Communicate with stakeholders

To complete this task, I am drafting an email to the Senior Manager Mike, ensuring all the data quality issues, ideas to tackle them, scaling concerns are clearly communicated.

Subject: - Data Quality and assurance issues and ideas to resolve them

Hello Mike,

I hope this email finds you well!

I am reaching out to you regarding some major concerns discovered in the Users, Receipts and brands data based on my thorough and deep exploratory data analysis. I want to provide you a brief description of the issues in all the data sets and provide clear overview of the data along with ideas to tackle the data related issues.

Firstly,

I want to list out the data quality issues in each dataset and ideas to tackle them.

1. **Missing data in certain columns: -**

    - The fields in the receipts data namely **'rewardsReceiptItemList'** (39.32% null values) and **'totalSpent'** (38.87% null values) and **'purchasedItemCount'** (43.25% null values) are some of the important field's w.r.t the business aspect which contain missing values. The missing data in these columns indicate that certain receipts did not have items purchased and hence the spent amount and purchased items count associated with them are also missing. This affects various business KPI’s which help in understanding the revenue gained by users and top items purchased by the users.

- The fields namely **‘bonusPointsEarned’** (51.39% null values) and **‘pointsEarned’** (45.58% null values) indicate that certain receipts did not include bonus points earned due to data capturing or human errors. These anomalies in data the respective data fields affect the business decisions like pricing, marketing strategies, cashbacks, and rewards planning.

- The fields ‘**Last_login’** and **‘state**’ in users data have significant null values. This is a major concern as it will affect in identifying the active users of the app and other KPI’s such as user activity trends and time spent on the website/app.

- The **‘topBrand’** column in brands data has significant missing values which affects in understanding the top brands of the business. Furthermore, this affects business decisions made in marketing, discounting and pricing of items.

These missing values issues can be handled by data governance, standards, business rules and validation. Also, efficient quality assurance checks in the databases using efficient and optimized SQL queries helps in tackling the issues.

2. **Inconsistency in data types: -**

- Most of the date fields in the data are not formatted based on the default format which can affect business model and KPI’s. This issue needs to be tackled by efficient handling of data sources, extraction and data transformation using various tools.

3. **Data redundancy and vague data(outliers)**

- The users data has more than half of redundant data, which needs to be resolved by data quality, entity relationship check and SQL query checks in the databases.

- Some of the values in the data vary largely from the related data values and do not make sense in terms of business. This needs to be evaluated to ensure there are no bugs in the app.

Furthermore, I would love to collaborate with the data engineers of the team to understand the data sources and data pipelines in depth to resolve the data issues and optimize the databases. I would also be very keen to understand the database design and business rules based on the data to preform root cause analysis for the issues regarding the anomalies and discrepancies in the data.

I strongly believe that it is critically important to handle these issues as they may affect the production and business scaling in the near future as the business grows further and more data needs to be handled.

Finally, I would like to get on a call to brief you about the concerns regarding the data in depth. Request you to let me know which time works best for you!

Looking forward to hearing from you!

Thanks,
Karthik
Data analyst