

**E-commerce Customer Prediction**

**Regression Analysis Using Python**

* **By**
* ***Pritam Bhakta***

**Acknowledgement**

**We** would like to express our special thanks of gratitude to our mentor Ms. Tania Chakraborty who gave us the golden opportunity to do this wonderful project on the topic ***E-commerce Customer Prediction Regression Analysis Using Python***, which also helped us in doing a lot of research and we came to know about so many new things as well as we learned how to use internet resources for our purpose, we are really thankful to our mentor and also to the internet resources.  
Secondly we would also like to thank our parents and group members who contributed their thoughts a lot in finalizing this project within the limited time frame.

**E-commerce**

**What is E-commerce?**

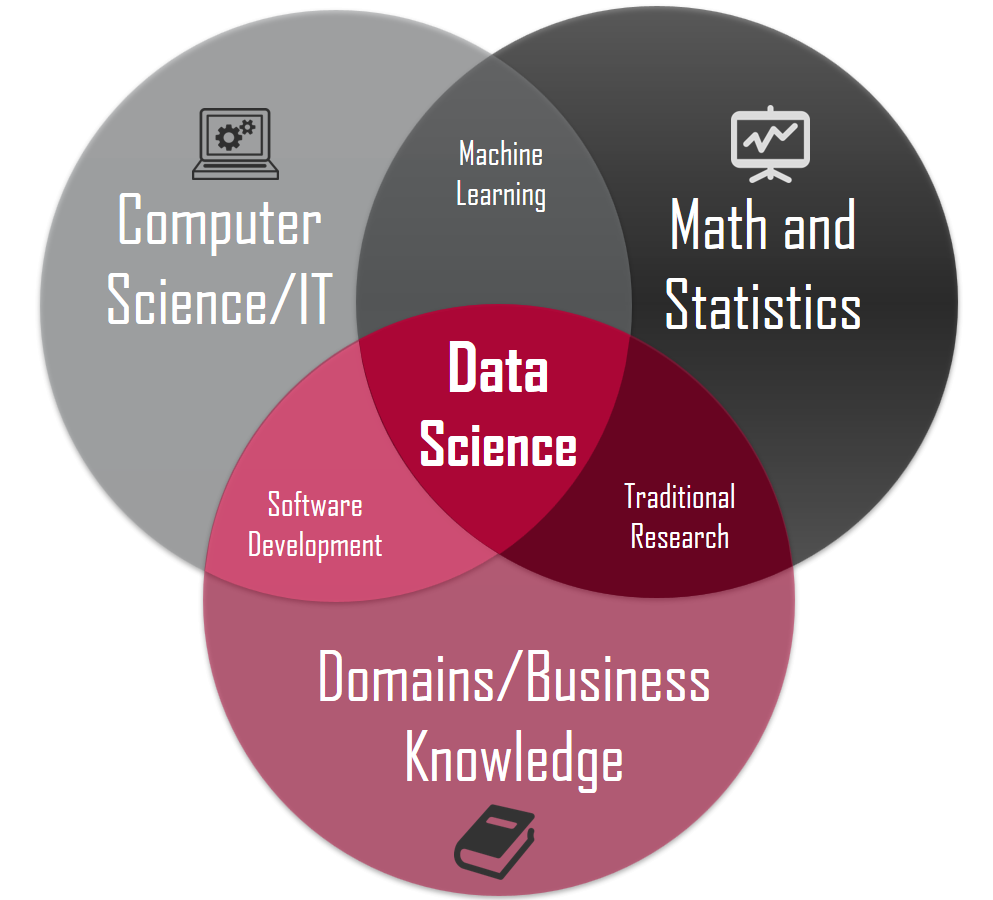
E-commerce refers to commercial transactions of [goods](https://corporatefinanceinstitute.com/resources/knowledge/accounting/cost-of-goods-sold-cogs/) or services conducted over the internet.  Over the past several years, e-commerce has rapidly evolved to become a combination of online and offline retail that is [vertically integrated](https://corporatefinanceinstitute.com/resources/knowledge/strategy/vertical-integration/).  You can find numerous e-commerce companies selling various types of products and services. Their avenues of doing business are typically divided into three main categories:



Some of the major players in the e-commerce industry, such as [Amazon](https://en.wikipedia.org/wiki/Amazon_(company)), Alibaba, and eBay, are well known by the public and own a large proportion of the [market share](https://corporatefinanceinstitute.com/resources/knowledge/economics/monopoly/).  These companies sell products of various brands, while other companies, such as Zalando, ASOS and MI also offer products of their own brands.

With an ever-increasing level of marketplace competition, there is more and more overlap of the kinds of goods and services retailers provide. For example, while still primarily a gateway for third-party sellers, Amazon is increasingly developing and marketing its own brand of products.

**Data Science**



# **Data Science**

## Definition - What does *Data Science* mean?

Data science is a broad field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data. It is geared toward helping individuals and organizations make better decisions from stored, consumed and managed data.

Data science is formerly known as datalogy.

*Data Science*

Data science enables the use of theoretical, mathematical, computational and other practical methods to study and evaluate data. The key objective is to extract required or valuable information that may be used for multiple purposes, such as decision making, product development, trend analysis and forecasting.

**Components of Data Science**

**A Definition of Data Management**

Data Management is an administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users. Organizations and enterprises are making use of Big Data more than ever before to inform business decisions and gain deep insights into customer behavior, trends, and opportunities for creating extraordinary customer experiences.

**Data Management ↔↔** **Types of Data**

* ***Structured Data***
* ***Unstructured Data***
* ***Semistructured Data***
* ***Structured Data***

**Structure Data** is **data** that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis.

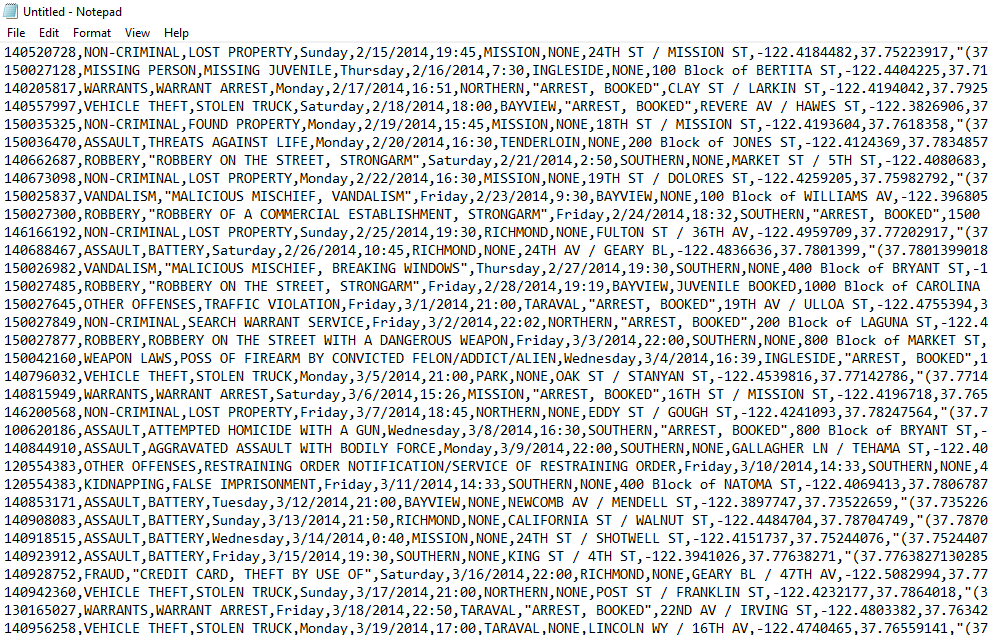
Example**:**

****

* ***Semi-structured Data***

**Semi-structured data** is data that is neither raw data, nor typed data in a conventional database system. Example CSV (comma Separated Value), XML, JSON (JavaScript Object Notation)

Example**:**

****

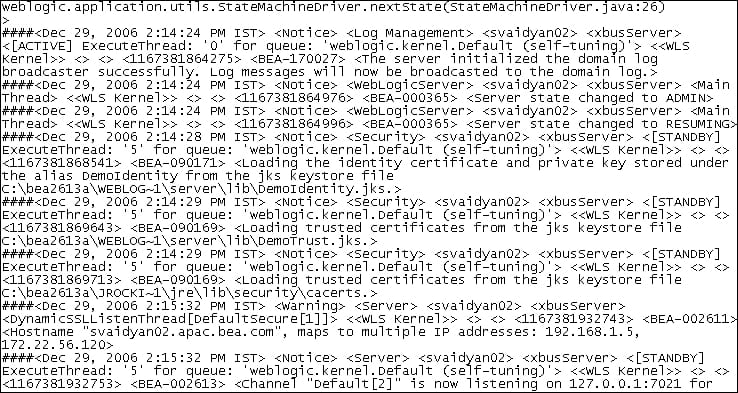
* ***Unstructured Data***

The phrase **Unstructured Data** usually refers to information that doesn’t reside in a traditional row-column database.

Unstructured Data files often include text and multimedia content.

Examples include e-mail messages, videos, photos , audio files , presentations , text corpus etc.

Example**:**

****

**Data Management Tools**

* ***Structured Data* 🡺 Oracle, MySQL, Sql, CSV/Excel.**
* ***Unstructured Data* ⮷**

**🡺****

* ***Semistructured Data* ⮵**

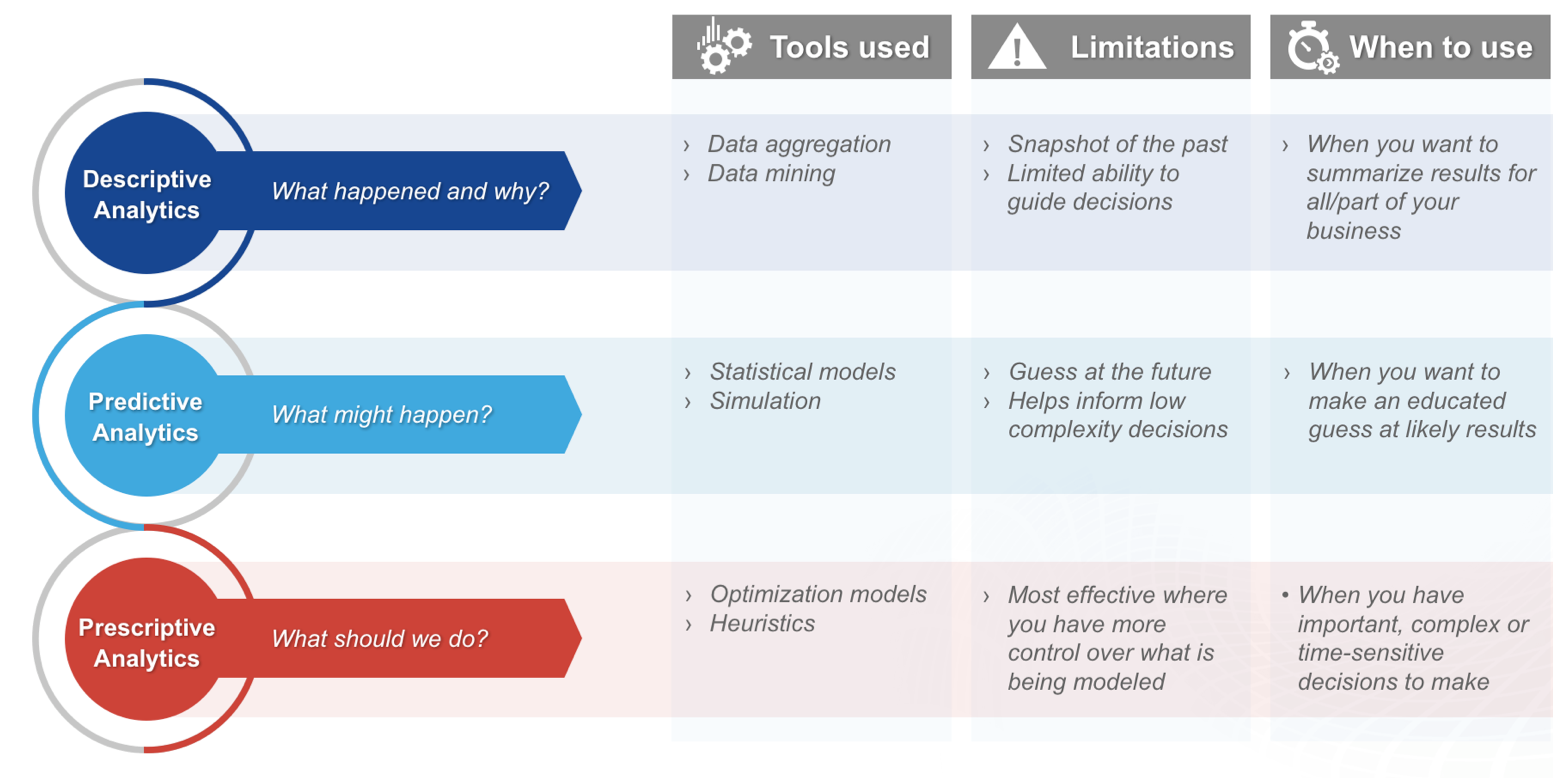
**PYTHON**

**Python is open source (free to use) high level language supported in most of the commonly used operating systems like** Mac, Windows, and Unix and has also been ported to [Java](https://techterms.com/definition/java) and .NET virtual machines.

‘Pandas’ is the **Python Data Analysis** Library, used for everything from importing **data** from Excel spreadsheets to processing sets for time-series **analysis**. SciPy is the scientific equivalent of NumPy, offering tools and techniques for **analysis** of scientific **data**. Stats models focuses on tools for **statistical analysis**.

