

Copy of 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 Quantum/TheForage

Available Datasets

[Link to Transactions data](#)

	P ₂ DATE	P ₂ STORE_NBR	P ₂ LYLTY_CARD_NBR	P ₂ TXN_ID	P ₂ PROD_NBR	A ₂ PROD_NAME	P ₂ PROD_QTY	I ₂ TOT_SALES
1	43390	1	1000	1	5	Natural Chip Company SeaSalt 175g	2	6
2	43399	1	1367	348	66	CCs Nacho Cheese 175g	3	4.3
3	43405	1	1343	383	62	Smiths Crinkle Cut Chips Chicken 170g	2	2.9
4	43429	2	2373	974	69	Smiths Chip Thinsly S/Cream&Onion 175g	5	15
5	43330	2	2426	1038	108	Kettle Tortilla ChipsHinyKijino 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	Gran Waves Sweet Chili 210g	1	3.6

[Link to Purchase Behaviour data](#)

	I ₃ LYLTY_CARD_NBR	A ₃ LIFESTAGE	A ₃ 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. 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:	<ul style="list-style-type: none">• Date → <i>Date of transaction</i>• STORE_NBR → <i>Store number</i>• LYLTY_CARD_NBR → <i>Loyalty Card number</i>• TXN_ID → <i>Transaction ID</i>

	<ul style="list-style-type: none"> • PROD_NBR → <i>Product number</i> • PROD_NAME → <i>Product name</i> • PROD_QTY → <i>Product quantity</i> • TOT_SALES → <i>Total sales</i>
Rows:	A total of 264,836 rows

The following is a breakdown of the Customer dataset with explanations

Observation	Details
Columns:	<ul style="list-style-type: none"> • LYLTY_CARD_NBR → <i>Loyalty Card number</i> • LIFESTAGE → <i>Life stage of customer</i> <ul style="list-style-type: none"> ◦ YOUNG SINGLES/COUPLES ◦ MIDAGE SINGLES/COUPLES ◦ OLDER SINGLES/COUPLES ◦ NEW FAMILIES ◦ YOUNG FAMILIES ◦ OLDER FAMILIES ◦ RETIREES • PREMIUM_CUSTOMER → <i>Customer category</i> <ul style="list-style-type: none"> ◦ Budget ◦ Mainstream ◦ Premium
Rows:	A total of 72,637 rows

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 )

```

DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES	IsOutlier
20 May 2019	226	226000	226210	4	Dorito Corn Chip Supreme 380g	200	650	1
19 August 2018	226	226000	226201	4	Dorito Corn Chip Supreme 380g	200	650	1

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