Capstone – Business Analysis and Recommender System(FMCG)

Ling Chong Gold

Agenda

- Business and Data Science Problem
- Background of Data
- Data Analysis
- Customer Segmentation
- Association Rules (Basket Analysis)
- Recommender System
- Next Step and Further Improvements

Business Problem

Analyze and understand our customers, recommend actions to boost sales

Problem Statement – Chain of Thoughts

- Gaining traction on data privacy
  - Challenge to collect customer demographic information
- Explicit feedback difficult to obtain and tends to be bias
  - Circumstances leading to how they are gathered
  - No governing standard
- According to a study by McKinsey, 75% of what consumers watch on Netflix comes from the company's recommender system
- Amazon credit 35% of their revenue to their recommender system
- Unawareness causes absence of interaction between customer and product

Source

Problem Statement – Data Analysis / Data Science

- Perform Analysis on product and customer purchase behavior to gather insights
- Perform basket analysis
- Create a recommender system with the absence of explicit feedback and customer demographic profile

Background of Data

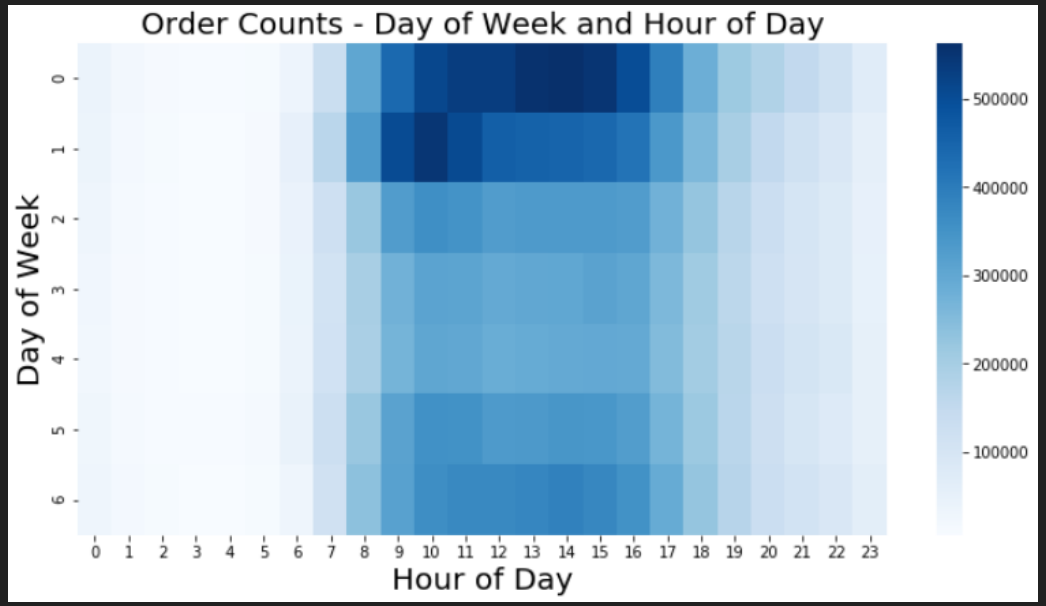
- Data was gathered from Kaggle

- Data was split into multiple csv files with corresponding primary and foreign keys similar to a relational database

- Combined dataset about 32M rows and 15 columns
  ~ 3.4M unique orders
  ~ 200K unique users
  ~ 49K unique products

- Validation set consisting of the last orders or customers which our model has not seen was prepared for evaluation of our recommender system

Data Analysis – Purchase Behaviour

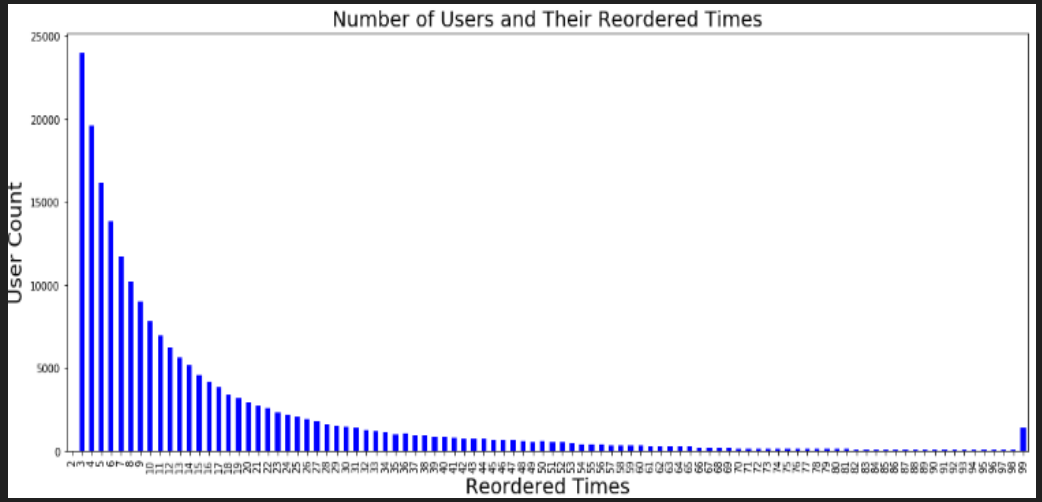


- Number of orders for day of the week and hour of the day
- Sunday (Day 0) between 9AM – 5PM
- Monday (Day 1) between 9AM – 11AM
- Marketing campaign or flash deals can be strategically scheduled for maximum outreach
- System Maintenance can be performed between 1AM - 5AM

Data Analysis – Purchase Behaviour

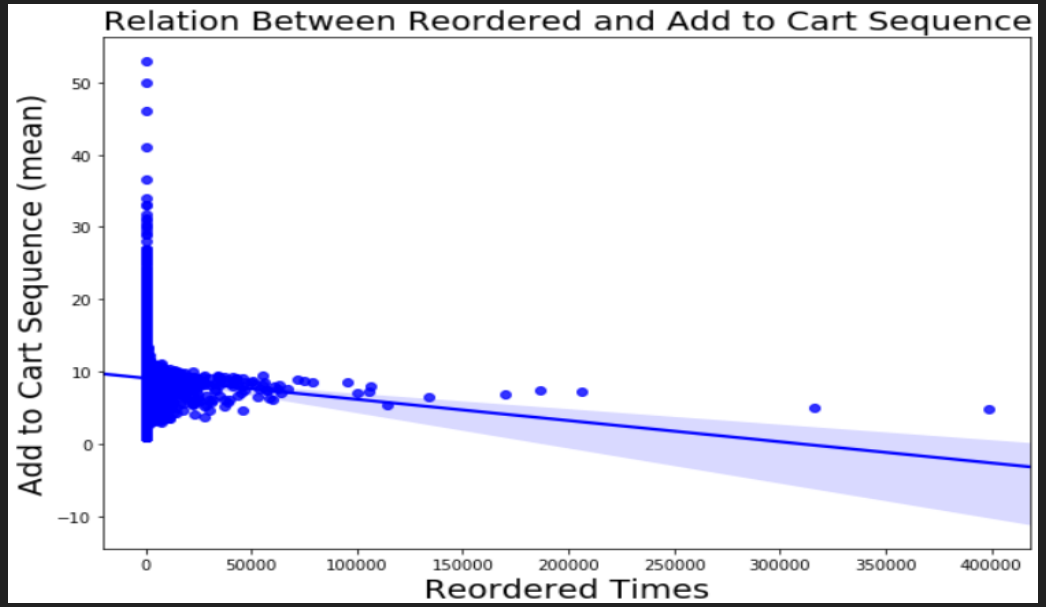
- Graph shows the number of orders and lapse in days since their last order
- Peak every 7 days
- Highest peak at 7 and 30 days after last order
- Most repeated orders comes in within the first 7 days
- Strategic push notifications personalized to customers and be implemented

Data Analysis – Purchase Behaviour



- Graph shows the number of customers and their number of reorders
- A group of loyal customers who have made 99 reorders
- Number of customers decreases as reordered times increase
- Explore on customer churn and customer retention success

Data Analysis – Purchase Behaviour

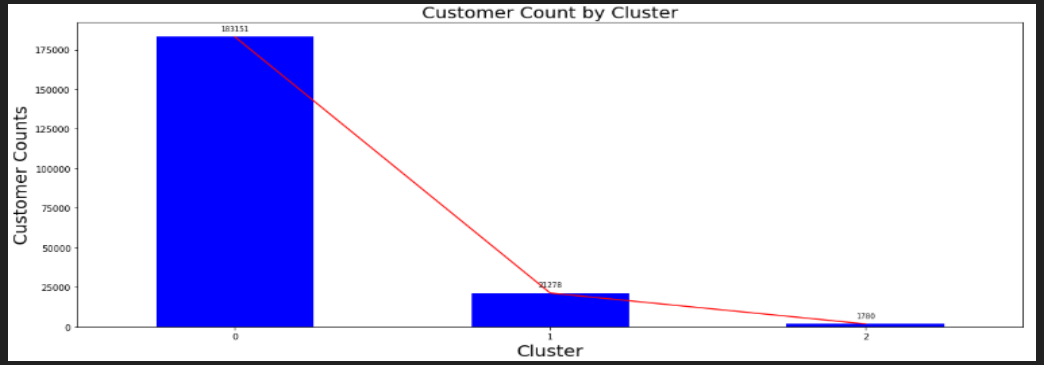


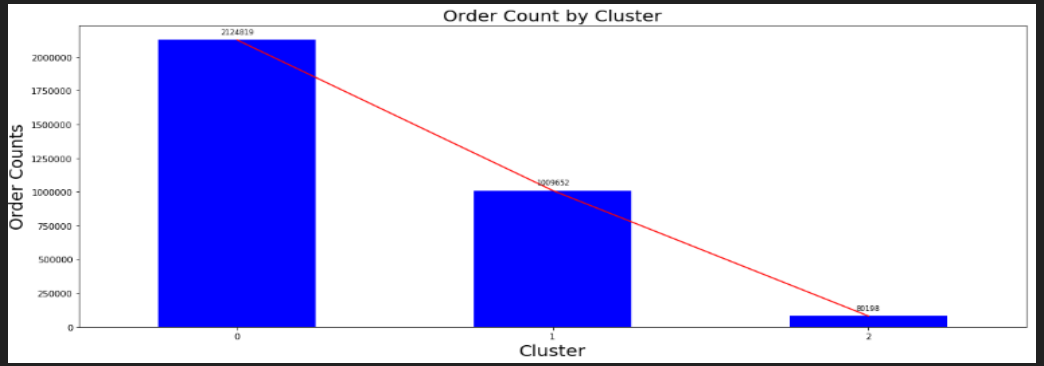
- Relation between the sequence the product is added to cart and their reordered times
- First 10 products added to cart are more likely to be reordered
- People tend to add items which they know they will buy first
- Items added later part of the cart could be to qualify for perks

Customer Segmentation - Explanation

- Relevant marketing campaigns can be targeted at individual clusters for effectiveness and cost saving
- Customers were segmented based on their purchase behavior on the sub-category of products through unsupervised learning
- PCA was utilized for feature extraction
- Clustering technique K-means was chosen
- Elbow method utilized to determined that the best number of clusters for our dataset
- EDA was then performed to gather insights for each cluster

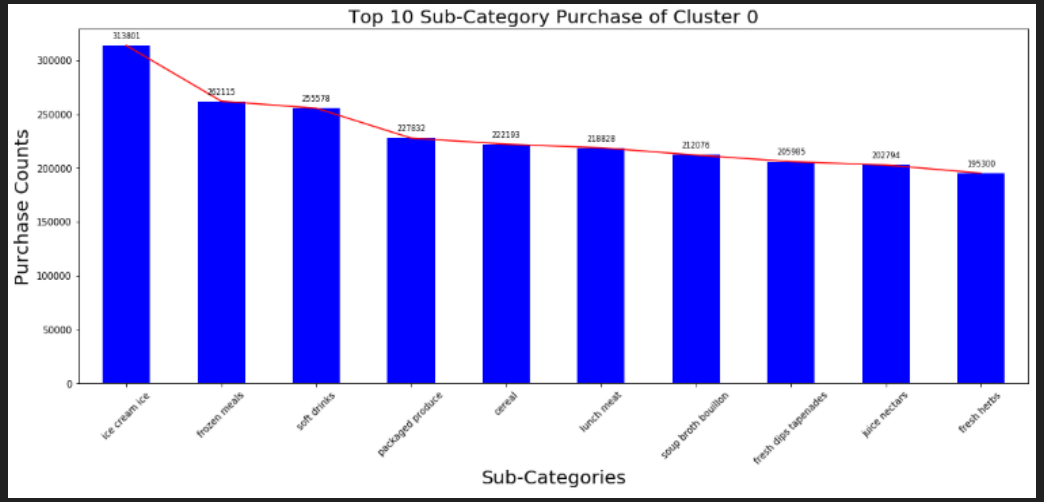
Customer Segmentation - Overview





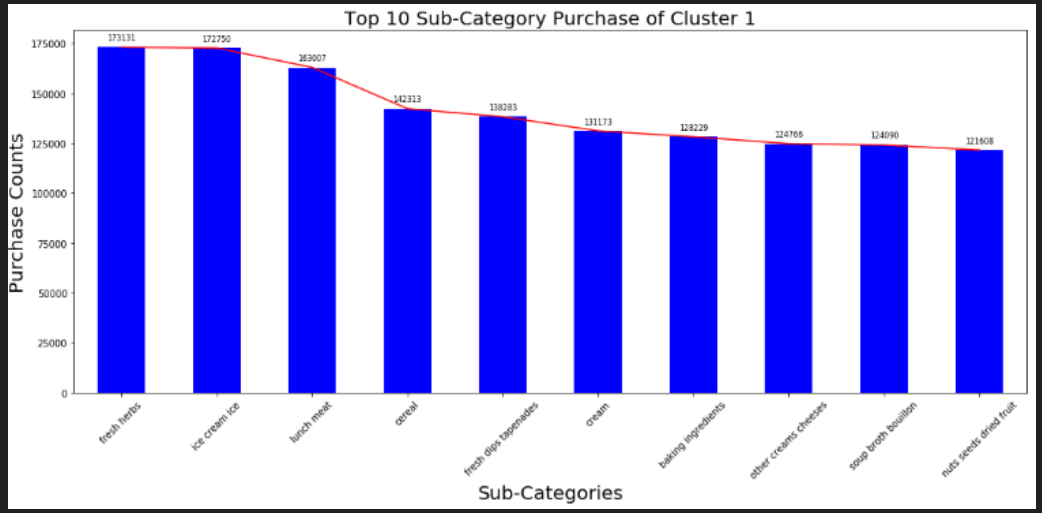
- Plot shows the number of customers and number of orders for each cluster.
  - Cluster 0 – 11.5 orders per customer
  - Cluster 1 – 47.6 orders per customer
  - Cluster 2 – 45 orders per customer
- Cluster 0 could contain most of the one time off customers, we can try to “move” repeated ones to other cluster
- Marketing strategies personalized to the clusters will improve customer experience and relevancy

Customer Segmentation – Cluster 0 (Convenience)



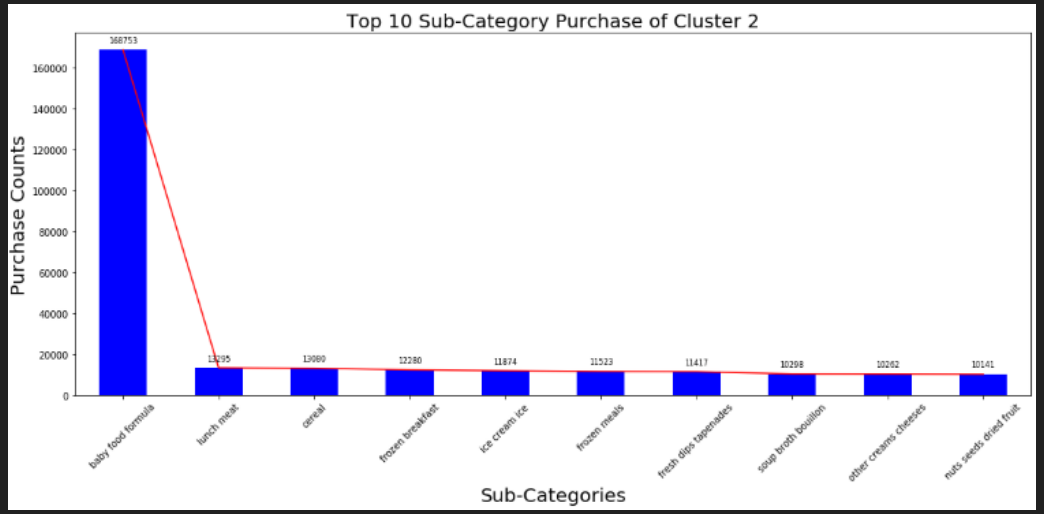
- Frozen meals, soft drink and packaged produces are the top sub-categories
- Promotional items related to convenience can be targeted at this cluster
- Easy to cook recipes with fresh products from our store made available in our site targeting at this cluster

Customer Segmentation – Cluster 1 (Fresh)



- Absence of ready-to-eat meals in this cluster's top purchase
- Fresh items are amongst the top purchases
- Membership perks like fixed routine delivery can be targeted at this cluster

Customer Segmentation – Cluster 2 (Babies)



- High purchase count of baby food formula
- Baby related promotions or products can be targeted at this cluster
- Provide toddler care tips with products from our stores to this cluster

Association Rules - Explanation

- Apriori library utilized to capture patterns of items association

- Strength of association
  - Support – Number of times the item appear out of the total transactions
  - Confidence – Likelihood that item B is bought when A is bought
  - Lift – The increase in ratio of the sale of B when A is sold

- Use cases
  - Recommender System – Recommend items upon adding to cart
  - New product – Product Y in flavors of X
  - Health Care – Diseases and their associated symptoms
  - Fraud Detection – Fraudulent transaction and their associated behaviors
  - Education – Students facing difficulties in a topic will likely face the same in another

Association Rules - Evaluation

	Antecedent	Consequents	Support	Confidence	Lift
253	Bag of Organic Bananas	Organic Strawberries	0.00180	0.23579	2.73546
140	Organic Strawberries	Organic Reduced Fat Milk	0.00232	0.20025	2.32326
150	Organic Strawberries	Whipped Cream Cheese	0.00209	0.20575	2.38702
412	Organic Whole Milk	Organic Strawberries	0.00135	0.27914	3.23843
397	Organic Lemon	Organic Hass Avocado	0.00103	0.29824	4.28357

We can consider introducing Organic Reduced Fat Milk in Strawberry flavor to the store

- Organic Strawberries appeared in 0.23% of our total transaction
- People are 20% more likely to buy Organic Reduced Fat Milk when they purchase Organic Strawberries
- People are 2 times more likely to buy Organic Reduced Fat Milk and Organic Strawberries compared to just buying Organic Strawberries

Recommender System - Explanation

- Leverage on implicit feedback gathered through customer’s purchase behavior using the Implicit library
- Interaction between customer and item is the basis of how our recommender system works
- An absence of interaction could mean that the customer do not like the item or the customer do not know about the item
- A good recommender system is able to identify features a user like based on their past behavior and behavior of similar users then match them with non-interacted products with these features.

Recommender System - Evaluation

```
Recommended items for User 1 are:

13517 Whole Wheat Bread 1.1968588
20063 Hazelnuts in Milk Chocolate, 33% Cocoa 1.1774435
26853 Complete Wheat 100% Whole Wheat Bread 1.1398611
15487 Raspberry English Tea Scones 1.1145719
============================================================
User 1 validation transactions are:

                     product_name  user_id
                             Soda        1
            Organic String Cheese        1
          0% Greek Strained Yogurt        1
 XL Pick-A-Size Paper Towel Rolls        1
           Milk Chocolate Almonds        1
                       Pistachios        1
            Cinnamon Toast Crunch        1
       Aged White Cheddar Popcorn        1
              Organic Whole Milk        1
            Organic Half & Half         1
              Zero Calorie Cola         1
```

Relevant Recommendations:

- Recommended Hazelnut in Milk Chocolate
- Purchased Milk Chocolate Almonds

Room for Improvement:

- Recommended bread twice

Recommender System - Evaluation

```
Recommended items for User 3754 are:

33502 Double Cheese Baked Snack Mix 1.1659267
45339 Men's Refresh Dandruff Shampoo 1.060647
29642 Ultra Soft Bath Tissue 1.0497061
13810 Reclosable Gallon Freezer Bags 1.0313741
============================================================
User 3754 validation transactions are:

                                    product_name  user_id
                             Twice Baked Potatoes     3754
                           Whipped Sweet Potatoes     3754
 100% Natural Skin & Hair Revitalizing Coconut Oil     3754
```

Relevant Recommendations:

- Shampoo
- Purchased Skin and Hair product

Recommender System - Evaluation

```
Recommended items for User 58144 are:

34172 Top Ramen Shrimp Flavor Instant Noodle Soup 1.1561155
39322 Caramel Almond and Sea Salt Nut Bar 1.1357532
35175 Mini Stuffers Hamburger Dill Chips 1.1336474
19604 Medium Scarlet Raspberries 1.0747831
============================================================
User 58144 validation transactions are:

                                  product_name  user_id
                       Electrolyte Enhanced Water    58144
                                           Banana    58144
 Air Chilled Organic Boneless Skinless Chicken ...    58144
                             Lime Sparkling Water    58144
                      Non Fat Raspberry Yogurt       58144
                                 Farfalle No. 93     58144
                Total 0% Nonfat Plain Greek Yogurt    58144
                           Original Orange Juice      58144
                  Best Sloppy Joe Skillet Sauce       58144
                    Organic Cauliflower Florets       58144
                    Grated Parmigiano Reggiano        58144
                                   Whole Almonds      58144
```

Relevant Recommendations:

- Recommended Raspberries
- Purchased Banana
- Purchased Raspberry Yoghurt

Recommender System - Evaluation

```
Recommended items for User 114401 are:

44898 Organic Mac And Trees Fun Shape Macaroni & Cheese 1.1370347
35488 Organic Dry Roasted Premium Flaxseed 1.1338583
2190 Spicy Red Lentil Sauce 1.1187676
21702 Puna Coconut Pineapple 1.1060191
============================================================
User 114401 validation transactions are:

                                product_name  user_id
                                  Whole Milk   114401
No Pulp Calcium & Vitamin D Pure Orange Juice   114401
                  Original Fresh Stack Crackers   114401
                          Cheddar Broccoli Rice   114401
                               Corn Pops Cereal   114401
                         Eggo Strawberry Waffles   114401
       Original 100% Pure No Pulp Orange Juice   114401
                             Orange Juice To-Go   114401
                  All Natural Peach Tea Bottles   114401
                           Hickory Smoked Bacon   114401
```

Relevant Recommendations:

- Recommended Puna Coconut Pineapple
- Purchased Juices

Recommender System - Evaluation

```
Recommended items for User 200372 are:

30890 MCT Oil 1.26377
8651 Shipping Packaging Tape Heavy Duty 1.2522316
17419 Sprouted Whole Wheat Bread 1.209813
17018 Ghee Vanilla Bean 1.1393975
============================================================
User 200372 validation transactions are:

                product_name  user_id
                   Diet Cola   200372
       Original Potato Chips   200372
  Salsa Con Queso Medium Dip   200372
        Pure Sport Body Wash   200372
          Snickers Ice Cream   200372
 Raspberry Cheesecake Gelato   200372
                   Rosemary   200372
                Red Potatoes   200372
   2% Low Fat Cottage Cheese   200372
```

Relevant Recommendations:

- None

Reasons for Recommendation:

- Recommended Ghee Vanilla Bean
- Previous orders includes a variety of bread spread

- Recommended MCT Oil
- Previous orders includes supplements and energy drinks

Next Steps and Further Improvements

- Evaluate the effectiveness of the recommender system after deployment and perform the next iteration of improvement

- Add features to products during the next iteration.

  e.g. organic, natural, convenient, fresh, price range, manufacturer etc

- Using neural network Sequential to predict customer's next purchase and perform through association rules

- Additional customers insights can be obtained with more data
  - Recency
  - Frequency
  - Monetary

