Code Book

Tasheena Narraidoo, Michael Shi, Reynaldo Pena

12/04/16

Contents


```
require(mosaic)
require(mosaicData)
require(MVA)
require(aplpack)
require(scatterplot3d)
require(MASS)
require(tourr)
require(plyr)
library(caTools)
options(digits=3)
```

```
whitewine <- read.csv("winequalitywhite.csv", sep = ";", header = TRUE)
redwine <- read.csv("winequalityred.csv", sep = ";", header = TRUE)
```

```
names(whitewine)
```

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

```
names(redwine)
```

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

We have 2 datasets, one for red wine and one for white wine. They have the same variable names. The datasets have each 12 variables.

Data Codebook

variables

white wine

We have 4,898 observations for white wine.

```
nrow(whitewine)
```

```
## [1] 4898
```

```
summary(whitewine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 3.80   Min.   :0.080    Min.   :0.000   Min.   : 0.6
##  1st Qu.: 6.30   1st Qu.:0.210    1st Qu.:0.270   1st Qu.: 1.7
##  Median : 6.80   Median :0.260    Median :0.320   Median : 5.2
##  Mean   : 6.85   Mean   :0.278    Mean   :0.334   Mean   : 6.4
##  3rd Qu.: 7.30   3rd Qu.:0.320    3rd Qu.:0.390   3rd Qu.: 9.9
##  Max.   :14.20   Max.   :1.100    Max.   :1.660   Max.   :65.8
##    chlorides     free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.009   Min.   :  2.0       Min.   :  9          Min.   :0.987
##  1st Qu.:0.036   1st Qu.: 23.0       1st Qu.:108          1st Qu.:0.992
##  Median :0.043   Median : 34.0       Median :134          Median :0.994
##  Mean   :0.046   Mean   : 35.3       Mean   :138          Mean   :0.994
##  3rd Qu.:0.050   3rd Qu.: 46.0       3rd Qu.:167          3rd Qu.:0.996
##  Max.   :0.346   Max.   :289.0       Max.   :440          Max.   :1.039
##        pH          sulphates       alcohol        quality
##  Min.   :2.72   Min.   :0.22   Min.   : 8.0   Min.   :3.00
##  1st Qu.:3.09   1st Qu.:0.41   1st Qu.: 9.5   1st Qu.:5.00
##  Median :3.18   Median :0.47   Median :10.4   Median :6.00
##  Mean   :3.19   Mean   :0.49   Mean   :10.5   Mean   :5.88
##  3rd Qu.:3.28   3rd Qu.:0.55   3rd Qu.:11.4   3rd Qu.:6.00
##  Max.   :3.82   Max.   :1.08   Max.   :14.2   Max.   :9.00
```

fixed.acidity

fixed.acidity takes values from 3.80 to 14.20. Its unit of measurement is g(tartaric acid)/dm3. The mean fixed.acidity is 6.85. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$fixed.acidity)
```

```
##  min  Q1 median  Q3  max mean    sd    n missing
##  3.8 6.3    6.8 7.3 14.2 6.85 0.844 4898       0
```

volatile.acidity

volatile.acidity takes values from 0.080 to 1.1. Its unit of measurement is g(acetic acid)/dm3. The mean volatile.acidity is 0.278. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$volatile.acidity)
```

```
##   min   Q1 median   Q3 max  mean    sd    n missing
##  0.08 0.21   0.26 0.32 1.1 0.278 0.101 4898       0
```

citric.acid

citric.acid takes values from 0 to 1.66. Its unit of measurement is g/dm3. The mean citric.acid is 0.334. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$citric.acid)
```

