**What is the most likely future release date of a shares outstanding report based on your analysis of the reports provided?**

**Problem analysis:**

To find the future release date of a shares outstanding report, we need to find the possible common features from the historical report. According to the observation of several files, almost words in the report are describe the regulation. These words seem no relevant to the questions. Therefore, we don’t have to use language model or other methods to analysis the semantic of words. Next, we will try to extract some important date value, numerical value and key words from the report.

There is some potential information that could be feature.

Common information in the report:

* 1. Company Name
  2. Report release date
  3. Report type
  4. Ordinary share average price (USD, EUR)
  5. Ordinary share high price (USD, EUR)
  6. Ordinary share low price (USD, EUR)
  7. Date of company cancellation announces (relevant to released date)
  8. Number of cancellation shares
  9. Number of Ordinary Shares in issue

The question can be transfer to find the date interval from the company cancellation announcement date to report release date.

**Methodology:**

This part is about how to extract the target information from a file. According to the observation, we find almost files are HTML documents except a file named “vAJpAaDA”. Then I use the python package BeautifulSoup to parse the html file in tree structure.

For the released date and report type, I found that these two information are stored in the div tag “marketnews-released” and “marketnews-h1” for all the files. Therefore, we can use search tag to extract the information.

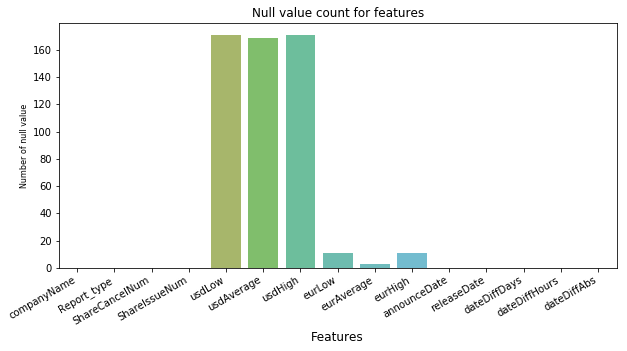
For the other information , I tried to use regular expression to locate the target sentences and then use regular expression again to filter the numerical value or date.

Then, all the extracted information is stored in a piece of dataframe and finally all the information will store in a large dataframe. After we collect the target features, I will try to analysis the parsed data. Since the dataset is very small, we only get 263 fils in total including some missing data entries. The machine learning algorithms may not work well in this question.

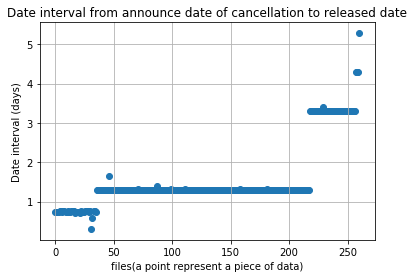
**Data analysis:**

According to the dataframe, we can find there are three type of report. ('Transaction in Own Shares', 'Share buy-back programme', 'Transaction in Own Shares - Replacement') The main type of report is “Transaction in Own Shares”. The report “Share buy-back programme” is different so it should be deleted.

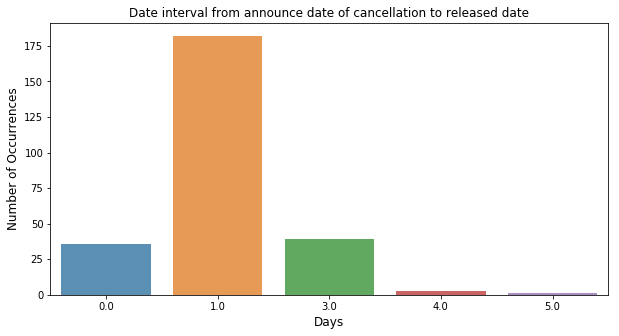
The following figure show the null value of each features.



Next, we add the extra feature which represent the date interval from the company cancellation announcement date to report released date (in days and in hours).



We can easily find that the data interval is very regular.



Hence, we can make a general conclusion. The released date of share outstanding report is in the next day of the company cancellation announcement.