Data Mining 1 Project 4

Ryan Canfield

2022-12-01

### Project 4

### In this project, you will build association rules from a dataset of your choice using the a priori algorithm and the FP-algorithm. You will identify three association rules generated by the algorithms, and make recommendations to your client based on these rules.

## To complete this project:

#Step 1 additon

2 +2

## [1] 4

# Step 2

# Complete homework assignment 9 due Tuesday November 29, which includes these steps:

# 1. Choose a dataset. Some recommended datasets are provided below. However, you may choose any dataset that addresses the goals of the project, as long as the data can be converted into a transaction object in R.

# Reading in the data found on kaggle https://www.kaggle.com/datasets/knightbearr/sales-product-data?resource=download  
# Chose the month of December to look at sales because of holidays and potential high sales.  
## Pre-Processing   
# Out of the six original columns: Order ID, Product, Quantity Ordered, Price Each, Order Date, Purchase Address, only used Order.ID and Product to convert to a transaction object.  
# To do this I created a new .csv with just the desired columns because when I deleted the undesired columns from a copy of the .csv it was giving weird errors.  
  
December\_Sales <- read.csv("C:/Users/RRC/OneDrive/Documents/CLASSES/YEAR 3/Data Mining 1/Data Sets/Sales\_by\_Month/Sales\_December\_2019.csv")  
head(December\_Sales) # Previews the data and make sure pre-processing worked.

## Order.ID Product Quantity.Ordered Price.Each Order.Date  
## 1 295665 Macbook Pro Laptop 1 1700 12/30/19 00:01  
## 2 295666 LG Washing Machine 1 600.0 12/29/19 07:03  
## 3 295667 USB-C Charging Cable 1 11.95 12/12/19 18:21  
## 4 295668 27in FHD Monitor 1 149.99 12/22/19 15:13  
## 5 295669 USB-C Charging Cable 1 11.95 12/18/19 12:38  
## 6 295670 AA Batteries (4-pack) 1 3.84 12/31/19 22:58  
## Purchase.Address  
## 1 136 Church St, New York City, NY 10001  
## 2 562 2nd St, New York City, NY 10001  
## 3 277 Main St, New York City, NY 10001  
## 4 410 6th St, San Francisco, CA 94016  
## 5 43 Hill St, Atlanta, GA 30301  
## 6 200 Jefferson St, New York City, NY 10001

# 2. Identify who your client is based on the dataset you have chosen, and what problem you are trying to help them solve. For example, if you have chosen a dataset of grocery transactions, then your client might be the manager of a local branch of the grocery store chain, and the problem you are trying to help them solve is how to increase sales of certain items in their store.

# Identify who your client is:  
# My Client could a general manager, store owner or any higher up of a electronic store like Best Buy or Radio Shack.  
  
# What problem you will try to help them solve:  
# The problem we are trying to help solve is how to sell less popular items like headphones and Monitors. Using association analysis we can see what items are frequently bought together, then according to the findings move said items closer together or prompt people to buy them using sales or deals. This will create more more revenue and make the store more money.

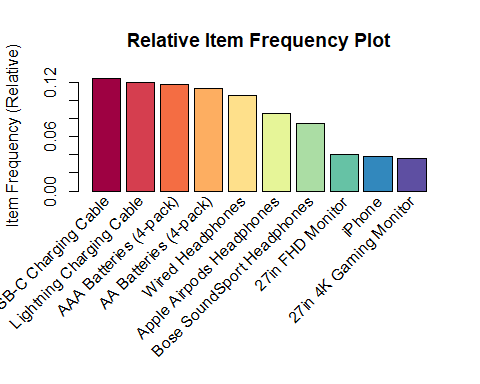
# 3. Use the arules() package in R to convert your data to a transaction object.

# Converting my data to a transaction object using the read.transactions function.  
December\_Transactions <- read.transactions(file = "C:/Users/RRC/OneDrive/Documents/CLASSES/YEAR 3/Data Mining 1/Data Sets/Sales\_by\_Month/Sales\_December3\_2019.csv",  
 header = TRUE, format = "single", sep = ",", cols = c("Order ID", "Product"))  
summary(December\_Transactions)

## transactions as itemMatrix in sparse format with  
## 24008 rows (elements/itemsets/transactions) and  
## 21 columns (items) and a density of 0.0494716   
##   
## most frequent items:  
## USB-C Charging Cable Lightning Charging Cable AAA Batteries (4-pack)   
## 2980 2888 2823   
## AA Batteries (4-pack) Wired Headphones (Other)   
## 2708 2533 11010   
##   
## element (itemset/transaction) length distribution:  
## sizes  
## 1 2 3 4   
## 23126 835 42 5   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 1.000 1.039 1.000 4.000   
##   
## includes extended item information - examples:  
## labels  
## 1   
## 2 20in Monitor  
## 3 27in 4K Gaming Monitor  
##   
## includes extended transaction information - examples:  
## transactionID  
## 1   
## 2 295665  
## 3 295666

# 4. Generate a frequency plot of the top 10 items in your dataset.

# We can also look at a plot of the Top 10 items with the highest item frequency  
arules::itemFrequencyPlot(December\_Transactions, topN = 10, col = brewer.pal(10, 'Spectral'),   
 main = 'Relative Item Frequency Plot', type = "relative", ylab = "Item Frequency (Relative)")



## Once your dataset is loaded properly as a transaction object and you have identified the client and the problem to solve:

## 5. Run the a priori algorithm to generate frequent itemsets. You should test a range of support and confidence thresholds, and provide some justification for your final choice of parameters. For example, you can use the arulesViz() package to visualize the performance of different support/confidence combinations, as we have done in class.

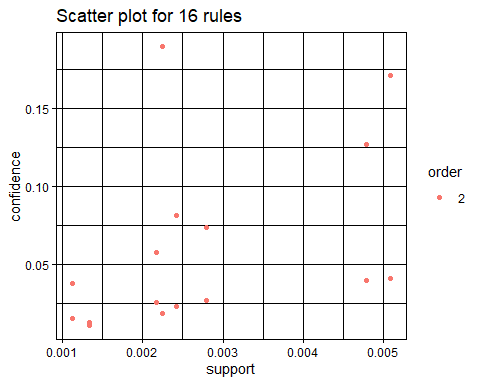
# Running the apriori algorithm using the transaction dataset.  
apriori.Rules <- apriori(December\_Transactions, parameter = list(supp = 0.001, conf = 0.01, minlen = 2))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.01 0.1 1 none FALSE TRUE 5 0.001 2  
## maxlen target ext  
## 10 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 24   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[21 item(s), 24008 transaction(s)] done [0.00s].  
## sorting and recoding items ... [19 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 done [0.00s].  
## writing ... [16 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

# Inspecting the rules generated  
inspect(apriori.Rules)

## lhs rhs support   
## [1] {Vareebadd Phone} => {USB-C Charging Cable} 0.002249250  
## [2] {USB-C Charging Cable} => {Vareebadd Phone} 0.002249250  
## [3] {Google Phone} => {Bose SoundSport Headphones} 0.001124625  
## [4] {Bose SoundSport Headphones} => {Google Phone} 0.001124625  
## [5] {Google Phone} => {Wired Headphones} 0.002415861  
## [6] {Wired Headphones} => {Google Phone} 0.002415861  
## [7] {Google Phone} => {USB-C Charging Cable} 0.005081639  
## [8] {USB-C Charging Cable} => {Google Phone} 0.005081639  
## [9] {iPhone} => {Apple Airpods Headphones} 0.002165945  
## [10] {Apple Airpods Headphones} => {iPhone} 0.002165945  
## [11] {iPhone} => {Wired Headphones} 0.002790736  
## [12] {Wired Headphones} => {iPhone} 0.002790736  
## [13] {iPhone} => {Lightning Charging Cable} 0.004790070  
## [14] {Lightning Charging Cable} => {iPhone} 0.004790070  
## [15] {Wired Headphones} => {USB-C Charging Cable} 0.001332889  
## [16] {USB-C Charging Cable} => {Wired Headphones} 0.001332889  
## confidence coverage lift count  
## [1] 0.18947368 0.01187104 1.5264712 54   
## [2] 0.01812081 0.12412529 1.5264712 54   
## [3] 0.03776224 0.02978174 0.5025476 27   
## [4] 0.01496674 0.07514162 0.5025476 27   
## [5] 0.08111888 0.02978174 0.7688520 58   
## [6] 0.02289775 0.10550650 0.7688520 58   
## [7] 0.17062937 0.02978174 1.3746543 122   
## [8] 0.04093960 0.12412529 1.3746543 122   
## [9] 0.05726872 0.03782073 0.6693805 52   
## [10] 0.02531646 0.08555482 0.6693805 52   
## [11] 0.07378855 0.03782073 0.6993744 67   
## [12] 0.02645085 0.10550650 0.6993744 67   
## [13] 0.12665198 0.03782073 1.0528604 115   
## [14] 0.03981994 0.12029324 1.0528604 115   
## [15] 0.01263324 0.10550650 0.1017781 32   
## [16] 0.01073826 0.12412529 0.1017781 32

# Plotting the rules  
plot(apriori.Rules, method = "two-key plot")



#Now, plot an interactive graph; the darkness of the nodes indicates how high the lift is  
# plot(apriori.Rules, method = "graph", engine = "htmlwidget")  
# need to comment this out even though it works to knit the file to a word doc

## 6. Run the FP-growth algorithm to generate frequent itemsets. As for the a priori algorithm, you should test a range of support and confidence thresholds, and provide some justification for your final choice of parameters.

# Kind of works but not included anymore!  
FP\_growth.rules <- rCBA::fpgrowth(December\_Sales, support = 0.001, confidence = 0.1, maxLength = 2, consequent = "Product")

## 2023-01-22 19:09:17 rCBA: initialized

## 2023-01-22 19:09:18 rCBA: data 25117x6

## took: 1.13 s

## 2023-01-22 19:09:18 rCBA: rules 44

## took: 0.54 s

inspect(FP\_growth.rules)

## lhs rhs support confidence lift  
## [1] {Purchase.Address=Purchase Address} => {Product=Product} 0.001911056 1.0000000 523.270833  
## [2] {Order.ID=Order ID} => {Product=Product} 0.001911056 1.0000000 523.270833  
## [3] {Order.Date=Order Date} => {Product=Product} 0.001911056 1.0000000 523.270833  
## [4] {Price.Each=Price Each} => {Product=Product} 0.001911056 1.0000000 523.270833  
## [5] {Quantity.Ordered=Quantity Ordered} => {Product=Product} 0.001911056 1.0000000 523.270833  
## [6] {Quantity.Ordered=} => {Product=} 0.003185094 1.0000000 313.962500  
## [7] {Order.Date=} => {Product=} 0.003185094 1.0000000 313.962500  
## [8] {Price.Each=} => {Product=} 0.003185094 1.0000000 313.962500  
## [9] {Order.ID=} => {Product=} 0.003185094 1.0000000 313.962500  
## [10] {Purchase.Address=} => {Product=} 0.003185094 1.0000000 313.962500  
## [11] {Price.Each=600.0} => {Product=LG Washing Machine} 0.003185094 0.4819277 151.307229  
## [12] {Price.Each=600.0} => {Product=LG Dryer} 0.003423976 0.5180723 151.307229  
## [13] {Quantity.Ordered=4} => {Product=AA Batteries (4-pack)} 0.001313851 0.3027523 2.798759  
## [14] {Quantity.Ordered=4} => {Product=AAA Batteries (4-pack)} 0.002866584 0.6605505 5.860490  
## [15] {Price.Each=400} => {Product=Vareebadd Phone} 0.011346897 1.0000000 88.129825  
## [16] {Quantity.Ordered=3} => {Product=AA Batteries (4-pack)} 0.005693355 0.3602015 3.329842  
## [17] {Quantity.Ordered=3} => {Product=AAA Batteries (4-pack)} 0.008201616 0.5188917 4.603675  
## [18] {Price.Each=999.99} => {Product=ThinkPad Laptop} 0.021539197 1.0000000 46.426987  
## [19] {Price.Each=109.99} => {Product=20in Monitor} 0.022614166 1.0000000 44.220070  
## [20] {Price.Each=1700} => {Product=Macbook Pro Laptop} 0.025640005 1.0000000 39.001553  
## [21] {Price.Each=300} => {Product=Flatscreen TV} 0.026356651 1.0000000 37.941088  
## [22] {Price.Each=600} => {Product=Google Phone} 0.028466775 1.0000000 35.128671  
## [23] {Price.Each=379.99} => {Product=34in Ultrawide Monitor} 0.033682367 1.0000000 29.689125  
## [24] {Price.Each=389.99} => {Product=27in 4K Gaming Monitor} 0.034239758 1.0000000 29.205814  
## [25] {Price.Each=700} => {Product=iPhone} 0.036150814 1.0000000 27.661894  
## [26] {Price.Each=149.99} => {Product=27in FHD Monitor} 0.038300752 1.0000000 26.109148  
## [27] {Quantity.Ordered=2} => {Product=USB-C Charging Cable} 0.009236772 0.1293921 1.090218  
## [28] {Quantity.Ordered=2} => {Product=AA Batteries (4-pack)} 0.021579010 0.3022867 2.794455  
## [29] {Quantity.Ordered=2} => {Product=AAA Batteries (4-pack)} 0.025361309 0.3552705 3.152006  
## [30] {Price.Each=99.99} => {Product=Bose SoundSport Headphones} 0.071983119 1.0000000 13.892146  
## [31] {Price.Each=150} => {Product=Apple Airpods Headphones} 0.081976351 1.0000000 12.198640  
## [32] {} => {Product=Wired Headphones} 0.101206354 0.1012064 1.000000  
## [33] {Quantity.Ordered=1} => {Product=Wired Headphones} 0.093880639 0.1041473 1.029059  
## [34] {Price.Each=11.99} => {Product=Wired Headphones} 0.101206354 1.0000000 9.880803  
## [35] {} => {Product=AA Batteries (4-pack)} 0.108173747 0.1081737 1.000000  
## [36] {Price.Each=3.84} => {Product=AA Batteries (4-pack)} 0.108173747 1.0000000 9.244387  
## [37] {} => {Product=AAA Batteries (4-pack)} 0.112712505 0.1127125 1.000000  
## [38] {Price.Each=2.99} => {Product=AAA Batteries (4-pack)} 0.112712505 1.0000000 8.872130  
## [39] {} => {Product=Lightning Charging Cable} 0.115220767 0.1152208 1.000000  
## [40] {Quantity.Ordered=1} => {Product=Lightning Charging Cable} 0.108014492 0.1198269 1.039976  
## [41] {Price.Each=14.95} => {Product=Lightning Charging Cable} 0.115220767 1.0000000 8.678991  
## [42] {} => {Product=USB-C Charging Cable} 0.118684556 0.1186846 1.000000  
## [43] {Quantity.Ordered=1} => {Product=USB-C Charging Cable} 0.108691325 0.1205777 1.015951  
## [44] {Price.Each=11.95} => {Product=USB-C Charging Cable} 0.118684556 1.0000000 8.425696

## 7. For each algorithm, sort the association rules by “lift” and report the 10 rules with the highest lift. Comment on similarities and differences in the top 10 rules across the two algorithms.

#Filter rules with confidence greater than 0.20  
subRules <- apriori.Rules[quality(apriori.Rules)$lift > 0]  
  
#Let's select 10 rules from subRules having the highest confidence  
top10subRules <- head(subRules, n = 16, by = "lift")  
inspect(top10subRules)

## lhs rhs support   
## [1] {Vareebadd Phone} => {USB-C Charging Cable} 0.002249250  
## [2] {USB-C Charging Cable} => {Vareebadd Phone} 0.002249250  
## [3] {Google Phone} => {USB-C Charging Cable} 0.005081639  
## [4] {USB-C Charging Cable} => {Google Phone} 0.005081639  
## [5] {iPhone} => {Lightning Charging Cable} 0.004790070  
## [6] {Lightning Charging Cable} => {iPhone} 0.004790070  
## [7] {Google Phone} => {Wired Headphones} 0.002415861  
## [8] {Wired Headphones} => {Google Phone} 0.002415861  
## [9] {iPhone} => {Wired Headphones} 0.002790736  
## [10] {Wired Headphones} => {iPhone} 0.002790736  
## [11] {iPhone} => {Apple Airpods Headphones} 0.002165945  
## [12] {Apple Airpods Headphones} => {iPhone} 0.002165945  
## [13] {Google Phone} => {Bose SoundSport Headphones} 0.001124625  
## [14] {Bose SoundSport Headphones} => {Google Phone} 0.001124625  
## [15] {Wired Headphones} => {USB-C Charging Cable} 0.001332889  
## [16] {USB-C Charging Cable} => {Wired Headphones} 0.001332889  
## confidence coverage lift count  
## [1] 0.18947368 0.01187104 1.5264712 54   
## [2] 0.01812081 0.12412529 1.5264712 54   
## [3] 0.17062937 0.02978174 1.3746543 122   
## [4] 0.04093960 0.12412529 1.3746543 122   
## [5] 0.12665198 0.03782073 1.0528604 115   
## [6] 0.03981994 0.12029324 1.0528604 115   
## [7] 0.08111888 0.02978174 0.7688520 58   
## [8] 0.02289775 0.10550650 0.7688520 58   
## [9] 0.07378855 0.03782073 0.6993744 67   
## [10] 0.02645085 0.10550650 0.6993744 67   
## [11] 0.05726872 0.03782073 0.6693805 52   
## [12] 0.02531646 0.08555482 0.6693805 52   
## [13] 0.03776224 0.02978174 0.5025476 27   
## [14] 0.01496674 0.07514162 0.5025476 27   
## [15] 0.01263324 0.10550650 0.1017781 32   
## [16] 0.01073826 0.12412529 0.1017781 32

## 8. Choose three association rules that were identified. Describe the rules in detail. Which items that appear in these rules also appeared in your frequency plot of top 10 items?

## Association rule #1  
#[1] {Vareebadd Phone} => {USB-C Charging Cable} with a lift of 1.5264712   
#[2] {USB-C Charging Cable} => {Vareebadd Phone} with a lift of 1.5264712  
  
## Association rule #2  
# [13] {Google Phone} => {Bose SoundSport Headphones} with a lift of 0.5025476   
# [14] {Bose SoundSport Headphones} => {Google Phone} with a lift of 0.5025476   
  
## Association rule #1  
# [15] {Wired Headphones} => {USB-C Charging Cable} with a lift of 0.1017781   
# [16] {USB-C Charging Cable} => {Wired Headphones} with a lift of 0.1017781   
  
  
# For the first rule we have we have Vareebadd Phone (which is an Asian brand of smartphone) are associated with USB-C Charging Cable and vise versa. The lift this rule shares is a little above 1.5 which means they start to correlate, so when someone buys either a cable or phone they are likely to buy the other one as well.  
# For the second rule we have a Google Phone which is associated with the Bose SoundSport Headphones and Bose SoundSport Headphones are associated with Google Phones. The lift this rule shares under 1 with a value of 0.503, so when someone buys either a headphones or a google phone they are likely to buy the other one as well.  
# For the third rule we have Wired Headphones associate with USB-C Charging Cable with the reverse as well. The lift value is 0.102 which means that these to accessories are very likely to be bought together.   
# Out of the 6 items I chose for the association rules, only 3 are in the top 10 frequency plot. The three items are USB-C Charging Cable, Wired Headphones, and Bose SoundSport Headphones. They rank USB-C as the most frequent, wired headphones as the fifth most purchased and Bose headphones as the seventh most purchased.

## 9. Based on the three association rules you have chosen, develop recommendations that you would make to your client. Be as creative as possible!

# I would recommend to the client that if they by a VPhone and a USB-C Charging Cable (Rule 1) then to give a discount on different lengths of USB-C cable like a 3ft 6ft and 10ft cord bundle deal. Another recommendations would be to give them a deal on a wireless plan for the first few months either give them a few at a discount or a couple months free if they buy a certain level.  
  
# I would recommend to the client that if they get a Google phone and the Bose headphones (Rule 2) to give them a spotify subscription for a discounted price. Another recommendation maybe could be to do something gym related since they are sports headphones. You could either do water bottles or gym memberships etc.  
  
# I would recommend to the client that if they get a USB-C Charging Cable and Wired Headphones (Rule 3) to put discounts on other accessories key boards or mice. You could also move the monitors closer to these products since they are already popular.

## There are two deliverables for this project:

## A 10-minute presentation in class on Thursday Dec 8, where you will summarize the work you have done on this project.

## An R-markdown submission and 1-2 page written report identifying the client, describing the problem you are helping the client solve, and outlining your recommendations to the client based on the three association rules you chose. This final report will be due on Wednesday Dec 14 by 11am. Office hours will be held Mon Dec 12 and Tues Dec 13 to help students finalize their reports.