Technical Report (Summary)

The following analysis is for the Fetch Take Home project. The files provided in this assignment are **USER_TAKEHOME.csv** (100,000 records), **PRODUCTS_TAKEHOME.csv** (845,552 records), and **TRANSACTION_TAKEHOME.csv** (50,000 records), which contains data from June 2024 to September 2024.

The tools used for this project include:

- **Excel/ Google Sheets**
- **Python**
- **Big Query (Google SQL)**

Summary of findings (Part I)

Question 1: Are there any data quality issues present?

Yes, all three tables contain data quality issues, although one table has more issues than the other two. The data quality issues found include: inconsistency in categories within variables, misspelling, different types of data within a variable, duplicate rows, missing data, and incorrect data types. A summary of performed checks and findings are listed below:

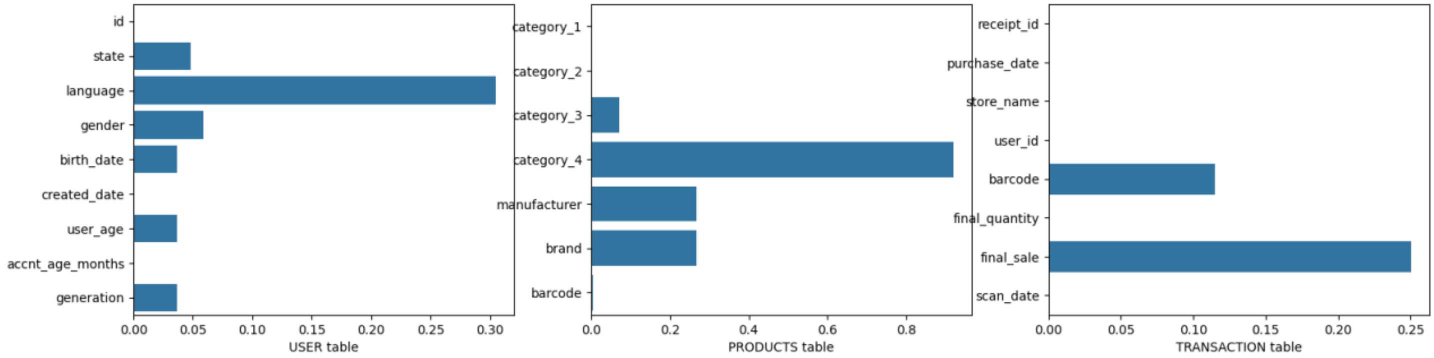


Figure 1Percent of missing data per variable post deduplication.

USER table:

- inconsistency in categories:
  - Non-Binary vs non_binary
  - 'My gender isn't listed' vs not_listed
  - 'Prefer not to say' vs prefer_not_to_say
- 5 variables have missing data e.g., contain NaNs/Null values.
  - state (4.8%)
  - language (30.5%)
  - gender (5.9%)
  - birth_date (3.7%)
  - note that *user_age* and *generation* were created based off *birth_date*
- no duplicates rows.
- Checked that all state acronyms correspond to an actual state. Note that PR and DC are listed in the
- states column.

PRODUCTS table:

- 7 variables have missing data e.g., contain NaNs/Null values.
    - category_1 (0.01%)
    - category_2 (0.17%)
    - category_3 (7.16%)
    - category_4 (92.02%)
    - manufacturer (26.8%)
    - brand (26.8%)
    - barcode (0.47%)
- typos found: e.g., Accesories, "Alchoholic"
- A total of 57 duplicate rows were found.

TRANSACTION table:

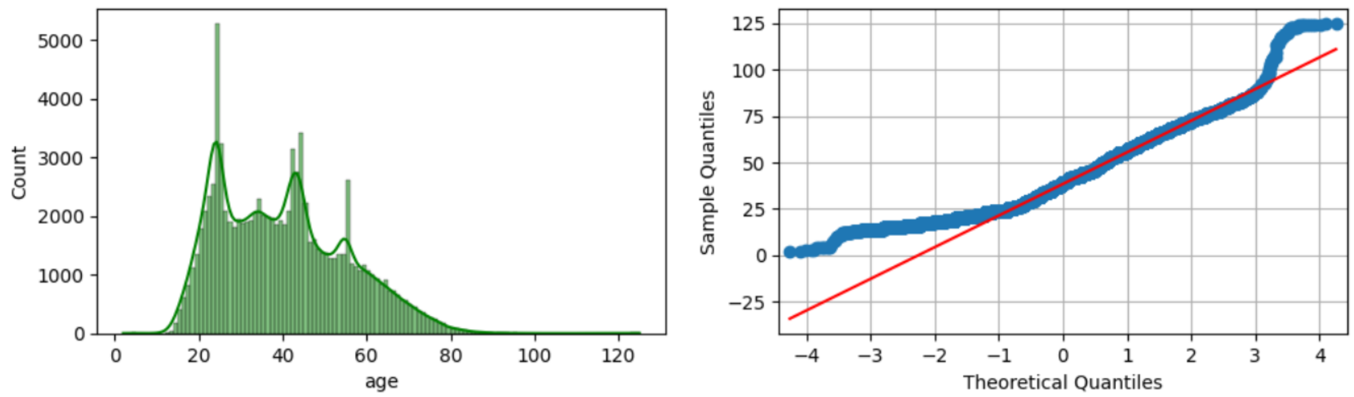
- 2 variables have missing data e.g., contain NaNs/Null values.
    - barcode (11.5%)
    - final_sale (25.1%)
        - This goes down to 0% post SQL analysis.
- found capitalization inconsistency e.g., "TINKER COMMISsARY"
- final_quantity variable data issues:
    - contains the string 'zero' instead of the value 0.
    - has non-integer values for quantities (e.g., 1.23), which does not make sense. These records (110 in total) were removed post analysis in SQL.
- A total of 171 duplicate rows were found.
- There are 94 records in which the products were scanned before they were purchased.

Question 2: Are there any fields that are challenging to understand?

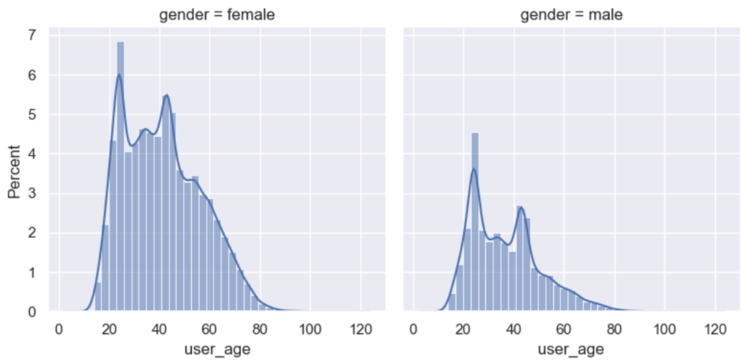
Yes, the TRANSACTION table contains non-integer values in a final_quantity variable. Moreover, there are two rows per transaction with each final_quantity and final_price variable combination, that made this table tricky to understand.

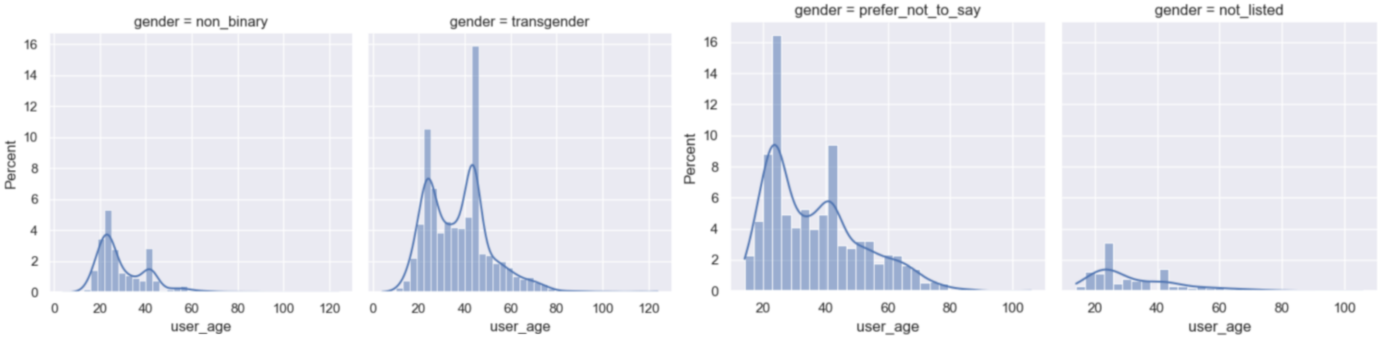
Additional observations

The distribution of **age** shows that a large percent of our users is in their 20s and 40s. The data does not follow a normal distribution, but is large enough to make population inferences.



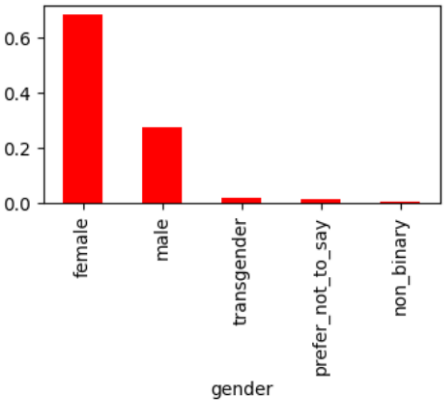
if we consider both **age** and **gender**, the age observation is also true within each gender group.

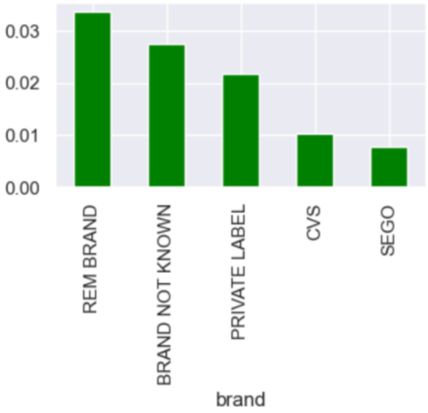


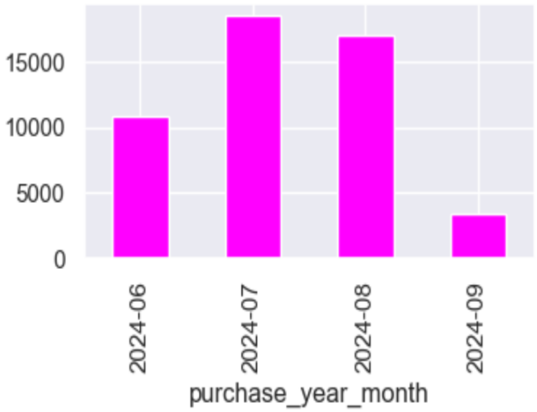


Our analysis also suggests that females account to close to 69% of users, while males are in the 28% range. The rest of the users account for less than 2% each. In terms of **brands**, close to 3.4% correspond to REM BRAND, 2.7% to BRAND UNKNOWN, and 2.1% to PRIVATE LABEL while all others are less than 1%. It's worth mentioning that we have a high number of missing data in this variable and that the data provided may only be a sample.

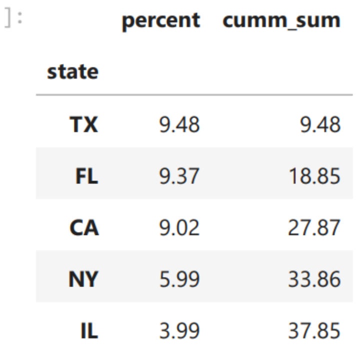
Finally, purchases peaked in July 2024 (37%) and declined all the way to September 2024, the latter being the lowest of the four months, with only about 6%.







