This week's assignment (D10) deals with machine learning.

We would like you to write code that

1) Builds on all previous data analysis reports. So the data from D1 should already be loaded. We'll reuse it here.

2) Does a PCA for all 3 "psychological user characteristics" – **sensation seeking** (columns 401-420), **personality** (columns 421 – 464), **movie experience** (columns 465-474).

3) Do a median split of the ratings of the movie "Saw (2004)". Label ratings below the median as "0" and the ratings above the median as "1". In other words, discretize the ratings, where 0 means "enjoyed the movie less than average" and 1 means "enjoyed the movie more than average". As these are ratings data, there will be quite a few ratings right on the median. Don't label these at all – for the purposes of this exercise, you can treat ratings that correspond to the median as missing values.

4) Creates three separate logistic regression models predicting the discretized liking or disliking of the movie "Saw (2004)" (from step 3) from the scores on the first principal component of sensation seeking, personality and movie experience, respectively.

Hints:
*There are missing value in these psychological user characteristics, as some users didn't fill these in. The best way to handle this is to remove these row-wise (but per characteristic).
*Recall that unless one z-scores data before doing a PCA, the first principal component will simply point to the mean (among other problems). This yields misleading results and should be avoided.
*For the prediction question (only): Note that 289 users have rated the movie Alien (1979). Whereas it is valid to create the new scores/rotated data (the scores of the users in the direction of the principal components) from all users who have answered the full set of question on that characteristic, it is important to only use these data from the 289 users who have seen the movie to predict who will like the movie. So it is important to keep track of user identity. Note: The reason we use a random movie here (Alien) is that I want to illustrate the principle, without giving away the answer (if I used Saw as an example. Moreover, it is prudent to create general purpose methods that work for movies other than Saw).
*Make sure to do the PCA before considering the outcome (the Y, the saw score), in other words do it on the predictors (the X only). In other words, mix in the predictor later. Do the PCAs on the predictors first.
*Note: "rotatedData" from the CS code is the data in terms of their principal component score.
*When considering the movie data (for the prediction), you need to (3x) find the set (row-wise) of people who have jointly rated the movie and provided their personal characteristics, per characteristic (in other words, without missing values).
*Make sure that your logistic regression functions returns both betas and the p-values, not just (as is the default) beta$_0$.