Final Advice

We have tried to conduct cluster analysis on the data set. Initially we used the k-means algorithm for the analysis. We set the value of *k* to 2 in the hope of finding roughly the same data division for buyers and non-buyers as represented by the labeled data set, unfortunately this was not the case. Only five data points were assigned to one of the two clusters, the other 459 were assigned to the other, whilst the labeled data is composed of the same number of buyers as non-buyers.

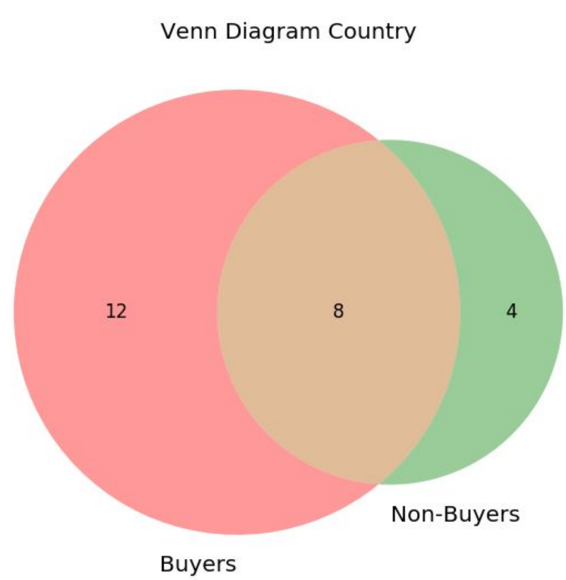
Therefore we decided to confine ourselves to clustering within the subset of the data labeled as non-buyers. The objective was to find outstanding features that might explain why certain groups are less likely to purchase Bomberbot. For example, a cluster with a relatively high value for Average Time might be less interested in continuing to use bomberbot after the trial, since the high value possibly implies that the children have already finished a great part of the game or that they find the game too difficult.

We applied the mean-shift algorithm to the set of non-buyers and found five clusters. One of these clusters represented the group of people that never actually used their accounts, i.e. the values for Amount Students, Average Amount Students and Average Time were all 0. A second cluster consisted of accounts that were used, but we didn’t find a striking correlation between the different users for any of the features. The remaining clusters consisted of only one data point each and were therefore disregarded.

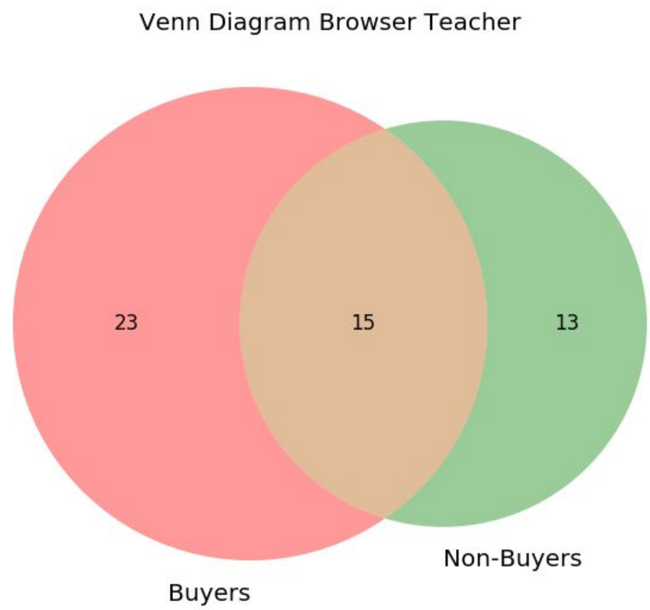
From the above can be concluded that applying an unsupervised algorithm to the given data set does not work well and therefore we have not been able to explain why certain people are less interested in purchasing Bomberbot, as we intended.

To still be able to explain some part of a lower tendency to continue with Bomberbot after the trial period, we have visualised the value distribution for the categorical features between buyers and non-buyers using venn diagrams. The diagrams are shown below.

