**ETL Project 2020**

**UCSD Extension Data Science and Data Visualization 2020**

**Team Nine**

Nicolas Tacos

Carl Patterson

1. Extract

The topic for our data was breweries within the US. Data was searched on ‘Kaggle.com’. There were three (3) files to review and use for our project.

* beers.csv
* breweries.csv
* reviews.csv

1. Transform

The files were downloaded, and cleaning was done with Jupyter’ and the files were ‘beers\_and\_breweries.ipynb’ and ‘reviews\_cleanup CP10Nov20.ipynb’. Data was checked for ‘null’ values and removed from each of the files. The required fields were extracted from the data and each file was saved for upload into a SQL database. The ‘reviews.csv’ was rather large. The null values were removed, and the data was sorted by state and we selected CA for the final csv file. The schemata for the database is given below.

A close up of a sign

Description automatically generated Table-1: Schemata for Database.

1. Load

The database was loaded and the interrelationships between each of the tables was noted. The ‘brewery\_id’ between the breweries and ‘beers table’ is the connection between these two tables. The connection between the ‘beers’ table and the ‘reviews’ is the beer\_id. There is no connection between the ‘breweries’ table and the’ reviews’ table. These connections were chosen due to the common fields of these tables.