Hello [Business Leader]

Here are my findings after analyzing the User, Brand and Receipt data shared with me.

I mapped the JSON data provided into a data model. Based on the data model I tried to identify potential data quality issues that may occur in the future.

**Questions about the data:**

* The format in which the data was shared didn’t seem valid ( JSON files shared had data appended one after the another)
* The data provided seems to be nested in places there for values like orderID(oid) which makes it complicated to normalize and map the data. Are these nesting values expected and what significance do they have?
* The data associated with Users has a duplicate information which can potentially lead to additional errors and missed opportunities when trying to analyze metrics and KPIs critical to the business.
* The data associated with Brands has a lot of missing data particularly the "BrandCode" which is a key element used to create relationships with our Receipt data. Is the BrandCode auto generated or manually populated?

**Discovery of data quality issues:**

* Duplicated data in Users was discovered by extracting unique UserIDs and then making sure that there are complete cases for each unique user.
* Since BrandCode is essential when creating relationships with Receipts checking for missing data is the first step. Upon checking about 20% brand data has missing BrandCode.

**Additional information needed to resolve the data quality issues:**

* A better approach would be to use a JSON list of objects instead of appending different JSON objects.  
  We would prefer flat JSON which can be easily ingested to a database for query retrieval or converted to a CSV.
* Making sure the User data is free from duplicates we must ensure to create constraints on UserID to make it a primary key.

**Optimizing the Data Model:**

* We can make BrandCode the primary key or mandatory when ingesting data to avoid missing values.
* This is to ensure the PrimaryKey-ForeignKey relationship with Receipt data is maintained accurately without loss of information.

**Performance and scaling concerns:**

* As the volume of data increases the nested JSON format can make it difficult to parse and ingest data to maintain integrity. This was already addressed in the script provided by using json\_normalize to flatten the nested JSONs
* Additionally, using a valid JSON format is necessary to make data processing easier. JSON list of objects should be used instead of appended JSON objects.
* When the volume of data to be ingested and analyzed is overwhelming traditional RDBMS we can use cloud technologies like AWS/Azure to store them in a data lake and use parallel processing to address performance and scaling concerns.