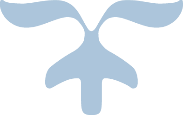


###### PAI: SP-Buy

### PAI CA1



By Group 1, Class: DAAA/FT/2B/03

Shaun Kwo Rui Yu – 2317933

Loh Yip Khai – 2317454

Koh Zi Hao – 2317483

Contents

###### Introduction [1](#_bookmark0)

Importing the Necessary Libraries and Datasets [1](#_bookmark1)

Ensuring Data Consistency and Error Prevention [1](#_bookmark2)

Big Data Resource Management [1](#_bookmark3)

Data Inspection [1](#_bookmark4)

Data Cleaning [1](#_bookmark5)

Data Preprocessing 2

[Database Creation](#_TOC_250003) 2

[SQL Queries](#_TOC_250002) 4

ETL Pipeline 7

Exploratory Data Analysis [8](#_bookmark8)

[Interactive Tableau Dashboard](#_TOC_250001) 12

[Conclusion](#_TOC_250001) 12

[Appendix](#_TOC_250000) 13

**Introduction**

###### The task is to develop a system to identify fraud activities on the SP-Buy platform, starting with the analysis of historical data and building an interactive dashboard for the business users.

# **Importing Libraries and Datasets**

###### The necessary libraries, **Pandas** and **NumPy**, were imported to facilitate data handling and analysis.

###### Three datasets—**Customer Features**, **Order Features**, and **Fraud Labels**—were imported into separate DataFrames named Customer, Order, and Labels, respectively.

# **Ensuring Data Consistency and Error Prevention**

# This ensures that customer\_id values are clean and free from unnecessary whitespace. It is particularly useful for:

# **Data consistency:** Avoids issues caused by unexpected spaces or paragraphing (\n) from the original datasets.

# **Error prevention:** Helps prevent bugs in downstream operations (e.g., filtering, merging datasets, or using IDs as keys).

# **Big Data Resource Management**

###### **Convert CSV to HDF5 files**

###### Advantage: Faster data loading and processing

###### Disadvantage: Large storage space is needed as HDF5 takes up more data

###### **Optimize data types for storage space and processing**

###### INT64 is converted to INT16 and FLOAT16 to FLOAT64

###### Advantage: Lesser memory used and faster loading and processing

###### Disadvantage: Data is less accurate if data is cut off.

# **Data Inspection**

###### **Missing Values Check**

###### In the **Customer** DataFrame, 1,240 entries have null values in first\_order\_datetime.

###### In the **Orders** DataFrame, 3 entries have null values in num\_items\_ordered.

###### The **Labels** DataFrame contains no null values.

###### **Duplicate Values Check**

###### The **Customer** DataFrame contains 38,620 duplicate customer\_id entries.

###### The **Orders** DataFrame includes 1,180 duplicate order\_id entries.

###### The **Labels** DataFrame has 441,705 duplicate customer\_id entries and 138,610 duplicate order\_id entries.

###### **Data Cleaning**

**Customer Dataset**

* All fully duplicated rows were removed to ensure data integrity.

**Order Dataset**

* Rows with missing values in num\_items\_ordered were dropped, affecting only three rows, to maintain consistency in order data.
* All fully duplicated rows were eliminated.

**Labels Dataset**

* Fully duplicated rows were removed to ensure the dataset’s accuracy.

**Data Preprocessing**

Ensuring that all Primary Keys exist with no duplicates

**Customer Dataset**

* For records with the same customer\_id, only the row with the highest num\_orders\_last\_50days was retained, as this row represents the most recent customer activity.
* We validated the uniqueness of each customer\_id, confirming it as the primary key.

**Order Dataset**

* A check was conducted to verify any duplicate combinations of country\_code and order\_id.

