### **Third: communicate with stakeholders**

#### **Construct an email or slack message that is understandable to a product or business leader who is not familiar with your day-to-day work. Summarize the results of your investigation. Include:**

1. Key Data quality issues:

Subject: Data Quality issues

Hi so after reviewing the 3 datasets i.e USERS\_TAKEHOME,TRANSACTION\_TAKEHOME and PRODUCTS\_TAKEHOME a few questions that are in top of my mind and I would like to further discuss and get your clarity on are:

1. In the TRANSACTION\_TAKEHOME data column FINAL\_SALE, some rows have Sale values despite FINAL\_QUANTITY being ‘0’. So my first question will be what does ‘0’ represent in the FINAL\_QUANTITY column. My first thought was Final\_Quantity being a categorical column but in the schema we have that as “NUMERIC”
2. While validating the common columns between all 3 datasets like BARCODE in TRANSACTION\_TAKEHOME and PRODUCTS\_TAKEHOME one of our findings are some of the BARCODES in TRANSACTION\_TAKEHOME do not match with PRODUCT\_TAKEHOME data. So is this something we should be concerned about?
3. In the USERS\_TAKEHOME data while analyzing I found a couple of entries where the birth\_date of USER is greater than the created\_date of the USER. Logically this will not make sense, so your inputs regarding this will be appreciated.

So overall OUTSTANDING Questions:

1. Should “0” be valid in the FINAL\_QUANTITY column or there are any errors?
2. To validate our data join any confirmation regarding barcodes not in PRODUCT\_TAKEHOME table
3. Do we need to make an exception about users where there created\_date<birth\_date.

2. Interesting Trend in the Data:

While analysing the data through sql queries one trend we wanted to communicate you is what are the top 5 brands by sales among users having their account since 6 months:

BRAND FINAL\_SALE

COCA-COLA 2592.10

ANNIE'S HOMEGROWN GROCERY 2383.92

DOVE 2327.47

BAREFOOT 2284.59

ORIBE 2085.93

COCO-COLA is the leading brand among users who have had their accounts atleast 6 months from now. We do believe understand such trends is very important in analysis part as it gives us insights on our customer preferences so that we can make strategies revolving around such analytics and insights.

3. Additional help and FINAL questions:

1. Handling Missing values: Should missing values in columns like FINAL\_SALE, BIRTH\_DATE, BARCODE be replaced with certain statistical methods or those NULL records can be excluded based on the business context
2. Relationship across 3 datasets: Should all BARCODES in TRANSACTION\_TAKEHOME exist in USER\_TAKEHOME data? How to handle this anomaly?
3. Duplicate values found in the datasets should they be removed or aggregated based on a certain logic depending on the business needs?