Eshaan Vora
[redacted] Data Analyst Case Study
11/28/22

Answers to Case Study Questions:
*Note: The SQL queries used to produce these answers are in the "CaseStudy.sql" file*

*The "DataUploader.py" file uploads the data from the "DA_CaseStudy.csv" file into the subsequently created 'DataAnalyst_CaseStudy' Database which exists on the host's local mySQL server. The program cleans the dataset in Python Pandas before uploading the data to the 'Survey' table.*

1. About the data:
   - There are 10 unique survey identifiers. The survey identifiers are not related to the questions being asked but instead represent the date that the survey was taken. (Higher survey ids reflect a more recent survey response)
   - 5,865 distinct users
   - 6,996 distinct surveys
   - 45,895 total rows in dataset
   - Each question in each survey response is recorded as a new row entry in the table
   - Survey Question 1 seems like it is from a drop-down menu option and each response to this question should be cleaned to extract only the relevant information which is the trip type

2. There are 8 distinct retail locations: *Albertsons, Costco, Kroger, Sam's Club, Target, Trader Joe's, Walmart, Whole Foods*

3. Please refer to SQL Queries.

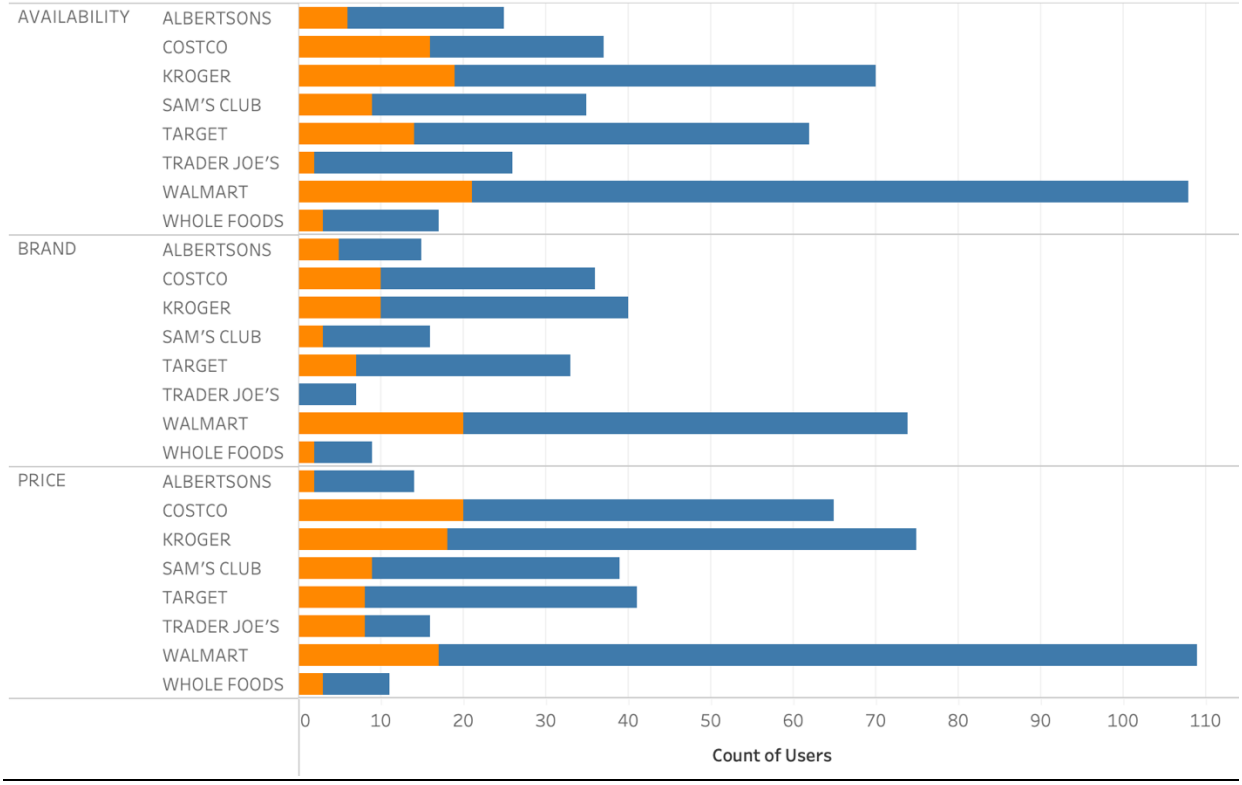
| | WALMART | TARGET | KROGER | COSTCO | SAM'S CLUB | ALBERTSO.. | TRADER JOE'S | WHOLE FOODS |
|---|---|---|---|---|---|---|---|---|
| GRAB & GO | 802 | 398 | 506 | 194 | 170 | 187 | 148 | 167 |
| STOCK-UP | 521 | 261 | 280 | 371 | 221 | 93 | 116 | 78 |
| FILL IN | 582 | 346 | 319 | 203 | 146 | 126 | 114 | 79 |
| OTHER | 132 | 149 | 58 | 47 | 25 | 8 | 6 | 7 |
| RETURNS & SERVICES | 54 | 38 | 12 | 9 | 5 | 3 | 4 | 11 |

4. Please refer to SQL Queries.

| USER_ID | Num_Distinct_Surve... |
|---|---|
| user_3244 | 4 |
| user_2184 | 4 |
| user_2359 | 4 |
| user_2355 | 4 |
| user_533 | 4 |
| user_2077 | 4 |
| user_2900 | 3 |
| user_3100 | 3 |

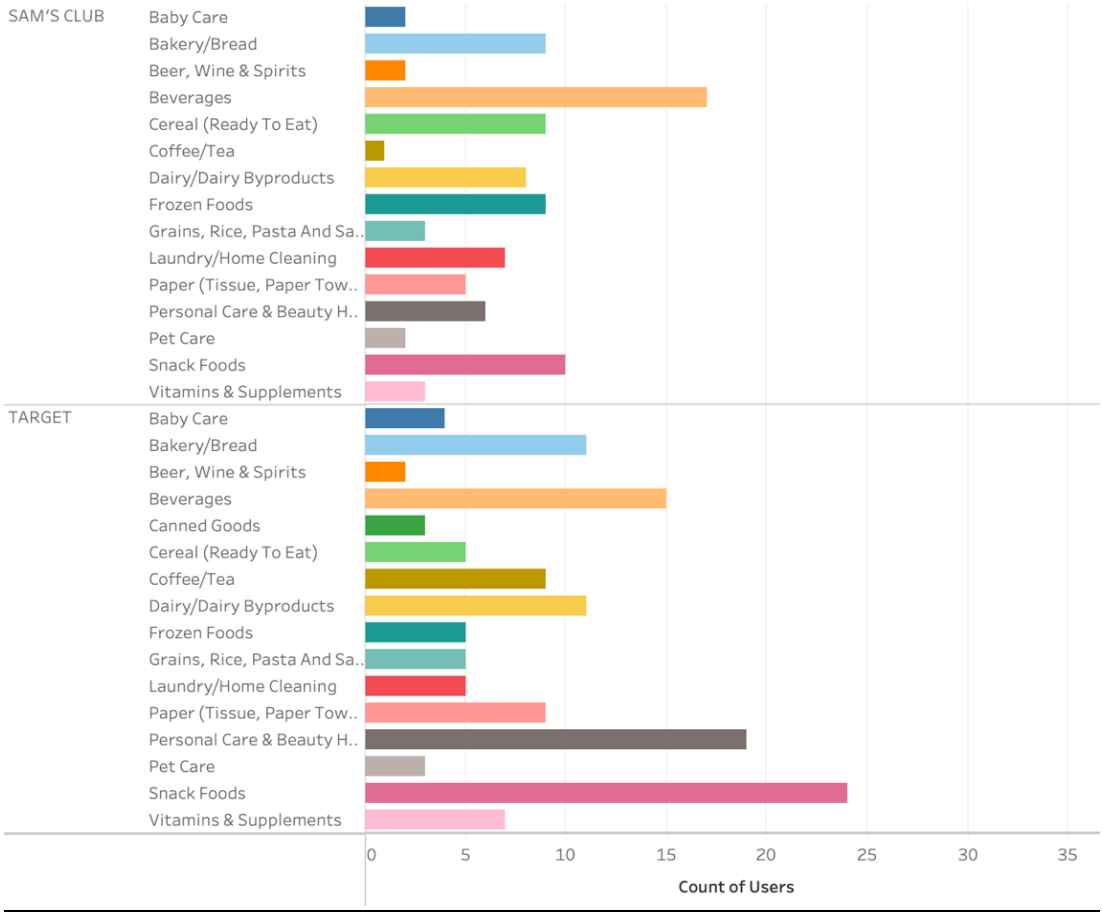
5. Please refer to SQL Queries.

