CS5228 Lab Assignment 2

Mining Frequent Patterns and Association Rules

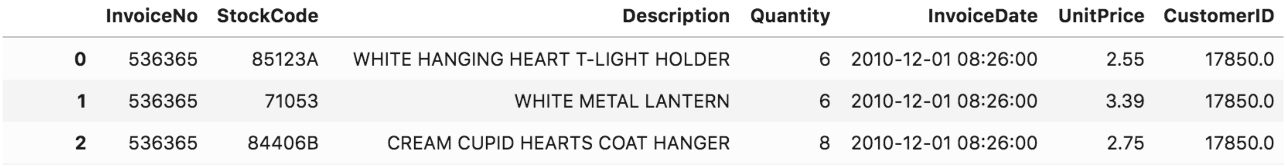
Semester 2 2018/19
Due date: 15 March 2019

You are now working as a data analyst for an online shopping company called Nozama. You have been asked to mine the current sales records to derive insights that can help the management develop new strategies to improve the product sales.

The dataset can be found in `record.csv`. It contains the following columns:

Column Name	Explanation
InvoiceNo	The ID of the transaction
StockCode	The ID of the item
Description	The name of the item
Quantity	The number of an item bought in the transaction
InvoiceDate	The date of the transaction
UnitPrice	The unit price of the item
CustomerID	The ID of customer

For example, the following records indicate that customer 17850 bought six “WHITE HANGING HEART T-LIGHT HOLDER”, which has stock code 85123A, on 1st Dec 2010. In the same transaction 536365, the customer also bought items 71053, 84406B etc.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0

To get started, you will need to install the following software packages:

- Python (version 3.6 or newer)
- Jupyter Notebook or Jupyter Lab
- Common python modules: pandas, numpy, matplotlib, seaborn
- efficient-apriori (https://pypi.org/project/efficient-apriori/)
- SPMF (http://www.philippe-fournier-viger.com/spmf/)

Here are some very useful webpages to find out more about the packages that you will be using for this assignment:

- https://pandas.pydata.org/pandas-docs/stable/indexing.html
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.groupby.html
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.apply.html
- https://stackoverflow.com/questions/39922986/pandas-group-by-and-sum

You can also google "<what you want to know> + stackoverflow" if the URLs above do not solve you problems.

Submission

1) An IPython Notebook file, `answer_sheet.ipynb`, is provided. You are expected to write all your codes and answers within the indicated spaces in the IPython notebook (answers to the conceptual questions can be embedded in the Notebook as markdown cells). Please fill in your answers to Sections 1~3 in `answer_sheet.ipynb`. Submit a single IPython notebook with the name "`YourNameInIVLE_YourIDInIVLE.ipynb`" to the submission folder in IVLE.

2) For Section 4, please submit a **one-page pdf file** named "`YourNameInIVLE_YourIDInIVLE.pdf"` to the submission folder in IVLE. Note that submissions that violate page limits or naming conventions will be considered void.

If you have further questions, you can email ziwei.xu@u.nus.edu .

1. Data Cleaning and Exploration (20 points)

Before continuing, let us examine the dataset for "dirty" records to do some data cleaning. There are at least two types of "dirty" records in the dataset. Please provide a description of each of the types of "dirty" records that you can find in the dataset, as well as the corresponding number of such records that are to be removed the dataset. (5 points)

After removing the "dirty" records, let us explore the dataset by getting "quick facts" such as those listed in the table below. Please provide the answers to the questions listed in the table. (3 points)

	Question	Answer
1)	Starting date of the dataset?	(YYYY-MM-DD)
2)	Ending date of the dataset?	(YYYY-MM-DD)
3)	Number of customers?	(Integer)
4)	Number of transactions?	(Integer)
5)	Number of different kind of items?	(Integer)
6)	Number of transactions customer ID 17850 have made?	(Integer)
7)	Which customer (ID) have made the most transactions?	(Integer)
8)	What is the item ID of the best-seller? We define "best-seller" as the item with the highest sales volume.	(Integer)

9) Next, let us get some general understanding about the transactions. Please make a histogram of the number of unique items per transaction (1 points) and describe one insight that you can observe from the plot, and explain why you find it interesting. (3 points)

