Course Notes Part 2--Data Exploration

#EDA

Part 2.1: Understand the Raw Dataset

```
import pandas as pd
import numpy as np

churn_df = pd.read_csv('bank.data.csv')
#Reading files
```

1. Read the head and it's data

```
churn_df.head()
#Print the head of the file to see the columns
```

Output:

RowNumber	CustomerId	Surname	CreditScore	Geography	NumOfProducts	Ha
1	15634602	Hargrave	619	France	1	1
2	15647311	Hill	608	Spain	1	0
3	15619304	Onio	502	France	3	1
4	15701354	Boni	699	France	2	0
5	15737888	Mitchell	850	Spain	1	1

2. Check the information and datatypes

```
churn_df.info()
#Check basic informations about the dataset
```

Output: Outputs the data columns and their type

Column	Non-Null Count	Dtype
RowNumber	10000 non-null	int64
CustomerId	10000 non-null	int64
Surname	10000 non-null	object
CreditScore	10000 non-null	int64
Geography	10000 non-null	object
Gender	10000 non-null	object
Age	10000 non-null	int64
Tenure	10000 non-null	int64
Balance	10000 non-null	float64
NumOfProducts	10000 non-null	int64
HasCrCard	10000 non-null	int64
IsActiveMember	10000 non-null	int64
EstimatedSalary	10000 non-null	float64
Exited	10000 non-null	int64

Data Types:

- 2 columns of `float64`
- 9 columns of `int64`
- 3 columns of `object`

Memory Usage: 1.1+ MB

3. Check Unique Values

```
print(churn_df.nunique())
#Check the unique informations about each columns
```

Output:

Column	Unique Values
RowNumber	10000
CustomerId	10000

Column	Unique Values
Surname	2932
CreditScore	460
Geography	3
Gender	2
Age	70
Tenure	11
Balance	6382
NumOfProducts	4
HasCrCard	2
IsActiveMember	2
EstimatedSalary	9999
Exited	2

Checks for the number of unique values.

Part 2.2: Understand the Features

1. Check missing variable

This is to check the missing variable

```
print(churn_df.isnull().sum())
```

Output:

Column	Null Values
RowNumber	0
CustomerId	0
Surname	0
CreditScore	0
Geography	0
Gender	0
Age	0

Column	Null Values
Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0

2. Check Basic Statistics

```
print(churn_df[['CreditScore', 'Age', 'Tenure', 'NumOfProducts',
'Balance', 'EstimatedSalary']].describe())
```

Output:

Metric	CreditScore	Age	Tenure	NumOfProducts	Balance	EstimatedSalary
count	10000	10000	10000	10000	10000	10000
mean	650.53	38.92	5.01	1.53	76485.89	100090.24
std	96.65	10.49	2.89	0.58	62397.41	57510.49
min	350.00	18.00	0.00	1.00	0.00	11.58
25%	584.00	32.00	3.00	1.00	0.00	51002.11
50%	652.00	37.00	5.00	1.00	97198.54	100193.92
75%	718.00	44.00	7.00	2.00	127644.24	149388.25
max	850.00	92.00	10.00	4.00	250898.09	199992.48

Numerical Features containing `discrete` features and `continuous` features

Discrete Features:

1. CreditScore
2. Age
3. Tenure
4. NumOfProducts

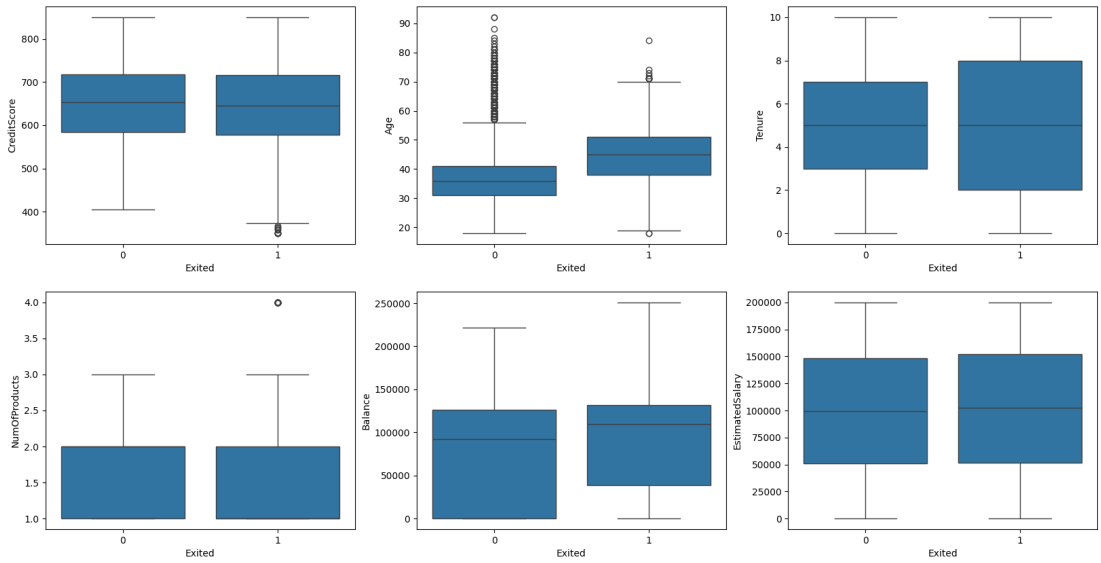
Continuous Features:

1. Balance
2. EstimatedSalary

3. Boxplot for Numerical Features(EDA)

```
#Check the feature distribution
#pandas.dataframe.describe()
#boxplot, distplot, countplot
import matplotlib.pyplot as plt
import seaborn as sns

_,axss = plt.subplots(2, 3, figsize = [20, 10])
sns.boxplot(x = 'Exited', y = 'CreditScore', data = churn_df, ax =
axss[0][0])

sns.boxplot(x = 'Exited', y = 'Age', data = churn_df, ax = axss[0][1])
sns.boxplot(x = 'Exited', y = 'Tenure', data = churn_df, ax = axss[0]
[2])
sns.boxplot(x = 'Exited', y = 'NumOfProducts', data = churn_df, ax =
axss[1][0])

sns.boxplot(x = 'Exited', y = 'Balance', data = churn_df, ax = axss[1]
[1])
sns.boxplot(x = 'Exited', y = 'EstimatedSalary', data = churn_df, ax =
axss[1][2])

plt.show()
```

Output:



Analysis:

And from this figure you can tell which column affects whether or not exited:

1. Age
2. Tenure
3. Balance
   You can tell these three factors are affecting the churns.

4. Checking Categorical Datasets(EDA)

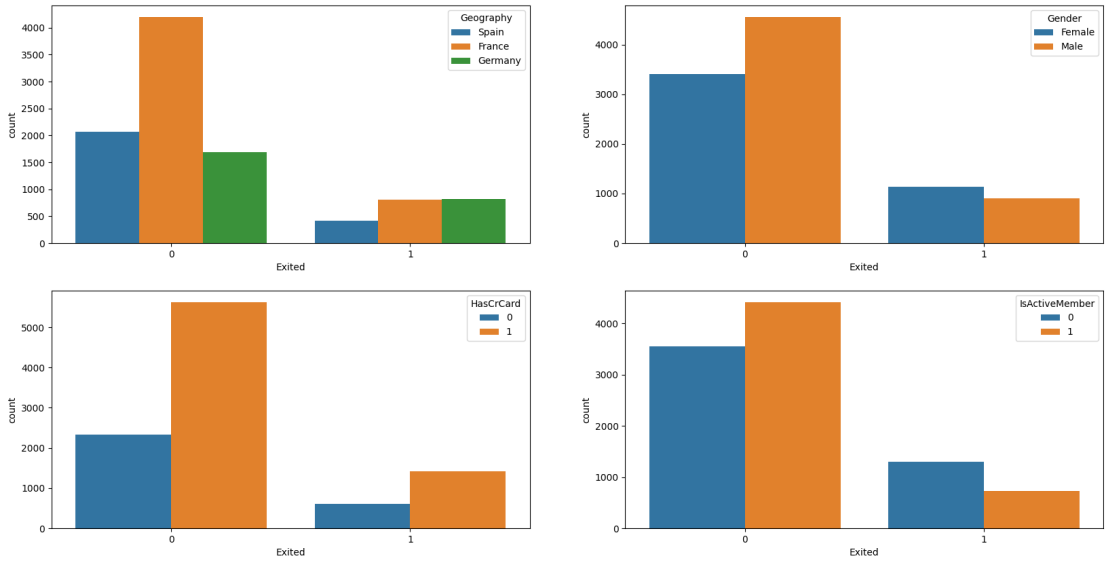
Categorical feature

1. Geography
2. Gender
3. HasCrCard
4. IsActiveMember

```
_,axss = plt.subplots(2, 2, figsize = [20, 10])
sns.countplot(x = 'Exited', hue = 'Geography', data = churn_df, ax =
axss[0][0])

sns.countplot(x = 'Exited', hue = 'Gender', data = churn_df, ax =
axss[0][1])
sns.countplot(x = 'Exited', hue = 'HasCrCard', data = churn_df, ax =
axss[1][0])
```

```
sns.countplot(x = 'Exited', hue = 'IsActiveMember', data = churn_df, ax
= axss[1][1])
plt.show()
```

Output:



Analysis:

From the graphs

1. Geography
2. Gender
3. IsActiveNumber

   These three factors are affecting the churn prediction

Summary, these EDAs give you a overview of the data and give you potential useful data.