ECON 607 Assignment 2

Andrew Girgis

September 2023

Contents

List of Figures

1 Generate Output

1.1 Using MarketBasketAnalysisData.csv, write and run the Market Basket Analysis script. Create the output and then insert it below.

Figure 1: Screen capture of the Market Basket data

	Oranges	Apples	Bananas	Berries	Pears	Grapes
0	1	1	0	0	0	0
1	1	0	1	1	1	0
2	0	1	1	1	0	1
3	1	1	1	1	0	0
4	1	1	1	0	0	1
5	1	1	0	0	0	0
6	1	0	1	1	1	0
7	0	1	1	1	0	1
8	1	1	1	1	0	0
9	1	1	1	0	0	1

Figure 2: Screen capture of the Market Basket Analysis output table

	antecedents	consequents	support	confidence	lift
0	(Grapes)	(Apples)	0.4	1.000000	1.250000
1	(Apples)	(Grapes)	0.4	0.500000	1.250000
2	(Berries)	(Bananas)	0.6	1.000000	1.250000
3	(Bananas)	(Berries)	0.6	0.750000	1.250000
4	(Grapes)	(Bananas)	0.4	1.000000	1.250000
5	(Bananas)	(Grapes)	0.4	0.500000	1.250000
6	(Berries, Oranges)	(Bananas)	0.4	1.000000	1.250000
7	(Bananas, Oranges)	(Berries)	0.4	0.666667	1.111111
8	(Berries)	(Bananas, Oranges)	0.4	0.666667	1.111111
9	(Bananas)	(Berries, Oranges)	0.4	0.500000	1.250000
10	(Berries, Apples)	(Bananas)	0.4	1.000000	1.250000
11	(Bananas, Apples)	(Berries)	0.4	0.666667	1.111111
12	(Berries)	(Bananas, Apples)	0.4	0.666667	1.111111
13	(Bananas)	(Berries, Apples)	0.4	0.500000	1.250000
14	(Grapes, Bananas)	(Apples)	0.4	1.000000	1.250000
15	(Grapes, Apples)	(Bananas)	0.4	1.000000	1.250000
16	(Bananas, Apples)	(Grapes)	0.4	0.666667	1.666667
17	(Grapes)	(Bananas, Apples)	0.4	1.000000	1.666667
18	(Bananas)	(Grapes, Apples)	0.4	0.500000	1.250000
19	(Apples)	(Grapes, Bananas)	0.4	0.500000	1.250000

2 Analyze the Results: Use the key terms to explain the most important observation(s) from the output above

2.1 Key Terms

- Antecdents: The preceding good "If" (A)
- Consequents: The following good "then" (B)
- Support: Frequency that the rules show up / A and B are bought together
  - high support = useful relationship
  - low support = probably not a useful relationship
- Confidence: Measures reliability of rule / chance to buy B if A was bought
  - $>.5$ confidence = 50% of the cases where A_1 and A_2 were purchased, the purchase also included B_1 and B_2
- Lift: The ratio of the observed support that is expected between A and B
  - 1: A and B are independent, and no rule can be derived from them
  - >1: A and B are dependent on each other
  - <1: A has a negative effect on B (items are substitutes)

2.2 Observations

In the code we set a minimum threshold of 1 for the lift, meaning that all observations seen in the output table in Figure 2 are dependent on eachother. Here we will take a deeper dive into the different cases we see in the output table. Starting off with the lift we can see 3 significant values; 1.11, 1.25, & 1.67. The greater the lift we know the more dependant A & B are on eachother.
Similarly with confidence we see the values; 0.5, 0.67, 0.75 and 1.0. The confidence measures the chance to buy B if A is bought so we know that when the confidence is equal to 0.5 then there is a 50% that the consumer will buy B with the purchase of A, however if the confidence is equal to 1.0 we know there is a 100% that the consumer will buy B with the purchase of A.
With regards to Support we see 2 values in the support column of our output table, those values being; 0.4 and 0.6. This tells us the frequency that the goods A and B are bought together and whether this is a useful relationship or potentially not a useful relationship. I will make the assumption in the rest of my analysis of 50% being the threshold of whether the relationship is useful or not. Based on this we know that the observations where the support is 0.6 or 60% are useful and the observations where the support is 0.4 or 40% are possibly not useful relationship. See Figure 3 for the filtered list of indices where the support was greater than 50%.

Figure 3: Screen capture of the useful relationships from the output

	antecedents	consequents	support	confidence	lift
2	(Berries)	(Bananas)	0.6	1.00	1.25
3	(Bananas)	(Berries)	0.6	0.75	1.25

Therefore the most important observations are indices 2 and 3, the relationship on Berries on Bananas both ways.

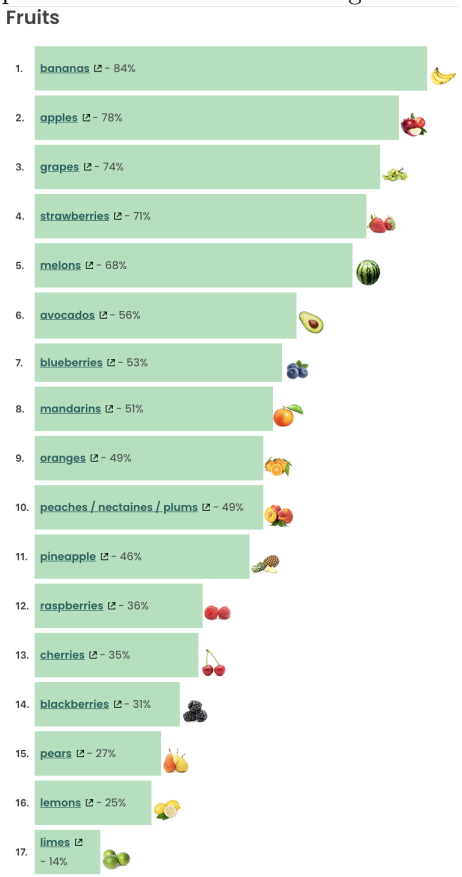
3 Provide Insights: Explain the relevance of these results

3.1 Why Bananas and berries?

Why would consumers buy bananas and berries together? Before doing any research I had no idea what the correlation could be, I kept thinking to myself 'If it were a fruit and a vegetable that would make sense since consumers would want to buy both families for a balanced diet', however bananas are fruits and berries are a subset of fruits. According to the article "What's the Difference Between Fruits and Berries?" written by Alina Petre from Healthline Media "...to be considered a berry, a fruit must develop from one single ovary and generally have a soft exocarp and fleshy mesocarp. The endocarp must also be soft and may enclose one or more seeds. Bananas fulfill all of these requirements." Under this definition a Banana is considered a berry. My first thought after reading this was about how grocery stores layout their produce sections, they will put similar produce next to each other since they are part of the same family, all the different lettuce will be on the refrigerated section on the wall next to all the similar spicing leaves like cilantro and parsley. So based on that, I concluded that since Bananas are considered berries and are in the same family they must be put next to each other in the grocery store and when consumers buy one they buy the other. However this conclusion didnt feel complete to me, and after doing some more research I came across 'Bananas are Berries. Raspberries are Not.' written by Ada McVean & Cassandra Lee that states "It turns out berry is actually a botanical term, not a common English one. Blackberries, mulberries, and raspberries are not berries at all, but bananas, pumpkins, avocados and cucumbers are." this made me realize they might not be in the same family at all so my original hypothesis didn't completely satisfy the question. It wasnt till I asked myself "what are the most bought fruits?" and the obvious next question "what are the least bought fruits?" that I realized a more probable answer to our correlation between bananas and berries. While doing research on this, I found a article written by the International Fresh Produce Association that provides a chart of the most bought to least bought fruits in the United States, See Figure 4. Thats when I realized that the reason why they are bought together so frequently IS because they are next to each other in the grocery stores but the reason why they are together in the grocery

store ISNT because they are part of the same berry family. As we can see from Figure 4, Bananas are the highest selling fruits in the United States, and berries are among the least with most of the berry family in the 12th to 14th spots on the chart (excluding blueberries and strawberrys who are in 7th and 4th respectively), due to this grocery stores will want to increase the sales of the fruits that dont do as well by putting them next to the fruits that are top sellers. The thought process is that the consumers who would go buy bananas and other top selling fruits are likely to be health conscious so they are likely to buy more fruits to balance their diet and since bananas are the highest sellers they must be a staple for most consumers who buy fruits so buy putting the fruits that dont do as well as the top sellers next to the fruits that are doing well the healthy consumers are likely to buy them in the same *basket*!

Figure 4: Screen capture of the most to least bought fruits in the United states



4 References

McVean, A., & Lee, C. (2022, November 14). *Bananas are berries. raspberries are not. Office for Science and Society.*

https://www.mcgill.ca/oss/article/did-you-know/bananas-are-berries-raspberries-are-not: :text=It%20turns%20out%20berry%20is,the%20ovary%20of%20a%20flower.

Petre, A. (2023, July 13). *Is a banana a berry or fruit? the surprising truth. Healthline.*

https://www.healthline.com/nutrition/bananas-berries-or-fruitsfruits-vs-berries

Top 20 fruits and vegetables sold in the U.S. 2022. International Fresh Produce Association. (2021, November 12).

https://www.freshproduce.com/resources/consumer-trends/top-20/

5 Code

```
# %% [markdown]
# # Import Libraries
#
#

# %%
import pandas as pd

from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

# %% [markdown]
# ### Import csv using pandas

# %%
basket_data = pd.read_csv("/Users/andrew/Downloads/UW courses/ECON 607/Assignment
2/MarketBasketAnalysisData.csv")
basket_data

# %% [markdown]
# ### Drop ID column from dataframe

# %%
basket_data = basket_data.drop(['ID'], axis=1)
basket_data

# %% [markdown]
# ### Use apriori tool to set minimum support to 40%

# %% [markdown]
# __Support__: Percentage of customers that buy the good

# %%
frequentsets = apriori(basket_data, min_support=0.4, use_colnames=True)
frequentsets


# %% [markdown]
# ### Use Association rules to get minimum lift of 1

# %% [markdown]
# __Antecdents__: The preceding good "If" <br>
# __Consequents__: The following good "then" <br>
# __Support__: The percentage of customers that bought both goods <br>
```

```
# __Confidence__: Those who bought 'Antecdent' also bought 'Consequent' x amount
of the time (in percentage) <br>

# %%

rules = association_rules(frequentsets, metric="lift", min_threshold=1)
rules

# %% [markdown]
# ### Create a new Dataframe consisting of the antecedents, consequents, support,
confidence, & lift

# %% [markdown]
# __Support__: Frequency that the rules show up / A and B are bought together
# <ul>
#     <li>high support = useful relationship
#     <li>low support = probably not a useful relationship
# </ul>
#
# __Confidence__: Measures reliability of rule / chance to buy B if A was bought
# <ul>
#     <li>0.5 confidence = 50% of the cases where Bread and Diapers were
#         purchased, the purchase also included Beer and Eggs
# </ul>
#
# __Lift__: The ratio of the observed support that is expected between A and B
# <ul>
#     <li> 1: A and B are independent, and no rule can be derived from them
#     <li> > 1: A and B are dependent on each other
#     <li> < 1: A has a negative effect on B (items are substitutes)
# </ul>
#

# %%
basket_data2 = pd.DataFrame(rules, columns=['antecedents', 'consequents', 'support',
'confidence', 'lift'])
basket_data2

# %%
filtered_data = basket_data2.loc[basket_data2['support'] >= 0.5]
filtered_data

# %%
```