***Section -2 EDA***

*About the dataset:*

The dataset consists of 2 excel sheets and 8 features in each sheet. I combined both the sheets to obtain a dataset of shape (1067371, 8)

*Exploring the features:*

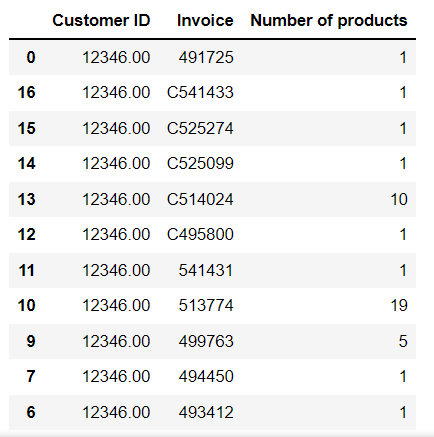
The features were categorised in this manner:

* Categorical Features (Nominal) : Invoice , Stock Code, Country, Description , Customer ID
* Numeric Features: Price, Quantity, InvoiceDate (date time)

The unique values for the underlying categories were as follows:

* Products: 5305
* Transactions: 53628
* Customers: 5942

After looking at the unique Invoice values:

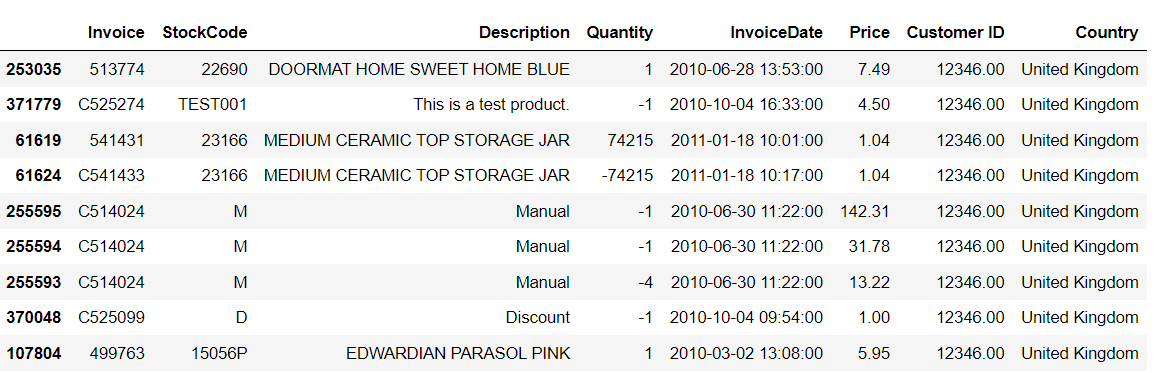


* I observed that there are some transactions with C in the beginning of the Invoice. Since the invoice was only a 6 digit number until now, but with the C it must mean the transaction was cancelled for some reason.
* Also we can see there are customers who purchased only one product and there were customers who purchased many products quite frequently.

*Orders cancellation analysis:*

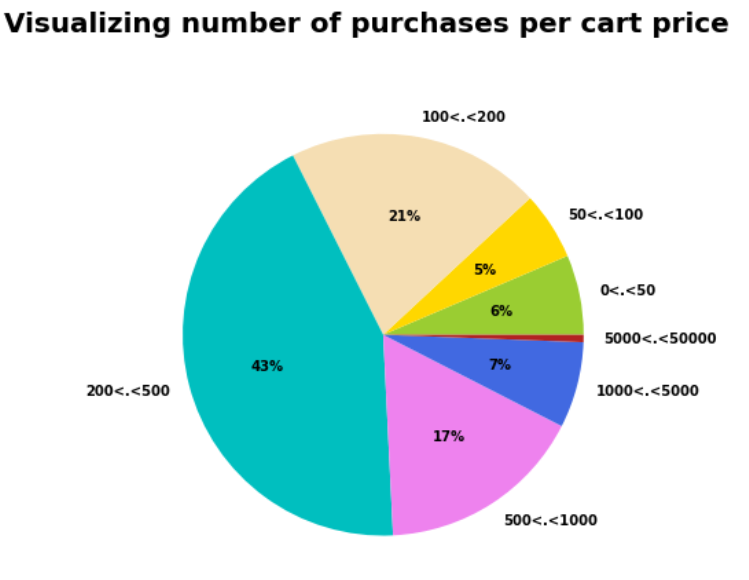
Number of orders cancelled: 7901/44876 (17.61%)

