Clearing of the Vinho Verde Data Set

1. Duplicates
2. Legal Limits
3. Outliers
4. Resulting Correlations

N
0 20 40mi
0 20 40km

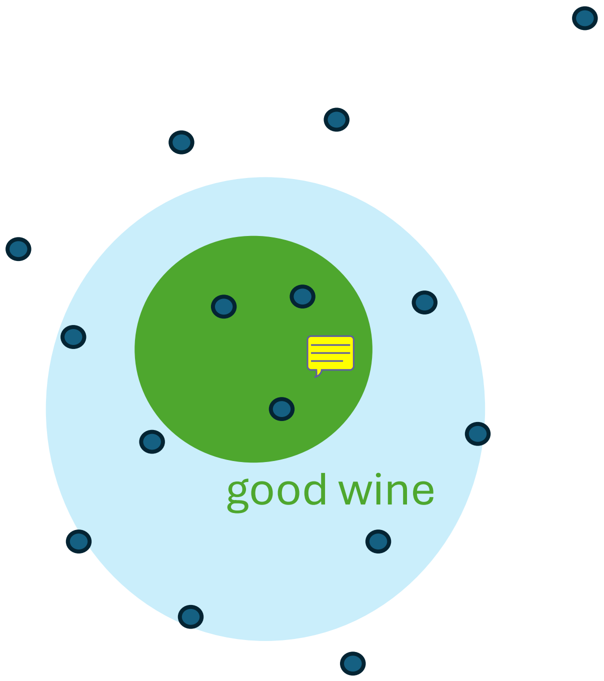
Data Cleaning

- The data set is very noisy
  - Naturally: wine and taste is very complex and interrelated
  - Shallow

- Duplicates
  - Wines with exactly the same values
- „Illegal“ Wines in EU
  - Total Sulfur Dioxide < 210 mg/l
  - Volatile Acidity < 1.1 g/l
- Outliers Input Quantities
  - values outside 1.5*IQR

Remove the above wines from the data set because

- Noisy and complex data set
- At least work on the main features



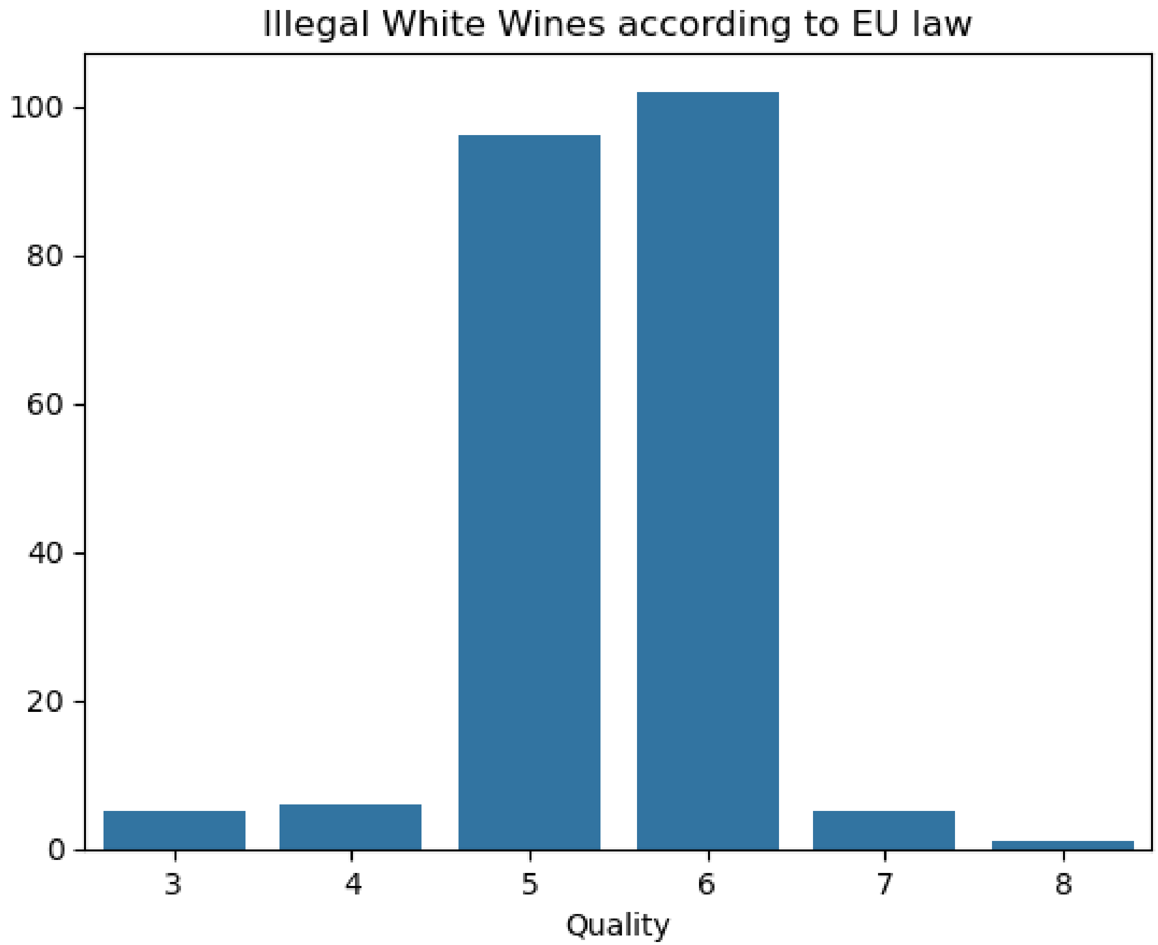
Duplicates

- 937 of 4898 (19%) entries were dropped.
