Exploratory Data Analysis

Fiona Huang

2025-04-18

```r
# import datasets
wine <- read.csv("C:/Users/xinya/Downloads/Cornell Classes/STSCI 3740/final project/wine-quality-white-a
```

```r
# look at the variables in the dataset
head(wine)
```

```
##    type fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 white           7.0             0.27        0.36           20.7     0.045
## 2 white           6.3             0.30        0.34            1.6     0.049
## 3 white           8.1             0.28        0.40            6.9     0.050
## 4 white           7.2             0.23        0.32            8.5     0.058
## 5 white           7.2             0.23        0.32            8.5     0.058
## 6 white           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

```r
names(wine)
```

```
##  [1] "type"                "fixed.acidity"       "volatile.acidity"
##  [4] "citric.acid"         "residual.sugar"      "chlorides"
##  [7] "free.sulfur.dioxide" "total.sulfur.dioxide" "density"
## [10] "pH"                  "sulphates"           "alcohol"
## [13] "quality"
```

```r
# Check missing values
colSums(is.na(wine))
```

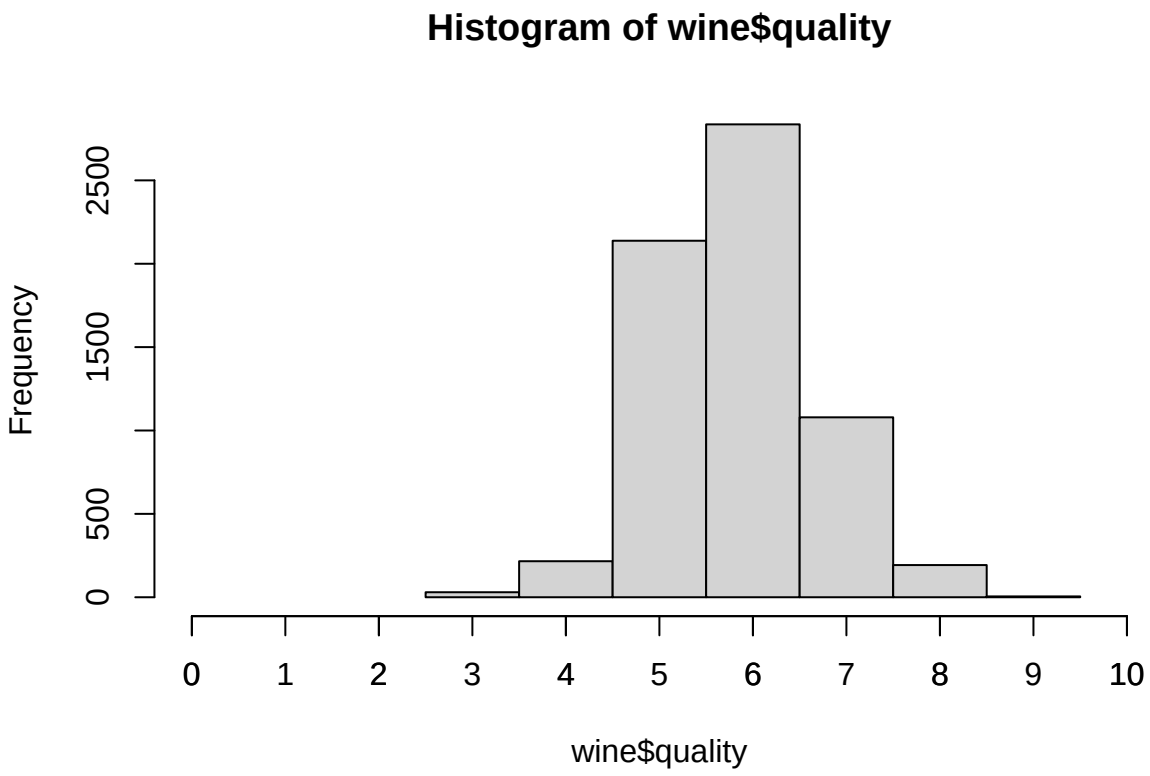
```
##                type       fixed.acidity    volatile.acidity
```