```
##  min   Q1 median   Q3  max  mean    sd    n missing
##    0 0.27   0.32 0.39 1.66 0.334 0.121 4898       0
```

residual.sugar

residual.sugar takes values from 0.6 to 65.8. Its unit of measurement is g/dm3. The mean residual.sugar is 6.39. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$residual.sugar)
```

```
##  min  Q1 median  Q3  max mean   sd    n missing
##  0.6 1.7    5.2 9.9 65.8 6.39 5.07 4898       0
```

chlorides

chlorides takes values from 0.009 to 0.346. Its unit of measurement is g(sodium chloride)/dm3. The mean chlorides is 0.0458. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$chlorides)
```

```
##    min    Q1 median   Q3   max   mean     sd    n missing
##  0.009 0.036  0.043 0.05 0.346 0.0458 0.0218 4898       0
```

free.sulfur.dioxide

free.sulfur.dioxide takes value from 2.0 to 289.0. Its unit of measurement is mg/dm3. The mean free.sulfur.dioxide is 35.3. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$free.sulfur.dioxide)
```

```
##  min Q1 median Q3 max mean sd    n missing
##    2 23     34 46 289 35.3 17 4898       0
```

total.sulfur.dioxide

total.sulfur.dioxide takes values from 9 to 440. Its unit of measurement is mg/dm3. The mean total.sulfur.dioxide is 138. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$total.sulfur.dioxide)
```

```
##  min  Q1 median  Q3 max mean   sd    n missing
##    9 108    134 167 440  138 42.5 4898       0
```

density

density takes values from .987 to 1.039. Its unit of measurement is g/cm3. The mean density is 0.994. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$density)
```

```
##    min    Q1 median    Q3  max  mean      sd    n missing
##  0.987 0.992  0.994 0.996 1.04 0.994 0.00299 4898       0
```

pH

pH takes values form 2.72 to 3.82. The mean pH is 3.19. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$pH)
```

```
##   min   Q1 median   Q3  max mean    sd    n missing
##  2.72 3.09   3.18 3.28 3.82 3.19 0.151 4898       0
```

sulphates

sulphates take values from .22 to 1.08. Its unit of measurement is g(potassium sulphate)/dm3. The mean sulphates is 0.49. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$sulphates)
```

```
##   min   Q1 median   Q3  max mean    sd    n missing
##  0.22 0.41   0.47 0.55 1.08 0.49 0.114 4898       0
```

alcohol

alcohol takes values from 8 to 14.2. Its unit of measurement is vol.%. The mean alcohol is 10.5 %. It is a continuous input variable in assessing wine quality.

```
favstats(whitewine$alcohol)
```

```
##  min  Q1 median   Q3  max mean   sd    n missing
##    8 9.5   10.4 11.4 14.2 10.5 1.23 4898       0
```

quality

quality (which is within the [0,10] range) takes value from 3 to 9. It is the value attributed to the quality of wine, 0 being the lowest quality and 10, the highest.

```
tally(~quality, data=whitewine)
```

```
## quality
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

red wine

```
nrow(redwine)
```

```
## [1] 1599
```

```
summary(redwine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.120    Min.   :0.000   Min.   : 0.90
##  1st Qu.: 7.10   1st Qu.:0.390    1st Qu.:0.090   1st Qu.: 1.90
##  Median : 7.90   Median :0.520    Median :0.260   Median : 2.20
##  Mean   : 8.32   Mean   :0.528    Mean   :0.271   Mean   : 2.54
##  3rd Qu.: 9.20   3rd Qu.:0.640    3rd Qu.:0.420   3rd Qu.: 2.60
##  Max.   :15.90   Max.   :1.580    Max.   :1.000   Max.   :15.50
##    chlorides     free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.012   Min.   : 1.0        Min.   :  6.0        Min.   :0.990
##  1st Qu.:0.070   1st Qu.: 7.0        1st Qu.: 22.0        1st Qu.:0.996
##  Median :0.079   Median :14.0        Median : 38.0        Median :0.997
##  Mean   :0.087   Mean   :15.9        Mean   : 46.5        Mean   :0.997
##  3rd Qu.:0.090   3rd Qu.:21.0        3rd Qu.: 62.0        3rd Qu.:0.998
##  Max.   :0.611   Max.   :72.0        Max.   :289.0        Max.   :1.004
##        pH          sulphates        alcohol        quality
##  Min.   :2.74   Min.   :0.330   Min.   : 8.4   Min.   :3.00
##  1st Qu.:3.21   1st Qu.:0.550   1st Qu.: 9.5   1st Qu.:5.00
##  Median :3.31   Median :0.620   Median :10.2   Median :6.00
##  Mean   :3.31   Mean   :0.658   Mean   :10.4   Mean   :5.64
##  3rd Qu.:3.40   3rd Qu.:0.730   3rd Qu.:11.1   3rd Qu.:6.00
##  Max.   :4.01   Max.   :2.000   Max.   :14.9   Max.   :8.00
```

fixed.acidity

fixed.acidity takes values from 4.6 to 15.9. Its unit of measurement is g(tartaric acid)/dm3. The mean fixed.acidity is 8.32. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$fixed.acidity)
```

```
##  min  Q1 median  Q3  max mean   sd    n missing
##  4.6 7.1    7.9 9.2 15.9 8.32 1.74 1599       0
```

volatile.acidity

volatile.acidity takes values from 0.12 to 1.58. Its unit of measurement is g(acetic acid)/dm3. The mean volatile.acidity is 0.528. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$volatile.acidity)
```

```
##   min   Q1 median   Q3  max  mean    sd    n missing
##  0.12 0.39   0.52 0.64 1.58 0.528 0.179 1599       0
```

citric.acid

citric.acid takes values from 0.09 to 1. Its unit of measurement is g/dm3. The mean citric.acid is 0.271. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$citric.acid)
```

```
##  min   Q1 median   Q3 max  mean    sd    n missing
##    0 0.09   0.26 0.42   1 0.271 0.195 1599       0
```

residual.sugar

residual.sugar takes values from 0.9 to 15.5. Its unit of measurement is g/dm3. The mean residual.sugar is 2.54. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$residual.sugar)
```

```
##  min  Q1 median  Q3  max mean   sd    n missing
##  0.9 1.9    2.2 2.6 15.5 2.54 1.41 1599       0
```

chlorides

chlorides takes values from 0.012 to 0.611. Its unit of measurement is g(sodium chloride)/dm3. The mean chlorides is 0.0875. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$chlorides)
```

```
##    min   Q1 median   Q3   max   mean     sd    n missing
##  0.012 0.07  0.079 0.09 0.611 0.0875 0.0471 1599       0
```

free.sulfur.dioxide

free.sulfur.dioxide takes value from 1 to 72. Its unit of measurement is mg/dm3. The mean free.sulfur.dioxide is 15.9. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$free.sulfur.dioxide)
```

```
##  min Q1 median Q3 max mean   sd    n missing
##    1  7     14 21  72 15.9 10.5 1599       0
```

total.sulfur.dioxide

total.sulfur.dioxide takes values from 6 to 289. Its unit of measurement is mg/dm3. The mean total.sulfur.dioxide is 46.5. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$total.sulfur.dioxide)
```

```
##  min Q1 median Q3 max mean   sd    n missing
##    6 22     38 62 289 46.5 32.9 1599       0
```

density

density takes values from 0.99 to 1. Its unit of measurement is g/cm3. The mean density is 0.997. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$density)
```

```
##   min    Q1 median    Q3 max  mean      sd    n missing
##  0.99 0.996  0.997 0.998   1 0.997 0.00189 1599       0
```

pH

pH takes values form 2.74 to 4.01. The mean pH is 3.31. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$pH)
```

```
##   min   Q1 median  Q3  max mean    sd    n missing
##  2.74 3.21   3.31 3.4 4.01 3.31 0.154 1599       0
```

sulphates

sulphates take values from 0.33 to 2. Its unit of measurement is g(potassium sulphate)/dm3. The mean sulphates is 0.658. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$sulphates)
```

```
##   min   Q1 median   Q3 max  mean   sd    n missing
##  0.33 0.55   0.62 0.73   2 0.658 0.17 1599       0
```

alcohol

alcohol takes values from 8.4 to 14.9. Its unit of measurement is vol.%. The mean alcohol is 10.4 %. It is a continuous input variable in assessing wine quality.

```
favstats(redwine$alcohol)
```

```
##  min  Q1 median   Q3  max mean   sd    n missing
##  8.4 9.5   10.2 11.1 14.9 10.4 1.07 1599       0
```

quality

quality (which is within the [0,10] range) takes value from 3 to 8. It is the value attributed to the quality of wine, 0 being the lowest quality and 10, the highest.

```
tally(~quality, data=redwine)
```

```
## quality
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

combined dataset

```
library(plyr)
nrow(redwine) #1599
```

```
## [1] 1599
```

```
nrow(whitewine) # 4898
```

```
## [1] 4898
```

```
redwine[,"type"]  <- c("red")
whitewine[,"type"]  <- c("white")

wine <- join(redwine, whitewine,type = "full")
```

```
## Joining by: fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dio
```

We have a total of 6497 observations for the combined dataset.

```
nrow(whitewine)
```

```
## [1] 4898
```

```
summary(whitewine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 3.80   Min.   :0.080    Min.   :0.000   Min.   : 0.6
##  1st Qu.: 6.30   1st Qu.:0.210    1st Qu.:0.270   1st Qu.: 1.7
##  Median : 6.80   Median :0.260    Median :0.320   Median : 5.2
##  Mean   : 6.85   Mean   :0.278    Mean   :0.334   Mean   : 6.4
##  3rd Qu.: 7.30   3rd Qu.:0.320    3rd Qu.:0.390   3rd Qu.: 9.9
##  Max.   :14.20   Max.   :1.100    Max.   :1.660   Max.   :65.8
##    chlorides     free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.009   Min.   :  2.0       Min.   :  9          Min.   :0.987
##  1st Qu.:0.036   1st Qu.: 23.0       1st Qu.:108          1st Qu.:0.992
##  Median :0.043   Median : 34.0       Median :134          Median :0.994
##  Mean   :0.046   Mean   : 35.3       Mean   :138          Mean   :0.994
##  3rd Qu.:0.050   3rd Qu.: 46.0       3rd Qu.:167          3rd Qu.:0.996
##  Max.   :0.346   Max.   :289.0       Max.   :440          Max.   :1.039
##        pH         sulphates       alcohol        quality
##  Min.   :2.72   Min.   :0.22   Min.   : 8.0   Min.   :3.00
```

```
##  1st Qu.:3.09   1st Qu.:0.41   1st Qu.: 9.5   1st Qu.:5.00
##  Median :3.18   Median :0.47   Median :10.4   Median :6.00
##  Mean   :3.19   Mean   :0.49   Mean   :10.5   Mean   :5.88
##  3rd Qu.:3.28   3rd Qu.:0.55   3rd Qu.:11.4   3rd Qu.:6.00
##  Max.   :3.82   Max.   :1.08   Max.   :14.2   Max.   :9.00
##      type
##  Length:4898
##  Class :character
##  Mode  :character
##
##
##
```

fixed.acidity

fixed.acidity takes values from 3.80 to 15.9. Its unit of measurement is g(tartaric acid)/dm3. The mean fixed.acidity is 7.22. It is a continuous input variable in assessing wine quality.

```
favstats(wine$fixed.acidity)
```

```
##  min  Q1 median  Q3  max mean  sd    n missing
##  3.8 6.4      7 7.7 15.9 7.22 1.3 6497       0
```

volatile.acidity

volatile.acidity takes values from 0.080 to 1.58. Its unit of measurement is g(acetic acid)/dm3. The mean volatile.acidity is 0.34. It is a continuous input variable in assessing wine quality.

```
favstats(wine$volatile.acidity)
```

```
##   min   Q1 median  Q3  max mean    sd    n missing
##  0.08 0.23   0.29 0.4 1.58 0.34 0.165 6497       0
```

