MSBX 5405 Project

Coffee Shop Data Analytics Report

Team 18:

Linyi Yao

Zhen Wang

Eunhye Beissinger

Huaiqian Yan

Xiaosong Fan

Project Scenario

Theme

For the purpose of this project, we want to find and develop a relational dataset that includes diverse data types and fields for research purposes. We aim to find a dataset that has sufficient numerical values to conduct statistical analysis and categorical fields for distinguishment and logical analysis. The dataset also needs to have a significant relationship between different tables and records so we may study the correlation among them and join them under SQL environment when needed. With these criteria in mind, we developed the following dataset that fulfills our requirement for this project.

This dataset contains representative retail data from a fictional coffee chain. They were created with three retail store locations in New York city. We are aiming to develop a sales data database to better understand the business and provide any additional feedback for its future development and profit growth. Once we have developed the database, it will be more accessible for us to analyze these historical data and find the weaknesses and strengths of the operation. Furthermore, the database can be applied to the retail data of other industries as long as it succeeds here. With the development and optimization of the database, it will create enormous value and infinite possibilities in the future. That's why we are interested in designing a database for retail data. This fictional coffee shop dataset is clean and neat, so this is a good start for us.

Our report and analysis will focus on client retention, profit & loss by different locations, and operation maximization by evaluating the performance of different products provided to the customers. Throughout our analysis, you will discover the theme of our team exploring the patterns of the data, thus finding how the business operates by different metrics and perspectives. Towards the end, we have developed a greater understanding of the business in a whole picture, and conclude to findings of how the business may optimize its operations for high efficiency.

Tables and Relationships

There are eleven tables in the database which are *receipt, date, customer, generations, product, product detail, pastry, sales outlet, sales target, staff, staff position*.

Receipt records the relevant information of each transaction. It contains index (a unique identifier of each record), transaction id, transaction date, transaction time, sales outlet id, staff id, customer id, the way of placing the order (instore or not), order, line-item id, product id, quantity of the sold items, total dollar amount of the sold items, unit price, and if-a-promoted item.

Dates summarizes the information of transaction dates. It contains date ID (the numeric format of the transaction date), transaction date, week ID (the numeric format of the week when transaction happened), week name, month ID (the numeric format of the month when transaction happened), month name, quarter ID (the numeric format of the quarter when transaction happened), quarter name, and year ID (the numeric format of the year when transaction happened).

Customer stores the personal information of customers. It contains customer id (a unique identifier of each customer), home store (the store where they registered as a loyal member), customer name, customer email, the date when the customer became a loyal member, loyalty card number, birthdate, gender, and birth year.

Generations is a subset of *customer* table to classify the customers to different generation according to their birth year. It contains birth year and generations.

Product stores the general information of the product sold in the coffee shops. It contains product id (a unique identifier of the products for sale), product groups (the first level of classification for the products), product categories (the second level of classification for the products), product types (the third level of classification for the products), product name, and product description.

Product detail stores the detailed information of the products. It contains product id, unit of measure, whole-sale price, retail price, if-a-taxed item, if-a-promoted item, and if-a-new product.

Pastry stores the information of only for the pastry-kind products. It contains product id (the amount offered each day in each shop).

Sales outlet stores the detailed information of each sales outlet. It contains sales outlet id (a unique identifier of each sales outlet), sales outlet type (warehouse or retail), area of each store, store addresses, store city, store state, store phone numbers, store postal code, store longitude, store latitude, manager id, and the neighborhood area of the store.

Sales target contains the target sales number of different products. It includes the period of the goal, the goal of beans, the goal of beverage, the goal of food, the goal of merchandise, and the goal of all products.

Staff stores the personal information of the staff. It contains staff id (a unique identifier of each employee), first name, last name, start date of their work, location, and position id (a unique identifier of each position).

Staff position stores the position information. It contains position id and position names.

Relevant Columns and Fields

To better utilize the dataset, we did some transformation of the original dataset.

In *receipt*, we found the transaction id is not unique, so we created a new column *index* to be the primary key of this table. We converted the data type of *transaction time* from “text” to “time.”

In *date,* we converted the data type of *Year_ID* from “integer” to “year.”

In *customer*, we got rid of “-” in column *loyalty_card_number*.

We split *product* into two tables *product* and *product detail*. It is clearer and more logical for the database.

We only kept the column *product_id* and *start_of_day* in *pastry* because other column can be found in other tables. It can save space and optimize the operation of the database.

Database Normalization

Brief summary

The raw data set includes a total 9 tables (*see ER diagram raw database in Appendix A*). All of them are considered like 1NF tables and we don’t find a single cell with multiple values nor duplicate rows. We consider the ‘receipt’ table as a fact table with most combined information from customers, staff, stores and products. After normalization, our ‘coffee shop’ database might be more like a star schema (*see NF2 ER diagram in Appendix A*). In our NF2 version database (total 11 tables), we keep single primary key for each table. Even though, we’ve done higher normalization for some of the tables (*see NF2.5 ER diagram in Appendix B*) by removing calculated fields and making detailed information tables (keep independency and less redundancy), we decide to the NF2 version as a better choice of balancing between data storage capacity and analysis convention.*(note: the main reason we accept the lower NF version is we have relatively small data base, so keeping redundancies and independencies is acceptable here).*

Detailed normalization steps and examples

Firstly, we have checked mostly like identity key(s) (e.g. the columns named with ‘id’) with its unique values for each table. We find that only ‘receipt’ table, there are partial keys but no single identity key. Since we prefer not to use composite primary key for analysis, we decide to add a new field named ‘index’ as its primary key.

We have removed some suspicious records in ‘receipt’ table as we find there are transaction records from 2 customer don’t have matching customer id in ‘customer’ table. It not only causes errors when we try to insert data with foreign key constraint, but also providing no meaningful information for our business analysis later. *(Note: regarding to data integrity, instead of removing the customer records from child table ‘receipt’, the alternative way is to insert 2 (faked) customer information in the parent table ‘customer’ before insert data into ‘receipt’ table)*

Here we provided 5 most normalized tables as examples for our process:

1. 'Product' table

The original 'product' table (Figure 1) includes both product unique characteristics (such as combination of product group, type and description etc.) and 6 common properties (such as measure unit, tax type, whole sale price, whether it is new product etc.). In order to minimize the repeated rows of the share properties among different product, we decide to split the orginal 'product' table into two tables as (new) 'product' and 'product detail'. Product _id performs as primary key for both tables and serves as a bridge to match product with its common properties.

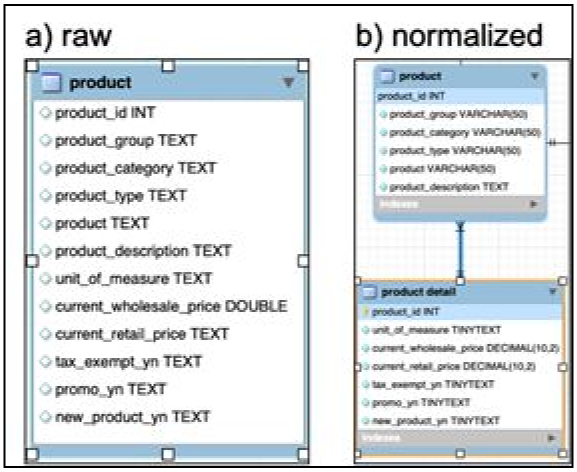


Figure 1. table 'product'.

2. 'staff table'

The old table is considered as NF2 (Figure 2. a) raw). It has unique identity key as 'staff_id' and all non-key attributes are depend on it, however both 'location' and 'position' fields contains repeated values. There multiple staffs have same positions (e.g. Coffee Wrangler) and one location (e.g. store or state) can have multiple staffs. In our analysis we decide to split out the position information to its own table as 'staff position' with 'position_id' as its primary key. For 'location' field, it has more unique values (or less repeated values) compare that of 'position' field. If we split out the location information into another new table, it finally requires more data storage capacity compared with the new table created for the 'position' field. Therefore, for the sake of balancing between database simplicity and normalization, we keep the 'location' field in original table for this report.

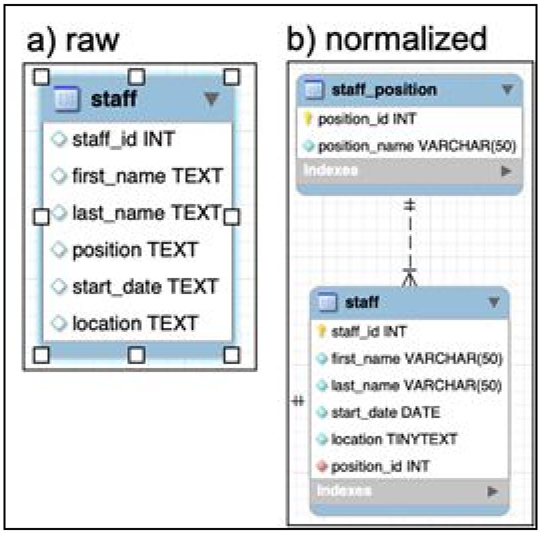


Figure 2. table 'staff'

3. 'date table'

In 'dates' table include all information about transection date. It store the date in week, month, quarter and year as separated columns. In addition, each of them is give in two format. For an example, week information are repeated in 'Week_ID' column (integer week number value, e.g. '4') and 'Week_Desc' column (string description of week, e.g, 'week 4'). At first, we think the whole table about date information is not necessary, since we can easily use build-in time functions (e.g. WEEK, DAYNAME, MONTH, and YEAR) for transaction date stored in 'receipt' table. But actually if we consider a large business data set in real, when we want to generate reports based on time bins (e.g. quarters), it may be much easier and faster to access information use join than using function to 'calculate' every single time. Finally, as a compromise, we decide to keep the new 'date' table with only one data type (integer) of each date information.

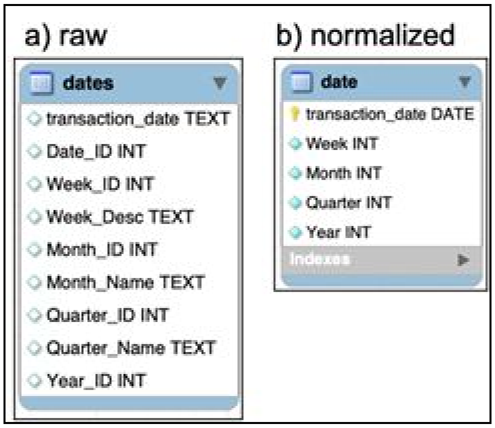
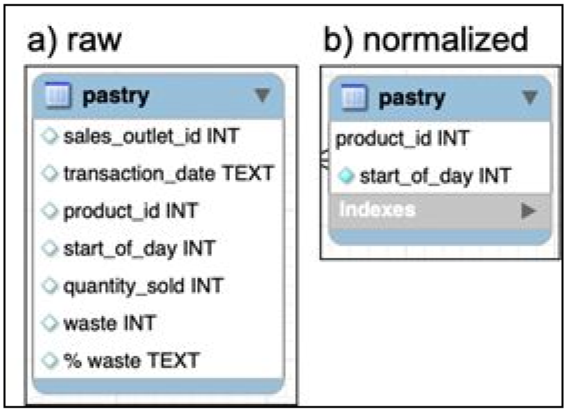


Figure 3. table 'date'

4. ‘pastry table’

Pastry table is the one we made the most changes. The purpose of this table is used to show daily waste percentage of different type food (pastry) sold for each store. There are only 5 unique kind of pastries provided among different stores on different days. The start of day quantity for each pastry from the supplier to store is constant. Therefore, we find a lot redundancy for each column in the ‘old’ pastry table. There are 3 calculated fields as ‘quantity_sold’ (daily total quantity), ‘waste’, and ‘%waste’. The daily pastry sold total quantity for each store is calculated from order details in ‘receipt’ table. The waste percentage is calculated as total sold quantity divided by start of day quantity.

After checking and computing, we find the critical information from pastry table is just unique product id with corresponding start of day quantity for 5 kind of pastries. The product id is used to link both ‘product’ and ‘receipt’ tables for product details and order details respectively. Comparing with convention for keeping redundancy and calculation using joining tables for this pastry daily waste, we decide to shrink down the table to only two essential fields as our better choice in this case.



a) raw

pastry

◇ sales_outlet_id INT
◇ transaction_date TEXT
◇ product_id INT
◇ start_of_day INT
◇ quantity_sold INT
◇ waste INT
◇ % waste TEXT

b) normalized

pastry

product_id INT
◆ start_of_day INT

Figure 4. table ‘pastry’

5. ‘store_outlet table’

It summarizes the store size and location information. The geographic data such state, city, postcode, and neighborhood all have redundancies. We’ve tried to take them into 3 separate tables and do normalization from 2NF to 2.5NF for ‘store_outlet’ table *(shown in ER diagram , Appendix B)*. But through with our analysis, we think the normalization doesn’t provide benefit regarding to the relatively small data set (only 9 stores) but make our queries unnecessarily complicated. Maybe such normalization is good for a larger dataset with much more stores in different locations (e.g. global chain business). At the end, we go for denormalization and keep 2NF here.

Business Questions

Product (6 questions) :

- Which products with order quantities greater than 100?
- How many coffee categories does the coffee shop have?
- Which product category is purchased the most number of times?
- What is the best selling product for each location?
- What is the best selling product overall?
- What is the top selling coffee beans? provide the product name, product type and description.

Sales (7 questions):

- What is the daily sales and profit?
- Which store has the most number of customers?
- Which store has the most number of sales before 7 am on average?
- Monthly total sale from all in-store purchases in 2019.
- What days of the week have the most sales for store No.3?
- Difference between total monthly sale and monthly sale goal for each store?
- Which staff's total sales is higher than the average total sales?

Customer (3 questions):

- What's the most common generation of customers?
- Who is the first customer in the record?
- How many times have customers visited on their birthday?

Staff (4 questions):

- Which staff are in the receipt table, what is their name and position?
- Which staff served the highest gross number of orders?
- Which staff has served the highest number of customers?
- What is the total value of sales by each staff?

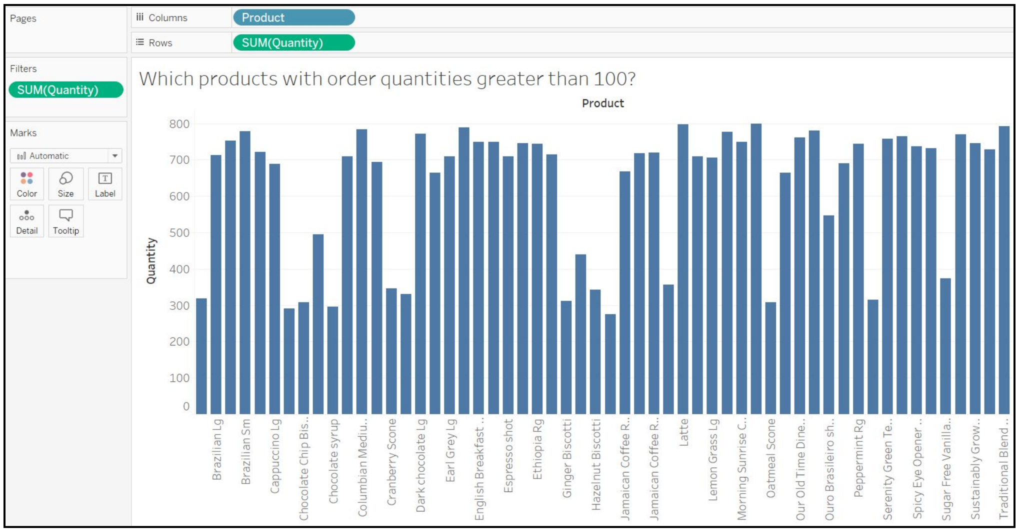
Pastry (5 questions):

- How many pastries are sold per day?
- What is the most popular pastry item and what is the least popular item?
- What is the best selling pastry for each location?
- For each pastry, what's the number of different customer gender purchased?
- What is the best selling day over the week for each type of pastries? What is the total quantity sold on that day?

Data Visualization

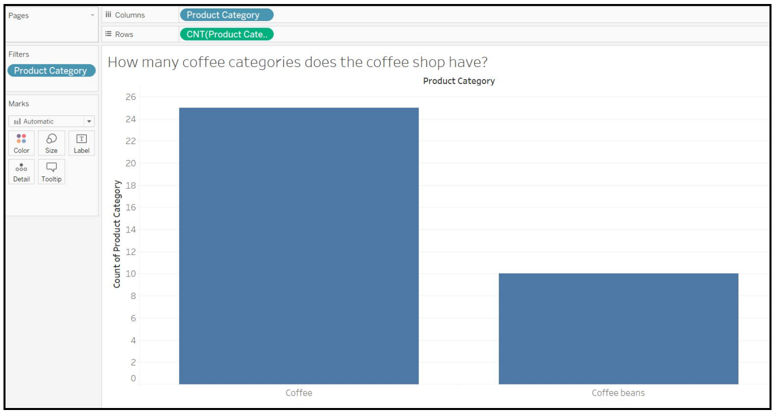
Product

- Which products with order quantities greater than 100?



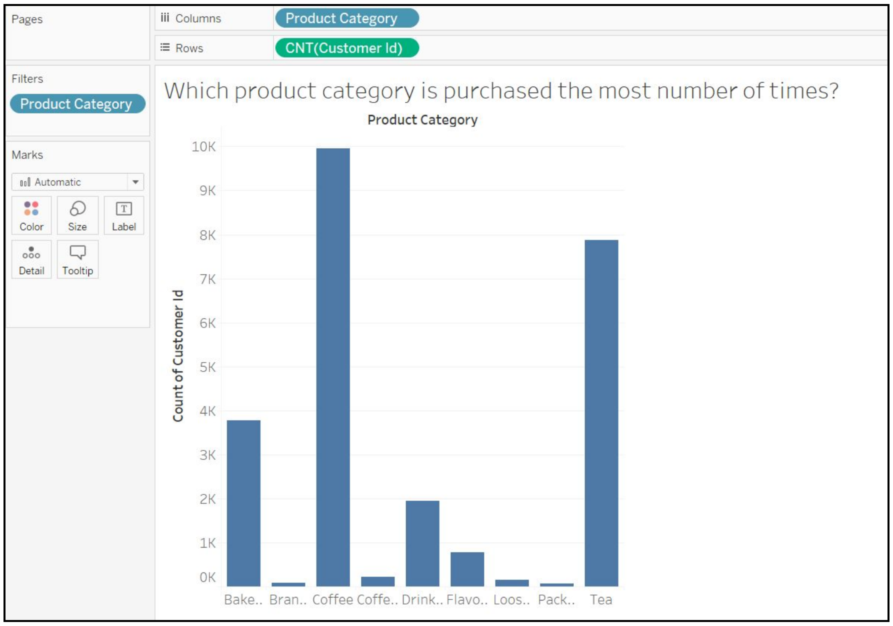
This bar chart entails all the products that have more than 100 sold in quantities. As observed, most products are coffee related products, coffee drinks and coffee beans.

- How many coffee categories does the coffee shop have?



This chart shows that overall, the shop sells both coffee and coffee beans. In general, coffee is sold more popular at the store.

- Which product category is purchased the most number of times?



Columns: Product Category

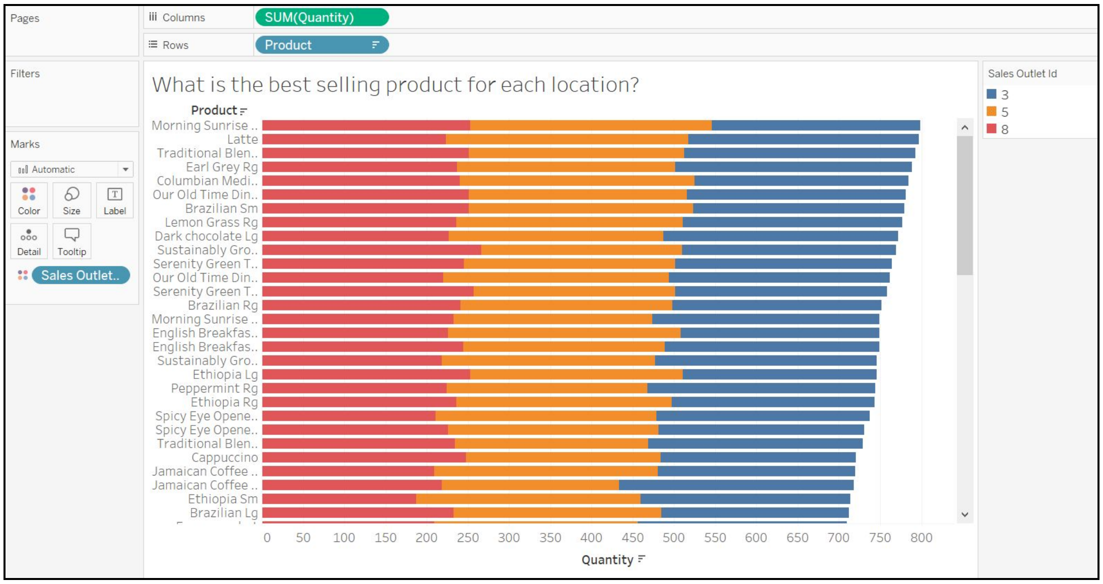
Rows: CNT(Customer Id)

Filters: Product Category

Marks: Automatic

In this table, it can be observed that coffee and tea are the most popular categories among the customers, followed by bakery items (pastry etc.)

- What is the best selling product for each location?



Columns: SUM(Quantity)

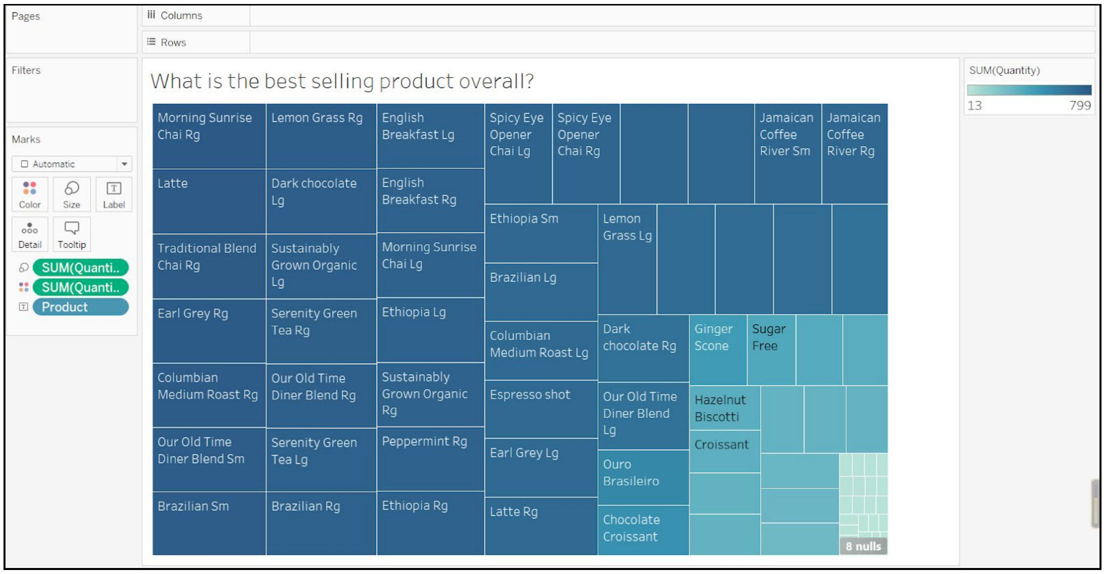
Rows: Product

Marks: Automatic

Sales Outlet..

This table demonstrates how each product is sold in different locations. As there are many items sold by the store, the numerical values can be more easily observed through MySQL.

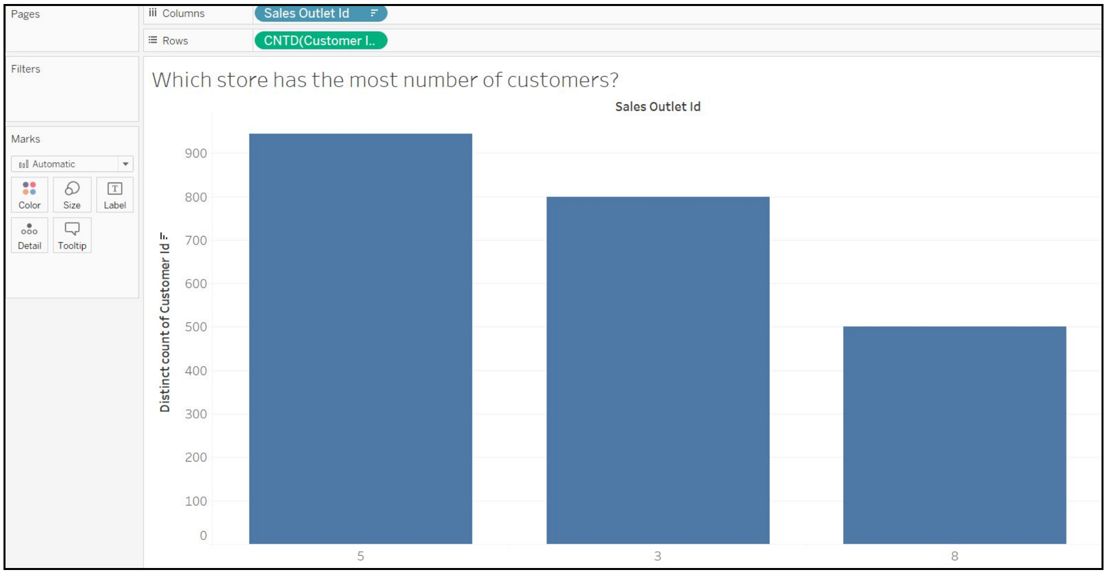
- What is the best selling product overall?



This visualization clearly demonstrates the sales performance of all products. It can be shown that Chait tea products, Latte, and other coffee beverages have the highest sales overall.

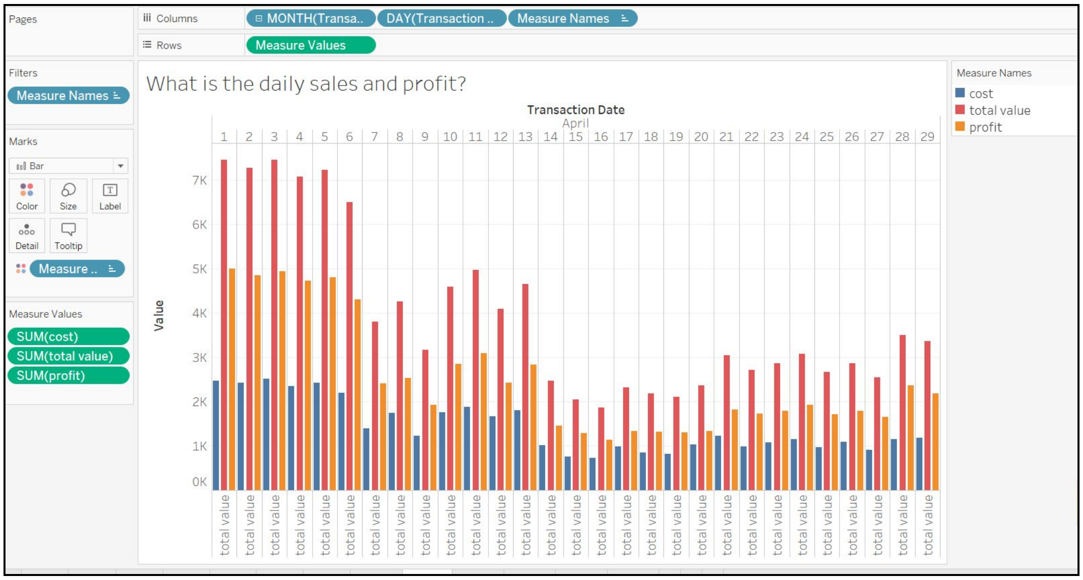
Sales

- Which store has the most number of customers?



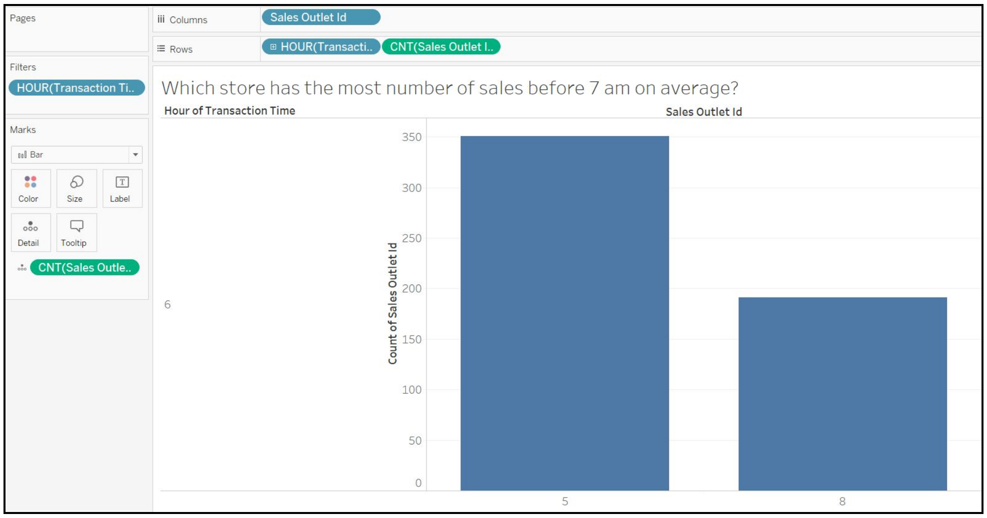
This table shows that store 5 has the most number of customers, followed by store 3 and 8. This result will be further discussed in the question evaluating sales per square foot.

- What is the daily sales and profit?



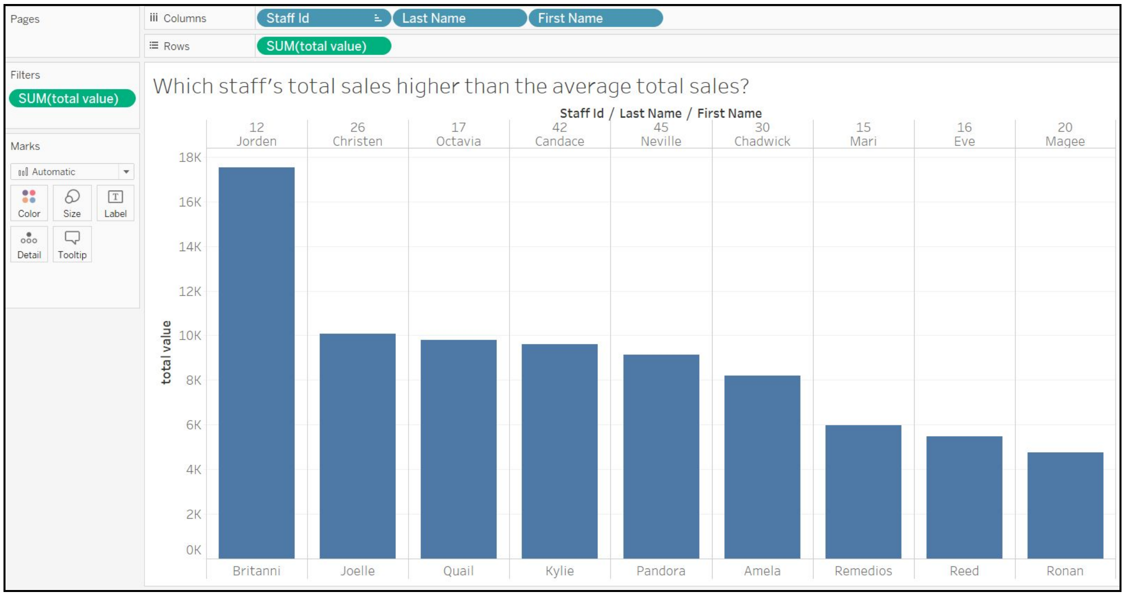
This table shows the overall Profit analysis for each day. It can be observed that on 4/1 - 4/6, the store had the highest sales, which also had the highest profit. 4/14 to 4/27, the stores generally have lower sales of the month.

- Which store has the most number of sales before 7am on average?



This table is to evaluate and demonstrate stores that start the earliest among all the stores. It can be observed that store 5 has the most number of orders before 7 am, followed by store 8.

- Which staff’s total sales is higher than the average total sales?



Columns: Staff Id, Last Name, First Name

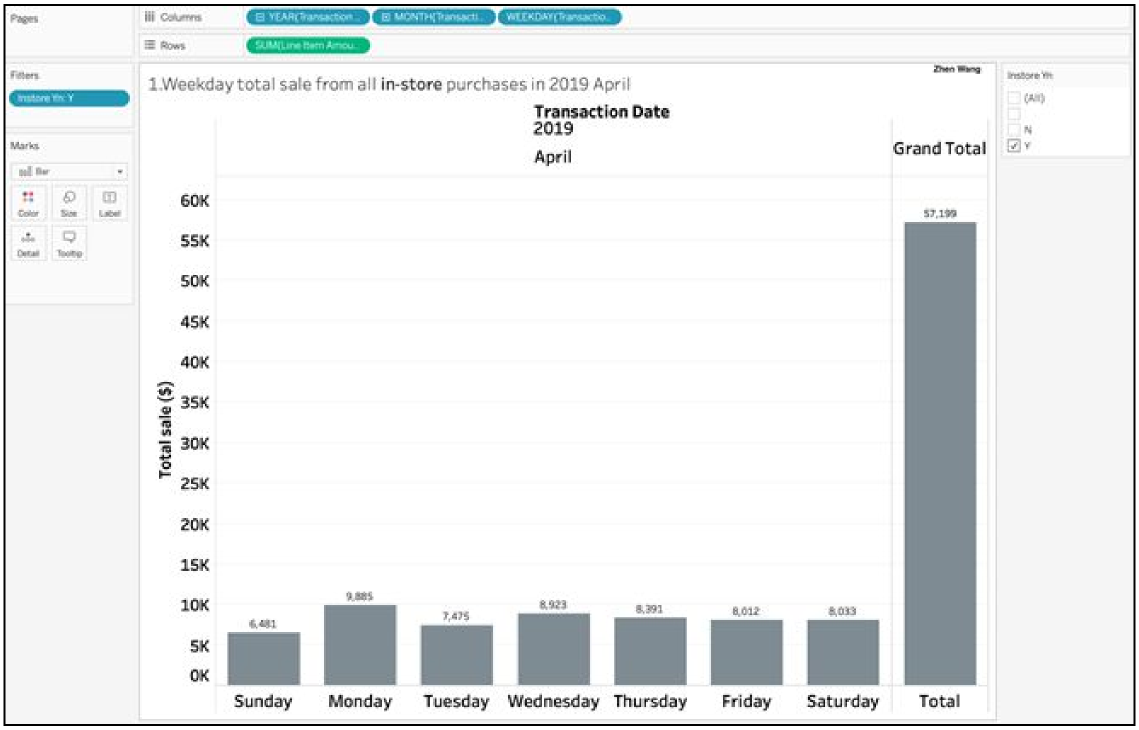
Rows: SUM(total value)

Filters: SUM(total value)

Marks: Automatic

This chart only shows the total value of sales by each customer, and only the ones with higher than average sales. It can be observed that Britanni has the highest sales, which confirms the visualization from the question “Which staff has the highest number of order?”. The highest sales staff is Britanni, followed by Joelle and Quail.

- What is the best selling day over the week for each type of pastries? What is the total quantity sold on that day?



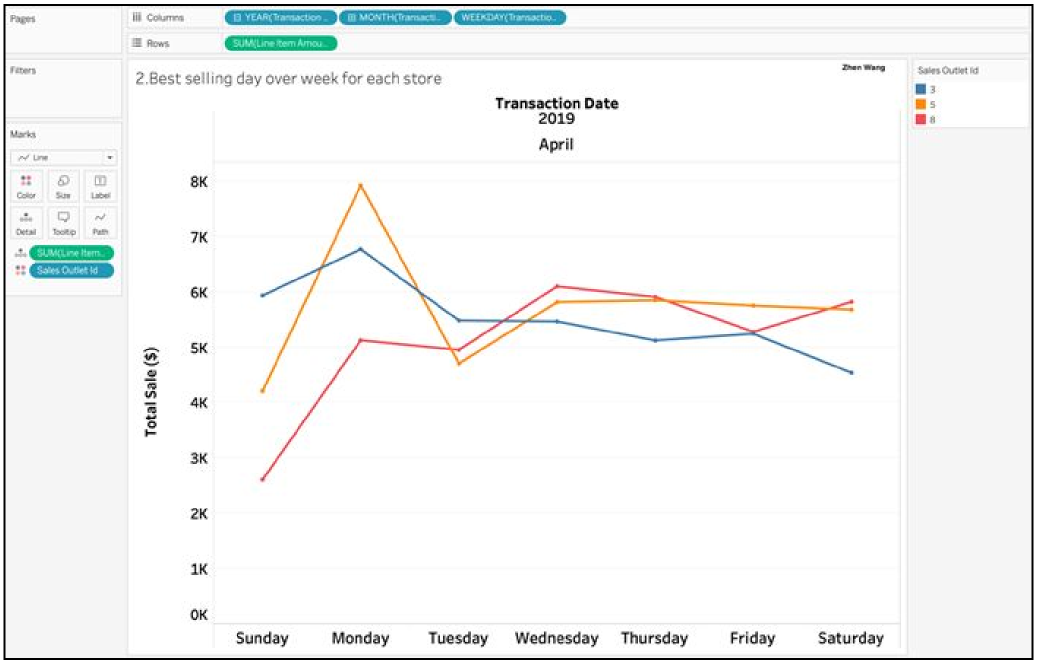
Columns: YEAR(Transaction... MONTH(Transacti... WEEKDAY(Transactio...

Rows: SUM(Line Item Amou...

Filters: Instore Yn: Y

Instore Yn: □ (All) □ □ N ☑ Y

The bar plot is used to show in-store total sale over weekdays, with grant total in the end. The monthly sale results is based on April 2019. We find the best-selling day over all is Monday. The plot uses 3 three discreate dimensions (year, month, weekday) with 1 continuous measures (‘total line amount’ =sum of total sale) as column and row respectively. The filter condition is used 1 discrete dimension as ‘instore Yn’ = ‘Y’ for ‘in-store’ only. Only one table ‘receipt’ is involved and no joins applied here.



Columns: YEAR(Transaction... MONTH(Transacti... WEEKDAY(Transactio...

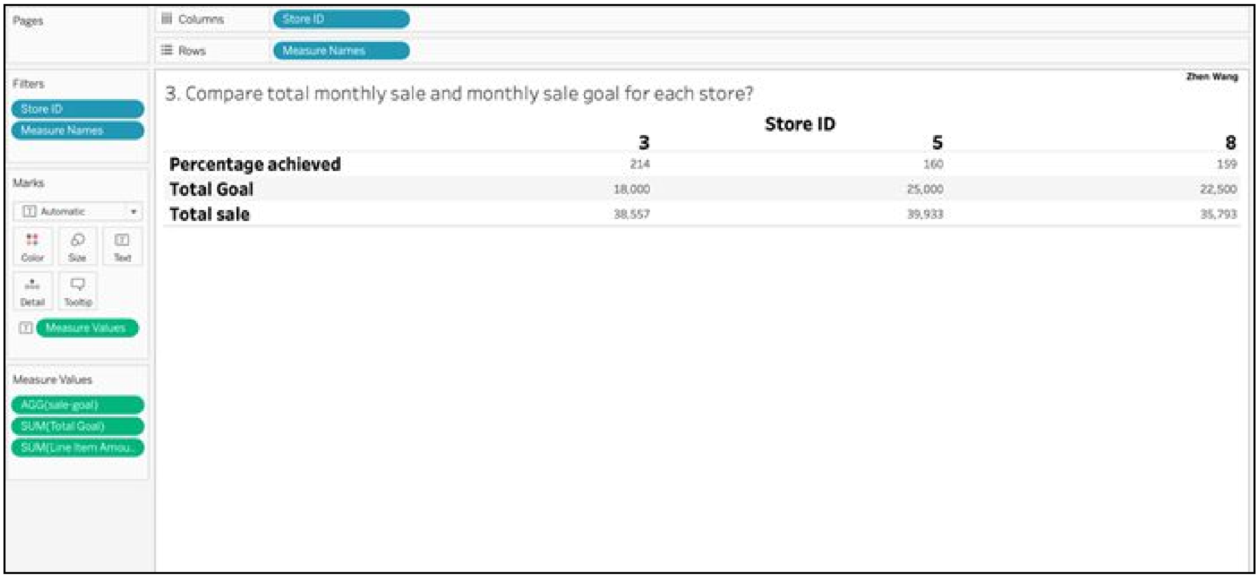
Rows: SUM(Line Item Amou...

Marks: Line; SUM(Line Item... ; Sales Outlet Id

Sales Outlet Id: 🟦 3 🟧 5 🟥 8

The line plot is used to find total selling for each store over weekdays. All information are included in ‘receipt’ table and no joins is required here. There are only 3 stores (id = 3,5,8) with available order details. We find the best-selling day over all is Monday for all stores and store No.5 yields the highest total sales on Monday.

The plot uses 3 three discreate dimensions (year, month, weekday) with 1 continuous measures (‘total line amount’ =sum of total sale) as column and row respectively. We use 1 continuous measure (‘total line amount’) in marker to combine 3 plots into one plot as below. 1 discrete dimension (‘sales_outlet_id’) is used as color-code marker for distinguishing each store.

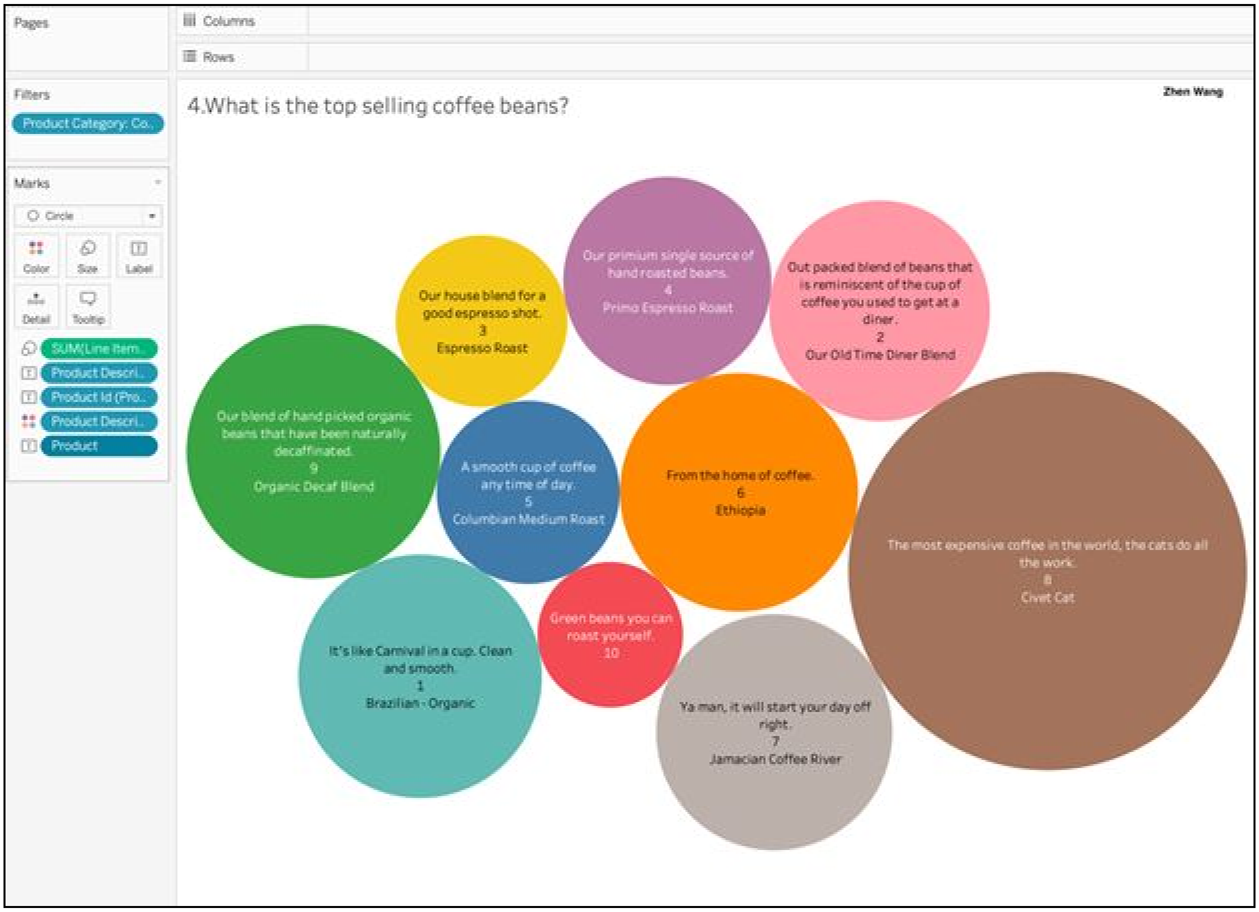


3. Compare total monthly sale and monthly sale goal for each store?

	Store ID		
	3	5	8
Percentage achieved	214	160	159
Total Goal	18,000	25,000	22,500
Total sale	38,557	39,933	35,793

We use table look to compare total monthly sale and total sale goal for each store. Also, we define a calculated measure as ‘percentage achieved’, in order to have a straightforward view of results. All 3 stores have achieved sale goal over 100% with Store No.3 yields the best performance (214%).

The table use 1 dimension (‘Store ID’) and 1 discreate measure(‘measure name’: combination of 3 continuous measures as ‘Total goals’, ‘Total sale’ and ‘Percentage achieved’) as row and columns. The 3 continuous measures are all summation values from group by, suing the filter conditions(‘Store ID’) and (‘measure name’). The results are based two joined (inner join) table as ‘receipt’ and ‘sale target’ through store id.

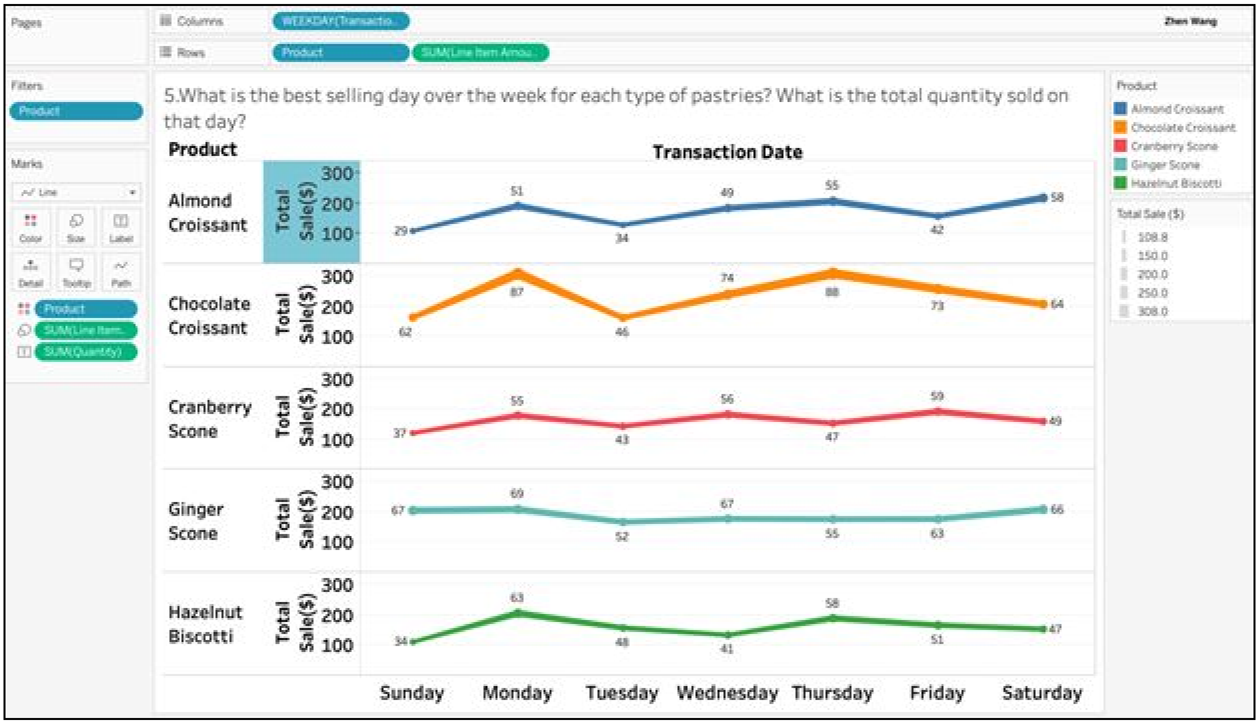


4.What is the top selling coffee beans?

Pages
Columns
Rows
Filters
Product Category: Co..
Marks
Circle
Color
Size
Label
Detail
Tooltip
SUM(Line Item..
Product Descri..
Product Id (Pro..
Product Descri..
Product
Zhen Wang
Our house blend for a good espresso shot.
3
Espresso Roast
Our primium single source of hand roasted beans.
4
Primo Espresso Roast
Out packed blend of beans that is reminiscent of the cup of coffee you used to get at a diner.
2
Our Old Time Diner Blend
Our blend of hand picked organic beans that have been naturally decaffinated.
9
Organic Decaf Blend
A smooth cup of coffee any time of day.
5
Columbian Medium Roast
From the home of coffee.
6
Ethiopia
The most expensive coffee in the world, the cats do all the work.
8
Civet Cat
It's like Carnival in a cup. Clean and smooth.
1
Brazilian - Organic
Green beans you can roast yourself.
10
Ya man, it will start your day off right.
7
Jamacian Coffee River

This bubble chart shows the best-selling whole coffee bean along with its product name, id and description. The results are based on joined information (inner join) from table 'receipt' and 'product'.

We see the larges bubble (brown color) is 'Civet Cat', which yields the highest sales. The chart use 1 discreate dimension ('Product Categroy' ='Coffee Bean') to find only whole coffee bean products. 1 continuous measures ('total line amount' =sum of total sale) is used for marker size. 3 product discreate dimensions are shown as the bubble labels for product details. The color code is based on different product description for a pleasant view.

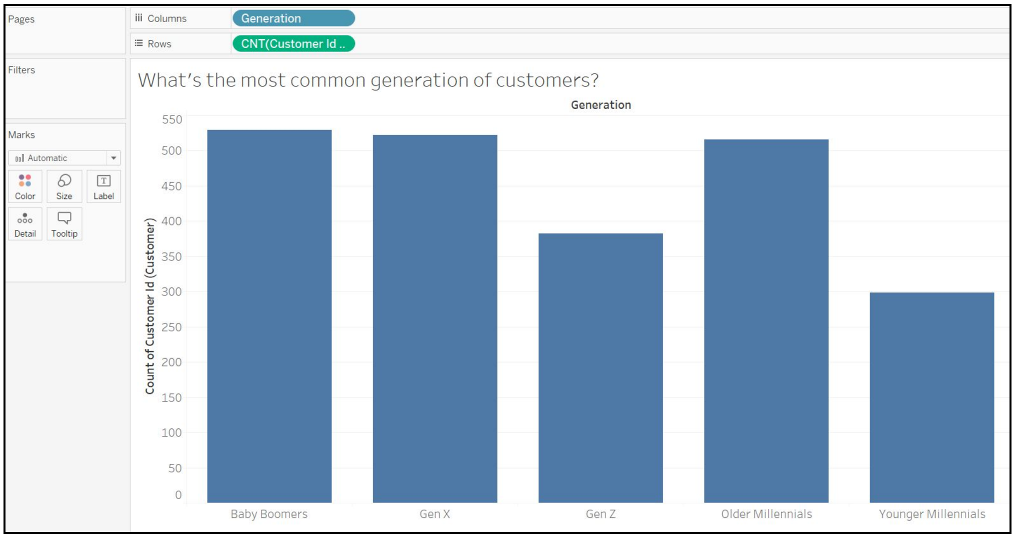

5.What is the best selling day over the week for each type of pastries? What is the total quantity sold on that day?

The line plots are used to find total quantity sold for 5 types of pastry over the weekdays. The information are based on (inner) join 3 tables as ‘receipt’, ‘pastry’ and ‘product’ table with group by pastry types. We find the most sold quantity for all pastries are most on Monday, except Cololate Crossant (Thursday).

The plots use 1 discreate dimensions (weekday) as X axis, and Y axis combines 1 continuous measures (‘total line amount’ =sum of total sale) and 1 discrete dimension(‘product’). 1 discrete dimension (‘product’) is color-code marker for distinguishing result of each pastry. 1 continuous measure (sum of ‘quantity’) is used to label data points for a straight view.

Customer

- What’s the most common generation of customers?



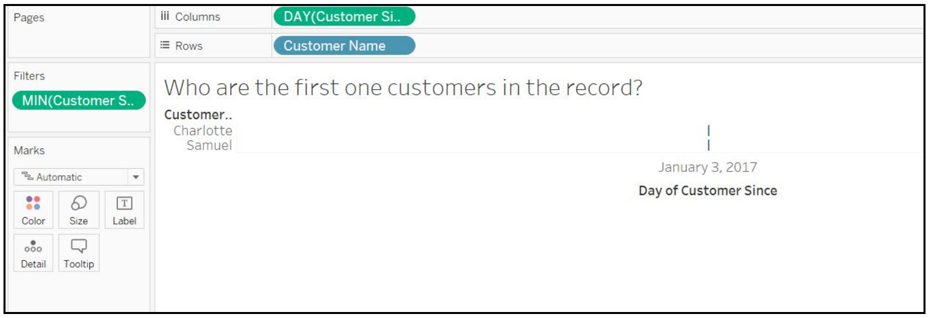
Columns: Generation

Rows: CNT(Customer Id ..

What’s the most common generation of customers?

As observed in this chart, the most common generation of customers are baby boomers, Gen X and Older Millennials. This indicates that the coffee shop has potential to grow among the younger age population (Gen Z, young millennials).

- Who is the first customer in the record?



Columns: DAY(Customer Si..

Rows: Customer Name

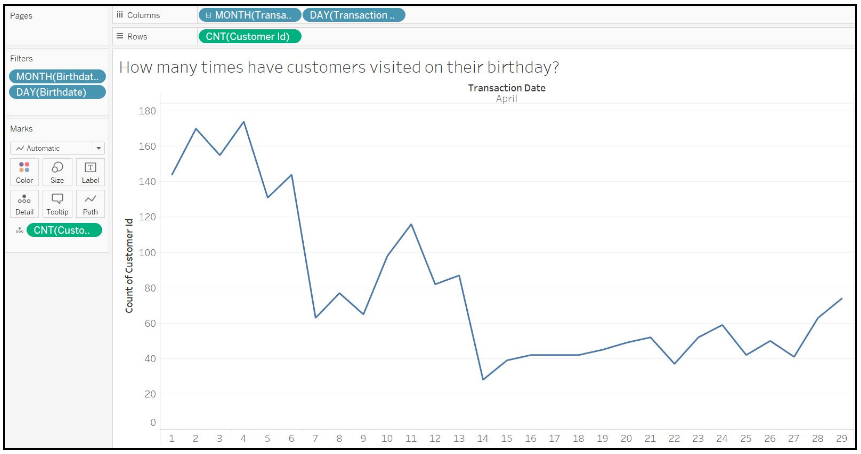
Filters: MIN(Customer S..

Who are the first one customers in the record?

Customer..	January 3, 2017
Charlotte	I
Samuel	I

The question is to find the first customer who ever visited the store. It is shown that Charlotte is the first ever customer. This could be used as a sentimental reference in the future to congratulate the first customer.

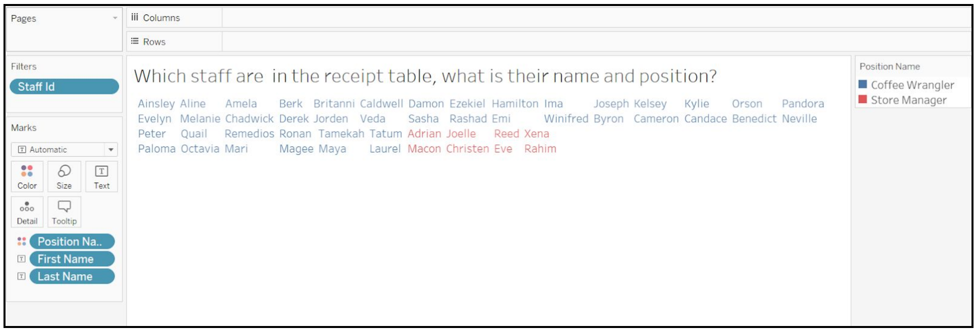
- How many times have customers visited on their birthday?



Since the dataset only includes sales records for the month of April, this table shows how many customers visited on that day are having their birthday. Sometimes it could be multiple visits on the same day if the customer purchased several orders.

Staff

- Which staff are in the receipt table? What are their names and positions?



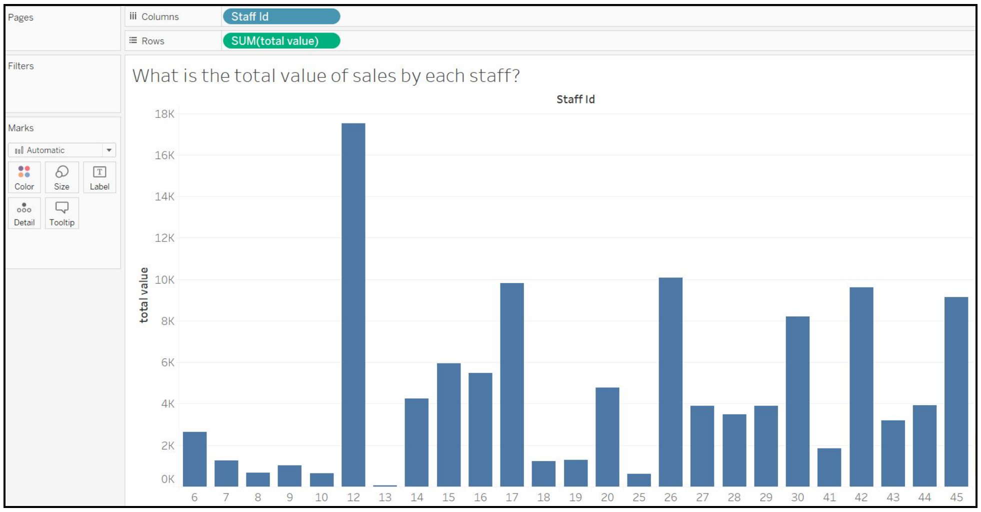
Which staff are in the receipt table, what is their name and position?

Ainsley Aline Amela Berk Britanni Caldwell Damon Ezekiel Hamilton Ima Joseph Kelsey Kylie Orson Pandora
Evelyn Melanie Chadwick Derek Jorden Veda Sasha Rashad Emi Winifred Byron Cameron Candace Benedict Neville
Peter Quail Remedios Ronan Tamekah Tatum Adrian Joelle Reed Xena
Paloma Octavia Mari Magee Maya Laurel Macon Christen Eve Rahim

Position Name
🟦 Coffee Wrangler
🟥 Store Manager

This visualization lists the names of all staff who have made sales in the coffee shop and their responsibilities at the store. This table can be used to track all staff who have interacted with customers and made sales.

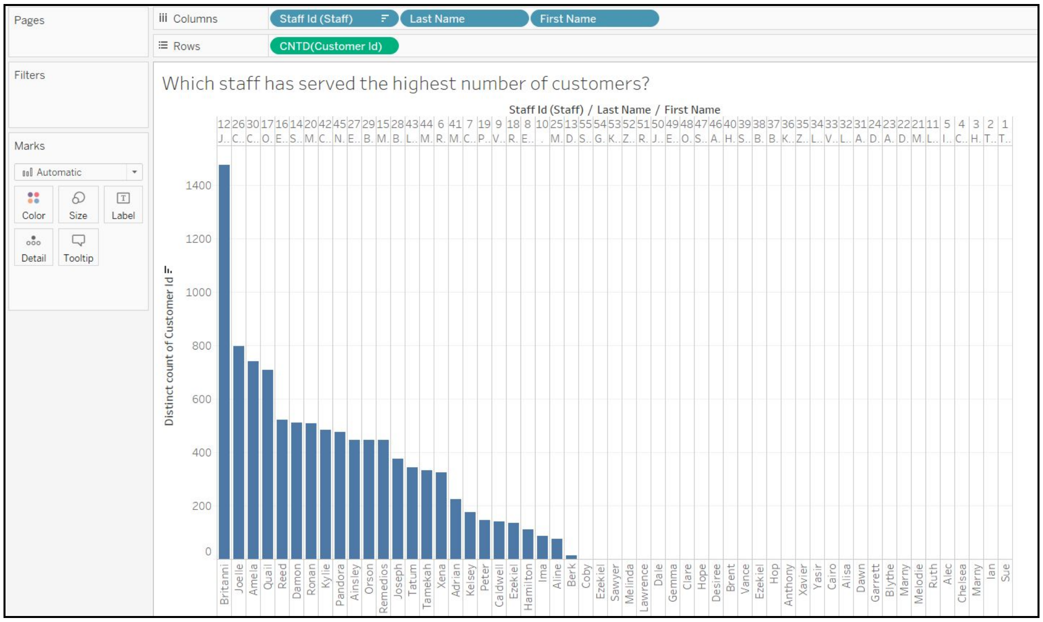
- What is the total value of sales by each staff?



What is the total value of sales by each staff?

This bar chart demonstrates the overall gross performance of all staff members. Staf 12, 17, 26, 30, 42, 45 have the best performance, where staff 12 has the leading performance among all. This analysis needs to be taken with the perspective that it does not include the length of time the staff has worked, as some staff are more experienced and have longer history than others.

- Which staff has served the highest number of customers?



Pages

Filters

Marks

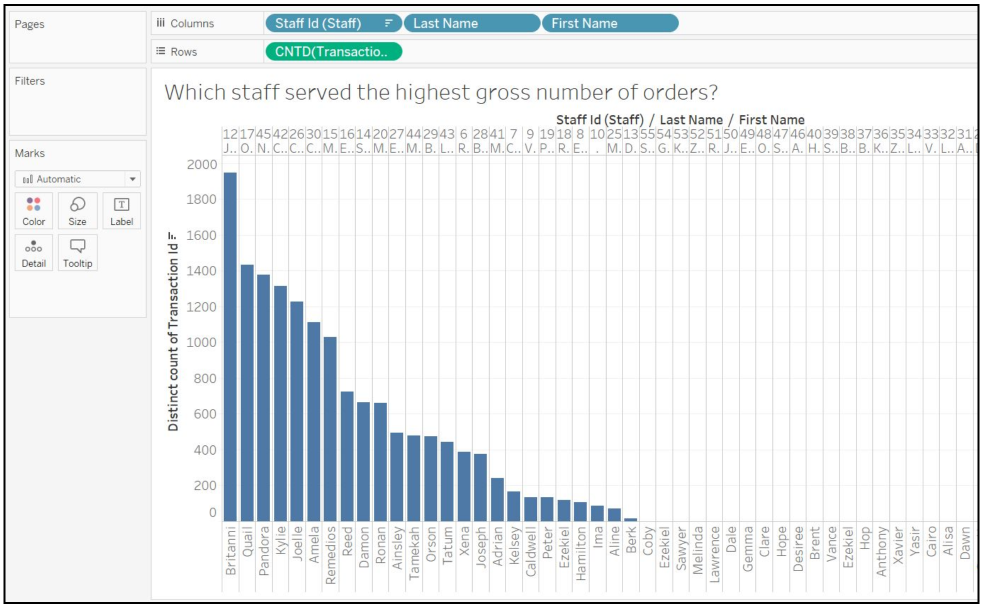
Columns: Staff Id (Staff) | Last Name | First Name

Rows: CNTD(Customer Id)

Which staff has served the highest number of customers?

This table shows the number of customers each staff has served. Some staff work in the office and do not interact with customers, thus the reason it’s 0 for some customers. It can be observed the staff Britanni served most number of customers, followed by Joelle and Amela.

