Wine_Preference_Prediction

Bhawneet Singh

11/2/2021

Importing all libraries

```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

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

```
library(nnet)
library(AER)
```

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.3     v purrr   0.3.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x gridExtra::combine() masks dplyr::combine()
## x dplyr::filter()      masks stats::filter()
## x dplyr::lag()         masks stats::lag()
## x car::recode()        masks dplyr::recode()
## x purrr::some()        masks car::some()
```

```
library(varhandle) # for unfactoring
library(reshape2) # for dcast
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

Important aspect: What chemical charateristics are most important in predicting the quality of a wine?

Reading both datasets

```
df_red <- read.table(file = 'C:/Users/bhawn/Downloads/red_wine_quality.csv', sep=";", header=T)


df_white <- read.table(file = 'C:/Users/bhawn/Downloads/winequality-white.csv', sep=";", header=T)
```

Merging both datasets(combining red wine and white wine into wine dataset)

```
color_red <- rep('red', 1599)
color_white <- rep('white', 4898)

df_red$color_red <- color_red
df_white$color_white <- color_white

## Now, we will combine white and red wine dataset into wine dataset that
## consists of 13 variables, with 6495 observations.

df_wine <- merge(df_red, df_white, all=TRUE)

df_wine$color_red <- ifelse(df_wine$color_red=='red', 1, 0)
df_wine$color_white <- ifelse(df_wine$color_white=='white', 1, 0)
df_wine[is.na(df_wine)] <- 0


df_wine$color_red <- factor(df_wine$color_red)
df_wine$color_white <- factor(df_wine$color_white)


df_wine$colors <- ifelse(df_wine$color_red=='1', 'red', 'white')


df_wine$colors <- factor(df_wine$colors)
str(df_wine)
```

```
## 'data.frame':    6495 obs. of  15 variables:
##  $ fixed.acidity       : num  3.8 3.9 4.2 4.2 4.4 4.4 4.4 4.5 4.6 4.6 ...
##  $ volatile.acidity    : num  0.31 0.225 0.17 0.215 0.32 0.46 0.54 0.19 0.445 0.52 ...
##  $ citric.acid         : num  0.02 0.4 0.36 0.23 0.39 0.1 0.09 0.21 0 0.15 ...
##  $ residual.sugar      : num  11.1 4.2 1.8 5.1 4.3 2.8 5.1 0.95 1.4 2.1 ...
##  $ chlorides           : num  0.036 0.03 0.029 0.041 0.03 0.024 0.038 0.033 0.053 0.054 ...
##  $ free.sulfur.dioxide : num  20 29 93 64 31 31 52 89 11 8 ...
##  $ total.sulfur.dioxide: num  114 118 161 157 127 111 97 159 178 65 ...
##  $ density             : num  0.992 0.989 0.99 0.997 0.989 ...
##  $ pH                  : num  3.75 3.57 3.65 3.42 3.46 3.48 3.41 3.34 3.79 3.9 ...
##  $ sulphates           : num  0.44 0.36 0.89 0.44 0.36 0.34 0.4 0.42 0.55 0.56 ...
##  $ alcohol             : num  12.4 12.8 12 8 12.8 13.1 12.2 8 10.2 13.1 ...
```

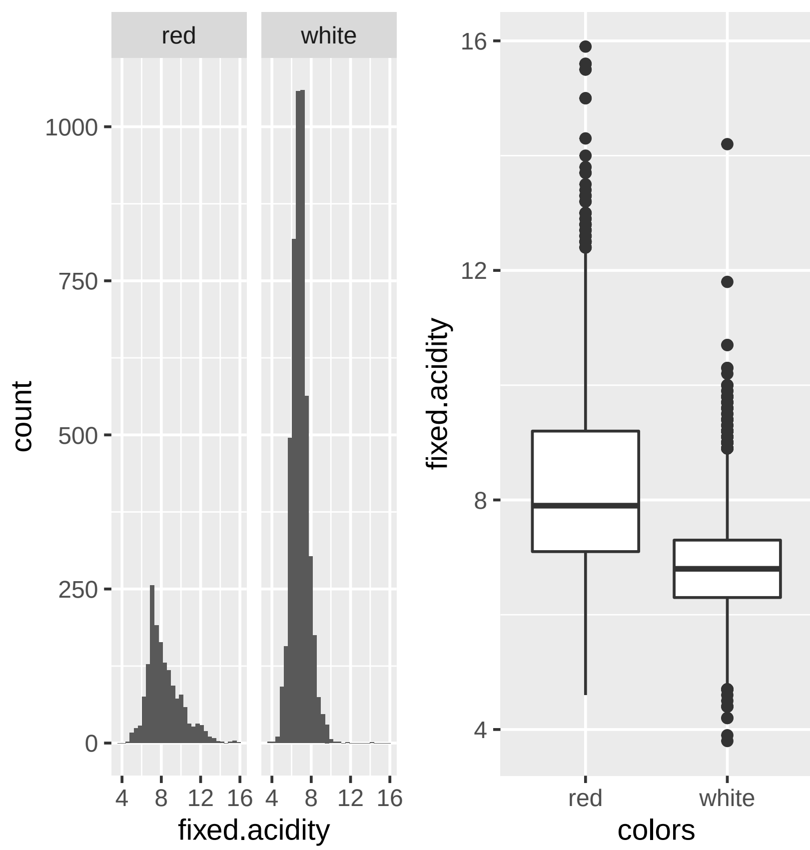
```
##  $ quality              : int  6 8 7 3 8 6 7 5 5 4 ...
##  $ color_red            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ color_white          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 1 ...
##  $ colors               : Factor w/ 2 levels "red","white": 2 2 2 2 2 2 2 2 2 1 ...
```

Plot every characterstic.

fixed.acidity

```
p2 <- ggplot(aes(x=fixed.acidity), data=df_wine) + geom_histogram() +
  facet_wrap(~colors)
p3 <- ggplot(aes(x=colors, y=fixed.acidity), data=df_wine) + geom_boxplot()
grid.arrange(p2, p3, ncol=3)
```

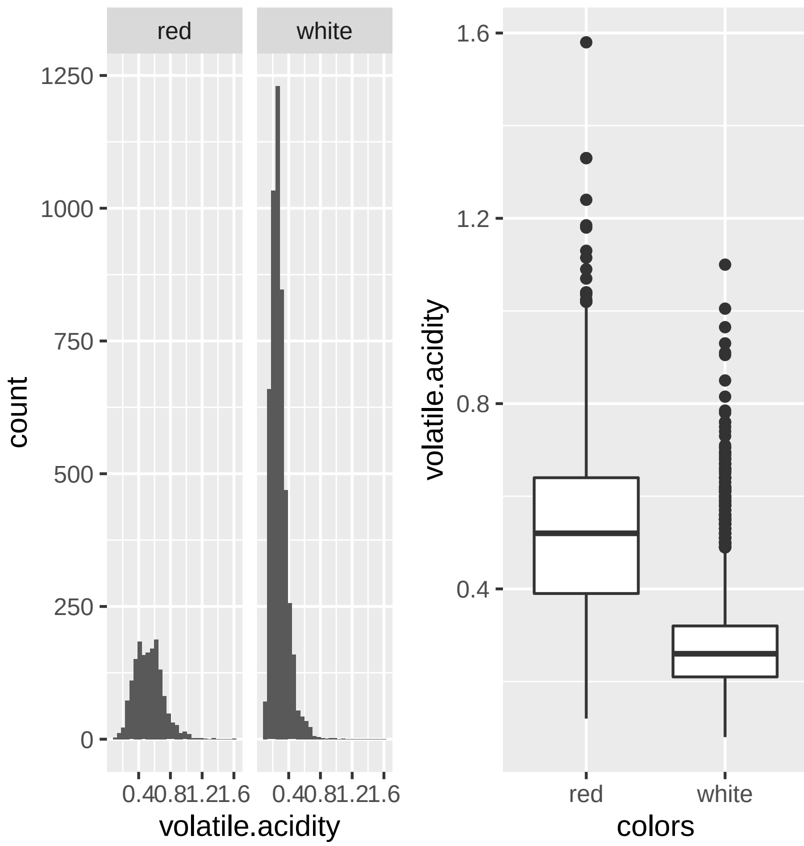
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



volatile.acidity

