Instacart:

Market Basket Analysis with Customer and Product Segmentation

What is Instacart?

An American company operating a grocery delivery service made abundantly necessary in the times of COVID-19.

"In a matter of a couple of weeks, we were already ahead of our end-of-year goal. A week later, we were ahead of our 2021 goals, and a few days after that, we were ahead of our 2022 goals."
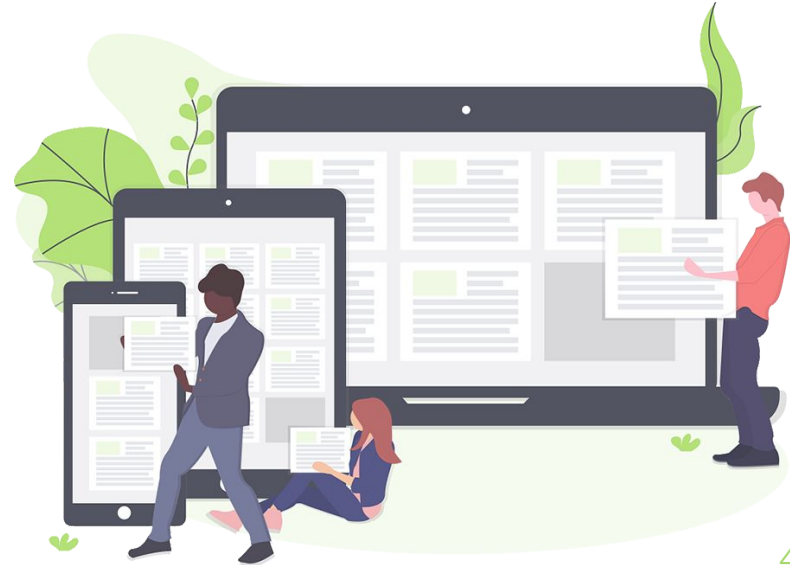
Apoorva Mehta, CEO, Instacart, Inc.
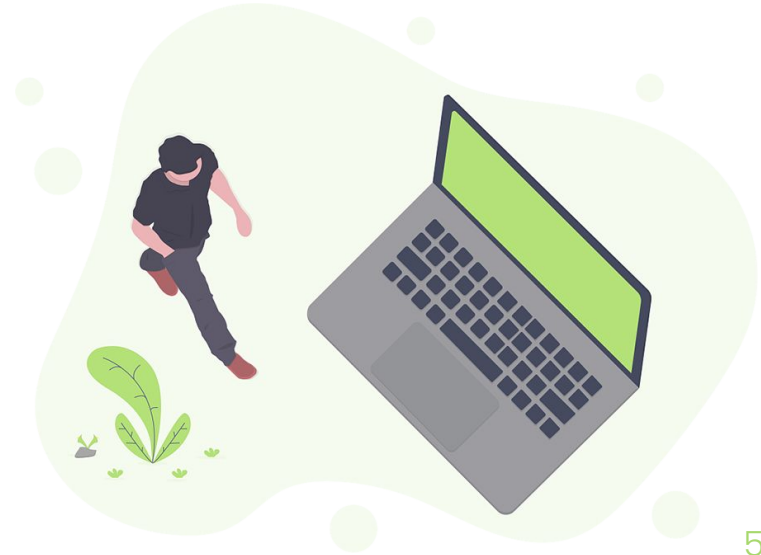via Bloomberg Businessweek

Our Selected Topic

- Market Basket Analysis

- Customer Segmentation

- Product Segmentation
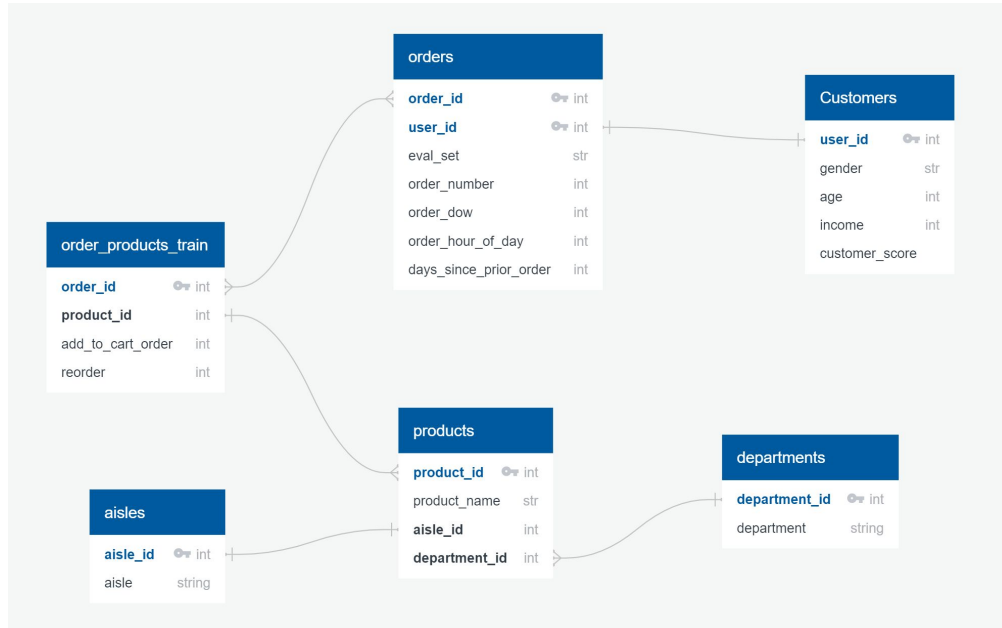
The Exploratory Phase

1,315,590 orders

49,688 products

12,950 customers

*All data initially in six tables

The Exploratory Phase Continued..

orders

order_id	🔑	int
user_id	🔑	int
eval_set		str
order_number		int
order_dow		int
order_hour_of_day		int
days_since_prior_order		int

Customers

user_id	🔑	int
gender		str
age		int
income		int
customer_score		

order_products_train

order_id	🔑	int
product_id		int
add_to_cart_order		int
reorder		int

products

product_id	🔑	int
product_name		str
aisle_id		int
department_id		int

departments

department_id	🔑	int
department		string

aisles

aisle_id	🔑	int
aisle		string

# Questions We Hope to Answer

Customer Segmentation:

1. What is the distribution of annual Income by age?

2. What is the distribution of gender for Instacart users?

3. What is the age group of most frequent Instacart users?

4. What is the overall distribution of income level for instacart users?

5. How do male and female spending scores differ?

6. Who is our targeted customer and who is our least active customer?

Questions We Hope to Answer

Product Segmentation:

1. What are the Top 10 most ordered products?
2. What are the Top 10 most reordered products?
3. Which 2 products were ordered together the most?
4. Which day and time of the week was the busiest in terms of number of orders received?
5. Which products were not frequently reordered? This could help determine which products did not meet customers expectations.
6. What were the top aisles with the least amount of orders? This could be a recommendation that we make to our vendors in terms of restructuring their aisles.
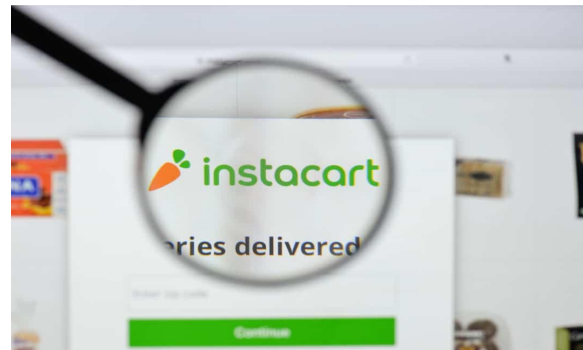
The Analysis Phase

Tableau Analysis

and

Machine Learning

Customer Analysis | **Product Analysis**
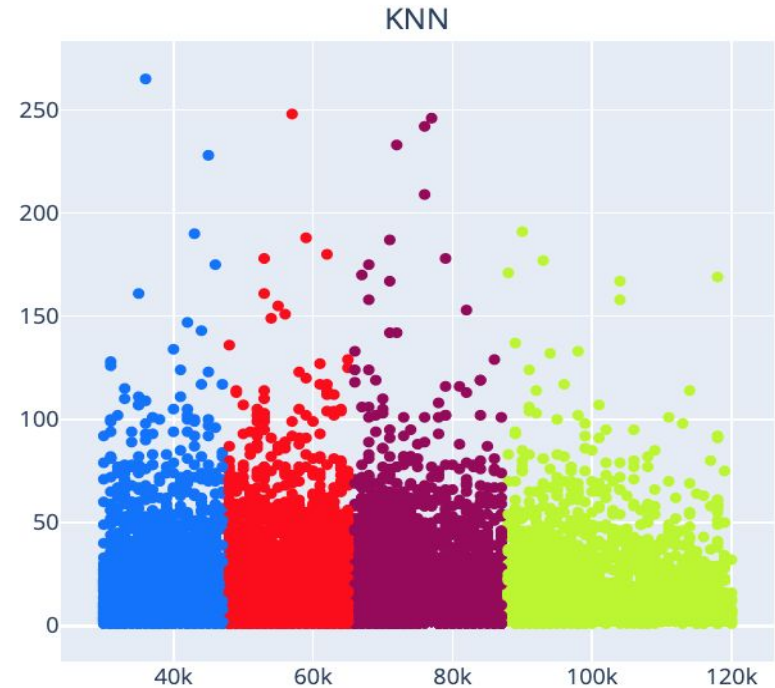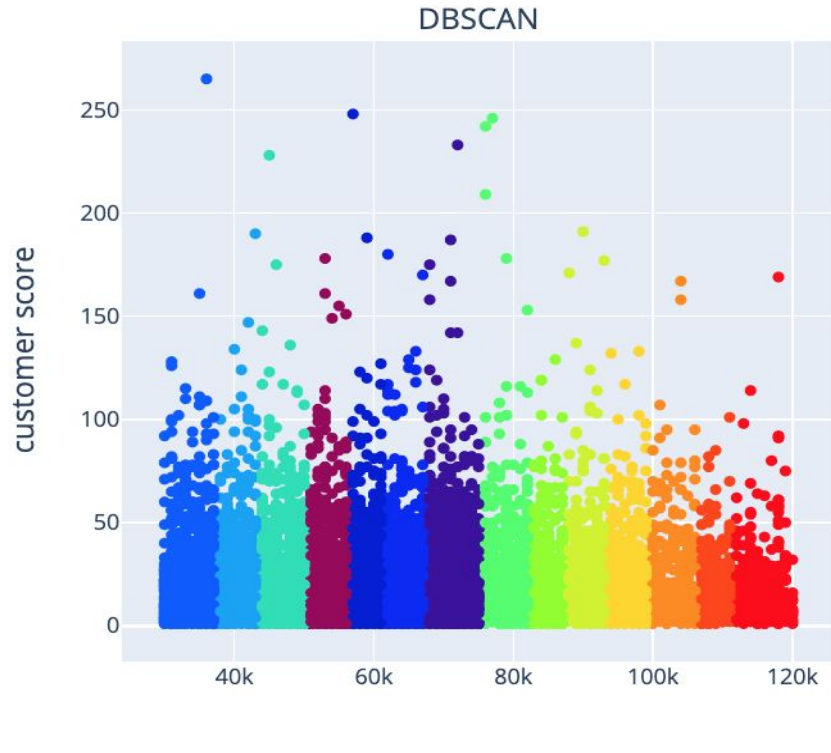
Machine Learning

Customer Segmentation:

