# Fetch Data Analyst – Take Home

**First: explore the data**

**Review the unstructured csv files and answer the following questions with code that supports your conclusions:**

**Q1.1. Are there any data quality issues present?**

Table: USERS

* BIRTH\_DATE:
  + Several cells in the BIRTH\_DATE field are null or blank, which might be expected since users may not be required to provide this information at sign-up.
  + There are a number of users listed as 100+ years old (e.g., 24 users are listed as 120 years old or older).
    - This may indicate fake or duplicate accounts created for rewards, as these users likely sped through the demographic questions.
  + One user’s BIRTH\_DATE occurs after their CREATED\_DATE, which is illogical and suggests either a data entry error or an issue with the field.
  + Can run this query to identify that record with the error:
    - SELECT \*

FROM USER\_TAKEHOME

WHERE BIRTH\_DATE > CREATED\_DATE

* STATE, LANGUAGE, GENDER:
  + These fields also have null or blank values, possibly due to their optional nature during sign-up.
* GENDER:
  + The values in this field lack standardization, with inconsistent naming conventions (e.g., "non\_binary" vs. "Non-Binary"). This inconsistency could lead to missing data or errors during analysis.
  + Examples of current values include:
    - female, male, non\_binary, Non-Binary, My gender isn’t listed, not\_listed, not\_specified, Prefer not to say, prefer\_not\_to\_say, transgender, unknown.
  + Noted action item to work with the data engineering team on the data being loaded here.
  + Can run this query to identify the misaligned values:
    - SELECT DISTINCT GENDER FROM USERS

Table: TRANSACTIONS

* Receipt Records:
  + Each receipt\_id has multiple associated records, likely representing individual items scanned or purchased. Fields such as BARCODE, FINAL\_QUANTITY, and FINAL\_SALE suggest this is the case, which makes sense as multiple items can be part of a single purchase.
* SCAN\_DATE:
  + Some SCAN\_DATE values occur after their corresponding PURCHASE\_DATE, which is illogical and may indicate data inaccuracies.
* FINAL\_QUANTITY:
  + Some items in the dataset have a FINAL\_QUANTITY of zero. This is unusual for a transactions table where items are assumed to be purchased. Further investigation is needed to determine whether this is expected (e.g., scanned items that weren’t fulfilled or purchased).
* BARCODE:
  + Several records have null or blank values, which is problematic since barcodes should uniquely identify items.
  + The barcodes for each receipt\_id are also identical across all of its records; I would assume barcodes should be unique per item purchased

File: PRODUCTS

* CATEGORY\_1:
  + Typically, a products table with multiple category fields would have data populated at the CATEGORY\_1 field, as that is the highest level of categorization. However, there are null or blank values for some products. Clarification is needed on whether this is expected.
* BARCODE:
  + The barcode should serve as the primary key to link products to transactions. However, there are null or blank values which will likely cause issues in analyses.
* CATEGORY\_2 through CATEGORY\_4:
  + Blank values in these fields seem reasonable, as they likely represent more specific categorizations that may not always apply.
* MANUFACTURER and BRAND:
  + These fields also have null or blank values. This could be expected if the data was imported from a third-party system without this information. If not, this would be a data quality issue requiring further investigation.

**Q1.2. Are there any fields that are challenging to understand?**

* Other than the assumptions and issues mentioned above, there are no additional fields that are difficult to understand.

**Second: Provide SQL Queries**

**Answer three of the following questions with at least one question coming from the closed-ended and one from the open-ended question set. Each question should be answered using one query.**

**Q2.1. What are the top 5 brands by sales among users that have had their account for at least six months? (Close-ended)**

Assumptions:

* I am assuming that top brands “by sales” is asking for total revenue of sales, not total quantity sold.

Steps:

1. Join the PRODUCTS and TRANSACTIONS tables to pull in the brand into each transaction
2. Sum the final\_sale and group by brand to get that sum aggregated by brand
3. Pull users with created\_date <= 6 months before today’s date
4. Filter the user\_id to users from the previous step
5. Order in descending order by the sum of final sales

Below is my query:

SELECT

brand,

SUM(final\_sale) as total\_sales\_brand

FROM TRANSACTIONS as t

JOIN PRODUCTS as p on p.barcode = t.barcode

WHERE user\_id IN (SELECT id FROM USERS WHERE created\_date <= dateadd('month', -6, CURRENT\_DATE()))

GROUP BY brand

ORDER BY total\_sales\_brand DESC

LIMIT 5

**Q2.2. Which is the leading brand in the Dips & Salsa category? (Open-ended)**

**Assumptions:**

* Defining "leading" as the brand with the most sales based on "final\_sale" revenue

Steps:

1. Join the PRODUCTS and TRANSACTIONS tables to pull in the brand
2. Sum the final\_sale and group by brand to get that sum aggregated by brand
3. Filter to the ‘Dips & Sales’ category on the correct category field
4. Order in descending order by the sum of final sales

Below is my query:

SELECT

brand,

SUM(final\_sale) as total\_sales\_brand

FROM transactions AS t

JOIN products as p on p.barcode = t.barcode

WHERE category\_2 = 'Dips & Salsa'

GROUP BY brand

ORDER BY total\_sales\_brand DESC

LIMIT 1

**Q2.3. What are the top 5 brands by receipts scanned among users 21 and over? (Close-ended)**

Steps:

1. Join the PRODUCTS and TRANSACTIONS tables to pull in the brand
2. Count the distinct number of receipt\_ids and group by brand
3. Calculate users who are 21+ and filter in the WHERE clause to those users
4. Order in descending order by the count distinct of receipt\_ids

Below is my query:

SELECT

brand,

COUNT(DISTINCT t.receipt\_id) as receipts\_scanned -- using COUNT(DISTINCT receipt\_id) to account for duplicate transaction records associated with the same receipt, ensuring that each receipt is only counted once in the final result

FROM TRANSACTIONS as t

JOIN PRODUCTS as p on p.barcode = t.barcode

WHERE t.user\_id IN

(SELECT id FROM users WHERE TIMESTAMPDIFF(YEAR, birth\_date::DATE, CURRENT\_DATE()) >= 21)

GROUP BY p.brand

ORDER BY receipts\_scanned DESC

LIMIT 5

**Third: communicate with stakeholders**

**Construct an email or slack message that is understandable to a product or business leader who is not familiar with your day-to-day work. Summarize the results of your investigation. Include:**

* **Key data quality issues and outstanding questions about the data**
* **One interesting trend in the data** 
  + **Use a finding from part 2 or come up with a new insight**
* **Request for action: explain what additional help, info, etc. you need to make sense of the data and resolve any outstanding issues**

-------------------------

Hi team,

I’ve reviewed the data tables related to our Users, Transactions, and Products and identified several quality issues that need clarification or further investigation. Below is a summary of my findings:

**Key Data Quality Issues**

USERS:

* Demographic Fields (e.g., BIRTH\_DATE, STATE, LANGUAGE, GENDER): Many fields have null or blank values, which might be expected since these inputs could be optional.
* BIRTH\_DATE:
  + There are 24 users marked as being 120+ years old. This suggests potential fake or duplicate accounts.
  + One user’s BIRTH\_DATE is listed after their account’s CREATED\_DATE, which is likely an error.
* GENDER Field: Inconsistent naming conventions (e.g., "non\_binary" vs. "Non-Binary") could cause errors in analysis. Suggest standardizing these values.

TRANSACTIONS:

* SCAN\_DATE and PURCHASE\_DATE: Some SCAN\_DATE values occur after the PURCHASE\_DATE, which doesn’t make sense.
* FINAL\_QUANTITY: Several records show items with zero quantities purchased. Need clarification on whether this is expected (e.g., scanned but not purchased).
* BARCODE: Blank or null barcodes and duplicate barcodes per receipt are problematic, as barcodes should be unique to track individual items.

PRODUCTS:

* CATEGORY\_1: This key category field contains null or blank values for several products. Clarification is needed on whether this is expected.
* BARCODE: Missing barcodes in this table raise concerns, as they are critical for linking products to transactions.
* MANUFACTURER/BRAND: These fields also have null or blank values, but this might be expected if data is from a third-party system.

**Interesting Trend**

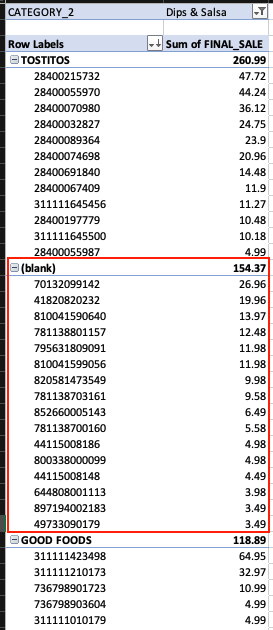
Below are the Top 10 Dips and Salsa Brands from my analysis:



While analyzing the top 10 brands, I noticed that the second highest-selling brand in the data consists of products with missing brand names. To ensure clarity and readability, I omitted these blank values from the chart. However, this is still an interesting finding that despite having no recorded brand, these products still generated significant sales.

This raises a potential data quality issue worth investigating further. It would be helpful to confirm whether these missing values are expected (e.g., due to incomplete product data) or if there’s an opportunity to improve the brand mapping in the dataset.

Below is a pivot table the team can use to identify the barcodes for brands with a missing brand name:



**Action Items:**

To move forward with the analysis and resolve these issues, I’d like to request the following:

1. Confirmation from the appropriate stakeholders and the data engineering team on each of the key data quality issues listed above
2. Brand name assignments for top dips & salsa brands with blank brand values

Let me know if you’d like to discuss further and I can set up a sync with the team.

Thank you,  
Anthony Nguyen