```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

```
transactions = pd.read_excel("QVI_transaction_data.xlsx")
customers = pd.read_csv("QVI_purchase_behaviour.csv", encoding='latin1')
```

```
transactions.head()
transactions.info()
customers.head()
customers.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 264836 entries, 0 to 264835
Data columns (total 8 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   DATE            264836 non-null  int64  
 1   STORE_NBR       264836 non-null  int64  
 2   LYLTY_CARD_NBR  264836 non-null  int64  
 3   TXN_ID          264836 non-null  int64  
 4   PROD_NBR        264836 non-null  int64  
 5   PROD_NAME       264836 non-null  object 
 6   PROD_QTY        264836 non-null  int64  
 7   TOT_SALES       264836 non-null  float64
dtypes: float64(1), int64(6), object(1)
memory usage: 16.2+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 72637 entries, 0 to 72636
Data columns (total 3 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   LYLTY_CARD_NBR    72637 non-null  int64 
 1   LIFESTAGE         72637 non-null  object
 2   PREMIUM_CUSTOMER  72637 non-null  object
dtypes: int64(1), object(2)
memory usage: 1.7+ MB
```

```
transactions.describe()
```

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_QTY	TOT_SALES
count	264836.000000	264836.00000	2.648360e+05	2.648360e+05	264836.000000	264836.000000	264836.000000
mean	43464.036260	135.08011	1.355495e+05	1.351583e+05	56.583157	1.907309	7.304200
std	105.389282	76.78418	8.057998e+04	7.813303e+04	32.826638	0.643654	3.083226
min	43282.000000	1.00000	1.000000e+03	1.000000e+00	1.000000	1.000000	1.500000
25%	43373.000000	70.00000	7.002100e+04	6.760150e+04	28.000000	2.000000	5.400000
50%	43464.000000	130.00000	1.303575e+05	1.351375e+05	56.000000	2.000000	7.400000
75%	43555.000000	203.00000	2.030942e+05	2.027012e+05	85.000000	2.000000	9.200000
max	43646.000000	272.00000	2.373711e+06	2.415841e+06	114.000000	200.000000	650.000000

```
customers['LIFESTAGE'].value_counts()
customers['PREMIUM_CUSTOMER'].value_counts()
```

	count
PREMIUM_CUSTOMER	
Mainstream	29245
Budget	24470
Premium	18922

dtype: int64

```
transactions.isnull().sum()
customers.isnull().sum()
```

```
                  0
LYLTY_CARD_NBR    0
     LIFESTAGE    0
PREMIUM_CUSTOMER  0

dtype: int64
```

```
transactions['DATE'] = pd.to_datetime(transactions['DATE'])
```

```
transactions.describe()
```

```
                              DATE     STORE_NBR  LYLTY_CARD_NBR        TXN_ID       PROD_NBR       PROD_QTY      TOT_SALES
count                       264836  264836.00000    2.648360e+05  2.648360e+05  264836.000000  264836.000000  264836.000000
 mean  1970-01-01 00:00:00.000043464     135.08011    1.355495e+05  1.351583e+05      56.583157       1.907309       7.304200
  min  1970-01-01 00:00:00.000043282       1.00000    1.000000e+03  1.000000e+00       1.000000       1.000000       1.500000
  25%  1970-01-01 00:00:00.000043373      70.00000    7.002100e+04  6.760150e+04      28.000000       2.000000       5.400000
  50%  1970-01-01 00:00:00.000043464     130.00000    1.303575e+05  1.351375e+05      56.000000       2.000000       7.400000
  75%  1970-01-01 00:00:00.000043555     203.00000    2.030942e+05  2.027012e+05      85.000000       2.000000       9.200000
  max  1970-01-01 00:00:00.000043646     272.00000    2.373711e+06  2.415841e+06     114.000000     200.000000     650.000000
  std                          NaN      76.78418    8.057998e+04  7.813303e+04      32.826638       0.643654       3.083226
```

```
transactions['PROD_QTY'].describe()
```

```
          PROD_QTY
count  264836.000000
 mean       1.907309
  std       0.643654
  min       1.000000
  25%       2.000000
  50%       2.000000
  75%       2.000000
  max     200.000000

dtype: float64
```

```
df = transactions.merge(customers, on='LYLTY_CARD_NBR', how='left')
```

```
df['TOTAL_SPEND'] = df['TOT_SALES']
```

```
segment_spend = df.groupby(
    ['LIFESTAGE', 'PREMIUM_CUSTOMER']
)['TOTAL_SPEND'].sum().reset_index()
```

```
avg_spend = df.groupby(
    ['LIFESTAGE', 'PREMIUM_CUSTOMER']
)['TOTAL_SPEND'].mean().reset_index()
```

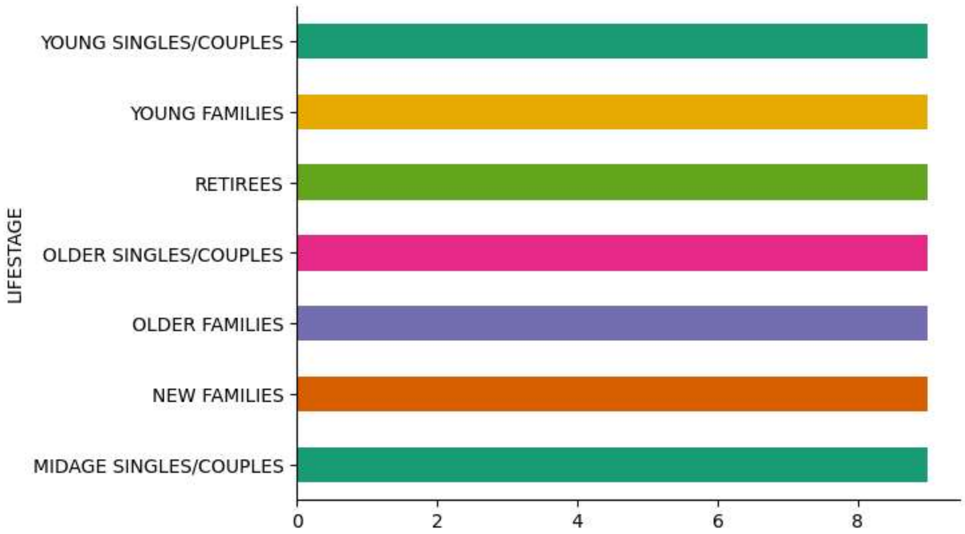
```
df['PACK_SIZE'] = df['PROD_NAME'].str.extract(r'(\d+)').astype(float)
pack_pref = df.groupby(
    ['LIFESTAGE', 'PREMIUM_CUSTOMER']
)['PACK_SIZE'].mean().reset_index()
```

```
df['BRAND'] = df['PROD_NAME'].str.split().str[0]
brand_pref = df.groupby(
    ['LIFESTAGE', 'PREMIUM_CUSTOMER', 'BRAND']
)['TOTAL_SPEND'].sum().reset_index()
```

```
brand_pref.sort_values('TOTAL_SPEND', ascending=False).groupby(
    ['LIFESTAGE', 'PREMIUM_CUSTOMER']
).head(3)
```

Show hidden output

```
from matplotlib import pyplot as plt
import seaborn as sns
_df_1.groupby('LIFESTAGE').size().plot(kind='barh', color=sns.palettes.mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
```


YOUNG SINGLES/COUPLES
YOUNG FAMILIES
RETIREES
OLDER SINGLES/COUPLES
OLDER FAMILIES
NEW FAMILIES
MIDAGE SINGLES/COUPLES
LIFESTAGE