Market Basket Analysis

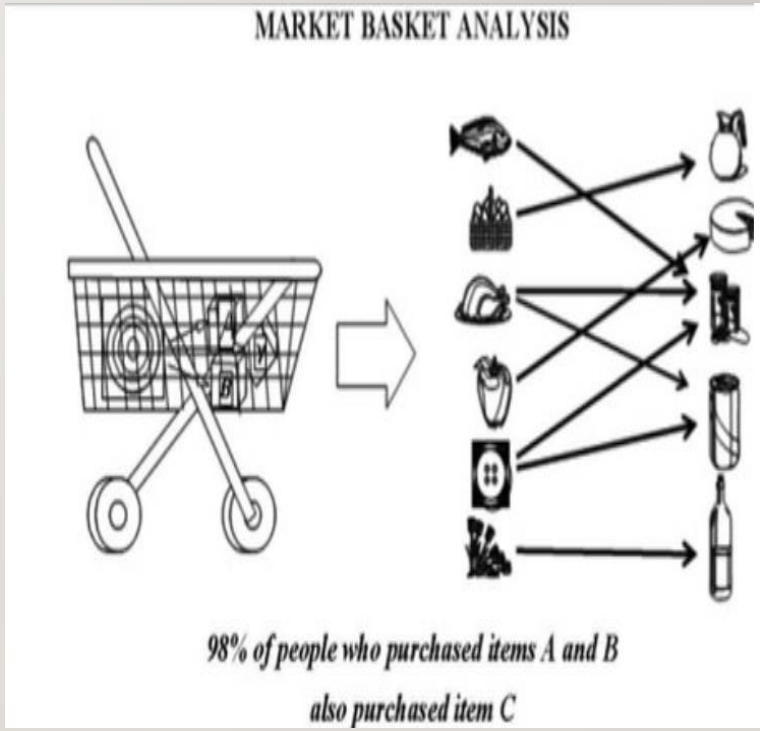
MARKET BASKET ANALYSIS USING APRIORI ALGORITHM

CHATHURA BRAHAKMANAGE

MARKET BASKET ANALYSIS USING APRIORI ALGORITHM

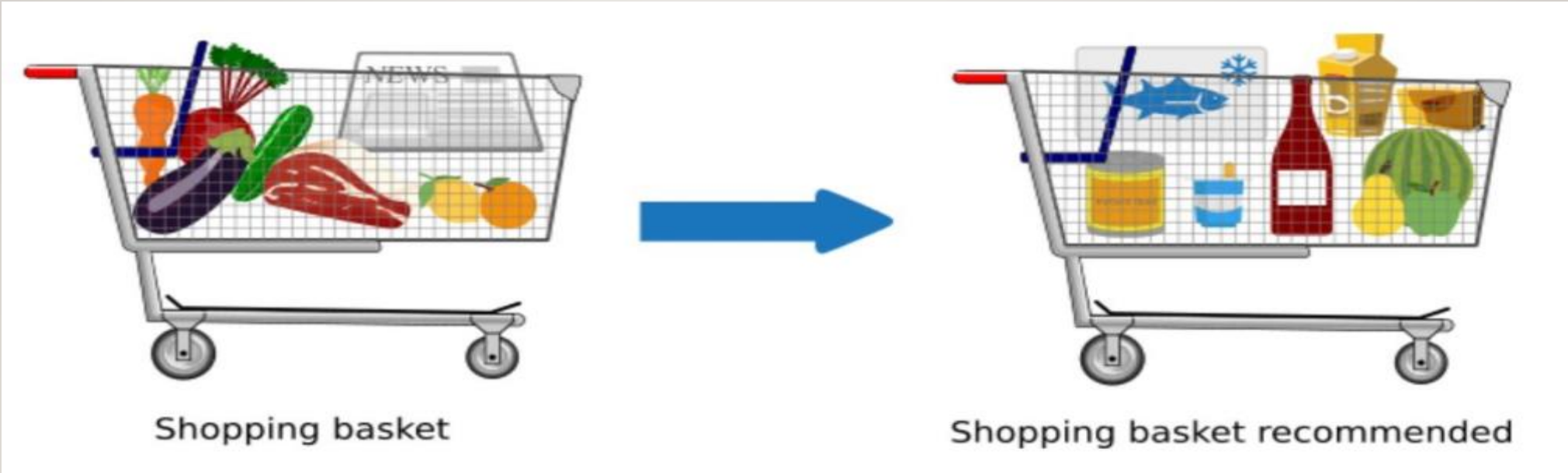
- MOTIVATION

- Determines the products which are bought together and to reorganize the supermarket layout
- Design promotional campaigns such that products' purchase can be improved
- Association Rule Mining is used to find an association between :
  - different objects in a set
  - find frequent patterns in a transaction database



GOAL

- Use sample of over 3 million grocery orders from more than 200,000 Instacart users and build a recommendation system for shopping websites using Apriori algorithm