**Data Analytics**

* **Descriptive** Analytics, which use data aggregation and data mining to provide insight into the past and answer: “What has happened?”
* **Predictive** Analytics, which use statistical models and forecasts techniques to understand the future and answer: “What could happen?”
* **Prescriptive** Analytics, which use optimization and simulation algorithms to advice on possible outcomes and answer: “What should we do?”



Relation Between Data Science and eCommerce Business

The nature of business in the eCommerce industry involves many trackable customer touchpoints like clicking an ad, viewing an item, making a purchase, or submitting a product rating. The resulting volume of information can be overwhelming to navigate, but it’s a data science gold mine. That’s because data scientists excel in sifting through big data to identify causal patterns or make predictions about what’s in store in the future for your business.

The biggest way that data science is useful in eCommerce is through [customer lifetime value (LTV) modeling](https://www.datascience.com/resources/video/customer-lifetime-value-modeling-data-science). With this approach, you can predict the future revenue that each customer will bring to your business in a given period as a function of the length of time an individual is likely to remain a customer, how often they’re likely to make purchases, and the average value of each purchase. Understanding how your most valuable customers are typically introduced to your business allows you to focus your marketing efforts on the right channels.

**Project Description**

Our project is mainly focused in making the prior decisions and also suggestions to company by analyzing the previously conserved Raw data given by an well-known *E-commerce* company of New York , USA .

We got a contract work with above mentioned company that sells clothing online but they also have in-store style and clothing advice sessions.

Now company is trying to decide whether to focus their efforts on their mobile app or their websites.

**Project OBJECTIVE**

**To Analyze the Best Option for the company to get more Profit by investing the Time and Money whether in App or Website**

**Project code**

**Data description**

We just got some contract work with 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 or their website.

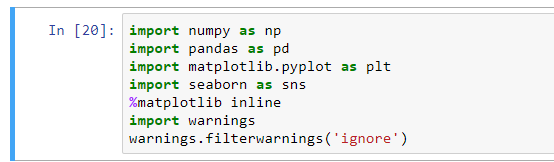
**The Dataset**

We'll work with the Ecommerce Customers csv file from the company. It has Customer info such as Email, Address, and their color Avatar and these data are categorical data. 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.
* Yearly Amount Spent: How much a customer spent in a year.

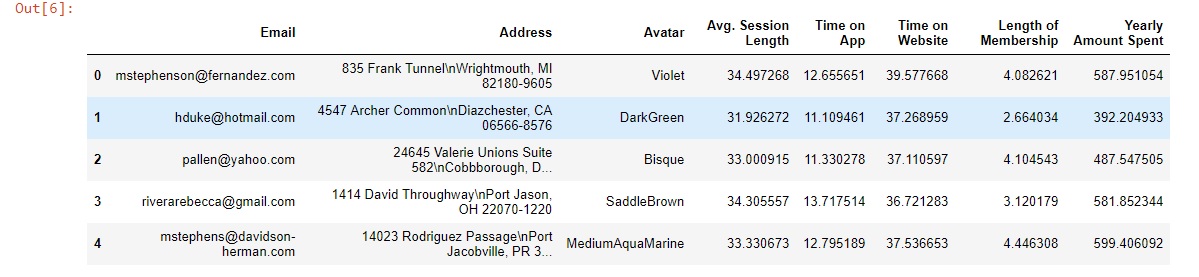
**Importing the dataset**

At first we have to import the packages of python for data visualization such as ‘NumPy’, ‘pandas’, ‘Matplotlib’, ‘seaborn’.



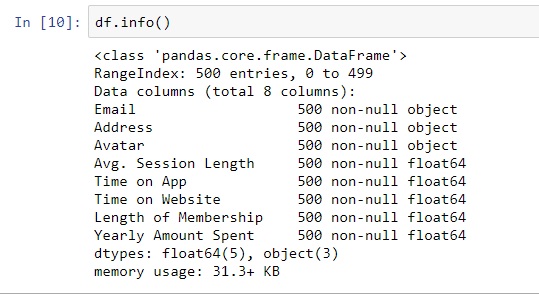
Next, we can import the dataset using read\_csv ( ) method from ‘pandas’ package and store it into df data frame.

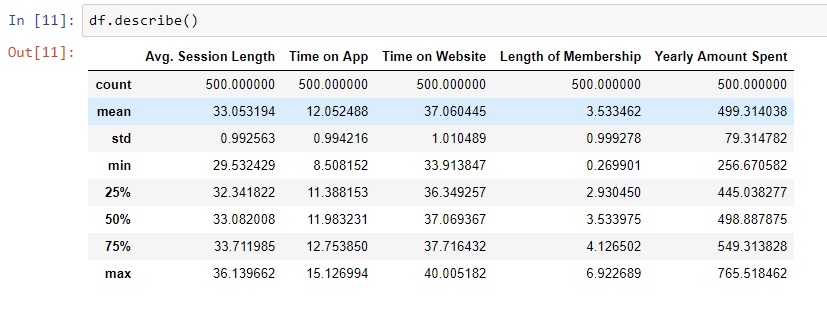


And the data frame is like 

Let use some methods.

Using info ( ) we can get total information of that dataset.



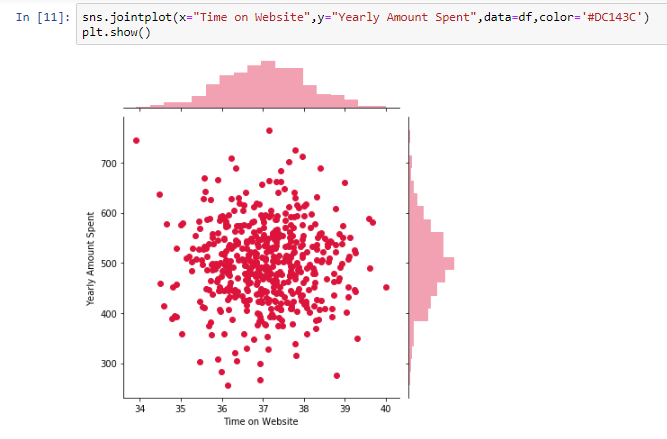
Using describe ( ) method we get values of count, mean, standard daviation, minimum and maximum values.

**Descriptive Analysis**

Now that we’ve imported the data, we’re ready to analyse it. We can plot some graphs using the [***Seaborn***](http://seaborn.pydata.org/) library and see if we can find any peculiar relationships between columns (or ‘features’) in our dataset.

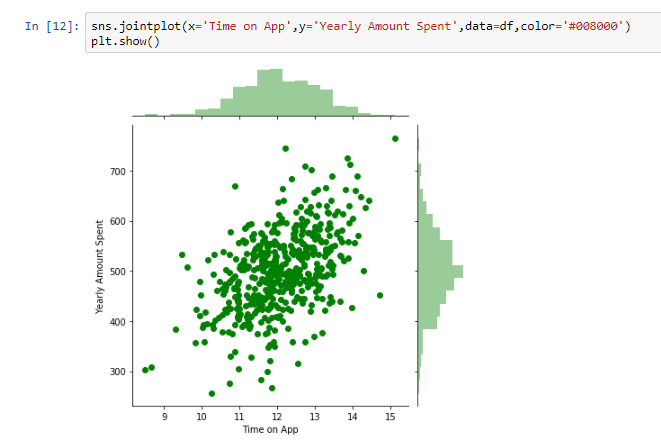
Seaborn allows us to create [***jointplots***](http://seaborn.pydata.org/generated/seaborn.jointplot.html?highlight=jointplot#seaborn.jointplot) comparing two different features.

1.jpg



In above plot we compare Time on Website and Yearly Amount Spent using jointplot. It seems like isn’t much of a correlation between these two features. Let’s build another jointplot to see if there’s any correlation between the time spent on the mobile app and the yearly amount spent.

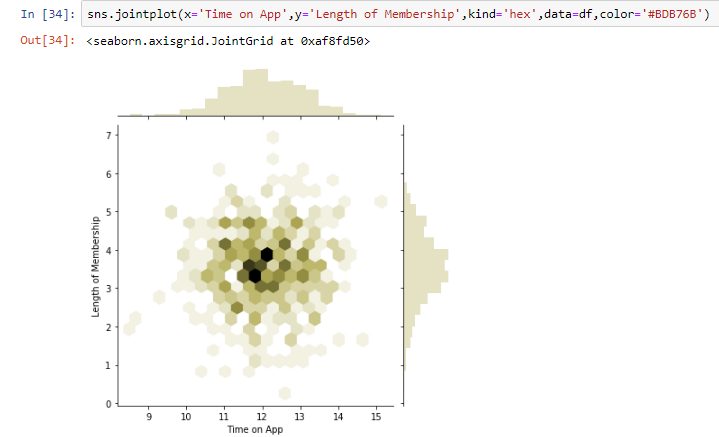
2.png



By this jointplot we compare Time on App with Yearly Amount Spent. There’s a slightly **stronger** correlation between these two features, compared to the previous plot.

There is another interesting feature in jointplot that is ‘hex’ kind plot.

4.png

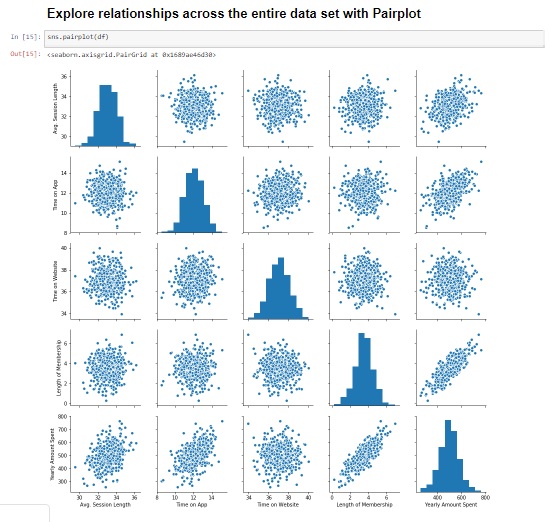


In above plot we create a 2D hex bin plot comparing Time on App with Length of Membership using jointplot.

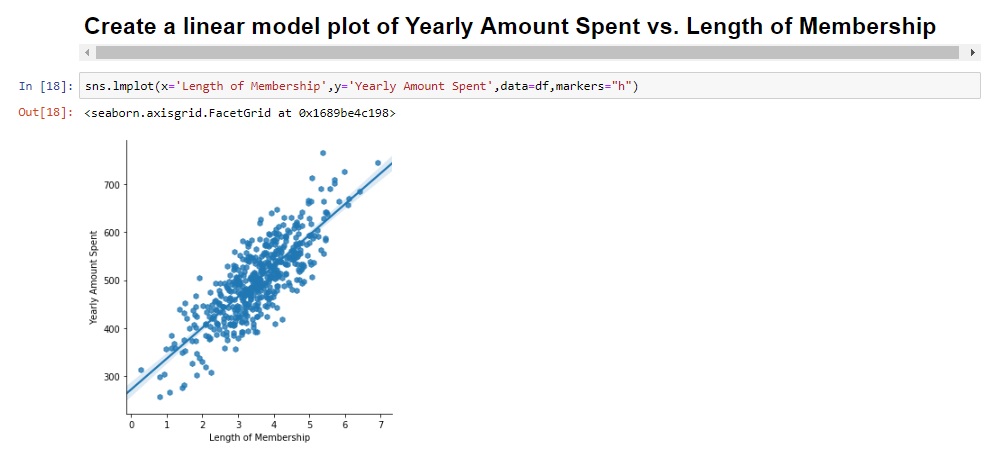
Another plot of ‘seaborn’ is ***pairplot*** which automatically create jointplots for all pair combinations of features in the dataset.



The output will be -



Based off this plot, it looks like the **length of membership** is the strongest correlated feature with the yearly amount spent



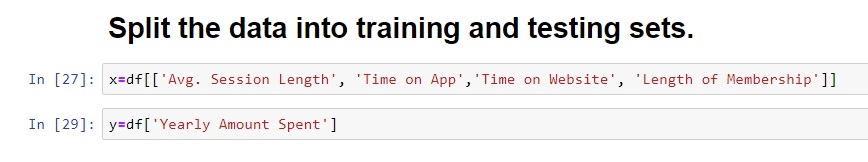
By this linear model plot we can say it is **strong** correlation.

**Predictive Analysis**

For predictive analysis at first we have divide the data into two parts: Train Data & Test Data.

**Training and Testing Data**

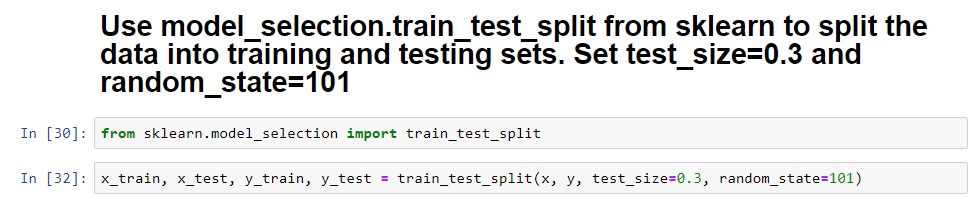
Our ultimate goal for the project is to boost the yearly amount spent for each customer, so we can use that feature as the dependent variable y for our regression. The other numeric columns will make up the set of independent variables x.



After splitting the dataset into a training set and a testing set, the training set will contain the values that the model will learn, and the testing set will contain the values that we can use to test the model’s accuracy.

We can split the X and y data frames into training and testing sets using train\_test\_split. While splitting the data, we can also specify what percent of the original dataset will be used as the testing set.

In our case, we will use **30%** of the dataset for testing, whereas the model will train on the remaining **70%**.

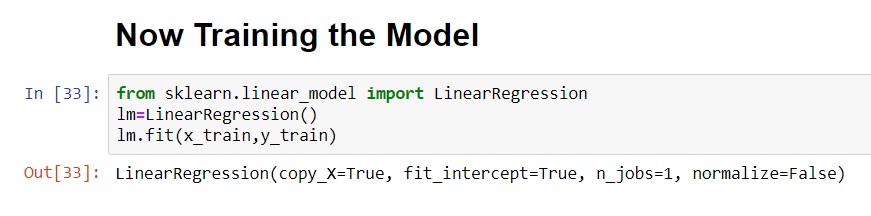


Now, our train data is contained in x\_train and y\_train, whereas our test data is contained in x\_test and y\_test.

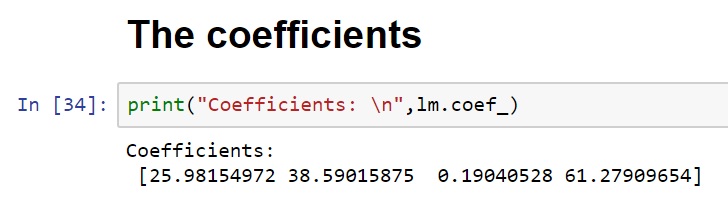
**Training our Model**

Now we should build our model. Since we want to fit a linear regression model on our data, let’s use the [**Linear** **Regression**](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html) module from **Sklearn**.

Now, we need to fit our model lm to the training set. This task sounds complicated, but Sklearn allows us to train the model easily using the .fit () method. We will pass in the training data *(*X\_train*and*y\_train*)* as the parameters.



Our model has now been trained. Let’s see what coefficients our model has chosen for each of our independent variables. It’s important to check the coefficients; because they tell us how influential each feature is over the yearly amount spent. We can take a look at the coefficients by calling .coef\_ on our model.

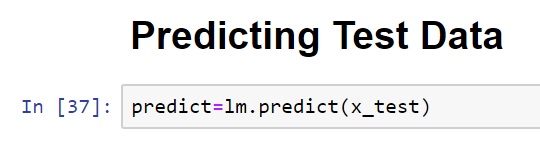
.

Now it’s time to test our model.

**Testing our Model**

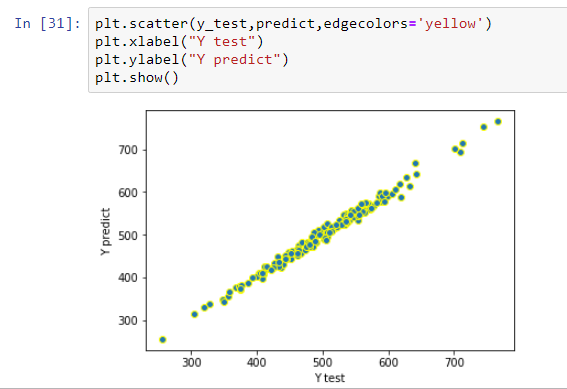
Now, let’s see how well our model performs on the test data. The main idea behind testing and evaluation is to give the model a completely new set of data it hasn’t seen before, and find out how well our model can predict the right outcomes.

To predict the corresponding yearly amounts spent for each observation in x\_test, we can simply call the .predict () method on the model. We will store these predictions as a separate data frame called predict.



Let’s see how accurate our model’s predictions are. We can build a **scatterplot** of the actual yearly amount spent (from y\_test) against the predicted yearly amount spent (from predict) using [***matplotlib***](https://matplotlib.org/index.html).

.

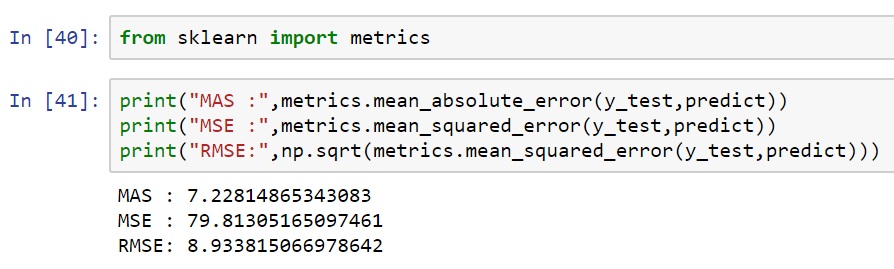


Evaluating our Model

There are different ways to measure the error between the predicted y-values and the actual y-values. We will calculate three kinds of errors using [*NumPy*](http://www.numpy.org/) and *Sklearn’s*metrics:

1. **Mean Absolute Error**
2. **Mean Squared Error**
3. **Root Mean Squared Error**

****



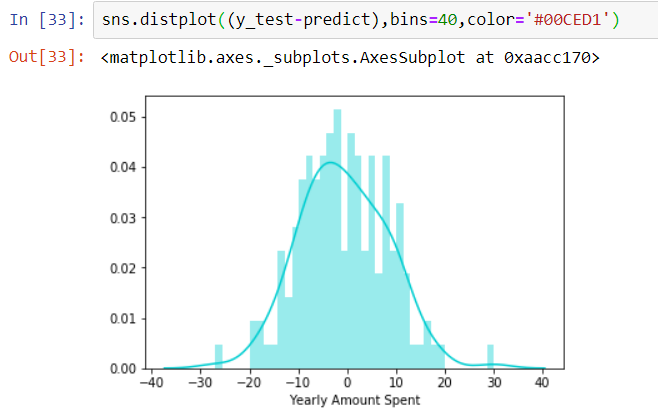
These errors seem fairly small, so we can conclude that our model is a pretty good fit.

**Residuals**

Although our model seems fairly good at making predictions, we need to make sure everything is okay with our data before the decision-making step.

To do this, let’s plot a histogram of the residuals and make sure it looks normally distributed, using Seaborn’s [**distplot**](http://seaborn.pydata.org/generated/seaborn.distplot.html?highlight=distplot#seaborn.distplot). The residuals are nothing but the **difference between the actual y-values and the predicted y-values**.

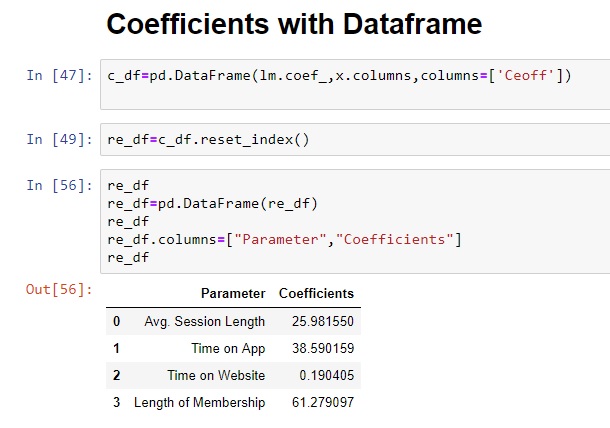
12.jpg



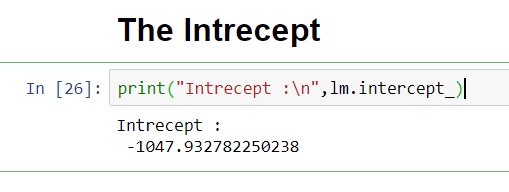
**Making the Decision**

Finally, we have to use our model to answer our original question: **Should the company focus more on their mobile app or on their website?**

Let’s recreate the coefficients as a data frame and see which feature (time on app or time on website) more influence on the yearly amount has spent.

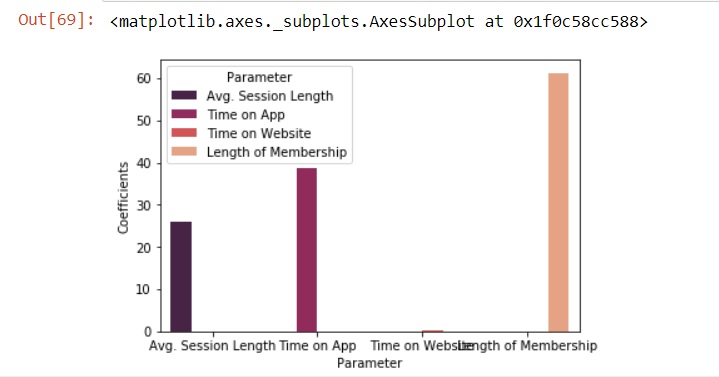


and the intercept is



Now if we plot a bar graph with the coefficient values then it will lool like below





From these coefficients, we can see that one minute on the app corresponds to **$38.59 in revenue**, whereas one minute on the website corresponds to just **$0.19 in revenue**. Therefore, it is pretty clear from our linear regression model that if the company wants to increase profits, they should focus their efforts more on their app.

**Conclusion**

After completing the project we want to conclude that any eCommerce company should use the Data science for analyzing the customer’s usage experiences and predict the best profitable options.