Tipología y Cilco de Vida de los Datos

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

La presente práctica ha sido desarrollada por: **Leandro Pájaro Fuentes**

Introducción

El vino es la bebida que se obtiene de la fermentación alcohólica total o parcial, del zumo de uvas maduras.

El vino esta formado por diferentes componentes, de los cuales el principal es el agua, que esta presente entre un 82% y un 88%. El segundo componente más importante es el alcohol, que surge gracias a la fermentación, y le da cuerpo y aroma al vino. La graduación del vino suele variar entre el 7% y el 17%, dependiendo del tipo de vino. El resto de componentes aparecen en menor cantidad, como azucares, influyen en el sabor del vino; taninos, que le dan color y textura al vino; sustancias volátiles, que constituyen parte del aroma; ácidos, que participan también en el sabor del vino; y algunos otros de menor importancia.

Todos estos componentes son los que hacen que cada vino sea diferente, pero la cantidad en la que aparecen estos en el vino, se debe sobre todo al clima, al suelo, y a la vid que da las uvas. Estos factores, influyen en la calidad de la uva, y por consiguiente, en que los componentes aparezcan en una cantidad u otra, y, por supuesto, en la calidad final del vino. Tomado de: https://touryvino.com

Descripcion del conjunto de datos

El conjunto de datos seleccionado corresponde a la variente roja del vino Portugués "Vinho Verde". Por cuestiones de privacidad y logística, sólo se dispone de variables fisicoquímicas (de entrada) y sensoriales (de salida) (por ejemplo, no hay datos sobre los tipos de uva, la marca del vino, el precio de venta del vino, etc.). Fuente

Con este conjunto de datos se pretende realizar un análisis para determinar si a partir de las medidas de caracterísitcas de ciertos componentes del vino es posible determinar la calidad de esto.

Referencia:

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, 2009.

Selección de los datos

Carga del conjunto de datos

El conjunto de datos es cargado desde repositorio Github.

```
import pandas as pd
import numpy as np
import scipy as sc
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
from scipy import stats
from scipy.stats import pearsonr
warnings.filterwarnings('ignore')
datos_vino = pd.read_csv('https://raw.githubusercontent.com/lpajarof/redwine/main/Archivos/wi
```

Revisión de las columnas del conjunto de datos

```
datos_vino.columns
```

```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

Las variables del conjunto de datos correspondientes a cada columna son las siguientes:

Variables de entrada:

1. fixed acidity: La mayoría de los ácidos que intervienen en el vino
2. volatile acidity: Cantidad de ácido acético en el vino, que en niveles demasiado altos puede dar lugar a un sabor desagradable a vinagre
3. citric acid: Ácido cítrico, que se encuentra en pequeñas cantidades, puede añadir "frescura" y sabor a los vinos
4. residual sugar: Cantidad de azúcar que queda después de la fermentación, es raro encontrar vinos con menos de 1 gramo/litro y los vinos con más de 45 gramos/litro se consideran dulces
5. chlorides: Cantidad de sal en el vino
6. free sulfur dioxide: Forma libre de SO2 existe en equilibrio entre el SO2 molecular (como gas disuelto) y el ion bisulfito; impide el crecimiento microbiano y la oxidación del vino
7. total sulfur dioxide: Cantidad de formas libres y ligadas de S02; en bajas concentraciones, el SO2 es casi indetectable en el vino, pero en concentraciones de SO2 libre superiores a 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino
8. density: Densidad, se aproxima a la del agua en función del porcentaje de alcohol y del contenido de azúcar
9. pH: Describe lo ácido o básico que es un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos se sitúan entre 3-4 en la escala de pH
10. sulphates: Un aditivo del vino que puede contribuir a los niveles de gas de dióxido de azufre (S02), que actúa como antimicrobiano y antioxidante
11. alcohol: Porcentaje de alcohol del vino

Variable de salida o clase

12. quality: Calidad del vino

Características del conjunto de datos

```
datos_vino.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157

Se evidencia en el conjunto de datos que todas las columnas son numéricas, son 1599 registros, ninguna columna presenta valores nulos y los valores máximos y mínimos de cada atributo. No existen valores en formato texto.

75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000

Revisión de valores nulos

```
datos_vino.isnull().sum()
```

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
```

Se observa que en conjunto de datos no se cuenta con valores nulos.

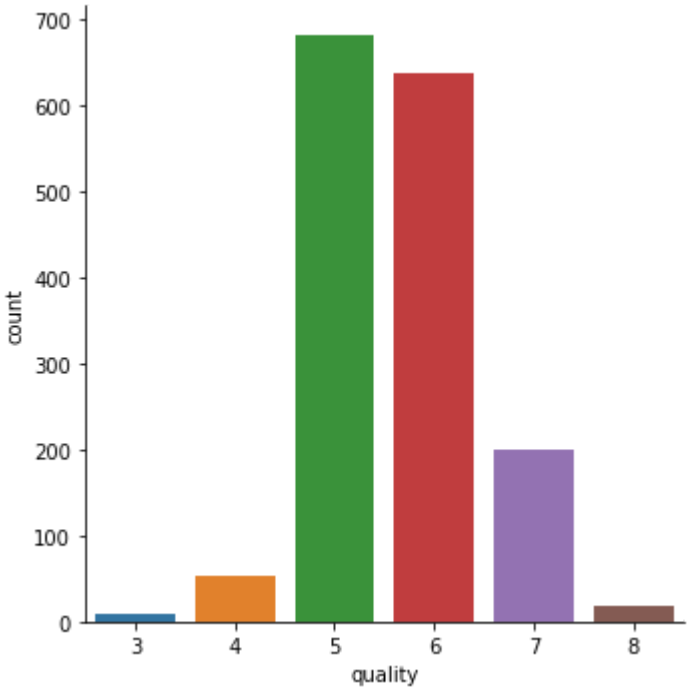
Revisión de la varaible 'quality'

```
datos_vino.quality.value_counts().sort_index()
```

```
3     10
4     53
5    681
6    638
7    199
8     18
Name: quality, dtype: int64
```

```
sns.catplot(x='quality', data=datos_vino, kind='count')
```

```
plt.show()
```



En la gráfica anterior se observa la cantidad de registros por cada valor de la variable 'quality' evidenciando que la mayor parte de los regisros se concentran para los valores de 'quality' 5 y 6. Para el presenta análisis se tomarán los valores iguales o inferiores a 5 como mala calidad y superiores a 5 vinos de buena calidad.

