Startup Genome - Python Test

https://github.com/analystsg/Coding-Exercise
Use this Github Repo to download the sample data under the “data” folder.

Overview:

We are including in this exercise a small subset of what we usually get from multiple funding sources.

An example of sources, for this dataset are:

1. Pitchbook
2. Dealroom
3. Crunchbase
4. Others

These sources present deals as per their own criteria. Therefore deal dates (*Original_Date)* are scattered. In addition funding rounds are tagged differently (one source would have series a while other might have series b.) (check columns *SourceType* to see this effect*)*

Example: Same deal is recorded differently by each of the sources

Hint: Look only at Type_combined column

Pitchbook:25th Mar’21: Series A
Crunchbase: 15th Apr’21: Series C
Dealroom: 20th Mar’21: Series A

For deduping we consider following constraints:

1. Naming: In the terminology the words “Series” & “Venture” are the same. For example “series a” and “venture a” are the same.

2. Confidence in sources: We assume certain sources to be more reliable than others. Hence we use this belief to dedupe when the same deal from multiple sources are present.
    a. The priority goes like this:
        i. Pitchbook
        ii. Crunchbase
        iii. Dealroom
        iv. Others(any)

3. Deal span: Also some sources are late in identifying/posting deals on their respective platforms. Even though the deal is the same the dates would be different but close. Hint: As per our analysis most same deals fall within 190 days of each other. Use this assumption to dedupe the same deals.

Objective:

Each part of the objective has its own point, be creative and enjoy the exercise.

1. Add code comments, step by step, which you are performing .
2. Deduplicate the funding rounds

Use any platform you feel comfortable with, to read csv file and perform the task. (Jupyter, Pycharm, spyder etc.)