Retail Sector domain Analysis

- **Team Name :: G3**
- **Team Members:**
  - **Rohith kumar R**
  - **Suresh**
  - **Priyanka**
  - **Anish R**
  - **Shilpa Kate**
  - **Sukanya**
  - **Sneha**

- **Mentor Name:- Mr. Rajashekhar**
  Mr. Himavanth Ila

- **Date :- 28 April 2021**

EXCELR
Raising Excellence

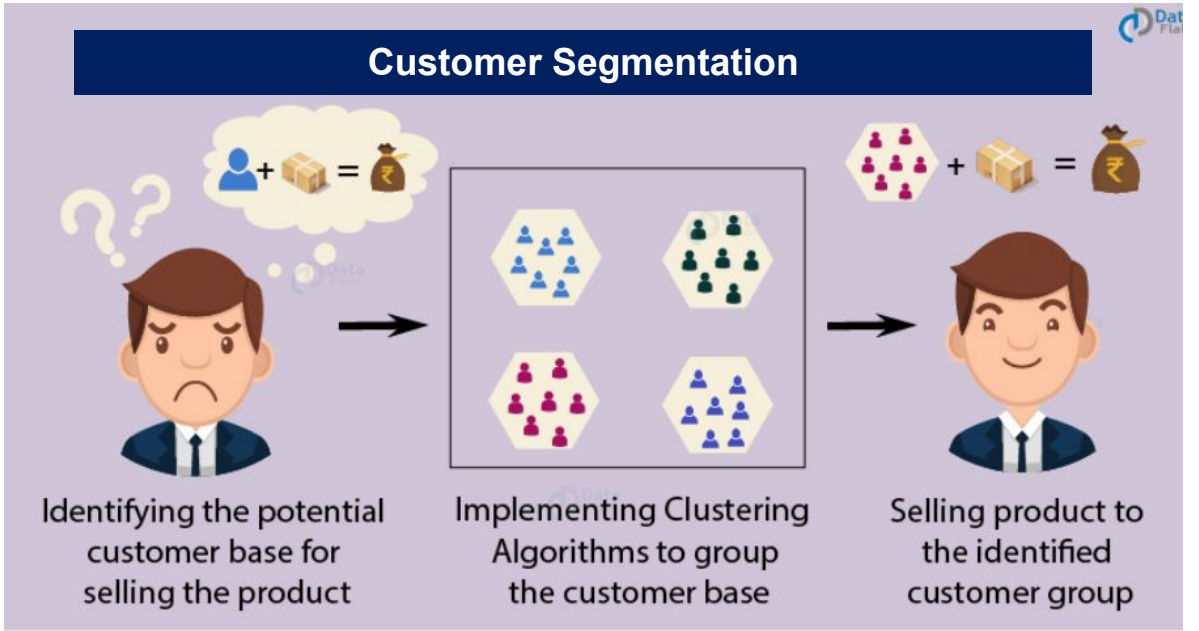
Business Problem:

For a medium to large size retail store, it is also imperative that they invest not only in acquiring new customers but also in customer retention. Many businesses get most of their revenue from their 'best' or high-valued customers. Since the resources that a company has, are limited, it is vital to find these customers and target them.

Objective:

To predict the likelihood of customer shopping next month.

Customer Segmentation



Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

Approaches:

1. Cohort Analysis
2. RFM Analysis

Cohort Analysis

A cohort is a set of users who share similar characteristics over time. Cohort analysis groups the users into mutually exclusive groups and their behaviour is measured over time. It can provide information about product and customer lifecycle.

There are three types of cohort analysis:

Time Cohorts: Using a timestamp to group customers; a good example would be grouping based on the month of first purchase.

Behaviour Cohorts: Here, customers are grouped based on previous activity like products bought or services subscribed to.

Size Cohorts: Categorizing customers based on amount spent over a specific period of time is a good instance of Size Cohorts.

Retention Rate

Retention rate is one of the most important metrics of Cohort Analysis .It is the measure of customer retention or the percentage of people who remain customers after some time.

Higher retention = more recurring paying users.

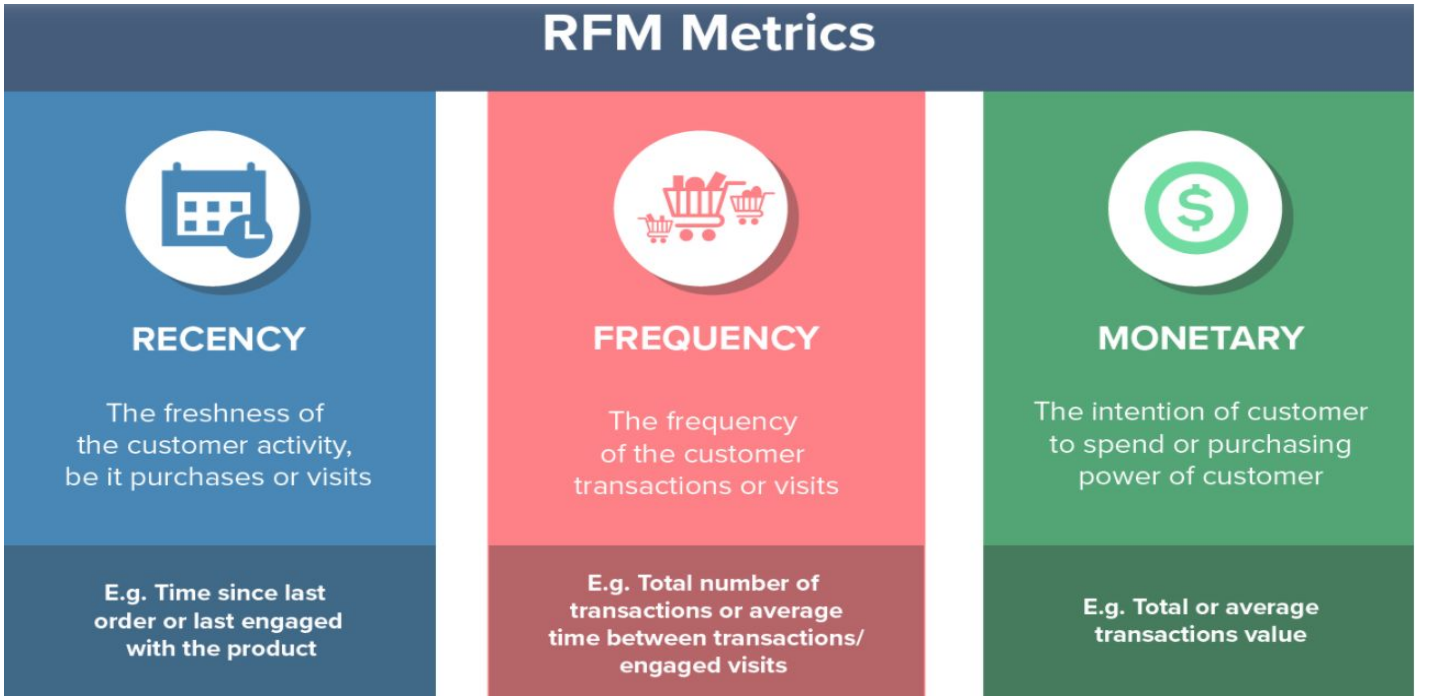
Limitations :-

•Cohort Analysis is a bit limited because it shows only one side of the customers' behaviour; either the number of customers who purchased and their retention, or the revenue per cohort.

•Even though this is very useful and gives great insights, this type of analysis is still one-dimensional.

•In other words, we cannot clearly explore the dependency between number of customers and their purchase value.

Customer RFM Segmentation



RFM segmentation is a technique used to get to know the customers better and to be able to divide them into groups which will make marketing targeting more effective and cost-efficient

Steps to find RFM:-

For Recency:- Calculate the number of days between present date and date of last purchase each customer.
For Frequency:-Calculate the number of orders for each customer.
For Monetary:-Calculate sum of purchase price for each customer.

RFM Analysis

•Choose the number of groups for Recency, Frequency and Monetary values These numbers are often between 3 and 5.
•We can choose the number of groups depending on the type of problem.
• If we choose **4** groups we can have 4x4x4 = 64 combinations ranging from 444 to 111.
• Higher the RFM score, the more valuable the customer.
•In this case the customers are split into quartiles (four equal groups)
• The top 25% a Recency score of 4 (most recent), the next 25% a score of 3 and so on.
•The same process is then undertaken for frequency and monetary.
• Finally, all customers are ranked by concatenating R, F, and M values.

•Taking CLV into account through RFM segmentation is important as it can help you increase customer value and loyalty

Customer Lifetime Value Prediction

RFM segmentation can help you get the data needed to estimate a customer's lifetime value (CLV), which is the monetary estimation of the value your business will derive from your relationship with any given customer.

Approaches to modelling the CLV problem

There are two broad approaches to modelling the *CLV* problem*:*

Historical Approach:

Aggregate Model — calculating the CLV by using the average revenue per customer based on past transactions. This method gives us a single value for the CLV.

Cohort Model — grouping the customers into different cohorts based on the transaction date, etc., and calculate the average revenue per cohort. This method gives CLV value for each cohort.

Predictive Approach:

Machine Learning Model — Using regression techniques to fit on past data to predict the CLV.

Probabilistic Model — It tries to fit a probability distribution to the data and estimates the future count of transactions and monetary value for each transaction.

