Credit Card Fraud Detection

Introduction

The dataset contains transactions made using credit cards by European cardholders in September 2013. All input features are numerical and have undergone a Principal Component Analysis (PCA) transformation. Due to confidentiality constraints, the original features and additional background details are not provided.

Feature Details

This dataset contains 283726 rows and 31 columns

V1, V2, ..., V28: These are the principal components derived from the PCA transformation.

Time: This feature indicates the number of seconds elapsed between each transaction and the first transaction in the dataset.

Amount: Represents the transaction amount, which can be used for cost-sensitive learning based on individual transactions.

Class: The target variable, where:
1 denotes a fraudulent transaction.
0 denotes a legitimate transaction.

Approach

1. **Library Imports:**
    a. Essential libraries like os, numpy, pandas, matplotlib, seaborn, and plotly are imported.
    b. matplotlib inline ensures plots are displayed inline in the notebook.
2. **File Handling:**
    a. The directory containing the dataset is specified, and the CSV file (creditcard.csv) is read into a DataFrame.
    b. The dataset is duplicated and written to a new CSV file named duplicate.csv.
3. **Data Combination:**
    a. Two DataFrames, df and duplicate_df, are concatenated using **pd.concat()** to form a combined DataFrame.
4. **Basic Exploration:**
    a. The shape (number of rows and columns) of the combined DataFrame is printed.
    b. The first few rows of the combined DataFrame are displayed using .head().
5. **Dataset Overview:**
    a. The notebook prints the dimensions of the dataset (shape) to provide a quick overview.
    b. Displays the first few rows using .head() and provides a summary with **.info()**.
6. **Handling Duplicates:**
    a. Focus on analyzing the DataFrame df_no_duplicates, suggesting preprocessing steps to clean the dataset.

Summary Stats

Amount:

- Mean: 88.47 (likely in a monetary unit such as dollars or euros).
- Std Dev: 250.4, indicating a wide spread in transaction amounts.
- Range: [0, 25,691.16], where the minimum is 0 (no transaction) and the maximum is a very high-value transaction.

Class (Target Variable):

- Mean: 0.001667, indicating that ~0.167% of transactions are fraudulent.
- Std Dev: 0.0408, reflecting the dataset's highly imbalanced nature.
- Range: [0, 1], with 0 being non-fraudulent and 1 being fraudulent.

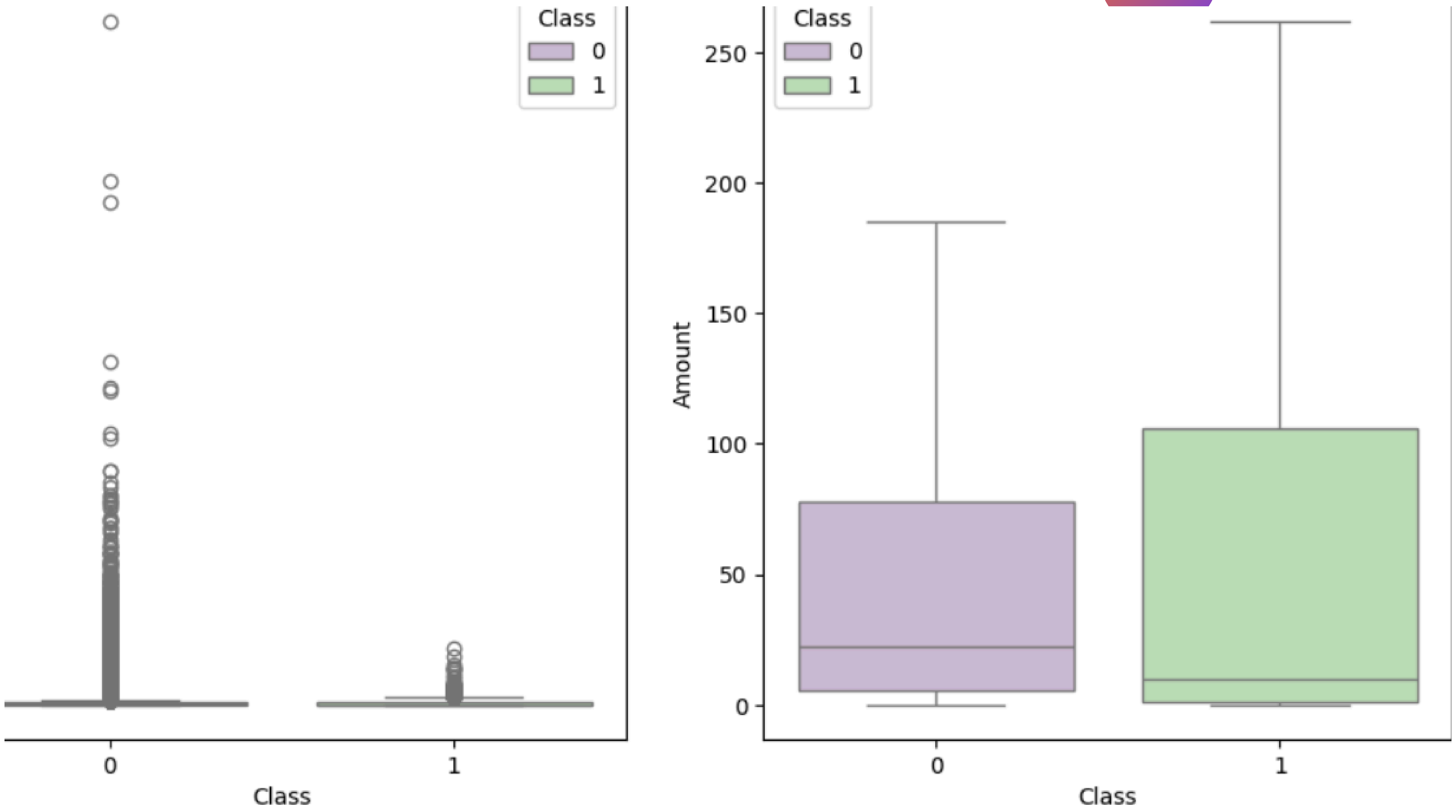
Observation

- **Right-Skewed Distribution:** The difference between the mean (88.41) and the median (22.00), along with the high maximum value, suggests that the data is heavily right-skewed. Most transactions are small, but a few very large transactions significantly affect the mean.

- **High Variability:** The large standard deviation (250.38) and the presence of outliers (evident from the max value) indicate significant variability in transaction amounts.

- **Potential Outliers:** The max value (25,691.16) and the range of data suggest that there are extreme outliers in the dataset. These might need special attention, especially if they are associated with anomalies or fraudulent activities.

Visualizations



Insights

Right Plot (Box Plot):

- Class 0 (Purple):
  - The interquartile range (IQR) is relatively narrow, indicating most of the data is concentrated within a small range of amounts.
  - Outliers are significant and extend to much higher amounts.
- Class 1 (Green):
  - The IQR is wider than that of Class 0, suggesting greater variability in transaction amounts.
  - The median value for Class 1 is higher than for Class 0.
  - Outliers are present but less extreme than in Class 0.

Insights:

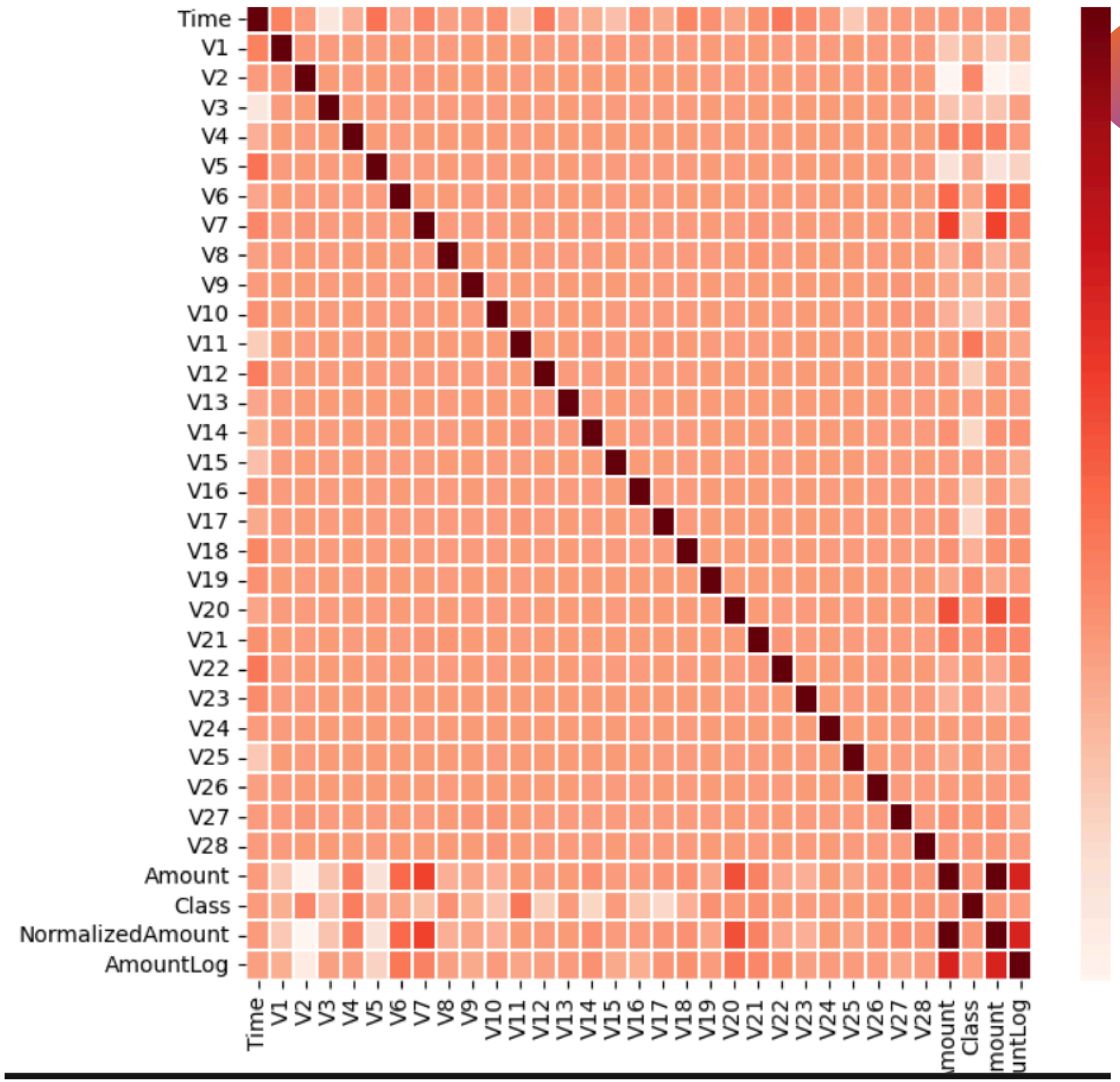
Fraud/Anomaly Detection Perspective:

- If Class 1 represents fraud or anomalies, higher median and variability could indicate fraudulent transactions often involve larger amounts.
- The fewer outliers in Class 1 suggest that fraud detection might focus more on the variability in amounts rather than extreme values.

Data Skewness:

- Both classes exhibit a high concentration of lower transaction amounts, indicating a right-skewed distribution.

Heat-Map



Time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Amount
Class
NormalizedAmount
AmountLog
Time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
mount
Class
mount
untLog

Insights

Diagonal Dominance:

- The diagonal cells are dark red, indicating a perfect correlation of 1.0 (variables are perfectly correlated with themselves).

High Correlation Areas:

- Some off-diagonal cells also show strong correlations (dark or deep red areas), indicating that certain variables are strongly related to each other.
- For example, the "Amount" and "NormalizedAmount" variables appear to have some correlation.

Low Correlation Areas:

- Lighter-colored areas (white to pale orange) indicate weaker or no correlation between the variables.
- Variables like "Class" (likely a categorical target variable) show minimal correlation with most features.

Negative Correlations:

- Any cells with shades of white or light pink represent negative correlations. These are weaker in this dataset.

Feature Relationships:

- Features with very low correlation to "Class" might not be as predictive.

STUDIO SHODWE

Thank You