① Users dataset

(i) Findings :- users dataset has 100,000 entries. It has 6 columns id, Created date, Birthdate, state, language, gender

→ Mainly Birth date, State, language and Gender has missing values.

a) Birth date column has 3,675 null ie, missing data

b) state column has 4812 missing values.

c) language column has 30% of missing data

d) Gender column has approximately 5% of null [missing data]

(ii) Assumptions

here in the dataset created_date & Birth_date have [illegible] String as datatype, We need to Consider as datetime objects

→ There are missing values in language & gender columns maybe because of incomplete user profiles.

(iii) Conclusions

There are lot of missing values in language column. Datatypes needs cleaning. Need to change to datetime datatype for Birth date & created date

Data Quality Issues

① As observed so far from above points, missing value are present in birth date, state, language & gender Columns.

② Inconsistent datatypes for columns like created_date & Birth date

③ There are lot of Blank entries in language & gender column.

Challenges

Here I didnot understand properly about language column As it has code like data [es-519]

② Products dataset

(i) Findings :- The dataset hes 845,552 rows and it hes Seven (7) columns → category_1, category_2, category_3, category_4, Manufactorer, Brand & Barcode.

→ I observed, columns like category_3, category_4, Manufacturer & Brand have missing values.

→ Barcode Column has mostly float values. As we know Barcodes are treated as strings, they are like kind of identifiers.

(ii) Assumptions

I observed that there are lot of category hierarchy in the dataset. But category1, category_2 have few missing values, whereas category3, category_4 hes more missing data.

→ There is also missing data in columns like Brand & Manufacturer maybe because of incomplete information in products dataset.

(iii) ~~Conclu~~ Conclusion :-

I observed approximately 92% of data is missing in category-4 column.

Data Quality Issues

① category_4, Manufacturer, Brand columns have more missing data.

② As observed, barcode column needs to a string datatype, but it is stored as float type.

③ lot of blank entries in Manufacturer & Brand columns.

Challenges

→ It is challenging to understand column category 4, as it hes more missing values.

③ Transaction dataset

(i) Findings :- dataset has 50,000 rows & 8 columns

⇒ Receipt_id, purchase_date, scan_date, store_name, user_id, Barcode, final_Quantity & final_scale.

⇒ lot of missing values in Barcode column. [approx 11%]

⇒ Columns like Final_Quality and Final_scale contains mixed datatypes. In Final_Quality column, contains values like "zero" instead of numerical data.

⇒ Barcode column consists of float values, but this column must be a string datatype.

⇒ purchase_date & scan_date are treated as strings, instead of datetime.

(ii) Assumptions :- Final_Quality & final_sale must have numerical entries.

(iii) Conclusions :-

(i) There are data Quality issues, including missing values in Barcode, mixed datatypes in final_Quality & Final_sale.

Data Quality issues

① missing values in Barcode column

② Inconsistent datatypes in final_Quality & final_sale

③ Invalid values in final_Quality ~~data~~ column

Challenges :- Finding text values like "zero" in numeric fields is bit confusing