

Closed-Ended

- What are the top 5 brands by receipts scanned among users 21 and over?
 - Nerds Candy, Trident, Dove, Meijer, Hershey's
- What are the top 5 brands by sales among users that have had their account for at least six months?
 - Nabisco, Great Value, Oreo, Hidden Valley (Only 4 returned based on number of matched user IDs)
- What is the percentage of sales in the Health & Wellness category by generation?
 - Millennials: 57.36%, Gen X: 26.12%, Baby Boomers: 16.52%

Open-Ended

- Who are Fetch's power users?
 - **Assumptions:** In the dataset there were ~177 matches on IDs between the user and transaction tables, significantly limiting the amount of data that can be used. There likely is a greater quantity of data of data that contains similar information, that could be leveraged for a more in depth analysis. But for this question I am assuming the number of matches are representative of Fetch's user base.
 - **Answer:** Based on the data available, English speaking Women ages 35-70 are the core power users. This also identifies the individual users through their user IDs. I based this off of the number of receipts each user had scanned through Fetch. With more variables (e.g. City, race, income) and user data, there are many opportunities to hone in on who the high-engaging customers are.
- Which is the leading brand in the Dips & Salsa category?
 - **Assumptions:** I am determining the leading brand in the category by looking at which has the greatest dollar sales. This is one way to identify category leaders, but it works best in tandem with other metrics. Such as market share change, customer loyalty, repeat purchase behavior, and fair share of shelf (an index of how a brand is performing based on its size).
 - **Answer:** Tostitos; followed by Pace, Marketside, Fritos, and Helluva Good!

Message to Stakeholders

Hello [Person's Name],

I hope everything's going great! I wanted to check in with an update on the analysis I'd been working on for you, as well as what the next best steps are for us to move ahead.

During the data exploration phase I noticed that there are quite a few issues and inconsistencies. The biggest problems I ran in to was with missing values, duplicate observations, and inconsistent data types within the same columns. I was able to identify these and fix many of them using Python. However, I am curious if there is a way in the future that some of these data obstacles could be eliminated prior to the data exploration phase.

When we spoke last, we discussed using this data to identify Fetch's power users and I've got some results for you! Our core users are mostly 35-70 year-old English-speaking Women- based on how many receipts they've submitted. It also returned a list of each individual power user's ID if you ever want to identify them directly. We could take this a step further with a more robust dataset. With more user data and other identifying variables (race, income, education level, etc.) we can further narrow down who our top users really are.

Looking ahead it would be a huge help if you could guide me to the person or team that oversees retrieving and storing our data. Each of the three files had some shared issues, as well as unique challenges. I would like to get a better understanding of how this data is collected to see if we can have more consistency, less missing values, and overall higher quality data.

Thank you for your assistance and please let me know if you have any questions or would like to jump on a call!

Best,

Andrew Ksiezyk