Bank Customers Churn Analysis

Term Project Milestone 2: Data Preparation.

In [2]:
```
#Importing data
import pandas as pd
df = pd.read_csv("Churn Modeling.csv")
```

In [3]:
```
#checking null values
df.isnull().sum()
```

Out[3]:
```
RowNumber          0
CustomerId         0
Surname            0
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

Observations

- There is no null value in any column of the dataset.

In [4]:
```
#removing unnecessary features
df.drop(columns = ['RowNumber', 'CustomerId', 'Surname'], axis = 1, inplace = True)
```

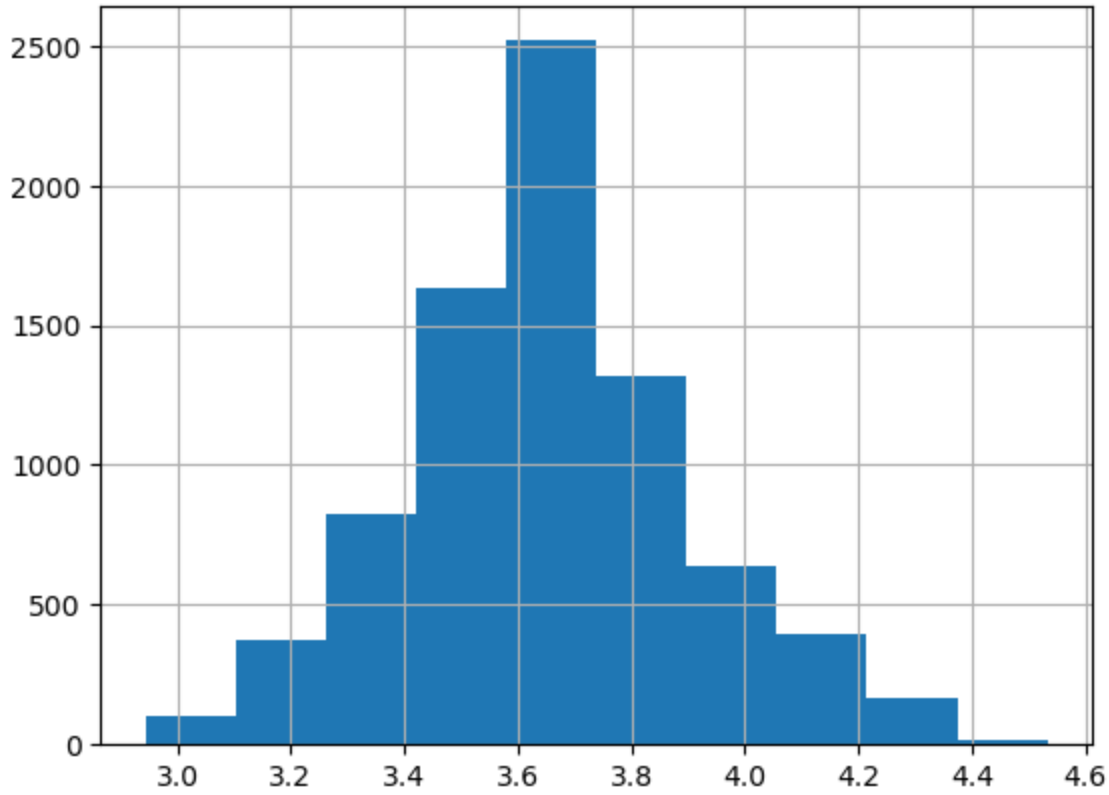
In [5]:
```
#separating input and output features
X = df.drop('Exited', axis = 1)
y = df['Exited']
```

In [6]:
```
#splitting data into train and test data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 1, test_size =
```

The variable age is skewed, so we will transform it by taking log of variable age.

In [7]:
```
import numpy as np
X_train['log_age'] = np.log1p(X_train['Age'])
X_test['log_age'] = np.log1p(X_test['Age'])
X_train['log_age'].hist()
```

Out[7]:
```
<AxesSubplot:>
```



In [8]:
```
#dropping original age variable from train and test data
X_train.drop('Age', axis = 1, inplace = True)
X_test.drop('Age', axis = 1, inplace = True)
```

In [9]:
```
#creating dummy variables for both training and test data
X_train = pd.get_dummies(X_train, drop_first = True)
X_test = pd.get_dummies(X_test, drop_first = True)
```

In [10]:
```
#checking shape of train and test data
print(X_train.shape)
print(X_test.shape)
```

```
(8000, 11)
(2000, 11)
```

Summary of Data Preparation

- Drop unnecessary columns
- Replace age variable with log of age
- split data into train and test data
- Create dummy variables for train and test data