Subject: Insights & Collaboration Needed for Data Quality & Performance

Hi Team,

I hope you're well. I've been working with our datasets and identified key areas where we could use your insights:

## I. Required Info to Optimize data set

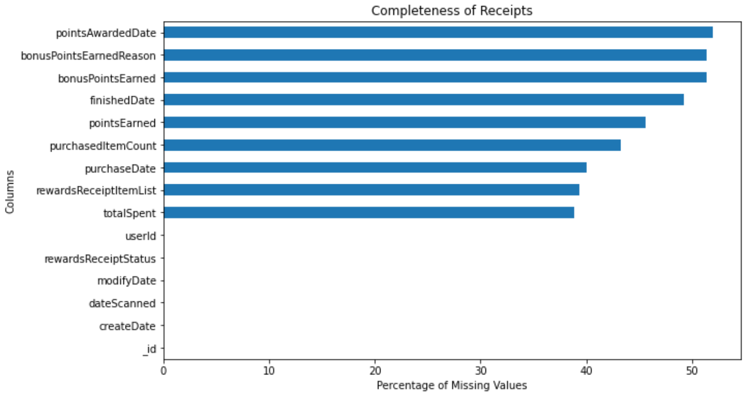
a. Data Optimization: Is a unique reference for Consumer Packaged Goods (CPG) necessary? Also, could you share the 'receiptItems' data schema?

b. Data Redundancy: We have both 'category' and 'category code' columns. Is there a specific reason for this duplication?

## II. Data Quality Issues

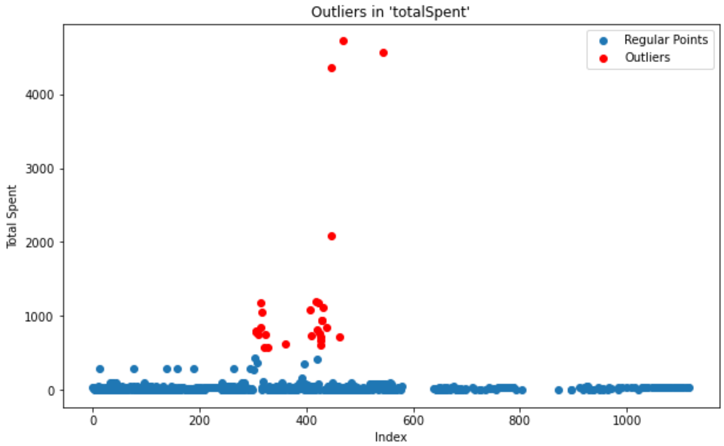
a. High Missing Data Rate: To address the over 40% missing data rate in six variables of the receipts dataset, we need:

1) An explanation of the data collection process and potential reasons for missing values,
2) An assessment of data source reliability
3) Clarification on whether any missing values are intentional for business reasons.



b. Outliers: I found significant outliers in transaction amounts, with the red dots in the graph indicating transactions far from the normal range. we need:

1) Access to data validation tools to ensure accuracy.
2) Insights from business analysts to understand transaction norms.



3) Duplicate Records: To resolve the issue of 283 duplicate IDs in the user dataset, we need criteria for identifying unique users during data processing.

## III. Performance and Scaling Solutions

1. Reduce Query Times: Enhance or update database indexes for faster queries.
2. Manage Data Growth: Adopt distributed storage for scalability.
3. Improve Scalability: Utilize caching for frequently accessed data under high user load

Your perspective would be invaluable in addressing these points. Let's discuss how we can work together to enhance our data practices.

Best,
Lili