In [ ]:
```python
# 1. Import libraries
# ----------------------
import pandas as pd
import re


# 2. Load data
# ----------------------
transactions = pd.read_excel("/Users/mykytaloiko/Downloads/Quantum/QVI_transaction_data.xlsx")
purchase_behaviour = pd.read_csv("/Users/mykytaloiko/Downloads/Quantum/QVI_purchase_behaviour.csv")

# Merge transactions with customer profiles
data = transactions.merge(purchase_behaviour, on="LYLTY_CARD_NBR", how="inner")


# 3. Data cleaning
# ----------------------
# Convert DATE column from Excel serial to datetime
data['DATE'] = pd.to_datetime(data['DATE'], origin='1899-12-30', unit='D')

# Keep only chip products
chip_keywords = ['chips', 'chip', 'chp']
def is_chip_product(name):
    words = re.findall(r"[A-Za-z]+", name.lower())
    return any(word in chip_keywords for word in words)

data = data[data['PROD_NAME'].apply(is_chip_product)]

# Remove customer with abnormal purchases (200 packs in single transaction)
data = data[data['LYLTY_CARD_NBR'] != 226000]



# 4. Feature Engineering
# ----------------------
# Extract pack size from product name
data['PACK_SIZE'] = data['PROD_NAME'].str.extract(r'(\d+)\s?g', expand=False)
data['PACK_SIZE'] = pd.to_numeric(data['PACK_SIZE'], errors='coerce')


# Extract brand name (first word of product name)
data['BRAND'] = data['PROD_NAME'].str.split().str[0]

# Standardize brand names
data['BRAND'] = data['BRAND'].replace({
    'RED': 'RRD',
    'Red': 'RRD',
    'Snbts': 'Sunbites',
    'Smith': 'Smiths',
    'WW': 'Woolworths',
    'NCC': 'Natural',
    'Dorito': 'Doritos',
    'Infzns': 'Infuzions',
    'SUNBITIES': 'Sunbites'
})


# 5. Customer analysis
# ----------------------
# Number of customers by lifestage
print(data['LIFESTAGE'].value_counts())

# Spending metrics and average units by lifestage and premium status
segment_metrics = data.groupby(["LIFESTAGE", "PREMIUM_CUSTOMER"]).agg(
    total_sales=("TOT_SALES", "sum"),
    avg_sales=("TOT_SALES", "mean"),
    transaction_count=("TOT_SALES", "count"),
    total_units=("PROD_QTY", "sum"),
    unique_customers=("LYLTY_CARD_NBR", "nunique")
).reset_index()

# Add metric: average number of units per customer
segment_metrics["avg_units_per_customer"] = (
    segment_metrics["total_units"] / segment_metrics["unique_customers"]
)

print(segment_metrics)


# 6. Export cleaned dataset with segment metrics
# ----------------------
data.to_csv("data_cleaned.csv", index=False)
segment_metrics.to_csv("segment_metrics.csv", index=False)
```