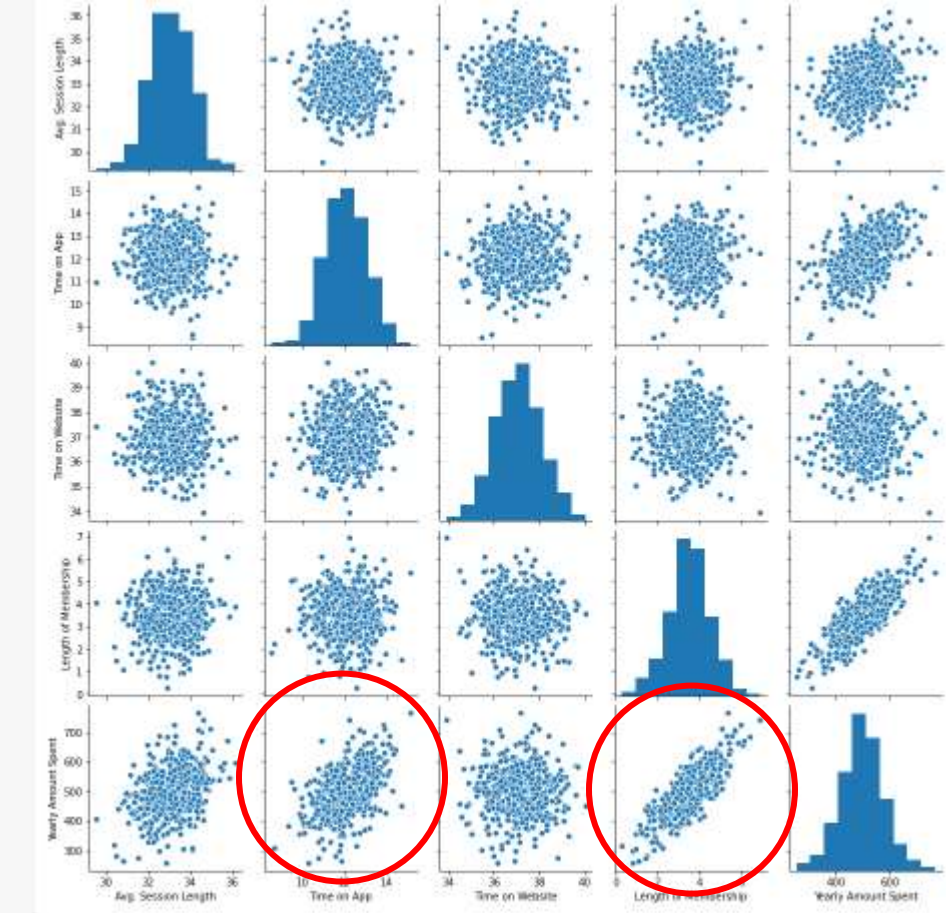
The implementation and steps of the answer are all in the Ecommerce Customers.ipynb file

I.

We have 5 numeric data columns. I set the last column, the Yearly Amount Spent column, as the target, y. After the graphs obtained:

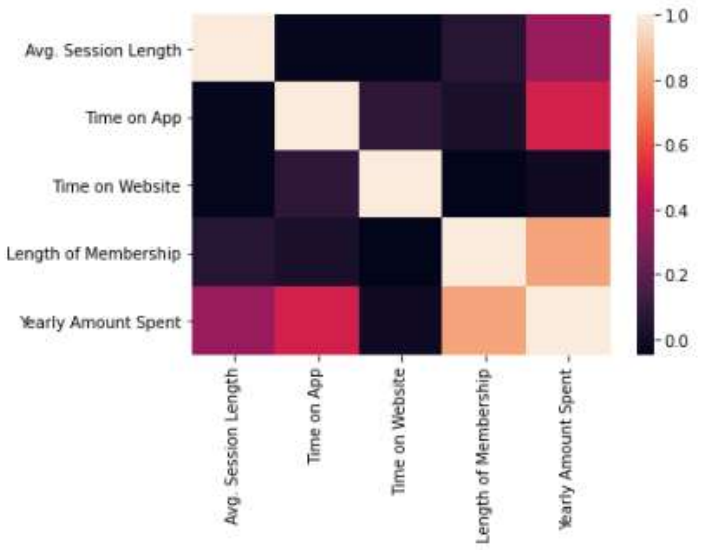
If we examine the first 4 rows and 4 columns, it shows that there is no linear or correlation between the xi variables. Regarding the last line, that is, the correlation between the target variable and each of the variables, it seems that there is only a significant linear relationship between the Yearly Amount Spent and the Time on App and Length of Membership columns. The second one seems more coherent.



For a more detailed analysis, we have the following diagram:

According to the Kanaar guide index, there is no correlation on the first 4 rows and columns, i.e., the evaluation variables (between Time on App and Time on Website, there is a negligible correlation - about 0.1, which can be ignored), but in the last row and column Apart from Time

on Website, the other 3 columns have good degrees of correlation with the target variable. This value for Time on App and Length of Membership is about 0.5 and 0.8, respectively, and about the Avg column. Session Length drops to 0.4. This graph more accurately confirms the results of the previous graph and in addition gives a numerical measure. From this initial result, it is clear that the Length of Membership has the greatest impact on the Annual Amount Spent, followed by Time on App and Avg. Session Length respectively and finally Time on Website have the least impact.

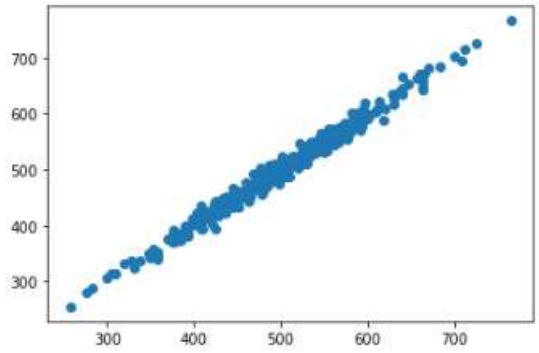


II.

First, I separate 25% of the data for testing. (The total data was 500) Now I fit the multiple linear regression model on the remaining 375 data. To calculate the training error, I get the MSE value and its square root only on this set of 375 training data. The amount of errors obtained is as follows:

	Train set
MSE	*103.18965151113609*
RMSE	*10.158230727402096*

And if we want to see the result of matching the real values of the target variable of the training data and their estimated value, the graph has a small deviation from the bisector line, which shows that we have an acceptable model.



Inside the green box, in the first line, the value of the width from the origin is equal to -1049, and in the next lines, the value of the coefficient of each variable is shown in front of it. As we expected from the previous part, the correlation coefficient with the Time on Website variable is a very small value close to zero, and the highest correlation coefficient is +60, which is related to the Length of Membership, the magnitude of which can be seen from the results of the previous part. It was a prediction.

Pi values are written inside the red box, which we use to evaluate the model. All the lines have shown a very small p value and equal to zero for 3 variables, and this means that the hypothesis of H zero can be rejected with great confidence. That is, there is a linear relationship between our y and x. The only thing that is problematic is the value like 0.173 in the fourth line. If, as usual, we assume the alpha value for the test to be equal to 0.05, then the null hypothesis cannot be rejected in this case. If we check the value of the confidence interval given for the coefficient like this variable in the next two columns (yellow box), the only coefficient that is zero inside the confidence interval is related to this case. The hypothesis of H zero in this case cannot be rejected, but apart from this variable, the rest have obtained acceptable results.

The last thing I check to evaluate the model (considering that it is multiple regression) is the value like the red arrow, which is equal to Adj. It is R-squared and the larger this value is and closer to one, the better the model is. This number is 0.981 and it is very good.

Dep. Variable:	Q("Yearly Amount Spent")	R-squared:	0.984
Model:	OLS	Adj. R-squared:	0.984
Method:	Least Squares	F-statistic:	5674.
Date:	Fri, 29 May 2020	Prob (F-statistic):	0.00
Time:	11:03:10	Log-Likelihood:	-1401.5
No. Observations:	375	AIC:	2813.
Df Residuals:	370	BIC:	2833.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1049.4457	27.160	-38.639	0.000	-1102.853	-996.038
Q("Avg. Session Length")	25.7315	0.516	49.900	0.000	24.717	26.745
Q("Time on App")	38.8872	0.540	71.975	0.000	37.825	39.950
Q("Time on Website")	0.3199	0.539	0.593	0.554	-0.741	1.381
Q("Length of Membership")	61.5267	0.531	115.868	0.000	60.483	62.571

Omnibus:	0.355	Durbin-Watson:	1.734
Prob(Omnibus):	0.837	Jarque-Bera (JB):	0.271
Skew:	0.064	Prob(JB):	0.873
Kurtosis:	3.028	Cond. No.	2.64e+03

III.

I used 5 categories. Russ, I ran the same 375 function training data to get an approximation of the error. The value of MSE and RMSE predicted for the test is equal to:

	Predicted
MSE	*104.86734425049633*
RMSE	*10.233088271425057*

IV.

The error obtained for the 125 test data that we left out is equal to:

	Test set
MSE	*84.86573160592083*
RMSE	*9.212259853365016*

And from the previous part we had for the training data:

	Train set
MSE	*103.18965151113609*
RMSE	*10.158230727402096*

If we compare the RMSE error, the difference is very small. That is, the test and training errors are very close to each other. For this reason, it can be said that the model is not overfit. Because in that case, the difference would be very large and the RMSE of the test data would be much larger than the training data. So here we don't have overfit and the chosen complexity level (linear) for estimating the goal is not more than what we need.

On the other hand, the error value obtained is not a large number. If we look at the numbers of the y vector, the difference of 10 for 500 numbers (average) is not a significant error value. But if I want to present a more definite reason, I will refer to the previous part. Usually we raise the underfit error when the complexity of the model for the data used and their relationship with the target variable is low. For example, when we fit a model that does not have a good linear correlation coefficient on a linear model, the obtained model has a large error because it is not complex enough to cover the relationships between variables. So the error eventually increases for both training and test data. Now, let's go back to the previous part and the value of pi that was obtained and showed the linearity of the relationship between x and y. And the model used here is also a linear model. It can be said that the complexity of the model is not less than the data and if we chose a more complex nonlinear model, we would have overfitting. So the result is that it is not underfit and the model has exactly the desired complexity and an acceptable error. Of course, in order to determine exactly where the best point of bias-variance tradeoff happens, I have to use more calculations that I don't know or use the cross-validation method, which is not requested here. So, based on the information I have on this level of the model and errors, I think the model is in the bias-variance balance and has a good performance.

V.

You should invest in a variable whose increase has a greater increase in the target value. It means that it has more dependence. If we consider this as an optimization problem, that variable should take a value that has a larger coefficient in the objective function. Assume the objective function is equal to the regression equation obtained in part b and insert the numbers from the green box of the table in part b:

$$y = 25.731489\, x_1 + 38.887210\, x_2 + 0.319876\, x_3 + 61.5267\, x_4 - 1049.44$$

Now the question between Time on Website and Time on App columns is looking for the most effective variable. Between the coefficient x2 and x3 we must choose the largest one. The result obtained from the model is that focusing on the application can increase the time of using it, and this value affects the target value by a factor of about 40, while the amount of using the website does not have much of a linear effect!