Subject: Data Quality Issues Discussion

Hi [Leader Name],

I hope you are doing well. This is Stacy. I'm writing this email to discuss the data quality issues I found in the Receipts, Brands, and Users data, which I believe are important for you to know.

1. The **duplicate records** were found in the **Barcode** column under Brands data and in the **user ID** column under Users data. Barcode and user ID should uniquely identify brand and user individually. I highly suggest removing duplicate records to avoid saving redundant records in the database. Please also note that we should remove the duplicate records carefully since I found some of the duplicate records were classified under different categories separately.

2. The **missing values** were found among Receipts, Brands, and Users data. Large-scale missing values may result in misleading results and bias. **CategoryCode** and **topBrand** column in Brands data has over 50% missing values. Missing category code values may affect brands classified incorrectly. Missing top Brand value makes it lose its classified function of whether a brand is a top brand or not.

   FinishedDate and **PointsAwardedDate** also have a large percentage of missing values. About 49.2% finished date value missing, which means almost half records we don't know if the receipt finished processing. Same as points awarded date missing. Almost 52% data is missing making us hard to know if we already awarded points for the transaction. In this situation, we are also hard to determine whether the issues are either coming from system processing error or just taking time to process.

3. User **CreatedDate** in Users data contains date/time from 2014 to 2021. However, there are a few years gaps that were not included in the data which are 2016, 2018, and 2019. I would recommend going back to the database to check whether theses missing years exist in database or whether the data downtime happen. Data downtime would cause erosion of data trust and even the risk of revenue loss.

4. **BonusPointsEarned** and **PointsEarned** columns in Receipts data not only exist a large proportion of missing values but also exist unreasonably large points earned in certain records. For example, there are 2 receipts having the same total spent of $1 dollar, but one receipt earned 5 points, and another one earned 500 points. I strongly recommend checking the rewards process of the system to see if there are some logical errors existing or if the rewards events make sense.

There are a bunch of data issues found when I explore the data. To solve the data quality issues, I'd like to have a meeting with you to discuss the issues further in detail and talk about my plan. Please let me know what time works best for you.

Warm regards,
Stacy