```
##                    0                    0                    0
##          citric.acid       residual.sugar            chlorides
##                    0                    0                    0
##  free.sulfur.dioxide total.sulfur.dioxide              density
##                    0                    0                    0
##                   pH            sulphates              alcohol
##                    0                    0                    0
##              quality
##                    0
```

```
# check the distirbution of each variable
summary(wine)
```

```
##      type           fixed.acidity    volatile.acidity  citric.acid
##  Length:6497        Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  Class :character   1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500
##  Mode  :character   Median : 7.000   Median :0.2900   Median :0.3100
##                     Mean   : 7.215   Mean   :0.3397   Mean   :0.3186
##                     3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900
##                     Max.   :15.900   Max.   :1.5800   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  1.00      Min.   :  6.0
##  1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0
##  Median : 3.000   Median :0.04700   Median : 29.00      Median :118.0
##  Mean   : 5.443   Mean   :0.05603   Mean   : 30.53      Mean   :115.7
##  3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0
##  Max.   :65.800   Max.   :0.61100   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50
##  Median :0.9949   Median :3.210   Median :0.5100   Median :10.30
##  Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49
##  3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30
##  Max.   :1.0390   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.818
##  3rd Qu.:6.000
##  Max.   :9.000
```

```
# check the distribution of wine quality (predicting variable)
hist(wine$quality, breaks = seq(2.5, 9.5, by = 1), xlim=c(0, 10))
axis(1, at = 0:10)
```

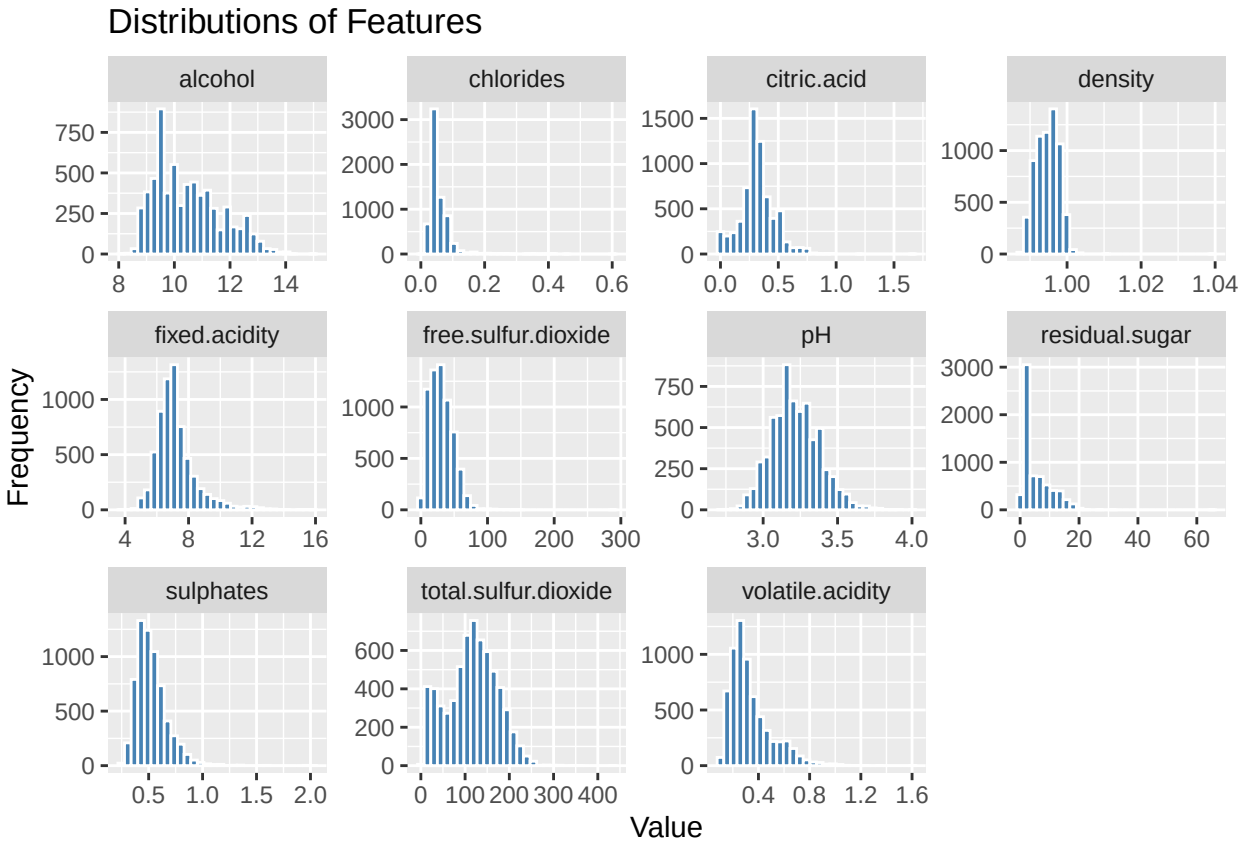


Look at the histogram of each variables

```
num_wine <- wine %>% select(-type)

wine_long <- pivot_longer(num_wine, -quality, names_to = "feature", values_to = "value")

ggplot(wine_long, aes(x = value)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  facet_wrap(~feature, scales = "free") +
  labs(title = "Distributions of Features", x = "Value", y = "Frequency")
```

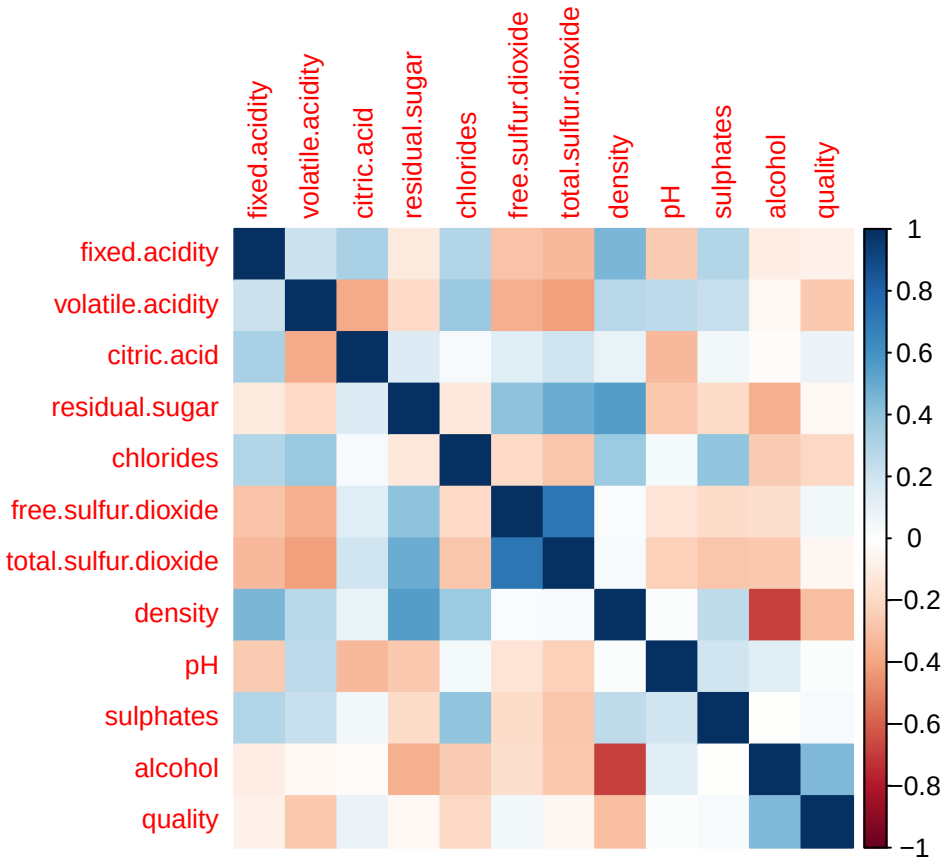



