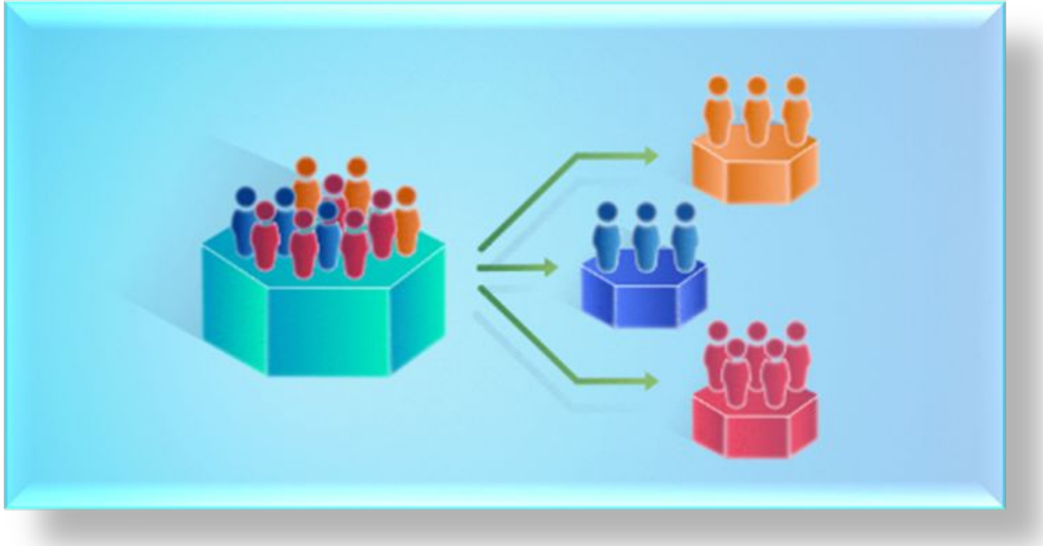
Capstone Project 4

Online Retail Customer Segmentation

Unsupervised ML - Clustering



Sayali Kamalapurkar

Points To Discuss

Problem Statement

In this part, we are going to understand the problem statement in terms of business context. Further we will discuss about the dataset we have.

Data Wrangling

In this section, we will be discussing about the how we have dealt with null and duplicate values in the data, also what could be done with outliers if present.

EDA on features

Before we start EDA on features, we have extracted some new meaningful features from the existing columns, which is called as feature engineering. Further, we will answer some questions through EDA.

Clustering Models

Here, we have built RFM model and we will be discussing about the K-Means clustering algorithm. For this we need to identify number of clusters beforehand, which could be done by using Elbow method and Silhouette score. We have also done cohort analysis that is customer behaviour over time.

Conclusions

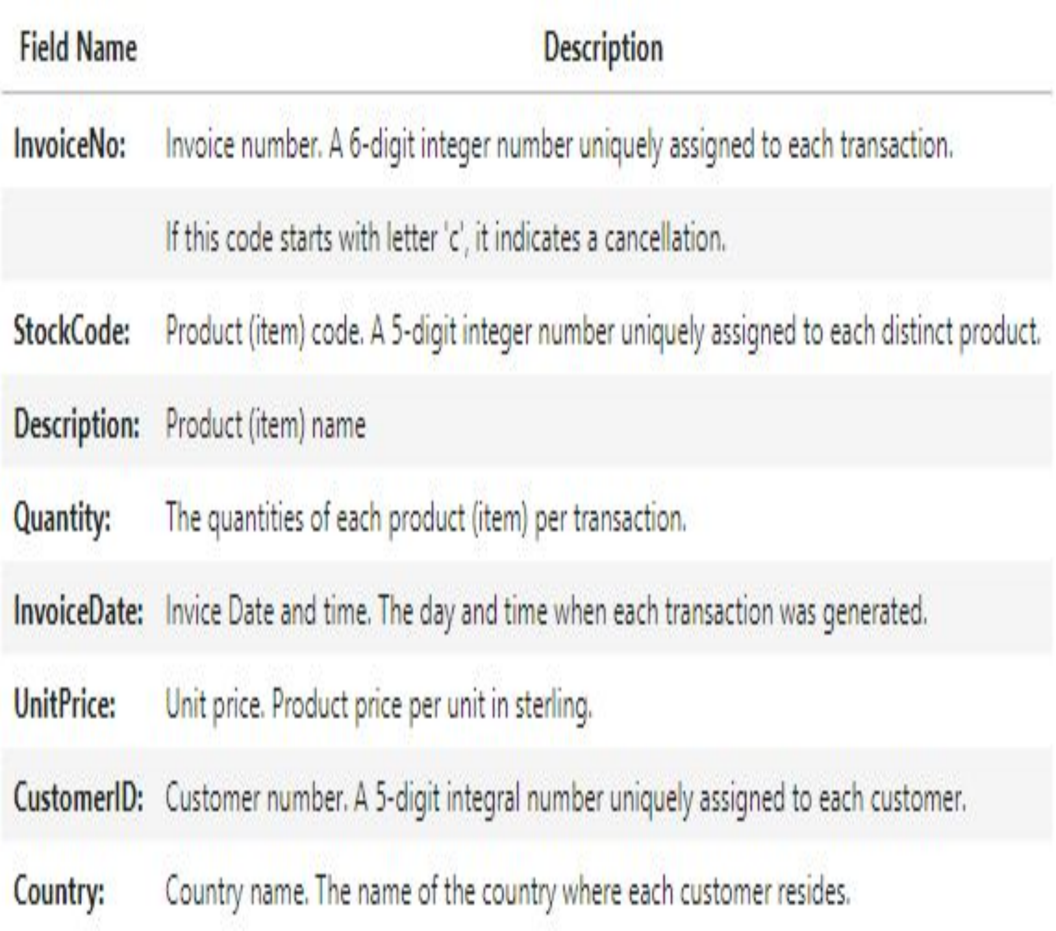
Finally, we will be discussing the conclusions drawn from the whole process. Further, we will discuss what benefits company will get through this clustering.

Problem Statement

Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

In this project, your task is to identify major customer segments on a transactional data set which contains all the transactions 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.

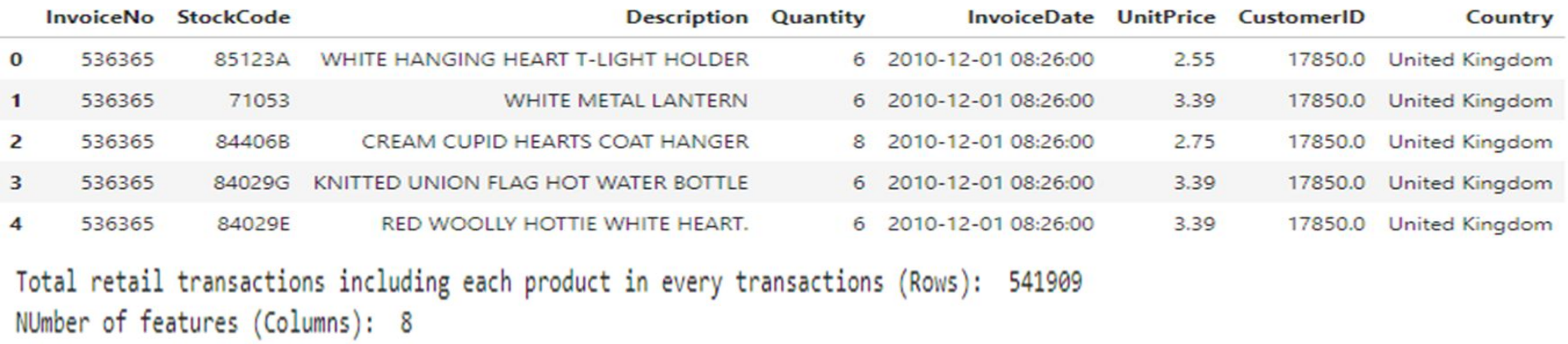
Data Summary



Field Name	Description
InvoiceNo:	Invoice number. A 6-digit integer number uniquely assigned to each transaction.
	If this code starts with letter 'c', it indicates a cancellation.
StockCode:	Product (item) code. A 5-digit integer number uniquely assigned to each distinct product.
Description:	Product (item) name
Quantity:	The quantities of each product (item) per transaction.
InvoiceDate:	Invice Date and time. The day and time when each transaction was generated.
UnitPrice:	Unit price. Product price per unit in sterling.
CustomerID:	Customer number. A 5-digit integral number uniquely assigned to each customer.
Country:	Country name. The name of the country where each customer resides.

Data Summary

Let us have a look at the dataset we have:



	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

```
Total retail transactions including each product in every transactions (Rows):  541909
NUmber of features (Columns):  8
```

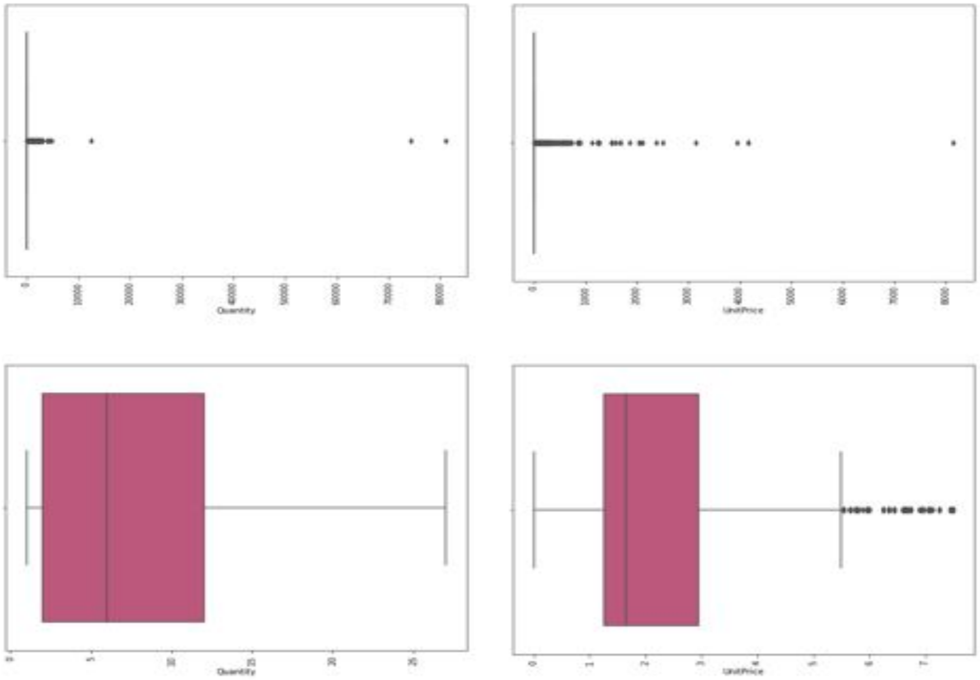
- **For 4070 products, there are 25900 transactions in the data. This means that each product is likely to have multiple transactions in the data. There are almost as many products as customers in the data as well.**
- **In the dataset, we have retail transactions from almost 38 different countries of the world.**
- **Almost 25900 overall transactions, this includes cancelled transactions too.**
- **Also, 4223 different items are listed in the dataset.**
- **Overall customer count in the available dataset is 4372.**

Data Wrangling

Dealing with null and duplicate values:

- There are 135080 null values in CustomerID column and 5268 duplicate values in the dataset.
- We have dropped duplicate rows and also rows containing null values in CustomerID column.

Outlier Treatment:



We have removed all outliers from Quantity and Unit Price columns.

Observations from boxplot:

- **Median of Quantity is 6 and most of the quantities lie between 2 to 12.**
- **50th percentile for unit price is approx 1.75, and most of the unit prices are between 1.25 to 3.**

Feature Engineering

As we have very less features about the customers, we created some new features from existing features.

They are as follows –

- Month, day_name, Day, Hour – all these new features derived from InvoiceDate column.

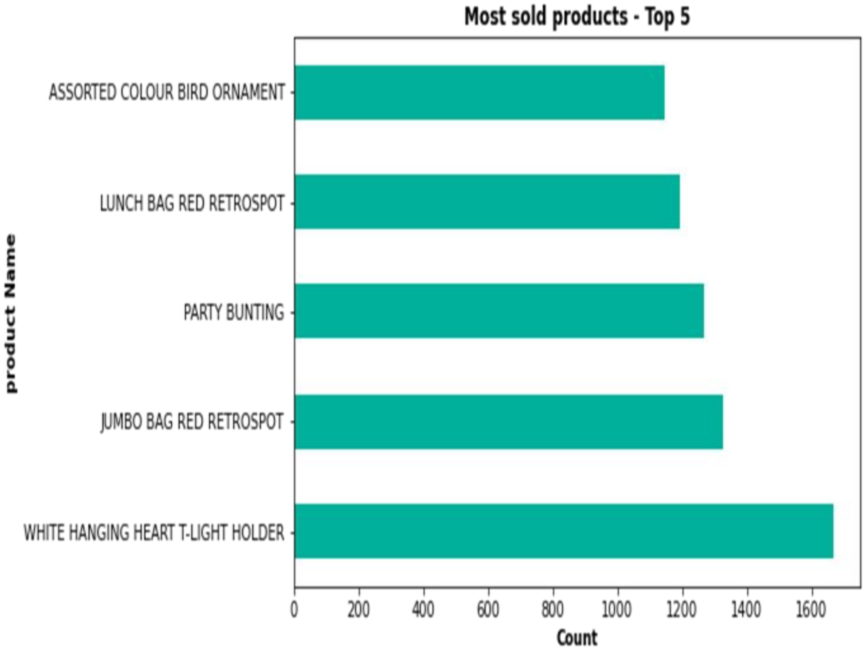
- Day_time – from Hour of the day, like Morning, Afternoon and Evening

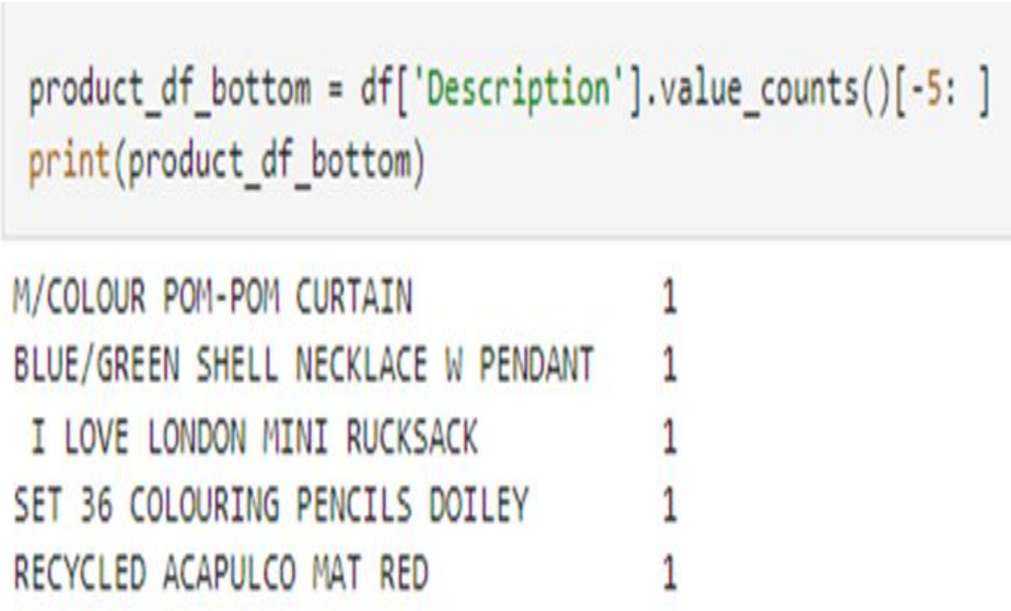
- Sales_Amount – from Quantity and UnitPrice columns by multiplying them to get total sale price for that particular row in the data.

