Cosette L. Hampton, Fall 2018
Assignment #6
MACS 30000, Dr. Evans
Due: 11/19/2018, 11:30am

1. Netflix Prize and Bell, Koren and Volinsky (2010) (3 points).
   (a) Contestants were expected to build a "collaborative filtering" model using 500,000 customers' ratings on 18,000 movies to predict a set of **3 million ratings** (Bell, et al., p. 1). With complex models it was important to reduce overfitting, especially where there was sparse data on ratings and the number of *dimensions* (or "subtle movie characteristics that affect users' ratings," Bell, et al., p. 27) was larger than 5. They were also judged on an extension that would allow predictive **parameters** (like a movie's popularity) **to change over time** (Bell, et al., p. 28).

   With all of these different models, Netflix would offer the prize to the team or participant that had, "…the greatest improvement in root mean squared error (RMSE) over Netflix's internal algorithm, Cinematch." (Bell, et al., p. 24) Netflix also required that the winning team have a **10% improvement over Cinematch**. When two teams had more than 10% improvement and they had very close **RMSE levels**, Netflix rounded RMSEs to four decimal places and used **submission time** as the tie-breaker.

   More judgement details from https://www.netflixprize.com/rules.html.
   Submissions had to qualify to be judged; qualifying was based on:
   - Algorithms providing "…predictions for all the withheld ratings for each customer/movie id pair in the qualifying set," and,
   - "…the RMSE of a Participant's submitted predictions on the test subset must be less than or equal to 90% of 0.9525, or **0.8572** (the "qualifying RMSE")."

   Additionally, other requirements for the winning submissions were that:
   - "The description of the algorithm must be written in English,"
   - It must be reproduceable by a computer scientist,
   - "It must describe substantially all the steps performed,"
   - "…it should also list any prior applications of the techniques employed by the algorithm," and finally,
   - "…it must provide an analysis of how the algorithm achieved its improvement over previous Progress Prize winners."

   (b) At the beginning of the Netflix Prize contest, the most commonly used method for predicting ratings (stars) on movies was called "nearest neighbors." (Bell, et al., p. 25) It essentially used a user's ratings to weigh similar movies/items and apply a rating (Bell, et al., 2010).

(c) According to Bell, et al., 2010, "Methods for optimal averaging of an ensemble based on limited information had a large impact on overall performance." The model that primarily worked to produce information about missing ratings and developed "extensions" to the main model was the matrix factorization model (Bell, et al., p. 27). The winning team's blend differed because each individual component did not need to share anything, and they did "nonlinear blends via neural networks," (Bell, et al., p. 29) and averaging two prediction sets always improved the better set "if it was not highly correlated with the other components." (Bell, et al., p. 28)

2. Collaborative problem solving: Project Euler (3 points).
   (a) Username: friend key –
   cosettelh: `1411036_vkojd7DAomEeT95sOyBYi8z8JY09TlzZ`
   (b) See .rmd file for R manipulation and .html file for markdown output. Done with problem #3, Largest prime factor.
   (c) I would like to win the "Gold Medal," "Ultimate Decimator," and "Valued Contributor" prizes. I like with gold medal the recognition that will come with being the first to solve a problem and the fact that this will be both about speed and innovation. I was choosing between Ultimate Decimator and Big Game Hunter because I wanted to show through problems solved that I was skilled, however I do not want to be unnecessarily skilled for my field but rather have a good grasp on various skill levels, so I think from 1-400 problems, the Ultimate Decimator prize shows that. Finally, I want a forum post award because it shows not only can I find a solution, but I can also explain how I found the solution to my problem or explain how to help others with their problems, and that takes more instructor-based and communication skill, as well as total grasp of the computational issue itself. I think 25 kudos shows that.
3. Human computation projects on Amazon Mechanical Turk (2 points).
   (a) Chose "Find Career Background of State Judges"
   (b) You receive $0.25 for completing information about one judge. You do more if you want there is no stated limit to how many times one person can complete this HIT, or how many judges are available.
   (c) I did not qualify to do this HIT, you are required to have a Masters. No other restrictions, qualifications or eligibility requirements are given.
   (d) You are allotted 30 minutes to complete the task. I could complete the task in 30 minutes. I think that I could find the requested information in 10 minutes, so I think I could do 6 of these tasks in one hour. The implied hourly rate is $0.50/hour.
   (e) The HIT expires in four days, or 11/22.
   (f) If 1 million people participated in the task, it would cost the HIT creator $250,000.
4. Kaggle open calls (2 points).
   (a) Done – profile is at https://www.kaggle.com/cosettelh
   (b) A current ongoing competition on Kaggle is one titled, "House Prices: Advanced Regression Techniques." Dean De Cock of Truman State University compiled the

dataset used for the competition, and it is considered a “getting started” competition, or one that was created by data scientists at Kaggle to engage people who are beginners in data science programming and analytics. Getting started exercises help new users get familiar with both Kaggle and other concepts in machine learning, and they are non-competitive in that they do not reward ranking points or prizes, and there is no end date or deadlines. Kaggle is a website that allows students, data scientists and those interested in machine learning to participate in learning-based competitions and more rigorous competitions for prizes like cash or publicity. It also helps users learn data analysis tools better and find data science related blogs, forums and even open job positions. The competition engages new machine learners to predict the sales price for each house in the given dataset. Submissions to this project are evaluated using the root mean squared error (RMSE), or the difference between the log of the predicted sales price for the house and the log of the observed sales price of the house.

There is no maximum team size and a team or individual can submit up to five entries per day, but only two final submissions to be judged. They ask that participants do not share information outside of teams and to not do this competition from multiple Kaggle accounts. As stated earlier this project has been open since 2016 and there is no entry or exit deadline.

(c) For this project there is no real winner, it is just judged by Kaggle data scientists on the RMSE, so they will likely give feedback to the team on how good their RMSE is, their technique to get there, and how they may improve if they choose to continue working on it. The data is public and older so others who do the project may use it for other datasets to make predictions about sales prices for homes, which may help with property tax calculations, gentrification/affordability estimates and other important housing-related issues.