# Using Pandas for data preprocessing

## Dataset Overview:

The adult dataset contains data extracted from the (U.S.) census bureau. It contains approx. 49,000 records of census information taken in 1994 from many diverse demographics. The dataset is made up of the following fields.

1. **age**: continuous

2. **workclass**: 8 values [Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.]

3. **fnlwgt**: continuous. The \# of people the census takers believe that observation represents. We will be ignoring this attribute.

4. **education**: 16 values The highest level of education achieved for that individual [Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.]

5. **educational-num**: continuous. Highest level of education in numerical form.

6. **marital-status**: 7 values [Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.]

7. **occupation**: 14 values [Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces]

8. **relationship**: 6 values. Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. We will be ignoring this attribute. We will be ignoring this attribute.

9. **race**: 5 values. [White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.]

10. **gender**: Male, Female

11. **capital-gain**: continuous. We will be ignoring this attribute

12. **capital-loss**: continuous. We will be ignoring this attribute

13. **hours-per-week**: continuous. Hours worked per week.

14. **native-country**: (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland,

France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.)

15**. income**: Yes, No. Whether or not the person makes more than $50,000 per annum income. Either ">50K" or "<=50K".

The table below shows the relationship between educ_num and education

| Educ_num | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* |
|---|---|---|---|---|---|---|---|---|---|---|
| Education | Preschool | 1st-4th | 5th-6th | 7th-8th | 9th | 10th | 11th | 12th | HS-grad | Some-college |
| Educ_num | *11* | *12* | *13* | *14* | *15* | *16* | | | | |
| Education | Assoc-voc | Assoc-acdm | Bachelors | Masters | Prof-school | Doctorate | | | | |

## Exercises:

1. Missing values in the adult data are presented as the string '?'.
   a. Using an replace function convert any '?' to a np.nan
   b. Now determine the number of missing values in the dataset
   c. Drop any rows containing missing values and verify that all missing values have been deleted.

2. Using the groupby function divide the data into groups based on gender.
   a. Calculate the mean educational-num value for each subgroup

| gender | |
|---|---|
| Female | 10.105886 |
| Male | 10.124513 |

   b. We are interested in looking at the gender balance between those earning greater than 50K and those earning less than 50K in each of the subgroups.
   Using a for loop iterate through each of the groups and print the normalized number of those in the current group earning more than 50K and those earning less than or equal to 50K (value_counts may be useful).

```
Group  Female
<=50K    0.886424
>50K     0.113576
Name: income, dtype: float64

Group  Male
<=50K    0.687523
>50K     0.312477
Name: income, dtype: float64
```

3. Write a function that will take in as a parameter the adult DataFrame and a String label (for income). You can assume that the value of this String parameter will be either ">50K" or "<=50K". In this question we want to determine the percentage of people that earn either greater than 50K or less than 50K (depending on the value of the String parameter) for each educational number.

   For example, the following is the output if we call the function and pass in >50K. A total of 1.3 of all respondents that achieved education num value of 1 earn greater than 50K. In contrast 73.3% of those with an educational num value of 16 earn greater than 50K.

```
For those with Income >50K
1.3888888888888888  have an educational num value of  1
3.6036036036036037  have an educational num value of  2
4.8997772828507795  have an educational num value of  3
6.682867557715674  have an educational num value of  4
5.621301775147929  have an educational num value of  5
6.704824202780049  have an educational num value of  6
5.497220506485485  have an educational num value of  7
7.452339688041595  have an educational num value of  8
16.343096800378813  have an educational num value of  9
20.10304071118295  have an educational num value of  10
25.727411944869832  have an educational num value of  11
26.410086264100862  have an educational num value of  12
41.981505944517835  have an educational num value of  13
55.40970564836913  have an educational num value of  14
75.4140127388535  have an educational num value of  15
73.34558823529412  have an educational num value of  16
```

4. Determine the median value for the educational-num column in the dataset. Next use the apply function to convert the educational-num column into a binary column. You should populate each entry in the column with the value True or False: if the existing value is less than the median then you should enter True, otherwise False. The new column (Series) should replace the existing educational-num column.

   Next we want to group based on the value of the new educational-num feature and determine the distribution of the income value for respondents in each group. In other words we are interested in seeing the distribution of income for the group that has achieved a lower level of education compared to the group that has attained a higher level of education. In the result below, the group labelled False are those that have attained the higher standard of education.

```
Group  False
<=50K    0.65881
>50K     0.34119
Name: income, dtype: float64

Group  True
<=50K    0.865291
>50K     0.134709
Name: income, dtype: float64
```