EDA Homework 10

Due: Friday, April 14, 5pm

- The data frame `hamster` in the `lattice.RData` file has log-transformed weights of six hamster organs: the lung, heart, liver, spleen, kidney, and testes.

  Do a PCA of the organ weights and create a PCA biplot. Do you think it's useful to standardize the variables (`scale. = TRUE` in the `prcomp` function)? Does the PCA with the standardized variables tell you something different than the PCA with variables on the un-standardized variables?

- The data set `http://jfukuyama.github.io/teaching/stat670/assignments/nyt_articles.csv` has tf-idf-normalized word frequencies for the set of articles. Do a PCA on this matrix and create a plot giving the sample points on the principal plane, colored by article type.

  Does the PCA with the standardized variables tell you something different than the PCA with the un-standardized variables? Which one is more useful for this data set? What does it tell you about the articles?

  Note: You might have to remove some columns for `prcomp` to work.

- The loadings of the variables (words) on the principal axes are stored in the `rotation` slot in the output from `prcomp`. There are too many of these to visualize all of them at once, so we'll just look at a set of them that have the largest loadings on the principal axes. Plot the biplot axes corresponding to the variables with the largest loadings on the principal axes, and describe what the axes suggest about the differences betwen articles about art and articles about music.

  Note: Because you are plotting just a subset of the biplot axes, you won't be able to use `ggbiplot`.

What to submit:

- An R code file to reproduce your plots.
- A write-up containing your plots and responses to the questions.