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

El dataset està format per diferents paràmetres químics de diferents varietats de vi blanc de la variant del vi portuguès “Vinho Verde”.

El dataset està format pels següent atributs:

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Els estadístics bàsics que descriuen cada columna són els següents:

-	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	4898.0	4898.0	4898.0	4898.0	4898.0	4898.0	4898.0	4898.0	4898.0	4898.0	4898.0	4898.0
mean	6.854787668 44	0.278241118 824	0.334191506 737	6.391414863 21	0.045772356 0637	35.308084932 6	138.36065 7411	0.994027376 48	3.188266639 44	0.489846876 276	10.51426704 78	5.877909350 76
std	0.843868227 688	0.100794548 425	0.121019804 203	5.072057784 01	0.021847968 0937	17.007137325 2	42.498064 5541	0.002990906 91694	0.151000599 615	0.114125833 949	1.230620567 776	0.885638574 968
min	3.8	0.08	0.0	0.6	0.009	2.0	9.0	0.98711	2.72	0.22	8.0	3.0
max	14.2	1.1	1.66	65.8	0.346	289.0	440.0	1.038998	3.82	1.08	14.2	9.0

On observem la quantitat d’atributs de cada valor (no hi ha missing values), la mitja, la desviació estàndard i el mínim i màxim de cada valor.

Per conèixer la qualitat del vi (i per tant, el seu preu), s'han de conèixer els components químics que el formen. A través d'aquest dataset tenim una sèrie de paràmetres químics dels quals al final se n'ha obtingut una qualitat.

Doncs a partir d'aquests paràmetres podem estudiar si alguns depenen d'altres i en quin grau per poder reduir la dimensionalitat dels factors que contribueixen a la qualitat del vi. També es pot buscar factors de dependència entre variables (regressions lineals), etc. També, i donat els paràmetres inicials (factors químics), saber si podem determinar la qualitat sense tastar-lo.

2. Neteja de les dades.

2.1. Selecció de les dades d'interès a analitzar. Quins són els camps més rellevants per tal de respondre al problema?

Per aquest dataset agafarem tots els atributs, ja que tots, a priori, són importants per determinar la qualitat del vi.

2.2. Les dades contenen zeros o elements buits? I valors extrems? Com gestionaries cadascun d'aquests casos?

Les dades no contenen elements buit ni zeros. En aquest cas si hi hagués zeros serien part de la mesura corresponent, és a dir, que la quantitat "d'atribut" que hi hauria en el vi seria zero i, per tant, no es tractaria; es deixaria tal com està.

Si hi hagués elements buits, que tampoc n'hi ha, el substituiríem per la mitja de valors d'aquella columa. D'altres formes de substitució podria ser utilitzar la moda, interpolar el valor, etc.

En aquest cas s'han estudiat els valors extrems amb els quartils:
Si definim el rang interquartil com:

- iqr = IQ3 -IQ1 (on IQn és l'interquantil i n en aquest cas es 3=0.75 i 1=0.25)

Podem definir "els bigotis" superiors i inferiors com:

- maxQuantil = IQ3 + 1.5*iqr
- minQuantil = IQ1 - 1.5*iqr

Tots els valors per sota de minQuantil i per damunt de maxQuantil s'han considerat outliers.
En el codi hi ha opcions per aquests outliers:

- *Esborrar els outliers (per defecte)*
- Cambiar el valor d'aquest outlier per la mitja (per no perdre la fila).
- Deixar-los sense tractar.

En el codi s'han trobat 758 files que contenen outliers que es distribueixen de la següent forma (el primer valor representa el total d'outliers per aquella columna, el segon valor els outliers per damunt del bigoti superior i el segon valor els outliers per davall del bigoti inferior):

fixed acidity --> 119 , 105 , 14
volatile acidity --> 186 , 186 , 0
citric acid --> 270 , 185 , 85
residual sugar --> 7 , 7 , 0
chlorides --> 212 , 201 , 11
free sulfur dioxide --> 50 , 50 , 0
total sulfur dioxide --> 19 , 14 , 5
density --> 5 , 5 , 0
pH --> 75 , 66 , 9
sulphates --> 124 , 124 , 0
alcohol --> 0 , 0 , 0

3. Anàlisi de les dades.

3.1. Selecció dels grups de dades que es volen analitzar/comparar.

Els grups de dades que es seleccionen seran principalment "quality" que conté un conjunt de 10 dades diferents (de l'1 al 10). Llavors, en funció de la normalitat i homogeneïtat de les dades s'aplicaran un estadístics o altres per analitzar/comparar les diferents columnes versus "quality".

