Context:

The **ForeignNames\_2019\_2020.csv and ForeignNames\_2021.csv** are datasets of commercial foreign firm names and their location countries. **ForeignNames\_2019\_2020.csv** contains firm names from 2019-2020. **ForeignNames\_2021.csv contains** firm names from 2021.

Data entry has been conducted manually by multiple people with no quality control, and as a result, the same company may have been recorded under different names, for instance: “Teaboard Ltd.” and “Teaboard Limited”.

There are two parts to this exercise:

**PART 1**

Use **ForeignNames\_2019\_2020.csv** for the following tasks:

* Clean the firm names and assign a unique identifier (ID) to each firm (include it in a new variable “cleaned\_ID”).  Give a corresponding unique cleaned name to all the names (include it in a new variable called “cleaned\_name”). For example, if Teaboard Limited, Teaboard LTd, Teaboar Ld all get assigned the same name, choose the best name to give to this group, lets say “Teaboard Limited”.
  + Please do this cleaning of firm names country by country (countries are identified by the ISO3 codes included in the file ForeignNames\_2019\_2020.csv and their mapping to country names is given in the separate file Country\_Name\_ISO3.csv we also sent you) as stopwords (words you discount in matching) may vary from country to country and also because it minimizes error.
* For all firm ID assignments, please follow a convention such that the country iso3 is attached in front of the numeric ID to identify the country of the firm.
* Divide the Data into Training and Test Sets and Apply Machine Learning: First, divide the cleaned data into training and test sets, typically using 70% for training and 30% for testing. Apply your initial cleaning algorithm to the training data, and then manually review and correct any errors in the cleaned names. Develop a machine learning model to predict the cleaned firm names (cleaned\_name) based on the manually corrected training data. The variable to be predicted with the ML algorithm is the cleaned\_name, as the goal is to standardize and accurately identify firm names despite inconsistencies in the original entries. Apply the trained model to the test data to see if it can generalize the cleaning rules and corrections. Finally, compare the accuracy of the cleaned names in the test data before and after applying the machine learning model by measuring the proportion of correctly predicted cleaned names. Comment on whether the accuracy has increased after applying the machine learning algorithm
* Include the output in a csv file that you will name **outputfile\_yourfirstname\_1** (this file will include all the observations and all the fields in the ForeignNames\_2019\_2020.csv dataset + a variable called “cleaned\_ID” (which is the ID you have given) + a variable called “cleaned\_name” (the firm name you have given))
  + Construct also a csv file that you will name **outputfile\_yourfirstname\_1\_changed** that includes ONLY the firm names that have changed as a result of your cleaning + the original firm name that cleaned firm names changed to. In our example, this will be a dataset with all the “Teaboard” firms. Teaboard LTd, Teaboar Ld are the firm names that have changed as a result of the cleaning while Teaboard Limited is the original firm name that cleaned firm names changed to. Please do NOT include in this file those firms whose names have not changed.)

**PART 2**

Use **ForeignNames\_2021.csv** for the following tasks:

* You did all the above cleaning work using a list of firm names until 2020. But the data on foreign names keeps getting updated, so now you need to expand the list of firm names to go up to 2021. You need to clean the names and assign the same IDs to firms that already exist prior to 2021 and new IDs to firms that appear just in 2021. For example, if Teaboard Ltd appears in 2021, it will get the same ID as before (2214, prefixed by the country 3-digit ISO codes). It is important to note that we again want all variations of Teaboard Ltd. to get the same unique ID just like before.
* The output from this task should be a file with all the original firm names, the variable “cleaned\_name”, the variable “cleaned\_ID”, a binary variable called “new” (taking value 1 if the firm name is new in 2021 and 0 otherwise). If the variable “new” takes a value 0, please include the name of the firm from the older data (i.e. the 2019-2020 data) in a variable called “old\_name”. Obviously, both the old and new IDs should be the same.
* Include the output in a csv file that you will name **outputfile\_yourfirstname\_2**

1. Now in the real world, this process will be continuously happening, ie, every year you will have to add new firms. How will you optimize your work? Will you run everything again whenever new firms are added or find a way to make frequent updates without rerunning everything? Please include a short description of your update process in a word or PDF document you will name **comments\_yourfirstname.**

**A few important points:**

1. We want to assign a unique ID to each unique firm. For example, if we see in the dataset Teaboard Ltd. and Teaboard Limited, we will want them to have the same ID. There is some degree of judgement involved in deciding what two firm names are the same. There is no perfect answer to it, but please **do describe your methodology**, as well as the assumptions, and tradeoffs you are making. Include all this description and discussion in a word or PDF document you will name **comments\_yourfirstname.**
2. We expect you to use **python** for this exercise.
3. Please send us your python codes, data outputs (3 csv files or 4 csv files if you do the bonus question), and a brief description of your methodology (the word or PDF file). Please write a paragraph on what you would do to improve your methods if you were given more time.
4. Please make sure that your exercise is fully replicable. Your code should **run without error on the reviewer’s computer**. (You can assume that the reviewer will have the correct versions of Python and Python packages installed. Do mention what packages are needed at the beginning of the code and the python version.)
5. Please submit your response as a public **Github repository** that the reviewer can clone and execute locally.