

**Research Project**

**On**

**Instacart Market Basket Analysis**

**By**

**Mohita Sharma**

**Nikhil Kamath**

**Labdhi Ghelani**

**Pranav Gudipally**

**Team name: Data Rangers**

[ALY6110 21068](https://northeastern.blackboard.com/webapps/blackboard/execute/launcher?type=Course&id=_2567936_1&url=) Data Management and Big Data

Winter 2019

Submitted on: **February 15th**, **2019**

Instructor**: Prof. Jamie Warner**

Introduction

In 2014, Market Basket which is a grocery chain in US combined with Instacart. Instacart operates in six largest cities nationwide, began working with Market Basket to deliver its food online. This app aims to make life easy, after selecting products through Instacart, shoppers can review the orders and do the in-store shopping which gets delivered at home. Instacart Data science team plays a big role in providing wonderful experience. The use of big data along with certain algorithms helps to develop model that predict which products the user will buy again, try for the first time or add to their cart next time. With ever increasing cost of acquiring new customers and increasing competition, leveraging existing customer is the best option at the disposal of businesses. The best solution for this turns out to be analyzing using big data techniques. This proposal includes methods of data collection, storing, cleaning, validating and analyzing various trends and patterns of Customer behavior.

Aim

The main aim is to build a recommendation engine which would recommend products to a user based on their previous purchase which shall help maximize the sale of the products. By using just, the analysis mentioned below, we would be able to help and recommend our customers that which product they would want to buy with the product already present in their cart, the business questions on which we shall do the analysis that lays a foundation in building the recommendation engine are as follows:

* Top 10 Highest Products sold
* What time of the day do people order the highest selling products?
* Top 5 users departmental purchase
* Reordering frequency distribution by days since prior order
* Reorder ratio for day of the week vs hour of the day
* On which day do people order the highest selling products?

Data Collection

We have collected data from Kaggle Data Sets which contains a sample over 3 million grocery orders from more than 200000 Instacart users structured in a relational set of files describing customer’s orders over time. It contains data such as: which time of the day the product was purchased, a relative measure between time and order of product and many more. Also, each entity (customer, products, order, and aisle) has an association unique id. This dataset includes orders from many different retailers and is a heavily biased subset of Instacart’s production data. The values and all the ID’s in the dataset are entirely randomized and they cannot be linked back to any other ID. Also, the information provided about the users is sequence of orders and product in that order. Retail ID is not provided, only products that are brought by multiple people at various retailers are included. Many interesting patters can be found using these datasets. The description of each file is as below:

a. aisles.csv: contains aisle\_id, aisle

b. department.csv: contains department\_id, department

c. orders\_product\_train.csv: contains order\_id, product\_id, add\_to\_cart\_order, reordered

d. orders\_product\_prior.csv: contains order\_id, product\_id, add\_to\_cart\_order, reordered

e. orders.csv: contains order\_id, user\_id, eval\_set, order\_number, order\_dow, order\_hour\_of\_day, order\_since\_prior\_order

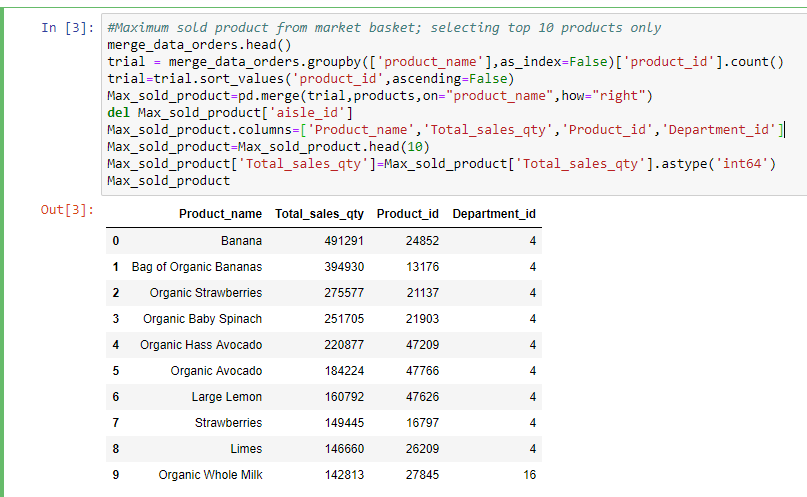
f. products.csv: contains product\_id, product\_name, aisle\_id, department\_id

g. sample\_submission.csv: contains order\_id, products

Analysis

**Maximum sold product from market basket (Top 10)**

Here we have done the analysis for the top 10 products (using the head(10) function) sold in the market basket, according to our outputs that products is “bananas” with a product id “24852” which were sold at a quantity of 491291 by consumers, stands out as the top product in terms of highest sales

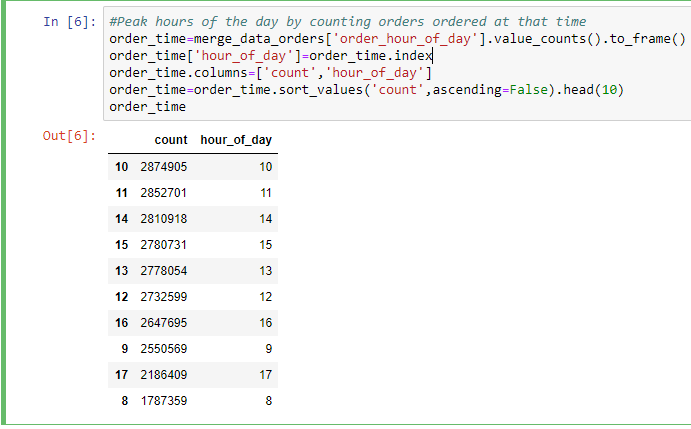


A screenshot of a cell phone

Description generated with very high confidence

**What time of the day do people order the highest selling products?**

We have calculated the peak hour by counting the number of orders ordered at that time and we can see that the highest orders are being done at 10AM which is around 2874905 orders which is when people usually state their meal and day



A close up of a map

Description automatically generated

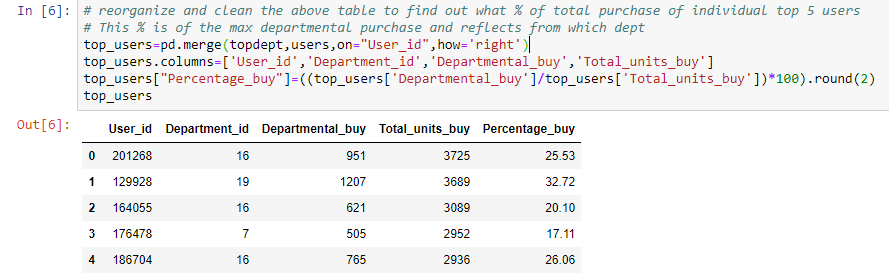
**For top 5 users, their departmental purchase**

This analysis is for the top 5 users and their consumer behavior, meaning, the product they purchase the most and what quantity and from which department (Department\_id, Departmental\_buy) of total units bought across all departments (Total\_units\_buy) which here is for the user having separate user id 201268 purchasing a particular product having an id 951 from the department having id 16



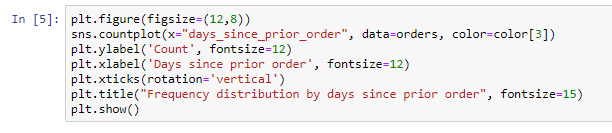
In terms of percentage

This particular analysis shows the maximum departmental purchase reflects from which department (all departments have different id) for top 5 users individually



**Reordering frequency distribution by days since prior order**

This analysis shows us the usual reordering frequency of consumers who order through instacart, looking at the output visualization we can deduce that customers order usually once a week or once a month (peaks at 7th and 30th day of a month) which shows weekly intervals

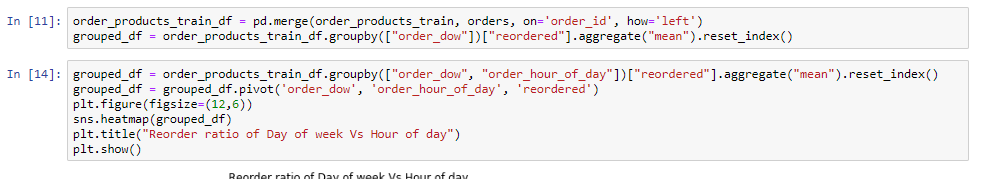


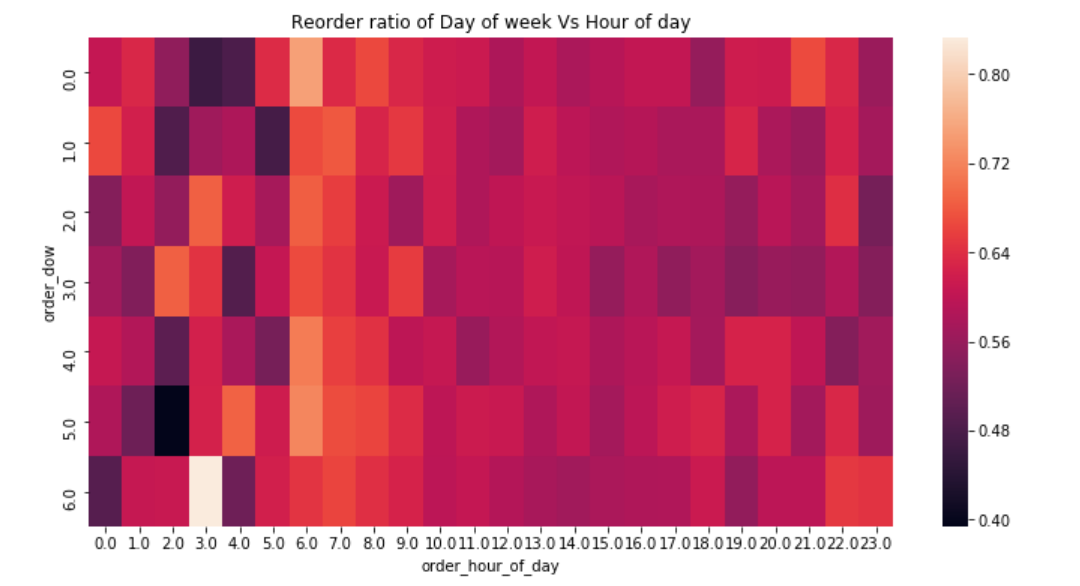
**A screenshot of a cell phone

Description generated with high confidence**

**Reorder ratio for day of the week Vs hour of the day**

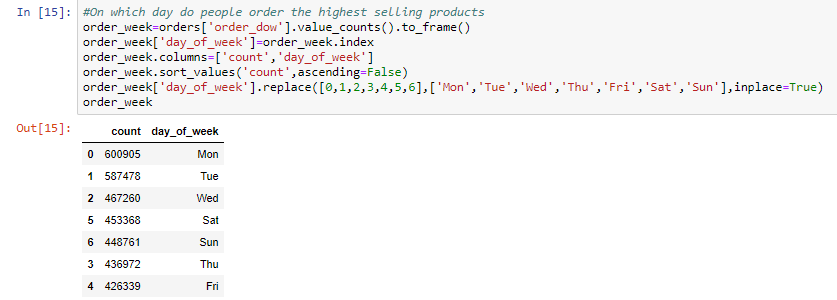
This analysis shows the reorder ratio between two sets of data (order\_dow and order\_hour\_of\_day) whereas the heat map used for visualization draws meaning out of the data where it shows that the reorder ratio were quite high during early mornings as compared to the later half of the day



****

**On which day do people order the highest selling products?**

This analysis shows on what day of the week are the products being sold the highest, clearly Monday is when there is highest sale in the products which means that this analysis shall help the retailers restock their stock by Sunday night for better sales and growth



A screenshot of a cell phone

Description generated with high confidence

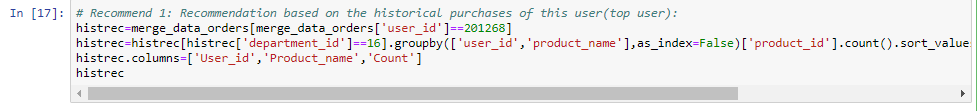
Recommendation Engine

Given that a user has placed products P1, . . . , Pn−1 in their shopping cart, what are the next best k products Pn, . . . , Pn+k−1 to advertise. The recommendation engine would help us to find the combo products that a user might like to purchase i.e determining what products might sell well together. Using this technique, Instacart can utilize target marketing during peak times.

By using the analyses made above, we would be able to help and recommend our customers that which product they should buy with the product already present in their cart. For example, users buying product like Baby diapers from the baby department would also like to buy baby shampoo or other products for a baby. So, a recommendation system predicts what a user is likely to buy based on what they have brought earlier.

**Recommendation 1**

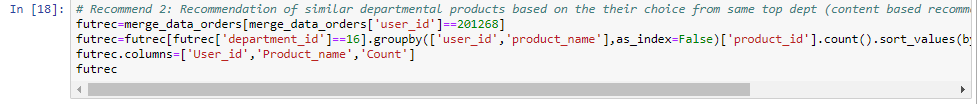
Here we provide a recommendation to our top user (user ID 201268) with the product that person has ordered the most (56 times) that is “Feta Cheese Crumbles” which this user is likely to order again

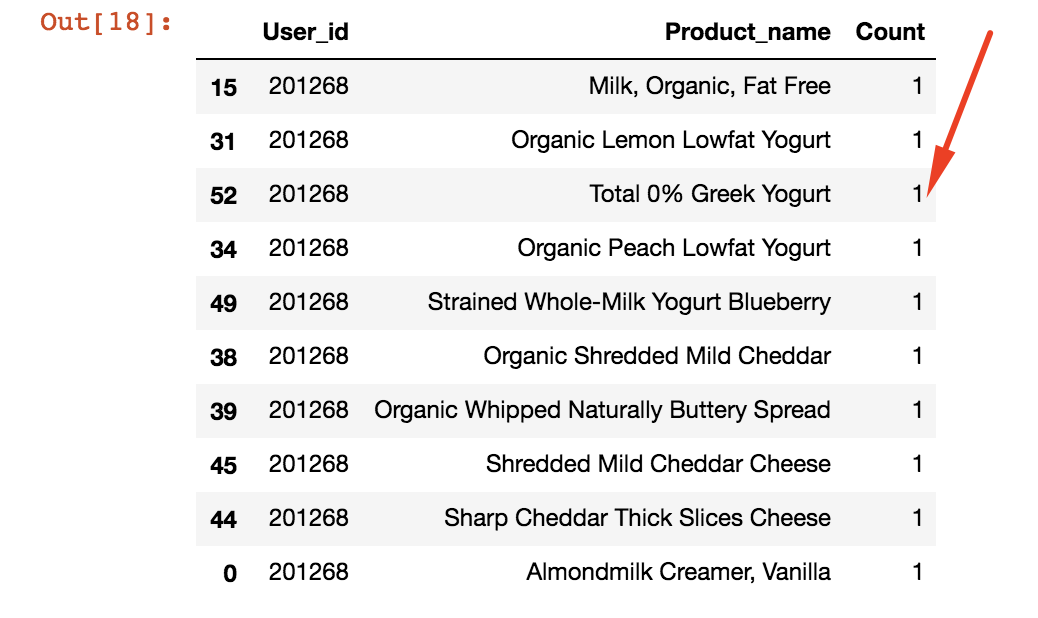
****

****

**Recommendation 2**

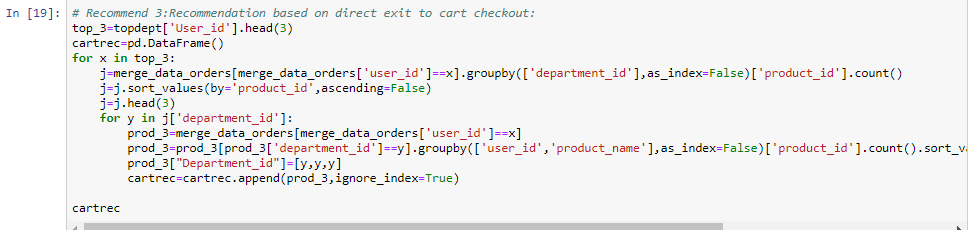
Here the personalized recommendation of products to the top user is considered, here is based on their choice from the top department

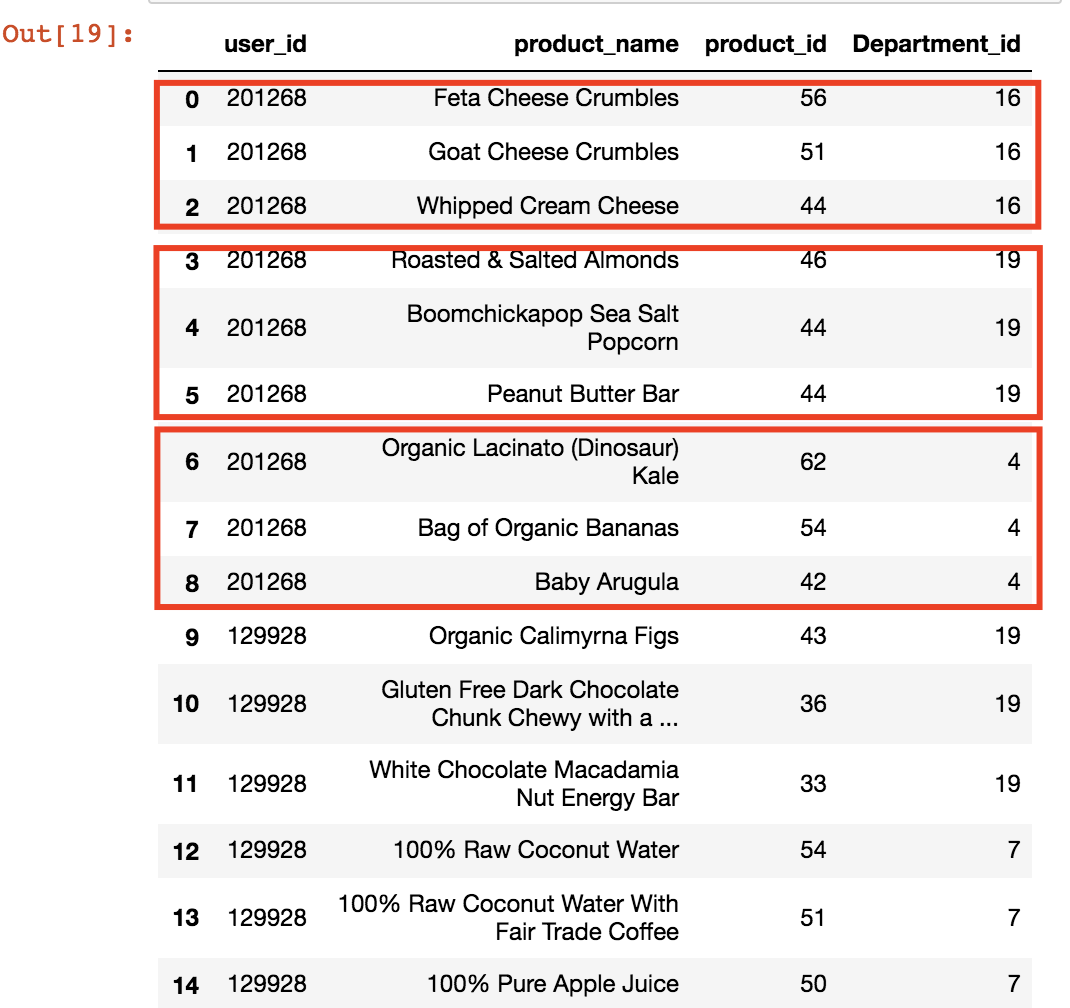


****

**Recommendation 3**

This recommendation is based on direct exit to cart checkout option which has been designed for the top user based on their purchase history



****

**Conclusion**

Effective recommendations are a valuable service to the customers and a profitable service to the retailer, so to help the customer have a hassle free efficient shopping experience and for retailers to increase their sales and manage their existing customers, we have decided to build a recommendation engine that can recommend items to a user which are most popular among users of same type, also we will divide users to multiple segments based on their preferences and recommend items based on where they belong to. Based on above done analysis of the data set we have developed the recommendation engine which will not only help the consumers for easy checkout but also by predicting this user-based purchase, retailers could run an inventory check stocking the most used department to retain their consumers which would act as a source for their commercial and organizational benefits.

**References**

* Alan Flores-Lopez, Skip Perry, Poorvi Bhargava 2018- Recommendations in online grocery shopping <http://cs229.stanford.edu/proj2017/final-reports/5240337.pdf>
* Ryan Aminollahi March 4th 2018 Use algorithms to recommend items to customers in python <https://towardsdatascience.com/use-algorithms-to-recommend-items-to-customers-in-python-347b769b21f3>
* Moorissa Tjokro Oct 15th 2018 How to build a recommendation system for purchase data <https://medium.com/datadriveninvestor/how-to-build-a-recommendation-system-for-purchase-data-step-by-step-d6d7a78800b6>
* Vladimik lazovskiy March 19th 2018 What’s in your customers shopping cart <https://towardsdatascience.com/whats-in-your-customer-s-next-shopping-cart-73d64287ec53>
* (n.d.). Retrieved from <https://www.kaggle.com/c/instacart-market-basket-analysis/data>