a) Business Context

eyos collects granular transaction receipt data from a panel of independent grocery retailers. The collected data is parsed from printed paper receipts and is subjected to both data collection errors and store – level nuances.

We have curated a dataset which consists of name groups (i.e. barcodes), each containing a number of product names parsed from different retailers. Each row of the dataset contains a variant version of the product name with its associated name group.

In this assessment, applicants are invited to suggest approaches to identify outlier instances, and develop a data quality system to flag out stores with potential issues.

b) Dataset details

Outlier Detection

Each name group may contain outlier product names that are significantly different from other product names captured within the group. Outliers include but are not limited to:

1. Product name is not available
2. Product name is conjoined with another product name
3. Product name looks clean, but is intuitively different from other names in the group, for example:

SEAGULL NAPTH 25g WRNA / PCS	8886012805206
SEA GULL WARNA RENTENG	8886012805206
MANGKOK SAMBAL ALL VAR	8886012805206
SEAQULL NAPT WARNA 25GR	8886012805206

Focusing on 3., propose an outlier detection model that identifies name-barcode matchings within each name group that are likely to be incorrect. More examples include:

KECAP MANIS SK GANDARIA R	8990090003772
SOKLIN LANTAI H MUDA 400ML REF	8992727001724
BAYGON LIQUID ELEC.SILKY JAS	8992727001724

Note that there may be noise in the product names i.e. "- *SEA GULL NAPHT 25GR SG-519W 1PCSX 1.500,00:*" but this does not qualify as a wrongly matched product name.

c) Evaluation

- Introduce your thought process in experimentation, understanding the problem, developing, and selecting a model
- Design a high level pipeline on how the model would be tested, deployed, and maintained
- Points to note:
    - Problem understanding & data investigation
    - Making assumptions and clarifications about dataset
    - Establishing baselines and curation of data
    - Model selection & training
    - Deployment, monitoring, and maintenance
    - Communication of insights and knowledge transfer