**EXPLANATION OF ASSIGNMENT 1**

**Abhranil Chandra, IIT Kharagpur**

**DATA CLEANING:**

I first imported the dataset as a pandas dataframe. But the dataset contained almost 25% of data with missing UserID(represented as -1). But deleting all of them would have led to loss of a huge amount of data and theoretically at most 5% of data can be deleted without affecting results much.

So I logically correlated identical TransactionIDs and filled in the missing UserIDs with values lower than the minimum valid UserId in proper order. I was tempted to drop the unknown UserId rows first but the loss of data would be huge so I took this unconventional step. But if needed the same analysis can be done on the correct UserId rows as well as I have stored them in a separate dataframe(df\_ok).

I removed duplicates, dropped about 2% rows with missing ItemCode and Description.

**BASIC EDA:**

I there after did some basic EDA on the final dataset like number of number of customers per country, on Quantity, on UnitPrice, handled Cancelled transactions ie. the ones with negative “Quantity”

**FEATURE ENGINEERING:**

I added “TotalPrice” column- product of Quantity and UnitPrice & plotted Revenue per Country vs Country, also plotted No. of Invoices per Country vs Country.

Label encoded the countries, converted the TransactionTime to standard timestamp format,created a Recency column to measure the time of last visit to website. Created separate columns for Day, Month, Year.

**ADVANCED FEATURE ENGINEERING:**

RFM Score Definition : RFM analysis is a data driven customer behavior segmentation technique. RFM stands for recency, frequency, and monetary value. The idea is to segment customers based on when their last purchase was, how often they've purchased in the past, and how much they've spent overall.

|  |  |  |  |
| --- | --- | --- | --- |
| Segment | RFM Score | Description | Marketing |
| Best Customers | 111 | Bought recently,frequently, high priced items | No price incentives, new products, loyalty programs |
| Loyal Customers | \*1\* | Buy most frequently | Use R and M to further decide |
| Big Spenders | \*\*1 | Spend the most | Market most expensive products |
| Almost Lost | 311 | Haven't shopped for some time, but spend high and used to do so frequently | Aggressive advertising and price incentives, special discounts and offers |
| Lost Customers | 4\*\* | Shopped long ago so may have switched to other sellers | Aggressive advertising and price incentives, special discounts and offers |
| Lost Cheap Customers | 444 | Shopped long ago,shop cheap,shop less | Not that profitable customers so no need to spend much to reacquire |

Added RFMScore column to store the RFM score, the main metric to evaluate customers.

Another crucial step to cluster customers is to first cluster the products and items. So I used nltk and word vectoriser to analyse the item descriptions and cluster the items into 135 different clusters by K-means clustering(silhouette score was highest for 135 so I chose that many cluster)

**CUSTOMER SEGMENTATION:**

I applied K-Means Clustering to classify the customers into segregated baskets. The silhouette score of 2 and 3 clusters were maximum but it is not that insightful to classify only into 2 or 3 clusters. So I chose 6 clusters as the optimal number of clusters. Further I did analysis of the major segment- cluster 0 which contained almost 90% of customers.

**INSIGHTS DRAWN FROM EDA ON CLUSTER 0:**

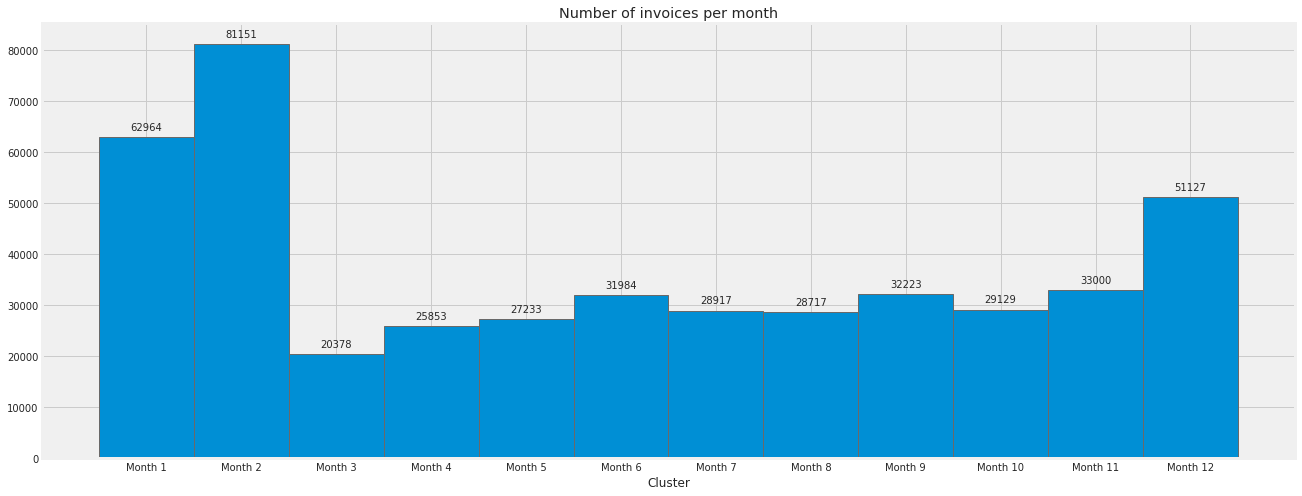
* Maximum number of customers in this cluster(almost 90%).
* Out of these customers quiet a substantial number are ideal customers(RFM score 111), but also almost equal amount of worst customers(RFM score 444), but overall the cluster has customers who visit the e-commerce site once in a while, last visited quite some time ago and spend medium to less amount of money.
* Key figures:

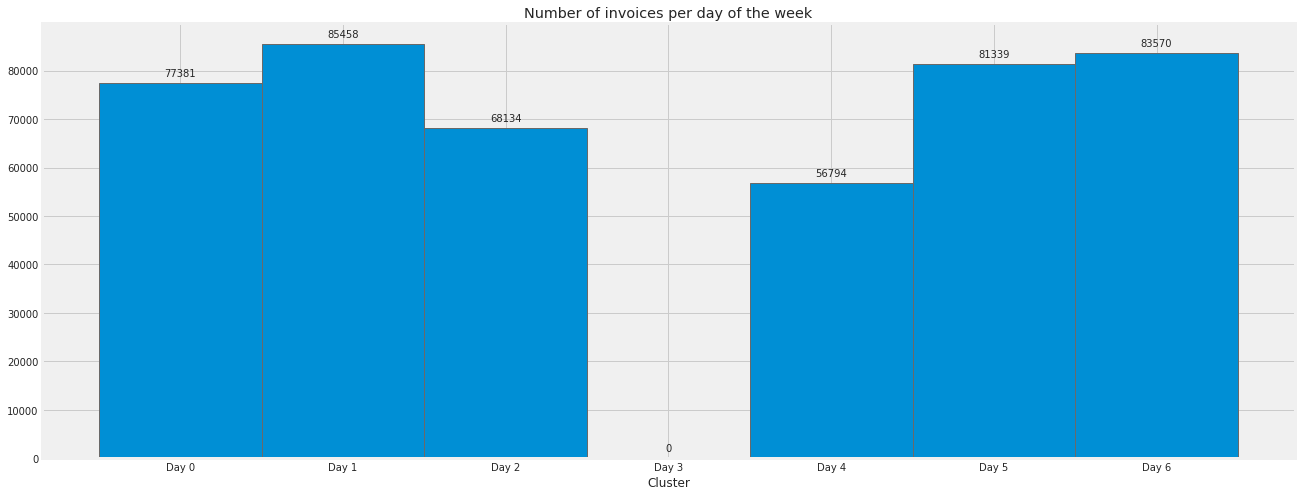
1. Quantity 25.749616
2. UnitPrice 4.617257
3. Total Price 66.660905
4. frequency 6.560935
5. min\_recency 89.545808(almost 3 months)

* TOP 10 bought products :

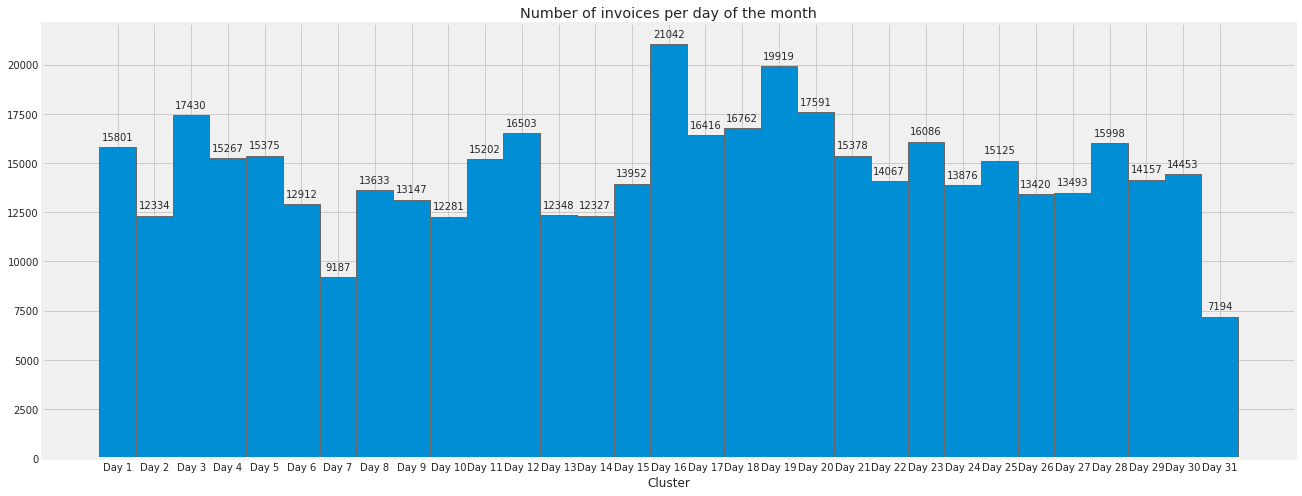
1. WHITE HANGING HEART T-LIGHT HOLDER: 1345
2. JUMBO BAG RED RETROSPOT: 1079
3. REGENCY CAKESTAND 3 TIER: 960
4. ASSORTED COLOUR BIRD ORNAMENT: 926
5. PARTY BUNTING: 924
6. LUNCH BAG RED RETROSPOT: 898
7. LUNCH BAG BLACK SKULL: 753
8. SET OF 3 CAKE TINS PANTRY DESIGN: 725
9. LUNCH BAG CARS BLUE: 679
10. LUNCH BAG PINK POLKADOT: 676

* Most customers were from U.K.
* Maximum sales accrued in the months of November, December, January & February which are the pre and post festivity months for the European countries where almost all of the sales occurred. Also the shopping spree is more dominant in the weekends- Friday, Saturday and Sunday with 80k+, 80k+ & 70k+ invoices pre day on these days.





* Also noticeable is the fact that sales peak during the middle of the months



Similarly we can also analyse the remaining five clusters with exactly the same code only replacing the cluster number where necessary. As the code and analysis is exactly the same and repetitive and would unnecessarily lengthen the jupyter notebook and its run time, so I excluded their code.