IV. DATA DESCRIPTION/ PREPARATION

1. Data Description

The dataset used in this project is from Dr. Marie Anderson, Machine Learning Engineer, Chief Data and Analytics Office, Lloyds Banking Group, Bristol. The raw transaction data provides a record of customer interactions, including date, time, amount, mode, and associated location. For reasons of confidentiality, the raw transaction data is subjected to agent-based simulations to generate artificial transactional data at the individual level. Our research is based on this dataset.

In the initial release of the dataset, there are 1,048,575 rows of transaction data and four variables: 'from\_totally\_fake\_account', 'monopoly\_money\_amount', 'to\_randomly\_generated\_account', and 'not\_happened\_yet\_date'. The following is a detailed description of the variables:

**from\_totally\_fake\_account**: virtual transfer out account. Character variable representing the trading account.

**monopoly\_money\_amount**: virtual transaction amount. Numeric variable representing the spend/transfer amount.

**to\_randomly\_generated\_account**: virtual to account. Character variable representing the transaction merchant/virtual personal account.

**not\_happened\_yet\_date**: virtual transaction date. Date type variable from 1 January 2025 to 31 December 2025.

1. Data Quality

The initial dataset has few variables, so the data preprocessing part is simpler. We firstly performed the operations of removing duplicate values and checking for missing values on the dataset. After that, we preprocessed the raw data based on the characteristics of the two columns: ‘to\_randomly\_generated\_account’ and ‘not\_happened\_yet\_date’.

**Removing Duplicate Values**: we removed duplicate values in the dataset considering that in the exact same transaction data (same time, place, amount) is unlikely in real life. The de-duplicated dataset contains 10,148,280 pieces of data.

**Handling Missing Values**: after checking, there are no missing values in the de-duplicated dataset.

**Data Inconsistency**: We observed that the ‘to\_randomly\_generated\_account’ column contains both merchant names, e.g. CAFÉ, DVD\_SHOP, and personal accounts (five digits), so in the subsequent exploratory analyses we categorised the consumer accounts and classified the personal accounts as "person" for ease of follow-up.

**Parse Dates**: We note that we can extract the transaction month from the date to see which months of the year are more active. We also plan to convert the transaction dates to Monday to Sunday, so that we can further analyse the weekly trading patterns and get more valuable information.

IV. 数据说明/准备

A. 数据说明

本项目中使用的数据集来自Marie博士。原始交易数据提供了客户互动记录，包括日期、时间、金额、方式和相关地点。出于保密性的原因，原始交易数据经过基于代理的模拟，生成个人层面的人工交易数据。我们的研究是基于这个数据集展开的。

在最初发布的数据集中，共有1048575行交易数据，和四个变量，分别是from\_totally\_fake\_account，monopoly\_money\_amount，to\_randomly\_generated\_account，not\_happened\_yet\_date。下面是变量的详细说明：

from\_totally\_fake\_account：虚拟的转出账户。字符型变量，代表了交易账户。

monopoly\_money\_amount：虚拟的交易金额。数值型变量，代表了消费/转账金额。

to\_randomly\_generated\_account：虚拟的收款账户。字符型变量，代表了交易商户/虚拟的个人账户

not\_happened\_yet\_date：虚拟的交易日期。日期型变量，从01/01/2025到31/12/2025。

B. 数据质量

初始数据集的变量不多，所以数据预处理部分较为简单。我们首先对数据集进行了去重、检查缺失值的操作。之后根据收款账户和交易日期两列的特征，对原始数据进行预处理。

数据去重：我们考虑到在完全相同的交易数据（相同的时间、地点、金额）在现实生活中可能性很小，所以我们删除了数据集中的重复值。去重后的数据集包含10148280条数据。

处理缺失数据：经过检查，去重后的数据无缺失值。

数据不一致：我们观察到收款账户列中既包含了商户名称，例如…，同时包含个人账户（五位数字），所以在之后的探索性分析中我们把消费账户进行分类，同时把个人账户归为“个人”类，以方便后续的工作。

交易日期：我们注意到我们可以在日期中提取交易月，来观察一年中哪些月份的交易较为活跃。同时，我们计划把交易日期转化为周一到周日，这样我们可以进一步分析每周中的交易规律，得到更有价值的信息。