Problem 2

1. Importing Data and Summary

Start by loading the transaction and customer data from the provided CSV files. We want to:

1. Generate summaries of the data to get an understanding of its shape, size, and general structure.
2. Use describe() for numeric columns and value_counts() for categorical columns.
3. Check for missing values with isnull().sum().

[5]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES
0	43390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	6.0
1	43599	1	1307	348	66	CCs Nacho Cheese 175g	3	6.3
2	43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	2.9
3	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	15.0
4	43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	13.8

[6]:
```
df1['TOT_SALES'].describe()
```

[6]:
```
count    264836.000000
mean          7.304200
std           3.083226
min           1.500000
25%           5.400000
50%           7.400000
75%           9.200000
max         650.000000
Name: TOT_SALES, dtype: float64
```

:
```
df1.isnull().sum()
```

:
```
DATE              0
STORE_NBR         0
LYLTY_CARD_NBR    0
TXN_ID            0
PROD_NBR          0
PROD_NAME         0
PROD_QTY          0
TOT_SALES         0
dtype: int64
```

:
```
df2.head()
```

:

	LYLTY_CARD_NBR	LIFESTAGE	PREMIUM_CUSTOMER
0	1000	YOUNG SINGLES/COUPLES	Premium
1	1002	YOUNG SINGLES/COUPLES	Mainstream
2	1003	YOUNG FAMILIES	Budget
3	1004	OLDER SINGLES/COUPLES	Mainstream
4	1005	MIDAGE SINGLES/COUPLES	Mainstream

:
```
df2.isnull().sum()
```

:
```
LYLTY_CARD_NBR      0
LIFESTAGE           0
PREMIUM_CUSTOMER    0
dtype: int64
```

2. Data Cleaning

1. Missing data: If there are any null values in critical columns(TOTAL_SALES), decide whether to replace missing values or drop the rows, depends on the data distribution.
2. Outliers: Use boxplots/standard deviation methods to detect any extreme values in TOTAL_SALES.

[11]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES
3	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	15.0
11	43332	8	8294	8221	114	Kettle Sensations Siracha Lime 150g	5	23.0
56	43601	74	74336	73182	84	GrnWves Plus Btroot & Chilli Jam 180g	5	15.5
72	43331	96	96203	96025	7	Smiths Crinkle Original 330g	5	28.5
100	43605	130	130108	134125	2	Cobs Popd Sour Crm &Chives Chips 110g	5	19.0
...	...	...	...	...	...	...	...	...
258715	43328	194	194381	194835	102	Kettle Mozzarella Basil & Pesto 175g	4	21.6
258721	43327	200	200248	199694	3	Kettle Sensations Camembert & Fig 150g	4	18.4
258726	43332	203	203253	203360	28	Thins Potato Chips Hot & Spicy 175g	5	16.5
258729	43601	208	208205	207318	37	Smiths Thinly Swt Chli&S/Cream175G	5	15.0
258788	43599	264	264149	262909	25	Pringles SourCream Onion 134g	5	18.5

578 rows × 8 columns

3. Feature Engineering

1. Extract PACK_SIZE from the Product Name
2. Calculate various values such as 'Average prices' and 'Sales per product unit'

[16]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES	PACK_SIZE
0	43390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	6.0	175.0
1	43599	1	1307	348	66	CCs Nacho Cheese 175g	3	6.3	175.0
2	43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	2.9	170.0
3	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	7.4	175.0
4	43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	7.4	150.0

[17]:
```
df1['PACK_SIZE'].isnull().sum()
```

[17]:
```
6064
```

[18]:
```
df1['PACK_SIZE'] = df1['PACK_SIZE'].fillna(df1['PACK_SIZE'].median())
```

[19]:
```
df1['SALES_PER_UNIT'] = df1['TOT_SALES'] / df1['PROD_QTY']
df1.head()
```

[19]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES	PACK_SIZE	SALES_PER_UNIT
0	43390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	6.0	175.0	3.000000
1	43599	1	1307	348	66	CCs Nacho Cheese 175g	3	6.3	175.0	2.100000
2	43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	2.9	170.0	1.450000
3	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	7.4	175.0	1.480000
4	43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	7.4	150.0	2.466667

4. Merging Datasets

Now, once the data cleaning and processing is done on both the datasets, we'll merge the datasets on a shared key(LYLTY_CARD_NBR)

