Wine Quality Data

Harsh Mehta

2022-11-30

Exploring the Data

```
str(df)
```

```
## 'data.frame':    6465 obs. of  16 variables:
##  $ type                : chr  "white" "white" "white" "white" ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3
0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34
0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045
0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22
...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49
0.45 ...
##  $ alcohol             : num  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 ...
##  $ BillOfMaterials     : int  558 618 630 630 624 612 642 582 558 570 ...
##  $ StorageCost         : int  264 240 252 288 264 240 240 300 300 294 ...
##  $ Price               : int  1788 1800 1818 1776 1818 1794 1776 1830 1836
1830 ...
```

```
head(df)
```

```
##    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 BillOfMaterials StorageCost Price
## 1       6             558         264  1788
## 2       6             618         240  1800
## 3       6             630         252  1818
## 4       6             630         288  1776
## 5       6             624         264  1818
## 6       6             612         240  1794
```

The df data frame consist of 6465 Columns and 15 Rows

- type
- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality
- Bill Of Materials(Cost for Making the wine)
- Cost of storage
- Price

Cleaning the Data

The Data frame 'df' consist tibbles with the value '0'. To get rid of this data we will use `na.omit()` Function. The new data frame now only consists of 6308 row.

```
df[df == 0] <- NA
winequality<-na.omit(df)
dim(winequality)
```

```
## [1] 6308   16
```

To understand the data more clearly we will add Two more columns to the data Indicating Profit and Total Cost

```
total_cost <- winequality$BillOfMaterials + winequality$StorageCost
winequality <- winequality %>%
  add_column(total_cost)

profit <- winequality$Price -winequality$total_cost
winequality <- winequality %>%
  add_column(profit)
```

Questions

Q1 -> What is the difference between different types of wines?

Q2 -> What are the costs and different attributes of that cost?

Q3 -> What determines the quality and what does the ideal wine look like?

Q4 -> What factors determines the price/profit of the wine and what is the percentage of the profit generated?

Q5 -> Is there any Correlation between different attributes of the wines?

Q1 -> What is the difference between different types of wines?

```
library("ggthemes")
library("gridExtra")
```

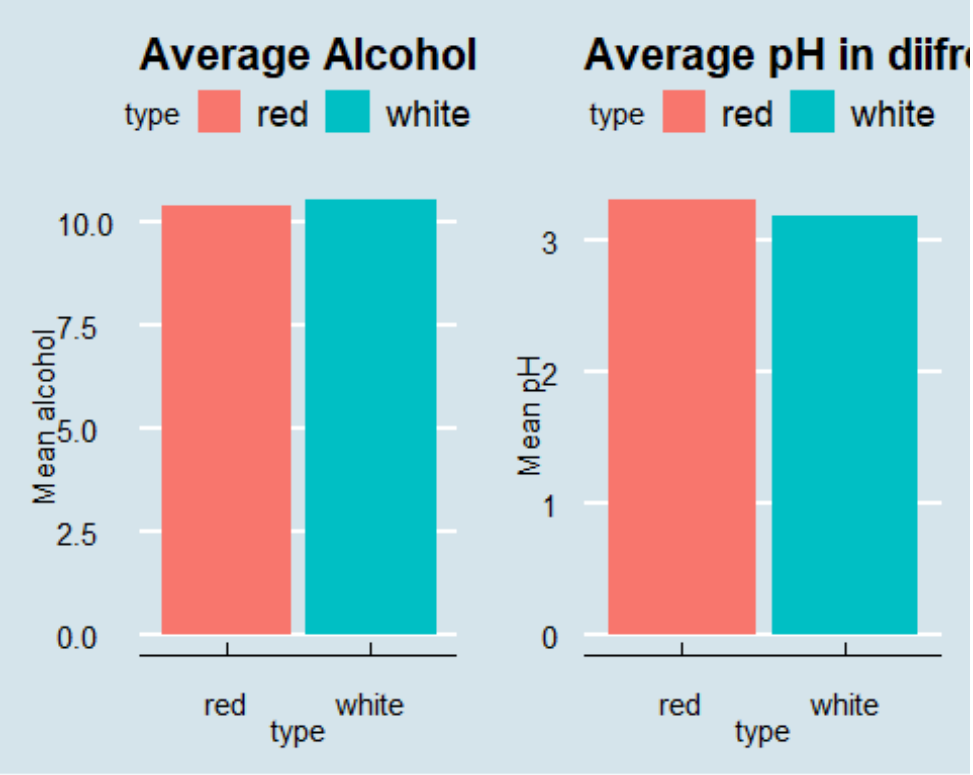
```
##
## Attaching package: 'gridExtra'
```

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

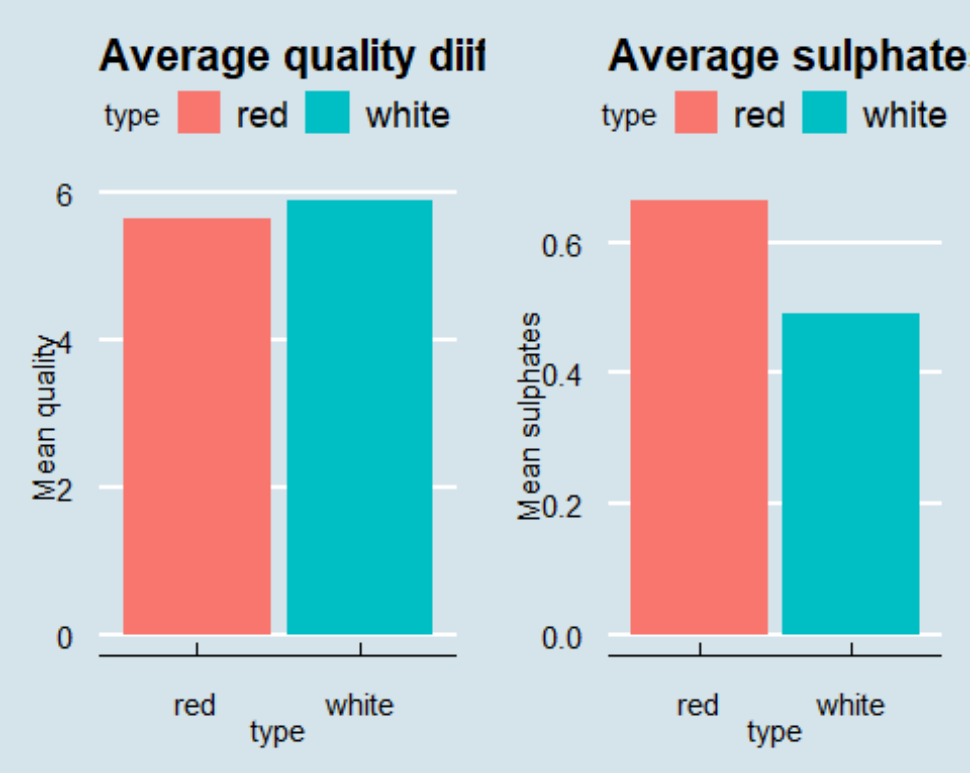
```
library(ggplot2)
g1 <- ggplot(winequality) +
  aes(x = type, y = alcohol, fill = type) +
  geom_bar(position="dodge",stat="summary",fun="mean") +

  labs(title = "Average Alcohol in diifrent types of wine",
```

```
       x = "type", y = "Mean alcohol")+
  theme_economist()

g2 <-  ggplot(winequality) +
  aes(x = type, y = pH, fill = type) +
  geom_bar(position="dodge",stat="summary",fun="mean") +

  labs(title = "Average pH in diifrent types of wine",
       x = "type", y = "Mean pH")+
  theme_economist()

g3 <- ggplot(winequality) +
  aes(x = type, y = fixed.acidity, fill = type) +
  geom_bar(position="dodge",stat="summary",fun="mean") +

  labs(title = "Average acidity in diifrent types of wine",
       x = "type", y = "Mean acidity")+
  theme_economist()

g4 <- ggplot(winequality) +
  aes(x = type, y = residual.sugar, fill = type) +
  geom_bar(position="dodge",stat="summary",fun="mean") +

  labs(title = "Average sugar in diifrent types of wine",
       x = "type", y = "Mean sugar")+
  theme_economist()

g5 <- ggplot(winequality) +
  aes(x = type, y = quality, fill = type) +
  geom_bar(position="dodge",stat="summary",fun="mean") +

  labs(title = "Average quality diifrent types of wine",
       x = "type", y = "Mean quality")+
  theme_economist()

g6 <-ggplot(winequality) +
  aes(x = type, y = sulphates, fill = type) +
  geom_bar(position="dodge",stat="summary",fun="mean") +

  labs(title = "Average sulphates in diifrent types of wine",
       x = "type", y = "Mean sulphates")+
  theme_economist()

g7 <-ggplot(winequality) +
  aes(x = type, y = Price, fill = type, fill = type) +
  geom_bar(position="dodge",stat="summary",fun="mean") +

  labs(title = "Average price of diifrent types of wine",
```

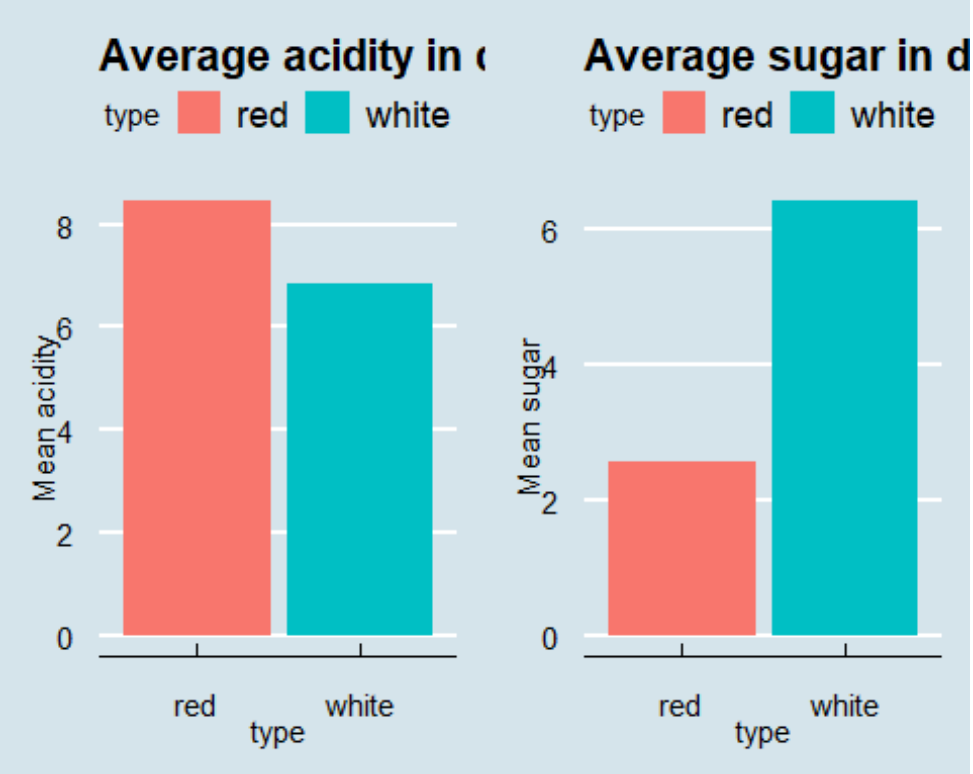
```
       x = "type", y = "Mean price")+
  theme_economist()

## Warning: Duplicated aesthetics after name standardisation: fill

g8 <-ggplot(winequality) +
  aes(x = type, y = total_cost, fill = type) +
  geom_bar(position="dodge",stat="summary",fun="median") +

  labs(title = "Median cost of diifrent types of wine",
       x = "type", y = "Mean cost")+
  theme_economist()

grid.arrange(g1, g2, ncol = 2)
```




```
grid.arrange(g5, g6, ncol = 2)
```




```
grid.arrange(g3, g4, ncol = 2)
```




```
grid.arrange(g7, g8, ncol = 2)
```

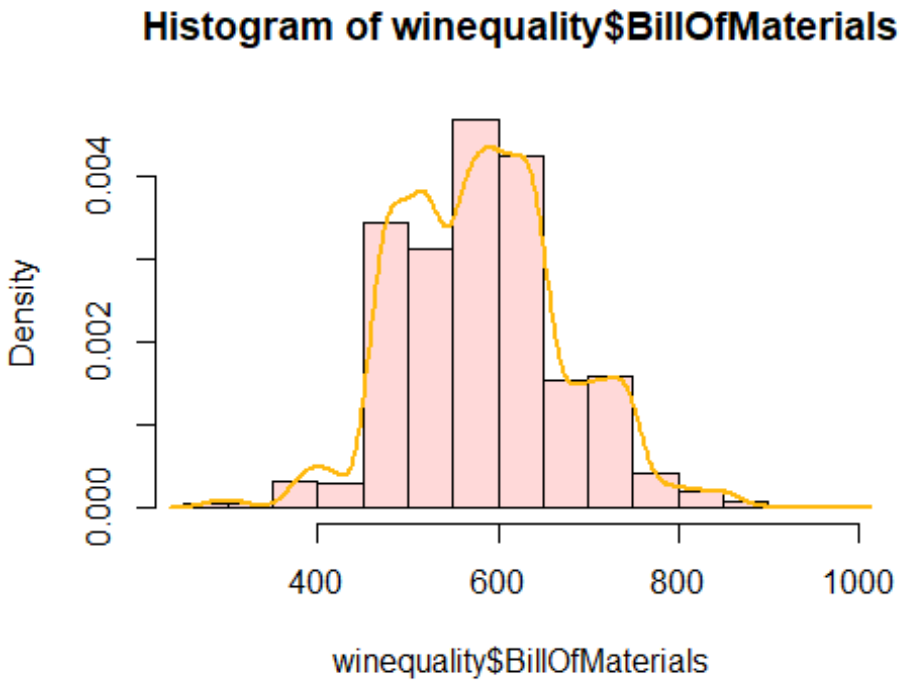
By Studying the Average contents of Red and White Wine We can Conclude that:

- White wine contains more sugar then red wine
- Red wine contains more sulfates then the white wine
- Red wine is slightly more acidic in nature

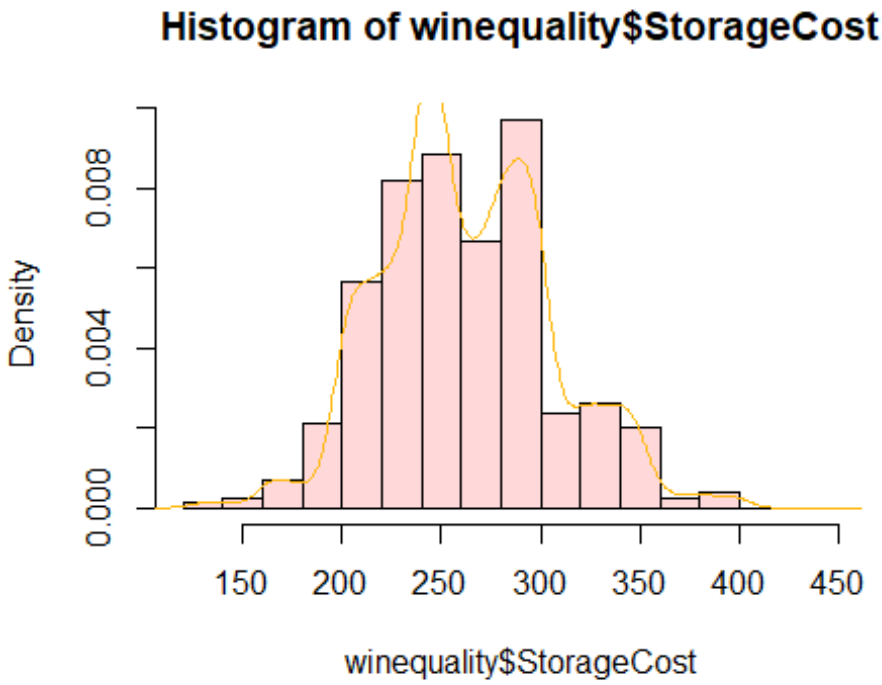
Q2 -> What are the costs and different attributes of that cost?

The total cost is based on the sum of of Bill of Materials and the Storage Cost. To explore this columns we will plot A histogram and will also see its probability density. The generic function `hist` computes a histogram of the given data values. `density()` function computes kernel density estimates.

```
hist(winequality$BillOfMaterials,freq = FALSE, col=rgb( 1,.25,.25, .2))
lines(density(winequality$BillOfMaterials), col="darkgoldenrod1", lwd=2)
```



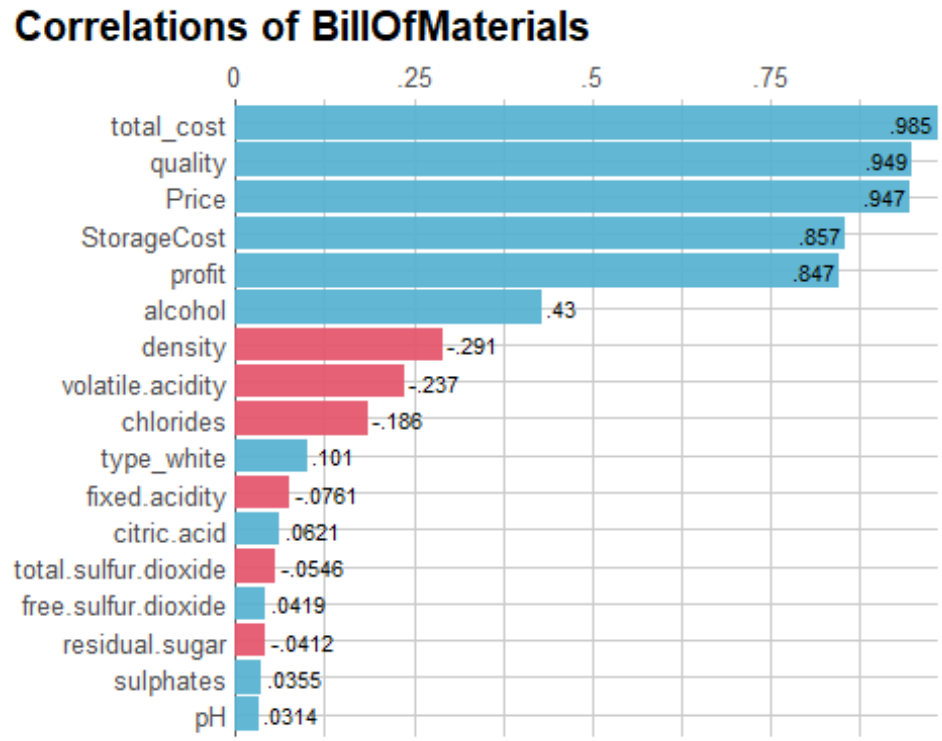

```
hist(winequality$StorageCost, freq = FALSE, col=rgb( 1,.25,.25, .2))
lines(density(winequality$StorageCost), col="darkgoldenrod1")
```



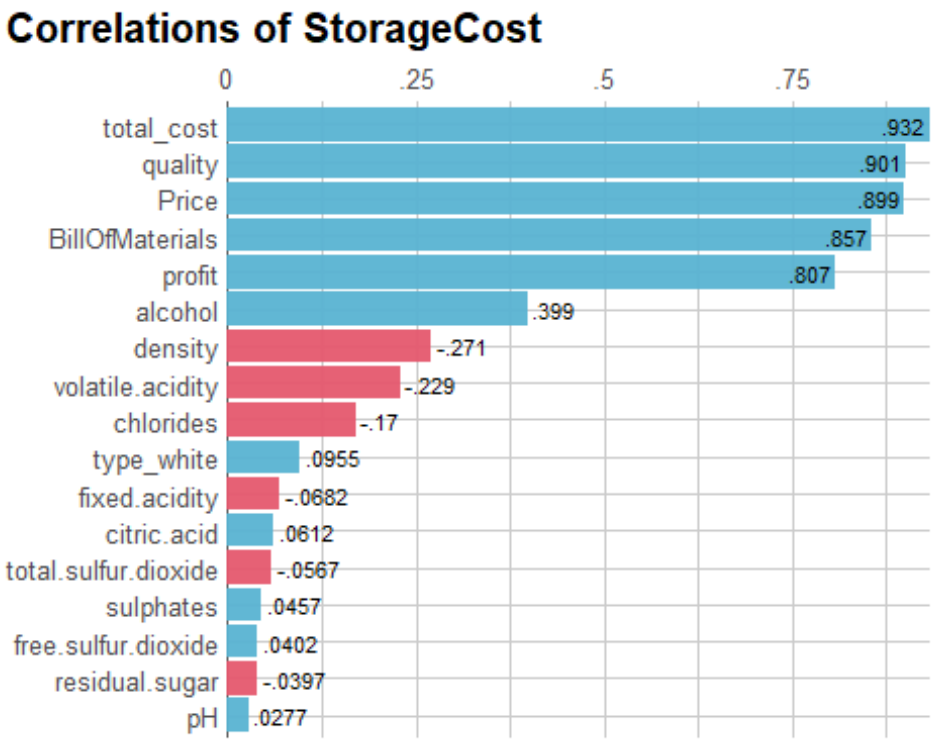
We will see the Correlation of the attributes and find out what factors affect the cost. `corr_var()`function correlates a whole dataframe with a single feature. It automatically runs `ohse` (one-hot-smart-encoding) so no need to input only numerical values.

```
library('lares')
corr_var(winequality,
         BillOfMaterials,
)
```

```
## Warning in .font_global(font, quiet = FALSE): Font 'Arial Narrow' is not
## installed, has other name, or can't be found
```




```
corr_var(winequality,
         StorageCost
)
```

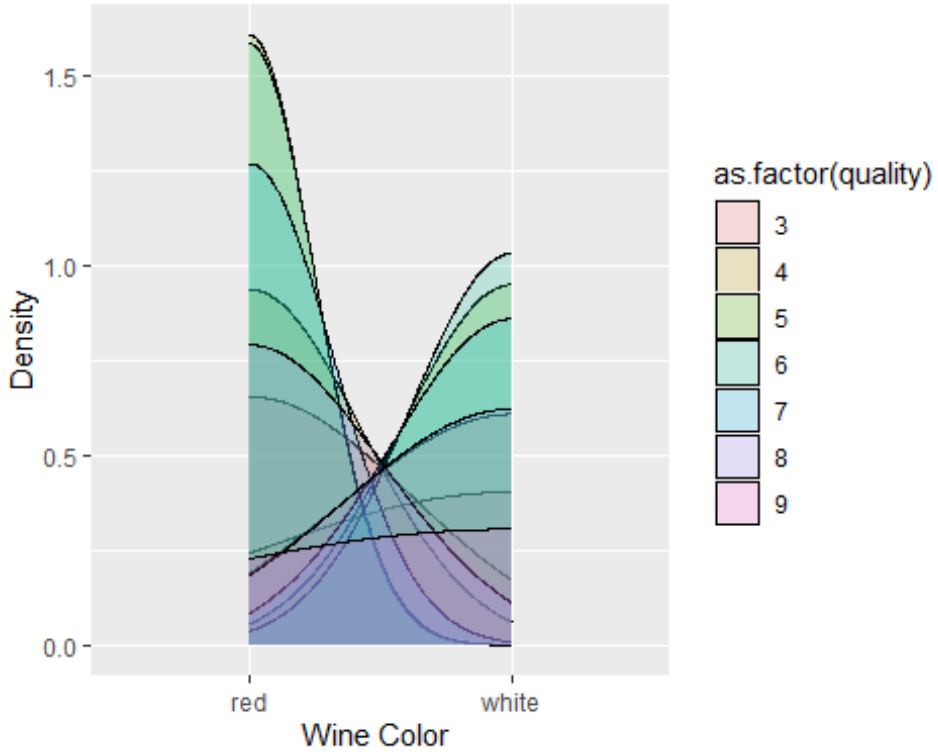


- Majority of the wines has the cost of ingredients from 500 to 700 rupees
- Majority of the wine has the storage cost of 225 to 325 rupees
- The major factor affecting the cost is the quality of the wine i.e the higher the rating of the wine, Higher its cost.

Q3 -> What determines the quality and what does the ideal wine look like?

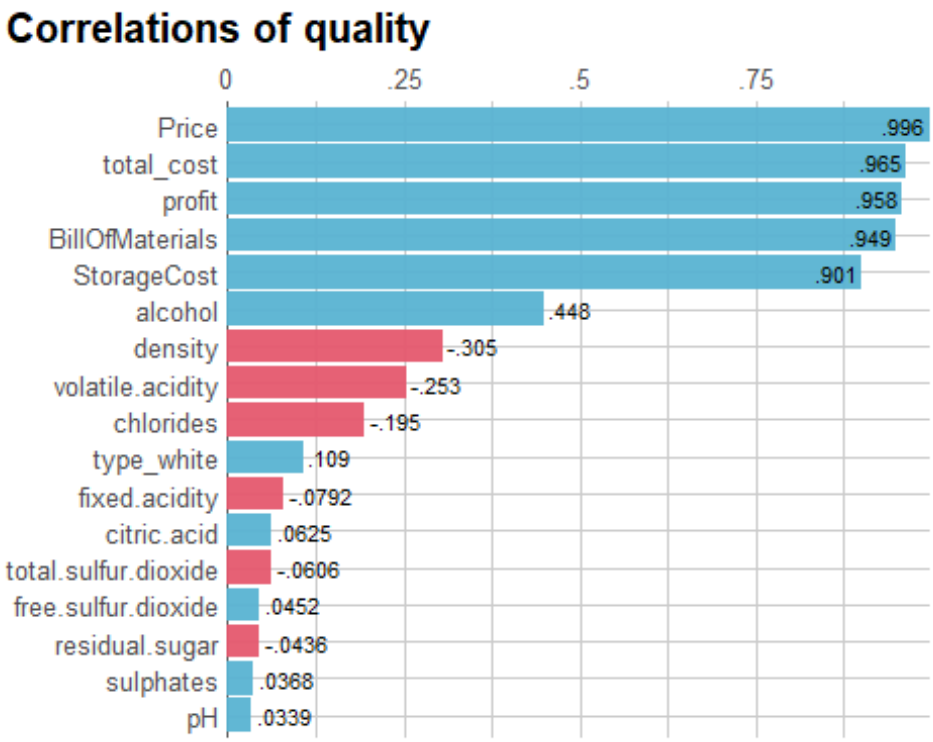
Lets Explore the quality of wine by type.

```
ggplot(winequality, aes(x = as.factor(winequality$type))) +
  geom_density(aes(fill = as.factor(quality)), alpha = 0.2)+  labs(x="Wine
Color ", y= "Density")
```



