Exploratory Data Analysis

1. Data Segmentation
2. Data Visualisation

In [24]:

```
import pandas as pd
import numpy as np
import seaborn as sns
```

In [25]:

```
#To load the data
data = pd.read_excel('Dataset EDA.xlsx')
```

In [26]:

```
data.head()
```

Out[26]:

	status	card_present_flag	bpay_biller_code	account	currency	long_lat	txn_description	merchant_id	merchant_c
0	authorized	1.0	NaN	ACC-1598451071	AUD	153.41 -27.95	POS	81c48296-73be-44a7-befa-d053f48ce7cd	
1	authorized	0.0	NaN	ACC-1598451071	AUD	153.41 -27.95	SALES-POS	830a451c-316e-4a6a-bf25-e37caedca49e	
2	authorized	1.0	NaN	ACC-1222300524	AUD	151.23 -33.94	POS	835c231d-8cdf-4e96-859d-e9d571760cf0	
3	authorized	1.0	NaN	ACC-1037050564	AUD	153.10 -27.66	SALES-POS	48514682-c78a-4a88-b0da-2d6302e64673	
4	authorized	1.0	NaN	ACC-1598451071	AUD	153.41 -27.95	SALES-POS	b4e02c10-0852-4273-b8fd-7b3395e32eb0	

5 rows × 23 columns

In [27]:

```
data.describe()
```

Out[27]:

	card_present_flag	merchant_code	balance	age	amount
count	7717.000000	883.0	12043.000000	12043.000000	12043.000000
mean	0.802644	0.0	14704.195553	30.582330	187.933588
std	0.398029	0.0	31503.722652	10.046343	592.599934
min	0.000000	0.0	0.240000	18.000000	0.100000
25%	1.000000	0.0	3158.585000	22.000000	16.000000

	card_present_flag	merchant_code	balance	age	amount
50%	1.000000	0.0	6432.010000	28.000000	29.000000
75%	1.000000	0.0	12465.945000	38.000000	53.655000
max	1.000000	0.0	267128.520000	78.000000	8835.980000

In [28]:

```
data.shape
```

Out[28]:

```
(12043, 23)
```

In [29]:

```
#To get average value of the 'amount'
a = data['amount']
avg_transaction_amt = a.sum()/a.count()
print(avg_transaction_amt)
```

```
187.93358797641784
```

In [30]:

```
data.columns
```

Out[30]:

```
Index(['status', 'card_present_flag', 'bpay_biller_code', 'account',
       'currency', 'long_lat', 'txn_description', 'merchant_id',
       'merchant_code', 'first_name', 'balance', 'date', 'gender', 'age',
       'merchant_suburb', 'merchant_state', 'extraction', 'amount',
       'transaction_id', 'country', 'customer_id', 'merchant_long_lat',
       'movement'],
      dtype='object')
```

In [31]:

```
#for Unique values
data.nunique()
```

Out[31]:

```
status                2
card_present_flag     2
bpay_biller_code      3
account             100
currency              1
long_lat            100
txn_description       6
merchant_id        5725
merchant_code         1
first_name           80
balance           12006
date                 91
gender                2
age                  33
merchant_suburb    1609
merchant_state        8
extraction         9442
amount             4457
transaction_id    12043
country               1
customer_id         100
merchant_long_lat  2703
movement              2
dtype: int64
```

In [32]:

```
#To get Unique value in 'movement' column
data['movement'].unique()
```

Out[32]:

```
array(['debit', 'credit'], dtype=object)
```

In [33]:

```
#for NULL value
data.isnull().sum()
```

Out[33]:

```
status                   0
card_present_flag     4326
bpay_biller_code     11158
account                  0
currency                 0
long_lat                 0
txn_description          0
merchant_id           4326
merchant_code        11160
first_name               0
balance                  0
date                     0
gender                   0
age                      0
merchant_suburb       4326
merchant_state        4326
extraction               0
amount                   0
transaction_id           0
country                  0
customer_id              0
merchant_long_lat     4326
movement                 0
dtype: int64
```

In [34]:

```
#To Drop 'merchant_code','country' & 'currency' from the data

fdata = data.drop(['merchant_code','country','currency'], axis= 1)
```

In [35]:

```
fdata.head()
```

Out[35]:

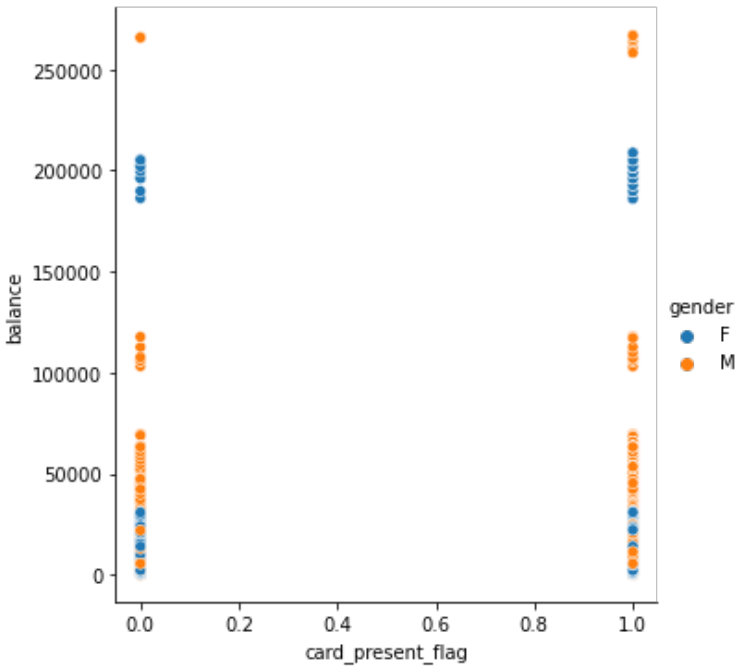
	status	card_present_flag	bpay_biller_code	account	long_lat	txn_description	merchant_id	first_name	balance
0	authorized	1.0	NaN	ACC-1598451071	153.41 -27.95	POS	81c48296-73be-44a7-befa-d053f48ce7cd	Diana	35.39
1	authorized	0.0	NaN	ACC-1598451071	153.41 -27.95	SALES-POS	830a451c-316e-4a6a-bf25-e37caedca49e	Diana	21.20
2	authorized	1.0	NaN	ACC-1222300524	151.23 -33.94	POS	835c231d-8cdf-4e96-859d-e9d571760cf0	Michael	5.71
3	authorized	1.0	NaN	ACC-1037050564	153.10 -27.66	SALES-POS	48514682-c78a-4a88-b0da-2d6302e64673	Rhonda	2117.22
4	authorized	1.0	NaN	ACC-1598451071	153.41 -27.95	SALES-POS	b4e02c10-0852-4273-b8fd-7b3395e32eb0	Diana	17.95

In [36]:

```
#Relational Plot - card_present_flag Vs balance
sns.relplot(x='card_present_flag', y='balance', hue = 'gender', data = fdata)
```

Out[36]:

```
<seaborn.axisgrid.FacetGrid at 0x24a81afe2b0>
```



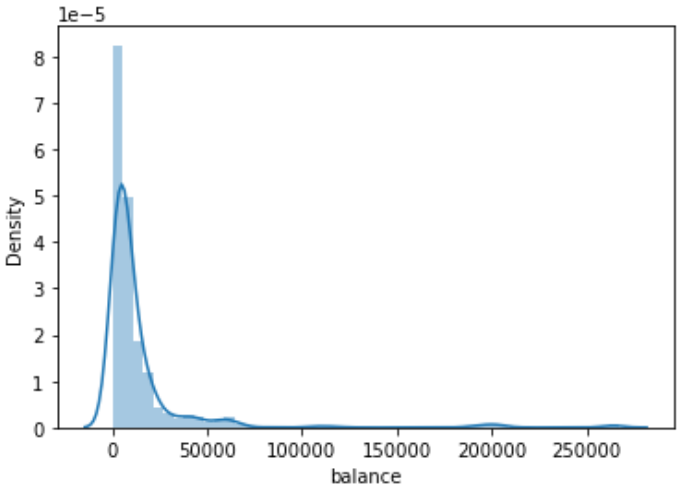
In [37]:

```
#Distribution Plot - 'balance'
sns.distplot(fdata['balance'])
```

```
C:\Users\Gururaj K\anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarni
ng: `distplot` is a deprecated function and will be removed in a future version. Please a
dapt your code to use either `displot` (a figure-level function with similar flexibility)
or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[37]:

```
<AxesSubplot:xlabel='balance', ylabel='Density'>
```



In [38]:

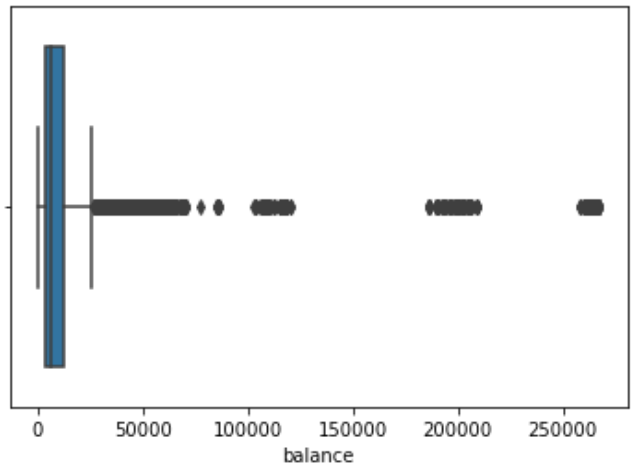
```
#Box Plot - 'balance'
sns.boxplot(fdata['balance'])
```

```
C:\Users\Gururaj K\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid positi
onal argument will be `data`, and passing other arguments without an explicit keyword wil
```

```
l result in an error or misinterpretation.
  warnings.warn(
```

Out[38]:

```
<AxesSubplot:xlabel='balance'>
```



In [39]:

```
#Distribution Plot - 'age'
sns.distplot(fdata['age'])
```

```
C:\Users\Gururaj K\anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarni
ng: `distplot` is a deprecated function and will be removed in a future version. Please a
dapt your code to use either `displot` (a figure-level function with similar flexibility)
or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[39]:

```
<AxesSubplot:xlabel='age', ylabel='Density'>
```