

Download datasets:

be sure you download Credit Card Fraud Detection—Cleaned & ML- Ready

```
import os
print(os.listdir("/kaggle/input"))
```

Code

```
# =====
# Credit Card Dataset Visualization
# Original vs Cleaned Comparison
# =====

from pandas import read_csv

from matplotlib.pyplot import figure
from matplotlib.pyplot import title
from matplotlib.pyplot import legend
from matplotlib.pyplot import show
from matplotlib.pyplot import subplot
from matplotlib.pyplot import tight_layout
from matplotlib.pyplot import xticks

from seaborn import set
from seaborn import kdeplot
from seaborn import boxplot
from seaborn import countplot

set(style="whitegrid")

# -----
# Load Datasets
# -----

original_df = read_csv(r"/kaggle/input/datasets/organizations/mlg-
ulb/creditcardfraud/creditcard.csv")
cleaned_df = read_csv(r"/kaggle/input/datasets/bornaetminan/credit-
card-fraud-detectioncleaned-and-ml-ready/creditcard_cleaned.csv")

print("Original Shape:", original_df.shape)
print("Cleaned Shape:", cleaned_df.shape)
```

```

# =====
# 1 Amount Distribution Comparison
# =====

figure(figsize=(10,5))
kdeplot(original_df["Amount"], label="Original", fill=True)
kdeplot(cleaned_df["amount"], label="Cleaned", fill=True)
title("Transaction Amount Distribution (Original vs Cleaned)")
legend()
show()

# =====
# 2 Boxplot (Outlier Visualization)
# =====

figure(figsize=(10,5))
subplot(1,2,1)
boxplot(y=original_df["Amount"])
title("Original Amount")

subplot(1,2,2)
boxplot(y=cleaned_df["amount"])
title("Cleaned Amount")

tight_layout()
show()

# =====
# 3 Duplicate Comparison
# =====

original_duplicates = original_df.duplicated().sum()
cleaned_duplicates = cleaned_df.duplicated().sum()

print("Original Duplicates:", original_duplicates)
print("Cleaned Duplicates:", cleaned_duplicates)

# =====
# 4 Hour Feature Distribution (New Feature)
# =====

figure(figsize=(8,5))
countplot(x=cleaned_df["hour"])
title("Transaction Distribution by Hour (Cleaned Dataset)")
xticks(rotation=45)
show()

# =====

```

`# 5 Class Distribution Comparison`

```
# =====
```

```
figure(figsize=(10,5))
```

```
subplot(1,2,1)
countplot(x=original_df["Class"])
title("Original Class Distribution")
```

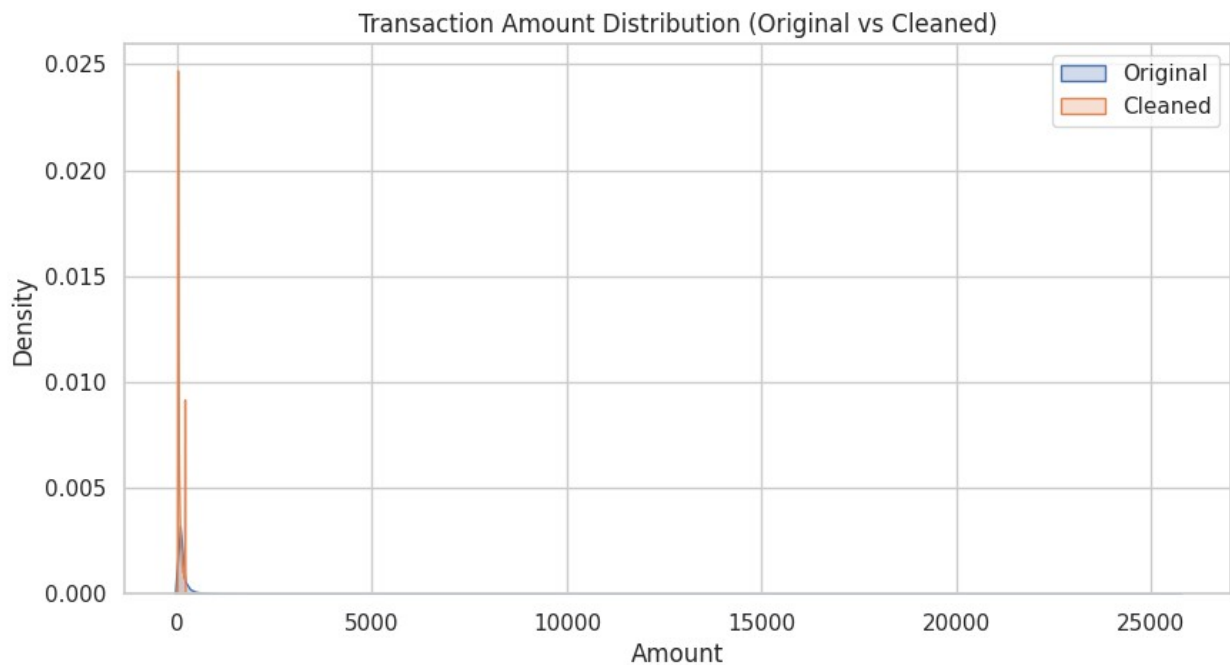
```
subplot(1,2,2)
countplot(x=cleaned_df["class"])
title("Cleaned Class Distribution")
```

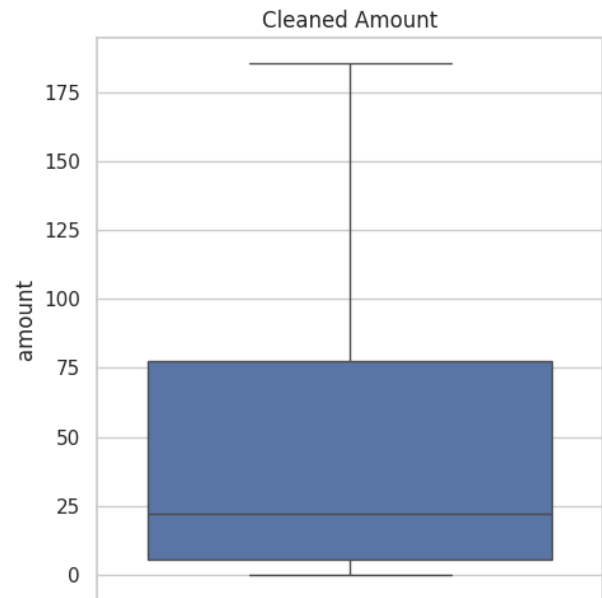
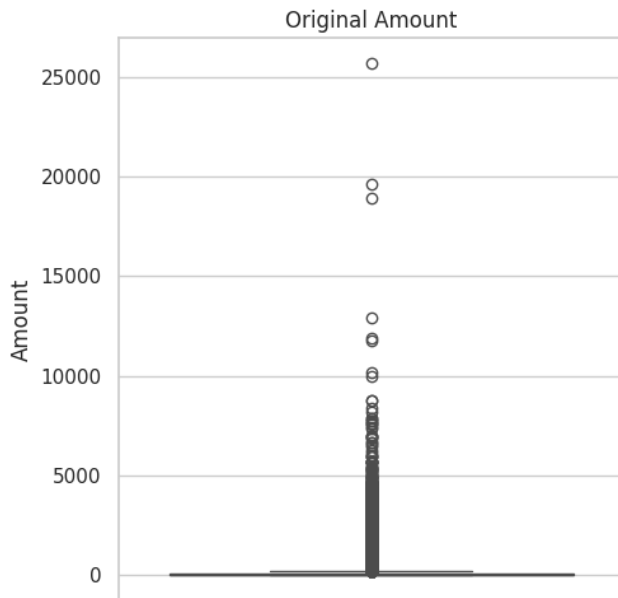
```
tight_layout()
show()
```

```
print("\n Visualization comparison completed successfully.")
```

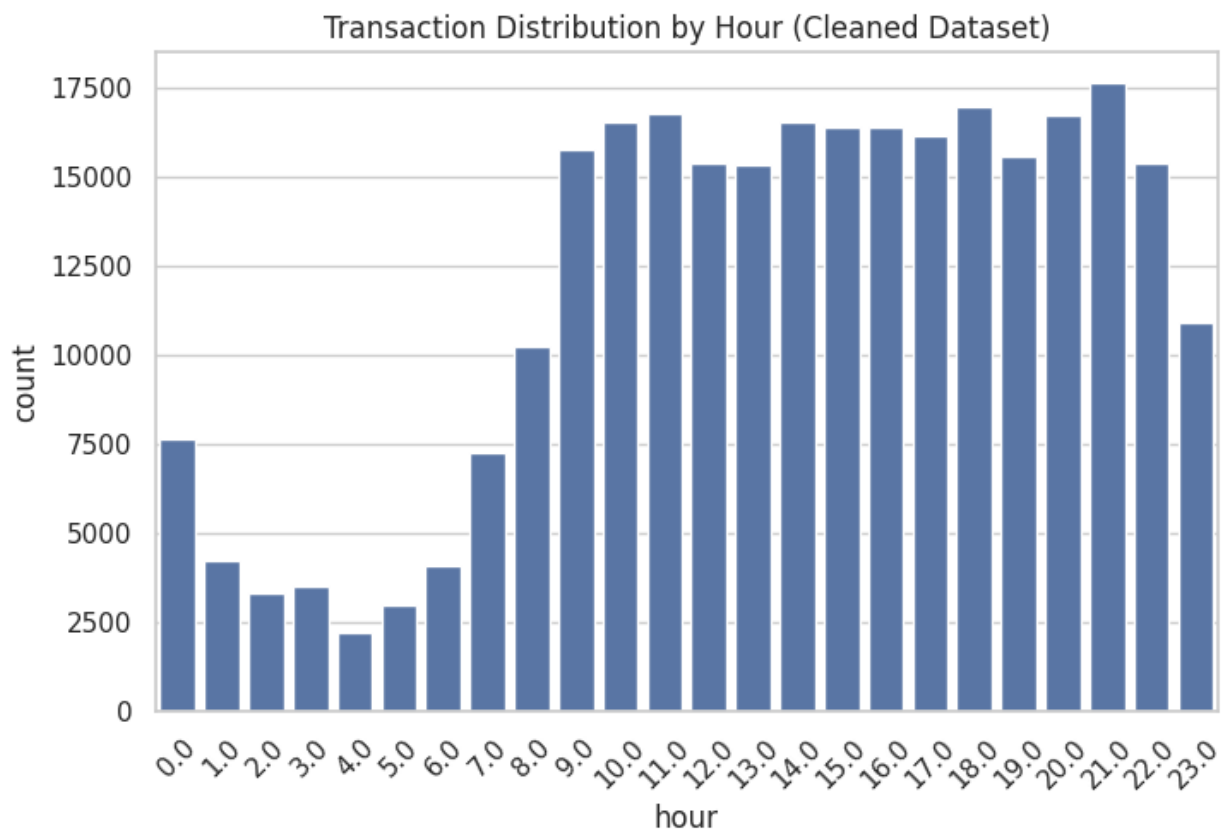
Original Shape: (284807, 31)

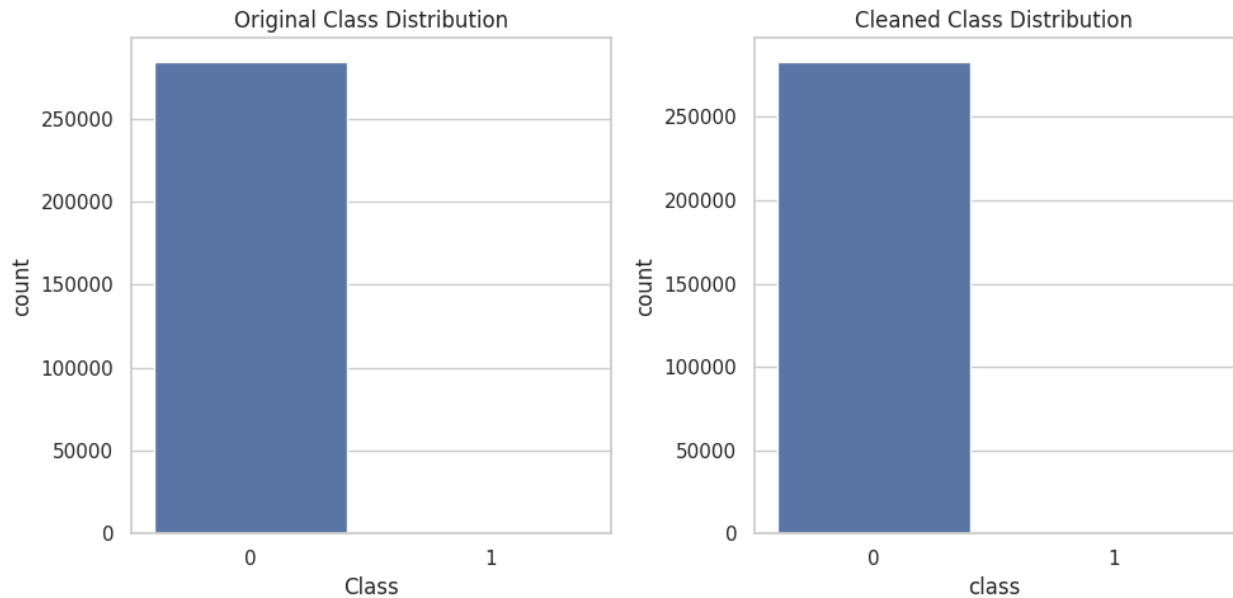
Cleaned Shape: (283726, 32)





Original Duplicates: 1081
Cleaned Duplicates: 0





Visualization comparison completed successfully.

Description

Credit Card Dataset — Original vs Cleaned Comparison

Overview

This section compares the original credit card transaction dataset with the cleaned version generated by the preprocessing pipeline.

The goal is to demonstrate the impact of data cleaning, outlier handling, and feature engineering through visual evidence.

Transaction Amount Distribution

The KDE distribution plot shows:

- Original dataset → Heavy tail and extreme values
- Cleaned dataset → Controlled distribution after IQR-based outlier treatment

This confirms successful normalization of extreme transaction amounts.

Boxplot Comparison

The boxplots clearly illustrate:

- Original dataset contains extreme outliers
- Cleaned dataset shows capped values within statistical bounds

This improves model stability and reduces skew influence.

Duplicate Removal

Duplicate rows were removed during preprocessing.

This ensures:

- Reduced bias
 - More reliable model training
 - Improved data integrity
-

Feature Engineering — Hour Extraction

A new feature (`hour`) was extracted from transaction time.

This enables:

- Behavioral pattern analysis
- Time-based fraud modeling
- Advanced feature engineering opportunities

⚖ Class Distribution

Class imbalance remains preserved after cleaning.

This is critical because:

- Fraud detection requires realistic distribution
 - Oversampling should be applied only during modeling
-

□ Conclusion

The cleaned dataset demonstrates:

- ✓ Reduced outlier impact
- ✓ Removed duplicates
- ✓ Added interpretable features
- ✓ Maintained realistic fraud distribution
- ✓ Improved modeling readiness

The preprocessing pipeline significantly enhances data quality while preserving the integrity of financial transaction behavior.
