Pràctica 2 – Tipologia i cicle de vida de les dades

COMPONENTS DEL GRUP: **Arnau Santos Ribelles** i **Ferran Pintó Haro**.

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

El dataset conté informació relacionada amb les variants negre i blanca de la varietat de vi portuguès "Vinho Verde". Segons els autors del conjunt de dades, per motius de privacitat i logística només hi ha variables fisicoquímiques (entrades) i sensorials (sortida), i no hi ha, per exemple, dades sobre tipus de raïm, marca de vi, preu de venta, etc.

Comptem amb dos datasets (un per la variant blanca i un per la negre) que ajuntem posteriorment. Les 14 variables que conté el dataset són:
"**type**": valor categòric corresponent al tipus de varietat (white / red).
"**fixed acidity**": valor numèric corresponent al la quantitat d'àcids implicats (que no s'evaporen fàcilment).
"**volatile acidity**": valor numèric corresponent a la quantitat d'àcid acètic en el vi.
"**citric acid**": valor numèric corresponent a la quantitat d'àcid cítric. Entre 0 i 1.
"**residual sugar**": valor numèric corresponent a la quantitat (grams/litre) de sucre que queda després de la parada de la fermentació.
"**chlorides**": valor numèric corresponent a la quantitat de sal al vi.
"**free sulfur dioxide**": valor numèric corresponent a la quantitat de la forma lliure del SO2.
"**total sulfur dioxide**": valor numèric corresponent a la quantitat de formes lliures i lligades de SO2.
"**density**": valor numèric corresponent a la densitat del vi.
"**pH**": valor numèric corresponent al PH del vi. Escala de 0 (molt àcid) al 14 (molt bàsic).
"**sulphates**": valor numèric corresponent a la quantitat de sulfats.
"**alcohol**": valor numèric corresponent als graus d'alcohol que té el vi (percentatge d'alcohol).
"**quality**": qualitat percebuda del vi. Valor numèric entre 0 i 10.
"quality_d": variable categòrica creada segons la qualitat del vi ("bo" si la qualitat és 7 o superior i "no bo" si és inferior a 7)

És un dataset rellevant per poder fer tasques de classificació i regressió, al tenir diferents variables numèriques corresponents a les propietats del vi i una variable corresponent a la qualitat percebuda. Així podrem observar quines característiques del vi estan més relacionades amb la qualitat.

És pretén respondre la pregunta següent:
Quines són les propietats fisicoquímiques que fan que un vi sigui bo?

Alhora que observar:
Les propietats que fan que un vi sigui bo són les mateixes per vins negres i blancs?

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.

Com s'ha comentat, originàriament tenim dos datasets, un corresponent a vi negre i un a vi blanc. En aquest pas ajuntem els dos datasets per tenir un únic dataset corresponent a vins.
Abans de fer-ho s'ha creat una nova columna als dos datasets anomenada "type" corresponent al tipus de vi que és: "White" si és vi blanc i "Red" si és negre. Es fa ja que posteriorment ens interessarà saber la varietat de vi que correspon a cada registre, com per exemple, per fer anàlisis per cada tipus de vi.

S'han integrat els dos datasets en un de sol mitjançant una fusió vertical, ja que tenim el mateix format de les bases de dades: mateixes columnes amb els mateixos tipus de variables. Així, tenim un dataset que

inclou els dos tipus de vins, per poder fer anàlisis conjunts de "vins", però permetent fer anàlisis per tipus de vi, segons la variable creada "type".

Ara tenim 6497 files, corresponent a la suma de les files dels dos datasets originals.

3. Neteja de les dades.

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

S'ha comprovat que el dataset no té valors absents (NA) en cap variable.
Al comprovar si les dades contenen zeros, s'ha vist que la columna "citric acid" té 151 zeros, però els considerem vàlids ja que està al rang de valors possibles de la variable: pot no haver-hi àcid cítric als vins.

3.2. Identifica i gestiona els valors extrems.

Hem vist que en gairebé totes les variables hi ha valors extrems (o atípics), i s'han gestionat els que han semblat més atípics.
Per fer-ho s'han creat boxplot de cada variable per observar els valors extrems, veient també la distribució amb l'histograma en algun cas. En algun cas també hem observat els valors dels valors extrems.

- Pel que fa a la variable "fixed acidity", s'observen molts valors atípics però no els considerem anòmals.
- Pel que fa a la variable "volatile acidity", també observem molts valors atípics i no els considerem anòmals.
- Per la variable "cítric acid", eliminem els valors atípics que són superiors a un valor de 1 gram per litre, ja que s'ha trobat en una cerca que la quantitat legal màxima d'àcid cítric en el vi és de 1 gram per litre, i per tant, s'han considerat anòmals els valors atípics a partir d'aquest valor. Dos registres han complert aquesta condició i hem eliminat les files, al ser un percentatge ínfim del total.
- Pel que fa a la variable "residual sugar", hem considerat anòmals els valors per sobre de 30. Hem eliminat 3 registres que complien aquesta condició.
- Per la variable "chlorides", no s'han considerat anòmals cap dels valors atípics detectats.
- Per la variable "free sulfur dioxide", s'han considerat atípics els valors a partir del 150, eliminant una fila que complia la condició.
- Per la variable "total sulfur dioxide" no s'han considerat anòmals cap dels valors atípics observats.
- Per la variable "density", no s'ha observat cap valor extrem.
- Per la variable "pH" no s'han considerat anòmals cap dels valors atípics observats, al ser valors lògics dins l'escala del pH.
- Per la variable "sulphates", no s'han considerat anòmals cap dels valors extrems detectats, al trobar-se que és possible tenir fins a 2 grams per litre de sulfats; i cap valor supera aquest límit.
- Pel que fa a "alcohol", no es consideren anòmals cap dels 3 valors extrems observats.
- Pel que fa a "quality", malgrat observar valors extrems, es troben dins el rang possible de la variable (el rang de la variable va del 0 al 10). Els donem com a valors vàlids.

Cal dir que a l'observar els valors extrems i anar eliminant els registres amb valors considerats anòmals a mesura d'anar-los veient en cada variable, pot ser que en alguna variable no hagin aparegut valors extrems o valors que haguéssim considerat extrems perquè en eliminar files amb el criteri d'una altra variable, hagin estat eliminades files que tenien valors extrems en altres columnes.

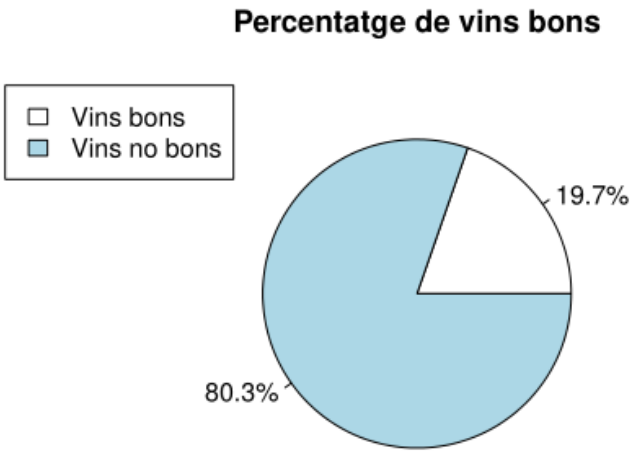
Ara el dataset s'ha reduït a 6491 files. Només hem eliminat 6 files del dataset per valors extrems anòmals, i per aquest motiu hem pogut eliminar els registres: 6 files és un percentatge molt petit del total.

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?).

Abans de començar amb l'anàlisi de dades volem discretitzar la variable qualitat del vi ("quality").

Per fer-ho, substituïm els valors numèrics per etiquetes (bo/no bo). Amb la variable nova podrem interpretar i comparar resultats. Classifiquem els 7 o superior com a "bo" i la resta com a "no bo".



Veiem que el 19.7% dels vins han estat classificats com a vins bons (puntuació 7 o més en qualitat) i el 80.3% com a vins no bons (puntuació inferior a 7).

A part d'altres anàlisis, ens interessarà comparar aquests grups de dades: els vins bons i els no bons. Volem comparar la diferencia en les característiques entre els vins bons i els vins no bons, i per fer-ho farem alguns test d'hipòtesis i una regressió logística.

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

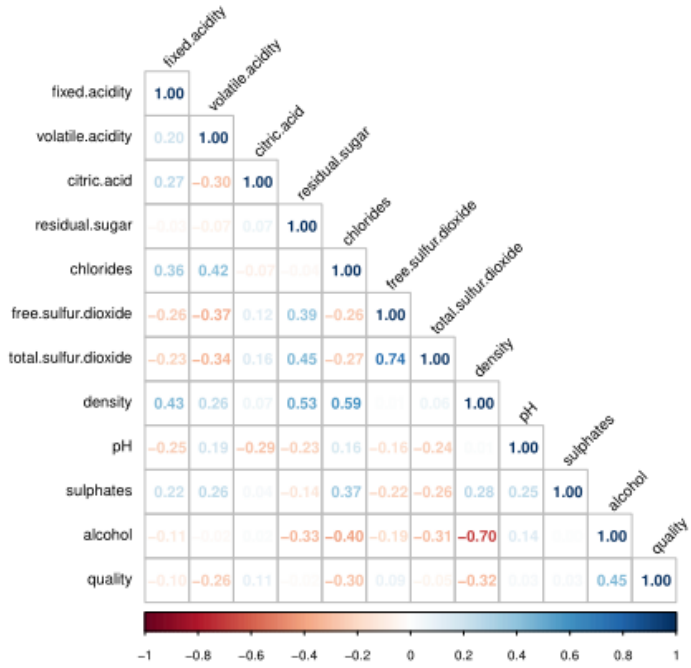
A l'estudiar la normalitat de cada variable, s'ha vist que la normalitat de totes les variables (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol i quality), tenen un valor p inferior a 0.05, que ens port a acceptar que les dades no provenen d'una distribució normal ja que rebutgem la hipòtesi nul·la, en totes les variables.

Pel que fa a la homoscedasticitat, hem contrastat la igualtat de variàncies entre grups
que necessitarem saber posteriorment. Per això s'ha contrastat la igualtat de variàncies entre el vi blanc i el negre pel que fa a la qualitat, i s'ha vist que les variàncies són iguals.

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.

CORRELACIONS

Comencem l'anàlisi realitzant una matriu de correlació entre les variables corresponents a les característiques i la variable quality. La correlació s'ha fet amb Spearman, ja que s'ha vist que cap variable segueix una distribució normal.

