First: explore the data

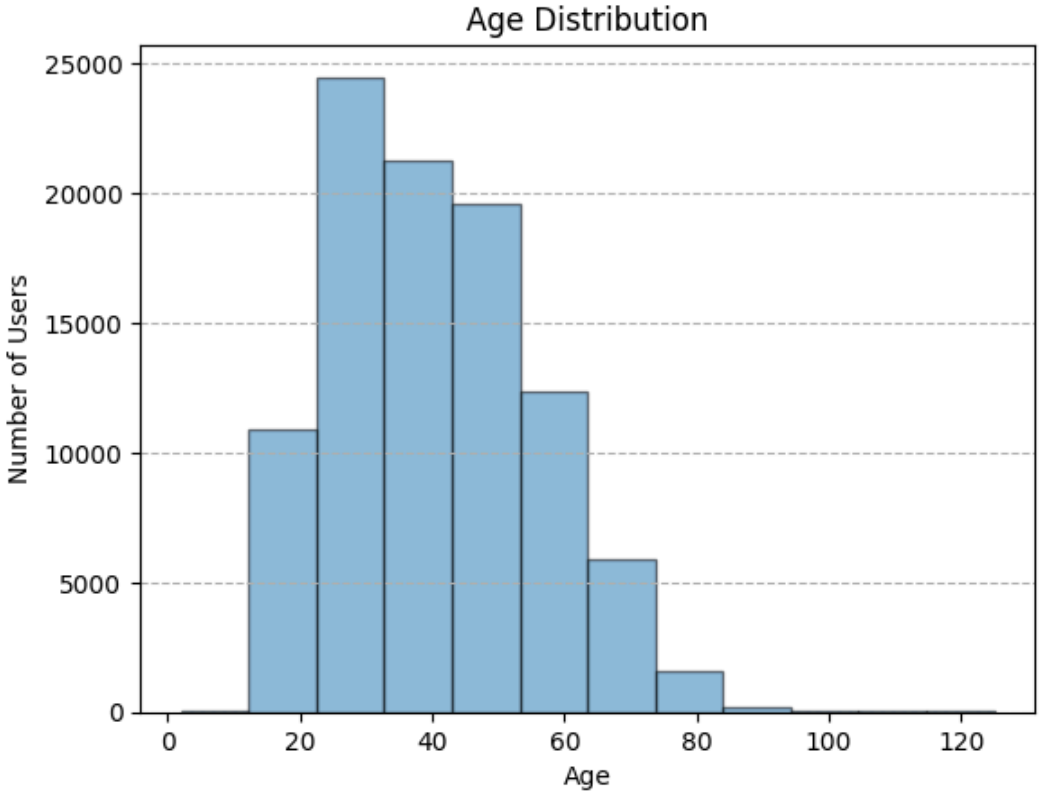
- **Are there any data quality issues present?**

There were data quality issues across all tables that needed to be addressed. In the Users table, the column names were incorrect, with the first row containing actual headers instead of metadata. From the histogram of user age distribution (attached below), I noticed users aged 125+, which led me to find 1900-01-01 birthdates. I assume that this is placeholders for missing values. I adjusted these accordingly to improve accuracy. In the Products table, I initially assumed BARCODE was the primary key, but I found duplicate values that shouldn't exist. Additionally, there were inconsistencies in barcode lengths, requiring assumptions based on the data distribution to standardize them. In the Transactions table, some RECEIPT_IDs appeared multiple times with different FINAL_QUANTITY or FINAL_SALE values, despite having the same ID and timestamp, indicating potential data discrepancies. There were also formatting inconsistencies, where "zero" was stored as text instead of a numeric value, which required standardization. To improve data integrity and usability, I standardized column formats across all tables, addressed missing values with interpretable placeholders, and ensured consistency for accurate reporting and analysis.

- **Are there any fields that are challenging to understand?**

A few fields in the dataset were challenging to interpret due to inconsistencies and missing context. In the Users table, the BIRTH_DATE field contained values like 1900-01-01, which I assumed to be a placeholder for missing data rather than an actual birthdate. However, this raised questions about whether it should be treated as NULL or if it held any specific meaning. In the Products table, the category columns (CATEGORY_1, CATEGORY_2, etc.) had some missing values, making it unclear whether certain classifications were incomplete or if these columns were optional. However, if I were to guess, it looks like each category column gets more specific as it goes from CATEGORY_1 to CATEGORY_4. Transactions table had discrepancies in RECEIPT_ID, where multiple rows shared the same ID but had different FINAL_QUANTITY and FINAL_SALE values. It was unclear whether these were partial transactions, refunds, or data entry errors, which required further investigation. To address these issues, I applied standardization techniques and made informed assumptions to improve data consistency while preserving as much accuracy as possible.

```python
import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
df_users = pd.read_csv("USER_TAKEHOME.csv")

# Convert BIRTH_DATE to datetime
df_users['BIRTH_DATE'] = pd.to_datetime(df_users['BIRTH_DATE'])

# Create a today variable so we can calculate the age based on the BIRTH_DATE
# Also convert it to UTC to make it timezone-aware
today = pd.Timestamp('today').tz_localize('UTC')

# Calculate age
df_users['AGE'] = (today - df_users['BIRTH_DATE']).dt.days // 365

# Drop null values in AGE
df_users = df_users.dropna(subset=['AGE'])

# Plot histogram of age distribution
# Made 12 bins - 1 bin for every 10 years
plt.hist(df_users['AGE'], bins=12, edgecolor='black', alpha=0.5)
plt.xlabel('Age')
plt.ylabel('Number of Users')
plt.title('Age Distribution')
plt.grid(axis='y', linestyle='--')
plt.show()
```



Second: provide SQL queries

(Note: Query results are in the appendix below. Assumptions in the comments of the SQL file)

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

Among users who have had their accounts for over six months, CVS leads in total sales, bringing in $72.00, more than double that of DOVE ($30.91) in second place. TRIDENT ($23.36), COORS LIGHT ($17.48), and TRESEMMÉ ($14.58) round out the Top 5. We can see 3 of the 5 as Health & Wellness category, while the other 2 are Snacks and Alcohol respectively.