```r
p2 <- ggplot(aes(x=volatile.acidity), data=df_wine) + geom_histogram() +
  facet_wrap(~colors)
p3 <- ggplot(aes(x=colors, y=volatile.acidity), data=df_wine) + geom_boxplot()
grid.arrange(p2, p3, ncol=3)
```

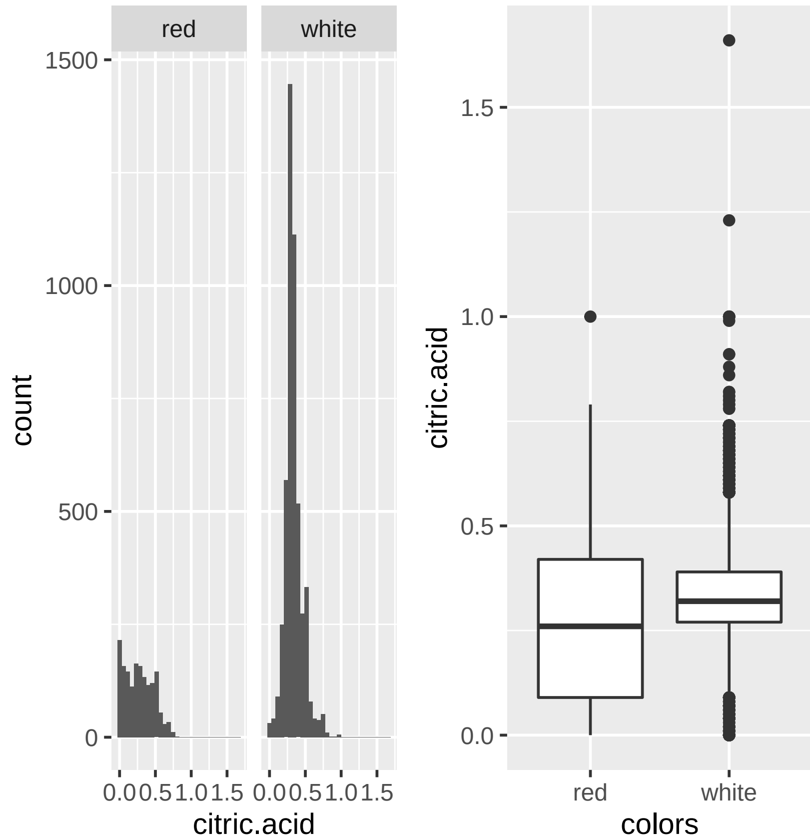
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



citric.acid

```r
p2 <- ggplot(aes(x=citric.acid), data=df_wine) + geom_histogram() +
  facet_wrap(~colors)
p3 <- ggplot(aes(x=colors, y=citric.acid), data=df_wine) + geom_boxplot()
grid.arrange(p2, p3, ncol=3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
by(df_wine$citric.acid, df_wine$colors, summary)
```

```
## df_wine$colors: red
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
## --------------------------------------------------------------
## df_wine$colors: white
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

quality

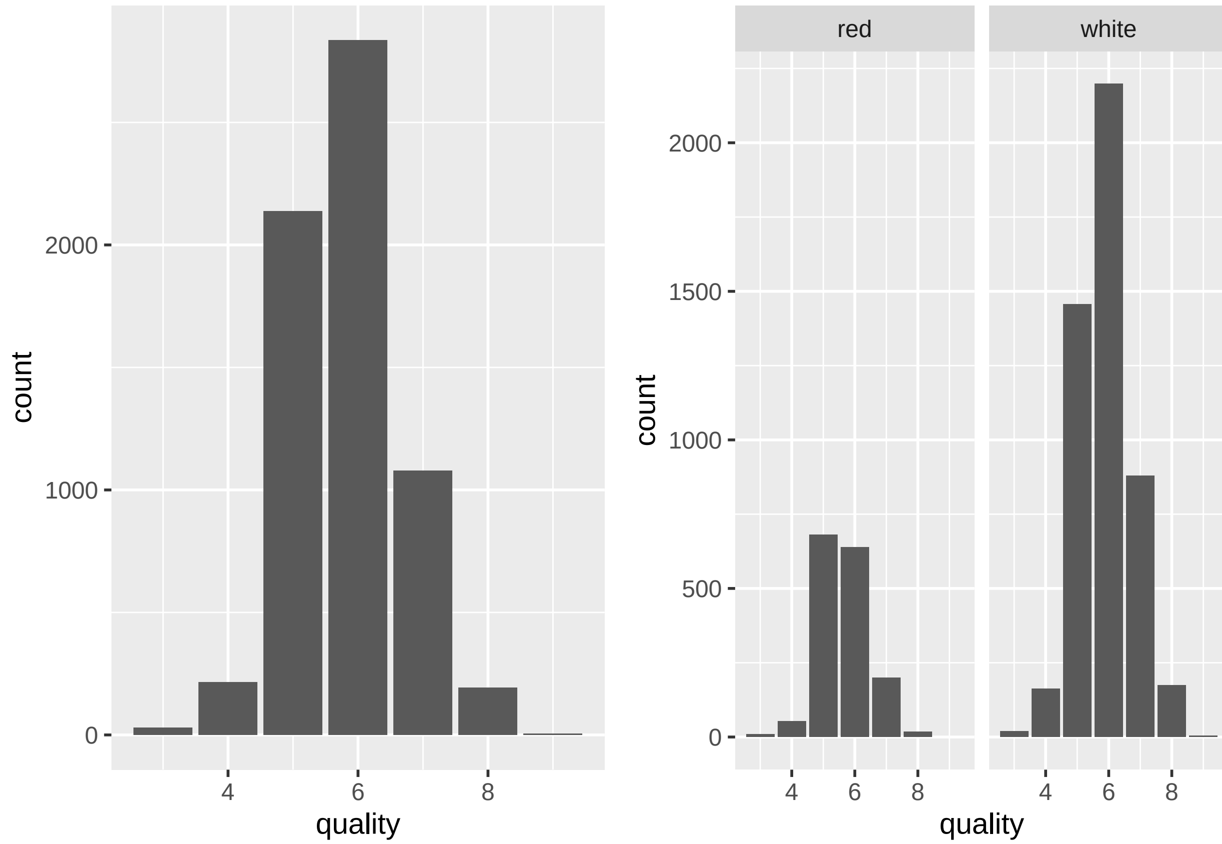
```
p1 <- ggplot(aes(x=quality), data=df_wine) + geom_histogram(stat = 'count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
p2 <- ggplot(aes(x=quality), data=df_wine) + geom_histogram(stat = 'count') +
  facet_wrap(~colors)
```

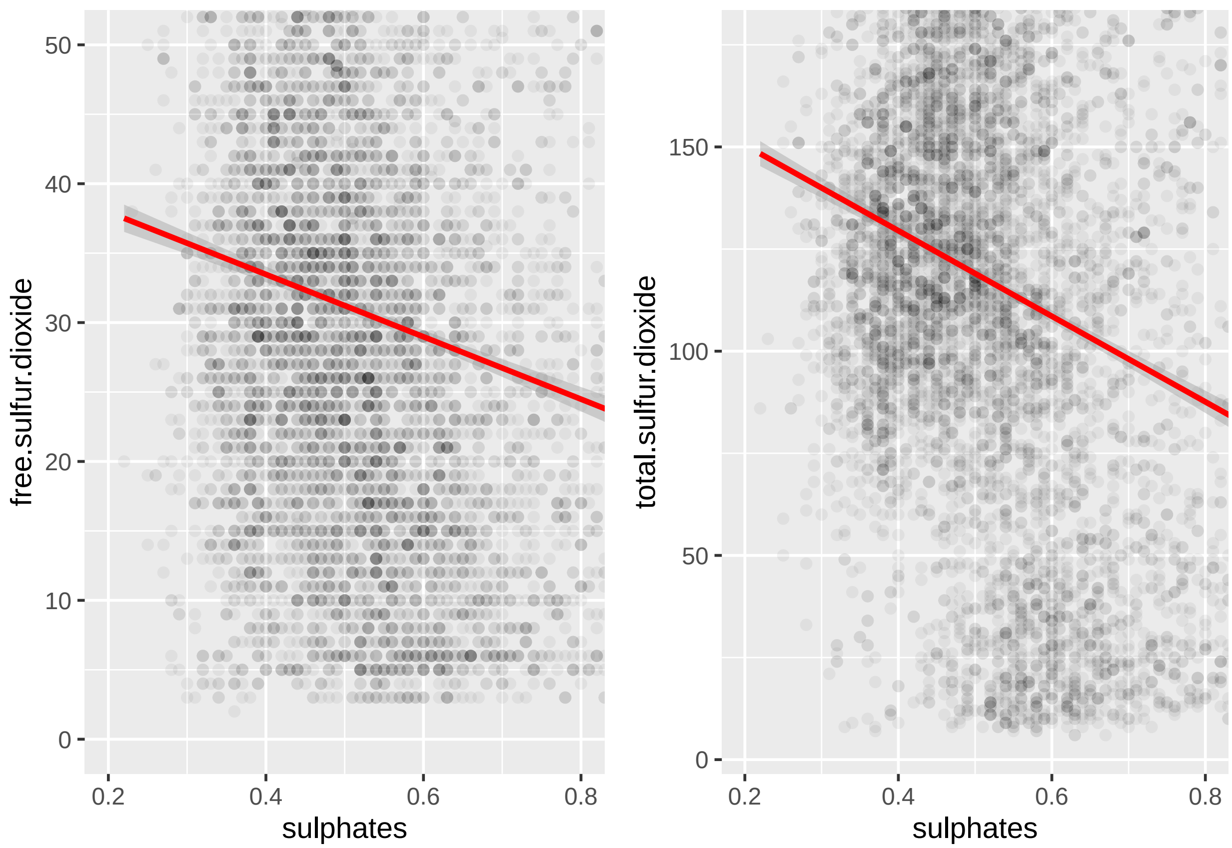
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```r
grid.arrange(p1, p2, ncol=2)
```




```r
p1<-ggplot(aes(x=sulphates, y=free.sulfur.dioxide), data=df_wine) +
  geom_point(alpha=1/20) +
  coord_cartesian(xlim= c(0.2,0.8), ylim = c(0,50)) +
  geom_smooth(method = 'lm', color='red')
p2<-ggplot(aes(x=sulphates, y=total.sulfur.dioxide), data=df_wine) +
  geom_point(alpha=1/20) +
  coord_cartesian(xlim= c(0.2,0.8), ylim = c(5,175)) +
  geom_smooth(method = 'lm', color='red')
grid.arrange(p1, p2, ncol=2)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

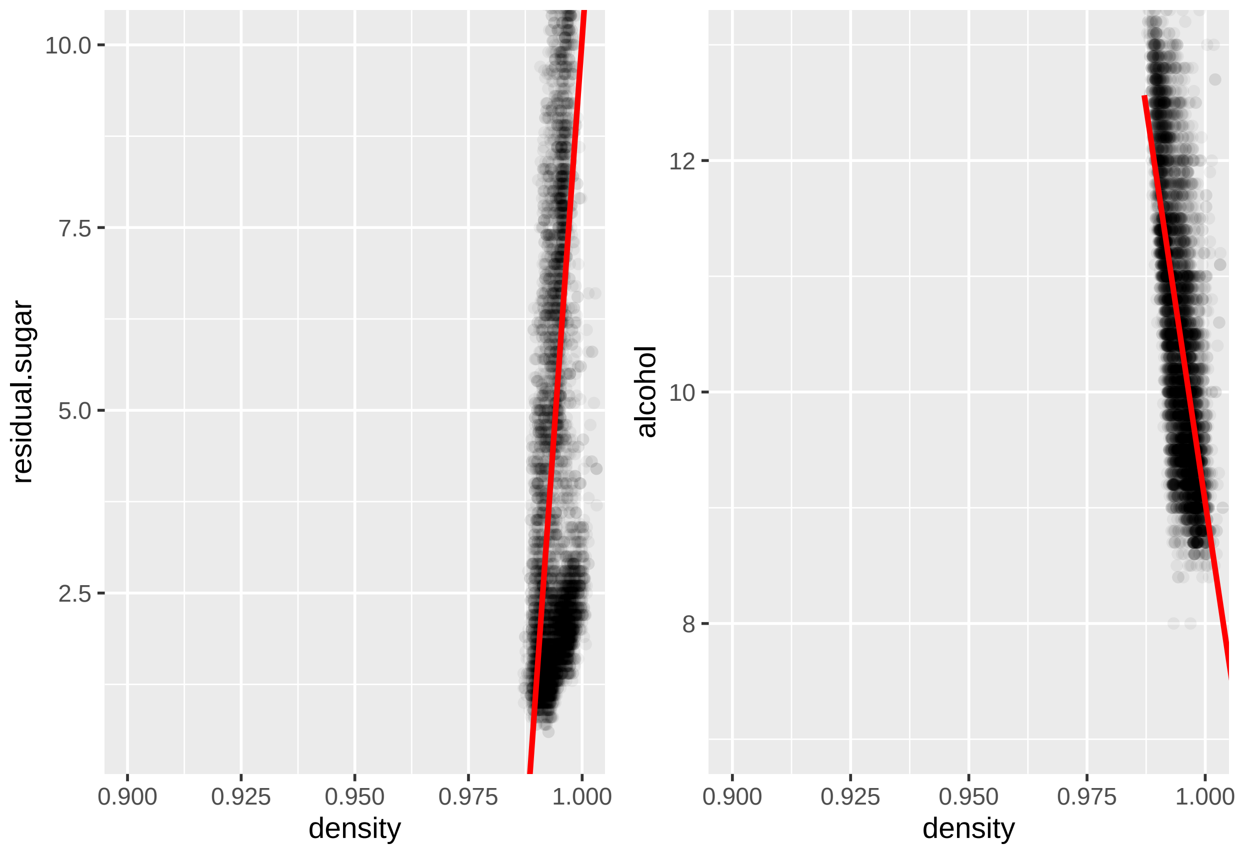



```
# Analysis: 'Free.sulfur.dioxide' and 'Total.sulfur.dioxide' behave in a similar way,
# because 'free.sulfur.dioxide' is a subset of 'total.sulfur.dioxide'.
```

We can see that 'density' is directly proportional to 'residual.sugar', while being inversely proportional to 'alcohol'

```
ds_plot<-ggplot(aes(x=density, y=residual.sugar), data=df_wine) +
  geom_point(alpha=1/20) +
  coord_cartesian(xlim= c(0.9,1), ylim = c(0.5,10)) +
  geom_smooth(method = 'lm', color='red')
da_plot<-ggplot(aes(x=density, y=alcohol), data=df_wine) +
  geom_point(alpha=1/20) +
  coord_cartesian(xlim=c(0.9,1), ylim = c(7,13)) +
  geom_smooth(method = 'lm', color='red')
grid.arrange(ds_plot, da_plot, ncol=2)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



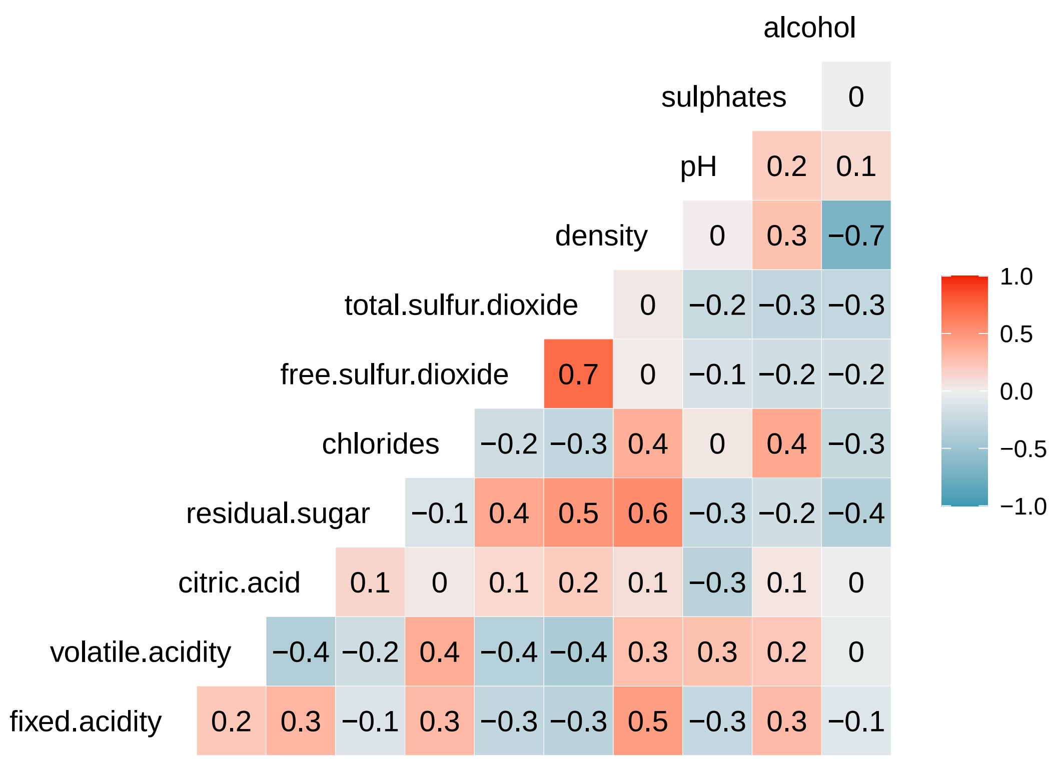
Correlation

```
cor(df_wine[, 1:11])
```

```
##                      fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity           1.00000000        0.2188802  0.32434804     -0.1121625
## volatile.acidity        0.21888021        1.0000000 -0.37814701     -0.1961870
## citric.acid             0.32434804       -0.3781470  1.00000000      0.1423526
## residual.sugar         -0.11216248       -0.1961870  0.14235264      1.0000000
## chlorides               0.29813441        0.3770866  0.03891721     -0.1291294
## free.sulfur.dioxide    -0.28255983       -0.3524202  0.13334041      0.4031857
## total.sulfur.dioxide   -0.32905272       -0.4144808  0.19528216      0.4955420
## density                 0.45887324        0.2712221  0.09606870      0.5524611
## pH                     -0.25261819        0.2615858 -0.32975275     -0.2672487
## sulphates               0.29947064        0.2258837  0.05609606     -0.1861283
## alcohol                -0.09554077       -0.0377080 -0.01054963     -0.3594917
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.29813441         -0.28255983          -0.32905272
## volatile.acidity      0.37708658         -0.35242024          -0.41448080
## citric.acid           0.03891721          0.13334041           0.19528216
## residual.sugar       -0.12912937          0.40318565           0.49554203
## chlorides             1.00000000         -0.19497275          -0.27970815
## free.sulfur.dioxide  -0.19497275          1.00000000           0.72100861
## total.sulfur.dioxide -0.27970815          0.72100861           1.00000000
## density               0.36243509          0.02588831           0.03236451
## pH                    0.04479595         -0.14604245          -0.23844723
```

```
## sulphates             0.39546984         -0.18832343          -0.27576660
## alcohol              -0.25682704         -0.17975147          -0.26569168
##                         density          pH    sulphates       alcohol
## fixed.acidity         0.45887324 -0.25261819  0.299470640 -0.095540771
## volatile.acidity      0.27122213  0.26158583  0.225883709 -0.037708002
## citric.acid           0.09606870 -0.32975275  0.056096064 -0.010549631
## residual.sugar        0.55246112 -0.26724866 -0.186128260 -0.359491718
## chlorides             0.36243509  0.04479595  0.395469839 -0.256827041
## free.sulfur.dioxide   0.02588831 -0.14604245 -0.188323432 -0.179751473
## total.sulfur.dioxide  0.03236451 -0.23844723 -0.275766596 -0.265691680
## density               1.00000000  0.01178381  0.259305929 -0.686786680
## pH                    0.01178381  1.00000000  0.192246554  0.121309406
## sulphates             0.25930593  0.19224655  1.000000000 -0.002960343
## alcohol              -0.68678668  0.12130941 -0.002960343  1.000000000
```

```r
ggcorr(df_wine[, 1:11], label = TRUE, hjust = 1.0, size = 4, layout.exp = 4)
```



Using multinomial logistic regression model

```
## # weights:  49 (36 variable)
## initial  value 12638.686418
## iter  10 value 9511.698674
## iter  20 value 9141.237511
## iter  30 value 8080.229120
```

```
## iter  40 value 7848.384051
## iter  50 value 7841.233046
## iter  60 value 7840.828686
## iter  70 value 7840.689781
## final  value 7840.665200
## converged

## Call:
## multinom(formula = quality ~ colors + sulphates + free.sulfur.dioxide +
##     total.sulfur.dioxide + colors:total.sulfur.dioxide, data = df_wine)
##
## Coefficients:
##   (Intercept) colorswhite sulphates free.sulfur.dioxide total.sulfur.dioxide
## 4  0.02178989    4.466482  1.855173        -0.085152265           0.04940783
## 5  1.27045347    3.141137  2.027555        -0.030692035           0.05909247
## 6  0.61055957    4.500301  3.947616        -0.012570660           0.03913702
## 7 -1.18584170    5.687751  5.442266        -0.002747937           0.02452636
## 8 -2.92717303    6.183397  4.538400         0.010209829           0.01858763
## 9 -9.52891973   11.101039  2.245185         0.010591527           0.05448470
##   colorswhite:total.sulfur.dioxide
## 4                      -0.05247857
## 5                      -0.05792717
## 6                      -0.05095674
## 7                      -0.04702913
## 8                      -0.04404284
## 9                      -0.08692202
##
## Std. Errors:
##   (Intercept) colorswhite sulphates free.sulfur.dioxide total.sulfur.dioxide
## 4   0.4496770   0.3533258 0.5270540         0.012158138           0.01577409
## 5   0.2987269   0.2236957 0.2654476         0.009902425           0.01514005
## 6   0.2896887   0.2133224 0.2414658         0.009787463           0.01514198
## 7   0.3147767   0.2415168 0.2730530         0.010050999           0.01535580
## 8   0.5231444   0.4360567 0.4938390         0.010913114           0.01774864
## 9   0.2934593   0.2902828 0.1298575         0.039039995           0.08456371
##   colorswhite:total.sulfur.dioxide
## 4                       0.01376033
## 5                       0.01307722
## 6                       0.01308157
## 7                       0.01330481
## 8                       0.01583138
## 9                       0.08395435
##
## Residual Deviance: 15681.33
## AIC: 15753.33

##
## z test of coefficients:
##
##                                     Estimate Std. Error  z value  Pr(>|z|)
## 4:(Intercept)                      0.0217899  0.4496770   0.0485 0.9613522
## 4:colorswhite                      4.4664820  0.3533258  12.6413 < 2.2e-16 ***
## 4:sulphates                        1.8551734  0.5270540   3.5199 0.0004317 ***
## 4:free.sulfur.dioxide             -0.0851523  0.0121581  -7.0037 2.492e-12 ***
```

```
## 4:total.sulfur.dioxide              0.0494078  0.0157741   3.1322 0.0017349 **
## 4:colorswhite:total.sulfur.dioxide -0.0524786  0.0137603  -3.8138 0.0001369 ***
## 5:(Intercept)                       1.2704535  0.2987269   4.2529 2.110e-05 ***
## 5:colorswhite                       3.1411366  0.2236957  14.0420 < 2.2e-16 ***
## 5:sulphates                         2.0275549  0.2654476   7.6382 2.202e-14 ***
## 5:free.sulfur.dioxide              -0.0306920  0.0099024  -3.0994 0.0019388 **
## 5:total.sulfur.dioxide              0.0590925  0.0151400   3.9031 9.499e-05 ***
## 5:colorswhite:total.sulfur.dioxide -0.0579272  0.0130772  -4.4296 9.440e-06 ***
## 6:(Intercept)                       0.6105596  0.2896887   2.1076 0.0350621 *
## 6:colorswhite                       4.5003014  0.2133224  21.0962 < 2.2e-16 ***
## 6:sulphates                         3.9476164  0.2414658  16.3486 < 2.2e-16 ***
## 6:free.sulfur.dioxide              -0.0125707  0.0097875  -1.2844 0.1990148
## 6:total.sulfur.dioxide              0.0391370  0.0151420   2.5847 0.0097473 **
## 6:colorswhite:total.sulfur.dioxide -0.0509567  0.0130816  -3.8953 9.807e-05 ***
## 7:(Intercept)                      -1.1858417  0.3147767  -3.7672 0.0001651 ***
## 7:colorswhite                       5.6877509  0.2415168  23.5501 < 2.2e-16 ***
## 7:sulphates                         5.4422665  0.2730530  19.9312 < 2.2e-16 ***
## 7:free.sulfur.dioxide              -0.0027479  0.0100510  -0.2734 0.7845462
## 7:total.sulfur.dioxide              0.0245264  0.0153558   1.5972 0.1102200
## 7:colorswhite:total.sulfur.dioxide -0.0470291  0.0133048  -3.5347 0.0004082 ***
## 8:(Intercept)                      -2.9271730  0.5231444  -5.5953 2.202e-08 ***
## 8:colorswhite                       6.1833972  0.4360567  14.1803 < 2.2e-16 ***
## 8:sulphates                         4.5384001  0.4938390   9.1900 < 2.2e-16 ***
## 8:free.sulfur.dioxide               0.0102098  0.0109131   0.9356 0.3495018
## 8:total.sulfur.dioxide              0.0185876  0.0177486   1.0473 0.2949746
## 8:colorswhite:total.sulfur.dioxide -0.0440428  0.0158314  -2.7820 0.0054026 **
## 9:(Intercept)                      -9.5289197  0.2934593 -32.4710 < 2.2e-16 ***
## 9:colorswhite                      11.1010387  0.2902828  38.2422 < 2.2e-16 ***
## 9:sulphates                         2.2451845  0.1298575  17.2896 < 2.2e-16 ***
## 9:free.sulfur.dioxide               0.0105915  0.0390400   0.2713 0.7861608
## 9:total.sulfur.dioxide              0.0544847  0.0845637   0.6443 0.5193787
## 9:colorswhite:total.sulfur.dioxide -0.0869220  0.0839543  -1.0353 0.3005061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Findings:- 'total.sulfur.dioxide' is the main feature distinguishes between 'Red' and 'White' wine.

