Kedar Pinge

Customer Retention Analysis

A detailed report about retaining retail customers for XYZ and its unforeseen scope for growth.

NMIMS, March 2021

Abstract

Having recorded an immense amount of data over a period of 36 months, we are at a posititon of deploying various types of Retention Analysis techniques to make the most out of one of the most overlooked aspects in Customer Analysis. In spite of a company's prime focus being given to customer acquisition, it is said that it's about seven times cheaper to retain a customer than to acquire a new one.

We will start off with some basic exploratory data analysis, and then move on to deeper retention techniques.

We will be using both, R and Python languages for this purpose, both of the codes being attached to the file.

Exploratory Data Analysis

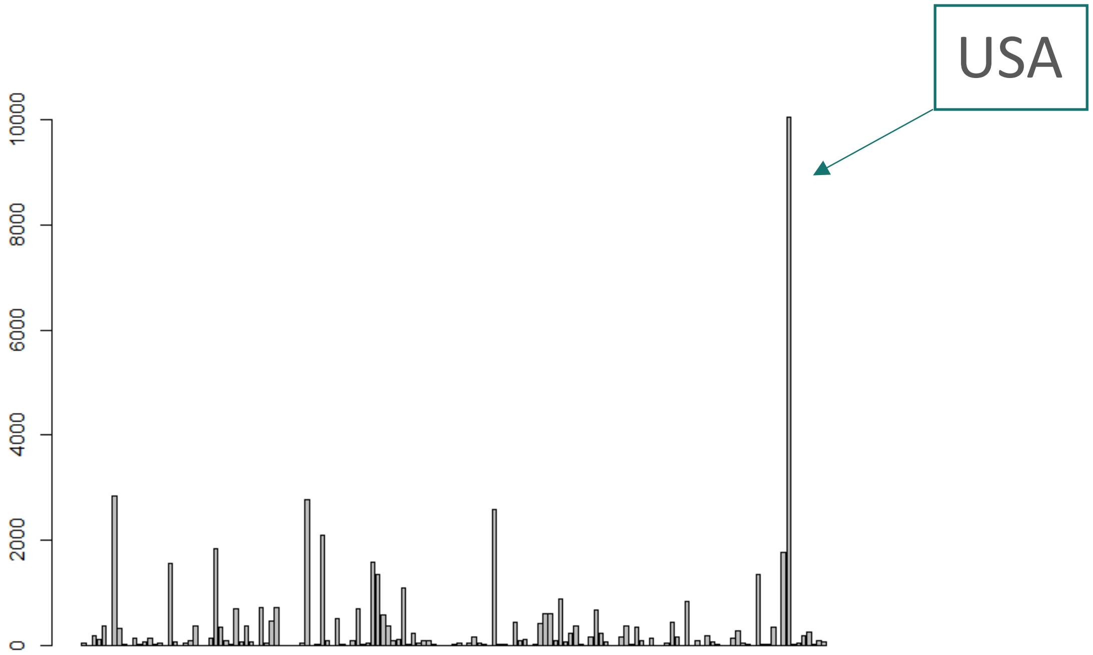
Let's start summarising the data and make basic inferences from it.

Checking for the missing values, 251 in the column of Unit price

Groups	Count	Proportion
Customer_ID	0	0.000000
Sno	0	0.000000
Order_Date	0	0.000000
Quantity	0	0.000000
Unit_Price	251	0.004894

