Assigned: 04/03/2022
Due: Fri. 04/22/2022, 11:59pm

Instructions: This project will cover topics of association analysis and recommender systems.

Name your main submission files as *A5_Group_<GroupName>.extension*. You will be submitting your "code" file and a PDF of the code + text + results of your code running.

The following are acceptable formats:

Language	"Code" File Format	PDF Generation
R	R Markdown, .Rmd	Rmd → PDF: Use `knitr` or `rmarkdown` to collect all text responses, figures, tables, and code in the R Markdown file and process it to produce a PDF file.
	R Sweave, .Snw	Snw → PDF: Use `R Sweave` to collect all text responses, figures, tables, and code in the Snw file and process it to produce a PDF.
Python	Jupyter notebook, .ipynb Colab notebook, .ipynb	.ipynb → PDF: Incorporate all text responses, figures, tables, and code in the notebook and process it to produce a PDF file.

Submission Requirements: Your answers must be computer generated (including text and diagrams). Your final document submission should include text responses to questions and description of your efforts, tables, R/Python code used to calculate answers, and figures.

Create a zip-file called A5_Group_<*GroupName*>.zip and submit on Canvas. Your group name, <*GroupName*> will be given on Canvas.
For example, if I was using R, I would submit either:

- A5_Group_<*GroupName*>.Rmd, A5_Group_<*GroupName*>.pdf, or
- A5_Group_<*GroupName*>.Snw, A5_Group_<*GroupName*>.pdf

For Python, I would submit:

- A5_Group_<*GroupName*>.ipynb, A5_Group_<*GroupName*>.pdf

Any other packages or tools, outside those listed in the assignments or Canvas, should be cleared by Dr. Brown before use in your submission.

Questions:

1. (16 points) Confirm the results from Part1 Q2 using data table below. For R, the `arules` package is available. Python has the 'mlxtend' library.

2. You will analyze a portion of the Instacart Online Grocery Shopping Dataset 2017[1]. The 2 data sets you are given contains just 20K or 500K items purchased, while the original data set has 3 million orders.

   You will only need to focuse on the following files: 'order_products__train_small.csv', 'order_products__train_med.csv', and 'products.csv' for this analysis. You can link the product number in the "order_products" file to the name of the product in the "products.csv" file.

[1]The full data set is available here: https://www.instacart.com/datasets/grocery-shopping-2017

TID	Items
T100	B, D, F, G, I, J
T200	C, B, D, G, I, J
T300	D, F, G, H
T400	A, F, G, J, K
T500	A, B, D, E, G

Table 1: Transaction Data

(a) (12 points) Create a histogram showing the number of products per order for both the 'order_products__train_small.csv' and 'order_products__train_med.csv' data sets. Indicate with a vertical line where the mean number of products per order lands.

(b) (6 points) For the 'order_products__train_small.csv' data, create an top 15 item frequency plot, that is plot the top 15 most frequently purchased items. This should be a bar plot with items vs. frequency (relative support).

(c) (6 points) For the 'order_products__train_med.csv' data, create an top 15 item frequency plot, that is plot the top 15 most frequently purchased items. This should be a bar plot with items vs. frequency (relative support).

(d) (24 points) For the 'order_products__train_small.csv' data, use Apriori to find association rules with a minimum support of 0.003 and confidence of 0.5. Report in a table the top 10 rules (sorted by lift) with the product names, the support, confidence and lift.

(e) (16 points) Rerun Apriori on the same data set with a minimum support of 0.0025 and confidence of 0.5. Create a scatterplot of the rules, plotting support vs. confidence colored by lift value.

(f) (16 points (bonus)) For the 'order_products__train_med.csv' data, use Apriori to find association rules with a minimum support of 0.003 and confidence of 0.5. Report in a table the top 10 rules (sorted by lift) with the product names, the support, confidence and lift.