**ETL Project: Wine!**

**Extract: your original data sources and how the data was formatted (CSV, JSON, MySQL, etc).**

In starting this project, we first looked at what data was out there and easily accessible. As anyone is this class is aware, there is a LOT of data out there. In browsing through countless sites, we found one with a downloadable csv file containing information on a variety of wines. It contained data on the country and region the wines came from as well as bottle price and a rating from Wine Enthusiast magazine among other things. We decided it might be interesting to look at the locations the wines came from a little bit more closely, to be exact, the GDP and population of the different countries. We also decided to split the US out by state as they are more comparable in size to countries across most of the rest of the world.

In order to get this data, we downloaded a list of countries worldwide and used that to pull GDP and population data from the World Bank API. This returned JSON objects which we wrote to a data frame and exported as a CSV file. For the state GDPs, we grabbed a table from a website (), went ‘Old School’ on it and edited it in an Excel Workbook. It was only 50 entries and so it was easier to manipulate the data and add a new column manually.

**Transform: what data cleaning or transformation was required.**

In our original data set, there were several rows that were missing data as well as some columns containing nothing or data, we felt was irrelevant. We created a new data frame that only included columns that we felt contained worthwhile data and then dropped any rows that were missing an entry in any column. This still left us with over 90,000 rows of data, which is a significant amount of information.

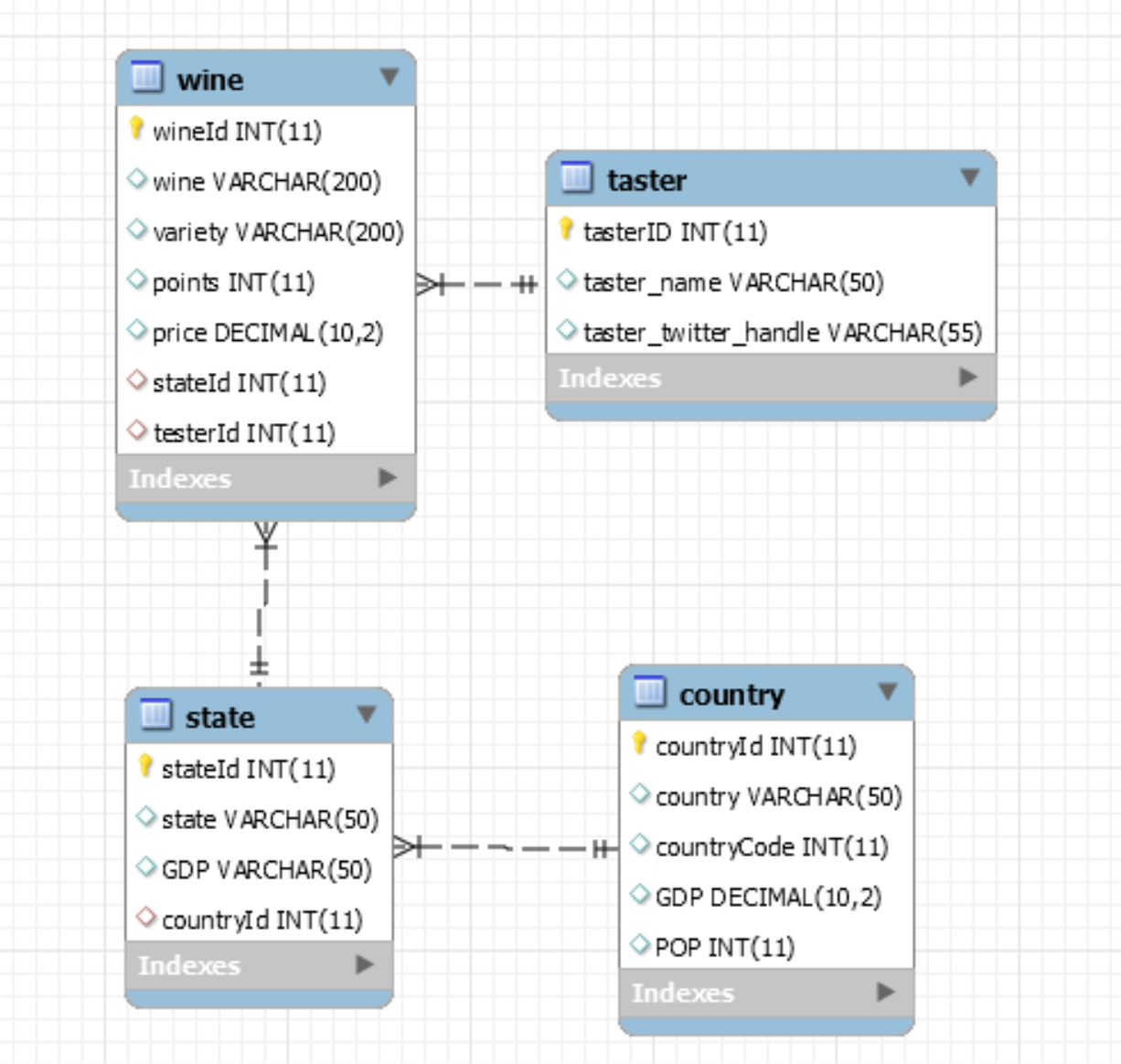
For the state data, we added a Country ID column in excel, adding in the Country ID for the USA from the country data. We also extracted the taster data (taster name and twitter handle) and used Python and Pandas to add an index and create a taster ID that we could use as a primary key for tasters and a foreign key in our main table. It was more complicated and frustrating to get pandas to let us create an ID column and set it as the index for the tables than I can possibly put into words, politely.

**Load: the final database, tables/collections, and why this was chosen.**

To finish things off we merged tables back together and deleted extra columns to reduce redundancy. Our primary table includes foreign keys to the other tables for ease of access and to maintain clear relationships between our data points. We felt this was the simplest way to store and show our data and its relationships within our database.

\*\*\* Note: In compiling the final database, we ran into the issue of special characters in SQL, that SQL refuses to recognize some characters used in our data. To get around this, we had to use a special code to get SQL to allow our special characters and then use that to alter our individual tables.

Database schema



Team WineTime:

Sandra, Sumaya, and Emily