MAKERERE  UNIVERSITY

SEMESTER ONE 2024/2025 ACADEMIC YEAR

SCHOOL COMPUTING AND IMFORMATICS TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

MASTER OF SCIENCE IN COMPUTER SCIENCE

MCS 7103

MACHINE LEARNING

ASSIGNMENT ONE

NALIKKA JOAN DEBORAH

2024/HD05/26060U

2400726060

Introduction:

This report highlights Exploratory Data Analysis and data wrangling on the wine recognition dataset. The aim of this dataset is to determine which physicochemical features influence the wine quality and whether this influence is positive or negative.

There are 13 features and each is studied closely to other features and the findings are reported below.

EXPLORATORY DATA ANALYSIS:

I performed in-depth Exploratory data analysis to discover the patterns and how different features interact with each other as well as their relationship with the 'Target' and below are my findings:

Questions and their answers that were asked before EDA:

1. How much data do I have?

I have 13 features and here is the summary of my features' metadata:

	name	role	type	demographic
0	class	Target	Categorical	None
1	Alcohol	Feature	Continuous	None
2	Malicacid	Feature	Continuous	None
3	Ash	Feature	Continuous	None
4	Alcalinity_of_ash	Feature	Continuous	None
5	Magnesium	Feature	Integer	None
6	Total_phenols	Feature	Continuous	None
7	Flavanoids	Feature	Continuous	None
8	Nonflavanoid_phenols	Feature	Continuous	None
9	Proanthocyanins	Feature	Continuous	None
10	Color_intensity	Feature	Continuous	None
11	Hue	Feature	Continuous	None
12	0D280_0D315_of_diluted_wines	Feature	Continuous	None
13	Proline	Feature	Integer	None

I also have 178 data entries.

```
print(X.shape)
```

```
(178, 13)
```

2. How much variability exists within each variable?

I generated the summary of the statistics for all my features and here are my key findings:

	Alcohol	Malicacid	Ash	Alcalinity_of_ash	Magnesium
count	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573
std	0.811827	1.117146	0.274344	3.339564	14.282484
min	11.030000	0.740000	1.360000	10.600000	70.000000
25%	12.362500	1.602500	2.210000	17.200000	88.000000
50%	13.050000	1.865000	2.360000	19.500000	98.000000
75%	13.677500	3.082500	2.557500	21.500000	107.000000
max	14.830000	5.800000	3.230000	30.000000	162.000000

	Total_phenols	Flavanoids	Nonflavanoid_phenols	Proanthocyanins
count	178.000000	178.000000	178.000000	178.000000
mean	2.295112	2.029270	0.361854	1.590899
std	0.625851	0.998859	0.124453	0.572359
min	0.980000	0.340000	0.130000	0.410000
25%	1.742500	1.205000	0.270000	1.250000
50%	2.355000	2.135000	0.340000	1.555000
75%	2.800000	2.875000	0.437500	1.950000
max	3.880000	5.080000	0.660000	3.580000

	Color_intensity	Hue	0D280_0D315_of_diluted_wines	Proline
count	178.000000	178.000000	178.000000	178.000000
mean	5.058090	0.957449	2.611685	746.893258
std	2.318286	0.228572	0.709990	314.907474
min	1.280000	0.480000	1.270000	278.000000
25%	3.220000	0.782500	1.937500	500.500000
50%	4.690000	0.965000	2.780000	673.500000
75%	6.200000	1.120000	3.170000	985.000000
max	13.000000	1.710000	4.000000	1680.000000

With the skewness of

Alcohol	-0.051482
Malicacid	1.039651
Ash	-0.176699
Alcalinity_of_ash	0.213047
Magnesium	1.098191
Total_phenols	0.086639
Flavanoids	0.025344
Nonflavanoid_phenols	0.450151
Proanthocyanins	0.517137
Color_intensity	0.868585
Hue	0.021091
0D280_0D315_of_diluted_wines	-0.307285
:Proline	0.767822

And Kurtosis of:

```
Alcohol                         -0.852500
Malicacid                        0.299207
Ash                              1.143978
Alcalinity_of_ash                0.487942
Magnesium                        2.104991
Total_phenols                   -0.835627
Flavanoids                      -0.880382
Nonflavanoid_phenols            -0.637191
Proanthocyanins                  0.554649
Color_intensity                  0.381522
Hue                             -0.344096
0D280_0D315_of_diluted_wines    -1.086435
Proline                         -0.248403
```

I also plotted the histograms for the various features to show the variability and they can be found in the code.

