Data Exploration

Struktur

Exploration:

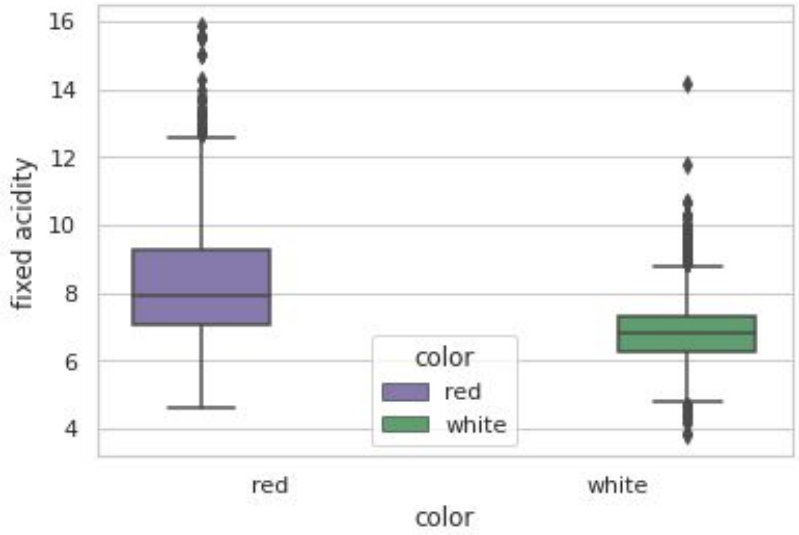
- Allgemeine Auffälligkeiten
- Mean Values und Min/Max
- Outlier und Noise
- Missing Values
- Verteilung
- Korrelation

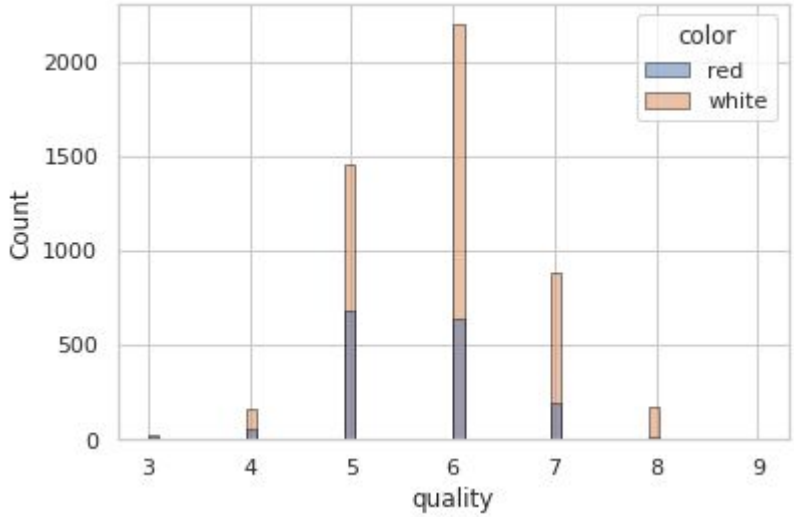
Fazit:

- Präliminäre/Spekulative Feature Selection

Allgemeine Auffälligkeiten

- Größere Unterschiede zwischen Rot- und Weißwein Datasets
  - Weißwein hat eine wesentlich höhere Anzahl an Instanzen
  - Große Unterschiede bei der Verteilung der Werte pro Feature
  - Verteilung der von Rot- und Weißwein ähnlich (Qualität)





Mean Values und Min/Max (1)

	ID	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	flavanoids	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	magnesium	alcohol	lightness	quality
count	6495.000000	6478.000000	6495.000000	6495.000000	6495.000000	6495.000000	6495.000000	6495.000000	6495.000000	6495.000000	6468.000000	6495.000000	6495.000000	6495.000000	6495.000000	6495.000000
mean	3248.755196	7.213453	0.339804	0.318670	5.443033	0.056123	0.416905	30.527329	115.747575	0.995130	3.252775	0.538531	0.492057	10.491844	0.105349	5.863279
std	1875.409590	1.295632	0.164960	0.145323	4.757924	0.035140	0.064610	17.751729	56.516565	0.003687	1.717880	0.270958	0.173181	1.192693	0.012140	2.188689
min	1.000000	3.800000	0.080000	0.000000	0.600000	0.009000	0.380000	1.000000	6.000000	0.987110	2.720000	0.000000	0.000000	8.000000	0.070000	3.000000
25%	1625.500000	6.400000	0.230000	0.250000	1.800000	0.038000	0.380000	17.000000	77.250000	0.992000	3.110000	0.330000	0.390000	9.500000	0.100000	5.000000
50%	3249.000000	7.000000	0.290000	0.310000	3.000000	0.047000	0.380000	29.000000	118.000000	0.994800	3.210000	0.570000	0.480000	10.300000	0.108000	6.000000
75%	4872.500000	7.700000	0.400000	0.390000	8.100000	0.066000	0.380000	41.000000	156.000000	0.998880	3.320000	0.740000	0.580000	11.300000	0.111000	6.000000
max	6496.000000	15.900000	1.580000	1.660000	65.800000	0.610000	0.530000	289.000000	440.000000	1.038980	99.990000	2.000000	1.080000	14.900000	0.140000	99.000000

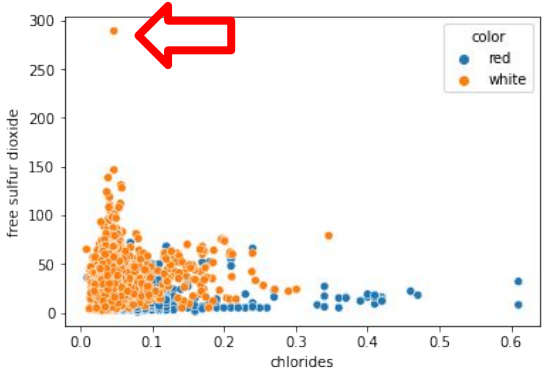
Auffälligkeiten:

- Werte nicht (einheitlich) normiert
- flavanoids konstanter Wert für je Rot- und Weißwein
- density Werte sehr nah aneinander

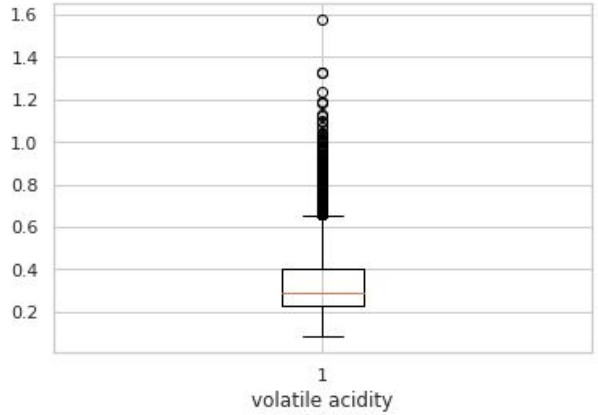
Mean Values und Min/Max (2)

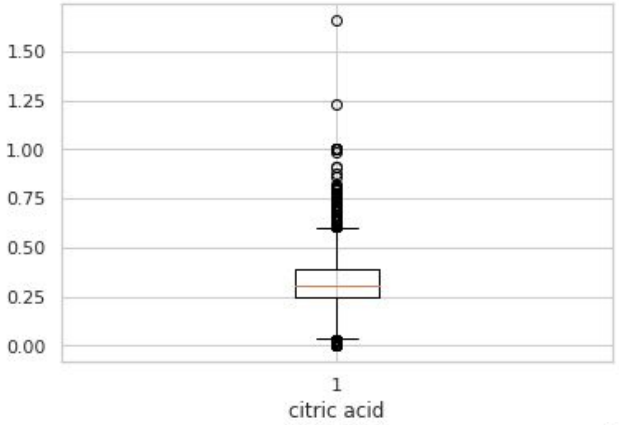
	ID	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	flavanoids	free sulfur dioxide	total sulfur dioxide	density	pH
count	6495.000000	6478.000000	6495.000000	6495.000000	6495.000000	6495.000000	6495.000000	6495.000000	6495.000000	6495.000000	6468.000000
mean	3248.755196	7.213453	0.339804	0.318670	5.443033	0.056123	0.416905	30.527329	115.747575	0.995130	3.252775
std	1875.409590	1.295632	0.164960	0.145323	4.757924	0.035140	0.064610	17.751729	56.516565	0.003687	1.717880
min	1.000000	3.800000	0.080000	0.000000	0.600000	0.009000	0.380000	1.000000	6.000000	0.987110	2.720000
25%	1625.500000	6.400000	0.230000	0.250000	1.800000	0.038000	0.380000	17.000000	77.250000	0.992000	3.110000
50%	3249.000000	7.000000	0.290000	0.310000	3.000000	0.047000	0.380000	29.000000	118.000000	0.994800	3.210000
75%	4872.500000	7.700000	0.400000	0.390000	8.100000	0.066000	0.380000	41.000000	156.000000	0.998880	3.320000
max	6496.000000	15.900000	1.580000	1.660000	65.800000	0.610000	0.530000	289.000000	440.000000	1.038980	99.990000

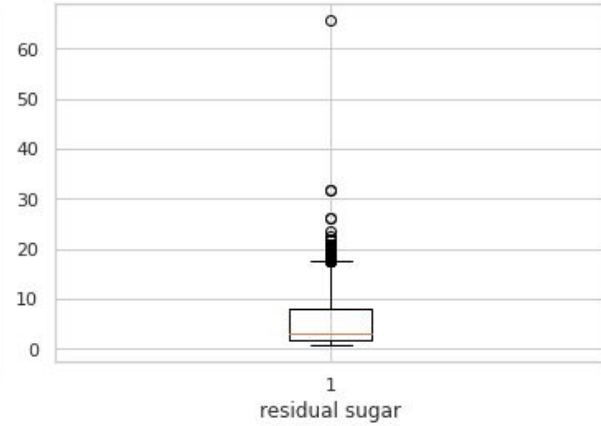
sulphates	magnesium	alcohol	lightness	quality
6495.000000	6495.000000	6495.000000	6495.000000	6495.000000
0.538531	0.492057	10.491844	0.105349	5.863279
0.270958	0.173181	1.192693	0.012140	2.188689
0.000000	0.000000	8.000000	0.070000	3.000000
0.330000	0.390000	9.500000	0.100000	5.000000
0.570000	0.480000	10.300000	0.108000	6.000000
0.740000	0.580000	11.300000	0.111000	6.000000
2.000000	1.080000	14.900000	0.140000	99.000000

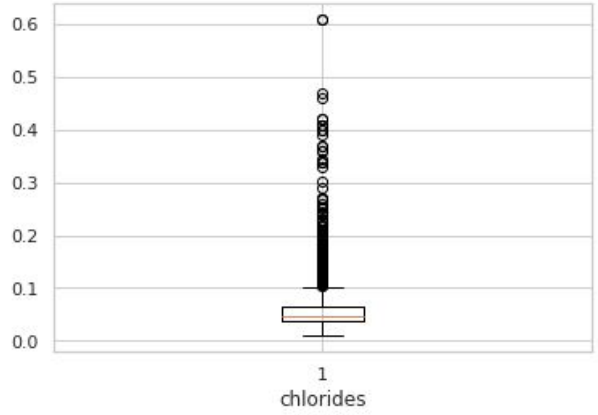


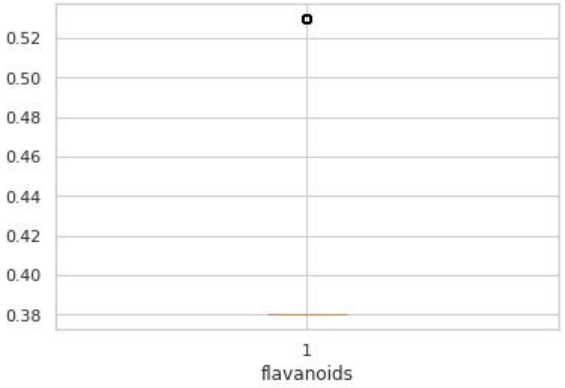
Outlier und Noise(1)

