Department of Computer Engineering

Academic Year: 2024-25 **Semester: VIII**
Class / Branch: BE Computer **Subject: Applied Data Science Lab**

Experiment No. 1

1. Aim: Explore the descriptive statistics on the given dataset.

Dataset: In this experiment, a fictitious data of Body Mass Index(BMI) containing 10 observations and 5 variables is used. The dataset contains Height, Weight, Age, BMI, and Gender columns.

2. **Software used:** Google Colaboratory/Jupyter Notebook

3. Theory :-

Descriptive Statistics:

Descriptive statistics can be defined as the measures that summarize a given data, and these measures can be broken down further

1. Measure of central tendency
2. Measure of spread/dispersion
3. Measure of symmetry/shape

Measure of Central Tendency

Measure of central tendency is used to describe the middle/centre value of the data. Mean, Median, Mode are measures of central tendency.

1. Mean

- Mean is the average value of the dataset.
- Mean is calculated by adding all values in the dataset divided by the number of values in the dataset.
- We can calculate the mean for only numerical variables.

2. Median

- The Median is the middle number in the dataset.
- Median is the best measure when we have outliers.

3. Mode

The mode is used to find the common number in the dataset.

Measure of spread

- The measure of spread/dispersion is used to describe how data is spread. It also describes the **variability** of the dataset.
- **Standard Deviation, Variance, Range, IQR,** are used to describe the measure of spread/dispersion
- The measure of spread can be shown in graphs like **boxplot**.

1.Variance

- Variance is used to describe how far each number in the dataset is from the mean.
- Formula to calculate population variance

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

2.Standard Deviation

- Standard Deviation is the measure of the spread of data from the mean.
- Standard deviation is the square root of variance.
- More the standard deviation, more the spread.

3.Range

- The range is the difference between the largest number and the smallest number.
- Larger the range, the more the dispersion.

4. Interquartile range (IQR)

- Quartiles describe the spread of data by breaking into quarters. The median exactly divides the data into two parts.
- **Q1(Lower quartile)** is the middle value in the first half of the sorted dataset.
- **Q2**– is the median value
- **Q3 (Upper quartile)** is the middle value in the second half of the sorted dataset
- The interquartile range is the difference between the 75th percentile(Q3) and the 25th percentile(Q1).
- 50% of data fall within this range.

Boxplot is used to describe how the data is distributed in the dataset. This graph represents five-point summary (minimum, maximum, median, lower quartile, and upper quartile) and is used to identify **outliers**.

- whiskers—denote the spread of data
- box—represents the IQR- 50% of data lies within this range.

Measure of shape

1.Skewness

Skewness, which is the measure of the symmetry, or lack of it, for a real-valued random variable about its mean. The skewness value can be positive, negative, or undefined. In a perfectly symmetrical distribution, the mean, the median, and the mode will all have the same value.

2.*Kurtosis*

Kurtosis provides a measurement about the extremities (i.e. tails) of the distribution of data, and therefore provides an indication of the presence of outliers. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.

4. Program

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
from scipy import stats
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
df=pd.read_csv(r'/content/drive/MyDrive/ADS LAB/bmi.csv')
df
```

	Gender	Height	Weight	bmi	Age
0	Male	174	80	26.4	25
1	Male	189	87	24.4	27
2	Female	185	80	23.4	30
3	Female	165	70	25.7	26
4	Male	149	61	27.5	28
5	Male	177	70	22.3	29
6	Female	147	65	30.1	31
7	Male	154	62	26.1	32
8	Male	174	90	29.7	27

```
df.mean()
```

```
<ipython-input-4-c61f0c8f89b5>:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is
  df.mean()
Height    168.222222
Weight     73.888889
bmi        26.177778
Age        28.333333
dtype: float64
```

```
df.median()
```

```
<ipython-input-5-6d467abf240d>:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is
  df.median()
Height    174.0
Weight     70.0
bmi        26.1
Age        28.0
dtype: float64
```

```
df.mode()
```

	Gender	Height	Weight	bmi	Age
0	Male	174.0	70.0	22.3	27.0
1	NaN	NaN	80.0	23.4	NaN
2	NaN	NaN	NaN	24.4	NaN
3	NaN	NaN	NaN	25.7	NaN
4	NaN	NaN	NaN	26.1	NaN
5	NaN	NaN	NaN	26.4	NaN
6	NaN	NaN	NaN	27.5	NaN
7	NaN	NaN	NaN	29.7	NaN
8	NaN	NaN	NaN	30.1	NaN

```
df["Age"].median()
```

```
28.0
```

```
df["Age"].mean()
```

```
28.333333333333332
```

```
df["Age"].mode()
```

```
0    27
dtype: int64
```

```
df.var()
```

```
<ipython-input-10-28ded241fd7c>:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') i
  df.var()
Height    236.194444
Weight    115.361111
bmi         6.966944
Age         5.500000
dtype: float64
```

```
df.std()
```

```
<ipython-input-11-ce97bb7eaef8>:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') i
  df.std()
Height    15.368619
Weight    10.740629
bmi        2.639497
Age        2.345208
dtype: float64
```

```
M1=df.max()
M1
```

```
Gender    Male
Height     189
Weight      90
bmi       30.1
Age         32
dtype: object
```

```
M2=df.min()
M2
```

```
Gender    Female
Height       147
Weight        61
bmi         22.3
Age           25
dtype: object
```

```
df.describe()
```

	Height	Weight	bmi	Age
count	9.000000	9.000000	9.000000	9.000000
mean	168.222222	73.888889	26.177778	28.333333
std	15.368619	10.740629	2.639497	2.345208
min	147.000000	61.000000	22.300000	25.000000
25%	154.000000	65.000000	24.400000	27.000000
50%	174.000000	70.000000	26.100000	28.000000
75%	177.000000	80.000000	27.500000	30.000000
max	189.000000	90.000000	30.100000	32.000000

```
df.describe(include="all")
```

	Gender	Height	Weight	bmi	Age
count	9	9.000000	9.000000	9.000000	9.000000
unique	2	NaN	NaN	NaN	NaN
top	Male	NaN	NaN	NaN	NaN
freq	6	NaN	NaN	NaN	NaN
mean	NaN	168.222222	73.888889	26.177778	28.333333

```
df["Age"].describe()
```

```
count     9.000000
mean     28.333333
std       2.345208
min      25.000000
25%      27.000000
50%      28.000000
75%      30.000000
max      32.000000
Name: Age, dtype: float64
```

```
df["Age"].var()
```

```
5.5
```

```
df["Age"].std()
```

```
2.345207879911715
```

```
m2=df["Age"].min()
m2
```

```
25
```

```
m1=df["Age"].max()
m1
```

```
32
```

```
range=m1-m2
range
```

```
7
```

```
Q1=df.quantile(0.25)
Q1
```

```
Height    154.0
Weight     65.0
bmi        24.4
Age        27.0
Name: 0.25, dtype: float64
```

```
Q1=df["Age"].quantile(0.25)
Q1
```

```
27.0
```

```
Q3=df["Age"].quantile(0.75)
Q3
```

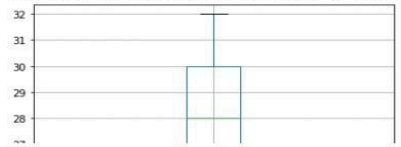
```
30.0
```

```
IQR=Q3-Q1
IQR
```

```
3.0
```

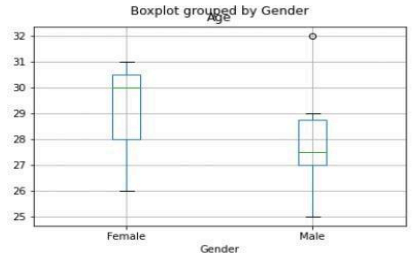
```
df.boxplot(column="Age")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd99716e340>
```




```
df.boxplot(column="Age",by="Gender")
```

```
/usr/local/lib/python3.8/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarning: Creating an ndarray from ragge
  X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
<matplotlib.axes._subplots.AxesSubplot at 0x7fd99726a610>
```




```
sd=df["Age"].std()
Kurtosis=df["Age"].kurtosis()
Skew=df["Age"].skew()
mean=df["Age"].mean()
DQ=sd/mean
Harmonic_Mean=stats.hmean(df["Age"])
Risk= Harmonic_Mean /mean
print("SD=",sd,"Kurtosis=",Kurtosis,"Skew=",Skew,"DQ=",DQ,"Risk=",Risk)
zscore=stats.zscore(df["Age"])
print(zscore)
```

```
SD= 2.345207879911715 Kurtosis= -1.041322314049585 Skew= 0.232582599660668 DQ= 0.08277204282041346 Risk= 0.9939701115655059
0   -1.507557
1   -0.603023
2    0.753778
3   -1.055290
4   -0.150756
5    0.301511
6    1.206045
7    1.658312
8   -0.603023
Name: Age, dtype: float64
```

5.Conclusion :- The measures of central tendency, measures of dispersion and measures of shape are explored on the given dataset.