1. Get the data

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 is real estate related to income?

How is educational experience related to income?

How the portion of investment income compared to total income for different stratum.

Is there any racial or reginal discrimination that influences educational level or income?

* How will you know if you have mined useful data from it?

We can try to grab some other data (both in domain and cross-domain) and see if the model we built on our dataset still works on these test data, a reasonable loss can be accepted, but a significant loss might suggest that our model is biased, either by the algorithm itself or by the sampling process.

* 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

## Description of fnlwgt (final weight)

The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau. We use 3 sets of controls. These are:

1. A single cell estimate of the population 16+ for each state.
2. Controls for Hispanic Origin by age and sex.
3. Controls by Race, age and sex.

We use all three sets of controls in our weighting program and "rake" through them 6 times so that by the end we come back to all the controls we used. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights. There is one important caveat to remember about this statement. That is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.[1]

. The continuous variable fnlwgt represents final weight, which is the number of units in the target population that the responding unit represents. [2]

Education, categorical, discrete

Education-num: numeric, continuous

The variable education\_num stands for the number of years of education in total, which is a continuous representation of the discrete variable education.[2]

Marital-status: categorical, discrete

Occupation: categorical, discrete

Relationship: categorical, discrete

The variable relationship represents the responding unit’s role in the family.[2]

Race: categorical, discrete

Sex: categorical, binary

Capital-gain: numeric, continuous

Capital-loss: numeric, continuous

The variable relationship represents the responding unit’s role in the family. capital\_gain and capital\_loss are income from investment sources other than wage/salary.[2]

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.

Standard deviation is a number used to tell how measurements for a group are spread out from the mean. A low standard deviation means that most of the numbers are close to the mean, and vice versa. The data greater than the mean plus 3 times std or the data less than the mean minus 3 times std may be the error data we should give another look at. We could see that capital-gain and capital-loss has a relative big std, which may represents the variance of the data is large.

* Visualize the most important or interesting attributes

1. Plot the histogram of the capital gain and capital loss. We can see from the histogram that almost all people’s capital gain and loss are 0, which means there’re no gains or losses that an individual experiences on the sale of a capital asset.
2. Plot the pie chart of races and workclasses. The dominate races is white which consists 86% of the people and the dominate workclass is private which consists 73.9% of the people.
3. Plot the boxplot of ages in different workclasses. We can see that the range, quatiles and median of age differs from workclass to workclass. People in their younger age may be more willing to work without pay.

* Explore the relationships among the attributes, excluding the class attribute

We use the correlation matrix. Each element in the correlation matrix is a correlation coefficient r which measures the strength and direction of a linear relationship between the row and column variables. The value of r is always between +1 and -1.

We can see a positive linear relationship between the education and income, which means a person may receive higher income with higher education level. There a slightly weak positive linear relationship between age and income, which means a person may receive higher income when he grows older. There are other attributes which has the positive linear relationship with income, including capital-gain, hours-per-week and loss. Fnlwgt has a very weak negative linear relationship, which is too small that we can say they don’t relate very much.

* Identify and explain any interesting relationships between the class attribute and the other attributes.

Source:

[1] Kaggle adult census income dataset. Last access: Sept. 2019. url: <https://www.kaggle.com/uciml/adult-census-income>

[2] Haojun Zhu, Predicting Earning Potential using the Adult Dataset. Dec. 2016. url: <https://rstudio-pubs-static.s3.amazonaws.com/235617_51e06fa6c43b47d1b6daca2523b2f9e4.html>