#### Assignment -2

## **Python Programming**

Assignment Date	17 September 2022
Student Name	Renuka G
Student Roll Number	211419104220
Maximum Marks	2 Marks

### 1. Download the dataset: Dataset

```
In [2]: import numpy as np import pandas as pd import matplotlib.pyplot as plt import seaborn as sns

In [3]: import warnings warnings.('ignore')
```

#### 2. Load the dataset.



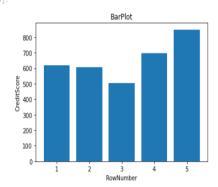
#### 3. Perform Below Visualizations

## **Univariate Analysis**

```
In [10]: sns.barplot(data['Gender'], data['Age'])
Out[10]: <AxesSubplot:xlabel='Gender', ylabel='Age'>
         35
         30
         25
       ∯ 20
         15
         10
In [11]: plt.plot(data['Age'].head(10))
Out[11]: [<matplotlib.lines.Line2D at 0x296a12b9f10>]
        45
        35
 In [12]: plt.pie(data['Age'].head(), autopct="%.2f")
 In [13]: sns.distplot(data['Age'].head(200))
Out[13]: <AxesSubplot:xlabel='Age', ylabel='Density'>
         0.05
         0.04
        Density
0.03
         0.02
         0.00
        BI - Variate Analysis
In [14]: plt.scatter(data['CreditScore'].head(100),data['Age'].head(100))
Out[14]: 
Out[14]:
        60
        30
                                                                                                                           Activ
```

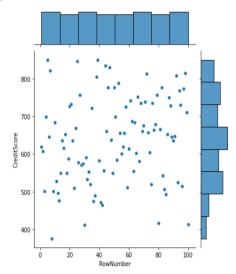
```
In [15]: plt.bar(data['RowNumber'].head() ,data['CreditScore'].head(), )
    plt.title('BarPlot')
    plt.xlabel('RowNumber')
    plt.ylabel('CreditScore')
```

Out[15]: Text(0, 0.5, 'CreditScore')



```
In [16]: sns.jointplot(data['RowNumber'].head(100) ,data['CreditScore'].head(100), )
```

Out[16]: <seaborn.axisgrid.JointGrid at 0x296a14df8b0>



In [17]:	da	ata.head()													
Out[17]:		RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
	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

# Multi - Variate Analysis



## 4. Perform descriptive statistics on the dataset.

n [21]:	data	.head()																		
ut[21]:	Ro	wNumber	CustomerId	Surname	CreditScore	Geogr	raphy G	iender	Age	Tenui	re E	Balance	NumOfProducts	HasCrCard	IsAc	ctiveMember	Estim	natedSalary	Exited	ı
	0	1	15634602	Hargrave	619	F	France F	emale	42		2	0.00	1	1		1		101348.88	1	
	1	2	15647311	Hill	608		Spain F	emale	41		1 83	3807.86	1	0		1		112542.58	0	
	2	3	15619304	Onio	502	F	France F	emale	42		8 159	9660.80	3	1		0		113931.57	1	
	3	4	15701354	Boni	699	F	France F	emale	39		1	0.00	2	0		0		93826.63	0	)
	4	5	15737888	Mitchell	850		Spain F	emale	43		2 12	5510.82	1	1		1		79084.10	0	)
n [22]:	data	.tail()																		
		RowNumb	per Custome		nme CreditS					-			ce NumOfProdo		Card	IsActiveMem			•	
	9995 9996	RowNumb	96 156062		aku	<b>core G</b> 771 516	<b>Geograph</b> Franc Franc	e Ma	ale	Age 1 39 35	<b>Tenure</b> 5 10	0.	00	icts HasCr 2	Card 1	IsActiveMem	ober E	<b>EstimatedSal</b> 96270 101699	.64	0 0
	9995	RowNumb	96 156062	29 Obiji 92 Johnst	aku	771	Franc	e Ma	ale ale	39	5	57369.	00	2	1	IsActiveMem	0	96270	.64	0
	9995 9996	<b>RowNumb</b> 99 99	96 156062 97 155698	29 Obiji 92 Johnst	aku one Liu	771 516	Franc	e Ma e Ma e Fema	ale ale	39 35	5 10 7	57369.	00 61 00	2	1	IsActiveMem	0	96270 101699	.64 .77 .58	0
	9995 9996 9997	<b>RowNumb</b> 99 99	156062 197 155698 198 155845 199 156823	29 Obiji 92 Johnst 32 55 Sabba	aku one Liu atini	771 516 709	Franco Franco Franco German	e Ma e Ma e Fema	ale ale ale	39 35 36	5 10 7 3	0. 57369.	00 61 00 31	2 1 1	1	IsActiveMem	0 1 1	96270 101699 42085	1.64 1.77 1.58 1.52	0
n [22]: ut[22]: n [23]:	9995 9996 9997 9998 9999	99 99 99	156062 197 155698 198 155845 199 156823	29 Obiji 92 Johnst 32 55 Sabba	aku one Liu atini	771 516 709 772	Franco Franco Franco German	e Ma e Ma e Fema y Ma	ale ale ale	39 35 36 42	5 10 7 3	0. 57369. 0. 75075.	00 61 00 31	2 1 1 2	1	IsActiveMem	0 1 1 0	96270 101699 42085 92888	1.64 1.77 1.58 1.52	0 0 1

```
In [20]: data.info()
             <class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
                                          Non-Null Count Dtype
                  Column
                                          10000 non-null
10000 non-null
10000 non-null
10000 non-null
10000 non-null
                    RowNumber
                                                               int64
                    CustomerId
Surname
CreditScore
                                                               int64
object
int64
                    Geography
Gender
                                                               object
                                          10000 non-null
10000 non-null
10000 non-null
10000 non-null
                                                               object
int64
int64
float64
                    Age
Tenure
Balance
                    NumOfProducts
                                          10000 non-null int64
             10 HasCrCard 10000 non-null int64

11 IsActiveMember 10000 non-null int64

12 EstimatedSalary 10000 non-null float64

13 Exited 10000 non-null int64

dtypes: float64(2), int64(9), object(3)

memory usage: 1.1+ MB
 In [24]:
## 4 movement of business decisions
data.mean()
             RowNumber
CustomerId
CreditScore
                                      5.000500e+03
                                      1.569094e+07
6.505288e+02
3.892180e+01
              Age
Tenure
                                      5.012800e+00
             Balance
                                       7 648589e+04
             NumOfProducts
HasCrCard
IsActiveMember
                                      7.848589E+04
1.530200E+00
7.055000E-01
5.151000E-01
             EstimatedSalary
                                      1.000902e+05
              Exited
                                      2.037000e-01
             dtype: float64
             data.median()
            RowNumber
CustomerId
                                      5.000500e+03
1.569074e+07
Out[25]:
            CreditScore
                                      6.520000e+02
            Age
Tenure
                                      3.700000e+01
                                      5.000000e+00
9.719854e+04
1.000000e+00
            Balance
NumOfProducts
            HasCrCard
                                      1.000000e+00
            TsActiveMember
                                      1 0000000+00
            EstimatedSalary
Exited
dtype: float64
                                     1.001939e+05
0.000000e+00
In [26]: data.mode()
                    RowNumber Customerld Surname CreditScore Geography Gender Age Tenure Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
                0
                               1 15565701
                                                                    850.0
                                                                                 France
                                                                                            Male 37.0
                                                                                                             2.0
                                                                                                                        0.0
                                                                                                                                           1.0
                                                                                                                                                        1.0
                                                                                                                                                                            1.0
                                                                                                                                                                                         24924.92
                                                                                                                                                                                                       0.0
                                                     Smith
                            2 15565706
                                                                    NaN
                                                                                            NaN NaN NaN
                                                                                                                                                       NaN
                                                                                                                                                                          NaN
                                                                                                                                                                                             NaN NaN
                                                   NaN
                                                                              NaN
                                                                                                                       NaN
                                                                                                                                          NaN
                               3
                                      15565714
                                                                                  NaN
                3
                    4 15565779
                                                      NaN
                                                                    NaN
                                                                            NaN
                                                                                           NaN NaN NaN
                                                                                                                       NaN
                                                                                                                                          NaN
                                                                                                                                                       NaN
                                                                                                                                                                          NaN
                                                                                                                                                                                             NaN
                                                                                                                                                                                                      NaN
                4
                              5 15565796
                                                      NaN
                                                                    NaN
                                                                                  NaN
                                                                                            NaN NaN
                                                                                                            NaN
                                                                                                                       NaN
                                                                                                                                          NaN
                                                                                                                                                       NaN
                                                                                                                                                                          NaN
                                                                                                                                                                                             NaN
                                                                                                                                                                                                      NaN
            ...
            9995
                            9996
                                     15815628
                                                      NaN
                                                                     NaN
                                                                                  NaN
                                                                                            NaN NaN
                                                                                                            NaN
                                                                                                                       NaN
                                                                                                                                          NaN
                                                                                                                                                                          NaN
                                                                                                                                                                                             NaN
                                                                                                                                                                                                      NaN
            9996
                           9997
                                     15815645
                                                      NaN
                                                                    NaN
                                                                                  NaN
                                                                                            NaN NaN
                                                                                                            NaN
                                                                                                                       NaN
                                                                                                                                          NaN
                                                                                                                                                       NaN
                                                                                                                                                                          NaN
                                                                                                                                                                                             NaN
                                                                                                                                                                                                      NaN
            9997
                            9998
                                     15815656
                                                      NaN
                                                                    NaN
                                                                                  NaN
                                                                                            NaN NaN
                                                                                                            NaN
                                                                                                                       NaN
                                                                                                                                          NaN
                                                                                                                                                       NaN
                                                                                                                                                                          NaN
                                                                                                                                                                                             NaN
                                                                                                                                                                                                      NaN
                      9999 15815660
                                                                    NaN NaN
            9998
                                                      NaN
                                                                                            NaN NaN
                                                                                                            NaN
                                                                                                                       NaN
                                                                                                                                          NaN
                                                                                                                                                       NaN
                                                                                                                                                                          NaN
                                                                                                                                                                                             NaN
                                                                                                                                                                                                      NaN
```

10000 rows × 14 columns

10000 15815690

NaN

NaN

NaN

NaN NaN NaN

NaN

NaN

NaN

NaN

NaN NaN

9999

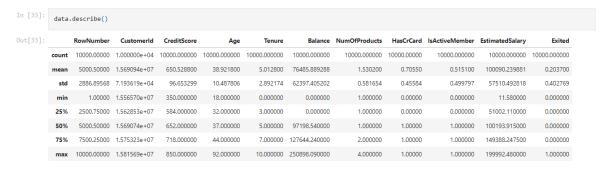
```
In [27]: data.var()
                   RowNumber
CustomerId
CreditScore
                                                         8.334167e+06
5.174815e+09
9.341860e+03
Out[27]:
                  CreditScore
Age
Tenure
Balance
NumOfFroducts
HasCrCard
ISACTiveNember
EstimatedSalary
Exited
dtype: float64
                                                         1.099941e+02
8.364673e+00
3.893436e+09
                                                          3.383218e-01
                                                         2.077905e-01
2.497970e-01
3.307457e+09
1.622225e-01
In [28]: data.std()
                  RowNumber
CustomerId
CreditScore
                                                          2886.895680
71936.186123
96.653299
10.487806
Out[28]:
                   Age
Tenure
                                                         2.892174
62397.405202
0.581654
0.455840
0.499797
                  Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
                                                        57510.492818
0.402769
                   Exited
dtype: float64
In [29]: data.skew()
                                                        0.000000
0.001149
-0.071607
1.011320
0.010991
-0.141109
                  RowNumber
CustomerId
                   CreditScore
Age
Tenure
Balance
                   NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
                                                         0.745568
-0.901812
-0.060437
0.002085
                   Exited
dtype: float64
                                                         1.471611
                                                                                                                                                                                                                                                                                                                     Δctivat
 In [30]: data.kurt()
                                                        -1.200000
-1.196113
-0.425726
1.395347
-1.165225
                    RowNumber
CustomerId
CreditScore
                   Age
Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
                                                          -1.489412
                                                         0.582981
-1.186973
-1.996747
                                                        -1.181518
0.165671
                    {\sf EstimatedSalary}
                    Exited
dtype: float64
                      quantile= data['Age'].quantile(q=[0.75, 0.25])
                      quantile
 Out[31]: 0.75 44.0
0.25 32.0
Name: Age, dtype: float64
 In [32]: sns.boxplot(data['Age'], color='green')
 Out[32]: <AxesSubplot:xlabel='Age'>
```

40

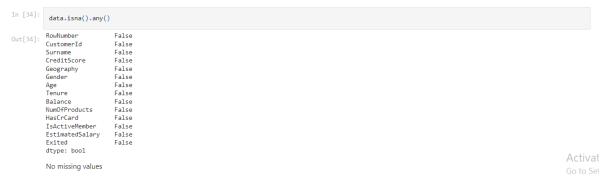
20 30

50 60 Age

70

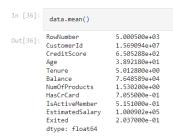


# 5. Handle the Missing values.



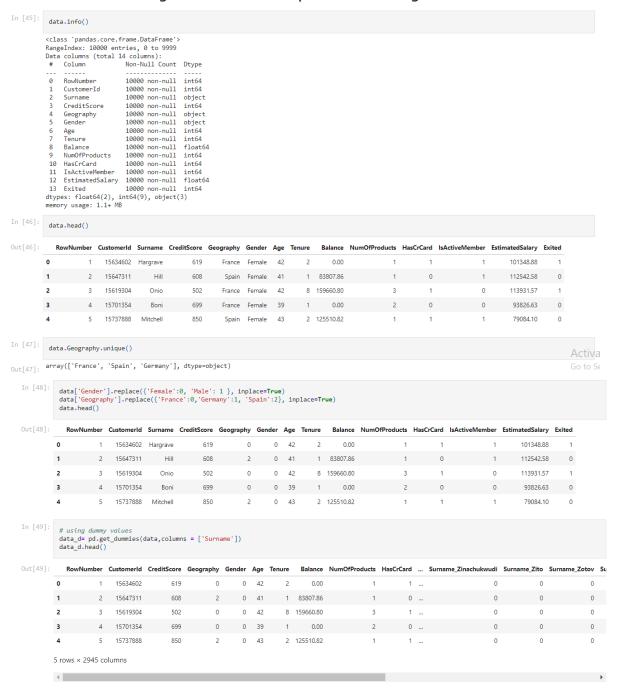
## 6. Find the outliers and replace the outliers





```
In [37]: qut= data.quantile(q=[0.25,0.75]) qut
Out[37]:
            RowNumber Customerld CreditScore Age Tenure Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
            0.25 2500.75 15628528.25 584.0 32.0 3.0 0.00
                                                                                                1.0
                                                                                                         0.0 0.0
                                                                                                                                        51002.1100 0.0
            0.75 7500.25 15753233.75 718.0 44.0 7.0 127644.24
                                                                                                  2.0
                                                                                                             1.0
                                                                                                                              1.0 149388.2475 0.0
In [38]: irq=qut.loc[0.75]- qut.loc[0.25] # q3 and q1
                            4999.5000
124705.5000
134.0000
12.0000
           RowNumber
CustomerId
CreditScore
Out[38]:
            Age
Tenure
                                         4.0000
            EstimatedSalary 98386.1375
Exited 0.0000
dtype: float64
In [39]: # Lower
lower= qut.loc[0.25]+(1.5*irq)
lower
           RowNumber
CustomerId
CreditScore
                                   1.000000e+04
1.581559e+07
7.850000e+02
Out[39]:
            Age
Tenure
Balance
                                   5.000000e+01
9.000000e+00
1.914664e+05
            NumOfProducts
                                   2.500000e+00
            HasCrCard
IsActiveMember
EstimatedSalary
                                   1.500000e+00
1.500000e+00
1.985813e+05
            Exited
dtype: float64
                                   0.000000e+00
  In [40]: #upper upper= qut.loc[0.75]+(1.5*irq) upper
                                    1.499950e+04
1.594029e+07
              RowNumber
CustomerId
  Out[40]:
              CreditScore
Age
Tenure
                                     9.190000e+02
6.200000e+01
1.300000e+01
              Balance
                                     3.191106e+05
              NumOfProducts
HasCrCard
IsActiveMember
                                    3.500000e+00
2.500000e+00
2.500000e+00
2.969675e+05
              EstimatedSalary
              Exited
dtype: float64
                                     0.000000e+00
  In [41]: sns.boxplot(data['Age'])
  Out[41]: <AxesSubplot:xlabel='Age'>
  In [42]: data['Age'].mean()
  Out[42]: 38.9218
                                                                                                                                                                                            Activa
```

## 7. Check for Categorical columns and perform encoding.



## 8. Split the data into dependent and independent variables.

```
In [50]: data.head()
Out [50]: RowNumber Customerld Surname CreditScore Geography Gender Age Tenure Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
                          1 15634602 Hargrave
                                                              619
                                                                           0
                                                                                       0 42
                                                                                                                                                                                    101348.88
            1 2 15647311 Hill 608 2 0 41 1 83807.86
                          3 15619304 Onio 502
                                                                              0 0 42 8 159660.80
            3 4 15701354 Boni 699 0 0 39 1 0.00
                                                                                                                                                                      0 93826.63 0
                                                                         2 0 43 2 125510.82
                          5 15737888 Mitchell 850
                                                                                                                                                                                     79084.10
In [51]: x=data_d.drop(columns= ['EstimatedSalary']).values
              y=data_d['EstimatedSalary'].values
Out[51]: array([[1.0000000e+00, 1.5634602e+07, 6.1900000e+02, ..., 0.0000000e+00, 0.0000000e+00, 0.0000000e+00], [2.0000000e+00, 0.0000000e+00], [2.0000000e+00, 1.5647311e+07, 6.0800000e+02, ..., 0.0000000e+00, 0.0000000e+00, 0.0000000e+00], [3.0000000e+00, 0.5019304e+07, 5.0200000e+02, ..., 0.0000000e+00, 0.0000000e+00], 0.0000000e+00],
                     ..., 0.0000000e+03, 1.5584532e+07, 7.0900000e+02, ..., 0.0000000e+00, 0.0000000e+00, 0.0000000e+00], [9.9990000e+03, 1.5682355e+07, 7.7200000e+02, ..., 0.0000000e+00, 0.0000000e+00, 0.0000000e+00], [1.0000000e+04, 1.5528319e+07, 7.9200000e+02, ..., 0.0000000e+00, 0.0000000e+00, 0.0000000e+00]])
In [52]: y
Out[52]: array([101348.88, 112542.58, 113931.57, ..., 42085.58, 92888.52, 38190.78])
```

### 9. Scale the independent variables

### 10. Split the data into training and testing

```
In [55]: from sklearn.model_selection import train_test_split
In [56]: x_train, x_test, y_train, y_test = train_test_split(x,y, test_size = 0.2)
In [57]: print(x_train.shape, x_test.shape)
(8000, 2944) (2000, 2944)
```