```
# compute a correlation matrix
cor_matrix <- cor(num_wine)
cor_matrix
```

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity            1.00000000       0.21900826  0.32443573    -0.11198128
## volatile.acidity         0.21900826       1.00000000 -0.37798132    -0.19601117
## citric.acid              0.32443573      -0.37798132  1.00000000     0.14245123
## residual.sugar          -0.11198128      -0.19601117  0.14245123     1.00000000
## chlorides                0.29819477       0.37712428  0.03899801    -0.12894050
## free.sulfur.dioxide     -0.28273543      -0.35255731  0.13312581     0.40287064
## total.sulfur.dioxide    -0.32905390      -0.41447619  0.19524198     0.49548159
## density                  0.45890998       0.27129565  0.09615393     0.55251695
## pH                      -0.25270047       0.26145440 -0.32980819    -0.26731984
## sulphates                0.29956774       0.22598368  0.05619730    -0.18592741
## alcohol                 -0.09545152      -0.03764039 -0.01049349    -0.35941477
## quality                 -0.07674321      -0.26569948  0.08553172    -0.03698048
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity          0.29819477         -0.28273543          -0.32905390
## volatile.acidity       0.37712428         -0.35255731          -0.41447619
## citric.acid            0.03899801          0.13312581           0.19524198
## residual.sugar        -0.12894050          0.40287064           0.49548159
## chlorides              1.00000000         -0.19504479          -0.27963045
## free.sulfur.dioxide   -0.19504479          1.00000000           0.72093408
## total.sulfur.dioxide  -0.27963045          0.72093408           1.00000000
## density                0.36261466          0.02571684           0.03239451
```

```
## pH                    0.04470798         -0.14585390          -0.23841310
## sulphates             0.39559331         -0.18845725          -0.27572682
## alcohol              -0.25691558         -0.17983843          -0.26573964
## quality              -0.20066550          0.05546306          -0.04138545
##                         density          pH    sulphates      alcohol
## fixed.acidity         0.45890998 -0.25270047  0.299567744 -0.095451523
## volatile.acidity      0.27129565  0.26145440  0.225983680 -0.037640386
## citric.acid           0.09615393 -0.32980819  0.056197300 -0.010493492
## residual.sugar        0.55251695 -0.26731984 -0.185927405 -0.359414771
## chlorides             0.36261466  0.04470798  0.395593307 -0.256915580
## free.sulfur.dioxide   0.02571684 -0.14585390 -0.188457249 -0.179838435
## total.sulfur.dioxide  0.03239451 -0.23841310 -0.275726820 -0.265739639
## density               1.00000000  0.01168608  0.259478495 -0.686745422
## pH                    0.01168608  1.00000000  0.192123407  0.121248467
## sulphates             0.25947850  0.19212341  1.000000000 -0.003029195
## alcohol              -0.68674542  0.12124847 -0.003029195  1.000000000
## quality              -0.30585791  0.01950570  0.038485446  0.444318520
##                         quality
## fixed.acidity        -0.07674321
## volatile.acidity     -0.26569948
## citric.acid           0.08553172
## residual.sugar       -0.03698048
## chlorides            -0.20066550
## free.sulfur.dioxide   0.05546306
## total.sulfur.dioxide -0.04138545
## density              -0.30585791
## pH                    0.01950570
## sulphates             0.03848545
## alcohol               0.44431852
## quality               1.00000000
```

```r
# graph a heatmap based on the correlation matrix
corrplot(cor_matrix, method="color", tl.cex=0.8)
```

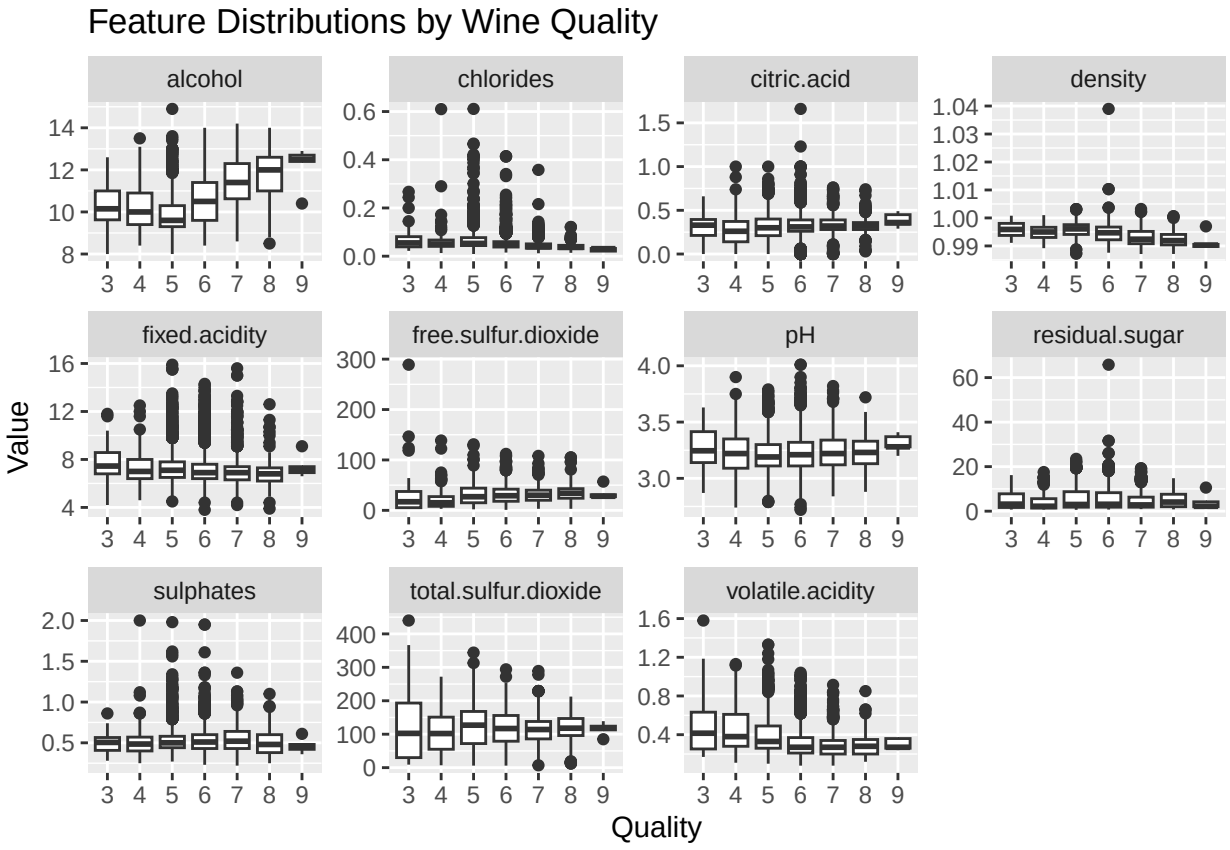


Look at the correlation with quality

```
quality_corr <- cor(num_wine)[, "quality"]
sort(quality_corr, decreasing = TRUE)
```

```
##              quality              alcohol          citric.acid
##           1.00000000           0.44431852           0.08553172
##  free.sulfur.dioxide            sulphates                   pH
##           0.05546306           0.03848545           0.01950570
##       residual.sugar total.sulfur.dioxide        fixed.acidity
##          -0.03698048          -0.04138545          -0.07674321
##            chlorides     volatile.acidity              density
##          -0.20066550          -0.26569948          -0.30585791
```

```
# plot the distribution of each variable vs wine quality
ggplot(wine_long, aes(x = factor(quality), y = value)) +
  geom_boxplot() +
  facet_wrap(~feature, scales = "free") +
  labs(title = "Feature Distributions by Wine Quality", x = "Quality", y = "Value") +
  theme(legend.position = "none")
```


Feature Distributions by Wine Quality
alcohol
chlorides
citric.acid
density
fixed.acidity
free.sulfur.dioxide
pH
residual.sugar
sulphates
total.sulfur.dioxide
volatile.acidity
Value
Quality