**Use Of Regression in E-Commerce**

E-commerce, or electronic commerce, refers to the buying and selling of goods and services over the internet. In today’s era, e-commerce has become an increasingly popular way for people to shop and do business. Some of the reasons for the bloom of e-commerce in present times are convenience, easy accessibility, a wide variety of products, and global reach. The COVID-19 pandemic has drastically changed the way people live, work, and shop, and this has had a significant impact on the e-commerce industry. Every day, more and more companies are going online, and they all want to increase their revenue. The revenue of ecommerce refers to the total amount of money earned by an ecommerce business through the sale of goods or services over a specific period of time. The revenue of an e-commerce business can be influenced by various factors, including the volume of sales, pricing strategy, time spent on a website etc. E-commerce has revolutionized the way we conduct business and it is expected to continue to grow and evolve in the future.

**A multiple linear regression model is been built.**

**Multiple Linear Regression -**

**Problem Statement:**

In this study, we try to predict whether the Yearly Amount Spent by the customer on the on the E-Commerce platform is dependent on various factors like the Average Session Length, the Time on App, the Time on Website and the Length of Membership.

**Dataset:**

* The dataset is taken from the open-source platform Kaggle.
* The dataset contains 500 observations for 8 variables out of which three are categorical variables and other five are numerical variables.
* There are four independent (input) variables and one dependent (output) variable.
* The independent variables are Average Session Length, Time on App, Time on Website & Length of Membership and the dependent variable is Yearly Amount Spent.
* The categorical variables Email, Address and Avatar are not included in the study.

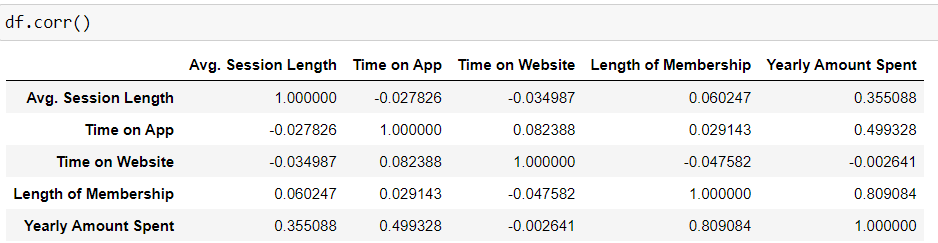
**Multiple Linear Regression Model:**

All the necessary libraries are imported in Python.Then, the dataset is loaded with the help of .read\_csv() command. The data.head() command is used to view the first five rows of the dataset.

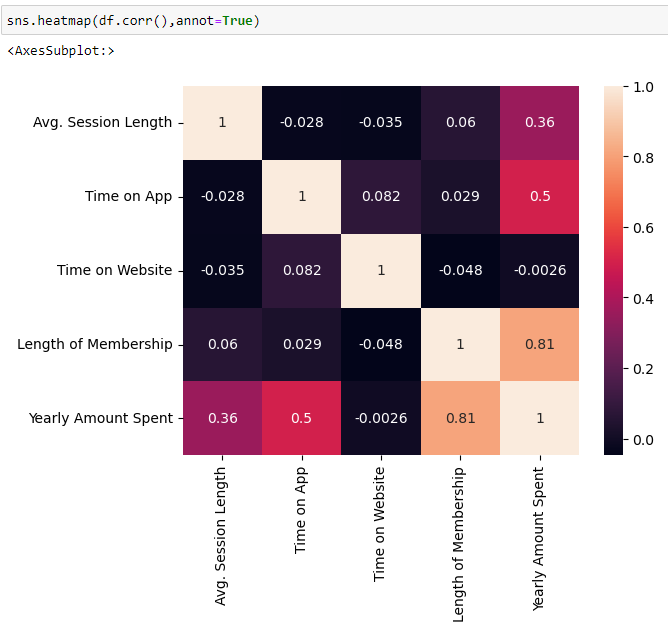
The presence of null values is checked using is.null() function and sum() function counts them.Our dataset contains no null values, so the count of the null values for all the variables is zero.

**Exploratory Data Analysis:**

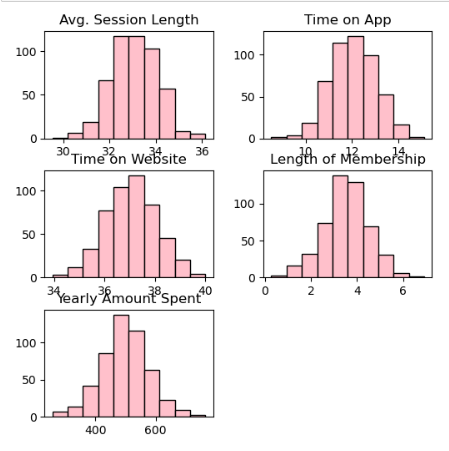
df.info() gives the overview of the dataset. df.describes() gives the statistical information of numerical column. It gives the count, mean, standard deviation, minimum & maximum values and the quartiles.

df.corr() gives the correlation matrix of the dataset.

