# BeNanna Bakery Products Report

Feb 2021

Prepared by: Youssef Razak

**Product Report**

# Content



# Introduction

A business's value proposition is its products, in essence, the business is the products thus becoming paramount for the business to keep a close eye on the performance of each of its products to maximize profits by identifying which products are performing well and which are dragging down the rest. In this report, we will take a summary view of products' performance and perform association analysis to figure out the interaction between the company's products in terms of sales.

By the end of the report, we expect to have a better understanding of BenNannas product line.

# The Data

## We have retrieved from the business POS system 5 years of transactional data, from 01-01-2016 to 31-12-2020. Which contained 421189 rows of data. How was the data proccesed?

### Categories

We started by looking at the product categories, most of the categories were unique and present throughout all of the 5 years of data.
We merged Cakes & Pies, Cakes and Pies and Cakes to one category: Cakes & Pies. We decided not to remove categories that were not present for the full 5 years of data in the cleaning phase and filter them out when necessary later down the analysis. We are left with 18 distinct product categories as follows: Pastry, Breads, Buns, Sourdough, Cakes & Pies, Puff Pastry, Specialty, Cookie, Promotion, Frozen, None, Seasonal, Baking supplies, Whole Sale, Dairy, Online Store, Courses, and lunch.

### Items

We had multiple instances of duplicated items for example :
mocha puffs and mocha puff or sour dough and sourdough being used to identify the same item. We corrected these inconsistencies by merging the items together and keeping the single version of the item name or correctly spelled version. After standardizing item names we procced to removing items in the list that are not 'valid', for example, we consider an item not valid if the name contains the word (Void) and (Voucher), this was true for 735 void items and 266 voucher items that were removed from the data. Promotional items such as 2 for $10 sourdough, free pastry, groupon, and non specific items such as custom amount and misc were removed from the data. Then the only thing left was to drop duplicated rows and rows with a negative quantity.
After removing these datapoints we are left with 149 unique items and 395380 rows of data which means we shed 25809(6.12%) rows of redundant data.

# An Example of the final Dataset

| date_time | Category | Item | Qty | Transaction ID | year | month | week_name | week | is_weekend |
|---|---|---|---|---|---|---|---|---|---|
| 2016-12-31 23:57:24 | Breads | swiss white | 2 | fdnGVxRa2XBw7Fnx8gtVqhxeV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:57:24 | Specialty | oliebollen | 1 | fdnGVxRa2XBw7Fnx8gtVqhxeV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:56:19 | Specialty | oliebollen | 1 | V4ri6t97vjyCzA7ldEQRyCleV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:54:09 | Specialty | oliebollen | 1 | RS1iJSk5N0CWNwrfKBhvKW4eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:52:46 | Specialty | oliebollen | 1 | b7V19z0TRWRpafHX8PCBzn0eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:52:46 | Buns | soft white dinner buns | 2 | b7V19z0TRWRpafHX8PCBzn0eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:50:53 | Specialty | oliebollen | 1 | lU9EZRVH9rflcYILVXACgvkeV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:50:53 | Buns | raisin buns | 1 | lU9EZRVH9rflcYILVXACgvkeV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:49:13 | Specialty | stollen | 1 | FaqDuiYzoM1mC18GwNcHz95eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:49:13 | Buns | raisin buns | 1 | FaqDuiYzoM1mC18GwNcHz95eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:49:13 | Specialty | oliebollen | 1 | FaqDuiYzoM1mC18GwNcHz95eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:47:29 | Specialty | stollen | 1 | nXqYMVV4wnzKrlD960SZgT3eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:47:29 | Breads | sourdough | 1 | nXqYMVV4wnzKrlD960SZgT3eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:42:39 | Specialty | oliebollen | 1 | NqC2LofkZlR9Tmc2bar36zieV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:41:33 | Puff Pastry | sausage rolls | 8 | xc8x0BQBNIWWyjeyGAEeowkeV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:41:33 | Pastry | serious coconut tart | 1 | xc8x0BQBNIWWyjeyGAEeowkeV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:40:56 | Puff Pastry | sausage rolls | 1 | PlnGDD5X4Jd1vesOEzY3ym3eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:40:56 | Breads | sourdough | 1 | PlnGDD5X4Jd1vesOEzY3ym3eV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:39:58 | Buns | raisin buns | 1 | 1MLnNxpaPVkv6PQT3zU9c8leV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:39:58 | Cookie | coconut loonie | 1 | 1MLnNxpaPVkv6PQT3zU9c8leV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:39:58 | Puff Pastry | apple turnover | 1 | 1MLnNxpaPVkv6PQT3zU9c8leV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:39:58 | Cakes & Pies | black forest cake | 1 | 1MLnNxpaPVkv6PQT3zU9c8leV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:39:11 | Puff Pastry | party rolls | 2 | HvExL7rWzMXw5kTWBoCuq8keV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:39:11 | Cakes & Pies | mocha cake | 1 | HvExL7rWzMXw5kTWBoCuq8keV | 2016 | 12 | Saturday | 52 | TRUE |
| 2016-12-31 23:39:11 | Cakes & Pies | fresh fruit variable | 1 | HvExL7rWzMXw5kTWBoCuq8keV | 2016 | 12 | Saturday | 52 | TRUE |

