**Missing Values**

For the categorical features we have missing values that are marked as ‘?’. There are 1735 question marks in the workclass column, 1742 in the occupation and 554 in the native-country column. The dataset shows that, where workclass is missing, the occupation will be missing as well. These can be considered as unemployed since there isn’t such a class.

For native-country, there are 554 ‘?’. Therefore, it will be left as it is and documented in the data quality report. In the capital gain and capital loss columns approximately 28500 values are 0 therefore these columns will be discarded.

**Outliers**

From examining the modes and second modes of the categorical features, the following can be deducted. Most of the subjects work privately, are white, males, and live in the United-States. More than half of the subjects has not finished a form of university. In case of the marital status they are either married or never-married.

In the dataset, there are only 600 people that are between 70 and 90 and they could possibly be outliers since the mean is 38 and the standard deviation is 13.6 for age.

People with a fnlwgt above 601253 are potential outliers since there is less than 200 of them. Also, people between the minimum fnlwgt(12285) and 25000 are potential outliers.

People that are working more than 79 hours are scarce and might skew the model.