# Abstract

With the rise in online transactions, companies are trying to leverage the humongous data generated by the transaction activities to transform it to meaningful insights. Data Mining techniques can be used to develop a cross selling strategy for the products. Data scientists use predictive analytics to improve the customer experience of shopping online by developing models that predict which products a user will buy again, or try for the first time, or which products are bought together.

In this paper, we analyze the trend in customer shopping behavior on Instacart website for buying groceries. The data set is made public by Instacart- a same day grocery delivery service, for 3 million transactions of over 200,000 users. Data set is explored using open source statistical learning tool R.

Market Basket Analysis is done using Apriori algorithm for various support levels, confidence and lift to suggest combination of products to be included in a basket to cross sell the products on the platform. The model is developed to predict which previously purchased products will be in a user’s next order. The F-score measures the model’s performance.

# Introduction

## Motivation

Data Science plays a huge role in the e-commerce sector. A large volume of customer data is available to companies which they can use to mine patterns and understand customer behavior. Supply chain can be improved, warehousing can be made more efficient, or the product that a user might wish to buy can be recommended based on the predictive models.

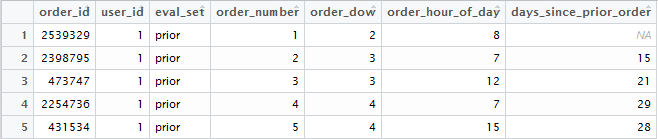
Instacart is an internet – based grocery delivery service with a slogan of *Groceries Delivered in an Hour.* The purpose of this paper is to analyze the trend in customer buying pattern on Instacart, suggest combination of products which can be sold together under various offers.

# Data Set

Instacart recently made their 3 million Instacart Orders data publicly available called “The Instacart Online Grocery Shopping Dataset 2017”. The data set is a relational set of files describing customers’ orders over time. The data set is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, between 4 and 100 of their order details is provided with the sequence of products purchased in each order, the week and hour of day the order was placed and a relative measure of time between orders.

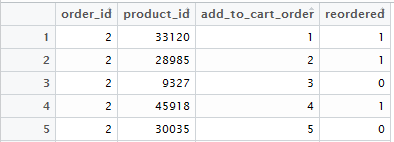
The name of the data files with the description of data contained in them is described below.

1)Orders



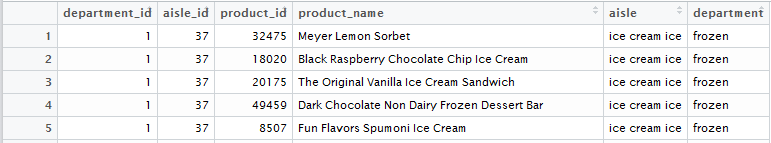
The data set contains list of unique order\_id for corresponding to order made by users. Order\_number gives the number of the order. Eval\_set denotes if the order is a prior order, train, or test. All but the last order of every user is classified as prior. Last order of every user is either classified as train or test. The ones classified as test are the order\_id for which we predict which products will be included in the next order. Order\_dow gives the day of the week and order\_hour\_of\_day denotes hour of the day. Days\_since\_prior\_order gives the time difference between two orders and contains NULL value for the first order of every user. There are 3 million plus order\_id for 200,000 plus different users.

2)Prior



Prior table contains product\_id for every order\_id. It thereby gives information about products included in every order. Add\_to\_cart\_order gives the order for product\_id by which it was added by customer to their shopping cart. Every product\_id is classified and coded as 1 under reordered column if it was previously ordered by customer and 0 otherwise. It is the largest table with over **32 million rows** of data.

3)Products, Aisles, and Department



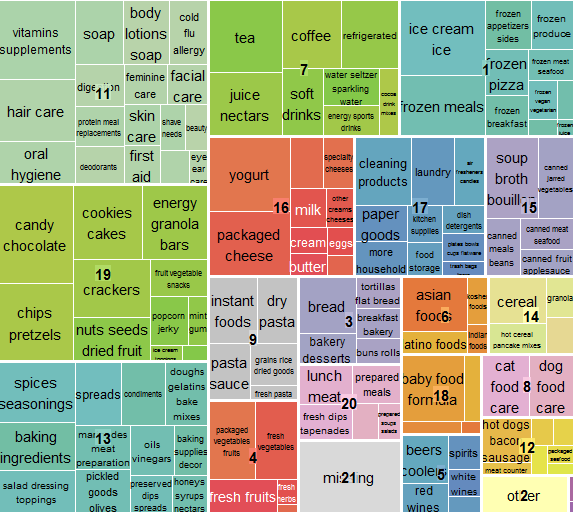
Products table contains product\_id and product name, aisles table gives aisle\_id and aisle name and so does department table.

## Product Offerings

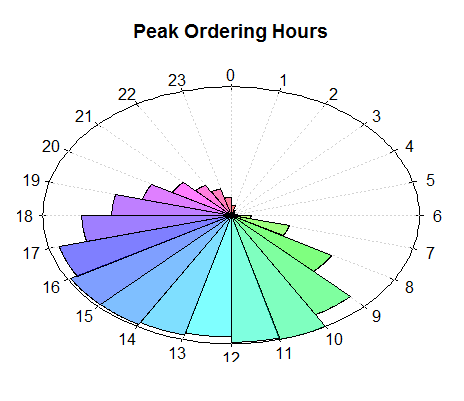
Instacart offers a wide variety to its customers comprising of about 50,000 different products spread across 134 aisles under 21 different departments. This can be visualized in the tree map below. All the aisles are common in color and their size is representing the different number of products contained in them.

It is found that candy chocolate, Ice cream ice and vitamins supplement aisles offer maximum number of products. Over 1250 products are found in aisle titled missing in department called missing. For the list of these products refer Appendix A. Top and Bottom 20 Aisles by number of product offering can be visualized in Appendix A. It can be inferred from the below table that Instacart has maximum number of product offerings across personal care and edible item departments. Detailed list is given in Appendix A.





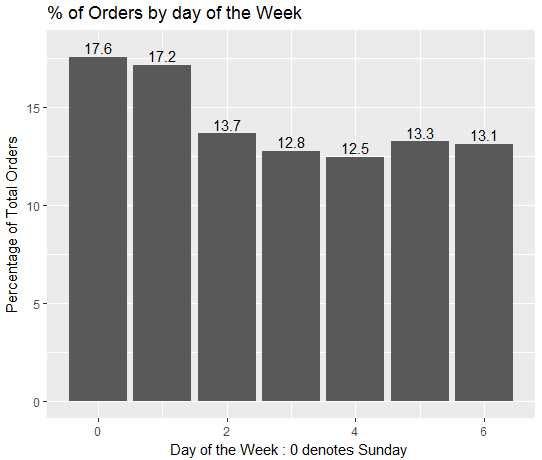
## Order Trend Analysis



Inference: From the figure above it is inferred that most people order between 9:00 AM to 6:00 PM in the evening. Instacart can accordingly plan to hire persons for delivery during days shifts. The visualization is plotted in R with colored portion representing the relative percentage of total orders across the day with 10:00 AM being the busiest hour (100% in vis).

## Day of the Week

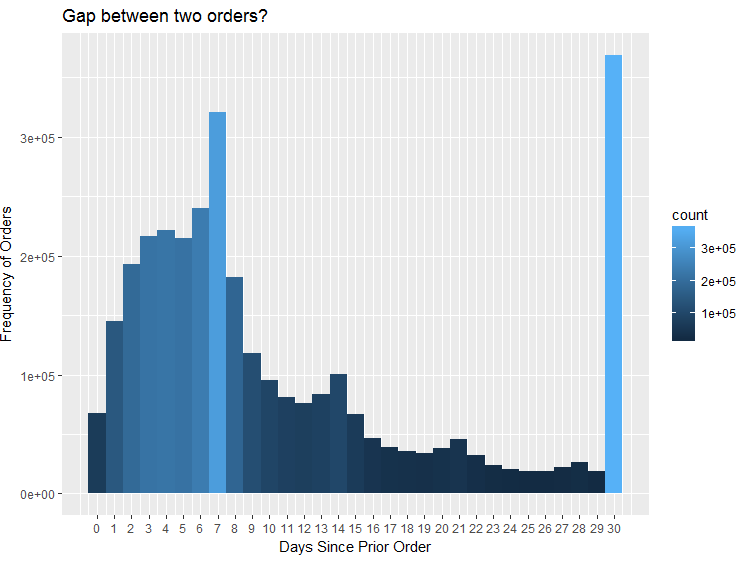
From the figure below it is deduced that Sunday and Monday are the days when people order most on Instacart.



It is found that 10:00 AM on Monday is the time of the day when most orders are placed. This can be thought of when people go to work they refill their groceries for the rest of the week. To visualize this percentage of orders made every hour for every day is plotted and compared. The figure can be referred in Appendix A for further details.

## Reordering Gap

We are given the gap between two orders for every user. When we plot it we find two categories of people? One that reorders monthly other who does weekly. This is based on the peaks formed at 30th day and 7th day.



# References

“The Instacart Online Grocery Shopping Dataset 2017”, Accessed from [https://www.instacart.com/datasets/grocery-shopping-2017 on June 2017](https://www.instacart.com/datasets/grocery-shopping-2017%20on%20June%202017)”

“Data Dictionary: <https://gist.github.com/jeremystan/c3b39d947d9b88b3ccff3147dbcf6c6b>”

“<https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2>”

“ <https://www.kaggle.com/philippsp/first-exploratory-analysis> “

“ <http://zoonek2.free.fr/UNIX/48_R/03.html> “

# Appendix A

