# Western Governors University

## D212 - Data Mining II - Rules and Lift Analysis

### Shane Boyce

### SCENARIO

One of the most critical factors in customer relationship management that directly affects a company's long-term profitability is understanding its customers. When a company can better understand its customer characteristics, it is better able to target products and marketing campaigns for customers, resulting in better profits for the company in the long term.

In [1]:
```python
#import statements
import numpy as np #linear algebra
import pandas as pd #dataframes
import matplotlib.pyplot as plt #visualization
import seaborn as sns #visualization
from mlxtend.preprocessing import TransactionEncoder   #association rules
from mlxtend.frequent_patterns import apriori, association_rules #association rules

# import data from csv
mb_df = pd.read_csv('teleco_market_basket.csv')
```

In [2]:
```python
# view data
mb_df.head()
```

Out[2]:

| | Item01 | Item02 | Item03 | Item04 | Item05 | Item06 | Item07 | Item08 | Item09 | Item10 | Item11 | Item12 | Item13 | Item14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **1** | 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 |
| **2** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **3** | Apple Lightning to Digital AV Adapter | TP-Link AC1750 Smart WiFi Router | Apple Pencil | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **4** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

In [3]:
```python
# view data
mb_df.shape
```

Out[3]:
```
(15002, 20)
```

In [4]:
```python
# view data summary
mb_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15002 entries, 0 to 15001
Data columns (total 20 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Item01  7501 non-null   object
 1   Item02  5747 non-null   object
 2   Item03  4389 non-null   object
 3   Item04  3345 non-null   object
 4   Item05  2529 non-null   object
 5   Item06  1864 non-null   object
 6   Item07  1369 non-null   object
 7   Item08  981 non-null    object
 8   Item09  654 non-null    object
 9   Item10  395 non-null    object
 10  Item11  256 non-null    object
 11  Item12  154 non-null    object
 12  Item13  87 non-null     object
 13  Item14  47 non-null     object
 14  Item15  25 non-null     object
 15  Item16  8 non-null      object
 16  Item17  4 non-null      object
 17  Item18  4 non-null      object
 18  Item19  3 non-null      object
 19  Item20  1 non-null      object
dtypes: object(20)
memory usage: 2.3+ MB
```

In [5]:
```python
mb_df.describe()
```

Out[5]:

| | Item01 | Item02 | Item03 | Item04 | Item05 | Item06 | Item07 | Item08 | Item09 | Item10 | Item11 | Item12 | Item13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 7501 | 5747 | 4389 | 3345 | 2529 | 1864 | 1369 | 981 | 654 | 395 | 256 | 154 | 87 |
| **unique** | 115 | 117 | 115 | 114 | 110 | 106 | 102 | 97 | 88 | 80 | 66 | 50 | 43 |
| **top** | Dust-Off Compressed Gas 2 pack | Dust-Off Compressed Gas 2 pack | Dust-Off Compressed Gas 2 pack | Dust-Off Compressed Gas 2 pack | Apple USB-C Charger cable | USB 2.0 Printer cable | Apple USB-C Charger cable | Apple USB-C Charger cable | Apple USB-C Charger cable | Apple USB-C Charger cable | TopMate C5 Laptop Cooler pad | Apple USB-C Charger cable | Apple USB-C Charger cable |
| **freq** | 577 | 484 | 375 | 201 | 153 | 107 | 96 | 67 | 57 | 31 | 22 | 15 | 8 |

## Part I: Research Question

### A. Describe the purpose of this data mining report by doing the following:

1. Propose one question relevant to a real-world organizational situation that you will answer using market basket analysis.

One of the recommendations made in previous analysis was to sell items in bundled packages to either alleviate cost or to encourage adoption of services or products. This market basket analysis asks: What are the most common bundles of products and services sold to customers? This will assist with recommendation of bundled packages to customers.

1. Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data. The goal of this analysis is to identify if buying an item in the data set is associated with buying another item in the data set. This will allow us to identify the most common bundles of products and services sold to customers.

## Part II: Market Basket Justification

### B. Explain the reasons for using market basket analysis by doing the following:

1. Explain how market basket analyzes the selected dataset. Include expected outcomes.

Market Basket analysis, or apriori rules and lift, provides snsight into how items or events might be grouped together. If you give a moose a muffin, he is going to want a glass of milk to wash it down. This type of analysis let's a business predict with the moose will want (antecedent) after getting his muffin (precedent) and how likely that ask is to happen.

