ETL Project: Craft Beers

Project scope:

We derived data from 2 sources, <https://www.kaggle.com/ritesaluja/beer-beer-beer/data> (Beer-beer-beer. csv) this contains the address of Craft breweries across the world, beer name and style this contains name, Beers types, styles and other related data. <https://www.kaggle.com/nickhould/craft-cans#beers.csv> (Craft Beers Data set.csv). this contains name, Beers types, styles and other related data.

Both files are in a CSV format for Ease of modification. Once cleaned of extraneous data in Python they will each be loaded in to PostgressSQL so they can be queried in a structured data base

# **E**xtract:

Data Sources:

**Beer-beer-beer**: Initially contained the following columns: Name, id, brewery\_id, cat\_id, style\_id, Alcohol By Volume, International Bitterness Units, Standard Reference Method, Universal Product Code, filepath, Description, add\_user, last\_mod, Style, Category, Brewer, Address, City, State

Once cleaned in Python it will contain the following columns: Name, Brewer, brewery\_id, id, Address, City, State, Country

**Craft Beers Data set**: initially contained the following columns: abv, ibu, id, name, style, brewery\_id, ounces

Once cleaned in Python it will contain(We decided to not remove any columns): name, style, brewery\_id, id, abv, ibu, ounces

# **T**ransform:

During this process that following will be done:

Duplicate data will be removed for both files using Python, it was found that Multiple records exists containing the same data

Extranious Columns will be removed from Beer-beer-beer data set: , cat\_id, style\_id, Alcohol By Volume, International Bitterness Units, Standard Reference Method, Universal Product Code, filepath, Description, add\_user, last\_mod. These columns are either duplicate or not needed in final data set

# **L**oad: