Statistic and Data Analysis - Basic

Bivariate Analysis

This report aims to investigate if there is a relationship between age and systolic blood pressure levels. Correlation coefficient is a commonly used approach. This may help to identify whether the two variables have a linear relationship.

Data URL: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data

Step-01: Import Data- 374 data was imported

```
[7] #Import libraries
    import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
```

```
[8] #Load data set
    file=pd.read_csv("/content/Sleep_health_and_lifestyle_dataset.csv")
    file.head()
```

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea

Step-02: Data preparation

```
#I am going to check the corelation between Age and Systolic blood pressure
#Therefore i have removed following columns
NewData=file.drop(columns=['Person ID', 'Gender','Occupation','BMI Category','Quality of Sleep',"Sleep Duration","Stress Level","Heart Rate","Daily
NewData.head()
```

	Age	Blood Pressure
0	27	126/83
1	28	125/80
2	28	125/80
3	28	140/90
4	28	140/90

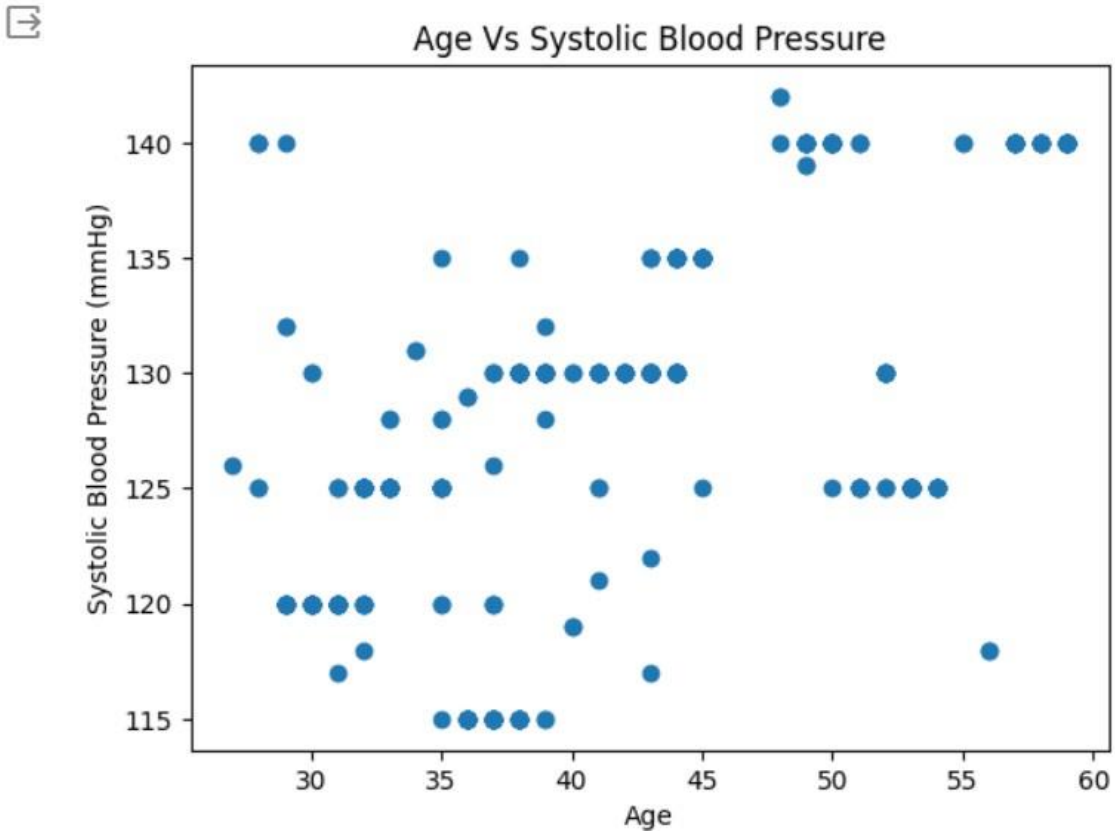
Next steps: View recommended plots

```
[51] # Check for missing values and handle them appropriately
     missing_values = NewData.isnull().sum()
     print(missing_values)

     Age              0
     Blood Pressure   0
     dtype: int64
```

First I have removed unnecessary columns from the dataset. After that missing value has been checked. Systolic blood pressure value was extracted from the blood pressure column to get deeper understanding of analysis.

Step-03: The scatter plot was drawn using imported data.

```
plt.scatter(NewData['Age'], NewData['Systolic'])
plt.xlabel('Age')
plt.ylabel('Systolic Blood Pressure (mmHg)')
plt.title('Age Vs Systolic Blood Pressure')
plt.show()
```



In this scatter plot,

X-axis represents Age

Y-axis represents Systolic blood pressure

Each data point represents a person.

Step-04: Calculate Correlation Coefficient

```
x=NewData['Age']
y=NewData['Systolic']
r=np.corrcoef(x,y)
print("The correlation between age and systolic blood pressure is approximately ",r[1,0])
```

```
The correlation between age and systolic blood pressure is approximately  0.6058784440490963
```

The correlation value of 0.606 shows that there is a connection between age and systolic blood pressure, which is positive. This means that there is a noticeable tendency for systolic blood pressure to increase as age increases.

Calculate the regression line's equation to determine its slope.

Y = B0 + B1X

Y	dependent variable
X	independent variable
B0	is a constant
B1	the regression coefficient

```
x=NewData['Age']
y=NewData['Systolic']
x1 = sm.add_constant(x)
results = sm.OLS(y, x1).fit()
results.summary()
```

OLS Regression Results

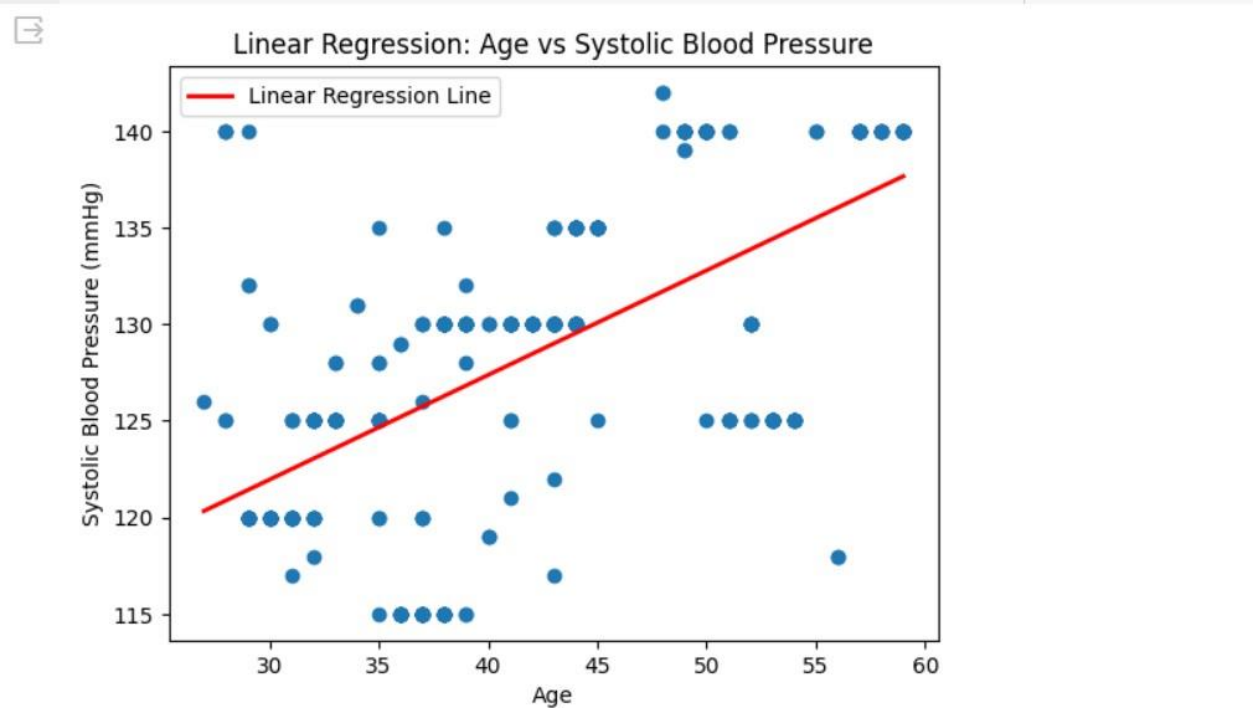
Dep. Variable:	Systolic	**R-squared:**	0.367
Model:	OLS	**Adj. R-squared:**	0.365
Method:	Least Squares	**F-statistic:**	215.8
Date:	Sun, 25 Feb 2024	**Prob (F-statistic):**	7.62e-39
Time:	01:26:59	**Log-Likelihood:**	-1210.4
No. Observations:	374	**AIC:**	2425.
Df Residuals:	372	**BIC:**	2433.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	105.7207	1.587	66.622	0.000	102.600	108.841
Age	0.5413	0.037	14.689	0.000	0.469	0.614