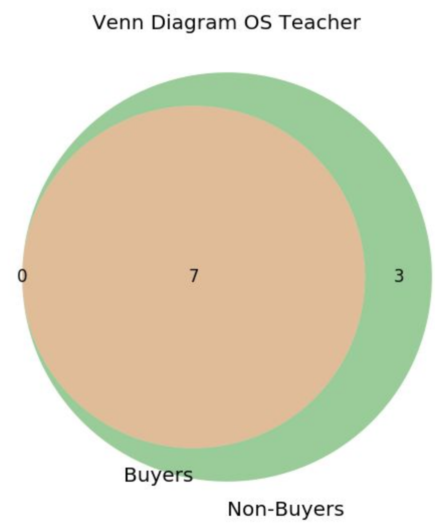
Country:
argentina 0.0%
belgium 0.01%
italy 0.0%
morocco 0.01%



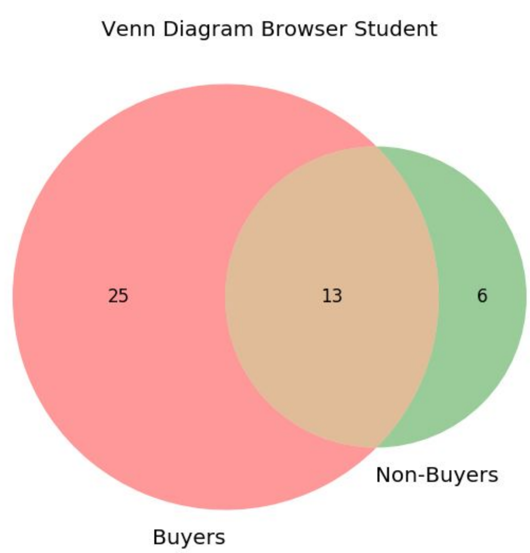
Browser Teacher:
Firefox 34.0 0.0%
Firefox 39.0 0.0%
Chrome 41.0.2272.118 0.0%
Firefox 40.0 0.02%
Chrome 44.0.2403.157 0.0%
Safari 4.0 0.0%
Chrome 45.0.2454.98 0.0%
Safari 7.0 0.0%
Chrome 46.0.2490.64 0.0%
Safari 7.1.8 0.0%
Safari 8.0.6 0.0%
Safari 8.0.8 0.0%
Chromium 45.0.2454.101 0.0%



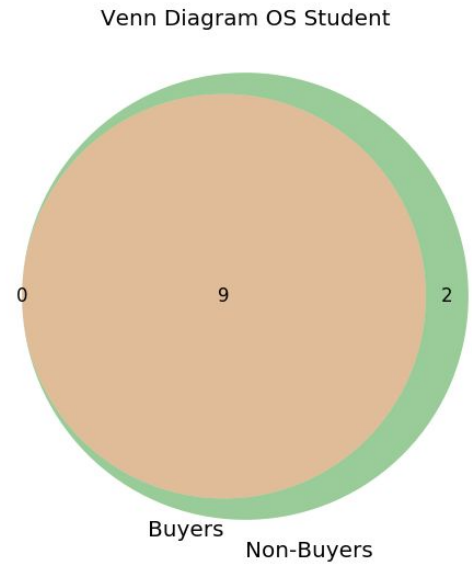
OS Teacher:
Windows XP 0.01%
Linux 0.0%
Windows 8 0.01%



Browser Student:
Firefox 41.0 0.0%
Firefox 42.0 0.01%
Chrome 45.0.2454.93 0.02%
Chrome 45.0.2454.98 0.01%
Chrome 45.0.2454.99 0.0%
Chrome 46.0.2490.71 0.0%



OS Student:
Windows Vista 0.0%
Windows XP 0.0%



From the venn diagrams, we cannot show a significant difference in the elements of the categories since the percentages are all below 0.02%. We do see that teachers and children that use old versions of browsers and operating systems don't buy the game. This is an indication that the game does not work well on old browsers and operating systems. When a customer would use an old version, they can be advised to use a newer version to have an optimal experience. This could reduce the amount of non-buyers. The countries of the customers that don't buy the game don't say much, since it is not significant. It is possible that the game is still new in those countries, which means that it is not bought yet.

We normalized as much features as we could over the trial period, but not all could be normalized. Some features do not have a date attribute which made it impossible to detect if it was in the trial period. To do this more precisely, we would suggest to add a date attribute to all the features in the database. A feature that shows the progress of the students would also be convenient to see if customers won't buy the game if the children already did enough/all exercises.