And as expected, 'sulphates' manifests the greatest significance.**

```
## # weights:  49 (36 variable)
## initial  value 12638.686418
## iter  10 value 8962.092968
## iter  20 value 7399.332637
## iter  30 value 7355.600126
## iter  40 value 7354.064539
## iter  50 value 7353.510045
## iter  60 value 7352.687304
## iter  70 value 7351.937606
## final  value 7351.813918
## converged
```

```
## Call:
## multinom(formula = quality ~ density * residual.sugar + density *
##     alcohol, data = df_wine)
##
## Coefficients:
##   (Intercept)     density residual.sugar   alcohol density:residual.sugar
## 4    5.645870   0.4471323     -3.5823568 17.291022              3.5541285
## 5   13.589435  -3.0489838     -0.5310360  9.367321              0.5442513
## 6    8.589446  -6.3642539     -1.2423905 16.671335              1.3309848
## 7   -9.156787   4.3627099     -0.9144025 21.458552              1.0295507
## 8   -5.117893  -3.0971129      1.0773468 32.624158             -0.8687465
## 9   -7.056620 -10.3267617     -0.7014090 24.162200              0.8546362
##   density:alcohol
## 4       -17.76241
## 5       -10.04452
## 6       -16.57723
## 7       -20.84743
## 8       -31.99894
## 9       -22.98071
##
## Std. Errors:
##   (Intercept)  density residual.sugar   alcohol density:residual.sugar
## 4    1.241520 1.150167      3.5505013 2.9378418              3.5597536
## 5    1.105439 1.087101      2.1421492 1.6265639              2.1448903
## 6    1.089940 1.082671      1.5989915 1.3947617              1.5993506
## 7    1.111482 1.090820      1.6964413 1.6544499              1.6977896
## 8    1.275878 1.165280      3.9065824 3.5026329              3.9264491
## 9    2.951701 2.969129      0.1940257 0.4692468              0.1810733
##   density:alcohol
## 4      2.99172455
## 5      1.64594492
## 6      1.41089175
## 7      1.67850171
## 8      3.57000045
## 9      0.04613767
##
## Residual Deviance: 14703.63
## AIC: 14775.63

##
## z test of coefficients:
##
##                           Estimate Std. Error   z value  Pr(>|z|)
## 4:(Intercept)             5.645870   1.241520    4.5475 5.427e-06 ***
## 4:density                 0.447132   1.150167    0.3888 0.6974581
## 4:residual.sugar         -3.582357   3.550501   -1.0090 0.3129880
## 4:alcohol                17.291022   2.937842    5.8856 3.966e-09 ***
## 4:density:residual.sugar  3.554128   3.559754    0.9984 0.3180758
## 4:density:alcohol       -17.762407   2.991725   -5.9372 2.900e-09 ***
## 5:(Intercept)            13.589435   1.105439   12.2932 < 2.2e-16 ***
## 5:density                -3.048984   1.087101   -2.8047 0.0050365 **
## 5:residual.sugar         -0.531036   2.142149   -0.2479 0.8042128
## 5:alcohol                 9.367321   1.626564    5.7590 8.463e-09 ***
## 5:density:residual.sugar  0.544251   2.144890    0.2537 0.7996940
```

```
## 5:density:alcohol        -10.044522   1.645945   -6.1026 1.044e-09 ***
## 6:(Intercept)              8.589446   1.089940    7.8807 3.257e-15 ***
## 6:density                 -6.364254   1.082671   -5.8783 4.145e-09 ***
## 6:residual.sugar          -1.242390   1.598991   -0.7770 0.4371683
## 6:alcohol                 16.671335   1.394762   11.9528 < 2.2e-16 ***
## 6:density:residual.sugar   1.330985   1.599351    0.8322 0.4052942
## 6:density:alcohol        -16.577230   1.410892  -11.7495 < 2.2e-16 ***
## 7:(Intercept)             -9.156787   1.111482   -8.2384 < 2.2e-16 ***
## 7:density                  4.362710   1.090820    3.9995 6.348e-05 ***
## 7:residual.sugar          -0.914402   1.696441   -0.5390 0.5898785
## 7:alcohol                 21.458552   1.654450   12.9702 < 2.2e-16 ***
## 7:density:residual.sugar   1.029551   1.697790    0.6064 0.5442448
## 7:density:alcohol        -20.847426   1.678502  -12.4203 < 2.2e-16 ***
## 8:(Intercept)             -5.117893   1.275878   -4.0113 6.039e-05 ***
## 8:density                 -3.097113   1.165280   -2.6578 0.0078646 **
## 8:residual.sugar           1.077347   3.906582    0.2758 0.7827191
## 8:alcohol                 32.624158   3.502633    9.3142 < 2.2e-16 ***
## 8:density:residual.sugar  -0.868747   3.926449   -0.2213 0.8248939
## 8:density:alcohol        -31.998945   3.570000   -8.9633 < 2.2e-16 ***
## 9:(Intercept)             -7.056620   2.951701   -2.3907 0.0168165 *
## 9:density                -10.326762   2.969129   -3.4780 0.0005051 ***
## 9:residual.sugar          -0.701409   0.194026   -3.6150 0.0003003 ***
## 9:alcohol                 24.162200   0.469247   51.4915 < 2.2e-16 ***
## 9:density:residual.sugar   0.854636   0.181073    4.7198 2.360e-06 ***
## 9:density:alcohol        -22.980709   0.046138 -498.0900 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Findings:- 'alcohol' shows the greatest significance.

```
## # weights:  70 (54 variable)
## initial  value 12638.686418
## iter  10 value 8398.635854
## iter  20 value 8121.768943
## iter  30 value 7945.961689
## iter  40 value 7919.930050
## iter  50 value 7915.314451
## iter  60 value 7913.583986
## iter  70 value 7913.043000
## iter  80 value 7912.756953
## iter  90 value 7912.478425
## iter 100 value 7912.336793
## final  value 7912.336793
## stopped after 100 iterations
```

```
## Call:
## multinom(formula = quality ~ pH + pH * fixed.acidity + pH * volatile.acidity +
##     pH * citric.acid + fixed.acidity * citric.acid, data = df_wine)
##
## Coefficients:
##   (Intercept)         pH fixed.acidity volatile.acidity citric.acid
## 4   9.5473164 -0.8899710    -2.8763384        27.743736   19.065407
```

```
## 5  15.3099519 -2.5453949    -3.8611268        32.145028   21.374148
## 6  21.6248448 -3.9059736    -3.7761148         9.759831    9.940744
## 7  19.4983870 -2.8099792    -2.8178292         3.603867  -13.006344
## 8  11.8125568 -0.3122835    -0.4575317         7.532426  -38.448091
## 9  -0.3893375 -0.9648470    -2.4438854        -1.504410    2.210338
##   pH:fixed.acidity pH:volatile.acidity pH:citric.acid fixed.acidity:citric.acid
## 4        0.6943591           -8.426852      -7.914596                 0.8150699
## 5        1.1083144          -10.320497      -7.090211                 0.1989721
## 6        1.0687090           -4.521053      -4.211498                 0.4624212
## 7        0.6797651           -3.048016       1.516342                 1.1399083
## 8       -0.1760237           -3.832924       9.195868                 1.3426940
## 9        0.8986836           -1.535693       1.881870                -0.8290740
##
## Std. Errors:
##   (Intercept)        pH fixed.acidity volatile.acidity citric.acid
## 4   2.2440799 0.8941306     0.7585761        3.8480163    2.383738
## 5   3.4469167 1.1548361     0.7608160        3.1178981    3.730452
## 6   3.1947815 1.0875119     0.7435463        3.0720717    3.495051
## 7   3.7043330 1.2156619     0.8167868        4.0111368    4.170552
## 8   3.4660833 1.2153312     0.9220529        1.5075096    1.412340
## 9   0.4826808 2.0815016     1.6120241        0.3296492    1.092586
##   pH:fixed.acidity pH:volatile.acidity pH:citric.acid fixed.acidity:citric.acid
## 4        0.2501772           1.1818274       1.375416                 0.5467973
## 5        0.2480521           0.9784102       1.595259                 0.5150349
## 6        0.2431047           0.9656030       1.544804                 0.5153720
## 7        0.2634840           1.2477370       1.697460                 0.5257718
## 8        0.2977012           0.5553913       1.386956                 0.5941350
## 9        0.4795651           1.0048244       4.429621                 2.1765185
##
## Residual Deviance: 15824.67
## AIC: 15932.67
```

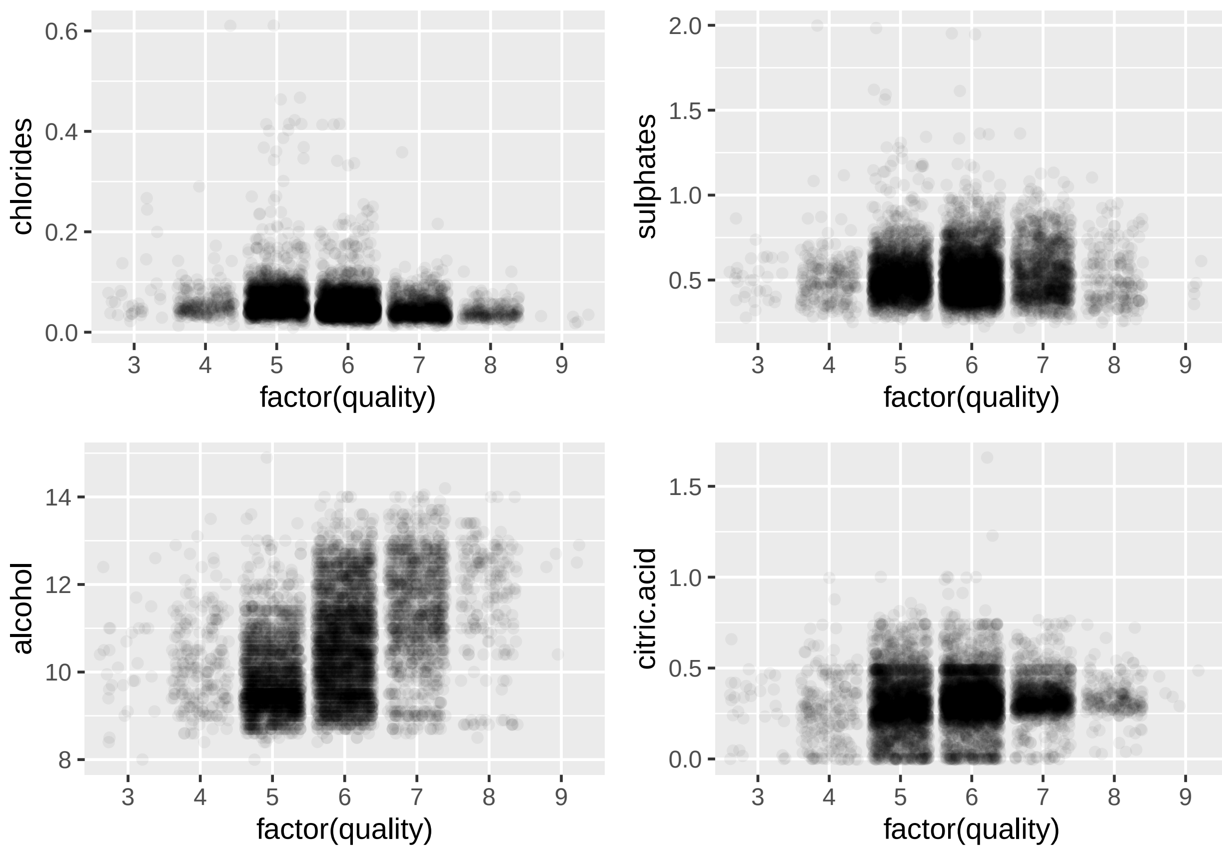
```
##
## z test of coefficients:
##
##                              Estimate Std. Error  z value  Pr(>|z|)
## 4:(Intercept)                 9.54732    2.24408   4.2544 2.096e-05 ***
## 4:pH                         -0.88997    0.89413  -0.9953 0.3195671
## 4:fixed.acidity              -2.87634    0.75858  -3.7918 0.0001496 ***
## 4:volatile.acidity           27.74374    3.84802   7.2099 5.600e-13 ***
## 4:citric.acid                19.06541    2.38374   7.9981 1.263e-15 ***
## 4:pH:fixed.acidity            0.69436    0.25018   2.7755 0.0055122 **
## 4:pH:volatile.acidity        -8.42685    1.18183  -7.1304 1.001e-12 ***
## 4:pH:citric.acid             -7.91460    1.37542  -5.7543 8.699e-09 ***
## 4:fixed.acidity:citric.acid   0.81507    0.54680   1.4906 0.1360599
## 5:(Intercept)                15.30995    3.44692   4.4416 8.928e-06 ***
## 5:pH                         -2.54539    1.15484  -2.2041 0.0275161 *
## 5:fixed.acidity              -3.86113    0.76082  -5.0750 3.875e-07 ***
## 5:volatile.acidity           32.14503    3.11790  10.3098 < 2.2e-16 ***
## 5:citric.acid                21.37415    3.73045   5.7296 1.006e-08 ***
## 5:pH:fixed.acidity            1.10831    0.24805   4.4681 7.893e-06 ***
## 5:pH:volatile.acidity       -10.32050    0.97841 -10.5482 < 2.2e-16 ***
## 5:pH:citric.acid             -7.09021    1.59526  -4.4446 8.807e-06 ***
## 5:fixed.acidity:citric.acid   0.19897    0.51503   0.3863 0.6992543
```

```
## 6:(Intercept)               21.62484    3.19478   6.7688 1.299e-11 ***
## 6:pH                        -3.90597    1.08751  -3.5917 0.0003286 ***
## 6:fixed.acidity             -3.77611    0.74355  -5.0785 3.804e-07 ***
## 6:volatile.acidity           9.75983    3.07207   3.1770 0.0014883 **
## 6:citric.acid                9.94074    3.49505   2.8442 0.0044518 **
## 6:pH:fixed.acidity           1.06871    0.24310   4.3961 1.102e-05 ***
## 6:pH:volatile.acidity       -4.52105    0.96560  -4.6821 2.839e-06 ***
## 6:pH:citric.acid            -4.21150    1.54480  -2.7262 0.0064062 **
## 6:fixed.acidity:citric.acid  0.46242    0.51537   0.8973 0.3695817
## 7:(Intercept)               19.49839    3.70433   5.2637 1.412e-07 ***
## 7:pH                        -2.80998    1.21566  -2.3115 0.0208063 *
## 7:fixed.acidity             -2.81783    0.81679  -3.4499 0.0005608 ***
## 7:volatile.acidity           3.60387    4.01114   0.8985 0.3689376
## 7:citric.acid              -13.00634    4.17055  -3.1186 0.0018170 **
## 7:pH:fixed.acidity           0.67977    0.26348   2.5799 0.0098826 **
## 7:pH:volatile.acidity       -3.04802    1.24774  -2.4428 0.0145724 *
## 7:pH:citric.acid             1.51634    1.69746   0.8933 0.3716961
## 7:fixed.acidity:citric.acid  1.13991    0.52577   2.1681 0.0301536 *
## 8:(Intercept)               11.81256    3.46608   3.4080 0.0006543 ***
## 8:pH                        -0.31228    1.21533  -0.2570 0.7972148
## 8:fixed.acidity             -0.45753    0.92205  -0.4962 0.6197464
## 8:volatile.acidity           7.53243    1.50751   4.9966 5.835e-07 ***
## 8:citric.acid              -38.44809    1.41234 -27.2230 < 2.2e-16 ***
## 8:pH:fixed.acidity          -0.17602    0.29770  -0.5913 0.5543354
## 8:pH:volatile.acidity       -3.83292    0.55539  -6.9013 5.153e-12 ***
## 8:pH:citric.acid             9.19587    1.38696   6.6303 3.351e-11 ***
## 8:fixed.acidity:citric.acid  1.34269    0.59414   2.2599 0.0238266 *
## 9:(Intercept)               -0.38934    0.48268  -0.8066 0.4198884
## 9:pH                        -0.96485    2.08150  -0.4635 0.6429816
## 9:fixed.acidity             -2.44389    1.61202  -1.5160 0.1295104
## 9:volatile.acidity          -1.50441    0.32965  -4.5637 5.027e-06 ***
## 9:citric.acid                2.21034    1.09259   2.0230 0.0430697 *
## 9:pH:fixed.acidity           0.89868    0.47957   1.8740 0.0609366 .
## 9:pH:volatile.acidity       -1.53569    1.00482  -1.5283 0.1264331
## 9:pH:citric.acid             1.88187    4.42962   0.4248 0.6709550
## 9:fixed.acidity:citric.acid -0.82907    2.17652  -0.3809 0.7032645
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

