Document amb les 7 qüestions plantejades a la pràctica:

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

Hem seleccionat el fitxer Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). Vivim un època on la cultura del vi creix. Són molts els nous amants del vi, augmenta notablement el turisme enològic amb visites a cellers, tast de diferents vins per conèixer les seves característiques, etc.

Aquest dataset inclou 12 dades importants d'una mostra de 1599 vins. La qualitat (*quality*) que serà el nostre target i 11 característique més:

1. fixed acidity (acidesa fixa). Numèric. g/l.
2. volatile acidity (acidesa volàtil) .Numèric. g/l.
3. citric acid (àcid cítric). Numèric. g/100ml
4. residual sugar (sucre residual). Numèric. g/l.
5. chlorides (clorurs) Quantitat de sal al vi.
6. free sulfur dioxide(Anhídrid sulfurós lliure) Prevenció del creixement microbià i l'oxidació del vi.
7. total sulfur dioxide (Diòxid de sofre total) quantitat de formes lliures + unides de SO2
8. density (densitat). Ppm (parts per milió)
9. pH. Descriu el nivell d'acidesa amb valors de 0 a 14.
10. Sulphates (sulfats) Additiu del vi que contribueix al SO2, antimicrobià i antioxidant. mg/l
11. Alcohol. Es mesura en graus.
12. quality (qualitat) Valors entre 0 i 10. Variable target.

Descripció del problema:

1. Conèixer les característiques més importants per determinar la qualitat del vi negre.
2. Quina és la influència de l'alcohol per establir si un vi es excel·lent o no.

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.

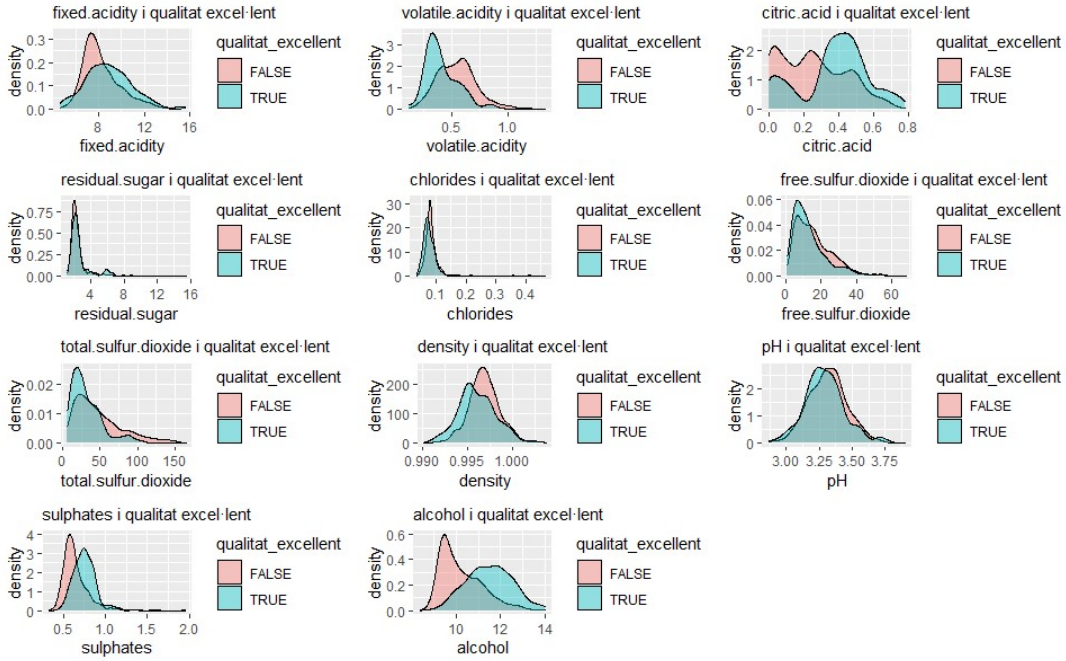
Seleccionem el fitxer winequality-red.csv inclòs a l'enllaç indicat a l'apartat anterior. SI mirem les 10 primeres fileres:

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
1	7.4	0.700	0.00	1.90	0.076	11	34	0.99780	3.51	0.56	9.4	5
2	7.8	0.880	0.00	2.60	0.098	25	67	0.99680	3.20	0.68	9.8	5
3	7.8	0.760	0.04	2.30	0.092	15	54	0.99700	3.26	0.65	9.8	5
4	11.2	0.280	0.56	1.90	0.075	17	60	0.99800	3.16	0.58	9.8	6
5	7.4	0.700	0.00	1.90	0.076	11	34	0.99780	3.51	0.56	9.4	5
6	7.4	0.660	0.00	1.80	0.075	13	40	0.99780	3.51	0.56	9.4	5
7	7.9	0.600	0.06	1.60	0.069	15	59	0.99640	3.30	0.46	9.4	5
8	7.3	0.650	0.00	1.20	0.065	15	21	0.99460	3.39	0.47	10.0	7
9	7.8	0.580	0.02	2.00	0.073	9	18	0.99680	3.36	0.57	9.5	7
10	7.5	0.500	0.36	6.10	0.071	17	102	0.99780	3.35	0.80	10.5	5

El fitxer conté 1.599 ocurrències, prou per fer l'anàlisi i amb les característiques principals d'un vi. Per tant, treballarem amb aquest fitxer per resoldre el nostre objectiu.

Un dels objectius és identificar les qualitats per tenir un vi excel·lent. Per això, creem una nova variable 'qualitat_excellent', binària, amb valor 1 per vins excel·lents, és a dir quality >=7, i 0 per la resta.

Veiem la corba de distribució de les variables segons aquesta nova 'qualitat_excellent'.



Veiem que quan volatile.acidity o totla-sulfure.dioxide tenen valors baixos la qualitat té més probabilitats de ser excel·lent. Tanmateix, també veiem que quan citric.acid, sulphates o alcohol tenen valors alts la qualitat també té més probabilitats de ser excel·lent.

Observem que les variables pH, free.sulfur.dioxide i residual sugar estan molt poc relacionades amb la qualitat i també estan molt poc relacionades amb qualitat excel·lent o no

3. Neteja de les dades.

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

El fitxer no conté valors buits o zeros.

```
colSums(is.na(df))
```

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

```
colSums(df == "")
```

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

Mirem també la possibilitat de dades duplicades:

