Here's a structured project plan with tasks for each analysis category. Each task includes objectives, suggested code, and potential insights to impress recruiters with a comprehensive analysis.

**1. Household/Customer Analysis**

**Objective:**

Understand customer purchasing patterns, average spending, and response to discounts. Identify loyal customers, high spenders, and discount-sensitive segments.

**Tasks:**

1. **Calculate Total and Average Spend per Household**

Code

household\_spending = transactions.groupby('household\_key')['sales\_value'].agg(['sum', 'mean']).rename(columns={'sum': 'total\_spend', 'mean': 'avg\_spend'}).reset\_index()

household\_spending = household\_spending.sort\_values(by='total\_spend', ascending=False)

* + **Insight**: Identify top-spending households and average spend, useful for targeting high-value customers.

1. **Analyze Discount Utilization by Household**

Code

household\_discounts = transactions.groupby('household\_key')['total\_discount'].sum().reset\_index()

household\_discounts = household\_discounts.sort\_values(by='total\_discount', ascending=False)

* + **Insight**: Determine households that benefit most from discounts, which can inform discount strategies.

1. **Determine Basket Size per Household**

Code

basket\_size = transactions.groupby('household\_key')['quantity'].sum().reset\_index().rename(columns={'quantity': 'total\_quantity'})

* + **Insight**: Find out if high-spending households also have larger basket sizes, helping with cross-sell and up-sell strategies.

**2. Product Analysis**

**Objective:**

Identify best-selling products, product categories that generate the most revenue, and product discount impact on sales.

**Tasks:**

1. **Identify Top-Selling Products by Quantity and Revenue**

Code

product\_sales = transactions.groupby('product\_id')['quantity', 'sales\_value'].sum().reset\_index()

product\_sales = product\_sales.sort\_values(by='sales\_value', ascending=False)

top\_products = product\_sales.head(10)

* + **Insight**: Recognize popular products and prioritize stock or promotions for these items.

1. **Analyze Discount Impact on Product Sales**

Code

product\_discount\_impact = transactions.groupby('product\_id')['total\_discount', 'sales\_value'].sum().reset\_index()

product\_discount\_impact['discount\_to\_sales\_ratio'] = product\_discount\_impact['total\_discount'] / product\_discount\_impact['sales\_value']

* + **Insight**: Determine which products are highly discount-driven, guiding discount allocation.

1. **Link with Product Details to Analyze Categories**

Code

product\_details = pd.merge(product\_sales, products[['product\_id', 'department', 'brand']], on='product\_id', how='left')

category\_sales = product\_details.groupby('department')['sales\_value'].sum().reset\_index().sort\_values(by='sales\_value', ascending=False)

* + **Insight**: Identify top-performing product categories to inform assortment and stocking strategies.

**3. Demographics Analysis**

**Objective:**

Analyze spending patterns and discount sensitivity based on demographics like income, household size, and age group.

**Tasks:**

1. **Calculate Average Spend per Income Group**

Code

demographics\_with\_spend = pd.merge(demographics, household\_spending, on='household\_key', how='left')

income\_spend = demographics\_with\_spend.groupby('income\_desc')['total\_spend'].mean().reset\_index().sort\_values(by='total\_spend', ascending=False)

* + **Insight**: Target high-income groups with premium products and low-income groups with discount offers.

1. **Analyze Spend by Household Size**

Code

size\_spend = demographics\_with\_spend.groupby('household\_size\_desc')['total\_spend'].mean().reset\_index().sort\_values(by='total\_spend', ascending=False)

* + **Insight**: Identify household sizes that spend the most, potentially useful for product bundling strategies.

1. **Analyze Spend by Age Group**

Code

age\_spend = demographics\_with\_spend.groupby('age\_desc')['total\_spend'].mean().reset\_index().sort\_values(by='total\_spend', ascending=False)

* + **Insight**: Discover spending patterns by age, which can inform age-targeted marketing.

**4. Customer Loyalty and Retention Analysis**

**Objective:**

Assess customer loyalty and develop strategies for retention by examining repeat purchases and changes in spending over time.

**Tasks:**

1. **Identify Repeat Customers**

Code

repeat\_customers = transactions.groupby('household\_key')['basket\_id'].nunique().reset\_index().rename(columns={'basket\_id': 'total\_baskets'})

repeat\_customers = repeat\_customers[repeat\_customers['total\_baskets'] > 1]

* + **Insight**: Quantify customer retention by finding households with multiple purchases.

1. **Analyze Monthly Spend Changes**

Code

transactions['month'] = transactions['day'] // 30 # Assuming 'day' is day count from the beginning of the dataset

monthly\_spend = transactions.groupby(['household\_key', 'month'])['sales\_value'].sum().reset\_index()

spend\_changes = monthly\_spend.pivot(index='household\_key', columns='month', values='sales\_value').fillna(0)

* + **Insight**: Track changes in household spending over time to identify trends and churn risks.

**5. Discount Sensitivity Analysis**

**Objective:**

Evaluate how discounts affect customer purchasing behavior across different demographics and product categories.

**Tasks:**

1. **Calculate Discount Utilization by Income Group**

Code

income\_discount = demographics\_with\_spend.groupby('income\_desc')['total\_discount'].sum().reset\_index().sort\_values(by='total\_discount', ascending=False)

* + **Insight**: Identify which income groups are more responsive to discounts, helping to optimize promotional spending.

1. **Analyze Average Percentage Discount per Product Category**

Code

product\_discounts = pd.merge(transactions, products[['product\_id', 'department']], on='product\_id', how='left')

category\_discount = product\_discounts.groupby('department')['percentage\_discount'].mean().reset\_index().sort\_values(by='percentage\_discount', ascending=False)

* + **Insight**: Find categories with higher average discounts, potentially indicating high price sensitivity.

**6. Visualizations and Reporting**

**Objective:**

Create visualizations to communicate insights effectively. Visualize key metrics like top products, demographic spending patterns, and discount impacts.

**Tasks:**

1. **Bar Chart of Top Product Categories by Sales Value**

Code

import matplotlib.pyplot as plt

category\_sales.plot(kind='bar', x='department', y='sales\_value', title='Top Product Categories by Sales Value', color='skyblue')

plt.show()

1. **Box Plot of Sales Value Distribution by Household Size**

Code

demographics\_with\_spend.boxplot(column='total\_spend', by='household\_size\_desc', vert=False, figsize=(10, 6), color='lightcoral')

plt.title('Sales Value Distribution by Household Size')

plt.suptitle('')

plt.show()

1. **Line Plot of Monthly Spend Changes for Top Households**

Code

spend\_changes.T.plot(figsize=(12, 6), title='Monthly Spend Changes for Top Households')

plt.xlabel('Month')

plt.ylabel('Total Spend')

plt.show()

These analyses will offer valuable insights into customer and product dynamics, including identifying high-value customers, top-selling products, and the impact of discounts. With code provided for each task, this project demonstrates strong Python and data analysis skills, plus the ability to translate data into actionable business insights—perfect for impressing recruiters!