WINE QUALITY ANALYSIS AND QUALITY SCORE PREDICTION

IDENTIFYING KEY WINE QUALITY DRIVERS AND PREDICTING WINE QUALITY

THOMAS LALOUX – PERSONAL PROJECT 11/2025

EXECUTIVE SUMMARY

- This analysis focuses on the **Portuguese 'Verdo Vinho' red wine**, with a dataset containing sensory/quality scores and physicochemical properties related to the latter
- **A detailed exploratory data analysis** shows that high quality scores are usually characterized by
  - significantly higher alcohol and citric acid concentrations, and a lower volatile acidity
  - lower density, slightly lower pH on average, and higher sulphate concentration also seem to contribute, but to a much lower extent
- **A random forest regression model** has been successfully fit to the data (and assessed on an independent test set), to predict the 'Verdo Vinho' wine quality score, with high accuracy measures:
  - a root mean squared error (RMSE) of 1.37%
  - a mean absolute error (MAE) of 8.67%
  - while also confirming the observations made in the exploratory data analysis, with exceptions
- **A SHAP analysis** has allowed to consolidate our conclusions
  - it shows how each explanatory variable contributes to the predictive model and a high/low wine quality score
  - it confirms the alcohol %, sulphate concentration and volatile acidity are key drivers in the regression model
- Finally, this analysis concludes with **perspectives and concrete applications**

OUTLINE

1. Context, dataset and scope
2. Exploratory data analysis
3. Regression model to predict wine quality
4. Conclusion and perspectives

1. CONTEXT, DATASET AND SCOPE

Context

- This analysis focused on a dataset containing sensory/quality scores and physicochemical properties for the Portuguese red ‘Verdo Vinho’ wine
- The dataset as such has first been referenced by Paulo Cortez et al. in *“Modeling wine preferences by data mining from physicochemical properties”* (2009) and is also available in the UCI machine learning repository (*https://archive.ics.uci.edu/ml/datasets/wine+quality*)

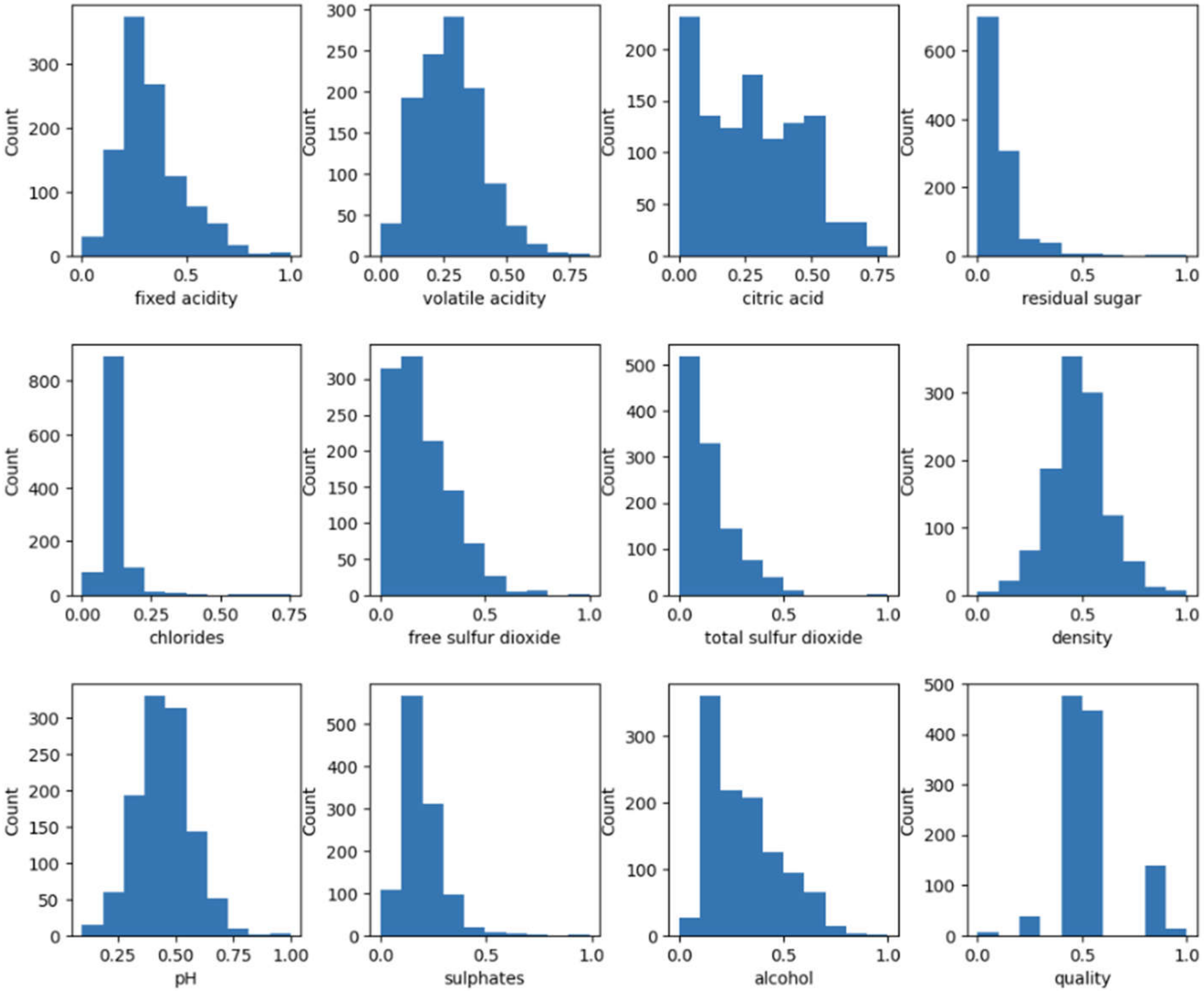
Dataset

- Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available; hence, there is no data about grape types, wine brand, or wine selling price.
- The output variable is: Quality, which is a score between 0 and 10 then scaled to 0-1
- The explanatory variables are: Fixed acidity, Volatile acidity, Critic acid, Residual sugar, Chlorides, Free sulphur dioxide, Total sulphur dioxide, Density, pH, Sulphates, Alcohol

Scope

- The goal of the project is to understand what makes a good wine, at least in the case of the ‘Verdo Vinho’ wine
- As a second objective, the project aims at fitting a regression model to predict wine quality scores with accuracy
- Finally, the project aims at properly interpreting the results, and providing relevant perspectives

2. EXPLORATORY DATA ANALYSIS – HISTOGRAMS



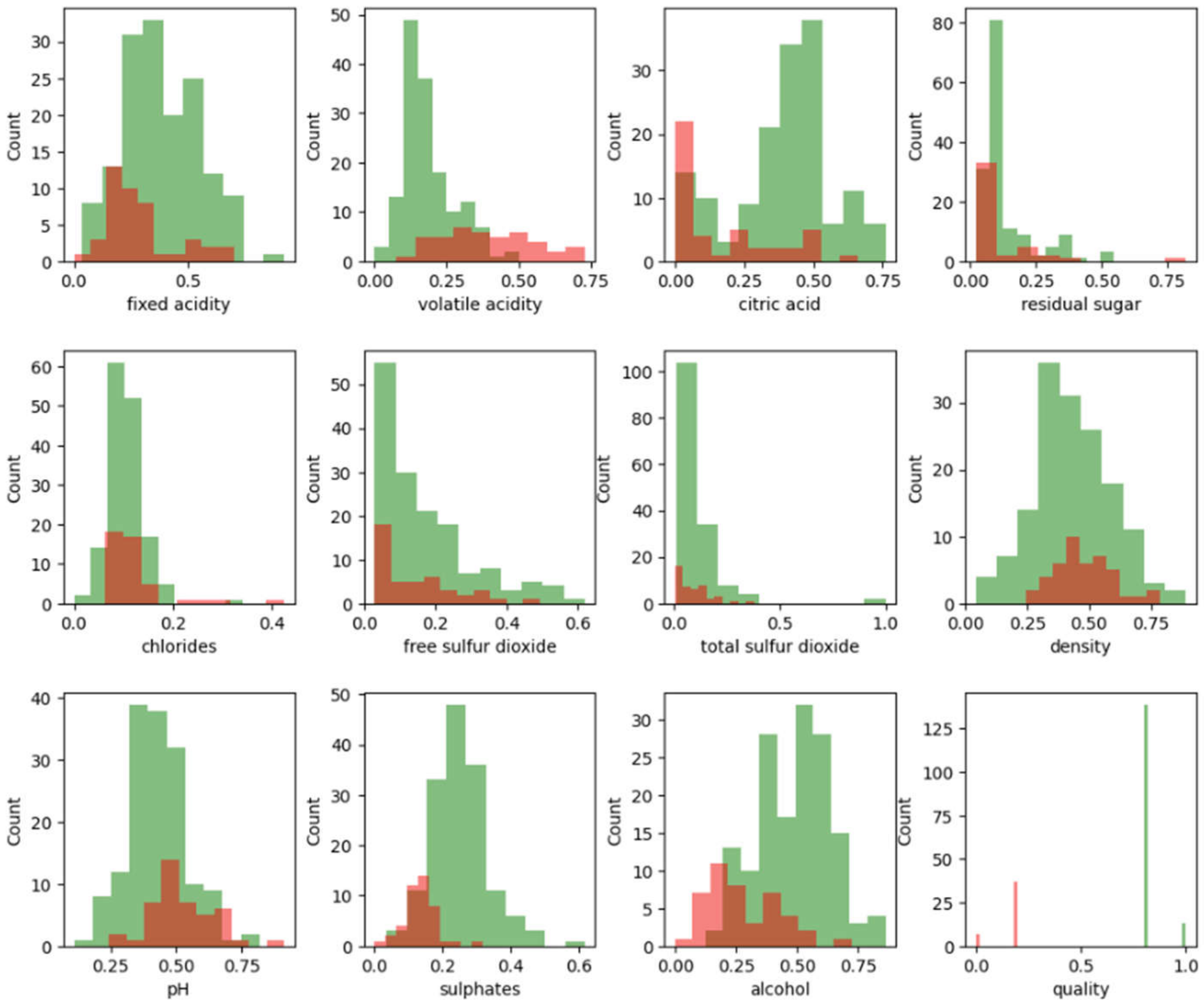
Observations on explanatory variables

- explanatory variables don't show any multimodality
- pH and density are likely normally distributed
- all other variables show a skewed distribution

Observations on wine quality

- most tested wines are average
- very limited number (proportionally) of good/bad wines, which might considerably affect the quality of any classifier
- modeling wine quality and differentiating good/bad wines might require to discard average wines if relevant and needed

2. EXPLORATORY DATA ANALYSIS – HISTOGRAMS WITH HIGH/LOW SCORE SPLIT



Focus: quality scores ≤ 0.2 and ≥ 0.8 only

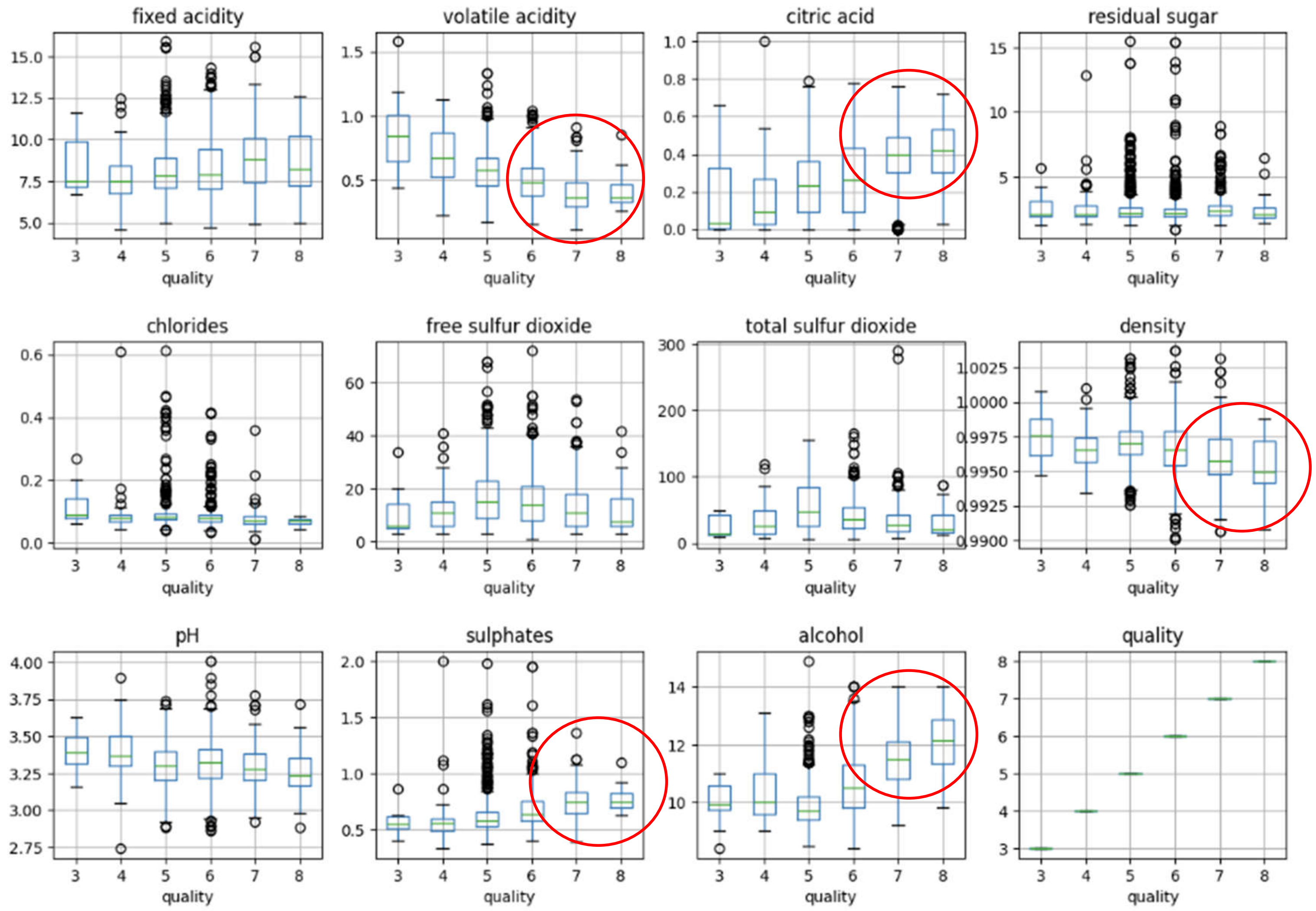
Factors that might explain high wine quality

- much more alcohol (++)
- much lower volatile acidity (++)
- quite higher fixed acidity (+)
- more citric acid (+)
- slightly more residual sugar (+/-)
- possibly less free sulphur dioxide (+/-)
- low total sulphur dioxide (+/-)
- slightly more sulphates (+/-)

Factors that don't seem to have a significant effect

- chlorides
- density (not so clear)
- pH or maybe slightly lower can have a positive effect

2. EXPLORATORY DATA ANALYSIS – BOXPLOTS PER QUALITY SCORE LEVEL

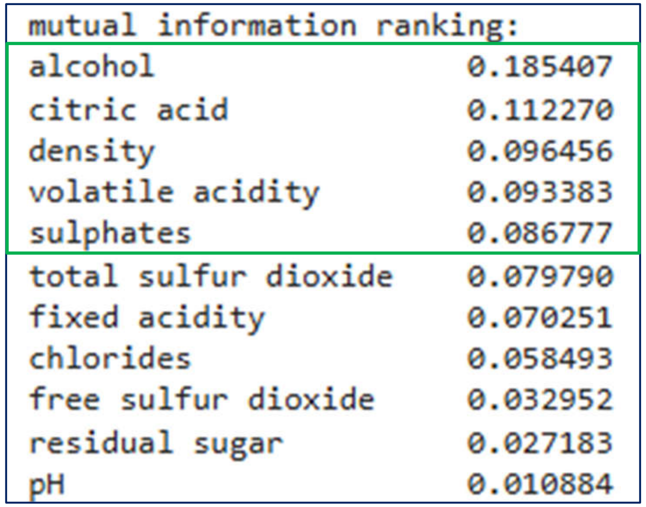


Better wines seem to show

- clearly more alcohol (++)
- clearly more citric acid (++)
- a clearly lower volatile acidity (++)
- more sulphates (+)
- a lower density (+)
- a higher average fixed acidity but might not be significative
- a slightly lower pH on average

2. EXPLORATORY DATA ANALYSIS – MUTUAL INFORMATION (MI)

Mutual information aims at quantifying the explanatory level of each factor in favour of the wine quality score

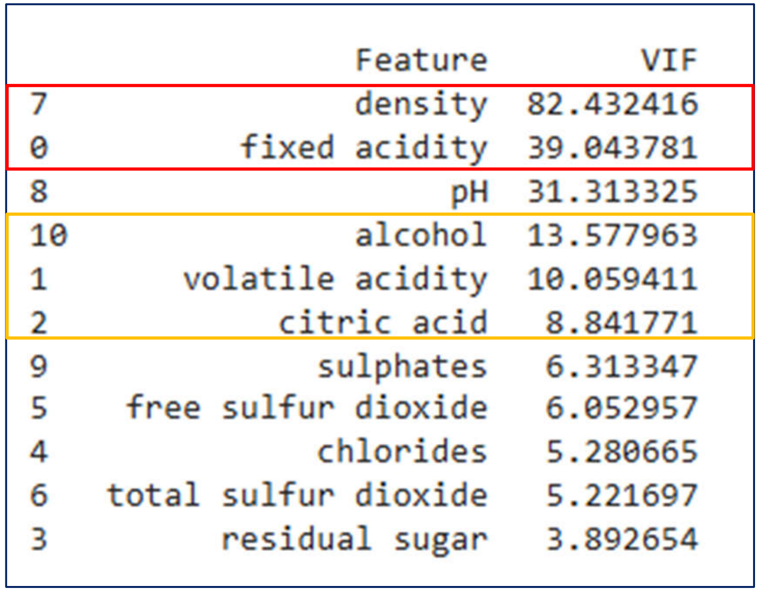


```
mutual information ranking:
alcohol                 0.185407
citric acid             0.112270
density                 0.096456
volatile acidity        0.093383
sulphates               0.086777
total sulfur dioxide    0.079790
fixed acidity           0.070251
chlorides               0.058493
free sulfur dioxide     0.032952
residual sugar          0.027183
pH                      0.010884
```

- The mutual information ranking confirms the observations previously made from histograms and box plots per wine quality score

3. REGRESSION MODEL TO PREDICT WINE QUALITY

Dimension reduction/features selection - variance inflation analysis

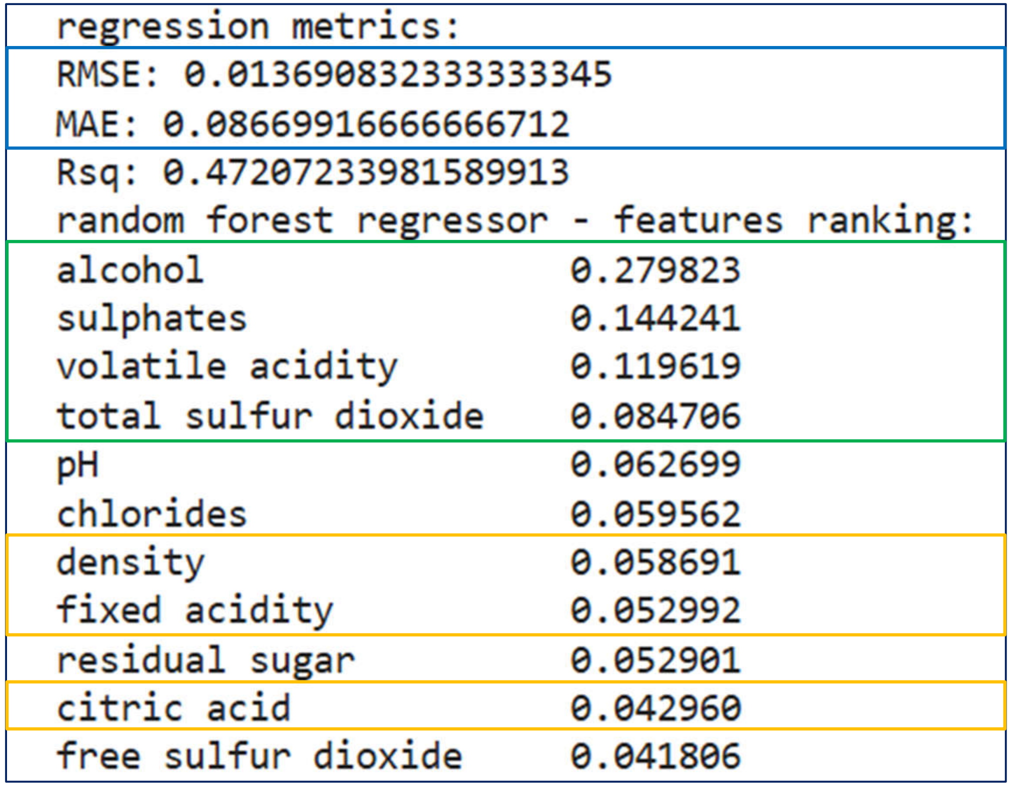


```
                 Feature        VIF
7                density  82.432416
0          fixed acidity  39.043781
8                     pH  31.313325
10               alcohol  13.577963
1       volatile acidity  10.059411
2            citric acid   8.841771
9              sulphates   6.313347
5    free sulfur dioxide   6.052957
4              chlorides   5.280665
6   total sulfur dioxide   5.221697
3         residual sugar   3.892654
```

- VIF values: higher = redundant
- Some of the variables with high explanatory power also seem to show high/average VIF values, i.e. a form of redundancy with other variables - which is not necessarily apparent in linear or rank correlation matrices (see Appendix)
- This aspect can be kept in mind, as some predictive models might be sensitive to correlated/redundant features
  - If sensitive (e.g. linear models, for which coefficients can become unstable/not reliable for interpretations), a proper filtering of variables can be done
  - As an alternative to features selection, the model can also be selected to be robust to correlated/redundant variables

3. REGRESSION MODEL TO PREDICT WINE QUALITY

Fitting and assessing a random forest regression model

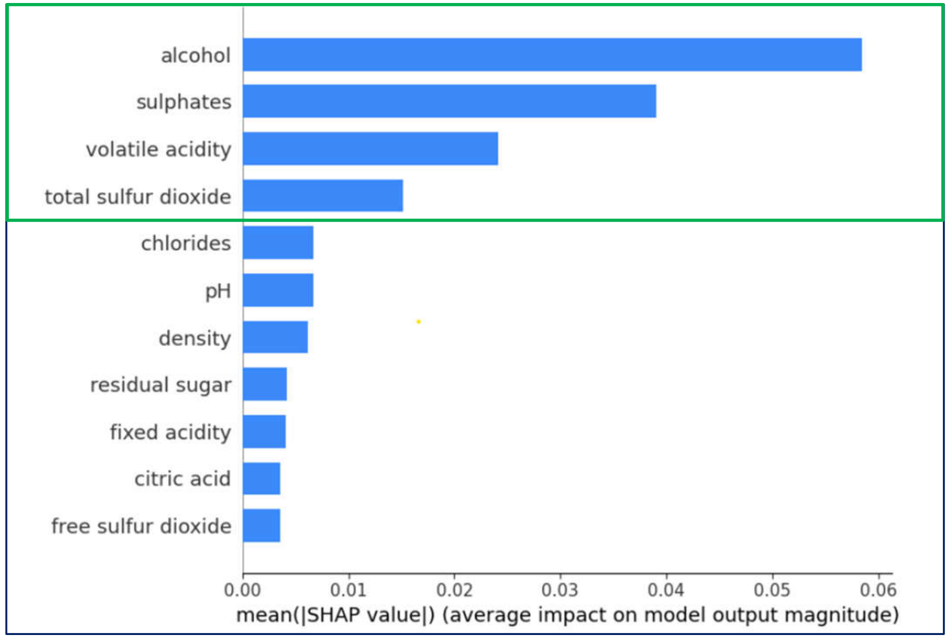


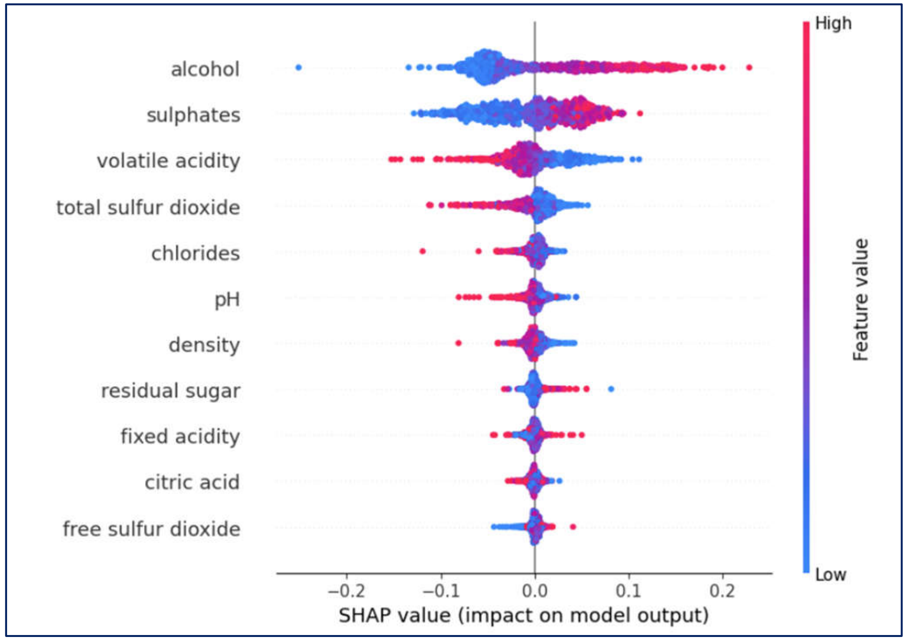
```
regression metrics:
RMSE: 0.01369083233333345
MAE: 0.0866991666666712
Rsq: 0.47207233981589913
random forest regressor - features ranking:
alcohol                 0.279823
sulphates               0.144241
volatile acidity        0.119619
total sulfur dioxide    0.084706
pH                      0.062699
chlorides               0.059562
density                 0.058691
fixed acidity           0.052992
residual sugar          0.052901
citric acid             0.042960
free sulfur dioxide     0.041806
```

- The root mean squared error is very good (1.37%)
- As an alternative, the mean absolute error is good as well, for scores defined from 0 to 1 by a step of 0.2
- The features ranking shows that alcohol, volatile acidity and total sulphur dioxide have, in decreasing order, the highest contribution to the model
- Since the model considers the redundancy between variables, this ranking slightly differs from the mutual information & EDA analysis
  - the first four factors are a confirmation
  - acid nitric and density seem to have a low added value once the first factors are already considered
- The regression model allows to predict the wine quality score, which in turns also allows a classification

3. REGRESSION MODEL TO PREDICT WINE QUALITY

SHAP interpretation





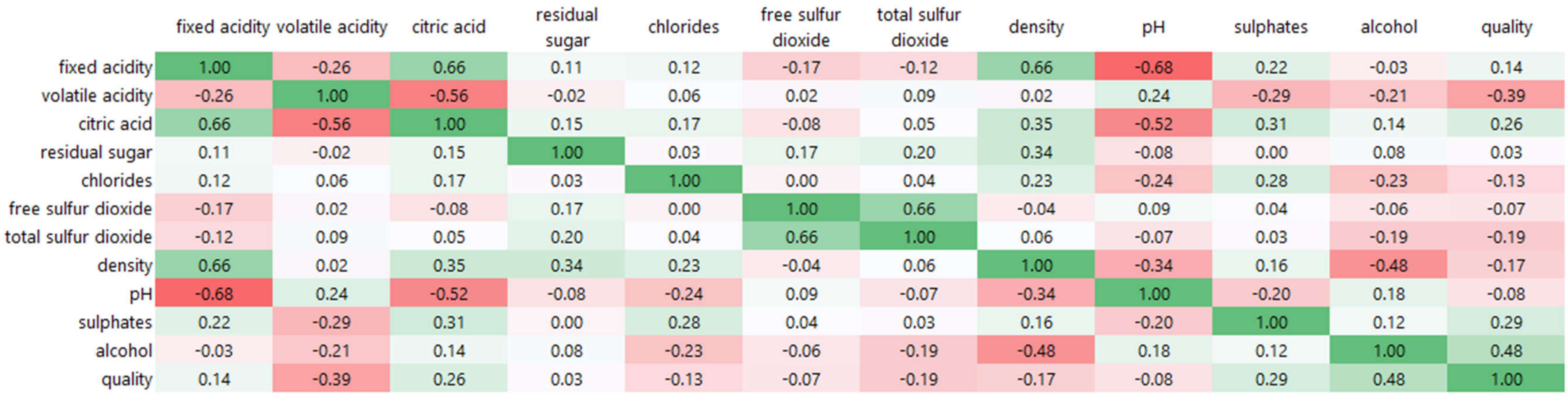
- The SHAP analysis allows to quantify and visualize the contributions of the features to the decision tree
- Features/variables contributing the most to predictive model are sorted by decreasing order
- Alcohol %, sulphate concentration and volatile acidity are the top three factors influencing wine quality in the model
- In pink (blue) are represented high (low) variable values, hence a high alcohol % and low volatile acidity both contribute to high wine quality score

4. CONCLUSION AND PERSPECTIVES

- **A detailed exploratory data analysis** shows that high quality scores are usually characterized by
  - significantly higher alcohol and citric acid concentrations, and a lower volatile acidity
  - with a more secondary effect, a lower density, higher sulphates concentration and a slightly lower pH on average also seem to contribute, to a lower extent
- **A random forest regression model** has been fit to the data (and assessed on an independent test set), and allowed to reach high accuracy measures in predicting the 'Verdo Vinho' wine quality score:
  - a root mean squared error (RMSE) of 1.37% and a mean absolute error (MAE) of 8.67%
  - while also confirming the observations made in the exploratory data analysis, with exceptions
- **Finally, a SHAP analysis** has allowed to consolidate our interpretations
  - it shows how each explanatory variable contributes to the predictive model and a high/low wine quality score
  - it confirms the alcohol %, citric acid concentration and volatile acidity as key drivers and explanatory variables in the regression model
- **Perspectives**
  - Such a predictive model can be useful to support oenologist wine tasting evaluations and improve wine production. Similar techniques can also help in target marketing by modeling consumer tastes
  - Classification models and other regression models could be tested as well, aside the robust random forest regression model used in this analysis
  - Extending the analysis to other wine types, to extract commonalities, is likely to generate interesting insights too

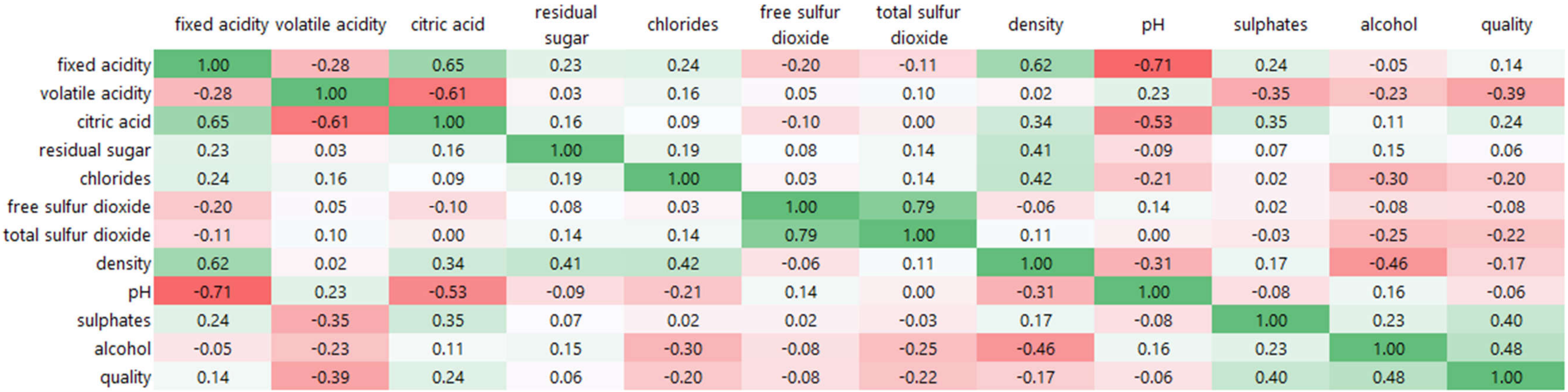
APPENDIX

2. EXPLORATORY DATA ANALYSIS – PEARSON’S LINEAR CORRELATION MATRIX



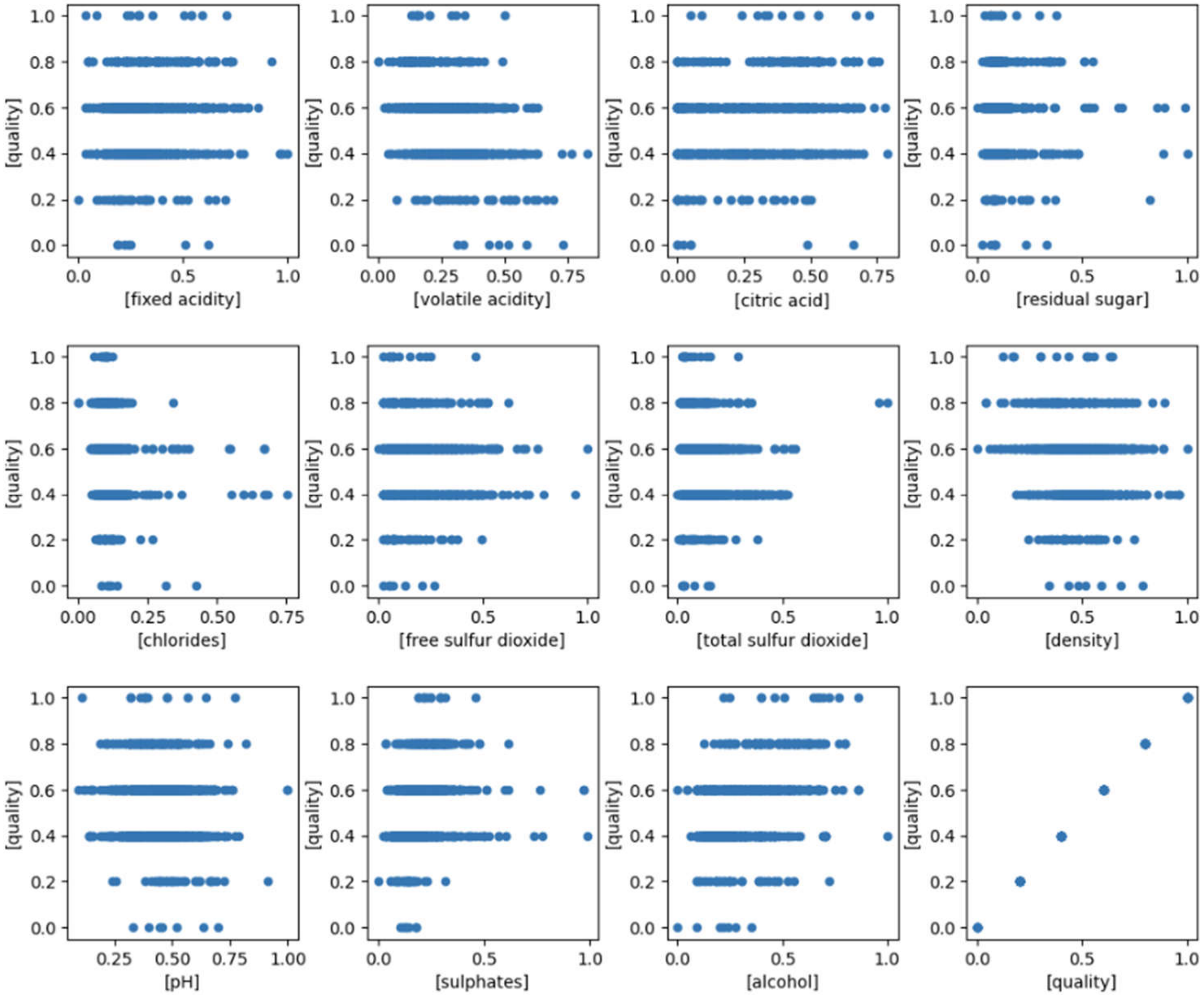
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.26	0.66	0.11	0.12	-0.17	-0.12	0.66	-0.68	0.22	-0.03	0.14
volatile acidity	-0.26	1.00	-0.56	-0.02	0.06	0.02	0.09	0.02	0.24	-0.29	-0.21	-0.39
citric acid	0.66	-0.56	1.00	0.15	0.17	-0.08	0.05	0.35	-0.52	0.31	0.14	0.26
residual sugar	0.11	-0.02	0.15	1.00	0.03	0.17	0.20	0.34	-0.08	0.00	0.08	0.03
chlorides	0.12	0.06	0.17	0.03	1.00	0.00	0.04	0.23	-0.24	0.28	-0.23	-0.13
free sulfur dioxide	-0.17	0.02	-0.08	0.17	0.00	1.00	0.66	-0.04	0.09	0.04	-0.06	-0.07
total sulfur dioxide	-0.12	0.09	0.05	0.20	0.04	0.66	1.00	0.06	-0.07	0.03	-0.19	-0.19
density	0.66	0.02	0.35	0.34	0.23	-0.04	0.06	1.00	-0.34	0.16	-0.48	-0.17
pH	-0.68	0.24	-0.52	-0.08	-0.24	0.09	-0.07	-0.34	1.00	-0.20	0.18	-0.08
sulphates	0.22	-0.29	0.31	0.00	0.28	0.04	0.03	0.16	-0.20	1.00	0.12	0.29
alcohol	-0.03	-0.21	0.14	0.08	-0.23	-0.06	-0.19	-0.48	0.18	0.12	1.00	0.48
quality	0.14	-0.39	0.26	0.03	-0.13	-0.07	-0.19	-0.17	-0.08	0.29	0.48	1.00

2. EXPLORATORY DATA ANALYSIS – SPEARMAN’S RANK CORRELATION MATRIX



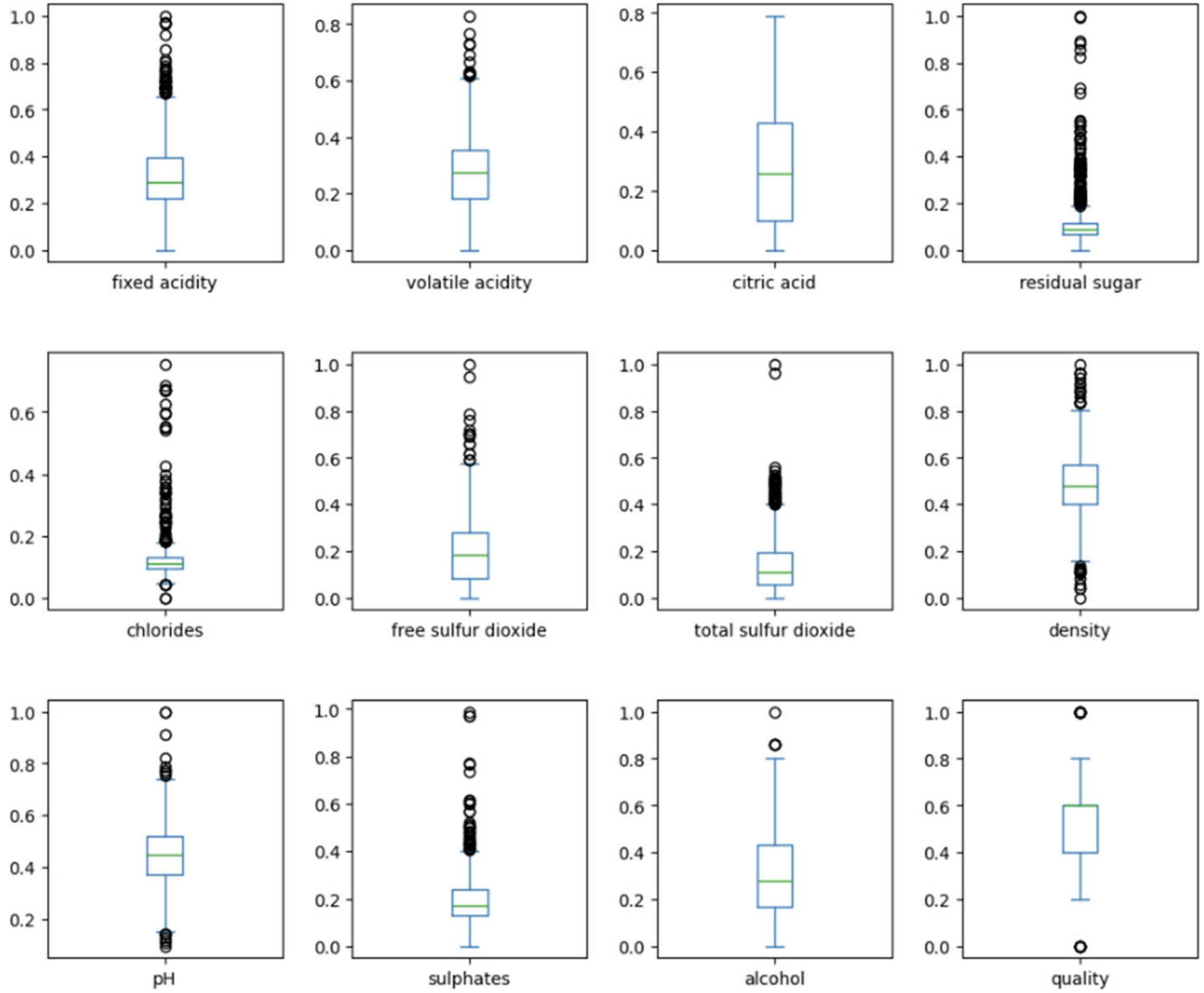
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.28	0.65	0.23	0.24	-0.20	-0.11	0.62	-0.71	0.24	-0.05	0.14
volatile acidity	-0.28	1.00	-0.61	0.03	0.16	0.05	0.10	0.02	0.23	-0.35	-0.23	-0.39
citric acid	0.65	-0.61	1.00	0.16	0.09	-0.10	0.00	0.34	-0.53	0.35	0.11	0.24
residual sugar	0.23	0.03	0.16	1.00	0.19	0.08	0.14	0.41	-0.09	0.07	0.15	0.06
chlorides	0.24	0.16	0.09	0.19	1.00	0.03	0.14	0.42	-0.21	0.02	-0.30	-0.20
free sulfur dioxide	-0.20	0.05	-0.10	0.08	0.03	1.00	0.79	-0.06	0.14	0.02	-0.08	-0.08
total sulfur dioxide	-0.11	0.10	0.00	0.14	0.14	0.79	1.00	0.11	0.00	-0.03	-0.25	-0.22
density	0.62	0.02	0.34	0.41	0.42	-0.06	0.11	1.00	-0.31	0.17	-0.46	-0.17
pH	-0.71	0.23	-0.53	-0.09	-0.21	0.14	0.00	-0.31	1.00	-0.08	0.16	-0.06
sulphates	0.24	-0.35	0.35	0.07	0.02	0.02	-0.03	0.17	-0.08	1.00	0.23	0.40
alcohol	-0.05	-0.23	0.11	0.15	-0.30	-0.08	-0.25	-0.46	0.16	0.23	1.00	0.48
quality	0.14	-0.39	0.24	0.06	-0.20	-0.08	-0.22	-0.17	-0.06	0.40	0.48	1.00

2. EXPLORATORY DATA ANALYSIS – SCATTER PLOTS



- Any type of relationship between y and x's is difficult to find, due to the discrete nature of the wine quality variable
- no variable seem to show obvious outliers, except maybe total sulphur dioxide and citric acid for one data point each
- data quality can be confirmed on the test set as well

2. EXPLORATORY DATA ANALYSIS – BOX PLOTS


fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality