Fourth: Communicate with Stakeholders

Construct an email or slack message that is understandable to a product or business leader who isn't familiar with your day-to-day work. This part of the exercise should show off how you communicate and reason about data with others. Commit your answers to the git repository along with the rest of your exercise.

Considering Seigel Lindsey as the business stakeholder and the email recipient.

Subject: Data quality analysis & business rules discussion.

Hello Lindsey,

From converting the raw data (.json) into excel format, creating database model and conducting extension exploratory data analysis, following are some of the data quality issues & business rules to be discussed:

I. Brands dataset:
   a) Missing values: Top_brand and Category_code are two variables with >50% null values. Top_brand (The Boolean indicator) shows if the sold products are top brand or not, which can help in analysing purchase behaviour and choices of users.
   b) Data collection: CPG variable values are collected as dictionary (undefined key-value pairs) with no guidelines or business rules documented.

II. Receipts dataset:
   a) Missing values: We have multiple null values for instances where Points_earned is "Null" i.e., the ReceiptRewardsStatus is {flagged, rejected, pending, submitted}, this leads to collecting irrelevant data and needs refined business rules to fasttrack the request or trigger red flags for such cases.
   b) Invalid instances: There are few records with very high points earned , while the total amount spend is well below the points earned (*which is rather impossible), also they have RewardsReceiptStatus as "rejected" or "flagged". Again, creating business rules and trigger points for such cases will help to clear requests faster and have better data quality for analysis.
   c) Distribution (Numerical Variables): Points_earned, PurchasedItemCount, and Total_spent have right skewed distribution (more zero values) with outliers well outside the acceptable range, which decreases prediction model efficiency and data quality.

d) RewardsReceiptItemList variable is collecting all information on product purchased and necessary user details. The barcodes scanned for the products in each receipt, does not match with any barcodes defined in the brands table, which will further hinder to connect the records across tables for analysis.

III. Users dataset:

a) Majority users are from Wisconsin and signup through email source, which tells us the market demographics and helps us target the users.