The University of Texas at Dallas

School of Management

BUAN 6383/MIS 6386 **Modeling for**
Syam Menon **Business Analytics**

Homework 02

Objective

- Gain an understanding of association rule mining and clustering

Instructions

- **Due Date: See Syllabus**
  eLearning will stop accepting submissions after the due date, and late submissions will not be accepted
- **Submit one report per group via eLearning as a Microsoft Word document**
  - The report should be named **hw02-group*xx*.docx**
    (for example, group 05 should name the report hw02-group05.docx)
  - Clearly identify your group number and all group members on the cover page
  - A professional quality report is expected – messy or hard-to-read reports will be penalized
- **Submit all the code you have developed as a jupyter notebook**
  - The file should be named **hw02-group*xx*.ipynb**
  - (for example, group 05 should name the notebook hw02-group05.ipynb)
  - If you prefer, you can submit separate jupyter notebooks for each question If you choose to do so, the files should be named **hw02-group*xx*-p*yy*q*zz*.ipynb** (for example, group 05 should name the notebook
  - for Part 2, question 3 hw02-group05-p02q03.ipynb)
  - Clearly identify which question each part of the code is for, and what it is supposed to do
  - Clear, detailed comments are required; I should be able to run the codes you submit
- **This homework counts for 70 points**

Data Sets

- `transactions.csv`
- `bankcustomers.csv`

Part I: Analyzing Transactions

A store is interested in determining relationships between items purchased from its *Stationary* and *Health and Beauty Aids* departments. They wish to conduct a market basket analysis of items purchased from these departments. `transactions.csv` contains information on 200,000 transactions made over a three-month period. The file has two columns – the transaction id (`Transaction`) and the product purchased (`Product`), and contains over 400,000 rows (as each row involves a single product, transactions involving multiple products span multiple rows). Seventeen products are represented in the data set: bar soap, bows, candy bars, deodorant, greeting cards, magazines, markers, pain relievers, pencils, pens, perfume, photo processing, prescription medications, shampoo, toothbrushes, toothpaste, and wrapping paper.

1. Read in the data and generate a file in which every row represents a transaction, with `True` identifying items that were part of that transaction, and `False` identifying items that were not (as in the example from class). Name the file **group*xx*transactions01.csv**, where ***xx*** is your group number.

2. Identify the frequent itemsets using a minimum support threshold of 1%. How many itemsets are frequent?

3. Identify all association rules with a minimum confidence of 10%. How many rules are generated?
4. Which rules have the highest lift? Using the results from the previous questions, show exactly how this lift value was calculated for one of the rules with highest lift.
5. For the same rule, show how leverage and conviction were obtained.
6. Interpret and discuss the 5 rules with
   a. the highest confidence,
   b. the highest lift,
   c. the highest leverage, and
   d. the highest conviction.

   If there are more than five meeting the required criterion, pick any five. Are any of these surprising? Comment on the extent of their redundancy and utility.
7. Do any of these metrics seem preferable to the others for this dataset? Discuss why or why not.
8. If you were in charge of these departments, how would you use the results of this analysis to come up with a strategic plan? Explain your reasoning. This question is open ended, and I am looking for innovative thinking.

Part I: Clustering Customers

A bank in the U.K. is interested in identifying categories of customers for potential future promotions. Data collected on 600 customers is in `bankcustomers.csv`; the variables involved are below.

Variables	Description
`id`	customer ID
`age`	age in years
`sex`	1 → female, 0 → male
`region`	1 → inner city, 2 → town, 3 → rural, 4 → suburban
`income`	income in $
`married`	1 → married, 0 → not married
`children`	number of children
`car`	1 → owns car, 0 → does not own car
`savings`	1 → has savings account, 0 → no savings account
`current`	1 → has checking account, 0 → no checking account
`mortgage`	1 → has mortgage, 0 → no mortgage
`pep`	1 → has personal equity plan, 0 → no personal equity plan

pep provided tax incentives to promote individual investment in stocks.

1. Notice that `region` is categorical; we need to do what is referred to as "one-hot encoding" – convert it into separate (binary) variables, one for each possible value of `region`. So you will need to create 4 new variables, corresponding to inner city, town, rural, and suburban (a 1 in a column would represent being from the associated region). You can do this either explicitly by writing your own code, or by using the OneHotEncoder option available in sklearn (preprocessing). Read in the data, create the four new columns, and drop `region`.
2. Apply hierarchical clustering (with Euclidian distance as the measure of distance) to the dataset using (i) centroid linkage, (ii) single linkage, (iii) complete linkage, (iv) average linkage, and (v) Ward linkage. For each of these, comment on whether you see any clear clusters, and how many clusters you would recommend (and why). Across all the linkage approaches tried, which one has worked best in this example (provide your reasoning)? What are some distinguishing characteristics of each cluster?
3. Apply *k*-means clustering to the dataset. Try different values of *k* (4, 5, 6, 7, and 8 at least); make sure you include the number of clusters you decided to use with hierarchical clustering. Are clear clusters visible for any value of *k*? As before, how many clusters would you recommend, and why? What are some distinguishing characteristics of

each cluster? How different are these results from those with hierarchical clustering? Which seems preferable in this case? Explain.

4. If you were the manager of this bank, how would you use the results of this analysis to come up with a strategic plan? Explain your reasoning. As in Part I, this question is open ended, and I am looking for innovative thinking.