Task 1: Data preparation and customer analytics

What you'll do

- Analyze transaction and customer data to identify trends and inconsistencies.
- Develop metrics and examine sales drivers to gain insights into overall sales performance.
- Create visualizations and prepare findings to formulate a clear recommendation for the client's strategy.

Background Information

As part of an Analytics Team of my company, we are tasked by a client to better understand the types of customers who purchase their products and their purchasing behaviour within the region.

The insights from your analysis will feed into the supermarket's strategic plan.

Provided Data

Link to Transactions data

Link to Purchase Behaviour data

Understand The Data

Transaction Dataset:

Observation	Details
Columns:	· Date → *Date of transaction* · STORE_NBR → *Store number* · LYLTY_CARD_NBR → *Loyalty Card number* · TXN_ID → *Transaction ID* · PROD_NBR → *Product number.* · PROD_NAME → *Product name* · PROD_QTY → *Product quantity* · TOT_SALES → *Total sales*
Rows:	A total of 264,836 rows
Data Cleanings Needed:	· Rename dataset query to Transactions · Date converted from excel number type to date type · Removed all records that are not for transactions involving "Chips" · Identified outlier by using → `IsOutlier = IF(ABS(ChipTransactions[PROD_QTY] - AVERAGE(ChipTransactions[PROD_QTY])) > 2 * STDEV.P(ChipTransactions[PROD_QTY]), 1, 0)`

DATA PREPARATION

==========

Objective: Identify Outliers

Based on the columns only quantity (PROD_QTY) sold can have meaningful outliers. We can identify outliers by using the following DAX query:

```
IsOutlier = IF(ABS(ChipTransactions[PROD_QTY] - AVERAGE(ChipTransactions[PROD_QTY])) > 2 *
STDEV.P(ChipTransactions[PROD_QTY]), 1, 0)
```

I checked to confirm that the same customer purchased items that produced the outliers. Also, the same customer only performed the 2 outlier transactions. I checked by filtering the LYLTY_CARD_NBR by 226000.

Objective: Remove Outliers

In Transform Pane, I remove records for customer with LYLTY_CARD_NBR 226000 by filtering out on LYTY_CARD_NBR column.

Objective: Chart Transactions Over Time

- Create a Date table in Table View
- Use `Dates = CALENDARAUTO()` to generate all dates within the existing data range
- Change date type in Date table to match date type in Transactions table
- Create Column chart using month field in Date table and TXN_ID (Transaction ID) column in Transactions table.

Objective: Analyze December Transactions:

- Filter the existing Transactions dataset by month: December
- Create a column chart with x-axis → Date and y-axis Transactions

Observations:

1. No transactions occurred on Dec 25. This is because the shop was closed on Christmas day.
1. Increase in sales occurs in the lead up to Christmas day.
1. Most transactions happened on Dec 24.
1. Least transactions happened on Dec 12.

Objective: Histogram of Pack Sizes

- In the Transform UI,

o Duplicate Transactions query dataset

o Rename duplicated dataset to “TxnPackSizes”

o Rename PROD_NAME column to PACK_SIZE

o Cleaned data on PACK_SIZE column leaving only the numeric size value

- In the Main UI

o Grouped PACK_SIZE field and created bins (PACK_SIZE_BINs)

o Used PACK_SIZE_BIN and TXN_ID fields to create Pack Size Histogram

Observations:

1. Bin with most transactions is bin 163-194 (106,133 transactions)
1. Bin with least transactions is bin 225-256 (3,169 transactions)

Objective: Add Brand Column to Transactions dataset

1. Duplicate Transactions table (query) and rename duplicated table to Brands
1. Duplicate PROD_NAME column in Brands table and rename duplicated column to BRAND
1. Run the following Python script to extract brand names from PROD_NAME values
1. Delete all columns from Brands table except for TXN_ID and BRAND columns
1. Change type of TXN_ID column to whole number and type of BRAND column to text
1. Merge a new query (table/dataset) from Transactions and Brands tables
1. Rename the new query (table) to Txn+Brands

```
# 'dataset' holds the input data for this script
```

```
import pandas as pd
```

```

```

```
# Assuming 'dataset' is your DataFrame
```

```
dataset['BRAND'].replace({
```

```
    'Burger Rings.*': 'Burger Rings',
```

```
    'CCs.*|Cheetos.*': 'Cheetos',
```

```
    'Cheezels.*': 'Cheezels',
```

```
    'Cobs Popd.*': 'Cobs Popd',
```

```
    'Dorito.*': 'Doritos',
```

```
    'French Fries.*': 'French Fries',
```

```
    'Grain Waves.*|GrnWves.*': 'Grain Waves',
```

```
    'Infuzions.*|Infzns.*': 'Infuzions',
```

```
    'Kettle.*': 'Kettle',
```

```
    'Natural.*|NCC.*': 'Natural Chip Company',
```

```
    'Old.*': 'Old El Paso',
```

```
    'Pringles.*': 'Pringles',
```

```
    'Red.*|RRD.*': 'Red Rock Deli',
```

```
    'Smit.*': 'Smiths',
    'Sunbites.*|Snbts.*': 'Sunbites',
    'Thins.*': 'Thins',
    'Tos.*': 'Tostitos',
    'Twi.*': 'Twisties',
    'Ty.*': 'Tyrrells',
    'Wool.*|WW.*': 'Woolworths'
}, regex=True, inplace=True)

```

```
dataset
```

Objective: Merge Customer Data to Transactions Data

1. Create a new query (dataset) from the customer data csv file, name it “Customer Data”.
1. Confirmed that there was no errors or nulls in the Customer Data dataset
1. Merge Customer Data to Transactions Data using Left Outer Join (on the LYLTY_CARD_NBR column)
1. Name the merged dataset “TxnCustomerData”

CUSTOMER ANALYTICS

==========

Objective: Who spends the most on chips (total sales), describing customers by lifestage and how premium their general purchasing behaviour is

Sales are coming mainly from Budget - older families, Mainstream - young singles/couples, and Mainstream - retirees

Objective: How many customers are in each segment

There are more Mainstream - young singles/couples and Mainstream - retirees who buy chips. This contributes to there being more sales to these customer segments

Objective: How many chips are bought per customer by segment

Older families and young families in general buy more chips per customer

Objective: What's the average chip price by customer segment

Mainstream midage and young singles and couples are more willing to pay more per packet of chips compared to their budget and premium counterparts. This may be due to premium shoppers being more likely to buy healthy snacks and when they buy chips, this is mainly for entertainment purposes rather than their own consumption. This is also supported by there being fewer premium midage and young singles and couples buying chips compared to their mainstream counterparts.

Objective: Customer's total spend over the period and total spend for each transaction to understand what proportion of their grocery spend is on chips

Objective: Proportion of customers in each customer segment overall to compare against the mix of customers who purchase chips

1.