# Normalizing company Names

* Manual: 2 hours
* Automated: 0 hours

**Skills Required**

* Python

**Project Summary**

For this project, we will create a script that creates a copy of a column in an Excel dataset, and then normalizes the data to reduce data variation in that specific column.

**Detailed Script Requirements**

**Step 1**

First, take a look at the “input.xlsx” file in the following google drive folder:

* <https://drive.google.com/drive/folders/1pJvES_3l3zfsU_6TVAtnUYDAdm_QdinU?usp=share_link>

This file is very large, so I would recommend downloading it to view, instead of viewing via google sheets.

This file contains a large dataset with a range of columns. For this project we are only interested in column E which contains company names.

**Step 2**

Create a new column to the right of column E and call this column “companyNameNormal”.

Copy all the values from column E in to the new “companyNameNormal” column.

**Step 3 – normalizing the data**

Now, we need to ensure that the data in the “companyNameNormal” column is normalized. This is because, there are a small number of companies found in the dataset, however, there are many different ways of spelling each company. For example, consider the company name “HSBC”, there are many variations of the name in the dataset, such as:

* HSBC Investments
* HSBC UK
* HSBC Holdings

Please review the following python library and tutorial video:

* Library: https://github.com/seatgeek/thefuzz
* Tutorial Video: <https://www.youtube.com/watch?v=1jNNde4k9Ng&ab_channel=NeuralNine>

In the “input.xlsx” file, there are 000,000 values for 000 company names.

Open the “input-[example-with-normal-names]” file in the google drive folder. Take a look at “column F” in this file, you will see that each company name in “column F” has been assigned to a normal company name which is consistent for all similar company names in the column E.

The script must now use this library to replicate the result seen in “input-[example-with-normal-names]”.

**Step 4 – Saving the output file**

Now, once the company names have been normalized. You must now save the file in the following location in my google drive:

* <https://drive.google.com/drive/folders/1KcKuQ2M9_EyBj-4QdcqFzaPR9lv78rSZ?usp=sharing>

And ensure that file is input-names-normal-{date].xlsx. Where [date] is the current date in DD-MM-YY.

**Project complete**

For this project, I would like you to develop a python program to complete this task. To complete this task, I must be able to run the program on my own local desktop. If this project is completed to an excellent standard, on time with no bugs on the first code review – you will be paid a 10% bonus.

Thank you, George