Utilizing KMeans, we hope to be successful in clustering and categorizing the Instacart customers into four main groups:
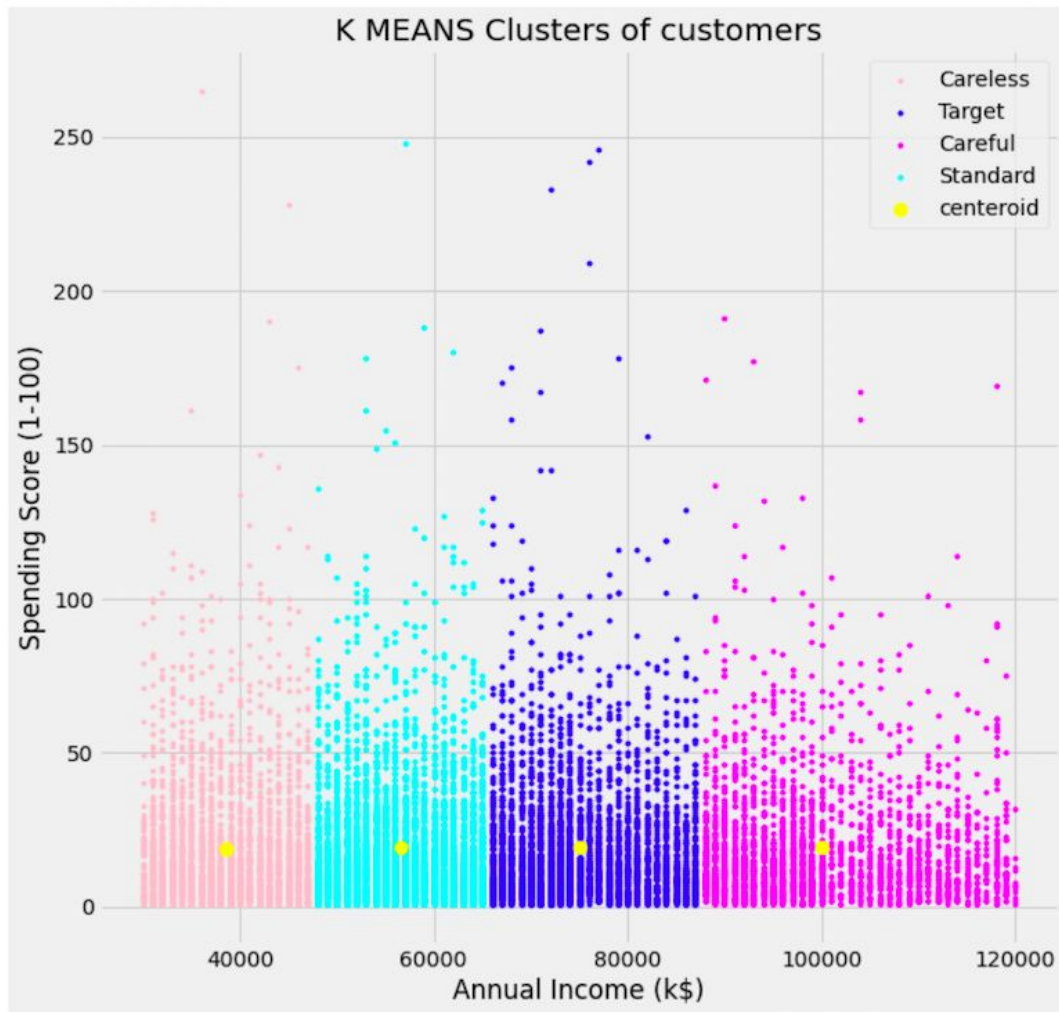
- Target
- Standard
- Careless
- Careful

Groceries + from stores you love + delivered to your doorstep + in as little as an hour =

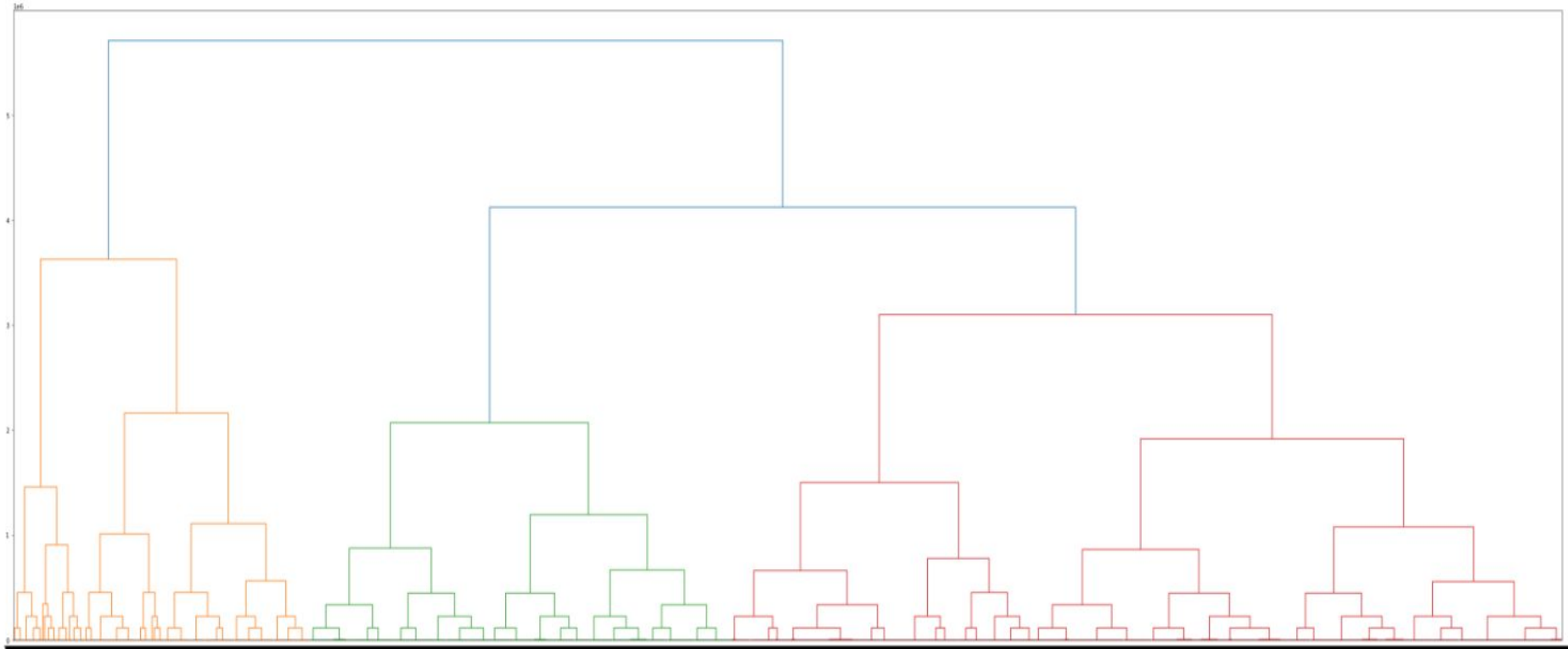# Density Based Clustering
K-Means

K MEANS Clusters of customers

Grouping the customers based on their Annual Income and Spending Score using K MEANS

Dendrogram (Hierarchical Clustering)

Results of Analysis

Customer Segmentation

- The highest incomes came from middle aged and older customers

- Females were greater represented in the dataset (female was 58.63%, male was 41.37%)

- Ages 50-65 were our most frequent Instacart users

- Income distribution ranged from 30-120k with 50-70k being the most common income among our users.

