Assignment 3 - Descriptive Statistics

Kaustubh Shrikant Kabra

ERP Number :- 38

TE Comp 1

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
df=pd.read_csv('iris flower.csv')
```

The **Iris Dataset** contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species

In [3]:
```
df
```

Out[3]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

In [4]:
```
df.head()
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa

	sepal_length	sepal_width	petal_length	petal_width	species
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Summary of data

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   sepal_length  150 non-null    float64
 1   sepal_width   150 non-null    float64
 2   petal_length  150 non-null    float64
 3   petal_width   150 non-null    float64
 4   species       150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

Mean:

The mean is the average or a calculated central value of a set of numbers and is used to measure the central tendency of the data.

Mean = (Sum of Observations) ÷ (Total Numbers of Observations)

Percentile:

A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.

Percentile Value = $\mu + z\sigma$

where: μ: Mean

z: z-score from z table that corresponds to percentile value

σ: Standard deviation

Standard Deviation :

The standard deviation of a random variable, sample, statistical population, data set, or probability distribution is the square root of its variance.

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{N-1}} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N-1}}$$

σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

In [6]:
```
np.mean(df['sepal_length'])
```

Out[6]:
```
5.843333333333335
```

In [37]:
```
np.median(df['petal_length'])
```

Out[37]:
```
4.35
```

Group By categories

In [40]:
```
print(np.min(df['sepal_width']))
print(np.max(df['sepal_width']))
print(np.std(df['petal_width']))
```

```
2.0
4.4
0.760612618588172
```

In [41]:
```
df['species'].value_counts()
```

Out[41]:
```
Iris-setosa        50
Iris-virginica     50
Iris-versicolor    50
Name: species, dtype: int64
```

In [64]:
```
df.describe()
```

Out[64]:

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
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

	sepal_length	sepal_width	petal_length	petal_width
max	7.900000	4.400000	6.900000	2.500000

In [65]:
```
df.groupby(['species']).mean()
```

Out[65]:

species	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	5.006	3.418	1.464	0.244
Iris-versicolor	5.936	2.770	4.260	1.326
Iris-virginica	6.588	2.974	5.552	2.026

In [66]:
```
df.groupby(['species']).min()
```

Out[66]:

species	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	4.3	2.3	1.0	0.1
Iris-versicolor	4.9	2.0	3.0	1.0
Iris-virginica	4.9	2.2	4.5	1.4

In [67]:
```
df.groupby(['species']).max()
```

Out[67]:

species	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	5.8	4.4	1.9	0.6
Iris-versicolor	7.0	3.4	5.1	1.8
Iris-virginica	7.9	3.8	6.9	2.5

In [68]:
```
df.groupby(['species']).std()
```

Out[68]:

species	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	0.352490	0.381024	0.173511	0.107210
Iris-versicolor	0.516171	0.313798	0.469911	0.197753
Iris-virginica	0.635880	0.322497	0.551895	0.274650

In [74]:
```
df.groupby(['species']).quantile(q=0.5)
```

Out[74]:

species	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	5.0	3.4	1.50	0.2
Iris-versicolor	5.9	2.8	4.35	1.3
Iris-virginica	6.5	3.0	5.55	2.0

In [75]:
```
df.groupby(['species']).quantile(q=0.75)
```

Out[75]:

species	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	5.2	3.675	1.575	0.3
Iris-versicolor	6.3	3.000	4.600	1.5
Iris-virginica	6.9	3.175	5.875	2.3

In [76]:
```
df.groupby(['species']).quantile(q=0.25)
```

Out[76]:

species	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	4.800	3.125	1.4	0.2
Iris-versicolor	5.600	2.525	4.0	1.2
Iris-virginica	6.225	2.800	5.1	1.8

In [96]:
```
df['sepal_length'].loc[(df['species']=='Iris-setosa')]
```

Out[96]:
```
0     5.1
1     4.9
2     4.7
3     4.6
4     5.0
5     5.4
6     4.6
7     5.0
8     4.4
9     4.9
10    5.4
11    4.8
12    4.8
13    4.3
14    5.8
15    5.7
16    5.4
17    5.1
18    5.7
19    5.1
20    5.4
```

```
21    5.1
22    4.6
23    5.1
24    4.8
25    5.0
26    5.0
27    5.2
28    5.2
29    4.7
30    4.8
31    5.4
32    5.2
33    5.5
34    4.9
35    5.0
36    5.5
37    4.9
38    4.4
39    5.1
40    5.0
41    4.5
42    4.4
43    5.0
44    5.1
45    4.8
46    5.1
47    4.6
48    5.3
49    5.0
Name: sepal_length, dtype: float64
```

In [97]:
```
np.mean(df['sepal_length'].loc[(df['species']=='Iris-setosa')])
```

Out[97]:
```
5.005999999999999
```

In [131…
```
data2=df[['sepal_length','species']].loc[(df['species']=='Iris-setosa') | (df['species'
data2=pd.DataFrame(data2)
data2.reset_index()
```

Out[131…

	index	sepal_length	species
0	0	5.1	Iris-setosa
1	1	4.9	Iris-setosa
2	2	4.7	Iris-setosa
3	3	4.6	Iris-setosa
4	4	5.0	Iris-setosa
...	...	...	...
95	145	6.7	Iris-virginica
96	146	6.3	Iris-virginica
97	147	6.5	Iris-virginica
98	148	6.2	Iris-virginica

	index	sepal_length	species
99	149	5.9	Iris-virginica

100 rows × 3 columns

In [120…
```
np.mean(df['sepal_length'].loc[(df['species']=='Iris-setosa') | (df['species']=='Iris-v
```

Out[120…
```
5.796999999999998
```

In [121…
```
data2['species'].value_counts()
```

Out[121…
```
Iris-setosa       50
Iris-virginica    50
Name: species, dtype: int64
```

In [129…
```
data2.iloc[99]
```

Out[129…
```
sepal_length              5.9
species         Iris-virginica
Name: 149, dtype: object
```