ASSIGNMENT 1

DUE DATE: Wednesday 22 September, 2010  11:59 PM

This assignment will contribute 7 points towards your final grade. The grading itself will be done out of 14 points distributed across three problems.

Problem 1 [5 points]:

Understanding population characteristics is an important step in any privacy setup. So, you will perform a minor analysis on the year 2000 census data for a few geographic regions in Colorado.

Download the data file (a comma-separated list) from the assignments page in the course website. This file has the number of individuals living in a few specific geographic areas of Colorado. The first row is a header row. The first column lists geography identifiers used by the Census Bureau. Every other column is for a specific age (between 10 years and 60 years, inclusive). Rows 2 to 6 correspond to males and rows 7 to 11 correspond to females.

If a cell has value x, it means that there are x individuals that have the same year of birth and gender, and live in the same geographic area. We can think of x as the level of anonymity for all those individuals since no two of them can be distinguished based on just their gender, year of birth and place where they live.

Your task will be to compute the 25th-percentile anonymity level corresponding to every age (i.e. age 10 years through age 60 years). The 25th-percentile anonymity level is a value such that 25% of the individuals have anonymity level less than or equal to that value. For e.g., if the value is 1500 for 20 years old individuals, then 25% of 20 years old individuals have an anonymity level less than or equal to 1500. You may use any software to compute the values (e.g. Excel, Matlab, etc.). Many of them already have built in functions to compute percentiles.

Create a scatterplot with the percentile values corresponding to every age. What do you observe in the plot? All I am looking for is an English interpretation of the plot. Look out for any deviant behaviors (such as a spike) and think of a probable cause for it.

Note: Your answer to this problem should just have the plot and your write up.
Problem 2 [4 points]:

Consider a data set with three attributes $A_1$, $A_2$ and $A_3$, modified as follows:

$A_1$: number of possible values = 4; generalized (but not completely suppressed)
$A_2$: number of possible values = 3; generalized (but not completely suppressed)
$A_3$: number of possible values = 2; suppressed

Assume that $N$ QI-groups are formed as a result of the modification, with sizes $k_1, ..., k_N$ as follows:

\[
\begin{array}{c|c|c}
\text{attributes} & \text{size } k_1 & \text{size } k_2 \\
\hline
\text{Original Data} & \text{size } k_1 & \text{size } k_2 \\
\hline
\text{records} & \cdot & \cdot \\
\hline
\text{size } k_N & \cdot & \cdot \\
\end{array}
\]

Let $k_{\text{min}} = \min(k_1, ..., k_N)$ and $k_{\text{max}} = \max(k_1, ..., k_N)$.

(a) (1 pt) What is the maximum number of QI-groups that you can have in any possible data set composed of the three attributes? Justify your answer.

(b) (1 pt) Assume that the attribute $A_3$ is not present in the data set. What will be $k_{\text{min}}$ and $k_{\text{max}}$ if $A_1$ and $A_2$ are generalized the same as before? Justify your answer.

(c) (2 pts) Say we add another attribute $A_4$ to the data set and generalize its values such that $A_4$ values in the first half of the records get modified to $m_1$ and those in the second half get modified to $m_2$. Explain under what condition will the previous QI-groups change.
Problem 3 [5 points]:

Consider the following 25 data points corresponding to some 2 attribute statistical database.

(a) (3 pts) Manually run the MDAV algorithm on these set of points and show the groups that are formed. Assume a minimum group size $k = 4$. Also mark the average record in each group.

(b) (2 pts) Can you do better? Create a grouping that is (or at least looks so) better than the one reported by MDAV. Mark the average record in each group and explain why you think your grouping is better.