TP Data Quality

PART I

You are given a dataset (jobs.csv) that contains data about doing different kind of works on apartments and houses of people (repairs, changes, etc.). The data that exists contains various errors of different types. Some errors are easy to spot even visually (e.g. having dates in the wrong format or dates and times in some places and dates only to others and so on). Based on the discussions we had in class you need to write some code that finds and reports those problems. Some of the problems can be checked for all columns (e.g. missing values), some are column specific and depend on the type of the data involved.

More precisely:

(1) **Check for missing values**: we should check for NULL values or empty strings. If the number of NULL values exceed the 80% of all values then the column should be reported as problematic.

(2) **Check for outliers**: outliers are values that are outside the normal value ranges. What is a normal value range depends on the specific column. For numerical values, we should compute the average and the standard deviation and if the standard deviation exceeds the average, we can declare this as a problematic column. In columns that are categorical, we can group by the different categories and identify categories that appear in small numbers and consider those as outliers. In columns that represent intervals, we should check that the beginning of the interval is not after the end of the interval, etc.

(3) **Check for formatting**: columns should be uniformly formatted, for example dates should be in the format of YYYY-MM-DD and if there is time in HH:MM:SS (this is just an example). The important point here is the uniformity and which representation is chosen.

(4) (**Extra) Identify the dependent columns**: the existence of values in a column might depend on values of another column, e.g. if someone puts in a column that she/he is married then the name of the spouse might appear. If he/she says that is not married we cannot expect the name. In this case, this is not a missing value but a value we cannot have. (A solution could be to put a specific value to indicate these situations).

PART II

You need now to try and correct some of these errors.

- **Missing values**: depending on the type of the column we can see various solutions, for example: (i) substitute the missing value with the average, (ii) substitute the missing values with the max or the min depending on the data, (iii) make ALL values NULL, (iv) substitute with the most popular value (the one that appears the most times), etc.
- **Outliers**: delete the outliers values and substitute them with NULLs
- **Format**: identify the proper format and transform all values to this format

At the end of this process, you need to create a new table (DataFrame) with values and columns that have been cleaned and those columns that remain very problematic removed.