Preprocessing & EDA Plan

Processing Script: Mark in yellow
EDA: Mark in green
Column Explanation: Mark in blue

1. Merge IDs with Transaction tables based on TransactionID
2. Drop TransactionID and unnecessary columns: Recipient_emaildomain, DeviceInfo
3. Process Null based on column Clusters (total 358 columns)

- isFraud: 1: Fraudulent, 0: Non-fraud
    + Distribution: Fraud: ~30k/ 500k
    + EDA: correlate and try to find relationships with other columns (try pairplot with a subset of columns after dropping V columns). Also use value_counts to get a more detailed distribution.

- Id: 12 - 38: Binary Flag showing Verification done during Transaction
    + Idea: Id columns mean the Verification log of the device when making a transaction. If the id is null then there can be various reasons why
        - no verification system
        - cannot verify due to error, lack information
        - anonymous/ private filtering of device (ID masking)

      If the device is verified then we can see its Type and Info are displayed, so it's a good practice to keep the NaN value and replace it with 'missing' to use as a actual type

    + Process: Replace all Null with ‘Missing’, and create a new Column name ‘VerificationSet’ to count number of Finished Verification based on Each ID return T or ‘Found’ or ‘New’
    + EDA: Number of Verification may vary based on amount of Transaction (TransactionAmount), More verification may mark less Fraud, vice versa, less verification may mark higher chance of fraud

- DeviceType: correlates with Verification, if a device is verified then its information will be saved. Replace all Null to Missing and keep it as a label alongside with ‘mobile’ and ‘desktop’
    + EDA: Can different types of Device: mobile/ desktop lead to changes in verificationSet? Or Fraudulent activities?

- Id 01 - 11: Numeric identities of the purchaser/ transaction
    + Matching falsehood between Distribution and metadata
    + All is 76% Null
    + Process: Drop all to ensure integrity

- TransactionDT & TransactionAmount
    + TransactionDT: Transaction time delta (differences) from an anchor point (Can use as Index for Time Series Analysis)
    + TransactionAmount: Amount of money recorded in Transaction

- ProductCD: Product Category (5 different, no null)
    + EDA: Look at the distribution of products

```
ProductCD
W    439654
C     65437
R     37699
H     33023
S     11628
Name: count, dtype: int64
```

give comment and show

do different products may lead to higher TransactionAmount (more expensive), which may lead to Fraud

- Card1- card6 (numerical)
    + Card 1: card id
    + Card 2: bank id
    + Card 3: card type/ code
    + Card 4: card brand
    + Card 5: card issue number
    + Card 6: card type
      Note: not sure, may need proper check for each column

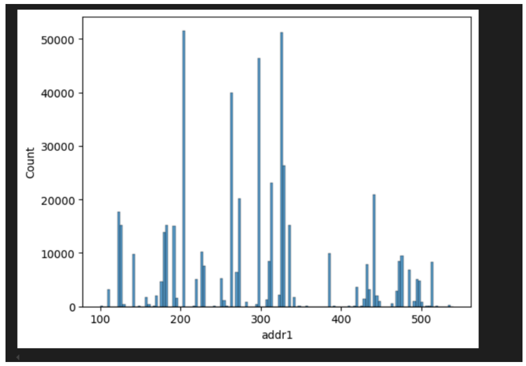
    + Process: Low proportion of Null: impute by Median
    + EDA: investigate more on Card 1 (card ID): is there a card ID flagged with isFraud? Rank cardID based on TransactionDT and TransactionAmount? What's the trend of activities that a CardID flagged with Fraud may do

- Addr1, addr2: Billing region, Country Code
    + May relate with other columns (distance, regions,...)
    + Process: both columns have more popular labels:

```
df['addr2'].value_counts()
```

```
addr2
87.0    520481
60.0      3084
96.0       638
32.0        91
65.0        82
         ...
49.0         1
14.0         1
25.0         1
22.0         1
93.0         1
Name: count, Length: 74, dtype: int64
```

        • For addr2:
          Keep the first 2, label others as 'others'

- For addr1: 

  Keep top 8 (>2000), label other as 'others'

- However, this will be conducted later

- Dist1, dist2: Distance
  - Dist1: Distance between billing and cardholder address
  - Dist2 (no null): Distance between billing and shipping address
  - Addr columns maybe useful to fillna for dist1
  - Process: Group by addr and fill with group mean
  - EDA: Does the increase in distance relate to any specific trend?

- P_emaildomain: Purchaser email domain
  - Process: Remove .com and only keep domain. Also Keep the first 2 popular: gmail and yahoo, label the rest as 'others'
  - EDA: is there any relationship between using gmail/ yahoo with other features? (Obviously not)

- C1 - C14: Count of different field
    + No Null so no need for process
    +

Column	Likely Meaning (Inferred)
C1	Count of transactions for this user/card in a short time window
C2	Count of successful online transactions
C3	Count of failed login/payment attempts
C4	Number of times a particular merchant or category has been used
C5	Count of transactions from the same email domain
C6	Frequency of transaction for a specific IP/device
C7	Count of past declined transactions or disputes
C8	Count of account logins or authentications
C9	Count of transactions from same billing address
C10	Count of previous transactions with similar amount/value
C11	Time-based count: e.g., transactions per hour/day/week
C12	Count of purchases in the same merchant category
C13	Count of recurring payments or subscriptions
C14	Number of transactions with the same card and amount (recurring, like utility bills)

    + EDA: Keep attention to C1, C3, C7, C8 if available

- D1 - D15: Time based delta
    + Columns with Null
      * D1 Days since last login (or first observed activity)
      * D2 Days since card was issued or account was created
      * D3 Days since last known address update
      * D4 Days since last transaction with same card/account
      * D5 Days since last transaction using the same email or browser
      * D10 Days since first transaction on current session
      * D11 Days since device was first associated with the account
      * D15 Days since last online interaction with the current device

    + Process:
        - D1 low NaN: median
        - D2,4,10 relies on Card1: Fill with group mean
        - D3 relies on addresses
        - D5 relies on email domain
        - D11, 15, dropped due to value mismatch/ high null

    + EDA: Check if these columns pair have any relationship (d3 - addr1 2, d5 - email domain, d2 4 10 - card1)

- M1 - M9: Binary match class
    + Show different matches of a Transaction: matching amount, matching time, id,.....
    + Most are binary showing T/F for matching/ not, only M4 has 3 labels M0, M1, M2 but relabel into M0 - F, M1, M2 - T
    + Process: Fill all Na with F and create a new Column 'MatchingCount' to count the number of T Matches
    + EDA: More matches may mark safer transactions, is there any relationship between Matches and VerificationSet?

- V1 - V339: Engineered anonymized features. Often contain predictive signals.

    + Just numerical-encoded data for Model evaluation, no analysis meaning.
    + Process: 3 ranges
        - <20% Null: Median impute (152 cols)
        - 20 - 70%: KNN (29 cols)
        - > 70% Null: Drop (120 cols)
        - Use PCA to reduce dimension to 14 dimensions
    + EDA: Screeplot, available already no need to remake