In [1]:

```
import pandas as pd

df = pd.read_csv('QVI_purchase_behaviour.csv')
df.info()
df.head()
```

```
<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
```

Out[1]:

	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

In [2]:

```
df.dtypes
```

Out[2]:

```
LYLTY_CARD_NBR       int64
LIFESTAGE           object
PREMIUM_CUSTOMER    object
dtype: object
```

In [3]:

```
#Handle Missing or Incorrect Data
# Check missing values
df.isnull().sum()

# Drop or fill missing values if any
df.dropna(inplace=True)
# OR
# df.fillna(method='ffill', inplace=True)
```

In [4]:

```
# Standardize text: strip whitespace and fix casing
df['LIFESTAGE'] = df['LIFESTAGE'].str.strip().str.title()
df['PREMIUM_CUSTOMER'] = df['PREMIUM_CUSTOMER'].str.strip().str.title()

# Convert to categorical for memory efficiency
df['LIFESTAGE'] = df['LIFESTAGE'].astype('category')
df['PREMIUM_CUSTOMER'] = df['PREMIUM_CUSTOMER'].astype('category')
```

In [5]:

```
df.duplicated(subset='LYLTY_CARD_NBR').sum()
```

Out[5]:

```
np.int64(0)
```

In [6]:

```
#Customer AnalyticsHere you start deriving insights from the cleaned data.
```

In [7]:

```
# Count by life stage
df['LIFESTAGE'].value_counts(normalize=True) * 100

# Count by premium segment
df['PREMIUM_CUSTOMER'].value_counts(normalize=True) * 100

# Crosstab to see segment overlaps
pd.crosstab(df['LIFESTAGE'], df['PREMIUM_CUSTOMER'])
```

Out[7]:

PREMIUM_CUSTOMER	Budget	Mainstream	Premium
LIFESTAGE			
Midage Singles/Couples	1504	3340	2431
New Families	1112	849	588
Older Families	4675	2831	2274
Older Singles/Couples	4929	4930	4750
Retirees	4454	6479	3872
Young Families	4017	2728	2433
Young Singles/Couples	3779	8088	2574

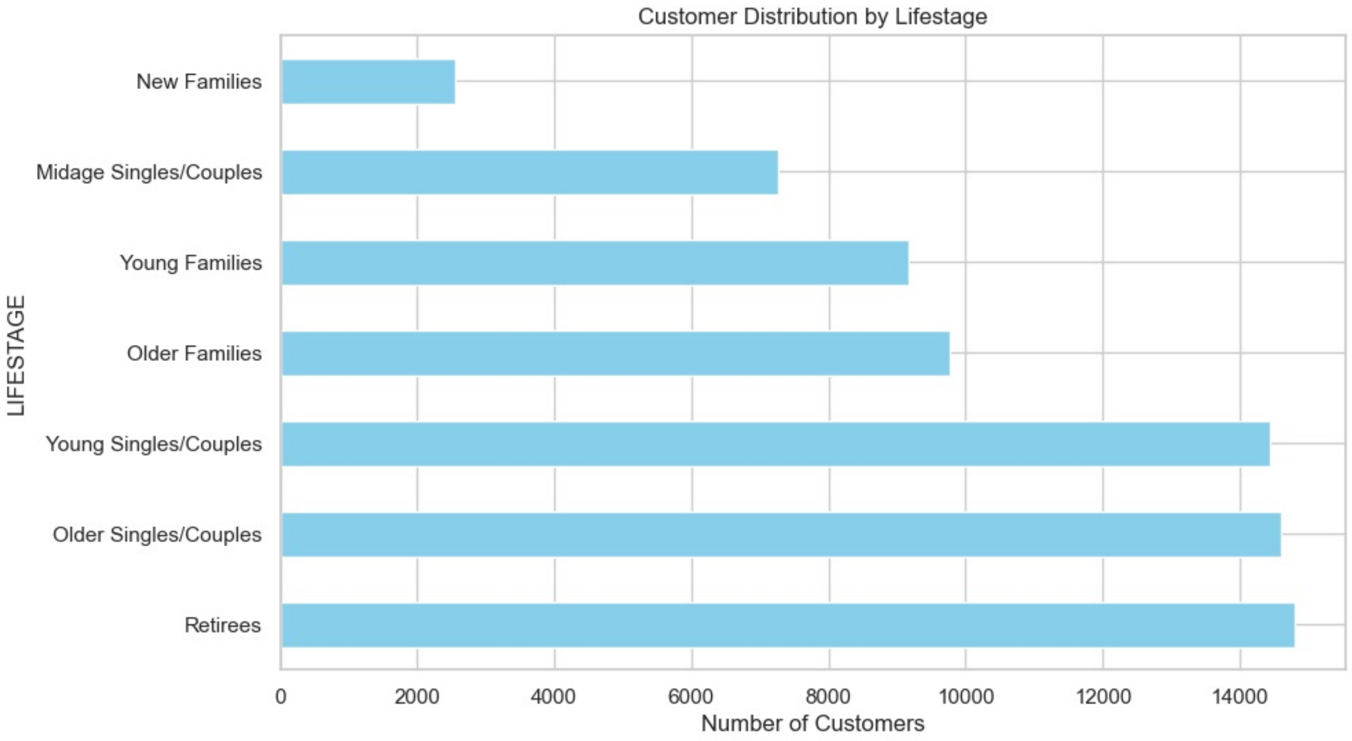
In [8]:

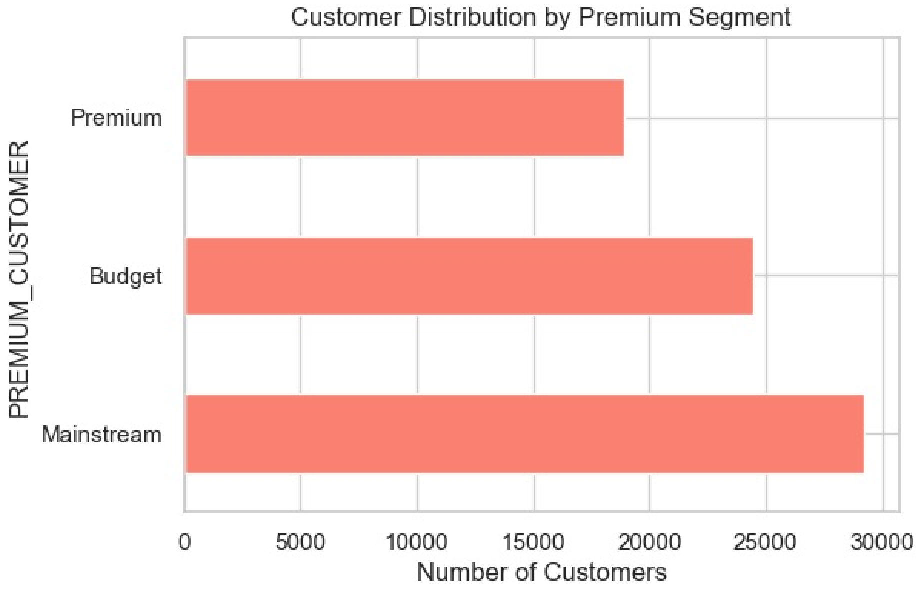
```python
import seaborn as sns
import matplotlib.pyplot as plt

sns.set(style="whitegrid")

# Lifestage Distribution
df['LIFESTAGE'].value_counts().plot(kind='barh', figsize=(10,6), color='skyblue')
plt.title('Customer Distribution by Lifestage')
plt.xlabel('Number of Customers')
plt.show()

# Premium Segment Distribution
df['PREMIUM_CUSTOMER'].value_counts().plot(kind='barh', figsize=(6,4), color='salmon')
plt.title('Customer Distribution by Premium Segment')
plt.xlabel('Number of Customers')
plt.show()
```





In [9]:
```
print(df.describe())
```

```
       LYLTY_CARD_NBR
count    7.263700e+04
mean     1.361859e+05
std      8.989293e+04
min      1.000000e+03
25%      6.620200e+04
50%      1.340400e+05
75%      2.033750e+05
max      2.373711e+06
```

In [ ]: