Data_607_Tidyverse_Extended_Assignment

Mahmud Hasan Al Raji extended by Enid Roman

2022-11-14

The main task here is to Create an example by using one or more TidyVerse packages, and any dataset from fivethirtyeight.com or Kaggle, create a programming sample "vignette" that demonstrates how to use one or more of the capabilities of the selected TidyVerse package with the selected dataset.Here, I have selected a data set from kaggle.com and put that data set on my github. The data set reflects the different properties of two types of wine.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df<-read.csv("https://raw.githubusercontent.com/Raji030/data607_tidyverse_create_assignment/main/winequa
glimpse(df)
```

```
## Rows: 6,497
## Columns: 13
## $ type                <chr> "white", "white", "white", "white", "white", "whi~
## $ fixed.acidity       <dbl> 7.0, 6.3, 8.1, 7.2, 7.2, 8.1, 6.2, 7.0, 6.3, 8.1,~
## $ volatile.acidity    <dbl> 0.27, 0.30, 0.28, 0.23, 0.23, 0.28, 0.32, 0.27, 0~
## $ citric.acid         <dbl> 0.36, 0.34, 0.40, 0.32, 0.32, 0.40, 0.16, 0.36, 0~
## $ residual.sugar      <dbl> 20.70, 1.60, 6.90, 8.50, 8.50, 6.90, 7.00, 20.70,~
## $ chlorides           <dbl> 0.045, 0.049, 0.050, 0.058, 0.058, 0.050, 0.045, ~
## $ free.sulfur.dioxide <dbl> 45, 14, 30, 47, 47, 30, 30, 45, 14, 28, 11, 17, 1~
## $ total.sulfur.dioxide <dbl> 170, 132, 97, 186, 186, 97, 136, 170, 132, 129, 6~
## $ density             <dbl> 1.0010, 0.9940, 0.9951, 0.9956, 0.9956, 0.9951, 0~
```

```
## $ pH                  <dbl> 3.00, 3.30, 3.26, 3.19, 3.19, 3.26, 3.18, 3.00, 3~
## $ sulphates           <dbl> 0.45, 0.49, 0.44, 0.40, 0.40, 0.44, 0.47, 0.45, 0~
## $ alcohol             <dbl> 8.8, 9.5, 10.1, 9.9, 9.9, 10.1, 9.6, 8.8, 9.5, 11~
## $ quality             <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 5, 5, 7, 5, 7, 6~
```

The dplyr package in tidyverse can be used to subset a data frame by subsetting rows using specific column value:

```
read_wine_data<-df %>% filter(type=="red")
glimpse(read_wine_data)
```

```
## Rows: 1,599
## Columns: 13
## $ type                <chr> "red", "red", "red", "red", "red", "red", "red", ~
## $ fixed.acidity       <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7.5~
## $ volatile.acidity    <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600, ~
## $ citric.acid         <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00, 0~
## $ residual.sugar      <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1,~
## $ chlorides           <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069, ~
## $ free.sulfur.dioxide <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, 16~
## $ total.sulfur.dioxide <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 102,~
## $ density             <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978, 0~
## $ pH                  <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39, 3~
## $ sulphates           <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47, 0~
## $ alcohol             <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 10.~
## $ quality             <int> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 5, 7~
```

The dplyr package in tidyverse package can also be used to count the number of times a column value occurs:

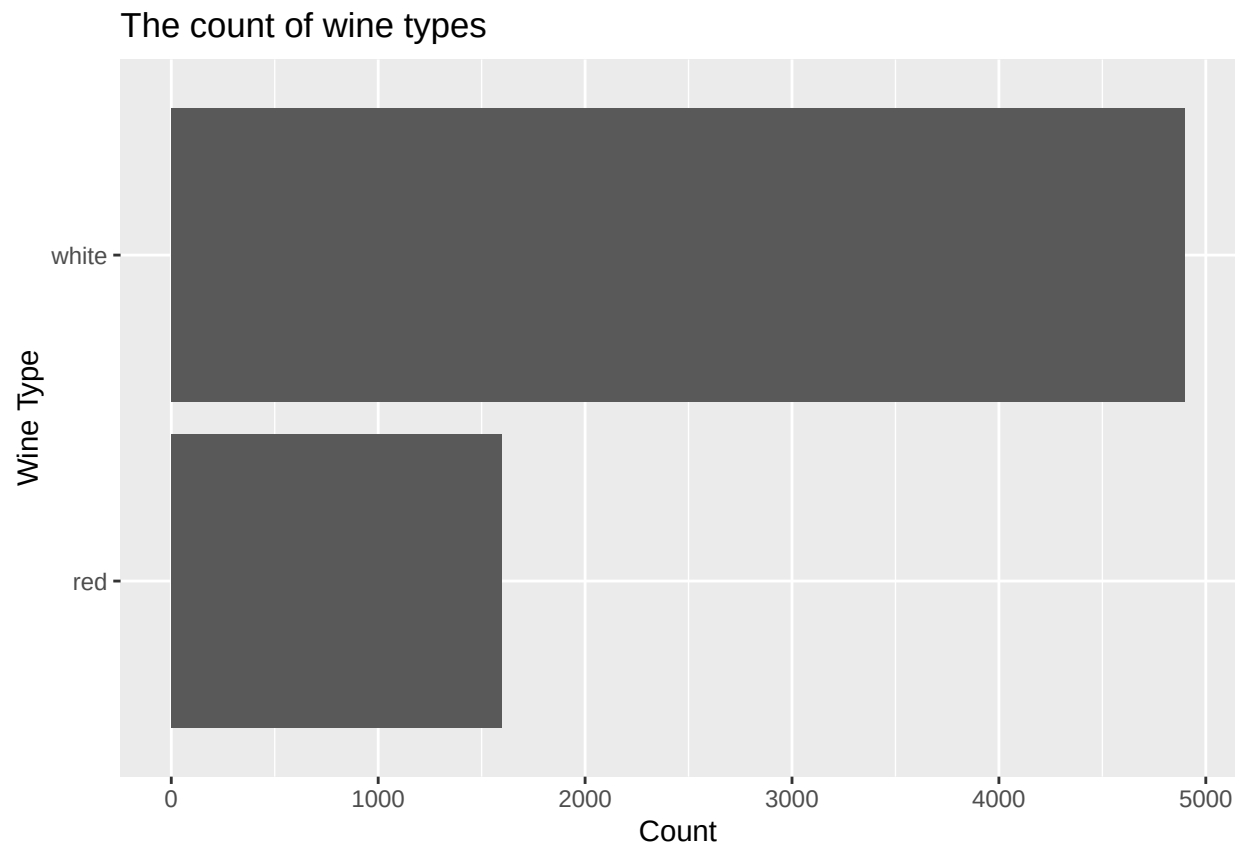
```
count_wine_type<-df %>% count(type)
count_wine_type
```

```
##    type    n
## 1   red 1599
## 2 white 4898
```

The ggplot2 package in tidyverse can be used to visualize relationship between variables of interest

```
# Creating horizontal plot to visualize the count by wine types
ggplot(data=count_wine_type, aes(x=type, y=n)) +
  geom_bar(stat="identity") +
```

```
    labs(
      x = "Wine Type",
      y = "Count",
      title = "The count of wine types"
    ) +
    coord_flip()
```



The purrr package is used to compute the summary of different variables

```
df %>% split(.$type) %>% # from base R
  map(summary)
```

```
## $red
##      type           fixed.acidity    volatile.acidity  citric.acid
##  Length:1599        Min.   : 4.600   Min.   :0.1200   Min.   :0.0000
##  Class :character   1st Qu.: 7.100   1st Qu.:0.3900   1st Qu.:0.0900
##  Mode  :character   Median : 7.900   Median :0.5200   Median :0.2600
##                     Mean   : 8.322   Mean   :0.5277   Mean   :0.2711
##                     3rd Qu.: 9.200   3rd Qu.:0.6400   3rd Qu.:0.4200
##                     Max.   :15.900   Max.   :1.5800   Max.   :1.0000
##                     NA's   :2        NA's   :1        NA's   :1
##  residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median : 2.200   Median :0.07900   Median :14.00       Median : 38.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87       Mean   : 46.47
```

```
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00       Max.   :289.00
##
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##                   NA's   :2       NA's   :2
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
##
##
## $white
##      type           fixed.acidity    volatile.acidity  citric.acid
##  Length:4898        Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  Class :character   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Mode  :character   Median : 6.800   Median :0.2600   Median :0.3200
##                     Mean   : 6.856   Mean   :0.2783   Mean   :0.3343
##                     3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##                     Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##                     NA's   :8        NA's   :7        NA's   :2
##  residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median : 5.200   Median :0.04300   Median : 34.00      Median :134.0
##  Mean   : 6.393   Mean   :0.04578   Mean   : 35.31      Mean   :138.4
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00      Max.   :440.0
##  NA's   :2        NA's   :2
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##                   NA's   :7       NA's   :2
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
##
```

Tidyverse Extended by Enid Roman

Distribution of Single Variables

Wine Quality

```
table(df$quality)
```

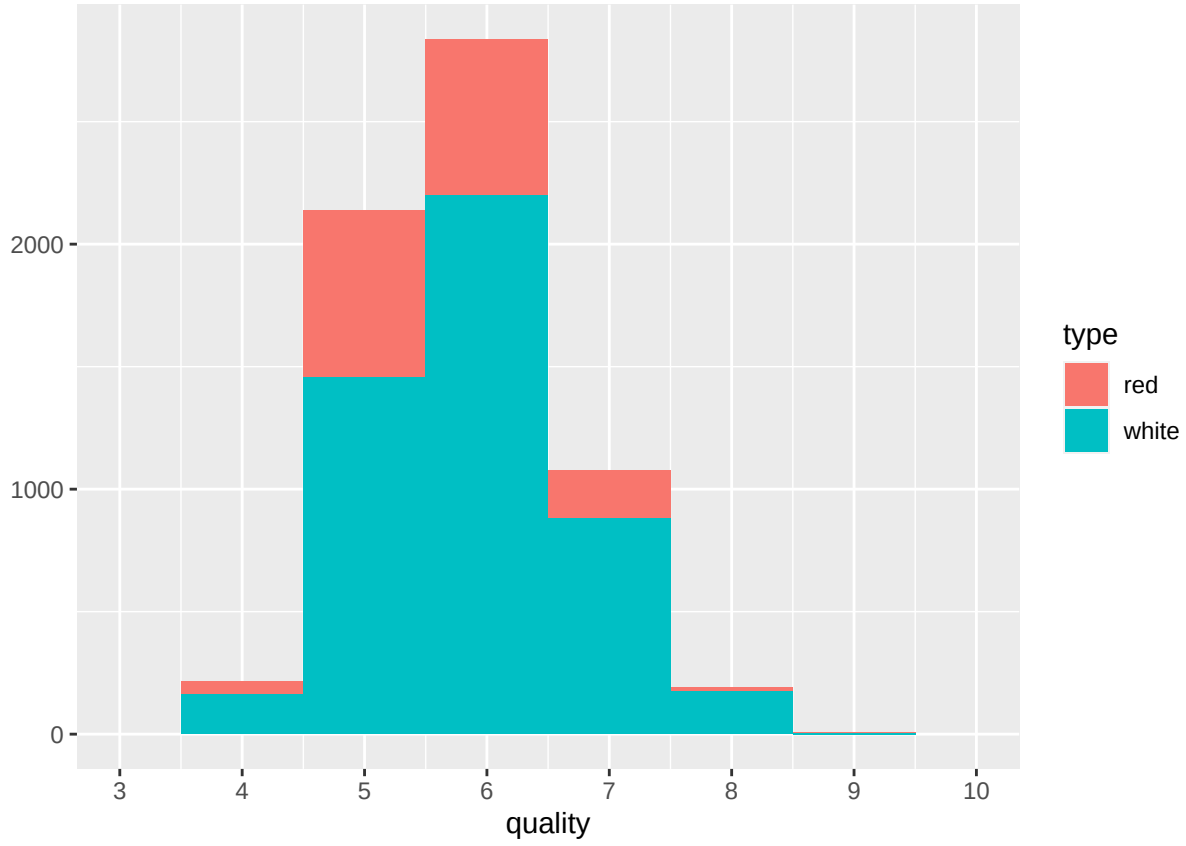
Red wine sample is smaller. We know that number of observations for red and white are different in the dataset, but still we can see that for both colors it's normal distribution with almost the same picks at 5 and 6 quality point.

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

```
library(ggplot2)
qplot(quality, data = df, fill = type, binwidth = 1) +
  scale_x_continuous(breaks = seq(3,10,1), lim = c(3,10))
```

Here we use the function qplot() in ggplot2 part of Tidyverse Package is very similar to the basic plot() function from the R base package. It can be used to create and combine easily different types of plots.

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```



Level of alcohol

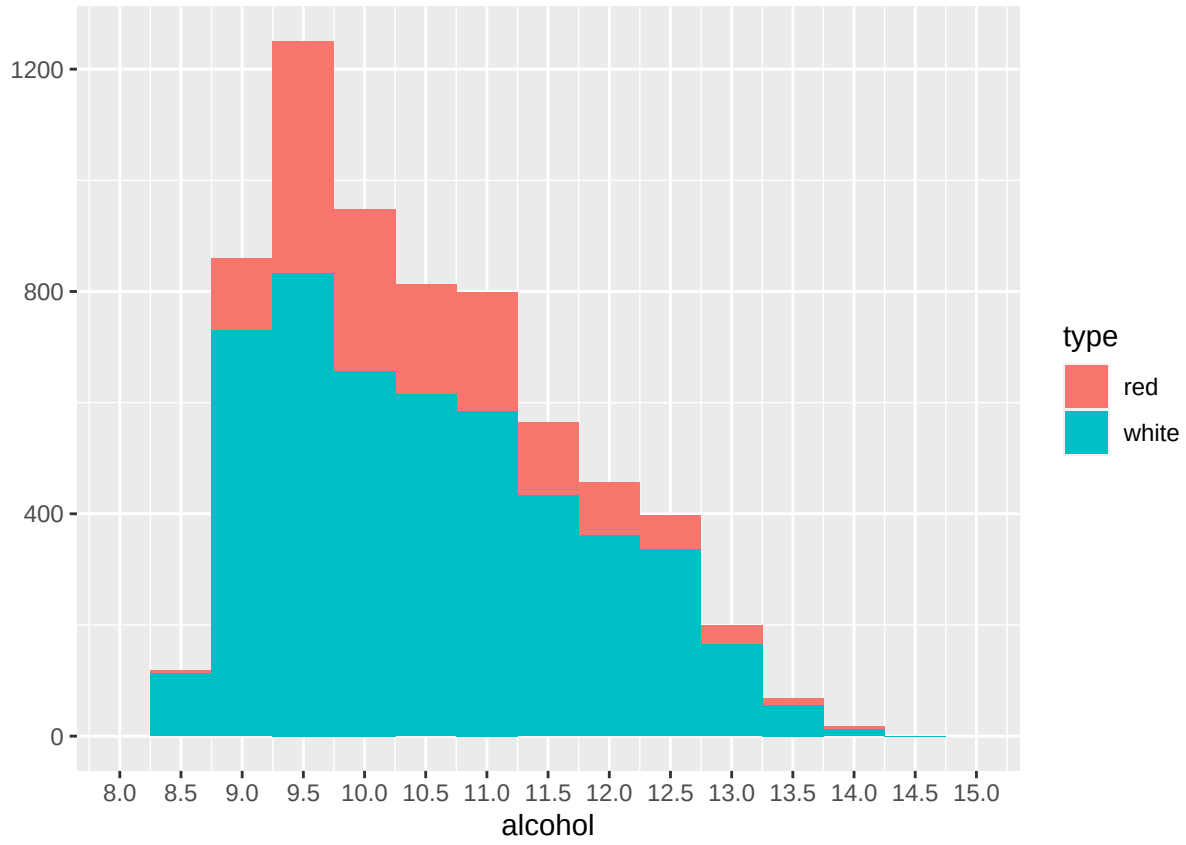
```
summary(df$alcohol)
```

Alcohol level distribution looks skewed. Again, red wine sample is smaller but it gives the same pattern of alcohol level distribution as white wines. Most frequently wines have 9.5%, mean is 10.49% of alcohol.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.30   10.49   11.30   14.90
```

```
qplot(alcohol, data = df, fill = type, binwidth = 0.5) +
    scale_x_continuous(breaks = seq(8,15,0.5), lim = c(8,15))
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```



Wine Density

```
summary(df$density)
```

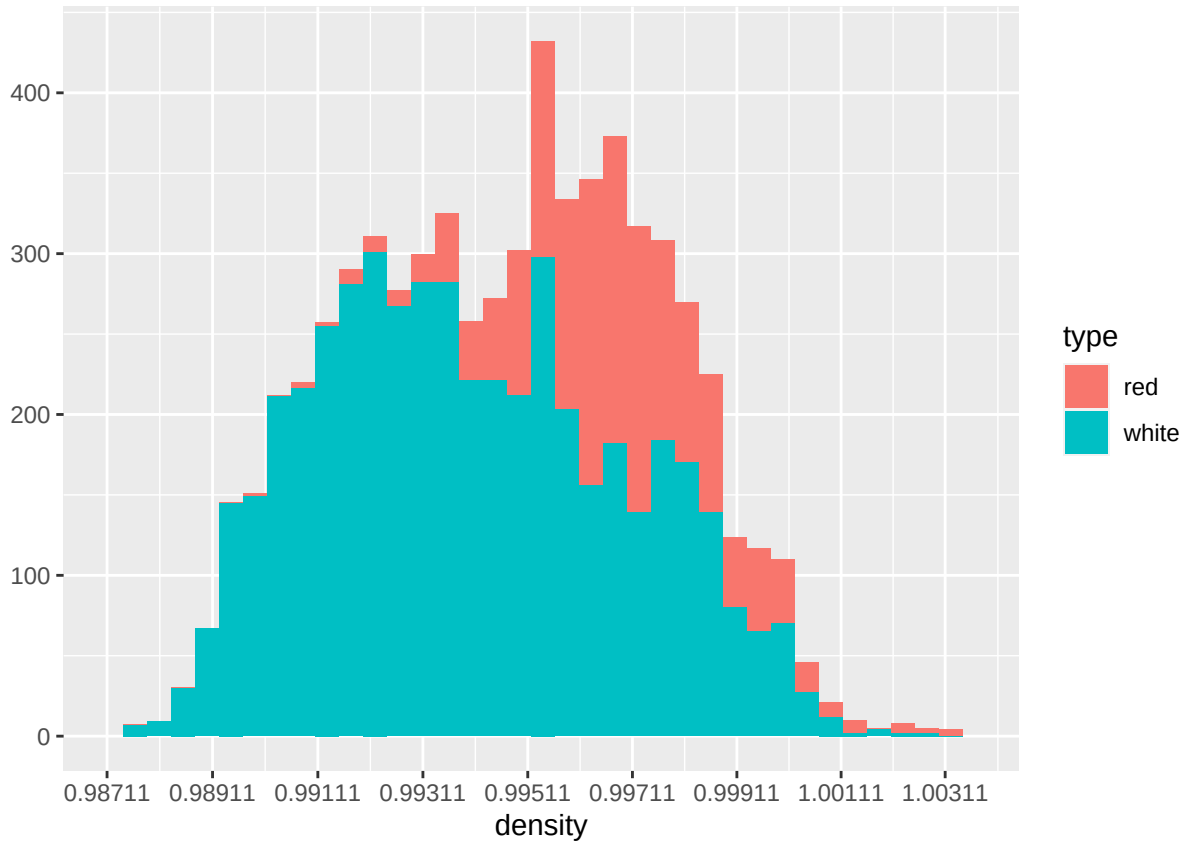
Looking at 'table' summary we see that there are two outliers: 1.0103 and 1.03898. To see the distribution of density clearer I used log10 and limited the data. Now we can see that density distribution of white wine is bimodal and of red wine is normal.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9923  0.9949  0.9947  0.9970  1.0390
```

```
qplot(density, data = df, fill = type, binwidth = 0.0002) +
    scale_x_log10(lim = c(min(df$density), 1.00370),
                  breaks = seq(min(df$density), 1.00370, 0.002))
```

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```



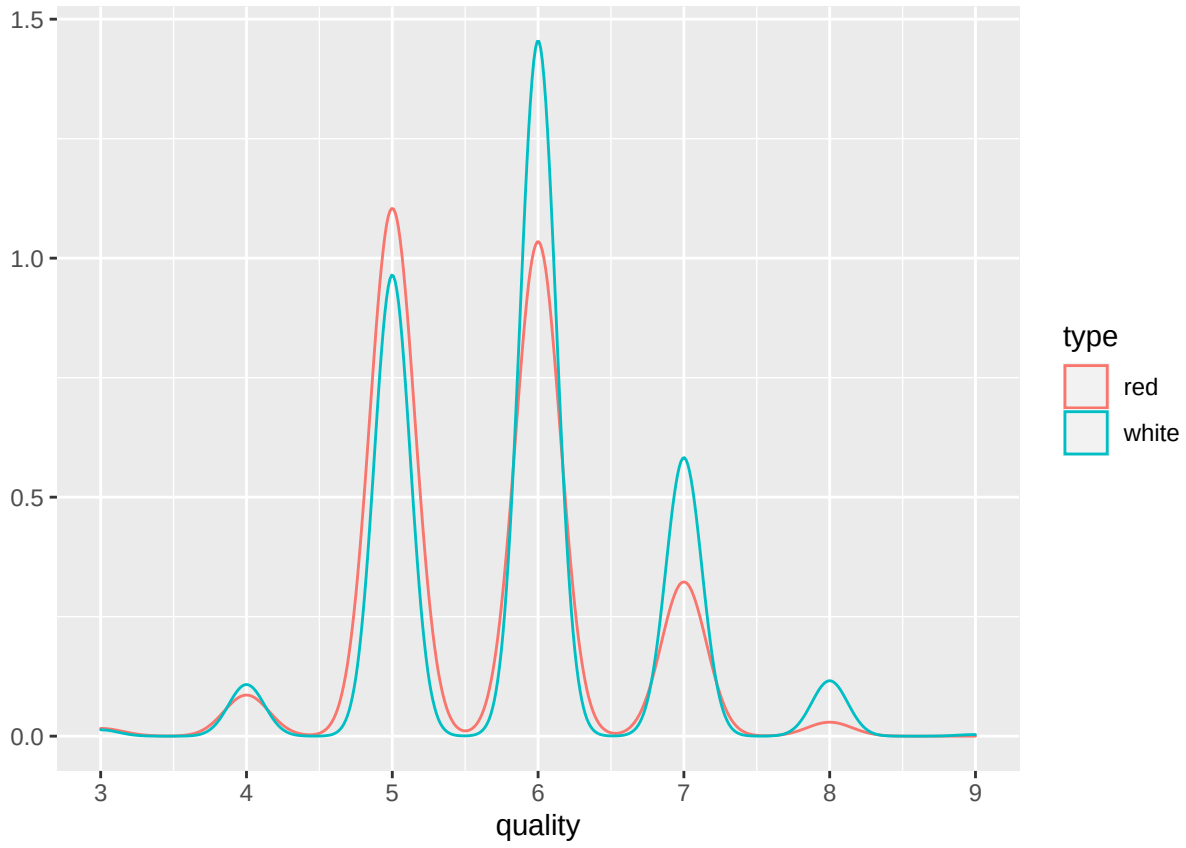
Distribution of Two and More Variables

Density of Quality by color

```
qplot(quality, data = df, binwidth = 1, color = type, geom = "density") +
    scale_x_continuous(breaks = seq(3, 9, 1))
```

In out sample we have almost the same amount of red and white wines with quality '3', '4' and '9', more red wines with quality '5' and more white wines with quality "6", "7" and "8".

```
## Warning: Ignoring unknown parameters: binwidth
```


1.5
1.0
0.5
0.0
3
4
5
6
7
8
9
quality
type
red
white