Omnibus:	7.988	**Durbin-Watson:**	0.871
Prob(Omnibus):	0.018	**Jarque-Bera (JB):**	7.892
Skew:	-0.347	**Prob(JB):**	0.0193
Kurtosis:	3.153	**Cond. No.**	214.

Y= 105.72 + (0.5413) * X

```
x=NewData['Age']
y=NewData['Systolic']
plt.scatter(x, y)
yreg=105.72 + (0.5413) * x
plt.plot(x, yreg, color='red', linewidth=2, label='Linear Regression Line')
plt.xlabel('Age')
plt.ylabel('Systolic Blood Pressure (mmHg)')
plt.title('Linear Regression: Age vs Systolic Blood Pressure')
plt.legend()

# Show plot
plt.show()
```



It reveals that systolic blood pressure increases with age. It also indicates that age is a key factor in determining blood pressure changes.

Multivariate Analysis

A study aims to investigate if there is a relationship between systolic blood pressure levels and multiple variables like age, gender, BMI.

Step-01: Import libraries and Data- 374 data was imported

```
[37] #Import libraries
     import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import statsmodels.api as sm
     import seaborn as sns
```

```
[38] #Load data set
     file=pd.read_csv("/content/Sleep_health_and_lifestyle_dataset.csv")
     file.head()
```

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea

Next steps: View recommended plots

Activate Windows
Go to Settings to activate Windows.

Step-02: Data preparation

```
[39] #remove following columns
     NewData=file.drop(columns=['Person ID','Occupation', 'Quality of Sleep',"Sleep Duration","Heart Rate","Daily S
     NewData.head()
```

	Gender	Age	Stress Level	BMI Category	Blood Pressure
0	Male	27	6	Overweight	126/83
1	Male	28	8	Normal	125/80
2	Male	28	8	Normal	125/80
3	Male	28	8	Obese	140/90
4	Male	28	8	Obese	140/90

Next steps: View recommended plots

```
[40] # Check for missing values and handle them appropriately
     missing_values = NewData.isnull().sum()
     print(missing_values)
```

```
Gender            0
Age               0
Stress Level      0
BMI Category      0
Blood Pressure    0
dtype: int64
```

Unnecessary columns were removed from the dataset. BMI category had same value with different name like “Normal” and “Normal Weight”. “Normal Weight” was replaced by “Normal”. 21 columns were replaced. Systolic blood pressure value was extracted from the blood pressure column to get deeper understanding of analysis.

Step-03: Calculate correlation matrix

- BMI function was created to convert categorical data into numerical. Average BMI value was considered for each category.

```
[106] #BMI_Category is a categorical variable. This function has written to make it numeric. Here average value was considered
      def BMI(Category):
          if Category =="Normal":
              return 21.7
          elif Category=="Overweight":
              return 27.45
          elif Category == "Obese":
              return 34.95
          else:
              return 18.5
```

- Gender function was created to convert categorical data into numerical.

```
[107] #Gender is a categorical variable. This function has written to make it numeric
      def Gender(Category):
          if Category =="Male":
              return 1
          else:
              return 0
```

