**1. Description & task**

For this task we are using a selected raw dataset of Clarivate Analytics’ Science Citation Index Expanded (SCIE). The data file contains information about scientific journal publications in science, technology, engineering, mathematics, and health in 2010. We focus on publications with at least one author who is affiliated to a German organization.

Your task is to present your coding skills and your ability to analyze the data — similar data will be used in the project in which you will be working.

First, please count the absolute number of articles for the given year. Keep in mind that there are different ways of counting (e.g. first author counting, whole counting, whole-normalized counting, complete-normalized counting). Describe the differences in the counting methods and justify your selection.

Second, we ask you to identify all articles which have at least one author who is affiliated to a university. Your task is to clean and re-code the variable “organization”. Use your programming skills to unify the names of the universities and make a suggestion for a codebook, e.g. “Humboldt-Universität zu Berlin” = “HU” = “HU Berlin” = “Humboldt Uni”. Please add the cleaned and re-coded names as a new variable called “OrganizationNew”.

**2. Dataset description**

The raw data is available as a .dta STATA file “Dusdal\_file4\_MergedData\_2010\_20170316\_CANDIDATES” and in addition as a .xls EXCEL file “Dusdal\_file4\_MergedData\_2010\_20170316\_CANDIDATES”. Both files contain the same information.

*Variables:*

“ut” = Accession number to identify an article. In case of a co-authored paper, each author gets the same ut number in the dataset.

“k\_code” = Identifies an entry as an original research article (k\_code==“@”)

“organization” = Organizational affiliation of an author 🡺 task: to clean

“suborganizations” = Additional information that helps to identify the organizational affiliation of an author

“city” = alternative variable to generate the variable “country”; can be used to identify the organizational affiliation of an author

“country”, “cntr”, “cntrcode” = The variable “country” was generated from the city name. It does not label the origin of the authors of a journal article, but the host country in which they wrote their article.

**3. Submission requirements**

You need to submit (1) one document and (2) all your codes. The document file describes your findings which you use to explain, e.g., the counting method you chose and how you re-coded and systematized the names of the universities. If you use supporting visualizations, please make sure that you label the figures and explain their content.

We have used STATA and SPSS in our calculations, but you may use any program of your choice; however, please justify why you think this program and your method of calculating will be optimal.