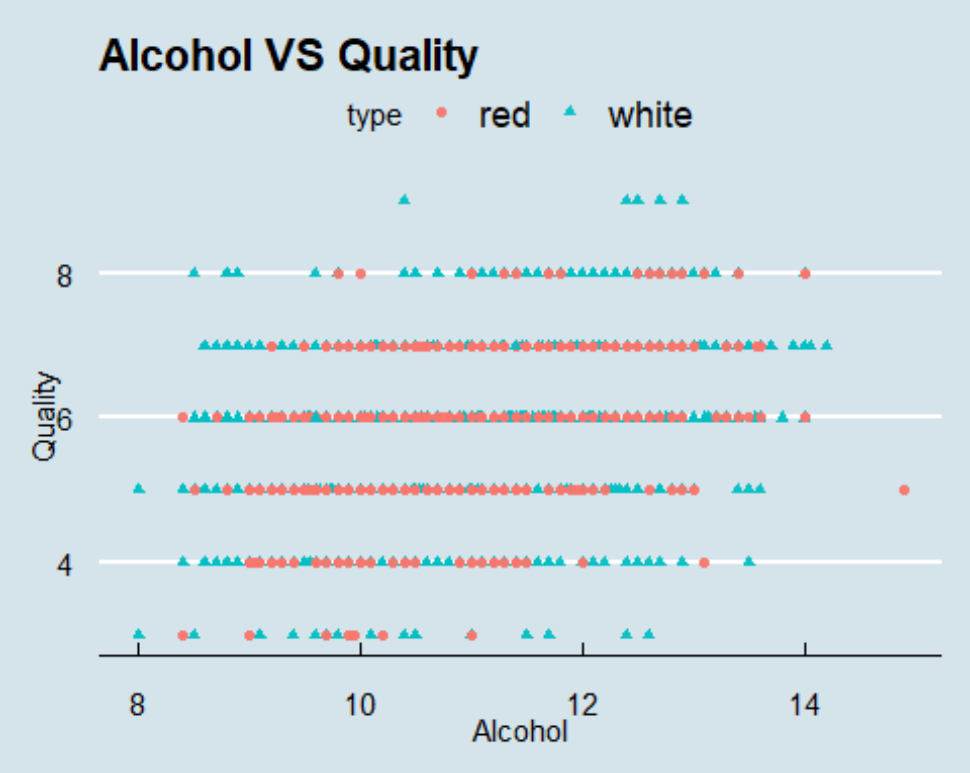
We will see the Correlation of the attributes and find out what factors affects the quality.

```
corr_var(winequality,
         quality)
```



Using ggpolt() we will plot some graphs to study the correlations.

```
g9 <-ggplot(winequality, aes(x = alcohol, y = quality, col = type))+
  geom_point(aes(shape = type))+
  labs(title = "Alcohol VS Quality",
          x = "Alcohol", y = "Quality")
g9 + theme_economist()
```




```
g10 <- ggplot(winequality, aes(winequality$Price, winequality$quality)) +
labs(title = "Price VS Quality", x = "Price", y = "Quality") +
geom_line(colour = "#0c4c8a", size = 2) + theme_economist()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.

g11 <- ggplot(winequality, aes(winequality$total_cost, winequality$quality))
+ labs(title = "Cost VS Quality",cx = "Cost", y = "Quality") +
geom_line(colour = "#87728c", size = 2) + theme_economist()
grid.arrange(g10, g11, ncol = 2)
```

We will make a different data frame with highest rated quality wine and observe it

```
library(summarytools)

##
## Attaching package: 'summarytools'

## The following object is masked from 'package:tibble':
##
##     view

max(winequality$quality)

## [1] 9

Ideal_Wine <- filter(winequality, winequality$quality == 9)
Ideal_Wine_Summarry <- summarytools::descr(Ideal_Wine, round.digits = 2,
transpose = TRUE)
Ideal_Wine_Summarry

## Non-numerical variable(s) ignored: type

## Descriptive Statistics
## Ideal_Wine
## N: 5
##
##                                  Mean    Std.Dev       Min        Q1
Median        Q3       Max
```

```
## ------------------------- --------- --------- --------- --------- -------
-- --------- ---------
##                     alcohol     12.18      1.01     10.40     12.40
12.50     12.70     12.90
##            BillOfMaterials    876.60     59.56    828.00    828.00
864.00    891.00    972.00
##                  chlorides      0.03      0.01      0.02      0.02
0.03      0.03      0.04
##                citric.acid      0.39      0.08      0.29      0.34
0.36      0.45      0.49
##                    density      0.99      0.00      0.99      0.99
0.99      0.99      1.00
##              fixed.acidity      7.42      0.98      6.60      6.90
7.10      7.40      9.10
##        free.sulfur.dioxide     33.40     13.43     24.00     27.00
28.00     31.00     57.00
##                         pH      3.31      0.08      3.20      3.28
3.28      3.37      3.41
##                      Price   2698.20     43.81   2655.00   2655.00
2700.00   2727.00   2754.00
##                     profit   1400.40     91.91   1278.00   1368.00
1395.00   1431.00   1530.00
##                    quality      9.00      0.00      9.00      9.00
9.00      9.00      9.00
##             residual.sugar      4.12      3.76      1.60      2.00
2.20      4.20     10.60
##                StorageCost    421.20     20.52    396.00    405.00
423.00    441.00    441.00
##                  sulphates      0.47      0.09      0.36      0.42
0.46      0.48      0.61
##       total.sulfur.dioxide    116.00     19.82     85.00    113.00
119.00    124.00    139.00
##                 total_cost   1297.80     58.81   1224.00   1269.00
1287.00   1332.00   1377.00
##           volatile.acidity      0.30      0.06      0.24      0.26
0.27      0.36      0.36
##
## Table: Table continues below
##
##
##
##                                  MAD     IQR     CV   Skewness   SE.Skewness
Kurtosis   N.Valid   Pct.Valid
## ------------------------- ------- ------- ------ ---------- -------------
---------- --------- -----------
##                     alcohol    0.30    0.30   0.08      -0.98          0.91
-1.03      5.00      100.00
##            BillOfMaterials   53.37   63.00   0.07       0.61          0.91
-1.51      5.00      100.00
##                   chlorides    0.01    0.01   0.27      -0.25          0.91
```

```
-2.12      5.00      100.00
##                citric.acid    0.10    0.11   0.21       0.14          0.91
-2.01      5.00      100.00
##                  density    0.00    0.00   0.00       1.04          0.91
-0.96      5.00      100.00
##             fixed.acidity    0.44    0.50   0.13       0.84          0.91
-1.18      5.00      100.00
##        free.sulfur.dioxide    4.45    4.00   0.40       0.98          0.91
-1.03      5.00      100.00
##                         pH    0.12    0.09   0.03       0.00          0.91
-1.90      5.00      100.00
##                     Price   66.72   72.00   0.02       0.09          0.91
-2.06      5.00      100.00
##                    profit   53.37   63.00   0.07       0.09          0.91
-1.57      5.00      100.00
##                   quality    0.00    0.00   0.00        NaN          0.91
NaN      5.00      100.00
##              residual.sugar    0.89    2.20   0.91       0.90          0.91
-1.16      5.00      100.00
##                StorageCost   26.69   36.00   0.05      -0.11          0.91
-2.12      5.00      100.00
##                  sulphates    0.06    0.06   0.20       0.43          0.91
-1.48      5.00      100.00
##       total.sulfur.dioxide    8.90   11.00   0.17      -0.44          0.91
-1.44      5.00      100.00
##                 total_cost   66.72   63.00   0.05       0.11          0.91
-1.81      5.00      100.00
##           volatile.acidity    0.04    0.10   0.19       0.21          0.91
-2.21      5.00      100.00
```

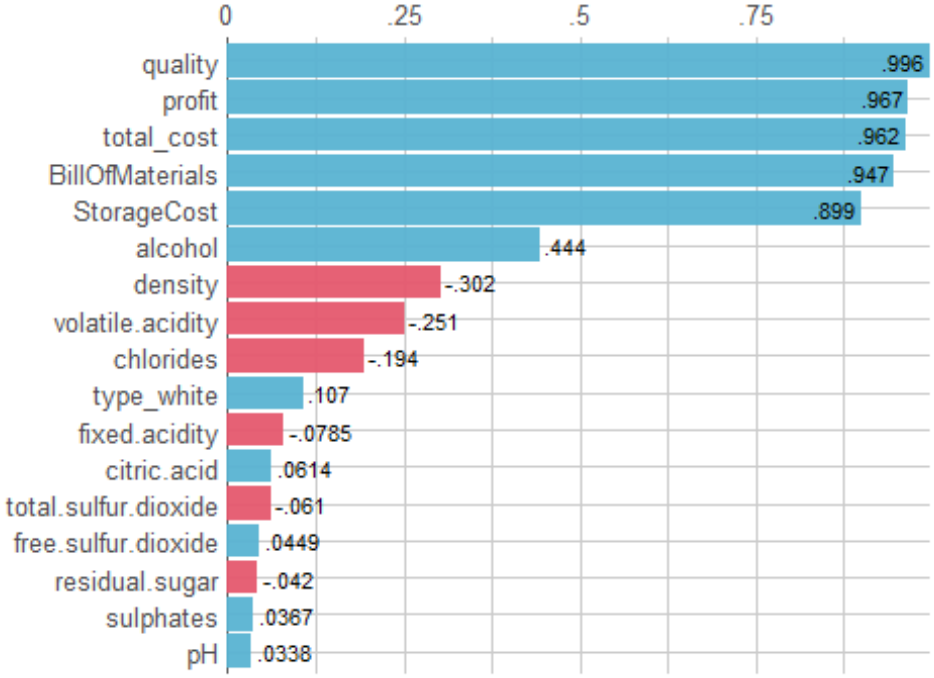
- All the highest rated wines are white
- The price and cost of the highest rated wines are also very high

Q4-> What factors determines the price/profit of the wine and what is the percentage of the profit generated?

We will see the Correlation of the attributes and find out what factors affect the price and profit.

