DNSC 6279 – Quiz 2

Name:

NetId/Email:

1.) **(2 pts.)** Consider the variable X1. Assume X1 is an interval variable and impute the missing values in the table with the mean value of X1.

X1
0
60
0
60
60

After performing the imputation, you are informed by a domain expert that X1 is in fact a binary variable. Now impute X1 with the mode (most common) value of X1.

X1
0
60
0
60
60

2.) **(3 pts.)** Consider the categorical variable X2 below. You would like to use X2 with a modeling algorithm that does not accept character inputs. Encode X2 into 3 binary, numeric variables with the values 0 or 1.

X2	X2_A	X2_B	X2_C
C			
A			
A			
B			
A			
B			
C			

3.) **(1 pt.)** Consider the variable X3 below. X3 is very predictive and you would like to include it in a model, but it contains several extreme values which may cause X3 to have undue influence on your model. Split X3 into 3 bins:

Bin A: X3 < 0
Bin B: 0 ≤ X3 < 1
Bin C: X3 > 1

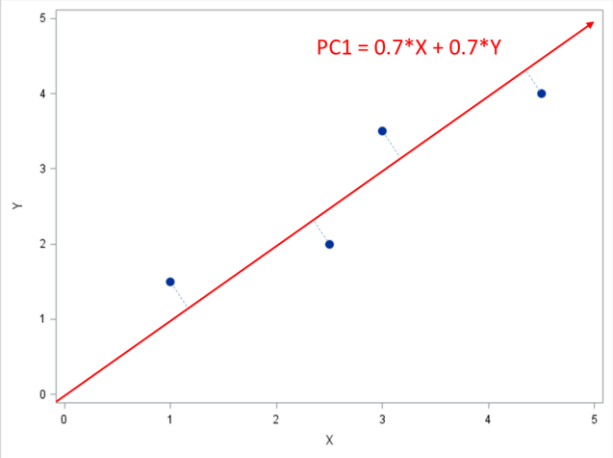
X3	Binned_X3
10347	
0.5	
0.7	
0.55	
-1.3	
100	
-0.1	

DNSC 6279 – Quiz 2

Name:

NetId/Email:

4.) **(4 pts.)** X and Y are highly correlated and can be accurately represented by a single principal component.



After running principal components analysis in your favorite software, you find the eigenvector that defines the direction of this principal component is (0.7, 0.7). Use this information to reduce X and Y into a single principal component and fill in the table below.

X	Y	PC1
1	1.5	
2.5	2	
3	3.5	
4.5	4	