In [1]:
```
import pandas as pd
import matplotlib as plt
import datetime as dt
```

In [2]:
```
data_copy = pd.read_csv(r"D:\Academics\DS\Semster 3\Data M\scanner_data.csv")
data = data_copy.copy()
```

In [3]:
```
type(data)
```

Out[3]:
```
pandas.core.frame.DataFrame
```

In [4]:
```
data.head(10)
```

Out[4]:

	Unnamed: 0	date	id_cliente	Transaction_ID	SKU_Category	SKU	Quantity	Price
0	1	2/1/2016	2547	1	X52	0EM7L	1.0	3.13
1	2	2/1/2016	822	2	2ML	68BRQ	1.0	5.46
2	3	2/1/2016	3686	3	0H2	CZUZX	1.0	6.35
3	4	2/1/2016	3719	4	0H2	549KK	1.0	5.59
4	5	2/1/2016	9200	5	0H2	K8EHH	1.0	6.88
5	6	2/1/2016	5010	6	JPI	GVBRC	1.0	10.77
6	7	2/1/2016	1666	7	XG4	AHAE7	1.0	3.65
7	8	2/1/2016	1666	7	FEW	AHZNS	1.0	8.21
8	9	2/1/2016	1253	8	0H2	9STQJ	1.0	8.25
9	10	2/1/2016	5541	9	N5F	7IE9S	1.0	8.18

In [5]:
```
data.tail(10)
```

Out[5]:

	Unnamed: 0	date	id_cliente	Transaction_ID	SKU_Category	SKU	Quantity	Price
131696	131697	4/7/2016	10468	32899	1VL	USW0M	1.0	5.87
131697	131698	4/7/2016	10468	32899	RML	EDZ1Y	1.0	6.07
131698	131699	4/7/2016	10468	32899	LSD	4AZHE	1.0	7.73
131699	131700	4/7/2016	20203	32900	J4R	LI0IX	1.0	6.25
131700	131701	4/7/2016	20203	32900	U5F	F7FQ5	3.0	7.27
131701	131702	4/7/2016	20203	32900	IEV	FO112	3.0	6.46
131702	131703	4/7/2016	20203	32900	N8U	I36F2	1.0	4.50
131703	131704	4/7/2016	20203	32900	U5F	4X8P4	1.0	5.19
131704	131705	4/7/2016	20203	32900	0H2	ZVTO4	1.0	4.57
131705	131706	4/7/2016	20203	32900	Q4N	QM9BP	1.0	13.68

In [6]:
```
data = data[data["Price"] > 0]
data["date"]= pd.to_datetime(data["date"])
```

In [7]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 131706 entries, 0 to 131705
Data columns (total 8 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Unnamed: 0      131706 non-null  int64
 1   date            131706 non-null  datetime64[ns]
 2   id_cliente      131706 non-null  int64
 3   Transaction_ID  131706 non-null  int64
 4   SKU_Category    131706 non-null  object
 5   SKU             131706 non-null  object
 6   Quantity        131706 non-null  float64
 7   Price           131706 non-null  float64
dtypes: datetime64[ns](1), float64(2), int64(3), object(2)
memory usage: 9.0+ MB
```

In [8]:
```
from datetime import timedelta
```

```
present  = data["date"].max() + timedelta(1)
```

In [9]:
```
group = data.groupby("id_cliente")
```

In [10]:
```
rfm = group.agg({"date": lambda x:(present-x.max()).days,
                 "Price": lambda x:x.sum()})
```

In [11]:
```
rfm
```

Out[11]:

id_cliente	date	Price
1	345	16.29
2	196	22.77
3	335	10.92
4	55	33.29
5	121	78.82
...	...	...
22621	81	9.69
22622	16	6.07
22623	11	128.01
22624	324	19.60
22625	1	83.62

22625 rows × 2 columns

In [12]:
```
s = data["id_cliente"].value_counts()
ss= s.sort_index()
```

```
rfm["frequency"] = ss
rfm
```

Out[12]:

id_cliente	date	Price	frequency
1	345	16.29	2
2	196	22.77	2
3	335	10.92	3
4	55	33.29	5
5	121	78.82	5
...	...	...	...
22621	81	9.69	2
22622	16	6.07	1
22623	11	128.01	2
22624	324	19.60	2
22625	1	83.62	9

22625 rows × 3 columns

In [13]:

```
rfm.columns = ['recency','monetary','frequency']
rfm
```

Out[13]:

id_cliente	recency	monetary	frequency
1	345	16.29	2
2	196	22.77	2
3	335	10.92	3
4	55	33.29	5

id_cliente	recency	monetary	frequency
5	121	78.82	5
...	...	...	...
22621	81	9.69	2
22622	16	6.07	1
22623	11	128.01	2
22624	324	19.60	2
22625	1	83.62	9

22625 rows × 3 columns

In [ ]:

```
```

In [14]:

```
rfm
```

Out[14]:

id_cliente	recency	monetary	frequency
1	345	16.29	2
2	196	22.77	2
3	335	10.92	3
4	55	33.29	5
5	121	78.82	5
...	...	...	...
22621	81	9.69	2
22622	16	6.07	1
22623	11	128.01	2

id_cliente	recency	monetary	frequency
22624	324	19.60	2
22625	1	83.62	9

22625 rows × 3 columns

In [15]:
```
rfm['frequency'].value_counts()
```

Out[15]:
```
1      6301
2      4363
3      2591
4      1800
5      1332
       ...
105       1
106       1
218       1
91        1
191       1
Name: frequency, Length: 118, dtype: int64
```

In [16]:
```
rfm["frequency_modified"] = rfm['frequency'].rank(method = 'first')
```

In [17]:
```
rfm["frequency_modified"]
```

Out[17]:
```
id_cliente
1         6302.0
2         6303.0
3        10665.0
4        15056.0
5        15057.0
          ...
22621    10662.0
22622     6301.0
22623    10663.0
22624    10664.0
22625    19226.0
Name: frequency_modified, Length: 22625, dtype: float64
```

In [18]:
```
rfm["frequency_modified"].sort_values()
```

Out[18]:
```
id_cliente
7            1.0
8            2.0
9            3.0
10           4.0
11           5.0
           ...
16905    22621.0
1685     22622.0
17104    22623.0
1665     22624.0
1660     22625.0
Name: frequency_modified, Length: 22625, dtype: float64
```

In [19]:
```
rfm['R_quartile'] = pd.qcut(rfm['recency'], 4,['1','2','3','4'])
rfm['M_quartile'] = pd.qcut(rfm['monetary'], 4,['4','3','2','1'])
rfm['f_quartile'] = pd.qcut(rfm['frequency_modified'], 4,['4','3','2','1'])
```

In [20]:
```
rfm = rfm[['recency','monetary','frequency','frequency_modified','R_quartile','M_quartile','f_quartile']]
```

In [21]:
```
rfm
```

Out[21]:

id_cliente	recency	monetary	frequency	frequency_modified	R_quartile	M_quartile	f_quartile
1	345	16.29	2	6302.0	4	3	3
2	196	22.77	2	6303.0	3	3	3
3	335	10.92	3	10665.0	4	3	3
4	55	33.29	5	15056.0	2	2	2
5	121	78.82	5	15057.0	2	1	2
...	...	...	...	...	...	...	...
22621	81	9.69	2	10662.0	2	4	3
22622	16	6.07	1	6301.0	1	4	3

id_cliente	recency	monetary	frequency	frequency_modified	R_quartile	M_quartile	f_quartile
22623	11	128.01	2	10663.0	1	1	3
22624	324	19.60	2	10664.0	4	3	3
22625	1	83.62	9	19226.0	1	1	1

22625 rows × 7 columns

In [22]:
```
rfm2=rfm.copy()
rfm2
```

Out[22]:

id_cliente	recency	monetary	frequency	frequency_modified	R_quartile	M_quartile	f_quartile
1	345	16.29	2	6302.0	4	3	3
2	196	22.77	2	6303.0	3	3	3
3	335	10.92	3	10665.0	4	3	3
4	55	33.29	5	15056.0	2	2	2
5	121	78.82	5	15057.0	2	1	2
...	...	...	...	...	...	...	...
22621	81	9.69	2	10662.0	2	4	3
22622	16	6.07	1	6301.0	1	4	3
22623	11	128.01	2	10663.0	1	1	3
22624	324	19.60	2	10664.0	4	3	3
22625	1	83.62	9	19226.0	1	1	1

22625 rows × 7 columns

In [23]:
```
rfm2=rfm[rfm.f_quartile=='3']
```

```
rfm2
```

Out[23]:

id_cliente	recency	monetary	frequency	frequency_modified	R_quartile	M_quartile	f_quartile
1	345	16.29	2	6302.0	4	3	3
2	196	22.77	2	6303.0	3	3	3
3	335	10.92	3	10665.0	4	3	3
6	276	25.55	3	10666.0	4	2	3
13	60	53.24	3	10667.0	2	2	3
...	...	...	...	...	...	...	...
22620	356	8.60	1	6300.0	4	4	3
22621	81	9.69	2	10662.0	2	4	3
22622	16	6.07	1	6301.0	1	4	3
22623	11	128.01	2	10663.0	1	1	3
22624	324	19.60	2	10664.0	4	3	3

5656 rows × 7 columns

In [24]:

```
rfm2
```

Out[24]:

id_cliente	recency	monetary	frequency	frequency_modified	R_quartile	M_quartile	f_quartile
1	345	16.29	2	6302.0	4	3	3
2	196	22.77	2	6303.0	3	3	3
3	335	10.92	3	10665.0	4	3	3
6	276	25.55	3	10666.0	4	2	3
13	60	53.24	3	10667.0	2	2	3
...	...	...	...	...	...	...	...

id_cliente	recency	monetary	frequency	frequency_modified	R_quartile	M_quartile	f_quartile
22620	356	8.60	1	6300.0	4	4	3
22621	81	9.69	2	10662.0	2	4	3
22622	16	6.07	1	6301.0	1	4	3
22623	11	128.01	2	10663.0	1	1	3
22624	324	19.60	2	10664.0	4	3	3

5656 rows × 7 columns

In [ ]:

In [ ]: