STATISTICAL ANALYSIS

In [1]:

```
import pandas as pd
```

In [26]:

```
df=pd.read_csv("Sleep_health_and_lifestyle_dataset.csv")
df.head()
```

Out[26]:

	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	NaN
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
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

In [27]:

```
df.info
```

Out[27]:

```
<bound method DataFrame.info of      Person ID  Gender  Age            Occupation  Sleep Duration  \
0            1    Male   27     Software Engineer             6.1   
1            2    Male   28                Doctor             6.2   
2            3    Male   28                Doctor             6.2   
3            4    Male   28  Sales Representative             5.9   
4            5    Male   28  Sales Representative             5.9   
..         ...     ...  ...                   ...             ...   
369        370  Female   59                 Nurse             8.1   
370        371  Female   59                 Nurse             8.0   
371        372  Female   59                 Nurse             8.1   
372        373  Female   59                 Nurse             8.1   
373        374  Female   59                 Nurse             8.1   

     Quality of Sleep  Physical Activity Level  Stress Level BMI Category  \
0                   6                       42             6   Overweight   
1                   6                       60             8       Normal   
2                   6                       60             8       Normal   
3                   4                       30             8        Obese   
4                   4                       30             8        Obese   
..                ...                      ...           ...          ...   
369                 9                       75             3   Overweight   
370                 9                       75             3   Overweight   
371                 9                       75             3   Overweight   
372                 9                       75             3   Overweight   
373                 9                       75             3   Overweight   

    Blood Pressure  Heart Rate  Daily Steps Sleep Disorder  
0           126/83          77         4200            NaN  
1           125/80          75        10000            NaN  
2           125/80          75        10000            NaN  
3           140/90          85         3000    Sleep Apnea  
4           140/90          85         3000    Sleep Apnea  
..             ...         ...          ...            ...  
369         140/95          68         7000    Sleep Apnea  
370         140/95          68         7000    Sleep Apnea  
371         140/95          68         7000    Sleep Apnea  
372         140/95          68         7000    Sleep Apnea  
373         140/95          68         7000    Sleep Apnea  

[374 rows x 13 columns]>
```

DESCRIPTIVE STATISTICS

- mean,median,mode,standard deviation,variance,skewness,kurtosis

In [38]:

```
print('\nMean of Physical Activity Level:',df['Physical Activity Level'].mean())
print('\nMedian of Stress Level:',df['Stress Level'].median())
print('\nMode of Quality of Sleep:',df['Quality of Sleep'].mode())
print('\nStandard deviation of Heart Rate:',df['Heart Rate'].std())
print('\nVariance of Daily Steps:',df['Daily Steps'].var())
print('\nSkewness of Sleep Duration:',df['Sleep Duration'].skew())
print('\nKurtosis of Age:',df['Age'].kurt())
```

```
Mean of Physical Activity Level: 59.17112299465241

Median of Stress Level: 5.0

Mode of Quality of Sleep: 0    8
Name: Quality of Sleep, dtype: int64

Standard deviation of Heart Rate: 4.135675535112214

Variance of Daily Steps: 2617651.144786458

Skewness of Sleep Duration: 0.037554389846484834

Kurtosis of Age: -0.9097795476259583
```

INTERPRETATION OF SKEWNESS AND KURTOSIS

Kurtosis value is almost zero:Data points follow normal distribution in Age

Skewness value is slightly positive i.e datapoints have heavy tailed normal distribution

INFERENTIAL STATISTICS

In [44]:

```
from scipy import stats
Physical_Activity=df['Physical Activity Level']
Physical_Activity_assumed=60.074745432
t_stat,p_value=stats.ttest_1samp(Physical_Activity,Physical_Activity_assumed)
print('t_statistic:',t_stat)
print('\np_value:',p_value)
```

```
t_statistic: -0.8389127244590534

p_value: 0.40205559391040013
```

INTERPRETATION

t_statistic represents that sample mean is 0.83 standard errors below the population mean

p_value is larger than 0.05,so we fail to reject the null hypothesis:Physical_Activity_assumed=60.074745432

CONFIDENCE INTERVAL

In [42]:

```
import numpy as np
from scipy import stats
sample_mean = np.mean(df['Physical Activity Level'])
standard_error = stats.sem(df['Physical Activity Level'])

# Compute 95% confidence interval for Physical Activity Level
confidence_interval = stats.norm.interval(0.95, loc=sample_mean, scale=standard_error)
print('95% confidence interval for Physical Activity Level:',confidence_interval)
```

```
95% confidence interval for Physical Activity Level: (57.059976767891335, 61.28226922141348)
```

Regression Analysis

In [46]:

```
import statsmodels.api as sm

# Define independent variable (add constant for intercept)
X = sm.add_constant(df['Physical Activity Level'])

# Define dependent variable
y = df['Quality of Sleep']

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

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

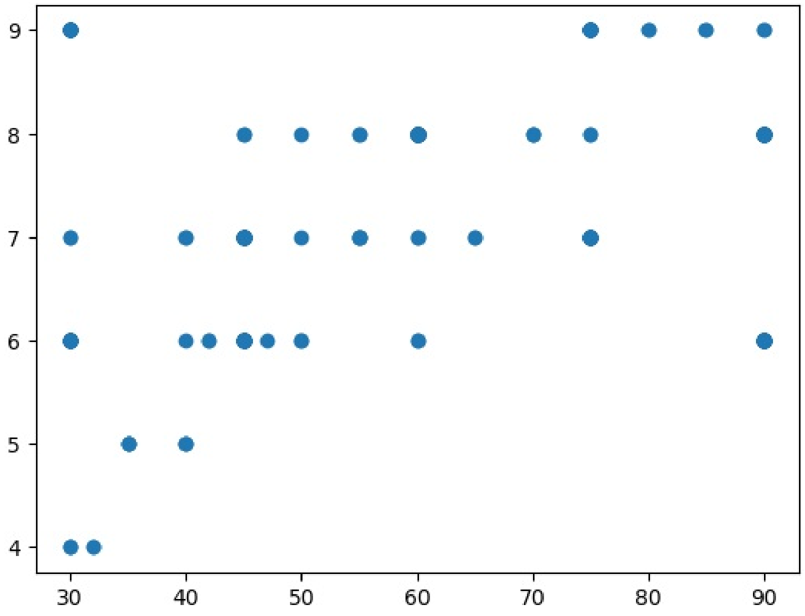
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:       Quality of Sleep   R-squared:                       0.037
Model:                            OLS   Adj. R-squared:                  0.035
Method:                 Least Squares   F-statistic:                     14.38
Date:                Sun, 08 Sep 2024   Prob (F-statistic):           0.000175
Time:                        23:34:50   Log-Likelihood:                -590.33
No. Observations:                 374   AIC:                             1185.
Df Residuals:                     372   BIC:                             1193.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                       6.6570      0.183     36.307      0.000       6.296       7.018
Physical Activity Level     0.0111      0.003      3.792      0.000       0.005       0.017
==============================================================================
Omnibus:                       19.171   Durbin-Watson:                   0.435
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                8.103
Skew:                          -0.044   Prob(JB):                       0.0174
Kurtosis:                       2.284   Cond. No.                         189.
==============================================================================

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

In [48]:

```
import matplotlib.pyplot as plt
```

In [51]:

```
plt.scatter(df['Physical Activity Level'] ,y)
plt.show()
```



CONCLUSION

- we have R-squared=0.037
- it means that 3.7% of the variance in quality of sleep is defined Physical Activity Level rate

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js