Stastistical Computing Final Project - Group 10

Team Members

1. Varsha Kuruva - M15128170
2. Snehal Mahajan – M14508438
3. Shankari Arulmani - M05950769
4. Chance Gough M15407540
5. Sree charan Reddy - M15235902

Introduction

Our research aims to help a large retailer better understand its customers. The goal of the study is to gain knowledge from the available database, which includes information on customers, transactions, and other topics. As part of our methodology, we intend to organize, examine, and analyze the dataset we obtained in order to finally respond to the following important questions.

In [ ]:
```
import pandas as pd                                  # This helps in analyzing the data
import seaborn as sns  # This helps in creating statistical inferences and plotting graphs
import matplotlib.pyplot as plt                       # Plotting graph library
from completejourney_py import get_data               #This provides access to data sets
import warnings                                       # To suppress warnings
import numpy as np #for numerical computing
import plotly.express as px #for tree maps
import squarify #for plotting

warnings.filterwarnings('ignore')

from completejourney_py import get_data
cj_data = get_data()

transactions=cj_data['transactions']
products=cj_data['products']
coupons=cj_data['coupons']
campaigns=cj_data['campaigns']
demographics=cj_data['demographics']
campaign_descriptions=cj_data['campaign_descriptions']
coupon_redemptions=cj_data['coupon_redemptions']
```

Business Problem

An in-depth analysis of customer shopping behavior using year-long shopping data has uncovered critical insights into business opportunities, growth factors, and inhibitors. Through exploratory data analysis of customer demographics, transaction history, coupon usage, income range, and other key factors, we aim to optimize the allocation of Regork's resources. Our analysis is structured around the following key questions:

What is working in Regork's favour? What is working agains them? What changes can be made to better their profit?

By answering these questions, we can develop a comprehensive strategy for driving growth and improving customer retention.

Approach

Analysed the demographics based on quantity and order value to briefly understand the most loyal customer base,this is to understand where to focus ‘Regork’s’ customer behavior based promotions should be and to recognize parameters that increase loyalty as they contribute the most to revenue.

Identified the areas of concern and proposed remedial measures for the same.

Proposed Solution

After conducting a cohort analysis, we have identified a highly loyal group of customers who fall within the age range of 45-54 and have an income range of $50-75k. In order to retain these customers and increase revenue for the company, we propose conducting experiments by dividing these groups into stratified samples and reducing the retail discount offered to them. Additionally, we plan to continuously improve the customer experience, product quality, and ease of shopping to increase brand value and word of mouth marketing.

Our analysis was based on various metrics such as the number of visits, and we will continue to target this cohort with new products and promotions to understand their purchase patterns, which can be applied to other cohorts to further increase growth opportunities.

Data Preparation

In our analysis, we have taken into account the fact that each demographic group has a different number of individuals. Therefore, calculating the total sales for each demographic group may not be an accurate measure. Instead, we have decided to calculate the average sales for each demographic group.

To do this, we have categorized the data based on income ranges. We have identified four income range categories: low, middle, upper-middle, and high. Customers with incomes falling into the following income range categories are included in each category:

1.Low income range category: customers with incomes falling into "Under 15K", "25-34K", and "35-49K" income ranges 2.Middle income range category: customers with incomes falling into "50-74K" and "75k-99K" income ranges 3.Upper-middle income range category: customers with incomes falling into "100-124K" and "125k-149K" income ranges 4.High income range category: customers with incomes falling into "150k-174K", "175-199K", "200-249K", and "250K+" income ranges.

By categorizing our data in this way, we can more easily understand the purchase patterns of our customers and make more informed decisions based on their income levels.

In [10]:
```
df = transactions.merge(demographics, how='inner')
df = df[df['quantity'] < 1000]

df = df[['household_id', 'basket_id', 'quantity', 'sales_value', 'household_size', 'income']]
df['income_char'] = df['income'].astype(str).str.strip()
df['income_range'] = np.select(
    [df['income_char'].isin(['Under 15K', '25-34K', '35-49K']),
     df['income_char'].isin(['50-74K', '75k-99K']),
     df['income_char'].isin(['100-124K', '125k-149K']),
     df['income_char'].isin(['150k-174K', '175-199K', '200-249K', '250K+'])],
    ['Low', 'Middle', 'Upper-Middle', 'High'])
df = df.groupby(['household_size', 'income_range']).agg(
    hsld_count=('household_id', 'nunique'),
    distinct_basket=('basket_id', 'nunique'),
    total_sales=('sales_value', 'sum')
).reset_index()
df = df.dropna(subset=['income_range'])
df['sales_per_basket'] = np.round(df['total_sales'] / df['distinct_basket'])
df['income_range'] = pd.Categorical(df['income_range'], categories=['Low', 'Middle', 'Upper-Middle', 'High'])
```

Exploratory Data Analysis

Initially, we conducted an exploratory analysis of our demographic and transaction data to gain insights about our customer base. We used tree maps to visualize the distribution of customers based on their age and income range.

To better understand our customers' purchasing behavior, we developed custom metrics, such as Average Quantity Per Basket and Average Revenue Per Basket, which we grouped by household size and income range.

We also analyzed customer loyalty using a custom metric called "Visit Number" to determine how frequently customers visited our website and placed orders. We segmented the data by different age groups and income ranges to identify our most loyal customers.

Next, we conducted a coupon analysis for the underperforming departments to evaluate the effectiveness of the coupons and determine potential improvements. Additionally, we performed a seasonality analysis on some of our important departments and used the findings to optimize the retail discounts offered.

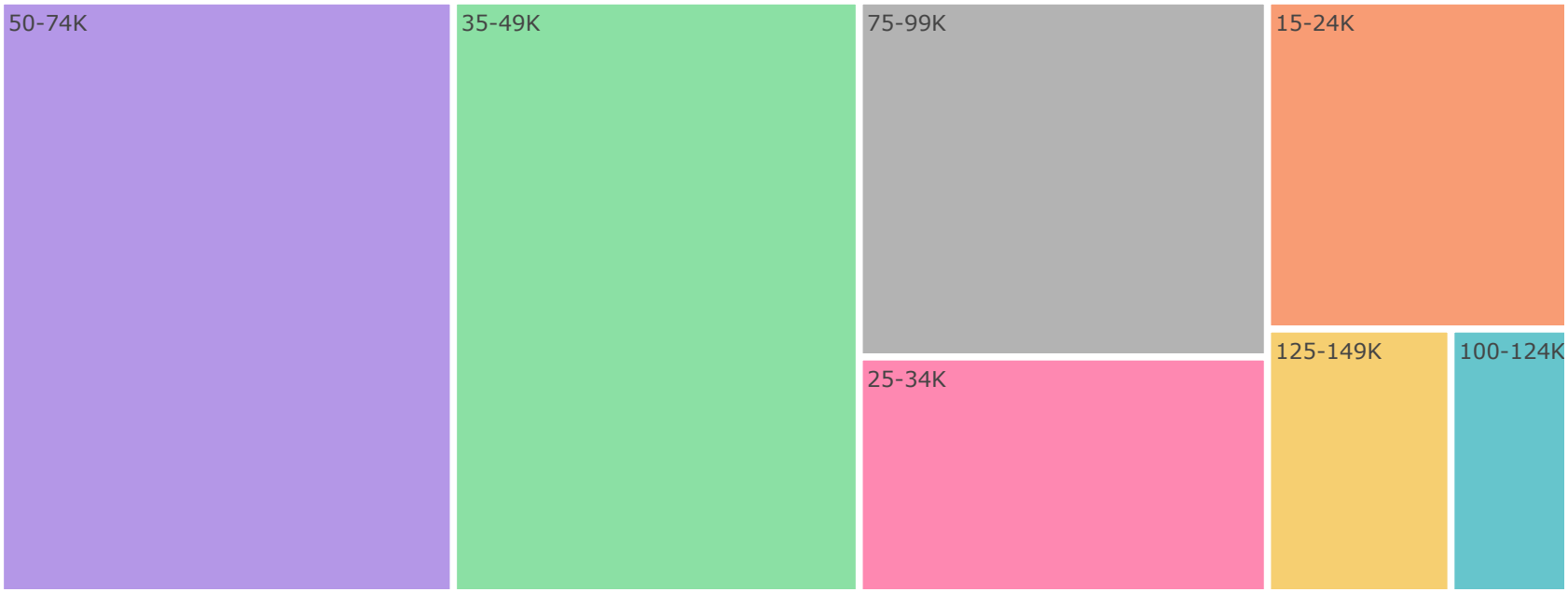
Demographic Analysis based on Customer Loyalty

The primary purpose of this analysis is to understand which customer base is loyal to 'Regork'. To identify that demographic,we are initially filtering data to plot the households by their income range.

So, the first thing we did was the plot the graphs for "Income Range" then we see where the majority of the household's belong, based on the graphs given below we see that the mass household belong to 50-74k range, followed by 35-49k bracket.

In [13]:
```
#Treemap for income range -1
import plotly.express as px



df = transactions.merge(demographics, on='household_id')
df = df.groupby(['income'])['household_id'].nunique().reset_index(name='count_hsld')

fig = px.treemap(df, path=['income'], values='count_hsld', color='income',
color_discrete_sequence=px.colors.qualitative.Pastel)

fig.update_layout(title_text="Households by Income Range")

fig.show()
```

Households by Income Range

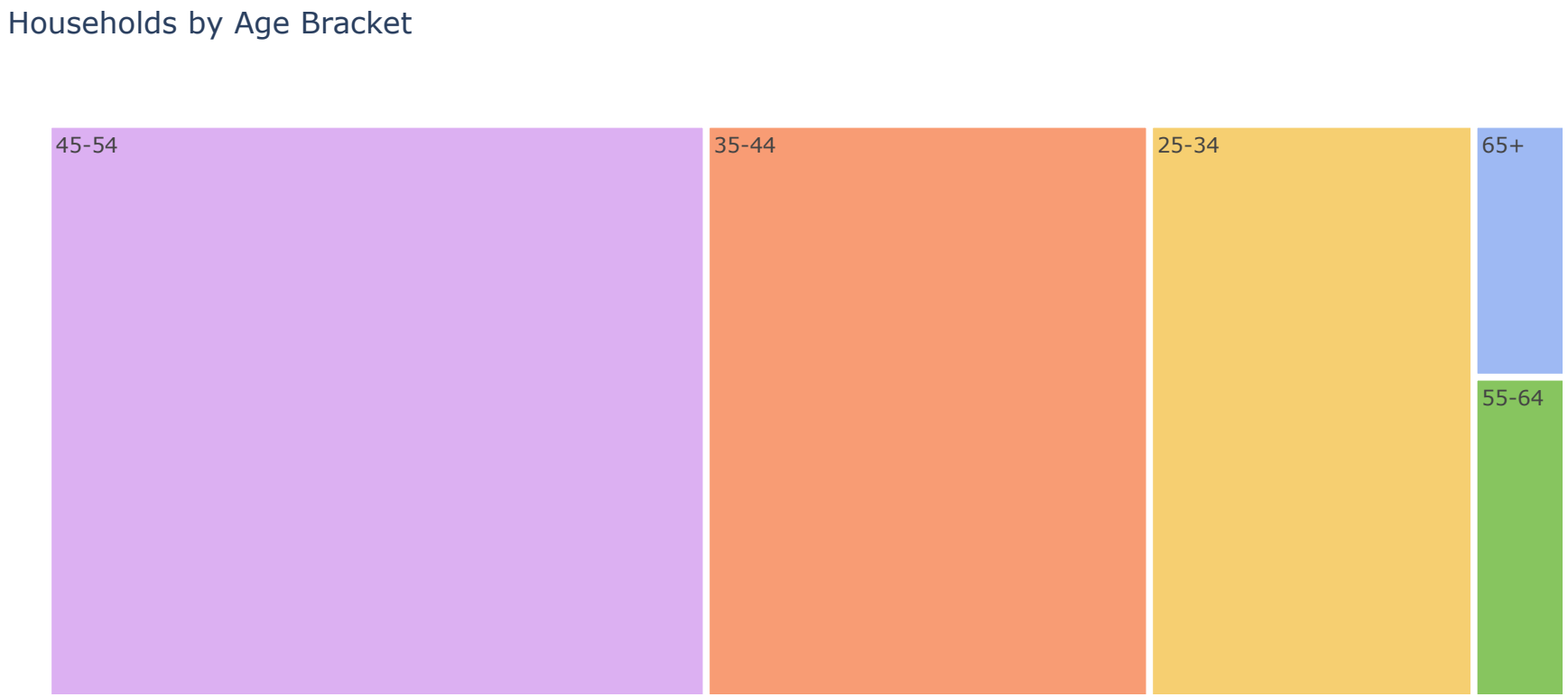


We concur from the above graph where does our customer base lie based on the household_id.

To dive deeper into our initial insight , the transaction data is then analysed by the age groups that are a part of our dataset.This will help us in our comparitive analysis further in our data exploration. Again, the agenda here is to check the distribution based on age group, we find majority of our households belong to the age group of 45-54 followed up by 35-44.

In [27]:
```
import pandas as pd
import matplotlib.pyplot as plt
import squarify


# get data
df = transactions.merge(demographics, on='household_id')
df['age_cat'] = pd.Categorical(df['age'])
```

```
df['age_code'] = df['age_cat'].cat.codes
age_counts = df.groupby('age')['household_id'].nunique().reset_index()

# create treemap
squarify.plot(sizes=age_counts['household_id'], label=age_counts['age'], color=label['age_code'], alpha=.8)
plt.axis('off')
plt.title('Households by Age Bracket')
plt.show()
```



We concur from the above graph where does our customer base lie based on the household_id.

The above two graphs plotted for income range and age gives us an idea about where our majority of population lies. Looking at our graph it is apparent that income ranges of "50-74K" and "75K-99K" are the income ranges that generate more revenue for 'Regork' - as in 'Middle' income range is the most loyal customer base.

Follow up by that we dissect the data by using a calculated metric of Average number of Quantity ordered by Household size and Average Revenue per order for household size, the hypothesis was the lower level income houses won't order more items into their basket apart from the one which is absolutely necessary and based on the graphs we plotted below we can see that. Hence, concluding the hypothesis was true

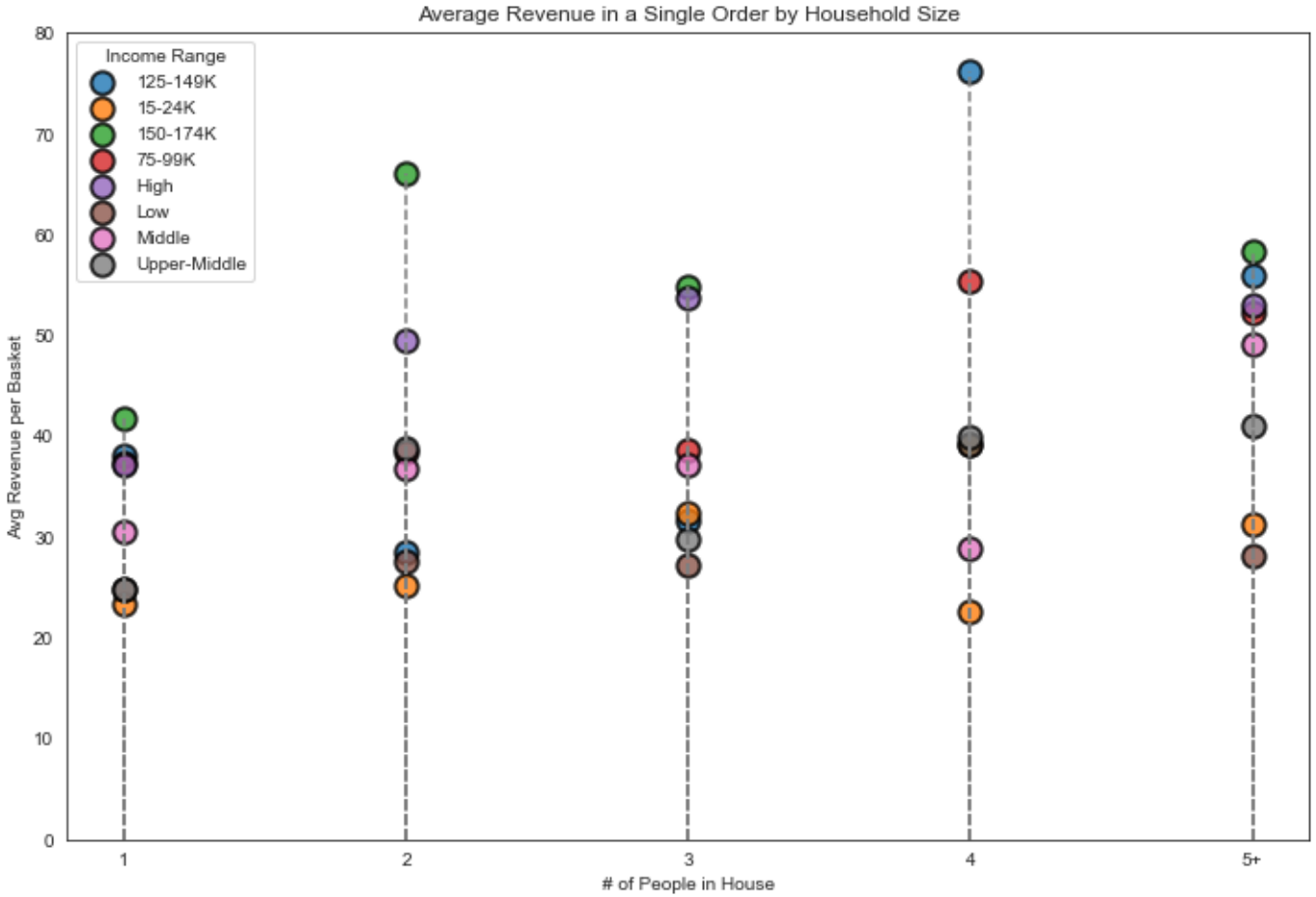
Exploring the data further we are establishing what the average order value of different customer bases is considering different income ranges as the primary parameter.

In [28]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# get data
df = pd.merge(transactions, demographics, on='household_id')
df = df[df['quantity'] < 1000]
df = df[['household_id', 'basket_id', 'quantity', 'sales_value', 'household_size', 'income']]
df['income_char'] = df['income'].astype(str).str.strip()
df['income_range'] = df['income_char'].map({
    'Under 15K': 'Low',
    '25-34K': 'Low',
    '35-49K': 'Low',
    '50-74K': 'Middle',
    '75k-99K': 'Middle',
    '100-124K': 'Upper-Middle',
    '125k-149K': 'Upper-Middle',
    '150k-174K': 'High',
    '175-199K': 'High',
    '200-249K': 'High',
    '250K+': 'High'
})
age_counts = df.groupby(['household_size', 'income_range']).agg(
    hsld_count=('household_id', pd.Series.nunique),
    distinct_basket=('basket_id', pd.Series.nunique),
    total_sales=('sales_value', sum)
).reset_index().dropna(subset=['income_range'])
age_counts['sales_per_basket'] = (age_counts['total_sales'] / age_counts['distinct_basket']).round(2)

# create plot
sns.set_style('white')
g = sns.FacetGrid(age_counts, col='income_range', height=4, aspect=0.8)
g.map(sns.scatterplot, 'household_size', 'sales_per_basket', color='red', alpha=0.5, s=50, edgecolor='grey', linewidth=1.5)
g.map(sns.lineplot, 'household_size', 'sales_per_basket', color='blue', alpha=1, linewidth=1)
g.set_axis_labels('# of People in House', 'Avg Revenue per Basket')
g.set_titles(col_template='{col_name}')
plt.subplots_adjust(top=0.9)
g.fig.suptitle('Average Revenue in a Single Order by Household Size', fontsize=14)
plt.show()
```

In [29]:

```
import pandas as pd
import matplotlib.pyplot as plt

# get data
df = transactions.merge(demographics, on='household_id')
df = df[df['quantity'] < 1000]
df = df[['household_id', 'basket_id', 'quantity', 'sales_value', 'household_size', 'income']]
df['income_char'] = df['income'].astype(str).str.strip()
df['income_range'] = df['income_char'].replace({
    'Under 15K': 'Low',
    '25-34K': 'Low',
    '35-49K': 'Low',
    '50-74K': 'Middle',
    '75k-99K': 'Middle',
    '100-124K': 'Upper-Middle',
    '125k-149K': 'Upper-Middle',
    '150k-174K': 'High',
    '175-199K': 'High',
    '200-249K': 'High',
    '250K+': 'High'
})
age_counts = df.groupby(['household_size', 'income_range'])['household_id'].nunique().reset_index()
sales_per_basket = df.groupby(['household_size', 'income_range'])['sales_value', 'basket_id'].agg({'sales_value': 'sum', 'basket_id': 'n
sales_per_basket.columns = ['total_sales', 'distinct_basket']
sales_per_basket['sales_per_basket'] = sales_per_basket['total_sales'] / sales_per_basket['distinct_basket']
sales_per_basket = sales_per_basket.reset_index()

# create lollipop plot
fig, ax = plt.subplots(figsize=(12, 8))

for i, grp in sales_per_basket.groupby('income_range'):
```

```
    ax.scatter(x='household_size', y='sales_per_basket', s=150, linewidth=2, edgecolors='black', alpha=0.8, data=grp, label=i)
    for row in grp.itertuples():
        ax.plot([row.household_size, row.household_size], [0, row.sales_per_basket], color='grey', linewidth=1.5, linestyle='--')

ax.set_ylim(bottom=0)
ax.legend(title='Income Range')
ax.set_xlabel('# of People in House')
ax.set_ylabel('Avg Revenue per Basket')
ax.set_title('Average Revenue in a Single Order by Household Size')

plt.show()
```

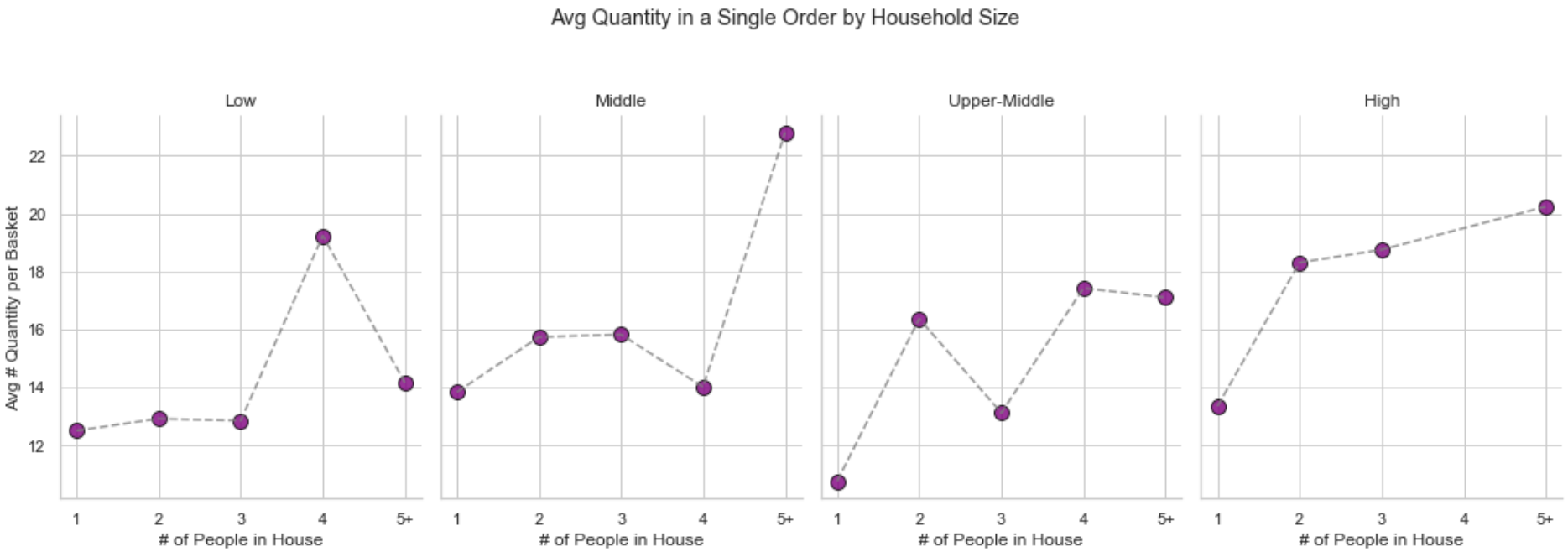


The hypothesis was for the lower income range customers usually don't spend more on unnecessary items and based on the graphs we see that this is followed the lower income household the avg revenue per basket is in the range of 20-30 and for the higher income household is higher than 40, indicating that we can target these household for the cross sell of the other items or items which are not frequently sold cause they've the purchasing power

Exploring the data further we are establishing what the average quantity per basket of different customer bases is considering their income ranges available .

In [39]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


# Merge data
df = transactions.merge(demographics, on='household_id')
df = df[df['quantity'] < 1000]
df = df[['household_id', 'basket_id', 'quantity', 'sales_value', 'household_size', 'income']]

# Clean income data
df['income_char'] = df['income'].astype(str).str.strip()
df['income_range'] = df['income_char'].replace({
    'Under 15K': 'Low',
    '25-34K': 'Low',
    '35-49K': 'Low',
    '50-74K': 'Middle',
    '75k-99K': 'Middle',
    '100-124K': 'Upper-Middle',
    '125k-149K': 'Upper-Middle',
    '150k-174K': 'High',
    '175-199K': 'High',
    '200-249K': 'High',
    '250K+': 'High'
})

# Aggregate data
agg_df = df.groupby(['household_size', 'income_range']).agg({
    'household_id': pd.Series.nunique,
    'basket_id': pd.Series.nunique,
    'quantity': sum
}).reset_index()

agg_df['qty_per_basket'] = agg_df['quantity'] / agg_df['basket_id']
agg_df = agg_df.dropna()

# Plot
sns.set_style('whitegrid')
g = sns.FacetGrid(agg_df, col='income_range', col_order=['Low', 'Middle', 'Upper-Middle', 'High'], height=5, aspect=0.75, margin_titles=
g.map(sns.scatterplot, 'household_size', 'qty_per_basket', s=100, color='purple', alpha=0.8, edgecolor='black', linewidth=1)
g.map(sns.lineplot, 'household_size', 'qty_per_basket', alpha=0.8, color='grey', linewidth=1.5, linestyle='--')
g.set_titles(col_template="{col_name}")
g.set_xlabels("# of People in House")
g.set_ylabels("Avg # Quantity per Basket")
g.fig.suptitle("Avg Quantity in a Single Order by Household Size", y=1.05)
```

```
plt.subplots_adjust(top=0.85)
plt.show()
```



Inference:

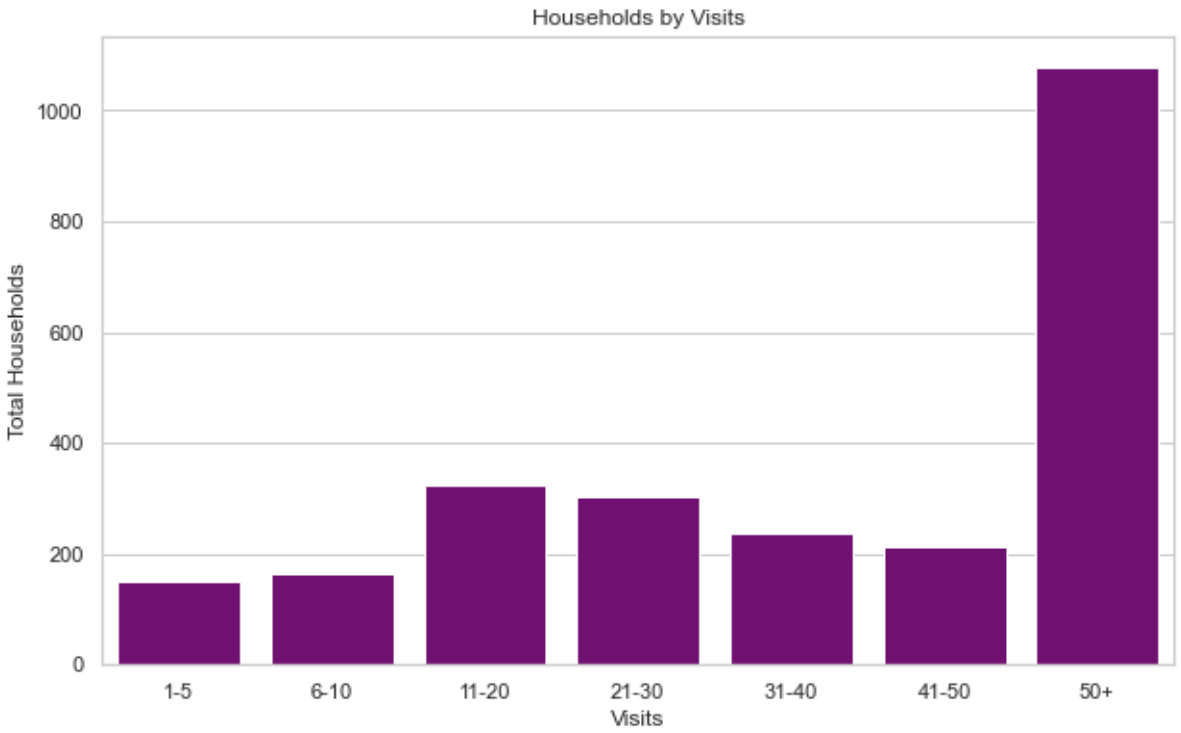
On combining the above two graphs we can say that for the lower income household who are purchasing from us, we cannot cross-sell any items and they not only buy items which are bare minimum but also the items which are cheaper else their average spend would be higher, we see that's not happening.

Diving deeper into our established insights from our data analysis:

The next analysis we went on was to identify the most loyal customer and to which cohorts do they belong. We calculated that by identifying the "Visit Number". This means that customers who frequently visit us are considered loyal and any customer whose visit number is more than 50 essentially translates to the fact that these customers are ordering once every week from us on an average and they can be further experimented with.

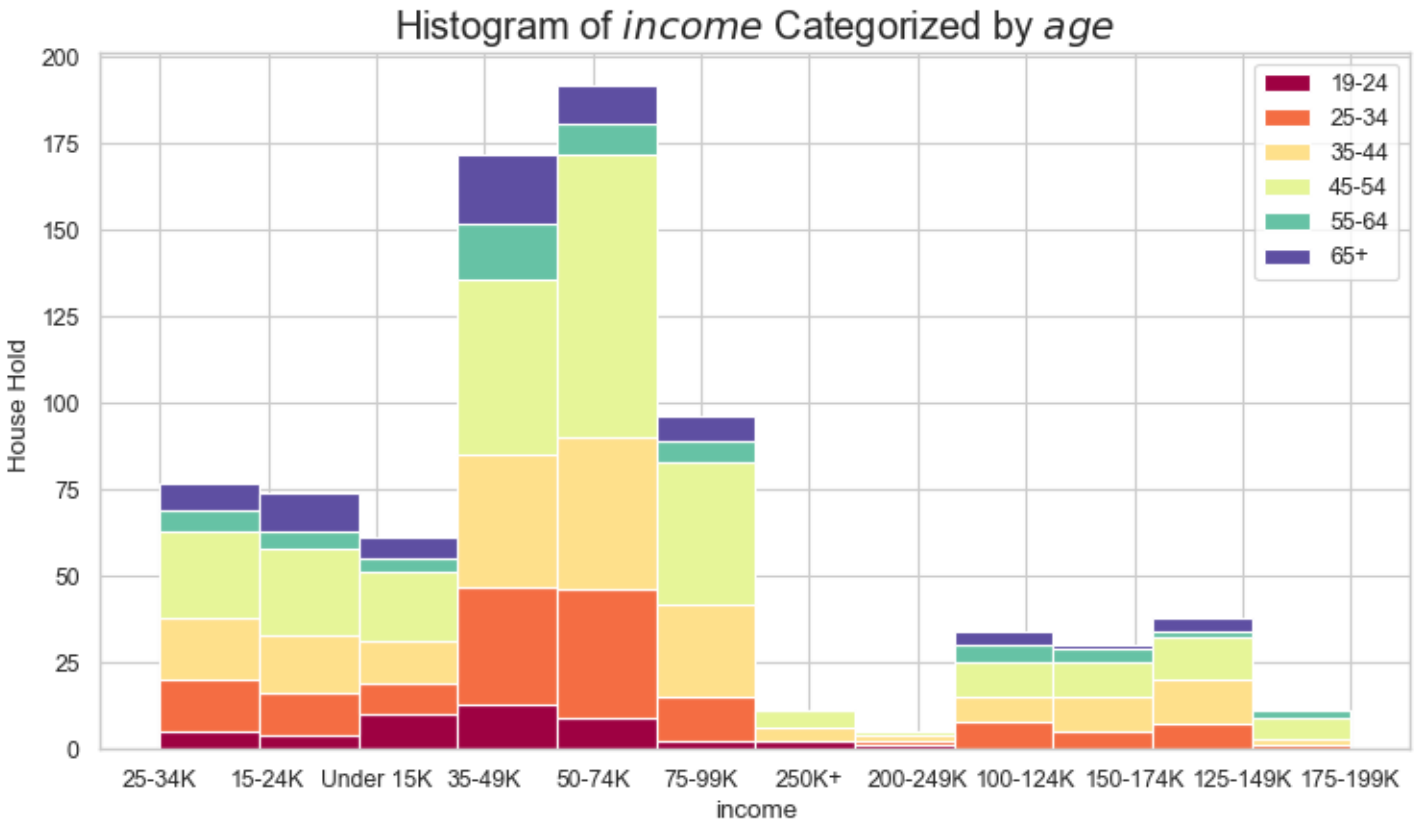
In [45]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns



visits_data = (transactions[['household_id', 'basket_id', 'quantity', 'sales_value']]
               .groupby('household_id')
```

```
                .agg(unique_orders=('basket_id', 'nunique'),
                     total_qty=('quantity', 'sum'),
                     total_sales=('sales_value', 'sum'))
                .reset_index())

visits_counts = (visits_data.groupby('unique_orders')
                 .agg(total_hsld=('household_id', 'nunique'))
                 .reset_index())

visits_counts['No_of_Visits'] = pd.cut(visits_counts['unique_orders'],
                                       bins=[0, 5, 10, 20, 30, 40, 50, float('inf')],
                                       labels=["1-5", "6-10", "11-20", "21-30", "31-40", "41-50", "50+"])

visits_counts_by_group = (visits_counts.groupby('No_of_Visits')
                          .agg(total_hsld=('total_hsld', 'sum'))
                          .reset_index())

plt.figure(figsize=(10, 6))
sns.barplot(x='No_of_Visits', y='total_hsld', data=visits_counts_by_group, color='purple')
plt.title('Households by Visits')
plt.xlabel('Visits')
plt.ylabel('Total Households')
plt.show()
```



So based on the graph plotted we see that most of the customers belong to the category of 50+ visits and our next step of the analysis is to find out the demographic information for the same.

In [46]:
```
df = demographics
x_var = 'income'
groupby_var = 'age'
df_agg = df.loc[:, [x_var, groupby_var]].groupby(groupby_var)
vals = [df[x_var].values.tolist() for i, df in df_agg]
plt.figure(figsize=(11, 6), dpi= 80)
colors = [plt.cm.Spectral(i/float(len(vals)-1)) for i in range(len(vals))]
n, bins, patches = plt.hist(vals, df[x_var].unique().__len__(), stacked=True, density=False, color=colors[:len(vals)])

plt.legend({group:col for group, col in zip(np.unique(df[groupby_var]).tolist(), colors[:len(vals)])})
plt.title(f"Histogram of ${x_var}$ Categorized by ${groupby_var}$", fontsize=18)
plt.xlabel(x_var)
plt.ylabel("House Hold ")
plt.show()
```



Based on the above graph, we can identify the age and income range of our most loyal customers. This group presents an opportunity to conduct various experiments by dividing them into test and control groups. 1.we could experiment with reducing the usage of discounts or coupons for this group and measuring the impact. Additionally, if we introduce a new category of products, we could offer additional discounts to this group to measure their sales of the new products. 2.Another option is to eliminate discounts for this group in the categories in which they regularly order, which could potentially increase revenue.

What is inhibiting our growth?

Campaign and Coupon Analysis

In [47]:
```
df4= (
    transactions
    .merge(products, how='inner', on='product_id')
    .query("department != 'FUEL' & department != 'MISCELLANEOUS'")
    .groupby('department', as_index=False)
    .agg({'sales_value':sum,'coupon_disc':sum,'retail_disc':sum,'coupon_match_disc':sum})
    .sort_values('sales_value', ascending=False) # set ascending to False
)

df4new=df4.head(10)
df4new
```

Out[47]:

	department	sales_value	coupon_disc	retail_disc	coupon_match_disc
12	GROCERY	2316393.89	14051.45	453642.78	3866.59
6	DRUG GM	596827.45	4159.05	58778.14	315.47
20	PRODUCE	322858.82	274.54	41984.71	76.80
13	MEAT	308575.33	54.10	102859.16	6.70
14	MEAT-PCKGD	232282.53	701.77	69009.01	209.60
5	DELI	148344.06	102.93	13761.45	16.35
16	PASTRY	69116.68	31.12	9633.67	1.35
15	NUTRITION	57261.22	110.82	7229.68	30.39
24	SEAFOOD-PCKGD	35977.46	136.22	12074.49	11.00
8	FLORAL	22303.18	62.93	675.28	0.00

In [48]:
```
bottom_sales= (
    transactions
```

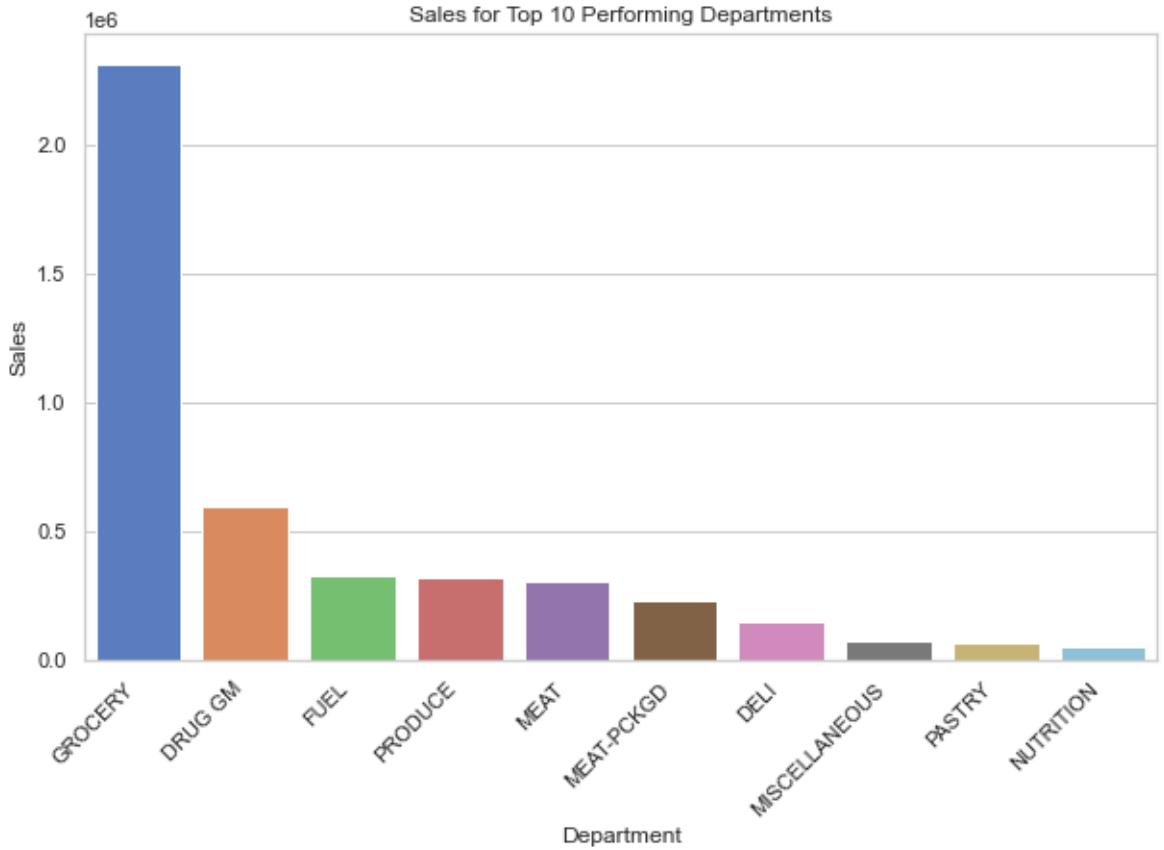
```
    .merge(products, how='inner', on='product_id')
    .query("department != 'FUEL' & department != 'MISCELLANEOUS'")
    .groupby('department', as_index=False)
    .agg({'sales_value':sum,'coupon_disc':sum,'retail_disc':sum,'coupon_match_disc':sum})
    .sort_values('sales_value', ascending=True)
)

bottom_sales=bottom_sales.head(10)
bottom_sales
```

Out[48]:

	department	sales_value	coupon_disc	retail_disc	coupon_match_disc
7	ELECT &PLUMBING	1.00	0.00	0.00	0.00
19	PROD-WHS SALES	2.52	0.00	0.00	0.00
26	TOYS	4.17	0.00	3.20	0.00
18	POSTAL CENTER	6.57	0.00	0.39	0.00
2	CNTRL/STORE SUP	44.95	0.00	0.00	0.00
17	PHOTO & VIDEO	51.70	1.75	8.66	0.75
11	GM MERCH EXP	67.02	0.00	0.00	0.00
0	AUTOMOTIVE	301.96	0.00	0.00	0.00
9	FROZEN GROCERY	419.73	0.00	22.65	0.00
4	COUPON	652.51	16.62	139.35	3.75

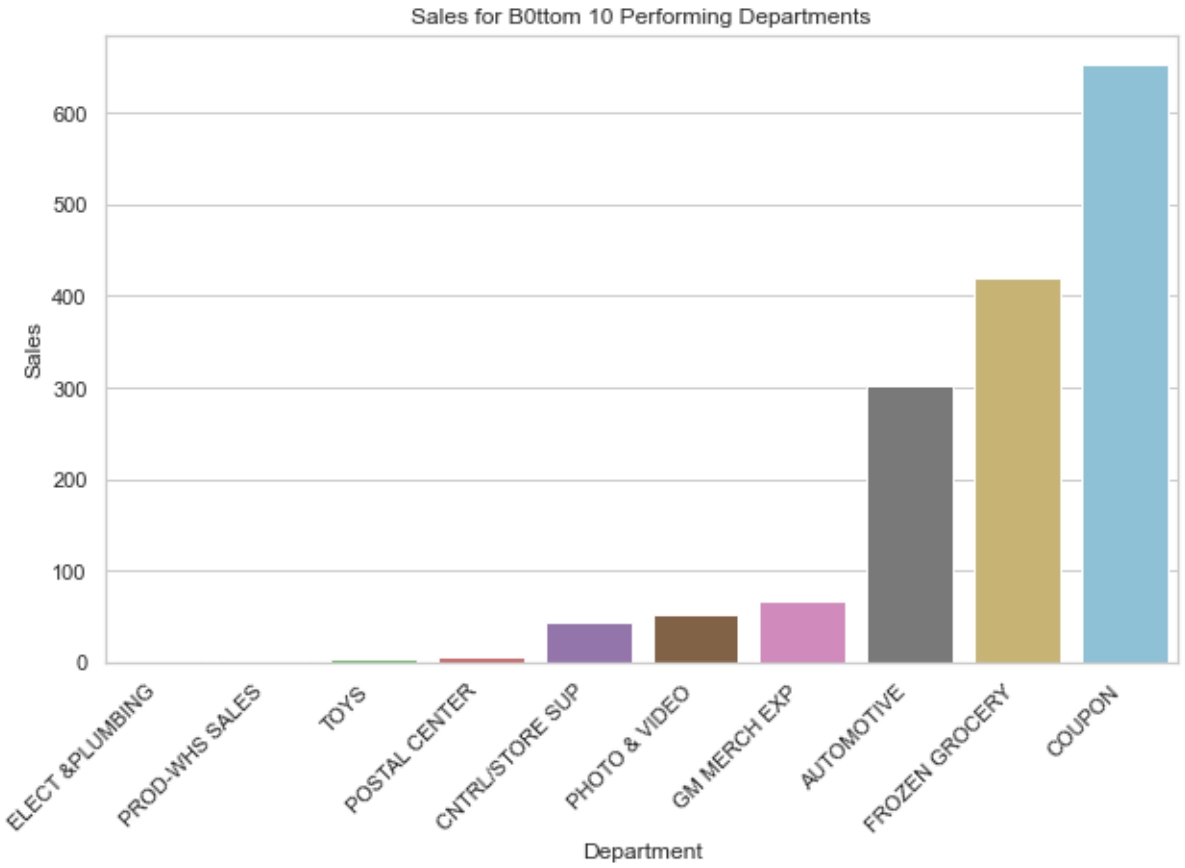
In [49]:

```
top_10_departments = (
    transactions
    .merge(products, on='product_id', how='inner')
    .groupby('department', as_index=False)
    .agg(dep_sales=('sales_value', 'sum')) # renamed column to dep_sales
    .nlargest(10, 'dep_sales')
    [['department', 'dep_sales']]
)

import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10,6))
sns.barplot(x='department', y='dep_sales', data=top_10_departments, palette='muted')
plt.ylabel('Sales')
plt.xlabel('Department')
plt.title('Sales for Top 10 Performing Departments')
plt.xticks(rotation=45, ha='right')
plt.show()
```



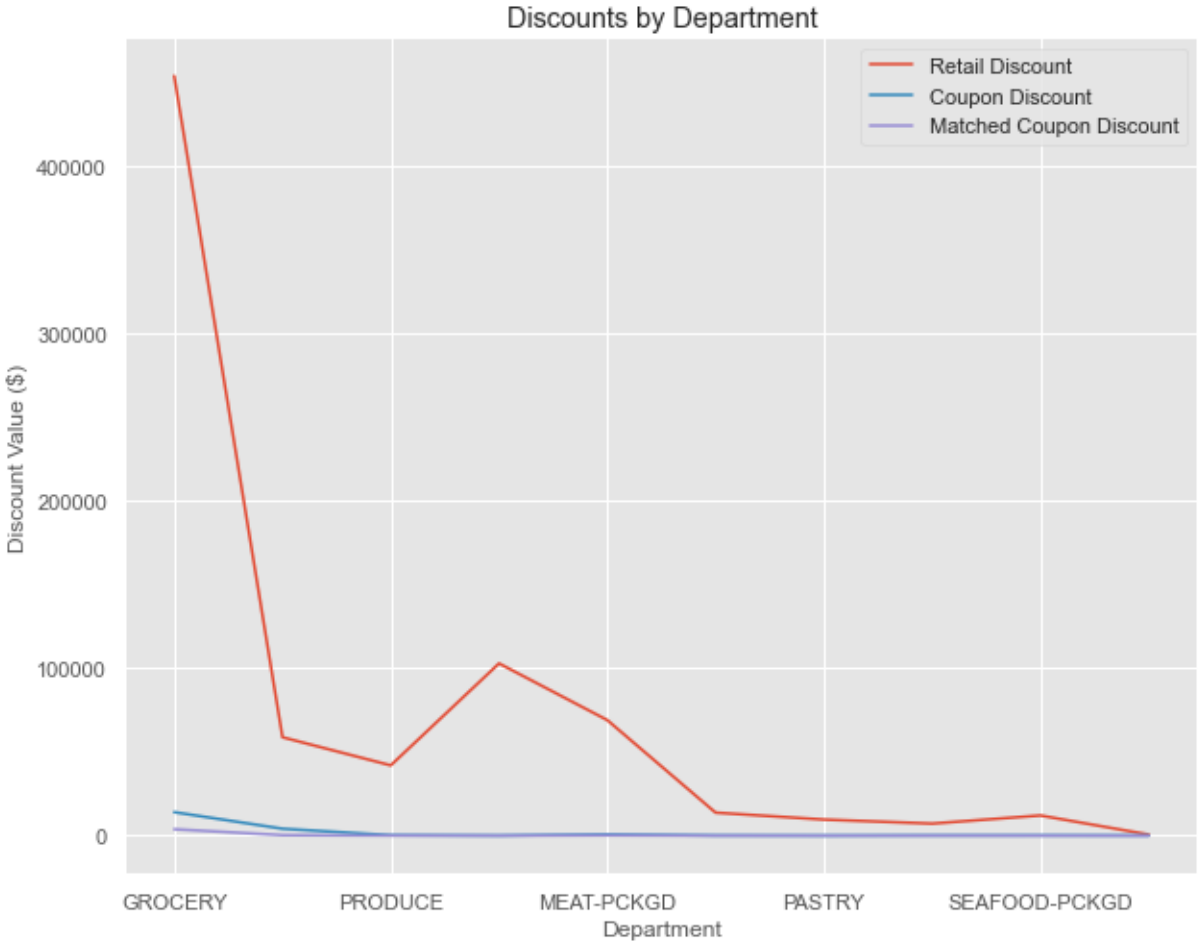
In [50]:

```
bottom_10_departments = (
    transactions
    .merge(products, on='product_id', how='inner')
    .groupby('department', as_index=False)
    .agg(dep_sales=('sales_value', 'sum')) # renamed column to dep_sales
    .nsmallest(10, 'dep_sales')
    [['department', 'dep_sales']]
)
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10,6))
sns.barplot(x='department', y='dep_sales', data=bottom_10_departments, palette='muted')
plt.ylabel('Sales')
plt.xlabel('Department')
plt.title('Sales for B0ttom 10 Performing Departments')
plt.xticks(rotation=45, ha='right')
plt.show()
```



In [51]:

```
import matplotlib.pyplot as plt

# Set plot style
plt.style.use('ggplot')

# Create plot
ax = df4new.plot.line(x='department', y=['retail_disc', 'coupon_disc','coupon_match_disc'], figsize=(10, 8))

# Set plot title and axis labels
ax.set_title('Discounts by Department')
ax.set_xlabel('Department')
ax.set_ylabel('Discount Value ($)')

# Set legend location and labels
ax.legend(loc='best', labels=['Retail Discount', 'Coupon Discount', 'Matched Coupon Discount'])

# Show plot
plt.show()
```



The above line graph assesses the top five department categories based on sales value. As we know from our prior graph grocery is the largest category in regards to sales value and now we know that this is an area where the retailer is providing the greatest amount of discounts. The retailer could benefit from issuing fewer discounts in these categories to create greater value as the goods which make up groceries, drugs, produce, meat, etc. are goods which are necessities and will be purchased regardless of coupon related price shifts. Retail discounts are the greater of the discounts, thus lowering this could increase profit.

Conclusion

The primary goal of this analysis is to establish a comprehensive framework that takes into account both internal and external factors driving and inhibiting growth. The analysis focuses on answering the following questions: Who are the Regork customers? How can we retain the customers that are driving growth? To achieve this, we aim to identify the demographics of loyal customers. Based on the graph below, we can observe that customers with income ranges of $35 - 49kand$50-75k show high levels of loyalty. Moreover, customers in the age group of 45-54 are among the most loyal customers.

Recommendations and Inferences

1. Our loyal customer base has been established as the age group of 45-54, if we want to increase our loyal customer base, apart from focusing on this demographic alone, we need to look at ways to cater to the other age groups.

2. Most of our top performing departments do expceptionally well, irrespective of whether or not coupons and discounts are being offered, therefore, we can avoid discounts and coupons for those departments and issue them in the least performing departments to foster growth.

3. Campaigns can be run on demographics that tend to buy less despite having substantial income sources to see if they yield better results.

Limitations

1) The primary limitation of this dataset is the limited time period for which it is valid, namely 2017,

2) There are only a few data points in the demographics table. The demographics table is only mapped to 32% of the transaction data set.

3) There is no detailed information on the current promotions, such as the percentage of discount offered, etc.