In [2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
Matplotlib is building the font cache; this may take a moment.
```

In [5]:
```
transactions = pd.read_csv("QVI_transaction_data.csv")
purchase_behaviour = pd.read_csv("QVI_purchase_behaviour.csv")
```

In [6]:
```
print(transactions.head())
print(purchase_behaviour.head())


print(transactions.info())
print(purchase_behaviour.info())


print(transactions.isnull().sum())
print(purchase_behaviour.isnull().sum())
```

```
    DATE  STORE_NBR  LYLTY_CARD_NBR  TXN_ID  PROD_NBR  \
0  43390          1            1000       1         5
1  43599          1            1307     348        66
2  43605          1            1343     383        61
3  43329          2            2373     974        69
4  43330          2            2426    1038       108

                                PROD_NAME  PROD_QTY  TOT_SALES
0    Natural Chip        Compny SeaSalt175g         2        6.0
1                  CCs Nacho Cheese    175g         3        6.3
2    Smiths Crinkle Cut  Chips Chicken 170g         2        2.9
3    Smiths Chip Thinly  S/Cream&Onion 175g         5       15.0
4  Kettle Tortilla ChpsHny&Jlpno Chili 150g         3       13.8
   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
<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
None
<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
None
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
LYLTY_CARD_NBR      0
LIFESTAGE           0
PREMIUM_CUSTOMER    0
dtype: int64
```

In [7]:
```
transactions['PACK_SIZE'] = transactions['PROD_NAME'].str.extract(r'(\d+)\s*G',


transactions['BRAND'] = transactions['PROD_NAME'].str.split().str[0]
```

In [10]:
```
merged = pd.merge(transactions, purchase_behaviour, how='left', on='LYLTY_CARD_N
```

In [12]:
```
merged = merged[merged['TOT_SALES'] < 100]


merged = merged[merged['PACK_SIZE'] < 500]
```

In [13]:
```
merged['TOTAL_SPEND'] = merged['PROD_QTY'] * merged['TOT_SALES']
```

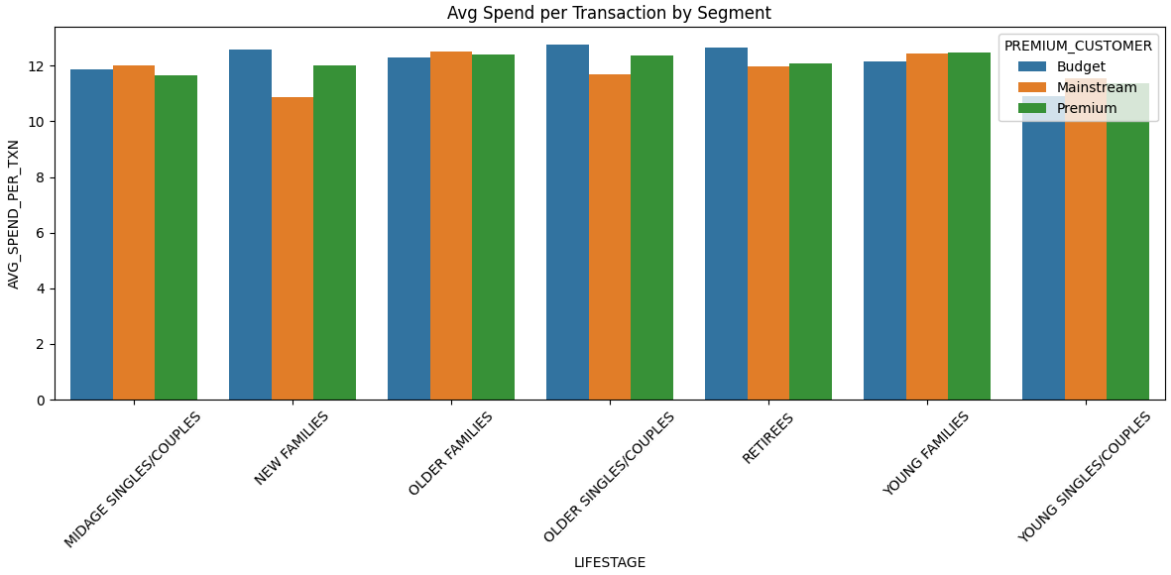
In [14]:
```
customer_summary = merged.groupby('LYLTY_CARD_NBR').agg({
    'TOTAL_SPEND': 'sum',
    'PROD_QTY': 'sum',
    'PACK_SIZE': lambda x: x.mode()[0] if not x.mode().empty else None,
    'BRAND': lambda x: x.mode()[0] if not x.mode().empty else None,
    'TXN_ID': 'count',
    'LIFESTAGE': 'first',
    'PREMIUM_CUSTOMER': 'first'
}).reset_index()

customer_summary['AVG_SPEND_PER_TXN'] = customer_summary['TOTAL_SPEND'] / custom
```

In [15]:
```
segment_analysis = merged.groupby(['LIFESTAGE', 'PREMIUM_CUSTOMER']).agg({
    'TOTAL_SPEND': 'sum',
    'PROD_QTY': 'sum',
    'TXN_ID': 'count'
}).reset_index()

segment_analysis['AVG_SPEND_PER_TXN'] = segment_analysis['TOTAL_SPEND'] / segmen
```

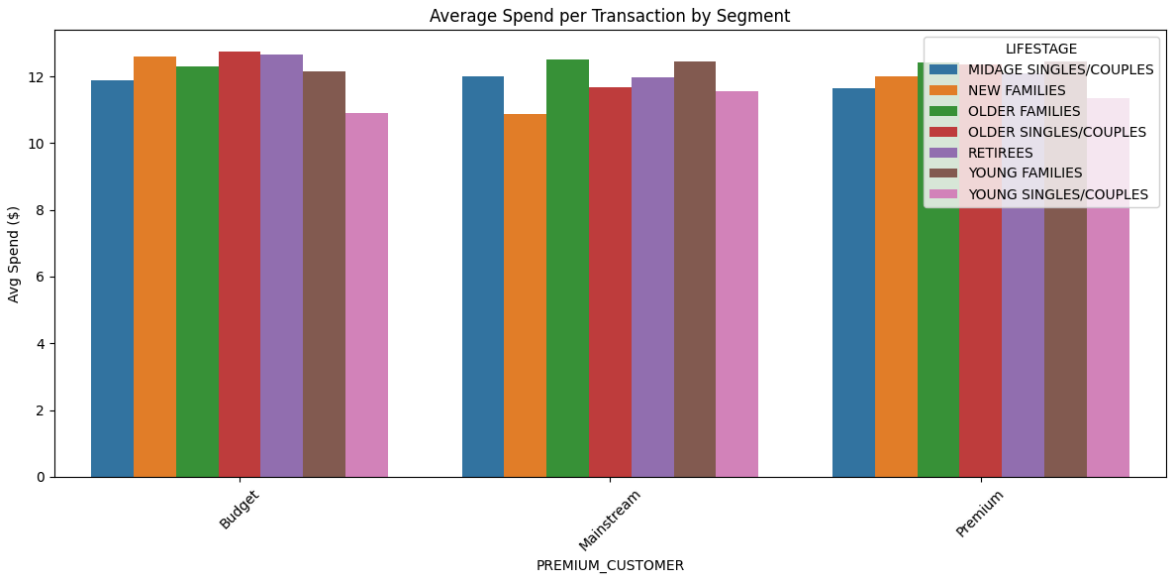
In [19]:
```
plt.figure(figsize=(12,6))
sns.barplot(data=segment_analysis, x='LIFESTAGE', y='AVG_SPEND_PER_TXN', hue='PR
plt.title("Avg Spend per Transaction by Segment")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



In [17]:
```
summary = segment_analysis.pivot(index='LIFESTAGE', columns='PREMIUM_CUSTOMER',
print(summary)
```

```
PREMIUM_CUSTOMER           Budget  Mainstream    Premium
LIFESTAGE
MIDAGE SINGLES/COUPLES  11.881818   12.004364  11.642246
NEW FAMILIES            12.595000   10.863830  12.016667
OLDER FAMILIES          12.302105   12.516578  12.407905
OLDER SINGLES/COUPLES   12.752632   11.679767  12.365468
RETIREES                12.667857   11.983190  12.090000
YOUNG FAMILIES          12.153776   12.449013  12.458955
YOUNG SINGLES/COUPLES   10.917814   11.550000  11.356291
```

In [18]:
```
plt.figure(figsize=(12, 6))
sns.barplot(data=segment_analysis, x='PREMIUM_CUSTOMER', y='AVG_SPEND_PER_TXN',
plt.title("Average Spend per Transaction by Segment")
plt.ylabel("Avg Spend ($)")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



In [ ]:
```
customer_summary.to_csv("customer_summary.csv", index=False)
segment_analysis.to_csv("segment_analysis.csv", index=False)
```