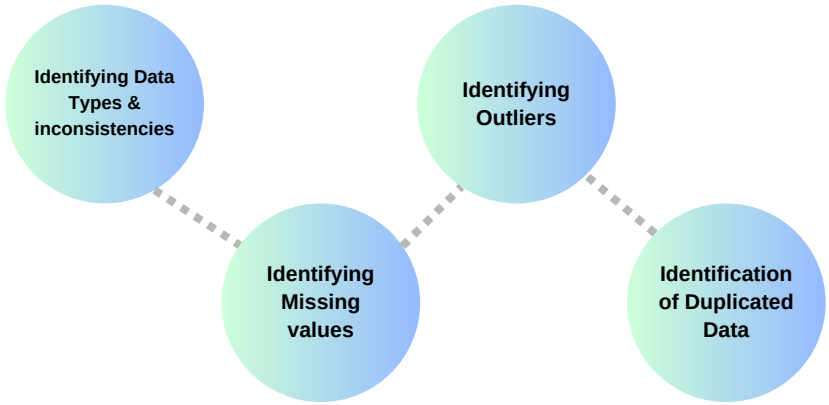
Final Project Data Science Sanbercode Batch 69 :
Transaction Data Analysis for Fraud Detection on "PayNow" Using the CRISP-DM Approach

1. BUSINESS UNDERSTANDING

- **Objective:** To identify the characteristics and patterns that distinguish fraudulent transactions from legitimate transactions in order to reduce financial losses and improve the security of the "PayNow" platform.
- **Problem:** There is an increase in fraud transactions whose pattern has not been clearly identified.
- **Question formulation:**
  a. What are the common characteristics of fraudulent transactions?
  b. Is there a specific time pattem?
  c. What factors (country of origin, card type, IP address, device, browser, etc.) are most commonly associated with fraudulent transactions?
  d. How does the transaction value compare between fraud and non-fraud?
  e. How does the value distribution of fraud transactions compare to normal transactions?

3. DATA PREPARATION/ DATA CLEANING



2. DATA UNDERSTANDING

The available data is in the form of transaction data of Fintech company "PayNow", whether detected as fraud or not. The dataset consists of 331 rows of 9 columns, with the identification of the column data as follows.

	transaction_id	user_id	transaction_time	transaction_amount	payment_method	device_id	ip_address	billing_country	is_fraud
0	TXN0001	USER043	2024-01-01 00:05:32	150.75	credit card	DEV076	192.168.1.1	USA	0.0
1	TXN0002	USER012	2024-01-01 01:15:45	25.50	paypal	DEV033	10.0.0.5	US	0.0
2	TXN0003	USER089	2024-01-01 02:30:10	$5000.00	gift card	DEV101	172.16.0.8	RU	1.0
3	TXN0004	USER043	2024-01-01 03:45:22	75.20	Credit Card	DEV076	192.168.1.1	USA	0.0
4	TXN0005	USER012	2024-01-01 04:55:30	120.00	debit card	NaN	10.0.0.5	US	0.0
5	TXN0006	USER077	2024-01-01 05:10:05	1.25	credit_card	DEV092	203.0.113.1	USA	1.0

Label	Keterangan	Tipe data	Non-null
transaction_id	ID unik transaksi	Object	331
user_id	ID pengguna	Object	331
transaction_time	waktu transaksi	Datetime	331
transaction_amount	nominal transaksi	Float	330
payment_method	metode pembayaran transaksi	Category	331
device_id	perangkat yg digunakan untuk transaksi	Object	329
ip_adress	alamat ip perangkat	Object	328
billing_country	negara asal tagihan	Category	330
is_fraud	transaksi legitimate (0) dan fraud (1)	Float	330

	is_fraud
count	330.000000
mean	0.230303
std	0.421666
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

- There are missing values on some labels
- Payment_method value is inconsistent (e.g. "Credit Card", "credit card", "c. card").
- Value billing_country has a variation of the spelling ("USA", "usa", "United States").
- There are potential outliers in transaction_amount

Based on the results of descriptive statistics on the is_fraud label:

- There are 330 transactions that have a fraud or non-fraud label.
- There are 23% of transactions detected as fraud, while 77% are genuine transactions.
- The STD value = 0.4217 indicates the degree of variation/spread between Fraud (1) and Normal (0) transactions. In this dataset, std = 0.42, meaning that the distribution between fraud vs normal is quite diverse, but not perfectly balanced (because fraud is only 23%).
- is_fraud label data is binary.

3. DATA PREPARATION/ DATA CLEANING

- **Identifying Data Type & Inconsistencies**

Problem: some labels have inconsistent data, such as date writing, payment method writing, country code writing, and transaction amount. In addition, some labels also do not have the appropriate data type.

Solution: inconsistent data is handled, date format is made YYYY-MM-DD, writing payment methods such as c. card to Credit Card, writing country codes such as USA to US, transaction nominal is limited to >=0 and is consistent only with numbers. The data type that was previously 8 objects and 1 float was changed as above.

- **Identifying Missing values**

Problem: there are missing values on some labels, such as transaction_time, transaction_amount, device_id, ip_address, billing_country, and6u is_fraud.

Solution: NaT is filled in the surrounding date, NaN transaction_amount in the drop, NaN in device_id and ip_address adjusted to user_id, billing_country filled in data afterwards, and NaN in is_fraud filled median.

- **Identifying Outliers**

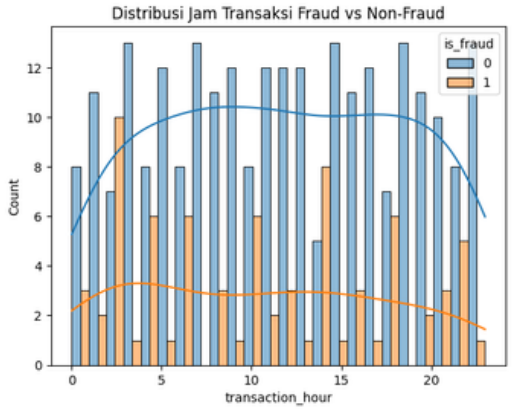
Problem: by using IQR, there are 16 outliers in the transaction_amount column, which are very extreme and affect visualization

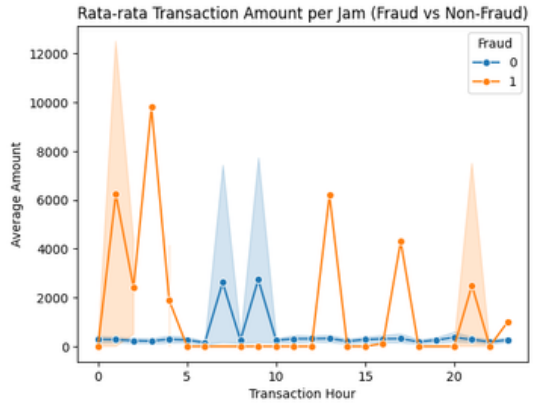
Solution: outliers are retained but restricted when performing analysis and visualization.

- **Identification of Duplicated Data**

Problem: there are 3 duplicated data detected with .duplicated()

Solution: duplicated data dropped from df





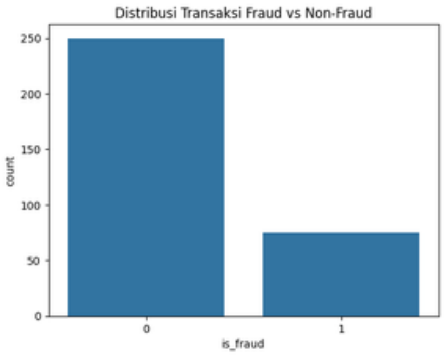
In numbers, **non-fraud transactions happen more often** almost every hour (about 8–13 transactions per hour on average), while fraud is relatively rare (around 2–3 transactions per hour), but it's spread out pretty evenly throughout the day.

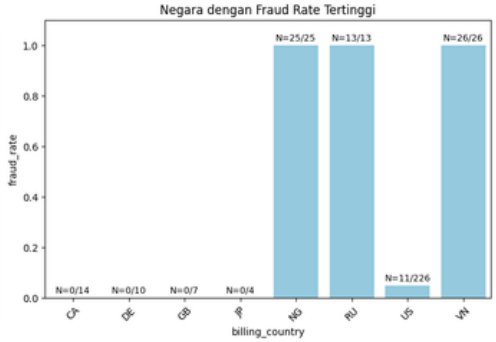
Fraud shows much more extreme fluctuations compared to non-fraud. Fraud tends to occur with much larger and more varied amounts, compared to normal transactions which are stable and consistent. This could be a strong indicator that high-value transactions with fluctuating patterns should be suspected.

Transaction time is not the main indicator, but the transaction amount is very important to distinguish between fraud and non-fraud.

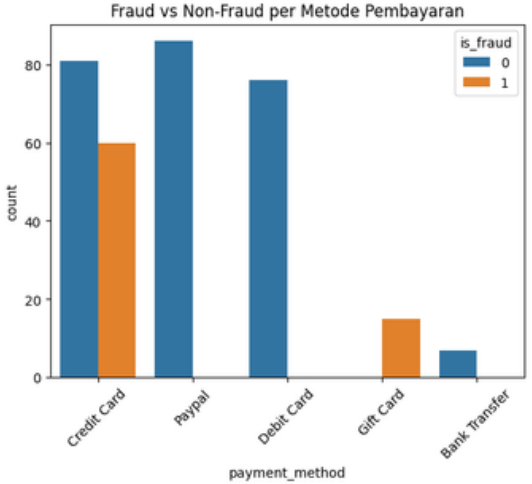
4. EXPLORATORY DATA ANALYSIS

A *countplot* visualization was performed to observe the distribution of non-fraud and fraud transactions. Based on the *countplot* shown, **the number of normal transactions is much higher** compared to fraud transactions. The proportion of fraud is approximately 20–25% of the total transactions.



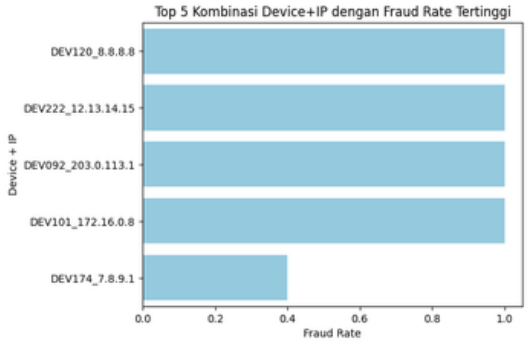


- Countries with a 100% Fraud Rate: NG, RU, VN, meaning that every transaction from these countries always results in fraud in this dataset.
- Countries with Low Fraud Rate but High Volume: US, very low fraud rate, but the total number of transactions is very large.



- **Credit Cards** dominate the overall number of transactions (both fraud and non-fraud). However, it is seen that almost 3/4 of fraud transactions occur here. This means that **credit cards are the most vulnerable method to fraud.**
- **Paypal & Debit Cards**, all transactions **are safe (non-fraud).** No fraud has been found through these methods.
- **Gift Card,** have fewer total transactions, but the proportion of fraud is much higher compared to non-fraud, as in the available dataset Gift Cards are 100% fraud. This indicates that gift cards are **often used as a method of fraud**, possibly due to their more anonymous nature.
- **Bank Transfer,** have a small number of transactions, and there is no history of fraud in the available dataset. Bank transfers are less frequently used in fraud, possibly because the process is more rigid (harder to fake compared to fast digital methods).

device_ip	total_tx	fraud_count	fraud_rate
DEV120_8.8.8.8	21	21	1.0
DEV222_12.13.14.15	26	26	1.0
DEV092_203.0.113.1	7	7	1.0
DEV101_172.16.0.8	8	8	1.0
DEV174_7.8.9.1	5	2	0.4



Four combinations of Device and IP show a 100% fraud rate, meaning certain Devices & IPs are used as a fraud method by perpetrators.

5. SYNTHESIS OF FINDINGS & RECOMMENDATION

Key Characteristics of Fraudulent Transactions

- The nominal value of fraudulent transactions is much higher and more variable compared to stable normal transactions.
- Fraud is not highly dependent on specific hours, but its distribution is more fluctuating than that of normal transactions.
- Certain countries have a 100% fraud rate (for example, NG, RU, VN), so they need to be monitored.
- Certain payment methods are prone to fraud: Credit Cards account for ~75% of fraud cases, while Gift Cards have a 100% fraud rate.
- Specific combinations of Device + IP show a 100% fraud rate, indicating consistent patterns of device/IP usage by perpetrators.

Initial Recommendations

- Mark transactions with amounts above a certain threshold for manual review.
- All transactions from countries with fraud rates approaching 100% (NG, RU, VN) should ideally be restricted or given additional verification.
- Be careful with payment methods: Gift Card transactions should be investigated and Credit Card transactions closely monitored.
- Blacklist device_id + ip_address combinations that have a high fraud rate because they exhibit very consistent patterns.

Conclusion

- **Fraud Characteristics:** transaction amounts tend to be larger and more varied, often occurring via Credit Card and Gift Card, and are repeatedly detected on certain device+IP combinations.
- **Time Pattern:** fraud occurs throughout the day with more extreme fluctuations compared to normal transactions, so time is not a primary indicator.
- **Important Factors:** the country of origin has a significant impact; for example, all transactions from NG, RU, and VN are fraudulent.
- **Value Comparison:** fraud has a much wider nominal distribution with many outliers, whereas normal transactions remain stable.