Tipologia i cicle de vida de les dades, PRAC 2

Joan Borràs

15/5/2022

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre

El dataset seleccionat es diu ‘Red Wine Quality’ i recull un conjunt de característiques associades al vi negre que pretenen descriure o determinar si es tracta d’un vi bo o contràriament un vi dolent.Aquestes característiques o variables són les següents:

- fixed acidity : acidesa del vi que no s’evapora fàcilment.
- volatile acidity: quantitat d’àcid acètic del vi- citric acid: quantitat d’àcid cítric
- residual sugar: quantitat de sucre remenen després de la fermentació- chlorides: quantitat de sal en el vi
- free sulfur dioxide: quantitat de diòxid de sulfur lliure resultant de l’equilibi amb el bisulfit.
- total sulfur dioxide: quantitat total de diòxid de sulfur
- density: quantitat i densitat de l’aigua depenent del percentatge d’alcohol i sucre
- PH: ph del vi- sulphates: quantitat de sulfats.
- Alcohol: quantitat d’alchol del vi
- quality: variable target que determina la qualitat en una escala del 1 al 10

```
# Carraguem el dataframe i mostrem les variables
wine_df <- read.csv("winequality-red.csv",header=T,sep=",", stringsAsFactors = FALSE)
str(wine_df)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

2. Integració i selecció de les dades d’interès a analitzar. Pot ser el resultat d’addicionar diferents datasets o una subselecció útil de les dades originals, en base a l’objectiu que es vulgui aconseguir.

El dataframe seleccionat es pot considerar prou complet i no seria necessàri ampliar. Tampoc resulta òptim simplificar ni reealitzar una subselecció de variables ja que el model predictiu perdria validesa al no disposar de totes les característiques o variables al complet.

No obstant, amb objecte d’estudiar i visualitzar millor les dades, podem crear una columna que categoritzi la qualitat entre molt bo, bo, normal, dolent o molt dolent. Tenint en compte que la variable quality es distribueix en un rang del 0 al 10, categoritzarem de la següent manera:

de [0 a 3) molt dolent, de [3 a 5) dolent, de [5 a 6) normal, de [6 a 8) bo, i de [8 a 10] molt bo.

```
# Categoritzem la variable quality
wine_df$quality_cat <- cut(wine_df$quality, breaks = c(0, 3, 5, 6, 8, 10),
                  labels = c("molt dolent", "dolent", "normal", "bo", "molt bo"))
```

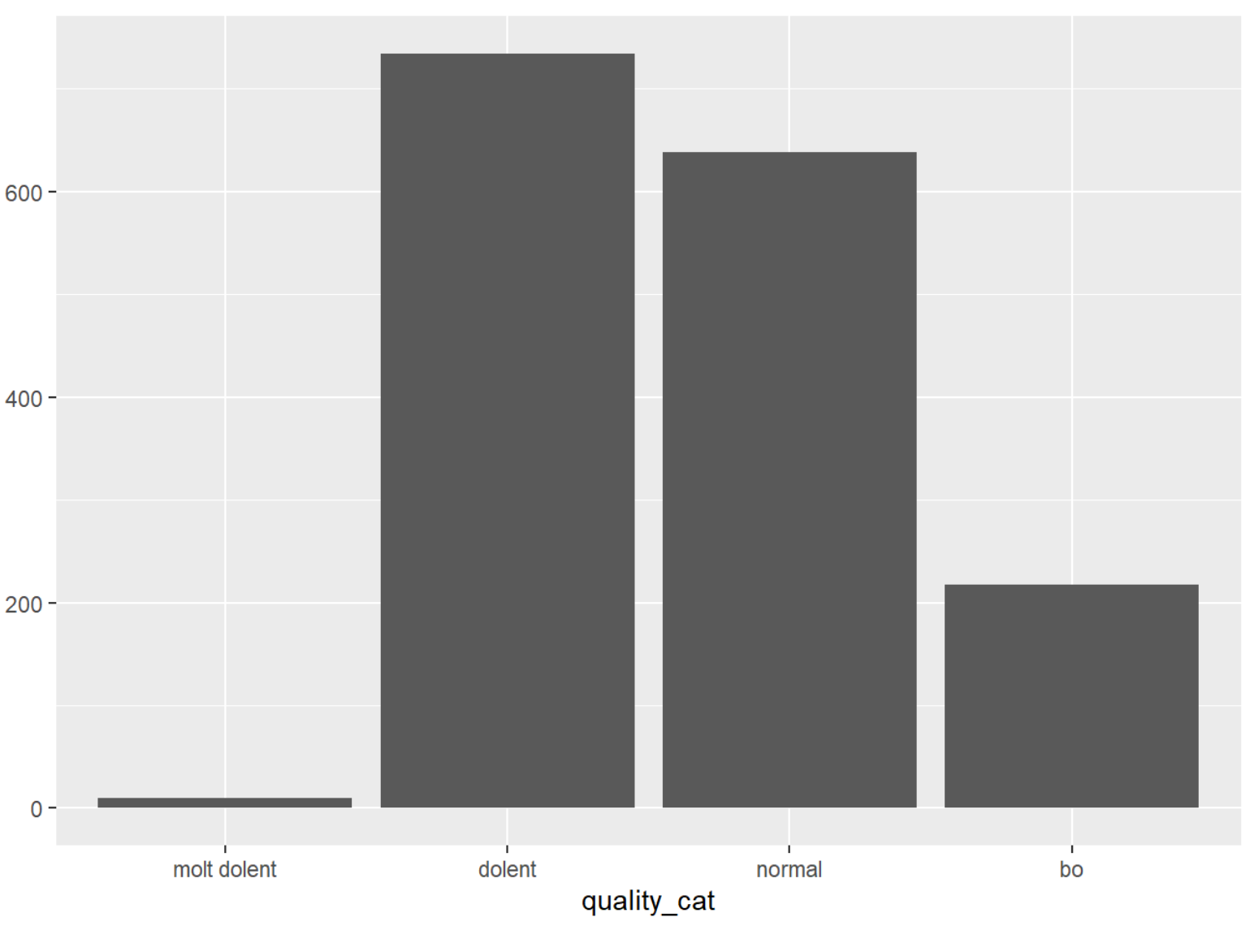
Comprovem que s’ha categoritzat correctament

```
# Mostrem el resultat
str(wine_df)
```

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
##  $ quality_cat         : Factor w/ 5 levels "molt dolent",..: 2 2 2 3 2 2 2 4 4 2 ...
```

Un cop categoritzades les dades, podem observar com es distribueixen

```
# Importem llibreria
library("ggplot2")

# Mostrem la distribució de la variable quality_cat
ggplot(wine_df, aes(x=quality_cat)) +
  geom_bar()
```



Podem observar que el major nombre de registres corresponen a l’etiqueta dolent, mentre que per l’etiqueta molt dolent existeixen molt pocs registres.

Com que ja no farem ús d’aquesta columna, guardem el csv modificat.

```
write.csv(wine_df,"wines_df_mod.csv", row.names = FALSE)
```

Seguidament, un cop realitzada la distribució de la variable quality_cat, proseguim a eliminar-la per tal de seguir estudiant el dataset amb solsament variables numèriques.

```
# Eliminem la variable quality_cat
wine_df$quality_cat <- NULL
```

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Gestiona cadascun d’aquests casos.

```
# Comprovem si existeixen elements nulls en el dataset
is.null(wine_df)
```

```
## [1] FALSE
```

```
# Comprovem si existeixen 0 en el dataset
colSums(wine_df==0)
```

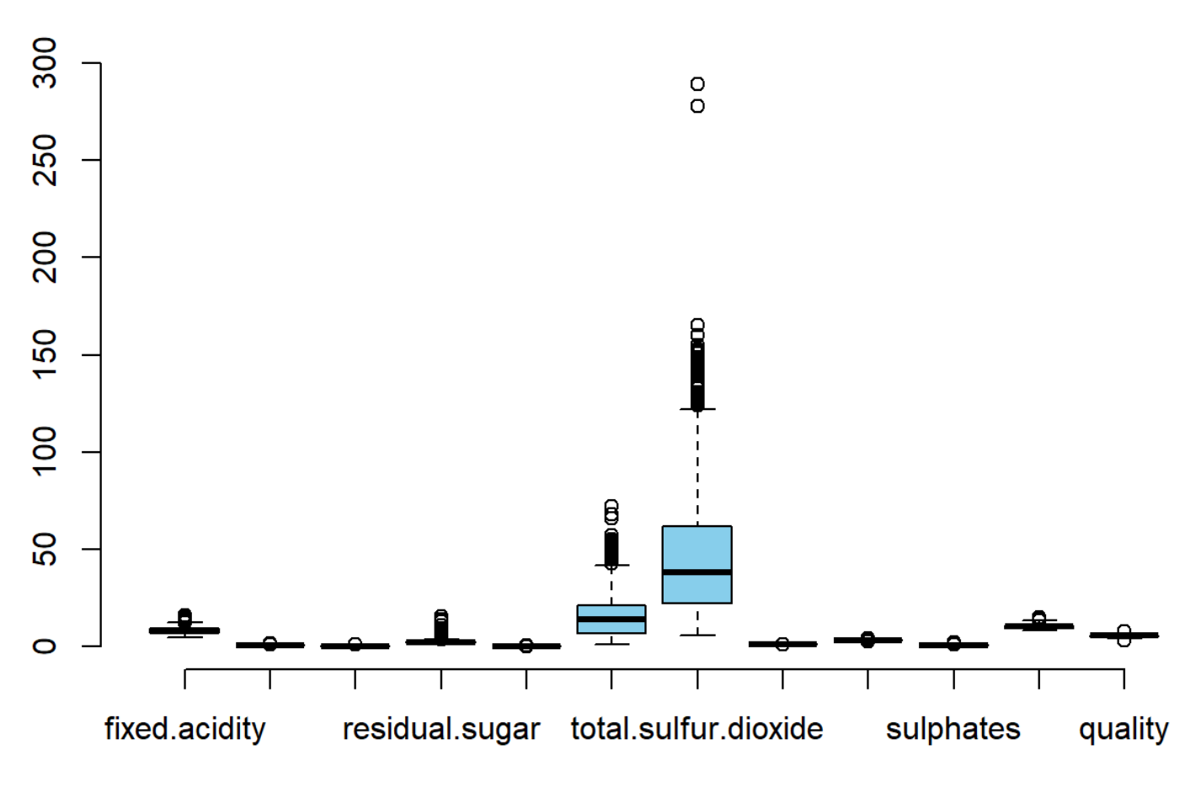
```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                  132
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
```

Podem observar que a la columna citric.acid hi ha una gran quantitat de 0. Al tractar-se d’una variable que raonablement pot contenir 0, és a dir, pot ser que en un vi no hi hagi l’existència de àcid cítric, en el tractament de aquestes dades no cadrà eliminar aquests registres o substituir-los per algun estadístic com per exemple, la mitjana d’aquest valor.

