Practica 2
Integración, limpieza, validación y análisis de datos

Carmen Beatriz Mora Gonzalez de la Huebra

1. Descripción del dataset. Por qué es importante y que pregunta/problema pretende responder?

   He seleccionado el dataset "**Wine Quality**" (Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)).

   Este dataset, fue creado por Cortez et at [1] a partir de 1599 muestras de variantes de vino tinto Portugués. Consta de 11 atributos físico-químicos medidos en cada muestra de vino, y una apreciación de calidad, con valores entre 0 (muy malo) y 10 (excelente), (mediana de al menos 3 evaluaciones hechas por expertos en vino):

   1. Acidez fija (fixed acidity)
   2. Acidez volátil (volatile acidity)
   3. Ácido cítrico (citric acid)
   4. Azúcar residual (residual sugar)
   5. Cloruros (chlorides)
   6. Dióxido de sulfuro libre (free sulfur dioxide)
   7. Dióxido de sulfuro total (total sulfur dioxide)
   8. Densidad (density)
   9. pH (pH)
   10. Sulfatos (sulphates)
   11. Alcohol (alcohol)
   12. Apreciación de calidad (quality)

   La pregunta que pretendo responder es si existe alguna relación entre los parámetros físico-químicos y el atributo sensorial de calidad. De ser así, sería muy interesante entender qué parámetros influyen más en la apreciación de calidad y obtener un modelo que represente dicha relación para "predecir" la calidad de un vino en una nueva muestra a partir de parámetros físico-químicos medibles objetivamente.

2. Integración y selección de los datos de interés a analizar.

   El dataset es auto contenido, por lo que no requiere integración con otros datos. Respecto a la selección de datos, a priori no planteo eliminar ningún atributo por razones de calidad de los datos, sin embargo, a partir del análisis de correlación entre atributos busco identificar atributos redundantes a fin de excluir alguno(s) de ellos como atributos predictivos a la hora de proponer un modelo mediante regresión múltiple lineal. También usaré la correlación de las variables físico-químicas con la variable de apreciación de calidad para identificar si hay alguna variable físico-química que no presente relación lineal con el atributo a predecir y de ser así, considerar excluirla en el modelaje.

3. Limpieza de los datos.

Como un primer paso para explorar los datos he procedido a realizar un resumen de valores estadísticos que me den una idea general del rango de valores presentes en cada grupo de atributos: mínimo, cuartiles, máximos, media y la desviación estándar:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599,000	1599,000	1599,000	1599,000	1599,000	1599,000	1599,000	1599,000	1599,000	1599,000	1599,000	1599,000
mean	8,320	0,528	0,271	2,539	0,087	15,875	46,468	0,997	3,311	0,658	10,423	5,636
std	1,741	0,179	0,195	1,410	0,047	10,460	32,895	0,002	0,154	0,170	1,066	0,808
min	4,600	0,120	0,000	0,900	0,012	1,000	6,000	0,990	2,740	0,330	8,400	3,000
25%	7,100	0,390	0,090	1,900	0,070	7,000	22,000	0,996	3,210	0,550	9,500	5,000
50%	7,900	0,520	0,260	2,200	0,079	14,000	38,000	0,997	3,310	0,620	10,200	6,000
75%	9,200	0,640	0,420	2,600	0,090	21,000	62,000	0,998	3,400	0,730	11,100	6,000
max	15,900	1,580	1,000	15,500	0,611	72,000	289,000	1,004	4,010	2,000	14,900	8,000

a. Los datos contienen ceros o elementos vacíos? Cómo gestionarías cada uno de estos casos?

Los datos no contienen elementos vacíos. Sí aparecen valores ceros que corresponden a valores válidos de datos en el atributo *citric acid*. La especificación del dataset origen indica que no tenemos valores faltantes con la sentencia: *Missing Attribute Values: None*

b. Identificación y tratamiento de valores extremos.