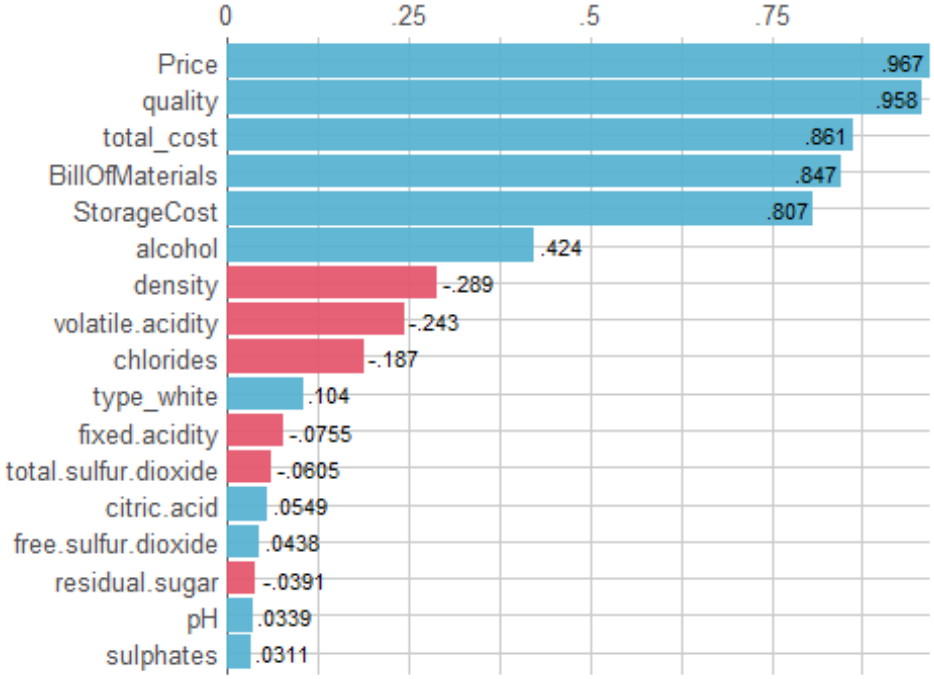
```
corr_var(winequality,
         Price)
```

Correlations of Price




```
corr_var(winequality,
         profit)
```

Correlations of profit

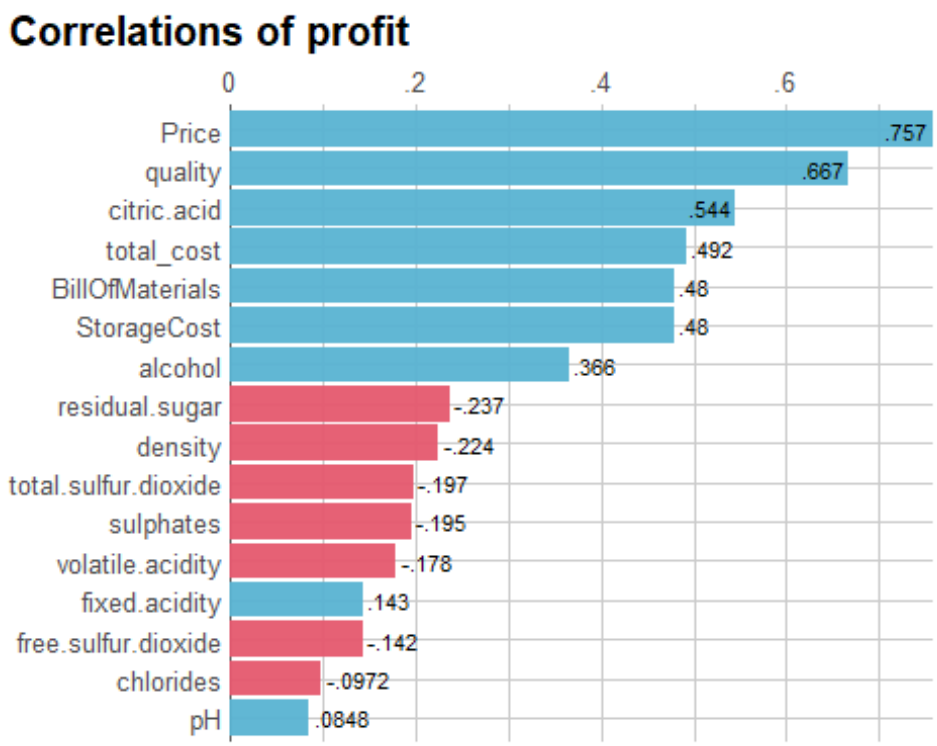


Profit persent with the different attributes.

```
Avereage_Profit_Persent <-
mean(winequality$profit)*100/mean(winequality$Price)
Avereage_Profit_Persent

## [1] 51.64962

min_Profit_Persent <- min(winequality$profit)*100/mean(winequality$Price)
min_Profit_Persent

## [1] 24.52926

Max_Profit_Persent <- max(winequality$profit)*100/mean(winequality$Price)
Max_Profit_Persent

## [1] 87.48199
```

We will now compute the top most profitable wine and analyze its summary and correlation

```
Max_Profit<-slice_max(winequality, n = 10, winequality$profit)
corr_var(Max_Profit,
         profit)
```




```
Max_profit_Wine_Summarry <- summarytools::descr(Max_Profit, round.digits = 4,
transpose = TRUE)
Max_profit_Wine_Summarry
```

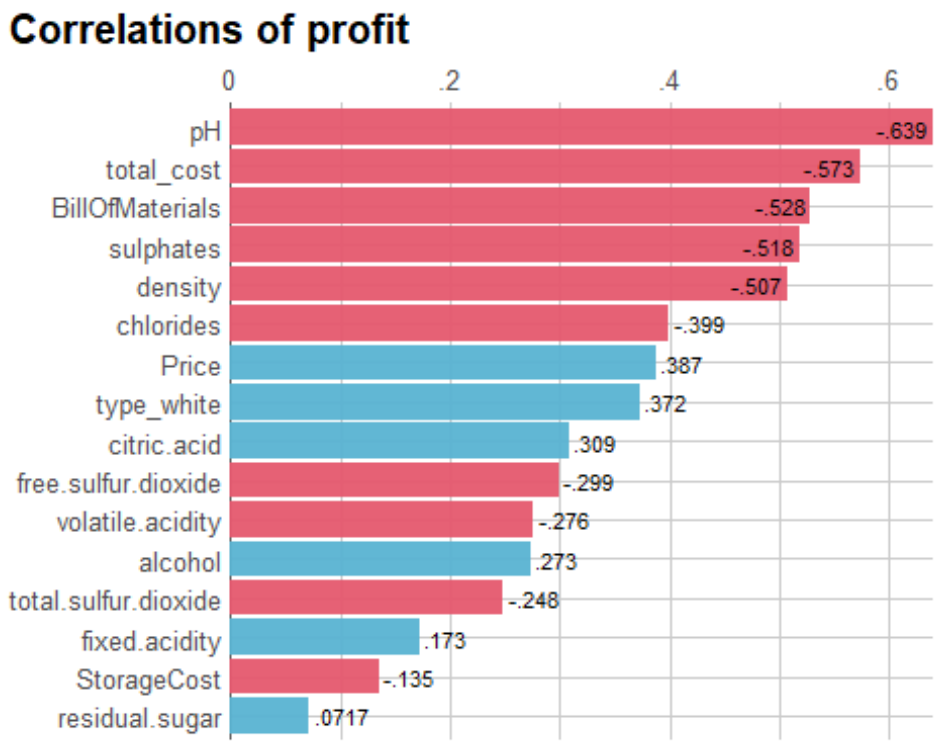
```
## Non-numerical variable(s) ignored: type

## Descriptive Statistics
## Max_Profit
## N: 11
##
##                                    Mean    Std.Dev         Min          Q1
Median          Q3
## -------------------------- ----------- ---------- ----------- -----------
----------- -----------
##                     alcohol     11.9636     0.8652     10.4000     11.2000
12.1000     12.7000
##            BillOfMaterials    785.7273    56.3916    736.0000    744.0000
752.0000    828.0000
##                  chlorides      0.0325     0.0088      0.0180      0.0240
0.0330      0.0410
##                citric.acid      0.3664     0.0780      0.2600      0.3100
0.3600      0.4500
##                    density      0.9922     0.0024      0.9892      0.9903
0.9919      0.9941
##               fixed.acidity      7.1455     0.8722      5.5000      6.7000
7.1000      7.4000
##        free.sulfur.dioxide     34.6364    12.1678     22.0000     28.0000
30.0000     38.0000
##                          pH      3.2945     0.1181      3.1200      3.2000
3.2800      3.3700
##                       Price   2540.0000   136.1022   2432.0000   2448.0000
2448.0000   2700.0000
##                      profit   1386.5455    52.8457   1352.0000   1360.0000
1368.0000   1395.0000
##                     quality      8.3636     0.5045      8.0000      8.0000
8.0000      9.0000
##             residual.sugar      5.4273     4.1807      2.0000      2.1000
4.2000      7.0000
##                 StorageCost    367.7273    48.3344    320.0000    328.0000
344.0000    423.0000
##                   sulphates      0.4900     0.1115      0.3600      0.4000
0.4800      0.5800
##       total.sulfur.dioxide    129.9091    23.4156     96.0000    113.0000
124.0000    142.0000
##                  total_cost   1153.4545   102.0915   1072.0000   1072.0000
1096.0000   1269.0000
##            volatile.acidity      0.2855     0.0826      0.1500      0.2400
0.2700      0.3400
##
## Table: Table continues below
##
##
##
##                                      Max       MAD        IQR       CV
```

```
Skewness   SE.Skewness
## ------------------------- ----------- --------- ---------- -------- -----
----- -------------
##                   alcohol     13.0000    0.8896     1.0500   0.0723    -
0.5310        0.6607
##           BillOfMaterials    891.0000   23.7216    84.0000   0.0718
0.6711        0.6607
##                 chlorides      0.0460    0.0119     0.0105   0.2705    -
0.1520        0.6607
##               citric.acid      0.4900    0.0741     0.0900   0.2129
0.4631        0.6607
##                   density      0.9970    0.0024     0.0029   0.0024
0.5779        0.6607
##             fixed.acidity      9.1000    0.4448     0.5500   0.1221
0.4174        0.6607
##        free.sulfur.dioxide     59.0000    2.9652     6.5000   0.3513
1.1274        0.6607
##                        pH      3.5500    0.1186     0.1250   0.0358
0.6205        0.6607
##                     Price   2754.0000   23.7216   229.5000   0.0536
0.5605        0.6607
##                    profit   1530.0000   11.8608    25.5000   0.0381
1.7791        0.6607
##                   quality      9.0000    0.0000     1.0000   0.0603
0.4914        0.6607
##            residual.sugar     14.8000    3.1135     4.6000   0.7703
0.9755        0.6607
##               StorageCost    441.0000   35.5824    77.5000   0.1314
0.4995        0.6607
##                 sulphates      0.7100    0.1186     0.1250   0.2276
0.6373        0.6607
##       total.sulfur.dioxide    177.0000   19.2738    24.5000   0.1802
0.5920        0.6607
##                total_cost   1332.0000   35.5824   170.5000   0.0885
0.6228        0.6607
##           volatile.acidity      0.4700    0.0445     0.0700   0.2895
0.6522        0.6607
##
## Table: Table continues below
##
##
##
##                             Kurtosis   N.Valid   Pct.Valid
## ------------------------- ---------- --------- -----------
##                   alcohol    -1.1877   11.0000    100.0000
##           BillOfMaterials    -1.3124   11.0000    100.0000
##                 chlorides    -1.2656   11.0000    100.0000
##               citric.acid    -1.2977   11.0000    100.0000
##                   density    -0.9944   11.0000    100.0000
##             fixed.acidity     0.4944   11.0000    100.0000
```

```
##        free.sulfur.dioxide    -0.3596   11.0000    100.0000
##                         pH    -0.3814   11.0000    100.0000
##                      Price    -1.7334   11.0000    100.0000
##                     profit     2.0356   11.0000    100.0000
##                    quality    -1.9079   11.0000    100.0000
##             residual.sugar    -0.3089   11.0000    100.0000
##                StorageCost    -1.6202   11.0000    100.0000
##                  sulphates    -0.9452   11.0000    100.0000
##       total.sulfur.dioxide    -0.7497   11.0000    100.0000
##                 total_cost    -1.5450   11.0000    100.0000
##           volatile.acidity     0.0270   11.0000    100.0000
```