Para el presente análisis se establecerá que los datos que en la columna "quality" tengan valores menores o iguales a 5 corresponderán a una calidad mala y los superiores a 5 a calidad buena.

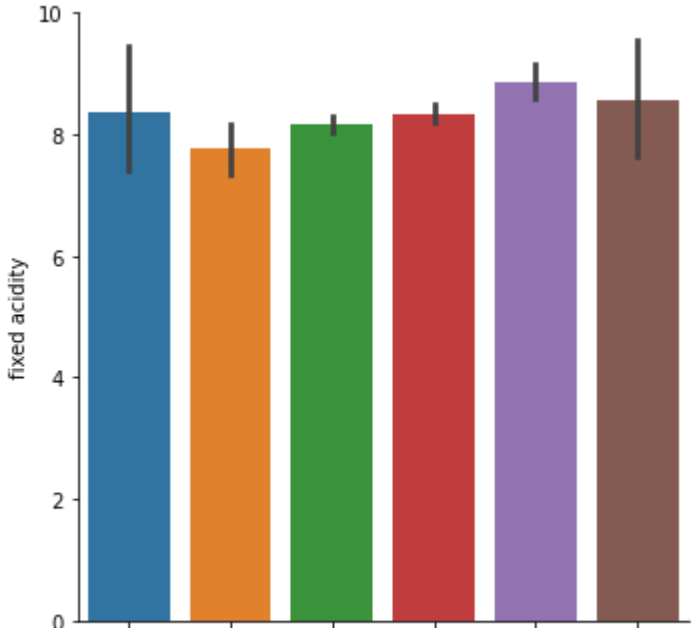
Análisis de los datos

Análisis bivariado

A continuación se procederá a realizar un análisis de cada una de las varibles de conjunto de datos en relación con la calidad.

Revisión 'fixed acidity' y 'quality'

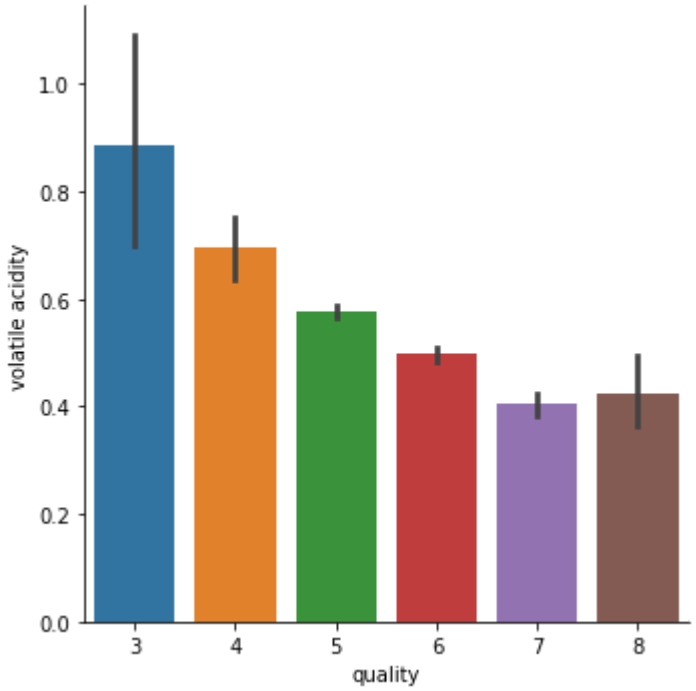
```
sns.catplot(x='quality', y = 'fixed acidity', data = datos_vino,kind='bar')
plt.show()
```



Se observa en la gráfica que los valores de 'fixed acidity' no marcan una clara diferencia en relación con la calidad 'quality'

Revisión 'volatile acidity' y 'quality'

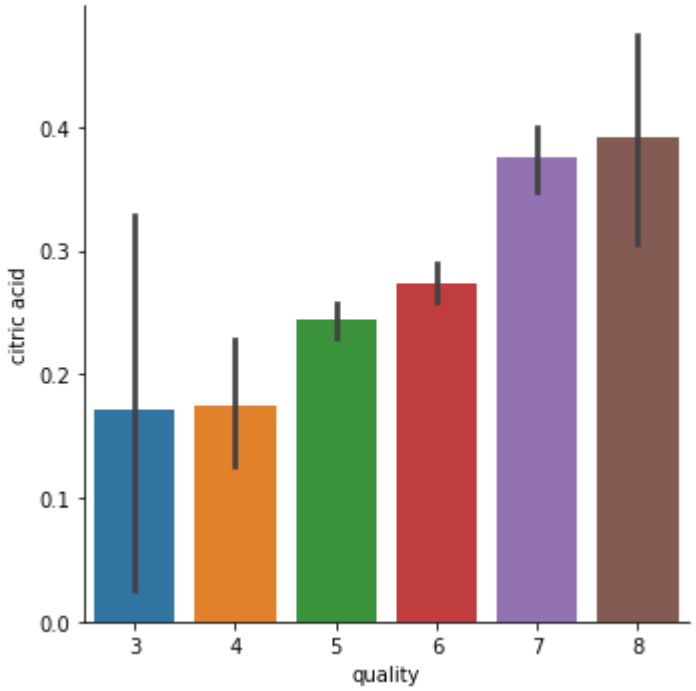
```
sns.catplot(x='quality', y = 'volatile acidity', data = datos_vino,kind='bar')
plt.show()
```



Se observa que a mayor 'volatile acidity' los valores de 'quality' son menores lo que indica que la calidad puede considerarse 'mala'

Revisión 'citric acid' y 'quality'

```
sns.catplot(x='quality', y = 'citric acid', data = datos_vino,kind='bar')
plt.show()
```



En reación a 'citric acid' a menores valores la calidad del vino puede considerarse buena.

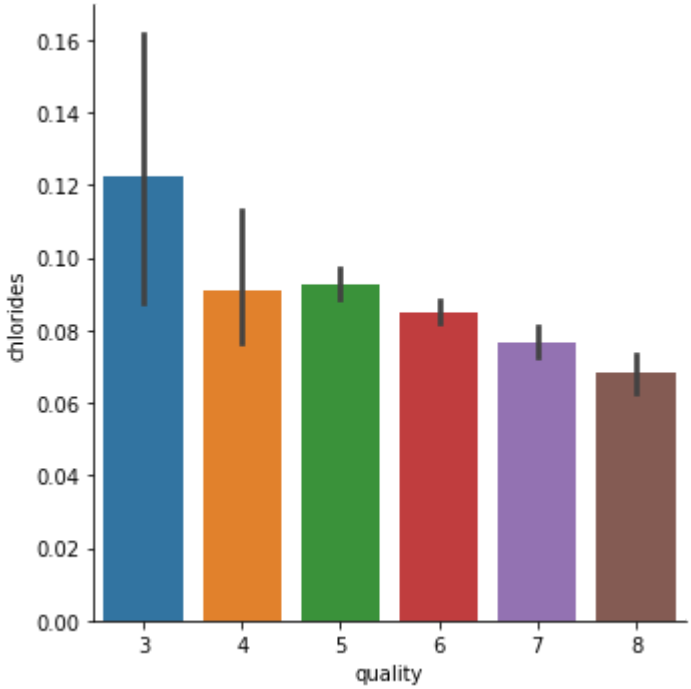
Revisión 'residual sugar' y 'quality'

```
sns.catplot(x='quality', y = 'residual sugar', data = datos_vino,kind='bar')
plt.show()
```

Los valores de 'residual sugar' no marcan una diferenciación clara para la calidad del vino.

Revisión 'chlorides' y 'quality'

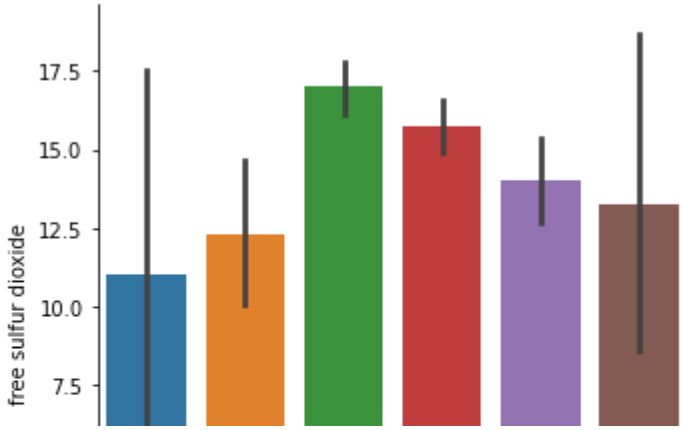

```
sns.catplot(x='quality', y = 'chlorides', data = datos_vino,kind='bar')
plt.show()
```



Para el caso de 'chlorides' se observa para valores por encima de 0.10 los valores de 'quality' se ubican en 3(una mala calidad)

Revisión 'free sulfur dioxide' y 'quality'

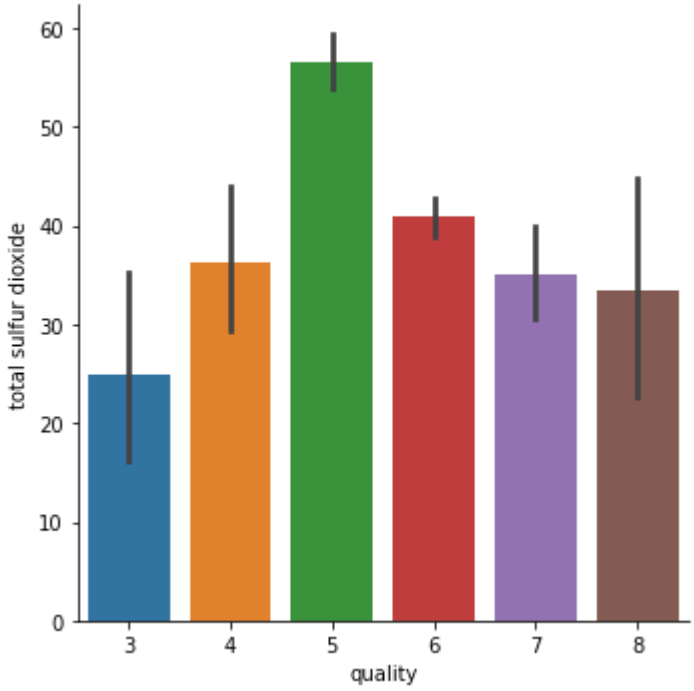
```
sns.catplot(x='quality', y = 'free sulfur dioxide', data = datos_vino,kind='bar')
plt.show()
```



Los valores de 'free sulfur dioxide' no evidencian una clara incidencia en la calidad del vino.

Revisión 'total sulfur dioxide' y 'quality'

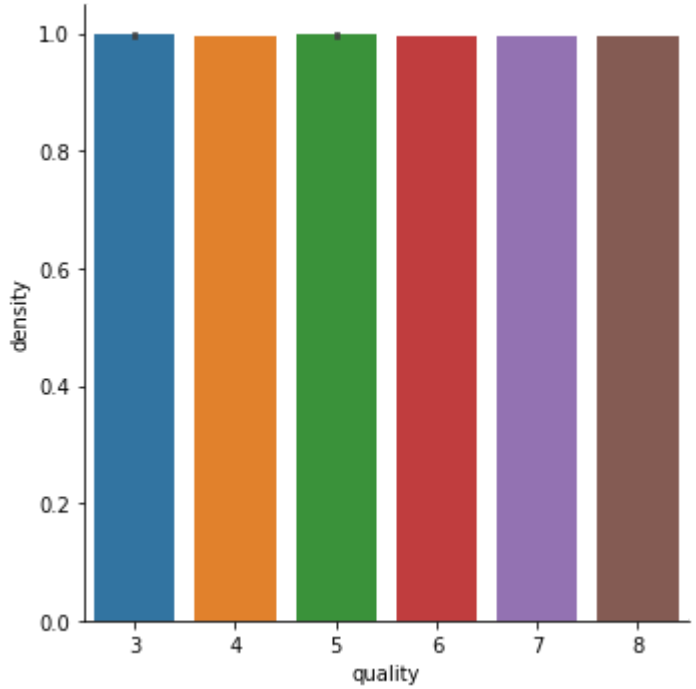
```
sns.catplot(x='quality', y = 'total sulfur dioxide', data = datos_vino,kind='bar')
plt.show()
```



Valores por encima de 40 en 'total sulfur dioxide' pueden incider en una calidad(5) mala del vino, para valores inferiores no es evidenciable de manera clara el impacto en la calidad del vino.

Revisión 'density' y 'quality'

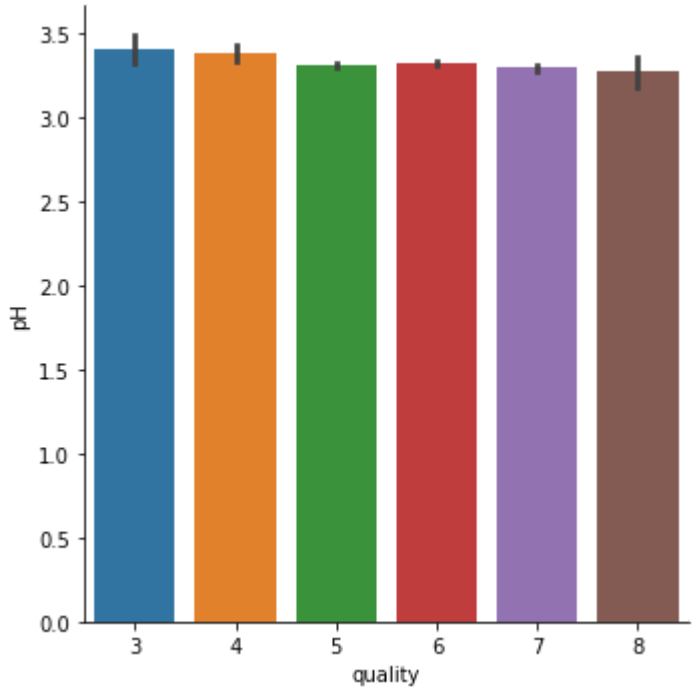
```
sns.catplot(x='quality', y = 'density', data = datos_vino,kind='bar')
plt.show()
```



Los valores de 'density' se ubican entre 0.9 y 1 y no marcan una diferencia en la calidad del vino.

Revisión 'pH' y 'quality'

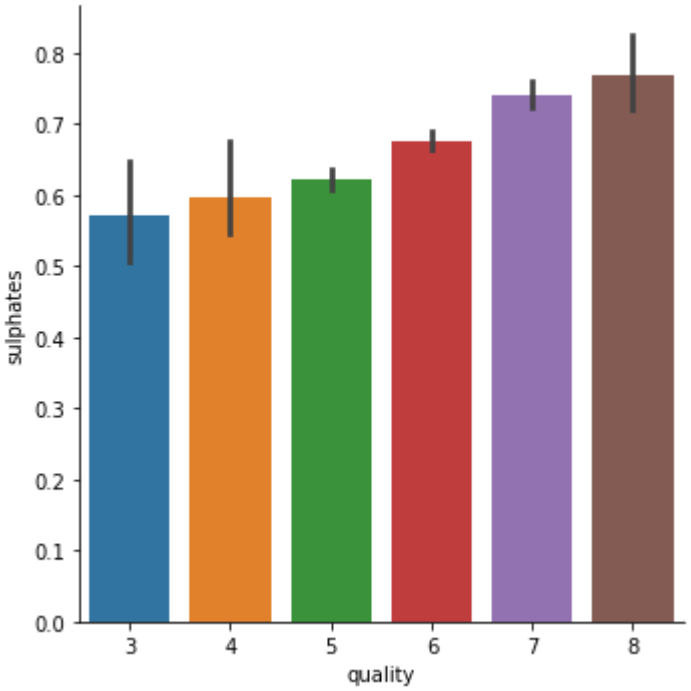
```
sns.catplot(x='quality', y = 'pH', data = datos_vino,kind='bar')
plt.show()
```



Los valores de 'pH' no marcan una diferencia en la calidad del vino

Revisión 'sulphates' y 'quality'

