

CUSTOMER BEHAVIOUR AND MARKET BASKET ANALYSIS



Presenter - Alka Bhambhu

AGENDA

BACKGROUND

Market Growth
Current Market



OBJECTIVE

Goals
Research
Questions



DATA

Source &
Description
Limitation



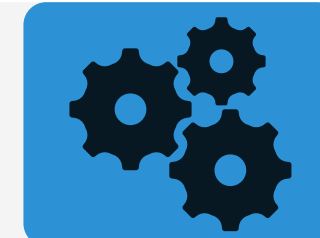
EXPLORATION

Data visualisation
Customer segmentation
Association



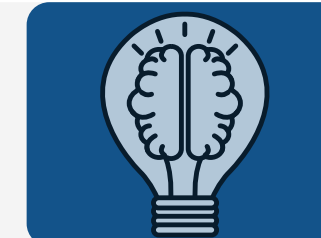
MODELING

Data Preparation
Performance



INFERENCE

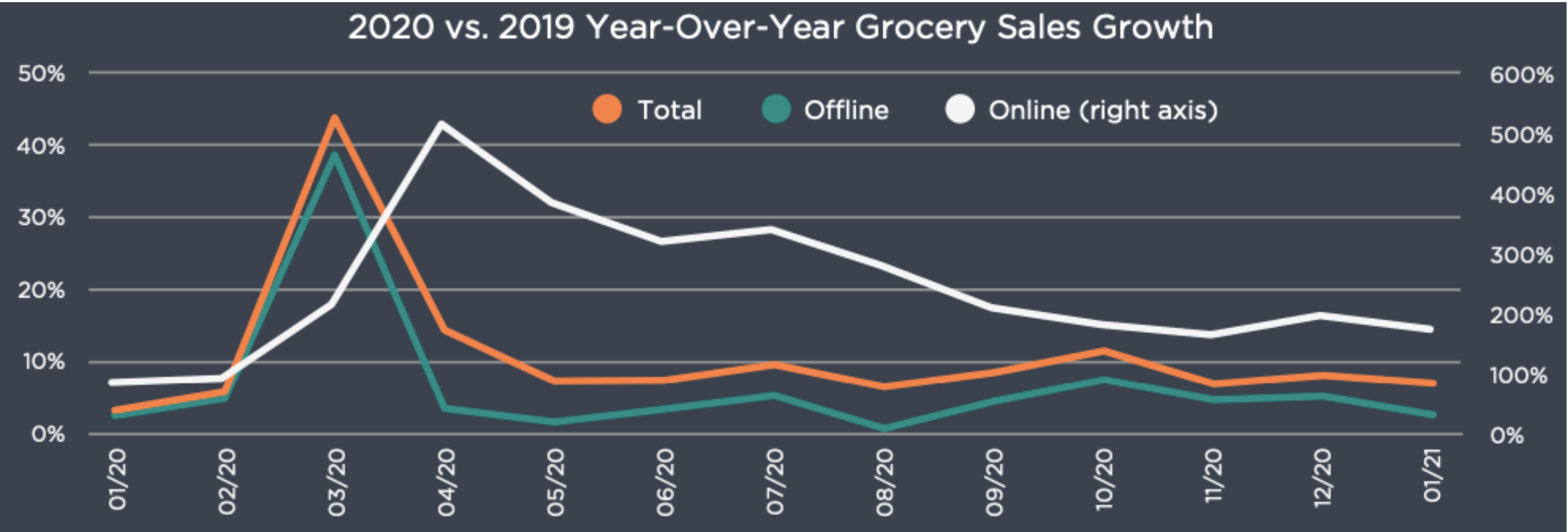
Recommendation
Future Work
Questions



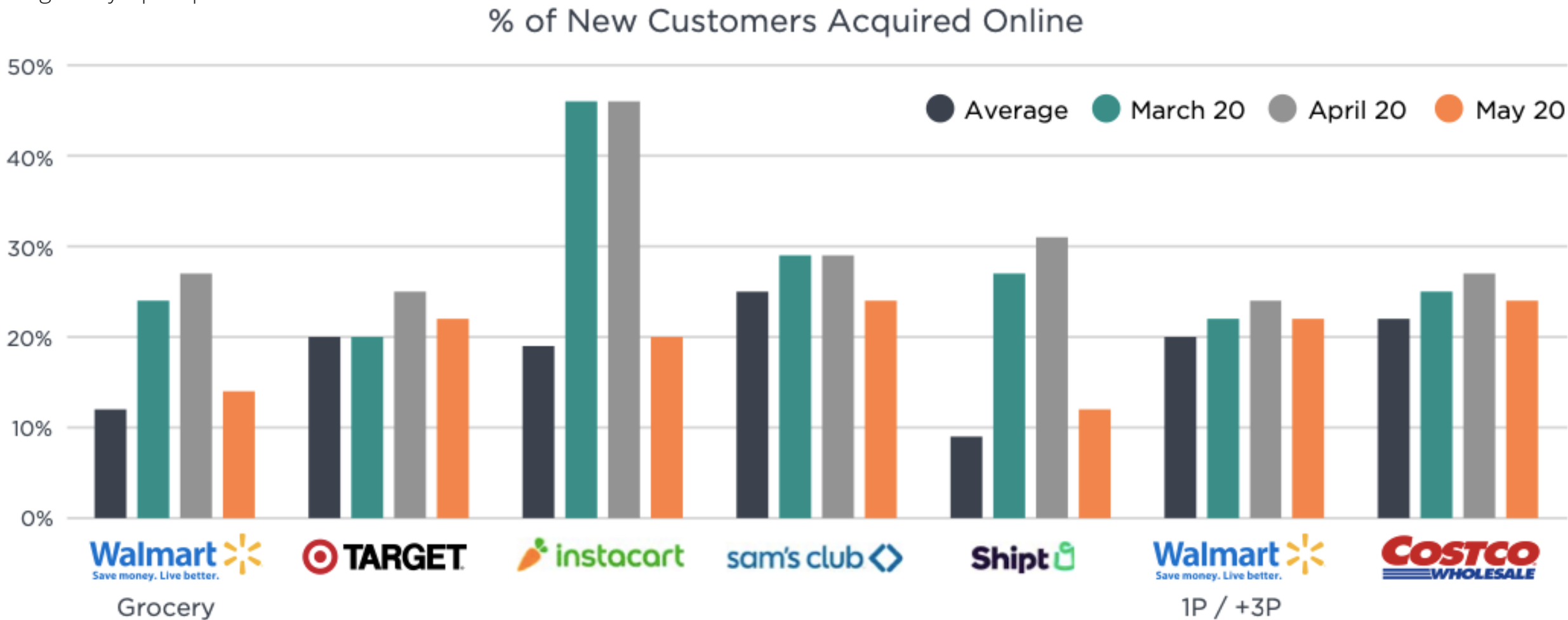
BACKGROUND

Market Growth
Current Market

Market Growth in 2020



https://1010data.com/media/3019/2021_1010data_stateofgroceryreport.pdf



https://1010data.com/media/3019/2021_1010data_stateofgroceryreport.pdf

Percentage of new customers acquired by Instacart



19%

The infographic consists of two light blue rounded rectangular boxes. The left box contains a white circle with '19%' inside, and the text 'Pre-pandemic' below it. The right box contains a white circle with '46%' inside, and the text 'During-pandemic' below it.

Pre-
pandemic

46%

During-
pandemic

Main reasons for ordering Groceries Online (%)

I mainly buy groceries online when I need to stock up on grocery supplies 21

I buy majority of my groceries by visiting the physical store, and use e-commerce mainly to top off my weekly or monthly grocery needs 21

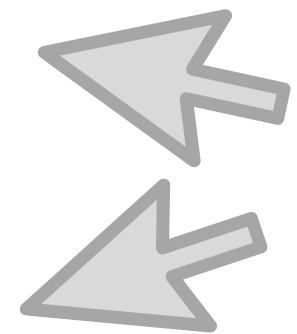
I use e-commerce as a primary way to fulfill all my weekly grocery needs 18

I buy groceries online only when I am time constrained and need my groceries fast 15

I buy groceries online to fulfill my daily meal needs (eg, pre-prepared meals for dinner or lunch) 14

I buy groceries online when I am hosting for an event or occasion (eg, Thanksgiving dinner, birthday party) 7

I use e-commerce when I want to discover and learn about new products 5



Importance of Online Interface Features (%)

■ Least important feature ■ Most important feature

Quick selection and checkout (with few clicks)

Easy access and reordering of past purchases

Accurate information on out-of-stock items

Ability to track progress of my order

Ability to make changes before/during picking

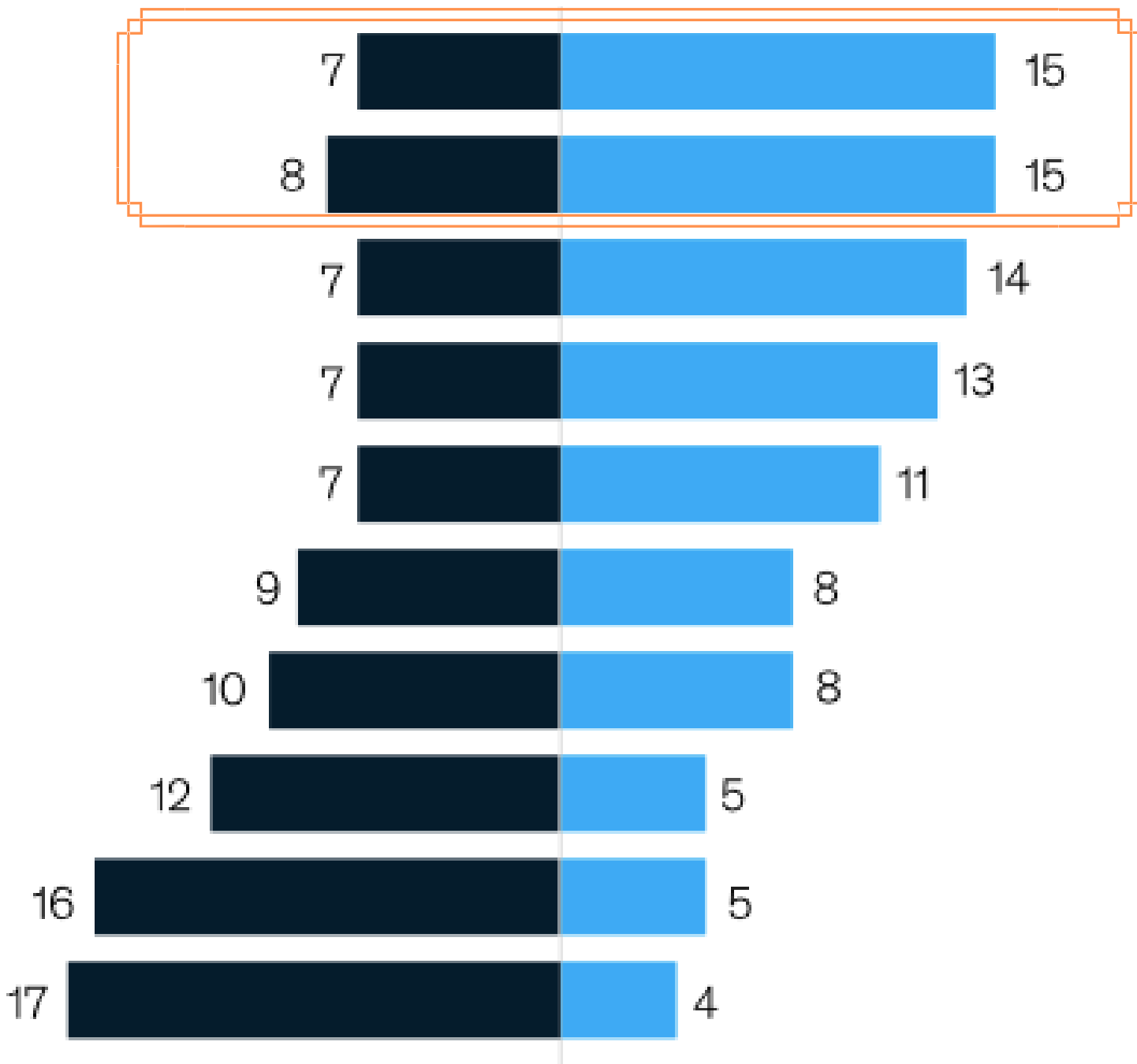
Ability to create customized lists and preferences

Display of promotions to help guide product selection

Useful product recommendations based on past purchases and items in current order

Chat feature with order picker

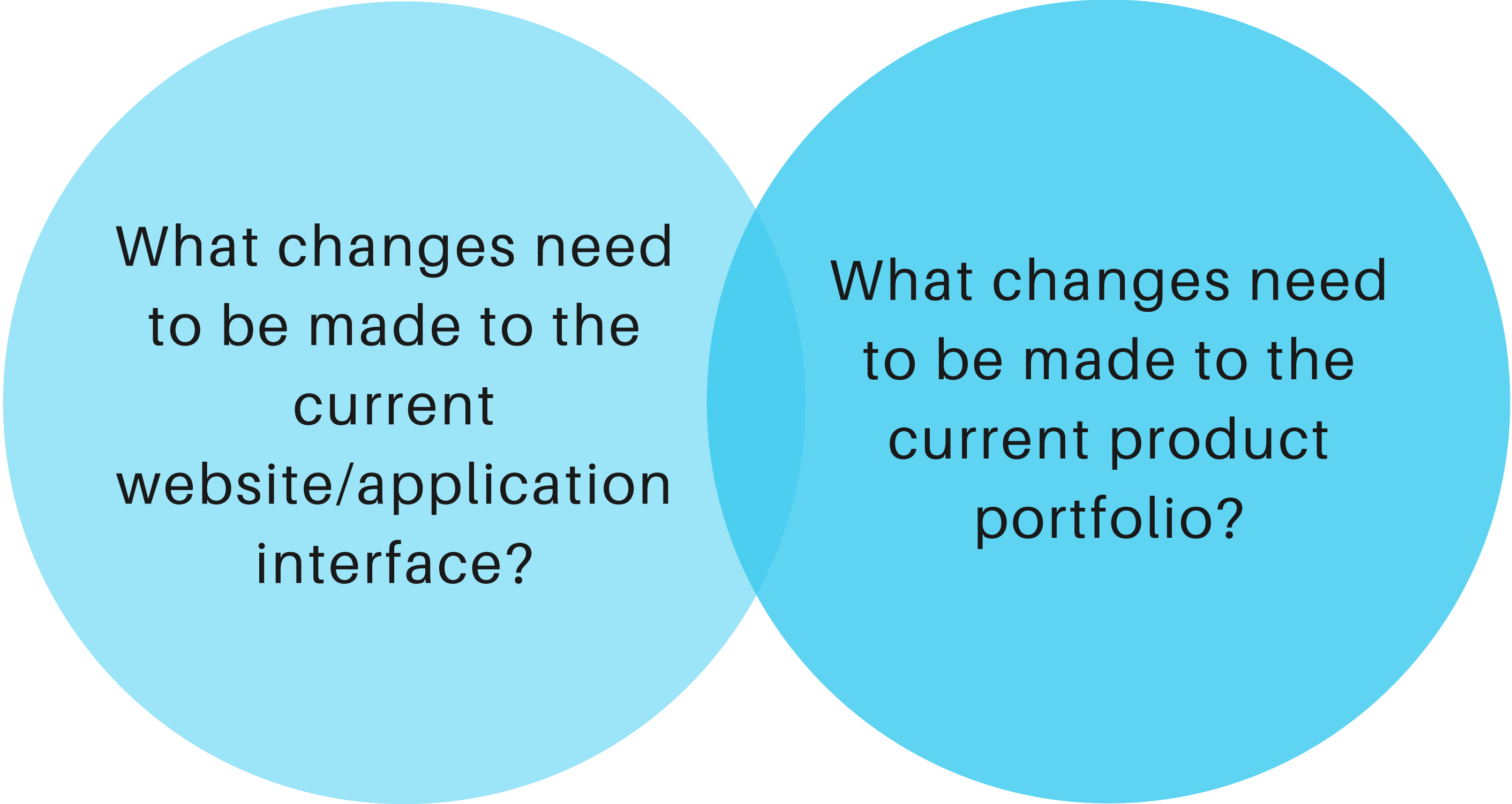
Other household members have ability to access and edit orders



OBJECTIVE

Research Questions
Goals

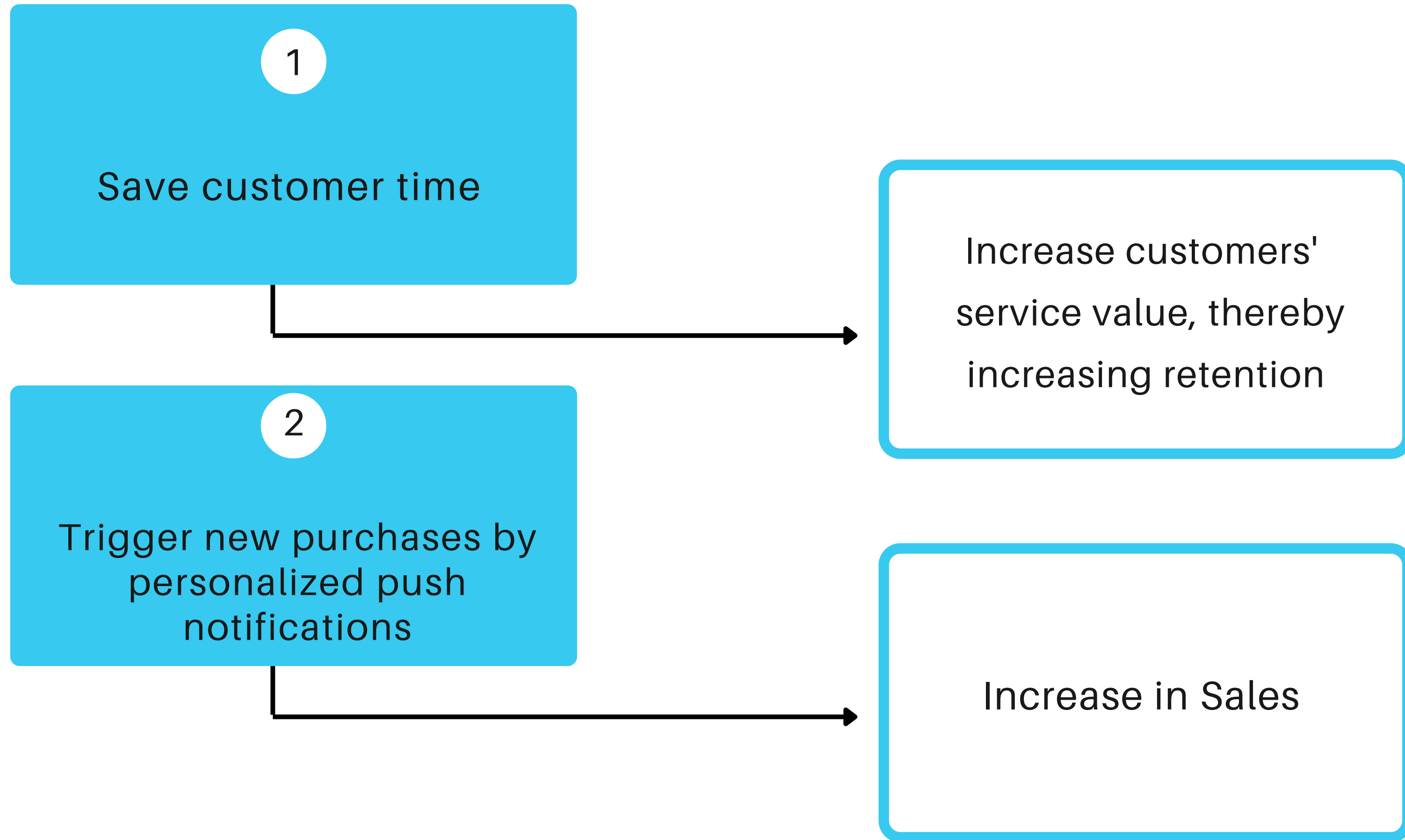
Research Question



What changes need
to be made to the
current
website/application
interface?

What changes need
to be made to the
current product
portfolio?

Determining a user's basket will help solve the following two business objectives:

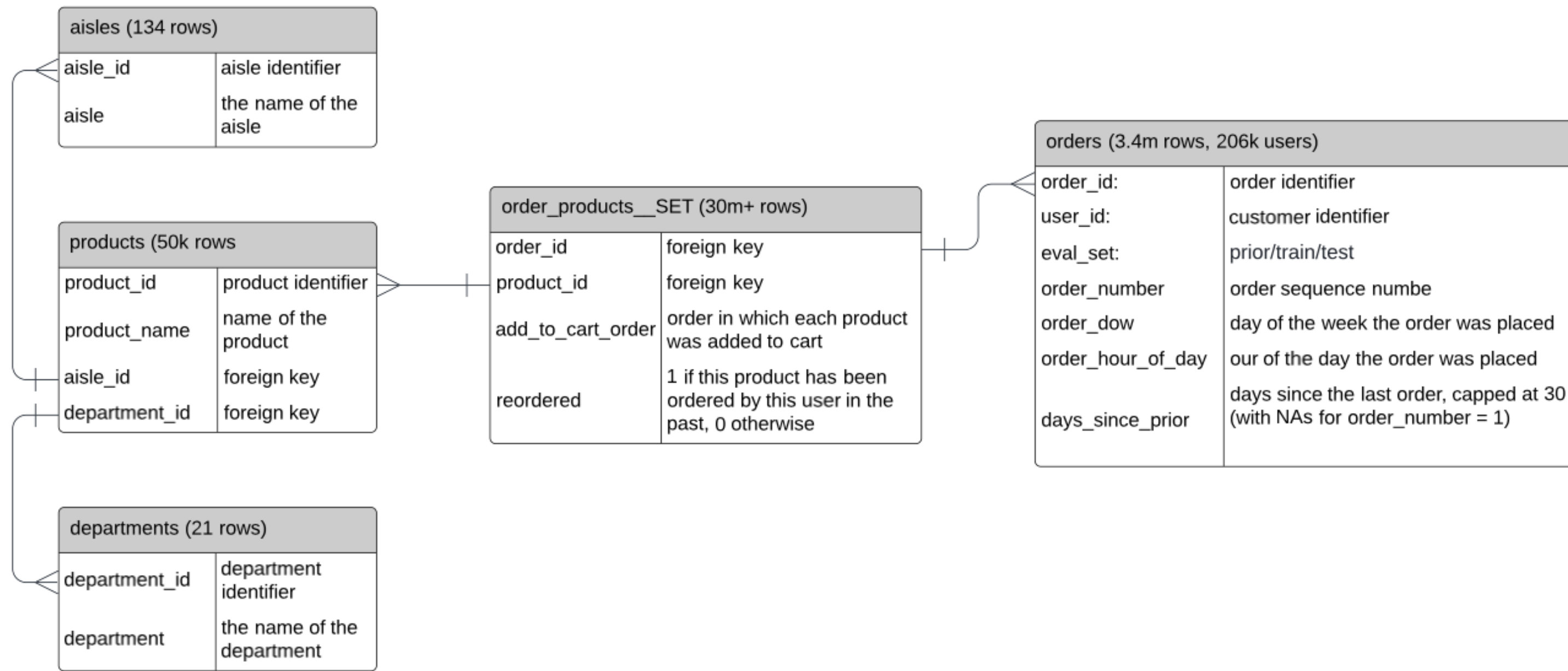


DATA

Source	Description	Limitation

SOURCE

The dataset contains a sample of over **three million** grocery orders in 2017 of more than **2,00,000** de-identified Instacart users. The information contains each user's order number between 4 and 100. The information provided is completely **anonymous** to protect the privacy of the users and retailers. User ID in the dataset is completely randomized and the only information provided is the user's sequence of order and products in that order. Similarly, only products ordered are provided and no retailer ID is stated.



