ETL Project Write-Up

**Subject:** Wine Reviews

**Question:** What are the top 5 & bottom 5 wine titles (combination of winery, vintage, and designation)?

**Data Sources:** <data.world> ([Wines.xlsx](https://github.com/Sasmita28/ETL_Project/blob/etl_project_sasmita/wine_reviews_data/Wines.xlsx))

[Kaggle.com](kaggle.com) ([winemag-data-130k-v2.csv](https://github.com/Sasmita28/ETL_Project/blob/etl_project_sasmita/wine_reviews_data/winemag-data-130k-v2.csv))

**Steps**

**Extract (Randall: CSV & JSON)**

* Put CSV files in a directory.
* Use Pandas.read\_csv to pull in CSV data.
* Create Pandas dataframes of CSV files.

**Transform (Ben: CSV, Sasmita: JSON)**

* Ben: Clean Pandas [winemag-data-130k-v2.csv](https://github.com/Sasmita28/ETL_Project/blob/etl_project_sasmita/wine_reviews_data/winemag-data-130k-v2.csv) dataframe.
* Sasmita: Clean Pandas [Wines.xlsx](https://github.com/Sasmita28/ETL_Project/blob/etl_project_sasmita/wine_reviews_data/Wines.xlsx) dataframe.
* Sasmita: Join the Pandas dataframes.
* Ben: Create equivalent tables in PgAdmin and assign Age as Primary Key.

**Load (Sasmita)**

* Use Pandas.to\_sql to load the joined table to the PostgreSQL database.

**Columns to Keep for Each Table**

* [winemag-data-130k-v2.csv](https://github.com/Sasmita28/ETL_Project/blob/etl_project_sasmita/wine_reviews_data/winemag-data-130k-v2.csv) (id, country, description, designation, points, price, title (Primary Key), variety, winery)
* [Wines.xlsx](https://github.com/Sasmita28/ETL_Project/blob/etl_project_sasmita/wine_reviews_data/Wines.xlsx) (Vintage (Year), country, designation, points, price, title (Primary Key), source, winery

**Tips for Handling Duplicates**

* **Take average of points & price for each title.**
* **If title has two sources, then change the source to avg.**