



PRÀCTICA 1:REGRESSIÓ

RED WINE QUALITY DATASET

Grup 105

Laia Rubio – NIU:1600830

Erik Villarreal – NIU:1599119

Raúl Villar – NIU:1596830

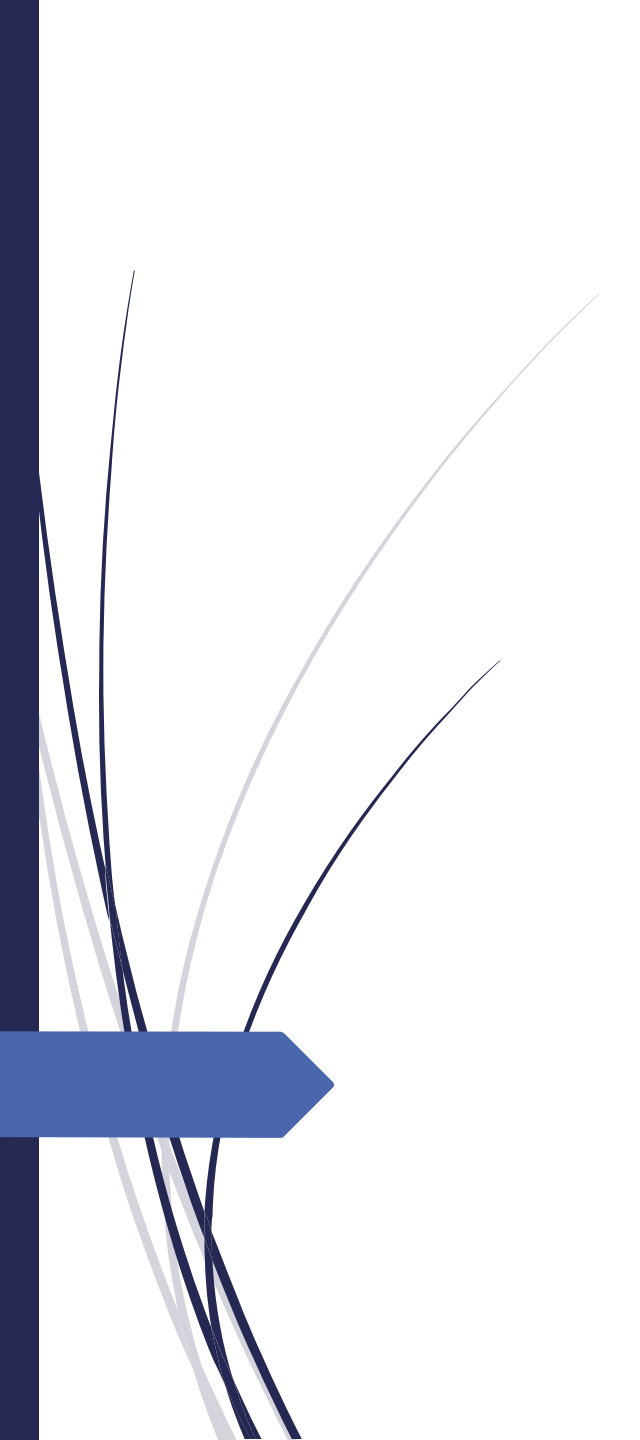
PLANTEJAMENT DE DADES

Taula de les primeres 5 mostres de la BD i els seus respectius valors

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.400	0.700	0.000	1.900	0.076	11.000	34.000	0.998	3.510	0.560	9.400	5
1	7.800	0.880	0.000	2.600	0.098	25.000	67.000	0.997	3.200	0.680	9.800	5
2	7.800	0.760	0.040	2.300	0.092	15.000	54.000	0.997	3.260	0.650	9.800	5
3	11.200	0.280	0.560	1.900	0.075	17.000	60.000	0.998	3.160	0.580	9.800	6
4	7.400	0.700	0.000	1.900	0.076	11.000	34.000	0.998	3.510	0.560	9.400	5

Taula obtinguda amb la funció describe() de la llibreria pandas

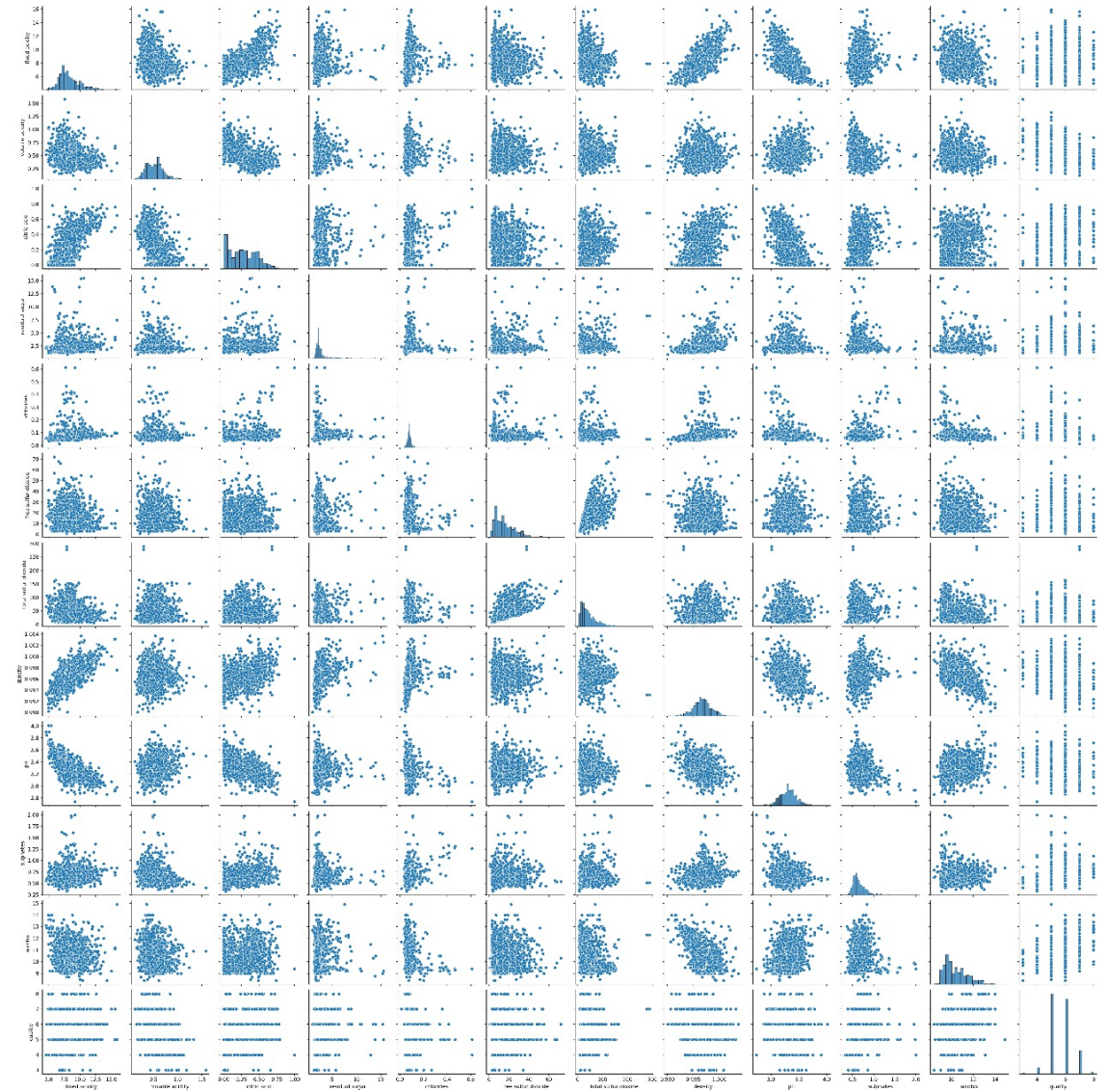
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000
mean	8.320	0.528	0.271	2.539	0.087	15.875	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



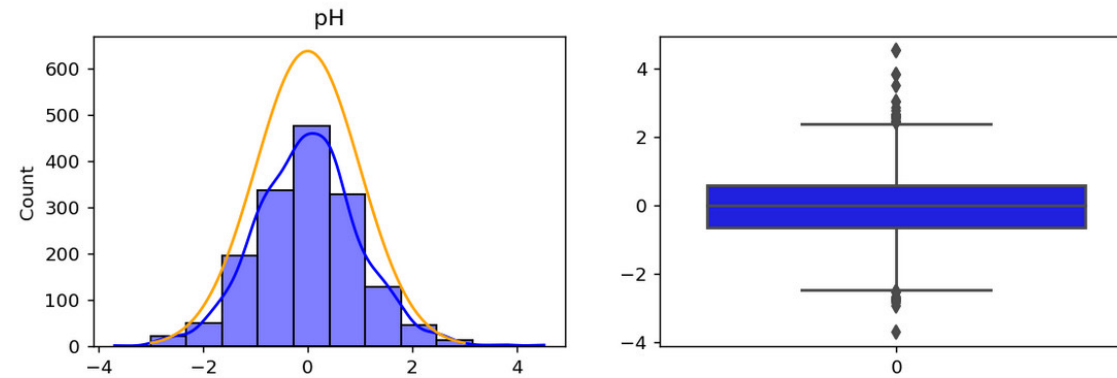
Nom Variable	Tipus de dada	Rang	Tipus de variable
fixed acidity	Float64	(4,600 – 15,900)	Continua
volatile acidity	Float64	(0,120 – 1,580)	Continua
citric acid	Float64	(0 – 1)	Continua
residual sugar	Float64	(0,900 – 15,500)	Continua
chlorides	Float64	(0,012 – 0,611)	Continua
free sulfur dioxide	Int64	(1 – 72)	Discreta
total sulfur dioxide	Int64	(6 – 289)	Discreta
density	Float64	(0,990 – 1,004)	Continua
pH	Float64	(2,740 – 4,010)	Continua
sulphates	Float64	(0,330 – 2,000)	Continua
alcohol	Float64	(8,400 – 14,900)	Continua
quality	Int64	(3 – 8)	Discreta

Taula d'informació sobre las variables

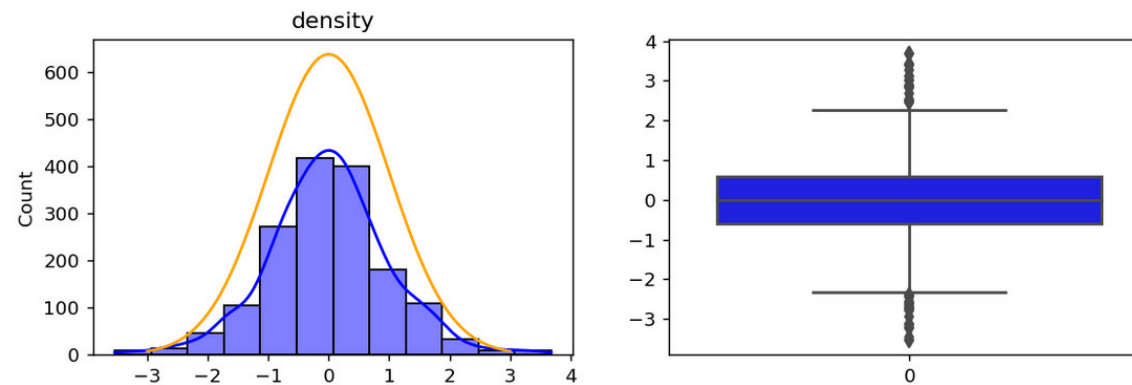
- Gràfiques generades amb la funció `pairplot()` de la llibreria `seaborn`
- Hipòtesis: atributs 'density' i 'pH' segueixen una distribució Gaussiana



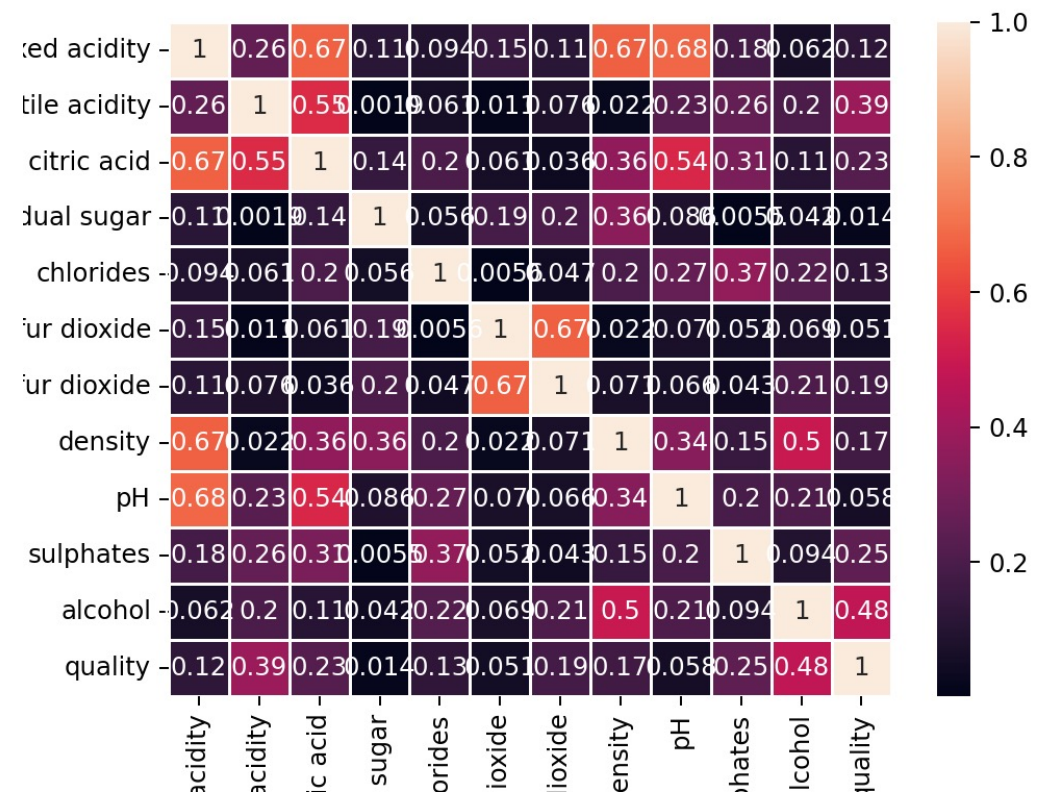
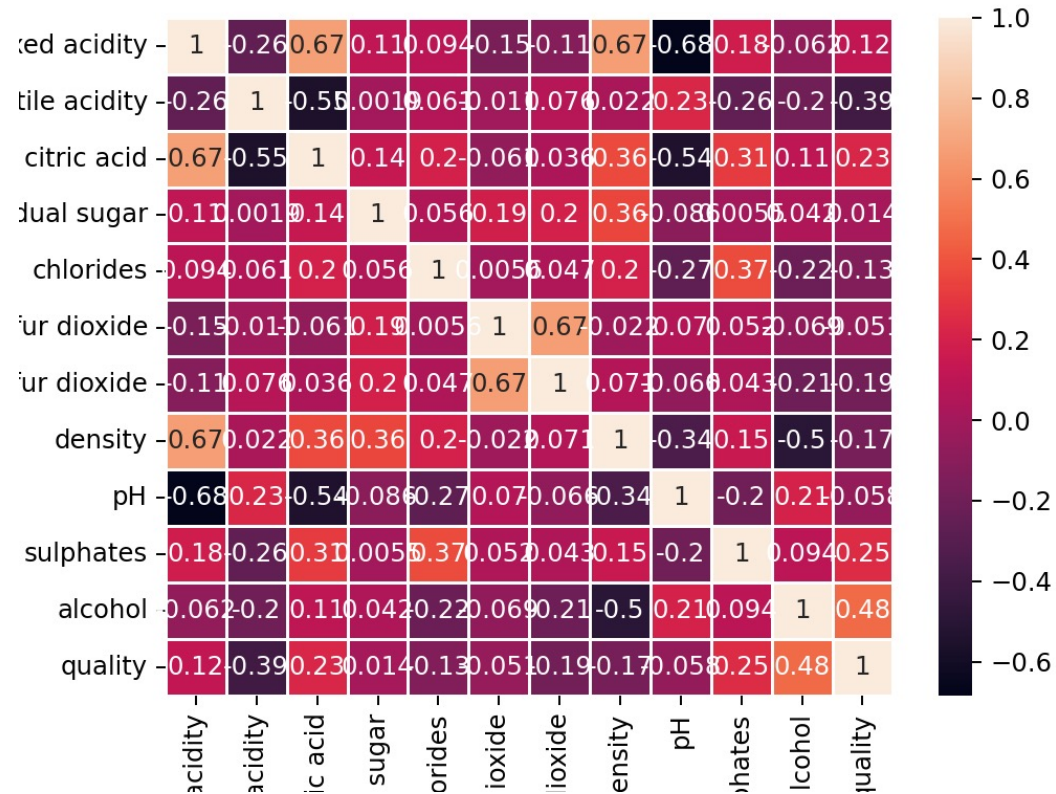
Gràfica de la distribució de l'atribut 'pH'.



Gràfica de la distribució de l'atribut 'density'.



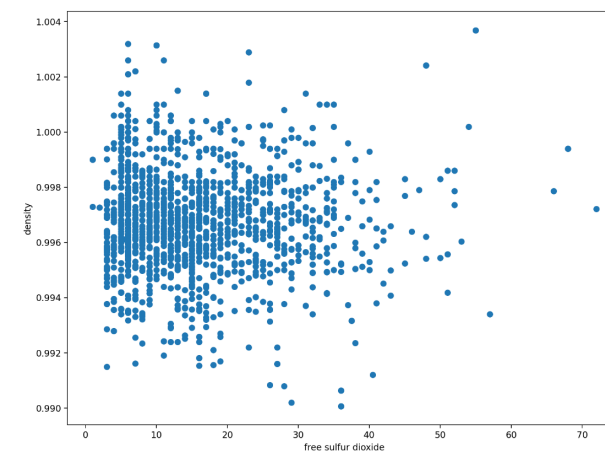
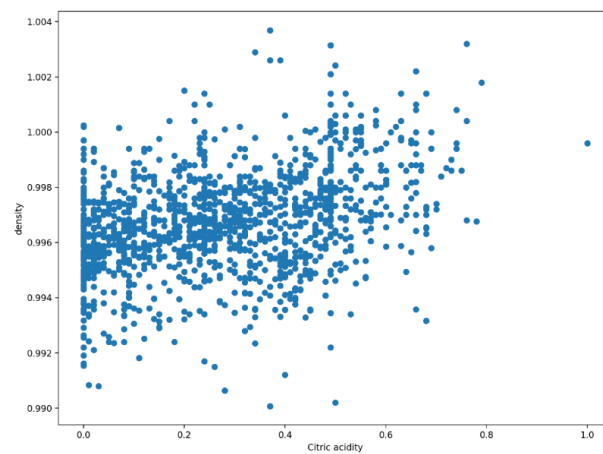
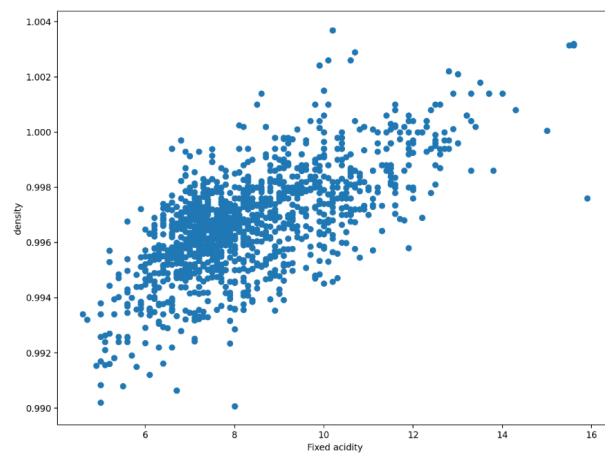
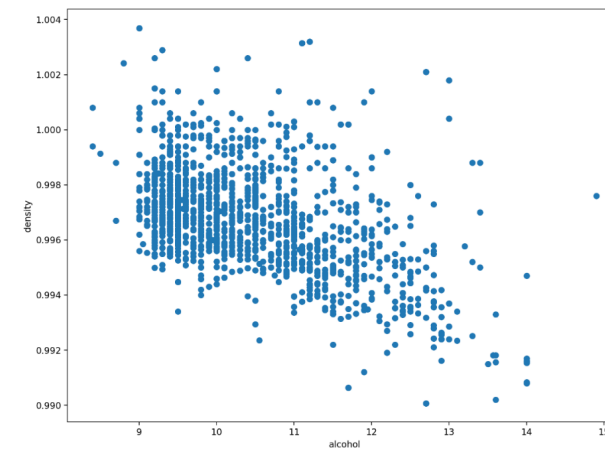
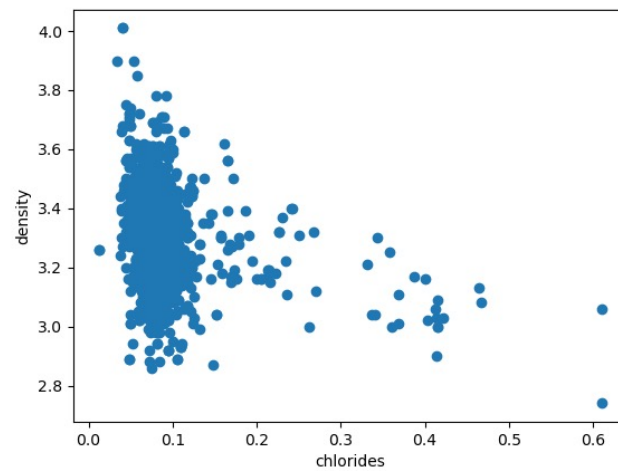
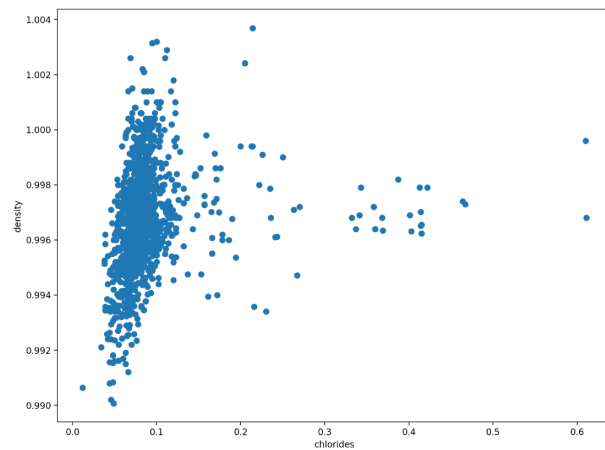
MATRIUS DE CORRELACIÓ

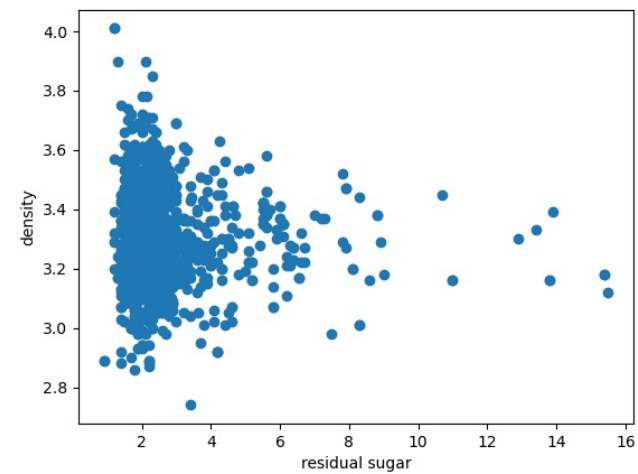
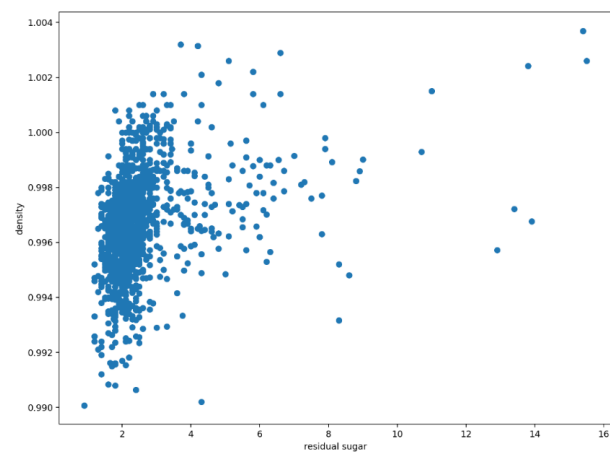
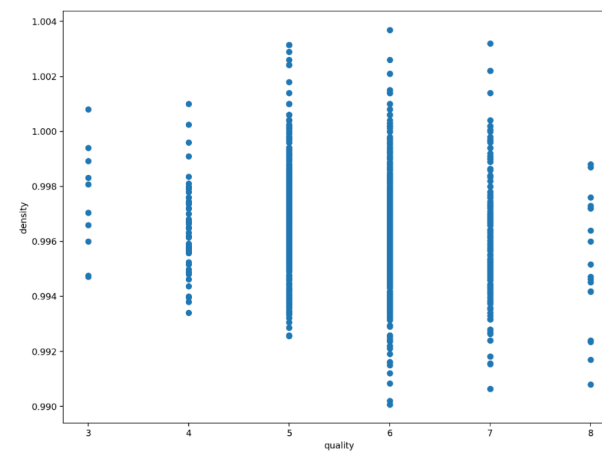
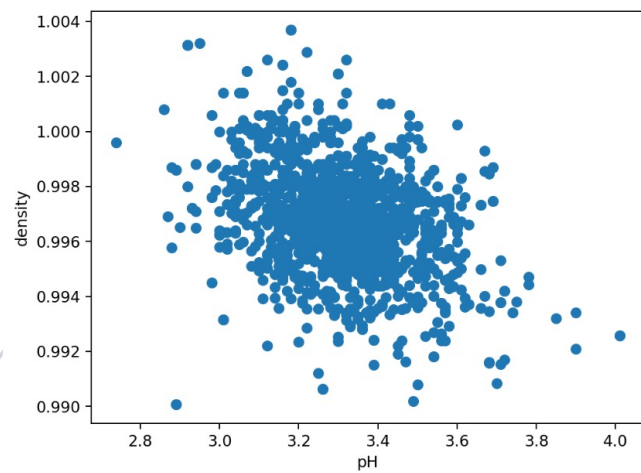


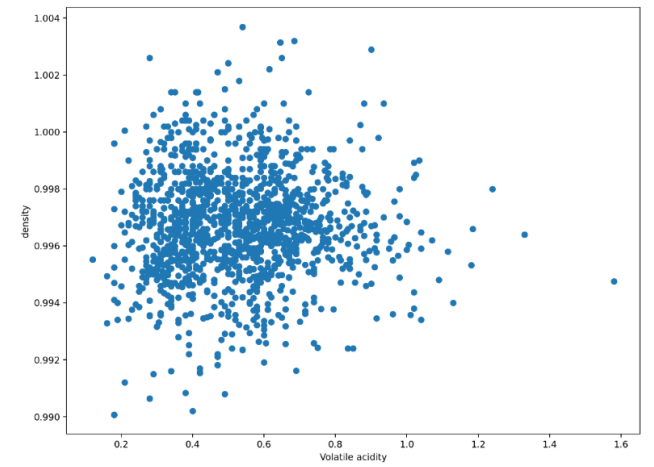
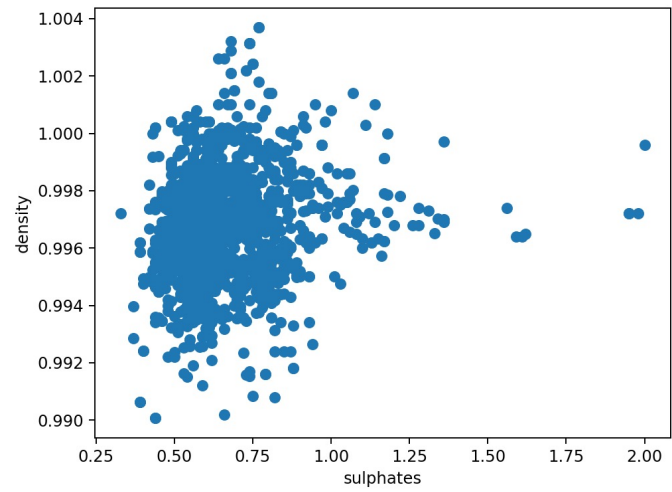
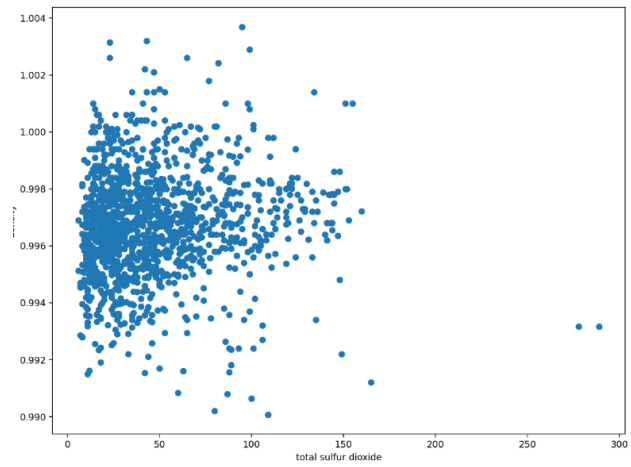


