### First: explore the data

**Review the unstructured csv files and answer the following questions with code that supports your conclusions:**

* Are there any data quality issues present?

I started off by looking over the three csv’s in Excel before uploading to a SQL Server database. Below are my initial observations:

* PRODUCTS
  + BARCODE has duplicate values although it is the primary key, meaning I will need to dedupe it
  + Every field has at least one NULL value
* TRANSACTION
  + RECEIPT \_ID and BARCODE combined have duplicate values and I assume that is the composite key, which means I will need to dedupe them together
  + BARCODE and FINAL\_SALE both have at least one NULL value
  + FINAL\_QUANTITY has data quality issues. We have some records with a value of “1” while others have a value of “zero”.
* USER
  + BIRTH\_DATE, STATE, LANGUAGE, and GENDER ALL HAVE AT LEAST ONE NULL VALUE
* Are there any fields that are challenging to understand?
* PRODUCTS
  + PRODUCTS was pretty straightforward. I was surprised that we had more PRODUCTS records than TRANSACTION records.
* TRANSACTION
  + RECEIPT\_ID might make someone think that this is “receipt” level. However, this data appears to be RECEIPT\_ID, BARCODE level data (essentially the representation of a receipt, assuming the receipt is grouping the same product into one record with a quantity). This is just an observation I made, not necessarily something challenging to understand.
  + FINAL\_QUANTITY and FINAL\_SALE definitely cause some initial confusion. FINAL\_QUANTITY has data quality issues as it has numbers and strings rather than one uniform field. FINAL\_SALE is also confusing because it has NULL values. It might just be an issue where there is duplicate data or we need to do some sort of aggregation.

I will upload this initial raw data into a SQL Server database to confirm and correct the comments I have mentioned above. I will also look for other instances of data inaccuracy or lack of data integrity.

I made some updates to take care of the data inaccuracies and quality issues. Below they are detailed:

* PRODUCTS
  + Excluded records where BARCODE is NULL or blank
  + Grouped data on BARCODE to remove records that were complete duplicates (multiple records where all the fields were identical)
  + For the remaining duplicates I grouped data on BARCODE and kept the record that had the most non-NULL values. If both records, had the same amount of non-NULL values, the program kept the record higher up in the order of the original dataset (essentially random for our purposes).
* TRANSACTION
  + Excluded records where BARCODE is NULL or blank
  + Converted values of FINAL\_QUANTITY that were “zero” to “0”. Also rounded decimals, assuming there are no partial quantities as the products with partial quantities did not appear to be items that were sold by a weight rather than a unit. Any numbers less than 1 round up, while the rest round normally.
  + There were duplicate records that had all the same descriptive information but differing values for the metrics (FINAL\_QUANTITY and FINAL\_SALE). Therefore, we grouped our data and kept the maximum of each of these fields. Made a choice on MAXIMUM with the idea that if both numbers existed then they were both correct depending on perspective. Perhaps the smaller was a per unit price rather than the total price for that product for FINAL\_SALE.
* USER
  + No duplicates found
  + NULLs were from fields describing the user. I cannot replace them with anything as I would be completely making things up without any logic or reason.

I will be using this processed data for the following questions.

**We recommend using SQL or python and data visualization to examine the data.**

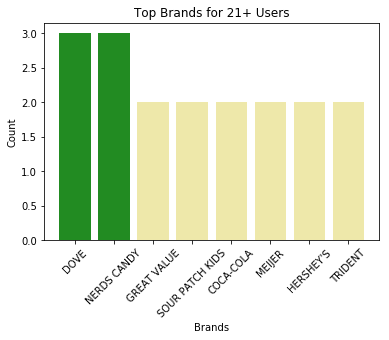
**Second: provide SQL queries**

**Answer three of the following questions with at least one question coming from the closed-ended and one from the open-ended question set. Each question should be answered using one query.**

**Closed-ended questions:**

* What are the top 5 brands by receipts scanned among users 21 and over?

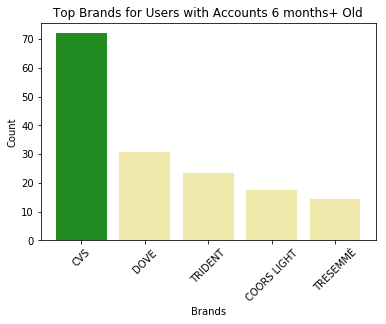
The top 5 brands for users 21 and over are as follows:



Notice that there are actually 8 brands rather than 5. That is because there are several brands that appear on receipts for these users that have the same number of occurrences. DOVE and NERDS CANDY both appear 3 times and are tied in the first place, while the remaining brands each appear twice.

* What are the top 5 brands by sales among users that have had their account for at least six months?

The top 5 brands by sales for users that have had their account for at least six months are as follows:



**Open-ended questions: for these, make assumptions and clearly state them when answering the question.**

* Which is the leading brand in the Dips & Salsa category?

The leading brand in the Dips & Salsa category is TOSTITOS. They rank as the number one brand in the four different metrics we calculated in order to decide the leading brand. Below are the four metrics:

* Total sales
* Total units sold
* Unique number of users
* Unique number of items

Here are the numbers for the top five brands ranked by total amount of sales:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Brand | Sales | Units | Users | Items | Rank: Sales | Rank: Units | Rank: Users | Rank: Items |
| TOSTITOS | 181.3 | 38 | 35 | 12 | 1 | 1 | 1 | 1 |
| GOOD FOODS | 94.91 | 9 | 9 | 5 | 2 | 11 | 10 | 7 |
| PACE | 85.75 | 24 | 24 | 12 | 3 | 2 | 2 | 1 |
| FRITOS | 67.16 | 20 | 19 | 5 | 4 | 3 | 3 | 7 |
| MARKETSIDE | 65.22 | 16 | 16 | 11 | 5 | 5 | 5 | 3 |

If we had to define the leading brand based on one metric it would be the total amount of sales. As previously mentioned, TOSTITOS also has the highest number of units sold, unique users purchasing, and sells the widest variety of products. They have the largest reach in terms of capturing an audience of users.

**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:**

* Key data quality issues and outstanding questions about the data
* One interesting trend in the data
  + Use a finding from part 2 or come up with a new insight
* Request for action: explain what additional help, info, etc. you need to make sense of the data and resolve any outstanding issues

Hello Mr. Boss,

I wanted to provide you with some findings and insights from my recent data investigation.

Before I get to that, I wanted to cover a couple data issues I had to correct. We may need to improve our data infrastructure, happy to take the lead myself, to ensure data integrity and quality for future work. Below are the main issues:

* Duplicate data: There was vast amounts of various duplication going on. We need to create a deduping process that is resilient and trustworthy.
* NULL values: A lot of data was lost somewhere in the process or we never had it to begin with. We may want to look into whether we can improve our data sources.

After correcting the above issues, I was able to dig in and uncover important information. The most significant finding was that TOSTITOS is a huge leader in the Dips & Salsa category. They ranked highest in all of our metrics, although there are some close competitors in PACE and FRITOS. This is a huge opportunity for us to take advantage by investing our time and effort towards the Dips & Salsa category. I am thinking more promotion to help acquire new users who are drawn to this category.

I would love to work on this project to help us increase our user base. I will spend more time looking into our existing data to see if I can draw further conclusions. Let me know if you are able to provide any additional data specific to this category.

Regards,

Alec

­­