3.2. Comprovació de la normalitat i homogeneïtat de la variància. Si és necessari (i possible), aplicar transformacions que normalitzin les dades.

Per comprovar la normalitat de les dades s'ha apliquen els mètodes següents (el programa dóna l'opció de triar):

- Shapiro test del paquet Scipy (basat en el test de Shapiro-Wilk) per cada columna. El p-value retornat per cada valor ha estat inferior a 0.05 el que implica que en cada columna es descarta la hipòtesi nul·la i per tant, la variable no segueix una distribució normal.
- normaltest(...) del paquet Scipy (basat en d'Agostino & Pearson test) en cada columna. El p-value retornat ha estat inferior a 0.05 per cada columna el que implica que es descarta la hipòtesi nul·la. Per tant, la variable no segueix una distribució normal.

Per tant, com que cap columna no passa cap test, no podem acceptar la hipòtesi nul·la i acceptar que vénen d'una població amb distribució normal.

Per comprovar la homogeneïtat de les dades s'ha aplicate el mètode:

- Levene(...) del paquet Scipy (basat en el test de Levene). S'ha aplicat a totes les columnes i ha retornat un p-value inferior a 0.05 el que implica que es descarta l'hipòtesi nul·la i s'assumeix la no homegeneïtat de les variàncies (dades heterogènies).
- El programa dóna l'opció de calcular el test de Fligner-Killeen, un test no-paramètric que serveix per calcular la similitud de variàncies (homegeneïtat) però en aquest cas aquest test és més robust que Levene pel que fa a la suposició de normalitat en les dades.
  Retorna el mateix resultat que Levene.

Per normalitzar les dades (s'empra la normalització basada en l'unitat) s'ha aplicat la següent fòrmula a cada valor per cada columna:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

(on X' es el nou valor, X el valor antic, X_{max} el valor màxim de la columna i X_{min} el valor mínim)

que transforma a cada valor de cada columna entre els valors 0 i 1.

3.3. Aplicació de proves estadístiques (tantes com sigui possible) per comparar els grups de dades.

Per al dataset s'han aplicat els següents estadístics:
:

- Regressió lineal: entre aquelles columnes que tenen el factor de correlació superior a 0.675 (regressió que comença a tenir qualitat).

- Kruskal-Wallis: com que no podem aplicar ANOVA degut a la no normalitat i homogeneïtat de les dades i a més tenim més de tres categories de dades (10 en concret per a la columna quality), s'ha aplicat Kruskall Wallis. El resultat d'aplicar Kruskal-Wallis (versió no-paramètrica de l'ANOVA que no assumeix normalitat) per a totes les columnes vs quality ha retornat un valor p-value inferior a 0.05 el que implica que es descarta l'hipòtesi nul·la. Aquest fet no implica obligatòriament que els grups de dades hagin de venir de diferents poblacions però sí ens diu que totes les columnes son importants per determinar la qualitat del vi.
  S'ha aplicat el Kruskal-Wallis al dataset sense esborrar outliers i reporta el mateixi resultat.

- Regressió múltiple I: s'ha provat de fer una regressió múltiple tal que l'atribut "quality" és la variable dependent i tots els altres atributs independents però s'ha obtingut una correlació de 0,251 pel que es descarta la equació obtinguda.

- Regressió mútiple II: s'ha provat de fer una regressió múltiple amb l'atribut "quality" com a variable dependent i llavors cada possible combinació de dos atributs com independents i també s'ha desestimat les rectes perquè el coeficient de correlació màxim obtingut per a totes les combinacions ha estat 0,202.

- *OPCIONAL*: Hi ha comentat el mètode *applyPCA* que redueix a 2 variables el dataset conservant la següent variabilitat en cada component:
  PCA var : [0.92126858 0.06736344] =98,8% que és un molt bon resultat.

4. Representació dels resultats a partir de taules i gràfiques.

S'han pintat els boxplot (o diagrama de caixes) basat en els quartils on es presenten els outliers que estan per damunt i davall dels bigotis. Es pinta el gràfic abans de tractar els outliers per a que es puguin veure. Es pot trobar els boxplot en la carpeta figures amb la següent llegenda: *nomColumna_BoxPlot.png*

S'han pintat les gràfiques de totes les regressions lineals de cada parell possible de columnes tals que tenen una correlació superior a 0.675. Les rectes obtingudes són les següents:

1. alcohol = -0.884768869402 * density + 0.778421836718
2. density = 0.713700099071 * residual sugar + 0.262145833371

Les gràfiques estan el la carpeta figures del Github amb noms : lineN.png on N es el nombre de recta (en aquest cas 1 ó 2).

Per calcular les rectes s'han utilitzat tots els valors però per pintar la gràfica i que quedi clara s'ha emprat una mostra aleatòria de 30 dades.

S'han calculat totes les equacions de regressió múltiple entre totes les possibles combinacions de 2 columnes (variable independent) i "quality" (com a variable dependent). Finalment s'ha calculat la recta de regressió entre totes les columnes com a variable independents i de "quality" com a variable dependent. S'han escrit els resultats en el fitxer RegressionMultipleValues.txt que està en la carpeta figures. La regressió múltiple amb totes les columnes es troba en últim lloc del fitxer.

També s'han pintat les distribucions normals de cada columna en forma d'histograma junt amb la corba tal que la suma de totes les freqüències és 1. Per pintar-ho s'han estandaritzat els valors de cada columna.

Les gràfiques estan el la carpeta figures del github amb noms : *nomDeLaColumna.png*

També s'ha pintat la gràfica Q-Q plot que permet observar com d'aprop està la distribució d'un conjunt da dades a alguna distribució ideal, en aquest cas sembla que les columnes no segueixen una distribució normal (tal com deia el test de Shapiro).

Les gràfiques estan en la carpeta figures del github amb *nom nomColumna_Q_Q_plot.png*

També es presenten en format taula els estadístics bàsics del conjunt inicial i del conjunt transformat (després de la eliminació d'outliers i de normalitzar-lo). Els conjunts tenen per nom wineStatistics.csv i wineTreatedStatistics.csv., respectivament i estan en la carpeta Datasets.

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

Les conclusions a les que arribem són les següents:

- S'han trobat 758 files amb outliers amb el mètode del boxplot.
- Les columnes del dataset no segueixen cap distribució normal. Tampoc són homogènies.
- El contrast d'hipòtesi amb Kruskal-Wallis entre cada columna i "quality" dóna que cap accepta la hipòtesi nul·la. Per tant, totes les columnes són necessàries per a determinar la qualitat del vi.
- Si no s'esborren els outliers, en contrasts d'hipòtesi (entre totes les columnes amb quality) amb Kruskall-Wallis, la hipòtesi nul·la també es descarta sempre.
- Degut a que totes les columnes són importants per determinar la qualitat del vi s'intenta buscar una regressió múltiple amb "quality" com a variable dependent i la resta de columnes com a variables independents.La recta de regressió múltiple entre totes les columnes i "quality" dóna una correlació molt baixa (0.251).
- Les regressions múltiples entre els possibles parells de columnes i "quality" sempre dóna una correlació molt baixa (0.202 màxim).
- Entre totes les possibles combinacions de equacions de regressió simple de totes les combinacions columnes, només dues combinacions donen una regressió superior a 0.675.

Es conclou doncs, que necessitem totes les variables per determinar la qualitat del vi (segons Kruskal-Wallis) però no podem determinar quina es la equació que determina la qualitat en funció de totes les variables (la regressió múltiple dóna una correlació molt

baixa). Per tant no podem saber la qualitat del vi en funció de les seves variables abans de tastar-lo.

6. 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.