residual.sugar) ) + geom_histogram(fill= "lightblue")
p6 <- ggplot(data = data, aes(x = chlorides) ) + theme(axis.text=element_text(size=11)) + geom_histogr
p7 <- ggplot(data = data, aes(x = free.sulfur.dioxide) ) + theme(axis.text=element_text(size=9)) + geom
p8 <- ggplot(data = data, aes(x = total.sulfur.dioxide) ) + theme(axis.text=element_text(size=9)) + ge
p9 <- ggplot(data = data, aes(x = density) ) + theme(axis.text=element_text(size=7.5)) + geom_histogr
p10 <- ggplot(data = data, aes(x = pH) ) + geom_histogram(fill= "lightblue")
p11 <- ggplot(data = data, aes(x = sulphates) ) + theme(axis.text=element_text(size=9)) + geom_histogr
p12 <- ggplot(data = data, aes(x = alcohol) ) + geom_histogram(fill= "lightblue")
```

```
plot_grid(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,ncol = 4,nrow = 3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Most of the 12 variables are pretty normally distributed. But variables like free.sulfur.dioxide, total.sulfur.dioxide, and alcohol are slightly skewed to the right.

## Variable Description

### Response Variable:

- **quality:** the quality of the wine (a score between 0 and 10)

### Explanatory Variables:

- **fixed.acidity:** ( $g(\text{tartaric acid})/dm^3$ ) the amount of acid in wine that's not volatile (do not evaporate fast)
- **volatile.acidity:** ( $g(\text{acetic acid})/dm^3$ ) the amount of acetic acid in wine; at high levels can lead to an unpleasant and vinegar taste in wines
- **citric.acid:** ( $g/dm^3$ ) citric acid is found in small quantities, and can add freshness and flavor to wines
- **residual.sugar:** ( $g/dm^3$ ) the amount of sugar left after fermentation stops; generally greater than 1 gram/liter in wines and wines with greater than 45 grams/liter are considered sweet
- **chlorides:** ( $g(\text{sodium chloride})/dm^3$ ) the amount of salt in the wine
- **free.sulfur.dioxide:** ( $mg/dm^3$ ) the free form of  $SO_2$  exists in equilibrium between molecular  $SO_2$  (as a dissolved gas) and bisulfite ion
- **total.sulfur.dioxide:** ( $mg/dm^3$ ) amount of free and bound forms of  $SO_2$ ; at free  $SO_2$  concentration over 50ppm,  $SO_2$  becomes evident in the smell and taste of the wine

- **density:** ( $g/cm^3$ ) the density of the liquid, which is close to the density of water depending on the percent alcohol and sugar content in the wine
- **pH:** the indicator of the acidity or basic property of the wine on a scale from 0 (very acidic) to 14 (very basic)
- **sulphates:** ( $g(potassium\ sulphate)/dm^3$ ) a wine additive which can contribute to sulfur dioxide gas  $SO_2$  levels
- **alcohol:** (vol.%) the percent alcohol content of the wine.

## Analysis

In our data set, the response variable quality is an ordinal data with a scale of 0 to 10. **Quality** is currently an integer variable in the dataset, but for modeling purposes, we might convert it into factor variable type. We will investigate 11 explanatory variables to understand their main effect and interaction effect on the wine quality (response variable).

Our **main objective** of the study is to understand which explanatory variables play a significant influence on the quality of the wine.

Our current hypothesis is that:

Proposed methods:

### Data Exploration

- Plot the relationships of each explanatory variable vs. the response variable quality and all other explanatory variables so that we can have a rough idea of variables' relationships, identify outliers, and needs for data transformation.
- Based on the findings from the plots, we will focus on strong linear/non-linear relationships/significant multicollinearity detected. Through inspection of variable definitions, we believe free sulphur dioxide could be directly related to total sulphur dioxide, which is one potential multicollinearity we will test on. A comprehensive test of strong correlation/multicollinearity will be conducted on other explanatory variables as well.
- Analyze the distribution of each variables, specifically histogram and summary statistics of mean, median, and standard deviation for numerical data types, to further inspect for missing value, outliers, and potential data transformation.
- Given our response variable (quality) is a categorical variable, we might transform it from an integer type to a factor variable. We can then create a bar graph and calculate the number of observations in each quality rating using the `group_by()` and `summarise()` functions. This analysis of the distribution of quality can then help us determine the most appropriate reference level.

### Regression Analysis

- Conduct linear regression models using explanatory variables and the response variable after adjusting for any findings in the Data Exploration stage (e.g. add a new variable for data transformation).
- We will use a stepwise model selection to choose variables for the model, including only main effects.
- We might also conduct model selection using forward and backward procedure, depending on the explanatory variables included in the previous model selection, and compare their differences
- After deciding on the model, we will conduct Nested F tests to determine any significant interaction effects among the explanatory variables. For example, density, residual sugar, density, alcohol, according to variable definitions.
- Create a final model based on previous results.
- Check for model fit (i.e.  $R^2$ value)

### Assumptions

- Check if the final linear regression model meets the assumptions for regression: linearity, constant variance, independence and normality using standardized residual plots, histogram (distribution of residuals), and QQ-plot.
- Try to improve the model if certain assumptions are not met.
- Check for potential influence points and see how they might affect the model fit. We will test for leverage and Cook's Distance and draw conclusions along with the standardized residual plots.
- Calculate multicollinearity using the `vif()` function.

## Conclusion

- After finalizing our model and making sure it meets all assumptions, we will draw the final conclusion on our research interest, and identify factors influencing the wine quality.

## Reference

<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

[https://ac.els-cdn.com/S0167923609001377/1-s2.0-S0167923609001377-main.pdf?\\_tid=949194e4-fb5f-4ecc-8572-25fb69b53955&acdnat=1541795552\\_dd074c301f5165bdb4ba634ffcd1534](https://ac.els-cdn.com/S0167923609001377/1-s2.0-S0167923609001377-main.pdf?_tid=949194e4-fb5f-4ecc-8572-25fb69b53955&acdnat=1541795552_dd074c301f5165bdb4ba634ffcd1534)

## Appendix