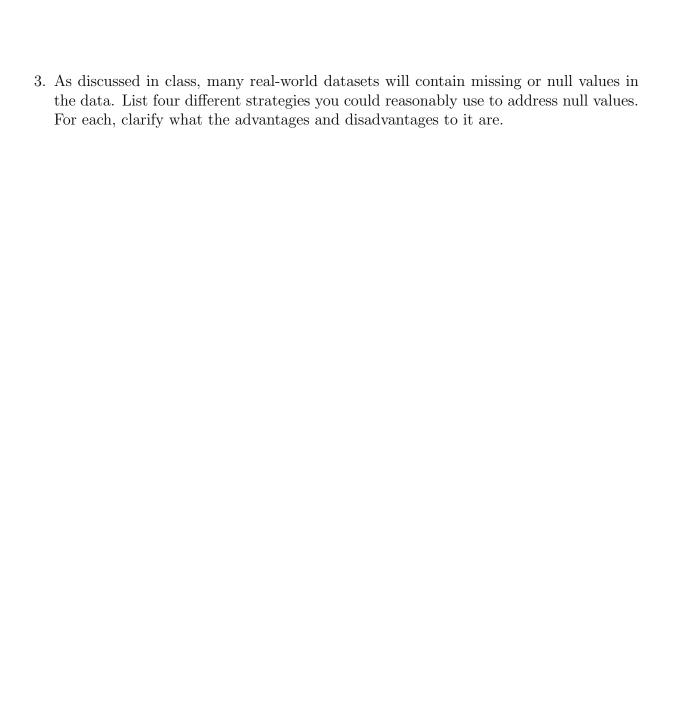
ECE M148 Introduction to Data Science Due: April 12, 12:00 PM Instructor: Lara Dolecek TAs: Harish GV, Jayanth Shreekumar

Please upload your homework to Gradescope by April 12, 12:00 PM. You can access Gradescope directly or using the link provided on BruinLearn. You may type your homework or scan your handwritten version. Make sure all the work is discernible.

Homework 1

1. Consider the following data set $A = \{1, 1, 5, 9, 9\}$. What are the mean and median of A? Now, consider $B = \{1, 1, 5, 9, 9, 11\}$. What are the mean and median of B? Using the mean and median, compare A and B.

- 2. In class, we discussed different ways to sample data. Explain in 1-2 sentences each the advantages and disadvantages of:
 - (a) Random sampling
 - (b) Stratified sampling
 - (c) Systematic sampling
 - (d) Cluster sampling



- 4. Consider the following sampling scenarios and determine which type of sampling bias is being demonstrated and explain your answer.
 - (a) Bob is a wealthy CEO who thinks taxes are too high. To confirm this hypothesis, he asks all his wealthy CEO friends their opinion.
 - (b) Sally is a teacher who wants to know how her class is performing. She sends out a survey with the following question: "Do you feel like you will get an A in the course or are you failing?"
 - (c) Constantine wants to know people's opinion about his website. He posts a survey link on his website asking for responses.

You may choose among the following options for the type of bias:

- i) Response Bias
- ii) Voluntary Bias
- iii) Convenience Bias
- iv) Under-coverage Bias
- v) Over-coverage Bias
- vi) Non-response bias

5. Perform KNN Regression on the following data set for different values of K: $(x, y) = \{(1, 1), (2, 4), (3.2, 6), (4, 3), (5, 2), (6, 2)\}$. Start by plotting the given points on a 2-D grid and then fitting a KNN regressor for the different values of K:

Make sure to draw the regression plot from 0 to 7.

- K = 1
- K = 2
- K = 3
- K = 6

Contrast and compare your findings over various choices of K. Is a larger K always better? Is K = 1 always better? Why or why not? Comment on what you think about the KNN performing regression on all x < 1.