# Inferential - Statistics

*Pratik Gandhi*

*May 16, 2016*

## Loading and reading the data:

```
library(tidyr)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.2.4
```

```
basedir <- "C:/Users/Pratik Gandhi/Documents/Data Science Stuff/Projects/Wine_Quality_Data_Set"
setwd(basedir)

# Loading the dataset
wh_wine <- read.csv("winequality-white.csv", header = TRUE)

# Looking at few observations of the data:
head(wh_wine, n = 5)
```

```
##   fixed.acidity.volatile.acidity.citric.acid.residual.sugar.chlorides.free.sulfur.dioxide.total.sulfu
## 1                                                                                                     7
## 2                                                                                                   6.3
## 3                                                                                                  8.1;(
## 4                                                                                                 7.2;0
## 5                                                                                                 7.2;0
```

## Data Munging Part:

### Few of the following things to be done:

1. Renaming the column names.
2. Adding more variables if necessary:

```
colnames(wh_wine) <- c("all_data")
wh_wine <- wh_wine %>% separate(all_data, c("fixed.acidity", "volatile.acidity",
    "citric.acid", "residual.sugar", "chlorides", "free.sulfur.dioxide", "total.sulfur.dioxide",
    "density", "pH", "sulphates", "alcohol", "quality"), sep = ";")

# Checking the dimension of the data:
dim(wh_wine)
```

```
## [1] 4898   12
```

```
# Checking the class of the variables:
str(wh_wine)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : chr  "7" "6.3" "8.1" "7.2" ...
```

```
##  $ volatile.acidity    : chr  "0.27" "0.3" "0.28" "0.23" ...
##  $ citric.acid         : chr  "0.36" "0.34" "0.4" "0.32" ...
##  $ residual.sugar      : chr  "20.7" "1.6" "6.9" "8.5" ...
##  $ chlorides           : chr  "0.045" "0.049" "0.05" "0.058" ...
##  $ free.sulfur.dioxide : chr  "45" "14" "30" "47" ...
##  $ total.sulfur.dioxide: chr  "170" "132" "97" "186" ...
##  $ density             : chr  "1.001" "0.994" "0.9951" "0.9956" ...
##  $ pH                  : chr  "3" "3.3" "3.26" "3.19" ...
##  $ sulphates           : chr  "0.45" "0.49" "0.44" "0.4" ...
##  $ alcohol             : chr  "8.8" "9.5" "10.1" "9.9" ...
##  $ quality             : chr  "6" "6" "6" "6" ...
```

- Now we have complete "tidy" dataset. The next part would be exploring the data and looking at each of them individually.
- There are 4898 observations and 12 variables.
- Looking at the variables it seems we need to change the class type of each of them if needed.
- Quality is a discrete variable while the others are continuous variables.
- Fixed acidity,volatile acidity and citric acid are related to pH.
- Free sulphur dioxide is contributing/related to total sulphur dioxide.

```
# Changing the class type and adding more levels
wh_wine$quality <- as.integer(wh_wine$quality)

# Changing the class type of all variables except 'Identifier' and 'Quality'
# variables:
remain_vars <- setdiff(names(wh_wine), c("quality"))
wh_wine[remain_vars] <- sapply(wh_wine[remain_vars], as.numeric)
str(wh_wine)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ 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 ...
```

```
# Summarizing the dataset:
summary(wh_wine)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
```

```
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

```
# Quality of wines:
table(wh_wine$quality)
```

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

**Initial investigation and observations from the data:**

- Less than 25% of wines have pH value less than or equal to 3.09. Most of the wines have pH value between 3-4. In other words most of the wines are very acidic.
- The mean residual sugar of wine is 6.391 g/l but the maximum is 65.8 which is clearly an outlier.
- Similarly, the free sulfur dioxide has mean of 35.31 ppm with almost 75% values having values less than or 46 ppm. The maximum value of 289 ppm is also high compared to these values and an outlier.
- Most of the wines are in the quality range of 5-7. The mean is 5.8 and the highest quality of wine is 9. There is no wine with 1,2 or 10.
- More than 75% (may be more less) of wines have the total sulphur dioxide content below 108 ppm. It is mentioned that at concentrations above 50 ppm SO2 becomes evident in taste. It would be interesting to see if it holds any correlation with the quality of wines.
- Interestingly, citric acid has minimum as 0. It would be really interesting to see if these values are missing, or not reported.

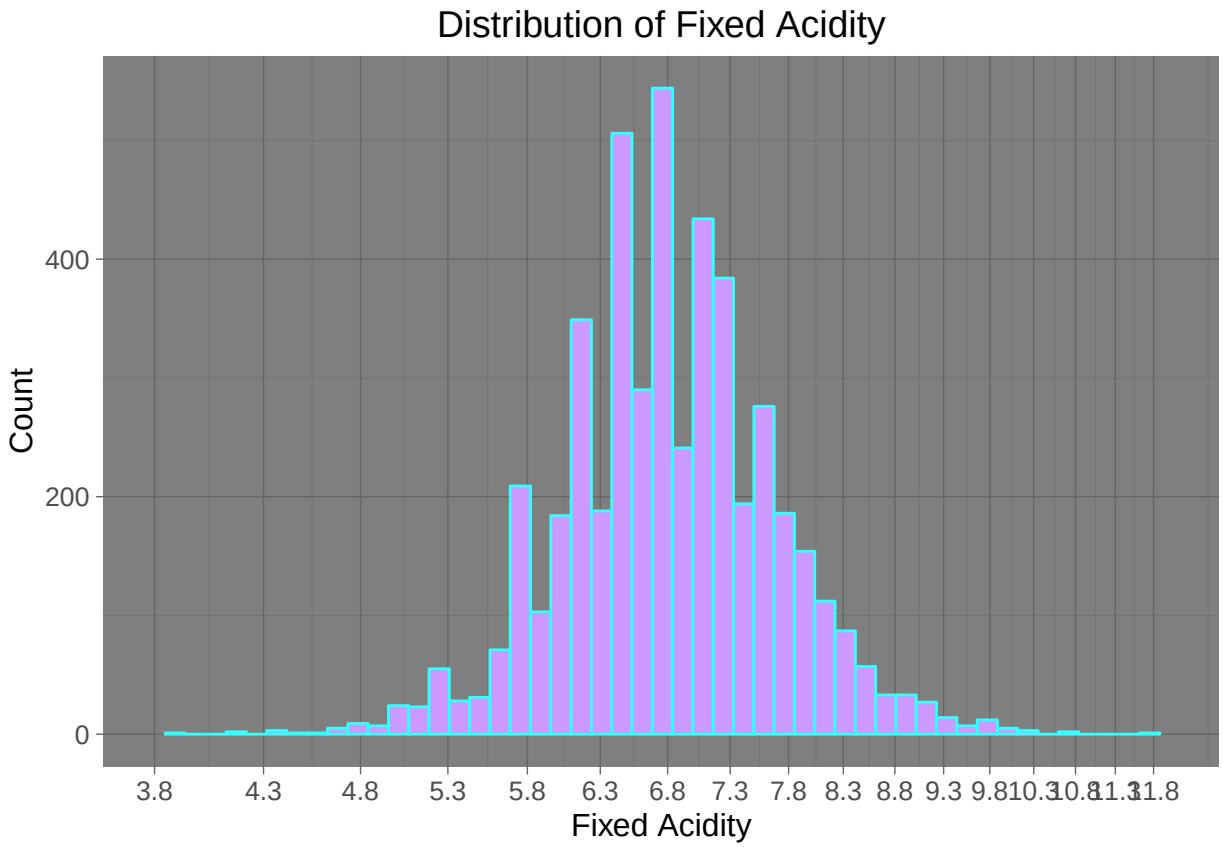
## Univariate Exploration and Analysis:

```
# ggpairs(wh_wine)

# Fixed Acidity:
ggplot(wh_wine, aes(fixed.acidity)) + geom_histogram(binwidth = 0.01, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Distribution of Fixed Acidity") + scale_x_log10(breaks = seq(min(wh_wi
    12, 0.5), lim = c(min(wh_wine$fixed.acidity), 12)) + xlab("Fixed Acidity") +
    ylab("Count") + theme_dark()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

