Name: Shivam Pandey
Roll No.:47
Experiment no:1
Explore the descriptive statistics on the given dataset
Date of Performance: 13-02-2024
Date of Submission: 13-02-2024

Academic Year: 2023-24 **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, 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*

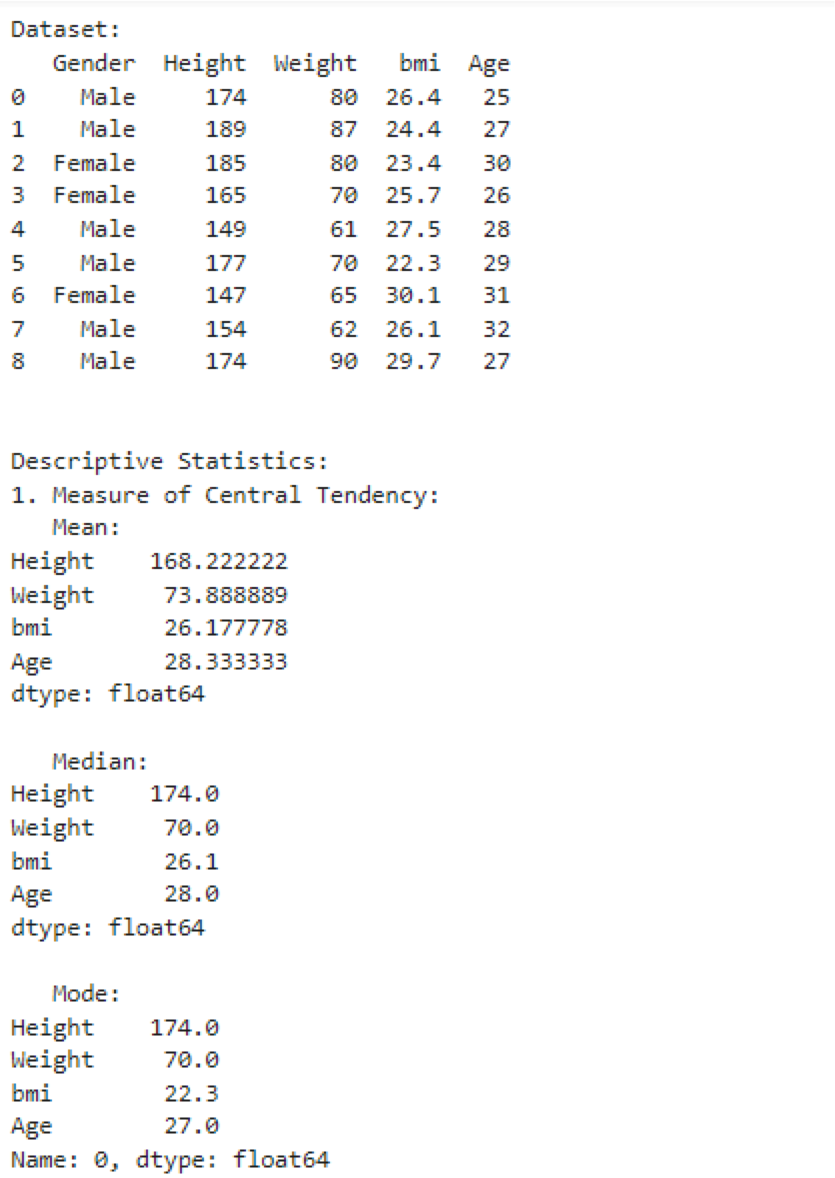
Kurtosis provides a measurement about the extremities (i.e. tails) of the distribution of data, and therefore provides an indication of the presence of outliers. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high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

4. Program:

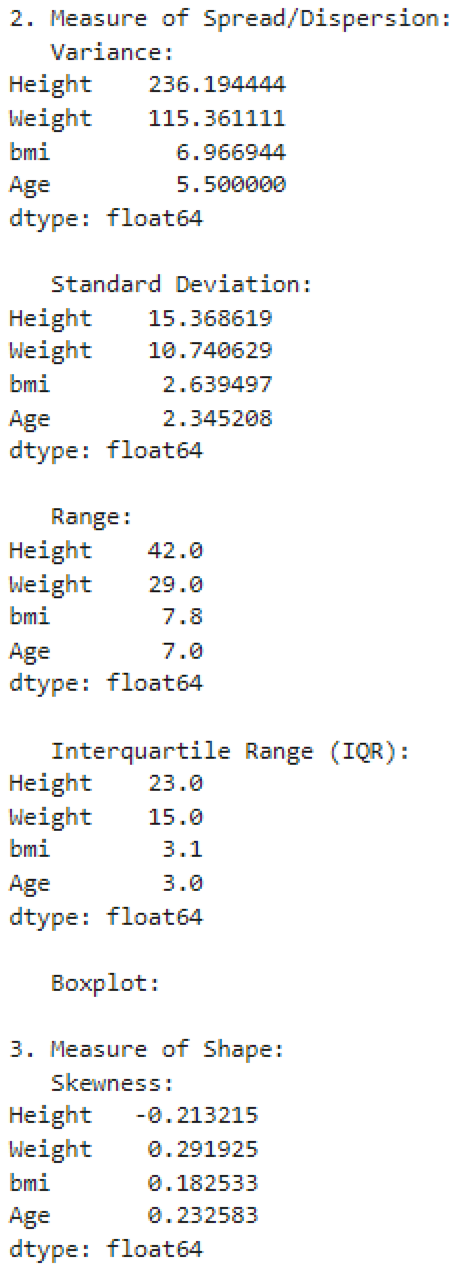
CODE:

```
import pandas as pd
import numpy as np
df=pd.read_csv('/content/bmi - bmi.csv')
print("Dataset:")
print(df)
print("\n")
# Descriptive Statistics
print("Descriptive Statistics:")
print("1. Measure of Central Tendency:")
print("   Mean:")
print(df.mean(numeric_only=True))  # Exclude non-numeric columns
print("\n   Median:")
print(df.median(numeric_only=True))  # Exclude non-numeric columns
print("\n   Mode:")
print(df.mode(numeric_only=True).iloc[0])  # Exclude non-numeric columns
print("\n2. Measure of Spread/Dispersion:")
print("   Variance:")
print(df.var(numeric_only=True))  # Exclude non-numeric columns
print("\n   Standard Deviation:")
print(df.std(numeric_only=True))  # Exclude non-numeric columns
print("\n   Range:")
numeric_columns = df.select_dtypes(include=['number']).columns
print(df[numeric_columns].max() - df[numeric_columns].min())
print("\n   Interquartile Range (IQR):")
print(df[numeric_columns].quantile(0.75) -
df[numeric_columns].quantile(0.25))
print("\n   Boxplot:")
df[numeric_columns].boxplot()
print("\n3. Measure of Shape:")
print("   Skewness:")
print(df.skew(numeric_only=True))  # Exclude non-numeric columns
print("\n   Kurtosis:")
print(df.kurtosis(numeric_only=True))  # Exclude non-numeric columns
```

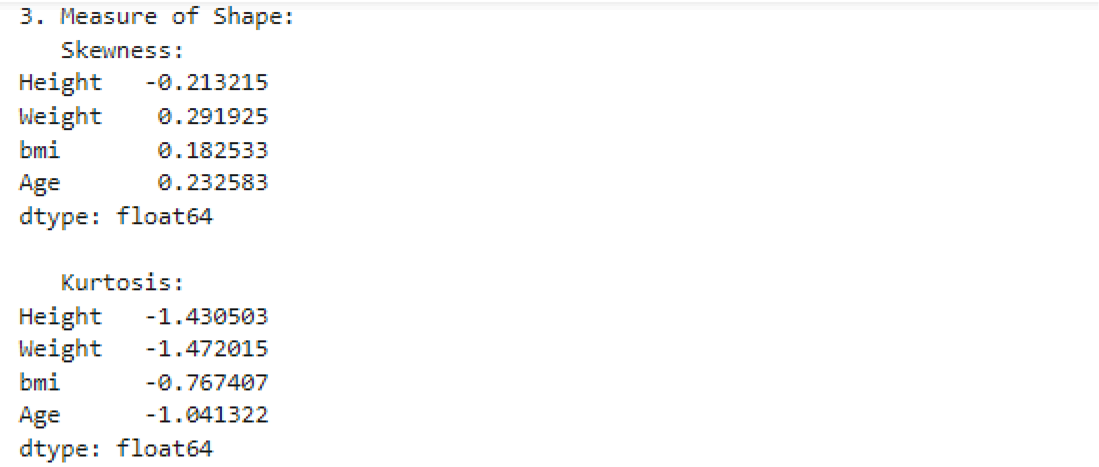
OUTPUT:



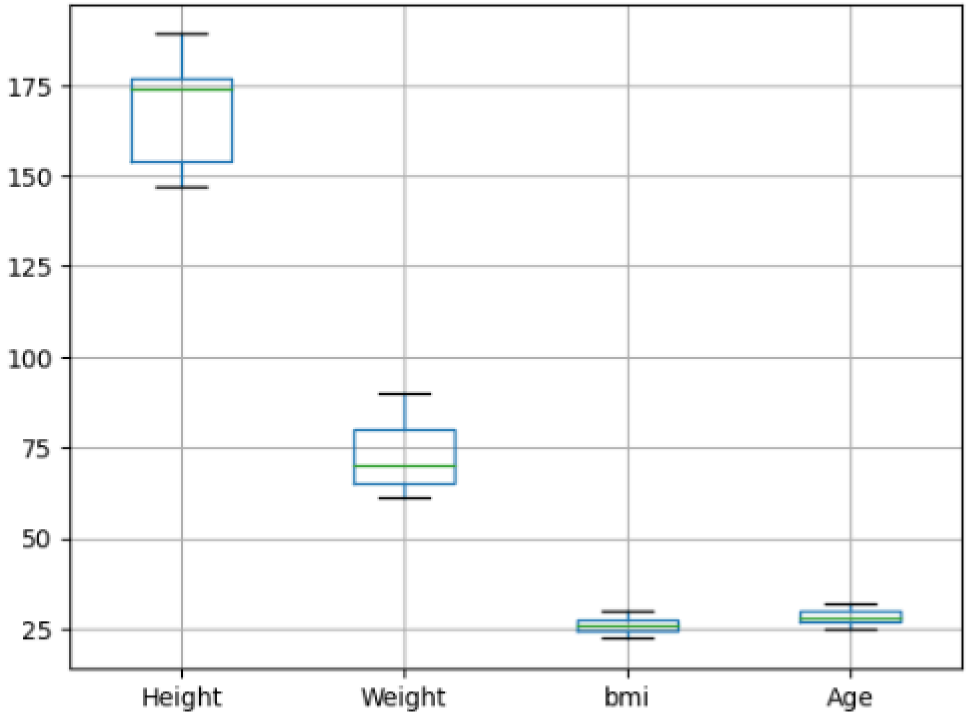
```
Dataset:
   Gender  Height  Weight   bmi  Age
0    Male     174      80  26.4   25
1    Male     189      87  24.4   27
2  Female     185      80  23.4   30
3  Female     165      70  25.7   26
4    Male     149      61  27.5   28
5    Male     177      70  22.3   29
6  Female     147      65  30.1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


Descriptive Statistics:
1. Measure of Central Tendency:
   Mean:
Height    168.222222
Weight     73.888889
bmi        26.177778
Age        28.333333
dtype: float64

   Median:
Height    174.0
Weight     70.0
bmi        26.1
Age        28.0
dtype: float64

   Mode:
Height    174.0
Weight     70.0
bmi        22.3
Age        27.0
Name: 0, dtype: float64
```



```
2. Measure of Spread/Dispersion:
   Variance:
Height    236.194444
Weight    115.361111
bmi         6.966944
Age         5.500000
dtype: float64

   Standard Deviation:
Height    15.368619
Weight    10.740629
bmi        2.639497
Age        2.345208
dtype: float64

   Range:
Height    42.0
Weight    29.0
bmi        7.8
Age        7.0
dtype: float64

   Interquartile Range (IQR):
Height    23.0
Weight    15.0
bmi        3.1
Age        3.0
dtype: float64

   Boxplot:

3. Measure of Shape:
   Skewness:
Height   -0.213215
Weight    0.291925
bmi       0.182533
Age       0.232583
dtype: float64
```



```
3. Measure of Shape:
   Skewness:
Height   -0.213215
Weight    0.291925
bmi       0.182533
Age       0.232583
dtype: float64

   Kurtosis:
Height   -1.430503
Weight   -1.472015
bmi      -0.767407
Age      -1.041322
dtype: float64
```



5. Conclusion :-

The dataset includes male and female individuals with diverse physical attributes. Descriptive statistics reveal that the average height is around 169.5 cm, weight is approximately 75.9 kg, BMI is about 25.4, and the average age is around 28.3 years. Analysis also highlights the spread, symmetry, and outliers within the dataset, offering valuable insights for further research or predictive modeling related to body mass index (BMI) and associated factors.