9.11.2022 Enna Augustin 050235634

Weekly assignment 2

The attributes in the dataset of the assignment are project id number, project pleage value goal in USD, project pleaged value in USD, number of tweets for a project, number of mentions of a project, number of retweets for a project, project owner retweet count, project owner follower count, project category, project introduction (texts), project introduction (pictures), project introduction (videos), project risk (texts), project owner's prior crowdfunding project and project success. From the domain knowledge view, the dimensions in the dataset are project and the project's owner. The features related to the "project" dimensions are for example project id number and project introductions. While the features related to the "project's owner" dimension are for example project owner follower count and project owner's prior crowdfunding projects. The data is two-dimensional.

The data in the dataset is quite diverse. From the characteristics view the data consists of both qualitative and quantitative data. Examples of qualitative data in the dataset are "project success", "project risk" and "project category". From these attributes nominal data are "project risk" and "project category" since they give information about the project but cannot be classified by a number. "Project success" on the other hand is ordinal data because it agrees or disagrees based on the project's success and is given a value of 1 or 2. Rest of the attributes are quantitative data because they give a numeric value for the attribute. Examples of discrete data are "project id number", "number of tweets" and "project introductions" because they can only integer numbers whereas for example project pleage value goal and pleaged value since the values are continuous and can get any values between zero and specified upper value.

The quality of the data is pretty good but there are some deprivations. There are some missing values in the data for example in the attributes "project owner's prior crowdfunding project" and "project owner follower count". When it comes to invalid values there is at least one in the attribute "project owner follower count" and one in the attribute "project owner's prior crowdfunding project". These values are invalid because the attributes are discrete and for example a person cannot have 2010,32 followers. Duplicate values or outliers I did not notice in the dataset. At the same time, I did notice that there was one data error in the set since in the attribute "project pledged value in USD" the decimal separator is supposed to be a period but in the row 67 the separator used is a comma. Besides the deprivations mentioned above I did not come across any other quality issues in the dataset.

I would like to start the cleaning of the dataset from deciding what attributes are relevant for the business issue and removing the irrelevant ones. But since we do not know what a possible business issue, I would start removing the attributes that contain text such as project introduction and project risk because I do not believe that from those attributes possible client could get any real value when performing data analysis. Next step in the data cleaning process would be removing all the rows that contain missing or invalid values. After this if there still are rows containing formatting errors would I correct those errors so that all the formatting of values in a column match. After these actions the dataset is already quite clean, but I personally feel like the data should be narrowed more down for example removing too small values from the attribute “project pledged value” so that the remaining values would be more comparable.