|  |  |  |
| --- | --- | --- |
|  |  |  |
| Ramses Meza  June 2021  <https://github.com/lordtable/UTDSGC_C3T4> |  | Data Analytics Course 3.4: Discover Associations between products |

Business Acumen

Blackwell Electronics is currently assessing the option of acquiring Electronidex, a start-up online electronics retailer. In order to better inform any acquisition decision, Blackwell wishes to first gain knowledge on the Electronidex clientele and understand their purchasing patterns, if any, via a market basket analysis. The impact of these insights can go beyond a potential acquisition: they may also drive the future sales strategy and further supplier partnerships.

**Data Management, Cleaning & Pre-processing**

The dataset was provided on a .csv file format and locally stored. The data on the file is un-structured: each row represents a transaction which contains the names of the items purchased, out of a pool of 125 available items. The file contains one (1) month worth of transaction data (9,835 transactions). This data was loaded into RStudio as a transactional dataset. The data required no editing.

**Market Basket Analysis**

Figure 1 depicts a plot showing the top-20 most-frequently purchased items in a decreasing order. For instance, the ***iMac*** is included on ~ 25% of the purchases, making it the most frequently purchased item.

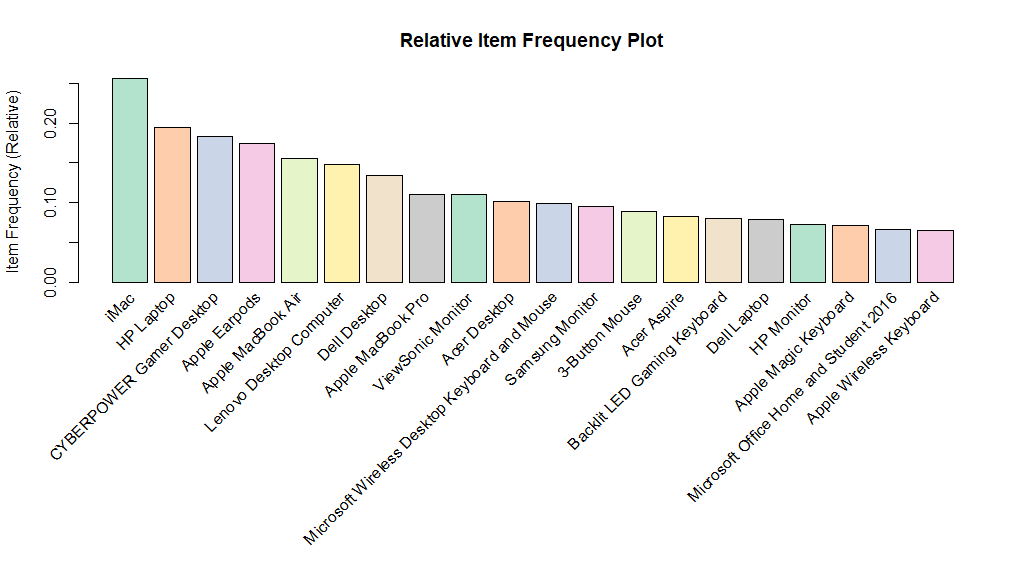


Figure 1: Top-20 most frequently purchased items

Another Apple-manufactured items are also included on this top-20, making it the most purchased brand on the Electronidex catalogue.

In order to perform the market basket analysis, association rules were built using the *arules* library. Sensitivity on the parameters was performed in order to yield enough association rules that can be meaningful. The two (2) main parameters are **support** (how often a rule is applicable to a dataset) and **confidence** (how reliable is a rule). The higher both parameters are, the more we can rely on an association rule, which is yielded as long as it meets minimum confidence and support values as provided by the analyst, trying to reach a balance between obtaining several meaningful rules without being computationally expensive. I used a minimum support=0.01 and minimum confidence=0.2, which means that generated rules are those that cover at least 1% of the transactions and are correct at least 20% of the time. This combination allowed generating 288 rules, with support ranging 1-25% and confidence 20-60%.

