Data Collection and Preprocessing Phase

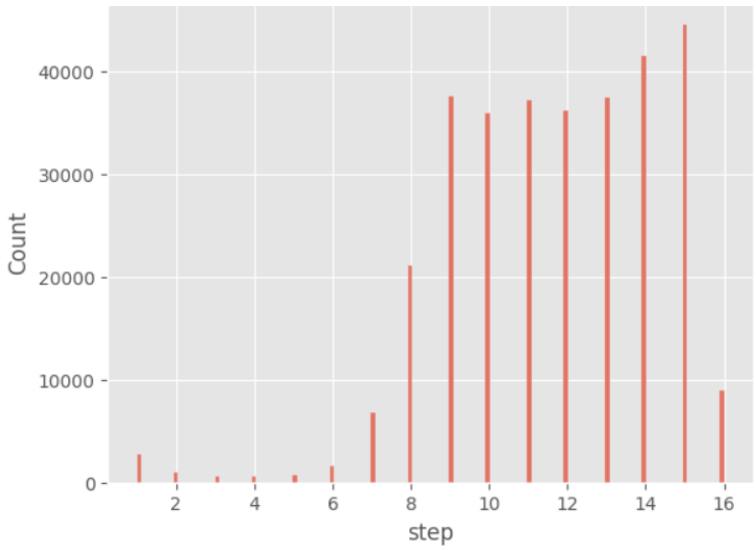
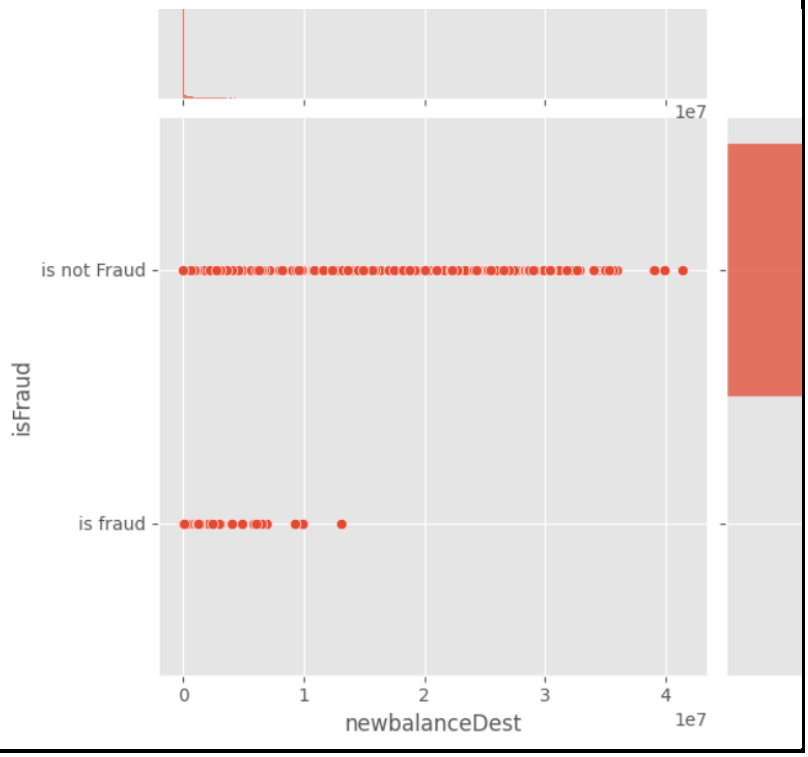
Date	10 June 2024
Team ID	739991
Project Title	Online Payment Fraud Detection
Maximum Marks	6 Marks

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.

Section	Description
Data Overview	<u>Dimension:</u> 314699rows × 11columns <u>Descriptive statistics:</u> data 314699 rows × 11 columns
Univariate Analysis	

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.00	0.00	0.0	0.0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.00	0.00	0.0	0.0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.00	0.00	1.0	0.0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.00	0.00	1.0	0.0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.00	0.00	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
314694	16	CASH_OUT	359950.11	C622142425	11719.0	0.00	C1848110361	2181900.90	2717203.17	0.0	0.0
314695	16	PAYMENT	4032.79	C861304824	3509.0	0.00	M982320760	0.00	0.00	0.0	0.0
314696	16	CASH_IN	114371.60	C829816201	20467.0	134838.60	C671501829	747527.75	633156.15	0.0	0.0
314697	16	PAYMENT	13415.82	C1030884245	82291.0	68875.18	M1025196660	0.00	0.00	0.0	0.0
314698	16	CASH_IN	65.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

	
Bivariate Analysis	

Descriptive Analysis	

```
data.describe(include='all')
```

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud
count	314699.000000	0.0	3.146990e+05	314698	3.146980e+05	3.146980e+05	314698	3.146980e+05	3.146980e+05	314698
unique	NaN	NaN	NaN	314680	NaN	NaN	140904	NaN	NaN	2
top	NaN	NaN	NaN	C1842781381	NaN	NaN	C985934102	NaN	NaN	is not Fraud
freq	NaN	NaN	NaN	2	NaN	NaN	85	NaN	NaN	314511
mean	11.636214	NaN	1.757438e+05	NaN	8.865920e+05	9.054732e+05	NaN	9.786492e+05	1.194413e+06	NaN
std	2.725433	NaN	2.986535e+05	NaN	2.867410e+06	2.904734e+06	NaN	2.367532e+06	2.587206e+06	NaN
min	1.000000	NaN	3.000000e-01	NaN	0.000000e+00	0.000000e+00	NaN	0.000000e+00	0.000000e+00	NaN
25%	10.000000	NaN	1.286067e+04	NaN	0.000000e+00	0.000000e+00	NaN	0.000000e+00	0.000000e+00	NaN
50%	12.000000	NaN	7.864058e+04	NaN	1.835857e+04	0.000000e+00	NaN	8.890536e+04	1.949259e+05	NaN
75%	14.000000	NaN	2.317330e+05	NaN	1.812713e+05	2.201000e+05	NaN	8.624901e+05	1.236109e+06	NaN
max	16.000000	NaN	1.000000e+07	NaN	3.893942e+07	3.894623e+07	NaN	4.133844e+07	4.138365e+07	NaN

<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>

```python
data=pd.read_csv("/content/PS_20174392719_1491204439457_log.csv")
```

```python
data
```

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.0	0.00	0.0	0.0
1	1	PAYMENT	1864.28	C1666544295	21249.00	19384.72	M2044282225	0.0	0.00	0.0	0.0
2	1	TRANSFER	181.00	C1305486145	181.00	0.00	C553264065	0.0	0.00	1.0	0.0
3	1	CASH_OUT	181.00	C840083671	181.00	0.00	C38997010	21182.0	0.00	1.0	0.0
4	1	PAYMENT	11668.14	C2048537720	41554.00	29885.86	M1230701703	0.0	0.00	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
56198	9	CASH_OUT	16024.60	C1088493558	442118.00	426093.40	C1084323592	5818.0	8074.67	0.0	0.0
56199	9	PAYMENT	20502.92	C410885495	3073.00	0.00	M1731153077	0.0	0.00	0.0	0.0
56200	9	CASH_IN	175858.36	C702220078	290164.69	466023.05	C65594254	24083.0	0.00	0.0	0.0
56201	9	PAYMENT	2955.89	C1632500548	466023.05	463067.17	M363811903	0.0	0.00	0.0	0.0
56202	9	PAYMEN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

56203 rows × 11 columns

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

```
step              0
amount            1
oldbalanceOrg     1
newbalanceOrig    1
oldbalanceDest    1
newbalanceDest    1
dtype: int64
step              0
amount            0
oldbalanceOrg     0
newbalanceOrig    0
oldbalanceDest    0
newbalanceDest    0
dtype: int64
```

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

```python
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
y_train1=le.fit_transform(y_train)
y_true = [0, 1, 1, 0]

y_pred = ['is fraud', 'is not Fraud', 'is not Fraud', 'is fraud']


label_encoder = LabelEncoder()
y_pred_numeric = label_encoder.fit_transform(y_pred)

accuracy = accuracy_score(y_true, y_pred_numeric)
print(f'Accuracy: {accuracy}')
```

</td></tr>
<tr><td>Feature Engineering</td><td>Attached the codes in final submission.</td></tr>
</table>

Save Processed Data	-