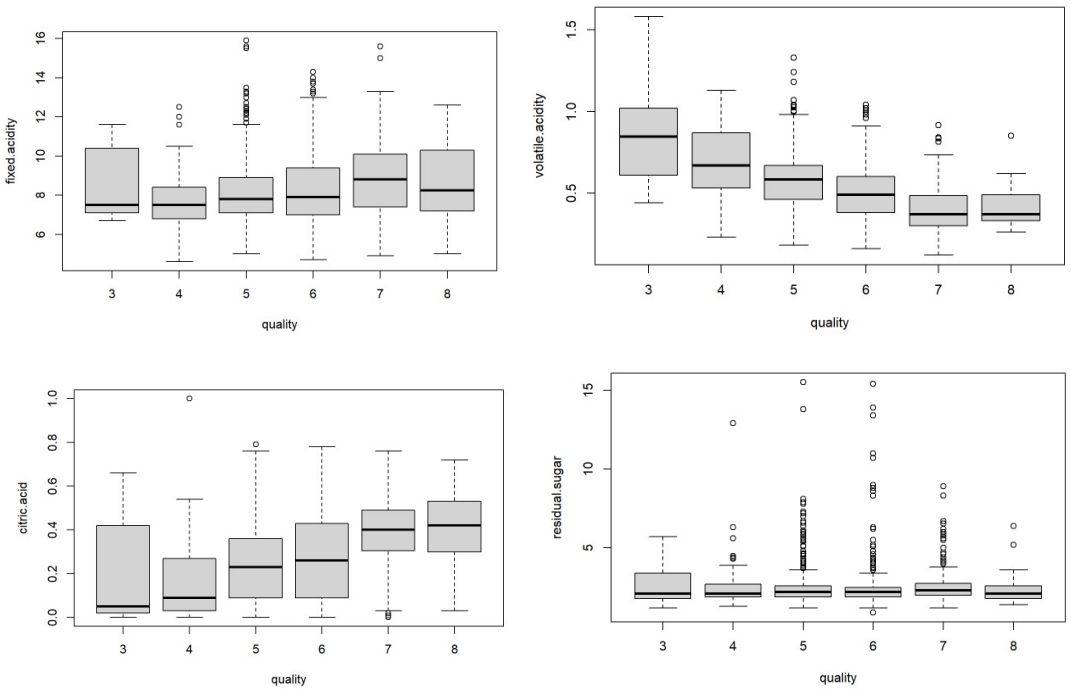
```
sum(duplicated(df))
```

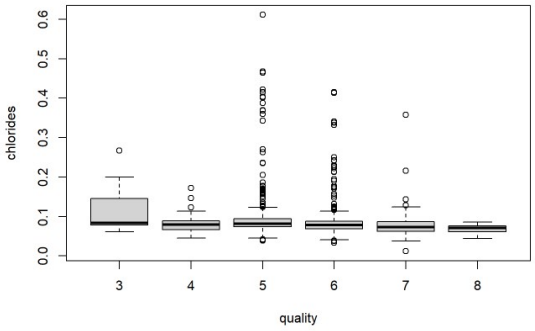
```
## [1] 240
```

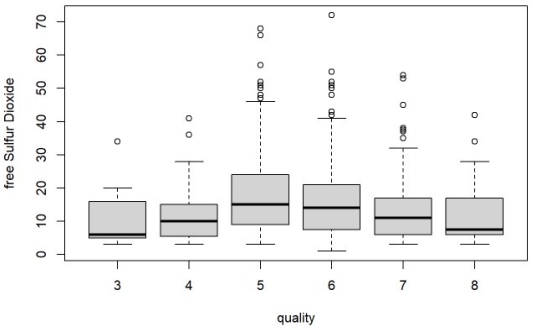
En tenim 240, però és possible que dos vins tinguin exactament els mateixos nivells de cada variable i no tenim cap indicador que ens identifiqui como duplicats, per tant, els mantindrem al fitxer.

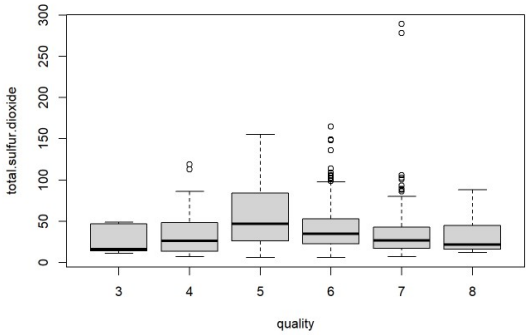
3.2. Identifica i gestiona els valors extrems.

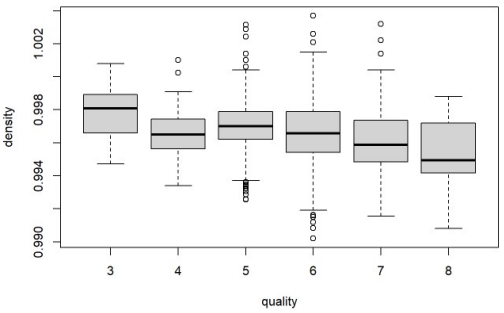
Mitjançant gràfics de caixes I bigotis, dibuixem els extrems per cada valor de 'quality':

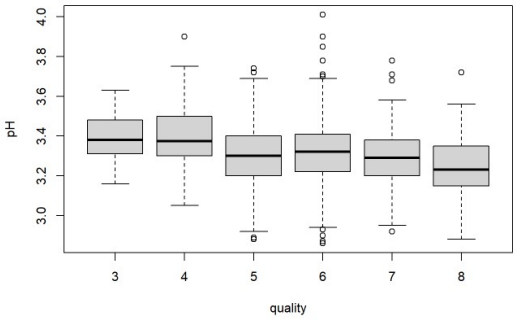


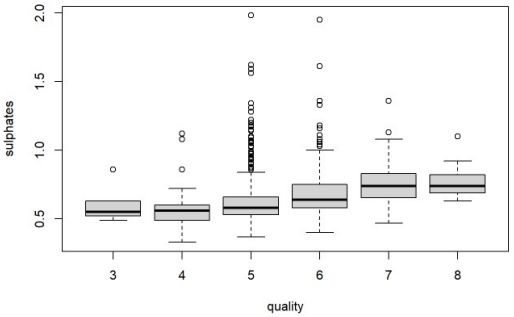


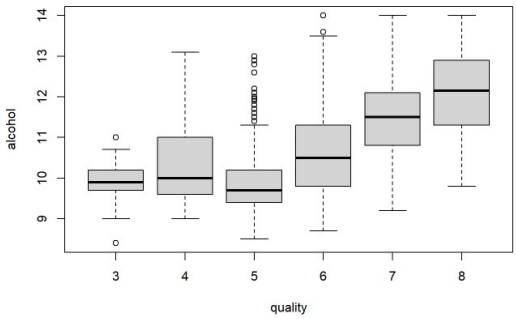












A totes les variables en tenim valors extrems, decidim eliminar aquells que es troben el el 0,01% inferior i superior dels valors de cadascuna de les variables. Així, esborrem 17 valors, passant el dataset a tenir-ne 1582 elements.

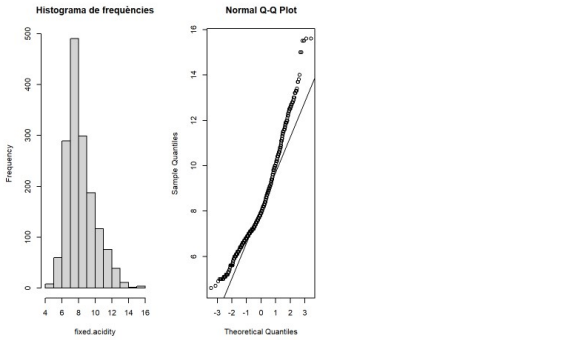
4. Anàlisi de les dades.

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?).

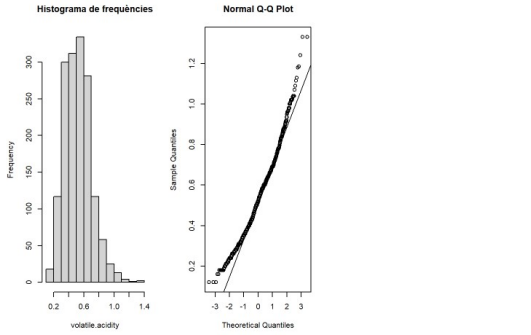
Les preguntes són referides a la qualitat del vi. Analitzarem totes les variables proporcionades pel dataset, principalment per obtenir quines són les més influents en la qualitat.

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

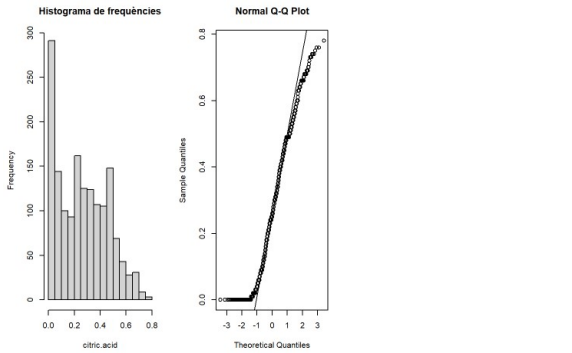
Mitjançant histogràmes, gràfiques Q-Q plot i el test Saphiro-Wilk . Resultats:



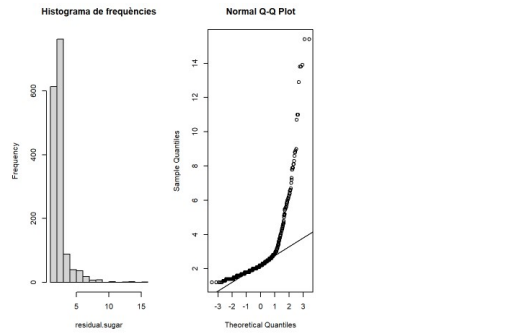

```
## [1] "test de Saphiro-Wilk de la variable  fixed.acidity  és  0.0000000000000000000000574049506505455"
```



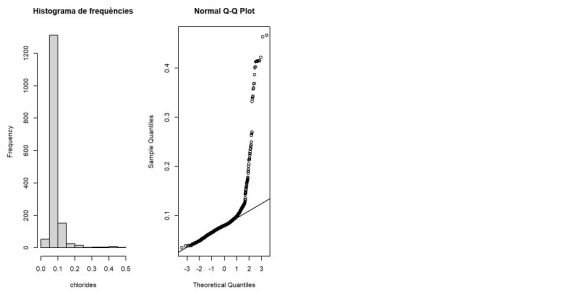

```
## [1] "test de Saphiro-Wilk de la variable  volatile.acidity  és  0.0000000000001786621470250 67"
```



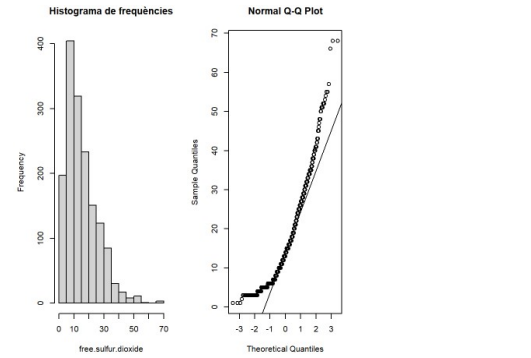

```
## [1] "test de Saphiro-Wilk de la variable  citric.acid  és  0.0000000000000000000070433409703446 8"
```



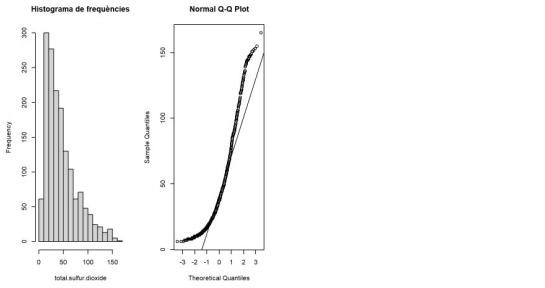

```
## [1] "test de Saphiro-Wilk de la variable  residual.sugar  és  0.0000000000000000000000000000000000000000000000000000000000575909 78574169"
```



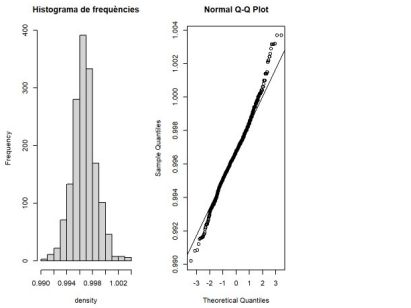

```
## [1] "test de Saphiro-Wilk de la variable  chlorides  és  0.0000000000000000000000000000000000000000000000000000000148713029 97056"
```



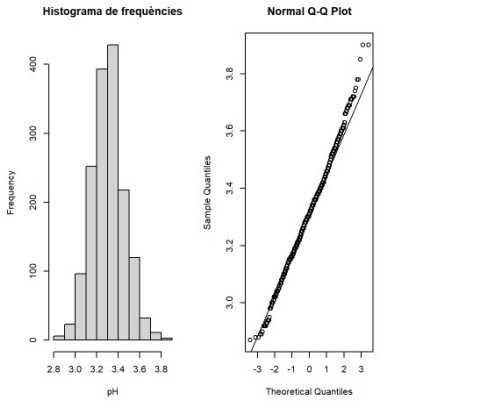

```
## [1] "test de Saphiro-Wilk de la variable  free.sulfur.dioxide  és  0.0000000000000000000000000000000000000000000000000000000000228038871510153"
```



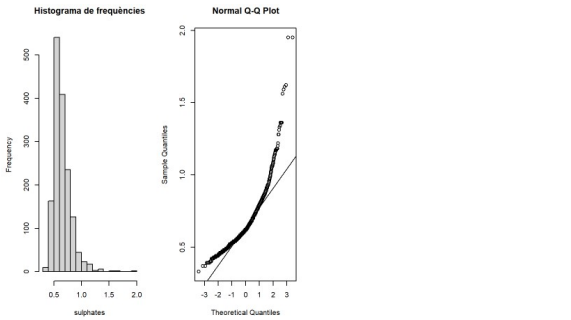

```
## [1] "test de Saphiro-Wilk de la variable  total.sulfur.dioxide  és  0.0000000000000000000000000000000034919732716075 1"
```



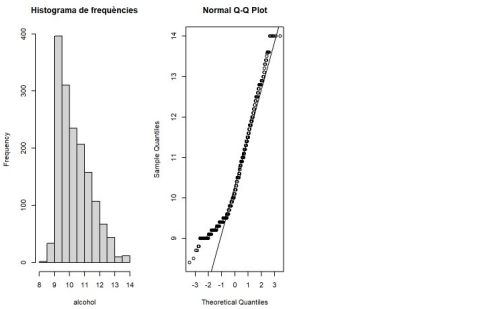

```
## [1] "test de Saphiro-Wilk de la variable  density  és  0.000000198457486312986"
```



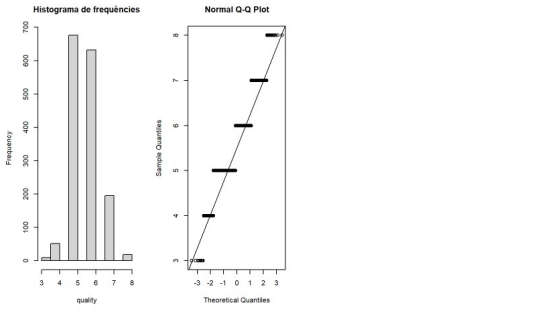

```
## [1] "test de Saphiro-Wilk de la variable  pH  és  0.000456029891682251"
```




```
## [1] "test de Saphiro-Wilk de la variable  sulphates  és  0.000000000000000000000000000000000171902264175559"
```




```
## [1] "test de Saphiro-Wilk de la variable  alcohol  és  0.000000000000000000000000000000810985011996217"
```




```
## [1] "test de Saphiro-Wilk de la variable  quality  és  0.000000000000000000000000000000000103947017448143"
```

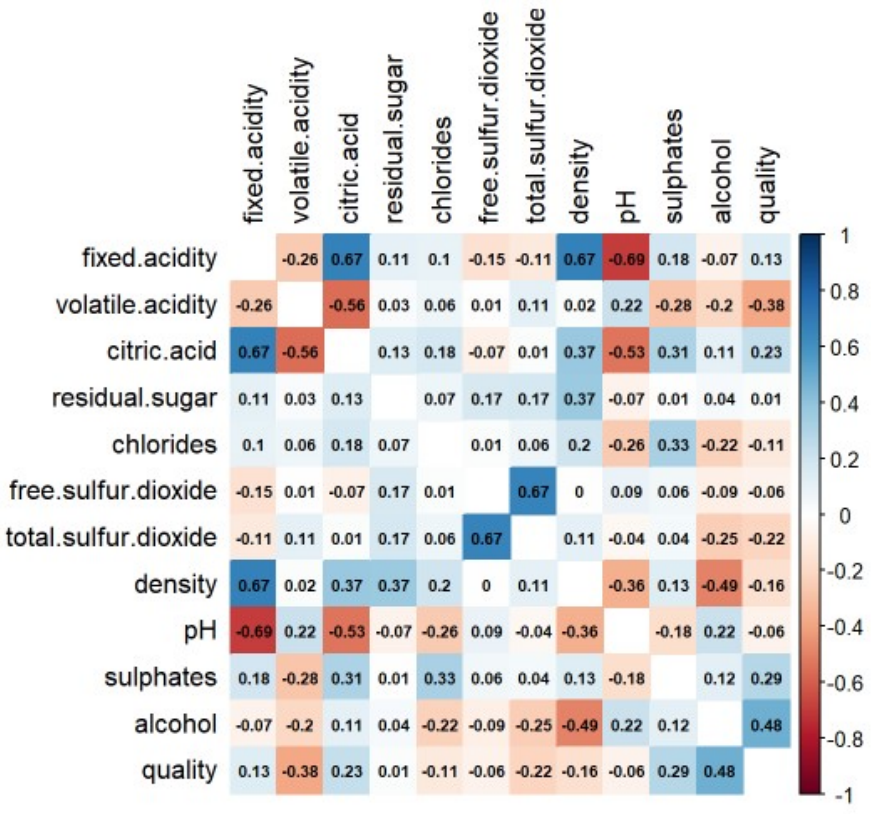
El Q-Q plot ens mostra com la majoria de les varibles s'ajusten a la línia de normalitat, però presenten una distribució exponencial. Els resultats del test Saphiro-Wilk ens donen valors per sona de 0,05 que ens indicarien situació de no normalitat.

