Automatic Data Cleaning

1. Problem formula:
Data almost alwalys come in a dirty way
Data cleaning is time consuming
Desire to clean data automatically

2. Main goal:
Design a tool which combines many fucntions to clean data automatically

3. Possible directions:
Inconsistent column names
Duplicated rows
Missing value
Outlier
Data type
Outliers
Werid features, age= -1

4. Tool capabilites:
Visualize:
    give a general idea of dataset, e.g, data type, featues meaning, distribution
Detect:
    detect common data problems, e.g, missing value, outliers
Clean:
    clean data problems, seems necessary in an interactive way, for example, users have to specify whether to fill the missing value by mean or 0.

5. Preparation focus
Literature study
Graduation plan
Case study
Tool function
    Data type
    Outliers detect
Improve seminar report
Presentation

6. Data type: If data type is determined, we can choose appropriate ways to visualize data.
    Boolean
    Numeric
    Categorical
        norminal
        ordinal

from simple to complex
Not necessary for each value to meet the requirement considering dirty data, enough if 98% data satisfy
reference: Automatic Discovery of the Statistical Types of Variables in a Dataset
Difficulties:
    norminal and ordinal

7. Outliers
    heuristic method: simple and one dimenstional
    advanced method:
        normal method, muti-dimentional, too much time
        novel method: Baysian, fast
reference: Automatic Outlier Detection: A Bayesian Approach
Difficulties:
    Add outliers to openml dataset
    Build and train your own model: Implentation?
    Is it necessary to detect outliers by heuristic methods first or just advanced method?