In [13]:
```
import pandas as pd
import numpy as np
from scipy import stats
# Load the dataset into Python (replace 'your_dataset.csv' with your file path or use an online dataset)
data=pd.read_csv("C:/Users/kovva/Downloads/dataset.csv")
data.head(5)
```

Out[13]:

| | S/N | Year | Age | Menopause | Tumor Size (cm) | Inv-Nodes | Breast | Metastasis | Breast Quadrant | History | Diagnosis Result |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2019 | 40 | 1 | 2 | 0 | Right | 0 | Upper inner | 0 | Benign |
| **1** | 2 | 2019 | 39 | 1 | 2 | 0 | Left | 0 | Upper outer | 0 | Benign |
| **2** | 3 | 2019 | 45 | 0 | 4 | 0 | Left | 0 | Lower outer | 0 | Benign |
| **3** | 4 | 2019 | 26 | 1 | 3 | 0 | Left | 0 | Lower inner | 1 | Benign |
| **4** | 5 | 2019 | 21 | 1 | 1 | 0 | Right | 0 | Upper outer | 1 | Benign |

In [7]:
```
temp={
    'Year':data["Year"],
    'Age':data["Age"],
    'Tumer':data["Tumor Size (cm)"]
}
new_data=pd.DataFrame(temp)
new_data.head(5)
```

Out[7]:

| | Year | Age | Tumer |
|---|---|---|---|
| **0** | 2019 | 40 | 2 |
| **1** | 2019 | 39 | 2 |
| **2** | 2019 | 45 | 4 |
| **3** | 2019 | 26 | 3 |
| **4** | 2019 | 21 | 1 |

## Calculate basic descriptive statistics

In [10]:
```
print("Mean:\n", new_data.mean())
print("\nMedian:\n", new_data.median())
print("\nMode:\n", new_data.mode().iloc[0])
print("\nStandard Deviation:\n", new_data.std())
print("\nVariance:\n",new_data.var())
```

```
Mean:
 Year     2019.521127
Age         39.826291
Tumer        4.262911
dtype: float64

Median:
 Year     2020.0
Age        40.0
Tumer        4.0
dtype: float64

Mode:
 Year     2020
Age        38
Tumer       3
Name: 0, dtype: int64

Standard Deviation:
 Year       0.500730
Age       14.092781
Tumer      2.567281
dtype: float64

Variance:
 Year        0.250731
Age       198.606475
Tumer       6.590929
dtype: float64
```

## Additional descriptive statistics

In [20]:
```
print("\nRange:\n",new_data.max() - new_data.min())
print("\nSkewness:\n",new_data.skew())
print("\nKurtosis:\n",new_data.kurt())
```

```
Range:
 Year      1
Age      64
Tumer    13
dtype: int64

Skewness:
 Year    -0.085184
Age       0.068217
Tumer     0.834117
dtype: float64

Kurtosis:
 Year    -2.011723
Age     -0.642520
Tumer    0.418773
dtype: float64
```

In [14]:
```
age_sample=new_data["Age"]

# Hypothetical population mean for BMI
population_mean = 0.05

# Perform one-sample t-test
t_stat, p_value = stats.ttest_1samp(age_sample, population_mean)

print(f"T-Statistic: {t_stat}")
print(f"P-Value: {p_value}")
```

```
T-Statistic: 41.19242738596511
P-Value: 3.928075881040218e-103
```

## Confidence Intervals

In [15]:
```
sample_mean = np.mean(age_sample)
standard_error = stats.sem(age_sample)

# Compute 95% confidence interval for BMI
confidence_interval = stats.norm.interval(0.95, loc=sample_mean, scale=standard_error)

print(f"95% Confidence Interval for BMI: {confidence_interval}")
```

```
95% Confidence Interval for BMI: (np.float64(37.933707833271704), np.float64(41.71887432635271))
```

## Regression Analysis

In [19]:
```
import statsmodels.api as sm

# Define independent variable (add constant for intercept)
X = sm.add_constant(new_data['Age'])

# Define dependent variable
y = new_data['Tumer']

# Fit linear regression model
model = sm.OLS(y, X).fit()

# Print model summary
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Tumer   R-squared:                       0.247
Model:                            OLS   Adj. R-squared:                  0.243
Method:                 Least Squares   F-statistic:                     69.05
Date:                Thu, 05 Sep 2024   Prob (F-statistic):           1.16e-14
Time:                        20:03:22   Log-Likelihood:                -472.41
No. Observations:                 213   AIC:                             948.8
Df Residuals:                     211   BIC:                             955.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.6603      0.460      1.436      0.152      -0.246       1.567
Age            0.0905      0.011      8.310      0.000       0.069       0.112
==============================================================================
Omnibus:                       16.484   Durbin-Watson:                   1.781
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.680
Skew:                           0.614   Prob(JB):                     8.78e-05
Kurtosis:                       3.772   Cond. No.                         127.
```

==========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.