Exploratory Data Analysis Report: Transaction Fraud Detection

Objective: The primary goal of this EDA is to understand the characteristics of fraudulent versus non-fraudulent transactions to inform the development of a One-Class Classification model. This model aims to identify fraudulent transactions by learning the patterns of "normal" (non-fraudulent) transactions and flagging deviations as anomalies.

1. Dataset Overview

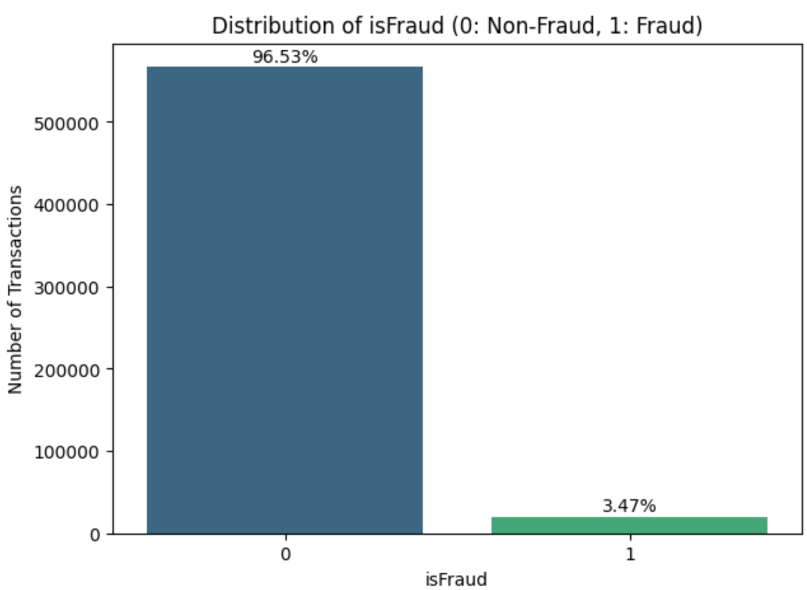
The provided dataset is a preprocessed and cleaned subset from a fraud detection competition, containing approximately 590,000 transactions and 48 features. Key feature categories include:

- **Transaction details**: *TransactionDT* (time delta), *TransactionAmount*.
- **Product information**: *ProductCD*.
- **Card details**: *Card1* - *Card6*.
- **Geographic information**: *Addr1*, *Addr2*, *Dist1*.
- **Purchaser email domain**: P*_emaildomain*.
- **Count-based features**: *C1* - *C14*.
- **Time-based delta features**: *D1* - *D5*, *D10*.
- **Anonymized engineered features**: *V0* - *V13*.

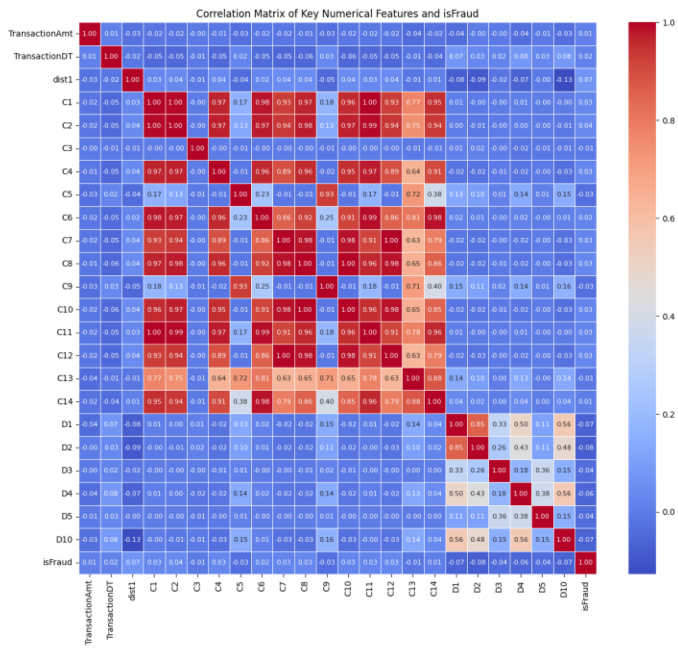
Missing values are present in *Dist1*, *P_emaildomain*, and several *C* and *D* columns, requiring careful handling during preprocessing.

