**Why are the dates and titles missing?**

Question: Why are the dates and titles missing in some entries/keyword matches in the final output?

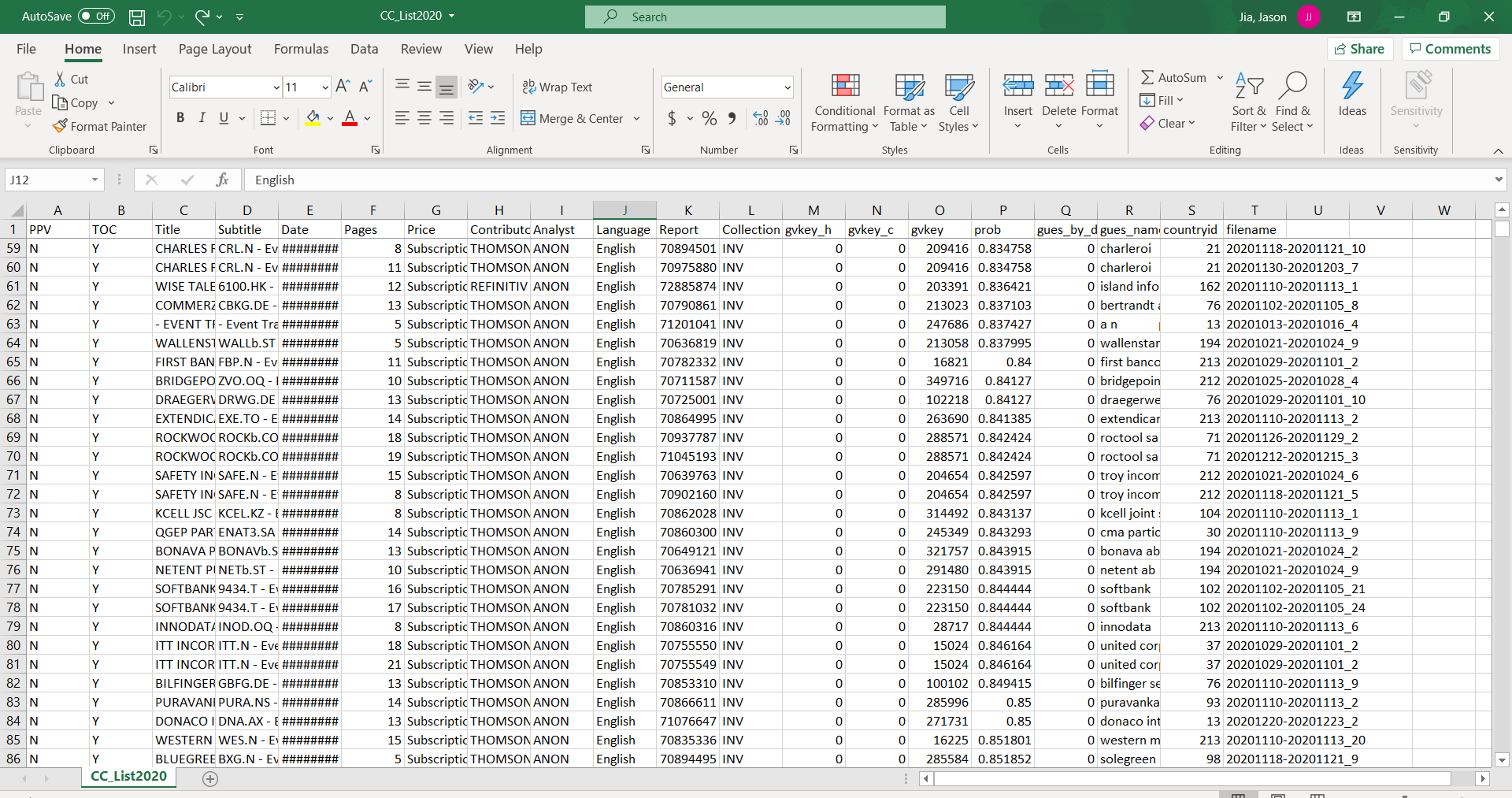
Answer: It’s because these entries belong to firms not matched with a gvkey, and dates and titles are extracted only from the dataset of gvkey-matched firms. This is also why the gvkeys are missing.

**Details**

The relevant stages of the processing pipeline are Stages 3 (Firm Identification) and 4 (Keyword Identification).