Se puede observar que existen valores extremos en atributos “sulfur dioxide” (free y total), pero dichos valores corresponden a valores altos de calidad, por lo que podrían ser válidos y considero que no procede tratarlos.

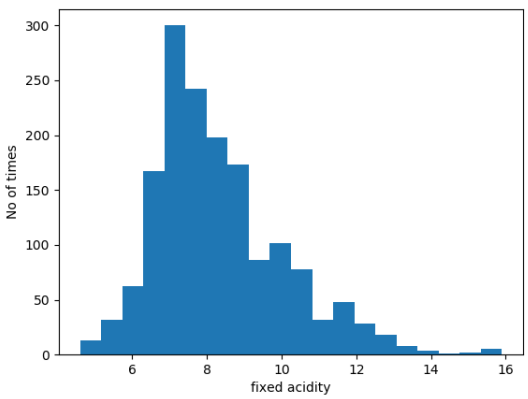
4. Análisis de los datos.

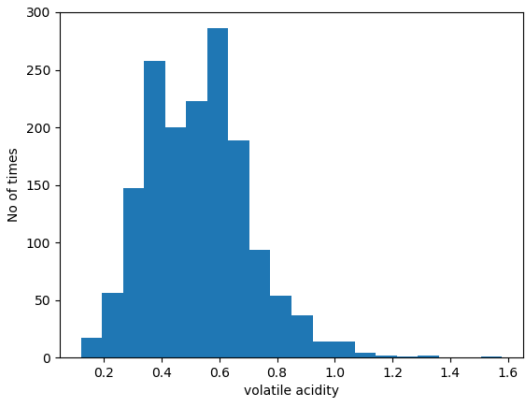
a. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

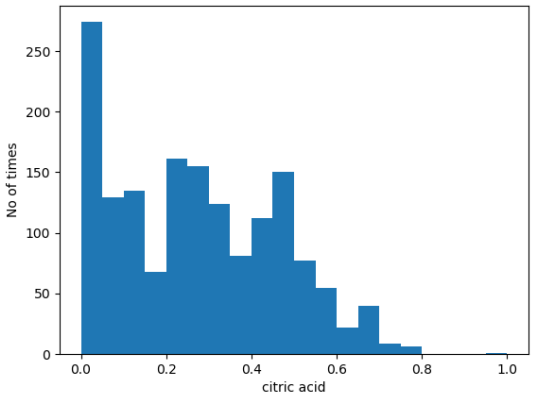
He planificado usar el análisis de correlación entre atributos para identificar si hay atributos dependientes entre si y, si fuera el caso, descartar los atributos redundantes en la construcción del modelo. Utilizaré regresión múltiple para modelar la relación entre los atributos físico-químicos y la apreciación de calidad.

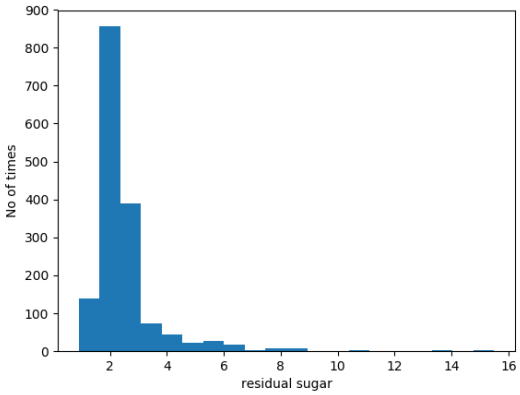
b. Comprobación de la normalidad y homogeneidad de la varianza.

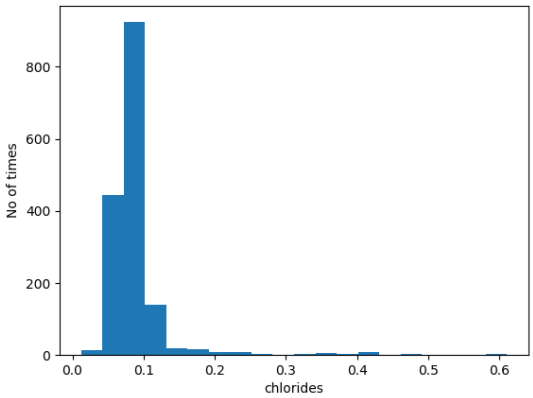
Para este apartado, he utilizado como herramienta los histogramas de todos a las variables involucradas y que se muestran en las figuras a continuación.

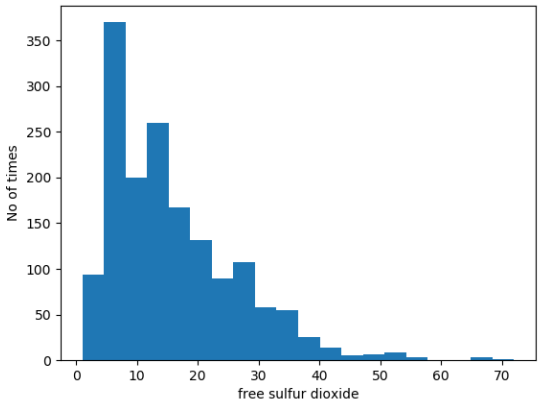


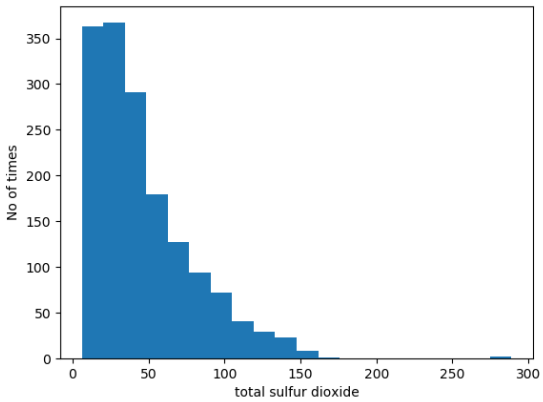


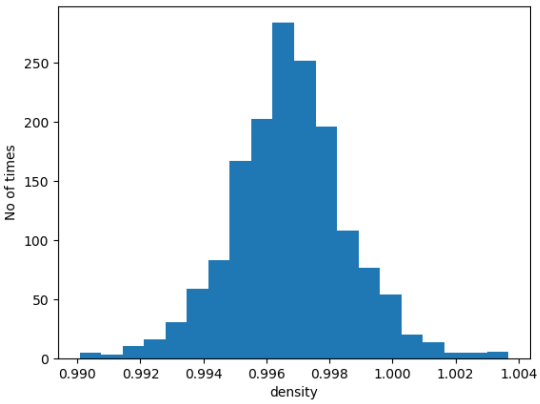

No of times
citric acid

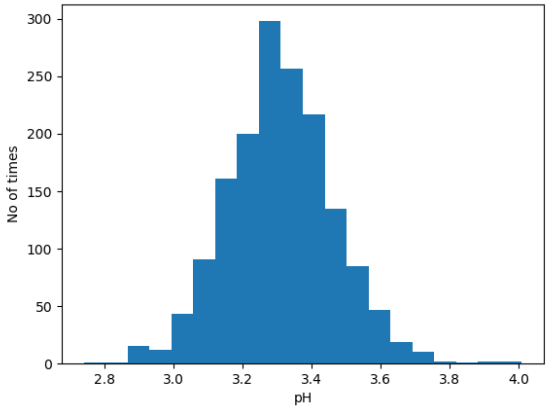

No of times
residual sugar

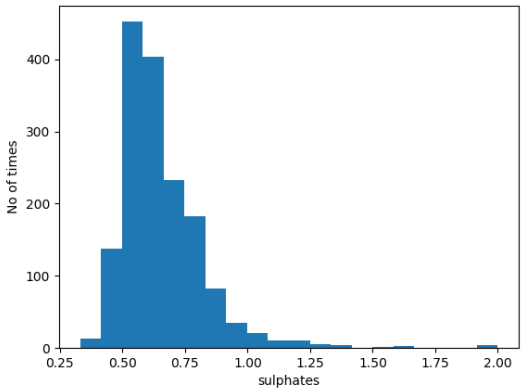

No of times
chlorides

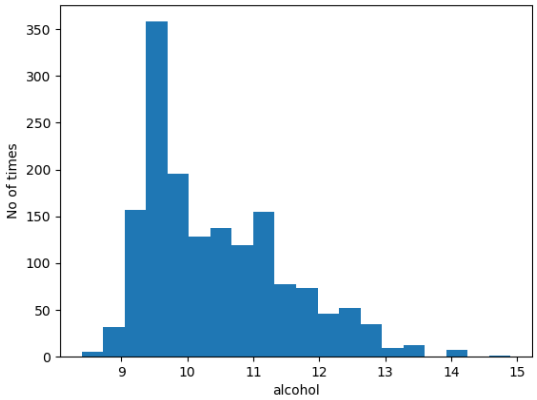

No of times
free sulfur dioxide

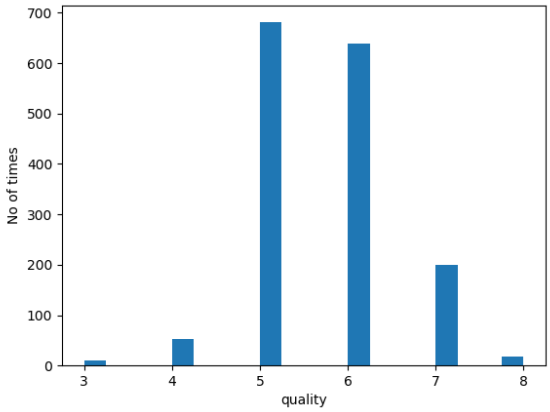

No of times
total sulfur dioxide


No of times
density









Podemos hacer las siguientes observaciones:

- El atributo de calidad, muestra una distribución normal, con picos en valor de calidad 5.
- Los atributos pH y densidad muestra una distribución normal bastante clara.
- Las variables free sulfur dioxide y total sulfur dioxide, muestran una distribución exponencial, una transformación (logarítmica posiblemente) haría más normal su distribución.

c. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

