Association Analysis

Exercises

1. Using the following Venn Diagram, determine the **confidence** of the following rules. (*note: there are 10,000 transactions total in this data set. 9,160 transactions did not involve any of the 3 items of interest.*

Bread $\Longrightarrow$ *Jelly*

PeanutButter $\Longrightarrow$ *Jelly*

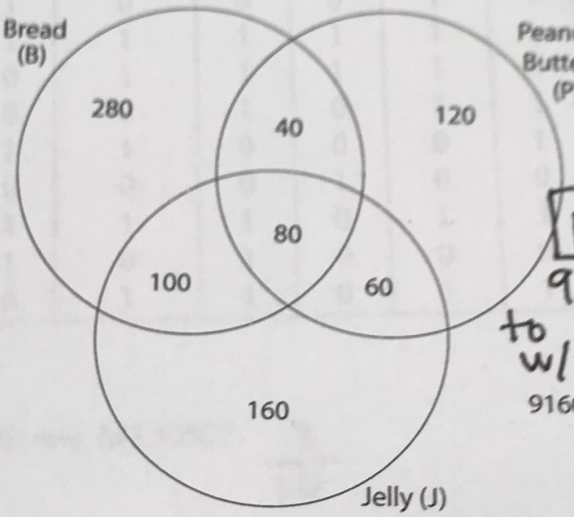
Also, determine the **lift** of the first rule, *Bread* $\Longrightarrow$ *Jelly*, and explain the result in a sentence.

$P(\text{Jelly} \mid \text{Bread}) = 0.36$

36% of those who bought bread also bought Jelly

$P(\text{Jelly} \mid \text{Peanut Butter}) = 0.47$

47% of those who bought PB also bought Jelly



Lift $B \rightarrow J = 9$

9 times more likely to see Jelly purchased w/ Bread than at random.

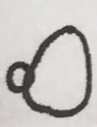
2. Give an example of an association rule that might satisfy the given conditions. For visual folks, it might help to draw a Venn diagram with two circles representing the antecendent and consequent of the rule Consider how the relative size of each circle and the overlap would be representative of support and confidence. Comment on whether such a rule might be subjectively interesting.

a) *A rule with high support and high confidence.*

cereal → milk. usually not so interesting

Many shopping trips include these items together!

b) *A rule with low support and low confidence.*



star fruit ~~cereal~~ → capers not so interesting

c) *A rule with low support and high confidence*

trailers flip over → tornado. could be interesting!

ppose you collected some data for association analysis, and you found the support of the rule

$$itemA \Longrightarrow itemB$$

was 100%. What can you say about the confidence, expected confidence and lift of this rule? What can you take away from this about rules with very high support? Are they likely to be interesting? Discuss briefly. Confidence, expected confidence = 100%.
Lift = 1
not interesting! common/obvious!

4. In an effort to learn more about visitors to news websites, you've collected data on whether or not individuals have visited certain sites in the past week. These are binary variables indicating if the website has been visited ("1") or not ("0"). You'd like to perform an association analysis to determine rules indicating whether an individual is likely to visit an alternative site, given their browsing history Using the following dataset, answer the questions below.

Name	NPR	CBS	MSNBC	ABC	FOX	POST	WSJ
Tito	1	1	0	0	0	0	0
Jojo	0	1	0	0	0	1	1
Loki	0	1	1	1	1	1	1
Niko	0	0	1	1	1	1	1
Yaya	1	0	0	1	0	1	1
Kujo	1	1	1	0	0	0	1
Arod	0	0	0	0	1	0	0
Park	1	1	1	1	0	1	1
Yorp	0	1	0	1	0	0	0
Hoki	0	0	1	1	0	1	1

a. What is the support of the rule $CBS \Longrightarrow MSNBC$? $\frac{3}{10}$

b. What is the confidence of the rule $NPR \Longrightarrow WSJ$? $\frac{3}{4}$

c. What is the confidence of the rule $WSJ \Longrightarrow FOX$? $\frac{2}{7}$

d. What is the lift of the rule $WSJ \Longrightarrow POST$? $10/7$

e. What is the lift of the rule $POST \Longrightarrow WSJ$? $10/7$

f. Suppose you put this binary data into a matrix, $\mathbf{X}$. Let $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ and explain the meaning behind $\mathbf{S}$
Hint: You should actually perform the multiplication of $\mathbf{X}\mathbf{X}^T$ and see what happens to answer this questio

S_{ij} = # stations person i and person j have in common.

S_{ii} = # stations used by persoin i