Patrick Tinsley
Data Science - HW1
Due Date: 2/6

1 Data Description

Code can be found in ptinsley-HW1-Q1.py

1. Calculate mean, median, and mode of Data Science scores.

   Mean: 86.0, Median: 84.0, Mode: [83]

2. Calculate variance and standard deviation of Data Science scores.

   Variance: 55.5, Standard Deviation: 7.44983221287567

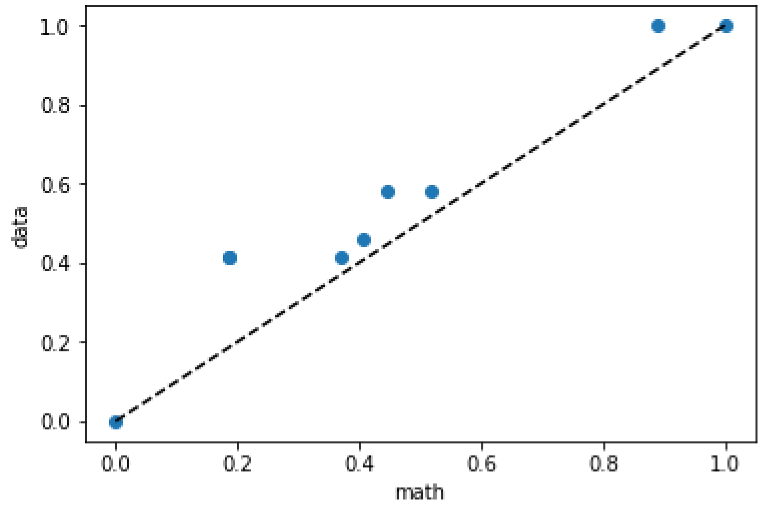
3. Please write down the function
   a. μ’ = f(μ, n, xn+1)
   b. v’ = g(v, μ, n, xn+1) (2)

   See ptinsley-HW1-Q1.py for code.

2 Data Visualization

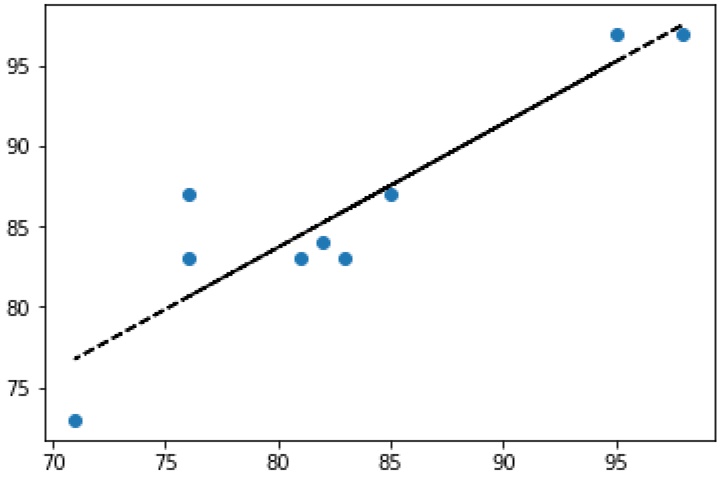
Code can be found in ptinsley-HW1-Q2.py

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?



Since the points line above the y=x line, the scores are higher in the data science class, which means data science is the easier class.

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)?

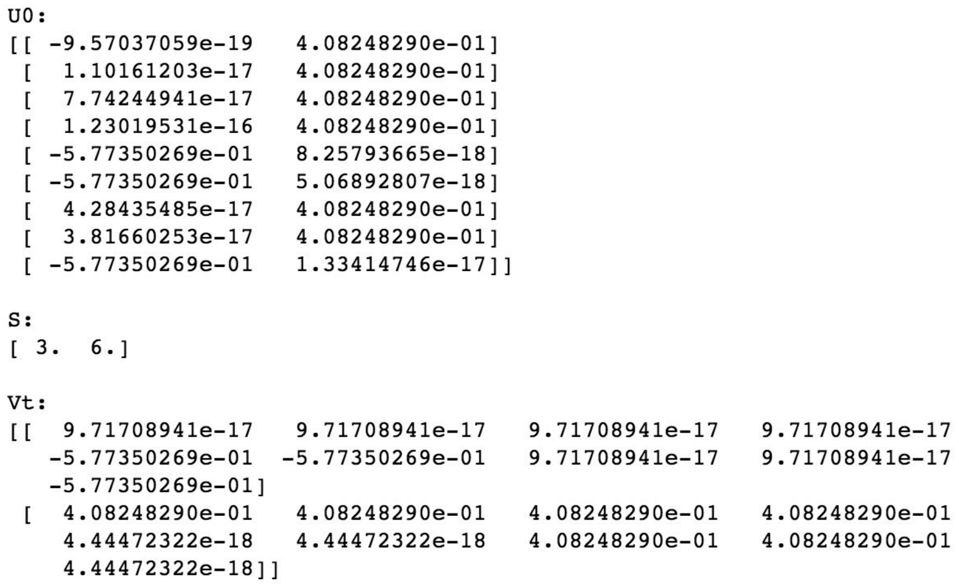


The furthest point appears to be located at x=76, y=87, which is associated with Joel Embiid. To verify, we can calculate residuals and find the maximum. After running the code in ptinsley-HW2-Q2.py, we see that our guess is indeed correct.

3 Data Reduction

Code can be found in ptinsley-HW1-Q3.py

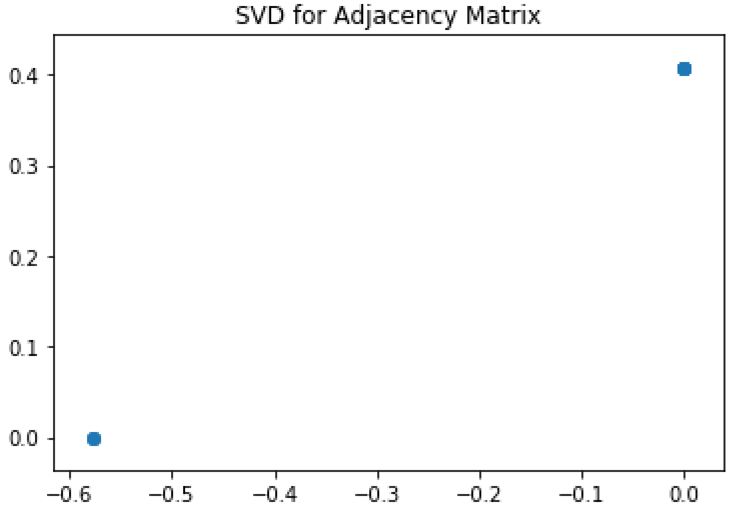
To generate the following figures from singular value decomposition, I used the linalg function from the scipy.sparse library. One can see that there are indeed k = 2 clusters. One cluster has **six** elements (or students), while the other has **three**; these are the singular values (λ_i) in the S matrix. In the U (or U0) matrix, we see the x- and y-coordinates that represent the data for each student in the lower, two-dimensional space; this is visualized in the x-y plot. It appears that the six students in the larger cluster have very small x-values (on the order of 10^{-17}) and “larger” y-values on the order of 10^{-1} while the three students in the smaller cluster have “larger” x-values (10^{-1}) and small y-values (10^{-17}). This can be seen in the cluster separation in the second plot.



```
U0:
[[ -9.57037059e-19   4.08248290e-01]
 [  1.10161203e-17   4.08248290e-01]
 [  7.74244941e-17   4.08248290e-01]
 [  1.23019531e-16   4.08248290e-01]
 [ -5.77350269e-01   8.25793665e-18]
 [ -5.77350269e-01   5.06892807e-18]
 [  4.28435485e-17   4.08248290e-01]
 [  3.81660253e-17   4.08248290e-01]
 [ -5.77350269e-01   1.33414746e-17]]

S:
[ 3.  6.]

Vt:
[[  9.71708941e-17   9.71708941e-17   9.71708941e-17   9.71708941e-17
   -5.77350269e-01  -5.77350269e-01   9.71708941e-17   9.71708941e-17
   -5.77350269e-01]
 [  4.08248290e-01   4.08248290e-01   4.08248290e-01   4.08248290e-01
    4.44472322e-18   4.44472322e-18   4.08248290e-01   4.08248290e-01
    4.44472322e-18]]
```



4 Course Project: Teaming

I will be working with Mark Giannini and Brian Tunnell, both graduate students in my program.