Instacart Market Basket Analysis

Capstone Project: Data Science and Analytics Certificate Bootcamp

Stack Education at Framingham State University

Deepika Dittakavi

June 18, 2020

Introduction

- Instacart is a grocery ordering and delivery app. Instacart's data science team uses transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session.

- Data Analytics can be used to increase user base, sales, customer satisfaction, efficiency etc.

- **Business Users:** Instacart Sales team, Market Basket Store Managers, Product Distributors.

Executive Summary

- **Goal:** Understand, evaluate and analyze the data by leveraging Supervised Classification algorithms to predict the products Instacart users are likely to reorder from past purchases.

- Data was analyzed using Tableau and R. Classification algorithms were evaluated in R on train and test data. Among many evaluated models, optimum performance was obtained by the Naïve Bayes model with an F1 score of 0.41.

- **Primary Recommendations:**
  - Leverage insights to plan for shopper schedules & supply of products
  - Build loyalty program and promotional sales to increase sales during off-peak days/hours.
  - Schedule additional workers on peak days/hours
  - Implement a 1-click cart to order for users to enable easy re-orders

- For future, geographic location details and time to delivery could be added to get more insights.

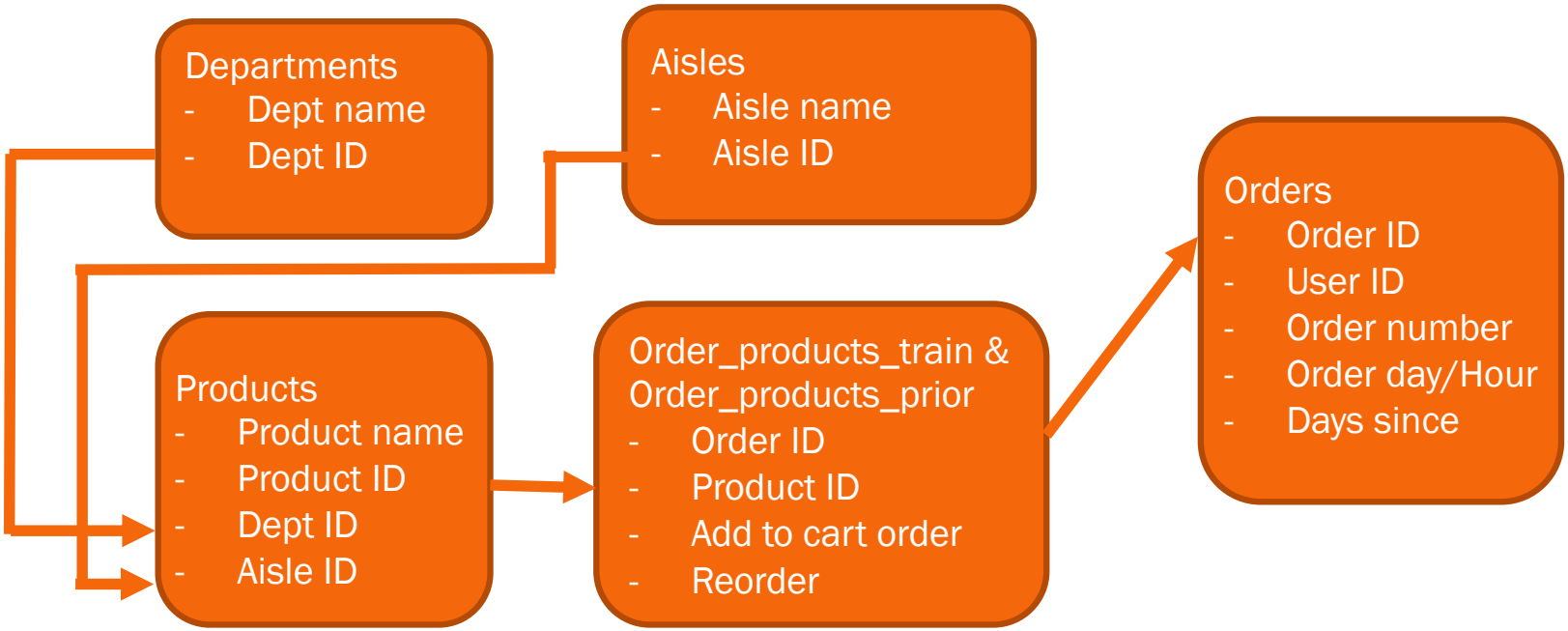
Data Analysis Approach

METHODOLOGY:

- Data Exploration & Insights
  - Understand data
  - Align with Stakeholders
  - Draw meaningful Insights
- Feature Engineering
  - Explore relationships and derive features
- Model Analysis and Performance
  - Evaluate different models, compare and apply best model to final data
- Conclusion
  - Final summary and recommendations

Data Exploration: Dataset

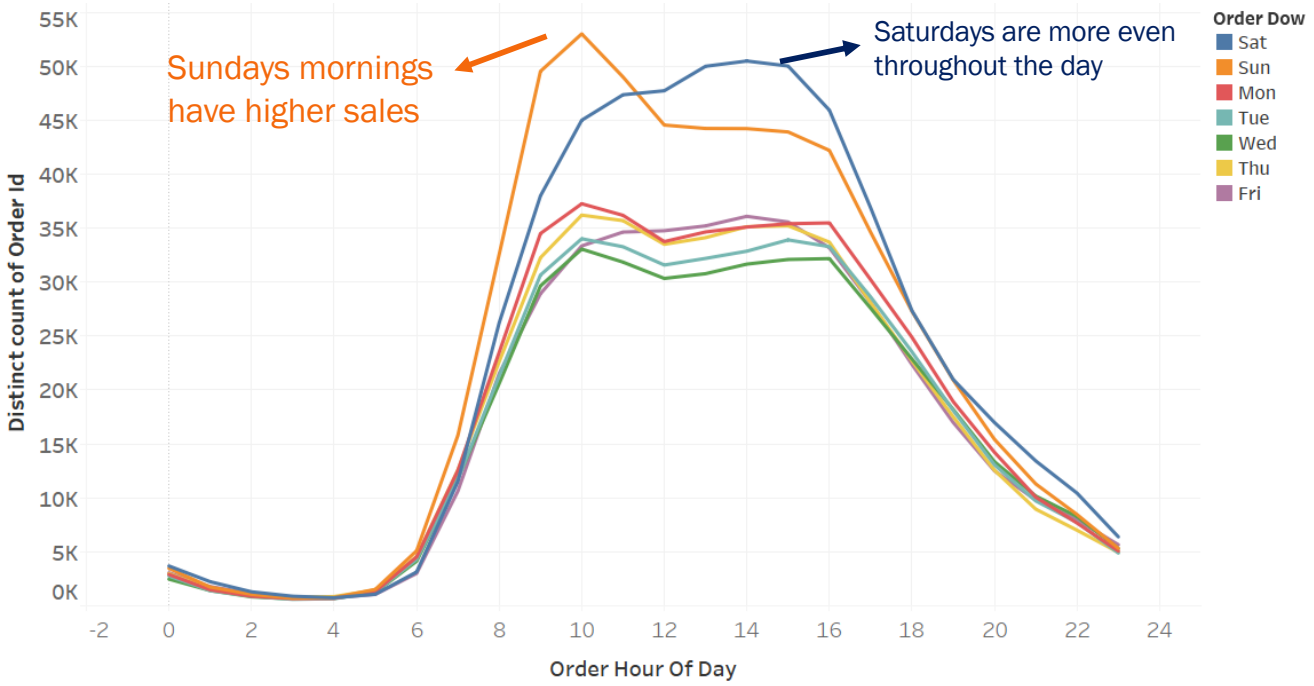
- Data Source: Kaggle (https://www.kaggle.com/c/instacart-market-basket-analysis)



- Challenges:
  - Multiple files need merging
  - Dependent Variable and features not given and need to be defined
  - Imbalanced Dataset
  - Huge data(3 million orders) can lead to infrastructure limitations

Data Exploration: Time

- Store managers schedule Grocery supply by coordinating with Product distributors
- Schedule shoppers by leveraging demand is critical for on time delivery and customer satisfaction



Data Exploration: Key Insights

Parameter	Fact	Insight	Recommendation to Business Users
Department	Produce, Dairy eggs and Beverages are top reorders	Basic food items are reordered Reorders have a pattern, that can be used for features in the model	Work with product distributors to keep up the supply of these high demand products
Aisle	Fruits & Vegetables, Milk and eggs top reorders	Lower shelf life items have high reorders	Schedule store reps to stock items frequently, monitor freshness of products
Product	Bananas most popular product. Organic products very popular	Use “organic” feature for model Reorders seem to be most basic items	Arrange frequent reordered items like Organic near store entrance for easy shopping
Day of week	Weekends high sales, Wednesday lowest	Users think about groceries on weekends and on Thu/Fri to shop for weekend	Schedule for more shoppers on weekend and lower in mid-week
Hour of day	Early morning and afternoon are peak hours	Users likely ordering in the morning for evening delivery	Schedule shoppers & Re-stock
Days since prior order	Users mostly buy weekly, local maximums at 2 wks and 3 wks	Cart size expected to be smaller for lower gap in days.	
Cart size	Increases with days since last order	User orders more products hence use as a feature in model	
Add to cart order	First placed products are most needed products	Use position of product in cart for model & determine demand of product	

Feature Engineering

- Identify User behavior
  - User reorder percent
  - Average time of order, frequency
  - How many orders/products
- Identify Product characteristics
  - Product reorder percent and frequency
  - Is it one of the popular products
- Identify User-product combination patterns
  - Reorder percent for a product by a user
  - User cart order of the product

Model Analysis

- **Output Measure Method:** Since data is imbalanced, accuracy will be biased one-way. F1-Score is used to measure performance.

		Prediction	
		True (1)	False (0)
Actual	True (1)	T_p	F_n
	False (0)	F_p	T_n

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + T_n}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

- Cut-off analysis was used for fine tuning.

- Data split into 3 sets:
  - Train set – used to run different models
  - Evaluation set – used to compare outputs and evaluate models
  - Test set – used to run final model on data(not seen by model)

Results

- Results on Train and Evaluation data sets:

Model*	Accuracy	Precision	Recall	F1 Score
CART	0.9085	0.1709	0.6120	0.2672
Random Forest	0.9081	0.1675	0.6047	0.2623
Naïve Bayes	0.8998	0.2978	0.4781	0.3669
Logistic Regression	0.9086	0.1734	0.5952	0.2686
Naïve Bayes(Tuned-cutoff**)	0.8628	0.4937	0.3543	0.4126

- * Models were analyzed on a random sample of data with 10% users (13K) due to infrastructure limitations – 850K observations
- ** Cutoff analysis showed 0.25 cutoff yields higher F1 score than default 0.5

- Final Results on Test data:

Model*	Accuracy	Precision	Recall	F1 Score
Naïve Bayes	0.863	0.4840	0.3562	0.4104

Conclusions and Future Work

Wrap up:

- The Naïve Bayes model showed optimum performance with an F1 score of 0.41.
- This model can be used by Business users to take appropriate actions to increase profitability, optimize efficiency and get customer satisfaction

Future Work:

- The model can include over sampling to address the imbalance in the data.
- Extend analysis to include first time buying of new products
- Include analysis based on Apriori theorem(products that are bought together).
- Extend analysis to complete dataset on powerful infrastructure

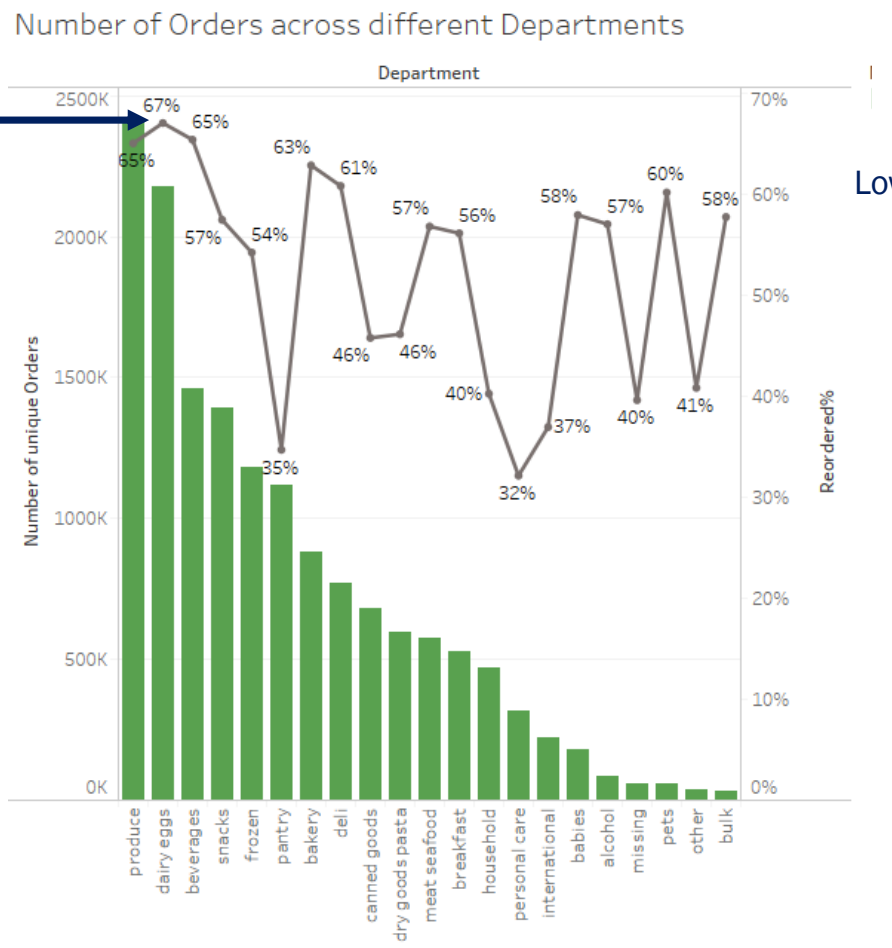
References

- https://www.kaggle.com/c/instacart-market-basket-analysis
- https://topepo.github.io/caret/model-training-and-tuning.html
- https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e
- https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/
- https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

Appendix: Departments

Dairy and Eggs most reordered

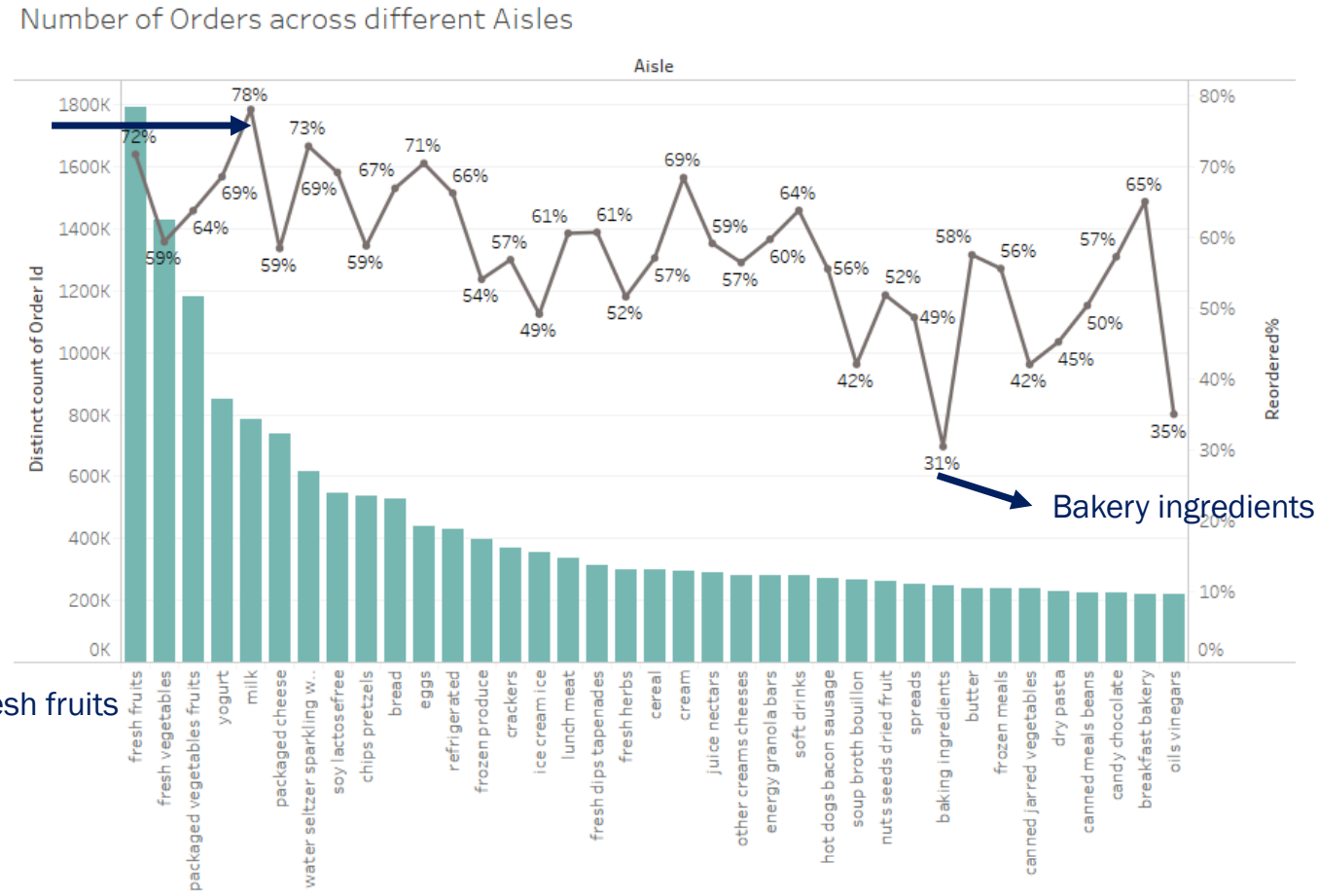
Most orders from Produce, Dairy & Eggs, Beverages & Snacks



Low shelf life items most re-ordered

Appendix: Aisles

Milk most reordered



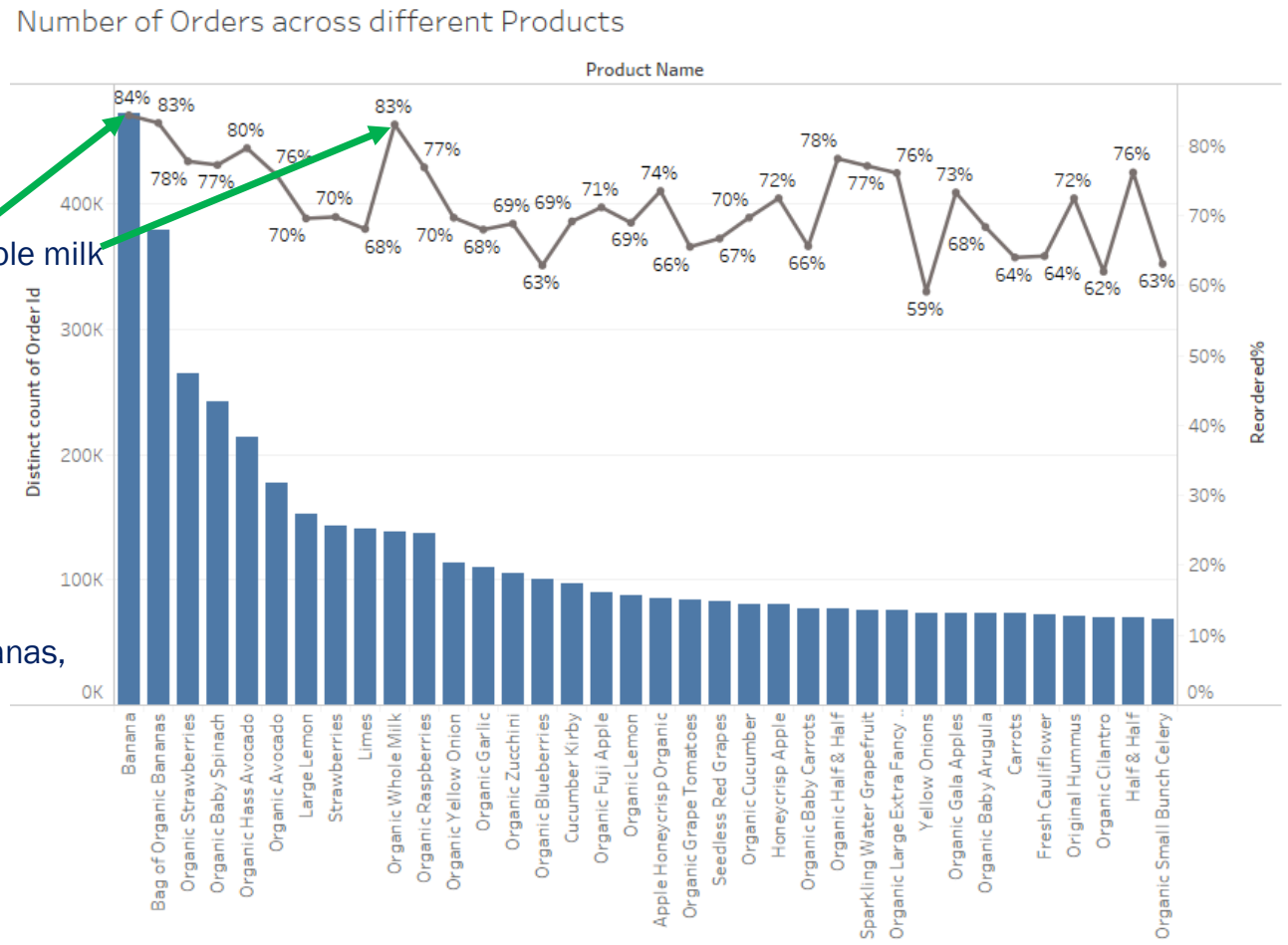
Bakery ingredients lower reorder rate

Most orders from Fresh fruits & vegetables

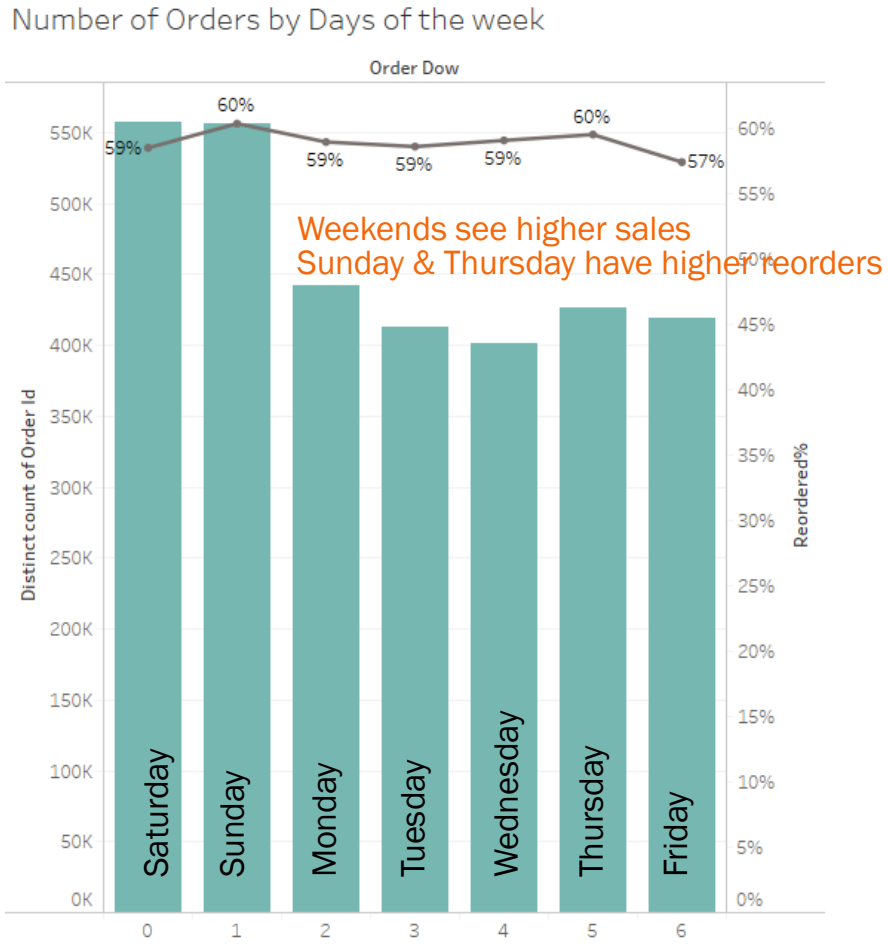
Appendix: Products

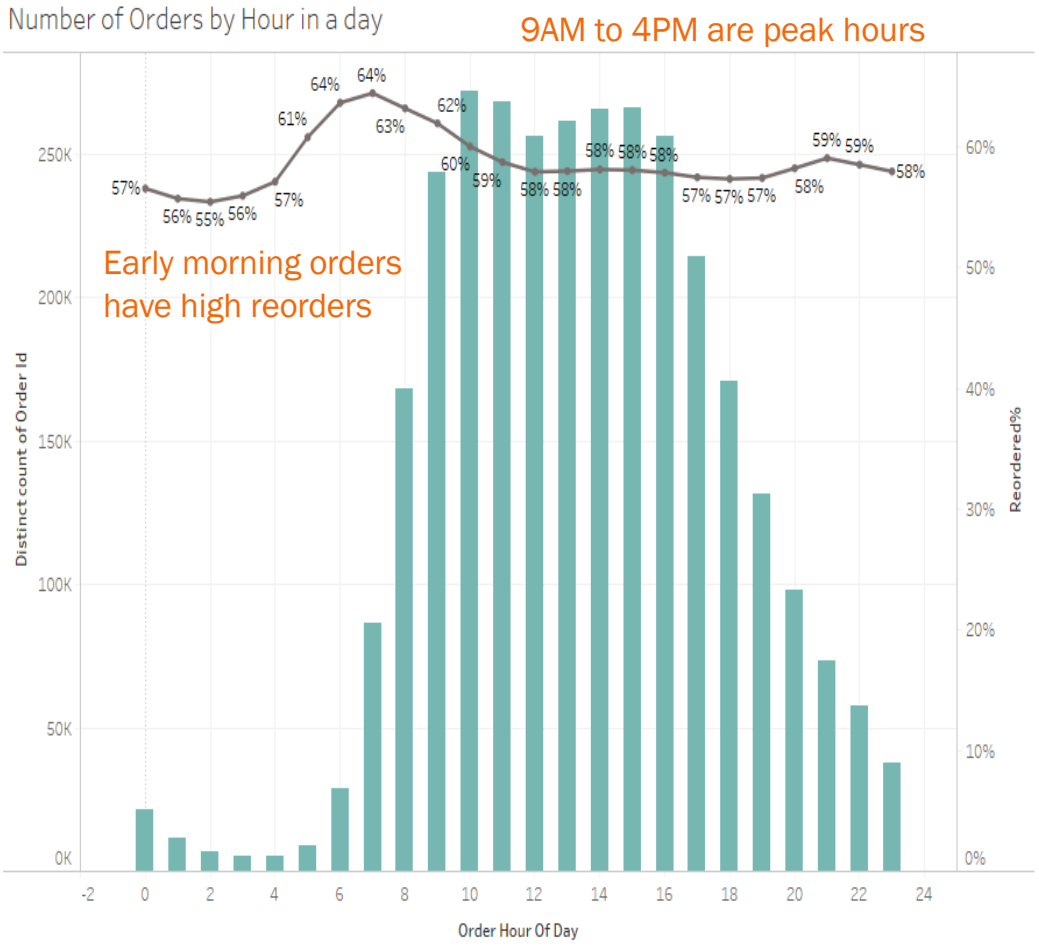
Bananas and organic whole milk most reordered

Most orders for bananas, Organic bananas, Organic strawberries



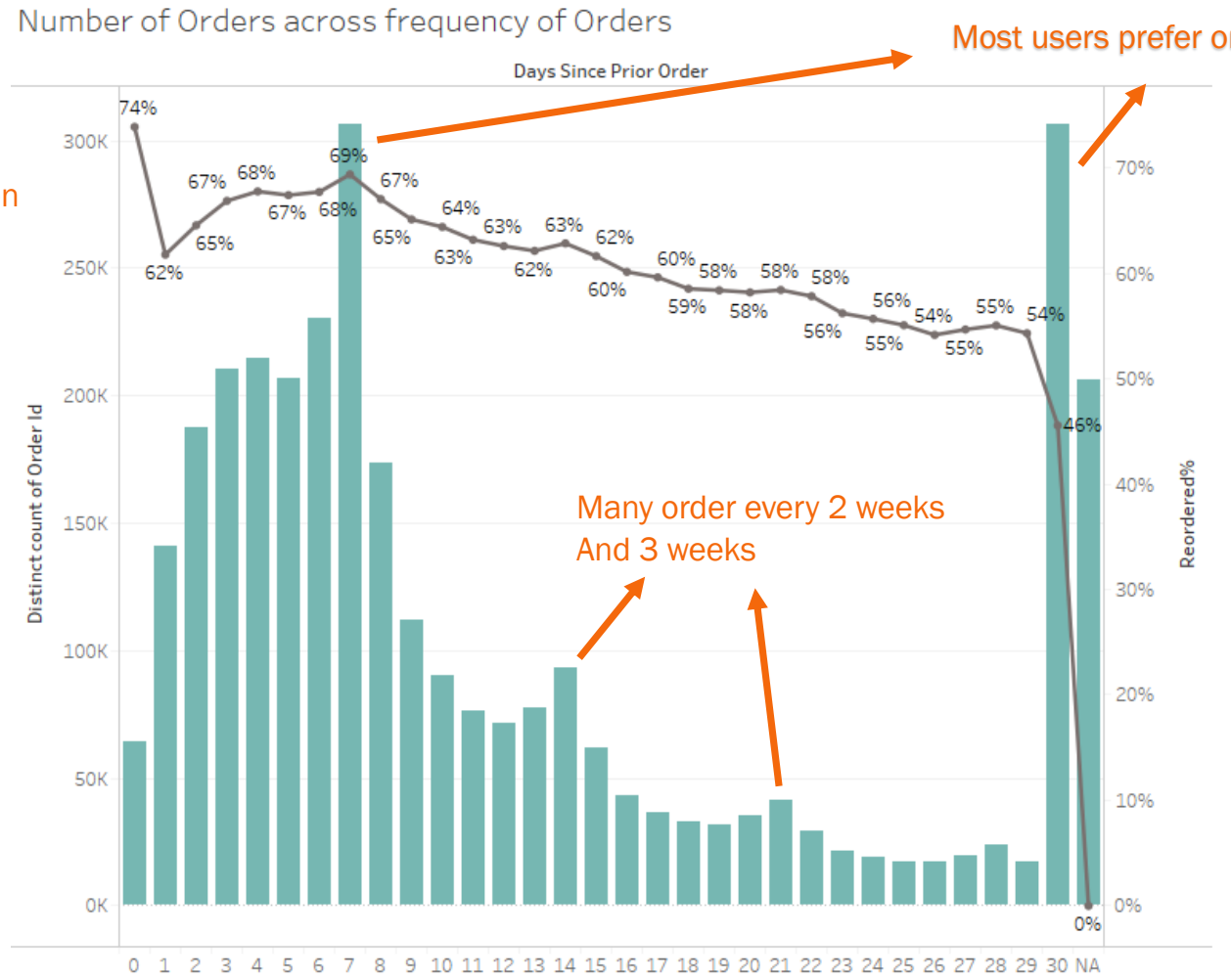
Appendix: Time





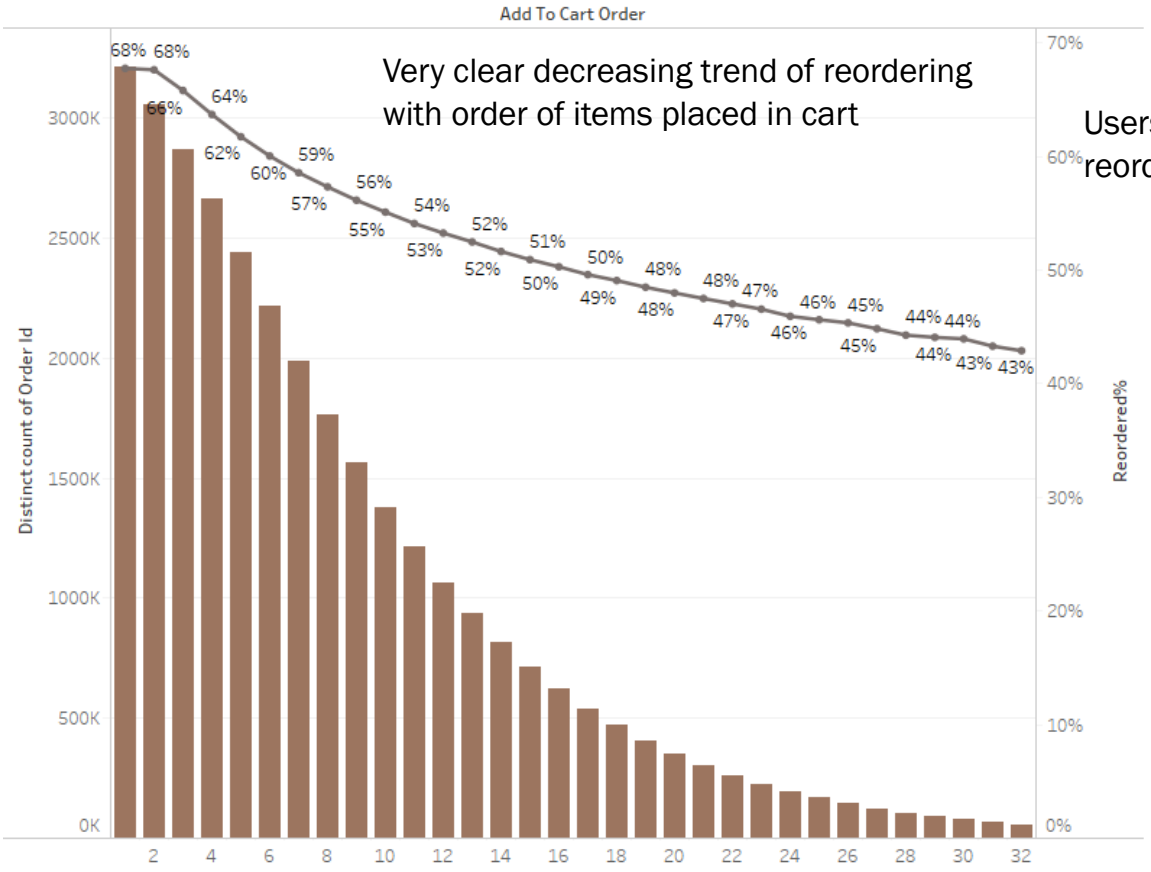
Appendix: Frequency of orders

Reorders most when ordering frequently



Appendix: Cart Order

Reorder % by the order placed in the cart



Very clear decreasing trend of reordering with order of items placed in cart

Users place most definitely needed reordering items in order of priority