Predicting wine quality from its chemical properties

Presented on 10/21/2019
By Marcin Kolaszewski
At Brown University

Git repo: `github.com/marklin36/1030-final-project.git`

Dataset and motivation

Predicting wine quality from its chemical properties

- Quality score is discrete → classification problem
- What are chemical properties of high quality white wine?
- Stakeholders:
  - Consumers → better choice
  - Producers → better offer
- To what extend wine quality is objective?

Source: UCL machine learning repository
`archive.ics.uci.edu/ml/datasets/wine+quality`

Variable	Units
Fixed acidity	pH
Volatile acidity	g/L
Residual sugar	g/L
Chlorides	g/L
Free sulfur oxide	ppm
Total sulfur oxide	ppm
Density	g/L
Sulphates	g/L
Acidity	pH
Alcohol	alcohol
Quality	1-12 score

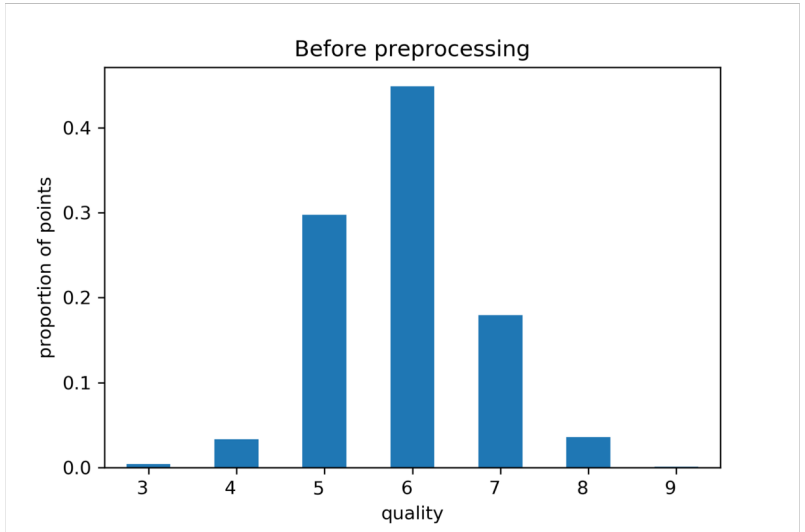
Preprocessing

All 12 features are continuous

Standard scaling was applied to all features

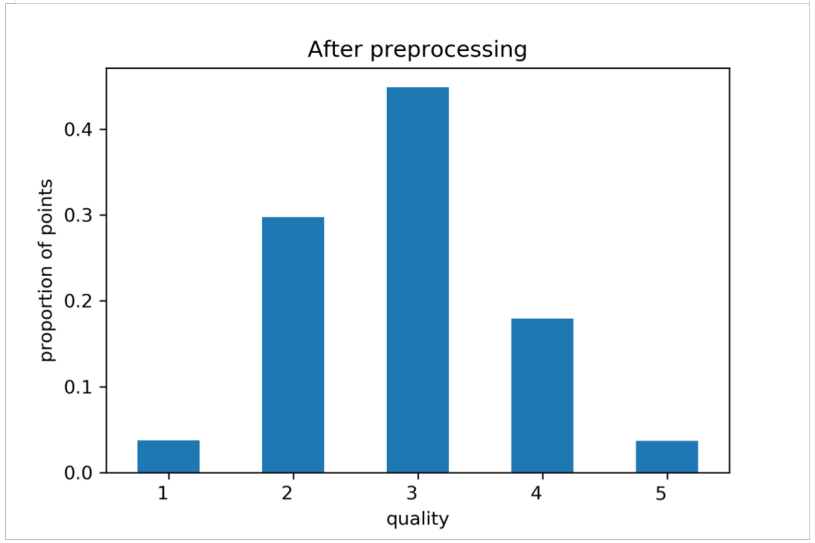
No missing values

Number of data points: 4898

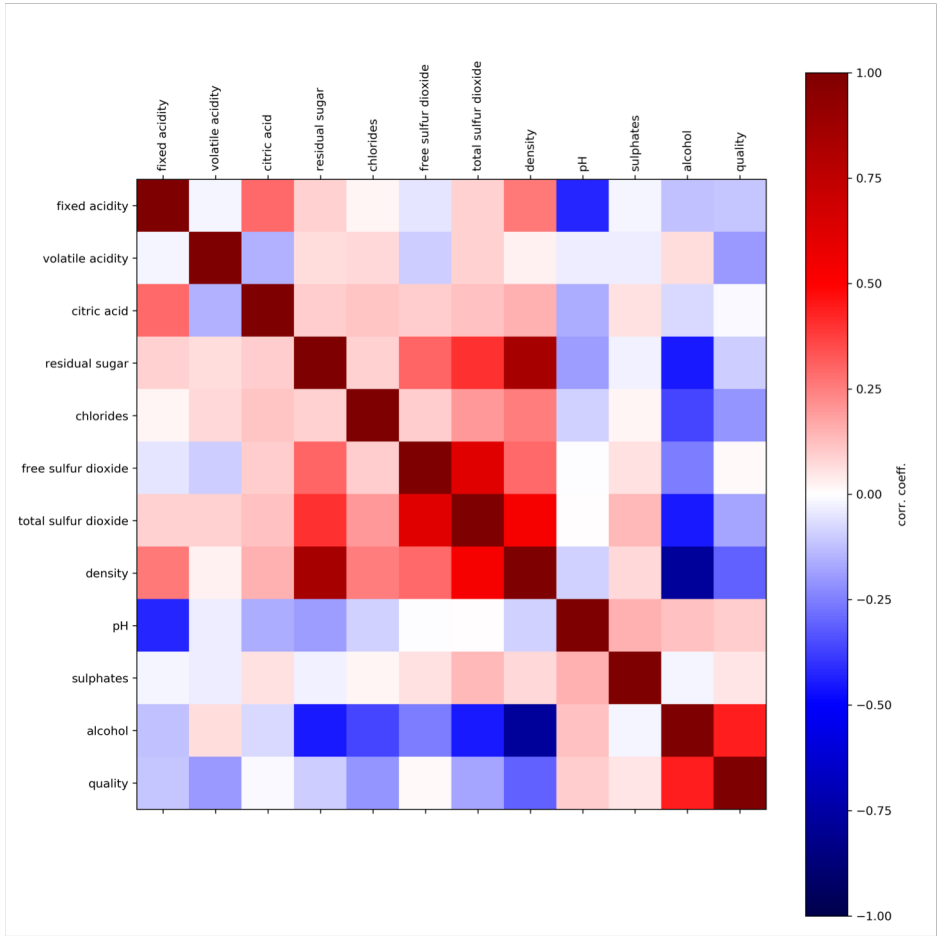


Label transformation:

- Less than 5 → 1
- 5 → 2
- 6 → 3
- 7 → 4
- More than 7 → 5



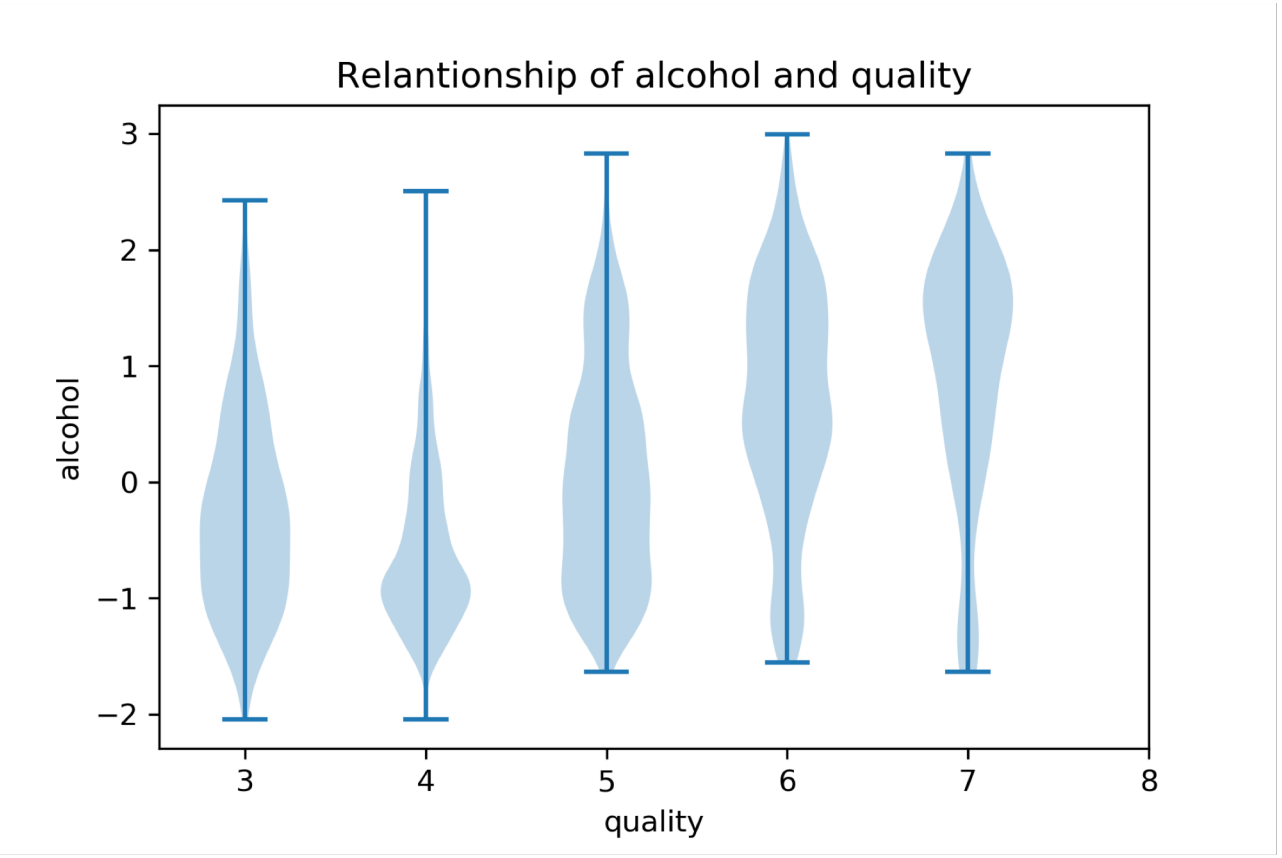
What are the most important features?

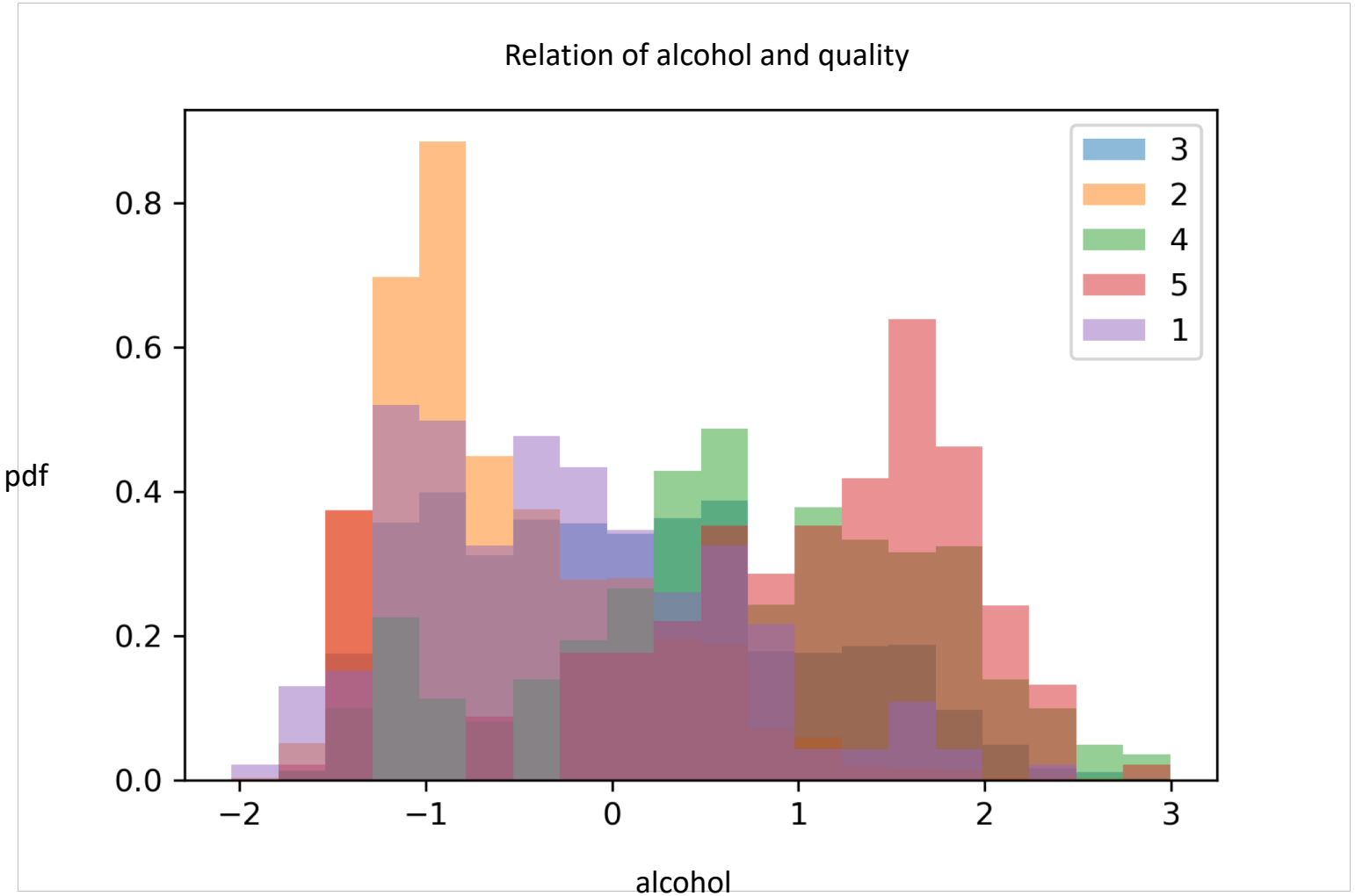


Visualization of the correlation matrix

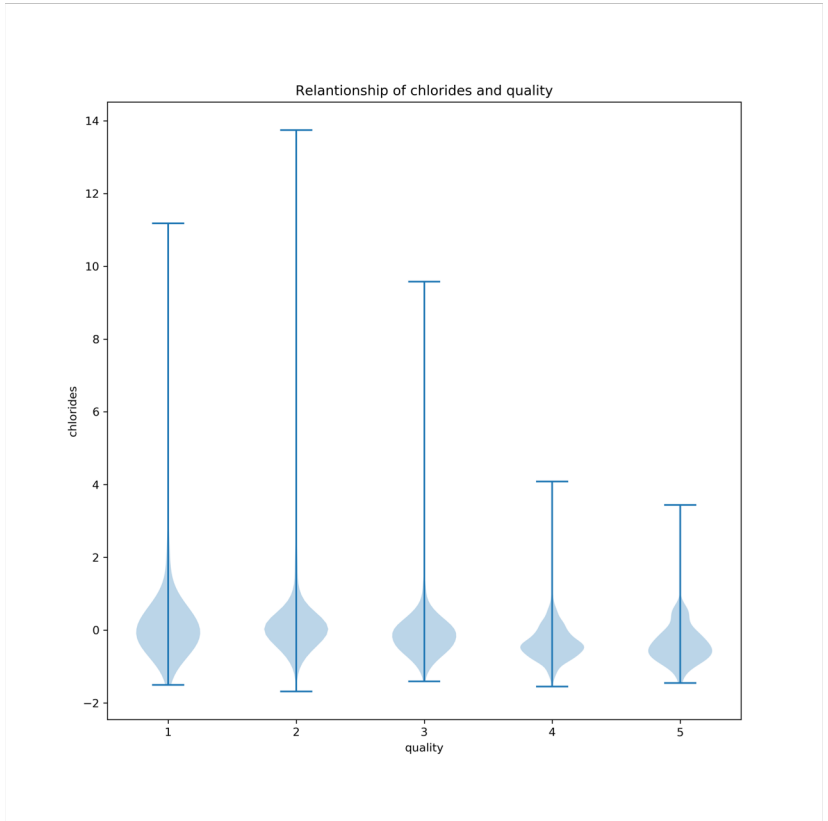
Variable	Correlation
Density	-0.309747
Chlorides	-0.210518
Volatile acidity	-0.195547
Total sulfur dioxide	0.173394
Fixed acidity	-0.112149
Residual sugar	-0.098631
Citric acid	-0.009790
Free sulfur dioxide	0.013386
Sulphates	0.054161
ph	0.100084
Alcohol	0.440394

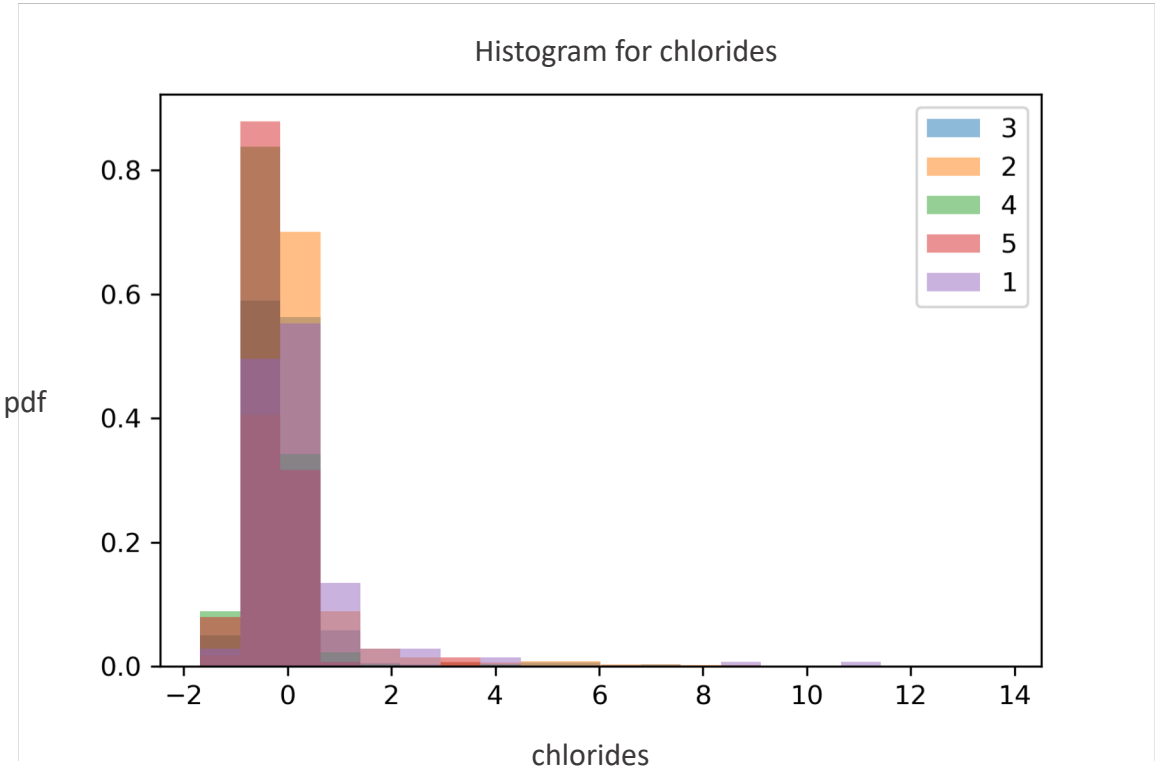
Relationship between alcohol and quality



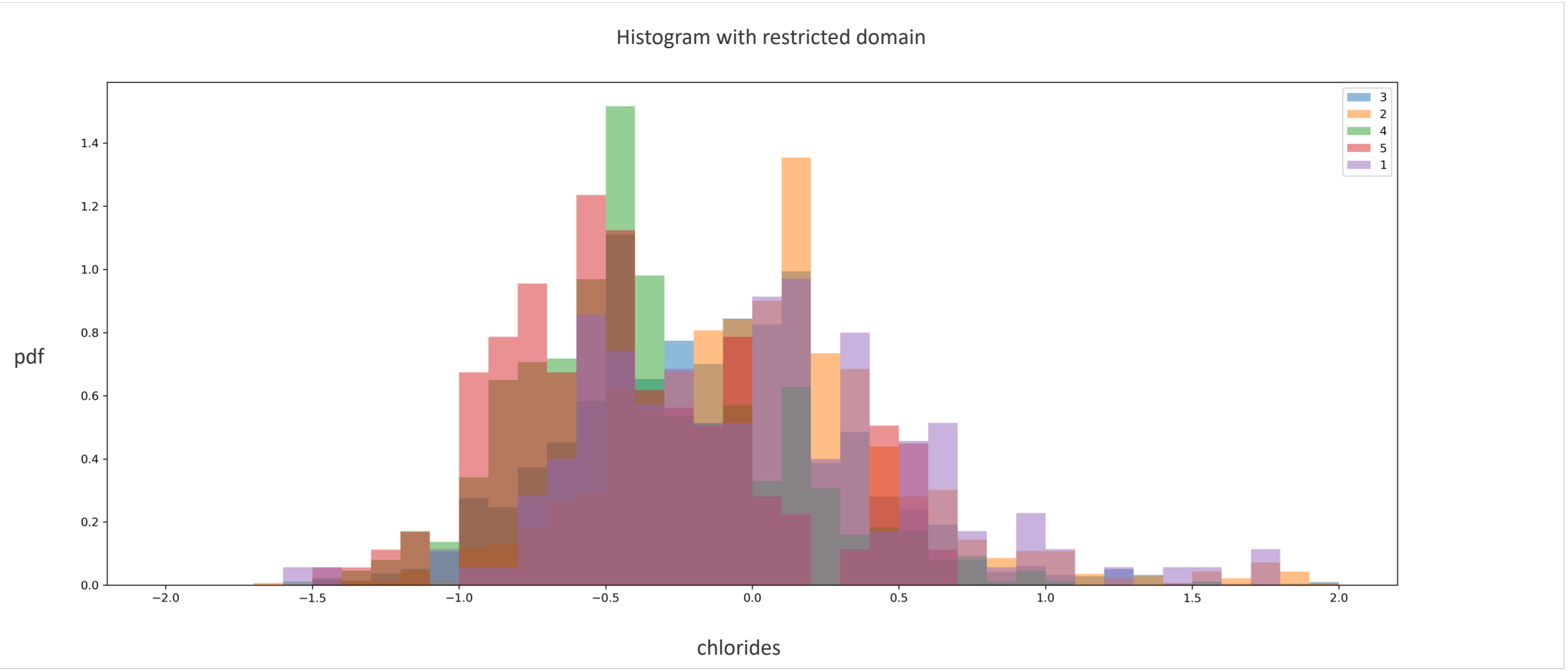

Relation of alcohol and quality
pdf
0.0
0.2
0.4
0.6
0.8
−2
−1
0
1
2
3
alcohol
3
2
4
5
1

Chlorides

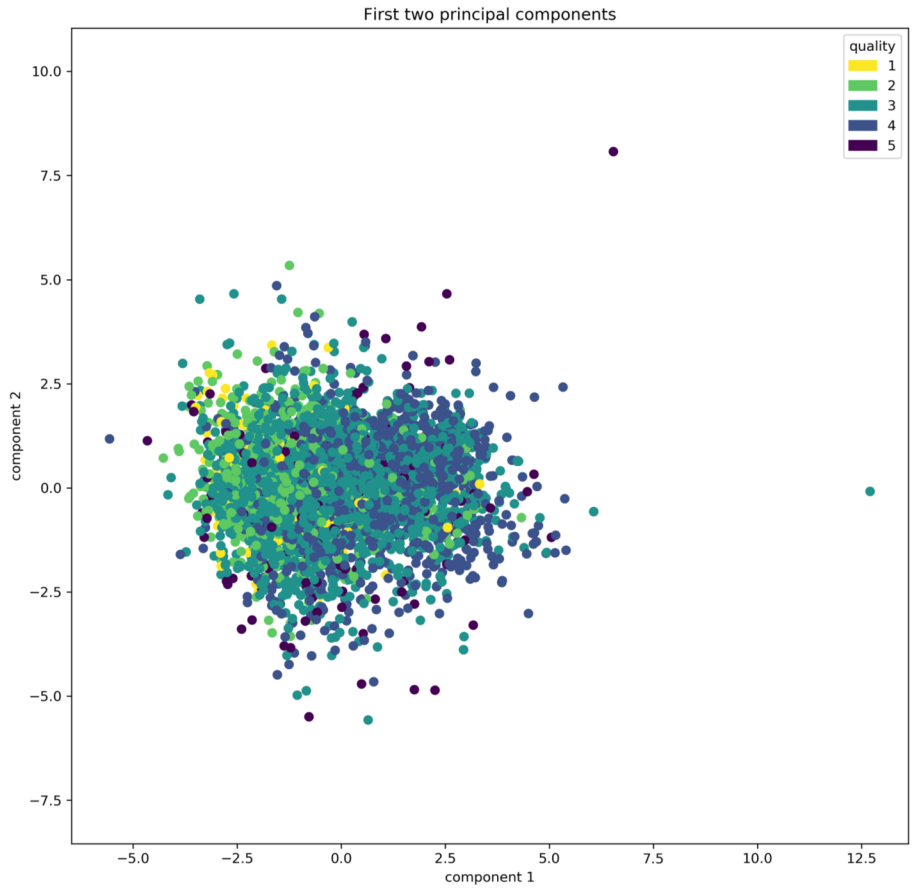


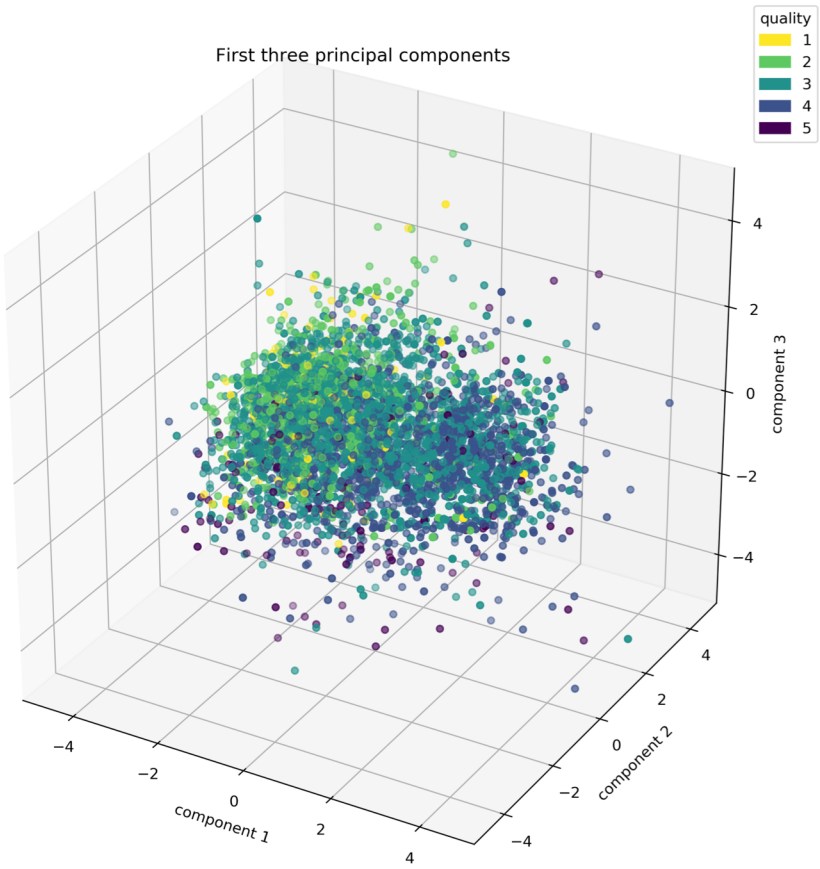


- The distribution of chlorides has a very long right tail
- Closer look is needed

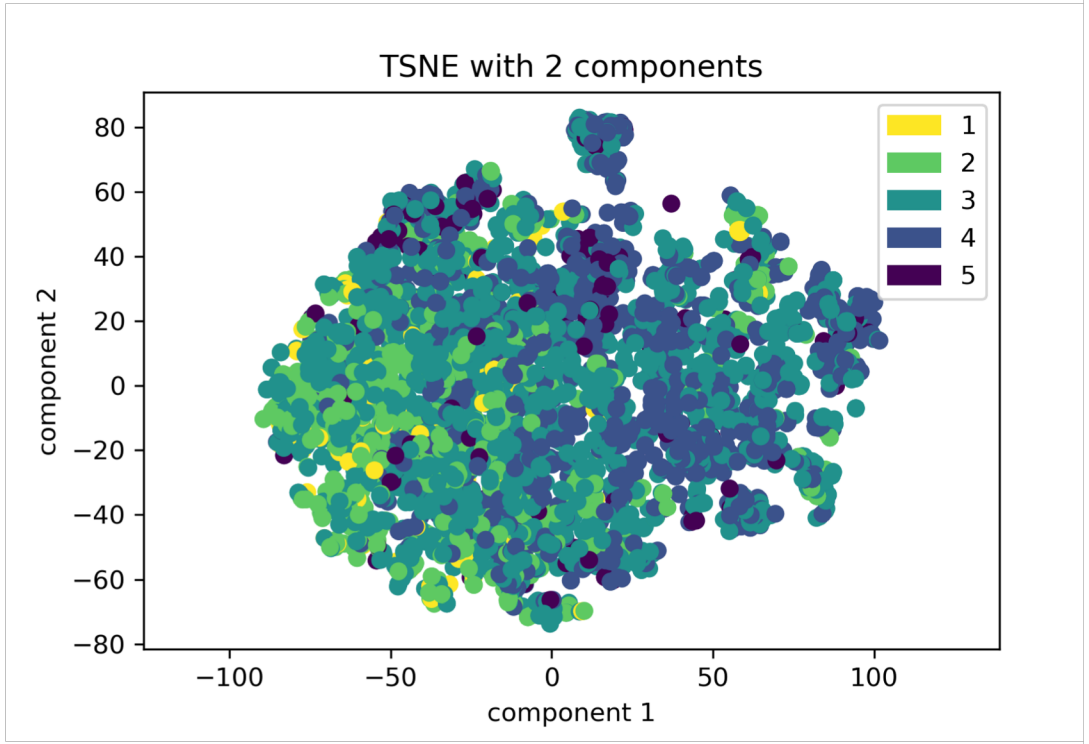

Histogram with restricted domain
pdf
chlorides
3
2
4
5
1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

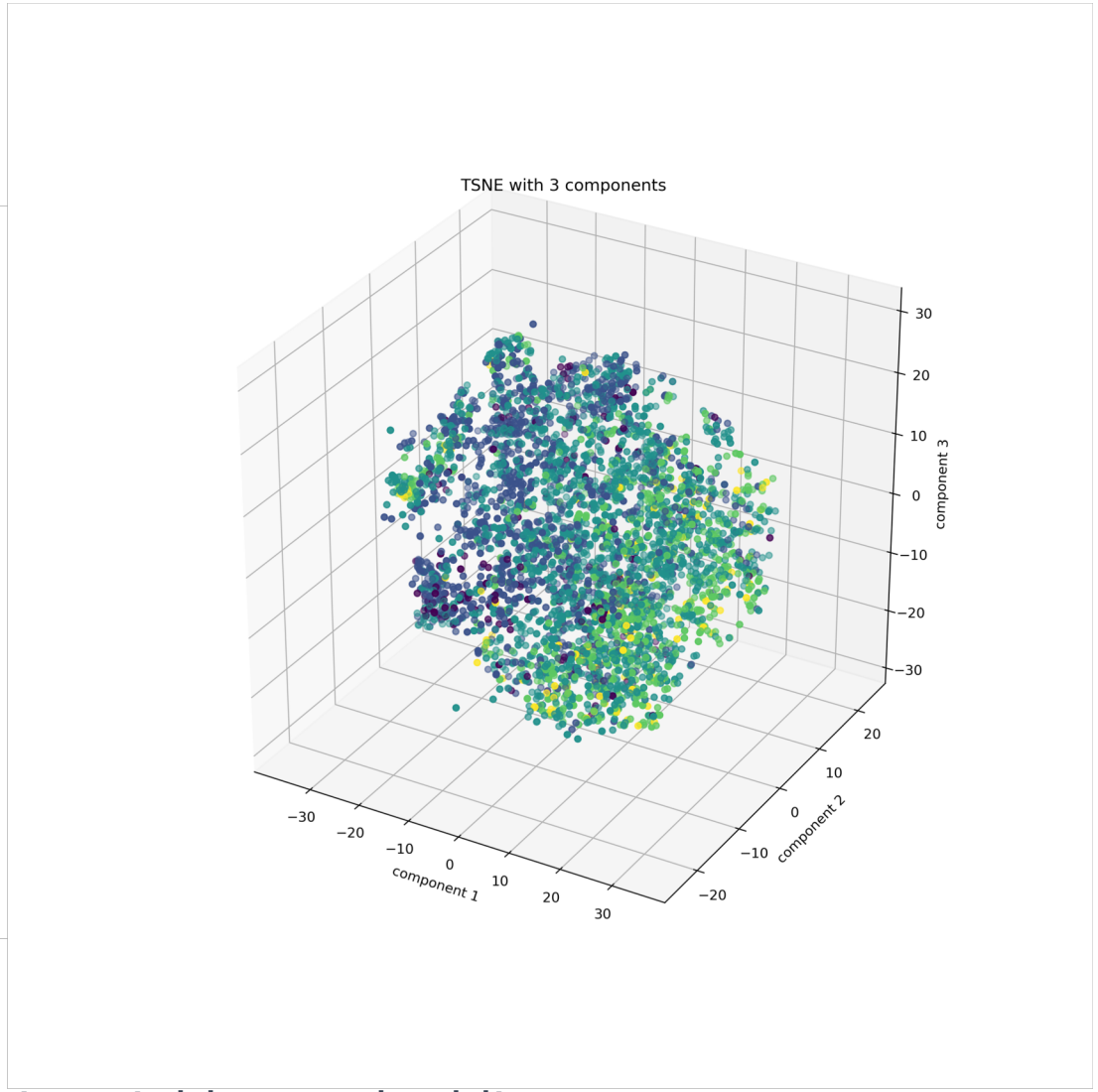
Principal component analysis



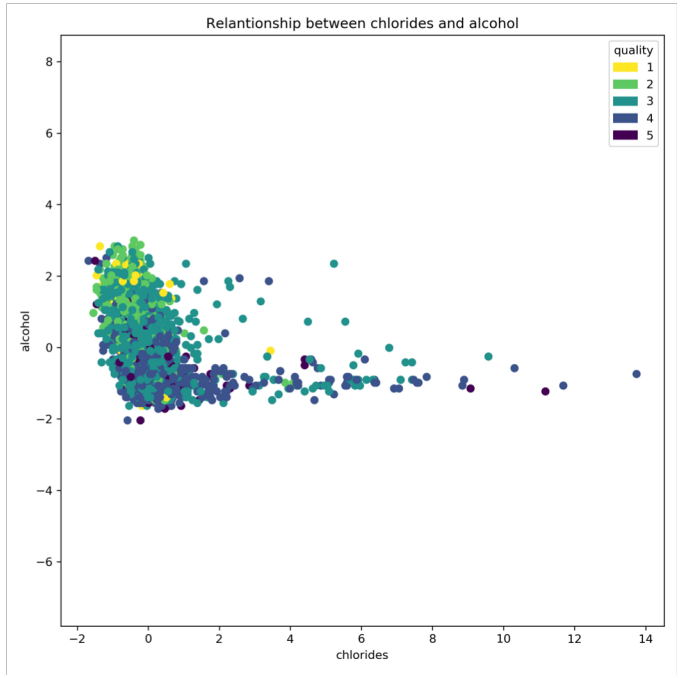


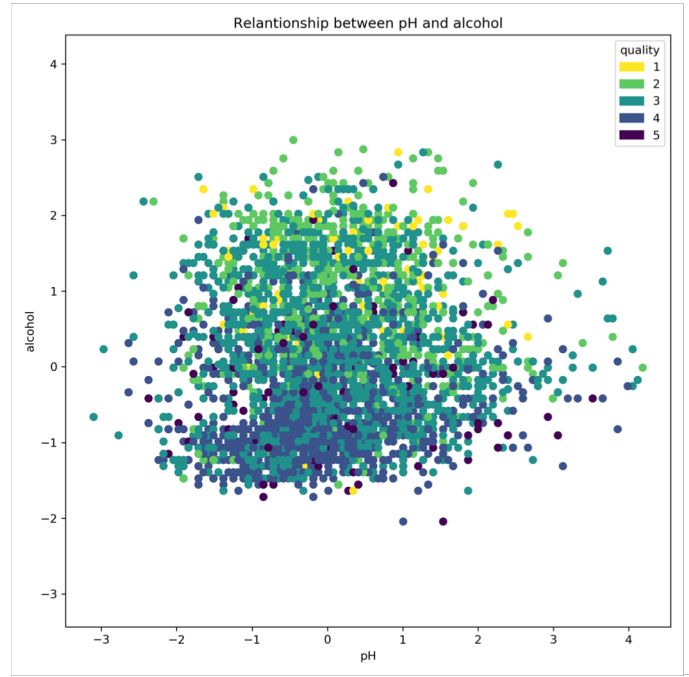
Non-linear dimensionality reduction

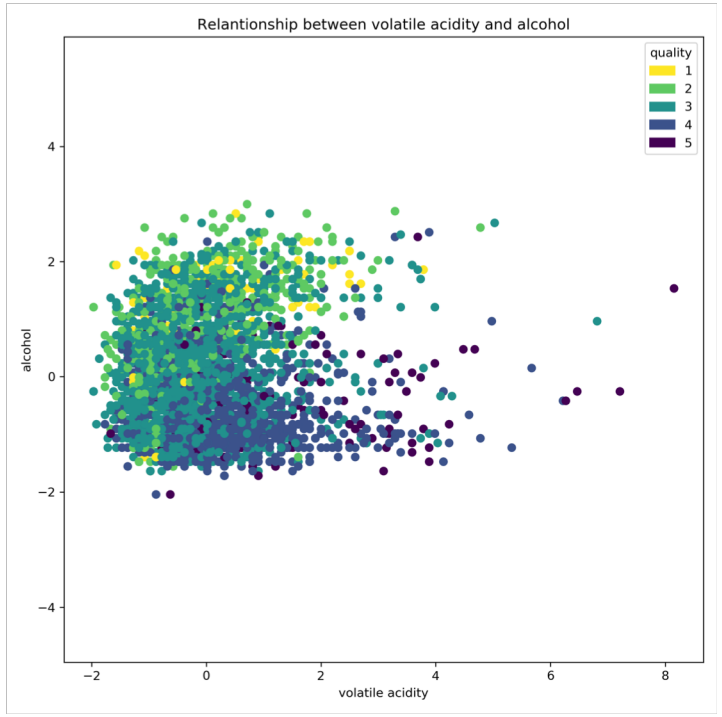


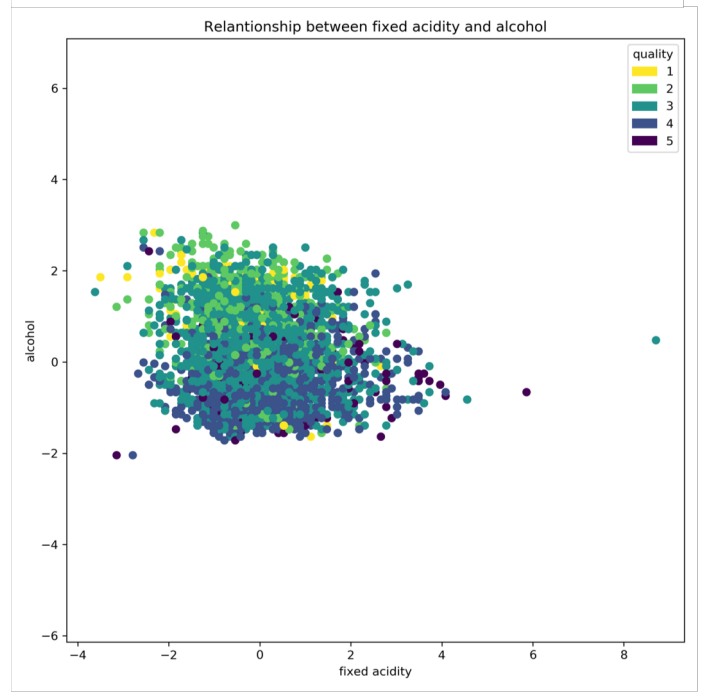


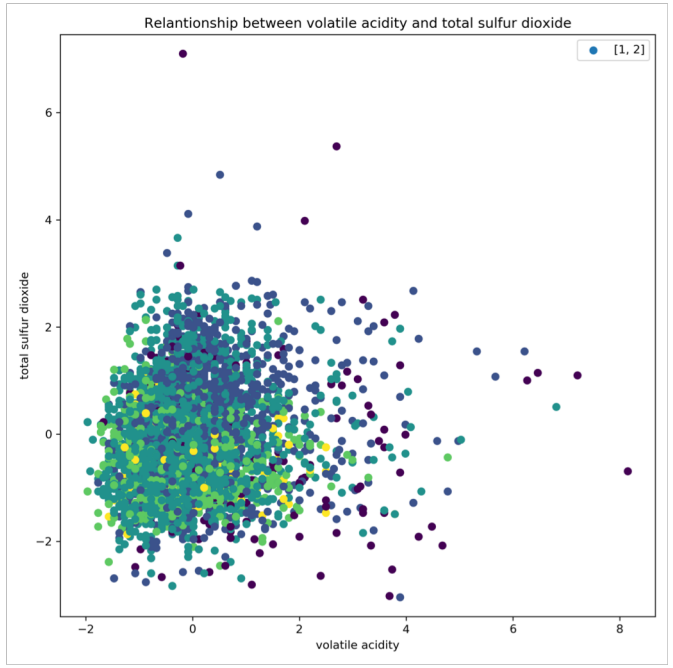
t-distributed stochastic neighbor embedding

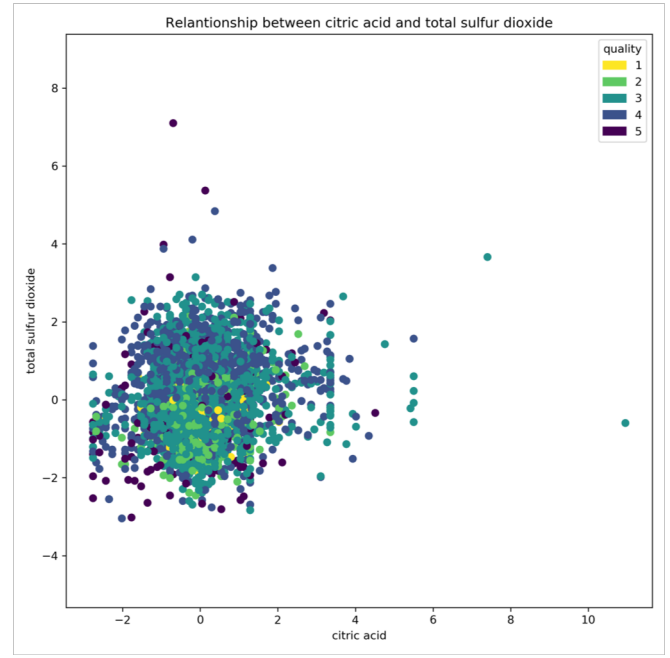

Relantionship between chlorides and alcohol
quality
1
2
3
4
5
alcohol
chlorides

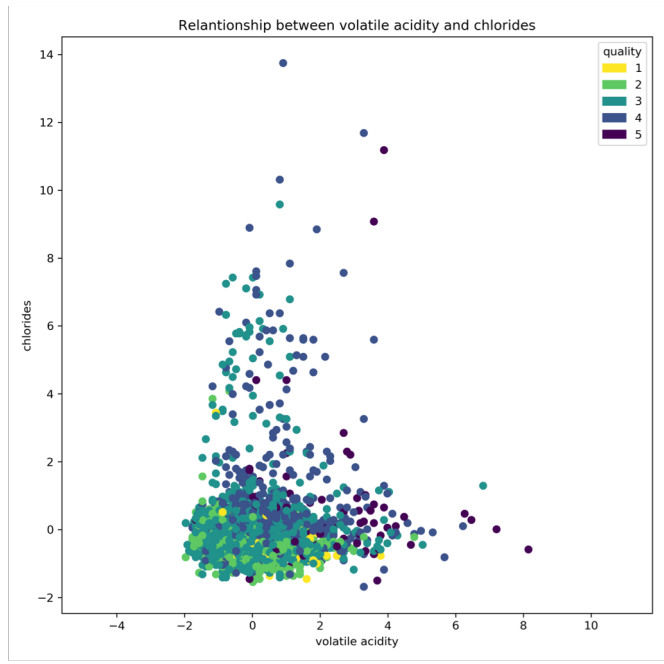

Relantionship between pH and alcohol
quality
1
2
3
4
5
alcohol
pH

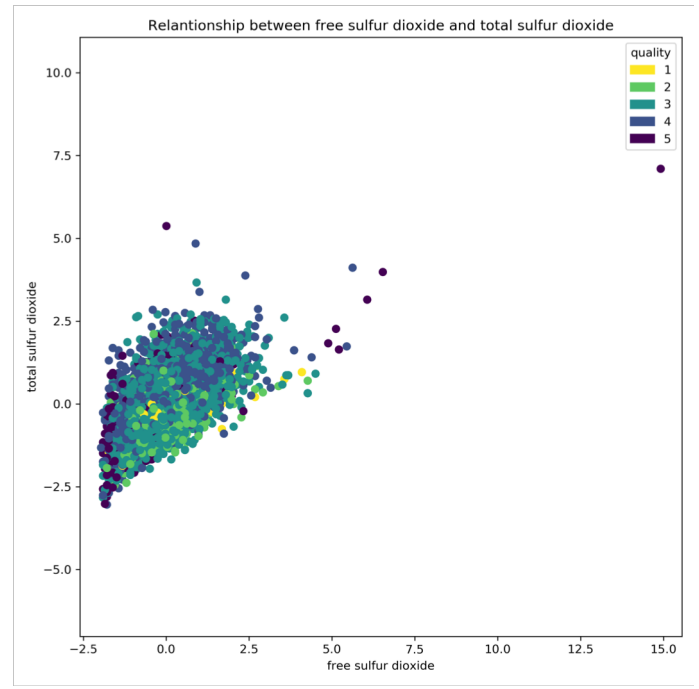

Relantionship between volatile acidity and alcohol
quality
1
2
3
4
5
alcohol
volatile acidity

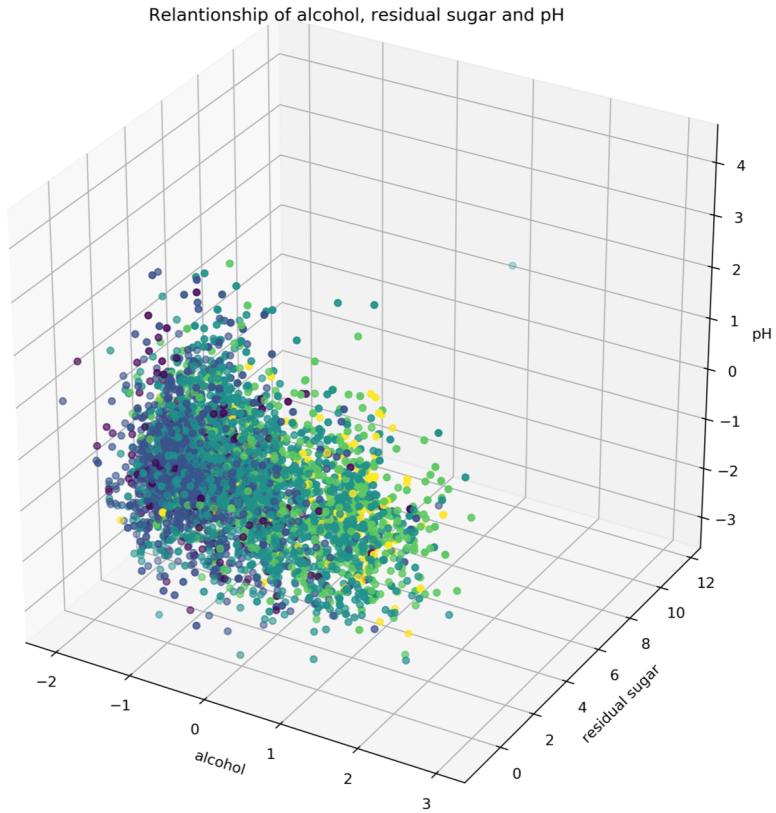

Relantionship between fixed acidity and alcohol
quality
1
2
3
4
5
alcohol
fixed acidity

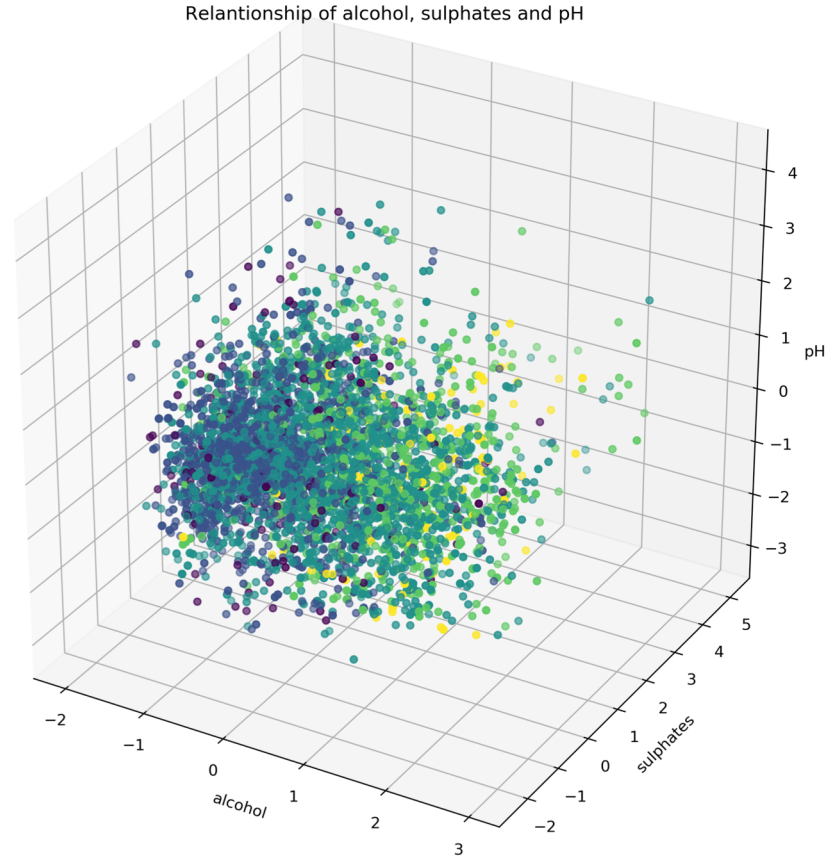

Relantionship between volatile acidity and total sulfur dioxide
[1, 2]
total sulfur dioxide
volatile acidity


Relantionship between citric acid and total sulfur dioxide
quality
1
2
3
4
5
total sulfur dioxide
citric acid


Relantionship between volatile acidity and chlorides
quality
1
2
3
4
5
chlorides
volatile acidity

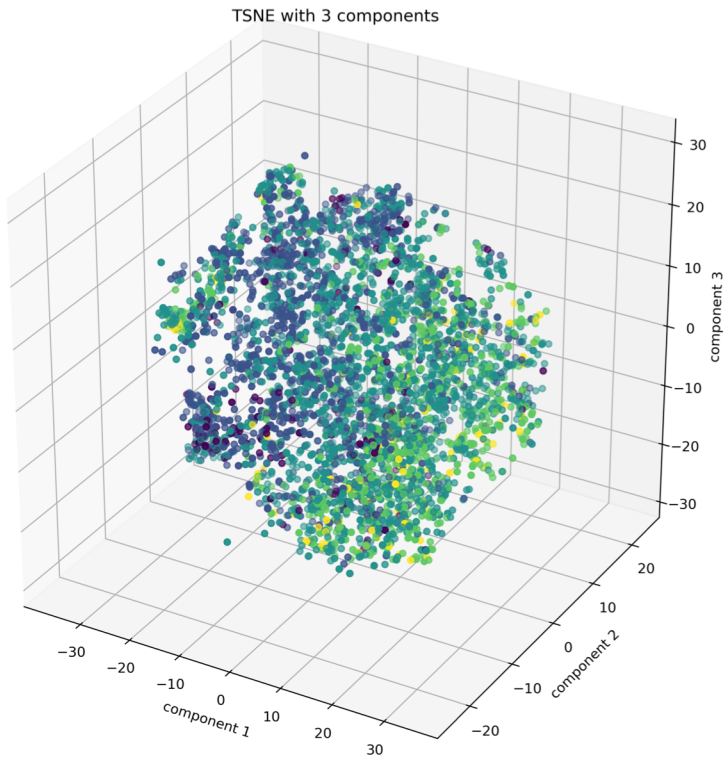

Relantionship between free sulfur dioxide and total sulfur dioxide
quality
1
2
3
4
5
total sulfur dioxide
free sulfur dioxide


Relantionship of alcohol, residual sugar and pH
alcohol
residual sugar
pH


Relantionship of alcohol, sulphates and pH
alcohol
sulphates
pH

Potential approaches to classification

- Logistic regression
- Support vector machines
- Decision trees



SVM with non-linear kernel might get decent results

Q&A