CSC591: Foundations of Data Science
HW4-5-R-project: Combined R mini-projects for topics covered across HW4 and HW5.
This h/w accounts for **2%** of the grade.

Released: **11/11/16**
Due: **11/22/16 (23:55pm);** (One day late: -25%; -100% after that).

Student Name:
Student ID:

Notes

- Submit Single zip file; follow naming convention for all files.
- Zip file should include
    - One .pdf file, showing instructions on how to run R programs; libraries used, etc.; Also answer any question specific items (for example, if question asks you to submit a plot, then include it here).
    - Create one R file per question; should include all functions required to execute the file.

Mini R project (25 points)

R1. Validation (**10 points**): Implement Q5 from the [**RT**] book (under 5.4 exercises)

R2. (**15 points**) Implement PCA based Eigenface generation (see class slides). Use the data given in the faces-corrected.zip). Please note that you have to implement your own PCA function first (can't simply use existing library call). Name the file as PCA.R

(Briefly describe your steps and include top 10 Eigenface images; submit code separately).

R3. (10+5 = **15 points**) (To answer this question, you can use any 2-d data, real or simulated for implementation; we will provide test data one day before submission to answer part **b** of this question)

(**a**) (10 points) Implement G-Means (paper is provided under additional resources) (Algorithm 1, listed on page 3). (use attached template file "Gmeans.R" for your implementation). Please note that your code should work for any number of dimensions.

(**b**) (5 points) Generate 2-d plots (scatter plots and draw Gaussian distribution as ellipsoids) (test data will be provided one day before submission deadline), include these plots as part of h/w solution)

R4. (2+3+5 = **10 points**) Explore various R packages for constructing Bayesian networks. Answer the following questions:

(1) list the packages you found (not based on just web search; but install, and explore various examples provided with those packages), describe major functionality of each package, and highlight differences if any (2 points); and

(2) using any package of your choice, and given data (bn-data.csv), answer the following:

(a) (3 points) construct the Bayesian network and draw the resulting network, and

(b) (5 points) compute conditional probabilities for each node (submit resulting conditional probabilities as tables – for each attribute – please format properly so that it's easy to read).