Clustering Methodology

* Our group decided that the best statistical measures with which to attempt to cluster the various industries in each Excel sheet were by mean, median, and standard deviation of all the values for each industry.
* There are instances where values are missing in places in many industries across all our Excel sheets. Our code was written to include only the values that were ascertained. As such, the actual number of values used to calculate each statistical measure will vary.
* To find mean values:
  + Utilized a package named “matrixStats”
  + Used function “colMeans” from package to find the mean value for each column of our data frames, each column being one industry.
  + Converted all the industry means into a list and added that list as a row to the bottom of the data frames.
  + Reordered the columns in each data frame by the newly created mean values for each industry, such that the means were in ascending order.
  + Separated each data frame into “Adjusted” and “Unadjusted” industries.
  + Wrote a function that plots each industry column as a time series from the sorted data frames. The function was written to cluster industries in groups of five.
  + Axis names, legend, different line types, and color were added to each individual time series to make the plot more accessible.
* To find median values:
  + Utilized a package named “matrixStats”
  + Used function “colMedians” from package to find the median value for each column of our data frames, each column being one industry.
  + Converted all the industry medians into a list and added that list as a row to the bottom of the data frames, after the row containing the mean values.
  + Reordered the columns in each data frame by the newly created median values for each industry, such that the medians were in ascending order. Note that this order may be similar but not the same as the order of the columns when sorted by the mean values.
  + Separated each data frame into “Adjusted” and “Unadjusted” industries.
  + Wrote a function that plots each industry column as a time series from the sorted data frames. The function was written to cluster industries in groups of five.

Axis names, legend, different line types, and color were added to each individual time series to make the plot more accessible.

* To find standard deviation values:
  + Utilized a package named “matrixStats”
  + Used function “colSds” from package to find the median value for each column of our data frames, each column being one industry.
  + Converted all the industry standard deviations into a list and added that list as a row to the bottom of the data frames, after the rows containing the mean and median values.
  + Reordered the columns in each data frame by the newly created standard deviation values for each industry, such that the standard deviations were in ascending order. Note that this order may be similar but not the same as the order of the columns when sorted by the mean or median values.
  + Separated each data frame into “Adjusted” and “Unadjusted” industries.
  + Wrote a function that plots each industry column as a time series from the sorted data frames. The function was written to cluster industries in groups of five.

Axis names, legend, different line types, and color were added to each individual time series to make the plot more accessible.

* Items of Note:
  + The number of industries in each table was not always divisible by five. Thus, some of the plots of the clusters may have slightly more or less than five industries included. The orders determined by mean, median, and standard deviation values have been preserved.
  + The best clusters will be selected purely by our judgement of their shape and likeness. We will not be using any other statistical or analytical measures to attempt to find the “best” clusters.
  + A data dictionary will be added so that the legends that are included with each plots may be quickly interpreted. The best clusters might need have captions underneath explaining the industries that the plot contains when it comes time to making our poster.