- Converted datatype of CustomerID column from float to integer, as it is just an identification number of the customers.

EDA on features

Which products are the most and least sold ones?



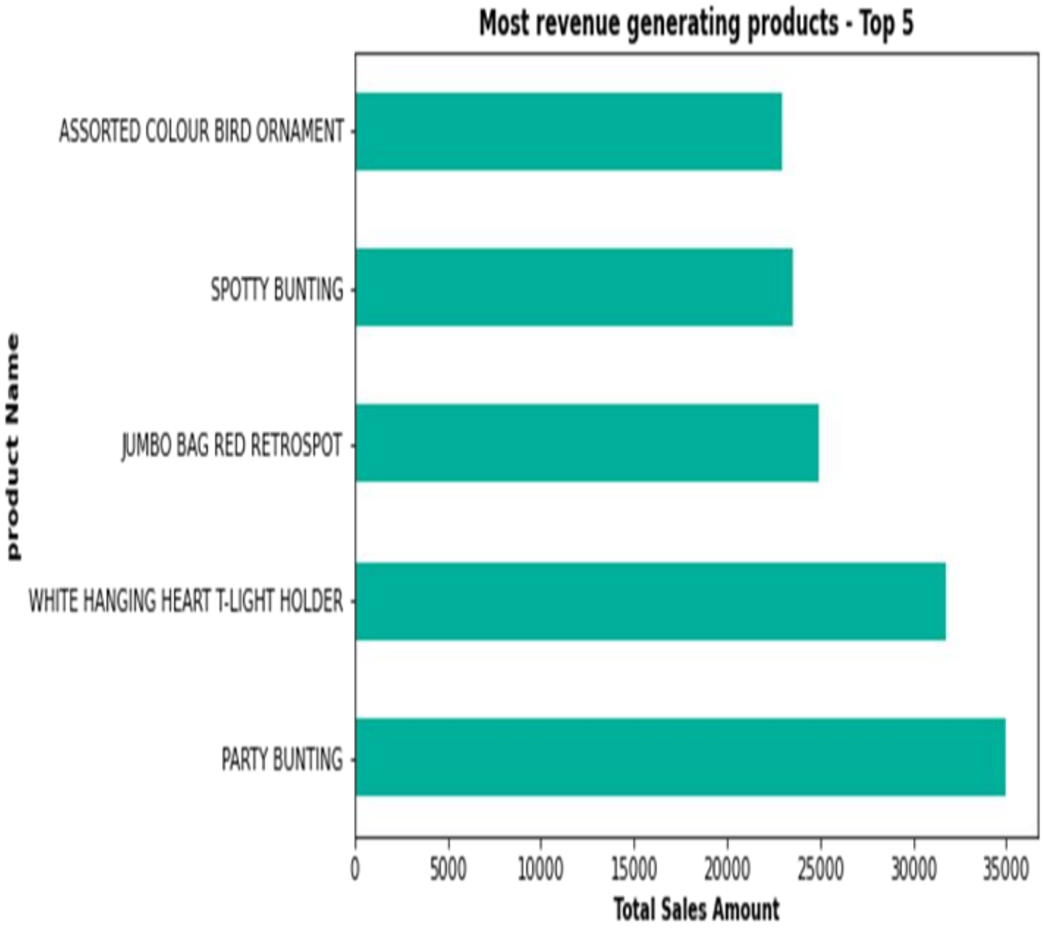


```
product_df_bottom = df['Description'].value_counts()[-5: ]
print(product_df_bottom)
```

```
M/COLOUR POM-POM CURTAIN              1
BLUE/GREEN SHELL NECKLACE W PENDANT   1
 I LOVE LONDON MINI RUCKSACK          1
SET 36 COLOURING PENCILS DOILEY       1
RECYCLED ACAPULCO MAT RED             1
```

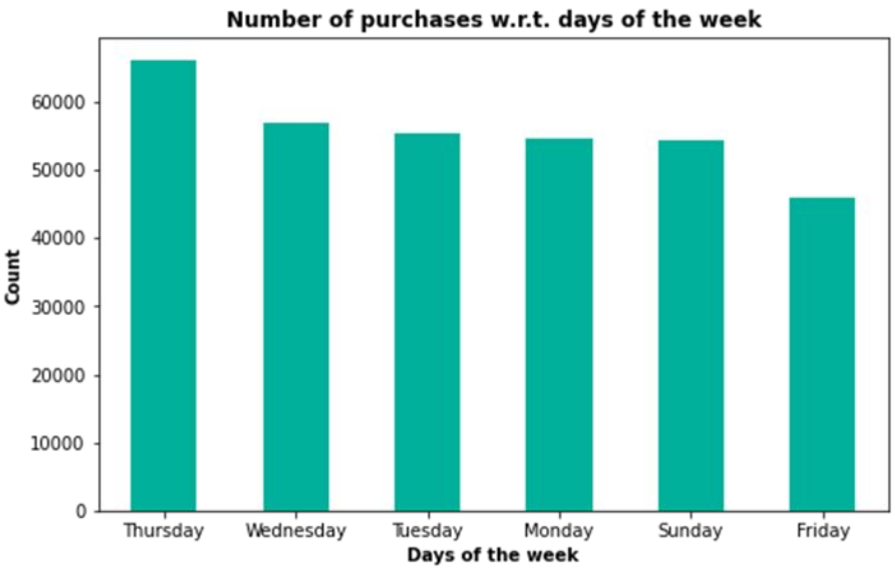
Most sold product is - **WHITE HANGING HEART T-LIGHT HOLDER**

Which are the most revenue generating products?

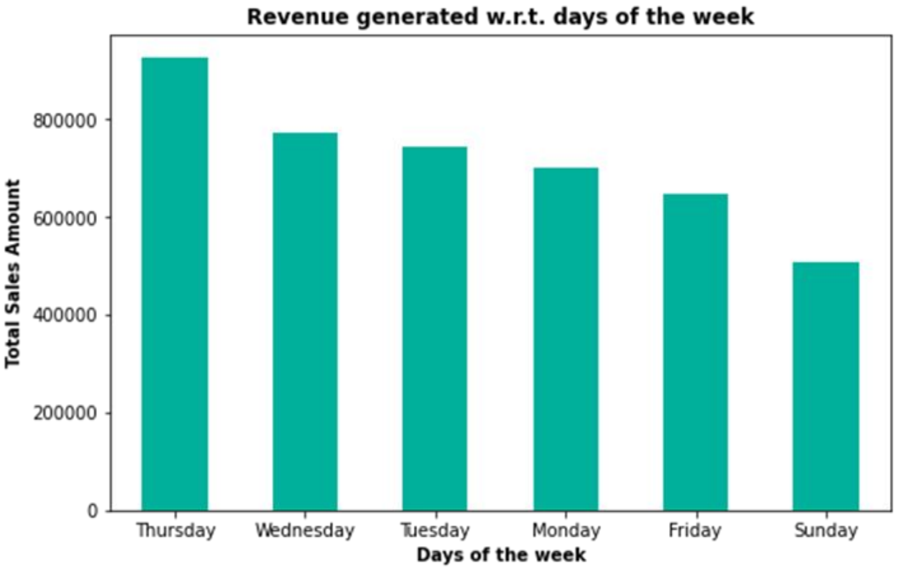


- Most revenue generating product is - **PARTY BUNTING**
- **PARTY BUNTING** is the third highest selling product with highest revenue generator
- **WHITE HANGING HEART T-LIGHT HOLDER** is the top selling product with second highest revenue generator

Which day had the most and least number of purchases?



On which days most of the revenue generated?

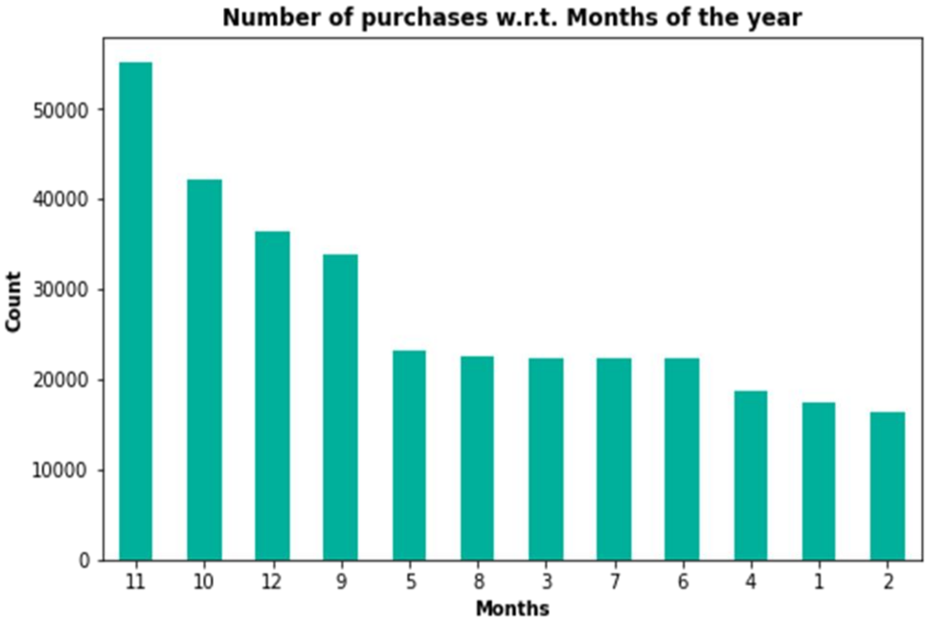


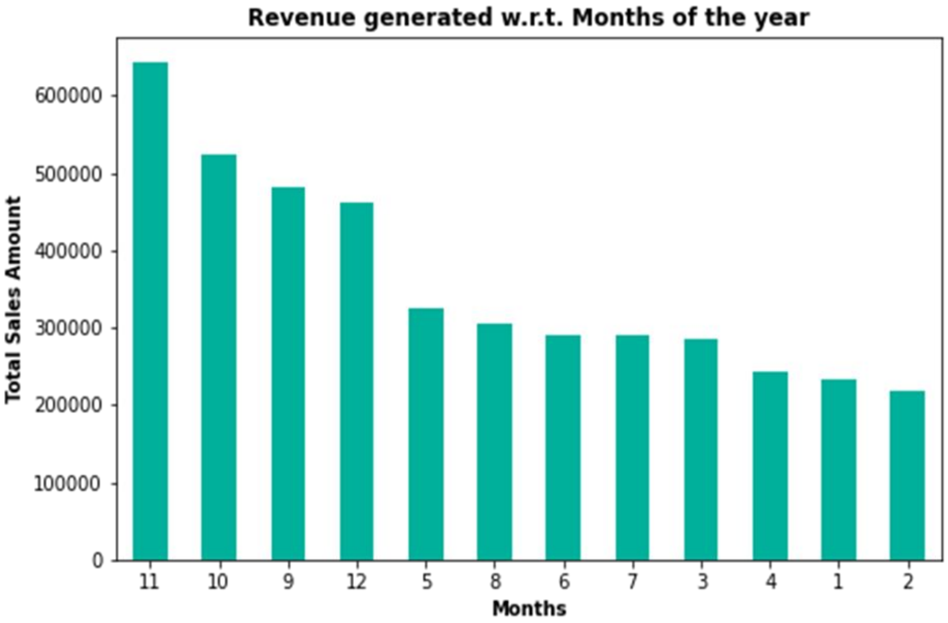
- **Thursday is the day on which highest number of purchases are done which resulted in highest revenue generation compared to other days of the week.**

- **One point to be noted that, Friday was the day on which lowest number of purchases are done, but still in terms of revenue generation Friday is second lowest day of the week. This means that on Friday, products sold were having higher unit price than Sunday.**

Which month had the most and least number of purchases?

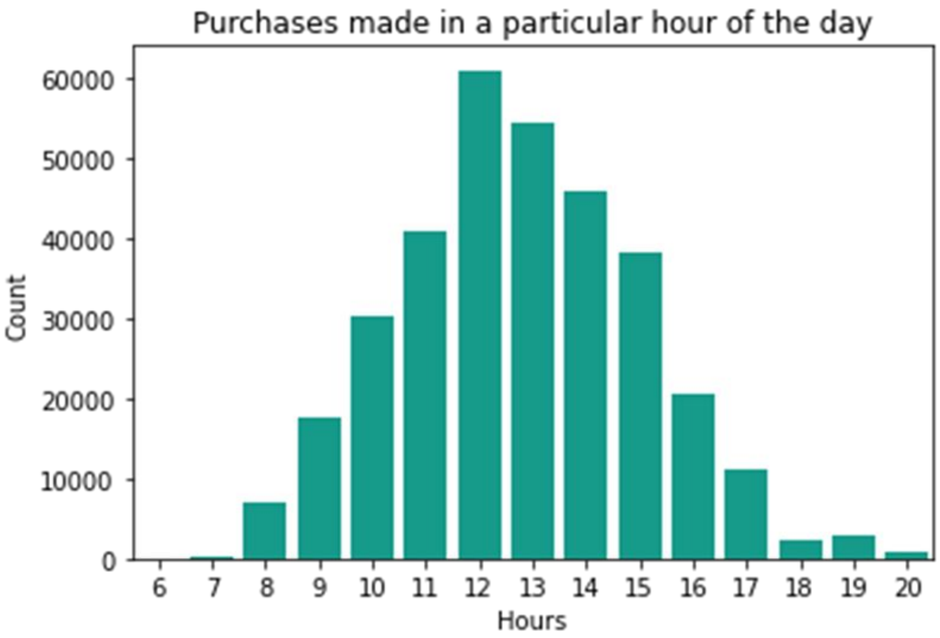
In terms of revenue generation, which month is most important?



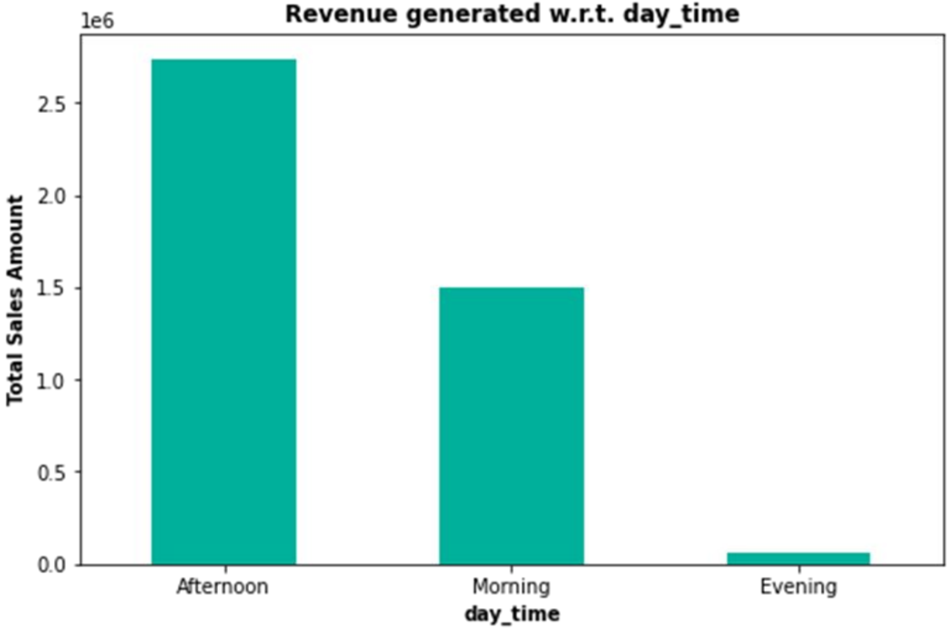


- September to December are the months in which highest number of purchases are happening.
- Same pattern we can see in revenue generation w.r.t. months. September to December, higher number of purchases results in higher revenue.

