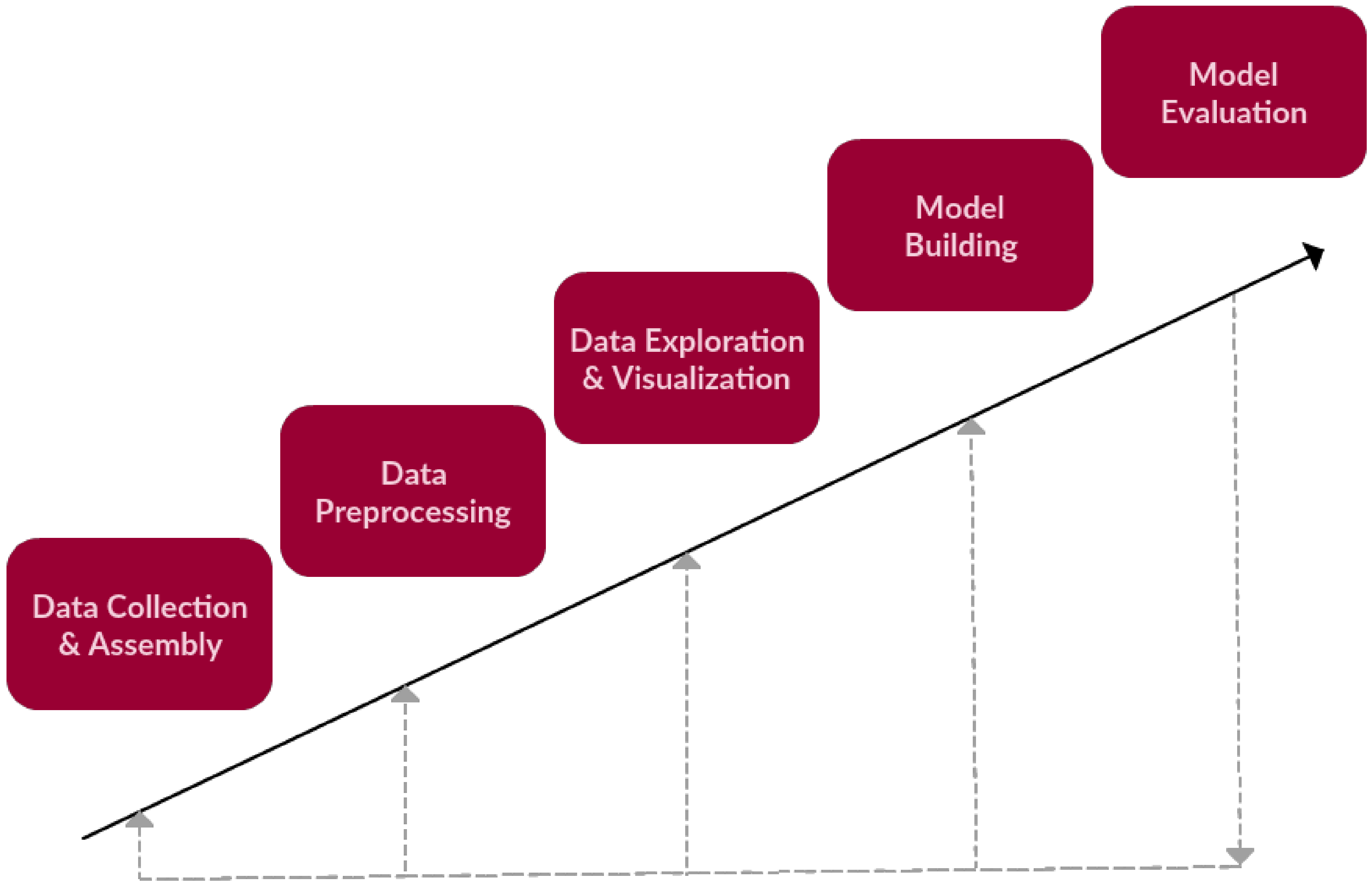
Data Analysis

Practice 1: Data Visualization and Preprocessing

Dr. Nataliya K. Sakhnenko

Data Analysis Steps



Data Collection & Assembly

Data Preprocessing

Data Exploration & Visualization

Model Building

Model Evaluation

Pandas lib

pandas is an open source library providing high-performance, easy-to-use data structures and data analysis tools for the Python

pandas

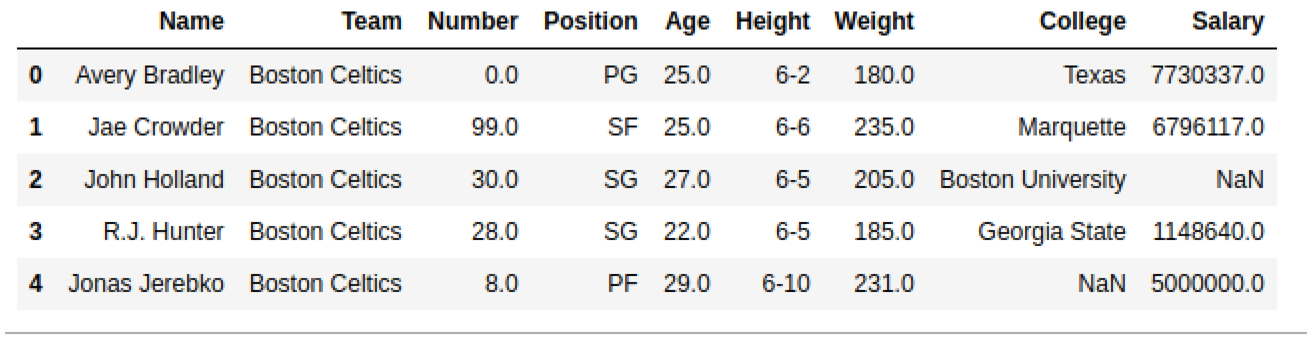
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

pandas.read_csv()

