PSTAT 126 Project Step 1

Hanya Ansari, Carina Yuen, Daren Aguilera

2023-10-19

Name and Source of Data:

Wine Quality Based on Physicochemical Tests from UCI Machine Learning Repository *No Missing Attribute Values*

Number of Instances: red wine: 1599

Number of Variables: 12 total, 11 continuous, 1 discrete.

Description of relevant variables and their observational unit:

Continuous:

- `fixed acidity` $[\frac{\text{(grams of tartaric acid)}}{dm^3}]$:
  - acids involved with wine or fixed or nonvolatile (do not evaporate readily)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

- `volatile acidity` $[\frac{\text{grams of acetic acid}}{dm^3}]$:
  - the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

- `citric acid` $[\frac{g}{dm^3}]$:
  - found in small quantities, citric acid can add 'freshness' and flavor to wines

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

- `residual sugar` $[\frac{g}{dm^3}]$:
  - the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

- `chlorides` [$\frac{\text{grams of NaCl}}{dm^3}$]:
  - the amount of salt in the wine

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

- `free sulfur dioxide` [$\frac{mg}{dm^3}$]:
  - the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

- `total sulfur dioxide` [$\frac{mg}{dm^3}$]:
  - amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

- `density` [$\frac{g}{cm^3}$]:
  - the density of water is close to that of water depending on the percent alcohol and sugar content

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

- `pH`:
  - describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

- `sulphates` [$\frac{grams_{\text{potassium sulfate}}}{dm^3}$]:
  - a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

- `alcohol` [volume percent]:
  - the percent alcohol content of the wine

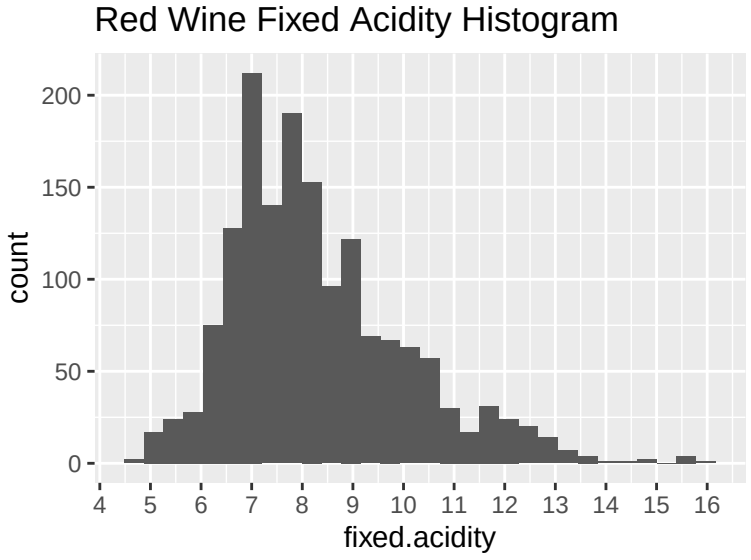
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

Discrete:

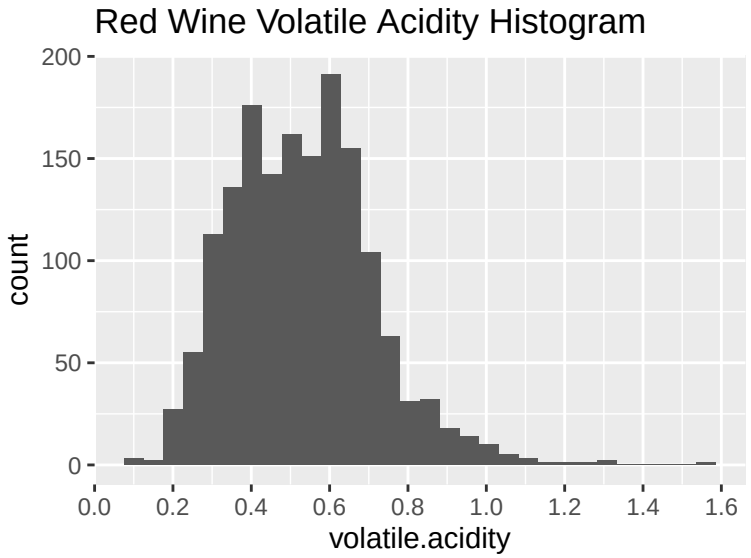
- `Quality` Quality was measured by independent scaling of certified individuals across 3 separate cultivators in Italy. Regarding the preferences, each sample was evaluated by a minimum of three sensory assessors (using blind tastes). It is acknowledged that quality rating via taste is not as clearly defined as continuous quantitative variable.
  - Observational Unit: Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

```
##  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
```

