SUREKHA

Instacart case studies

EXPLORATORY ANALYSIS

CONTEXT

Instacart is an online grocery store that aims to boost sales through targeted marketing efforts.

KEY QUESTIONS

Sales and marketing teams require insights into consumer buying habits and customer demographics.

TOOLS

Python, Jupyter Notebook, Pandas, Seaborn and Matplotlib libraries, Excel

DATA

- **Instacart** open source dataset
- Customer and prices data set created by Career Foundry

SKILLS

- Data cleaning
- Data wrangling
- Data merging
- Deriving variables
- Grouping and aggregating data
- Reporting

THE PROCESS

1. **The data preparation phase**

   For my analysis, I cleaned the original Instacart data and the CF-generated data, which involved checking for missing values, and inconsistent data types, and deleting unnecessary columns. I then merged the two datasets. Since I already knew the questions that needed to be answered, I derived the necessary variables to conduct the analysis.

2. **The analysis phase**

   To organize the list of questions, I created a structure that consisted of four sections: time analysis, product analysis, customer analysis and customer profiling. As the dataset had limitations, each section had its own set of constraints that I had to overcome.

3. **The results**

   Despite the sound reasoning behind the analysis, the data limitations were apparent in all four sections of the results. The stakeholders were informed of the analysis results, accompanied by a clear disclaimer concerning the data limitations.

TIME ANALYSIS

QUESTIONS/TASKS

- Determine the days of the week and hours of the day when fewer orders are placed.
- Identify the specific periods in a day when individuals tend to spend the highest amount of money?

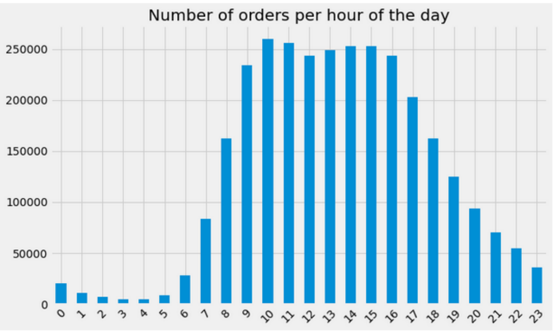
CONSTRAINTS — Fabricated prices data set and no data about the quantities of the ordered items

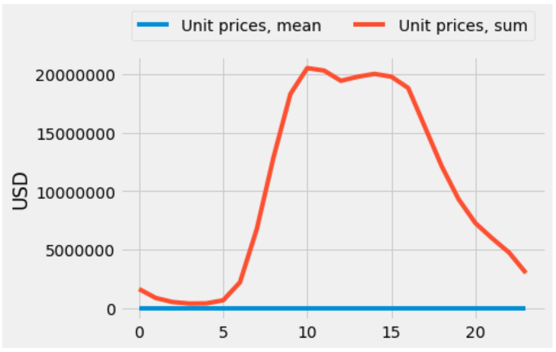
SOLUTION — Analysis was based on the number of unique orders and their distribution throughout the week, including the hours of the day.

TIME ANALYSIS

The bar charts utilized the original Instacart data to display the distribution of orders per day of the week and hour of the day. Surprisingly, the orders were distributed during the week in a narrow range. However, no obvious signs of bias were visible in the charts.

After analyzing the prices of products throughout the day, it was found that the average price per product remains stable regardless of the time of day. Moreover, the graph of total sales values during the day follows the same pattern as the chart showing the number of orders per hour.





PRODUCTS ANALYSIS

QUESTIONS/TASKS

- Simplify pricing by adding a price flag corresponding to price ranges such as low, medium, and high.
- Determine which departments are most popular with customers based on sales volume.

CONSTRAINTS

Fabricated prices data set and no data about quantities of the ordered items

SOLUTION

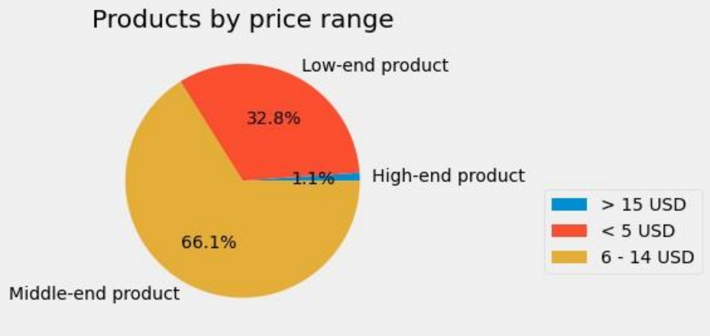
Analysis was based on the number of times the **items** were ordered

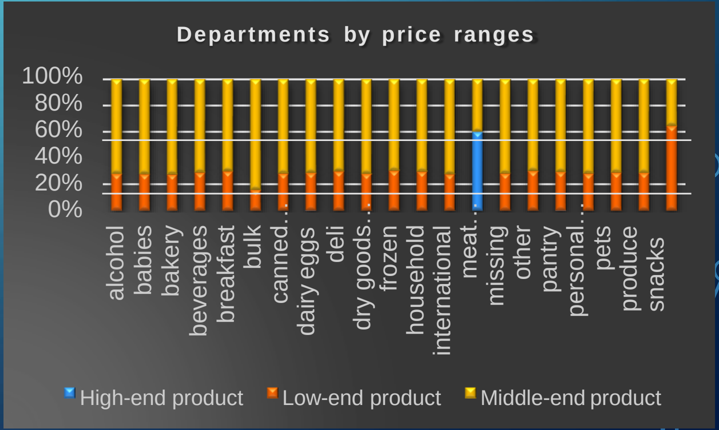
PRODUCTS ANALYSIS

Based on their price, products were classified with over 65% of products being middle-end with prices between 6 and 14 USD. Only 1% of the products cost more than 15 USD.

Ideally, I would gauge the popularity of departments by analyzing the quantities of sold products. However, since that data is unavailable, I have arranged the departments based on the total number of times products from a specific department were ordered.

After cross-referencing departments and price flags, I was able to connect the price ranges and departments. Upon analysis, I found that there were no significant differences between the proportion of the price groups for the entire dataset and the department subsets, except for meat/seafood and snacks.





CUSTOMER ANALYSIS

QUESTIONS/TASKS

- Determine the ordering habits and behavior of customers.
- Determine if there are differences in ordering habits based on brand loyalty, demographics, and geography.

CONSTRAINTS —

Fabricated prices data set and fabricated demographical and geographical data about the customers

SOLUTION —

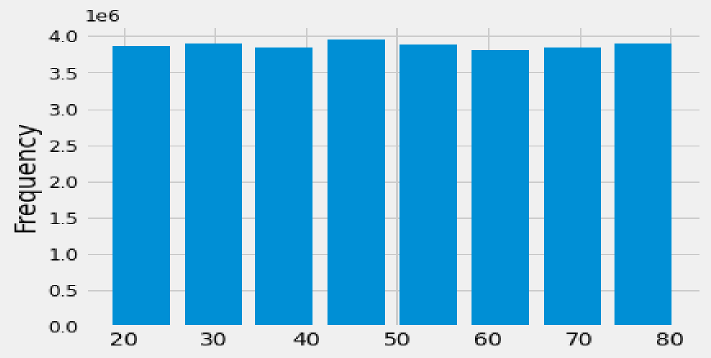
Unfortunately, there were no other options but to use the data as it is.

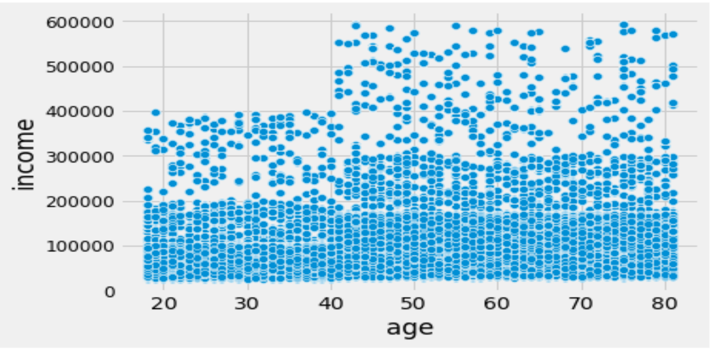
CUSTOMER ANALYSIS

Defining "ordering habits" was challenging due to data bias, so I used three criteria.

- Brand loyalty
- Frequency of ordering
- Spending habits

Although this section of the analysis was the most extensive, the final conclusion drawn was that there were no significant differences in terms of above 3 criterias between the complete dataset and any of its subgroups, irrespective of the filtering, grouping, or aggregation done on it. The bias present in the data was most evident in the uniformity of the age and income variables..





CUSTOMER PROFILING

QUESTIONS / TASKS

- Identify particular customer demographics and characteristics.
- Determine if there are any differences in the ordering habits between the profiles?"

CONSTRAINTS

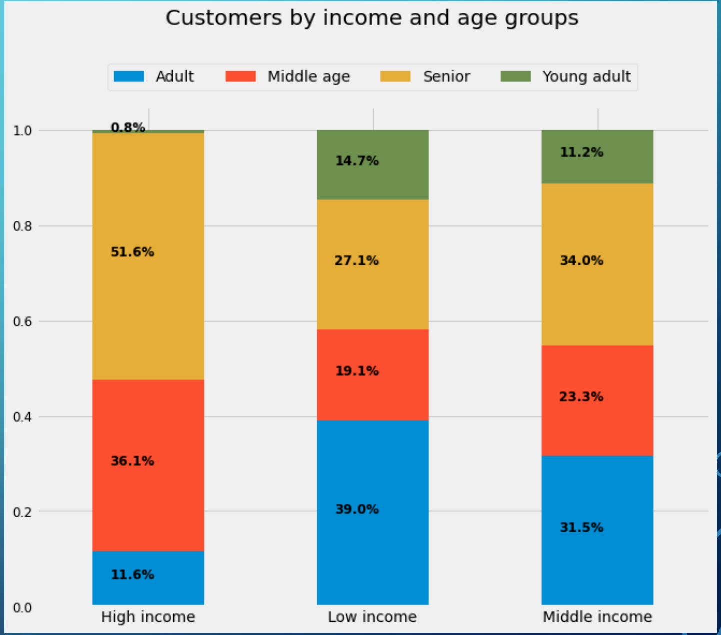
Fabricated prices data set and fabricated demographical and geographical data about the customers

SOLUTION

Unfortunately, there where no other options here but to use the data as it is.

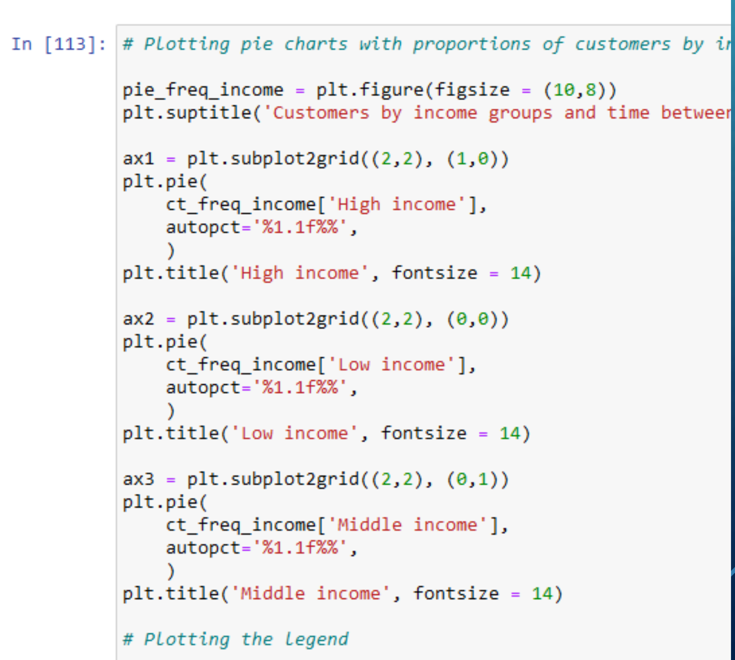
CUSTOMER PROFILING

As per the task requirement to compare the ordering habits based on the three main criteria established earlier, the analysis revealed no significant differences between the subsets. Paid special attention to the income-age grouping and found that more than 50% of the high-income group consists of seniors. This revealed the problem with the demographic data, as the purchasing power usually decreases after retirement. Lastly, I performed micro-profiling by grouping the customers based on their gluten tolerance, which was crudely determined by whether or not they ordered a product from the "bakery" department.



CONCLUSION

The primary challenge I faced during the project was how to overcome the limitations of the data and produce an analysis that aligned with the project outline's questions. Even though I provided answers to the questions, I acknowledged that they were incomplete, leaving me curious about the actual results of the analysis with the original data. From a technical perspective, working with a dataset that had over 30 million rows posed several problems, particularly when running code with high computational demands. Consequently, I quickly adopted various RAM usage optimization techniques to mitigate such problems.



```
In [113]: # Plotting pie charts with proportions of customers by i

          pie_freq_income = plt.figure(figsize = (10,8))
          plt.suptitle('Customers by income groups and time between

          ax1 = plt.subplot2grid((2,2), (1,0))
          plt.pie(
              ct_freq_income['High income'],
              autopct='%1.1f%%',
              )
          plt.title('High income', fontsize = 14)

          ax2 = plt.subplot2grid((2,2), (0,0))
          plt.pie(
              ct_freq_income['Low income'],
              autopct='%1.1f%%',
              )
          plt.title('Low income', fontsize = 14)

          ax3 = plt.subplot2grid((2,2), (0,1))
          plt.pie(
              ct_freq_income['Middle income'],
              autopct='%1.1f%%',
              )
          plt.title('Middle income', fontsize = 14)

          # Plotting the Legend
```

THANKYOU

My Github:
https://github.com/rekhajuttiga