- Males and females had incredibly similar spending scores and spending habits
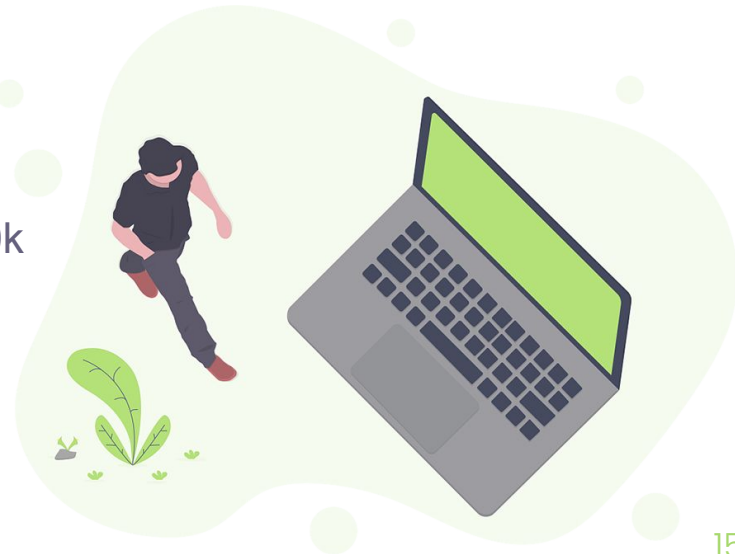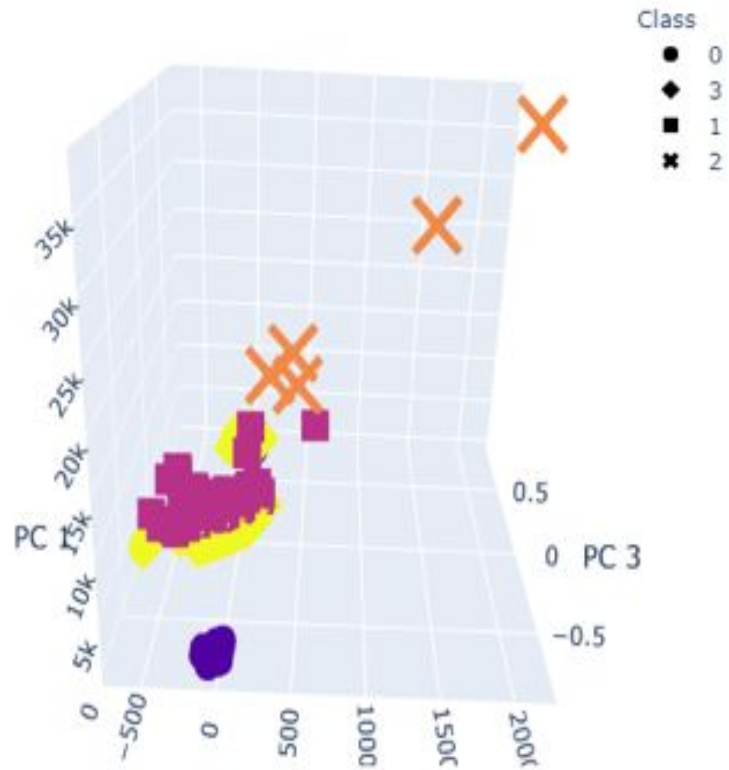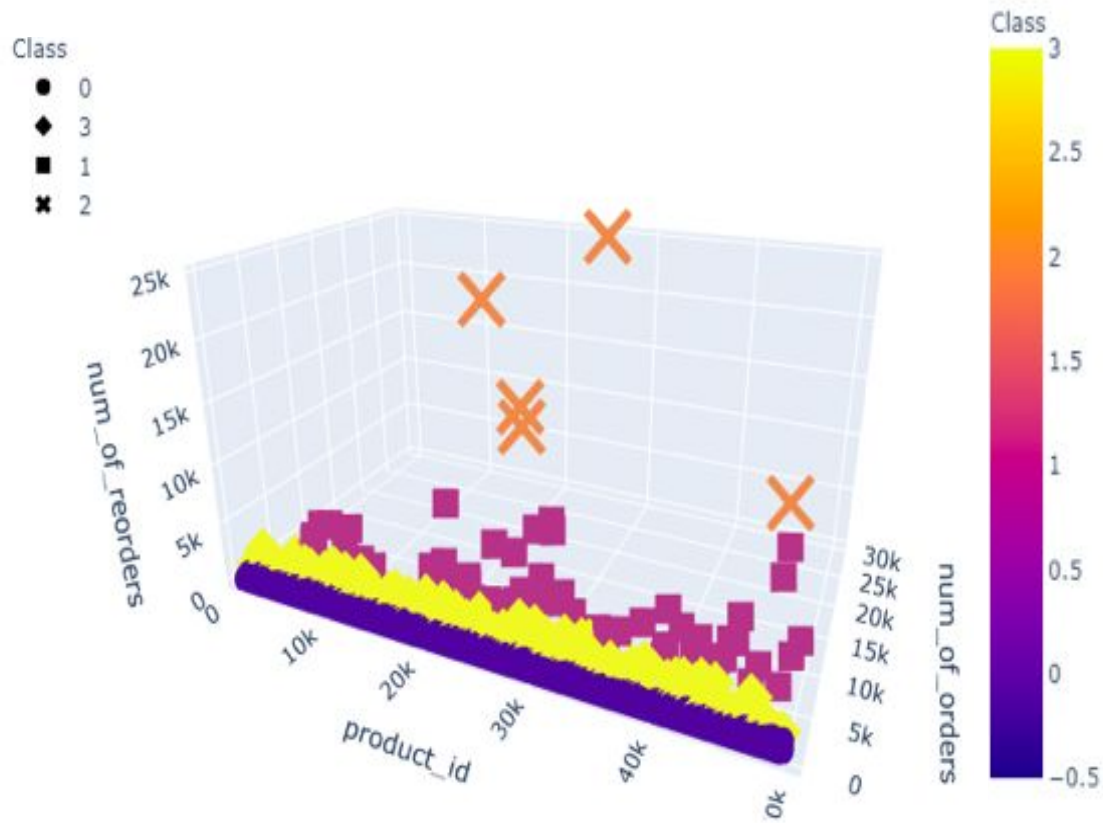
Machine Learning

