Retail Store Product Category Sales Analysis

Background Information

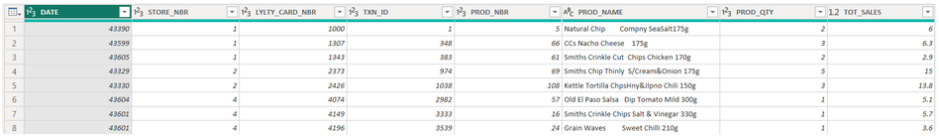
The client is a major retail store business.

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



	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES
1	43390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	6
2	43599	1	1307	348	66	CCs Nacho Cheese 175g	3	6.3
3	43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	2.9
4	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	15
5	43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	13.8
6	43604	4	4074	2982	57	Old El Paso Salsa Dip Tomato Mild 300g	1	5.1
7	43601	4	4149	3333	16	Smiths Crinkle Chips Salt & Vinegar 330g	1	5.7
8	43601	4	4196	3539	24	Grain Waves Sweet Chilli 210g	1	3.6

Link to Purchase Behaviour data

	LYLTY_CARD_NBR	LIFESTAGE	PREMIUM_CUSTOMER
1	1000	YOUNG SINGLES/COUPLES	Premium
2	1002	YOUNG SINGLES/COUPLES	Mainstream
3	1003	YOUNG FAMILIES	Budget
4	1004	OLDER SINGLES/COUPLES	Mainstream
5	1005	MIDAGE SINGLES/COUPLES	Mainstream
6	1007	YOUNG SINGLES/COUPLES	Budget
7	1009	NEW FAMILIES	Premium
8	1010	YOUNG SINGLES/COUPLES	Mainstream
9	1011	OLDER SINGLES/COUPLES	Mainstream
10	1012	OLDER FAMILIES	Mainstream

Primary Objectives

1. Understand the datasets
2. Clean and prepare the datasets
3. What is the most popular packet size
4. Who spends the most on chips (total sales) by life-stage and customer category
5. What are the total number of customer transactions by life-stage and customer category
6. How many chips (quantity) are bought by life-stage and customer category
7. What is the average chip price by life-stage and customer category
8. Summary of Observations

TASK 1: Understand the datasets

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* • 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

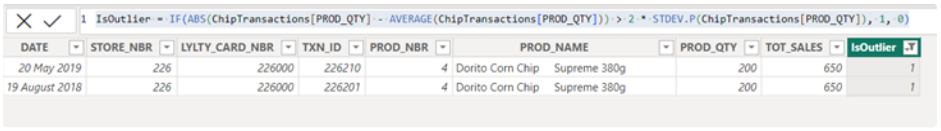
TASK 2: Clean and prepare datasets

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:

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



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

```
= 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:

```python
# '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
```

Next - Merge customer data to transactions data

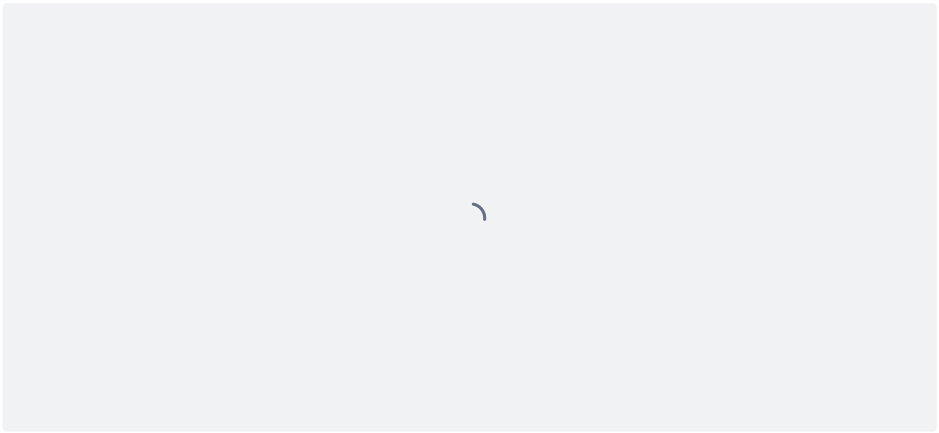
- Create a new query (dataset) from the customer data csv file, name it “Customer Data”.
- Confirmed that there was no errors or nulls in the Customer Data dataset
- Merge Customer Data to Transactions Data using Left Outer Join (on the LYLTY_CARD_NBR column)
- Name the merged dataset “TxnCustomerData”

At this point we should have a unified table with the following columns:

TASK 3: What is the most popular packet size

TASK 3: What is the most popular packet size

- On TxnPackSizes dataset -
  - Group PACK_SIZE field and created bins (PACK_SIZE_BINs)
  - Use PACK_SIZE_BIN and TXN_ID fields to create Pack Size Histogram:



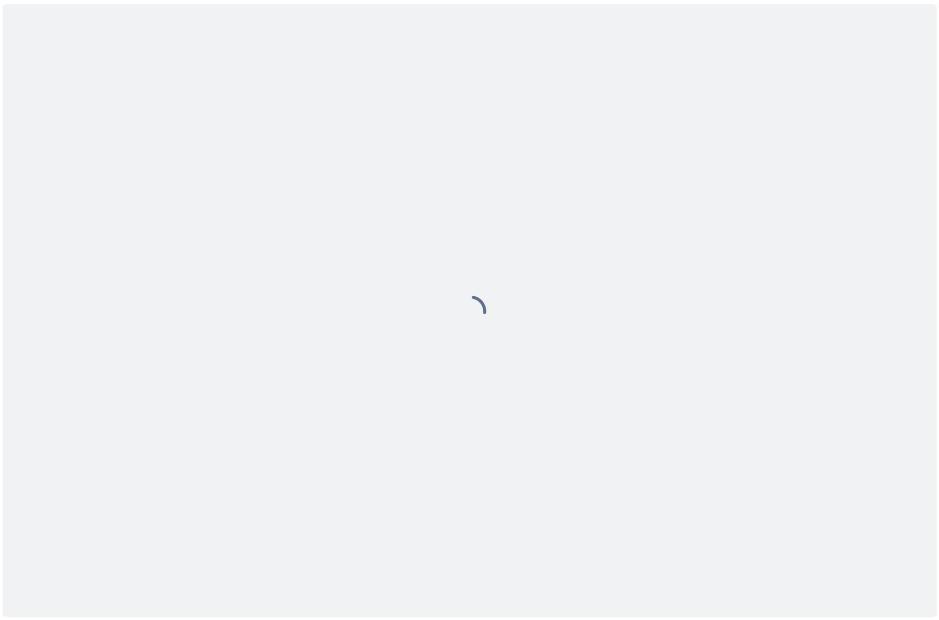
Observations:

- Most popular pack sizes are within bin 163-194 (106,133 transactions)
- Least popular pack sizes are within bin 225-256 (3,169 transactions)

TASK 4: Who spends the most on chips (total sales) by life-stage and customer category

TASK 4: Who spends the most on chips (total sales) by life-stage and customer category

Create a multicolumn column chart with customer life-stage as the x-axis, total sales as the y-axis and life-stage segmented by customer category:

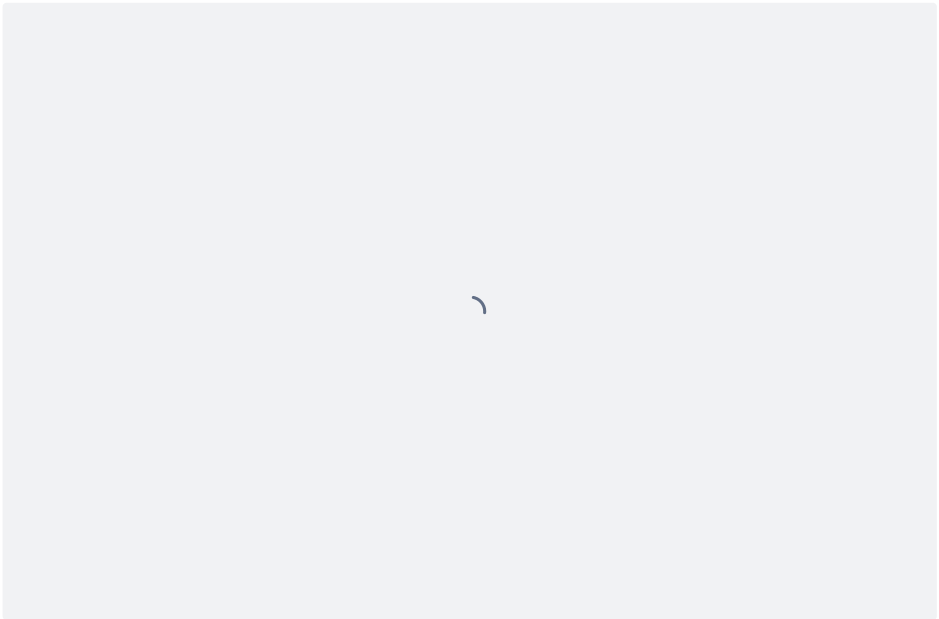


Observations: Sales are coming mainly from Budget - older families, Mainstream - young singles/couples, and Mainstream - retirees

TASK 5: What are the total number of customer transactions by life-stage and customer category

TASK 5: What are the total number of customer transactions by life-stage and customer category

Create another multicolumn column chart but this time with customer life-stage as the x-axis, total transactions as the y-axis and life-stage segmented by customer category:

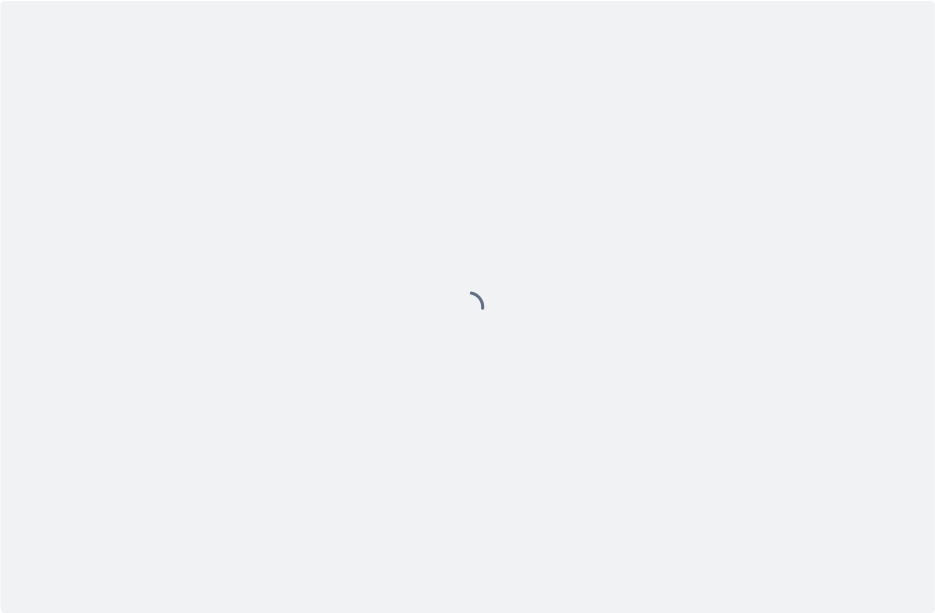


Observations: There are more Mainstream - young singles/couples and Mainstream - retirees who buy chips. This contributes to there being more sales to these customer segments

TASK 6: How many chips (quantity) are bought by life-stage and customer category

TASK 6: How many chips (quantity) are bought by life-stage and customer category

Create another multicolumn column chart but this time with customer life-stage as the x-axis, Average “Price Per Unit” as the y-axis and life-stage segmented by customer category:

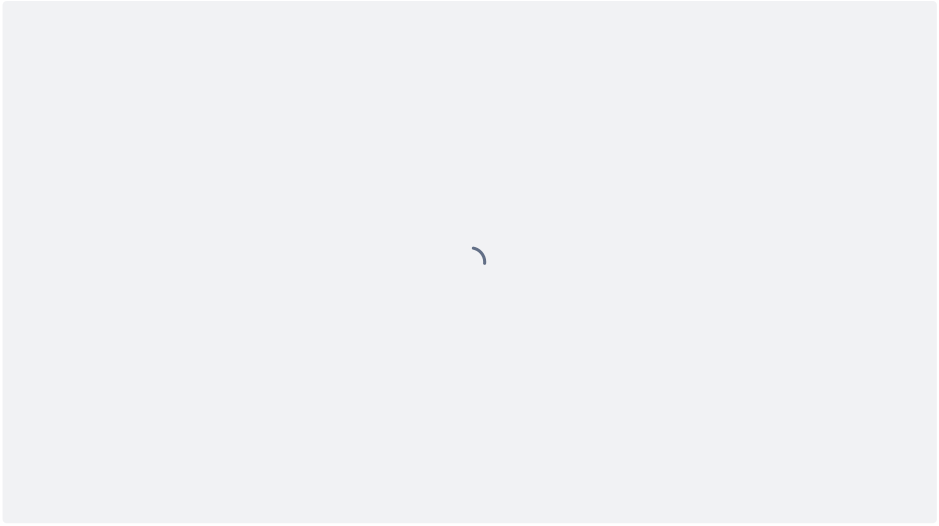


Observations: Older families and young families in general buy more chips per customer.

TASK 7: What is the average chip price by life-stage and customer category

TASK 7: What is the average chip price by life-stage and customer category

- Add another column to the TxnCustomerData dataset named “Price Per Unit”. It would be a measure computed by the following DAX formula:
- ```
  Price Per Unit = DIVIDE(TxnCustomerData[TOT_SALES], TxnCustomerData[PROD_QTY])
  ```
- Create another multicolumn column chart but this time with customer life-stage as the x-axis, Average “Price Per Unit” as the y-axis and life-stage segmented by customer category:
  ```



Observations: 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.

Summary of Observations

- Most popular pack sizes are within bin 163-194 (106,133 transactions)
- Least popular pack sizes are within bin 225-256 (3,169 transactions)
- Sales are coming mainly from Budget - older families, Mainstream - young singles/couples, and Mainstream - retirees
- There are more Mainstream - young singles/couples and Mainstream - retirees who buy chips. This contributes to there being more sales to these customer segments
- Older families and young families in general buy more chips per customer.
- 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.