2. Key Insights from Exploratory Data Analysis

2.1. *isFraud* Distribution & Correlation

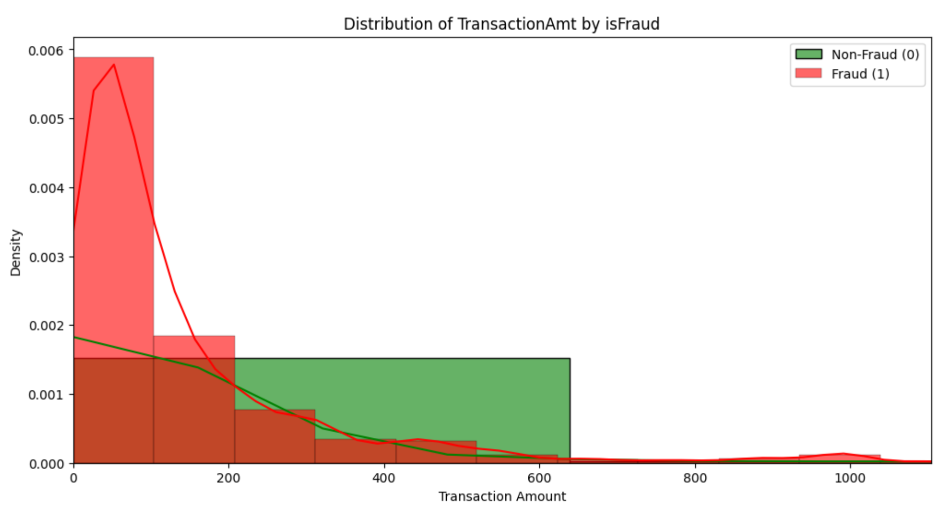


- Severe Class Imbalance: Only **3.47%** of transactions are fraudulent, while **96.53%** are non-fraudulent. This imbalance is crucial for model training and evaluation, making One-Class Classification a suitable approach.



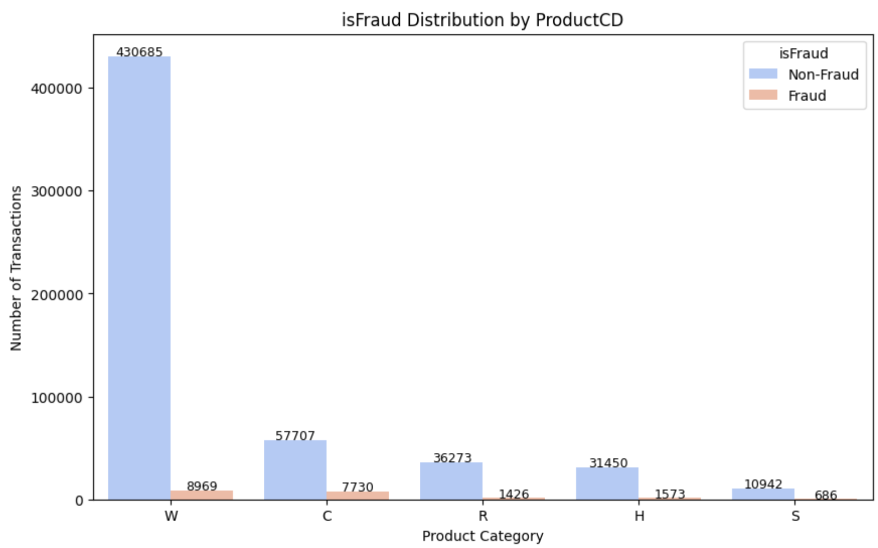
- Feature Correlations:
  - Most numerical features (e.g., *TransactionAmount, TransactionDT, C-features, D-features*) show **very low linear correlations** with *isFraud* (typically between -0.07 and 0.04). This implies that fraudulent patterns are likely non-linear and multi-dimensional.
  - *C-features* (C1-C14) and some *D-features* exhibit **extremely high inter-correlations** (close to 1.00), suggesting redundancy.