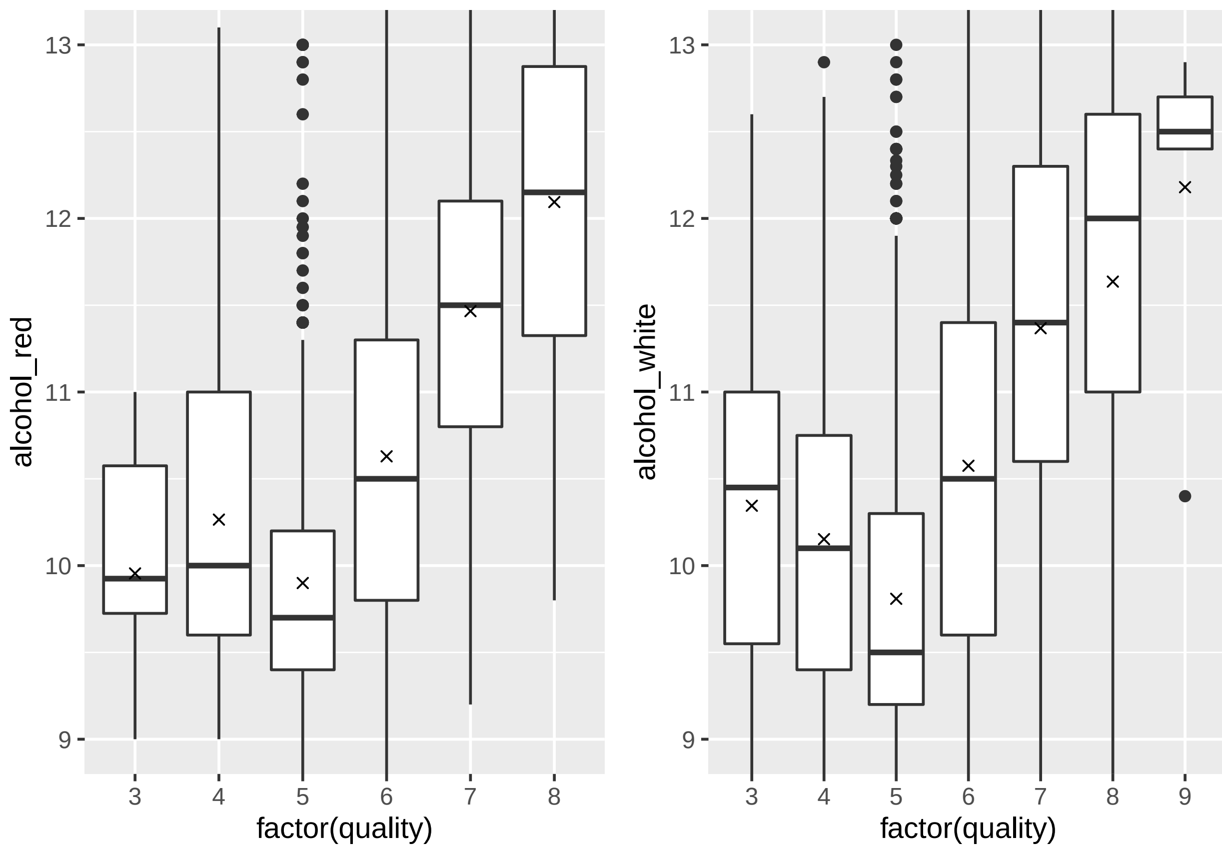
Findings:- Here 'citric.acid' shows the greatest significance.

Now, Let's look at how the categorical features of 'quality' vary with 4 different selected predictors -

chlorides, sulphates, alcohol, citric.acid.



Findings:-'alcohol' definitely most likely to have a positive impact on 'quality'. But it seems also worth investigating 'chlorides', 'sulphates' and 'citric.acid' further.




```
## # A tibble: 13 x 7
## # Groups:   quality [7]
##    quality colors mean_alcohol median_alcohol up_quantile low_quantile     n
##      <int> <fct>         <dbl>          <dbl>       <dbl>        <dbl> <int>
##  1       3 red            9.96           9.93        10.6         9.72    10
##  2       3 white         10.3           10.4         11           9.55    20
##  3       4 red           10.3           10           11           9.6     53
##  4       4 white         10.2           10.1         10.8         9.4    163
##  5       5 red            9.90           9.7         10.2         9.4    681
##  6       5 white          9.81           9.5         10.3         9.2   1456
##  7       6 red           10.6           10.5         11.3         9.8    638
##  8       6 white         10.6           10.5         11.4         9.6   2198
##  9       7 red           11.5           11.5         12.1        10.8    199
## 10       7 white         11.4           11.4         12.3        10.6    879
## 11       8 red           12.1           12.2         12.9        11.3     18
## 12       8 white         11.6           12           12.6        11      175
## 13       9 white         12.2           12.5         12.7        12.4      5
```

Findings:- Amazingly, in both red and white wine, the quality range from 5 to 9 is hugely affected by the amount of alcohol.

As the amount of alcohol increases, the samples get higher rated. Here,'colors' does not make any difference.

But in case of red wine, as the additive(sulphates) increases, the samples get higher rated, but this does not hold true in white wine; therefore, 'colors' is suspected of a certain relationship with the quality rating. The additive does not work in white wine.**

We've suspected that 'alcohol' is the most distinguishable property that affect the quality rating. To investigate further, we organize two groups of wine samples:

- low.alcohol(less than the median)

- high.alcohol(greater than or equal to the median)

Then, find the mean-quality rating of each group. First, get the median amount of alcohol content: 10.3, then select samples with alcohol content less than the median, and greater than or equal to the median. Lastly, get the mean-quality rating for the low alcohol and high alcohol groups.

```
## [1] 10.3
```

```
## [1] 5.476071
```

```
## [1] 6.145827
```

Findings: Judging from the series of the plots above, 'colors' of wine (red or white) is

somewhat associated with quality-rating when it comes to a certain properties such as 'sulphates' and 'citric.acid'.**

Seemingly, in red wine, more amount of 'sulphates' and 'citric.acid' promote higher ratings,

as compared to white wine.

Another Important Investigation :In terms of higher quality, which wine is dominant?

```
qual_color <- group_by(df_wine, quality, colors)
qual_total <- group_by(df_wine, colors)
df_qc.1 <- summarise(qual_color, fixed.acidity=length(fixed.acidity),
                     pH=length(pH), n=n()); df_qc.1
```

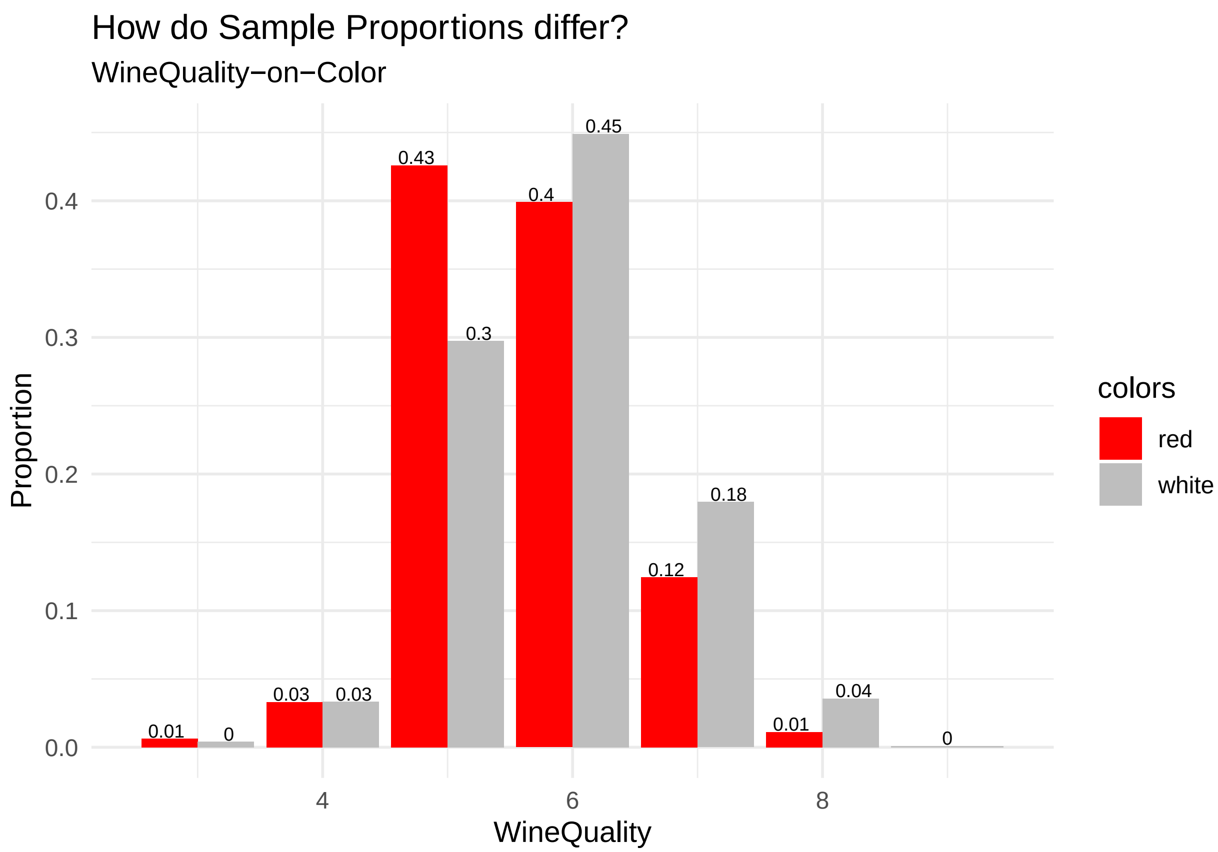
```
## `summarise()` has grouped output by 'quality'. You can override using the `.groups` argument.
```

```
## # A tibble: 13 x 5
## # Groups:   quality [7]
##    quality colors fixed.acidity    pH     n
##      <int> <fct>          <int> <int> <int>
##  1       3 red               10    10    10
##  2       3 white             20    20    20
##  3       4 red               53    53    53
##  4       4 white            163   163   163
##  5       5 red              681   681   681
##  6       5 white           1456  1456  1456
##  7       6 red              638   638   638
##  8       6 white           2198  2198  2198
##  9       7 red              199   199   199
## 10       7 white            879   879   879
## 11       8 red               18    18    18
## 12       8 white            175   175   175
## 13       9 white              5     5     5
```

```
df_qc.2 <- summarise(qual_total, fixed.acidity_t=length(fixed.acidity),
                     pH_t=length(pH), n_t=n()); df_qc.2
```

```
## # A tibble: 2 x 4
##   colors fixed.acidity_t  pH_t   n_t
##   <fct>            <int> <int> <int>
## 1 red               1599  1599  1599
## 2 white             4896  4896  4896
```

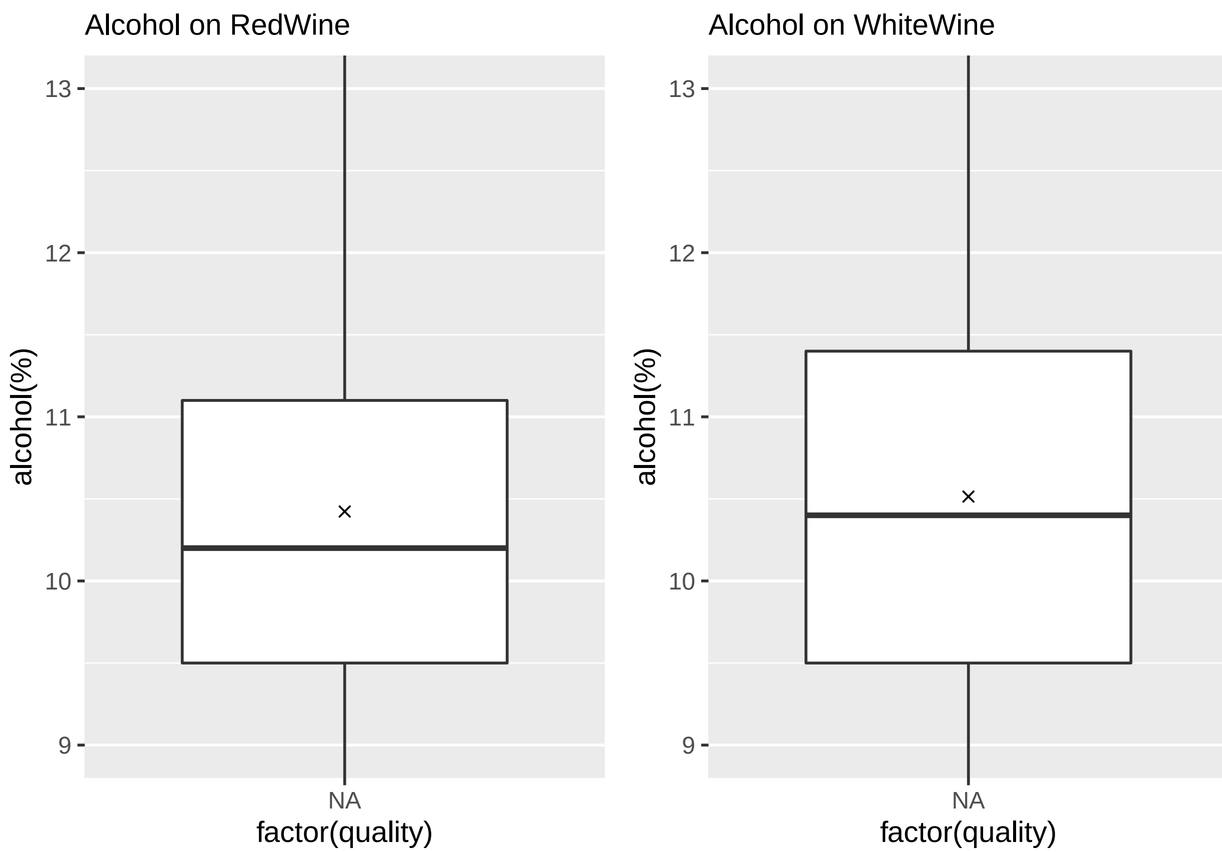
```r
df_qc.3 <- merge(df_qc.1, df_qc.2, by='colors')
df_qc.3$pH_pro <- df_qc.3$pH/df_qc.3$pH_t
df_wine$quality = unfactor(as.data.frame(df_wine$quality))
ggplot(aes(x=quality, y=pH_pro, fill=colors), data=df_qc.3) +
  geom_bar(stat="identity", position=position_dodge()) +
  scale_fill_manual(values=c("red", "grey")) +
  geom_text(aes(label=round(pH_pro, digits = 2)),
            vjust=-0.1, size=2.5, position = position_dodge(width = 1)) +
  labs(x = "WineQuality", y = "Proportion",
       title = "How do Sample Proportions differ?",
       subtitle = "WineQuality-on-Color") +
  theme_minimal()
```



Findings:- Above plot clearly shows how wine color is associated with the quality ratings. ## For the lower ratings -3/4/5, 'red' shows higher proportion. and for the higher ratings, ## the reverse is true. This means the white wine in general receives higher rating.

Final Plot

Based on P-value from the multinomial regression, and the correlation matrix, alcohol is another powerful predictor.



Findings: In the quality ratings range from 5 to 9, alcohol in general helps wines get rated higher.

Final Plot Summary:

We are looking for the features that help our wine get higher rated. Based on the above discussion, we can definitely

say that white wine is generally easy to get higher rated, but if we maintain enough level of 'sulphates' and 'citric.acid',

red wine can be compatible. In addition, alcohol is the powerful element that renders higher rating.

[Observations]

The difference of the red and white in their sample size caused some confusion, for which I tried to use

the proportion of the samples rather than their size.

Here, I made a set of some fascinating arguments: In general, red wines are rated lower than white wines, but if one added more additive(sulphates to increase S02 gas) and citric.acid, red wines would also be able to get rated higher. pH-level(3.11 ~ 3.21) is the range where 'citric.acid' and' fixed.acidity' work best. From the quality rating of 5, the advantage of alcohol in wines kicks off. However, there are still several questions that we need to answer. One can suspect that there might be a threshold or ceiling where alcohol does not work. The same is true when it comes to 'citric.acid' and 'sulphates'. Since there is a lack of wine samples with much higher alcoholic content or more 'citric.acid' and 'sulphates', we cannot say these properties are always the right ones to receive higher rating.