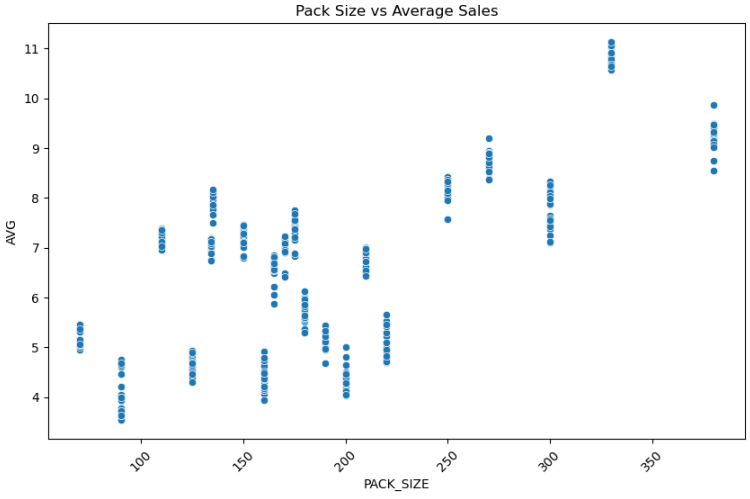
[27]:

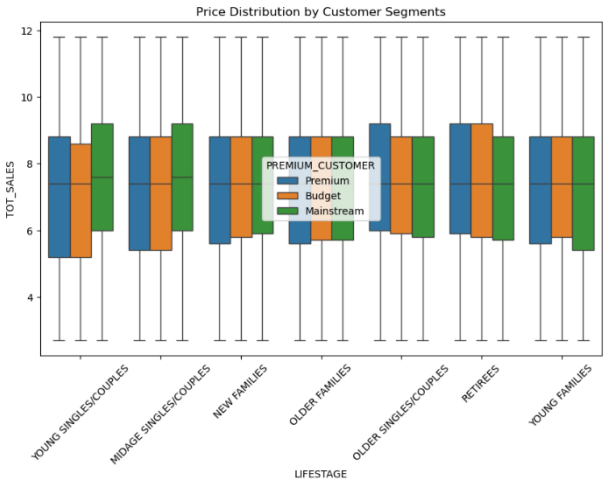
DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES	PACK_SIZE	SALES_PER_UNIT	LIFESTAGE	PREMIUM_CUSTOMI
43390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	6.0	175.0	3.000000	YOUNG SINGLES/COUPLES	Premiu
43599	1	1307	348	66	CCs Nacho Cheese 175g	3	6.3	175.0	2.100000	MIDAGE SINGLES/COUPLES	Budg
43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	2.9	170.0	1.450000	MIDAGE SINGLES/COUPLES	Budg
43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	7.4	175.0	1.480000	MIDAGE SINGLES/COUPLES	Budg
43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	7.4	150.0	2.466667	MIDAGE SINGLES/COUPLES	Budg
43604	4	4074	2982	57	Old El Paso Salsa Dip Tomato Mild 300g	1	5.1	300.0	5.100000	MIDAGE SINGLES/COUPLES	Budg
43601	4	4149	3333	16	Smiths Crinkle Chips Salt & Vinegar 330g	1	5.7	330.0	5.700000	MIDAGE SINGLES/COUPLES	Budg
43601	4	4196	3539	24	Grain Waves Sweet Chilli 210g	1	3.6	210.0	3.600000	MIDAGE SINGLES/COUPLES	Budg

5. Customer Data Analysis

[30]:

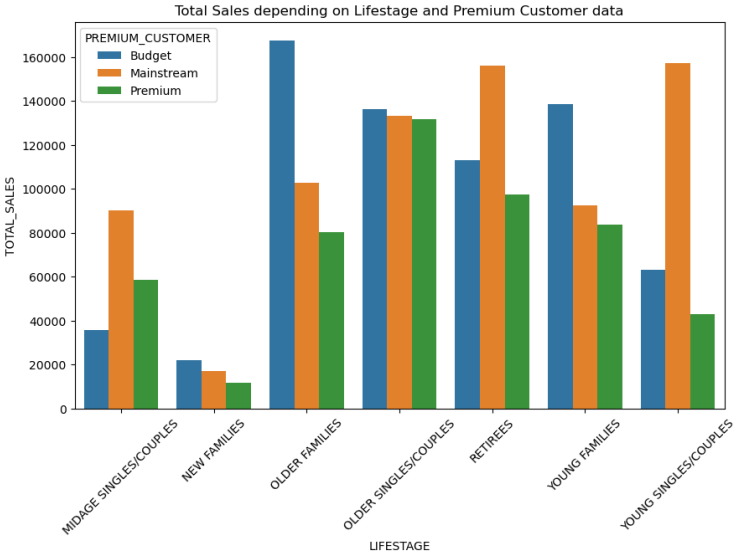
	LIFESTAGE	PREMIUM_CUSTOMER	TOTAL_SALES	AVG_SALES
0	MIDAGE SINGLES/COUPLES	Budget	35698.10	7.111175
1	MIDAGE SINGLES/COUPLES	Mainstream	90203.10	7.596690
2	MIDAGE SINGLES/COUPLES	Premium	58764.95	7.152501
3	NEW FAMILIES	Budget	22141.05	7.368070
4	NEW FAMILIES	Mainstream	17153.35	7.377785
5	NEW FAMILIES	Premium	11590.90	7.294462
6	OLDER FAMILIES	Budget	167540.85	7.234061
7	OLDER FAMILIES	Mainstream	102766.20	7.214701
8	OLDER FAMILIES	Premium	80373.70	7.181353
9	OLDER SINGLES/COUPLES	Budget	136460.40	7.413506
10	OLDER SINGLES/COUPLES	Mainstream	133350.40	7.279747
11	OLDER SINGLES/COUPLES	Premium	131759.15	7.421378
12	RETIREES	Budget	113048.40	7.436905
13	RETIREES	Mainstream	156238.95	7.278438
14	RETIREES	Premium	97287.10	7.428765
15	YOUNG FAMILIES	Budget	138694.50	7.253138
16	YOUNG FAMILIES	Mainstream	92473.85	7.164628
17	YOUNG FAMILIES	Premium	83591.60	7.229231
18	YOUNG SINGLES/COUPLES	Budget	63095.70	6.827061
19	YOUNG SINGLES/COUPLES	Mainstream	157254.40	7.540731
20	YOUNG SINGLES/COUPLES	Premium	42908.30	6.831444

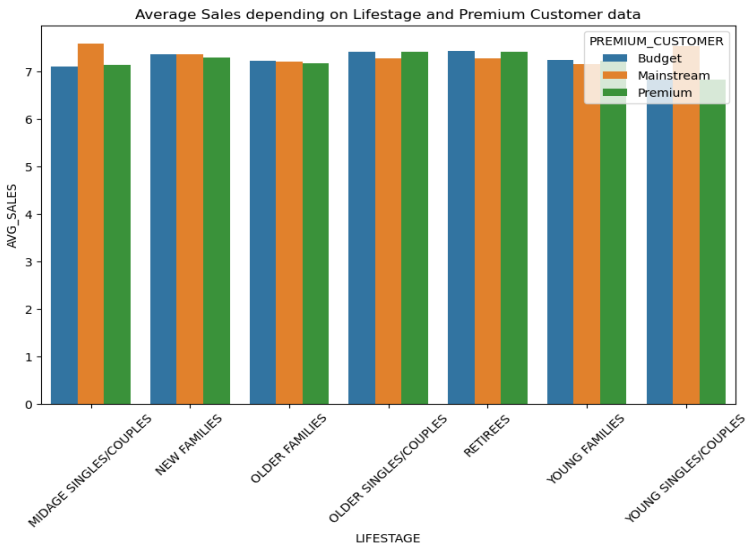




2]:

	PACK_SIZE	LIFESTAGE	PREMIUM_CUSTOMER	TOTAL	AVG
0	70.0	MIDAGE SINGLES/COUPLES	Budget	137.40	5.088889
1	70.0	MIDAGE SINGLES/COUPLES	Mainstream	250.40	5.008000
2	70.0	MIDAGE SINGLES/COUPLES	Premium	345.80	5.403125
3	70.0	NEW FAMILIES	Budget	65.00	5.000000
4	70.0	NEW FAMILIES	Mainstream	41.00	5.125000
...	...	...	...	...	...
436	380.0	YOUNG FAMILIES	Mainstream	2681.75	9.476148
437	380.0	YOUNG FAMILIES	Premium	2421.00	9.135849
438	380.0	YOUNG SINGLES/COUPLES	Budget	1601.40	8.750820
439	380.0	YOUNG SINGLES/COUPLES	Mainstream	5676.90	9.068530
440	380.0	YOUNG SINGLES/COUPLES	Premium	1315.50	9.010274





Objective:

Provide a data-driven strategic recommendation for an upcoming category review by analyzing chip purchasing behaviors across different customer

Overview:

We worked with two datasets:

1. Transaction Data: Contains sales transactions including product names, quantities, total sales, and other details.
2. Customer Data: Contains customer information like lifestage and premium customer status.

Data Cleaning:

1. Handled outliers using IQR method.
2. Addressed missing values by filling with medians where appropriate.
3. Extracted additional features such as PACK_SIZE from product names and computed SALES_PER_UNIT.

Feature Engineering:

1. Created metrics like SALES_PER_UNIT to better understand per-product performance.
2. Merged transaction data with customer information for enhanced analysis.

Customer Data Analysis: Grouped customers by LIFESTAGE and PREMIUM_CUSTOMER to examine total and average sales.

Pack Size and Sales Correlation: Analyzed the relationship between chip pack sizes and total/average sales.

Visualization:

1. Bar plots of total and average sales by customer segment and premium status.
2. Box plots highlighting price distribution across customer segments.
3. Scatter plots of packet sizes across the average sales.

Key Results:

1. Older Families and Retirees in both mainstream and premium categories showed high total sales.
2. Young Singles/Couples exhibited lower average sales compared to other categories.
3. Sales were relatively affected by the size of chip packs.

Conclusion:

Targeting Older Families and Retirees in marketing campaigns may lead to higher returns due to their demonstrated purchasing power. Additional promotional efforts for Young Singles/Couples could help boost sales in that category.