Q & A

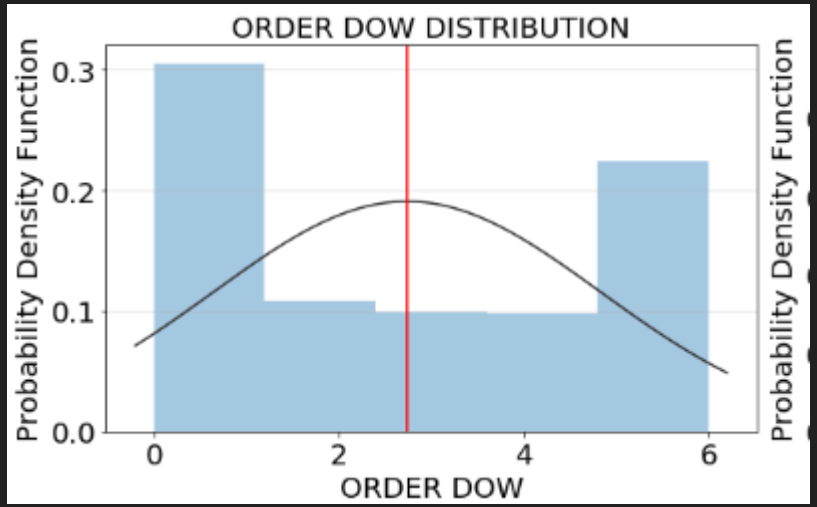
Thank You

Divya, Ryan and Zi Liang
for mothering us the last 3 months

“Python Does it For You” - Divya

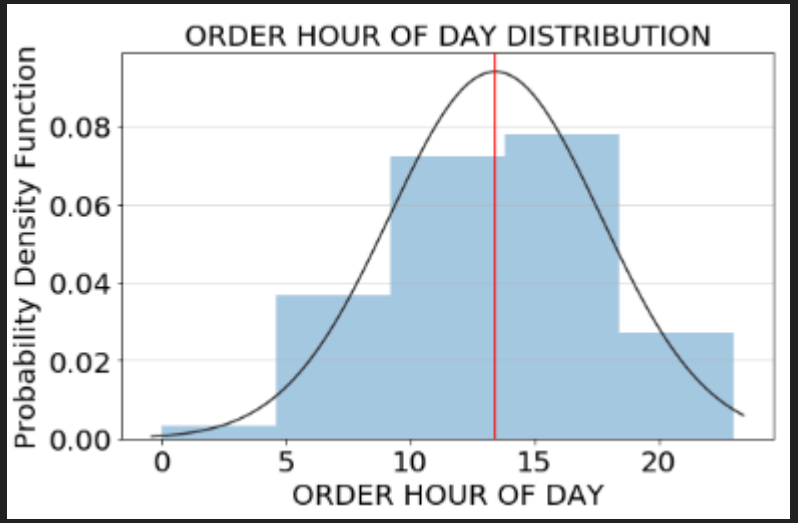
Backup Slides

Backup Slides



- Distribution of orders for Day of the Week
- Most orders are placed on Sundays(Day 0) and Mondays(Day 1)

Backup Slides

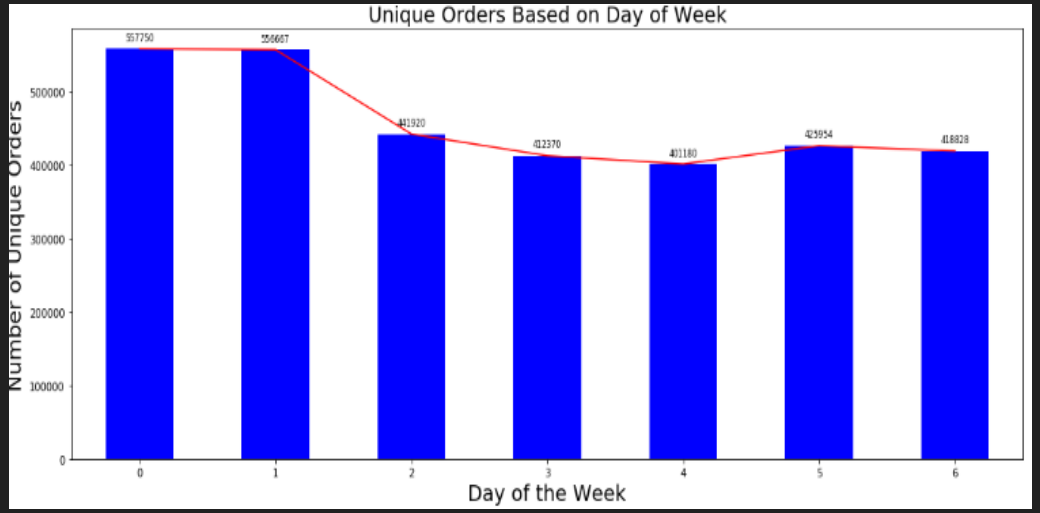


- Distribution of orders for hour of day
- Most orders are placed between 9AM and 5PM

Backup Slides

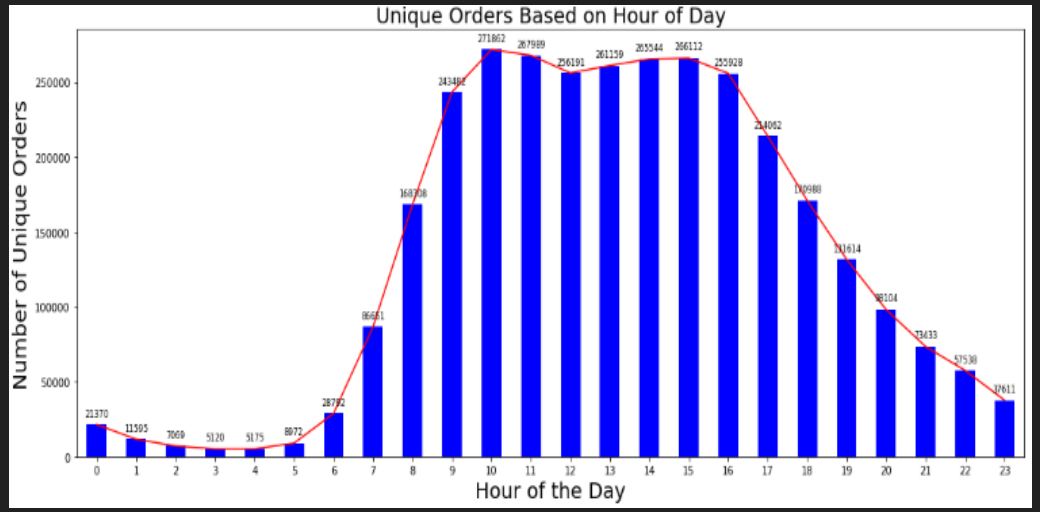
- Distribution of orders for days since prior order
- Most repeated orders are placed within 12 days of their last order

Backup Slides



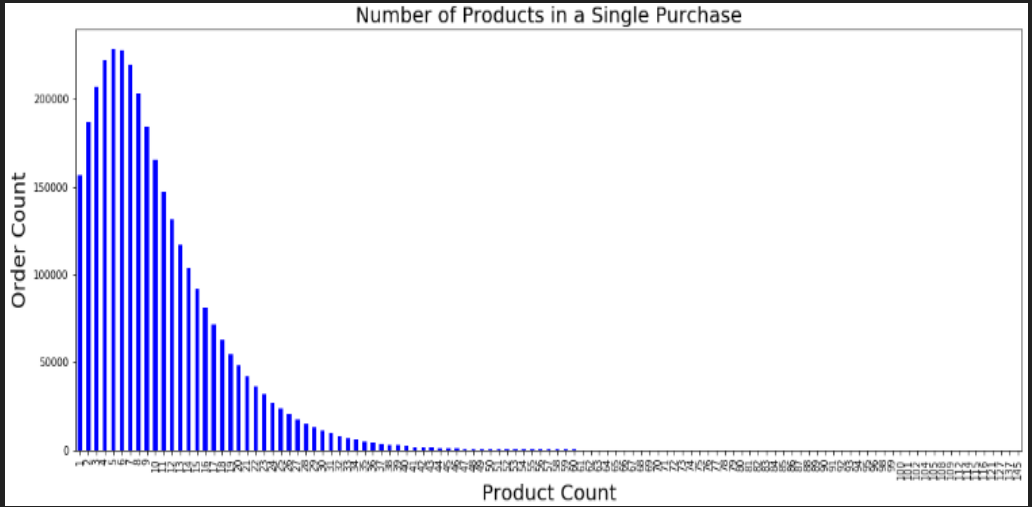
- Unique Order Count Based on Day of the Week
- Sundays(Day 0) and Mondays(Day 1) have most unique orders placed

Backup Slides



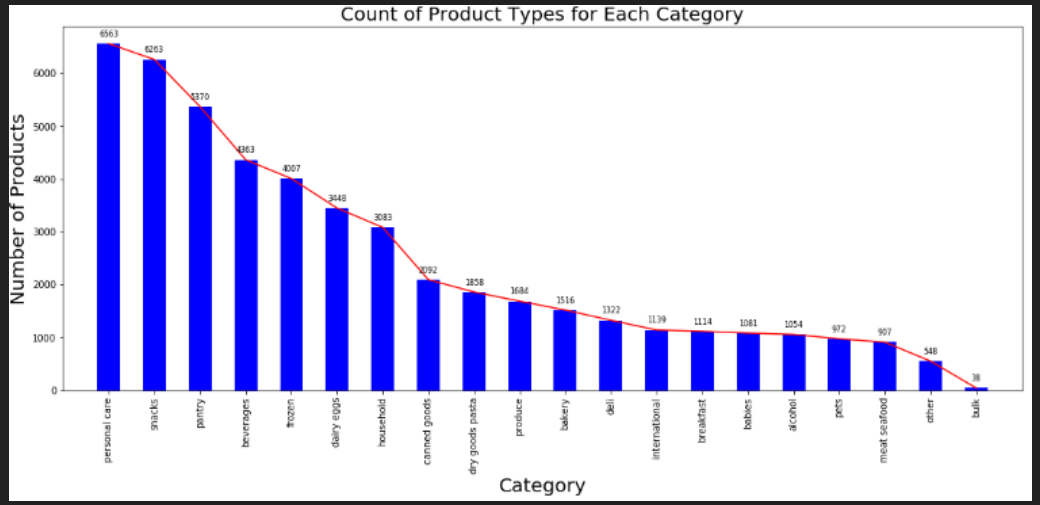
- Unique Order Count Based on Hour of the day
- Most orders are placed between 9AM – 5PM

Backup Slides



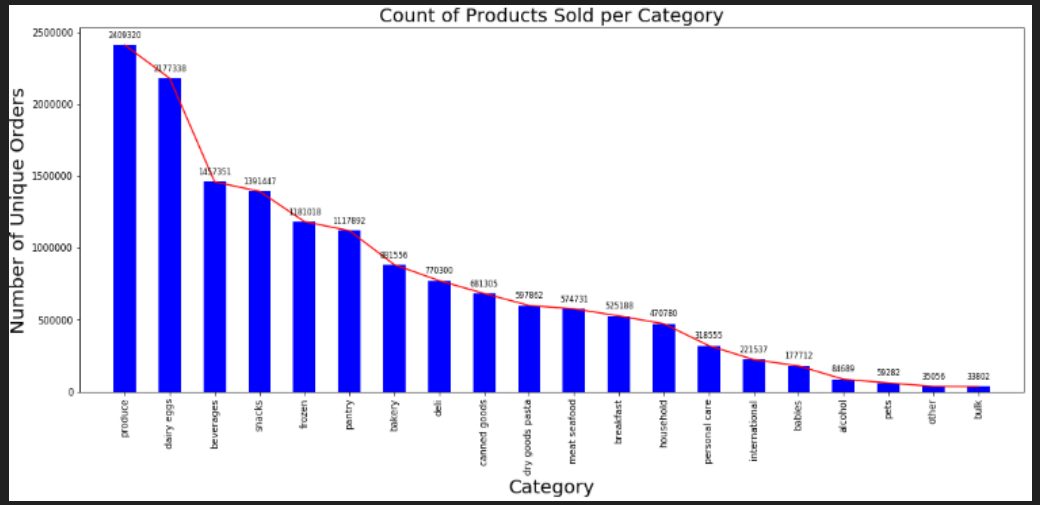
- Number of products in a single purchase
- Most orders contains 3-10 products

Backup Slides



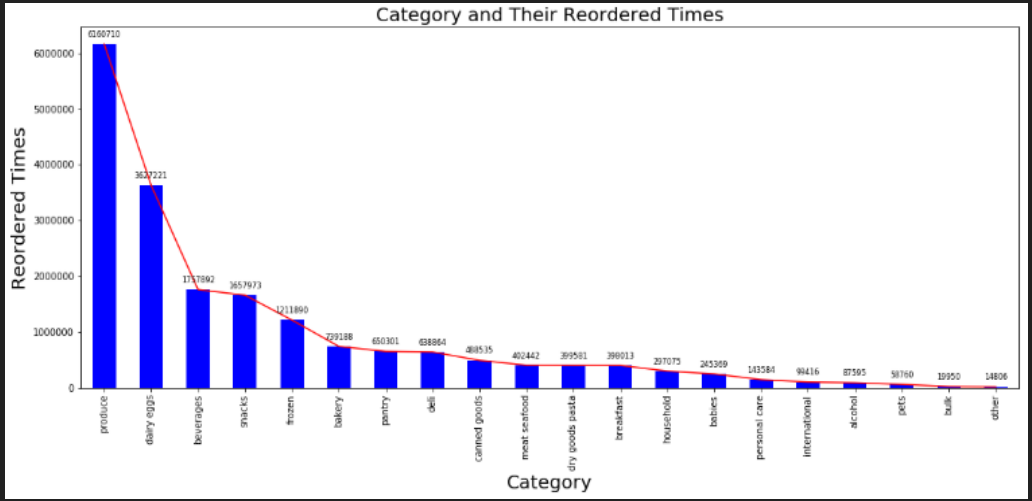
- Count of Products for Each Category

Backup Slides



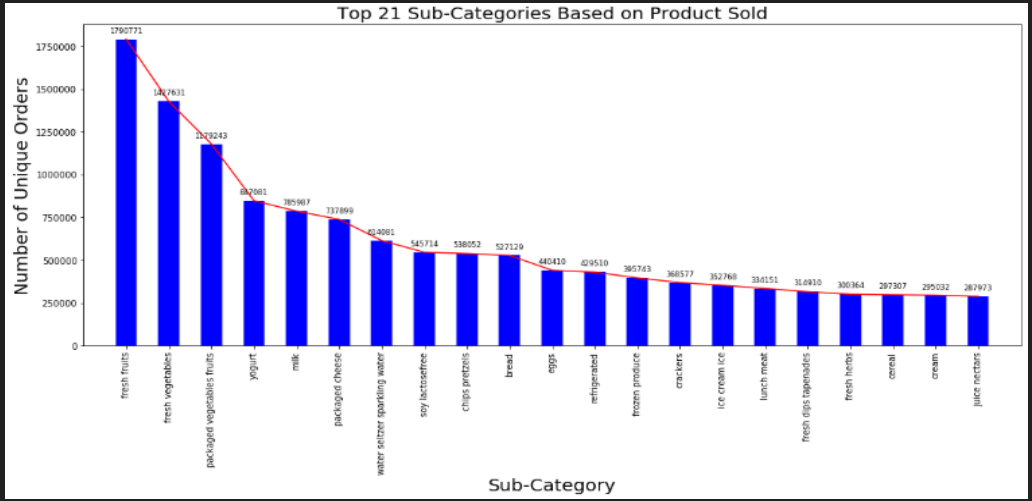
- Count of Products sold for Each Category

Backup Slides



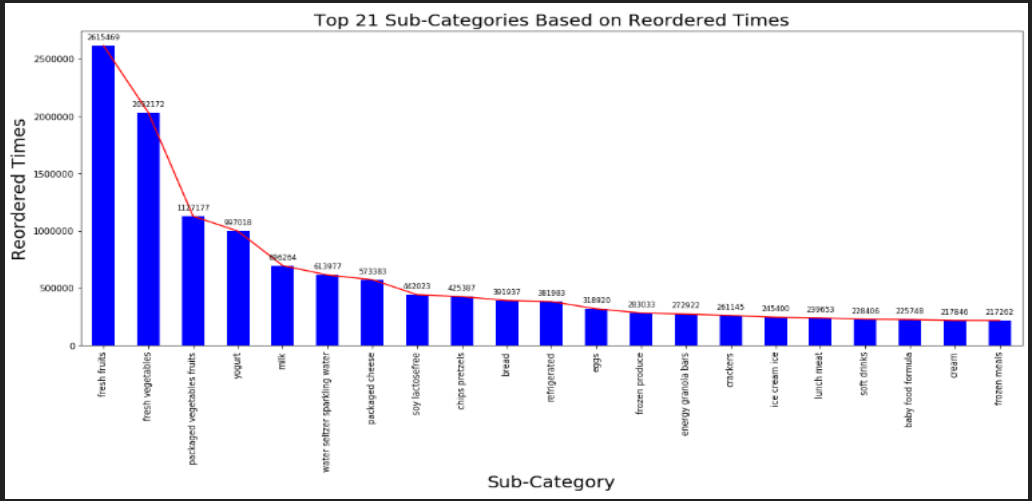
- Category and their reordered times

Backup Slides



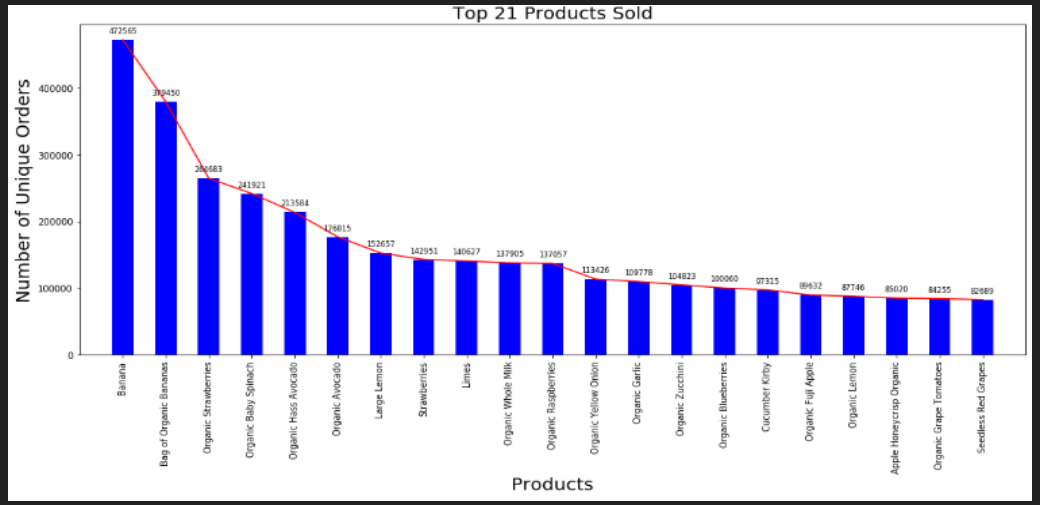
- Top 21 Sub-Category Sold

Backup Slides



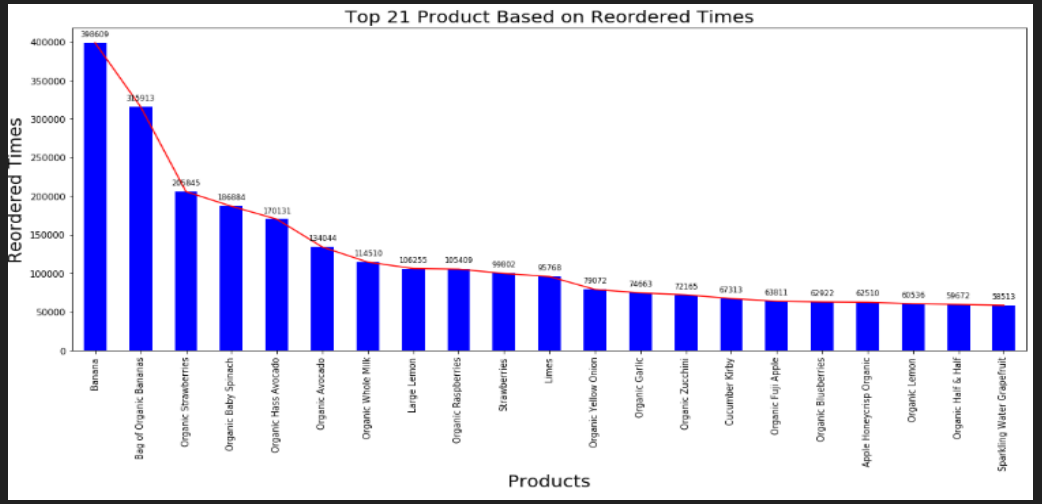
- Top 21 Sub-Category reordered

Backup Slides



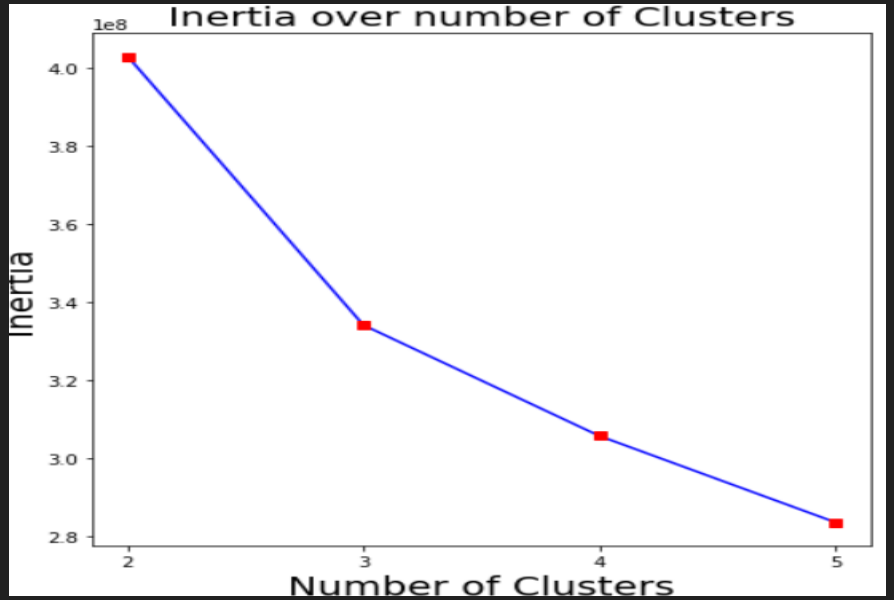
- Top 21 Products Sold

Backup Slides



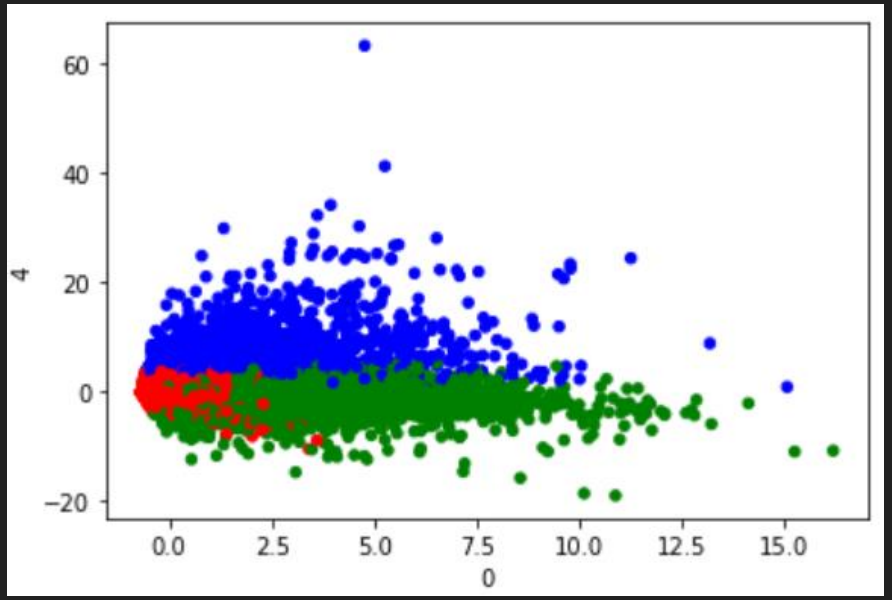
- Top 21 Products Reordered

Backup Slides



- Elbow method to determine number of clusters

Backup Slides



- Visualisation on clusters (Customer Segmentation)