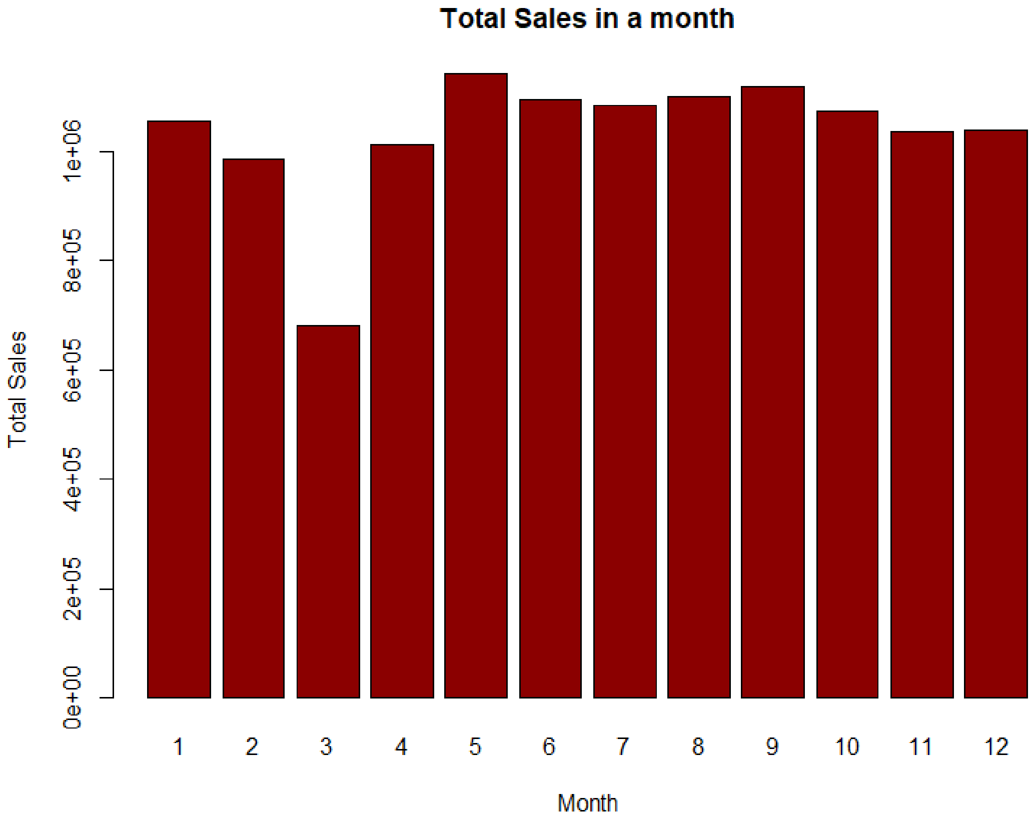
Since it's a very small number compared to the data, we can omit those.

Looking at the distribution of Sales grouped by countries, we can tell USA has the highest amount of customers by an extremely large margin



USA

We can also take a look at the total sales by first creating a column called 'Total Sales' by multiplying the Quanitity with Unit Price, and then grouping the values by each month. The following barplot shows a significant reduction of sales in March, with other months having similar amounts in total sales. This is because of the one time we have skipped march while calculating, since the available data is from April of 2018



COHORT MODEL

Cohort Analysis is proficiently used to group customers in retail analysis and study their patterns based on the groups they belong to. For example, we can create cohorts based on the month these customers made their first purchase.

CUSTOMER LIFETIME VALUE (CLV)

We will be using this concept thoroughly within this report in order to assess the total amount of money a customer is ready to spend on your products in his lifetime. Retail Industry doesn't have customers that pay on a subscription basis, so we don't see windows of opportunities to retain them before we lose them completely once they churn.

A generic formula for calculating the CLV is :

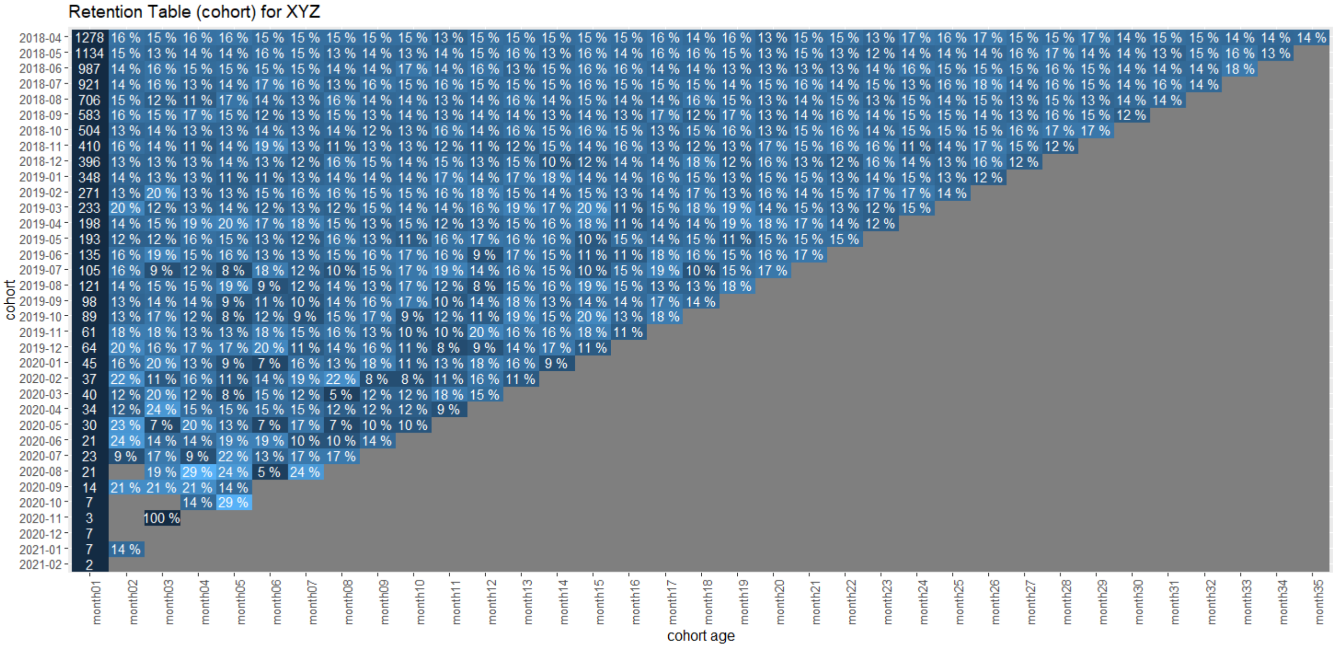
CLV = ((Average Sales X Purchase Frequency) / Churn) X Profit Margin

Purely Based on this formula, we have a CLV value of **$19036.05.**

This is definetely an absurd value, caused purely because of the few customers that spent a lot of money. So this approach needs to be manipulated in some ways just like we did earlier.

Values for each cohort :

	Months	CLV
0	Jan	6756.54
1	Feb	8731.56
2	March	4952.77
3	Apr	52062.80
4	May	21568.15
5	Jun	42777.30
6	Jul	39594.57
7	Aug	17870.69
8	Sep	11138.14
9	Oct	10259.25
10	Nov	10208.46
11	Dec	6386.78



Retention Table (cohort) for XYZ

cohort	month01	month02	month03	month04	month05	month06	month07	month08	month09	month10	month11	month12	month13	month14	month15	month16	month17	month18	month19	month20	month21	month22	month23	month24	month25	month26	month27	month28	month29	month30	month31	month32	month33	month34	month35
2018-04	1278	16 %	15 %	16 %	16 %	15 %	15 %	15 %	15 %	15 %	13 %	15 %	15 %	15 %	15 %	15 %	16 %	14 %	16 %	13 %	15 %	15 %	13 %	17 %	16 %	17 %	15 %	15 %	17 %	14 %	15 %	15 %	14 %	14 %	14 %
2018-05	1134	15 %	13 %	14 %	14 %	16 %	15 %	13 %	14 %	13 %	14 %	15 %	16 %	13 %	16 %	14 %	16 %	16 %	15 %	13 %	15 %	13 %	12 %	14 %	14 %	14 %	16 %	17 %	14 %	14 %	13 %	15 %	16 %	13 %	
2018-06	987	14 %	16 %	15 %	15 %	15 %	15 %	14 %	14 %	17 %	14 %	16 %	13 %	15 %	16 %	16 %	14 %	14 %	13 %	13 %	13 %	13 %	14 %	16 %	15 %	15 %	15 %	16 %	15 %	14 %	14 %	14 %	18 %		
2018-07	921	14 %	16 %	13 %	14 %	17 %	16 %	13 %	16 %	15 %	16 %	15 %	15 %	15 %	16 %	15 %	15 %	15 %	14 %	15 %	16 %	14 %	15 %	13 %	16 %	18 %	14 %	16 %	15 %	14 %	16 %	14 %			
2018-08	706	15 %	12 %	11 %	17 %	14 %	13 %	16 %	14 %	14 %	13 %	14 %	16 %	14 %	15 %	14 %	14 %	16 %	15 %	13 %	14 %	15 %	13 %	15 %	14 %	15 %	13 %	15 %	13 %	14 %	14 %				
2018-09	583	16 %	15 %	17 %	15 %	12 %	13 %	15 %	13 %	14 %	13 %	14 %	14 %	13 %	14 %	13 %	17 %	12 %	17 %	13 %	14 %	16 %	14 %	15 %	15 %	14 %	13 %	16 %	15 %	12 %					
2018-10	504	13 %	14 %	13 %	13 %	14 %	13 %	14 %	12 %	13 %	16 %	14 %	16 %	15 %	16 %	15 %	13 %	15 %	16 %	13 %	15 %	16 %	14 %	15 %	15 %	15 %	16 %	17 %	17 %						
2018-11	410	16 %	14 %	11 %	14 %	19 %	13 %	11 %	13 %	13 %	12 %	11 %	12 %	15 %	14 %	16 %	13 %	12 %	13 %	17 %	15 %	16 %	16 %	11 %	14 %	17 %	15 %	12 %							
2018-12	396	13 %	13 %	13 %	14 %	13 %	12 %	16 %	15 %	14 %	15 %	13 %	15 %	10 %	12 %	14 %	14 %	18 %	12 %	16 %	13 %	12 %	16 %	14 %	13 %	16 %	12 %								
2019-01	348	14 %	13 %	13 %	11 %	11 %	13 %	14 %	14 %	14 %	17 %	14 %	17 %	18 %	14 %	14 %	16 %	15 %	13 %	15 %	15 %	13 %	14 %	15 %	13 %	12 %									
2019-02	271	13 %	20 %	13 %	13 %	15 %	16 %	16 %	15 %	15 %	16 %	18 %	15 %	14 %	15 %	13 %	14 %	17 %	13 %	16 %	14 %	15 %	17 %	17 %	14 %										
2019-03	233	20 %	12 %	13 %	14 %	12 %	13 %	12 %	15 %	14 %	14 %	16 %	19 %	17 %	20 %	11 %	15 %	18 %	19 %	14 %	15 %	13 %	12 %	15 %											
2019-04	198	14 %	15 %	19 %	20 %	17 %	18 %	15 %	13 %	15 %	12 %	13 %	15 %	16 %	18 %	11 %	14 %	14 %	19 %	18 %	17 %	14 %	12 %												
2019-05	193	12 %	12 %	16 %	15 %	13 %	12 %	16 %	13 %	11 %	16 %	17 %	16 %	16 %	10 %	15 %	14 %	15 %	11 %	15 %	15 %	15 %													
2019-06	135	16 %	19 %	15 %	16 %	13 %	13 %	15 %	16 %	17 %	16 %	9 %	17 %	15 %	11 %	11 %	18 %	16 %	15 %	16 %	17 %														
2019-07	105	16 %	9 %	12 %	8 %	18 %	12 %	10 %	15 %	17 %	19 %	14 %	16 %	15 %	10 %	15 %	19 %	10 %	15 %	17 %															
2019-08	121	14 %	15 %	15 %	19 %	9 %	12 %	14 %	13 %	17 %	12 %	8 %	15 %	16 %	19 %	15 %	13 %	13 %	18 %																
2019-09	98	13 %	14 %	14 %	9 %	11 %	10 %	14 %	16 %	17 %	10 %	14 %	18 %	13 %	14 %	14 %	17 %	14 %																	
2019-10	89	13 %	17 %	12 %	8 %	12 %	9 %	15 %	17 %	9 %	12 %	11 %	19 %	15 %	20 %	13 %	18 %																		
2019-11	61	18 %	18 %	13 %	13 %	18 %	15 %	16 %	13 %	10 %	10 %	20 %	16 %	16 %	18 %	11 %																			
2019-12	64	20 %	16 %	17 %	17 %	20 %	11 %	14 %	16 %	11 %	8 %	9 %	14 %	17 %	11 %																				
2020-01	45	16 %	20 %	13 %	9 %	7 %	16 %	13 %	18 %	11 %	13 %	18 %	16 %	9 %																					
2020-02	37	22 %	11 %	16 %	11 %	14 %	19 %	22 %	8 %	8 %	11 %	16 %	11 %																						
2020-03	40	12 %	20 %	12 %	8 %	15 %	12 %	5 %	12 %	12 %	18 %	15 %																							
2020-04	34	12 %	24 %	15 %	15 %	15 %	15 %	12 %	12 %	12 %	9 %																								
2020-05	30	23 %	7 %	20 %	13 %	7 %	17 %	7 %	10 %	10 %																									
2020-06	21	24 %	14 %	14 %	19 %	19 %	10 %	10 %	14 %																										
2020-07	23	9 %	17 %	9 %	22 %	13 %	17 %	17 %																											
2020-08	21		19 %	29 %	24 %	5 %	24 %																												
2020-09	14	21 %	21 %	21 %	14 %																														
2020-10	7			14 %	29 %																														
2020-11	3		100 %																																
2020-12	7																																		
2021-01	7	14 %																																	
2021-02	2																																		

