Intro to Data Science, Fall 2019
Homework 2: First Experience with Analyzing Data

Due Sep 15 at 11:59 PM

1. Let's revisit the Python notebook we discussed in class called `MovieLens1M.ipynb`. Create your own notebook in Python to answer the following questions:

   (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").

2. Install RStudio on your computer by downloading from `https://www.rstudio.com/products/rstudio/download/` and follow appropriate instructions. RStudio is a set of integrated tools designed to help you code and execute R programs. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. There is a very large number of tools writen in R that are accessible to you once you have RStudio installed. Learn how to install R packages.

3. Redo probem (1) above using R.

4. The EPA publishes air quality data on a continuous basis. It is possible to download data on fine particulate matter air pollution, also referred to as PM2.5, which refers to the amount of particulate matter that is smaller than 2.5 micrometers in the air. Higher this number, more is the pollution. The fields are called the following: *RD, Action.Code, State.Code, County.Code, Site.ID, Parameter, POC, Sample.Duration, Unit Method, Date, Start.Time, Sample.Value*. The air quality index is in the column titled *Sample.Value*. The two data files are called:

   `https://users.cs.fiu.edu/~giri/teach/5768/F18/RD_501_88101_1999-0.txt`

   and

   `https://users.cs.fiu.edu/~giri/teach/5768/F18/RD_501_88101_2012-0.txt`

   More details on the descriptions of the data can be found at:

   `https://aqs.epa.gov/aqsweb/airdata/FileFormats.html`. Create a notebook in Python or R to analyze how air quality index has improved from 1999 to 2012.