INSTACART

Grocery Delivery Service

1

Instacart, an online grocery store that operates through an app, is looking to uncover more information about their customers background and purchasing behavior to optimize their targeted marketing strategy.

Click the links to view:

Excel Report | Github Repository

Project Overview

Customer Profile Analysis

- Strategies suggestion for better customer segmentation with the right product ads
- Key Factors influencing on customer's ordering behaviors identification
- Demographic information classification for customer profile's establishment

Data

- Datasets: Orders, Orders_Products_prior, Products, Customers (created for this project only)
- Data source: Accessed from The Instacart Online Grocery Shopping Dataset on March 2022
- Investigated factors: Ordering time/day, Total spending, Age, Family Status, Income, Department preference, Purchasing price, Ordering frequency, Max. ordering number, Regions

Skills

• Python • Anaconda • Jupyter Notebook • Data wrangling • Subsetting • Data merging • Deriving new variables • Grouping data • Aggregating data • Reporting in Excel • Population flows

Python Coding & Output

Write python statements to clean prior order and product dataset

(1.) Check for mixed-type data

```
for col in df_ords_prior.columns.tolist():
    weird = (df_ords_prior[[col]].applymap(type) != df_ords_prior[[col]].iloc[0].apply(type)).any(axis = 1)
    if len (df_ords_prior[weird]) > 0:
        print (col + ' has mixed type data')
    else:
        print (col + ' is uniform')
```

```
order_id is uniform
product_id is uniform
add_to_cart_order is uniform
reordered is uniform
```

No mixed type data is found.

(2.) Check for missing value

```
df_ords_prior.isnull().sum()
```

```
order_id             0
product_id           0
add_to_cart_order    0
reordered            0
dtype: int64
```

No missing value is found.

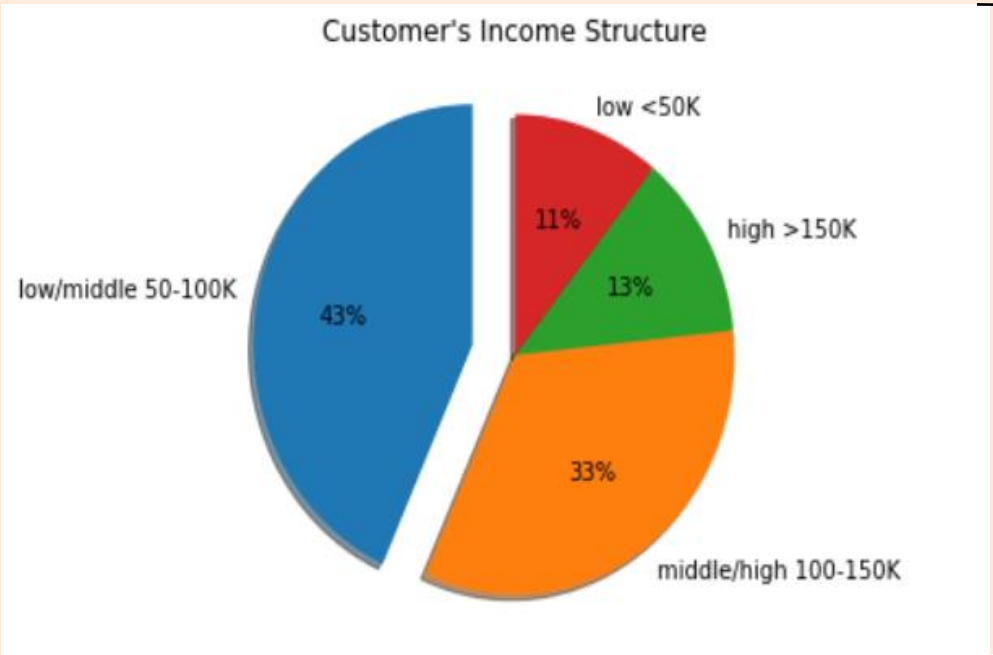
(3.) Check for duplicates

```
df_ords_prior_dups= df_ords_prior[df_ords_prior.duplicated()]
```

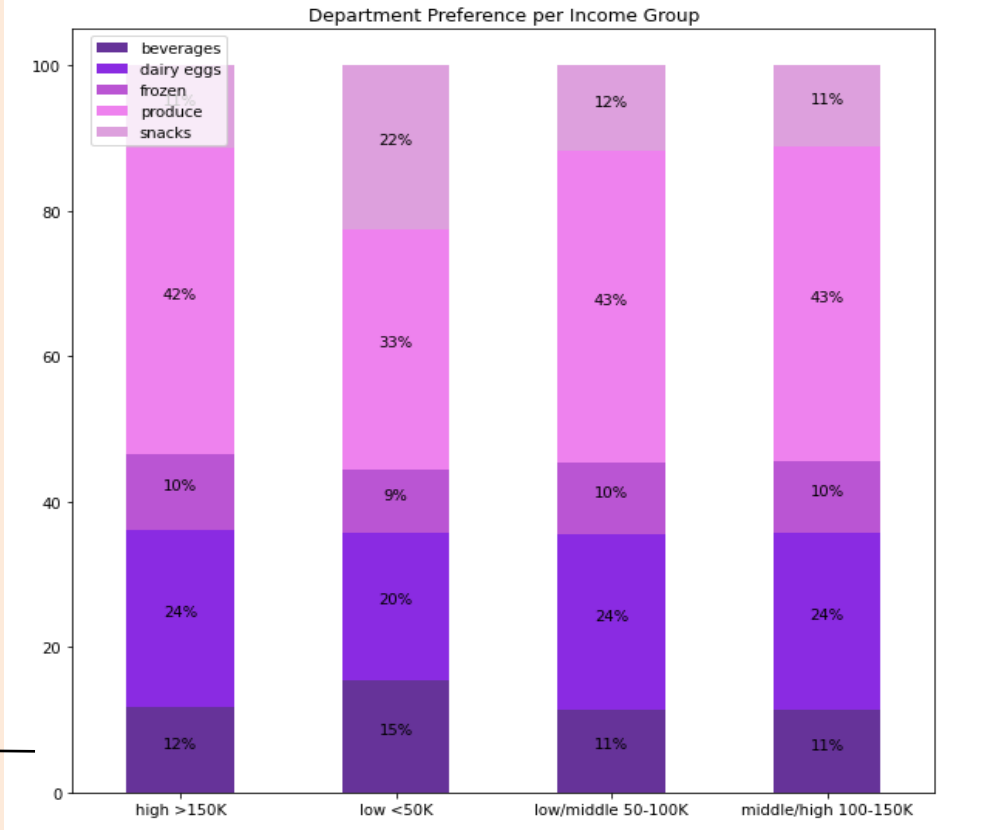
```
df_ords_prior.dups
```

order_id	product_id	add_to_cart_order	reordered

No duplicate is found.



I have used bar chart to represent the customer's income structure. We can see that 76% of customer's earning is in the middle class and more than half of them earn 50-100k. Similar analyses are also applied to age structure, family status, purchasing frequency and regions to investigate customers' background and behavior.



100% stacked bar chart is used to exhibit the influence of Income on department preference. Low-income customers have especially different purchasing behavior than the other income groups. They tend to buy more snacks and beverage, less dairy eggs, frozen and produce products.

Conclusion

Insights and Recommendations

Insights

- Busiest days are **Saturday and Sunday** while the busiest hours of the day are between **9-16**.
- Around 98% of goods that customers buy cost between **1-15 USD**.
- Department popularity: **produce > dairy > snacks > beverages > frozen**
- 68% customers purchase groceries **within 10 days**, while only 9% customers over 20 days.
- Sales performance in South is relatively poor than the other regions.
- Low-income customers prefer more snacks and beverage, less dairy eggs, frozen and produce products.
- High-income young parents tend to buy more dairy eggs and frozen products but less products.

Recommendations

- Schedule advertisement especially on **Tuesday and Wednesday starting from 7 am** that there would be special discounts on popular goods or shipping cost these days to enhance sales.
- **Different levels of membership** can have different extent of benefits to retain the frequent customers and intrigue regular customers to become frequent customers.
- For low-income customers: Advertise mainly snacks and beverages to them. If there is limited or special discounts on diary or produce products, it should also be included in the ads.
- For high-income young parents: Advertise mainly dairy and frozen products to them. Ads can also include produce products which are convenient to eat such as salat.

Project Reflection

4 datasets were downloaded to investigate customer's purchasing behavior in this case. When I merged the datasets to have a general view among the purchasing factors, I met serious memory shortage issue. In order to solve the problem, I created an empty shell to save the appended result at the beginning, then removed the relatively larger dataset to save the memory. In the end, I have deleted a function that would merge the chunks. However, it didn't work.

I checked the memory usage again before and after I applied this method and surprisedly found out that this method would use even more memory and cannot solve the problem efficiently, so I searched in the Stack Overflow and found out a method altering the datatype to save the memory. During the process, try and error is needed since too much space available for the data would cause memory collapse but too little space saved for our data cannot demonstrate descriptive analysis properly. In the end, I successfully assigned proper datatype to each attribute. For example, I have changed the order number from int64 to int8 while changing float64 to float16 for the column of days since prior order. The memory issue is then solved and I can perform descriptive analysis well.

From this project, I not only learnt how to dealt with memory issue but also learnt how to transform data to valuable insights such as deriving new variables, and visualize this insight into different types of charts by Python.