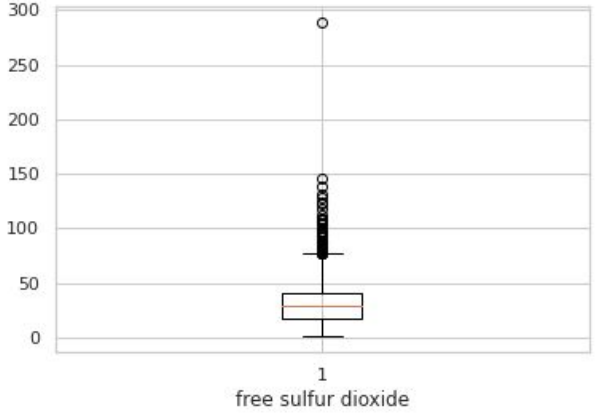




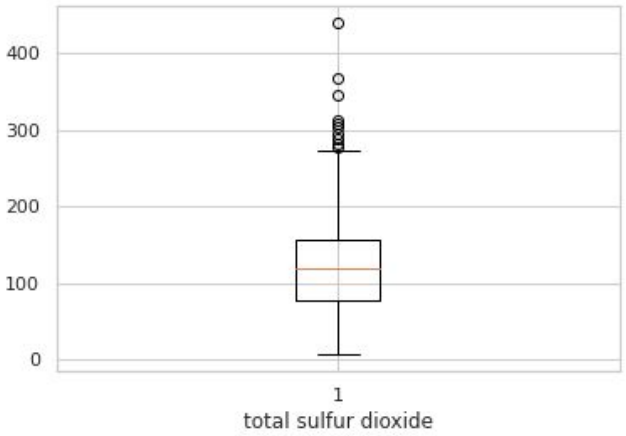


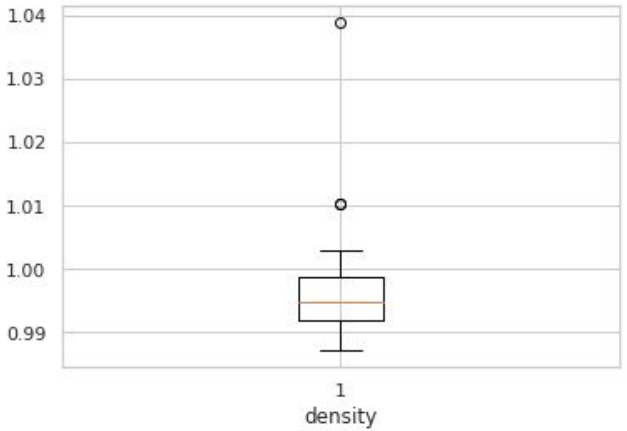


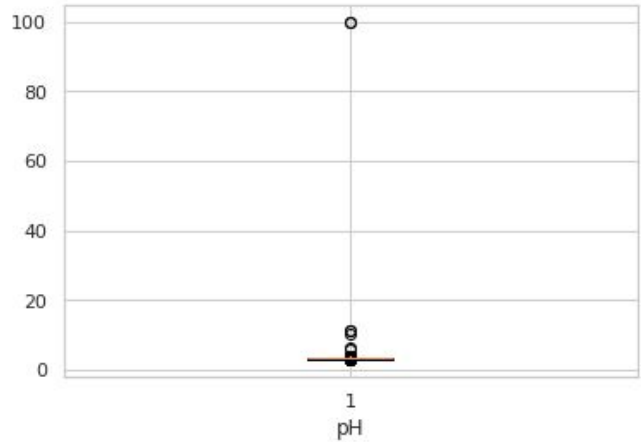


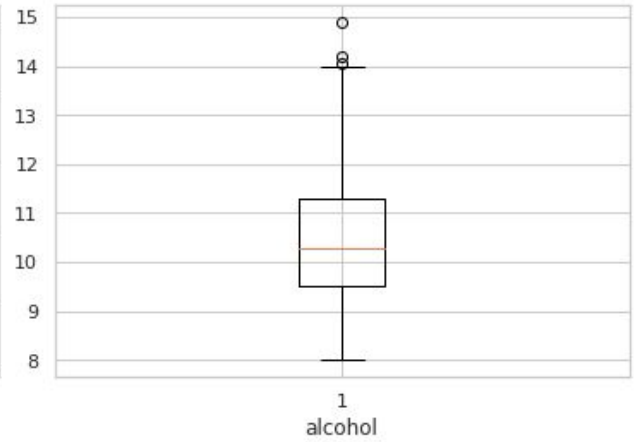


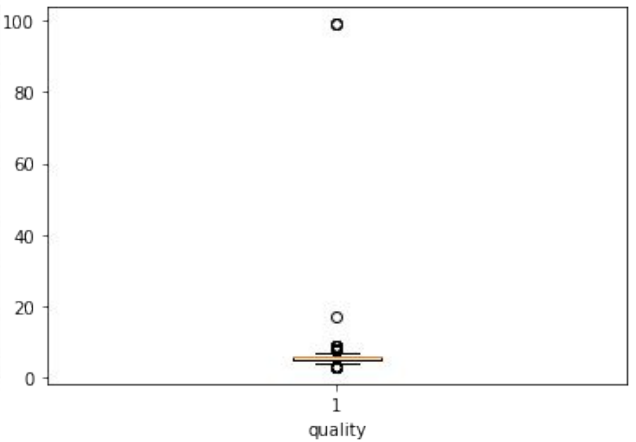
Outlier und Noise (2)











Missing Values

- Fehlende Werte bei ‘fixed acidity’ (nur Rotwein) und pH-Wert
- Teilweise Missing Values bei aufeinanderfolgenden IDs

```
Missing values:
 ID                    0
fixed acidity         17
volatile acidity       0
citric acid            0
residual sugar         0
chlorides              0
flavanoids             0
free sulfur dioxide    0
total sulfur dioxide   0
density                0
pH                    27
sulphates              0
magnesium              0
alcohol                0
lightness              0
quality                0
color                  0
```

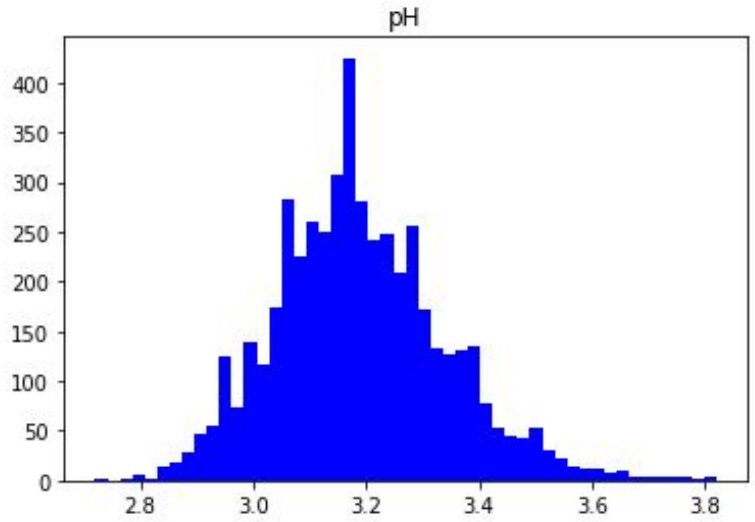
409	410	NaN	0.46	0.49	4.5	0.070	0.53	26.0	49.0	1.0000	3.05
410	411	NaN	0.43	0.34	2.5	0.080	0.53	26.0	86.0	1.0000	3.38
411	412	NaN	0.45	0.35	2.4	0.080	0.53	23.0	78.0	1.0000	3.38
412	413	NaN	0.74	0.16	1.9	0.100	0.53	15.0	77.0	1.0000	3.27
606	607	9.4	0.41	0.48	4.6	0.070	0.53	10.0	20.0	1.0000	NaN
607	608	8.8	0.48	0.41	3.3	0.090	0.53	26.0	52.0	1.0000	NaN
608	609	10.1	0.65	0.37	5.1	0.110	0.53	11.0	65.0	1.0000	NaN

Verteilung der Daten

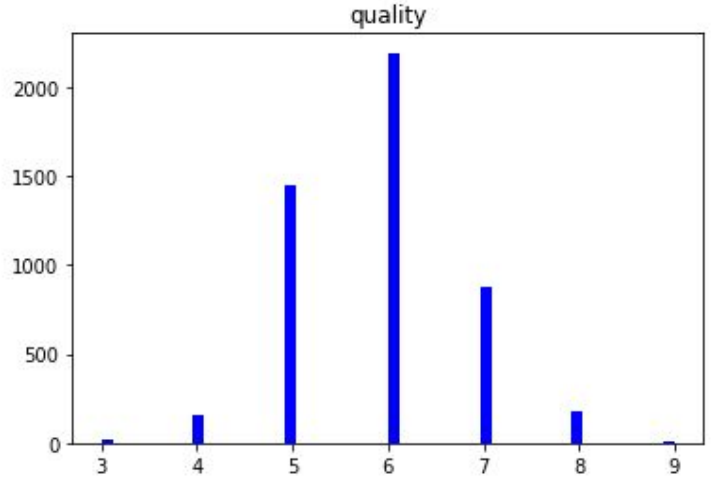
Verteilungsarten im Datensatz:

- Normalverteilung
- Uniformverteilung
- Log-Normalverteilung/Skewed distribution
- Sonstiges

Klassische Gauß-Verteilung

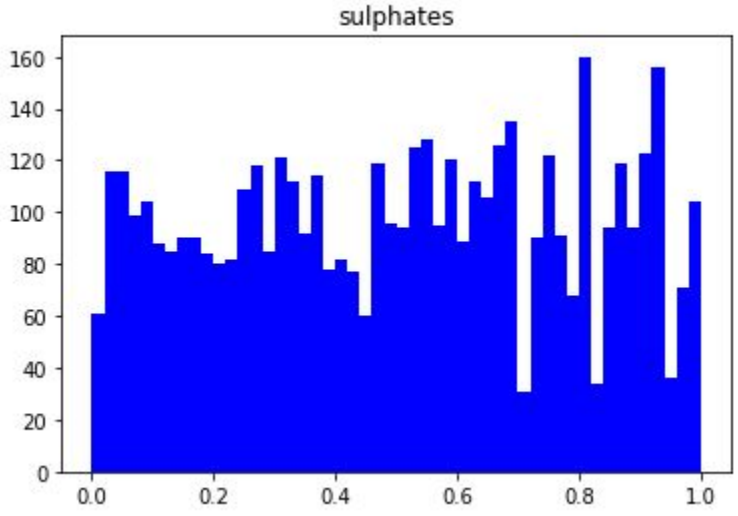


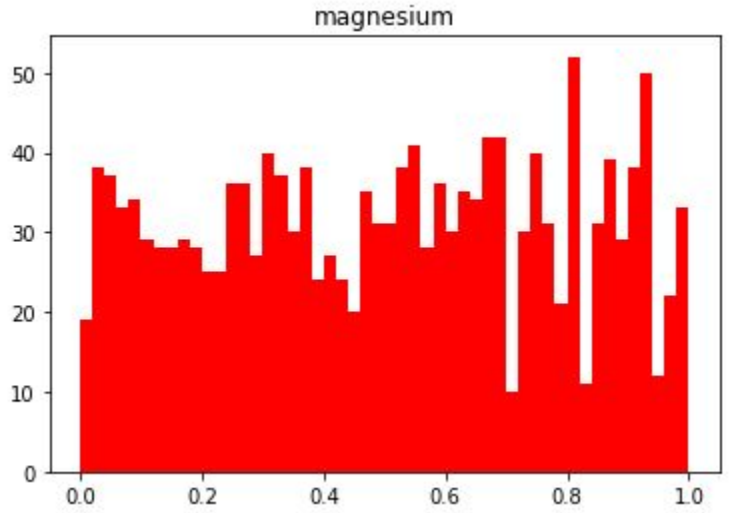
pH 0.458768



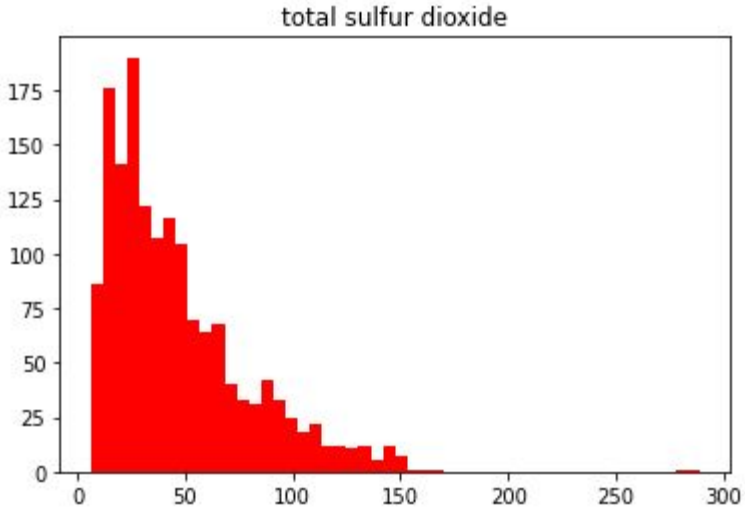
quality 0.155848

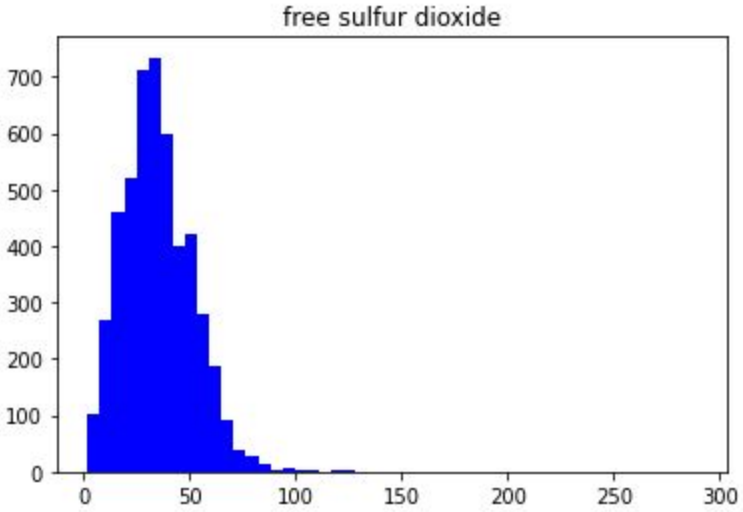
Uniformverteilung



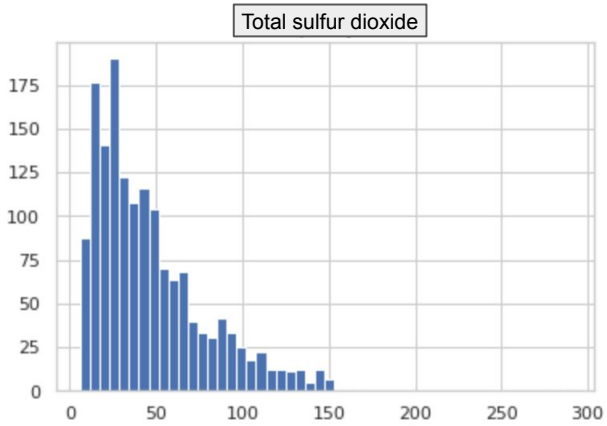


Log-Normalverteilung/Skewed distribution

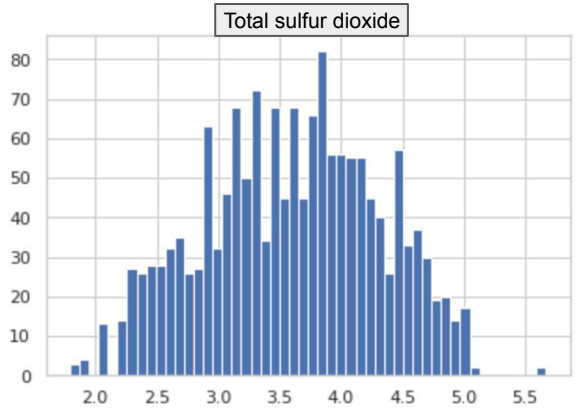




Log-Normalverteilung/Skewed distribution



x



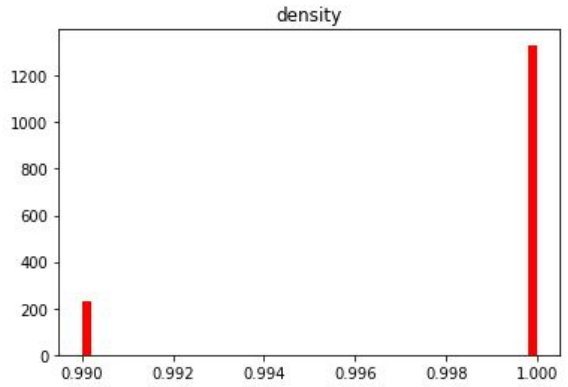
ln(x)

sonstige

Bimodal?

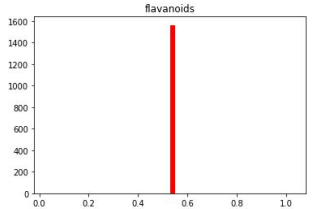
Es wurde offensichtlich auf Werte gerundet, die für diese Skala keinen Sinn macht

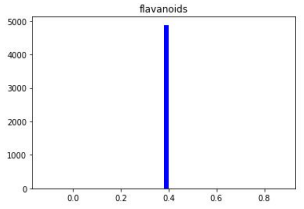
→ Sollte vermutlich entfernt werden



Konstanter Wert

→ Definiert das Label/die Weinsorte

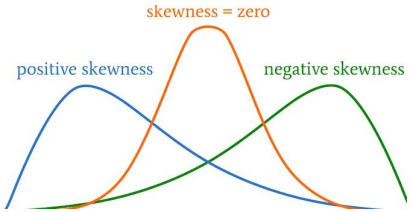




flavanoids 0.000000

flavanoids 0.000000

Überblick



Rotwein Skew:

fixed acidity	0.980387
volatile acidity	0.692556
citric acid	0.324888
residual sugar	4.595426
chlorides	5.718728
flavanoids	0.000000
free sulfur dioxide	1.259800
total sulfur dioxide	1.525189
density	-1.980676
pH	15.867907
sulphates	2.463123
magnesium	-0.024204
alcohol	0.866436
lightness	-0.410118
quality	0.212963

dtype: float64

Weißwein Skew:

fixed acidity	0.647363
volatile acidity	1.574424
citric acid	1.284795
residual sugar	1.075824
chlorides	5.025558
flavanoids	0.000000
free sulfur dioxide	1.406552
total sulfur dioxide	0.387699
density	0.976748
pH	0.458768
sulphates	-0.025326
magnesium	0.980032
alcohol	0.486121
lightness	-0.117372
quality	0.155848

dtype: float64

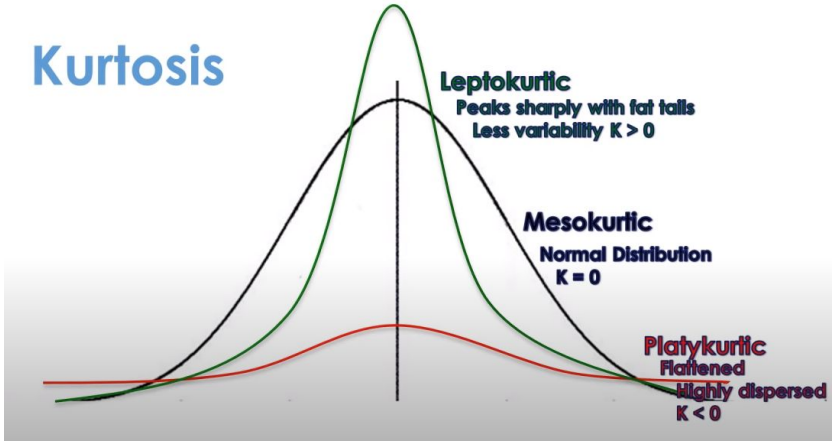
Fairly symmetrical/ symmetrical

Moderately skewed

Highly skewed

Nicht normalverteilt

Wölbung



Weißwein Kurtosis:

fixed acidity	2.167770
volatile acidity	5.075559
citric acid	6.170920
residual sugar	3.473933
chlorides	37.558571
flavanoids	0.000000
free sulfur dioxide	11.481252
total sulfur dioxide	0.573850
density	9.798207
pH	0.533035
sulphates	-1.177286
magnesium	1.611151
alcohol	-0.699119
lightness	-0.614685
quality	0.218117

dtype: float64

Rotwein Kurtosis:

fixed acidity	1.110358
volatile acidity	1.235069
citric acid	-0.785483
residual sugar	29.121961
chlorides	43.332935
flavanoids	0.000000
free sulfur dioxide	2.039520
total sulfur dioxide	3.847713
density	1.925538
pH	311.278175
sulphates	12.011295
magnesium	-1.186654
alcohol	0.228303
lightness	-0.449995
quality	0.318302

dtype: float64

Korrelation

Zusammenhänge zwischen Attribute

→ wichtig für die Feature Selection

Korrelationsmatrix

- Pearson
- Spearman

Zusammenfassung der Korrelationen

Die Korrelationsmatrix

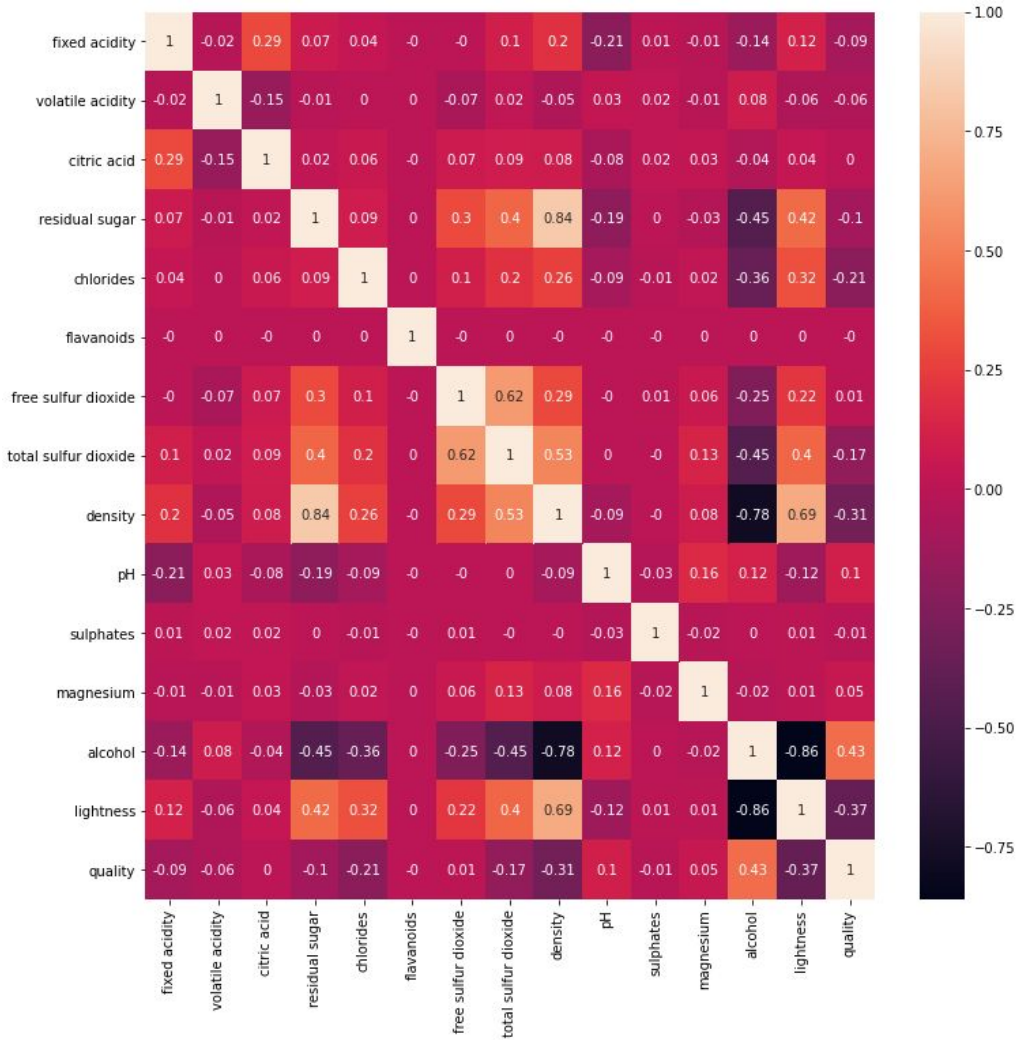
Veranschaulichung der Zusammenhänge unter den Attributen

[-1;0[- neg. Korrelation

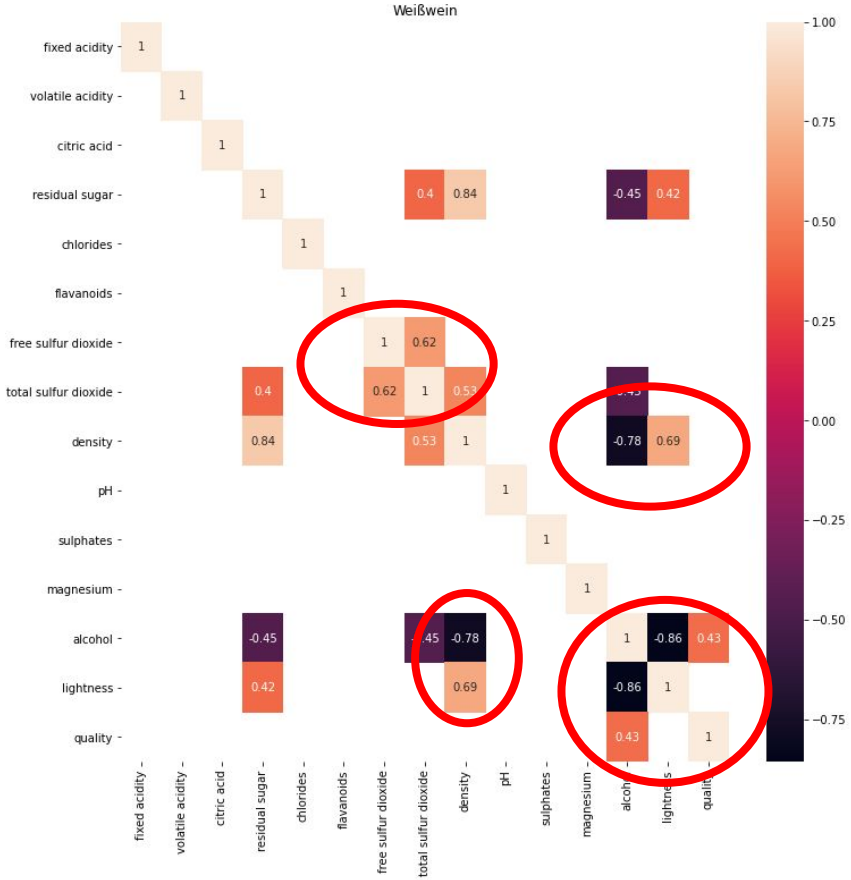
]0;1] - pos. Korrelation

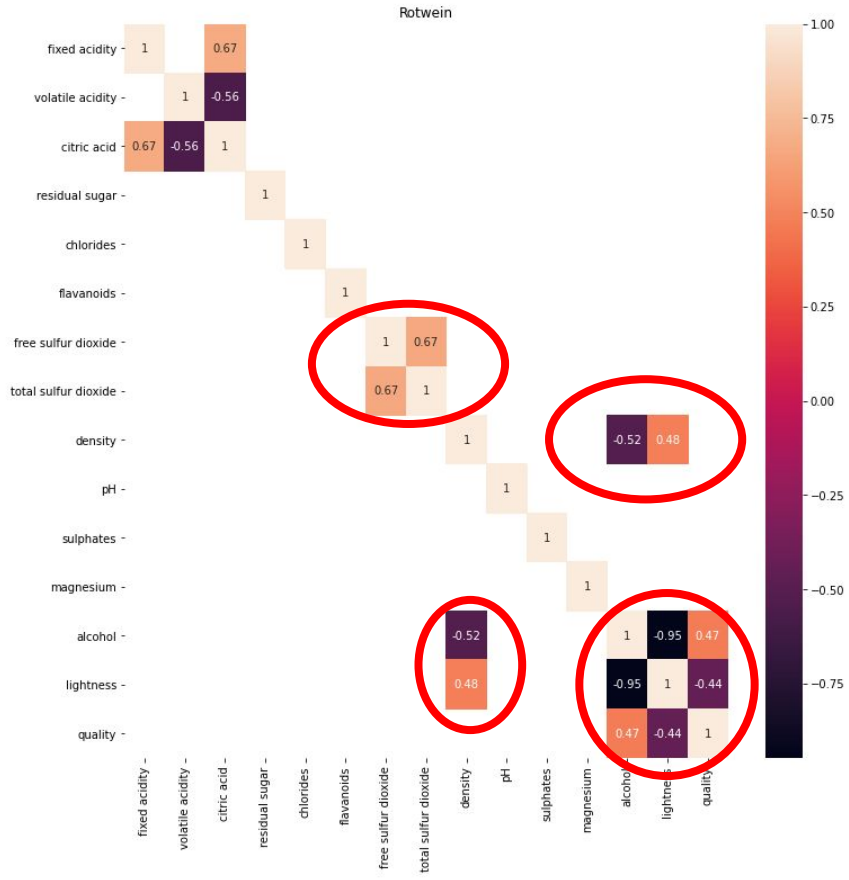
Einstufung:

Bis ± 0.4 → keine/triviale Korrelation
Ab ± 0.4 → schwache Korrelation
Ab ± 0.6 → moderate/gute Korrelation
Ab ± 0.8 → starke Korrelation



Gefilterte Matrix (Pearson Korrelation)