**Stage 3: Firm Identification (gvkey Matching)**

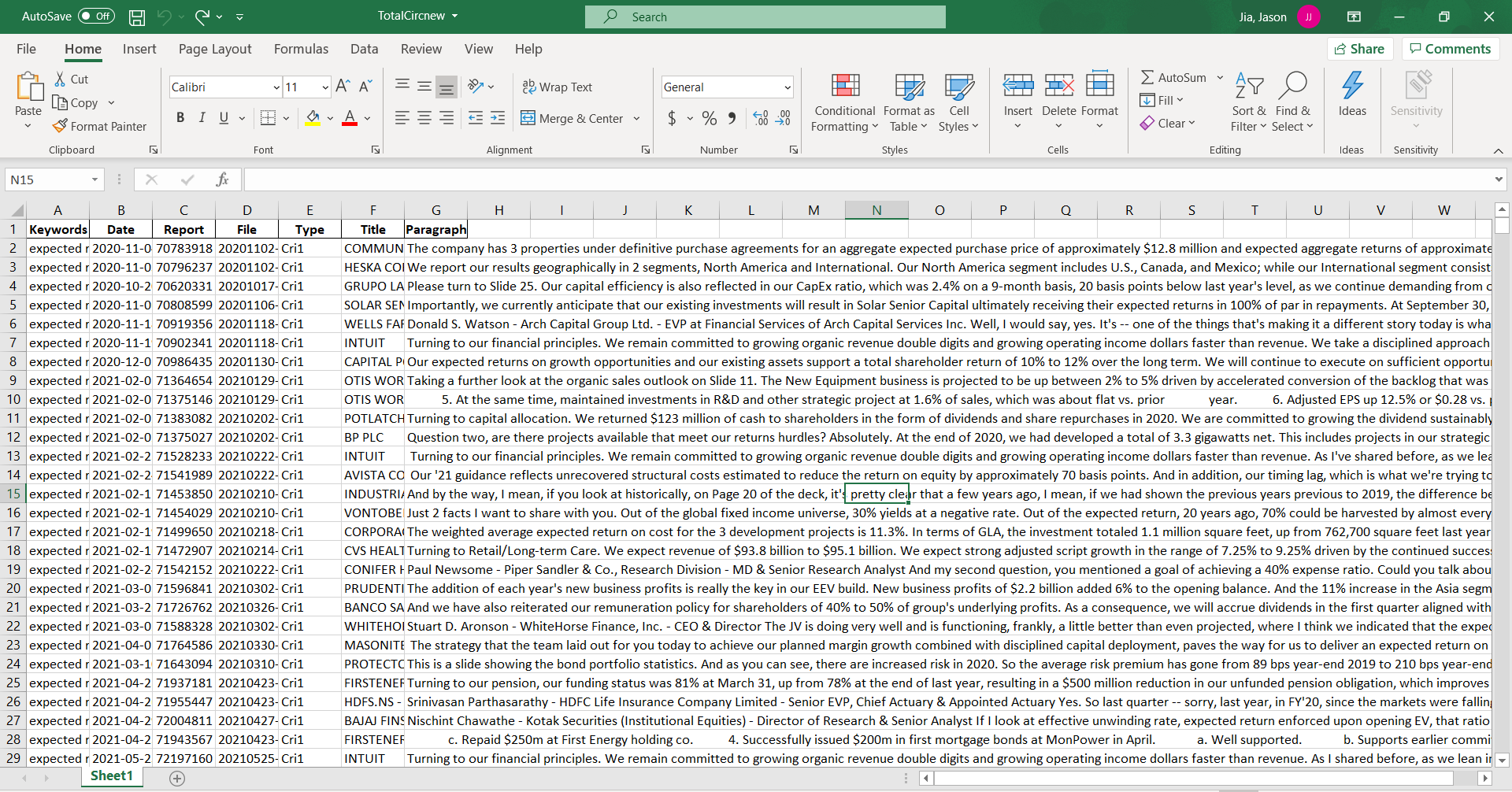
* Output: *CC\_List.csv*, a table of gvkey-matched firms with identifying information (e.g. dates, titles, report ID).



**Stage 4: Keyword Identification (find paragraphs containing keywords)**

Stages 4.1 and 4.2: Run *keyword\_ident\_1.py* and *keyword\_ident\_2.py*.

* Output: *TotalCircnew.xlsx*, a table of keyword matches with identifying information (e.g. report ID)



Stage 4.3: Run *mergeclean.do*.

* Inputs: *CC\_List.csv*, *TotalCircnew.xlsx*
* Purpose: Merge the 2 types of datasets using report ID to get a consolidated dataset with keyword matches, gvkeys and identifying information. Keyword matches in *TotalCircnew.xlsx* with report IDs that don’t appear in *CC\_List.csv* are still retained.
* Output: *cric1\_newtotal.xlsx*
  + Columns in blue come from *CC\_List.csv* and columns in orange come from *TotalCircnew.xlsx*.
  + Dates and titles are missing in about 2.5%-3% of the entries. This happens in both the old (2001-2020) and new (2020-2021) compilations.
  + After comparing the report IDs, it is found that most (>99%) of the missing dates in *…\Jason-Kilian\Conference Calls with missing dates\paragraph\_datemissing.xlsx* belong to this pool of incomplete entries.

