**STAT 3280 Fall 2021 HW3**

**Due by the end of Oct 3, eastern time**. Submit your homework by sending it to our GTA, Ruizhong Miao (rm9dd@virginia.edu), with the subject "STAT 3280-HW3: names", where the "names" should be replaced by your last name(s) of the group. Each group only has to submit it once. Make sure you include everyone's **name AND computing id on the first page**. **Missing any part of these will result in missing grades**. Please use a separate page for each problem. And the answer to each problem cannot be longer than one page (with reasonable font size, line space, margins etc.). You can explain how you did it in R by submitting your code with detailed explanations, but only include this part in an appendix. The GTA will not be guaranteed to look at your appendix, so make sure you explains things clearly in your main text. Notice that you are working on a visualization task. So, for each problem, make sure that your **plain language explainations should not exceed 1/2 of the paper in total** for each problem. The main results would be your figures. **You can use any software or packages for this homework**

Total points: 10 + bonus points.

1. (4 pts) You worked on the data set of face images last time. The data set contains only 4 faces in HW2. It is a subset of a larger data set with 698 images. This time, we will work on the full data set and use PCA/MDS to deal with the data. In the file "Full_Faces_Data.Rda", there is a matrix object `face`, in the dimension of $698 \times 4096$. You can use the R command

   ```
   load("Full_Faces_Data.Rda‘")
   ```

   to load it into your R environment. Each row of the matrix is a 4096 dimensional vector, representing the $64 \times 64$ image, as you have tried last time. For this problem, first embed all the 698 images into a two-dimensional space by either PCA or MDS (up to your choice). After doing this, you will have a two-dimensional coordinate for each image. Let's call them Coord1 and Coord2. Now, calculate the min, 25% quantile, median, 75% quantile, and max of the Coord1 values, and locate which images correspond to these five values (if you have ties, pick any one with the same value). Now visualize these five images as you did last time, but arrange them in one row with the increasing order of their Coord1. The uniform variation pattern of them will give you a sense of what is the feature captured by your Coord1 (e.g. from facing-left to facing right, or from facing-down to facing-up, whatever pattern you could observe). Summarize the pattern. Then do the same thing, but this time focus on Coord2. (Visualization 1.5 pts for each of Coord1 and Coord 2. 0.5 pt for summarizing pattern in each of Coord1 and Coord2)

2. (6 pts) We would introduce the data set about diving competition judgement in class. The target for the analysis will be evaluating the nationality bias from the judges. Are the judges tend to be biased towards divers from their own countries? Who are more

biased in this setting? Answer these question and potentially explore other patterns by a one-page visualization results. Only **brief** descriptions by plain language are allowed. (2 pts for informative answer to each of the two questions. 2 pts for insightful analysis beyond the two questions.)

3. (7 bonus pts). **Challenge:** About the face data again. Still use the same 2-dimensional coordinates. Pick the first 20 of them (row 1 to row 20). Generate a figure similar to the figure below: For each of the 20 images, visualize the image at the same location as their two-dimensional coordinates, up to tiny rounding errors for the position. For this problem, you have to completely rely on R. You can use different packages, but cannot us other softwares. (Any successfully generated figures that satisfy the basic requirements would earn 2 bonus points and an invitation to presentation in class. The audience will wait for up to 5 bonus points for the presenter team.)