where SET is one of the two following evaluation sets (eval_set in orders):

"prior": orders prior to that users most recent order (~3.2m orders)

"train": most recent orders (~131k orders)

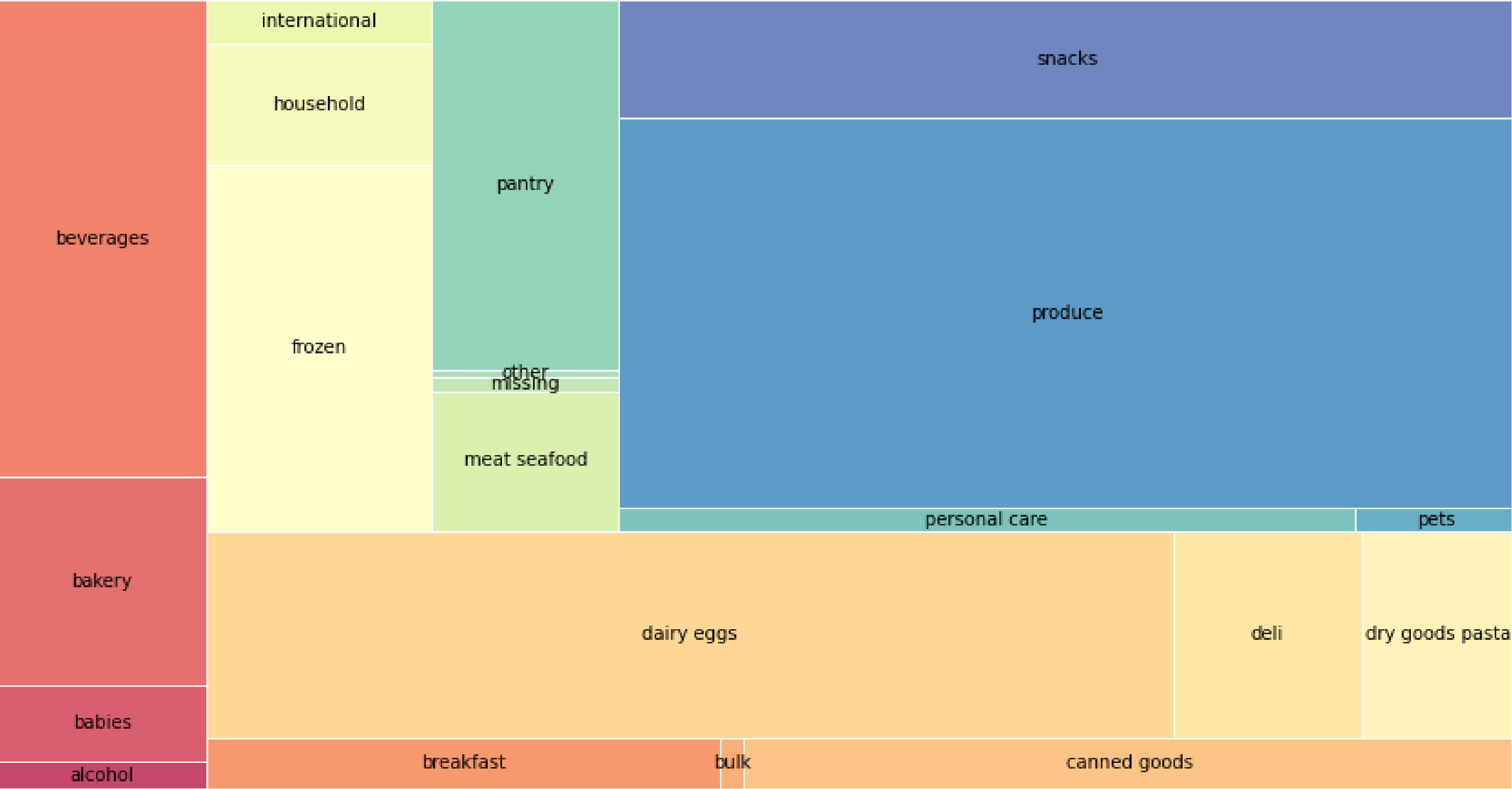
LIMITATION

- Limited RAM & Large Dataset
- 2017 Data
- No Demographic information

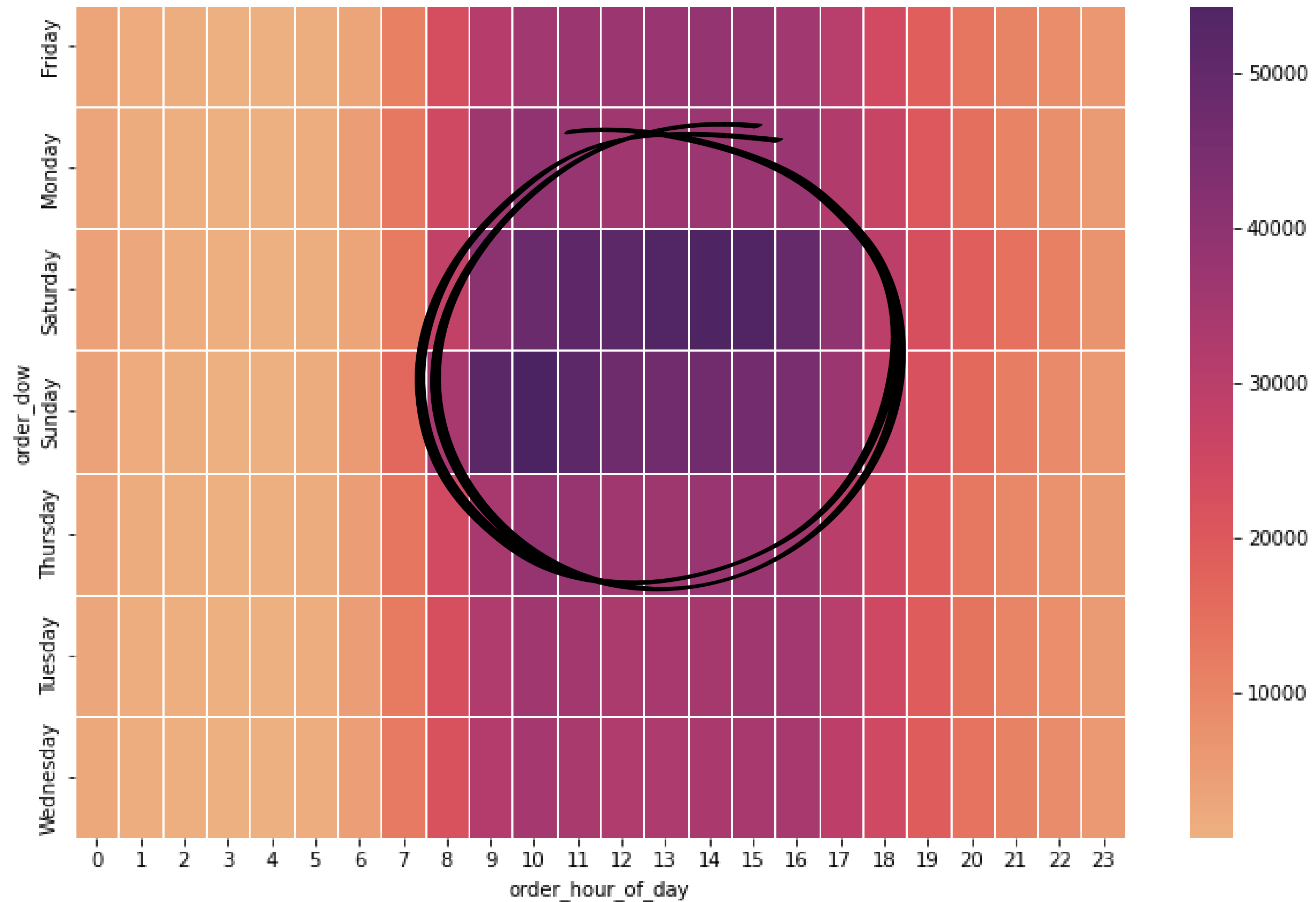
EXPLORATION

Data visualisation
Customer segmentation
Association

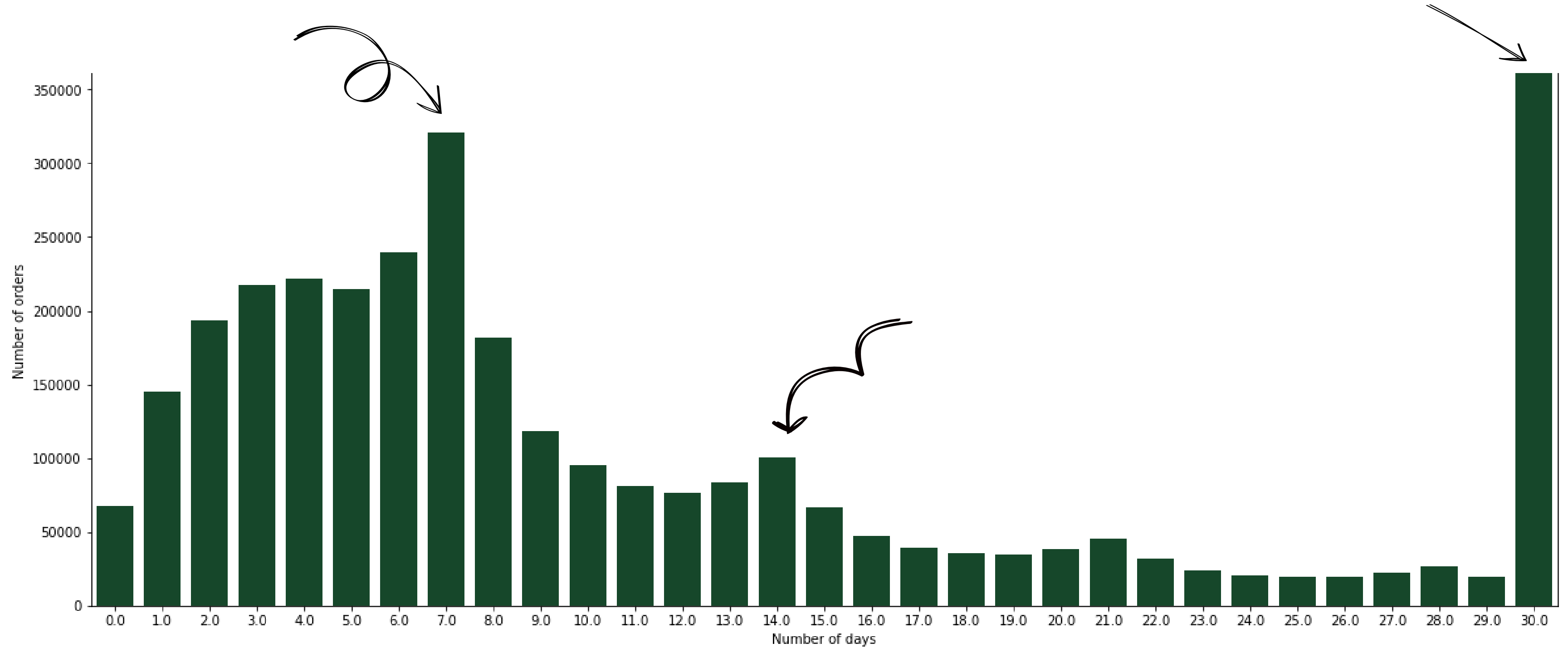
Department wise Order Distribution



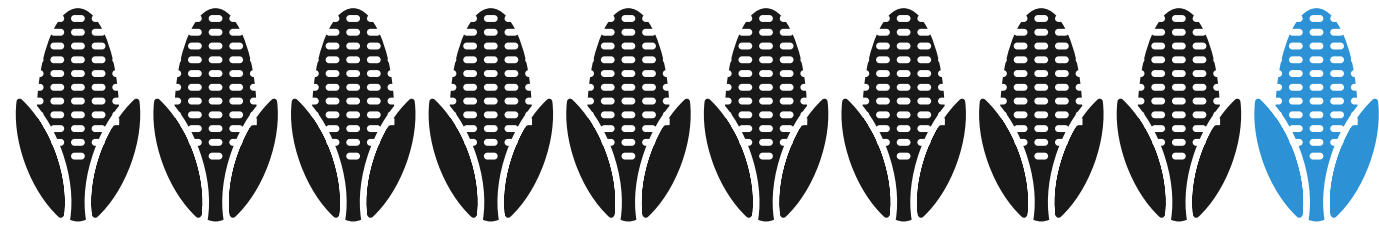
Order Distribution based on hour & days



Number of Days Since Prior Order



Organic Products



10% OF PRODUCTS



31% OF ORDERS

63%
Reorder Ratio

CUSTOMER SEGMENTATION

In K-Mean Clustering the similarity between data points is derived based on the distance between data points and the centroid of clusters.

