Data cleaning steps:

Since our data cleaning involves doing some string formatting and string stripping, we have done some pre work using Python.

1. Convert ‘Salary Estimate’ column into two separate columns: 'salaryLB' and 'salaryUB' and remove unnecessary characters in the column, for example, $, (Glassdoor est.), etc. And then convert values in 'salaryLB' and 'salaryUB' to int.
2. Convert ‘Revenue’ column into two separate columns: revenueLB and revenueUB and remove unnecessary characters in the column, for example, $, (USD), etc. And then add ‘000’ if revenue is in billion instead of million. Convert values in 'salaryLB' and 'salaryUB' to int. And replace 'Unknown / Non-Applicable' into null.
3. Separate ‘Locations’ into ‘city’ and ‘state’, and 'Headquarters' into ‘city’ and ‘state’, and then combine them and remove all duplicate locations and save that into a Location Table.
4. Remove all unnecessary characters after ‘\n’ in ‘Company Name’.
5. Remove all duplicate combinations of ‘Sector’ and ‘Industry’ and save that into Sector Table.
6. Separate each single competitor company in ‘Competitors’ into a new row in the Competition Table, since in the original data set, if a company has multiple competitors, all competitors stay in a single cell.

After fixing the basic format problem using Python, we are trying to import the data to psql. Here we are going to delete the data which violates our constraint, the details of the clean step can be seen in clean_data.sql.

Headquarters:

Removing data which revenueLB or revenueUB is less than 0.
Removing data which founded time is less than 0.
Check whether the size of the headquarter is in the given set we provided in our schema.ddl, remove it if not.
Here we don’t need to check whether the locationID is in Location since we didn’t remove any data from the Location table.

Company:

Check if the headID is in the table Headquarters, remove it if not.
Check if the value of the locationID is valid, remove it if not.
Check if the rating of the company is valid, which needs to >= 1 and <= 5. Remove it if not.
Check whether the ownership of the company is in the given set we provided in our schema.ddl, remove it if not.

Competition:

Check if aID and bID is in the table Company, remove it if not.

Job:

Check if companyID is in the table Company, remove it if not.