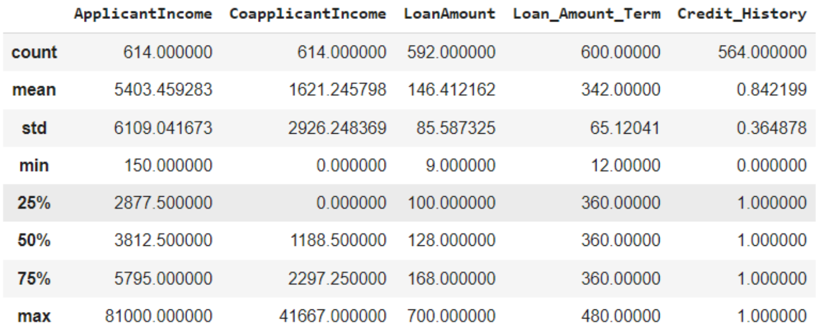
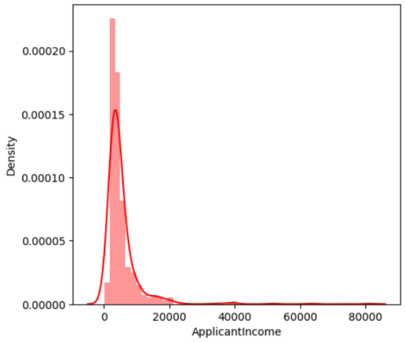
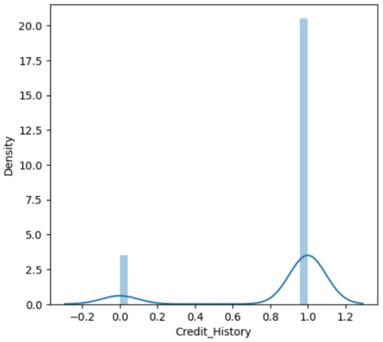
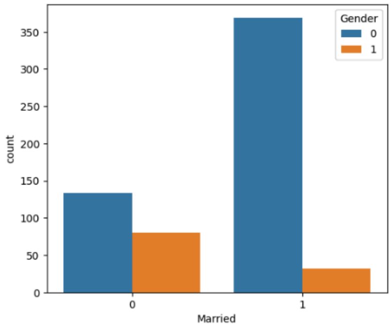
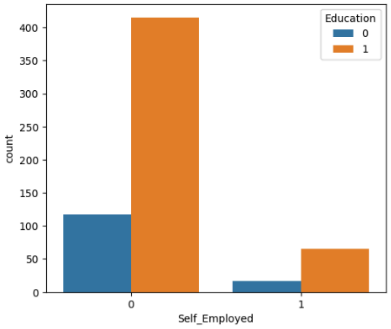
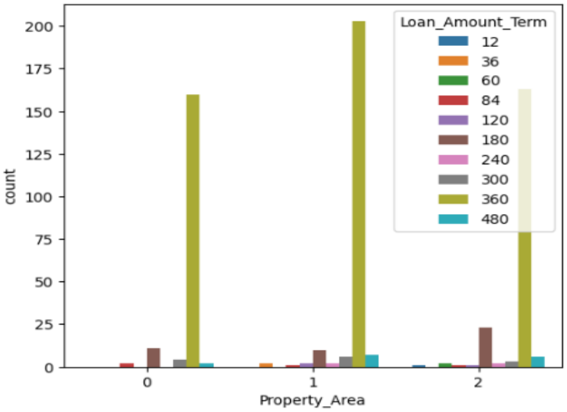
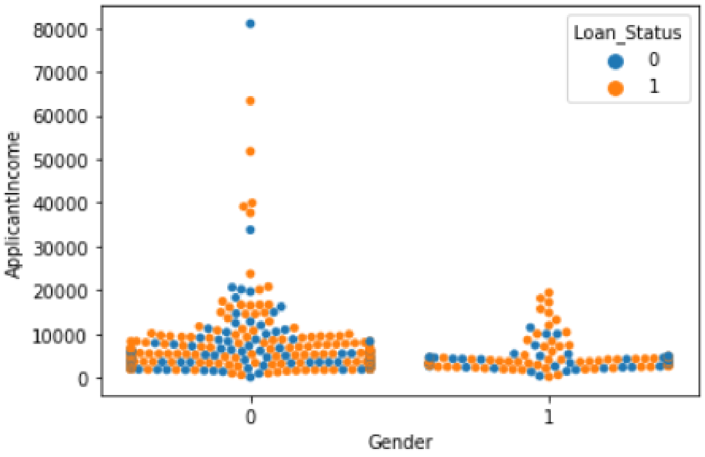
Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	PNT2022TMID124356
Project Title	SmartLender - Applicant Credibility Prediction for Loan Approval
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

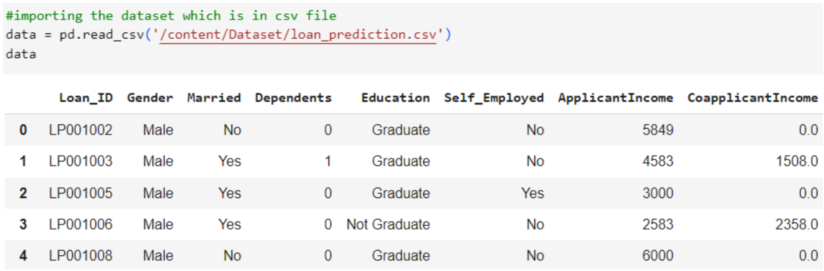
<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>Dimension:
614 rows × 13 columns
Descriptive statistics:

<table>
<tr><th></th><th>ApplicantIncome</th><th>CoapplicantIncome</th><th>LoanAmount</th><th>Loan_Amount_Term</th><th>Credit_History</th></tr>
<tr><td>count</td><td>614.000000</td><td>614.000000</td><td>592.000000</td><td>600.00000</td><td>564.000000</td></tr>
<tr><td>mean</td><td>5403.459283</td><td>1621.245798</td><td>146.412162</td><td>342.00000</td><td>0.842199</td></tr>
<tr><td>std</td><td>6109.041673</td><td>2926.248369</td><td>85.587325</td><td>65.12041</td><td>0.364878</td></tr>
<tr><td>min</td><td>150.000000</td><td>0.000000</td><td>9.000000</td><td>12.00000</td><td>0.000000</td></tr>
<tr><td>25%</td><td>2877.500000</td><td>0.000000</td><td>100.000000</td><td>360.00000</td><td>1.000000</td></tr>
<tr><td>50%</td><td>3812.500000</td><td>1188.500000</td><td>128.000000</td><td>360.00000</td><td>1.000000</td></tr>
<tr><td>75%</td><td>5795.000000</td><td>2297.250000</td><td>168.000000</td><td>360.00000</td><td>1.000000</td></tr>
<tr><td>max</td><td>81000.000000</td><td>41667.000000</td><td>700.000000</td><td>480.00000</td><td>1.000000</td></tr>
</table></td></tr>
<tr><td>Univariate Analysis</td><td></td></tr>
</table>