```
sns.catplot(x='quality', y = 'sulphates', data = datos_vino,kind='bar')
plt.show()
```



Los valores de 'sulphates' por encima 0.6 pueden incidir en valores de calidad buena

Revisión del 'alcohol' y 'quality'

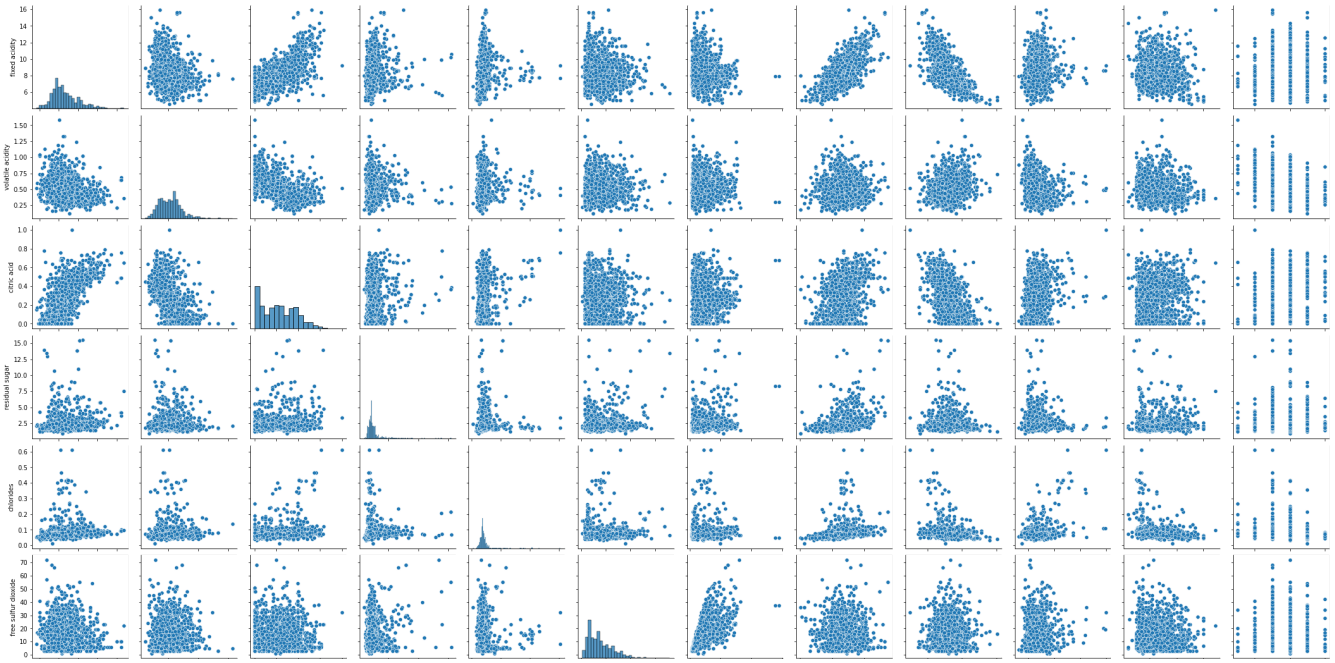
```
sns.catplot(x='quality', y = 'alcohol', data = datos_vino,kind='bar')
plt.show()
```

Se puede evidenciar en la gráfica que valores de 'alcohol' por encima de 10 pueden incidir en la calidad buena del vino.

Analisis de correlación de las variables

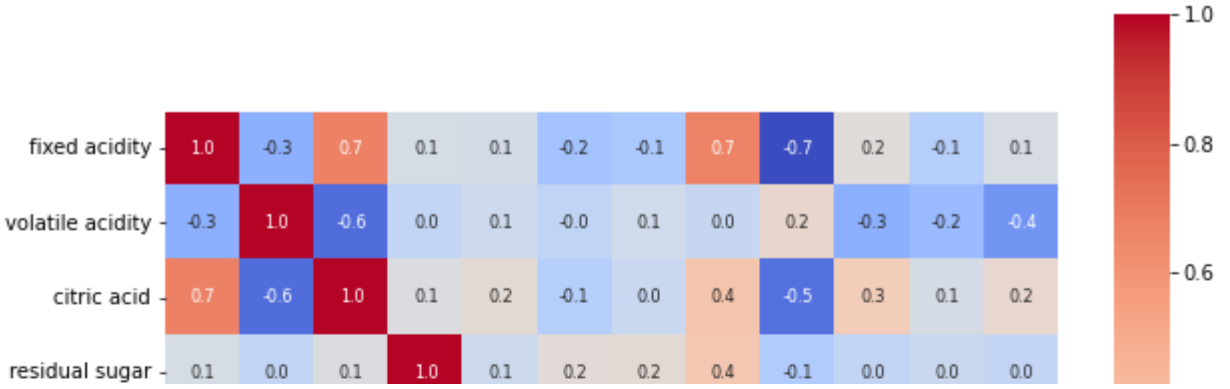
Se realiza a continuación un análisis de todas las variables para identificar cuales pueden estar correlacionadas con la calidad del vino.


```
sns.pairplot(datos_vino)
plt.show()
```



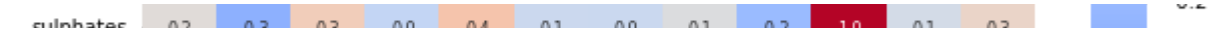
La gráfica anterior nos muestra que no existe a simple vista una relación directa entre la calidad del vino ya lguno de los otros atributos. A continuación, se realizarán otro tip de análisis para poder validar si es posible determinar la calidad del vino en función de otras variables.

```
correlation = datos_vino.corr()
plt.figure(figsize=(10,10))
sns.heatmap(correlation, cbar=True, square=True, fmt = '.1f', annot = True, annot_kws={'size'
plt.show()
```



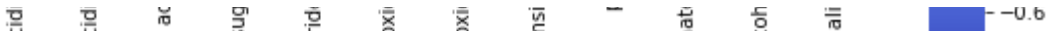
Al revisar la correlación entre la variable 'quality' y las demas varaibles se identifica que con la varaible 'alcohol' existe una correlación de 0.5 lo cual indica un efecto medio de los valores de alcohol en función de la calidad del vino. De igual forma la 'volatile acidity' se relaciona de forma inversa con la 'quality' indicando un efecto medio a mayor valor de 'volatile acidity' se podrá obtener una menor incidencia en la buena calidad del vino.

De acuerdo con lo anterior se tomaran los atributos 'alcohol' y 'volatile acidity' junto con 'quality' para realizar los siguientes análisis.

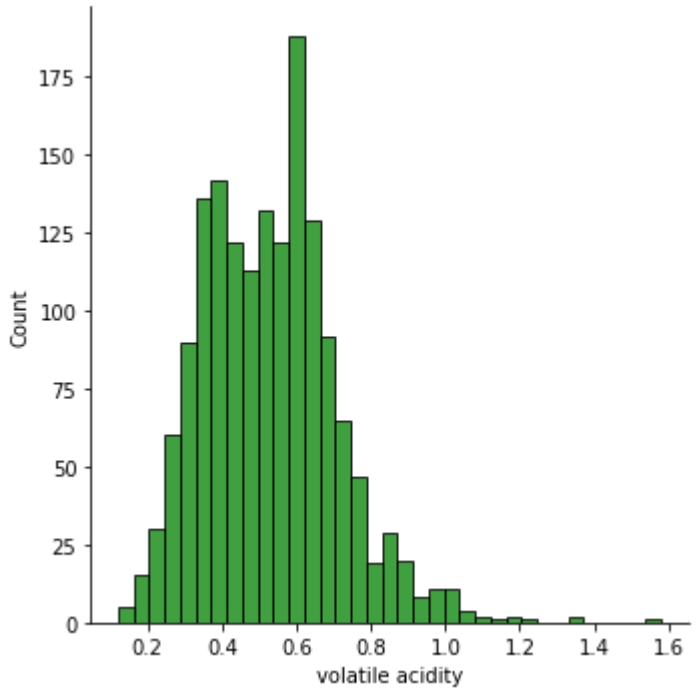


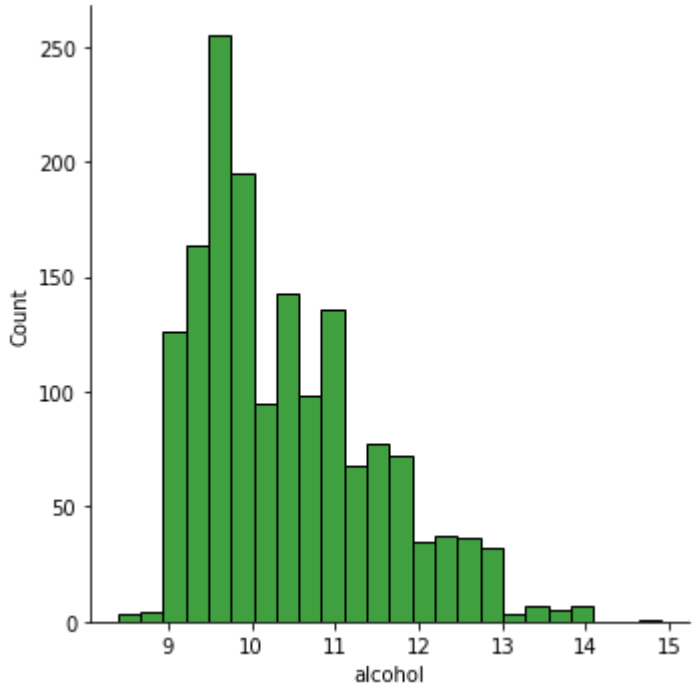
Comprobación de la normalidad y homogeneidad de la varianza

A continuacíon se realizará la comprobación de la normaliad de los datos para cada una de las variables.



```
sns.displot(x = 'volatile acidity', data = datos_vino,kind='hist',color="green")
sns.displot(x = 'alcohol', data = datos_vino,kind='hist',color="green")
sns.displot(x = 'quality', data = datos_vino,kind='hist',color="green")
plt.show()
```





Gráficamente evidenciamos que las vairables 'alcohol', 'volatile acidity' y 'quality' no tienen una distribución normal