citric.acid

citric.acid takes values from 0 to 1.66. Its unit of measurement is g/dm3. The mean citric.acid is 0.319. It is a continuous input variable in assessing wine quality.

```
favstats(wine$citric.acid)
```

```
##  min   Q1 median   Q3  max  mean    sd    n missing
##    0 0.25   0.31 0.39 1.66 0.319 0.145 6497       0
```

residual.sugar

residual.sugar takes values from 0.6 to 65.8. Its unit of measurement is g/dm3. The mean residual.sugar is 5.44. It is a continuous input variable in assessing wine quality.

```
favstats(wine$residual.sugar)
```

```
##  min  Q1 median  Q3  max mean   sd    n missing
##  0.6 1.8      3 8.1 65.8 5.44 4.76 6497       0
```

chlorides

chlorides takes values from 0.009 to 0.611. Its unit of measurement is g(sodium chloride)/dm3. The mean chlorides is 0.056. It is a continuous input variable in assessing wine quality.

```
favstats(wine$chlorides)
```

```
##    min    Q1 median    Q3   max  mean    sd    n missing
##  0.009 0.038  0.047 0.065 0.611 0.056 0.035 6497       0
```

free.sulfur.dioxide

free.sulfur.dioxide takes value from 1 to 289.0. Its unit of measurement is mg/dm3. The mean free.sulfur.dioxide is 30.5. It is a continuous input variable in assessing wine quality.

```
favstats(wine$free.sulfur.dioxide)
```

```
##  min Q1 median Q3 max mean   sd    n missing
##    1 17     29 41 289 30.5 17.7 6497       0
```

total.sulfur.dioxide

total.sulfur.dioxide takes values from 6 to 440. Its unit of measurement is mg/dm3. The mean total.sulfur.dioxide is 116. It is a continuous input variable in assessing wine quality.

```
favstats(wine$total.sulfur.dioxide)
```

```
##  min Q1 median  Q3 max mean   sd    n missing
##    6 77    118 156 440  116 56.5 6497       0
```

density

density takes values from .987 to 1.04. Its unit of measurement is g/cm3. The mean density is 0.995. It is a continuous input variable in assessing wine quality.

```
favstats(wine$density)
```

```
##    min    Q1 median    Q3  max  mean    sd    n missing
##  0.987 0.992  0.995 0.997 1.04 0.995 0.003 6497       0
```

pH

pH takes values form 2.72 to 4.01. The mean pH is 3.22. It is a continuous input variable in assessing wine quality.

```
favstats(wine$pH)
```

```
##   min   Q1 median   Q3  max mean    sd    n missing
##  2.72 3.11   3.21 3.32 4.01 3.22 0.161 6497       0
```

sulphates

sulphates take values from .22 to 2. Its unit of measurement is g(potassium sulphate)/dm3. The mean sulphates is 0.531. It is a continuous input variable in assessing wine quality.

```
favstats(wine$sulphates)
```

```
##   min   Q1 median  Q3 max  mean    sd    n missing
##  0.22 0.43   0.51 0.6   2 0.531 0.149 6497       0
```

alcohol

alcohol takes values from 8 to 14.9. Its unit of measurement is vol.%. The mean alcohol is 10.5 %. It is a continuous input variable in assessing wine quality.

```
favstats(wine$alcohol)
```

```
##  min  Q1 median   Q3  max mean   sd    n missing
##    8 9.5   10.3 11.3 14.9 10.5 1.19 6497       0
```

quality

quality (which is within the [0,10] range) takes value from 3 to 9. It is the value attributed to the quality of wine, 0 being the lowest quality and 10, the highest.

```
tally(~quality, data=wine)
```

```
## quality
##    3    4    5    6    7    8    9
##   30  216 2138 2836 1079  193    5
```