Tipologia i cicle de vida de les dades

Pràctica 2: Neteja i validació de les dades

L'objectiu d'aquesta activitat serà el tractament del data set "Red Wine Quality", obtingut a partir de la següent font https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009.

1 Descripció del Dataset

Aquest dataset està composat per un conjunt de camps que intenten respondre **quines característiques dels vins son les que més incideixen en la qualitat del vi**.

Les característiques capturades dels vins son les següents:

- **Fixed Acidity**: la majoria dels àcids implicats en la elaboració del vi son fixes o no volàtils. Son àcids que no s'evaporen fàcilment, i que per tant perduren més en el temps.
- **Volatile Acidity**: la quantitat d'àcid acètic en el vi, que a massa nivells pot portar a un gust desagradable del vinagre.
- **Citric Acid**: l'àcid cítric es troba en petites quantitats, i pot afegir frescor i gust als vins.
- **Residual sugar**: fa referencia a la quantitat de sucre restant després de la fermentació, és rar trobar vins amb menys d'un gram per litre, i els vins amb més de 45 grams per litre es consideren dolços.
- **Chlorides**: la quantitat de sal al vi.
- **Free sulfur dioxide**: la quantitat de diòxid de sofre lliure al vi. La forma lliure del SO2 existeix en equilibri entre el SO2 molecular (com un gas dissolt) i el ion bisulfit; que impedeix el creixement microbial i l'oxidació del vi.
- **Total sulfur dioxide**: quantitat total de diòxid de sofre, quantitat de formes lliures i limitades de SO2; en baixes concentracions, el SO2 no es pot detectar en el vi, però a concentracions de SO2 lliures superiors a 50 ppm, SO2 es fa evident al nas i al gust del vi.
- **Density**: la densitat de l'aigua es pròxima a la de l'aigua en funció del percentatge d'alcohol i el contingut de sucre.
- **pH**: el pH descriu com és un vi àcid o bàsic en una escala del 0 (molt àcid) a 14 (molt bàsic); la majoria dels vins són entre 3 i 4 en l'escala de pH.

- **Sulphates**: un additiu vitivinícola que pot contribuir als nivells de gas de diòxid de sofre (SO2), que actua com a antimicrobià i antioxidant.
- **Alcohol**: el percentatge de contingut d'alcohol del vi.
- **Quality**: variable de sortida (basada en dades sensorials, puntuació entre 0 i 10)

2 Cárrega de dades

Descarreguem l'arxiu CSV disponible a Kaggle.com i l'obrim amb Pandas.

Mostrem només els 10 primers registres.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
5	7.4	0.66	0.00	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
6	7.9	0.60	0.06	1.6	0.069	15.0	59.0	0.9964	3.30	0.46	9.4	5
7	7.3	0.65	0.00	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7
8	7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
9	7.5	0.50	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.80	10.5	5

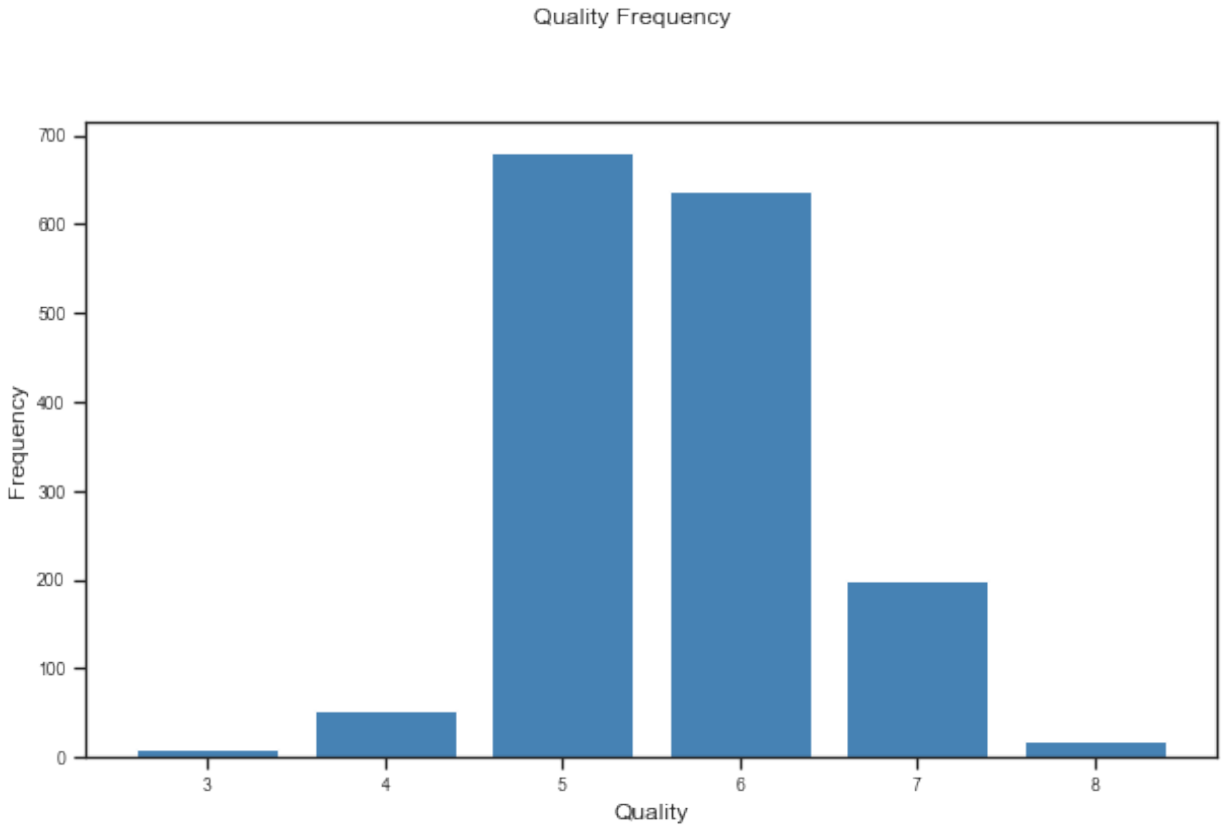
Fem una breu descripció dels valors que pren cada variable.

	count	mean	std	min	25%	50%	75%	max	Data Type
fixed acidity	1599.0	8.319637	1.741096	4.60000	7.1000	7.90000	9.200000	15.90000	float64
volatile acidity	1599.0	0.527821	0.179060	0.12000	0.3900	0.52000	0.640000	1.58000	float64
citric acid	1599.0	0.270976	0.194801	0.00000	0.0900	0.26000	0.420000	1.00000	float64
residual sugar	1599.0	2.538806	1.409928	0.90000	1.9000	2.20000	2.600000	15.50000	float64
chlorides	1599.0	0.087467	0.047065	0.01200	0.0700	0.07900	0.090000	0.61100	float64
free sulfur dioxide	1599.0	15.874922	10.460157	1.00000	7.0000	14.00000	21.000000	72.00000	float64
total sulfur dioxide	1599.0	46.467792	32.895324	6.00000	22.0000	38.00000	62.000000	289.00000	float64
density	1599.0	0.996747	0.001887	0.99007	0.9956	0.99675	0.997835	1.00369	float64
pH	1599.0	3.311113	0.154386	2.74000	3.2100	3.31000	3.400000	4.01000	float64
sulphates	1599.0	0.658149	0.169507	0.33000	0.5500	0.62000	0.730000	2.00000	float64
alcohol	1599.0	10.422983	1.065668	8.40000	9.5000	10.20000	11.100000	14.90000	float64
quality	1599.0	5.636023	0.807569	3.00000	5.0000	6.00000	6.000000	8.00000	int64

Tots els camps son numèrics, les propietats mesurades son valors numèrics continus i la variable de sortida **quality** es entera/discreta.

2.1 Discretització del camp quality

El camp **quality** es més un camp categòric que numèric, amb valors discrets entre 3 i 8.

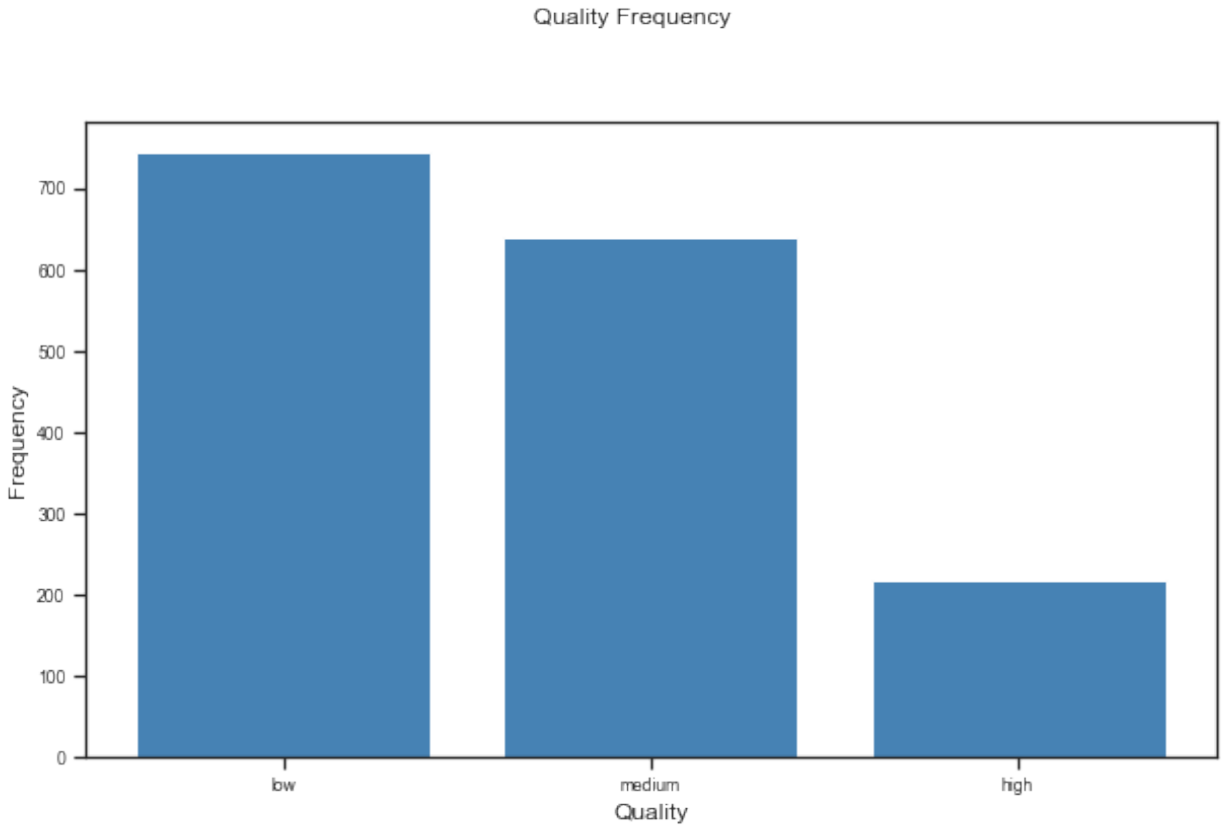


La major part dels vins que estan presents al dataset son d'una qualitat mitja. Això pot comportar que les dades disponibles no siguin suficients per tal de classificar els vins de baixa i alta qualitat.

Per tal de facilitar l'anàlisi crearem una característica derivada del camp **qualitat**.

Nou atribut/característica afegides al data set original:

- **quality_cat**: es tracte d'un atribut derivat de la característica **quality**. Creem 3 grups diferenciats de vins en funció de la seva qualitat. Els vins que tenen una qualitat igual o inferior a 5 s'etiquetaran com vins d'una **qualitat baixa**, els vins amb una qualitat enregistrada de 6 s'etiquetaran com vins d'una **qualitat mitja**, i per últim, els vins amb una qualitat igual o superior a 7 els considerarem vins d'una **alta qualitat**.



Amb la discretització aconseguim equilibrar els casos de vins amb baixa i mitja qualitat, però encara continuem tenint un número molt baix de vins de alta qualitat.

3 Data Cleaning

Procedirem a crear un dataset amb els següents camps que ens ajudaran a identificar la qualitat de les dades amb les que treballem:

- Tipus de dades
- Número de valors nuls (missings)
- Número de zeros
- Número de calors extrems
- Número de valors únics

	Types	Uniques	Missings	Zeros	Outliers
fixed acidity	float64	96	0	0	49
volatile acidity	float64	94	0	0	19
citric acid	float64	80	0	132	1
residual sugar	float64	91	0	0	155
chlorides	float64	37	0	0	151
free sulfur dioxide	float64	60	0	0	30
total sulfur dioxide	float64	144	0	0	55
density	float64	2	0	0	0
pH	float64	89	0	0	35
sulphates	float64	96	0	0	59
alcohol	float64	65	0	0	13
quality	int64	6	0	0	28
quality_cat	category	3	0	0	0

Observacions:

- Els camps **residual sugar** (155) i **chlorides** (151) son tres camps que presenten un número molt elevat de valors atípics. Els valors atípics poden afectar negativament al càlcul de les mitges i les seves variàncies.
- El camp **citric_acid** es l'únic camp que presenta zeros en una quantitat elevada (132). Podrien ser errors de captura de les dades o d'entrada de dades. Encara que també podrien ser valor vàlids. En cas de no ser valors correctes hauríem de: obviar el camp, eliminar totes les entrades amb zeros, imputar els valors en base als altres casos (fent servir, per exemple, l'algoritme *K-Nearest Neighbors*)

Les dades **no presenten cap valor buit**. Per tal de resoldre valors buits hauríem d'aplicar les mateixes regles que en el punt anterior.

3.1 Imputar valors pel camp "Citric Acid"

Anem a imputar els valors dels casos que contenen zero com a valor pel camp **citric_acid**, i farem servir l'algoritme K-NN amb K = 5.

[**Suposició!!**] L'àcid cítric es un àcid fixe que prové de la pròpia polpa del raïm. Això vol dir que es estrany que aquest no estigui present i que s'hagi evaporat absolutament durant el procés de fermentació. Per aquest motiu intentarem regenerar els valors no informats del àcid cítric (aquells que contenen 0) a partir de les altres variables relacionades amb el nivell d'acid: **fixed acid**, **volatile acidity** i **citric acid**.

Una vegada aplicat l'algoritme K-NN sobre els valors nulls, obtenim un nou dataset sense zeros al camp **citric acid**:

	Types	Uniques	Missings	Zeros	Outliers
fixed acidity	float64	96	0	0	49
volatile acidity	float64	94	0	0	19
citric acid	float64	79	0	0	1
residual sugar	float64	91	0	0	155

	Types	Uniques	Missings	Zeros	Outliers
chlorides	float64	37	0	0	151
free sulfur dioxide	float64	60	0	0	30
total sulfur dioxide	float64	144	0	0	55
density	float64	2	0	0	0
pH	float64	89	0	0	35
sulphates	float64	96	0	0	59
alcohol	float64	65	0	0	13
quality	int64	6	0	0	28
quality_cat	category	3	0	0	0

4 Anàlisi Descriptiu

Les estadístiques descriptives proporcionen resums senzills de les dades.

- La mitja (aritmètica) calcula el valor típic del nostre conjunt de dades. No és robust, sent impactat pels valors extrems.
- La mitjana és el valor mitjà exacte del nostre conjunt de dades. És robust, no sent impactat pels valors extrems.
- La moda és el valor que més apareix.
- El rang és la diferència entre el valor més gran i el més petit del nostre conjunt de dades.
- La variància i la desviació estàndard són la distància mitjana de la mitja.

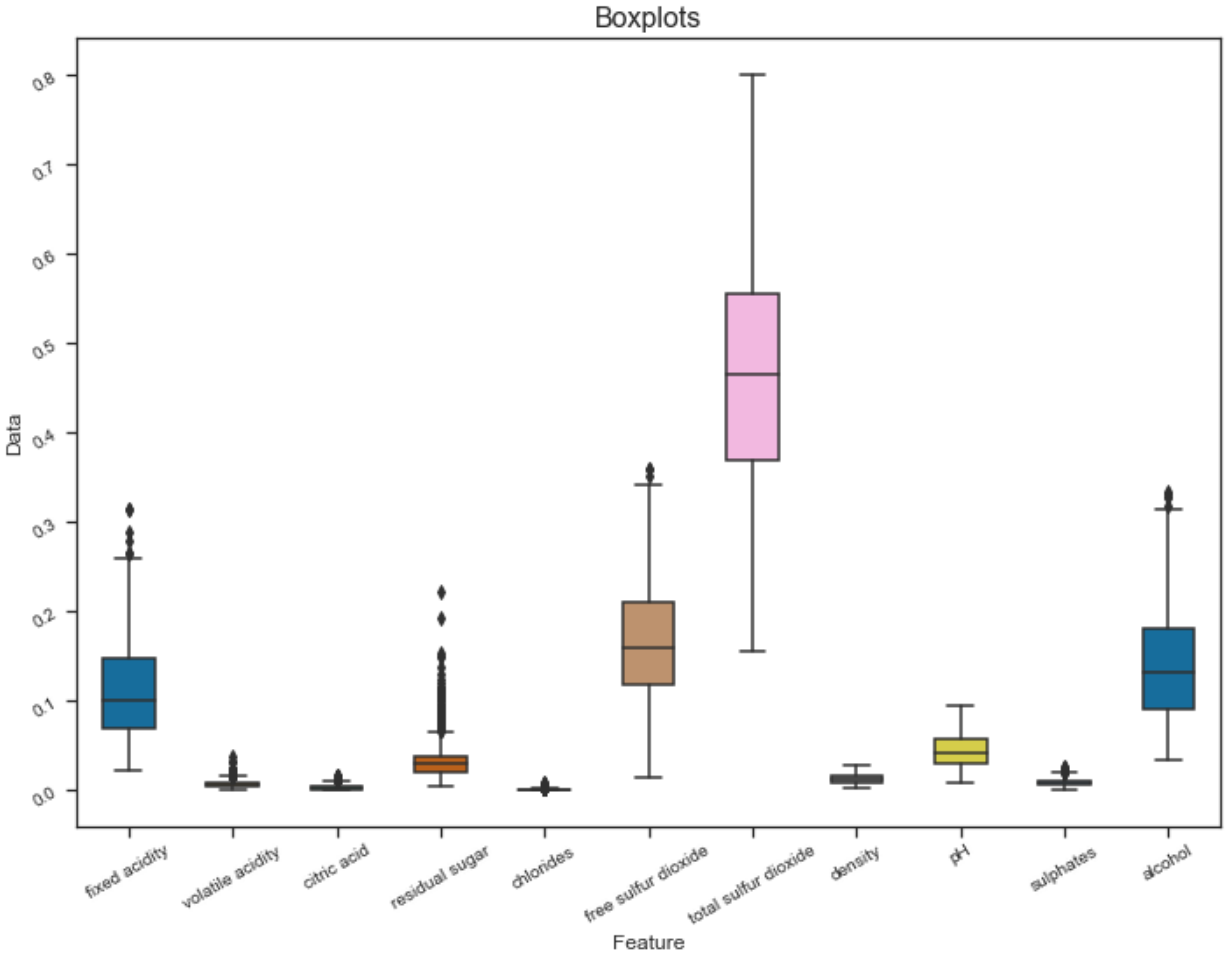
	count	mean	std	min	25%	50%	75%	max	median	var	mode
fixed acidity	1599.0	8.319637	1.741096	4.60000	7.1000	7.90000	9.200000	15.90000	7.90000	3.031416	7.2
volatile acidity	1599.0	0.527821	0.179060	0.12000	0.3900	0.52000	0.640000	1.58000	0.52000	0.032062	0.6
citric acid	1599.0	0.281119	0.184689	0.01000	0.1200	0.26000	0.420000	1.00000	0.26000	0.034110	0.49

	count	mean	std	min	25%	50%	75%	max	median	var	mode
residual sugar	1599.0	2.538806	1.409928	0.90000	1.9000	2.20000	2.600000	15.50000	2.20000	1.987897	2
chlorides	1599.0	0.087467	0.047065	0.01200	0.0700	0.07900	0.090000	0.61100	0.07900	0.002215	0.08
free sulfur dioxide	1599.0	15.874922	10.460157	1.00000	7.0000	14.00000	21.000000	72.00000	14.00000	109.414884	6
total sulfur dioxide	1599.0	46.467792	32.895324	6.00000	22.0000	38.00000	62.000000	289.00000	38.00000	1082.102373	28
density	1599.0	0.996747	0.001887	0.99007	0.9956	0.99675	0.997835	1.00369	0.99675	0.000004	0.9972
pH	1599.0	3.311113	0.154386	2.74000	3.2100	3.31000	3.400000	4.01000	3.31000	0.023835	3.3
sulphates	1599.0	0.658149	0.169507	0.33000	0.5500	0.62000	0.730000	2.00000	0.62000	0.028733	0.6
alcohol	1599.0	10.422983	1.065668	8.40000	9.5000	10.20000	11.100000	14.90000	10.20000	1.135647	9.5
quality	1599.0	5.636023	0.807569	3.00000	5.0000	6.00000	6.000000	8.00000	6.00000	0.652168	5

5 Anàlisi Univariant

Normalitzem els resultats per tal de poder visualitzar-los amb la mateixa escala.

Generem els boxplots amb les dades normalitzades.



Observacions:

- Els camps **residual sugar** i **chlorides** son els camps que presenten més valors extrems. Haurem de fixar-nos en aquests casos, igual pertanyen a una població de vins diferent, sigui per la regió a la que pertanyi el vi, pel tipus de raïm, etc.

Anàlisis de la Normalitat

El **skewness** mesura la proporció entre les cues (tails) de dades, es una mesura de simetria. Un skew positiu +X vol dir que la cua de la dreta de la distribució acumula X% mes de valors que la de la dreta. I -X al contrari.

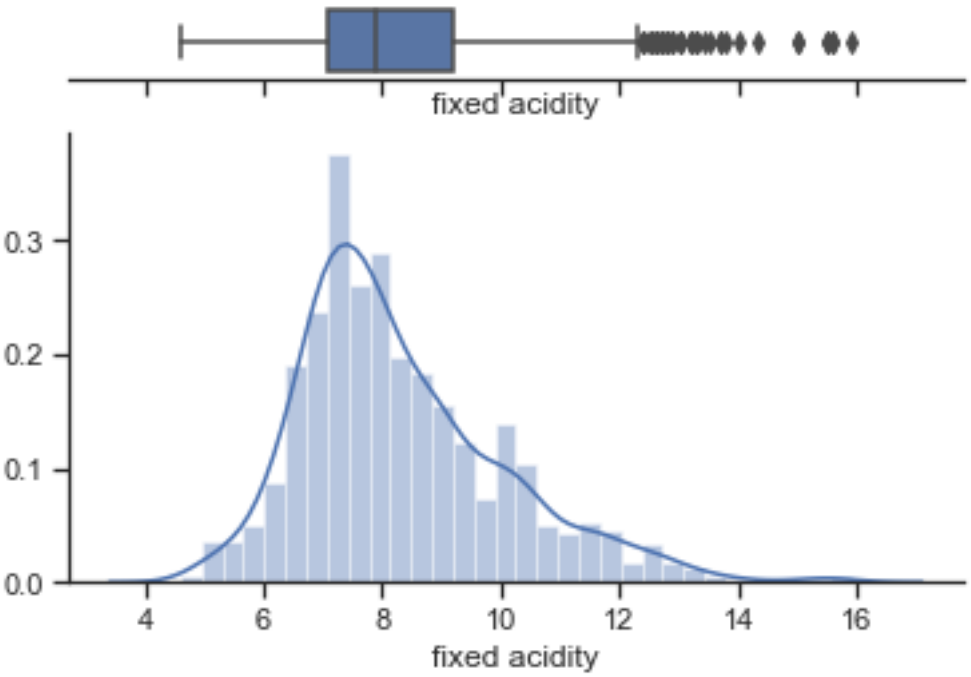
Kurtosis mesura la quantitat de probabilitat que s'acumula a les cues (tails). El valor de kurtosis de una distribució normal es 3, llavors si el nostre kurtosis es superior voldrà dir que les nostres cues son molt més grans, acumulen molt més valors (probabilitat), i un valor menor voldrà dir que existeix una major concentració de les dades al voltant de la part central, sent les cues menors (menor probabilitat).

