

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Loading Data

In [2]:

```
from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive
```

In [3]:

```
df_teamlead = pd.read_csv("/content/drive/MyDrive/Churn_Modelling.csv")
df_teamlead.head()
```

Out[3]:

	RowN umbe r	Custo merI d	Sur nam e	Credi tScor e	Geog raph y	Ge nde r	A g e	Te nur e	Bala nce	NumOf Product s	HasC rCar d	IsActive Membe r	Estimat edSalar y	Ex ite d
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Descriptive Statistics

In [4]:

```
df_teamlead.describe()
```

Out[4]:

	RowN umber	Custo merId	Credit Score	Age	Tenur e	Balanc e	NumOf Product s	HasC rCard	IsActive Member	Estimate dSalary	Exited
co un t	10000. 00000	1.0000 00e+04	10000. 000000	10000. 000000	10000. 000000	10000.0 00000	10000.00 0000	10000. 00000	10000.00 0000	10000.00 0000	10000. 000000
me an	5000.5 0000	1.5690 94e+07	650.52 8800	38.921 800	5.0128 00	76485.8 89288	1.530200	0.7055 0	0.515100	100090.2 39881	0.2037 00
std	2886.8 9568	7.1936 19e+04	96.653 299	10.487 806	2.8921 74	62397.4 05202	0.581654	0.4558 4	0.499797	57510.49 2818	0.4027 69
mi n	1.0000 0	1.5565 70e+07	350.00 0000	18.000 000	0.0000 00	0.00000 0	1.000000	0.0000 0	0.000000	11.58000 0	0.0000 00
25 %	2500.7 5000	1.5628 53e+07	584.00 0000	32.000 000	3.0000 00	0.00000 0	1.000000	0.0000 0	0.000000	51002.11 0000	0.0000 00
50 %	5000.5 0000	1.5690 74e+07	652.00 0000	37.000 000	5.0000 00	97198.5 40000	1.000000	1.0000 0	1.000000	100193.9 15000	0.0000 00
75 %	7500.2 5000	1.5753 23e+07	718.00 0000	44.000 000	7.0000 00	127644. 240000	2.000000	1.0000 0	1.000000	149388.2 47500	0.0000 00
m ax	10000. 00000	1.5815 69e+07	850.00 0000	92.000 000	10.000 000	250898. 090000	4.000000	1.0000 0	1.000000	199992.4 80000	1.0000 00

In [5]:

```
df_teamlead[["NumOfProducts" ,"HasCrCard", "IsActiveMember"]] =  
df_teamlead[["NumOfProducts" ,"HasCrCard", "IsActiveMember"]].astype(str)
```

In []:

```
df_teamlead.info()  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 10000 entries, 0 to 9999  
Data columns (total 14 columns):  
#   Column                Non-Null Count  Dtype  
---  -  
0   RowNumber             10000 non-null  int64  
1   CustomerId            10000 non-null  int64  
2   Surname                10000 non-null  object  
3   CreditScore            10000 non-null  int64  
4   Geography              10000 non-null  object  
5   Gender                 10000 non-null  object
```

```
6   Age          10000 non-null  int64
7   Tenure       10000 non-null  int64
8   Balance      10000 non-null  float64
9   NumOfProducts 10000 non-null  object
10  HasCrCard    10000 non-null  object
11  IsActiveMember 10000 non-null  object
12  EstimatedSalary 10000 non-null float64
13  Exited       10000 non-null  int64
dtypes: float64(2), int64(6), object(6)
memory usage: 1.1+ MB
```

In []:

```
df_teamlead.drop("RowNumber",axis=1,inplace=True)
```

Visualization

In []:

```
plt.figure(figsize=(8,8))
plt.hist(df_teamlead.EstimatedSalary,bins=int(np.sqrt(len(list(df_teamlead.EstimatedSalary))))))
plt.show()
```

In []:

```
df2_tm = df_teamlead.sort_values(by="EstimatedSalary")
df2_tm.head()
```

Out[]:

	CustomerId	Surname	Credit Score	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
2362	15791053	Lucciano	709	Germany	Male	45	4	122917.71	1	1	1	11.58	1
9647	15679693	Walker	625	France	Male	31	5	0.00	2	0	1	90.07	0
9010	15786463	Hsing	645	Germany	Female	59	8	121669.93	2	0	0	91.75	1
5548	15639662	Phillips	710	France	Male	38	2	0.00	2	1	0	96.27	0
4989	15602851	Ozoma	629	France	Male	40	9	0.00	1	1	0	106.67	0

In []:

```
sns.barplot(y=
"EstimatedSalary",x="Gender",hue="HasCrCard",data=df2_tm[:100])
```

Out[]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4cd3b981d0>
```

In []:

```
sns.lmplot(x = "EstimatedSalary",y =
"Balance",hue="Gender",data=df2_tm[:100])
```

Out[]:

```
<seaborn.axisgrid.FacetGrid at 0x7f4cd3b98d10>
```

Handling Missing Value

There is no missing value so no need to do any preprocessing for null values

In []:

```
df_teamlead.isna().sum()
```

Out[]:

```
CustomerId      0
Surname         0
CreditScore     0
Geography       0
Gender          0
Age            0
Tenure         0
Balance         0
NumOfProducts  0
HasCrCard       0
IsActiveMember  0
EstimatedSalary 0
Exited          0
dtype: int64
```

In []:

```
num = df_teamlead.select_dtypes(include=["float64","int64"])
num.head()
```

Out[]:

	CustomerId	CreditScore	Age	Tenure	Balance	EstimatedSalary	Exited
0	15634602	619	42	2	0.00	101348.88	1
1	15647311	608	41	1	83807.86	112542.58	0
2	15619304	502	42	8	159660.80	113931.57	1

	CustomerId	CreditScore	Age	Tenure	Balance	EstimatedSalary	Exited
3	15701354	699	39	1	0.00	93826.63	0
4	15737888	850	43	2	125510.82	79084.10	0

Outlier Remover

In []:

```
figure = plt.figure(figsize = (8,10))
for i,j in enumerate(num.columns):
    plt.subplot(3,3,i+1)
    sns.boxplot(num[j])
    plt.title(j)
```

In []:

```
df_teamlead.head()
```

Out[]:

	Custo merId	Surn ame	Credit Score	Geogr aphy	Gen der	A ge	Ten ure	Bala nce	NumOfP roducts	HasCr Card	IsActive Member	Estimate dSalary	Exi ted
0	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

In []:

```
df_teamlead.size
```

Out[]:

```
130000
```

```
In [ ]:
Q1 = df_teamlead.quantile(0.25)
Q3 = df_teamlead.quantile(0.75)
IQR = Q3 - Q1
df_teamlead = df_teamlead[~((df_teamlead < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5
* IQR)))]

df_teamlead.size

99801

df_teamlead.head()
```

```
Out [ ]:
```

	Custo merId	Surn ame	Credit Score	Geogr aphy	Gen der	A ge	Ten ure	Bala nce	NumOfP roducts	HasCr Card	IsActive Member	Estimate dSalary	Exi ted
1	15647 311	Hill	608	Spain	Fem ale	4 1	1	8380 7.86	1	0	1	112542.5 8	0
3	15701 354	Boni	699	France	Fem ale	3 9	1	0.00	2	0	0	93826.63	0
4	15737 888	Mitc hell	850	Spain	Fem ale	4 3	2	1255 10.82	1	1	1	79084.10	0
6	15592 531	Bartl ett	822	France	Mal e	5 0	7	0.00	2	1	1	10062.80	0
8	15792 365	He	501	France	Mal e	4 4	4	1420 51.07	2	0	1	74940.50	0

Scaling the data

```
In [ ]:
from sklearn.preprocessing import StandardScaler
num = num.drop("Exited",axis=1)
df_teamlead[num.columns] =
StandardScaler().fit_transform(df_teamlead[num.columns])
```

OneHotEncoding

```
In [ ]:
cat = df_teamlead.select_dtypes(exclude=["int64","float64"])
```

```
cat.columns
```

Out[]:

```
Index(['Surname', 'Geography', 'Gender', 'NumOfProducts', 'HasCrCard',  
      'IsActiveMember'],  
      dtype='object')
```

In []:

```
mod = df_teamlead.drop(["Surname","CustomerId"],axis=1)  
df_teamlead = pd.get_dummies(drop_first=True,data=mod)  
df_teamlead.head()
```

Out[]:

	CreditScore	Age	Te nur e	Bal anc e	Estima tedSal ary	Ex ite d	Geograp hy_Ger many	Geogra phy_Sp ain	Gend er_M ale	NumOf Product s_2	NumOf Product s_3	HasC rCar d_1	IsActive Membe r_1
1	- 0.45 5004	0.5 998 17	- 1.4 031 86	0.1 740 70	0.2211 81	0	0	1	0	0	0	0	1
3	0.49 6133	0.3 502 14	- 1.4 031 86	- 1.1 597 14	- 0.1049 37	0	0	0	0	1	0	0	0
4	2.07 4392	0.8 494 20	- 1.0 554 81	0.8 377 64	- 0.3618 20	0	0	1	0	0	0	1	1
6	1.78 1735	1.7 230 29	0.6 830 46	- 1.1 597 14	- 1.5644 88	0	0	0	1	1	0	1	1
8	- 1.57 3373	0.9 742 21	- 0.3 600 70	1.1 009 99	- 0.4340 20	0	0	0	1	1	0	0	1

Splitting The Data

In []:

```
from sklearn.model_selection import train_test_split
```

In []:

```
X_train, X_test, y_train, y_test =
train_test_split(df_teamlead.drop("Exited",axis=1),df_teamlead["Exited"],
test_size=0.33, random_state=42)
```

In []:

X_train

Out[]:

	CreditScore	Age	Tenure	Balance	EstimatedSalary	Geography_Germany	Geography_Spain	Gender_Male	NumberOfProducts_2	NumberOfProducts_3	HasCreditCard_1	IsActiveMember_1
5699	0.339352	-0.148991	0.335341	1.002652	-0.220065	0	0	1	0	0	1	1
9362	0.339352	-0.148991	-0.707775	1.023467	-1.090516	0	0	1	0	0	1	1
7166	-0.350483	0.475016	1.030751	-1.159714	-0.342299	0	1	0	1	0	1	0
6009	-1.322524	-0.398594	-0.012365	-1.159714	1.065572	0	1	1	1	0	1	0
1470	-0.873086	-0.648196	-1.403186	-0.179488	-0.663047	1	0	0	1	0	1	1
...
6833	0.245283	-0.273792	1.726162	0.192161	0.808182	0	0	0	0	0	0	0
7039	-1.343428	2.971042	-1.403186	1.448630	-0.745551	1	0	1	0	0	0	1

	CreditScore	Age	Tenure	Balance	EstimatedSalary	Geography_Germany	Geography_Spain	Gender_Male	NumberOfProducts_2	NumberOfProducts_3	HasCreditCard_1	IsActiveMember_1
1127	-1071675	1.099022	-0.707775	1.127783	-0.060021	1	0	0	1	0	1	1
9911	0.318448	0.974221	-0.707775	0.674980	-1.325708	0	0	1	0	0	0	0
9486	1.060543	-0.523395	0.335341	1.668808	1.076720	0	0	1	0	0	1	1

5143 rows × 12 columns

In []: