



Shop the products you love



Save time & money



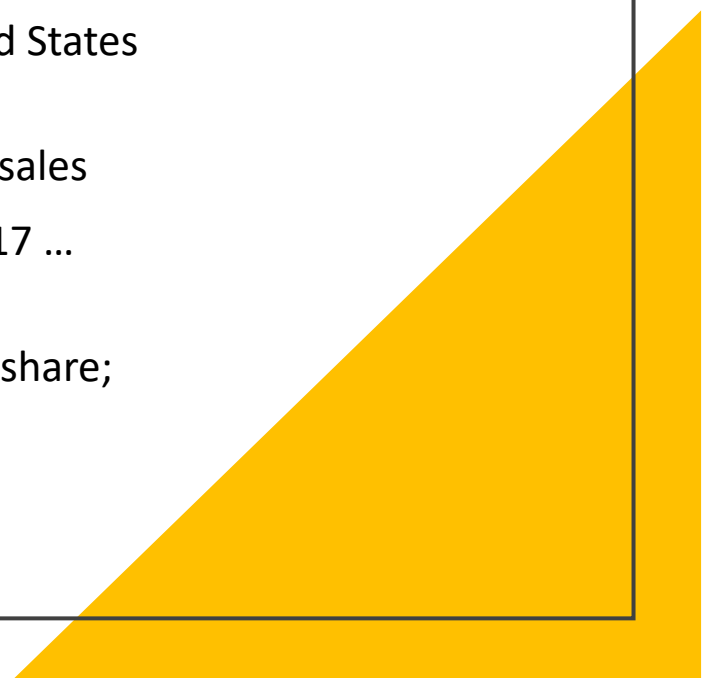
Get same-day delivery or
pickup



Improving Instacart 1 customer at a time

Presented by Paulette
Rodrigues

Introduction

- Instacart was **launched** in June 2012 as a grocery delivery and pick-up service
 - It is essentially a **delivery platform** that partners with retailers across the United States and Canada to provide food products to their customers
 - By 2017 the retailer Whole Foods was responsible for **more** than 10 % of there sales
 - Their competitor Amazon acquired Whole Foods, for \$13.7 billion in August 2017 ... putting them now in second to Amazon (Tom's Guide)
 - For the past three years Instacart was able to **maintain** and **retain** their market share; based on 2019 figures, membership has **tripled** in 2020
- 
- A large yellow triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

Historical Data on Instacart Shoppers

Instacart Shoppers

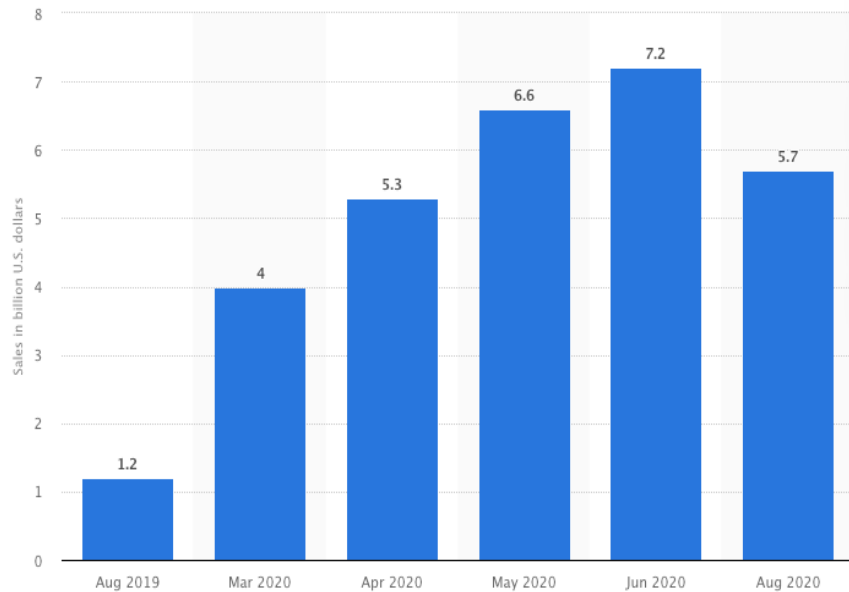
2014	5,000
2016	20,000
2018	70,000
2019	130,000
2020	500,000





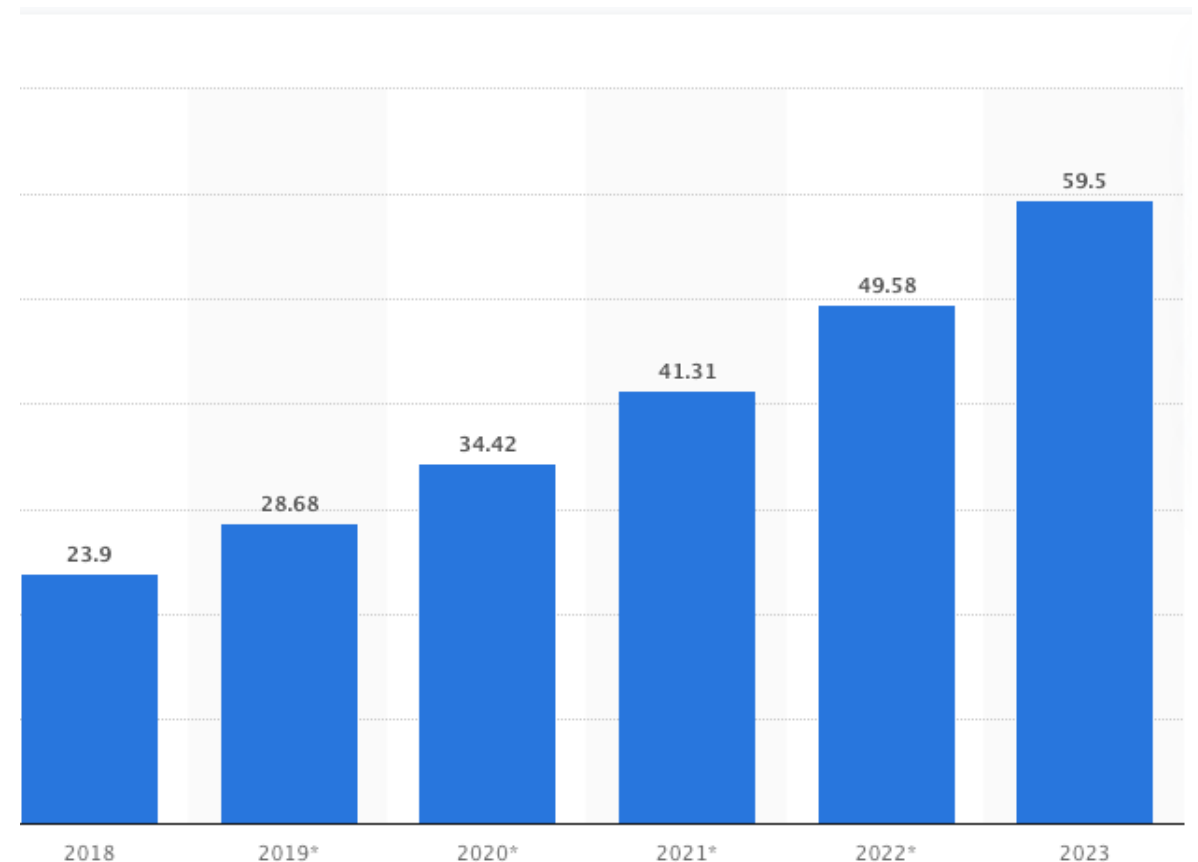
Current Trends

(in billion U.S. dollars)



