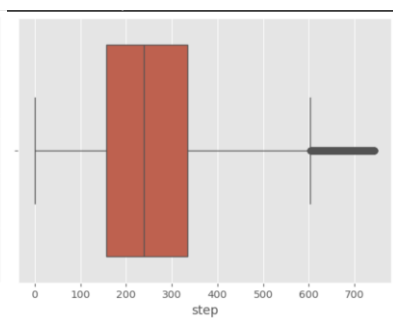
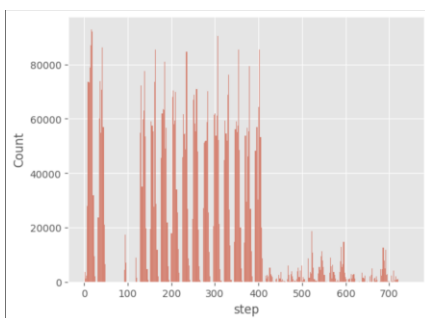


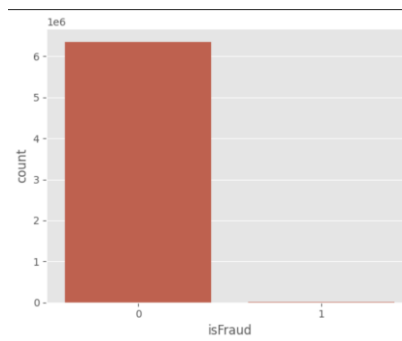
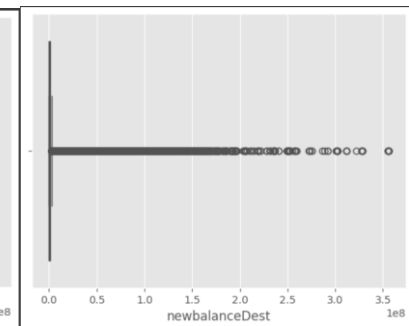
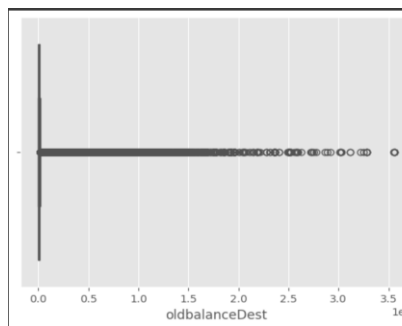
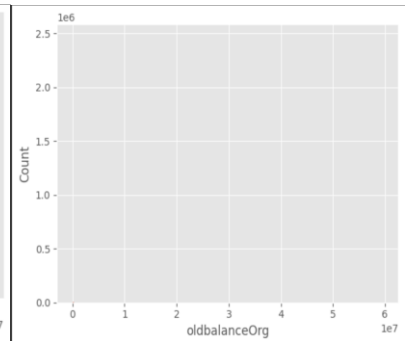
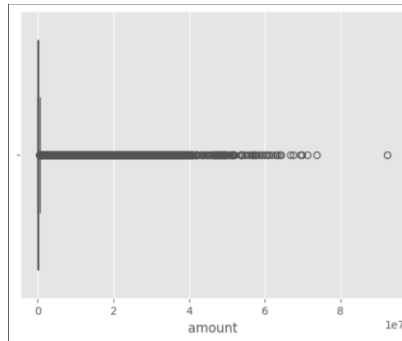
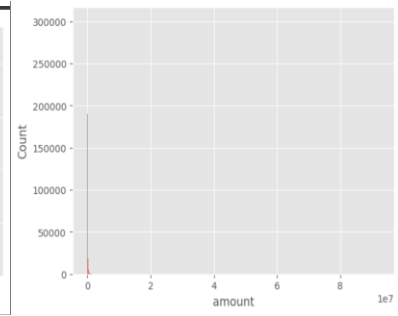
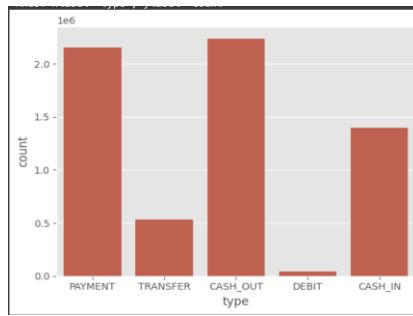
## Data Collection and Preprocessing Phase

Date	13 June 2025
Team ID	SWTID1749662491
Project Title	Online Payments Fraud Detection using Machine Learning
Maximum Marks	6 Marks

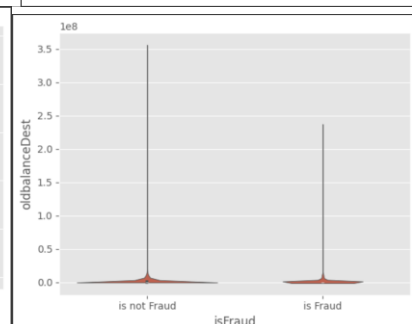
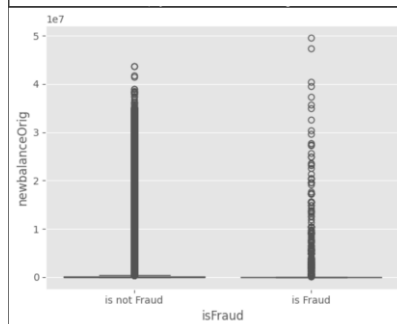
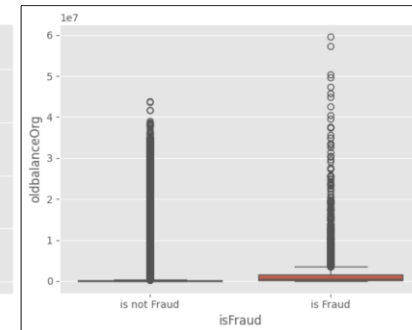
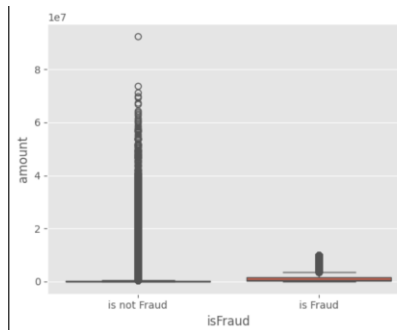
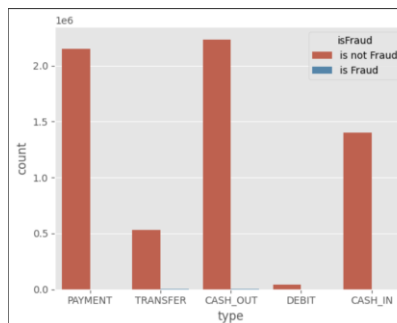
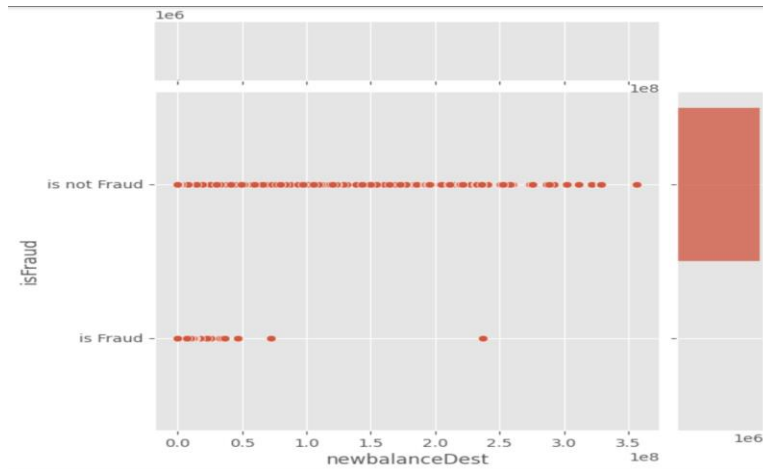
## Data Exploration and Preprocessing

