DNSC 6279
Quiz 02

Name:

1. **(1 pt.)** Name the R package that is currently considered to be the most scalable and efficient for data manipulation.

data.table

2. **(1 pt.) True or False**: Data manipulation using R and Python packages is typically as scalable and production-ready as data manipulation using Spark.

False

3. **(1 pt.) True or False**: Data manipulation using Base SAS is often considered production-ready simply because SAS sells numerous tools to productize Base SAS code.

True

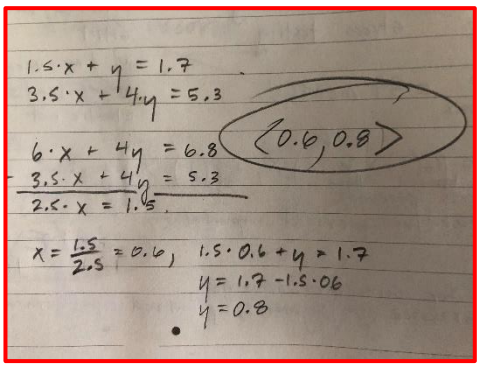
3.) **(3 pts.)** Consider the variable X below. X is very predictive and you would like to include it in a model, but it contains many categorical levels. Target-encode X into a numeric variable based on its per-level average with respect to the target Y.

X	Y	TE_X
A	1	2.5
B	2	3.5
C	3	4.5
A	4	2.5
B	5	3.5
C	6	4.5

(One point for each correct encoded level: 2.5, 3.5, 4.5.)

4. **(2 pts.)** Given the two variables and their first principal component values below, calculate the eigenvector from which the un-centered, un-scaled principal component values were derived. Show your work or describe how you completed this problem using your calculator.

X	Y	PC_1
1.5	1	1.7
2.5	3	3.9
3.5	4	5.3
4.5	5	6.7



(One point for each correct element of eigenvector.)

DNSC 6279 Name:
Quiz 02

5. **(2 pts.)** Discretize the column vector below such that all values less than the median are in a 'low' bin, all values including and above the median are in a 'high' bin, and missing values are included in the most frequent bin.

X	BIN_X
1	Low
2	Low
3	High
	High
5	High
1000000	High

(One point each for correct high labels and correct low labels.)