AUCTUS: The Grocery Store Analysis

DOM 304: Contemporary Business Analytics

THE TEAM

Aamaya Kumar

Rahul Madan

Hriday Pradhan

Mehak Agrawal

Kathyayani Rana

Eshaan Sharma

1. Problem Statement
2. Problem Description
3. Understanding the Data
4. Data Preparation
5. Model Building and Evaluation
6. Deployment
7. Recommendations & Conclusion
8. Acknowledgement

PROBLEM STATEMENT

The Grabbo convenience store at Shiv Nadar University seeks to boost sales. We'll perform Market Basket Analysis. This process identifies the purchasing patterns of customers by finding associations between the different items that customers place in their "shopping baskets". Retailers and marketers will benefit from the finding of this kind of relationship since it will provide them insight into which products people regularly purchase together, which will help them design marketing strategies.

PROBLEM DESCRIPTION

Our analysis has a number of advantages that can help our in-house convenience store. They are as follows:

INCREASED PROFITS

When products are placed in locations with most traction, they are sold more. This will increase the profits of Grabbo.

INCREASED TRACTION

If Grabbo sells products that are not available throughout the campus, it will increase traction in their store premises.

GREATER PRODUCT MIX

By understanding the needs of the consumers through our model, Grabbo can increase their product mix.

APRIORI ALGORITHM

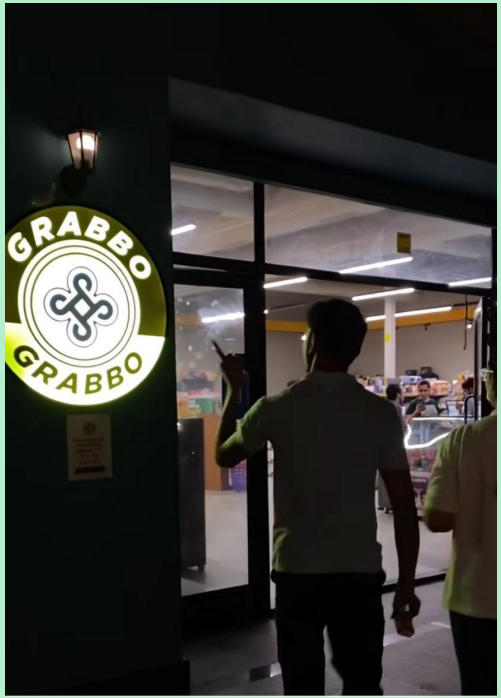
Apriori algorithm refers to an algorithm that is used in rule mining for frequent products sets and relevant association rules. Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers buy at a Convenience Store.

Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store by suggesting combos which are mostly sold together.

BUSINESS UNDERSTANDING: ABOUT GRABBO

- Conventional stores are a one-stop-shop for all households.
- Grabbo, the conventional store of Shiv Nadar University, is the only one on a 297 acre residential campus. Students, faculty and staff pay visit to the store for their daily needs that include snacks, drinks, packaged food items, daily grocery and more.
- Grabbo's invoice data set can help us analyse various aspects to increase profitability of the business.
- The invoice data can help us correlate the reasons behind frequent purchases and group purchases along with ascertaining the most sold items.
- With the help of the invoice data we aim to analyse, interpret and evaluate the combinations best sold together, strategic placement of products in the premise and strategic discounts that can help increase sales.

BUSINESS UNDERSTANDING: ABOUT GRABBO



BUSINESS UNDERSTANDING: GRABBO

6336

Total Unique Items

90515

Total Unique Bills

266557

Total Data Points

Components of Apriori Algorithm

SUPPORT

Support refers to the default popularity of any product. You find the support as a quotient of the division of the number of transactions comprising that product by the total number of transactions.

CONFIDENCE

Confidence refers to the possibility that the customers bought both biscuits and chocolates together. So, you need to divide the number of transactions that comprise both biscuits and chocolates by the total number of transactions to get the confidence.

LIFT

Lift measures the performance of a targeting model (known as an association rule) at predicting a specific outcome, compared with a random choice. Therefore, Lift is the ratio between target response and average response. That is to say that Lift is a ratio between confidence and expected confidence.