Cancelled orders had negative quantity , so we must remove them for further analysis of data



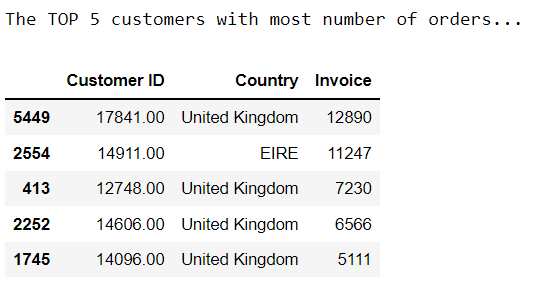
After obtaining a cleaned data frame which has no cancelled orders i.e negative quantity or products and price, I created a column Total\_Price which has the total amount of a transaction.

*Analysis on Total price denoted as cart price:*



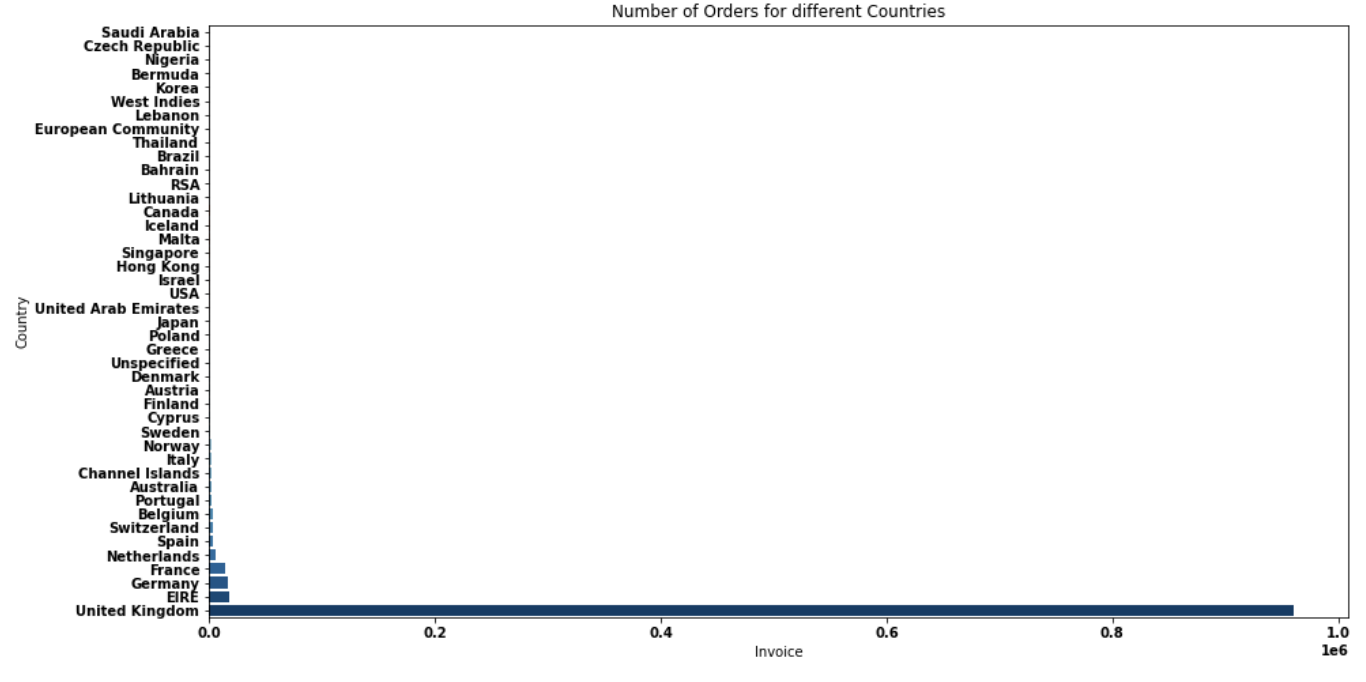
* We can clearly see that 43% of the total purchases have a total cart price in between 200 and 500 currency and around 60% of the total orders have a total cart price more than 200 currency.

