

report on arvilla ecommerce clothing line customers experience using mobile app or website

Zep Internship Task



November 9, 2022

linear regression project by jerry poland

NOTE: This was originally a capstone Machine learning project by Pierian Data (A Data science Training company)

**1. PROBLEM STATEMENT**

Arhilla is an Ecommerce company based in New York City that 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 company is trying to decide whether to focus their efforts on their mobile app experience or their website. They've hired an analyst on contract to help them figure it out!

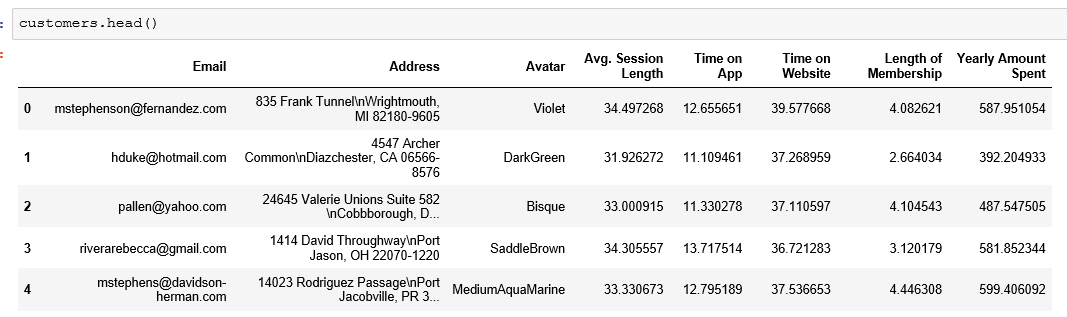
The key objective of this analysis is to use Linear regression model, to develop a modelling framework that will tell whether the company should focus their efforts on their mobile app or their website.

**2. THE DATASET**

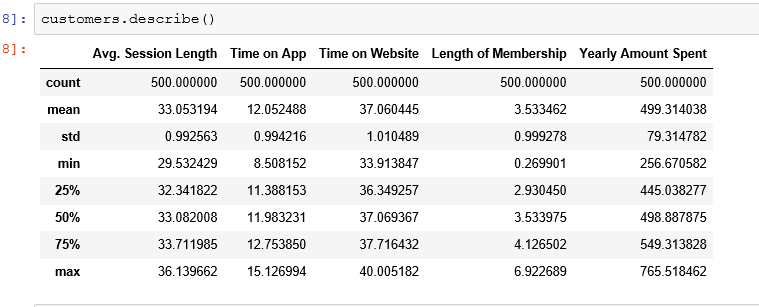
The dataset used is the Ecommerce Customers csv file from the company. It has Customer info, such as Email, Address, and their color Avatar. Then it also has numerical value columns:

* Avg. Session Length: Average session of in-store style advice sessions.
* 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.

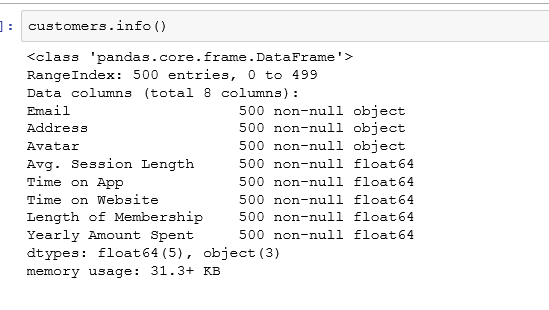
Lets have a view of the first five rows of the dataset



Lets look at the description of the dataset



Let us as well consider the general information of the dataset



**3. DATA CLEANING**

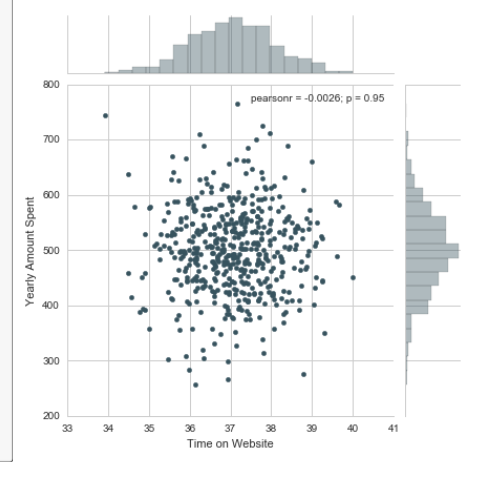
Data cleaning is a very important aspect in any data science project. The cleanness of our data determines the accuracy of our model. Looking at the ‘customers.info()’, we can tell the dataset is clean as there are no null values.

**4. EXPLORATORY DATA ANALYSIS**

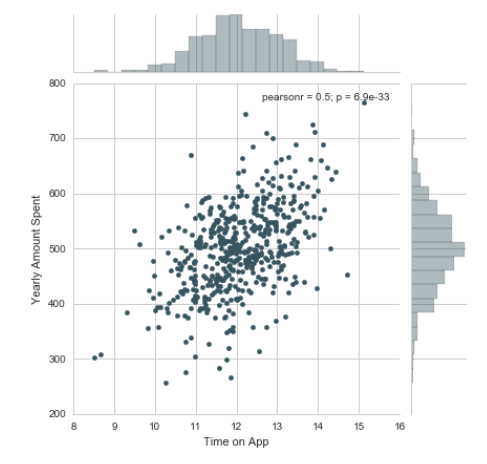
Exploratory Data Analysis is mandatory to explore in depth our data in other to draw meaningful insights from it. We will be using the numerical data of the csv file.

**Correlation between columns.**

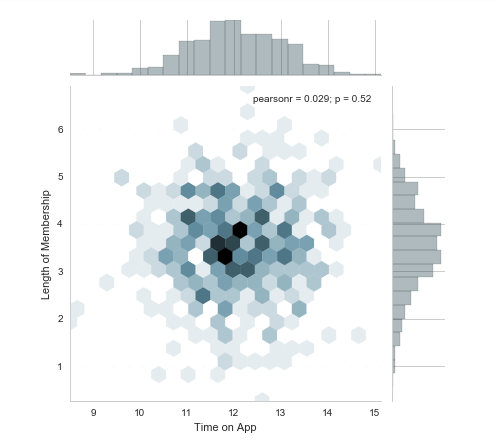
We will use seaborn to create a jointplot in order to make these comparisons.



*Correlation between time on website and yearly amount spent columns*

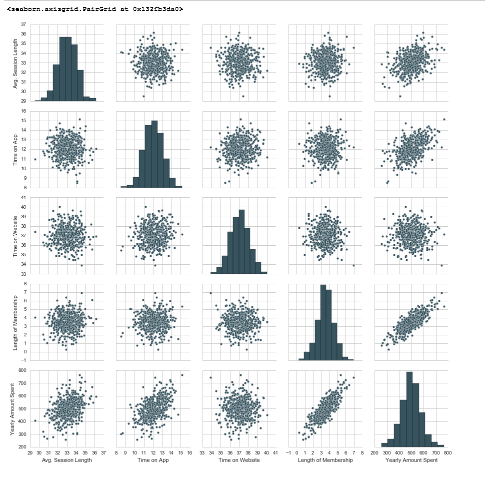


*Correlation between time on app and yearly amount spent columns*



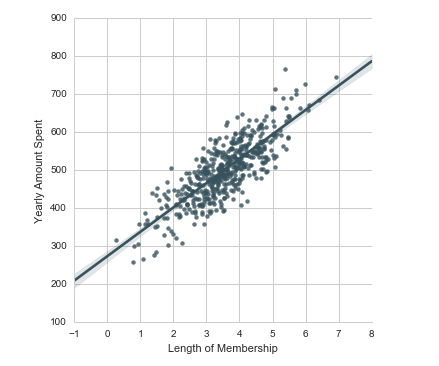
*2D hex bin plot comparing Time on App and Length of Membership*

**Let's explore these types of relationships across the entire data set using pairplot**



**Based off the plot above what looks to be the most correlated feature with Yearly Amount Spent is the length of membership.**

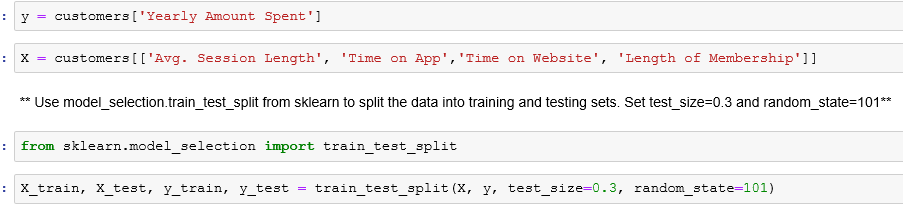
Thus we will create a linear model plot (using seaborn's lmplot) of Yearly Amount Spent vs. Length of Membership.



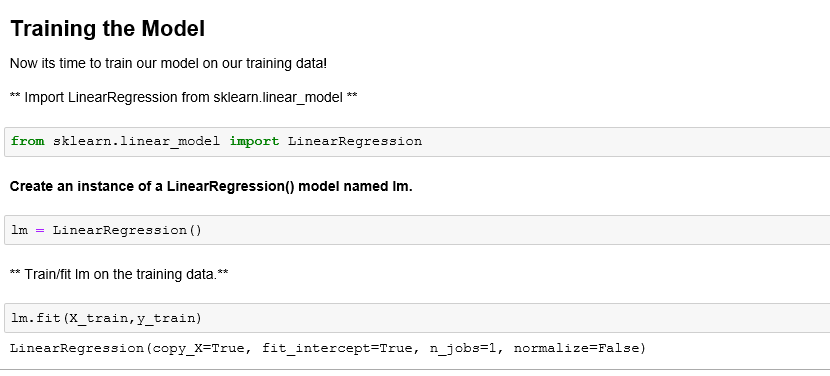
**5. MODELING BUILDING**

Now that we've explored the data a bit, let's go ahead and split the data into training and testing sets. Set a variable X equal to the numerical features of the customers and a variable y equal to the "Yearly Amount Spent" column.

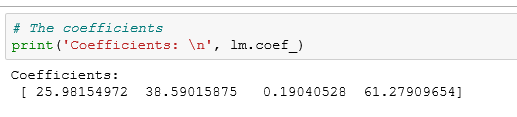
**Training and testing data**

****

**Training the model**

****

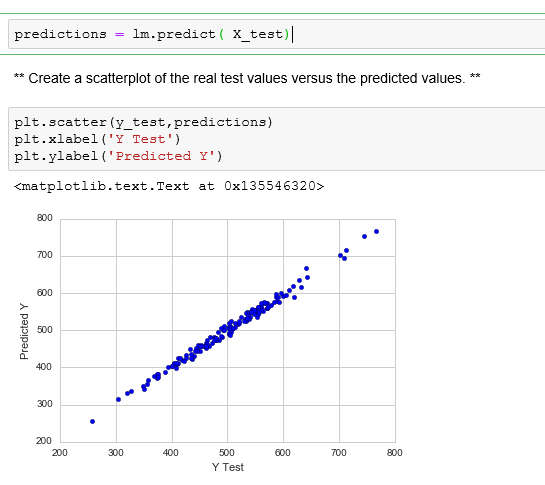
**Print out the coefficient of the model**

****

**Predicting Test Data**

Now that we have fit our model, let's evaluate its performance by predicting off the test values!

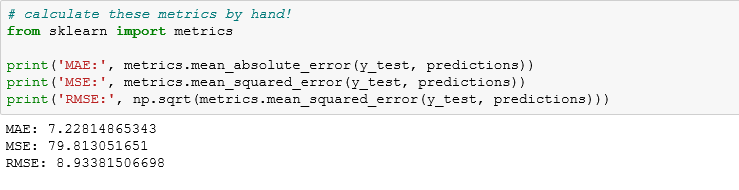
\*\* Using lm.predict() to predict off the X\_test set of the data.\*\*



**6. EVALUATING THE MODEL**

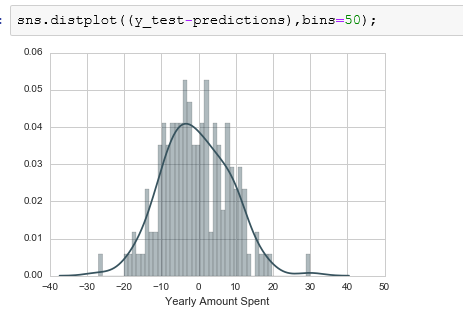
Let's evaluate our model performance by calculating the residual sum of squares and the explained variance score (R^2).

\*\* Calculate the Mean Absolute Error, Mean Squared Error, and the Root Mean Squared Error.\*\*

****

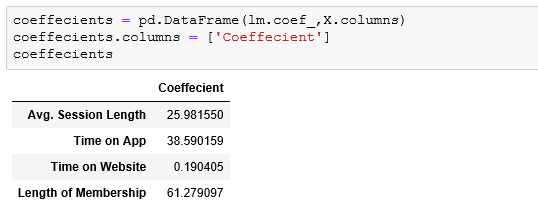
**Residual**

**Plot a histogram of the residuals and make sure it looks normally distributed. Use either seaborn distplot**

****

**7. FINAL INTERPRETATION AND CONCLUSION**

We still want to figure out the answer to the original question, do we focus our effort on mobile app or website development? Or maybe that doesn't even really matter, and Membership Time is what is really important. Let's see if we can interpret the coefficients at all to get an idea.

****

Interpreting the coefficients:

* Holding all other features fixed, a 1 unit increase in **Avg. Session Length** is associated with an **increase of 25.98 total dollars spent**.
* Holding all other features fixed, a 1 unit increase in **Time on App** is associated with an **increase of 38.59 total dollars spent**.
* Holding all other features fixed, a 1 unit increase in **Time on Website** is associated with an **increase of 0.19 total dollars spent**.
* Holding all other features fixed, a 1 unit increase in **Length of Membership** is associated with an **increase of 61.27 total dollars spent**.

**Do you think the company should focus more on their mobile app or on their website?**

This is tricky, there are two ways to think about this: Develop the Website to catch up to the performance of the mobile app, or develop the app more since that is what is working better. This sort of answer really depends on the other factors going on at the company, we would probably want to explore the relationship between Length of Membership and the App or the Website before coming to a conclusion