**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.

**Logistic Regression –**

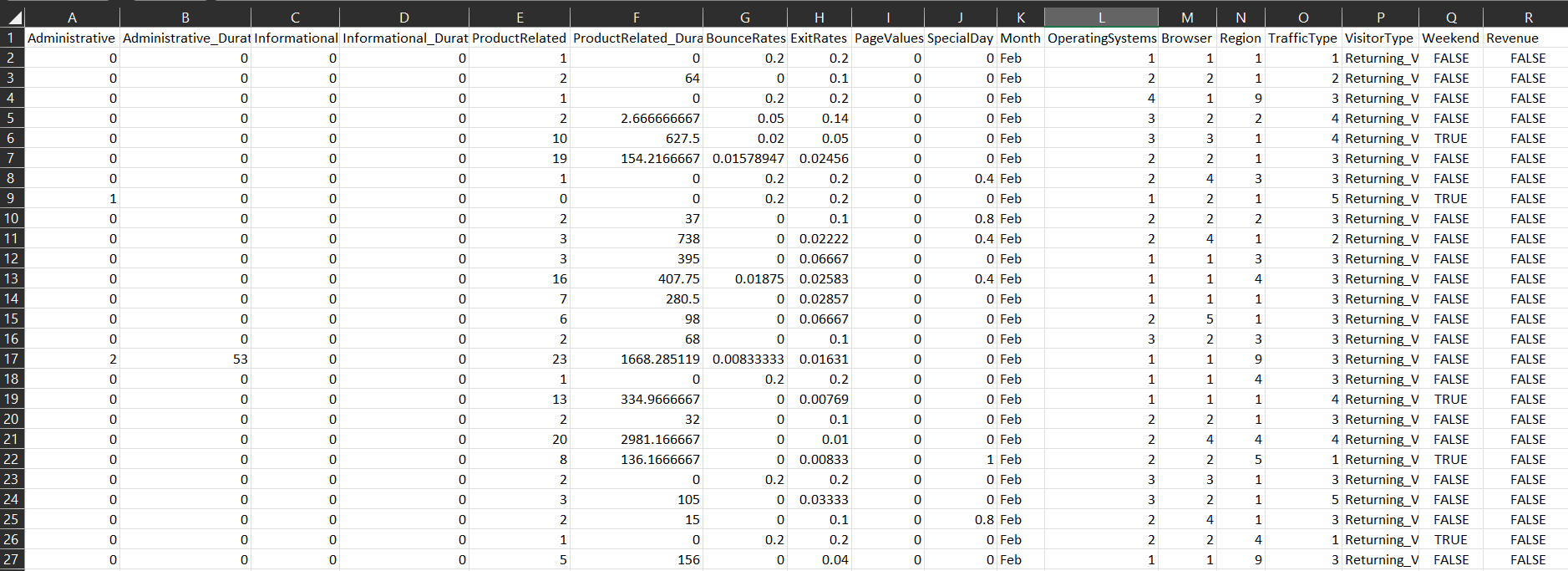
**Problem Statement:**

Due to the recent boom in e-commerce, there is fierce competition among shopping websites to boost their earnings. In this study we predict whether customer will buy product (spend money) on a shopping website or not.

**Dataset:**

* Data from UCI Machine Learning Repository is used for analysis. There are 12330 rows and 18 variables in the dataset.
* Revenue is the dependent variable and the rest other variables are independent variables.
* Administrative, Informational and Product Related represent the number of different types of pages visited by the visitor in that session while Administrative Duration, Product Related Duration and Informational Duration represent the time spent by user in each of these page categories.
* Bounce Rate represents the percentage of visitors who arrive on a website and leave before going to second page while Exit Rate is the percentage of visitors who leave website after a particular page.
* The Page Value represents the average value for a page that a user has visited before landing on goal page or completing an e-commerce transaction or both.
* The Bounce Rate, Exit Rate and Page Value represent the metrics measured by Google Analytics for each page in e-commerce website.
* The special day represents a few days before and after any special day like Mother’s Day, Valentine’s Day etc. as usually sales go up during these special days.
* The visitor type variable tells us whether the customer is a returning customer or a new customer. The Traffic Type variable means how did the user land on a particular website. There are various traffic types like direct, Organic, Referral, Social, Email, Paid etc.
* Other variables include the month, region type, browser, weekend and revenue. Revenue is a binary variable.

