Part I: Research Question

A1. Proposal of Question

To understand customers' purchasing habits and provide relevant discounts, can we confidently identify what items are purchased together using Market Basket Analysis?

A2. Defined Goals

As we know, it costs ten times more to acquire new customers than retain old ones and the industry's annual churn rate hovers around twenty-five percent. Though the goal of this analysis is to not predict Churn, our stakeholders have requested an analysis of our customer's purchasing habits to provide relevant discounts to mitigate churn. Therefore, the goal of this Market Basket analysis is to use key performance measurements such as confidence, support, and lift to identify what products are purchased together.

Part II: Market Basket Justification

B1. Explanation of Market Basket

Market Basket Analysis is a data mining technique that uses the Apriori algorithm to uncover purchasing patterns in a retail setting. The general purpose of a Market Basket Analysis is to provide better customer service, improve cross-selling opportunities, and increase direct mail response rates (GeeksforGeeks, 2022b). For our purposes, Market Basket analysis will be used to possibly identify items that are purchased together by our customers. The expected outcome is that we will be able to confidently define which products are generally purchased together by our customers, based on certain key performance measurements.

B2. Transaction Example

An example of a transaction comes directly from the first row of the dataset. In it, we can see a customer who purchased 20 items. The items are listed below:

- Logitech M510 Wireless mouse
- HP 63 Ink

- HP 65 ink
- nonda USB C to USB Adapter
- 10ft iPHone Charger Cable
- HP 902XL ink
- Creative Pebble 2.0 Speakers
- Cleaning Gel Universal Dust Cleaner
- Micro Center 32GB Memory card
- YUNSONG 3pack 6ft Nylon Lightning Cable
- TopMate C5 Laptop Cooler pad
- Apple USB-C Charger cable
- HyperX Cloud Stinger Headset
- TONOR USB Gaming Microphone
- Dust-Off Compressed Gas 2 pack
- 3A USB Type C Cable 3 pack 6FT
- HOVAMP iPhone charger
- SanDisk Ultra 128GB card
- FEEL2NICE 5 pack 10ft Lighning cable
- FEIYOLD Blue light Blocking Glasses

B3. Market Basket Assumption

To perform Market Basket Analysis, we use the apriori algorithm. The Apriori Algorithm assumes that all subsets of a frequent itemset must be frequent (Apriori property). If an itemset is infrequent, all its supersets will be infrequent (GeeksforGeeks, 2022a). A summary of this is we form a strong association between two items based on how often customers purchase those two items together, separately, or at a later date. If they are not often purchased together, separately, or at a later date, then they are not nearly as associated.

Part III: Data Preparation and Analysis

C1. Transforming the Dataset

To prepare our dataset for our market basket analysis, we first dropped rows that had only null values and then filled in null values with empty strings. We then converted the dataframe into a list of lists so that we can use TransactionEncoder to transform the list into an array of True and False. We then took that array and converted it back into a dataframe so that we can perform our Apriori Algorithm and Association Rules on it.

C2. Code Execution

A screenshot of the Apriori execution code has been provided below.

Apriori Algorithm

```
In [19]: #Train data using apriori on the list_df
         frequent_itemsets = apriori(df_model, min_support= 0.03 , use_colnames=True)

In [20]: rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.2)
```

C3. Association Rules Table

A screenshot of the association rules table has been provided below. The values for support, lift, and confidence can be found in each of their respective columns. The table has been sorted by Lift as it is a relative strength indicator that shows the association between two objects (Nandakumar, 2022).

```
In [21]: result = pd.DataFrame(rules)
         result.sort_values(by='lift',inplace=True,ascending=False)
         result
```

Out[21]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
23	(VIVO Dual LCD Monitor Desk mount)	(SanDisk Ultra 64GB card)	0.174110	0.098254	0.039195	0.225115	2.291162	0.022088	1.163716
24	(SanDisk Ultra 64GB card)	(VIVO Dual LCD Monitor Desk mount)	0.098254	0.174110	0.039195	0.398915	2.291162	0.022088	1.373997
12	(SanDisk Ultra 64GB card)	(Dust-Off Compressed Gas 2 pack)	0.098254	0.238368	0.040928	0.416554	1.747522	0.017507	1.305401
11	(Nylon Braided Lightning to USB cable)	(Dust-Off Compressed Gas 2 pack)	0.095321	0.238368	0.035729	0.374825	1.572463	0.013007	1.218270
25	(VIVO Dual LCD Monitor Desk mount)	(Screen Mom Screen Cleaner kit)	0.174110	0.129583	0.035462	0.203675	1.571779	0.012900	1.093043
26	(Screen Mom Screen Cleaner kit)	(VIVO Dual LCD Monitor Desk mount)	0.129583	0.174110	0.035462	0.273663	1.571779	0.012900	1.137061
14	(Screen Mom Screen Cleaner kit)	(Dust-Off Compressed Gas 2 pack)	0.129583	0.238368	0.047994	0.370370	1.553774	0.017105	1.209650
13	(Dust-Off Compressed Gas 2 pack)	(Screen Mom Screen Cleaner kit)	0.238368	0.129583	0.047994	0.201342	1.553774	0.017105	1.089850
18	(Screen Mom Screen Cleaner kit)	(HP 61 ink)	0.129583	0.163845	0.032129	0.247942	1.513276	0.010898	1.111823
15	(Stylus Pen for iPad)	(Dust-Off Compressed Gas 2 pack)	0.095054	0.238368	0.033729	0.354839	1.488616	0.011071	1.180529
17	(Dust-Off Compressed Gas 2 pack)	(VIVO Dual LCD Monitor Desk mount)	0.238368	0.174110	0.059725	0.250559	1.439085	0.018223	1.102008
16	(VIVO Dual LCD Monitor Desk mount)	(Dust-Off Compressed Gas 2 pack)	0.174110	0.238368	0.059725	0.343032	1.439085	0.018223	1.159314
22	(HP 61 ink)	(VIVO Dual LCD Monitor Desk mount)	0.163845	0.174110	0.039195	0.239219	1.373952	0.010668	1.085581
21	(VIVO Dual LCD Monitor Desk mount)	(HP 61 ink)	0.174110	0.163845	0.039195	0.225115	1.373952	0.010668	1.079070
9	(HP 61 ink)	(Dust-Off Compressed Gas 2 pack)	0.163845	0.238368	0.052660	0.321400	1.348332	0.013604	1.122357
10	(Dust-Off Compressed Gas 2 pack)	(HP 61 ink)	0.238368	0.163845	0.052660	0.220917	1.348332	0.013604	1.073256
3	(Screen Mom Screen Cleaner kit)	(Apple Pencil)	0.129583	0.179709	0.030796	0.237654	1.322437	0.007509	1.076009
19	(HP 61 ink)	(USB 2.0 Printer cable)	0.163845	0.170911	0.034395	0.209927	1.228284	0.006393	1.049383
20	(USB 2.0 Printer cable)	(HP 61 ink)	0.170911	0.163845	0.034395	0.201248	1.228284	0.006393	1.046827
0	(Apple Pencil)	(Dust-Off Compressed Gas 2 pack)	0.179709	0.238368	0.050927	0.283383	1.188845	0.008090	1.062815
1	(Dust-Off Compressed Gas 2 pack)	(Apple Pencil)	0.238368	0.179709	0.050927	0.213647	1.188845	0.008090	1.043158
5	(USB 2.0 Printer cable)	(Apple Pencil)	0.170911	0.179709	0.036395	0.212949	1.184961	0.005681	1.042232
4	(Apple Pencil)	(USB 2.0 Printer cable)	0.179709	0.170911	0.036395	0.202522	1.184961	0.005681	1.039640
7	(Apple Pencil)	(VIVO Dual LCD Monitor Desk mount)	0.179709	0.174110	0.036528	0.203264	1.167446	0.005239	1.036592
6	(VIVO Dual LCD Monitor Desk mount)	(Apple Pencil)	0.174110	0.179709	0.036528	0.209801	1.167446	0.005239	1.038081
2	(HP 61 ink)	(Apple Pencil)	0.163845	0.179709	0.033196	0.202604	1.127397	0.003751	1.028711
8	(Apple USB-C Charger cable)	(Dust-Off Compressed Gas 2 pack)	0.132116	0.238368	0.031063	0.235116	0.986357	-0.000430	0.995748

C4. Top Three Rules

A screenshot of the top three rules based on Lift has been provided below.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
23	(VIVO Dual LCD Monitor Desk mount)	(SanDisk Ultra 64GB card)	0.174110	0.098254	0.039195	0.225115	2.291162	0.022088	1.163716
24	(SanDisk Ultra 64GB card)	(VIVO Dual LCD Monitor Desk mount)	0.098254	0.174110	0.039195	0.398915	2.291162	0.022088	1.373997
12	(SanDisk Ultra 64GB card)	(Dust-Off Compressed Gas 2 pack)	0.098254	0.238368	0.040928	0.416554	1.747522	0.017507	1.305401

The top three rules for this Market Basket Analysis are:

1. **Rule 1: Antecedent:** VIVO Dual LCD Monitor Desk mount. **Consequent:** SanDisk Ultra 64GB card

   a. Confidence = 0.22, Support = 0.39, Lift = 2.3
   b. Of all customers who buy a VIVO Dual LCD Monitor Desk mount, 22% also purchased a SanDisk Ultra 64GB. Of all transactions, 3.9% of them contain both items and when customers purchase a VIVO Dual LCD Monitor Desk mount, they are 2.3 times more likely to purchase a SanDisk Ultra 64GB later.
2. **Rule 2: Antecedent:** SanDisk Ultra 64GB card. **Consequent:** VIVO Dual LCD Monitor Desk mount.
   a. Confidence = 0.40, Support = 0.39, Lift = 2.3
   b. Of all customers who buy a SanDisk Ultra 64GB card, 40% also purchased a VIVO Dual LCD Monitor Desk mount. Of all transactions, 3.9% of them contain both items and when customers purchase a SanDisk Ultra 64GB card, they are 2.3 times more likely to purchase a VIVO Dual LCD Monitor Desk later.
3. **Rule 3: Antecedent:** SanDisk Ultra 64GB card. **Consequent:** Dust-Off Compressed Gas 2 pack.
   a. Confidence = 0.41, Support = 0.4, Lift = 1.7
   b. Of all customers who buy a SanDisk Ultra 64GB card, 41% also purchased a Dust-Off Compressed Gas 2 pack. Of all transactions, 4% of them contain both items and when customers purchase a SanDisk Ultra 64GB card, they are 1.7 times more likely to purchase a Dust-Off Compressed Gas 2 pack later.

Part IV: Data Summary and Implications

D1. Significance of Support, Lift, and Confidence Summary

Support tells us what percentage of all transactions will occur together. It is calculated by dividing the frequency by the number of transactions (*Market Basket Analysis [Association Analysis]*, 2022). The significance of our support values for our three rules is around 4% which means they show up in 4% of all total transactions. This support value meets our minimum requirement of 0.03 which indicates that these items are strong and have the potential to be discounted (Kadlaskar, 2022).

The lift indicates the factor by which the probability of customers buying the Consequent product if, the Antecedent product has already been bought. (*Market Basket Analysis [Association Analysis]*, 2022). Lift is the ratio between the target response and the average response (Rodriguez, 2021). The significance of the lift in our three rule is 1.7, 2.1, and 2.1, which indicates that there is some positivity here that when the corresponding Antecedent product is purchased, customers are over 1.7 times more likely to purchase the Consequent product. Therefore, we could then expect customers who purchase one of the items in this rule, will later purchase its consequent item making these rules possibly viable for our discount plan.

Confidence tells us the percentage of customers who purchased the Consequent product out of the total number of customers who purchased an Antecedent product. Confidence is a measure of accuracy and is calculated with the equation below:

$$Confidence(X \rightarrow Y) = \frac{Support(X,Y)}{Support(X)}$$

The significance of our confidence value from our three rules was 22%, 40%, and 41% which though not entirely great, does meet our minimum confidence which indicates that they are strong (Kadlaskar, 2022). An indication of strength though doesn’t indicate a good fit as it would be ideal to have a confidence level closer to 1 or 100%.

D2. Practical Significance of Findings

Though the results of all our findings meet or exceed our minimum requirements, the practical significance of this is that we cannot be confident in the applications of these rules. With a low confidence value hovering at 20-40%, we can't be certain that our discount plan will make any significant impact on improving customer purchases and therefore help mitigate customer churn. Our support did meet the minimum requirement, but it too was low and doesn't give us much information to go by. Our lift was much more positive in that we can assume that once any of these products in these rules are purchased, their consequent product will have a higher likely hood of being purchased. However, as the confidence and support values are low, we cannot make a decision on lift alone. Also, just looking at the rules, we cannot realistically understand why a customer would purchase a desk mount with a SanDisk Ultra 64GB. If the SanDisk Ultra 64GB was purchased with an item that can make use of it such as a camera peripheral or some other computer peripheral, we can then logically conclude that the two items were properly associated.

D3. Course of Action

Though we did find rules that met our minimum requirements, we could not find any practical significance in those rules. Therefore, we could not identify with confidence, which items are purchased together using Market Basket analysis. This may be due to having an insufficient amount of data as we did have a lot of empty rows. When we removed those rows, it essentially cut our dataset in half so we only had half of our dataset to work with. This could mean that there was an error in the data collection and that the error may need to be rectified before proceeding forward. If there is no error, then we may just need more data so that we can get a better understanding of our customers. Since we could not find confidently identify products that are purchased together, our next course of action would be to understand if there was any error in the data collection, and if there is no error, continue data collection and then reiterate over the dataset again.

Part V. Attachments

F. Panopto Recording

G. Web Sources

Market Basket Analysis In Python|How to implement market basket analysis in Python|apriori algorithm. (2020, March 13). YouTube. https://www.youtube.com/watch?v=4QIWJVVWJdQ&t=725s

Market Basket Optimization using Association Rule Mining. (2020, July 21). DataR Labs. https://www.datarlabs.com/post/market-basket-optimisation-using-association-rule-mining

H. Sources

GeeksforGeeks. (2022a, January 13). *Apriori Algorithm*. https://www.geeksforgeeks.org/apriori-algorithm/

GeeksforGeeks. (2022b, March 15). *Market Basket Analysis in Data Mining*. https://www.geeksforgeeks.org/market-basket-analysis-in-data-mining/

Kadlaskar, A. (2022, July 27). *A Comprehensive Guide on Market Basket Analysis*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/

Market Basket Analysis – The Apriori Algorithm. (2019, November 18). The Science of Data. https://scienceofdata.org/2019/10/27/market-basket-analysis-the-apriori-algorithm/

Market Basket Analysis [Association Analysis]. (2022). Datatab. https://datatab.net/tutorial/market-basket-analysis

Nandakumar, S. (2022, May 13). *Why does lift have a bigger role than confidence in Association rules?* Medium. https://blog.devgenius.io/why-lift-has-bigger-role-than-confidence-in-association-rules-619324fc21ab

Rodriguez, R. (2021, June 10). *What Is Lift in Market Basket Analysis?* Tech Business Guide. https://techbusinessguide.com/what-is-lift-in-market-basket-analysis/