¿Que propiedades del vino se relacionan con alta calidad?

Unai Famoso

Sobre el *Dataset*

Los *datasets* provienen de muestras de vino portugués "Vinho Verde".

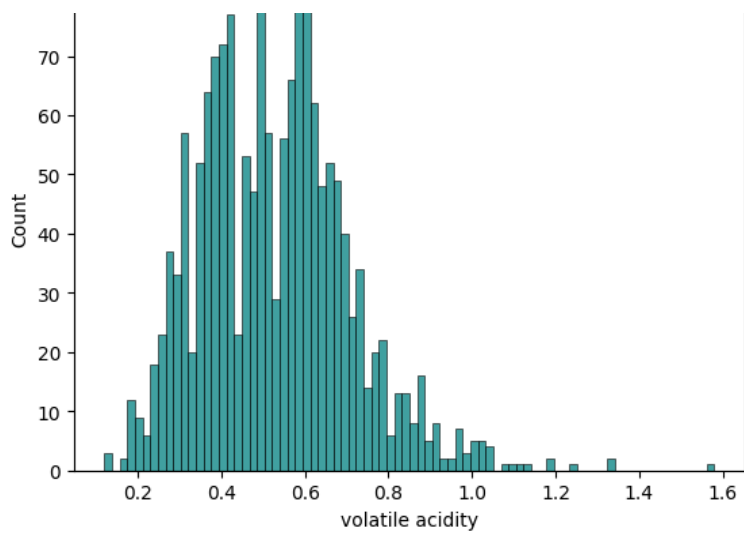
Los datos están divididos en dos *datasets*: tinto y blanco, en esta presentación solo hablaremos del tinto

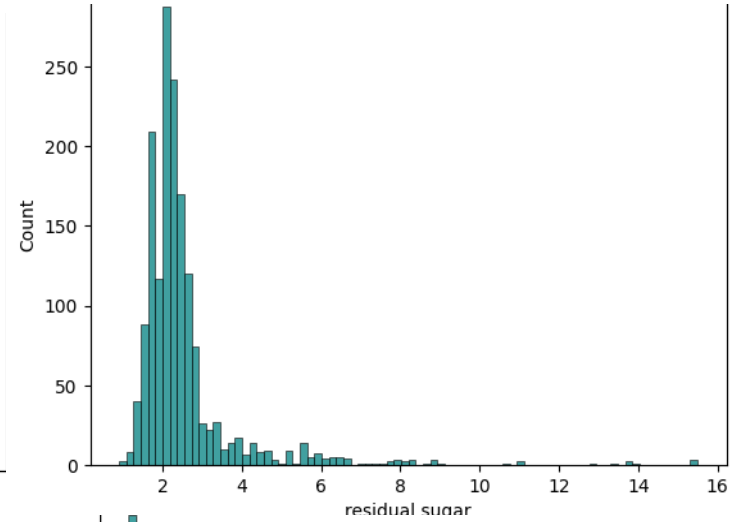
- *Acidez fija (g/l)*
- *Acidez volátil (g/l)*
- *Acido cítrico (g/l)*
- *Cloruros (g/l)*
- *Dióxido de Azufre total (SO2 total) (mg/l)*
- *Dióxido de Azufre libre (SO2 libre) (mg/l)*
- *Densidad (g/l)*
- *pH*
- *Sulfatos (g/l)*
- *Alcohol (vol %)*

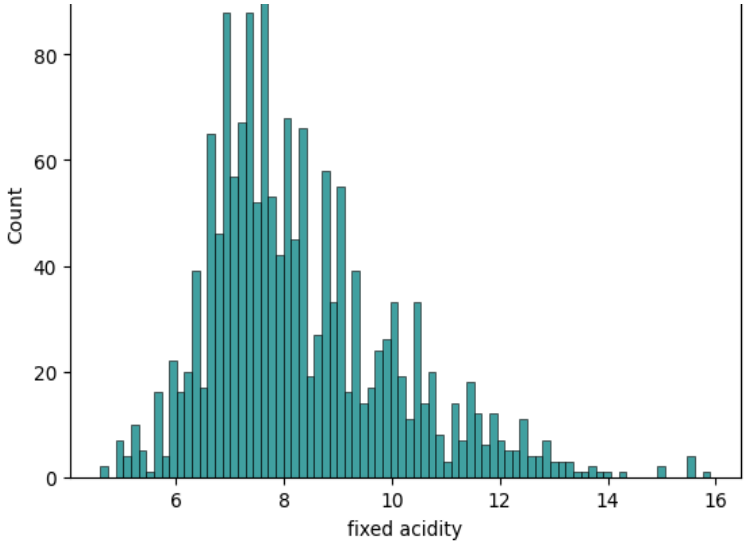
Y finalmente una nota de calidad (1-10) de una cata a ciegas de la muestra.

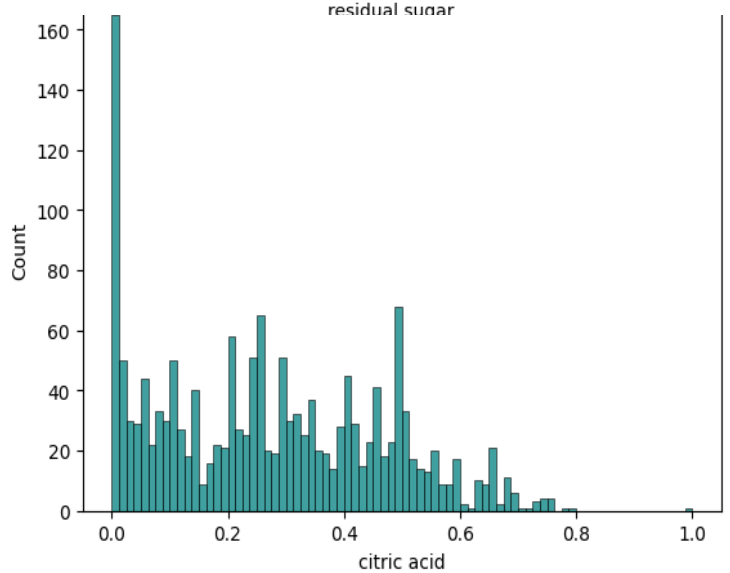
Preliminar primeras 4 variables

- No se han considerado ningún *outlier* estas 4 graficas



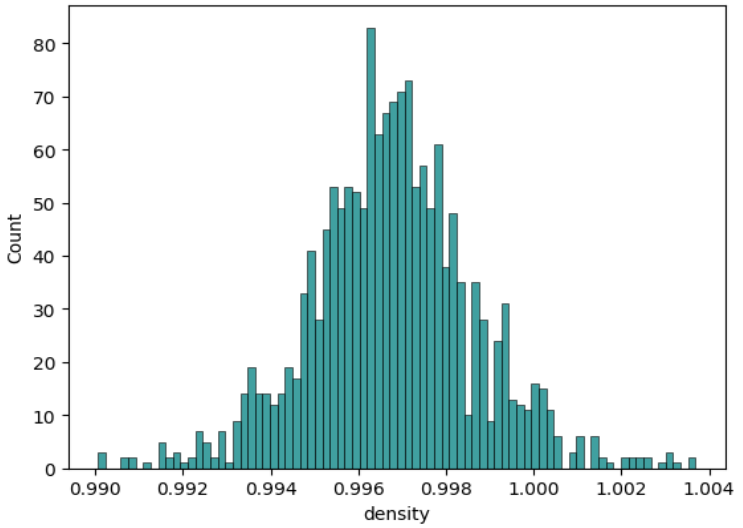


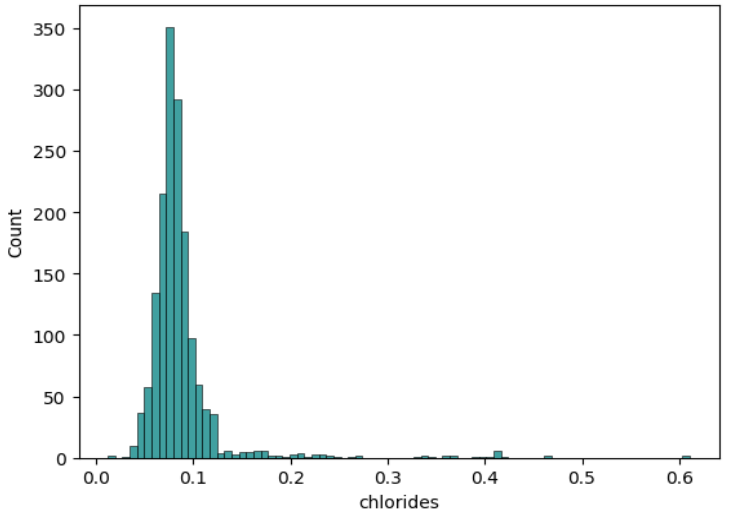


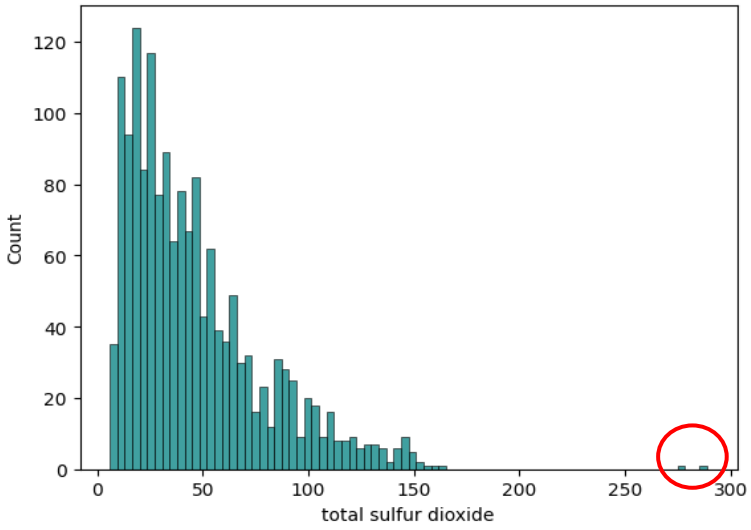


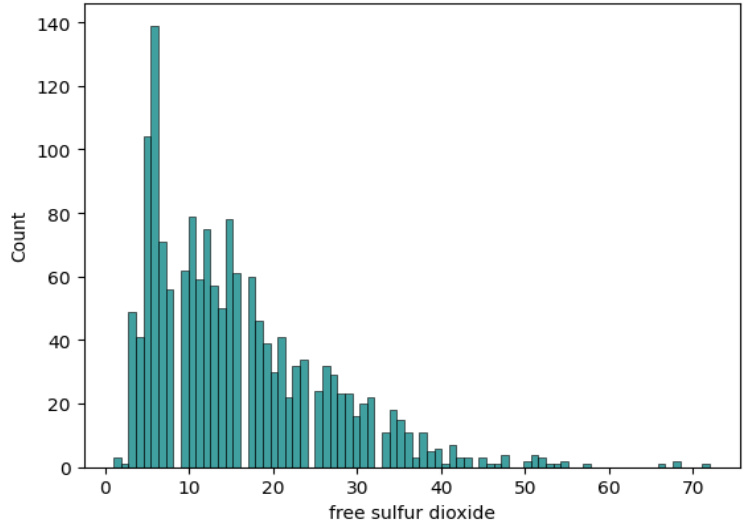
Preliminar segundas 4 variables

- Aquí si que se eliminaron dos valores anómalos SO2 total.



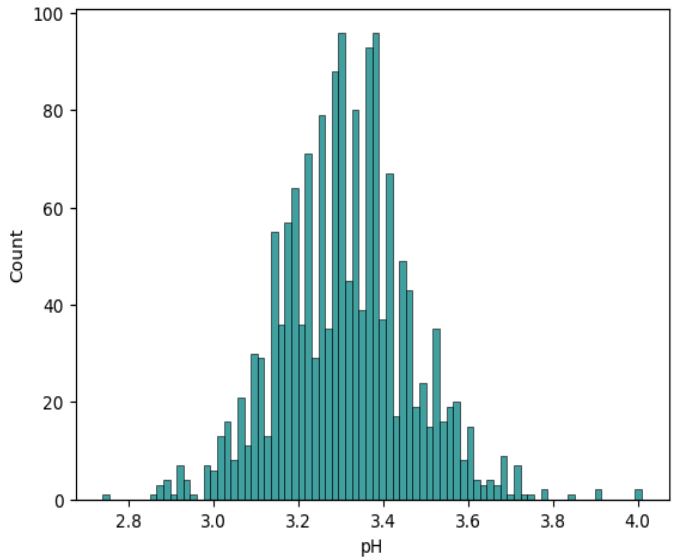


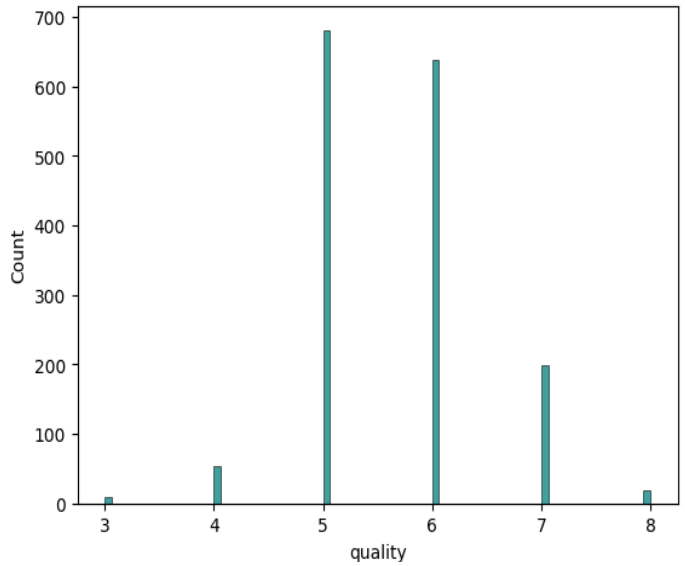


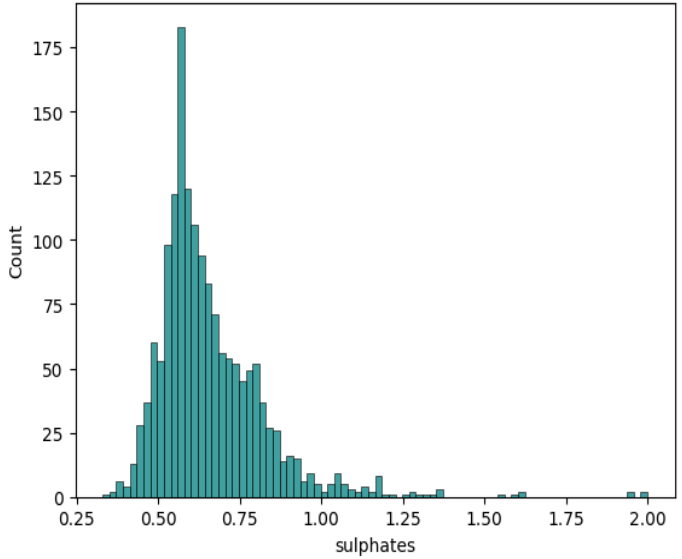


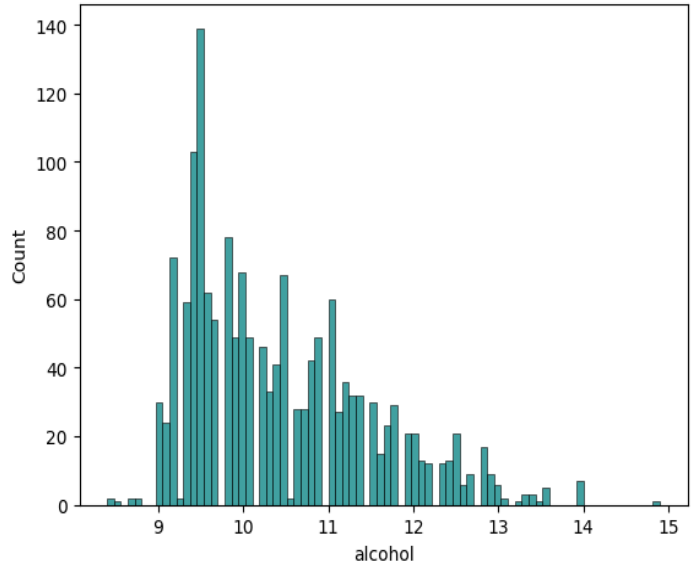
Preliminar ultimas 4 variables

- No se han considerado *outliers*.
- Otra cosa que podemos ver es la falta de vinos de alta calidad, ya que hay pocas muestras



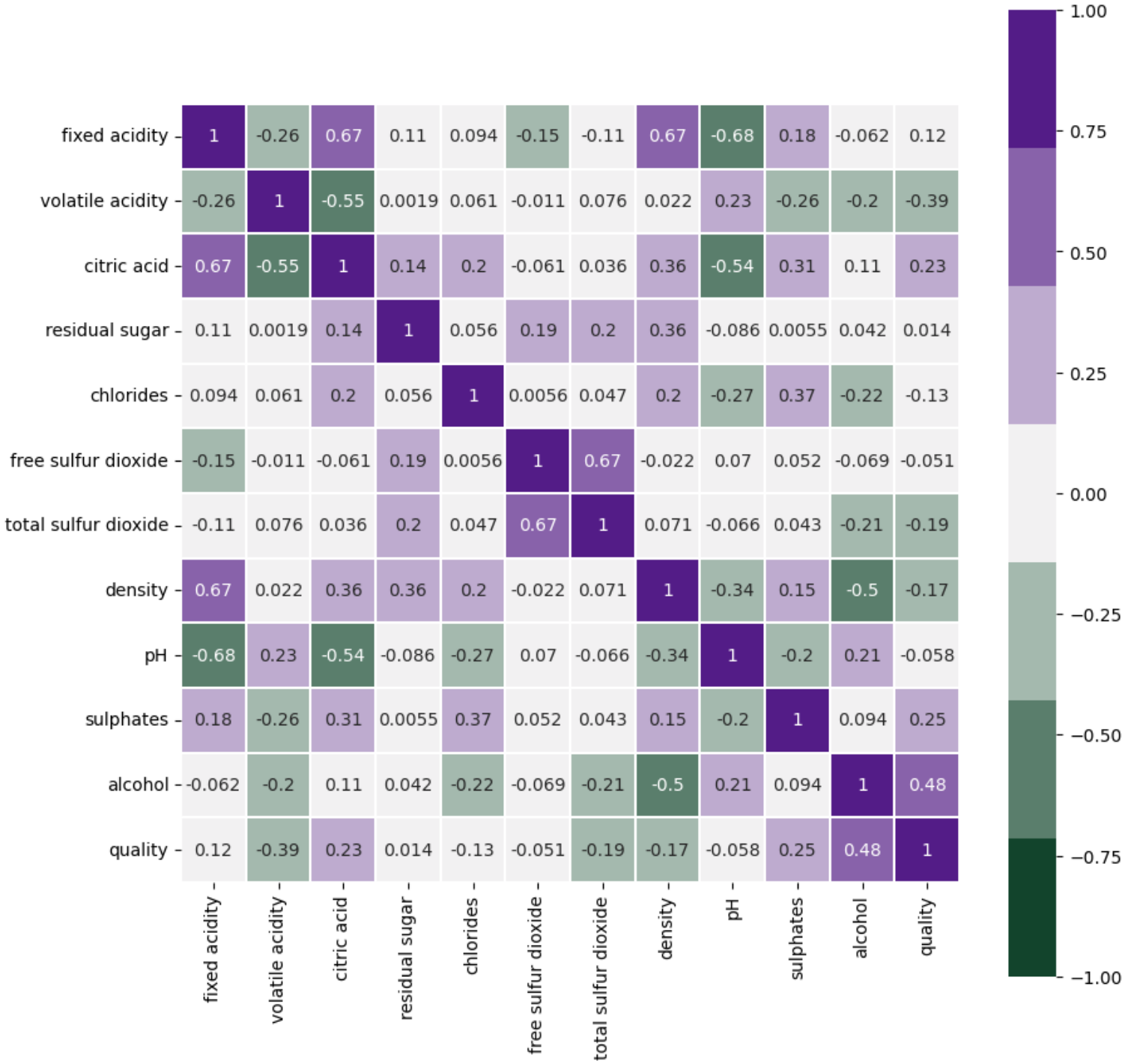




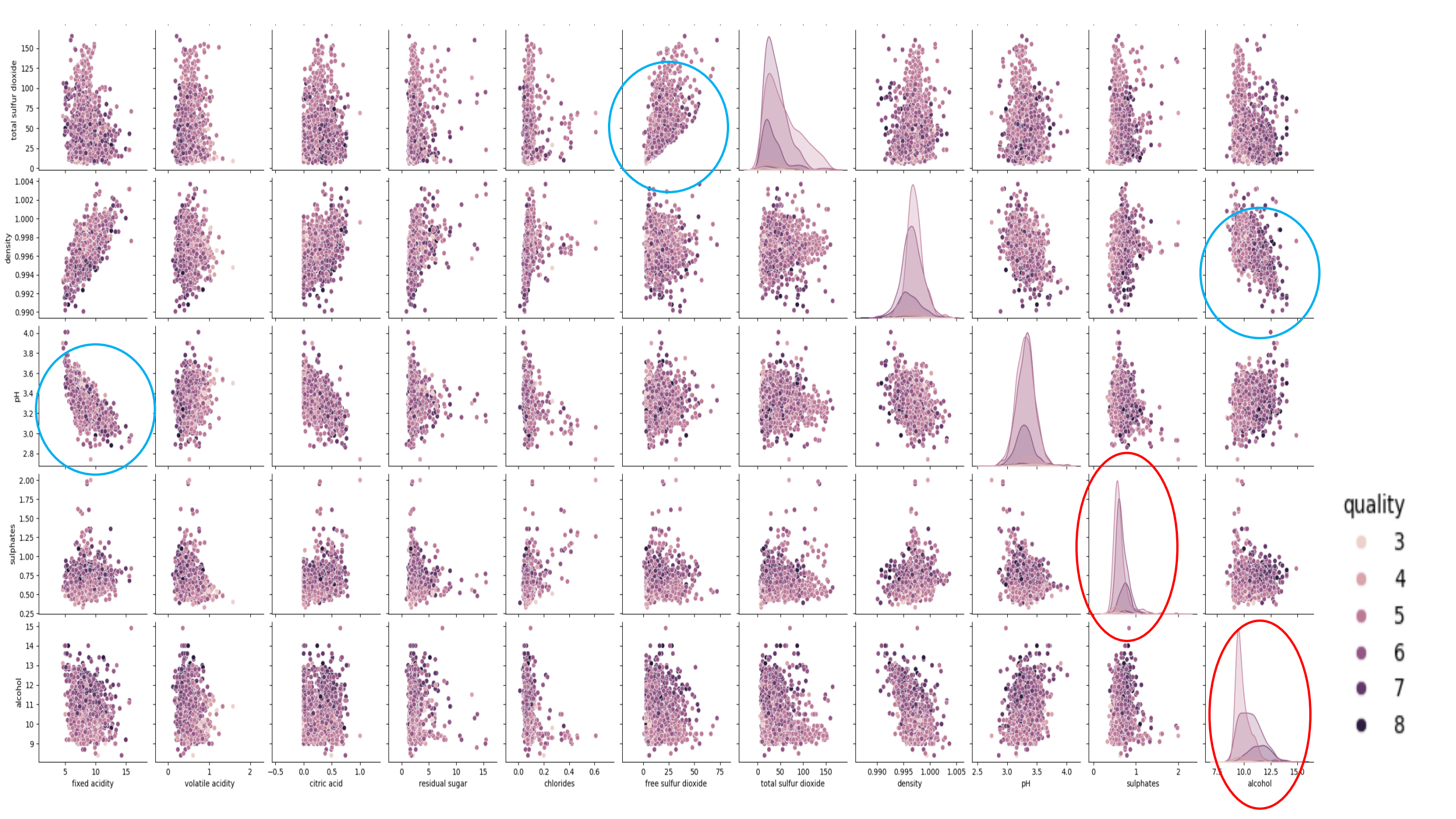


Correlaciones

- Considerando un *cut-off* de 0.15 y solo estudiando en la calidad
- Y finalmente la calidad esta relacionada:
  - **Negativamente Acidez volátil (-0.39)**
  - **Acido cítrico (0.23)**
  - **Negativamente con al SO2 total (-0.19)**
  - **Negativamente con la densidad (-0.17)**
  - **Sulfatos (0.25)**
  - **Alcohol (0.48)**

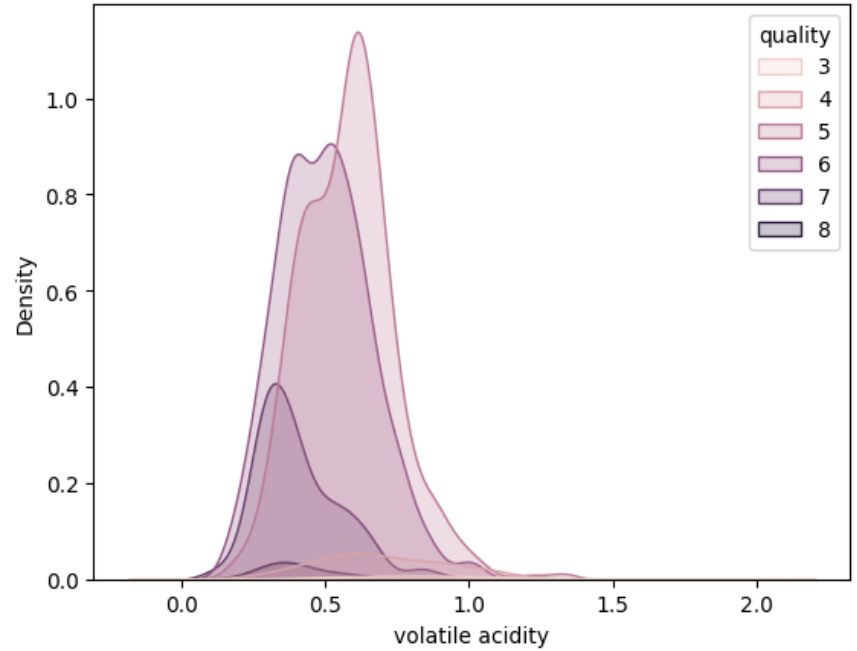


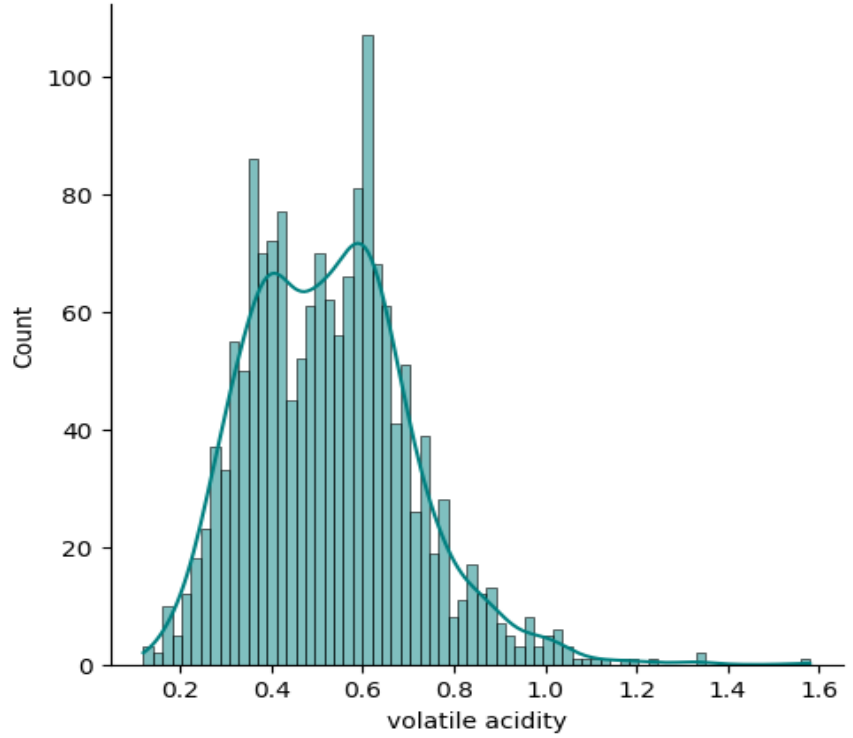

fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
3
4
5
6
7
8


total sulfur dioxide
density
pH
sulphates
alcohol
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
3
4
5
6
7
8

Acidez volátil

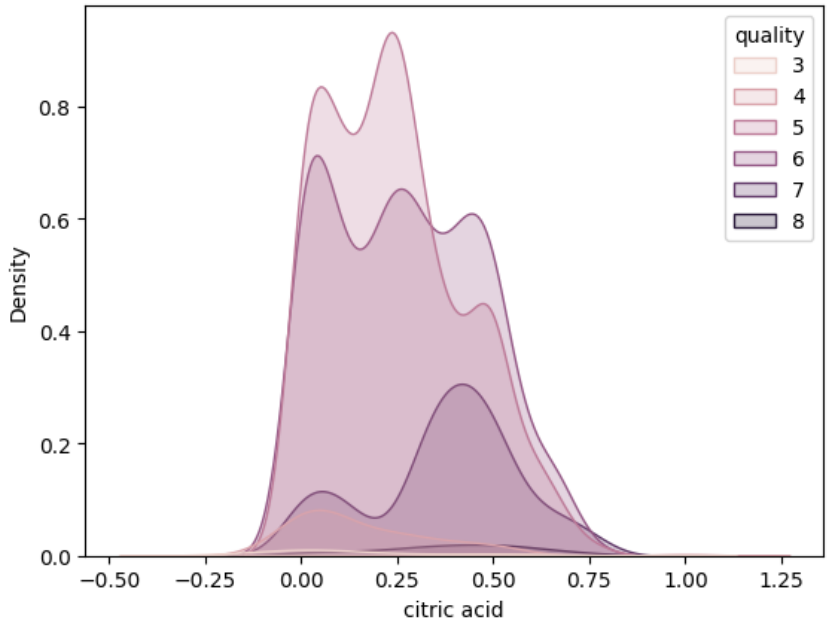
- Menor acidez volátil se asocia a una mejor calidad.
- Todas las calidades tienen varias concentraciones.
- En el histograma se ve mas claro.

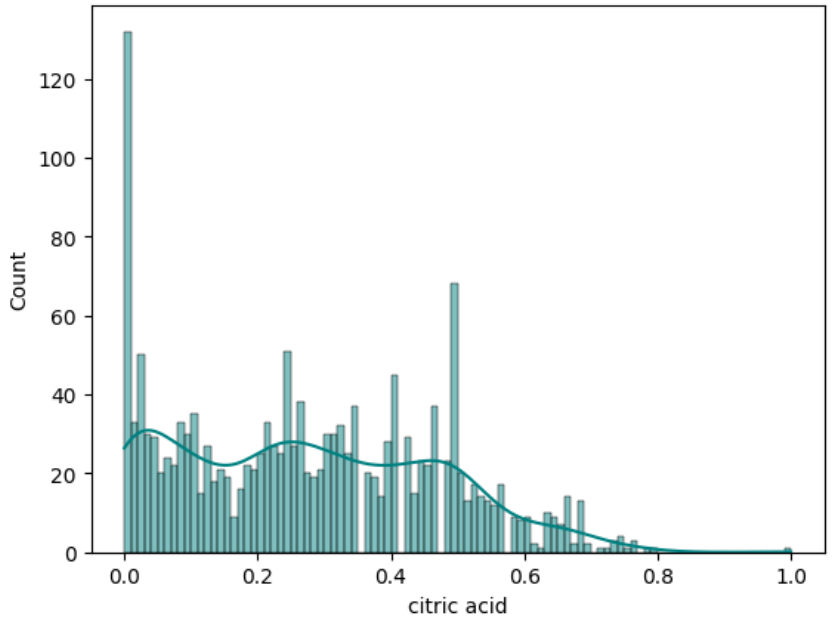




Acido cítrico

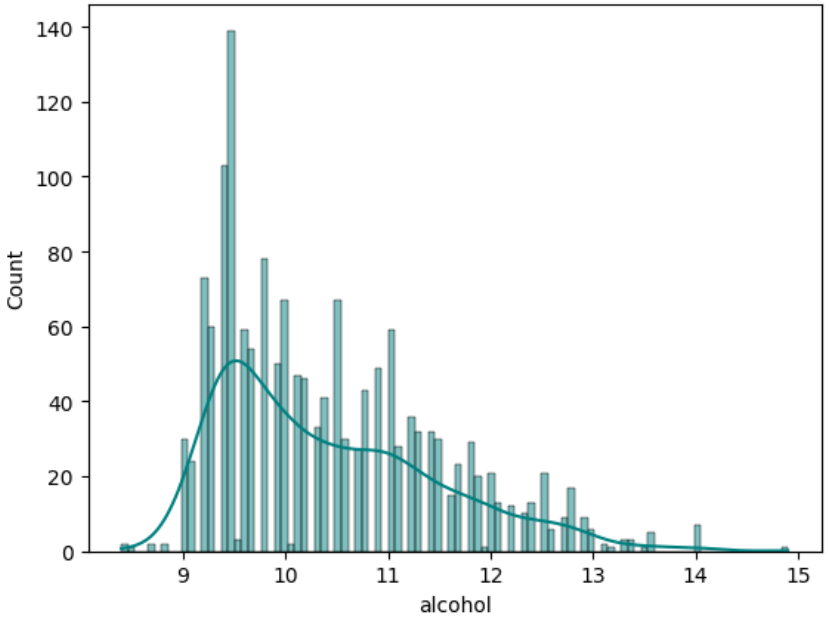
- En las curvas se ven 3 niveles, pero en el histograma no se ve tan claro y es mas probable que la muestra sea demasiado pequeña.
- Estos "tres niveles" seria cercano a 0, 0.25, y cerca de 0.50.

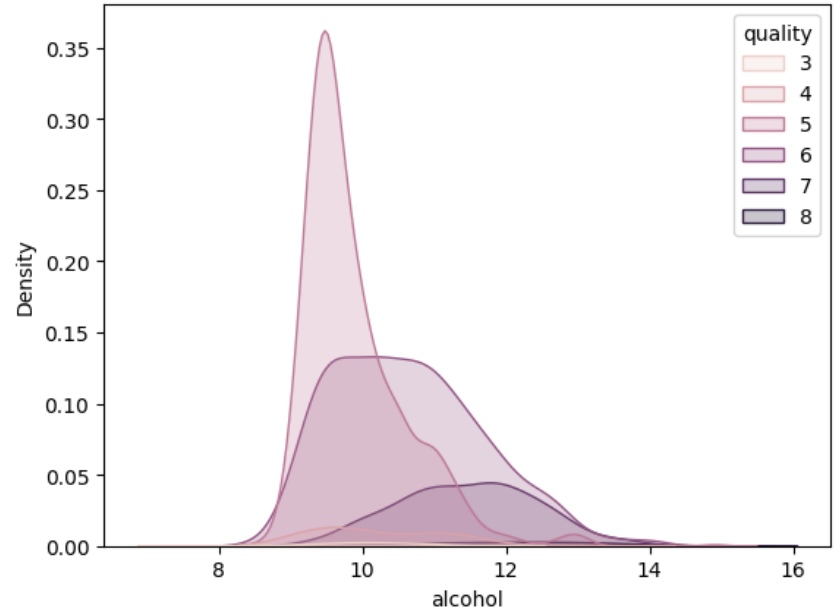




Alcohol

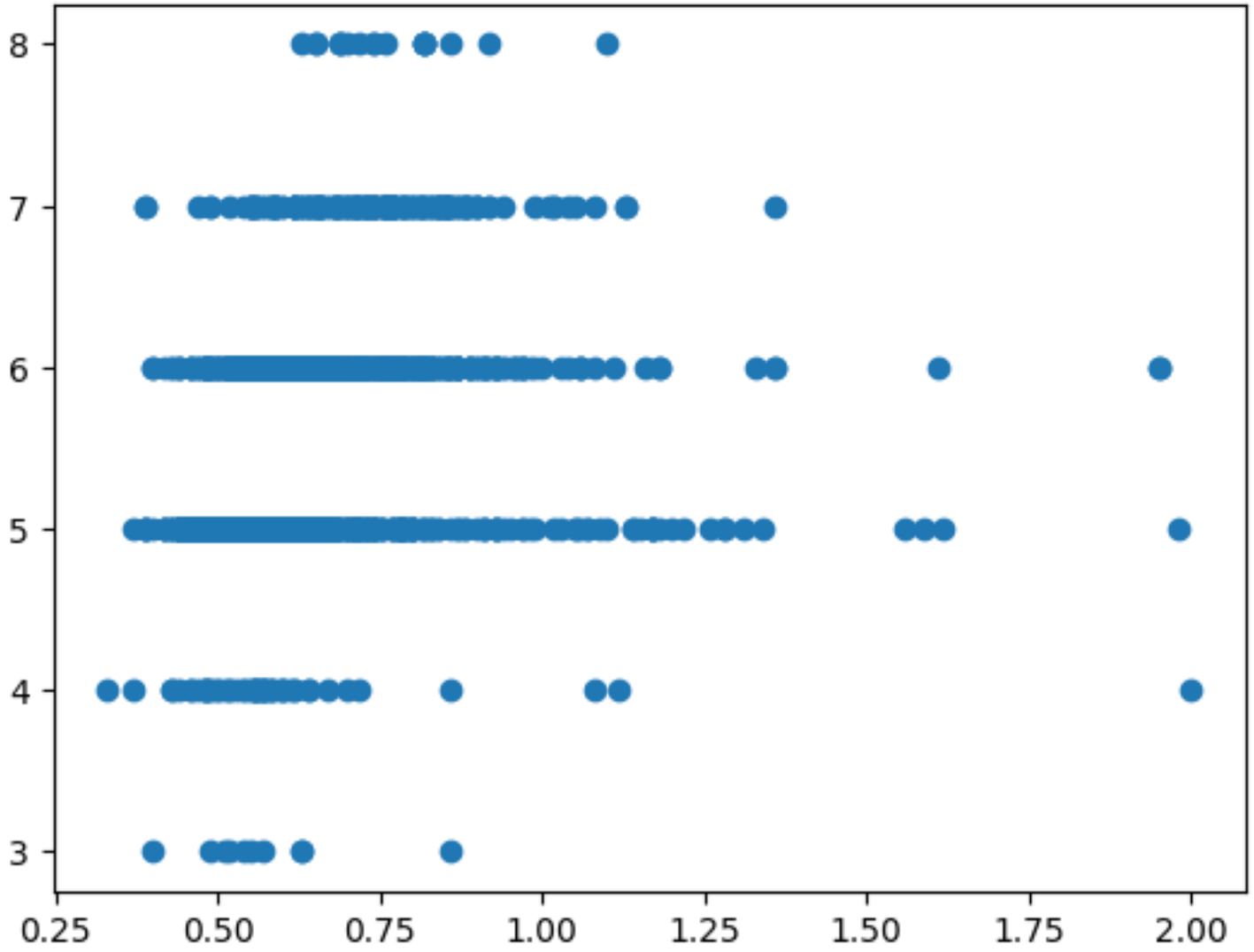
- Vinos de baja calidad están en concentración mas baja torno al 9,5%.
- Los vinos de mayor calidad muestran una distribución mucho más dispersa, pero siempre mas concentrado.





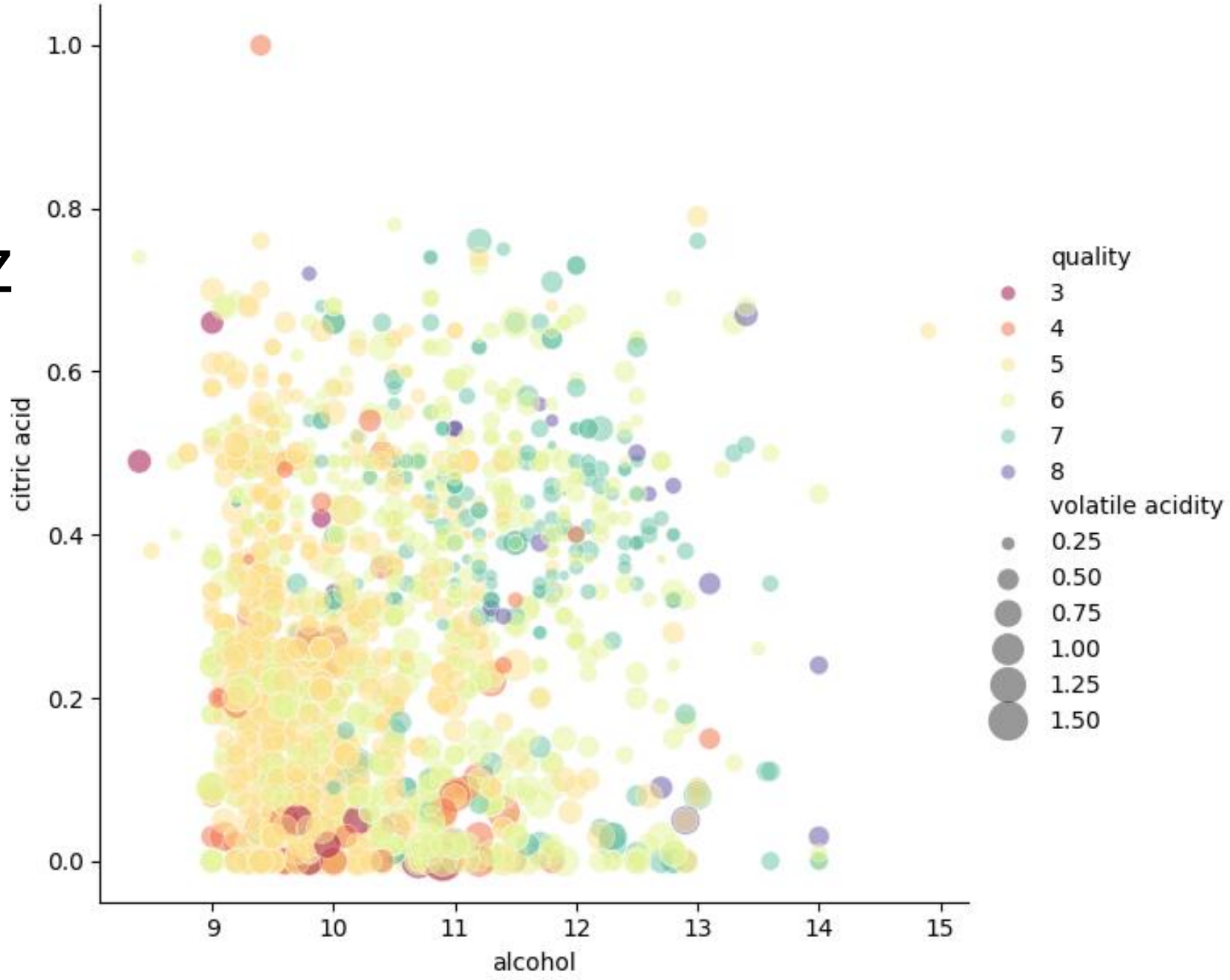
Sulfatos

- Vinos de calidad inferior tienen mas dispersión. Podría indicar un menor control durante su elaboración.
- Los de calidad mayor son menos dispersos 0.74 g/ml.



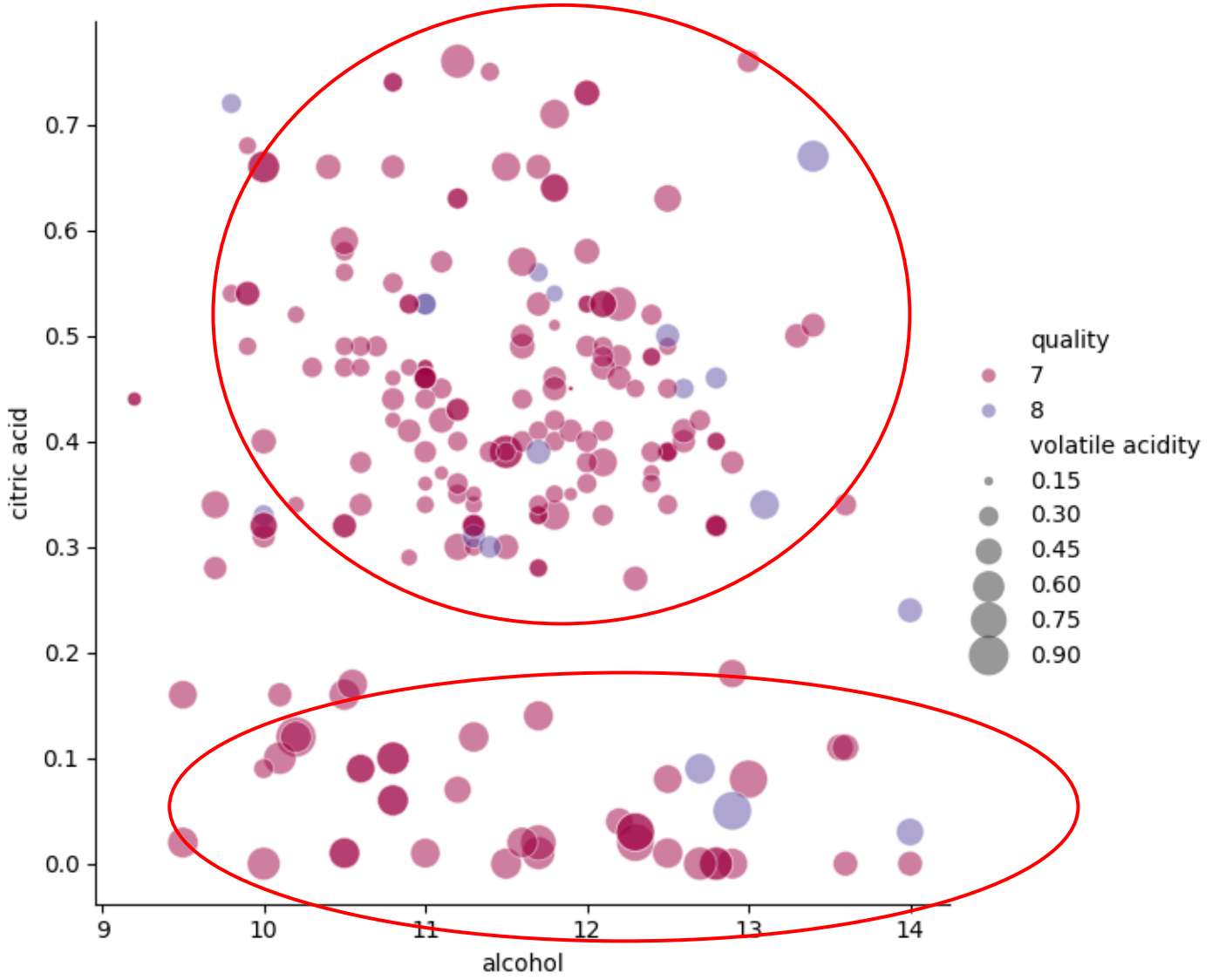
Todas las variables a la vez

- Demasiados de baja calidad.
- Filtramos para ver los de mayor calidad.