- Instacart earned its first monthly net profit in April 2020, **netting** \$10 million...**reaching** its 2022 goals
- Covid19 contributed to an **increased** in demand ... which has ballooned their net worth from \$7.6 to \$13.7 billion
- In 2020, it has **expanded** its support staff from 1,200 to 18,000
- They are **on track** to processing more than \$35 billion in grocery sales this year

Expected online growth to 2023



Why have you chosen to make your grocery purchases online?

(Multiple responses allowed)

Response	Percentage
As part of COVID-19, I didn't want to venture to the store	54.4%
It saves me time	47.3%
I don't have to deal with long lines in the store	43.9%
Convenience; 24x7 shopping	37.0%
As part of COVID-19, many items were out of stock in the store	31.1%
Online search makes shopping simpler	26.5%
I don't buy things I don't need	23.9%
I don't enjoy going to the grocery store	22.1%
I can avoid the parking lot	21.5%
I don't have to shop in bad weather	18.6%
As part of COVID-19, I was unable to get to a store	15.2%
It is easy to replenish items	14.3%
It saves me money	14.0%
Access to online specials and coupons	13.2%
Price transparency	10.6%
I receive the freshest items	4.5%
None of the these	2.5%

Online Shopping Customer Survey

Question?

How can Instacart's customers **decrease** the time spent by to complete their full order, **improve** their online experience and have them **return** as customers post Covid19?



Your stores





Problem Statement

We need to predict what a customer may purchase on their next order (therefore directing them to the retailer that will fulfill their shopping needs)

Data Description

Orders (3.4 million, 206 K users):

- Describe the customers orders e.g. order number, day of week/hour of day, days since prior order etc.

Products (50 k):

- Name of the products available for purchase

Aisles (134):

- Contains the aisle available

Departments (21):

- Contains the department in the stores

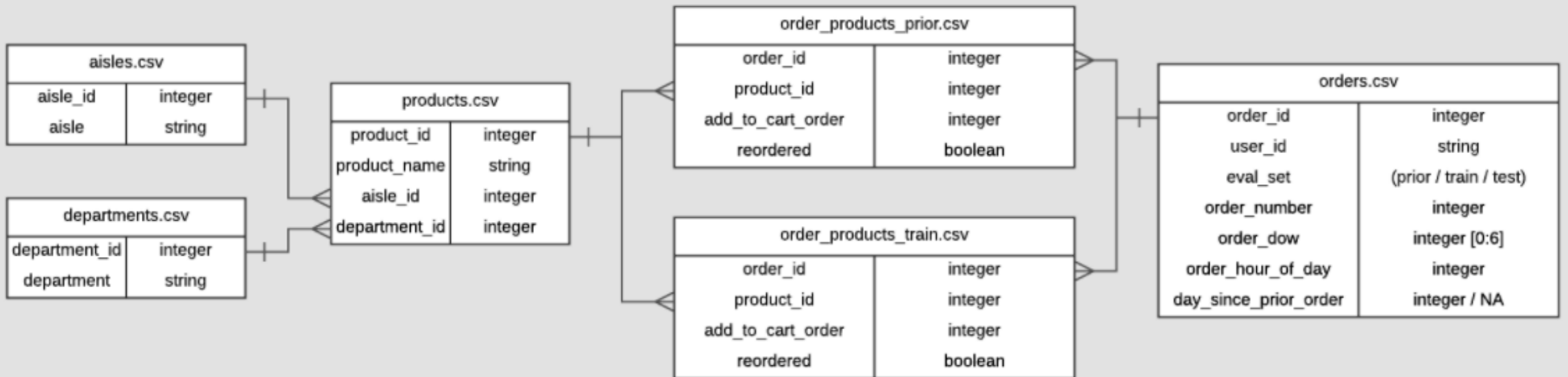
Order_Products (30 million):

- Listing of all items ordered by customers per order
 - "prior": orders prior to that users most recent order (~3.2m orders)
 - "train": training data supplied to participants (~131k orders)

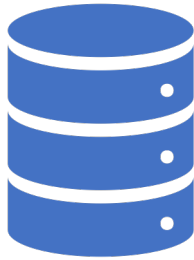
Instacart Dataset: Entity Relationship Diagram

- <https://i.imgur.com/R7c37Yw.png>

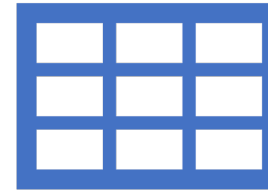
Instacart DBMS Diagram



Data Cleaning Process



SQLite was utilized to generate the final dataset for modeling



Data was aggregated to create one row per customer's per item purchased

Aggregation was done based on the following conditions:

- By items purchased from all customer's orders
- By items purchased from the last 5 orders of the customer

Department of the item was also kept as the natural category of the items

Null values were set to 0

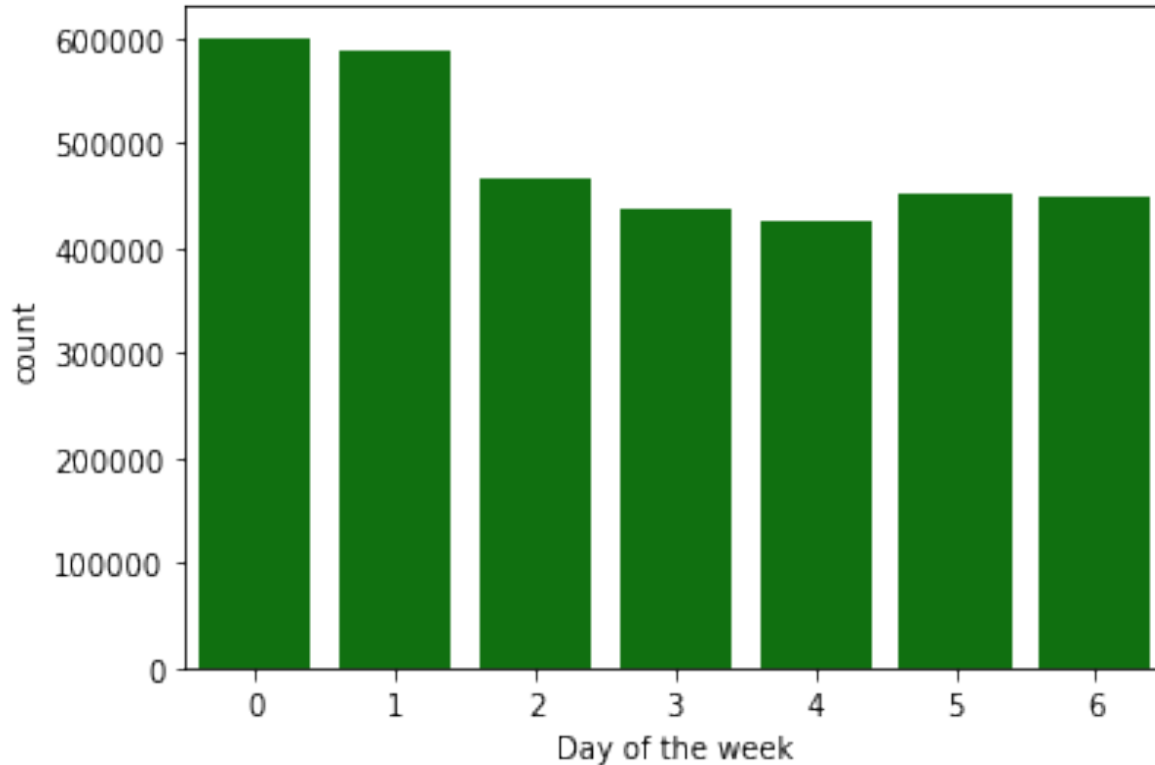
The background is a dark, blurred image featuring various financial data visualizations. On the left, there's a line graph with white circular markers connected by a thin white line. In the center, a bar chart with orange bars is visible, with a white rectangular box highlighting one of the bars. To the right, another line graph with blue markers and a white line is partially visible. The overall aesthetic is professional and data-oriented.

Exploratory Data Analyst

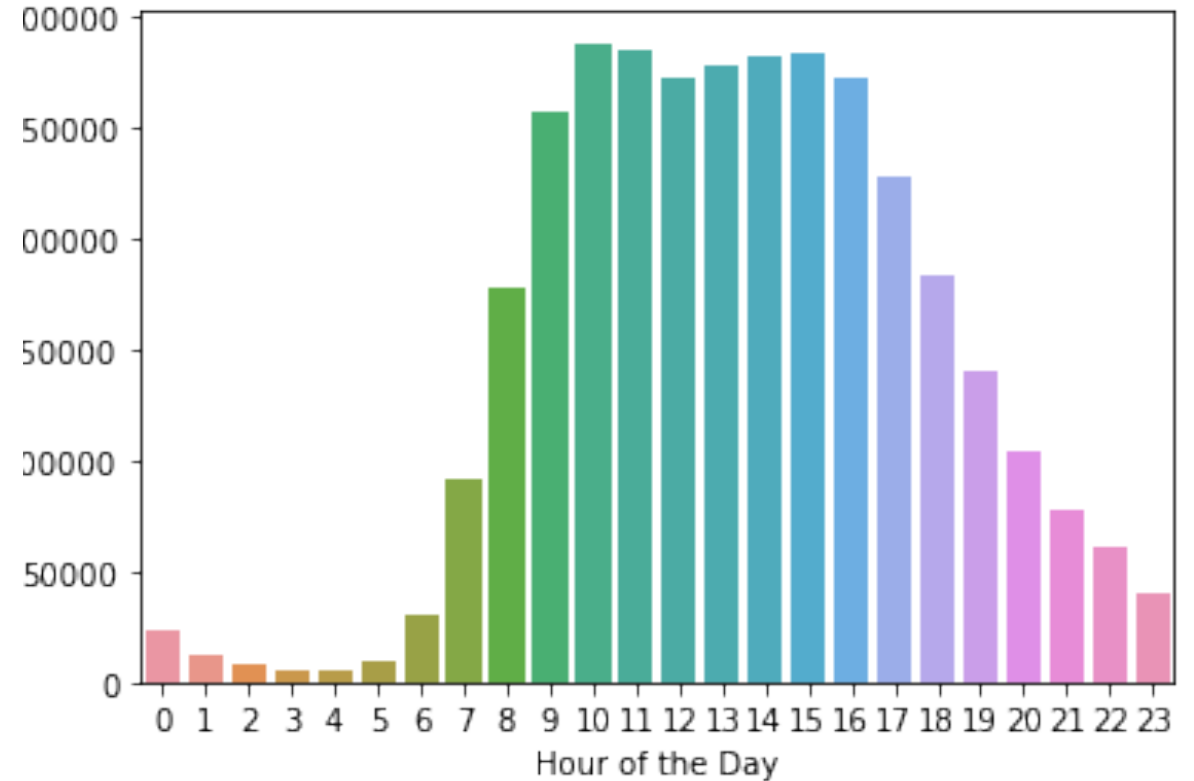
EDA: Distribution of Day of the week and hour items are purchased

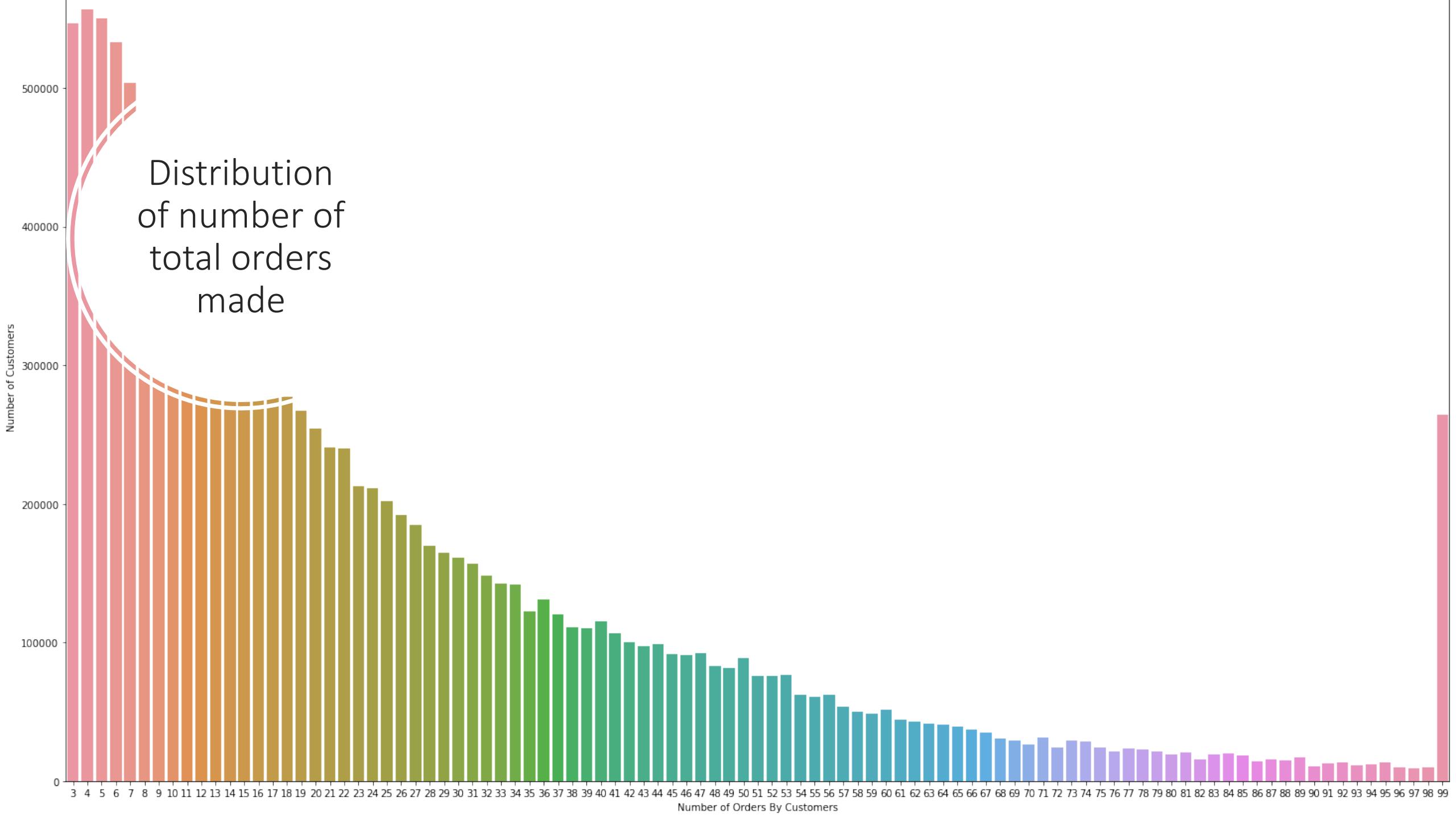


Purchase Day of the Week Distribution

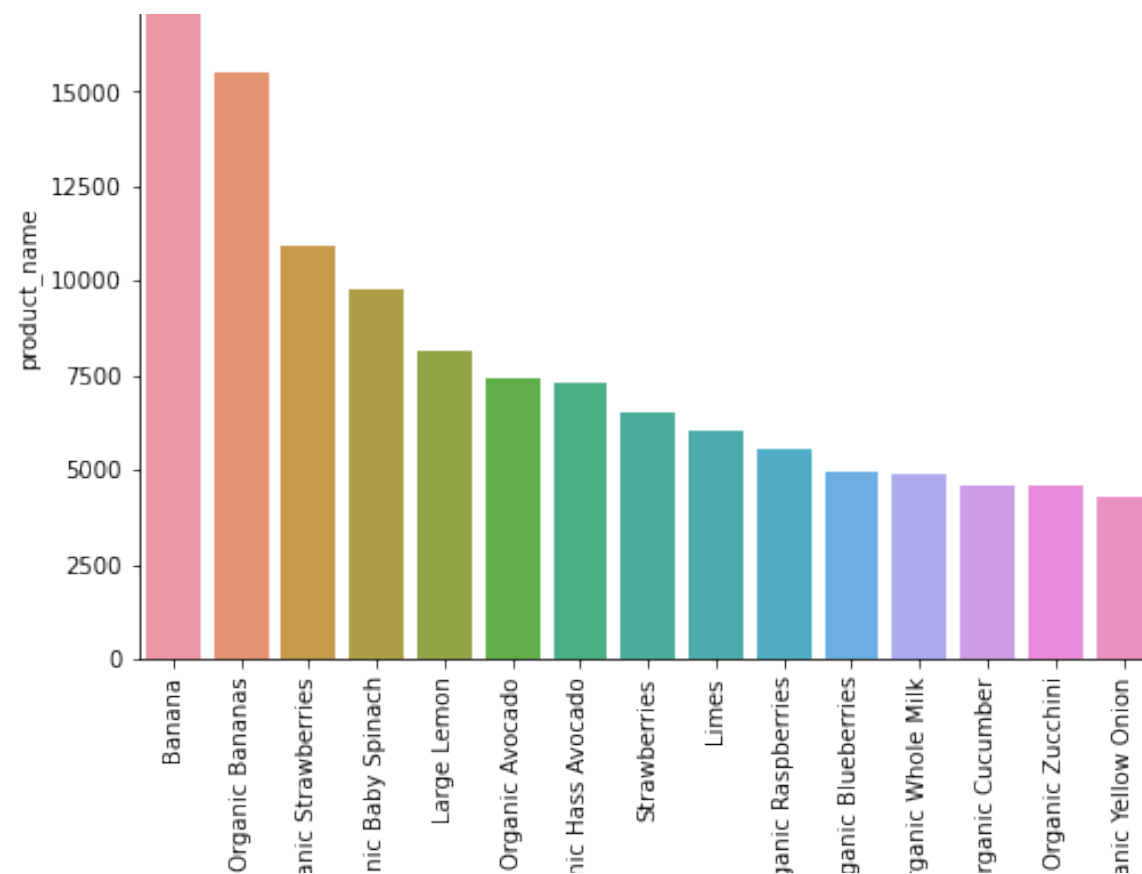
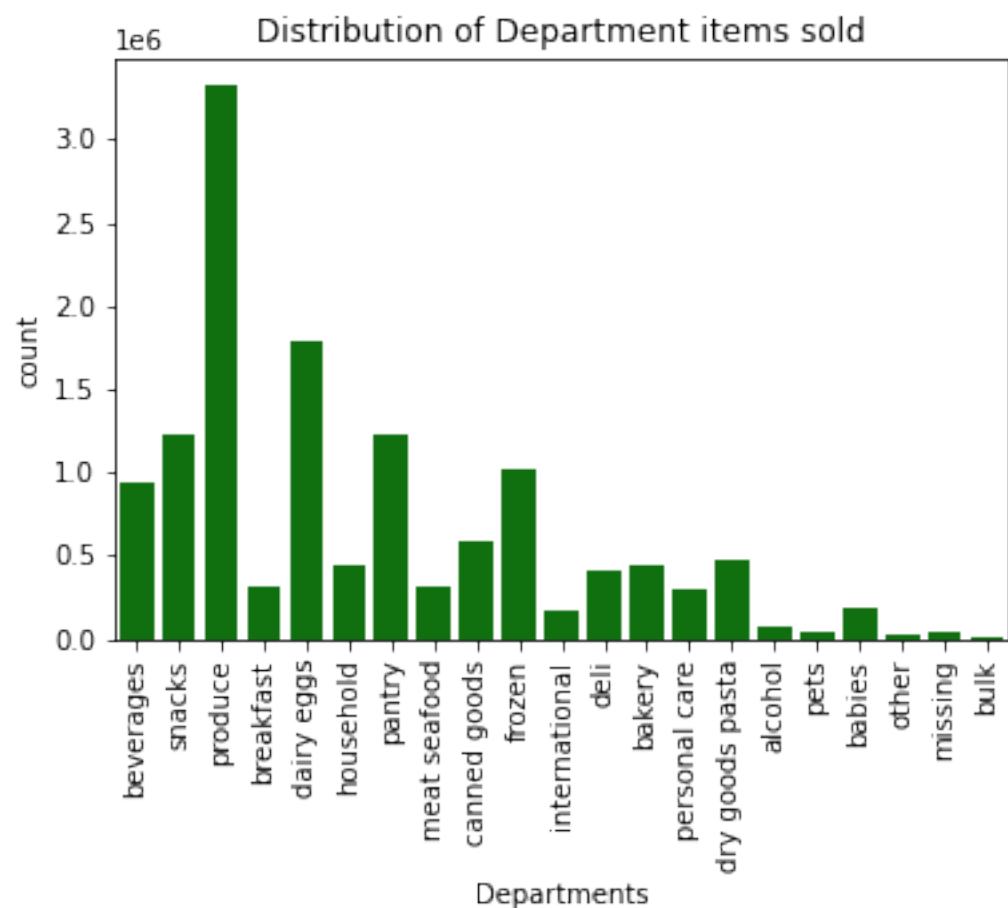


