SQL Queries

I am going to be writing SQL queries using the MYSQL syntax. Tables will be created based on the ER diagram given.

Assumptions for designing the table schema:

- All string 'zero' or ' ' values are already cleaned and cast to FLOAT.
- All missing values and duplicates have been removed.
- Each row in the Transactions table now represents a unique product purchased (line item) within a receipt
- The combination of RECEIPT_ID and BARCODE can be used as a composite primary key for the transactions table
- PURCHASE_DATE column in the TRANSACTIONS TABLE has the time component added to it
- With the removal of duplicate values, each barcode now maps to a unique product record

Schema:

USER TABLE:

```
CREATE TABLE Users (
    ID VARCHAR(50) PRIMARY KEY,
    CREATED_DATE DATETIME,
    BIRTH_DATE DATETIME,
    STATE VARCHAR(10),
    LANGUAGE VARCHAR(10),
    GENDER VARCHAR(10)
);
```

TRANSACTION TABLE:

```
CREATE TABLE Transactions (
```

```
    RECEIPT_ID VARCHAR(50),
    PURCHASE_DATE DATETIME,
    SCAN_DATE DATETIME,
    STORE_NAME VARCHAR(100),
    USER_ID VARCHAR(50),
    BARCODE VARCHAR(20),
    FINAL_QUANTITY FLOAT,
    FINAL_SALE FLOAT,
    PRIMARY KEY (RECEIPT_ID, BARCODE),
    FOREIGN KEY (USER_ID) REFERENCES Users(ID),
    FOREIGN KEY (BARCODE) REFERENCES Products(BARCODE));
```

PRODUCTS TABLE:

```
CREATE TABLE Products (
    BARCODE BIGINT PRIMARY KEY,
    CATEGORY_1 VARCHAR(100),
    CATEGORY_2 VARCHAR(100),
    CATEGORY_3 VARCHAR(100),
    CATEGORY_4 VARCHAR(100),
    MANUFACTURER VARCHAR(100),
    BRAND VARCHAR(100)
);
```

Closed Ended Questions:

1. What are the top 5 brands by receipts scanned among users 21 and over?

```sql
--What are the top 5 brands by receipts scanned among users 21 and over?

-- Get the top 5 brands by number of receipts scanned among users aged 21 and over
SELECT
    p.BRAND,  -- Brand name from product metadata
    COUNT(DISTINCT t.RECEIPT_ID) AS receipt_count  -- Total unique receipts where the brand appeared
FROM
    USER_TAKEHOME u
JOIN
    TRANSACTION_TAKEHOME t ON u.ID = t.USER_ID  -- Join transactions to users
JOIN
    PRODUCTS_TAKEHOME p ON t.BARCODE = p.BARCODE -- Join products via barcode
WHERE
    TIMESTAMPDIFF(YEAR, u.BIRTH_DATE, CURDATE()) >= 21  -- Filter: users must be at least 21 years old
    AND p.BRAND IS NOT NULL  -- Exclude records without a brand value
GROUP BY
    p.BRAND  -- Aggregate by brand
ORDER BY
    receipt_count DESC  -- Show highest receipt counts first
LIMIT 5;  -- Return only the top 5 brands
```

Open Ended Questions:

1. Who are Fetch's power users?

Assumption: To answer this question, I will assume power users as those users who are either in the top 1% of transaction count or top 1% of total spend.

SQL Query (Screenshots using Notepad++):

```sql
-- Step 1: Aggregate stats per user
WITH user_statitics AS (
  SELECT
    USER_ID,
    COUNT(*) AS number_of_transactions,  -- Total number of transactions by user
    SUM(CAST(FINAL_SALE AS DECIMAL(10,2))) AS total_amount_spent  -- Total amount spent by user
  FROM Transactions
  WHERE FINAL_SALE IS NOT NULL --Exclude records without a final sale value
  GROUP BY USER_ID -- Aggregate by User ID
),

-- Step 2: Rank users by transaction count and total spend
txn_ranks AS (
  SELECT *,
         NTILE(100) OVER (ORDER BY number_of_transactions) AS txn_percentile,  -- Break users into 100 groups by transaction count
         NTILE(100) OVER (ORDER BY total_amount_spent) AS spend_percentile       -- Break users into 100 groups by amount spent
  FROM user_statistics
)

-- Step 3: Select users who fall in the top 1% of either metric
SELECT
  USER_ID,
  number_of_transactions,
  total_amount_spent
FROM txn_ranks
WHERE txn_percentile = 100  -- Top 1% by number of transactions
   OR spend_percentile = 100  -- Top 1% by total amount spent
ORDER BY total_spent DESC;
```

2. Which is the leading brand in the Dips & Salsa category?

Assumption: Leading brand means the brand with the **highest total sales** in the "Dips & Salsa" category. I will match this by filtering on Category_3 = Dips & Salsa

```sql
-- Fetch Leading Brands in Dips & Salsa Category
-- Find the leading brand in the 'Dips & Salsa' category by total sales
SELECT
  p.BRAND,  -- The brand of the product
  SUM(CAST(t.FINAL_SALE AS DECIMAL(10,2))) AS total_sales  -- Sum of all final sales, cast to decimal to better handle financial data
FROM Transactions t
JOIN Products p
  ON t.BARCODE = p.BARCODE  -- Join on barcode to link transactions and product tables
WHERE
  p.CATEGORY_3 = 'Dips & Salsa'  -- Focus only on products in the 'Dips & Salsa' subcategory
  AND t.FINAL_SALE IS NOT NULL  -- Exclude rows where sales data is missing
GROUP BY p.BRAND  -- Aggregate total sales by brand
ORDER BY total_sales DESC  -- Rank brands from highest to lowest based on total sales
LIMIT 1;  -- Return only the top-selling brand
```