|  |  |  |
| --- | --- | --- |
| **EX:3** | **EDA-Data Cleaning** | **AD23632** |

**Aim:**

• Handling missing values: detection, filling, and dropping

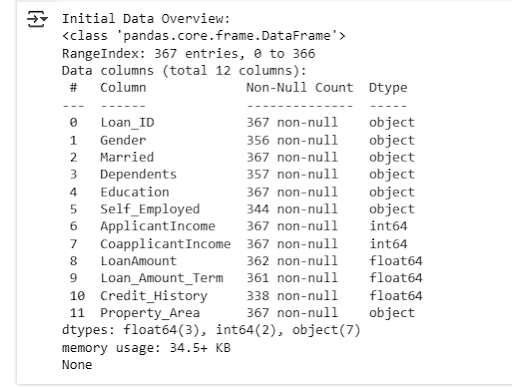
• Removing duplicates and unnecessary data

• Data type conversion and ensuring consistency

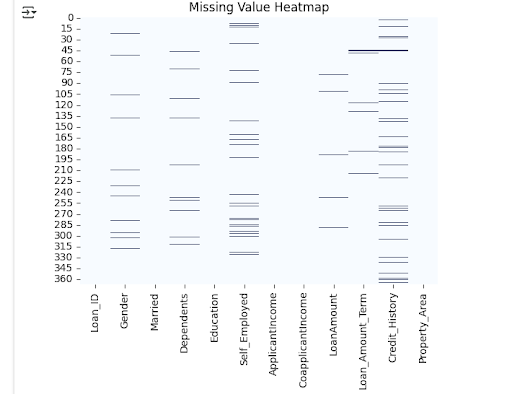
• Normalize data (e.g., standardization, min-max scaling).

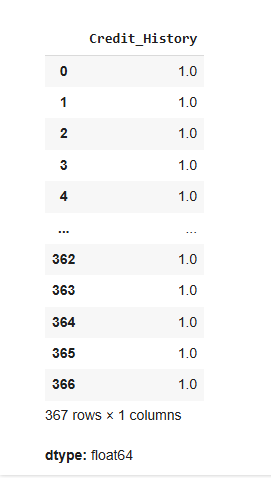
**CODE:**

import pandas as pd  
import numpy as np  
import seaborn as sns  
import matplotlib.pyplot as plt  
from sklearn.preprocessing import MinMaxScaler, StandardScaler  
df = pd.read\_csv('/content/test\_Y3wMUE5\_7gLdaTN.csv')  
print("Initial Data Overview:")  
print([df.info](http://df.info)())

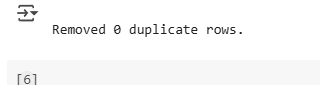
­­­­

print("\nMissing Values in Each Column:\n", df.isnull().sum())  
sns.heatmap(df.isnull(), cbar=False, cmap="Blues")  
plt.title("Missing Value Heatmap")  
plt.show()  
for col in ['Gender', 'Married', 'Dependents', 'Self\_Employed']:  
    df[col].fillna(df[col].mode()[0])  
df['LoanAmount'].fillna(df['LoanAmount'].median())  
df['Loan\_Amount\_Term'].fillna(df['Loan\_Amount\_Term'].mode()[0])  
df['Credit\_History'].fillna(df['Credit\_History'].mode()[0])

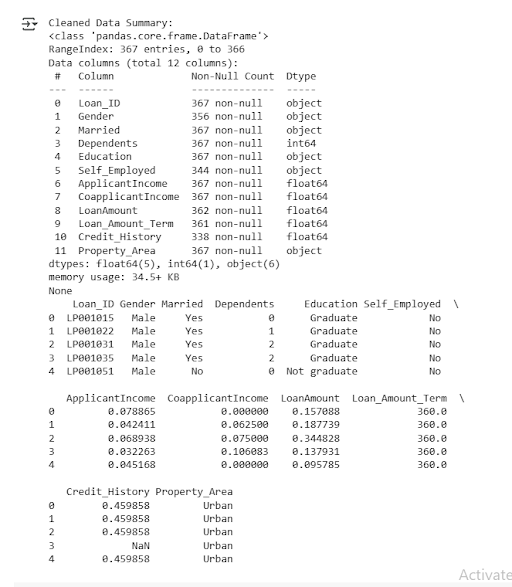




initial\_rows = df.shape[0]  
df.drop\_duplicates(inplace=True)  
print(f"\nRemoved {initial\_rows - df.shape[0]} duplicate rows.")



df['Dependents'] = df['Dependents'].replace('3+', 3).fillna(0).astype(int)  
for col in ['Gender', 'Married', 'Education', 'Self\_Employed', 'Property\_Area']:  
    df[col] = df[col].str.strip().str.capitalize()  
min\_max\_scaler = MinMaxScaler()  
scale\_cols = ['ApplicantIncome', 'CoapplicantIncome', 'LoanAmount']  
df[scale\_cols] = min\_max\_scaler.fit\_transform(df[scale\_cols])  
scaler = StandardScaler()  
df[['Credit\_History']] = scaler.fit\_transform(df[['Credit\_History']])  
print("\nCleaned Data Summary:")  
print([df.info](http://df.info)())  
print(df.head())



**Result:**

Thus the EDA-Data Cleaning is done successfully.