Com que la mostra és gran, pel Teorema del Límit Central, podem assumir que la mitjana mostral segueix una distribució normal.

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.

Correlació.

Fem una matriu de correlació amb totes les variables per veure com es relacionen amb el target quality.



Resultat: Les variables més relacionades amb la quality són l'alcohol, volatile.acidity, sulphates, citric.acid i total.sulfur.dioxide .

Hipòtesi de contrast.

Ens farem la següent qüestió: diferència entre la graduació de alcohol per veure si els bons vins tenen més alcohol que els dolents.

Es tracta d'una comparació de mitjanes en poblacions normals independents:

H0 : $\mu 1 - \mu 2 = 0$

H1 : $\mu 1 - \mu 2 > 0$

on $\mu 1$ denota la mitjana de graus alcohol per vi bo i $\mu 2$ la mitjana de graus alcohol per vi dolent.

Per això crearem dos conjunts separats amb les dades d'alcohol per vins bons i una altre per vins dolents. Hem vist la normalitat i calculem que podem assumir la igualtat de variàncies.

```r
# contrast
d <- 0 # Diferència entre mitjanes
t.test(vi_bo, vi_dolent, alternative="greater", mu = d, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  vi_bo and vi_dolent
## t = 17.244, df = 274.6, p-value < 0.0000000000000022
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.145774      Inf
## sample estimates:
## mean of x mean of y
##  11.50900  10.24195
```

El pvalor del test (0) és inferior al nivell de significació (0.05).

Resultat: podem concloure que la mitjana de graduació d'alcohol als vins bons és més gran que als vins dolents.

Regressió

Primer farem un estudi per avaluar un model lineal que expliqui la variable quality en funció de l'alcohol.

```
##
## Call:
## lm(formula = quality ~ alcohol, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8527 -0.4065 -0.1850  0.5164  2.5902
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  1.79226    0.17619   10.17 <0.0000000000000002 ***
## alcohol      0.36914    0.01684   21.93 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7043 on 1580 degrees of freedom
## Multiple R-squared:  0.2333, Adjusted R-squared:  0.2328
## F-statistic: 480.8 on 1 and 1580 DF,  p-value: < 0.00000000000000022
```

La mitjana dels residus és -0,185, molt proper a 0, un dels supòsits per validar el mètode dels mínims quadrats

El coeficient de determinació té un valor 0,2333, això vol dir que el nostre model ens explica només el 23,33% de variància de les observacions.

Afegim les altres 4 variables més relacionades per veure si millora el model

```
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     citric.acid + total.sulfur.dioxide, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75874 -0.37625 -0.05829  0.44337  2.10871
##
## Coefficients:
##                        Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           2.6772111  0.2077515  12.887 < 0.0000000000000002 ***
## alcohol               0.2954504  0.0163043  18.121 < 0.0000000000000002 ***
## volatile.acidity     -1.1032193  0.1140315  -9.675 < 0.0000000000000002 ***
## sulphates             0.8857491  0.1069826   8.279 0.00000000000000261 ***
## citric.acid           0.0143854  0.1037951   0.139                0.89
## total.sulfur.dioxide -0.0026209  0.0005368  -4.883 0.000001152727724423 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6472 on 1576 degrees of freedom
## Multiple R-squared:  0.3543, Adjusted R-squared:  0.3523
## F-statistic:   173 on 5 and 1576 DF,  p-value: < 0.00000000000000022
```

La mitjana dels residus és -0.05829, inferior a l'anterior model (-0,185), molt més proper a 0.

El coeficient de determinació té un valor 0.3543, això vol dir que el nostre model ens explica només el 35,43 % de variància de les observacions. Millora el model fent servir només l'alcohol

Resposta: agafant els valors més influents a la correlació: l'alcohol, volatile.acidity, sulphates, citric.acid i total.sulfur.dioxide, el model ens explicaria el 35,43% de la variància de les observacions.

Model predicitu amb Support Vector Machine

Prenem els valors més influents a la correlació: l'alcohol, volatile.acidity, sulphates, citric.acid i total.sulfur.dioxide i apliquen un model SVM. Aquest model l'apliquen sobre la target binària 'qualitat_excellent'Creem un fitxer de train i test amb el dataset de treball en proporció 80/20.

Aquestes són les dades obtingudes:

```
## Confusion Matrix and Statistics
##
##
## pred   0   1
##    0 290  38
##    1  25  12
##
##               Accuracy : 0.8274
##                 95% CI : (0.7846, 0.8647)
##    No Information Rate : 0.863
##    P-Value [Acc > NIR] : 0.9774
##
##                  Kappa : 0.1804
##
##  Mcnemar's Test P-Value : 0.1306
##
##            Sensitivity : 0.9206
##            Specificity : 0.2400
##         Pos Pred Value : 0.8841
##         Neg Pred Value : 0.3243
##             Prevalence : 0.8630
##         Detection Rate : 0.7945
##   Detection Prevalence : 0.8986
##      Balanced Accuracy : 0.5803
##
##       'Positive' Class : 0
```

La especificitat és molt baixa, el nombre de falsos positius i falsos negatius és massa alt.

Mirem de fer-ho amb la variable quality, on trobem els següents resultats:

```
## Confusion Matrix and Statistics
##
##
## pred   3   4   5   6   7   8
##    3   0   0   0   0   0   0
##    4   0   0   2   1   0   0
##    5   1   5 105  64  12   0
##    6   2   5  53  50  26   3
##    7   0   1   5  11  15   4
##    8   0   0   0   0   0   0
##
## Overall Statistics
##
##               Accuracy : 0.4658
```

```
##                      Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          0.000000 0.000000   0.6364   0.3968  0.28302  0.00000
## Specificity          1.000000 0.991525   0.5900   0.6276  0.93269  1.00000
## Pos Pred Value            NaN 0.000000   0.5615   0.3597  0.41667      NaN
## Neg Pred Value       0.991781 0.969613   0.6629   0.6637  0.88450  0.98082
## Prevalence           0.008219 0.030137   0.4521   0.3452  0.14521  0.01918
## Detection Rate       0.000000 0.000000   0.2877   0.1370  0.04110  0.00000
## Detection Prevalence 0.000000 0.008219   0.5123   0.3808  0.09863  0.00000
## Balanced Accuracy    0.500000 0.495763   0.6132   0.5122  0.60786  0.50000
```

5. Representació dels resultats a partir de taules i gràfiques. Aquest apartat es pot respondre al llarg de la pràctica, sense la necessitat de concentrar totes les representacions en aquest punt de la pràctica.

   S'ha anat presentant a cada apartat.

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

   Amb les proves realitzades podem establir que l'alcohol, volatile.acidity, sulphates, citric.acid i total.sulfur.dioxide són les característiques més influents en la qualitat del vi.

   Sempre haurem de tenir en compte la resta de característiques, amb aquestes 5 no podem pedir resultats excel·lents, el nombre de falsos positius i falsos negatius és elevat.

   Com a conclusió també podem dir que la mitjana d'alcohol als vins bons és més alta que als vins dolents.

7. Codi.

   Al repositori Github creat per aquesta pràctica:

   https://github.com/rpadUOC/prac2tipologiadades

Contribucions	Signatura
Investigació prèvia	RPD / AFG
Redacció de les respostes	RPD / AFG
Desenvolupament del codi	RPD / AFG