3.2. Identifica i gestiona els valors extrems.

Tractarem de detectar els valors outliers mitjaçant un un gràfic de caixes, que mostra per una banda els valors dintre de la normalitat(caixa) i, per altra banda, els valors fora de aquesta normalitat (fora de la caixa).

```
# Mostrem la gràfica dels valors outliers
win_out <-boxplot(wine_df, col="skyblue", frame.plot=F)
```



Mostrem aquests valors outliers

```
# Mostrem els valors
win_out$out
```

```
##   [1]  12.80000  12.80000  15.00000  15.00000  12.50000  13.30000  13.40000
##   [8]  12.40000  12.50000  13.80000  13.50000  12.60000  12.50000  12.80000
##  [15]  12.80000  14.00000  13.70000  13.70000  12.70000  12.50000  12.80000
##  [22]  12.60000  15.60000  12.50000  13.00000  12.50000  13.30000  12.40000
##  [29]  12.50000  12.90000  14.30000  12.40000  15.50000  15.50000  15.60000
##  [36]  13.00000  12.70000  13.00000  12.70000  12.40000  12.70000  13.20000
##  [43]  13.20000  13.20000  15.90000  13.30000  12.90000  12.60000  12.60000
##  [50]   1.13000   1.02000   1.07000   1.33000   1.33000   1.04000   1.09000
##  [57]   1.04000   1.24000   1.18500   1.02000   1.03500   1.02500   1.11500
##  [64]   1.02000   1.02000   1.58000   1.18000   1.04000   1.00000   6.10000
##  [71]   6.10000   3.80000   3.90000   4.40000  10.70000   5.50000   5.90000
##  [78]   5.90000   3.80000   5.10000   4.65000   4.65000   5.50000   5.50000
##  [85]   5.50000   5.50000   7.30000   7.20000   3.80000   5.60000   4.00000
##  [92]   4.00000   4.00000   4.00000   7.00000   4.00000   4.00000   6.40000
##  [99]   5.60000   5.60000  11.00000  11.00000   4.50000   4.80000   5.80000
## [106]   5.80000   3.80000   4.40000   6.20000   4.20000   7.90000   7.90000
## [113]   3.70000   4.50000   6.70000   6.60000   3.70000   5.20000  15.50000
## [120]   4.10000   8.30000   6.55000   6.55000   4.60000   6.10000   4.30000
## [127]   5.80000   5.15000   6.30000   4.20000   4.20000   4.60000   4.20000
## [134]   4.60000   4.30000   4.30000   7.90000   4.60000   5.10000   5.60000
## [141]   5.60000   6.00000   8.60000   7.50000   4.40000   4.25000   6.00000
## [148]   3.90000   4.20000   4.00000   4.00000   4.00000   6.60000   6.00000
## [155]   6.00000   3.80000   9.00000   4.60000   8.80000   8.80000   5.00000
## [162]   3.80000   4.10000   5.90000   4.10000   6.20000   8.90000   4.00000
## [169]   3.90000   4.00000   8.10000   8.10000   6.40000   6.40000   8.30000
## [176]   8.30000   4.70000   5.50000   5.50000   4.30000   5.50000   3.70000
## [183]   6.20000   5.60000   7.80000   4.60000   5.80000   4.10000  12.90000
## [190]   4.30000  13.40000   4.80000   6.30000   4.50000   4.50000   4.30000
## [197]   4.30000   3.90000   3.80000   5.40000   3.80000   6.10000   3.90000
## [204]   5.10000   5.10000   3.90000  15.40000  15.40000   4.80000   5.20000
## [211]   5.20000   3.75000  13.80000  13.80000   5.70000   4.30000   4.10000
## [218]   4.10000   4.40000   3.70000   6.70000  13.90000   5.10000   7.80000
## [225]   0.17600   0.17000   0.36800   0.34100   0.17200   0.33200   0.46400
## [232]   0.40100   0.46700   0.12200   0.17800   0.14600   0.23600   0.61000
## [239]   0.36000   0.27000   0.03900   0.33700   0.26300   0.61100   0.35800
## [246]   0.34300   0.18600   0.21300   0.21400   0.12100   0.12200   0.12200
## [253]   0.12800   0.12000   0.15900   0.12400   0.12200   0.12200   0.17400
## [260]   0.12100   0.12700   0.41300   0.15200   0.15200   0.12500   0.12200
## [267]   0.20000   0.17100   0.22600   0.22600   0.25000   0.14800   0.12200
## [274]   0.12400   0.12400   0.14300   0.22200   0.03900   0.15700   0.42200
## [281]   0.03400   0.38700   0.41500   0.15700   0.15700   0.24300   0.24100
## [288]   0.19000   0.13200   0.12600   0.03800   0.16500   0.14500   0.14700
## [295]   0.01200   0.01200   0.03900   0.19400   0.13200   0.16100   0.12000
## [302]   0.12000   0.12300   0.12300   0.41400   0.21600   0.17100   0.17800
## [309]   0.36900   0.16600   0.16600   0.13600   0.13200   0.13200   0.12300
## [316]   0.12300   0.12300   0.40300   0.13700   0.41400   0.16600   0.16800
## [323]   0.41500   0.15300   0.41500   0.26700   0.12300   0.21400   0.21400
## [330]   0.16900   0.20500   0.20500   0.03900   0.23500   0.23000   0.03800
## [337]  52.00000  51.00000  50.00000  68.00000  68.00000  43.00000  47.00000
## [344]  54.00000  46.00000  45.00000  53.00000  52.00000  51.00000  45.00000
## [351]  57.00000  50.00000  45.00000  48.00000  43.00000  48.00000  72.00000
## [358]  43.00000  51.00000  51.00000  52.00000  55.00000  55.00000  48.00000
## [365]  48.00000  66.00000 145.00000 148.00000 136.00000 125.00000 140.00000
## [372] 136.00000 133.00000 153.00000 134.00000 141.00000 129.00000 128.00000
## [379] 129.00000 128.00000 143.00000 144.00000 127.00000 126.00000 145.00000
## [386] 144.00000 135.00000 165.00000 124.00000 124.00000 134.00000 124.00000
## [393] 129.00000 151.00000 133.00000 142.00000 149.00000 147.00000 145.00000
## [400] 148.00000 155.00000 151.00000 152.00000 125.00000 127.00000 139.00000
## [407] 143.00000 144.00000 130.00000 278.00000 289.00000 135.00000 160.00000
## [414] 141.00000 141.00000 133.00000 147.00000 147.00000 131.00000 131.00000
## [421] 131.00000   0.99160   0.99160   1.00140   1.00150   1.00150   1.00180
## [428]   0.99120   1.00220   1.00220   1.00140   1.00140   1.00140   1.00140
## [435]   1.00320   1.00260   1.00140   1.00315   1.00315   1.00315   1.00210
## [442]   1.00210   0.99170   0.99220   1.00260   0.99210   0.99154   0.99064
## [449]   0.99064   1.00289   0.99162   0.99007   0.99007   0.99020   0.99220
## [456]   0.99150   0.99157   0.99080   0.99084   0.99191   1.00369   1.00369
## [463]   1.00242   0.99182   1.00242   0.99182   3.90000   3.75000   3.85000
## [470]   2.74000   3.69000   3.69000   2.88000   2.86000   3.74000   2.92000
## [477]   2.92000   2.92000   3.72000   2.87000   2.89000   2.89000   2.92000
## [484]   3.90000   3.71000   3.69000   3.69000   3.71000   3.71000   2.89000
## [491]   2.89000   3.78000   3.70000   3.78000   4.01000   2.90000   4.01000
## [498]   3.71000   2.88000   3.72000   3.72000   1.56000   1.28000   1.08000
## [505]   1.20000   1.12000   1.28000   1.14000   1.95000   1.22000   1.95000
## [512]   1.98000   1.31000   2.00000   1.08000   1.59000   1.02000   1.03000
## [519]   1.61000   1.09000   1.26000   1.08000   1.00000   1.36000   1.18000
## [526]   1.13000   1.04000   1.11000   1.13000   1.07000   1.06000   1.06000
## [533]   1.05000   1.06000   1.04000   1.05000   1.02000   1.14000   1.02000
## [540]   1.36000   1.36000   1.05000   1.17000   1.62000   1.06000   1.18000
## [547]   1.07000   1.34000   1.16000   1.10000   1.15000   1.17000   1.17000
## [554]   1.33000   1.18000   1.17000   1.03000   1.17000   1.10000   1.01000
## [561]  14.00000  14.00000  14.00000  14.00000  14.90000  14.00000  13.60000
## [568]  13.60000  13.60000  14.00000  14.00000  13.56667  13.60000   8.00000
## [575]   8.00000   8.00000   8.00000   8.00000   3.00000   8.00000   8.00000
## [582]   8.00000   3.00000   8.00000   3.00000   8.00000   3.00000   3.00000
## [589]   8.00000   8.00000   8.00000   8.00000   8.00000   3.00000   3.00000
## [596]   8.00000   8.00000   3.00000   3.00000   3.00000   8.00000
```

Eliminem aquests valors per tal de no esbiaixar l’estudi de les dades

```
# Eliminem els outliers
wine_df <-wine_df[!(wine_df %in% win_out$out),]
```

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d’anàlisi s’aplicaran?).

En primer lloc estudiarem les diverses distribucions de les variables en funció de la variable objectiu.

Tenint en compte la naturalesa del dataset, en primer lloc estudiarem les correlacions entre totes les variables.En el cas que s’observin variables altament relacionades, aplicarem un test Xi2 per tal de contrastar si aquestes són independents.

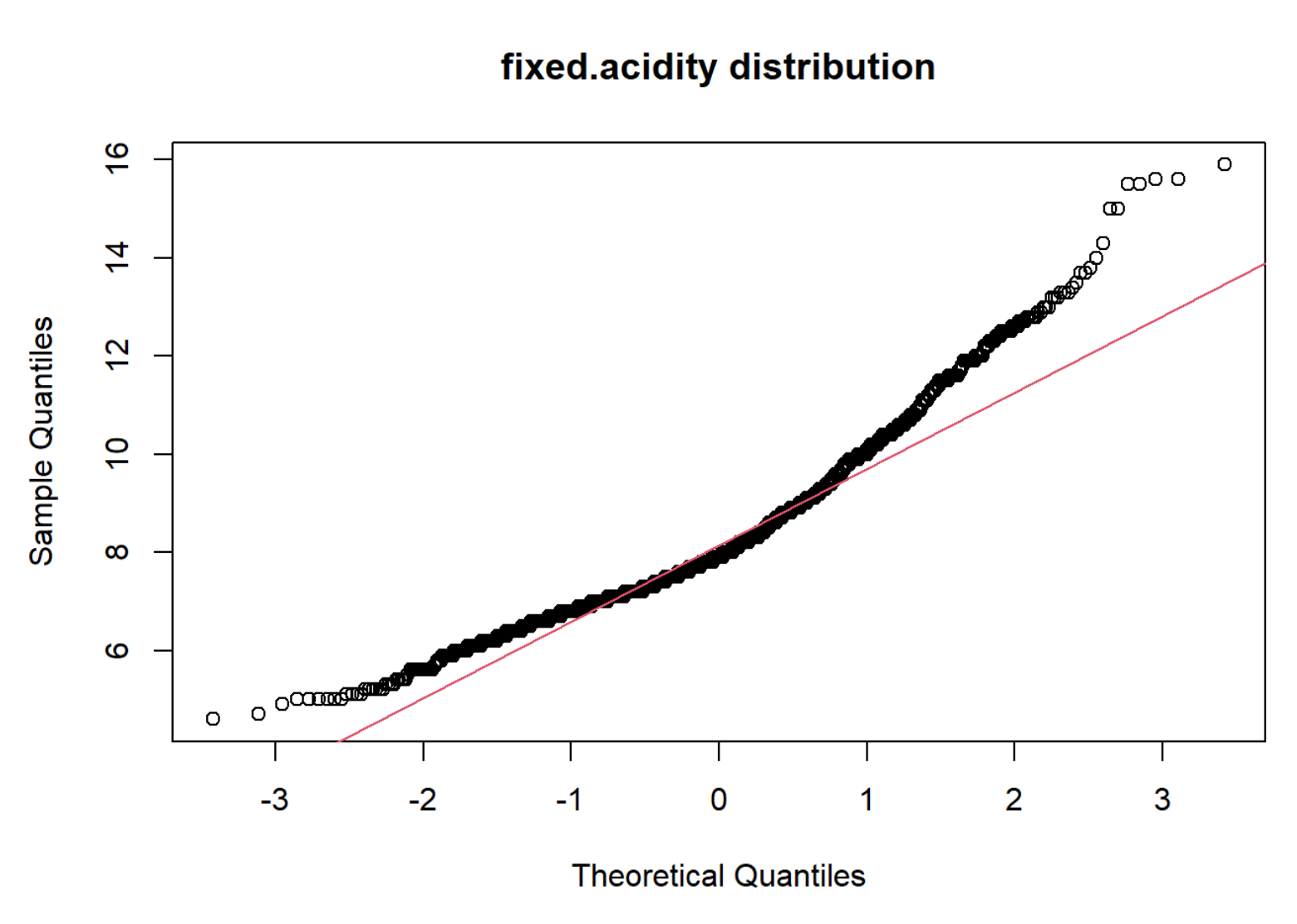
Finalment realitzarem una regressió lineal múltiple per veure l’influència de cada variable en la variable target, en aquest cas, Quality.

També veurem les distribucions que segueixen les diferents variables segons la qualitat del vi. Resultarà interessant comprobar també

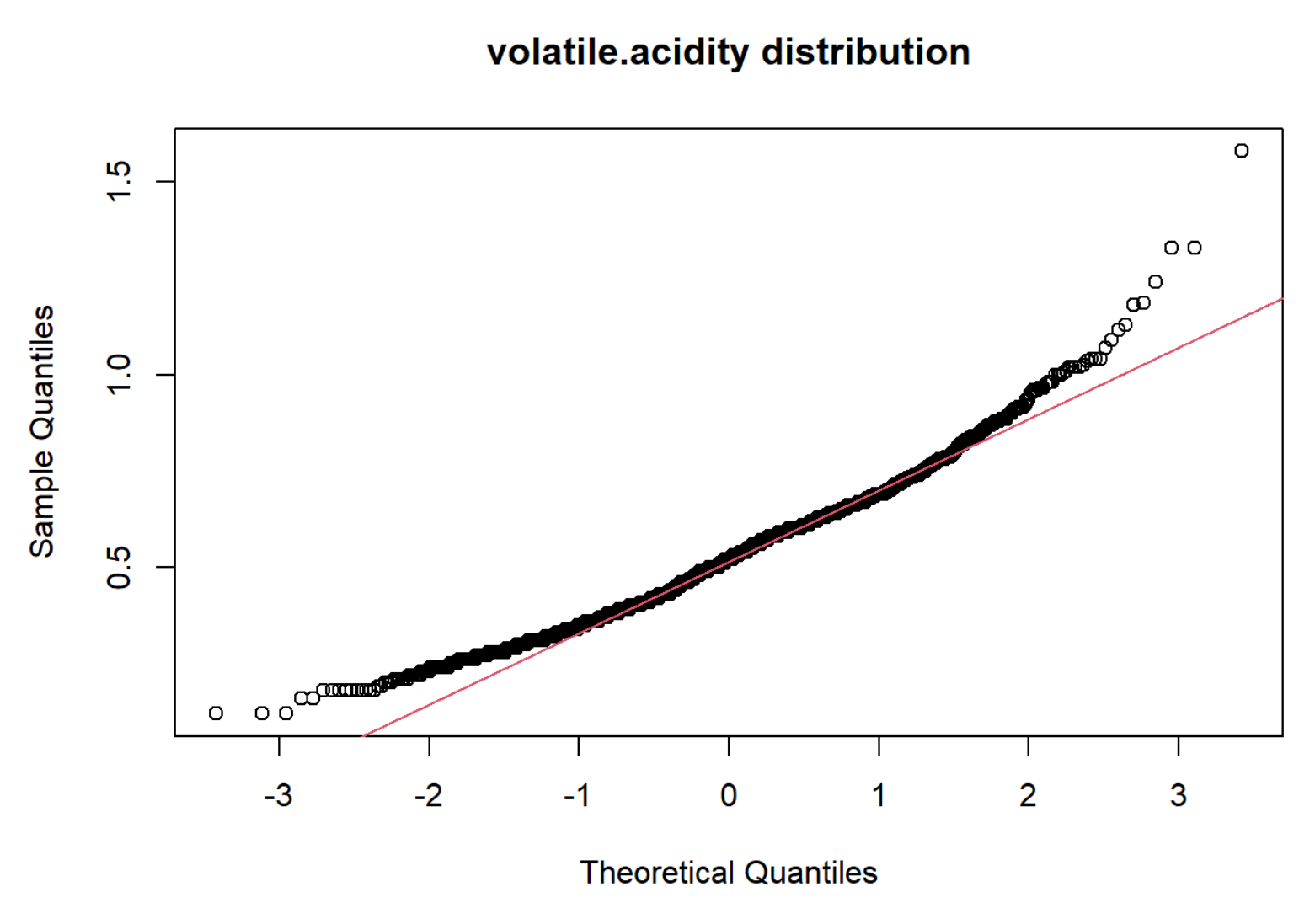
4.2. Comprovació de la normalitat i homogeneïtat de la variància.

En primer lloc, per tal d’estudiar la normalitat de les variables farem servir un gràfic quantil-quantil que ens dona informació sobre la distribució dels valors de cada variable.

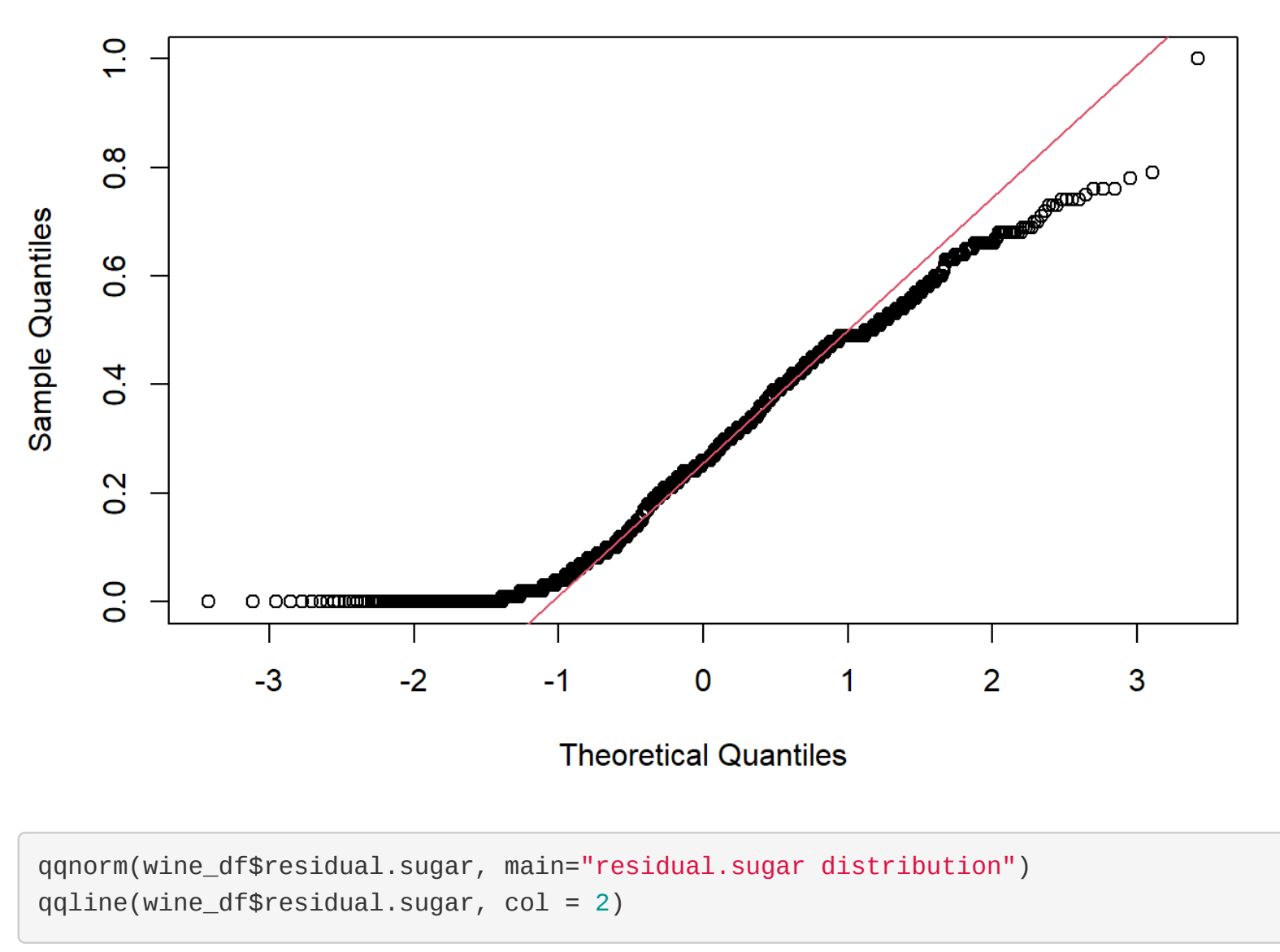
```
# Creem les gràfiques de distribució qq (quantil quantil)
qqnorm(wine_df$fixed.acidity, main="fixed.acidity distribution")
qqline(wine_df$fixed.acidity, col = 2)
```



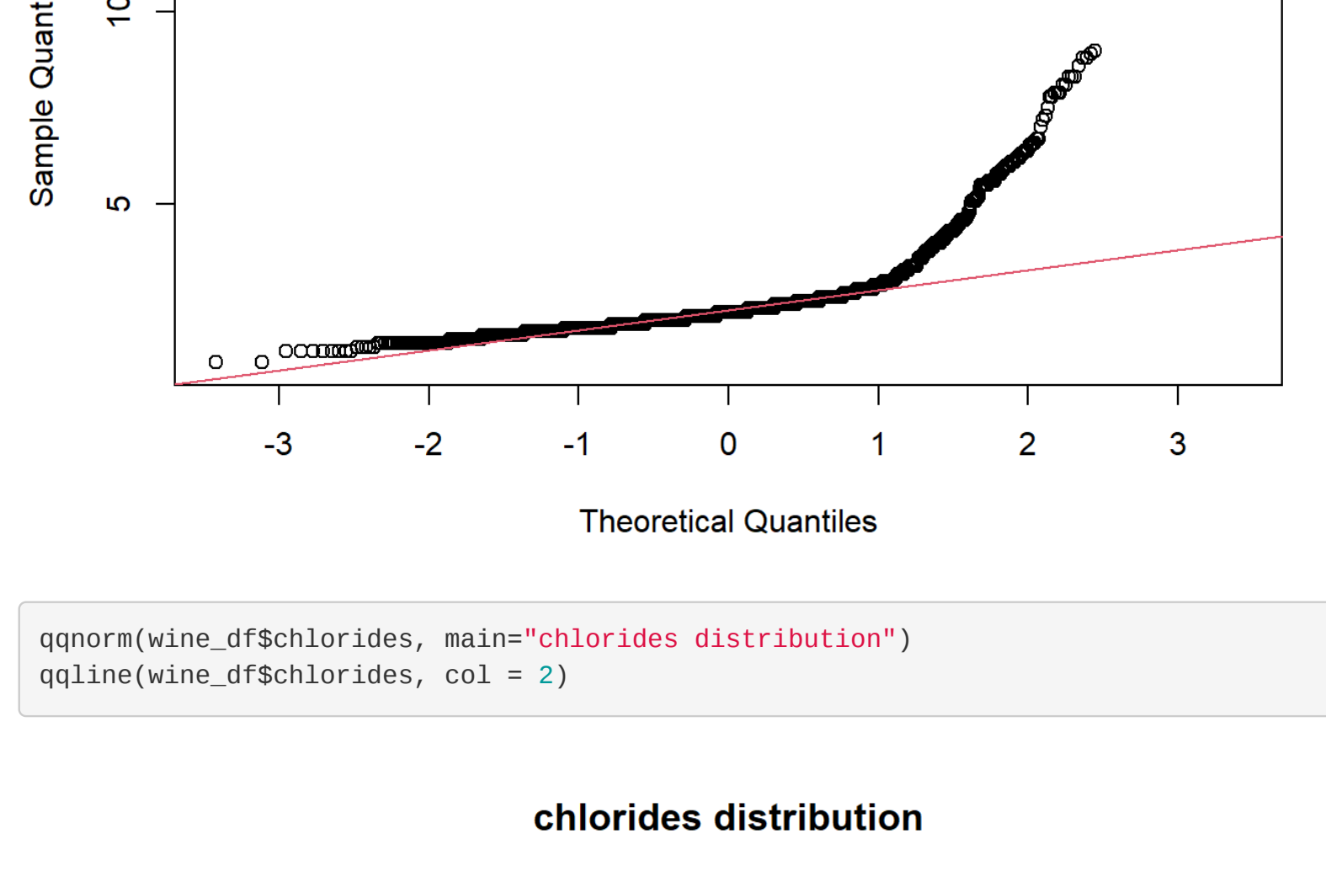
```
qqnorm(wine_df$volatile.acidity, main="volatile.acidity distribution")
qqline(wine_df$volatile.acidity, col = 2)
```



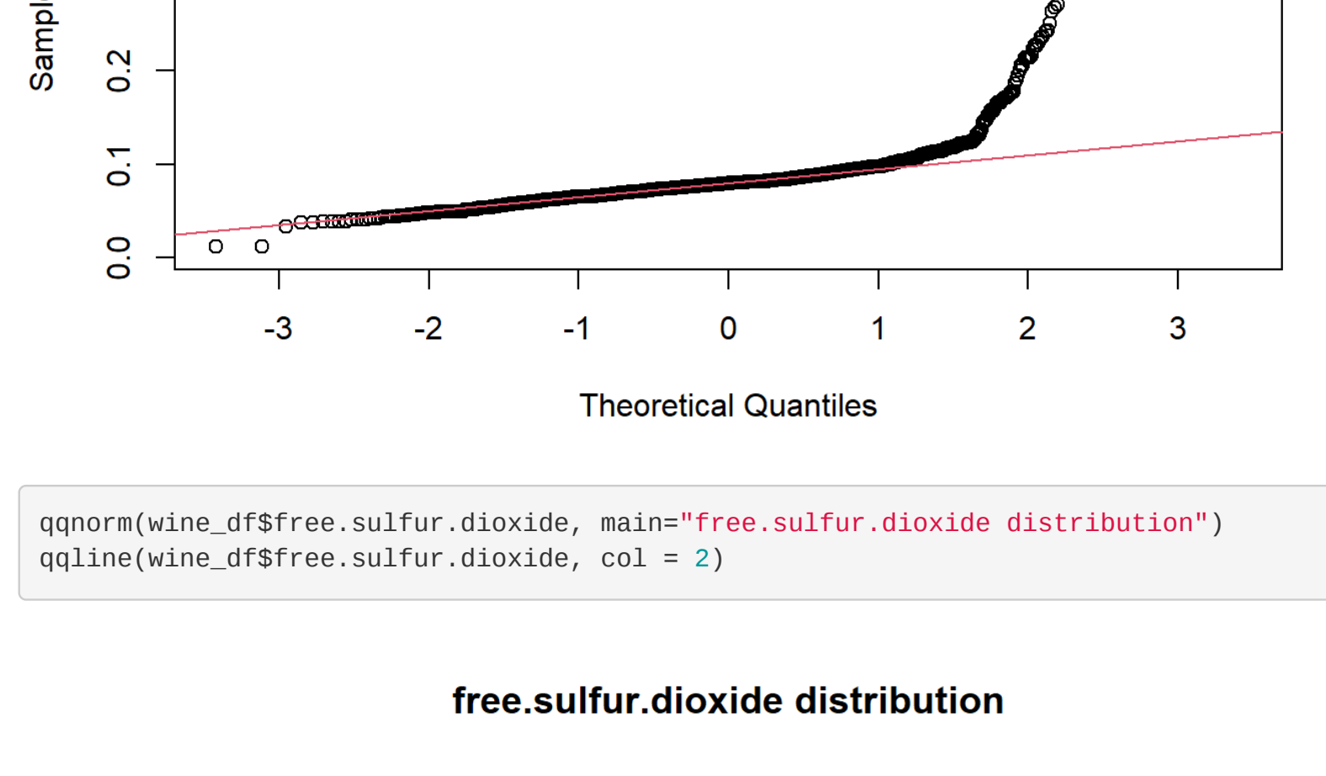
```
qqnorm(wine_df$citric.acid, main="citric.acid distribution")
qqline(wine_df$citric.acid, col = 2)
```



