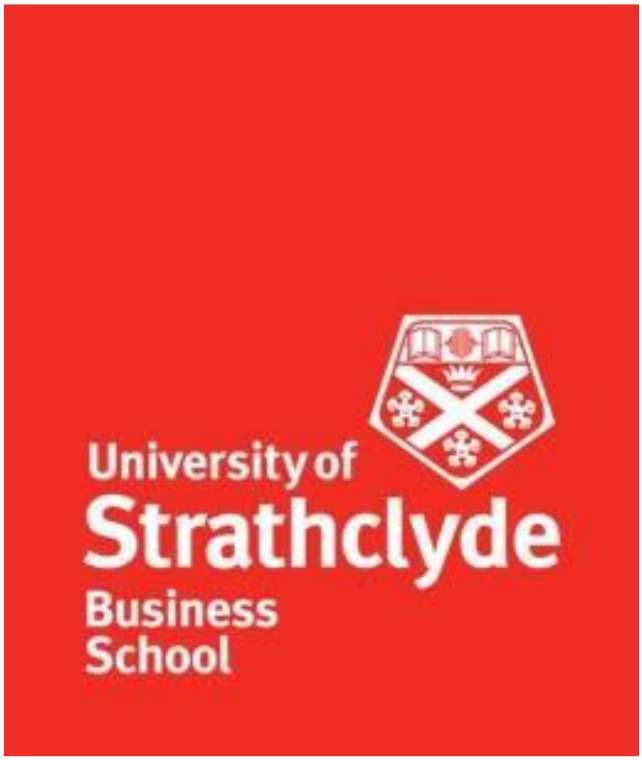


ANALYSIS OF ECOMMERCE DATASET USING PYTHON

CS989: Big Data Fundamentals Python Assignment. word count: 2895

Sourabh Mahajan
MSc. Data Analytics

Abstract

Today, in the competitive business world, data analytics plays a very important role. E-commerce website like Flipkart, Amazon, etc uses data analytical tools and techniques which help them to focus on market strategies, trend, customer pattern, etc for higher returns in their day to functioning. The results achieved by analysis of customer, transaction data, product and sales data strongly influence their profit growth and customer satisfaction. Nowadays in this growing digital market, the number of online users has increased, and the era has moved into a digital era. The number of vendors and retailers using the internet to commence with their business have seen an exponential growth. Having said this, to understand the impact of customer transactions a specific data set was chosen to dissect various statistics

Analysis of Ecommerce Dataset using Python

Sourabh Mahajan

November 2018

Index

1. Introduction to dataset

1.1 Getting Started

Kaggle and machine learning dataset website have tremendous collection of datasets, the very beginning task to choose one or either from them and start with the small experiments. 3 to 4 types of dataset were encountered and experimented with the technical stuff learned during lecture and lab session of python. Working with Brazil ecommerce dataset is challenging which was very huge and complex one. After doing small analysis, it was stuck and was not able to move forward with same and left. After same sort of stories with dataset like bakery transaction data, finally came up with UK ecommerce website dataset.

1.2 Introduction to Ecommerce dataset

E-commerce data actual transaction from UK retailer is available at both websites at UCI Machine Learning Repository with title **"Online Retail"** and on Kaggle as E-commerce data *(kaggle, 2018).* This data typically contains transactional data 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. Dataset has 8 columns and 541909 rows. The table below is the output of reading the csv file and the head command applied on data.csv. Each row of dataset indicates a product of an order. An order can contains more than one product with different quantities. InvoiceNo and InvoiceDate indicates order number purchased date respectively and summation of all rows with same invoice number can be considered as one complete order.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/01/2010 08:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/01/2010 08:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/01/2010 08:26	2.75	17850	United Kingdom

Table1: Dataset Overview

	Quantity	UnitPrice	CustomerID
count	541909	541909	406829
mean	9.55	4.61	15287.69
std	218.08	96.76	1713.6
min	-80995	-11062.1	12346
25%	1	1.25	13953
50%	3	2.08	15152
75%	10	4.13	16791
max	80995	38970	18287

Table2: Dataset info

Coloumn	Null Value
CustomerID	135037
Description	1454
Country	0
UnitPrice	0
InvoiceDate	0
Quantity	0
StockCode	0
InvoiceNo	0

Table3: Null value Count

Describe command results show quantity and unitprice negative figure implies that some invoice/ orders have been cancelled. Data needs to be clean as we can see null values in two columns.

2. Identification and description of key challenges

It was always predicted that there are many different insights you can extract transactional data.

For the purpose this assignment following analysis challenges were concluded

1. Percentage of Cancelled order and its impact on total revenue?
2. Basket analysis
    i. Total product sold till date?
    ii. Total Revenue till date?
    iii. Average order value of dataset?
    iv. Average basket size of dataset?
3. Orders according to countries
4. Customer patterns
    i. Top 5 customer spender
    ii. Top 5 customer according to order
    iii. Clustering customer according to their purchase quantity
    iv. Prediction of future sale according to country and customer

It has been noted that if analysis is done on above questions, the ecommerce company can make some positive changes on management and marketing of business. Business can be made more growing and profitable.

3. Methodology

After having played with 4 datasets I came up with my current (UK ecommerce) dataset and I made step by step process to analysis data.

a) Overview of dataset look for what information can be pulled out from dataset and how it will help in respective cases
b) Look for size, information of dataset and impurities in dataset
c) Finally apply python knowledge and collect the requirements placed at step a.

4. Cleaning of data set and adding informative columns

Dataset contains unique variable customer id which has almost 135037 null values. and it is dropped from dataset. Secondly dataset is missing with description field of gifts, and it is converted to meaningful value as **"Other gifts".** Dataset has two columns unit price (specify price of each quantity) and quantity, the product of these two columns results into a new column Total_Price of Invoice. To perform deeper analysis, countryCode. *(GeeksforGeeks, 2018)* starting from 1 to 38 have been added. After cleaning dataset, we can find picture of dataset as shown below table:

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Total_Price	Country Code
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/01/2010 08:26	2.55	17850	United Kingdom	15.3	35
536365	71053	WHITE METAL LANTERN	6	12/01/2010 08:26	3.39	17850	United Kingdom	20.34	35
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/01/2010 08:26	2.75	17850	United Kingdom	22	35

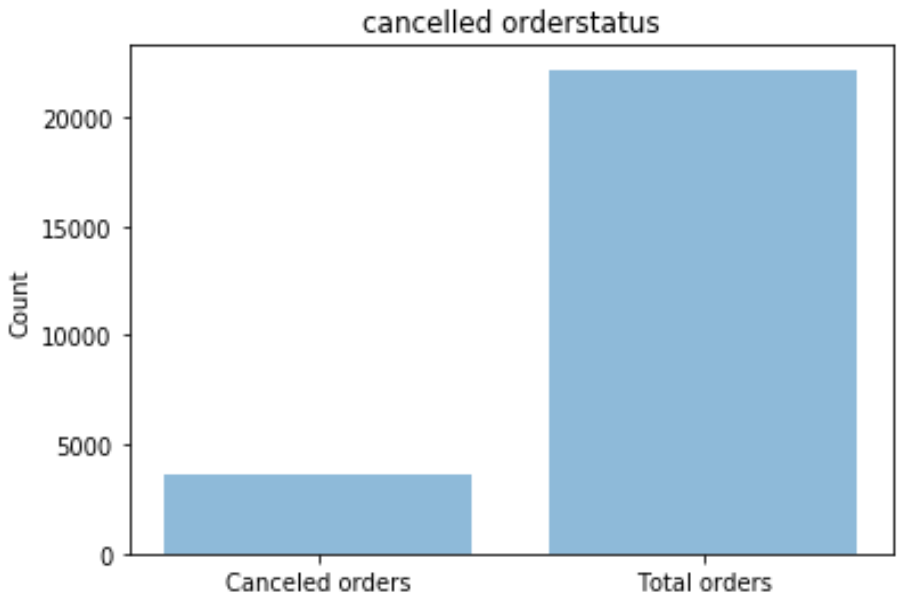
Table 4.1 overview of clean dataset

```
Int64Index: 401604 entries, 0 to 541908
Data columns (total 10 columns):
InvoiceNo      401604 non-null object
StockCode      401604 non-null object
Description    401604 non-null object
Quantity       401604 non-null int64
InvoiceDate    401604 non-null object
UnitPrice      401604 non-null float64
CustomerID     401604 non-null int64
Country        401604 non-null object
Total_Price    401604 non-null float64
CountryCode    401604 non-null int64
dtypes: float64(2), int64(3), object(5)
memory usage: 33.7+ MB
```

5. Analysis and Finding of Dataset

5.1 Cancelled orders Analysis

Any kind of loss is threat to a growing business. Cancelled orders are always an important part of transaction analysis. To derive the cancelled order percentage data was group by on the account of invoice number and customer id. It has been found that 16.47% orders were cancelled for one year. 16.47% is huge ratio of cancelled orders which needs to put some steps and further analysis on factors contributing to cancellation of order.



Total Number of cancelled orders: 3654
Total number of orders: 22190
Percent of cancelled orders: 16.47%

5.2 Basket Analysis:

By simple analysis of the raw transaction data, one can get deeper with overall business health metrics. The best metrics are the ones that are simple to calculate and directly represent real business activity. The table below illustrates some of the metrics calculated from the raw order data. *(Demacmedia.com, 2018)* Basic mathematical formulae have been applied on various column of dataset (refer code for details). Before applying logic to get below information cancelled order has been removed from dataset.