2.2. *TransactionDT & TransactionAmount*



- ***TransactionAmount*:** Fraudulent transactions tend to have a **higher density at lower transaction amounts** compared to non-fraudulent ones, but also show a broader spread.
- ***TransactionDT*:** The **ratio of fraudulent transactions over time is highly volatile**, with daily spikes exceeding 6-7% (compared to an overall average of 3.47%). This temporal variability is a significant indicator.

2.3. ProductCD



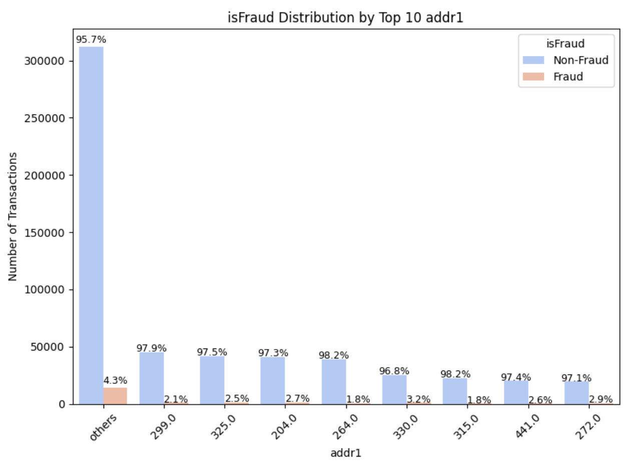
- *ProductCD* is a **highly discriminative feature**:
  - *ProductCD == 'C'* has the **highest fraud rate at 11.81%,** making it a strong signal for fraud.
  - *ProductCD == 'S'* also shows a relatively high fraud rate at **5.90%.**
  - *ProductCD == 'W'* has the lowest fraud rate at **2.04%.**

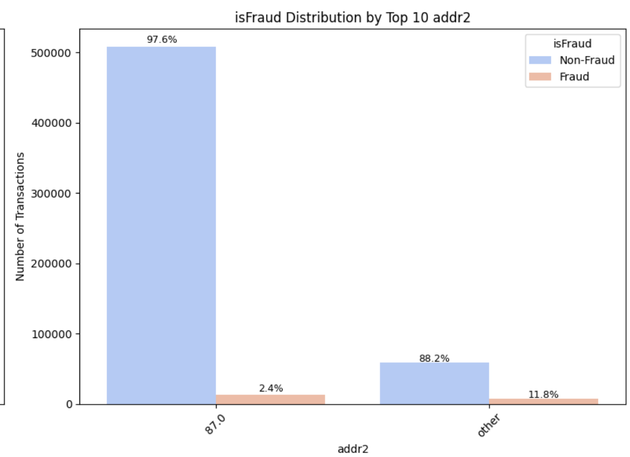
2.4. *Card1 - Card6*

Card 1		Card 2		Card 3		Card 5	
Unique ID	**is Fraud**	**Unique ID**	**is Fraud**	**Unique ID**	**is Fraud**	**Unique ID**	**is Fraud**
9633	719	321	48928	150	522765	226	300308
9500	528	111	45189	185	53383	224	80600
15885	436	555	41948	other	11293	166	57132
9026	388	490	38107			102	28950
15063	319	583	21798			117	25939
2616	314						
15066	313						
9917	305						
5812	297						
6019	294						

- ***Card1* (Card ID):** Specific Card1 IDs are repeatedly involved in multiple fraudulent transactions (e.g., *Card1 == 9633* has 719 fraud cases), highlighting compromised cards as a source of fraud.
- ***Card3* (Card Type/Code):** While *150.0* is dominant, the *other* category, despite its small count, appears to have a relatively higher proportion of fraudulent transactions.

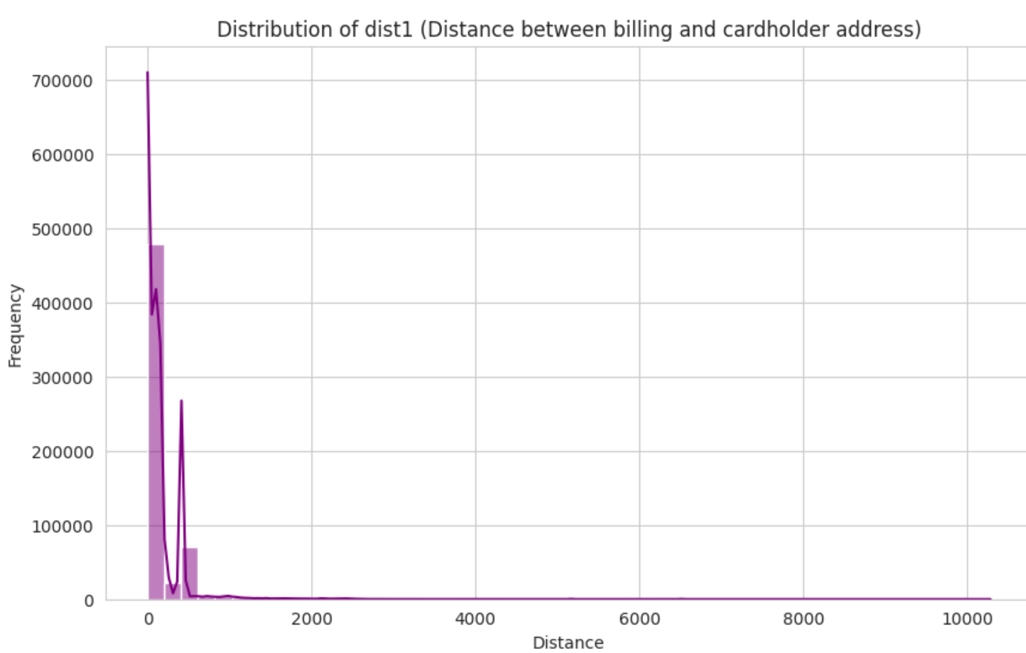
2.5. *Addr1*, *Addr2* (Region, Country)

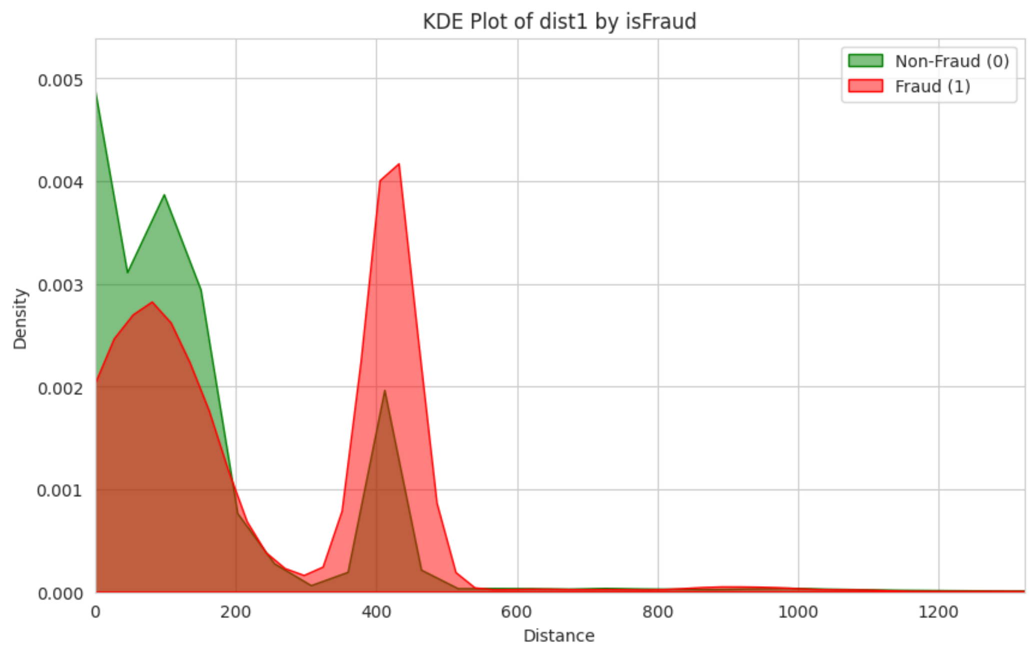




- The distribution of fraud largely mirrors the overall distribution for *Addr1* (billing region) and *Addr2* (country code). However, visually, the "others" categories for both *Addr1* and *Addr2* appear to have a slightly higher proportion of fraud, suggesting transactions from less common regions might be riskier.

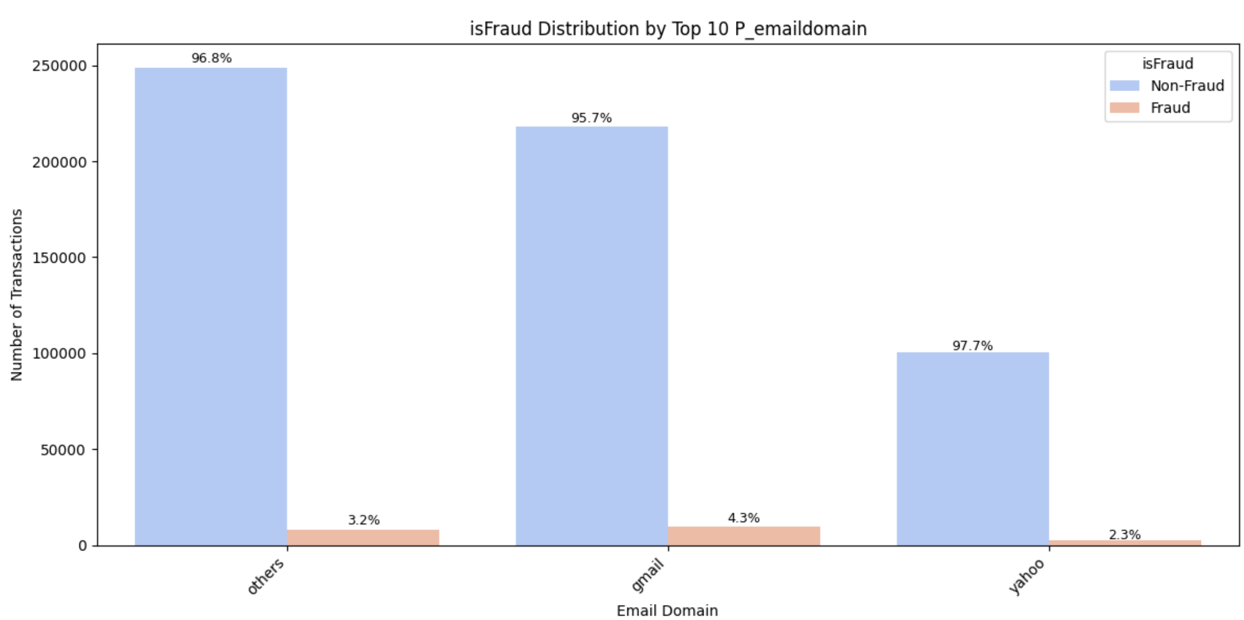
2.6. *Dist1* (Distance between billing and cardholder address)





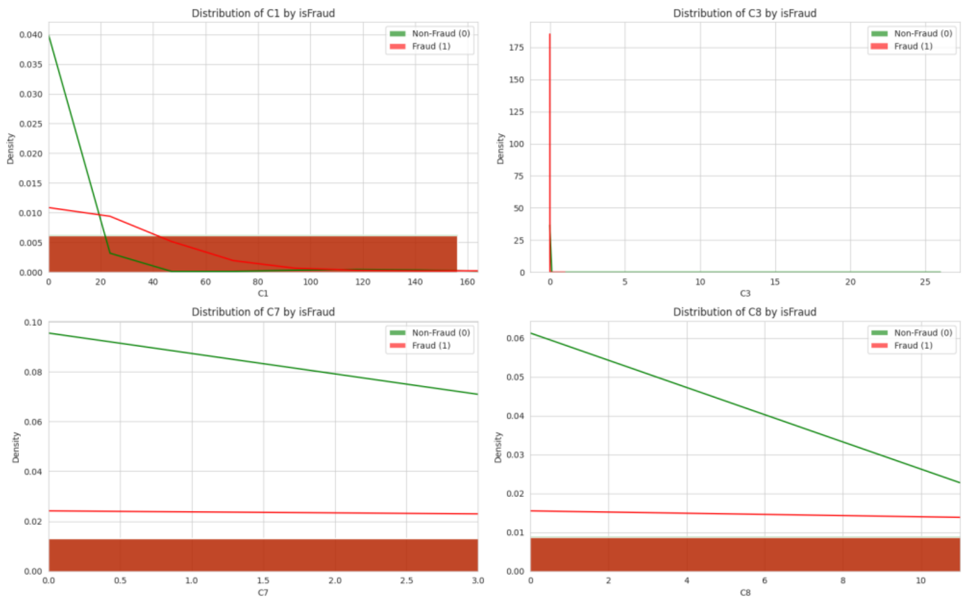
- *Dist1* is a **strong indicator of fraud**: Fraudulent transactions show a **much higher median *Dist1*** and a broader distribution compared to non-fraudulent ones. There's a **prominent peak for fraudulent transactions at a distance of approximately 400**, suggesting fraudsters use cards from a different location.

2.7. *P_emaildomain* (Purchaser email domain)



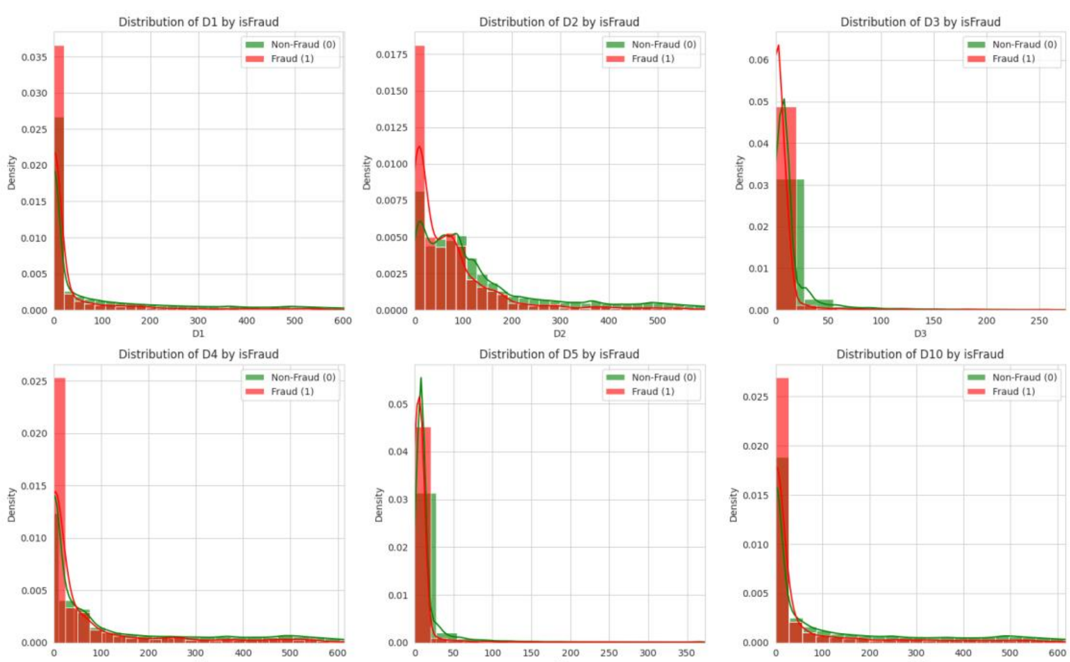
- Among the top domains, *gmail.com* has the highest fraud ratio at **4.31%,** followed by *others* (3.22%) and *yahoo.com* (2.25%). This suggests fraudsters may prefer common, easily accessible email services.

2.8. *C1 - C14* (Count Features)



- ***C3* (Failed Login/Payment Attempts), *C7* (Past Declined Transactions/Disputes), and *C8* (Account Logins/Authentications)** are highly promising fraud indicators. Fraudulent transactions consistently show **higher values** in these features, indicating more suspicious activity (e.g., more failed attempts, more logins).
- High multicollinearity exists among many C-features, implying they measure similar aspects of activity.

2.9. *D1 - D5*, *D10* (Time-based Delta Features)



- Fraudulent transactions tend to have **higher median values and a broader spread** for most *D*-columns (time deltas since previous activities). This suggests fraudulent transactions often occur with a larger time delta, potentially indicating the use of older or less frequently used accounts/cards.
- For fraudulent transactions, *yahoo.com* email domains are associated with a slightly longer average *D5* (time since last transaction using the same email/browser).

2.10. *V-columns* (Anonymized Features)

- As in preprocessing steps, these features have undergone PCA and are expected to contain complex predictive signals crucial for the One-Class Classification model.

3. Overall Implications for One-Class Classification

The EDA highlights that while linear relationships with fraud are weak, several features exhibit distinct patterns for fraudulent transactions. An OCC model can leverage these insights:

1. **Imbalance Handling**: OCC is inherently suited for the severe class imbalance, focusing on learning the distribution of the majority (non-fraudulent) class.
2. **Key Anomaly Drivers**: *ProductCD* ('C' and 'S'), *Dist1* (higher values, especially around 400), and *C3*, *C7*, *C8* (higher count values) are strong candidates for defining "anomalous" behavior.
3. **Complex Patterns**: The low linear correlations suggest that OCC models, particularly those capable of capturing non-linear boundaries (e.g., Isolation Forest, One-Class SVM), will be effective in identifying deviations from normal multi-dimensional patterns.
4. **Temporal & Behavioral Anomalies**: The volatility in fraud ratio and the distinct *D*-feature patterns for fraud suggest that temporal and behavioral anomalies (e.g., unusual time gaps between activities) are important.
5. **Multicollinearity and Outliers**: OCC models are robust to multicollinearity and are designed to detect outliers, which are prevalent in this dataset.

4. Next Steps for One-Class Classification

1. **Feature Engineering**:
    - Create interaction features (e.g., *ProductCD* and *Dist1*).
    - Derive cyclical time features (*TransactionDT* as day of week, hour of day) and time-windowed aggregates.
    - Aggregate features by *Card1*, *Addr1*, *P_emaildomain* (e.g., average *TransactionAmount*, fraud rate for that group).
    - Explore sums or ratios for *C* and *D* columns.

2. **Handling Categorical Features**: Apply One-Hot Encoding for *ProductCD*, *Card4*, *Card6*. For high-cardinality features like *P_emaildomain*, *Addr1*, *Addr2*, consider grouping less frequent categories into "other" or using target encoding.
3. **Scaling**: Numerical features should be scaled (e.g., *StandardScaler*) as many OCC algorithms are distance-based.
4. **Model Selection**: Evaluate algorithms like Isolation Forest, One-Class SVM, or Local Outlier Factor (LOF).
5. **Evaluation:** Focus on metrics like Precision, Recall, F1-score, and AUC-PR (Precision-Recall curve) for the minority class (fraud), as accuracy will be misleading.
6. **Thresholding:** Carefully tune the anomaly score threshold to balance false positives and false negatives based on business requirements.

This EDA provides a strong foundation for developing an effective One-Class Classification system for transaction fraud detection.