Calculem també els índex de *skewness* i *kurtosis*:

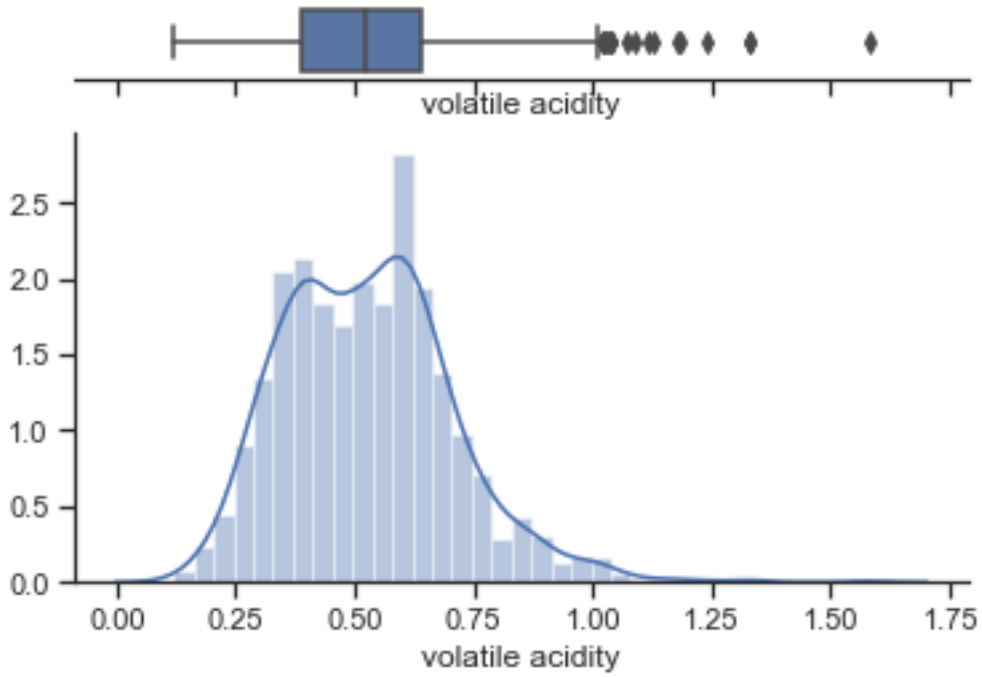
	Skewness	Kurtosis
fixed acidity	0.982751	1.13214
volatile acidity	0.671593	1.22554
citric acid	0.394456	-0.664411
residual sugar	4.54066	28.6176
chlorides	5.68035	41.7158
free sulfur dioxide	1.25057	2.02356
total sulfur dioxide	1.51553	3.80982
density	0.0712877	0.934079
pH	0.193683	0.806943
sulphates	2.42867	11.7203
alcohol	0.860829	0.200029
quality	0.217802	0.296708
quality_cat	-	-

Histogrames

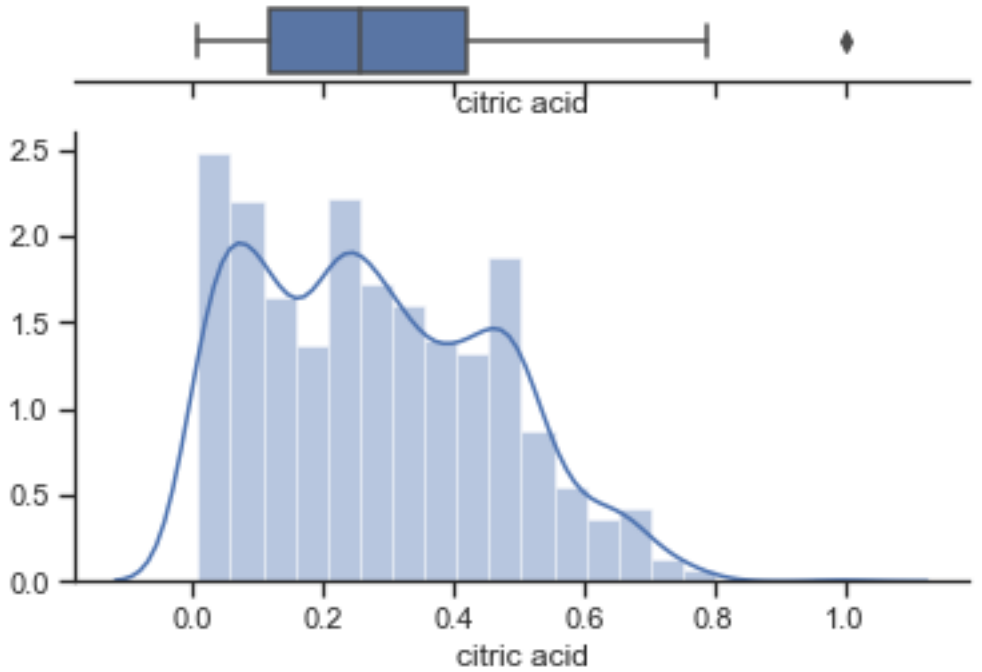
Generem les visualitzacions de *histogrames*, *densitat* i *boxplots* dels atributs d'entrada.



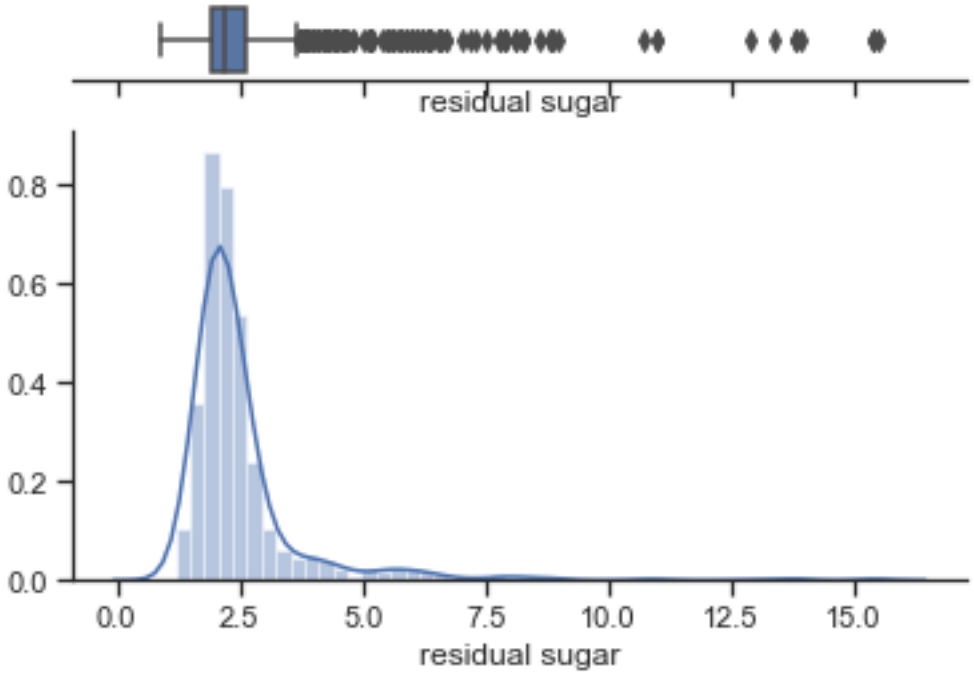
Fixed acidity: La distribució mostra un skewness positiu (0.98), asimètrica. La mitjana està al voltant de 7.9, i la mitja es troba al 8.32. La mitja està moguda cap a la dreta pels valors atípics que tiren una mica d'ella. La gran quantitat de vins es concentren al voltant de la mitja en base a la seva acidesa fixa.



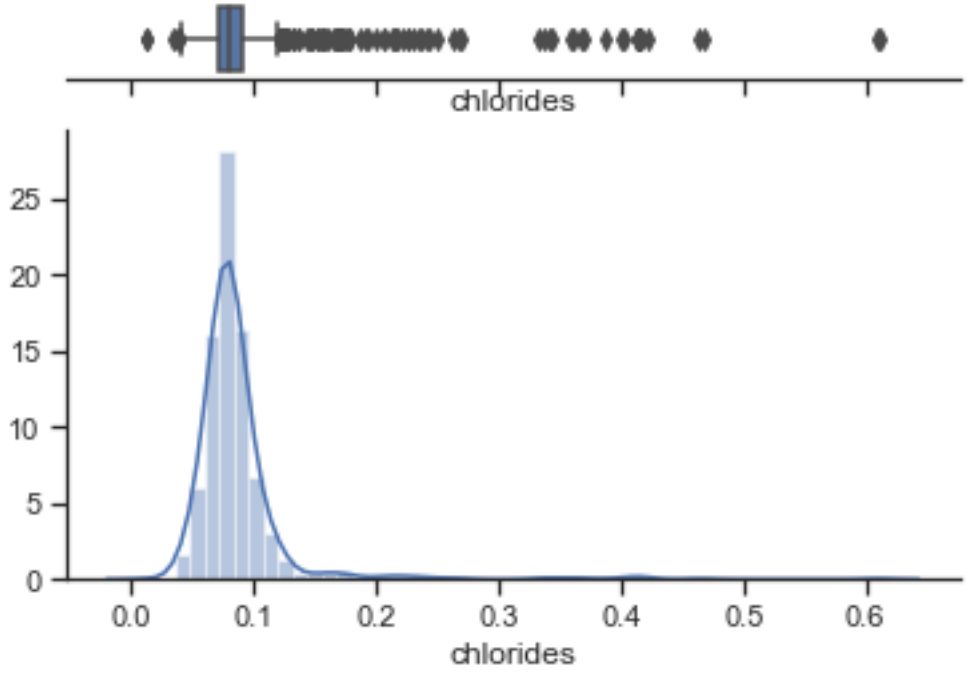
Volatile acidity: presenta una distribució més del tipus bimodal amb dos pics al voltant del 0.38 i 0.63.



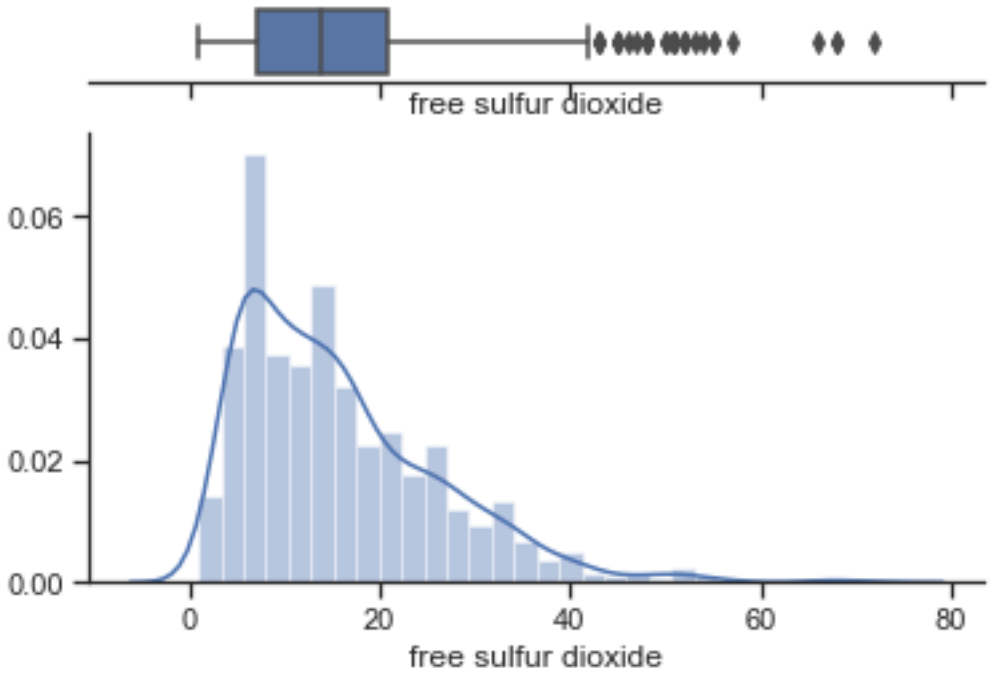
Citric acid: mostra una distribució sense una forma molt clara, els valors semblen no seguir cap distribució. També presenta pocs valors molt a l'extrem. Pot-ser aquest camp no ens ajudi massa a l'hora de identificar la qualitat d'un vi. Es un candidat a ficar fora de la selecció final.



