Hi Ibrahim,

Happy Monday!!

My name is Aman Chouksey I am from the Data Science team at Fetch Rewards. I have been working on extracting the users, brands and receipts data for creating a pipeline to store the data into warehouse and make it ready for analysis. As a staring step I did a throw Exploratory Data Analysis and found out some discrepancies that I would like to highlight.

1. All the three data has significant amount of missing values and duplicate values as well, such cause of data discrepancy generates a biased dataset, I am guessing that probably either the system which is capturing the data is mulfucntioning or its the human error while entering, either ways its impacting our business.

2. The user data receipts data has duplicate values with approximately 57% duplicate values in user data and hence a major reason for data redundancy that is costing us a lot when it comes to paying for data warehouse thus, I recommend that we should explore other potentials datasets as well to reduce such issues.

3. Additionally the id columns in the datasets are a combination of numeric and alphabets, thus when dealing with multiple datsets and joining them it might misguide our analysis, thus it would be better if we can create numeric unique id for each datasets.

4. Also the json files captures multiple datasets in the same data files, anyways the amount of space required to store that data remains same thus it would be better if we could capture them separately like, the receipt file has a column which has dictionary as values which contains information about another rewardsReceiptItemList thus I suggest a new snowflake scheman for proper dataware house management

I am still trying to explore the data and thus would be obliged if someone from the team could connect with me to help me understand the data and how the team uses it further this would help me model the data better.

Looking forward to hearing from you.

Have a nice day!!
Regards,
Aman Chouksey