Practicum Data v1.0 Schema

accounts.csv (75,252 rows):

- **global_transaction_id:** Global transaction identifier for this dataset
- **ticket_num:** Original ticket number for the transaction
- **account_num_hash:** Hashed account number to link transactions

transactions.csv (124,934 rows):

- **global_transaction_id:** Global transaction identifier for this dataset
- **store_num:** Physical store number
- **ticket_num:** Original ticket number for the transaction
- **date:** Date of the transaction (YYYY-MM-DD)
- **transaction_start_time:** Time the transaction began (HH:MM:SS)
- **transaction_end_time:** Time the transaction concluded (HH:MM:SS)
- **num_items:** The sum of 'qty_sold' of items that make up this transaction. Note: if qty_is_weight = 1, the total is incremented by 1 and not the weight value.
- **ticket_total_value:** The total value in cents of the transaction/ticket

items_transactions.csv (1,293,520 rows):

- **global_transaction_id:** Global transaction identifier for this dataset
- **item_id:** Unique item identifier
- **dept_num:** Unique department identifier
- **qty_sold:** The total number of this item being sold, can be weight if qty_is_weight = 1, else it's a count
- **item_price:** Price of the item in cents
- **qty_is_weight:** Boolean signifying if the qty_sold field is weight (1) or count (0)
- **ticket_num:** Original ticket number for the transaction
- **date:** Date of the transaction
- **time_scanned:** The time that the particular item was scanned/rung up

items_descriptions.csv (48,551 rows):

- **item_id**: Unique item identifier
- **description**: Item description
- **ecomm_description**: E-commerce item description, not available for all items
- **category**: A category id number - we don't have the mapping for what the id stands for however. Could help with categorization.
- **item_type**: Another id number for which we don't have the mapping for what it actually means. Could help with categorization.
- **upc**: The given Universal Product Code. However - does not seems to be globally registered UPCs. Item_id is derived from this column.

Dataset Notes:

• The dataset containing item names (items_descriptions.csv) is large! If you want to work with a smaller set to start, focus on the items that appear in the transactions data first.

• If using pandas, it is recommended to read in the CSVs with dtype=str to avoid auto-formatting to ints. You will notice that Excel will do this as well. Use a text editor to view the raw data.

• Not all items in items_transactions.csv have a description in items_descriptions.csv (but about 97% of them do).