5509 Final

Sumit Kumar. (2024). student lifestyle dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/9876359

```
df.head()
```

	Student_ID	Study_Hours_Per_Day	Extracurricular_Hours_Per_Day	Sleep_Hours_Per_Day	Social_Hours_Per_Day	Physical_Activity_Hours_Per_Day	GPA	Stress_Level
0	1	6.9	3.8	8.7	2.8	1.8	2.99	Moderate
1	2	5.3	3.5	8.0	4.2	3.0	2.75	Low
2	3	5.1	3.9	9.2	1.2	4.6	2.67	Low
3	4	6.5	2.1	7.2	1.7	6.5	2.88	Moderate
4	5	8.1	0.6	6.5	2.2	6.6	3.51	High

```
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Student_ID                      2000 non-null   int64
 1   Study_Hours_Per_Day             2000 non-null   float64
 2   Extracurricular_Hours_Per_Day   1981 non-null   float64
 3   Sleep_Hours_Per_Day             2000 non-null   float64
 4   Social_Hours_Per_Day            1989 non-null   float64
 5   Physical_Activity_Hours_Per_Day 1998 non-null   float64
 6   GPA                             2000 non-null   float64
 7   Stress_Level                    2000 non-null   object
dtypes: float64(6), int64(1), object(1)
memory usage: 125.1+ KB
```

	Student_ID	Study_Hours_Per_Day	Extracurricular_Hours_Per_Day	\
count	2000.000000	2000.000000	1981.000000	
mean	1000.500000	7.475800	2.009187	
std	577.494589	1.423888	1.144749	
min	1.000000	5.000000	0.000000	
25%	500.750000	6.300000	1.000000	
50%	1000.500000	7.400000	2.000000	
75%	1500.250000	8.700000	3.000000	
max	2000.000000	10.000000	4.000000	

	Sleep_Hours_Per_Day	Social_Hours_Per_Day	\
count	2000.000000	1989.000000	
mean	7.501250	2.719507	
std	1.460949	1.681118	
min	5.000000	0.000000	
25%	6.200000	1.300000	
50%	7.500000	2.600000	
75%	8.800000	4.100000	
max	10.000000	6.000000	

	Physical_Activity_Hours_Per_Day	GPA
count	1998.000000	2000.000000
mean	4.332633	3.115960
std	2.511633	0.298674
min	0.000000	2.240000
25%	2.400000	2.900000
50%	4.100000	3.110000
75%	6.100000	3.330000
max	13.000000	4.000000

Student_ID	0
Study_Hours_Per_Day	0
Extracurricular_Hours_Per_Day	19
Sleep_Hours_Per_Day	0
Social_Hours_Per_Day	11
Physical_Activity_Hours_Per_Day	2
GPA	0
Stress_Level	0

```
df=df.fillna(0)
print(df.isnull().sum())
df.info()
```

Student_ID	0
Study_Hours_Per_Day	0
Extracurricular_Hours_Per_Day	0
Sleep_Hours_Per_Day	0
Social_Hours_Per_Day	0
Physical_Activity_Hours_Per_Day	0
GPA	0
Stress_Level	0

```python
df = df.drop(columns=['Student_ID'])
df = df.replace('Low', 1)
df = df.replace('Moderate', 2)
df = df.replace('High', 3)
print(df.describe())
```

```
       Study_Hours_Per_Day  Extracurricular_Hours_Per_Day  \
count          2000.000000                    2000.000000
mean              7.475800                       1.990100
std               1.423888                       1.155855
min               5.000000                       0.000000
25%               6.300000                       1.000000
50%               7.400000                       2.000000
75%               8.700000                       3.000000
max              10.000000                       4.000000

       Sleep_Hours_Per_Day  Social_Hours_Per_Day  \
count          2000.000000           2000.000000
mean              7.501250              2.704550
std               1.460949              1.688514
min               5.000000              0.000000
25%               6.200000              1.200000
50%               7.500000              2.600000
75%               8.800000              4.100000
max              10.000000              6.000000

       Physical_Activity_Hours_Per_Day          GPA  Stress_Level
count                      2000.00000  2000.000000   2000.000000
mean                          4.32830     3.115960      2.366000
std                           2.51411     0.298674      0.727536
min                           0.00000     2.240000      1.000000
25%                           2.40000     2.900000      2.000000
50%                           4.10000     3.110000      3.000000
75%                           6.10000     3.330000      3.000000
max                          13.00000     4.000000      3.000000
```

```
df.corr()
```

	Study_Hours_Per_Day	Extracurricular_Hours_Per_Day	Sleep_Hours_Per_Day	Social_Hours_Per_Day	Physical_Activity_Hours_Per_Day	GPA	Stress_Level
Study_Hours_Per_Day	1.000000	-0.002629	0.026717	-0.137820	-0.488113	0.734468	0.738843
Extracurricular_Hours_Per_Day	-0.002629	1.000000	0.008844	-0.139081	-0.369989	-0.032174	-0.006099
Sleep_Hours_Per_Day	0.026717	0.008844	1.000000	-0.193556	-0.470302	-0.004278	-0.298917
Social_Hours_Per_Day	-0.137820	-0.139081	-0.193556	1.000000	-0.417142	-0.085677	-0.054702
Physical_Activity_Hours_Per_Day	-0.488113	-0.369989	-0.470302	-0.417142	1.000000	-0.341152	-0.205207
GPA	0.734468	-0.032174	-0.004278	-0.085677	-0.341152	1.000000	0.550395
Stress_Level	0.738843	-0.006099	-0.298917	-0.054702	-0.205207	0.550395	1.000000

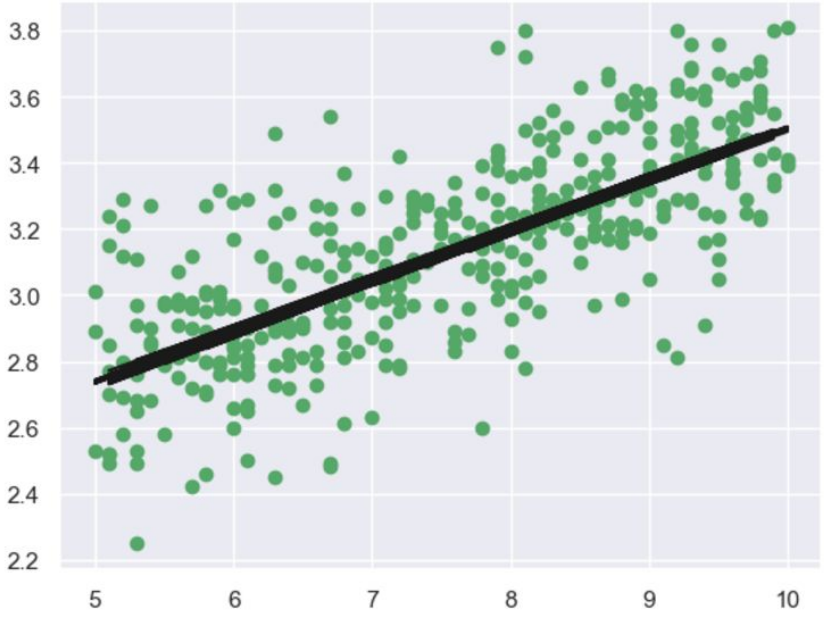
```
model = smf.ols('GPA~Study_Hours_Per_Day', X_train)
res = model.fit()
print(res.summary())
```

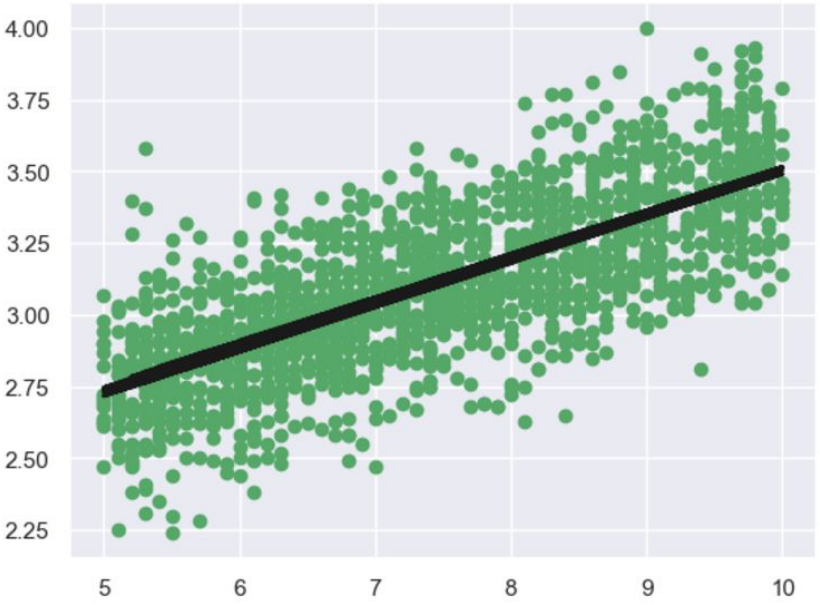
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    GPA   R-squared:                       0.537
Model:                            OLS   Adj. R-squared:                  0.537
Method:                 Least Squares   F-statistic:                     1855.
Date:                Tue, 26 Nov 2024   Prob (F-statistic):          1.14e-269
Time:                        17:55:54   Log-Likelihood:                285.87
No. Observations:                1600   AIC:                            -567.7
Df Residuals:                    1598   BIC:                            -557.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.9631      0.027     72.452      0.000       1.910       2.016
Study_Hours_Per_Day     0.1538      0.004     43.071      0.000       0.147       0.161
==============================================================================
Omnibus:                        0.260   Durbin-Watson:                   2.023
Prob(Omnibus):                  0.878   Jarque-Bera (JB):                0.246
Skew:                           0.030   Prob(JB):                        0.884
Kurtosis:                       3.002   Cond. No.                         41.3
==============================================================================

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

```
y_pred = model.predict(X_test)
plt.scatter(X_test['Study_Hours_Per_Day'], X_test['GPA'],color='g')
plt.plot(X_test['Study_Hours_Per_Day'], y_pred, color='k')
plt.show()

y_pred = model.predict(X_train)
plt.scatter(X_train['Study_Hours_Per_Day'], X_train['GPA'],color='g')
plt.plot(X_train['Study_Hours_Per_Day'], y_pred, color='k')
plt.show()
```

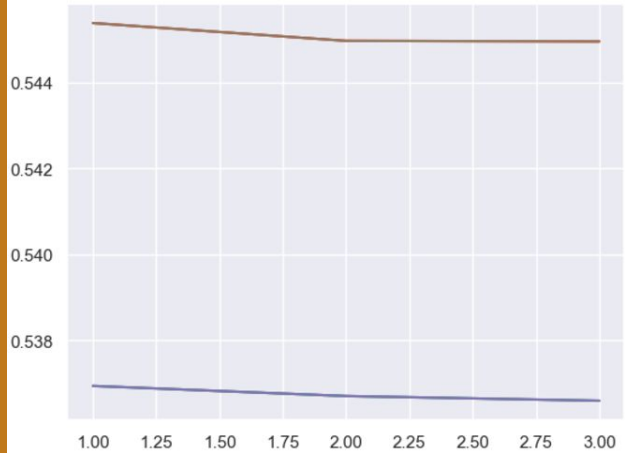




```
variables = list(df.columns.values)
dict = {}
for each in variables:
    test = 'GPA~' + each
    model = smf.ols(test, df)
    model = model.fit()
    r = model.rsquared_adj
    dict[each] = r
{k: v for k, v in sorted(dict.items(), key=lambda item: item[1])}
```

```
{'Sleep_Hours_Per_Day': -0.00048218628185647816,
 'Extracurricular_Hours_Per_Day': 0.0005351537302717979,
 'Social_Hours_Per_Day': 0.006843745857941452,
 'Physical_Activity_Hours_Per_Day': 0.11594275395217035,
 'Stress_Level': 0.3025854820843027,
 'Study_Hours_Per_Day': 0.5392127057073811,
 'GPA': 1.0}
```

```
num = [1, 2, 3]
test1 = smf.ols('GPA~Study_Hours_Per_Day', X_test).fit().rsquared_adj
one = smf.ols('GPA~Study_Hours_Per_Day', X_train).fit().rsquared_adj
test2 = smf.ols('GPA~Study_Hours_Per_Day+Stress_Level', X_test).fit().rsquared_adj
two = smf.ols('GPA~Study_Hours_Per_Day+Stress_Level', X_train).fit().rsquared_adj
test3 = smf.ols('GPA~Study_Hours_Per_Day+Stress_Level+Physical_Activity_Hours_Per_Day', X_test).fit().rsquared_adj
three = smf.ols('GPA~Study_Hours_Per_Day+Stress_Level+Physical_Activity_Hours_Per_Day', X_train).fit().rsquared_adj
adjr2_test = [test1, test2, test3]
adjr2_train = [one, two, three]
plt.plot(num, adjr2_train)
plt.plot(num, adjr2_test)
plt.show()
```



```
formula = 'GPA~Study_Hours_Per_Day'
model = smf.ols(formula, df)
res = model.fit()
print(1, res.rsquared)

for x in range(2,10):
    annoying = '+ np.power(Study_Hours_Per_Day,' + str(x) + ')'
    formula = formula + annoying
    model = smf.ols(formula, df)
    res = model.fit()
    print(x, res.rsquared)
```

```
1 0.5394432146089783
2 0.5394956556791237
3 0.5395720056131703
4 0.5395744382788368
5 0.5398204132008578
6 0.5398971743690786
7 0.5399087608478204
8 0.5400553182706231
9 0.5401765380529989
```

Out[295... OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.541
Model:	OLS	**Adj. R-squared:**	0.540
Method:	Least Squares	**F-statistic:**	470.0
Date:	Mon, 02 Dec 2024	**Prob (F-statistic):**	0.00
Time:	06:59:09	**Log-Likelihood:**	358.07
No. Observations:	2000	**AIC:**	-704.1
Df Residuals:	1994	**BIC:**	-670.5
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0232	0.000	106.218	0.000	0.023	0.024
Study_Hours_Per_Day	0.2370	0.005	47.703	0.000	0.227	0.247
Extracurricular_Hours_Per_Day	0.0752	0.004	19.446	0.000	0.068	0.083
Sleep_Hours_Per_Day	0.0781	0.003	28.206	0.000	0.073	0.084
Social_Hours_Per_Day	0.0840	0.003	33.014	0.000	0.079	0.089
Physical_Activity_Hours_Per_Day	0.0827	0.002	51.490	0.000	0.080	0.086
Stress_Level	0.0002	0.010	0.020	0.984	-0.020	0.021

In [149…

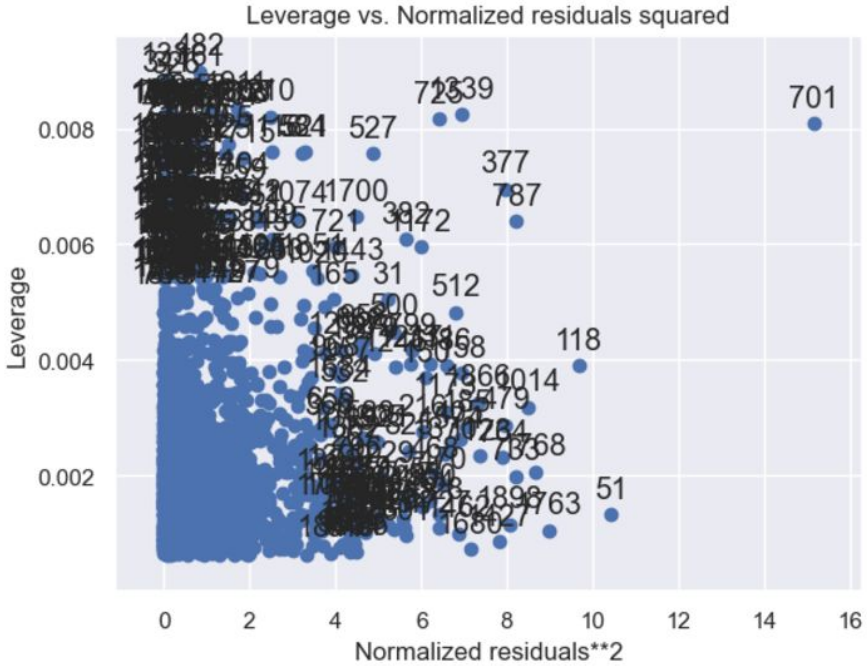
```
model = smf.ols('GPA~Study_Hours_Per_Day+Extracurricular_Hours_Per_Day*Stress_Level', df).fit()
model.summary()
```

Out[149…

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.540
Model:	OLS	**Adj. R-squared:**	0.540
Method:	Least Squares	**F-statistic:**	586.6
Date:	Tue, 26 Nov 2024	**Prob (F-statistic):**	0.00
Time:	19:54:28	**Log-Likelihood:**	357.02
No. Observations:	2000	**AIC:**	-704.0
Df Residuals:	1995	**BIC:**	-676.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.9837	0.036	55.867	0.000	1.914	2.053
Study_Hours_Per_Day	0.1514	0.005	32.051	0.000	0.142	0.161
Extracurricular_Hours_Per_Day	-0.0082	0.013	-0.625	0.532	-0.034	0.017
Stress_Level	0.0066	0.014	0.465	0.642	-0.021	0.034
Extracurricular_Hours_Per_Day:Stress_Level	0.0002	0.005	0.030	0.976	-0.010	0.011




```
(array([   1,    2,    8,   36,   46,   54,   58,   62,   81,   86,   88,
          99,  113,  121,  133,  149,  151,  160,  161,  168,  171,  180,
         210,  220,  237,  246,  250,  269,  272,  276,  289,  325,  326,
         329,  339,  341,  343,  354,  377,  382,  401,  409,  428,  461,
         482,  488,  521,  527,  545,  584,  588,  592,  629,  642,  648,
         651,  668,  694,  701,  706,  713,  714,  721,  725,  729,  742,
         749,  756,  759,  778,  784,  787,  815,  838,  840,  915,  917,
         933,  936,  939,  952,  955,  973,  988, 1009, 1011, 1039, 1045,
        1057, 1074, 1077, 1080, 1081, 1092, 1094, 1109, 1116, 1117, 1130,
        1172, 1174, 1175, 1177, 1180, 1194, 1211, 1253, 1268, 1303, 1304,
        1339, 1347, 1408, 1410, 1445, 1466, 1487, 1488, 1496, 1510, 1561,
        1603, 1604, 1629, 1648, 1651, 1676, 1700, 1712, 1715, 1716, 1726,
        1743, 1758, 1785, 1798, 1818, 1830, 1835, 1854, 1857, 1890, 1900,
        1909, 1911, 1920, 1927], dtype=int64),)
```

Conclusion

- Simple model with study hours and GPA was best model
- High variability in the data
- Future steps: categorical regression between stress level and GPA