Data Management Assignment Report

Table of contents

Introduction

There are two common workflows for data integration: ELT (Extract, Load, Transform) and ETL (Extract, Transform, Load)

In ELT , normalization and data validation are performed after the data is injected into the database whereas in the ETL both transformations are performed after data injection. In our work, we chose ETL workflow and used SQL and R interchangeably to perform transformations.

Part 1: Database Design and Implementation

Task 1.1: E-R Diagram

Entities

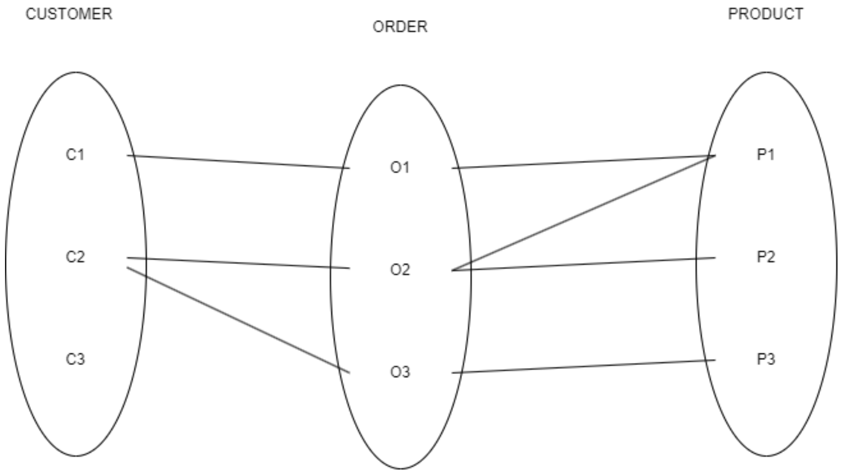
Our E-R Diagram for the E-commerce database consisted of five main entities:

1. CUSTOMER: Details of buyers, including attributes such as customer ID, customer name, contact information, and gender.
2. PRODUCT: Details of sale items, including product ID, product name, description, rating, price, and available stock.
3. ADVERTISEMENT: Details of promotional activities used to advertise products, including advertisement ID, number of times the advertisement is shown, cost of the advertisement, and advertisement placement.
4. SUPPLIER: Details of product providers including supplier ID, supplier name, and contact information.
5. CATEGORY: Represents product classifications, including category ID, name, and fee percentage.

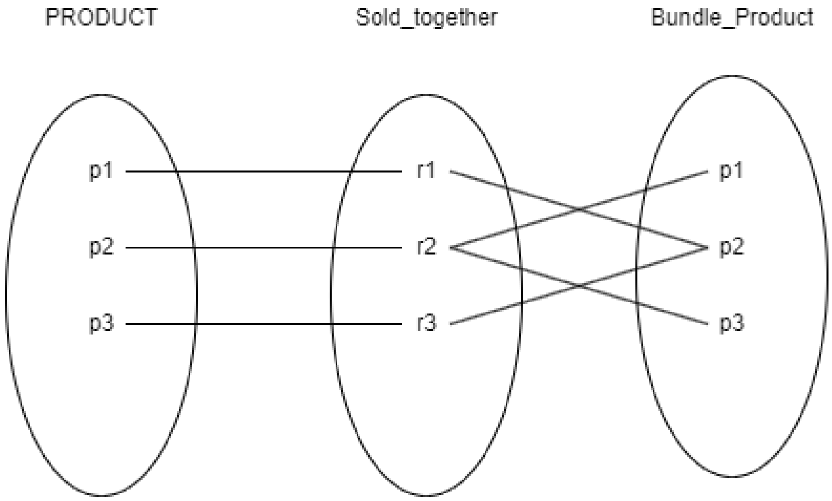
Relationships

The relationships among these entities are as follows:

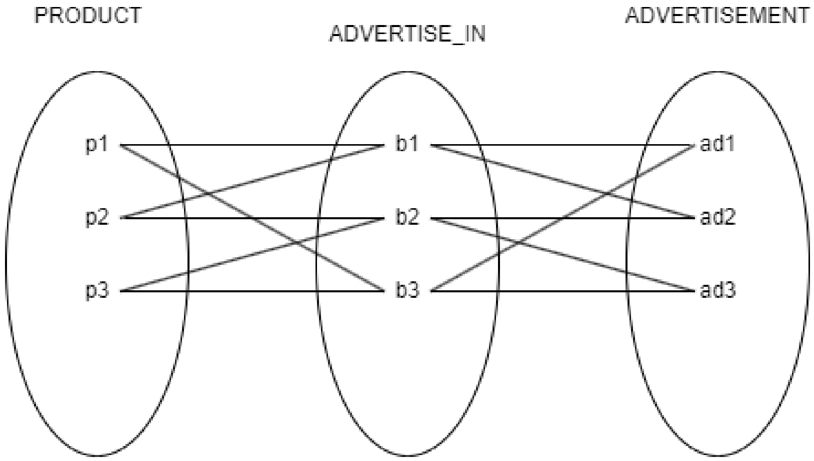
- CUSTOMER and PRODUCT entities have an M-N relationship where one customer can order many products, and one product can be ordered by many customers. This relationship is represented as ORDER, which contains the details of each order, including order ID, order date, order status, order quantity, promo code, payment method, and delivery fee.



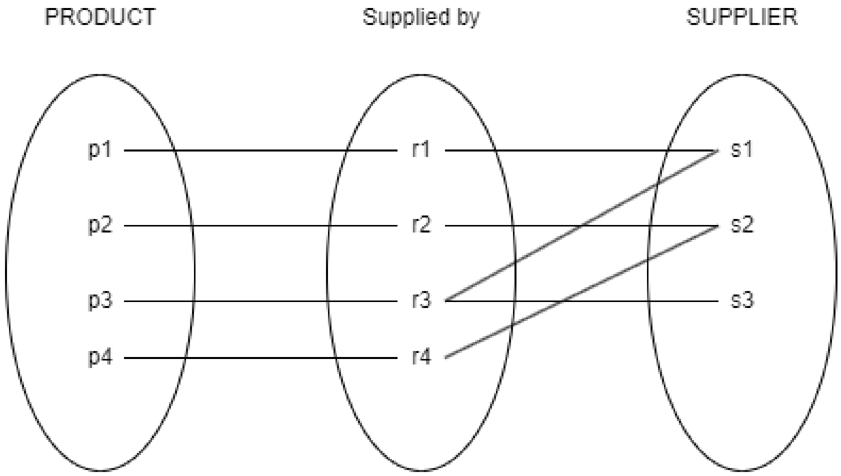
- PRODUCT entity has a self-referencing relationship Sold_Together of 1 to many. The relationship indicates that one product can be sold together with others as a bundle.



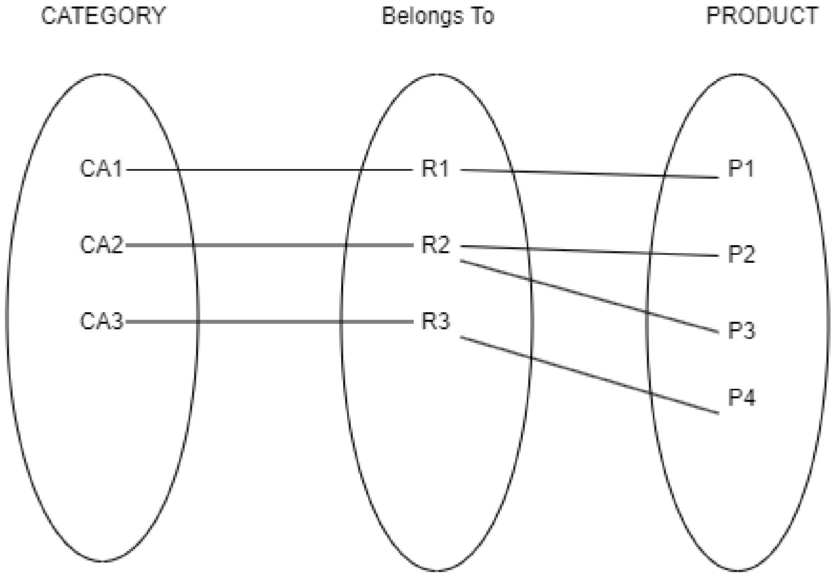
- PRODUCT and ADVERTISEMENT have an M-N relationship. ADVERTISE_IN represents this relationship where multiple advertisements can promote one product, and one advertisement can promote multiple products.



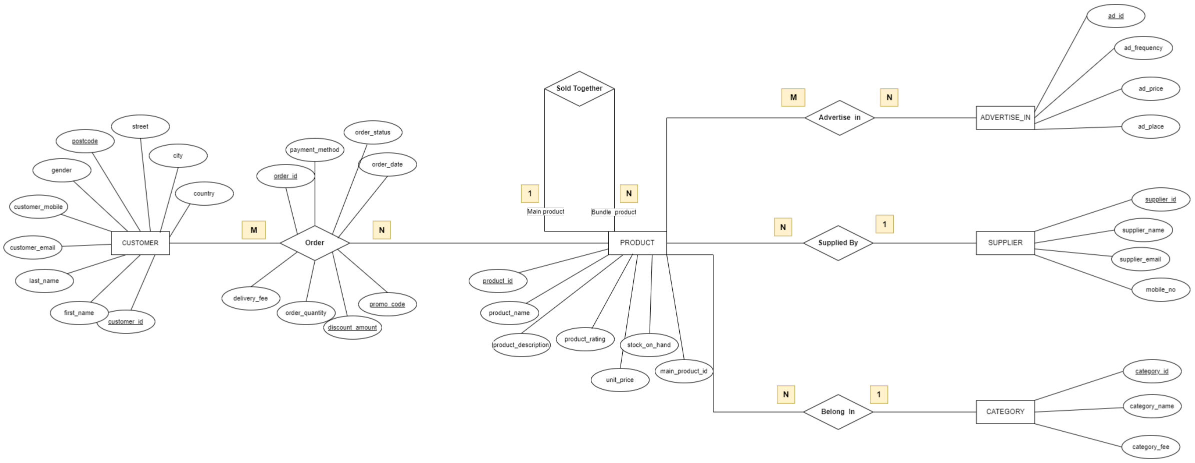
- PRODUCT and SUPPLIER have an N-1 relationship, indicating that one product can be supplied by only one supplier, but one supplier can provide multiple products.



- CATEGORY and PRODUCT have 1-N relationship, where one product belongs to one category, but one category can contain multiple products.



As a result, this is our E-R diagram.



Task 1.2: SQL Database Schema Creation

Normalization:

Our initial ER Diagram required two updates for 3NF compliance. The first adjustment was on the ORDER table, which initially did not adhere to 3NF.

Functional Dependency:

- {order_id, customer_id, product_id } -> { order_date , order_status, order_quantity, promo_code, discount amount, payment_method, delivery_fee}
- {promo_code} -> { discount_amount}

Assuming order_id duplication for simultaneous customer orders led to UPDATE anomalies. If there is an update on any attributes, such as payment_methods, we must alter more than one row in case there is more than one product in that order_id. Additionally, discount_amount is also transitively dependent on promo_code.

Therefore, we divided the ORDER table into three entities:

1. ORDER_DETAIL holds individual order details with attributes including product_id (primary key), customer_id (foreign keys), promo_code (foreign keys), order_date, order_status, and payment_method.
2. ORDER_ITEM contains product IDs and the quantities per order using order_id (foreign keys) and product_id (foreign keys) as a composite primary key, and quantity.
3. DISCOUNT stores promotional discount data, with promo_code as the primary key and discount_amount.

The next adjustment was on the CUSTOMER table, which was initially on 2NF. Based on its functional dependency, street, city, and country can also be determined by both postcode and customer_id.

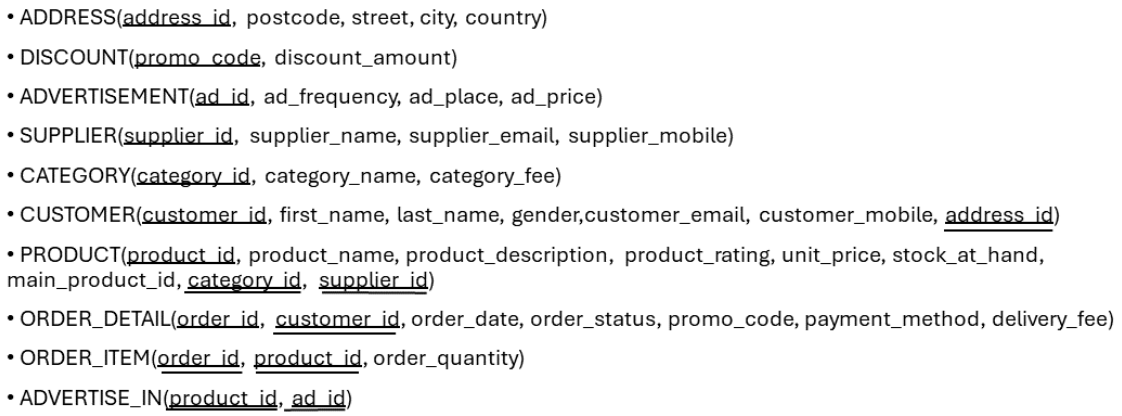
Functional Dependency:

- {customer_id} ->{ first_name, last_name, gender, email, customer_mobile, street, city, country, postcode }
- {postcode} -> { street, city, country }

This indicated that street, city, and country were transitively dependent on postcode. As a result, a new table is created as ADDRESS, which has postcode as a primary key and stores street, city, and country as other attributes.

