This is a short blog. I think the gitlab project updates would be more interesting to look at and it seems to be more promising because I finally make the pipeline work! Although the status is pending/stuck. I will look tomorrow and hopefully to address the problem.

In a very high level about what I was doing from last Friday to today.

* Successfully transfer data from local to spark then into delta lake on my local machine
* Successfully make the gitlab pipeline run (from “failed” status to “pending”)
  + app.py, config.ini, input output, requirement.txt, ci.yaml and setting up gitlab runner. It helps me learn git
* Study how to dockerize an application or project (like how do multiple container can communicate each other under the same docker network, how to access one container from local host via ports, etc)
* Study soda.io in terms of how to write rules for profiling or DQ

My next steps:

* Learn more about the delta lake, specifically need to know how to manage and possibly configure the data in delta lake (like where I can access the data in terms of what directory they live in, then would be able to adapt the profiling codes or would be able to know if it is possible to adapt the profiling codes)
* Find a public postgres database to practice the same procedure of loading data into spark and into delta lake, and how would this way of inputting data modify my pipeline
* Make the pipeline run successfully
* Dockerize everything I mentioned above so that others can test