Instacart Grocery Basket Analysis Case Study

Marketing strategy for an online grocery store.

Instacart Basket Grocery Analysis

What is Instacart?

- Instacart is a grocery store that uses an app to operate.

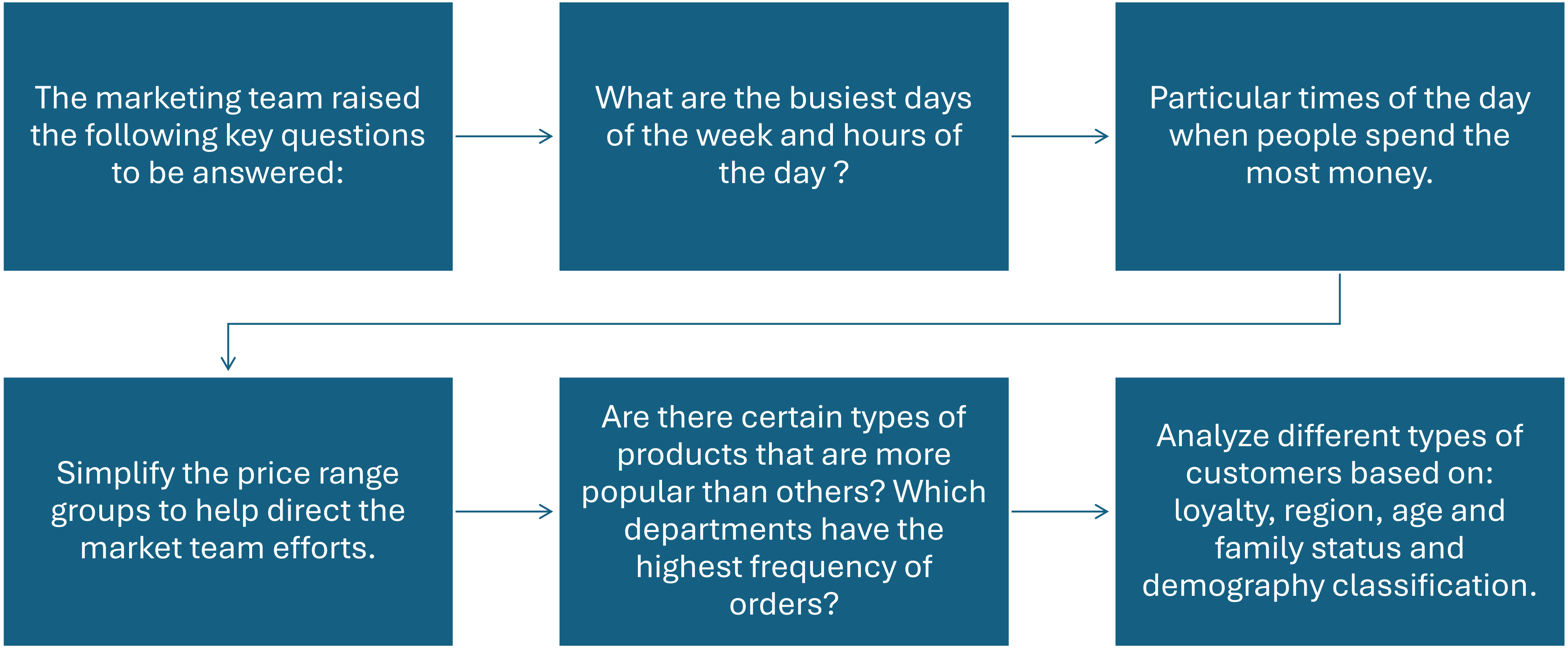
What was this project for?

- The main objective of this project was the execution of a data and exploratory analysis in order to improve Instacart sales and suggest strategies for client segmentation working directly with the marketing department.

Context

- This was a project carried out for a fictitious company with the aim of enabling practical learning of the Python tool. The idea was to use open-source data to identify customer consumption patterns recorded in the Instacart database, creating tags for each consumer group. With these tags, it would be easier to target specific marketing campaigns and thus increase Instacart sales.
- My role as an analyst was to outline how this strategy would be implemented based on the results of the analysis, confirming that consumers have the correct corresponding tags.