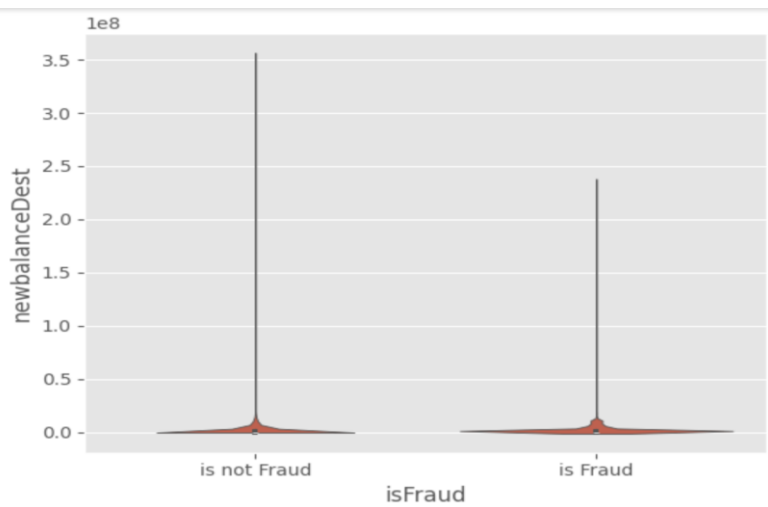
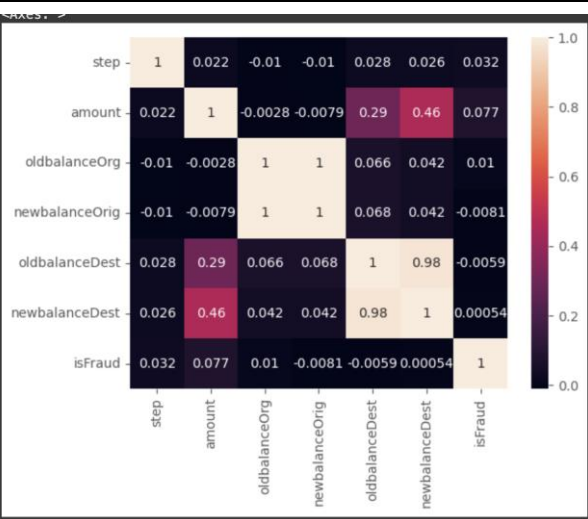
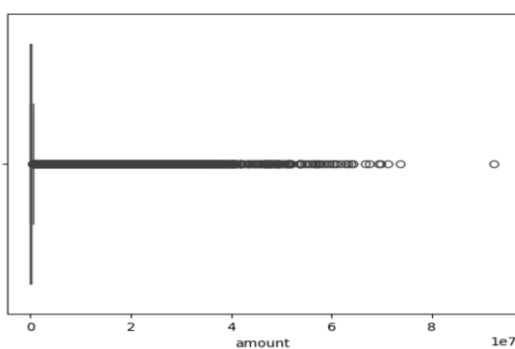
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description																																																																																																																																				
Data Overview	Dimensions: 6362620 rows × 10 columns																																																																																																																																				
	Descriptive Statistics:																																																																																																																																				
	<table><tr><th></th><th>step</th><th>type</th><th>amount</th><th>nameOrig</th><th>oldbalanceOrig</th><th>newbalanceOrig</th><th>nameDest</th><th>oldbalanceDest</th><th>newbalanceDest</th><th>isFraud</th></tr><tr><td>count</td><td>6.362620e+06</td><td>6362620</td><td>6.362620e+06</td><td>6362620</td><td>6.362620e+06</td><td>6.362620e+06</td><td>6362620</td><td>6.362620e+06</td><td>6.362620e+06</td><td>6362620</td></tr><tr><td>unique</td><td>NaN</td><td>5</td><td>NaN</td><td>6353307</td><td>NaN</td><td>NaN</td><td>2722362</td><td>NaN</td><td>NaN</td><td>2</td></tr><tr><td>top</td><td>NaN</td><td>CASH_OUT</td><td>NaN</td><td>C1530544995</td><td>NaN</td><td>NaN</td><td>C1286084959</td><td>NaN</td><td>NaN</td><td>Is not Fraud</td></tr><tr><td>freq</td><td>NaN</td><td>2237500</td><td>NaN</td><td>3</td><td>NaN</td><td>NaN</td><td>113</td><td>NaN</td><td>NaN</td><td>6354407</td></tr><tr><td>mean</td><td>2.433972e+02</td><td>NaN</td><td>1.798619e+05</td><td>NaN</td><td>8.338831e+05</td><td>8.551137e+05</td><td>NaN</td><td>1.100702e+06</td><td>1.224995e+06</td><td>NaN</td></tr><tr><td>std</td><td>1.423320e+02</td><td>NaN</td><td>6.038582e+05</td><td>NaN</td><td>2.888243e+06</td><td>2.924049e+06</td><td>NaN</td><td>3.399180e+06</td><td>3.674129e+06</td><td>NaN</td></tr><tr><td>min</td><td>1.000000e+00</td><td>NaN</td><td>0.000000e+00</td><td>NaN</td><td>0.000000e+00</td><td>0.000000e+00</td><td>NaN</td><td>0.000000e+00</td><td>0.000000e+00</td><td>NaN</td></tr><tr><td>25%</td><td>1.560000e+02</td><td>NaN</td><td>1.338957e+04</td><td>NaN</td><td>0.000000e+00</td><td>0.000000e+00</td><td>NaN</td><td>0.000000e+00</td><td>0.000000e+00</td><td>NaN</td></tr><tr><td>50%</td><td>2.390000e+02</td><td>NaN</td><td>7.467194e+04</td><td>NaN</td><td>1.420800e+04</td><td>0.000000e+00</td><td>NaN</td><td>1.327057e+05</td><td>2.146614e+05</td><td>NaN</td></tr><tr><td>75%</td><td>3.350000e+02</td><td>NaN</td><td>2.087215e+05</td><td>NaN</td><td>1.073152e+05</td><td>1.442584e+05</td><td>NaN</td><td>9.430367e+05</td><td>1.111909e+06</td><td>NaN</td></tr><tr><td>max</td><td>7.430000e+02</td><td>NaN</td><td>9.244552e+07</td><td>NaN</td><td>5.958504e+07</td><td>4.958504e+07</td><td>NaN</td><td>3.560159e+08</td><td>3.561793e+08</td><td>NaN</td></tr></table>		step	type	amount	nameOrig	oldbalanceOrig	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	count	6.362620e+06	6362620	6.362620e+06	6362620	6.362620e+06	6.362620e+06	6362620	6.362620e+06	6.362620e+06	6362620	unique	NaN	5	NaN	6353307	NaN	NaN	2722362	NaN	NaN	2	top	NaN	CASH_OUT	NaN	C1530544995	NaN	NaN	C1286084959	NaN	NaN	Is not Fraud	freq	NaN	2237500	NaN	3	NaN	NaN	113	NaN	NaN	6354407	mean	2.433972e+02	NaN	1.798619e+05	NaN	8.338831e+05	8.551137e+05	NaN	1.100702e+06	1.224995e+06	NaN	std	1.423320e+02	NaN	6.038582e+05	NaN	2.888243e+06	2.924049e+06	NaN	3.399180e+06	3.674129e+06	NaN	min	1.000000e+00	NaN	0.000000e+00	NaN	0.000000e+00	0.000000e+00	NaN	0.000000e+00	0.000000e+00	NaN	25%	1.560000e+02	NaN	1.338957e+04	NaN	0.000000e+00	0.000000e+00	NaN	0.000000e+00	0.000000e+00	NaN	50%	2.390000e+02	NaN	7.467194e+04	NaN	1.420800e+04	0.000000e+00	NaN	1.327057e+05	2.146614e+05	NaN	75%	3.350000e+02	NaN	2.087215e+05	NaN	1.073152e+05	1.442584e+05	NaN	9.430367e+05	1.111909e+06	NaN	max	7.430000e+02	NaN	9.244552e+07	NaN	5.958504e+07	4.958504e+07	NaN	3.560159e+08	3.561793e+08	NaN
	step	type	amount	nameOrig	oldbalanceOrig	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud																																																																																																																											
count	6.362620e+06	6362620	6.362620e+06	6362620	6.362620e+06	6.362620e+06	6362620	6.362620e+06	6.362620e+06	6362620																																																																																																																											
unique	NaN	5	NaN	6353307	NaN	NaN	2722362	NaN	NaN	2																																																																																																																											
top	NaN	CASH_OUT	NaN	C1530544995	NaN	NaN	C1286084959	NaN	NaN	Is not Fraud																																																																																																																											
freq	NaN	2237500	NaN	3	NaN	NaN	113	NaN	NaN	6354407																																																																																																																											
mean	2.433972e+02	NaN	1.798619e+05	NaN	8.338831e+05	8.551137e+05	NaN	1.100702e+06	1.224995e+06	NaN																																																																																																																											
std	1.423320e+02	NaN	6.038582e+05	NaN	2.888243e+06	2.924049e+06	NaN	3.399180e+06	3.674129e+06	NaN																																																																																																																											
min	1.000000e+00	NaN	0.000000e+00	NaN	0.000000e+00	0.000000e+00	NaN	0.000000e+00	0.000000e+00	NaN																																																																																																																											
25%	1.560000e+02	NaN	1.338957e+04	NaN	0.000000e+00	0.000000e+00	NaN	0.000000e+00	0.000000e+00	NaN																																																																																																																											
50%	2.390000e+02	NaN	7.467194e+04	NaN	1.420800e+04	0.000000e+00	NaN	1.327057e+05	2.146614e+05	NaN																																																																																																																											
75%	3.350000e+02	NaN	2.087215e+05	NaN	1.073152e+05	1.442584e+05	NaN	9.430367e+05	1.111909e+06	NaN																																																																																																																											
max	7.430000e+02	NaN	9.244552e+07	NaN	5.958504e+07	4.958504e+07	NaN	3.560159e+08	3.561793e+08	NaN																																																																																																																											
Univariate Analysis	<div></div>																																																																																																																																				



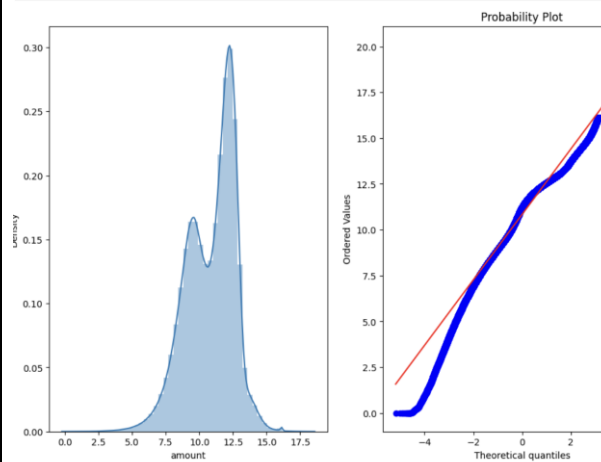
## Bivariate Analysis



	
Multivariate Analysis	
Outliers and Anomalies	<pre>sns.boxplot(x=df['amount'])</pre> <p>&lt;Axes: xlabel='amount'&gt;</p> 

```
q1 : 13389.57
q2 : 208721.4775
IQR : 195331.9075
Upper Bound : 501719.33875
Lower Bound : -279608.29125
Skewed data : 338078
Skewed data : 0
```

```
transformationPlot(np.log1p(df['amount']))
```



## Data Preprocessing Code Screenshots

### Loading Data

```
[2] import kagglehub

# Download latest version
path = kagglehub.dataset_download("rupakroy/online-payments-fraud-detection-dataset")

print("Path to dataset files:", path)

Path to dataset files: /kaggle/input/online-payments-fraud-detection-dataset

[3] import os
for root,dirs,files in os.walk(path):
    for file in files:
        print(file)

PS_20174392719_1491204439457_log.csv

[4] data=os.path.join(path,"PS_20174392719_1491204439457_log.csv")
df=pd.read_csv(data)
```

Handling Missing Data	<pre>df.isnull().sum()</pre>  <p>There was no missing data.</p>
Data Transformation	<p><b>OBJECT DATA LABEL_ENCODING</b></p> <pre>from sklearn.preprocessing import LabelEncoder le=LabelEncoder() df['type']=le.fit_transform(df['type']) df['type'].value_counts()</pre> 
Feature Engineering	-
Save Processed Data	-