Friday, July 12, 2024

Mason Zhang - Pre-Doc Researcher Test

Overview

In this repository, I have included data processing scripts and output datasets for the two data tasks. For a detailed step-by-step working process please refer to the comments within the script files. In this README file, I will summarize what I did and highlight several key points noticed during data manipulation.

Files in the Repository

- task1_script.R
- task2_script.R
- task1_result.csv
- task2_result.csv
- README.pdf
- README.md
- Theory_Task.pdf

Task 1: Earnings Announcements for European Firms

Summary

To start with Task 1, I loaded data from the EuropeEAs, Europefirmquarters, and tradingdates files. I then converted the date columns to the same format to ensure consistency. Next, I made sure all CUSIP values were 9 digits long by adding zeros at the start if necessary. I also extracted the last 6 digits of CUSIP to create SEDOL.

After formatting, I merged the datasets based on sedol and fiscal quarter end dates. If an earnings announcement was made after 4 PM, I adjusted the date to the next day. Then, I used the tradingdates file to find the next available trading day if the adjusted date was not a trading day.

Finally, I created the output dataset by removing invalid CUSIP entries and duplicates, ensuring each firm-quarter was unique and had the correct trading date.

Data Manipulation

To manage data errors, I ensured uniform date and time formatting. I removed duplicates to maintain unique firm-quarters. For missing data, I filtered out rows with missing or invalid `CUSIP` values and included only those with valid dates.

Task 2: Non-ESG News

Summary

For Task 2, I started by loading data from the allnews and ESGnews files and converting the date columns to the same format. I identified ESG-related topics, groups, and types based on the ESGnews dataset and a manual inspection of the allnews dataset.

Note: In order to identify whether news in the allnews dataset is ESG-related, we need to first check its TOPIC. If its TOPIC is one of the ESG-related topics, in this case, Environment, Society, and Politics (Governance), it is ESG-related news. If its TOPIC is business or economy, we then need to check its GROUP. Lastly, we check the TYPE.

I then identified firms with ESG news events and extracted all news events for these firms from the allnews dataset. I marked ESG-related news. Non-ESG news events were identified by excluding marked ESG news.

I filtered out non-ESG news events that occurred on the same date as ESG news events for the same firm. The final output dataset included unique firm-dates with non-ESG news events.

Data Manipulation

For data errors, I ensured consistent date formatting. I addressed duplicates by removing them to maintain unique firm-dates in the final dataset. Missing data was handled by ensuring only rows with valid dates were included.