OVERVIEW

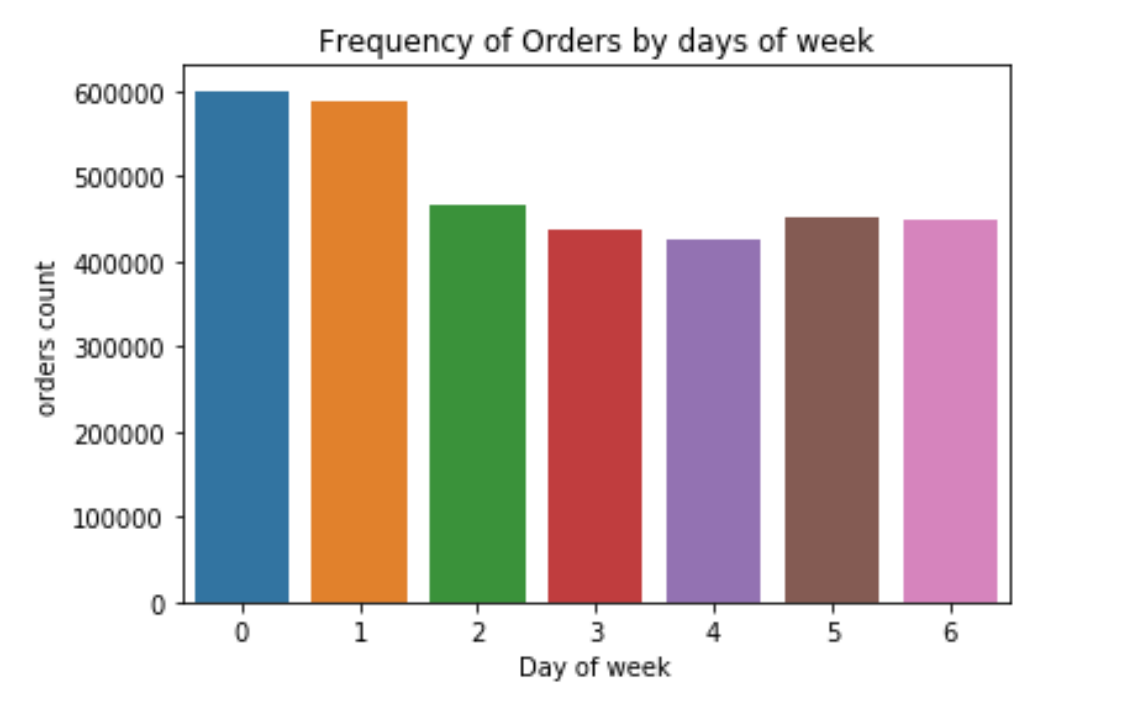
- DATASET
- EXPLORATORY ANALYSIS
- ASSOCIATOIN RULE
- APRIORI ALGORITHM
- RESULT
- FUTURE WORK

DATA SET

- Contains a sample of over 3 million grocery orders from more than 200,000 Instacart users
- For each user, we provide between 4 and 100 of their orders, with the sequence of products purchased in each order
- 6 Data Tables with:
  - aisles
  - departments
  - orders
  - products
  - order_products_prior
  - order_products_train

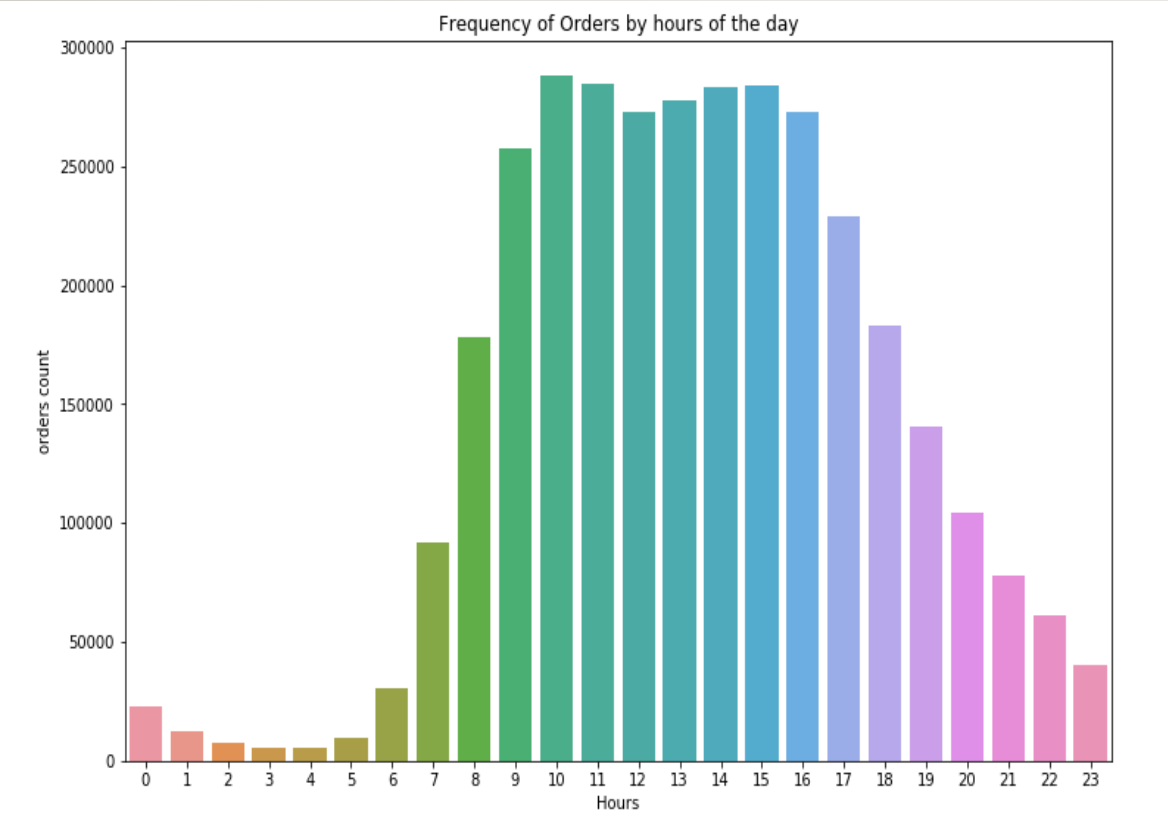
EXPLORATORY ANALYSIS

FREQUENCY OF ORDERS BY DAYS OF WEEK



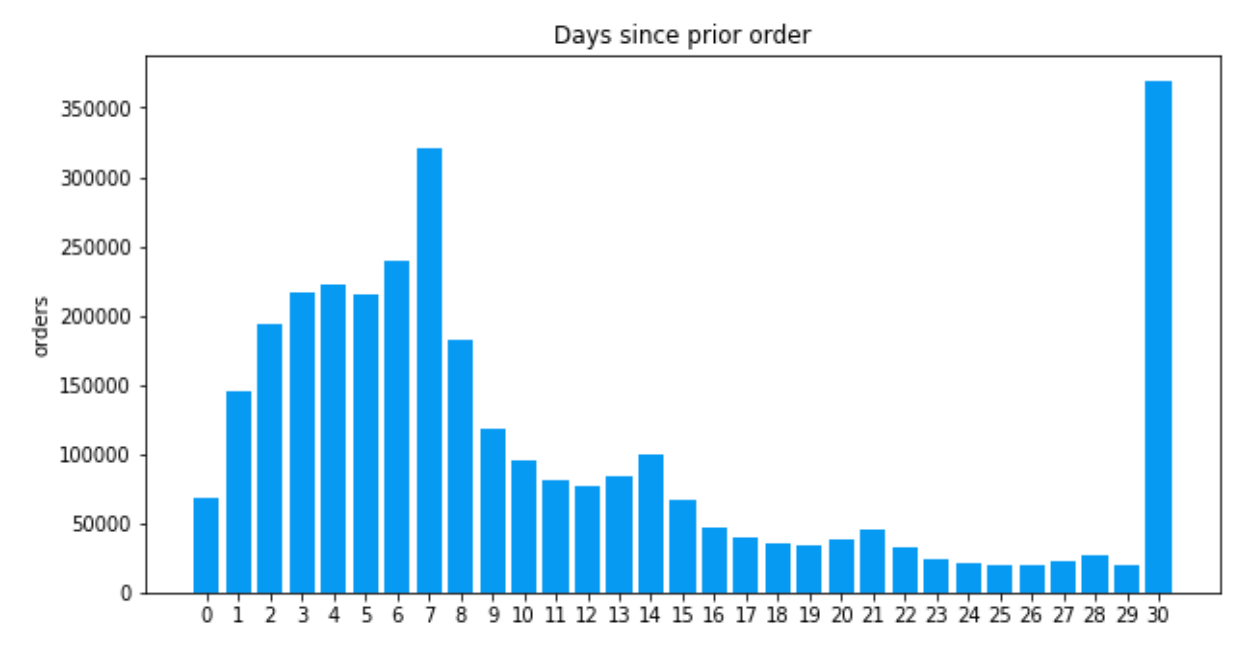
- Most orders are placed in 0 and 1st day of the week.
- May be Saturday and Sunday.
- Need to restock before the weekend

FREQUENCY OF ORDERS BY HOURS OF THE DAY



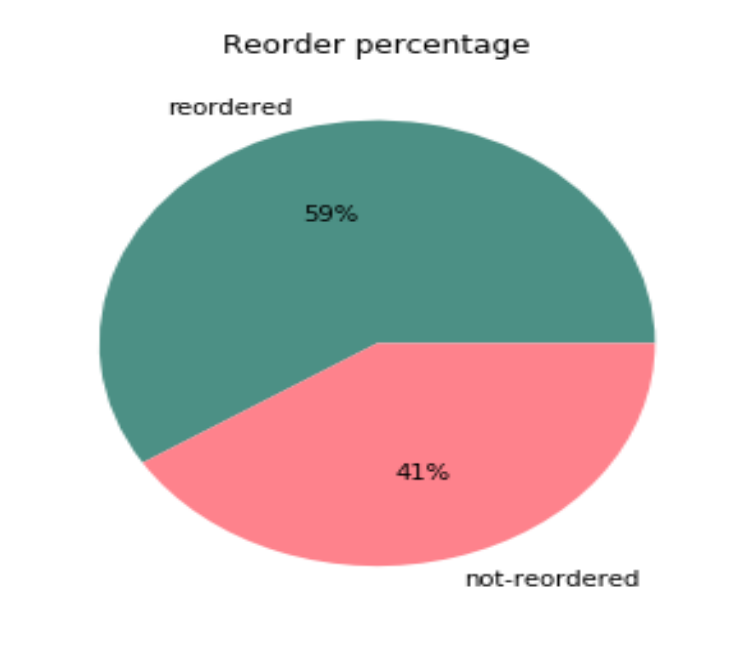
- Majority of orders are made during day time and that also in the morning.
- Website will have a high traffic during that time.
- Frequency of orders start decreasing after 16th hour.
- Good to add any promotional items to increase sales on off hours.

DAYS SINCE PRIOR ORDER



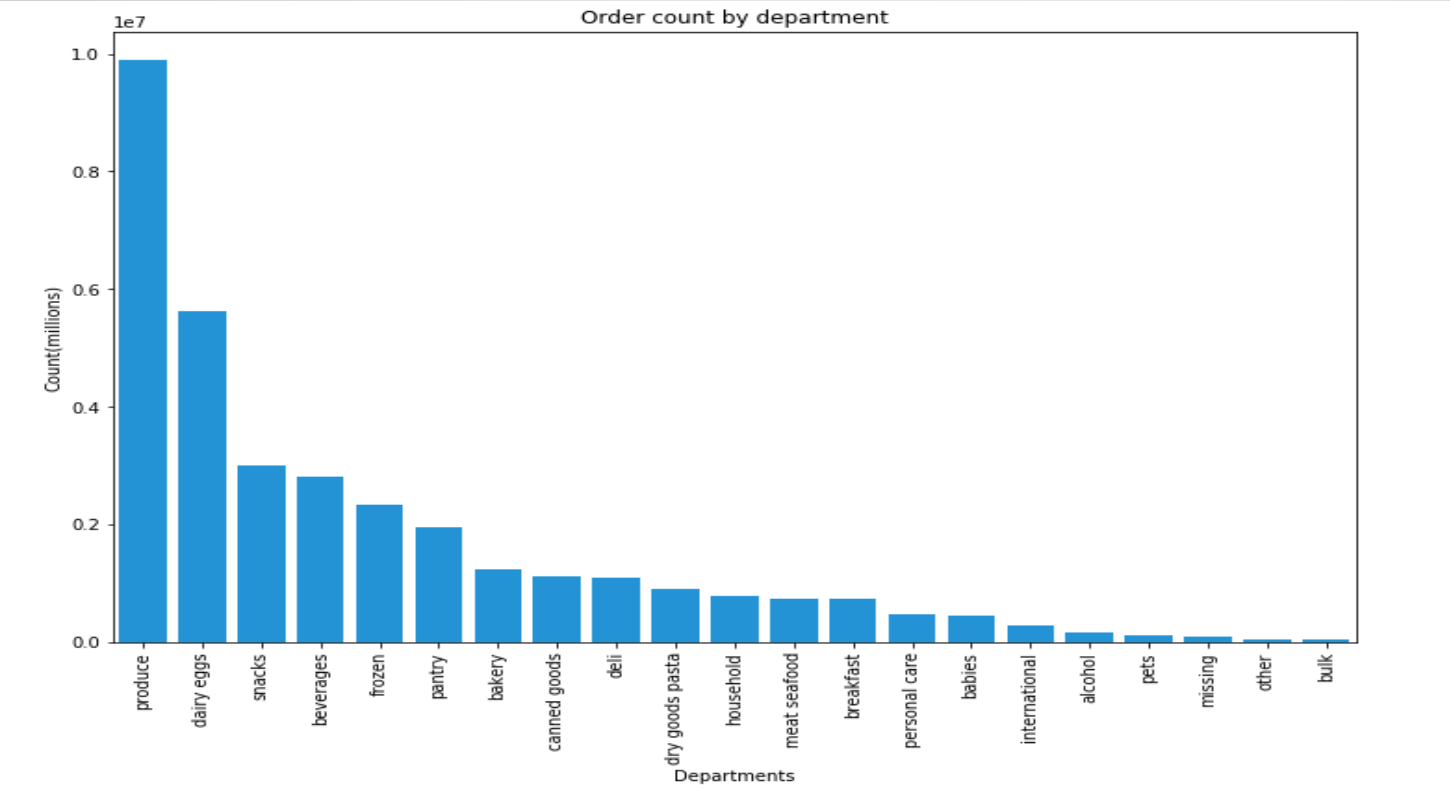
- Weekly and monthly order count has a hike.

REORDER PERCENTAGE



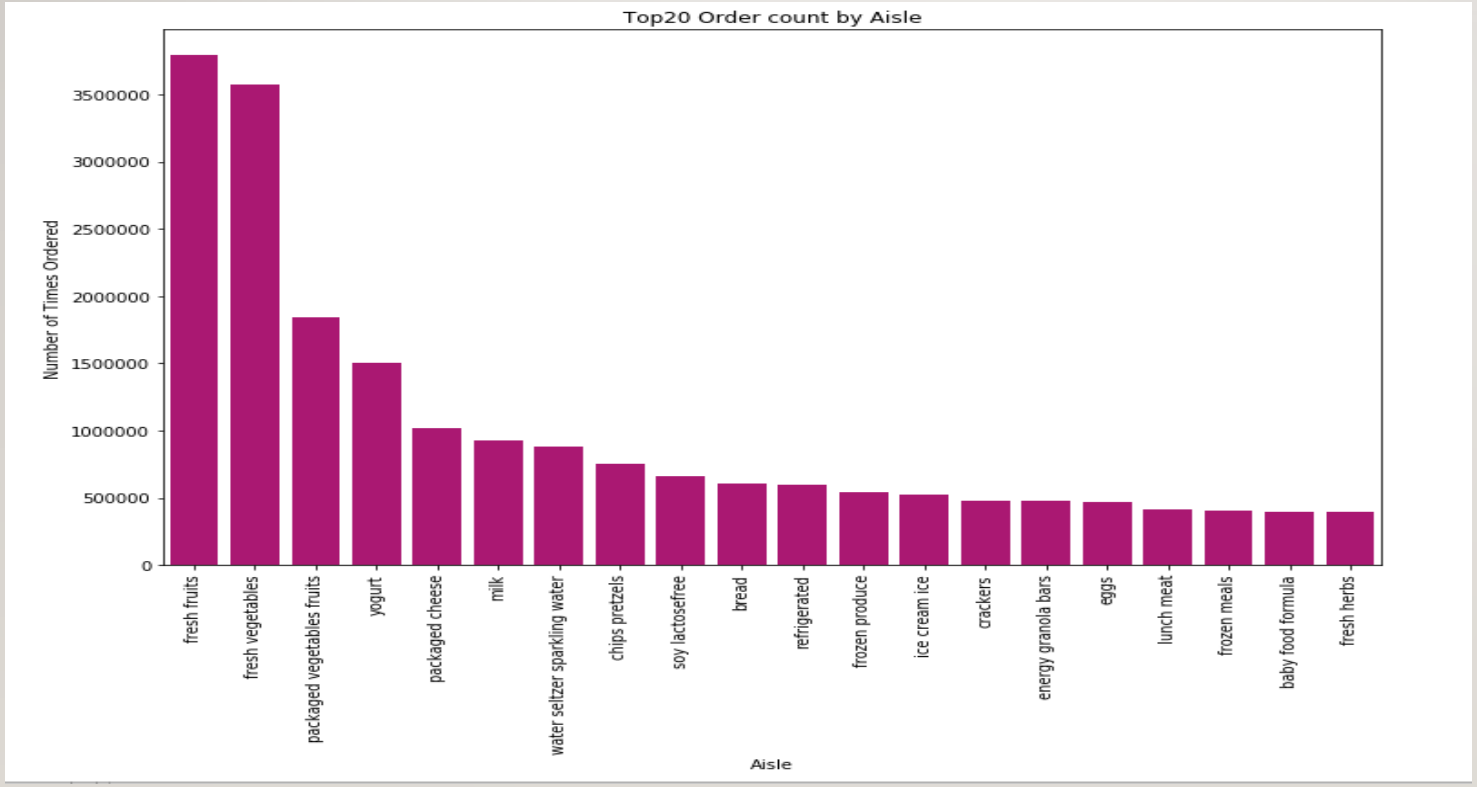
- 59% of products are reordered!

ORDER COUNT BY DEPARTMENT



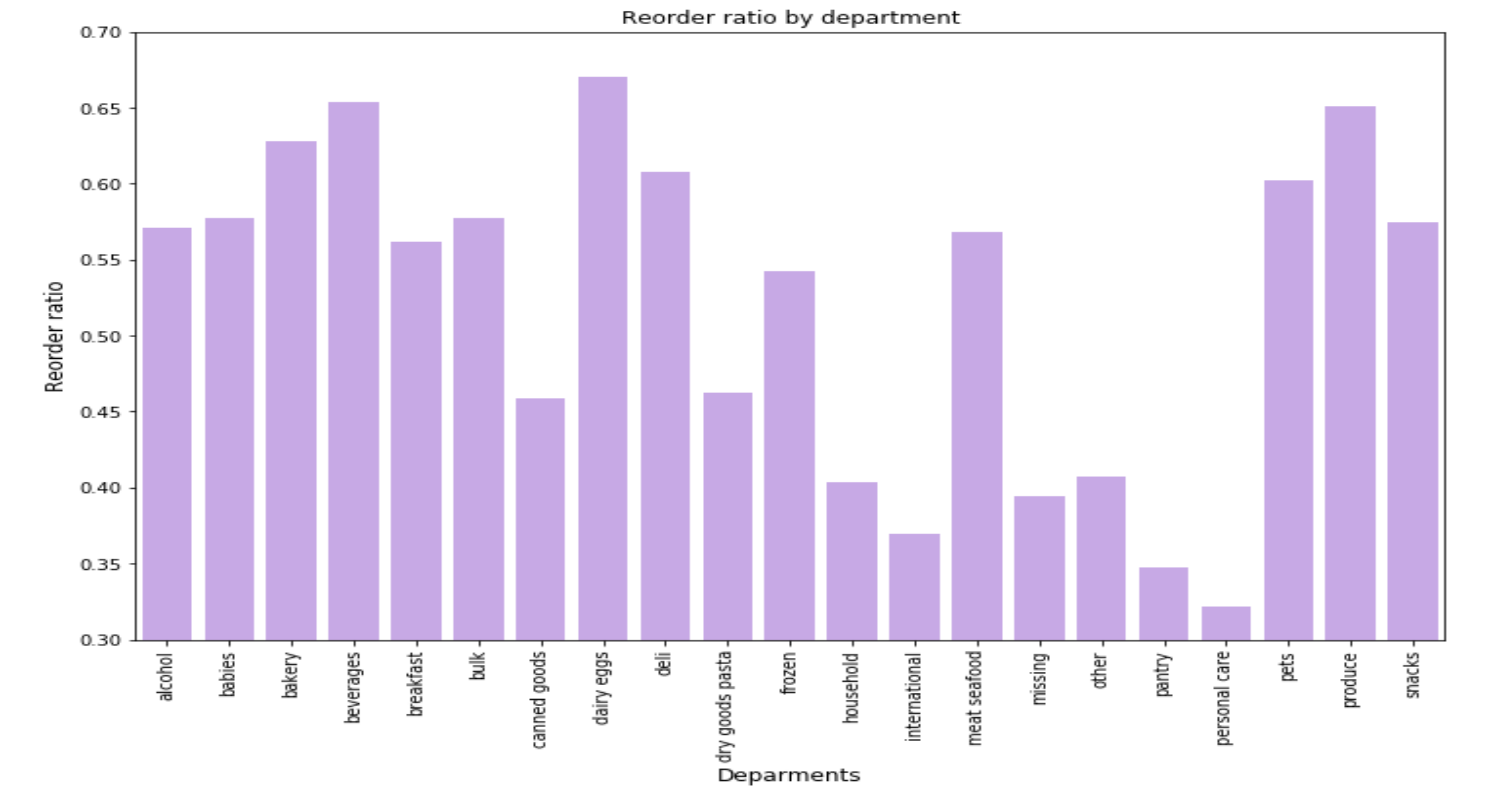
- Produce department has the highest order count.
- Diary eggs and snacks departments are the second and thirds highest order count.

ORDER COUNT BY AISLE



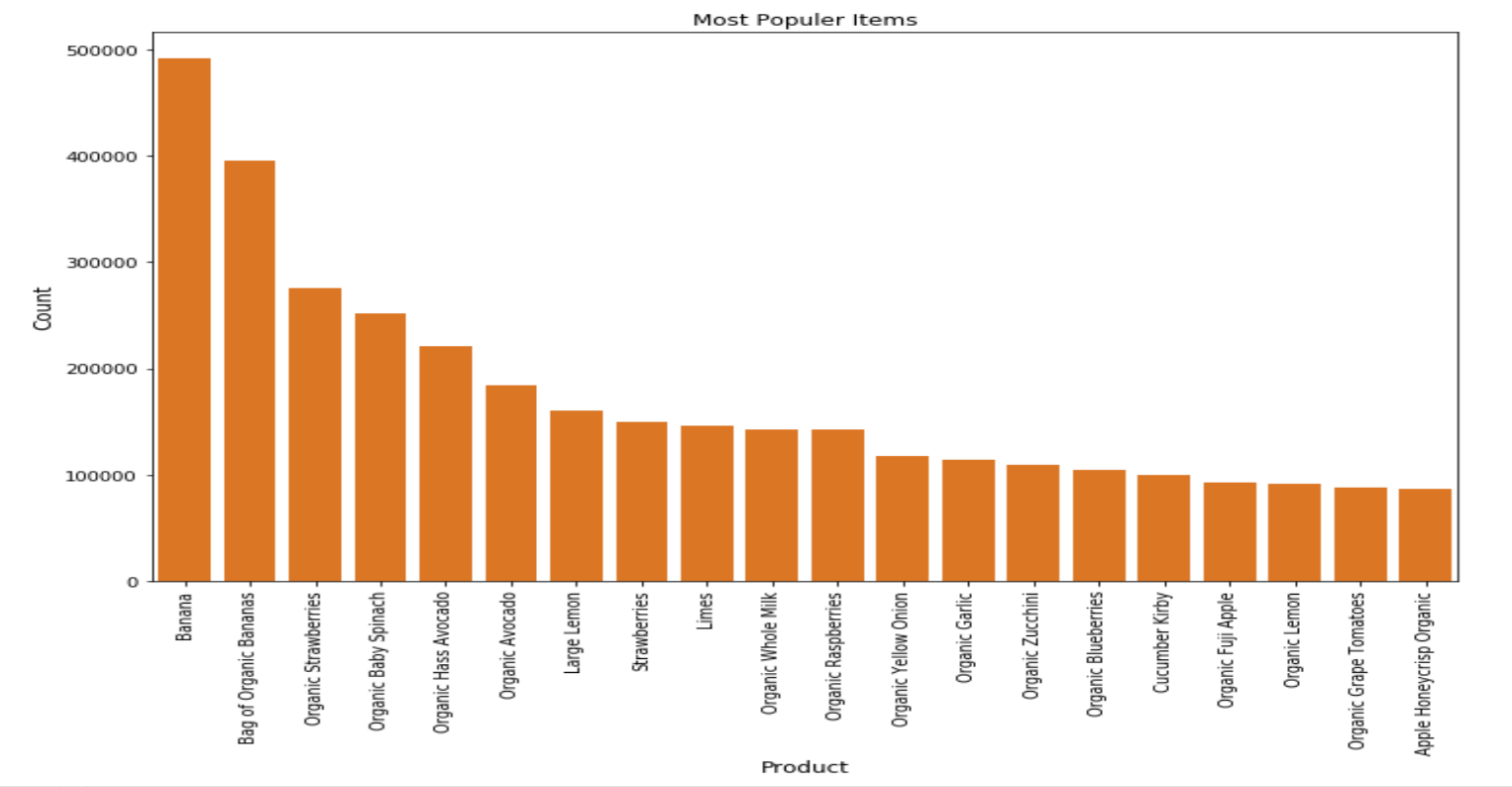
- Aisle with fresh foods has most number of times ordered count.
- Placing some law sales aisles near to these aisles may be helpful to increase sales.

REORDER RATIO BY DEPARTMENT



- Diary Eggs department has the highest reorder ratio.
- Beverages and Produce departments have the second and third highest reorder ratios.

MOST POPULAR ITEMS



- Banana is the most popular product.
- Most of the organic fruits are popular among the buyers.
- 15 out of 5 of top 20 products are organic

ASSOCIATOIN RULE

Definition: Association Rule

- Association Rule
  - An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
  - Example: {Milk, Diaper} → {Beer}

- Rule Evaluation Metrics
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

APRIORI ALGORITHM

A key concept in Apriori algorithm is the anti-monotonicity of the support measure. It assumes that

- All subsets of a frequent itemset must be frequent
- Similarly, for any infrequent itemset, all its supersets must be infrequent too

Step 1: Create a frequency table of all the items that occur in all the transactions.

Step 2: We know that only those elements are significant for which the support is greater than or equal to the threshold support.

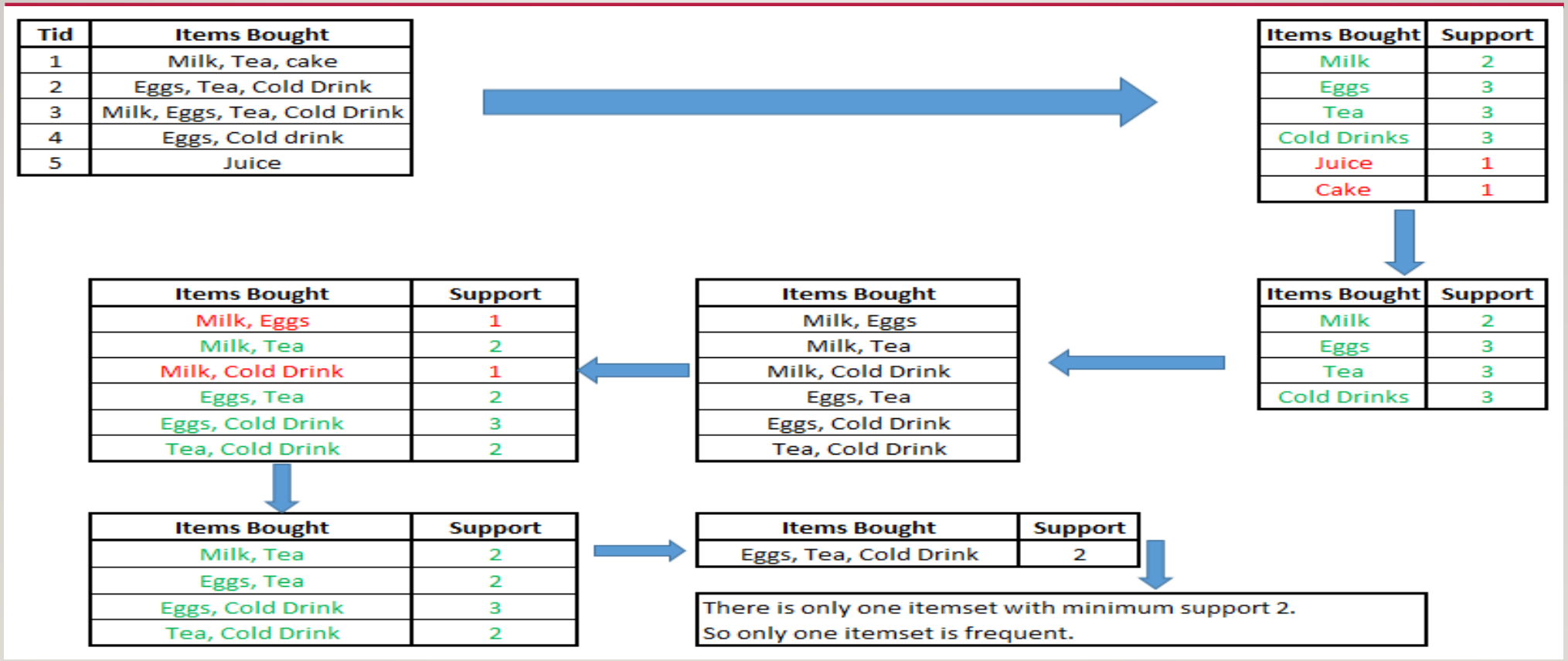
Step 3: The next step is to make all the possible pairs of the significant items keeping in mind that the order doesn't matter, i.e., AB is same as BA.

Step 4: We will now count the occurrences of each pair in all the transactions.

Step 5: Again only those itemsets are significant which cross the support threshold

Step 6: Now let's say we would like to look for a set of three items that are purchased together. We will use the itemsets found in step 5 and create a set of 3 items.

APRIORI ALGORITHM



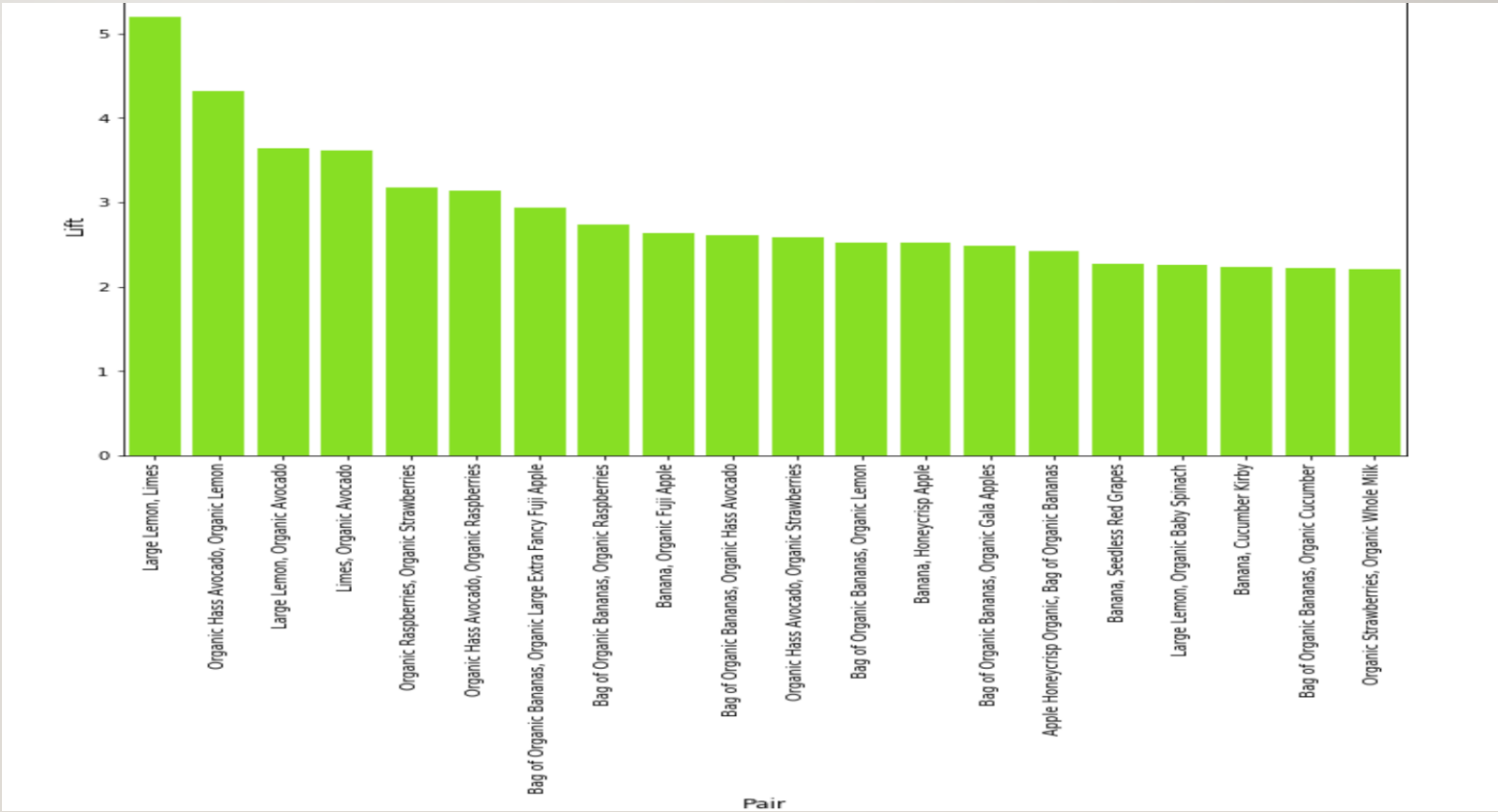
MODEL METRICS

- Number of items is huge, minimum support value set as 0.005 in order to include most items sets
- If minsup is too high => miss item sets with rare items
- If minsup is too low => computationally expensive

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
53	Limes	Large Lemon	0.032882	0.036593	0.006263	0.190476	5.205192
59	Organic Lemon	Organic Hass Avocado	0.023023	0.059994	0.005973	0.259446	4.324557
48	Organic Avocado	Large Lemon	0.047612	0.036593	0.006350	0.133374	3.644744
65	Limes	Organic Avocado	0.032882	0.047612	0.005654	0.171958	3.611635
24	Organic Raspberries	Organic Strawberries	0.038188	0.071273	0.008641	0.226272	3.174710

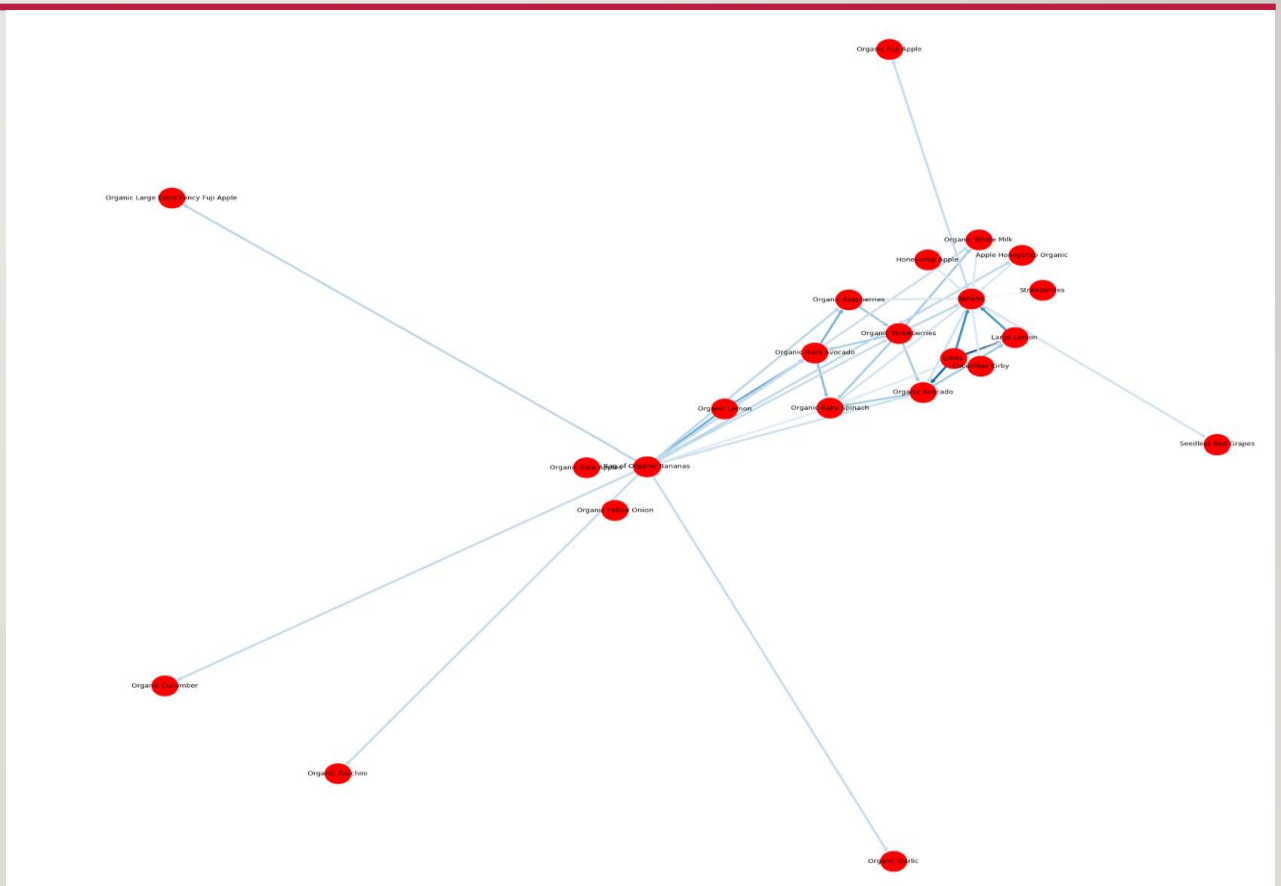
TOP20 ITEM PAIRS BY LIFT METRIC

- Large lemon and limes mostly bought together.
- Organic fruits items mostly bought together
- Similar items in a same category are bought together.



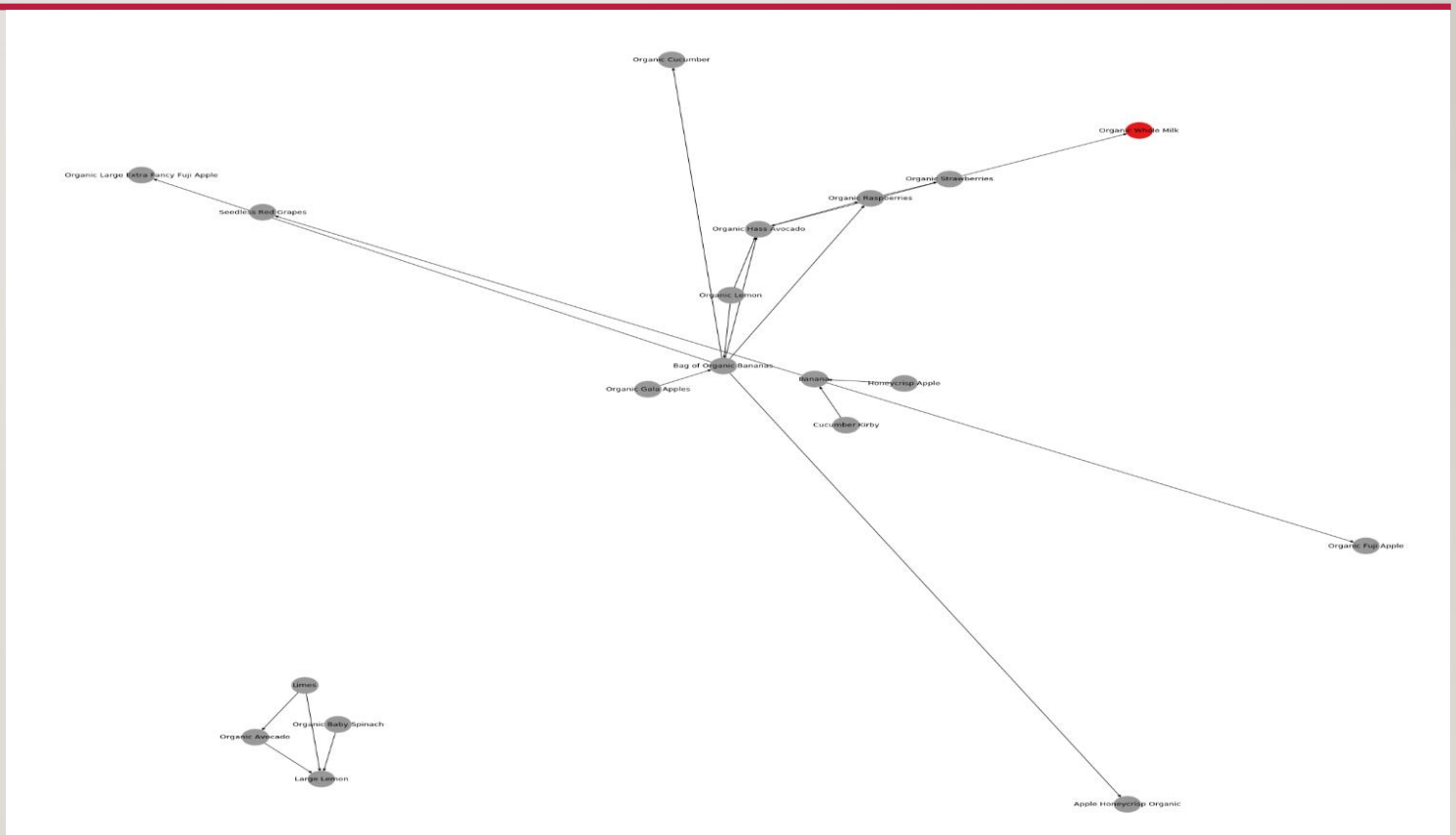
TOP 20 LIFT ITEMS WITH RELATIONSHIP

- Acyclic graph (directed) Edges are weighted by lift
- Organic items tend to be connected to each other with high lift value



TOP 20 LIFT ITEMS AND THEIR RELATIONSHIPS BY DEPARTMENTS

- Organic fruits tend to cluster and they are in the top 20 items set.



FUTURE WORK: HOW TO IMPROVE APRIORI'S EFFICIENCY

- The upcoming researchers can explore such methods/algorithms which can produced single level and multiple level association rules without candidate set generation approach so that it may consumes less time and memory.
- Investigate such rules set theory which can answer mobile users query promptly.
- Design such algorithm which can work efficiently on existing data structure for efficient utilization of memory.

THE END