Our team determined what data do we want to search for and where. At first, we kept our ideas of a data source very vague on purpose; to not constrict ourselves to come up with other ideas relating to a dataset. We focused on macro view point, and decided we wanted to get data for the whole United States; that could be algaculture, weather, or anything of that nature, relating to the United States.

With that in mind, we decided to conduct our search through Kaggle. We found many datasets and tried to hypothesize what dataset can correlations with another. After many hypotheses attempts and datasets, we came a across a dataset relating to the stock market in the api yahoo finance. An idea developed; if the stock market index, could be affected by weather, or extreme weather in a yearly basis. The weather data was found at Kaggle. We realized that it would be hard to determine stock trends solely based on weather, so we added another source; natural disaster.

Our sources are as follows:

* Yahoo Finance API.
* Historical Surface Temperature of the World per Country, csv data.
* Natural Disasters table in the website Wikipedia.

The transformation was straight forward for stock and weather. The stock and natural data have correlations in year. We attempted to create an associative entity table for stock and natural disaster which resolve many-to-many relationship, but had trouble with this step, and discarded it all together. Instead we normalized year for all the tables, and natural disaster. We went with this approach since data such as weather, stock, disaster shall, hypothetically, consistently be updated in a yearly basis. The more normalized tables shall make it easy to update some of the disasters that occur, and the years passing by.

We realized that normalizing the natural disaster data was going to be an undertaking. We only normalized the year and disaster for the table weather damage. The death toll, and damage cost has a inconsistent format with damage cost containing numbers and word. Location also was not normalized due to time constraints.

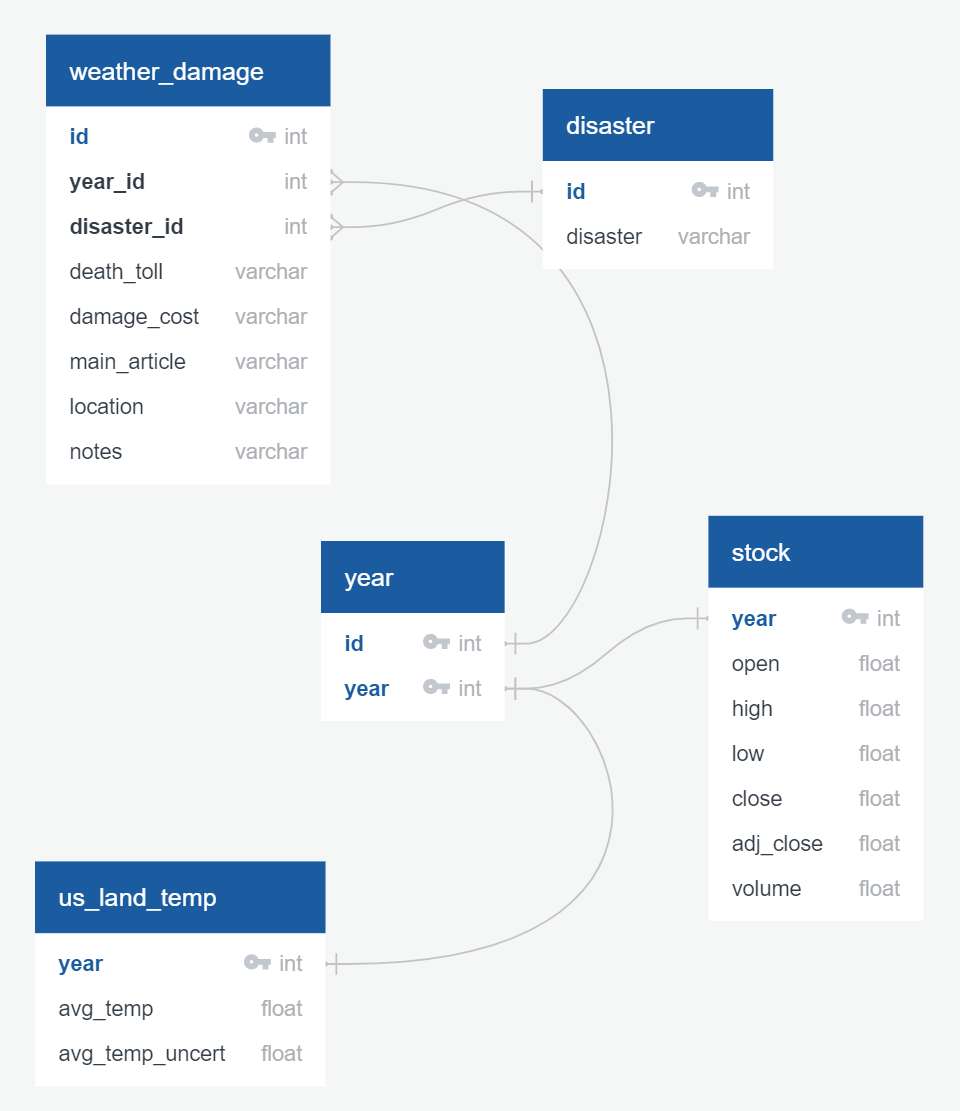


Figure Schema of the tables