	 
Bivariate Analysis	  
Multivariate Analysis	

<table>
<tr><td>Outliers and Anomalies</td><td>-</td></tr>
<tr><td colspan="2">

Data Preprocessing Code Screenshots

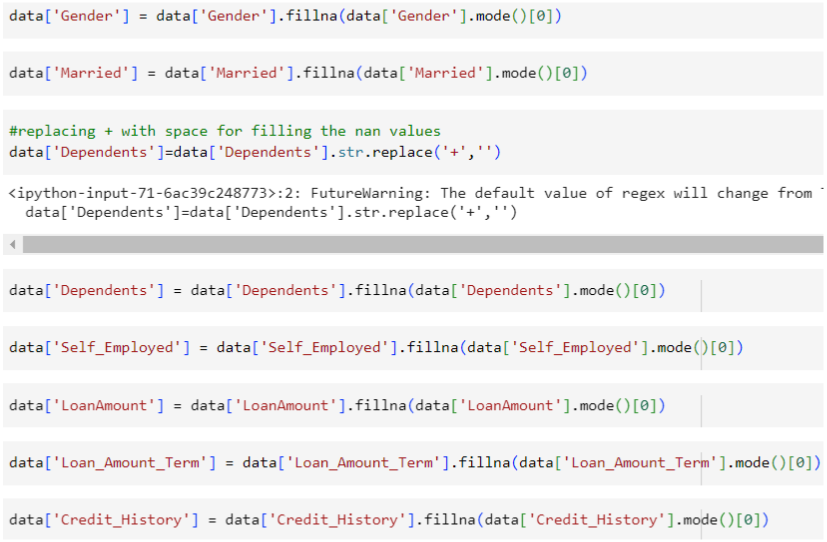
</td></tr>
<tr><td>Loading Data</td><td>



```
#importing the dataset which is in csv file
data = pd.read_csv('/content/Dataset/loan_prediction.csv')
data
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome
0	LP001002	Male	No	0	Graduate	No	5849	0.0
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0
4	LP001008	Male	No	0	Graduate	No	6000	0.0

</td></tr>
<tr><td>Handling Missing Data</td><td>



```
data['Gender'] = data['Gender'].fillna(data['Gender'].mode()[0])
```

```
data['Married'] = data['Married'].fillna(data['Married'].mode()[0])
```

```
#replacing + with space for filling the nan values
data['Dependents']=data['Dependents'].str.replace('+','')
```

```
<ipython-input-71-6ac39c248773>:2: FutureWarning: The default value of regex will change from
  data['Dependents']=data['Dependents'].str.replace('+','')
```

```
data['Dependents'] = data['Dependents'].fillna(data['Dependents'].mode()[0])
```

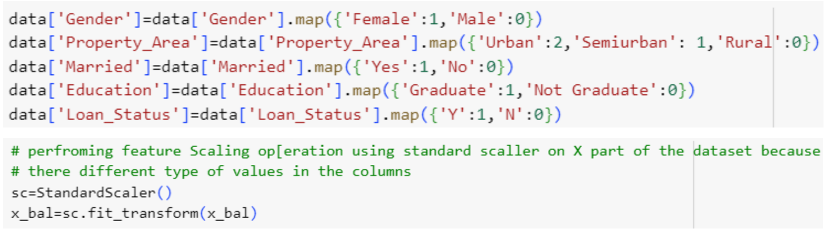
```
data['Self_Employed'] = data['Self_Employed'].fillna(data['Self_Employed'].mode()[0])
```

```
data['LoanAmount'] = data['LoanAmount'].fillna(data['LoanAmount'].mode()[0])
```

```
data['Loan_Amount_Term'] = data['Loan_Amount_Term'].fillna(data['Loan_Amount_Term'].mode()[0])
```

```
data['Credit_History'] = data['Credit_History'].fillna(data['Credit_History'].mode()[0])
```

</td></tr>
<tr><td>Data Transformation</td><td>



```
data['Gender']=data['Gender'].map({'Female':1,'Male':0})
data['Property_Area']=data['Property_Area'].map({'Urban':2,'Semiurban': 1,'Rural':0})
data['Married']=data['Married'].map({'Yes':1,'No':0})
data['Education']=data['Education'].map({'Graduate':1,'Not Graduate':0})
data['Loan_Status']=data['Loan_Status'].map({'Y':1,'N':0})
```

```
# perfroming feature Scaling op[eration using standard scaller on X part of the dataset because
# there different type of values in the columns
sc=StandardScaler()
x_bal=sc.fit_transform(x_bal)
```

</td></tr>
<tr><td>Feature Engineering</td><td>Attached the codes in final submission.</td></tr>
<tr><td>Save Processed Data</td><td>-</td></tr>
</table>