Logical Schema

According to the E-R diagram and normalizations, we can list the logical schema as follows:



- ADDRESS(address_id, postcode, street, city, country)
- DISCOUNT(promo_code, discount_amount)
- ADVERTISEMENT(ad_id, ad_frequency, ad_place, ad_price)
- SUPPLIER(supplier_id, supplier_name, supplier_email, supplier_mobile)
- CATEGORY(category_id, category_name, category_fee)
- CUSTOMER(customer_id, first_name, last_name, gender,customer_email, customer_mobile, address_id)
- PRODUCT(product_id, product_name, product_description, product_rating, unit_price, stock_at_hand, main_product_id, category_id, supplier_id)
- ORDER_DETAIL(order_id, customer_id, order_date, order_status, promo_code, payment_method, delivery_fee)
- ORDER_ITEM(order_id, product_id, order_quantity)
- ADVERTISE_IN(product_id, ad_id)

Physical Schema

Firstly, we created a connection to our database named "database.db"

```r
connect <- dbConnect(RSQLite::SQLite(), "database.db")
```

Then, we first created entities without foreign keys, including ADDRESS, DISCOUNT, ADVERTISEMENT, SUPPLIER, and CATEGORY.

1. Create ADDRESS entity

```sql
CREATE TABLE IF NOT EXISTS ADDRESS (
  address_id VARCHAR(50) PRIMARY KEY,
  postcode VARCHAR(20) NOT NULL,
  street VARCHAR(50) NOT NULL,
  city VARCHAR(100) NOT NULL,
  country VARCHAR(100) NOT NULL
  );
```

2. Create DISCOUNT entity

    ```sql
    CREATE TABLE IF NOT EXISTS DISCOUNT (
      promo_code VARCHAR(20) PRIMARY KEY,
      discount_percent INT NOT NULL
    );
    ```

3. Create ADVERTISEMENT entity

    ```sql
    CREATE TABLE IF NOT EXISTS ADVERTISEMENT (
      ad_id VARCHAR(50) PRIMARY KEY,
      ad_frequency INT NOT NULL,
      ad_place VARCHAR(50) NOT NULL,
      ad_price DECIMAL(10, 2) NOT NULL
    );
    ```

4. Create SUPPLIER entity

    ```sql
    CREATE TABLE IF NOT EXISTS SUPPLIER (
      supplier_id VARCHAR(50) PRIMARY KEY,
      supplier_name VARCHAR(50) NOT NULL,
      supplier_email VARCHAR(50) NOT NULL UNIQUE,
      supplier_mobile VARCHAR(20) NOT NULL UNIQUE
    );
    ```

5. Create CATEGORY entity

    ```sql
    CREATE TABLE IF NOT EXISTS CATEGORY (
      category_id VARCHAR(50) PRIMARY KEY,
      category_name VARCHAR(50) NOT NULL,
      category_fee INT NOT NULL
    );
    ```

Then, we created entities that are children entities and entities that have referential integrity, which consist of CUSTOMER, PRODUCT, ORDER_DETAIL, ORDER_ITEM, and ADVERTISE_IN

6. Create CUSTOMER entity

```
CREATE TABLE IF NOT EXISTS CUSTOMER (
  customer_id VARCHAR(50) PRIMARY KEY,
  first_name VARCHAR(50) NOT NULL,
  last_name VARCHAR(50) NOT NULL,
  gender VARCHAR(10),
  customer_email VARCHAR(50) NOT NULL UNIQUE,
  customer_mobile VARCHAR(15) NOT NULL UNIQUE,
  address_id VARCHAR(50) NOT NULL,
  FOREIGN KEY (address_id) REFERENCES ADDRESS (address_id)
);
```

7. Create PRODUCT entity

```
CREATE TABLE IF NOT EXISTS PRODUCT (
  product_id VARCHAR(50) PRIMARY KEY,
  product_name VARCHAR(50) NOT NULL,
  product_description VARCHAR(50),
  product_rating DECIMAL(5,2),
  unit_price DECIMAL(10,2) NOT NULL,
  stock_on_hand INT NOT NULL,
  main_product_id VARCHAR(50),
  category_id VARCHAR(50) NOT NULL,
  supplier_id VARCHAR(50) NOT NULL,
  FOREIGN KEY (supplier_id) REFERENCES SUPPLIER (supplier_id),
  FOREIGN KEY (category_id) REFERENCES CATEGORY (category_id)
);
```

8. Create ORDER_DETAIL entity

```
CREATE TABLE IF NOT EXISTS ORDER_DETAIL (
  order_id VARCHAR(50) PRIMARY KEY,
  customer_id VARCHAR(50),
  order_date DATE NOT NULL,
  order_status VARCHAR(50) NOT NULL,
  promo_code VARCHAR(20),
  payment_method TEXT NOT NULL,
  delivery_fee DECIMAL(10, 2) NOT NULL,
```

```sql
  FOREIGN KEY (customer_id) REFERENCES CUSTOMER (customer_id),
  FOREIGN KEY (promo_code)  REFERENCES DISCOUNT (promo_code)
);
```

9. Create ORDER_ITEM entity

```sql
CREATE TABLE IF NOT EXISTS ORDER_ITEM (
  order_id VARCHAR(50),
  product_id VARCHAR(50),
  order_quantity INT NOT NULL,
  PRIMARY KEY (order_id, product_id),
  FOREIGN KEY (order_id)   REFERENCES ORDER_DETAIL (order_id)
  FOREIGN KEY (product_id) REFERENCES PRODUCT (product_id)
);
```

10. Create ADVERTISE_IN entity

```sql
CREATE TABLE IF NOT EXISTS ADVERTISE_IN (
  product_id VARCHAR(50),
  ad_id VARCHAR(50),
  PRIMARY KEY (product_id, ad_id),
  FOREIGN KEY (product_id) REFERENCES PRODUCT (product_id),
  FOREIGN KEY (ad_id)      REFERENCES ADVERTISEMENTS (ad_id)
);
```

Part 2: Data Generation and Implementation

Task 2.1: Synthetic Data Generation

Synthetic data from the normalized schema was created for each table using Mockaroo, a mock data generator platform. Data was initially generated for the parent entities to ensure that the referential integrity exists on the children entities or entities that need a reference for foreign keys from parent entities. The number of observations are as follows:

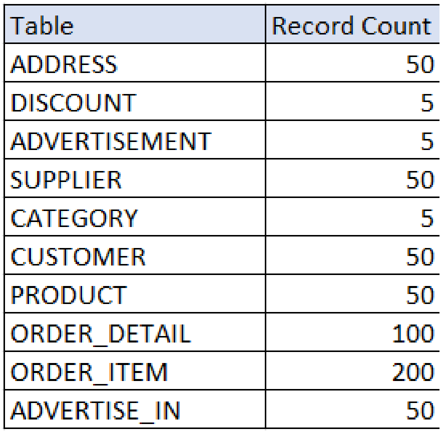
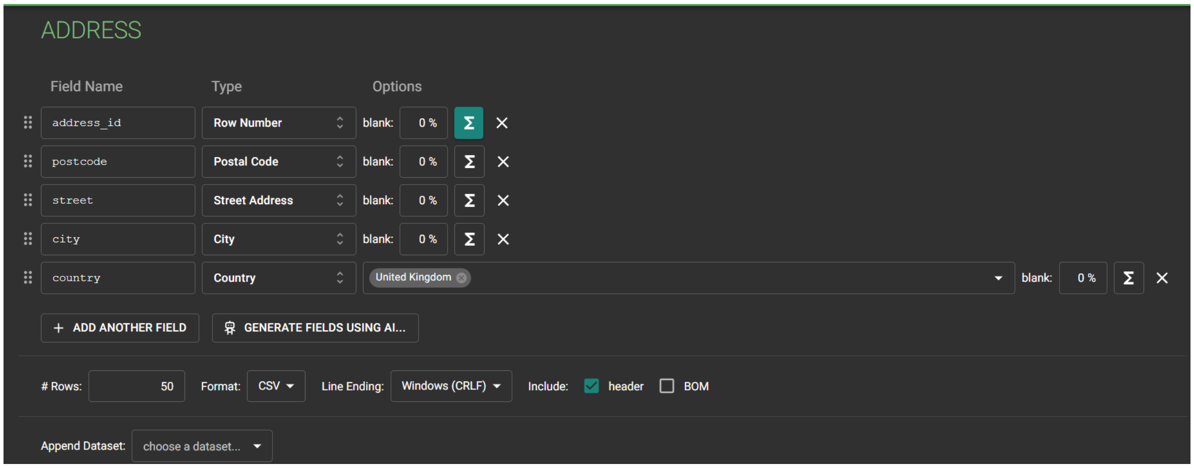


Table	Record Count
ADDRESS	50
DISCOUNT	5
ADVERTISEMENT	5
SUPPLIER	50
CATEGORY	5
CUSTOMER	50
PRODUCT	50
ORDER_DETAIL	100
ORDER_ITEM	200
ADVERTISE_IN	50

Data generation for parent entities

1. ADDRESS



ADDRESS

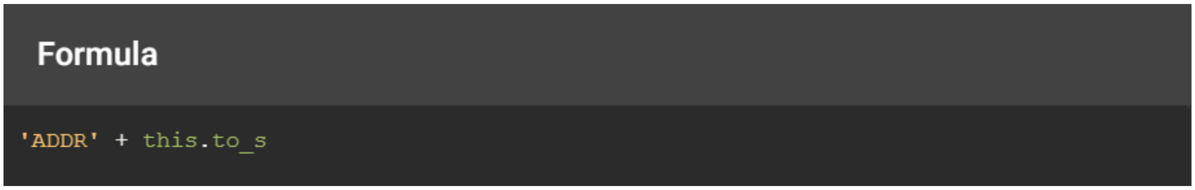
Field Name	Type	Options
address_id	Row Number	blank: 0 %
postcode	Postal Code	blank: 0 %
street	Street Address	blank: 0 %
city	City	blank: 0 %
country	Country	United Kingdom blank: 0 %

Rows: 50 Format: CSV Line Ending: Windows (CRLF) Include: ☑ header □ BOM

Append Dataset: choose a dataset...

As we would like to focus on the customers who live in the United Kingdom, the generated address information will only be in the United Kingdom.

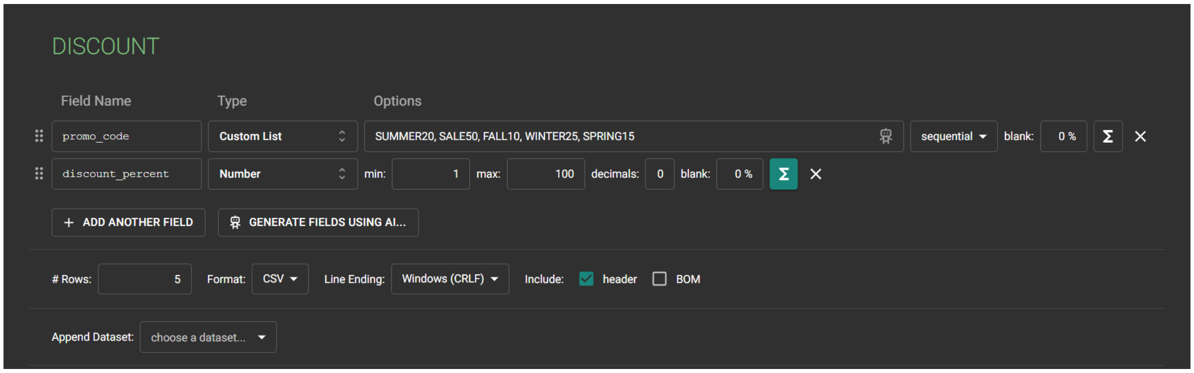
In addition, we have set the address_id, which is the primary of the table, to start with ‘ADDR’ and followed by the row number.



Formula

```
'ADDR' + this.to_s
```

2. DISCOUNT



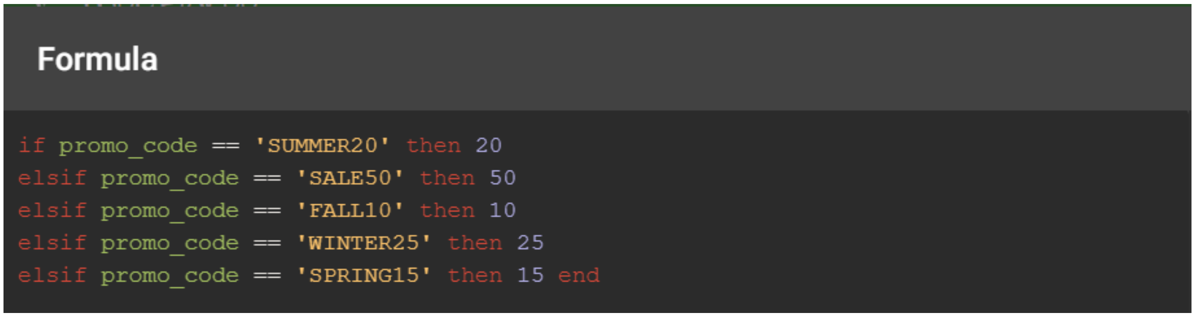
DISCOUNT

Field Name	Type	Options
promo_code	Custom List	SUMMER20, SALE50, FALL10, WINTER25, SPRING15 · sequential · blank: 0 %
discount_percent	Number	min: 1 · max: 100 · decimals: 0 · blank: 0 %

Rows: 5 · Format: CSV · Line Ending: Windows (CRLF) · Include: ☑ header □ BOM

Append Dataset: choose a dataset...

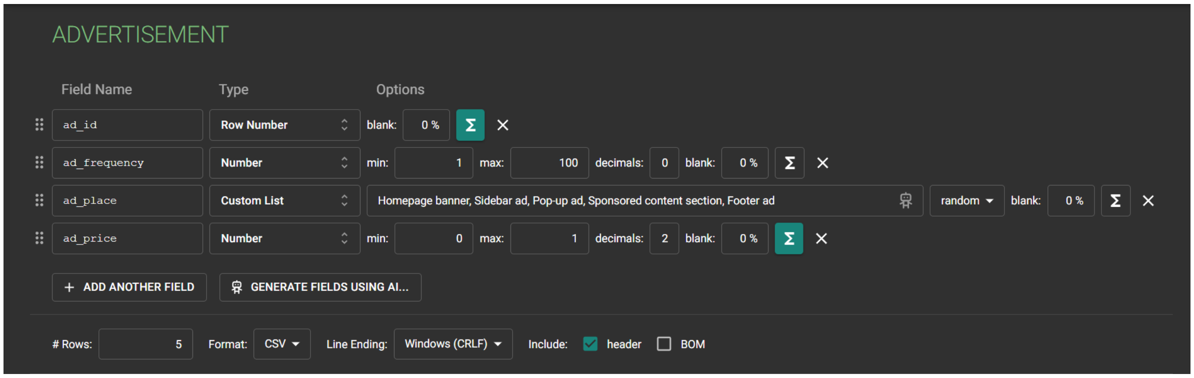
For the promo_code, the values are created using an AI generator, and the values of the discount_percent, which stores the discount amount as percentages stores, are assigned manually to a particular promo_code using this formula:



Formula

```
if promo_code == 'SUMMER20' then 20
elsif promo_code == 'SALE50' then 50
elsif promo_code == 'FALL10' then 10
elsif promo_code == 'WINTER25' then 25
elsif promo_code == 'SPRING15' then 15 end
```

3. ADVERTISEMENT

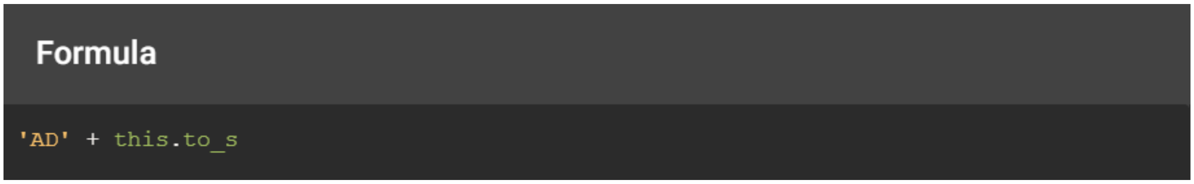


ADVERTISEMENT

Field Name	Type	Options
ad_id	Row Number	blank: 0 %
ad_frequency	Number	min: 1 · max: 100 · decimals: 0 · blank: 0 %
ad_place	Custom List	Homepage banner, Sidebar ad, Pop-up ad, Sponsored content section, Footer ad · random · blank: 0 %
ad_price	Number	min: 0 · max: 1 · decimals: 2 · blank: 0 %

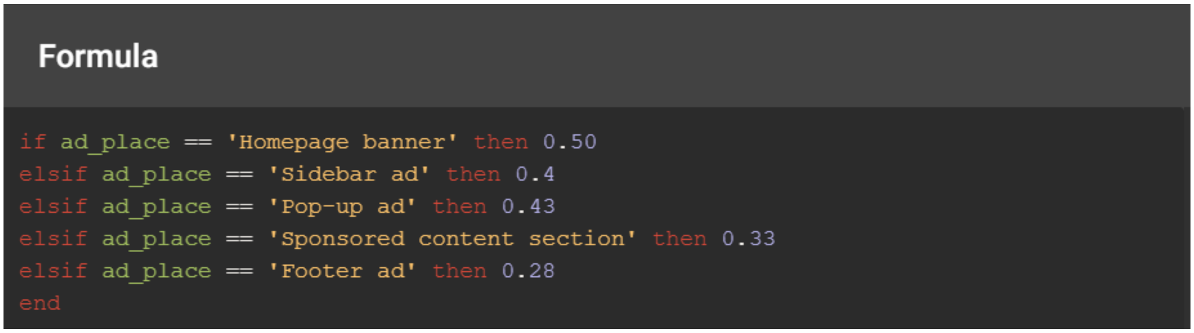
Rows: 5 · Format: CSV · Line Ending: Windows (CRLF) · Include: ☑ header □ BOM

We set the condition for the primary key, ad_id, to start with ‘AD’ followed by the row number. For the ad_place attribute, we used an AI generator to generate places where advertisements can be shown. We assumed that different places would cause different amounts of ad_price, so we manually assigned ad_price values based on different ad_place.



Formula

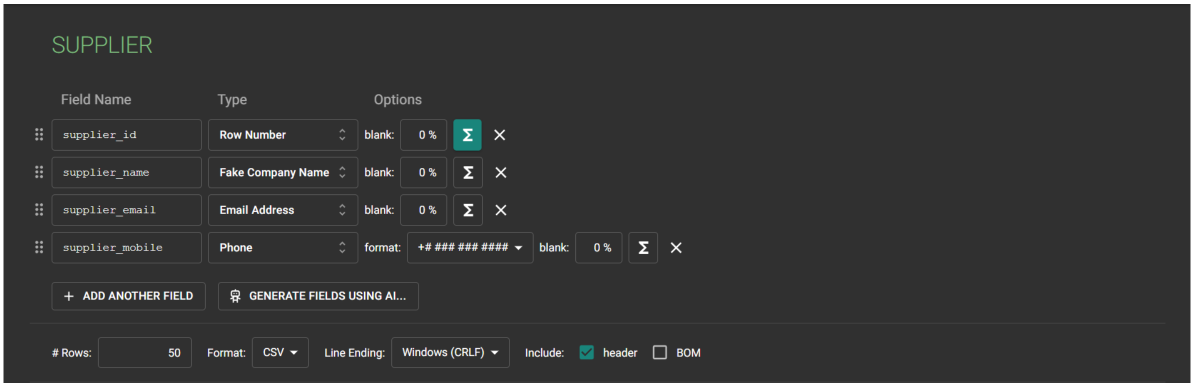
```
'AD' + this.to_s
```



Formula

```
if ad_place == 'Homepage banner' then 0.50
elsif ad_place == 'Sidebar ad' then 0.4
elsif ad_place == 'Pop-up ad' then 0.43
elsif ad_place == 'Sponsored content section' then 0.33
elsif ad_place == 'Footer ad' then 0.28
end
```

4. SUPPLIER

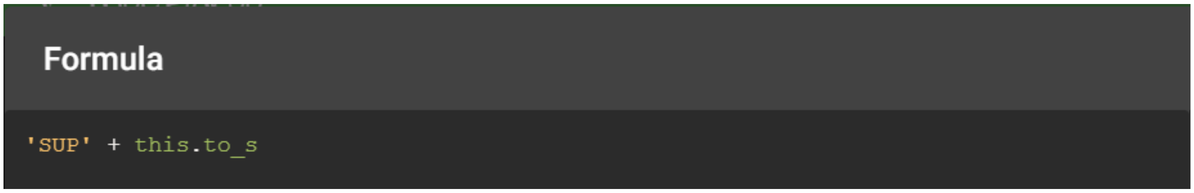


SUPPLIER

Field Name	Type	Options
supplier_id	Row Number	blank: 0 %
supplier_name	Fake Company Name	blank: 0 %
supplier_email	Email Address	blank: 0 %
supplier_mobile	Phone	format: +# ### ### #### blank: 0 %

Rows: 50 Format: CSV Line Ending: Windows (CRLF) Include: ☑ header ☐ BOM

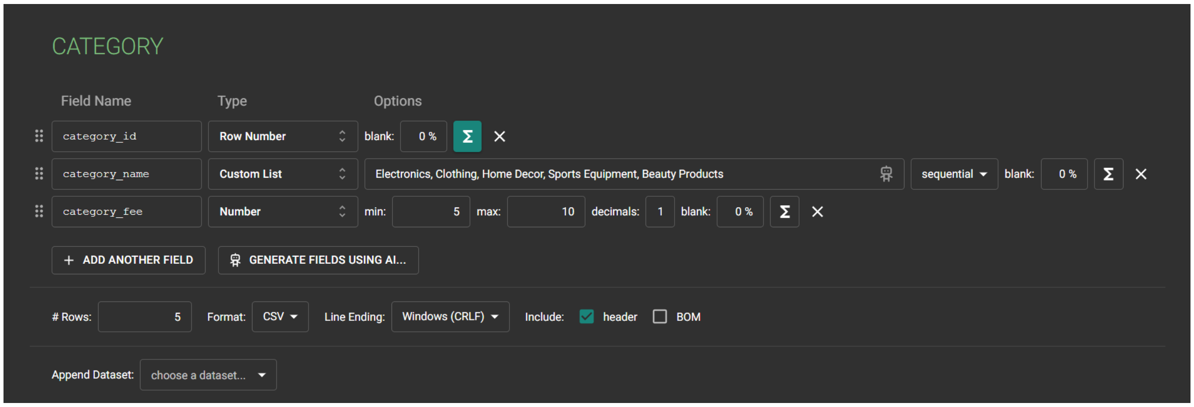
We have set the supplier_id, which is the primary of the table, to start with 'SUP' and then follow with the row number.



Formula

```
'SUP' + this.to_s
```

5. CATEGORY



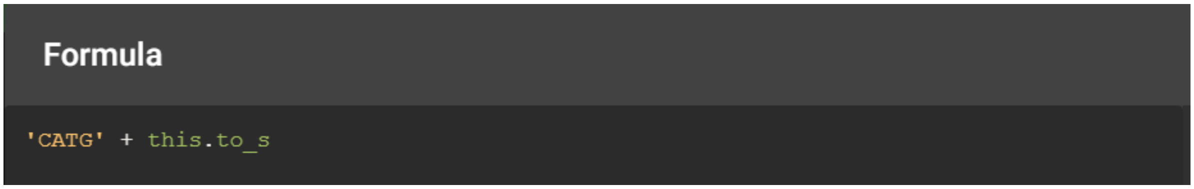
CATEGORY

Field Name	Type	Options
category_id	Row Number	blank: 0 %
category_name	Custom List	Electronics, Clothing, Home Decor, Sports Equipment, Beauty Products; sequential; blank: 0 %
category_fee	Number	min: 5 max: 10 decimals: 1 blank: 0 %

Rows: 5 Format: CSV Line Ending: Windows (CRLF) Include: ☑ header ☐ BOM

Append Dataset: choose a dataset...

We have set the category_id, which is the primary of the table, to start with ‘CATG’ and then follow with a number.

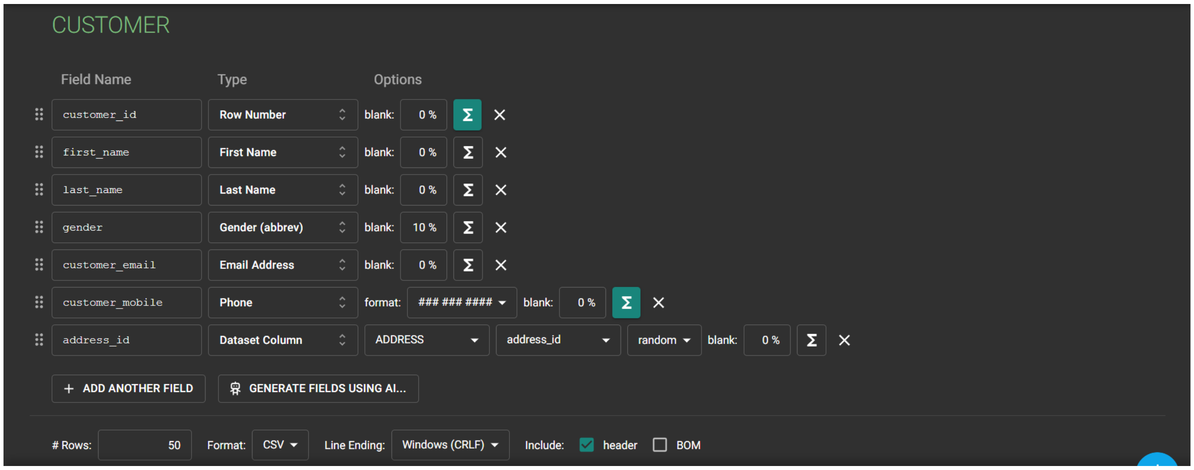


Formula

```
'CATG' + this.to_s
```

Data generation for children entities

6. CUSTOMER

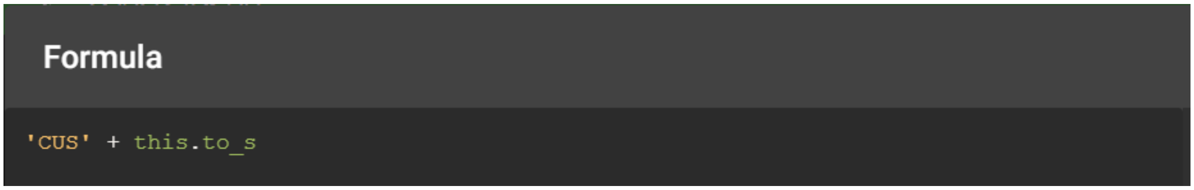


CUSTOMER

Field Name	Type	Options
customer_id	Row Number	blank: 0 %
first_name	First Name	blank: 0 %
last_name	Last Name	blank: 0 %
gender	Gender (abbrev)	blank: 10 %
customer_email	Email Address	blank: 0 %
customer_mobile	Phone	format: ### ### #### blank: 0 %
address_id	Dataset Column	ADDRESS; address_id; random; blank: 0 %

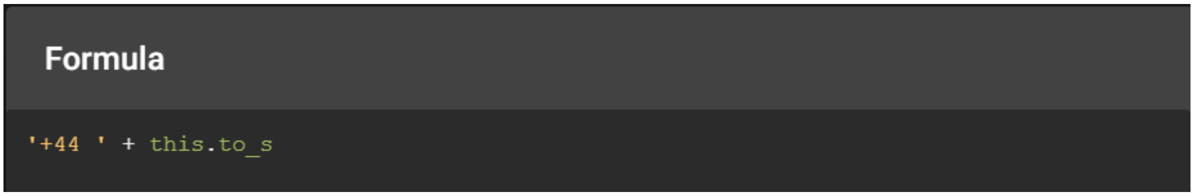
Rows: 50 Format: CSV Line Ending: Windows (CRLF) Include: ☑ header ☐ BOM

There were additional constraints that we set in the formula. Firstly, we set to condition the customer_id to start with ‘CUS’ and followed by the row number. Moreover, we gave a condition on customer_mobile to start with +44, which is the United Kingdom country code, before generating the rest of the numbers to comply with our focus on United Kingdom customers only. The formulas for those conditions are as followed



Formula

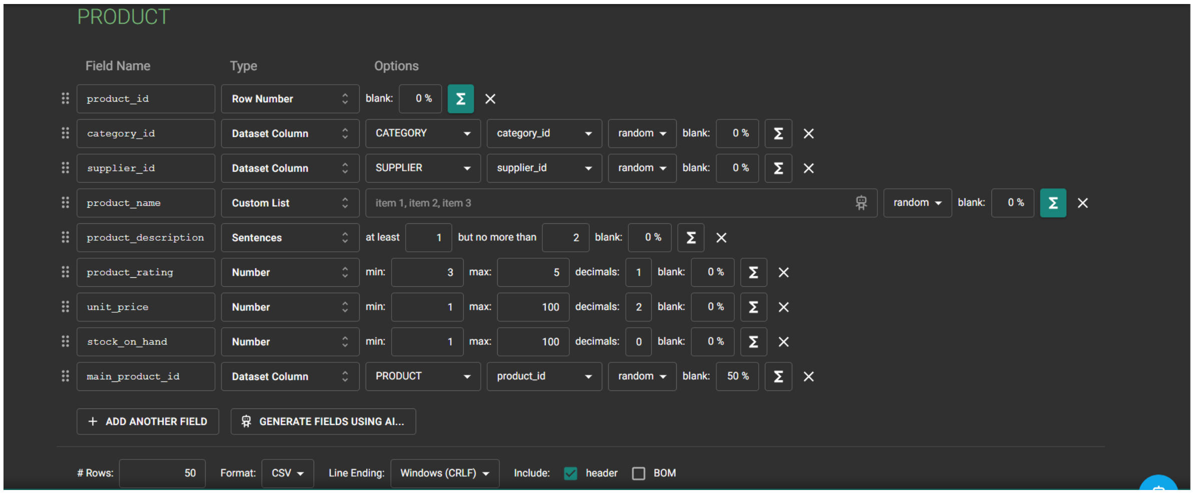
```
'CUS' + this.to_s
```



Formula

```
'+44 ' + this.to_s
```

7. PRODUCT

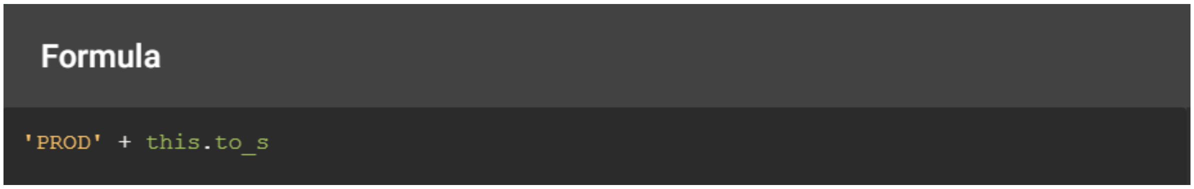


PRODUCT

Field Name	Type	Options				
product_id	Row Number	blank: 0 %				
category_id	Dataset Column	CATEGORY	category_id	random	blank: 0 %	
supplier_id	Dataset Column	SUPPLIER	supplier_id	random	blank: 0 %	
product_name	Custom List	item 1, item 2, item 3	random	blank: 0 %		
product_description	Sentences	at least 1	but no more than 2	blank: 0 %		
product_rating	Number	min: 3	max: 5	decimals: 1	blank: 0 %	
unit_price	Number	min: 1	max: 100	decimals: 2	blank: 0 %	
stock_on_hand	Number	min: 1	max: 100	decimals: 0	blank: 0 %	
main_product_id	Dataset Column	PRODUCT	product_id	random	blank: 50 %	

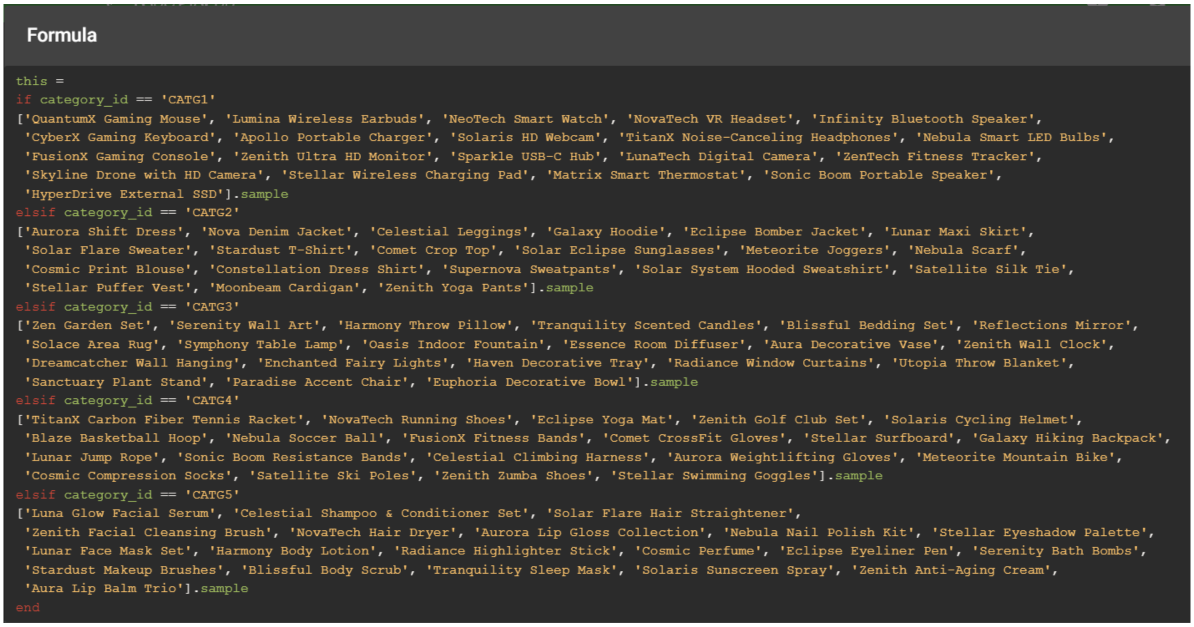
Rows: 50 Format: CSV Line Ending: Windows (CRLF) Include: ☑ header ☐ BOM

We set two conditions in this table. The first condition is on the primary key, product_id, to start the value with 'PROD' and then athe row number. Next, to ensure consistency between the products and categories, we manually assigned the values of product_name to the categories they were supposed to belong to.



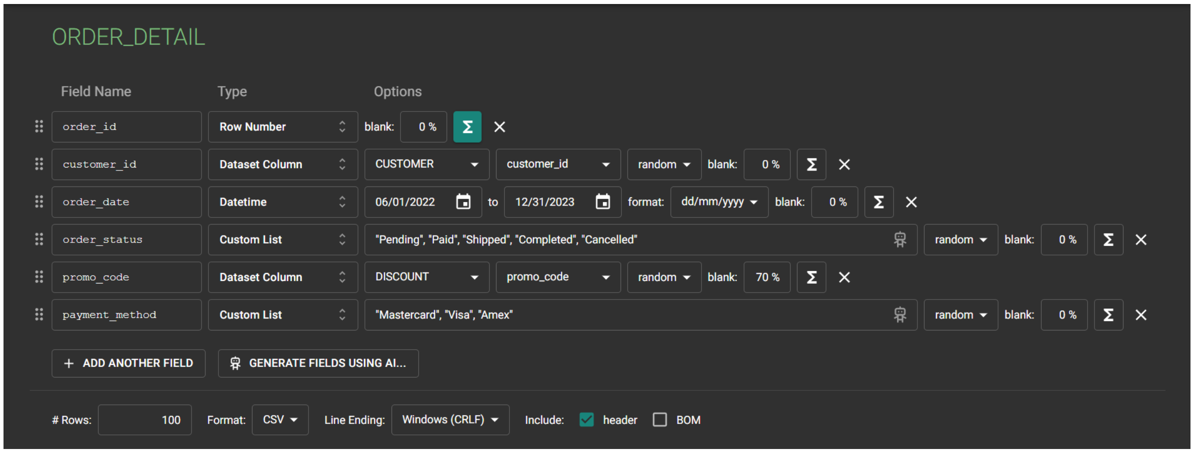
Formula

```
'PROD' + this.to_s
```



```
this =
if category_id == 'CATG1'
['QuantumX Gaming Mouse', 'Lumina Wireless Earbuds', 'NeoTech Smart Watch', 'NovaTech VR Headset', 'Infinity Bluetooth Speaker',
 'CyberX Gaming Keyboard', 'Apollo Portable Charger', 'Solaris HD Webcam', 'TitanX Noise-Canceling Headphones', 'Nebula Smart LED Bulbs',
 'FusionX Gaming Console', 'Zenith Ultra HD Monitor', 'Sparkle USB-C Hub', 'LunaTech Digital Camera', 'ZenTech Fitness Tracker',
 'Skyline Drone with HD Camera', 'Stellar Wireless Charging Pad', 'Matrix Smart Thermostat', 'Sonic Boom Portable Speaker',
 'HyperDrive External SSD'].sample
elsif category_id == 'CATG2'
['Aurora Shift Dress', 'Nova Denim Jacket', 'Celestial Leggings', 'Galaxy Hoodie', 'Eclipse Bomber Jacket', 'Lunar Maxi Skirt',
 'Solar Flare Sweater', 'Stardust T-Shirt', 'Comet Crop Top', 'Solar Eclipse Sunglasses', 'Meteorite Joggers', 'Nebula Scarf',
 'Cosmic Print Blouse', 'Constellation Dress Shirt', 'Supernova Sweatpants', 'Solar System Hooded Sweatshirt', 'Satellite Silk Tie',
 'Stellar Puffer Vest', 'Moonbeam Cardigan', 'Zenith Yoga Pants'].sample
elsif category_id == 'CATG3'
['Zen Garden Set', 'Serenity Wall Art', 'Harmony Throw Pillow', 'Tranquility Scented Candles', 'Blissful Bedding Set', 'Reflections Mirror',
 'Solace Area Rug', 'Symphony Table Lamp', 'Oasis Indoor Fountain', 'Essence Room Diffuser', 'Aura Decorative Vase', 'Zenith Wall Clock',
 'Dreamcatcher Wall Hanging', 'Enchanted Fairy Lights', 'Haven Decorative Tray', 'Radiance Window Curtains', 'Utopia Throw Blanket',
 'Sanctuary Plant Stand', 'Paradise Accent Chair', 'Euphoria Decorative Bowl'].sample
elsif category_id == 'CATG4'
['TitanX Carbon Fiber Tennis Racket', 'NovaTech Running Shoes', 'Eclipse Yoga Mat', 'Zenith Golf Club Set', 'Solaris Cycling Helmet',
 'Blaze Basketball Hoop', 'Nebula Soccer Ball', 'FusionX Fitness Bands', 'Comet CrossFit Gloves', 'Stellar Surfboard', 'Galaxy Hiking Backpack',
 'Lunar Jump Rope', 'Sonic Boom Resistance Bands', 'Celestial Climbing Harness', 'Aurora Weightlifting Gloves', 'Meteorite Mountain Bike',
 'Cosmic Compression Socks', 'Satellite Ski Poles', 'Zenith Zumba Shoes', 'Stellar Swimming Goggles'].sample
elsif category_id == 'CATG5'
['Luna Glow Facial Serum', 'Celestial Shampoo & Conditioner Set', 'Solar Flare Hair Straightener',
 'Zenith Facial Cleansing Brush', 'NovaTech Hair Dryer', 'Aurora Lip Gloss Collection', 'Nebula Nail Polish Kit', 'Stellar Eyeshadow Palette',
 'Lunar Face Mask Set', 'Harmony Body Lotion', 'Radiance Highlighter Stick', 'Cosmic Perfume', 'Eclipse Eyeliner Pen', 'Serenity Bath Bombs',
 'Stardust Makeup Brushes', 'Blissful Body Scrub', 'Tranquility Sleep Mask', 'Solaris Sunscreen Spray', 'Zenith Anti-Aging Cream',
 'Aura Lip Balm Trio'].sample
end
```

8. ORDER_DETAIL

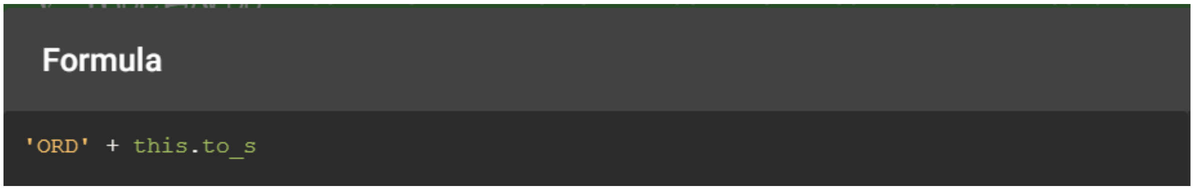


ORDER_DETAIL

Field Name	Type	Options
order_id	Row Number	blank: 0 %
customer_id	Dataset Column	CUSTOMER · customer_id · random · blank: 0 %
order_date	Datetime	06/01/2022 to 12/31/2023 · format: dd/mm/yyyy · blank: 0 %
order_status	Custom List	"Pending", "Paid", "Shipped", "Completed", "Cancelled" · random · blank: 0 %
promo_code	Dataset Column	DISCOUNT · promo_code · random · blank: 70 %
payment_method	Custom List	"Mastercard", "Visa", "Amex" · random · blank: 0 %

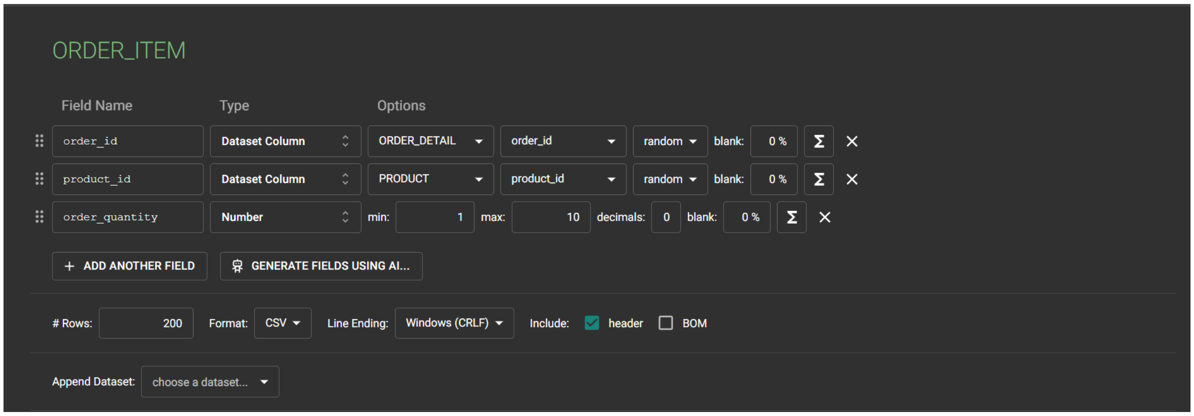
Rows: 100 Format: CSV Line Ending: Windows (CRLF) Include: ☑ header ☐ BOM

We have set the order_id, which is the primary of the table, to start with 'ORD' and then follow with the row number.



```
'ORD' + this.to_s
```

9. ORDER_ITEM



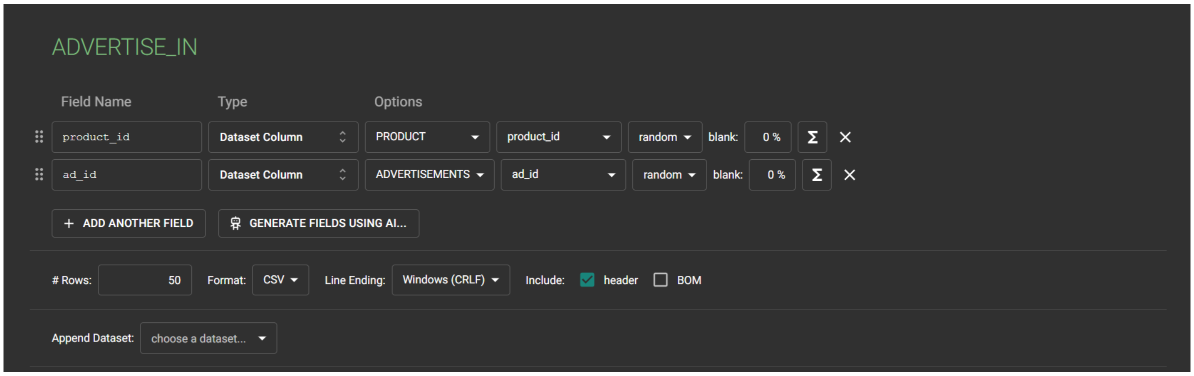
ORDER_ITEM

Field Name	Type	Options							
order_id	Dataset Column	ORDER_DETAIL	order_id	random	blank: 0 %				
product_id	Dataset Column	PRODUCT	product_id	random	blank: 0 %				
order_quantity	Number	min: 1	max: 10	decimals: 0	blank: 0 %				

Rows: 200 Format: CSV Line Ending: Windows (CRLF) Include: ☑ header ☐ BOM

Append Dataset: choose a dataset...

10. ADVERTISE_IN



ADVERTISE_IN

Field Name	Type	Options			
product_id	Dataset Column	PRODUCT	product_id	random	blank: 0 %
ad_id	Dataset Column	ADVERTISEMENTS	ad_id	random	blank: 0 %

Rows: 50 Format: CSV Line Ending: Windows (CRLF) Include: ☑ header ☐ BOM

Append Dataset: choose a dataset...

Task 2.2: Data Import and Quality Assurance

Data Import

Loading all generated data sets using read_csv to each entity

```r
read_csv_files <- function(directory) {
  csv_files <- list.files(directory, pattern = "\\.csv$", full.names = TRUE)
  data_frames <- list()

  for (csv_file in csv_files) {
    # Extract the base file name without extension
    file_name <- tools::file_path_sans_ext(basename(csv_file))

    stem <- gsub("\\d+$", "", file_name)

    # Read the CSV file
    data <- read.csv(csv_file)

    # Check if data frame already exists
    if (stem %in% names(data_frames)) {
      # If exists, append data to existing data frame
      data_frames[[stem]] <- rbind(data_frames[[stem]], data)
    } else {
      # Otherwise, create new data frame
      data_frames[[stem]] <- data
    }
  }
  return(data_frames)
}

## Read csv files and assign to "data_frames"
data_frames <- read_csv_files("data_upload")
```

Quality Assurance

The imported data sets were validated before being added to the database to ensure they aligned with the nature of the keys and data type specified in the physical schema.

The following data quality checks were performed in all entities included:

1. Check if the values of the primary key are unique

```
# Check for duplicate key in each table
##List primary key for each table
primary_keys <- list(
  "ADDRESS" = c("address_id"),
  "DISCOUNT" = c("promo_code"),
  "SUPPLIER" = c("supplier_id"),
  "CATEGORY" = c("category_id"),
  "ADVERTISEMENT" = c("ad_id"),
  "CUSTOMER" = c("customer_id"),
  "PRODUCT" = c("product_id"),
  "ADVERTISE_IN" = c("ad_id", "product_id"),
  "ORDER_DETAIL" = c("order_id"),
  "ORDER_ITEM" = c("order_id","product_id")
)

## Function to check for duplicate primary key and remove duplicates
remove_duplicate_primary_keys <- function(df, primary_key_cols) {
  unique_rows <- !duplicated(df[, primary_key_cols])
  return(df[unique_rows, ])
}


## Apply the function for each data frame
for (table_name in names(data_frames)) {
  if (table_name %in% names(primary_keys)) {
    data_frames[[table_name]] <- remove_duplicate_primary_keys(data_frames[[table_name]],
  }
}
```

2. Check if there is any missing data in the attributes that have NOT NULL constraint

```
# Check missing data
## Function to check for missing values in columns except those specified to skip
check_missing_values <- function(df, columns_to_skip, table_name) {
  columns <- setdiff(names(df), columns_to_skip)
  missing <- sapply(df[columns], function(x) any(is.na(x) | x == ""))
  if (table_name == "ORDER_DETAIL") {
    missing["promo_code"] <- FALSE  # Allow missing values in "promo_code" for "ORDER_DETA
  }
  return(missing)
}

## Columns to skip
columns_to_skip <- c("gender", "product_description", "product_rating", "main_product_id")

## Apply the function for each data frame
for (table_name in names(data_frames)) {
  if (table_name %in% names(primary_keys)) {
    missing_values <- check_missing_values(data_frames[[table_name]], columns_to_skip, tab
    if (any(missing_values)) {
      empty_row_indices <- apply(data_frames[[table_name]][, -which(names(data_frames[[tab
      data_frames[[table_name]] <- data_frames[[table_name]][!empty_row_indices, ]
    }
  }
}
```

3. Check if there are any values of a foreign key in the children table that do not exist in the parent table

```r
# Foreign key check
##CUSTOMER - ADDRESS
invalid_address_fk <- data_frames$CUSTOMER[!data_frames$CUSTOMER$address_id %in% data_fram
data_frames$CUSTOMER <- data_frames$CUSTOMER[data_frames$CUSTOMER$address_id %in% data_fra

##PRODUCT - CATEGORY
invalid_category_fk <- data_frames$PRODUCT[!data_frames$PRODUCT$category_id %in% data_fram
data_frames$PRODUCT <- data_frames$PRODUCT[data_frames$PRODUCT$category_id %in% data_frame

##PRODUCT - SUPPLIER
invalid_supplier_fk <- data_frames$PRODUCT[!data_frames$PRODUCT$supplier_id %in% data_fram
data_frames$PRODUCT <- data_frames$PRODUCT[data_frames$PRODUCT$supplier_id %in% data_frame

##ORDER_DETAIL - CUSTOMER
invalid_customer_fk <- data_frames$ORDER_DETAIL[!data_frames$ORDER_DETAIL$customer_id %in%
data_frames$ORDER_DETAIL <- data_frames$ORDER_DETAIL[data_frames$ORDER_DETAIL$customer_id

##ORDER_DETAIL - DISCOUNT
discounted_order <- data_frames$ORDER_DETAIL[!is.na(data_frames$ORDER_DETAIL$promo_code) &
invalid_promo_fk <- discounted_order[!discounted_order$promo_code %in% data_frames$DISCOUN
data_frames$ORDER_DETAIL <- data_frames$ORDER_DETAIL[is.na(data_frames$ORDER_DETAIL$promo_
```

```
Warning in is.na(data_frames$ORDER_DETAIL$promo_code) |
trimws(data_frames$ORDER_DETAIL$promo_code) == : longer object length is not a
multiple of shorter object length
```

```r
##ORDER_ITEM - PRODUCT
invalid_product_fk <- data_frames$ORDER_ITEM[!data_frames$ORDER_ITEM$product_id %in% data_
data_frames$ORDER_ITEM <- data_frames$ORDER_ITEM[data_frames$ORDER_ITEM$product_id %in% da

##ORDER_ITEM - ORDER_DETAIL
invalid_order_fk       <- data_frames$ORDER_ITEM[!data_frames$ORDER_ITEM$order_id %in% dat
data_frames$ORDER_ITEM <- data_frames$ORDER_ITEM[ data_frames$ORDER_ITEM$order_id %in% dat

##ADVERTISE_IN - ADVERTISEMENT
invalid_ad_fk <- data_frames$ADVERTISE_IN[!data_frames$ADVERTISE_IN$ad_id %in% data_frames
data_frames$ADVERTISE_IN <- data_frames$ADVERTISE_IN[data_frames$ADVERTISE_IN$ad_id %in% d

##ADVERTISE_IN - PRODUCT
invalid_adproduct_fk <- data_frames$ADVERTISE_IN[!data_frames$ADVERTISE_IN$product_id %in%
data_frames$ADVERTISE_IN <- data_frames$ADVERTISE_IN[data_frames$ADVERTISE_IN$product_id %
```

4. CUSTOMER: Remove customer name Remove name that does not match pattern

```r
# Check name format
## Function to check for invalid name format in columns
check_name_format <- function(df) {
  invalid_format <- rep(FALSE, nrow(df))
  for (i in 1:nrow(df)) {
    if (!grepl("^[A-Z][a-z]*$", df[i, "first_name"]) || !grepl("^[A-Z][a-z]*$", df[i, "las
      invalid_format[i] <- TRUE
    }
  }
  return(invalid_format)
}

## Apply the function for the "customer" table
if ("CUSTOMER" %in% names(data_frames)) {
  customer_df <- data_frames$CUSTOMER
  if (nrow(customer_df) > 0) {
    invalid_format <- check_name_format(customer_df)
    if (any(invalid_format)) {
      data_frames$CUSTOMER <- customer_df[!invalid_format, ]
    }
  }
}
```

5. PRODUCT/CATEGORY/SUPPLIER: Remove name that does not match pattern

```r
# Check product/category/supplier name
## Function to keep records if the specified column matches the regex pattern
naming_format <- function(df, column_to_check, regex_pattern) {
  matching_records <- grepl(regex_pattern, df[[column_to_check]], perl = TRUE)
  return(df[matching_records, ])
}

## Define regex pattern
regex_pattern <- "^[A-Za-z,.-]+( [A-Za-z,.-]+)*$"

## Apply the function for the specified columns in the respective tables
if ("CATEGORY" %in% names(data_frames)) {
  data_frames[["CATEGORY"]] <- naming_format(data_frames[["CATEGORY"]], "category_name", r
}
```

```
if ("SUPPLIER" %in% names(data_frames)) {
  data_frames[["SUPPLIER"]] <- naming_format(data_frames[["SUPPLIER"]], "supplier_name", r
}

if ("PRODUCT" %in% names(data_frames)) {
  data_frames[["PRODUCT"]] <- naming_format(data_frames[["PRODUCT"]], "product_name", rege
}
```

6. CUSTOMER/SUPPLIER: Remove email that does not match pattern and duplicate emails

```
# Email check
##Define a function to check email format
check_email_format <- function(email) {
  grepl("^\\S+@\\S+\\.\\S+$", email)
}

##Check email format for CUSTOMER table
if ("CUSTOMER" %in% names(data_frames)) {
  invalid_emails_customer <- !sapply(data_frames$CUSTOMER$customer_email, check_email_form
  if (any(invalid_emails_customer)) {
    data_frames$CUSTOMER <- data_frames$CUSTOMER[!invalid_emails_customer, ]
    data_frames$CUSTOMER <- data_frames$CUSTOMER[!duplicated(data_frames$CUSTOMER$customer
  }
}

##Check email format for SUPPLIER table
if ("SUPPLIER" %in% names(data_frames)) {
  invalid_emails_supplier <- !sapply(data_frames$SUPPLIER$supplier_email, check_email_form
  if (any(invalid_emails_supplier)) {
    data_frames$SUPPLIER <- data_frames$SUPPLIER[!invalid_emails_supplier, ]
    data_frames$SUPPLIER <- data_frames$SUPPLIER[!duplicated(data_frames$SUPPLIER$supplier
  }
}
```

7. CUSTOMER/SUPPLIER: Remove mobile that does not match pattern and duplicate emails

```
# Mobile number check
##Define a function to check mobile format
check_mobile_format <- function(mobile) {
  grepl("^\\+\\d{1,3}\\s[0-9]{3}\\s[0-9]{3}\\s[0-9]{4}$", mobile)
}

##Check mobile format for CUSTOMER table
if ("CUSTOMER" %in% names(data_frames)) {
  invalid_mobiles_customer <- !sapply(data_frames$CUSTOMER$customer_mobile, check_mobile_f
  if (any(invalid_mobiles_customer)) {
    data_frames$CUSTOMER <- data_frames$CUSTOMER[!invalid_mobiles_customer, ]
    data_frames$CUSTOMER <- data_frames$CUSTOMER[!duplicated(data_frames$CUSTOMER$customer
  }
}

##Check mobile format for SUPPLIER table
if ("SUPPLIER" %in% names(data_frames)) {
  invalid_mobiles_supplier <- !sapply(data_frames$SUPPLIER$supplier_mobile, check_mobile_f
  if (any(invalid_mobiles_supplier)) {
    data_frames$SUPPLIER <- data_frames$SUPPLIER[!invalid_mobiles_supplier, ]
    data_frames$SUPPLIER <- data_frames$SUPPLIER[!duplicated(data_frames$SUPPLIER$supplier
  }
}
```

8. PRODUCT: Remove product with invalid main_product_id

```
# Check self reference product
invalid_main_product <- data_frames$PRODUCT[!is.na(data_frames$PRODUCT$main_product_id) &
                                              !(data_frames$PRODUCT$main_product_id %in% d
                                                  data_frames$PRODUCT$main_product_id != d
                                              trimws(data_frames$PRODUCT$main_product_id)
data_frames$PRODUCT <- data_frames$PRODUCT[!row.names(data_frames$PRODUCT) %in% row.names(
```

Once the dataset was validated, it checks if records exits in database before writing into the database. Only new records that does not match will be written.

```
# Define table names
table_names <- c("CUSTOMER", "PRODUCT", "ADDRESS", "DISCOUNT", "SUPPLIER",
                 "CATEGORY", "ADVERTISEMENT", "ADVERTISE_IN", "ORDER_ITEM", "ORDER_DETAIL"

# Connect to the database
connect <- dbConnect(RSQLite::SQLite(), "database.db")

# Iterate over table names
```

```
for (table_name in table_names) {
  # Get data frame from data_frames list
  df <- data_frames[[table_name]]

  cat("Total records in", table_name, ":", nrow(df), "\n")

  # Retrieve existing records from the database
  db_data <- dbGetQuery(connect, paste("SELECT * FROM", table_name))

  # Compare and find new records
  new_records <- df[!df[, 1] %in% db_data[, 1], ]

  # Write new records to the database
  if (nrow(new_records) > 0) {
    dbWriteTable(connect, table_name, new_records, append = TRUE, row.names = FALSE)
  }
}
```

```
Total records in CUSTOMER : 46
Total records in PRODUCT : 48
Total records in ADDRESS : 50
Total records in DISCOUNT : 5
Total records in SUPPLIER : 50
Total records in CATEGORY : 5
Total records in ADVERTISEMENT : 5
Total records in ADVERTISE_IN : 50
Total records in ORDER_ITEM : 198
Total records in ORDER_DETAIL : 100
```

```
# Disconnect from the database
dbDisconnect(connect)
```

Part 3: Data Pipeline Generation

Task 3.1: GitHub Repository and Workflow Setup

Repository named **DM_Group3** was created on GitHub with public access and following files were created:

1. **README:** Includes all student ID numbers of our group.

2. **database_schema.R :** R Script to create a database and build the database schema.
3. **data_validation_and_load.R:** R Script for data validation and data injection into the database.
4. **data_analysis.R:** R script to perform data analysis for necessary data visualizations
5. **data_upload and images:** Folders to store 10 datasets generated by Mockaroo, and images to store E-R diagrams.
6. **Report:** Quarto Markdown file to report the execution of tasks.

New project was created in RStudio for each user. Github was used for version control to track changes, and collaboration between users.

Github actions were utilized to automate the entire ETL workflow process such as running tests, building, and deploying tasks after pushing code, which can improve development efficiency and ensure code stability

Workflow file **etl.yaml** was created in the **.github/workflows** directory in **DM_Group3** repository.

Task 3.2: GitHub Actions for Continuous Integration

Workflow was designed to execute a task named "ETL workflow for group 3", after a trigger by pushing events to the 'main' branch. Sequence of steps were executed on the latest version of Ubuntu:

1. Use GitHub Actions checkout action to check out code into the working directory.
2. Set up R environment and specify the R version as '4.3.3'.
3. Cache R packages to avoid reinstalling them on each run.
4. Install required packages if the cache misses.
5. Run database_schema.R to render database schema.
6. Run data_validation_and_load.R to render validation and load data into database.
7. Run data_analysis.R to render data analysis.
8. Utilize an authentication token to push changes to the 'main' branch.

Part 4: Data Analysis and Reporting with Quarto in R

Task 4.1: Advanced Data Analysis in R

```r
# Connecting to the database
connect <- dbConnect(RSQLite::SQLite(), "database.db")
```

```r
# Report 1

# Top 10 products based on the Profit Generated

# Getting the tables from the database
product <- RSQLite::dbGetQuery(connect,'SELECT * FROM PRODUCT')
order_item <- RSQLite::dbGetQuery(connect,'SELECT * FROM ORDER_ITEM')
order_detail <- RSQLite::dbGetQuery(connect,'SELECT * FROM ORDER_DETAIL')
discount <- RSQLite::dbGetQuery(connect,'SELECT * FROM DISCOUNT')
# Creating a profit table that includes a total_profit column for each product
#incorporating the discount percent
#filtering out all "cancelled" orders.

profit_data <- order_item %>%
  inner_join(order_detail, by = "order_id") %>%
  filter(order_status != "Cancelled") %>%
  inner_join(product, by = "product_id") %>%
  left_join(discount, by = c("promo_code" = "promo_code")) %>%
  mutate(
    discount_percentage = ifelse(is.na(discount_percent), 0, discount_percent),
    total_profit = (order_quantity * unit_price) *
      (1 - discount_percentage / 100)
  ) %>%
  group_by(product_id, product_name) %>%
  summarise(
    total_profit = sum(total_profit),
    .groups = 'drop'
  ) %>%
  arrange(desc(total_profit))

# Selecting the top 10 profitable products
top_10_profit_data <- head(profit_data, 10)


# Visualizing the results
ggplot(top_10_profit_data, aes(x = reorder(product_id, total_profit),
                               y = total_profit,
                               fill = product_name)) +
  geom_bar(stat = "identity") +
  scale_fill_discrete(name = "product Name") +
  labs(title = "Top 10 Most Profitable Products", x = "Product ID",
       y = "Total Profit", caption="Figure 1") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.title = element_text(size = 12),
        legend.text = element_text(size = 10),
        plot.title = element_text(hjust = 0.5))
```

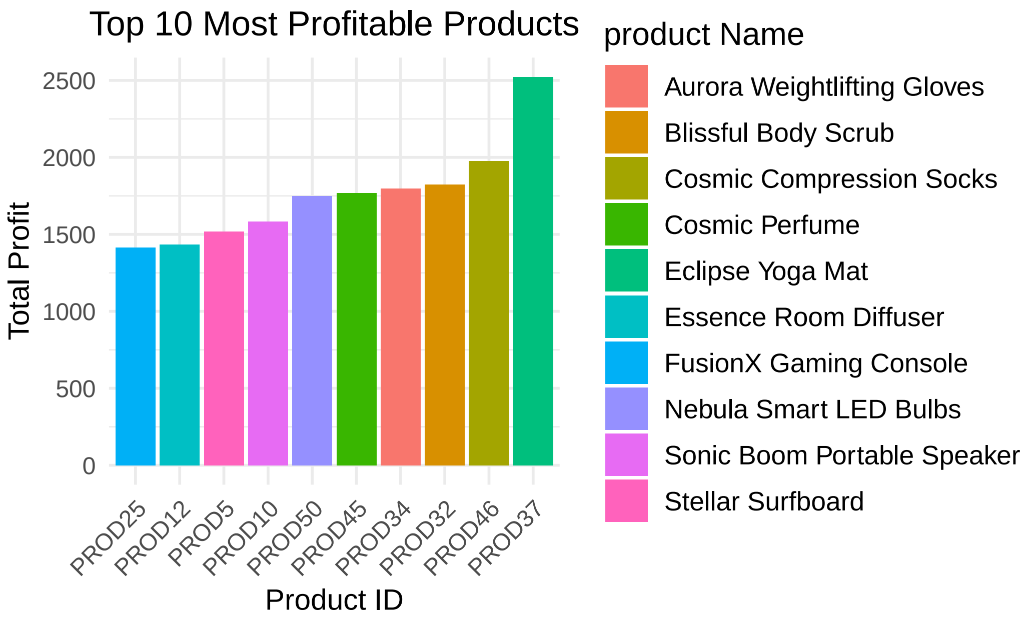


Figure 1

```
# Top Most Selling Products

# Creating a sales that that includes the total quantity sold for each product
# selecting the top 10.
sales_data <- order_item %>%
  group_by(product_id) %>%
  summarise(Total_Sales = sum(order_quantity, na.rm = TRUE)) %>%
  arrange(desc(Total_Sales))

top_selling_products <- sales_data %>%
  left_join(product, by = "product_id") %>%
  select(product_id, product_name, Total_Sales) %>%
  top_n(10, Total_Sales)
```

```
# Visualizing the results
ggplot(top_selling_products, aes(x = reorder(product_id, Total_Sales),
                                 y = Total_Sales, fill = product_name)) +
  geom_bar(stat = "identity") +
  scale_fill_discrete(name = "product Name") +
  labs(title = "Top Most Selling Products", x = "Product ID",
       y = "Total Sales", caption="Figure 2") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.title = element_text(size = 12),
        legend.text = element_text(size = 10),
        plot.title = element_text(hjust = 0.5))
```

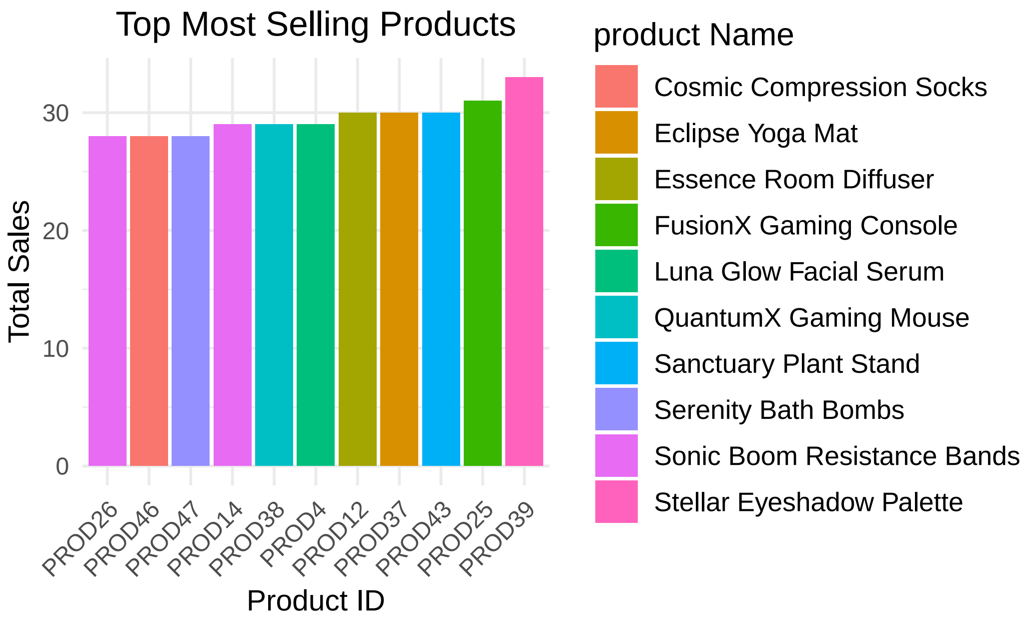


Figure 2

Figure 1 represents the profitability of individual products. The “Eclipse Yoga Mat” generates the highest profit for the company indicating a higher margin or a premium pricing strategy. Conversely, Figure 2 illustrates sales volume, where items like the “Stellar Eyeshadow Palette” and “Fusion X Gaming Console” lead, suggesting they are popular among customers. Some products are both popular and profitable implying a successful product strategy. However, there are also noticeable discrepancies. For example, “QuantumX Gaming Mouse” appears as a top seller, yet it’s absent from the most profitable items, implying a lower profit margin. In contrast, “Nebula Smart LED Bulbs” are among the most profitable but not the top sellers, suggesting a high margin compensating for lesser sales.

The e-commerce company could focus on marketing strategies for high-margin products to boost their volume of sales, increasing overall profitability. For products that sell well but are less profitable, the company may need to assess whether they can improve margins through better supplier negotiations.

