ETL Project: Wine Review Data

Vipul Aggarwal, Liliana Joya, Kannika Phadougxath

# Project Scope

The scope of this project is to perform ETL on a dataset regarding wine reviews and complementary information obtained from different sources. The wine data is extracted from Kaggle, Wikipedia and the Geopy library, then cleaned, transformed, and normalized to a form where it is ready to be loaded into a database, in this case, via SQLAlchemy.

# Extraction: Dataset and Sources

The dataset chosen is a wine reviews table from Wine Magazine. The dataset was found in Kaggle. It contains variety, location, winery, score, as well as information about the person reviewing the wine. The data was checked for integrity as we looked for duplicated values and NaN values among others.

Additionally to the wine reviews table, location data was used together with the Geopy library to expand the information and include longitude and latitude of the region and a CSV file including that information was created and used as a secondary source of data.

Finally, a table of wine color (red, white, rose) was extracted from Wikipedia to complement our grape variety information.

# Transformation: Data Cleanup & Normalization

## Transformation

Wine Reviews Data

1. Keep Relevant Data: First step was to remove unnecessary columns for the relevant data such as “description” and “designation”.
2. NaN Values:Check for NaN values in “country” and dropped when the country information is not available. Location is an important factor and we need to have it.
3. Duplicate Values: The original data apparently contained no duplicated values, but this was due to the first column “id” having different values, regardless of some rows containing the same information thereafter. Once we reset the index and removed the id column, we identified that there were indeed duplicated values indeed. There were approximately 10 thousand duplicated values. We dropped the duplicated values.
4. Integrity check: Through out the check, types of data, unique counts of data and total counts of data were preformed to check the integrity of the dataset.
5. Creating new “year” column: Even though the year information was presented in the title, it was considered important to obtain that information in a separate column, which is why the date was stripped from the title to create a new “year” column.
6. Creating new “location” column: The data included the different notations in terms of location, we wanted to present all this information together in one column so “region1”, “region2”, “province” and “country were combined to obtain complete location information for every row. This would help us avoid to deal with NaN values in either one of the other column because it would provide at least one row of accurate location information.

Latitude and Longitude Data

1. To create the latitude and longitude data it was decided to create the list in a different jupyter notebook, create a csv out of the list and read the csv directly to the main jupyter notebook to avoid large running times.
2. NAN Values.
3. Preventing duplicated new values for the final dataset: We found there were some grapes that could be used to produce either red or white wines.( Arnsburger, Cereza, Grignolino, Pignerol, Shoenburger) were producing double entries when joining to our main dataset. In this case after a short google process we identified that most of those produce actually white wines so we dropped the first values in the grape color dataset, which coincidently were all the red values.
4. NAN values in the taster table: We decided we don’t need to drop the wine information just because there is no taster information available so we are not dropping the value and assigning the NAN value to the review table when the taster information is not available.
5. Setting the length of

## Normalization

## Type of final production database

Adsfadsf