Final exam (max 70 points)

The file `90returns.csv` contains returns of 90 most active stocks over 523 days (the full names are in file `mostactive.csv`). The presentation matters (imagine yourself as a data analyst presenting the data-driven decision on the optimal investment in front of venture capitalists/shark tank TV show).

1. (10 points). Project returns onto the line. Use `rug` and `density` to display the projections and their pdf. Use text with the options `srt=90` and `adj=0` to display the symbols. Report the variance explained by projection onto the line in the title.

2. (15 points). Validate the PCA ranking of the returns from the lowest to the highest using the mean of the returns over the entire period. Display the regression line and the symbols using the `text` command with the option `adj=0`. Use the `legend("bottomright"...)` with the value of the correlation coefficient rounded up to the second decimal. Show the vertical and horizontal lines with zero return. Print out the stock symbols, the full names of the companies, the value of the positive projection returns, and the mean of the returns (create a data frame and print it out).

3. (10 points). Project returns onto the plane. Use `rug` on both axes to show the projections. Display the symbols as in part 2. Display the % variance explained by the PCA projection onto the 1st and 2nd component at x- and y-axis, respectively, and onto plane at the title. Plot the PCA points using red color with `pch=16,cex=1.5` for which the mean of the returns is positive. Explain the result.

4. (10 points). Apply hierarchical clustering to the return data analyzed above. Does the dendrogram match the companies' standing from Problem 1? Give a qualitative answer.

5. (15 points). Find the optimal K using the broken-line algorithm. Compute the p-value for testing the hypothesis that there are no clusters (print out the observed `r` and the range of the simulated `r`s, use `nSim=1000`).

6. (10 poibts). Display the K-means clusters on the PCA projection onto plane as in Problem 3 (use different colors for different clusters), and commment of the difference with the positive returns.