ImportantFactors




**ORANGE** = NON-BUYER
**BLUE** = BUYER

Costco seems to have the highest ratio of non-buyers, indicating there might be more barriers to completing a purchase there.




6. Please refer to SQL Queries.

| Location | Recommendation_Score |
|---|---|
| target | 27700 |
| walmart | 20400 |
| Kroger | 12700 |
| costco | 12000 |
| whole foods | 9200 |
| trader joe's | 9200 |
| Albertsons | 7800 |
| sam's club | 7300 |

Recommendation_Score = (Yes Responses - No Responses) * 100

7. 'Availability' is the most important purchasing factor among the buyers from Walmart.

8. For 'Coffee/Tea' Buyers, Walmart is the most popular location, followed by Target. Additionally, among all 'Coffee/Tea' Buyers, the most important factor driving the purchase was 'Availability'.

9. Walmart was the most preferred location for ‘Grab & Go’ followed by Kroger’s.

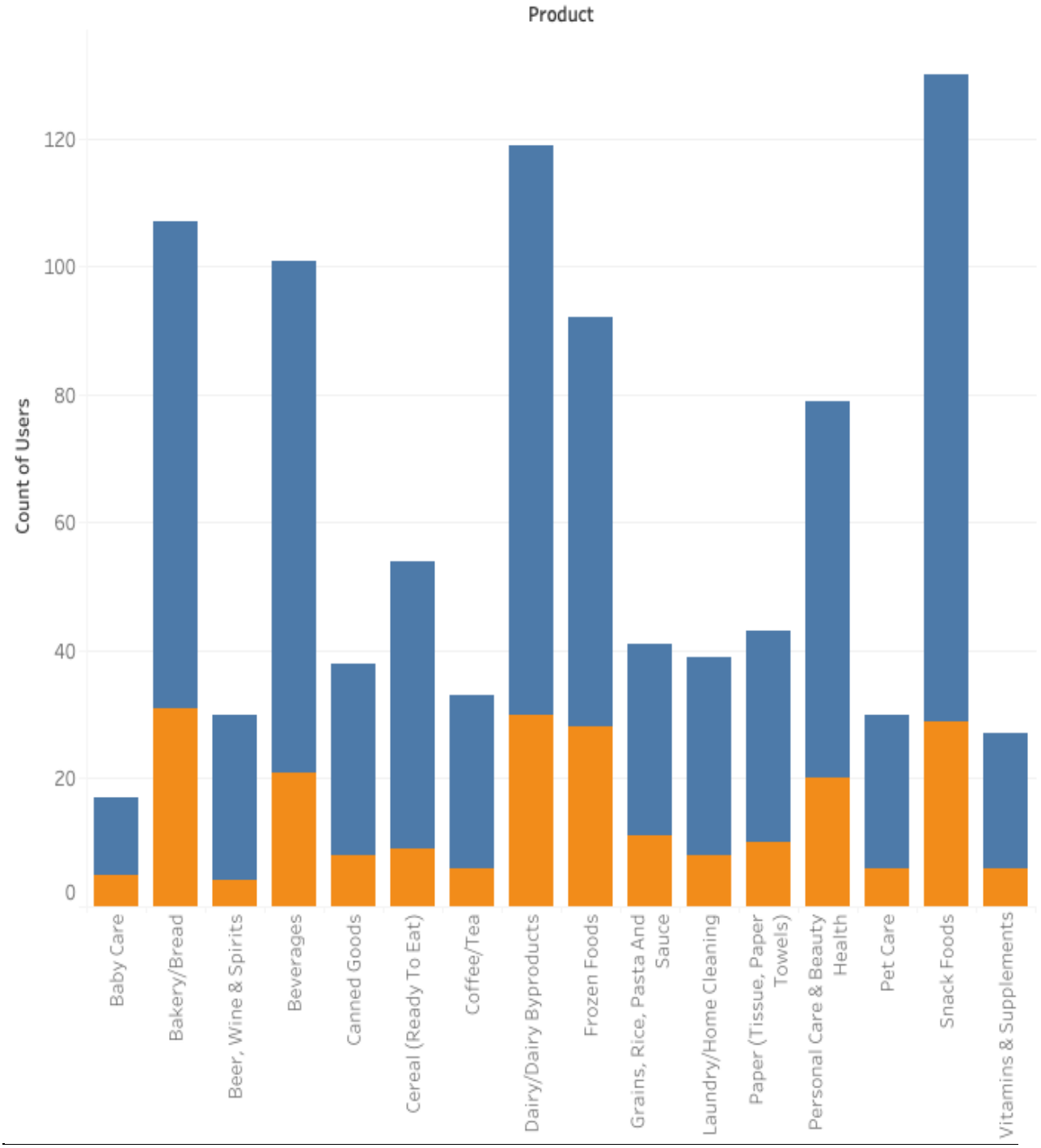
10. I believe a Time-Series plot would work well to study the relationship between the “Number of Views & Subscription Date”. Plotting the subscription date along the X-axis and the number of views on the Y-axis would allow use to see the relationship over time and if we notice a pattern, we can investigate those dates further or run an event study to check if a particular date is an anomaly. We could also aggregate the Number of Views for each day and model the aggregated relationship.

11. The best option for the growth director, keeping in mind the limited budget and retention goals, would be to pay users $.50 incentive per week for sharing survey opinions. Paying users more, at $1 per week, produced the same amount of retention after 3 months of the 7-month study. While we do not know the magnitude of the effect that the incentive paid to users has on the overall budget, we can infer that the $.50 incentive is the better option because the retention rates become equal between the two options despite the difference in price paid.

12. I would complete Task B first, as I would prioritize getting the simpler task out of the way so I could shift my full attention to the longer tasks. I would then work on Task C to give ample time to work on the tasks with ambiguous timeframes. Lastly, I would transition to working on Task A, hopefully having finished or having gained more clarity on Task C. Because Task A requires 2-3 days, I would try to finish the task with 1 day to spare for troubleshooting.

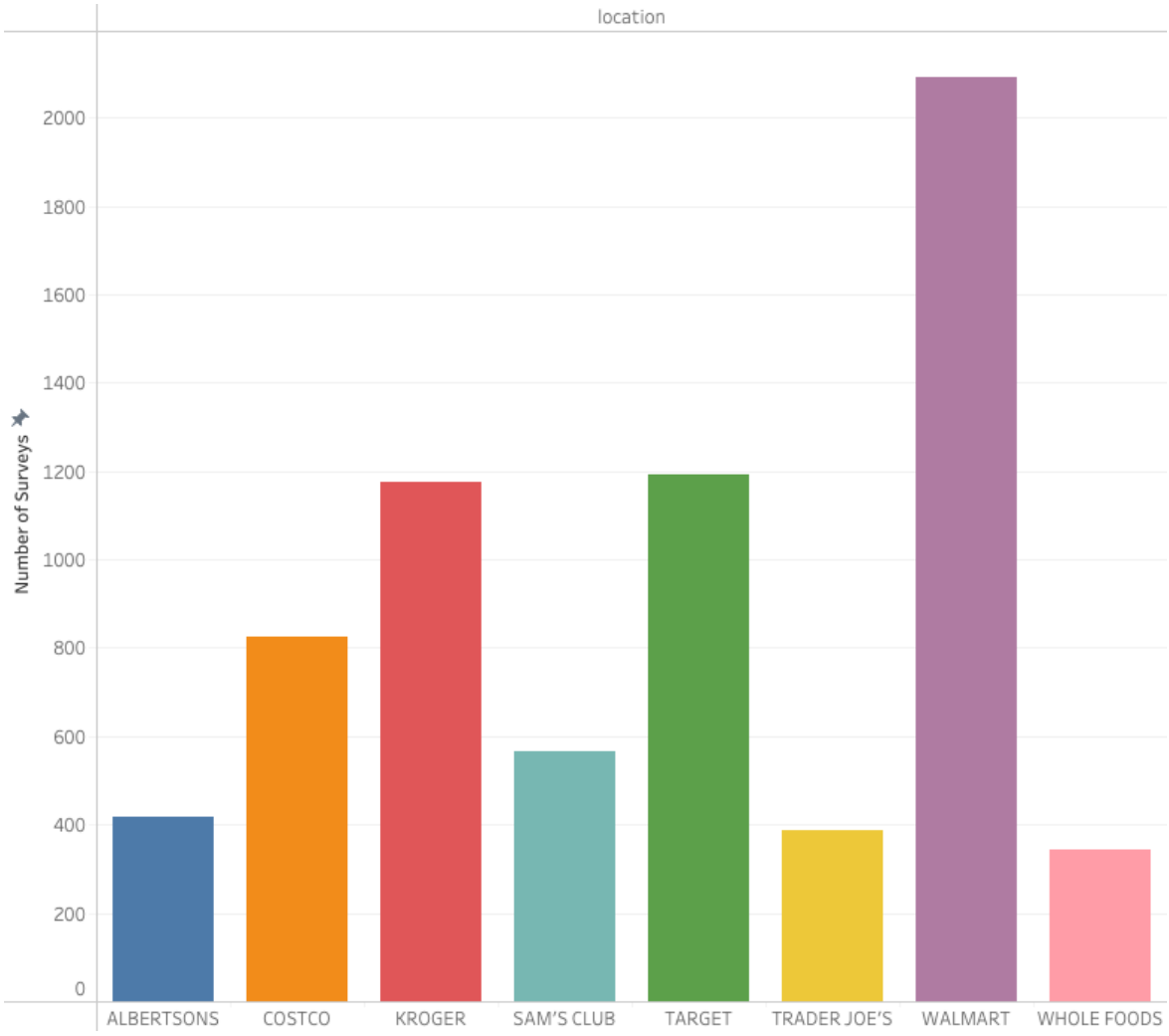
    Alternatively, I could start working on Task A before working on Task C and complete the straightforward tasks first. In this scenario, I would make sure to provide updates and communicate progress with the team, especially because the timeline for the task is already so uncertain. This way, the straightforward tasks would be prioritized and completed before undertaking the possibly lengthier problem-solving tasks. I would still complete Task B first, as I would prefer to get the simpler, but no-less important task out of the way.

Appendix:




**ORANGE** = NON-BUYER

**BLUE** = BUYER


location
Number of Surveys
2000
1800
1600
1400
1200
1000
800
600
400
200
0
ALBERTSONS
COSTCO
KROGER
SAM'S CLUB
TARGET
TRADER JOE'S
WALMART
WHOLE FOODS