1. Brief description of your data set (age of the data, number of observations and attributes)

Data is from the 1994 census database and was use for the training and testing of a machine learning algorithm to predict if citizens were earning greater than or less than $50,000 a year. There was a total of seventeen different algorithms tested and each one had an error rate of between 15% and 20%, one of the tests crashed for unknown reasons. There is a total of fifteen parameters, eight of which were categorical, the other seven were continuous variables. There is a total of 32,561 observations.

1. At least 1 yes-no or binary-choice question (Does it...? Is it...?)

Does it contain a satisfactory number of parameters to properly predict a person’s income? Yes and no. Firstly, I think it shows a large component of a person’s income—profession, years of education, relationship, race, sex. Whether or not those should to be parameters (ie, sex and race) is not in the scope of this summary. One parameter I think would be important is the income of parents and marital status of parent during childhood. I fully realize those are not included or generally relevant in census data, but I believe those to be important determining factors in the future success of individuals.

1. At least 1 non-binary question (What is...? How many...? When does...?)

This set has just over thirty-two thousand observations. In 1994 the United States has over 263 million citizens (and currently has over 325 million citizens), so this survey only has 0.012% of all population at the time. This is a very small sample size.

1. Future improvements to the data set: What other related data could be collected?

Building off of the prior questions, I think that it can have a larger sample size. In addition to that, region or state could also be added with the expanded survey size. I think that is important because the overall reason for this survey was to predict if a person earned over $50,000 a year. A salary of $50k is very regionally dependent—Loudoun Country, VA (highest median income at $115,574 a year) and Blackwater, AZ (lowest median income in the US at $9,491 a year) would yield very different results based on what occupation, age, etc., you have. With all that said, the vast differences within and the size of the United States might make it difficult to have one grandiose algorithm to estimate if a person earns more or less than a certain dollar value.

1. Source citation for your data set

The data was extracted from the census data bureau at http://www.census.gov/ftp/pub/DES/www/welcome.html donated by Ronny Kohavi and Barry Becker, Data Mining and Visualization, Silicon Graphics. Full dataset and other commentary analysis can be found at https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data.­­­