The expected outcome will be to link rules with initial purchase antecedents with antecedent/consequent purchase. From here, we can measure the support of each item and the group of items (the rate at which is it purchased compared to the entire dataset). After determing the support, a confiedence ration is given to the items or collection of items. A lift score ios given which is the ration of transactions pertaining to the rule. Low lift means the items are independent an not bought together (IE coloring books and car oil) whereas high lift means they are related packaged items (IE Italian food and breadsticks).

1. Provide one example of transactions in the dataset.

One transaction the dataset provides is precedent {buying an apple pencil} with an antecedent {Buying a 2 pack of compressed air duster}

1. Summarize one assumption of market basket analysis.

Market Basket Analysis assumes that there must be a measurable and meaningful frequency to items to make meaningful conclusion scores. For example, your restaurant might sell breadsticks with spaghetti every time but if those are not popular menu items market basket will fail to pick up the connection.

## Part III: Data Preparation and Analysis

### C. Prepare and perform market basket analysis by doing the following:

1. Transform the dataset to make it suitable for market basket analysis. Include a copy of the cleaned dataset.

In [6]:
```python
#flatten transaction data
transactions = mb_df.stack().groupby(level=0).apply(list).tolist()
```

In [7]:
```python
#instantiate encoder
te = TransactionEncoder()
te.fit(transactions)
```

Out[7]:
```
TransactionEncoder()
```

In [8]:
```python
#onehote encode data
onehot = te.fit_transform(transactions)
```

In [9]:
```python
#convert to dataframe
market_basket = pd.DataFrame(onehot, columns=te.columns_)
```

In [10]:
```python
#view data
market_basket.head()
```

Out[10]:

| | 10ft iPHone Charger Cable | 10ft iPHone Charger Cable 2 Pack | 3 pack Nylon Braided Lightning Cable | 3A USB Type C Cable 3 pack 6FT | 5pack Nylon Braided USB C cables | ARRIS SURFboard SB8200 Cable Modem | Anker 2-in-1 USB Card Reader | Anker 4-port USB hub | Anker USB C to HDMI Adapter | Apple Lightning to Digital AV Adapter | ... | hP 65 Tri-color ink | iFixit Pro Tech Toolkit | iPhone 11 case | iPhone 12 Charger cable | iPhone 12 Pro case |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | True | False | False | True | False | False | False | False | False | False | ... | False | False | False | False | False |
| **1** | False | False | False | False | False | False | False | False | False | True | ... | False | False | False | False | False |
| **2** | False | False | False | False | False | False | False | False | False | False | ... | False | False | False | False | False |
| **3** | False | False | False | False | False | False | False | False | False | False | ... | False | False | False | False | False |
| **4** | False | False | False | False | False | False | False | False | False | False | ... | False | False | False | False | False |

5 rows × 119 columns

In [11]:
```python
#to csv
market_basket.to_csv('market_basket.csv', index=False)
```

1. Execute the code used to generate association rules with the Apriori algorithm. Provide screenshots that demonstrate the error-free functionality of the code.

In [12]:
```python
# association rules

frequent_itemsets = apriori(market_basket, min_support=0.01, use_colnames=True)
# sort by support
frequent_itemsets.sort_values(by='support', ascending=False, inplace=True)
```

In [13]:
```python
# top 10 frequent itemsets
frequent_itemsets.head(10)
```

Out[13]:

| | support | itemsets |
|---|---|---|
| **18** | 0.238368 | (Dust-Off Compressed Gas 2 pack) |
| **8** | 0.179709 | (Apple Pencil) |
| **66** | 0.174110 | (VIVO Dual LCD Monitor Desk mount) |
| **63** | 0.170911 | (USB 2.0 Printer cable) |
| **23** | 0.163845 | (HP 61 ink) |
| **9** | 0.132116 | (Apple USB-C Charger cable) |
| **58** | 0.129583 | (Screen Mom Screen Cleaner kit) |
| **57** | 0.098254 | (SanDisk Ultra 64GB card) |
| **42** | 0.095321 | (Nylon Braided Lightning to USB cable) |
| **59** | 0.095054 | (Stylus Pen for iPad) |

In [14]:
```python
# to csv
frequent_itemsets.to_csv('frequent_itemsets.csv', index=False)
```

1. Provide values for the support, lift, and confidence of the association rules table.

In [15]:
```python
# association rules
ruleset = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
# sort by support
ruleset = ruleset.sort_values('support', ascending=False)
```

In [16]:
```python
ruleset.sample(20)
```

Out[16]:

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|---|---|---|---|---|---|---|---|---|---|
| **334** | (Nylon Braided Lightning to USB cable, Screen ... | (Dust-Off Compressed Gas 2 pack) | 0.023597 | 0.238368 | 0.011065 | 0.468927 | 1.967236 | 0.005440 | 1.434136 |
| **366** | (Stylus Pen for iPad) | (Logitech M510 Wireless mouse) | 0.095054 | 0.071457 | 0.010532 | 0.110799 | 1.550572 | 0.003740 | 1.044245 |
| **314** | (FEIYOLD Blue light Blocking Glasses) | (Nylon Braided Lightning to USB cable) | 0.065858 | 0.095321 | 0.011332 | 0.172065 | 1.805116 | 0.005054 | 1.092693 |
| **229** | (Dust-Off Compressed Gas 2 pack) | (HP 61 ink, Screen Mom Screen Cleaner kit) | 0.238368 | 0.032129 | 0.013998 | 0.058725 | 1.827780 | 0.006340 | 1.028255 |
| **405** | (10ft iPHone Charger Cable 2 Pack) | (HP 61 ink) | 0.050527 | 0.163845 | 0.010132 | 0.200528 | 1.223888 | 0.001853 | 1.045884 |
| **177** | (Dust-Off Compressed Gas 2 pack) | (VIVO Dual LCD Monitor Desk mount, Screen Mom ... | 0.238368 | 0.035462 | 0.015731 | 0.065996 | 1.861024 | 0.007278 | 1.032691 |
| **31** | (Screen Mom Screen Cleaner kit) | (Apple Pencil) | 0.129583 | 0.179709 | 0.030796 | 0.237654 | 1.322437 | 0.007509 | 1.076009 |
| **303** | (Stylus Pen for iPad) | (VIVO Dual LCD Monitor Desk mount, Dust-Off Co... | 0.095054 | 0.059725 | 0.011465 | 0.120617 | 2.019529 | 0.005788 | 1.069244 |
| **218** | (Apple Pencil) | (Logitech M510 Wireless mouse) | 0.179709 | 0.071457 | 0.014131 | 0.078635 | 1.100450 | 0.001290 | 1.007790 |
| **212** | (10ft iPHone Charger Cable 2 Pack) | (VIVO Dual LCD Monitor Desk mount) | 0.050527 | 0.174110 | 0.014265 | 0.282322 | 1.621513 | 0.005468 | 1.150780 |
| **269** | (USB Type C to USB-A Charger cable) | (Apple USB-C Charger cable) | 0.080389 | 0.132116 | 0.011998 | 0.149254 | 1.129720 | 0.001378 | 1.020145 |
| **296** | (HP 65 ink) | (Dust-Off Compressed Gas 2 pack) | 0.033329 | 0.238368 | 0.011598 | 0.348000 | 1.459926 | 0.003654 | 1.168147 |
| **215** | (Syntech USB C to USB Adapter) | (Apple USB-C Charger cable) | 0.081056 | 0.132116 | 0.014131 | 0.174342 | 1.319617 | 0.003423 | 1.051143 |
| **114** | (Screen Mom Screen Cleaner kit) | (Logitech M510 Wireless mouse) | 0.129583 | 0.071457 | 0.017598 | 0.135802 | 1.900474 | 0.008338 | 1.074457 |
| **351** | (HP 61 ink) | (VIVO Dual LCD Monitor Desk mount, Screen Mom ... | 0.163845 | 0.035462 | 0.010932 | 0.066721 | 1.881480 | 0.005122 | 1.033494 |
| **384** | (VIVO Dual LCD Monitor Desk mount, Dust-Off Co... | (FEIYOLD Blue light Blocking Glasses) | 0.059725 | 0.065858 | 0.010265 | 0.171875 | 2.609786 | 0.006332 | 1.128021 |
| **107** | (Syntech USB C to USB Adapter) | (VIVO Dual LCD Monitor Desk mount) | 0.081056 | 0.174110 | 0.018131 | 0.223684 | 1.284728 | 0.004018 | 1.063858 |
| **309** | (Premium Nylon USB Cable) | (Screen Mom Screen Cleaner kit) | 0.051060 | 0.129583 | 0.011465 | 0.224543 | 1.732817 | 0.004849 | 1.122457 |
| **338** | (Screen Mom Screen Cleaner kit) | (Nylon Braided Lightning to USB cable, Dust-Of... | 0.129583 | 0.035729 | 0.011065 | 0.085391 | 2.389991 | 0.006435 | 1.054299 |
| **254** | (TopMate C5 Laptop Cooler pad) | (Screen Mom Screen Cleaner kit) | 0.076523 | 0.129583 | 0.013198 | 0.172474 | 1.330994 | 0.003282 | 1.051831 |

1. Identify the top three rules generated by the Apriori algorithm. Include a screenshot of the top rules along with their summaries.

In [17]:
```python
# note that I pulled 6 here because each rule has two rows, one for the antecedent and one for the consequent
ruleset.head(6)
```

Out[17]:

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|---|---|---|---|---|---|---|---|---|---|
| **0** | (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 |
| **1** | (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 |
| **2** | (HP 61 ink) | (Dust-Off Compressed Gas 2 pack) | 0.163845 | 0.238368 | 0.052660 | 0.321400 | 1.348332 | 0.013604 | 1.122357 |
| **3** | (Dust-Off Compressed Gas 2 pack) | (HP 61 ink) | 0.238368 | 0.163845 | 0.052660 | 0.220917 | 1.348332 | 0.013604 | 1.073256 |
| **4** | (Apple Pencil) | (Dust-Off Compressed Gas 2 pack) | 0.179709 | 0.238368 | 0.050927 | 0.283383 | 1.188845 | 0.008090 | 1.062815 |
| **5** | (Dust-Off Compressed Gas 2 pack) | (Apple Pencil) | 0.238368 | 0.179709 | 0.050927 | 0.213647 | 1.188845 | 0.008090 | 1.043158 |

In [18]:
```python
# to csv
ruleset.to_csv('ruleset.csv', index=False)
```

## Part IV: Data Summary and Implications

### D. Summarize your data analysis by doing the following:

1. Summarize the significance of support, lift, and confidence from the results of the analysis.

The support of a rule is the proportion of transactions in the dataset that contain the antecedent and consequent. The confidence of a rule is the proportion of transactions in the dataset that contain the antecedent and consequent divided by the proportion of transactions in the dataset that contain the antecedent. The lift of a rule is the confidence of a rule divided by the proportion of transactions in the dataset that contain the consequent. The lift of a rule is a measure of the strength of the association between the antecedent and consequent. A lift of 1 indicates that the antecedent and consequent are independent. A lift greater than 1 indicates that the antecedent and consequent are positively associated. A lift less than 1 indicates that the antecedent and consequent are negatively associated.

In the top three rules in this dataset indicates ( `Dust-Off Compressed Gas 2 pack` & `VIVO Dual LCD Monitor Desk mount` ), ( `Dust-Off Compressed Gas 2 pack` & `HP 61 ink` ), ( `Dust-Off Compressed Gas 2 pack` & `Apple Pencil` ) have a higher chance of being bought together. This may indicate that `Dust-Off Compressed Gas 2 pack` is generally sold with many items. Each of these transactions have a support of 0.05:0.059 meaning these top 3 rules compromise about 5% of all purchases. Where these 6 rules differ is the confidence. Since each of these items has a different rate the likelihood of x being purchased is different than y even if with a support score. the confidence from these 6 rules shows a range of 0.21:0.34 with higher confidence scores being given to rules where `Dust-Off Compressed Gas 2 pack` is the antecedent item.

1. Discuss the practical significance of the findings from the analysis.

This summary of the top rules shows that people come in for a specific item such as a monitor and end up purchasing dust off compressed air. Likely, this is from dust off air being near the check out aisle and a last minute purchase that might not have a real association with the purchased item. The dataset should be cleaned of door busters in order to build accuracy ffff

1. Recommend a course of action for the real-world organizational situation from part A1 based on your results from part D1.

## Sources

No external sources used for this analysis