**Homework-1**

You are allowed to use any of the popular statistical software package for this homework - examples include but not limited to R-studio, MATLAB, WEKA, Python, etc.

You are given two csv files - one from amazon product and the other from google products. Answer the following questions.

1. For all the microsoft products in amazon file, find out if there is any outlier wrt price.

a. Present your source code.

b. Present a visualization.

(3 points = 1.5 + 1.5)

2. For all the game related products in googleproducts.csv, present a histogram, where X-axis is the manufacturer name and Y-axis is the no of products by the manufacturer.

a. Present your source code.

b. Present a visualization.

(3 points = 1.5 + 1.5)

3. There are products in these two files who have different name and description in these two files but referring to the same entity.

a. Present your entity resolution algorithm with appropriate pseudo-code and description.

b. Present your source code.

c. Present a csv file as the output of your algorithm which contains one to one mapping between the ids of the products (one from amazon and the other from google), as they are identified as the same entity by your algorithm.

(9 points = 3+3+3)

|  |  |
| --- | --- |
|  | Amazon.csv [Amazon.csv](http://njit2.mrooms.net/pluginfile.php/601424/mod_assign/introattachment/0/Amazon.csv?forcedownload=1) |

|  |  |
| --- | --- |
|  | |
|  | GoogleProducts.csv [GoogleProducts.csv](http://njit2.mrooms.net/pluginfile.php/601424/mod_assign/introattachment/0/GoogleProducts.csv?forcedownload=1) | |