Glimpse of the dataset:



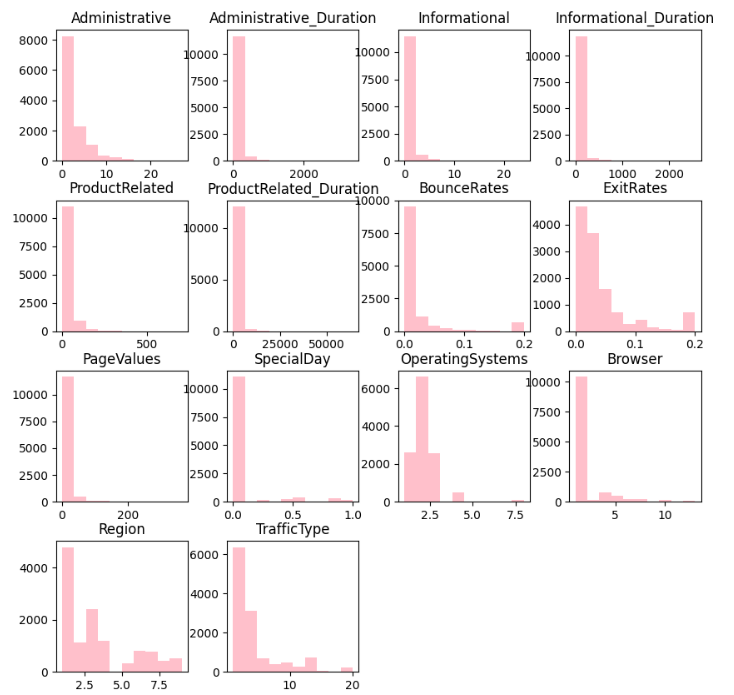
**Logistic Regression Model:**

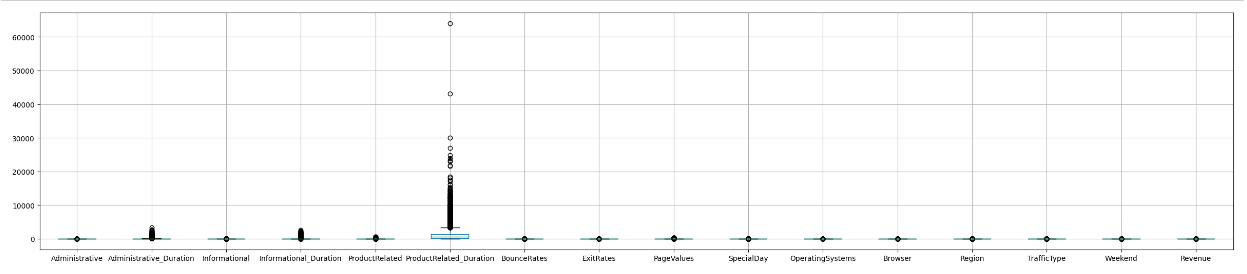
All the necessary libraries are imported in Python. Then, the dataset is loaded with the help of .read\_csv() command. The data.head() command is used to view the first five rows of the dataset.

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.

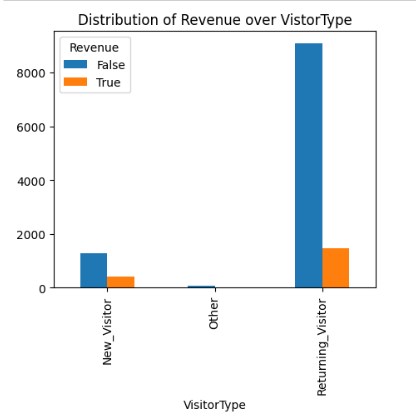
Next histogram is plotted. From histogram we can observe that all variables are positively skewed. 

Boxplot is plotted to detect the presence of outliers. Outliers are present in Administrative Duration, Product Related Duration and Informational Duration. 

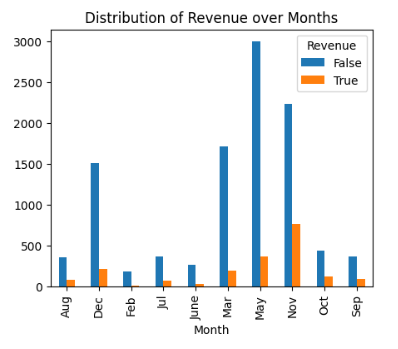
The counter function from collections library is used to count the number of specific types of values in Visitor Type, Revenue, Weekend and Region. Using it we find out that:

* There are 10551 Returning Visitors, 1694 New Visitors and 85 Other Visitors
* Out of 12330 people visiting the website 10422 people do not buy anything (spend money) while only 1908 spend money
* 9462 people visit the website on weekends while 2868 people visit during weekdays.
* 4780 people are from region 1, 1136 people are from region 2, 2403 people are from region 3, 1182 people are from region4, 318 people are from region 5, 805 people are from region 6, 761 people are from region 7 while 434 people are from region 8 and 511 people are from region 9.

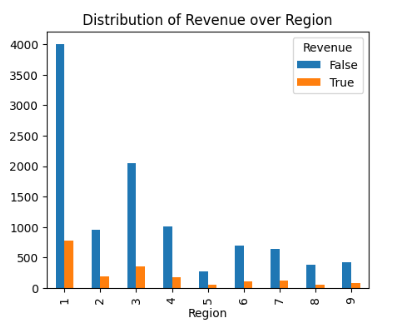
Bar plot and line graph are plotted to see the distribution of revenue over Visitor Type, Month, Region, and Traffic Type.



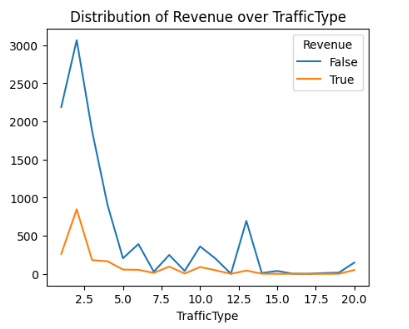
The proportion of new customers visiting the website and buying products is higher than the proportion of returning customers visiting the website and purchasing products (spending money). Other types of visitors hardly visit the website or make any purchase.



Most people visited the website in the month of May, November & December, whereas the most purchases that occurred happened in November.

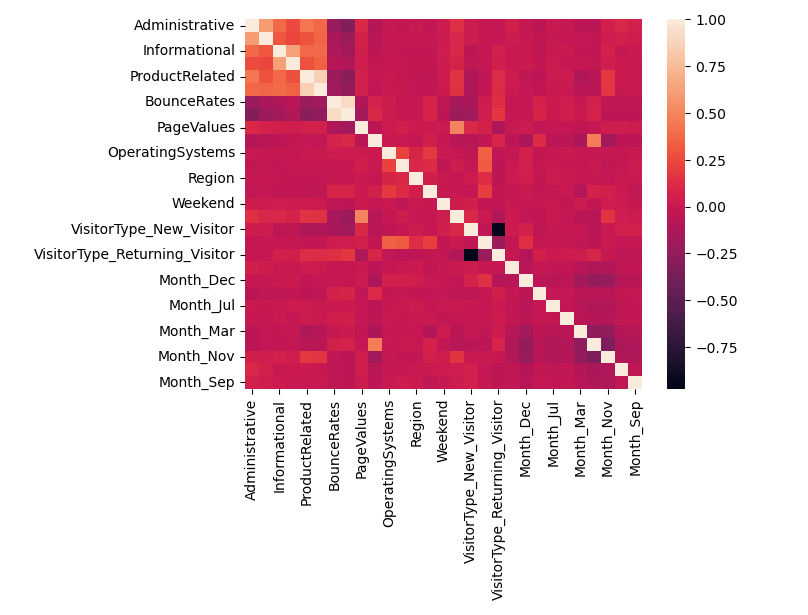


The website is most visited by people from Region 1. Most of the customers are from region 1 to region 5. They are more prone to making purchases.



We can see that different Traffic Type browsed the web page and out of which few of them made the purchase from the website.

Then a heatmap is plotted in order to understand the correlation between all variables.



The next step is to assign independent variable X and dependent variable Y their respective values. Revenue is the dependent variable and all other variables are independent variables. There are categorical variables in our dataset. We use dummy variables for transforming Month and Visitor Type. Label encoding is used on dependent variable Y and it is fitted.

**Building a Logistic Regression Model:**

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. Feature scaling is applied in order to reduce the dominance of some features. Since our data is not normal, standardisation is used. The goal of standardization is to have all values in the same range.

Then logistic regression model is fitted.

**Conclusion:**

The accuracy score of our logistic regression model is 87.10% indicating that the model makes correct predictions 87% of the times. Hence our model is a good fit for the data.

From confusion matrix we can observe that there are 1992 True Positives (TP), 52 False Positives (FP), 266 False Negatives (FN) and 156 True Negatives (TN) values. This means 1992 times we predict that user spend money on our website and 1992 they actually spent money on the website. 52 times we predicted that the user would not spend money but they did spend money. 266 times we predicted that the user will spend money but they did not spend any money on the website and 156 times we predicted that the user will not spend money on the website and in fact they did not spend any money.

Further other regression models like Decision Trees and Random Forest can also be fitted in order to compare which type of regression model gives the best accuracy. Classification techniques like KNN and Naïve Bayes can also be used.