The ODI-2022 dataset is categorized as a survey dataset. It contains 304 different answers from students who were attending the first data mining lecture in period 5. The survey consist 16 questions. Together with the index of each answer, there are a total of 17 different variables. Some of the questions allowed the participant to fill in an open field, makes it difficult to analyze since it is not possible to assign a specific type to that column/attribute (also the range could be problematic). Even if the question asked to answer in a specific type, there are still different types of answers possible. For example when asked: “What is your stress level (0-100)?” We see that some answered “over 9000”, which is of course funny, but makes it difficult to analyze this specific attribute. We could clean these answers so that the attribute is of a specific type, but that could lead to losing too much information. Transformation of the variables to some type is also possible, but that could require advanced techniques, and somebody has to assign how to interpret answers such as “over 9000”. It could be interpreted as a stress level of 100 or infinity for example (i.e. “to what range should we assign them to?”). These kind of problems occurred routinely in the dataset. We also have the problem that similar answers can be stated in different ways which could skew results. For example, when asked: “What program are you in?” It is possible to answer either computer science or CI. Both mean the same thing, but are stated differently. So when working with this dataset we will need a way to these different formulations as one answer.