Assignment 1: Data Preprocessing and Visualization

Rumeha Asim

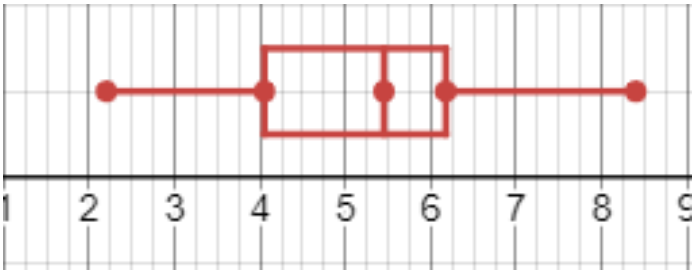
July 2024

1 Problem 1.1

1.1 Question 1 solution:

Minimum = 2.2
First Quartile = 4.0
Median = 5.45
3rd Quartile = 6.3
Maximum = 13.4

1.2 Question 2 solution:



1.3 Question 3 solution:

- **Equal Frequency Partitioning:**
  Bin 1: [2.2, 2.5, 2.5, 2.7 4.0]
  Bin 2: [4.0, 4.2, 4.3, 4.8, 5.4]
  Bin 3: [5.5, 5.5, 5.5, 5.8, 6.3]
  Bin 4: [6.5, 7.9 7.9 8.4 13.4]

- **Equal Width Partitioning:**
  Bin 1: [2.2, 2.5, 2.5, 2.7, 4.0, 4.0 4.2, 4.3, 4.8]
  Bin 2: [5.4, 5.5, 5.5, 5.5, 5.8, 6.2, 6.5]
  Bin 3: [7.9, 7.9, 8.4]
  Bin 4: [13.4]

1.4 Question 4 solution:

Any value that exceeds the range of 1.5 times the inter-quartile range of the box plot is consdiered an outlier, which in this case is 13.4.

1.5 Question 5 solution:

- **Deletion:** This involves removing rows or columns with missing values. The resultant data set would be
- **Imputation:** This method finds the closest data points based on available features and uses their values to estimate the missing values. Such as using the mean, median or mode.

2 Problem 1.2

2.1 Part(a) solution:

Programming language: Python

```
import numpy as np
import statistics as st
from scipy import stats

marks = [56.7, 45, 65, 77, 75, 78, 56, 73, 75, 24,
0, 10, 100, 95, 45, 52, 55, 56, 58, 58.5, 59, 61]

mean = round(np.mean(marks), 1)
print("Mean of the dataset:", mean)

median = round(np.median(marks), 1)
print("Median of the dataset:", median)

mode = st.mode(marks)
print("Mode of the dataset:", mode)

trimMean5 = round(stats.trim_mean(marks, 0.05), 1)
print("5% trimmed mean of the data set:", trimMean5)
```

```
trimMean10 = round(stats.trim_mean(marks, 0.1), 1)
print("10% trimmed mean of the dataset:", trimMean10)

RESULTS FROM CODE:
    Mean of the dataset: 58.8
    Median of the dataset: 60.0
    Mode of the dataset: 45
    5% trimmed mean of the data set: 59.7
    10% trimmed mean of the dataset: 60.5
```

I believe **mean** is the best measure of central tendency as it takes all values of the dataset into consideration for finding the central tendency and any changes in the values will effect the mean. Therefore it is more representative of the dataset values than other measurements.
The most robust measure of centrality would be the **median** as it is unaffected by extreme values, making it a reliable measure in the presence of outliers.

2.2 Part(b) solution:

Programming language: Python

```
import numpy as np
import statistics as st
from scipy import stats

np.random.seed(0)
data = np.random.uniform(1, 5, 22)
print("Randomly generated values:", data)

mean = round(np.mean(data), 1)
print("Mean of the dataset:", mean)

median = round(np.median(data), 1)
print("Median of the dataset:", median)

mode = round(st.mode(data), 1)
print("Mode of the dataset:", mode)

trimMean5 = round(stats.trim_mean(data, 0.05), 1)
print("5% trimmed mean of the data set:", trimMean5)

trimMean10 = round(stats.trim_mean(data, 0.1), 1)
print("10% trimmed mean of the dataset:", trimMean10)
```

```
RESULTS FROM CODE:
Randomly generated values: [3.19525402 3.86075747 3.4110535  3.17953273
2.6946192 3.58357645 2.75034885 4.567092   4.85465104 2.53376608 4.16690015
3.11557968 3.27217824 4.70238655 1.28414423 1.3485172  1.08087359 4.33047938
4.112627   4.48004859 4.91447337 4.19663426]
Mean of the dataset: 3.4
Median of the dataset: 3.5
Mode of the dataset: 3.2
5% trimmed mean of the dataset: 3.5
10% trimmed mean of the dataset: 3.5
```

3 Problem 1.3

3.1 Part(a) solution:

Categorical categories (count of distinct values):

- Class(3)

Numerical categories(count of distinct values):

- Alcohol(127)
- Malicacid(134)
- Ash(79)
- Alcalinity of Ash(63)
- Magnesium (37)
- Total Phenols (97)
- Flavanoids(132)
- Non-flavanoid Phenols (35)
- proanthocyanins(101)
- Colour intensity(132)
- Hue (78)
- 0D280 0D315 of diluted wines(122)
- Proline(121)

3.2 Part(b) solution:

Alcohol:
Mean: 13.13
Median: 13.05
Mode: 13.05
Minimum: 11.03
Maximum: 37
Q1: 12.335
Q3: 13.675

Malicacid:
Mean: 2.35
Median: 1.88
Mode: 1.73
Minimum: 0.74
Maximum: 5.8
Q1: 1.61
Q3: 3.083

Ash:
Mean: 2.37
Median: 2.36
Mode: 2.3
Minimum: 1.36
Maximum: 3.23
Q1: 2.21
Q3: 2.558

Alkalinity of Ash:
Mean: 19.49
Median: 19.5
Mode: 20
Minimum: 10.6
Maximum: 30
Q1: 17.2
Q3: 21.5

Magnesium:
Mean: 99.74
Median: 98.0
Mode: 88
Minimum: 70
Maximum: 162
Q1: 88
Q3: 107

Total Phenols:
Mean: 2.3
Median: 2.36
Mode: 2.2
Minimum: 0.89
Maximum: 3.88
Q1: 1.742
Q3: 2.8

Flavanoid:
Mean: 2.03
Median: 2.13
Mode: 2.65
Minimum: 0.34
Maximum: 5.08
Q1: 1.205
Q3: 2.875

Non-Flavanoid Phenols:
Mean: 0.36
Median: 0.34
Mode:0.26
Minimum: 0.13
Maximum: 0.66
Q1:0.27
Q3: 0.438

Proanthocyanins:
Mean: 1.59
Median: 1.56
Mode: 1.35
Minimum: 0.41
Maximum: 3.58
Q1: 1.25
Q3: 1.95

Colour intensity:
Mean: 5.06
Median: 4.69
Mode: 3.8
Minimum: 1.28
Maximum: 13
Q1: 3.22
Q3: 6.2

Hue:
Mean: 0.96
Median: 0.96
Mode: 1.04
Minimum: 0.48
Maximum: 1.71
Q1: 0.782
Q3: 1.12

0D280 0D315 of diluted wines:
Mean: 2.16
Median: 2.78
Mode: 2.87
Minimum:1.27
Maximum: 4
Q1: 1.938
Q3: 3.17

Proline:
Mean: 746.89
Median: 673.5
Mode: 680
Minimum: 278
Maximum: 1680
Q1: 500.5
Q3: 985

3.3 Part(c) solution:

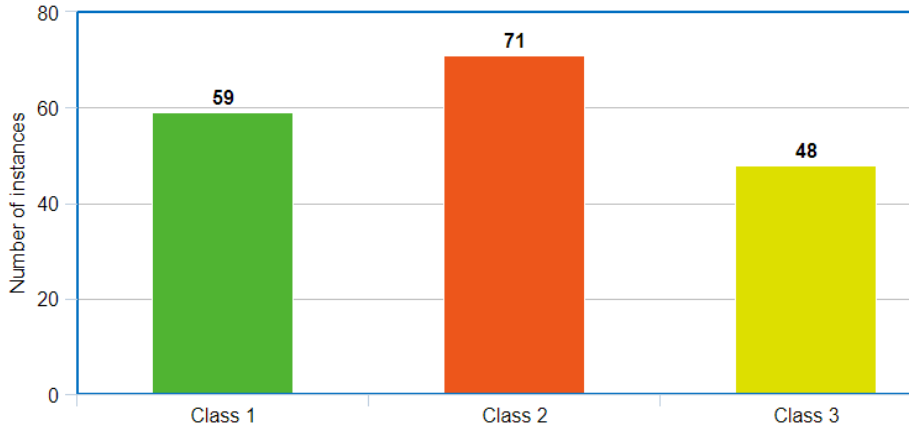


Figure 1: Bar chart for class

3.4 Part(d) solution:

No missing values.

3.5 Part(e) solution:

Attribute name(mean, standard deviation)
Alcohol (13.13, 1.96)
Malicacid (2.35, 1.12)
Ash (2.37, 0.27)
Alkalinity of Ash (19.49, 3.33)
Magnesium (99.74, 14,24)
Total phenols (2.3, 0.62)
Flavanoids (2.03, 1.00)
Non-flavanoid phenols (0.36, 0.12)
Proanthocyanins (1.59, 0.57)
Colour intensity (5.06, 2.3)
Hue (0.96, 0.23)
0D280 0D315 of diluted wines (2.16, 0.71)
Proline (746.84, 314.02)

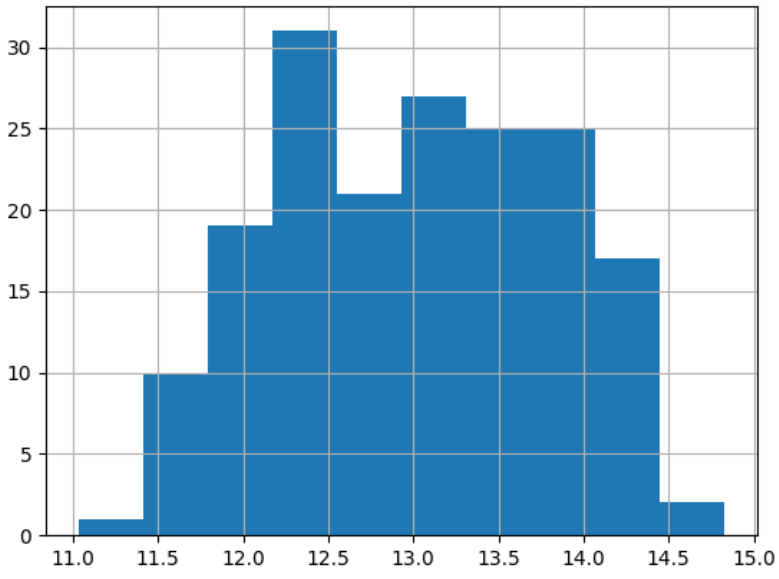


Figure 2: Alcohol quantity histogram

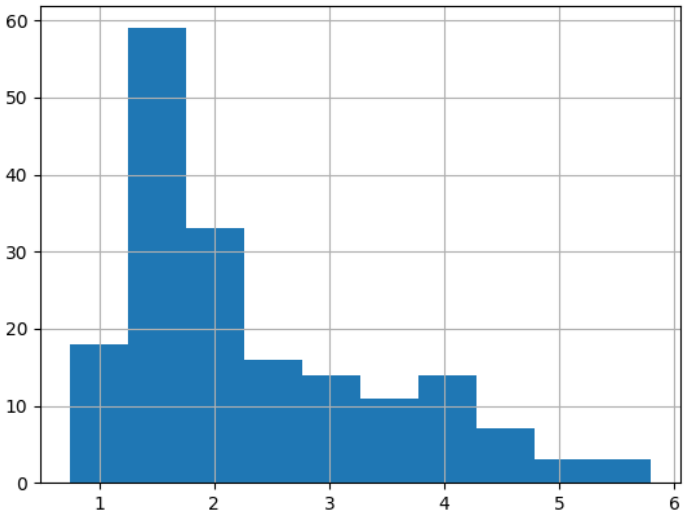


Figure 3: Malicacid quantity histogram

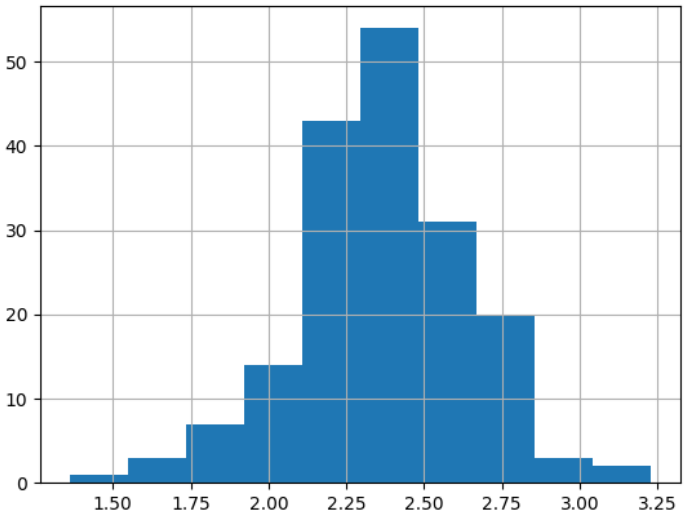


Figure 4: Ash quantity histogram

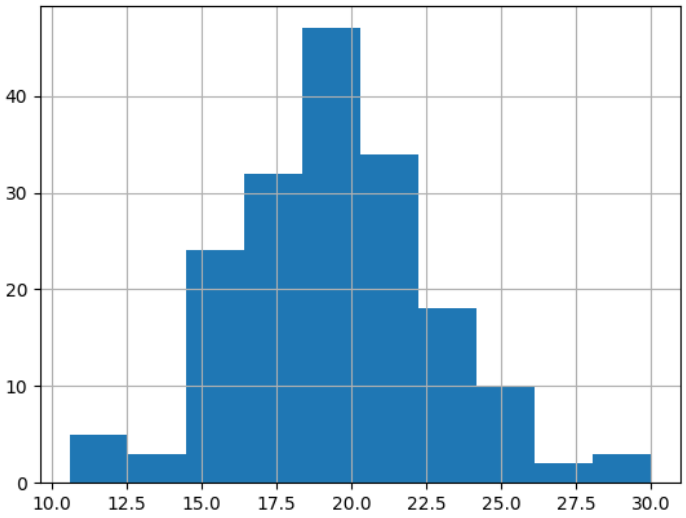


Figure 5: Alkalinity of ash histogram

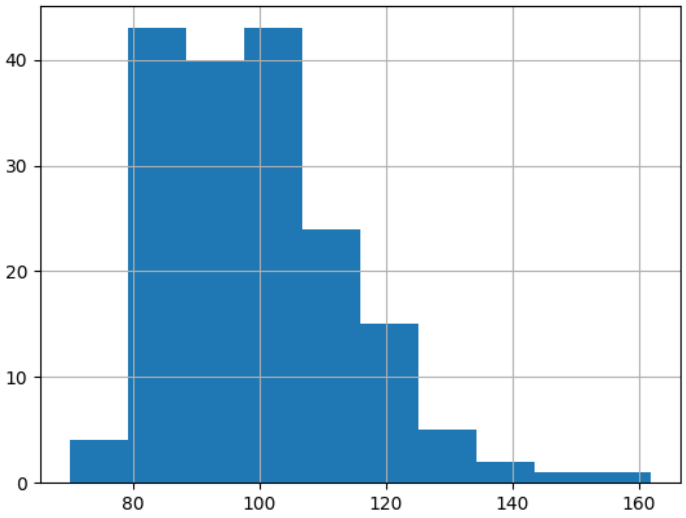


Figure 6: Magnesium quantity histogram

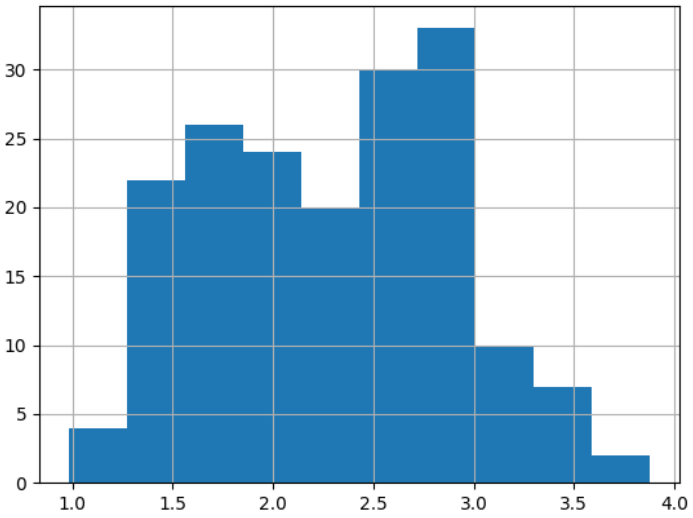


Figure 7: Total phenols quantity histogram

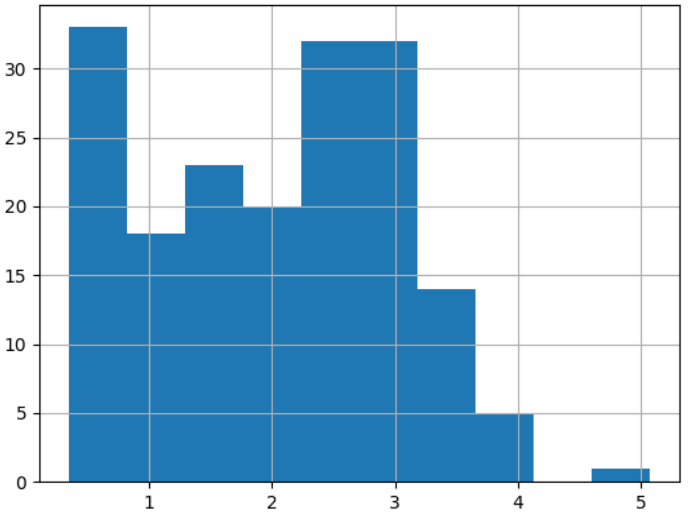


Figure 8: Flavanoid quantity histogram

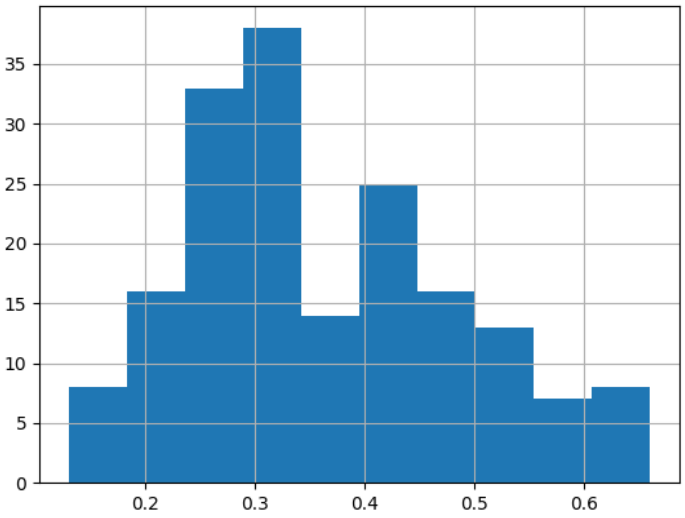


Figure 9: Non-flavanoid phenols quantity histogram

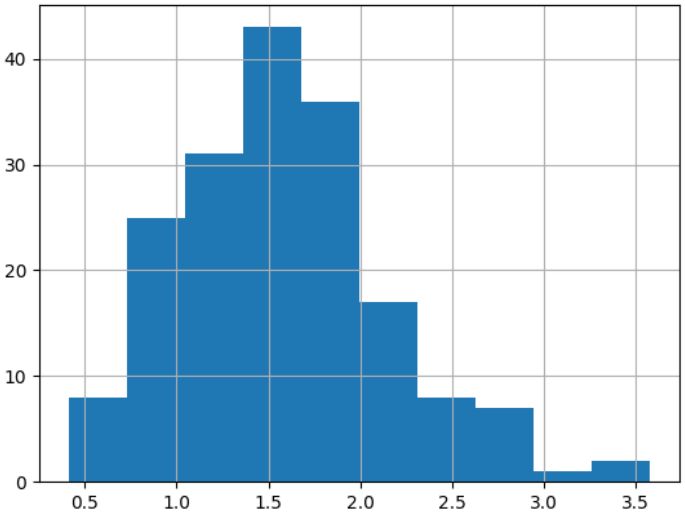


Figure 10: Proanthocyanins quantity histogram

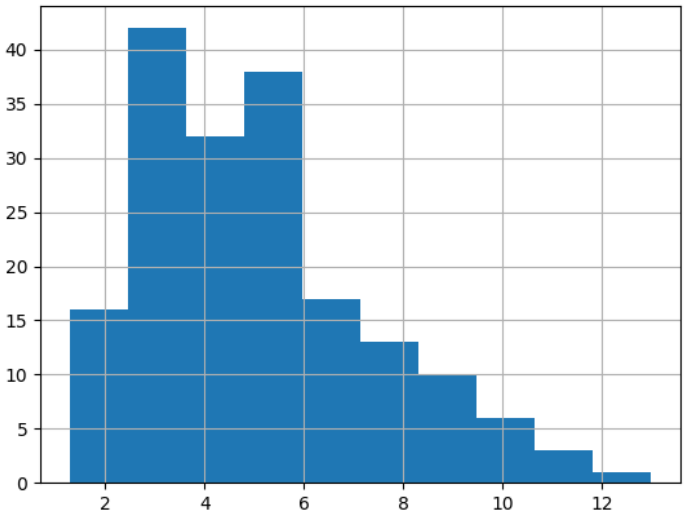


Figure 11: Colour intensity histogram

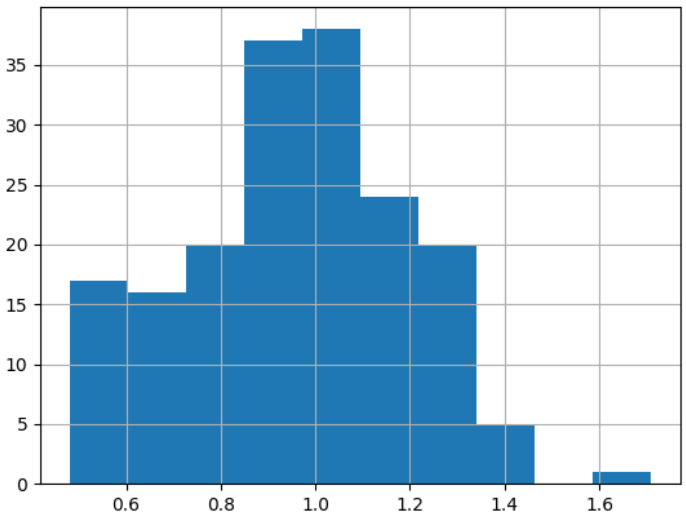


Figure 12: Hue histogram

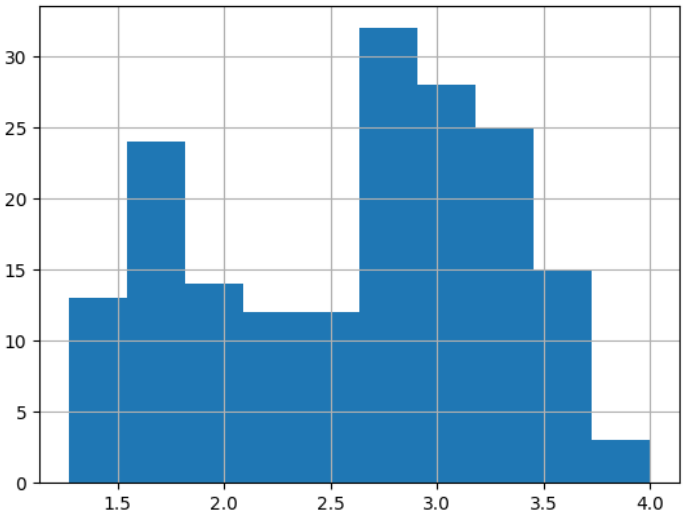


Figure 13: 0D280 0D315 of diluted wines histogram

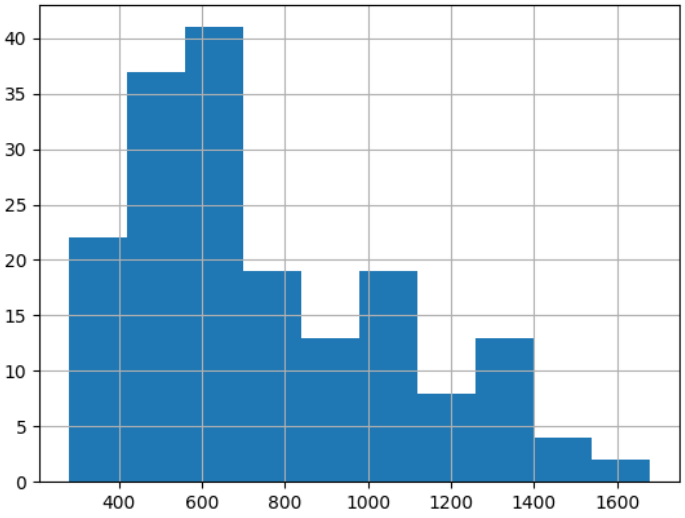
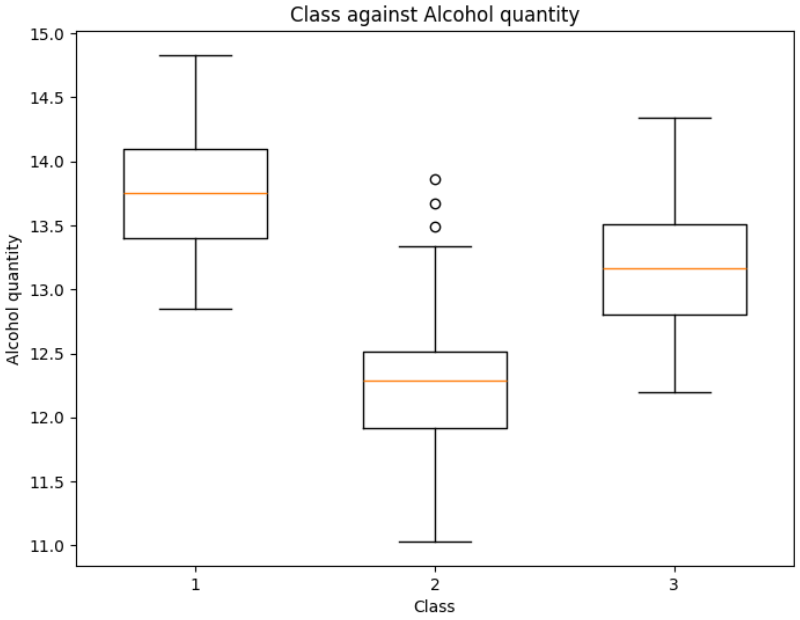
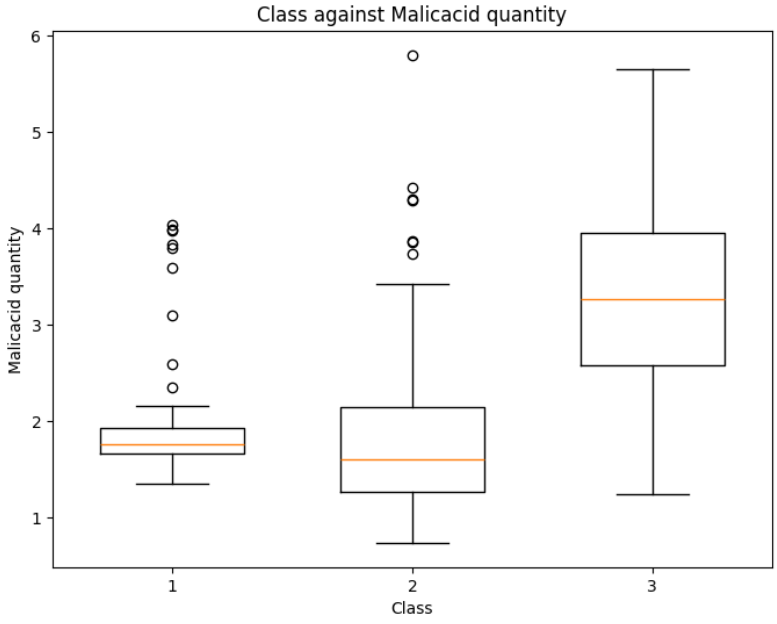
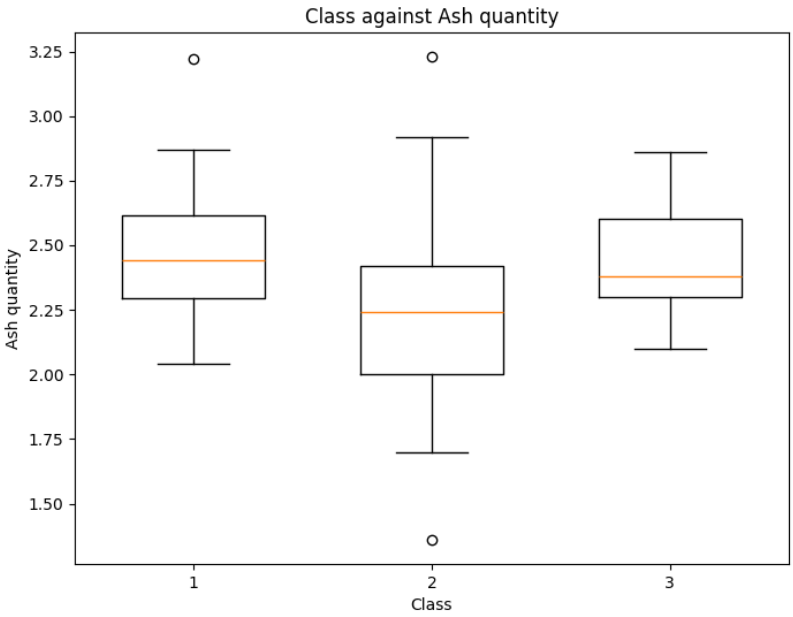


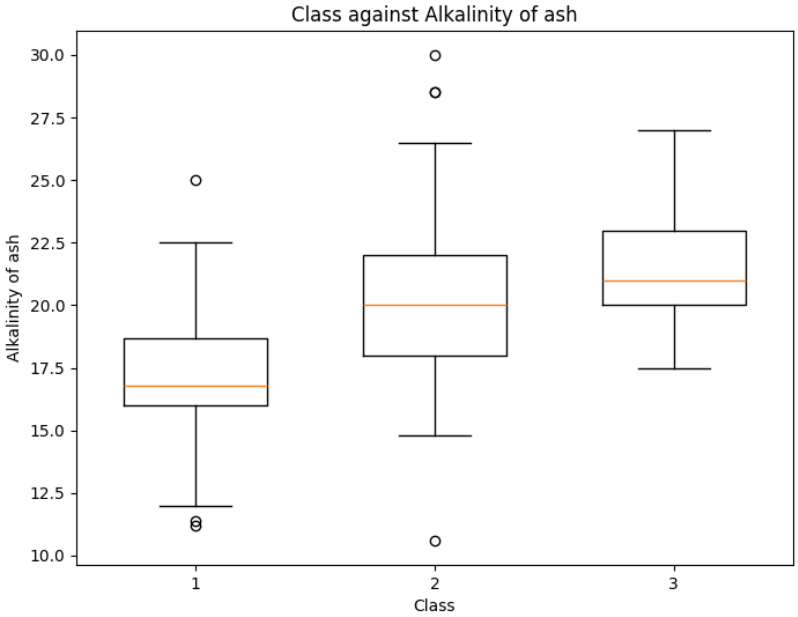
Figure 14: Proline quantity histogram

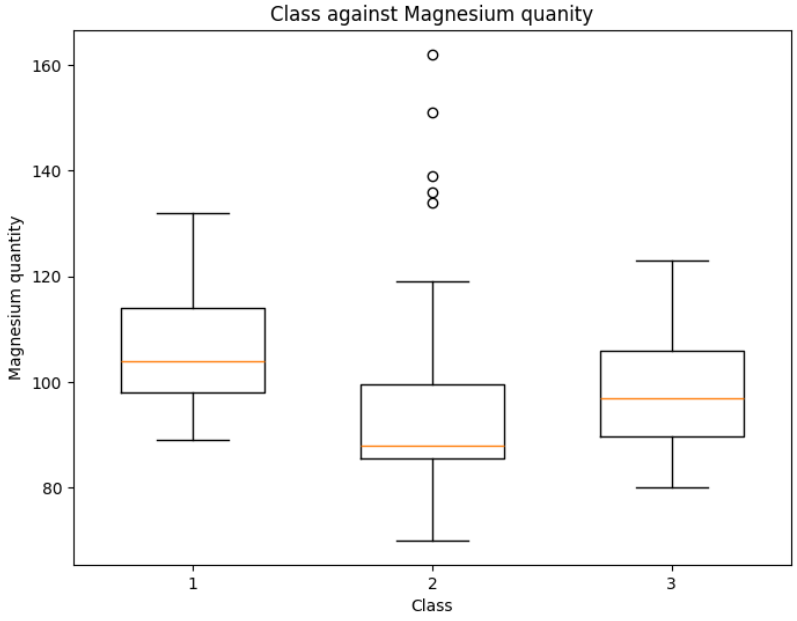
3.6 Part(f) solution:

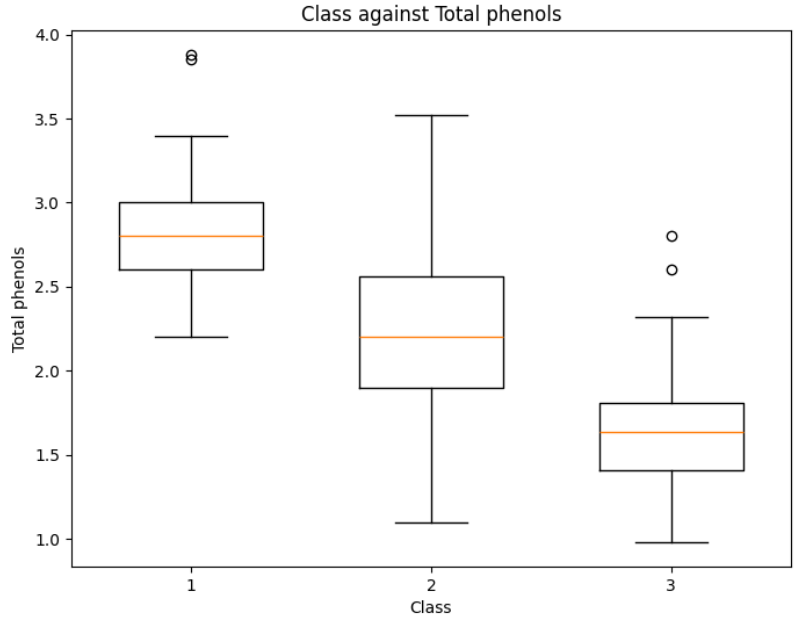


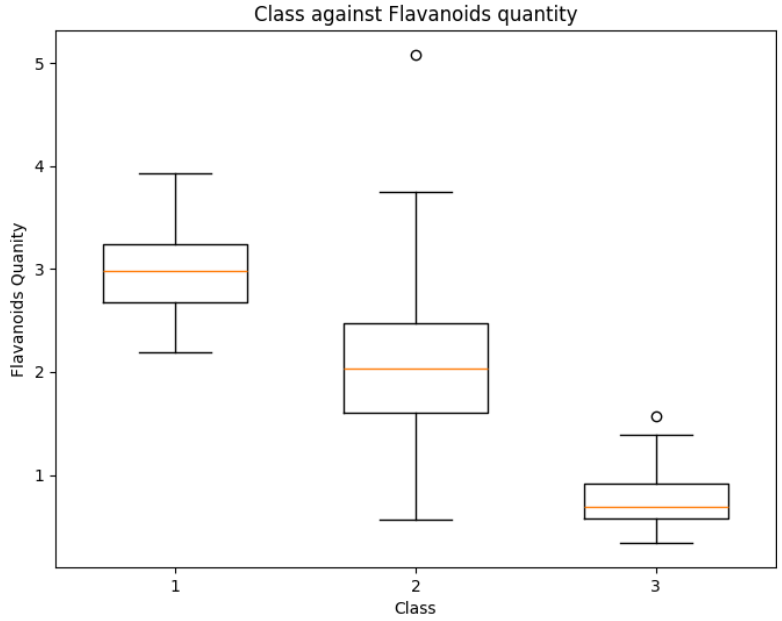


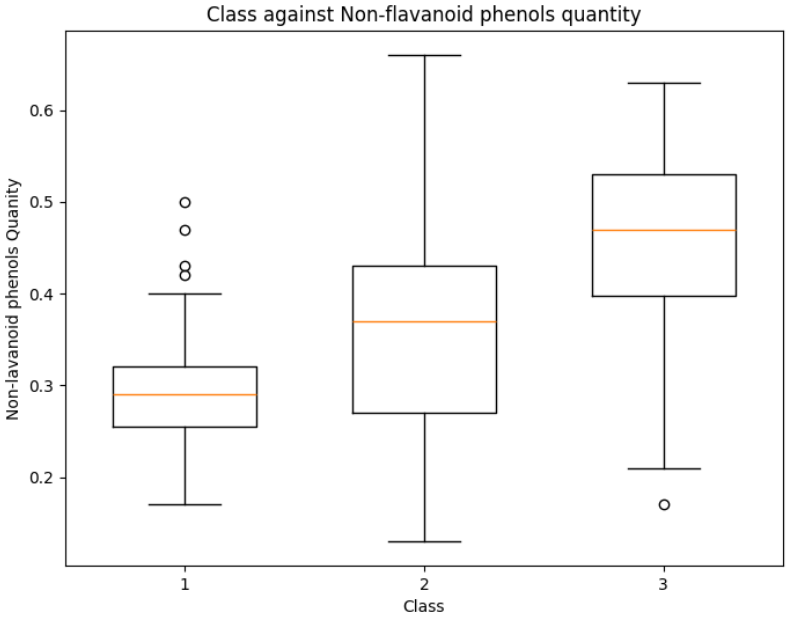

Class against Ash quantity
Ash quantity
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1
2
3
Class

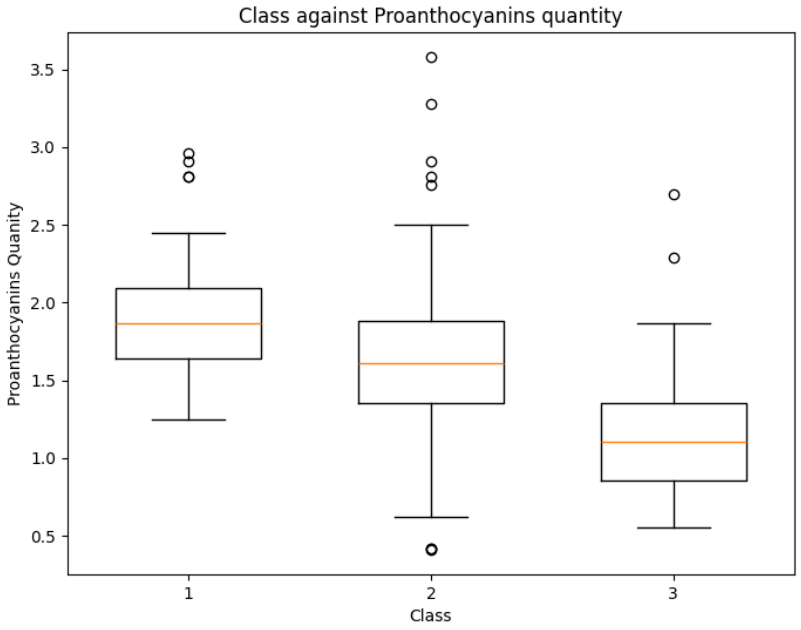

Class against Alkalinity of ash
Alkalinity of ash
30.0
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
1
2
3
Class

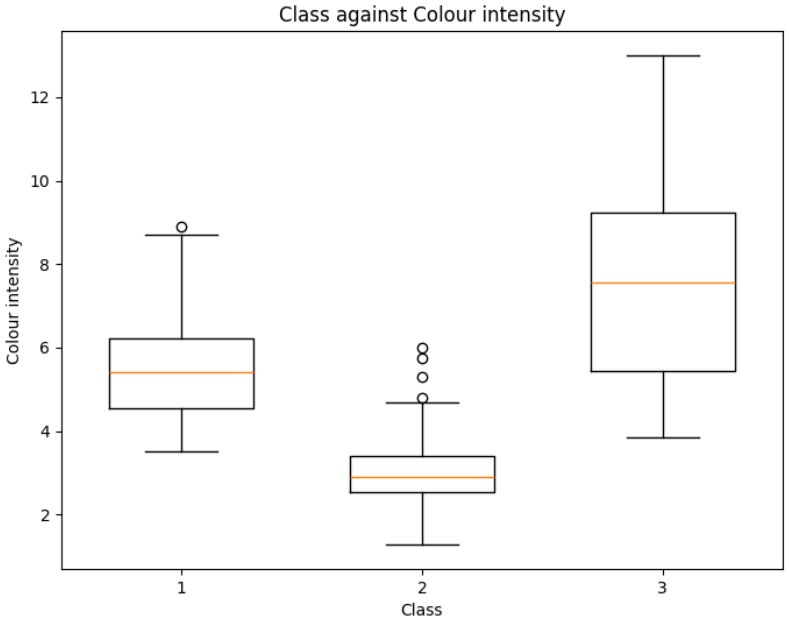

Class against Magnesium quanity
Magnesium quantity
160
140
120
100
80
1
2
3
Class

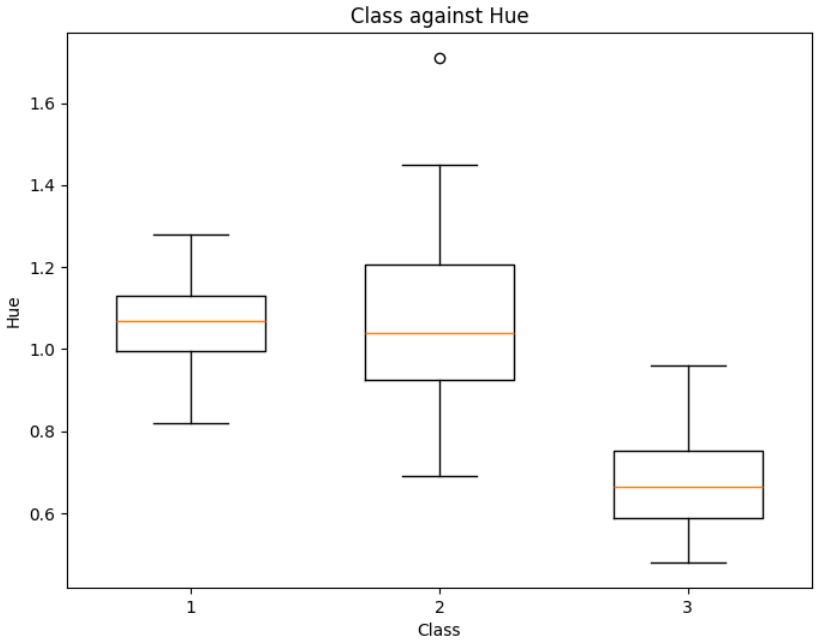

Class against Total phenols
Total phenols
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1
2
3
Class

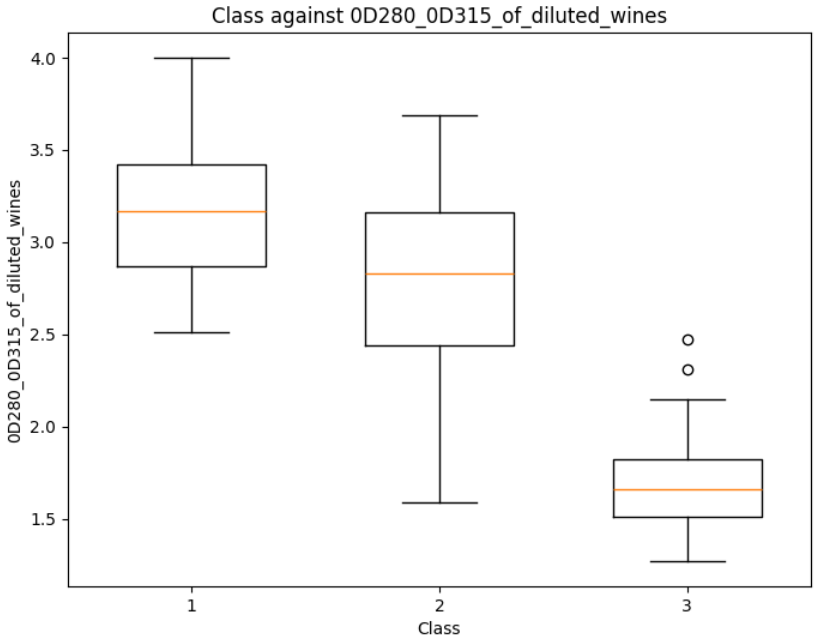

Class against Flavanoids quantity
Flavanoids Quanity
5
4
3
2
1
1
2
3
Class

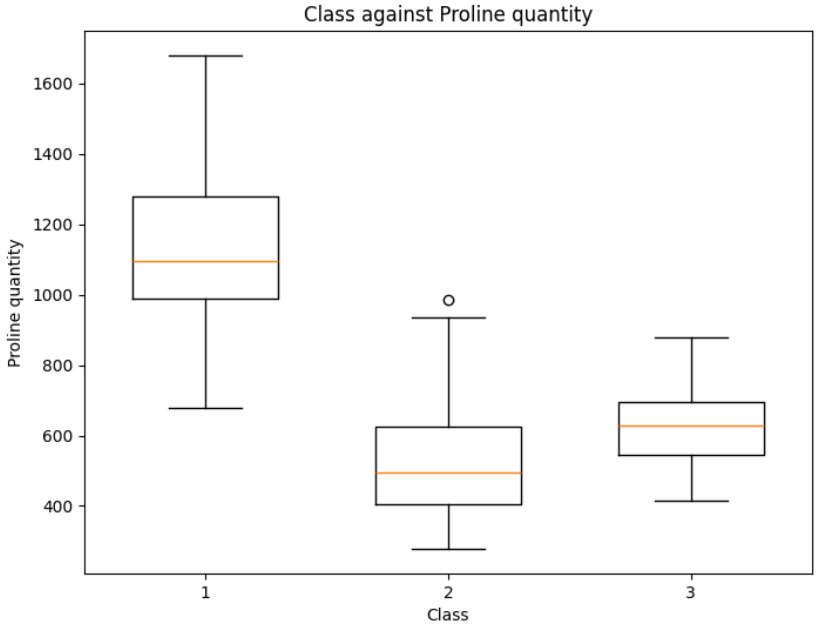

Class against Non-flavanoid phenols quantity
Non-lavanoid phenols Quanity
0.6
0.5
0.4
0.3
0.2
1
2
3
Class


Class against Proanthocyanins quantity
Proanthocyanins Quanity
3.5
3.0
2.5
2.0
1.5
1.0
0.5
1
2
3
Class


Class against Colour intensity
Colour intensity
12
10
8
6
4
2
1
2
3
Class


Class against Hue
Hue
1.6
1.4
1.2
1.0
0.8
0.6
1
2
3
Class


Class against 0D280_0D315_of_diluted_wines
0D280_0D315_of_diluted_wines
4.0
3.5
3.0
2.5
2.0
1.5
1
2
3
Class



3.7 Part(g) solution:

```
Alcohol  11.03  11.41  11.45  11.46  11.56  ...  14.37  14.38  14.39  14.75  14.83
Class                                        ...
1             0      0      0      0      0  ...      1      2      1      1
 1
2             1      1      1      1      1  ...      0      0      0      0
 0
3             0      0      0      0      0  ...      0      0      0      0
 0


Malicacid  0.74  0.89  0.90  0.92  0.94  ...  5.04  5.19  5.51  5.65  5.80
Class                                    ...
1             0     0     0     0     0  ...     0     0     0     0     0
2             1     1     1     1     2  ...     0     0     0     0     1
3             0     0     0     0     0  ...     1     1     1     1     0
```

Ash	1.36	1.70	1.71	1.75	1.82	1.88	...	2.84	2.86	2.87	2.92	3.22	3.23
Class							...						
1	0	0	0	0	0	0	...	1	0	1	0	1	0
2	1	2	1	1	1	1	...	0	0	0	1	0	1
3	0	0	0	0	0	0	...	0	1	0	0	0	0

Alkalinity of ash	10.6	11.2	11.4	12.0	12.4	...	26.0	26.5	27.0	28.5	30.0
Class						...					
1	0	1	1	1	1	...	0	0	0	0	0
2	1	0	0	0	0	...	1	1	0	2	1
3	0	0	0	0	0	...	0	0	1	0	0

Magnesium	70	78	80	81	82	84	85	...	128	132	134	136	139	151	162
Class								...							
1	0	0	0	0	0	0	0	...	1	1	0	0	0	0	0
2	1	3	4	1	1	3	5	...	0	0	1	1	1	1	1
3	0	0	1	0	0	0	1	...	0	0	0	0	0	0	0

Total phenols	0.98	1.10	1.15	1.25	1.28	...	3.40	3.50	3.52	3.85	3.88
Class						...					
1	0	0	0	0	0	...	1	0	0	1	1
2	0	1	0	0	0	...	0	1	1	0	0
3	1	0	1	1	1	...	0	0	0	0	0

Flavanoids	0.34	0.47	0.48	0.49	0.50	...	3.69	3.74	3.75	3.93	5.08
Class						...					
1	0	0	0	0	0	...	1	1	0	1	0
2	0	0	0	0	0	...	0	0	1	0	1
3	1	2	1	1	2	...	0	0	0	0	0

Non-flavanoid phenols	0.13	0.14	0.17	0.19	...	0.60	0.61	0.63	0.66
Class					...				
1	0	0	3	1	...	0	0	0	0
2	1	2	1	1	...	1	1	1	1
3	0	0	1	0	...	2	2	3	0

Colour intensity	1.28	1.74	...	11.75	13.00
Class			...		
1	0	0	...	0	0
2	1	1	...	0	0
3	0	0	...	1	1

Hue	0.48	0.54	0.55	0.56	0.57	...	1.36	1.38	1.42	1.45	1.71
Class						...					
1	0	0	0	0	0	...	0	0	0	0	0
2	0	0	0	0	0	...	2	1	1	1	1
3	1	1	1	2	5	...	0	0	0	0	0

OD280_OD315_of_diluted_wines	1.27	1.29	1.30	1.33	...	3.71	3.82	3.92	4.00
Class					...				
1	0	0	0	0	...	1	1	1	
1									
2	0	0	0	0	...	0	0	0	
0									
3	1	2	1	3	...	0	0	0	
0									

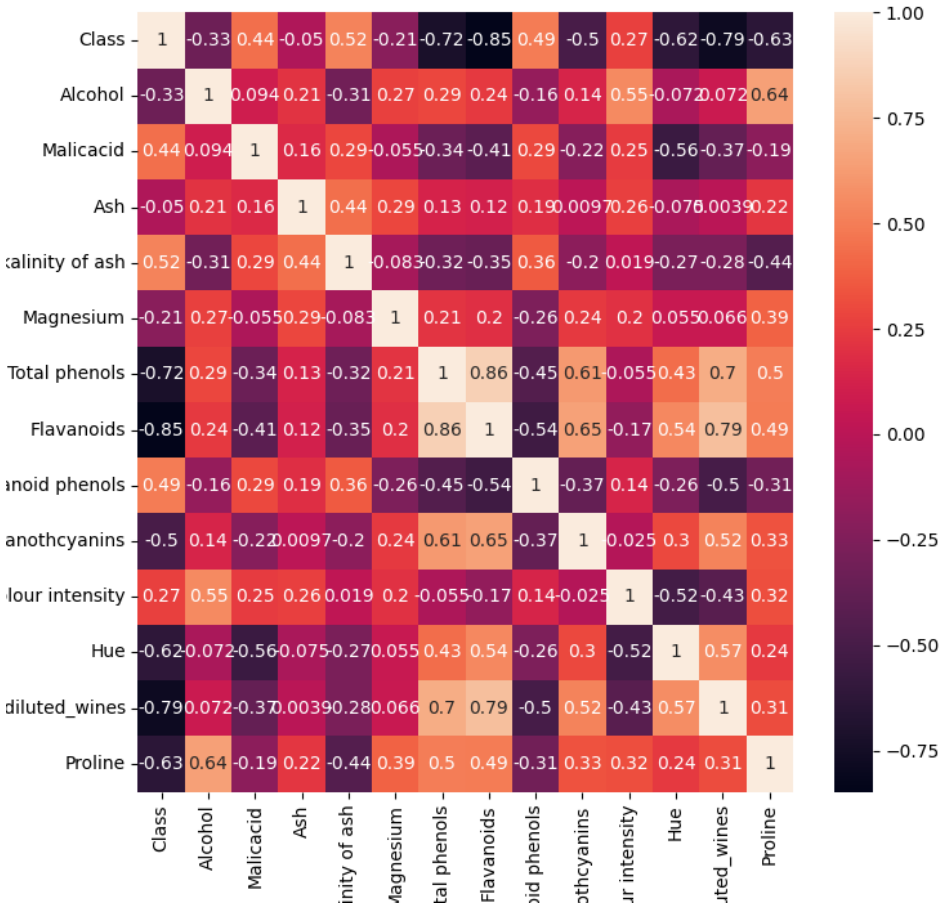
Proline	278	290	312	315	325	342	...	1450	1480	1510	1515	1547	1680
Class							...						
1	0	0	0	0	0	0	...	1	1	1	1	1	1
2	1	1	1	1	1	1	...	0	0	0	0	0	0
3	0	0	0	0	0	0	...	0	0	0	0	0	0

3.8 Part(i) solution:

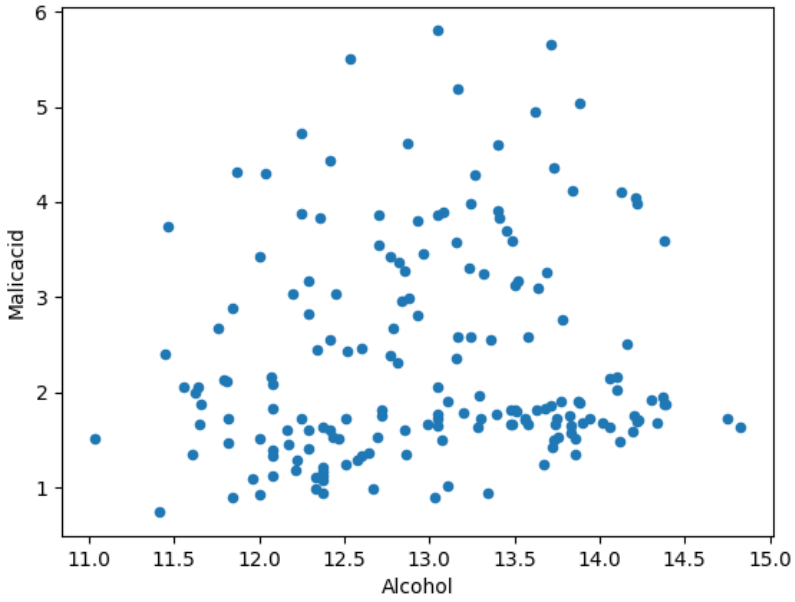
Covariance table

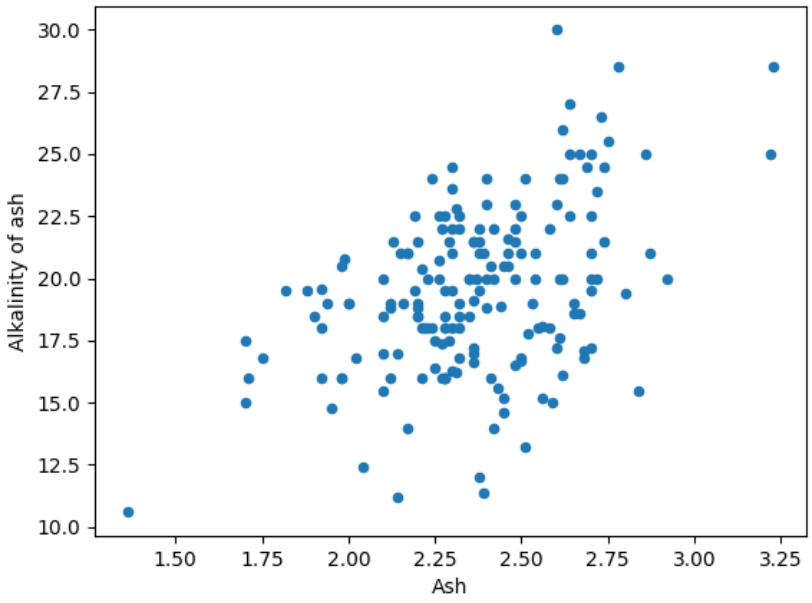
	Class	...	Proline
Class	0.600679	...	-154.667651
Alcohol	-0.206515	...	164.567185
Malicacid	0.379039	...	-67.548867
Ash	-0.010555	...	19.319739
Alkalinity of ash	1.340364	...	-463.355345
Magnesium	-2.315495	...	1769.158700
Total phenols	-0.348835	...	98.171057
Flavanoids	-0.656091	...	155.447492
Non-flavanoid phenols	0.047177	...	-12.203586
Proanothcyanins	-0.221413	...	59.554334
Colour intensity	0.477339	...	230.767480
Hue	-0.109368	...	17.000223
OD280_OD315_of_diluted_wines	-0.433737	...	69.927526
Proline	-154.667651	...	99166.717355

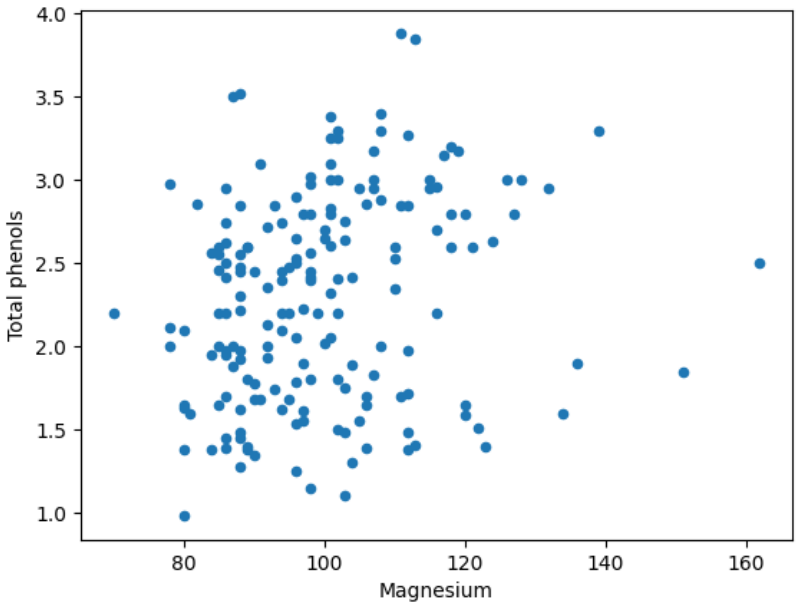
3.9 Part(j) solution:

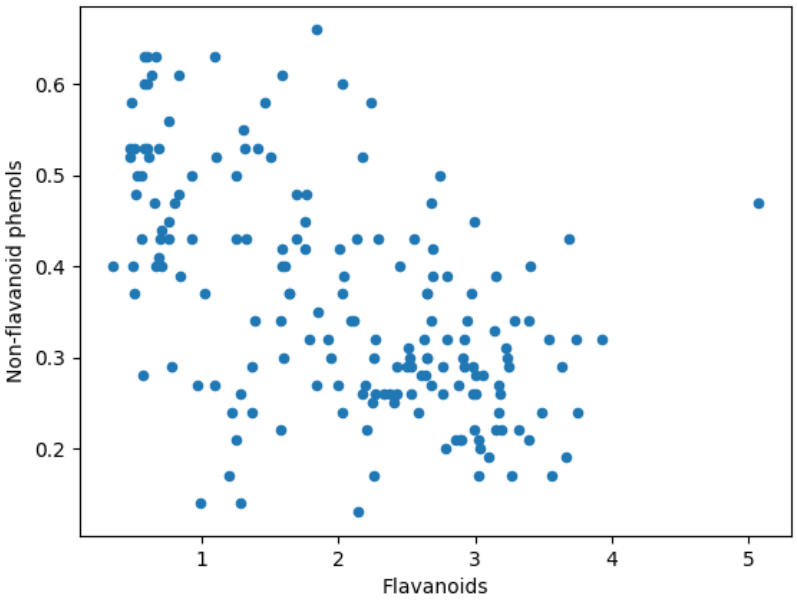


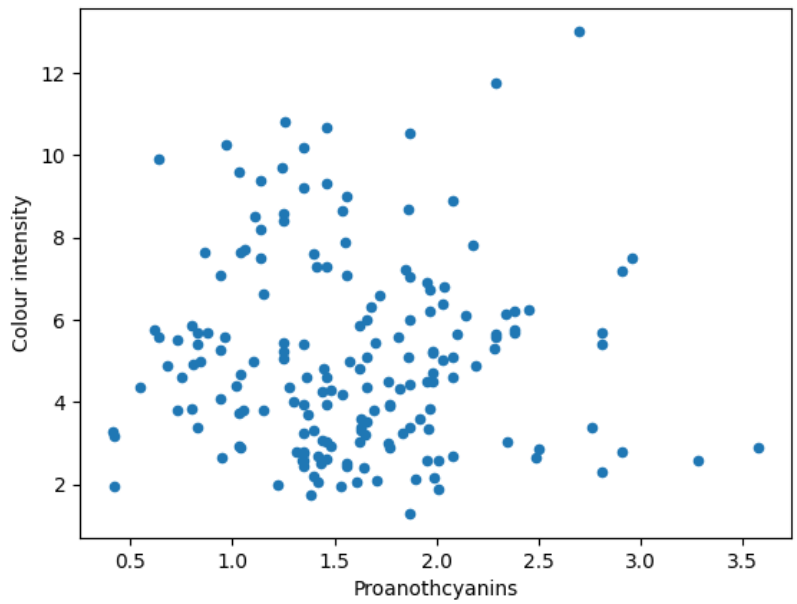
3.10 Part(l) solution:

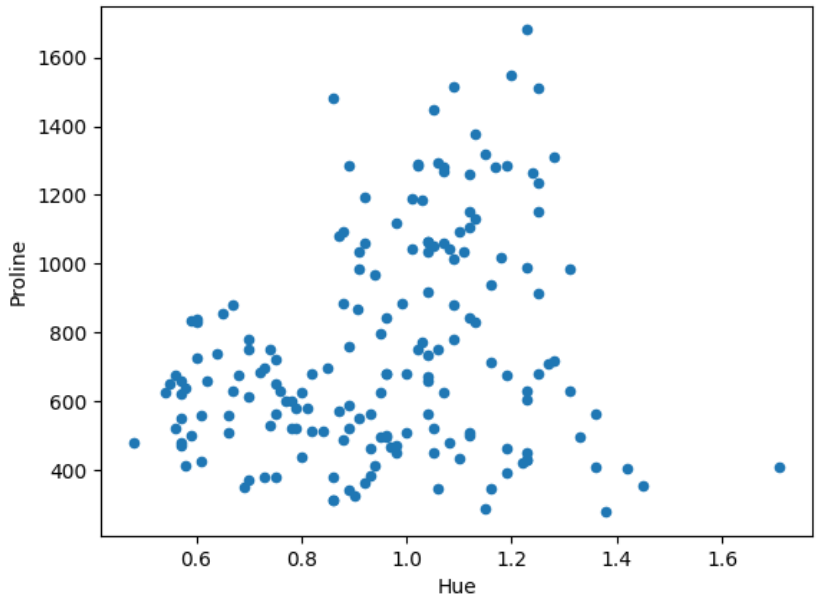



Alkalinity of ash
30.0
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
1.50
1.75
2.00
2.25
2.50
2.75
3.00
3.25
Ash


Total phenols
4.0
3.5
3.0
2.5
2.0
1.5
1.0
80
100
120
140
160
Magnesium


Non-flavanoid phenols
0.6
0.5
0.4
0.3
0.2
1
2
3
4
5
Flavanoids


Colour intensity
12
10
8
6
4
2
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Proanothcyanins



3.11 Part(m) solution:

	Class	Alcohol	Malicacid	Ash	Alkalinity of ash	...	Proanothcyanins
0	0.0	0.842105	0.191700	0.572193	0.257732	...	0.593060
1	0.0	0.571053	0.205534	0.417112	0.030928	...	0.274448
2	0.0	0.560526	0.320158	0.700535	0.412371	...	0.757098
3	0.0	0.878947	0.239130	0.609626	0.319588	...	0.558360
4	0.0	0.581579	0.365613	0.807487	0.536082	...	0.444795
..	...	...	...	...	...	...	...
173	1.0	0.705263	0.970356	0.582888	0.510309	...	0.205047
174	1.0	0.623684	0.626482	0.598930	0.639175	...	0.315457
175	1.0	0.589474	0.699605	0.481283	0.484536	...	0.296530
176	1.0	0.563158	0.365613	0.540107	0.484536	...	0.331230
177	1.0	0.815789	0.664032	0.737968	0.716495	...	0.296530

Colour intensity	Hue	0D280_0D315_of_diluted_wines	Proline
0.372014	0.455285	0.970696	0.561341
0.264505	0.463415	0.780220	0.550642
0.375427	0.447154	0.695971	0.646933
0.556314	0.308943	0.798535	0.857347
0.259386	0.455285	0.608059	0.325963
...	...	...	...
0.547782	0.130081	0.172161	0.329529
0.513652	0.178862	0.106227	0.336662
0.761092	0.089431	0.106227	0.397290
0.684300	0.097561	0.128205	0.400856
0.675768	0.105691	0.120879	0.201141

3.12 Part(n) solution:

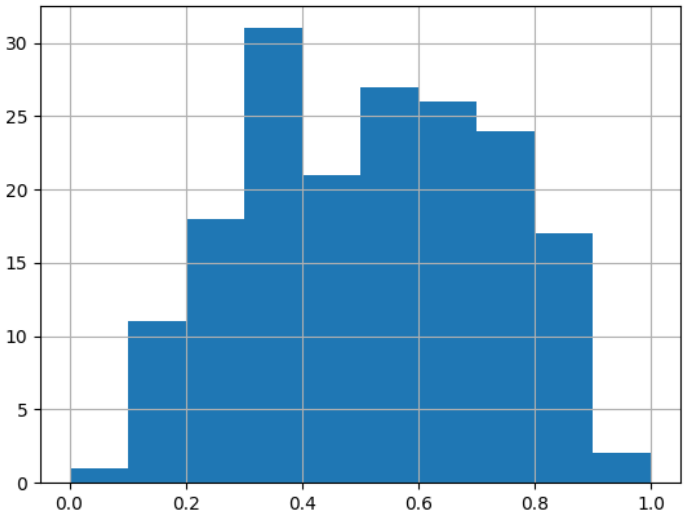


Figure 15: Normalised alcohol quantity histogram

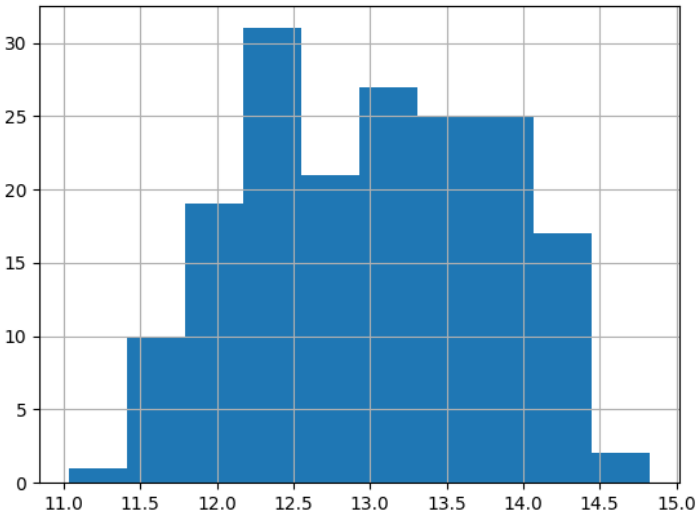


Figure 16: Not normalised histogram

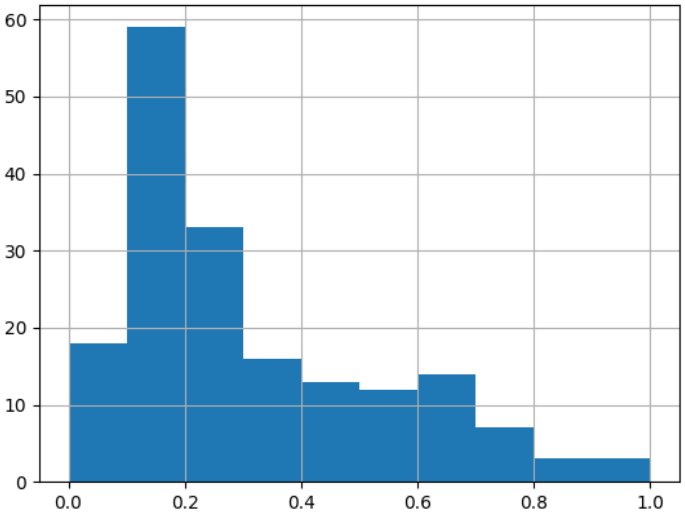


Figure 17: Normalised malicacid histogram

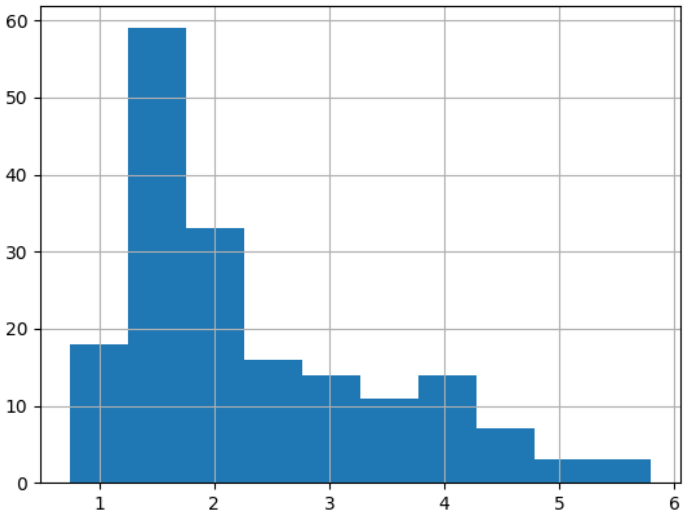


Figure 18: Not normalised histogram

The shape of the normalised histograms and those that were not normalised is the same. But the x-axis values in the normalised histogram do not go above 1 as the min-max approach, used to normalise the datasets to make the histograms, rescales the feature to a range of [0,1].

3.13 Part(p) solution:

The results of this database show that there is a significant relationship between the class of the wine and the quantities of different constituents in the wine. But there is no significant relationship between each of the constituents of the wines. This data base had 13 attributes out of which 1 was a categorical attribute and the others were numeric attributes. This is a total of 178 instances recorded in the database: Class 1, 59; Class 2, 71; Class 3, 48. Line graphs between numerical data could not be made as there was no relationship between the constituents of the wine. Heat maps between 2 numerical attributes also could not be made as there were hundreds of distinct values.

3.14 Part(q) solution:

Outliers of numeric attributes:
Alcohol: NA
Malicacid : 5.8, 5.51, 5.65
Ash : 3.22, 1.36, 3.23

Alkalinity of ash : 10.6, 30.0, 28.5, 28.5
Magnesium: 151, 139, 136, 162
Flavanoids : NA
Non-flavanoid phenols : NA
Proanthocyanins: 3.28, 3.58
Colour intensity: 10.8, 13, 11.75, 10.68
Hue : 1.71
0D280 0D315 of diluted wines: NA
Proline: NA

3.15 Part(r) solution:

Malicacid:
Mean: 2.29
Median: 1.87
Mode: 1.73
Q1:1.61
Q3:3.01
Min:0.74
Max:5.19

Ash:
Mean: 2.36
Median: 2.36
Mode: 2.3
Q1:2.21
Q3:2.55
Min:1.7
Max:2.9

Proanthocyanins:
Mean: 1.57
Median:1.54
Mode:1.35
Q1:1.25
Q3:1.95
Min:0.41
Max:2.96

Alkalinity of ash:
Mean: 19.44
Median: 19.5
Mode: 20
Q1:17.2
Q3:21.5

Min:11.2
Max:30

Magnesium:
Mean: 98.66
Median: 97.5
Mode: 88
Q1:88
Q3:106.8
Min:70
Max:134

Colour intensity:
Mean: 4.91
Median: 4.6
Mode: 3.8
Q1:3.18
Q3:6.08
Min:1.28
Max:10.52

Hue:
Mean: 0.95
Median: 0.96
Mode: 1.04
Q1:0.78
Q3:1.12
Min:0.48
Max:1.45