Email, Communicate with Stakeholders

Subject: Identifying Data Quality Issues - Discussion Request

Dear Mananger’s Name:

I hope this email finds you well. I wanted to bring your attention to some important observations I’ve made regarding our data quality. After conducting a thorough analysis, I’ve identified certain areas that require our attention and possibly some corrective actions.

The main quality issue present in our data is the occurrence of missing values. Across the users, brands, and receipts datasets, some columns have missing values exceeding 10%, with the highest figure reaching up to 52%. Apart from the overall quantity of missing values, there are also specific details that require attention:

- In the brand dataset, the number of missing values for certain features is unreasonable. For instance, the count of missing values for 'categoryCode' is significantly higher than that for 'category'.
- In the receipt dataset, certain columns exhibit a notably high proportion of missing values. However, due to the presence of meaningful zeroes, the actual count of missing values should be somewhat reduced. For instance, 'pointsEarned' and 'bonusPointsEarned'.

In addition to the issue of missing values, our data also presents the following data quality issues:

- Excessive duplication is observed within the user dataset.
- Within the brand dataset, the 'category' column contains duplicated or similar values.
- The receipts dataset exhibits uneven data distribution.

I have undertaken initial steps to address the aforementioned issues. I believe that discussing these findings with you would be highly beneficial for further steps. Please let me know a suitable time for us to meet or discuss this matter.

Thank you for your time and consideration.

Best regards,
Ruoxin Wang