Hello!

I hope everyone is doing well. I wanted to ask a couple of questions about the dataset that I was given. First, how was the data collected, and who collected the data? I am asking because when I first started to look over the data I found that some values such as how much the customer spent was listed as “”. I also noticed that the dataset did not include any receipt data on users created within the past 6 months. In order to resolve these data quality issues, I wanted to know if it was possible to retrieve this data, recollect the data, or to exclude user data that had values such as “” listed earlier.

Second, is it possible to get more data on the brands scanned by our customers. I noticed that within the brand dataset each brand was only listed once, so it’s impossible to determine the top five brands scanned by customers in the most recent month.

Finally, I wanted to share some concerns I have about managing the data we have so far. As of this moment, each document can be negotiated with in a smaller program like Excel or Google Sheets. However, once we acquire more user data it will be too much for such program. I recommend we start looking at storing our user data in data lakes.

Thank you for your time,  
  
Taron Melvin