- Correlation matrix was created

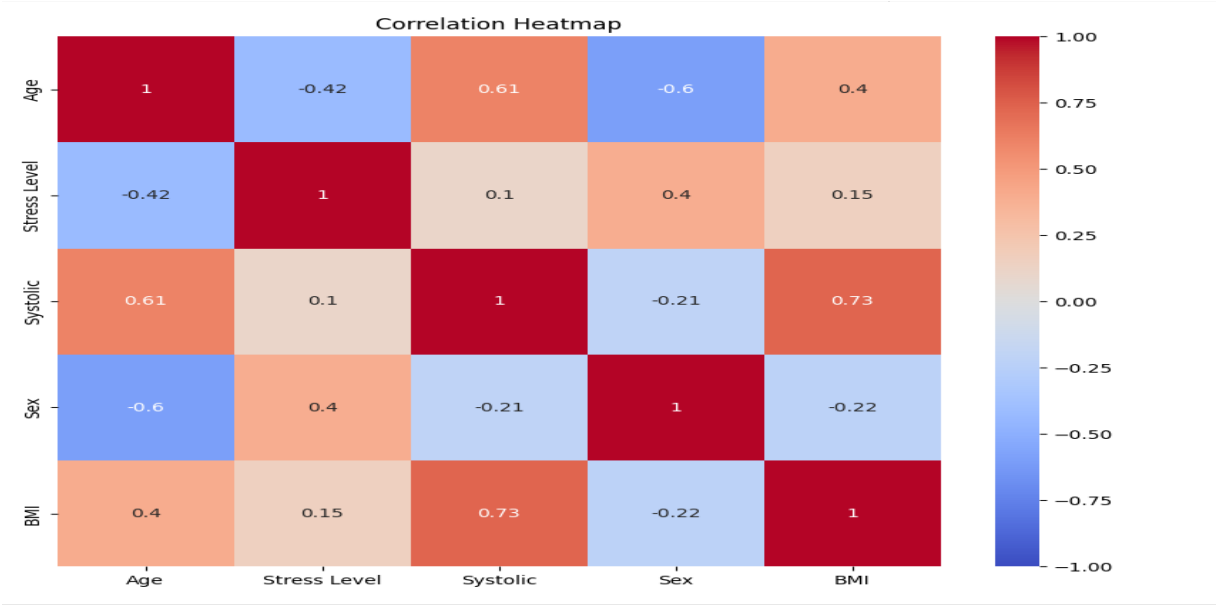
```
[112] #Correlation matrix was created
      df = pd.DataFrame(NewData)

      correlation_matrix = df.corr()

      # Print the correlation matrix
      print("Correlation Matrix:")
      print(correlation_matrix)
```

```
Correlation Matrix:
                   Age  Stress Level  Systolic       Sex       BMI
Age           1.000000     -0.422344  0.605878 -0.596358  0.402452
Stress Level -0.422344      1.000000  0.102818  0.396018  0.153393
Systolic      0.605878      0.102818  1.000000 -0.210527  0.727903
Sex          -0.596358      0.396018 -0.210527  1.000000 -0.221098
BMI           0.402452      0.153393  0.727903 -0.221098  1.000000
```

Interpretation

- **Age:** has a positive correlation with 'Systolic' (0.605878) and 'BMI' (0.402452), suggesting that older individuals tend to have higher systolic blood pressure and higher BMI.
- **Stress Level:** The correlation with 'Systolic' (0.102818) and 'BMI' (0.153393) is relatively low
- **Systolic BP:** has a strong positive correlation with 'BMI' (0.727903), indicating that higher systolic blood pressure tends to be associated with higher BMI. Systolic BP has a negative correlation with 'Sex' (-0.210527), indicating that lower systolic blood pressure tends to be associated with being categorized as 'Female'. In this dataset Female is 0 and Male is 1.

```
#Draw heat map
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Correlation Heatmap')
plt.show()
```



In the heat map:

- Dark red indicates a strong positive correlation.
- Dark blue indicates a strong negative correlation.
- Light colors (near white) indicate little to no correlation.

Interpretation

- 'Age' and 'Systolic' have a relatively strong positive correlation (dark red).
- 'BMI' and 'Systolic' also have a strong positive correlation (dark red), indicating that higher BMI tends to be associated with higher systolic blood pressure.