The top-10 most frequently purchased items are expected to be part of several rules, as shown on Table 1. Although not completely diagnostic, it indicates that some items (such as ***iMac*** and ***HP Laptop***) are very pervasive, rules-wise (for all rules generated). This may be an indication that these high-count items are not strongly conditioned to the purchased of other particular item(s).

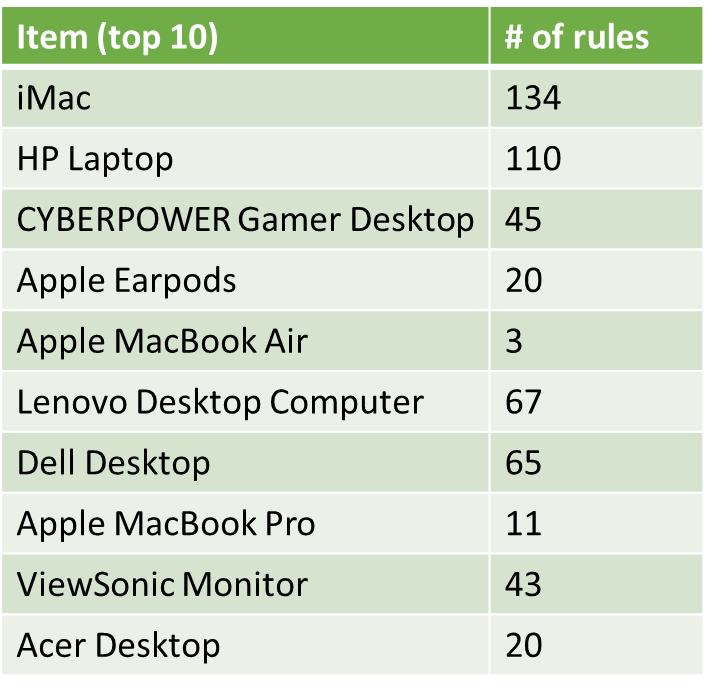


Table 1: Top-10 most frequently-purchased items and # rules they are part of.

We can also visualize all the 288 generated rules by scatter-plotting based on their support, confidence and **lift** (importance of a rule), as shown on Figure 2. It depicts that the vast majority of rules are skewed towards very low support with a more balanced distribution along the confidence axis. Highest lift rules have low support and a distributed range of confidence.

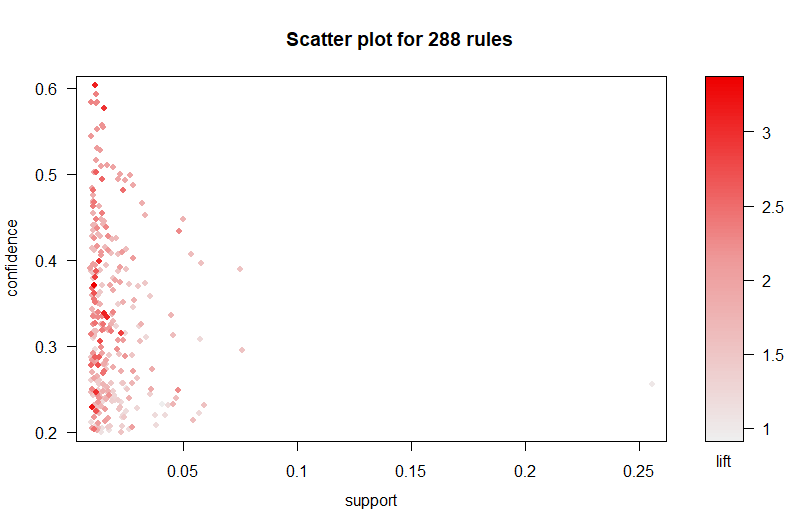


Figure 2: Scatter plot of all the 288 generated association rules

All these observations indicate that a significant number of the generated rules are not meaningful/useful. Therefore, all rules can be sorted based on either confidence, support or lift, as shown on Tables 2,3 and 4, respectively.

