

Grocery Basket Analysis

Chaitali Limbhore
Career Foundry
Data Analytics Portfolio



Introduction

OVERVIEW

Instacart is an online grocery store that operates through an application. It already has very good sales and wants to uncover more information about their sales pattern based on customer profiles with the appropriate products.

OBJECTIVE

Analyze the data and perform an initial data and exploratory analysis in order to derive insights and suggest strategies for better segmentation based on the provided criteria.

PURPOSE & CONTEXT

I have done this project as a part of my Data Analytics course at Career Foundry to demonstrate my skills in data analysis and visualization using Python.

Project Plan

PHASE 1

Data and Business Understanding

PHASE 2

Data Preparation

PHASE 3

Analysis and Visualizations

My Role

Data Analyst

Project Duration

30 Days

Tools Used

Jupyter Notebook(Anaconda)
for Python, Excel

Credits

Tutor : Mohsun Hajiyev
Mentor : Madhuri Joshi

Phase 1:

Data and Business Understanding

- In this phase of project, I tried to understand business goals, objective and requirements from BRD. (Business Requirement Document)
- I did check for different datasets from Instacart by looking at it's size, columns, data types etc.
- Below are the key questions from Sales and Marketing teams of Instacart:

What are the busiest days of week and hours of the day?



What are the hours of day when people spend most money?



Are there certain types of products that are more popular than others?



Analyze different types of customers based on :

- **Loyalty**
- **Region**
- **Age and Family status**

Data Sources and Python Skills

- For this project, we have used number of open-source data sets from Instacart.
- We also have a customer data set (created and included for the purpose of this project)
- Below are the sources for data sets and python skills used to understand the data.

DATA SOURCES

- ❑ **The Instacart Online Grocery Shopping Dataset 2017”, Accessed from**
<https://www.instacart.com/datasets/grocery-shopping-2017> on May 15, 2021
- ❑ **Data Dictionary:** [The Instacart Online Grocery Shopping Dataset 2017 Data Descriptions \(github.com\)](https://github.com/instacart/instacart-data-dictionary)
- ❑ **Customers Data Set:**
https://s3.amazonaws.com/coach-courses-us/public/courses/data-immersion/A4/A4_Data_Assets/customers.zip

PYTHON SKILLS USED

- ❑ Importing Python Libraries
- ❑ Importing and Exporting Instacart Data Sets
- ❑ Checking the size of data and different columns with their data types.
- ❑ Understanding necessary columns which will be useful for further analysis.

Phase 2:

Data Preparation

STEPS INVOLVED

- In data preparation phase, data has been checked for its quality and consistency so that clean dataset is available for further analysis.
- Missing values, duplicate records and mixed-type columns from data have been checked and corrected.
- Renaming column names so that it can be easily understandable.
- Data wrangling, subsetting and deriving new variables to create customer profiles.
- Combining different datasets using integration, so that only one data set can be used for analysis.

PYTHON SKILLS USED

- ❑ Data wrangling and subsetting
- ❑ Data Consistency Check
- ❑ Deriving new variables
- ❑ Merge the data using integration

Bringing Datasets Together

Below population flow shows integration of different datasets and their total count at each stage.

Orders, Products and Customers datasets have been checked for consistency and then combined and merged to get a final integrated dataset.

Population flow

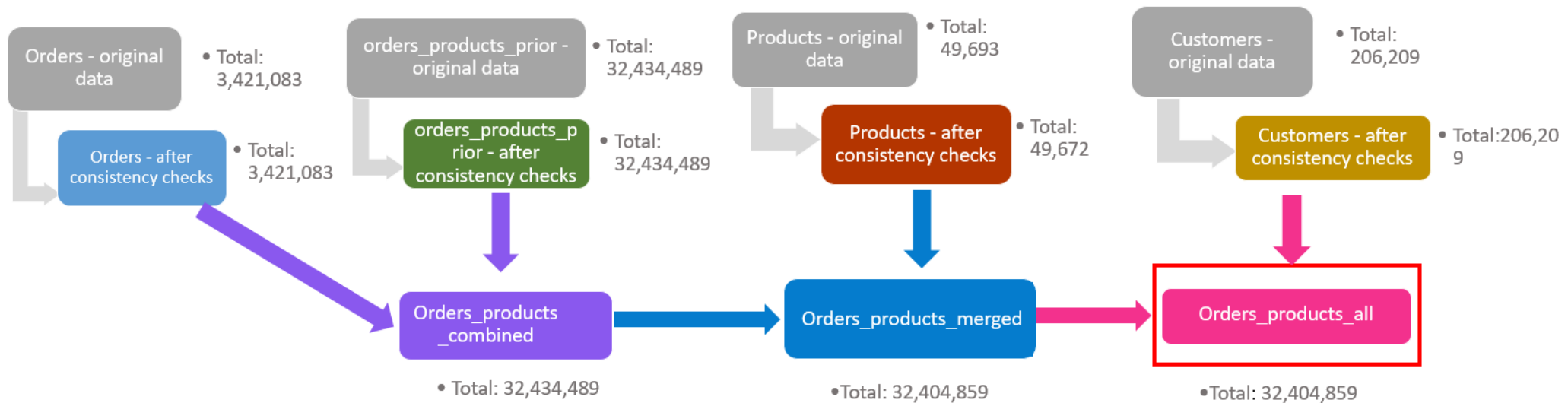


Fig.1 Integration of datasets

Phase 3:

Analysis and Visualization

STEPS INVOLVED

- The last phase of project is Analysis and Visualization.
- In this phase, different statistical and descriptive analysis have been performed on final dataset to answer the business questions.
- Data has been grouped and aggregated to check the relationship between variables.
- Final results are presented in the visual form using python code for visualization.

PYTHON SKILLS USED

- ❑ Descriptive and Statistical Analysis
- ❑ Grouping Data
- ❑ Aggregating Data
- ❑ Visualizations using Python libraries.

Results and Recommendations

Saturday & Sunday have the highest number of orders placed while Tuesday & Wednesday have the lowest number of orders placed.

I would recommend ads to be run on weekdays to help with Mon - Fri order volume.

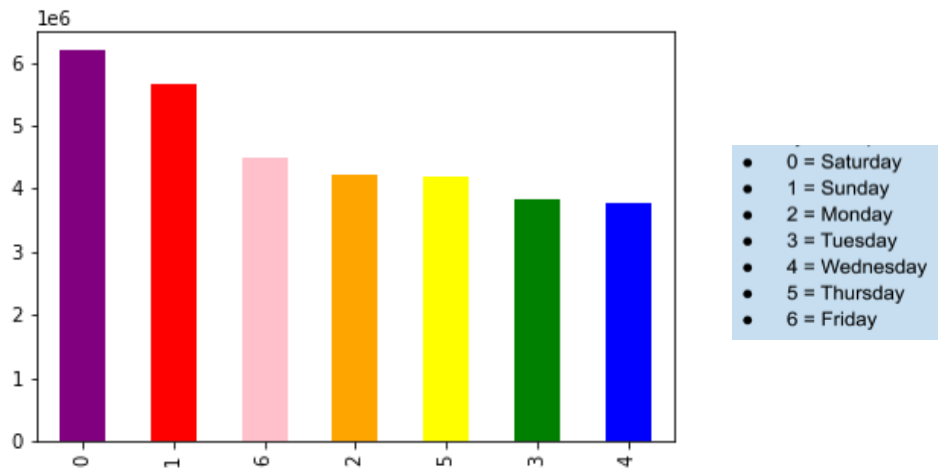


Fig.2 Number of Orders Per Day

Very few orders are placed through the hours of night.

With the most popular hours being 10am - 3pm & least popular hours being 10pm - 6am, company should focus on ads during 6pm - 10pm.

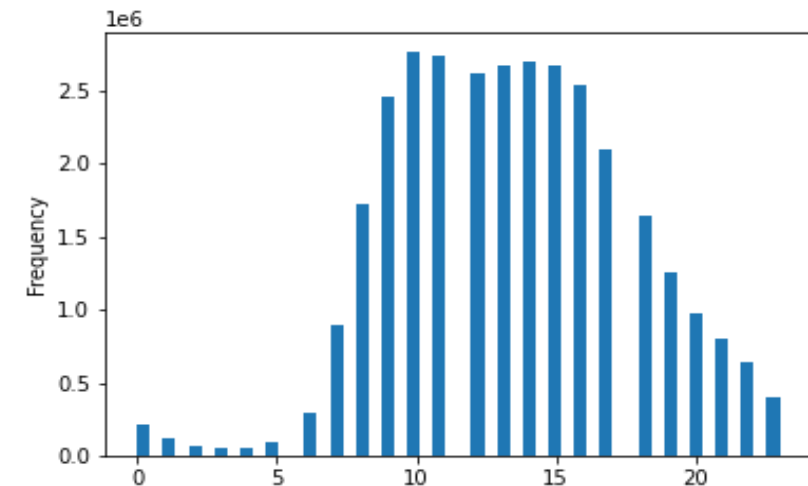


Fig.3 Number of Orders Per Hour

Results and Recommendations

Surprisingly orders at 2am result in the highest average prices paid (perhaps emergency purchases? or people may spend more time on shopping during the night time).

Because marketing at such an uncommon time may not be worthwhile, I would recommend focusing advertising on the 2nd highest paid times at 6am and other hours which are having average price of products purchased.

