

Assignment -2

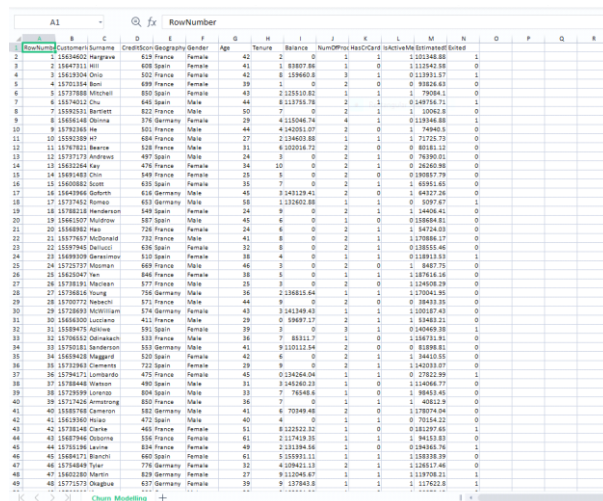
DATA VISUALISATION AND PRE-PROCESSING

ASSIGNMENT DATE	28 September 2022
STUDENT NAME	THANGA PANDI.P
STUDENT ROLLNUMBER	612419104022
MAXIMUM MARKS	2 Marks

Question-1:

Download the Dataset

SOLUTION:



RowNumber	Customer's Name	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfOpenAccounts	IsActive	IsActiveM	EstimatedSalary								
1	15634621 Merve	628	France	Female	42	2	0	1	1	1	105124.80	1							
2	15647311 Wei	608	Spain	Female	41	1	83807.86	1	0	1	111043.08	0							
3	15618304 Chiu	582	France	Female	42	8	109660.6	3	1	0	0113931.37	1							
4	15701034 Bern	689	France	Female	39	1	0	2	0	0	93824.63	0							
5	15773888 Mitchell	800	Spain	Female	43	215510.82	1	1	1	1	79084.1	0							
6	15674622 Chin	640	Spain	Male	44	612173.18	2	1	0	0	388784.71	1							
7	15826251 Bertien	820	France	Male	50	7	0	2	1	1	10062.6	0							
8	15685648 Oliveira	770	Germany	Female	29	412506.74	4	1	0	0	123746.80	1							
9	15782385 Wei	501	France	Male	44	4142051.07	2	0	1	1	74940.5	0							
10	15582889 vt	634	France	Male	27	2134603.80	1	1	1	1	71725.73	0							
11	151978015 Benme	518	France	Male	31	6103018.72	2	0	0	0	80181.12	0							
12	15737173 Andrews	487	Spain	Male	24	3	0	2	1	0	78390.01	0							
13	15652284 Kay	470	France	Female	34	10	0	2	1	0	32626.90	0							
14	15683483 Chin	548	France	Female	25	3	0	2	0	0	930837.79	0							
15	15680882 Suen	610	Spain	Female	35	7	0	2	1	1	65951.60	0							
16	15643966 Gethun	614	Germany	Male	45	3148129.45	2	0	1	1	64327.26	0							
17	15773422 Romeo	633	Germany	Male	38	1132800.80	1	1	0	0	3097.47	1							
18	15178812 Henderson	549	Spain	Female	24	8	0	2	1	1	14404.41	0							
19	15661307 Mathews	387	Spain	Male	45	4	0	1	0	0	018884.81	0							
20	15648982 Wei	758	France	Female	24	4	0	2	1	1	14724.03	0							
21	15573487 McDonald	752	France	Male	41	8	0	2	1	1	1170866.17	0							
22	15573482 Dorland	610	Spain	Female	32	8	0	2	1	0	0135551.46	0							
23	15689928 Gerasimov	510	Spain	Female	38	4	0	1	1	0	014913.33	1							
24	15125737 Moisan	669	France	Male	46	3	0	2	0	1	8467.70	0							
25	15615057 Yee	646	France	Female	38	5	0	1	1	1	187616.16	0							
26	15783181 Maclean	377	France	Male	25	3	0	2	0	1	1124050.29	0							
27	15758820 Young	756	Germany	Male	36	2135635.64	2	1	1	1	1170041.85	0							
28	15700772 Natchez	571	France	Male	44	9	0	2	0	0	38433.35	0							
29	15753885 MacMillan	574	Germany	Female	43	31481348.43	1	1	1	1	1100187.43	0							
30	15686820 Luciano	411	France	Male	29	0	59697.17	2	1	1	83483.21	0							
31	15689745 Oliveira	590	Spain	Female	39	8	0	3	1	0	0140469.36	1							
32	15708512 Orlowski	553	France	Male	36	7	83311.7	1	0	0	1166751.91	0							
33	15750181 Sanderson	533	Germany	Male	41	9105112.84	2	0	0	0	81896.81	0							
34	15683428 Maguire	510	Spain	Female	42	4	0	2	1	1	13443.05	0							
35	15732983 Clements	722	Spain	Female	29	9	0	2	1	1	142031.07	0							
36	15734612 Lombardi	470	France	Female	45	0134264.04	1	1	0	0	07821.99	1							
37	15738848 Watson	480	Spain	Male	31	31481260.23	1	0	1	1	114066.77	0							
38	15739938 Lomax	806	Spain	Male	33	7	78948.6	1	0	1	8845.46	0							
39	15717424 Armstrong	850	France	Male	36	7	0	1	1	1	40812.9	0							
40	15683768 Cameron	583	Germany	Male	41	4	70949.48	2	0	1	1178074.04	0							
41	15613960 Hean	470	Spain	Male	40	4	0	2	1	0	76161.21	0							
42	15738148 Corne	480	France	Female	51	8123232.32	1	0	0	0	0181297.85	1							
43	15687946 Ouellette	558	France	Female	41	2117418.18	1	1	1	1	14151.83	0							
44	15735136 Lavine	834	France	Female	49	2131394.58	1	0	0	0	0194685.76	1							
45	15684671 Bouchet	660	Spain	Female	41	5105931.11	1	1	1	1	1189338.39	0							
46	15734688 Tyer	776	Germany	Female	32	4108421.13	2	1	1	1	1126517.46	0							
47	15682280 Martin	829	Germany	Female	27	9122046.87	1	1	1	1	1137026.21	1							
48	15717575 Oughden	627	Germany	Female	39	9	137863.8	1	1	1	117502.6	1							