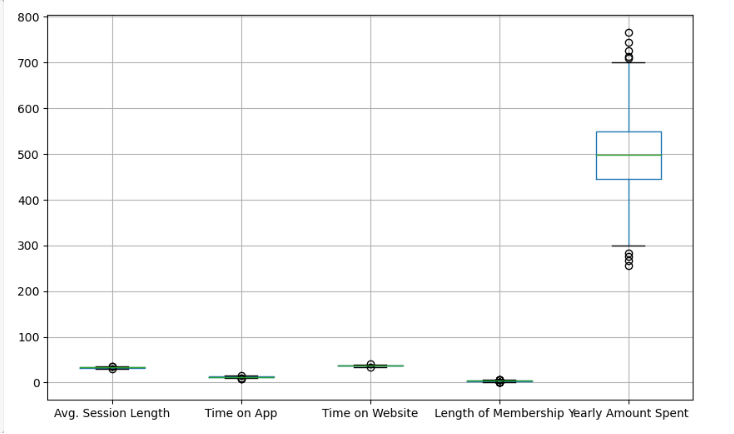
The correlation is also found with the help of a heatmap. There is no high correlation between the independent variables, so multicollinearity is not present in the data.



Next histogram is plotted. From histogram we can observe that all variables are normally distributed.



Boxplot is plotted to detect the presence of outliers. Outliers are present in the Yearly Amount Spent.



**Building a Multiple Linear Regression Model:**

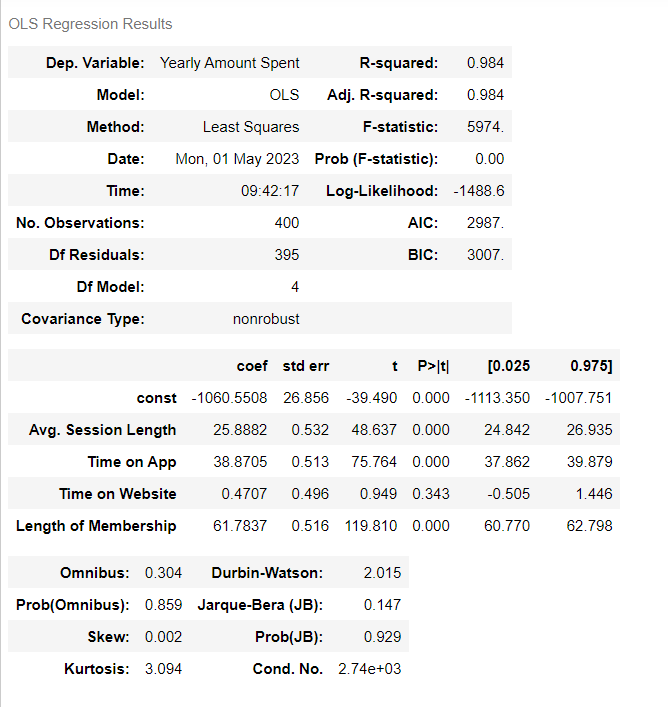
The independent variables x and dependent variable y are defined. Here, the independent variables are Average Session Length, Time on App, Time on Website & Length of Membership and the dependent variable is Yearly Amount Spent.

The dataset is split into train and test set in the ratio of 80:20. This means that 20% of our data points will be in test set and 80% of them will be in the train set. Random state is taken to be 0 indicating that we will get the same train and test sets across different executions.

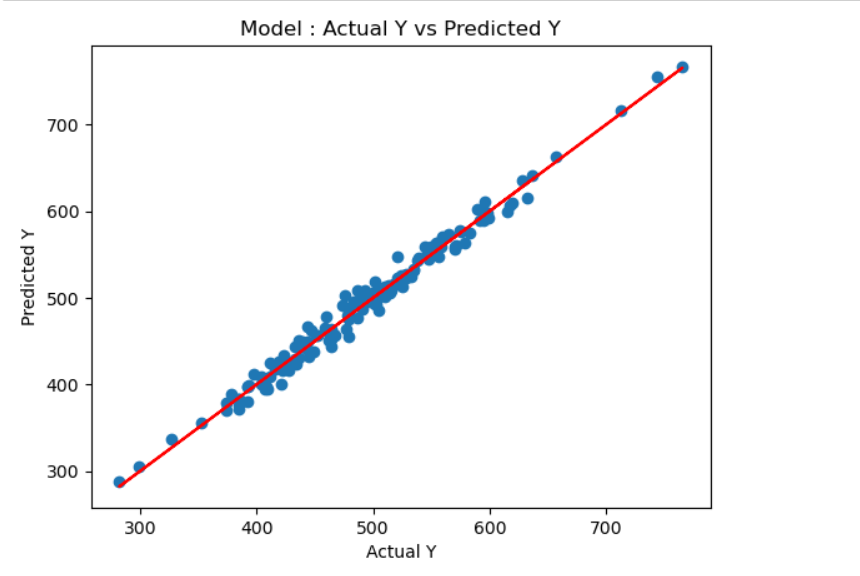
Then, the values of x\_train, x\_test, y\_train and y\_test are printed.

Then, the multiple linear regression model is built using OLS Regression.

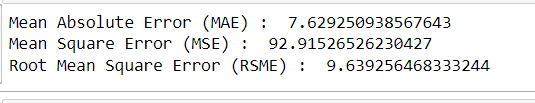
The output of the OLS Regression is:



After fitting the model, the scatter plot for Actual values vs Predicted values is plotted. The actual values and the predicted values have almost perfect linearity.

****

Finally, the errors like Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE) are checked using metrics package from scikit-learn library.



**Conclusion:**

From the results of the OLS Regression, the Adjusted R2 value is 0.984. This means that 98.4% of the variance can explained by the model. It is very high, so our model is a good fit.

From the results of the OLS Regression, the value of Durbin-Watson is 2.015. Since this is within the range of 1.5 and 2.5, autocorrelation is not problematic in the model.