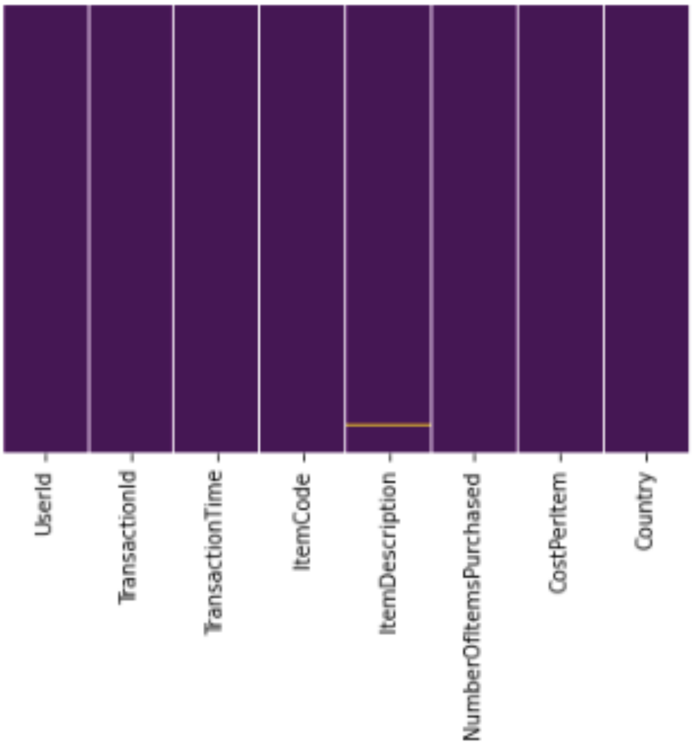
Problem
To explore and identify different segments present in the customer transition data

Process
1. exploratory data analysis
1.1 missing data and corrected it so we will use seaborn library to create a simple heat map to see whether there is any missing data



Conclusion- Roughly 0.27% (2908 out of 1083818) of the ItemDescription data is missing. As this fraction is very low so it's safe to drop these nan rows for now
1.2 Now after removing missing data will look after unknown values (-1) as in data set, I correlated TransactionId with UserId and some of -1 data in UserId was taken care of.

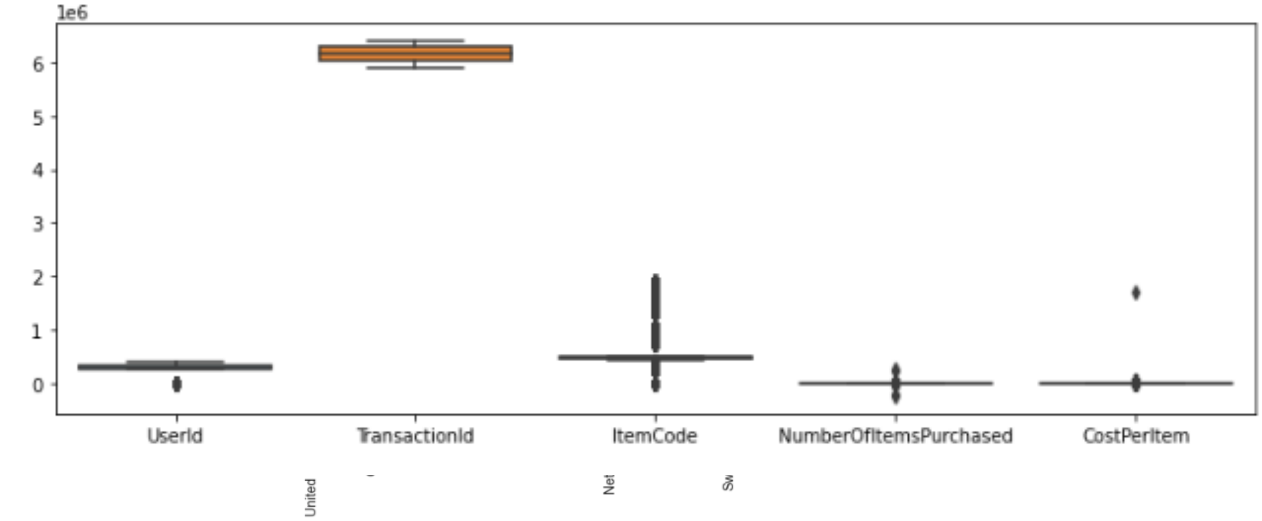
2 **Analysis**

2.1 Analysis on Categorical Attributes

- During analysis we found out that <90% of data is from UK country
- We will group data into two parts one with only US country and rest with others countries

2.2 Analysis of Distribution

- Will see how ItemDescription distribution takes place by dividing the whole column of ItemDescription into low, medium and high using logic 'low' if x<17 else 'high' if x>29 else ‘medium' again which was formed by calculating the mean of the column
-



2.3 Analysis on Numerical Attributes

- CostPerItem has mean of approx 9 value , now there were some negative value which needs to be removed

Along with that I have studied 4 reports to check the correlation, mean , duplicate rows, data distribution.

3. Data Cleaning

3.1 Will clean the missing values
3.2 Drop the duplicates
3.3 Check any resemblance of unknown value, will try to find it because data is very important. If they are still unknown we will drop that row
3.4 Removing negative value data where ever key value can't have value less than zero

Now we will save this final cleaned Dataset

Identifying different segments in the data

Segmentation is a method of dividing customers into groups or clusters on the basis of common characteristics

We will use RFM (Recency, Frequency, Monetary) analysis which is a behaviour-based approach grouping customers into segments.

- Recency (R): Who have purchased recently?
- Frequency (F): Who has purchased frequently?
- Monetary Value(M): Who have high purchase amount?

STEPS

1. Calculate the Recency, Frequency, Monetary values for each customer.
2.Add segment bin values to RFM table using quartile
3. Concate all scores in single column(RFM_Score).

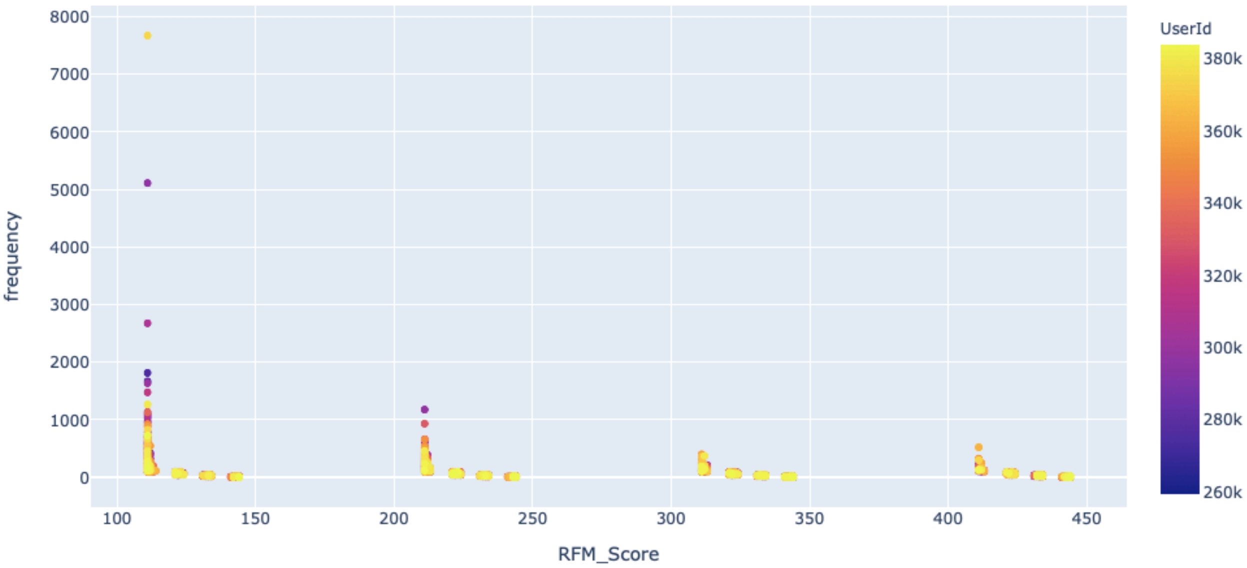
Now we will look how RFM Analysis works
Here, you are going to perform following opertaions:

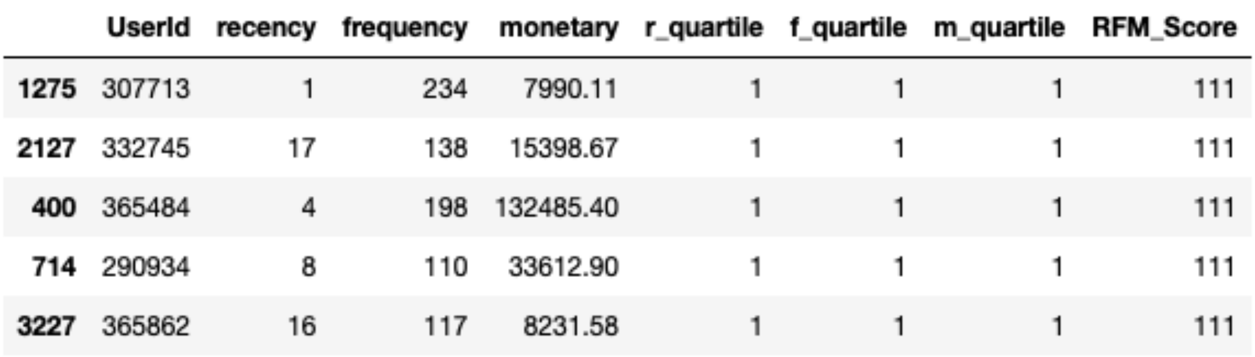
- For Recency, Calculate the number of days between present date and date of last purchase each customer.
- For Frequency, Calculate the number of orders for each customer.
- For Monetary, Calculate sum of purchase price for each customer.

Customers with the lowest recency, highest frequency and monetary amounts considered as top customers.

Results

RFM_Score with 111 represents the most valuable customer I have plotted the scatter plot also (you can go to the point and it will give you the details UserID, and formed groups which represents similar kind of customers
as you can see below (example) and combined CSV file is attached.





	UserId	recency	frequency	monetary	r_quartile	f_quartile	m_quartile	RFM_Score
1275	307713	1	234	7990.11	1	1	1	111
2127	332745	17	138	15398.67	1	1	1	111
400	365484	4	198	132485.40	1	1	1	111
714	290934	8	110	33612.90	1	1	1	111
3227	365862	16	117	8231.58	1	1	1	111