CSE 40647/60647: Data Science	**Spring 2018**
Homework 1	
Handed Out: January 18, 2018	*Due: February 6, 2018 11:59 pm*

- This assignment is due at **11:59 PM** on the due date. Contact TA if you have technical difficulties in submitting it on **Sakai**. We shall NOT accept any late submission!
- Homework must be submitted in ZIP format (including .pdf and .py). Name your ZIP file as **YourNetid-HWx.zip**. Handwritten answers must be scanned into PDF.
  – YourNetid-HWx.zip
  —- YourNetid-HWx.pdf
  —- YourNetid-HWx-Qy.py
  —- ... (and any supplementary materials)
- Please use **Piazza** if you have any question about the homework.

Data set

Suppose we have $N = 1,000$ students who have taken both *Math* and *Data Science* classes. We sample $n = 9$ of them and list their names (fake as NBA player's names), *Math* scores, and *Data Science* scores as below. We will do some data processing in this homework on this sample data set. Good luck!

Student name	***Math* score**	***Data Science* score**
Giannis **A**ntetokounmpo	82	84
Kobe **B**ryant	98	97
Stephen **C**urry	83	83
Kevin **D**urant	95	97
Joel **E**mbiid	76	87
Markelle **F**ultz	71	73
Manu **G**inobili	81	83
James **H**arden	85	87
Brandon **I**ngram	76	83

1 Data Description (25 points)

1. Calculate *mean*, *median*, and *mode* of *Data Science* scores.
2. Calculate *variance* and *standard deviation* of *Data Science* scores.
3. We denote the i-th student's *Data Science* score as x_i ($1 \leq i \leq n$), denote the *mean* of these n scores as μ, and denote the *variance* as v. Suppose we sample one more student "Michael **J**ordan" whose *Data Science* score is x_{n+1}. Now we denote the new

mean (of the $n+1$ students′ *Data Science* scores) as μ' and the new variance as v'. Please write down the function

$$\mu' = f(\mu, n, x_{n+1}) \tag{1}$$

and the function

$$v' = g(v, \mu, n, x_{n+1}) \tag{2}$$

to incrementally calculate μ' and v'. Note that none of x_i ($1 \leq i \leq n$) or μ' is allowed to be used in the functions as input variable. You may assume $x_{n+1} = 100$ and use the given data points to verify if your functions are correct or not.

2 Data Visualization (35 points)

Use Python to generate the following two plots with the sample data set:

1. Q-Q plot. The X-axis is *Math* score. The Y-axis is *Data Science* score. Add a proper dashed line to answer the question: Which course is easier for the students, *Math* or *Data Science*?

2. Scatter plot. The X-axis is *Math* score. The Y-axis is *Data Science* score. Draw a *linear regression* dashed line to answer the question: Which student is more likely to be an outlier (farthest from the line)?

Please submit your code as **YourNetid-HW1-Q2.py**. Attach your figures and write down your answers in the PDF.

3 Data Reduction (35 points)

Suppose the matrix **X** (size: $n \times n$) below is the adjacency matrix of student-student social graph: $X_{i,j}$ is "1" if the two students are the same ($i = j$) or connected; "0" if they are different ($i \neq j$) and not connected. We consider the $n = 9$ students as data objects (rows) and as features (columns) themselves.

	A	B	C	D	E	F	G	H	I
Antetokounmpo	1	1	1	1	0	0	1	1	0
Bryant	1	1	1	1	0	0	1	1	0
Curry	1	1	1	1	0	0	1	1	0
Durant	1	1	1	1	0	0	1	1	0
Embiid	0	0	0	0	1	1	0	0	1
Fultz	0	0	0	0	1	1	0	0	1
Ginobili	1	1	1	1	0	0	1	1	0
Harden	1	1	1	1	0	0	1	1	0
Ingram	0	0	0	0	1	1	0	0	1

Use Python to call a Singular Value Decomposition (SVD) package and calculate *left singular vector* **U** (size: $n \times k$) and *singular values* λ_i ($i = 1 \ldots k$) where the number of singular values k is set as 2. The goal is to reduce the number of features from n to k.

Please submit your code as **YourNetid-HW1-Q3.py**. Write down in the PDF **U**, λ_i, and your observations on the two new features: Can you find the two student clusters?

4 Course Project: Teaming (5 points)

Who is/are your project partner(s)? List their names. Note that the number of names can only be 1–3. Please refer to the project policy: **The students may work in team of 2–4 (minimum 2 members are required) for the class project.**