1. Netflix Prize and Bell, Koren, and Volinsky (2010) (3 points).

a) The submissions to the Netflix Prize contest would be judged by the level of improvement over Netflix's movie recommendation system, Cinematch. And the mathematical way to calculate the improvement is RMSE (root mean squared error). I think if the improvement of the algorithm is less than 10%, the submission may be declined. Because the article states that "The movie rental company declared it would pay $1 million to the contestant who could improve its website's movie recommendation system by 10% or more. (Steve Lohr, The Contest That Shaped Careers and Inspired Research Papers)"

b) In the beginning of this contest, the most commonly used method is nearest neighbors. This method uses the weighted average of the users' evaluation of similar movies (this is defined as "neighbors") to make predictions. Here is the mathematical formula for calculating the rating:

$$\hat{r}_{ui} = \frac{\sum_{j \in N(i;u)} s_{ij} r_{uj}}{\sum_{j \in N(i;u)} s_{ij}}$$

c) I think it's the correlation between this model with others. If the correlation between them is lower, the better the effect of the two models will be combined. As the author mentioned, "Indeed, a hopelessly inferior method often improved a blend if it was not highly correlated with the other components. (Netflix Prize and Bell, Koren, and Volinsky, 2010)"

2. Collaborative problem solving: Project Euler (3 points).

a) My Project Euler user name and my friend key are:

SixueL: 1407604_K1sB72Jllh2AMC4AmjDLuOxAfpWCknrZ

b) I chose the problem "Sum Square Difference", and the answer is 25164150. Here is the code (Python):

```python
import numpy as np

a = np.arange(1, 101)
square = a * a
sum_of_square = sum(square)
square_of_sum = (a.sum()) ** 2
diff = square_of_sum - sum_of_square
print(diff)
```

c) All the rewards for solving the problems on Project Euler are small badges. Each of them has a unique design and meaning. The three awards that I would most aspire to achieving are "CC for Continued Commitment", "D for Dedication", and "Master of Archives". The reason that I like the first two badges is the picture with fruits and leaves looks so cute! Also, if I solve two-hundred and then five-hundred problems, I will enjoy the enthusiastic sense of accomplishment. I like the last one because it

is a symbol of the highest honor, which could motivate me to solve more problems using my programming and statistical skills.

3. Human computation projects on Amazon Mechanical Turk (2 points).

a) I chose the project with the title "Type the text from the images, carefully. Productivity and bonuses guaranteed."

b) The reward for doing this project once is $0.01. Also, you need to follow the instructions carefully, or your result won't be approved.

c) The instruction describes all the requirements for this project cautiously. For example, you need to pay attention to all the special symbols carefully. The instruction provides some of the symbols, and you can paste it. Also, if you encounter the situation with white spaces, you should only tape 1 space.

d) The allotted time for this task is 10 minutes. I think I could do 4 or 5 times in an hour. Therefore, the implied hourly rate for me is $0.04 or $0.05.

e) This job will expire at 05:36pm on 11/22/2018.

f) This job will cost the creator $10,000 if 1 million people will participate in the task.

4. Kaggle open calls (2 points).

a) My user name is Sixue and the link to my homepage is https://www.kaggle.com/sixueliu

b) This competition is called "Google Analytics Customer Revenue Prediction". In real business, only a small fraction of customers constitutes most of the revenue. Therefore, it is meaningful to develop different sales strategies based on customer segmentation. The sponsors of this competition are RStudio, Google Cloud, and Kaggle. RStudio is an open tool for R programming and also provides enterprise-ready professional software. Google Cloud is a platform which offers a suite of cloud computing services. The submissions will be judged by Root Mean Squared Error (RMSE). The mathematical formula is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

This formula characterizes the difference between predicted revenue and actual revenue. The award consists of two parts: the first one is "Leaderboard Prizes", and the other is "Special R Usage Prizes". Each award has three different levels of bonuses: which are $12,000, $8,000, and $5,000 and $10,000, $7,000, and $3,000. The first award is based on how close the algorithm design of the contestant is comparing to the predicted value. The second award is based on the top-ranking solutions using R. You can win the both awards in this competition. There are a few precautions that all the contestants should pay attention to. This competition doesn't allow submission of results from multiple accounts. One person can only

register under one account. You cannot communicate the code with people who are not in your group unless the code is made available to all participants. The deadline for registration is November 23, 2018. The final submission deadline is November 30, 2018. All the code will be tested using the real transaction data in December 1st, 2018 to January 31st ,2019. You may submit a maximum of 5 entries per day and may select up to 2 final submissions in this competition.

c) First, Google Merchandise Store will have a well-designed programming code to predict revenue per customer, which could help them to successfully analyze sales data, accurately predict diverse customer needs, and provide future sales strategies. Then, this will be a good incentive for those who are good at writing code with R. I am sure RStudio will attract more users if they continue to arouse people's interest in learning R language based on those competitions.