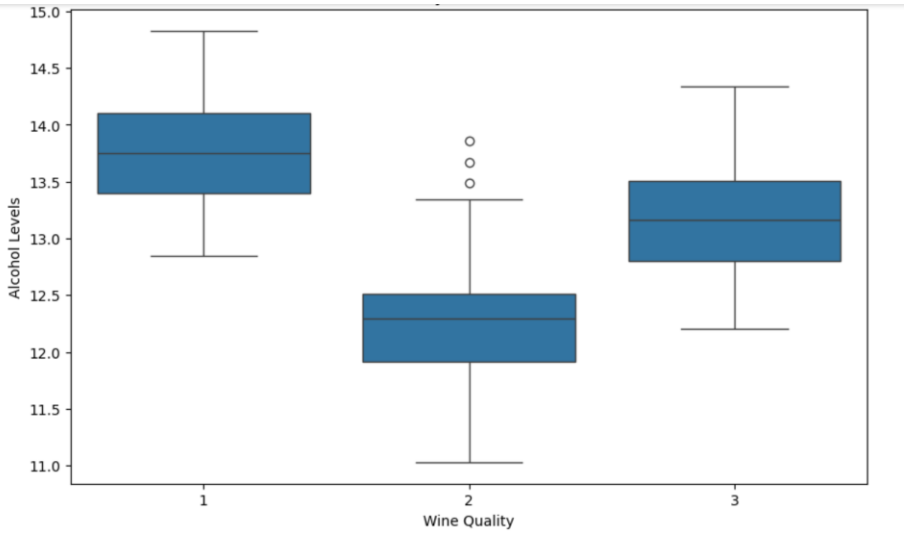
3. How do the different features interact with each other? / What are the relationships between chemical properties and wine quality?

   I created a boxplot for this relationship and the findings can be found in the code.

4. How is wine quality distributed across different alcohol levels or acidity?

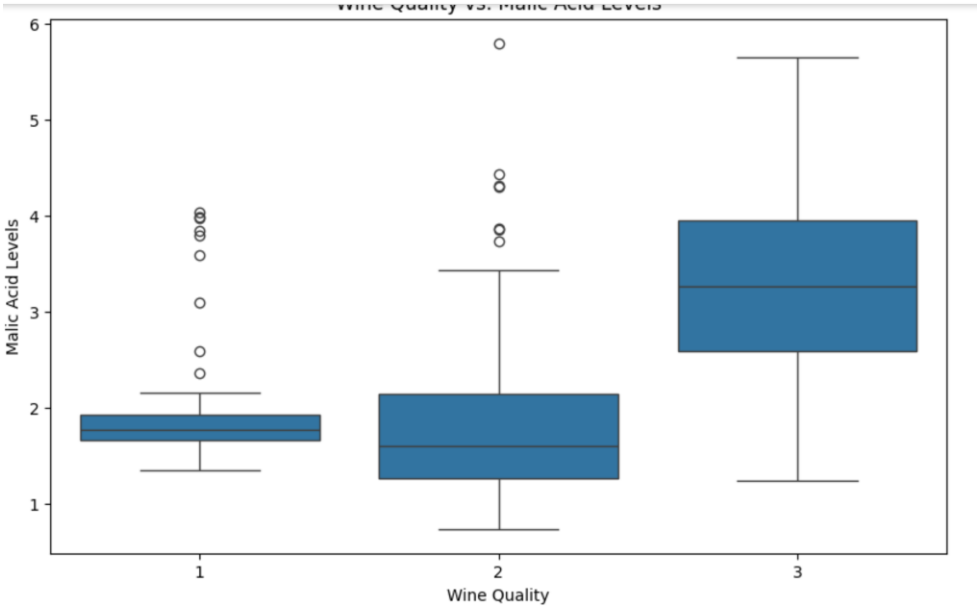
The wine is classified in 3 classes namely: 1,2 and 3.

a) Alcohol levels:



For alcohol content above 13.4, they are classified as class 1, for alcohol content between 12.4 and 13.4 are classified as class 3 and below 12.4 is classified as class 2.

b) Acidity:



For acid levels from 2.5 above, they are classified as class 3, class 2 has acid levels between 1.5 and 2.3 while class 1 is between 1.7 and 1.9 with occasional presence between 2. 3 and 4. 2.

According to the data representation, considering only the Malicacid to determine the wine class can lead to inaccurate class assignments in some models.

DATA WRANGLING:

The following are the questions that guided me during the data wrangling stage, and there corresponding answers in regards to my findings from my dataset:

1. Are there any missing values?

```
print(X.isnull().sum())
```

```
Alcohol                         0
Malicacid                       0
Ash                             0
Alcalinity_of_ash               0
Magnesium                       0
Total_phenols                   0
Flavanoids                      0
Nonflavanoid_phenols            0
Proanthocyanins                 0
Color_intensity                 0
Hue                             0
0D280_0D315_of_diluted_wines    0
Proline                         0
dtype: int64
```

There are no missing values.

2. Are there any duplicates?

```
print(X.duplicated().sum())
```

```
0
```

There are no duplicates in the dataset.

3. Are there outliers in the chemical properties?
   I used the Interquartile Range (IQR), to determine if there are any outliers in my values and this is a visual representation of how that was achieved:

```
import numpy as np

# Define a function to calculate outliers using IQR
def detect_outliers_iqr(df):
    outliers = {}
    for feature in df.columns:
        Q1 = df[feature].quantile(0.25)
        Q3 = df[feature].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR
        outliers[feature] = df[(df[feature] < lower_bound) | (df[feature] > upper_bound)].index.tolist()
    return outliers

# Detect outliers in the wine dataset
outliers = detect_outliers_iqr(X)
print(outliers)
```

These are my results:

```
{'Alcohol': [], 'Malicacid': [123, 137, 173], 'Ash': [25, 59, 121], 'Alcalinity_of_ash': [59, 73, 121, 127],
'Magnesium': [69, 73, 78, 95], 'Total_phenols': [], 'Flavanoids': [], 'Nonflavanoid_phenols': [], 'Proanthocyanins': [95, 110],
 'Color_intensity': [151, 158, 159, 166], 'Hue': [115], '0D280_0D315_of_diluted_wines': [], 'Proline': [],
```

I have outliers in Malicacid, Ash, Alcalinity_of_ash, magnesium, Total_phenols, Proanthocyanins, color_intensity, hues and 0D280_0D315_of_diluted_wines.

4. How were the outliers handled?

The outliers were capped, whereby the values that were below the lowest limit were given the value of the limit and the values that were above the highest limit were given the value of the limit respectively.

5. Is the data clean?

   After answering all the above questions satisfactorily, I can confirm that my data is clean.

Findings/Insights: -

- There were no missing values or duplicate values which mean the dataset is complete and ready for modelling.

- There were outliers in some features including Magnesium and Malicacid. There may have been measurement errors. This was handled by capping the outliers' values to their respective limit values.
- There was a strong positive correlation between Flavanoids and Total_phenols which means these values are closely related.
- There is also a strong negative correlation between Hue and Malicacid which means these values have limited predictive power for some models.
- The dataset heatmap shows that Alcohol and Flavanoids are the most influential features in determining wine quality. Therefore, models that rely on these features are the most likely to perform well.