DMDW ASSIGNMENT 2 – REPORT

SRUJAN R - BT18CSE041

Iris dataset (built-in python dataset):

- Each data object contains information about a unique iris flower and there are 150 data objects in total.
- Attributes of the data object are – Sepal length, Sepal width, petal length, petal width.
- All the four attributes are quantitative and are ratio scales type with a “true zero” of 0 cm.
- As all of them are represented using floating point numbers, they are continuous attributes.
- The dataset doesn’t contain any Nan values.
- As these ratio scale values, all measures of central tendency (mean, median, mode) along with measures of dispersion and percentiles can be found.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

- The mode of each attribute is given below. As we can see, all attributes are unimodal except “petal length” which is bimodal.

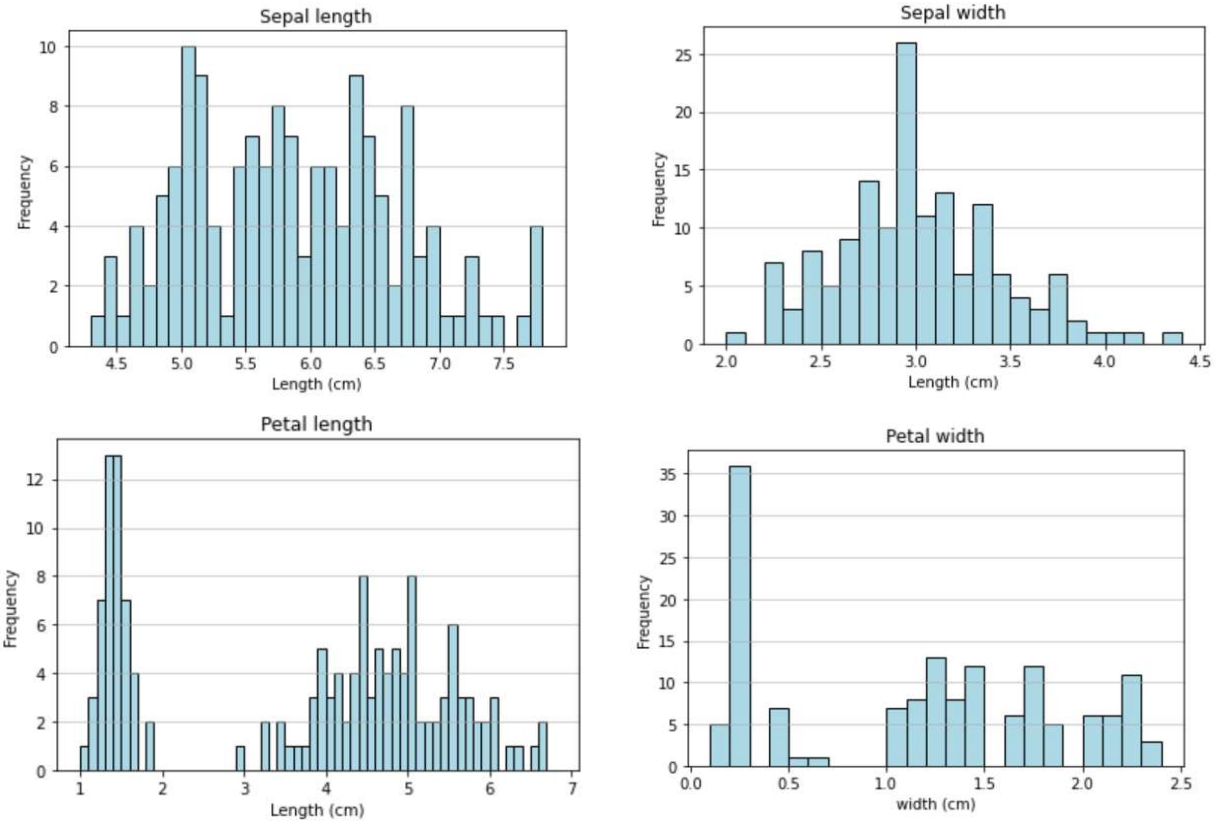
	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.0	3.0	1.4	0.2
1	NaN	NaN	1.5	NaN

- The median of each attribute is:

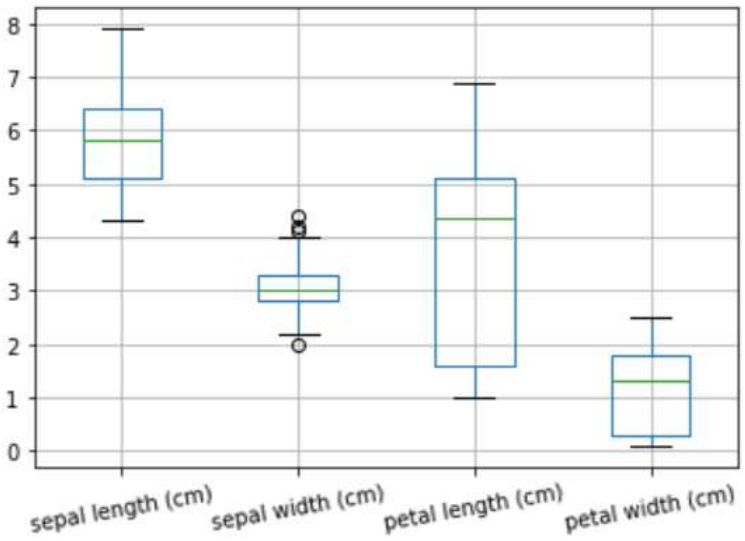
```
sepal length (cm)    5.80
sepal width (cm)     3.00
petal length (cm)    4.35
petal width (cm)     1.30
dtype: float64
```

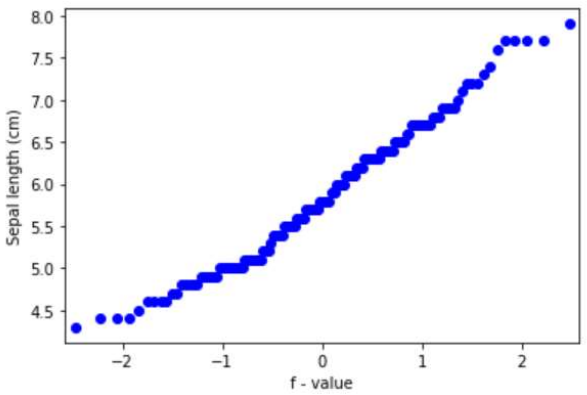
Attribute	Range (cm)	Midrange (cm)	Inter-Quartile Range (cm)
Sepal length	3.60	6.10	1.30
Sepal width	2.40	3.20	0.50
Petal length	5.90	3.95	3.50
Petal width	2.40	1.30	1.50

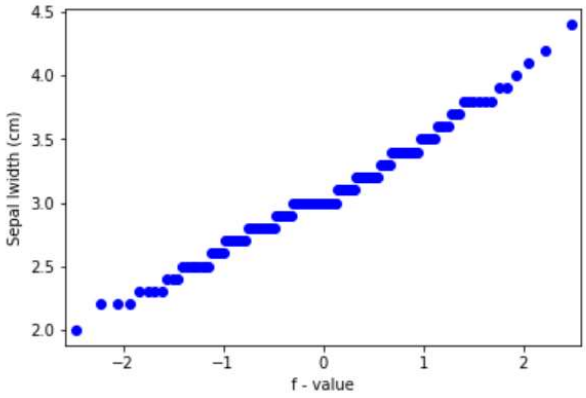
- As petal length has a very high standard deviation, a large Inter-quartile range is seen in its case, so contains some outliers which is also confirmed from the box plot.

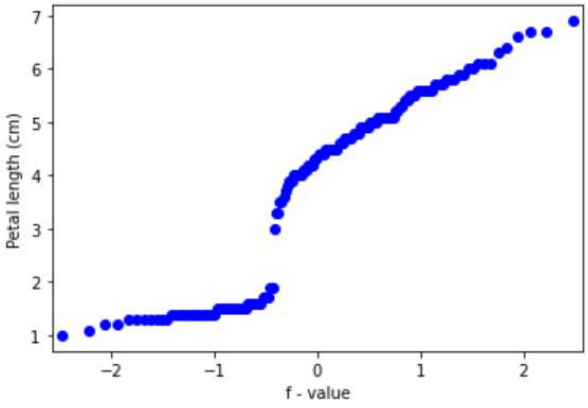


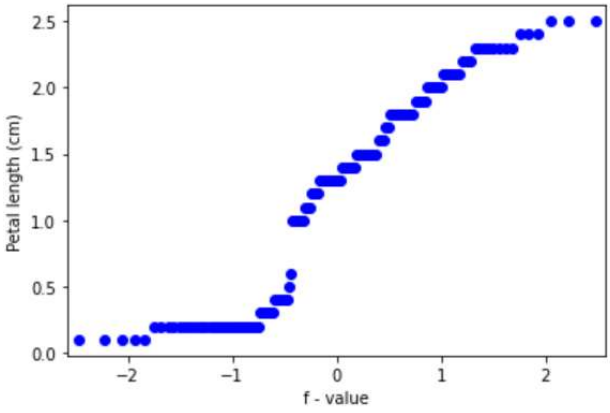
- As we can see from the histograms, sepal width is the only attribute not having any skew (approximately), while the rest are positively skewed attributes, because the mode is lesser than the mean and the median in them.
- The box plot and the Quantile-Quantile plot are given below:











- The cosine similarity and Manhattan distance show that the first 2 data objects/rows are very similar in terms of their numerical values.

```
Cosine Similarity between A and B: 0.9985791635040219
Cosine Distance between A and B: 0.0014208364959781283
```

```
Manhattan Distance between A and B: 0.6999999999999993
```

Student's Performance dataset:

- The original dataset has been modified to 5 columns.
- Each data object refers to academic performance of a particular student.

- The attributes of the data object in the modified version are – Gender (Nominal), Race/ethnicity (Nominal), math score (Ratio scale), reading score (Ratio scale) and writing score (Ratio scale).
- There are no Nan values in the dataset to be filled.

	math score	reading score	writing score
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000

- The mode of each attribute is given below. As we can see, all attributes are unimodal:

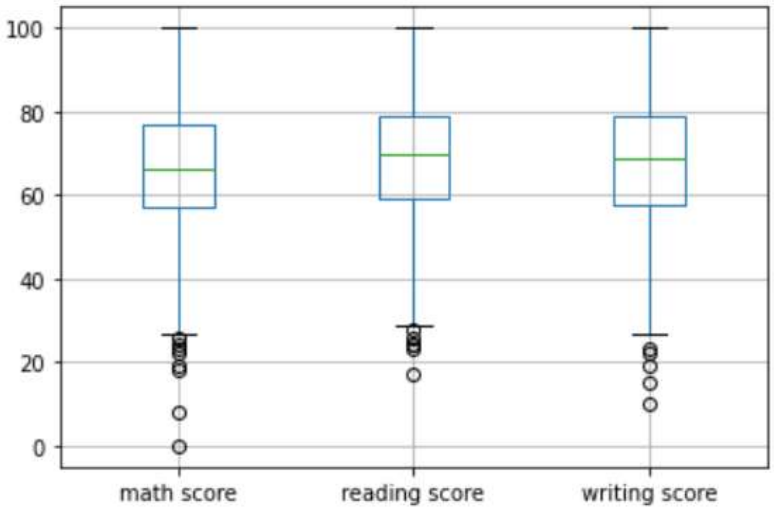
	gender	race/ethnicity	math score	reading score	writing score
0	female	group C	65	72	74

- The median of each attribute is:

```
math score       66.0
reading score    70.0
writing score    69.0
dtype: float64
```

Attribute	Range	Midrange	Inter-Quartile Range
Math score	100.00	50.00	20.00
Reading score	83.00	58.50	20.00
Writing score	90.00	55.00	21.25

- All three attributes have large inter-quantile range, meaning they contain outliers which is also confirmed in the box plot.

