Data Profiling

https://advancedsqlpuzzles.com

Here is a script that I wrote to perform a detailed data profile on a user defined table. There are times when I need to find which columns have NULL markers, tabs, or commas, etc.… so I created one do-all script that performs all sorts of valuable profile checks.

It's robust, feel free to customize it to fit your needs. It's easiest if you just grab the code from the GitHub repository below and test it for yourself. Once inside the code, you will see where to input the table name into a variable

At some point I hope to show examples here of the script's output and test it out in Databricks (where it might be the most useful), but for now this will need to suffice.

GitHub - Data Profiling

I would also recommend doing an internet search on "data profiling tools". There is an ever-increasing number of tools out there that have some interesting features (like creating test data) that are worth checking out.

Even if you are not new to data profiling, the Wikipedia article has a nice summary about it that I think everyone could benefit from. Here it is below.

Data profiling is the process of examining the data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that data. The purpose of these statistics may be to:

- *Find out whether existing data can be easily used for other purposes*

- *Improve the ability to search data by tagging it with keywords, descriptions, or assigning it to a category*
- *Assess data quality, including whether the data conforms to particular standards or patterns*
- *Assess the risk involved in integrating data in new applications, including the challenges of joins*
- *Discover metadata of the source database, including value patterns and distributions, key candidates, foreign-key candidates, and functional dependencies*
- *Assess whether known metadata accurately describes the actual values in the source database*
- *Understanding data challenges early in any data intensive project, so that late project surprises are avoided. Finding data problems late in the project can lead to delays and cost overruns.*
- *Have an enterprise view of all data, for uses such as master data management, where key data is needed, or data governance for improving data quality.*

Happy coding!