We also observed that four **states** take over 30% of our users: TX (9.48%), FL (9.37%), CA (9.02%), and NY (5.99%) and this percent increases to over 50% from 9 out of the 50 states and territories.

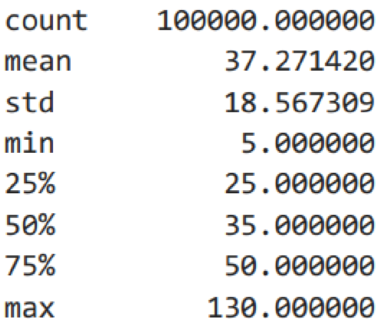


]:

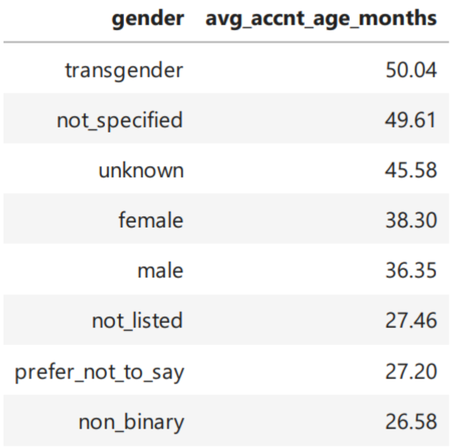
state	percent	cumm_sum
TX	9.48	9.48
FL	9.37	18.85
CA	9.02	27.87
NY	5.99	33.86
IL	3.99	37.85

Another interesting observation is the number of days from the time of purchase to the time a product is scanned: over 60% of products are scanned within 2 days. More precisely, within the same day (48%), 1-day (18%), 2-days (9%), and 3-days (6%) to name a few.

Finally, the average **account age** is about 37 months, with half of the users below 35 months. The distribution by gender is also shown.



```
count    100000.000000
mean         37.271420
std          18.567309
min           5.000000
25%          25.000000
50%          35.000000
75%          50.000000
max         130.000000
```



gender	avg_accnt_age_months
transgender	50.04
not_specified	49.61
unknown	45.58
female	38.30
male	36.35
not_listed	27.46
prefer_not_to_say	27.20
non_binary	26.58

Closed-Ended Questions (Part II)

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

brand	num_scanned_receipts
DOVE	3
NERDS CANDY	3
GREAT VALUE	2
COCA-COLA	2
SOUR PATCH KIDS	2
HERSHEY'S	2
TRIDENT	2
MEIJER	2

We can use another measure to break ties, such as sales.

Question 2: What are the top 5 brands by sales among users that have had their account for at least six months?

brand	total_sale
CVS	72
DOVE	30.91
TRIDENT	23.36
COORS LIGHT	17.48
TRESEMMÉ	14.58

Question 3: What is the percentage of sales in the Health & Wellness category by generation?

generation	percent_sales
Baby Boomers	54.26%
Gen X	23.7%
Millennials	22.04%

Open-Ended Questions (Part III)

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

brand	total_quantity	total_sales	receipt_count
TOSTITOS	38	181.3	36
NULLs	22	100.97	21
GOOD FOODS	9	94.91	9
PACE	24	85.75	24

Here we assume that by leading brand, we mean that it is leading in terms of total quantity, sales, and receipt scans. We cannot use a single variable, such as sales, because products have different costs. It would be useful to understand whether single items were scanned as opposed to bulk items. For example, a user can buy a bag of chips from brand X but also scan another item from brand Y, which comes with several smaller bags or is at a discount.

Finally, it is worth noting that records with no brand are taking second place in our list, meaning that we are missing out on a good amount of information for our analysis.