Marketing and Retail Analytics project

MILESTONE 2 – LAVANYA N RAO

CONTENTS

Conducting Exploratory Data Analysis.

Presenting an Introduction to Market Basket Analysis.

Identification of Association Rules.

Offering Recommendations based on Analysis.

NTRODUCTION

The aim is to pinpoint the most favored combinations, which could then be recommended to the Grocery Store chain following a comprehensive examination of the frequently encountered item sets within customer orders.

DATA EXPLORATION PROCESS

	count	mean	std	min	25%	50%	75%	max
Order_id	20641.0	575.986289	328.557078	1.0	292.0	581.0	862.0	1139.0

```
4730
```

```
Date        0
Order_id    0
Product     0
dtype: int64
```

- The dataset contains 4730 duplicate values, with no missing values present.

DATA EXPLORATION PROCESS

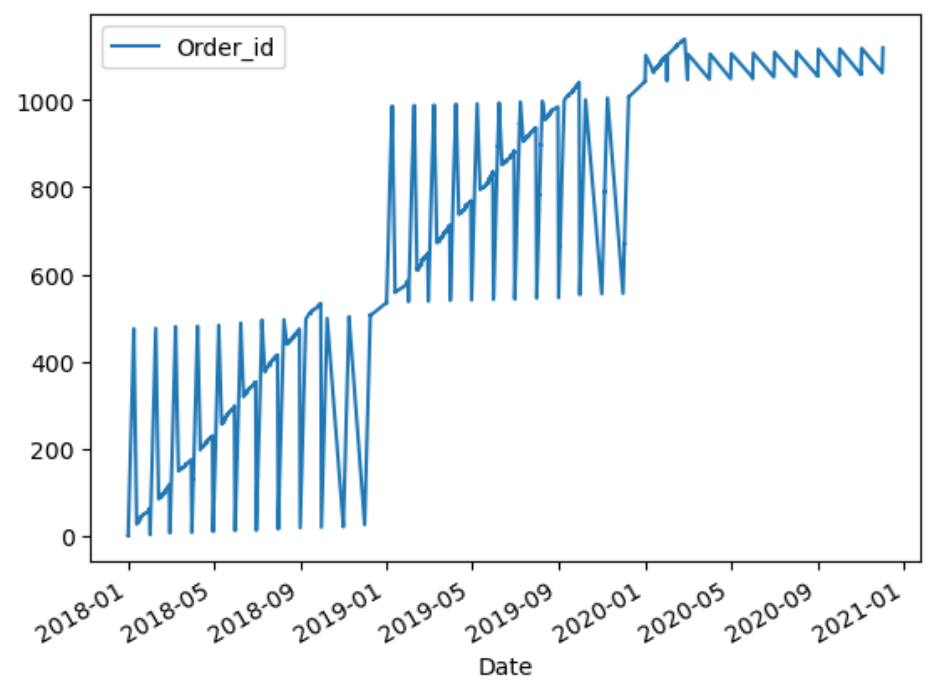
	Date	Order_id	Product
0	01-01-2018	1	yogurt
1	01-01-2018	1	pork
2	01-01-2018	1	sandwich bags
3	01-01-2018	1	lunch meat
4	01-01-2018	1	all- purpose

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20641 entries, 0 to 20640
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      20641 non-null  object
 1   Order_id  20641 non-null  int64
 2   Product   20641 non-null  object
dtypes: int64(1), object(2)
memory usage: 483.9+ KB
```

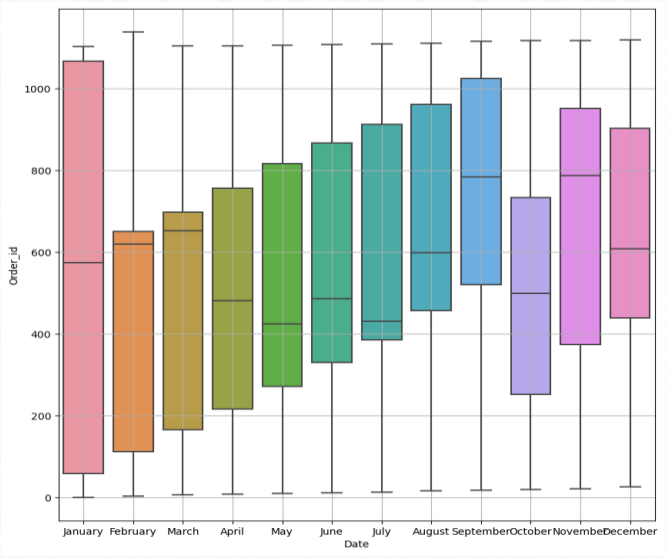
- The store's dataset comprises 20,641 observations.
- The data file includes three columns: date, Order_id, and Product.
- Within the dataset, there are two object data types and one integer data type.

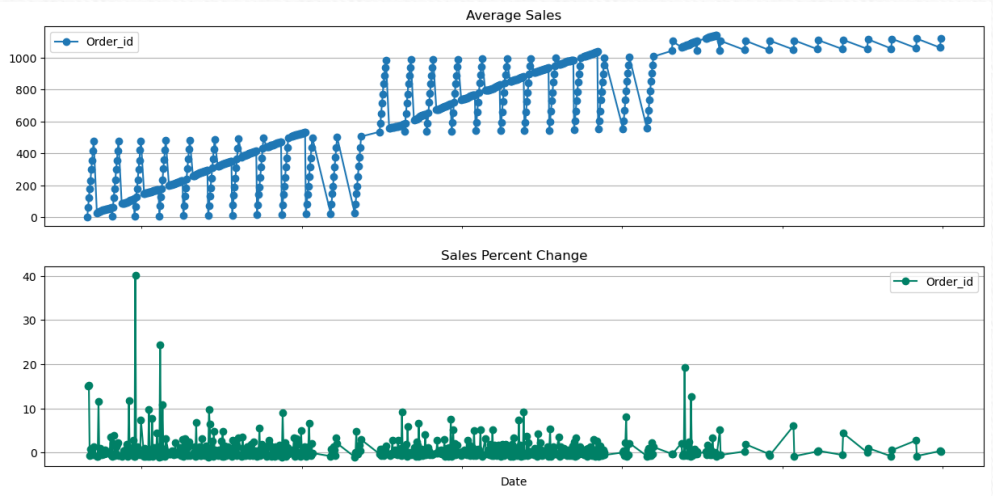
TRENDS ACROSS THE YEARS

- The sales show a steady yearly increase.

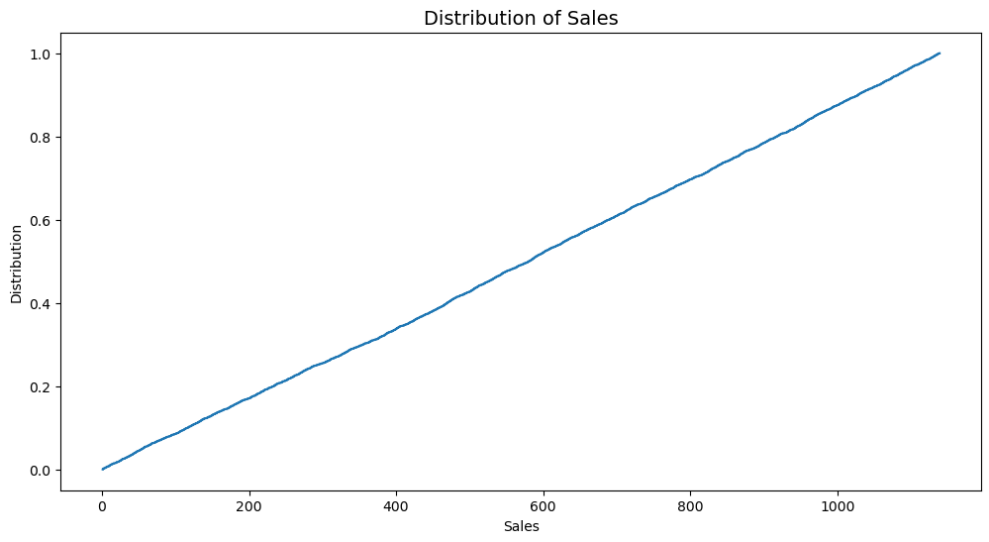


TRENDS ACROSS THE YEARS

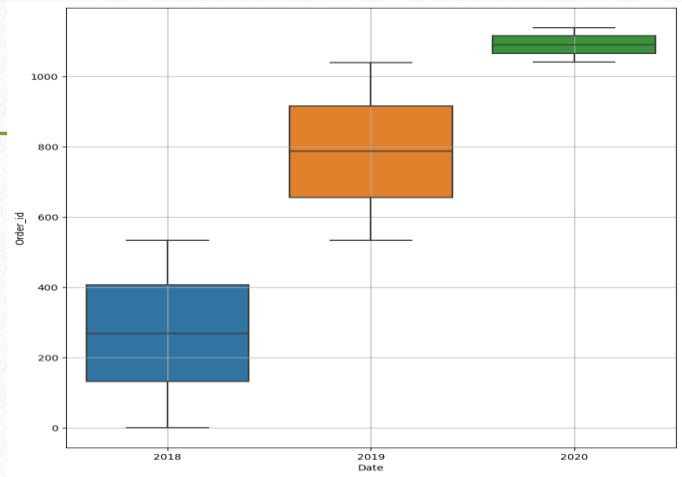




DISTRIBUTION OF SALES

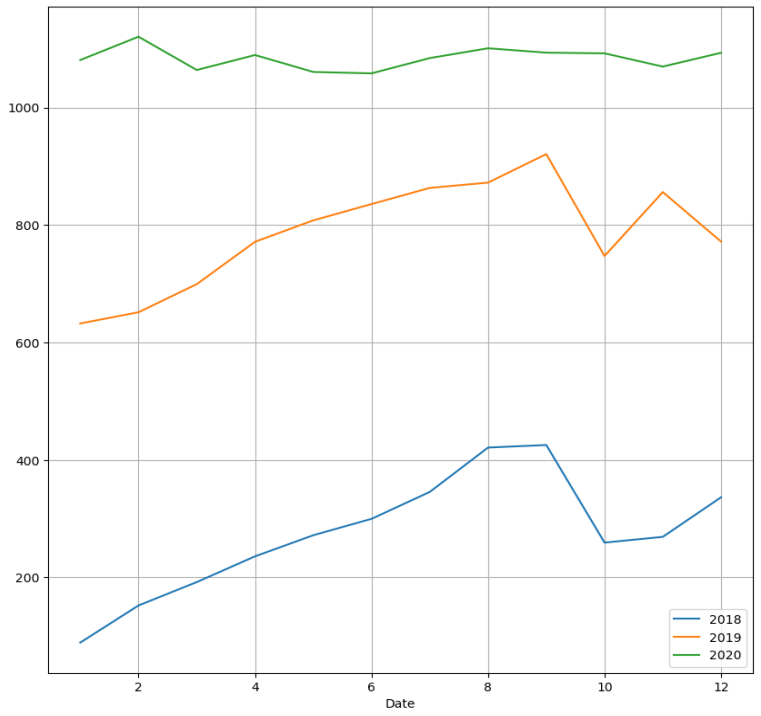


TRENDS ACROSS THE YEARS



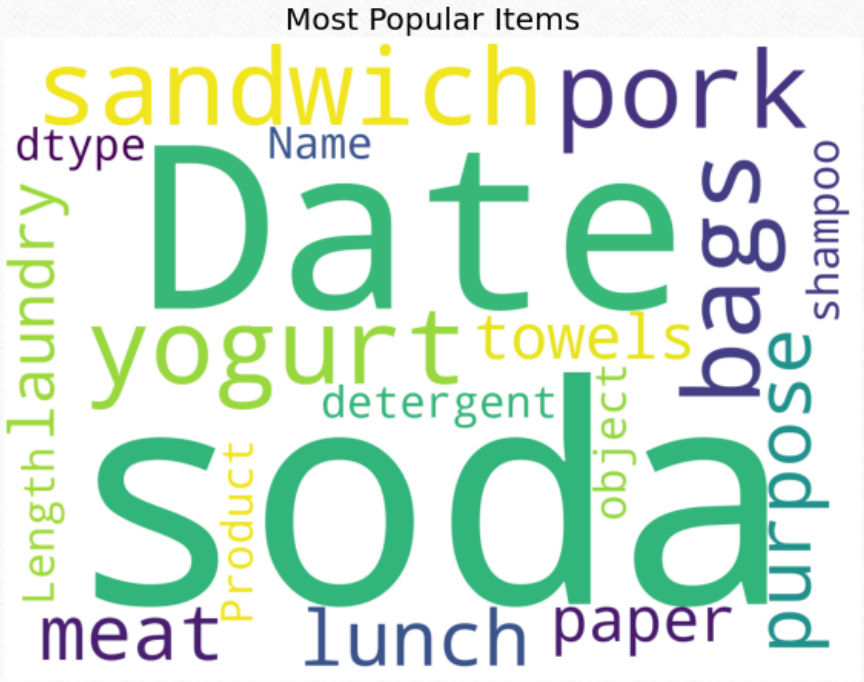
In 2020, the orders peaked, contrasting sharply with the lowest point experienced in 2018.

Date	2018	2019	2020
Date			
1	89.125609	632.501699	1081.015815
2	152.317186	651.580328	1120.710280
3	192.152731	699.412500	1064.000000
4	235.959402	771.545455	1089.533333
5	271.914336	807.997906	1060.687500
6	299.839923	835.753138	1058.311475
7	345.662681	863.277293	1084.321429
8	421.204698	872.293137	1101.020408
9	425.521368	920.856092	1093.545455
10	259.387850	747.331288	1092.405405
11	269.150000	856.179641	1069.830189
12	336.559259	771.931034	1093.350000



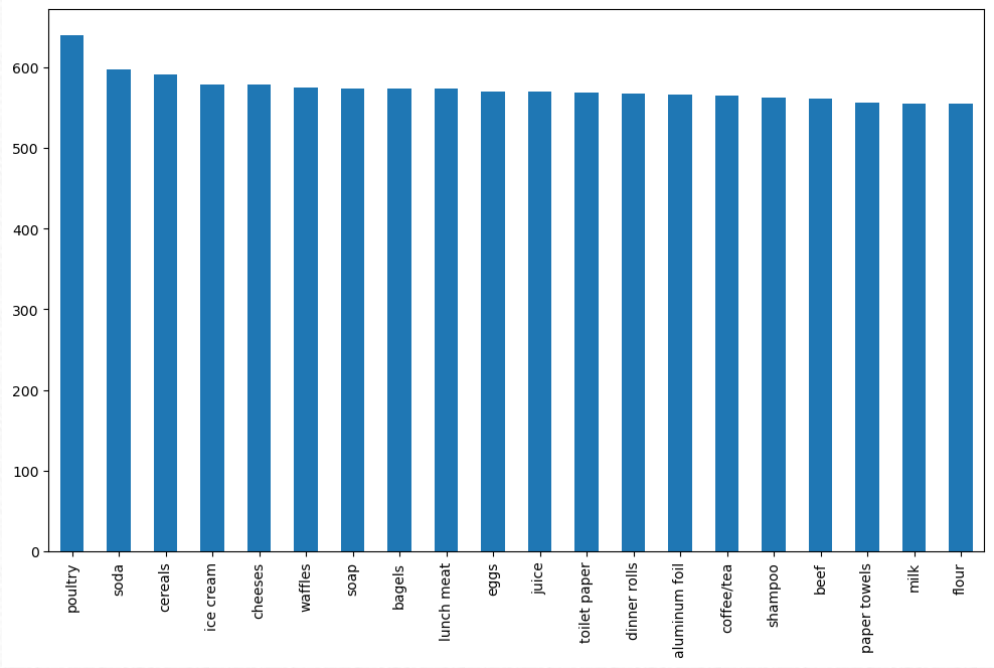
Most Popular Items

- Yogurt, soda, pork, etc., are the products most frequently purchased



TOP 20 PRODUCTS

- Upon examining the bar plot, it becomes clear that poultry, soda, and celery are the top 3 products in the store.



Grocery Store Analysis

Association rules are to be constructed.

The frequent identification of items purchased together is essential.

Recommendations need to be constructed based on this analysis.

ASSOCIATION RULES

Revealing underlying relationships between various items in a dataset, this technique uncovers associations among seemingly unrelated elements..

By identifying hidden relationships among diverse elements in a dataset, this method exposes the interconnectedness of seemingly unrelated items.

Association rules are typically expressed in modified format as i(j) -> i(k), indicating a robust connection between the acquisition of item i(j) and i(k). This signifies that both items were bought concurrently within a single transaction.

Support Threshold Value

The support of an itemset X, supp(X), denotes the proportion of transactions within the database where item X is present, with N representing the total transaction count.

$$Support = \frac{Frequency\ (X,Y)}{N}$$

This metric serves as an indicator of an itemset's popularity, indicating that higher support values correlate with more frequent occurrences of items within the dataset. Conversely, a low support value can aid in uncovering concealed relationships among the items.

Confidence Threshold Value

- The confidence of a rule indicates the probability of item Y being bought alongside item X. For instance, a confidence of .5 implies that in half of the instances where Baby Gel and Soap were bought, Cookies and Chips were also purchased.

$$Confidence\ (A \rightarrow B) = \frac{Support\ (A \cup B)}{Support\ (A)}$$

- A higher confidence value suggests greater reliability of the rule.

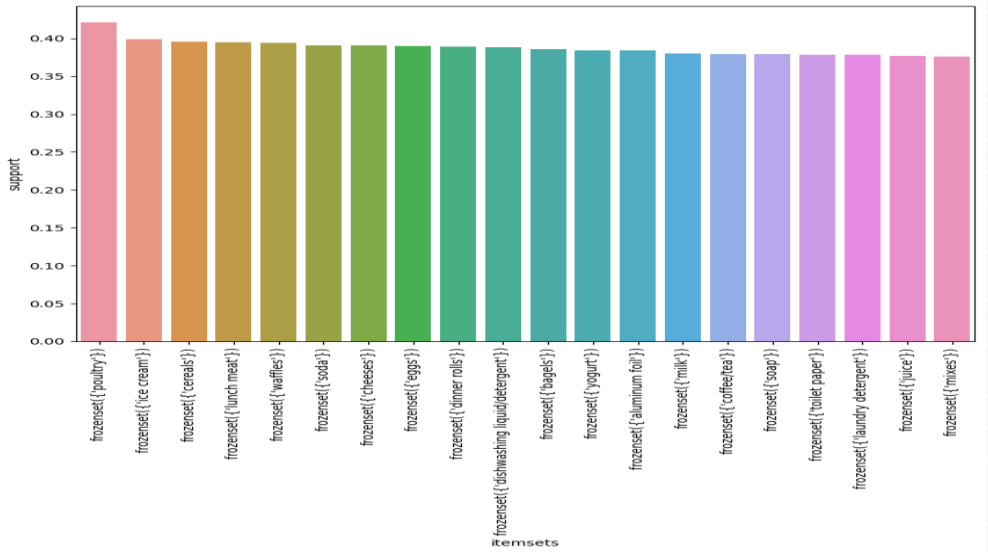
Association Rules

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
21916458	(ketchup, sugar, sandwich bags, all-purpose)	(laundry detergent, soap, flour)	0.023705	0.058824	0.011414	0.481481	8.185185	0.010019	1.815126
21916523	(laundry detergent, soap, flour)	(ketchup, sugar, sandwich bags, all- purpose)	0.058824	0.023705	0.011414	0.194030	8.185185	0.010019	1.211329
21907008	(toilet paper, fruits, all- purpose, coffee/tea)	(butter, beef, cereals)	0.022827	0.058824	0.010536	0.461538	7.846154	0.009193	1.747899
21907073	(butter, beef, cereals)	(toilet paper, fruits, all- purpose, coffee/tea)	0.058824	0.022827	0.010536	0.179104	7.846154	0.009193	1.190374
22026341	(pork, milk, individual meals, ice cream)	(sandwich loaves, shampoo, cereals)	0.020193	0.067603	0.010536	0.521739	7.717674	0.009170	1.949557
22026384	(sandwich loaves, shampoo, cereals)	(pork, milk, individual meals, ice cream)	0.067603	0.020193	0.010536	0.155844	7.717674	0.009170	1.160694
22029873	(ketchup, cheeses, lunch meat, milk)	(pork, soap, coffee/tea)	0.022827	0.065847	0.011414	0.500000	7.593333	0.009910	1.868306
22029908	(pork, soap, coffee/tea)	(ketchup, cheeses, lunch meat, milk)	0.065847	0.022827	0.011414	0.173333	7.593333	0.009910	1.182064
21916456	(ketchup, laundry detergent, sugar, all-purpose)	(flour, soap, sandwich bags)	0.025461	0.059701	0.011414	0.448276	7.508621	0.009893	1.704291
21916525	(flour, soap, sandwich bags)	(ketchup, laundry detergent, sugar, all-purpose)	0.059701	0.025461	0.011414	0.191176	7.508621	0.009893	1.204885
22002793	(pasta, lunch meat, beef, sandwich bags)	(shampoo, fruits, spaghetti sauce)	0.026339	0.053556	0.010536	0.400000	7.468852	0.009125	1.577407
22002808	(shampoo, fruits, spaghetti sauce)	(pasta, lunch meat, beef, sandwich bags)	0.053556	0.026339	0.010536	0.196721	7.468852	0.009125	1.212109
21916490	(flour, laundry detergent, soap, sandwich bags)	(ketchup, sugar, all- purpose)	0.027217	0.057068	0.011414	0.419355	7.348387	0.009860	1.623939
21916491	(ketchup, sugar, all- purpose)	(flour, laundry detergent, soap, sandwich bags)	0.057068	0.027217	0.011414	0.200000	7.348387	0.009860	1.215979
21986699	(pasta, pork, soap)	(soda, ketchup, waffles, bagels)	0.047410	0.030729	0.010536	0.222222	7.231746	0.009079	1.246206
21986646	(soda, ketchup, waffles, bagels)	(pasta, pork, soap)	0.030729	0.047410	0.010536	0.342857	7.231746	0.009079	1.449593
21916481	(flour, sugar, sandwich bags, all- purpose)	(ketchup, laundry detergent, soap)	0.022827	0.069359	0.011414	0.500000	7.208861	0.009830	1.861282

Support for itemsets using Apriori

	support	itemsets
0	0.374890	(all- purpose)
1	0.384548	(aluminum foil)
2	0.385426	(bagels)
3	0.374890	(beef)
4	0.367867	(butter)
...	...	...
610567	0.010536	(ketchup, mixes, spaghetti sauce, pork, soap, ...
610568	0.011414	(ketchup, waffles, mixes, spaghetti sauce, soa...
610569	0.010536	(sandwich loaves, laundry detergent, soap, lun...
610570	0.011414	(yogurt, mixes, milk, sandwich bags, lunch mea...
610571	0.010536	(yogurt, tortillas, mixes, milk, sandwich bags...

610572 rows × 2 columns



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
21916458	(ketchup, sugar, sandwich bags, all- purpose)	(laundry detergent, soap, flour)	0.023705	0.058824	0.011414	0.481481	8.185185	0.010019	1.815126
21916523	(laundry detergent, soap, flour)	(ketchup, sugar, sandwich bags, all- purpose)	0.058824	0.023705	0.011414	0.194030	8.185185	0.010019	1.211329
21907008	(toilet paper, fruits, all- purpose, coffee/tea)	(butter, beef, cereals)	0.022827	0.058824	0.010536	0.461538	7.846154	0.009193	1.747899

Running the algorithm with the parameters set as follows, minimum_support = 0.01, enables us to observe all items with 2 or more appearances in the frequent itemset, alongside their corresponding rules. The calculation methodology for metrics is detailed in preceding slides.

The initial itemset unveils the association: "If Ketchup and Sugar, then Laundry Detergent, Soap, Flour," with a support value of 0.011, approximately 1.1% of all transactions exhibiting this combination.

The confidence value registers at 0.48, indicating a 48% likelihood that the sales of the first item set's antecedents occur whenever the consequents are purchased.

Lift metric serves to assess the relationship between items. A lift value below 1 suggests negligible correlation between Ketchup and Soap items. However, with a lift value of 8.18, it is apparent that the purchase of Ketchup is highly correlated with the acquisition of Soap/Flour items.

RECCOMMENDATIONS

Providing discounts on low-selling items can potentially boost their sales.

Providing discounts on low-selling items can boost their sales significantly.

Combo offers, when implemented, have the potential to boost sales within the store.