- Do not make sense.
- Effect: Duplicates shift correlation coefficients.

Illegal Stuff

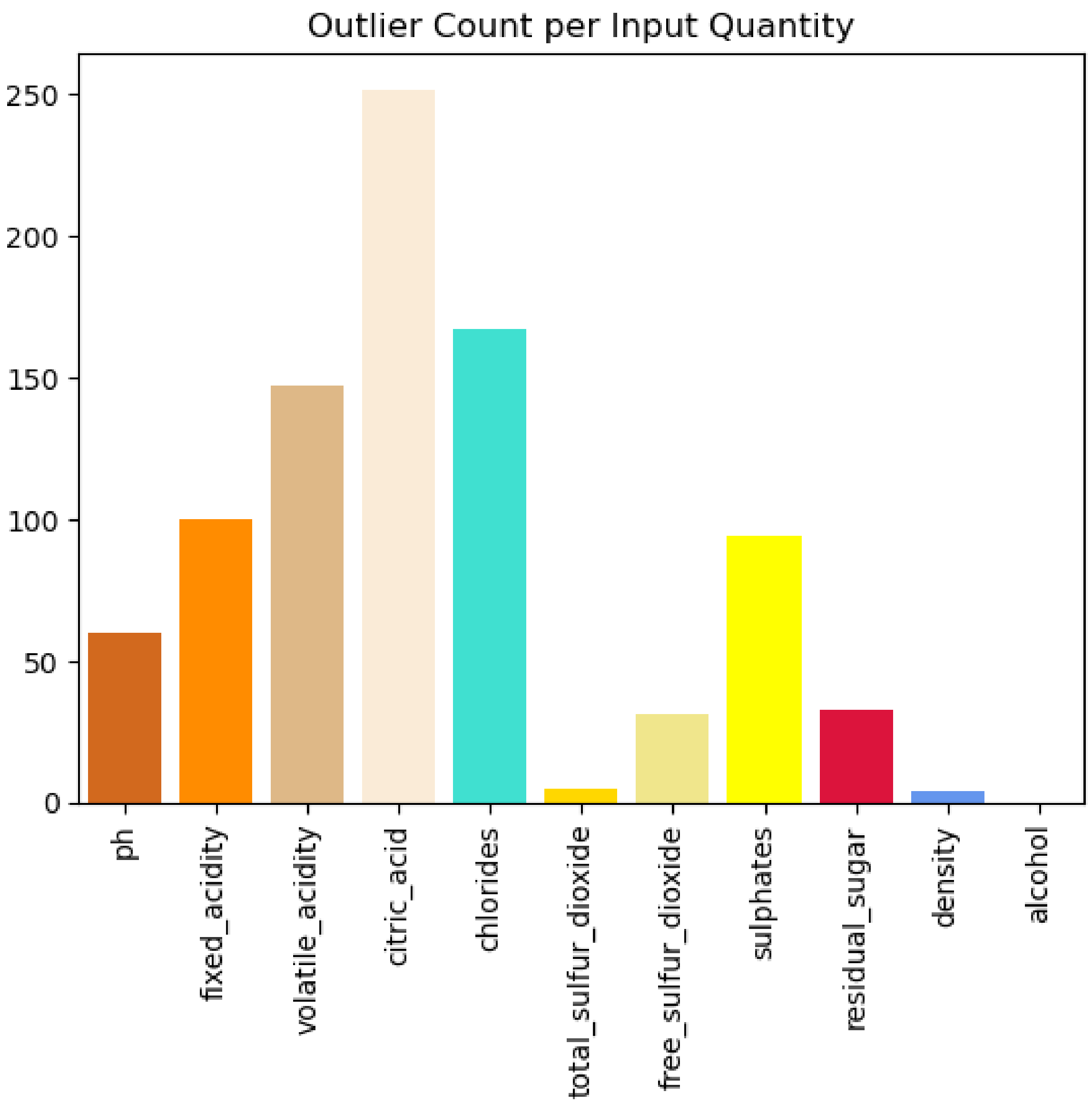
Quantity	Limit	Count
Total Sulfur Dioxide	< 210mg/l	215
Volatile Acidity	< 1.1g/l	0

