Copy of Retail Store Product Category Sales Analysis

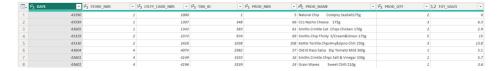> **ℹ Background Information**
>
> The client is a major retail store business.
>
> As part of a consultant analytics team, we are tasked by the client to better understand the types of customers who purchase their products under the "chips" category.
>
> dataset sourced from Quantium/TheForage

Available Datasets

Link to Transactions data

Link to Purchase Behaviour data

Primary Objectives

1. Understand the datasets

2. Clean and prepare the datasets

3. Who spends the most on chips (total sales) by life-stage and customer category

4. What are the total number of customer transactions by life-stage and customer category

5. How many chips (quantity) are bought by life-stage and customer category

6. What is the average chip price by life-stage and customer category

---

> ⌄ Task 1
>
> ## TASK 1: Understand the datasets
>
> The following is a breakdown of the Transactions dataset with explanations
>
> | Observation | Details |
> |---|---|
> | **Columns:** | • Date → *Date of transaction*
 • STORE_NBR → *Store number*
 • LYLTY_CARD_NBR → *Loyalty Card number*
 • TXN_ID → *Transaction ID* |

Observation	Details
	- PROD_NBR → *Product number* - PROD_NAME → *Product name* - PROD_QTY → *Product quantity* - TOT_SALES → *Total sales*
Rows:	A total of 264,836 rows

The following is a breakdown of the Customer dataset with explanations

Observation	Details

Columns:	- LYLTY_CARD_NBR → *Loyalty Card number*
	- LIFESTAGE → *Life stage of customer*
	    ◦ YOUNG SINGLES/COUPLES
	    ◦ MIDAGE SINGLES/COUPLES
	    ◦ OLDER SINGLES/COUPLES
	    ◦ NEW FAMILIES
	    ◦ YOUNG FAMILIES
	    ◦ OLDER FAMILIES
	    ◦ RETIREES
	- PREMIUM_CUSTOMER → *Customer category*
	    ◦ Budget
	    ◦ Mainstream
	    ◦ Premium

Rows:	A total of 72,637 rows

⌄ Task 2: Clean and prepare datasets

TASK 2: Clean and prepare datasets

✅ **First - 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:

```
1  IsOutlier =
2    IF(
3      ABS(ChipTransactions[PROD_QTY] - AVERAGE(ChipTransactions[PROD_QTY])
4          ) > 2 * STDEV.P(ChipTransactions[PROD_QTY])
5      ), 1, 0
6    )
```


📋 **Observation:** It appears one customer purchased 200 units of Dorito Corn Chip twice.

This is an outlier as all other single purchase quantities range between 1 and 5. This outlier needs to be removed from the dataset so it does not skew results.

Use the following power query m-code targeting the outlier transaction id to filter out the outliers

```
1  = Table.SelectRows(#"Filtered Rows", each [LYLTY_CARD_NBR] <> 226000)
```

✅ **Next - Create a "TransactionPackSizes" table from Transactions table**

- In the Power BI Transform UI
  - Duplicate Transactions table/query and rename duplicate table to TransactionPackSizes
  - Rename PROD_NAME column to PACK_SIZE
  - Clean data on the PACK_SIZE column leaving only the numeric size value

✅ **Next - Create a "Brand" column extracting product brands from PROD_NAME column**

- Duplicate PROD_NAME column and rename duplicate column to BRAND
- Observe that the brand part of the product name is usually at the start
- Also observe that some brand names or parts of brand names are abbreviated
- Use the following python script to extract only the brand part from the product names in the brand columns:

```
1  # 'dataset' holds the input data for this script
2  import pandas as pd
3
4  # Assuming 'dataset' is your DataFrame
5  dataset['BRAND'].replace({
6      'Burger Rings.*': 'Burger Rings',
7      'CCs.*|Cheetos.*': 'Cheetos',
8      'Cheezels.*': 'Cheezels',
9      'Cobs Popd.*': 'Cobs Popd',
10     'Dorito.*': 'Doritos',
11     'French Fries.*': 'French Fries',
12     'Grain Waves.*|GrnWves.*': 'Grain Waves',
13     'Infuzions.*|Infzns.*': 'Infuzions',
14     'Kettle.*': 'Kettle',
15     'Natural.*|NCC.*': 'Natural Chip Company',
16     'Old.*': 'Old El Paso',
17     'Pringles.*': 'Pringles',
18     'Red.*|RRD.*': 'Red Rock Deli',
19     'Smit.*': 'Smiths',
20     'Sunbites.*|Snbts.*': 'Sunbites',
21     'Thins.*': 'Thins',
22     'Tos.*': 'Tostitos',
23     'Twi.*': 'Twisties',
24     'Ty.*': 'Tyrrells',
25     'Wool.*|WW.*': 'Woolworths'
26  }, regex=True, inplace=True)
27
28  dataset
```

✅ **Next - Merge customer data to transactions data**

- Create a new query (dataset) from the customer data csv file, name it "Customer Data".
- Confirmed that there was no errors or nulls in the Customer Data dataset

- Merge Customer Data to Transactions Data using Left Outer Join (on the LYLTY_CARD_NBR column)
- Name the merged dataset "TxnCustomerData"

TASK 1: Understand the datasets

> ❌ TODO: MOVE TO ANALYSIS PART
> - In the Power BI Main UI
>   - On the TransactionPackSizes dataset
>   - Group PACK_SIZE field and created bins (PACK_SIZE_BINs)
>   - Use PACK_SIZE_BIN and TXN_ID fields to create Pack Size Histogram