HW 4 question 2

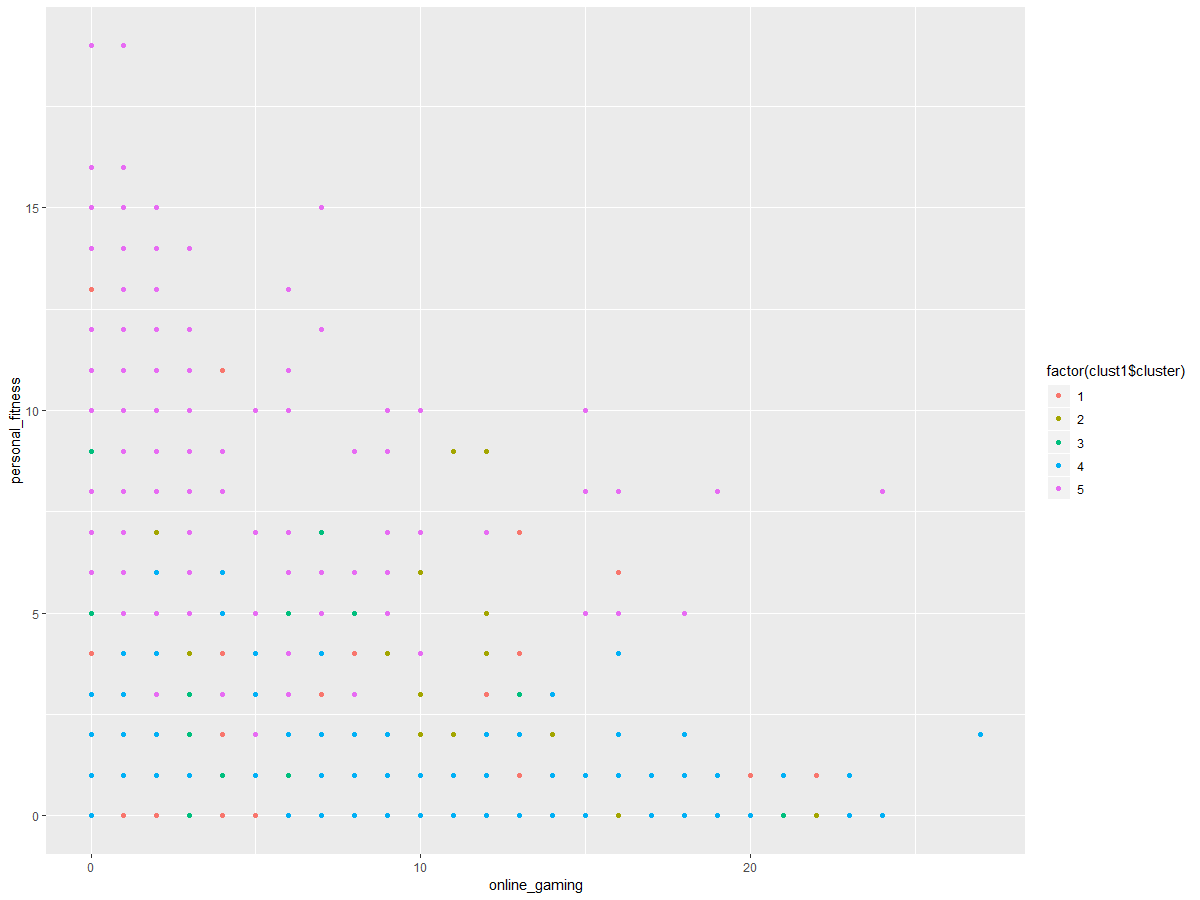
The goal of this question is to identify the market segments. We accomplished this goal in several steps

The first step is raw data processing. We think that “chatter”,” uncategorized” and “spam” are helpless in cluster, so we drop the columns. And, as reminded, we removed some users that maybe are bots.

The second step is center and scale the data and observe the characteristics of data. We established the correlation matrix between each label. We found that some labels have high correlation with some other labels, such as personal\_fitness and health\_nutrition. We predicted that such users pay more attention to health and may be grouped together. We have a general idea about how to market segment.

The third step is cluster. Even the optimal K seems to be 6 by the method, we still just want to classify these users into five different market types.

The fourth step. We run k-means with 5 clusters and 50 starts. Finally, we got 5 clusters, however, we do not know who is under the which market. Thus, we use several labels to test our clusters. For example, we use online\_gaming and personal\_fitness to drew the picture. We found that group 4 prefer online\_gaming and group 5 like personal\_fitness more.



We test some users’ data in group 4 and group 5, it seems that the clusters we get are reasonable.

Finally，we use PCA to get the labels in each cluster.

Group 1: "religion" "food" "parenting" "sports\_fandom" "school" "family" "beauty" "crafts" "cooking" "fashion

Group 2: "sports\_fandom" "religion" "parenting" "food" "school" "family" "automotive" "news" "crafts" "politics"

Group 3: "politics" "travel" "computers" "news" "automotive" "business" "small\_business" "college\_uni" "tv\_film" "online\_gaming"

Group 4 "health\_nutrition" "personal\_fitness" "outdoors" "politics""news" "eco" "computers" "travel" "food" "automotive"

Group 5 "beauty" "fashion" "cooking" "photo\_sharing" "shopping" "computers" "politics" "school" "travel" "business