CONTRUCCIÓ DEL REGRESSOR LINEAL

SELECCIÓ D'ATRIBUTS: 'DENSITY' COM A ATRIBUT OBJECTIU

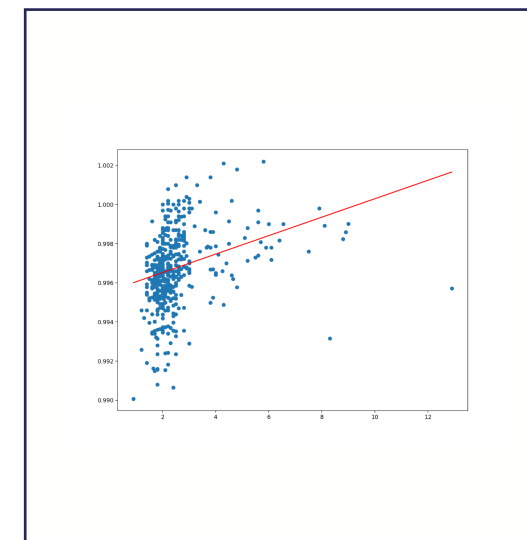
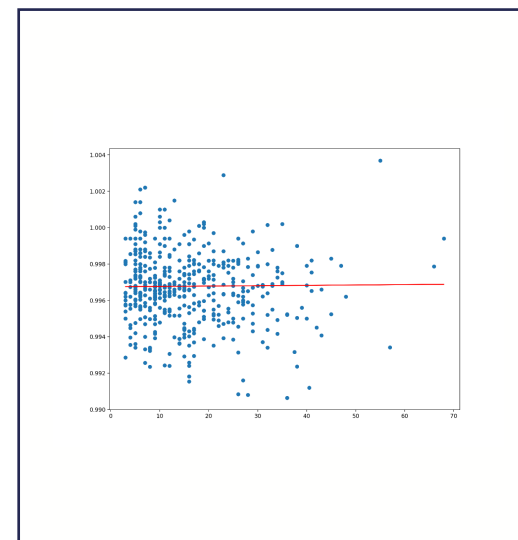
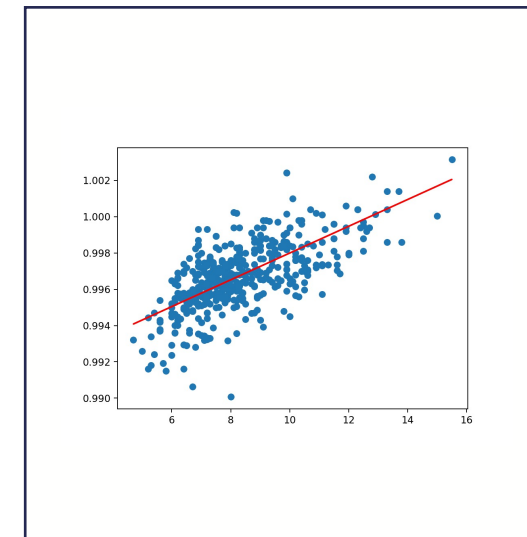
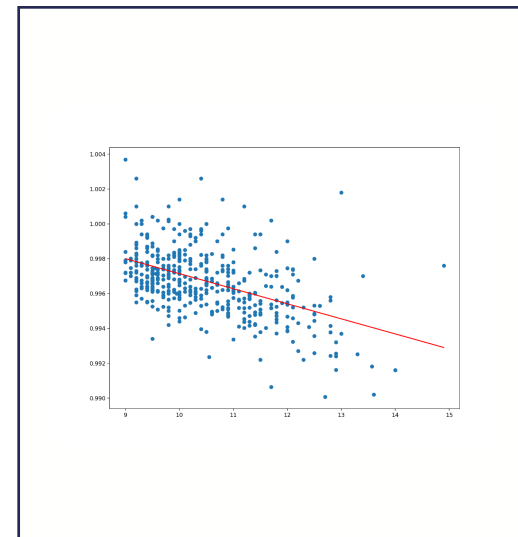