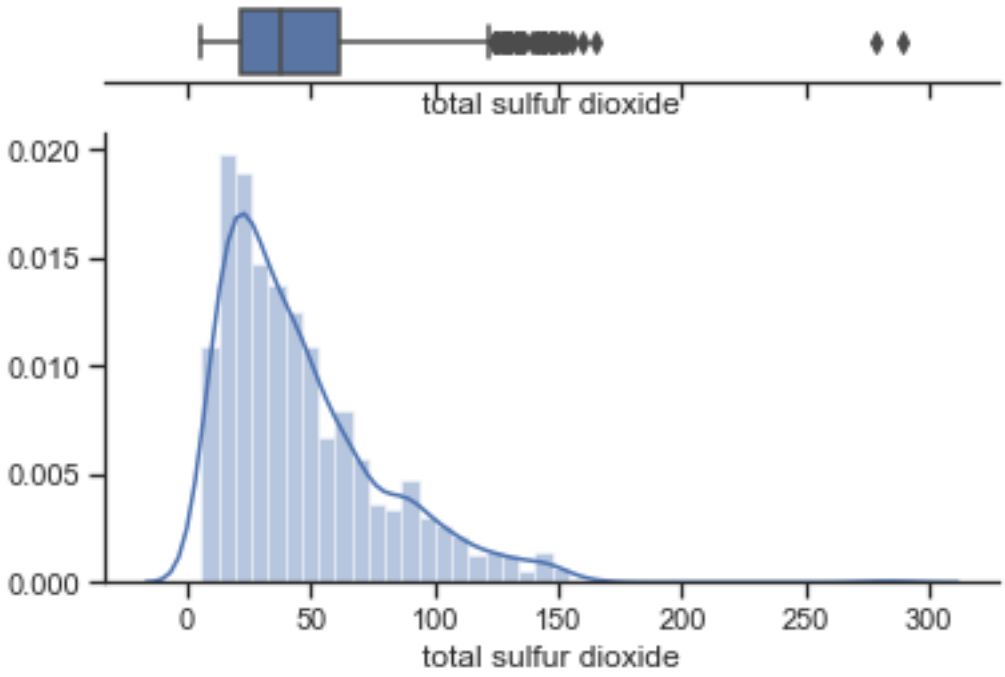
Residual sugar: la distribució mostra un skewness positiu bastant important (4.54), i un índex de kurtosis molt gran (28.62) ja que la cua de la dreta conté molts més valors que la cua de l'esquerra, presentant una cua bastant llarga, producte també de l'existència de molts valors grans extrems. Molts dels valors es concentren al pic de 2.2 (mitjana).



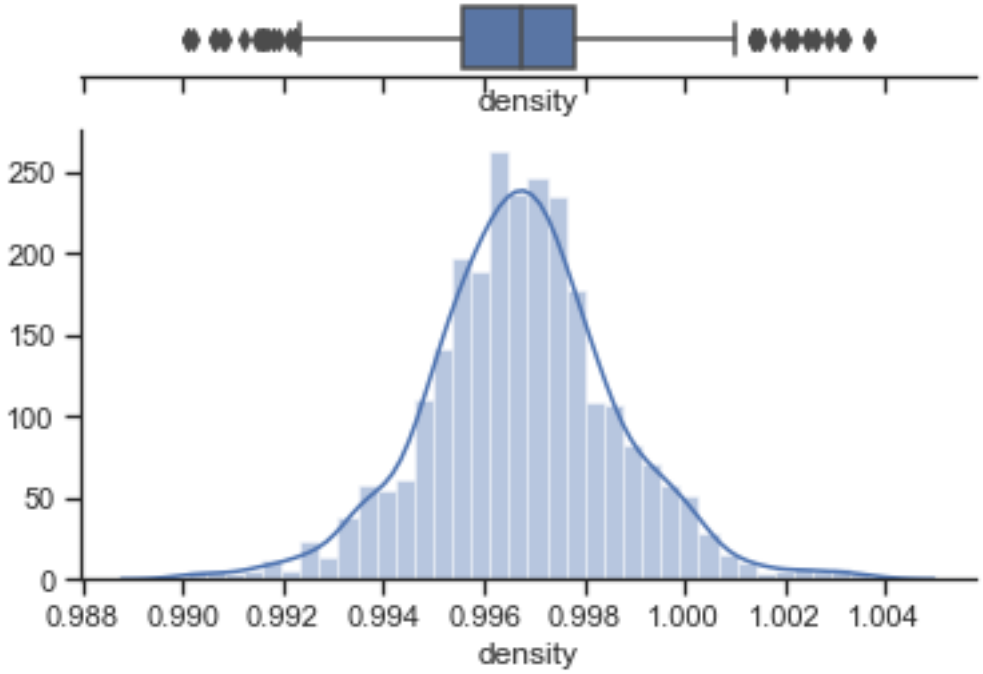
Chlorides: al igual que en el cas del **residual sugar**, la distribució presenta una gran distorsió positiva de (5.86) i un índex de kurtosis molt més elevat (45) provocat per una gran quantitat de valors extrems a la part dreta de la distribució. La major part dels vins es concentren al pic de 0.08.



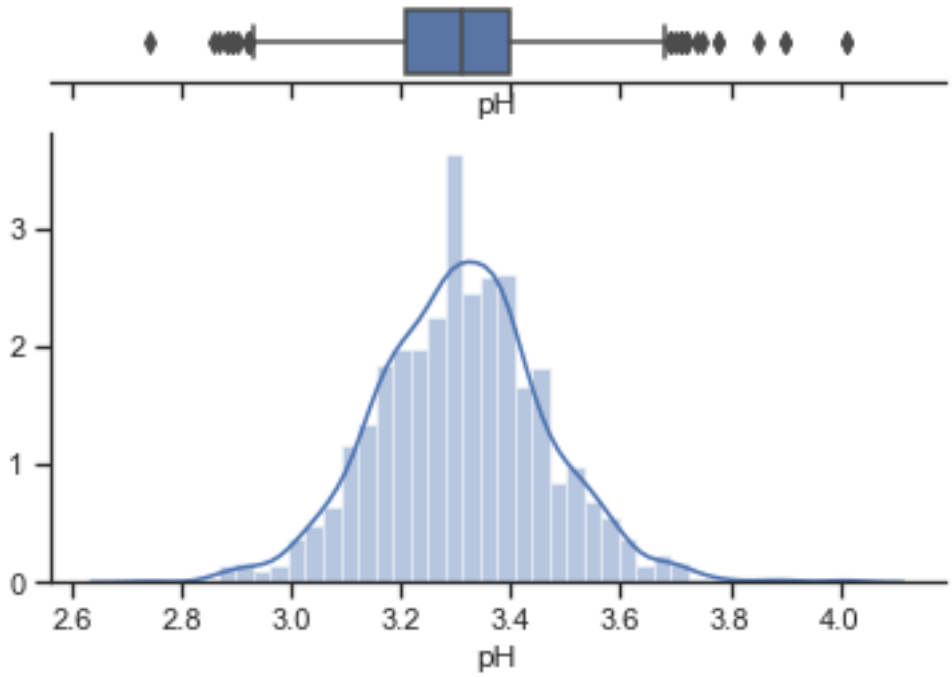
Free sulfur dioxide: mostra tb una distribució normal amb una forta distorsió positiva (1.26). Sembla que existeix una concentració important de vins al pic al voltant de 7, i que conforme els valors es fan mes grans la distribució va decreixent lentament, no es una caiguda brusca. La qual del costat de valors grans es una cua llarga. Això i l'existència d'alguns valors extrems fan que la mitja de quasi 16 es trobi molt més cap a la dreta de la mitjana al 14.



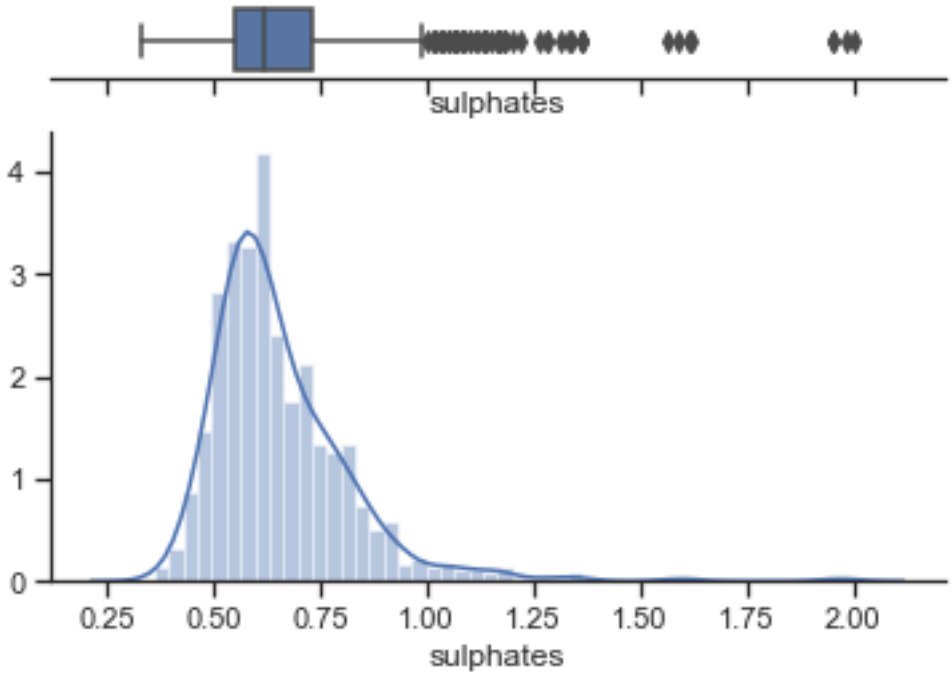
Total sulfur dioxide: distribució molt similar a la presentada al camp **free sulfur dioxide**. Presenta una distorsió positiva important amb una cua al cantó dels valors grans bastant llarga.



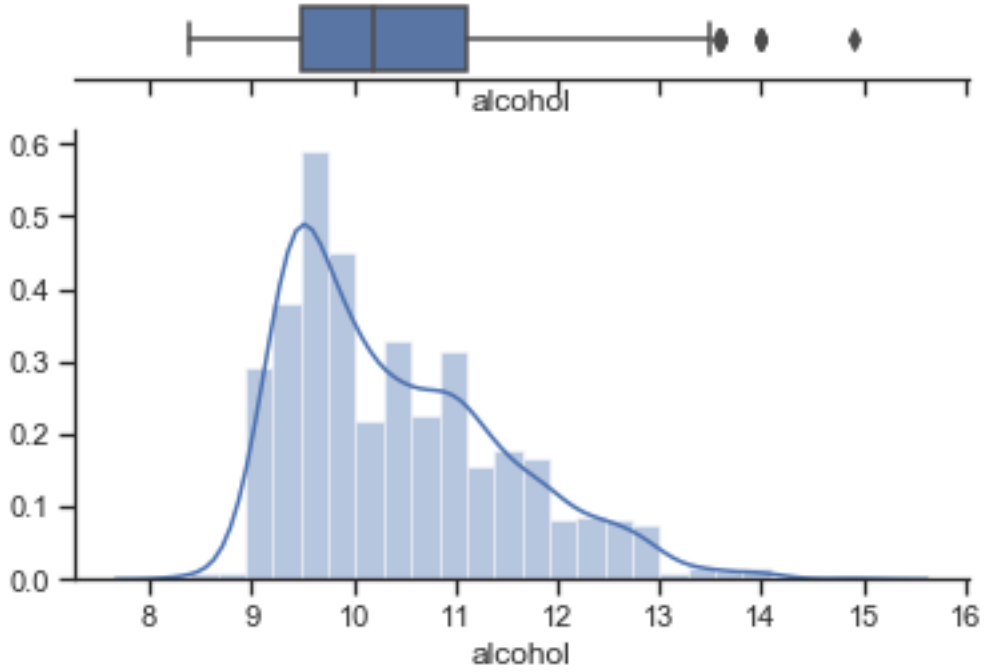
Density: mostra un comportament de distribució normal, amb un índex de skew de quasi 0.



pH: també presenta un comportament de distribució normal.



Sulphates: presenta una distribució normal amb una distorsió positiva (2.43) i una llarga cua cap als valors més grans, generada principalment per l'existència de pocs valors bastant extrems. El tipus de distorsió es bastant similar a les presentades pels camps **free sulfur dioxide** (1.25) i **total sulfur dioxide** (1.52).



Alcohol: presenta una distribució normal amb una lleugera distorsió positiva (0.87) en la línia dels camps **fixed acidity** o **volatile acidity.**

5.1 Resum

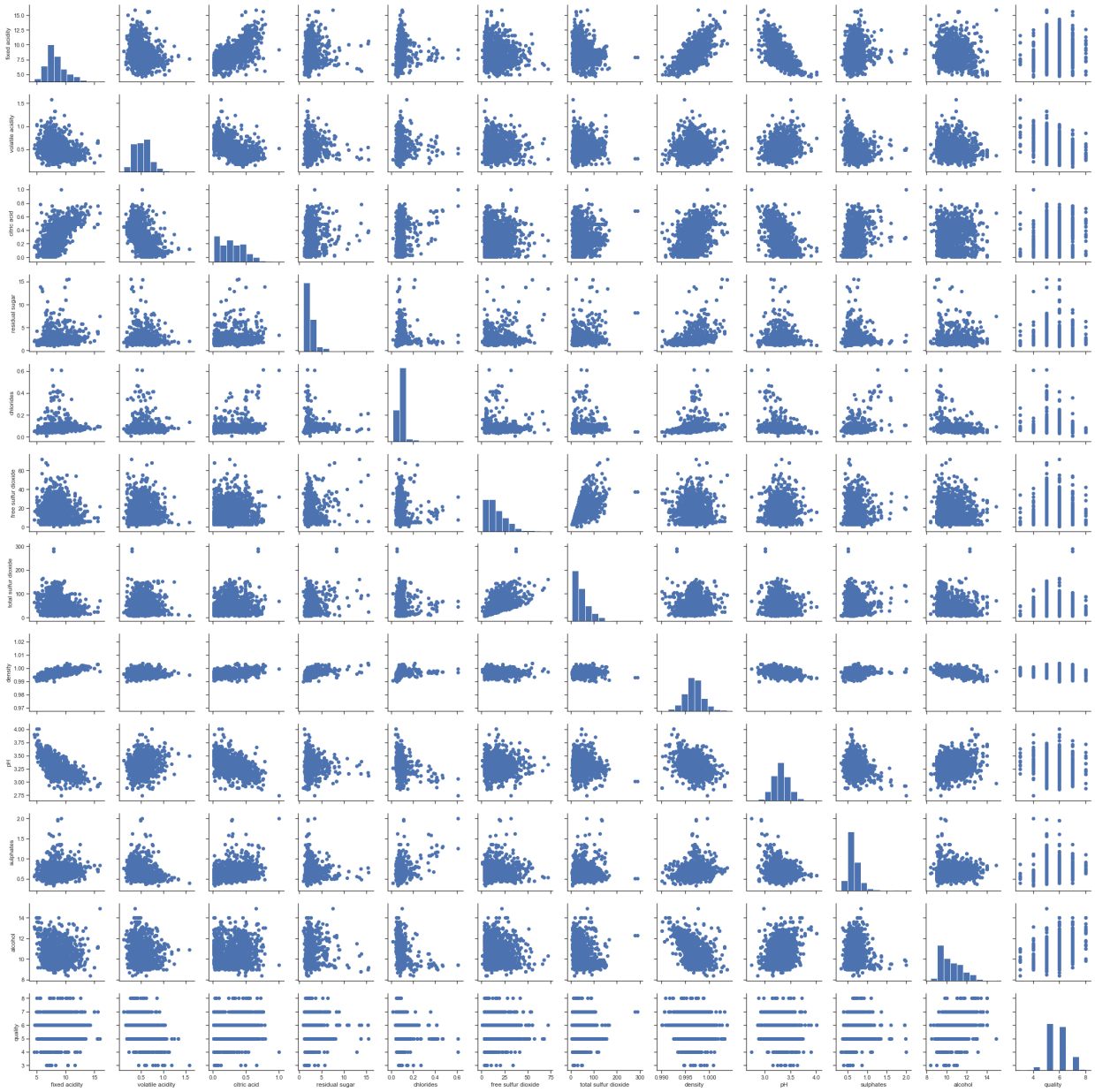
- En relació a les distribucions trobem que els camps **density**, **pH**, i **volatile acidity** mostren un comportament de distribució normal estàndard.
- Els camps **alcohol**, **fixed acidity**, **free sulfur dioxide**, **sulphates**, **total sulfur dioxide** mostren unes distribucions normals amb una tendència a la distorsió positiva, inclinades cap al cantó esquerra (valor més petits).
- El camp **citric acid** es el camp que presenta una distribució molt anormal. De tots els camps sembla el camp menys informatiu.
- Els camps **chlorides** i **residuals** mostren una concentració molt gran al voltant de la mitjana amb una quantitat gran de valors extrems que poden dificultar el anàlisis.

6 Anàlisis Bivariant

Farem un anàlisis de correlació per tal de veure els camps que estan fortament correlacionat.

6.1 Calculant la correlació entre camps

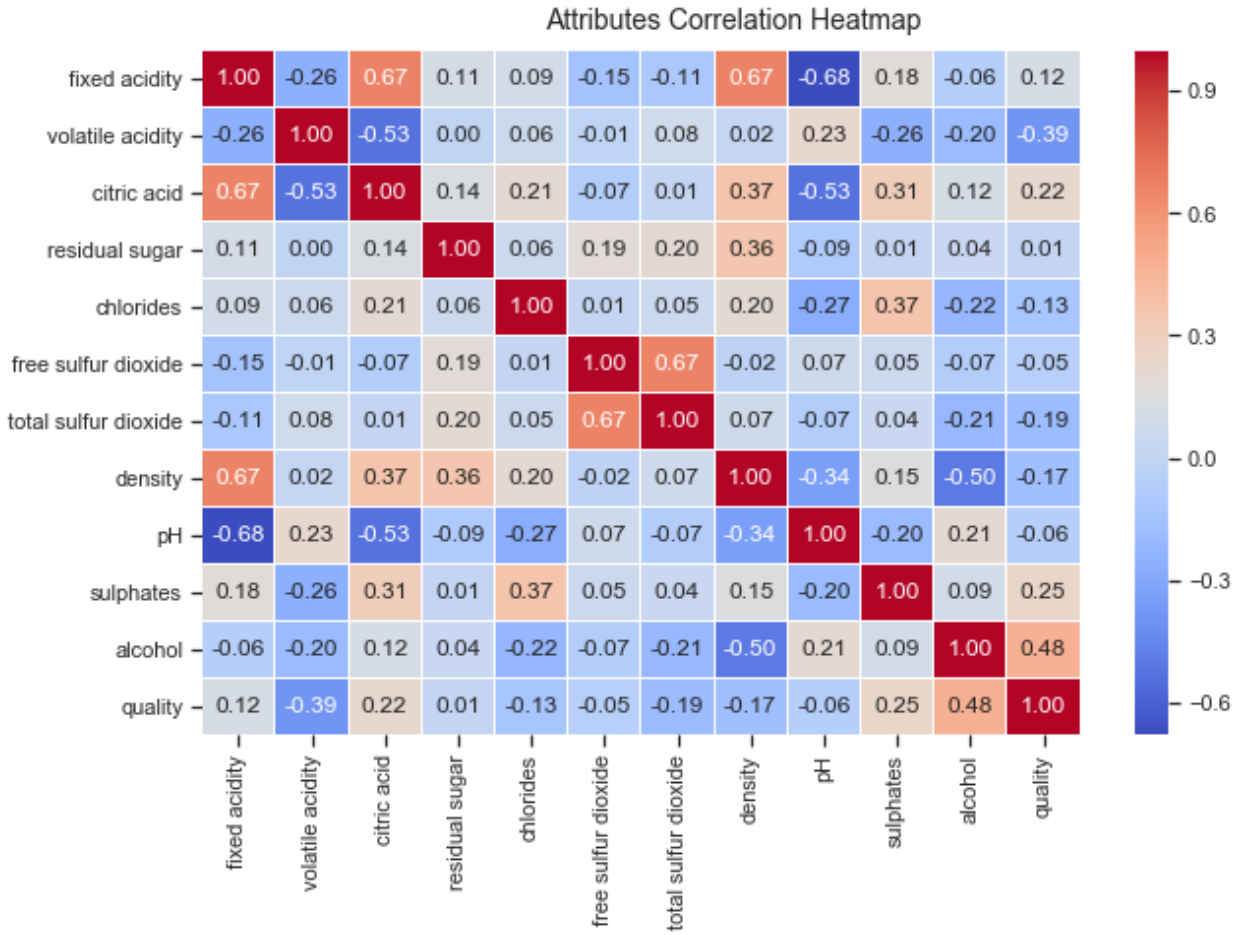
Els camps correlacionats, en general, no ajuden a millorar la qualitat dels models (això dependrà molt del problema així com el número de variables i el seu grau de correlació). Per aquest motiu pot ser bastant útil eliminar camps correlacionats abans d'entrenar qualsevol model de classificació o regressió (quality_cat o quality respectivament).



Observacions:

- Gràficament amb els scatter es fa una mica complicat veure les correlacions existents.

Calcularem la matriu de correlacions i mostrarem un heatmap per visualitzar-les millor.



Observacions:

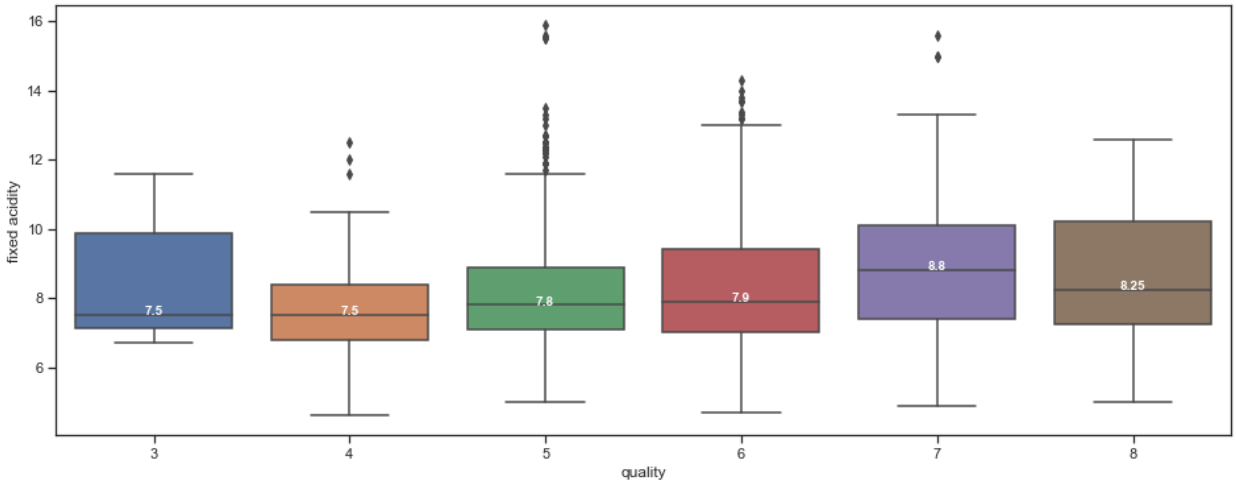
- El camp **fixed acidity** sembla que està positivament correlacionat amb els camps **citric acid** i **density**, a més **fixed acidity** major la **densitat** i la quantitat de **citric_acid**. Lo que es normal, ja que el àcid cítric es un dels tipus de àcids fixes que estan presents als raïms (tartàric, màlic, cítric, i succínic son els més importants).
- Al mateix temps, sembla que el camp **fixed acidity** i **citric acid** està negativament correlacionat amb el **pH**. Una cosa completament normal ja contra més acida es una solució més pròxim de 0 es troba el pH.
- Sembla que existeix una correlació negativa entre els camps **volatile acidity** i el **citric acid**. Els àcids volàtils son creats durant el procés de fermentació. A més gran proporció d'àcids volàtils més petita es la proporció d'àcid cítric.

- Com es també lògic, existeix una forta correlació positiva entre els camps **free sulfur dioxide** i **total sulfur dioxide**, ja que un està inclòs a l'altre.
- Existeix una molt forta correlació positiva molt forta entre la **density** i el grau de **alcohol** dels vins. A més alcohol la densitat del vi augmenta.
- Els camps que semblen estar més directament correlacionats amb la qualitat del vi son:
  - correlació positiva: vins més alcohòlics (**alcohol**) presenten una millor qualitat.
  - correlació negativa: vins que presenten una quantitat d'àcids volàtils obtenen una pitjor qualitat. Això es degut a que quantitats apreciables d'àcid volàtil son un signe de deterioració del vi.