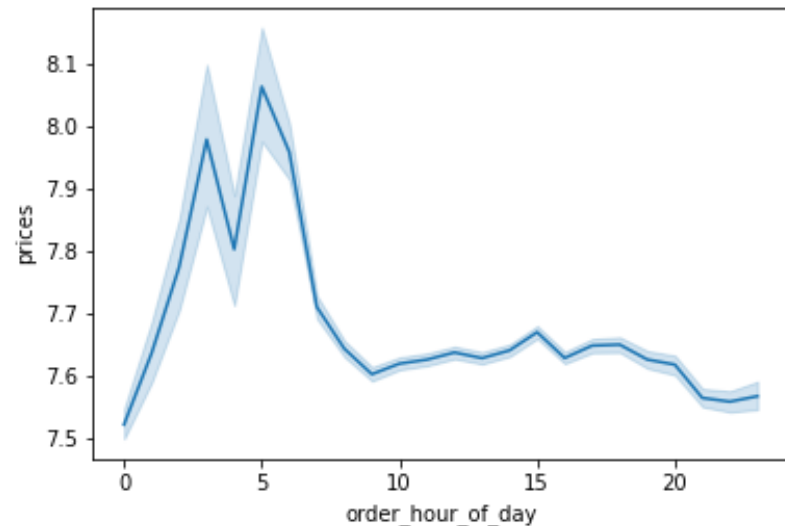


Fig.4 Average Price of Products Purchased Each Hour

Produce is the most popular department followed by dairy/eggs, snacks, frozen then beverages.

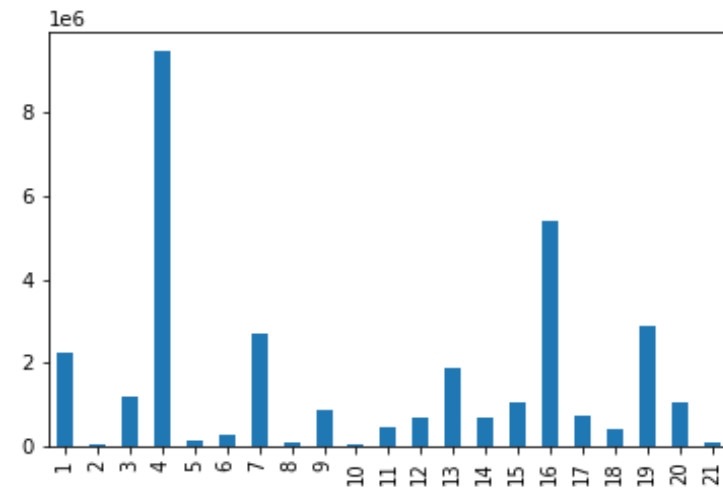


Fig.5 Frequency of Orders Placed in Each Department

department_id	department
1	frozen
2	other
3	bakery
4	produce
5	alcohol
6	international
7	beverages
8	pets
9	dry goods pasta
10	bulk
11	personal care
12	meat seafood
13	pantry
14	breakfast
15	canned goods
16	dairy eggs
17	household
18	babies
19	snacks
20	deli
21	missing

Results and Recommendations

The min and max ordering price is same for all customer categories.

But average price for ordering is the highest for new customers followed by regular and then loyal customers.

loyalty_flag	prices		
	mean	min	max
Loyal customer	10.386336	1.0	99999.0
New customer	13.294670	1.0	99999.0
Regular customer	12.495717	1.0	99999.0

Fig. 6 Distribution of orders among customers in terms of loyalty

South region has the largest number of total orders followed by West, Midwest region.

Northeast has lowest number of orders as compared to other regions.

Region	order_number sum
Midwest	128585728.0
Northeast	98521079.0
South	185091277.0
West	143295881.0

Fig.7 Ordering habits based on a customer's region

Customers Ordering Habits



Busiest Days: Saturday and Sunday
Busiest Hours : Between 10 AM and 3 PM



Most Popular Department : Produce



Times of the day when people spend the most money: Early morning 6 AM



Average price of ordering is the highest for new customers followed by regular and then loyal customers.



South region has the largest number of total orders followed by West, Midwest region.



Challenges Faced



This was my favourite as well as most challenging project. Since Python was completely new for me, I enjoyed working with it.

While working with large datasets, Jupyter Notebook slowed down dramatically and did not produce all the necessary outputs. It also started throwing some memory errors while integrating and exporting datasets. My laptop got hanged many times due to this time consuming execution for large datasets.

Solution: I discussed this issue with my tutor and he shared one article which explained this memory issue due to big data types for columns. I then converted data types for all necessary columns and then this issue got resolved.



Thank You!!

Glad you made it through! Please feel free to reach out to me using below links 😊