**Each row of data corresponds to an individual item sale, the row contains information on what the product is, what category it belongs to, and which transaction its part of.**

## Our Cleaned data set contains:

- 395380 rows.
- 149 unique items.
- 16 unique categories.
- 5 unique years
- 165435 unique transactions

**WALT DISNEY, CEO**

We keep moving forward, opening new doors, and doing new things because we're curious and curiosity keeps leading us down new paths.

# The Analysis

## 1st - Exploratory data analysis.
## 2nd - Market basket analysis.

***Exploratory data analysis***

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.

- What categories sold the most units?
- What categories are growing the fastest?
- What items sold the most units?
- What items sales are growing the most?
- How many units per average transaction?
- What's the average number of unit sales per day of the week.

***Market basket analysis***

Market Basket Analysis is a technique that identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence. A co-occurrence is when two or more things take place together.

Market Basket Analysis creates If-Then scenario rules, for example, if item A is purchased then item B is likely to be purchased. The rules are probabilistic in nature or, in other words, they are derived from the frequencies of co-occurrence in the observations. Frequency is the proportion of baskets that contain the items of interest. The rules can be used in pricing strategies, product placement, and various types of cross-selling strategies.

# Exploratory Data Analysis

## Statistics from Jan-2016 through Dez-2020

**130k**
Average Units sold in a Year

**10k**
Average Units sold in a Month

**400**
Average Units sold in day

## General Statistics

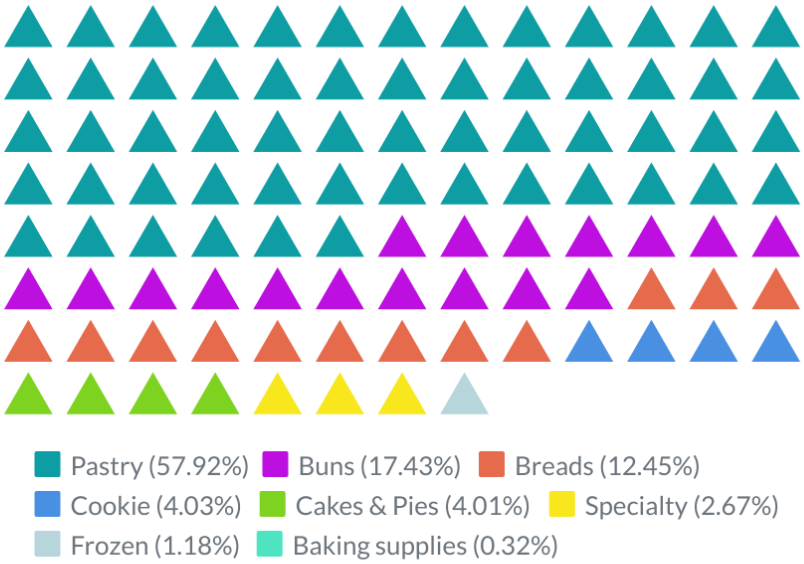
**102**
Average Transactions per day

**3**
Average units per transaction

**3**
Average minutes between transactions

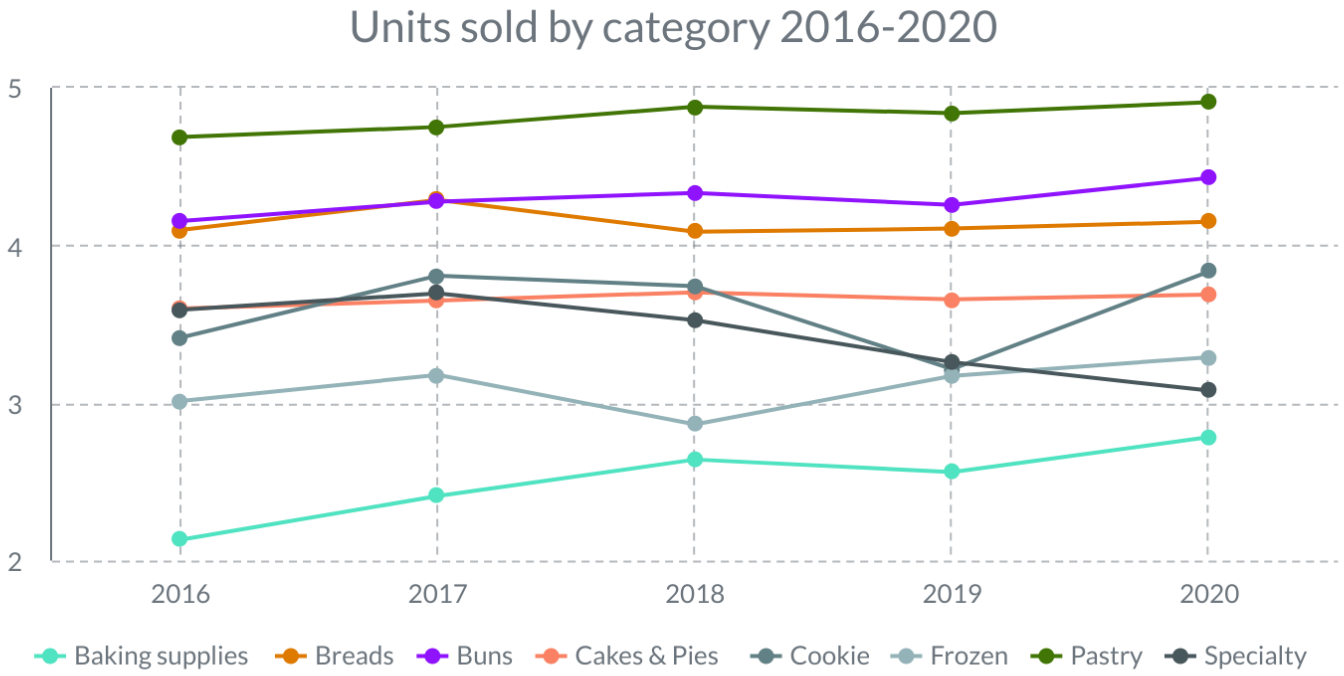
# Categories

Let's take a look at how the categories performed in the past 5 years. For comparability, we will filter out the categories that do not exist in all of the 5 years.

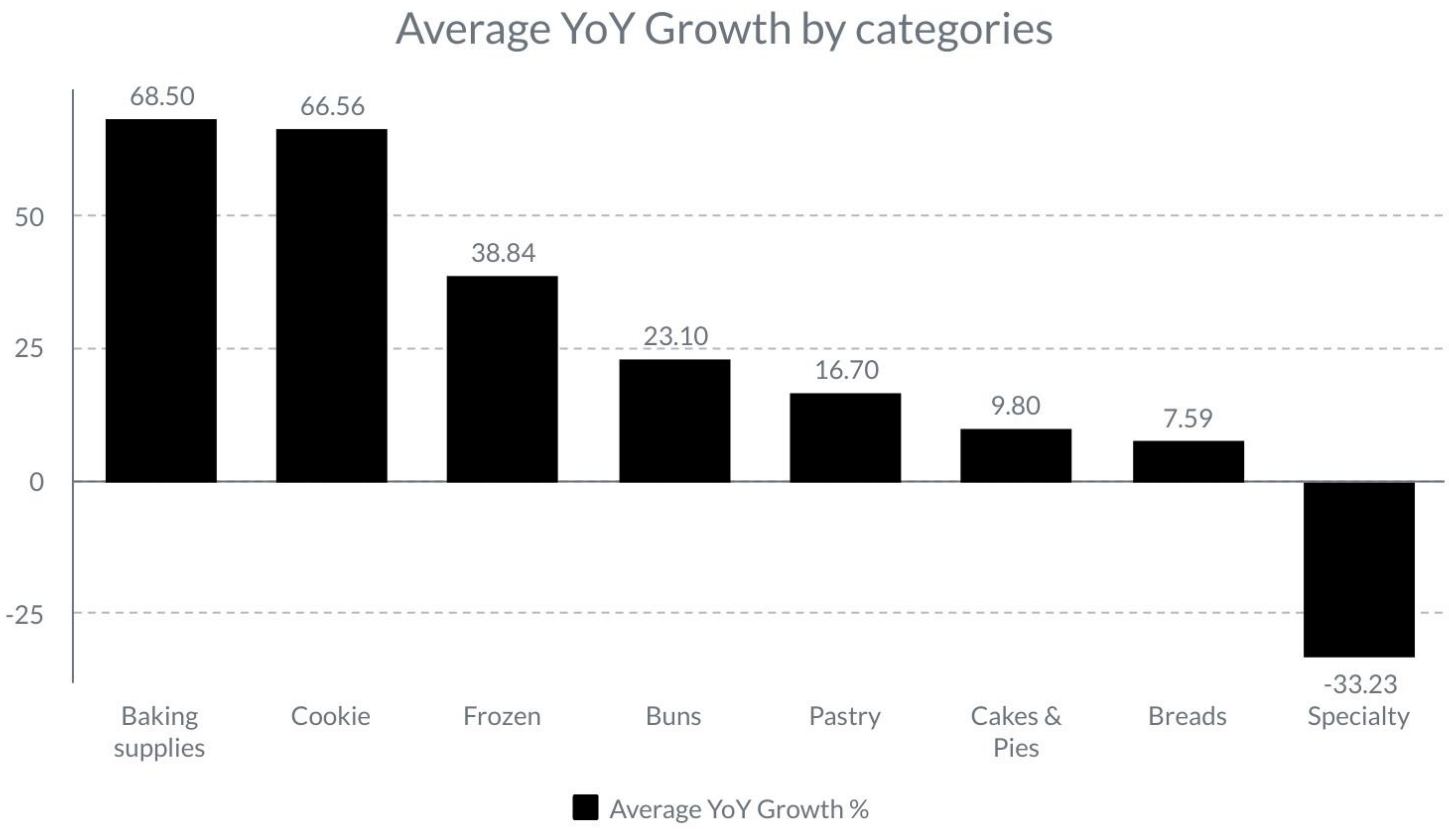



**58** **% of sales are from pastries**

Pastries are by far the most sold items the closest category sold less 200k units.



