REPORT

CUSTOMER SEGMENTS

Quantium's retail Analysis team

by: khanh huyen thai

khanhhuyen114

khanhhuyen1142000@gmail.com

About Report

You are part of ***Quantium's retail analytics team*** and have been approached by your client, the ***Category Manager for Chips***, who wants to better ***understand the types of customers who purchase Chips and their purchasing behaviour within the region.***

The insights from your analysis will feed into the supermarket's strategic plan for the chip category in the next half year.

What we do?

1. **Examine transaction data** – look for inconsistencies, missing data across the data set, outliers, correctly identified category items, numeric data across all tables. If you determine any anomalies make the necessary changes in the dataset and save it. Having clean data will help when it comes to your analysis.

2. **Examine customer data** – check for similar issues in the customer data, look for nulls and when you are happy merge the transaction and customer data together so it's ready for the analysis ensuring you save your files along the way.

3. **Data analysis and customer segments** – in your analysis make sure you define the metrics – look at total sales, drivers of sales, where the highest sales are coming from etc. Explore the data, create charts and graphs as well as noting any interesting trends and/or insights you find. These will all form part of our report to Julia.

4. **Deep dive into customer segments** – define your recommendation from your insights, determine which segments we should be targeting, if packet sizes are relative and form an overall conclusion based on your analysis.

Data Overview

The original dataset contains:

- **data_transaction:** 264,836 rows, 8 columns
- **purchase_behaviour:** 72,637 rows, 3 colums

In dataset **data_transaction,** there are 8 attributes, including:

DATE: date of buy product
STORE_NBR: index of stores
LYLTY_CARD_NBR: index lylty card
TXN_ID: index of txn
PROD_NBR: number id of product
PROD_NAME: name of product
PROD_QTY: number of items of order
TOT_SALES: total price of order

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES
1	2018-10-17	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	6.00
2	2019-05-14	1	1307	348	66	CCs Nacho Cheese 175g	3	6.30
3	2019-05-20	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	2.90
4	2018-08-17	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	15.00
5	2018-08-18	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	13.80
6	2019-05-19	4	4074	2982	57	Old El Paso Salsa Dip Tomato Mild 300g	1	5.10
7	2019-05-16	4	4149	3333	16	Smiths Crinkle Chips Salt & Vinegar 330g	1	5.70
8	2019-05-16	4	4196	3539	24	Grain Waves Sweet Chilli 210g	1	3.60
9	2018-08-20	5	5026	4525	42	Doritos Corn Chip Mexican Jalapeno 150g	1	3.90
10	2018-08-18	7	7150	6900	52	Grain Waves Sour Cream&Chives 210G	2	7.20

Figure 1.1: Fisrt 10 rows and several colomns of dataset data_transaction

In dataset **purchase_behaviour**, there are 3 attributes, including:

LYLTY_CARD_NBR: index lylty card
LIFESTAGE: customer attribute
PREMIUM_CUSTOMER: customer segmentation

	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

Figure 1.2: Fisrt 10 rows and several colomns of dataset purchase_behaviour

Data Exploration

Before jumping to the analysis process, we should do data pre-processing, including cleaning and transformation steps.

Firstly, check for missing values in data_transaction and purchase_behaviour; ensure there are no NULL values.

Secondly, examine outlet values through min, max, and mean, revealing that PROD_QTY has a sudden increase at the max value of 200, deviating significantly from the median. Investigate the outlet, discovering that this anomaly originates from a single customer with LYLTY_CARD_NBR = 226000. Check the DATE column to find timestamps that do not fall within a specific timeframe. Consequently, these instances are attributed to wholesale customers (without excluding outlet values from the dataset).

Thirdly, the DATE has values from '2018-07-01' to '2019-06-30', but when counting the days, there are only 364 days (while, on average, a year has 365 days). Therefore, it is necessary to check the outlet values within the DATE to identify which days have no customer transactions.

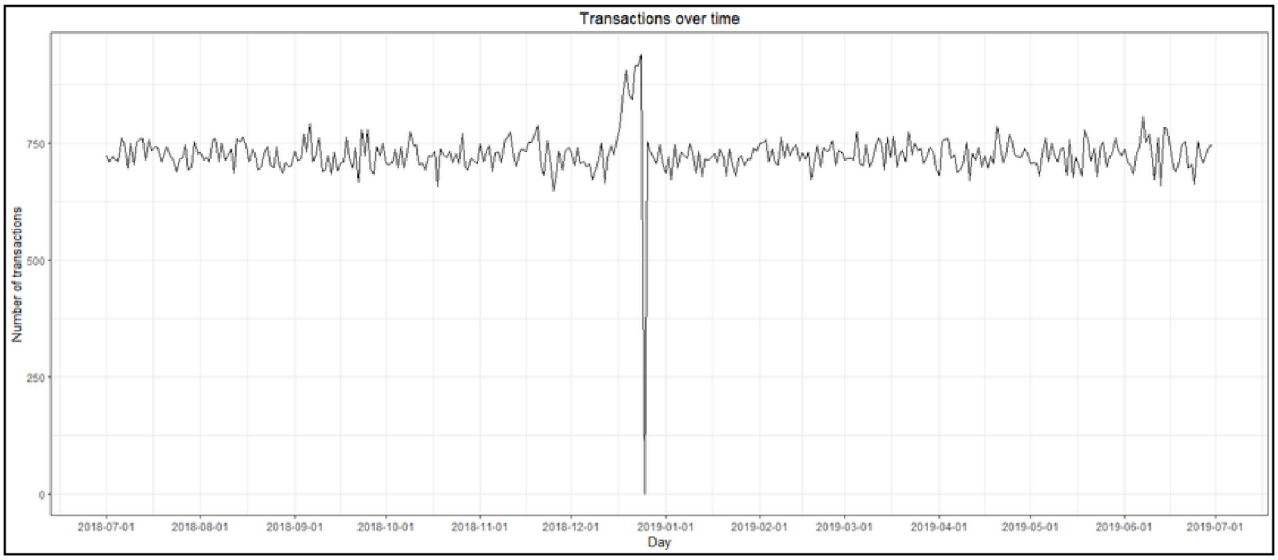


Figure 1.3: Sales Analysis by Month

Analyzing the DATE reveals that in December, there is a day with zero sales. Further detailed analysis will be conducted on each day within the month of December.

Data Exploration

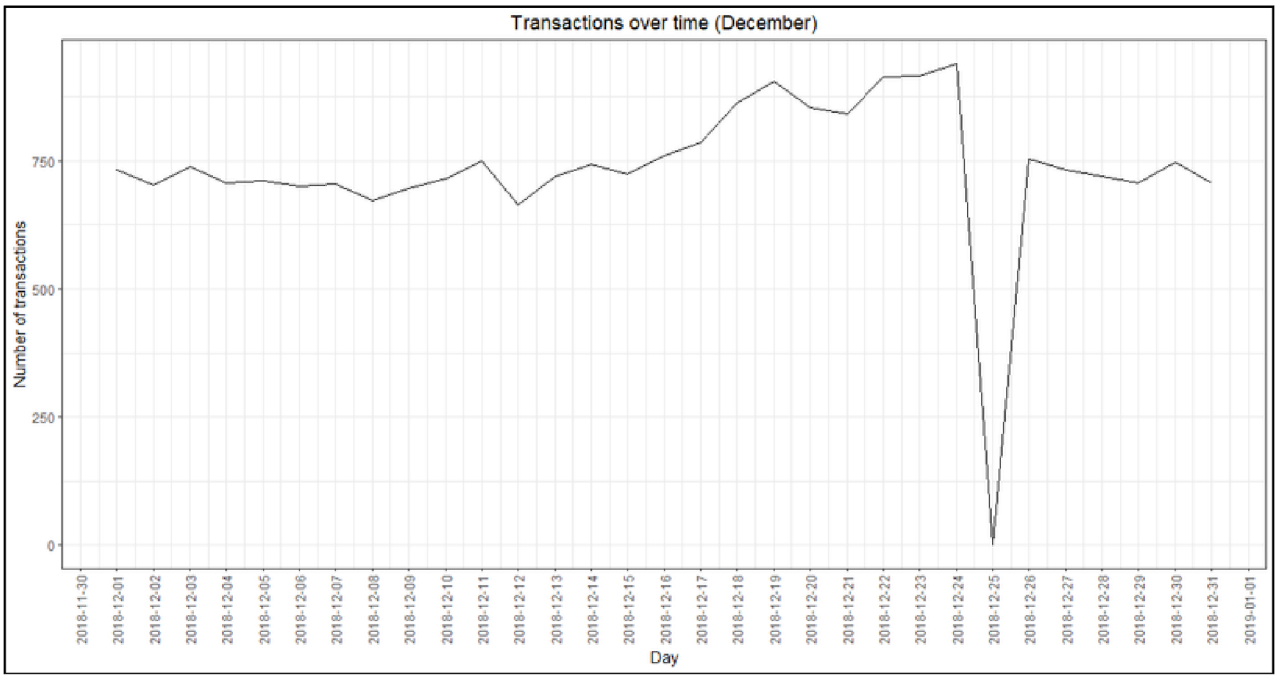
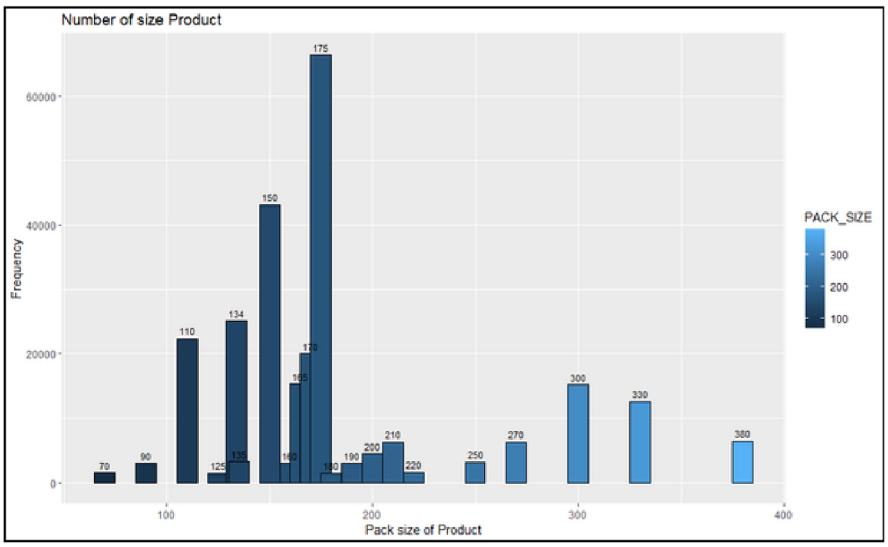


Figure 1.4: Sales Analysis by day in December

We can see that the increase in sales occurs in the lead-up to Christmas and that there are zero sales on Christmas day itself. This is due to shops being closed on Christmas day.

Finally, the PROD_NAME column contains information about BRAND, NAME PRODUCT, and PACKSIZE. I analyze the PROD_NAME column and split it into two columns: BRAND and PACKSIZE, serving for a more in-depth analysis.



The largest size is 38g and the smallest size is 70g - seems sensible.

Figure 1.5: Volume Chart

Data Exploration

When looking at the remaining attributes of dataset **data_transaction**:

- **DATE** value range from “2018-07-01” to “2019-06-30”
- **STORE_NBR** value range from 1.0 to 272.0 and median = 130.0
- **LYLTY_CARD_NBR** value range from 1,000 to 2,373,711 and median = 130,358
- **TXN_ID** value range from 1 to 202701 and median = 135,138
- **PROD_NBR** value range from 1 to 114.0 and median = 56.0
- **PROD_NAME** have length 264,836
- **PROD_QTY** value range from 1.000 to 200.000
- **TOT_SALES** value range from 1.500 to 650.000
- **BRAND** have leghth 264,836
- **PACKSIZE** value range from 70.0 to 380.0 and median = 170.0

After the data validation, the dataset transaction contains 264,836 rows and 10 columns without missing values.

In the case of the purchase_behaviour dataset, check for the absence of NULL values as well as outlet values. Therefore, retain the original purchase_behaviour dataset, which contains 72,637 rows and 3 columns:

- **LYLTY_CARD_NBR** value range from 1,000 to 2,373,711 and median = 134,040
- **LIFESTAGE** have length 72,637
- **PREMIUM_CUSTOMER** have length 72,637

Data analysis on customer segments

Now that the data is ready for analysis, we can define some metrics of interest to the client:

- Who spends the most on chips(total sales), describing customer by lifestage and how premium their
- How many customers are in each segment
- How many chips are bought per customer by segment
- What's the average chip price by customer segment

We could also ask our data team for more information. Examples are:

- The customer's total spend over the period and total spend for each transaction to understand what proportion of their grocery spend is on chips
- Proportion of customers in each customer segment overall to compare against the mix of customers who purchase chips.

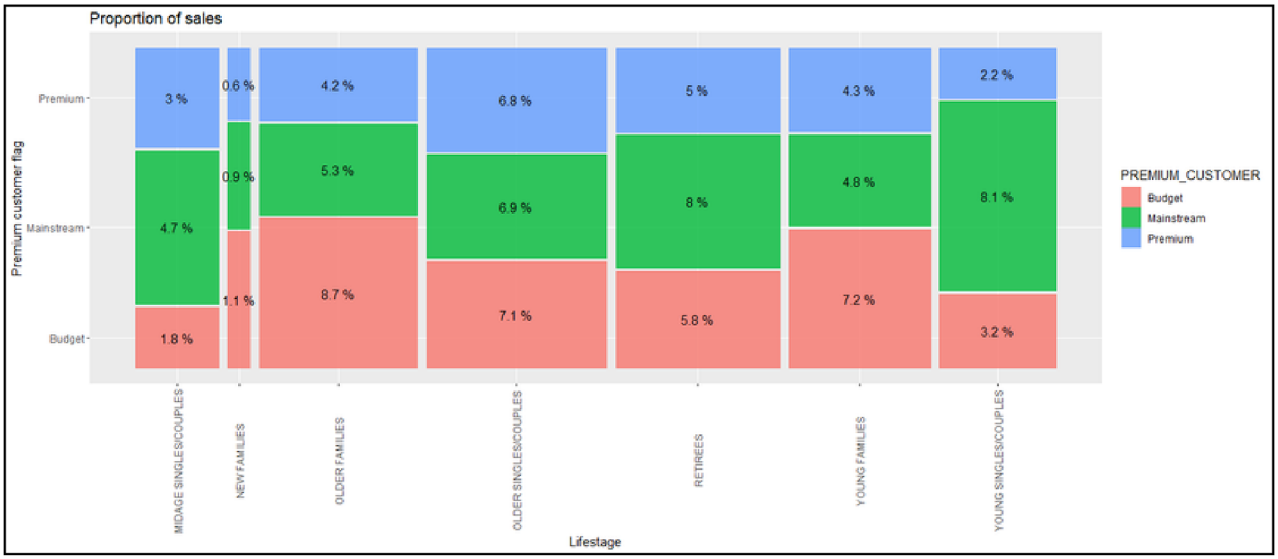


Figure 1.6: Sales Chart Across Customer Segments

Sales mainly come from 3 main sources:

- 8.7% from the OLDER FAMILY - Budget segment
- 8.2% from YOUNG SINGLE/COUPLES - Mainstream segment
- 8.1% from RETIREES - Mainstream segment

Sales are coming mainly from Budget - Older families, Mainstream - Young singles/Couples and Mainstream retirees.

Data analysis on customer segments

Analyzing the customer segments, we have:

- 11.1% of customers are YOUNG SINGLES/COUPLE with a buying source of Mainstream. 8.9% of customers are RETIREES with a buying source of Mainstream.
-

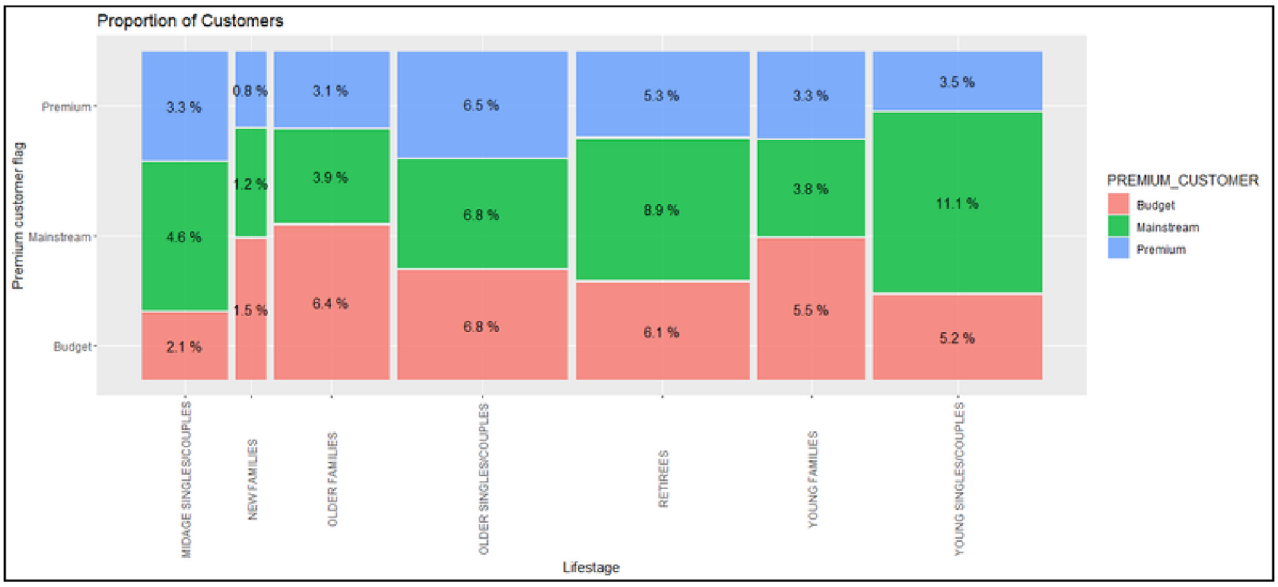


Figure 1.7: Customer Ratio in Each Segment

From the chart depicting sales and customer quantities across segments, we observe that Older Families, despite having the highest sales, do not have the highest customer count. A high customer count is not the factor contributing to the highest sales in the Older Families segment compared to other segments.

So, where does the reason lie?

This could be explained by the fact that individual customers in the Older Families segment make more significant purchases in a single shopping instance. To verify this assertion, I calculated the percentage of the average quantity purchased per customer."

Data analysis on customer segments

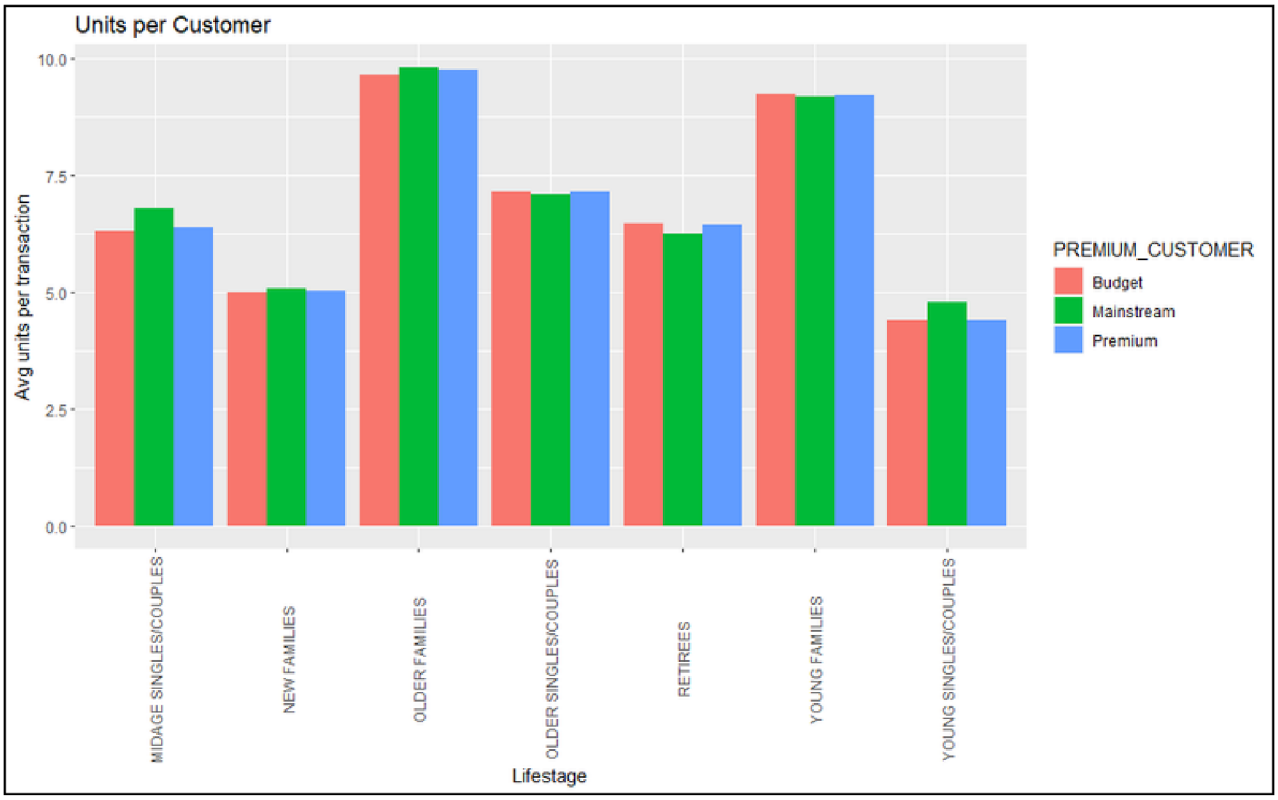


Figure 1.8: Customer Ratio in Each Segment

The average number of products purchased per person in the Older Families and Young Families segments is the highest (consistent with the hypothesis mentioned above).

Examining the average price for each product in each segment, we find that Mainstream Midage and Young Singles and Couples are more willing to pay more per packet of chips compared to their budget and premium counterparts.

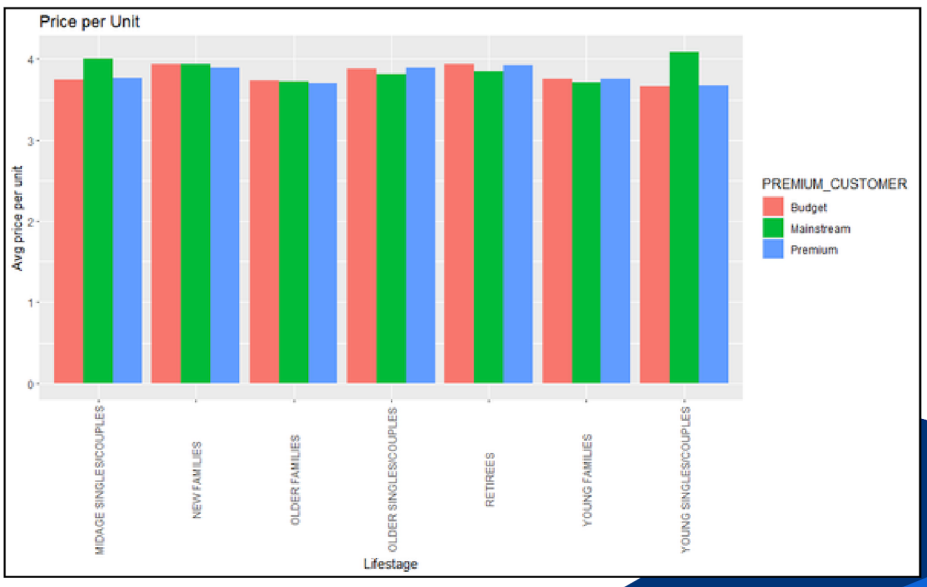


Figure 1.9: Customer Ratio in Each Segment

Data analysis on customer segments

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. As the difference in average price per unit isn't large, we can check if this difference is statistically different.

Perform an independent t-test between mainstream vs premium and budget midage and young singles and couples.

```
        Welch Two Sample t-test

data:  data_total[LIFESTAGE %in% c("YOUNG SINGLES/COU
NPLES", "MIDAGE SINGLES/COUPLES") & PREMIUM_CUSTOMER
== "Mainstream", price] and data_total[LIFESTAGE %in%
c("YOUNG SINGLES/COUNPLES", "MIDAGE SINGLES/COUPLES")
& PREMIUM_CUSTOMER != "Mainstream", price]
t = 18.619, df = 25044, p-value < 2.2e-16
alternative hypothesis: true difference in means is g
reater than 0
95 percent confidence interval:
 0.2337913       Inf
sample estimates:
mean of x mean of y
 4.000101  3.743654
```

Figure 1.10: t-test

The t-test results in a p-value < 2.2e-16, i.e. the unit price for mainstream, young and midage singles and couples are significantly higher than that of budget or premium, young and midage singles and couples.

Deep dive into specific customer segments for insights

We've uncovered several interesting insights that warrant further investigation.

We may consider targeting customer segments that contribute the most to sales to retain them or drive further sales growth. Let's focus on the Mainstream - young singles/couples segment. For example, let's examine whether they show a preference for a specific brand of chips.

Firstly, segregate the segment into 'segment 1' with Lifstage = 'Young Singles/Couples' and Premium = 'Mainstream,' and all other segments within segment 1 as 'other.'

Next, calculate the average quantity of products each customer purchases in 'segment 1' and 'other.'

The brand purchase ratio is as follows:

	BRAND	targetSegment	other	affinityToBrand
1:	TYRRELLS	0.029586871	0.024180309	1.2235936
2:	TWISTIES	0.043306068	0.035620248	1.2157711
3:	TOSTITOS	0.042581280	0.035362737	1.2041285
4:	PRINGLES	0.111979706	0.093573162	1.1967075
5:	KETTLE	0.185649203	0.155549690	1.1935042
6:	OLD	0.041597639	0.035304797	1.1782433
7:	DORITOS	0.122877407	0.106931560	1.1491220
8:	INFUZIONS	0.060649203	0.053053762	1.1431650
9:	COBS	0.041856492	0.037641712	1.1119710
10:	THINS	0.056611100	0.052886380	1.0704287
11:	GRNWVES	0.030674053	0.029446415	1.0416906
12:	CHEEZELS	0.016851315	0.016931367	0.9952720
13:	SMITHS	0.093419963	0.120753478	0.7736420
14:	FRENCH	0.003701595	0.005150226	0.7187247
15:	RRD	0.045376890	0.067519458	0.6720565
16:	NATURAL	0.018378546	0.027740403	0.6625191
17:	CHEETOS	0.007532615	0.011562257	0.6514832
18:	CCS	0.010483537	0.017343385	0.6044689
19:	SUNBITES	0.005953614	0.011221054	0.5305752
20:	WOOLWORTHS	0.028189066	0.056214713	0.5014535
21:	BURGER	0.002743839	0.006012888	0.4563263
	BRAND	targetSegment	other	affinityToBrand

Figure 1.11: The purchase ratio for each brand

The **Young Singles/Couples segment has a 23%** higher likelihood of purchasing **TYRRELLS** compared to the general population and a 56% lower likelihood of purchasing BURGER KING than the general population.

Deep dive into specific customer segments for insights

Next, find out which product weight is most preferred by the Mainstream segment among these Young Singles/Couples customers.

	PACK_SIZE	targetSegment	other	affinityToPack
1:	270	0.029845724	0.023761854	1.2560352
2:	380	0.030156347	0.024025803	1.2551650
3:	330	0.057465314	0.046422846	1.2378671
4:	134	0.111979706	0.093573162	1.1967075
5:	110	0.099658314	0.085043101	1.1718565
6:	210	0.027308967	0.024000052	1.1378712
7:	250	0.013460344	0.011858395	1.1350899
8:	135	0.013848623	0.012605177	1.0986456
9:	170	0.075740319	0.076390722	0.9914858
10:	175	0.239102299	0.248208687	0.9633116
11:	300	0.054954442	0.058126734	0.9454246
12:	150	0.155130462	0.164356576	0.9438653
13:	165	0.052184717	0.058513001	0.8918482
14:	180	0.003365086	0.005446364	0.6178592
15:	190	0.007014910	0.011742514	0.5973942
16:	160	0.006005384	0.010873414	0.5522998
17:	90	0.005953614	0.011221054	0.5305752
18:	125	0.002821495	0.005375548	0.5248758
19:	70	0.002847380	0.005652373	0.5037496
20:	200	0.008412715	0.016789736	0.5010630
21:	220	0.002743839	0.006012888	0.4563263
	PACK_SIZE	targetSegment	other	affinityToPack

From the analysis, the **Mainstream Young Singles/Couples segment is more likely to purchase the 270g chip packet** by over 27% compared to the rest of the population. This implies that the Mainstream Young Singles/Couples group is more inclined to buy the 270g chip packet compared to others.

Figure 1.12: Most Preferred Product Weight Ratio

However, to better understand why this group prefers the 270g packet, further examination is needed regarding the specific brands that offer this packet. Specifically, it is important to identify which products are associated with the 270g packet.

Upon analysis, it is found that, in reality, the **270g packet is exclusively associated** with the **Twisties brand**. Therefore, the higher likelihood of purchasing this packet may reflect the Mainstream Young Singles/Couples group's preference for Twisties products.

CONCLUSION

1. The main sales contributors are the BUDGET - OLDER FAMILY, MAINSTREAM - YOUNG SINGLES/COUPLES, and MAINSTREAM - RETIREES customer segments.

- For the MAINSTREAM - YOUNG SINGLES/COUPLES and MAINSTREAM - RETIREES segments, high sales are attributed to a larger customer base.
- In contrast, for the BUDGET - OLDER FAMILY segment, high sales result from a fewer number of customers making larger purchases.

2. The MAINSTREAM - RETIREES and YOUNG SINGLES/COUPLES segments tend to be willing to spend more per packet of chips, reflecting impulsive shopping behavior.

3. After analysis, the most promising segment is the MAINSTREAM - YOUNG SINGLES/COUPLES group, showing the potential to purchase chips 23% more than the other segments.

Therefore, to enhance product portfolio performance, product managers may consider allocating TYRRELLS products and smaller packages in more visible and accessible locations at places frequently visited by YOUNG SINGLES/COUPLES shoppers to create excitement and stimulate impulse buying behavior.

Particularly, the popularity of the 270g product is solely attributed to the development of the TWISTIES brand. Hence, the high purchasing likelihood for this product reflects the brand preference of the MAINSTREAM - YOUNG SINGLES/COUPLES segment for TWISTIES products.