Understanding the Data

Details

We collected transaction data from Grabbo convenience store - SNU

Dates: August 2021 to November 2022

Total months: 15 months

```
In 3  1  df
```

Out 3

	No	Bill number	Date	Item	Code	HSN Code	Unit	Qty
0	1361...	1361722	01/01/22	Cadbury - Oreo Chocolatey Sandwich ...	3095	19053100.0	PAC	2.0
1	1361...	1361822	01/01/22	Surf Excel Washing Soap Bar - 80 gms	1898	34011930.0	PAC	1.0
2	1361...	1361922	01/01/22	Vim - Lemons - 145 ml	2491	34022010.0	PAC	2.0
3	1361...	1361922	01/01/22	Nescafe - Classic Coffee Sachet - 7...	3043	210111.0	PAC	3.0
4	1362...	1362022	01/01/22	B Fizz- 160 ml	2435	NaN	CAN	1.0
...	...	...	...	...	...	...	...	...
265582	4014...	4014823	30/09/22	Catch - Fruit Juice - Apple - 200 m...	5517	22029020.0	BTL	2.0
265583	4014...	4014923	30/09/22	Catch-Flavoured drink- Orange	5835	22021020.0	PCS	1.0
265584	4015...	4015023	30/09/22	Epigamia - Milkshake - Strawberry -...	4696	22029930.0	PAC	1.0
265585	4015...	4015023	30/09/22	Galaxy Crispy - 18 gms	5878	18069010.0	PAC	1.0
265586	4015...	4015023	30/09/22	Maggi - Veg Atta Noodles - 72.5 gms	6036	19023010.0	PAC	2.0

Understanding the Data

COLUMNS (Total = 8)

- **No: string** - Unique bill number for each financial year (composition = <bill number>'\'<financial year>)
- Bill number: integer - derived from 'No' by removing "/"
- Date: Date - date of invoice in format DD/MM/YY
- Item:string - purchased product name
- Code: integer- unique product ID
- HSN code: integer - Internationally recognized product codes
- Unit: string(3) - unit of measure of item (eg. BTL for bottle, CAN for tin can etc)
- Qty: float - quantity of that item purchased (eg. 0.9 Kg Apples, 1.0 Pepsi)

Exploratory Data Analysis

```
In 5   1  len(df['Code'].value_counts())
       2  # total number of unique items in dataset

Out 5      6336
```

TOTAL NUMBER OF UNIQUE ITEMS IN DATASET = 6336

Exploratory Data Analysis

```
In 29   1  len(bill_and_items)
        2  # total number of bills

Out 29      90515
```

Total number of Invoices in data = 90515

Exploratory Data Analysis

```
In 13   1  df.shape[0]
        2  # Total rows in dataset

Out 13     265587
```

Total number of rows in data = 265587

Exploratory Data Analysis

```
In 6   1  print(df.isna().sum())

       No                0
       Bill number       0
       Date              0
       Item              0
       Code              0
       HSN Code      12265
       Unit              0
       Qty               0
       dtype: int64
```

Null values in all columns

5 Number summary for Quantity of items purchased

```
df[['Item','Qty']].describe()
```

	Qty
count	265536.000000
mean	1.822661
std	20.836134
min	0.010000
25%	1.000000
50%	1.000000
75%	2.000000
max	10000.000000

5 Number Summary for quantity:

1. Min = 0.01 (Nestle creamer)
2. Max = 10000 (Bisleri 250ml)
3. Lower quartile = 1.0
4. Median = 1.0
5. Upper quartile = 2

Data preparation

- We created new column named ‘Bill Number’ which combines “No” and “date” to give a unique bill number for bills in 2021 and 2022.
- Observed HSN code has null values using isnull() function
- Removed - No, date, HSN code, unit and quantity from data as they are not required in Apriori algorithm
- Changed data type of product code “Code” to int64
- Removed null rows

```
In 10   1  excl_merged.dtypes

Out 10     No              object
           Bill number      int64
           Date            object
           Item            object
           Code             int64
           HSN Code       float64
           Unit            object
           Qty            float64
           dtype: object
```

Challenges in data

Challenge: Due to nan value in 1 row, we were not able to change the data type of product code “Code”

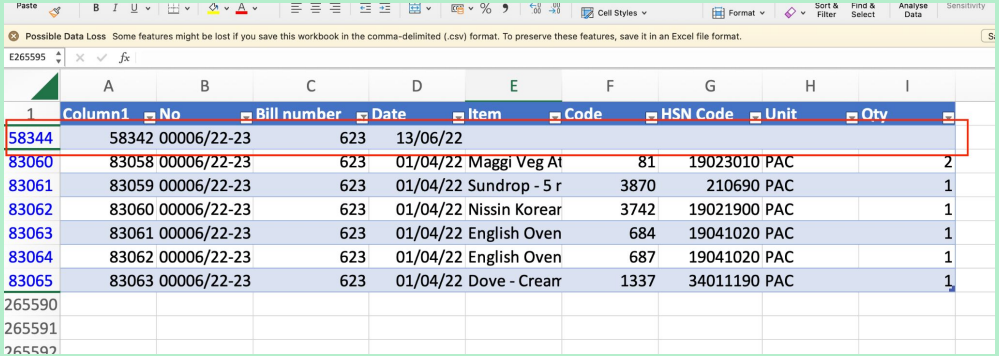
Solution: We removed the row from data

```
In 51  1  # change product Code data type
       2
       3  excl_merged = excl_merged.astype({"Code": int})

Traceback…
IntCastingNaNError: Cannot convert non-finite values (NA or inf) to integer

In 54  1  nan_in_col  = excl_merged[excl_merged['Code'].isna()]

In 55  1  nan_in_col
```

Out 55

	Unnamed: 0	No	Bill number	Date	Item	Code	HSN Code	Unit	Qty
58342	58342	00006/22-23	623	2022-06-13	NaN	NaN	NaN	NaN	NaN

1 rows × 9 columns Open in new tab



	A	B	C	D	E	F	G	H	I
1	Column1	No	Bill number	Date	Item	Code	HSN Code	Unit	Qty
58344	58342	00006/22-23	623	13/06/22					
83060	83058	00006/22-23	623	01/04/22	Maggi Veg At	81	19023010	PAC	2
83061	83059	00006/22-23	623	01/04/22	Sundrop - 5 r	3870	210690	PAC	1
83062	83060	00006/22-23	623	01/04/22	Nissin Korear	3742	19021900	PAC	1
83063	83061	00006/22-23	623	01/04/22	English Oven	684	19041020	PAC	1
83064	83062	00006/22-23	623	01/04/22	English Oven	687	19041020	PAC	1
83065	83063	00006/22-23	623	01/04/22	Dove - Cream	1337	34011190	PAC	1
265590									
265591									

Challenges in data

Challenge: Items having same names are having different product code “Code”

Solution: similar product grouping using python

	Item	Code
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460
22	Aashirvaad Shudh Chakki Atta - 5 Kg	3541
22	Aashirvaad Shudh Chakki Atta - 5 Kg	460

Similar product grouping

Results from Apriori Algorithm can be improved significantly by grouping similar/duplicate products into single product.

We used FUZZY STRING MATCHING in python to group similar item names into 1 category

Threshold for string match:
We used 80% similarity threshold for item names to be grouped together after manually looking at the results on different threshold values

FUZZY STRING MATCHING IN PYTHON

Used for grouping similar/ duplicate items

Fuzzy string matching works on Levenshtein distance between strings. We used library named ‘fuzzywuzzy’ to get the similarity score between 2 strings.

We used fuzzywuzzy.sort_token_ratio which does not consider placement of words in the strings.

Eg. ‘Real Fruit Juice - Guava 1 ltr’ & ‘1 ltr Real Guava Fruit Juice’ would be assigned a score of 100 ie. full match.

```
In 2
1  from fuzzywuzzy import fuzz
2  value = fuzz.token_sort_ratio('1 ltr Real Guava Fruit Juice', 'Real Fruit Juice - Guava 1 ltr')
3  print('Match score: ' + str(value))

Match score: 100
```

Example of grouped Items in data

```
2491: ['Vim - Lemons - 145 ml',
 'Vim Gel - Lemon - 250 ml',                    Group 1
 'Vim - Lemons - 155 ml'],

3043: ['Nescafe - Classic Coffee Sachet - 7.5 gms',
 'Nescafe Classic Coffe Powder - 25 gms',
 'Nescafe Classic Soluble Coffee Powder Stick Sachet - 1.5 gms',
 'Nescafe - All In 1 Instant Coffee Sachet - 16 gms',
 'Nescafe Classic Cooffee Jar- 50 gms',
 'Nescafe - Classic Coffee Pouch - 5.5 gms',              Group 2
 'Nescafe Classic Coffee Powder - 100 gms',
 'Nescafe Classic Pack - 50 gms'],
```

Result of grouping similar/duplicate items

Old number of unique Items in dataset = 6336

Similarity Threshold = 80%

New number = 4200

Reduction of 2136 items

34% reduction in unique items by grouping similar/ duplicate items.

```
1  len(excl_merged['Code'].unique())

   6336
```

```
len(excl_merged['Code'].unique())

4200
```

Association rules on different confidence and support values

```
rule = apriori(transactions = input_for_internet_code, min_support = 0.0004, min_confidence = 0.4, min_lift = 1.5, min_length = 2, max_length
  = 2)
```

	Left_Hand_Side	Right_Hand_Side	Support	Confidence	Lift
0	Potato (Aaloo)	Onion (Pyaaz)	0.000818	0.490066	249.204182
1	Tomato (Tamatar)	Onion (Pyaaz)	0.000873	0.456647	232.210333
2	Royal Homeware - VM Plastic Bucket - 16...	Aucto Plastic Mug - 1.5 Lit.	0.000895	0.536424	173.408586
3	Bauli Moonfils - Caramel Creme - 45 gms	Bauli Moonfils - Vanilla - 45 gms	0.000464	0.461538	37.035597

Association rules on different confidence and support values

```
rule = apriori(transactions = input_for_internet_code, min_support = 0.001, min_confidence = 0.03, min_lift = 1.5, min_length = 2, max_length
  = 2)
```

	Left_Hand_Side	Right_Hand_Side	Support	Confidence	Lift
0	Cup Noodles Veggie Monchow - 70 gms	Cup Noodles - Spiced Chicken - 70 gms	0.001050	0.111765	11.092525
1	Cup Noodles Veggie Monchow - 70 gms	Cup Noodles - Mazedaar Masala - 70 gms	0.001922	0.204706	13.963039
2	Haldiram's Plain Bhujia - 200 gms	Parle Hide & Seek Chocolate Chip Cookie...	0.001094	0.075630	3.097589
3	Banana (Kela)	Apple (Seb) - Fuji	0.001757	0.071237	13.213068
4	Banana (Kela)	English Oven - Brown Bread - 400 gms	0.001447	0.058692	2.876277
5	Banana (Kela)	Amul Gold Full Cream Milk - 500 ml	0.001116	0.045251	2.449692
6	Banana (Kela)	Amul Taaza Toned Milk - 500 ml	0.001922	0.077957	2.173168

Association rules on different confidence and support values

```
rule = apriori(transactions = input_for_internet_code, min_support = 0.0003, min_confidence = 0.35, min_lift = 1.5, min_length = 2,
 max_length = 2)
```

```
output_DataFrame
```

	Left_Hand_Side	Right_Hand_Side	Support	Confidence	Lift
0	Potato (Aaloo)	Tomato (Tamatar)	0.000663	0.397351	207.897255
1	Potato (Aaloo)	Onion (Pyaaz)	0.000818	0.490066	249.204182
2	Tomato (Tamatar)	Onion (Pyaaz)	0.000873	0.456647	232.210333
3	Papaya (Papeeta)	Banana (Kela)	0.000331	0.441176	17.891169
4	Nutella Hazelnut Sprread with Cocoa - 3...	English Oven - Brown Bread - 400 gms	0.000376	0.354167	17.356468
5	Kissan Mixed Fruit Jam - 100 gms	English Oven - Brown Bread - 400 gms	0.000630	0.372549	18.257322
6	Royal Homeware - VM Plastic Bucket - 16...	Aucto Plastic Mug - 1.5 Lit.	0.000895	0.536424	173.408586
7	Aucto - VM Transparent Plastic Bucket -...	Aucto Plastic Mug - 1.5 Lit.	0.000320	0.547170	176.882412
8	Mangal Kalash Plastic Bucket - 20 Ltr.	Aucto Plastic Mug - 1.5 Lit.	0.000309	0.482759	156.060345
9	Dove - Body Love Nourished Radiance Bod...	Veto Plast - Plastic Buckets - 16 Lit.	0.000342	0.645833	679.739583
10	Bauli Moonfils - Caramel Creme - 45 gms	Bauli Moonfils - Vanilla - 45 gms	0.000464	0.461538	37.035597

Association rules on different confidence and support values

```
rule = apriori(transactions = input_for_internet_code, min_support = 0.0004, min_confidence = 0.3, min_lift = 1.5, min_length = 2, max_length = 2)
```

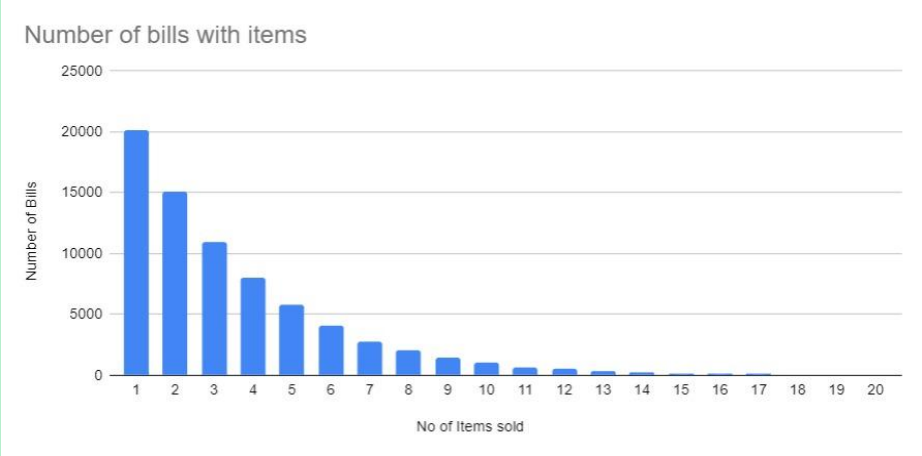
	Left_Hand_Side	Right_Hand_Side	Support	Confidence	Lift
0	Potato (Aaloo)	Tomato (Tamatar)	0.000663	0.397351	207.897255
1	Potato (Aaloo)	Onion (Pyaaz)	0.000818	0.490066	249.204182
2	Tomato (Tamatar)	Onion (Pyaaz)	0.000873	0.456647	232.210333
3	Apple (Seb) - Fuji	Banana (Kela)	0.001757	0.325820	13.213068
4	Orange (Santatra) Nagpur	Banana (Kela)	0.000497	0.338346	13.721047
5	Apple - Royal Del	Banana (Kela)	0.001370	0.322917	13.095341
6	Kissan Mixed Fruit Jam - 100 gms	English Oven - Brown Bread - 400 gms	0.000630	0.372549	18.257322
7	Royal Homeware - VM Plastic Bucket - 16...	Aucto Plastic Mug - 1.5 Lit.	0.000895	0.536424	173.408586
8	Maggi Masala e Magic - 6 gms	Maggi Masala Noodles - 280 gms	0.003712	0.344615	13.869658
9	Tetley Orijinal - 100 Tea Bags	Nestle Everyday Dairy Creamer - 3 gms x...	0.000431	0.312000	109.037375
10	Tetley Orijinal - 100 Tea Bags	Uttam White Sugar Sachet - Double Refin...	0.000420	0.304000	125.075273
11	Smoodh - Milk Shake - Coffee Frappe - ...	Smoodh Chocolate Milk - 85 ml	0.001282	0.329545	40.583411
12	Bauli Savoriz Puff Roll - Cheese Oregan...	Bauli Moonfils - Vanilla - 45 gms	0.000519	0.305195	24.489989
13	Bauli Moonfils - Caramel Creme - 45 gms	Bauli Moonfils - Vanilla - 45 gms	0.000464	0.461538	37.035597
14	Lotte - Chewits - 23 gms	Lotte - Caramilk - 23 gms	0.000718	0.312500	116.884039

All Association rules

We have stored all 2000+ rules retrieved from data analysis in a CSV file on GITHUB LINK HERE

We can apply filters according to our need and get the relevant association rules.

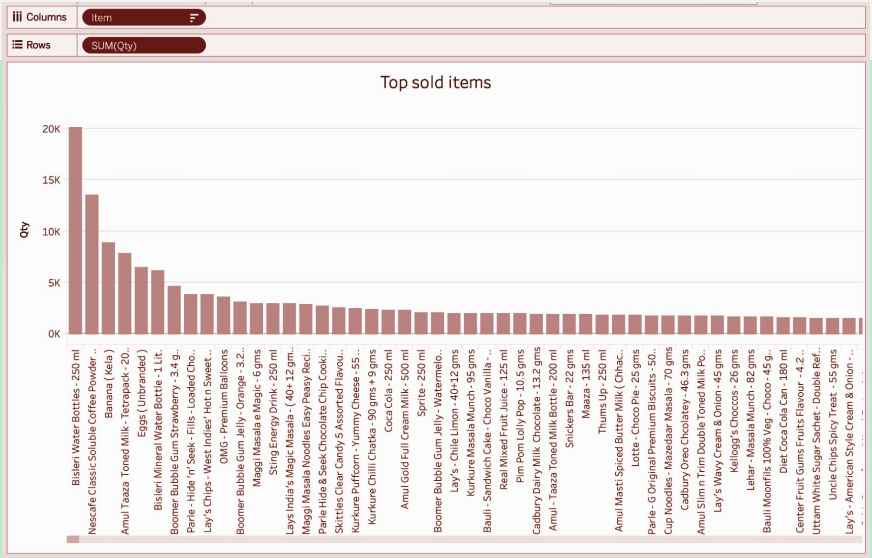
Data Analysis



Most of the customers are buying 1 item and leaving arcade shopping store. As we increase the number of items purchased in 1 bill, we can see a constant decrease in number of instances.

21% customers bought only 1 item
16% customers bought 2 items, and so on...

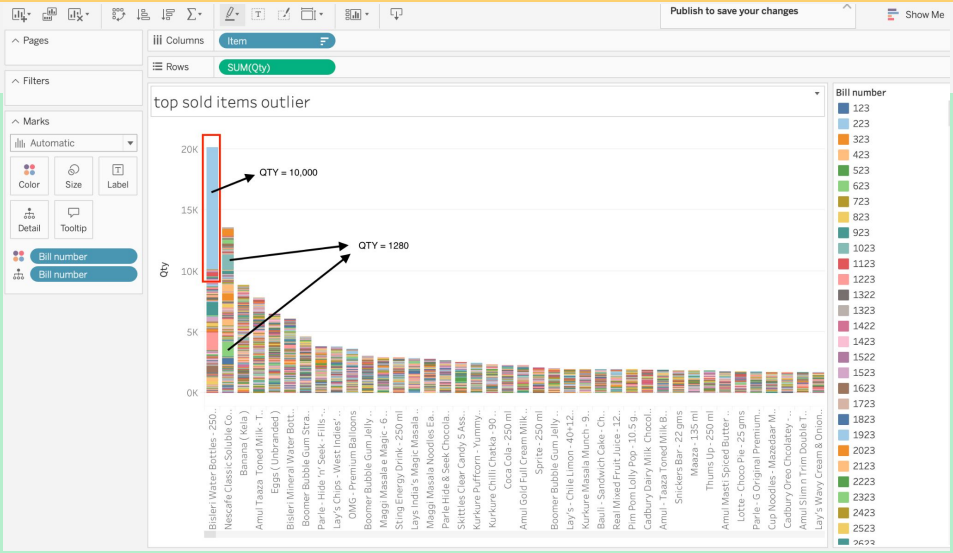
Data Analysis



We can observe that The top most sold items by quantity are:

1. Bisleri 250 ml water bottle
2. Nescafe Coffee powder sachets 2gm
3. Banana (Kela)
4. Amul taaza toned Milk 250ml
5. Eggs (unbranded)

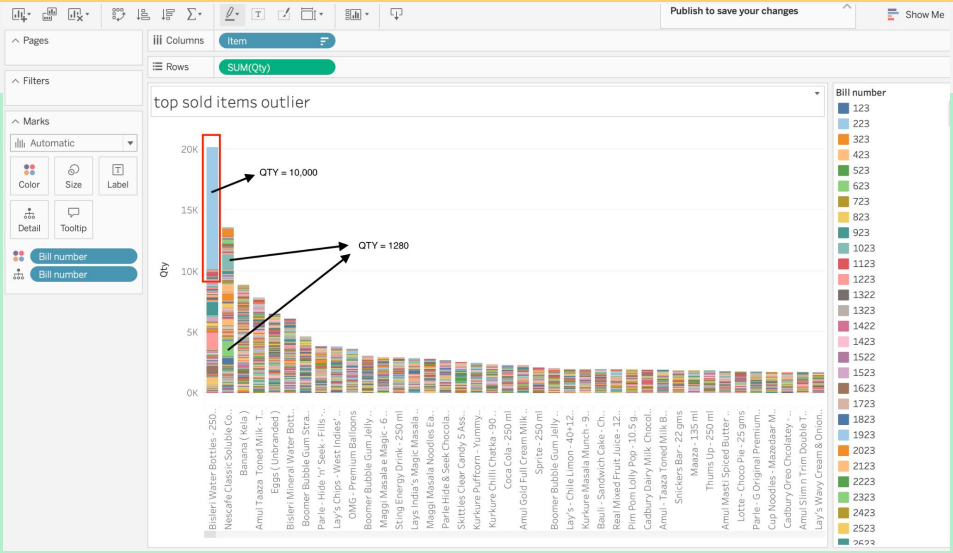
Data Analysis



On putting colour on bill number, we observed outliers in the data which were giving skewed results:

1. 1 bill having 10000 bottles of bisleri 250ml
2. Multiple bills having exact 1280 Nescafe Coffee Sachets

Data Analysis : Inference



Inference

1. 1 bill having 10000 bottles of bisleri 250ml : **Bought by SNU for Convocation ceremony**
2. Multiple bills having exact 1280 Nescafe Coffee Sachets: **full packs of qty 1280 bought by students/ gym enthusiasts.**

Data Analysis : Inference

Inference

1. Number of customers rose from the month of Jan to march as those were initial months of Arcade Grabbo store on SNU
2. There is a steep dip in number of customers from May to July as there are very few students on campus in those months
3. As students came back in August, the number of customers increased and at this time Grabbo added multiple new products in the store which further increased the customers.

RECOMMENDATIONS AND CONCLUSION

- The retail industry's unspoken motto of cross-selling and upselling is what encourages customers to make larger purchases. For such companies, using market basket analysis in data mining to identify patterns and extract customer insights to improve the performance of their brands, has become a thriving component.
- According to an urban tale, a grocery store saw an increase in sales after grouping beer and diapers together since a market basket analysis revealed that both products were frequently purchased by men.
- Companies are gaining billions by manipulating client thought processes when they employ this strategy strategically. It is a practical strategy to increase sales without spending more time or money on marketing that won't yield the same amazing results. So go ahead and give it a try on all the data you have stored to find patterns that can completely surprise you.

ACKNOWLEDGEMENT

We would like to extend our sincere gratitude to Professor Vallurupalli Vamsi, School of Management and Entrepreneurship, Shiv Nadar University for providing us with the opportunity for taking this course and the project. We are thankful for his constant guidance throughout the course and pushing us to do better. We got a practical learning experience with the help of this project.

THANK YOU!