Read a comma-separated values (csv) file into DataFrame.

pandas.DataFrame.head()

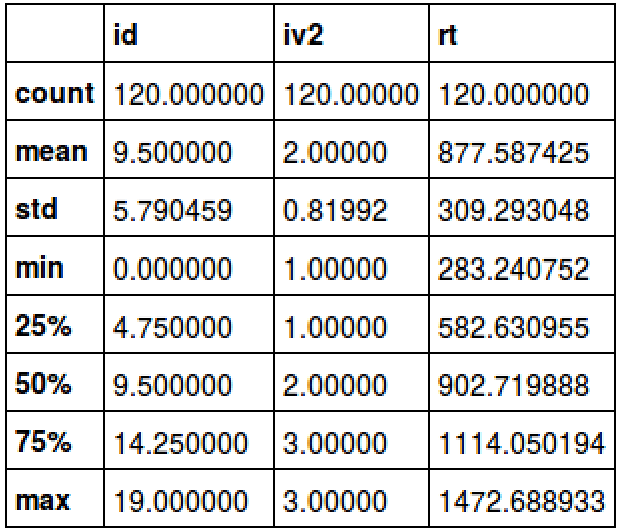
Return the first *n* rows



	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
2	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0

pandas.DataFrame.describe()

Generate descriptive statistics

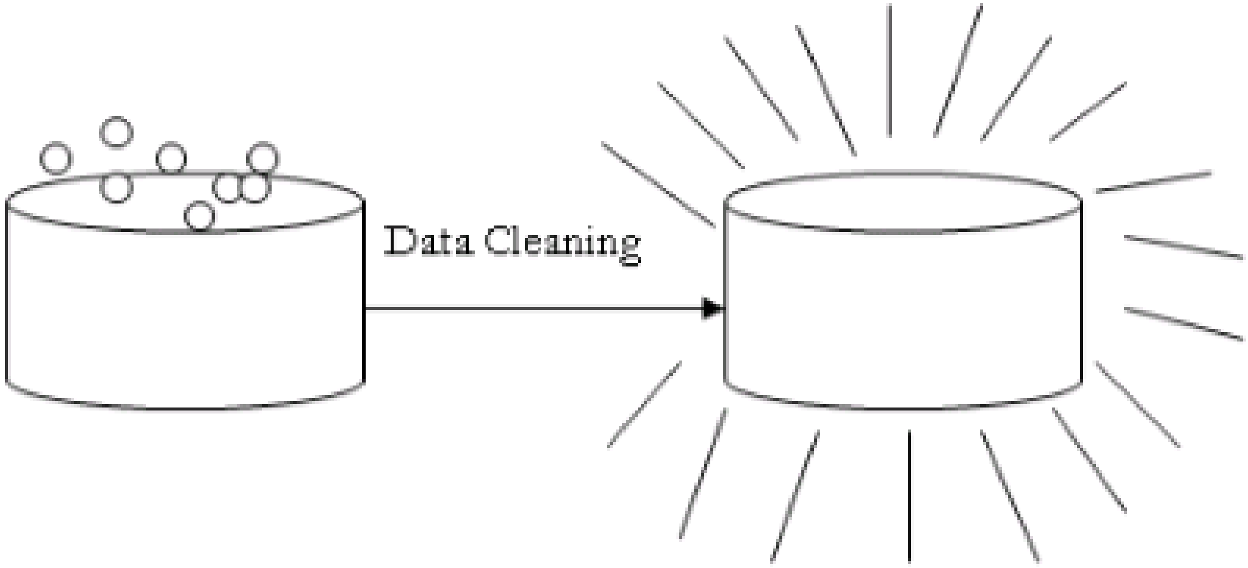


	id	iv2	rt
count	120.000000	120.00000	120.000000
mean	9.500000	2.00000	877.587425
std	5.790459	0.81992	309.293048
min	0.000000	1.00000	283.240752
25%	4.750000	1.00000	582.630955
50%	9.500000	2.00000	902.719888
75%	14.250000	3.00000	1114.050194
max	19.000000	3.00000	1472.688933

