Data Preprocessing Homework

Do not use a program to answer these questions. You can use a calculator.

1. Given the following data set which is part of a grade report, fill in missing values as you see appropriate and covert each column to have a mean of 0 and standard deviation of 1. Remove outliers if any and update the missing value.

Test-Ch3	Attendance	Class Activity	Test-Ch5	A3	Midterm Exam
7.5	25	24	5.5	92	79
9.5		46	8.5	106	92
9	25	28		85	
9.5	25	46	8.5	96	92
8.5	25	34	8.5	70	84
8	9	20	7.5	60	89
8	25		9	96	90
7.5		44	9	95	
10		46		90	95
9.5	25	48	7.5	90	
9.5	25	42	5.5	-5	78

2. Given the following dataset, use one hot encoding for the categorical data.

A6	Test-Ch8	Test-Ch10	Final Exam	Extra_Credit HW	
20	8.5	0	87	C	
93	9.5	9	103	A	
75	8.5	9	77	B	
88	10	8.5	NaN	A	
100	6.5	6.5	96	B	
0	0	6	43	F	

95	8.5	8.5	95	B	
20	9	7.5	97	A	
88	8.5	7.5	90	A	
60	10	9	77	B	
25	9	9.5	98	A	

3. Given the following dataset which is part of a grade report, use binning to put column named Test-Ch8 as a categorical variable and then convert it to numeric value using one hot encoding

A6	Test-Ch8	Test-Ch10	Final Exam	Extra_Credit HW
20	8.5	0	87	C
93	9.5	9	103	A
75	8.5	9	77	B
88	10	8.5	NaN	A
100	6.5	6.5	96	B
0	0	6	43	F
95	8.5	8.5	95	B
20	9	7.5	97	A
88	8.5	7.5	90	A
60	10	9	77	B
25	9	9.5	98	A

4. Given the following dataset, identify irrelevant columns by doing a univariate analysis between each column and target. The target is the weighted total.

Username	Test-Ch1	Test-Ch2	A2	Test-Ch3	Test-Ch5	Weighted Total
2208	9	8	85	7.5	5.5	78.00905
8434	10	8.5	100	9.5	8.5	97.14583
2179	8.5	10	88	9	8.5	81.37862
6096	9.5	9.5	95	9.5	8.5	94.45833
6779	9.5	7.5	90	8.5	8.5	86.2038
1220	9.5	8.5	80	8	7.5	58.62699
9384	8.5	8	85	8	9	86.86322
6355	8	8.5	90	7.5	9	91.58605
2298	9.5	10	93	10	8	93.14583

8193	8	7.5	75	9.5	7.5	80.80344
1552	9	8	70	9.5	5.5	84.96376