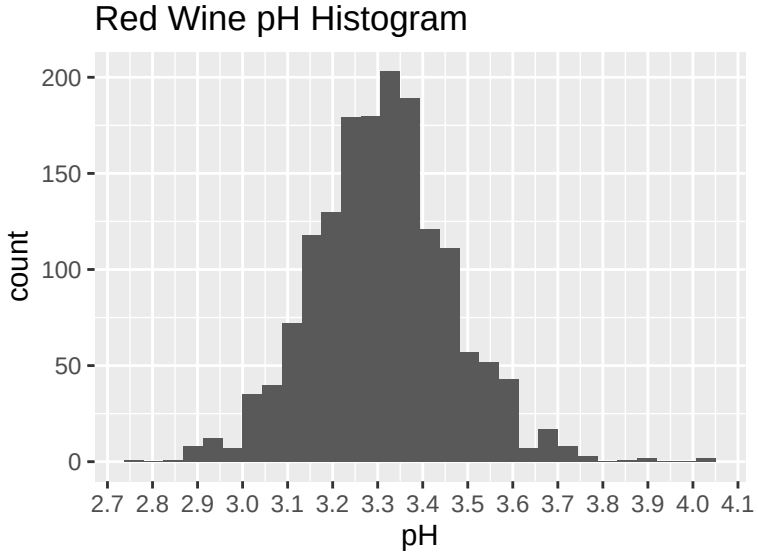
Graphical Displays and Comments



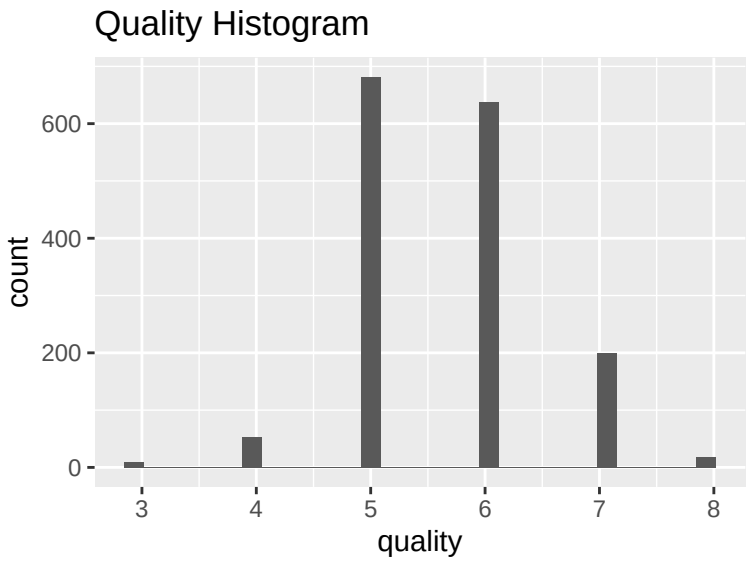
The Fixed Acidity histogram suggests that the data may have a gamma distribution, as there is a positive skew, with much of the fixed.acidity values concentrated around 7-8. Comparing the fixed acidity to the volatile acidity, the similar distribution shapes suggest that changing acidity settings does not make a large difference in wine quality in this scenario.



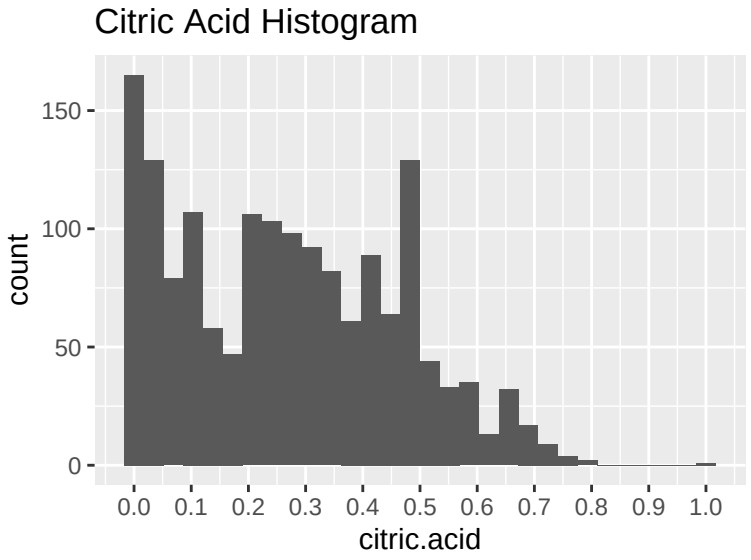
The data for volatile acidity also has a positive skew, with values concentrated around 0.4 and 0.6.



The pH data has a bell shaped and symmetric distribution, with values centered around 3.3.



The quality data is discrete.



The citric acid data roughly follows a gamma distribution with most of the values concentrated at 0 and 0.475. The presence of two peaks was unexpected.

Linear Model Fit model where y, the outcome variable, is the quality score, and each input variable (acidity, density etc.) is assigned as following.

$$fixedacidity = x_f, volatileacidity = x_v, citricacid = x_c, residualsugar = x_r, chlorides = x_{ch}, freesulfur = x_{fs}$$

$$total\ sulfur = x_{ts},\ density = x_d,\ pH = x_p,\ sulphates = x_s,\ alcohol = x_a$$

```
##                      fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity          1.00000000     -0.256130895  0.67170343    0.114776724
```

```
## volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882
## citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162
## residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000
## chlorides               0.09370519      0.061297772  0.20382291    0.055609535
## free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995
## total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882
## density                 0.66804729      0.022026232  0.36494718    0.355283371
## pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422
## sulphates               0.18300566     -0.260986685  0.31277004    0.005527121
## alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437
## quality                 0.12405165     -0.390557780  0.22637251    0.013731637
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.093705186        -0.153794193          -0.11318144
## volatile.acidity      0.061297772        -0.010503827           0.07647000
## citric.acid           0.203822914        -0.060978129           0.03553302
## residual.sugar        0.055609535         0.187048995           0.20302788
## chlorides             1.000000000         0.005562147           0.04740047
## free.sulfur.dioxide   0.005562147         1.000000000           0.66766645
## total.sulfur.dioxide  0.047400468         0.667666450           1.00000000
## density               0.200632327        -0.021945831           0.07126948
## pH                   -0.265026131         0.070377499          -0.06649456
## sulphates             0.371260481         0.051657572           0.04294684
## alcohol              -0.221140545        -0.069408354          -0.20565394
## quality              -0.128906560        -0.050656057          -0.18510029
##                          density          pH    sulphates     alcohol
## fixed.acidity         0.66804729 -0.68297819  0.183005664 -0.06166827
## volatile.acidity      0.02202623  0.23493729 -0.260986685 -0.20228803
## citric.acid           0.36494718 -0.54190414  0.312770044  0.10990325
## residual.sugar        0.35528337 -0.08565242  0.005527121  0.04207544
## chlorides             0.20063233 -0.26502613  0.371260481 -0.22114054
## free.sulfur.dioxide  -0.02194583  0.07037750  0.051657572 -0.06940835
## total.sulfur.dioxide  0.07126948 -0.06649456  0.042946836 -0.20565394
## density               1.00000000 -0.34169933  0.148506412 -0.49617977
## pH                   -0.34169933  1.00000000 -0.196647602  0.20563251
## sulphates             0.14850641 -0.19664760  1.000000000  0.09359475
## alcohol              -0.49617977  0.20563251  0.093594750  1.00000000
## quality              -0.17491923 -0.05773139  0.251397079  0.47616632
##                          quality
## fixed.acidity         0.12405165
## volatile.acidity     -0.39055778
## citric.acid           0.22637251
## residual.sugar        0.01373164
## chlorides            -0.12890656
## free.sulfur.dioxide  -0.05065606
## total.sulfur.dioxide -0.18510029
## density              -0.17491923
## pH                   -0.05773139
## sulphates             0.25139708
## alcohol               0.47616632
## quality               1.00000000
```

We did not expect a lot of variation within the data, but upon looking at the summary statistics outputted from skimr, there are certain parameters such as fixed acidity (Min: 4.6 Max: 15.9) and total sulfur dioxide (Min: 6 Max: 289.00) that had a large range compared to the other variables. On the other hand, comparing

fixed and volatile acidity, we noticed that variation in acidity levels does not make a large difference in wine quality.