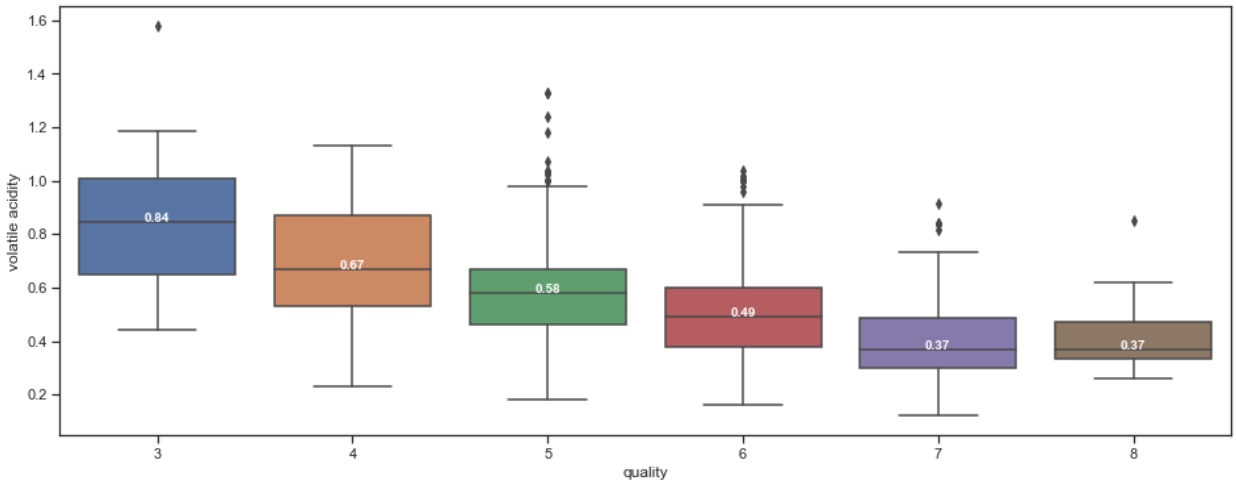
6.2 Visualitzar relació dels camps amb la qualitat dels vins

Ara generarem uns boxplots per veure gràficament com varia la qualitat dels vins en funció de cada característica. Compararem la mitjana a cada nivell de qualitat, si existeix una variació considerable, llavors podem deduir que la característica té un impacte rellevant en la qualitat dels vins.

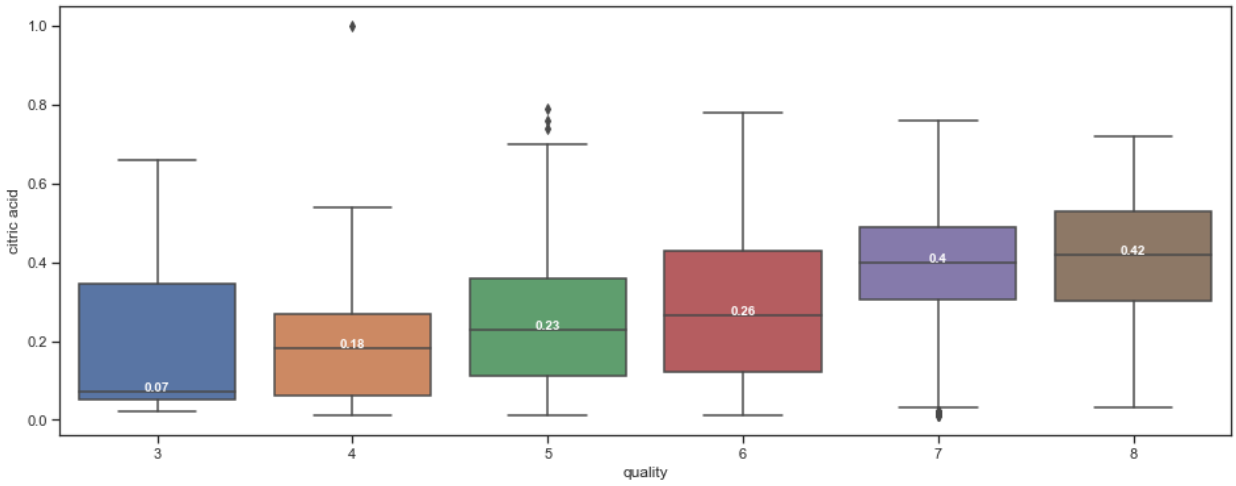
Comparem les mitjanes perquè és una mesura de centre robusta, que no s'altera amb l'existència de valors extrems.



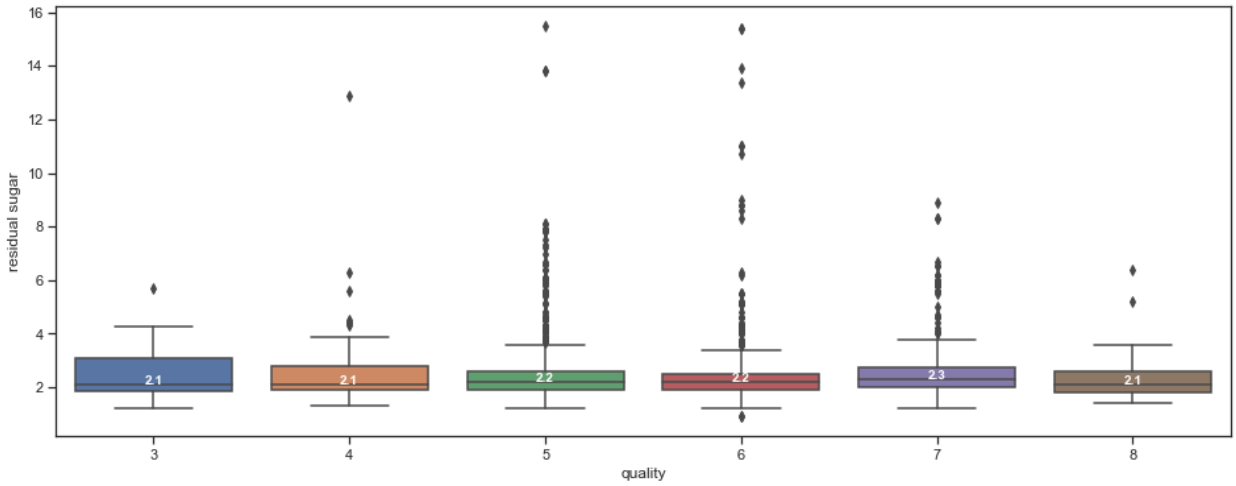
Fixed acidity: no sembla que aquest camp tingui incidència en la qualitat dels vins. La mitjana no presenta valors diferents (substancialment) a cada nivell de qualitat.



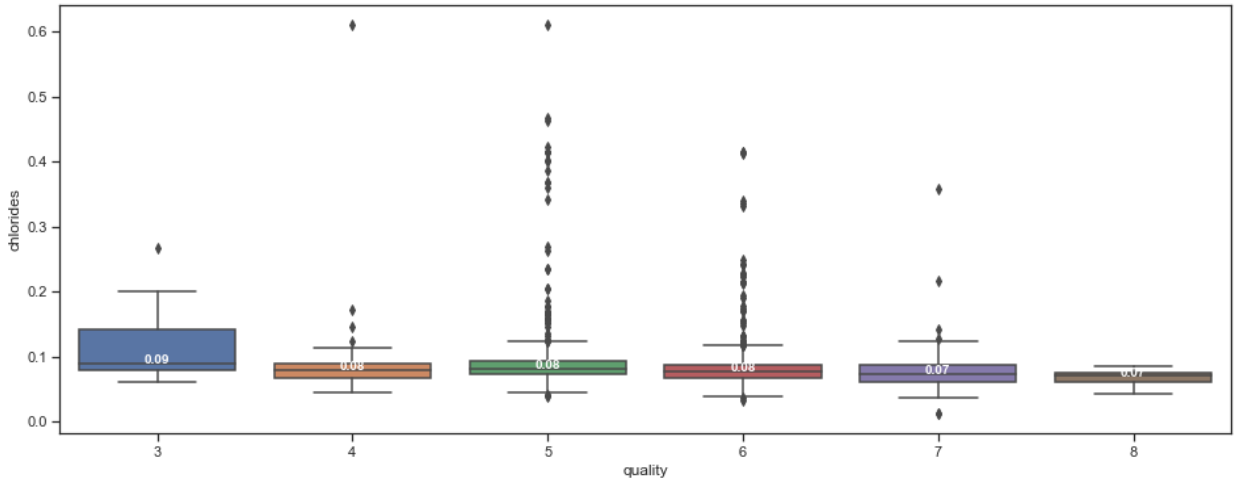
Volatile acidity: aquest camp si que incideix de manera molt directe en la qualitat dels vins, a valors menors major es la qualitat del vi. Això ja va quedar patent en l'apartat anterior, quan vàrem veure que existia una correlació negativa entre els camps, ja que uns nivells apreciables d'aquests àcids impliquen una deterioració del vi.



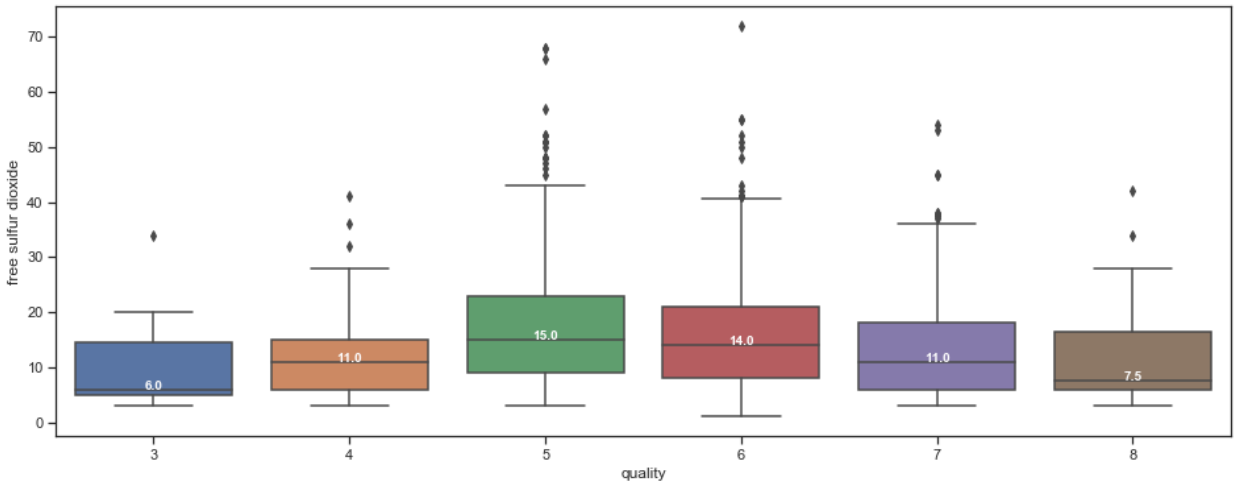
Citric acid: sembla que aquest àcid té un impacte en la qualitat, a volums més elevats d'aquest àcid millor es la qualitat. Per lo tant, raïms amb una concentració més elevada d'aquest àcid implicarà una millor qualitat del vi



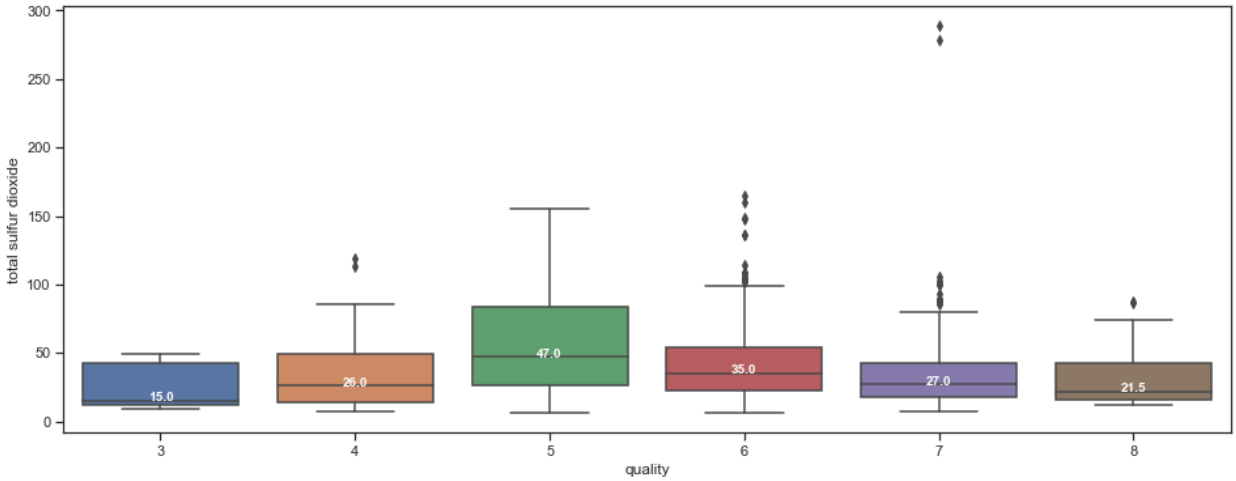
Residual sugar: no sembla que tingui incidència directa, les mitjanes no varien.



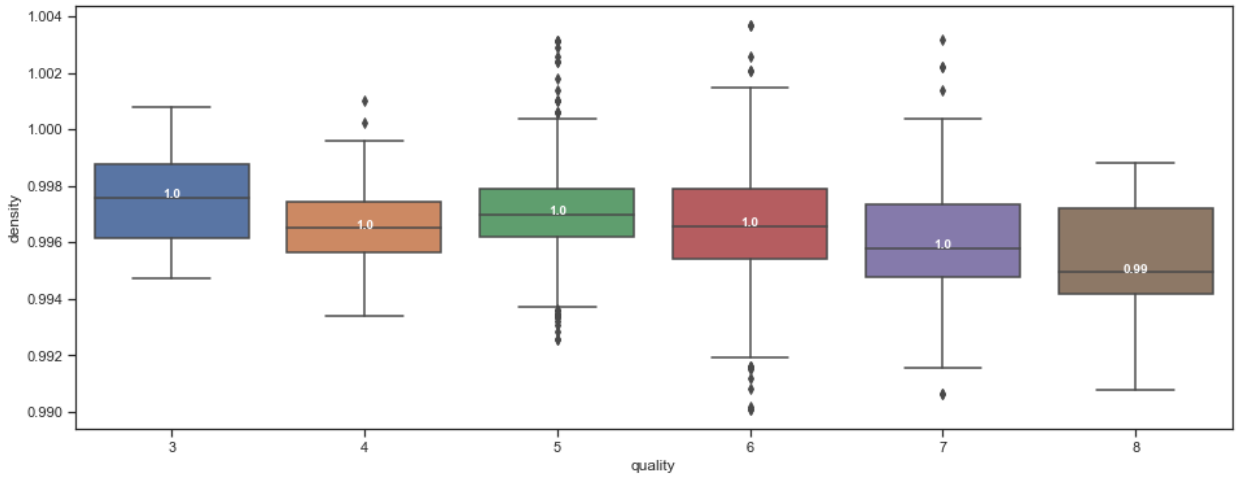
Chlorides: tampoc sembla que tingui cap incidència directa en la qualitat dels vins, la mitjana també sembla similar.



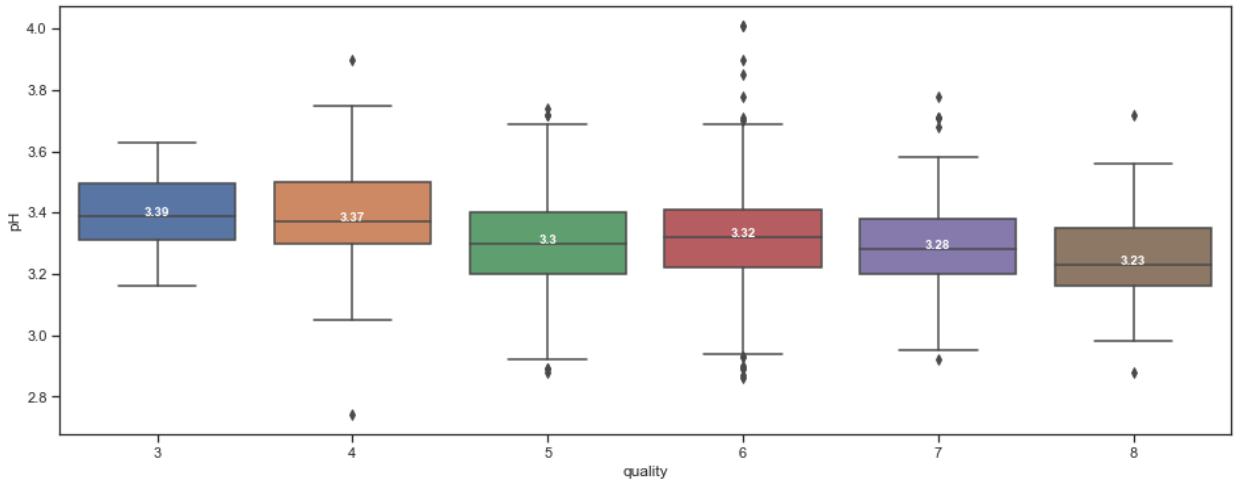
Free sulfur dioxide: el que observem es que tant les pitjors qualitats com les millors tenen un mateix (quasi) nivell de SO2 lliure, per lo tant no ajuda a l'hora d'identificar la qualitat dels vins.



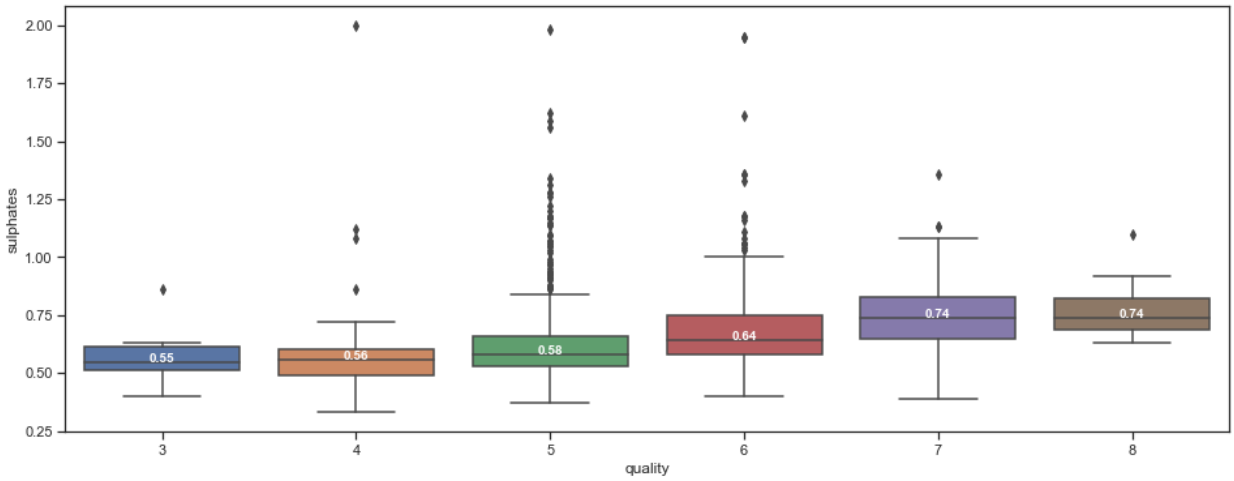
Total sulfur dioxide: es lo mateix que amb el nivell lliure de SO2.



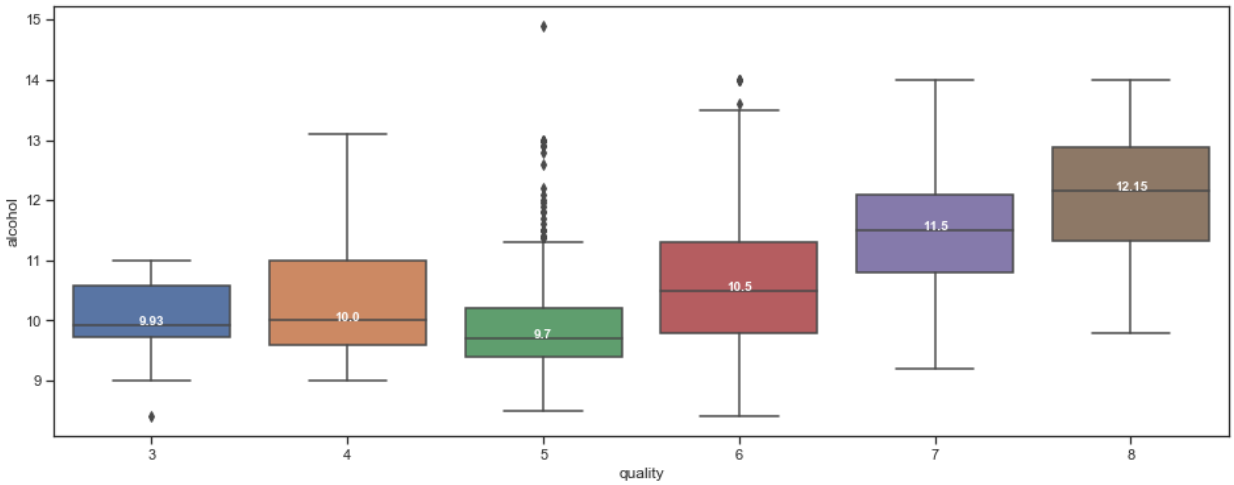
Density: les mitjanes son pràcticament iguals, no té cap incidència.



pH: les mitjanes son pràcticament iguals, no té cap incidència.



Sulphates: sembla que existeix una correlació positiva entre els sulfats i la qualitat del vi. A més sulfats millor es la qualitat del vi.



Alcohol: també sembla que pot tenir un impacte en la qualitat, ja que a valors mes grans d'alcohol millor es la qualitat dels vins.

Observacions:

- Visualment, els camps Alcohol, Sulphates, Citric Acid, i Volatile Acidity mostren una variació significativa de les mitjanes. A priori sembla que podrien ser bons classificadors dels vins per determinar la seva qualitat.
- Els gràfics dels camps **Free sulfur dioxide** i **Total sulfur dioxide** no mostren variacions aparents en els vins de baixa i alta qualitat.
- Els demés camps no mostren unes variacions significatives (visualment).

6.3 Calcular l'índex de correlació amb la qualitat dels vins (comprovació de les dades anteriors)

Ara calcularem a l'índex de correlació existent entre cada camp numèric i la qualitat dels vins, per veure si efectivament es compleix el que hem vist als gràfics del apartat anterior.

	quality
fixed acidity	0.124052
volatile acidity	-0.390558
citric acid	0.220846
residual sugar	0.013732
chlorides	-0.128907
free sulfur dioxide	-0.050656
total sulfur dioxide	-0.185100
density	-0.174919
pH	-0.057731
sulphates	0.251397
alcohol	0.476166
quality	1.000000

Efectivament, es confirmen les dades de l'apartat anterior. Els camps **Alcohol**, **Sulphates**, **Citric Acid**, i **Volatile Acid**, son els camps que mostren un major índex de correlació amb la qualitat dels vins.

6.4 Calcular la importància de cada camp en relació a la qualitat

Ja hem identificat que 4 camps tenen una incidència directe en la qualitat dels vins.

Ara anem a calcular un model de regressió sobre cada una d'aquestes variables per veure quant % de variància expliquen del model final, o sigui, quant importants son a l'hora de preveure la qualitat dels vins.

Ordre de càlcul:

- Citric acid
- Alcohol
- Volatile acidity
- Sulphates

Importancia del Acid Citric (R-squared)

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.049
Model:	OLS	**Adj. R-squared:**	0.048
Method:	Least Squares	**F-statistic:**	81.88
Date:	Sun, 06 Jan 2019	**Prob (F-statistic):**	4.08e-19
Time:	22:21:12	**Log-Likelihood:**	-1886.7
No. Observations:	1599	**AIC:**	3777.
Df Residuals:	1597	**BIC:**	3788.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.3646	0.036	149.466	0.000	5.294	5.435

citric acid	0.9657	0.107	9.049	0.000	0.756	1.175

Omnibus:	11.563	**Durbin-Watson:**	1.738
Prob(Omnibus):	0.003	**Jarque-Bera (JB):**	12.476
Skew:	0.159	**Prob(JB):**	0.00195
Kurtosis:	3.294	**Cond. No.**	5.86

Importancia del Alcohol (R-squared):

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.227
Model:	OLS	**Adj. R-squared:**	0.226
Method:	Least Squares	**F-statistic:**	468.3
Date:	Sun, 06 Jan 2019	**Prob (F-statistic):**	2.83e-91
Time:	22:21:18	**Log-Likelihood:**	-1721.1
No. Observations:	1599	**AIC:**	3446.
Df Residuals:	1597	**BIC:**	3457.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	1.8750	0.175	10.732	0.000	1.532	2.218
alcohol	0.3608	0.017	21.639	0.000	0.328	0.394

Omnibus:	38.501	**Durbin-Watson:**	1.748
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	71.758
Skew:	-0.154	**Prob(JB):**	2.62e-16
Kurtosis:	3.991	**Cond. No.**	104.

