Data Collection and Preprocessing Phase

Date	18-06-2025
Team ID	SWTID1749841176
Project Title	Online Payments Fraud Detection using Machine Learning
Maximum Marks	6 Marks

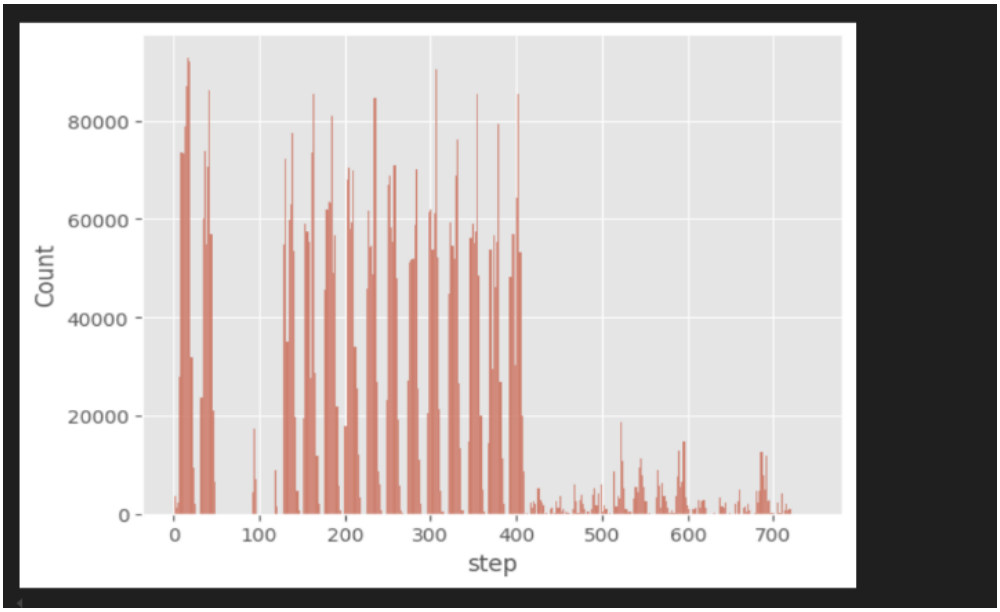
Data Exploration and Preprocessing Report

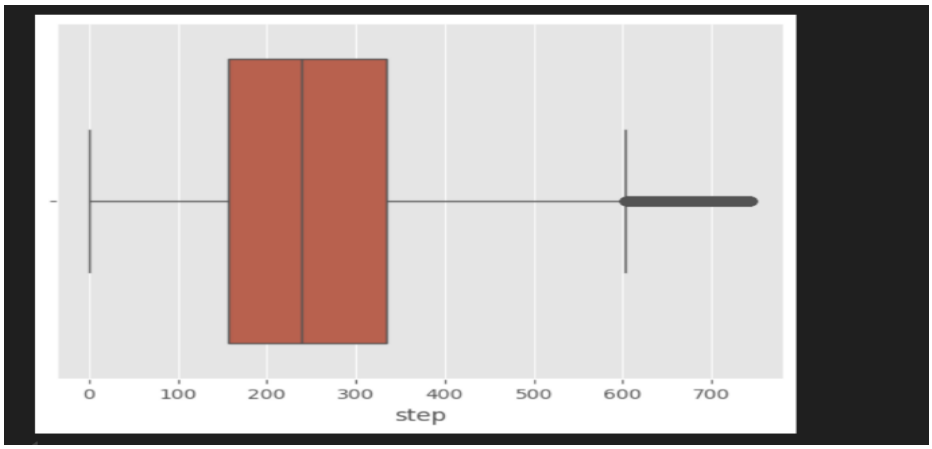
This project focuses on detecting online payment fraud using machine learning. We explored the dataset to understand feature distributions, detect outliers, and observe class imbalance. Key visualizations and statistical summaries highlighted patterns useful for model training. In preprocessing, we normalized features, encoded categorical variables, and ensured the dataset was clean and suitable for accurate fraud detection.