We will now compute least profitable wine and analyze its summary and correlation

```
Min_Profit<-slice_min(winequality, n = 10, winequality$profit)
corr_var(Min_Profit,
         profit)
```




```
Min_profit_Wine_Summarry <- summarytools::descr(Min_Profit, round.digits = 4,
transpose = TRUE)
Min_profit_Wine_Summarry

## Non-numerical variable(s) ignored: type

## Descriptive Statistics
## Min_Profit
## N: 11
##
```

```
##                                  Mean   Std.Dev        Min         Q1
Median         Q3
## ------------------------- ---------- --------- ---------- ---------- ----
------ ----------
##                    alcohol    10.5045    1.0845     9.4000     9.7000
10.1000    11.0000
##               BillOfMaterials   309.0000   12.7984   282.0000   306.0000
309.0000   318.0000
##                  chlorides     0.0535    0.0342     0.0220     0.0340
0.0410     0.0610
##                citric.acid     0.3073    0.1476     0.0200     0.2600
0.3400     0.4200
##                    density     0.9947    0.0023     0.9911     0.9926
0.9949     0.9970
##               fixed.acidity     7.8636    1.5227     6.1000     6.7000
7.3000     9.4000
##          free.sulfur.dioxide    37.7727   43.5002     5.0000     5.0000
20.0000    42.0000
##                           pH     3.2291    0.2250     2.8900     3.0500
3.2400     3.3800
##                        Price   895.9091   11.8697   882.0000   885.0000
894.0000   906.0000
##                       profit   448.0909   12.2430   429.0000   438.0000
453.0000   459.0000
##                    quality     3.0000    0.0000     3.0000     3.0000
3.0000     3.0000
##                residual.sugar     4.4000    4.6447     1.1500     1.4000
1.8000     8.5000
##                  StorageCost   138.8182    6.5851   129.0000   132.0000
141.0000   144.0000
##                   sulphates     0.5018    0.1684     0.2800     0.3700
0.5200     0.6300
##         total.sulfur.dioxide   100.0909   80.7025    12.0000    33.0000
57.0000   201.0000
##                   total_cost   447.8182   13.3553   423.0000   435.0000
453.0000   456.0000
##             volatile.acidity     0.4091    0.2527     0.1700     0.2400
0.3300     0.4800
##
## Table: Table continues below
##
##
##
##                                     Max       MAD        IQR       CV
Skewness   SE.Skewness
## ------------------------- ---------- --------- ---------- -------- ------
---- -------------
##                    alcohol    12.6000    0.5930     0.9500   0.1032
0.9311        0.6607
##               BillOfMaterials   324.0000   13.3434    12.0000   0.0414    -
```

```
0.8149        0.6607
##                 chlorides     0.1450    0.0119     0.0250   0.6394
1.6305        0.6607
##                citric.acid     0.4700    0.1186     0.1350   0.4803    -
0.9207        0.6607
##                   density     0.9983    0.0032     0.0033   0.0024    -
0.0635        0.6607
##             fixed.acidity    10.4000    0.8896     2.1000   0.1936
0.6066        0.6607
##        free.sulfur.dioxide   124.0000   22.2390    34.5000   1.1516
1.1223        0.6607
##                        pH     3.5500    0.2076     0.2500   0.0697    -
0.0839        0.6607
##                     Price   918.0000   13.3434    18.0000   0.0132
0.3865        0.6607
##                    profit   462.0000   13.3434    19.5000   0.0273    -
0.2303        0.6607
##                   quality     3.0000    0.0000     0.0000   0.0000
NaN         0.6607
##             residual.sugar    15.1000    0.8896     5.1000   1.0556
1.1388        0.6607
##                StorageCost   147.0000    4.4478    10.5000   0.0474    -
0.3666        0.6607
##                  sulphates     0.8600    0.1631     0.2150   0.3356
0.5536        0.6607
##       total.sulfur.dioxide   216.0000   57.8214   141.0000   0.8063
0.3814        0.6607
##                total_cost   465.0000    8.8956    13.5000   0.0298    -
0.5784        0.6607
##           volatile.acidity     0.9800    0.1631     0.2150   0.6178
1.0937        0.6607
##
## Table: Table continues below
##
##
##
##                             Kurtosis   N.Valid   Pct.Valid
## -------------------------- ---------- --------- -----------
##                   alcohol    -0.6949   11.0000    100.0000
##           BillOfMaterials    -0.5524   11.0000    100.0000
##                 chlorides     1.7535   11.0000    100.0000
##               citric.acid    -0.6408   11.0000    100.0000
##                   density    -1.4081   11.0000    100.0000
##              fixed.acidity    -1.3458   11.0000    100.0000
##        free.sulfur.dioxide    -0.3980   11.0000    100.0000
##                        pH    -1.3014   11.0000    100.0000
##                     Price    -1.2874   11.0000    100.0000
##                    profit    -1.7066   11.0000    100.0000
##                   quality        NaN   11.0000    100.0000
##             residual.sugar    -0.1531   11.0000    100.0000
```

```
##                StorageCost    -1.6296   11.0000    100.0000
##                  sulphates    -0.6193   11.0000    100.0000
##       total.sulfur.dioxide    -1.7659   11.0000    100.0000
##                 total_cost    -1.0967   11.0000    100.0000
##           volatile.acidity    -0.1342   11.0000    100.0000
```

- The gross profit percent lies between 25% to 87%.
- The least profitable wines have a very high pH value.
- The factor affecting the price and profit most is quality.

Q5-> Is there any Correlation between different attributes of the wines?

cov() form the variance-covariance matrix calculate the covariance of the numeric values . **cor()** forms the correlation matrix. The `unlist()` function and `lapply()` is used to select only numeric values from the data set.

```
cov_matrix <- cov(winequality [,unlist(lapply(winequality, is.numeric))])
cov_matrix
```

```
##                       fixed.acidity volatile.acidity   citric.acid
## fixed.acidity           1.69897482     4.957306e-02  5.954645e-02
## volatile.acidity        0.04957306     2.434418e-02 -6.698600e-03
## citric.acid             0.05954645    -6.698600e-03  1.914868e-02
## residual.sugar         -0.73964774    -1.329299e-01  7.732452e-02
## chlorides               0.01390243     2.005147e-03  3.770861e-04
## free.sulfur.dioxide    -6.73606510    -9.328458e-01  2.435773e-01
## total.sulfur.dioxide  -25.08721097    -3.301192e+00  1.062766e+00
## density                 0.00180574     1.280681e-04  4.960217e-05
## pH                     -0.04998417     5.191455e-03 -6.080721e-03
## sulphates               0.06044022     5.252073e-03  1.781810e-03
## alcohol                -0.13705513    -7.674091e-03 -1.287149e-03
## quality                -0.09004539    -3.444308e-02  7.545150e-03
## BillOfMaterials        -9.09064056    -3.391435e+00  7.881829e-01
## StorageCost            -3.86223490    -1.553875e+00  3.682515e-01
## Price                 -26.81288839    -1.028053e+01  2.227130e+00
## total_cost            -12.95287546    -4.945309e+00  1.156434e+00
## profit                -13.86001293    -5.335219e+00  1.070696e+00
##                       residual.sugar     chlorides free.sulfur.dioxide
## fixed.acidity           -0.739647744  1.390243e-02        -6.736065098
## volatile.acidity        -0.132929945  2.005147e-03        -0.932845756
## citric.acid              0.077324521  3.770861e-04         0.243577310
## residual.sugar          22.873635383 -2.027486e-02        33.772684276
## chlorides               -0.020274857  1.225154e-03        -0.115392034
## free.sulfur.dioxide     33.772684276 -1.153920e-01       314.736406025
## total.sulfur.dioxide   130.471798887 -5.181609e-01       709.465024638
## density                  0.008106957  3.799576e-05         0.001777052
## pH                      -0.191205877  1.417604e-04        -0.344414831
## sulphates               -0.130199316  2.088218e-03        -0.494219145
## alcohol                 -2.081981113 -1.082318e-02        -3.896254187
```

```
## quality                -0.181695565 -5.942316e-03         0.699833423
## BillOfMaterials       -18.037715145 -5.967167e-01        68.083536222
## StorageCost            -8.254756966 -2.589798e-01        30.998340436
## Price                 -52.667278436 -1.777731e+00       208.659509848
## total_cost            -26.292472112 -8.556965e-01        99.081876658
## profit                -26.374806324 -9.220340e-01       109.577633190
##                      total.sulfur.dioxide       density            pH
## fixed.acidity              -2.508721e+01  1.805740e-03 -4.998417e-02
## volatile.acidity           -3.301192e+00  1.280681e-04  5.191455e-03
## citric.acid                 1.062766e+00  4.960217e-05 -6.080721e-03
## residual.sugar              1.304718e+02  8.106957e-03 -1.912059e-01
## chlorides                  -5.181609e-01  3.799576e-05  1.417604e-04
## free.sulfur.dioxide         7.094650e+02  1.777052e-03 -3.444148e-01
## total.sulfur.dioxide        3.107508e+03  7.526085e-03 -1.799926e+00
## density                     7.526085e-03  9.093365e-06  2.576445e-06
## pH                         -1.799926e+00  2.576445e-06  2.466035e-02
## sulphates                  -2.270694e+00  1.167202e-04  4.208743e-03
## alcohol                    -1.821183e+01 -2.477500e-03  2.147721e-02
## quality                    -2.946859e+00 -8.027195e-04  4.642167e-03
## BillOfMaterials            -2.787294e+02 -8.038253e-02  4.518977e-01
## StorageCost                -1.373005e+02 -3.552964e-02  1.892295e-01
## Price                      -8.910566e+02 -2.385777e-01  1.391318e+00
## total_cost                 -4.160299e+02 -1.159122e-01  6.411272e-01
## profit                     -4.750268e+02 -1.226655e-01  7.501912e-01
##                          sulphates       alcohol       quality
BillOfMaterials
## fixed.acidity         0.0604402196  -0.137055134 -9.004539e-02   -
9.090641e+00
## volatile.acidity      0.0052520729  -0.007674091 -3.444308e-02   -
3.391435e+00
## citric.acid           0.0017818100  -0.001287149  7.545150e-03
7.881829e-01
## residual.sugar       -0.1301993163  -2.081981113 -1.816956e-01   -
1.803772e+01
## chlorides             0.0020882177  -0.010823184 -5.942316e-03   -
5.967167e-01
## free.sulfur.dioxide  -0.4942191455  -3.896254187  6.998334e-01
6.808354e+01
## total.sulfur.dioxide -2.2706942576 -18.211831753 -2.946859e+00   -
2.787294e+02
## density               0.0001167202  -0.002477500 -8.027195e-04   -
8.038253e-02
## pH                    0.0042087430   0.021477214  4.642167e-03
4.518977e-01
## sulphates             0.0222526272  -0.001523289  4.782495e-03
4.859461e-01
## alcohol              -0.0015232893   1.423372757  4.664907e-01
4.705566e+01
## quality               0.0047824950   0.466490711  7.601806e-01
7.586371e+01
```

```
## BillOfMaterials       0.4859461038  47.055656298  7.586371e+01
8.400039e+03
## StorageCost           0.2960992401  20.692747672  3.415859e+01
3.412356e+03
## Price                 1.4356414399 138.981230489  2.277715e+02
2.275242e+04
## total_cost            0.7820453439  67.748403969  1.100223e+02
1.181240e+04
## profit                0.6535960959  71.232826519  1.177492e+02
1.094002e+04
##                        StorageCost         Price    total_cost
profit
## fixed.acidity        -3.862235e+00   -26.8128884   -12.9528755   -
13.8600129
## volatile.acidity     -1.553875e+00   -10.2805280    -4.9453091    -
5.3352189
## citric.acid           3.682515e-01     2.2271304     1.1564344
1.0706960
## residual.sugar       -8.254757e+00   -52.6672784   -26.2924721   -
26.3748063
## chlorides            -2.589798e-01    -1.7777305    -0.8556965    -
0.9220340
## free.sulfur.dioxide   3.099834e+01   208.6595098    99.0818767
109.5776332
## total.sulfur.dioxide -1.373005e+02  -891.0566147  -416.0298555  -
475.0267591
## density              -3.552964e-02    -0.2385777    -0.1159122    -
0.1226655
## pH                    1.892295e-01     1.3913183     0.6411272
0.7501912
## sulphates             2.960992e-01     1.4356414     0.7820453
0.6535961
## alcohol               2.069275e+01   138.9812305    67.7484040
71.2328265
## quality               3.415859e+01   227.7715459   110.0223035
117.7492424
## BillOfMaterials       3.412356e+03 22752.4163044 11812.3953229
10940.0209815
## StorageCost           1.888673e+03 10242.7541834  5301.0290743
4941.7251090
## Price                 1.024275e+04 68737.1145200 32995.1704877
35741.9440323
## total_cost            5.301029e+03 32995.1704877 17113.4243972
15881.7460905
## profit                4.941725e+03 35741.9440323 15881.7460905
19860.1979417

cor_matrix <- cor(winequality[,unlist(lapply(winequality, is.numeric))])
cor_matrix
```

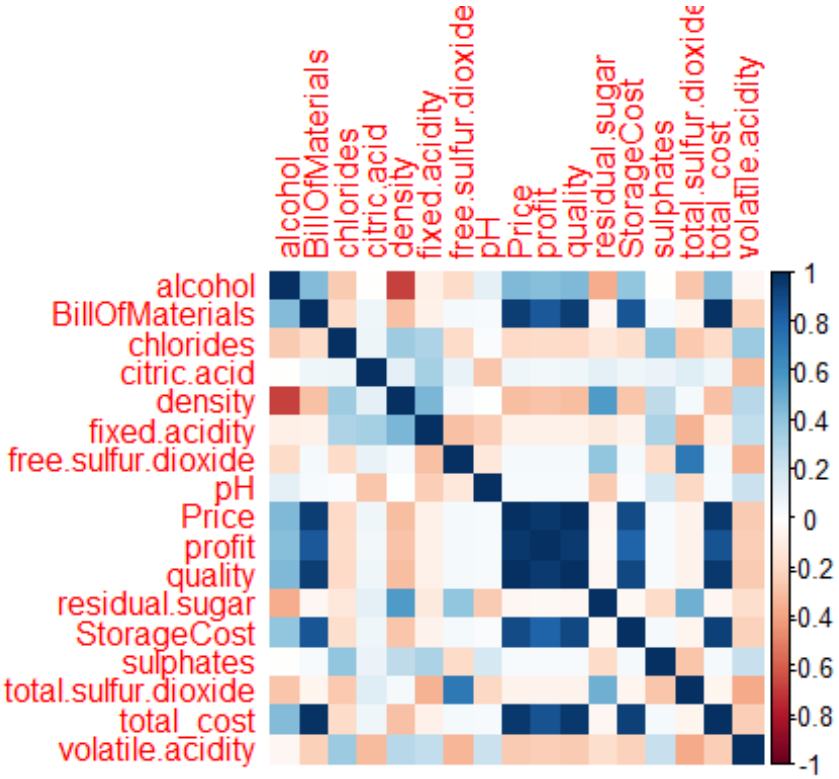
```
##                       fixed.acidity volatile.acidity  citric.acid
residual.sugar
## fixed.acidity           1.00000000       0.24375563  0.330136077    -
0.11864894
## volatile.acidity        0.24375563       1.00000000 -0.310253791    -
0.17813844
## citric.acid             0.33013608      -0.31025379  1.000000000
0.11683696
## residual.sugar         -0.11864894      -0.17813844  0.116836960
1.00000000
## chlorides               0.30472068       0.36715837  0.077853074    -
0.12111415
## free.sulfur.dioxide    -0.29129921      -0.33700689  0.099218794
0.39803796
## total.sulfur.dioxide   -0.34526516      -0.37954828  0.137772421
0.48937627
## density                 0.45940885       0.27219553  0.118869088
0.56211864
## pH                     -0.24419647       0.21188097 -0.279824715    -
0.25458571
## sulphates               0.31084359       0.22565373  0.086317996    -
0.18249479
## alcohol                -0.08813380      -0.04122589 -0.007796505    -
0.36487984
## quality                -0.07923365      -0.25318969  0.062537438    -
0.04357309
## BillOfMaterials        -0.07609574      -0.23716206  0.062146539    -
0.04115033
## StorageCost            -0.06818155      -0.22916066  0.061234594    -
0.03971535
## Price                  -0.07846112      -0.25131719  0.061387592    -
0.04200274
## total_cost             -0.07596338      -0.24228549  0.063882683    -
0.04202380
## profit                 -0.07545332      -0.24264027  0.054904127    -
0.03913181
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.30472068         -0.29129921          -0.34526516
## volatile.acidity      0.36715837         -0.33700689          -0.37954828
## citric.acid           0.07785307          0.09921879           0.13777242
## residual.sugar       -0.12111415          0.39803796           0.48937627
## chlorides             1.00000000         -0.18582624          -0.26556033
## free.sulfur.dioxide  -0.18582624          1.00000000           0.71738345
## total.sulfur.dioxide -0.26556033          0.71738345           1.00000000
## density               0.35997931          0.03321729           0.04477140
## pH                    0.02579051         -0.12362569          -0.20561220
## sulphates             0.39993538         -0.18674777          -0.27306238
## alcohol              -0.25917936         -0.18408333          -0.27383450
## quality              -0.19471617          0.04524419          -0.06063101
## BillOfMaterials      -0.18600827          0.04187241          -0.05455517
```

```
## StorageCost          -0.17025200          0.04020561          -0.05667449
## Price                -0.19372023          0.04486101          -0.06096822
## total_cost           -0.18687706          0.04269257          -0.05704924
## profit               -0.18692175          0.04382850          -0.06046724
##                          density          pH    sulphates      alcohol
## fixed.acidity         0.459408850 -0.244196473  0.310843590 -0.088133803
## volatile.acidity      0.272195529  0.211880965  0.225653733 -0.041225893
## citric.acid           0.118869088 -0.279824715  0.086317996 -0.007796505
## residual.sugar        0.562118636 -0.254585710 -0.182494786 -0.364879836
## chlorides             0.359979307  0.025790508  0.399935382 -0.259179362
## free.sulfur.dioxide   0.033217294 -0.123625692 -0.186747772 -0.184083332
## total.sulfur.dioxide  0.044771399 -0.205612203 -0.273062377 -0.273834495
## density               1.000000000  0.005440751  0.259473546 -0.688639523
## pH                    0.005440751  1.000000000  0.179664448  0.114635464
## sulphates             0.259473546  0.179664448  1.000000000 -0.008559185
## alcohol              -0.688639523  0.114635464 -0.008559185  1.000000000
## quality              -0.305311451  0.033904923  0.036771004  0.448461620
## BillOfMaterials      -0.290843060  0.031397846  0.035543223  0.430340146
## StorageCost          -0.271112865  0.027727506  0.045673922  0.399098869
## Price                -0.301767103  0.033793350  0.036707913  0.444325440
## total_cost           -0.293831431  0.031208730  0.040074935  0.434081506
## profit               -0.288647944  0.033898505  0.031090451  0.423671321
##                         quality BillOfMaterials StorageCost       Price
## fixed.acidity        -0.07923365     -0.07609574 -0.06818155 -0.07846112
## volatile.acidity     -0.25318969     -0.23716206 -0.22916066 -0.25131719
## citric.acid           0.06253744      0.06214654  0.06123459  0.06138759
## residual.sugar       -0.04357309     -0.04115033 -0.03971535 -0.04200274
## chlorides            -0.19471617     -0.18600827 -0.17025200 -0.19372023
## free.sulfur.dioxide   0.04524419      0.04187241  0.04020561  0.04486101
## total.sulfur.dioxide -0.06063101     -0.05455517 -0.05667449 -0.06096822
## density              -0.30531145     -0.29084306 -0.27111287 -0.30176710
## pH                    0.03390492      0.03139785  0.02772751  0.03379335
## sulphates             0.03677100      0.03554322  0.04567392  0.03670791
## alcohol               0.44846162      0.43034015  0.39909887  0.44432544
## quality               1.00000000      0.94936903  0.90149473  0.99642686
## BillOfMaterials       0.94936903      1.00000000  0.85671310  0.94687190
## StorageCost           0.90149473      0.85671310  1.00000000  0.89896495
## Price                 0.99642686      0.94687190  0.89896495  1.00000000
## total_cost            0.96461471      0.98521012  0.93242388  0.96202480
## profit                0.95831424      0.84700451  0.80687908  0.96736581
##                       total_cost      profit
## fixed.acidity        -0.07596338 -0.07545332
## volatile.acidity     -0.24228549 -0.24264027
## citric.acid           0.06388268  0.05490413
## residual.sugar       -0.04202380 -0.03913181
## chlorides            -0.18687706 -0.18692175
## free.sulfur.dioxide   0.04269257  0.04382850
## total.sulfur.dioxide -0.05704924 -0.06046724
## density              -0.29383143 -0.28864794
## pH                    0.03120873  0.03389851
```

```
## sulphates              0.04007494  0.03109045
## alcohol                0.43408151  0.42367132
## quality                0.96461471  0.95831424
## BillOfMaterials        0.98521012  0.84700451
## StorageCost            0.93242388  0.80687908
## Price                  0.96202480  0.96736581
## total_cost             1.00000000  0.86146572
## profit                 0.86146572  1.00000000
```

To plot the correlation we will use `corrplot()` library.

```
library(corrplot)

## corrplot 0.92 loaded

corrplot(cor_matrix, method = 'color', order = 'alphabet')
```



As per the analysis we can concluded that the alcohol is inversely correlated to density. To dig deep into it we will only focus on correlation matrix of alcohol.

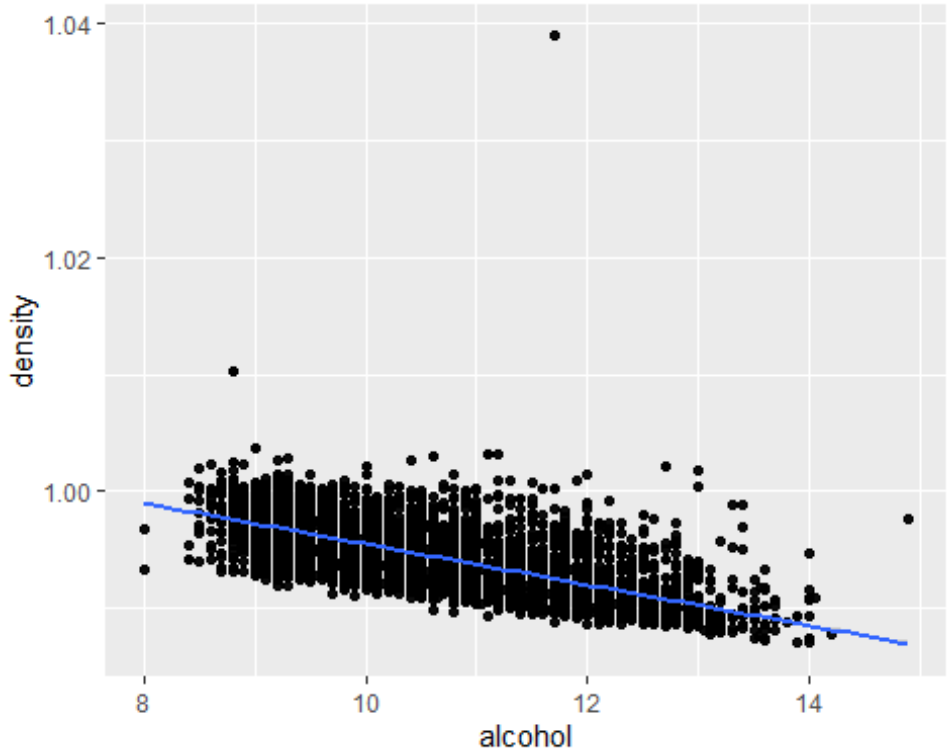
The `lm()` function is used to fit linear models to data frames. After doing so we will also summarize it.

```
alcohoal_density = lm(winequality$alcohol~winequality$density)
summary(alcohoal_density)

##
## Call:
```

```
## lm(formula = winequality$alcohol ~ winequality$density)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8613 -0.6030 -0.1028  0.5154 13.2788
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           281.493      3.593   78.34   <2e-16 ***
## winequality$density -272.451      3.613  -75.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8652 on 6306 degrees of freedom
## Multiple R-squared:  0.4742, Adjusted R-squared:  0.4741
## F-statistic:  5688 on 1 and 6306 DF,  p-value: < 2.2e-16
```

Also we will plot a linear distribution graph to understand the distribution

```
ggplot(winequality, aes(x = alcohol, y = density)) + geom_point() +
geom_smooth(method = "lm")

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



- The more the Alcohol contents the lesser the density of the alcohol
- There are few outlines in density column