In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import seaborn as sns
sns.set() #reset the current graph style, we need to re-run everythign again
```

In [2]:
```
data = pd.read_csv('1.01. Simple linear regression.csv')
```

In [3]:
```
data
```

Out[3]:

	SAT	GPA
0	1714	2.40
1	1664	2.52
2	1760	2.54
3	1685	2.74
4	1693	2.83
...	...	...
79	1936	3.71
80	1810	3.71
81	1987	3.73
82	1962	3.76
83	2050	3.81

84 rows × 2 columns

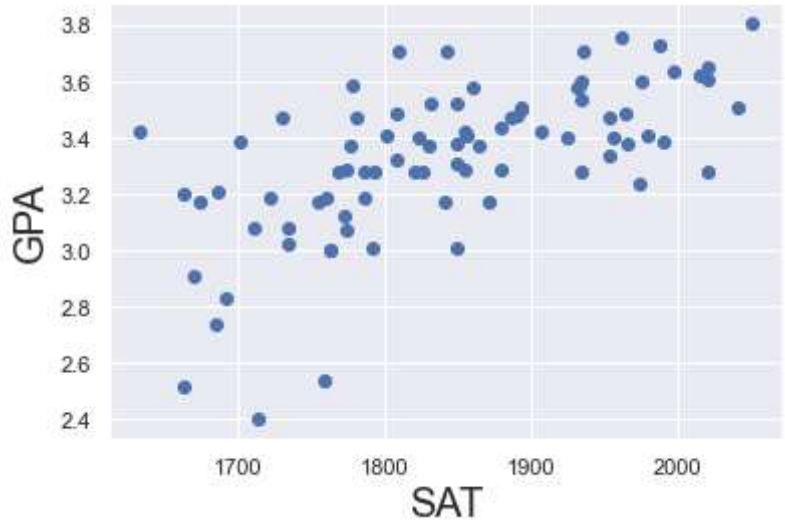
In [4]:
```
#provides statics data from each column
data.describe()
```

Out[4]:

	SAT	GPA
count	84.000000	84.000000
mean	1845.273810	3.330238
std	104.530661	0.271617
min	1634.000000	2.400000
25%	1772.000000	3.190000
50%	1846.000000	3.380000
75%	1934.000000	3.502500
max	2050.000000	3.810000

In [5]:
```
#We are going to create a linear regression which predits GPA based on the SAT sc
y = data['GPA']
x1 = data['SAT']
```

In [6]:
```
plt.scatter(x1,y)
plt.xlabel('SAT',fontsize=20)
plt.ylabel('GPA',fontsize=20)
plt.show()
```



In [7]:
```
x = sm.add_constant(x1)
results = sm.OLS(y,x).fit() #fit will apply a specific estimation technique (OLS
results.summary()
```

Out[7]:

OLS Regression Results

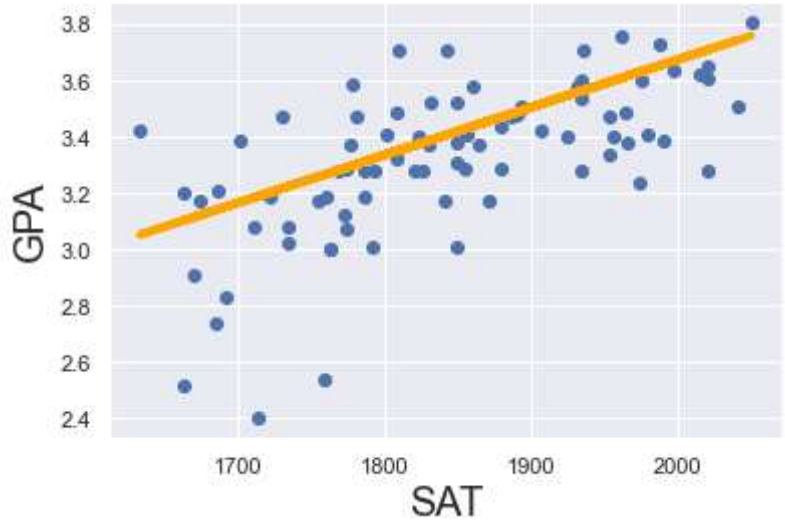
Dep. Variable:	GPA	**R-squared:**	0.406
Model:	OLS	**Adj. R-squared:**	0.399
Method:	Least Squares	**F-statistic:**	56.05
Date:	Tue, 24 Aug 2021	**Prob (F-statistic):**	7.20e-11
Time:	21:29:57	**Log-Likelihood:**	12.672
No. Observations:	84	**AIC:**	-21.34
Df Residuals:	82	**BIC:**	-16.48
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2750	0.409	0.673	0.503	-0.538	1.088
SAT	0.0017	0.000	7.487	0.000	0.001	0.002

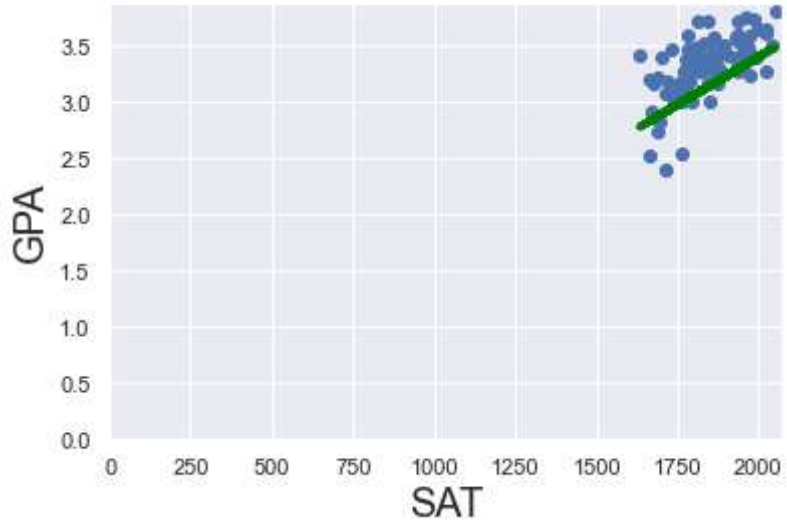
Omnibus:	12.839	**Durbin-Watson:**	0.950
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	16.155
Skew:	-0.722	**Prob(JB):**	0.000310
Kurtosis:	4.590	**Cond. No.**	3.29e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.29e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

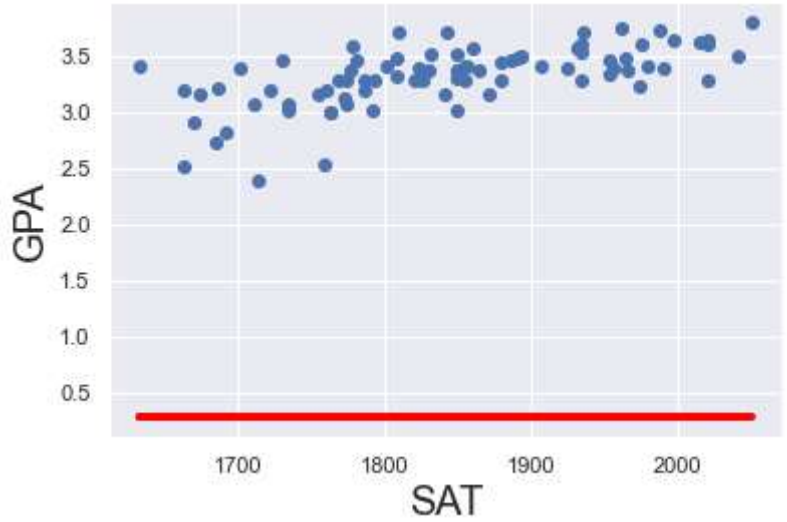
In [8]:
```
plt.scatter(x1,y)
yhat = 0.0017*x1 + 0.275
fig = plt.plot(x1, yhat, lw=4, c='orange', label='regression line')
plt.xlabel('SAT', fontsize=20)
plt.ylabel('GPA', fontsize=20)
plt.show()
```



In [9]:
```
plt.scatter(x1,y)
yhat = 0.0017*x1
fig = plt.plot(x1, yhat, lw=4, c='green', label='regression line')
plt.xlabel('SAT', fontsize=20)
plt.ylabel('GPA', fontsize=20)
plt.xlim(0)
plt.ylim(0)
plt.show()
```



In [10]:
```
plt.scatter(x1,y)
yhat = 0*x1 + 0.275
fig = plt.plot(x1, yhat, lw=4, c='red', label='regression line')
plt.xlabel('SAT', fontsize=20)
plt.ylabel('GPA', fontsize=20)
plt.show()
```



In [ ]:
```
#are thosse variables useful?
# Does it help us explain the variability we have in this case?
# p-value "P>\t\ = p-value < 0.05" means tha variable is significant
#SAT p-value is = 0.000
#we could observed that the p-value didn't match what we are looking and SAT is c
```