Due Date 1 Feb (Wed)

Cluster the customers and products for the retail dataset.

We want to cluster customers based on:

* number of products bought
* number of distinct products bought
* revenues
* number of visits
* additional attributes that you think are appropriate

We want to cluster products based on:

* number of distinct customers who buy the product
* revenues
* number of visits in which the product is bought
* additional attributes that you think are appropriate

As part of the process, we may find outliers, which may lead to some data cleaning. You should submit a report of your findings that may be beneficial to the business owner in understanding the nature of customers. The report should also briefly outline the data cleaning and preparation.

Suggested detailed algorithm and report contents:

1. Make sure that your data does not have values that are too high or too low. You can find these values with the help of max and min functions in a spreadsheet. You could also look at the distributions of each attributes through R or spreadsheet. If you choose to delete some records, please describe how you found the outliers and what were the values.
2. Once all the values are in reasonable ranges, normalize them so that no values are larger than 100. You have to be careful with such a normalization.
3. Experiment with different number of clusters using simple K-means in R. I would suggest using 5, 10, 15 as number of clusters. Clustering may help you further identify outlying objects. If you go further and find the optimal number of clusters, that will be even better, especially if you can justify your answer.
4. When you are presenting the results use the centroid data produced by R. You may also plot the cluster distribution by varying X and Y values in R. Show only those graphs that are more meaningful.

**Using your judgement instead of blindly following instructions is an important trait. Creativity with justification is encouraged.**

**Note for submission:** Submit a pdf (no word processor files) file describing your data processing efforts and analysis. Add the programs, R commands, and queries in appendix.