Question 1:

The Census-Income (KDD) Data Set (cf., https://archive.ics.uci.edu/ml/datasets/Census-Income+ (KDD)) was originally extracted from the census bureau database found at http://www.census.gov/ . For your convenience, please find a copy of the data set (census-income.data.bz2) and its description (census-income.names) – as well as SQL scripts to load it into MonetDB (0-create.sql, 1-load.sql) – on the course website (https://homepages.cwi.nl/~manegold/DBDM/#Assignments); the SQL scripts should in principle also work for other DBMSs, possibly requiring minor syntax changes. This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables (dimensions), one of which is the “target” indicating whether or not a person earns more than 50k per year, plus a measurement called “instance weight” that indicates the number of people in the population that each record represents due to stratified sampling. For more information about the data set see above mentioned website and files. To answer the following questions, you can use MonetDB, another database management system of your choice, or any other software or programming/scripting language you are familiar with and prefer to use. Please report in detail which software you use, and submit all code/queries/scripts/programs you produce and use.

Consider the given data set as 41-d base cuboid of a data cube, with “instance weight” as measurement.
(a) Explain how many 1-d aggregate cuboids there are in total, and calculate how many cells each of them consists of.

(b) Using \texttt{sum()} as aggregation function, calculate the Apex \((0\text{-d aggregate cuboid})\), the smallest and the largest 1-d aggregate cuboid \((\text{excluding “target” as dimensions})\), and for each 1-d aggregate cuboid the respective 2-d aggregate cuboid by adding “target” as second dimension. Store each \(X\)-d aggregate cuboid as table with \(X+1\) columns \((X\text{ dimensions plus one aggregated measurement})\) in a separate CSV file. The records of the 1-d aggregate cuboids should be sorted on ascending dimension values. The records of the 2-d aggregate cuboids should be clustered on “target” and for each “target” value sorted on ascending dimension values.

**Question 2:**

Suppose that a base cuboid has three dimensions, \(A, B, C\), with the following number of cells: \(|A| = 1,000,000\), \(|B| = 100\), and \(|C| = 1000\). Suppose that each dimension is evenly partitioned into 10 portions for chunking.

(a) Assuming each dimension has only one level, draw the complete lattice of the cube.

(b) If each cube cell stores one measure with four bytes, what is the total size of the computed cube if the cube is \textit{dense}?

(c) State the order for computing the chunks in the cube that requires the least amount of space and compute the total amount of main memory space required for computing the 2-dimensional planes.

**Question 3:**

Assume that a 9-dimensional cuboid contains only three base cells: \((1)\ (a_1, d_2, d_3, d_4, \ldots, d_9),\ (2)\ (d_1, b_2, d_3, d_4, \ldots, d_9),\ and\ (3)\ (d_1, d_2, c_3, d_4, \ldots, d_9)\), where \(a_1 \neq d_1, b_2 \neq d_2,\) and \(c_3 \neq d_3\). The measure for the cube is \texttt{count()}. 

(a) How many non-empty cuboids will a full data cube contain?

(b) How many non-empty aggregate (i.e., non-base) cells will a full cube contain?

(c) How many non-empty aggregate cells will an iceberg cube contain if the condition of the iceberg cube is \texttt{count >= 2}? 

(d) A cell \(c\) is a \textit{closed cell} if there exists no cell \(d\) such that \(d\) is a specialization of cell \(c\) (i.e., \(d\) is obtained by replacing a * in \(c\) by a non-* value) and \(d\) has the same measure value as \(c\). A \textit{closed cube} is a data cube consisting of only closed cells. How many closed cells are in the full cube?

**Question 4:**

Consider a data cube \(C\) with \(D\) dimensions where each dimension has exactly \(V\) distinct values in the base cuboid. Assume there are no concept hierarchies associated with the dimensions.

(a) What is the \textit{minimum} number of (non-empty) cells possible in the base cuboid?

(b) What is the \textit{maximum} number of (non-empty) cells possible in the base cuboid?

(c) What is the \textit{minimum} number of (non-empty) cells possible in the entire data cube \(C\) \((including both base cells and aggregate cells)\)?

(d) What is the \textit{maximum} number of (non-empty) cells possible in the entire data cube \(C\) \((including both base cells and aggregate cells)\)?