TL;DR OF EDA

Mor Hananovitz
PyData Israel 2024

About Me

Head of Data & Data Scientist

PARAZERO
DRONE SAFETY SYSTEMS

Senior Lecturer | Huji, Data Analytics

WiDS community manager

Snowboarder

CrossFitter

Agenda

01
Whats EDA?

02
Classic method

03
Ydata profiling

04
Clustering

05
LLM

What is Exploratory Data Analysis?

- Scope
- Issues
- Characteristics

Data Playground

Correlations

Detecting anomalies

Data characteristics

Distributions

Parameters

Segments

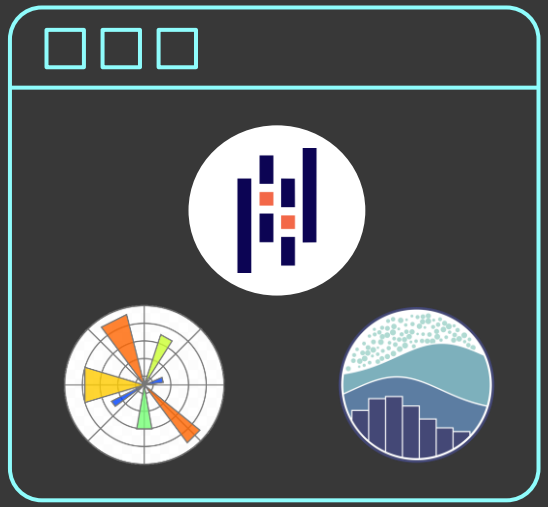
The Dataset

- UCI Wine Data set
- 13 chemical features + class column

Alcohol	Alcalinity of ash	Flavanoids	Color intensity	Proline
Malic acid	Magnesium	Nonflavanoid phenols	Hue	Class
Ash	Total phenols	Proantho cyanins	OD280/OD315 of diluted wines	

Manual

EDA



```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

df.info()

```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Alcohol                 178 non-null    float64
 1   Malicacid               178 non-null    float64
 2   Ash                     178 non-null    float64
 3   Alcalinity_of_ash       178 non-null    float64
 4   Magnesium               178 non-null    int64
 5   Total_phenols           178 non-null    float64
 6   Flavanoids              178 non-null    float64
```

df.describe(include = ‘all’)

	count	mean	std	min	25%
Alcohol	178.0	13.000618	0.811827	11.03	12.3625
Malicacid	178.0	2.336348	1.117146	0.74	1.6025
Ash	178.0	2.366517	0.274344	1.36	2.2100
Alcalinity_of_ash	178.0	19.494944	3.339564	10.60	17.2000
Magnesium	178.0	99.741573	14.282484	70.00	88.0000
Total_phenols	178.0	2.295112	0.625851	0.98	1.7425

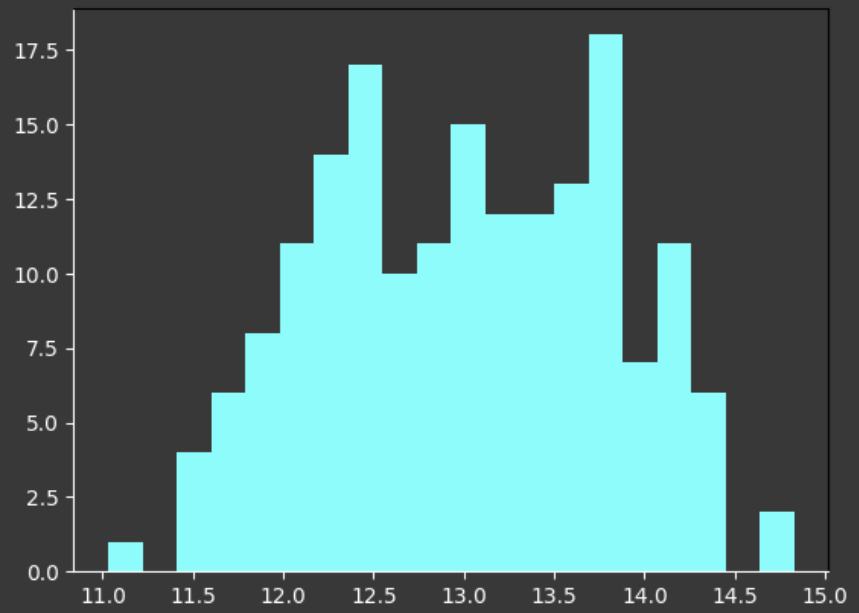
```
df['col_name'].value_counts()
```

class	count
2	71
1	59
3	48

Visualizations

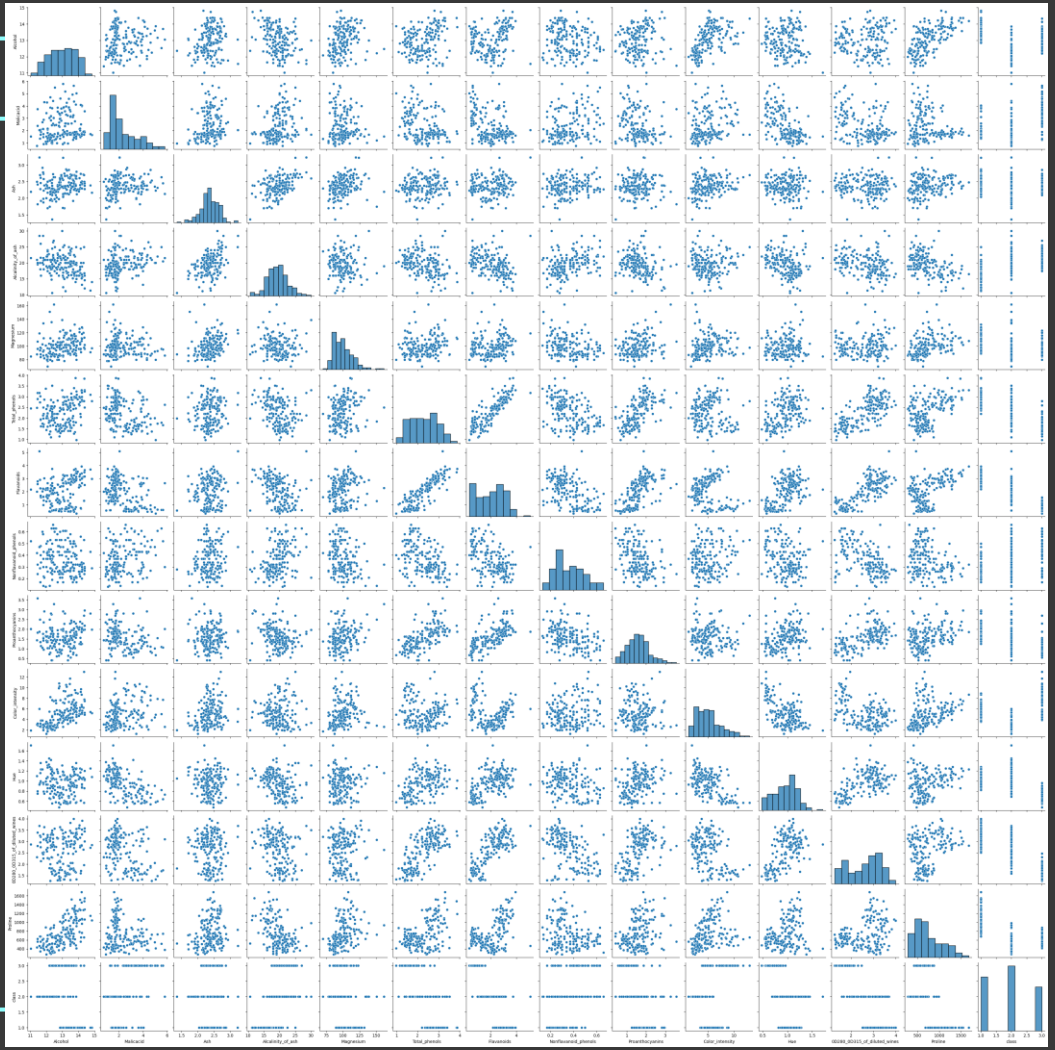
- Distributions
  - Histograms
  - Bar / line plots