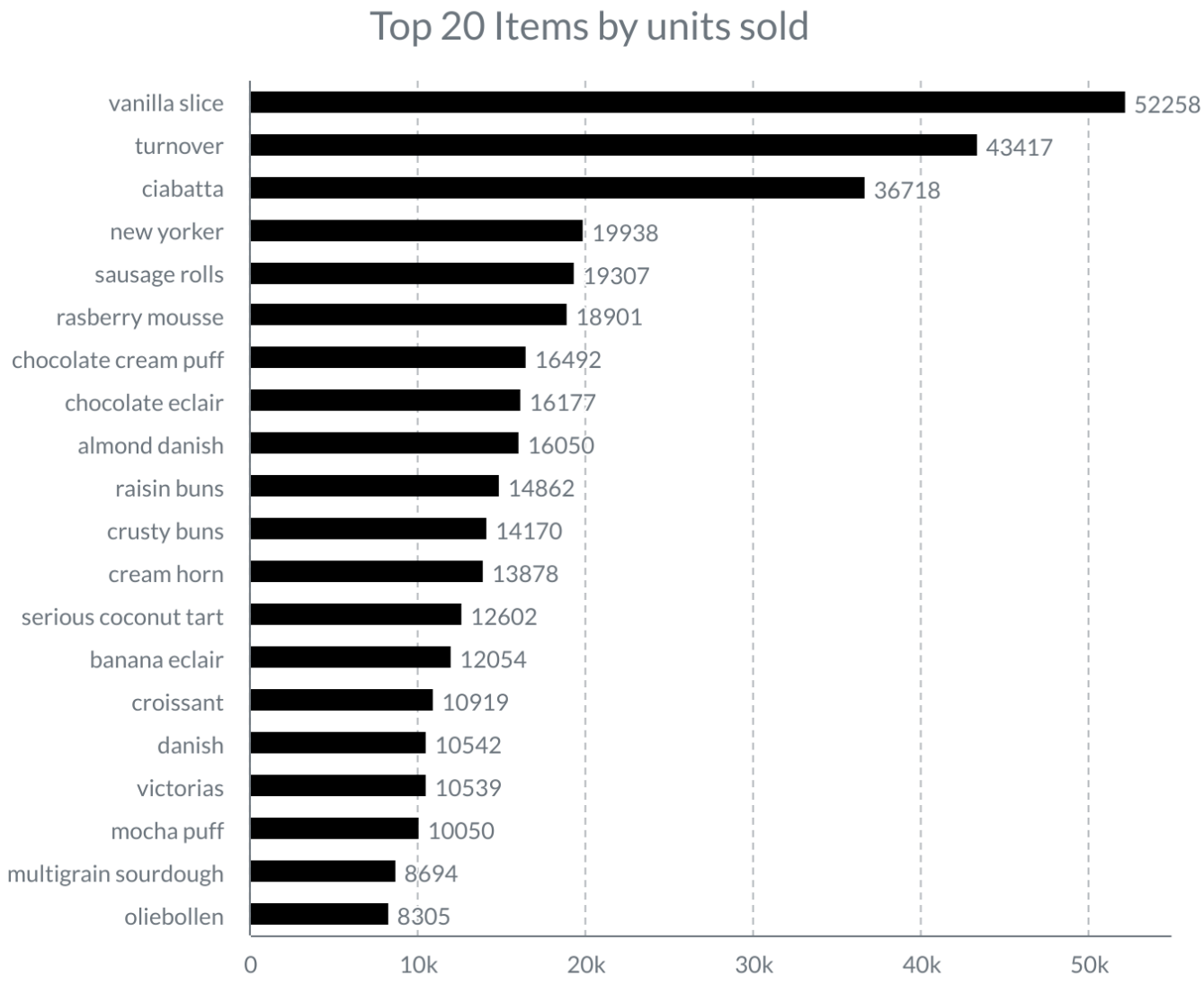

All categories show an increase in units sold for the past 5 years. But what is the fastest-growing category? To figure out that we can calculate the abs diff in units sold year over year and calculate the average of all the years. Lets do that and plot the results.




Speciality category has been seeing a decline in the past 5 years with it being the only category that is not growing.
Interestingly enough our two biggest categories Pastry and Buns have registered an average of 16% and 23% growth respectively.
While baking supply is our smallest category it is registering the fastest growth and might be a signal that that category deserves attention.

# Items

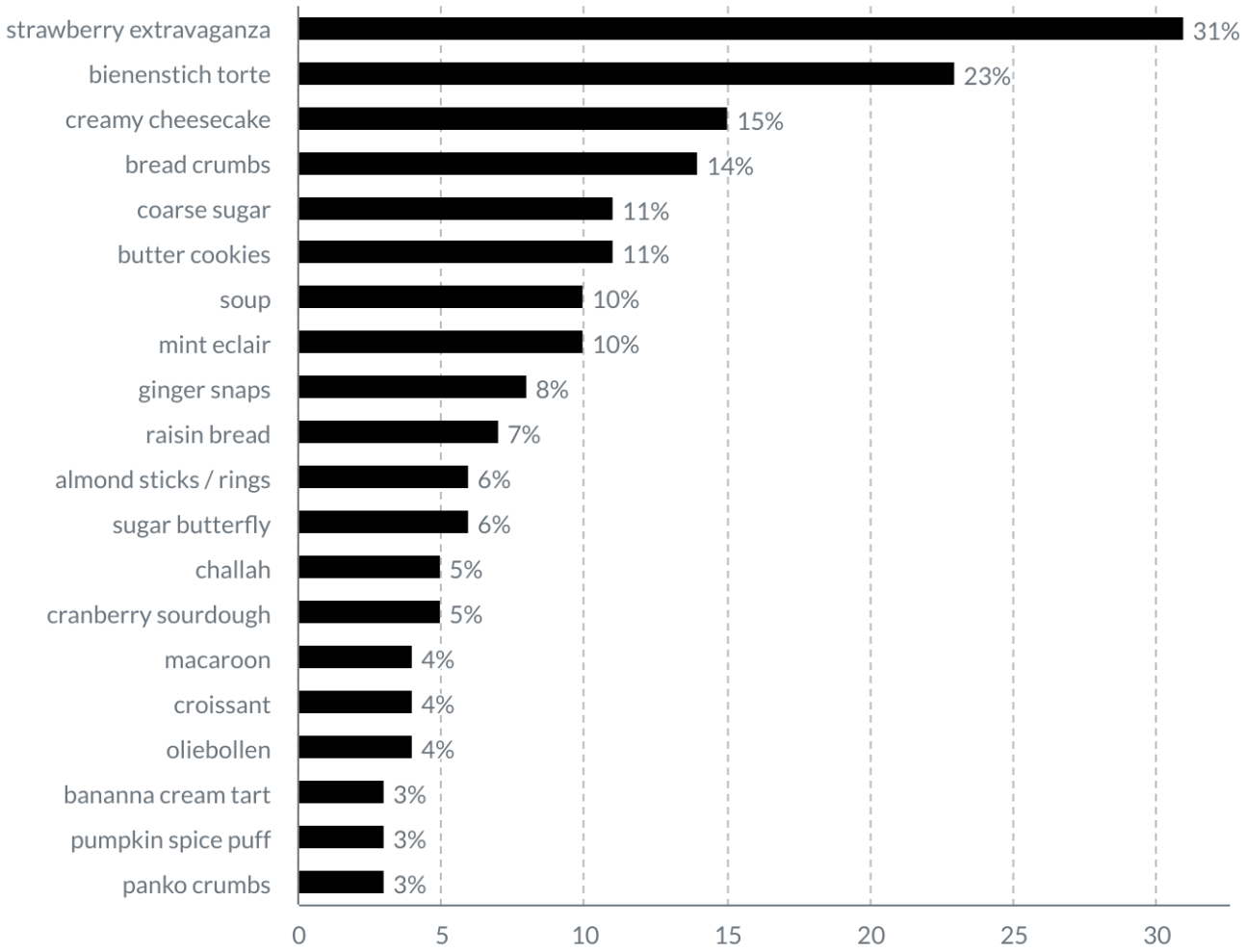
For items, we filtered out the items that do not appear in at least 4 years. That gave us a list of 61 unique items.  As we did for categories let's take a look at what has been the best selling product for the past five years.



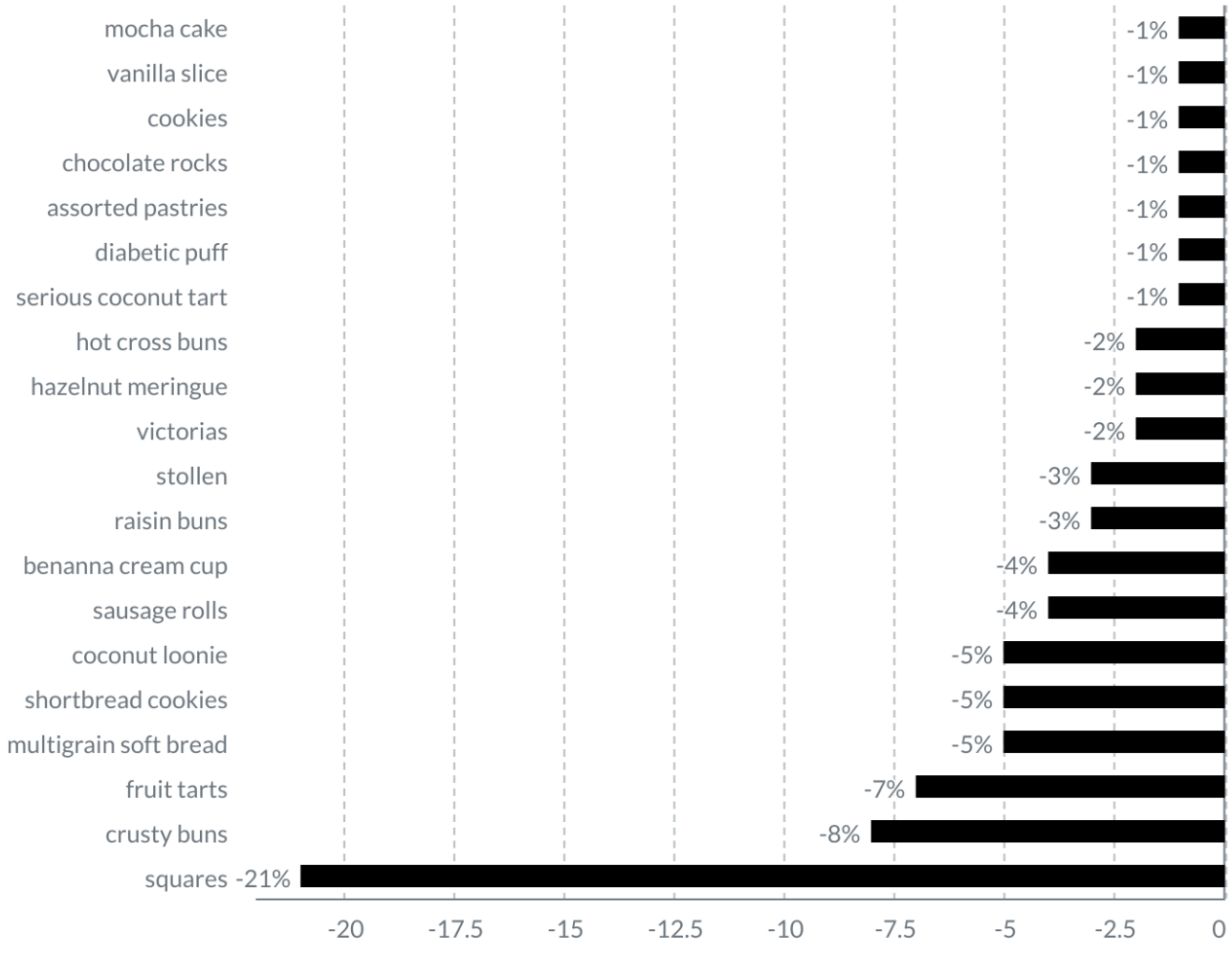

**34** **% of items are responsible for 80% of sales**

**The top 3 items make up for 28% of total units sold.**

## Top 20 Items Average YoY Growth




## Bottom 20 Items Average YoY Growth



# Market Basket Analysis

## What Is market basket analysis?

Market basket analysis is a technique to identify underlying relations between different items. Take an example of a Super Market where customers can buy variety of items. Usually, there is a pattern in what the customers buy. For instance, mothers with babies buy baby products such as milk and diapers. Damsels may buy makeup items whereas bachelors may buy beers and chips etc. In short, transactions involve a pattern. More profit can be generated if the relationship between the items purchased in different transactions can be identified.

For instance, if items A and B are bought together more frequently then several steps can be taken to increase the profit. For example:

1. A and B can be placed together so that when a customer buys one of the products he doesn't have to go far away to buy the other product.
2. People who buy one of the products can be targeted through an advertisement campaign to buy the other.
3. Collective discounts can be offered on these products if the customer buys both of them.
4. Both A and B can be packaged together.

The process of identifying an associations between products is called association rules mining.

## Apriori Algorithm for Association Rule Mining

Different statistical algorithms have been developed to implement association rule mining, and Apriori is one such algorithm. In this project we will implement the apriori algorithim to mine for association rules.
To better understand the outcome of the analysis we must understand the theory of the apriori algorithim.
There are three major components of Apriori algorithm:

- Support
- Confidence
- Lift

We will explain these three concepts with the help of an example.

Suppose we have a record of 1 thousand customer transactions, and we want to find the Support, Confidence, and Lift for two items e.g. burgers and ketchup. Out of one thousand transactions, 100 contain ketchup while 150 contain a burger. Out of 150 transactions where a burger is purchased, 50 transactions contain ketchup as well. Using this data, we want to find the support, confidence, and lift.

