**Abstract:**

Customer segmentation is the process of dividing customers into groups based on common characteristics so that companies can target each group efficiently. With the increase in businesses coming up every day, it has become significantly important for the old businesses to apply marketing strategies to stay in the market as the competition is increasing. This project is an attempt to classify the customers into different categories using unsupervised learning and train supervised learning classifiers using this data and decide the best classifier based on the values of accuracy. In this project many machine learning libraries present in Python are used to perform different operations.Logistic Regression ,Random Forest Classifier, XG Boost Classifier are the unsupervised learning algorithms used and an accuracy of 0.91, 0.98 and 0.97 was achieved respectively during testing.

1. **Problem Statement:**

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers..

**2. Dataset Description:**

* InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this

code starts with letter 'c', it indicates a cancellation.

* StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
* Description: Product (item) name. Nominal.
* Quantity: The quantities of each product (item) per transaction. Numeric.
* InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.
* UnitPrice: Unit price. Numeric, Product price per unit in sterling.
* CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
* Country: Country name. Nominal, the name of the country where each customer resides.

**3. Missing Value Treatment:**

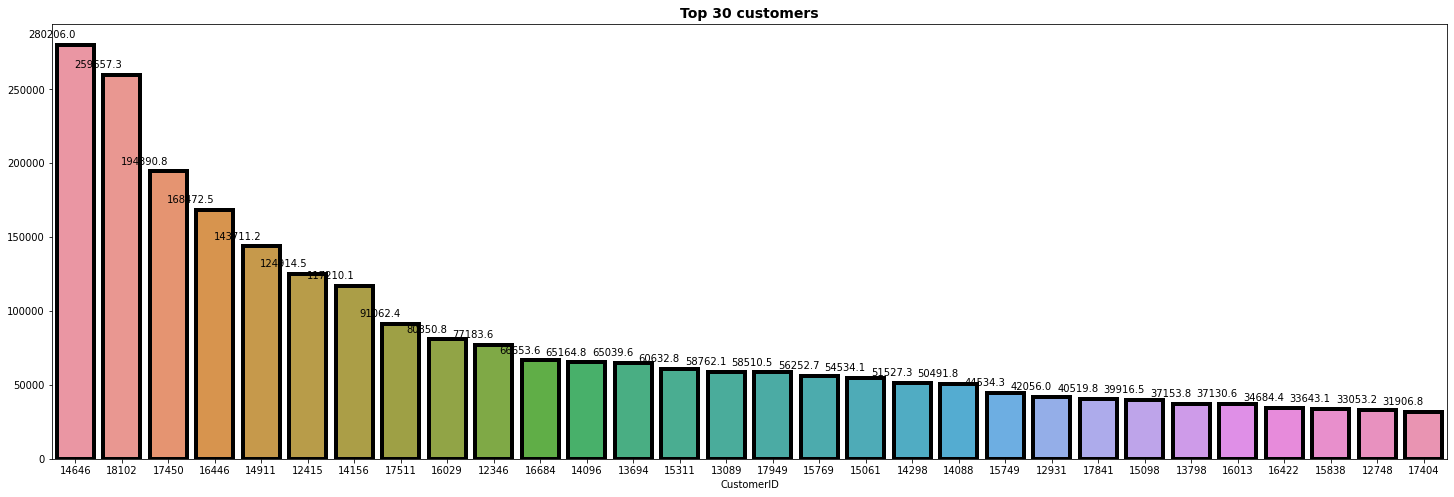
Missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Continuing to that we found missing observations in seven columns which we further treated with the median value that corresponds to that column.

**4.Exploratory Data Analysis (EDA):**

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

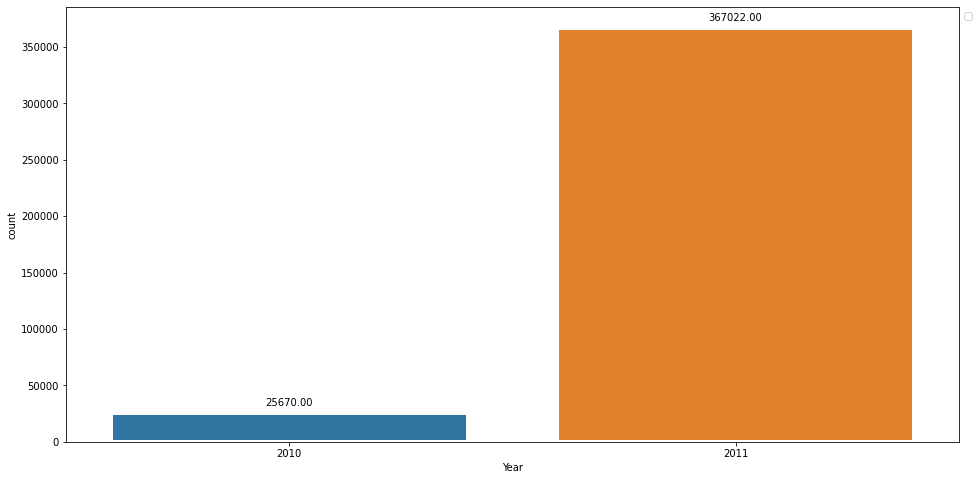
This section consists of details regarding the visual results:

**4.1 Top 30 Customers of our Retail Chain:**

****

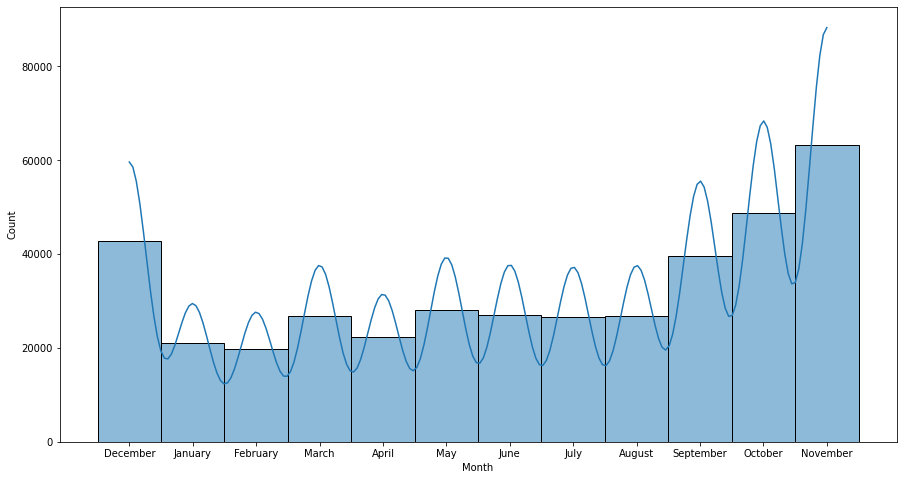
**4.2 Purchase Per Year:**

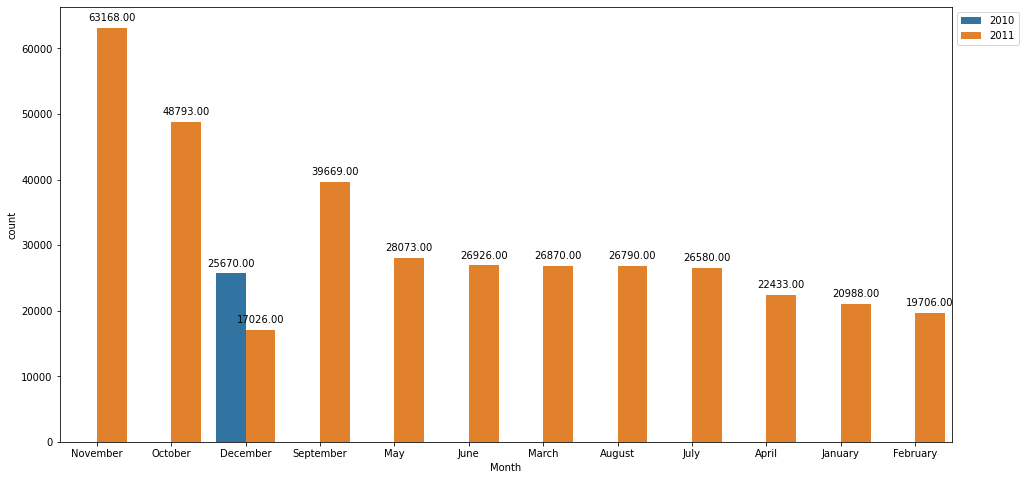




As observed , we have data of only december month of year 2010 thats why we can see that there is huge spike in purchase in year 2011.