```
levene_test = stats.levene(datos_vino['volatile acidity'],datos_vino['alcohol'],datos_vino['c
levene_test
```

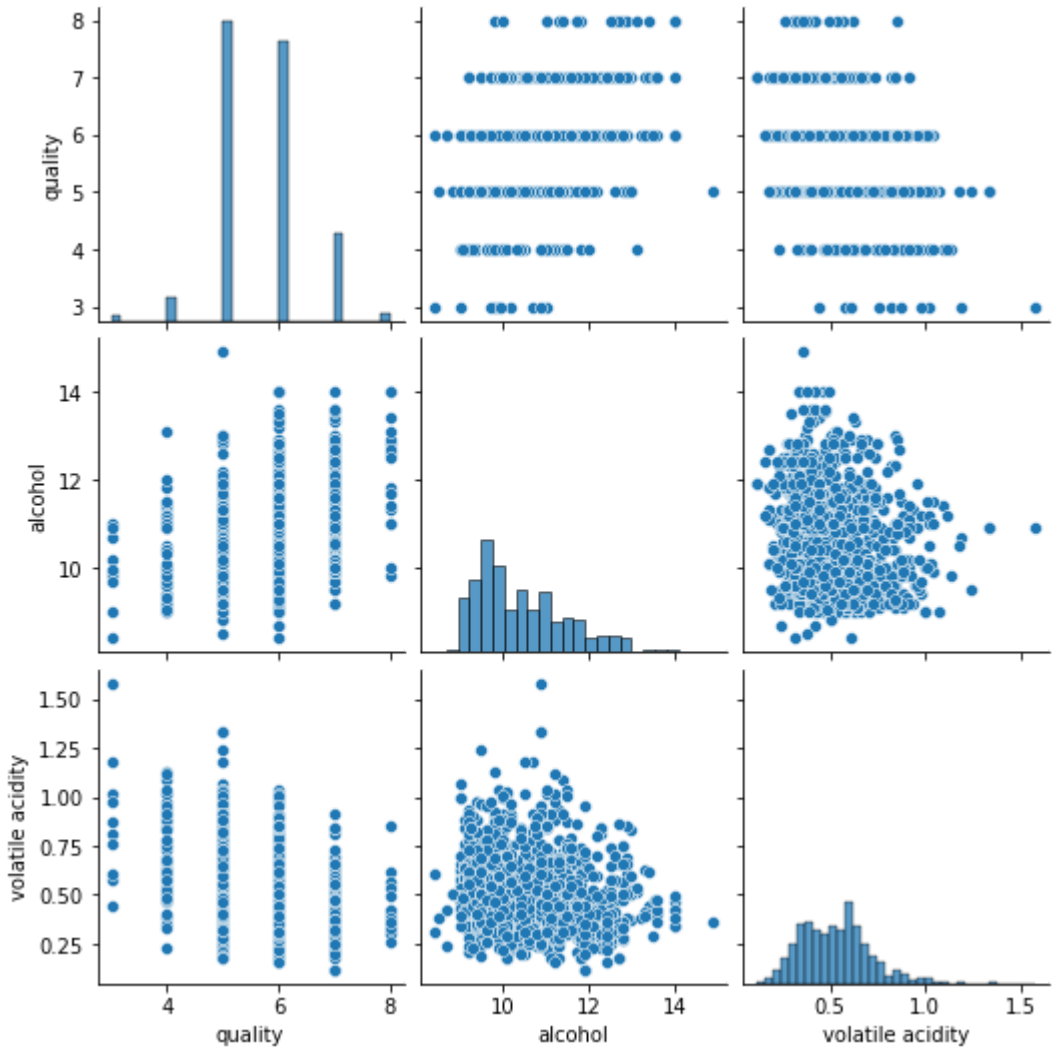
```
LeveneResult(statistic=811.9834632541177, pvalue=2.0018441210194983e-304)
```

Resultado del test el pvalue es menor a 0.05 por lo tanto se rechaza la hipótesis de que los grupos

Pruebas estadísiticas

Regresión lineal

```
sns.pairplot(data=datos_vino[['quality','alcohol','volatile acidity']])
plt.show()
```



De forma gráfica se evidencia que no es posible utilizar un modelo de regresión lineal para estimar la calidad del vino en función de las variables alcohol y acidez.

Correlación pearson entre las variables 'alcohol' y 'quality'

```
corr_test = pearsonr(x = datos_vino['alcohol'], y =  datos_vino['quality'])
print("Coeficiente de correlación de Pearson: ", corr_test[0])
print("P-value: ", corr_test[1])
```

```
Coeficiente de correlación de Pearson:  0.47616632400113607
P-value:  2.831476974778582e-91
```

El índice de correlación es 0.47 con lo cual no es recomendable utilizar un modelo de regresión lineal para predecir la calidad en función de alcohol.

Correlación pearson entre las variables 'volatile acidity' y 'quality'

```
corr_test = pearsonr(x = datos_vino['volatile acidity'], y =  datos_vino['quality'])
print("Coeficiente de correlación de Pearson: ", corr_test[0])
print("P-value: ", corr_test[1])
```

```
Coeficiente de correlación de Pearson:  -0.39055778026400717
P-value:  2.0517148070151443e-59
```

El índice de correlación es -0.39 con lo cual no es recomendable utilizar un modelo de regresión lineal para predecir la calidad en función de la acidez

Conclusiones

Con base en la información analizada de cada variable no se evidencia que la calidad del vino pueda estar relacionada con alguna variable de forma directa de tal manera que se pueda predecir el nivel de calidad en función de alguna de ellas.

Con el análisis de correlación de variables se identificó que unas posibbles variables pueden incidir en la calidad del vino las cuales son: 'alcohol' y 'volatile acidity' sin embargo, al realizar unas pruebas de regresión lineal y correlación de perason no se puede concluir que estás variables tengan una fuerte incidencia en la calidad del vino.