Para ayudarme a entender como cada atributo se relaciona con cada otro he procedido a calcular la correlación entre las variables presentes en el dataset. Con ello busco identificar:

- Variables fisico-quimicas correlacionadas entre sí a fin de plantear la posibilidad de descartar alguna de ellas en la predicción
- Variables que presenten una relación no lineal con la variable a predecir, lo cual sería otro criterio para no incluirlas en el modelo

El siguiente cuadro muestra los valores de correlación de todas las variables físico-quimica vs. el atributo de calidad, ordenados de mayor a menor (valor absoluto) y resaltando los máximos y mínimos:

	quality
alcohol	0,476
volatile acidity	-0,391
sulphates	0,251
citric acid	0,226
total sulfur dioxide	-0,185
density	-0,175
chlorides	-0,129
fixed acidity	0,124
pH	-0,058
free sulfur dioxide	-0,051
residual sugar	0,014

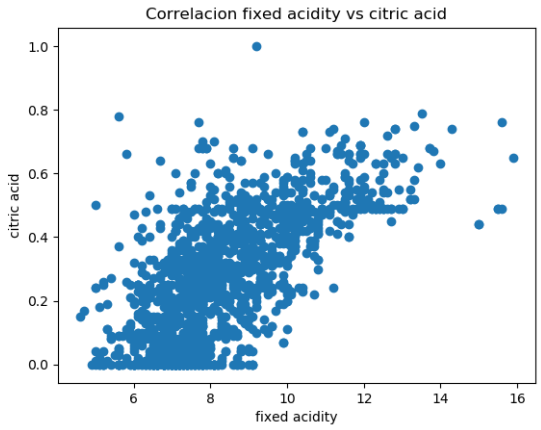
La matriz de correlación para los atributos físico-químicos:

