**Gvkey Matching Update 1**

Check if Hassan is better than Compustat:

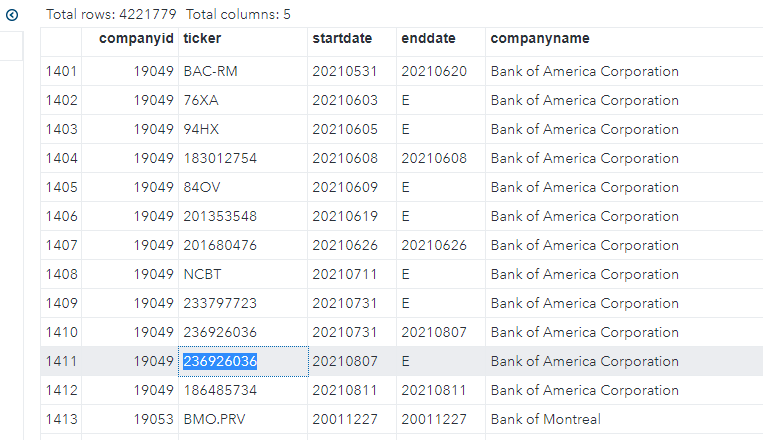
* From Sixun’s documentation: “In addition, we have Hassan’s Firm-Level Political Risk Dataset, which also provides us with GVKEY and firm names. We think that this data set contains almost all the company’s names that we have, since we use the same data source. In addition, in contrast to the Compustat dataset, Hassan’s dataset has English spelling for all company’s names, which helps us, since Conference Calls also has English spelling names.”
* I’m not sure about the English spelling part, since I’ve looked at about 500 firm names on Compustat and they’re all in English too.

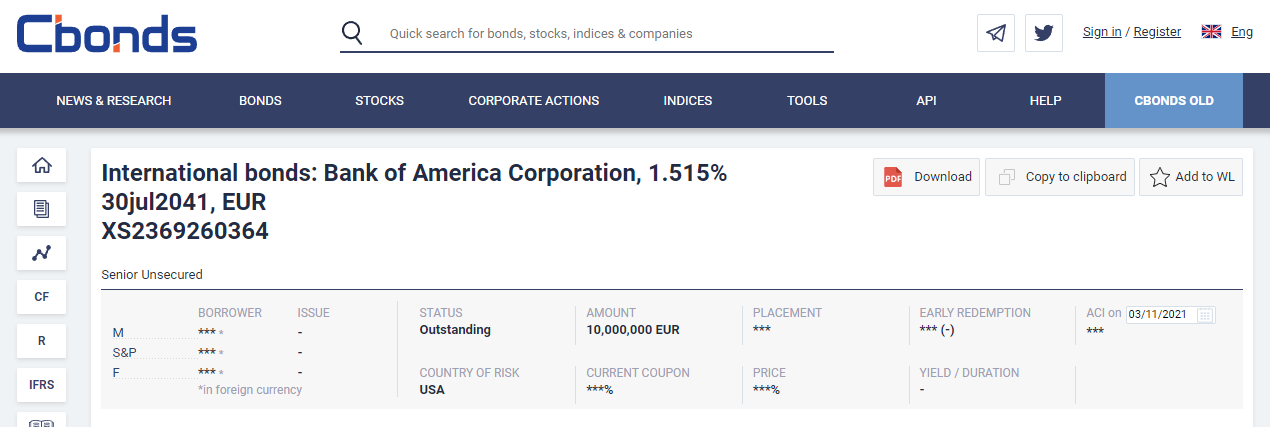
Check firm ticker – what it means to keep the first / last entry.

* Generally, yes, the tickers for a firm are sorted by date, so the last entry is the ‘most recent’.
* But to be exact, tickers are first sorted by the startdate, then by the enddate.



* Also, both stock and bond tickers are included, so in general, the last entry may not be the most recent *stock* ticker for the firm.





Check what the ‘country’ columns represent

* In the ciqcompany dataset, there are 2 country columns:
  + Countryid: likely the country of HQ.
  + Incorporationcountryid: country in which the company is incorporated or legally registered.
  + I’m using Countryid as the country variable.
  + The manuals I found didn’t contain these variables, so I have emailed the Capital IQ Helpdesk to confirm.

Country information is available in the S&P Capital IQ packages in the ciqCompany table in Xpressfeed™ as follows:

• incorporationCountryID: Country in which the entity was incorporated

o Represents the country where a company is registered as a legal entity

o Commonly called Country of Incorporation, Country of Registration, or Country of Domicile

• countryId: Country identifier for the entity headquarters

o Represents the country where a company’s global business address is located or where a company is headquartered

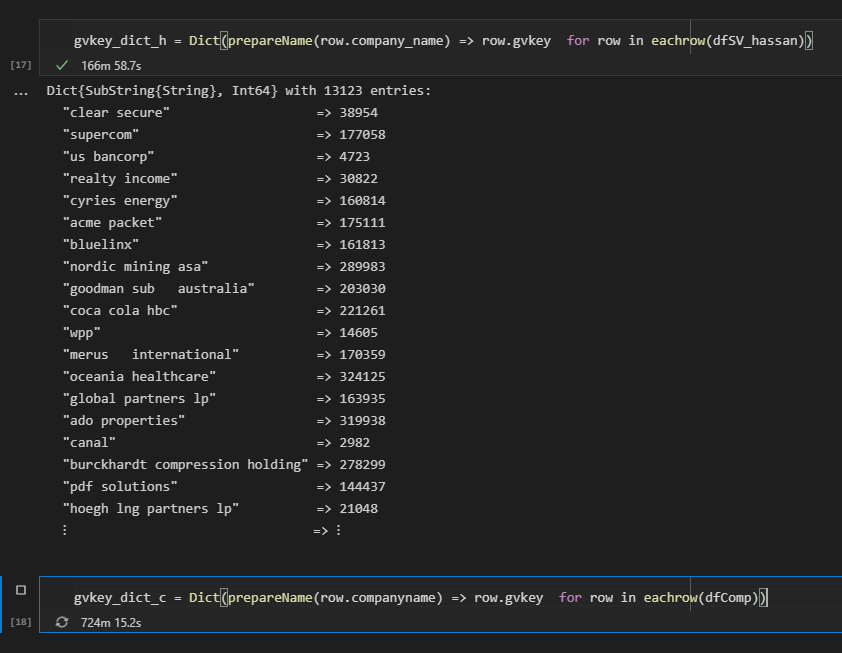
o Commonly called Country of Headquarters or Country of Issuer

These two data items may or may not be the same.

**In progress:**

Produce the gvkeys using the original code/method

* The current bottleneck is the runtime. I’ve been running the code for 15+ hours, and the program is still executing a particular line, which is to turn a dataframe into a dictionary. The runtimes are displayed below: the first one is completed in about 3 hours, and the second is still running after 12 hours.



Produce some statistics after getting the gvkeys

* Number of perfect matches
* Number of imperfect matches > 80%
* Number of imperfect matches >90%
* Whether it’s matched by fuzzy or by identifying the name perfectly.

Check for duplicates

* Some firms have multiple gvkeys. We want to avoid the situation where the same firm is matched to different gvkeys and thus looking different in our dataset.
* Stata duplicates report to check if the firm name appears twice