Estimation of Annual Average sales to improve business

- Course: CE463
- Course Instructor: Prof. Gopal R. Patil

TEAM MEMBERS:

Ravi Bhasuri:	170040013
Pradeep Seervi:	170040055
Adarsh Kumawat:	170040056
Shyopal Kumawat:	170040061
Sunil Chawla:	170040062

INTRODUCTION:

We have data from an E-Commerce company in New York City that sells clothes online but they also have in-store style and clothing advice sessions. Customers come into the store, have sessions/ meetings with a personal stylist, then they can go home and order either on a mobile app or website for the clothes they want. The company is trying to decide whether to focus their efforts on their mobile app experience or their website.

We have performed regression analysis on the data available to understand user behaviour and thus to predict the sale based on the following parameters:

- Avg. Session Length: Average session of in-store style advice sessions.
- Time on App: Average time spent on App in minutes
- Time on Website: Average time spent on Website in minutes
- Length of Membership: How many years the customer has been a member

The owner wants to know where he should focus more among the above listed things so that he can increase his sales efficiently. As we want to know the effect that the efforts of the owner have on sales we performed linear regression to know if the following activities are linearly related to sales or not.

VARIABLES USED:

- Avg. Session Length
- Time on App
- Time on Website
- Length of Membership

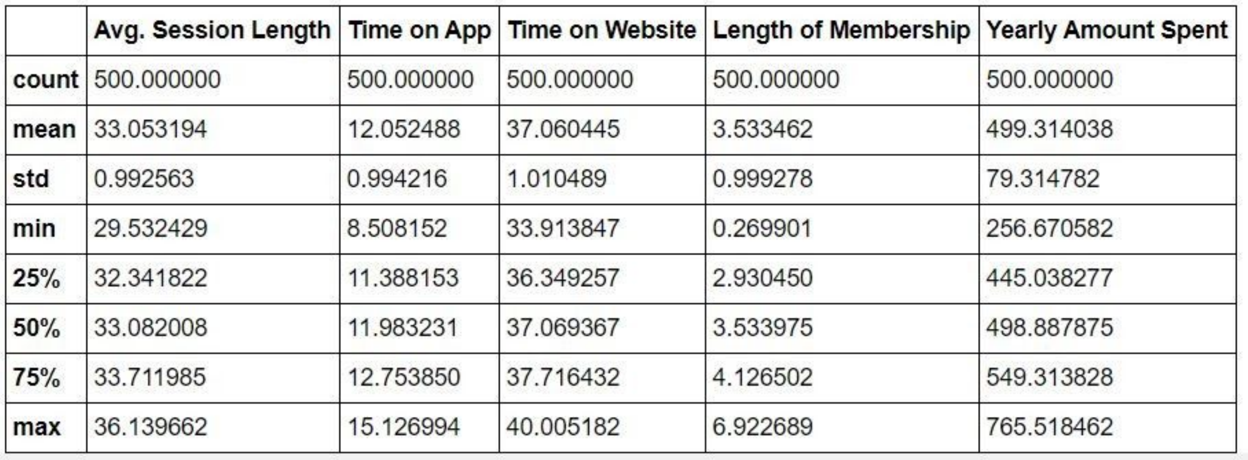
Reason:

The data available had other data also such as address and email id we rejected because no relation suspected in those. As the owner is investing his resources in the above listed parameters the effect/ result of such efforts should be analysed as a part of retrospection so that business can be done efficiently and profit be maximised. We suspect that the amount of time a person spend on app or website should be somewhat linearly related to sales as the more the time spent more is the probability of purchasing clothes, similarly length of membership is large when a person in a regular buyer of clothes also since the sessions are meant to guide people about styling this should also have some effect on purchase pattern of users.

DATA:

A total of 500 user data was available
We decided to use 350 for modelling
Remaining 150 were used for validating
The analysis of the data is tabulated below and interpretations can be taken from it for further proceedings.

