I hope this e-email finds you well. I want to provide some insights on the data I have been working with recently, the data was in the form on JSON files which normally is an important element of the dataset, because it is one of the forms of the raw data.

While performing the analysis and data processing I came across of the several data quality issues which impacts the accuracy & reliability of the data.

The quality issues are listed below:

Missing Values: Some fields contain null or missing values like null which potentially leading to incomplete analysis.

Inconsistent Datatypes: Having an inconsistent datatype affected the analysis and calculations. To get a smooth and error free analysis the datatypes must be common.

Nested Data Structure: The JSON files contain nested structure, making it challenging to extract and analyze data efficiently.

Data Integrity concern: There were inconsistencies and errors in the data, such as formatting issues and duplicate records. Few other issues related to data redundancy and schema evolution.

Since the problem statement is not known the data imputation becomes difficult since there are so many variables with different datatypes, which variable is important for the decision making is unknown and hence analysis becomes difficult. Based on the KPI we can use Mean, Median, Mode to fill out the missing data.

In the given data there is no common thread between brand and receipt table and hence becomes difficult to link, that is why the main concern was the Primary key and foreign key related. We have created another table called “Reward\_receipt” to make sure the table is interlinked. Also, the barcode information given in the table Receipt and Reward\_receipt did not match with each other which is another concern.

Some solutions to avoid Data related issues -

* Coordinate with Stakeholders and know what the exact problem statement is.
* Develop data cleaning and transformation pipelines using tools like Python libraries (Pandas, NumPy) to preprocess and clean the data.
* Implement automated data validation scripts and checks to identify and flag data quality issues during ingestion and processing.
* Collaborate with domain experts and data engineers to validate data assumptions and refine data processing workflows.
* Document data quality rules and procedures to maintain consistency and transparency in data processing.
* Collaborate with data engineers and analysts to validate and optimize the data assets.
* Enhance data transformation processes to handle nested structures and optimize data extraction.

I believe addressing these data quality issues and optimizing our data assets will enable us to derive more valuable insights and drive better-informed decisions for the business.

Please let me know if you have any questions or if there's anything else you'd like to discuss regarding our data strategy and initiatives.