CAP 5768: Intro to Data Science, Fall 2023
Homework 2: Python and R

Due Sep 28 via Canvas by 6:00 PM. Submit file named: `YourLastName-HW2.pdf`

---

6. (15 points) Carefully study the Python notebook called `MovieLens1M.ipynb`. The data file required to run this notebook is in the `ml-1m` directory under `data` provided to you on Canvas. After you have understood the code in the notebook, modify the notebook to answer the following questions. You can do the work in Python or R. Submit the modified Python/R notebook.

   (a) Each movie in the database has a genre (`comedy, animations`, etc.) associated with it. Show the top 20 genres with the highest number of responses from users.

   (b) Show the top 20 genres sorted by average ratings.

   (c) Show the top 20 movies sorted by descending mean female ratings for a specific genre (say `Drama`).

7. (**Extra Credit**) If you submitted a Python solution for the above problem, then rewrite the code in R and submit the R notebook. (Obviously, if you submitted an R notebook for the above problem, then rewrite and submit a Python notebook.) Make sure your program runs properly and gives the same answer as the Python program it emulates. Submit the R notebook as a separate file. The name of all files you submit should include your name. For example, my files would be named `GiriNarasimhan HW2.pdf` and `GiriNarasimhan HW2.Rmd`

8. (10 points) In class, we discussed the MDCPS data set (see `5.3-MDCPS-Grades-2017.ipynb`). In class, we discussed clustering of the data set based on 4 features English Language Arts Achievement, Social Studies Achievement, Mathematics Achievement, Science Achievement (see `7.2-Clustering.ipynb`). Study K-Means options from:

   `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`. Redo the K-Means clustering by choosing (a) n init = 8, and (b) `elkan` algorithm instead of `lloyd`. Remember to set the seed to achieve reproducibility of your results.

9. (10 points) For the above problem, implement a reasonable quality measure to compare the different clusterings above. Display your results.

10. (10 points) In class, we discussed `6.1-PCA.ipynb`, where we ran PCA on the MDCPS data set. Follow it up by doing K-Means clustering on the points transformed by PCA (i.e., using `pca_0` and `pca_1`). State and explain your parameter choices. Display your results by coloring points according to their cluster.

11. (5 points) Redo the PCA experiment with the MDCPS data, but color the points according to `School Type`. Try it again, but coloring the points according to `Percent of Economically Disadvantaged Students`. Discuss your conclusions from what you observe with these different color schemes. Note `School Type` is a categorical variable, but the other one is a real number.