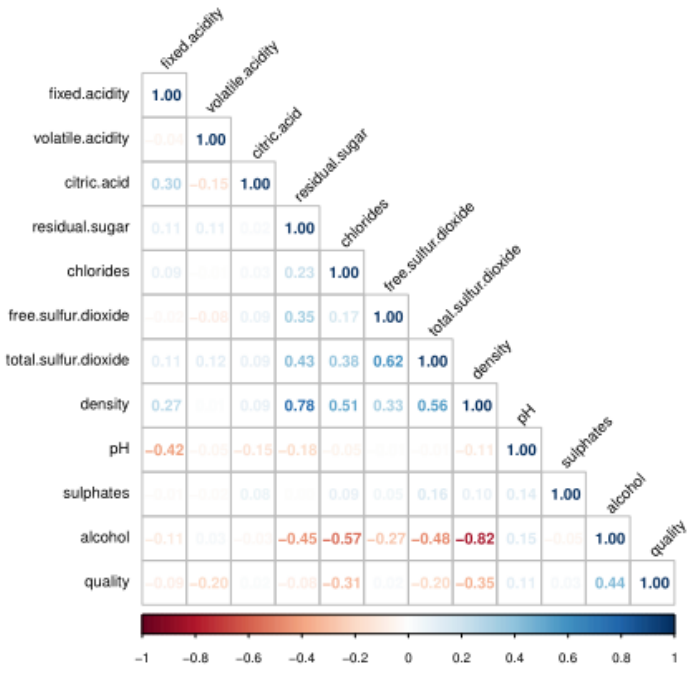


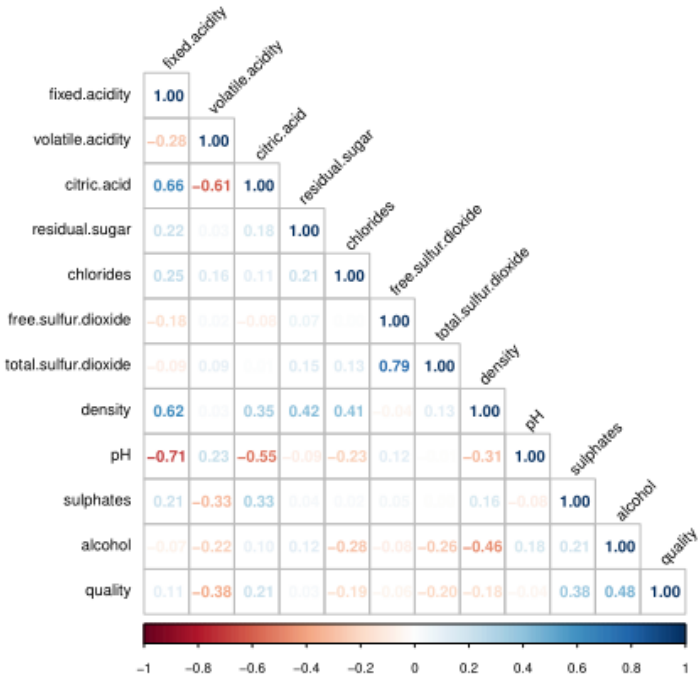
Atenent als coeficients de correlació (method = "spearman"), observem que les relacions lineals més destacables amb la qualitat són amb alcohol (0.45) i density (-0.32). També veiem una relació negativa moderada entre chlorides i la qualitat (-0.30). La resta són relacions dèbils/baixes. Les tres correlacions anomenades són correlacions moderades: la primera implica que a mesura que en certa mesura, a mesura que augmenta l'alcohol, augmenta la qualitat percebuda del vi; la segona implica que a mesura que augmenta la densitat, disminueix la qualitat; i la tercera que a mesura que augmenten els chlorides, disminueix en certa manera la qualitat.

Cal destacar també la relació entre les dues variables següents: la densitat (density) i l'alcohol (alcohol), amb un coeficient de -0.70. Aquesta forta correlació negativa ens fa pensar que degut a la química, la quantitat d'alcohol redueix la densitat, i aquesta relació pot afectar els anàlisis. La quantitat d'alcohol serà la millor opció com a predictor de la qualitat del vi. En anàlisis posteriors, possiblement no sigui útil introduir ambdós predictors als models.

El SO2 lliure i el SO2 total estan altament correlacionats entre si, com podíem esperar.

Ara fem la matriu de correlacions pels vins blancs i pels vins negres, per observar si aquestes correlacions es veuen accentuades en algun dels dos tipus de vins.





Matriu correlacions vi blanc	Matriu correlacions vi negre

Observem que en el vi blanc les correlacions relacionades amb la qualitat són molt similars.
En el cas dels vins negres, la correlació de la densitat amb la qualitat és de -0.18 (molt dèbil), la relació de l'alcohol amb la qualitat augmenta una mica respecte el total (0.48) i apareix una altra correlació mitjana, entre la volatilitat de l'acidesa (volatile acidity) i la qualitat (-0.38), en el sentit que a més acidesa volàtil, menys qualitat. Tampoc té gaire rellevància la correlació entre els chlorides i la qualitat (coeficient de -0.19).

TEST D'HIPÒTESIS

Com que s'ha observat que sembla que l'alcohol és la variable més correlacionada amb la qualitat del vi (tant en els blancs com en els negres), estudiarem ara si la mitjana d'alcohol és diferent pels vins bons i pels vins no bons.

Veient la distribució sembla que els vins bons tenen més alcohol que els vins no bons.

Cal analitzar estadísticament si la mitjana d'alcohol és diferent pels vins bons i els vins no bons.

Farem un contrast d'hipòtesi per la diferència de mitjanes de alcohol.

La pregunta de recerca és: "La quantitat d'alcohol és diferent en els vins bons i els vins no bons?"

- La *hipòtesi nul·la (H0)* és que la mitjana d'alcohol és iguals entre vins bons i dolents.
- La *hipòtesi alternativa (H1)* és que la mitjana d'alcohol és diferents entre vins bons i dolents.

Apliquem un test de dues mostres sobre la mitjana amb variàncies desconegudes.
Pel teorema del límit central podem assumir normalitat. Com s'ha vist al punt 4.2 en la igualtat de variàncies, les variàncies són diferents amb un nivell de confiança del 95%.
El test aplicat és un test de dues mostres independents sobre la mitjana amb variàncies desconegudes diferents.

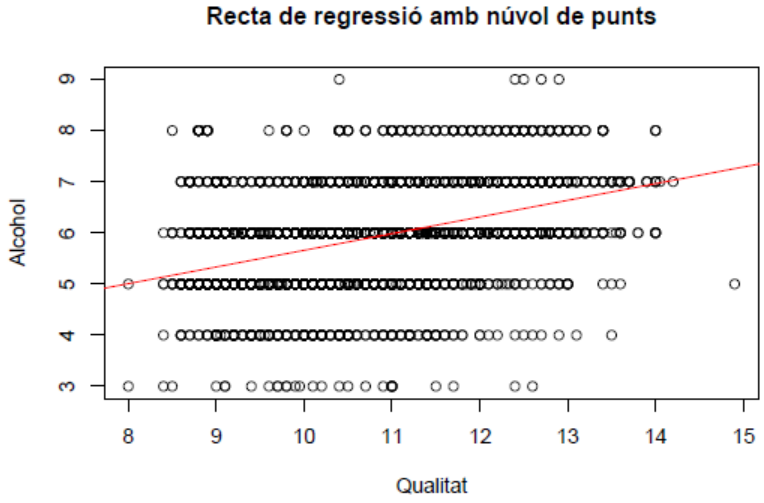
Rebutgem la hipòtesi nul·la del test d'igualtat de mitjanes, afirmant que la mitjana de alcohol és estadísticament diferent entre els vins bons i els no bons. Si veiem les mitjanes, veiem que **la mitjana d'alcohol és més alta en els vins bons (11.43 graus) que en els vins no bons (10.26 graus)**.

Més per curiositat que per resoldre preguntes, ens interessa saber si la qualitat és la mateixa en els vins blancs i els vins negres. Per comprovar-ho, comparem les mitjanes de qualitat entre ambdós tipus de vins. Apliquem el teorema central del límit i considerem que les dades segueixen una distribució normal al tenir mides de les mostres grans (n>30). Pel que fa a l'homoscedasticitat, a l'apartat 4.2 hem vist que les variàncies entre grups (tipus de vi) són iguals pel que fa a la variable qualitat.

De la prova es conclou que existeixen diferències estadísticament significatives entre el vi blanc i el negre en la qualitat percebuda. El vi blanc es percep amb una millor qualitat (5.9 vs 5.6).

REGRESSIÓ

Volem veure amb una regressió lineal simple la relació entre la variable explicativa alcohol i la variable qualitat (variable resposta). El model de regressió resulta significatiu: com havíem vist, a mesura que augmenta l'alcohol, augmenta la qualitat percebuda del vi.



El coeficient de determinació apunta que la variable alcohol és capaç d'explicar el 19.81% de la variabilitat present en la variable de resposta (quality).

En els models de regressió lineal múltiple, totes les variables explicatives ajuden a explicar la variabilitat de la qualitat excepte l'àcid cítric. Al comparar entre els vins blancs i negre, hem vist que hi ha diferents variables que prediuen millor la qualitat del vi.

En el vi blanc, les variables que millor prediuen la qualitat del vi són: l'alcohol, l'acidesa volàtil, el sucre residual, el "free sulfur dioxide", la densitat, el pH, els sulfats i l'acidesa fixe.
En canvi, en el vi negre les variables que millors prediuen la qualitat del vi són: l'alcohol, l'acidesa volàtil, el "free sulfur dioxide", el pH, els sulfats, el "total sulfur dioxide" i la quantitat de sal.

La qualitat del vi blanc és ben predita per varies variables, tres de les quals no prediuen bé la qualitat del vi negre: el sucre residual, la densitat i l'acidesa fixe. En canvi, la quantitat de sal i el "total sulfur dioxide" prediuen bé la qualitat del vi negre però no del blanc.

L'arbre de decisions amb una precisió del 82.6% ens indica que les variables més usades per predir si el vi és bo o no són: l'alcohol, l'acidesa volàtil, l'acidesa fixe, el "free sulfur dioxide" i el sucre residual.....

La regressió logística mostra que totes les variables excepte l'àcid cítric són significatives per predir la probabilitat que un vi sigui bo o no.

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'han anat incloent taules i gràfiques al llarg de la pràctica.

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

Les variables més correlacionades amb la qualitat són l'alcohol i els chlorides (obviant la densitat, al haver observat la seva relació amb l'alcohol), malgrat tinguin correlacions moderades. A més alcohol es perceb més qualitat, i a més chlorides se'n percep menys. Pel que fa al vi negre, l'acidesa volàtil es troba força correlacionada de manera negativa amb la qualitat, els sulfats tenen una correlació positiva amb la qualitat, i l'alcohol també hi té una correlació positiva.

Pel que fa a les correlacions ja veiem que les variables relacionades amb la qualitat són diferents entre els vins blancs i negre: en els blancs l'alcohol i la quantitat de sal són les que tenen més relació, i en els negres, l'alcohol, els sulfats i l'acidesa volàtil.

A l'estudiar en més profunditat la relació entre l'alcohol i ser un vi bo o no, s'ha observat que els vins considerats bons tenen, de mitjana, més graus d'alcohol que els vins que no són bons.

S'ha determinat que els vins blancs són percebuts com de més qualitat que els vins negres.

Totes les variables explicatives excepte l'àcid cítric ajuden a explicar predir la qualitat d'un vi. Al comparar entre els vins blancs i negre, hem vist que hi ha diferents variables que prediuen millor la qualitat del vi.

En el vi blanc, les variables que millor prediuen la qualitat del vi són:

- Alcohol
- Acidesa volàtil
- Sucre residual
- Free sulfur dioxide
- Densitat
- pH
- Sulfats
- Acidesa fixe

En canvi, en el vi negre les variables que millors prediuen la qualitat del vi són:

- Alcohol
- Acidesa volàtil
- Free sulfur dioxide
- pH
- Sulfats
- Total sulfur dioxide
- Quantitat de sal

Tornem a la pregunta/problema que es pretenia resoldre: les característiques (propietats fisicoquímiques) més relacionades amb que un vi sigui bo. L'alcohol és la característica que en tots els anàlisis s'observa com la més relacionada amb la qualitat del vi, i més determinant per a que un vi sigui bo. L'acidesa volàtil sembla ser la segona variable més relacionada amb la qualitat del vi.... Pràcticament totes les variables ajuden a determinar la qualitat dels vins, i a determinar si un vi es bo o no, com s'ha vist en els anàlisis.
La característica menys relacionada amb que un vi sigui bo o no sembla ser l'àcid cítric, al haver estat exclosa en la selecció de la regressió lineal múltiple, al no ser significativa en la regressió logística i a l'observar una correlació molt baixa.

D'altra banda, les propietats fisicoquímiques que determinen la qualitat del vi són diferents en el vi blanc i el vi negre, malgrat hi ha propietats comunes rellevants per ambdues tipologies de vins.

...

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

S'ha adjuntat el codi en R al github, amb el nom de **codi/Prac2 - Qualitat del vi.Rmd,** amb el que s'ha realitzat la neteja, anàlisi i representació de les dades

Pràctica 2 – Tipologia i cicle de vida de les dades

TAULA DE CONTRIBUCIONS

Contribucions	Signatura
Investigació prèvia	FPH, ASR
Redacció de les respostes	FPH, ASR
Desenvolupament del codi	FPH, ASR

Recursos

1. Calvo, M., Pérez, D., Subirats, L. (2019). Introducció a la neteja i anàlisi de dades. Editorial UOC.