10) We can also explore the data based on the items. Let us make a bar plot of the items with support higher than 5%. (2 points) Please describe one insight that you can observe from the plot, and explain how it can be related to rule mining. (3 points)

11) Compare the “best-seller” that you have previously found in (8) and the item with the highest support that you have just found in (10). Which item do you think is more popular? Here we define the “popular” items as items that are bought by many customers. Explain you answer and describe any assumptions you’ve made. (3 points)

2. Mining Association Rules (30 points)

After taking some efforts to explore the dataset to gain a good degree of familiarity with the data, you are now ready to mine the dataset for frequent patterns and association rules.

Please note that questions (2)~(4) below can require fairly long computation time to complete. We suggest that you use python’s built-in module `pickle` to save your intermediate results once you get them.

1) Let us first consider whether the “brute-force” counting method (i.e. counting all possible itemsets) is feasible. Suppose we can count 2^{36} itemsets per second. Will we complete the counting before the sun burns out (the sun has another $5 \times 10^9 < 2^{33}$ years to burn)? (4 points)

2) Run efficient-apriori in python with **min_support**=0.5%, **min_confidence**=20%, max_length=4. Write down the rule with the highest lift (denoted as r_1). (5 points)

3) Run efficient-apriori in python with **min_support**=1%, **min_confidence**=20%, max_length=4. Write down the rule with the highest lift (denoted as r_2). (4 points)

4) Run efficient-apriori in python with **min_support**=0.5%, **min_confidence**=40%, max_length=4. Write down the rule with the highest lift (denoted as r_3). (4 points)

5) You must have noticed numerous differences between the two runs in (2) and (3). List at least 3 differences you’ve found. You may want to consider the elapsed time and the quality of the results. (3 points)

6) Which one in $r_1, r_2,$ and r_3 do you think is better? Explain your answer. (4 points)

7) From your observation, what are the effects of increasing/reducing **min_support** and **min_confidence**? Support your answer with evidence. (6 points)

3. Mining Sequential Rules (30 points)

A customer may buy things in different transactions over time. It will be interesting if we can enable Nozama to know what customers tend to buy in the future, given what they have bought at present.

Let us define a sequence as a list of all transactions made by a certain user, ordered by time. Each transaction in the sequence is a set of stock codes. For example, given a sequence [{1,3}, {2}, {4}], we can know that the customer had bought item 4 after he/she had bought items 1,2,3, and that items 1 and 3 were bought together in the same transaction.

1) First, let's organize the dataset into sequences and fill the table below (4 points):

CustomerID	Beginning Date of Sequence	Ending Date of Sequence	Number of transactions in the sequence
12356	(YYYY-MM-DD)	(YYYY-MM-DD)	(Integer)

2) In this section, you are going to use the **ERMiner** algorithm in the SPMF software to mine sequential rules. You do not need to know the details of the algorithm for this exercise. Just follow the example in http://www.philippe-fournier-viger.com/spmf/ERMiner.php to gain a basic understanding on using the algorithm for mining sequence patterns. Run ERMiner with **min_support**=0.5%, **min_confidence**=60%, maximum length of antecedent and consequence being 1. Write down the rule(s) with the highest confidence. (8 points)

3) Do you think the parameters used in (2) are good in practice? (2 points) If yes, explain your answer. If no, give a better set of parameters and explain why it is better. (6 points)

4) What are the differences between sequential rules and association rules? Give 2 of them. (6 points). Describe a circumstance where sequential rules apply but association rules do not. (4 points)

4. Insights to Actions (20 points)

As mentioned, your task is to mine the current sales records to derive insights that can help inform the Nozama management to develop new strategies to improve the product sales. For example, are there any interesting rules discovered that can be exploited? Are there any insights about the possible profiles of Nozama's customers? Note that this is an open question. However, we do require that you should support your arguments with evidence. If necessary, you might want to explore the data further.

Please prepare a **one-page report** on the insights that you have derived from the dataset and how Nozama management can turn the insights into actions that will help it improve its sales and services. Please provide at least 4 insights and the corresponding actions. The insights should not overlap with what you have found in Sections 1~3.