Name: ______________________________

Homework 1: CSCI 347: Data Mining

Show your work. Include any code snippets you used to generate an answer, using comments in the code to clearly indicate which problem corresponds to which code.

1) [2 points] What are the two main types of attributes typically found in data?

2) [14 points] Consider the following data matrix D:

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.3 & 23 & 5.6 \\ x_2 & 0.4 & 1 & 5.2 \\ x_3 & 1.8 & 4 & 5.2 \\ x_4 & 6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.7 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

(A) [2 points] What is the sample mean of X_3?

(B) [2 points] What is the sample covariance between X_1 and X_3 ?

(C) [2 points] What is the (multivariate) sample mean $\hat{\mu}$ of the data set (your answer should be a vector)?

(D) [2 points] What is the sample variance $\hat{\sigma}_2^2$ of X_2?

(E) [2 points] What is the covariance matrix for this data?

(F) [2 points] What is the correlation between X_1 and X_3?

(G) [2 points] What is the total variance of D?

3) [6 points] Let **a** and **b** be two 4-dimensional vectors:

$a = (2,5, - 2.6,6)$ and $b = (15,2.5,4,4)$

(A) [2 points] What is $||a - b||_2$?

(B) [2 points] What is $||a - b||_1$?

(C) [2 points] What is the cosine of the angle between a and b?

4) [3 points] The following questions reference the *Heart Disease* data set from the UCI Machine Learning Repository:

https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Answer the following questions about the data set:

(A) [1 point] One attribute is named "cigs" What information is stored in the "cigs" attribute?

(B) [1 point] How many rows (entities/instances) are there in this data set?

(C) [1 point] How many attributes are there in this data set?