Regression Lab - Team Lightning Ridge - Properties of Wine and Their Effect on Quality

w203: Statistics for Data Science

Andrew Bailey, Arun Surendranath, Luc Robitaille, Visveswaran (Vish) Sivakumar

Abstract

For our regression lab the Lightning Ridge team chose to analyze a wine quality dataset from the University of California - Irvine. The dataset consisted of 4,898 observation for white wine and 1,599 observations for red wine. Both datasets recored the same 12 attributes across both red and white wines. The purpose of our analysis was to identify which physiochemical properties of a wine increase the overall quality of each type of wine.

Research Questions

> What are the most important attributes that create a great wine? Can we build a model which takes the chemical profile of wine and explains the overall quality of the product?

```
# df_wine_red <- read.delim("winequality-red.csv", header = TRUE, sep = "\t", dec = ".")
df_wine_red <- read.delim("Data/winequality-red.csv", header = TRUE, sep = ";")
#df_wine_red

df_wine_white <- read.delim("Data/winequality-white.csv", header = TRUE, sep = ";")
#df_wine_white
```

Approach

Our first attempts to address the research questions focused on identifying relationships between attributes of a wine prepping the dataset for any types of conflicts that might undermind our regression analysis.

```
summary(df_wine_red)
```

```
##  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    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
```

```
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

```
summary(df_wine_red)
```

```
##  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    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Data Wrangling and Analysis

Running some models

```
red_wine_model1 <- lm(quality ~ residual.sugar + citric.acid + chlorides + total.sulfur.dioxide + sulpha
summary(red_wine_model1)
```

```
##
## Call:
## lm(formula = quality ~ residual.sugar + citric.acid + chlorides +
##     total.sulfur.dioxide + sulphates + alcohol, data = df_wine_red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63704 -0.36195 -0.06023  0.48907  2.00045
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.100166   0.190418  11.029  < 2e-16 ***
## residual.sugar        0.004444   0.012335   0.360    0.719
## citric.acid           0.608102   0.092583   6.568 6.88e-11 ***
## chlorides            -2.682252   0.403452  -6.648 4.06e-11 ***
```

```
## total.sulfur.dioxide -0.002859   0.000535  -5.344 1.04e-07 ***
## sulphates             1.110084   0.112217   9.892  < 2e-16 ***
## alcohol               0.287505   0.017016  16.896  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.67 on 1592 degrees of freedom
## Multiple R-squared:  0.3142, Adjusted R-squared:  0.3117
## F-statistic: 121.6 on 6 and 1592 DF,  p-value: < 2.2e-16
```

```
white_wine_model1 <- lm(quality ~ residual.sugar + citric.acid + chlorides + total.sulfur.dioxide + sulp
summary(white_wine_model1)
```

```
##
## Call:
## lm(formula = quality ~ residual.sugar + citric.acid + chlorides +
##     total.sulfur.dioxide + sulphates + alcohol, data = df_wine_red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63704 -0.36195 -0.06023  0.48907  2.00045
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.100166   0.190418  11.029  < 2e-16 ***
## residual.sugar        0.004444   0.012335   0.360    0.719
## citric.acid           0.608102   0.092583   6.568 6.88e-11 ***
## chlorides            -2.682252   0.403452  -6.648 4.06e-11 ***
## total.sulfur.dioxide -0.002859   0.000535  -5.344 1.04e-07 ***
## sulphates             1.110084   0.112217   9.892  < 2e-16 ***
## alcohol               0.287505   0.017016  16.896  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.67 on 1592 degrees of freedom
## Multiple R-squared:  0.3142, Adjusted R-squared:  0.3117
## F-statistic: 121.6 on 6 and 1592 DF,  p-value: < 2.2e-16
```