```
df['Alcohol'].hist(bins = 20)
```



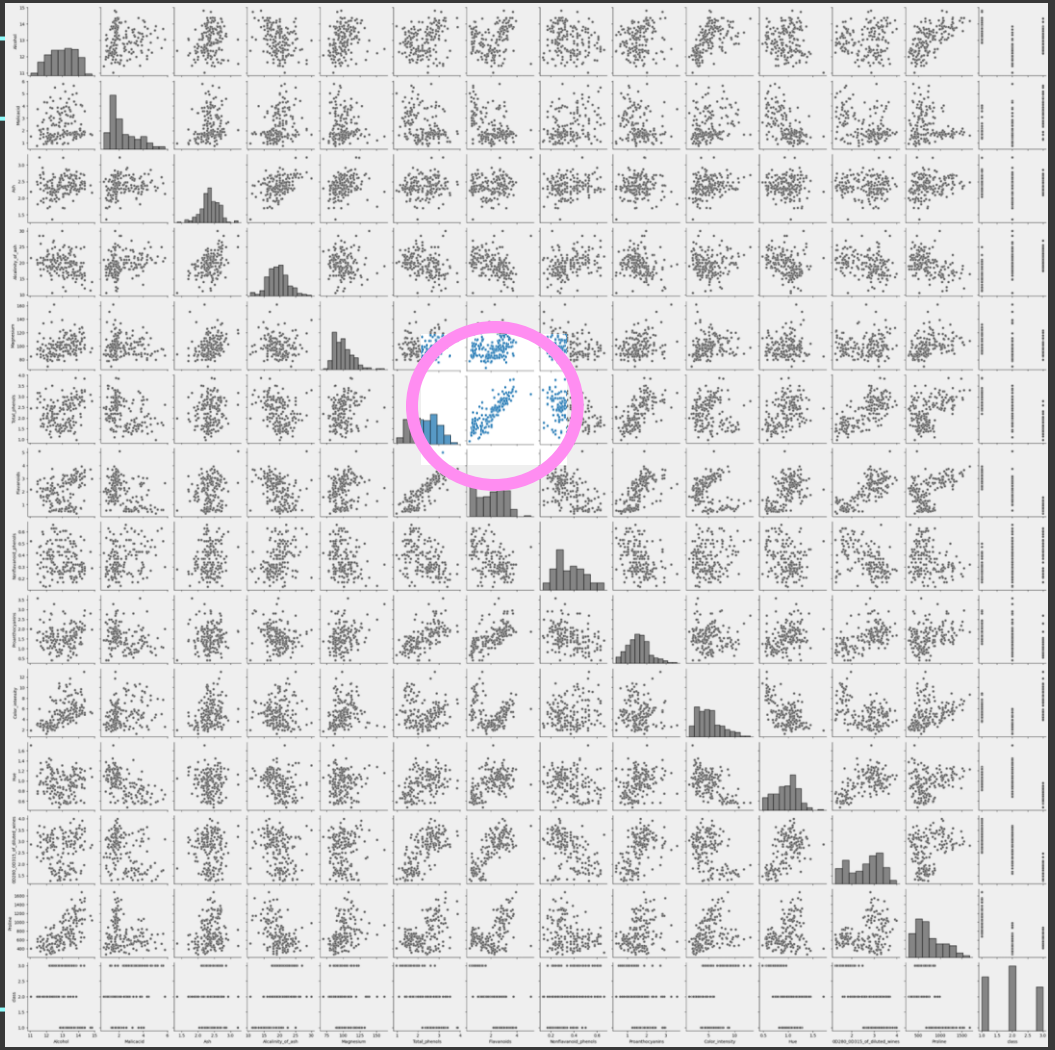
```
import
seaborn as
sns
```

```
[ ]  sns.pairplot(df)
```

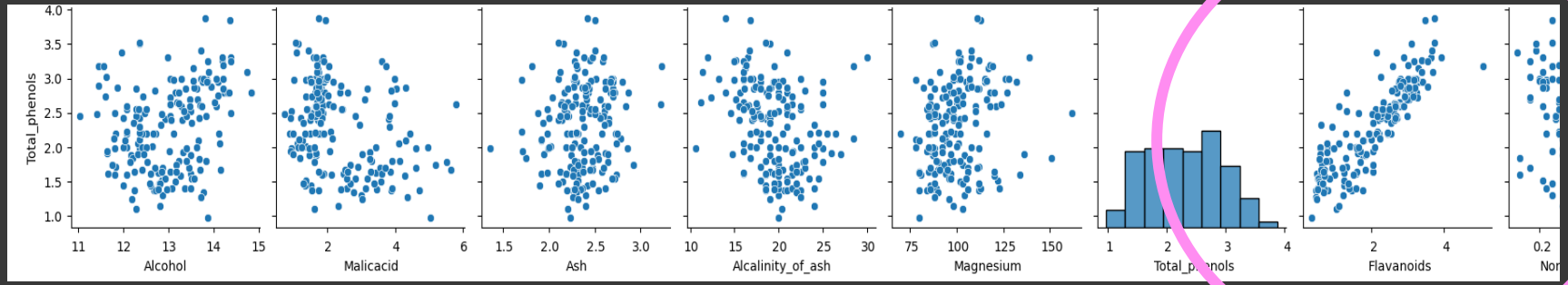


```
import
seaborn as
sns
```

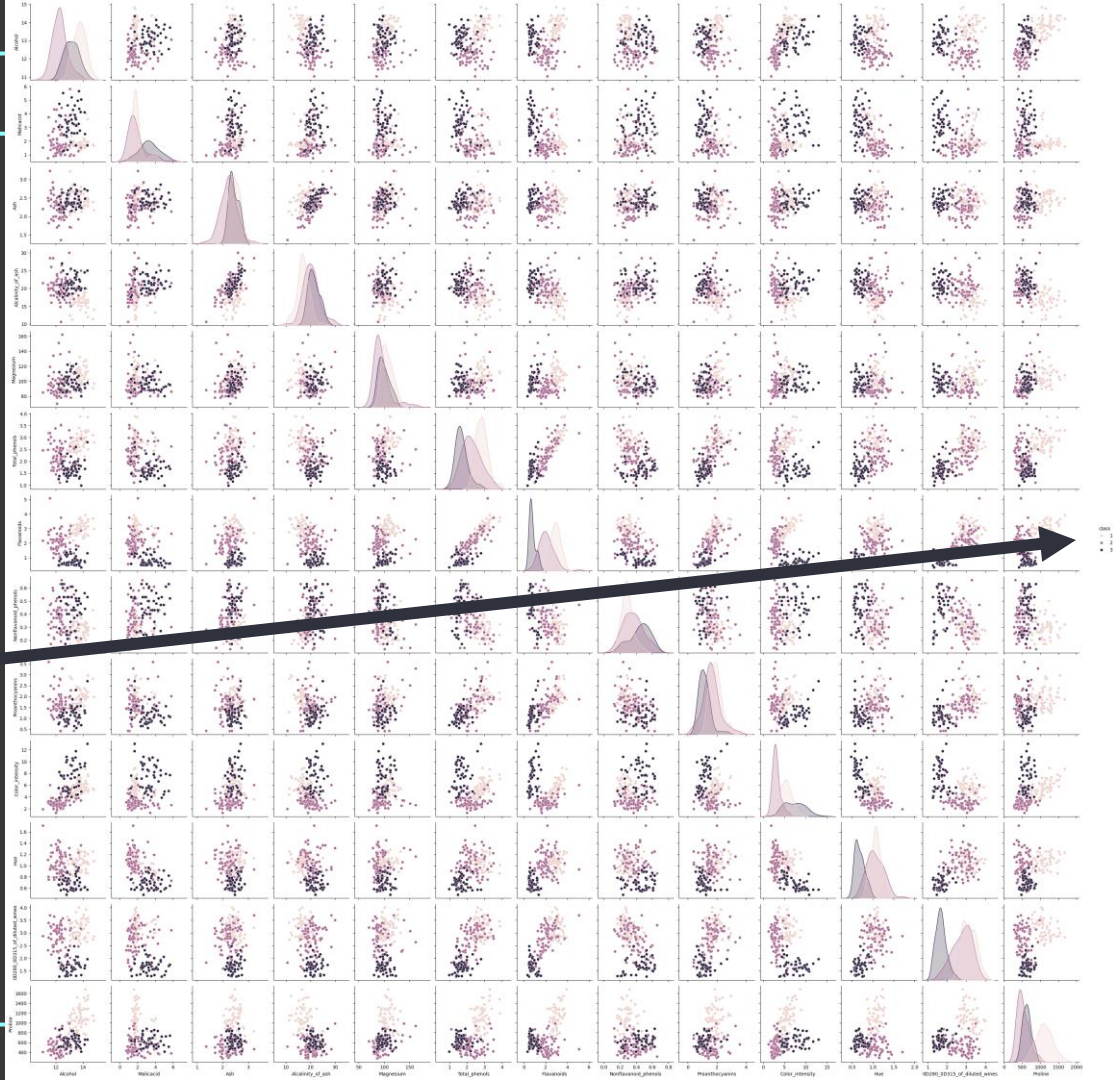
```
[ ]  sns.pairplot(df)
```

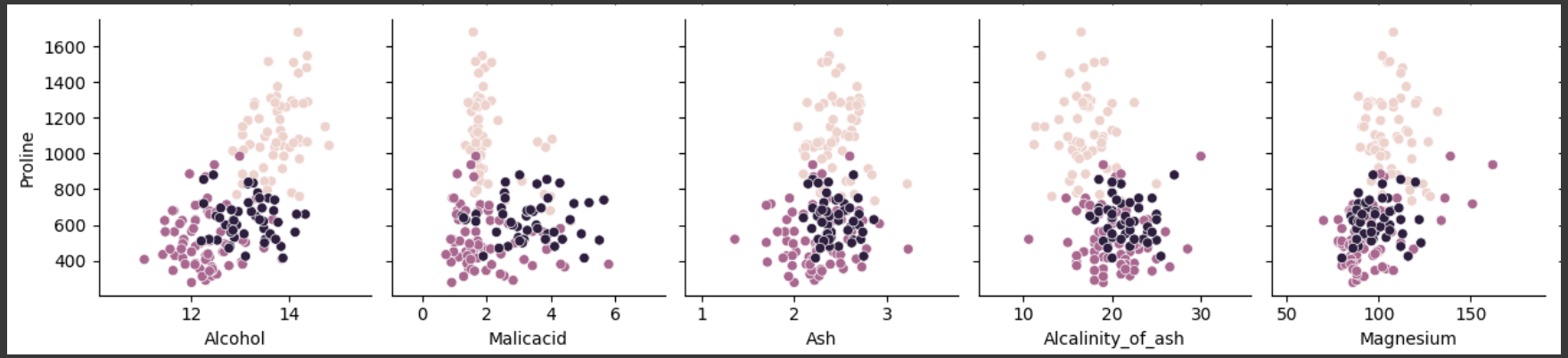


Pair Plot Analysis



```
sns.pairplot(df,
hue = 'class')
```




Proline
1600
1400
1200
1000
800
600
400
12
14
Alcohol
0
2
4
6
Malicacid
1
2
3
Ash
10
20
30
Alcalinity_of_ash
50
100
150
Magnesium

Transition to Automated EDA

- Low-code
- Fast development
- Consistent across projects
- Scalability (big data)

Automated

EDA Report

Ydata-Profiling

Generates an extensive EDA report in seconds:

- Statistics
- Correlations
- Missing values
- And more..

```
%%capture
! pip install -U ydata-profiling
```

```
from ydata_profiling import ProfileReport
```

```
ProfileReport(df, title='Profile Report')
```

Overview

Brought to you by YData

Overview | Alerts 12 | Reproduction

Dataset statistics

Number of variables	14
Number of observations	178
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	19.6 KiB
Average record size in memory	112.7 B

Variable types

Numeric	13
Categorical	1

Alerts

0D280_0D315_of_diluted_wines is highly overall correlated with Flavanoids and 3 other fields	High correlation
Alcohol is highly overall correlated with Color_intensity and 2 other fields	High correlation
Color_intensity is highly overall correlated with Alcohol and 1 other fields	High correlation
Flavanoids is highly overall correlated with 0D280_0D315_of_diluted_wines and 5 other fields	High correlation
Hue is highly overall correlated with Flavanoids and 2 other fields	High correlation
Magnesium is highly overall correlated with Proline	High correlation
Malicacid is highly overall correlated with Hue	High correlation
Nonflavanoid_phenols is highly overall correlated with Flavanoids	High correlation
Proanthocyanins is highly overall correlated with 0D280_0D315_of_diluted_wines and 2 other fields	High correlation
Proline is highly overall correlated with Alcohol and 2 other fields	High correlation

Variables

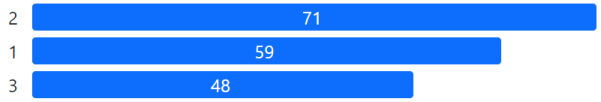
class

class

Categorical

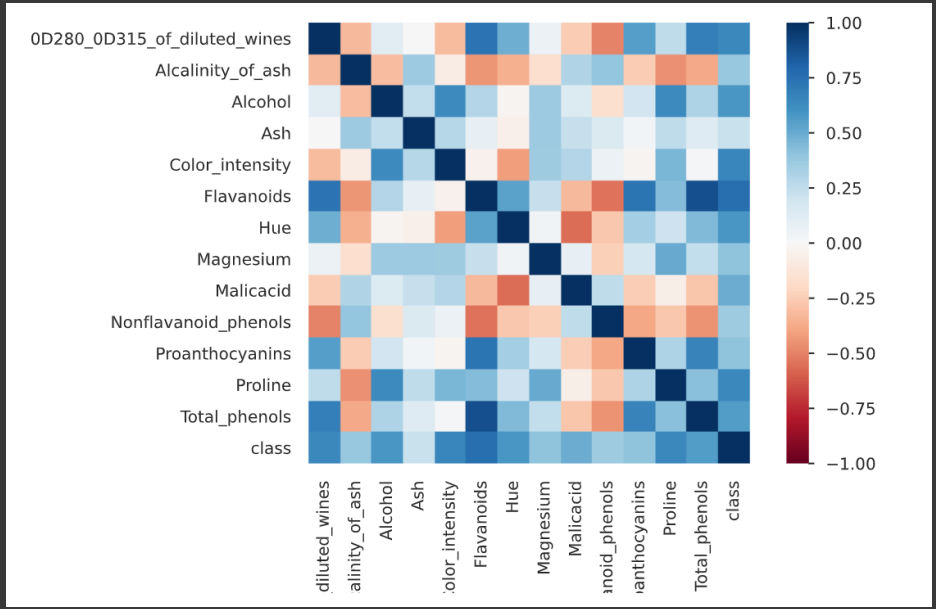
High correlation

Distinct	3
Distinct (%)	1.7%
Missing	0
Missing (%)	0.0%
Memory size	1.5 KiB

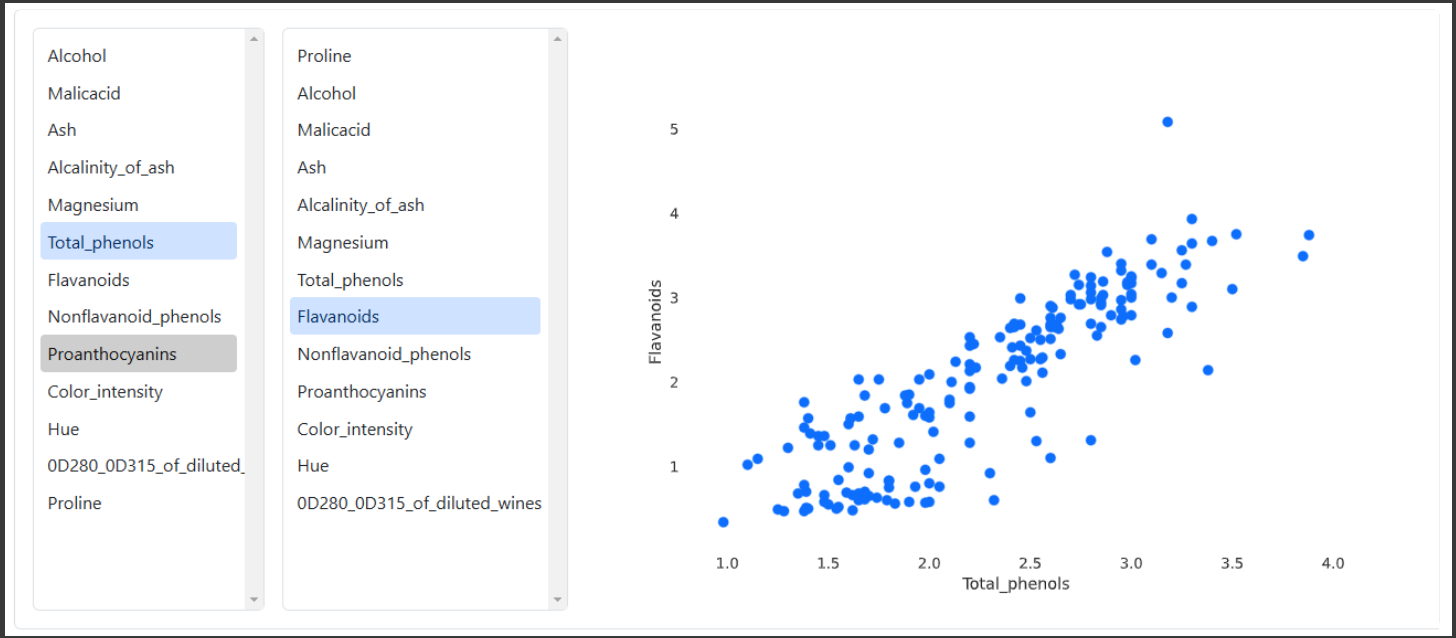


More details

Correlations



Interactions



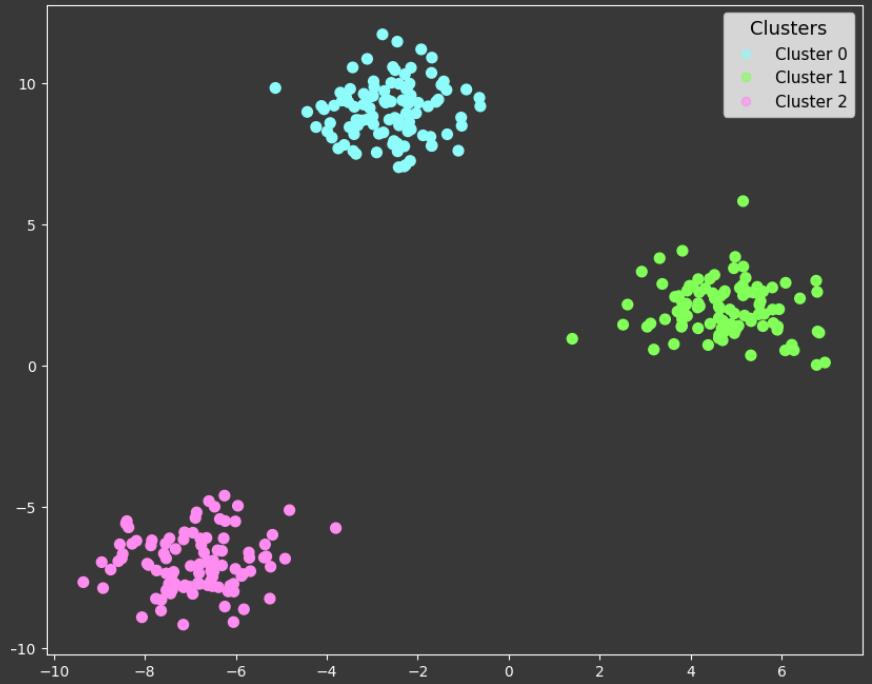
Ydata_profiling vs Pandas

	Ydata Profiling	Pandas
PROS	• Fast • Low-code • Intuitive	• Easy to change • Flexible
CONS	• Cannot be changed • Requires manual exploration	• Slow • Requires coding skills • Need to know what to look

Automated

EDA with ML

Clustering for EDA



Criteria for Clustering

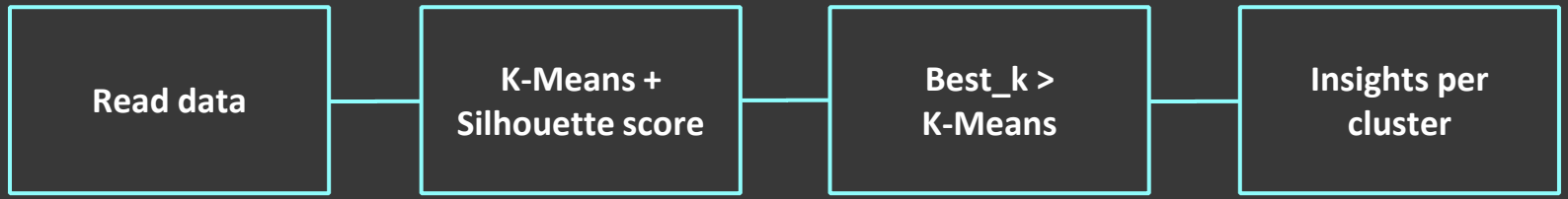
Labeling

Segments

Outliers

Applying Clustering for EDA

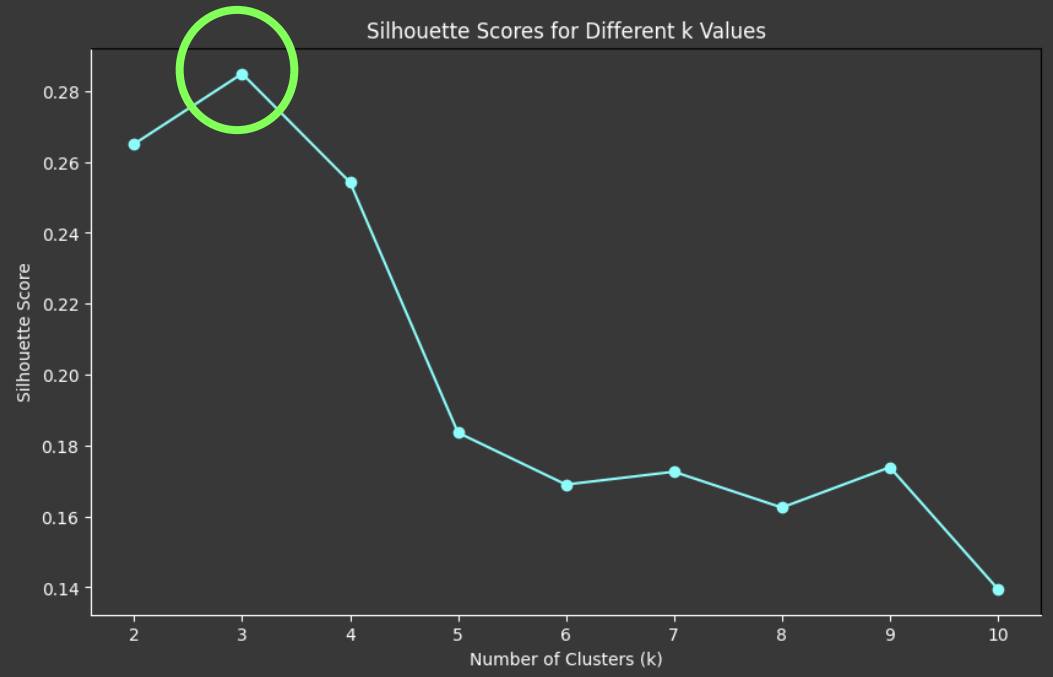
- Uncover segments within the data automatically



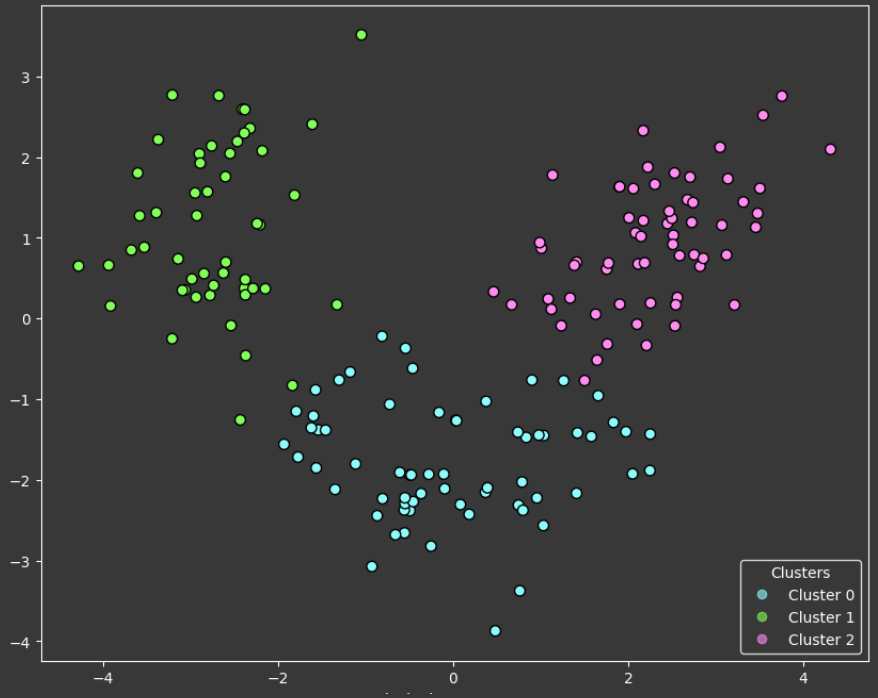
K-Means

```
[32] scaler = StandardScaler()
     data_scaled = scaler.fit_transform(df_cluster)

     range_k = range(2, 11)
     silhouette_scores = []

     for k in range_k:
         kmeans = KMeans(n_clusters=k, random_state=42)
         cluster_labels = kmeans.fit_predict(data_scaled)
         score = silhouette_score(data_scaled, cluster_labels)
         silhouette_scores.append(score)

     best_k = range_k[silhouette_scores.index(max(silhouette_scores))]
     best_score = max(silhouette_scores)
```

```
best_k =(max(silhouette_scores))
```



Cluster Visualization



Model Evaluation

Model Accuracy: 96.62%

```
[49] crosstab = pd.crosstab(df['class'], df['Cluster Labels'])

     crosstab
```

Cluster Labels	0	1	2
class			
1	0	0	59
2	65	3	3
3	0	48	0

Descriptive Statistics

Principal Component Analysis PCA

- Dimensionality reduction
- Explain variance
- 2D Visualization

PCA Explainability

pca.components_

	Component 1	Component 2
1	Flavanoids	Color_intensity
2	Total_phenols	Alcohol
3	0D280_0D315_of_diluted_wines	Proline

PCA Explainability

```
pca.explained_variance_ratio_
```

```
array([0.36198848, 0.1920749 ])
```

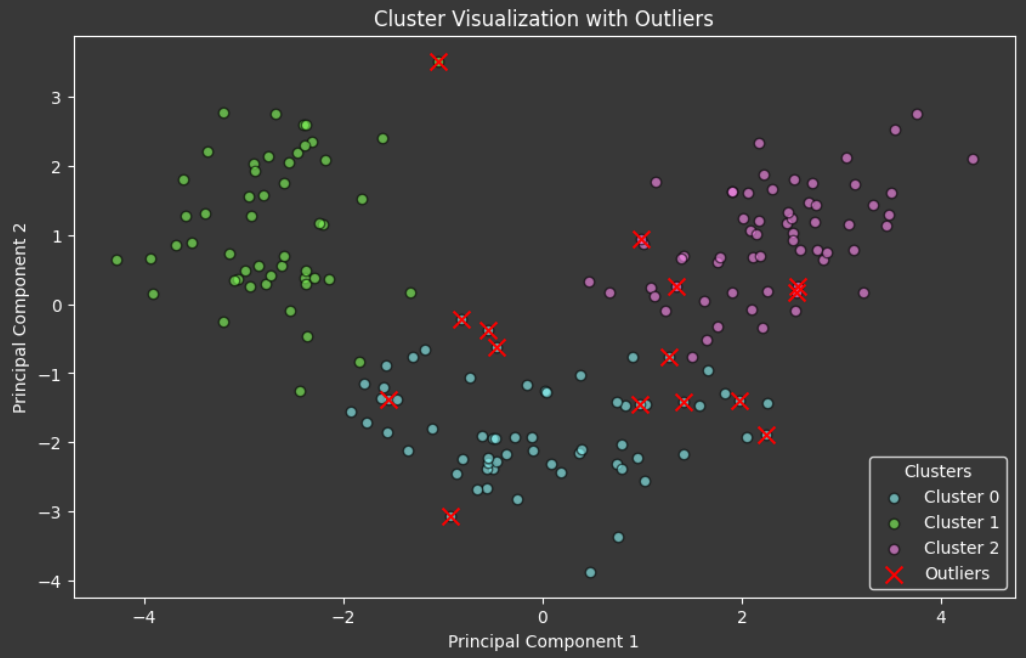
```
sum(pca.explained_variance_ratio_)
```

```
0.5540633835693527
```

Outliers Analysis

- Cluster distribution
- Z-Score

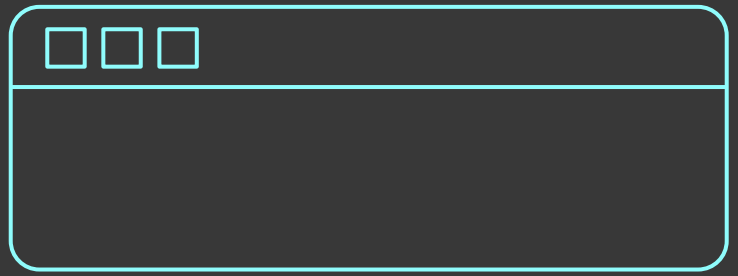
Cluster	Outlier [%]
0	15.384615
1	1.960784
2	6.451613



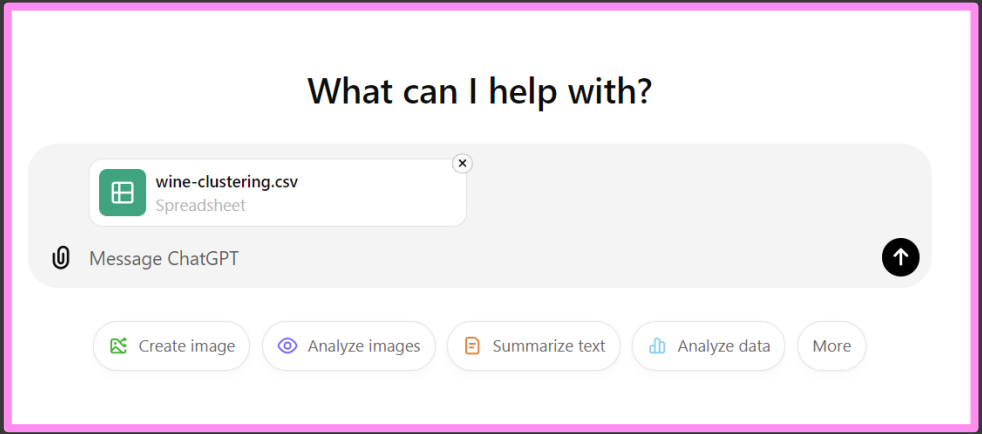
What are the outputs?

- Clusters
- T-test comparison
- Explainability:
  - Top 3 features
  - Confidence interval
- Outliers

But what does it all mean?



Using Large Language Models for EDA



Analyze Clusters

```
cluster_stats = df.groupby('Cluster Labels').describe()

cluster_stats.transpose()
```

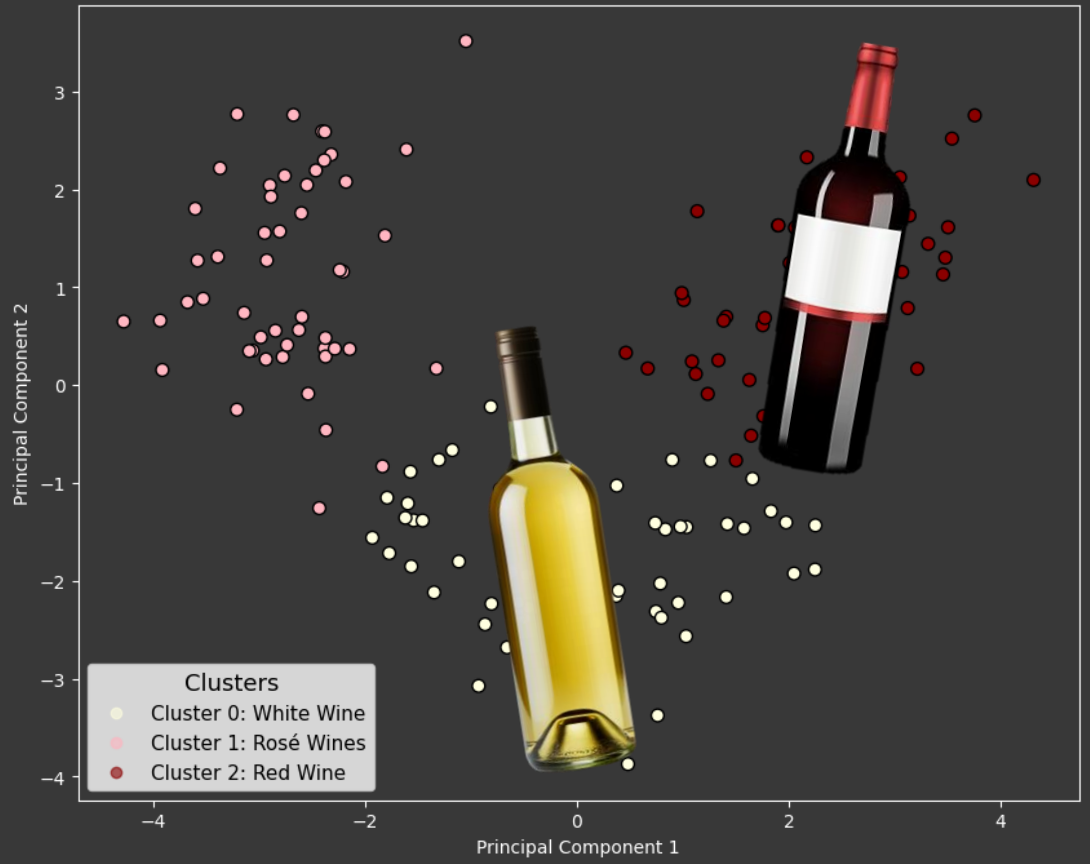
Analyze the results and provide classification of wine type per group

LLM-Based EDA

from the 3 clusters - can you try to guess which cluster is which type of wine?

High Color Intensity:

Likely indicates red wines due to deeper color from longer skin contact during fermentation.


Clusters
Cluster 0: White Wine
Cluster 1: Rosé Wines
Cluster 2: Red Wine
Principal Component 1
Principal Component 2
−4
−2
0
2
4
3
2
1
0
−1
−2
−3
−4

PCA Explainability

- pca.components_

	Component 1	Component 2
1	Flav	Color_int
2	phenols	ol
3	0D315_of_diluted_wines	Proline

Flavor features

Physical features

Conclusions

- No one tool to rule them all
- Mix and match

Thanks!

mor.hananovitz@gmail.com