**Support**

Support refers to the default popularity of an item and can be calculated by finding number of transactions containing a particular item divided by total number of transactions. Suppose we want to find support for item B. This can be calculated as:

***Support(B) = (Transactions containing (B))/(Total Transactions)***

For instance if out of 1000 transactions, 100 transactions contain Ketchup then the support for item Ketchup can be calculated as:

***Support(Ketchup) = (Transactions containingKetchup)/(Total Transactions) Support(Ketchup) = 100/1000 = 10%***

**Confidence**

Confidence refers to the likelihood that an item B is also bought if item A is bought. It can be calculated by finding the number of transactions where A and B are bought together, divided by total number of transactions where A is bought. Mathematically, it can be represented as:

***Confidence(A→B) = (Transactions containing both (A and B))/(Transactions containing A)***

Coming back to our problem, we had 50 transactions where Burger and Ketchup were bought together. While in 150 transactions, burgers are bought. Then we can find likelihood of buying ketchup when a burger is bought can be represented as confidence of Burger -> Ketchup and can be mathematically written as:

***Confidence(Burger→Ketchup) = (Transactions containing both (Burger and Ketchup))/(Transactions containing A) Confidence(Burger→Ketchup) = 50/150 = 33.3%***

**Lift**

Lift(A -> B) refers to the increase in the ratio of sale of B when A is sold. Lift(A –> B) can be calculated by dividing Confidence(A -> B) divided by Support(B). Mathematically it can be represented as:

***Lift(A→B) = (Confidence (A→B))/(Support (B))***

Coming back to our Burger and Ketchup problem, the Lift(Burger -> Ketchup) can be calculated as:

***Lift(Burger→Ketchup) = (Confidence (Burger→Ketchup))/(Support (Ketchup))***
***Lift(Burger→Ketchup) = 33.3/10 = 3.33***

Lift basically tells us that the likelihood of buying a Burger and Ketchup together is 3.33 times more than the likelihood of just buying the ketchup. A Lift of 1 means there is no association between products A and B. Lift of greater than 1 means products A and B are more likely to be bought together. Finally, Lift of less than 1 refers to the case where two products are unlikely to be bought together.

## How will we implement the Apriori Algorithm?

Our data set is quite large and because the Apriori algorithm tries to extract rules for each possible combination of items. For instance, Lift can be calculated for item 1 and item 2, item 1 and item 3, item 1 and item 4 and then item 2 and item 3, item 2 and item 4 and then combinations of items e.g. item 1, item 2 and item 3; similarly item 1, item2, and item 4, and so on, this process can be extremely slow due to the number of combinations. To speed up the process, we will perform the following steps:

**1.** Set a minimum value for support. This means that we are only interested in finding rules for the items that have certain default existence.
**2.** Extract all the subsets having higher value of support than minimum threshold.
**3.** Select all the rules from the subsets with confidence value higher than minimum threshold.
**4.** Order the rules by descending order of Lift.

# Implementing Apriori Algorithm and mining for association rules

**Step 1: Set  a minimum value for support**

To maximixe the probability of finding associations that are meaningfull we will filter the items to only those that have support greater that 0.01 ~ 10%

**Step 2: Extract all the subsets having higher value of support than minimum threshold**

We have 81 items that meet that criteria here are the top 3 and bottom 3 organized as subset - support:

(vanilla slice) -  0.13
(turnover) -  0.11
(new yorker) -  0.07
---
(mocha puff, vanilla slice) -  0.01
(mocha puff, vanilla slice) -  0.01
(chocolate cream puff, victorias) - 0.01

The full table will be provided as an apendix.

**Step 3: Select all the rules from the subsets with confidence value higher than minimum threshold.**

We have decided to set the confidence threshhold at 5% giving us enough rules to filter on later based on lift. The algorithim produces 38 rules,  that contain 19 item sets.

**Step 4: Order the rules by descending order of Lift.**

By order of lift the top associations were :
Banana Eclair  & Chocolate Eclair ----------lift : 4.26
Victorias & Chocolate Cream Puff ----------lift:  4.11
Cream Horn Chocolate Cream Puff --------lift:  4.09

The algorithim produced 3 item sets with very strong positive relationships & 4 with very strong relationships.

As we discussed above we know that lift indicates whether there is a relationship between A and B, or whether the two items are occuring together in the same orders simply by chance (ie: at random). In the case of the 19 item sets we can conclude that most items do occur together by chance and there is clear customer preference.

# Results & Conclustion

The algortithim produced results that indicate that there are underlying relationships between items.
Intresting enough we see that the strongest rule was between items of the same type but of different flavoures(Bannana and Chocolate Eclair) which is expected.
We can also note that Chocolate Cream Puffs are often bought together with Victorias, Cream Horns or Rasberry Mousse.

As a logical next step once item pairs have been identified as having positive relationship, recommendations can be made to customers in order to increase sales. And hopefully, along the way, also introduce customers to items they never would have tried before or even imagined existed!