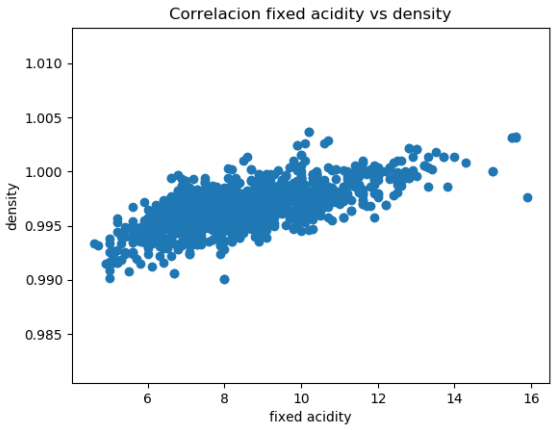
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
fixed acidity		-0,256	0,672	0,115	0,094	-0,154	-0,113	0,668	-0,683	0,183	-0,062
volatile acidity	-0,256		-0,552	0,002	0,061	-0,011	0,076	0,022	0,235	-0,261	-0,202
citric acid	0,672	-0,552		0,144	0,204	-0,061	0,036	0,365	-0,542	0,313	0,110
residual sugar	0,115	0,002	0,144		0,056	0,187	0,203	0,355	-0,086	0,006	0,042
chlorides	0,094	0,061	0,204	0,056		0,006	0,047	0,201	-0,265	0,371	-0,221
free sulfur dioxide	-0,154	-0,011	-0,061	0,187	0,006		0,668	-0,022	0,070	0,052	-0,069
total sulfur dioxide	-0,113	0,076	0,036	0,203	0,047	0,668		0,071	-0,066	0,043	-0,206
density	0,668	0,022	0,365	0,355	0,201	-0,022	0,071		-0,342	0,149	-0,496
pH	-0,683	0,235	-0,542	-0,086	-0,265	0,070	-0,066	-0,342		-0,197	0,206
sulphates	0,183	-0,261	0,313	0,006	0,371	0,052	0,043	0,149	-0,197		0,094
alcohol	-0,062	-0,202	0,110	0,042	-0,221	-0,069	-0,206	-0,496	0,206	0,094	
quality	0,124	-0,391	0,226	0,014	-0,129	-0,051	-0,185	-0,175	-0,058	0,251	0,476

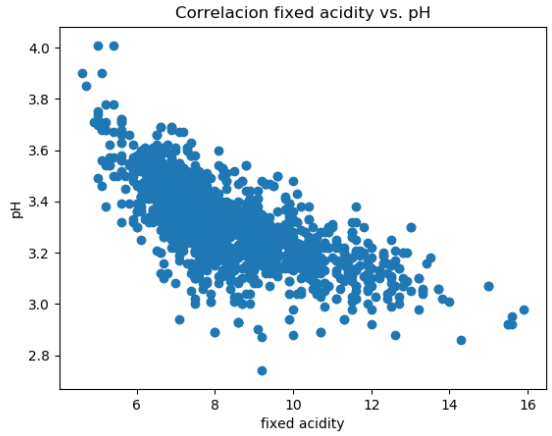
De estos resultados iniciales de análisis de correlación podemos extraer las siguientes observaciones:

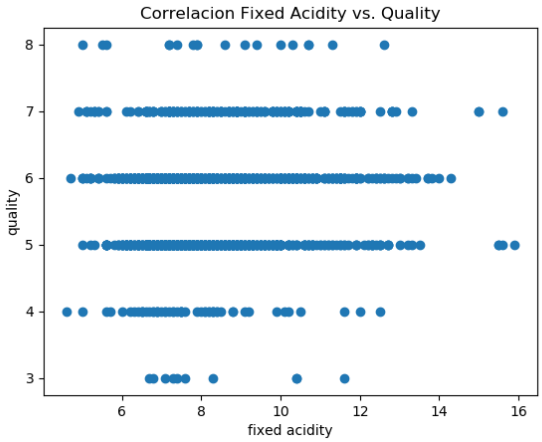
- No observo ningún atributo físico-químicos que por sí solo tenga una alta correlación (positiva o negativa) con el atributo sensorial de calidad.
- Los atributos que tienen una mayor relación lineal con el atributo de calidad son: *alcohol* (corr= 0,476), y *volatile acidity* (corr= -0.390).
- Los atributos que tienen menor relación lineal con el atributo de calidad son *residual sugar* y *free sulfur dioxide*.
- El atributo *fixed acidity* está medianamente correlacionado con *pH* (corr = 0.68) y *citric acid* (corr= 0,67) por lo que posiblemente la información que aporte sea redundante y no sea útil introducir esta variable como predictor en el modelo.
- Asimismo, el atributo *free sulfur dioxide* esta medianamente correlacionado con *total sulfur dioxide*, adicionalmente este atributo presenta una baja correlación con el atributo

cuantitativo de calidad que queremos predecir, por lo cual también sería un atributo candidato a excluir como atributo predictor.









Como próximo paso he procedido a realizar el modelaje de los datos usando regresión lineal múltiple (OLS de la librería **statsmodels** de Python). Para ello he descartado los atributos *fixed acidity* y *free sulfur dioxide* por las razones expuestas anteriormente, obteniendo el siguiente resultado:

volatile acidity	-1.133637
citric acid	-0.203062
residual sugar	0.009834
chlorides	-1.910650
total sulfur dioxide	-0.002378
density	4.582765
pH	-0.521500
sulphates	0.898146
alcohol	0.298361

Como ejercicio de comprobación he procedido realizado el cálculo del modelo sin extraer ningún atributo, en el modelo resultante obtengo pesos muy bajos en los atributos que he descartado en el modelaje inicial. Para el resto de los atributos, obtenemos resultados muy similares a los obtenidos extrayendo los atributos que suponía redundantes o no correlacionados con el de calidad:

fixed acidity	0.004194 ←
volatile acidity	-1.099743
citric acid	-0.184146
residual sugar	0.007071
chlorides	-1.911419
free sulfur dioxide	0.004548 ←
total sulfur dioxide	-0.003319
density	4.529146
pH	-0.522898
sulphates	0.887076
alcohol	0.297023

He procedido a calcular el valor del coeficiente de determinación R-squared (el cual mide cuan bien la regresión se aproxima a los datos reales), obteniendo 0.987 para el primer caso (descartando atributos *fixed acidity* y *free sulfur dioxide*) y 0.987 en el segundo (sin descartar atributos). Siendo los 2 valores idénticos, no me queda claro que obtenga mejoría en el modelo al descartar atributos (al menos con el algoritmo OLS utilizado).

5. Representación de los resultados a partir de tablas y gráficas.

Para un análisis más visual, además de los histogramas de cada variable, he realizado los scatterplots de cada variable contra todas las demás. El conjunto completo de gráficos están incluidos en la carpeta "*Fig*" del repositorio.

La siguiente tabla muestra métricas del modelo obtenido usando la función *summary* disponible en la librería *statsmodels*:

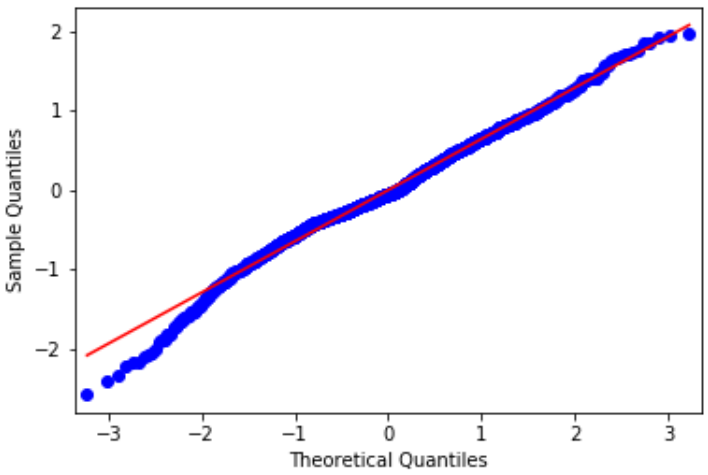
OLS Regression Results

Dep. Variable:	quality	R-squared:	0.987
Model:	OLS	Adj. R-squared:	0.987
Method:	Least Squares	F-statistic:	1.352e+04
Date:	Mon, 11 Jun 2018	Prob (F-statistic):	0.00
Time:	02:18:44	Log-Likelihood:	-1572.0
No. Observations:	1599	AIC:	3162.
Df Residuals:	1590	BIC:	3210.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
volatile acidity	-1.1336	0.115	-9.846	0.000	-1.359	-0.908
citric acid	-0.2031	0.123	-1.654	0.098	-0.444	0.038
residual sugar	0.0098	0.012	0.821	0.412	-0.014	0.033
chlorides	-1.9107	0.403	-4.742	0.000	-2.701	-1.120
total sulfur dioxide	-0.0024	0.001	-4.579	0.000	-0.003	-0.001
density	4.5828	0.459	9.979	0.000	3.682	5.484
pH	-0.5215	0.133	-3.924	0.000	-0.782	-0.261
sulphates	0.8981	0.111	8.122	0.000	0.681	1.115
alcohol	0.2984	0.017	17.398	0.000	0.265	0.332

Omnibus:	25.692	Durbin-Watson:	1.753
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37.304
Skew:	-0.168	Prob(JB):	7.93e-09
Kurtosis:	3.669	Cond. No.	1.74e+03

Una vez obtenido el modelo, también he realizado el Q-Q plot (quantile-quantile plot) como una ayuda grafica para comprobar la distribución normal de la variable dependiente en nuestro modelo:



6. Resolución del problema. A partir de los resultados obtenidos, cuales son las conclusiones? Los resultados permiten responder al problema?

   Trabajar con variables de un dominio de datos con los que no estamos familiarizados y en los que desconocemos a priori sus valores de referencia, dificulta en gran medida el análisis previo de los datos.

   Antes de realizar esta práctica daba por sentado que la complejidad mayor estaría en el modelaje. El mayor esfuerzo ha estado dedicado a analizar y entender los datos, más que en la realización del modelo predictivo en sí.

   Inicialmente esperaba obtener coeficientes mayores para los atributos que presentaban mayor correlación con el atributo de calidad (por ejemplo alcohol). También esperaba que el descartar atributos medianamente correlacionados tuviera un mayor impacto en indicadores del modelaje como el coeficiente de determinación.

   La diferencia de escalas de en los distintos atributos es un punto para haber estudiado mas. Podría haber normalizado los datos para utilizar las mismas escalas y que, de esta manera, los coeficientes obtenidos si fuesen más representativos del peso relativo de cada atributo, pero al normalizar los datos dificultaría la interpretación de su significado físico-quimico al estar estos en una escala distinta a la observada en las mediciones.

   Sería interesante repetir el análisis aquí presentado con otro set de datos, por ejemplo, el de vino blanco de los mismos autores. Comprobar si se mantienen los valores altos o bajos de correlación entre los mismos atributos.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.
   Este trabajo ha sido realizado utilizando el lenguaje y librerías escritas en Python. Todos los gráficos y datos expuestos pueden reproducirse mediante la ejecución de las fuentes que he incluido en la carpeta *Src*.

8. Referencias
   [1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
   Modeling wine preferences by data mining from physicochemical properties.
   In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.
   [2] Wes McKinney (2017, 2nd Edition). Python for Data Analysis. O'Reilley Media, Inc.
   [3] Joel Grus (2015). Data Science from Scratch. O'Reilley Media, Inc.