Project: Capstone Project 1: Milestone Report

Three questions can be answered from this data that can be considered once machine learning is implemented. The first question was the question asked in the Kaggle competition in which this data was created. Based on all of the items purchased, what is a proper item one can predict will be bought in the next order. More specifically, what is the following product a user will buy based on all of the items purchased in the history of the users' orders? The second question that can be answered from the data deals with an instance of when a customer is exploring the items to be purchased. If a customer is exploring the things available, what items can be recommended that will allow them to 'discover' a new product? It is similar to the recommendations that Netflix and Amazon offer on their websites. Perhaps a customer is a frequent buyer of ice cream. I think it would be a safe bet to recommend a product to accompany this ice cream, maybe some sprinkles or chocolate syrup. This idea leads me to the third question that can be answered from the data, which includes, "What products can usually be purchased together?' Some products are purchased together all the time, such as taco shells and chop meat or chicken. One thing cannot be consumed without the other. Let's say a purchase of a Keurig Coffee Machine exists within the purchase history of a user, to aid in increasing one's sales, recommendations can be made for the user to discover new coffee cups or can make it easy for the user to purchase their favorite brand. A common practice with recommendation systems involves making an upsell. This would be recommending a higher end product that ultimately costs more. A sound recommendation system increases sales improves customer experience and provides insight for the customer, the company, and its employees.

I obtained this data from the Kaggle website smoothly and efficiently. The link for the data downloaded a BZip2 Compressed Tar file than contained six CSV files. This was a relational set of files that deals with more than 200,000 Instacart users. Each user had anywhere from 4 to one hundred of their orders. Patoolib is a library that makes extracting data from an archive such as, tar.gz simple. It consists of seven methods to be used and is easily controlled through the os library from python. I used Patoolib to deflate the files from its compression and create data frames. Initially, I created a function that produced the data frames for these six files, but due to assigning an index, I lost the order_id column in one of the data frames. This prevented me from combining all the files and linking user_id with order_id.

After discovering this problem, after reviewing the data again on Kaggle's website, I decided not to use a function and create the six data frames individually. Next came inspecting the data. There were not any null or NaN values that required attention. NaN values existed for first purchases within the days_since_prior column. There was missing information regarding aisle and department for some products, but luckily, they have been filed as missing within the aisle and department data frames. To combine the data into one, more manageable, data frame, the

first data frames combined was the prior and train files. This completes the overall sample and includes all of the orders sampled together in one data frame. The columns consist of order id, product id, add to cart order and reorder and were named full. To eliminate the product id and replace it with a recognizable name, first the aisles and departments were merged with the products data frame. The aisle id and department id were also replaced with the recognizable names of each aisle and department associated with the product name and the numeric identifiers were eliminated. Next, the products data frame was merged with the full data frame to remove the inconspicuous numbers and allow a person to read the names of the products, aisles and departments instead of some arbitrary numbers. The next merge performed allowed the user id to be matched with the order id. This allowed me to give context to each order and each user. Aside from the order id and user id, the new columns included in the full data frame consisted of the days since prior order, the time of day the order was purchased, the day of the week the order was purchased and identified which file this information came from, it could have come from the prior history data or the train data files. It was essential to keep the eval set column for later use within machine learning. After some more thought, I will include even more information to the train data set. This information will be saved for an explanation in a future report. Now that one data frame was constructed, with non-relevant id numbers eliminated and names included to identify instances, exploratory analysis can be completed, and a story can be told from the data.

There were many things one could infer from the data. For example, it was necessary to start simple with some simple questions. How many products do we have to choose from, aisles and departments? There are forty-nine thousand six hundred and eighty-eight available to customers from one hundred thirty-four aisles and twenty-one departments accessible. There is a total of two hundred six thousand two hundred nine customers and three million four hundred twenty-one thousand eighty-three orders. Included in the information, one thousand two hundred and fifty-eight products are not assigned to an aisle or department. There are seventy-seven thousand three hundred ninety-six instances of these products being purchased. The aisles and departments are not completely important, but they should be considered in future exploration. To compare the aisles and departments concerning the products available a bar chart (Fig 2) was included with color-coded instances to determine popular aisles or departments these items come from. This bar chart showed the popularity of each department and what aisle was involved. Missing data was the most populated, but that is a reasonable consideration. The top twenty items purchased (Fig 3) were computed and it seems that organic items are trendy, making up half of the top ten purchased items list. Bananas and organic bananas were by far the most purchased items and had the highest probability of being selected. I was interested in whether or not the most frequent items purchased were also the most frequent items reordered (Fig 4). One could deduce this from speculation, but it can be proved with the data. After all, with almost half

of the items on both lists, it should be tested to see if there is a difference. It was also essential to look at the items that were chosen first. (Fig 5). This information was included in the add to cart order column. Were the items reordered and purchased the most also the first pick for a customer? Again, there were similar items, but I will hold off on more exploration until the statistics portion of this analysis. Fifty-four percent of the items available have been reordered before in this sample, and the rest have not been reordered. (Fig 6) I think the reorder column and the order purchased is critical in informing us about what should be recommended to the customer.

Continuing exploration, there was a mean of sixteen orders per users (Fig 7) and a mean of ten items per purchase. (Fig 8). This can give us a basket size to think of for later use, depending on how we decide the recommendation system will work. An additional column had not been analyzed in terms of distribution. Days since prior column had a mean of approximately sixteen and had a sharp increase at the end of its distribution. This was due to the customers who no longer purchased after thirty days. (Fig 9). Next, a scatter plot was used to determine if there was a relationship between the number of days since the last order and the basket size of the order, in other words, how many items were bought in relation to the number of days it had been since the last order.(Fig 10). This plot distribution showed a fan-shaped distribution and a noticeable non-linear relationship. An appealing aspect of the data to measure is the relationship of time with orders and the items purchased. A heat map was created to visualize the relationship of the time of day and the day of the week these orders took place. (Fig 11). This showed that most of the orders occur within the hours of seven in the morning to about eight at night. I would consider these the peak hours meaning that the ten percent threshold of orders were crossed.(Fig 13). Most of these orders occur early in the week.(Fig 12). One noticeable difference is the time of day on the weekend in which these orders take place. Perhaps because people like to sleep in on the weekends, most orders made occur in the afternoon rather than the morning. The frequency of such orders does increase on the weekends. Saturday has the most, Sunday has second most and Monday has the third most orders during the week. Perhaps this is when the customers like to stock their refrigerator for the week.

After some exploratory data analysis, one type of product was showing up frequently. Organic food seemed to be trendy and considering produce was the most popular department, this idea had to be explored further (Fig 14) to test if there was a difference in reordering rate based on a product being organic or non-organic, a Bayesian approach was chosen to be the most efficient approach. Probability within this sample would allow us to quantify our beliefs on how different organic and non-organic products are. In the initial step to using a Bayesian method, a beta distribution was created. Looking at the situation as binary, alpha was used as the products that were organic and beta as the non-organic products. Samples were generated from this data

one thousand times using Numpy's Beta method. Once the distributions were plotted from the samples it was easy to see that the graphs of the two products never converged. (Fig 15). This was evidence enough to conclude that these two products had a significant likelihood of having different reorder rates. The distribution of each graph was input into a histogram to endure its normal distribution.(Fig 16, 17). A binomial distribution was not needed to multiply with the beta distribution. To get a percentage of likelihood, the two samples were subtracted from each other and the mean was found of the difference. Because of the large difference in the distributions, every value was over zero and therefore a one hundred percent confidence interval could be established.

Another hypothesis was tested using a Bayesian method. In this case, the data was not binary, and a different distribution needed to be used. To trade off the distributions used, instead of a beta distribution, a Dirichlet distribution was used. To trade the binomial distribution for the distribution of the likelihood, a multinomial distribution was used keeping in mind the conjugate distribution. The hypothesis tested was if the top thirty items purchased were also the top thirty items reordered. Again, dealing with likelihood, a Bayesian method was more appropriate. First, a Dirichlet distribution provided samples. (Fig 18). This could have been done multiple times, but there was enough evidence to use just one instance. Also, because the graphs of the data converged so closely, it was evident that this was very likely. The multinomial distribution samples were calculated, and its values counted to establish the likelihood of the equation. (Fig 19). The Dirichlet prior was then multiplied with the multinomial distribution, likelihood, and confidence was set. It was found that only forty percent of the values were over zero maintaining the null hypothesis that there was no difference in the top items purchased and top items reordered. There was little difference between the two. (Fig 20).

Two other hypotheses test were performed using Frequentists statistics. These two hypotheses tested if the order in which the product was picked affected its reorder rate. Items that were selected first were to be tested first, and then any order number was examined. The means of each sample were calculated, and its standard deviation computed by pooling together the variance. Once the standard error could be computed, a T-value was chosen from a T- table and used in the two-tailed T-test. The confidence interval for both hypotheses tested did not contain zero, and therefore, the null hypotheses were rejected. It was found that item's chosen order does affect whether they are reordered.

Obtaining the data for this particular project was easy compared to my project regarding NYC parking tickets. There is so much to continue to explore but I am now looking forward to using the information gathered during the EDA to create an efficient model. A sound recommendation system increases sales improves customer experience and provides insight for the customer, the company, and its employees. Given a large set of sample data, it will be interesting to continue to explore and improve on the machine learning model to come.

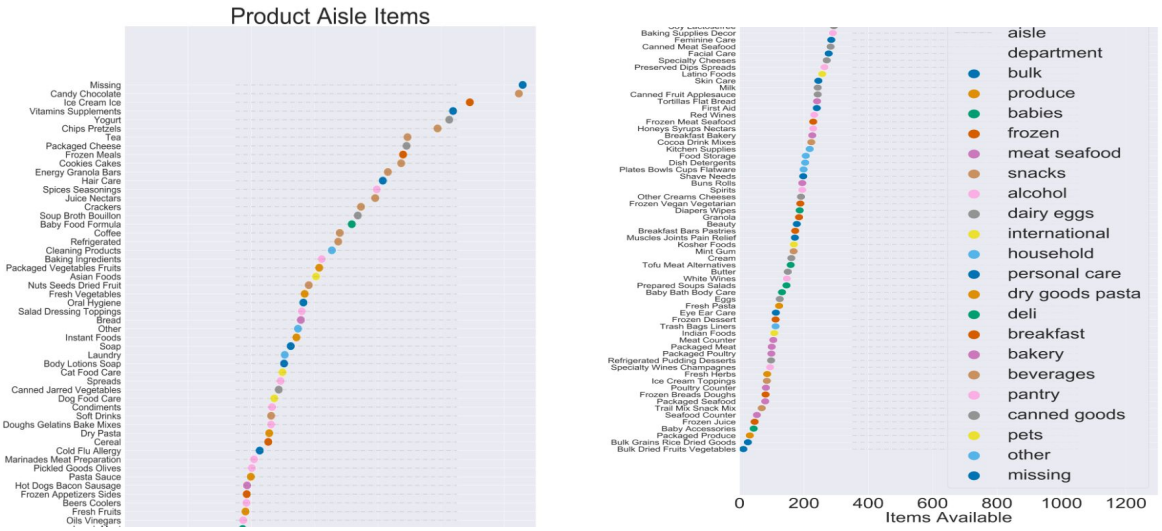
Product Aisle Items
aisle
department
bulk
produce
babies
frozen
meat seafood
snacks
alcohol
dairy eggs
international
household
personal care
dry goods pasta
deli
breakfast
bakery
beverages
pantry
canned goods
pets
other
missing
0
200
400
600
800
1000
1200
Items Available

Fig 2

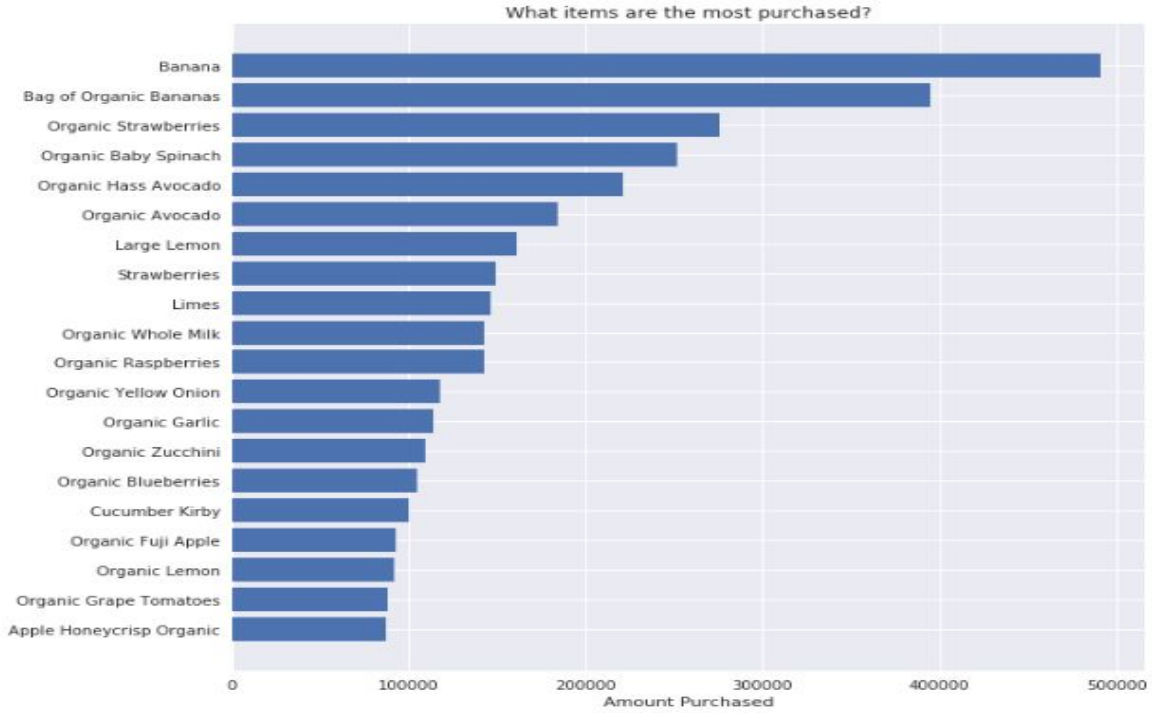
What items are the most purchased?
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Organic Hass Avocado
Organic Avocado
Large Lemon
Strawberries
Limes
Organic Whole Milk
Organic Raspberries
Organic Yellow Onion
Organic Garlic
Organic Zucchini
Organic Blueberries
Cucumber Kirby
Organic Fuji Apple
Organic Lemon
Organic Grape Tomatoes
Apple Honeycrisp Organic
0
100000
200000
300000
400000
500000
Amount Purchased

Fig 3

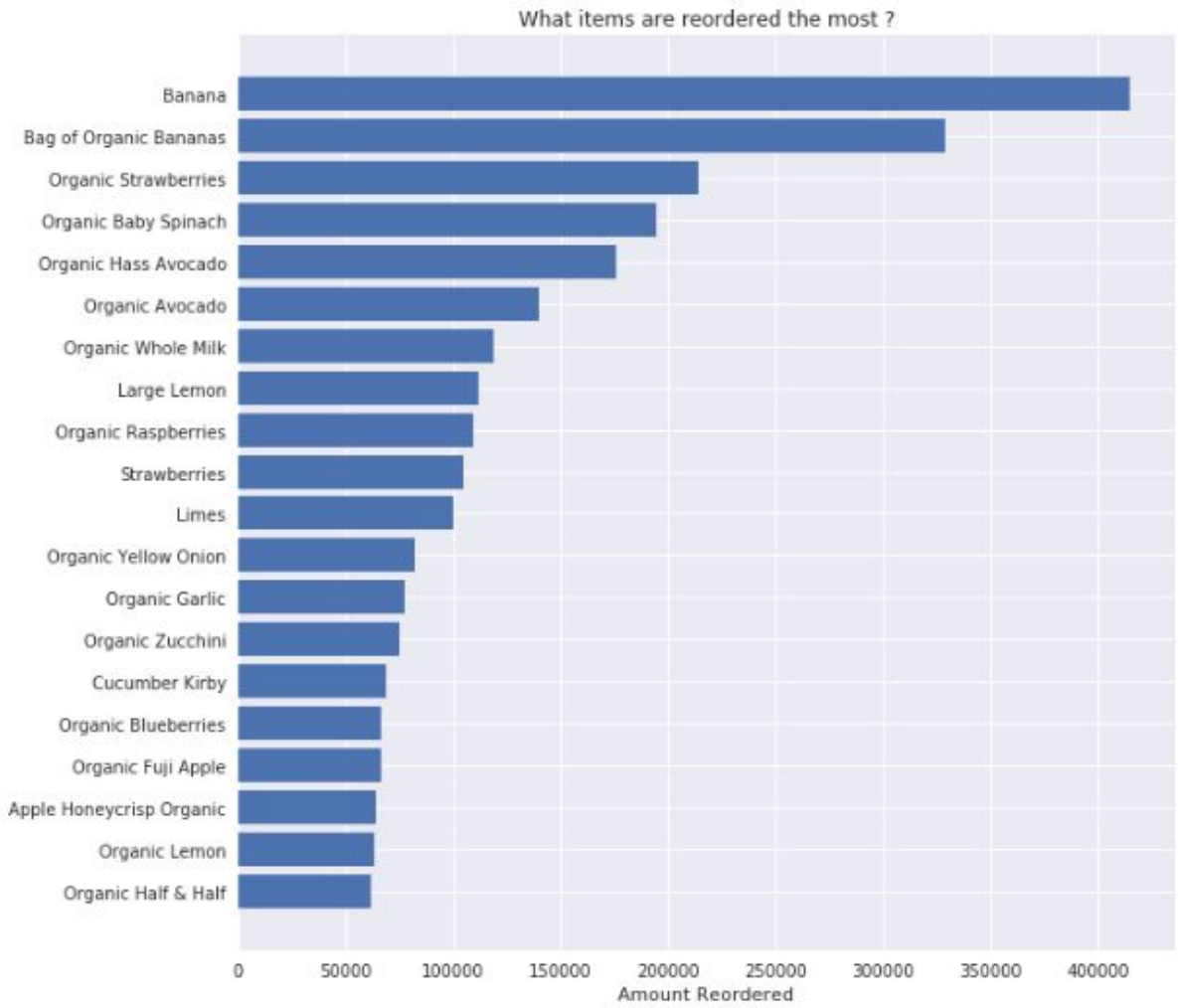
What items are reordered the most ?
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Organic Hass Avocado
Organic Avocado
Organic Whole Milk
Large Lemon
Organic Raspberries
Strawberries
Limes
Organic Yellow Onion
Organic Garlic
Organic Zucchini
Cucumber Kirby
Organic Blueberries
Organic Fuji Apple
Apple Honeycrisp Organic
Organic Lemon
Organic Half & Half
0
50000
100000
150000
200000
250000
300000
350000
400000
Amount Reordered

Fig 4

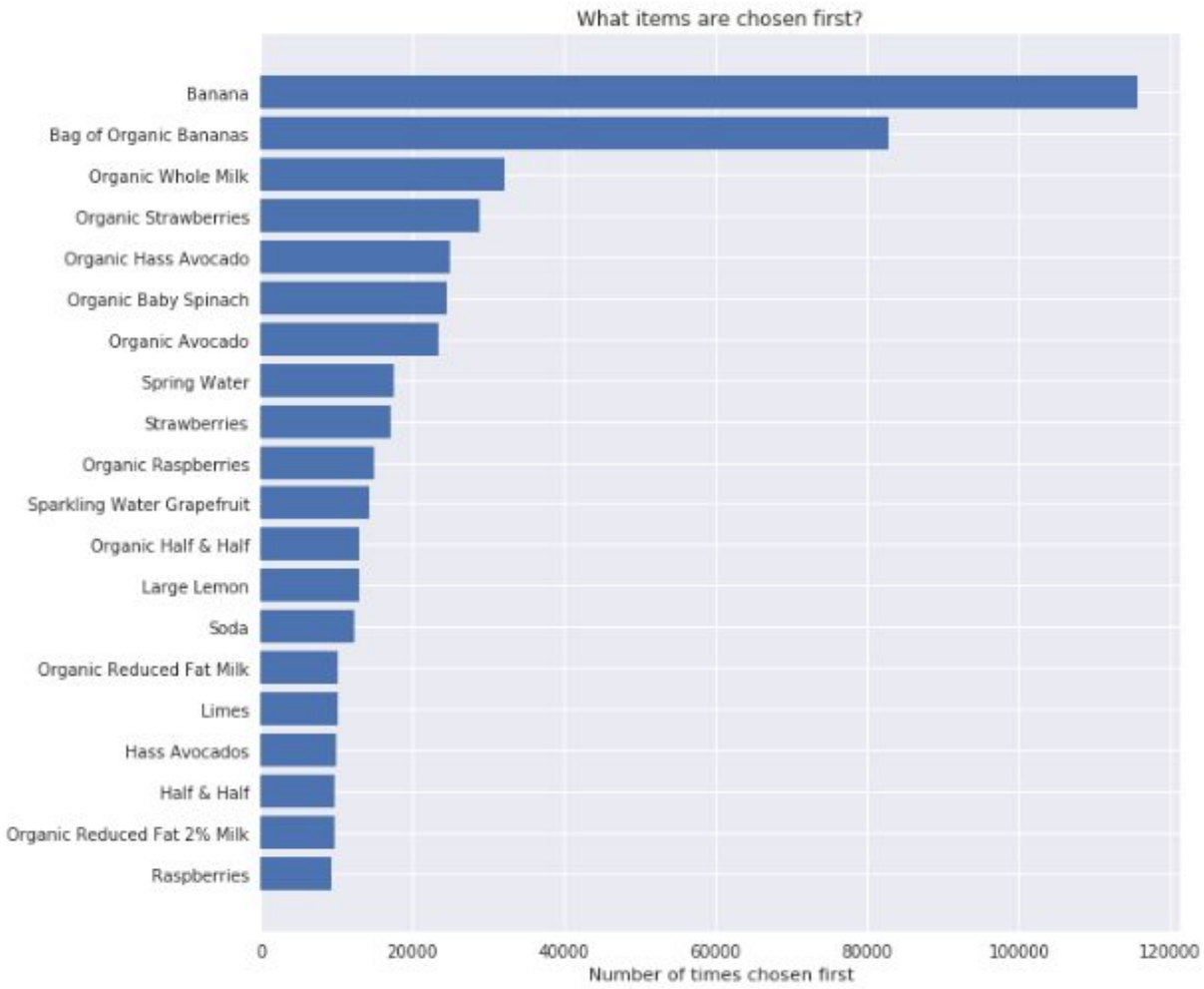
What items are chosen first?
Banana
Bag of Organic Bananas
Organic Whole Milk
Organic Strawberries
Organic Hass Avocado
Organic Baby Spinach
Organic Avocado
Spring Water
Strawberries
Organic Raspberries
Sparkling Water Grapefruit
Organic Half & Half
Large Lemon
Soda
Organic Reduced Fat Milk
Limes
Hass Avocados
Half & Half
Organic Reduced Fat 2% Milk
Raspberries
0
20000
40000
60000
80000
100000
120000
Number of times chosen first

Fig 5

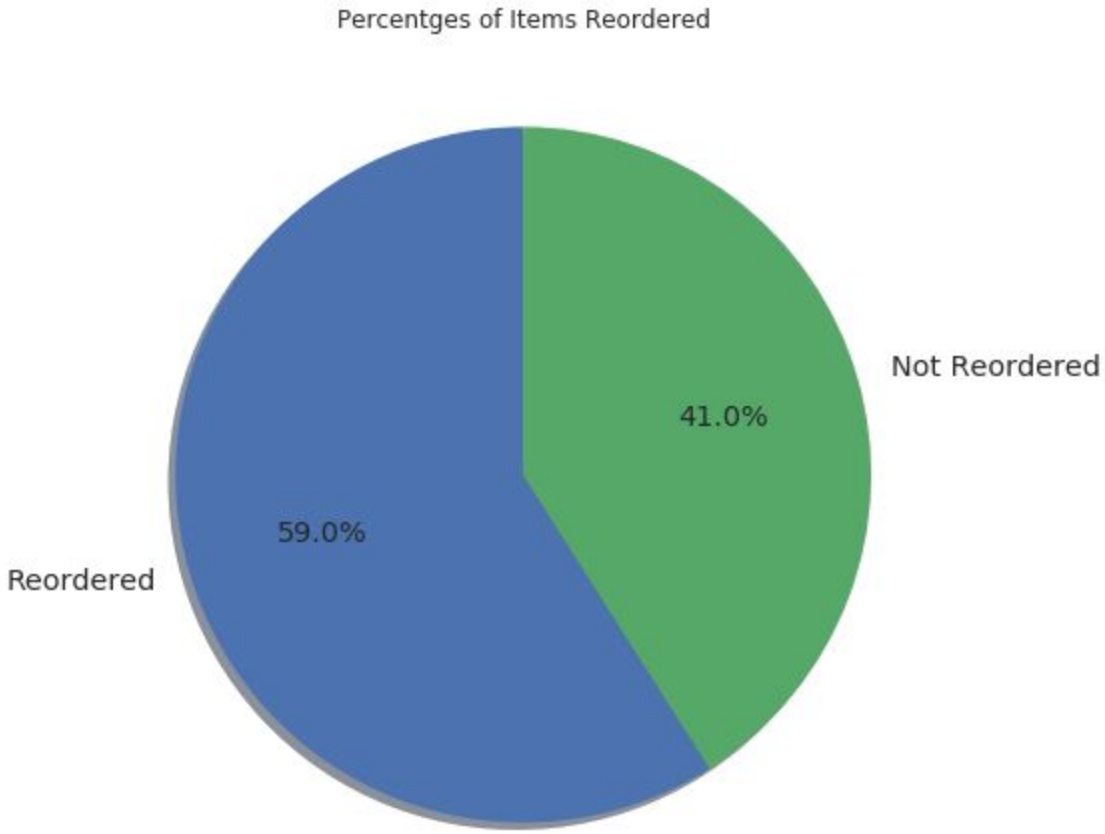
Percentges of Items Reordered
Not Reordered
41.0%
59.0%
Reordered

Fig 6

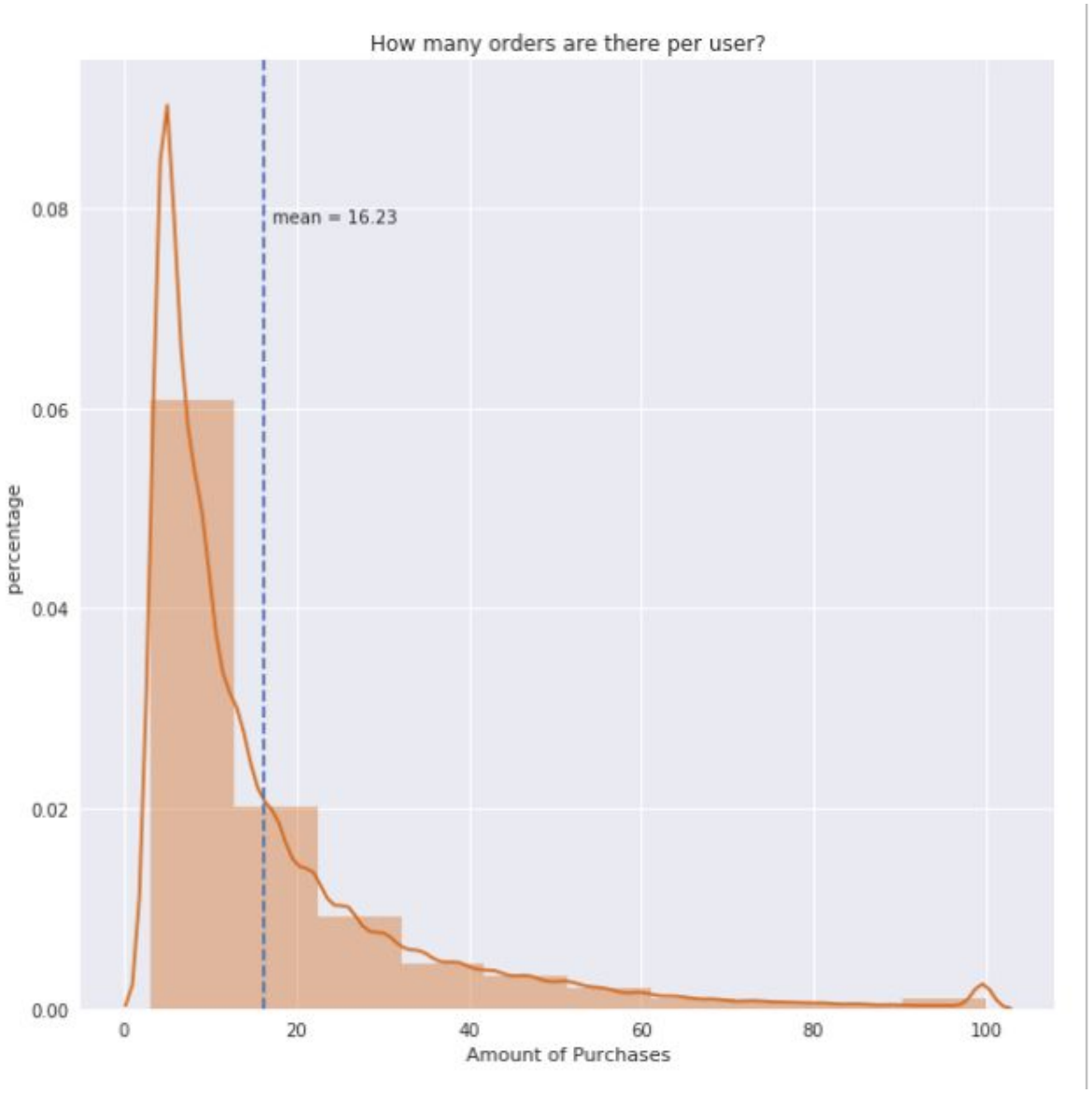
How many orders are there per user?
mean = 16.23
percentage
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
Amount of Purchases

Fig 7

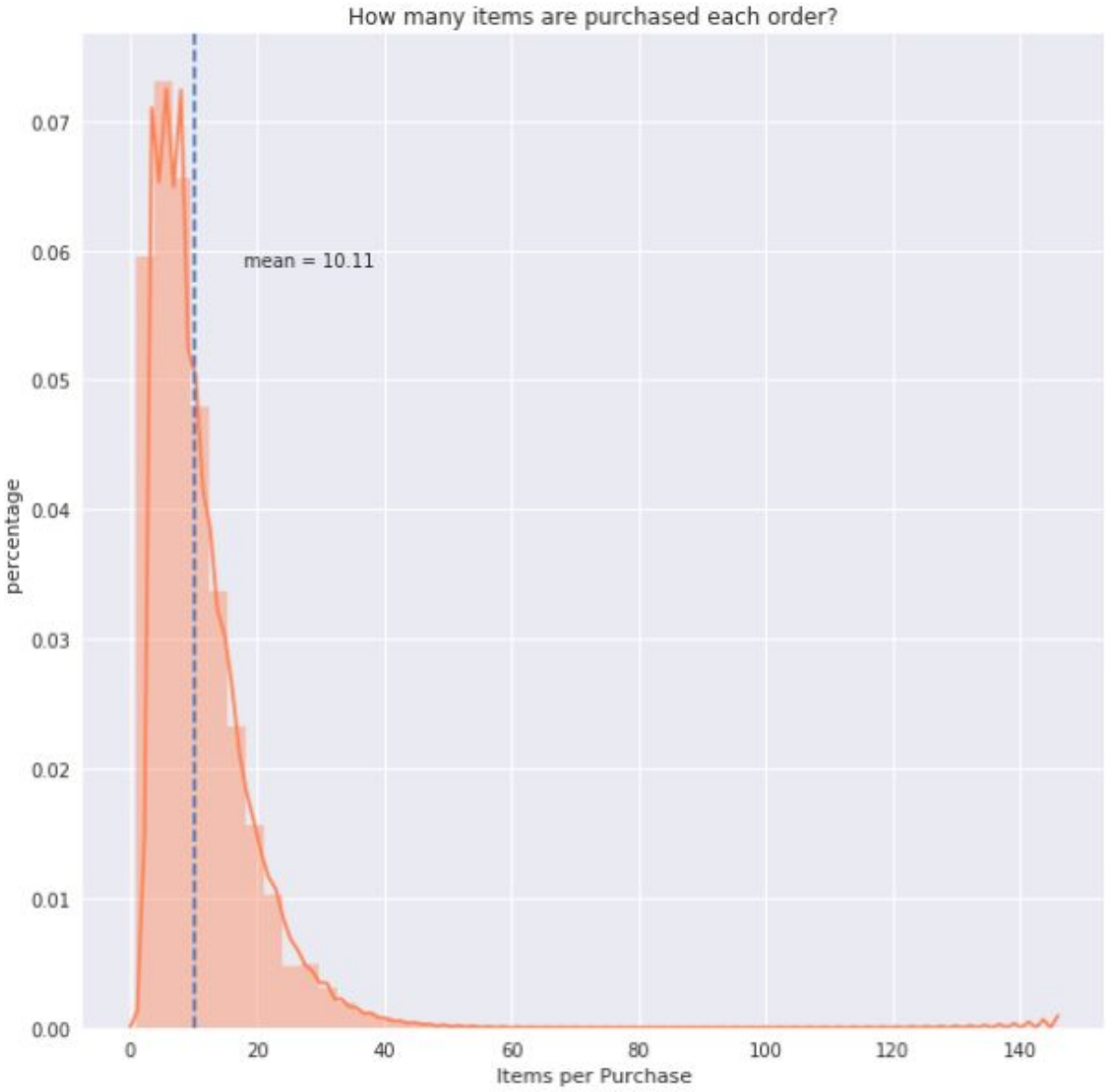
How many items are purchased each order?
mean = 10.11
percentage
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
20
40
60
80
100
120
140
Items per Purchase

Fig 8

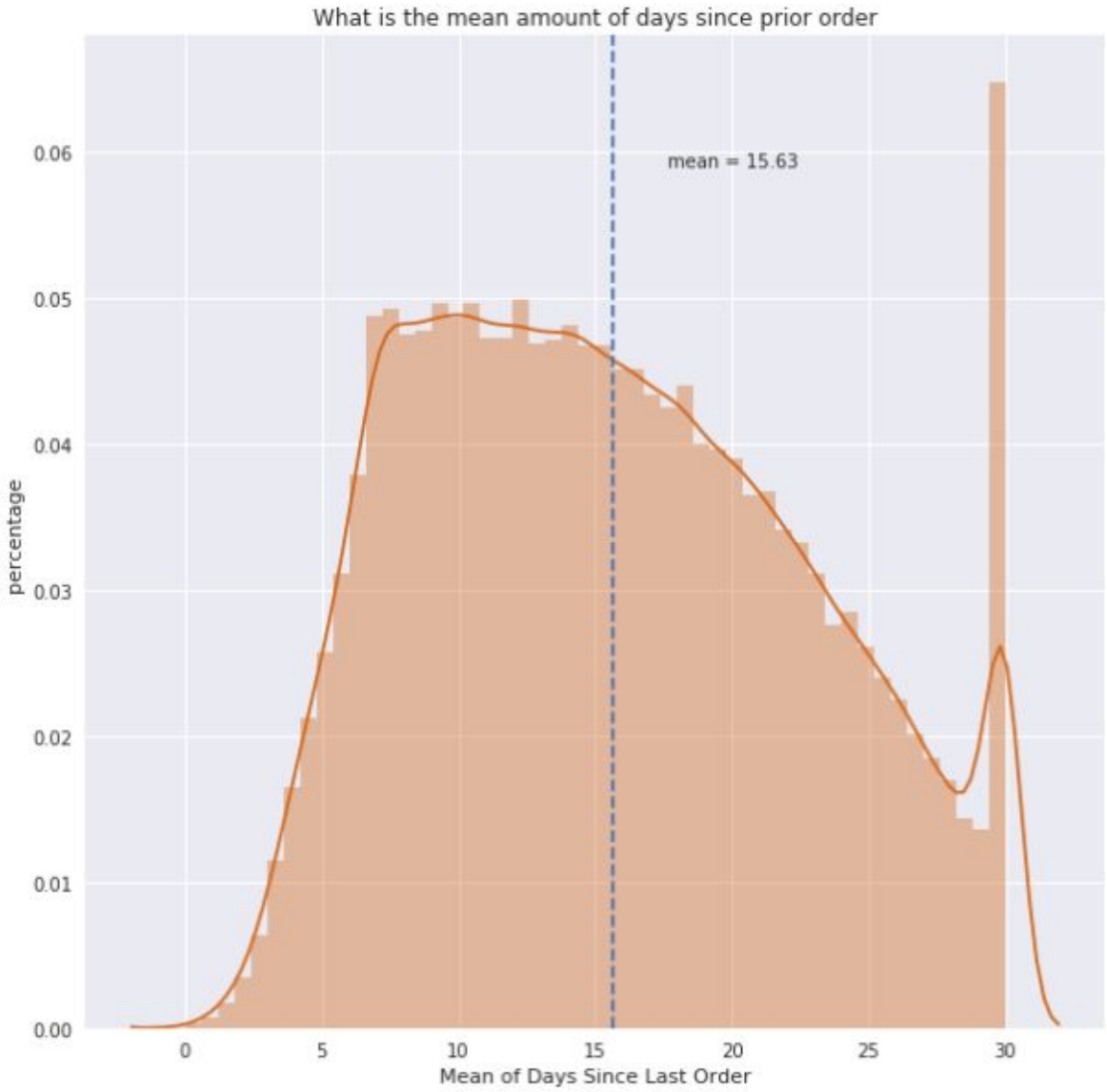
What is the mean amount of days since prior order
mean = 15.63
percentage
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
5
10
15
20
25
30
Mean of Days Since Last Order

Fig 9

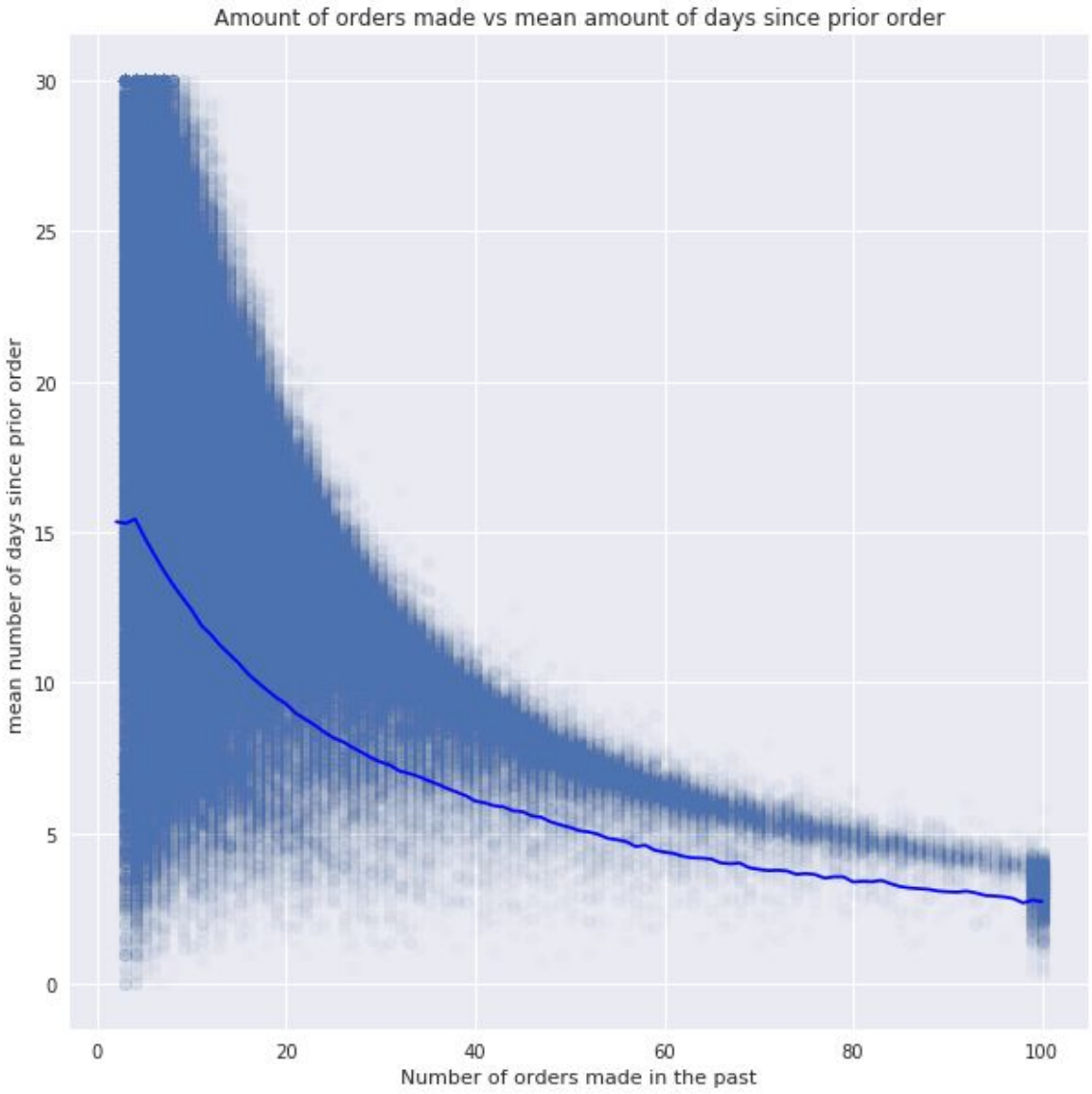
Amount of orders made vs mean amount of days since prior order
mean number of days since prior order
Number of orders made in the past
0
5
10
15
20
25
30
0
20
40
60
80
100

Fig 10

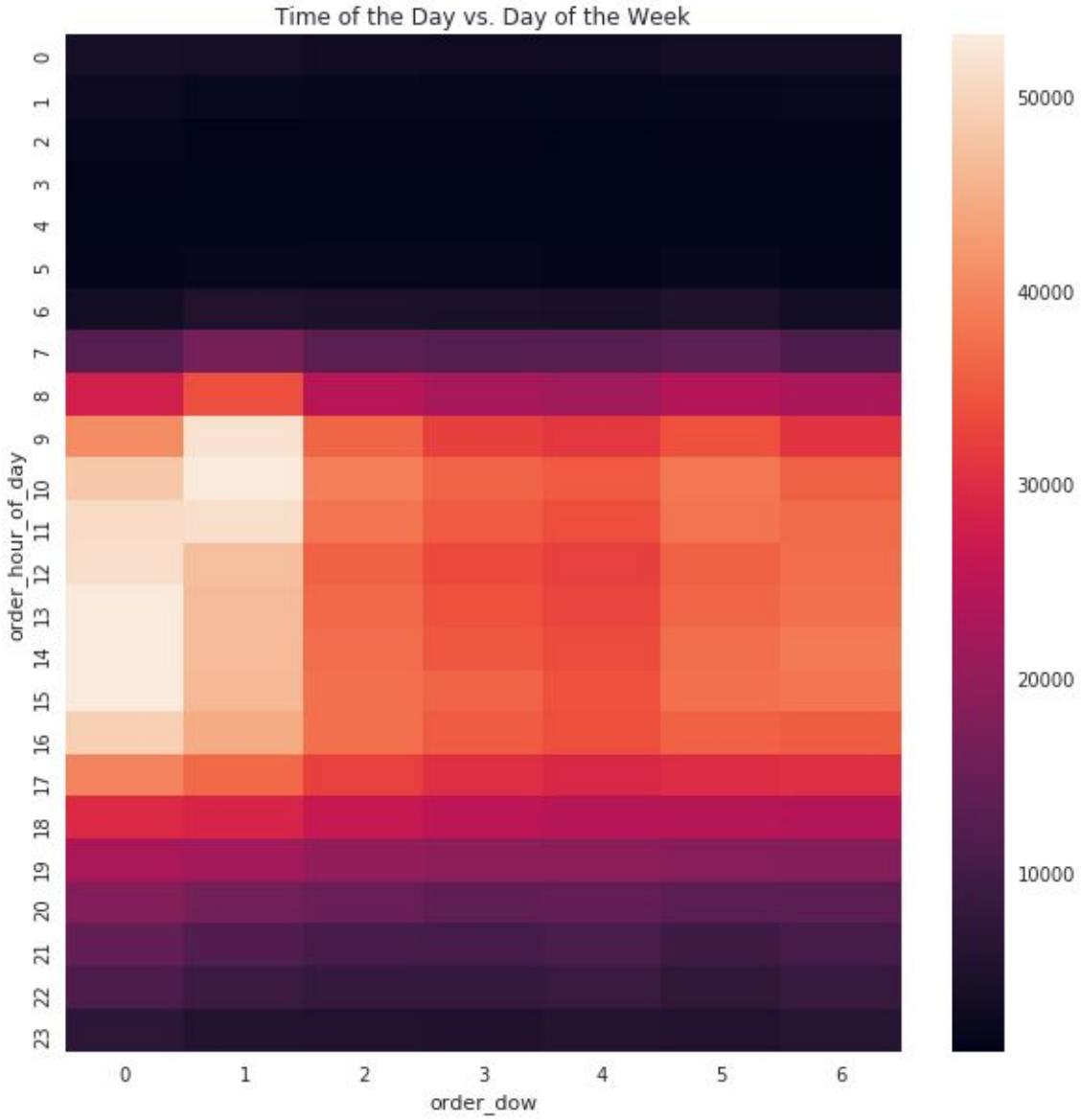
Time of the Day vs. Day of the Week
order_hour_of_day
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
order_dow
0
1
2
3
4
5
6
50000
40000
30000
20000
10000

Fig 11

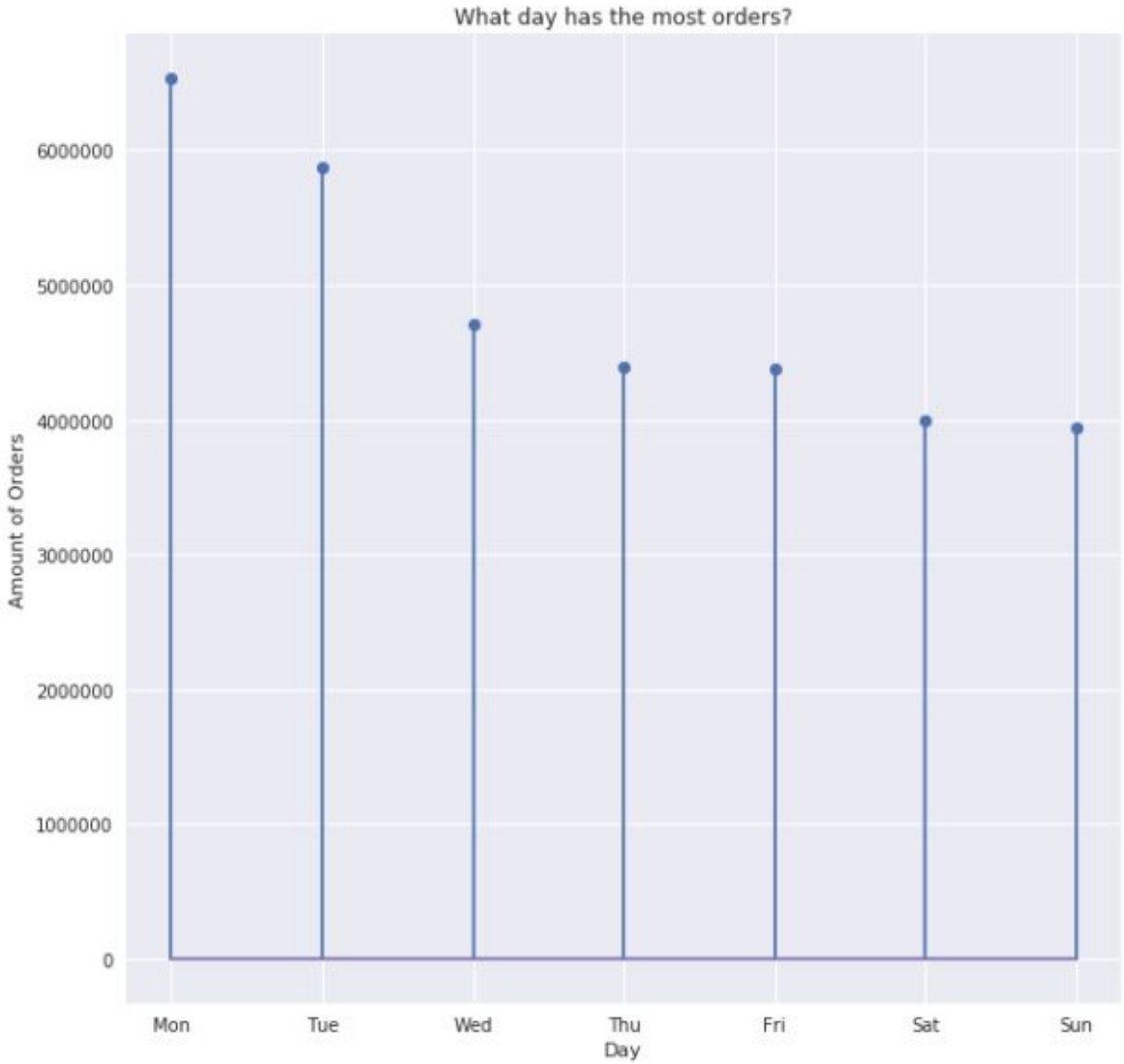
What day has the most orders?
Amount of Orders
6000000
5000000
4000000
3000000
2000000
1000000
0
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Day

Fig 12

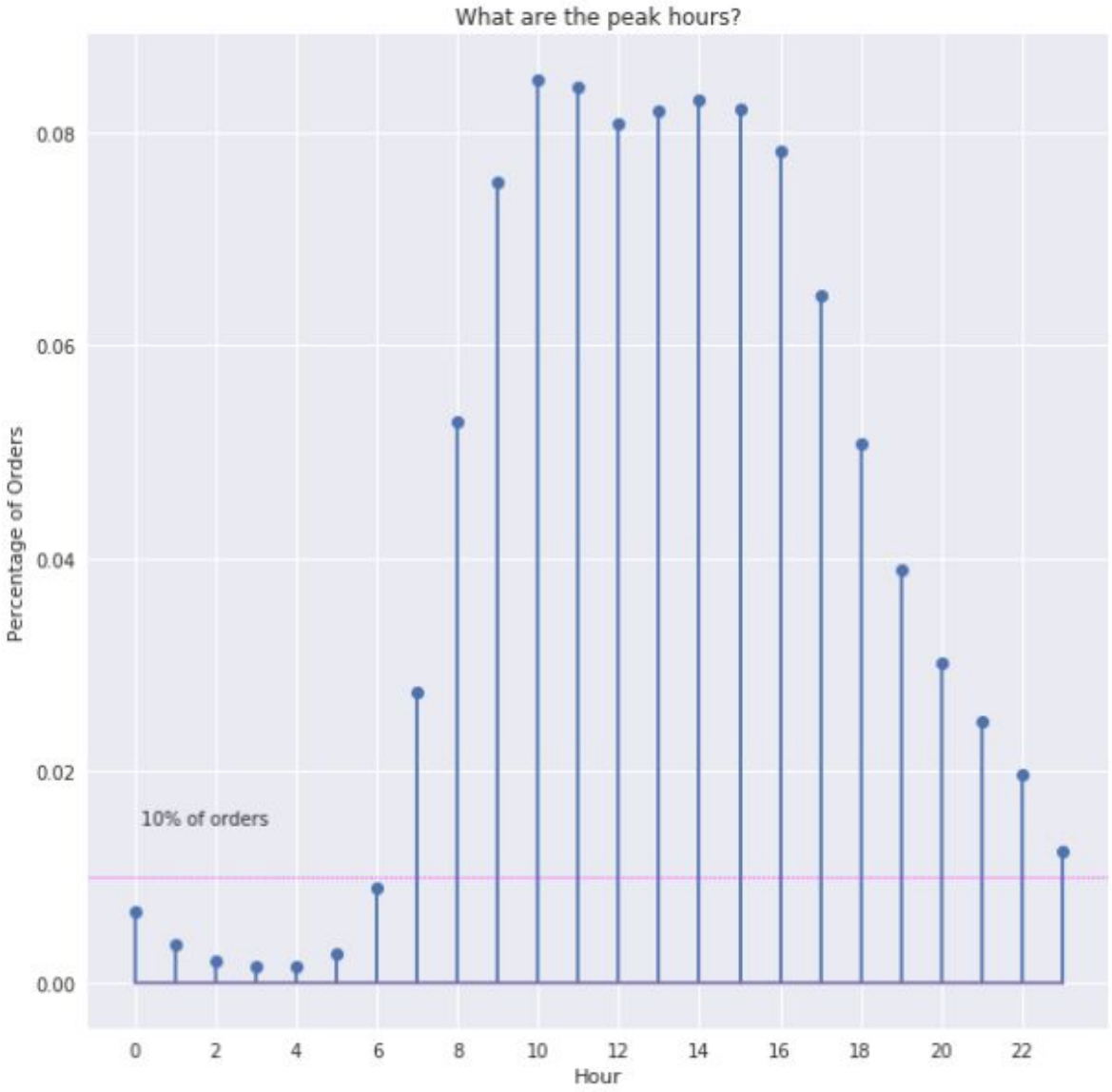
What are the peak hours?
Percentage of Orders
0.08
0.06
0.04
0.02
0.00
10% of orders
0
2
4
6
8
10
12
14
16
18
20
22
Hour

Fig 13

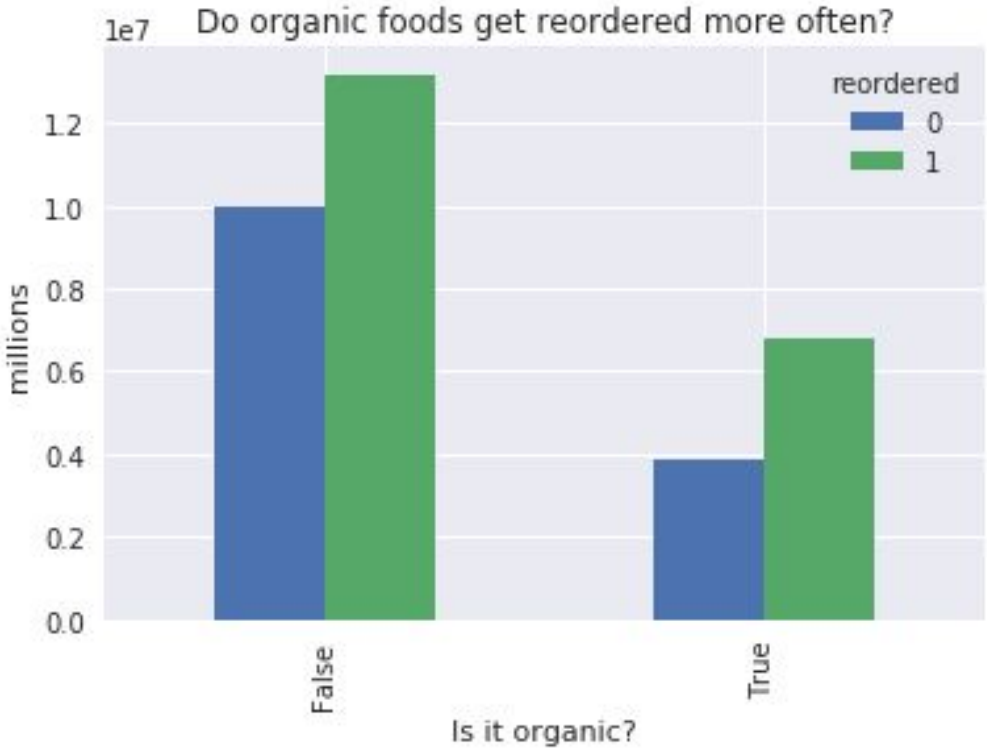
Do organic foods get reordered more often?
1e7
millions
1.2
1.0
0.8
0.6
0.4
0.2
0.0
reordered
0
1
False
True
Is it organic?

Fig 14

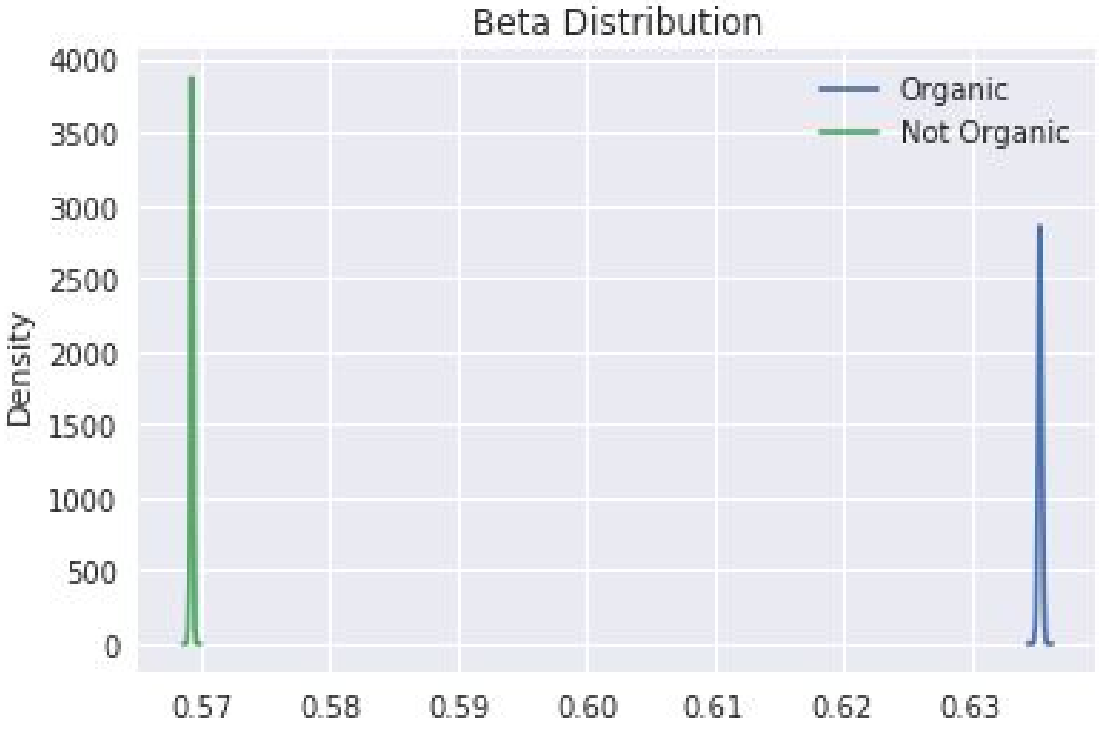
Beta Distribution
Organic
Not Organic
Density
4000
3500
3000
2500
2000
1500
1000
500
0
0.57
0.58
0.59
0.60
0.61
0.62
0.63

Fig 15

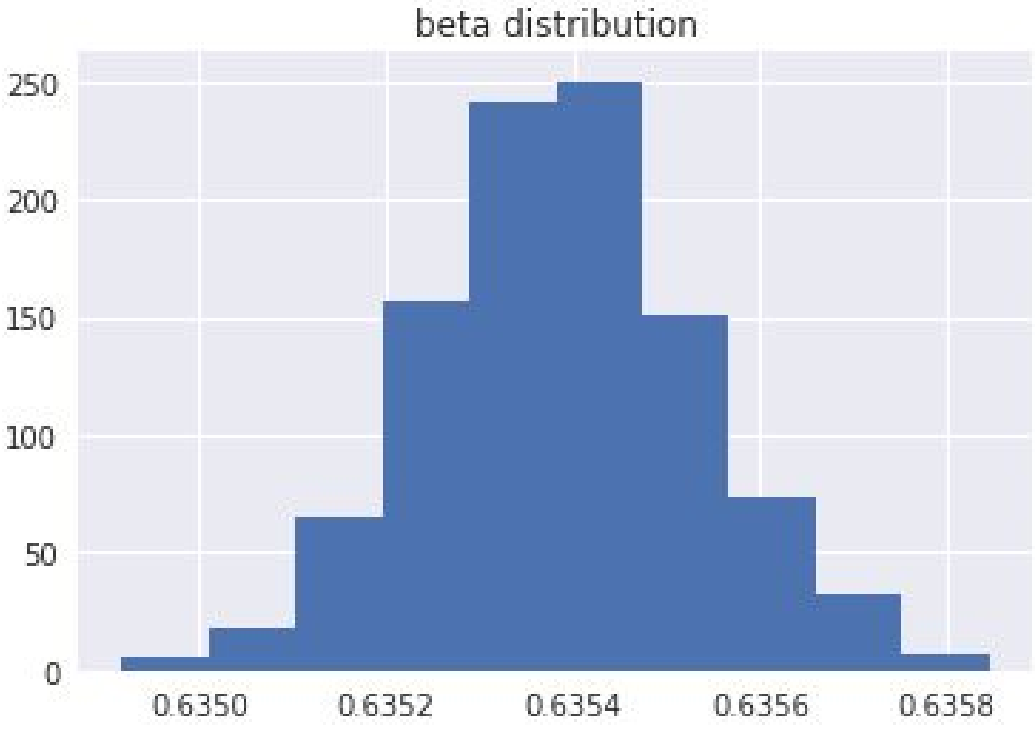
beta distribution
250
200
150
100
50
0
0.6350
0.6352
0.6354
0.6356
0.6358

Fig 16

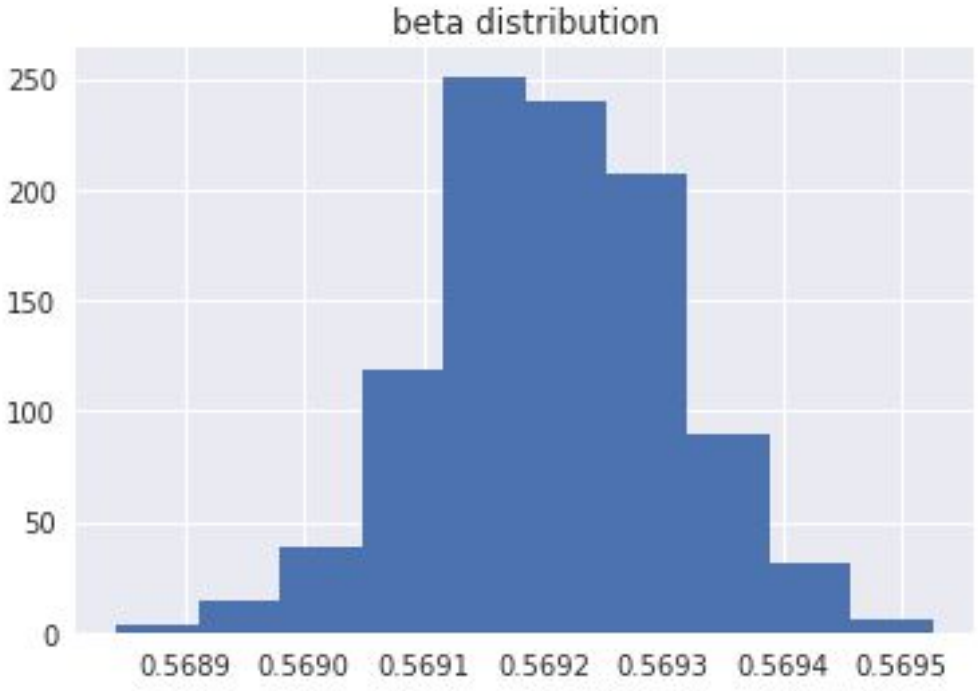
beta distribution
250
200
150
100
50
0
0.5689
0.5690
0.5691
0.5692
0.5693
0.5694
0.5695

Fig 17


Beta Distribution
Purchase
Reorder
Density
25
20
15
10
5
0
−0.05
0.00
0.05
0.10
0.15

Fig 18

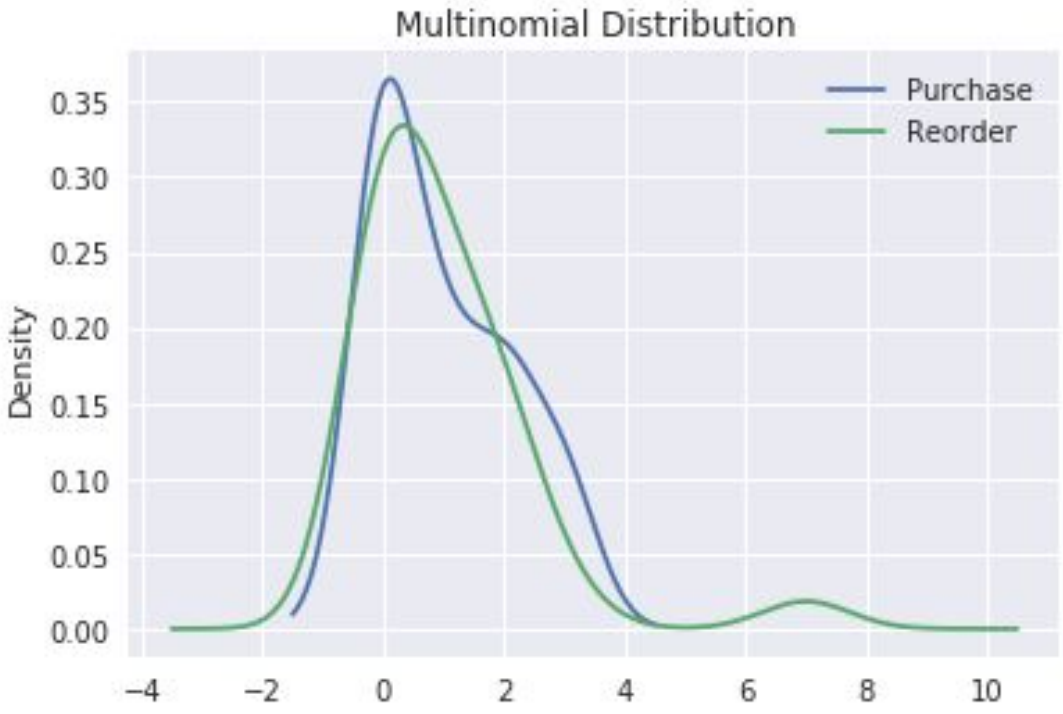
Multinomial Distribution
Purchase
Reorder
Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4
6
8
10

Fig 19

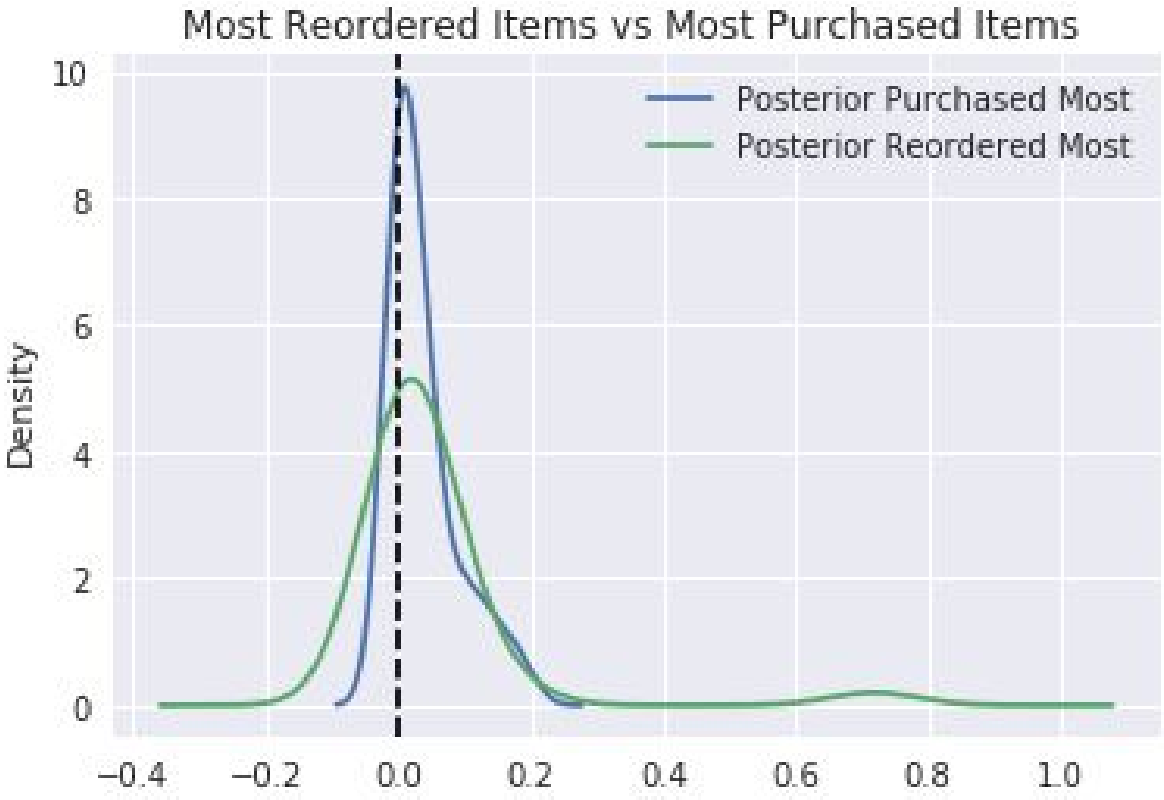
Most Reordered Items vs Most Purchased Items
Posterior Purchased Most
Posterior Reordered Most
Density
10
8
6
4
2
0
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0

Fig 20