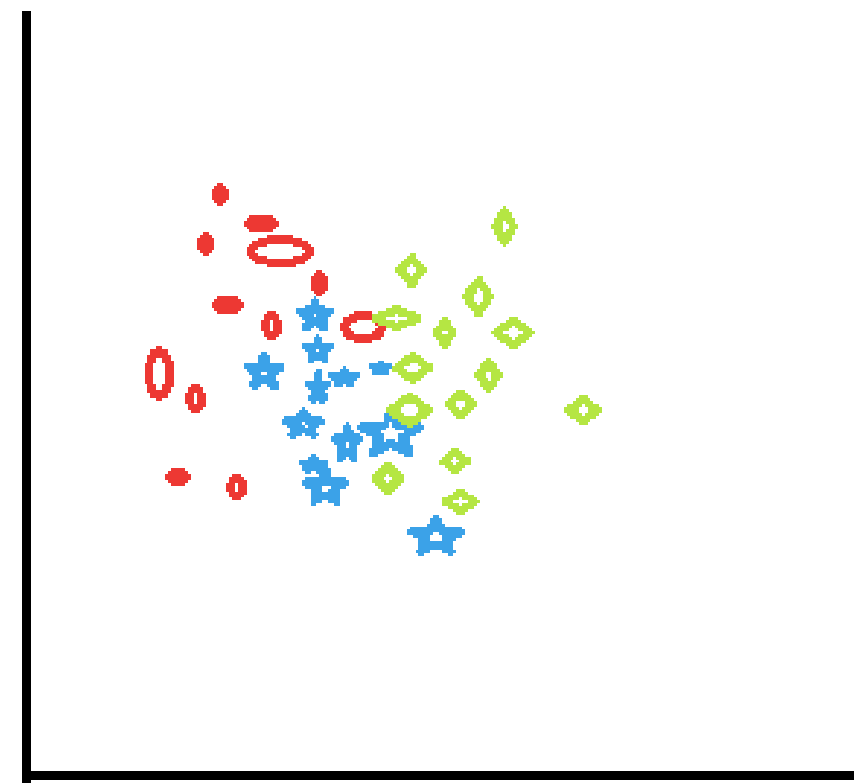


fig 1: before applying k-means clustering

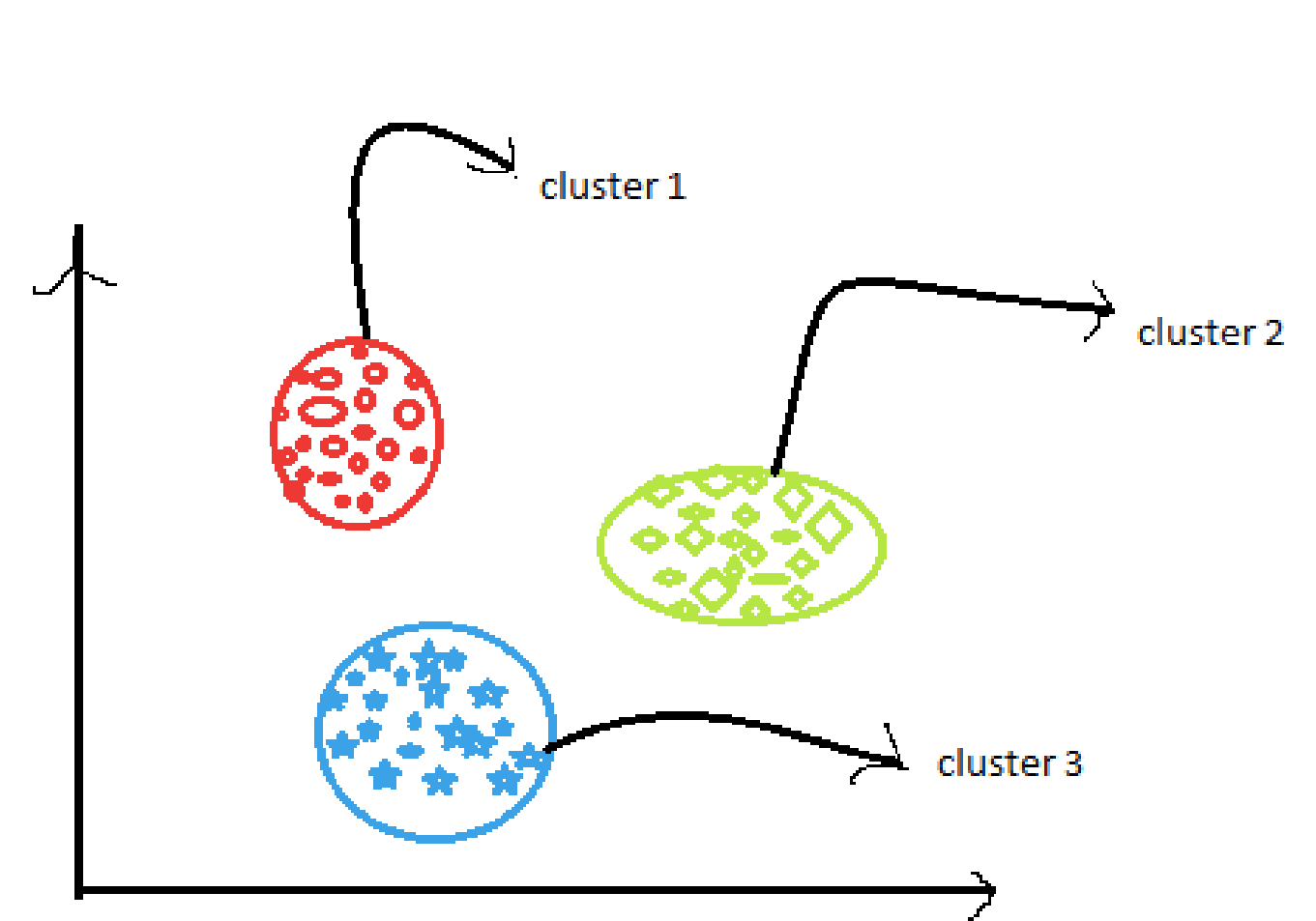


fig 2: After applying K-means clustering

Cluster 3

Preference - frozen produce, oils, vinegar, and nuts, seeds & dried fruits

Order frequency - Less

Cart Size - Small

Cluster 0 (In 20's or younger)

Preference - seltzer, sparkling water, soft drinks, juices, candies & chocolate

Order frequency - Low

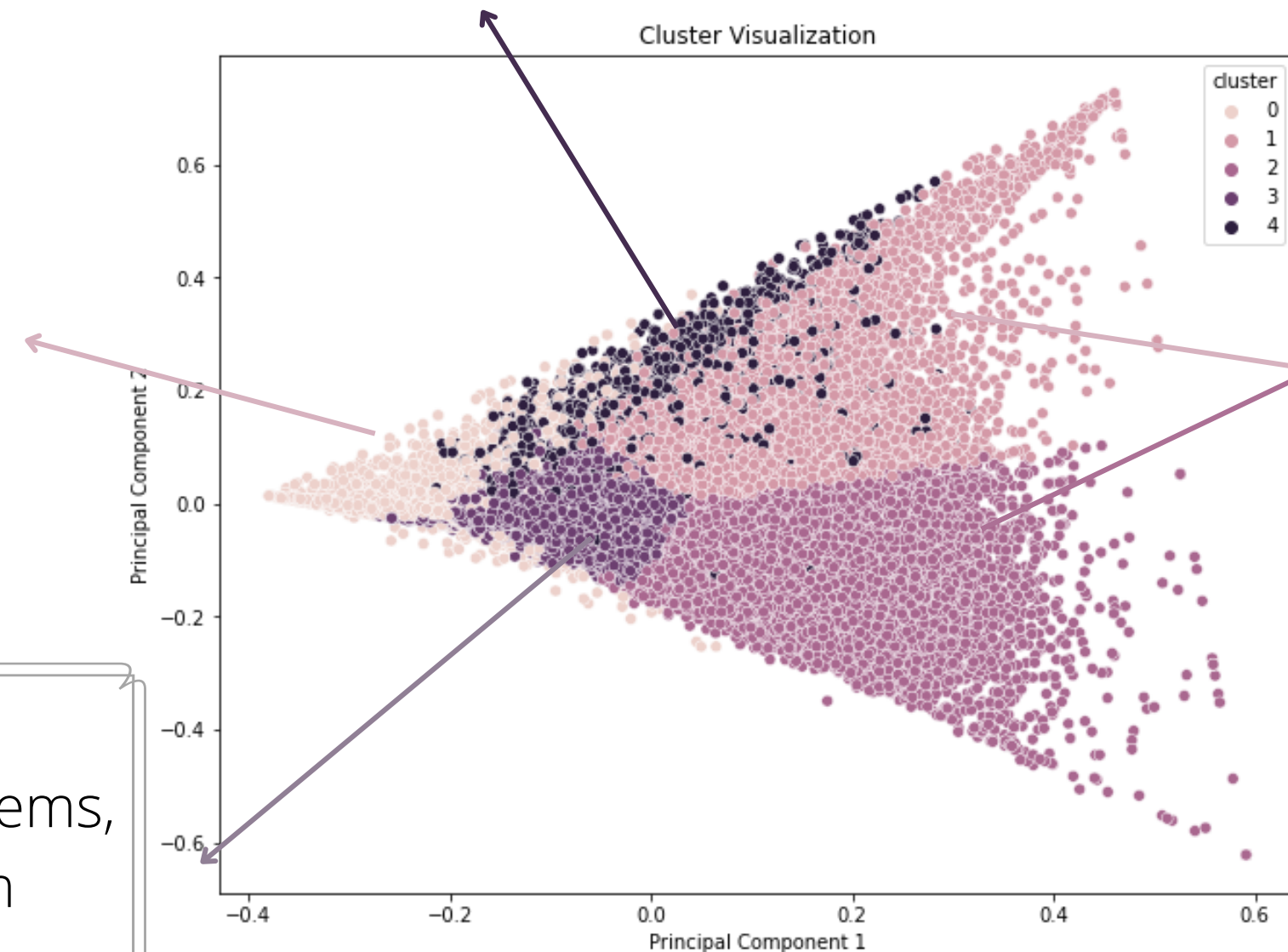
Cart Size - Small

Cluster 3

Preference - refrigerated items, frozen meals, and ice cream aside from staples

Order frequency - Less

Cart Size - Large



Cluster 1 & 2 - Families

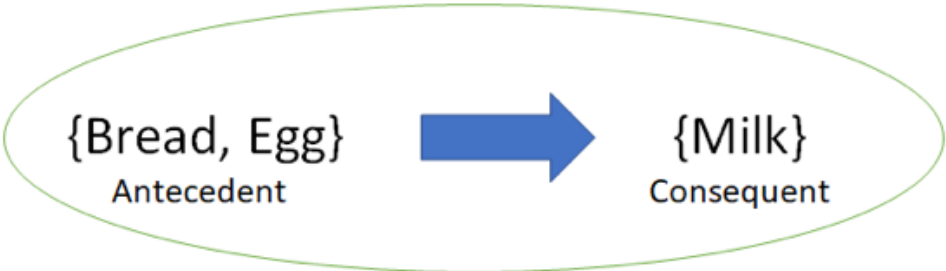
Preference - Regular grocery items such as packaged vegetables, fruits, bread, milk, egg & yogurt. Cluster 2 differs from cluster 1 in terms of its preference for fresh herbs.

Order frequency - High

Cart Size - Large

PRODUCT ASSOCIATION

Association rules are 'if-then' statements that reflect the probability of relationships among the dataset.



Itemset = {Bread, Egg, Milk}

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	length
(Lime Sparkling Water)	(Sparkling Water Grapefruit)	0.013527	0.022917	0.003886	0.287278	12.535775	0.003576	1.370918	
(Organic Lemon)	(Organic Hass Avocado)	0.023709	0.066632	0.005860	0.247175	3.709565	0.004280	1.239821	
(Organic Raspberries)	(Organic Strawberries)	0.041283	0.080619	0.009827	0.238035	2.952579	0.006499	1.206591	
(Organic Kiwi)	(Organic Strawberries)	0.014007	0.080619	0.003253	0.232210	2.880332	0.002123	1.197438	
(Organic Blueberries)	(Organic Strawberries)	0.024643	0.080619	0.005464	0.221719	2.750198	0.003477	1.181297	
(Organic Cilantro)	(Limes)	0.016881	0.037508	0.003669	0.217363	5.795179	0.003036	1.229807	
(Organic Cucumber)	(Organic Hass Avocado)	0.021952	0.066632	0.004737	0.215803	3.238743	0.003275	1.190222	
(Organic Whole String Cheese)	(Organic Strawberries)	0.017875	0.080619	0.003786	0.211836	2.627615	0.002345	1.166485	
(Raspberries)	(Strawberries)	0.015473	0.039087	0.003199	0.206748	5.289370	0.002594	1.211359	
(Organic Tomato Cluster)	(Organic Hass Avocado)	0.016552	0.066632	0.003364	0.203237	3.050148	0.002261	1.171450	
(Organic Red Bell Pepper)	(Organic Strawberries)	0.014870	0.080619	0.003020	0.203092	2.519151	0.001821	1.153685	
(Organic Cucumber)	(Organic Strawberries)	0.021952	0.080619	0.004312	0.196428	2.436493	0.002542	1.144118	
(Michigan Organic Kale)	(Organic Baby Spinach)	0.018345	0.073193	0.003589	0.195662	2.673226	0.002247	1.152260	
(Organic Small Bunch Celery)	(Organic Baby Spinach)	0.016830	0.073193	0.003262	0.193852	2.648494	0.002031	1.149673	
(Organic Large Extra Fancy Fuji Apple)	(Organic Strawberries)	0.022403	0.080619	0.004300	0.191920	2.380569	0.002493	1.137734	
(Organic Cucumber)	(Organic Baby Spinach)	0.021952	0.073193	0.004127	0.187990	2.568402	0.002520	1.141373	
(Organic Garlic)	(Organic Yellow Onion)	0.029242	0.030969	0.005464	0.186840	6.033198	0.004558	1.191685	
(Organic Raspberries)	(Organic Hass Avocado)	0.041283	0.066632	0.007615	0.184462	2.768382	0.004864	1.144482	
(Organic Zucchini)	(Organic Baby Spinach)	0.028263	0.073193	0.005189	0.183607	2.508526	0.003121	1.135246	
(Apple Honeycrisp Organic)	(Organic Hass Avocado)	0.024482	0.066632	0.004436	0.181203	2.719465	0.002805	1.139926	

A prominent pattern in the above rule is the lack of association between organic and non-organic products. This shows that customers who buy organic products have stronger preference for organic produce over non-organic produce. Such customers will appreciate and prefer the wide variety & selection of organic products on Instacart.

MODELING

Data Preparation
Performance

FEATURE EXPANSION

User Specific:

- total_orders_per_user - Number of orders by a user
- total_products_per_user - Number of products bought by a user
- unique_products_per_user - Number of unique products bought by a user
- total_reorders_per_user - Number of reordered products by a user
- reorder_ratio_per_user - Average of reorders by a user
- avg_gap_in_orders_per_user - Average number of days since prior order

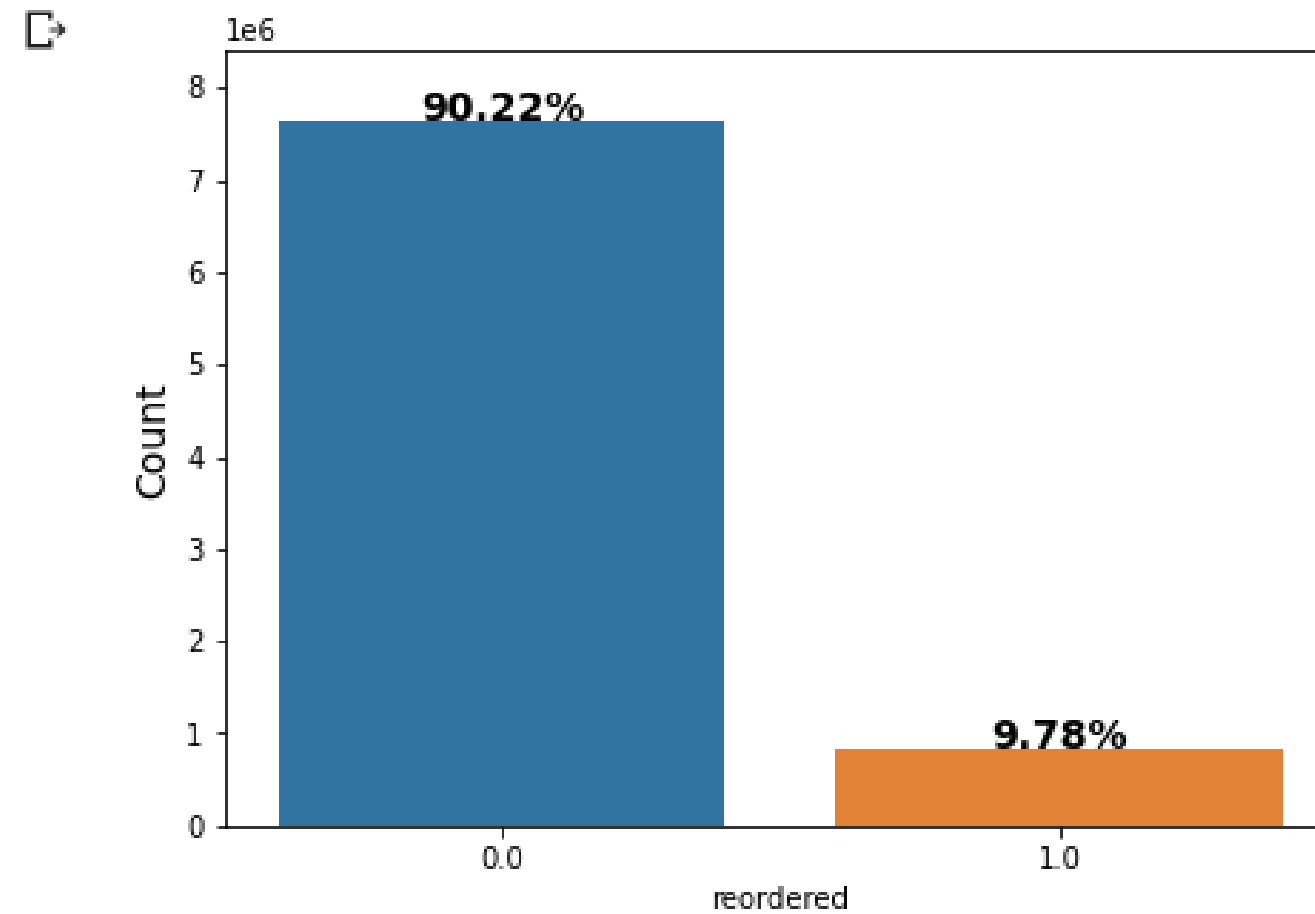
Product Specific:

- prod_reorder_ratio - Average reorders for a product
- num_orders_for_prod - Total number of orders for a product
- avg_cart_order_of_product - Average sequence in which a product is added to the cart
- is_organic - 1 if a product is organic, else 0
- is_vegan - 1 if a product is vegan, else 0
- is_Gluten-free - 1 if a product is gluten-free, else 0

User-product specific :

- total_prod_by_user - Number of orders per product by a user
- prod_reorder_by_user - Number of reorders per product by a user
- prod_reorder_ratio_by_user - Average reorders per product by a user
- days_since_prior_prod_user - Average days since prior order for a product by a user

CLASS IMBALANCE



Random Undersampling

```
#after randomly undersampling  
print(train_x.shape, train_y.shape)  
print(test_x.shape, test_y.shape)
```

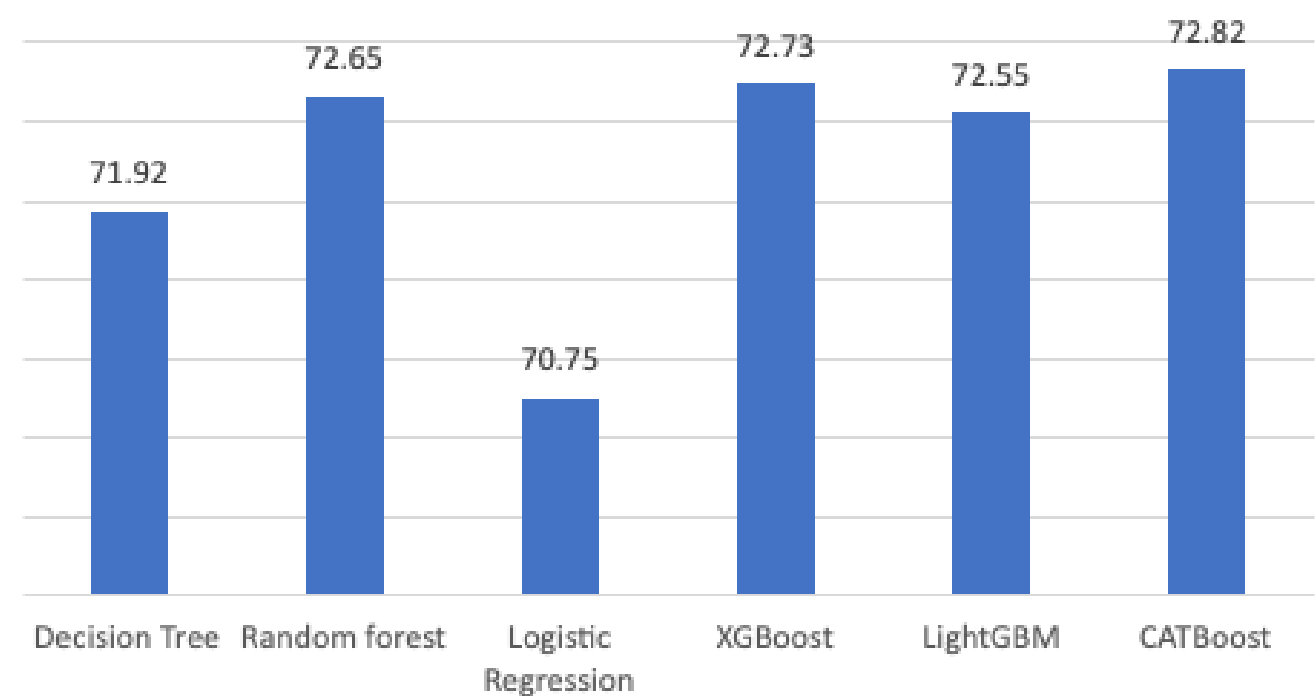
```
(1161314, 25) (1161314,)  
(496334, 25) (496334,)
```

ALGORITHMS

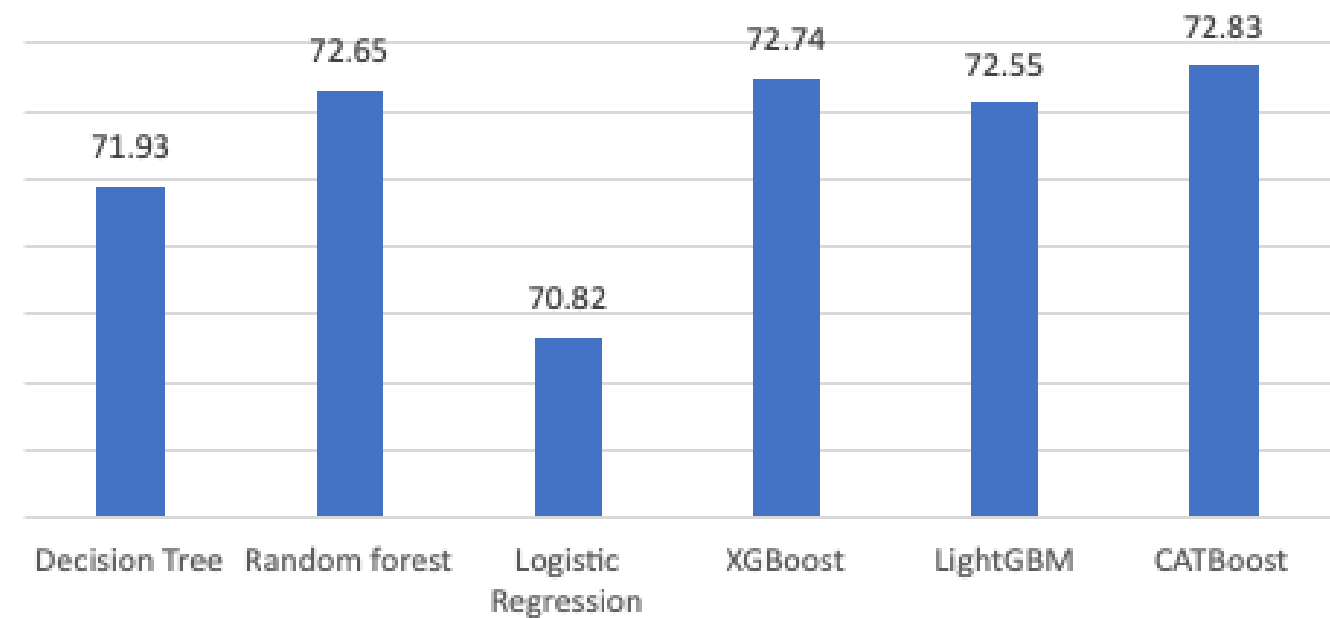
- Decision Tree
- Random Forest
- Logistic Regression
- Extreme Gradient Boosting (XGBoost)
- LightGBM
- CATBoost

PERFORMANCE METRICS

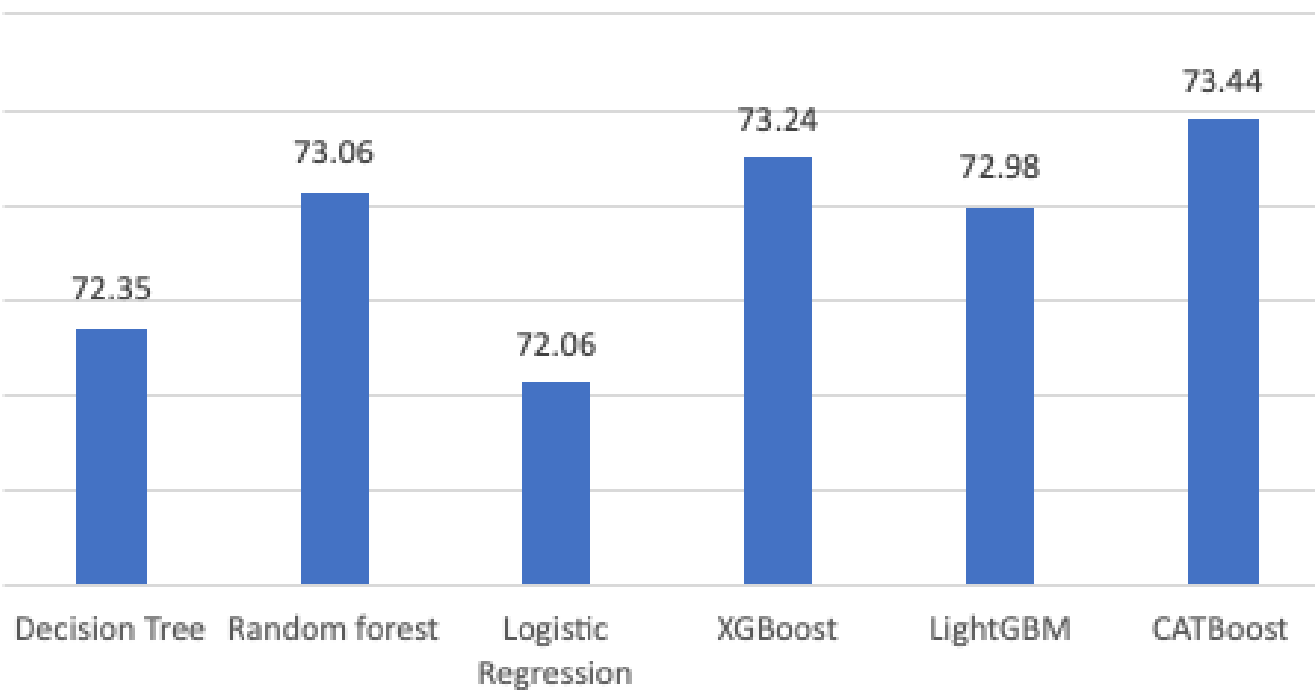
Accuracy



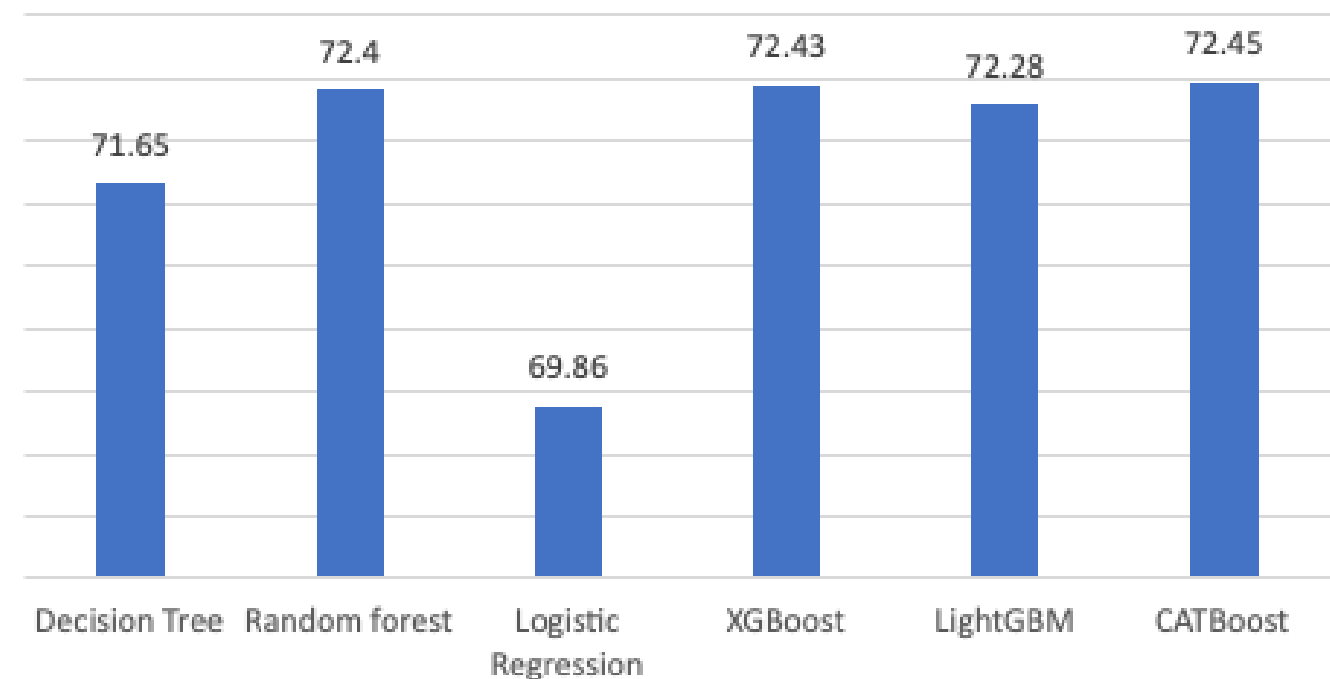
ROC-AUC



Recall

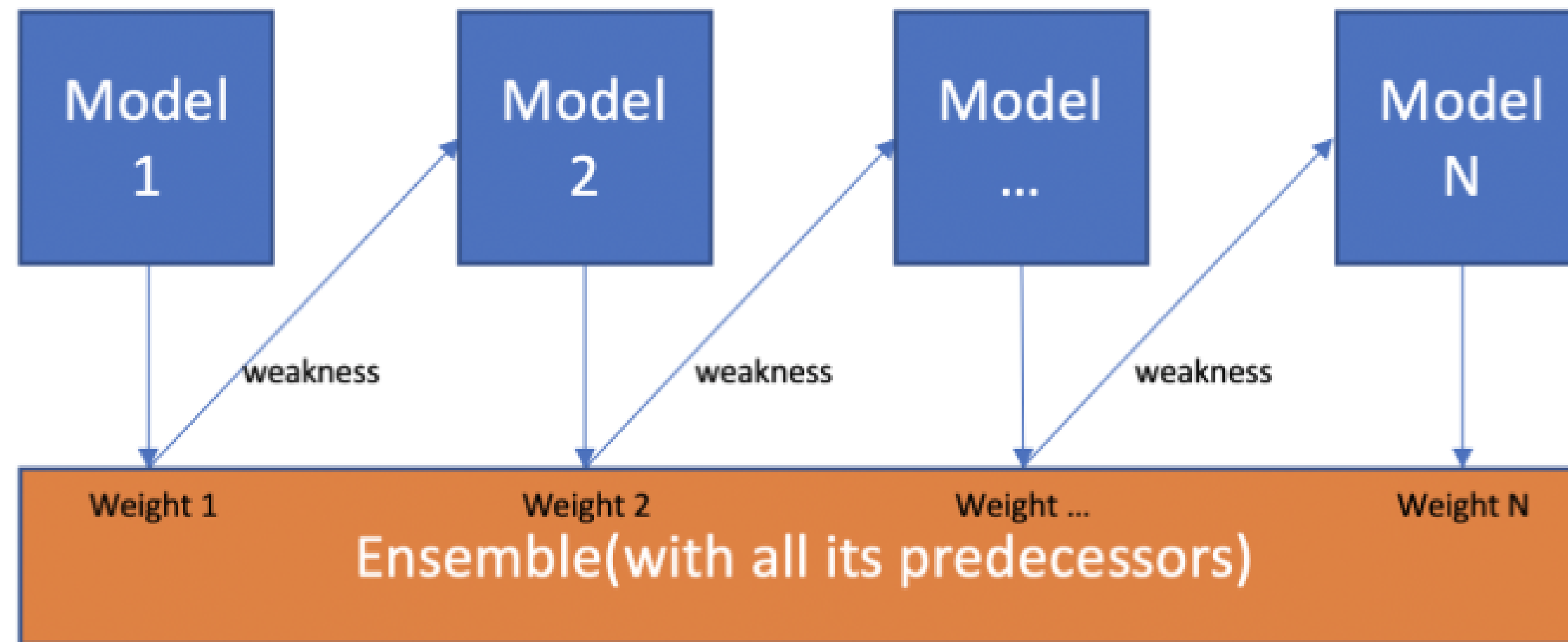


F1-Score



BOOSTING ALGORITHMS

Model 1,2,..., N are individual models (e.g. decision tree)



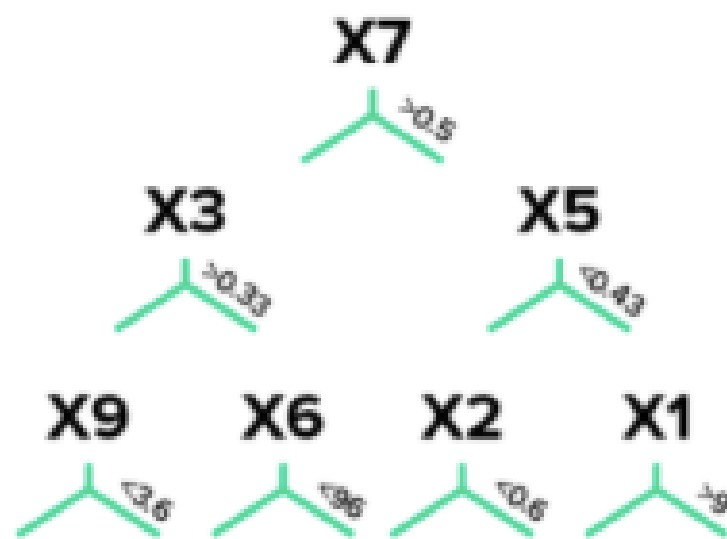
<https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30>

Boosting algorithms seek to improve the prediction power by training a sequence of weak models, each compensating the weaknesses of its predecessors. Types Boosting algorithms are:

- XGBoost
- LightGBM
- CATBoost

XGBoost

It looks for features and split-point that maximizes the gain. The splitting process is slower as it does not use weighted technique.

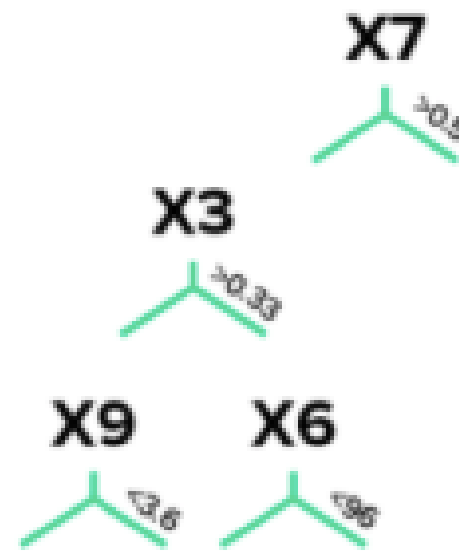


XGBoost

It splits according to the max_depth hyperparameter and then starts pruning until there is no more gain.

LightGBM

It used Gradient based one-side sampling, this maintains a balance between high speed and accuracy.

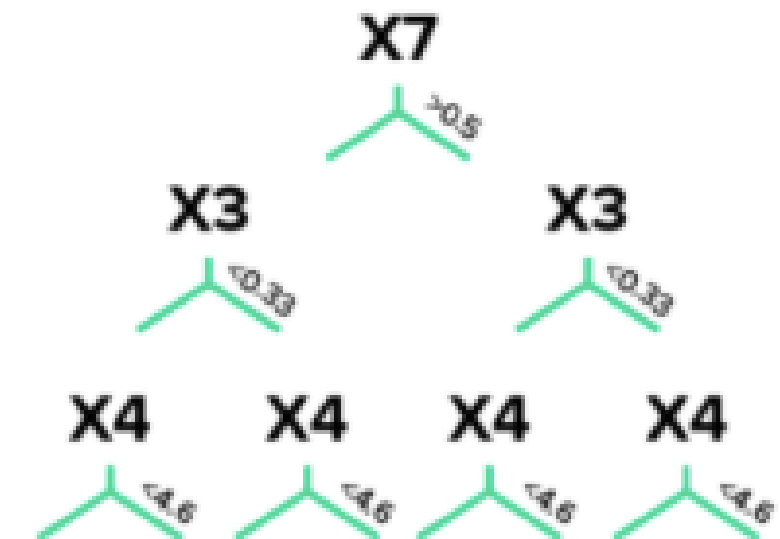


LightGBM

It grows tree leaf-wise. This results in unbalanced trees and overfitting in small datasets.

CATBoost

It used minimal variance sampling, in which weighted sampling is done on tree level instead of split level.



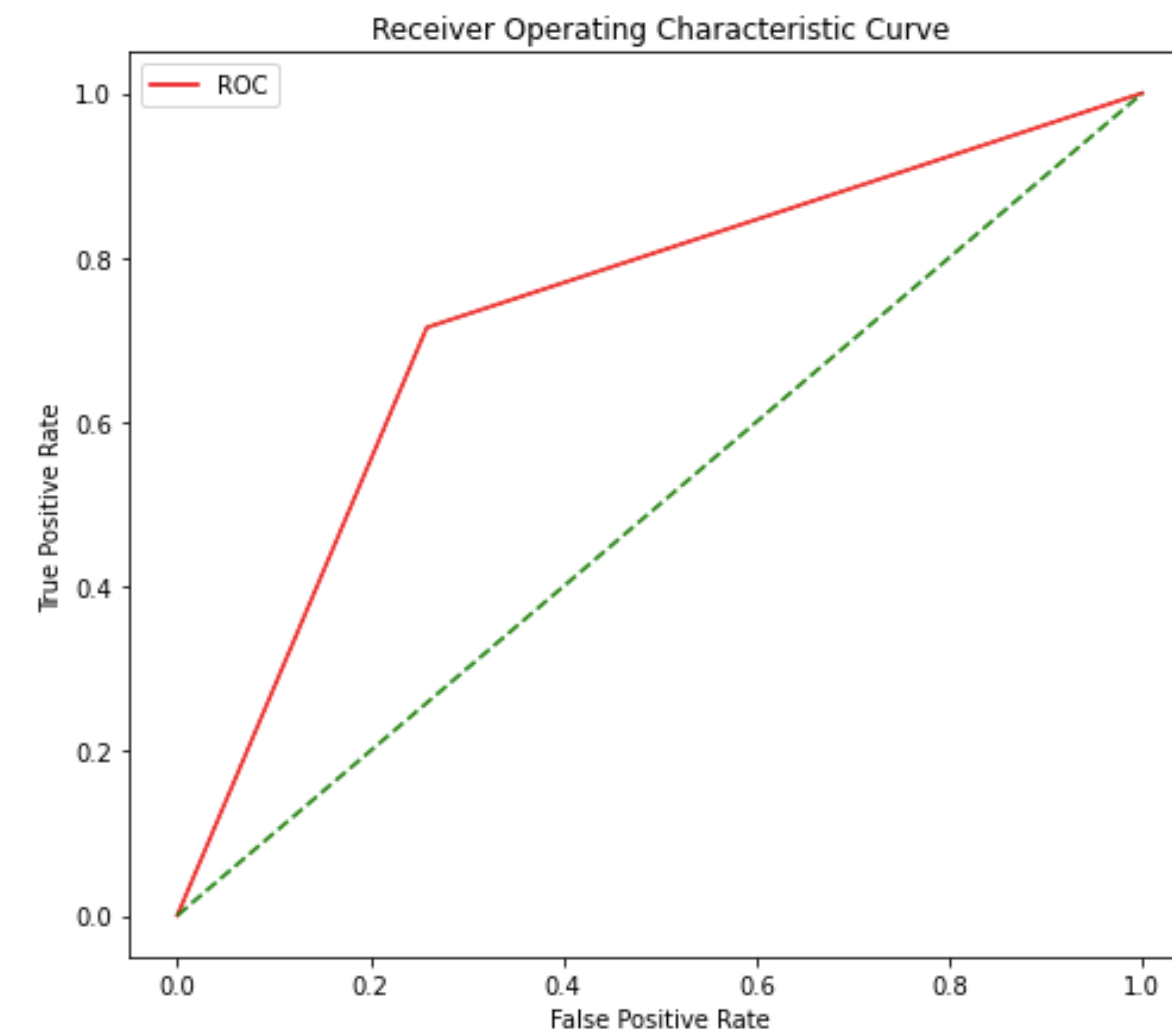
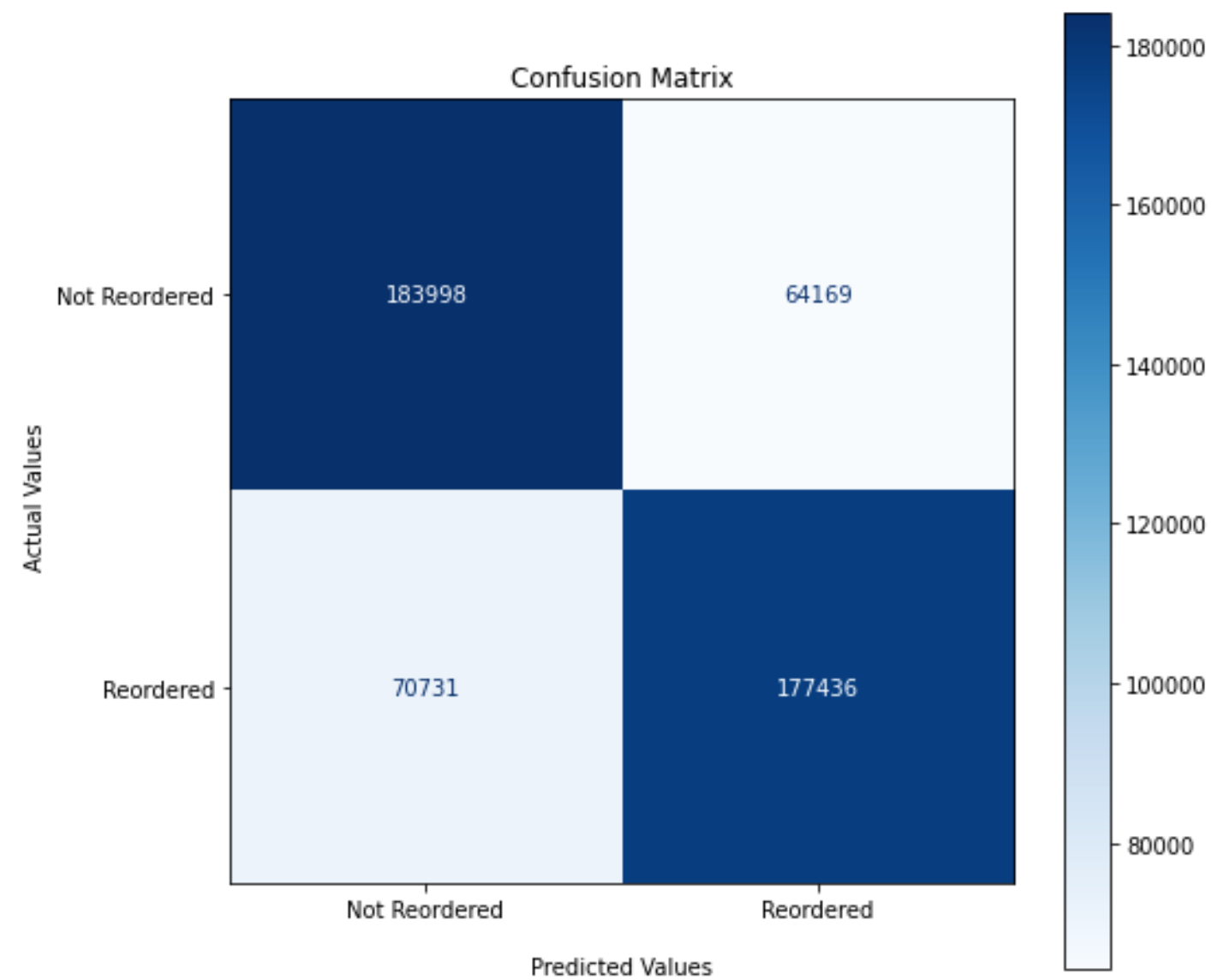
CatBoost

In CATBoost the same splitting criterion is used across an entire level of the tree. Such trees are balanced, less prone to overfitting, and faster.

MODEL SELECTION

CATBoost is preferred model because not only it outperforms other models in terms of performance metrics, it also has great speed and computational power.

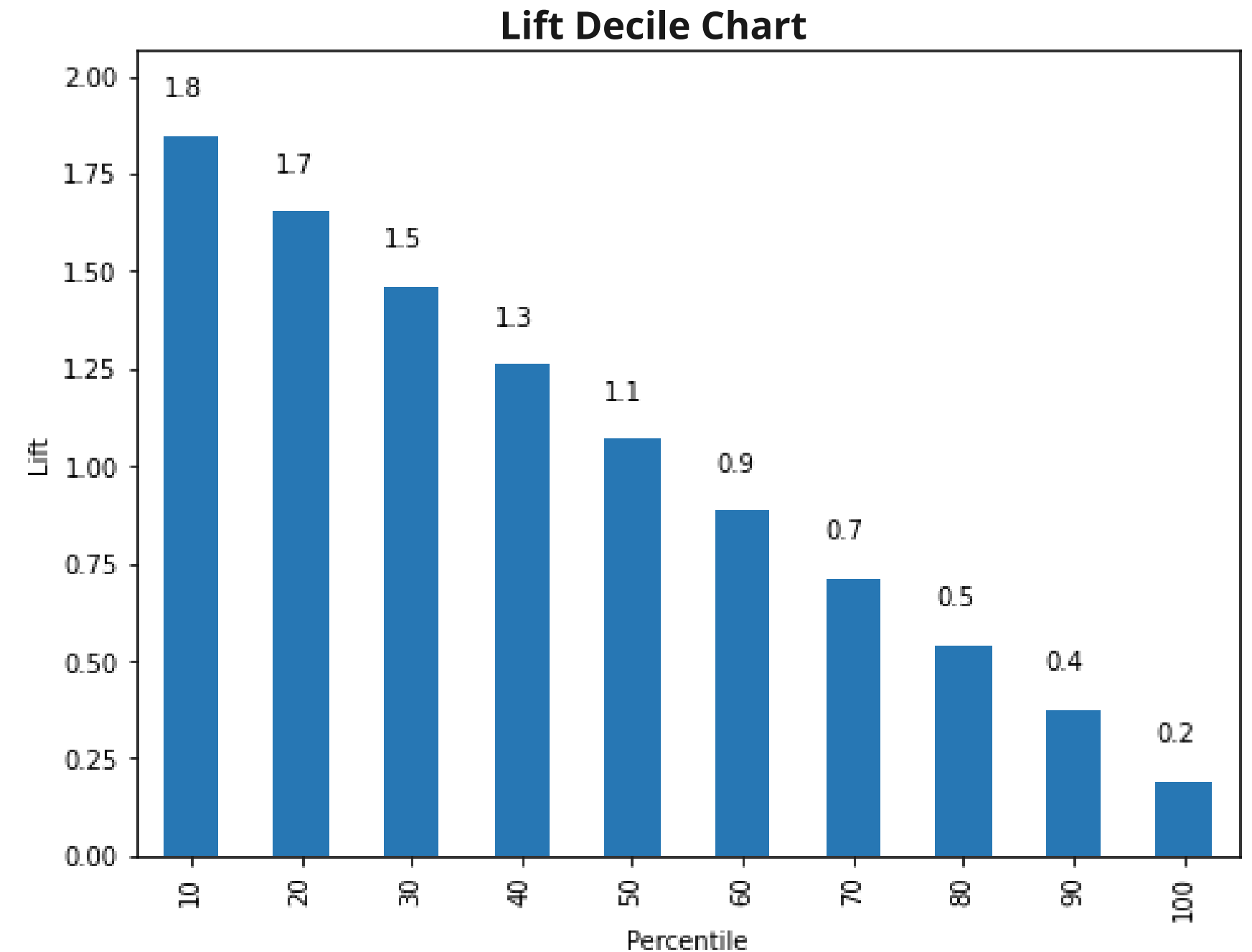
Accuracy: 72.8207%
Precision: 71.4986%
Recall: 73.4405%
F1 Score: 72.4566%
ROC AUC: 72.8367%



MODEL EFFECTIVENESS

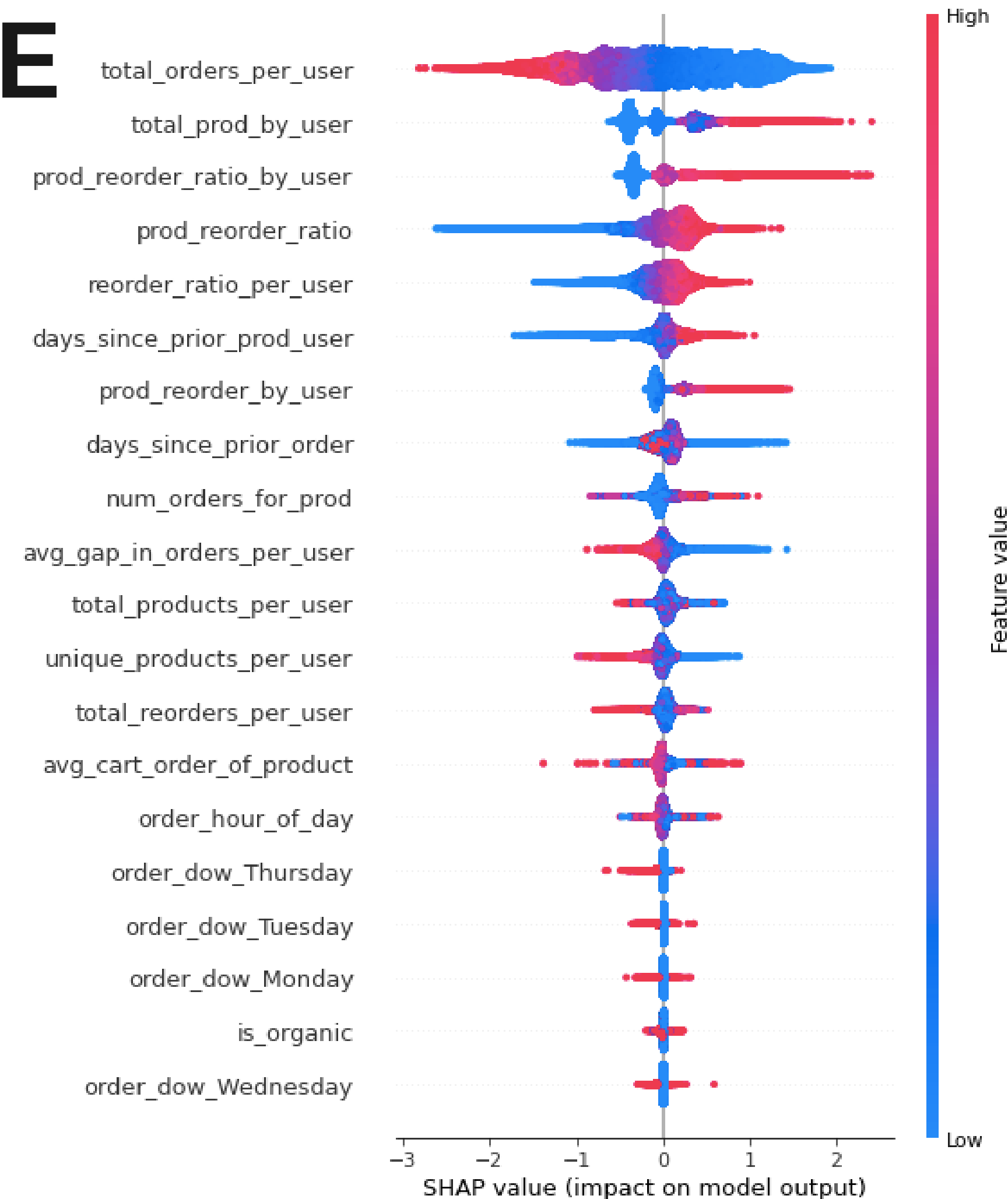
Lift score is the ratio between the result obtained with a model and the result obtained without a model. The 'staircase' effect in the decile chart shows the viability of the model by presenting from most likely to least likely.

30% of the records that are ranked by the model as "the most probable reordered" yields 5 times as many reordered as would a random selection of 30% of the records.



FEATURE IMPORTANCE

- Total number of orders are is some what evenly split, which indicates that while overall it is still important, its interaction is dependent on other variable
- High values of 'total_product_per_user' & 'product_reorder_ratio_by_user' have a positive impact on the model. While they are not the top important features, they have a huge impact on a subset of the users.
- Lower value of 'prod_reorder_ratio' and 'reorder_ratio_per_user' will have greater negative impact on model.
- Lower value of 'days_since_prior_prod_user' will have a negative impact. This will be an important factor in determining the best day to send push notification to increase call to action.



INFERENCE

RECOMMENDATIONS
FUTURE WORK

RECOMMENDATIONS

- Easy access to past orders for customers.
- Adding cart pre-fill options for quick checkout.
- Expanding organic products portfolio.

FUTURE WORK

- Leverage complete and current data.
- Using SMOTE technique to reduce bias in the data.
- Adding demographic data to the clustering technique

THANK YOU

Any Questions?

CREDITS

Special thanks to all the people who made and released these awesome resources for free:

- Canva
- State of Grocery Report by [1010Data](#)
- [McKinsey Report](#)
- K-mean clustering by [Analytics Vidhya](#)
- Boosting Algorithm by [Zixuan Zhang](#)