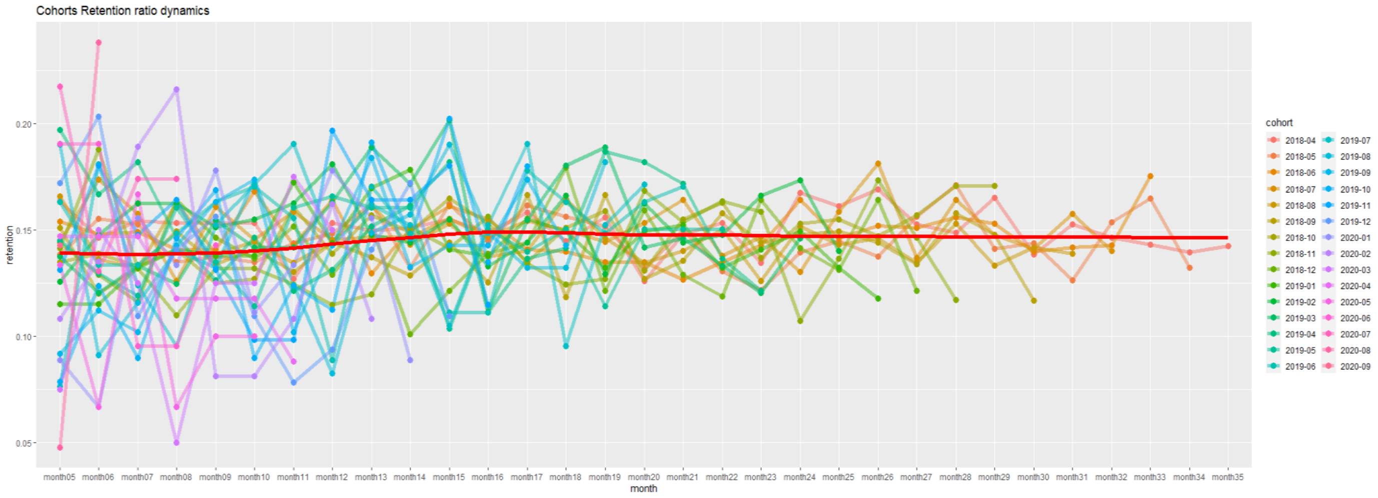
The
a

Above diagram displays the retention by the customers in percentage for further analysis which can be used to find the points at which we lose customers, and where we can introduce various strategies such as promo codes and discounts.

We can calculate the Retention Ratio by the formula :

Retention ratio = # clients in particular month / # clients in 1st month of life-time

Firstly, we will visualize Cohort Retention Ratio Dynamics :



There’s a consistent Retention Rate that we’re observing in this scenario, so we don’t really segregate people based on the months, especially in the retail industry since people will be buying the products all year round. However, this analysis can help us determine which month we’re able to maintain our customer retention ratio in other industries.

BETA GEOMETRIC / NEGATIVE BINOMIAL MODEL

This is one of the most commonly used probabilistic model for predicting CLV (as defined earlier). This model is used to predict the future transactions of a customer, which combines with the Gamma-Gamma model, which then adds the monetary transactions of a customer to finally give us the CLV.

There are however, certain **assumptions** I’ll be making about the dataset provided for predicting future transactions :

1)When a user is active, number of transactions in a time t is described by *Poisson distribution* with rate lambda.

2)Heterogeneity in transaction across users (difference in purchasing behavior across users) has *Gamma distribution* with shape parameter r and scale parameter a.

3)Users may become inactive after any transaction with probability p and their dropout point is distributed between purchases with *Geometric distribution*.

4)Heterogeneity in dropout probability has *Beta distribution* with the two shape parameters alpha and beta.

