**Wine**

Wine has evolved as part of life, culture and diet since time immemorial. As an enduring cultural symbol of fine life, the role of wine has evolved over time, changing from an important source of nutrition to a cultural complement to food and conviviality compatible with a healthy lifestyle.

In the western world today and for some decades now, wine & food go hand in hand. Most times wine is paired based on the food being served, more often than not when wine is the focus it is the other way around.  
  
The broadest way to categorize wines is red or white, sparking or not. When it comes to food pairing wines are identified by the acidity, its body & fruit, tannin etc. When sommeliers discuss wine the year of its make & the **place** it was made is equally if not more important. These are the factors for a bottle Merlot to range from $50 to $1000.

**ETL Project**  
  
Here we are sourcing data from Kaggle database(<https://www.kaggle.com/zynicide/wine-reviews>), we have taken three source files(csv, json and excel), describing country of origin, primary language, region, variety of wine, winery, price, point etc.

Global wine market was valued at approximately USD 302.02 billion in 2017, it has important economic implications for regions and countries in terms of production for both the local market and export (approximately 25% of the wine consumed was purchased outside the country where the wine was produced). While few countries are responsible for a large portion of the wine produced in the world (France 21%, Italy 20.4%, Spain 11.7%), in the past decades wine production has expanded substantially in many region of the world.

**ETL cycle consists of the following execution steps:**

1.    Cycle initiation  
2.    Build reference data  
3.    Extract (from sources)  
4.    Validate  
5.    Transform (clean, apply business rules, check for data integrity, create aggregates or disaggregates)  
6.    Stage (load into staging tables, if used)  
7.    Audit reports (for example, on compliance with business rules. Also, in case of failure, helps to diagnose/repair)  
8.    Publish (to target tables)  
9.    Archive

**Extraction:**

**csv datasource:** *'winemag-data\_first150k.csv'*

'country', 'description', 'designation', 'points', 'price', 'province',

'region\_1', 'region\_2', 'variety', 'winery'

**Json datasource:** *'winemag-data-130k-v2.json'*

'country', 'description', 'designation', 'points', 'price', 'province',

'region\_1', 'region\_2', 'taster\_name', 'taster\_twitter\_handle', 'title',

'variety', 'winery'

**Excel datasource:** *Wine\_Countries.xlsx*

'Country', 'Capital', 'Continent', 'Region', 'Primary Language',

'Secondary Language'

**Transformation:**

Steps taken for transformation:

1. Read source files and store in data frames
2. Select the columns of interest
3. Remove all the records with missing data and compare datatypes of columns to be joined.
4. Verify clean-up.
5. Join json and csv data frames
6. Read excel file and store in data frames
7. Transform the data so that it can be merged

a) Replacing column ‘country’ to ‘Country’ to support merge.

1. Merge the third file with previous files.

**Load:**

The load phase loads the data into the end target, which can be any data store including a simple delimited flat file or a data warehouse tables.  
  
As the load phase interacts with a database, the constraints defined in the database schema — as well as in triggers activated upon data load — apply (for example, uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.  
  
Here we are using Sqlalchemy is used to load table in our postgres database table. We create a database in postgres, create target table ‘wine\_country ’where the data used for analysis will be stored, define constraints.  
  
Connection to postgres is created in Python and final data frame with merge data is loaded to target table.

Target table is loaded to start analysis on wine database, as per requirement we can also decide on data load frequency.