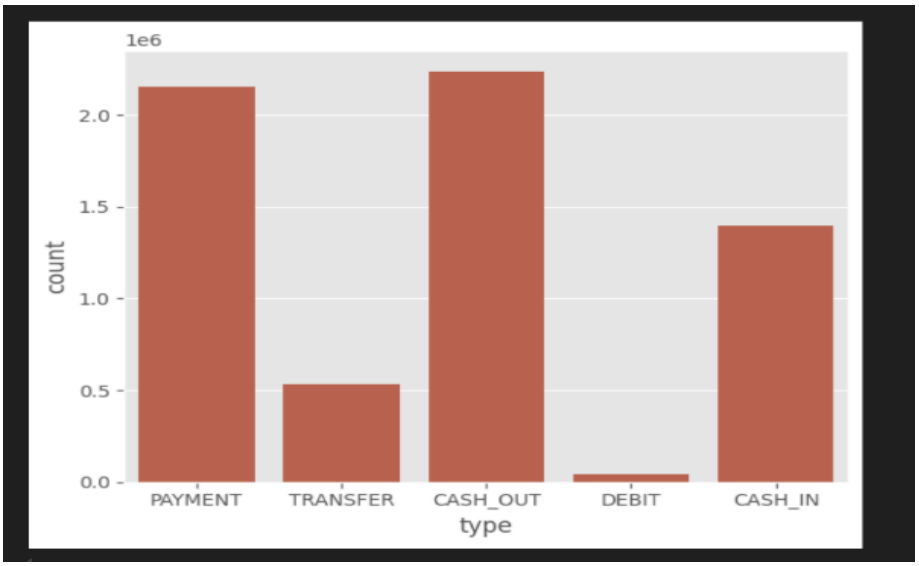
Section	Description
Data Overview	Dimension 6362620 rows × 11 columns Descriptive statistics:

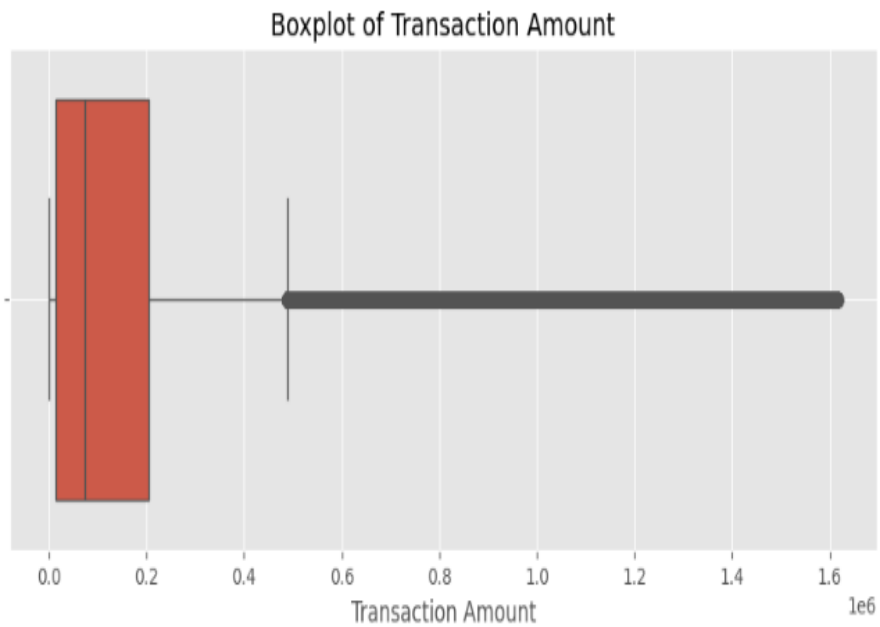
	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	
count	6.362620e+06	6362620	6.362620e+06	6362620	6.362620e+06	6.362620e+06	6362620	6.362620e+06	6.362620e+06	6.3626
unique	NaN	5	NaN	6353307	NaN	NaN	2722362	NaN	NaN	
top	NaN	CASH_OUT	NaN	C1677795071	NaN	NaN	C1286084959	NaN	NaN	
freq	NaN	2237500	NaN	3	NaN	NaN	113	NaN	NaN	
mean	2.433972e+02	NaN	1.798619e+05	NaN	8.338831e+05	8.551137e+05	NaN	1.100702e+06	1.224996e+06	1.2908
std	1.423320e+02	NaN	6.038582e+05	NaN	2.888243e+06	2.924049e+06	NaN	3.399180e+06	3.674129e+06	3.5904
min	1.000000e+00	NaN	0.000000e+00	NaN	0.000000e+00	0.000000e+00	NaN	0.000000e+00	0.000000e+00	0.0000
25%	1.560000e+02	NaN	1.338957e+04	NaN	0.000000e+00	0.000000e+00	NaN	0.000000e+00	0.000000e+00	0.0000
50%	2.390000e+02	NaN	7.487194e+04	NaN	1.420800e+04	0.000000e+00	NaN	1.327057e+05	2.146614e+05	0.0000
75%	3.350000e+02	NaN	2.087215e+05	NaN	1.073152e+05	1.442584e+05	NaN	9.430367e+05	1.111909e+06	0.0000
max	7.430000e+02	NaN	9.244552e+07	NaN	5.958504e+07	4.958504e+07	NaN	3.560159e+08	3.561793e+08	1.0000

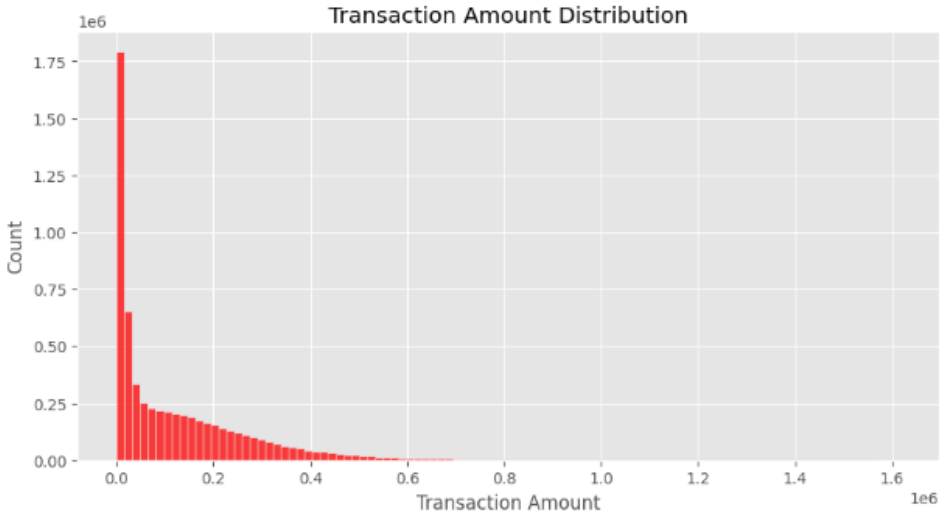
Univariate Analysis

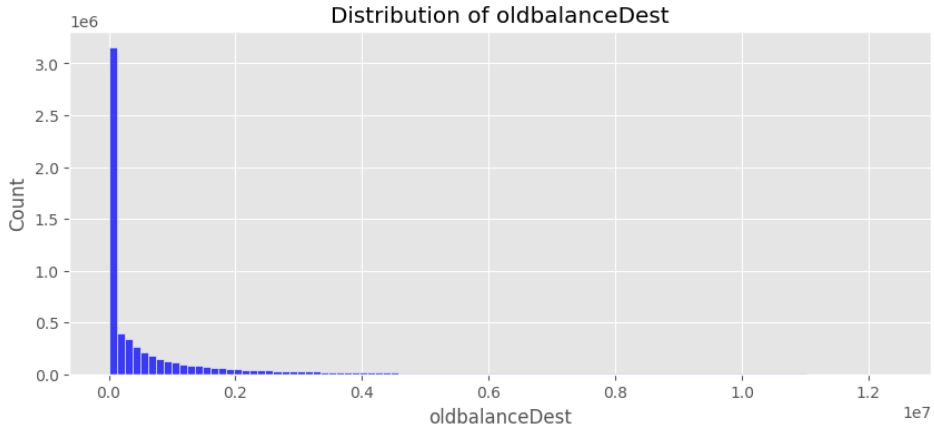


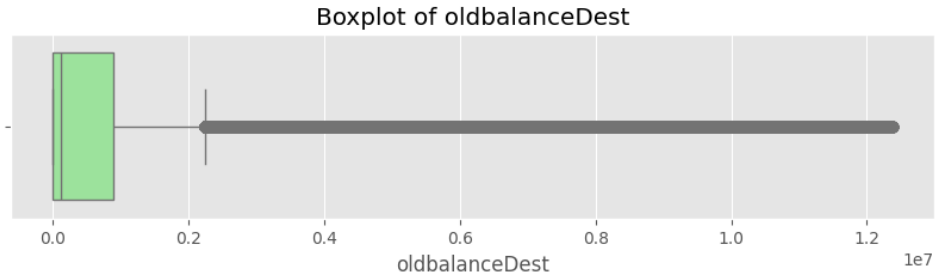

0
100
200
300
400
500
600
700
step

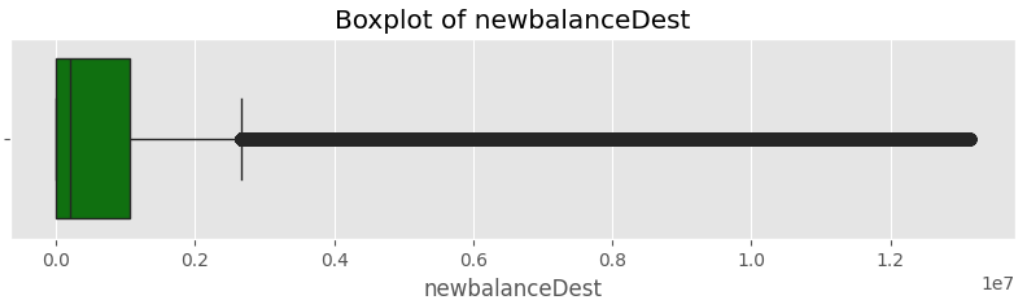

1e6
2.0
1.5
1.0
0.5
0.0
count
PAYMENT
TRANSFER
CASH_OUT
DEBIT
CASH_IN
type

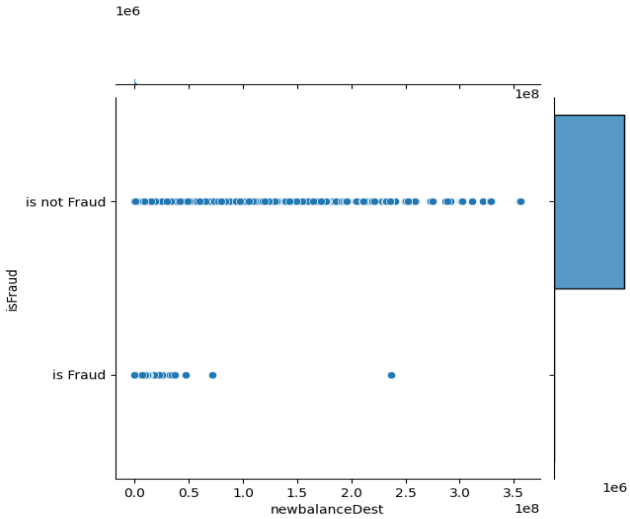
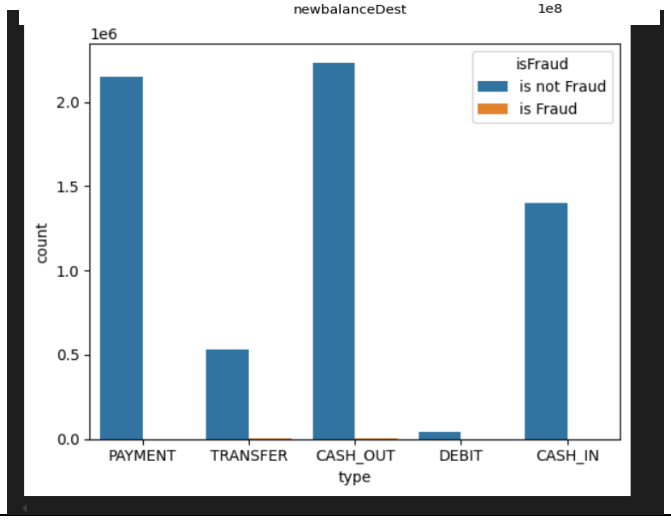

Boxplot of Transaction Amount
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Transaction Amount
1e6

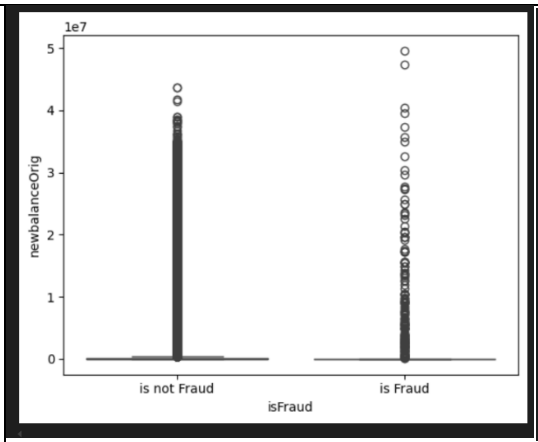
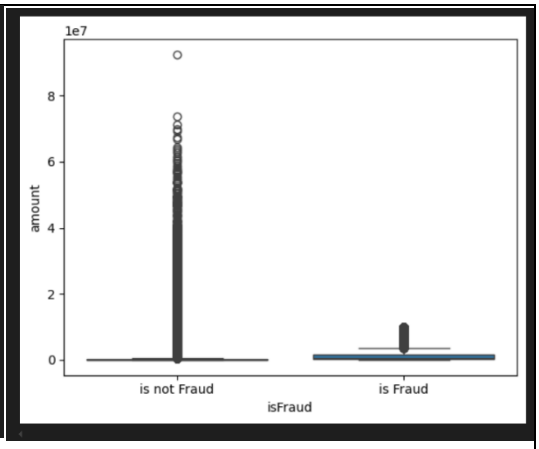
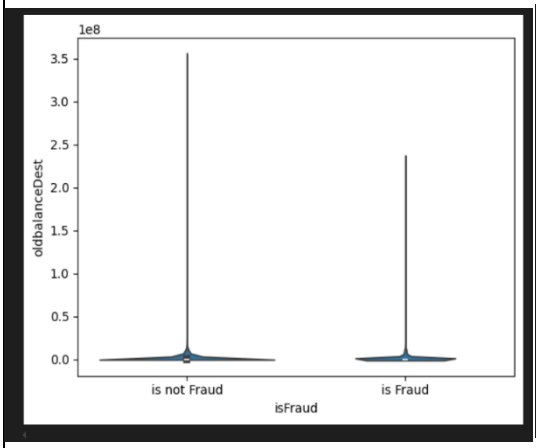
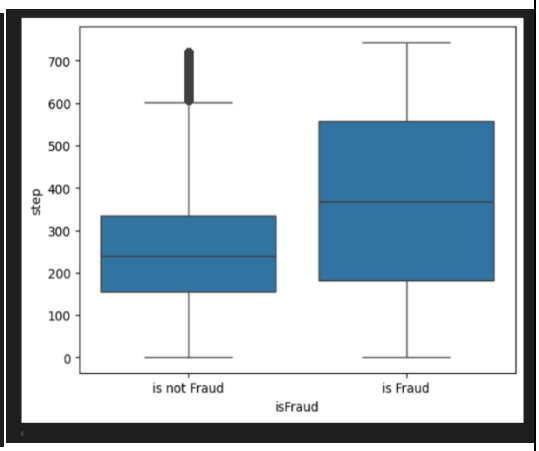
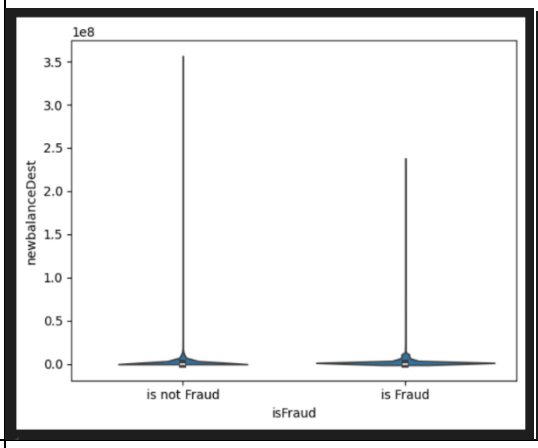
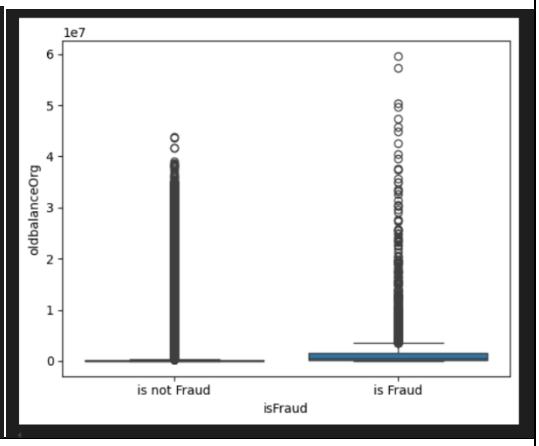

Transaction Amount Distribution
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Count
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Transaction Amount
1e6

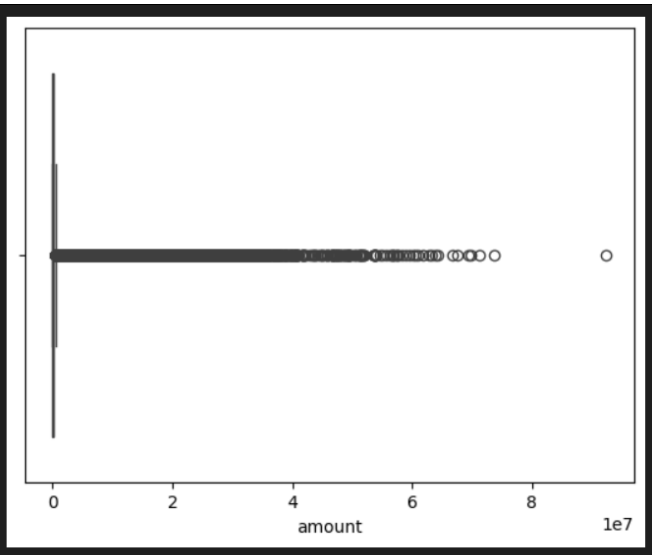

Distribution of oldbalanceDest
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Count
0.0
0.2
0.4
0.6
0.8
1.0
1.2
oldbalanceDest
1e7


Boxplot of oldbalanceDest
0.0
0.2
0.4
0.6
0.8
1.0
1.2
oldbalanceDest
1e7


Boxplot of newbalanceDest
0.0
0.2
0.4
0.6
0.8
1.0
1.2
newbalanceDest
1e7

Bivariate Analysis	 

	     
Outliers And Anomalies	



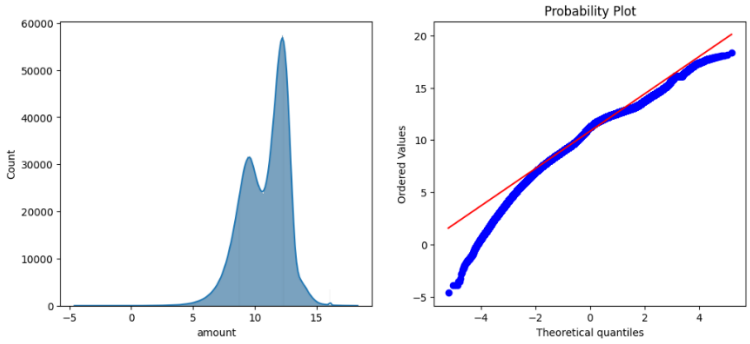

```python
def transformationPlot(feature):
    plt.figure(figsize=(12,5))
    plt.subplot(1,2,1)
    sns.histplot(feature, kde=True)
    plt.subplot(1,2,2)
    stats.probplot(feature, plot=plt)
    plt.show()

filtered_amount = df['amount'][df['amount'] > 0]
log_amount = np.log(filtered_amount)

transformationPlot(log_amount)
```

```python
    q1 = np.quantile(df['amount'],0.25)
    q3 = np.quantile(df['amount'],0.75)
    IQR = q3-q1
    upper_bound = q3+(1.5*IQR)
    lower_bound = q1-(1.5*IQR)
    print('Q1: ',q1)
    print('Q3: ',q3)
    print('IQR: ',IQR)
    print('Upper Bound: ',upper_bound)
    print('Lower Bound: ',lower_bound)
    print('Skewed Data: ',len(df[df['amount']>upper_bound]))
    print('Skewed Data: ',len(df[df['amount']<lower_bound]))
```

```
Q1:  13389.57
Q3:  208721.4775
IQR:  195331.9075
Upper Bound:  501719.33875
Lower Bound:  -279608.29125
Skewed Data:  338078
Skewed Data:  0
```



Data Preprocessing Code Screenshots

Loading Data

```
df = pd.DataFrame(data)

df.head()
```

✓ 0.0s

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

Handling Missing Values

-

Data Transform a-tion

```
la = LabelEncoder()
df['type'] = la.fit_transform(df['type'])
df['type'].value_counts()
```

```
type
1    2237500
3    2151495
0    1399284
4     532909
2      41432
Name: count, dtype: int64
```

```
df.loc[df['isFraud'] == 1,'isFraud'] = 'Fraud'
df.loc[df['isFraud'] == 0,'isFraud'] = 'Not Fraud'
```

Feature Engineerin g

Attached the codes in final submission.

Save Processed Data

-