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 i negre de la variant del vi portugués “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)

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 excepte l’últim, la qualitat, ja que aquest ultim atribut es dependent de tots els altres estudiats.

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 la següent fòrmula:

iqr = IQ3 -IQ1 (on IQn és l'interquantil i n en aquest cas es 3=0.75 i 1=0.25)
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 dos opcions per aquests outliers:

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

3. Anàlisi de les dades.

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

S'analitzaran totes les columnes del dataset excepte l'última columna "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 aplicat el mètode:

- 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 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 no-paramètric test 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.

Per normalitzar les dades 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 all 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 grups de dades, 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 excepte per a la "qualitat" 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.

- S'ha calculat els estadístics bàsics de cada columna que es pot trobar en la carpeta Dataset del Github amb noms wineStatistics.csv i wineTreatedStatistics.csv (després de normalitzar les dades)

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

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 empleat una mostra aleatòria de 30 dades.

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 la distribucio normal. Les gràfiques mostren que s'ajusten bastant a la distribució normal.

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

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

???????????

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.