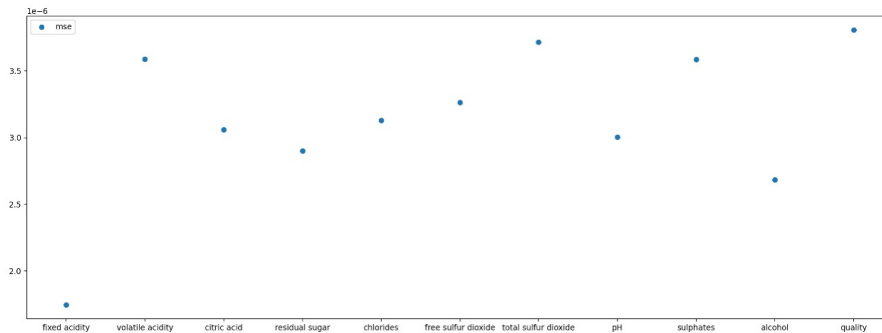
PRIMERES REGRESSIONS

- Amb les dades sense tractar i utilitzant l'error quadràtic mitjà com a mesura de la precisió del regressor

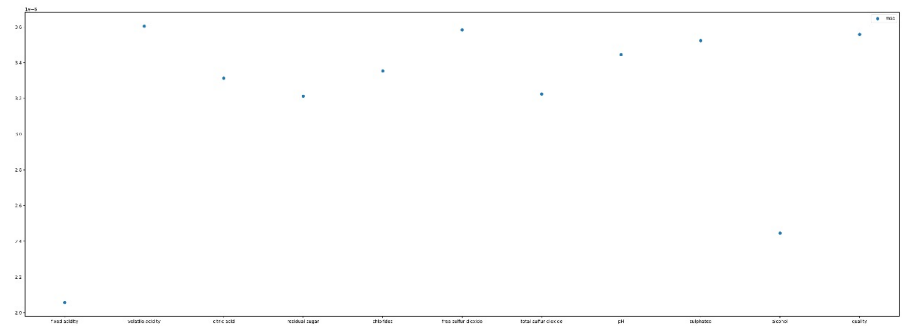


GRÀFIQUES AMB VALORS MSE

Gràfica amb l'error quadràtic mitjà respecte a l'atribut 'density'

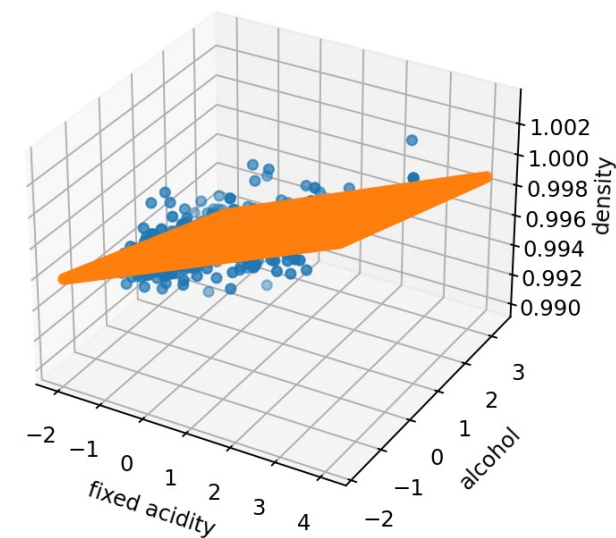
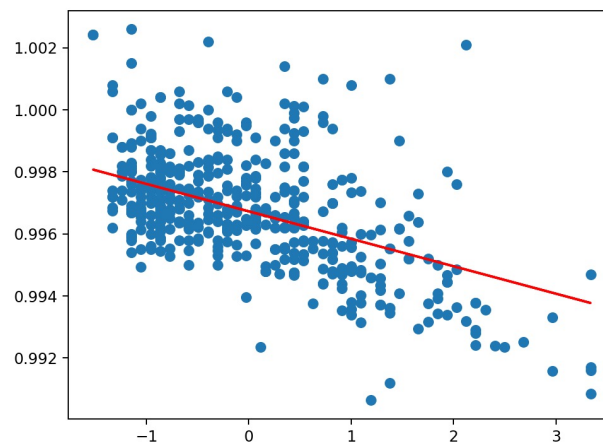
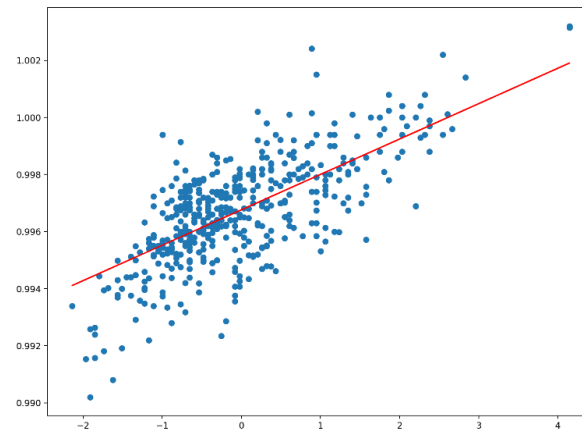


Gràfica amb l'error quadràtic mitjà amb dades estandaritzades



RESULTATS

- L'atribut amb més correlació és 'fixed acidity', el qual obté molt bons resultats



CONCLUSIONS

Anàlisi d'un dataset

Us de 2 mètodes per
modificar dades

Manera més eficient de
realitzar regressions lineals