

1.importing package

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

2. Loading dataset

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

	Row Num ber	Cust omer Id	Sur na me	Cred itSco re	Geo grap hy	Ge nd er	A g e	Te nu re	Bal anc e	NumO fProdu cts	Has CrC ard	IsActiv eMem ber	Estima tedSal ary	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	Bonini	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
...
9995	9996	15606229	Obijaku	771	France	Male	39	5	0.00	2	1	0	96270.64	0
9996	9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	101699.77	0

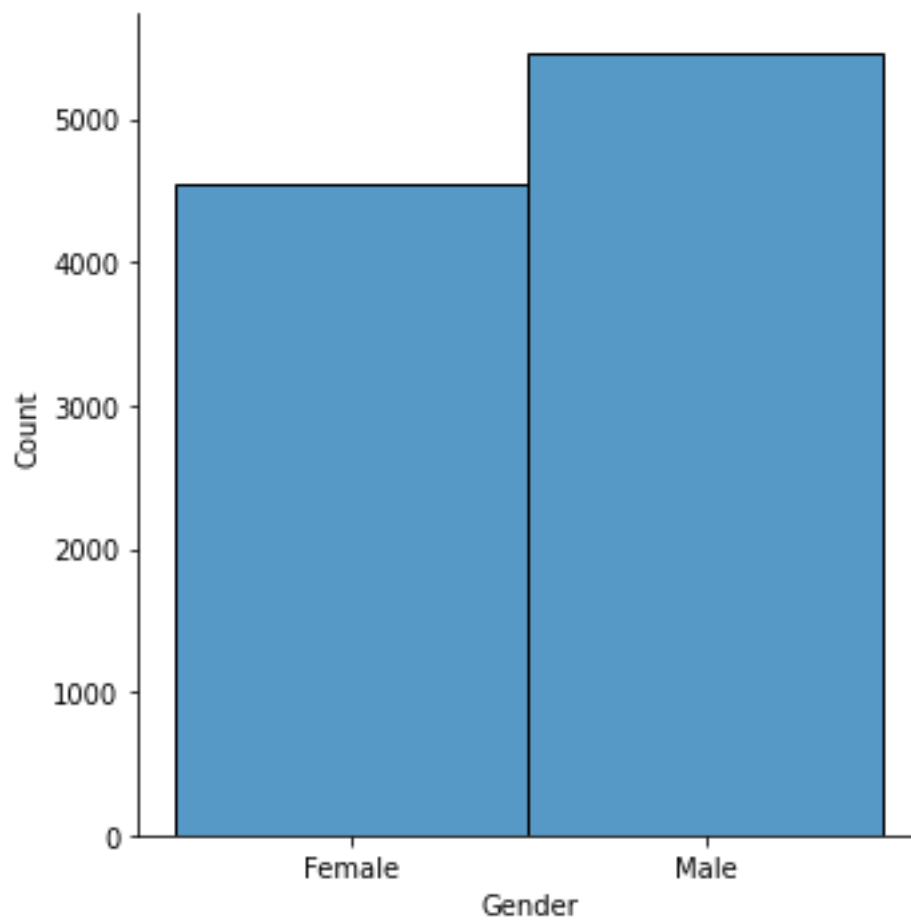
	Row Num ber	Cust omer Id	Sur na me	Cred itSco re	Geo grap hy	Ge nd er	A g e	Te nu re	Bal anc e	NumO fProdu cts	Has CrC ard	IsActiv eMem ber	Estima tedSal ary	Ex ite d
9 9 9 7	9998	1558 4532	Liu	709	Fran ce	Fe ma le	3 6	7	0.00	1	0	1	42085. 58	1
9 9 9 8	9999	1568 2355	Sab bati ni	772	Ger man y	Ma le	4 2	3	750 75.3 1	2	1	0	92888. 52	1
9 9 9 9	1000 0	1562 8319	Wal ker	792	Fran ce	Fe ma le	2 8	4	130 142. 79	1	1	0	38190. 78	0

10000 rows × 14 columns

3. Visualizations

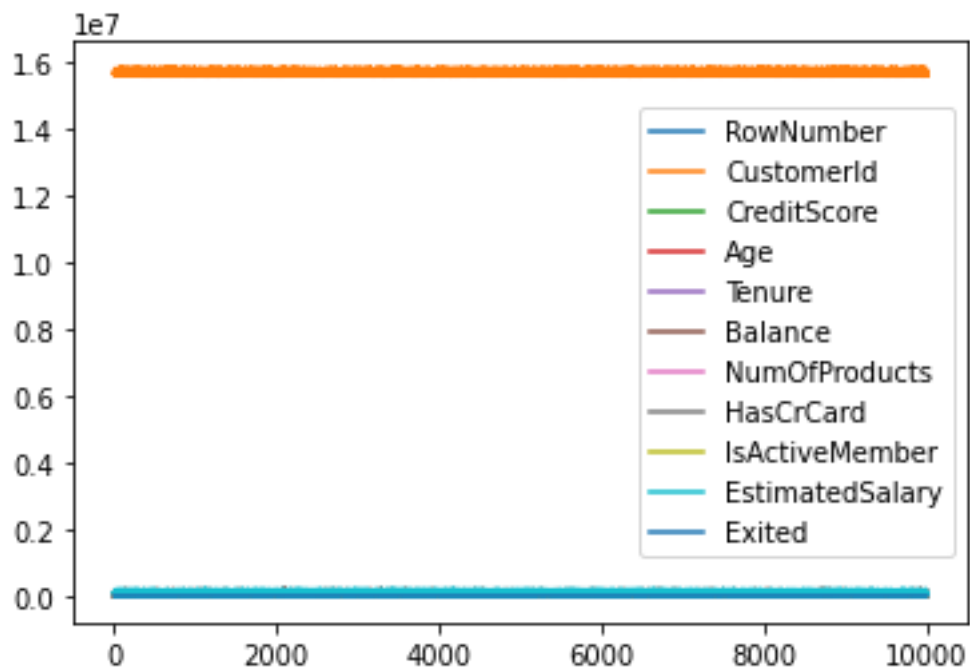
a) Univariate Analysis

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



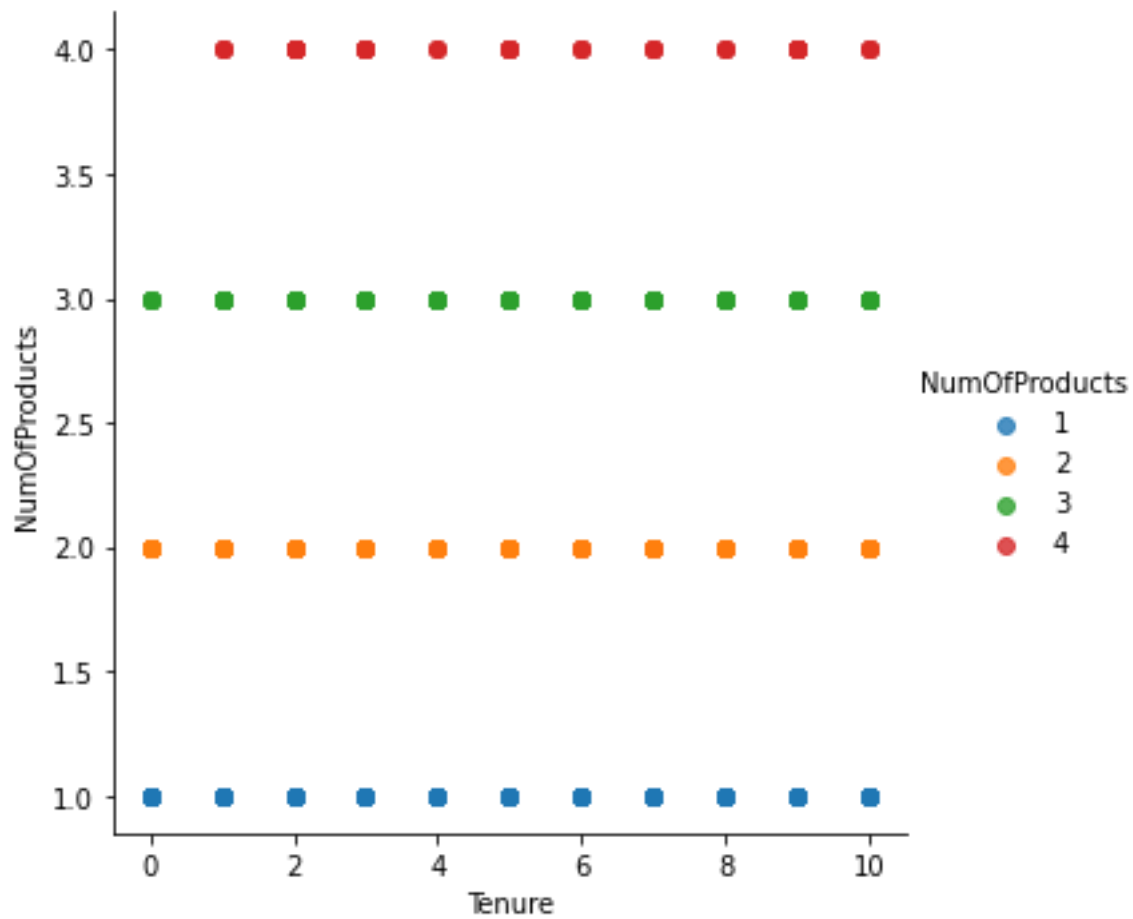
b) Bi-Variate Analysis

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



c) Multi - Variate Analysis

```
In [ ]:
sns.lmplot("Tenure", "NumOfProducts", df, hue="NumOfProducts", fit_reg=False);
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variables as keyword args: x, y, data. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
FutureWarning
```



4. Perform descriptive statistics on the dataset.

```
df.describe()
```

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.000000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	5000.500000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.705500	0.515100	100090.239881	0.203700
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCreditCard	IsActiveMember	EstimatedSalary	Exited
min	1.0000	1.55657e+07	350.0000	18.0000	0.000000	0.000000	1.000000	0.000000	0.000000	11.580000	0.000000
25%	2500.750000	1.562853e+07	584.0000	32.0000	3.000000	0.000000	1.000000	0.000000	0.000000	51002.100000	0.000000
50%	5000.500000	1.569074e+07	652.0000	37.0000	5.000000	97198.540000	1.000000	1.000000	1.000000	100193.915000	0.000000
75%	7500.250000	1.575323e+07	718.0000	44.0000	7.000000	127644.240000	2.000000	1.000000	1.000000	149388.247500	0.000000
max	10000.000000	1.581569e+07	850.0000	92.0000	10.000000	250898.090000	4.000000	1.000000	1.000000	199992.480000	1.000000

5. Handle the Missing values.

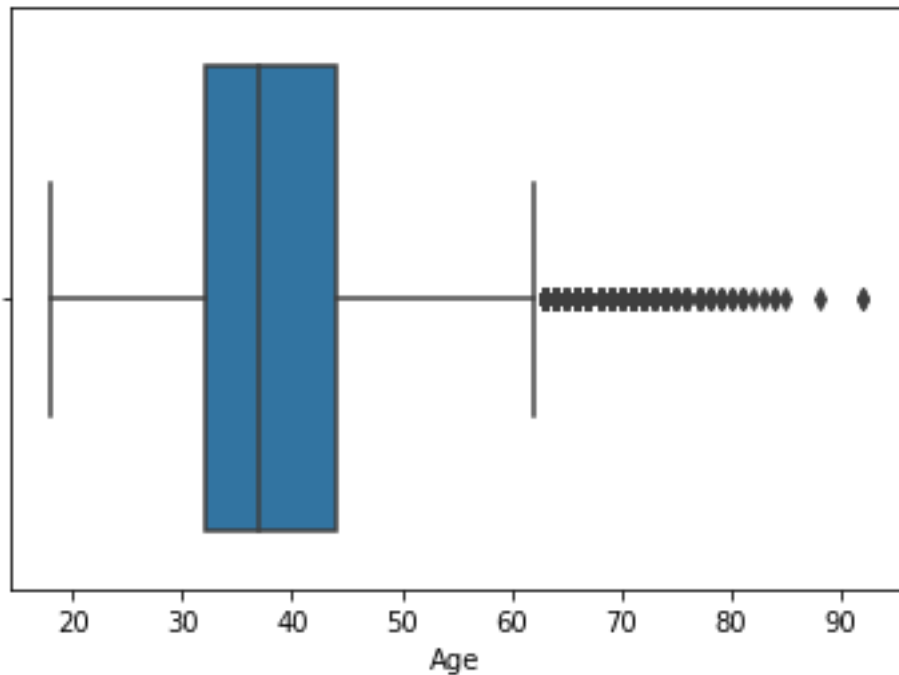
```
data = pd.read_csv("Churn_Modelling.csv")
pd.isnull(data["Gender"])
0      False
1      False
2      False
3      False
4      False
...
9995   False
9996   False
9997   False
9998   False
9999   False
Name: Gender, Length: 10000, dtype: bool
```

6. Find the outliers and replace the outliers.

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

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

FutureWarning



```
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
```

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

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation
```

```
.
```

```
FutureWarning
```


		R o w N u m b e r	C u s t o m e r I d	S u r n a m e	C r e d i t S c o r e	G e o g r a p h y	T e n u r e	B a l a n c e	N u m O f P r o d u c t s	H a s C r C a r d	I s A c t i v e M e m b e r	.	G e n d e r	G e n d e r	G e n d e r	G e n d e r	G e n d e r	G e n d e r	G e n d e r	G e n d e r	G e n d e r	G e n d e r	G e n d e r
														4 1	4 2	4 3	4 4	4 5	4 6	4 7	4 8	4 9	5 0
				v e																			
								8 3 8 0 7. 8 6															
1	2		15 64 73 11	H i l l	60 8	S p a n i n	1		1	0	1	.	1	0	0	0	0	0	0	0	0	0	0
								1 5 9 6 6 0. 8 0															
2	3		15 61 93 04	O n i o	50 2	F r a n c e	8		3	1	0	.	0	1	0	0	0	0	0	0	0	0	0
								0. 0 0															
3	4		15 70 13 54	B o n i	69 9	F r a n c e	1		2	0	0	.	0	0	0	0	0	0	0	0	0	0	0
								1 2 5 5 1 0. 8 2															
4	5		15 73 78 88	M i t c h e l l	85 0	S p a n i n	2		1	1	1	.	0	0	1	0	0	0	0	0	0	0	0

5 rows × 45 columns

8. 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.

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

[1 0 1 ... 1 1 0]
```

9. 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	Hill	608	Spain	Female	41
2	3	0.214421	Onio	502	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	Obijiaku	771	France	Male	39
9996	9997	0.016765	Johnstone	516	France	Male	35
9997	9998	0.075327	Liu	709	France	Female	36
9998	9999	0.466637	Sabbatini	772	Germany	Male	42
9999	10000	0.250483	Walker	792	France	Female	28

	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	\
0	2	0.00	1	1	1	
1	1	83807.86	1	0	1	
2	8	159660.80	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]
```

10. 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)