# Get the data

The data was fetched from adult.csv from <https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data> into a pandas dataFrame

# 2. Understand the data in context

* 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

* 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?

The analysis should have internal consistency. When we use different sets of data in the same area, the analysis should give us similar results. In that case, we can find some good quality data sets in the same area as test cases for our analysis and see if they have the similar results.

# 3. Understand the data

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

1. **age**: numeric, continuous.
2. **workclass**: categorical, discrete.
3. **fnlwgt**: numeric, continuous
4. **education**: ordinal, discrete
5. **education-num**: numeric, continuous
6. **marital-status**: categorical, discrete
7. **occupation**: categorical, discrete
8. **relationship**: categorical, discrete
9. **race**: categorical, discrete
10. **sex**: categorical, binary
11. **capital-gain**: numeric, continuous
12. **capital-loss**: numeric, continuous
13. **hours-per-week**: numeric, continuous
14. **native-country**: categorical, discrete
15. **>50k**: categorical, binary

*3.1 subsection: explanation for non-self-explanatory attributes*

* **fnlwgt**: represents final weight. according to [2], it's the number of units that this record could represent in the target population. In [1], the author explained that the final weight is controlled by three factors: a single cell estimatation of the population 16+ for each state; controls for Hispanic Origin by age and sex; controls by race, age and sex.
* **education\_num**: represents the number of years of education in total.[2]
* **relationship**: represents the individual's role in it's family.[2]
* **capital\_gain** and **capital\_loss**: represents the income and loss from non-salary ways, e.g. investment.[2]
* **education**: due to lack of domain knowledge in U.S. education system, we decided to refer to [3]. In which they order the education level in the following manner:  
  Preschool < 1st-4th < 5th-6th < 7th-8th < 9th < 10th < 11th < 12th < HS-grad < Prof-school < Assoc-acdm < Assoc-voc < Some-college < Bachelors < Masters < Doctorate.  
  [3] further merges "Preschool", "1st-4th", "5th-6th", "7th-8th", "9th", "10th", "11th" and "12th" groups to "dropout" group, "Assoc-acdm" and "Assoc-voc" groups to "Associates" group,“HS-grad” and “Some-college” groups to “HS-Graduate” group.  
  For us, we would adopt this strategy, merging Prof-school into HS-Graduate group, and instead of arranging the education level like shown above, we have the following order:  
  dropout < hs-graduate < associates < bachelor < master < doctorate.
* Verify data quality
  1. Duplicate or conflicting instances

We checked and dropped the duplication via pandas built-in function.

There are 24 duplicated data entries found.(One copy of the data entry will be kept in the dataset) After dropped duplicate, there are 32537 data entries remains.

* 1. 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. After dropped the missing values, there are 30162 data entries remains.

* 1. Outliers

To find outlier, we only focus on numeric values. As we can see in the following, the calculated lower bound and upper bound for both capital\_gain and capital loss is both 0.

In 3.3 section we will see that the median for these two value are 0 either. Which means in this dataset most of the people doesn't do investment and thus investment does not incur gain or loss to their income. So for this two attribute we just don't do anything to the outliers

* Provide appropriate basic statistics

Before cacluate the statistics, we set income > 50K to 1 and income <= 50K to 0 to better interpret the relations between the numeric values and the income.

For columns with numeric values, we 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 workclasses. 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.
4. Plot the histogram with KDE for age. We can see that the age is almost equally distributed compared to the capital-loss and capital-gain. It is just a little skew to left. The variation is still significant. We can see the frequency varies from age to age.
5. The previous graph shows the feature importance calculated by a random forest classifier.  
   As we can see from this graph, unlike our domain knowledge, fnlwgt plays the most significant role in predicting final income. Also age plays an important part inside this. Capital-gain, workclass and capital loss plays non-trival role in predicting final income, but they are not the most important features. I personally infer that this might because we are modeling the classificaiton problem as a binary classification with threshold at 50K, if we have multiple threshold or the threshold is set to a higher value, investment related income might play a more important role in this.

* 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

One thing we discovered about how relationship in family is related with income through cross tabulations. As we can see fromthe following table, class husband and wife tends to have a much higher portion of high income individuals compared to other portion. Perhaps income is one of important things that support family.

We group the statstics by the workclass. We can see the statistics varies between different work classes. We can see that in self-emp-inc workclass, the mean of the capital-gain of that workclass is the highest but the median is still 0. This can be further reflected from the standard deviation. The statistic tells us that the self-emp-inc has the high profit but also has a high risk.

* Attributes could add to this dataset

We could divide ages into different groups. Ages which only differs for 1 or 2 years may not be a striking difference but ages from 30-40 and ages from 50-60 may seem more representative and explainable.

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>