Data Cleaning - Assignment 2

Andrea Armani

Data are originally stored in 498 files, where each file represents a Pandas DataFrame with the same structure. Therefore, I was able to import and merge all of them using OpenRefine.

In the following, I list the transformation that I applied to such a dataset.

1. Importing all the table with the appropriate Data type per column;
2. Removed Stock Splits column: all the rows have the same value ( 0 ) except for one;
3. Removed Dividend column: it always maintains the same value ( 0 ) for each row;
4. Renamed File column into Company;
5. Replace string in Company column: every row is in the form
   1. 2019-09-01\_2020-09-07\_
   2. Company name
   3. .csv

Therefore I can remove a. and c. in order to obtain the actual name of the company.

1. Generaldynamics rows lack a sector value. Since the company operates in aerospatial and defense, it can be included into the Technology sector, so I massively updated this value.
2. Round values in columns Open, High, Low, Close: since OpenRefine doesn’t have a proper function to round a float number up to a number of digits, i solved the problem by casting the values into string and then looping the string in order to save only the relevant digits. After that, i casted the string back to numeric.
3. Finally, I didn’t convert the Date into DateType because it would add the minutes and the seconds to the value of each cell and such granularity is useless.