```
import pandas as pd
import math
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

```
cd /content/drive/MyDrive/Econometrics/Simple Regression
```

```
/content/drive/MyDrive/Econometrics/Simple Regression
```

```
Data = pd.read_csv('TestExer1-holiday expenditures-round2.csv')
Data.head()
```

	Observation	Age	Expenditures
0	1	49	95
1	2	15	104
2	3	43	91
3	4	45	98
4	5	40	94

```
Y = Data.Expenditures # the dependent variable
X = Data.Age # the independent variable
```

```
#coefficient b
b = ((X*Y).mean() - X.mean()*Y.mean()) / ((X**2).mean() - (X.mean())**2)
print("Value of b is: ",b)
```

```
Value of b is:  -0.33359609660627854
```

```
X_bar = Data.Age.mean() # sample mean of age
Y_bar = Data.Expenditures.mean() # sample mean of expenditures print("Mean Age : ", X_bar)
print("Mean Expenditure : ", Y_bar)
```

```
Mean Expenditure :  101.11538461538461
```

```
a = Y_bar - b*X_bar
print("Value of a : ", a)
```

```
Value of a :  114.24110795493165
```

```
Data["error"] = Data.Expenditures - a - b*Data.Age
Data.head()
```

	Observation	Age	Expenditures	error
0	1	49	95	-2.894899
1	2	15	104	-5.237167
2	3	43	91	-8.896476
3	4	45	98	-1.229284

```
sum_sq_error = (Data.error ** 2).sum() # calculating the sum of squares
```

```
## calclating ci in the dataset
Data["c"] = (Data.Age - X_bar) / ((Data.Age - X_bar)**2).sum()
Data.head(6) # showing the first few rows of the enhanced dataset
```

	Observation	Age	Expenditures	error	c
0	1	49	95	-2.894899	0.003411
1	2	15	104	-5.237167	-0.008603
2	3	43	91	-8.896476	0.001291
3	4	45	98	-1.229284	0.001998
4	5	40	94	-6.897264	0.000231
5	6	35	107	4.434755	-0.001536

```
beta = b - (Data.c * Data.error).sum()
print("The value of beta is: ", beta)
```

```
The value of beta is:  -0.33359609660628065
```

```
n = Data.shape[0] # number of entries
s_b_sq = np.sqrt((((Data.error)**2).sum()) / ((n-2) * (((X - X_bar)**2).sum())))
print("The value of standard error is: ", s_b_sq)
```

```
The value of standard error is:  0.09536918278863911
```

```
t_b = (b)/s_b_sq
print("The t value of b is: ", t_b)

```

```
The t value of b is:  -3.4979443762835545
```

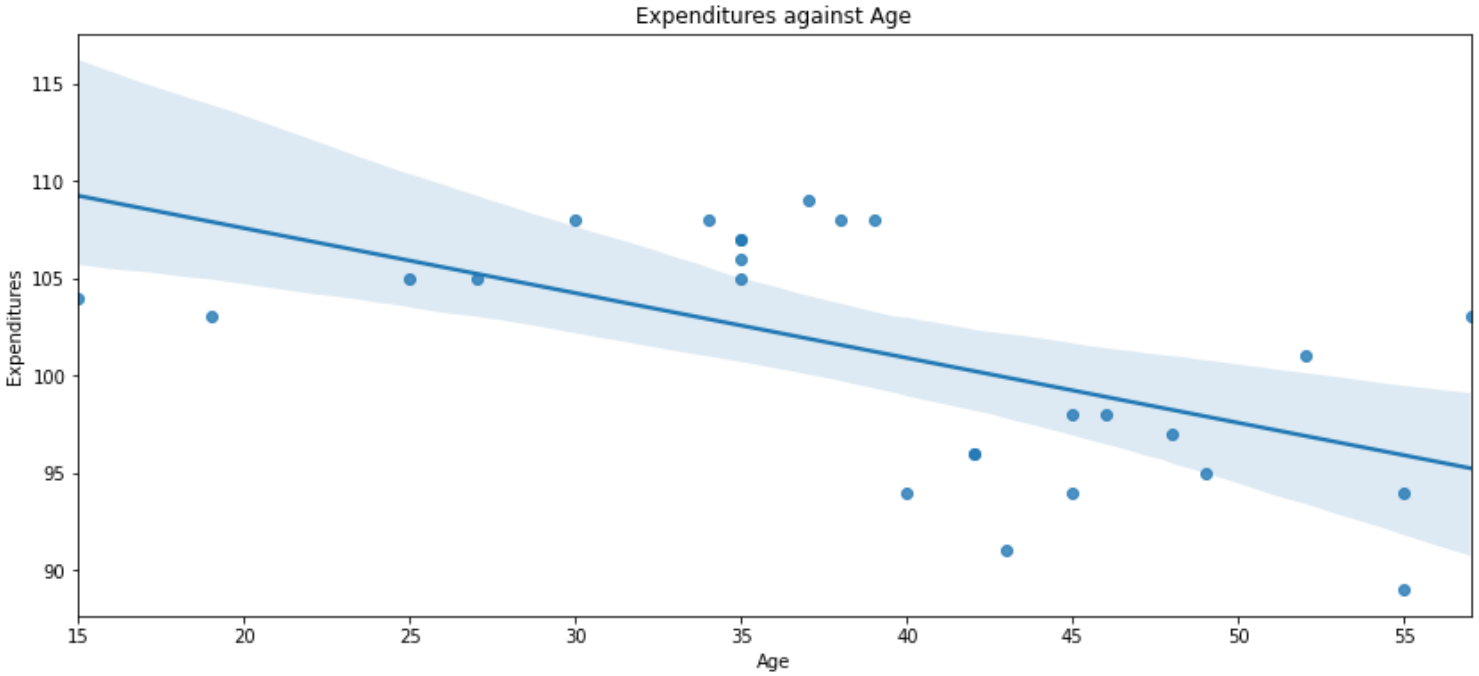
```
#Answer 1 summary
print("Summary of Answer a results\n")
print("Value of a : ", a)
print("Value of b : ",b)
print("The standard error is: ", s_b_sq)
print("The t value of b is: ", t_b)
```

```
7
```

```
Summary of Answer a results

Value of a :  114.24110795493165
Value of b :  -0.33359609660627854
The standard error is:  0.09536918278863911
The t value of b is:  -3.4979443762835545
```

```
#Question b
plot = sns.regplot(data=Data, x= "Age", y= "Expenditures")
plot.figure.set_size_inches(14,6)
plot.axes.set_title('Expenditures against Age')
```

```
Text(0.5, 1.0, 'Expenditures against Age')
```




```
1
2  › see two clusters of data around age groups less than 40 and greater than 40. These clusters will
```

```
lt40 = Data.Age < 40
Data_lt40 = Data[lt40].copy()
Data_lt40
```

	Observation	Age	Expenditures	error	c
1	2	15	104	-5.237167	-0.008603
5	6	35	107	4.434755	-0.001536
7	8	38	108	6.435544	-0.000476
9	10	30	108	3.766775	-0.003303
13	14	25	105	-0.901206	-0.005070
14	15	35	107	4.434755	-0.001536
15	16	35	106	3.434755	-0.001536
16	17	35	105	2.434755	-0.001536

```
## calculating the value of b which is needed to derive a
Y = Data_lt40.Expenditures # the dependent variable
X = Data_lt40.Age # the independent variable
b = ((X*Y).mean() - X.mean()*Y.mean()) / ((X**2).mean() - (X.mean())**2)
X_bar = X.mean() # sample mean of age
Y_bar = Y.mean()
a = Y_bar - b*X_bar
print("Summary of Answer c - part 1 results\n")
print("Value of a is: ", a)
print("Value of b is", b)
## calculate error from a and b
Data_lt40["error"] = Y - a - b*X
sum_sq_error = (Data_lt40.error ** 2).sum() # calculating the sum of squares
n = Data_lt40.shape[0] # number of entries
s_b_sq = np.sqrt((((Data_lt40.error)**2).sum()) / ((n-2) * (((X - X_bar)**2).sum())))
t_b = (b)/s_b_sq
print("The standard error is: ", s_b_sq)
print("The t value of b is: ", t_b)
# sample data set with errors and c
print("\n\n Sample data for the final dataset for Age less than 40 with error and c")
Data_lt40.head()
```

```
Summary of Answer c - part 1 results

Value of a is:  100.23227718258495
```

```
gt40 = Data.Age >= 40
Data_gt40 = Data[gt40].copy()
Data_gt40
```

	Observation	Age	Expenditures	error	c
0	1	49	95	-2.894899	0.003411
2	3	43	91	-8.896476	0.001291
3	4	45	98	-1.229284	0.001998
4	5	40	94	-6.897264	0.000231
6	7	42	96	-4.230072	0.000938
8	9	46	98	-0.895688	0.002351
10	11	52	101	4.105889	0.004472
11	12	55	89	-6.893323	0.005532
12	13	42	96	-4.230072	0.000938
18	19	48	97	-1.228495	0.003058
20	21	45	94	-5.229284	0.001998
22	23	57	103	7.773870	0.006238
23	24	55	94	-1.893323	0.005532

```
## calculating the value of b which is needed to derive a
Y = Data_gt40.Expenditures # the dependent variable
X = Data_gt40.Age # the independent variable
b = ((X*Y).mean() - X.mean()*Y.mean()) / ((X**2).mean() - (X.mean())**2)
X_bar = X.mean() # sample mean of age
Y_bar = Y.mean()
a = Y_bar - b*X_bar
print("Summary of Answer c - part 2 results\n")
print("Value of a is: ", a)
print("Value of b is", b)
## calculate error from a and b
Data_gt40["error"] = Y - a - b*X
sum_sq_error = (Data_gt40.error ** 2).sum() # calculating the sum of squares
n = Data_gt40.shape[0] # number of entries
s_b_sq = np.sqrt((((Data_gt40.error)**2).sum()) / ((n-2) * (((X - X_bar)**2).sum())))
t_b = (b)/s_b_sq
print("The standard error is: ", s_b_sq)
print("The t value of b is: ", t_b)

# sample data set with errors and c
print("\n\n Sample data for the final dataset for Age greater than or equal to 40 with error and c
Data_gt40.head()
```

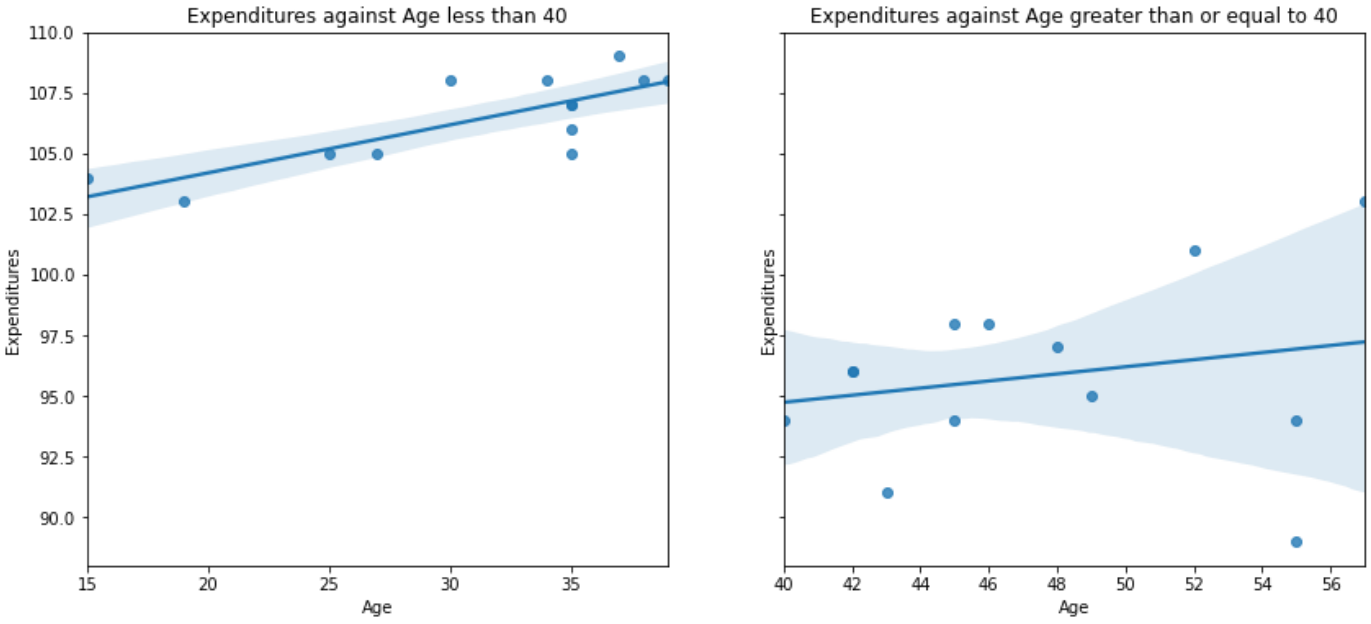
```
Summary of Answer c - part 2 results

Value of a is:  88.87188902488657
Value of b is 0.14647082823339977
The standard error is:  0.19738441872591267
The t value of b is:  0.7420587155705977
```

Sample data for the final dataset for Age greater than or equal to 40 with error and c

	Observation	Age	Expenditures	error	c
0	1	49	95	-1.048960	0.003411
2	3	43	91	-4.170135	0.001291
3	4	45	98	2.536924	0.001998
4	5	40	94	-0.730722	0.000231
6	7	42	96	0.976336	0.000938

```
1  fig, (ax1, ax2) = plt.subplots(ncols=2, sharey=True)
2  sns.regplot(x = Data_lt40.Age, y = Data_lt40.Expenditures, ax = ax1)
3  ax1.figure.set_size_inches(14,6)
4  ax1.axes.set_title('Expenditures against Age less than 40')
5  sns.regplot(x = Data_gt40.Age, y = Data_gt40.Expenditures, ax = ax2)
6  ax2.figure.set_size_inches(14,6)
7  ax2.axes.set_title('Expenditures against Age greater than or equal to 40')
```

```
Text(0.5, 1.0, 'Expenditures against Age greater than or equal to 40')
```




```
1  # # Answer d
2  #
3  # Splitting the data into the two clusters mentioned in answer b gives opposite inference to what
```

```
# Splitting the data into the two clusters mentioned in answer b gives opposite inference to what
# within the two clusters, people with age less than 40 have more sensitive spending habits. The
```

```
```