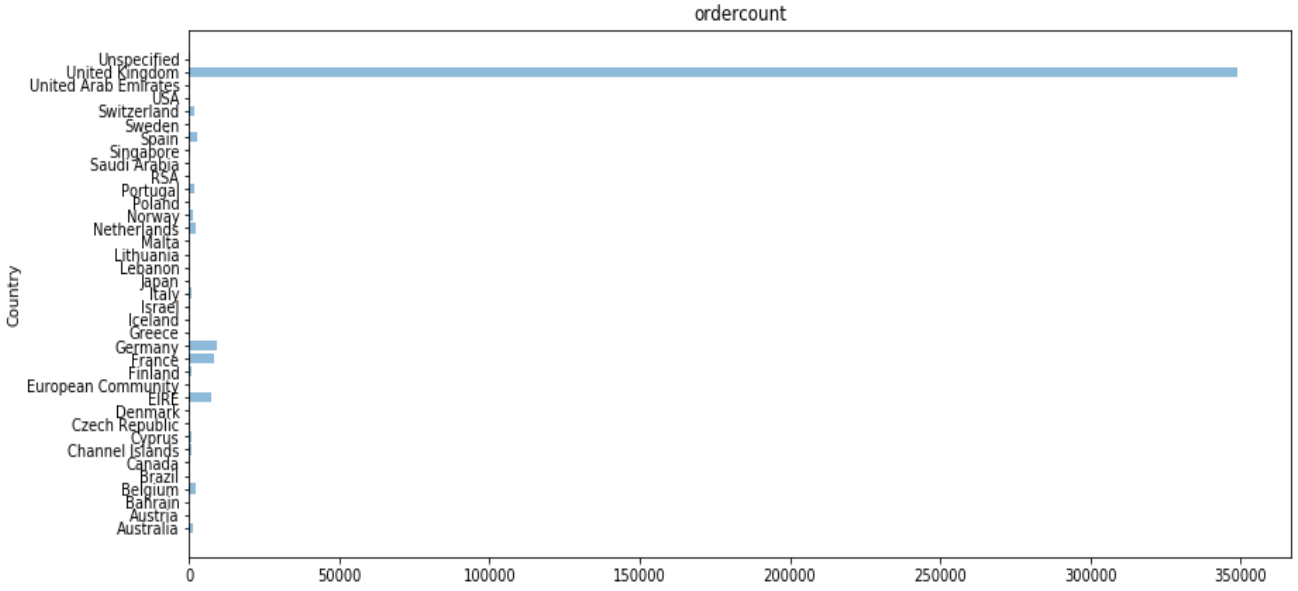
```
## deleting cancelled orders from dataset as it will not required for further analysis
dataset_test_index= dataset[dataset.InvoiceNo.str.match('^[a-zA-Z]')].index

dataset.drop(dataset_test_index, axis=0, inplace=True)
totalproduct_sold = dataset["Quantity"].sum().round(2)
print("Total product sold till date is ", totalproduct_sold)

#Total_revenue
Total_revenue= dataset["Total_Price"].sum().round(2)
print("Total Revenue till date is ", Total_revenue)

##avg order value
avgOrder_value= (Total_revenue/total_orders).round(2)
print("Avg order value of dataset is ", avgOrder_value)

##Average basket size = totalproduct sold/ number of order
avg_basketSize= (totalproduct_sold/total_orders).round(2)
print("Avg basket size of dataset is ", avg_basketSize)
```

Total product sold till data	5165886
Total Revenue till date	$ 8887208.89
Average order value of dataset	400.51
Average basket size of dataset	232.8

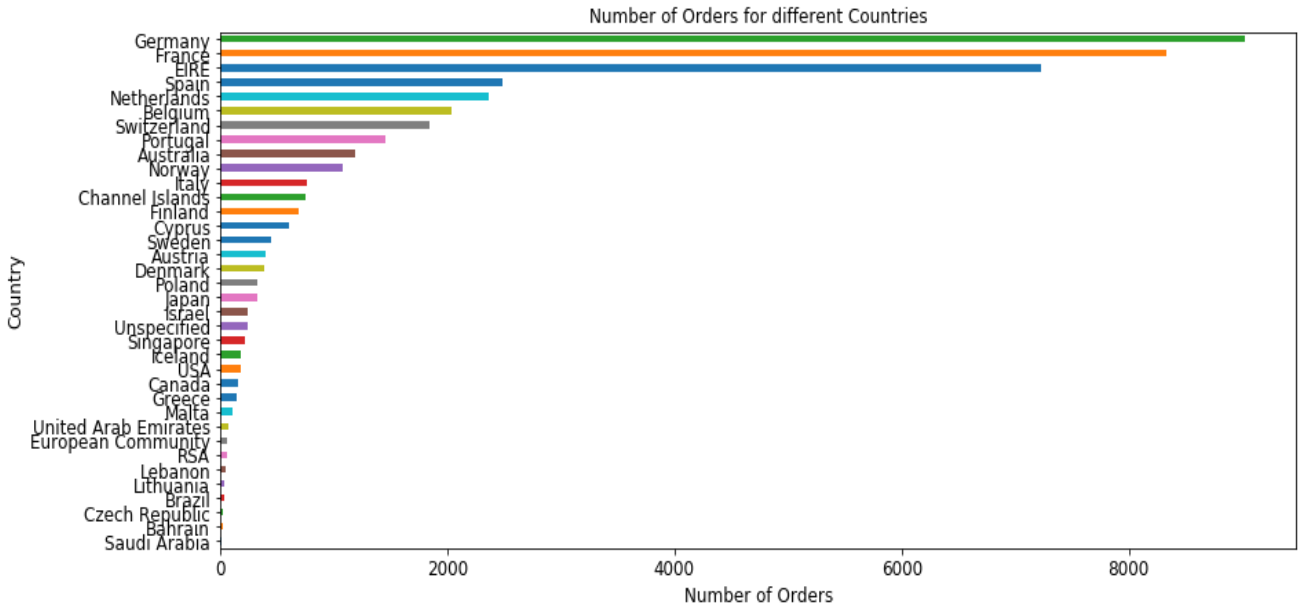
Table 5.2.1 Basket analysis

5.3 Orders according to country

Country wise order analysis has been done and it was found that UK is far ahead with 349227 orders than any other country. Germany rank 2nd with 9027 orders followed by France, EIRE, Spain respectively. Below Table 5.3.1 represents horizontal bar graph shows order count of all countries. Due to large count of UK orders one cannot see order count of any other country clearly. Hence UK order is neglected in Table 5.3.2. This is analysed by applying group by function on Invoice number and country column and taking count of invoice number in python



Graph 5.3.1 order count in different countries



Graph 5.3.2 order count in different countries (Without UK)

5.4 Customer pattern

Knowing your customer is the most significant thing in any kind of business. In ecommerce it plays a vital role. If you know your customer patterns, their needs, their changing trend you will be always two steps ahead than competition. This report has analysed top customers contributing to the revenue. Top customer who placed maximum number of orders

- **Top Five customer according to amount spend**
  The customer 14646, who made purchase of $ 280206.02 for one year comes from Netherlands. Below tabular and graphical represent shows top customer according to amount speeded on orders. This data is derived by grouping dataset by customer id, country and apply sum of total price and finally sort according to descending order.

- **Top 5 customers according to order count**

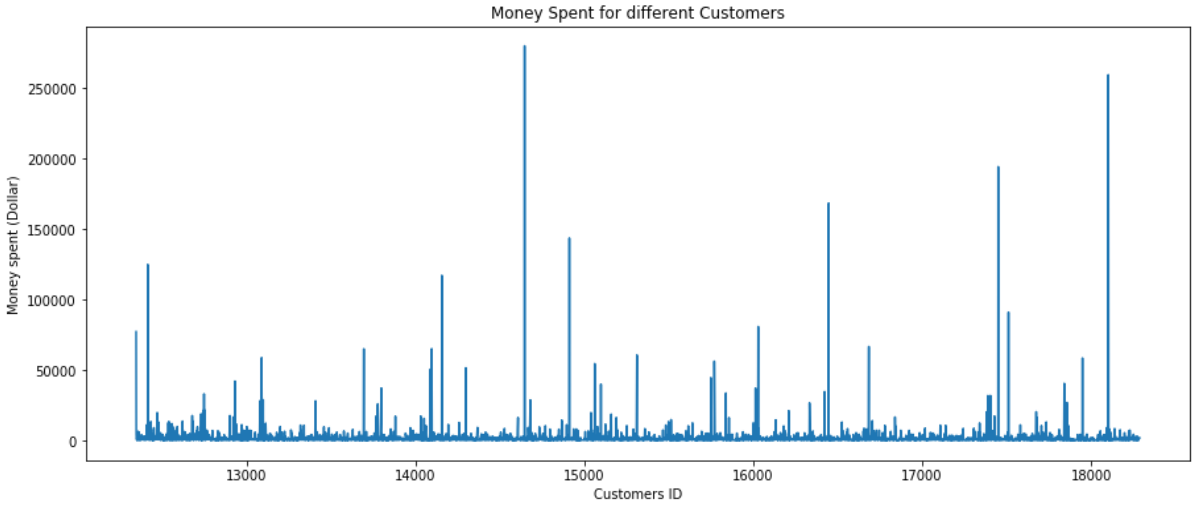
  Top customer who has ordered 7676 order for one year come from UK followed by EIRE with 5672 orders. This analysis is done by grouping the dataset by customer id and invoice number and apply sum of Invoice no and finally sort according to descending order.

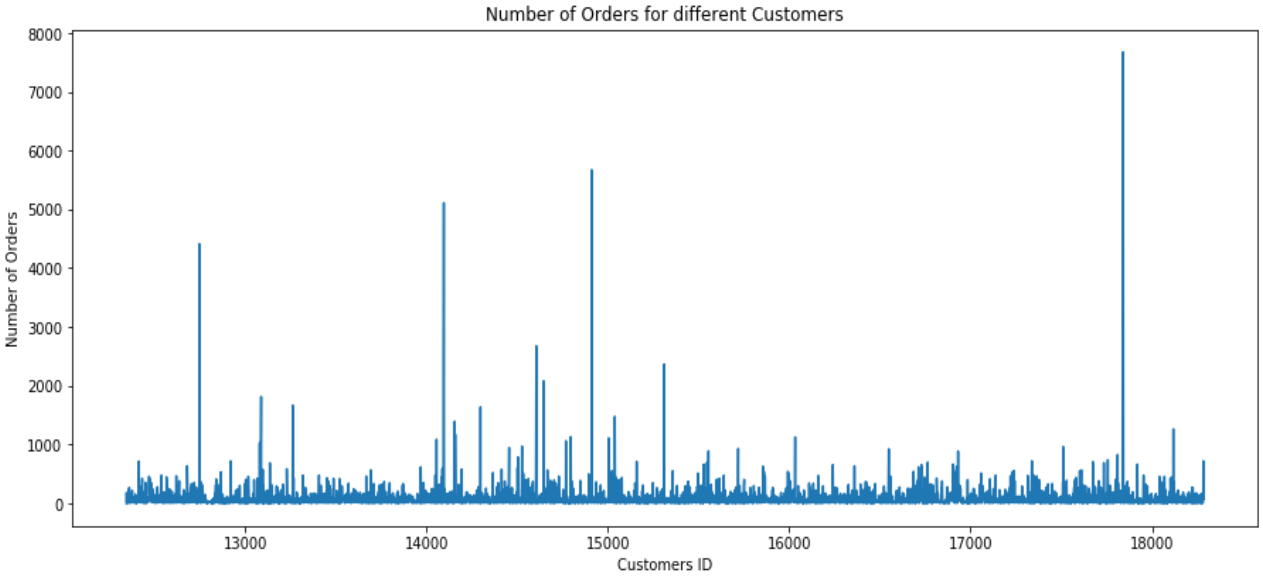
CustomerID	Country	Total_Price
14646	Netherlands	280206.02
18102	United Kingdom	259657.3
17450	United Kingdom	194390.79
16446	United Kingdom	168472.5
14911	EIRE	143711.17

5.4.1 Top 5 customer according to purchase

CustomerID	Country	InvoiceNo
17841	United Kingdom	7676
14911	EIRE	5672
14096	United Kingdom	5111
12748	United Kingdom	4413
14606	United Kingdom	2677

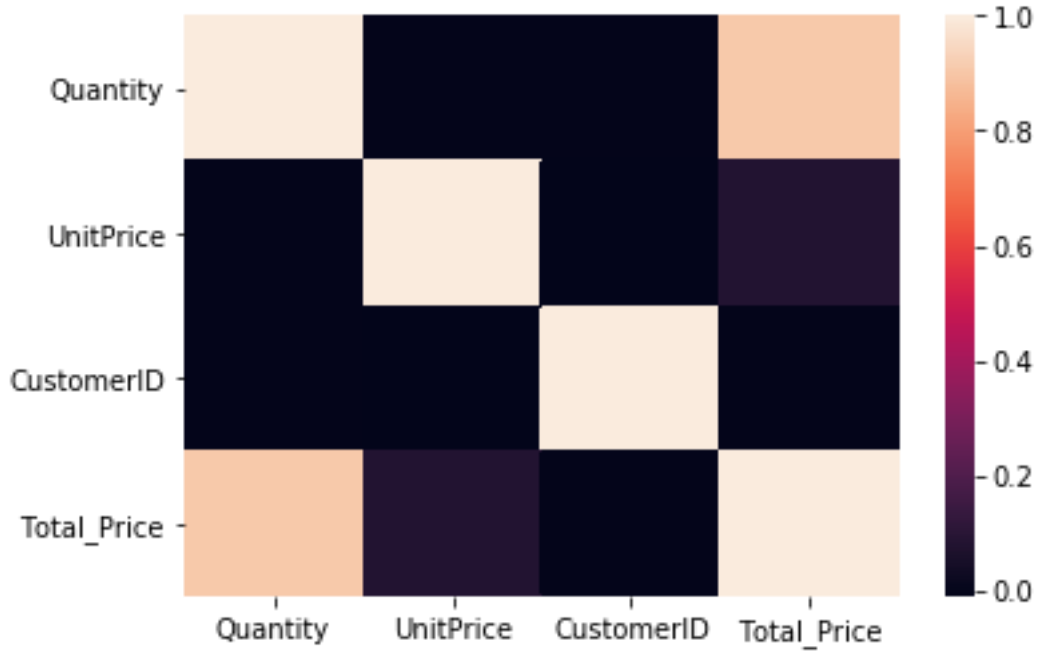
5.4.2 Top 5 customers according to order count





6. Analysis using unsupervised machine learning technique

Correlation of dataset can be tested by using heat map as shown below:



Graph 6: Heatmap showing correlation between columns of dataset

Heat map implies strong correlation between total_price and quantity purchased. Knowing the cluster of your customer will be always be better to implement marketing strategies. It helps to provide different sort of promotions to customer from different cluster. Using python kmeans clustering algorithm, it is found that customer can be clustered out according their purchased quantity and total bill of invoice. Initial step was to combine data of each customer and find out how many total quantities they purchased irrespective of orders and what was the total amount of the same. Few rows from the subset of the data looks as shown below in table 6.1. In general words, customer 12346 has spent $77183 by p urchasing 74215 products.

CustomerID	Total_Quantity	Total_Amount
12347	2458	4310
12348	2341	1797.24
12349	631	1757.55
12350	197	334.4

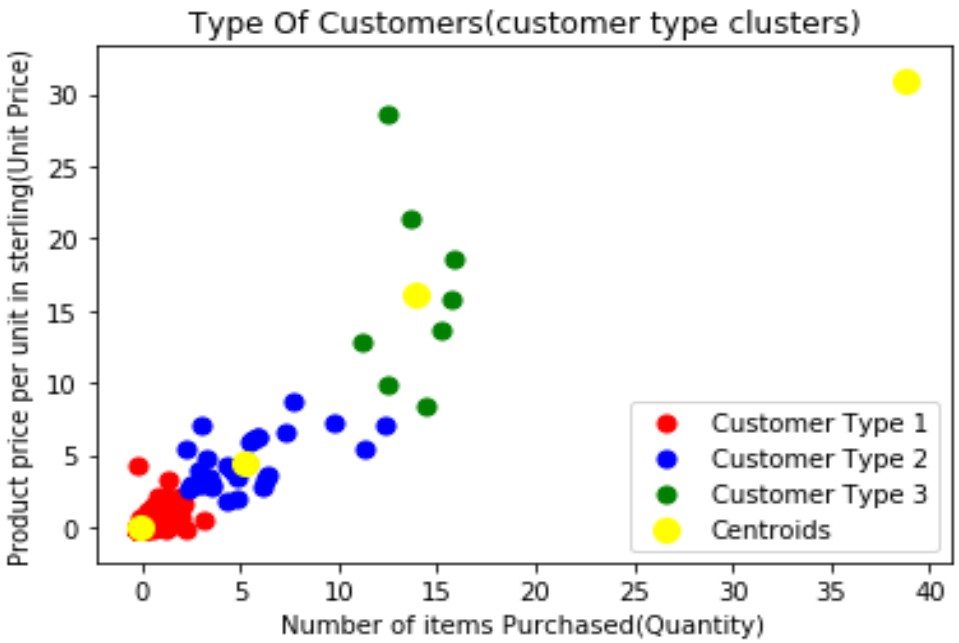
Table 6.1 customer total purchase information

Clustering is the task of dividing data points into several groups such that data points in the same groups are more like other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Using K means clustering algorithm.

K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). Here the two columns Total_Quantity and Total_Amount are treated as unlabelled and trying to cluster customer groups based on their number of quantities purchased and total bill of the same. The goal of this algorithm is to find customer groups (cluster) in the data. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. *(Trevino, 2018)*

The idea behind the use of StandardScaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1. Given the distribution of the data, each value in the dataset will have the sample mean value subtracted, and then divided by the standard deviation of the whole dataset. After trying different number to divide cluster it is found 3 is most perfect as input to algorithm as the number of clusters. Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.



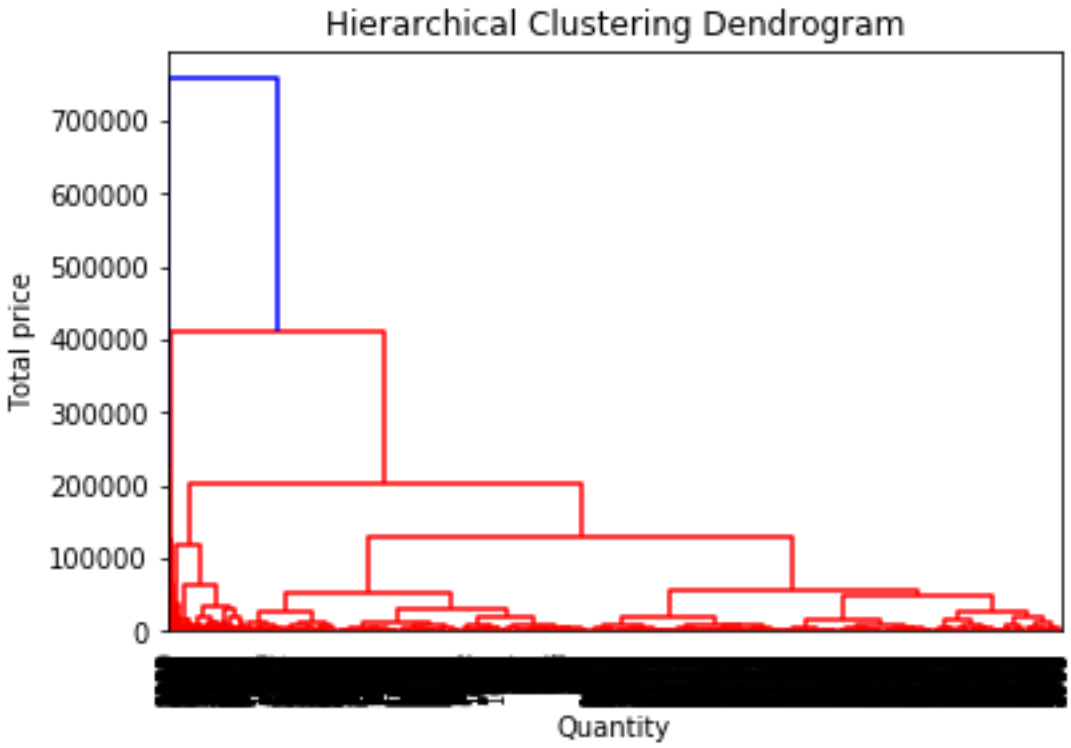
The above clustering graph is a resulting graph of customers clustered according their purchased quantity and amount spend on it. It represents mainly 3 type of customer. First type of cluster implies that there is group of customers who has highest population purchase less quantity of gifts having low prices. Second cluster tells that this type of customer purchased gifts of unit price around 6-10 pounds and quantity varying between 6 to 15 and so on.

Application of clustering result

This analysis helps in promotion, advertisement and marketing filed. For example, you can give some small discount vouchers to customer if they buy gifts having quantity more than 5. This small promotion can lead your customer of type1 to become type 2 and as a result sale will grow, so will the revenue and profit. Same rotation can be done for customer type2 and type3.

Hierarchical clustering

Hierarchical clustering, as the name suggests is an algorithm that builds hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. The results of hierarchical clustering can be shown using dendrogram. The dendrogram in this case can be interpreted in below graph



As hierarchical clustering does not give proper and clear analysis to this problem, k means cluster is best to analyse this ecommerce dataset.

7. Analysis using supervised machine learning technique

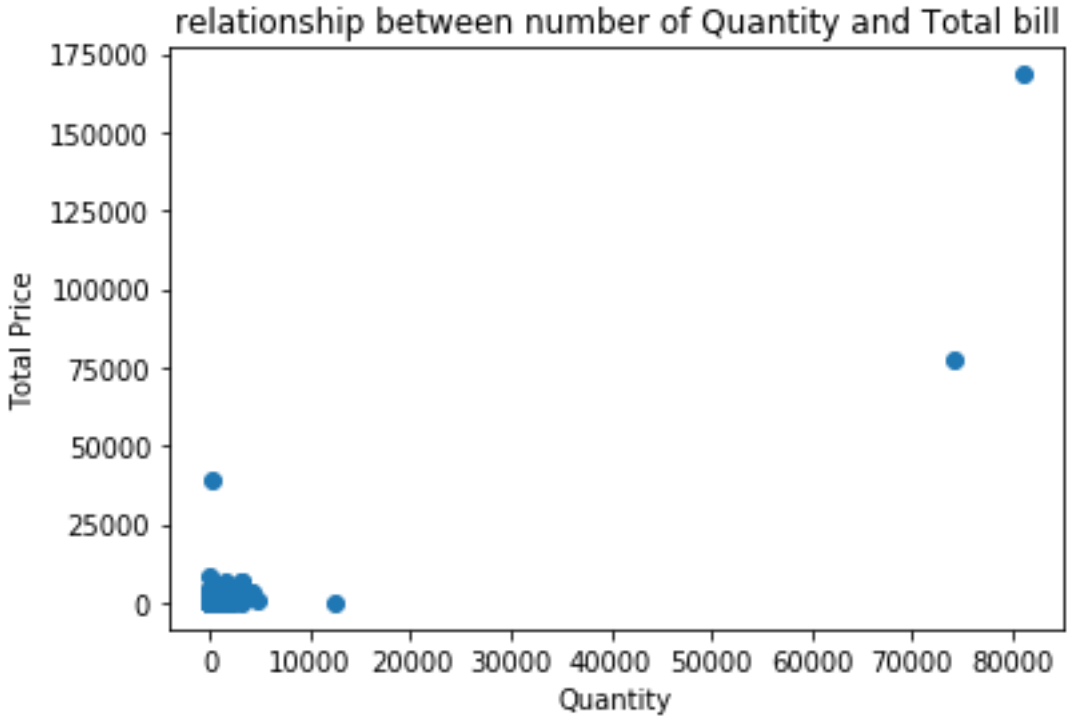
Supervised machine learning is type of algorithm in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. It provides the learning algorithms with known quantities to support future judgments. Training data for supervised learning includes a set of examples with paired input subjects and desired output. It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable. *(Analytics and clustering, 2018)*

Linear Regression is of mainly two types: Simple Linear Regression and Multiple Linear Regression. Simple Linear Regression is characterized by one independent variable. And, Multiple Linear Regression (as the name suggests) is characterized by multiple (more than 1) independent variables. While finding best fit line, you can fit a polynomial or curvilinear regression. And these are known as polynomial or curvilinear regression.

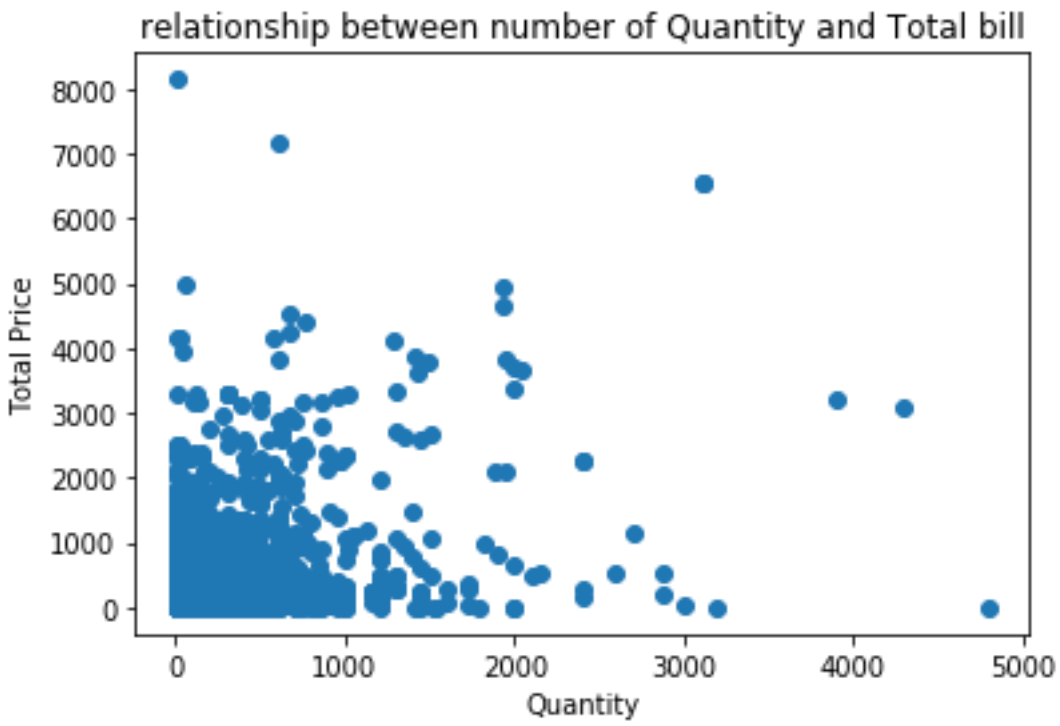
```
a= dataset[['Quantity']]
#a= a[a.Quantity <6000]
b= dataset[['Total_Price']]
#b=b[b.Total_Price< 20000]
X_train, X_test, Y_train, Y_test = sklearn.model_selection.train_test_split(a,b, test_size=0.30)
X_train.shape
Y_train.shape
lm = LinearRegression()
lm.fit(X_train, Y_train)
lm.fit(a,b)

plt.scatter(a,b)
plt.xlabel("Quantity")
plt.ylabel("Total Price")
plt.title("relationship between number of Quantity and Total bill" )
plt.show()
```

Fig7.1: linear regression code part

Here, in this analysis the relationship is established between independent (Quantity of purchase) and dependent variables (Total bill of order) by fitting a best line. Below graphs is plotted for sale prediction but due but some out linear it cannot be visualize practically. By removing some outliner, the graph is replotted as shown in Graph7.2



Graph7.1 Linear regression graph quantity vs total bill

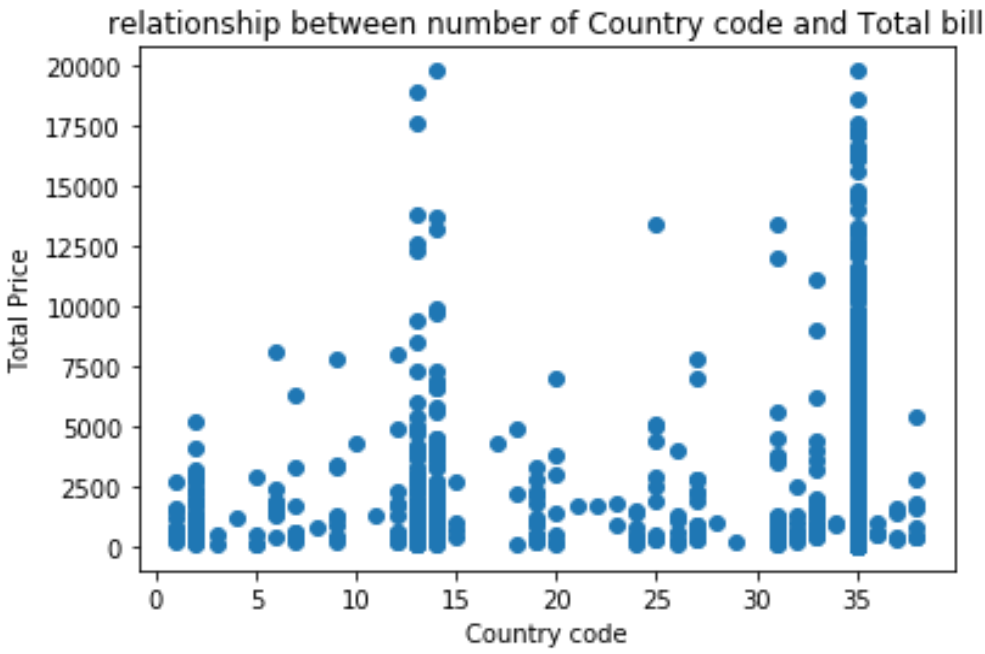


Graph7.2 Linear regression graph quantity vs total bill without out linear

Scatter plot does not give more clear prediction. Further analysis is required to predict future sale of company

Another attempt to get some linear regression from dataset

This time total price is plotted vs countrycode and analysis is done to predict future sale of specific country. As result below graph as been plotted using linear regression algorithm. It seems no relationship and it is assumed that furthermore analsysis required to predict future sale of company.



Graph7.3 Linear regression graph Country code vs total bill without out linear

8. Reflection on methods used for analysis:

Python numpy, pandas have a vast collection of statistics methods that can be used on any complex type of dataset. numpy, has long been a cornerstone of numerical computing in Python. It provides the data structures, algorithms, and library glue needed for most scientific applications involving numerical data in Python. Pandas provide high-level data structures and functions designed to make working with structured or tabular data fast, easy and expressive. This analysis also used many basic methods like groupby, sortby, mean, max etc. method used for analysis on numpy array, dataframes etc. K clustering works like a magic for unlabelled dataset. One can put many interesting analyses using k means from the any sort of dataset. Further to predict future sale liner regression algorithms is used.

Challenges, limitation and future analysis:

Some challenges faced during analsysis of ecommerce data were to clean data without much loss of data, finding best variables on which supervised algorithms can be applied. Dealing with memory error while working with huge dataset and some matrix algorithms. Some gift has zero-unit price which predicts they are given as gift with some orders. More clear analysis is required on same.
Future analsysis can be done on date and time series predicting sale of each day, each month etc.

9. Conclusions:

Various type of analysis method in python are used to analyse ecommerce transaction data in this assignment, which is beneficial to company for its growth. It can be concluded using simple statistics method on transaction data can give very clear and strong picture of business. Use of graphical libraries of python helped to visualize analysis process. Performing unsupervised machine learning algorithms opens doors of many facts which can be used for marketing and advertisement purpose.

References

1. kaggle. (2018). [online] Available at: https://www.kaggle.com/carrie1/ecommerce-data [Accessed 5 Nov. 2018]. (Source of datset) (kaggle, 2018)
2. Demacmedia.com. (2018). Making your eCommerce Data Analysis More Effective. [online] Available at: https://www.demacmedia.com/effective-data-analysis/ [Accessed 15 Oct. 2018].
   (Demacmedia.com, 2018)

3. GeeksforGeeks. (2018). Replacing strings with numbers in Python for Data Analysis - GeeksforGeeks. [online] Available at: https://www.geeksforgeeks.org/replacing-strings-with-numbers-in-python-for-data-analysis/ [Accessed 14 Oct. 2018]. (GeeksforGeeks, 2018)

4. Trevino, A. (2018). Introduction to K-means Clustering. [online] Datascience.com. Available at: https://www.datascience.com/blog/k-means-clustering [Accessed 24 Oct. 2018]. (Trevino, 2018)
5. Analytics, B. and clustering, A. (2018). An Introduction to Clustering & different methods of clustering. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/ [Accessed 5 Nov. 2018]. (Analytics and clustering, 2018)
6. Python for Data Analysis (online pdf)
7. 2015DataScience&BigDataAnalytics
8. 2017_Book_IntroductionToDataScience

Appendix

- Python version: Python 3.6.1
- Jupyter notebook version:

  3.7.0 (default, Jun 28 2018, 08:04:48) [MSC v.1912 64 bit (AMD64)]

- Packed used:
  Numpy,pandas, matplotlib.pyplot, seaborn, sklearn, scipy.cluster.hierarchy, dendrogram, linkage. Sklearn sub packages: metrics, linear_model, LinearRegression, StandardScaler KMeans