Which hour in a day had the most and least number of purchases?

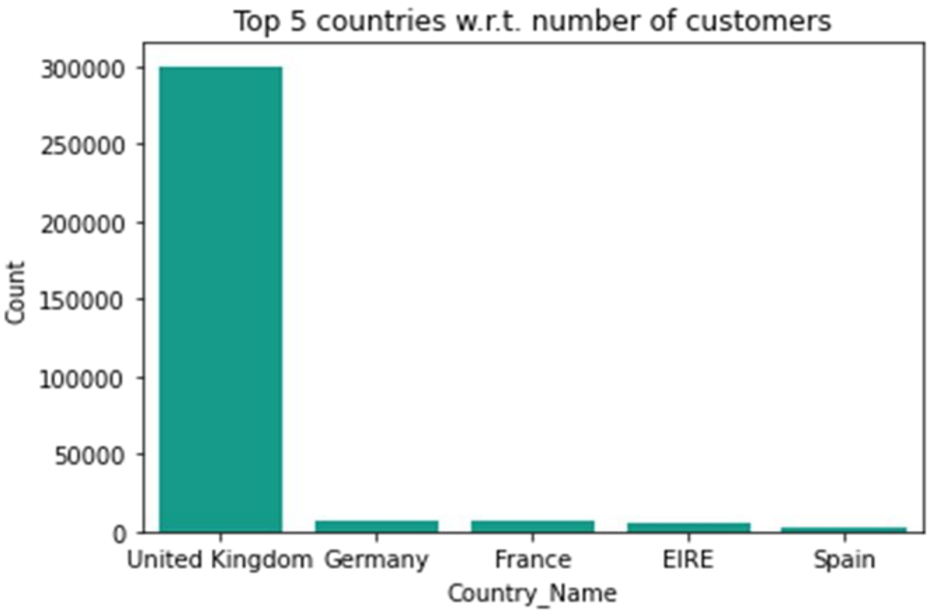


In which time of the day highest revenue is generated?

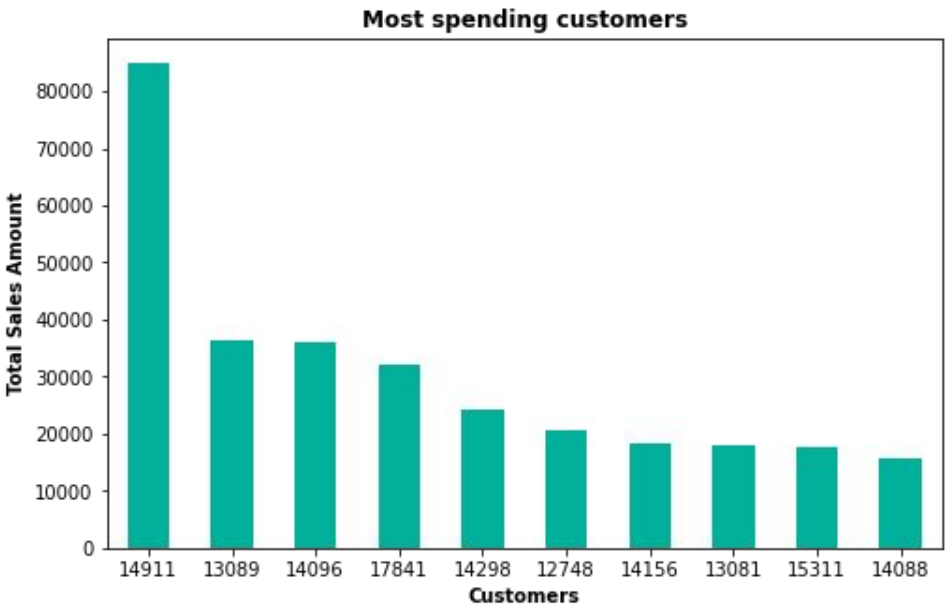


- Most of the purchases are happening between 10 AM to 3 PM.
- As we can see, maximum revenue is generated from the purchases made in the afternoon.

Which countries has the most number of customers?

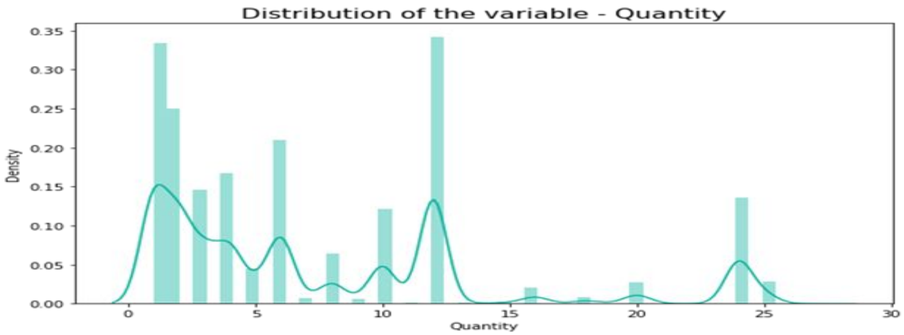


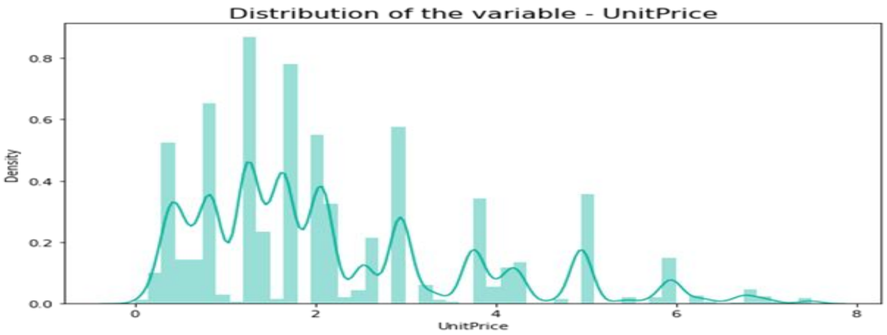
Which are the most spending customers?

