ASSIGNMENT 1:

DATA EXPLORATION AND PREPARATION USING EXCEL

**This assignment is done in group and due on Sunday, 13 September 2018. Each group submits the assignment using Elisa (zip file). All calculation should be done using Excel’s functions. *NOTE: You can use Open Office instead of MS Office.***

Consider the data collected by a hypothetical video store for 50 regular customers. This data consists of a table which, for each customer, records the following attributes: *Gender, Income, Age, Rentals* (total number of video rentals in the past year), *Avg. per visit* (average number of video rentals per visit during the past year), *Incidentals* (whether the customer tends to buy incidental items such as refreshments when renting a video), and *Genre* (the customer's preferred movie genre). This data is available in Elisa as an Excel spreadsheet (Video\_Store.xls). Using spreadsheet do the following tasks:

1. Explore the general characteristics of the data, by computing the means and standard deviations of the numerical attributes.
2. Perform the the following data preparation steps on the data (for each task add a new column to the original table for comparison purposes)

* Use smoothing by bin means to smooth the values of the Age attribute. Use a bin depth of 4.
* Use min-max normalization to transform the values of the Income attribute onto the range [0.0-1.0].
* Use z-score normalization to standardize the values of the Rentals attribute.
* Discretize the (original, non-normalized) Income attribute based on the following categories: High = 60K+; Mid = 25K-59K; Low = less than $25K.

1. Convert the original table (not the results of part 2) into the standard spreadsheet format. Note that this requires converting each categorical attribute into multiple attributes (one for each values of the categorical attribute) and assigning binary values corresponding to the presence or not presence of the attribute value in the original record). For example, the Gender attribute will be transformed into two attributes, "Genre=M" and "Genre=F". The numerical attributes will remain unchanged. This process should result in a new table with 12 attributes (one for Customer ID, two for Gender, one for each of Income, Age, Rentals, Avg. Per Visit, two for Incidentals, and three for Genre).

1. Using the standardized data set (from part 3), perform basic correlation analysis among the attributes. Discuss your results by indicating any strong correlations (positive or negative) among pairs of attributes. You need to construct a complete Correlation Matrix. Be sure to first remove the Customer ID column before creating the correlation matrix.

1. Perform a cross-tabulation of the two "gender" variables versus the three "genre" variables. Show this as a 2 x 3 table with entries representing the total counts. Then, use a graph or chart that provides the best visualization of the relationships between these sets of variables. Can you draw any significant conclusions?

1. Select all "good" customers with a high value for the Rentals attribute ( a "good customer is defined as one with a Rentals value of greater than or equal to 30). Then, create a summary (e.g., using means, medians, and/or other statistics) of the selected data with respect to all other attributes. Can you observe any significant patterns that characterize this segment of customers? Explain. Note: to know whether your observed patterns in the target group are significant, you need to compare them with the general population using the same metrics.

1. Suppose that because of the high profit margin, the store would like to increase the sales of incidentals. Based on your observations in previous parts discuss how this could be accomplished (e.g., should customers with specific characteristics be targeted? Should certain types of movies be preferred? Etc.). Explain your answer based on your analysis of the data.

1. Using the second sheet in the file (*Sheet No. 8*), determine the closest customer (the most similar) customer to the customer ID 30 (excluding customer ID 30), when the data is normalized and not normalized. Is there any difference in your result?