During my last 6 years as a management consultant and later financial planning professional, excel was my go to software for everything – from data analysis, data collection, to dashboards or even sometimes writing notes.

As I transition to the data science world and abruptly change my weapon of preference to python I was shocked by how much faster dealing with data can be, which led me into developing this simple experiment – a comparison of the performance of Excel’s Power Query vs. Python in a task that was recurrent in my day to day: loading, transforming, merging, and saving data.

You’ll find all the files for this experiment in this repo in my github: https://github.com/morilucas/excelvspython

The experiment

The general idea is to compare the performance of both methods with loading 2 dataframes 30 columns each and variable number of rows ranging from 10k to 1M, merging and saving the results. For that we will:

1. Generate dummy data - Write a python script to generate dummy data with a range of number of rows (dataframe\_generator.ipynb), that will:
   1. Generate a single dataframe with nrows and 60 columns
   2. Split in two dataframes
   3. Shuffle the second dataframe
   4. Save both in a folder
   5. \*I fixed both dataframes with 30 columns (an arbitrary size that seems reasonable for the tasks I used to perform)
2. Python merging (merge\_dataframes.ipynb) - Write a python script to load, merge and save the results as a csv.
3. Excel merging (merge\_dataframes.xlsx) - Develop a spreadsheet to load, merge and save the results as a .xlsx file.

The results

My takeaways