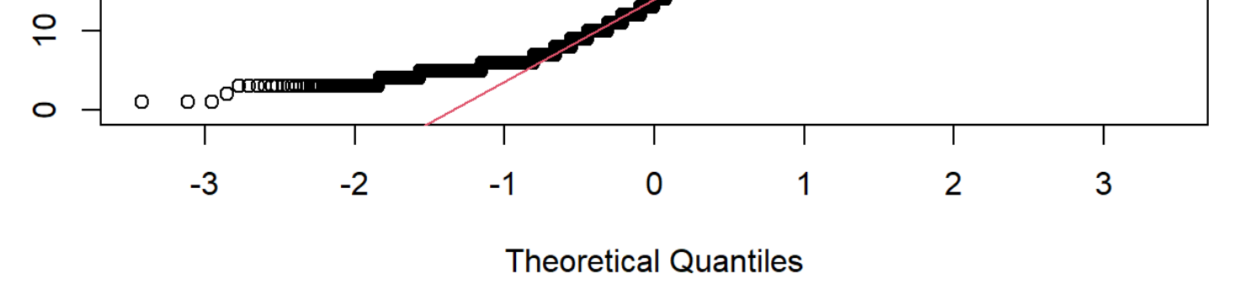
```
qqnorm(wine_df$residual.sugar, main="residual.sugar distribution")
qqline(wine_df$residual.sugar, col = 2)
```



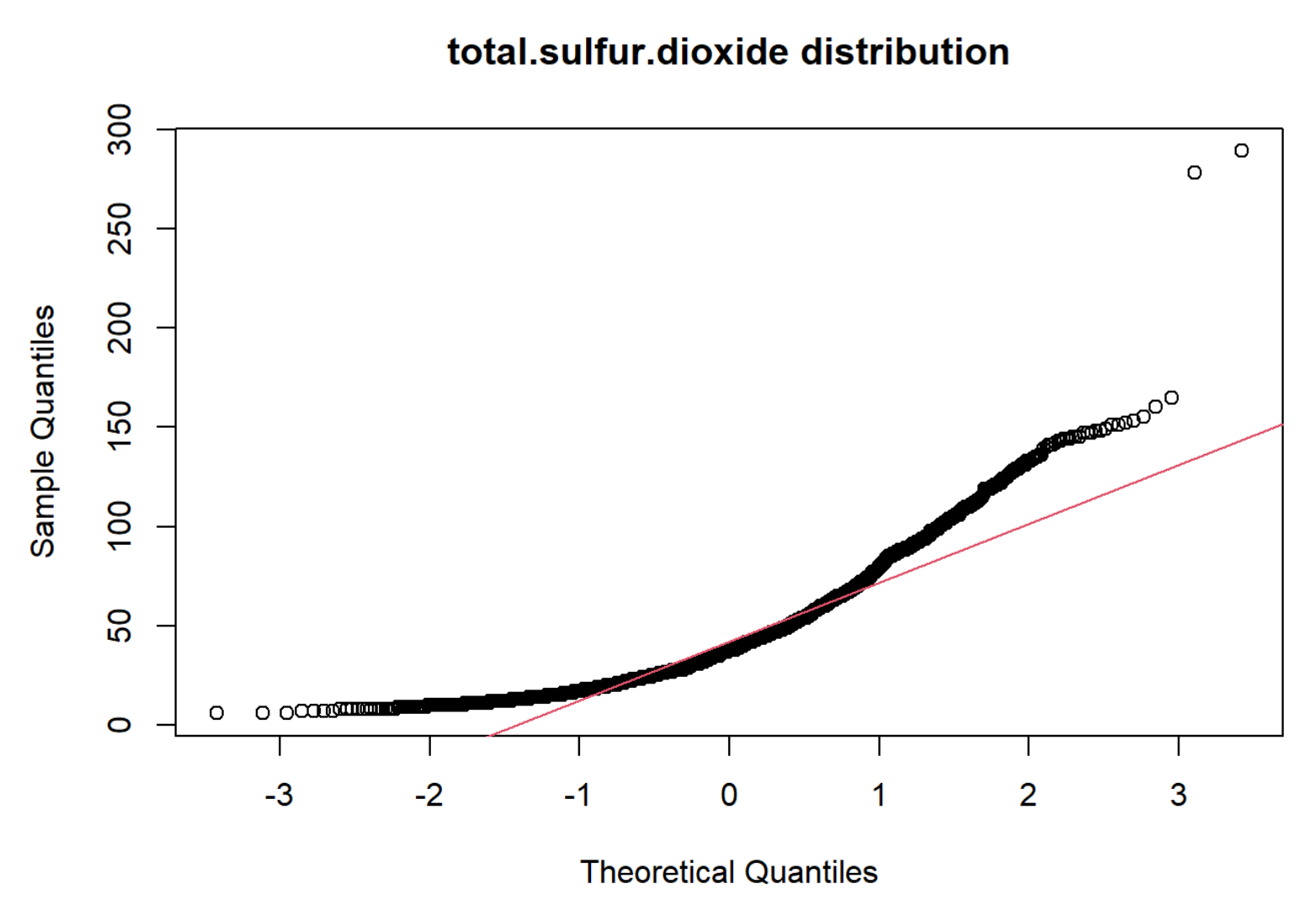
```
qqnorm(wine_df$chlorides, main="chlorides distribution")
qqline(wine_df$chlorides, col = 2)
```



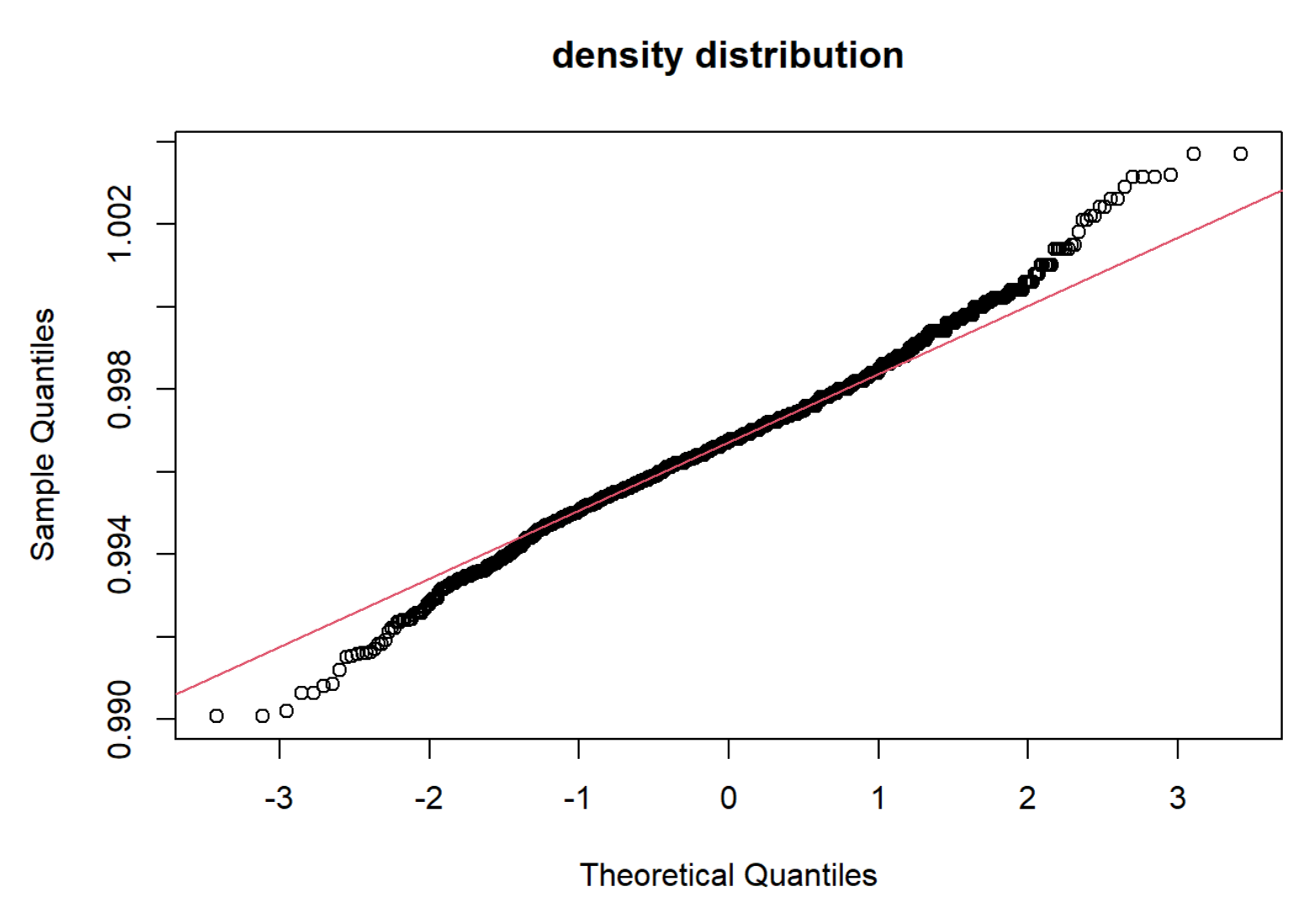
```
qqnorm(wine_df$free.sulfur.dioxide, main="free.sulfur.dioxide distribution")
qqline(wine_df$free.sulfur.dioxide, col = 2)
```



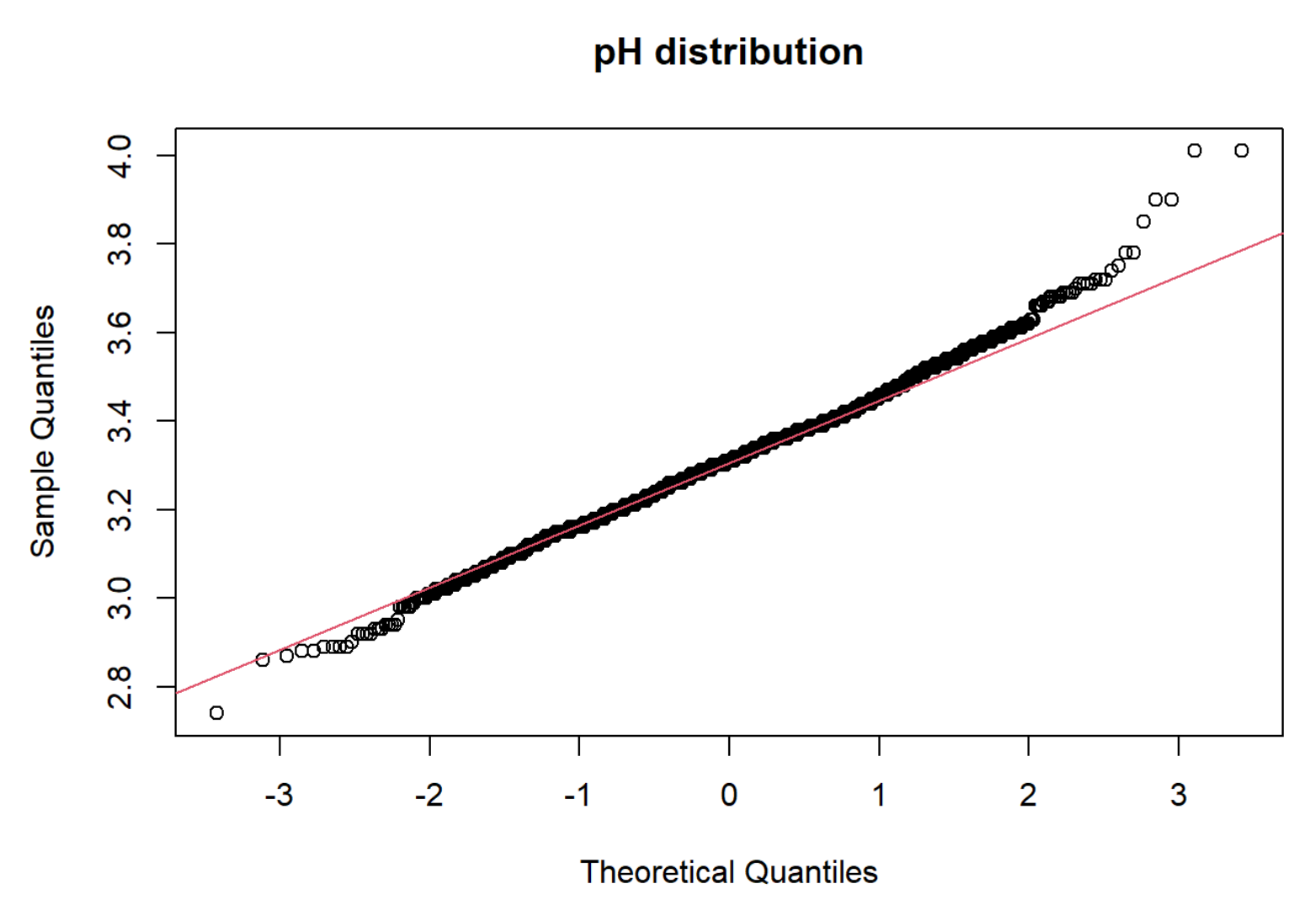
```
qqnorm(wine_df$total.sulfur.dioxide, main="total.sulfur.dioxide distribution")
qqline(wine_df$total.sulfur.dioxide, col = 2)
```



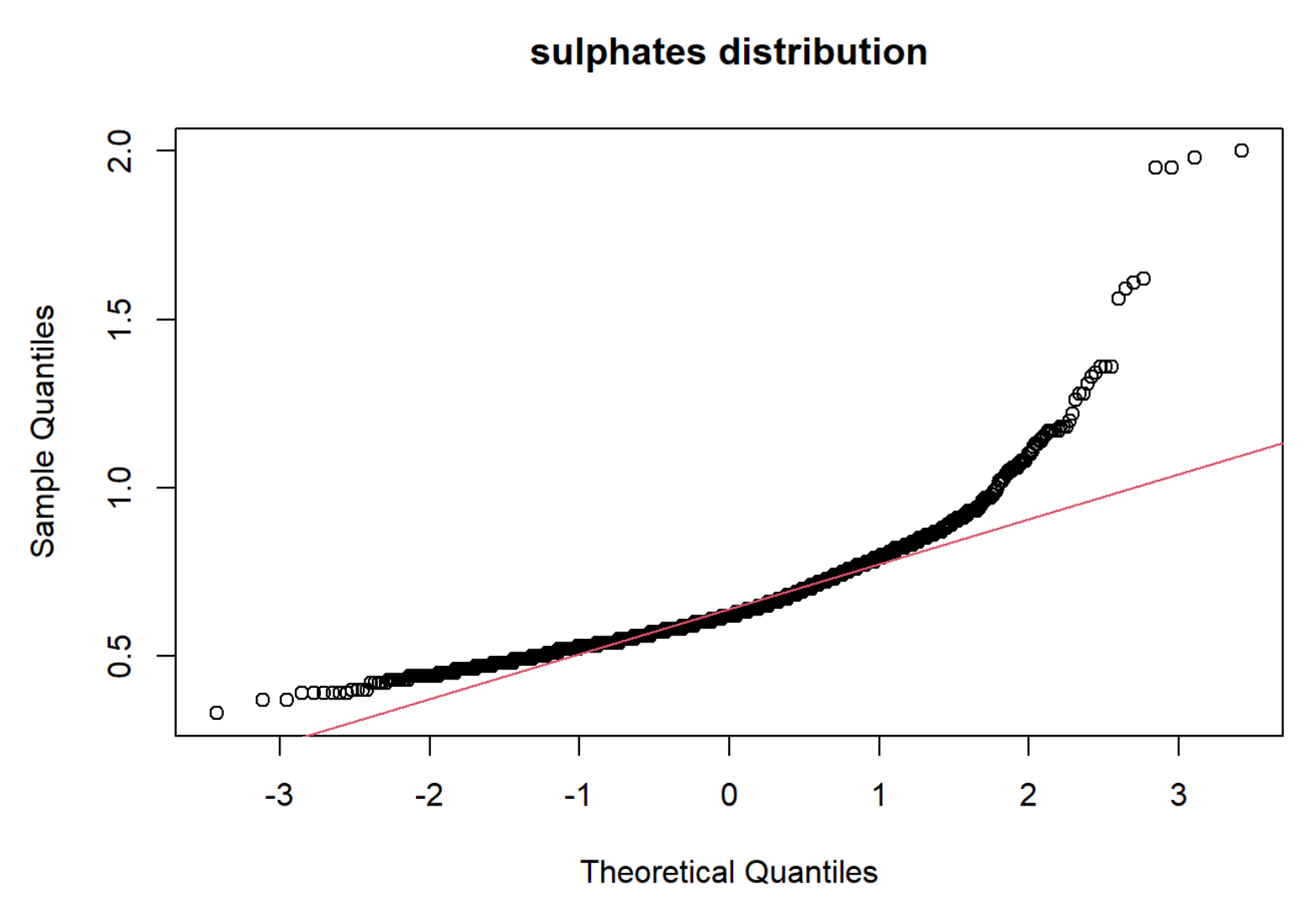
```
qqnorm(wine_df$density, main="density distribution")
qqline(wine_df$density, col = 2)
```



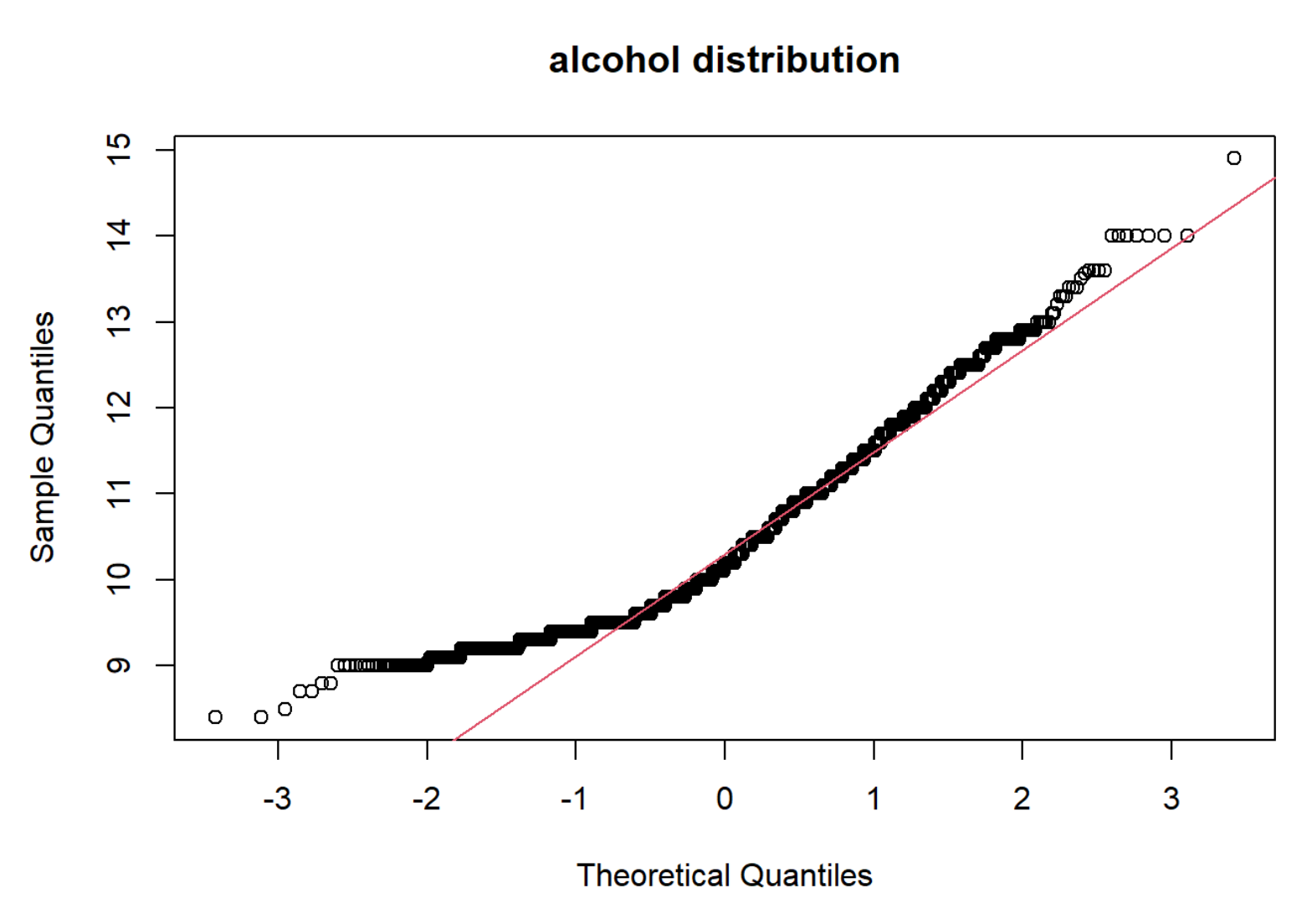
```
qqnorm(wine_df$pH, main="pH distribution")
qqline(wine_df$pH, col = 2)
```



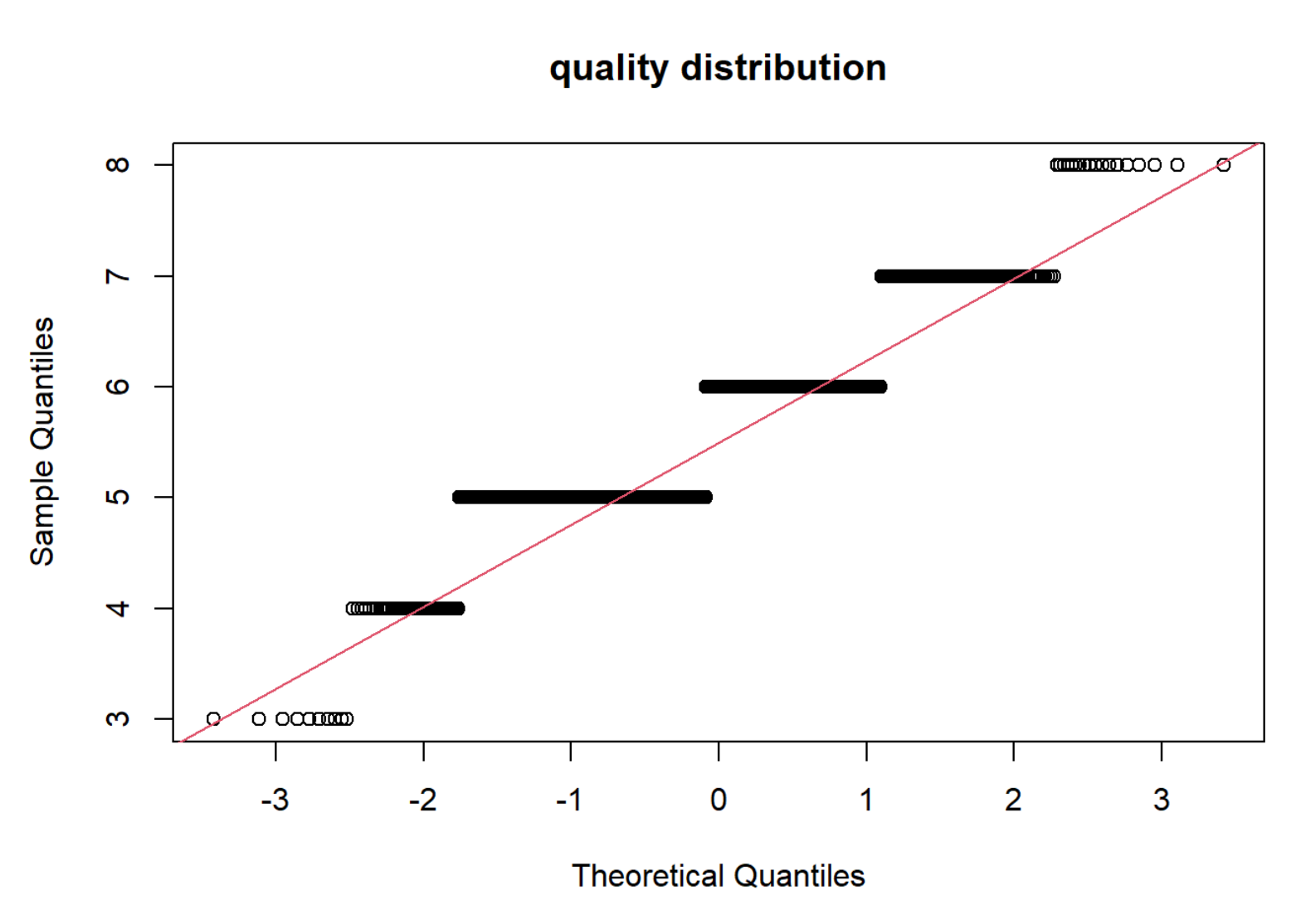
```
qqnorm(wine_df$sulphates, main="sulphates distribution")
qqline(wine_df$sulphates, col = 2)
```



```
qqnorm(wine_df$alcohol, main="alcohol distribution")
qqline(wine_df$alcohol, col = 2)
```



```
qqnorm(wine_df$quality, main="quality distribution")
qqline(wine_df$quality, col = 2)
```



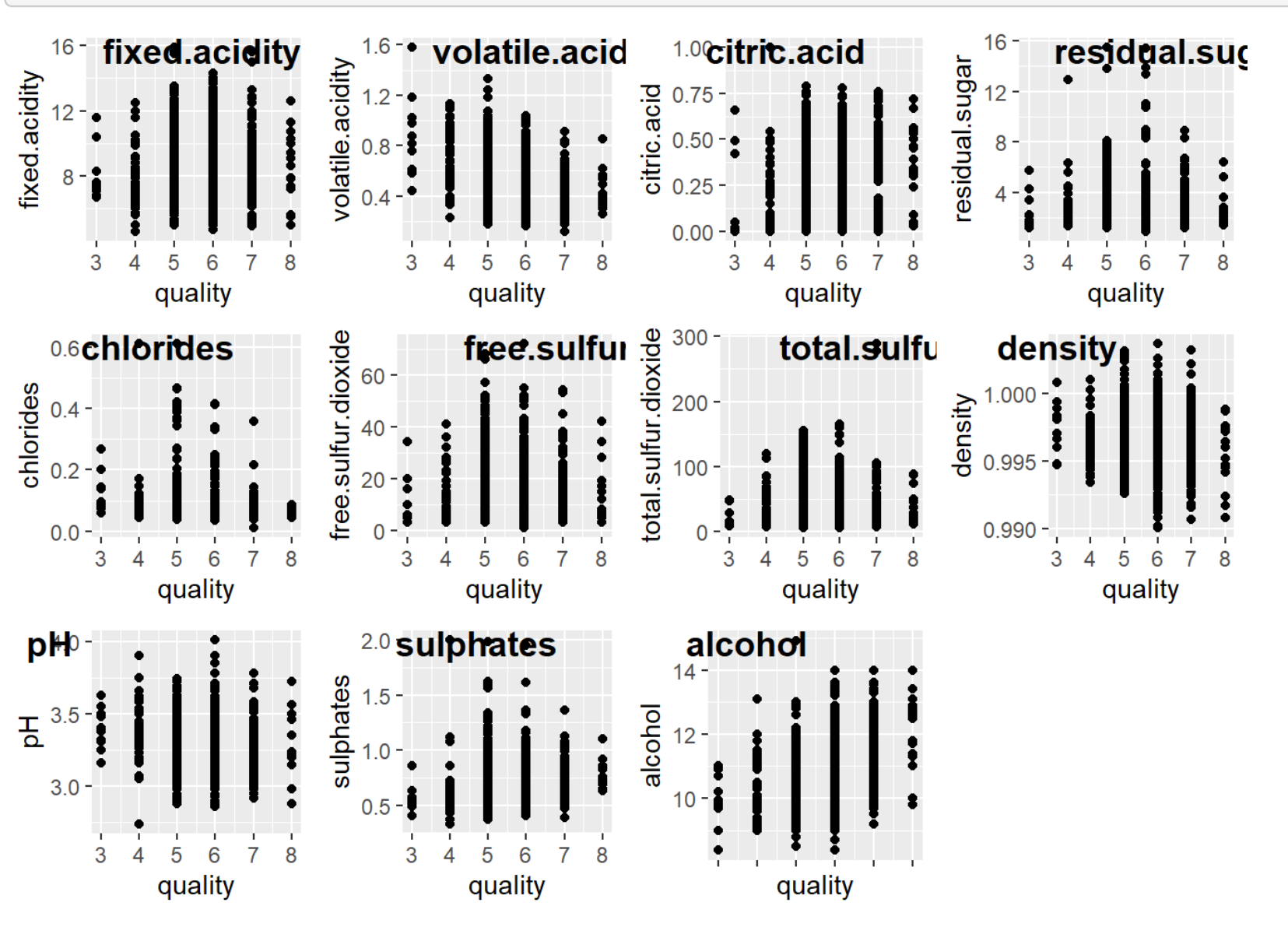
Pel les variables independetns, tot i que les seves cues no segueixen exactament la linea transversal, la major part de les dades semblen estar distribuïdes de forma normal. No obstant, la variable qualitat no sembla seguir aquesta distribució.