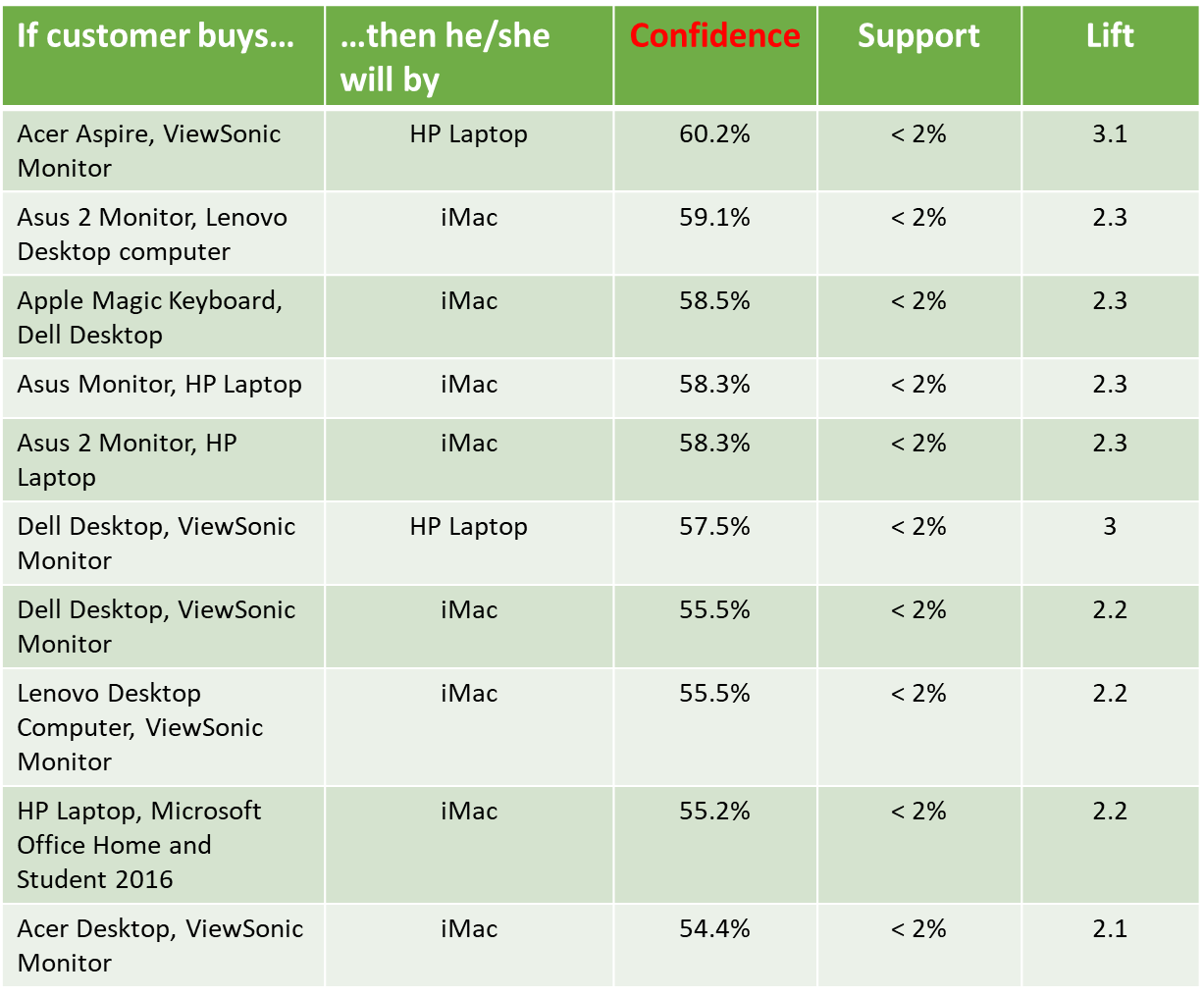


Table 2: Top-10 association rules sorted by Confidence.

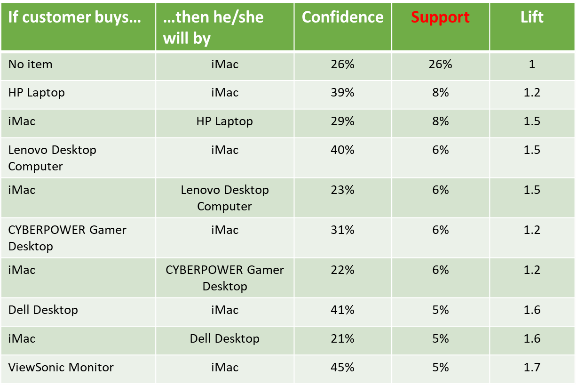


Table 3: Top-10 association rules sorted by Support.

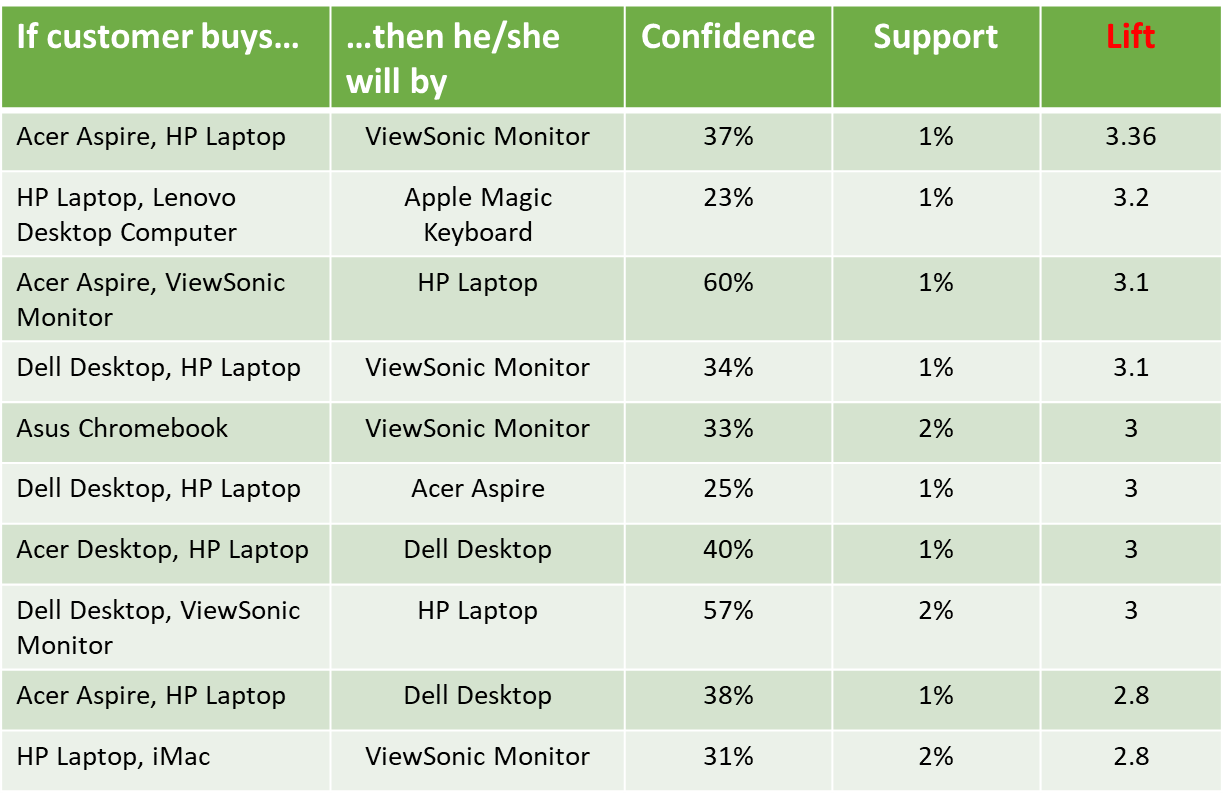


Table 4: Top-10 association rules sorted by Lift.

When sorted by Confidence, the generated rules do not exhibit a very high confidence, and support levels are close to the minimum threshold. Another observation is that the *rhs* of almost all these rules lead to ***iMac***. This sort of ubiquity is confirmed by the rules sorted by support, which also exhibits confidence levels well below 50% and support slightly above the minimum threshold. Contrary to the confidence-based sorting, the support-based sorting *lhs* is composed of single-item itemsets.

Lift-based sorting seems to reach a more “balanced” selection of rules, with much lower pervasiveness of ***iMac***, allowing to obtain rules that describe purchase patterns of more subtle or least-frequently purchased products. This becomes more evident when observing these lift-based sorting rules on graph form on Figure 3, depicting more intricated paths of association.

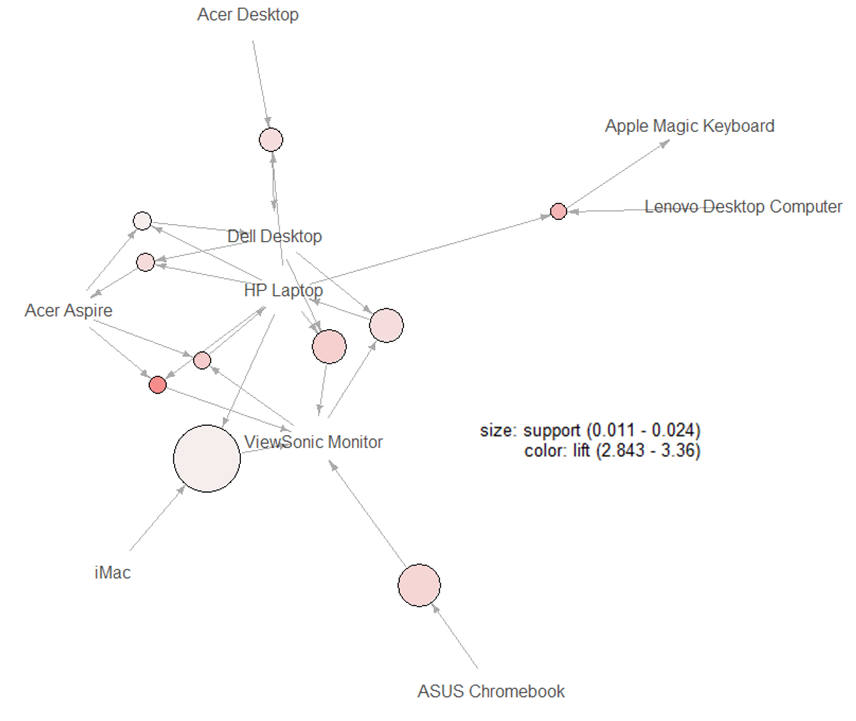


Figure 3: Top-10 association rules sorted by Lift: Graphic form

**Take-aways**

* ***iMac*** is present in the *rhs* of many generated rules instead of a few, globally indicating that customers purchase ***iMac*** with little regard of what else is on their shopping carts.
* ***HP Laptop***, on the other hand, seems to be present of the *rhs* of much fewer (top) rules.
* The remaining of the most frequently purchased products seem to be part of more subtle purchasing patterns, as indicated by the overall lower confidence and support of the respective rules.
* One (1) month worth of transaction data may not be enough to correctly unveil purchasing patterns, so a risk needs to be assigned to this fact during Electronidex acquisition decision-making.
* Blackwell needs to inquiry whether a special marketing-sales strategy exists for ***iMac*** as a stand-alone product, and if it is reproducible to the rest of the catalogue. If yes, then this would represent a strong support for acquisition, if not, then probably Electronidex high (and ubiquitous, basket-wise) sales of ***iMac*** is not due to the company’s proprietary marketing, reducing its attractiveness.