Subject: Update on Fetch Rewards Data Analysis & Request for Clarification

Greetings,

I hope you are doing well. My name is Likhita, and I am reaching out to provide an update on the progress of our data analysis for Fetch Rewards, highlight key data quality issues, and seek clarification on a few outstanding questions.

Summary of Work Completed:

I have been working with the **Users, Transactions, and Products** tables provided as unstructured CSV files. My approach involved:

1. **Data Cleaning & Transformation** – Handling missing values, standardizing formats, and resolving inconsistencies, deleting duplicates using advanced Python libraries.

2. **Exploratory Analysis** – Solving few business questions and deriving insights using SQL and Power BI

Key Data Quality Issues Identified:

- **Missing Data & Inconsistencies** across all three tables and duplicate records leading to skewed results. Missing data distribution is as follows:

  - **Products:** Category_3 (~7%), Category_4 (~92%), Manufacturer (~27%), Brand (~27%)

  - **Users:** BirthDate (~4%), State (~5%), Language (~31%), Gender (~6%)

  - **Transactions:** Barcode (~11.52%), Final Sale (~25%)

- **Data Type Issues:** All the date columns in three datasets (Scan Date, Purchase Date, Birth Date, Created Date) and Barcode were given as object datatypes.

- **Inconsistent Values:**

  - **Products:** Around 11% of Manufacturers, 3% of Brands are placeholder values

  - **Users:** Gender column has inconsistent and duplicate values because of the free text entry in that column.

    There are 7612 users whose birth date is between 1900-1960 and this might be because of fake accounts or wrong data collection.

Interesting Trends:

Weekly Purchase Activity Across Generations: This graph gives us the purchase activity of generations over the week. We also added Purchase activity by Product visual (bar chart) to check out which product

in each generation is contributing to more activity. We can see that **Snacks Category** dominates every day for every generation.

Gen X (1961-1980) Dominates - Purchase activity high all over the week but peaks on Friday and Saturday

Millennials (1981-1996) prefer weekdays – Balanced trend through out the week, slight peak on Monday and Friday. This means that they prefer to shop on weekdays (possibly at the start or end of their work)

GenZ (1997-2012) Almost absent - The almost complete absence of data points for Gen Z raises questions. Are they genuinely less active shoppers, or is this a data collection issue? Further exploration into these demographics is needed to see if they're underrepresented or if their purchasing habits are different (e.g., prefer online shopping).

Older Generation (1900-1960) most even spread - Their activity declines on Wednesday and starts to rise again towards the end of the week.

Gen Alpha (2013 +) – Only one point on Saturday, this is expected, given their age and minimal purchasing power.
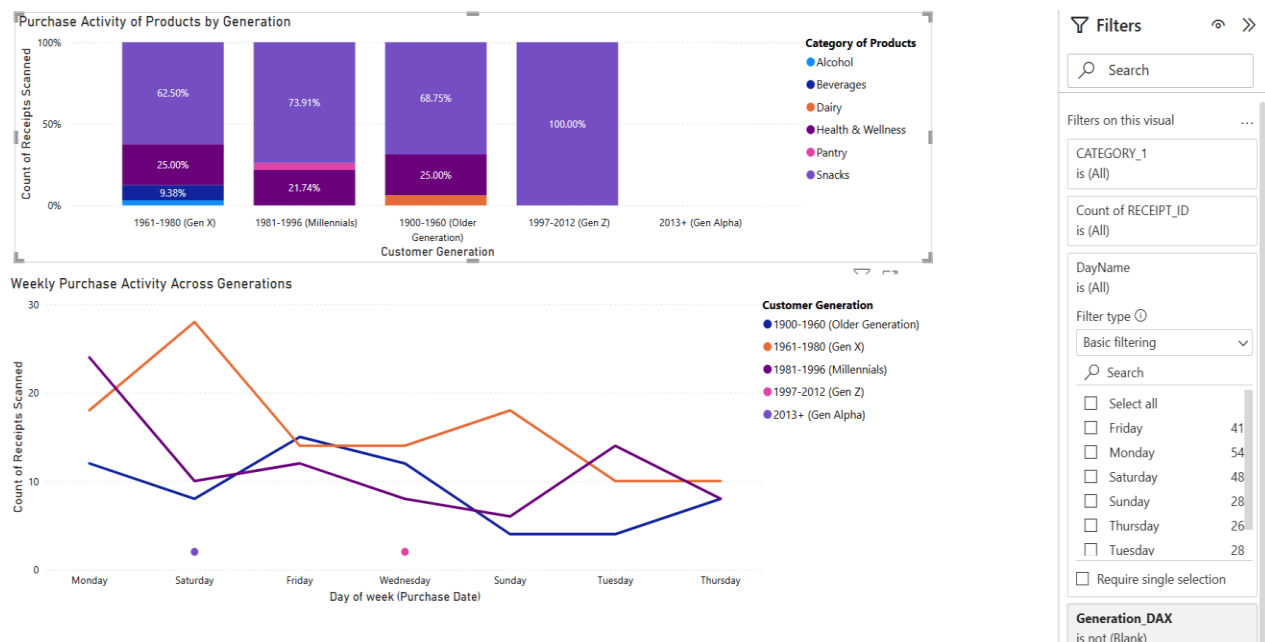
Trend Analysis:

- Interestingly, the **weekend does not show as strong a dominance as expected**. For some generations (like Millennials), weekday sales on **Monday and Friday** are nearly on par with weekend shopping. This suggests that workweek promotions might be effective.
- **Tuesday shows the least activity** across all generations. Identifying why customers are less active on Tuesday and implementing strategies/campaigns to attract customers on this day might help boost overall sales.

- Across multiple generations (Gen X, Millennials, and Older Generation), **Friday** and **Sunday** emerge as crucial shopping days. We could consider targeting promotions or discounts for these days to maximize sales.

Outstanding Questions & Request for Action:

To ensure accurate analysis, I would appreciate further clarification on the following questions and a request for action on all the data quality issues that I have mentioned:

- **Final Sale Discrepancy:** There are instances where the **Final Sale** value is zero while the quantity is non-zero. Could you confirm whether this is due to discounts, scanning errors, or another business logic?

  Assumption: I considered it to be due to discounts and so did not eliminate such values for analysis

- **Final Quantity Interpretation:** Given that Final Quantity contains float values, should we interpret it as **product weight** rather than the number of items purchased?

  Assumption: I considered it to be product weight since it has float values in it for our analysis

- **Barcode-Multiple Manufacturer Relationship:** In cases where **the same barcode is linked to multiple manufacturers**, is this an expected business practice, or should barcodes be unique to a single manufacturer?

  Assumption: I considered there might be other datasets or business reasons for not providing unique barcodes for products for my analysis.

- **Users born between 1900-1960** : There is a significant number of users from this generation with recorded valid activity. Could you confirm if this is expected, or if there may be instances of potential data inaccuracies, such as fake accounts or erroneous data entry?

These are my initial thoughts on the data. Thank you for your attention and continued support. Your insight on these questions would be valuable in refining the analysis. Please let me know if a meeting would be helpful to discuss further.

Looking forward to your response.

Best regards,
Likhita