Key Questions



The marketing team raised the following key questions to be answered:

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

Particular times of the day when people spend the most money.

Simplify the price range groups to help direct the market team efforts.

Are there certain types of products that are more popular than others? Which departments have the highest frequency of orders?

Analyze different types of customers based on: loyalty, region, age and family status and demography classification.

Instacart Supermarket

Data Set

Open data source:

[Customers dataset]

[Data dictionary]

[Dataset]

Tools

Microsoft Excel

Python

Skills

Python

Data wrangling

Data merging

Deriving variables

Grouping data

Aggregating data

Reporting in Excel
Population flows

Project Process, Results and Challenges

Using Jupyter notebook after loading Pandas, Numpy and Os, the datasets provided were cleaned and wrangled

A descriptive exploratory analysis was also conducted

Data Consistency Checks: Fixed mixed-type variables, uncovered and deal with missing values, remove duplicates

Exploratory Data Analysis to derive key customer insights

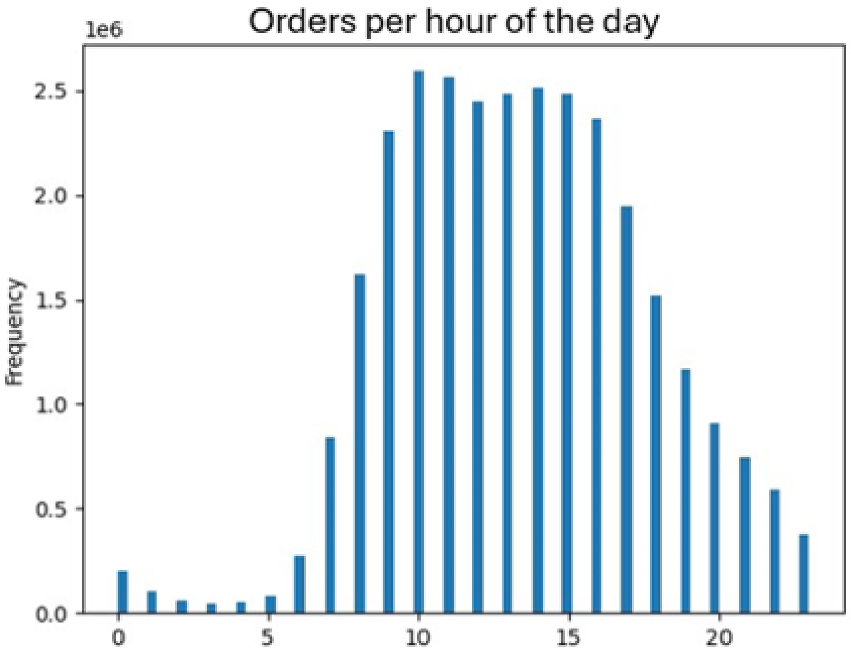
Merged datasets to discover variables relationships

Created new columns using conditional logic to derive new variables for analysis

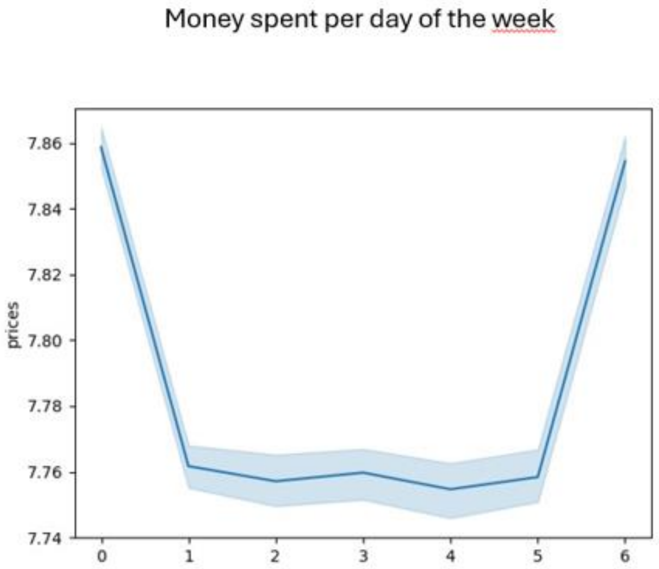
Grouped and aggregated key data to create various customer profiles.

Project Process, Results and Challenges

- The first stage involved installing Python and the Anaconda interface. Subsequently, using Jupyter notebook, the main project folder was created, and the necessary libraries were installed. The Pandas library was used to perform basic descriptive analysis, including determining the mean, median, total number of rows, standard deviation, minimum, and maximum.
- Next, data wrangling was conducted, involving changing necessary data types, renaming columns, analyzing data types using a data dictionary, and then performing subsetting by creating new dataframes based on certain criteria. At this point, some questions were already addressed, such as identifying the time of day with the highest number of orders.

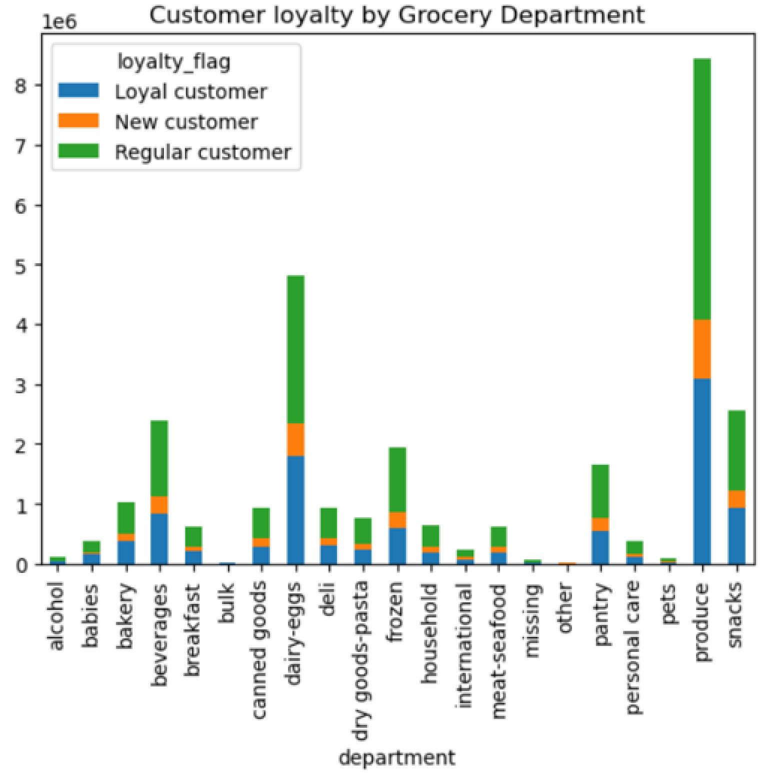


- The following stage included consistency checks, correcting mixed variables, detecting and handling missing values, and removing duplicates. This step was crucial for producing a reliable dataset with a lower risk of errors or bias in the final analysis.
- After executing these steps for each dataframe, they were merged to create a final dataframe used for building visualizations and statistical analyses. This step posed challenges due to the large size of the dataframes, causing computer slowdowns and crashes. To overcome this, unnecessary columns were identified and deleted, reducing file size. Additionally, sampling using Numpy was employed in specific situations, such as for line charts.



Legenda:
0 : Saturday
1: Sunday
2: Monday
3: Tuesday
4: Wednesday
5: Thursday
6: Friday

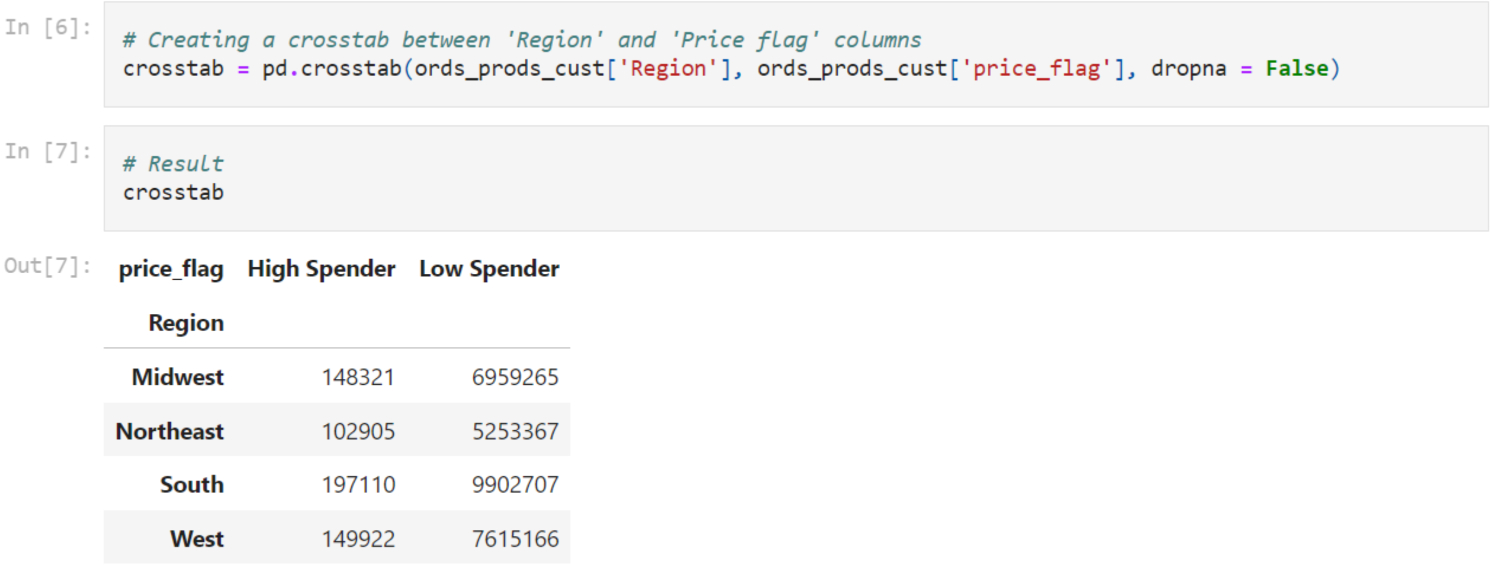
- With the final dataset ready, the process of creating new columns based on certain criteria using logical conditionals with if-statements, user-defined functions, loc() function, and for-loops began. This involved creating columns with specific flags, such as price labels, orders per day of the week, busiest period of the day, etc.
- From this point, the work took shape, and some key questions could be answered, such as "What are the busiest days of the week and hours of the day?" Various visualizations were then constructed using the Seaborn and Matplotlib libraries, providing a better understanding of the results and facilitating the presentation of findings through charts and graphs.
- The final results were presented in a report using Excel.



- For more details about the code and project process as well as the final report, access the Github links here:
- Analyze order behavior of different customer groups
- Final Report on Excel

Examples of Tables and Charts Created

Expending difference between US Regions



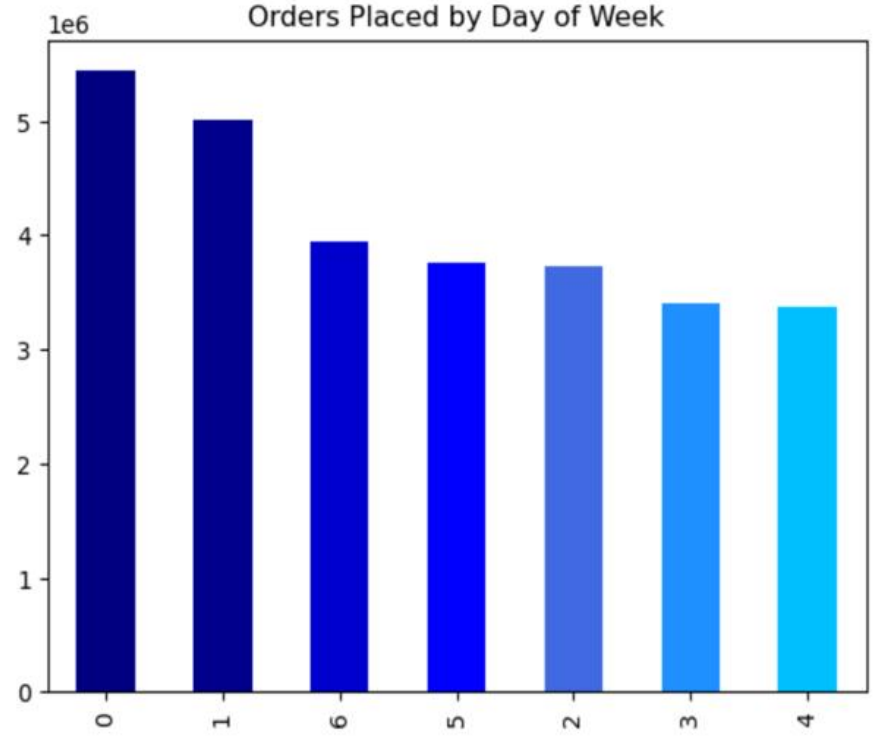
```
In [6]: # Creating a crosstab between 'Region' and 'Price flag' columns
        crosstab = pd.crosstab(ords_prods_cust['Region'], ords_prods_cust['price_flag'], dropna = False)
```

```
In [7]: # Result
        crosstab
```

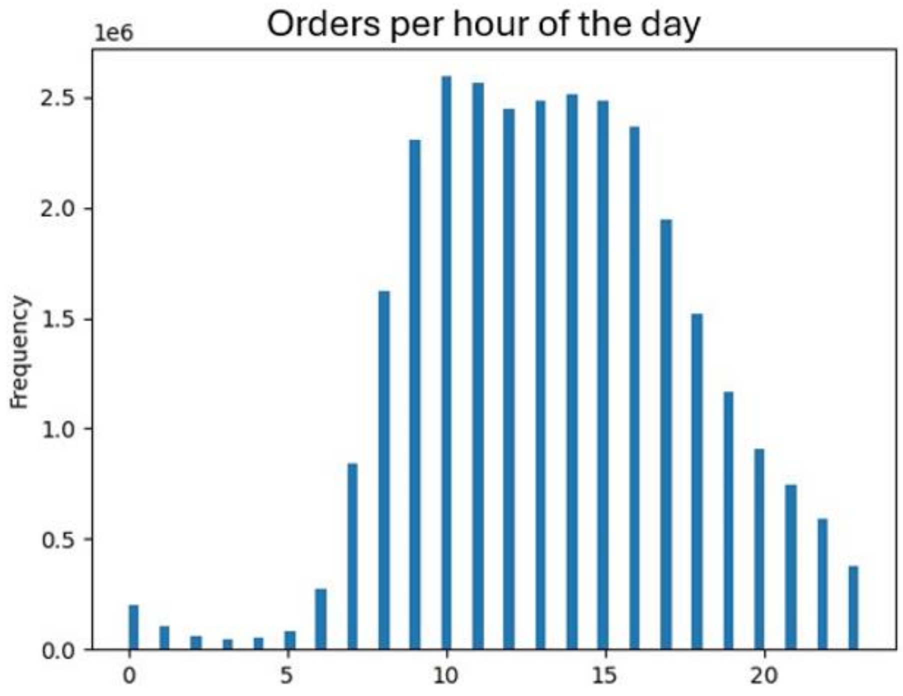
Out[7]:

price_flag	High Spender	Low Spender
Region		
Midwest	148321	6959265
Northeast	102905	5253367
South	197110	9902707
West	149922	7615166

Histograms for Busiest Days and Hours of the Week

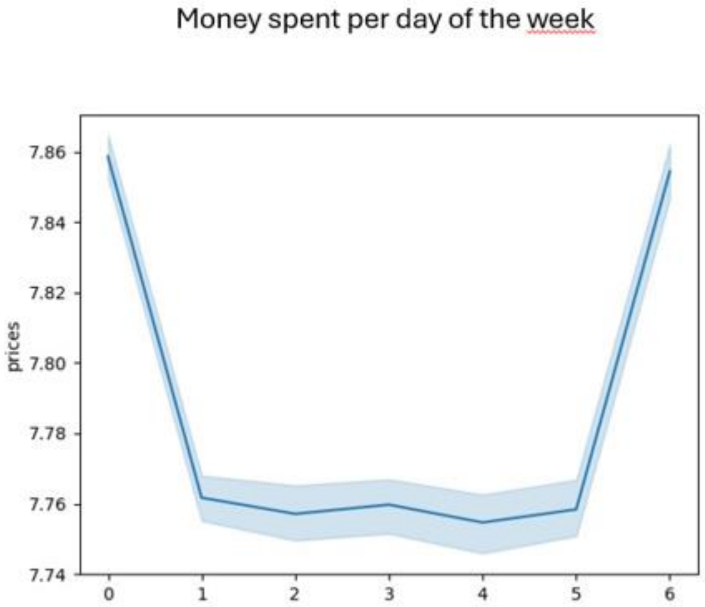


0 – Saturday. 1-Sunday, 2-Monday, 3-Tuesday, 4-Wednesday, 5-Thursday, 6-Friday.



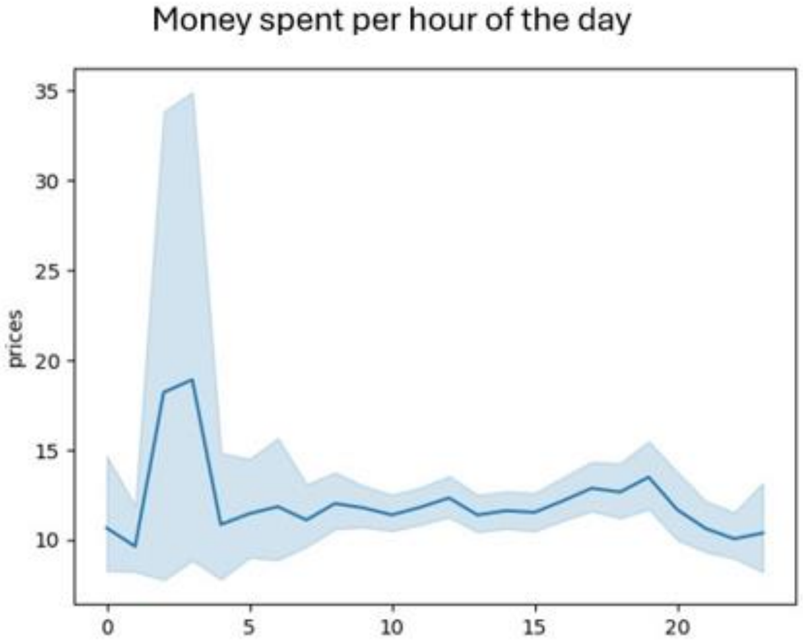
Saturday was the busiest day and 10:00 am was the busiest time of the day.

