In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
data=pd.read_csv("Churn_model.csv")
```

In [3]:
```
data.shape
```

Out[3]: (10000, 15)

In [5]:
```
data.head()
```

Out[5]:

| | Unnamed: 0 | RowNumber | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | Esti |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 15634602 | Hargrave | 619.0 | France | Female | 42 | 2 | 0.00 | 1 | 1 | 1 | |
| 1 | 1 | 2 | 15647311 | Hill | 608.0 | Spain | Female | 41 | 1 | 83807.86 | 1 | 0 | 1 | |
| 2 | 2 | 3 | 15619304 | Onio | 502.0 | France | Female | 42 | 8 | 159660.80 | 3 | 1 | 0 | |
| 3 | 3 | 4 | 15701354 | Boni | 699.0 | France | Female | 39 | 1 | 0.00 | 2 | 0 | 0 | |
| 4 | 4 | 5 | 15737888 | Mitchell | 850.0 | Spain | Female | 43 | 2 | 125510.82 | 1 | 1 | 1 | |

In [6]:
```
data.head(11)
```

Out[6]:

| | Unnamed: 0 | RowNumber | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | Est |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 15634602 | Hargrave | 619.0 | France | Female | 42 | 2 | 0.00 | 1 | 1 | 1 | |
| 1 | 1 | 2 | 15647311 | Hill | 608.0 | Spain | Female | 41 | 1 | 83807.86 | 1 | 0 | 1 | |
| 2 | 2 | 3 | 15619304 | Onio | 502.0 | France | Female | 42 | 8 | 159660.80 | 3 | 1 | 0 | |
| 3 | 3 | 4 | 15701354 | Boni | 699.0 | France | Female | 39 | 1 | 0.00 | 2 | 0 | 0 | |
| 4 | 4 | 5 | 15737888 | Mitchell | 850.0 | Spain | Female | 43 | 2 | 125510.82 | 1 | 1 | 1 | |
| 5 | 5 | 6 | 15574012 | Chu | 645.0 | Spain | Male | 44 | 8 | 113755.78 | 2 | 1 | 0 | |
| 6 | 6 | 7 | 15592531 | Bartlett | 822.0 | France | Male | 50 | 7 | 0.00 | 2 | 1 | 1 | |
| 7 | 7 | 8 | 15656148 | Obinna | 376.0 | Germany | Female | 29 | 4 | 115046.74 | 4 | 1 | 0 | |
| 8 | 8 | 9 | 15792365 | He | 501.0 | France | Male | 44 | 4 | 142051.07 | 2 | 0 | 1 | |
| 9 | 9 | 10 | 15592389 | H? | 684.0 | France | Male | 27 | 2 | 134603.88 | 1 | 1 | 1 | |
| 10 | 10 | 11 | 15767821 | Bearce | 528.0 | France | Male | 31 | 6 | 102016.72 | 2 | 0 | 0 | |

In [7]:
```
formatted_data=data.drop(data.columns[[0,1]],axis=1)
formatted_data.head()
```

Out[7]:

| | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 15634602 | Hargrave | 619.0 | France | Female | 42 | 2 | 0.00 | 1 | 1 | 1 | 101348.88 | 1 |
| 1 | 15647311 | Hill | 608.0 | Spain | Female | 41 | 1 | 83807.86 | 1 | 0 | 1 | 112542.58 | 0 |
| 2 | 15619304 | Onio | 502.0 | France | Female | 42 | 8 | 159660.80 | 3 | 1 | 0 | 113931.57 | 1 |
| 3 | 15701354 | Boni | 699.0 | France | Female | 39 | 1 | 0.00 | 2 | 0 | 0 | 93826.63 | 0 |
| 4 | 15737888 | Mitchell | 850.0 | Spain | Female | 43 | 2 | 125510.82 | 1 | 1 | 1 | 79084.10 | 0 |

In [10]:
```
formatted_data.shape
```

Out[10]: (10000, 13)

In [12]:
```
formatted_data.dtypes
```

Out[12]:
```
CustomerId           int64
Surname             object
CreditScore        float64
Geography           object
Gender              object
Age                  int64
Tenure               int64
Balance            float64
NumOfProducts        int64
HasCrCard            int64
IsActiveMember       int64
EstimatedSalary    float64
Exited               int64
dtype: object
```

In [14]:
```
formatted_data.isnull().sum()
```

Out[14]:
```
CustomerId           0
Surname              0
CreditScore        250
Geography            0
Gender               0
Age                  0
Tenure               0
Balance            250
NumOfProducts        0
HasCrCard            0
IsActiveMember       0
EstimatedSalary      0
Exited               0
dtype: int64
```

In [15]:
```
null_values=formatted_data['CreditScore'].isnull().sum()+ formatted_data['Balance'].isnull().sum()
print(null_values)
```

```
500
```

In [19]:
```
missing_percent=round((null_values/(10000*13)*100),2)
print( '% Missing values =',missing_percent)
```

```
% Missing values = 0.38
```

In [23]:
```
newdata=formatted_data.dropna()
newdata.shape
#only rows delete by this
```

Out[23]:
```
(9505, 13)
```

In [25]:
```
available_records=100-missing_percent
print('Available Records=',available_records)
```

```
Available Records= 99.62
```

In [31]:
```
prob=(available_records/100)**7
print('Probability of record not containing missing =',prob)
print('percentage of records remove due to missing value=',round((1-prob)*100,2))
```

```
Probability of record not containing missing = 0.9737013267613582
percentage of records remove due to missing value= 2.63
```

In [32]:
```
#if there is small fraction of data is null in data , by removing records thre will be large impact in loss of data, more no. of
```

In [33]:
```
#replacing missing values with median
#why median? outliers maybe exist thats why ..if thre will be no outliers then we take mean ..if thre is not needdded (farak ni p
```

In [39]:
```
median1=formatted_data['CreditScore'].median()
median2=formatted_data['Balance'].median()
```

In [44]:
```
formatted_data['CreditScore']=formatted_data['CreditScore'].fillna(value=median1)
formatted_data['Balance']=formatted_data['Balance'].fillna(value=median2)
formatted_data.shape
```

Out[44]:
```
(10000, 13)
```

In [ ]: