3. Communication to the Stakeholders

-> For this question, I am writing an email to Jon – the Data Architect, informing him about the data quality issues I found.

Hello Jon,
After performing an Exploratory Data Analysis of the provided datasets, I came across various data quality issues such as columns with null values, missing values for fields that I believe must be required, outliers/fault data entries, etc. Please find them mentioned below in detail.

1. **Dataset: brands.csv**

   Number of columns: 9
   Number of rows: 406

   Columns with number of null values:

COLUMN NAME	NUMBER OF NULL VALUES
BRAND_CODE	25
CATEGORY	27
CATEGORY_CODE	31
ROMANCE_TEXT	103
RELATED_BRAND_ID	243

   - For some records the BRAND_CODE, CATEGORY and CATEGORY_CODE fields are missing. But every item must be associated with some category and hence this is considered as missing data.

2. **Dataset: receipt_items.csv**

   Number of columns: 12
   Number of rows: 360377

   Columns with number of null values:

COLUMN NAME	NUMBER OF NULL VALUES

DESCRIPTION	1091
BAR_CODE	135369
BRAND_CODE	205488
QUANTITY_PURCHASED	7756
TOTAL_FINAL_PRICE	692
POINTS_EARNED	341425
REWARDS_GROUP	298440
ORIGINAL_RECEIPT_ITEM TEXT	1680

- Total final price is 0 for 6325 rows.
- Total final price is 0 for many items where quantity purchased is not zero or null. This could be due to the coupons, or some offer on the product, but needs to be specified in the dataset.
- POINTS_EARNED = TOTAL_FINAL_PRICE * 10
- Some items have a difference in the total final price for the same quantity and the same brand.
    - Ex: For BRAND_CODE = 'HERSHEYS' and Description = 'Hrsh Mlk Chc Bar Wrp 1.55 Oz' there are 3 different TOTAL_FINAL_PRICE of 0.88, 0.99, 1.19, respectively, which could be due to an offer on the item or due to the location but needs to be specified in the dataset.
- BRAND_CODE is different for same item.
    - Ex: HERSHEYS & HERSHEY'S MILK CHOCOLATE for the same item.
- Outliers in the TOTAL_FINAL_PRICE for the same barcode.
    - Ex: for the brand STARBUCKS:

QUANTITY_ PURCHASED	TOTAL_FINAL _PRICE	DESCRIPTION
1	310059.90	Starbucks Iced Coffee PremiumCoffee Beverage Unsweetend Blonde Roast 48 Oz 1 Ct
1	310059.90	Starbucks Iced Coffee PremiumCoffee Beverage Unsweetend Blonde Roast 48 Oz 1 Ct
1	5.99	Starbucks Iced Coffee PremiumCoffee Beverage Unsweetend Blonde Roast 48 Oz 1 Ct

3. **Dataset: receipts.csv**

Number of columns: 21
Number of rows: 70601

Columns with number of null values:

COLUMN NAME	NUMBER OF NULL VALUES
STORE_NAME	1836
PURCHASE_DATE	2066
PURCHASE_TIME	4947
TOTAL_SPENT	1492
USER_VIEWED	6465
PURCHASED_ITEM_COUNT	1452
PENDING_DATE	1453
MODIFY_DATE	2
FLAGGED_DATE	66576
PROCESSED_DATE	70601
FINISHED_DATE	6252
REJECTED_DATE	66217
NEEDS_FETCH_REVIEW	70276
DELETED	69733
NON_POINT_EARNING_RECEIPT	8986

- DATE_SCANNED is less than PURCHASE_DATE for 6 items or values which is not possible since DATE_SCANNED should always be greater than PURCHASE_DATE (Assuming that DATE_SCANNED is the date when the receipt is being scanned after the purchase).
- PURCHASED_ITEM_COUNT is 0 for many records where TOTAL_FINAL_PRICE is not 0.

4. **Dataset: users.csv**

Number of columns: 8
Number of rows: 164

Columns with number of null values:

COLUMN NAME	NUMBER OF NULL VALUES
SIGN_UP_PLATFORM	45

- As per my analysis, most of the users were born in 1972 followed by 1968.
- Number of users based on their gender is equally split between females and transgenders with 41 each. Males are the lowest with 35. But majority preferred not to disclose (47) their gender.
- Most users are from Florida with 16, followed by New York with 15, and Pennsylvania, Texas, California rounding out the top 5.
- Amongst the known numbers, the number of users signing up was dominated by Android with 53, defeating iOS by 14.

I believe that these data issues must be addressed immediately to ensure we are not missing out on or wrongly identifying patterns and thus ensure better and quality analysis which will serve our consumers better and ensure we are staying ahead of the curve.

Please let me know a suitable time that we can set-up to discuss further and in detail and consult with the other stakeholders to come to a decision on the way forward.

Thank you,
Prachi Yadav