Class 07: ETL Lab

Please download the [PortMap.csv](https://miamioh.instructure.com/courses/179824/files/25343426?wrap=1) (and feel free to skim [Description of the Dataset](https://www.unb.ca/cic/datasets/ddos-2019.html) for more details),  and write a script that would: (a) read only the first 1000 rows, and (b) reduce your dataset to the following columns:

*'Flow ID', 'Source IP', 'Destination IP', 'Timestamp', 'Flow Duration', 'Total Fwd Packets', 'Total Backward Packets', 'Max Packet Length', 'Average Packet Size', 'Active Mean',  'SimillarHTTP', 'Label'*

Upon completing the task above, please answer the following questions:

1. Is the resulting frame technically correct? If not, please convert any columns that need to be converted
2. Please provide a plot of the top 12 Flow IDs and their corresponding frequency
3. Compute % begnin and malicious labels
4. Compute median flow durations
5. What is the third quartile of the average column length variable?
6. Create a dataframe titled, 'df\_grouped', where you are grouping the data.frame by the label column. Once you create this dataframe, please provide a nice compact summary table containing a column for each of the following:
   1. means, standard deviations of each of the following variables:
   2. Flow Duration, Total Fwd Packets, Total Backward Packets, Max Packet Length and Active Mean

Upload a knitted RMarkdown (HTML) which shows your code and solutions to each of the questions above.