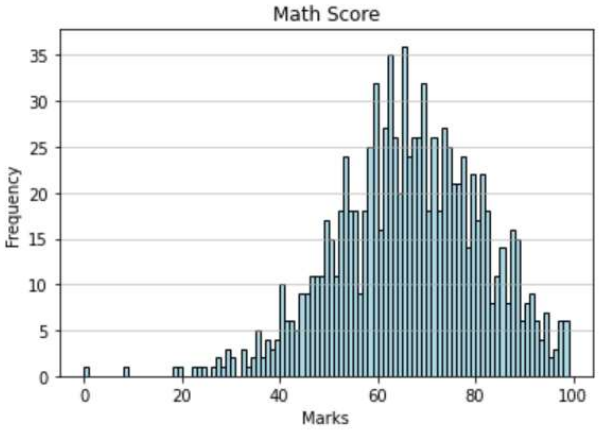


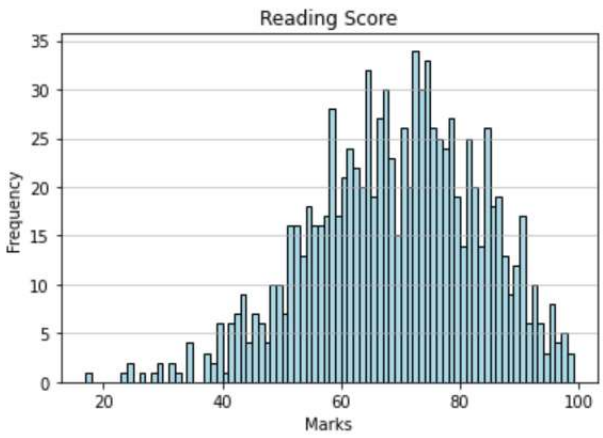
- The nominal attribute gender has a higher frequency of Female values.
- The nominal attribute race/ethnicity has a higher frequency of Group C values.

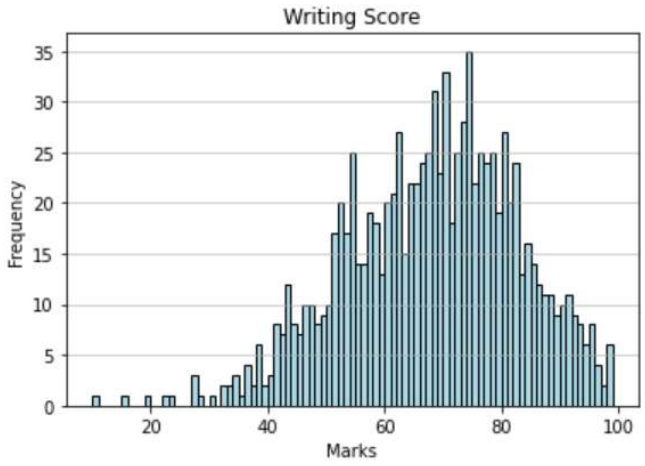
female	518
male	482

group C	319
group D	262
group B	190
group E	140
group A	89

- As we can see from the histograms, math score is the only attribute not having any skew (approximately), while the rest are negatively skewed attributes, because the mode is greater than mean and median in them.
- The Quantile-Quantile plot of all the 3 ratio scale values is given below along with their respective histograms:







- Even though the cosine similarity may show that the academic performance of first and second data objects/rows, a higher a Manhattan distance shows that their performance varies from subject to subject.

```
Cosine Similarity between A and B: 0.9940027960507773
Cosine Distance between A and B: 0.005997203949222651
```

```
Manhattan Distance between A and B: 35
```