Hello Stakeholder,

I am writing to communicate the issues and insights that I found while developing the business case I was assigned on. While doing data quality analysis, I identified two main issues in the data sources:

- Missing values:
  - The percentage of columns in the different datasets that have 40% or more of missing values is as follows:
    - Users: 0%. However, for this data source there are two columns with 11% of missing values.
    - Brand: 25%.
    - Receipts: 60%.
  - For some of the columns in the receipts dataset, having a missing value makes business sense. For example, a transaction which have a missing value in the points column represents that the user did not get any point in that transaction. However, it might be more appropriate to change these kinds of policies and use much more representative values.
  - It seems that most of the missing value problems come from integration issues, systems errors, or processing errors in the ETL process as there is no relation between the missing values and the rest of the data.
- Duplicated values: The users dataset has multiple rows that represent only one user.

In the spirit of solving the current issues, I might require information about the dataflow across the different systems and information about the data warehouse design and its data management policies. Also, I have some enquires about the data:

- Is this data being generated inside the company or does it come from third parties?
- Is there any initiative or effort of data governance around it especially when it comes to integrating systems?
- How is the data quality being monitored currently?

Problems that might come up

The excessive amount of redundancy in the tables will cause performance and scalability issues. First, having many duplicated rows will slow the process of deleting and updating records. Second, the system will need to scale faster in memory and processing capabilities, which result in unexpected costs. On the other hand, the missing values might cause bugs or unexpected behaviors when interacting or integrating with other systems. Additionally, the applications developed to handle this kind of data will need to be more complex as they must handle redundant and missing data.

Please feel free to reach out to me for any clarifications.

Best Regards,

Guillermo Lobaton.