Exercise 4
Frequent Itemset Extraction and Association and Correlation Rules

Data format:

D is said to be a relevant dataset for the Frequent Items extraction task, when it groups a set of transactions $T = \{T_1, T_2, \dots T_n\}$ from a database where each transaction T_i is a non-empty set of elements with a number of Items from I , with $I = \{I1, I2, I3, \dots Im\}$.
The data in a dataset is not always formatted adequately for Frequent Items extraction. Therefore, it is sometimes necessary to detect what represents a Transaction and what represents an Item in our Dataset. Therefore, the dataset will have to undergo changes in order to group the Items of each transaction together before extracting any information from the dataset.

Questions :

1- Eliminate the first 7 columns and do the necessary preprocessing.

2- Study the dataset and deduce what are the Transactions and Items.

3- Give the number of Transactions and the number of Items in this dataset.

4- Build a dataset “DatasetExos_2” from the dataset to obtain the correct transactional format necessary for the extraction of Frequent Items.

Support : *supp_min*

The support of the 1-itemset $I1$ = Percentage of transactions containing item $I1$.
= number of Transactions in which item $I1$ appears / number of transactions of D.

The support of the 2-itemset $\{I1, I2\}$ = Percentage of transactions containing item $I1$ and $I2$ at the same time.
= number of Transactions in which items $I1$ and $I2$ appear / number of transactions of D.

… and so on.

At each Iteration of the Apriori algorithm, a list of candidate k-itemsets C_k is built. And from each C_k a list of frequent k-itemsets L_k is created, keeping only the k-itemsets of C_k having a support >= supp_min (variable to be fixed).

- **C_1 is the list of 1-itemset candidates.**
- Generation of C_1 through the listing of all Items.
- **L_1 is the list of frequent 1-itemsets of C_1**
- Compute support of all elements of C_1 from base dataset D.
- Copy only elements with support >= supp_min.

- **C_2 is the list of 2-itemset candidates.**
- Generate C_2 through a join operation:
  For each item Ii of L_1
    For each item Ij of L_1 (Ii <> Ij)
      Add { Ii , Ij } to C_2 ;
    Done;
  Done;
- **L_2 is the list of frequent 2-itemsets of C_2.**
- Compute support of all elements of C_2 from base dataset D.
- Copy only elements with support >= supp_min.

… and so on.

Questions:

1- Write a function to generate the k-itemset candidates C_k.

2- Write a function to calculate the support of the k-itemsets C_k.

3- Write a function to generate the frequent k-itemsets L_k.

Confidence : *conf_min*

Let A and B be k-itemsets and A ⇒ B be an association rule.
Example: A = {$I1, I2$} and B = {$I3$}. The rule: {$I1, I2$} ⇒ {$I3$}.
Confidence(A ⇒ B) = The percentage of transactions in D containing A that also contain B.
This is the conditional probability, P(B/A).

Confidence(A ⇒ B) = P(B/A) = Support (A U B) / Support (A)

Questions:

1- Write a function to generate all the association rules of a L_k.

2- Write a function to calculate the confidence of an association rule.

Dataset: DatasetExos.csv is available on this link.

Have fun !