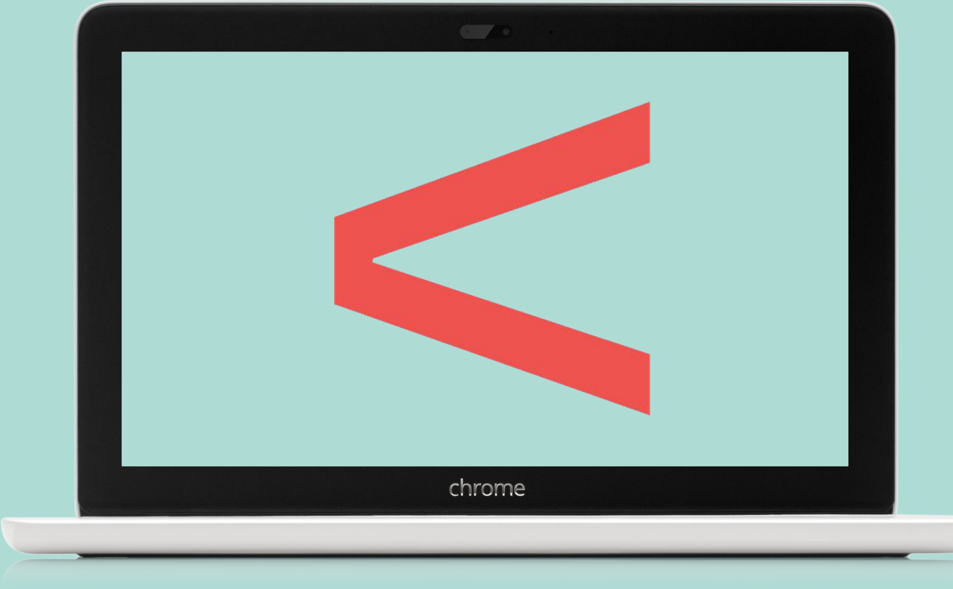
Introduction to Exploratory Data Analysis (EDA)



Exploratory Data Analysis
What is it?

EDA is an important piece of the Machine Learning puzzle.

During the EDA phase, we ‘explore’ our dataset with the goal of discovering patterns or trends, to identify outliers and test a hypothesis.

Statistics and visualisations are important tools in this initial analysis.

Main components of EDA

1. Understand the data and variables
2. Cleaning your dataset
3. Identify data patterns and correlations
4. Create new features or filter out unnecessary features (feature engineering)
5. Testing hypotheses

What is the data telling us about itself and regarding the problem we're trying to solve?

Important insights from EDA

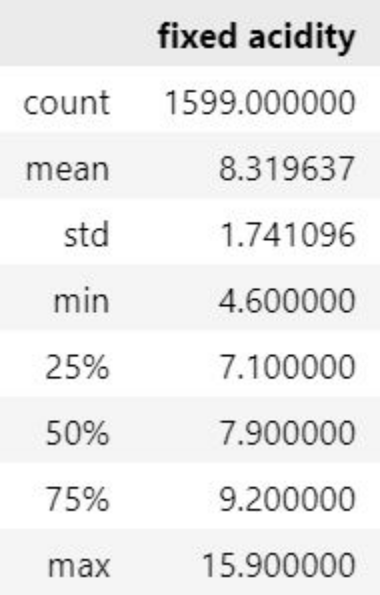
Some of the most important insights you'll find during this stage are:

1. Understand our data
2. Identify data patterns
3. Better understanding of the problem statement we're trying to solve
4. Filter out unnecessary features
5. Create new features
6. Testing hypotheses

What is the data telling us about itself and regarding the problem we're trying to solve?

Different types of EDA (examples)

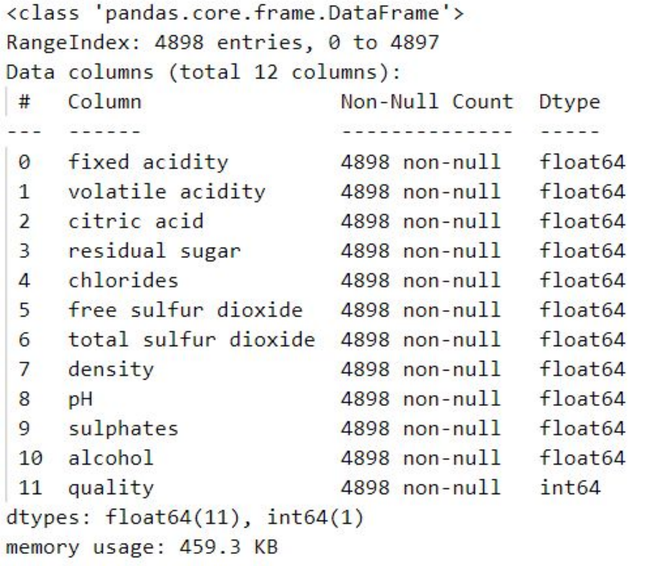
1. Univariate non-graphical: single dataset



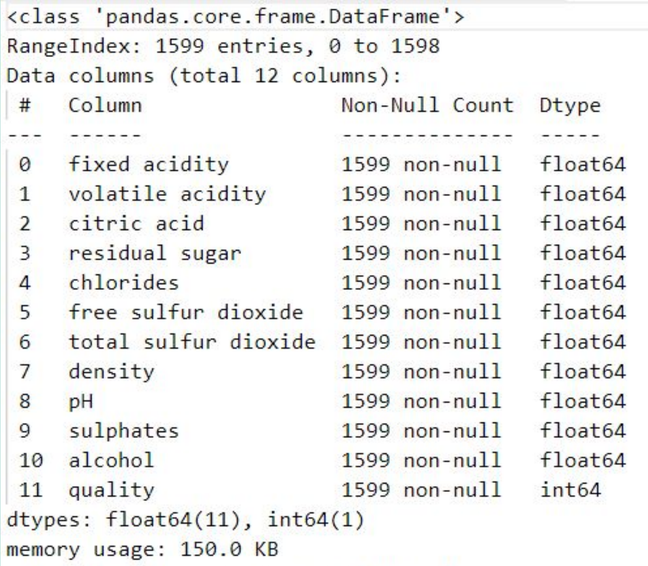
	fixed acidity
count	1599.000000
mean	8.319637
std	1.741096
min	4.600000
25%	7.100000
50%	7.900000
75%	9.200000
max	15.900000

Different types of EDA (examples)

1. Univariate non-graphical: compare the two datasets



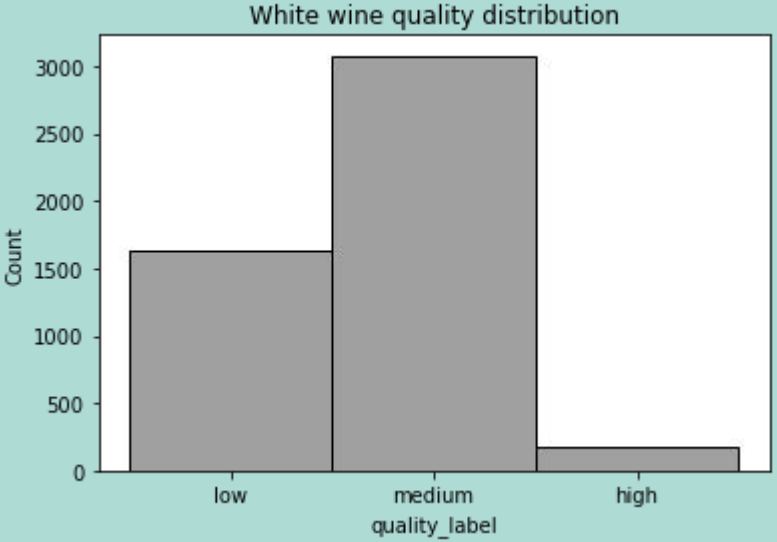
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         4898 non-null   float64
 1   volatile acidity      4898 non-null   float64
 2   citric acid           4898 non-null   float64
 3   residual sugar        4898 non-null   float64
 4   chlorides             4898 non-null   float64
 5   free sulfur dioxide   4898 non-null   float64
 6   total sulfur dioxide  4898 non-null   float64
 7   density               4898 non-null   float64
 8   pH                    4898 non-null   float64
 9   sulphates             4898 non-null   float64
 10  alcohol               4898 non-null   float64
 11  quality               4898 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

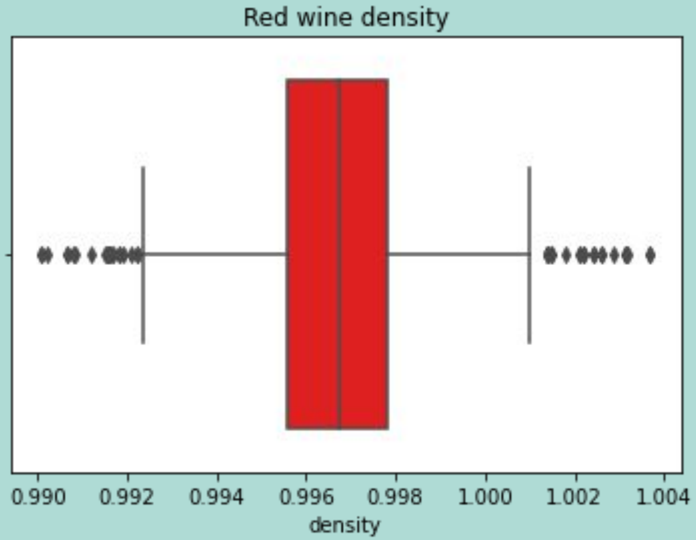
Different types of EDA (examples)

1. Univariate graphical: single dataset



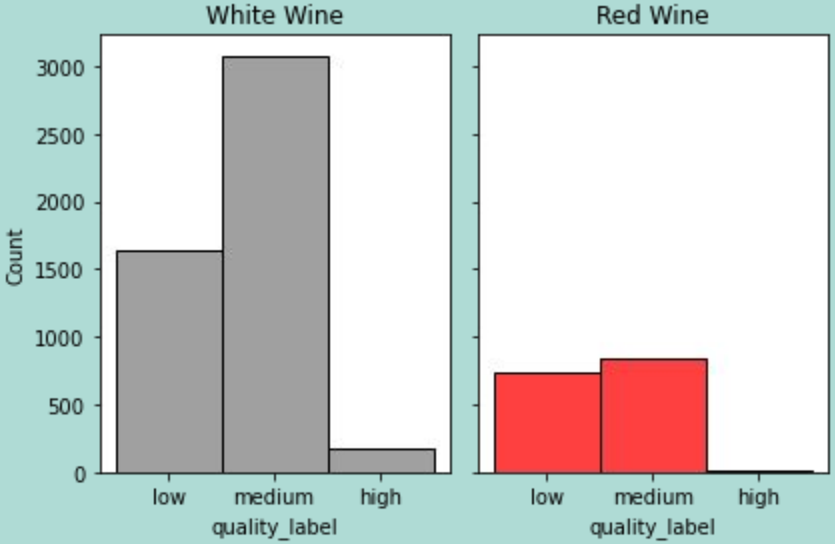
Different types of EDA (examples)

2. Univariate graphical: single dataset



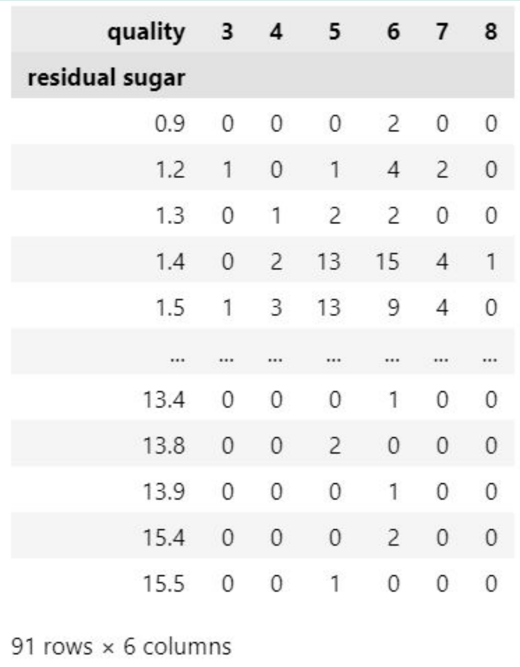
Different types of EDA (examples)

2. Univariate graphical: compare two datasets



Different types of EDA (examples)

3. Multivariate non-graphical (frequency table with cross-tabulation)

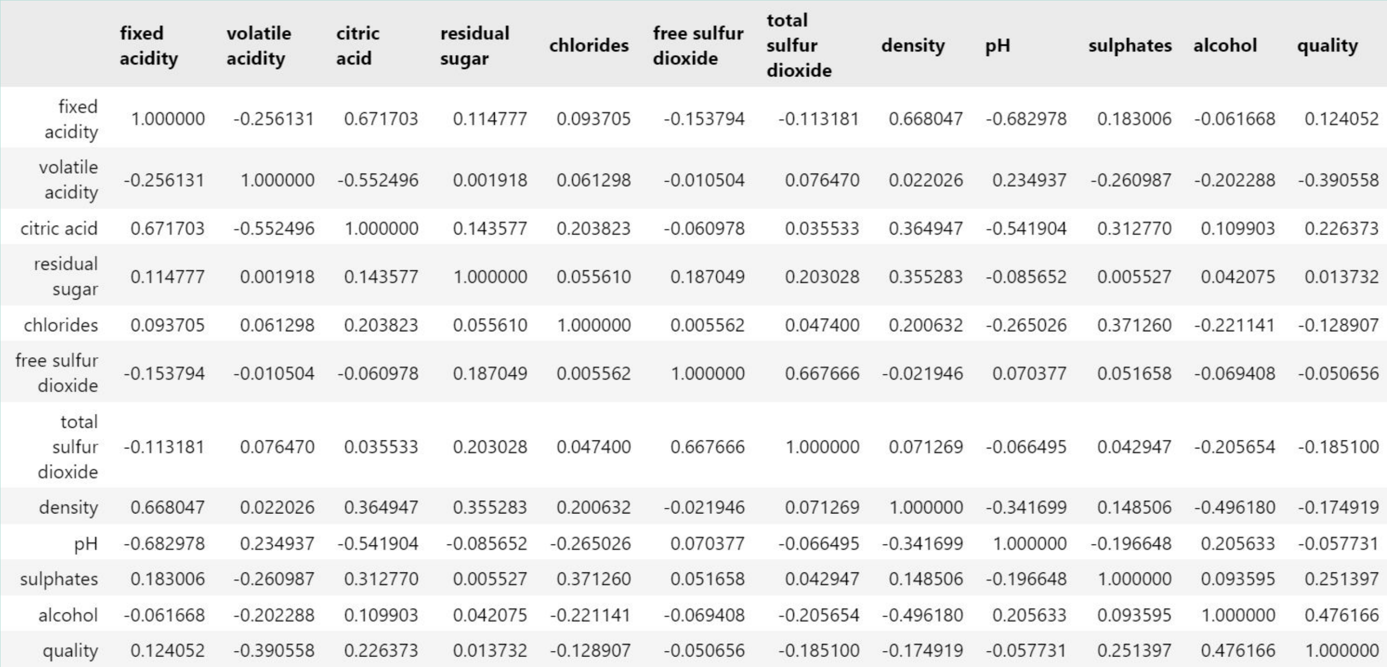


quality	3	4	5	6	7	8
residual sugar						
0.9	0	0	0	2	0	0
1.2	1	0	1	4	2	0
1.3	0	1	2	2	0	0
1.4	0	2	13	15	4	1
1.5	1	3	13	9	4	0
...	...	...	...	...	...	...
13.4	0	0	0	1	0	0
13.8	0	0	2	0	0	0
13.9	0	0	0	1	0	0
15.4	0	0	0	2	0	0
15.5	0	0	1	0	0	0

91 rows × 6 columns

Different types of EDA (examples)

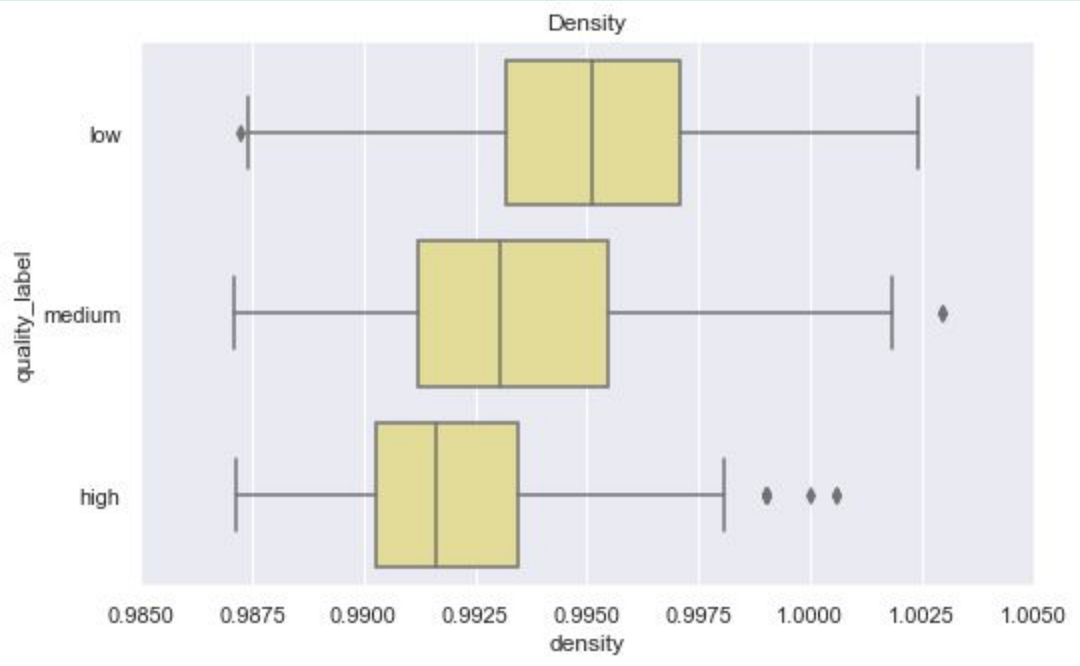
4. Multivariate non-graphical (correlation matrix)



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.000000	-0.256131	0.671703	0.114777	0.093705	-0.153794	-0.113181	0.668047	-0.682978	0.183006	-0.061668	0.124052
volatile acidity	-0.256131	1.000000	-0.552496	0.001918	0.061298	-0.010504	0.076470	0.022026	0.234937	-0.260987	-0.202288	-0.390558
citric acid	0.671703	-0.552496	1.000000	0.143577	0.203823	-0.060978	0.035533	0.364947	-0.541904	0.312770	0.109903	0.226373
residual sugar	0.114777	0.001918	0.143577	1.000000	0.055610	0.187049	0.203028	0.355283	-0.085652	0.005527	0.042075	0.013732
chlorides	0.093705	0.061298	0.203823	0.055610	1.000000	0.005562	0.047400	0.200632	-0.265026	0.371260	-0.221141	-0.128907
free sulfur dioxide	-0.153794	-0.010504	-0.060978	0.187049	0.005562	1.000000	0.667666	-0.021946	0.070377	0.051658	-0.069408	-0.050656
total sulfur dioxide	-0.113181	0.076470	0.035533	0.203028	0.047400	0.667666	1.000000	0.071269	-0.066495	0.042947	-0.205654	-0.185100
density	0.668047	0.022026	0.364947	0.355283	0.200632	-0.021946	0.071269	1.000000	-0.341699	0.148506	-0.496180	-0.174919
pH	-0.682978	0.234937	-0.541904	-0.085652	-0.265026	0.070377	-0.066495	-0.341699	1.000000	-0.196648	0.205633	-0.057731
sulphates	0.183006	-0.260987	0.312770	0.005527	0.371260	0.051658	0.042947	0.148506	-0.196648	1.000000	0.093595	0.251397
alcohol	-0.061668	-0.202288	0.109903	0.042075	-0.221141	-0.069408	-0.205654	-0.496180	0.205633	0.093595	1.000000	0.476166
quality	0.124052	-0.390558	0.226373	0.013732	-0.128907	-0.050656	-0.185100	-0.174919	-0.057731	0.251397	0.476166	1.000000

Different types of EDA (examples)

5. Multivariate graphical: one dataset



Resources:

- National Institute of Standards and Technology’s handbook with a chapter dedicated to the topic of EDA.

- Howard Seltman’s ‘Experimental Design and Analysis’, chapter 4.