Intern/JDS Recruitment 3 Question Paper:

Total Points: 100

Instructions:

- Life will give you many options; but here, you'll have to answer all.
- You'll use R/python to accomplish the tasks
- This is an open-book exam.
- Second part requires you to compete in our **private Kaggle competition.** Invitation to enter the competition is given below.
- You need to send your code and results of part one in a pdf format. Markdown is highly encouraged.
- For part 2, you'll send your best .py/.R/.ipynb file for Kaggle modelling competition
- Submission deadline: 26th March
- Have faith. At the very least, you'll get your script back.

PART 1: (50 Points)

1. Optimization & Applied Calculus: (10 Points)

I. A local copy center needs to buy white paper and yellow paper. They can buy from three suppliers. Supplier 1 sells a package of 20 reams of white and 10 reams of yellow for $60. Supplier 2 sells a package of 10 reams of white and 10 reams of yellow for $40. Supplier 3 sells a package of 10 reams of white and 20 reams of yellow for $50. The copy center needs 350 reams of white and 400 reams of yellow. Using Python/R, determine (1) how many packages they should buy from each supplier in order to minimize cost and (2) the minimum cost.

II. Researchers have shown that the number of successive dry days that occur after a rainstorm for a particular region is a random variable that is distributed exponentially with a mean of 9 days. Using Python/R, determine the (separate) probabilities that 13 or more successive dry days occur after a rainstorm, and fewer than 2 dry days occur after a rainstorm. Round the probabilities to four decimal places.

III. Use Python/R to graph the function.

$$f(x) = -\frac{1}{(x+2)^2} + 4$$

2. Simple Statistics: (15 Points)

Use the data provided and construct the data frame "test". The data frame contains test results for 49 students on two standardized tests. Each student took both tests. Do not change the order of the two test entries or the matching per student will not be correct.

testA: 58,49.7,51.4,51.8,57.5,52.4,47.8,45.7,51.7,46,50.4,61.9,49.6,61.6,54,54.9,49.7,

47.9,59.8,52.3,48.4,49.1,53.7,48.4,47.6,50.8,58.2,59.8,42.7,47.8,51.4,50.9,49.4,

64.1,51.7,48.7,48.3,46.1,47.3,57.7,41.8,51.5,46.9,42,50.5,46.3,44,59.3,52.8

testB: 56.1,51.5,52.8,52.5,57.4,53.86,48.5,49.8,53.9,49.3,51.8,60,51.4,60.2,53.8,52,

49,49.7,59.9,51.2,51.6,49.3,53.8,50.7,50.8,49.8,59,56.6,47.7,47.2,50.9,53.3,

50.6,60.1,50.6,50,48.5,47.8,47.8,55.1,44.9,51.9,50.3,44.3,52,49,46.2,59,52

I. Determine a two-sided 95% confidence interval for the Pearson Correlation Coefficient of the data in "test". Present the code and the confidence interval for the Pearson Correlation Coefficient.

II. Use Bootstrapping for an estimated confidence interval. The process involves resampling with replacement of rows from "test." The first step is to randomly sample with replacement the 49 rows of "test". Each sample will consist of 49 rows for which a sample correlation coefficient is calculated. This step should be repeated 10,000 times resulting in 10,000 sample correlation coefficients. Find the 95% Percentile Bootstrap confidence interval. Set your seed at 123.

III. Plot two histogram side by side using results from I & II. Add density curves & report the confidence Intervals. State how bootstrapping results compare to the results from Part 1.

3. Creative Visualization: (15 Points)

From your DataShall R3 modelling competition, visualize the testfile dataset. **At least,** two visualizations is expected from you. (For example, you may choose to visualize the data from its original text format, after vectorization, or even the topics). There is no boundary in creativity. But don't make a dumb wordcloud. Show your code.

4. Programming Problem Solving: (10 Points)

Text = "Dude!!!! And I thought I knew a lotttt. Phewwwww! I won't back down. At least I understand now Daaata Science is much more than what we are taught in MOOOCs. That is allllright. I won't get demotivated. I'll work harder and in noooo time, I'll get better & be backkk next time."

Use Python/R to write a function that will remove the redundant letters **after the second consecutive letter** in the above variable. That means, ["goooood", "nicccce"] should turn into ["good", "nicce"]. 4 points for solving the problem & 6 points for your coding efficiency. Show your time complexity analysis.

PART 2: (50 points)

Kaggle Competition: DataShall R3

You must have a Kaggle account in order to participate in the competition.

Invitation: https://www.kaggle.com/t/ee437dabd037401dad83808b6a47c65e

You'll find competition details in Kaggle.

Welcome.

Team DataShall