**Labels Dataset**

* We reviewed the data for duplicate combinations of customer\_id and order\_id as well as country\_code and order\_id.
* An integrity check identified eight cases where order\_id entries lacked a corresponding customer\_id; these rows were dropped to maintain data consistency.
* In cases of duplicate rows with identical country\_code, order\_id, and customer\_id but different is\_fraud values, we retained the row with the largest index, as it represents the most recent and relevant fraud assessment

**Database Creation:**

**Creating the Database**

* Created the database using Microsoft SQL Server

|  |
| --- |
| `customer\_features` table |
|  |
| `order\_features` table |
|  |
| `fraud\_labels` table |
|  |
| Created the ERD Diagram and specified the relationships |
|  |

**SQL Queries:**

**Query 1: List the countries with the highest fraud incidents, along with the average and total order values for fraud and non-fraud cases.**

This query identifies countries with the highest fraud incidents and analyzes fraud patterns by summarizing key metrics from the order and fraud datasets. It:

* **Counts Fraud Cases:** Total number of fraudulent orders (fraud\_count) per country.
* **Fraud Metrics:** Calculates the average (avg\_fraud\_order\_value) and total (total\_fraud\_order\_value) order values for fraudulent cases.
* **Non-Fraud Metrics:** Computes the average (avg\_non\_fraud\_order\_value) order value for non-fraudulent cases.
* **Groups and Ranks:** Groups data by country and ranks them by fraud\_count in descending order.

**Purpose:**  
To identify regions with high fraud prevalence, providing insights to develop targeted fraud prevention strategies, optimize detection systems, and reduce revenue losses.

**Query 2: Determine the most common collection types and payment methods in fraud cases.**

This query analyzes patterns in collection types and payment methods associated with fraud cases. It:

* **Counts Fraud Cases:** Computes the total number of fraudulent transactions (fraud\_count) for each combination of collect\_type (e.g., delivery or pickup) and payment\_method.
* **Calculates Order Metrics:** Determines the average order value (avg\_order\_value) for fraudulent transactions.
* **Filters Fraudulent Transactions:** Focuses exclusively on fraud cases (is\_fraud = 1).
* **Groups and Ranks:** Groups data by collect\_type and payment\_method and ranks combinations by fraud\_count in descending order.

**Purpose:**  
To identify high-risk collection and payment method combinations, enabling the development of targeted fraud prevention strategies. Insights from this query support efforts to reduce fraud and ensure secure transaction practices.

**Query 3: This query identifies top 10 customers with the highest cancel\_rate, refund\_rate, and fraud\_order\_count in their recent order history, potentially signaling fraudulent behavior**

### This query identifies the top 10 customers with suspicious behaviors indicative of potential fraud by analyzing their recent order history. It:

### **Calculating Key Metrics using Subquery:**

### Computes the **cancellation rate (cancel\_rate)** as the ratio of canceled orders to total orders in the last 50 days.

### Computes the **refund rate (refund\_rate)** as the ratio of refunded orders to total orders in the last 50 days.

### Counts the number of fraudulent orders (fraud\_order\_count) for each customer.

### **Filters High-Cancellation or Refund Cases:** Includes customers with more than 5 canceled or refunded orders and focuses on fraud-related cases (is\_fraud = 1).

### **Groups and Aggregates:** Groups by customer ID to calculate relevant metrics.

### **Prioritizes Suspicious Activity:** Filters customers with cancellation or refund rates exceeding 30% and sorts by cancel\_rate, refund\_rate, and fraud\_order\_count in descending order.

### **Limits to Top 10:** Selects the top 10 customers with the most suspicious patterns.

### **Purpose:** To flag high-risk customers exhibiting suspicious patterns like frequent cancellations, refunds, or a history of fraudulent orders. These insights help prioritize fraud investigations and inform prevention strategies.

**Query 4: cluster customers by their order and refund behavior in the last 50 days, particularly focusing on customers with associations to others.**

This query aims to classify customers based on their recent order and refund behavior over the last 50 days, particularly focusing on customers who have associations with others. It identifies high-activity and high-refund customers while also incorporating fraud-related insights.

1. Columns Selected:

* **customer\_id**: Unique identifier for each customer.
* **num\_associated\_customers**: Number of other customers linked to this customer (e.g., shared accounts or activities).
* **num\_orders\_last\_50days**: Number of orders placed by the customer in the last 50 days.
* **num\_refund\_orders\_last\_50days**: Number of orders refunded in the last 50 days.
* **activity\_level**: Categorizes customers into:
* **High Activity**: More than 150 orders in the last 50 days.
* **Moderate Activity**: 70 to 150 orders in the last 50 days.
* **Low Activity**: Fewer than 70 orders in the last 50 days.
* **refund\_pattern**: Categorizes customers into:
* **High Refund**: 5 or more refunded orders in the last 50 days.
* **Low Refund**: Fewer than 5 refunded orders.
* **fraud\_count**: Total count of fraudulent orders associated with the customer.

1. Fraud Data Integration:

* Joins the fraud\_labels table to count fraudulent orders for each customer (fraud\_count).

1. Grouping:

* Groups data by customer to calculate aggregate metrics like fraud count and apply categorizations.

1. Filters (HAVING Clause):

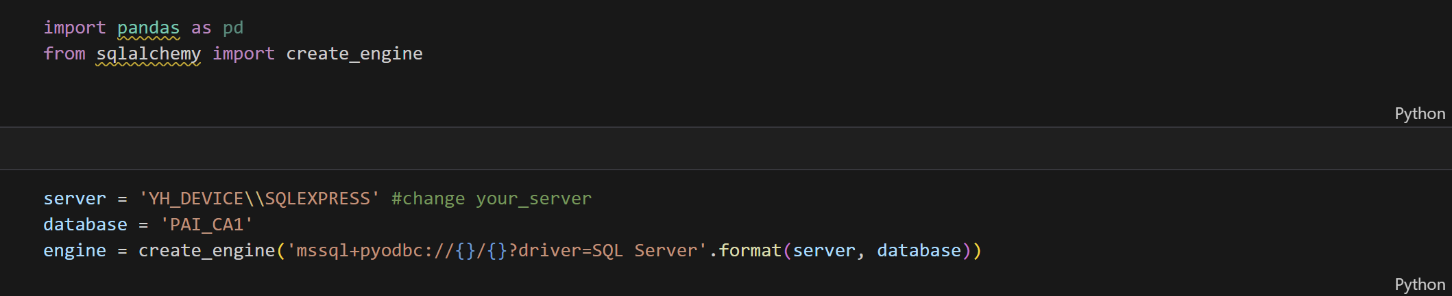
* Includes only customers with:
  + **More than 50 orders** in the last 50 days.
  + **5 or more refunded orders**, indicating a pattern of frequent refunds.