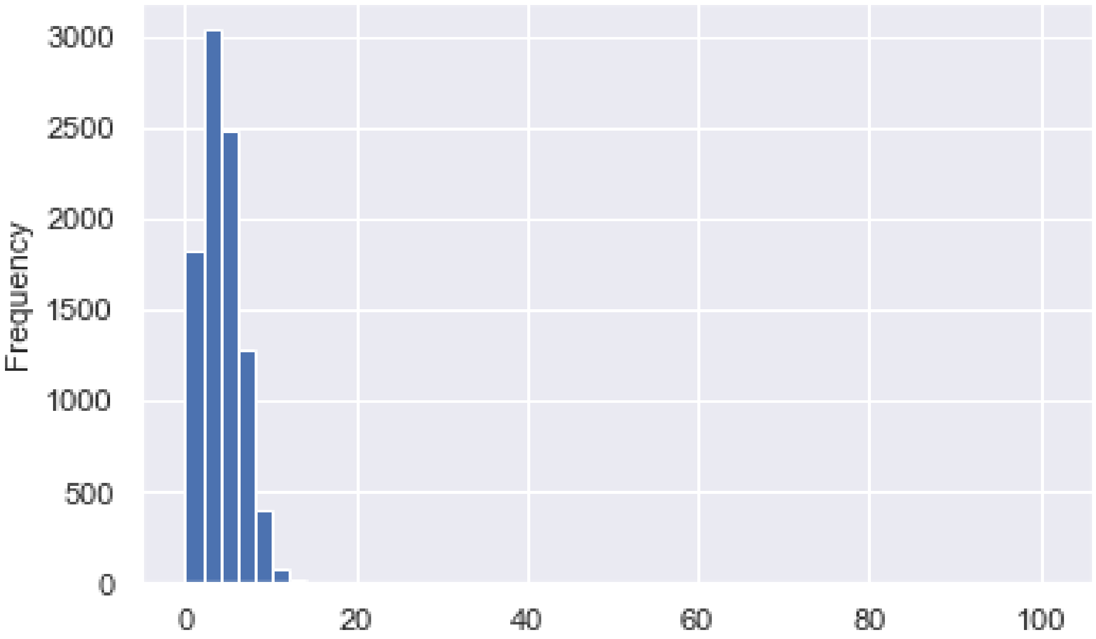
5)Transaction rate and dropout probability vary independently across users.

I’ll be using the package lifetimes in python to calculate CLV and predicting future transactions.

We create a summary table based on Recency(Time between first and last transaction) , Frequency(Number of Repeat Purchases) and Monetary Value (Mean of the Customer Sales value

Using the frequency table, we find out :

```
Percentage of customers that purchase the item only once: 2.27 %
count    9126.000000
mean        4.576266
std         3.079484
min         0.000000
25%         3.000000
50%         4.000000
75%         6.000000
max       101.000000
Name: frequency, dtype: float64
----------------------------------------
```

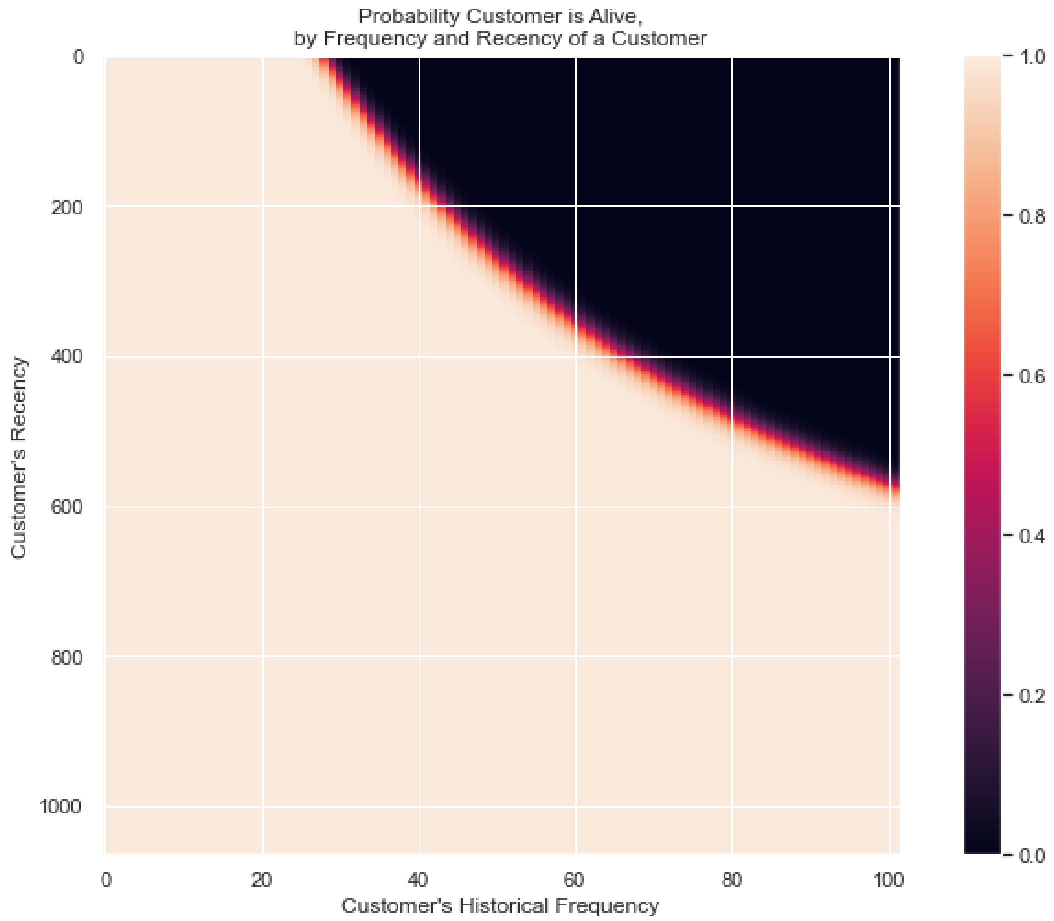


Now we fit the model using the in-package function.

We add a penalizer coefficient of 0.05 since we have a lot of retaining customers.

We then calculate the customer alive probability

Visual representation of relationship between recency and frequency :



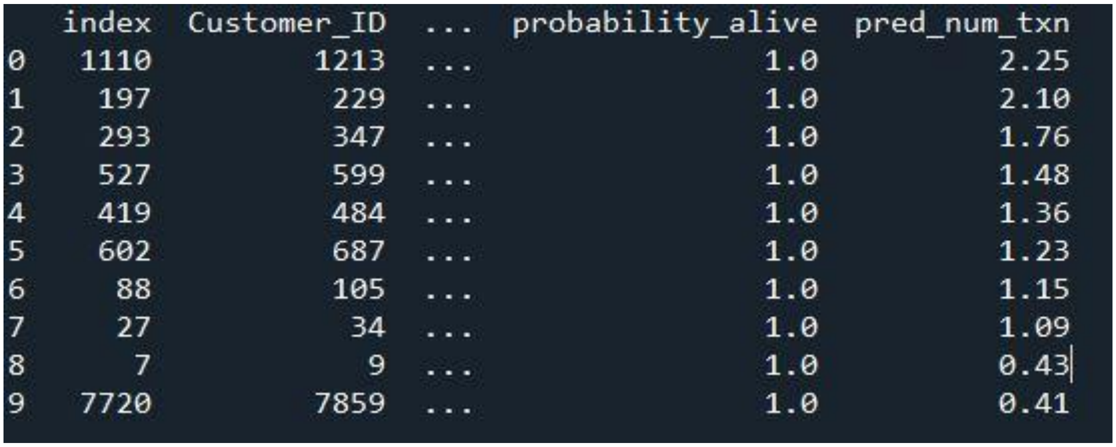
The probability of being alive is calculated based on the recency and frequency of a customer.

If a customer has bought multiple times (frequency) and the time between first & last transaction is high (recency), then his/her probability being alive is high.

Similarly, if a customer has less frequency (bought once or twice) and the time between first & last transaction is low (recency), then his/her probability being alive is high.

Next thing we can do with this trained model is to predict the likely future transactions for each customer.

Using the function : **model.conditional_expected_number_of_purchases_up_to_time()**



```
   index  Customer_ID  ...  probability_alive  pred_num_txn
0   1110         1213  ...                1.0          2.25
1    197          229  ...                1.0          2.10
2    293          347  ...                1.0          1.76
3    527          599  ...                1.0          1.48
4    419          484  ...                1.0          1.36
5    602          687  ...                1.0          1.23
6     88          105  ...                1.0          1.15
7     27           34  ...                1.0          1.09
8      7            9  ...                1.0          0.43
9   7720         7859  ...                1.0          0.41
```

This is the predicition of the number of purchases a customer makes in the next 10 days

Now, we have predicted the future transactions, so we move on to predicting the monetary value for each customer using the Gamma Gamma model. However, we are considering only customers who made repeat purchases with the business i.e., frequency > 0. Because, if frequency is 0, it means that they are one time customer and are considered already 'dead ' .

Some of the key assumptions of Gamma-Gamma model are:

The monetary value of a customer's given transaction varies randomly around their average transaction value.

Average transaction value varies across customers but do not vary over time for any given customer.

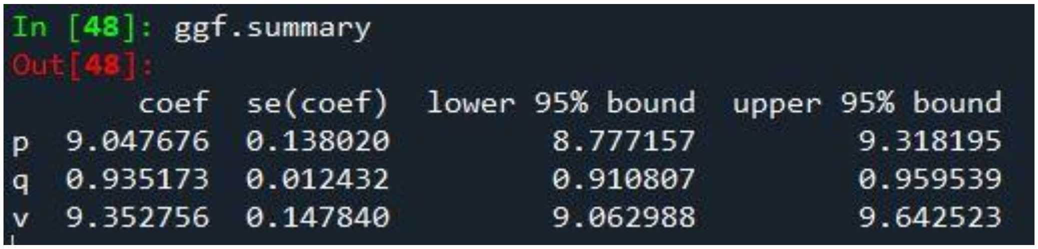
The distribution of average transaction values across customers is independent of the transaction process.

Finding correlation between frequency and monetary value :

	Frequency	monetary_value
frequency	1.000000	0.008634
monetary_value	0.008634	1.000000

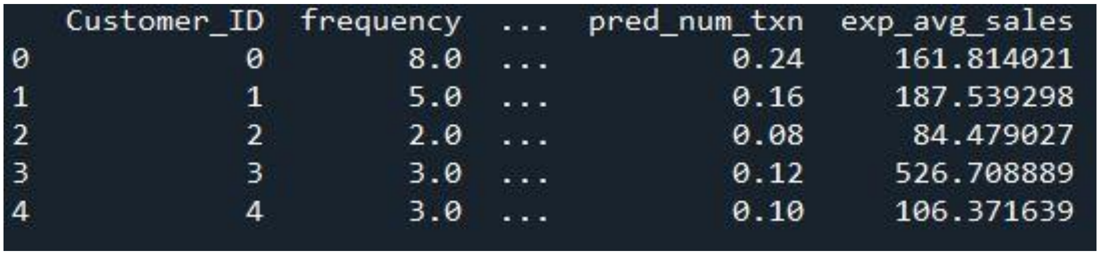
Given how weak this correlation is, our assumptions are satisfied.

Modelling the monetary value using Gamma Gamma,



```
In [48]: ggf.summary
Out[48]:
       coef  se(coef)  lower 95% bound  upper 95% bound
p  9.047676  0.138020         8.777157         9.318195
q  0.935173  0.012432         0.910807         0.959539
v  9.352756  0.147840         9.062988         9.642523
```

Now we predict expected average profit for each transaction



```
   Customer_ID  frequency  ...  pred_num_txn  exp_avg_sales
0            0        8.0  ...          0.24     161.814021
1            1        5.0  ...          0.16     187.539298
2            2        2.0  ...          0.08      84.479027
3            3        3.0  ...          0.12     526.708889
4            4        3.0  ...          0.10     106.371639
```

We also have the following by finding mean of the columns Exp_Avg_Sales and Monetary Value :

Expected Average Sales: **245.96736950892875**

Actual Average Sales: **242.65626685478682**

Now we finally calculate the Customer Lifetime Value by time, frequency and discount rate (Based on Discounted Cash Flow concept)

CONCLUSIONS

Based on this thorough analysis, we need to find customers eligible for gift hampers.

Assumptions being made :

This hamper campaign, we have a budget of $10,000, where each hamper costs $50. So we have 200 Hampers to give.

We will be using the 'expected average sales' column that we've created for each customer predicting their future sales.

In order to retain our profit, we can have expected average sales as low as 250 dollars for the customers. Also, we have observed the surge in sales in the USA, with customers bringing in the most sales from that country, we can choose that particular country to gift hampers, that are also bought from that country, since that reduces our shipping cost.

The customers that have high expected average sales are going to stay retained based on their behavior. So we need to find customers with as low expected sales as possible to incentivize them to stay a customer, but not with as low expected sales as to us not making a profit.

So we find the lowest possible 200 average sales customers where we can still make the profit.

Eventually we've found 200 customers and exported the a file with their Customer IDs called as **Hamper Customers**.

We will be using the Customer Lifetime Value that we have calculated in order to find the customers to be retained. In order to maximize the customers to stay loyal and keep retained, we have to choose the customers with the minimum CLV value. A company with currently 9126 customers have a lot of one time buyers as well. So we need customers that are one time buyers as well as customers in the 0.2 percentile of the CLV value (assuming those are the maximum number of people the company can afford to retain.

Both the files are attached to the report along with the codes.