Business Analytics: Marketing and Decisions

Homework 2

Due date: **23:59PM April 16th** (Thursday), 2020

This is a group homework. Please upload your answers to homework problems (stored in **.pdf**) and the associated R code (stored in **.R**) to WM5. The file names must be BA_HW2_X, where X is your group ID. In addition, please make sure you specify your ID and name in the answer sheet. The above requirements and HW clarity are part of grading criteria. TAs have the rights to deduct points if you make the grading difficult, e.g. cannot tell the ownership of HW.

Q1. Cosmetics Purchases. The data shown in Table 14.14 and the output in Table 14.15 are based on a subset of a dataset on cosmetic purchases (*Cosmetics.csv*) at a large chain drugstore.-25pts

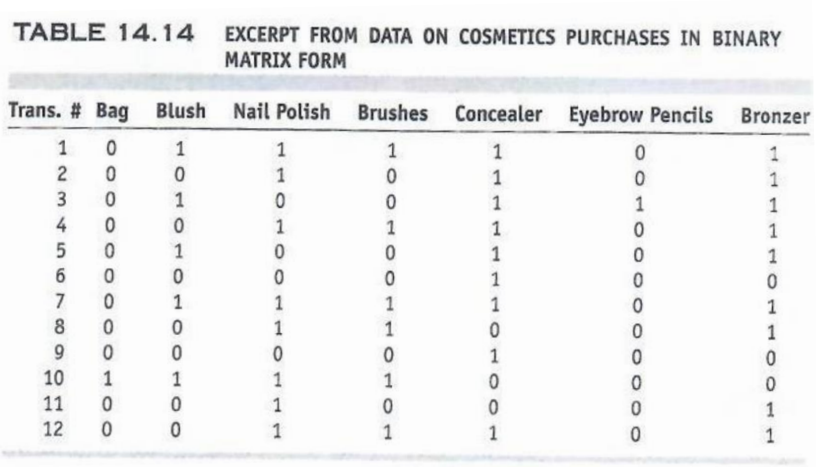

TABLE 14.14 EXCERPT FROM DATA ON COSMETICS PURCHASES IN BINARY MATRIX FORM

Trans. #	Bag	Blush	Nail Polish	Brushes	Concealer	Eyebrow Pencils	Bronzer
1	0	1	1	1	1	0	1
2	0	0	1	0	1	0	1
3	0	1	0	0	1	1	1
4	0	0	1	1	1	0	1
5	0	1	0	0	1	0	1
6	0	0	0	0	1	0	0
7	0	1	1	1	1	0	1
8	0	0	1	1	0	0	1
9	0	0	0	0	1	0	0
10	1	1	1	1	0	0	0
11	0	0	1	0	0	0	1
12	0	0	1	1	1	0	1

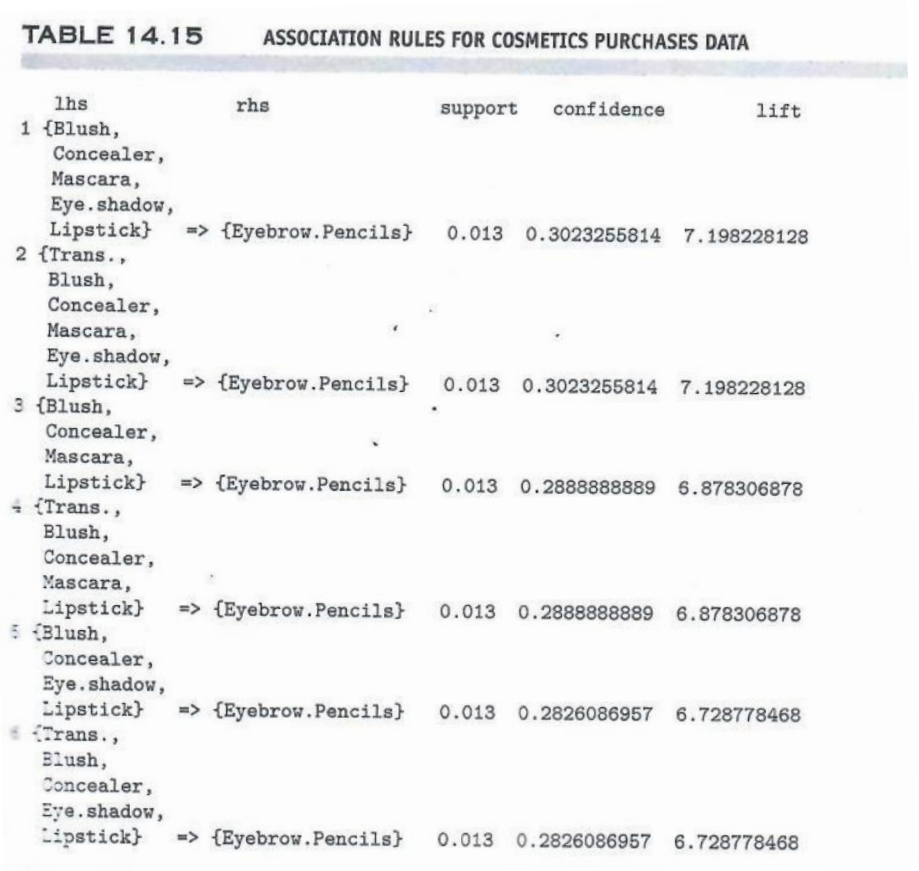

TABLE 14.15 ASSOCIATION RULES FOR COSMETICS PURCHASES DATA

```
   lhs              rhs                  support  confidence      lift
1 {Blush,
   Concealer,
   Mascara,
   Eye.shadow,
   Lipstick}    => {Eyebrow.Pencils}   0.013  0.3023255814  7.198228128
2 {Trans.,
   Blush,
   Concealer,
   Mascara,
   Eye.shadow,
   Lipstick}    => {Eyebrow.Pencils}   0.013  0.3023255814  7.198228128
3 {Blush,
   Concealer,
   Mascara,
   Lipstick}    => {Eyebrow.Pencils}   0.013  0.2888888889  6.878306878
4 {Trans.,
   Blush,
   Concealer,
   Mascara,
   Lipstick}    => {Eyebrow.Pencils}   0.013  0.2888888889  6.878306878
5 {Blush,
   Concealer,
   Eye.shadow,
   Lipstick}    => {Eyebrow.Pencils}   0.013  0.2826086957  6.728778468
6 {Trans.,
   Blush,
   Concealer,
   Eye.shadow,
   Lipstick}    => {Eyebrow.Pencils}   0.013  0.2826086957  6.728778468
```

The store wants to analyze associations among purchases of these items for purposes of point-of-sale display, guidance to sales personnel in promoting cross-sales, and guidance for piloting an eventual time-of-purchase electronic recommender system to boost cross-sales. Consider first only the data shown in Table 14.14, given in binary matrix form.

a. Select several values in the matrix and explain their meaning.-4pts

b. Consider the results of the association rules analysis shown in Table 14.15.-12pts

- **i.** For the first row, explain the "confidence" output and how it is calculated.
- **ii.** For the first row, explain the "support" output and how it is calculated.
- **iii.** For the first row, explain the "lift" and how it is calculated.
- **iv.** For the first row, explain the rule that is represented there in words.

c. Now, use the complete dataset on the cosmetics purchases (in the file *Cosmetics.csv*). Using R, apply association rules to these data (use the default parameters).-9pts

- **i.** Interpret the first three rules in the output in words.
- **ii.** Reviewing the first couple of dozen rules, comment on their redundancy and how you would assess their utility.

2. Course ratings. The Institute for Statistics Education at Statistics.com asks students to rate a variety of aspects of a course as soon as the student completes it. The Institute is contemplating instituting a recommendation system that would provide students with recommendations for additional courses as soon as they submit their rating for a completed course. Consider the excerpt from student ratings of online statistics courses shown in Table 14.16, and the problem of what to recommend to student E.N.-40pts

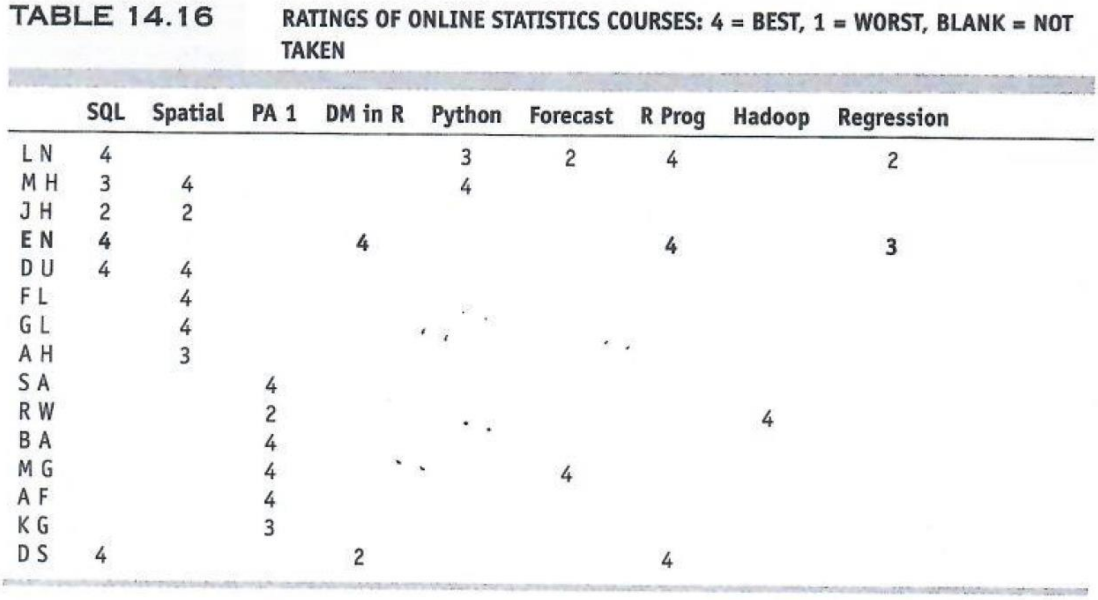

TABLE 14.16 RATINGS OF ONLINE STATISTICS COURSES: 4 = BEST, 1 = WORST, BLANK = NOT TAKEN

	SQL	Spatial	PA 1	DM in R	Python	Forecast	R Prog	Hadoop	Regression
L N	4				3	2	4		2
M H	3	4			4				
J H	2	2							
E N	4			4			4		3
D U	4	4							
F L		4							
G L		4							
A H		3							
S A			4						
R W			2					4	
B A			4						
M G			4			4			
A F			4						
K G			3						
D S	4			2			4		

a. First consider a user-based collaborative filter. This requires computing correlations between all student pairs. For which students is it possible to compute correlations with E.N.? Compute them.-9pts

b. Based on the single nearest student to E.N., which single course should we recommend to E.N.? Explain why.-4pts

c. Use R (function *similarity()*) to compute the cosine similarity between users (*CouseRatings.csv*).-4pts

d. Based on the cosine similarities of the nearest students to E.N., which course should be recommended to E.N.? Explain why -5pts

e. What is the conceptual difference between using the correlation as opposed to cosine similarities? -3pts

f. With large datasets, it is computationally difficult to compute user-based recommendation in real time, and an item-based approach is used instead. Returning to the rating data, let's now take that approach.

i. If the goal is still to find a recommendation for E.N., for which course pairs is it possible and useful to calculate correlations?-5pts

ii. Just looking at the data, and without yet calculating course pair correlations, recommend one course to E.N. based on item-based filtering. Then, calculate and report the correlation of your guess.-6pts

g. Apply item-based collaborative filtering to this dataset (*CouseRatings.csv* using R) and based on the results, recommend a course to E.N.-4pts

3. Clustering. Use the "movieLens.txt" data in Lecture 3 and perform cluster analysis of the 1664 non-duplicate movies. -35pts

a. Apply k-means clustering to the 19 binary (0/1) variables that capture movies attributes. To ensure the analysis results are stable, set the rounds of re-start to 100 and the number of maximum iterations in each round to 200. Set *k*=6 and show 5 movies in each cluster.-10pts

b. In class we identified a couple of clusters as Action/Romance (see Lecture 3 R code). Try to make sense and label each of the 6 clusters in part a. -10pts

c. Are you happy with *k*=6 and the categorization results in part b? Would you increase or decrease *k*? Explain why you increase or decrease *k* and articulate the changes in movie clusters. What is the impact of those changes for making recommendations? -15pts