CS 345: Data Mining Names:_____

In-Class Activity 1 _____

_____

_____

1. In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem.

2. Suppose we have the following 2-D data set. Consider the data as 2-D data points. Given a new data point, x = (2.5, 2.7, 2.1) as a query, rank the database points based on similarity using Manhattan distance.

	A_1	A_2	A_3
X_1	0.5	2.6	2.9
X_2	2.0	1.9	4.3
X_3	1.2	2.1	6.1
X_4	0.7	3.4	2.3

3. Briefly explain how to compute the dissimilarity between objects described by the following:

    a. Nominal attributes:

    b. Numeric attributes:

    c. Asymmetric binary attributes:

4. The below table shows how many transactions containing beer and/or nuts among 10000 transactions. (roughly) calculate χ2 (chi-square) and tell me if beer and nuts are correlated.

$$\chi^2 = \Sigma \frac{(observed - expected)^2}{expected}$$

	Beer	No Beer	Total
Nuts	50	800	850
No Nuts	150	9000	9150
Total	200	9800	10000

5. Calculate the covariance of Economic Growth % and S & P 500 Returns % using the data table below. How are these attributes related?

Economic Growth %	**S & P 500 Returns %**
2.1	8
2.5	12
4.0	14
3.6	10