4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

```
# Mostrem les distribucions de les diferents variables en funció de la qualitat
fx_ac <- ggplot(aes(x=quality,y=fixed.acidity),data=wine_df)+
  geom_point()

vlt_ac <- ggplot(aes(x=quality,y=volatile.acidity),data=wine_df)+
  geom_point()

ctrc_ac <- ggplot(aes(x=quality,y=citric.acid),data=wine_df)+
  geom_point()


r_sugr <- ggplot(aes(x=quality,y=residual.sugar),data=wine_df)+
  geom_point()


chlor <- ggplot(aes(x=quality,y=chlorides),data=wine_df)+
  geom_point()


f_slr_dox <- ggplot(aes(x=quality,y=free.sulfur.dioxide),data=wine_df)+
  geom_point()


ttl_sl_dox <- ggplot(aes(x=quality,y=total.sulfur.dioxide),data=wine_df)+
  geom_point()


densty <- ggplot(aes(x=quality,y=density),data=wine_df)+
  geom_point()

ph <- ggplot(aes(x=quality,y=pH),data=wine_df)+
  geom_point()


suplh <- ggplot(aes(x=quality,y=sulphates),data=wine_df)+
  geom_point()

alco <-ggplot(aes(x=quality,y=alcohol),data=wine_df)+
  geom_point()


library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.0.5
```

```
ggarrange(fx_ac, vlt_ac, ctrc_ac, r_sugr, chlor, f_slr_dox, ttl_sl_dox, densty, ph, suplh, alco + rremove("x.tex
t"),
          labels = c("fixed.acidity", "volatile.acidity", "citric.acid", "residual.sugar","chlorides", "free.sulf
ur.dioxide","total.sulfur.dioxide", "density","pH", "sulphates", "alcohol"),
          ncol = 4, nrow = 3, heights = c(25, 25))
```



De les anteriors gràfiques en podem destacar que els vins de alta qualitat porten menys volatile-acid, residual.sugar, chlorides, són menys densos, tenen un ph equilibrat, i no tenen més de 1.3 de sulfats.

Proseguim estudiant les diverses correlacions entre variables.

```
# Creem una matriu de correlacions
cor(wine_df)
```

```
##                      fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724
## volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882
## citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162
## residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000
## chlorides               0.09370519      0.061297772  0.20382291    0.055609535
## free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995
## total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882
## density                 0.66804729      0.022026232  0.36494718    0.355283371
## pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422
## sulphates               0.18300566     -0.260986685  0.31277004    0.005527121
## alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437
## quality                 0.12405165     -0.390557780  0.22637251    0.013731637
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.093705186        -0.153794193          -0.11318144
## volatile.acidity      0.061297772        -0.010503827           0.07647000
## citric.acid           0.203822914        -0.060978129           0.03553302
## residual.sugar        0.055609535         0.187048995           0.20302788
## chlorides             1.000000000         0.005562147           0.04740047
## free.sulfur.dioxide   0.005562147         1.000000000           0.66766645
## total.sulfur.dioxide  0.047400468         0.667666450           1.00000000
## density               0.200632327        -0.021945831           0.07126948
## pH                   -0.265026131         0.070377499          -0.06649456
## sulphates             0.371260481         0.051657572           0.04294684
## alcohol              -0.221140545        -0.069408354          -0.20565394
## quality              -0.128906560        -0.050656057          -0.18510029
##                          density          pH    sulphates     alcohol
## fixed.acidity         0.66804729 -0.68297819  0.183005664 -0.06166827
## volatile.acidity      0.02202623  0.23493729 -0.260986685 -0.20228803
## citric.acid           0.36494718 -0.54190414  0.312770044  0.10990325
## residual.sugar        0.35528337 -0.08565242  0.005527121  0.04207544
## chlorides             0.20063233 -0.26502613  0.371260481 -0.22114054
## free.sulfur.dioxide  -0.02194583  0.07037750  0.051657572 -0.06940835
## total.sulfur.dioxide  0.07126948 -0.06649456  0.042946836 -0.20565394
## density               1.00000000 -0.34169933  0.148506412 -0.49617977
## pH                   -0.34169933  1.00000000 -0.196647602  0.20563251
## sulphates             0.14850641 -0.19664760  1.000000000  0.09359475
## alcohol              -0.49617977  0.20563251  0.093594750  1.00000000
## quality              -0.17491923 -0.05773139  0.251397079  0.47616632
##                          quality
## fixed.acidity         0.12405165
## volatile.acidity     -0.39055778
## citric.acid           0.22637251
## residual.sugar        0.01373164
## chlorides            -0.12890656
## free.sulfur.dioxide  -0.05065606
## total.sulfur.dioxide -0.18510029
## density              -0.17491923
## pH                   -0.05773139
## sulphates             0.25139708
## alcohol               0.47616632
## quality               1.00000000
```

De la taula de correlacions anterior, podem veure que entre fixed.acidity i ph hi ha una correlació inversament proporcional del 68.29%. Això ens pot fer pensar que existeix algún tipus de relació entre aquestes variables i que, per tant, no són independents.

Per tal de verificar aquest supòsit, realitzarem un test xi2 on les hipòtesis seràn les següents:

H0 : pH i fixed.acidity són independents H1 : pH i fixed.acidity no són independents

En primer lloc, construim la taula de contingència

```
# Creem la taula de contingència
cont_table <- table(wine_df$pH, wine_df$fixed.acidity)
```

```
# Executem el test xi2
chisq.test(cont_table)
```

```
## Warning in chisq.test(cont_table): Chi-squared approximation may be incorrect
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  cont_table
## X-squared = 20246, df = 8360, p-value < 2.2e-16
```

Els resultats mostren com el p-value és menor a 0.05 i per tant, rebutgem la hipòtesis nul.la. Per tant, podem afirmar que les variables ph i fixed.acidity no són independents.

Finalment, realitzarem una regressó lineal múltiple per tal d'observar com varia la qualitat del vi en funció de les variables independents.

```
# Creem una regressió lienal múltiple
mlr_wine <- lm(quality ~ fixed.acidity + volatile.acidity +citric.acid + residual.sugar + chlorides +free.sulfur.
dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol, data = wine_df )
summary(mlr_wine)
```

```
## 
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid + 
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + 
##     density + pH + sulphates + alcohol, data = wine_df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.68911 -0.36652 -0.04699  0.45202  2.02498 
## 
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           2.197e+01  2.119e+01   1.036   0.3002    
## fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357    
## volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150    
## residual.sugar        1.633e-02  1.500e-02   1.089   0.2765    
## chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *  
## total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density              -1.788e+01  2.163e+01  -0.827   0.4086    
## pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *  
## sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561 
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Tot i que el model no sigui molt acurat, ja que la R2 ajusted és tan sols del 35.61% podem veure que el nivell de sulfats és una de les variables més rellevants a l'hora de determinar la qualitat del vi. També podem veure per altra banda, que el ph determina de forma inversa la qualitat del vi de forma rellevant.

a continuació, podem predir quin resultat tindrà un vi segons les caracteristiques que li imputem

```
# Creem el dataframe amb les noves dades
predict_df = data.frame(fixed.acidity = 7.9,
                    volatile.acidity = 0.67,
                    citric.acid = 0.08,
                    residual.sugar = 2.2,
                    chlorides = 0.071,
                    free.sulfur.dioxide =14,
                    total.sulfur.dioxide = 67,
                    density = 0.998,
                    pH = 3.22,
                    sulphates = 0.60,
                    alcohol = 10.6)



# Predim
predict(mlr_wine, predict_df)
```

```
##        1 
## 5.467345
```

El vi eleborat, no serà apte pels paladars més fins ja que tindrà una qualitat de 5.46 segons el model.