

**ALY6015- Intermediate Analytics - 21284**

**Final Project**

Instructor: Dr **Justin Rodgers**

**RED WINE QUALITY REGRESSION ANALYSIS**

Diana Puerta

Ashwin Nair

Report Analysis:

**Background**

The dataset consists of red variants of the Portuguese "Vinho Verde" wine.  The physicochemical (inputs) and sensory (the output) variables are available for us to assess the quality metrics of the variant of wine.

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

**Input variables (based on physicochemical tests):***fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide,*

*total sulphur dioxide, density, pH, sulphates, alcohol***Output variable (based on sensory data):**  
*quality (score between 0 and 10)*

Number of Instances: red wine – 1599

Number of Attributes: 11 + output attribute

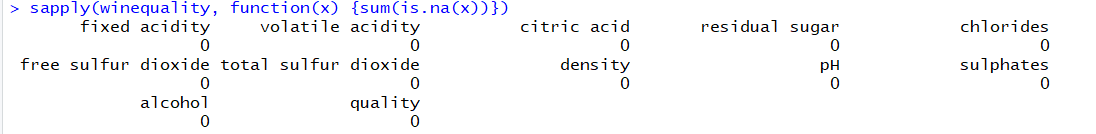
**Business Statement | Objective**

**The goal is to find a regression model of wine quality with the various physicochemical variables**

* *Three regression techniques were applied, under a computationally efficient procedure that performs simultaneous variable and model selection and that is guided by the sensitivity analysis*
* The business objective is to support the wine expert evaluations and ultimately improve the quality
* Figure out the significant physiochemical variables that affect the quality of the red wine
* Find out the correlation (positive & negative) between the quality of wine & physiochemical variables
* Obtain predictions from the fitted generalized linear model for most significant physiochemical variables

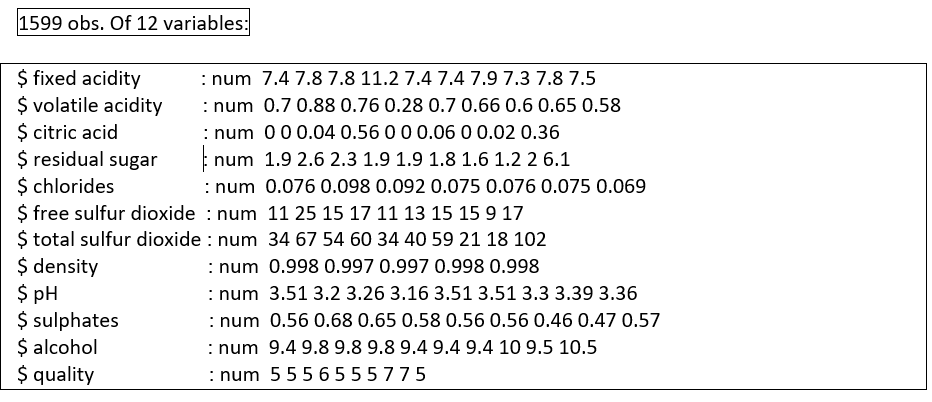
**Data Pre-Processing**

We upload the file from our local system and clean the data for null & redundant values.

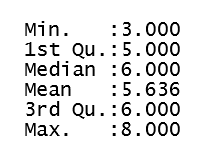


No null values were found in the dataset.

**Structure of the dataset:**



**Summarizing the quality variable:**

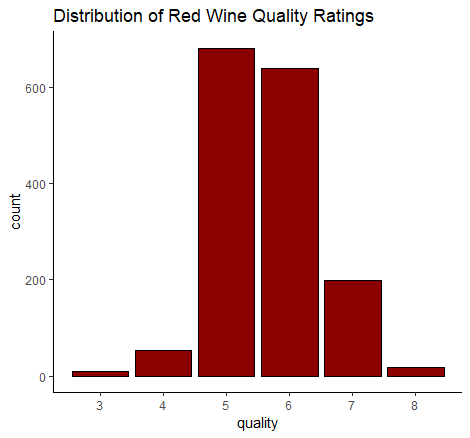


The dataset is highly unbalanced. Most of the observations are for '5' and '6' quality wines

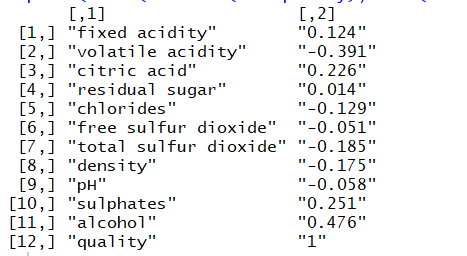
Quality - 3 4 5 6 7 8

Count - 10 53 681 638 199 18

Mean of **alcohol**, **sulphates**, **volatile acidity**, **total sulfur dioxide** and **density** looks significantly different from each other for wine quality '5' and '6' and hence they could be significant variable to determine the quality

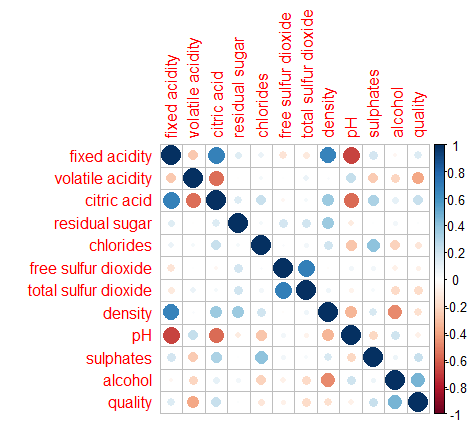


Finding Correlation with variables in the dataset:



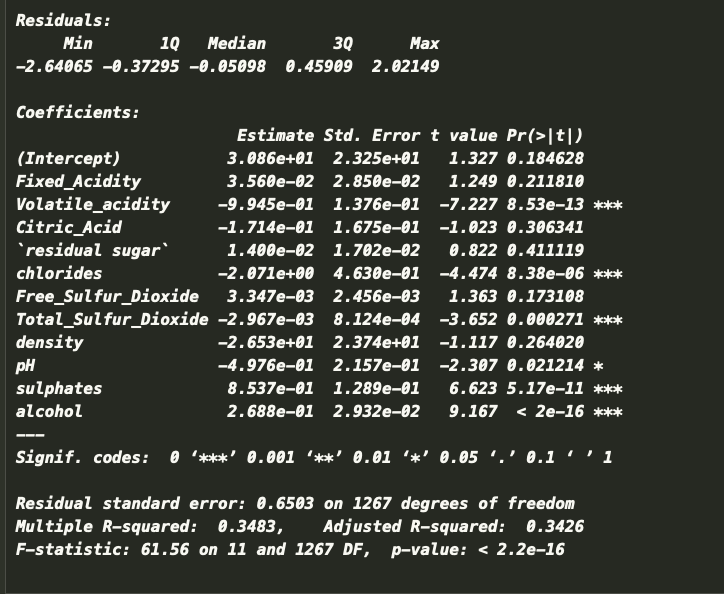
No variable is highly correlated with a quality variable. Splitting dataset into training and test dataset, we assess the frequency of each wine variant quality as shown in the table.





Acidic variables are correlated with each other and with pH and density. We find that alcohol is the most positive correlated variable and volatile acidity is the most negative correlated variable to quality.

