Task 3: Association and Lift Analysis

Data Preparation

In [1]:
```
# import libraries and packages

import pandas as pd
import numpy as np
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules
```

In [2]:
```
# load data set

df = pd.read_csv('teleco_market_basket.csv')
```

In [3]:
```
df.shape
```

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

In [4]:
```
# inspect first 5 rows

df.head()
```

Out[4]:

	Item01	Item02	Item03	Item04	Item05	Item06	Item07	Item08	Item09	Item10	Item11	Item12	Item13	Item1
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
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	TONC US Gamin Micropho
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
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	Na
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na

In [5]:
```
# look for null values
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 [6]:
```
# Drop records with all missing values
df.dropna(how = 'all', inplace = True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7501 entries, 1 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: 1.2+ MB
```

In [7]:
```
df.describe()
```

Out[7]:

	Item01	Item02	Item03	Item04	Item05	Item06	Item07	Item08	Item09	Item10	Item11	Item12
count	7501	5747	4389	3345	2529	1864	1369	981	654	395	256	154
unique	115	117	115	114	110	106	102	97	88	80	66	50
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
freq	577	484	375	201	153	107	96	67	57	31	22	15

In [8]:
```
# replace NaN with 0
df.fillna(0, inplace = True)
df.head()
```

Out[8]:

	Item01	Item02	Item03	Item04	Item05	Item06	Item07	Item08	Item09	Item10	Item11	Item12	Item13
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
3	Apple Lightning to Digital AV Adapter	TP-Link AC1750 Smart WiFi Router	Apple Pencil	0	0	0	0	0	0	0	0	0	0
5	UNEN Mfi Certified 5-pack Lightning Cable	0	0	0	0	0	0	0	0	0	0	0	0
7	Cat8 Ethernet Cable	HP 65 ink	0	0	0	0	0	0	0	0	0	0	0
9	Dust-Off Compressed Gas 2 pack	Screen Mom Screen Cleaner kit	Moread HDMI to VGA Adapter	HP 62XL Tri-Color ink	Apple USB-C Charger cable	0	0	0	0	0	0	0	0

In [9]:
```
# find new shape after changes
df.shape
```

Out[9]:
```
(7501, 20)
```

Code from http://rasbt.github.io/mlxtend/user_guide/preprocessing/TransactionEncoder/

In [10]:
```
# convert datset into list form
telelist = []
for i in range(0, 7501):
    telelist.append([str(df.values[i, j]) for j in range(0, 20)])
```

In [11]:
```
# convert lists into one-hot encoded array
# convert boolean to int for memory efficiency
encoder = TransactionEncoder()
encoded_array = encoder.fit(telelist).transform(telelist)
encoded_array.astype('int')
```

Out[11]:
```
array([[0, 1, 0, ..., 0, 1, 0],
       [1, 0, 0, ..., 0, 0, 0],
       [1, 0, 0, ..., 0, 0, 0],
       ...,
       [1, 0, 0, ..., 0, 0, 0],
       [1, 0, 0, ..., 0, 0, 0],
       [1, 0, 0, ..., 0, 0, 0]])
```

In [12]:
```
# check list form is correct
telelist[:1]
```

Out[12]:
```
[['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']]
```

In [13]:
```
# turn array into DataFrame
tele_df = pd.DataFrame(encoded_array, columns = encoder.columns_)
tele_df
```

Out[13]:

	0	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	...	hP 65 Tri-color ink	iFixit Pro Tech Toolkit	iPhone 11 case	iPhone 12 Charger cable
0	False	True	False	False	True	False	False	False	False	False	...	False	False	False	False
1	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False
2	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False
3	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False
4	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7496	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False
7497	True	False	False	False	False	False	True	False	False	False	...	False	False	False	False
7498	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False
7499	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False
7500	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False

7501 rows × 120 columns

In [14]:
```
# remove unneeded column
tele_df = tele_df.drop(['0'], axis = 1)
```

In [15]:
```
df.to_csv('association_rules_prepared_data.csv')
```

Data Analysis

In [16]:
```
# run apriori rule on data set
frequent_items = apriori(tele_df, min_support = 0.01, max_len = 3, use_colnames= True)
frequent_items.head()
```

Out[16]:

	support	itemsets
0	0.050527	(10ft iPhone Charger Cable 2 Pack)
1	0.042528	(3A USB Type C Cable 3 pack 6FT)
2	0.019064	(5pack Nylon Braided USB C cables)
3	0.010932	(ARRIS SURFboard SB8200 Cable Modem)
4	0.029463	(Anker 2-in-1 USB Card Reader)

In [17]:
```
# print number of rules created
association_results = list(frequent_items)
print(len(frequent_items))
```

