1. **Review unstructured JSON data and diagram a new structured relational data model.**

Power query tool in Excel allowed me to import the data from the JSON file and identify unique keys present in the Schema.

Following are images for the reference:

Graphical user interface, application, Word

Description automatically generated A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, application, Word

Description automatically generated Graphical user interface, application, Word

Description automatically generated

A new structured relational data model is attached to a github repository.

With the help of the Power Query I was able to clean, transform and reshape the data as needed and converted the data to CSV format. CSV format made easy to import the data in Mysql Workbench.

Steps for Transforming Data in Power Query:

* Converted the records into table.
* Checked the data type of each column in all the three schemas.
* Cleaned mixed data type and assigned correct data type.
* Converted Unix time into MM/DD/YYYY format.
* Custom column for date - #datetime(1970, 1, 1, 0, 0, 0) + #duration(0, 0, 0, [#”lastLogin”]/1000)

convert the number of seconds to the number of days and then add it to the date value of 1/1/1970.

* Removed duplicates from the table.
* Replaced empty data with the null values.
* Changed the name of all the columns to Standard column anmes.

I used ER diagram to create a new structed relational data model.

1. **Generate a query that answers a predetermined business question.**

I used MySQL workbench to write queries requested by business stakeholders. File attached to a github repository.

1. **Generate a query to capture data quality issues against the new structured relational data model.**

I used Python to discover the data quality issues such as

* Count of Missing values in all columns.
* Duplicate data
* Data Entry Errors
* Invalid Data type
* Geography names like names like state.

In Users Schema:

* Four columns have data type as object – need to assign correct data type such as Character or String.
* There is more than 50% duplicated data.

In Brands Schema:

* Barcode column has 7 duplicate values.
* All the columns have data type as object - need to assign correct data type such integer, float, boolean, character/string,etc.
* Category and Categorycode have different unique values
* Data entry errors: barcode is entered in brandcode.

These data quality issues can be resolved by performing one of the following strategies:

* Replace incorrect values or missing values by mean, mode or median.
* Alternate Data Source.
* Estimating incorrect/missing values.

You will find more details on functionality in the jupyter notebook attached to a github repository.

1. **Write a short email or Slack message to the business stakeholder.**

Hello [Nameofstakeholder],

I hope this email finds you well. I understand you are concerned results and I would be glad to provide as much information about the project status as I can.

I have a couple of questions for you regarding the data:

* What is the ultimate goal or business problem we want our data to answer?
* What standard KPIs shall we use to make profitable decisions to help company thrive?

While answering requests made by you, I discovered the data has poor quality that might provide misleading information. Data has lot of missing values, inaccurate information. When I was finding the top 5 brands for the recent month, I figured that the data is not telling anything about 50% of the brands whether it should be featured as ‘top brand’. Is there any alternate data source? If, it will be really useful for me to resolve the data quality issues.

We have insufficient user data as compared to the receipts scanned and purchases made in the application. So, to get the actionable insights, I need more information on users and also geographical data on users, since many decisions are driven by geographic data. I also need accurate information on category name and category code that are used to classify the products.

I plan to send out a full report soon. This report will include the finalized timeline and share the latest findings, including any new insights we have found.

I understand you are waiting on our report, and I want to make sure that we can get you the information you need. If you would like to meet before the next report, please let me know what time would work best for you.