Chose a dataset from which we could gather business insights. Designed the database, cleaned the data and tested the database.

Retail data of a Coffee Chain

Project 1 – Database Foundations for Business Analytics (BUAN 6320.005)

Group 11:
Maneka Prabhu (MXP210102)
Rishapp Rajesh (RXR210108)
Shivanag Avula (SXA162630)
Venkata Durga Sai Kowshik Vooda (VXV210032)
Vikhil Baswaraju (VXB210053)

Step 1: Choose a Dataset

The dataset we chose contains representative retail data for a coffee chain. The dataset is well structured and does not have a significant number of missing values. The dataset contains the following tables that will assist us in analysing the business and will provide sufficient data through the following tables:

1. Sales Receipts
2. Pastry Inventory
3. Sales Targets
4. Customers
5. Dates
6. Products
7. Sales Outlets
8. Staff
9. Generation

In Step 3, we will demonstrate the data contained in each table further. We have sourced the dataset from Kaggle.com and have attached the source link in the footnote for your reference[1].

Step 2: Business Understanding

1. Why the data has been gathered?
    - To analyze the different KPIs (key performance indicators) that are significant to a coffee shop business, like sales, profit, margins, customers gained, retained, or lost, the volume of orders, etc.
    - To assist in proactive decision-making for the future. For instance, understanding which outlet is contributing to the majority of total sales and which is contributing the least. This helps the business owners decide which region to focus on, in the coming quarter.

2. What can be done with the data? What can be achieved?
    - Generate reports exhibiting the trends of various metrics over time. For example, since we have historical customer data, we can create month-on-month and year-on-year comparisons of customer retention and growth.
    - Build prediction models to forecast sales and margins. For example, we can analyze the sales of different products and predict future sales to understand which products to discontinue.
3. What are some of the goals/targets we have regarding the business that we can achieve by investigating this data?
    - Identify the products that are not profitable to decide whether processes need to be changed or the product needs to be discontinued altogether.
    - Efficient inventory management by analyzing the waste % for each product and deciding the optimal range.
    - Choosing from customer retention or growth strategies based on past trends.
    - Identify anomalies or significant patterns that can signal the need for immediate action.
    - Conducting appropriate research can help in formulating strategies to increase ROI (Return on Investment).

[1]Dataset Source: Coffee shop sample data (11.1.3+) | Kaggle

4. What insightful information can this data provide us that can be used to improve the business?
    - Top 10 bestselling products in terms of profits and sales
    - Top 10 Customers in terms of profits and order frequency
    - Outlet contributing to the majority of total sales
    - Peak months during which sales are comparatively higher and peak times during the day when order frequency is highest
    - Periodically assessing the sales targets set and whether they are being met

5. Why are we studying this data?
   This data helps in business decision-making. Analysis of the data brings attention to certain issues that need to be timely tackled. The business can improvise and make changes to the strategies implemented based on the performance of the KPIs.

6. Are there any problems in our business (based on the given data)?
   We weren't able to find any major problems with the business based on the given data. We noticed the %waste in the pastryinventory table was pretty high which can probably be linked to the sales targets set. Additionally, while working on the dataset, we felt the data collection process could be improved. We observed that certain files did not describe accurately the data being collected. This would lead to incorrect inferences and can adversely impact the decisions taken.

7. Can we find any solutions to these problems by studying this data?
   By understanding which category of products leads to maximum wastage and by narrowing down the sales trends, we can better estimate the amount of inventory to stock up. Inventory is stocked up based on the certain sales targets that are set, since we have the sales targets for each category and the sales generated from each, the business can modify the sales targets for each outlet. These reasonable targets might further help reduce the %waste.

8. What are some of the things we can optimize/improve in our business by studying this data?
    - The pricing structure can be improved based on the demand and inventory analysis
    - Promotions can be offered on items that are in less demand, alternatively, it can also be confirmed if the products for which promotions are currently being offered are in demand
    - Sales-based commission can be offered to staff as an incentive plan to improve performance
    - Based on weekly data, peak and lean days can be identified to offer additional offers

Step 3: Data Understanding

Below are the columns contained in each table along with their data types and respective statistics:

1. '**salesreceipts**' Table

Column Name	Data Type	Range	Mean	Std Dev	Variance	Frequency
transaction_id	INT	NA	NA	NA	NA	49,894
transaction_date	TEXT	4/1/2019 - 4/29/2019	NA	NA	NA	49,894
transaction_time	BIGINT	NA	NA	NA	NA	49,894
sales_outlet_id	INT	NA	NA	NA	NA	49,894
staff_id	DOUBLE	NA	NA	NA	NA	49,894
customer_id	INT	NA	NA	NA	NA	49,894
instore_yn	TEXT	Y/N	NA	NA	NA	49,894
order	INT	1 - 9	1.17	1.03	1.05	49,894
line_item_id	INT	NA	NA	NA	NA	49,894
product_id	BIGINT	NA	NA	NA	NA	49,894
quantity	DOUBLE	1 - 8	1.44	0.54	0.29	49,894
line_item_amount	DOUBLE	0 - 360	4.68	4.44	19.68	49,894
unit_price	DOUBLE	0.8 - 45	3.38	2.68	7.20	49,894
promo_item_yn	TEXT	Y/N	NA	NA	NA	49,894

2. '**customer**' Table

Column Name	Data Type	Range	Mean	Std Dev	Variance	Frequency
customer_id	INT	NA	NA	NA	NA	2,246
home_store	INT	NA	NA	NA	NA	2,246
customer_first-name	TEXT	NA	NA	NA	NA	2,246
customer_email	TEXT	NA	NA	NA	NA	2,246
customer_since	TEXT	1/3/2017 - 4/9/2019	NA	NA	NA	2,246
loyalty_card_number	TEXT	NA	NA	NA	NA	2,246
birthdate	TEXT	NA	NA	NA	NA	2,246
gender	TEXT	M, F, N	NA	NA	NA	2,246
birth_year	INT	1950 - 2001	NA	NA	NA	2,246

3. '**dates**' Table

Column Name	Data Type	Range	Mean	Std Dev	Variance	Frequency
transaction_date	TEXT	4/1/2019 - 4/29/2019	NA	NA	NA	30
Date_ID	INT	NA	NA	NA	NA	30
Week_ID	INT	NA	NA	NA	NA	30
Week_Desc	TEXT	14 - 18 weeks	NA	NA	NA	30
Month_ID	INT	NA	NA	NA	NA	30
Month_Name	TEXT	April	NA	NA	NA	30
Quarter_ID	INT	2	NA	NA	NA	30
Quarter_Name	TEXT	Q2	NA	NA	NA	30
Year_ID	INT	2019	NA	NA	NA	30

4. '**generations**' Table

Column Name	Data Type	Range	Mean	Std Dev	Variance	Frequency
birth_year	INT	1946 - 2015	NA	NA	NA	70
generation	TEXT	NA	NA	NA	NA	70

5. '**pastryinventory**' Table

Column Name	Data Type	Range	Mean	Std Dev	Variance	Frequency
sales_outlet_id	INT	NA	NA	NA	NA	307
transaction_date	TEXT	4/1/2019 - 4/27/2019	NA	NA	NA	307
product_id	INT	NA	NA	NA	NA	307
start_of_day	INT	18 - 48	24.06	12.04	145.05	307
quantity_sold	INT	0 - 32	9.30	5.43	29.50	307
waste	INT	0 - 47	14.66	11.18	125.08	307
% waste	TEXT	0% - 96%	58%	21%	4%	307

6. '**product**' Table

Column Name	Data Type	Range	Mean	Std Dev	Variance	Frequency
product_id	INT	NA	NA	NA	NA	88
product_group	TEXT	NA	NA	NA	NA	88
product_category	TEXT	NA	NA	NA	NA	88
product_type	TEXT	NA	NA	NA	NA	88
product	TEXT	NA	NA	NA	NA	88
product_description	TEXT	NA	NA	NA	NA	88
unit_of_measure	TEXT	NA	NA	NA	NA	88
current_wholesale_pr	DOUBLE	0.04 - 36	3.89	5.62	31.56	88
current_retail_price	TEXT	0.8 - 45	6.58	7.12	50.67	88
tax_exempt_yn	TEXT	Y/N	NA	NA	NA	88
promo_yn	TEXT	Y/N	NA	NA	NA	88
new_product_yn	TEXT	Y/N	NA	NA	NA	88

7. '**salestargets**' Table

Column Name	Data Type	Range	Mean	Std Dev	Variance	Frequency
sales_outlet_id	INT	NA	NA	NA	NA	8
year_month	TEXT	NA	NA	NA	NA	8
beans_goal	INT	720 - 1000	777.50	102.68	10,543.75	8
beverage_goal	INT	13,500 - 18,000	14,578.13	1,925.30	3,706,787.11	8
food_goal	INT	3,420 - 4,750	3,693.13	487.74	237,893.36	8
merchandise _goal	INT	360 - 500	388.75	51.34	2,635.94	8
total_goal	INT	18,000 - 25,000	19,437.50	2,567.07	6,589,843.75	8

8. '**sales_outlet**' Table

Column Name	Data Type	Range	Mean	Std Dev	Variance	Frequency
sales_outlet_id	INT	NA	NA	NA	NA	9
sales_outlet_type	TEXT	NA	NA	NA	NA	9
store_square_feet	INT	900 - 3,400 sqft	1,544.44	700.97	491,358.02	9
store_address	TEXT	NA	NA	NA	NA	9
store_city	TEXT	NA	NA	NA	NA	9
store_state_province	TEXT	NA	NA	NA	NA	9
store_telephone	TEXT	NA	NA	NA	NA	9
store_postal_code	INT	NA	NA	NA	NA	9
store_longitude	DOUBLE	NA	NA	NA	NA	9
store_latitude	DOUBLE	NA	NA	NA	NA	9
manager	TEXT	6 - 41	23.50	11.46	131.25	9
Neighorhood	TEXT	NA	NA	NA	NA	9

9. '**staff**' Table

Column Name	Data Type	Range	Mean	Std Dev	Variance	Frequency
staff_id	INT	NA	NA	NA	NA	55
first_name	TEXT	NA	NA	NA	NA	55
last_name	TEXT	NA	NA	NA	NA	55
position	TEXT	NA	NA	NA	NA	55
start_date	TEXT	3/8/2001 - 11/22/2018	NA	NA	NA	55
location	TEXT	NA	NA	NA	NA	55

The above tables represent the summary information that each of the tables in the dataset contain. We used the SQL commands **MIN()** and **MAX()** to compute the range, **AVG()** for mean, **STDDEV()** for standard deviation, **VARIANCE()** for variance, and **COUNT()** for frequency.

Quality of the data

The column names didn't have to be changed for any. With regard to missing values, only 1 value in the sales_outlet table was missing. The value missing corresponds to the manager of the warehouse.

Step 4: Design a Database

Schema Design

Entity: customer

Primary key: customer_id
Foreign key: birth_year

Entity: generations

Primary key: birth_year

Entity: product

Primary key: product_id

Entity: staff

Primary key: staff_id

Entity: pastryinventory

Primary key: product_id
Foreign key: sales_outlet_id, transaction_date, product_id

Entity: salesreceipts

Primary key: N/A
Foreign key: transaction_date, product_id, sales_outlet_id, staff_id, customer_id

