W205

Theresa Azinge

2022-12-01

Exploration

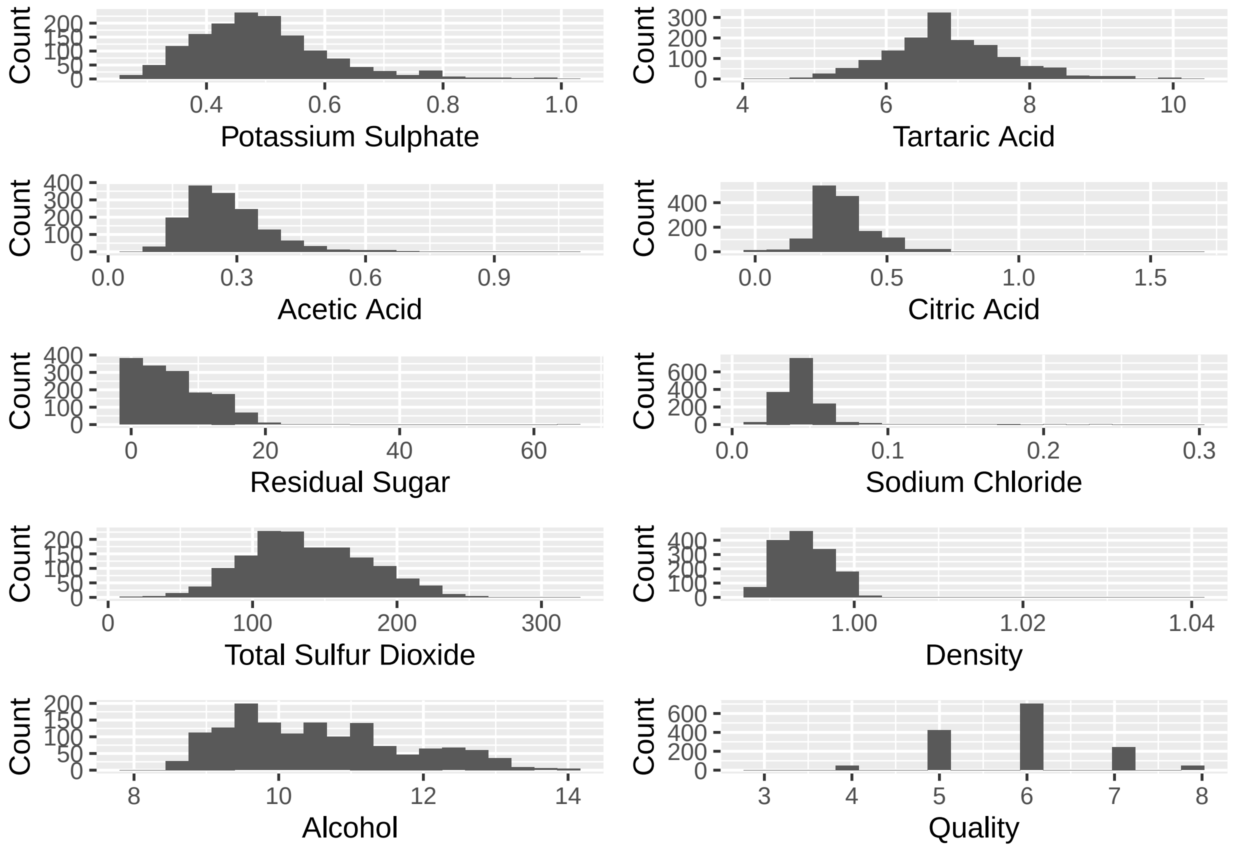
```r
setwd("~/Desktop/git_test/andamooka")
wine <- read.csv('./data/processed/wine_explore.csv', header=TRUE)
```

```r
# To assemble multiple plots
library(gridExtra)
```

```
##
## Attaching package: ’gridExtra’

## The following object is masked from ’package:dplyr’:
##
##     combine
```

```r
p1 <- ggplot(data = wine, aes(x = potassium_sulphate)) +
  geom_histogram(bins = 20) +
  labs(x = "Potassium Sulphate",y = "Count")


p2 <- ggplot(data = wine, aes(x = tartaric_acid)) +
  geom_histogram(bins = 20) +
  labs(x = "Tartaric Acid",y = "Count")

p3 <- ggplot(data = wine, aes(x = acetic_acid)) +
  geom_histogram(bins = 20) +
  labs(x = "Acetic Acid",y = "Count")

p4 <- ggplot(data = wine, aes(x = citric_acid)) +
  geom_histogram(bins = 20) +
  labs(x = "Citric Acid",y = "Count")

p5 <- ggplot(data = wine, aes(x = residual_sugar)) +
  geom_histogram(bins = 20) +
  labs(x = "Residual Sugar",y = "Count")

p6 <- ggplot(data = wine, aes(x = sodium_chloride)) +
  geom_histogram(bins = 20) +
  labs(x = "Sodium Chloride",y = "Count")
```

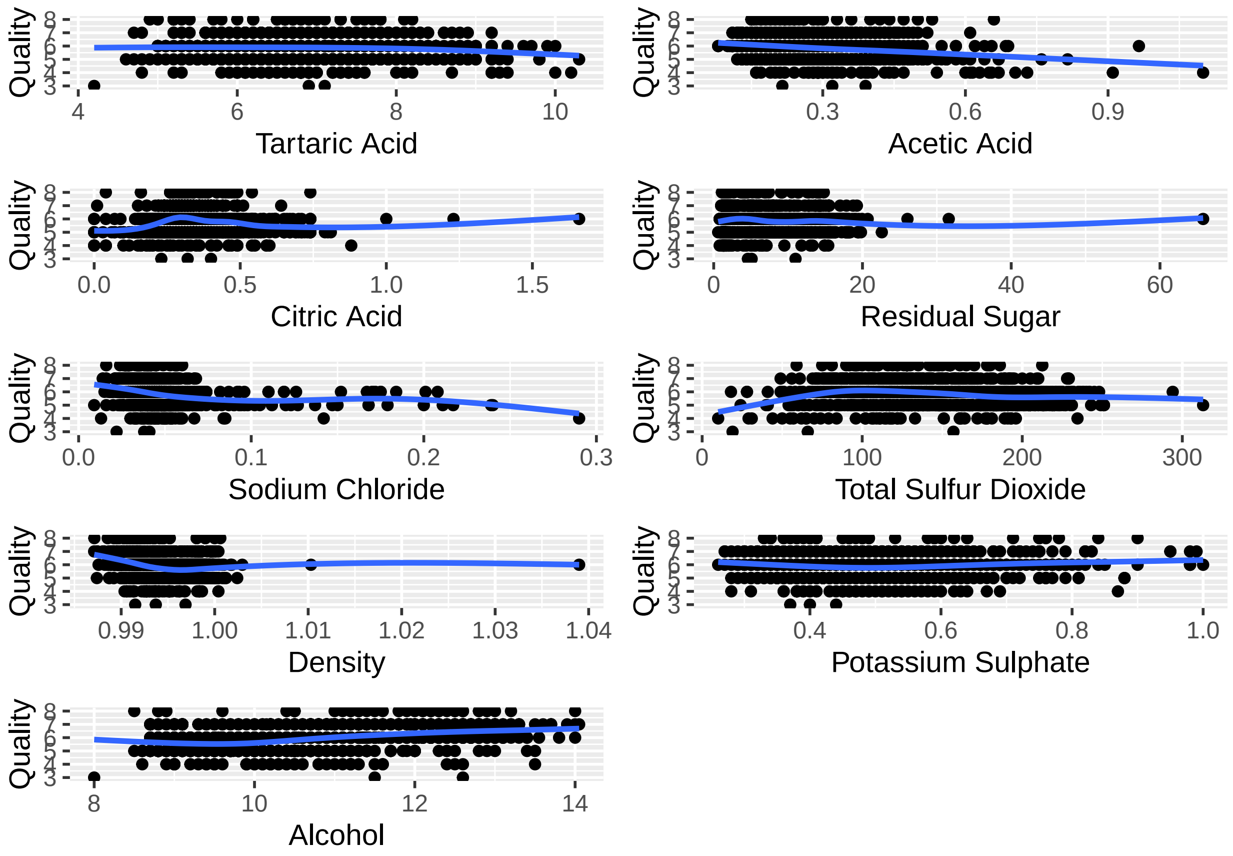
```
p7 <- ggplot(data = wine, aes(x = total_sulfur_dioxide)) +
  geom_histogram(bins = 20) +
  labs(x = "Total Sulfur Dioxide",y = "Count")

p8 <- ggplot(data = wine, aes(x = density)) +
  geom_histogram(bins = 20) +
  labs(x = "Density",y = "Count")

p9 <- ggplot(data = wine, aes(x = alcohol)) +
  geom_histogram(bins = 20) +
  labs(x = "Alcohol",y = "Count")

p10 <- ggplot(data = wine, aes(x = quality)) +
  geom_histogram(bins = 20) +
  labs(x = "Quality",y = "Count")

grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10, nrow = 5, ncol = 2)
```



```
# To assemble multiple plots
library(gridExtra)

p1 <- ggplot(data = wine, aes(x = tartaric_acid, y = quality)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = "Tartaric Acid", y = "Quality")
```

```
p2 <- ggplot(data = wine, aes(x = acetic_acid, y = quality)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = "Acetic Acid", y = "Quality")

p3 <- ggplot(data = wine, aes(x = citric_acid, y = quality)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = "Citric Acid", y = "Quality")


p4 <- ggplot(data = wine, aes(x = residual_sugar, y = quality)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = "Residual Sugar", y = "Quality")


p5 <- ggplot(data = wine, aes(x = sodium_chloride, y = quality)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = "Sodium Chloride", y = "Quality")


p6 <- ggplot(data = wine, aes(x = total_sulfur_dioxide, y = quality)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = "Total Sulfur Dioxide", y = "Quality")

p7 <- ggplot(data = wine, aes(x = density, y = quality)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = "Density", y = "Quality")

p8 <- ggplot(data = wine, aes(x = potassium_sulphate, y = quality)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = "Potassium Sulphate", y = "Quality")

p9 <- ggplot(data = wine, aes(x = alcohol, y = quality)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = "Alcohol", y = "Quality")

grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9, nrow = 5, ncol = 2)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



```
names(wine)
```

```
##  [1] "tartaric_acid"        "acetic_acid"          "citric_acid"
##  [4] "residual_sugar"       "sodium_chloride"      "total_sulfur_dioxide"
##  [7] "density"              "potassium_sulphate"   "alcohol"
## [10] "quality"
```

```
model1 <- lm(quality ~ tartaric_acid + acetic_acid + citric_acid + residual_sugar + sodium_chloride + t
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                         Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)           8.3729e+01  5.2967e+01   1.5808 0.1141503
## tartaric_acid         1.7001e-02  3.8452e-02   0.4421 0.6584615
## acetic_acid          -2.1168e+00  2.0943e-01 -10.1075 < 2.2e-16 ***
## citric_acid          -2.4928e-01  1.6110e-01  -1.5474 0.1219871
## residual_sugar        5.9312e-02  1.6557e-02   3.5823 0.0003518 ***
## sodium_chloride      -9.2812e-01  8.3884e-01  -1.1064 0.2687255
## total_sulfur_dioxide  1.1472e-03  6.3703e-04   1.8009 0.0719302 .
## density               -8.1262e+01  5.2976e+01  -1.5339 0.1252602
## potassium_sulphate     6.3540e-01  1.9046e-01   3.3361 0.0008711 ***
## alcohol                2.5306e-01  6.7834e-02   3.7306 0.0001983 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficients for acetic acid, residual sugar, potassium sulphate, and alcohol are statistically significant. So, we can conclude that there is a statistically meaningful relationship between these four variables and the quality of the white wine. This result might change if we include polynomial terms which control for possible nonlinearity between the variables. The next model will check for polynomial terms.

```
model2 <- lm(quality ~ tartaric_acid + I(tartaric_acid^2) + acetic_acid + I(acetic_acid^2) + citric_aci
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                              Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)                4.4554e+03  3.4784e+03  1.2809  0.200439
## tartaric_acid              5.3940e-01  2.8004e-01  1.9262  0.054277 .
## I(tartaric_acid^2)        -3.5262e-02  1.9606e-02 -1.7985  0.072302 .
## acetic_acid               -3.0570e+00  1.1463e+00 -2.6670  0.007739 **
## I(acetic_acid^2)           1.2637e+00  1.7089e+00  0.7395  0.459738
## citric_acid                3.1695e-01  8.3981e-01  0.3774  0.705930
## I(citric_acid^2)          -6.5630e-01  1.0351e+00 -0.6341  0.526141
## residual_sugar             8.9855e-02  2.0500e-02  4.3831 1.254e-05 ***
## I(residual_sugar^2)       -1.8353e-03  1.1508e-03 -1.5948  0.110979
## sodium_chloride           -3.7063e+00  3.6698e+00 -1.0099  0.312688
## I(sodium_chloride^2)       1.2272e+01  1.8145e+01  0.6763  0.498940
## total_sulfur_dioxide       1.4941e-02  2.8344e-03  5.2713 1.559e-07 ***
## I(total_sulfur_dioxide^2) -4.7577e-05  9.0828e-06 -5.2381 1.861e-07 ***
## density                   -8.8361e+03  6.9977e+03 -1.2627  0.206897
## I(density^2)               4.3849e+03  3.5185e+03  1.2462  0.212876
## potassium_sulphate        -1.0034e+00  9.7791e-01 -1.0261  0.305031
## I(potassium_sulphate^2)    1.4840e+00  8.6908e-01  1.7075  0.087937 .
## alcohol                   -4.6023e-01  3.6051e-01 -1.2766  0.201944
## I(alcohol^2)               2.9519e-02  1.6745e-02  1.7629  0.078128 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficients of total_sulfur_dioxide and total_sulfur_dioxide^2 are statistically significant and positive and negative respectively, which means an additional total_sulfur_dioxide is associated with an increase in quality but at a diminishing rate. Also, the non-significant coefficients have been removed.

```
model3 <- lm(quality ~ acetic_acid + residual_sugar + total_sulfur_dioxide + I(total_sulfur_dioxide^2) +
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                              Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)                1.3789e+00  3.1814e-01   4.3342 1.563e-05 ***
## acetic_acid               -2.1259e+00  1.9571e-01 -10.8621 < 2.2e-16 ***
## residual_sugar             2.7638e-02  4.3536e-03   6.3484 2.896e-10 ***
## total_sulfur_dioxide       1.4233e-02  2.7528e-03   5.1701 2.663e-07 ***
## I(total_sulfur_dioxide^2) -4.6710e-05  8.8807e-06  -5.2597 1.657e-07 ***
## potassium_sulphate         4.0925e-01  1.7447e-01   2.3457   0.01913 *
## alcohol                    3.5225e-01  1.9056e-02  18.4847 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
se.model1 = coeftest(model1, vcov = vcovHC)[ , "Std. Error"]
se.model2 = coeftest(model2, vcov = vcovHC)[ , "Std. Error"]
se.model3 = coeftest(model3, vcov = vcovHC)[ , "Std. Error"]

stargazer(model1, model2, model3, type = "text", #omit.stat = "f",
          se = list(se.model1, se.model2,se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001), title = "Table 1: The relationship between quality of wh
```

```
##
## Table 1: The relationship between quality of white wine and physical/chemical characteristics
## ====================================================================================================
##                                                       Dependent variable:
##                           ---------------------------------------------------------------------------
##                                                            quality
##                                    (1)                       (2)                      (3)
## ----------------------------------------------------------------------------------------------------
## tartaric_acid                     0.017                     0.539
##                                  (0.038)                   (0.280)
##
## I(tartaric_acid2)                                          -0.035
##                                                            (0.020)
##
## acetic_acid                     -2.117***                 -3.057**                 -2.126***
##                                  (0.209)                   (1.146)                  (0.196)
##
## I(acetic_acid2)                                             1.264
##                                                            (1.709)
##
## citric_acid                       -0.249                    0.317
##                                  (0.161)                   (0.840)
##
## I(citric_acid2)                                            -0.656
##                                                            (1.035)
##
## residual_sugar                   0.059***                 0.090***                  0.028***
##                                  (0.017)                   (0.021)                  (0.004)
##
## I(residual_sugar2)                                         -0.002
##                                                            (0.001)
##
## sodium_chloride                   -0.928                   -3.706
##                                  (0.839)                   (3.670)
##
## I(sodium_chloride2)                                        12.272
##                                                           (18.145)
```

```
##
## total_sulfur_dioxide              0.001                   0.015***                  0.014***
##                                  (0.001)                   (0.003)                  (0.003)
##
## I(total_sulfur_dioxide2)                                 -0.00005***              -0.00005***
##                                                           (0.00001)                (0.00001)
##
## density                          -81.262                 -8,836.058
##                                  (52.976)                (6,997.706)
##
## I(density2)                                               4,384.916
##                                                          (3,518.506)
##
## potassium_sulphate               0.635***                  -1.003                    0.409*
##                                  (0.190)                   (0.978)                  (0.174)
##
## I(potassium_sulphate2)                                       1.484
##                                                            (0.869)
##
## alcohol                          0.253***                  -0.460                   0.352***
##                                  (0.068)                   (0.361)                  (0.019)
##
## I(alcohol2)                                                  0.030
##                                                            (0.017)
##
## Constant                          83.729                  4,455.379                 1.379***
##                                  (52.967)                (3,478.360)                (0.318)
##
## ----------------------------------------------------------------------------------------------------
## Observations                      1,470                     1,470                    1,470
## R2                                0.249                     0.284                    0.255
## Adjusted R2                       0.244                     0.275                    0.252
## Residual Std. Error         0.736 (df = 1460)        0.721 (df = 1451)       0.732 (df = 1463)
## F Statistic              53.781*** (df = 9; 1460) 31.913*** (df = 18; 1451) 83.472*** (df = 6; 1463)
## ====================================================================================================
## Note:                                                                  *p<0.05; **p<0.01; ***p<0.001
```