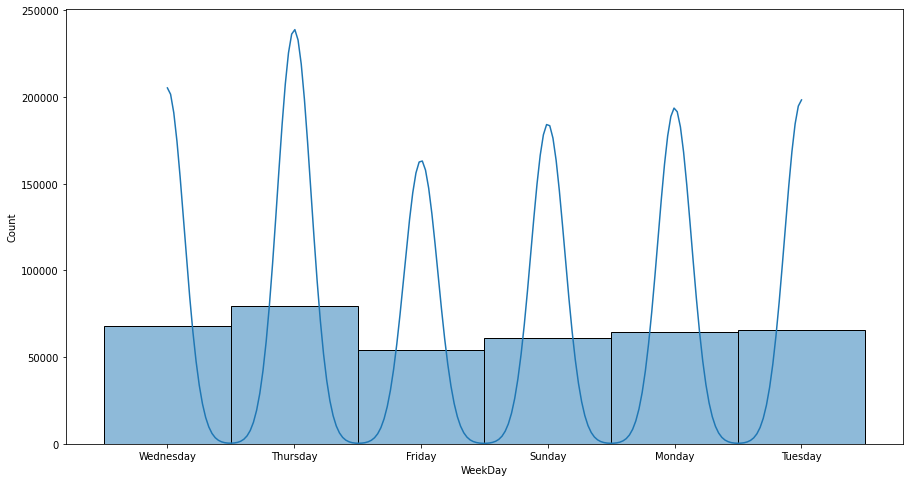
**4.3 Purchase Per Month:**

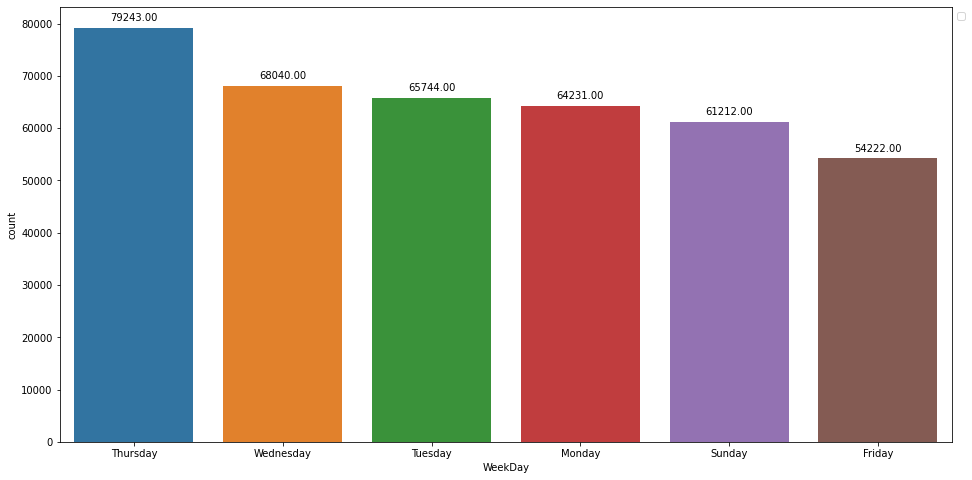


****

We can see that November and October month have highest number of purchase.

