This NoteBook Contains the Steps Require to PreProcessing and Data Cleaning the Dataset

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [13]:
```
data = pd.read_csv('Customer.csv')
```

Description:-

--> `Customer.csv` contains the data regarding the 5000+ Customers who uses the service of a telecom multimedia company Functionalities.

--> Data has been generated synthetically

--> Total Columns in Dataset are 12 and names of the columns are: `['CustomerID', 'Age', 'Gender', 'ContractType', 'MonthlyCharges','TotalCharges', 'TechSupport', 'InternetService','Tenure','PaperlessBilling', 'PaymentMethod', 'Churn']`

In [17]:
```
data.drop(columns=['Unnamed: 0'],inplace=True)
```

In [21]:
```
data.shape
```

Out[21]:
```
(5020, 12)
```

In [22]:
```
df = data.copy()
```

Data Cleaning

In [23]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5020 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   CustomerID        5020 non-null   int64  
 1   Age               4920 non-null   float64
 2   Gender            4970 non-null   object 
 3   ContractType      5020 non-null   object 
 4   MonthlyCharges    5000 non-null   float64
 5   TotalCharges      4972 non-null   float64
 6   TechSupport       4960 non-null   object 
 7   InternetService   4980 non-null   object 
 8   Tenure            5020 non-null   int64  
 9   PaperlessBilling  4990 non-null   object 
 10  PaymentMethod     4990 non-null   object 
 11  Churn             4970 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 470.8+ KB
```

In [24]:
```
df.sample(10)
```

Out[24]:

	CustomerID	Age	Gender	ContractType	MonthlyCharges	TotalCharges	TechSupport	Int
3415	3352	34.0	Female	Two year	45.967546	91.935091	Yes	
114	249	39.0	Female	Month-to-month	46.369826	3245.887788	No	
809	3009	75.0	Male	One year	38.521572	2234.251165	Yes	
667	92	89.0	Female	Month-to-month	47.122152	848.198737	Yes	
1774	60	61.0	Female	Month-to-month	89.072289	5522.481944	No	
1995	3804	81.0	Female	Two year	101.928504	5809.924732	No	
2378	50	NaN	Male	Two year	105.698381	4650.728745	Yes	
2553	4162	72.0	Female	Two year	35.885208	2368.423703	Yes	
1960	807	26.0	Female	Two year	85.457651	4358.340222	No	
2505	4609	88.0	Female	Month-to-month	72.221368	4766.610286	Yes	

Handling the missing values except the Churn feature

In [27]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5020 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   CustomerID       5020 non-null   int64  
 1   Age              4920 non-null   float64
 2   Gender           4970 non-null   object 
 3   ContractType     5020 non-null   object 
 4   MonthlyCharges   5000 non-null   float64
 5   TotalCharges     4972 non-null   float64
 6   TechSupport      4960 non-null   object 
 7   InternetService  4980 non-null   object 
 8   Tenure           5020 non-null   int64  
 9   PaperlessBilling 4990 non-null   object 
 10  PaymentMethod    4990 non-null   object 
 11  Churn            4970 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 470.8+ KB
```

In [29]:
```
# For Age column
df[df.Age.isna()]
```

Out[29]:

	CustomerID	Age	Gender	ContractType	MonthlyCharges	TotalCharges	TechSupport	Int
117	2155	NaN	Male	One year	44.260702	132.782105	Yes	
137	1749	NaN	Female	One year	26.383099	659.577480	Yes	
185	972	NaN	Male	Two year	71.746555	3802.567434	Yes	
219	1897	NaN	Female	Two year	79.555070	954.660841	No	
229	1624	NaN	Male	One year	44.681175	89.362349	Yes	
...	...	...	...	...	...	...	...	
4821	4320	NaN	Female	Month-to-month	119.865392	7791.250471	Yes	
4826	366	NaN	Male	Two year	99.221590	992.215904	No	
4857	1289	NaN	Female	One year	113.489645	7717.295850	Yes	
4991	776	NaN	Female	Two year	88.504518	708.036141	Yes	
5003	2212	NaN	Female	One year	52.869739	1744.701378	No	

100 rows × 12 columns

In [50]:

```
df = df[~(df.Age.isna())]
```

In [51]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4920 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   CustomerID        4920 non-null   int64  
 1   Age               4920 non-null   float64
 2   Gender            4873 non-null   object 
 3   ContractType      4920 non-null   object 
 4   MonthlyCharges    4900 non-null   float64
 5   TotalCharges      4872 non-null   float64
 6   TechSupport       4860 non-null   object 
 7   InternetService   4880 non-null   object 
 8   Tenure            4920 non-null   int64  
 9   PaperlessBilling  4890 non-null   object 
 10  PaymentMethod     4891 non-null   object 
 11  Churn             4870 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 499.7+ KB
```

In [54]:

```
# For Gender Column

df.Gender.value_counts()
```

Out[54]:

```
Gender
Female    2480
Male      2393
Name: count, dtype: int64
```

In [61]:

```
temp = df[df.Gender.isna()].index
```

In [64]:

```
df.Gender = df.Gender.fillna(method='ffill')
```

In [65]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4920 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   CustomerID       4920 non-null   int64  
 1   Age              4920 non-null   float64
 2   Gender           4920 non-null   object 
 3   ContractType     4920 non-null   object 
 4   MonthlyCharges   4900 non-null   float64
 5   TotalCharges     4872 non-null   float64
 6   TechSupport      4860 non-null   object 
 7   InternetService  4880 non-null   object 
 8   Tenure           4920 non-null   int64  
 9   PaperlessBilling 4890 non-null   object 
 10  PaymentMethod    4891 non-null   object 
 11  Churn            4870 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 628.7+ KB
```