Why data preprocessing

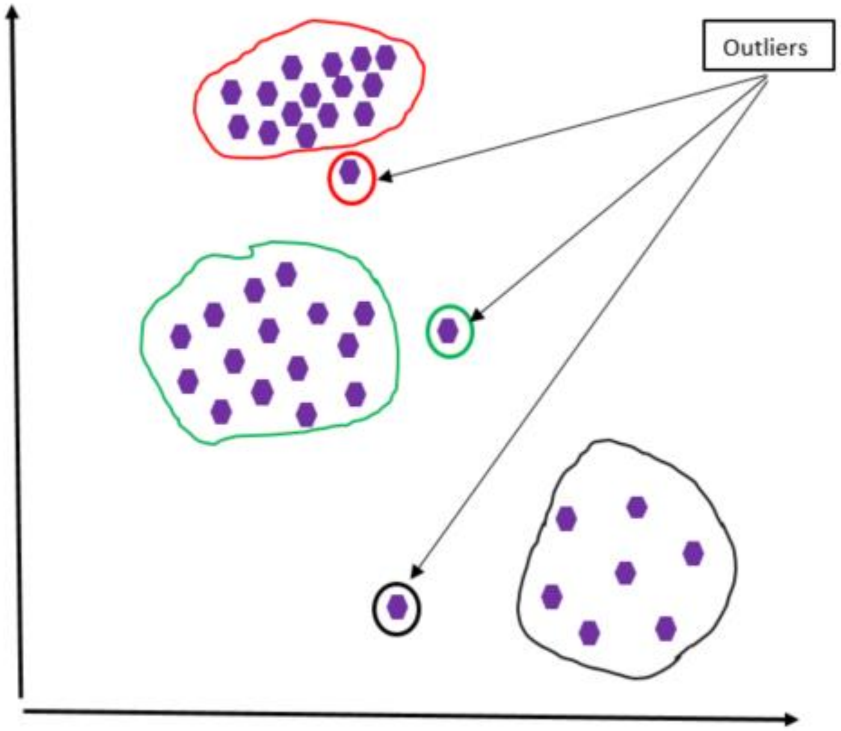
Data may be

- Incomplete (missing values)
- Noisy (containing errors or outliers)
- Inconsistent (containing discrepancies in dates, names, rates)

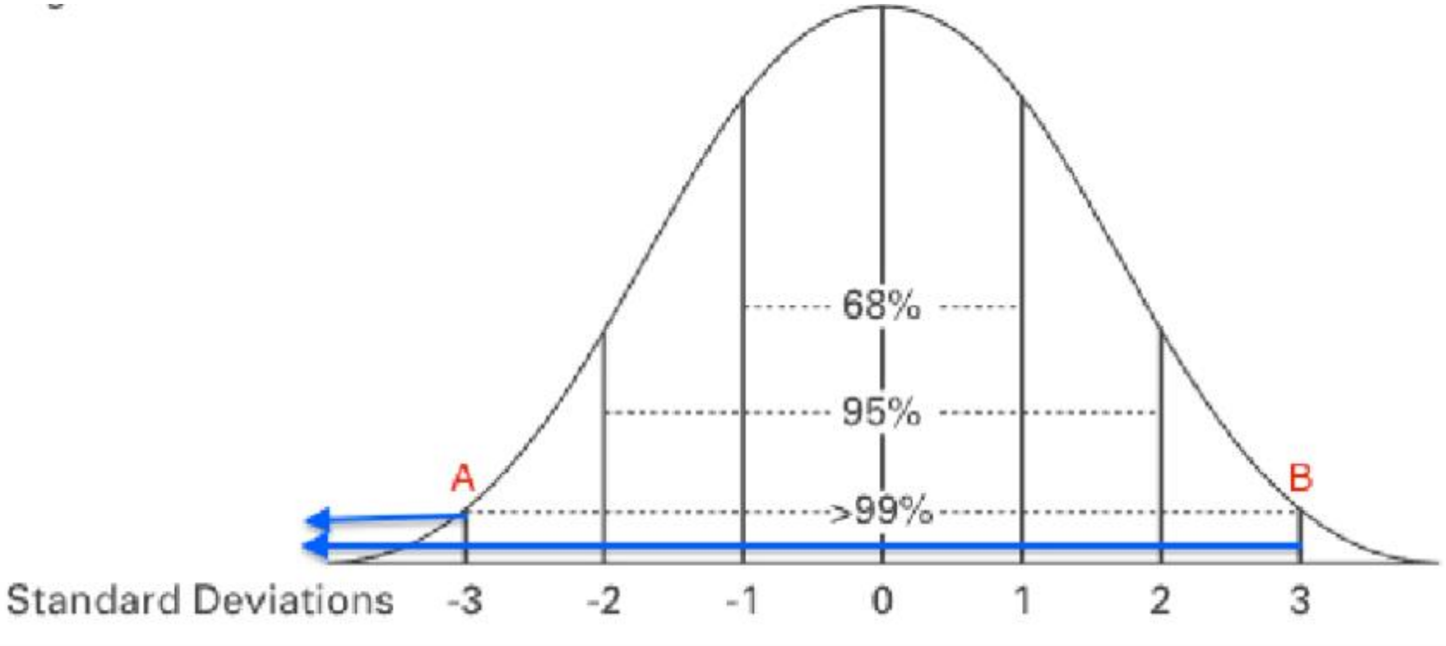


Data cleaning

Outliers detection



Three sigma rule



Normally data with $|x-\mu|>3\sigma$ are considered as outliers

Data cleaning

Missing data

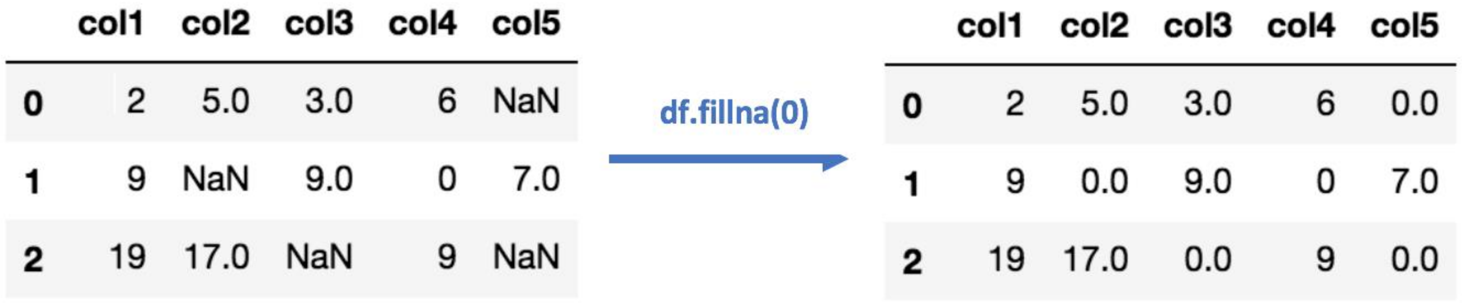
- may be deleted
- may be filled by:
  - ✓ the attribute mean
  - ✓ the attribute mean for all samples belonging to the same class
  - ✓ or other

`pandas.DataFrame.dropna()`

Remove missing values.

`pandas.DataFrame.fillna()`

Fill missing values



	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

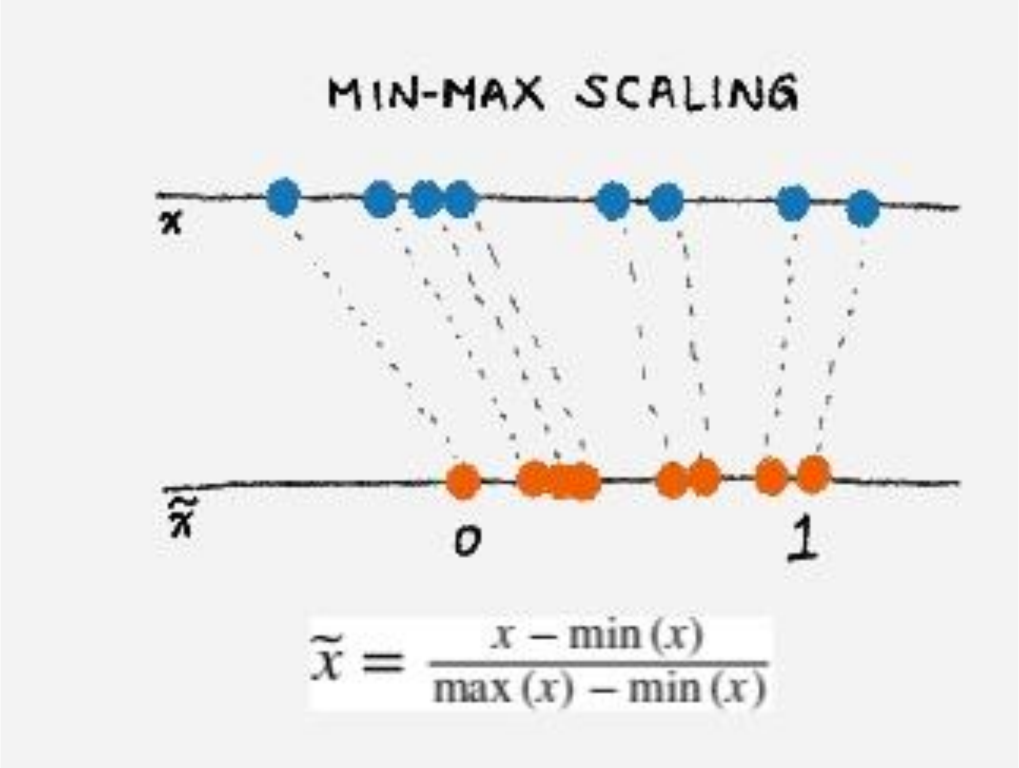
df.fillna(0)

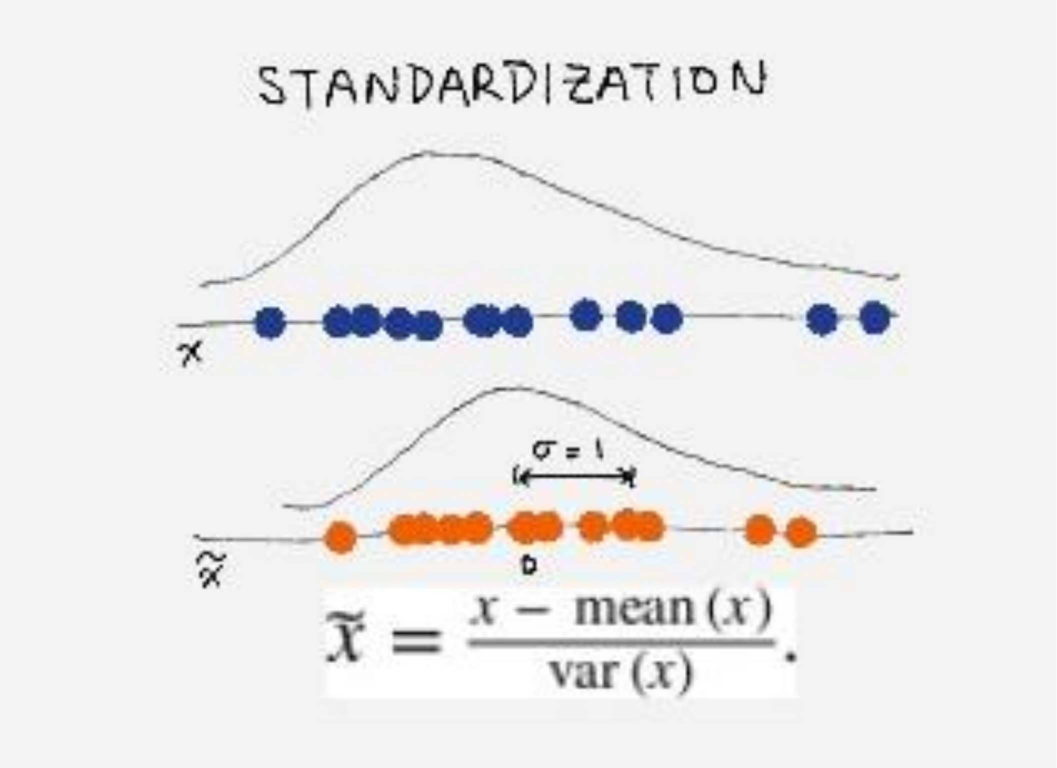
	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

Data Normalization

After min-max scaling, all feature values are within the [0, 1] range

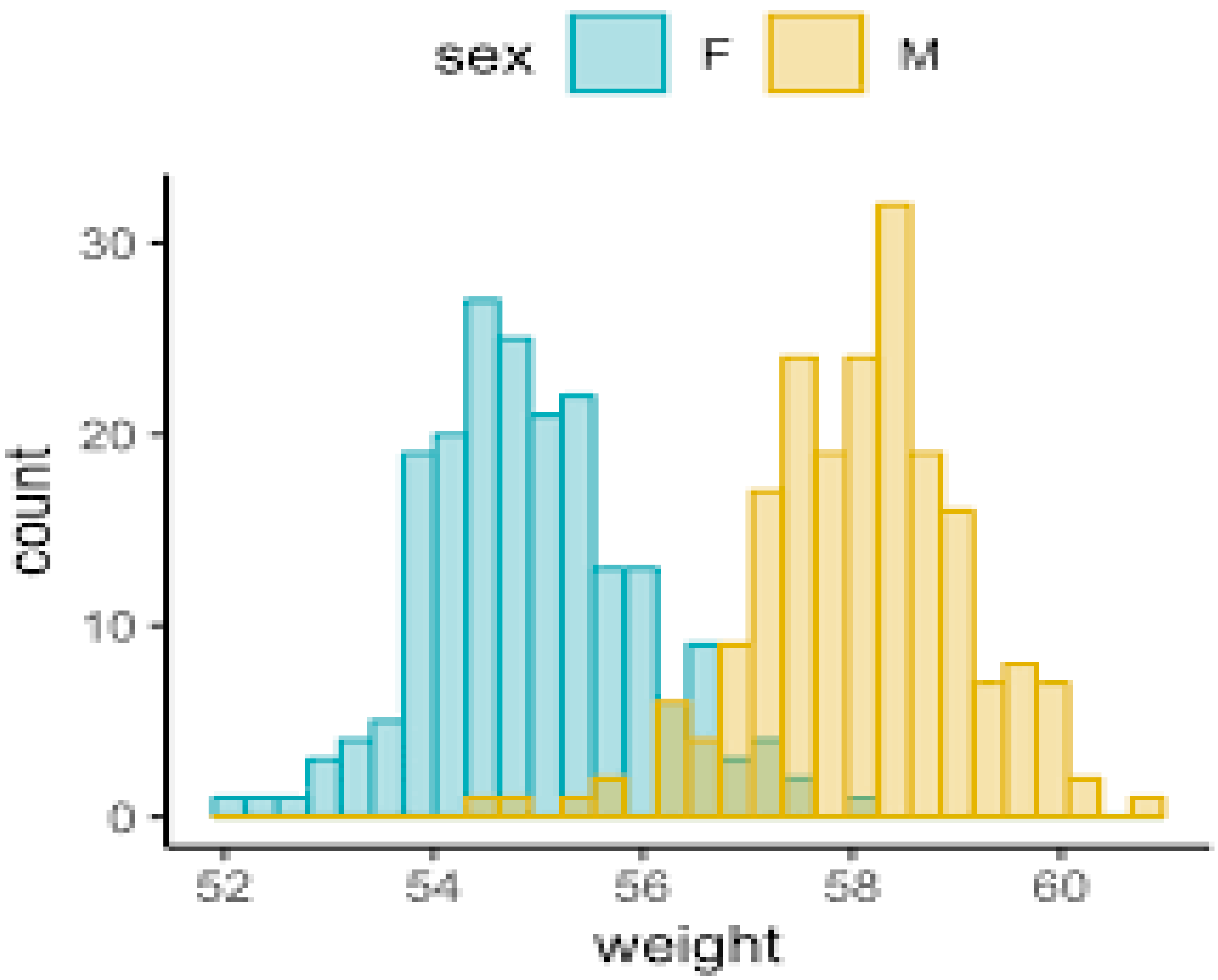
After standardization, a feature has mean 0 and variance 1





https://www.slideshare.net/gabrielspmoreira/feature-engineering-getting-most-out-of-data-for-predictive-models

Data Visualization: Histogram



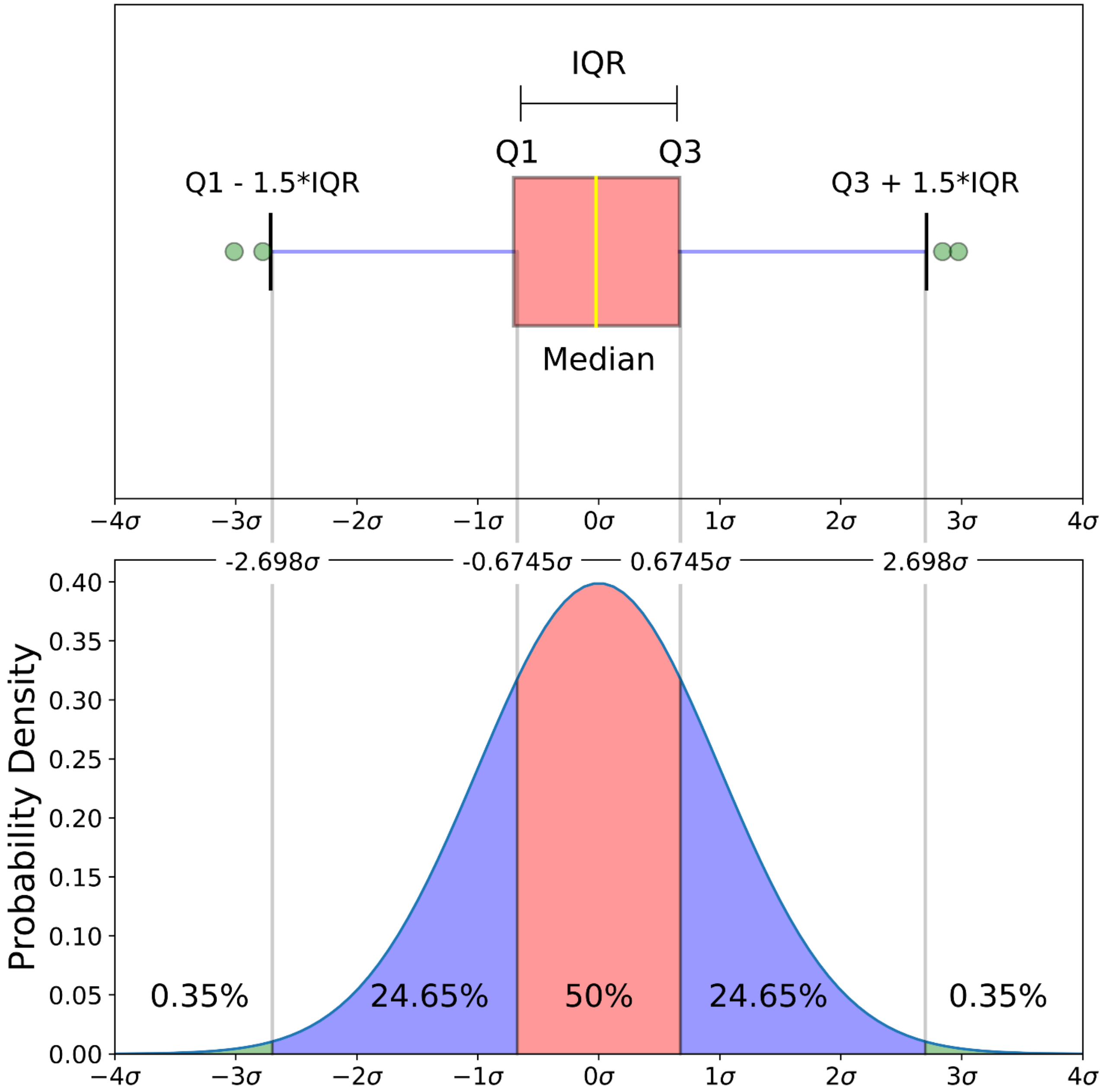
A **histogram** is an estimate of the probability distribution of a continuous variables.

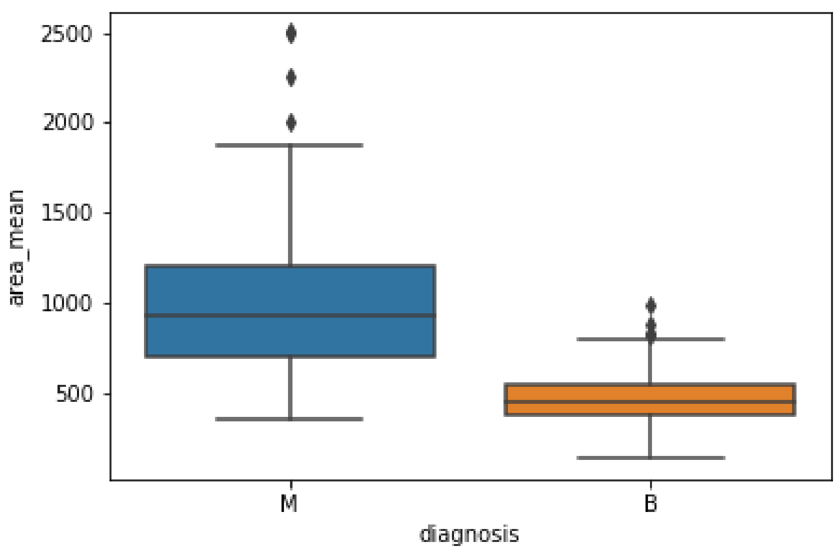
To construct a histogram:

- bin the range of values;
- plot a rectangle over each bin with height proportional to the frequency

seaborn.distplot()

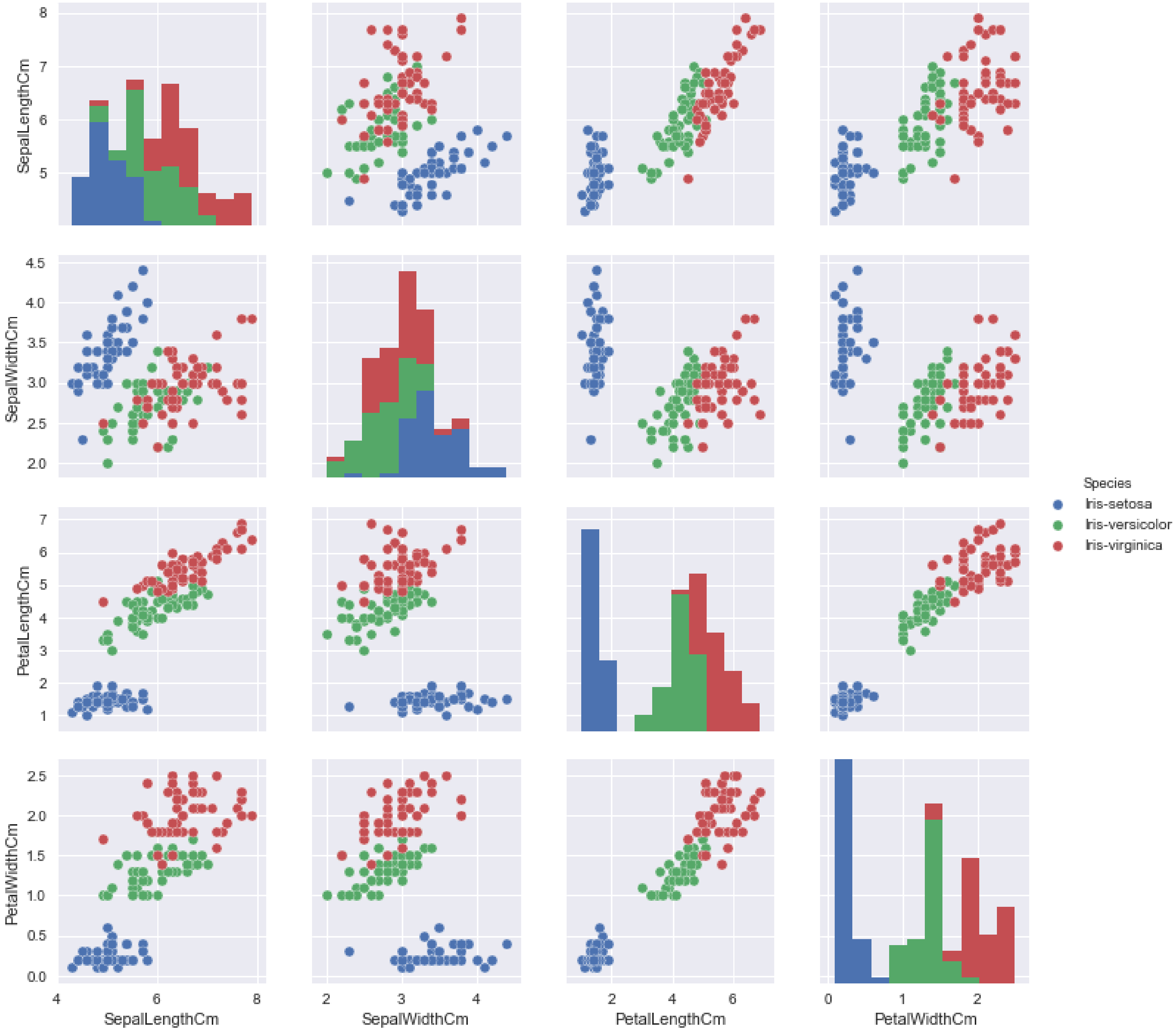
Data Visualization: Box Plot





A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

seaborn.boxplot()



Iris dataset

Plot pairwise relationships in a dataset

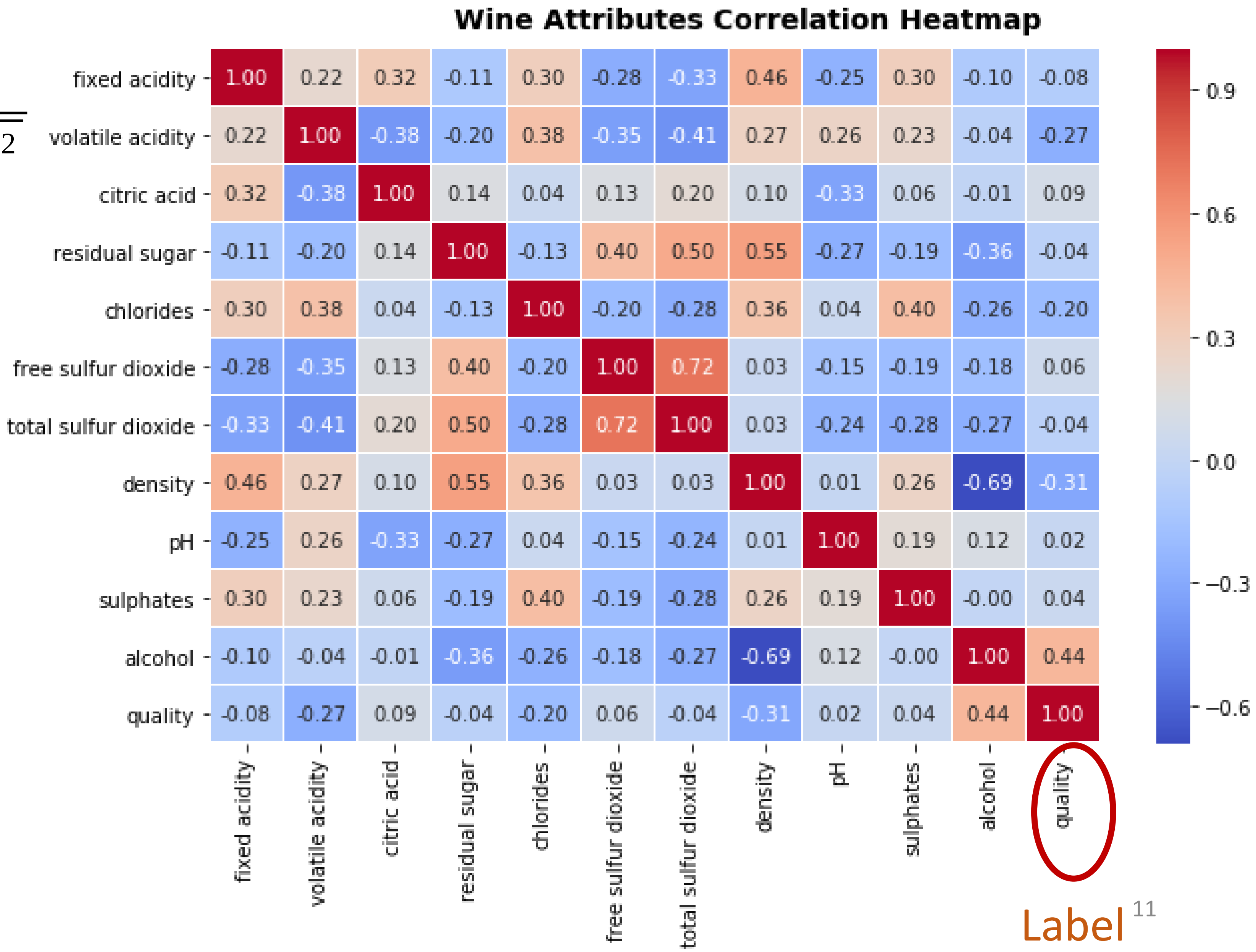
seaborn.pairplot()

Pearson correlation coefficient

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

Pearson's correlation coefficient is
the covariance of the two variables divided by the product of their standard deviations

seaborn.heatmap()



Wine Attributes Correlation Heatmap

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	0.22	0.32	-0.11	0.30	-0.28	-0.33	0.46	-0.25	0.30	-0.10	-0.08
volatile acidity	0.22	1.00	-0.38	-0.20	0.38	-0.35	-0.41	0.27	0.26	0.23	-0.04	-0.27
citric acid	0.32	-0.38	1.00	0.14	0.04	0.13	0.20	0.10	-0.33	0.06	-0.01	0.09
residual sugar	-0.11	-0.20	0.14	1.00	-0.13	0.40	0.50	0.55	-0.27	-0.19	-0.36	-0.04
chlorides	0.30	0.38	0.04	-0.13	1.00	-0.20	-0.28	0.36	0.04	0.40	-0.26	-0.20
free sulfur dioxide	-0.28	-0.35	0.13	0.40	-0.20	1.00	0.72	0.03	-0.15	-0.19	-0.18	0.06
total sulfur dioxide	-0.33	-0.41	0.20	0.50	-0.28	0.72	1.00	0.03	-0.24	-0.28	-0.27	-0.04
density	0.46	0.27	0.10	0.55	0.36	0.03	0.03	1.00	0.01	0.26	-0.69	-0.31
pH	-0.25	0.26	-0.33	-0.27	0.04	-0.15	-0.24	0.01	1.00	0.19	0.12	0.02
sulphates	0.30	0.23	0.06	-0.19	0.40	-0.19	-0.28	0.26	0.19	1.00	-0.00	0.04
alcohol	-0.10	-0.04	-0.01	-0.36	-0.26	-0.18	-0.27	-0.69	0.12	-0.00	1.00	0.44
quality	-0.08	-0.27	0.09	-0.04	-0.20	0.06	-0.04	-0.31	0.02	0.04	0.44	1.00

Label