Hello XYZ,

Thanks for sharing the data files of Receipts, Brands & Users. Based on my initial data exploration of the provided data below are some of the key insights observed which I need to highlight.

- **Quick Questions Regarding the Data:**
    1. It was surprising to see the receipts having totalSpend == 0 and some points being earned on the receipt scan. What kind of scenarios are these? I assume that the totalSpend == 0 matches the transactions in which the gift card or promotions were applied by the user.
    2. Business difference between submitted - finished rewardsReceiptStatus & rejected - flagged rewardsReceiptStatus?
    3. The data includes some of the test accounts created by the fetch staff, are these test accounts supposed to be cleaned before data loading?

- **Data validation techniques to check the Quality of Data:**
    1. Identifying the Missing values in the attributes for the Users, Receipts and Brands data file.
    2. Visualizing the data attributes to understand the frequencies in case of discrete attributes.
    3. Studying the descriptive statistics of mean, median, mode, min & max of the continuous numerical attributes.
    4. Understanding the distribution for the numerical data attributes.

- **Resolving the data quality issues:**

    Below mentioned are the key points we need to address before we proceed for resolving the data quality issues:

    1. The user's data is more biased towards the state of WI, more details needed as to understand why this kind of behavior is observed.
    2. Significance of the test accounts that were created by the fetch staff.
    3. Additional details around the cpg attribute in the brands data file.
    4. Since this data is limited to specific months, we need to know the mean/median/mode of the data attributes like (bonusPointsEarned, purchasedItemCount etc.) in the receipts data over the whole customer base in order to impute the missing values/NULLS appropriately.

- **Optimizing the data assets:**

    I will be needing the below information for appropriate optimization of the data assets.

    1. Understanding of rewardsReceiptStatus Lifecycle.
    2. Details around what % of the brands are topBrands over the complete database.
    3. More data for additional 6 months that will help in interpreting better results.
    4. What were the different digital marketing strategies being adopted that resulted in maximum User signups via email.

- **Anticipated performance & scaling concerns in production & Resolutions:**

Below are some of the considerations for the proposed data warehouse schema w.r.t performance & scaling concerns:

1. As we scale more data would be coming up in the Receipts_Scanned_Fact, purchased_Items_Dim tables and with new users coming to the platform the Users_Dim table will also be heavy on records so efficient table indexing and table partitioning for these tables need to be done for efficient reading of data.
2. The data tables that will be created while designing the Datawarehouse will have appropriate key constraints that will restrict the garbage data records thus improving the efficiency of the end data analysis.
3. As we scale the database can slow down while extracting the results thus SQL queries will be further optimized by studying the execution plans.

I look forward to learn more about the next steps with the implementation.

Thanks,
Ashish