Dear, [Product/Business Leader],  
  
I hope this message finds you well.

As the new Analytics Engineer on board, I've been analyzing our datasets on brands, receipts, and users extracted from our data lake, aiming to address certain business inquiries. My examination has revealed some initial findings and prompted several questions.

Firstly, I would like to understand the business processes that define brands, receipts, and users. Understanding their interrelation within our operational framework is key to developing an accurate data model, enhancing data reliability and integrity.  
  
In order to create a data model to support leadership's analysis and answer business questions, I used Python for data loading and exploration and SQL queries to answer business questions. While doing so, I have encountered notable data quality issues.

Particularly the prevalence of duplicate entries of user\_ids in user dataset. Clarifying the mechanism of user data extraction from our source systems would greatly aid in rectifying these discrepancies.

Additionally, from the data, it's clear that each receipt is linked to a user via a user\_id, showing who made each purchase. However, linking brands to receipts has been challenging. Each transaction listed in the receipts table can include several items, each with a unique barcode. While each brand in our brand data is also associated with a barcode, tracing the exact brand for each purchased item hasn't been straightforward. I'm keen on understanding how to reliably connect these dots, ensuring every item on a receipt is accurately matched to its corresponding brand.

In addressing performance and scaling concerns for our data processing in production, transitioning from Python notebooks to a more robust ELT (Extract, Load, Transform) pipeline is crucial. Leveraging a modern data warehouse like Snowflake, combined with powerful data transformation tools, can significantly enhance our data handling capabilities.

Incorporating dbt (data build tool) into our ELT pipeline adds immense value by enabling data testing, which can help prevent issues like duplication. It also offers benefits like comprehensive documentation and insight generation, aiding in understanding our datasets better. Moreover, dbt supports lineage tracking, which is crucial for visualizing data flow and dependencies. Integrating dbt with Git allows for seamless collaboration among team members, making it easier to manage versions and review changes. This combination of tools ensures that our data infrastructure is not only scalable and efficient but also robust and collaborative.

Please feel free to reach at your convenience to discuss further.

Thank You,  
Pranay Gawas