Data Quality Report Assignment

Machine Learning - April 2018

Team A: Simon Bonnaud, Arthur Chevallier, Anaïs Pignet, Eliott Vincent

A. Analysis of the dataset

a. Missing values

For each continuous feature, we have a miss percentage of 0.0%. However, if we take a look at our dataset, we observe that for certain features, such as the "capital-gain" one, there are some weird values (0; 99,999): we can imagine they aren't properly set in the original dataset.
Moreover, if take a look at the computed statistics for continuous features, we can notice that the median for the "capital gain" feature is equal to 0. This means that we have at least $\frac{30940}{2}$ or 15 470 values equals to 0. If we take a look at the third quartile, we can make a similar conclusion: 23 205 values, at least, are equals to 0. The maximum value, evaluated as 99,999 may look suspicious, according to our previous conclusion.

For categorical features, we can notice that we have some missing values (for "workclass", "occupation", and "native country" features). Those missing values correspond to the values set to "?" into the initial Dataset file.

For both features, we observe that there are no miss percentages higher than 60 %. This means that we don't have to remove any feature from our ABT.

b. Outliers

Regarding outliers, only continuous features have several issues. We can sum them up in the table below :

Feature name	Outliers	Observation
Capital gain	Unusual Maximum value and gap between maximum value and third quartile value	Max value : 99999, 3rd quartile : 0, Median : 0 Max - 3rd quartile = 99999 >> 3rd quartile - median = 0 So as a result of this calcul, the maximum value is unusual and is **likely** to be a valid outlier. Min value : 0, 1st quartile : 0, Median : 0 median - 1st quartile = 0 == 1st quartile - min = 0 So as a result of this calcul, the maximum value is unusual but is **unlikely** to be a valid outlier The maximum value is 99,999. It looks like a default value for a non specified input but we don't have enough information to make a conclusion on this. Gap: The gap between third quartile and the maximum is 99,999. Moreover, the third quartile value is still 0.0. It looks like lot of values aren't properly set here.
Capital loss	Gap between maximum value and third quartile value	Max value : 4356, 3rd quartile : 0, Median : 0 Max - 3rd quartile = 4356 >> 3rd quartile - median = 0 So as a result of this calcul, the maximum value is unusual and is **likely** to be a valid outlier.

		The gap between third quartile and the maximum is 4,356. The third quartile value is still 0.0. It looks like lot of values aren’t properly set here.
Hours-per-week	Maximum value	The maximum value is 99. 99 hours per week seems to be a lot of work hours in a week. It may look as a default input.
Fnlgwt	Gap between maximum value and third quartile value	Max value : 1484705, 3rd quartile : 237318, Median : 178384 Max - 3rd quartile = 1247387 >> 3rd quartile - median = 58934 So as a result of this calcul, the maximum value is unusual and is **likely** to be a valid outlier. The gap is really high but this feature describe a special value that can vary a lot and the maximum value isn’t suspicious.
Age	Unusual maximum value	Just like previous feature, we can see a gap between maximum and 3rd quartile values. However, 90 seems like a fine number for an age so it looks like a valid outlier.

c. Feature cardinality

For both continuous and categorical features, we can notice that we don’t have any cardinality of 1. We can now look at cardinality values for continuous and categorical features:

Continuous features

Feature name	**Cardinality**	**Observation**
Age	72	Even if the cardinality seems low, it should not be considered as suspicious: people may have the same age.
FNLWGT	20880	This value is close to the number of instances in the dataset.
Education num	16	Even if the cardinality seems low, it should not be considered as suspicious: in reality, we don’t have 30 940 different education levels.

Capital gain	119	This value is far from the number of instances in the dataset. The capital gain should differ from people to people. We may have some errors in this features.
Capital loss	91	This value is far from the number of instances in the dataset. The capital loss should differ from people to people. We may have some errors in this feature.
Hours per week	93	Even if the cardinality seems low, it should not be considered as suspicious: in reality, we may have a standard for working hours. It may even look a little bit too much for working-hours.

For categorical features, we don't seem to have features with a higher cardinality that we would expect. This means that we did not use multiple values to represent the same value.

d. How to address quality issues

To handle the above quality issues, we can either:

- drop any feature that have missing value
- create a missing indicator feature from features that have missing feature
- apply complete case analysis
- imputation

The below table describes which approach best fits each data quality issue:

Data Quality Issue	Approach
Capital-gain	This feature seems to have a lot of weird values (0 and 99,999). Around 28K of the values are equal to 0. It seems irrelevant in our context: we should update the dataset with more coherent values. We could also interpret these values as missing values, and then create an indicator “missing capital gain”, interpreted as no capital gain.
Capital-loss	In this feature, around 29,5K of values are equal to 0. We may consider them as missing values, and then create an indicator “missing capital loss” interpreted as no capital loss.
Hours per week	The maximum value for this feature is 99. It looks irrelevant, as the average hours worked in the U.S. is around 33. We should **clamp down** the “99” values to a lower value (e.g. 40).