Third Exercise – Communication with Stakeholders

Hello Team!

I am writing to discuss some data quality issues I discovered in the data pertaining to users, products and transactions. Furthermore, I'd love to discuss some interesting trends I have observed while exploring the data. I would also like to seek some clarity regarding some data anomalies that I came across while examining the data.

Data Quality Concerns:

Firstly, before diving into the data analysis, I reviewed the data at hand, specifically focusing on the barcodes of the products. During this preliminary review, I noticed a few flagged records where barcodes started with either "0" or "00," and these entries included an unusual apostrophe (') symbol at the beginning of the barcode. Out of curiosity, when I looked up online for legitimacy, I found that in each case I checked, barcodes starting with "00" were invalid. I'm wondering if there's a specific reason for this discrepancy. Could the apostrophe have a particular significance related to the products, or might this indicate a data entry issue? An additional context would aid me greatly in understanding the data better.

Secondly, I would like to inform you that I followed a consistent methodology to ensure a standardized and an all-round analysis to identify concerning issues related to data quality in all the three sets of data. In this process, I have discovered some major data quality concerns that could lead to misleading business insights.

In Users Data:

1) It was observed that more than 30% of the users' language was unknown. A more concerning issue was that 3.67% of the users didn't have their birth dates recorded. Furthermore, it was observed that around 1272 users had the same birth date that was 01-01-1970. Is this because the application's default date is set to 1st January 1970?
2) Another major find was that when the number of created accounts where the birth dates were missing was compared to the total number of accounts created monthly, a serious data collection concern was noticed. In 2023, the percentage of accounts created with missing birthdates started at 0.5% but skyrocketed to 55.75% by December. This means that over half of the accounts created in December of 2023 had no recorded birthdate. Additionally, more than 45% of accounts created in the last quarter of 2023 and the first month of 2024 were also missing this crucial data. This might be either because of the users having an option to skip to record their birthdate while creating their account or the birthdate is not being recorded properly by the UI. This is a major concern worth addressing! An interesting insight was that out of all the records where birthdate is missing, 97.98% of them have missing gender information and one reason might be that both the gender and the birthdate details were asked on the same page of the application.

3) According to the data it was found that there were 56 users who were above 100 years old at the time of creation of their accounts though none of them used the account for transactions.

One peculiar thing that I also observed when exploring the Users data for potential issues is that there were several instances where 2 accounts were created on the same day at the same time but by different individuals. Could this be due to a system glitch, or perhaps an issue with the way timestamps are being recorded? Is there a possibility of users unknowingly being assigned the same timestamp, or might this behavior indicate some sort of duplication in the account creation process? I am curious as to what could the reason be for this anomaly.

In Products Data:

1) During the initial steps of analysis, I found that 57 records were duplicated.
2) While analyzing I realized that all the products were categorized multiple times in a branch like fashion. This caused the product to remain uncategorized in the subsequent categories if it is not properly categorized in the earlier categories. This is one of the reasons for 92.02% of the products not labelled according to the 4th subcategory. While this process of branched categorizing is neat and organized, is there a better way to arrange these products to prevent a loss of information? Another thing that I would like to highlight is that more than 94% of the products whose barcodes have not been recorded have not been categorized in the 4th subcategory.
3) 26.78% of the products' brand and manufacturer details are unknown. And all these products belong to two broad categories – Health and Wellness, Snacks. It seems that the application is facing a problem in categorizing the products belonging to either of the sectors.
4) Another significant data quality issue I identified is that 8.71% of product barcodes do not adhere to the standard 12-digit format. Is there an alternative barcode format in use that I might not be aware of? And there were almost 4000 products whose barcodes were unknown and most of the products belonged to the snacks and health and wellness categories. Are both the issues mentioned above somehow related?

There were 547 products labelled as 'Needs Review' and when I checked for the brand information almost all those products belonged to snacks and beverages companies. A little more insight into why those products were categorized as 'Needs Review' would be helpful. Furthermore, there were 4 products belonging to the brand 'Henkel' but manufactured by 2 different companies resulting in 2 different sets of identical barcodes. I found it difficult to comprehend the reason behind this issue. While this is a minor issue in the present scenario, drawing business insights might become challenging if there are a lot of products that are labelled similarly.

In Transactions Data

1) There were 171 duplicate transactions present.
2) Out of the 49,829 recorded entries, only a limited number of values were unique across each column. Upon closer inspection, it became clear that some inconsistencies might be skewing the data. For instance, in the 'SCAN_DATE' column, which includes timestamps, only 24,440 of the 49,829 records were unique. This suggests that many items were scanned multiple times, resulting in repeated entries.
3) Just the barcode column had missing values with a lot of recurring values. This could turn into a major issue when merging the tables for analyses.
4) The ‘FINAL_QUANTITY' values and the 'FINAL_SALE' values were ambiguous. The ‘FINAL_SALE’ column had a lot of rows with one space character (“ ”) and the corresponding ‘FINAL_QUANTITY’ rows had a value. Conversely, when 'FINAL_SALE' had values, 'FINAL_QUANTITY' was often zero. Further analysis was performed to understand the columns better.

   I performed a deeper analysis to understand the reason for these discrepancies listed above and I found a major data quality concern. There was a presence of duplicate entries! It seems that every item got scanned more than once at the same time and date during every transaction! An interesting observation here was, wherever the sale value and the quantity were recorded correctly for an item at a particular time, then the rest of the duplicate records that got recorded at the same scanned time had the following changes:

   Either the sale value was left blank, or the sale quantity was mentioned as "zero".
5) There are 2856 transactions where the barcode values of the products didn’t get recorded properly. Apart from that, 63 barcodes did not follow the 12-digit format. This might be because of the application not scanning the barcodes properly or because of an improper use by the customer. On checking for an overlap between these records and the product records, it was found that there were 56 products that were bought which had invalid barcodes.
6) There were 94 instances when the product was scanned before it was purchased. There might be several reasons for this. Either the date and time didn’t get recorded correctly either during the scanning or during purchasing. Another grave concern is that it could be a result of fraudulent activities. Thirdly, it would be because of a software glitch. But I feel this issue needs to be given importance to resolve the unknown bottlenecks.

One main observation that was difficult to comprehend was that there were records which indicated that 2 identical products, bought at the same time in the same quantity by the same user had two different recorded prices. It might be because a user bought two or more identical products but somehow the quantity got entered wrong. Another hypothesis is that the product might have been on offer - something like, buy 1 for 5 dollars and get the second pack for 2.5 dollars. But a little more insight is needed into the matter.

An interesting trend I found while answering the business questions was that the company's growth was the highest during the years 2019 and 2020 with an annual growth in the new users count of 239.16% and 167.79% respectively. It was during the COVID-19 pandemic. It is interesting to notice that while the retail industry took a hit during that time, the online services dependent on retail industry grew in demand. This can be attributed to the fact that during that time, a lot of customers got accustomed to online shopping space and the services dependent on them.

Finally, I would like to bring to your attention that out of 17694 users who were a part of multiple transactions in the transactions data, only 130 of them have their personal information recorded in the Users data. This has an adverse impact on the business insights drawn from the Users data, for example age or gender. This is because, many valid transactions made by the users whose details were not present in the Users data would go completely unaccounted for when business insights are drawn. This further affects the decisions that would follow.

I look forward to hearing back your thoughts about the issues I presented. Please feel free to share any information that you feel would help me understand the situation better or aid me in my future analyses.

Best Regards,

Adarsh V.