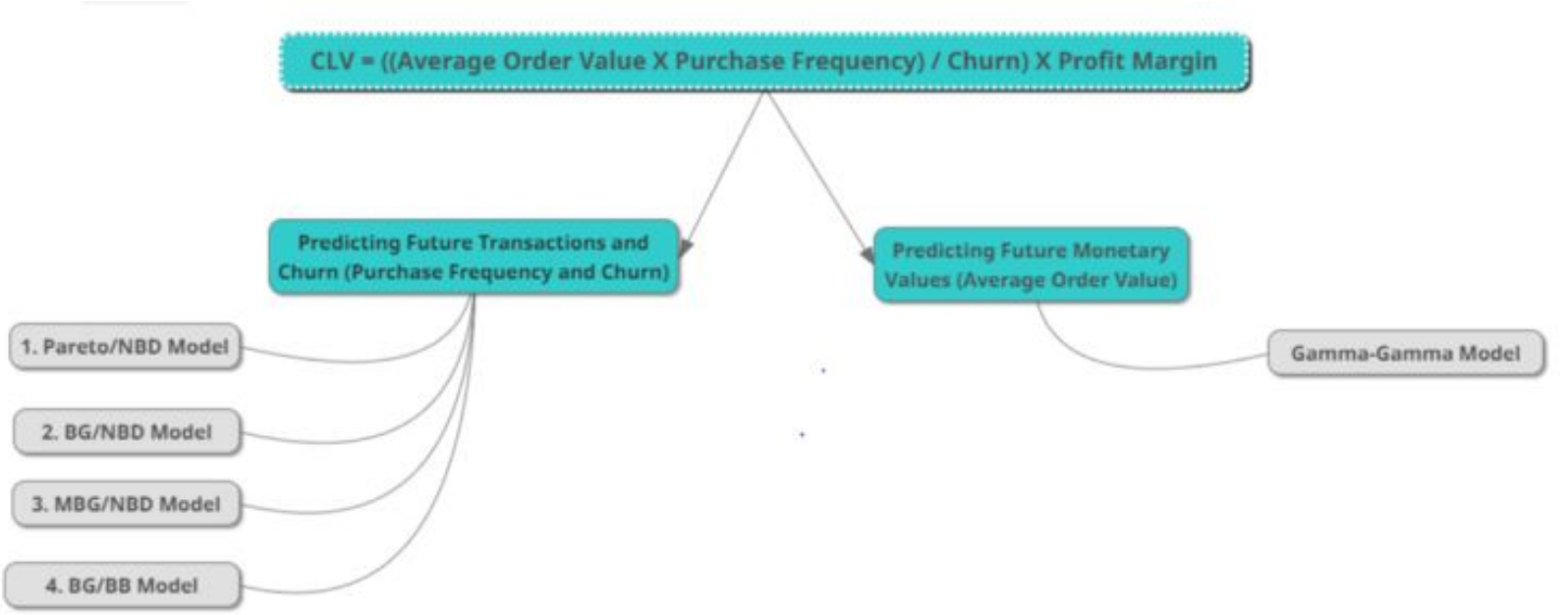
Predictive Approach: predict customer likelihood

Probabilistic models

• BG/NBD stands for Beta Geometric/Negative Binomial Distribution. This is one of the most commonly used probabilistic model for predicting the CLV.

•The BG/NBD and Pareto/NBD model actually tries to predict the future transactions of each customer. It is then combined with the Gamma-Gamma model, which adds the monetary aspect of the customer transaction and we finally get the customer lifetime value (CLV).

•There is a Python package called Lifetimes which makes our life much simpler. This package is primarily built to aid customer lifetime value calculations, predicting customer churn, etc. It has all the major models and utility functions which are needed for CLV calculations.



Predictive Approach:-Machine Learning Model

Building an ML model to predict your customer's lifetime value features certain key steps:

Collect & clean customer data: First, you need to have clean data for each customer. You must have a customer ID that's used to differentiate individual customers and a purchase amount for each purchase that customer has made.

Build a model: Next, you will need to implement ML algorithms to search your dataset for patterns. Once you have a list of patterns, you can design steps to analyze and understand those patterns:
To prepare for training the models, you must **choose a threshold date.**

Check if the model is successful: After all the previous steps, it is extremely important to evaluate that the model is working correctly.

Data set details

EXCELR
Raising Excellence

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Before removing duplicate entries, we had 541909 entries in the dataset.

No of Columns	No of Rows
8	541909

By inspection, we can understand that we have 5268 nos,ie. 1% rows are duplicate entries.

After removing duplicates, we have only 536641 entries in the dataset.

No of Columns	No of Rows
8	536641

Data set details: Data attributes

Attribute	Purpose	Data type
InvoiceNo	Invoice number	Nominal(Object)
StockCode	Product (item) code.	Nominal(Object)
Description	Product (item) name	Nominal(Object)
Quantity	The quantities of each product (item) per transaction	Numeric(int64)
InvoiceDate	the day and time when each transaction was generated. i.e. Invoice Date and time	datetime64
UnitPrice	Product price per unit in sterling.	Numeric(float64)
CustomerID	Customer number	Numeric(float64)
Country	Country name	Nominal(Object)

Data Preprocessing

1.Removing cancelled orders

•As per the data, if the invoice number code starts with the letter ‘c’, it indicates a cancelled order.

•A flag column was created to indicate whether the order corresponds to a cancelled order.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	order_canceled
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	0
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	0
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	0
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	0
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	0

No of orders cancelled ::9251/ 527390(1.72%)

Data Preprocessing

2.To check are there negative quantities against any InvoiceNo

```
1 df.loc[df['Quantity'] < 0,:]
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	order_canceled
2336	536589	21777	NaN	-10	2010-12-01 16:50:00	0.0	NaN	United Kingdom	0
4249	536764	84952C	NaN	-38	2010-12-02 14:42:00	0.0	NaN	United Kingdom	0
7003	536996	22712	NaN	-20	2010-12-03 15:30:00	0.0	NaN	United Kingdom	0
7004	536997	22028	NaN	-20	2010-12-03 15:30:00	0.0	NaN	United Kingdom	0
7005	536998	85067	NaN	-6	2010-12-03 15:30:00	0.0	NaN	United Kingdom	0
...	...	...	...	...	...	...	...	...	...
520928	581210	23395	check	-26	2011-12-07 18:36:00	0.0	NaN	United Kingdom	0
520930	581212	22578	lost	-1050	2011-12-07 18:38:00	0.0	NaN	United Kingdom	0
520931	581213	22576	check	-30	2011-12-07 18:38:00	0.0	NaN	United Kingdom	0
522503	581226	23090	missing	-338	2011-12-08 09:56:00	0.0	NaN	United Kingdom	0
524440	581422	23169	smashed	-235	2011-12-08 15:24:00	0.0	NaN	United Kingdom	0

1336 rows × 9 columns

In above fig CustomerID values are NaNs. These cases were also need to remove from the data.

We can infer that there are 1336 entries with negative values.

Data Preprocessing

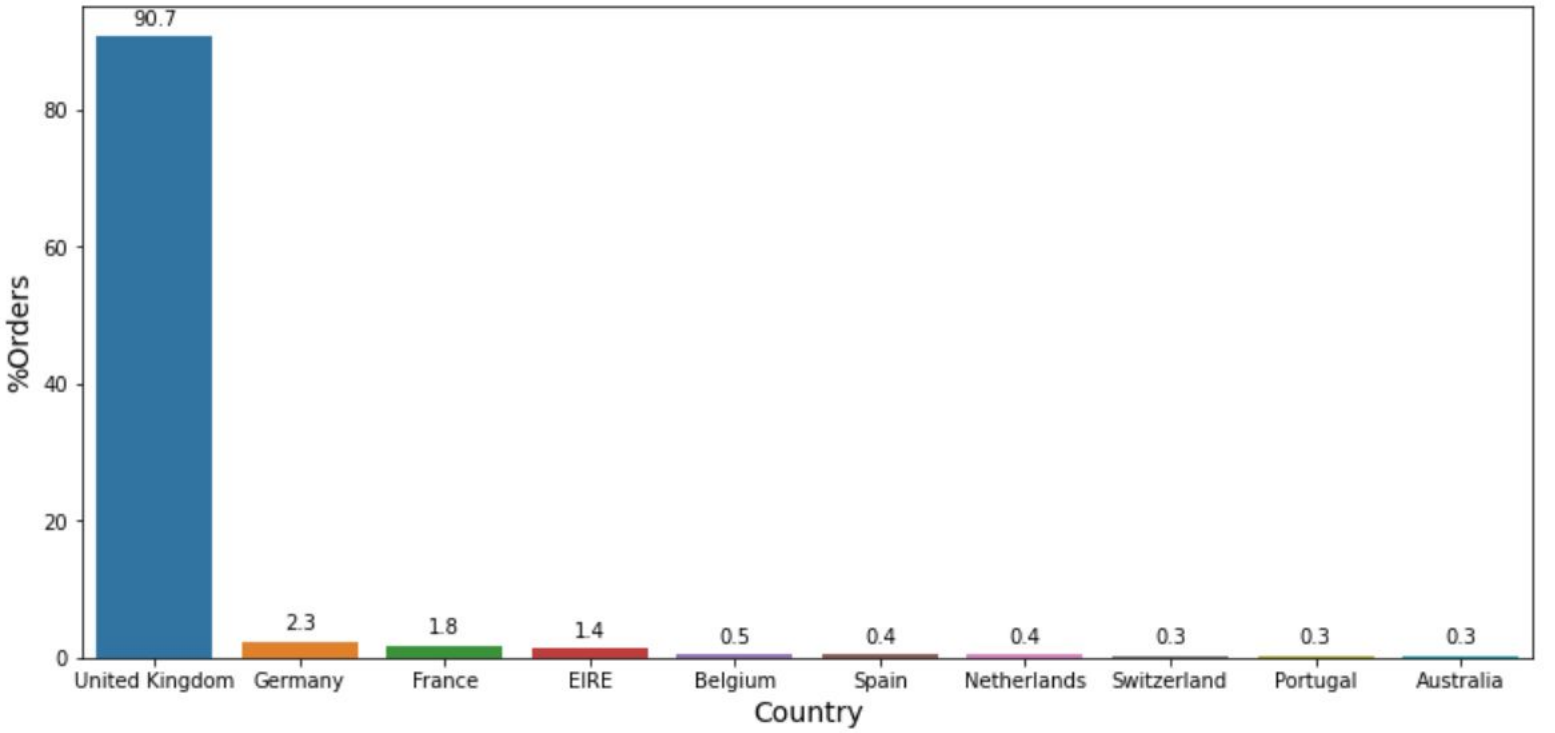
After Preprocessing, the data is cleaned from duplicate entries, cancelled orders and negative quantities which consists of 349227 rows of non-null values.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 349227 entries, 0 to 392716
Data columns (total 9 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   InvoiceNo       349227 non-null  object
 1   StockCode       349227 non-null  object
 2   Description     349227 non-null  object
 3   Quantity        349227 non-null  int64
 4   InvoiceDate     349227 non-null  datetime64[ns]
 5   UnitPrice       349227 non-null  float64
 6   CustomerID      349227 non-null  float64
 7   Country         349227 non-null  object
 8   order_canceled  349227 non-null  int32
dtypes: datetime64[ns](1), float64(2), int32(1), int64(1), object(4)
memory usage: 25.3+ MB
```

The number of products, transactions, and customers in our cleaned data:

	products	transactions	customers
quantity	3665	18536	4339

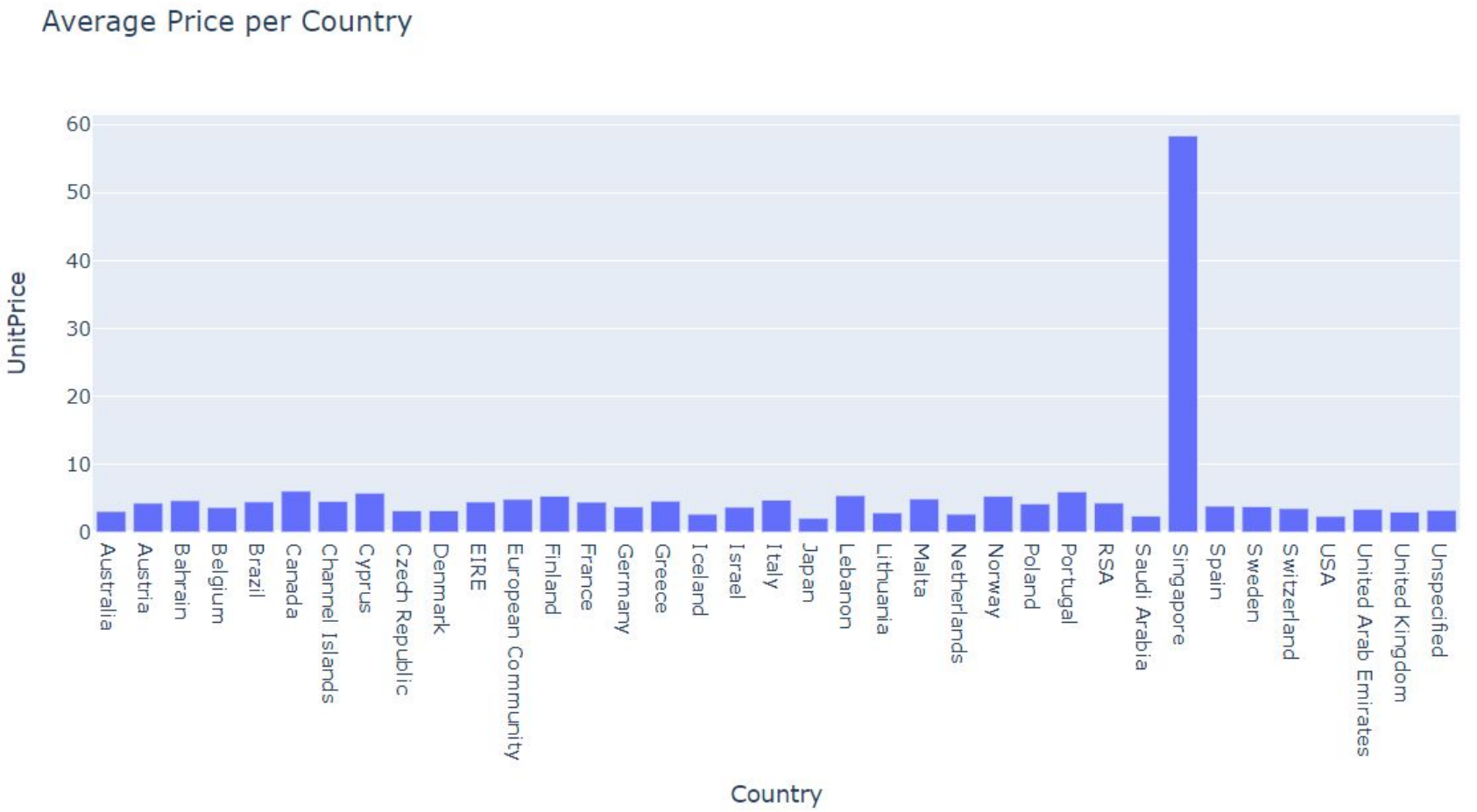
EDA



Percentage of orders from each country in the data.

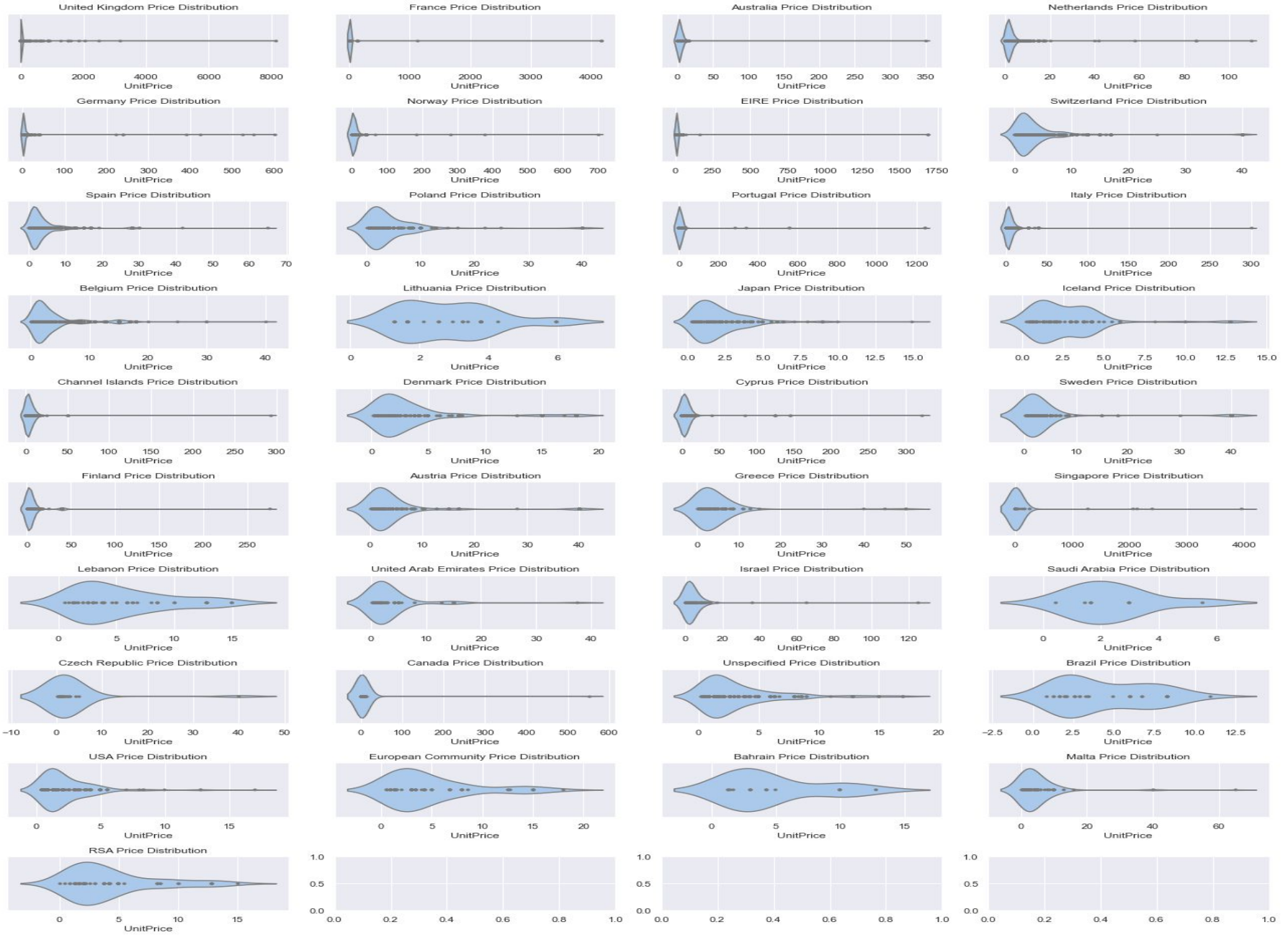
From this graph we can analyse that nearly 91% of the orders are coming from **United Kingdom** followed by **Germany, France & EIRE**.

EDA



From the above plot we can conclude that the **Singapore** has the highest average price followed by the **Portugal and Cyprus**.

EDA

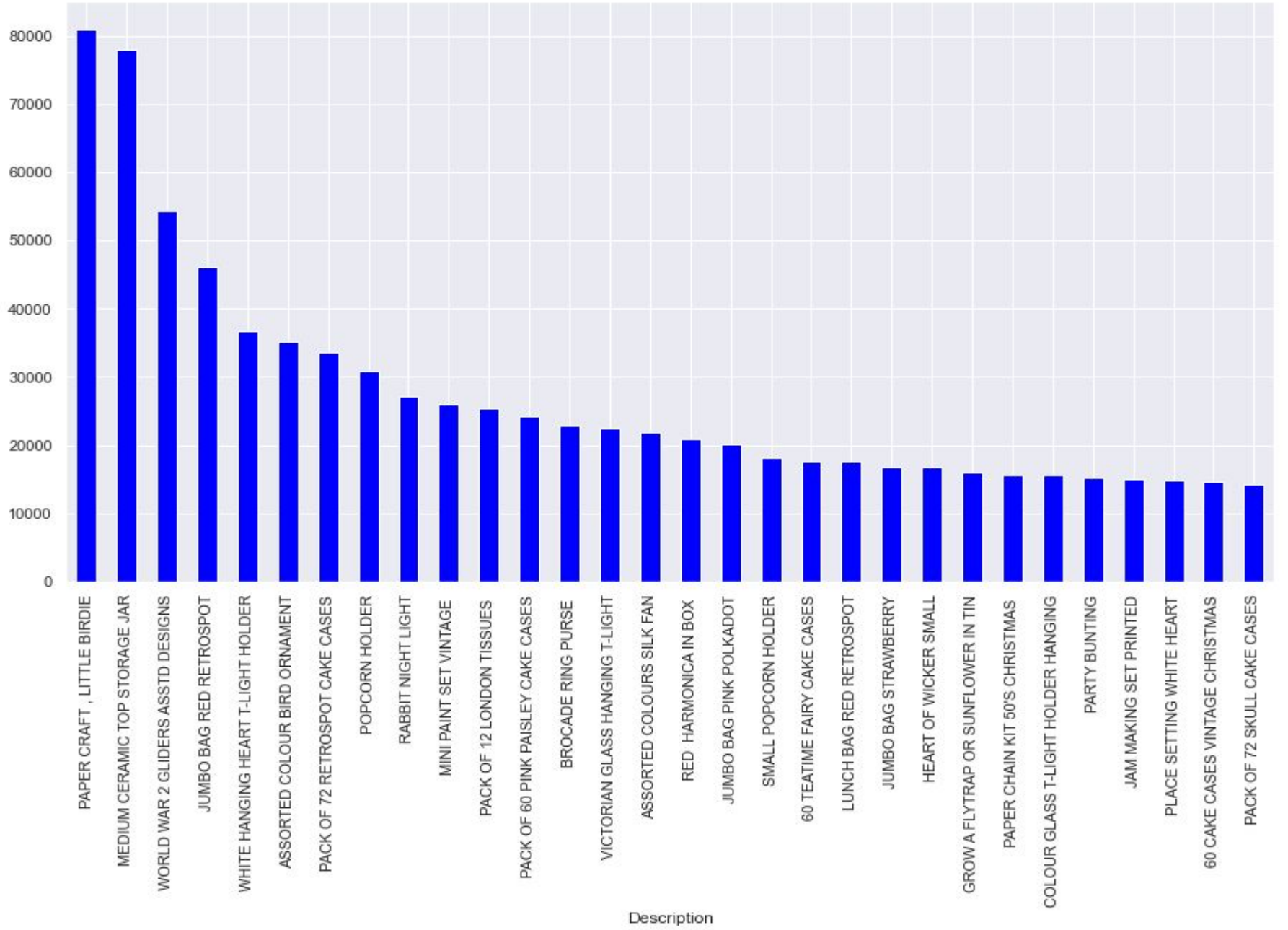

United Kingdom Price Distribution
France Price Distribution
Australia Price Distribution
Netherlands Price Distribution
Germany Price Distribution
Norway Price Distribution
EIRE Price Distribution
Switzerland Price Distribution
Spain Price Distribution
Poland Price Distribution
Portugal Price Distribution
Italy Price Distribution
Belgium Price Distribution
Lithuania Price Distribution
Japan Price Distribution
Iceland Price Distribution
Channel Islands Price Distribution
Denmark Price Distribution
Cyprus Price Distribution
Sweden Price Distribution
Finland Price Distribution
Austria Price Distribution
Greece Price Distribution
Singapore Price Distribution
Lebanon Price Distribution
United Arab Emirates Price Distribution
Israel Price Distribution
Saudi Arabia Price Distribution
Czech Republic Price Distribution
Canada Price Distribution
Unspecified Price Distribution
Brazil Price Distribution
USA Price Distribution
European Community Price Distribution
Bahrain Price Distribution
Malta Price Distribution
RSA Price Distribution
UnitPrice

EDA

The above graph is a Violin Plot to understand the skewness of Price variable for different commodities.

From the plot we can infer that the Price data is skewed and that is due to the fact that there's a lot of high extreme values present in data.

EDA

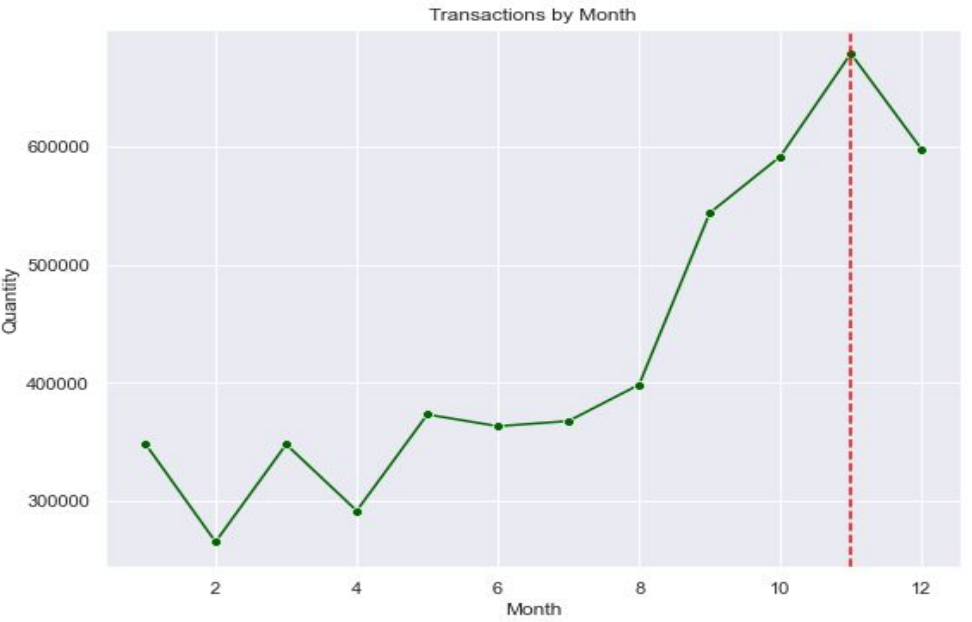

80000
70000
60000
50000
40000
30000
20000
10000
0
PAPER CRAFT , LITTLE BIRDIE
MEDIUM CERAMIC TOP STORAGE JAR
WORLD WAR 2 GLIDERS ASSTD DESIGNS
JUMBO BAG RED RETROSPOT
WHITE HANGING HEART T-LIGHT HOLDER
ASSORTED COLOUR BIRD ORNAMENT
PACK OF 72 RETROSPOT CAKE CASES
POPCORN HOLDER
RABBIT NIGHT LIGHT
MINI PAINT SET VINTAGE
PACK OF 12 LONDON TISSUES
PACK OF 60 PINK PAISLEY CAKE CASES
BROCADE RING PURSE
VICTORIAN GLASS HANGING T-LIGHT
ASSORTED COLOURS SILK FAN
RED HARMONICA IN BOX
JUMBO BAG PINK POLKADOT
SMALL POPCORN HOLDER
60 TEATIME FAIRY CAKE CASES
LUNCH BAG RED RETROSPOT
JUMBO BAG STRAWBERRY
HEART OF WICKER SMALL
GROW A FLYTRAP OR SUNFLOWER IN TIN
PAPER CHAIN KIT 50'S CHRISTMAS
COLOUR GLASS T-LIGHT HOLDER HANGING
PARTY BUNTING
JAM MAKING SET PRINTED
PLACE SETTING WHITE HEART
60 CAKE CASES VINTAGE CHRISTMAS
PACK OF 72 SKULL CAKE CASES
Description

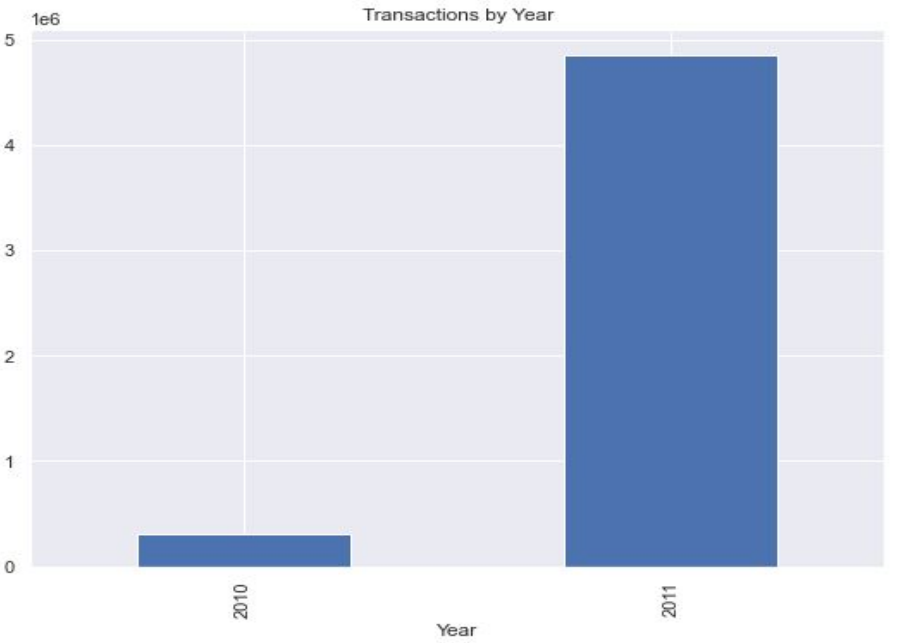
EDA

From the above plot we can analyse that some of the most sold commodities are:

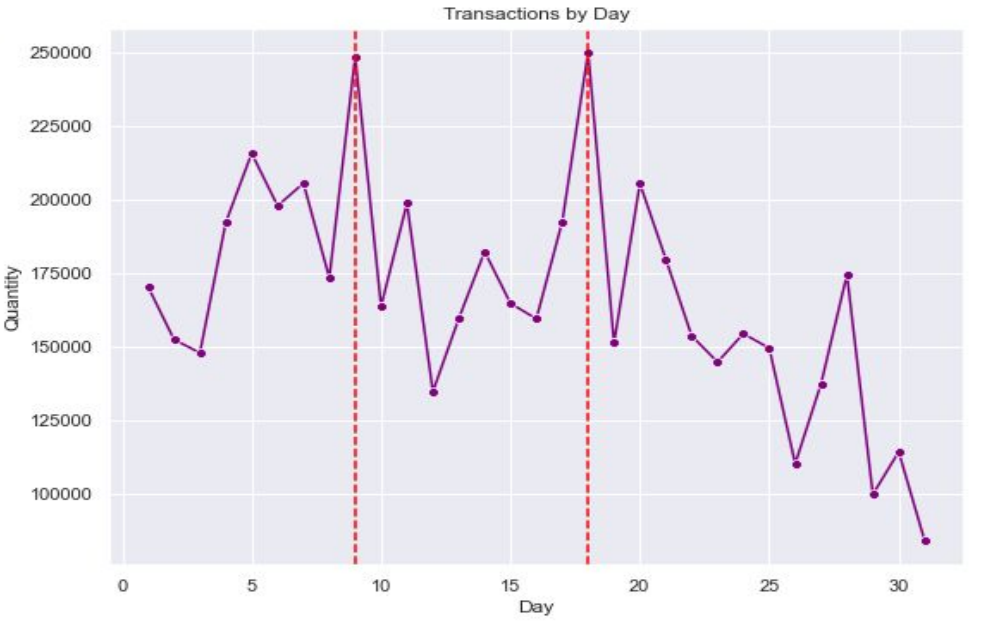
1. Paper Craft Little Birdie
2. Medium Ceramic Top Storage Jar
3. World War 2 Gliders
4. Jumbo Bag Red
5. White Hanging Heart

EDA


Transactions by Month
Quantity
600000
500000
400000
300000
2
4
6
8
10
12
Month


Transactions by Year
1e6
5
4
3
2
1
0
2010
2011
Year


Transactions by Quarter
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1
2
3
4
Quarter


Transactions by Day
Quantity
250000
225000
200000
175000
150000
125000
100000
0
5
10
15
20
25
30
Day

EDA

From the time plot we can infer that

- Most Transactions happened in the month of November which is evident due to festive seasons.
- Q4 being the highest when it comes transactions.
- It also observed that in the end of the 1st week and starting of the 3rd week, people tends to buy more.

Customer Segmentation

- We use RFM clustering technique is used for customer segmentation. In RFM analysis, we segregate the total into various segments so that it will be easier for us to monitor each group.

Theoretically we will segments like:

- Low Value: Customers who are less active than others, not very frequent buyer/visitor and generates very low - zero - maybe negative revenue.

- Mid Value: In the middle of everything.Fairly frequent and generates moderate revenue.

- High Value: The group we don't want to lose. High Revenue, Frequency and low Inactivity.

RFM Analysis - Receny

To calculate recency, we need to find out most recent purchase date of each customer and see how many days they are inactive for. After having no. of inactive days for each customer, we will apply K-means clustering to assign customers a recency score.

The table below gives us the Recency of each customers.

	CustomerID	MaxPurchaseDate	Recency
0	12346.0	2011-01-18 10:17:00	325
1	12747.0	2011-12-07 14:34:00	1
2	12748.0	2011-12-09 12:20:00	0
3	12749.0	2011-12-06 09:56:00	3
4	12820.0	2011-12-06 15:12:00	2
...	...	...	...
3945	18280.0	2011-03-07 09:52:00	277
3946	18281.0	2011-06-12 10:53:00	180
3947	18282.0	2011-12-02 11:43:00	7
3948	18283.0	2011-12-06 12:02:00	3
3949	18287.0	2011-10-28 09:29:00	42

3950 rows × 3 columns

We can infer that the CustomerID 12346 has been away from the store for 325 days and that his recency is very low, whereas CustomerID 12747 made a purchase one day before and that his recent is very high.

RFM Analysis - Receny

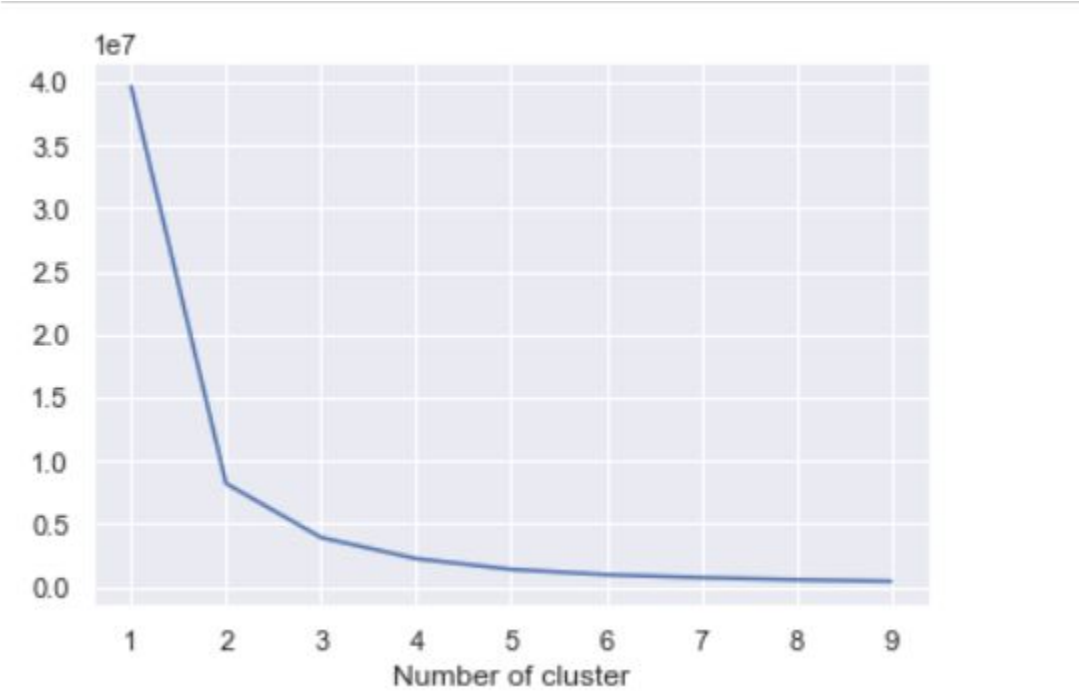
Descriptive Statistics for Recency

```
count    3950.000000
mean       90.778481
std       100.230349
min         0.000000
25%        16.000000
50%        49.000000
75%       142.000000
max       373.000000
Name: Recency, dtype: float64
```

We see that even though the average is 91 day recency, median is 49.

RFM Analysis - Receny

No. of Clusters using K-Means



Here it looks like 3 is the optimal one. Based on business requirements, we can go ahead with less or more clusters. We will be selecting 4 as the no. of clusters.

RFM Analysis - Receny

By applying k = 4, we can obtain the below Recency Cluster Table.

RecencyCluster	count	mean	std	min	25%	50%	75%	max
0	478.0	304.393305	41.183489	245.0	266.25	300.0	336.00	373.0
1	568.0	184.625000	31.753602	132.0	156.75	184.0	211.25	244.0
2	954.0	77.679245	22.850898	48.0	59.00	72.5	93.00	131.0
3	1950.0	17.488205	13.237058	0.0	6.00	16.0	28.00	47.0

From the table we can infer that the cluster 0 has the most inactive customers who had not turned upon for very long time whereas cluster 3 has the most active customers.

RFM Analysis - Frequency

To create frequency clusters, we need to find total number orders for each customer. We can achieve this by counting the Invoice Dates each CustomerID has done the transactions at the store.

	CustomerID	Recency	RecencyCluster	Frequency
0	17850.0	301	0	309
1	15100.0	329	0	6
2	18074.0	373	0	13
3	16250.0	260	0	24
4	13747.0	373	0	1
5	17908.0	373	0	54
6	16583.0	373	0	14
7	18085.0	329	0	29
8	17968.0	373	0	81
9	14729.0	373	0	71

From the table above, we can find out that the CustomerID 17850 has visited 309 times whereas CustomerID 13747 has visited only 1 time.

RFM Analysis - Frequency

By applying K-Means clustering with k = 4, we can obtain the below Frequency Cluster Table.

FrequencyCluster	count	mean	std	min	25%	50%	75%	max
0	3497.0	48.930798	44.240913	1.0	15.0	33.0	72.00	187.0
1	428.0	326.294393	131.375687	188.0	222.0	283.5	394.00	789.0
2	22.0	1298.363636	506.645320	849.0	968.0	1126.5	1429.25	2759.0
3	3.0	5799.666667	1774.543415	4459.0	4793.5	5128.0	6470.00	7812.0

From the table we can infer that the cluster 0 has the most inactive customers who are not frequent visitors of the store whereas cluster 3 has the most active customers who frequently visit the store.

RFM Analysis - Monetary

To find out the Monetary aspect, we have go calculate the revenue generated for each customer. For that, we have to multiply the total quantity each customer buys with the respective unit prices. Then apply the same clustering method.

	CustomerID	Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue
0	17850.0	301	0	309	1	5303.48
1	15808.0	305	0	197	1	3641.07
2	13047.0	31	3	196	1	3079.10
3	14688.0	7	3	356	1	5055.61
4	16029.0	38	3	270	1	53168.69
5	13408.0	1	3	501	1	27487.41
6	13767.0	1	3	399	1	16945.71
7	13448.0	16	3	198	1	3460.99
8	15513.0	30	3	314	1	14520.08
9	17920.0	3	3	665	1	4119.50

The above table shows how much revenue the store generated from individual customers throughout the transaction time.

RFM Analysis - Monetary

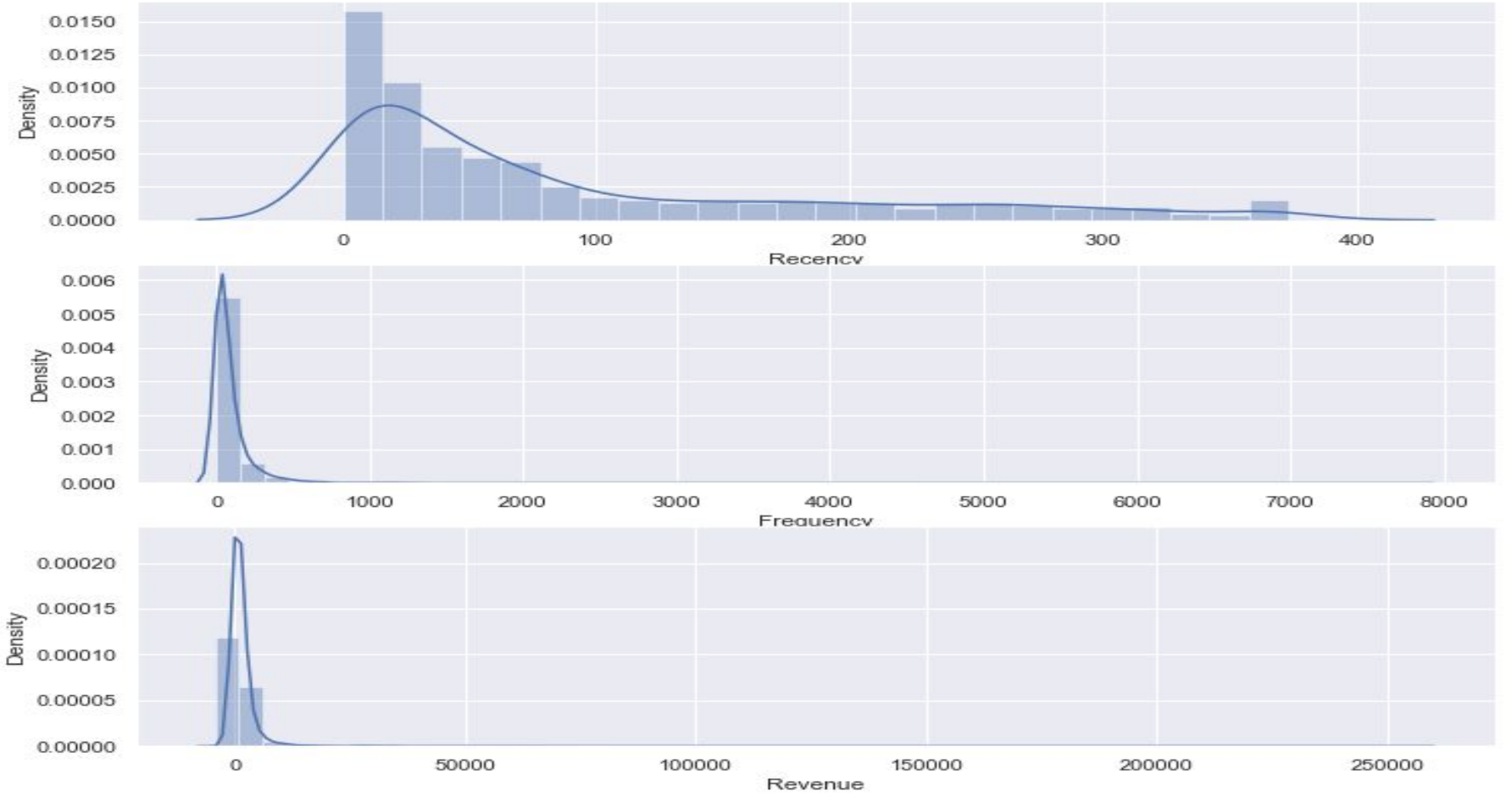
By applying K-Means clustering with k = 4, we can obtain the below Revenue Cluster Table.

RevenueCluster	count	mean	std	min	25%	50%	75%	max
0	3757.0	975.767643	1054.256714	-4287.63	266.5900	584.07	1313.66	5367.80
1	172.0	9838.626744	5295.787730	5411.87	6376.4225	7798.14	10942.97	28658.88
2	19.0	49746.982105	14576.171999	31300.08	35079.2800	51823.72	57221.52	88125.38
3	2.0	221880.330000	48872.618563	187322.17	204601.2500	221880.33	239159.41	256438.49

From the table we can infer that the cluster 0 has the customers with least revenue generated whereas cluster 3 has the customers with maximum revenue generated.

RFM Analysis

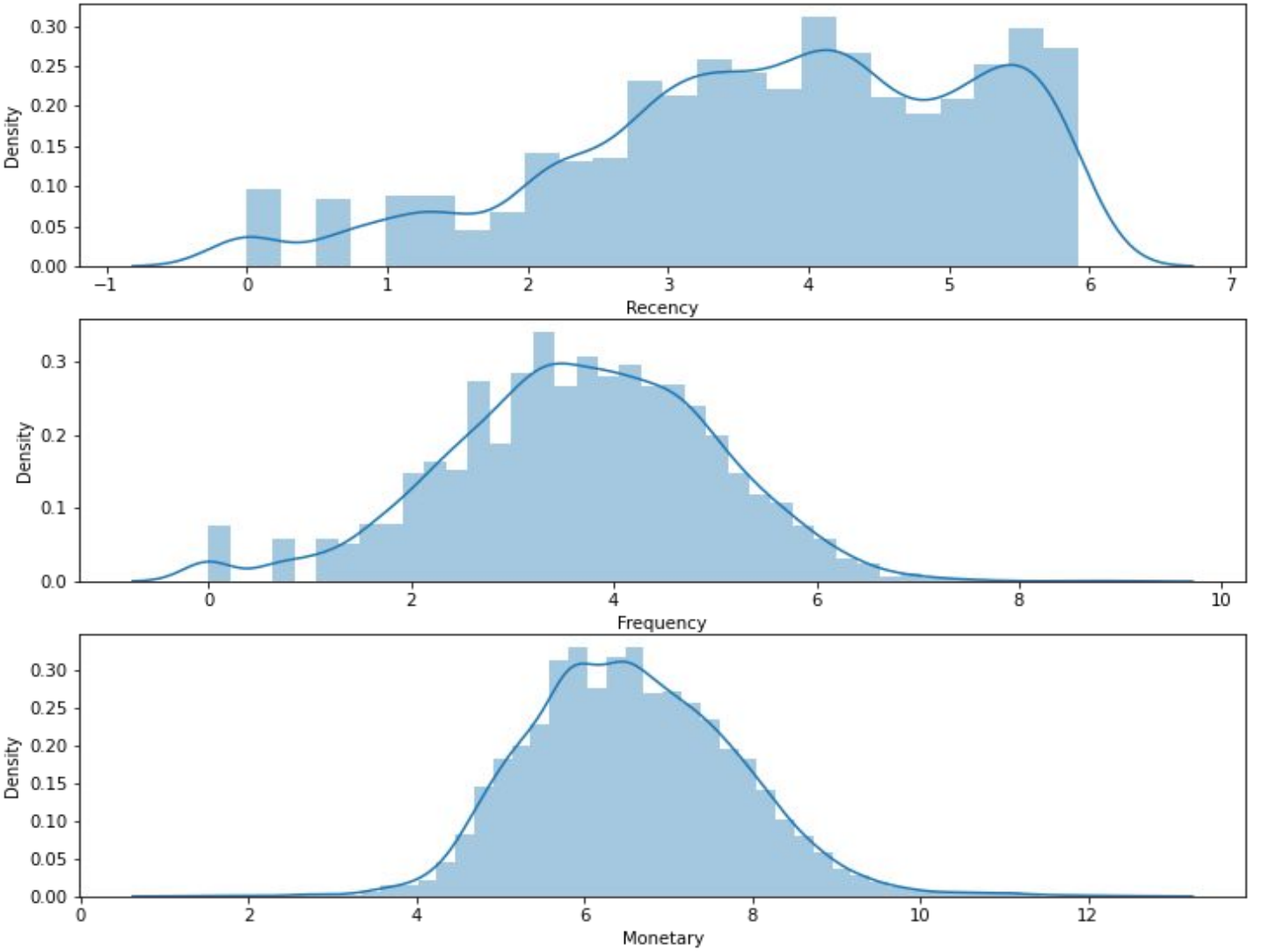
Checking the distribution of Recency, Frequency and Monetary Value variables for skewness.



From the graph we can evaluate that Recency, Frequency and Monetary variables are positively skewed. This may be due to some high extreme values in the data.

RFM Analysis

For removing skewness we can use “log transformation technique”.



So we have removed skewness from RFM data for further analysis.

RFM Analysis - Overall Score

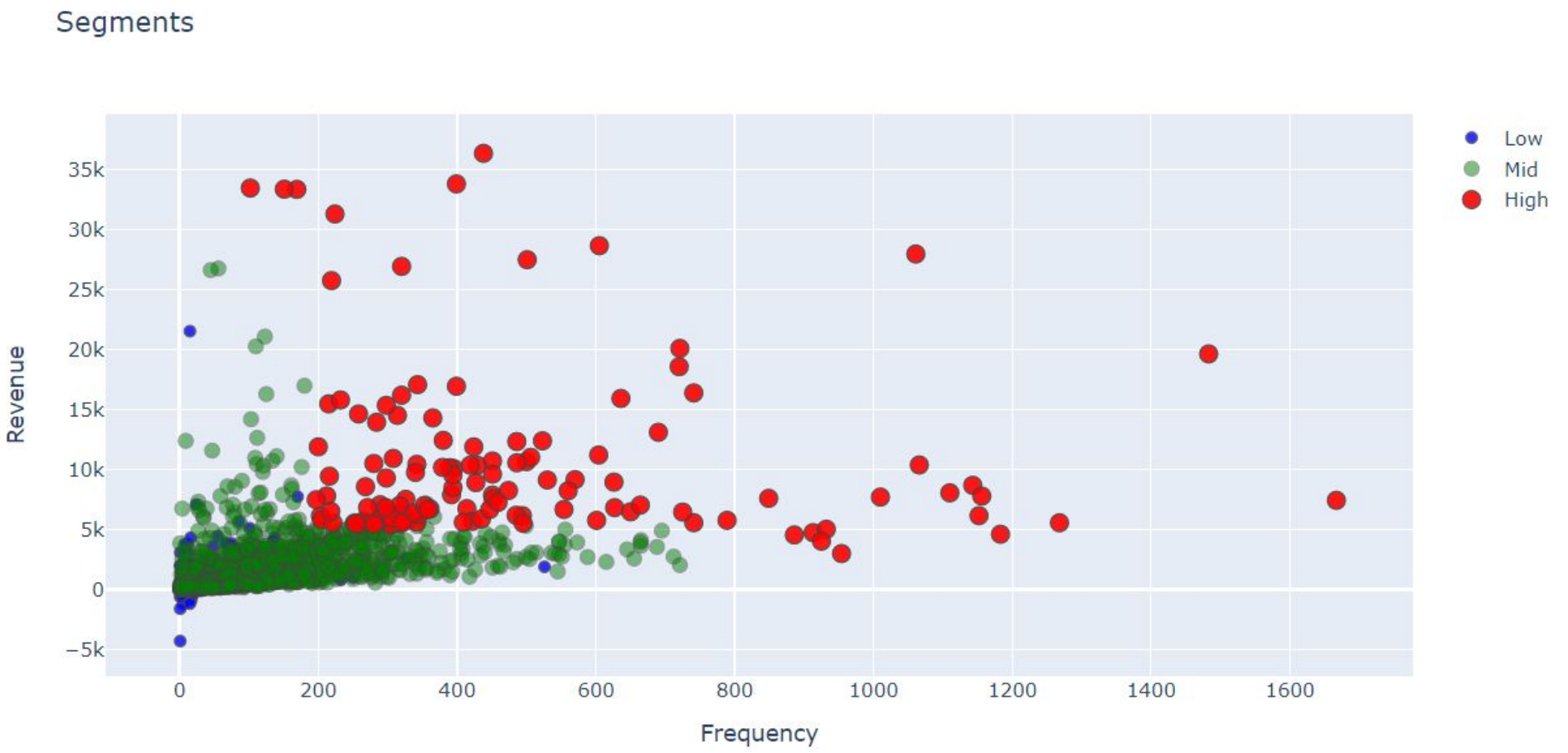
We have scores (cluster numbers) for recency, frequency & revenue. So we can create an overall score out of them.

OverallScore	Recency	Frequency	Revenue
1	5.422563	1.548658	4.738050
2	5.076672	2.512788	5.282074
3	4.470487	2.915835	5.730313
4	4.003375	3.430021	6.239459
5	3.622591	3.958655	6.681170
6	3.020159	4.336624	7.059050
7	2.672204	4.846097	7.638607
8	2.083681	5.401597	8.086950
9	0.781346	5.734829	8.531679

The table above clearly shows that customers with score 8 is our best customers whereas 0 is the worst. To keep things simple, better we name these scores:

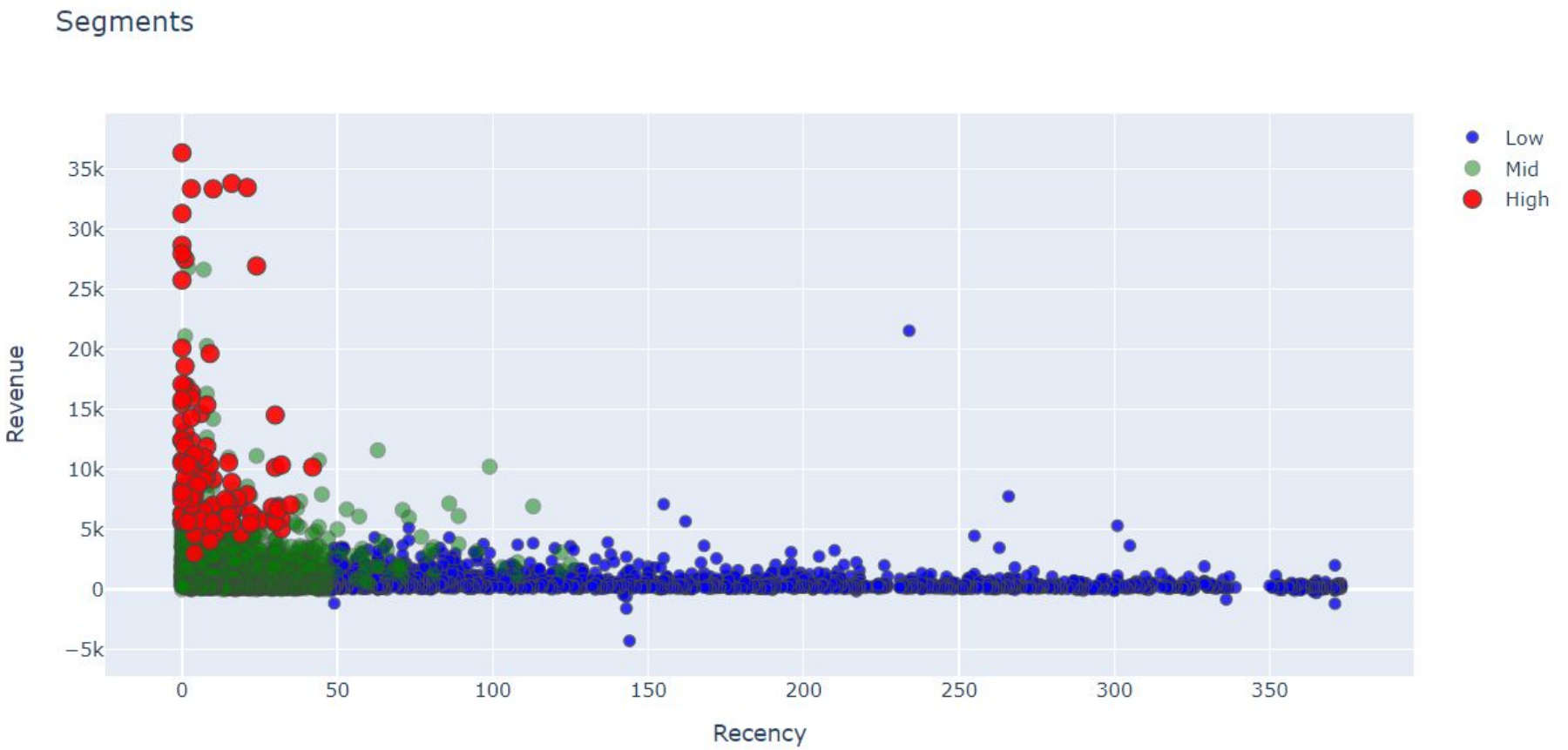
0 to 2: Low Value
3 to 5: Mid Value
5+: High Value

RFM Analysis - Overall Score



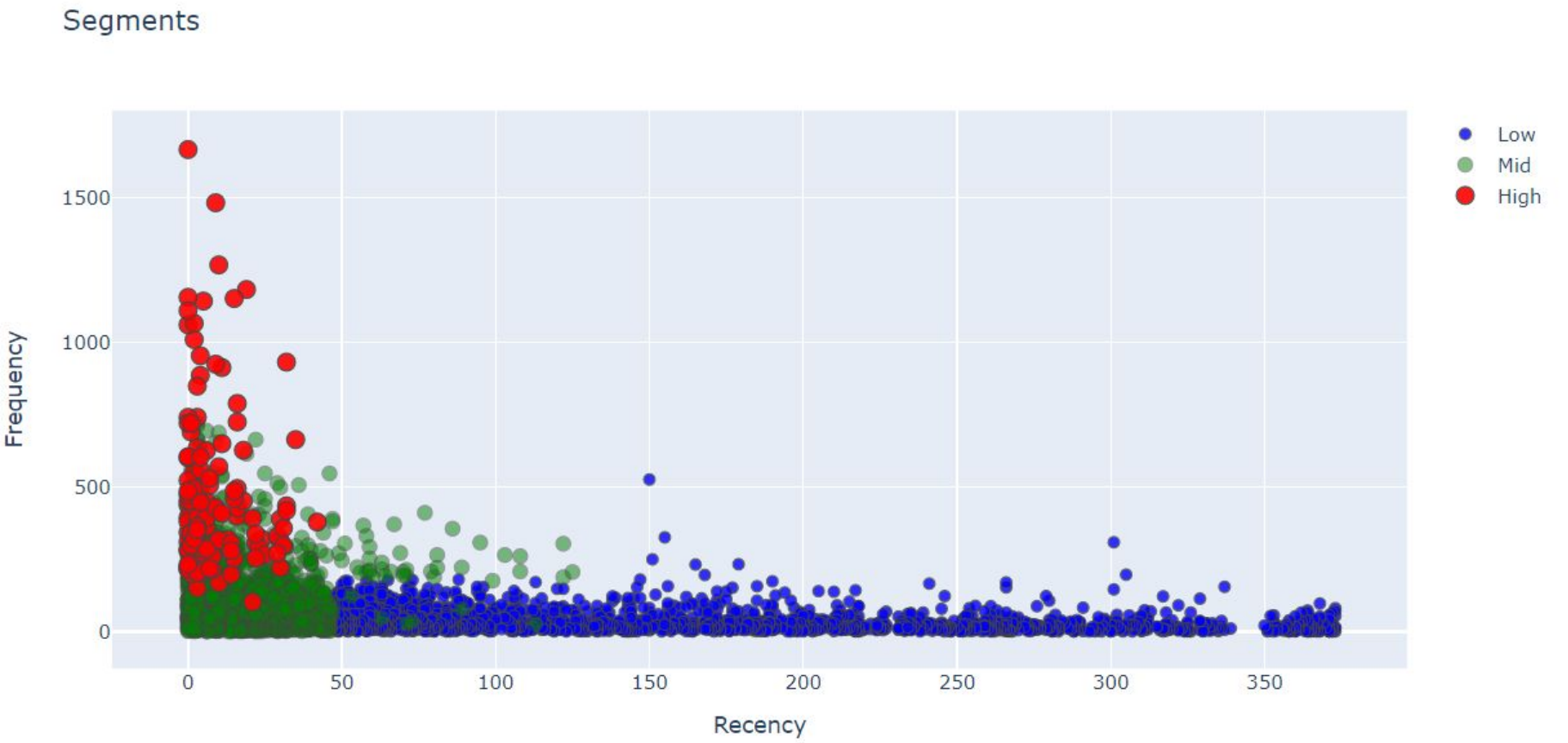
This is the plot of Revenue vs Frequency. We can see that as the Frequency increases the Revenue also increases.

RFM Analysis - Overall Score



This is the plot of Revenue vs Recency. We can see that as the Recency increases the Revenue decreases.

RFM Analysis - Overall Score



This is the plot of Recency vs Frequency. We can see that as the Recency increases the Frequency decreases.

RFM Analysis - Overall Score

From the RFM analysis of the dataset, we can strategize our action plan to reduce the customer churn out for the organization.

The main strategies we can take based on the RFM scores are:

High RFM Value: Improve Retention

Mid RFM Value: Improve Retention + Increase Frequency

Low RFM Value: Increase Frequency

Model Evaluation

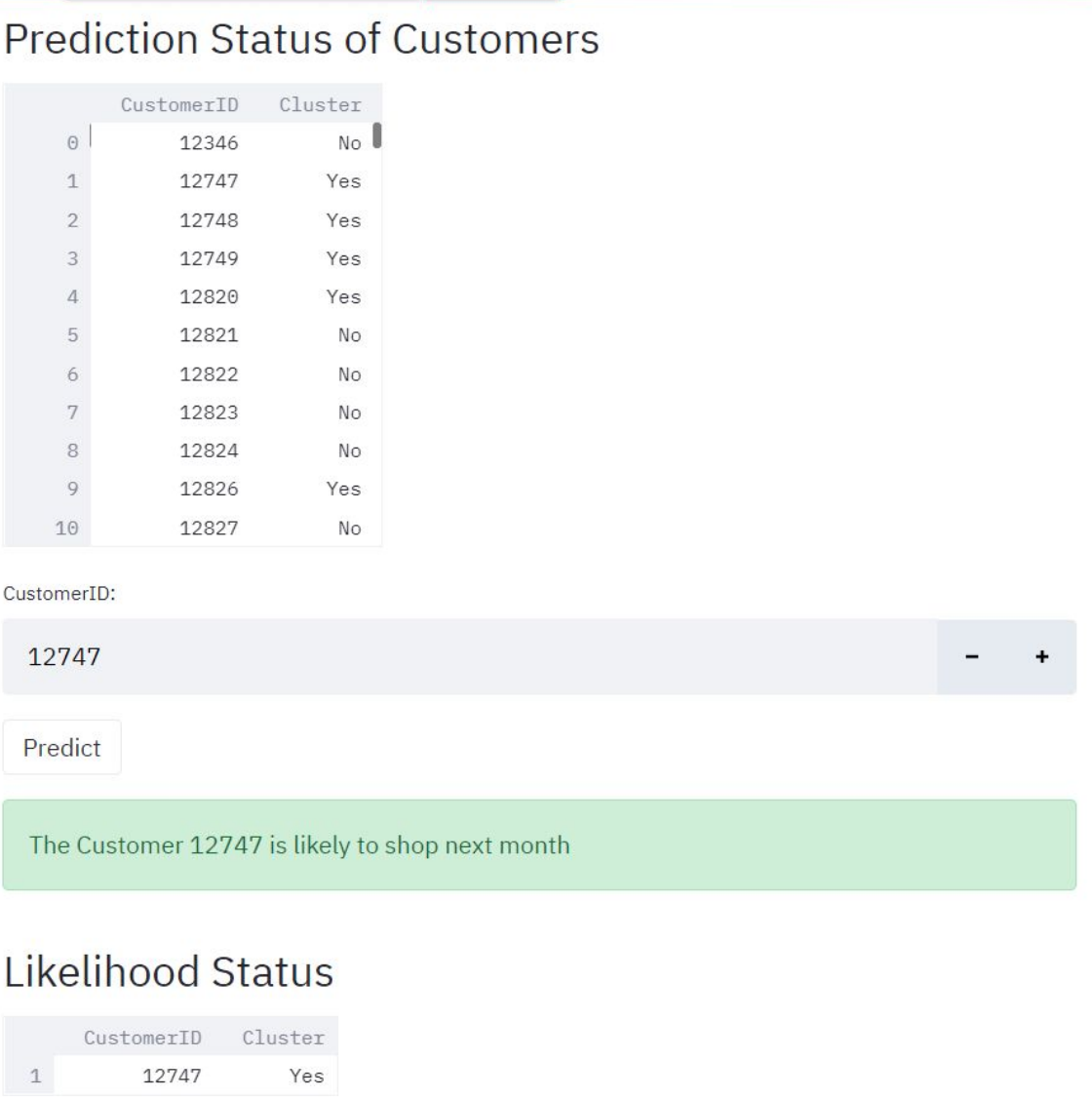
From the RFM analysis of the dataset, we can strategize our action plan to reduce the customer churn out for the organization.

```
LogReg| Mean=0.939000 STD=0.042532
DecTree| Mean=0.913000 STD=0.051778
NB| Mean=0.927000 STD=0.045177
SVM| Mean=0.932000 STD=0.050951
KNN| Mean=0.926000 STD=0.044317
RF| Mean=0.942000 STD=0.041665
XGBOOST| Mean=0.944000 STD=0.038000
```

From the table we can conclude that XGBoost model has little more accuracy compared to other models.

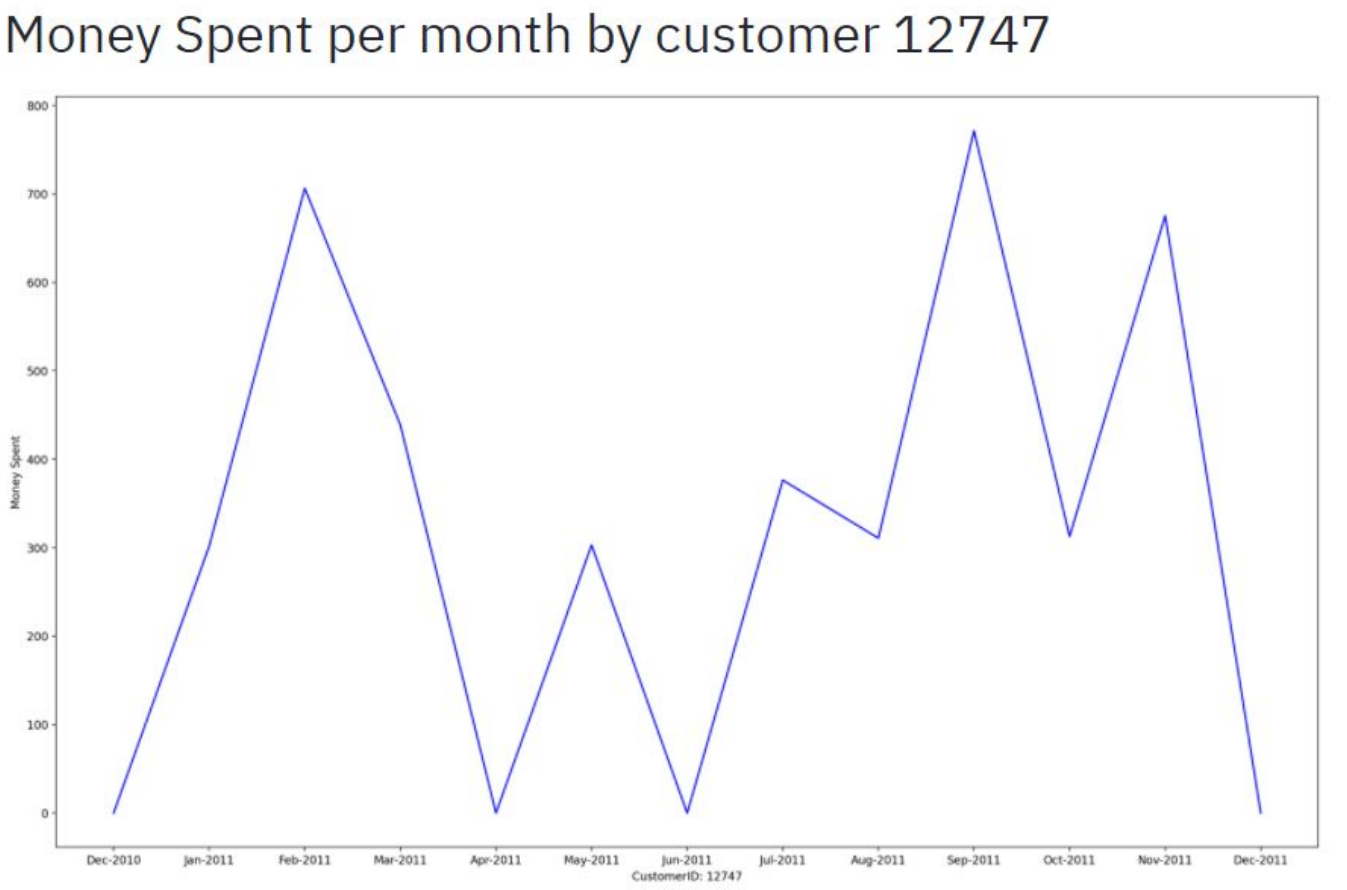
Deployment

We used Streamlit for the deployment part. The final output page is:



Deployment

The graph below explains the amount spent by a particular customer per month.



Thank you