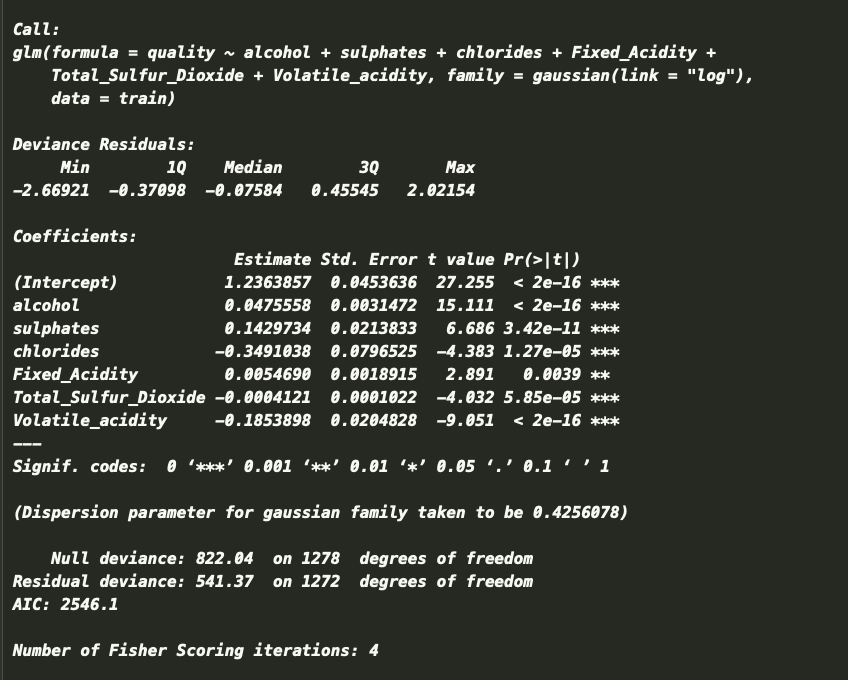
**Linear Regression:**



Linear Regression confirms that **alcohol**, **sulphates**, **volatile acidity**, and **total Sulphur dioxide** are significant variables.

Since all these P values are extremely small, we can discard pH value since it displays just one star.

**Generalized linear regression(Gaussian)**

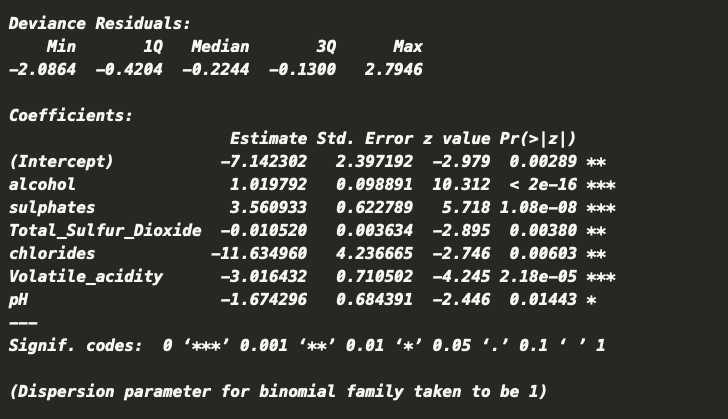


Gaussian models the dependency between a response yy and its covariates vector xx as a linear function:

The model is fitted by solving the least squares problem, which is equivalent to maximizing the likelihood for the Gaussian family.

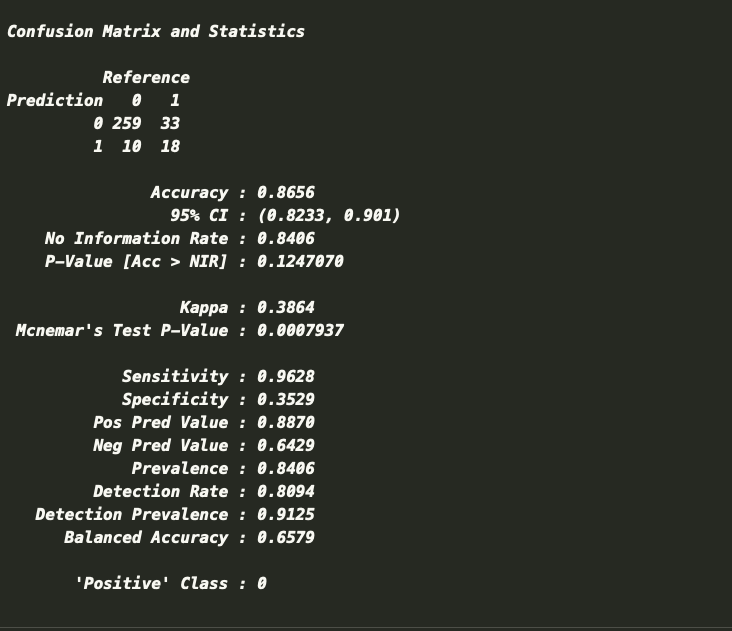
We use the ordinary least squares (OLS) to estimate the difference between the actual value and predicted value. The difference might be positive and negative.

**Binomial GLM**



The volatile acidity, alcohol and sulphates are the most significant variables for deciding if the wine is good or not. (Good wine = Quality >6)

**Prediction Accuracy**

****

With the model we used, we confirm an 86.56% prediction accuracy when considering the most significant physicochemical variables for the red wine variants.

Good wine is more than the perfect combination of different chemical components. Future improvement can be made if more data can be collected on both low-quality and high-quality wine.

**----------------------------------------------------------------------------------------------------------------**

**R Script**

# Final Project Week 6

library(readr)

library(dplyr)

library(ggplot2)

library(nnet)

library(caret)

library(MASS)

library(corrplot)

library(caTools)

library(InformationValue)

getwd()

setwd("C:/Users/Hp/Desktop/CPS Analytics/ALY 6105 Intermediate Analytics/Final Project")

winequality <- read.csv("WinequalityRed.csv")

head(winequality)

str(winequality)

summary(winequality)

sapply(winequality, function(x) {sum(is.na(x))})

mydata <- na.omit(winequality)

mydata<-scale(mydata)

mydata

table(winequality$quality)

ggplot(winequality,aes(x=quality))+geom\_bar(stat = "count",position = "dodge", color = "black", fill = "darkred")+ scale\_x\_continuous(breaks = seq(3,8,1))+ ggtitle("Distribution of Red Wine Quality Ratings")+ theme\_classic()

# finding correlation

wine\_corr <- data.frame(cor(winequality))

print(cbind(colnames(winequality),round(wine\_corr$quality, 3)))

# splitting data in training and test dataset

set.seed(1)

row.number <- sample(x=1:nrow(winequality), size=0.8\*nrow(winequality))

train = winequality[row.number,]

test = winequality[-row.number,]

head(train)

head(test)

winequality\_dist <- data.frame(table(winequality$quality))

train\_dist <- data.frame(table(train$quality))

test\_dist <- data.frame(table(test$quality))

class\_dist <- cbind(winequality\_dist, train\_dist[2], test\_dist[2])

class\_dist

# simple linear regression with all variables

quality\_lm <- lm(quality ~ ., data=train)

summary(quality\_lm)

# checking for correlation among significant variables

corrplot(cor(train))

head(train)

# generalized linear regression

quality\_glm <- glm(quality ~ alcohol + sulphates + chlorides + fixed.acidity+

total.sulfur.dioxide +volatile.acidity,

data = train, family = gaussian(link='log'))

summary(quality\_glm)

train$quality = as.factor(train$quality)

test$quality = as.factor(test$quality)

# Setting quality of wine more than 6 as 1 and remaining 0

winequality$good\_wine <- ifelse(winequality$quality > 6,1,0)

head(winequality)

View(winequality)

winequality$good\_wine <- as.factor(winequality$good\_wine)

# splitting into test & train

set.seed(1)

row.number <- sample(x=1:nrow(winequality), size=0.8\*nrow(winequality))

l\_train = winequality[row.number,]

l\_test = winequality[-row.number,]

head(l\_train)

head(l\_test)

#Binomial

quality\_bi\_glm <- glm(good\_wine ~ alcohol + sulphates + total.sulfur.dioxide

+ chlorides + volatile.acidity+ pH,

data = l\_train, family=binomial(link='logit'))

summary(quality\_bi\_glm)

View(winequality)

# constructing confusing matrix

prediction\_bi\_mult <- predict.glm(quality\_bi\_glm, newdata = l\_test, type = 'response')

predicted\_values <- ifelse(prediction\_bi\_mult > 0.5,1,0)

str(predicted\_values)

str(l\_test$good\_wine)

l\_test$good\_wine<-as.factor(l\_test$good\_wine)

predicted\_values<- as.factor(predicted\_values)

confusionMatrix(predicted\_values, l\_test$good\_wine)

----------------------------------------------------------------------------------------------------------------