**AIM:** Explore the descriptive and inferential statistics on the given dataset.

**THEORY:**

The experiment aims to delve into both descriptive and inferential statistics on a given dataset. Descriptive statistics involve the organization, summarization, and interpretation of data to uncover patterns, trends, and characteristics. On the other hand, inferential statistics aim to make predictions and draw conclusions about a larger population based on a representative sample.

**Descriptive Statistics:**

- Descriptive statistics provide a **comprehensive overview of the dataset's main features**. This involves **measures of central tendency** such as mean, median, and mode, which offer insights into the typical value or centre of the data. Additionally, **measures of variability** like standard deviation and range help to understand the dispersion or spread of the data points.
- The experiment will explore the **distribution of the data**, displaying the frequency of different outcomes either numerically or graphically. **Central tendency** measures like mean, median, and mode will be **calculated to understand where the data tends to cluster**. **Variability** measures such as standard deviation will **provide insights into how much the data points deviate from the mean**.
- Moreover, considerations of kurtosis and skewness will provide information on the shape and symmetry of the dataset. **Kurtosis** indicates whether **extreme values exist in the tails** of the distribution, while **skewness measures the asymmetry of the data**.

**Inferential Statistics:**

- Inferential statistics **aim to conclude a larger population based on a sample from that population**. Techniques such as **regression analysis** will be employed to reveal relationships between independent and dependent variables within the dataset. This analysis helps predict the value of the dependent variable based on different values of the independent variables.
- Furthermore, **hypothesis tests** will be conducted to determine whether the relationships observed in the sample data hold for the entire dataset. These tests involve making educated guesses (hypotheses) about the population parameters and using statistical methods to assess the likelihood of these hypotheses being correct.

**CONCLUSION:** In conclusion, descriptive and inferential statistics play a crucial role in analyzing a dataset. Descriptive statistics provide insights into the characteristics of the data, while inferential statistics help to make predictions about the entire population based on the sample data. By conducting hypothesis tests, we can assess the likelihood of the relationships observed in the sample data being correct for the entire dataset. These statistical techniques help to draw informed conclusions and make predictions that can be useful in various fields.

# ADS EXP 1

## Importing required libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

## Printing the CSV data

```
data = pd.read_csv('ADS_Exp_1_Dataset.csv')
data.head(10)
```

| | Timestamp | Email Address | Full Name | Division | Have you opted Applied Data Science ? | Have attended ADS lectures ? | at A |
|---|---|---|---|---|---|---|---|
| 0 | 1/22/2024 22:06:29 | 2020.shubham.gupta@ves.ac.in | Shubham Gupta | D17C | Yes | Yes | |
| 1 | 1/22/2024 22:08:14 | 2020.siddhant.kodolkar@ves.ac.in | Siddhant Kodolkar | D17C | Yes | Yes | |
| 2 | 1/22/2024 22:09:14 | 2020.mansi.bellani@ves.ac.in | Mansi Bellani | D17C | Yes | Yes | |
| 3 | 1/22/2024 22:09:21 | 2020.aditya.mundas@ves.ac.in | Aditya Mundas | D17C | Yes | Yes | |
| 4 | 1/22/2024 22:09:47 | 2020.vishakha.kulkarni@ves.ac.in | Vishakha Kulkarni | D17C | Yes | Yes | |
| 5 | 1/22/2024 22:10:49 | 2020.mihir.bhatkar@ves.ac.in | Mihir Bhatkar | D17C | Yes | Yes | |
| 6 | 1/22/2024 22:10:56 | 2020.sahil.kishnani@ves.ac.in | Sahil Kishnani | D17C | Yes | Yes | |
| 7 | 1/22/2024 22:11:39 | 2020.harsh.karira@ves.ac.in | Harsh Shankar Karira | D17C | Yes | Yes | |
| 8 | 1/22/2024 22:12:32 | 2020.khusboo.kimtani@ves.ac.in | Khusboo Harpal Kimtani | D17C | No | NaN | |
| 9 | 1/22/2024 22:14:54 | 2020.sachin.choudhary@ves.ac.in | Sachin Choudhary | D17C | Yes | Yes | |

10 rows × 24 columns

## Statiscal measures

```
data.describe()
```

| | What number of lecture did you attend? | What number of labs did you attend? | How well did you understand the concept of applied data science ? | How well did you understand the difference between data science and data analytics ? | How well did you understand the data wrangling? | How well do you know about data lakes and different tools about data analytics? | How well did you understand the concept of mean median and mode |
|---|---|---|---|---|---|---|---|
| count | 34.000000 | 34.000000 | 34.000000 | 34.000000 | 34.000000 | 34.000000 | 34.000000 |
| mean | 3.411765 | 1.147059 | 3.764706 | 3.970588 | 3.323529 | 3.529412 | 4.441176 |
| std | 0.957194 | 0.702047 | 1.046171 | 0.904041 | 0.976096 | 0.991946 | 0.785905 |
| min | 2.000000 | 0.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 3.000000 |
| 25% | 3.000000 | 1.000000 | 3.250000 | 3.250000 | 3.000000 | 3.000000 | 4.000000 |
| 50% | 3.000000 | 1.000000 | 4.000000 | 4.000000 | 3.000000 | 3.500000 | 5.000000 |
| 75% | 4.000000 | 2.000000 | 4.000000 | 5.000000 | 4.000000 | 4.000000 | 5.000000 |
| max | 5.000000 | 2.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 |

## Checking for missing values

```
print(data.isnull().sum())
```

```
Timestamp                                                                       0
Email Address                                                                   0
Full Name                                                                       0
Division                                                                        0
Have you opted Applied Data Science ?                                           0
Have attended ADS lectures ?                                                    6
Have attended ADS Lab ?                                                         6
What number of lecture did you attend?                                          6
What number of labs did you attend?                                             6
How well did you understand the concept of applied data science ?               6
How well did you understand the difference between data science and data analytics ?    6
How well did you understand the data wrangling?                                 6
How well do you know about data lakes and different tools about data analytics?         6
How well did you understand the concept of mean, median and mode?               6
How well did you understand the concept of cumulative frequency?                6
How well did you understand features related to iris flower dataset?            6
How well are you able to solves numerical problems based on cumulative frequency?       6
How well are you able to understand the different graphs like bar and histogram ?       6
How well are you able to find arithmetic median based on class interval ?       6
Explains concepts in understandable way                                         6
Solves doubts willingly                                                         6
How is the structuring of course                                                6
How well was the use of presentation                                            6
Provide support for student going above and beyond                              6
dtype: int64
```

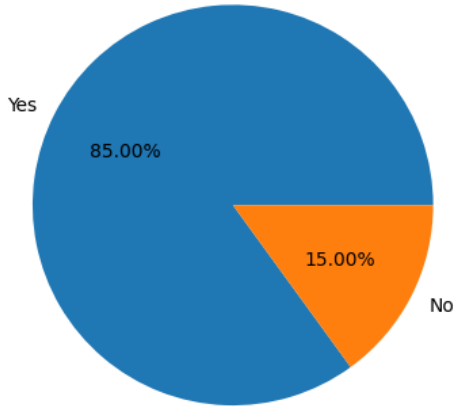
## Plotting the pie chart

```
opted = data['Have you opted Applied Data Science ?'].value_counts()
opted
```

```
Yes    34
No      6
Name: Have you opted Applied Data Science ?, dtype: int64
```

```
plt.pie(opted,labels=['Yes','No'],autopct="%0.2f%%")
```

```
([<matplotlib.patches.Wedge at 0x7830f9655b40>,
  <matplotlib.patches.Wedge at 0x7830f9655a50>],
 [Text(-0.9801072140121811, 0.4993894763020951, 'Yes'),
  Text(0.980107260768394, -0.4993893845378326, 'No')],
 [Text(-0.5346039349157351, 0.27239425980114274, '85.00%'),
  Text(0.534603960419124, -0.2723942097479087, '15.00%')])
```




## Cleaning the missing values

```
data = data.dropna()
print(data.isnull().sum())
```

```
Timestamp                                                                   0
Email Address                                                               0
Full Name                                                                   0
Division                                                                    0
Have you opted Applied Data Science ?                                       0
Have attended ADS lectures ?                                                0
Have attended ADS Lab ?                                                     0
What number of lecture did you attend?                                      0
What number of labs did you attend?                                         0
How well did you understand the concept of applied data science ?           0
How well did you understand the difference between data science and data analytics ?    0
How well did you understand the data wrangling?                             0
How well do you know about data lakes and different tools about data analytics?         0
How well did you understand the concept of mean, median and mode?           0
How well did you understand the concept of cumulative frequency?            0
How well did you understand features related to iris flower dataset?        0
How well are you able to solves numerical problems based on cumulative frequency?       0
How well are you able to understand the different graphs like bar and histogram ?       0
How well are you able to find arithmetic median based on class interval ?   0
Explains concepts in understandable way                                     0
Solves doubts willingly                                                     0
How is the structuring of course                                            0
How well was the use of presentation                                        0
Provide support for student going above and beyond                          0
dtype: int64
```

## Descriptive Analysis

```
selected_columns = [
    'How well did you understand the concept of applied data science ?',
    'How well did you understand the difference between data science and data analytics ?',
    'How well did you understand the data wrangling?',
    'How well do you know about data lakes and different tools about data analytics?',
    'How well did you understand the concept of mean, median and mode?',
    'How well did you understand the concept of cumulative frequency?',
    'How well did you understand features related to iris flower dataset?',
    'How well are you able to solves numerical problems based on cumulative frequency?',
    'How well are you able to understand the different graphs like bar and histogram ?',
    'How well are you able to find arithmetic median based on class interval ?',
]
selected_data = data[selected_columns]
```

```
selected_data_mean = selected_data.mean()
selected_data_mean
```

```
How well did you understand the concept of applied data science ?                    3.764706
How well did you understand the difference between data science and data analytics ?  3.970588
How well did you understand the data wrangling?                                      3.323529
How well do you know about data lakes and different tools about data analytics?      3.529412
How well did you understand the concept of mean, median and mode?                    4.441176
How well did you understand the concept of cumulative frequency?                     4.117647
How well did you understand features related to iris flower dataset?                 3.882353
How well are you able to solves numerical problems based on cumulative frequency?   3.970588
How well are you able to understand the different graphs like bar and histogram ?    4.147059
How well are you able to find arithmetic median based on class interval ?            4.147059
dtype: float64
```

```
selected_data_median = selected_data.median()
selected_data_median
```

```
How well did you understand the concept of applied data science ?                    4.0
How well did you understand the difference between data science and data analytics ?  4.0
How well did you understand the data wrangling?                                      3.0
How well do you know about data lakes and different tools about data analytics?      3.5
How well did you understand the concept of mean, median and mode?                    5.0
How well did you understand the concept of cumulative frequency?                     4.0
How well did you understand features related to iris flower dataset?                 4.0
How well are you able to solves numerical problems based on cumulative frequency?   4.0
How well are you able to understand the different graphs like bar and histogram ?    4.0
How well are you able to find arithmetic median based on class interval ?            4.0
dtype: float64
```

```
selected_data_mode = selected_data.mode()
selected_data_mode
```

| How well did you understand the concept of applied data science ? | How well did you understand the difference between data science and data analytics ? | How well did you understand the data wrangling? | How well do you know about data lakes and different tools about data analytics? | How well did you understand the concept of mean, median and mode? | How well did you understand the concept of cumulative frequency? | How well did you understand features related to iris flower dataset? |
|---|---|---|---|---|---|---|

```
selected_data.std()
```

```
How well did you understand the concept of applied data science ?                    1.046171
How well did you understand the difference between data science and data analytics ?  0.904041
How well did you understand the data wrangling?                                      0.976096
How well do you know about data lakes and different tools about data analytics?      0.991946
How well did you understand the concept of mean, median and mode?                    0.785905
How well did you understand the concept of cumulative frequency?                     1.007989
How well did you understand features related to iris flower dataset?                 1.094468
How well are you able to solves numerical problems based on cumulative frequency?   0.936961
How well are you able to understand the different graphs like bar and histogram ?    1.018982
How well are you able to find arithmetic median based on class interval ?            0.925476
dtype: float64
```

```
selected_data.var()
```

```
How well did you understand the concept of applied data science ?                    1.094474
How well did you understand the difference between data science and data analytics ?  0.817291
How well did you understand the data wrangling?                                      0.952763
How well do you know about data lakes and different tools about data analytics?      0.983957
How well did you understand the concept of mean, median and mode?                    0.617647
How well did you understand the concept of cumulative frequency?                     1.016043
How well did you understand features related to iris flower dataset?                 1.197861
How well are you able to solves numerical problems based on cumulative frequency?   0.877897
How well are you able to understand the different graphs like bar and histogram ?    1.038324
How well are you able to find arithmetic median based on class interval ?            0.856506
dtype: float64
```

```
selected_data.max()
```

```
How well did you understand the concept of applied data science ?                    5.0
How well did you understand the difference between data science and data analytics ?  5.0
How well did you understand the data wrangling?                                      5.0
How well do you know about data lakes and different tools about data analytics?      5.0
How well did you understand the concept of mean, median and mode?                    5.0
How well did you understand the concept of cumulative frequency?                     5.0
How well did you understand features related to iris flower dataset?                 5.0
How well are you able to solves numerical problems based on cumulative frequency?   5.0
How well are you able to understand the different graphs like bar and histogram ?    5.0
```

```
How well are you able to find arithmetic median based on class interval ?              5.0
dtype: float64
```

```
selected_data.min()
```

```
How well did you understand the concept of applied data science ?                      1.0
How well did you understand the difference between data science and data analytics ?   1.0
How well did you understand the data wrangling?                                        1.0
How well do you know about data lakes and different tools about data analytics?        1.0
How well did you understand the concept of mean, median and mode?                      3.0
How well did you understand the concept of cumulative frequency?                       1.0
How well did you understand features related to iris flower dataset?                   1.0
How well are you able to solves numerical problems based on cumulative frequency?     1.0
How well are you able to understand the different graphs like bar and histogram ?     1.0
How well are you able to find arithmetic median based on class interval ?              1.0
dtype: float64
```

```
range = selected_data.max() - selected_data.min()
range
```

```
How well did you understand the concept of applied data science ?                      4.0
How well did you understand the difference between data science and data analytics ?   4.0
How well did you understand the data wrangling?                                        4.0
How well do you know about data lakes and different tools about data analytics?        4.0
How well did you understand the concept of mean, median and mode?                      2.0
How well did you understand the concept of cumulative frequency?                       4.0
How well did you understand features related to iris flower dataset?                   4.0
How well are you able to solves numerical problems based on cumulative frequency?     4.0
How well are you able to understand the different graphs like bar and histogram ?     4.0
How well are you able to find arithmetic median based on class interval ?              4.0
dtype: float64
```

```
selected_data.kurtosis()
```

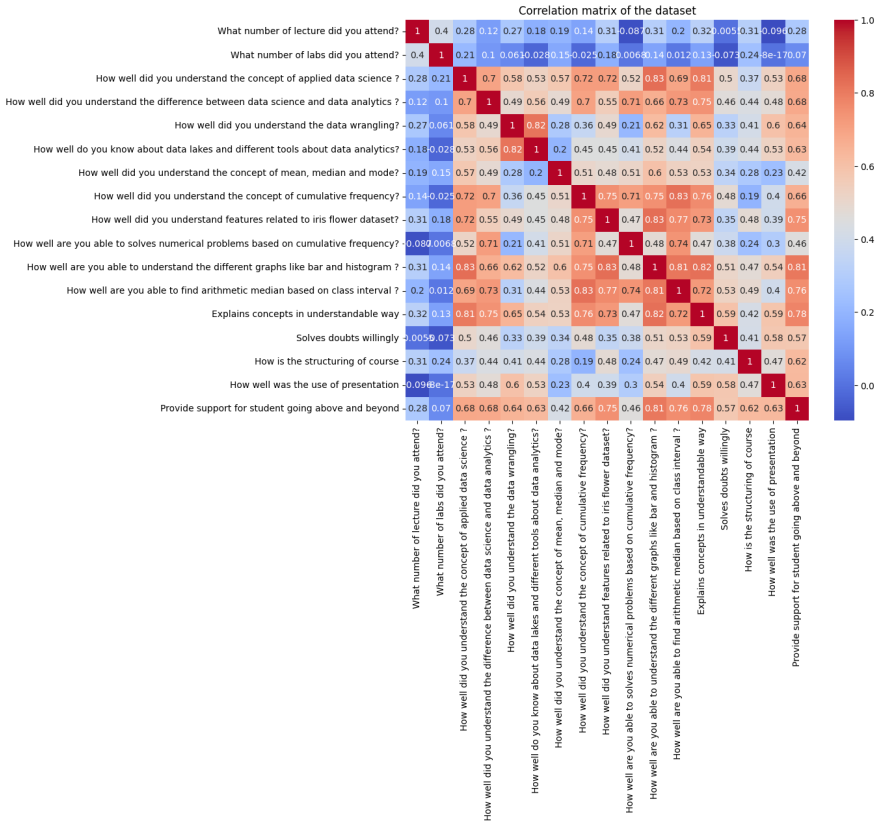
```
How well did you understand the concept of applied data science ?                      1.389446
How well did you understand the difference between data science and data analytics ?   1.941915
How well did you understand the data wrangling?                                        0.465444
How well do you know about data lakes and different tools about data analytics?       -0.027929
How well did you understand the concept of mean, median and mode?                     -0.606549
How well did you understand the concept of cumulative frequency?                       1.391304
How well did you understand features related to iris flower dataset?                   1.064355
How well are you able to solves numerical problems based on cumulative frequency?     1.338851
How well are you able to understand the different graphs like bar and histogram ?     1.890013
How well are you able to find arithmetic median based on class interval ?              2.449309
dtype: float64
```

```
selected_data.skew()
```

```
How well did you understand the concept of applied data science ?                     -1.182768
How well did you understand the difference between data science and data analytics ?  -0.985669
How well did you understand the data wrangling?                                       -0.507812
How well do you know about data lakes and different tools about data analytics?       -0.284218
How well did you understand the concept of mean, median and mode?                     -0.988052
How well did you understand the concept of cumulative frequency?                      -1.191141
How well did you understand features related to iris flower dataset?                  -1.080413
How well are you able to solves numerical problems based on cumulative frequency?    -0.878474
How well are you able to understand the different graphs like bar and histogram ?    -1.406668
How well are you able to find arithmetic median based on class interval ?             -1.282631
dtype: float64
```

```
correlation_matrix = data.corr()
# Plot the correlation matrix as a heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation matrix of the dataset')
plt.show()
```

```
<ipython-input-103-c79572ca6bba>:1: FutureWarning: The default value of numeric_only
  correlation_matrix = data.corr()
```




## Inferential Analysis

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.impute import SimpleImputer
```

```
column_means = data.mean()
top_5_means = column_means.nlargest(5)
top_5_means
```

```
<ipython-input-105-3835f8382530>:1: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a future
  column_means = data.mean()
How well did you understand the concept of mean, median and mode?                  4.441176
How well are you able to understand the different graphs like bar and histogram ?  4.147059
How well are you able to find arithmetic median based on class interval ?          4.147059
How well did you understand the concept of cumulative frequency?                   4.117647
How well was the use of presentation                                               4.000000
dtype: float64
```

```
data['Target'] = data[top_5_means.index].mean(axis=1)
```

```
X = data[top_5_means.index]
y = data['Target']
```

```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.25,random_state=42)
model=LinearRegression()
model.fit(X_train,y_train)
# Making predictions on the test set
y_pred = model.predict(X_test)
```

```
# Evaluating the performance of the model
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred, squared=False))
```

```
Mean Absolute Error: 1.3816108750890837e-15
Mean Squared Error: 2.454233927354259e-30
Root Mean Squared Error: 1.5665994789205884e-15
```

```
plt.scatter(y_test, y_pred)
```

```
<matplotlib.collections.PathCollection at 0x7830fa0ce6e0>
```