Table 1:Data Description



	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	33.053194	12.052488	37.060445	3.533462	499.314038
std	0.992563	0.994216	1.010489	0.999278	79.314782
min	29.532429	8.508152	33.913847	0.269901	256.670582
25%	32.341822	11.388153	36.349257	2.930450	445.038277
50%	33.082008	11.983231	37.069367	3.533975	498.887875
75%	33.711985	12.753850	37.716432	4.126502	549.313828
max	36.139662	15.126994	40.005182	6.922689	765.518462

Regression:

To analyse the data we plotted graphs of the data sets between the yearly amount spent and 4 parameters mentioned earlier i.e. average session length, time spent on app, time spent on website, and length of membership.
The plotted graphs are shown below:

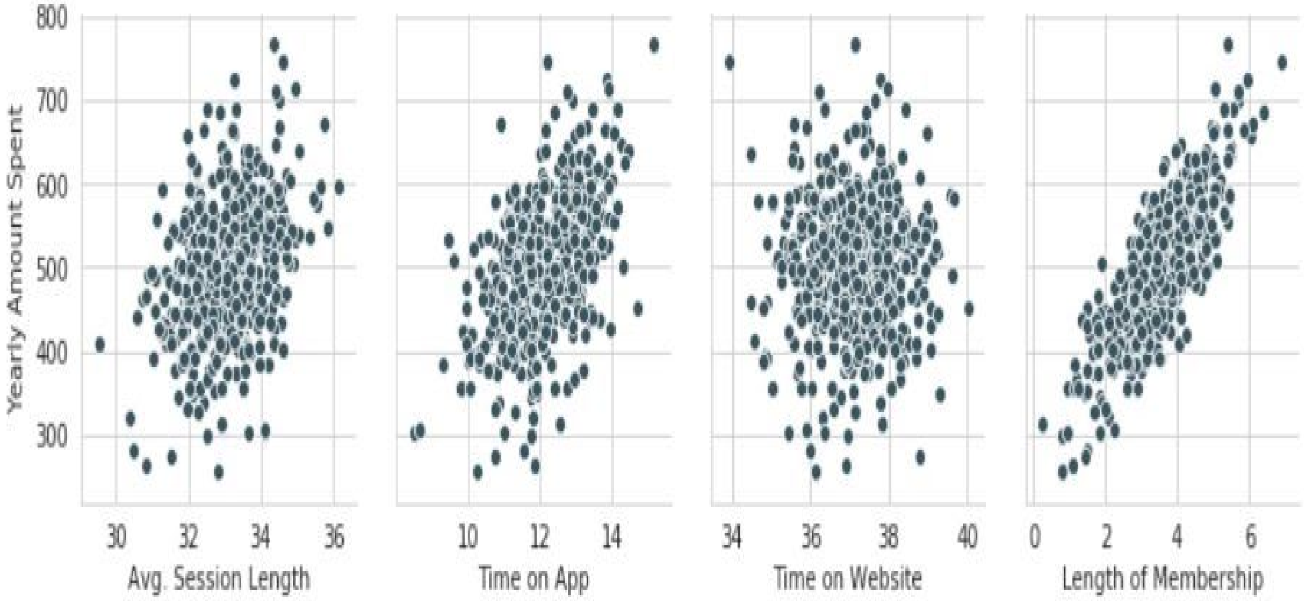
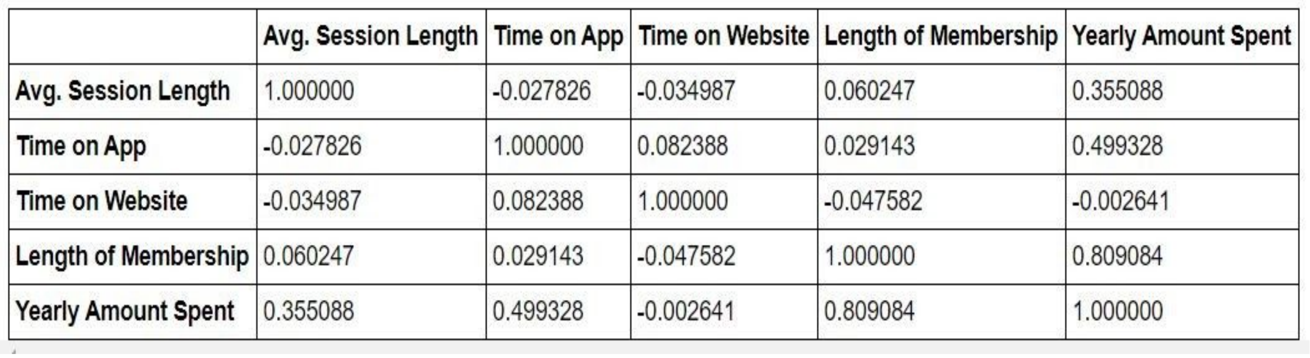


Figure1. Relation Graphs

From the above graphs we can see that the length of membership is linearly related with annual sales and also the relation between sales and time spent on mobile is also somewhat linear.

Now linear regression is performed:

Table2: Correlation matrix



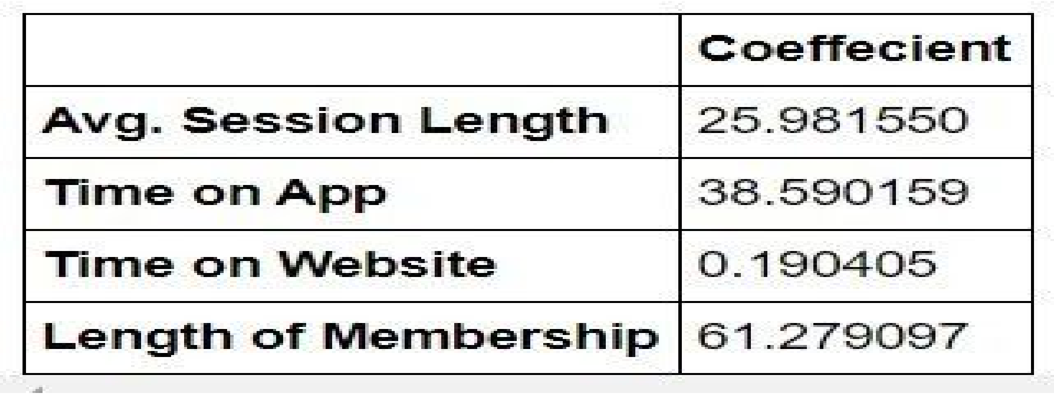
	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
Avg. Session Length	1.000000	-0.027826	-0.034987	0.060247	0.355088
Time on App	-0.027826	1.000000	0.082388	0.029143	0.499328
Time on Website	-0.034987	0.082388	1.000000	-0.047582	-0.002641
Length of Membership	0.060247	0.029143	-0.047582	1.000000	0.809084
Yearly Amount Spent	0.355088	0.499328	-0.002641	0.809084	1.000000

As we can see there is a decent correlation between the parameter length of membership and yearly amount spent from the above correlation matrix. Also the parameter time on website has very low correlation with yearly sales so it doesn’t have much effect on sales.

Scenario 1

Now taking all the parameters because we want to get know the effect of all on sales, we get:

Table3: Coefficient Matrix



	Coeffecient
Avg. Session Length	25.981550
Time on App	38.590159
Time on Website	0.190405
Length of Membership	61.279097

From the above table we can get the coefficients of the parameters

TESTING THE MODEL:

A graph between predicted values from the model and the values from the data kept for evaluation is plotted(figure2).

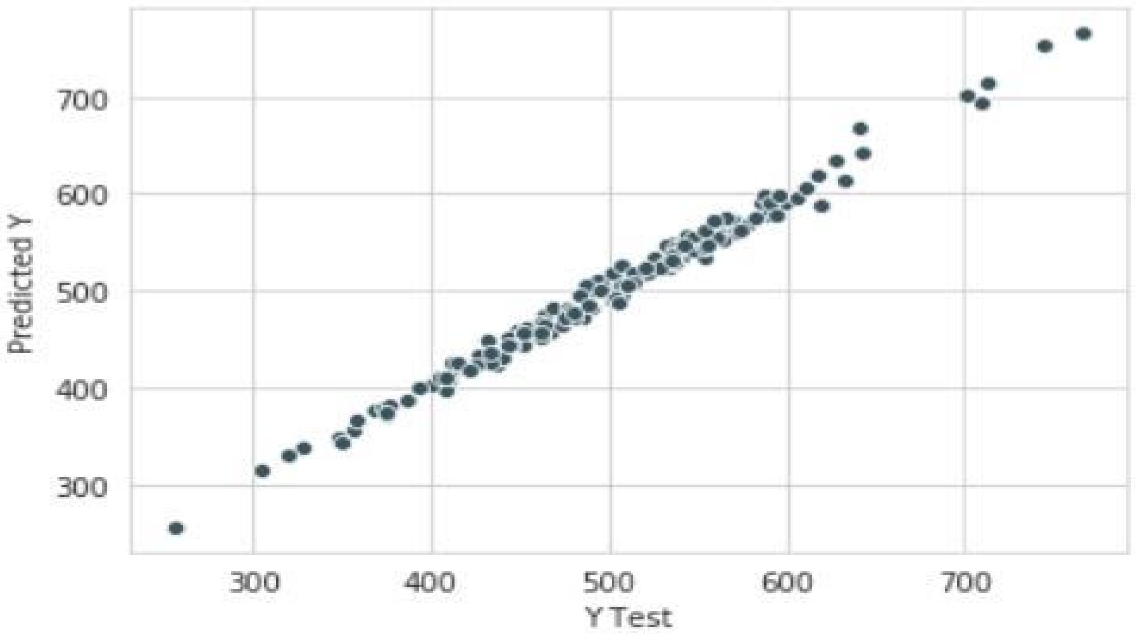


Figure2.

As we can see that it is a linear graph therefore the prediction is almost accurate

EVALUATION:

Evaluating our model’s performance by calculating the residual sum of squares and the explained variance score (R^2)

```
MAE: 7.228148653430853
MSE: 79.81305165097487
RMSE: 8.933815066978656
```

Where MAE stands for absolute mean error and RMSE stands do Root Mean Square Error
Since the values of MAE and RMSE are very low therefore our model is good prediction of the data available

Now Coefficient of determination
R Sq.=0.9890046246741233
We can see that the coefficient of determination is very close to one so our model is very good prediction of actual data set

Residuals:

Residual plot(figure3) is given below.

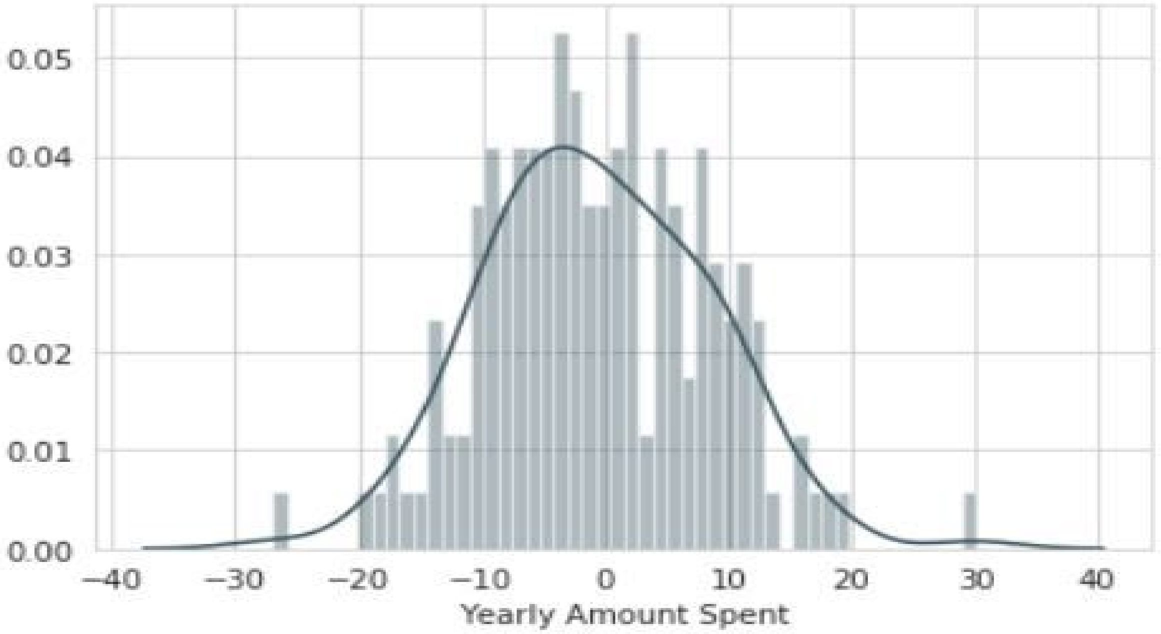
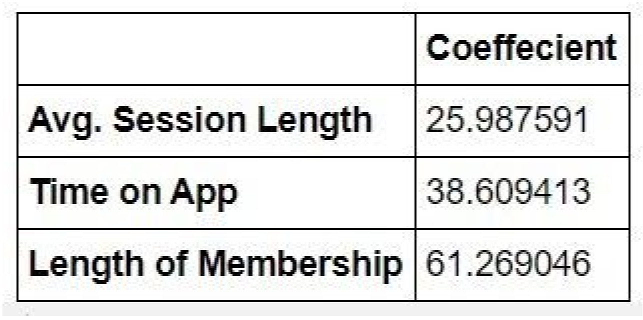


Figure 3.

Scenario 2:

Now Dropping Time on website as it has almost no correlation with sales, we get:

Table 4: Coefficient Matrix



	Coeffecient
Avg. Session Length	25.987591
Time on App	38.609413
Length of Membership	61.269046

From the above table we can get the coefficients.

TESTING THE MODEL:

A graph between predicted values from the model and the values from the data kept for evaluation is plotted.

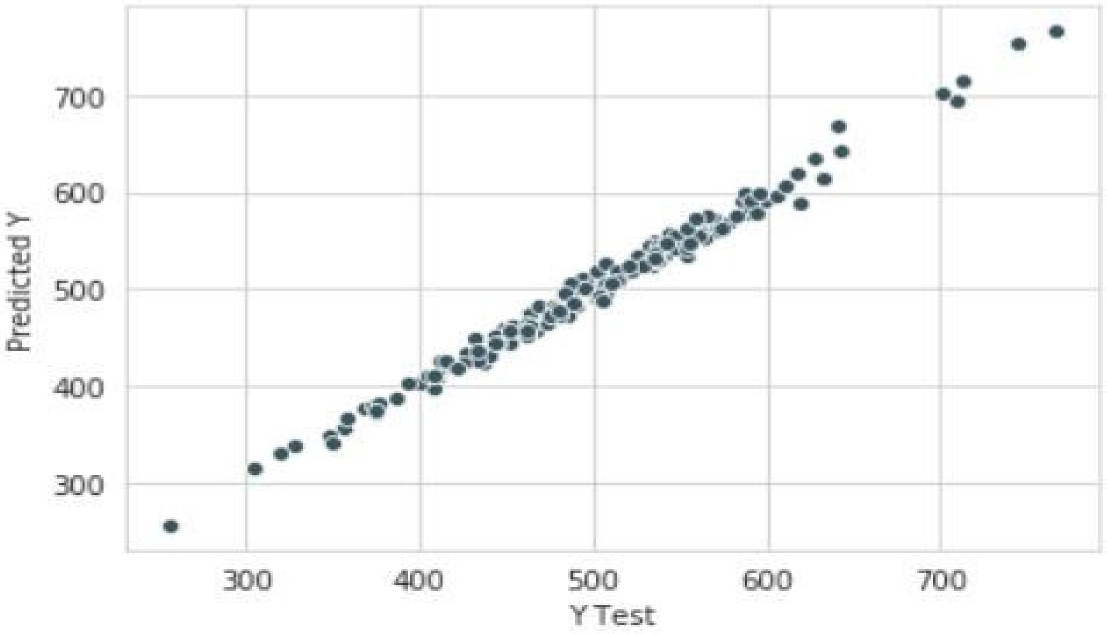


Figure 4.

As we can see from figure-4 that it is a linear graph therefore the prediction is almost accurate

EVALUATION:

Evaluating our model's performance by calculating the residual sum of squares and the explained variance score (R^2)

```
MAE: 7.23694905609139
MSE: 80.19544609397654
RMSE: 8.955191013818553
```

Since the values of MAE and RMSE are very low therefore our model is good prediction of the data available

Now Coefficient of determination

R Sq.=0.9889519444377916

We can see that the coefficient of determination is very close to one so our model is very good prediction of actual data set

Residuals:

Residual plot(figure5) is given below.

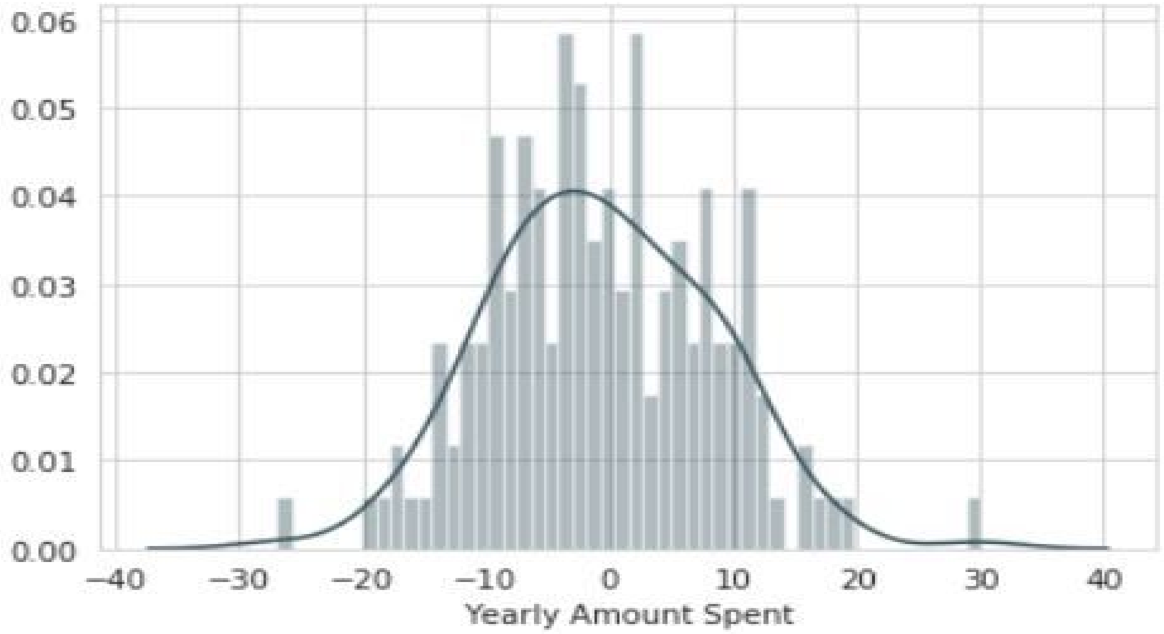


Figure 5.

Conclusion:

Since we can see that there is a linear graph between predicted value and the test data in both the scenarios also the value of R square is very high our models represent data accurately. Now taking scenario 1 because it has a higher value of R square.
Therefore:
y= 25.98 X1 + 38.59 X2 + 0.19 X3+ 61.279 X4
Where,
Y=Yearly amount spent
X1= Average Session length
X2= Time on App
X3= Time on website
X4= Length of membership
Hence the owner should focus more on improving his app and should continue the sessions conducted by him and should divert his focus from the website as there are very less effects of the amount of time spent on the website on his sales.

References:

- https://github.com/sunil-chawla/CE-463-Project/blob/main/463%20Project(2).ipynb(Our Working)
- https://github.com/ravi1728/ce463/blob/main/463%20Project.ipynb(our working)
- https://www.kaggle.com/carrie1/ecommerce-data

Data was sorted through ML due to inadequacy