- Which staff served the highest gross number of orders?



Pages

Filters

Marks

Columns: Staff Id (Staff) | Last Name | First Name

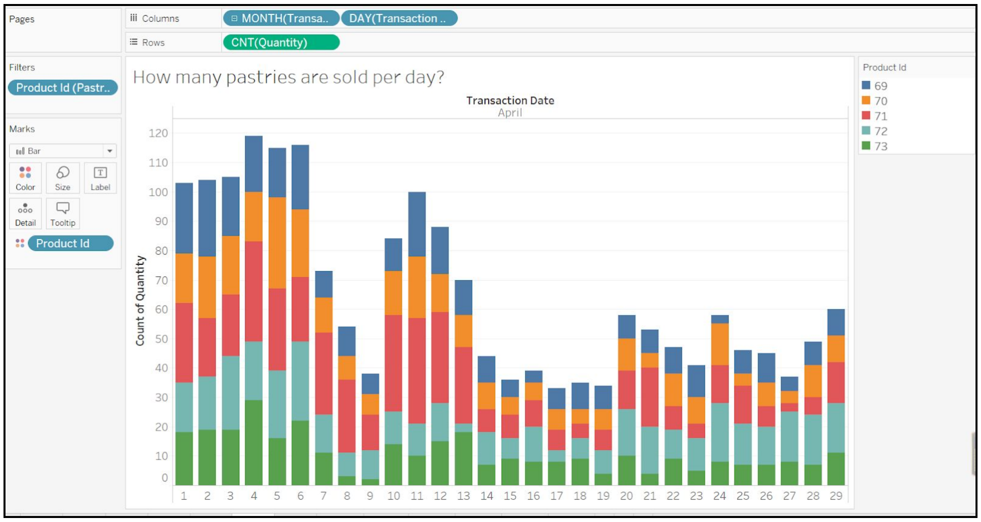
Rows: CNTD(Transactio..

Which staff served the highest gross number of orders?

Contrary to the question of the number of customers served, this question patients a different picture showing the sales performance. Britanni still have the highest number of orders, followed by Quail and Pandora.

Pastry

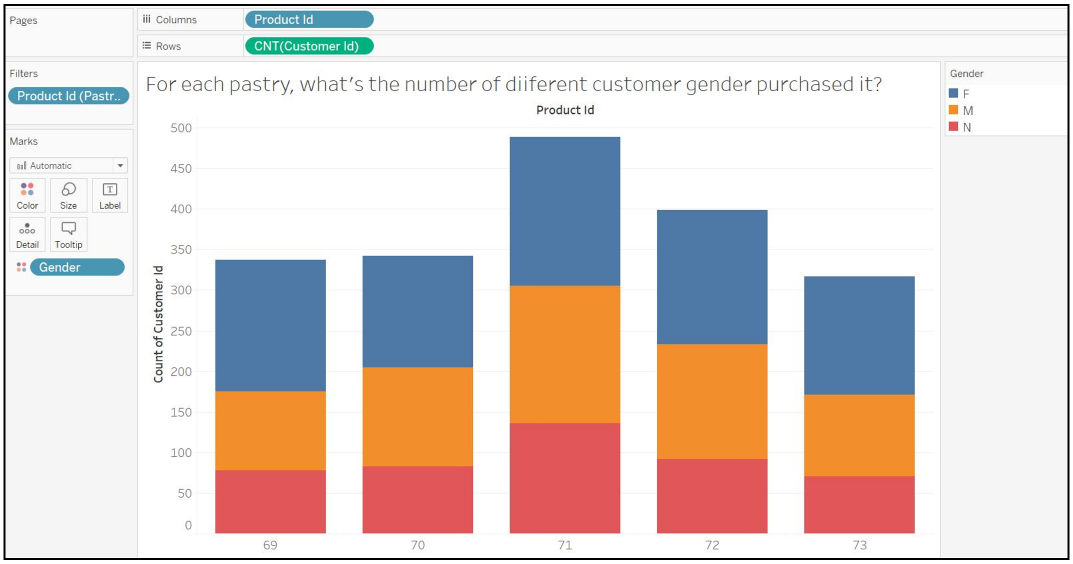
- How many pastries are sold per day?



Columns: MONTH(Transa.. DAY(Transaction ..
Rows: CNT(Quantity)
Filters: Product Id (Pastr..
Marks: Bar
Product Id

Through this bar chart, it can be observed that most pastry items are sold evenly through. Though in some instances, product 69 (blue) and product 73 (green) are sold less on some days.

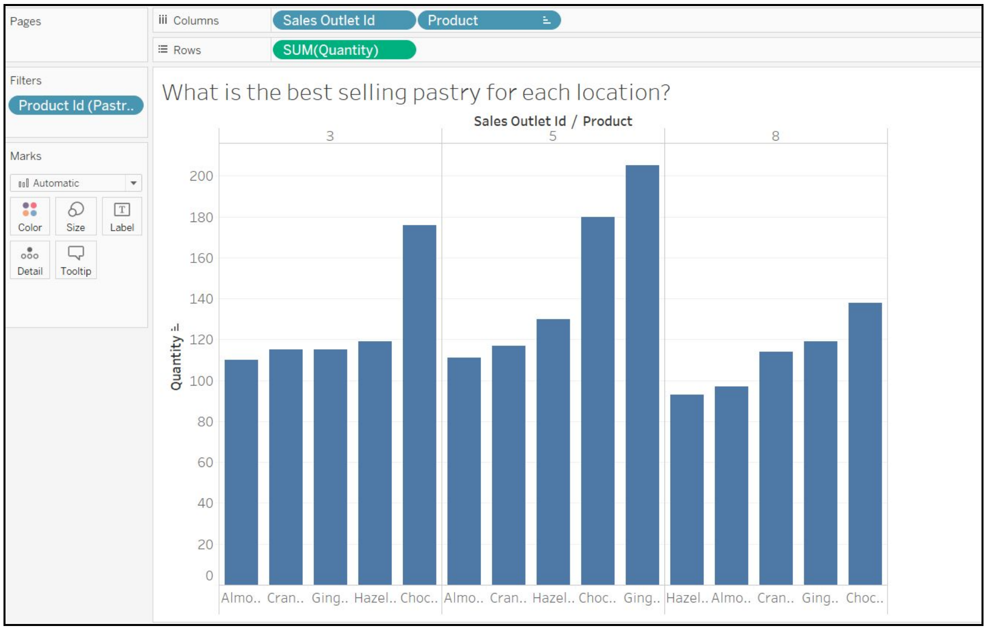
- For each pastry, what's the number of different customer gender purchased?



Columns: Product Id
Rows: CNT(Customer Id)
Filters: Product Id (Pastr..
Marks: Automatic
Gender

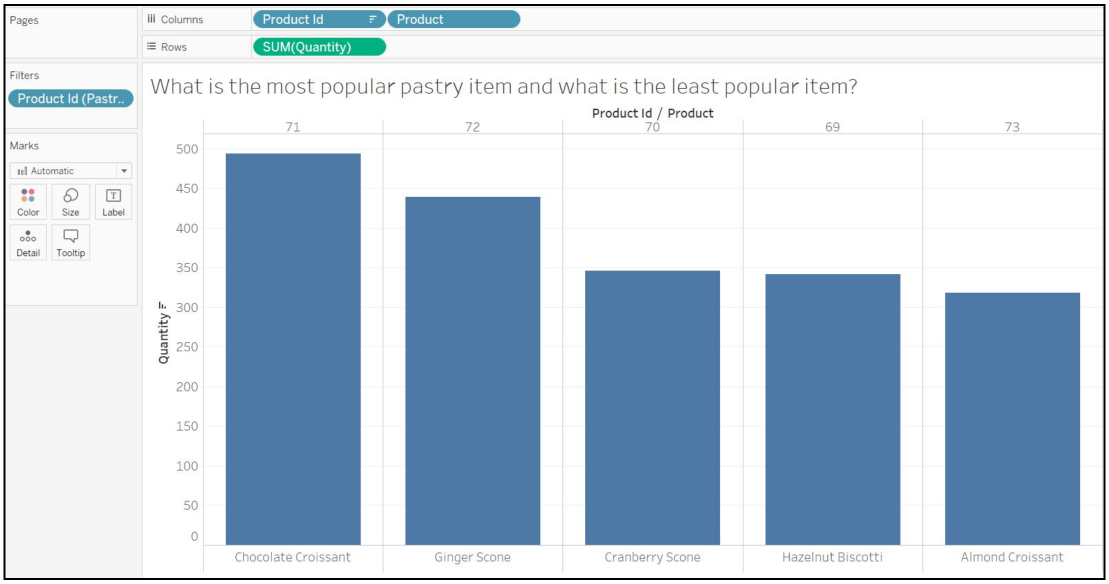
This table demonstrates that pastry item 69 and item 73 are sold more popular among female customers than male customers. Other items are sold more evenly among female or male customers. Some customers also didn't indicate their gender identity, thus the red part in the bar graph.

- What is the best selling pastry for each location?



This table filters the sales performance of pastry items by store location. It can be shown that Chocolate croissants are sold best in store 3 and 8, and Ginger scone is sold the best in store 5.

- What is the most popular pastry item and what is the least popular item?



This bar chart ranks the total quantity of sales for each pastry item. Chocolate croissants sell the best and almond croissants are sold the worst overall.