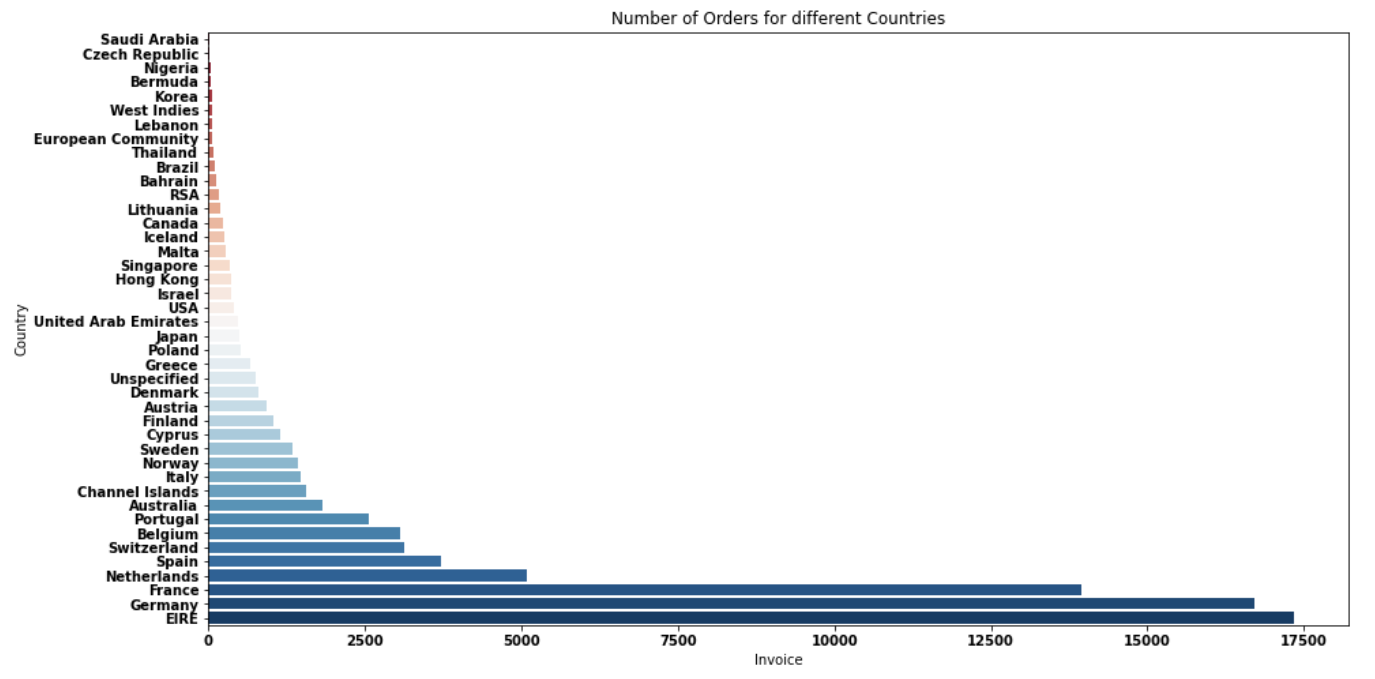
*Distribution of orders for different Customers*



*Exploring Patterns for Each country :*

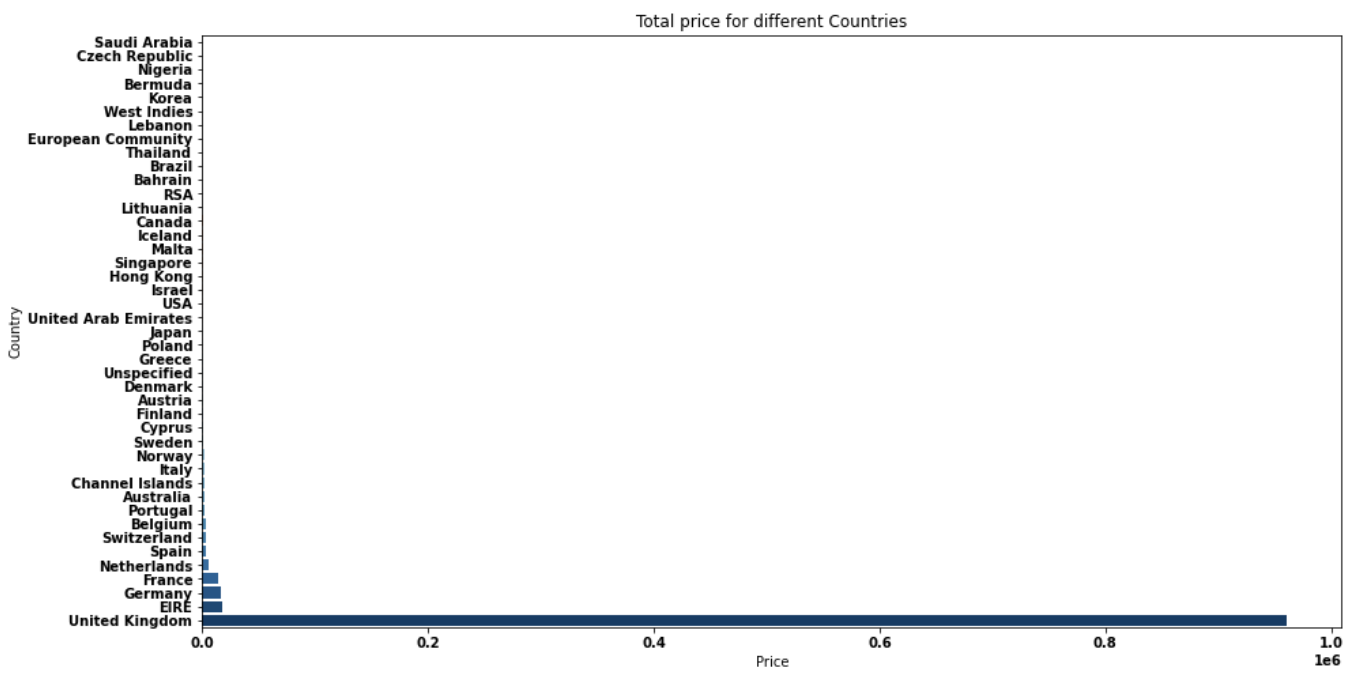
Customer distribution:



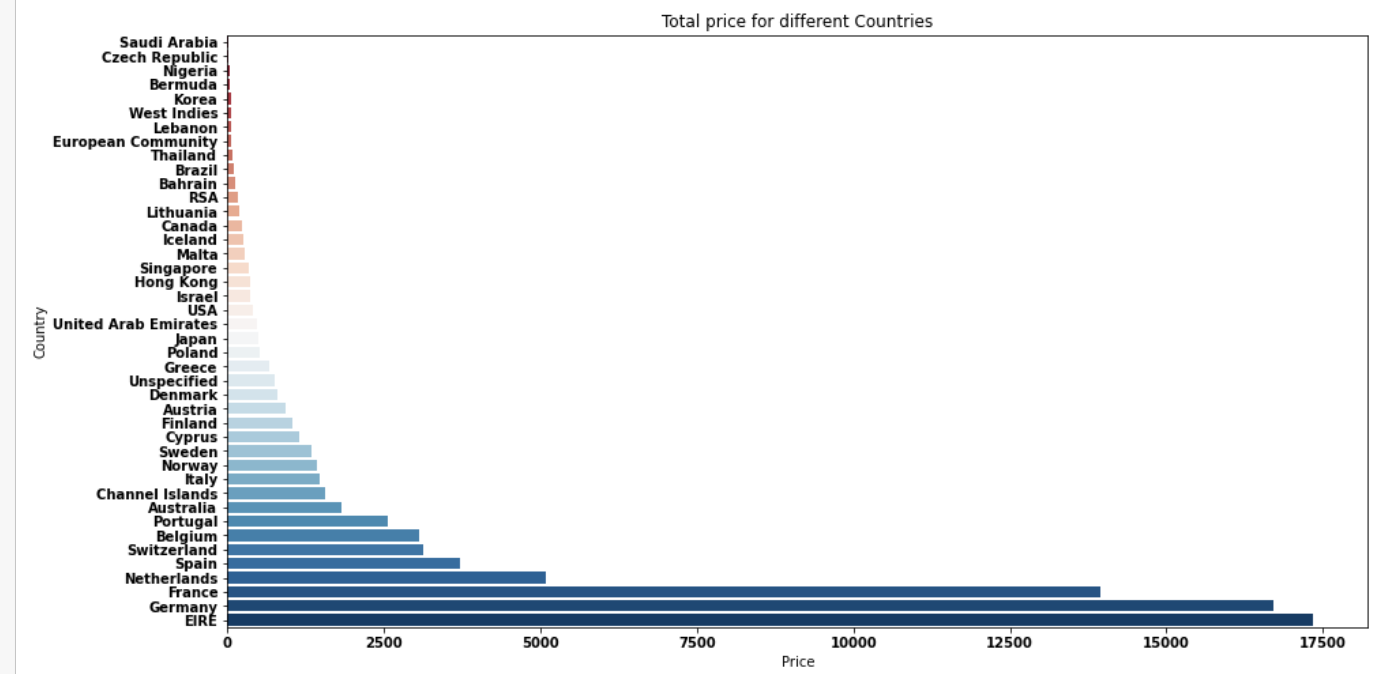


* The company receives high number of orders from UK as compared to other countries.
* As we can see the difference in overall distribution after excluding UK , this means that the company must be UK based.

*Price distribution*



Without UK



*Creating more features:*

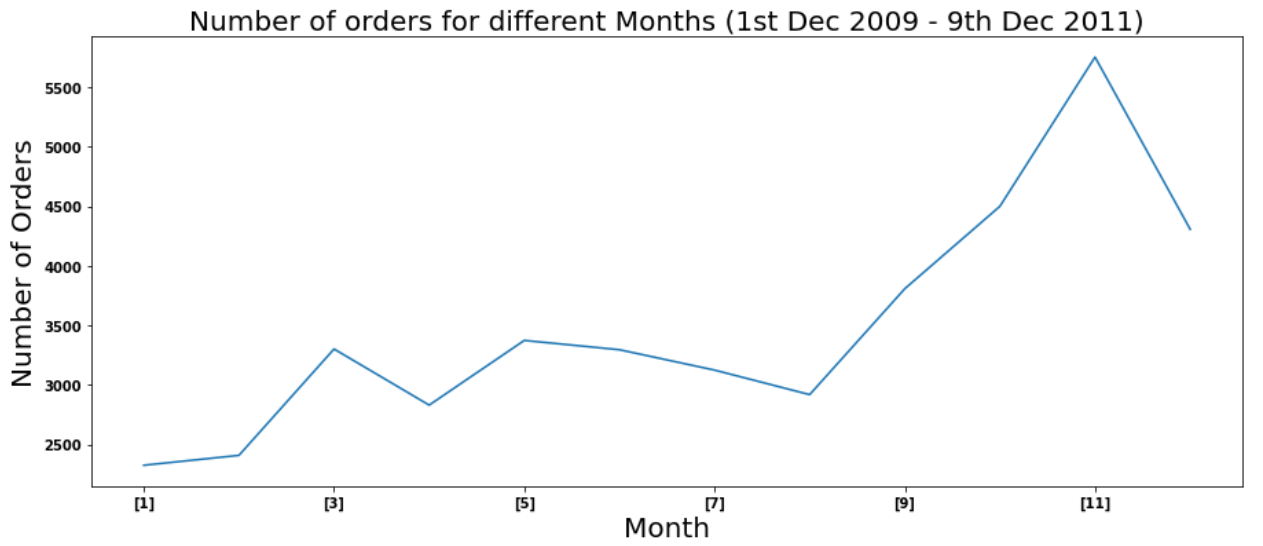
When ever there is a DateTime feature, Day , Month, Year, Year\_month, Weekday, Hour these can be extracted .

So that is what I did next.

*Analyzing the number of orders according to the Year Month:*

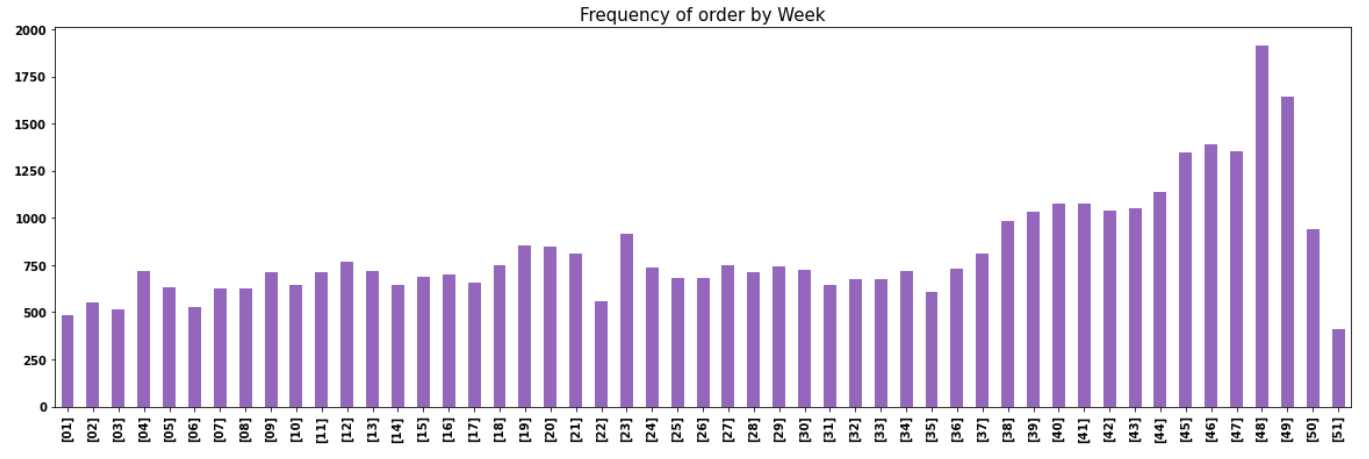


* 2011-November had the highest all time sales between December 2009 and December 2011



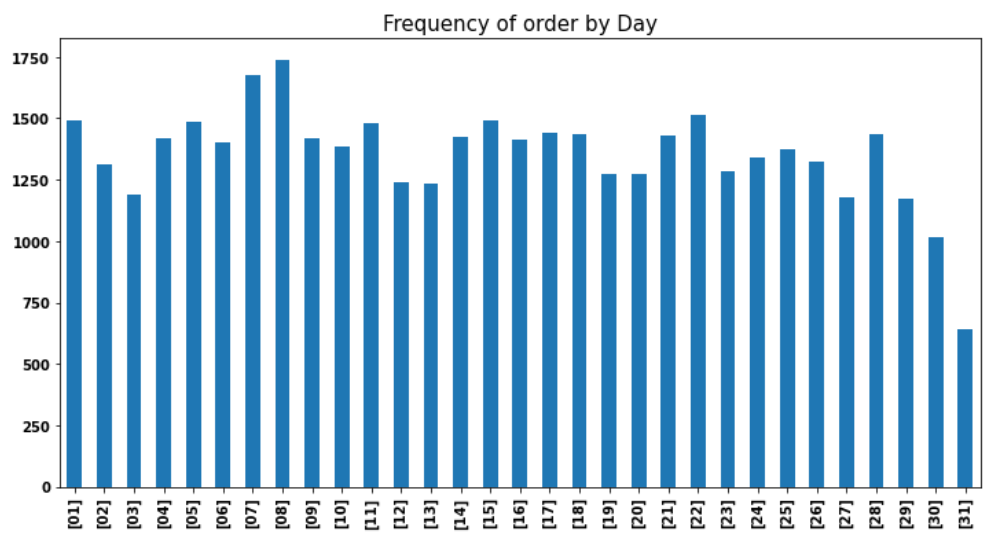
* There is seasonality in the orders as the sales increase from September till end of Early december. Hence this period is the busiest time for the company maybe because of End of summer Sale Period..

*Analyzing the number of orders according to the Week:*

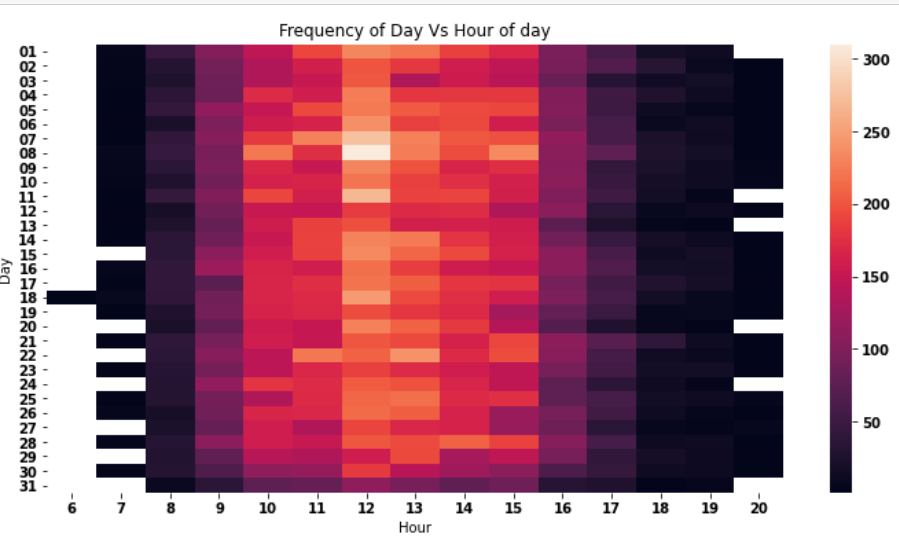


* 48th and 49th week of a year have the highest orders i.e. Early weeks of December.

*Analyzing the number of orders according to Day of the month:*

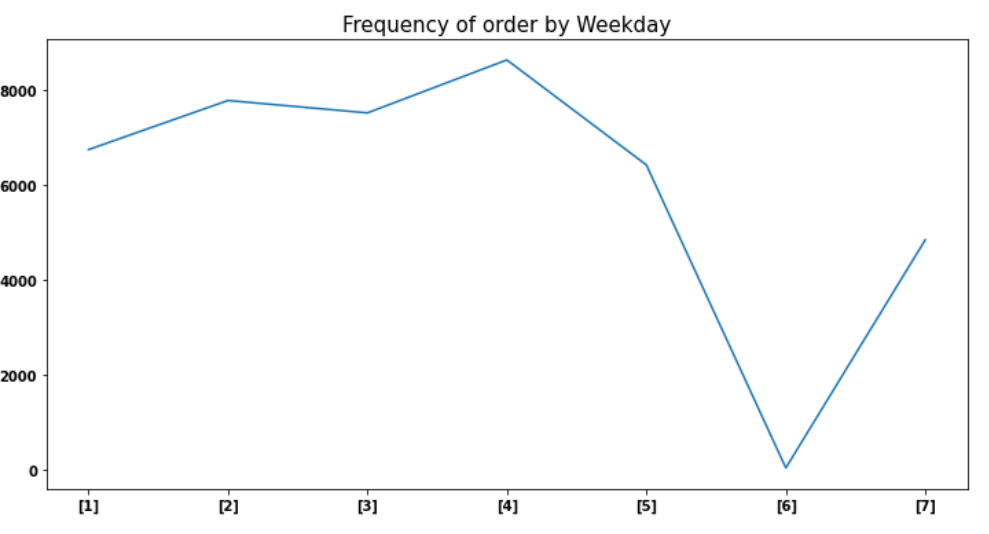


*Analyzing the Frequency of orders on a day vs Hour of the day*



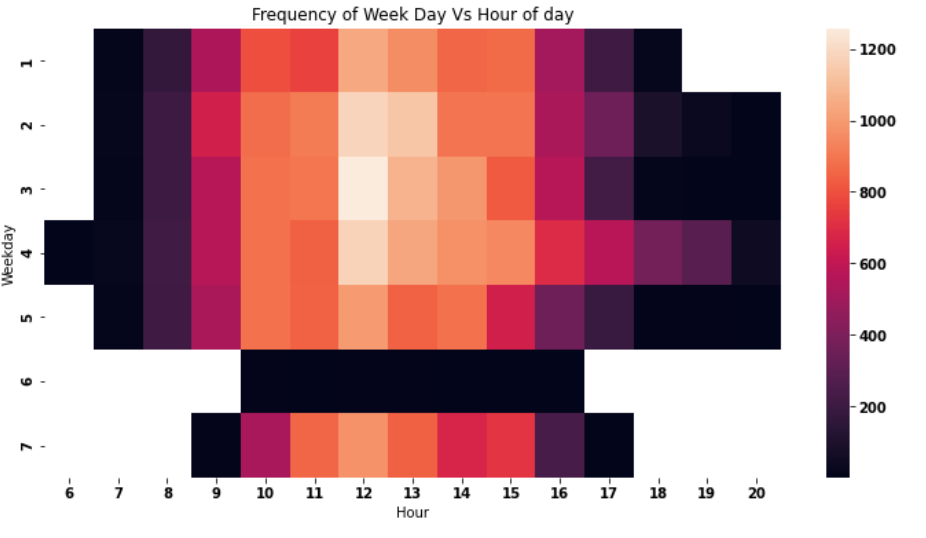
* This shows that people tend to order mostly around 11am to 2pm , this must be because of the lunch hours free time.

*Analyzing the number of orders according to Weekday*



* 6th day that is Saturday has almost no orders.

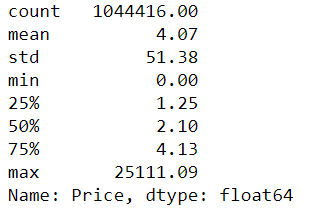
*Analyzing the Frequency of orders on a Weekday vs Hour of the day.*



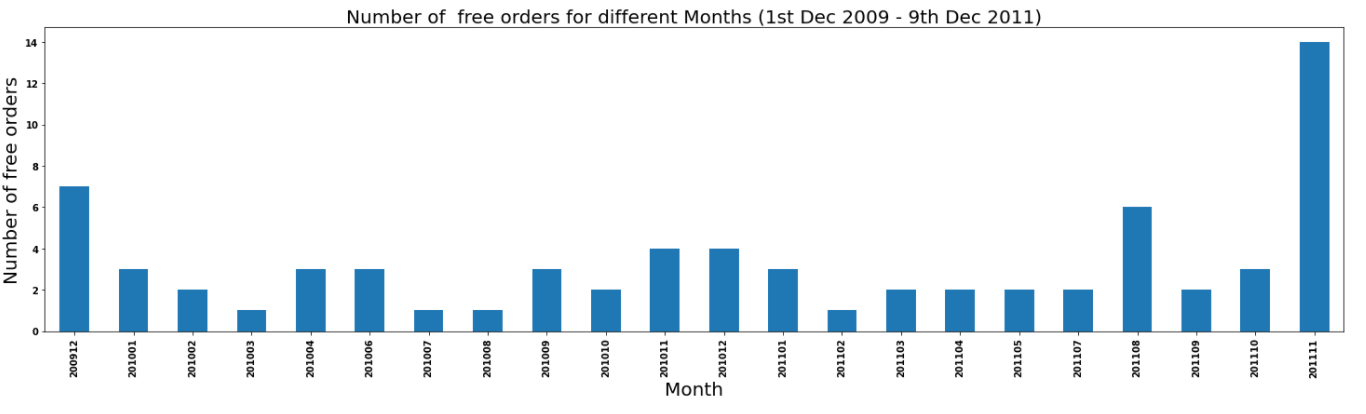
* Wednesday 12pm is the busiest time of the week in terms of customer orders.

*Product Analysis*

Since negative prices and quantities were removed the minimum of prices feature should be > 0 but:



* Minimum is 0 , this could mean some items must be sold for free,
* There were 2674 transactions which had 0 price.
* On further analysis , I saw that most of them had no Customer ID as they were TEST transactions and DOTCOM postages.
* So I removed them to target FREE transactions



* I didn’t understand why there are FREE items given to certain customers. It may be due to the sale season.
* The company gave away maximum no of free items on the month of November 2011.
* The company gave out 2-4 times FREE items to customers each month (Except in June 2011) on average.

***Conclusion of my EDA:***

* Number of orders cancelled: 7901/44876 (17.61%)
* 43% of the total purchases have a total price in between 200 and 500 currency and around 60% of the total orders have a total cart price more than 200 currency.
* The company receives high number of orders and highest Total\_price from UK as compared to other countries, this means that the company must be UK based.
* There is seasonality in the orders as the sales increase from September till end of Early December. Hence this period is the busiest time for the company maybe because of End of summer Sale Period.
* The number of orders received by the company tends to increase from Monday to Thursday and decrease afterward.
* The company receives the highest number of orders at **12:00pm**, this must be because of the lunch hours free time.
* The company gave away maximum no of free items on the month of November 2011, they also gave out 3-6 FREE items to customers each month (Except in June 2011) on average.

***Potential Analysis:***

1. This dataset also requires a deeper analysis of the cancelled orders.
2. The customers can be segmented according to the Description of the transactions. The keywords can be extracted from the description and the number of customers can be distributed among them. Clusters of keywords could be created and customers could be categorised according to the clusters.
3. Due to the high dimensionality of the data frame a
4. Principal Component Analysis would be required followed by the K means clustering technique to categorise the data.

***Ending Note:***

I would have utilised the Description feature and performed the above potential analysis but due to the time constraint and 2 coding questions I wouldn’t have completed the above EDA on time.

--------------------------------------------THE END---------------------------------------------