```
257
```

In [18]:
```
# sort frequent items by support value
frequent_items.sort_values(by = 'support', ascending = False, inplace = True)
frequent_items.head()
```

Out[18]:

	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)

In [19]:
```
# generate rules for lift in descending order
lift_rules = association_rules(frequent_items, metric = 'lift', min_threshold = 1)
lift_rules.sort_values(by = 'lift', ascending = False, inplace = True)
lift_rules
```

Out[19]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
162	(SanDisk 128GB Ultra microSDXC card)	(SanDisk Ultra 64GB card)	0.049460	0.098254	0.015998	0.323450	3.291994	0.011138	1.332860
163	(SanDisk Ultra 64GB card)	(SanDisk 128GB Ultra microSDXC card)	0.098254	0.049460	0.015998	0.162822	3.291994	0.011138	1.135410
131	(VIVO Dual LCD Monitor Desk mount, Dust-Off Co...	(SanDisk Ultra 64GB card)	0.059725	0.098254	0.017064	0.285714	2.907928	0.011196	1.262445
134	(SanDisk Ultra 64GB card)	(VIVO Dual LCD Monitor Desk mount, Dust-Off Co...	0.098254	0.059725	0.017064	0.173677	2.907928	0.011196	1.137902
387	(FEIYOLD Blue light Blocking Glasses)	(VIVO Dual LCD Monitor Desk mount, Dust-Off Co...	0.065858	0.059725	0.010265	0.155870	2.609786	0.006332	1.113898
...	...	...	...	...	...	...	...	...	...
252	(TopMate C5 Laptop Cooler pad)	(USB 2.0 Printer cable)	0.076523	0.170911	0.013332	0.174216	1.019340	0.000253	1.004003
266	(Apple Pencil)	(FEIYOLD Blue light Blocking Glasses)	0.179709	0.065858	0.011998	0.066766	1.013783	0.000163	1.000973
267	(FEIYOLD Blue light Blocking Glasses)	(Apple Pencil)	0.065858	0.179709	0.011998	0.182186	1.013783	0.000163	1.003029
231	(HP 63XL Ink)	(VIVO Dual LCD Monitor Desk mount)	0.079323	0.174110	0.013998	0.176471	1.013557	0.000187	1.002866
230	(VIVO Dual LCD Monitor Desk mount)	(HP 63XL Ink)	0.174110	0.079323	0.013998	0.080398	1.013557	0.000187	1.001169

406 rows × 9 columns

In [20]:
```
# generate confidence rules in descending order
confidence_rules = association_rules(frequent_items, metric = 'confidence', min_threshold = 0.1)
confidence_rules.sort_values(by = 'confidence', ascending = False, inplace = True)
confidence_rules
```

Out[20]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
310	(SanDisk Ultra 64GB card, Apple Pencil)	(Dust-Off Compressed Gas 2 pack)	0.019997	0.238368	0.010132	0.506667	2.125563	0.005365	1.543848
264	(SanDisk Ultra 64GB card, Screen Mom Screen Cl...	(Dust-Off Compressed Gas 2 pack)	0.021997	0.238368	0.011065	0.503030	2.110308	0.005822	1.532552
273	(HP 61 ink, SanDisk Ultra 64GB card)	(Dust-Off Compressed Gas 2 pack)	0.023064	0.238368	0.010932	0.473988	1.988472	0.005434	1.447937
271	(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
67	(10ft iPhone Charger Cable 2 Pack)	(Dust-Off Compressed Gas 2 pack)	0.050527	0.238368	0.023064	0.456464	1.914955	0.011020	1.401255
...	...	...	...	...	...	...	...	...	...
56	(Dust-Off Compressed Gas 2 pack)	(Apple Lightning to Digital AV Adapter)	0.238368	0.087188	0.024397	0.102349	1.173883	0.003614	1.016889
214	(Screen Mom Screen Cleaner kit)	(TopMate C5 Laptop Cooler pad)	0.129583	0.076523	0.013198	0.101852	1.330994	0.003282	1.028201
218	(Screen Mom Screen Cleaner kit)	(Apple Pencil, Dust-Off Compressed Gas 2 pack)	0.129583	0.050927	0.013065	0.100823	1.979774	0.006466	1.055491
59	(Dust-Off Compressed Gas 2 pack)	(TopMate C5 Laptop Cooler pad)	0.238368	0.076523	0.023997	0.100671	1.315565	0.005756	1.026851
145	(HP 61 ink)	(FEIYOLD Blue light Blocking Glasses)	0.163845	0.065858	0.016398	0.100081	1.519657	0.005607	1.038030

320 rows × 9 columns

In [21]:
```
# find top three rules
rules = association_rules(frequent_items, metric = 'lift', min_threshold = 1)
top_three = rules[(rules['support'] > 0.003) & (rules['confidence'] > 0.4) & (rules['lift'] > 1.00)]
top_three.head(3)
```

Out[21]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
8	(SanDisk Ultra 64GB card)	(Dust-Off Compressed Gas 2 pack)	0.098254	0.238368	0.040928	0.416554	1.747522	0.017507	1.305401
39	(FEIYOLD Blue light Blocking Glasses)	(Dust-Off Compressed Gas 2 pack)	0.065858	0.238368	0.027596	0.419028	1.757904	0.011898	1.310962
62	(10ft iPhone Charger Cable 2 Pack)	(Dust-Off Compressed Gas 2 pack)	0.050527	0.238368	0.023064	0.456464	1.914955	0.011020	1.401255

In [ ]:

In [ ]: