Customer Churn Prediction

Assignment Description: At Sunbase, we prioritize understanding our customers and ensuring their satisfaction. To achieve this, we want to develop a machine learning model that predicts customer churn. Your task as a Machine Learning Intern is to work on this project, following the guidelines and responsibilities outlined in the job description.

Objective: Develop a machine learning model to predict customer churn based on historical customer data. You will follow a typical machine learning project pipeline, from data preprocessing to model deployment.

- **First, we have loaded the dataset and after loading it we got the shape and size of the dataset.**

```
# Reading the excel file from dataset
df = pd.read_excel("customer_churn_large_dataset.xlsx")
df
```

	CustomerID	Name	Age	Gender	Location	Subscription_Length_Months	Monthly_Bill	Total_Usage_GB	Churn
0	1	Customer_1	63	Male	Los Angeles	17	73.36	236	0
1	2	Customer_2	62	Female	New York	1	48.76	172	0
2	3	Customer_3	24	Female	Los Angeles	5	85.47	460	0
3	4	Customer_4	36	Female	Miami	3	97.94	297	1
4	5	Customer_5	46	Female	Miami	19	58.14	266	0
...	...	...	...	...	...	...	...	...	...
99995	99996	Customer_99996	33	Male	Houston	23	55.13	226	1
99996	99997	Customer_99997	62	Female	New York	19	61.65	351	0
99997	99998	Customer_99998	64	Male	Chicago	17	96.11	251	1
99998	99999	Customer_99999	51	Female	New York	20	49.25	434	1
99999	100000	Customer_100000	27	Female	Los Angeles	19	76.57	173	1

100000 rows × 9 columns

- **It Contain 100000 rows and 9 columns.**
- **The columns of the dataset are of different data types :-**

```
[ ]  df.dtypes

CustomerID                      int64
Name                           object
Age                             int64
Gender                         object
Location                       object
Subscription_Length_Months      int64
Monthly_Bill                  float64
Total_Usage_GB                  int64
Churn                           int64
dtype: object
```

So here we can see that our target variable is in int64 format which makes it a Regression Model

EDA

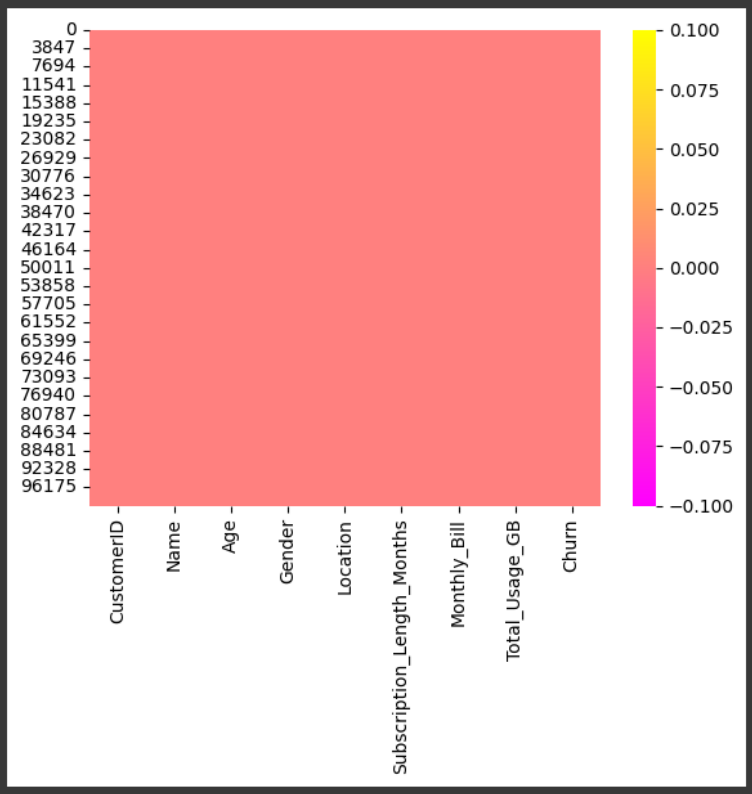
```
[ ]  df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   CustomerID                  100000 non-null  int64
 1   Name                        100000 non-null  object
 2   Age                         100000 non-null  int64
 3   Gender                      100000 non-null  object
 4   Location                    100000 non-null  object
 5   Subscription_Length_Months  100000 non-null  int64
 6   Monthly_Bill                100000 non-null  float64
 7   Total_Usage_GB              100000 non-null  int64
 8   Churn                       100000 non-null  int64
dtypes: float64(1), int64(5), object(3)
memory usage: 6.9+ MB
```

Here we have 3 columns which are cataagorical , so we will encode categorical column later

- **There are 0 NaN values present in the dataset**

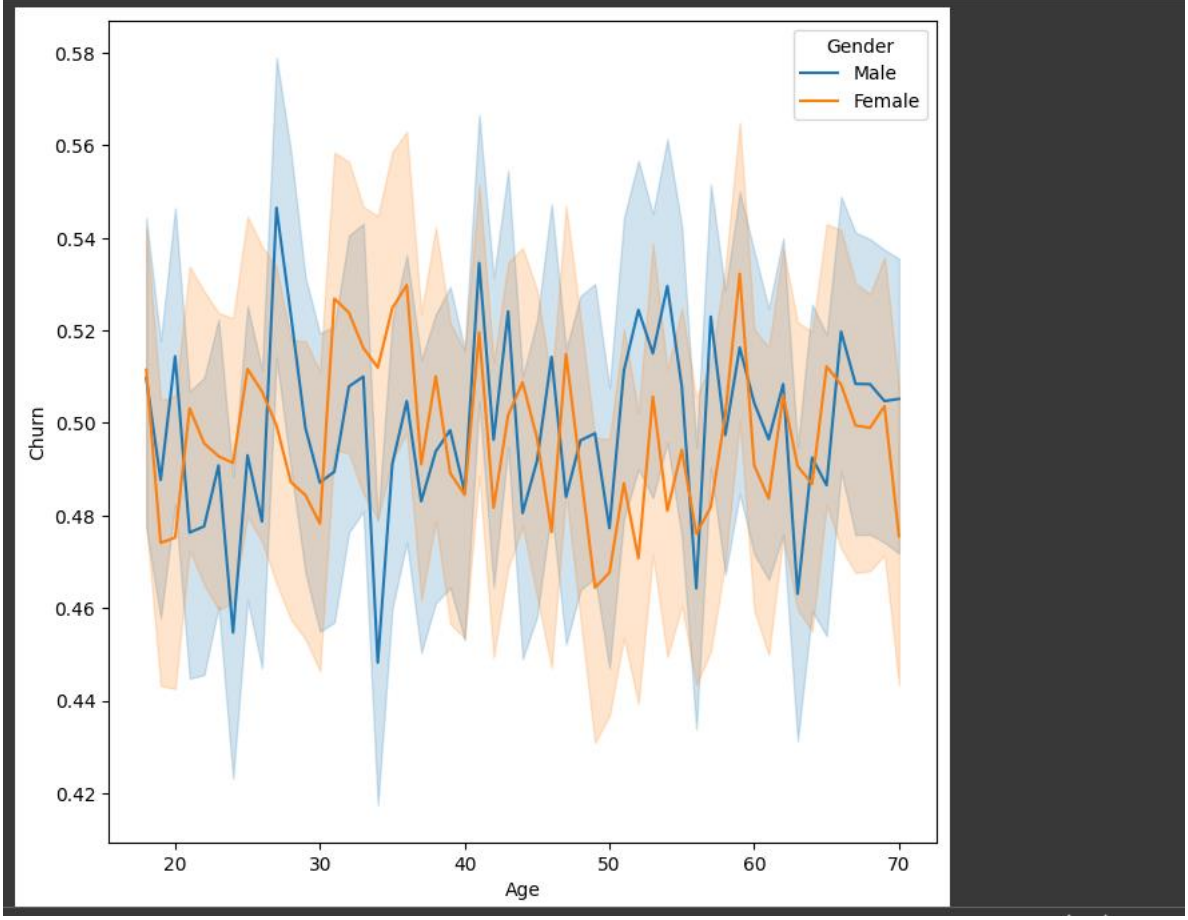
```
[ ]  # Let's visualize NaN values
     sns.heatmap(df.isnull(),cmap="spring")
     plt.show()
```



No Missing values or NaN value is present in the dataset

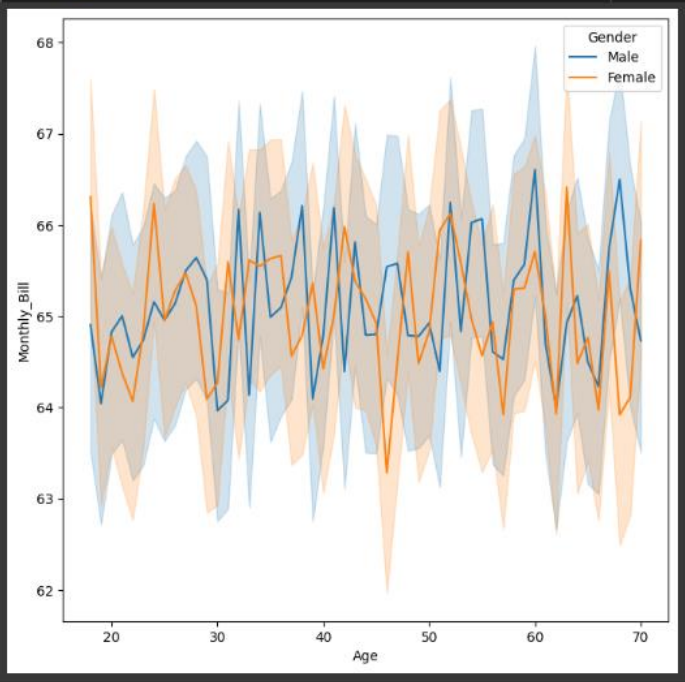
- **Did Univariate analysis and Bivariate analysis on the dataset**
- **We can notice that maximum churn happen to be in male and to in the age group of 25-30**

```
# Comparing Age and Churn and Gender
plt.figure(figsize=(8,8))
sns.lineplot( x = "Age",y = "Churn", data = df,hue= "Gender" ,color='m')
plt.show()
```



- We can see that the monthly bill seems to be high in both gender during the certain age groups :- In Female (age) - 20-30 and then 55-65 , In Males (age) - 30-43 than 52-70.

```
# Comparing Age and Monthly bill and Gender
plt.figure(figsize=(8,8))
sns.lineplot( x = "Age",y = "Monthly_Bill", data = df,hue= "Gender" ,color='m')
plt.show()
```



We can see that the monthly bill seems to be high in both gender during the certain age groups :- In Female (age) - 20-30 and than 55-65 , In Males (age) - 30-43 than 52-70.

- we come to know that Los Angeles has incurred the higgest monthly bills and the lowest have taken place in Chicago

- **I did Correlation in the dataset**
- **I did Label encoding on all the categorical columns**
- **Removed Outliers using ZSCORE**
- **Removed Skewness using PowerTransform**
- **Scaled the data using Standard Scaler on the Test data**
- **Applied all the Regression algorithm.**
- **Out of all the Algo. Gradient Bosting Regressor work Well so I took it too hyper parameter tuning.**
- **And Finally saved the model and Made prediction out of the model.**