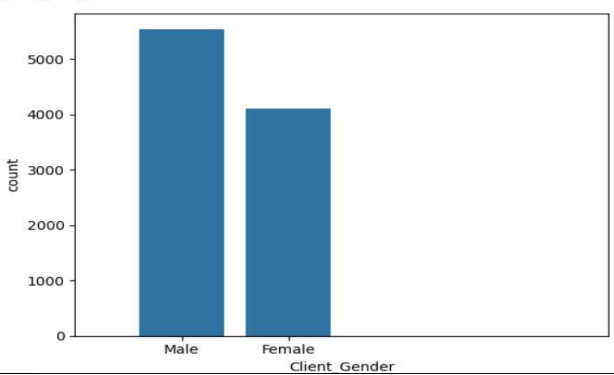
Data Collection and Preprocessing Phase

Date	15th July 2024
Team ID	739823
Project Title	Auto Foresight : A Predictive Model for Streamlining Car Loan Repayment Planning
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, andimplements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	dataset ID Client_Income Car_Owned Bike_Owned Active_Loan House_Own Child_Count Credit_Amount Loan_Annuity Accompany_Client ... Client_Permanent 0 12142509 6750 0.0 0.0 1.0 0.0 0.0 61190.55 3416.85 Alone ... 1 12138936 20250 1.0 0.0 1.0 NaN 0.0 15282 1826.55 Alone ... 2 12181264 18000 0.0 0.0 1.0 0.0 1.0 59527.35 2788.2 Alone ... 3 12188929 15750 0.0 0.0 1.0 1.0 0.0 53870.4 2295.45 Alone ... 4 12133385 33750 1.0 0.0 1.0 0.0 2.0 133988.4 3547.35 Alone ... ... 121851 12207714 29250 0.0 0.0 NaN 1.0 0.0 107820 3165.3 Relative ... 121852 12173765 15750 0.0 1.0 1.0 0.0 0.0 104256 3388.05 Alone ... 121853 12103937 8100 0.0 1.0 0.0 1.0 1.0 55107.9 2989.35 Alone ... 121854 12170623 38250 1.0 1.0 0.0 1.0 0.0 45000 2719.35 Alone ... 121855 12105610 9000 1.0 1.0 1.0 1.0 1.0 62428.95 4201.65 Alone ... 121856 rows × 40 columns
Univariate Analysis	print("Client Income") churn_customers['Client_Income'].describe() Client Income count 9566 unique 322 top 13500 freq 1058 Name: Client_Income, dtype: object print("Credit Amount") churn_customers['Credit_Amount'].describe() Credit Amount count 9570.0 unique 2251.0 top 45000.0 freq 303.0 Name: Credit_Amount, dtype: float64
Bivariate Analysis	sns.countplot(x = "Client_Gender",data = churn_customers) plt.xlim(-1,4) (-1.0, 4.0) 

Data Preprocessing Code Screenshots

Loading Data

```
#Reading the dataset

dataset=pd.read_csv(r"D:\Documents\dataset\train.csv")

C:\Users\Sharan\AppData\Local\Temp\ipykernel_5604\3481427543.py:1: DtypeWarning: Columns (1,7,8,16,17,18,19,20,35) have mixed types. Specify dtype option o
n import or set low_memory=False.
  dataset=pd.read_csv(r"D:\Documents\dataset\train.csv")

dataset
```

	ID	Client_Income	Car_Owned	Bike_Owned	Active_Loan	House_Own	Child_Count	Credit_Amount	Loan_Annuity	Accompany_Client	...	Client_Permanent
0	12142509	6750	0.0	0.0	1.0	0.0	0.0	61190.55	3416.85	Alone	...	
1	12138936	20250	1.0	0.0	1.0	NaN	0.0	15282	1826.55	Alone	...	
2	12181264	18000	0.0	0.0	1.0	0.0	1.0	59527.35	2788.2	Alone	...	
3	12188929	15750	0.0	0.0	1.0	1.0	0.0	53870.4	2295.45	Alone	...	
4	12133385	33750	1.0	0.0	1.0	0.0	2.0	133988.4	3547.35	Alone	...	
...	...	...	...	...	...	...	...	...	...	...	...	...
121851	12207714	29250	0.0	0.0	NaN	1.0	0.0	107820	3165.3	Relative	...	

Handling Missing Data

```
# [Data Pre-Processing] -Handling missing values

dataset= dataset.drop(['Credit_Bureau','Social_Circle_Default','Age_Days','Employed_Days','Score_Source_1','Score_Source_2','Score_Source_3','Registration_
dataset.head()
```

	ID	Client_Income	Car_Owned	Bike_Owned	Active_Loan	House_Own	Child_Count	Credit_Amount	Loan_Annuity	Accompany_Client	...	Client_Housing_Type	F
0	12142509	6750	0.0	0.0	1.0	0.0	0.0	61190.55	3416.85	Alone	...	Home	
1	12138936	20250	1.0	0.0	1.0	NaN	0.0	15282	1826.55	Alone	...	Home	
2	12181264	18000	0.0	0.0	1.0	0.0	1.0	59527.35	2788.2	Alone	...	Family	
3	12188929	15750	0.0	0.0	1.0	1.0	0.0	53870.4	2295.45	Alone	...	Home	
4	12133385	33750	1.0	0.0	1.0	0.0	2.0	133988.4	3547.35	Alone	...	Home	

5 rows × 25 columns

Data Transformation

```
# -Handling Categorical Values

dataset['Client_Income'] = pd.to_numeric(dataset['Client_Income'],errors='coerce')

dataset['Credit_Amount'] = pd.to_numeric(dataset['Credit_Amount'],errors='coerce')

dataset['Population_Region_Relative'] = pd.to_numeric(dataset['Population_Region_Relative'],errors='coerce')

dataset['Loan_Annuity'] = pd.to_numeric(dataset['Loan_Annuity'],errors='coerce')
```

Feature Engineering

```
# -Filling Missing Values and Creating data frame

from mlxtend.preprocessing import TransactionEncoder
column_names=['ID','Client_Income','Car_Owned','Bike_Owned','Active_Loan','House_Own','Child_Count','Credit_Amount','Loan_Annuity','Accompany_Client','Clie

#Create dataframes

loan_data=pd.DataFrame(result,columns=column_names)
loan_data
```

	ID	Client_Income	Car_Owned	Bike_Owned	Active_Loan	House_Own	Child_Count	Credit_Amount	Loan_Annuity	Accompany_Client	...	Client_Housing_
0	12132045.0	27000.0	0.0	1.0	1.0	1.0	0.0	60750.00	7222.50	1.0	...	
1	12196654.0	13500.0	0.0	1.0	1.0	1.0	0.0	28440.00	1851.30	1.0	...	
2	12201738.0	13500.0	1.0	0.0	1.0	1.0	0.0	18000.00	900.00	1.0	...	
3	12131195.0	15750.0	0.0	1.0	1.0	1.0	0.0	59301.00	1746.90	1.0	...	
4	12214557.0	13500.0	0.0	1.0	1.0	0.0	0.0	30234.15	1840.05	1.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	...
224017	12136406.0	12150.0	0.0	0.0	1.0	0.0	0.0	78192.00	2383.65	1.0	...	
224018	12173765.0	15750.0	0.0	1.0	1.0	0.0	0.0	104256.00	3388.05	1.0	...	
224019	12103937.0	8100.0	0.0	1.0	0.0	1.0	1.0	55107.90	2989.35	1.0	...	
224020	12170623.0	38250.0	1.0	1.0	0.0	1.0	0.0	45000.00	2719.35	1.0	...	
224021	12105610.0	9000.0	1.0	1.0	1.0	1.0	1.0	62428.95	4201.65	1.0	...	

224022 rows × 25 columns