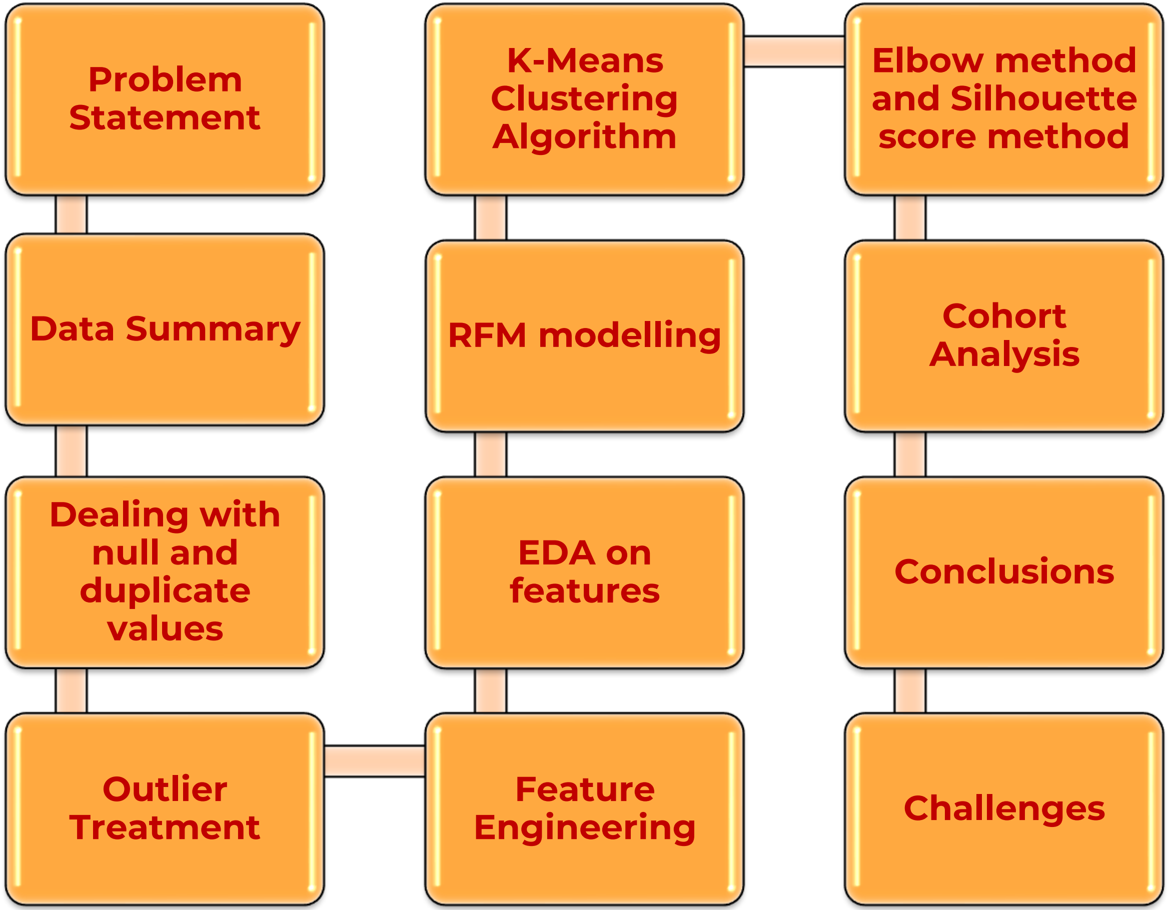
Capstone Project 3

Online Retail Customer Segmentation

Unsupervised ML - Clustering

Sayali Kamalapurkar

Table Of Contents



Problem Statement

Data Summary

Dealing with null and duplicate values

Outlier Treatment

Feature Engineering

EDA on features

RFM modelling

K-Means Clustering Algorithm

Elbow method and Silhouette score method

Cohort Analysis

Conclusions

Challenges

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 each and 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 transnational 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.

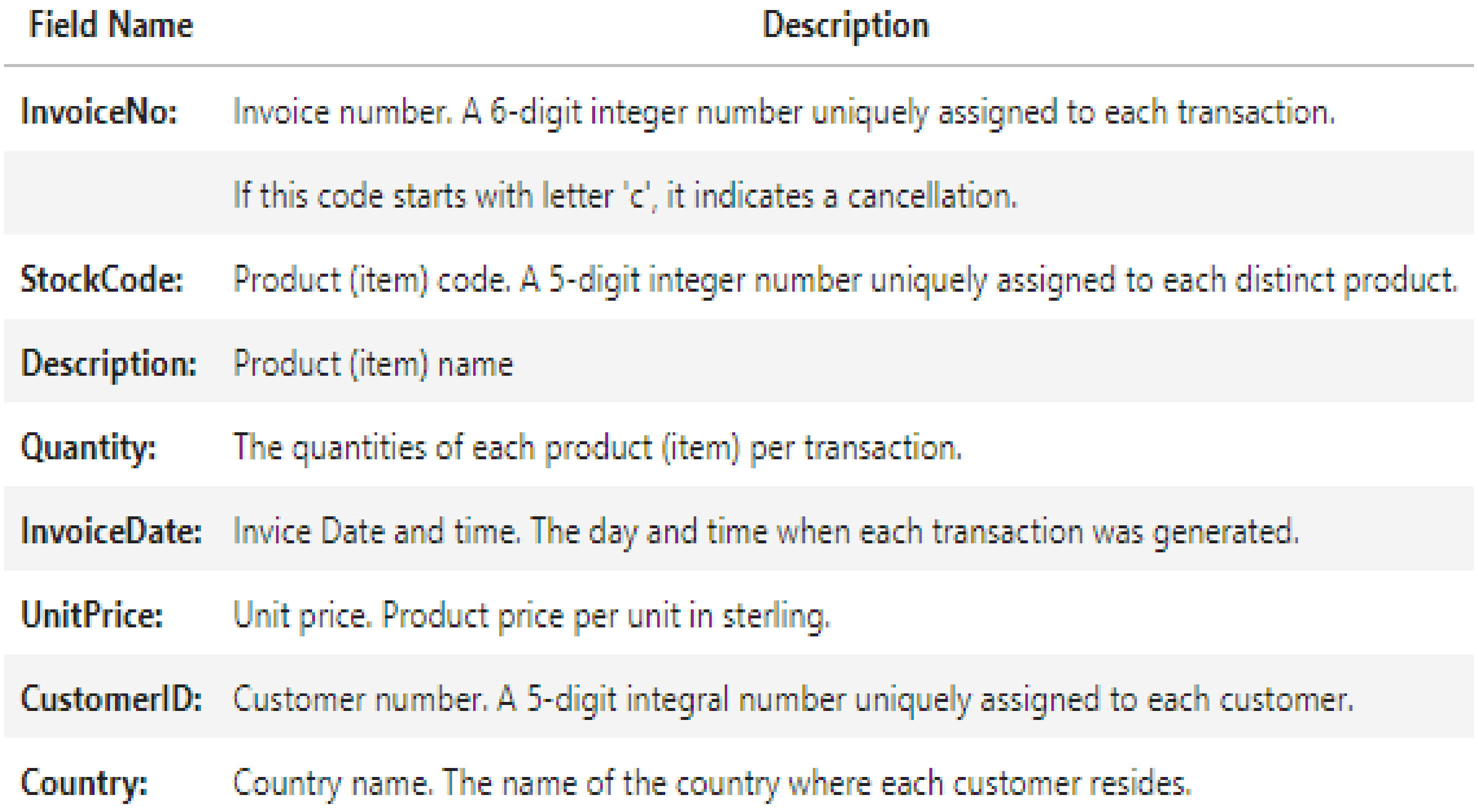
In this project, we are required to :

•Understand the Dataset & cleanup (if required)

•Build a clustering model to segment the customer-based similarity

•Also fine-tune the hyperparameters & compare the evaluation metrics of various clustering algorithms

Data Summary

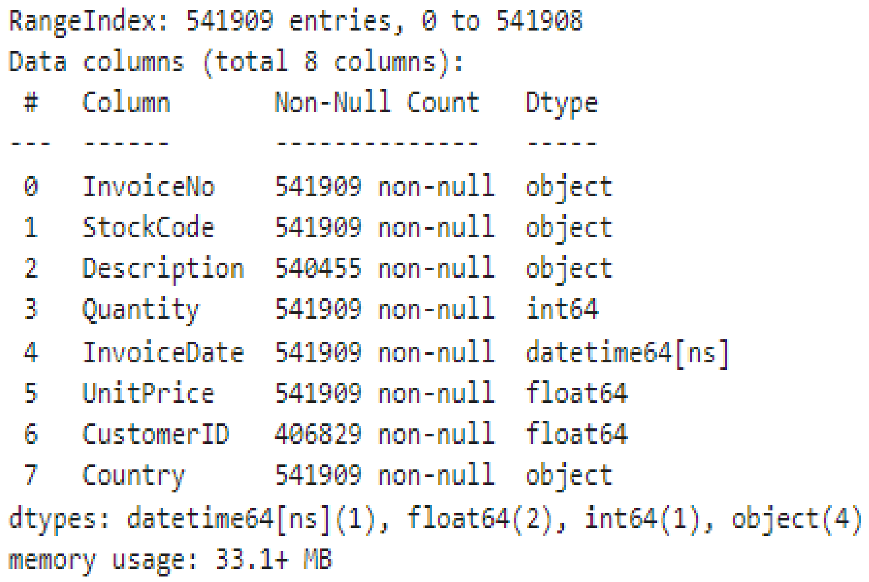


Field Name	Description
InvoiceNo:	Invoice number. A 6-digit integer number uniquely assigned to each transaction.
	If this code starts with letter 'c', it indicates a cancellation.
StockCode:	Product (item) code. A 5-digit integer number uniquely assigned to each distinct product.
Description:	Product (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.

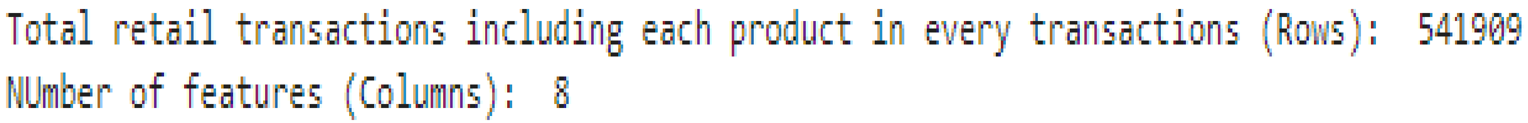
Data Summary

There are 4 features whose datatype is object, 2 features whose datatype is float64, 1 feature whose datatype is datetime64 and 1 feature whose datatype is int64.

Memory usage by the dataset is 33.1 MB.



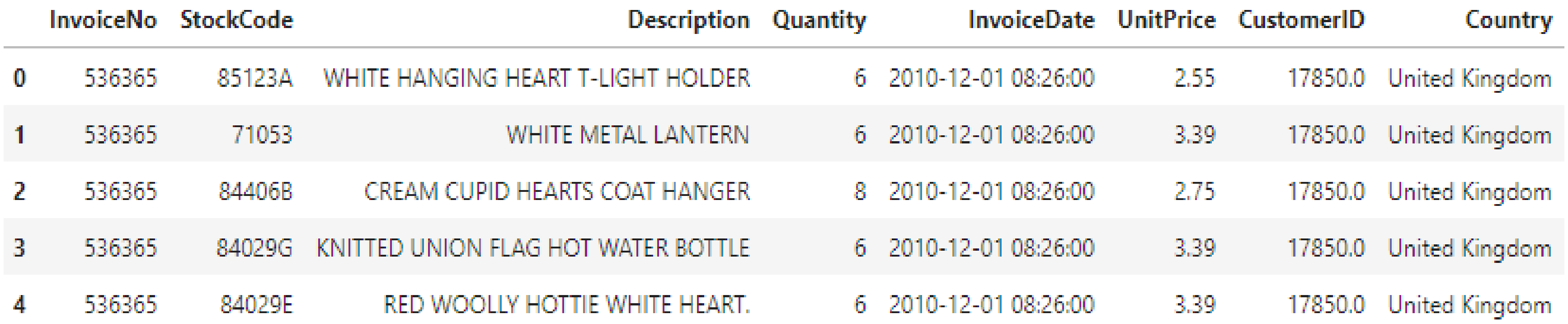
```
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```



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

Data Summary(Continued...)

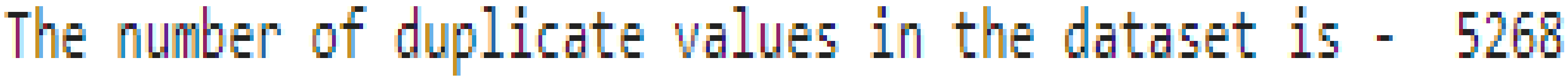
Let us have a look at our dataset -



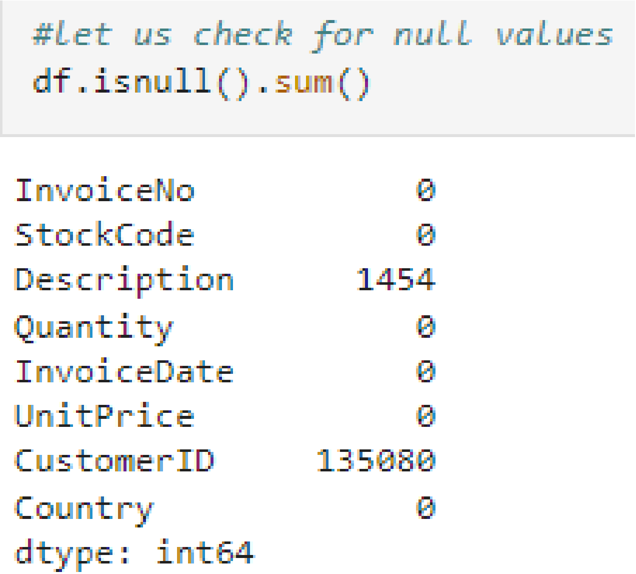
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS 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

- **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.**

Dealing with null and duplicate values



```
The number of duplicate values in the dataset is -  5268
```



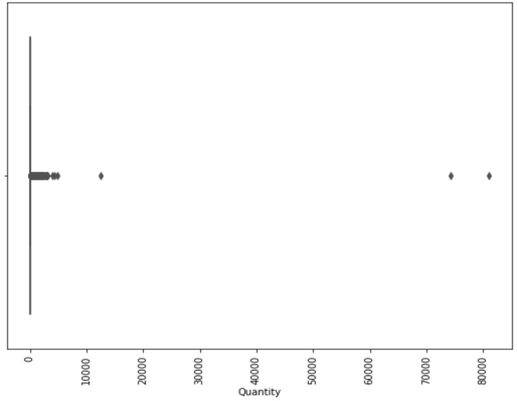
```
#let us check for null values
df.isnull().sum()
```

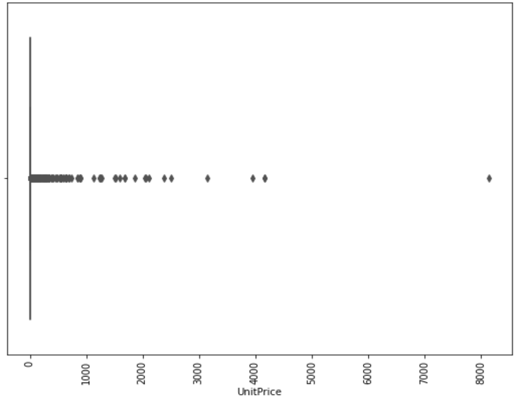
```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

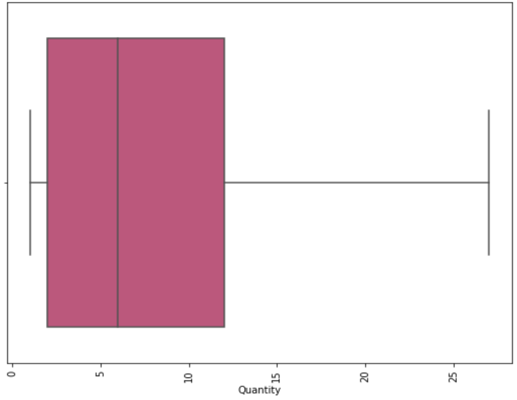
There are 135080 null values in CustomerID column.

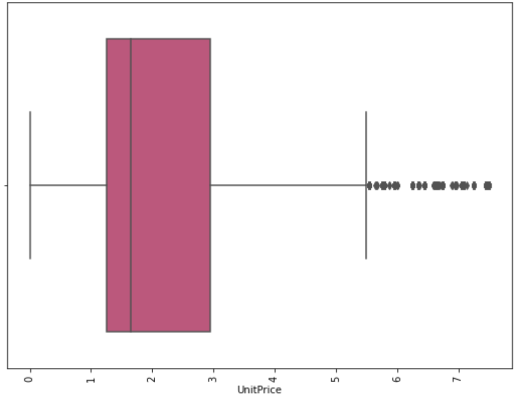
We have dropped duplicate rows and also rows containing null values in CustomerID column.

Outlier Treatment









We 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

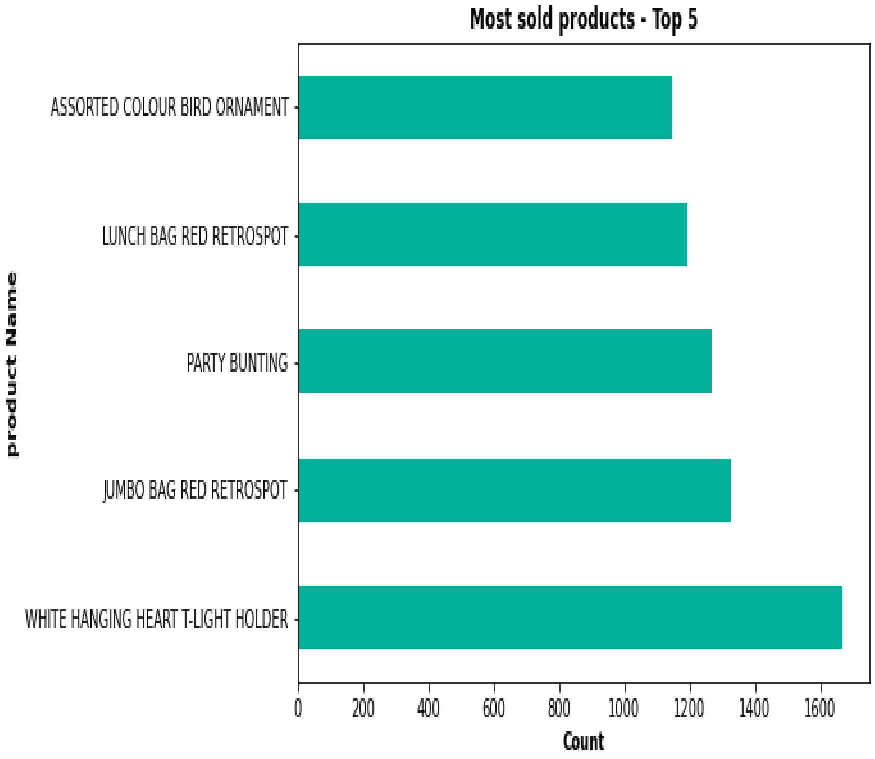
As we have very less features about the customers, we created some new features from existing features.

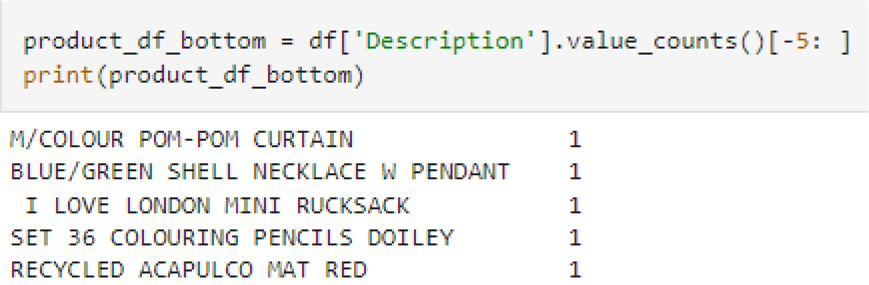
They are as follows –

- **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?





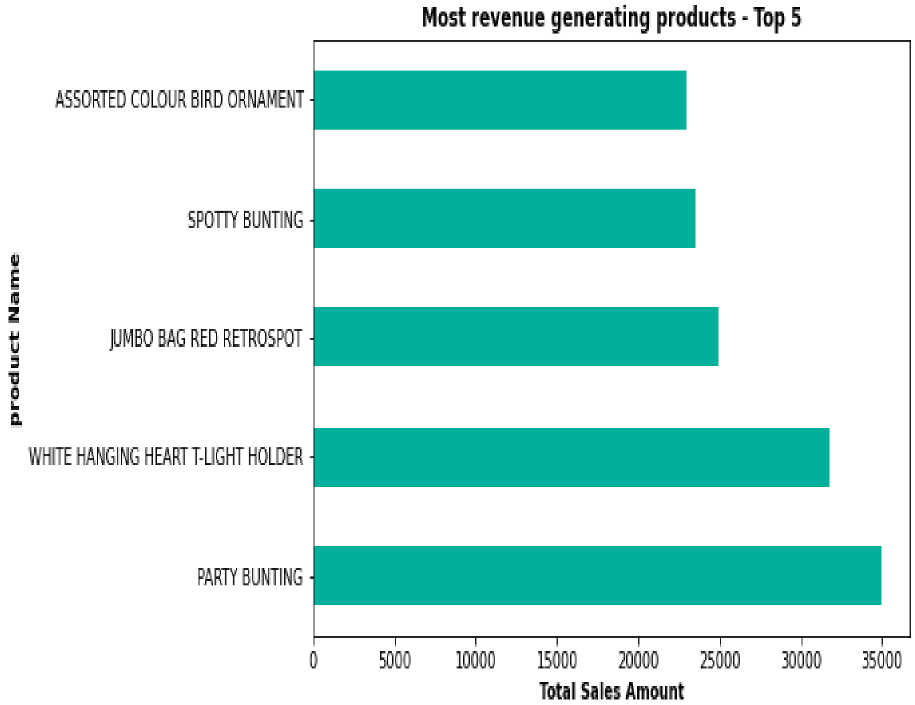
```
product_df_bottom = df['Description'].value_counts()[-5: ]
print(product_df_bottom)
```

```
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**

Continued...

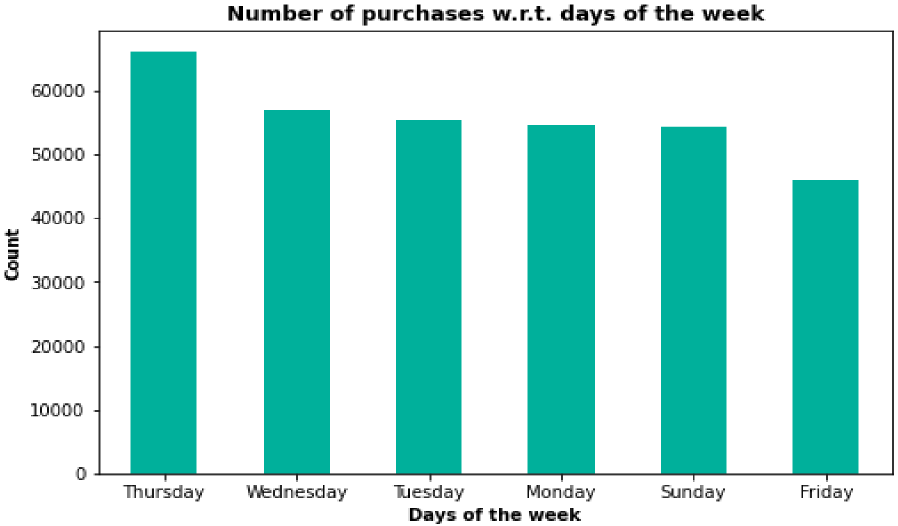
Which are the most revenue generating products?

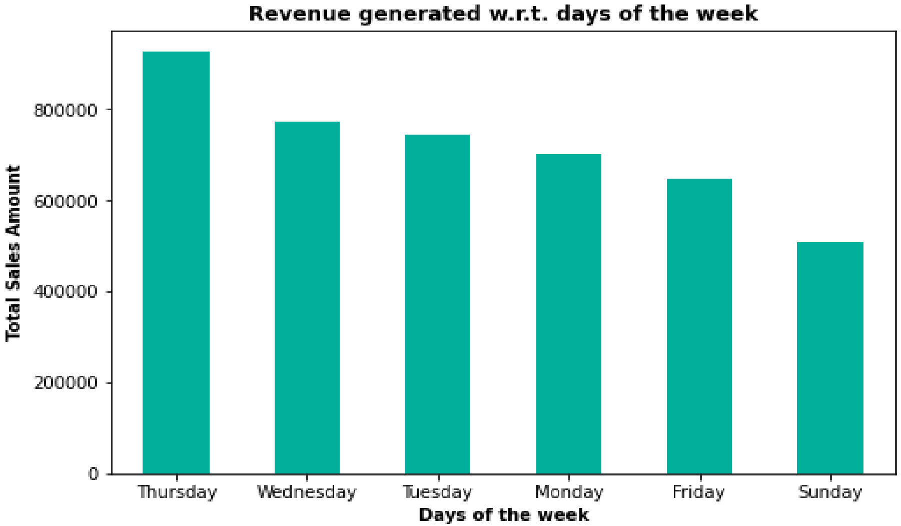


- Most revenue generating product is - **PARTY BUNTING**
- **PARTY BUNTING** is the third highest selling product with highest revenue generator
- **WHITE HANGING HEART T-LIGHT HOLDER** is the top selling product with second highest revenue generator

Continued…

Which day had the most and least number of purchases?
On which days most of the revenue generates?



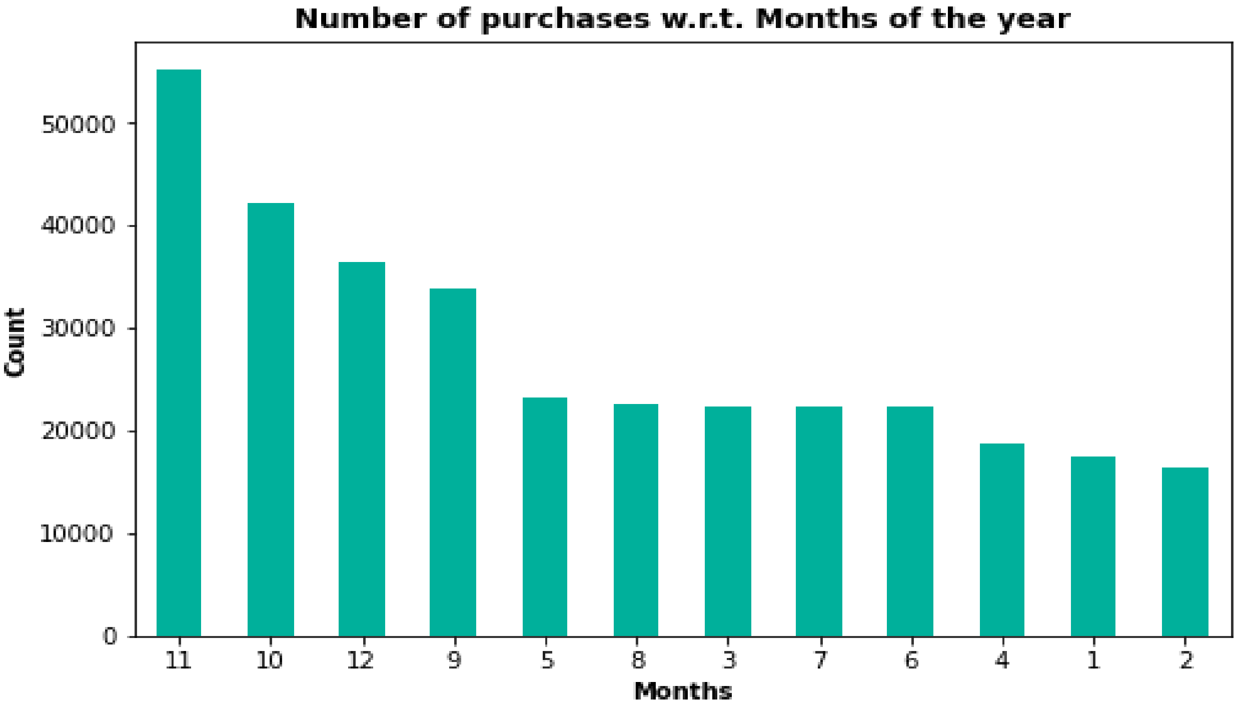


- **Thursday is the day on which highest number of purchases are done which resulted in highest revenue generation compared to other days of the week.**

- **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.**

Continued...

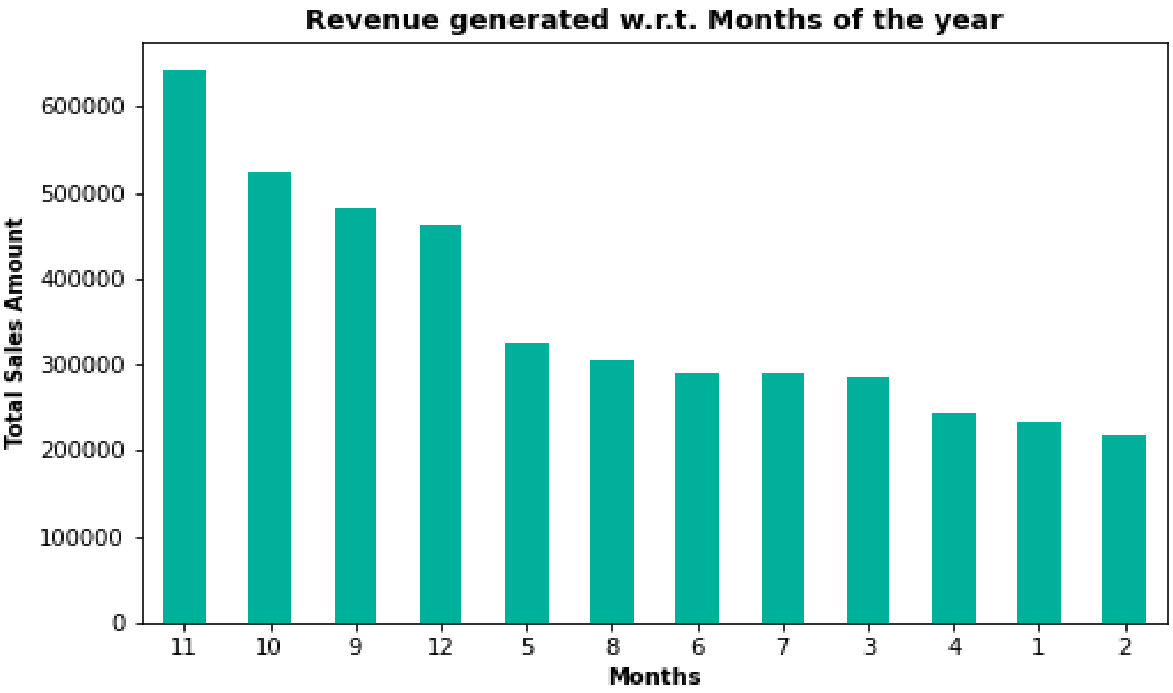
Which month had the most and least number of purchases?



September to December are the months in which highest number of purchases are happening.

Continued...

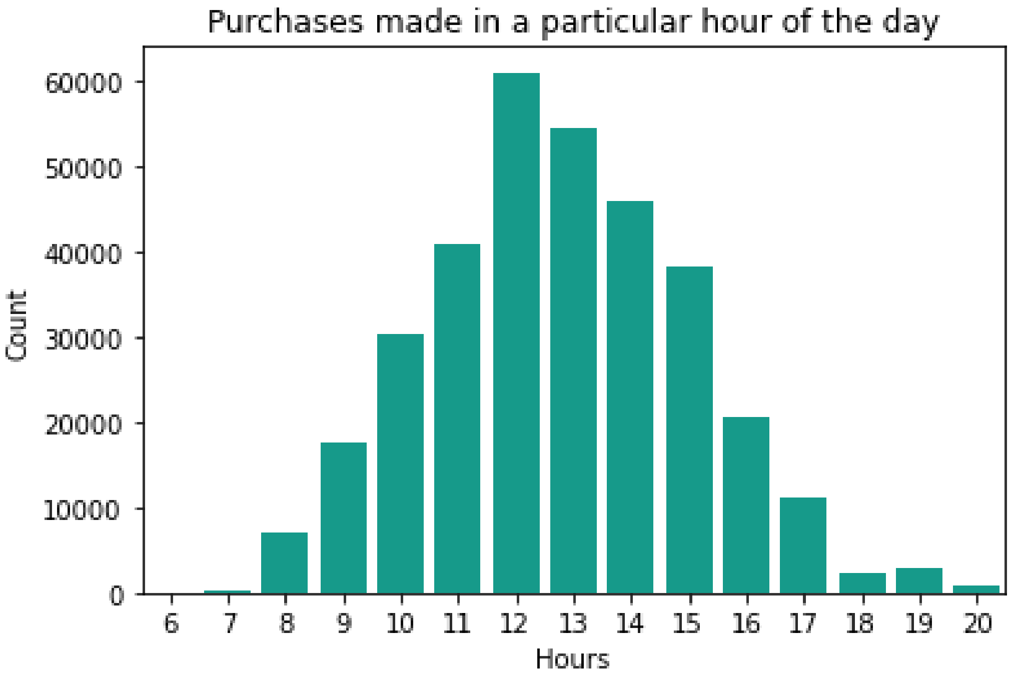
In terms of revenue generation, which month is most important?



Same pattern we can see in revenue generation w.r.t. months. September to December, higher number of purchases results in higher revenue.

Continued...

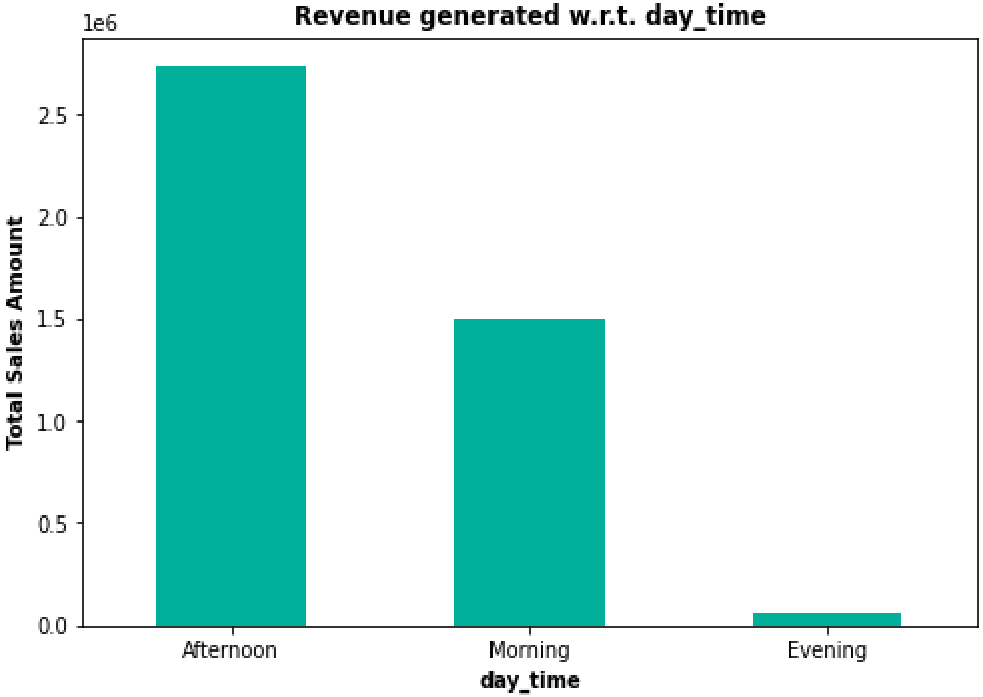
Which hour in a day had the most and least number of purchases?



Most of the purchases are happening between 10 AM to 3 PM.

Continued...

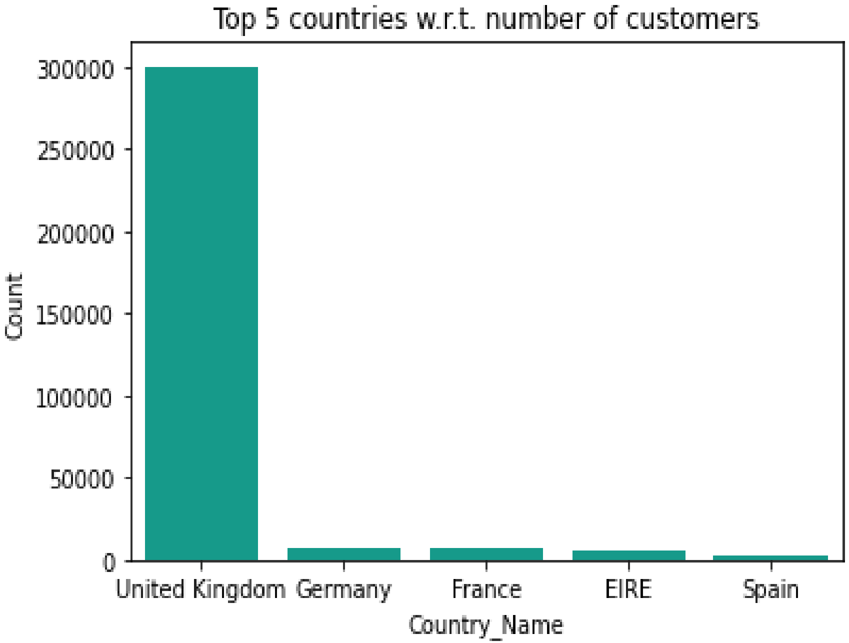
In which time of the day highest revenue is generated?



As we can see, maximum revenue is generated from the purchases made in the afternoon.

Continued...

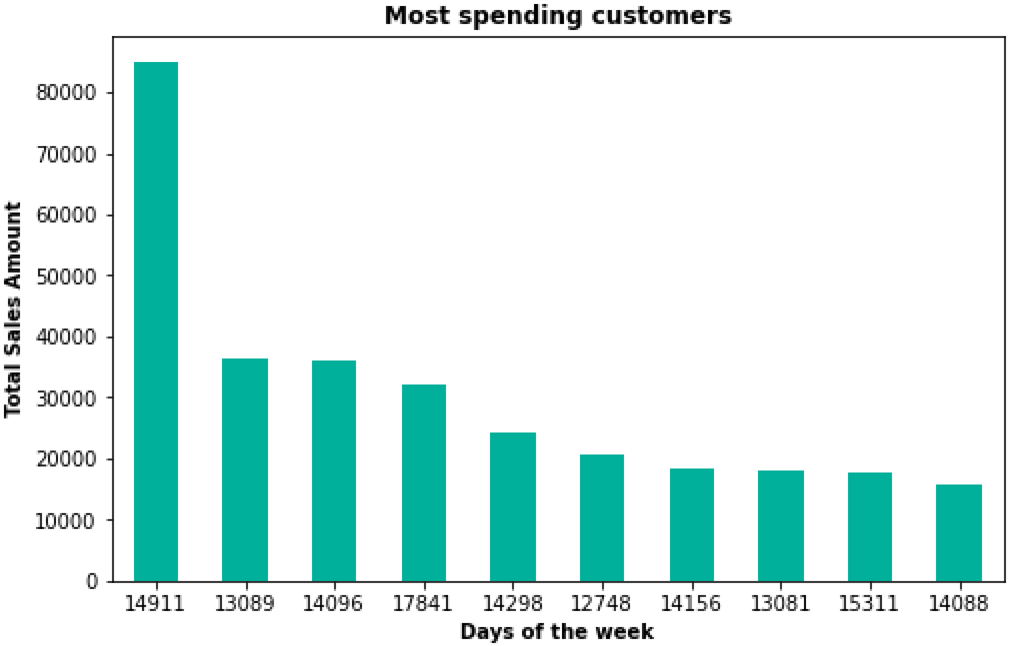
Which countries has the most number of customers?



Out of total customers, lakhs of customers are from United Kingdom, whereas customers from other countries are hardly some thousands.

Continued...

Which are the most spending customers?

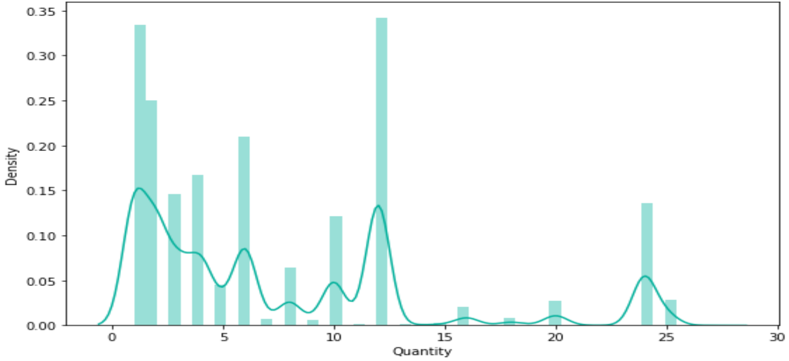


CustomerID - 14911 has spent over 80k, which is more than double the amount spent by any customer.

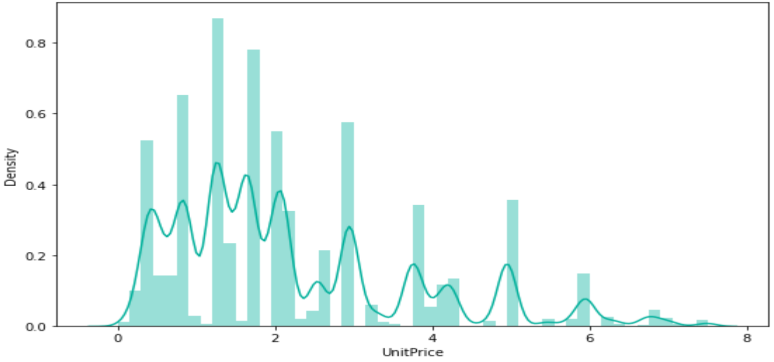
Continued...

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

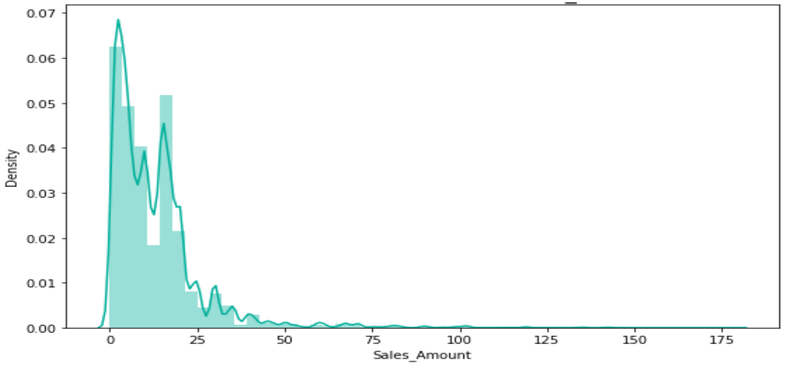
Distribution of the variable - Quantity



Distribution of the variable - UnitPrice



Distribution of the variable - 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.**

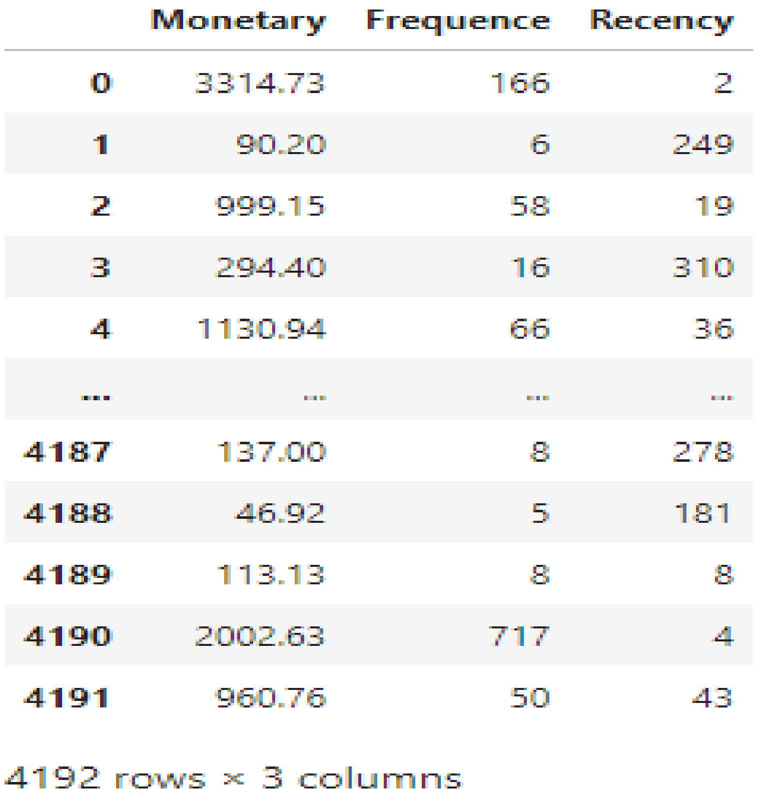
RFM Modelling

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.

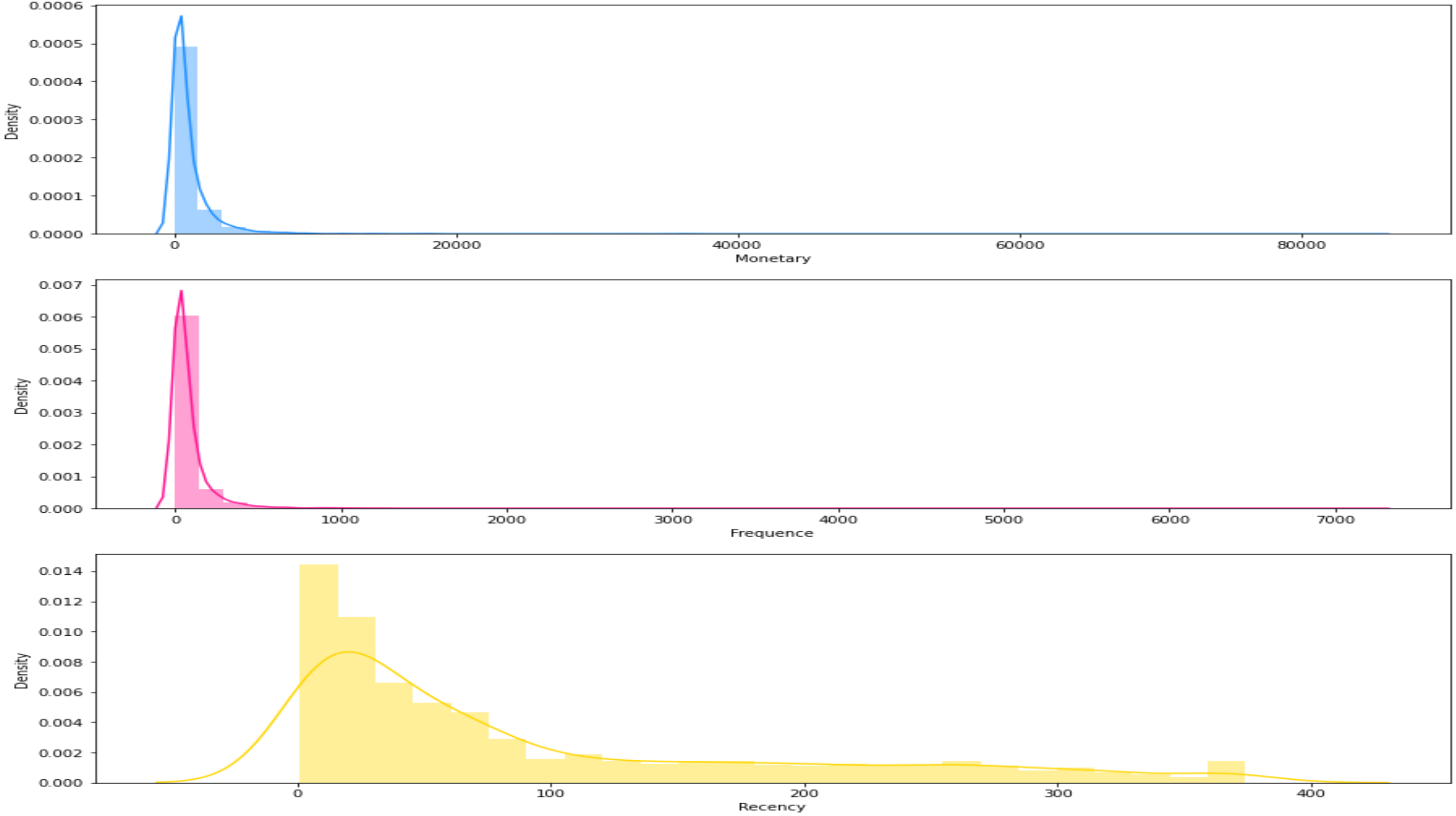
Monetary:- It represents how much an customer spends.



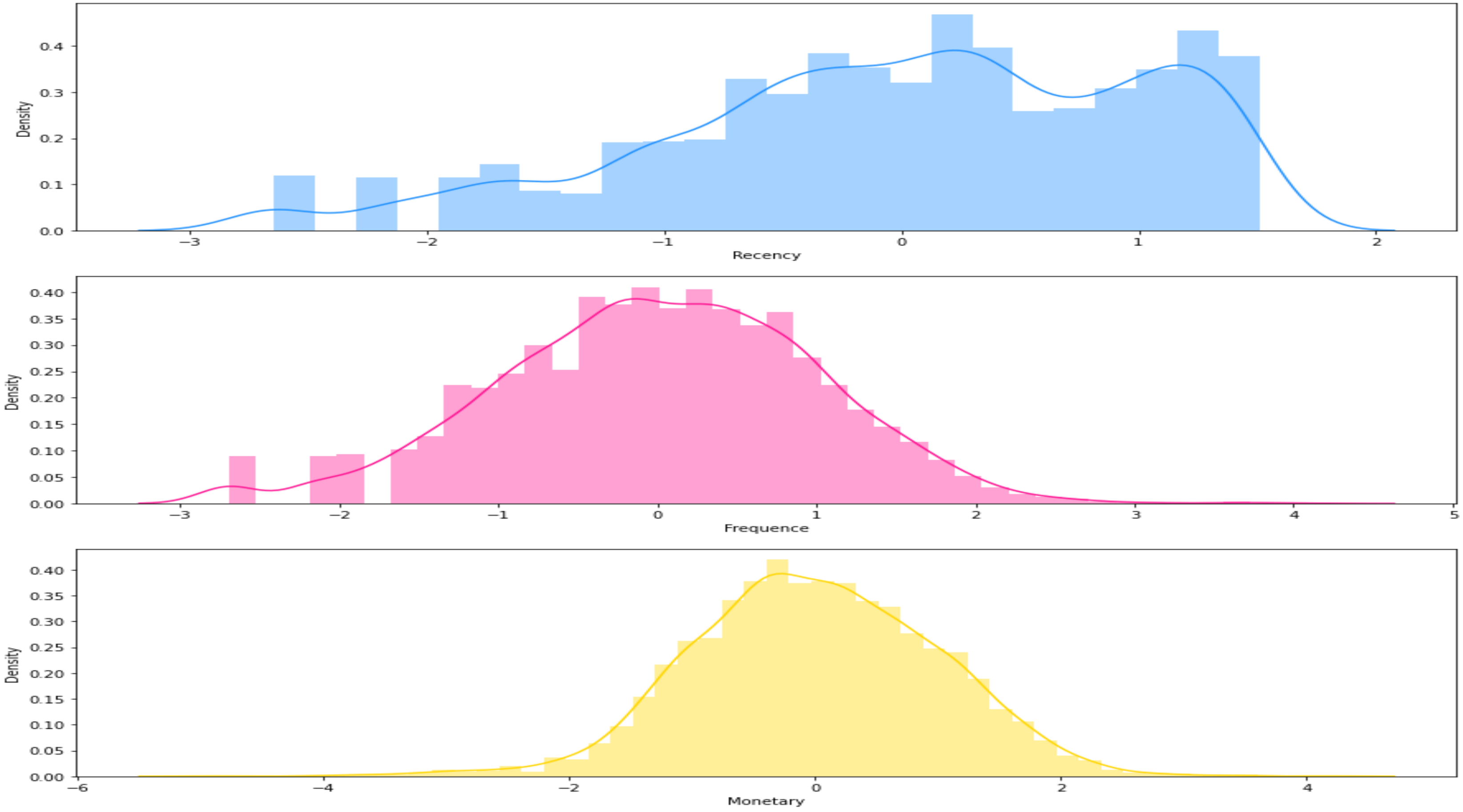
	Monetary	Frequence	Recency
0	3314.73	166	2
1	90.20	6	249
2	999.15	58	19
3	294.40	16	310
4	1130.94	66	36
...	...	...	...
4187	137.00	8	278
4188	46.92	5	181
4189	113.13	8	8
4190	2002.63	717	4
4191	960.76	50	43

4192 rows × 3 columns

Distributions of RFM

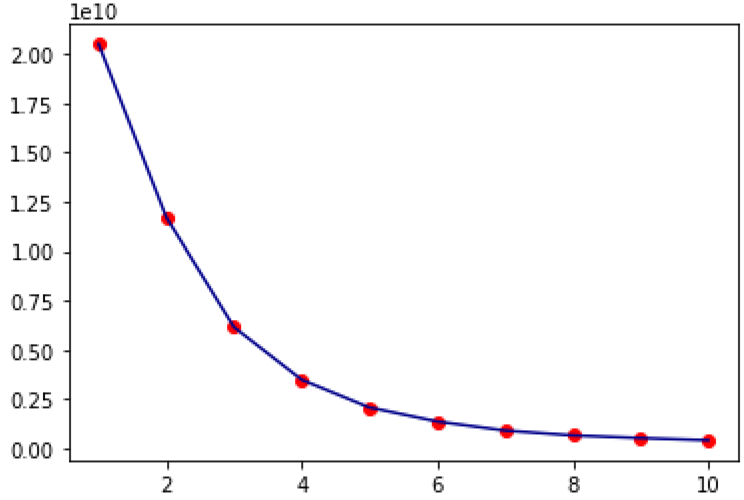


RFM after Log Transformation



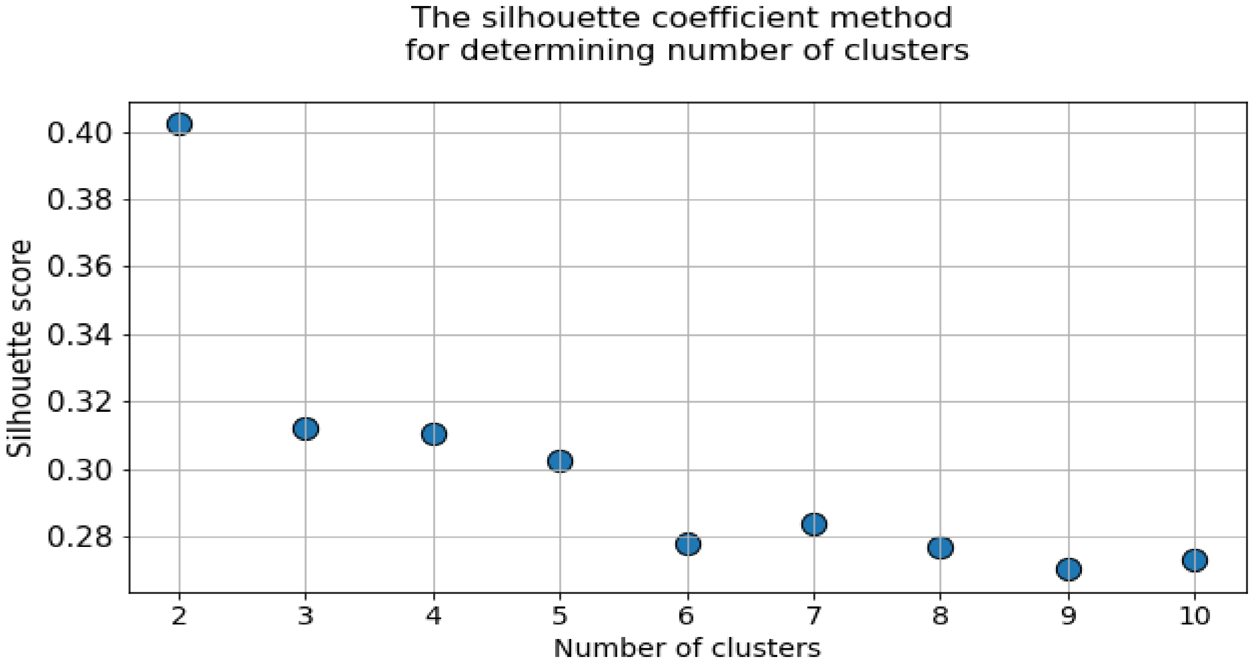
Deciding on number of clusters

To find optimal number of clusters, we are going to use Elbow Method.



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.

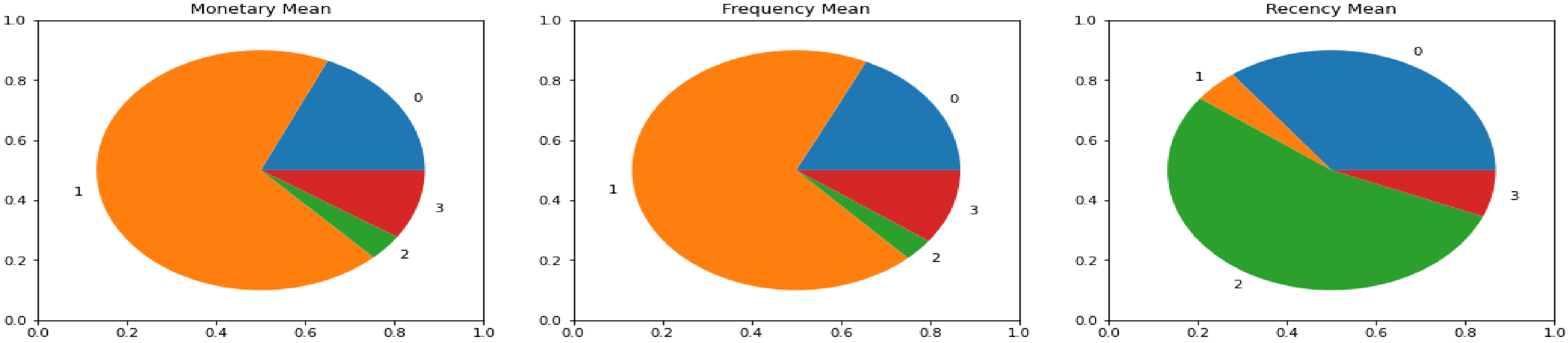
Silhouette Score Method



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 Analysis



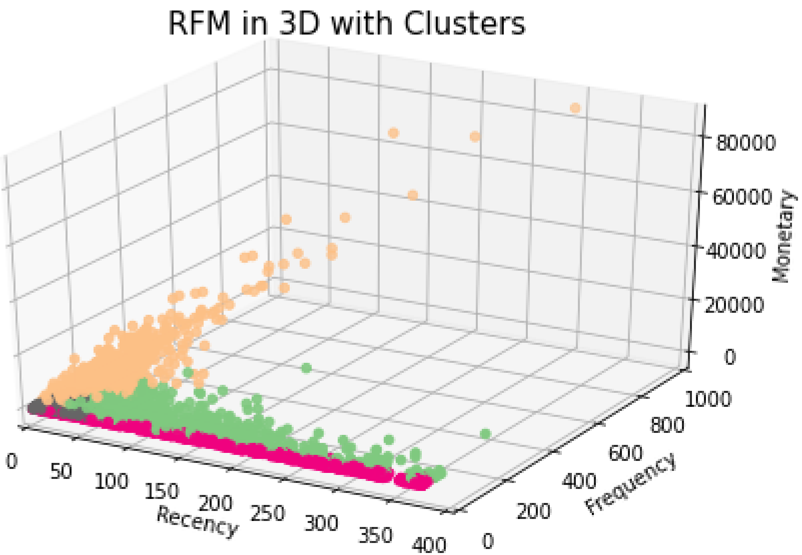
•Group 1 is the group of customers who spends maximum amount of money and also has a good frequency and very low recency rate.

•Group 0 is the group of customers whose frequency rate and monetary value are good and recency rate is also quite good.

•Group 2 the group of customers who has a very high recency rate means they have not purchased anything from the past. 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, and have little higher recency rate compared to group 1 customers.

Continued...



Observations :

1. In the above 3D graph,we put all the three variable into 3 axis and added the cluster variable to differentiate the points.

2. Dark pink points is the group of customers whose Recency is high, Frequency is low and Monetary value is also low.

3. Light pink points are the group of customers whose Recency is low, Frequency is better than grey ones and Monetary is good.

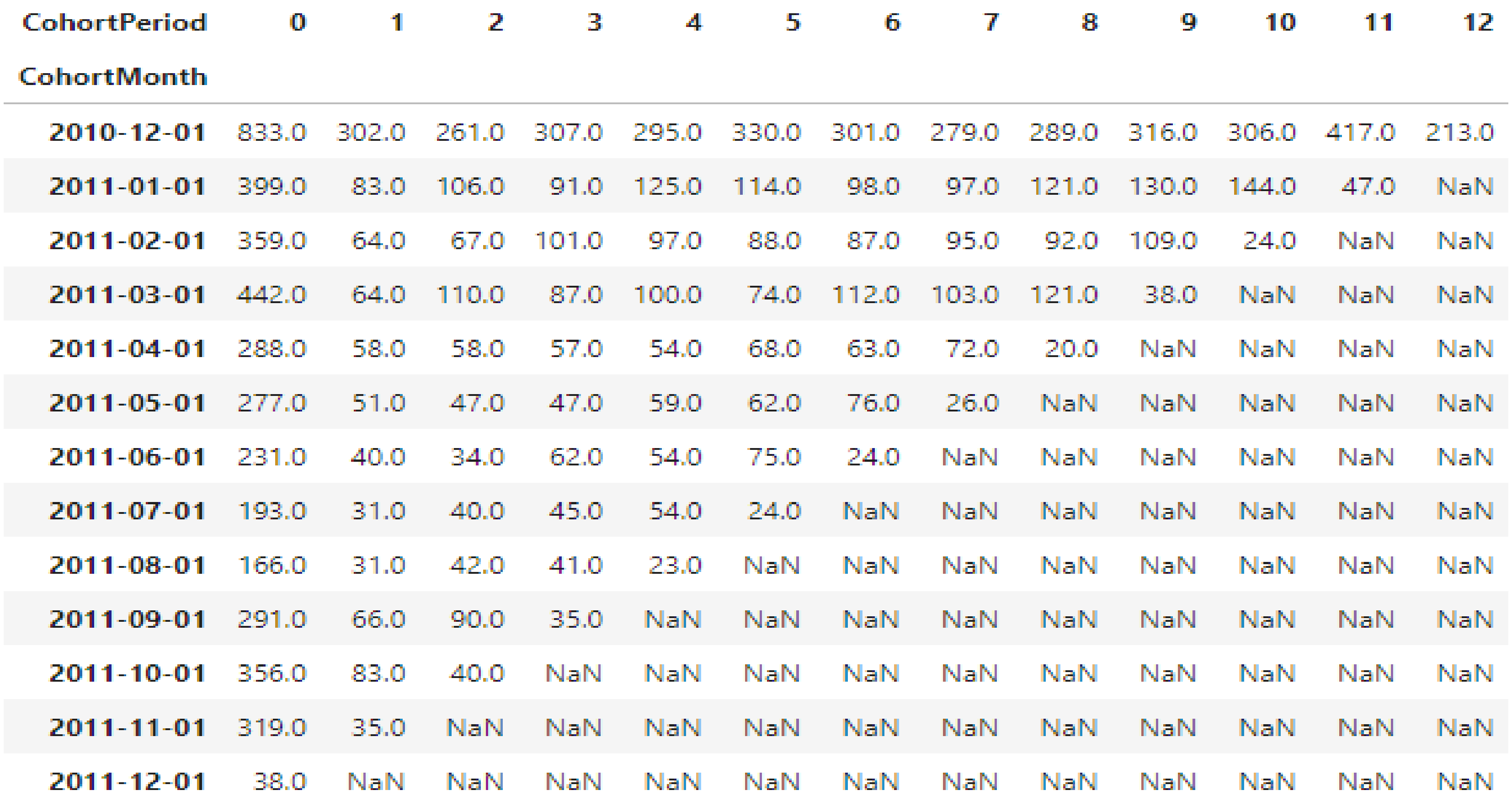
4. Green points are the group of customers whose Recency is high, Frequency and Monetory are better than the dark pink ones.

Cohort Analysis

Types of Cohort –

1. **Time Cohorts** are customers who signed up for a product or service during a particular time frame. Analyzing these cohorts shows the customers' behavior depending on the time they started using the company's products or services. The time may be monthly or quarterly even daily.
2. **Behavior cohorts** are customers who purchased a product or subscribed to a service in the past. It groups customers by the type of product or service they signed up. Customers who signed up for basic level services might have different needs than those who signed up for advanced services. Understaning the needs of the various cohorts can help a company design custom-made services or products for particular segments.
3. **Size cohorts** refer to the various sizes of customers who purchase company's products or services. This categorization can be based on the amount of spending in some periodic time after acquisition or the product type that the customer spent most of their order amount in some period of time

Retention and Acquisition Customers Table



CohortPeriod	0	1	2	3	4	5	6	7	8	9	10	11	12
CohortMonth													
2010-12-01	833.0	302.0	261.0	307.0	295.0	330.0	301.0	279.0	289.0	316.0	306.0	417.0	213.0
2011-01-01	399.0	83.0	106.0	91.0	125.0	114.0	98.0	97.0	121.0	130.0	144.0	47.0	NaN
2011-02-01	359.0	64.0	67.0	101.0	97.0	88.0	87.0	95.0	92.0	109.0	24.0	NaN	NaN
2011-03-01	442.0	64.0	110.0	87.0	100.0	74.0	112.0	103.0	121.0	38.0	NaN	NaN	NaN
2011-04-01	288.0	58.0	58.0	57.0	54.0	68.0	63.0	72.0	20.0	NaN	NaN	NaN	NaN
2011-05-01	277.0	51.0	47.0	47.0	59.0	62.0	76.0	26.0	NaN	NaN	NaN	NaN	NaN
2011-06-01	231.0	40.0	34.0	62.0	54.0	75.0	24.0	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	193.0	31.0	40.0	45.0	54.0	24.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	166.0	31.0	42.0	41.0	23.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	291.0	66.0	90.0	35.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	356.0	83.0	40.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	319.0	35.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	38.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

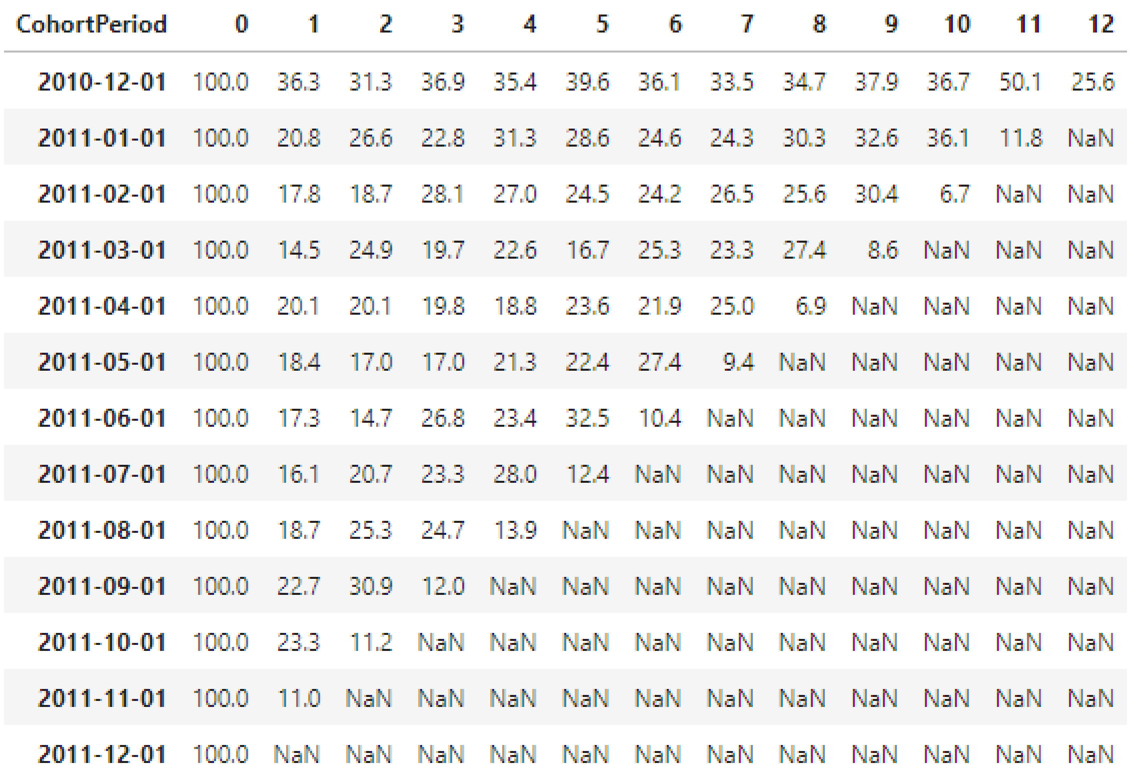
Observations from the table

1. The last table show retention and acquisition of customers.

2. Vertically i.e. the first column '0' tells how many new customers the business acquired in a particular month. ex: 833 is the number of customers business acquired in Dec'2010, 399 is the number of customers(different from previous month) business acquired in Jan'2011, and so on.

3. Horizontally i.e the first row tells the number of customers who is continuing to be part of business since their first purchase i.e. Dec'2010. ex: 302 is the number of customers out of 833 that continue to purchase one month after their first purchase, 261 is the number of customers that continue to purchase two months after their first purchase, and so on.

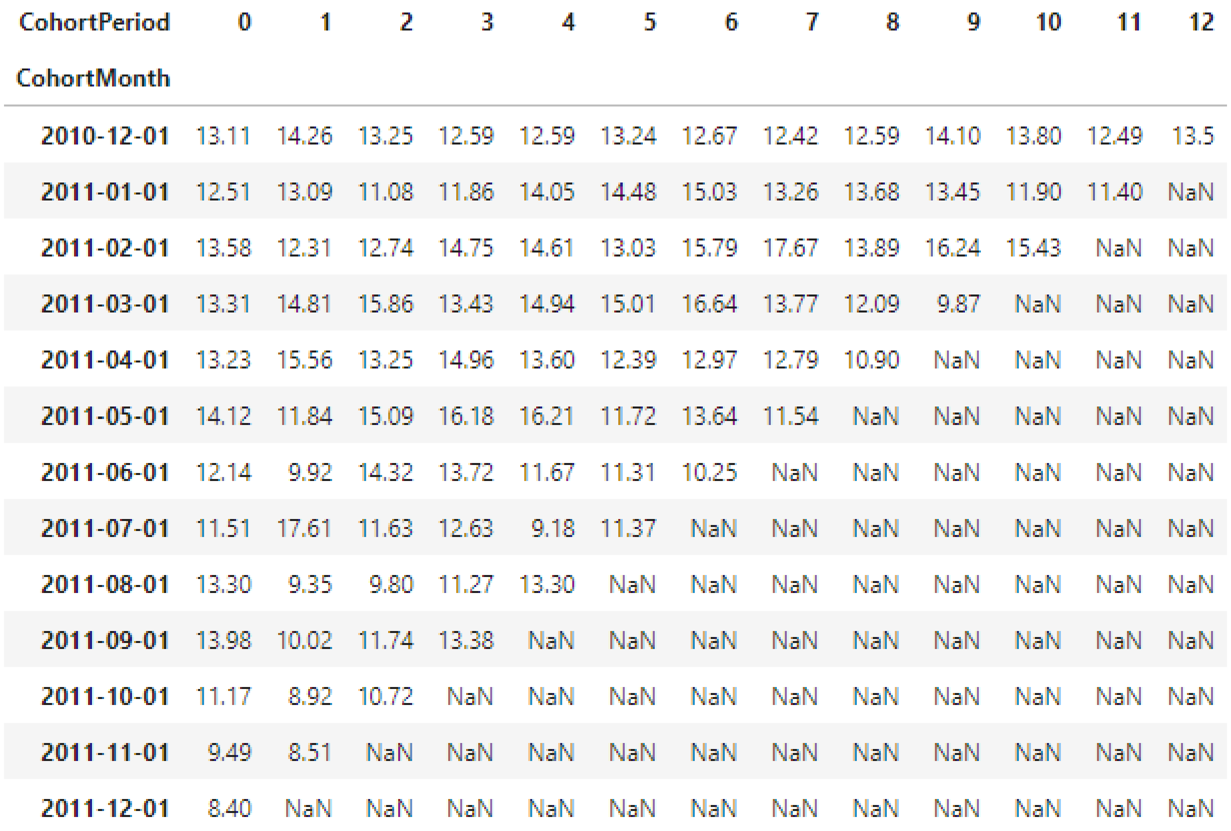
Retention table with percentages



CohortPeriod	0	1	2	3	4	5	6	7	8	9	10	11	12
2010-12-01	100.0	36.3	31.3	36.9	35.4	39.6	36.1	33.5	34.7	37.9	36.7	50.1	25.6
2011-01-01	100.0	20.8	26.6	22.8	31.3	28.6	24.6	24.3	30.3	32.6	36.1	11.8	NaN
2011-02-01	100.0	17.8	18.7	28.1	27.0	24.5	24.2	26.5	25.6	30.4	6.7	NaN	NaN
2011-03-01	100.0	14.5	24.9	19.7	22.6	16.7	25.3	23.3	27.4	8.6	NaN	NaN	NaN
2011-04-01	100.0	20.1	20.1	19.8	18.8	23.6	21.9	25.0	6.9	NaN	NaN	NaN	NaN
2011-05-01	100.0	18.4	17.0	17.0	21.3	22.4	27.4	9.4	NaN	NaN	NaN	NaN	NaN
2011-06-01	100.0	17.3	14.7	26.8	23.4	32.5	10.4	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	100.0	16.1	20.7	23.3	28.0	12.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	100.0	18.7	25.3	24.7	13.9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	100.0	22.7	30.9	12.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	100.0	23.3	11.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	100.0	11.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	100.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

We can see that over the period of time how the customer interact with the business. ex- In Jan'2011 the business acquire some new customers but after one month only 20.8% are retained or say revisit again. Then the number rise to 26.6% which means some customers back and purchase again and the reason could be an invitation/offers is sent to those group of customers.

Average spending by the customers over the time period



CohortPeriod	0	1	2	3	4	5	6	7	8	9	10	11	12
CohortMonth													
2010-12-01	13.11	14.26	13.25	12.59	12.59	13.24	12.67	12.42	12.59	14.10	13.80	12.49	13.5
2011-01-01	12.51	13.09	11.08	11.86	14.05	14.48	15.03	13.26	13.68	13.45	11.90	11.40	NaN
2011-02-01	13.58	12.31	12.74	14.75	14.61	13.03	15.79	17.67	13.89	16.24	15.43	NaN	NaN
2011-03-01	13.31	14.81	15.86	13.43	14.94	15.01	16.64	13.77	12.09	9.87	NaN	NaN	NaN
2011-04-01	13.23	15.56	13.25	14.96	13.60	12.39	12.97	12.79	10.90	NaN	NaN	NaN	NaN
2011-05-01	14.12	11.84	15.09	16.18	16.21	11.72	13.64	11.54	NaN	NaN	NaN	NaN	NaN
2011-06-01	12.14	9.92	14.32	13.72	11.67	11.31	10.25	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	11.51	17.61	11.63	12.63	9.18	11.37	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	13.30	9.35	9.80	11.27	13.30	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	13.98	10.02	11.74	13.38	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	11.17	8.92	10.72	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	9.49	8.51	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	8.40	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

For the group of customers of Jan'2011 they initially spent 12.51 but after one month they spent 13.09 higher than the previous, then they spend 11.08 after two months and so on.

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 spend changes over the time period.

- 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 prefer to buy often, finding out customer lifetime value (clv) and much more.

Challenges

❑ Doing data preprocessing – dealing with null values, there were some transactions which was cancelled and reversed entry was there, deal with such huge dataset was a difficult task

❑ Deciding on which clustering algorithms to use

Thank You