Importancia del Acidesa Volatil (R-squared):

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.153
Model:	OLS	**Adj. R-squared:**	0.152
Method:	Least Squares	**F-statistic:**	287.4
Date:	Sun, 06 Jan 2019	**Prob (F-statistic):**	2.05e-59
Time:	22:21:26	**Log-Likelihood:**	-1794.3
No. Observations:	1599	**AIC:**	3593.
Df Residuals:	1597	**BIC:**	3603.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	6.5657	0.058	113.388	0.000	6.452	6.679
volatile acidity	-1.7614	0.104	-16.954	0.000	-1.965	-1.558

Omnibus:	20.577	**Durbin-Watson:**	1.736

Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21.905
Skew:	0.242	**Prob(JB):**	1.75e-05
Kurtosis:	3.306	**Cond. No.**	7.18

Importancia dels Sulfats (R-squared):

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.063
Model:	OLS	**Adj. R-squared:**	0.063
Method:	Least Squares	**F-statistic:**	107.7
Date:	Sun, 06 Jan 2019	**Prob (F-statistic):**	1.80e-24
Time:	22:21:29	**Log-Likelihood:**	-1874.4
No. Observations:	1599	**AIC:**	3753.
Df Residuals:	1597	**BIC:**	3764.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	4.8477	0.078	61.818	0.000	4.694	5.002
sulphates	1.1977	0.115	10.380	0.000	0.971	1.424

Omnibus:	12.685	**Durbin-Watson:**	1.712
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	17.098

Skew:	0.083	**Prob(JB):**	0.000194
Kurtosis:	3.479	**Cond. No.**	8.51

Podem observar que tots els camps son estadísticament significatius a l'hora de preveure la qualitat del vi, P>|t| = 0.0000.

Relació de camps fortament correlacionats ordenats pel nivell d'importància (més importants primer):

- **Alcohol** amb un R-squared del **22%**. Aquest camp explica el 22% de la variació del dataset.
- **Volatile Acidity** amb un R-squared del **15%**.
- **Sulphates** amb un R-squared del **6%**.
- **Citric Acid** amb un R-squared del **5%**.

6.5 Resum

Els camps amb un impacte directe a la qualitat dels vins son:

- Positivament correlacionats: citric acid, alcohol
- Negativament correlacionats: volatile acidity, sulphates

De entre tots els camps anteriors, l'**alcohol** es el que té un major impacte en la qualitat dels vins, seguit del **volatile acidity**.

7 Anàlisi de la variància

Hem vist que l'**alcohol** es el camp més fortament correlacionat amb la qualitat dels vins.

El que farem ara es veure si veritablement las mitges i desviacions de les mostres d'alcohol per als 3 tipus de qualitats (baixa, bona, molt bona) son les mateixes (no hi ha diferencia entre qualitats) o realment pertanyen a poblacions diferents (hi ha diferencia entre qualitats).

La hipòtesis nul·la es que el grau alcohol no identifica de manera unívoca la qualitat dels vins ja que les seves mitjas i variacions son iguals:

- Ho: μ(low) = μ(medium) = μ(high)

Pel contrari, la hipòtesis alternativa serà que el grau d'alcohol si que identifica de manera clara la qualitat del vi, i per lo tant les seves mitges son diferents:

- H1: μ(low) ≠ μ(medium) ≠ μ(high)

Per fer aquest estudi farem el test ANOVA fent servir **statsmodel**, que ens ajudarà a obtenir més d'informació i introduïm el model com una fórmula de regressió.

Resultats del model:

OLS Regression Results

Dep. Variable:	alcohol	**R-squared:**	0.259
Model:	OLS	**Adj. R-squared:**	0.259
Method:	Least Squares	**F-statistic:**	279.6
Date:	Sun, 06 Jan 2019	**Prob (F-statistic):**	8.01e-105
Time:	20:13:54	**Log-Likelihood:**	-2129.9
No. Observations:	1599	**AIC:**	4266.
Df Residuals:	1596	**BIC:**	4282.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept *(quality_cat[T.low])*	9.9265	0.034	295.061	0.000	9.860	9.992
quality_cat[T.medium]	0.7030	0.050	14.199	0.000	0.606	0.800
quality_cat[T.high]	1.5916	0.071	22.481	0.000	1.453	1.730

Omnibus:	162.565	**Durbin-Watson:**	1.528
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	230.769

Skew:	0.777	**Prob(JB):**	7.75e-51
Kurtosis:	4.024	**Cond. No.**	3.69

El grup de qualitat *low* està capturat al coeficient *Intercept*.

En general, el model és significatiu, F-statistic (2, 1596) = 279.6, p = 0,0000. Això ens diu que hi ha una diferència significativa en les mitges dels grups. Els coeficients (coef a la taula) són la diferència de la mitja entre el grup de control (low) i els grups respectius (medium i high).

La intercepció és la mitja del grup de baixa qualitat (low), el coeficient del grup high = 1.5916, i el coeficient de qualitat mitja = 0.7030. Si observem els valors p ara (P>|t| a la taula), podem observar que les diferències entre els grups son significatives, p = 0.0000.

No hi ha cap comparació entre el grup de qualitat mitja i el grup de qualitat alta.

Procedent del marc **ANOVA**, la informació que realment ens interessa d'aquesta taula és la F-statistic i el corresponent p-value. I ens indica si hem explicat una quantitat significativa de la variància global. Per fer proves entre grups, hem de fer algunes proves post-hoc on podrem comparar tots els grups un contra l'altre.

Amb aquest mètode encara ens falta informació útil, necessitarem una taula **ANOVA**.

	sum_sq	df	F	PR(>F)
quality_cat	470.843637	2.0	279.579864	8.009998e-105
Residual	1343.920900	1596.0	NaN	NaN

Anem a inspeccionar la taula **ANOVA**.

La fila de **qualitat_cat** és l'efecte entre qualitats que és l'efecte experimental general.

La suma de quadrats del model (SSM té un valor de 470.84 a la taula) és la quantitat de variància que explica el model. El model actual explica una quantitat significativa de variància, F (2, 1596) = 279.58, p < 0.0000.

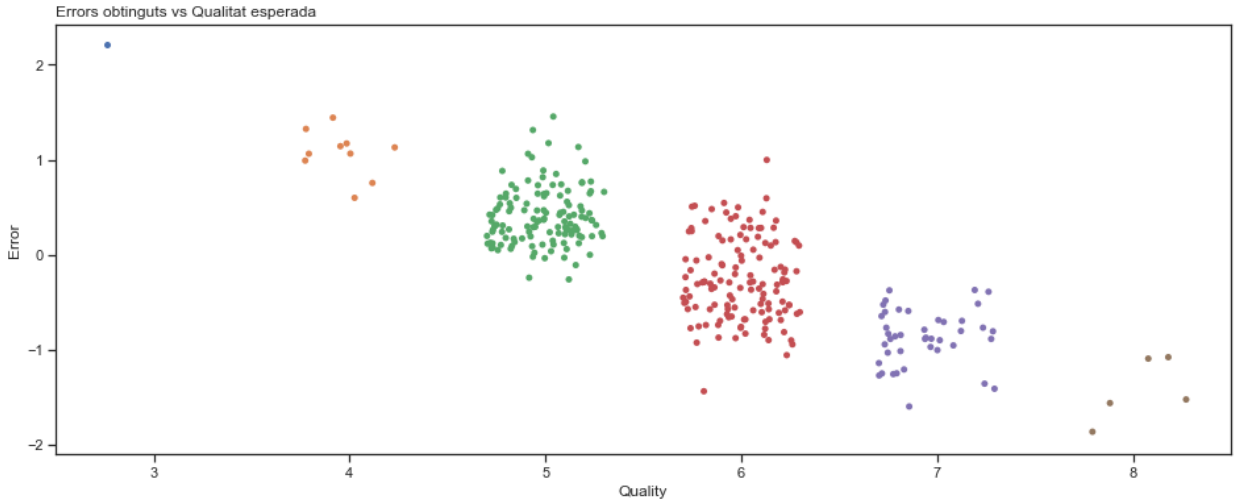
La fila **Residual** és la variació no sistemàtica de les dades (SSR, també anomenada variància inexplicable, el valor 1343.92 a la taula). En aquest cas, la variació no sistemàtica representa les diferències individuals naturals en el alcohol i les diferents qualitats del vi.

8 Regressió

Hem identificat els camps que estan correlacionats amb la qualitat dels vins.

Hem confirmat que el alcohol es la variable més correlacionada amb la qualitat, i on les variàncies de cada grup de qualitat (low, medium, high) representen poblacions independents.

Ara el que farem es crear un model de regressió amb tots aquests camps i en el gràfic següent mostrem els errors que s'han obtingut en base a la qualitat esperada i les prediccions generades pel model (hem fet servir un sample del 20% del data set original com a data set de test).



Podem veure una densitat més gran en les seccions de qualitat mitja que en les seccions de qualitat alta i baixa. Això es degut a que el data set contenia molts mes vins de qualitat mitja.

Podem observar que el model comet molts menys errors amb els vins de qualitat mitja que no pas amb els de baixa i alta qualitat, on els errors son molt més grans.

El model amb els camps seleccionats només es capaç d'explicar el 40% (R-squared) del canvi en la qualitat dels vins. No sembla un model adequat per predir la qualitat dels vins de baixa i alta qualitat.

9 Conclusions

El data set està molt balancejat, oferint pocs casos de vins de qualitat baixa i alta, i molts vins de qualitat mitja. Aquesta realitat pot afectar als models de classificació o regressió de la qualitat dels vins. La solució pot ser ampliar el número de vins en aquestes categories, o aplicar tècniques de boosting per a les classes de qualitat baixa i alta, donant més pes a aquest casos en front dels de qualitat mitja.

Podríem tenir problemes en la lectura/captura dels valors de la quantitat d'àcid cítric, ja que ens hem trobat un número molt elevat de zeros. Seria necessari comprovar que això es així i que les captures

son correctes. En aquest anàlisis hem pres la decisió de assumir que les dades no eren correctes i hem procedit a la seva imputació automàtica en base als casos on aquests valors estaven informats (K-NN amb k=5).

Com hem pogut comprovar, només dos camps presentaven un comportament de distribució normal estàndard, els camps density i pH. La resta de distribucions presenten una distorsió positiva (positive skewness) i un número important de valors atípics. Els camps **volatile acidity** i **citric acid** son els que mostren unes distribucions no normals (binomial la primera i desconeguda la segona)

Fent un anàlisis de mitges i variàncies hem trobat que els camps Alcohol, Volatile Acidity, Sulphates, i Citric Acid son els camps que més correlacionats estan amb la qualitat del vi. En el cas de l'alcohol i el àcid cítric la correlació es positiva, a valors més grans millor es la qualitat dels vins. En el cas del àcids volàtils i els sulfats, la correlació es negativa, a valors més grans pitjors qualitats.

Entrenant models de regressió lineal (fent servir OLS) hem pogut verificar que tots els camps anteriors - els correlacionats - son estadísticament significatius, i que el camp **alcohol** es el que més important, sent capaç d'explicar un 22% de la variació total del model. Els altres camps junts només son capaços de cobrir un 26% de la variació total del model.

Tots aquest camps junts, en el moment d'entrenat un model de regressió lineal, no son suficients per oferir unes prediccions de qualitat, cobrint només al voltant d'un 40% de la variació total del model.

En base a totes aquestes dades podem concloure que **l'actual data set NO ES SUFICIENT per tal de classificar o calcular (regressió) la qualitat de nous vins**.

Per millorar podríem explorar les següents accions:

- Oferir més casos de vins de baixa i alta qualitat
- Assegurar que les dades son totalment correctes (veure el problema del camp de l'àcid cítric)
- Analitzar en profunditat els valors extrems que hem trobat en alguns dels camps per veure si son producte d'errors o realment son vàlids. En cas de ser vàlids veure si es possible eliminar-los o crear grups de vins (data sets) independents.
- Afegir més característiques que puguin donar més informació al model.