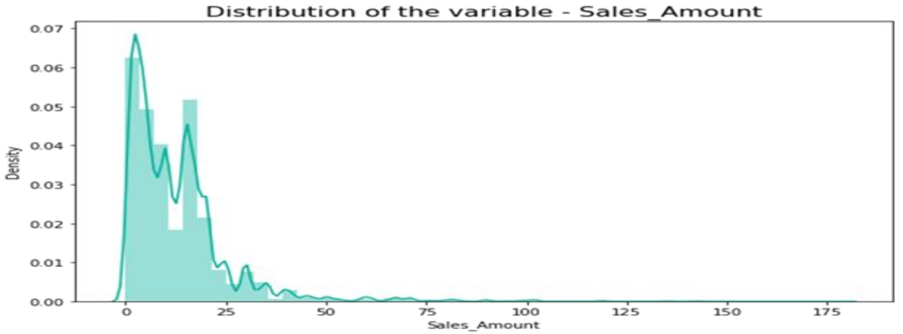


- **Out of total customers, lakhs of customers are from United Kingdom, whereas customers from other countries are hardly some thousands.**
- **CustomerID - 14911 has spent over 80k, which is more than double the amount spent by any customer.**

Distributions of Numerical features such as - Quantity, UnitPrice and Sales_Amount

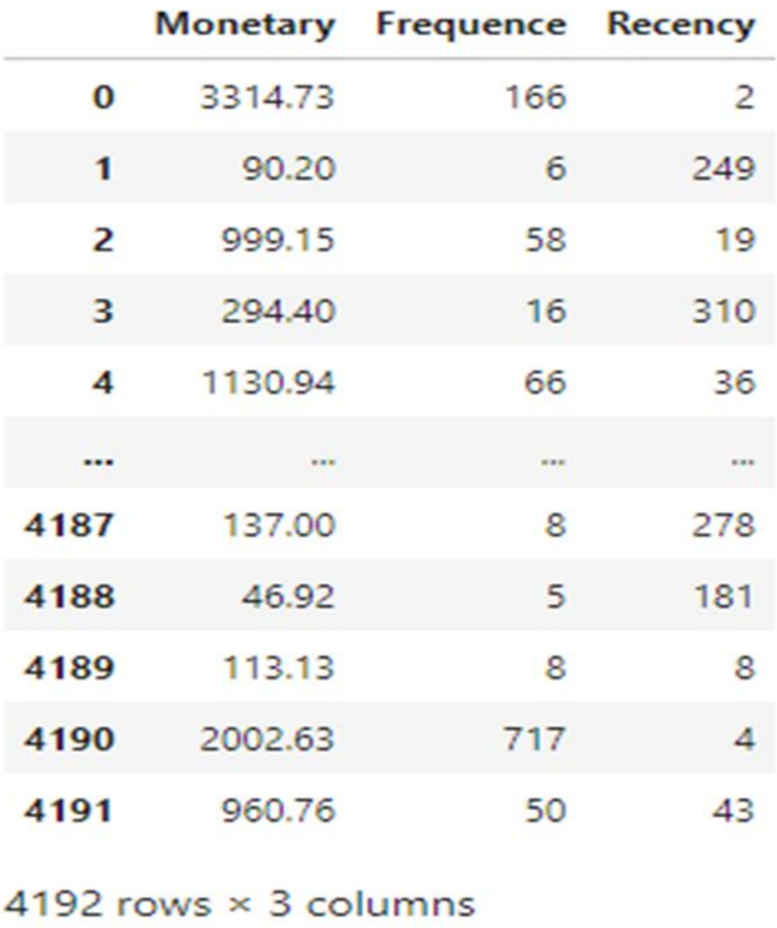






- Quantity and UnitPrice are discrete numerical variables. So, from the distribution plot, we can say that unit price for most of the products ranges between 0.5 to 2.5
- Distribution of Sales_Amount is highly right skewed.

RFM Modeling



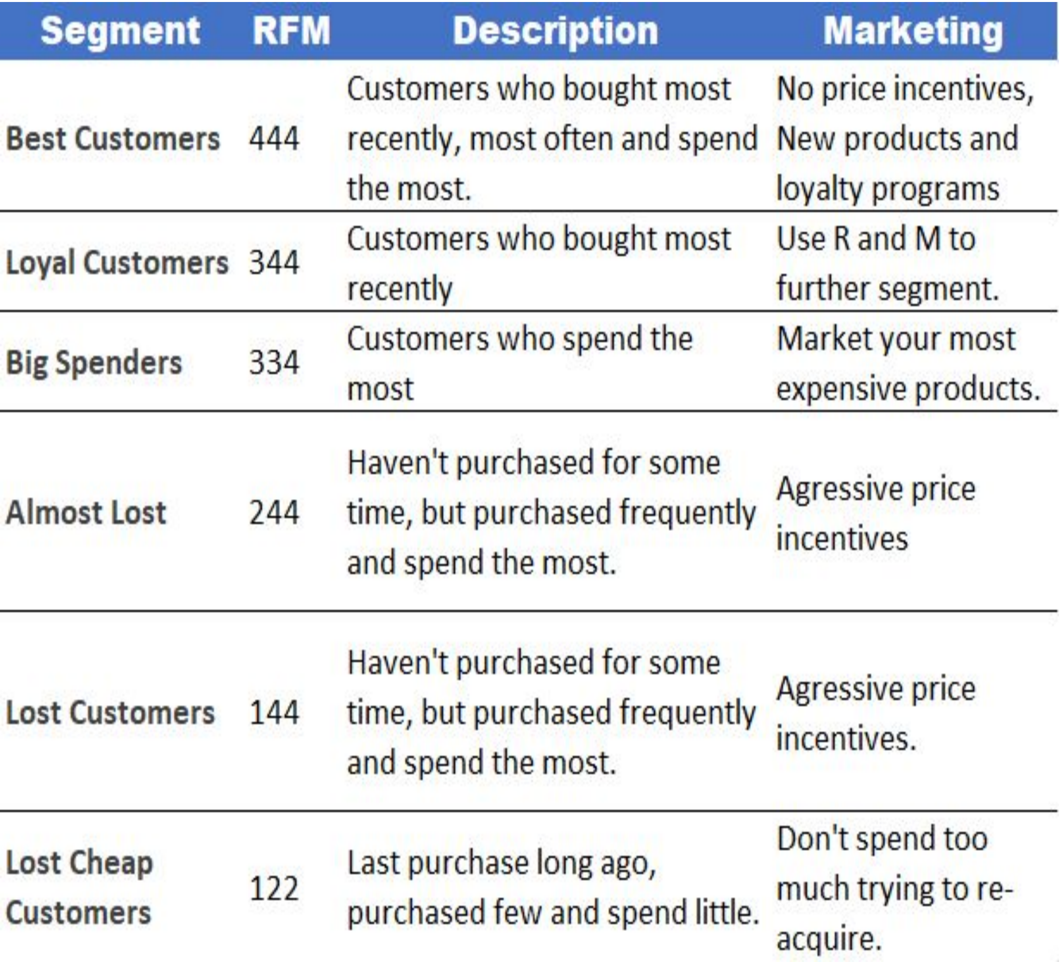
	Monetary	Frequence	Recency
0	3314.73	166	2
1	90.20	6	249
2	999.15	58	19
3	294.40	16	310
4	1130.94	66	36
...	...	...	...
4187	137.00	8	278
4188	46.92	5	181
4189	113.13	8	8
4190	2002.63	717	4
4191	960.76	50	43

4192 rows × 3 columns

We are dividing our customers on the basis of 3 factors

- **Recency:-** It represents how recently a customer purchased a product.
- **Frequency:-** It represents how often a customer purchased a product. The more frequent will be the better score.
- **Monetary:-** It represents how much an customer spends.

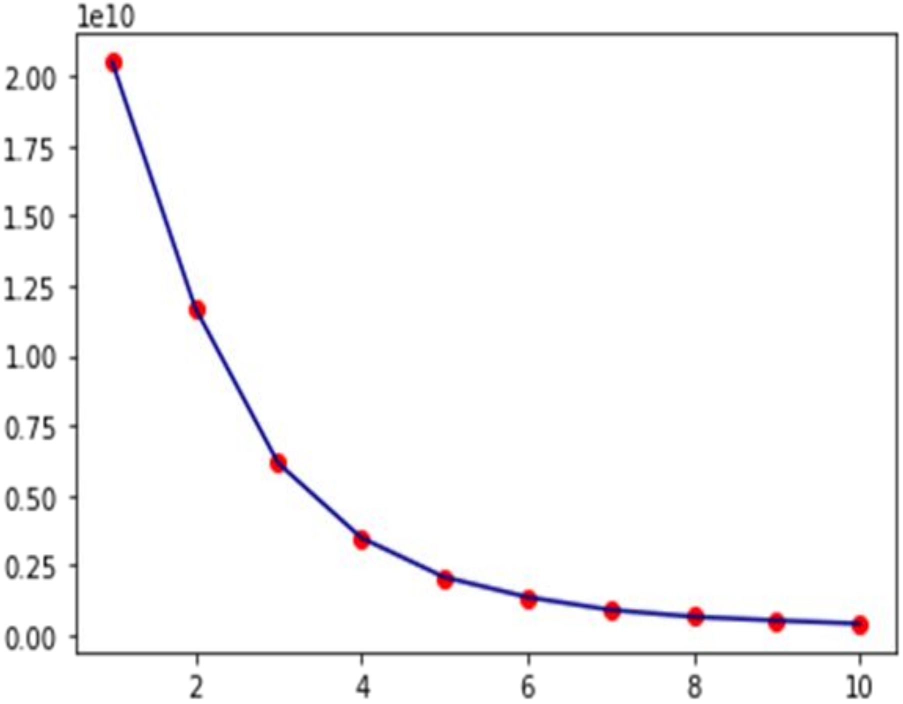
RFM Analysis



Segment	RFM	Description	Marketing
Best Customers	444	Customers who bought most recently, most often and spend the most.	No price incentives, New products and loyalty programs
Loyal Customers	344	Customers who bought most recently	Use R and M to further segment.
Big Spenders	334	Customers who spend the most	Market your most expensive products.
Almost Lost	244	Haven't purchased for some time, but purchased frequently and spend the most.	Agressive price incentives
Lost Customers	144	Haven't purchased for some time, but purchased frequently and spend the most.	Agressive price incentives.
Lost Cheap Customers	122	Last purchase long ago, purchased few and spend little.	Don't spend too much trying to re-acquire.

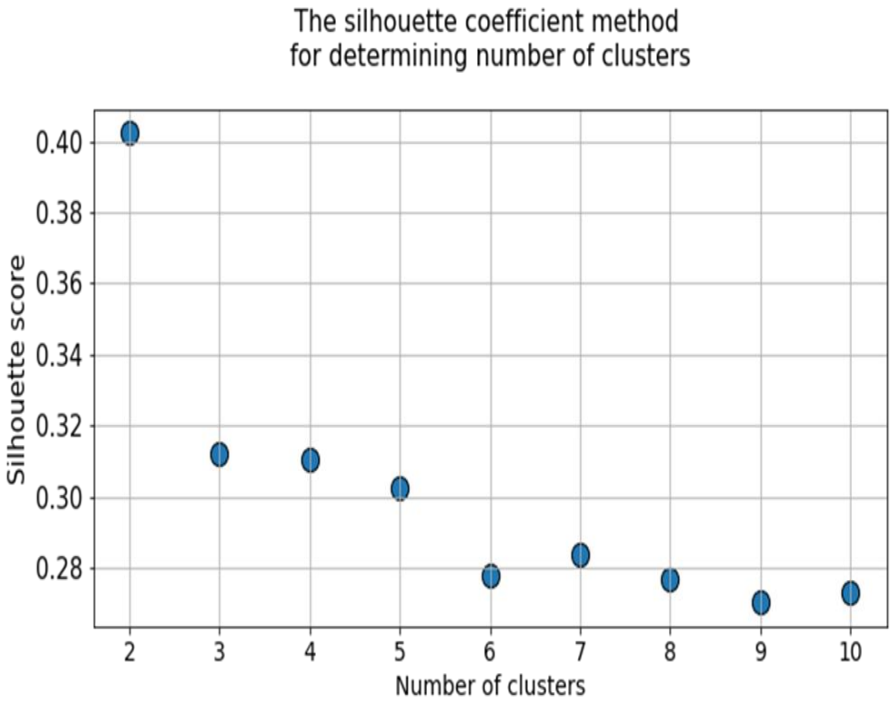
- **Best Recency score = 4 (most recently purchase)**
- **Best Frequency score = 4 (most frequently purchase)**
- **Best Monetary score = 4 (who spent the most)**

Elbow method to decide number of clusters



From the elbow graph, it seems that good number of cluster would be either 2 or 3 as after that, its a smooth curve i.e. no change of orientation. but to overcome that confusion, we will use silhouette score method to find the optimum number of clusters because it is often much better in figuring out the number of valid clusters than the elbow method.

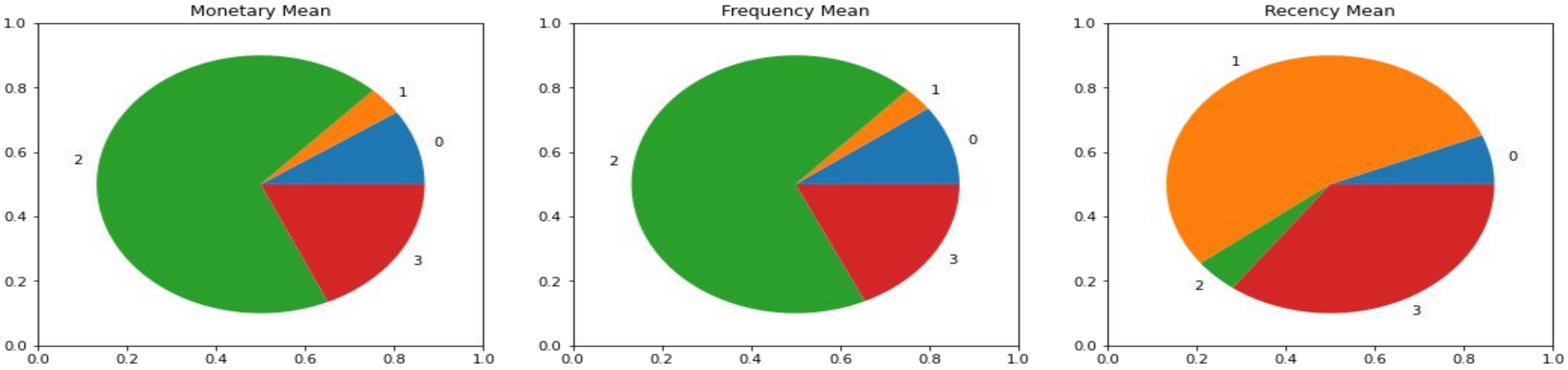
Silhouette Score Method



Here we can clearly see that optimum number of cluster should be 4 not 2 or 3. Because that is the only point after which the mean cluster distance looks to be plateaued after a steep downfall. So we will assume the 4 number of clusters as best for grouping of customer segments.

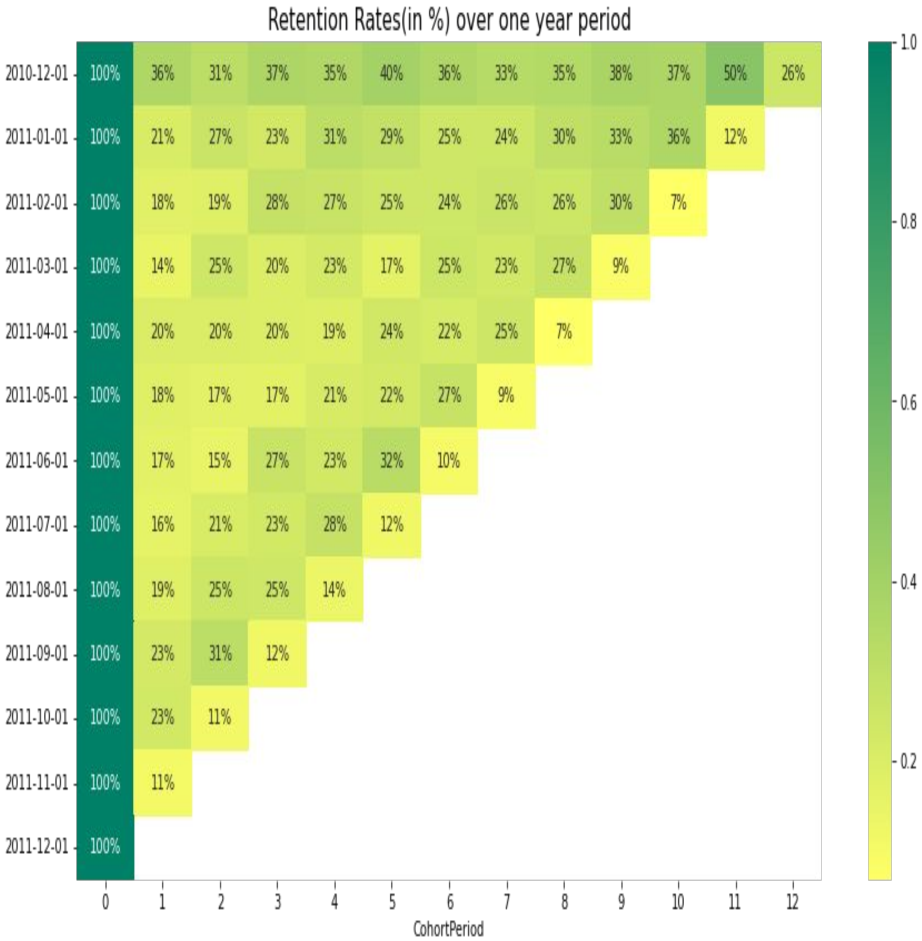
Now let's apply K-Means on 4 clusters to segregate the customer base.

K-Means Clustering



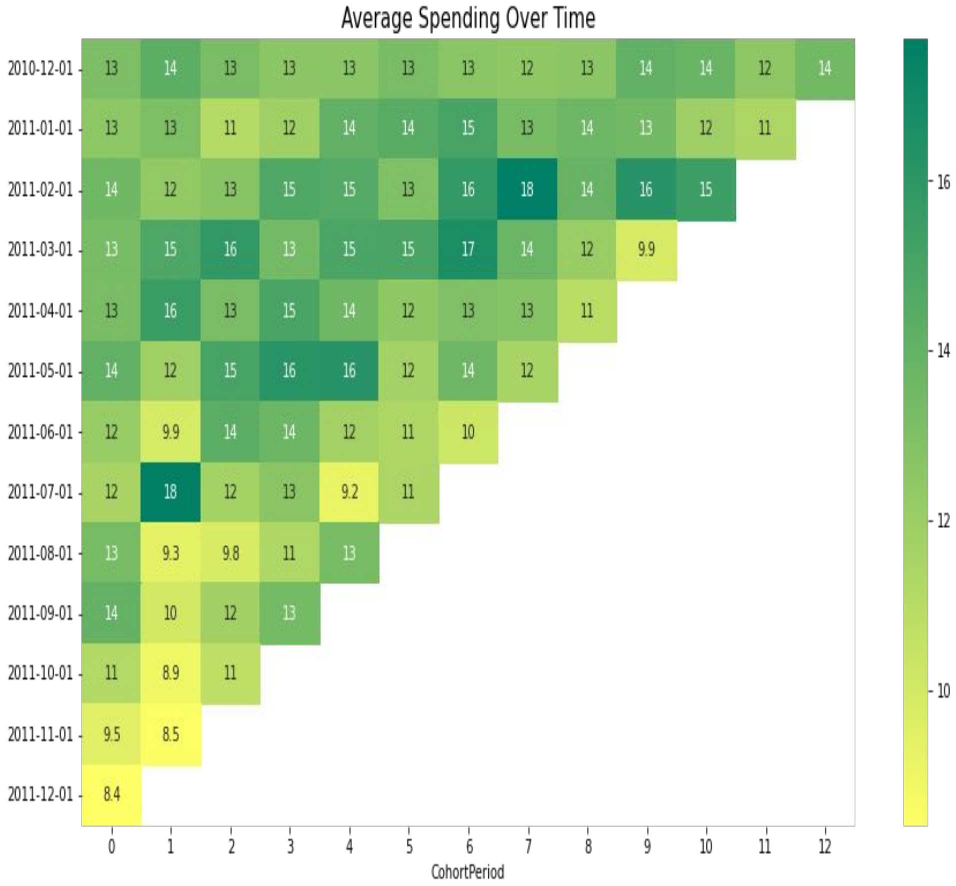
- Group 2 is the group of customers who spends maximum amount of money and also has a good frequency and very low recency rate. These are most important customers to the company with respect to revenue building.
- Group 0 is the group of customers whose frequency rate and monetary value are good and has low frequency rate compared to group 3 customers.
- Group 1 is the group of customers who has a very high recency rate means they have not purchased for a long time. Also, they have very less purchasing power and frequency is very low.
- Group 3 is the group of customers who has medium spending capacity and frequency but, they haven't purchased for a long time now.

Cohort Analysis



Retention Rates(in %) over one year period

	0	1	2	3	4	5	6	7	8	9	10	11	12
2010-12-01	100%	36%	31%	37%	35%	40%	36%	33%	35%	38%	37%	50%	26%
2011-01-01	100%	21%	27%	23%	31%	29%	25%	24%	30%	33%	36%	12%	
2011-02-01	100%	18%	19%	28%	27%	25%	24%	26%	26%	30%	7%		
2011-03-01	100%	14%	25%	20%	23%	17%	25%	23%	27%	9%			
2011-04-01	100%	20%	20%	20%	19%	24%	22%	25%	7%				
2011-05-01	100%	18%	17%	17%	21%	22%	27%	9%					
2011-06-01	100%	17%	15%	27%	23%	32%	10%						
2011-07-01	100%	16%	21%	23%	28%	12%							
2011-08-01	100%	19%	25%	25%	14%								
2011-09-01	100%	23%	31%	12%									
2011-10-01	100%	23%	11%										
2011-11-01	100%	11%											
2011-12-01	100%												



Average Spending Over Time

	0	1	2	3	4	5	6	7	8	9	10	11	12
2010-12-01	13	14	13	13	13	13	13	12	13	14	14	12	14
2011-01-01	13	13	11	12	14	14	15	13	14	13	12	11	
2011-02-01	14	12	13	15	15	13	16	18	14	16	15		
2011-03-01	13	15	16	13	15	15	17	14	12	9.9			
2011-04-01	13	16	13	15	14	12	13	13	11				
2011-05-01	14	12	15	16	16	12	14	12					
2011-06-01	12	9.9	14	14	12	11	10						
2011-07-01	12	18	12	13	9.2	11							
2011-08-01	13	9.3	9.8	11	13								
2011-09-01	14	10	12	13									
2011-10-01	11	8.9	11										
2011-11-01	9.5	8.5											
2011-12-01	8.4												

Conclusions

- Throughout the exercise, we went through various steps to perform customer segmentation. We started with importing data and important libraries. Then, did rigorous data wrangling.

- We have performed RFM Analysis on the data, where we clustered customers based on Recency, Monetary and Frequency aspect. We used Elbow method, Silhouette score method to find appropriate number of clusters. We discovered 4 clusters based on RFM data.

- Further, did cohort analysis to understand how retention and acquisition rate, average amount spends changes over the time.

- However, there can be more modifications on this analysis. One may choose to cluster into more no. depending on company objectives and preferences. The labelled feature after clustering can be fed into classification supervised machine learning algorithms that could predict the classes for new set of observations. The clustering can also be performed on new set of features such as type of products each customer prefers to buy often, finding out customer lifetime value (clv) and much more.