- Remove if one of the legal limits is surpassed.
- Irrelevant, because we cannot sell those wines!

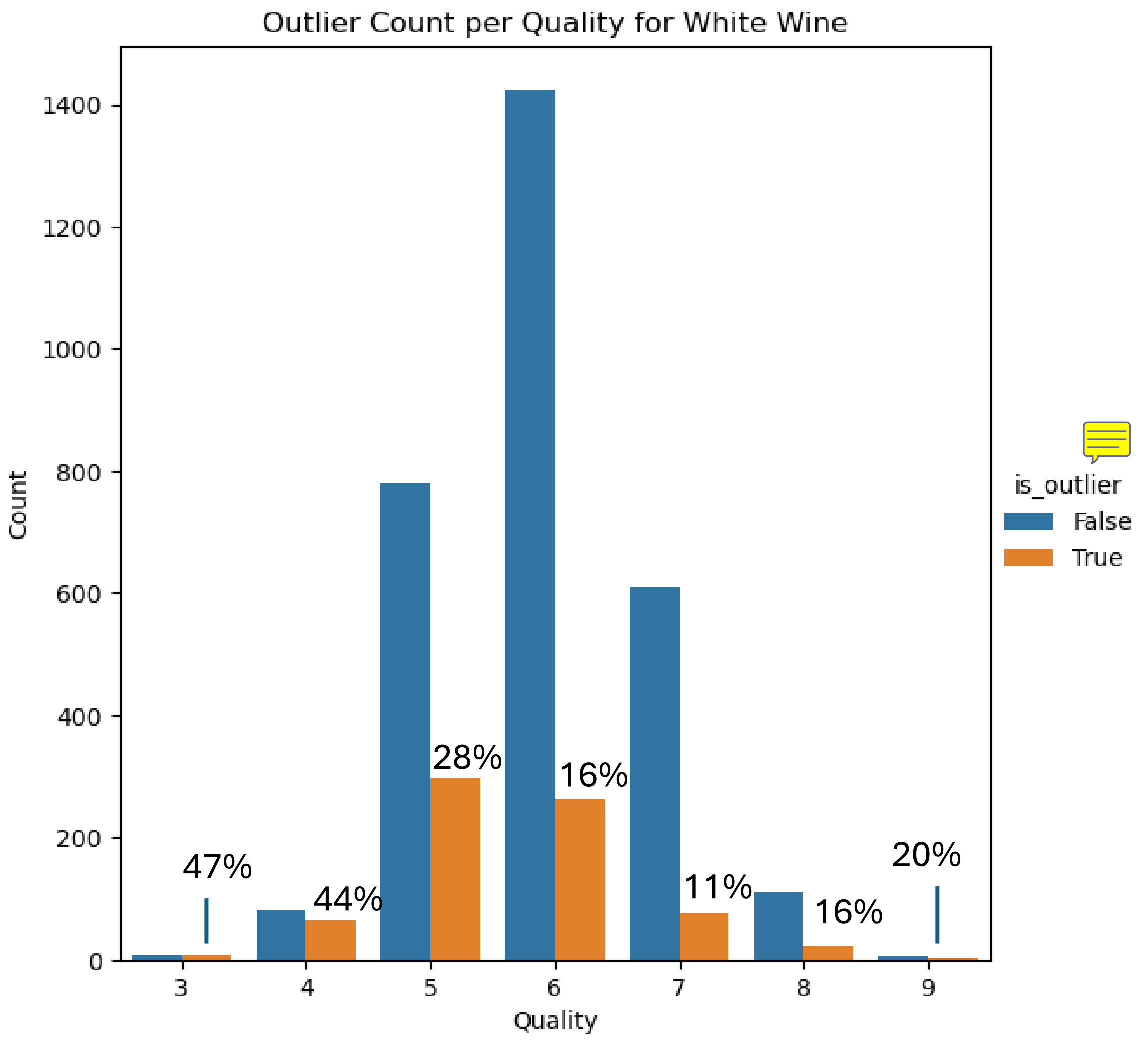


Outlier Wines

- 893 outliers in total
- **Acidity** seems to be a wi(l)dely varying quantity in wines.
- **Chlorides** (Terroir) varies strongly as well.
- **Sulfur Dioxide, Sulphates:** The use of preservatives varies a lot.

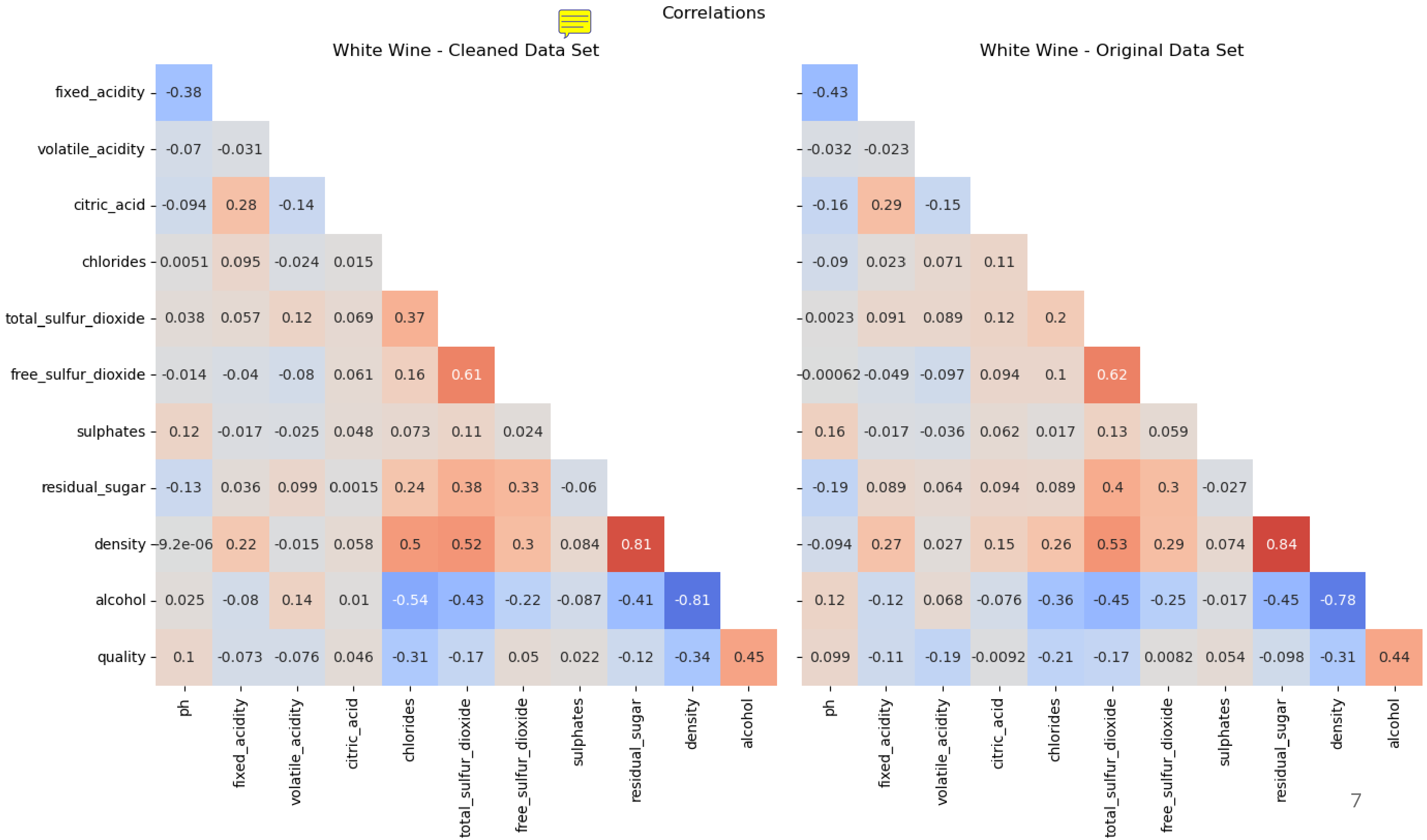


Outlier Wines



- Outlier Wine is a wine with at least one outlier in the input values
- Wine with
  - 1 outlier: 729
  - 2 outliers: 145
  - 3 outliers: 13
- First approach: Drop all outlier wines!

Result



Correlations

White Wine - Cleaned Data Set

	ph	fixed_acidity	volatile_acidity	citric_acid	chlorides	total_sulfur_dioxide	free_sulfur_dioxide	sulphates	residual_sugar	density	alcohol
fixed_acidity	-0.38										
volatile_acidity	-0.07	-0.031									
citric_acid	-0.094	0.28	-0.14								
chlorides	0.0051	0.095	-0.024	0.015							
total_sulfur_dioxide	0.038	0.057	0.12	0.069	0.37						
free_sulfur_dioxide	-0.014	-0.04	-0.08	0.061	0.16	0.61					
sulphates	0.12	-0.017	-0.025	0.048	0.073	0.11	0.024				
residual_sugar	-0.13	0.036	0.099	0.0015	0.24	0.38	0.33	-0.06			
density	-9.2e-06	0.22	-0.015	0.058	0.5	0.52	0.3	0.084	0.81		
alcohol	0.025	-0.08	0.14	0.01	-0.54	-0.43	-0.22	-0.087	-0.41	-0.81	
quality	0.1	-0.073	-0.076	0.046	-0.31	-0.17	0.05	0.022	-0.12	-0.34	0.45

White Wine - Original Data Set

	ph	fixed_acidity	volatile_acidity	citric_acid	chlorides	total_sulfur_dioxide	free_sulfur_dioxide	sulphates	residual_sugar	density	alcohol
	-0.43										
	-0.032	-0.023									
	-0.16	0.29	-0.15								
	-0.09	0.023	0.071	0.11							
	0.0023	0.091	0.089	0.12	0.2						
	-0.00062	-0.049	-0.097	0.094	0.1	0.62					
	0.16	-0.017	-0.036	0.062	0.017	0.13	0.059				
	-0.19	0.089	0.064	0.094	0.089	0.4	0.3	-0.027			
	-0.094	0.27	0.027	0.15	0.26	0.53	0.29	0.074	0.84		
	0.12	-0.12	0.068	-0.076	-0.36	-0.45	-0.25	-0.017	-0.45	-0.78	
	0.099	-0.11	-0.19	-0.0092	-0.21	-0.17	0.0082	0.054	-0.098	-0.31	0.44

Conclusions

- Data set with reduced noise mainly in Acidity
  - Acidity of wine is a very complex quantity and is not fully available in the data set.
  - It‘s good to get rid of it in order to find the main features of a good wine.

- Main features
  1. Alcohol
  2. Density
  3. Chlorides
  4. pH
  5. Residual Sugar

- Improvements
  - keep 5% of outliers per input quantity
  - keep the outliers in certain quantities, e.g. chloride
  - Select outliers more carefully (e.g.: take relation of sugar and density into account)
  - Use a different method to find outliers