Products that are both top sellers and highly profitable should be kept in optimum stock to avoid lost sales opportunities. For less profitable items, the company may consider keeping lower stock levels or discontinuing them if they do not contribute significantly to overall profits.

```r
#Report 2

# Bundles vs Individual products Comparison by Sales Volume

# Join product with order item
product_order_item <- order_item %>%
  inner_join(product, by = "product_id")

# Calculate total quantity sold for bundled products
bundled_sales <- product_order_item %>%
  filter(!is.na(main_product_id)) %>%
  group_by(main_product_id) %>%
  summarise(Total_Quantity_Sold = sum(order_quantity, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(Product_Type = "Bundled")

# Calculate total quantity sold for individual products
individual_sales <- product_order_item %>%
  filter(is.na(main_product_id)) %>%
  group_by(product_id) %>%
  summarise(Total_Quantity_Sold = sum(order_quantity, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(Product_Type = "Individual")

# Combining the results
total_sales_by_type <- bind_rows(bundled_sales, individual_sales) %>%
  group_by(Product_Type) %>%
  summarise(Total_Quantity_Sold = sum(Total_Quantity_Sold, na.rm = TRUE)) %>%
  ungroup()

print(total_sales_by_type)
```

```
# A tibble: 1 x 2
```

```
  Product_Type Total_Quantity_Sold
  <chr>                      <int>
1 Bundled                     1043
```

```
# Visualizing the results
ggplot(total_sales_by_type, aes(x = Product_Type, y = Total_Quantity_Sold,
                                fill = Product_Type)) +
  geom_bar(stat = "identity", width = 0.5) +  # Adjust width here
  scale_fill_manual(values = c("Bundled" = "skyblue",
                               "Individual" = "salmon")) +
  labs(title = "Total Quantity Sold by Product Type",
       x = "Product Type",
       y = "Total Quantity Sold") +
  theme_minimal() +
  theme(legend.position = "none")
```

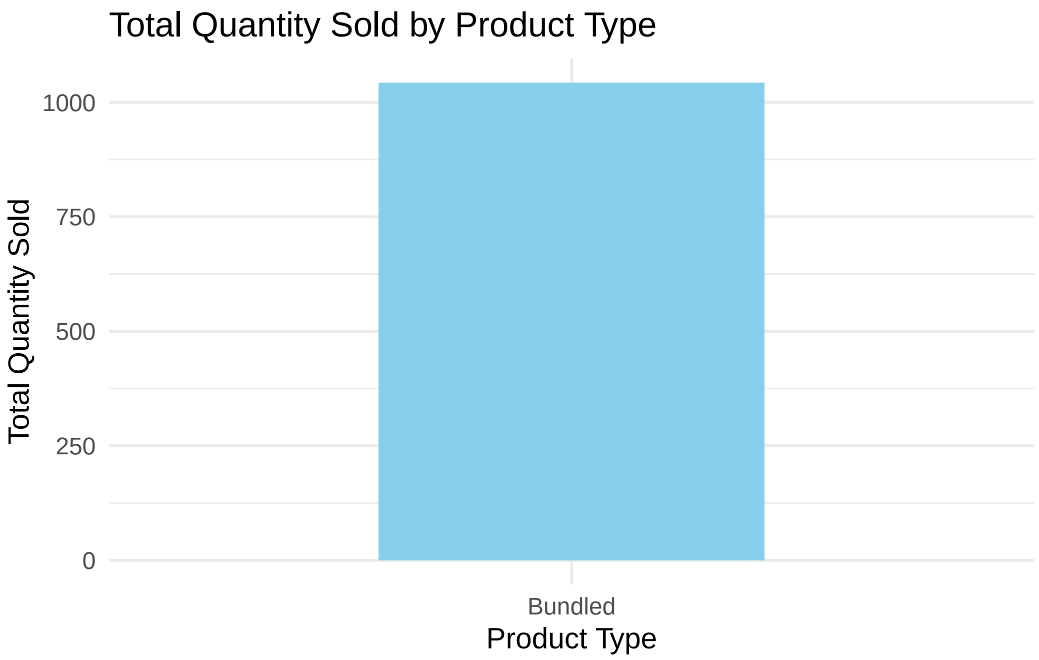


Figure 3

As the table shows, the total sales for individual products (products sold without bundles) and bundled products yield nearly equal quantities sold. This suggests that the e-commerce company's bundling strategy to boost sales was not as effective. However, it is difficult to conclude the efficacy of this deal based solely on the volume of sales.

Optimizing bundle offerings by analyzing and aligning them with customer preferences based on product ratings can significantly enhance their appeal. Experimenting with different bundle types and conducting a cost-benefit analysis might uncover ways to refine the strategy. The results (Figure 3) suggest a market demand for both types of products, indicating that customers prefer flexibility in their buying options. Rather than discontinuing the bundle products, the company should look deeper into the performance metrics of this deal, considering the inventory turnover and profit margins.

```
#Report 3

#SQL Query to get the sales over a time period

time_period_sales<- RSQLite::dbGetQuery(connect,'SELECT ORDITM.order_id,
                                        order_quantity,order_date FROM
                                        ORDER_DETAIL ORDDET INNER JOIN
                                        ORDER_ITEM ORDITM ON ORDITM.order_id =
                                        ORDDET.order_id')
#test

#Conversion of order_date field to appropiate format
time_period_sales$order_date <- as.Date(time_period_sales$order_date,
                                        format="%d/%m/%Y")

time_period_sales$order_date_mnth_yr <- format(
  time_period_sales$order_date,'%Y/%m')

#Conversion of order_date field to factors with levels

time_period_sales$order_date_mnth_yr <- factor(
  time_period_sales$order_date_mnth_yr, levels=c('2022/06',
'2022/07','2022/08',
'2022/09','2022/10','2022/11','2022/12',
'2023/01','2023/02','2023/03','2023/04',
'2023/05','2023/06','2023/07','2023/08'
,'2023/09','2023/10','2023/11','2023/12'))

#Group by sales for each year
time_period_sales_group_by<- time_period_sales %>%
group_by(order_date_mnth_yr)%>%
summarise(quantity=sum(order_quantity))

# Time series graph to show the quantity sold over a given time period
```

```
(ggplot(time_period_sales_group_by, aes(x = order_date_mnth_yr,
y = quantity,group=1)) +
geom_point()+geom_line()+
xlab('Time Period(year/month)')+ylab('Quantity Sold'))+
theme(axis.text.x=element_text(angle=45,hjust = 1))
```

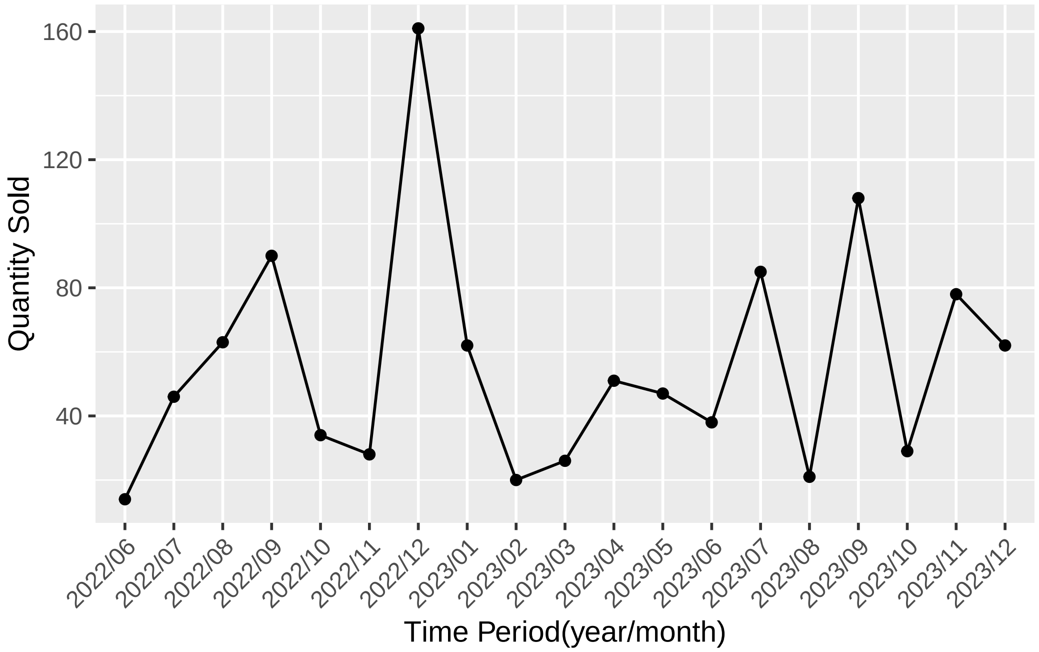


Figure 4: Time Series Graph to depict the seasonal trends in sales

Figure 4 helps us understand the purchase patterns of the customers over different periods of time. Purchase pattern analysis is significant for an e-commerce company as different strategies could be deployed to maximize sales and revenue. Figure 1 illustrates that the sales were maximum in September and December months over the time period from 2022 to 2023. Hence, appropriate pricing strategies could be deployed to increase revenue from top-selling products. Also, a sharp decline in sales was identified in October, which was later followed by an increase in succeeding months. This study could help maintain efficient inventory management for perishable products. Overall, this variation in sales across different seasons provides important insights for better management decisions.

```
#Report 4:

#SQL query to join customer and order tables to get the sales for each city
```

```
cust_order_join<- RSQLite::dbGetQuery(connect,
                                      'SELECT city,ORDITM.order_quantity
                                      FROM ADDRESS ADR
                                      INNER JOIN CUSTOMER CUST
                                      ON ADR.address_id= CUST.address_id
                                      INNER JOIN ORDER_DETAIL ORDDET
                                      ON ORDDET.customer_id = CUST.customer_id
                                      INNER JOIN ORDER_ITEM ORDITM
                                      ON ORDITM.order_id = ORDDET.order_id')

#cust_order_join<- inner_join(Customer,Order,by="customer_id")

#Grouping by city to get the overall sales per city

sales_per_region<- cust_order_join %>% group_by(city)%>%
  summarise(quantity=sum(order_quantity))

#Get the top 10 cities for sales
top_10_sales_per_region <- sales_per_region[order(-sales_per_region$quantity),]

top_10_sales_per_region <- head(top_10_sales_per_region,10)

#Bar graph to show the sales per region

ggplot1 <- ggplot(top_10_sales_per_region,
aes(x = reorder(city,-quantity), y = quantity,fill=quantity)) +
geom_bar(stat='identity') +
scale_fill_gradient(low = "lightblue", high = "darkblue")+
xlab('Top 10 cities where sales is maximum')+ylab('Quantity Sold')+
theme(axis.text.x=element_text(angle=45,
hjust = 1))

#Table to get the revenue per city

cust_order_product_join <-RSQLite::dbGetQuery(connect,
'SELECT city,ORDITM.order_quantity,unit_price
FROM ADDRESS ADR
INNER JOIN
CUSTOMER CUST ON ADR.address_id= CUST.address_id
INNER JOIN
ORDER_DETAIL ORDDET ON ORDDET.customer_id = CUST.customer_id
INNER JOIN
ORDER_ITEM ORDITM ON ORDITM.order_id = ORDDET.order_id
INNER JOIN
PRODUCT PRD ON PRD.product_id = ORDITM.product_id')

#SQL Query to get the top 5 products

top_5_products <-RSQLite::dbGetQuery(connect,
'SELECT city,ORDITM.order_quantity,unit_price,
PRD.product_id,PRD.product_name FROM
ADDRESS ADR INNER JOIN
CUSTOMER CUST ON ADR.address_id= CUST.address_id
INNER JOIN
ORDER_DETAIL ORDDET ON ORDDET.customer_id = CUST.customer_id
INNER JOIN
ORDER_ITEM ORDITM ON ORDITM.order_id = ORDDET.order_id
INNER JOIN
PRODUCT PRD ON PRD.product_id = ORDITM.product_id')

#Revenue calculation

cust_order_product_join$revenue<- cust_order_product_join$order_quantity*
cust_order_product_join$unit_price

#Grouping by the get the revenue for each city
revenue_per_region <- cust_order_product_join %>% group_by(city) %>%
  summarise(revenue= sum(revenue))

#Get the top 10 cities for revenue

top_5_revenue_per_region <-
revenue_per_region[order(-revenue_per_region$revenue),]

top_5_revenue_per_region <- head(top_5_revenue_per_region,10)

#Bar graph to show the revenue per region

ggplot2 <- ggplot(top_5_revenue_per_region,aes(x = reorder(city, -revenue),
y=revenue,fill=revenue))+
geom_bar(stat='identity')+
scale_fill_gradient(low = "lightblue", high = "darkblue")+
xlab('Top 10 cities with the maximum revenue')+ylab('Revenue')+
theme(axis.text.x=element_text(angle=45,hjust = 1))



#Revenue calculation to show the top 5 products for each city
#with maximum revenue

top_5_products$revenue <-
top_5_products$order_quantity*top_5_products$unit_price

top_5_products_region <- top_5_products %>% group_by(city,product_name) %>%
summarise(rev= sum(revenue))

#Table to show the products that are
#sold for the top 2 revenue producing cities
(top_5_products_region <- top_5_products_region %>%
filter(city %in% head(top_5_revenue_per_region$city,2))
%>% select(product_name))
```

```
# A tibble: 25 x 2
# Groups:   city [2]
   city    product_name
   <chr>   <chr>
 1 Bristol Aurora Lip Gloss Collection
 2 Bristol Celestial Leggings
 3 Bristol Cosmic Compression Socks
 4 Bristol Enchanted Fairy Lights
 5 Bristol FusionX Gaming Console
 6 Bristol Solar Flare Sweater
 7 Bristol Stellar Surfboard
 8 Bristol Supernova Sweatpants
 9 Bristol Zenith Ultra HD Monitor
10 Bristol Zenith Yoga Pants
# i 15 more rows
```

```
grid.arrange(ggplot1,ggplot2)
```

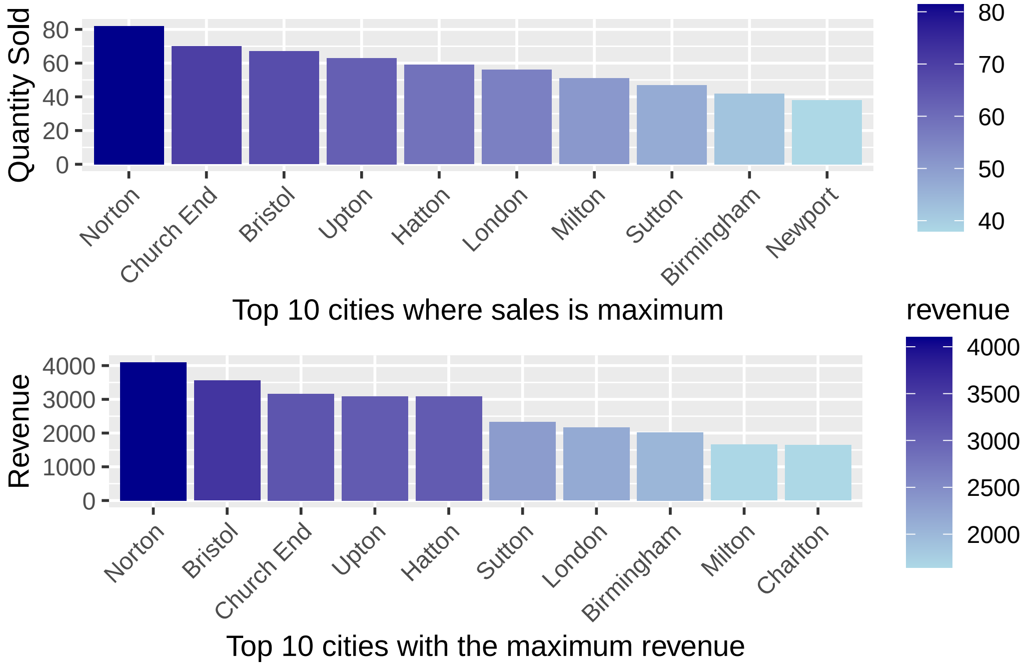


Figure 5: Top 10 cities with maximum revenue and sales

Figure 5 gives a crucial insight into cities that generated significant revenue with a lesser quantity of products sold. Top Products from these cities could be identified as shown in Table 1, as their sales could primarily improve revenue. Also, cities like Ford, Manchester, and Milton generated outstanding revenue despite limited sales. This indicates that customers in this region had a good willingness to pay, which could be capitalised by targeted marketing.

```
#Report 5
# Joining Product with advertise_in and advertisement to get advertisement
#details
advertise_in <- RSQLite::dbGetQuery(connect,'SELECT * FROM ADVERTISE_IN')
advertisement <- RSQLite::dbGetQuery(connect,'SELECT * FROM ADVERTISEMENT')

advertisement_data <- product %>%
  inner_join(advertise_in, by = "product_id") %>%
  inner_join(advertisement, by = "ad_id") %>%
  group_by(product_id, product_name, ad_place) %>%
  summarise(
    total_frequency = sum(ad_frequency),
    .groups = 'drop' )
```

```r
# Joining product with order to get the number of sales per product
number_of_sales <- product %>%
  inner_join(order_item, by = "product_id") %>%
  group_by(product_id, product_name) %>%
  summarise(sales_count = n(), .groups = 'drop')



merged_data <- merge(number_of_sales, advertisement_data, by =
c("product_id", "product_name"))



# Analyzing which ad place is most effective by calculating a
#ratio of total sales to total ad frequency
effective_ad_type <- merged_data %>%
  group_by(ad_place) %>%
  summarise( total_sales = sum(sales_count),
             total_frequency = sum(total_frequency), .groups = 'drop') %>%
  mutate( effectiveness = total_sales / total_frequency) %>%
  arrange(desc(effectiveness))




ggplot(effective_ad_type, aes(x = ad_place, y = effectiveness,
fill = ad_place))+geom_col(show.legend = FALSE, width = 0.5) +
# Adjust bar width here
scale_fill_brewer(palette = "Paired") +
# Use a more appealing color palette
labs(title = "Effectiveness of Advertisement Types",
x = "Ad Place",
y = "Total Sales / Total Frequency") +
theme_minimal(base_size = 14) +
# Increase base text size for better readability
theme(plot.title = element_text(hjust = 0.5, size = 20),
# Center and style title
axis.title = element_text(size = 16),
# Style axis titles
axis.text = element_text(size = 12),
# Style axis texts
panel.grid.major = element_blank(),
# Remove major grid lines
panel.grid.minor = element_blank(),
# Remove minor grid lines
panel.background = element_rect(fill = "white", colour = "grey50"))
```

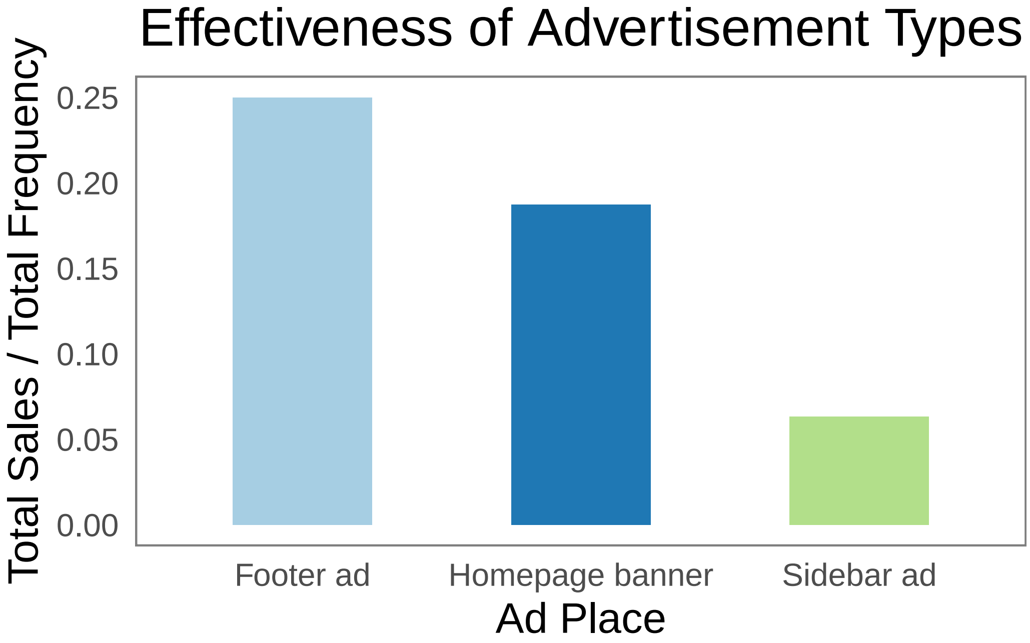


Figure 6: Sales of products per 1 advertisement of each type

We analysed 3 types of advertisements on the Marketplace to find that Products that are advertised with the Footer ads, on average, have higher sales in units. The intention was to compare the number of units sold depending on the type of advertisement this product was in. This finding can be used to assign a price for ads to be charged from suppliers.

```
#Report 6

## Calculating Average Review for each category

category <- RSQLite::dbGetQuery(connect,'SELECT * FROM CATEGORY')

# Join product_df with category_df to include category names
product_with_category <- product %>%
  inner_join(category, by = "category_id")


avg_rating_by_category <- product_with_category %>%
  group_by(category_id, category_name) %>%
  summarise(Average_Rating = round(mean(product_rating, na.rm = TRUE),2),
            .groups = 'drop') %>%
  arrange(category_id)


#Calculating Marketplace fee
merged_product_fee <- product %>%
  inner_join(select(category, category_id, category_fee, category_name),
             by = "category_id")
category_fee <- order_item %>%
  inner_join(select(order_detail, order_id, order_status), by = "order_id") %>%
  inner_join(select(merged_product_fee, product_id, product_name, unit_price,
                    category_fee, category_id, category_name),
             by = "product_id") %>%
  filter(order_status == "Completed") %>%
  mutate(marketplace_fee = order_quantity * unit_price * category_fee/100) %>%
  mutate(cat_total_sales = order_quantity * unit_price) %>%
  group_by(category_id, category_name) %>%
  summarise(total_fee = sum(marketplace_fee),
            total_sales = sum(cat_total_sales),
            total_sales_unit = sum(order_quantity)) %>%
  inner_join(select(avg_rating_by_category, category_id, Average_Rating),
               by= 'category_id')


#Visualise


#graph to compare category sales vs category fee
ggplot(category_fee, aes(x = reorder(category_name, -total_fee))) +
  geom_bar(aes(y = total_sales, fill = "Total Sales"), stat = "identity",
           position = position_dodge(width = 0.9), alpha = 0.7) +
  geom_bar(aes(y = total_fee, fill = "marketplace fee"), stat = "identity",
           position = position_dodge(width = 0.9)) +
  scale_fill_manual(name = "category", values =
                      c("marketplace fee" = "#3182bd",
                              "Total Sales" = "#31a354")) +
  labs(title = "Marketplace Fee vs Total Sales by Category", x = "Category",
       y = "Amount, GBP") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

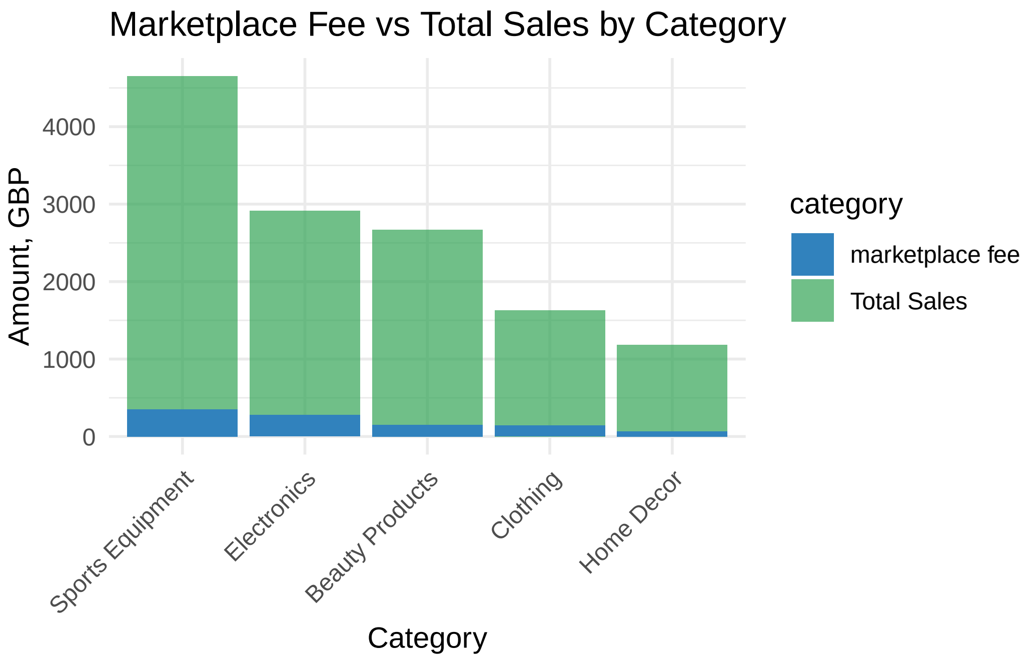


Figure 7: Comparing Marketplace fees to the sales in each Category of products

High sales in the category relate to a higher Marketplace fee. Increasing fees for the categories with higher sales and decreasing fees for the categories with low sales will maximise Marketplace profit by selling more in less popular categories and getting higher fees for already popular categories. Sports Equipment has the highest sales in terms of money. However, as we will see later, most units are sold in the Electronics category.

```
#Report 7
# Graph to compare Category sales VS Category average rating
ggplot(category_fee, aes(x = reorder(category_name, -Average_Rating))) +
  geom_bar(aes(y = total_sales_unit, fill = "Total Sales in Units"),
           stat = "identity", position = position_dodge(width = 0.9),
           alpha = 0.6) +
  geom_bar(aes(y = (Average_Rating-3)*20, fill = "Average Rating"),
           stat = "identity", position = position_dodge(width = 0.9),
           alpha = 1) +
  geom_text(aes(label = sprintf("%.2f", Average_Rating),
                y = Average_Rating, x = category_name),
            color = "black", size = 4, hjust = 0.5) +
  scale_fill_manual(name = "Category", values =
                      c("Average Rating" = "#FFCC80",
                        "Total Sales in Units" = "#31a354")) +
  labs(title = "Average Rating vs Total Sales in Units by Category",
       x = "Category", y = "Total Sales in Units") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(name = "Total Sales in Units")
```

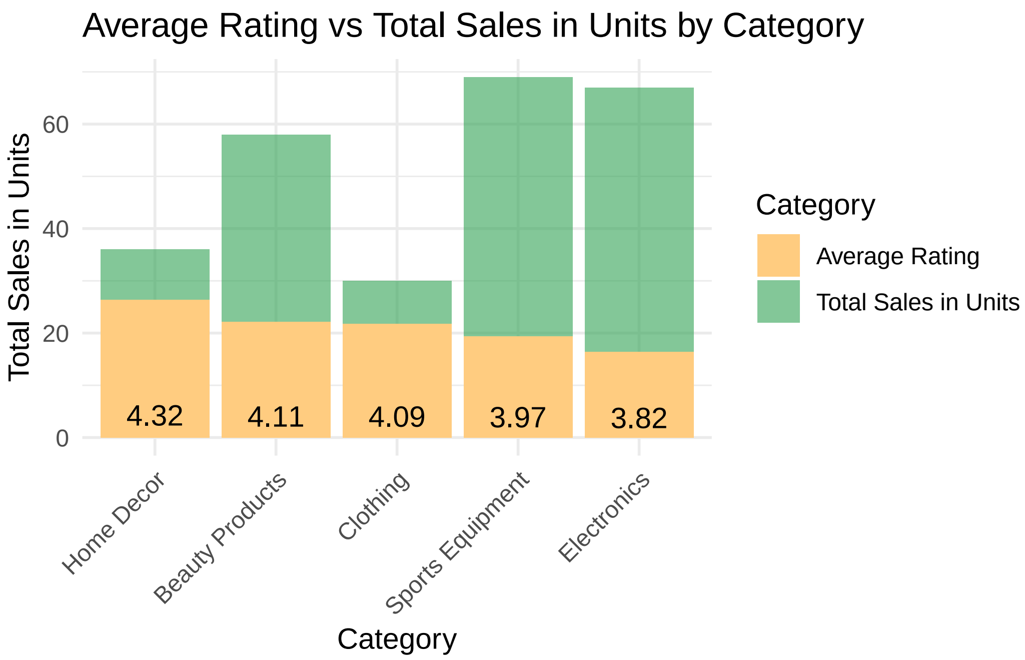


Figure 8: Average customers' rating for the category and category's sales in units

Categories that have products with higher ratings have lower sales in units. This means that with the increase in sales, the number of negative reviews increases at a higher rate. Marketplace may want to reconsider suppliers for the categories with lower average ratings.