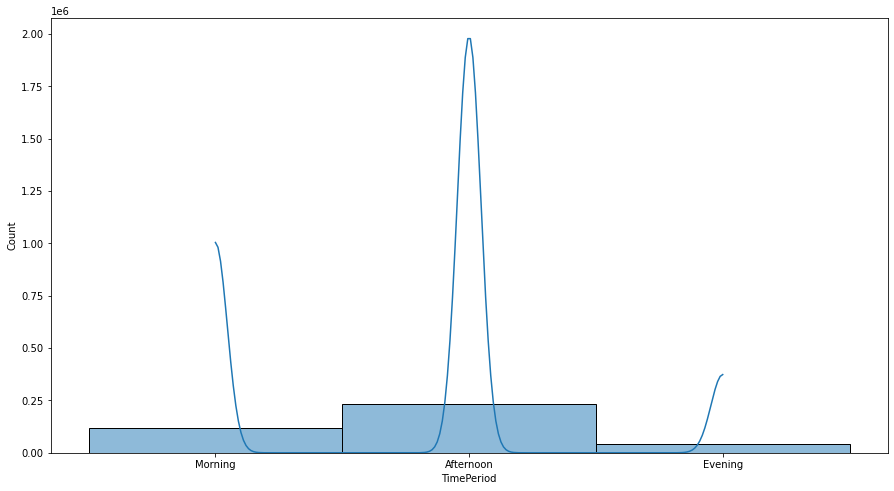
**4.4 Purchase Per Weekday**

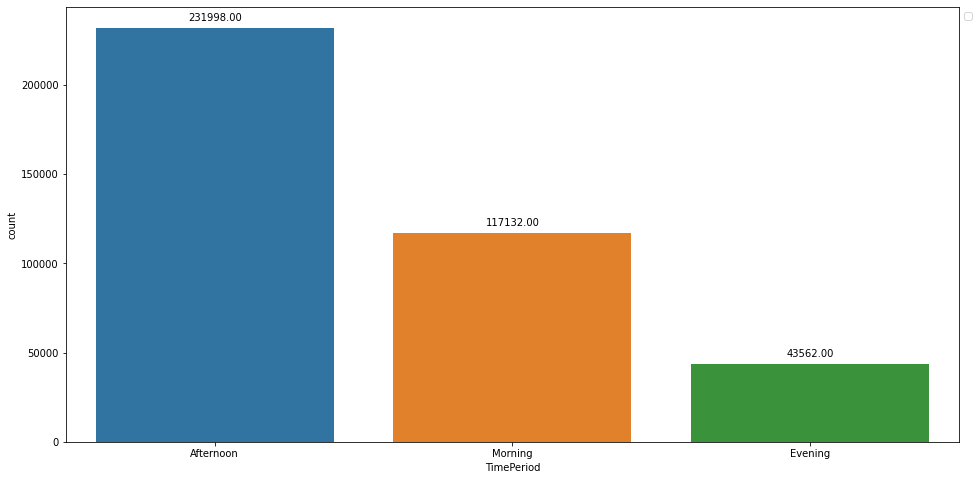
****

****

We can see that Thursday and Wednesday have the highest number of purchase.

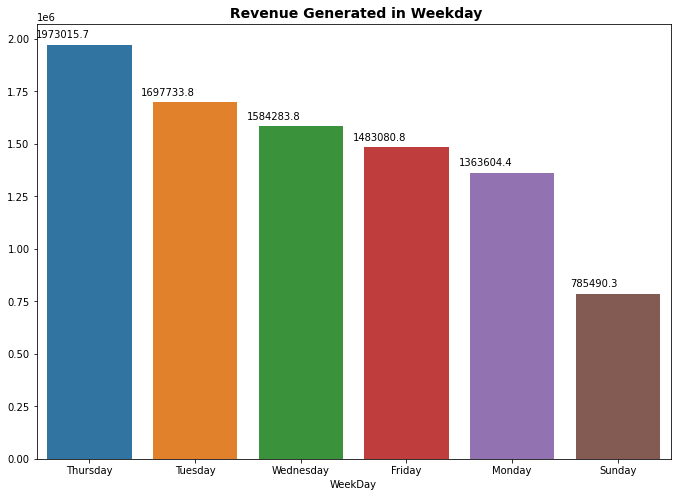
**4.5 Purchase Per Time Period:**

****



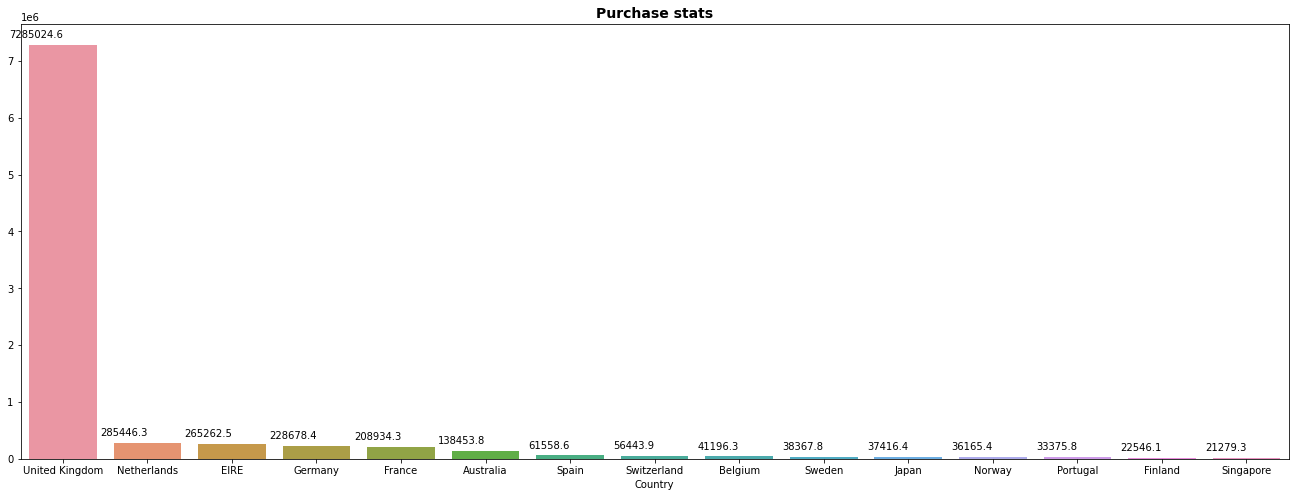
We can see that the most of people like to go for purchasing in afternoon time period.

**4.6 Revenue Generated in Weekdays:**

****

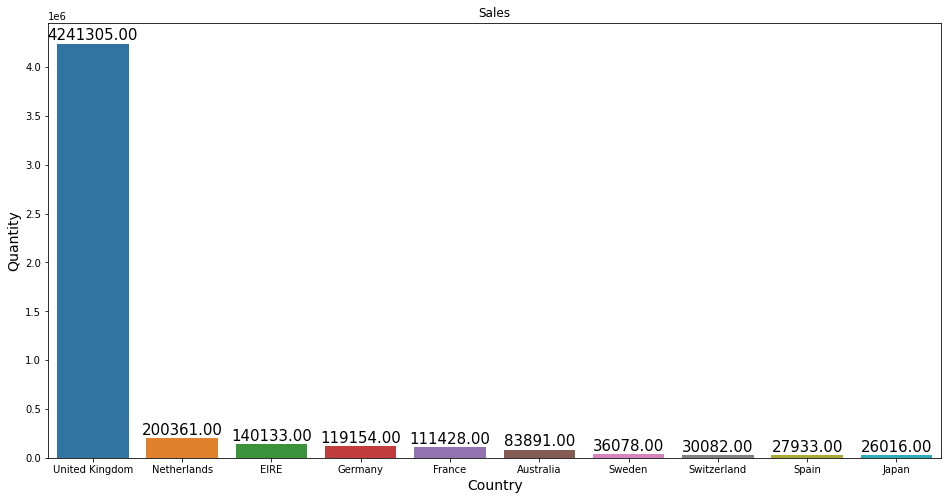
The highest Revenue was generated on Thursday and the Lowest Revenue was generated on Sunday.

**4.7 Purchase Satistics:**

****

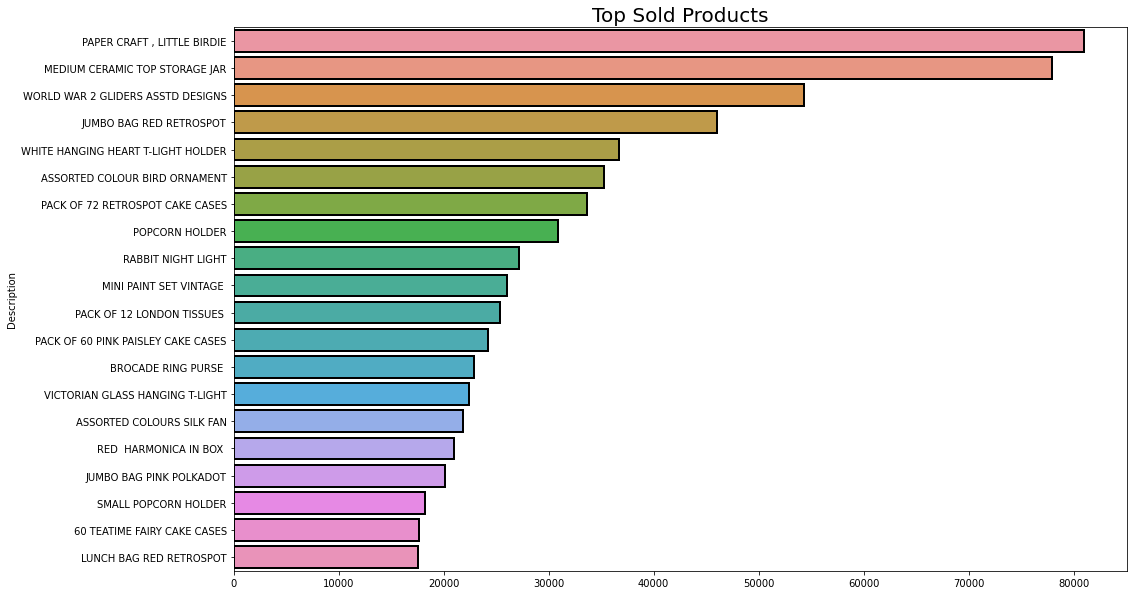
It can be seen that the country which have purchased more number of items as compared to other countries is United Kingdom and the country which have purchased least items is Singapore.

**4.8 Top 10 Purchasing Countries:**



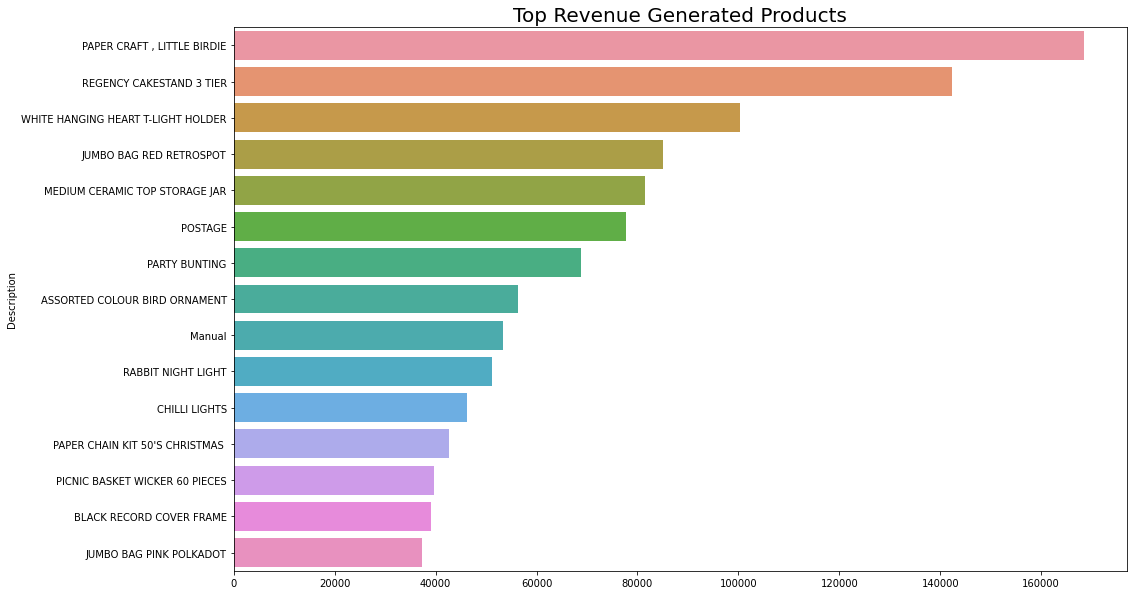
We can see that United Kingdom is at top in the list of top 10 purchasing countries and Japan is at the bottom in the list.

**4.9 Top Sold Products :**

****

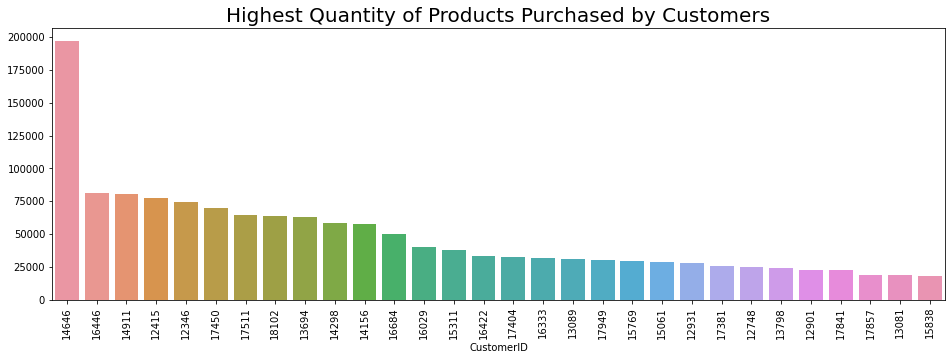
It can be observed that most sold product is "Paper craft, little birdie" and least sold product is "Lunch Bag Red Retrospot".

**4.10 Top Revenue Generated Products:**

****

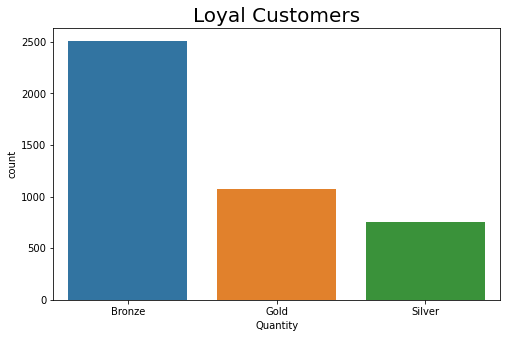
It can be seen that most revenue generated product is "Paper craft, little birdie" and least revenue generated product is "Jumbo Bag Pink Polkadot".

**4.11 Customer Statistics:**

****

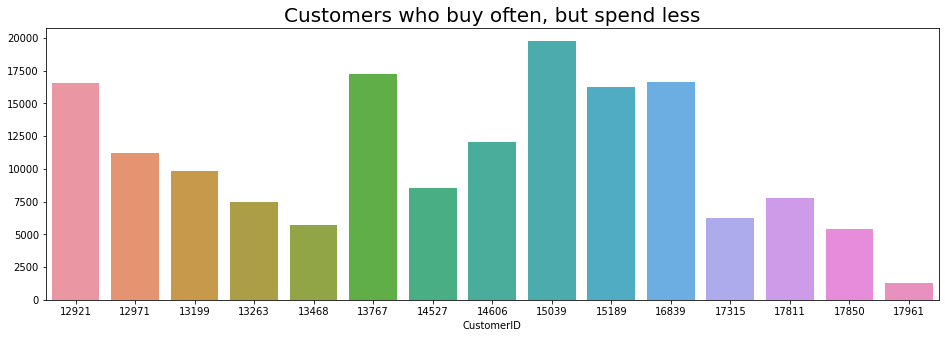
The bar plot shows the Customer ID of people who have purchased the highest quantity of products.

Since we don't have the purchase stats of our customers with the other companies, Let us segregate them into 3 Categories based on their purchases as Bronze, Silver and Gold.

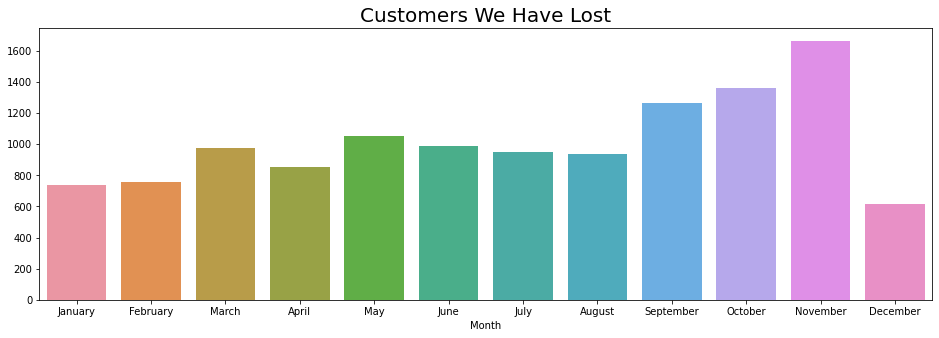


We can see that we have more number of bronze customers.

**4.12 Customers who buy often but spend less:**

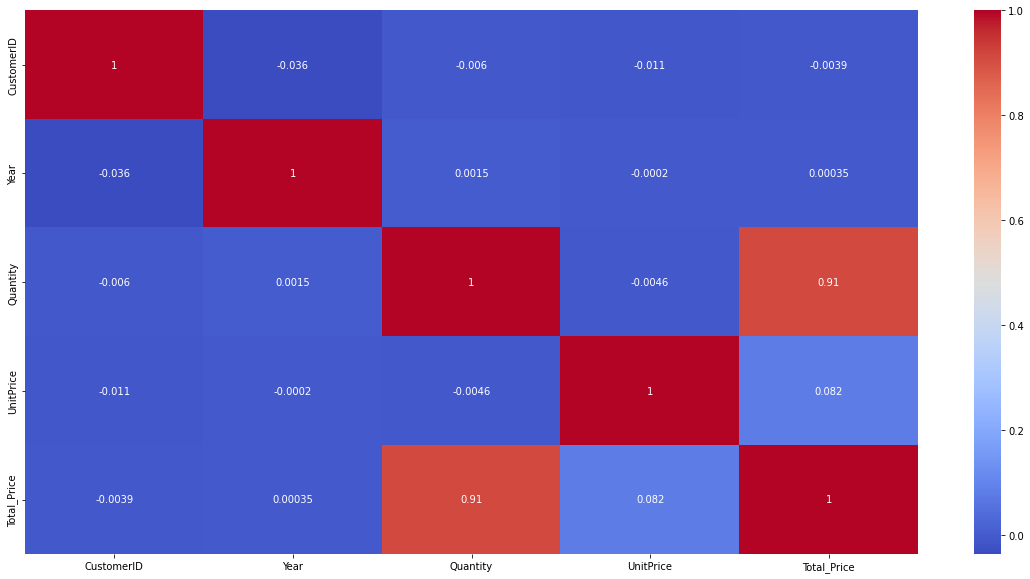
****

**4.13 Customers we have lost:**



It can be observed from above that we have lost the maximum customers in the month of November or we can also say that we have lost maximum number of customers in the end of the year.

**4.14 Correlation Analysis**

****

**5. Model Building:**

In this project, we used machine learning algorithms to get better result and to know which machine algorithm will be the best fit for this project.Machine Learning we used in this project are as follows:

**5.1 Logistic Regression:**

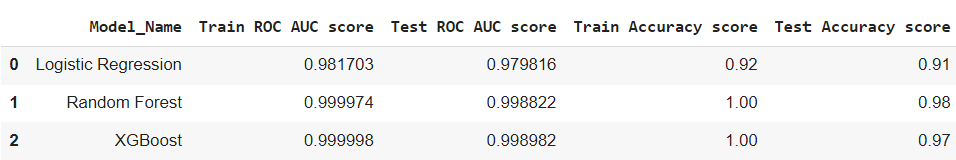
Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or interval type.

**5.2 Random Forest Classifier:**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

**5.3 XGBoost Classifier:**

[XGBoost](https://xgboost.ai/), which stands for Extreme Gradient Boosting, is a scalable, distributed [gradient-boosted](https://en.wikipedia.org/wiki/Gradient_boosting) decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

****

Random Forest have performed really well and got the best scores with Random Forest as compared to other Models, so I conclude Random Forest is my optimal model for use and we can use this model for further in online reatail customer segmentaion.

**6. Technologies used:**

**Python:** Python is a high-level interpreted language that supports different platforms like Windows, Linux, Mac, Raspberry Pi, etc. Python can be used for creating web applications, database systems, handling big data, and performing complex mathematical calculations. Python can be treated in an object-oriented, functional or procedural way.

**Google Colab:** Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

**Python packages:** Following are some of the python packages used in this project.

**Matplotlib:** Matplotlib is a visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualiz- ation library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram, etc.

**Pandas:** Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

**NumPy:** It provides structures for multiple dimensional array objects and tools for related operations. NumPy is usually used for high performance scientific computational tasks.

**Seaborn:** Seaborn is a visualization library for statistical graphics plotting in Python. It provides default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for the same variables for better understanding of the dataset**.**

**6. Conclusion:**

* Sales are very high in October, November & December as compared to other months.
* We have sales data of only december month from year 2010.
* The Retail Store is Closed on Saturday as per the observation.
* Maximum number of sales are happening on Tuesday, Wednesday and Thursday in ascending order respectively.
* The highest Revenue was generated on Thursday and the Lowest Revenue was generated on Sunday.
* It can be seen that the country which have purchased more number of items as compared to other countries is United Kingdom and the country which have purchased least items is Singapore.
* We can see that United Kingdom is at top in the list of top 10 purchasing countries and Japan is at the bottom in the list.
* It can be observed that most sold product is "Paper craft, little birdie" and least sold product is "Lunch Bag Red Retrospot".
* It can be seen that most revenue generated product is "Paper craft, little birdie" and least revenue generated product is "Jumbo Bag Pink Polkadot".
* We can see that we have more number of bronze customers.
* It can be observed from above that we have lost the maximum customers in the month of November or we can also say that we have lost maximum number of customers in the end of the year.
* The count of unique stock unit ids and their descriptions should have matched but they do not. This implies some stock units might have more than one descriptions.
* We noticed in the exploratory data analysis phase that majority of the transactions belonged to UK, so it makes sense to consider only this country data for maximum impact
* Most of the people are visiting the store in the afternoon as compared to morning and evening time period.
* The total number of sales in December of year 2010 is higher than December of year 2011.
* Management needs to concentrate on the decrease of sales in December month.
* With logistic regression we got the accuracy score of 0.91 on train data and 0.92 on test data.
* With Random forest classifier we got the train accuracy of 1.00 and test accuracy of 0.98.
* With XGBoost classifier we got the train accuracy score of 1.00 and test accuracy of 0.97

**References:**

1. GeeksforGeeks

2. Analytics Vidhya