Product Segmentation:

Utilizing KMeans, we hope to be successful in clustering and

categorizing the Instacart products into four main groups:

- Class 0 - Never Reordered

- Class 1 - Occasionally Reordered

- Class 2 - Often Reordered:
  "Popular"

- Class 3 - Frequently Reordered:
  "High Demand"

Groceries + from stores you love + delivered to your doorstep + in as little as an hour =

Product Clustering

Deep Learning Neural Networks to
Predict Product Reordering

	order_dow	order_hour_of_day	days_since_prior_order	add_to_cart_order	reordered	num_of_reorders	num_of_orders	department	aisle
0	5	0	15.0	1	1	6180	8882	produce	fresh fruits
1	5	0	15.0	2	1	3831	6083	produce	packaged vegetables fruits
2	5	0	15.0	3	1	2461	3595	produce	packaged produce
3	5	0	15.0	4	0	211	387	meat seafood	hot dogs bacon sausage
4	5	0	15.0	5	0	208	317	produce	fresh fruits
...	...	...	...	...	...	...	...	...	...
1864657	2	9	7.0	16	1	146	275	snacks	nuts seeds dried fruit
1864658	2	9	7.0	17	1	699	1120	bakery	bread
1864659	2	9	7.0	18	1	224	393	snacks	chips pretzels
1864660	2	9	7.0	19	0	13	37	pantry	pickled goods olives
1864661	2	9	7.0	20	0	290	714	pantry	oils vinegars

1812546 rows × 9 columns

Training the Model

```
Model: "sequential_1"
_____
Layer (type)                 Output Shape              Param #
=================================================================
dense_6 (Dense)              (None, 570)               108870
_____
dense_7 (Dense)              (None, 950)               542450
_____
dense_8 (Dense)              (None, 950)               903450
_____
dense_9 (Dense)              (None, 950)               903450
_____
dense_10 (Dense)             (None, 950)               903450
_____
dense_11 (Dense)             (None, 1)                 951
=================================================================
Total params: 3,362,621
Trainable params: 3,362,621
Non-trainable params: 0
_____
```

- **Received an accuracy of 70.5%**

```
8114/8114 - 88s - loss: 0.5889 - accuracy: 0.7007
Loss: 0.5889468789100647, Accuracy: 0.7007383704185486
```
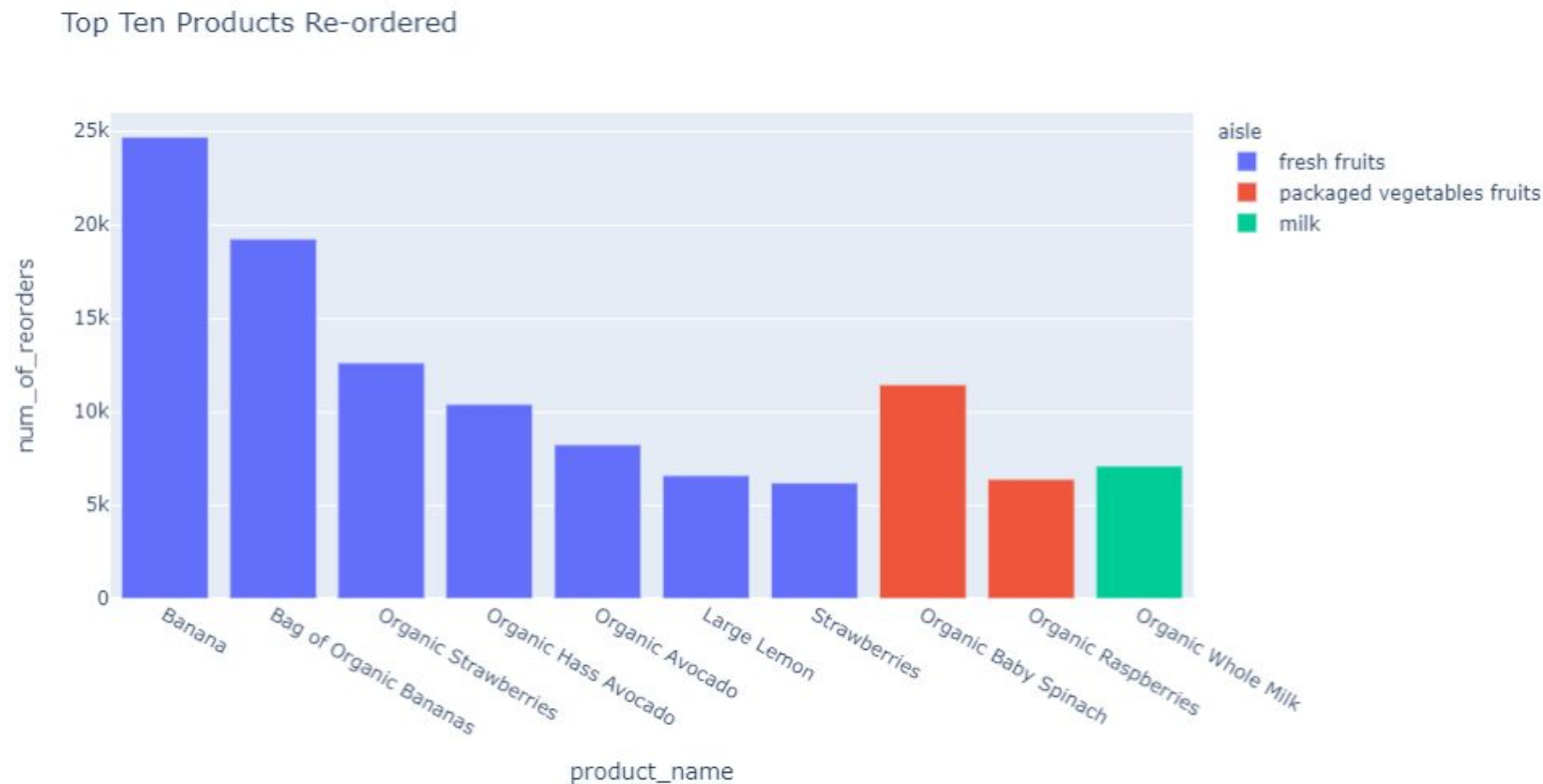
Top 10 Ordered Products

Top Ten Products Ordered

Top 10 Reordered Products
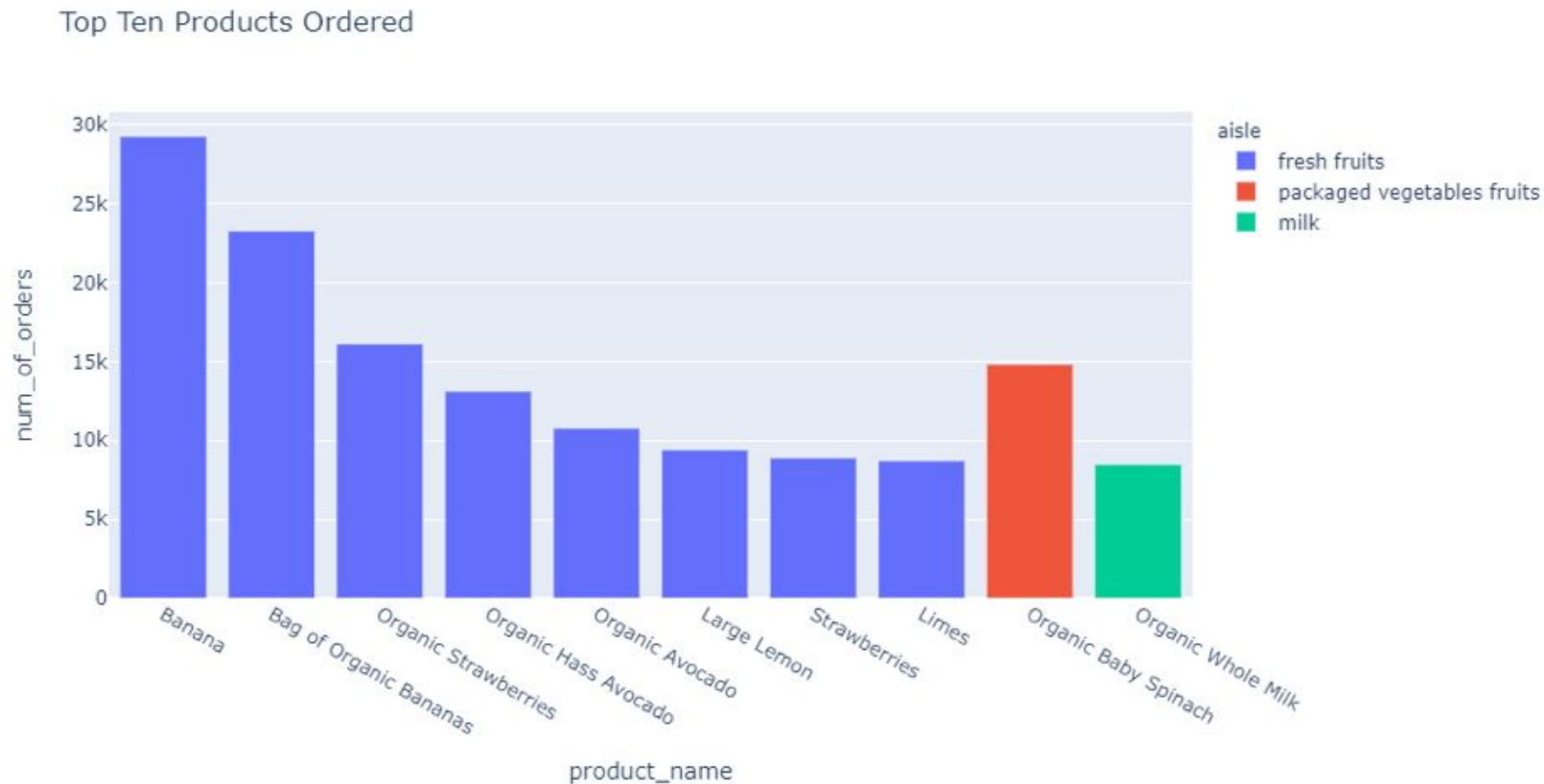
Top Ten Products Re-ordered

Results of Analysis

Product Segmentation

- Busiest times are on weekends:
  - Saturday(12-2 PM) and on Sundays(10-12 PM)
- Products that were not reordered frequently:
  - Corn Starch, Paprika, Bay leaves, Cayenne Pepper, Ground Ginger
- Aisles that had the least orders:
  - Beauty, Baby Accessories, Vitamin Supplements, First Aid, Kitchen Supplies
- Top products ordered together:
  - Bags of Organic Bananas + Organic Avocados
  - Bags of Organic Bananas + Organic Strawberries
  - Spinach + Bananas
- Top Departments with most orders:
  - Produce and Dairy Eggs and Beverages

Technologically Speaking

- Methodologies:
    - Jupyter Notebook, PGAdmin, Tableau, Google Slides, SQL, Python, Postgres, AWS, VS Code, GitHub
- Algorithms:
    - K-Means Clustering
    - Neural Networks
    - DB Scan
    - KNN
    - Dendrogram (Hierarchical Clustering)
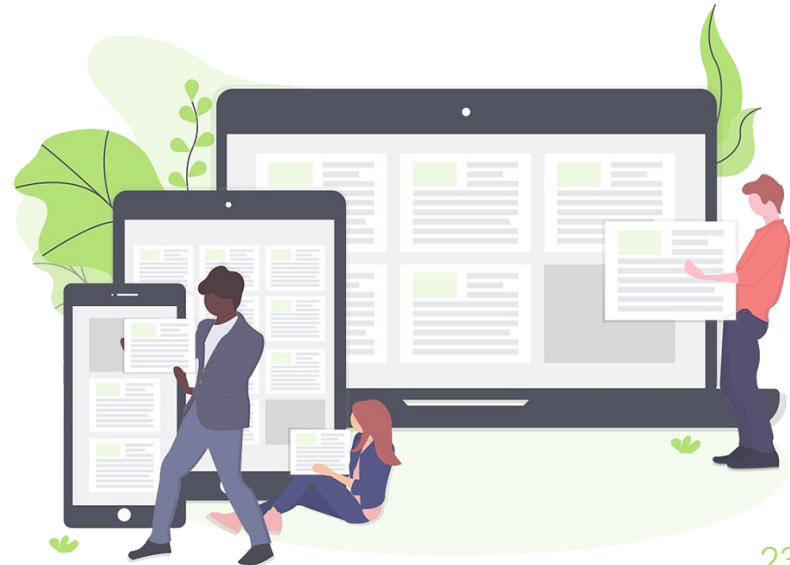
What We May Have Done Differently

- Customer buying prediction model

- Build an API (time restrictions)

- Choosing a project topic based on available datasets instead of choosing a topic and having to scramble to find a suitable dataset
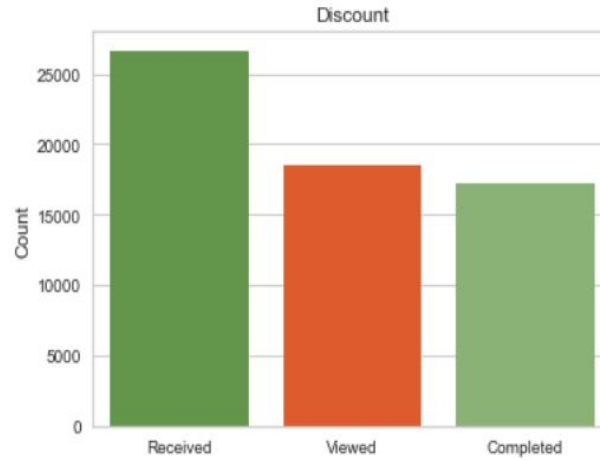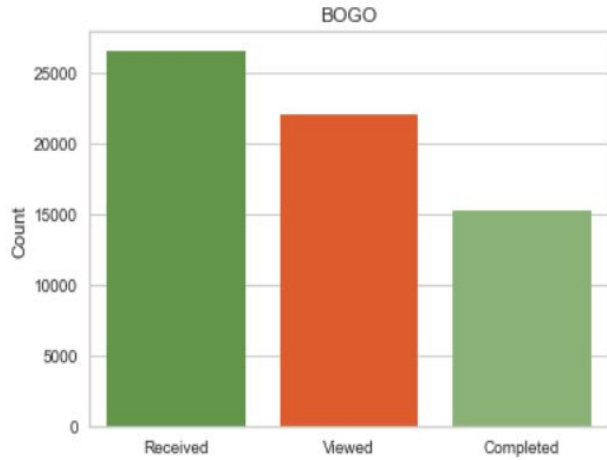
Recommendations for Future Analysis

- Combine more features from the customers dataset so that we can have a better prediction for specific customer groups
- Promos for under engaged demographics
- Promos for targeted customers
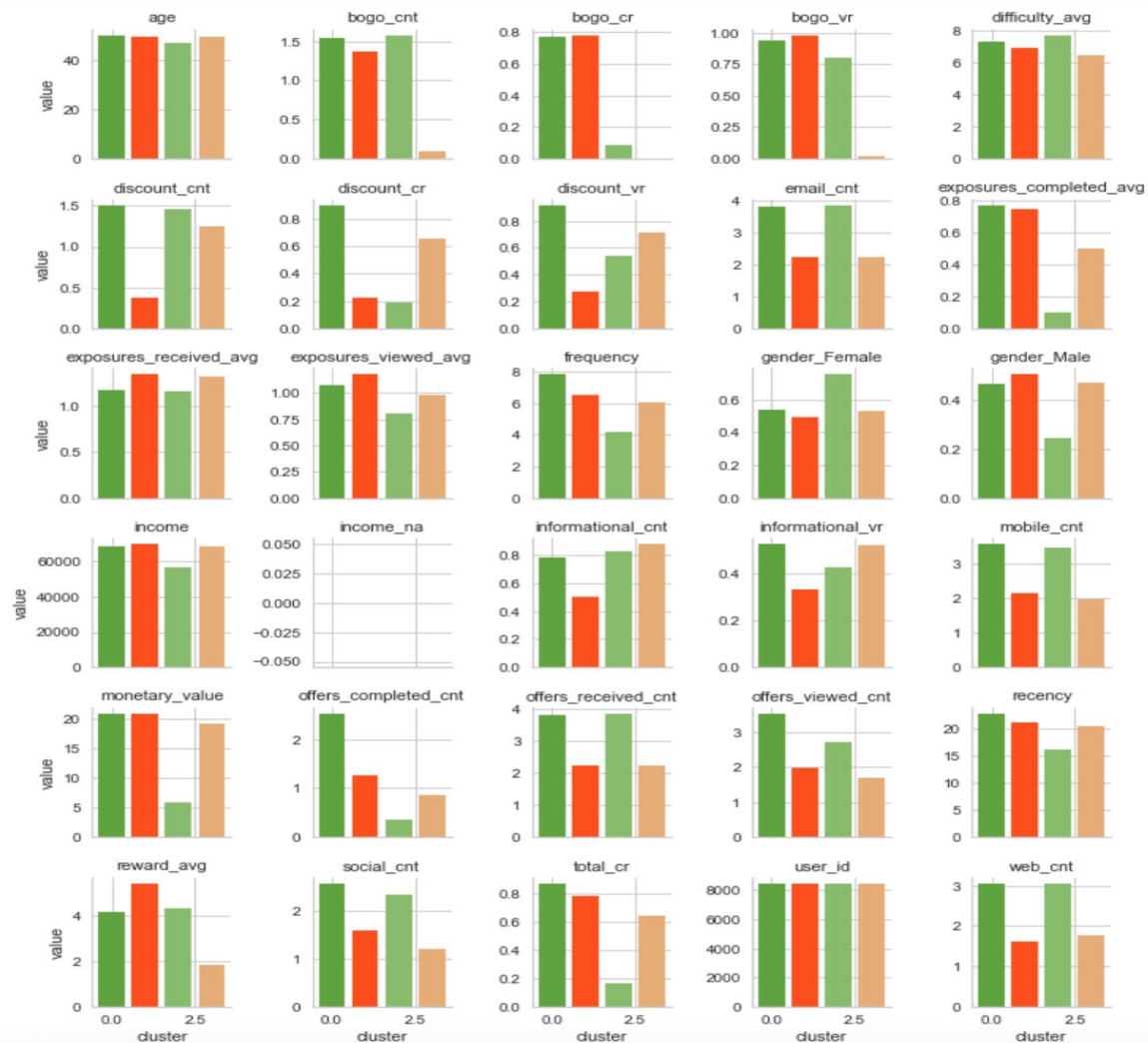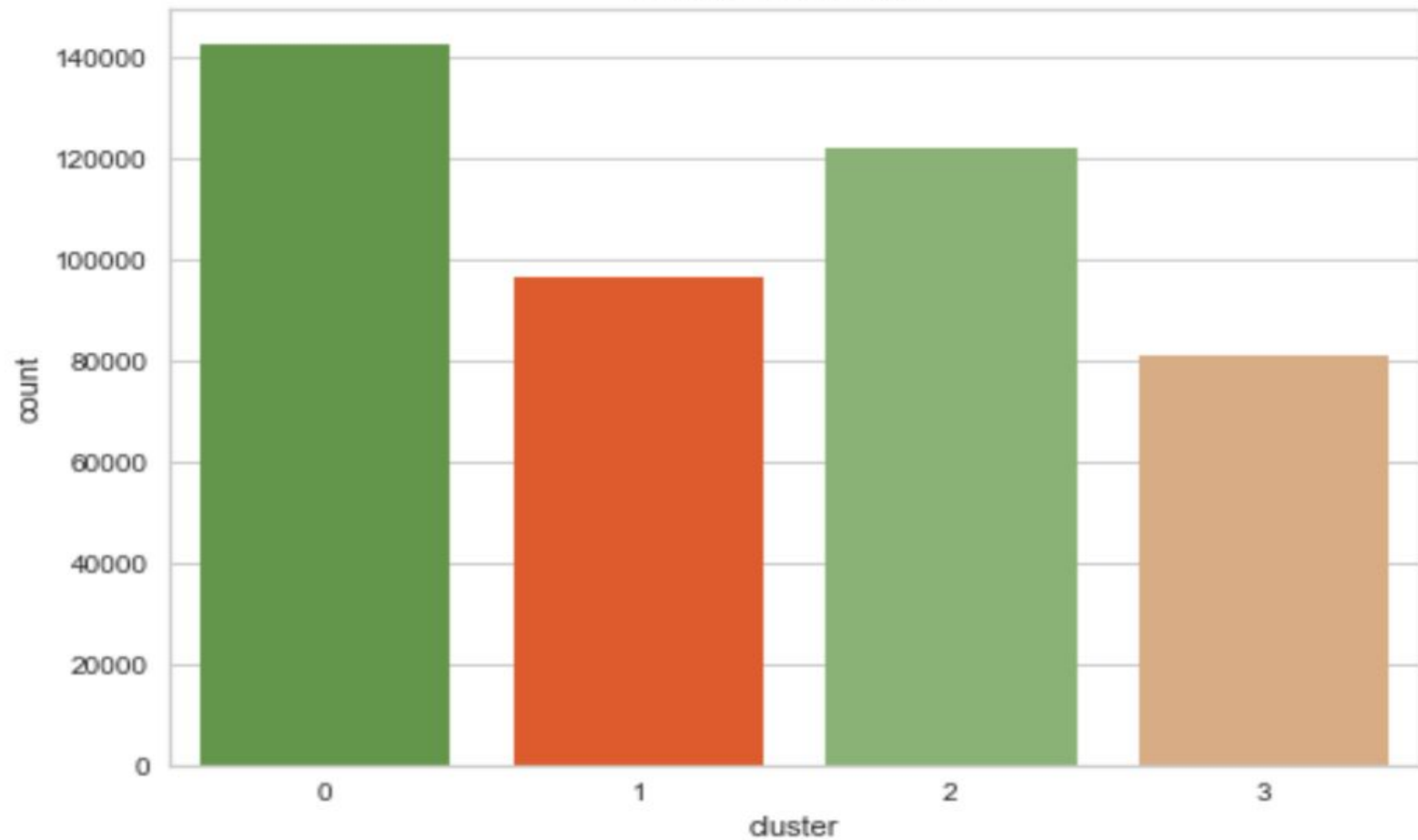- Promos for targeting low selling products

Promotions

Offer conversion funnel by type
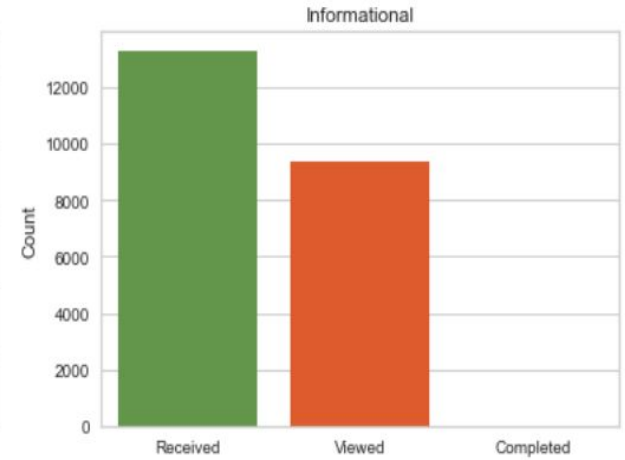
P R O M O S

Customers by cluster

Team Presentation

Ketaki Pradhan
Team Member

Leora Talmor
Team Member

Surekha Chandramouli
Team Member

Edward Rivera
Team Member

Allison Corrales
Team Member