If I were to create a dashboard/chart of this insight, I would create a bar chart that allows the stakeholders to easily and clearly see the total sales by each brand.

- **What is the percentage of sales in the Health & Wellness category by generation?**

Baby Boomers lead in Health & Wellness spending with $84.09, making up 37.32% of their total purchases ($225.30). Gen X follows, spending $37.81 on Health & Wellness, which accounts for 22.36% of their $169.13 total spend. Millennials, despite having $189.61 in total purchases, allocate only 18.55% ($35.17) to Health & Wellness. This trend suggests that older generations are spending more on the Health & Wellness category, while younger consumers focus more on other categories. This insight can help us tailor product offerings and marketing strategies to specific generations based on their buying behaviors.

If I were to create a dashboard/chart for this insight, I would create a stacked bar chart that shows both the health sales and total sales for each generation. I would also create a line chart of the health sales and total sales comparison over birth years and see if there are any interesting results there.

- **Which is the leading brand in the Dips & Salsa category?**

TOSTITOS is the clear leader in the Dips & Salsa category, generating $181.30 in total sales with 38 units sold across 36 transactions, making it both the highest-grossing and most frequently purchased brand. It significantly outperforms GOOD FOODS ($94.91, 9 units) and PACE ($79.73, 22 units), which rank second and third, respectively, in both revenue and purchase volume, further solidifying its dominance in the category.

If I were to create a dashboard/chart for this insight, I would create a triple bar chart that has the three aggregations broken down by each brand, so the stakeholders can see the differences easily. If one of the metrics is of a higher priority, then I would do a normal bar chart to make things less confusing for the stakeholders.

Third: communicate with stakeholders

Assuming that the team leader has some technical knowledge and also knows the context on the data sources

Hey team leader,

Hope you're doing well and having a great week.

I've been analyzing our transaction, user, and product data and wanted to highlight key findings along with a few data quality issues that need clarification.

I have some clarifying questions regarding our data quality and I was wondering if you know anything about it or who I should reach out to in order to find out.

1. Some users have a BIRTH_DATE of '1900-01-01', which seems like a placeholder.
    a. Should these be treated as missing values or NULLs, or do they have significance?
2. Some RECEIPT_IDs appear multiple times with different FINAL_QUANTITY and FINAL_SALE values.
    a. Are these partial purchases, refunds, or errors?
3. Some products have missing values in CATEGORY_1 through CATEGORY_4.
    a. Do you know if these fields are optional, or do we need to address missing classifications?

Switching over to some key insights that I have found regarding the data.

1. Top brands among users with accounts 6+ months old
    - CVS leads in sales ($72.00), followed by DOVE ($30.91), TRIDENT ($23.36), COORS LIGHT ($17.48), and TRESEMMÉ ($14.58).
    - Three of the top five brands fall under Health & Wellness, showing that long-term users tend to buy from this category more than others.

2. Health & Wellness spending broken down by generation
    - Baby Boomers spend the most on Health & Wellness, with $84.09, making up 37.32% of their total purchases.
    - Gen X follows, spending $37.81 (22.36%), while Millennials spend $35.17 (18.55%), allocating a smaller percentage of their total purchases to this category.
    - It looks like older generations prioritize Health & Wellness more, while younger consumers may be spending more on other categories.

3. Top brand in the Dips & Salsa Category

- TOSTITOS leads in sales with $181.30, selling 38 units across 36 transactions, significantly ahead of second-place GOOD FOODS ($94.91, 9 units) and third-place PACE ($79.73, 22 units).

Some next steps might be:

1. Health & Wellness appears to be a high-demand category that we can capitalize on. I can dive deeper into the category breakdown to confirm, but it makes sense to start focusing efforts here.
2. I think that tailoring product offerings and marketing strategies to specific generations based on their buying behaviors might be a good idea to consider. I can analyze other categories to see if certain demographics show strong preferences.
3. Since TOSTITOS is the clear leading brand in the Dips & Salsa category, we could explore potential promotional strategies and/or partnerships with TOSTITOS to further increase engagement and sales.

Let me know what your thoughts are and how you'd like to proceed!

Best,
Samuel

Appendix

Section 2 Query Results

1.

	BRAND	CATEGORY_1	total_sales
1	CVS	HEALTH & WELLNESS	72
2	DOVE	HEALTH & WELLNESS	30.910000000000004
3	TRIDENT	SNACKS	23.36
4	COORS LIGHT	ALCOHOL	17.48
5	TRESEMMÉ	HEALTH & WELLNESS	14.58

2.

	generation	total_health_sales	total_all_sales	health_sales_percentage
1	Gen X	37.81	169.12999999999997	22.36
2	Millennials	35.17	189.61	18.55
3	Baby Boomers	84.09	225.30000000000004	37.32

3.

	BRAND	total_sales	total_quantity	count_receipts
1	TOSTITOS	181.29999999999998	38	36
2	GOOD FOODS	94.90999999999998	9	9
3	PACE	79.73	22	22