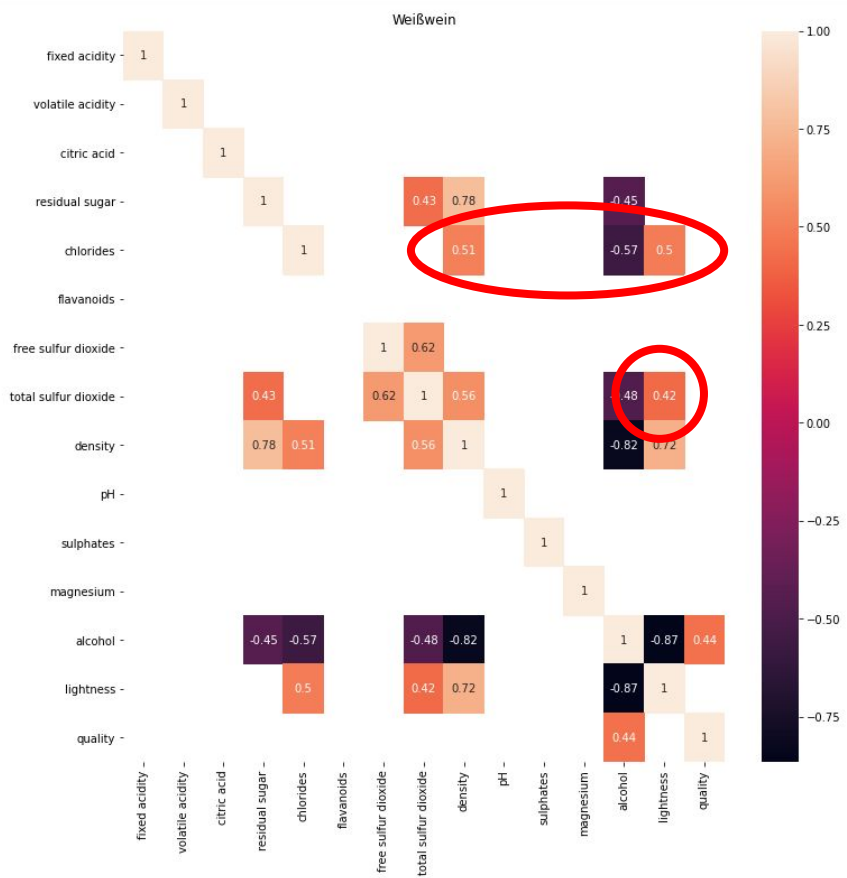


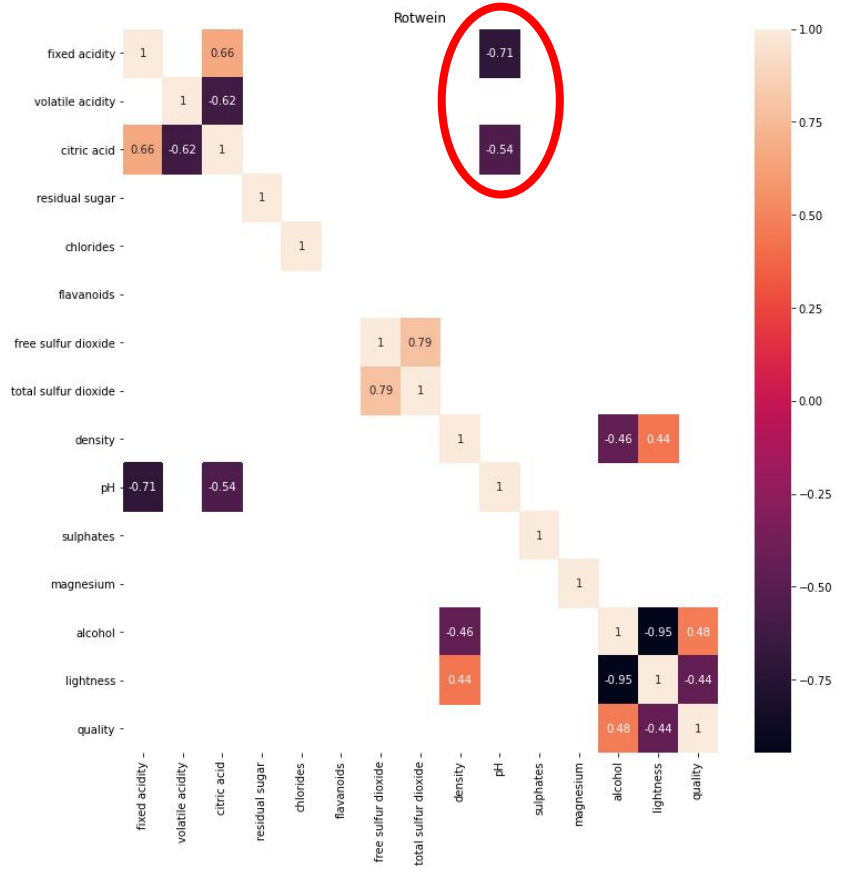
Spearman Korrelationskoeffizient

- Pearson Korrelationskoeffizient zeigt lediglich lineare Zusammenhänge auf!
  - → nicht-lineare Korrelationen (z.B. quadratische) werden nicht aufgedeckt
- Anderes Maß muss verwendet werden:

- → Spearman’sche Rangkorrelationseffizient
  - Bestimmung der Korrelation nicht direkt den kontinuierlichen Werten der Attribute...
  - ...sondern mit den Rangfolge der Werte der Features
    - Ordinalwerte zur Bestimmung der Korrelation

Gefilterte Matrix (Spearman Korrelation)





Übersicht - Korrelationen

Pearson Spearman

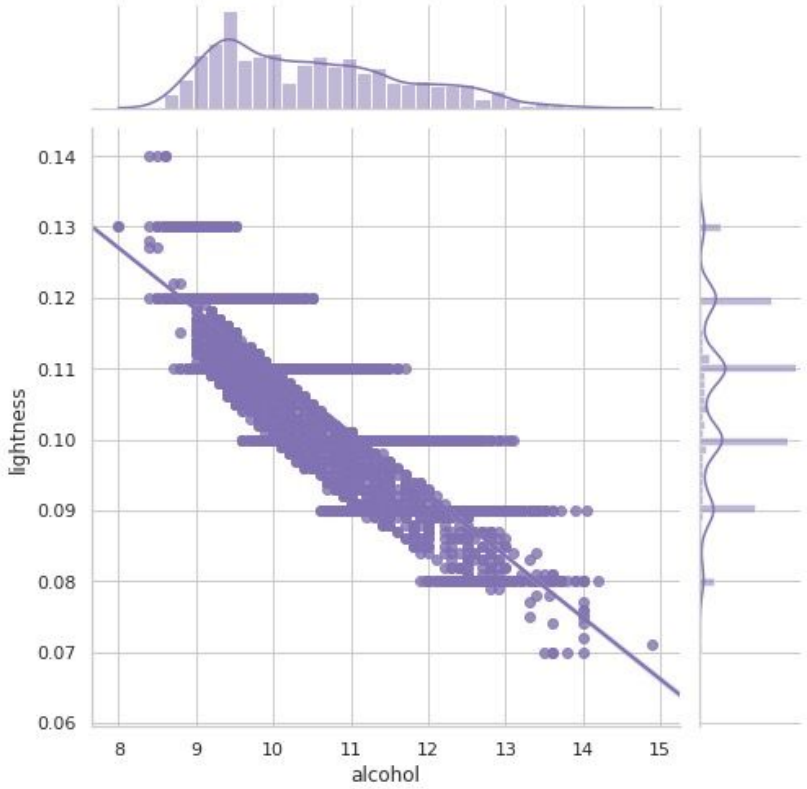
Korrelation	Weißwein	Rotwein
stark	residual sugar & density (0.84) alcohol & lightness (-0.86)	alcohol & lightness (-0.95)
moderat	free sulfur dioxide & total sulfur dioxide (0.62) density & lightness (0.69) density & alcohol (-0.78)	fixed acidity & citric acid (0.67) free sulfur dioxide & total sulfur dioxide (0.67) fixed acidity & pH (-0.71) citric acid & pH (-0.57)
schwach	total sulfur dioxide & residual sugar (0.4) total sulfur dioxide & density (0.53) total sulfur dioxide & lightness (0.4) residual sugar & lightness (0.42) residual sugar & alcohol (-0.45) total sulfur dioxide & alcohol (-0.45) chlorides & density (0.51) chlorides & lightness (0.5) chlorides & alcohol (-0.57)	density & lightness (0.48) alcohol & quality (0.47) volatile acidity & citric acid (-0.56) density & alcohol (-0.52) lightness & quality (-0.44)

Übersicht - Korrelationen

Pearson Spearman

Korrelation	Weißwein	Rotwein
stark	residual sugar & density (0.84) alcohol & lightness (-0.86)	alcohol & lightness (-0.95)
moderat	free sulfur dioxide & total sulfur dioxide (0.62) density & lightness (0.69) density & alcohol (-0.78)	fixed acidity & citric acid (0.67) free sulfur dioxide & total sulfur dioxide (0.67) fixed acidity & pH (-0.71) citric acid & pH (-0.57)
schwach	total sulfur dioxide & residual sugar (0.4) total sulfur dioxide & density (0.53) total sulfur dioxide & lightness (0.4) residual sugar & lightness (0.42) residual sugar & alcohol (-0.45) total sulfur dioxide & alcohol (-0.45) chlorides & density (0.51) chlorides & lightness (0.5) chlorides & alcohol (-0.57)	density & lightness (0.48) alcohol & quality (0.47) volatile acidity & citric acid (-0.56) density & alcohol (-0.52) lightness & quality (-0.44)

Korrelation: Alcohol & Lightness veranschaulicht



Fazit

- Starke negative Korrelation zwischen “alcohol” & “lightness” (& “density”)
- Nur bei Rotwein: starke Korrelation zwischen “residual sugar” & “density”

Intuitive Entscheidung, welche Attribute reduziert werden:

- Bei Rotwein: “lightness”, “lightness”, “total sulfur dioxide”
- Bei Weißwein: “lightness”, “citric acid”, “fixed acidity”, “total sulfur dioxide”

→ Zur Bestätigung: Information Gain Ratio anwenden

Key Takeaway: Unterschiede bei der Korrelation zwischen Rot- und Weißwein
→ separates Modell je Weinsorte