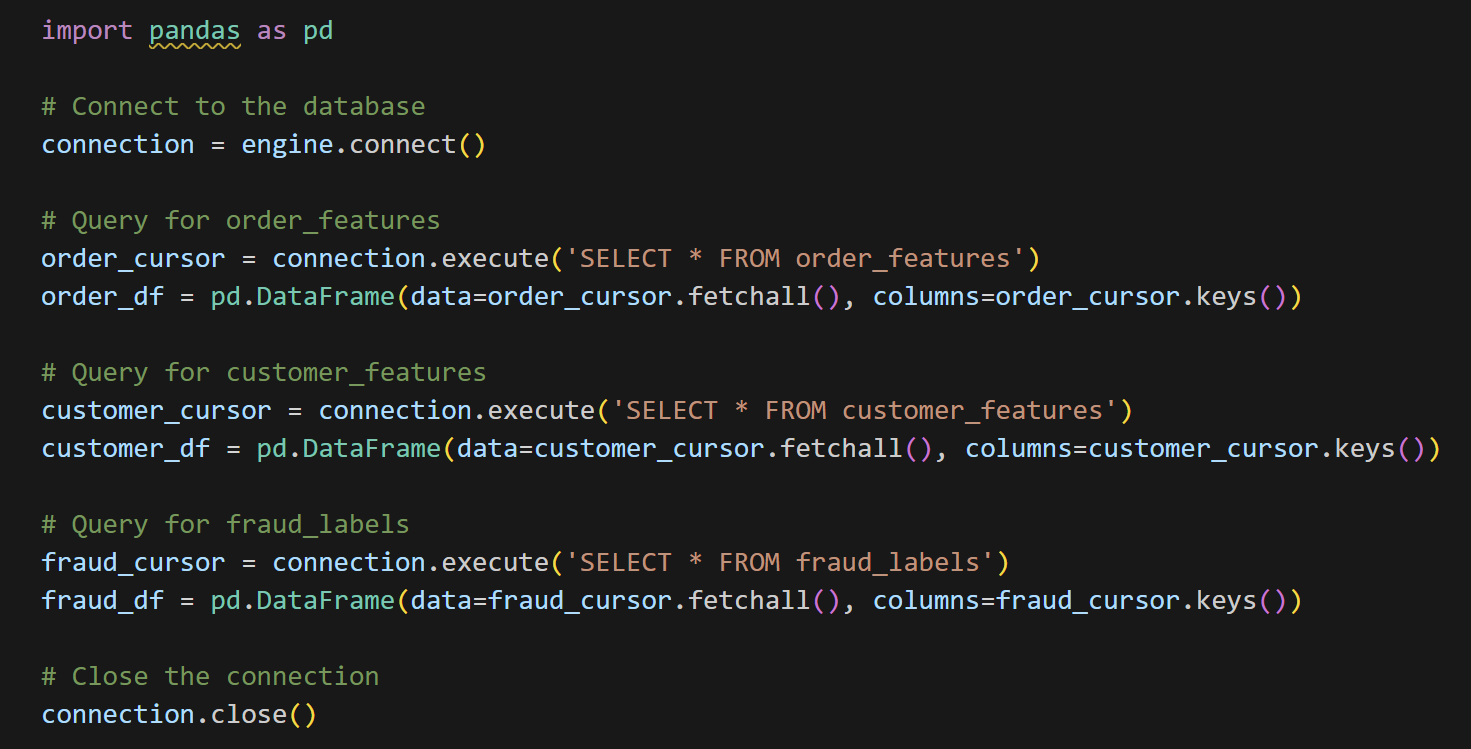
### **Purpose:**

This query identifies clusters of customers based on their behavior:

* Highlights customers with frequent orders and refunds.
* Allows for targeted analysis of high-risk customers (e.g., those with high refund rates or fraud associations).
* Provides actionable insights for monitoring or segmenting customer groups for fraud prevention or further investigation.

**ETL Pipeline**





**Extraction (E):**

* The script connects to the PAI\_CA1 database on a local SQL Server instance (YH\_DEVICE\SQLEXPRESS) using the sqlalchemy library.
* It fetches data from three tables: order\_features, customer\_features, and fraud\_labels.
* Raw query results are obtained with .fetchall().

**Transformation (T):**

* Data is structured into pandas DataFrames (order\_df, customer\_df, fraud\_df) using column names obtained from the database cursor.
* (Optional) Further transformations, like filtering or merging, can be added here if needed.

**Loading (L):**

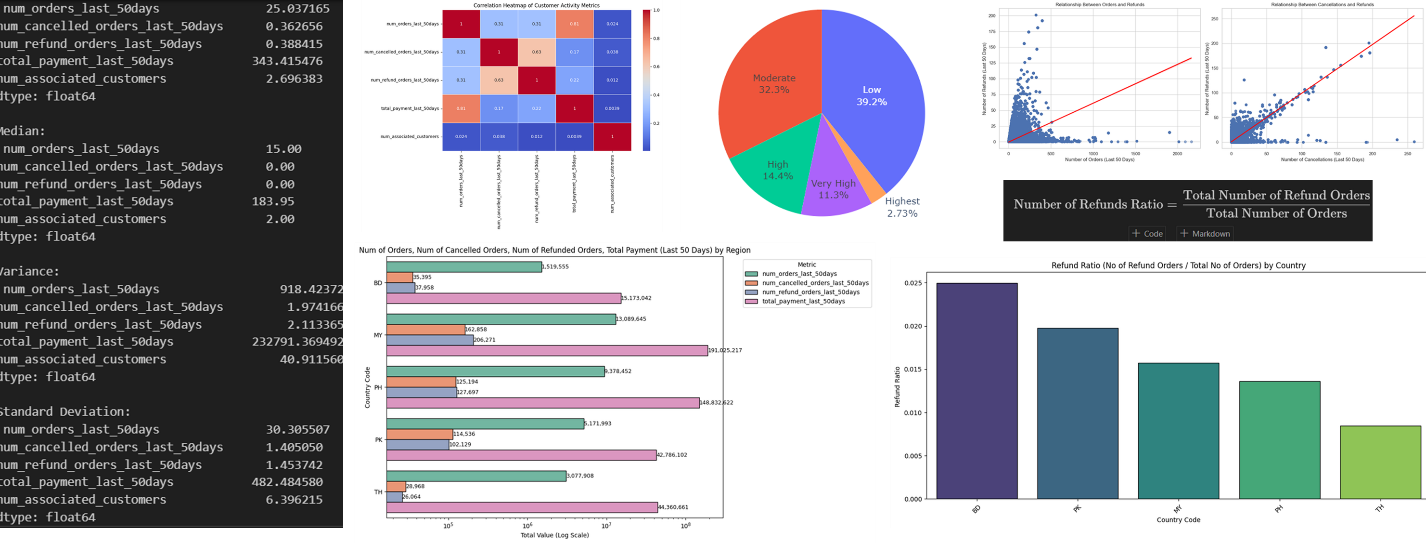
* The DataFrames are printed to the console, making them ready for analysis in the Python environment.
* They could also be saved to files (e.g., CSV or Excel) or exported to other systems for further use.

**Additional Notes:**

* The database connection is closed explicitly using connection.close() to free up resources.
* The pipeline can scale for larger data volumes with enhancements like batch processing and robust error handling.

**Exploratory Data Analysis (EDA)**

For `customers` dataset:



**Comparative Bar Chart:**

**Key Insights**

**1.High Total Payments in Malaysia (MY) and Philippines (PH):**

- Malaysia and the Philippines have significantly higher total payments compared to other regions, with Malaysia leading at over 190 million SGD. This could be due to either a higher number of active users or potentially larger transaction sizes.

- High total payments in these countries may also indicate higher transaction volumes or greater purchasing power, which could be associated with higher fraud activity as fraudsters may target high-value regions.

**2. High Numbers of Refunds in Malaysia (MY) and Philippines (PH):**

- Malaysia and the Philippines also have a higher number of refund orders. In Malaysia, refunds are the highest, even exceeding total payments in most other regions. This could be a red flag, as a high refund rate often correlates with fraudulent behavior, where users exploit refund policies.

- If Malaysia and the Philippines have a disproportionately high number of refunds relative to total payments or the number of orders, it may indicate refund fraud.

**3. Moderate Levels of Cancellations:**

- Cancellations do not show the same high volumes as refunds, but Malaysia and the Philippines again have relatively high numbers. While cancellations are not as directly associated with fraud, they could indicate user dissatisfaction or potentially suspicious activity if they coincide with other fraud indicators (like frequent refunds).

**4. Thailand (TH) Shows Lower Activity Overall:**

- Thailand has the lowest number of orders, cancellations, and refunds, and the lowest total payments. This may indicate a smaller user base in this region or less engagement with the platform. While lower activity doesn’t immediately suggest fraud, it might be useful to monitor if certain low-activity countries have unusual spikes in fraudulent orders in the future.

**5. Potential Fraud Risk in High-Refund Regions:**

- A general takeaway is that regions with high total payments and high refund counts (like Malaysia and the Philippines) could represent a fraud risk. Monitoring these regions more closely and applying stricter verification methods (like requiring mobile verification for new users) could help mitigate potential fraud.

**6. Bangladesh (BD) Shows High Activity with Fewer Refunds:**

- Although Bangladesh has a high number of orders and significant total payment, it has a lower number of refunds compared to Malaysia and the Philippines. This could indicate genuine transactions, suggesting that high order volume alone is not always indicative of fraud risk.

**Fraud Ratio Bar Graph:**  
**Key Insights**

**1. Highest Refund Ratio in Bangladesh (BD):**

- Bangladesh has the highest refund ratio, meaning a larger portion of orders result in refunds compared to other regions. This could indicate a higher tendency for customers in Bangladesh to request refunds, which might suggest potential misuse of the refund policy.

- Given that Bangladesh also had significant order and payment volumes (from the previous chart), the high refund ratio is a red flag for possible fraud. Investigating specific customers in Bangladesh who frequently request refunds could be beneficial.

**2. Pakistan (PK) and Malaysia (MY) Show Moderate Refund Ratios:**

- Pakistan and Malaysia have the next highest refund ratios. Malaysia, with its high total payments and order volume, has a noticeable rate of refunds, which aligns with earlier observations and further supports the need for closer monitoring in this region.

- Pakistan’s moderate refund ratio could indicate some risk, though it may not be as pressing as in Bangladesh. However, additional monitoring of refund patterns in Pakistan could still be useful.

**3. Lower Refund Ratios in the Philippines (PH) and Thailand (TH):**

- The Philippines and Thailand have lower refund ratios compared to the other countries, indicating that fewer orders result in refunds. This suggests a relatively lower refund-related risk for these regions.

- Thailand, in particular, shows the lowest refund ratio. This, combined with its low order volume, suggests minimal risk of refund fraud in this region at present.

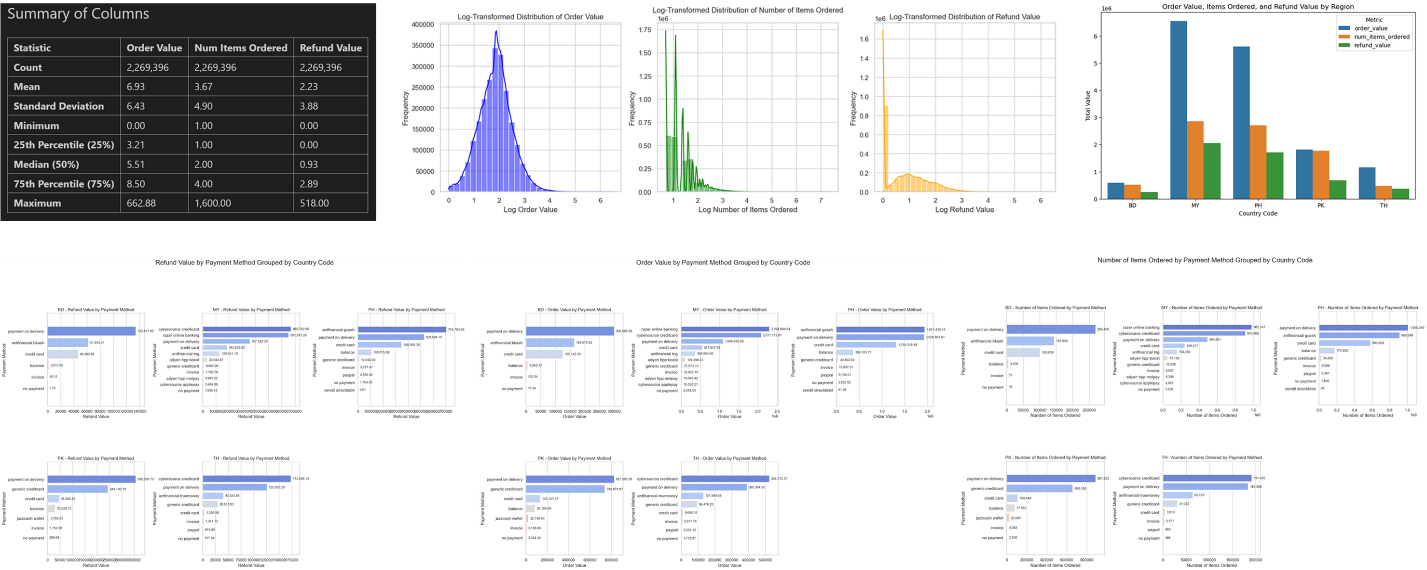
**4. Potential Focus on High-Refund Regions for Fraud Prevention:**

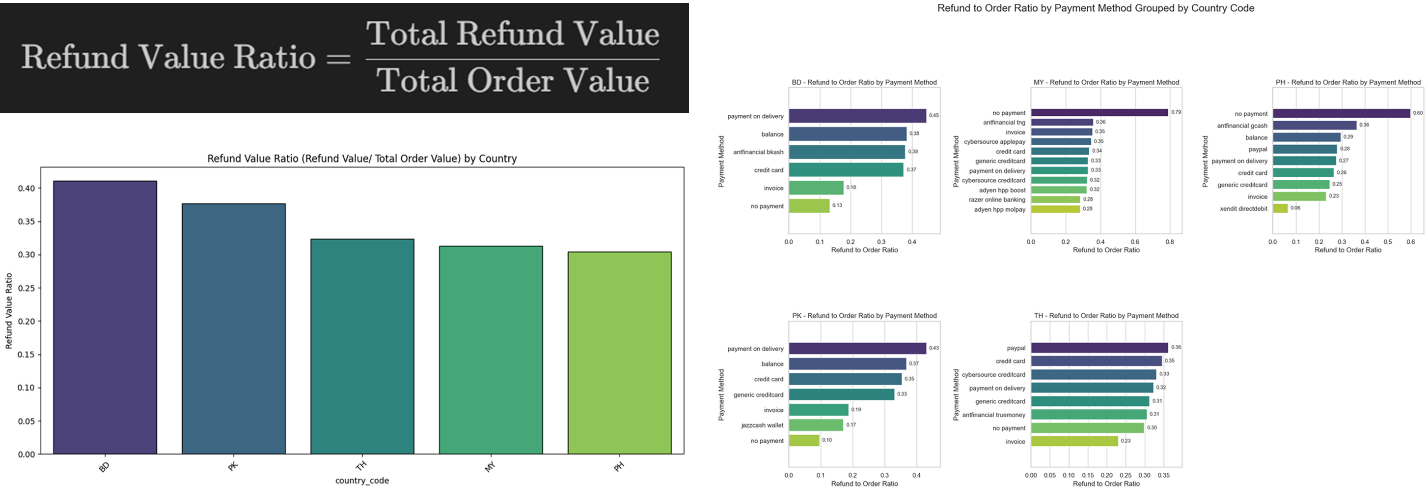
- Countries with high refund ratios, particularly Bangladesh, should be prioritized for fraud prevention measures. Implementing stricter policies for refunds, such as requiring additional verification steps or limiting the number of refunds allowed within a specific period, could help reduce fraud risk.

- These insights could also help in creating a region-specific fraud detection model that weighs refund activity more heavily in high-risk regions.

Overall, this chart suggests that Bangladesh and Malaysia have relatively high refund ratios, which could indicate potential refund fraud. Monitoring and investigating these regions more closely could help SP-buy better protect against fraud.

For `orders` dataset:





For `labels` dataset:

A close-up of a pie chart

Description automatically generated

**1. Overall Fraud vs. Non-Fraud Transactions (Pie Chart)**

* **Fraud Proportion**:
  + **11% of transactions** are **flagged as fraud**, while **89% are non-fraud**.
  + **Fraud cases** are a **small but potentially significant portion**, depending on industry standards.
* **General Fraud Trend**:
  + **Most transactions (89%)** are **legitimate**, suggesting effective fraud prevention, but there is **room for improvement**.

**2. Fraud by Country (Sunburst Chart)**

* **Country-Specific Insights**:
  + **PH (Philippines)** shows noticeable fraud activity.
  + **PK (Pakistan), MY (Malaysia), TH (Thailand), and BD (Bangladesh)** show varying fraud levels, with some countries having minimal fraud.
* **Regional Patterns**:
  + High-fraud countries need stronger anti-fraud measures, while low-fraud regions can maintain current efforts.

A graph of a bar chart

Description automatically generated with medium confidence

**Fraud Ratio (Flagged Fraud Orders / Total Orders) *(Bar Chart)***

* **Key Observations**:
  + **BD (Bangladesh)** has the highest fraud ratio (~30%), followed by **PK (Pakistan)** (~25%).
  + **PH (Philippines)** shows a moderate fraud ratio (~12%), while **TH (Thailand)** and **MY (Malaysia)** have lower ratios (~7% and ~5%).
* **Regional Patterns**:
  + **Bangladesh and Pakistan** require **stronger anti-fraud measures** due to higher fraud rates.
  + **Malaysia and Thailand** can focus on **monitoring efforts** to maintain low fraud levels.

**Interactive Tableau Dashboard**

A screenshot of a computer

Description automatically generated

**Conclusion**

The project effectively demonstrates the application of data science skills, focusing on fraud detection and analysis through SQL, Python, and interactive dashboard tools. Key takeaways include:

* **Fraud Insights:** Identification of fraudulent patterns and leveraging data-driven decisions for business strategies.
* **Technical Expertise:** Strong capabilities in SQL, ETL workflows, data preprocessing, and visualization.
* **Lessons Learned:** Emphasis on data quality, collaboration, iterative refinement, and mastering the end-to-end data science process.
* **Project Management:** Use of Agile methodologies, teamwork, and effective time management for successful project execution.

This highlights the integration of technical and management skills for impactful fraud analysis and business insights.

## Appendix

-

~ The End ~