2.

* Why was this data collected?

Demographic census

* What does each record represent?

Each record is a list of information of an adult

* Where did the data originally come from?

This data was extracted from the census bureau database found at

<http://www.census.gov/ftp/pub/DES/www/welcome.html>

Extraction was done by Barry Becker from the 1994 Census database. (line 15)

* What is the principal question that our data mining task seeks to answer?

Prediction task – determine whether a person makes over 50k a year

* Are there other questions that we might be able to answer with this data?
* How will you know if you have mined useful data from it?
* How would you measure the effectiveness of a good analysis?

3.

* Describe the meaning and type of the data for each attribute

Age: numeric, continuous

Workclass: categorical, discrete

Fnlwgt: numeric, continuous

Education, categorical, discrete

Education-num: numeric, continuous

Marital-status: categorical, discrete

Occupation: categorical, discrete

Relationship: categorical, discrete

Race: categorical, discrete

Sex: categorical, binary

Capital-gain: numeric, continuous

Capital-loss: numeric, continuous

Hours-per-week: numeric, continuous

Native-country: categorical, discrete

>50k, <=50k: categorical, binary

* Verify data quality

Duplicate or conflicting instances: 6

Unknown values:

There are unknown values in workclass category, occupation category and native-country category that are converted to “?”. The unknown values should be removed before we conduct further calculations.

A set of reasonably clean records was extracted using the following conditions:

((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)) (line 16, 17)

* Provide appropriate basic statistics

For columns with numeric values, I calculated mean, median, mode, trimmed mean, min, max, range, std.

The mean and the median can measure the central tendency of the data. However, if there are outliers, the mean has a larger possibility of being affected by the outliers than the median. If the data are symmetric and there are not many outliers, the mean and median of one column in the dataframe should be similar. We can see from the calculation that the mean and median of the age columns are similar, so do fnlwgt, education-num and hours-per-week. However, there are huge differences between mean and median of column capital-gain and capital-loss, which means that the dataset of capital-gain and capital-loss may not be symmetric.

We can also see this problem from the mode. The modes of age, fnlwgt, education-num and hours-per-week are similar with their median and mean, but the modes of capital-gain and capital-loss fall far away from the mean and median, which means the distribution of capital-gain and capital-loss are skewed.

Trimmed mean can also tell us that problem. After removing 20% of the largest and smallest values, the means of capital-gain and capital-loss vary greatly.