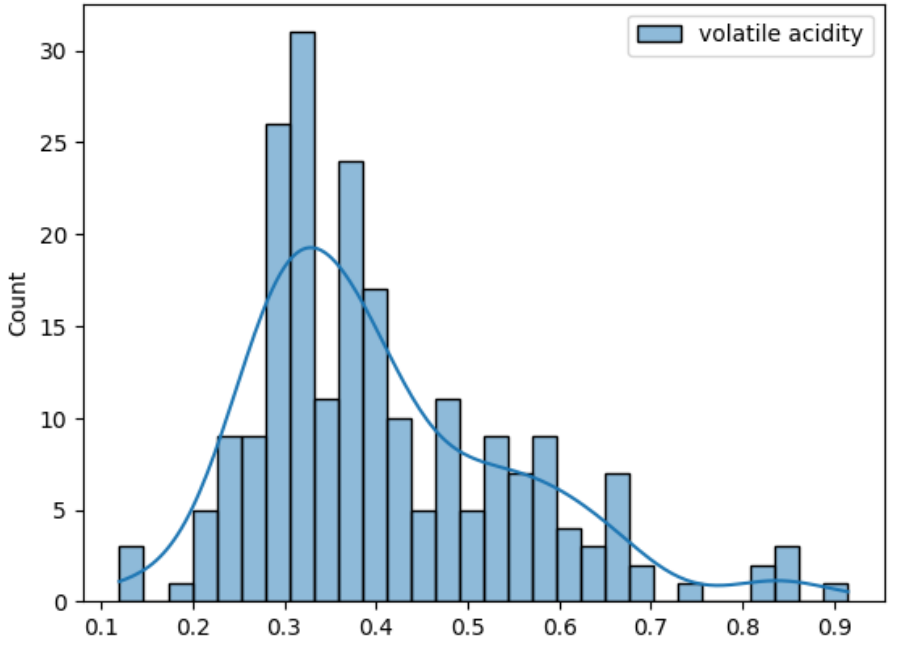
Todas las variables a la vez

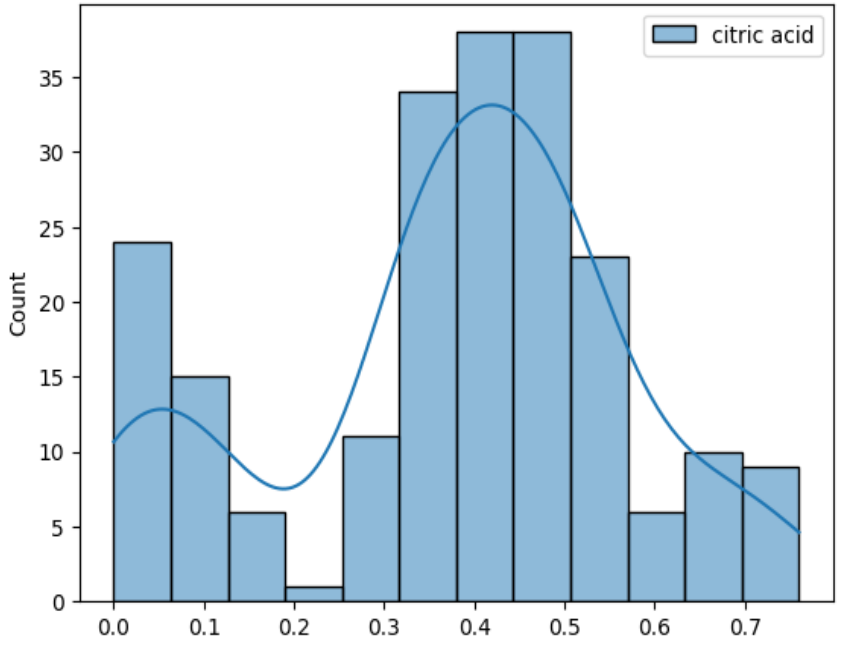
- Si solo ponemos los de máxima Calidad no hay suficientes muestras, es necesario meter mas datos.
- Los datos siguen siendo dispersos, pero las tendencias ahora son más claras y no se observan *outliers*. Parece que se pueden identificar dos grupos:
  - Los vinos con una concentración media de **ácido cítrico y acidez media**.
  - Los vinos con una baja concentración de **ácido cítrico y una acidez volátil** más alta.
- Es necesario usar análisis multivariante para asegurarse ya que ambos están muy correlacionados negativamente.

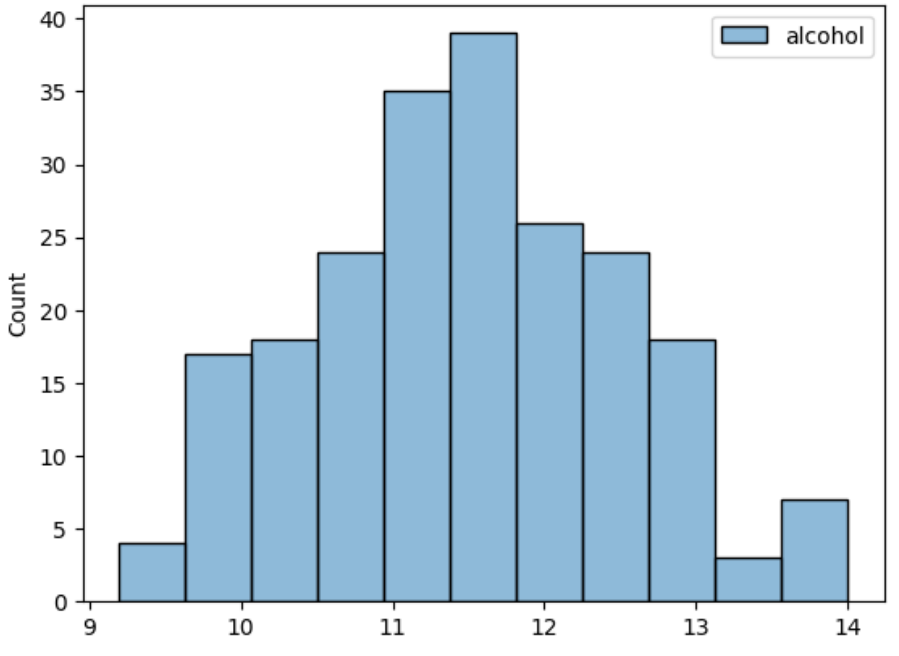


Estimación de concentración optima

- Vamos a intentar a calcular la concentración optima de las variables mas importantes.
- Pero para saber si los datos son fiables hay que comprobar si la distribución normal.







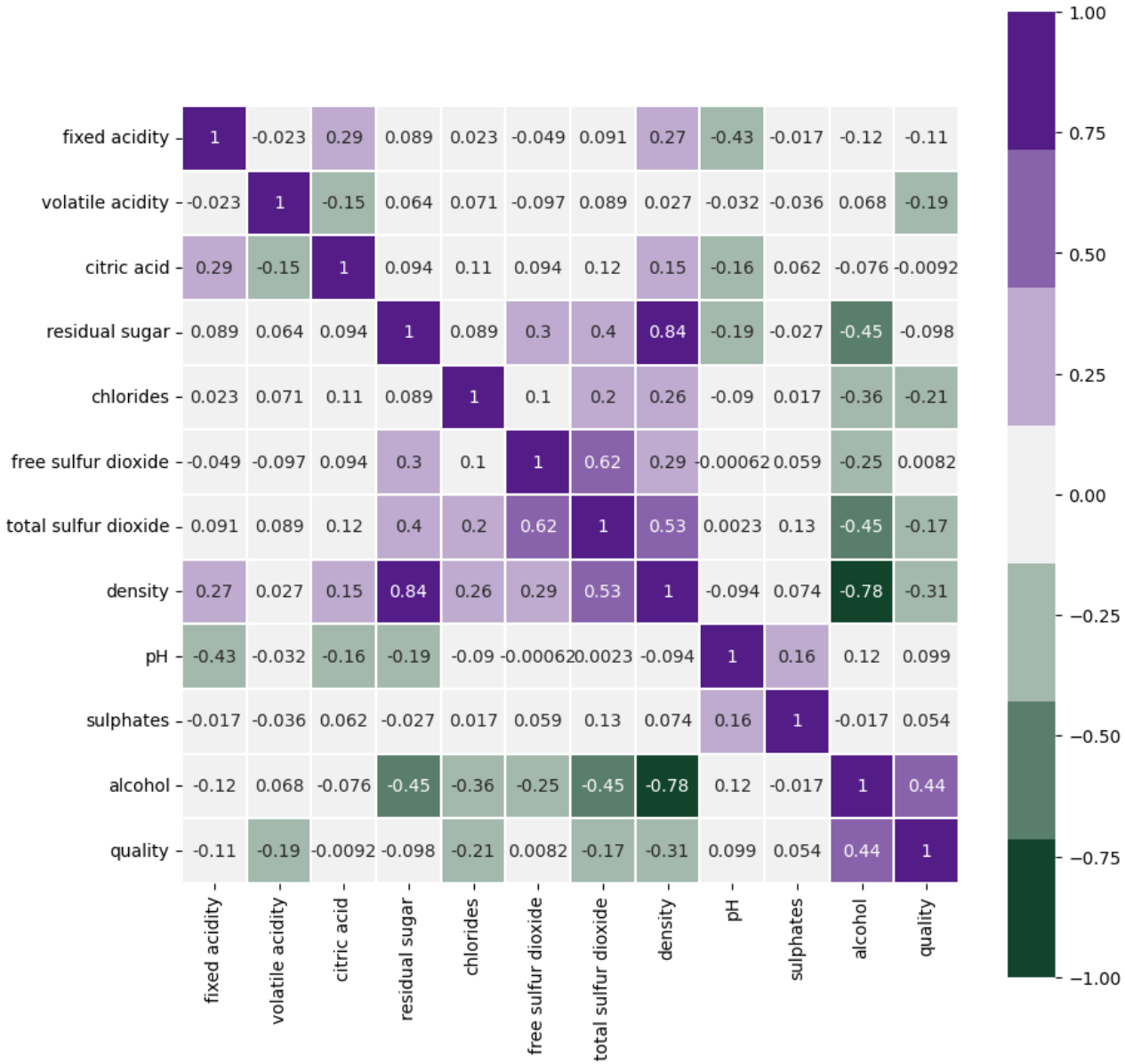
Considerando una significancia del 5%, ninguna de las variables sigue una distribución normal excepto el contenido de alcohol y densidad.
Realmente es algo esperable, ya que bastantes variables tienen un *skew* o son demasiado “planas” o “afiladas”.
Si es así los valores de los percentiles no son del todo fiables para estimar una concentración optima.
El acido cítrico y el alcohol tienen concentraciones muy similares, mientras tanto los sulfatos, y la acidez volátil son mas pequeñas.

	volatile acidity	citric acid	alcohol	sulphates
p_0	0.120	0.00	9.2	0.390
p_25	0.310	0.30	10.8	0.655
p_50	0.370	0.40	11.6	0.740
p_75	0.490	0.49	12.2	0.825
p_100	0.915	0.76	14.0	1.360

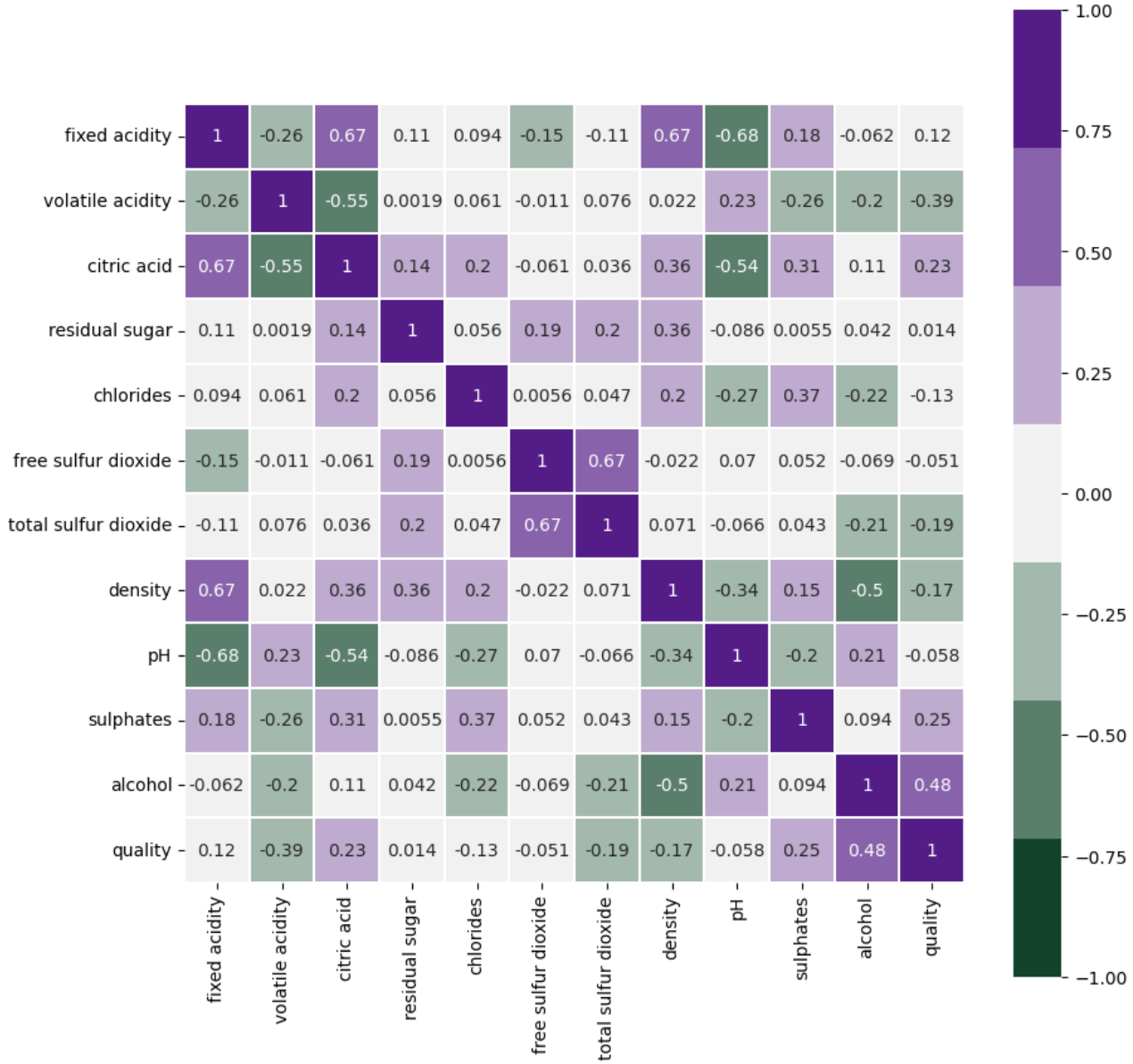
	statistic	pvalue
fixed acidity	8.350177	1.537383e-02
volatile acidity	32.004587	1.122774e-07
citric acid	7.388273	2.486892e-02
residual sugar	103.852678	2.809813e-23
chlorides	266.871920	1.120725e-58
free sulfur dioxide	65.491076	6.008879e-15
total sulfur dioxide	56.227478	6.171040e-13
density	3.824202	1.477696e-01
pH	8.940140	1.144651e-02
sulphates	28.121259	7.826113e-07
alcohol	1.984523	3.707373e-01
quality	149.058290	4.289470e-33

Conclusiones

- Todos los **ácidos y el pH** están estrechamente **relacionado** entre si.
- La **densidad** esta **relacionada negativamente** con el **alcohol,** pero **positivamente** para el resto de **los compuestos**.
- Los **ácidos volátiles** están **relacionados negativamente** con **alcohol**.
- La calidad esta relacionado positivamente con el contenido de **alcohol** y la concentración de **ácido cítrico** en el vino. Además, una mayor concentración de **sulfatos** también se asocia con una mejor calidad percibida. Por el contrario, los **ácidos volátiles**, **SO_2** y una baja **densidad** presentan una correlación negativa con la calidad.
- Se ha estudiado de que la concentración optima de las variables:
  - Acidez volátil 0.370 g/l
  - Acido cítrico 0.40 g/l
  - Alcohol 11.6 %
  - Sulfatos 0.740 g/l
- Realísticamente todas las variables y la relación de estas afectarían a la calidad. En futuros análisis seria recomendable **utilizar técnicas de análisis multivariante** como PCA.
- El conjunto de datos tiene algunos problemas:
  - Algo antiguo
  - Solo una variante de vino
  - No incluye pocos vinos de calidad mayor de 8, y tiene pocas muestras de vinos de calidad de 7 o mas
  - No contiene todos los compuestos relevantes en el sabor y aroma
- Estas conclusiones solo se aplican al vino tinto, el vino blanco tiene otras correlaciones



Blanco



Tinto