Ecommerce Customer Study

An Ecommerce company based in New York City sells clothing online, but they also have in-store style and clothing advice sessions. Customers come in to 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 goal of the project is to help the company deciding whether to focus their efforts on their mobile app experience or their website.

Data Exploration

The dataset contains 500 datapoints, each representing one customer; It has 8 columns (7 features, and 1 labels column) as shown below:

```
[8 rows x 5 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Email                 500 non-null    object 
 1   Address               500 non-null    object 
 2   Avatar                500 non-null    object 
 3   Avg. Session Length   500 non-null    float64
 4   Time on App           500 non-null    float64
 5   Time on Website       500 non-null    float64
 6   Length of Membership  500 non-null    float64
 7   Yearly Amount Spent   500 non-null    float64
dtypes: float64(5), object(3)
memory usage: 31.4+ KB
```

The dataset has Customer info, such as Email, Address, and their color Avatar. Then it also has numerical value columns:

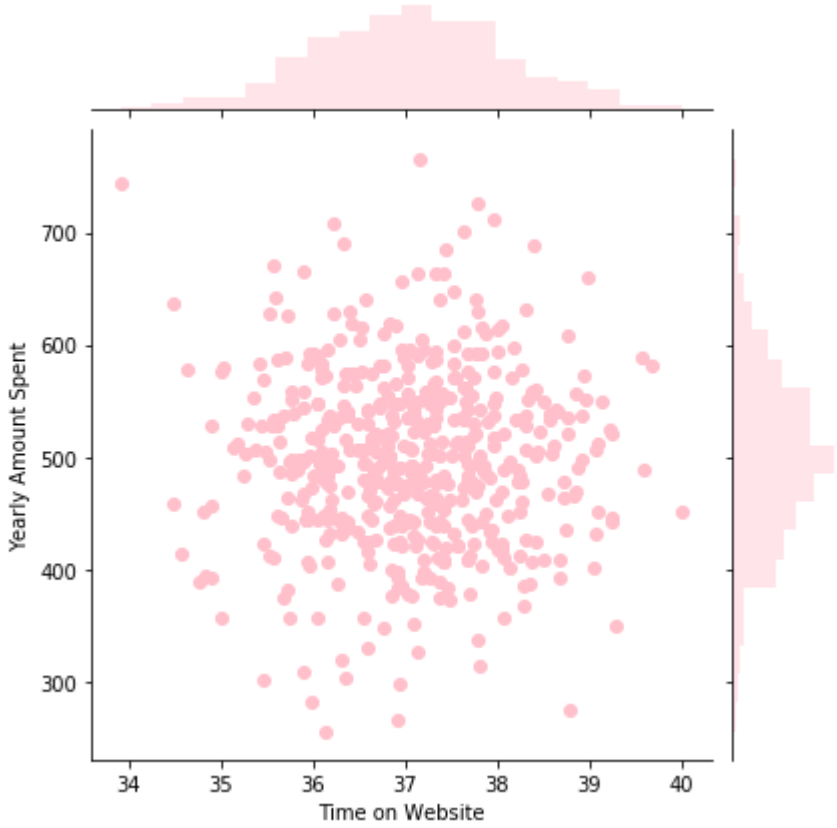
- Avg. Session Length: Average session of in-store style advice sessions in minutes
- Time on App: Average time spent on App in minutes
- Time on Website: Average time spent on the Website in minutes
- Length of Membership: How many years the customer has been a member

The figure below shows dataset head:

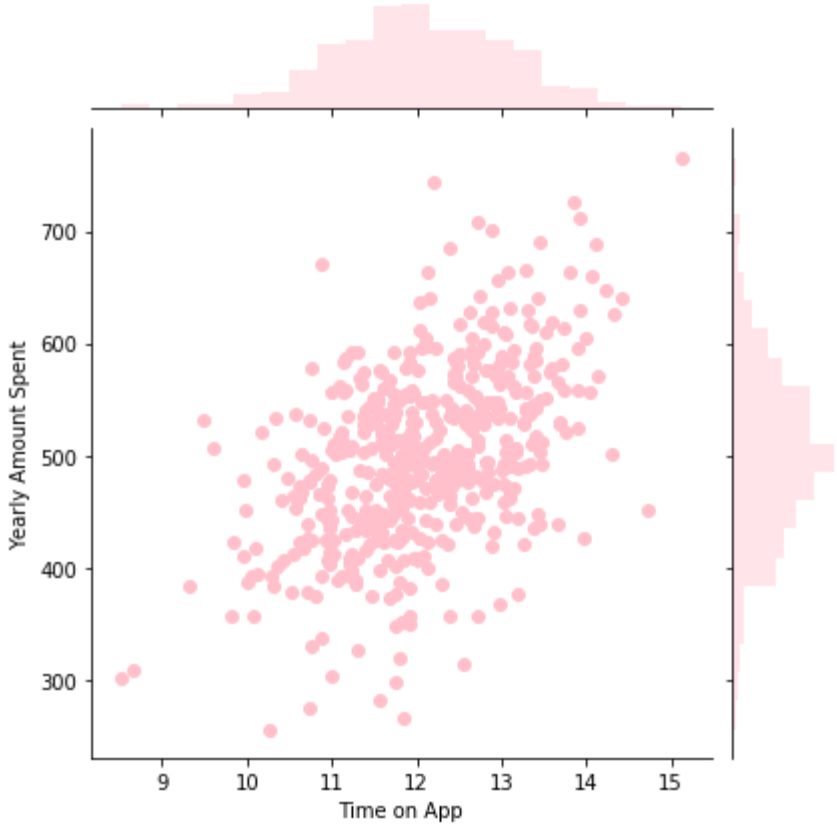
Index	Email	Address	Avatar	Avg. Session Length	Time on App	Time on Website	Length of Membership	Amount
0	mstephenson@fernandez.com	835 Frank Tunnel Wrightmouth, MI 82180-9605	Violet	34.4973	12.6557	39.5777	4.08262	587.951
1	hduke@hotmail.com	4547 Archer Common Diazchester, CA 06566-8576	DarkGreen	31.9263	11.1095	37.269	2.66403	392.205
2	pallen@yahoo.com	24645 Valerie Unions Suite 582 Cobbborough, DC 99414-7564	Bisque	33.0009	11.3303	37.1106	4.10454	487.548
3	riverarebecca@gmail.com	1414 David Throughway Port Jason, OH 22070-1220	SaddleBrown	34.3056	13.7175	36.7213	3.12018	581.852
4	mstephens@davidson-herman.com	14023 Rodriguez Passage Port Jacobville, PR 37242-1057	MediumAquaMarine	33.3307	12.7952	37.5367	4.44631	599.406
5	alvareznancy@lucas.biz	645 Martha Park Apt. 611 Jeffreychester, MN 67218-7250	FloralWhite	33.871	12.0269	34.4769	5.49351	637.102
6	katherine20@yahoo.com	68388 Reyes Lights Suite 692 Josephbury, WV 92213-0247	DarkSlateBlue	32.0216	11.3663	36.6838	4.68502	521.572
7	awatkins@yahoo.com	Unit 6538 Box 8980 DPO AP 09026-4941	Aqua	32.7391	12.352	37.3734	4.43427	549.904
8	vchurch@walter-martinez.com	860 Lee Key West Debra, SD 97450-0495	Salmon	33.9878	13.3862	37.5345	3.27343	570.2
9	bonnie69@lin.biz	PSC 2734, Box 5255 APO AA 98456-7482	Brown	31.9365	11.8141	37.1452	3.20281	427.199

After importing the dataset, we created data analysis displays to explore it.

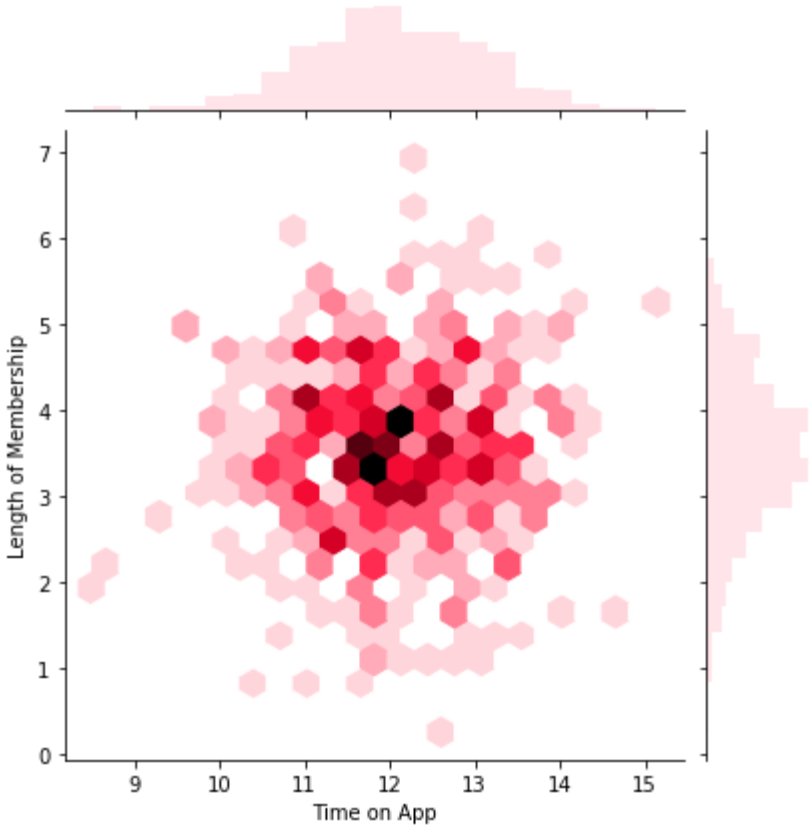
- Comparing the ‘Time on Website’ feature and ‘Yearly Amount Spent’ column (the labels). It looks like there is no visible correlation between this feature and the labels.



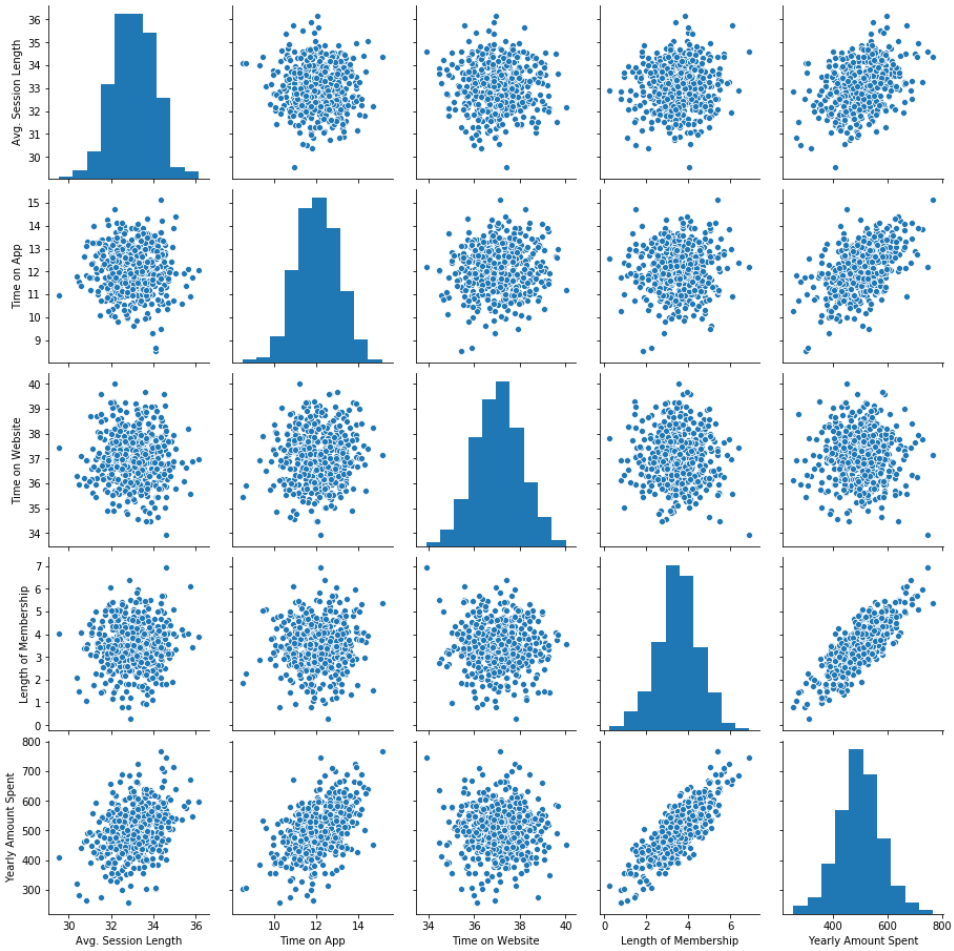
- Comparing the ‘Time on App’ feature and ‘Yearly Amount Spent’ column. Here we can see some correlation that indicates the more time costumers spend on the app, the more money they spend on purchases.



- Comparing ‘Time on App’ feature and ‘Length of Membership’ feature. Again, no strong correlation is visible.



- Exploring the types of relationships across the entire data set Using pairplots. Here we can see that most features are not strongly correlated, except of 'length of membership' feature, which correlates with our labels ('Yearly Amount Spent') as seen on the bottom right of the figure below.



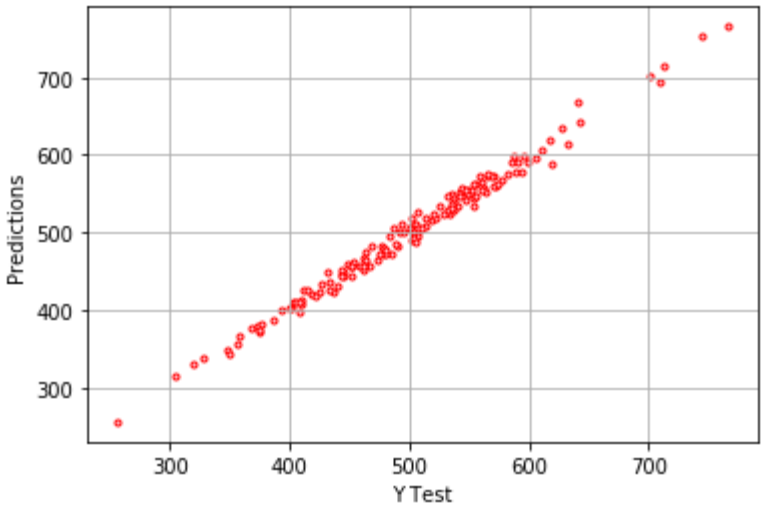
Data Preparation and Training

After exploring the dataset, and making sure no datapoint values are missing; now we split the data into training and testing sets; then we used the training data and labels to train our linear regression model. The resultant model has the coefficients shown below:

model coeffs:	Coefficient
Avg. Session Length	25.981550
Time on App	38.590159
Time on Website	0.190405
Length of Membership	61.279097

Model Testing

Now that we have fit our model, we went to evaluating its performance by predicting off the test labels. The figure below shows the relationship between our predicted labels, and the true values of the labels on the test dataset, this figure shows that our predictions are actually very accurate.



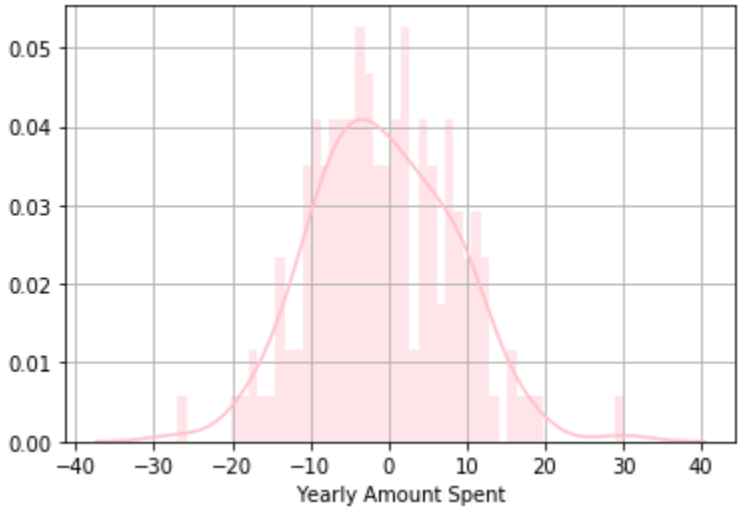
Model Evaluation

Residual sum of squares are calculated to measure the accuracy of our model.

- MAE: Mean absolute error
- MSE: Mean squared error
- RMSE: Square roof of mean squared error

```
MAE: 7.22814865343083
MSE: 79.81305165097461
RMSE: 8.933815066978642
```

To make sure everything was okay with our data, we went to explore the residuals as shown in the figure below. Following a normal distribution indicates a very good model with a good fit.



Conclusion

It is clear from our model, that 'time on website' feature does not have a sizable effect on money spent by customers, so the company can save money by non focusing on their website. On the other hand, 'time on app' is actually effective, a minute more on the app associates to $38.5 increase in spending, in case no other

features have changed. Finally, 'length of membership' is the feature with the highest correlation to spending, so the company should make sure to take care of its old customers.

Also it is worth mentioning that the company can focus its efforts on its website, in order to improve the spending of its website visitors.