PRODUCT TRACING

One of the key aspects of the program is the construction of the path of the product from the importer to the final seller. To achieve this product tracing logic was implemented. This part of the program is responsible for finding the most probable paths by which the product reached the final seller using various metrics.

Tracing Workflow

1. Input
As an input we take the raw and processed datasets of HDM, invoice and Declarations.
Next we take user input

- Timeframe: start and end date for each data which the user wants to explore.
- HDM Tin: Tin of the seller that the user wants to explore.
- Category: Category which the user wants to explore
- Brand: Brand which the user wants to explore

2. Data Preparation
In this part we receive data from two different sources. First source is the raw data and the second source is the processed data received from the NLP pipeline. We aggregate the raw data with the metrics that are needed in the tracing and merge it with the processed data to get the additional information required for tracing. As a result we receive three working dataframes, containing information from

1. Hdm Checks
2. Invoiced
3. Declarations

3. Declaration Processing
In this part we process and prepare the declaration data. According to the input timeframe we extract the quantity of the products imported and normalise them with the same logic as in Unit Measure Extraction for HDM and Invoice. This process should be done within the tracing process since it is dependent on the timeframe of a specific tracing instance.
4. Price Processing
Next we process the prices from all data frames, we normalise them according to the extracted unit of measure. This process is also dependent on the timeframe of a specific tracing instance and should be completed within the tracing process.
5. Tracing
After preparing data with and adding all the required information we proceed to the tracing and path construction which consists of several key steps.

1. For start we isolate the data. HDM and Declaration data is isolated based on the input Brand and Category while the Invoice data is isolated only based on the Category. This decision was made based on the Invoice data specification. A lot of sold products in the Invoice do not include brand name and therefore the brand extraction is not efficient on this data.

2. Next we separate these datasets into three parts based on the categorization score. This is done to separate the results of the predictions into three different confidence levels. The separation is done with following product categorization scores.
   High Confidence: 0.8 - 1

Medium Confidence: 0.5 - 0.8

Low Confidence: 0 - 0.5

All three parts of the datasets are passing through the path construction process separately. At this step we report the distribution of the data based on the score.

1. Next part is the preparation of data for path construction. First we receive all the unique importer tins for the category brand pair. Next we isolate the invoice data (invoice data separated by only threshold, NO CATEGORY) where we specify that the supplier tin should be in the unique importer tins for the category brand pair OR the buyer tin should correspond to the input buyer tin. We keep this data and add category isolation to the invoice data.
   As a result we get two invoice data. First includes all the transactions between category brand pairs of the importer and the final seller. Here we do not have high confidence that the product is the exact same product that we are looking for.

Second data is also isolated by the category (in the invoice). Here we have higher confidence that we found the right product.

1. From the two invoice data frames received from the previous step we construct directed graphs using the NetworkX library. One graph represents all transactions (the full graph), and another represents transactions within the specified category (the category graph). In these graphs, nodes represent companies, and edges represent transactions. This Network model allows us to find all the possible paths between the importer and Final Seller.
2. From the two graphs received in the previous step we extract top shortest paths. From the full graph we will see the top transactions between the importer and seller, while from the category graph we will see the top transaction between the importer and seller within our required category.

In this step we report both sets of the paths.
Next we take the paths from the category graph that are the subset of the full paths from the full graph. In this way we ensure they take the most probable correct paths. This will be our working path from now on.

1. After reporting the working paths we start exploring each path in detail.

For each path we calculate and report the following information

1. Unit Price of product in Declaration, Invoice and HDM.
2. Whole weight of the imported product, whole weight of the supplied product and whole weight of the sold product
3. The marginal growth of the price from declaration to invoice from invoice to HDM and from declaration to HDM.
4. If there are intermediaries we report the number of intermediaries and report all the above information for the intermediaries as well.

All of these steps are done for all probable paths.

1. Steps from 3 to 6 are repeated for all three parts of the data separated by categorization score.

Example of final tracing report.

{'timeframe': {'HDM': {'start_date': '01-MAY-2024 00:00:00',
'end_date': '31-MAY-2024 23:59:59'},
'Dec': {'start_date': '2023-10-01', 'end_date': '2024-06-01'},
'Inv': {'start_date': "01-APRIL-2024 00:00:00',
'end_date': '31-MAY-2024 23:59:59'}},
'HDM_tin': '1282006',
'category': 'rice',
'brand': 'gallo',
'data_division_by_threshold': {'threshold_0.8_2': {'data_distribution': { 'Dec_percentage': 100.0,
'INV_percentage': 60.08222144731771,
'HDM_percentage': 100.0,},
'importers': ['29448'],
'29448': {'top_paths_from_all': ([['1282006', '29448'],
['1282006', '2556388', '29448'],
['1282006', '1021368', '29448'],
['1282006', '2524256', '29448'],
['1282006', '2538542', '29448']],),
'top_paths_from_category': [['1282006', '29448'],
['1282006', '2538542', '29448']],
'intersection': [['1282006', '29448'], ['1282006', '2538542', '29448']],
'Product_statistcs': {
"exploring path ['1282006', '29448']": {'UnitPrice': {'Declaration': {'kg': 1164.9932492905778},
'HDM': {'kg': 4348.6238997089495},
'Invoice': {'kg': 3650.8074615384617}},
'Weight': {'whole_imported': (8545000.0, 'gram'),
'whole_sold': (125000.0, 'gram'),
'whole_supplied': (146500.0, 'gram')},
'PriceMargins': {'Declaration_to_HDM': {'Margin for kg': 273.2746007203941},
'Declaration_to_Invoice': {'Margin for kg': 213.37584692114052},
'Invoice_to_HDM': {'Margin for kg': 19.114030129554564}},
'transaction_type': 'direct',
'number_of_intermediaries': 0,
'brand_in_invoice': 'found'}, "exploring path ['1282006', '2538542', '29448']": {'UnitPrice': {'Declaration': {'kg':
1164.9932492905778},
'HDM': {'kg': 4348.6238997089495},
'invoice_transaction_1': {'kg': 3250.0},
'invoice_transaction_2': {'kg': 1041.6667}},
'Weight': {'whole_imported': (8545000.0, 'gram'),
'whole_sold': (125000.0, 'gram'),
'whole_supplied_transaction_1': (12000.0, 'gram'),
'whole_supplied_transaction_2': (351000.0, 'gram')},
'PriceMargins': {'Declaration_to_HDM': {'Margin for kg': 273.2746007203941},
'Declaration_to_Interemediary_1': {'Margin for kg': 178.97157361032663},

'Interemediary_2_to_HDM': {'Margin for kg': 317.467881013087}}
'transaction_type': 'intermediary',
'number_of_intermediaries': 1,
'brand_in_invoice': 'NOT found'}}}},
'threshold_0.5_0.8': {'data_distribution': {},
'Dec_percentage': 0.0,
'INV_percentage': 38.20544350028447,
'HDM_percentage': 0.0,
'importers': []},
'threshold_0_0.5': {'data_distribution': {},
'Dec_percentage': 0.0,
'INV_percentage': 1.7123350523978196,
'HDM_percentage': 0.0,
'importers': []}}}