Money Spent by Day of the Week and Hour of the Day



Money spent per day of the week

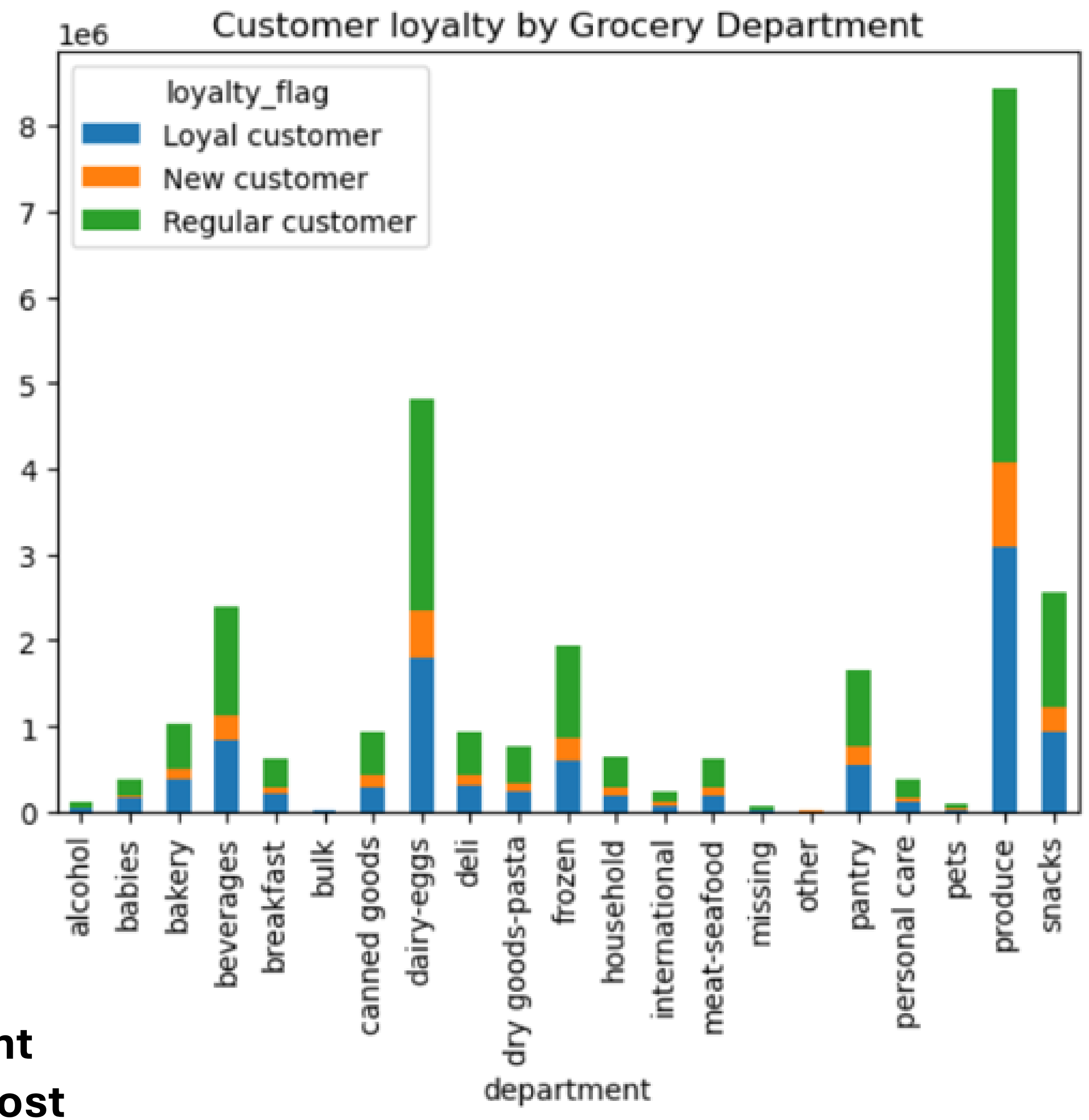
Legenda:
0 : Saturday
1: Sunday
2: Monday
3: Tuesday
4: Wednesday
5: Thursday
6: Friday



Money spent per hour of the day

Clients used to spend more money on weekends and during the night (00h until 05h).

Sales by Grocery Department Accordly to Customer Loyalty



Produce department was the most popular.

Conclusion

- The project allowed me to acquire important skills in Python and was a highly motivating process throughout the entire work. As I already had some knowledge of programming in R, comparing the two tools was also interesting, and today, I find that Python has a more user-friendly and pleasant interface, especially for beginners.
- I didn't encounter significant difficulties with the code, as I had the support of my mentor and tutor. I also found that Python has a vast online community, which I frequently turned to in case of challenges. This tool enables the execution of quality work even when in the learning process and working independently.
- I believe the biggest challenge of the project was dealing with very large datasets, as my computer did not have sufficient memory for this type of work. Additionally, choosing which visualizations to use from various libraries was quite challenging. For a beginner data analyst, determining what is most important can be complicated.
- I had hoped to conduct a geographic analysis of consumers by state, but the dataset had an equal number of consumers per state, which unfortunately did not provide relevant information.
- Overall, this project was very rewarding and sparked my interest in programming for data analysis, which I aim to deepen every day.

Complete Project Links

- Instacart
- Contact: mariana.oliveiria.data@gmail.com