In [72]:

```
# MonthlyCharges

df = df[~(df.MonthlyCharges.isna())]
```

In [73]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4900 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   CustomerID       4900 non-null   int64  
 1   Age              4900 non-null   float64
 2   Gender           4900 non-null   object 
 3   ContractType     4900 non-null   object 
 4   MonthlyCharges   4900 non-null   float64
 5   TotalCharges     4872 non-null   float64
 6   TechSupport      4840 non-null   object 
 7   InternetService  4860 non-null   object 
 8   Tenure           4900 non-null   int64  
 9   PaperlessBilling 4870 non-null   object 
 10  PaymentMethod    4872 non-null   object 
 11  Churn            4850 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 497.7+ KB
```

In [74]:

```
# TotalCharges

df = df[~(df.TotalCharges.isna())]
```

In [75]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4872 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   CustomerID        4872 non-null   int64  
 1   Age               4872 non-null   float64
 2   Gender            4872 non-null   object 
 3   ContractType      4872 non-null   object 
 4   MonthlyCharges    4872 non-null   float64
 5   TotalCharges      4872 non-null   float64
 6   TechSupport       4812 non-null   object 
 7   InternetService   4833 non-null   object 
 8   Tenure            4872 non-null   int64  
 9   PaperlessBilling  4843 non-null   object 
 10  PaymentMethod     4844 non-null   object 
 11  Churn             4823 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 494.8+ KB
```

In [76]:

```
# techsupport

df.TechSupport.value_counts()
```

Out[76]:

```
TechSupport
Yes    2414
No     2398
Name: count, dtype: int64
```

In [77]:

```
df = df[~(df.TechSupport.isna())]
```

In [78]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4812 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   CustomerID        4812 non-null   int64  
 1   Age               4812 non-null   float64
 2   Gender            4812 non-null   object 
 3   ContractType      4812 non-null   object 
 4   MonthlyCharges    4812 non-null   float64
 5   TotalCharges      4812 non-null   float64
 6   TechSupport       4812 non-null   object 
 7   InternetService   4773 non-null   object 
 8   Tenure            4812 non-null   int64  
 9   PaperlessBilling  4783 non-null   object 
 10  PaymentMethod     4785 non-null   object 
 11  Churn             4764 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 488.7+ KB
```

In [83]:

```
# For Internetaservice

df[df.InternetService.isna()].shape
```

Out[83]:

```
(39, 12)
```

In [84]:

```
df = df[~(df.InternetService.isna())]
```

In [85]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4773 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   CustomerID        4773 non-null   int64  
 1   Age               4773 non-null   float64
 2   Gender            4773 non-null   object 
 3   ContractType      4773 non-null   object 
 4   MonthlyCharges    4773 non-null   float64
 5   TotalCharges      4773 non-null   float64
 6   TechSupport       4773 non-null   object 
 7   InternetService   4773 non-null   object 
 8   Tenure            4773 non-null   int64  
 9   PaperlessBilling  4744 non-null   object 
 10  PaymentMethod     4747 non-null   object 
 11  Churn             4726 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 484.8+ KB
```

In [98]:

```
# for PaperlessBilling  and  PaymentMethod

df = df[~(df['PaymentMethod'].isna() | df['PaperlessBilling'].isna())]
```

In [99]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4718 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   CustomerID        4718 non-null   int64  
 1   Age               4718 non-null   float64
 2   Gender            4718 non-null   object 
 3   ContractType      4718 non-null   object 
 4   MonthlyCharges    4718 non-null   float64
 5   TotalCharges      4718 non-null   float64
 6   TechSupport       4718 non-null   object 
 7   InternetService   4718 non-null   object 
 8   Tenure            4718 non-null   int64  
 9   PaperlessBilling  4718 non-null   object 
 10  PaymentMethod     4718 non-null   object 
 11  Churn             4671 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 479.2+ KB
```

In [102]:

```
# for churn

df = df[~(df.Churn.isna())]
```

In [103]:

```
df.shape
```

Out[103]:

```
(4671, 12)
```

In [104]:

```
df.isna().sum()
```

Out[104]:

```
CustomerID          0
Age                 0
Gender              0
ContractType        0
MonthlyCharges      0
TotalCharges        0
TechSupport         0
InternetService     0
Tenure              0
PaperlessBilling    0
PaymentMethod       0
Churn               0
dtype: int64
```

In [106]:

```
df.drop_duplicates(inplace=True)
```

```
C:\Users\Mangukiya Ansh\AppData\Local\Temp\ipykernel_26216\3006716147.py:1: S
ettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
table/user_guide/indexing.html#returning-a-view-versus-a-copy (https://panda
s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
sus-a-copy)
  df.drop_duplicates(inplace=True)
```

In [107]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4651 entries, 0 to 5019
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   CustomerID        4651 non-null   int64  
 1   Age               4651 non-null   float64
 2   Gender            4651 non-null   object 
 3   ContractType      4651 non-null   object 
 4   MonthlyCharges    4651 non-null   float64
 5   TotalCharges      4651 non-null   float64
 6   TechSupport       4651 non-null   object 
 7   InternetService   4651 non-null   object 
 8   Tenure            4651 non-null   int64  
 9   PaperlessBilling  4651 non-null   object 
 10  PaymentMethod     4651 non-null   object 
 11  Churn             4651 non-null   object 
dtypes: float64(3), int64(2), object(7)
memory usage: 472.4+ KB
```

In [109]:

```
data = df
```

In [110]:

```
data.to_csv('Cleaned_Customer.csv')
```