Analytics report of data set

Analysis of continuous

From all the histograms, we observed the following suspicious values ;

- **capital-gain** : 154 around 100k and lot of 0 → For the amount around 100k, it can be values that represent big amounts out of the limit of the histogram. The lot of amount of 0 can be suspicious but it can represent a lot of society without capital gain.
- **fnlwgt** : 4 over 1.2M → There is a lot of values between 0M and 0.8M and very few values after 1M. These values can be suspicious but they can represent some society with a big fnlwgt too.
- **age** : lot of 90 → For the ages, the maximum age (90) has a high value compared to the nearly other values. It can be value that represent the age of people out of the limit of the histogram.
- **capital** : Nothing really suspicious but we noticed much company with their capital reach 0 per years. It should be verify but can be normal.
- **hours per week** : We observed that a lot of people work around 40 hours per week and it can be strange that it's not that much distributed but it can be normal. However there is more than 600 person who work more than 70 hour per week and it represent a lot of work for a person so it looks like outlier values. In addition we have 82 person around 99 hours per weeks. It can be values that represent big amounts out of the limit of the histogram.

For all the outliers data we could delete the values to not biaise the histogram and the analyse. For all the data with a big population whe should check if it's due to wrong values in the data set or if it's just a normal dispatching.

Analysis of categorical

With the different bar plots we've noticed only two suspicious data. First in the **workclass** feature, there is a lot of unknown data because of unspecified data. The second is similar to **workclass** but for the **occupation** feature.

To correct those missing values, we suggest two solutions:

- The first is to dispatch equitably the number of missing values between all others values to dissimulate those one.
- The second solution is to dispatch proportionally the missing values according to the others values' rates.

Feature cardinality

We don't see anything suspicious about the cardinality of continuous and categorical values.