**Hi,**

Hope this email finds you well!

After performing data modeling and data cleaning regarding the dataset, I find out there are some data quality issues. Please see below for the questions that I have, the process that I took to find out the issues, the information I need to resolve the issues, the concerns about the anticipation in production and some of my suggestions.

**Questions about the data**

1) We’re missing a lot of data in the dataset.

2) The data formats are not fully defined and are not consistent, which will be hard for the data analyst to perform his/her work in the future.

3) There are so many duplicated data in each table.

**Process of discovering the data quality issues**

Step1: select a data sample to review.

Step2: review the NULL values of the data.

Step3: review the format of the data.

Step4: write some SQL queries, like COUNT() clause, to check if there are any duplicated rows, then compare the results with original data.

**Things that we need to know to resolve the data quality issues**

1) I need to know where the dataset came from and who’s responsible for creating the data, so that I can know what happen to the missing values and get the original data to correct.

2) I need to know the rules of data modeling from our data engineers.

3) I also need to know the standard format of the schema, so that I can know what I should do for the incorrect data.

**Other information that you need to help us optimize the data assets we’re trying to create**

1) A separate dataset that contains the information about itemlist which is related to receipts and brands.

2) The data that we can delete directly from the existing tables.

**Performance and scaling concerns that I anticipate in production and the way how I plan to address them**

1) The data format not defined correctly will lead to spending a lot of time doing data cleaning. Thus, we have to set the rules for the data format before we do the data cleaning.

2) Receipt table containing so many columns will lead to confusions, and it is better to separate the receipt table from itemlist table.

Please let me know if you have any questions and I will get them answered as soon as I can!

Thanks,

Sankai Zhou