Hello,

I am writing this email to illustrate the issues I faced while processing and cleaning the data. Data quality is a key component of your business’s long-term success, especially in the data-driven business world we live in. High quality data can drive better customer experiences, increasing retention and driving higher top-line revenue; poor data quality, meanwhile, leads to analytics problems and insights that don’t accurately reflect customers, misaligns moments of engagement, and creates negative brand experiences. All these consequences can lead to missed revenue and extensive challenges in an increasingly competitive world. I have outlined the data quality issues and possible solutions to it.

Resolving Data Quality Issues:

I have looked at the data in two parts:

1. Cleaning the data
2. Structuring the data

**Data Cleaning Issues:** Data is in JSON format. Read the data in python and used Jupyter as the IDE. There are multiple parameters that had to be cleaned like Ids, Datetime etc. One of the major issues was with receipts data. It had a column named receipt item which had 18 more attributes that could be used for further analysis. Therefore, I extracted the column to use it as a table. It has a foreign key named receipt id that can connect with receipt data. Similar aggregation had to be done for users and brands data.

Upon checking the percentage of missing data, there were many areas where data needs to be corrected.

* In receipt data, total spent attribute had 38.8% missing data that caught my eye.
* In brands and reward\_receipt\_items data, the one attribute could be highlighted is brand code. It had 20% and 88% missing data. This also impacted results in finding the brand that had most spend and transactions. Most of the count was null or blank
* Users data has 28.6% duplicate data. Users and receipt is one to many relation. One user can have multiple receipts but one receipt cannot have multiple users

**Structuring the data**: Reward receipt item list was one such attribute that had to be used as a table as it had 18 more attributes in it. Data had to be normalized and structured to make sure that it is scalable.

Future case: If this is the data that we are getting from the source system. I would automate the whole cleaning and structuring process by using an ETL tool like Airflow to ingest data into a data lake like redshift.

* Help in maintaining the data in a data lake thereby increasing the performance and scalability.
* Other team member and colleagues can get access to the data.
* Different business purpose could be solved.
* Integration with machine learning, given the schema less structure and ability to store large amounts of data.
* Flexibility. A DL allows you to create large heterogeneous, multiregional, and microservices environments

These were some of my findings and insights. I hope I was able to communicate my ideas.

Thanks,

Nikhil Kamath