Data Mining Exercise 4, Question 2

In order to determine what market segments that NutrientH20 should focus on, we began by dropping all of the users that have been marked with spam, and then dropping the spam column altogether. Next, we found which users had more than 25% of their tweets relates to adult content so that any potential pornography bots that the original human annotators of the data may not have captured and deleted.

Due to the multiple dimensions amongst tweets from their followers, we used principal component analysis to reduce the dimensionality of the data. We then used hierarchical clustering on these principal components to find correlational clusters amongst their followers tweets in order to segment the market into different groups. In doing so we found 6 main marketing groups based on correlated interests. We believe that by focusing their attention on these six groups, NutrientH20 can direct their content in order to relate to these subgroups to further their sales.