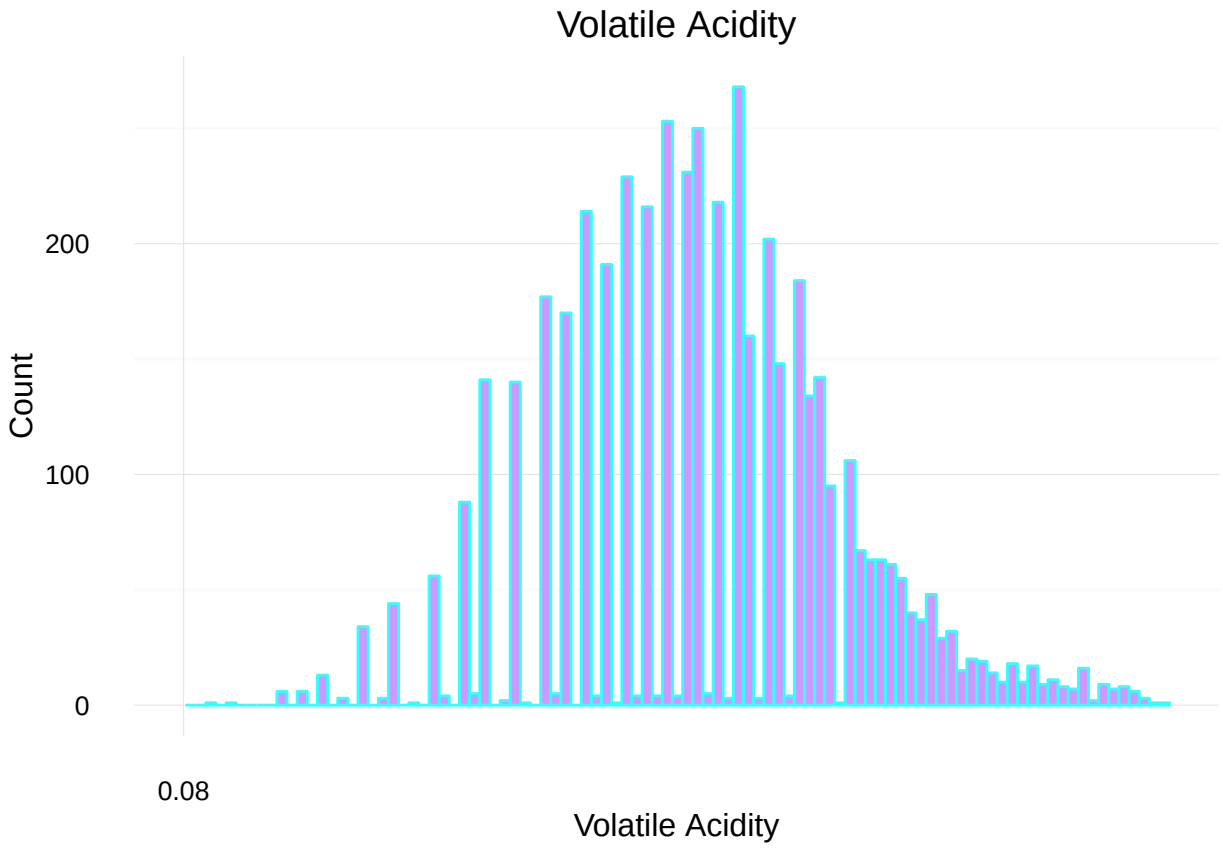



**The distribution of fixed acidity looks pretty normal. There might be couple of outliers on both tails.**

```
# Volatile Acidity:
ggplot(wh_wine, aes(volatile.acidity)) + geom_histogram(binwidth = 0.01, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Volatile Acidity") + scale_x_log10(breaks = seq(min(wh_wine$volatile.a
    0.75), lim = c(min(wh_wine$volatile.acidity), 0.75)) + xlab("Volatile Acidity") +
    ylab("Count") + theme_minimal()
```

```
## Warning: Removed 12 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

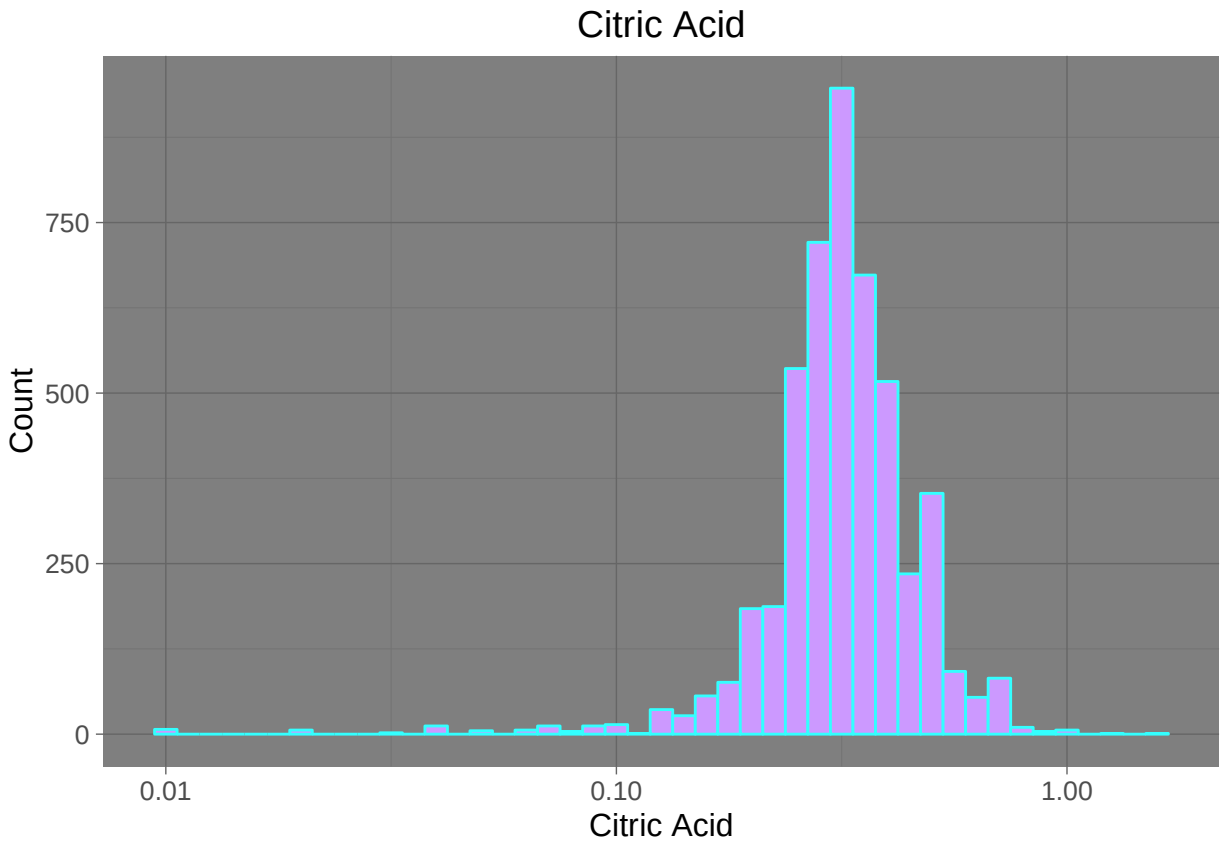



**Similar to Fixed Acidity this distribution also looks normal.**

## Need to make it more proper

```
# Citric Acid:
ggplot(wh_wine, aes(citric.acid)) + geom_histogram(binwidth = 0.05, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Citric Acid") + scale_x_log10() + xlab("Citric Acid") +
    ylab("Count") + theme_dark()
```

```
## Warning: Removed 19 rows containing non-finite values (stat_bin).
```




**Fixed acidity and volatile acidity look normally distruted on log10 scale while citric acid does not. As mentioned before, these behaviour of citric acid is because it has some missing values or non reported values.**

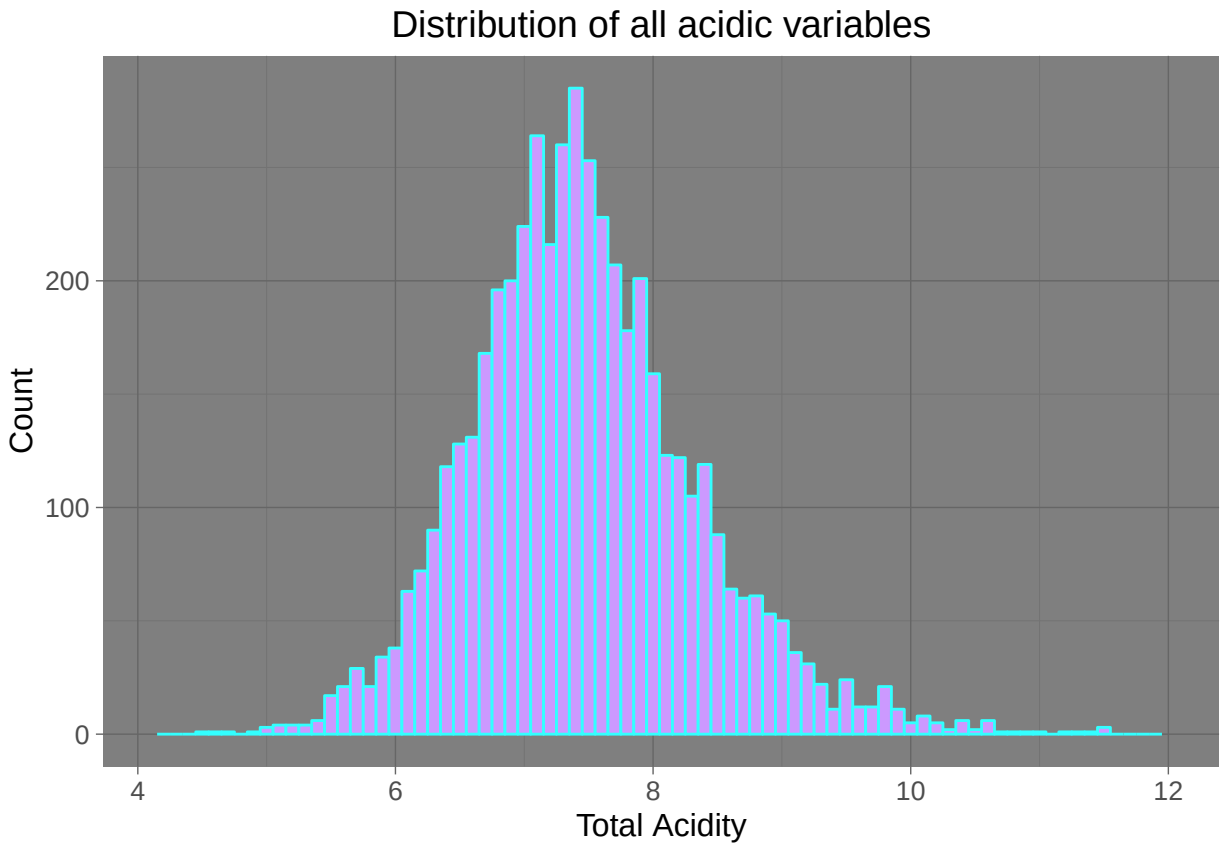
**Here log10 scale gives better results because pH is calculated on log10 and these components together contribute to the final pH value.**

```
## Combining them together we get:
wh_wine$tot.acidity = wh_wine$fixed.acidity + wh_wine$volatile.acidity + wh_wine$citric.acid

# Plotting them together:
ggplot(wh_wine, aes(tot.acidity)) + geom_histogram(binwidth = 0.1, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Distribution of all acidic variables") + xlab("Total Acidity") +
    ylab("Count") + xlim(min(wh_wine$tot.acidity), 12) + theme_dark()
```

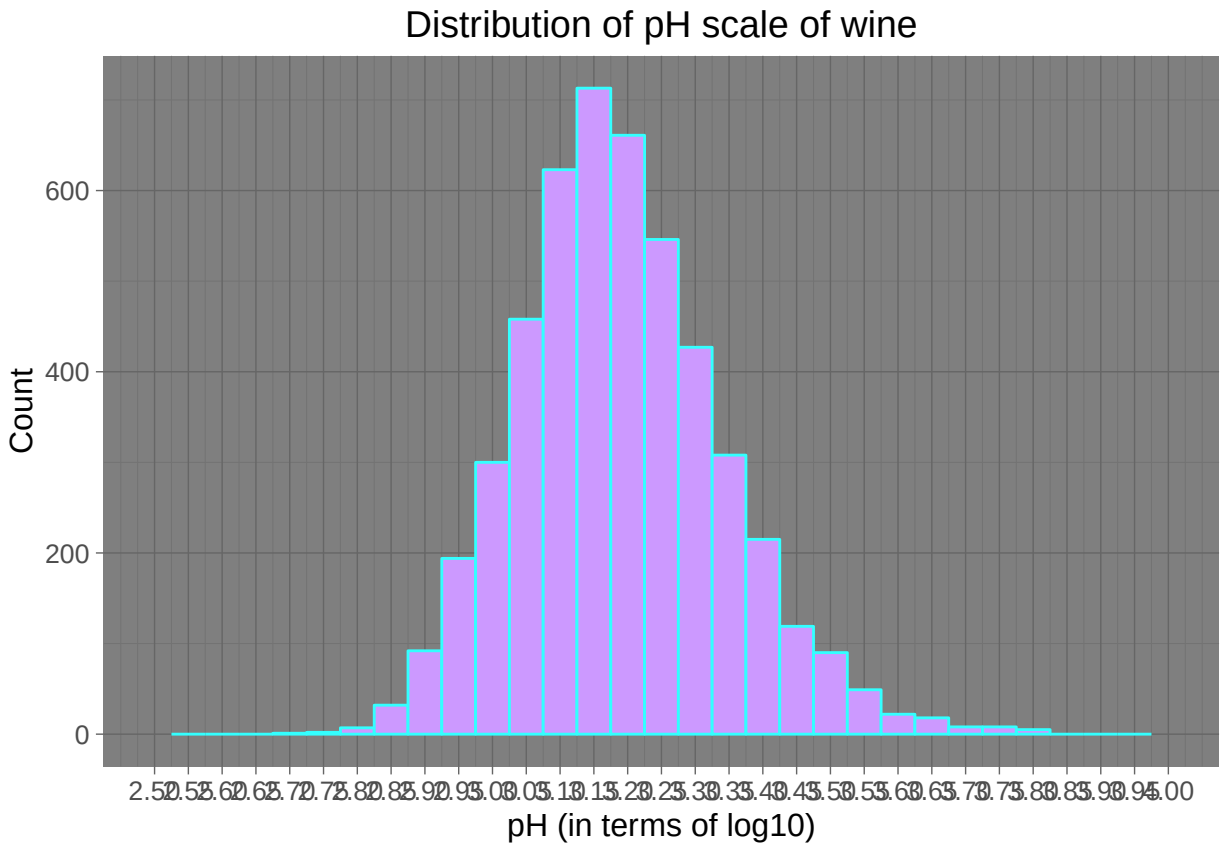
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```




**We get a normal distribution on combining them all together with few outliers.**

```
# pH:
ggplot(wh_wine, aes(pH)) + geom_histogram(binwidth = 0.05, color = "#33FFFF", fill = "#CC99FF") +
    ggtitle("Distribution of pH scale of wine") + xlab("pH (in terms of log10)") +
    ylab("Count") + scale_x_continuous(breaks = seq(2.5, 4, 0.05), lim = c(2.5, 4)) +
    theme_dark()
```

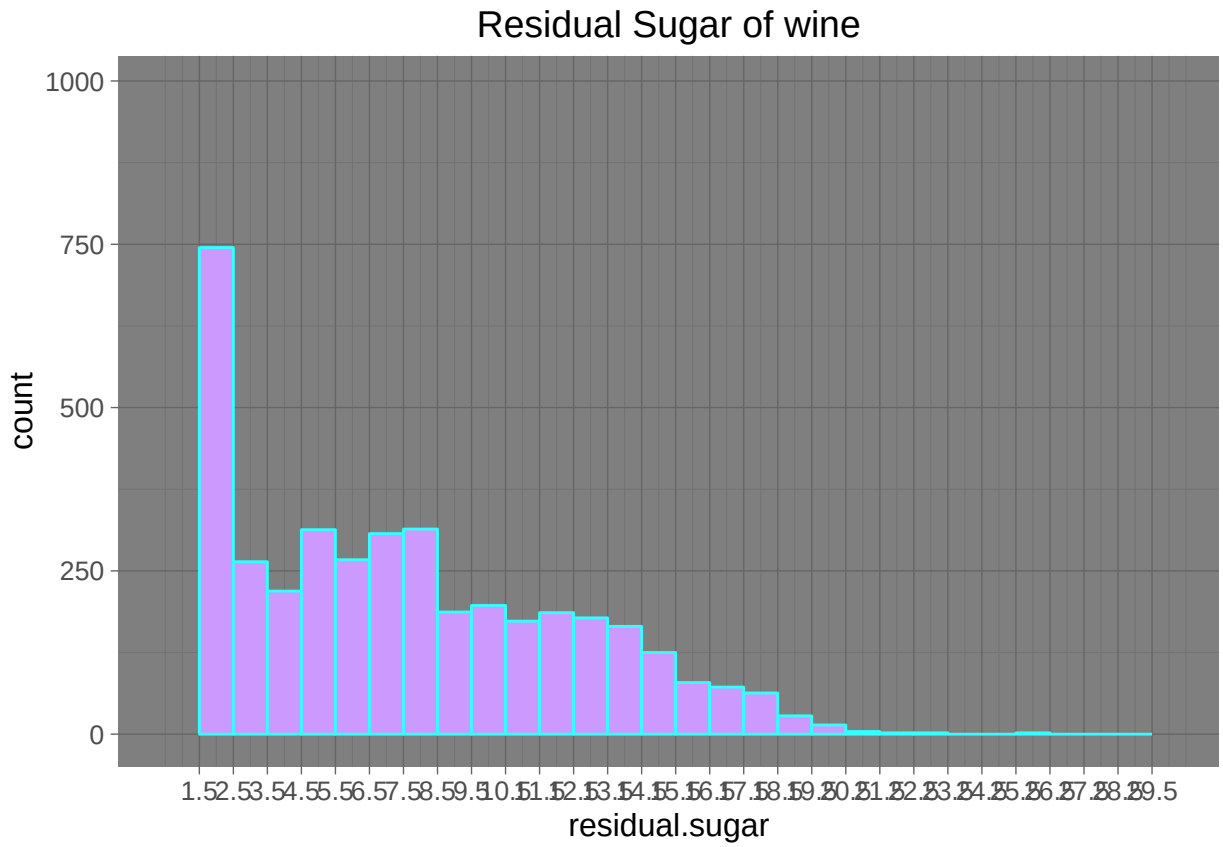



**The distribution is pretty normal with few outliers.**

**Here the term "acidity" refers to the fresh, tart and sour attributes of the wine and it is important component for wines.**

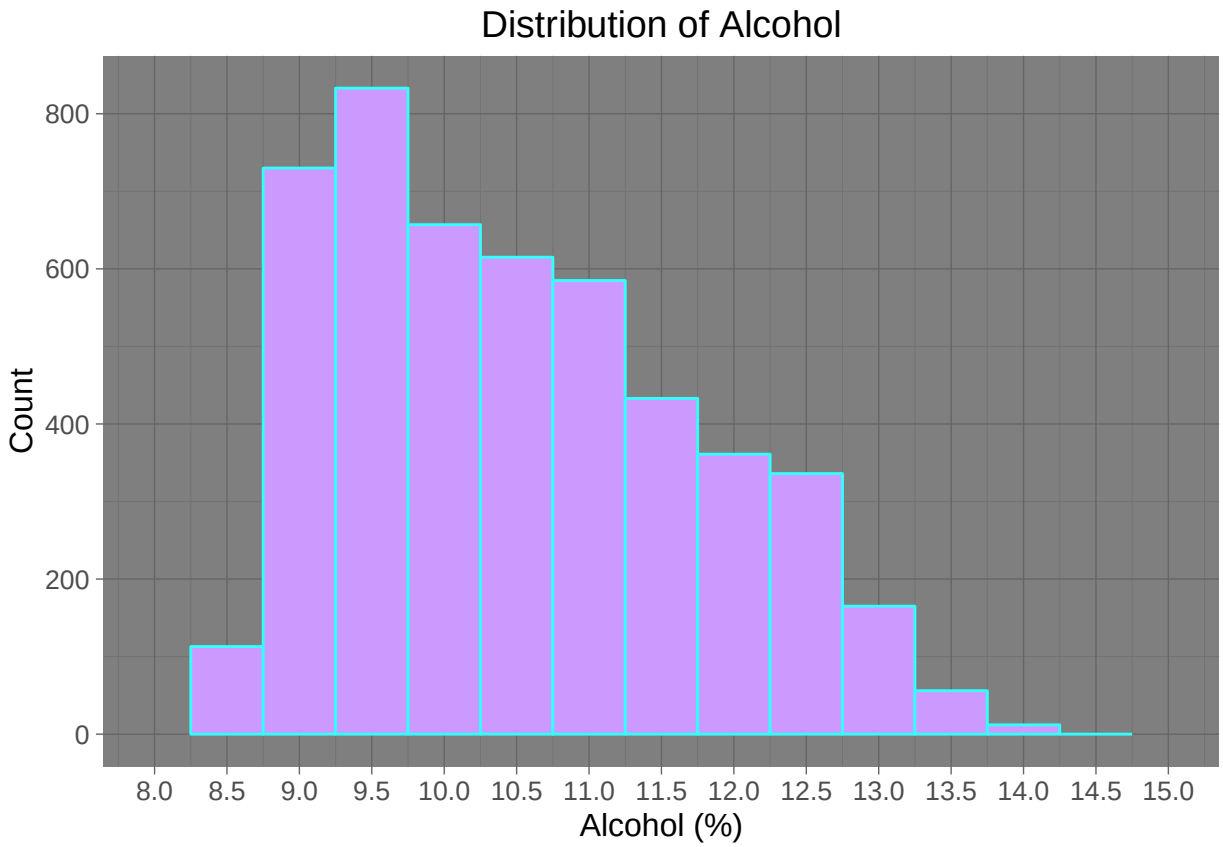
```
# Residual Sugar:
ggplot(wh_wine, aes(residual.sugar)) + geom_histogram(binwidth = 1, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Residual Sugar of wine") + scale_x_continuous(breaks = seq(1.5,
    30, 1), lim = c(min(wh_wine$residual.sugar), 30)) + theme_dark()
```

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




**The distribution is not normal and is positively skewed.**

```
# Levels of Alcohol:
ggplot(wh_wine, aes(alcohol)) + geom_histogram(binwidth = 0.5, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Distribution of Alcohol") + xlab("Alcohol (%)") +
    ylab("Count") + scale_x_continuous(breaks = seq(8, 15, 0.5), lim = c(8, 15)) +
    theme_dark()
```



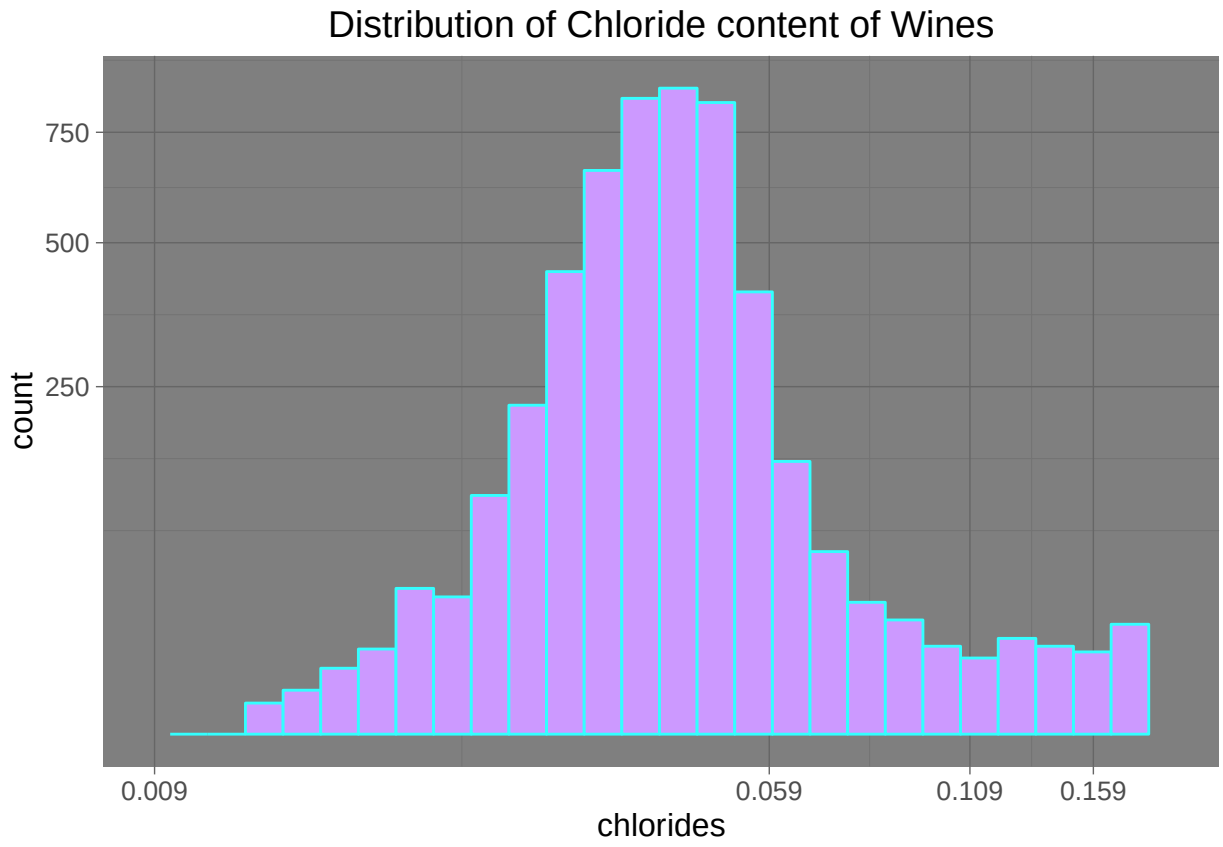

**The alcohol content for most of the wines is between 9-10%.**

**The distribution has long tail and is positively skewed as well.**

```
# Chlorides:
ggplot(wh_wine, aes(chlorides)) + geom_histogram(binwidth = 0.05, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Distribution of Chloride content of Wines") + scale_x_log10(breaks = s
    0.2, 0.05), lim = c(min(wh_wine$chlorides), 0.2)) + scale_y_sqrt() + theme_dark()
```

```
## Warning: Removed 17 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

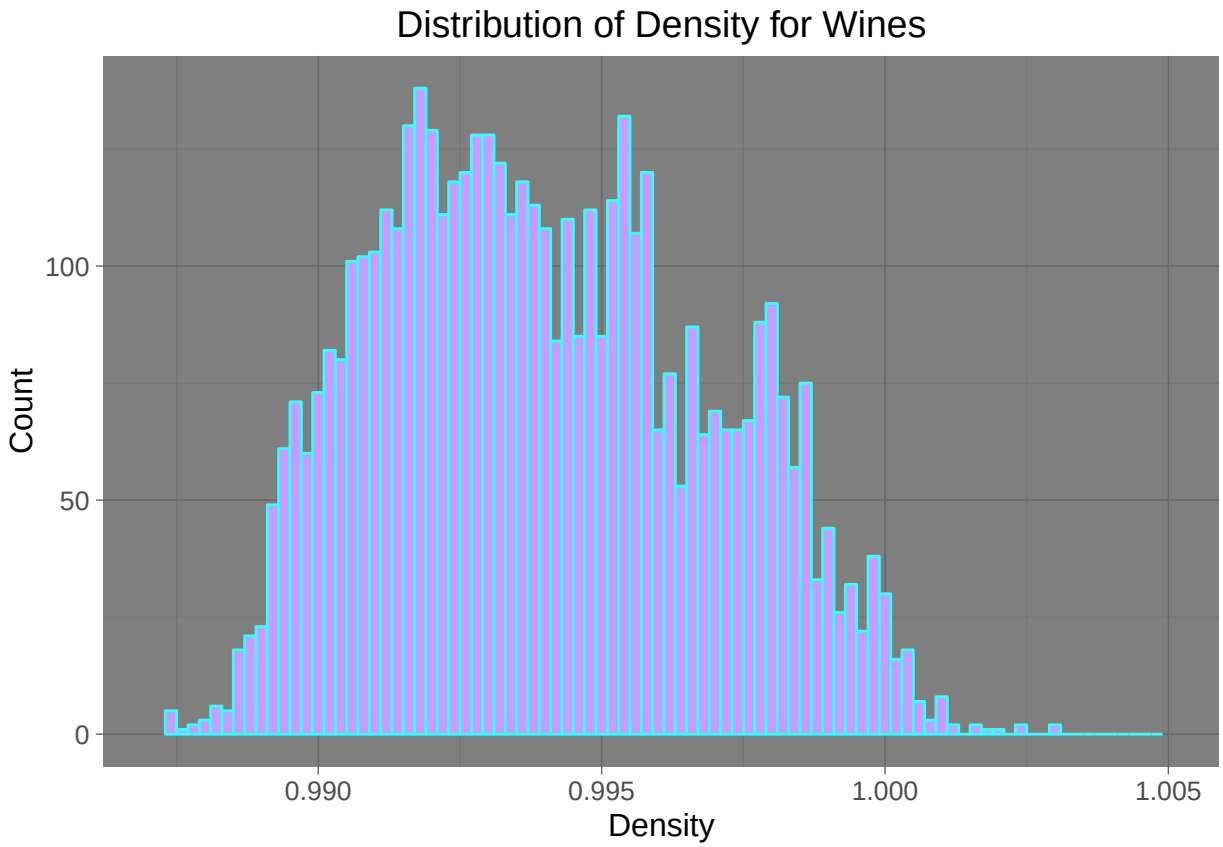



**The distribution looks normal but it has some extreme outliers.**

```
# Density:
ggplot(wh_wine, aes(density)) + geom_histogram(binwidth = 2e-04, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Distribution of Density for Wines") + xlab("Density") +
    ylab("Count") + scale_x_continuous(lim = c(min(wh_wine$density), 1.005)) + theme_dark()
```

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

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



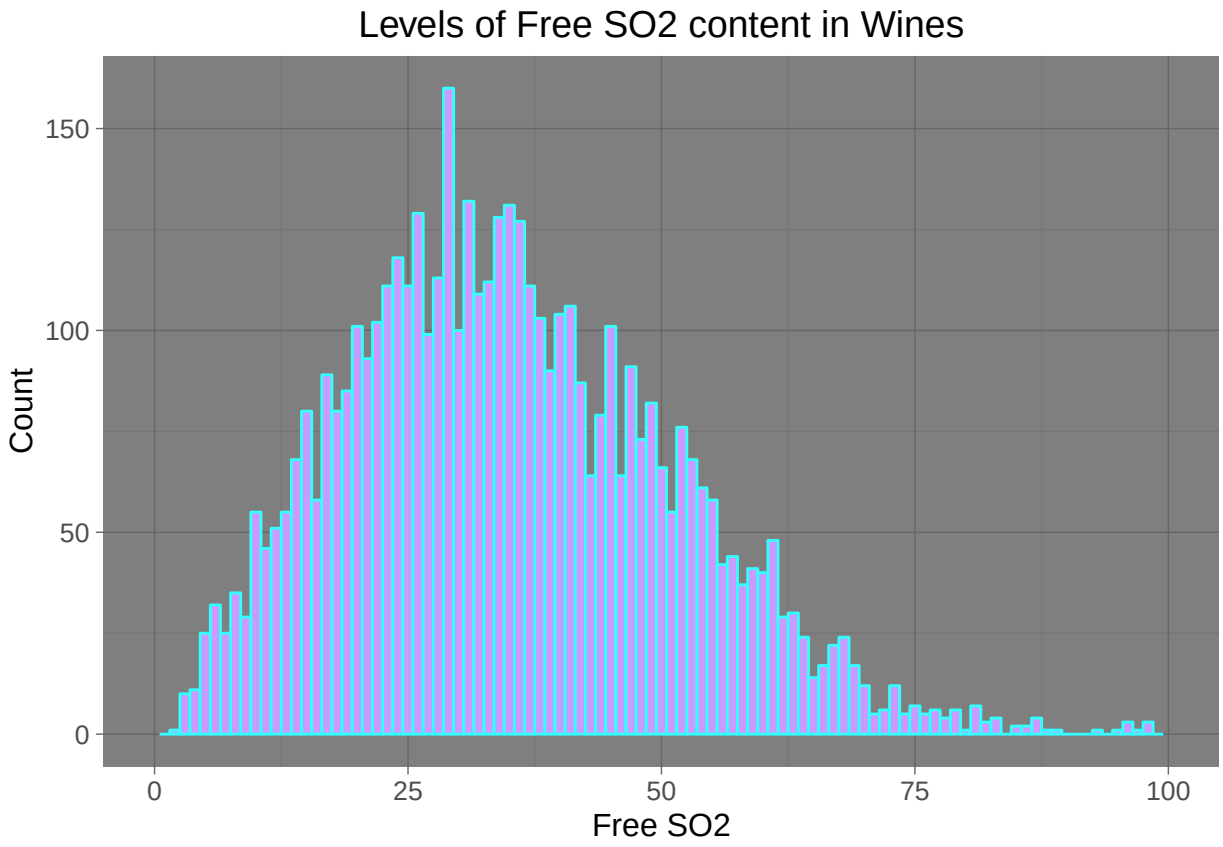

**The distribution for density is normal.**

**The density of wines is between 0.98 and 1**

**Even it has pretty normal distribution with few outliers.**

```
# Free Sulphur Dioxide
ggplot(wh_wine, aes(free.sulfur.dioxide)) + geom_histogram(binwidth = 1, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Levels of Free SO2 content in Wines") + xlab("Free SO2") +
    ylab("Count") + xlim(0, 100) + theme_dark()
```

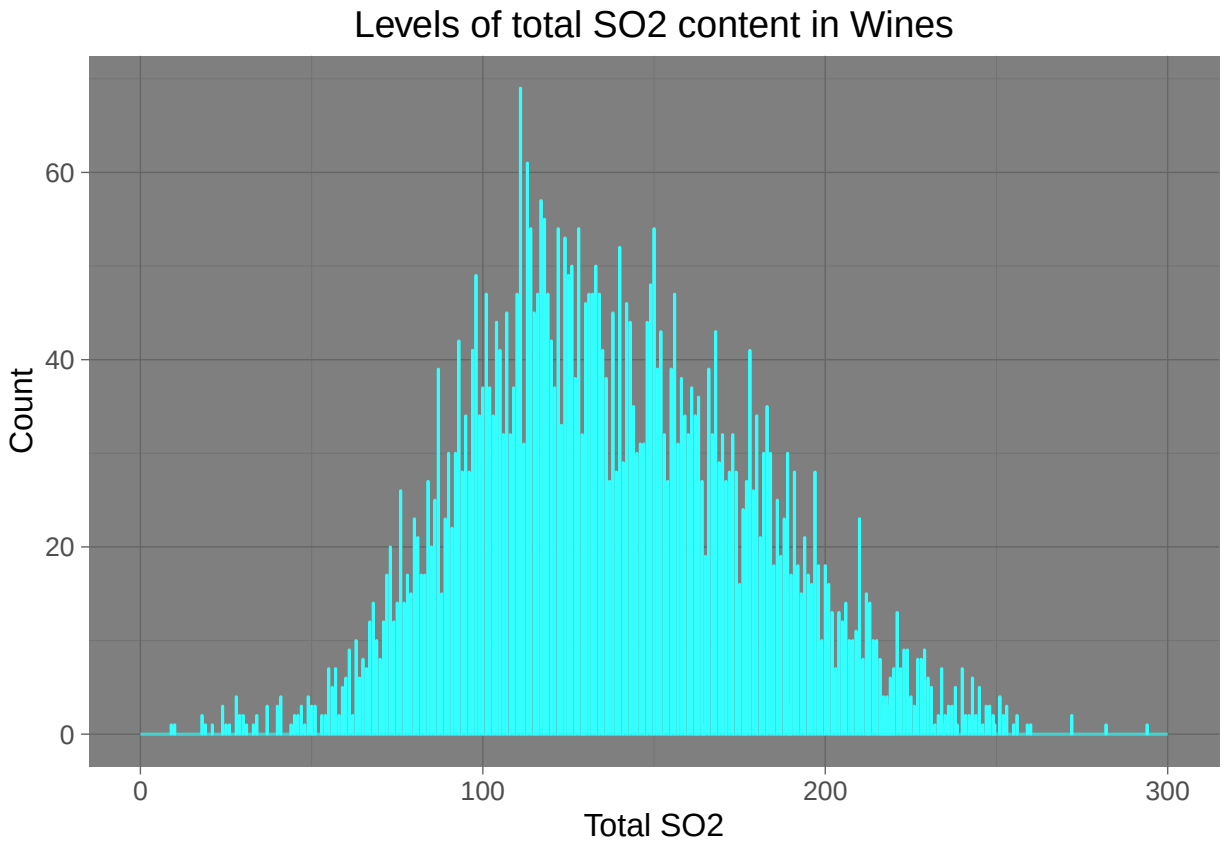
```
## Warning: Removed 17 rows containing non-finite values (stat_bin).
```




**The distribution is normal with a lot of outliers.**

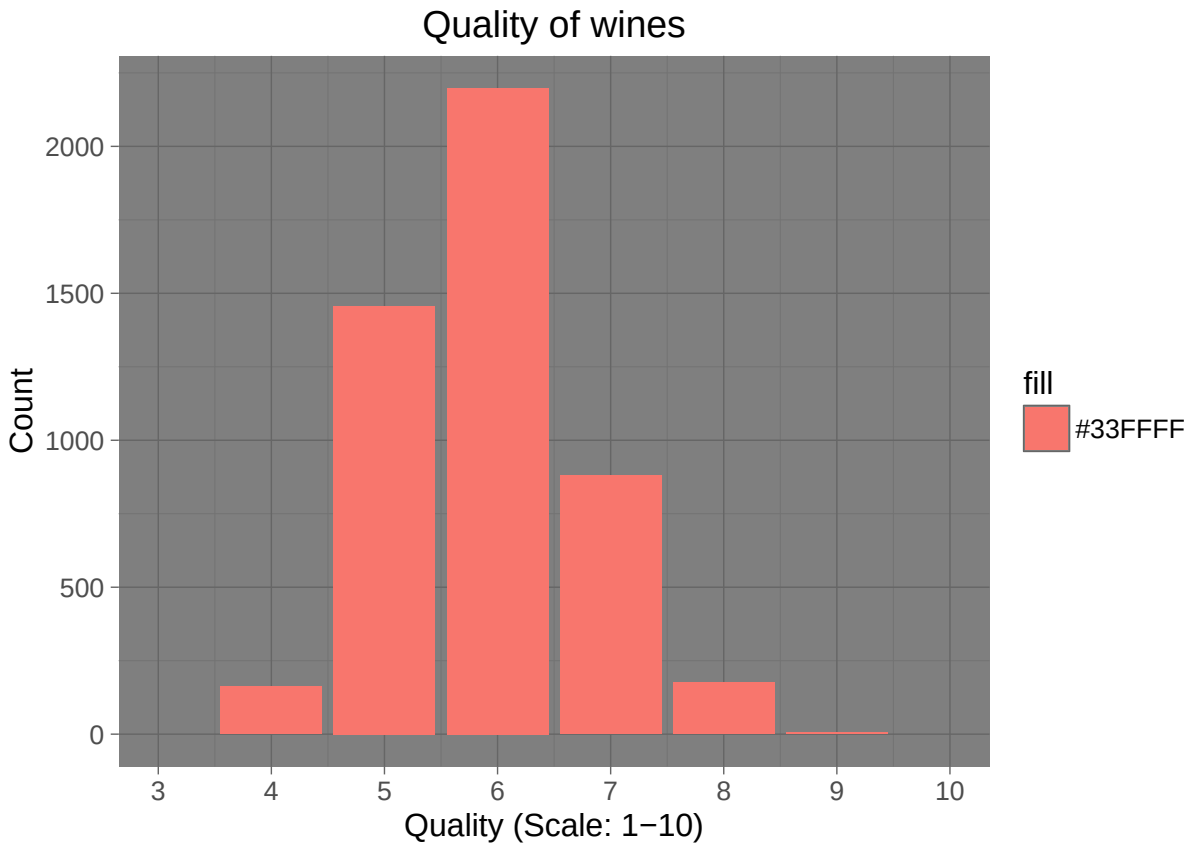
```
# Total Sulphur Dioxide
ggplot(wh_wine, aes(total.sulfur.dioxide)) + geom_histogram(binwidth = 0.1, color = "#33FFFF",
    fill = "#CC99FF") + ggtitle("Levels of total SO2 content in Wines") + xlab("Total SO2") +
    ylab("Count") + xlim(0, 300) + theme_dark()
```

```
## Warning: Removed 6 rows containing non-finite values (stat_bin).
```




**This distribution is also normal with couple of outliers.**

```
## Quality:
ggplot(wh_wine, aes(quality, binwidth = 1, fill = "#33FFFF")) + geom_bar() + ggtitle("Quality of wines"
    xlab("Quality (Scale: 1-10)") + ylab("Count") + scale_x_continuous(breaks = seq(3,
    10, 1), lim = c(3, 10)) + theme_dark()
```




**There are lot of average wines and quite a few excellent and poor wines**

## Doing a basic investigation to see correlations:

```
cor(wh_wine)
```

```
##                       fixed.acidity volatile.acidity  citric.acid residual.sugar
## fixed.acidity           1.00000000      -0.02269729  0.289180698     0.08902070
## volatile.acidity       -0.02269729       1.00000000 -0.149471811     0.06428606
## citric.acid             0.28918070      -0.14947181  1.000000000     0.09421162
## residual.sugar          0.08902070       0.06428606  0.094211624     1.00000000
## chlorides               0.02308564       0.07051157  0.114364448     0.08868454
## free.sulfur.dioxide    -0.04939586      -0.09701194  0.094077221     0.29909835
## total.sulfur.dioxide    0.09106976       0.08926050  0.121130798     0.40143931
## density                 0.26533101       0.02711385  0.149502571     0.83896645
## pH                     -0.42585829      -0.03191537 -0.163748211    -0.19413345
## sulphates              -0.01714299      -0.03572815  0.062330940    -0.02666437
## alcohol                -0.12088112       0.06771794 -0.075728730    -0.45063122
## quality                -0.11366283      -0.19472297 -0.009209091    -0.09757683
## tot.acidity             0.98717874       0.07157062  0.394143356     0.10473749
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.02308564       -0.0493958591          0.091069756
## volatile.acidity      0.07051157       -0.0970119393          0.089260504
## citric.acid           0.11436445        0.0940772210          0.121130798
```

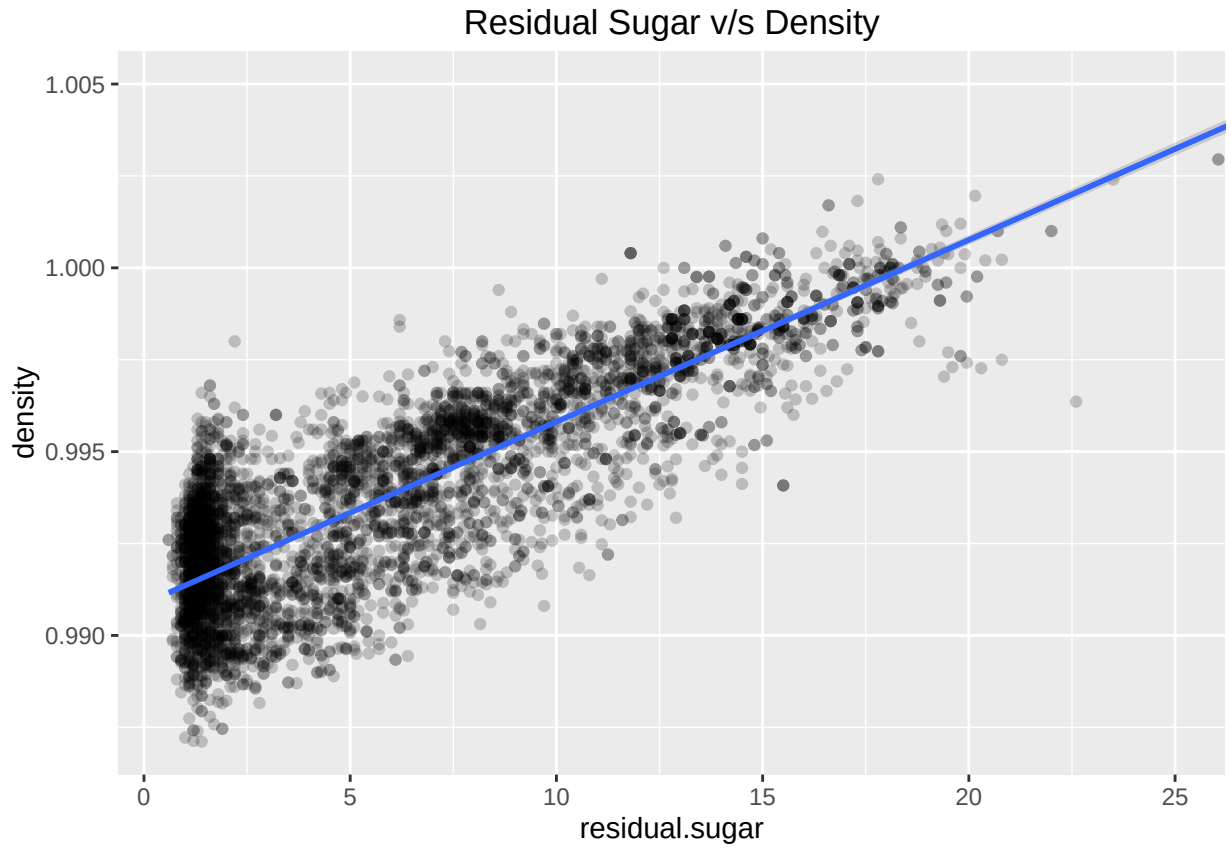
```
## residual.sugar        0.08868454        0.2990983537          0.401439311
## chlorides             1.00000000        0.1013923521          0.198910300
## free.sulfur.dioxide   0.10139235        1.0000000000          0.615500965
## total.sulfur.dioxide  0.19891030        0.6155009650          1.000000000
## density               0.25721132        0.2942104109          0.529881324
## pH                   -0.09043946       -0.0006177961          0.002320972
## sulphates             0.01676288        0.0592172458          0.134562367
## alcohol              -0.36018871       -0.2501039415         -0.448892102
## quality              -0.20993441        0.0081580671         -0.174737218
## tot.acidity           0.04552987       -0.0451333172          0.113188502
##                          density            pH   sulphates     alcohol
## fixed.acidity         0.26533101 -0.4258582910 -0.01714299 -0.12088112
## volatile.acidity      0.02711385 -0.0319153683 -0.03572815  0.06771794
## citric.acid           0.14950257 -0.1637482114  0.06233094 -0.07572873
## residual.sugar        0.83896645 -0.1941334540 -0.02666437 -0.45063122
## chlorides             0.25721132 -0.0904394560  0.01676288 -0.36018871
## free.sulfur.dioxide   0.29421041 -0.0006177961  0.05921725 -0.25010394
## total.sulfur.dioxide  0.52988132  0.0023209718  0.13456237 -0.44889210
## density               1.00000000 -0.0935914935  0.07449315 -0.78013762
## pH                   -0.09359149  1.0000000000  0.15595150  0.12143210
## sulphates             0.07449315  0.1559514973  1.00000000 -0.01743277
## alcohol              -0.78013762  0.1214320987 -0.01743277  1.00000000
## quality              -0.30712331  0.0994272457  0.05367788  0.43557472
## tot.acidity           0.27560881 -0.4306513315 -0.01185225 -0.11751272
##                          quality tot.acidity
## fixed.acidity        -0.113662831  0.98717874
## volatile.acidity     -0.194722969  0.07157062
## citric.acid          -0.009209091  0.39414336
## residual.sugar       -0.097576829  0.10473749
## chlorides            -0.209934411  0.04552987
## free.sulfur.dioxide   0.008158067 -0.04513332
## total.sulfur.dioxide -0.174737218  0.11318850
## density              -0.307123313  0.27560881
## pH                    0.099427246 -0.43065133
## sulphates             0.053677877 -0.01185225
## alcohol               0.435574715 -0.11751272
## quality               1.000000000 -0.13137721
## tot.acidity          -0.131377207  1.00000000
```

## Initial Correlation Analysis:

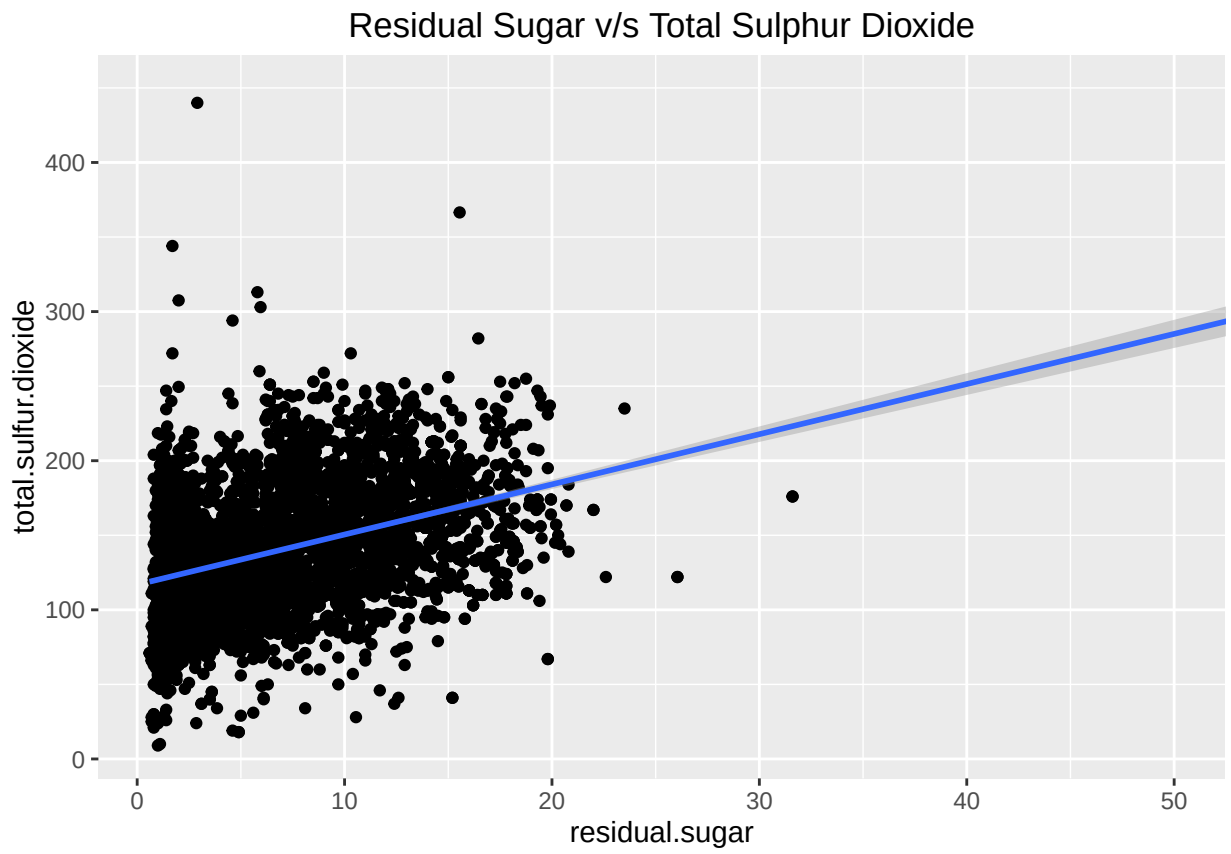
1. Quality which is the outcome variable has only stronger positive correlation with alcohol and negative correlation with density.
2. As hypothesised, free sulphur and total sulphur dioxide have strong positive correlation with each other.
3. pH has negative correlation with fixed acidity and also with total acidity if considered.
4. Residual sugar has positive correlation with total sulphur dioxide while strong positive correlation with density. While it has negative correlation with alcohol.
5. Density have strong negative correlation with alcohol and positive correlation with total sulphur dioxide.

### Bivariate Plots- if variables are related

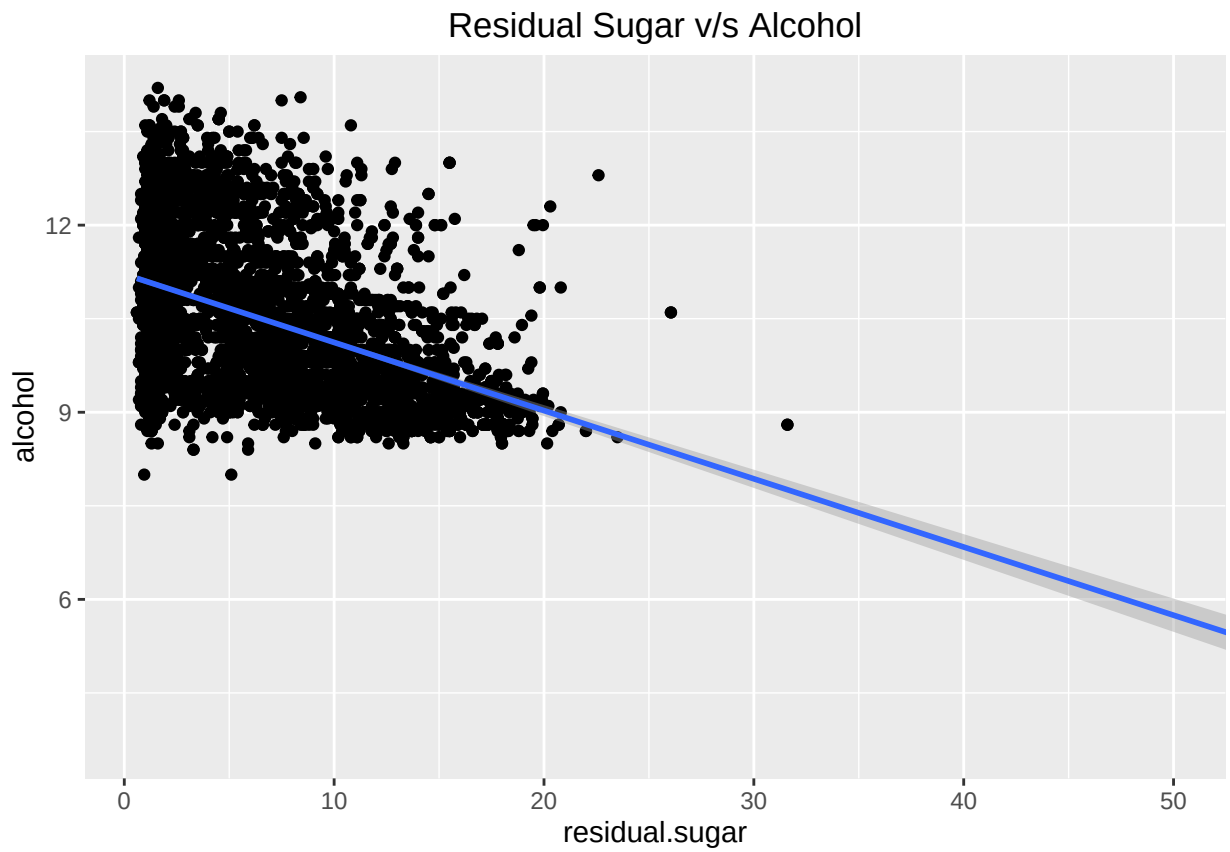
```
## Residual sugar v/s Density :
ggplot(wh_wine, aes(residual.sugar, density)) + geom_point(alpha = 0.2) + geom_smooth(method = "lm") +
    ggtitle("Residual Sugar v/s Density") + coord_cartesian(xlim = c(min(wh_wine$residual.sugar),
    25), ylim = c(min(wh_wine$density), 1.005))
```



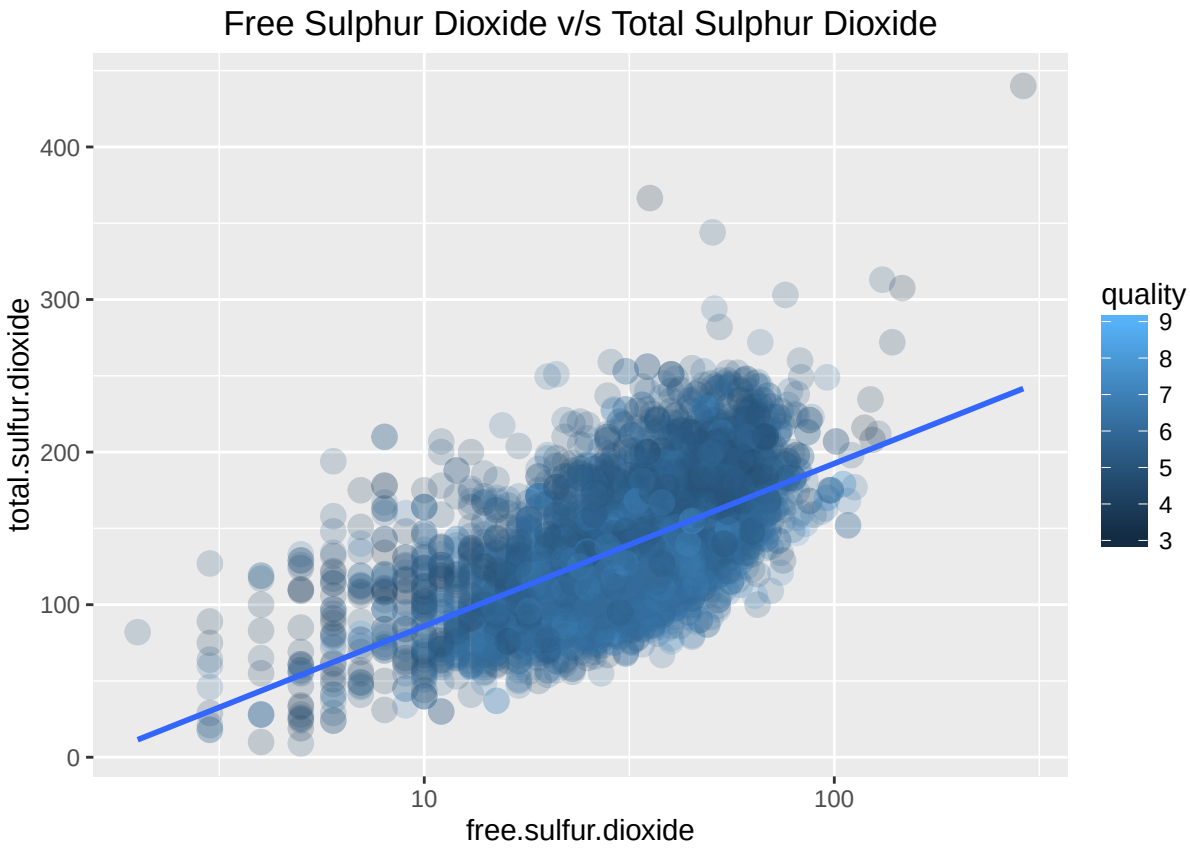

```
# Residual Sugar v/s Total Sulphur Dioxide:
ggplot(wh_wine, aes(residual.sugar, total.sulfur.dioxide)) + geom_point() + geom_smooth(method = lm) +
    ggtitle("Residual Sugar v/s Total Sulphur Dioxide") + coord_cartesian(xlim = c(min(wh_wine$residual
    50), ylim = c(min(wh_wine$total.sulfur.dioxide), 450))
```



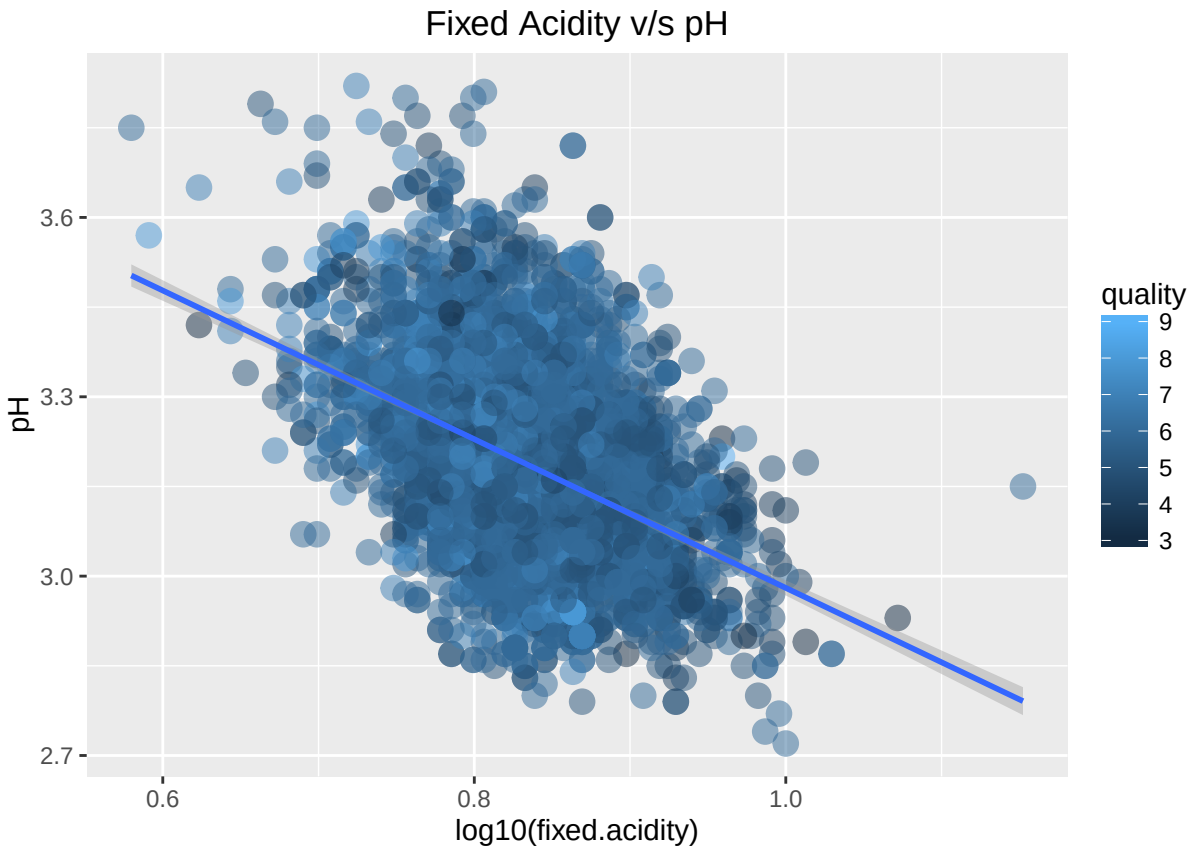

```
# Residual Sugar v/s Alcohol:
ggplot(wh_wine, aes(residual.sugar, alcohol)) + geom_point() + geom_smooth(method = lm) +
    ggtitle("Residual Sugar v/s Alcohol") + coord_cartesian(xlim = c(min(wh_wine$residual.sugar),
    50))
```



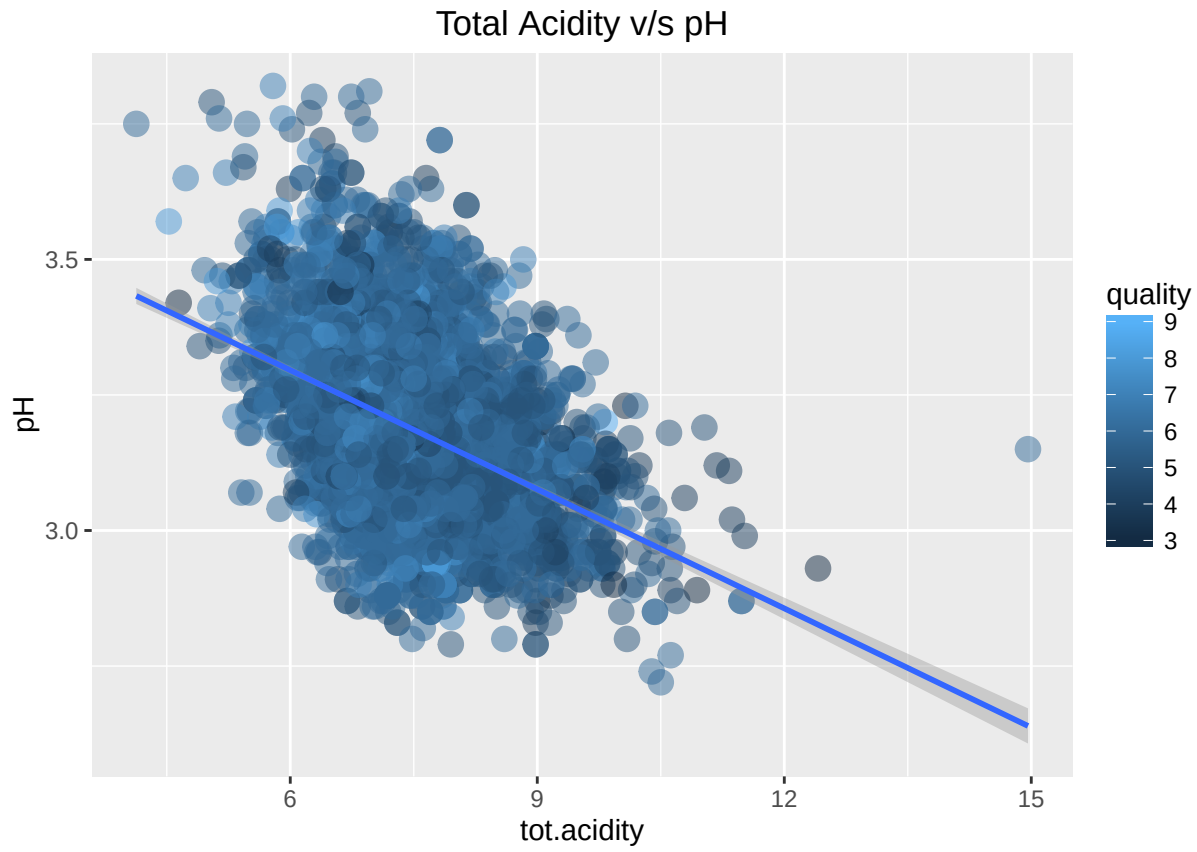

```
# Free sulphur dioxide v/s Total sulphur dioxide:
ggplot(wh_wine, aes(free.sulfur.dioxide, total.sulfur.dioxide, color = quality)) +
    geom_point(size = 4, alpha = 0.2) + geom_smooth(method = lm, se = FALSE) + ggtitle("Free Sulphur Di
    scale_x_log10()
```



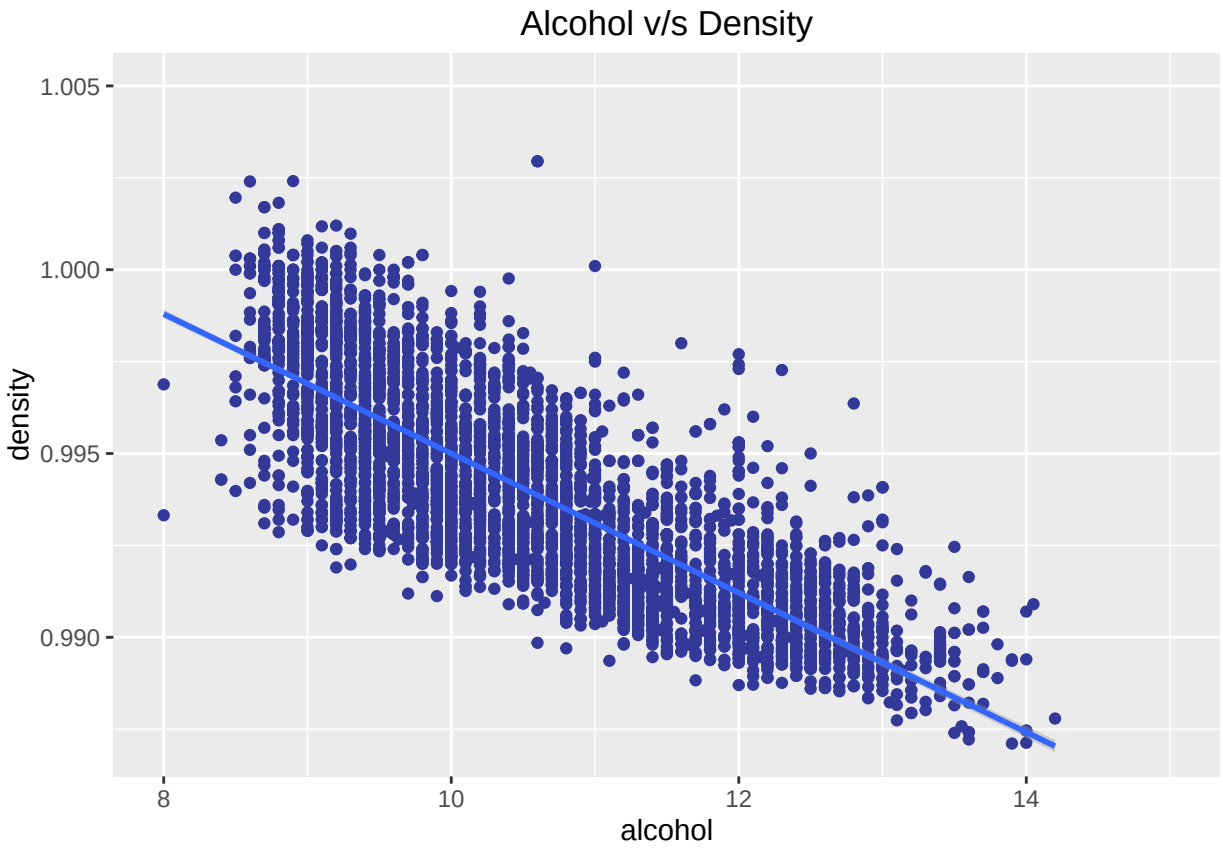

```
# Fixed Acidity v/s pH:
ggplot(wh_wine, aes(log10(fixed.acidity), pH, color = quality)) + geom_point(size = 4,
    alpha = 0.5) + geom_smooth(method = "lm") + ggtitle("Fixed Acidity v/s pH")
```



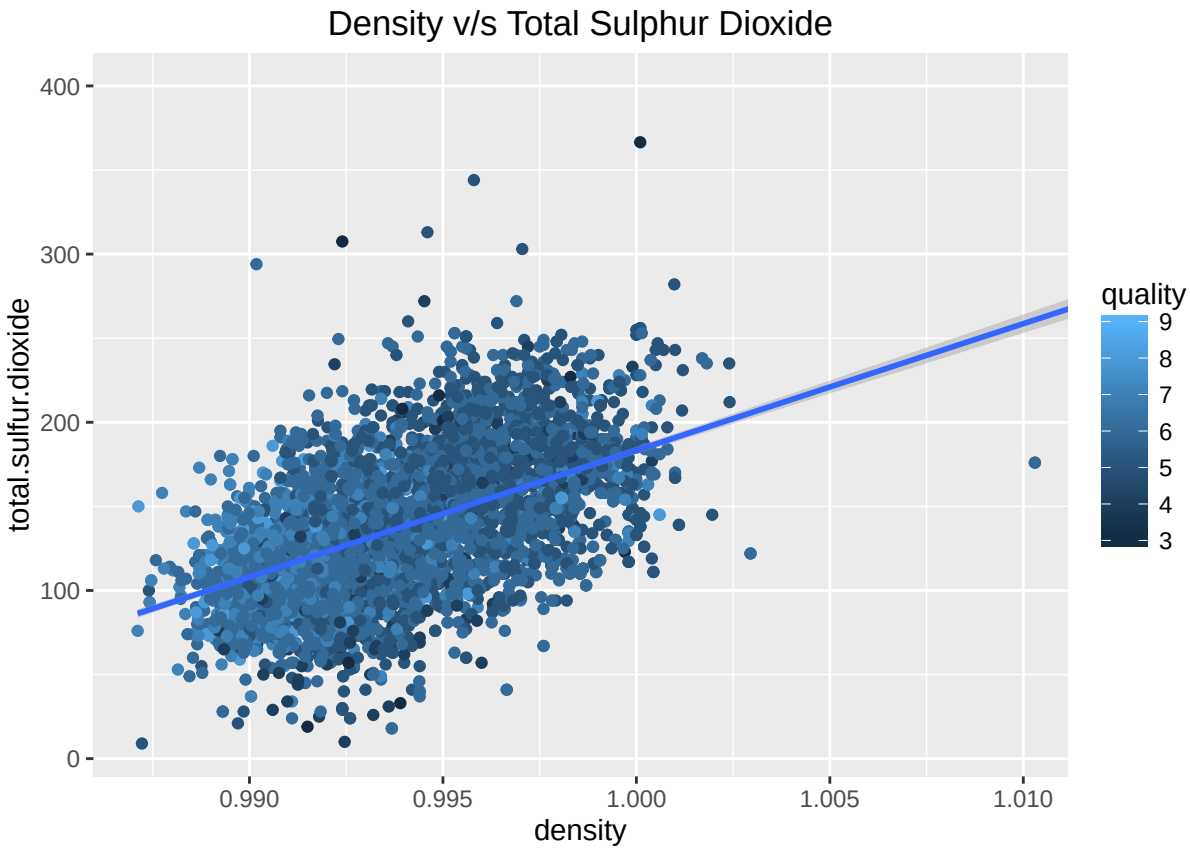

```
# Total Acidity v/s pH:
ggplot(wh_wine, aes(tot.acidity, pH, color = quality)) + geom_point(size = 4, alpha = 0.5) +
    geom_smooth(method = "lm") + ggtitle("Total Acidity v/s pH")
```




```
# Alcohol v/s Density:
ggplot(wh_wine, aes(alcohol, density)) + geom_point(color = "#333999") + geom_smooth(method = lm) +
    ggtitle("Alcohol v/s Density") + coord_cartesian(ylim = c(min(wh_wine$density),
    1.005), xlim = c(min(wh_wine$alcohol), 15))
```

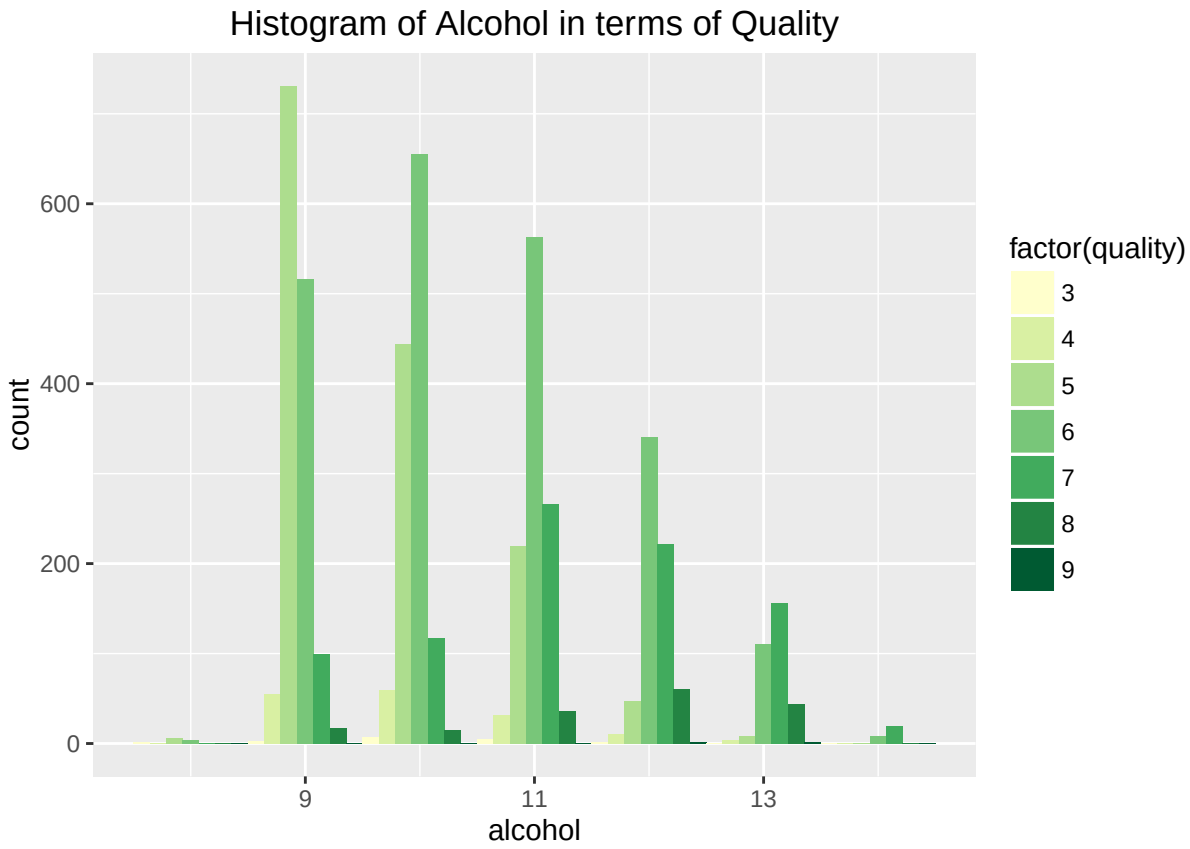



```
# Density v/s Total Sulphur Dioxide:
ggplot(wh_wine, aes(density, total.sulfur.dioxide, color = quality)) + geom_point() +
    geom_smooth(method = lm) + ggtitle("Density v/s Total Sulphur Dioxide") + coord_cartesian(xlim = c(
    1.01), ylim = c(min(wh_wine$total.sulfur.dioxide), 400))
```

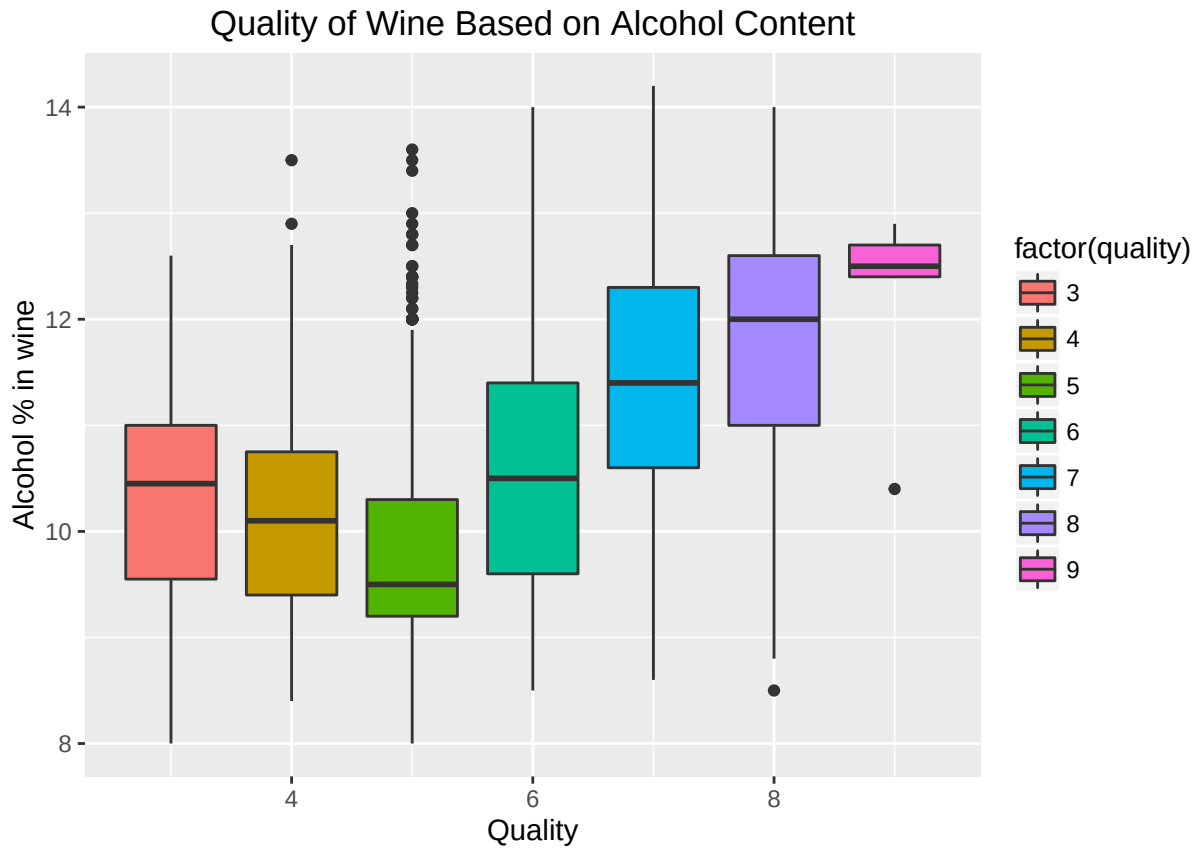



### Multivariate Plots:

```
# Alcohol v/s Quality:
ggplot(wh_wine, aes(alcohol)) + geom_histogram(aes(fill = factor(quality)), binwidth = 1,
    position = "dodge") + scale_fill_brewer(type = "seq", palette = 15) + ggtitle("Histogram of Alcohol
```




```
# Alcohol v/s Quality:
ggplot(wh_wine, aes(quality, alcohol, fill = factor(quality))) + geom_boxplot() +
    ggtitle("Quality of Wine Based on Alcohol Content") + xlab("Quality") + ylab("Alcohol % in wine")
```




```
## Both the above plots show how alcohol content affects the quality. The box plot
## shows how higher content of alcohol can give higher quality of wine. However
## there are some outliers, other variables which might be affect the quality of
## wines. Alcohol content alone would not produce higher quality.
```

**Reflection:**

1. Alcohol has the strongest correlation with the quality of wine being ....
2. Wine with the highest alcohol percentage is of level 7 while the one with least alcohol percentage is of level 5.
3.

**From the analysis I have done here, it is very clear that the quality of wines could not only be estimated on one's individual taste but other factors like viticulture (way grapes are grown) and how they are turned into wine, type of grapes used, region and many more. It is more of subjective measure.**