**ETL Final Report**

Nishi Gadireddy

Alex Young

Jonathan Cornish

In this project our team set about scraping web data on the prices, types, and tasters’ ratings on various wines. The goal was to create a database containing all the relevant information that we could then analyze to determine if there is a statistical relationship between wine price and wine ratings.

***Extract:***

The original data sources that we used are as follows:

1. <https://top100.winespectator.com/lists/>
   1. The Top 100 Wine Spectator List
   2. We scraped data from this website and put it into a table using pandas
2. <https://www.kaggle.com/zynicide/wine-reviews>
   1. This kaggle website has various large CSV files that we downloaded and then worked off of

***Transform:***

The first data set that was transformed was the web scraping data from the wine spectator list. The data was cleaned using Pandas and Jupyter Notebook. In the cleaning process unnecessary columns were removed and some columns were renamed. The next set of data that was cleaned was the Kaggle CSV. Here unnecessary columns were removed and titles were renamed to make merging easier. Additionally, the Kaggle data was cut to only show the first 150 wine reviews. Then all null values from the cut data set were removed. For the kaggle data we had to reformat the price information, which was a string with a ‘$’ sign at the front of the string. We reformatted it using the pandas ‘map’ function and turned the price strings into numeric values. After cleaning both of the datasets the two were merged with the common column names of price points, and wine name. The method we used for combining the data tables was a ‘union’ - whereby we put one data table directly underneath the other. Finally, the columns were reordered for better readability.

***Load:***

We used MYSQL to create the final tables by first creating tables that matched the columns from the Pandas dataframe. We used SQL instead of MongoDB because the format of the data that we were working with was already in two tables and thus we thought that it would be an easier way to analyze the data. Then we connected the database and loaded the final result from Pandas. To confirm that the data was properly loaded we performed a series of queries.

***Conclusion:***

In this project we were able to put to use course material we’d learned regarding scraping web data, using pandas to format data and make edits to tables, and SQL Alchemy to union two tables into one large table.