Purchase Hour Distribution

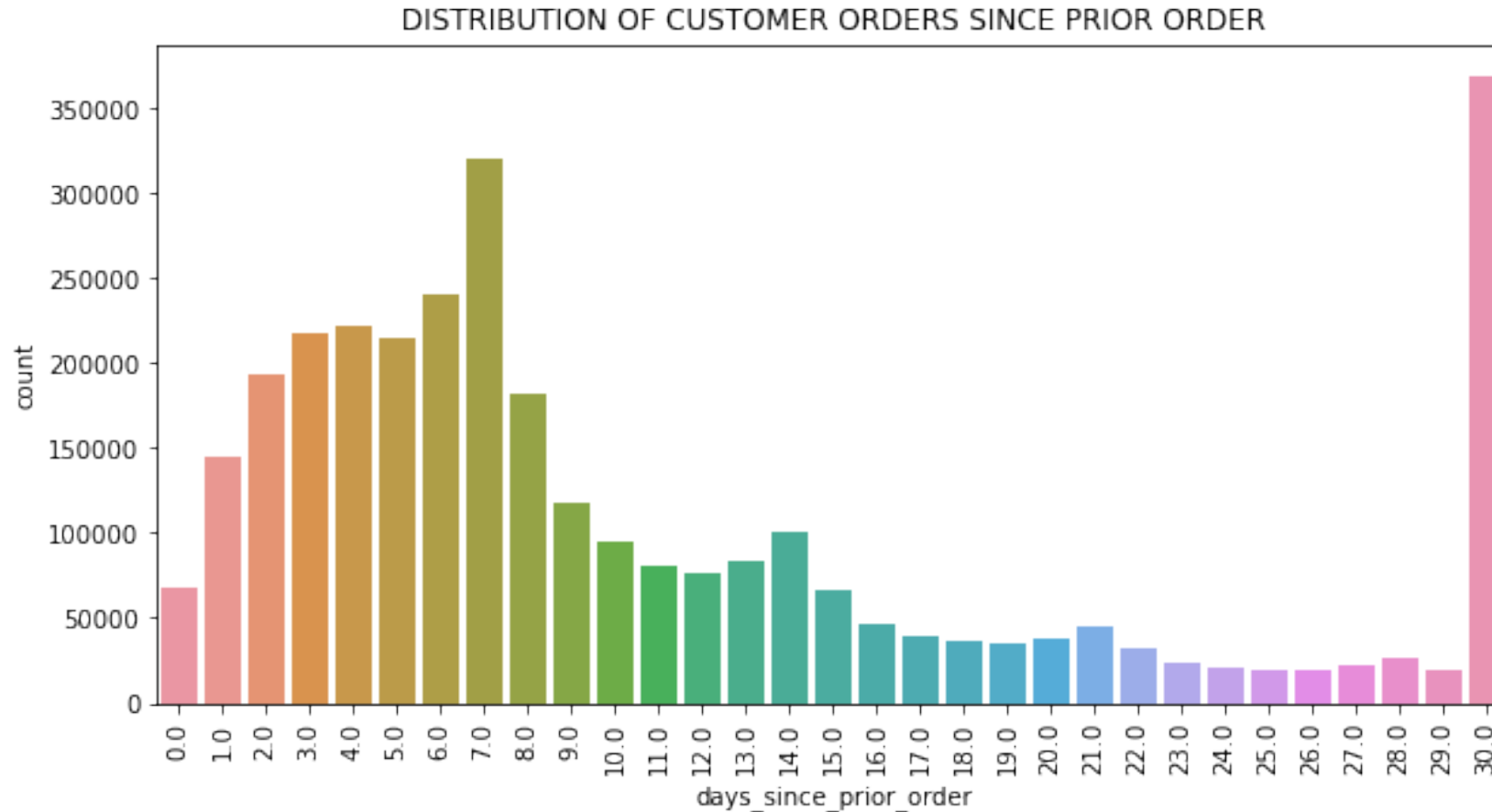




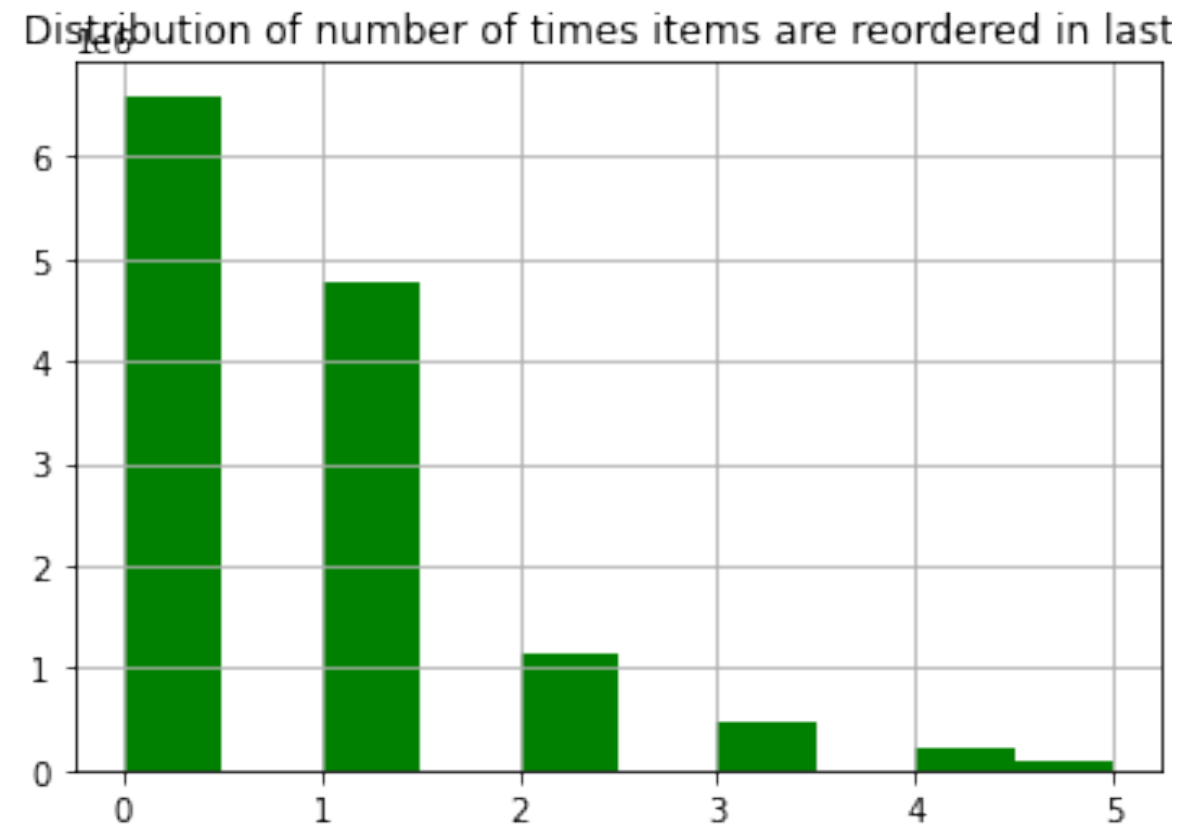
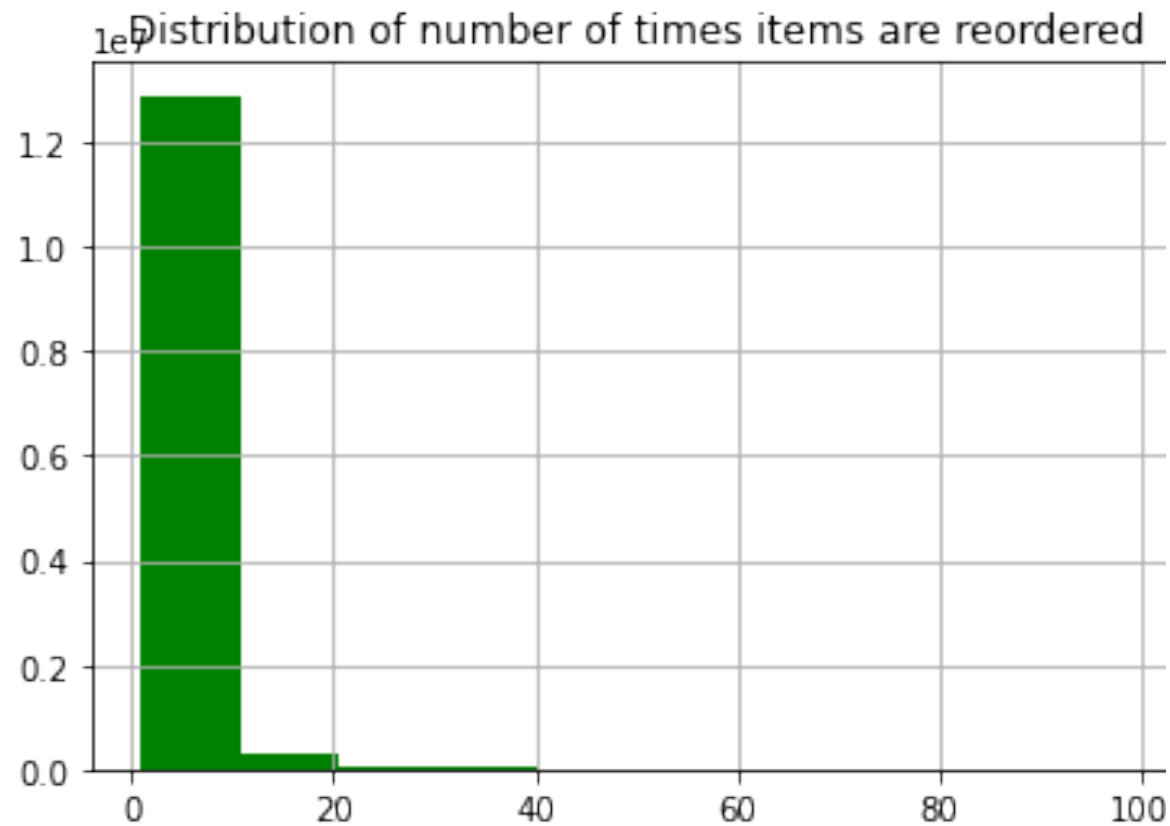
Distribution of number of times items are sold and of departments sold

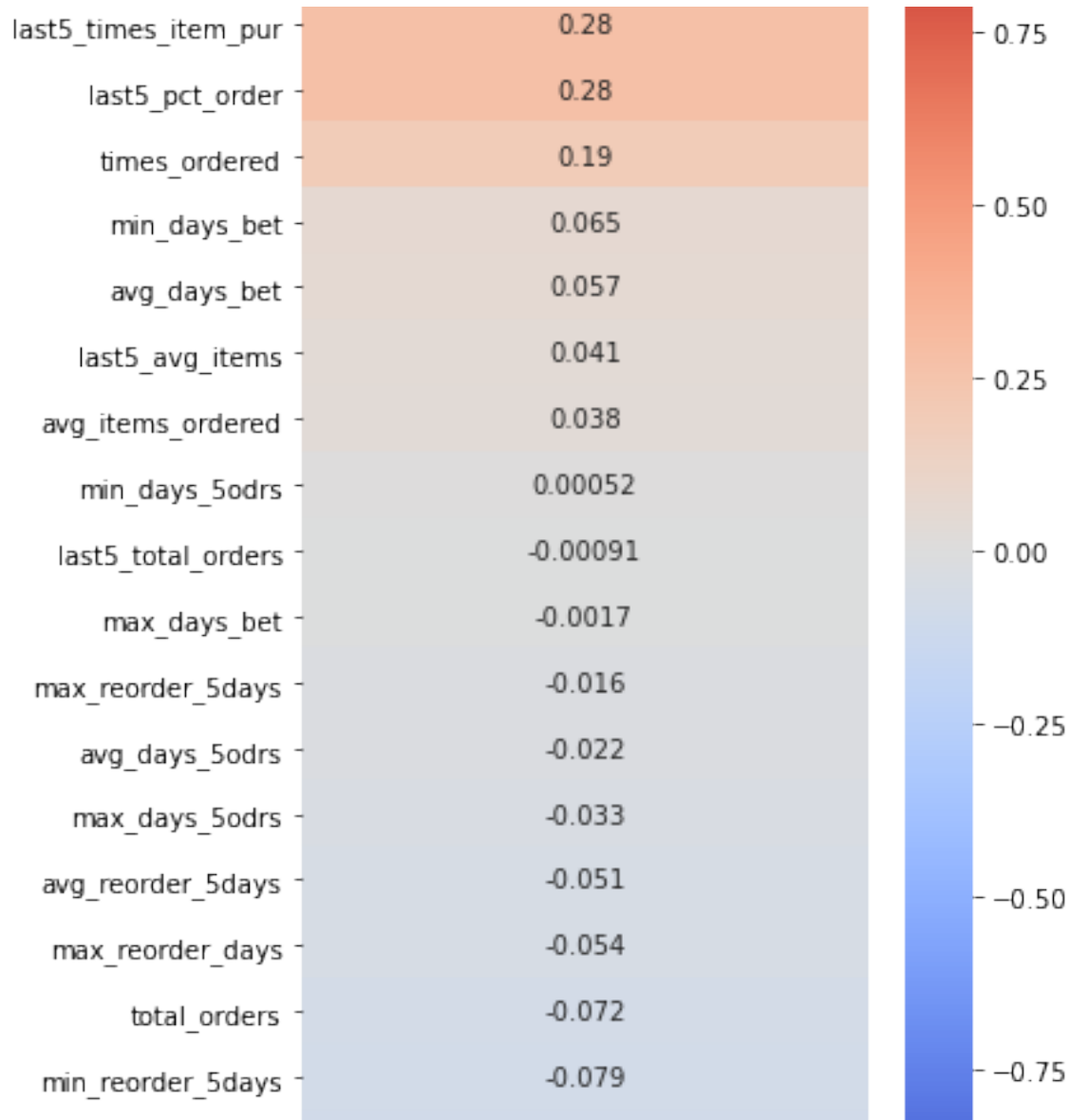


Maximum days between customer's orders



Distribution of number of orders by customer last 5 times vs all orders

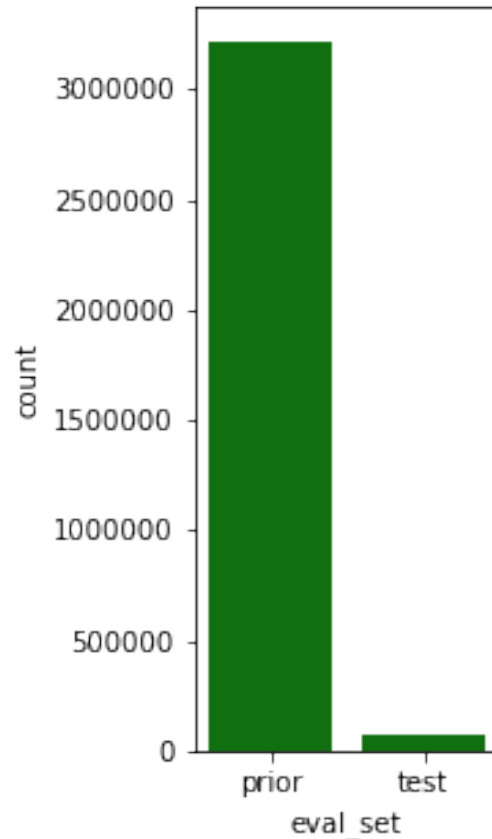




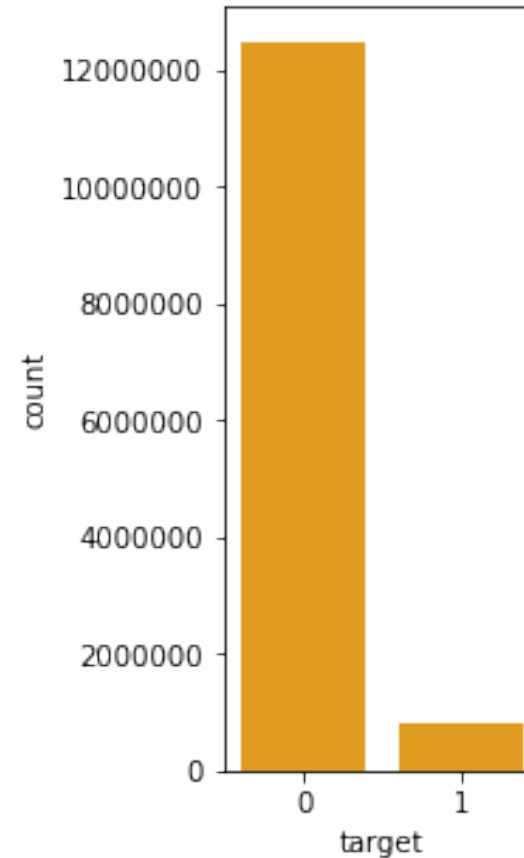
Pct_reorder,
last5_times_item_purc
hased, last5_pct_order
seems to be the most
correlated with the
target (reordered)

Comparing the original dataset vs final...

The train / test set has over 30 million rows with 2.2 % for training / testing model



The aggregated dataset has over 12 million rows with 6.2% of the items recorded





Modeling



Modeling – Specifications



Since we have an imbalanced dataset, the model must have an accuracy score of more than 93.8 to be effective



We will do a train/test split of 80/20 and report on the following

Accuracy

F1-score

Sensitivity (False Positive)

Recall (False Negative)



We also want to minimize False Negatives

Initial Results



On generation of the first model the score was as expected; all predictions were 0 and we therefore got a 93.7 % accuracy



Oversampling was then implemented using SMOTE, this produces a dataset of almost 20 million rows which resulted in long processing times



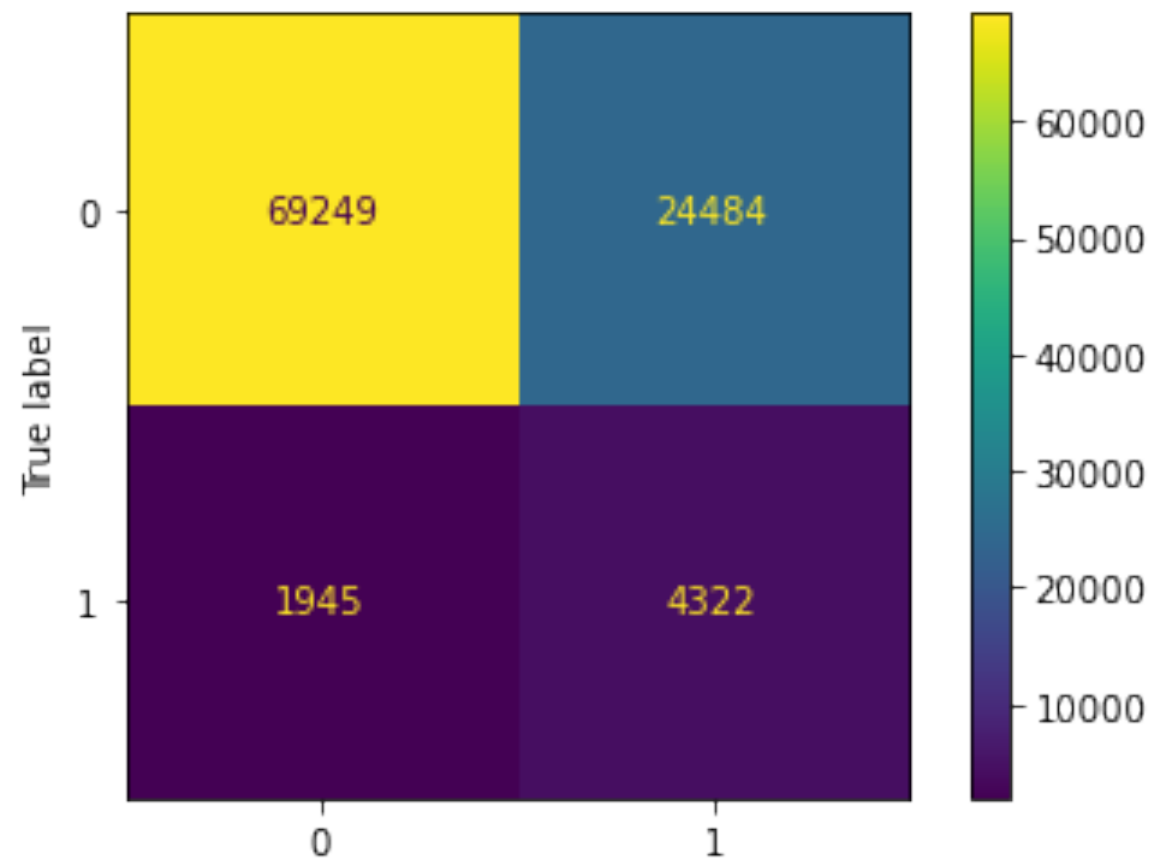
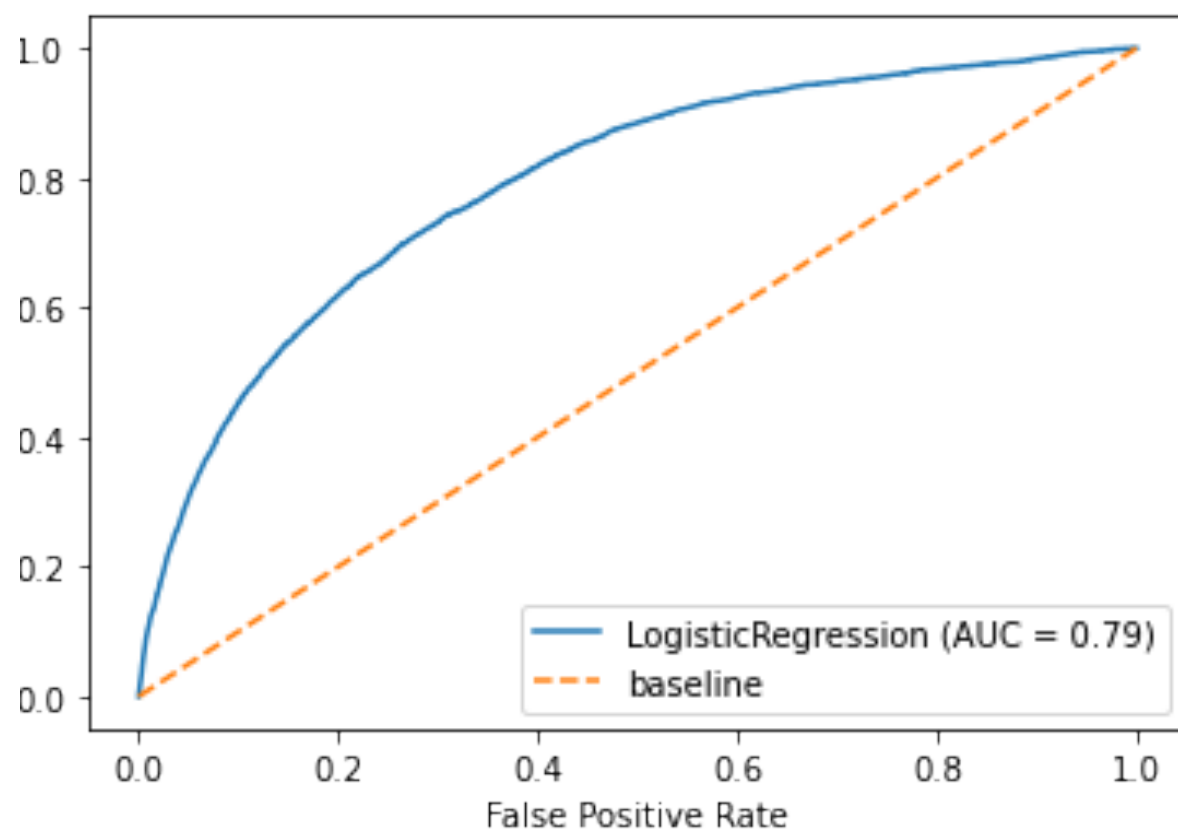
In the end a sample of 500,000 rows were taken SMOTE was then applied after the data was scaled

Modeling Results

Results Based on all numeric aggregate features

Model	Accuracy, Train	Accuracy, Test	Precision	Recall	F1-Score	False Negative
Logistic	0.71	0.74	0.15	0.69	0.25	1945
Kneighbors	0.91	0.77	0.14	0.48	0.21	3296
Decision Tree	0.99	0.89	0.17	0.2	0.19	4939
Random Forest	0.99	0.92	0.27	0.17	0.21	5148
AdaBoost	0.77	0.82	0.18	0.57	0.28	3649
Bagging	0.99	0.92	0.23	0.12	0.16	5426

Best Model
Logistic Regression
Sensitivity: 0.6896 Specificity: 0.7388





Next Steps

- Train the data with other models
- Apply grid search to the best model
- Add / remove other features to get the optimal solution



References

- <https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/33205>
- ["Instacart Market Basket Analysis, Winner's Interview: 2nd place, Kazuki Onodera"](#) by Edwin Chen, dated September 21, 2017.
- <https://www.tomsguide.com/best-picks/best-grocery-delivery-services>
- <https://www.businessofapps.com/data/instacart-statistics/#:~:text=The%20next%20few%20months%20accelerated,had%20passed%20its%202022%20goals>

Thank You!!!!