Question-2:

Loading dataset

SOLUTION:

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

from matplotlib import pyplot as plt

%matplotlib inline
```

```
df = pd.read_csv("Churn_Modelling.csv")
df
```

```
[ ] import pandas as pd
import seaborn as sns
import numpy as np
from matplotlib import pyplot as plt
%matplotlib inline
```

1.Loading dataset

```
df = pd.read_csv("Churn_Modelling.csv")
```

```
[ ] df
```

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	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.88	1	0	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	156680.80	3	1	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	93828.83	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	79084.10	0
...
9995	9998	15606229	Obijaku	771	France	Male	39	5	0.00	2	1	98270.84	0
9996	9997	15598992	Johnstone	516	France	Male	35	10	57389.81	1	1	101899.77	0
9997	9998	15594532	Liu	709	France	Female	38	7	0.00	1	0	42085.58	1
9998	9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	92888.52	1
9999	10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	38190.78	0

10000 rows x 14 columns

Question-3:

Visualizations

a) Univariate Analysis

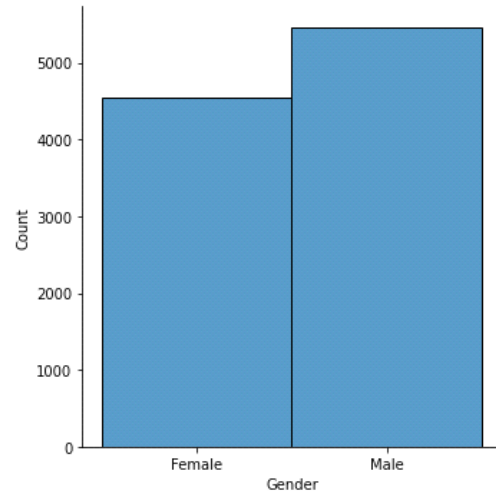
SOLUTION:

```
sns.displot(df.Gender)
```

a) Univariate Analysis

```
sns.displot(df.Gender)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fbbaf922250>
```



b) Bi-Variate Analysis

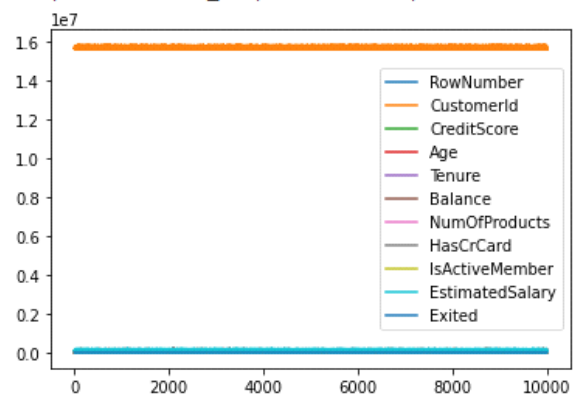
SOLUTION:

```
df.plot.line()
```

b)Bi-Variate Analysis

```
df.plot.line()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbbabcbc810>
```

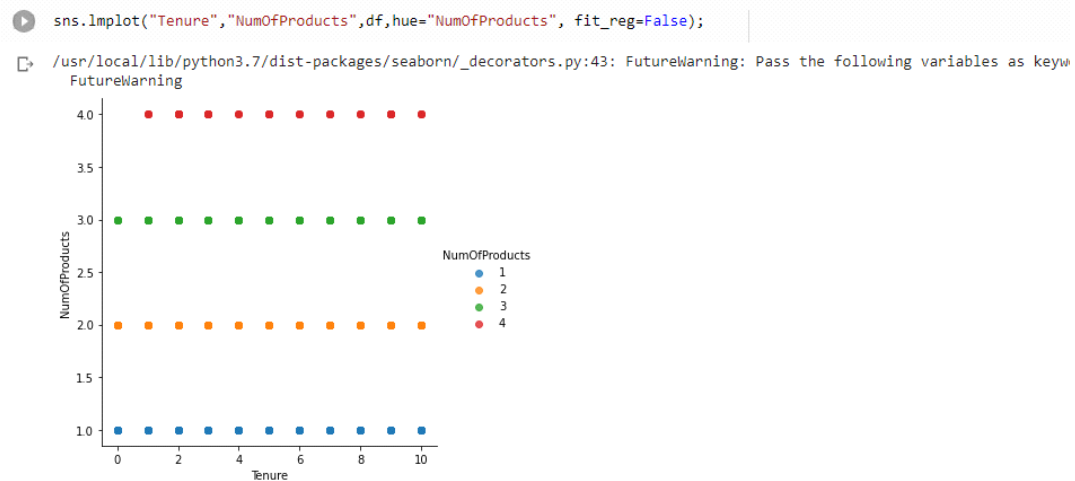


c) Multi - Variate Analysis

SOLUTION:

```
sns.lmplot("Tenure", "NumOfProducts", df, hue="NumOfProducts", fit_reg=False);
```

c)Multi - Variate Analysis



Question-4:

Perform descriptive statistics on the dataset.

SOLUTION:

```
df.describe()
```

1.Perform descriptive statistics on the dataset

```
[ ] df.describe()
```

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	2888.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

Question-5:

Handle the Missing values.

SOLUTION:

```
data = pd.read_csv("Churn_Modelling.csv")
pd.isnull(data["Gender"])
```

1.Handle the Missing values.

```
[ ] data = pd.read_csv("Churn_Modelling.csv")
pd.isnull(data["Gender"])
```

```
0      False
1      False
2      False
3      False
4      False
...
9995   False
9996   False
9997   False
9998   False
9999   False
Name: Gender, Length: 10000, dtype: bool
```

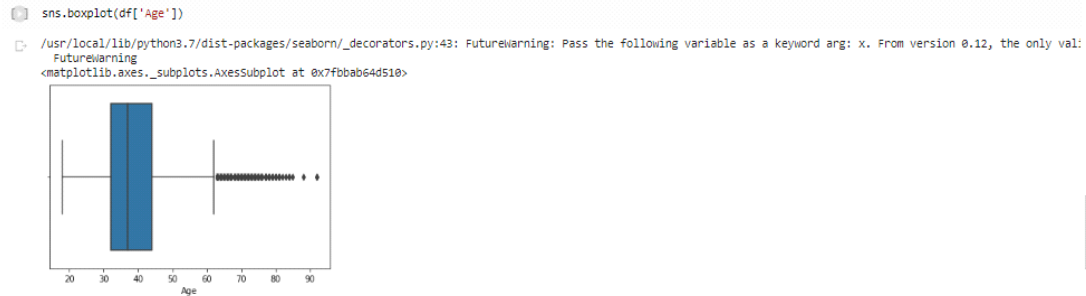
Question-6:

Find the outliers and replace the outliers.

SOLUTION:

```
sns.boxplot(df['Age'])
```

1. Find the outliers and replace the outliers.



SOLUTION:

```
df['Age']=np.where(df['Age']>50,40,df['Age'])  
df['Age']
```

```
[ ] df['Age']=np.where(df['Age']>50,40,df['Age'])  
    df['Age']
```

```
0      42  
1      41  
2      42  
3      39  
4      43  
..  
9995   39  
9996   35  
9997   36  
9998   42  
9999   28  
Name: Age, Length: 10000, dtype: int64
```

Question-7:

Check for Categorical columns and perform encoding.

SOLUTION:

```
pd.get_dummies(df, columns=["Gender", "Age"], prefix=["Age", "Gender"])).head()
```

1. Check for Categorical columns and perform encoding

```
pd.get_dummies(df, columns=["Gender", "Age"], prefix=["Age", "Gender"]).head()
```

RowIndex	RowNumber	CustomerId	Surname	CreditScore	Geography	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	...	Gender_41	Gender_42	Gender_43	Gender_44	Gender_45	Gender_46	Gender_47	Gender_48	Gender_49	Gender_50
0	1	15634602	Hargrave	619	France	2	0.00	1	1	1	...	0	1	0	0	0	0	0	0	0	0
1	2	15647311	Hill	608	Spain	1	83807.86	1	0	1	...	1	0	0	0	0	0	0	0	0	0
2	3	15619304	Onio	502	France	8	159660.80	3	1	0	...	0	1	0	0	0	0	0	0	0	0
3	4	15701354	Boni	699	France	1	0.00	2	0	0	...	0	0	0	0	0	0	0	0	0	0
4	5	15737888	Mitchell	850	Spain	2	125510.82	1	1	1	...	0	0	1	0	0	0	0	0	0	0

5 rows x 47 columns

Question-8:

- Split the data into dependent and independent variables.
- (a) Split the data into Independent variables.

SOLUTION:

```
X = df.iloc[:, :-1].values
print(X)
```

1. Split the data into dependent and independent variables.

a) Split the data into Independent variables.

```
X = df.iloc[:, :-1].values
print(X)

[[1 15634602 'Hargrave' ... 1 1 101348.88]
 [2 15647311 'Hill' ... 0 1 112542.58]
 [3 15619304 'Onio' ... 1 0 113931.57]
 ...
 [9998 15584532 'Liu' ... 0 1 42085.58]
 [9999 15682355 'Sabbatini' ... 1 0 92888.52]
 [10000 15628319 'Walker' ... 1 0 38190.78]]
```

(b) Split the data into Dependent variables

SOLUTION:

```
Y = df.iloc[:, -1].values  
print(Y)
```

b)Split the data into Dependent variables.

```
[ ] Y = df.iloc[:, -1].values  
    print(Y)  
  
[1 0 1 ... 1 1 0]
```

Question-9:

Scale the independent variables

SOLUTION:

```
import pandas as pd  
from sklearn.preprocessing import MinMaxScaler  
scaler = MinMaxScaler()  
df[["CustomerId"]] = scaler.fit_transform(df[["CustomerId"]])  
print(df)
```


1. Scale the independent variables

```
[ ] import pandas as pd
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
df[["CustomerId"]] = scaler.fit_transform(df[["CustomerId"]])

[ ] print(df)
```

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age
0	1	0.275616	Hargrave	619	France	Female	42
1	2	0.326454	Mill	608	Spain	Female	41
2	3	0.214421	Onio	582	France	Female	42
3	4	0.542636	Boni	699	France	Female	39
4	5	0.688778	Mitchell	850	Spain	Female	43
...
9995	9996	0.162119	Obijaku	771	France	Male	39
9996	9997	0.016765	Johnstone	516	France	Male	35
9997	9998	0.075327	Liu	789	France	Female	36
9998	9999	0.466637	Sabbatini	772	Germany	Male	42
9999	10000	0.258483	Walker	792	France	Female	28

	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember
0	2	0.00	1	1	1
1	1	83887.86	1	0	1
2	8	159650.80	3	1	0
3	1	0.00	2	0	0
4	2	125510.82	1	1	1
...
9995	5	0.00	2	1	0
9996	10	57369.61	1	1	1
9997	7	0.00	1	0	1
9998	3	75075.31	2	1	0
9999	4	130142.79	1	1	0

	EstimatedSalary	Exited
0	101348.88	1
1	112542.58	0
2	113931.57	1
3	93826.63	0
4	79084.10	0
...
9995	96270.64	0
9996	101699.77	0
9997	42085.58	1
9998	92888.52	1
9999	38190.78	0

[10000 rows x 14 columns]

Question-10:

Split the data into training and testing

SOLUTION:

```
from sklearn.model_selection import train_test_split

train_size=0.8
X = df.drop(columns = ['Tenure']).copy()
y = df['Tenure']
X_train, X_rem, y_train, y_rem = train_test_split(X,y, train_size=0.8)
test_size = 0.5
```

```

X_valid, X_test, y_valid, y_test = train_test_split(X_rem,y_rem, test_size=0.5)
print(X_train.shape), print(y_train.shape)
print(X_valid.shape), print(y_valid.shape)
print(X_test.shape), print(y_test.shape)

```

1.Split the data into training and testing

```

[ ] from sklearn.model_selection import train_test_split
    train_size=0.8
    X = df.drop(columns = ['Tenure']).copy()
    y = df['Tenure']
    X_train, X_rem, y_train, y_rem = train_test_split(X,y, train_size=0.8)
    test_size = 0.5
    X_valid, X_test, y_valid, y_test = train_test_split(X_rem,y_rem, test_size=0.5)
    print(X_train.shape), print(y_train.shape)
    print(X_valid.shape), print(y_valid.shape)
    print(X_test.shape), print(y_test.shape)

```

```

(8000, 13)
(8000,)
(1000, 13)
(1000,)
(1000, 13)
(1000,)
(None, None)

```