1. **Extract**

For this ETL project, I wanted to find datasets that were large enough to allow me to practice cleaning and manipulating data. I searched through two websites (Kaggle and data.world) and ended up with two CSV files with wine and beer data. The wine file was about 130,000 rows and the beer data had a little bit over 1,000,000.

1. **Transform**

Both data sets were large and had columns that needed cleaning. The wine data had columns that were unnecessary like “twitter handle”, “taster name”, and “description”. I used pandas and python to drop those columns and kept the ones I needed. Once I had all the data I wanted, I renamed the columns to be more descriptive.

The process with the beer data was roughly the same. I loaded the csv file into the notebook then cleaned the file using pandas. The only difference is that there were many duplicates in the dataset, so I dropped any duplicates from the data frame I wanted to use.

1. **Load**

I decided to load the two files into pgAdmin and use SQL to make queries and analyze. The relational database worked well with these files because it was structured with rows and columns. I created two tables (wine & beer) before creating the schemas below:

CREATE TABLE wine(

Id INT PRIMARY KEY,

Points INT,

Price FLOAT,

Variety VARCHAR,

Winery VARCHAR,

Province VARCHAR,

Region VARCHAR,

Country VARCHAR,

Designation VARCHAR,

)

CREATE TABLE wine(

Id INT PRIMARY KEY,

Overall\_review FLOAT,

brewery VARCHAR,

beer\_style VARCHAR,

beer\_name VARCHAR,

ABV VARCHAR,

)

Once my tables were created, I then created a connection to my pgadmin in jupyter notebook. I then uploaded my CSVs to the tables and made sure they worked by using sqlalchemy.