**Retail E-commerce Store**

A retail e-commerce website based in UK that sells clothing online but also have in-store style and clothing advice sessions.

Often times, Customers come into the store, have sessions with a personal stylist, then they can go home and order either on a website or mobile app for the clothes they want.

**The company is trying to decide whether to focus their efforts on their mobile app experience or their website.**

Before starting with the project, there is a need to import the necessary libraries and in this project pandas, numpy, seaborn, matplotlib and scikit-learn for training the machine learning model were used.

After importing the necessary libraries to make the analysis fun, there is a need to import the data (which was done using pandas). Lot of work of a data scientist revolve around cleaning datasets that need to be analyzed such as (checking for missing value, converting categorical features to numerical, standardized features and the likes). In this project, the data has been cleaned so we don’t need to spend most of the time on data cleaning.

Now let’s get some statistical information about the dataset..

**Firstly**, we observed that the dataset consists of eight (8) columns and 500 rows.

**Secondly,** out of the eight columns, five are numerical features and the remaining three are object. So we check for those features that are needed for the analysis which was segment into three: **1. Important 2. Not Important 3. In-between.**

**Important features :** Time spent on App, Time spent on website, Length of Membership, Yearly amount spent (which is the dependent variable).

**Not Important features :** Email, Address and Avatar.

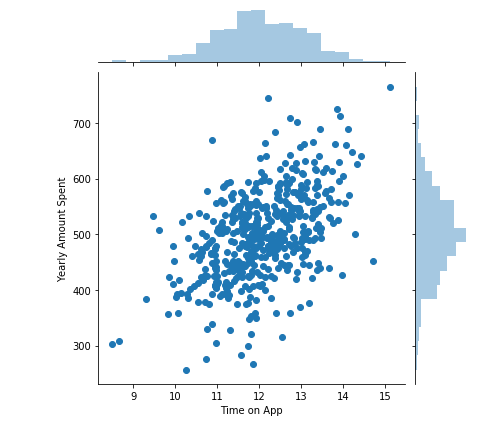
**In-between :** Avg Session Length

So we need to drop the **Not important** features since they’re not that important to the business problem.

Now that we’ve an idea about the dataset that we need for the project, let’s do some **Exploratory Data Analysis (EDA).**

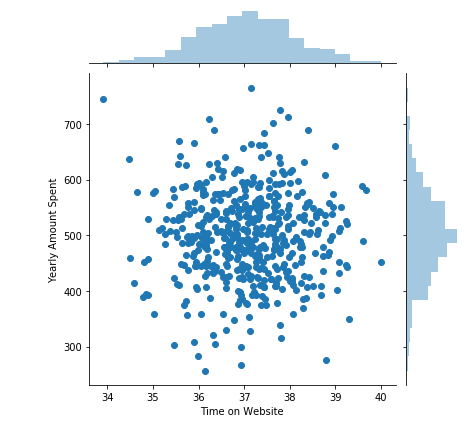
EDA allows us to gain deeper insights about the datasets alongside its features.

The first EDA (visualization) that was performed was to compare the **Time on App** and **Yearly amount Spent** by each customers to check for the correlation.



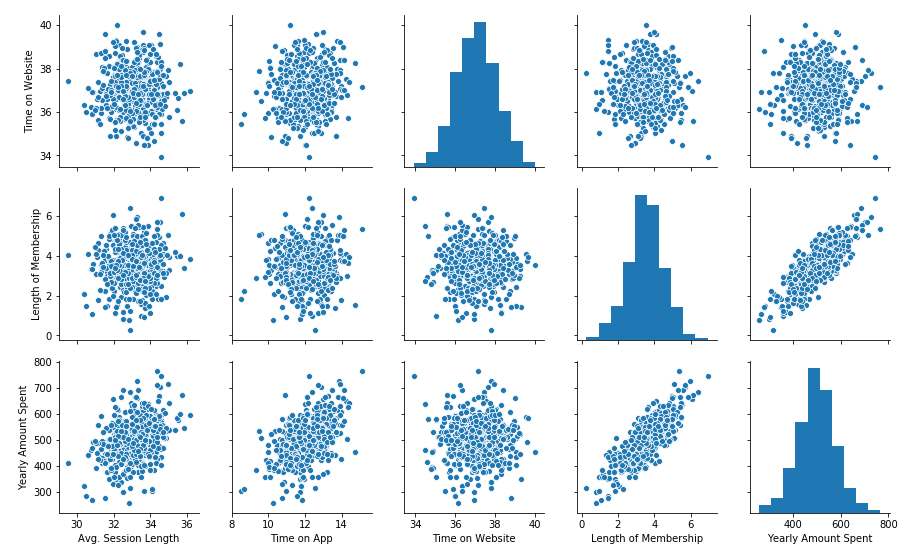
From the plot, there seems to be some sought of correlation between Time on App and Yearly amount Spent. That is, an increase in the time spent on app also increases the Yearly Amount Spent by the customers.

Let’s check for the **Time on website** and the **Yearly amount spent** by the customers.

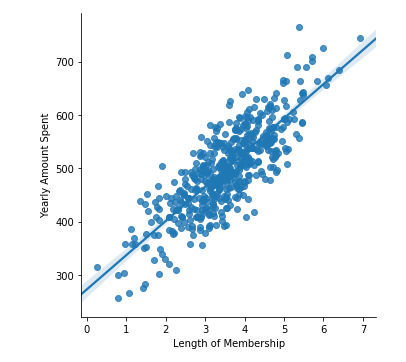


From the plot, there seems to be no correlation between **Time on Website** and **Yearly amount Spent**. Which tends to give us some clue about the problem statement.

Now let’s check the whole dataset and check for the feature that has a high correlation with the **Yearly amount spent.**



The **Length of Membership** is highly correlated with the **Yearly amount Spent** which make sense. That is, the longer you stay as member affect the yearly amount the customer spent. Let’s visualize it to see it clearly.



**Training and Testing sets**

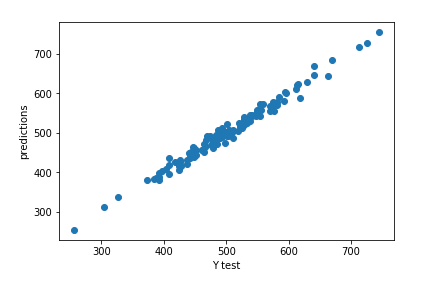
Now that the dataset has been explored, let’s split the data into training and testing sets .

Since the variable **“Yearly amount spent”** is our dependent feature, we drop it and set the remaining features (independent features) to **X variable** while the dependent feature was assigned to **y variable.**

The training and testing sets has been done, and the LinearRegression model has been fit on the training set. The coefficient and the intercept has also been checked. Let’s proceed.

One good thing about applying machine learning model on a dataset is to be able to evaluate how that model performed on the dataset. We evaluate the model performance off the test values (unknown values).

So a scatter plot was created to check for the real test values versus the predicted values.



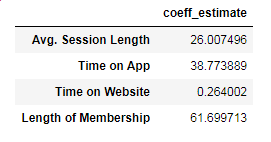
Notice that the model is actually doing quiet well which the scatter plot perfectly shown a linear line and that make sense.

**Evaluating the Model**

The model was evaluated using the Mean Absolute Error (MAE), Root Mean Squared Error(RMSE), Mean Squared Error (MSE).

Now that we’ve done the data exploration, fit our model on the dataset, we still want to figure out the answer to the original question. Which is, should the e-commerce shop focus their effort on mobile app or website development? Or perhaps may be the length of membership is what really matters.

**The coefficients was created to interpret the data.**



**In conclusion**, the average session length if you hold the other features fixed, a unit increase in average session length is related to an increase in 26 dollars spent. Likewise a unit increase in Time spent on the app is related to 39 dollars spent per year. The time on website only show an increase in 26 cent - which is not that great. By far the strongest feature which is Length of Membership. If you hold other features fixed, a unit increase in length of membership is related to around 62 dollars spent.

The question now is do you think the company should focus their time on the **website** or on the **mobile app**? There are numerous ways to think about this, but based on the result of the coefficient on the data and the model, the company could focus their efforts on redeveloping the website to meet up to the mobile app standards or better still since the mobile app is working already the company can still improve it.

The relationship between length of membership and the time spent on app can also be explored before coming to a conclusion!