Hi XYZ,

My name is ABC from XXX department. I was recently provided with some data to process and review, and I have some questions and thoughts I would like to share with you.

When I processed the data, I firstly looked at the size, columns, data types, amount of Null values and duplicate records of each table, and then the relationship between tables. The major data quality issues I found are listed as below:

- Data type and format are not consistent (for example: date is not in the datetime data type).
- Tremendous amount of null values (in some tables, some columns have more than 50% of null values).
- The id column which is supposed to be the primary key, however, has duplicate values (for example: the users table has multiple rows for the same user id).
- It contains contextually conflicting information (for example: for the receipts whose count of purchased items is zero, the bonus points are still approved).

To be able to solve the above issues, I have some questions here to better understand the data:

- The business importance of the columns which has more than 10% of missing values - We would either remove the column or add data validation constrain that the input can not be null to avoid the redundancy of the table and potentially reduce the storage.
- The definition of certain columns and their relationship. Knowing that will help answer the questions like, what is the reason why the brand name and the brandcode is not 1-to-1 matched.
- What is the business goal you anticipate from analyzing these data sets.

To increase database performance, I would use indexing to optimize the query execution. I would also defragment data when needed, like what I did for separating the 'rewardsReceiptItemList' column from the 'receipts' table. These all help to increase the efficiency.

Thank you for reading my email and please let me know if you need any further clarifications. Look forward to your feedback and thought.

Best,
ABC