Below you will find three exercises. Exercise 1 is mandatory, then you can choose to do either exercise 2 or 3 (or all three if you want).

Your answers for each exercise are to be provided on GitHub with instructions in a README file. The code should run on Python 3.6+ and the requirements file should be provided.
Add a one page description file that summarises the logic and code used for the assignment.
Assignments will be graded based on code and thought process.
Please make your GitHub URL public before sending it over so we can access it.

Good luck and send us an email if you have any questions at pierre@dealroom.co.

Exercise 1: G2 Scraper

In the attachments you will find a file called "data_scientist_intern_g2_scraper.csv", which contains 1,350 companies. Each company has an id, name, website, short description, address and industry.

Your task is to build a scraper for https://www.g2.com/ (G2 is a website providing product reviews). Your script should look up the companies from the file and return the corresponding product information for each company.
The scraper should get the product URL (e.g. https://www.g2.com/products/trello/reviews), website, rating, number of reviews, description, product and seller details, alternatives (optional) and pricing (also optional). You can save the scraped data to the initial companies file and send it to us together with the script you've built to scrape the data.

If you get blocked by G2, it's ok, just share the results you scraped. The point is the code and the logic.

Exercise 2: Revenue prediction model

The file "data_scientist_intern_revenue_model.csv" contains data on 27k companies. For each company there is historical revenue data for the period 2015-2019 and historical employee data for 2017-2020 (beware that there are missing values). You need to create a regression, random forest, or neural network model, which accurately predicts the revenue figure for each company for 2020, based on the historical data. Add the prediction to the initial dataset and send it to us together with your code and brief accuracy report.

Exercise 3: Duplicate detection

In "data_scientist_intern_duplicate_detection.csv" you will find a sample dataset of Dealroom company data containing some duplicates. Your task is to identify the duplicate profiles based on the information provided in the file. Please create a separate file to store the duplicate profiles that you identified. In this file you should show which of the profiles (data rows) are duplicates and which fields they are duplicated on.