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 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.

## Summary

# 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. There were approximately 10 thousand duplicated values, which were dropped.
4. **Integrity check:** Throughout 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 “combined 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 dealing with NaN values in either one of the other columns because it would provide at least one row of accurate location information.
7. **NAN values in the taster table:** The NaN values in the taster column were kept. 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.

Latitude and Longitude Data

1. **Create different dataset:** To create the latitude and longitude data it was decided to create the lists in a different jupyter notebook.
2. **Create CSV:** A csv was created out of the list and read the csv directly to the main jupyter notebook.
3. **Import CSV:** The CSV with Lat and Long data was imported to the jupyter notebook to avoid long running times.
4. **Join CSV:** The CSV joined on “province” for all our data in the main dataframe.

Wine variety and color data

1. **To\_html:** After reading the Wikipedia information in pandas, three dataframes were created from the tables, one for each wine grape color.
2. **Cleaning the rows:** Rows with multiple variations were split into different items of a series and then stacked into a new dataframe.
3. Finally, the color for each dataframe was included as a new column and all three dataframes were merged into one combined wine color df.
4. To allow a smoother join to the original wine dataframe, in the grape varieties columns all letters were set to lower case and spaces were stripped from the values. This allowed 80% of the data could be matched to Wikipedia data.
5. **Preventing duplicated new values for the final dataset**: It was observed that some grapes could be used both to produce either red or white wines( Arnsburger, Cereza, Grignolino, Pignerol, Shoenburger) which could produce multiple entries when joining to our main dataset, once as a red wine and once as a white wine, which would be incorrect. In this case after a short google process it was identified that most of these produce white wines, so the first entry in the grape color series were dropped, which coincidently were all the red grapes.
6. **Remaining NaN values:** The remainder of data in the variety column that had no matching color were further analyzed by creating a CSV file. It was observed that the varieties not being matched to a color were mostly blends. No color was assigned to such grape varieties; therefore, NaN values were kept as such.

## Normalization

As a first step in our loading preparation, since there was a set of data that was repeated for different reviews, a normalization exercise was introduced in the project. Upon analyzing the data, we found the data can be divided into the sets of tables pictured below.

Once the final unified dataset was ready, QuickDBD was used to create the ERD (Figure 1. ERD for Wine review Database)

The normalization process was as follows:

1. Dataframes were created out of the main dataframe for each of the master table defined in the ERD.
2. The columns that were needed in each table were taken from main dataframe and new dataframe was created using those columns with the duplicated data deleted.
3. The index was reset to create primary key for table after incrementing value by 1 to start primary key for the table from 1
4. The unique values for the dataframe was checked against unique values for the dataset from main dataframe to ensure extra records were added or data removed.
5. There were some records with no taster data/winery data and hence a record was created with null values in taster and winery dataframe.
6. To create foreign keys of master dataframe in our wine review dataframe the main dataframe was merged with individual master dataframe and the primary key in the master dataframe assigned to foreign key in wine review dataframe.

# Load

1. Created tables for each dataframe as per ERD (Figure 1. ERD for Wine review Database)
2. Used sqlalchemy to connect to database and inserted data into our tables from the dataframe.
3. A new database wine\_review\_db was created in postgresql
4. Created tables for each dataframe as per ERD
5. Used sqlalchemy to connect to database and inserted data into our tables from the dataframe

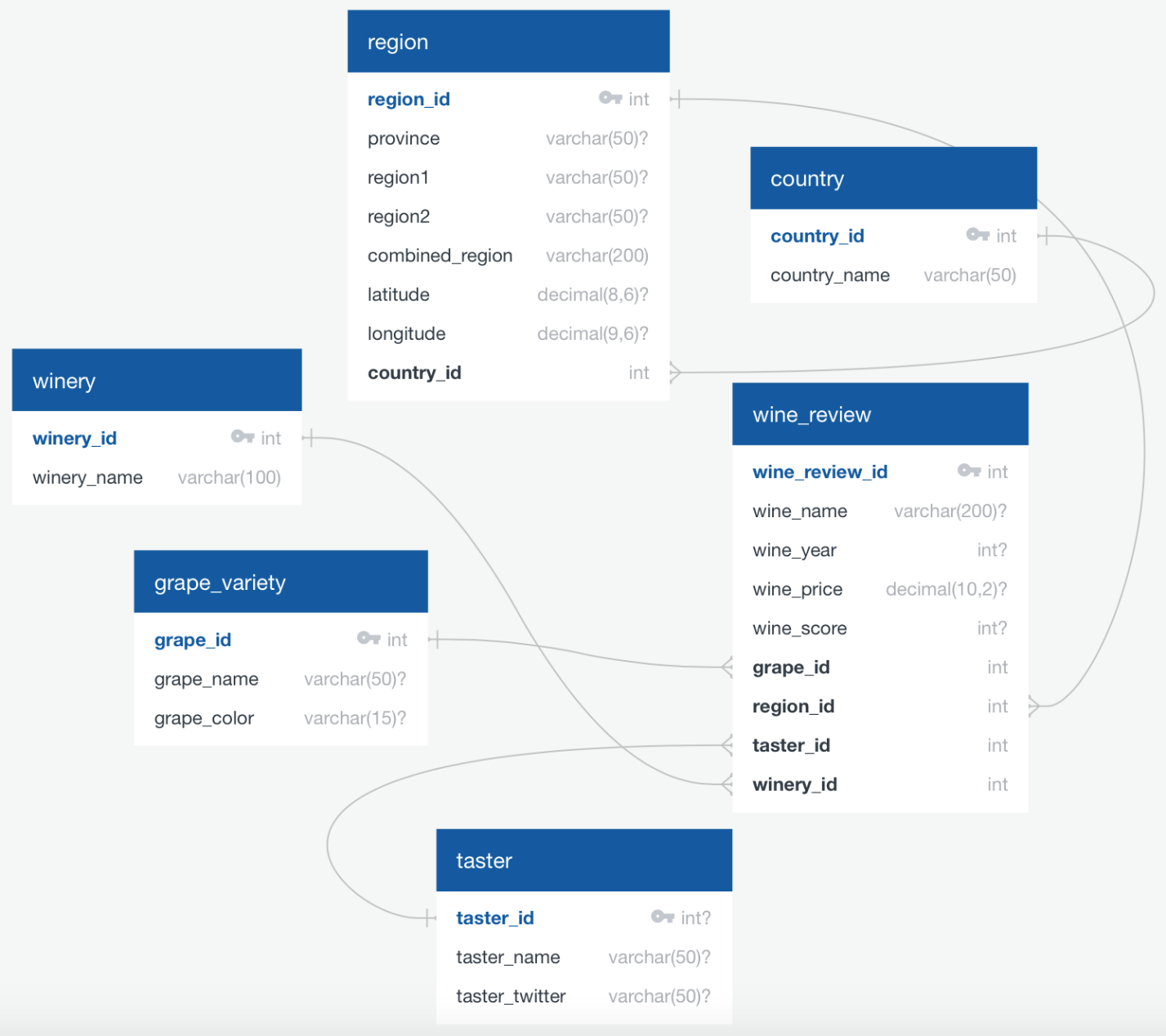


Figure 1. ERD for Wine review Database

## Resources:

1. Wine Review Data: <https://www.kaggle.com/zynicide/wine-reviews>
2. Geopy:   
   https://www.geeksforgeeks.org/how-to-find-longitude-and-latitude-for-a-list-of-regions-or-country-using-python/
3. Wine Color Wikepedia: <https://en.wikipedia.org/wiki/List_of_grape_varieties#Red_table_grapes>