*As per the fourth task of the challenge, below is an email drafted to a product leader*.

Hi Josh!

Hope you are doing well. This is Vishruti Patel and I'm an Analytics Engineer here at Fetch. In my current project, I've been working on various datasets for analysing Fetch Rewards. And I wanted to update you on some key observations and bring to your attention some data quality issues that I've discovered while working with Users, Brands, and Receipts datasets.

To begin with, in accordance with Users data - The data shows active consumer engagement, especially from users in Wisconsin, with regular logins and receipt submissions. However, the user data seems to contain repeated entries with identical details which suggests potential issues with either data duplication or incorrect data entry. This leads to my first question about data: Is there any possible way where we can find where the actual data is coming from and how it was collected? This is important for us in order to ensure that each user is uniquely represented in our system for accuracy reasons.

Additionally, when reviewing the brand data, I noticed multiple entries with brand codes "TEST BRANDCODE" alongside items flagged as new. Is there a possibility that these multiple entries could refer to test entries that are never updated with real data? If that is the case, it's crucial that we clean up this data to avoid any confusion when referencing actual product brands.

The receipt dataset reveals few anomalies as well, relating to instances where product descriptions appear as "ITEM NOT FOUND" and some occurrences of flagged barcodes such as "4011". In such cases, do you prefer applying any specific logic to categorize these entries, or would you prefer escalating them for further review?

All the above stated issues and inconsistencies became apparent to me during our routine checks on newly submitted receipts and while validating product data against expected classifications. To effectively address these data quality issues, I would require clarification on several key aspects. For instance, as stated above:

- Cases where product descriptions appear as "ITEM NOT FOUND", should we implement a fallback mechanism?
- For certain barcodes that have been flagged, could you provide any insights on whether they should be excluded or needed to be validated manually?
- Are there any predefined rules for brand-category associations or does our classification model needs to be refined to better align with business expectations

To help optimize the data assets, it would be beneficial to have clear guidelines on how to handle flagged or unrecognized entries, especially in the brand and receipt data. Additionally, understanding how to manage and resolve duplicates in the user dataset would improve the accuracy of our reports.

As we move towards production, there are some performance and scaling factors that need to be considered. With growing data volumes, the receipt data and user interaction continues to grow as well. One key challenge is ensuring efficient processing of high-frequency receipt

submissions without causing delays or system bottlenecks. To address this concern, we may need to optimize our data ingestion pipeline, potentially using batch processing or real-time streaming solutions to handle spikes in data flow. Moreover, when dealing with large-scale receipts and product lookup, database performance and query execution time could become a bottleneck. As a solution for maintaining optimal performance, I suggest we can use indexing strategies, caching mechanisms or database sharding.

Another concern could be maintaining data consistency and integrity, particularly in handling missing product descriptions and flagged barcodes. To solve this concern, I believe we could implement automated data validation rules and anomaly detection which can help identify and rectify inconsistencies early in the pipeline.

I look forward to your thoughts and any additional insights on how we can address these issues. Please let me know if you need any further details in the provided findings, I'd be happy to sync over a meeting.

Best regards,

Vishruti Patel