Entity: sales_outlet

Primary key: sales_outlet_id

Entity: salestargets

Primary key: sales_outlet_id
Foreign key: sales_outlet_id

Entity: dates

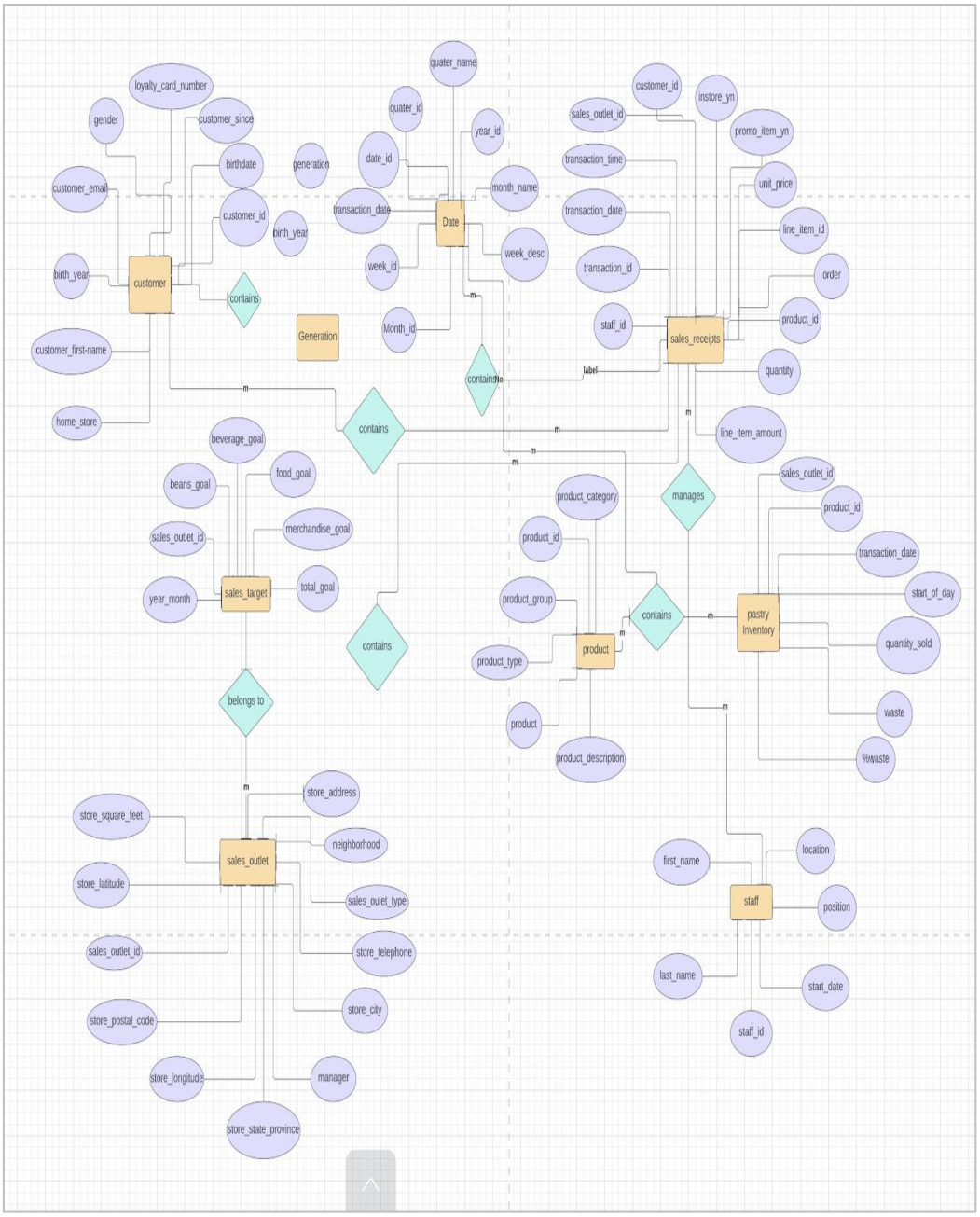
Primary key: transaction_date, Date_id

Relationships:

1. product to pastryinventory:
   Each product can contain one or more pastry inventory
   Each row in the pastry inventory table can correspond to one or many product(s)
2. pastryinventory to dates:
   Each inventory entry can have only one date
3. salesreceipts to dates:
   Each sales receipt can have only one date
4. sales_outlet to salestargets:
   Each sales outlet is allocated a sales target

5. customer to generation:
   Each customer belongs to one Generation (Baby Boomers, GenX, GenZ, Older Millenials, and Younger Millenials)
6. salesreceipts to Customer:
   Each sales receipt corresponds to one customer
7. salesreceipts to staff:
   Each sales receipt corresponds to one staff member
8. salesreceipts to sales_outlet:
   Each sales receipt matches with a transaction that occurred at one sales outlet

ER Diagram:



Schema Normalization:

Functional Dependencies from the schema

- product table
  - **{** product_group, product_category, product_type, product_description **} ->** product_id
- customer table
  - **{** customer_email, loyalty_card_number **} ->** customer_id
- Salesreciepts table
  - **{** transaction_time, product_id, transaction_date, customer_id**} ->** transaction_id
- sales_outlet table
  - **{** sales_outlet_type, store_address **} ->** sales_outlet_id
- staff table
  - **{** first_name, last_name, position, location**} ->** staff_id

Check if your schema is in BCNF (Boyce-Codd Normal Form)

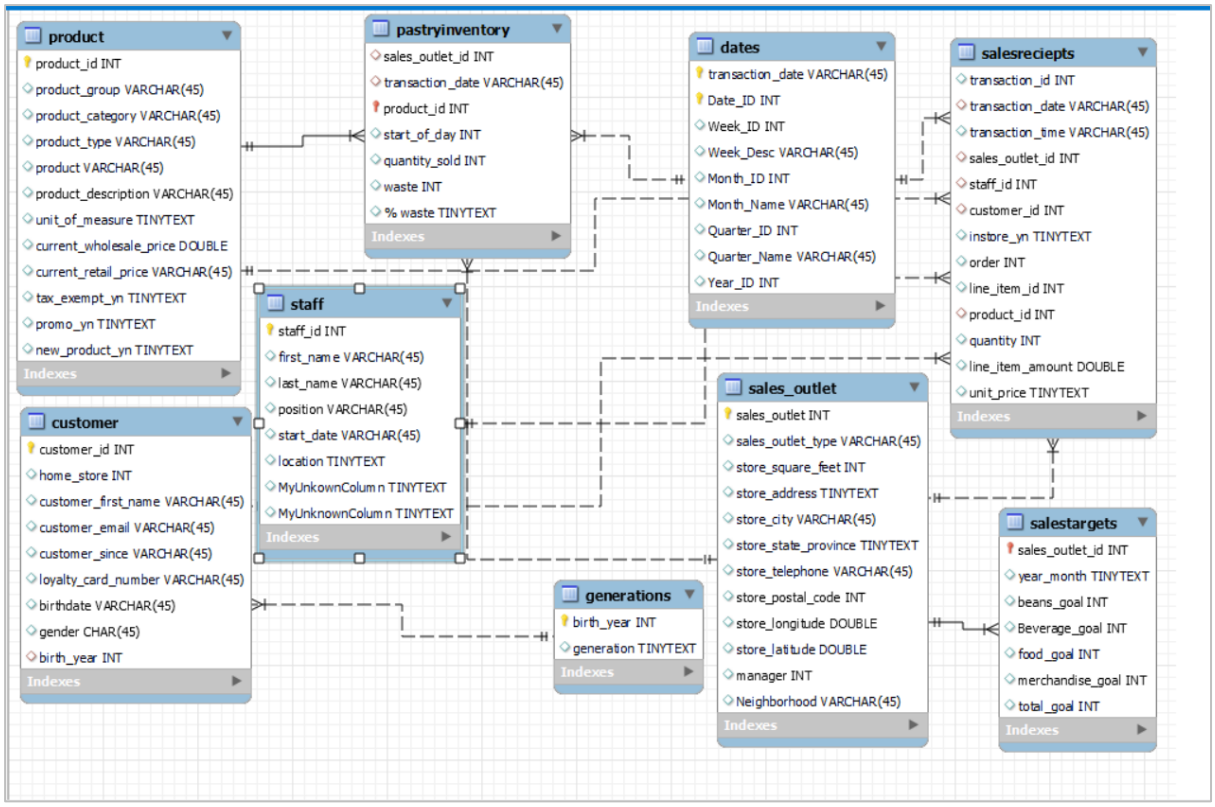
A schema is in BCNF, if a table is in 3NF and has the table's super key for each functional dependency.
In our dataset, there is a super key for each functional dependency to verify that the schema is in BCNF, such as

- product id, product group, product category, product type, product description
- transaction_id, transaction_time, product_id, transaction_date, customer_id
- customer id, customer email, and loyalty card number
- sales outlet id, sales outlet type, store address
- staff id, first name, last name, position, and location.

Every functional dependence has a super key, which demonstrates that the schema is in BCNF: product id, transaction_id, customer id, sales outlet id, staff id

There were no errors while importing data into the database

Schema Design



Step 5: Data Cleaning and Database Testing

A. Statistics

We have carried out the basic statistical analysis including range, mean, variance, and frequency in Step 3. Some of the inferences we can make based on it are:

a. From the 'salesreceipts' table, we can infer that the sales for the month of April 2019 are being analyzed, based on the range of the column 'transaction_date'. The column line_item_amount (net selling price of an order), where the mean is $4.68, implies that an order placed by the customer in these coffee shops, on average amount to $4.68.
b. From the 'customer' table, we can see that the customers who visited the coffee shop in the given time period were born between 1950 and 2001.
c. From the 'pastryinventory' we can say that inventory value at the start of the day on average is $24.06 with a high variability over the month, the average quantity sold is $9.3, and the % waste was around 58%.
d. The product table shows that the average current_retail_price is 69% higher than the average current_wholesale_price. (% change calculated from $3.89 to $6.58)
e. On average, each sales outlet is expected to meet a sales target of $19437.5, including beans, beverages, food, and merchandise.

B. General Queries for importing and sorting the dataset

The queries used are as follows:

1. 201904 sales receipts (Salesreciepts) –
    i. Alter table name from **201904 sales receipts** to **Salesreciepts** –
    ALTER TABLE `Project1`.`201904 sales receipts`
    RENAME TO `Project1`.`Salesreciepts`

ii. Alter **transaction date** datatype from **TEXT** to **DATE**
ALTER TABLE `Project1`.`Salesreciepts`
CHANGE COLUMN `transaction_date` `transaction_date`
DATE NULL DEFAULT NULL

iii. Set **transaction_id** as a primary key
ALTER TABLE `project`.`salesreciepts`
CHANGE COLUMN `transaction_id` `transaction_id` INT NOT NULL ,
ADD PRIMARY KEY (`transaction_id`)

2. customer (Customer)
   i. Set **customer id** as a **primary key**
   ALTER TABLE `Project1`.`customer`
   CHANGE COLUMN `customer_id` `customer_id` INT NOT NULL,
   ADD PRIMARY KEY (`customer_id`)

3. Generations
   i. Set **birth_year** as a **primary key**
   ALTER TABLE `Project1`.`generations`
   CHANGE COLUMN `birth_year` `birth_year` INT NOT NULL,
   ADD PRIMARY KEY (`birth_year`)

4. Pastry inventory
   i. ALTER TABLE `project`.`pastry inventory`
   ALTER TABLE `project`.`pastry inventory`
   ADD CONSTRAINT `sales_outlet_idFK`
   FOREIGN KEY (`sales_outlet_id`)
   REFERENCES `project`.`salesoutlet` (`sales_outlet_id`)
   ON DELETE NO ACTION
   ON UPDATE NO ACTION,
   ADD CONSTRAINT `product_idFK`
   FOREIGN KEY (`product_id`) REFERENCES `project`.`product` (`product_id`)

5. Product
   i. Set product_id as a primary key
   ALTER TABLE `Project1`.`product`
   CHANGE COLUMN `product_id` `product_id` INT NOT NULL,
   ADD PRIMARY KEY (`product_id`)

6. Sales_outlet
   i. Set sales_outlet_id as a primary key
   ALTER TABLE `Project1`.`sales_outlet`
   CHANGE COLUMN `sales_outlet_id` `sales_outlet_id` INT NOT NULL,
   ADD PRIMARY KEY (`sales_outlet_id`)

   ii. Add missing figure **'30' under column manager where row sales_out_id is 2**
   UPDATE `Project1`.`sales_outlet` SET `manager` = '30'
   WHERE (`sales_outlet_id` = '2')

7. Sales target (Salestargets)
    i. Set sales_outlet_id as a primary key and table name to Salestargets
    ALTER TABLE `Project1`.`sales targets`
    CHANGE COLUMN `sales_outlet_id` `sales_outlet_id` INT NOT NULL,
    ADD PRIMARY KEY (`sales_outlet_id`); , RENAME TO `Project1`.`Sales_targets`

8. Staff
    i. Set staff_id as a primary key
    ALTER TABLE `Project1`.`staff`
    CHANGE COLUMN `staff_id` `staff_id` INT NOT NULL,
    ADD PRIMARY KEY (`staff_id`)

    ii. Delete 2 unknown column
    ALTER TABLE `Project1`.`staff`
    DROP COLUMN `MyUnknownColumn_[0]`,
    DROP COLUMN `MyUnknownColumn

9. Dates
    i. Alter transaction date datatype from **TEXT** to **DATE**
    ALTER TABLE `Project1`.`dates`
    CHANGE COLUMN `transaction_date` `transaction_date` DATE NULL DEFAULT NULL

    ii. Set transaction_date as a primary key
    ALTER TABLE `Project1`.`dates`
    CHANGE COLUMN `transaction_date` `transaction_date` DATE NOT NULL,
    ADD PRIMARY KEY (`transaction_date`)

Conclusion

Using SQL, we were able to observe, clean, and manipulate the retail data of a coffee chain to generate valuable insights and inferences. The dataset required minimal cleansing which we were able to do with commands like CREATE, ALTER, and DROP. The overall data quality was good. We were able to understand the data better by computing the basic statistical measures like mean, std dev, and variance. However, for a more comprehensive and exhaustive analysis of the business, we would need more data points like sales data across all months. Additionally, better clarity on some of the columns will help ensure more appropriate business decisions.