Work sample: Data quality check and improvements

This work sample is about checking the data quality of fictitious companies and developing suggestions for improvement. The choice of your working tools is up to you, example SQL statements for solving the tasks in part 1 are an advantage.

Part 1: Data Quality Check

a) Check the data integrity of the company file, especially for missing or inconsistent values in important fields such as name, address and so on. How do you identify these values?
b) Identify and remove duplicate entries in the file to improve data quality.
c) Check the data format in the company file and correct any errors, e.g. incorrect date formats.
d) What else do you notice about the data set?

Part 2: Concept development and recommendations for action

After the data quality check, it is a very important task for us to improve the quality of the data pipelines and drive forward technical developments. Make a detailed proposal and develop a concept (e.g. through a flowchart) on how this can be achieved.

The starting point is the delivery of a new file from one of our suppliers that needs to be imported into the production system. The file has to run through several databases (STAGE, LIVE) and is delivered every day.

This file contains data on German companies and their directors. The company itself and the directors are not pre-matched to our dataset. Please suggest what might be good matching keys for the company data and the person data.

Which tools would you be using, which KPIs make sense?