Data Extraction Project : Wine Map

Barthelemy Charlier, Daniel Gagliardi, Adrien Goldszal

Objective : Collecting wine data with their precise location, such as coordinates, to be able to show them on a map.

Reasons/motivations : Our motive is quite simple: the wine industry is not only old but very sparse (at least in France/Europe) making it surprisingly complex to find complete and intuitive open source dataset of wine online with not only name and location but also features (taste, grapes,...).

Data Source 1 : The Vivino website (https://www.vivino.com/fr/). Vivino is a large website collecting thousands of up to date wines and their market prices from around the world. On the website, one can interact with filters and clickable cards for each wine. These wine card webpages often give details on the vineyard, grapes, region, wine style, alcohol content, price, allergens, and even food pairings.

Objective 1 : Scraping the webpages of vivino to collect the wine information to have a dataset of the sort :

Wine name	Vineyard	Grapes	Region Keywords	Wine Style	Alcohol content	Price

Dataset cleaning and preprocessing : Price needs to handle different currencies and be converted to a numeric data type, grapes are often lists that can be stored as JSON arrays of strings.

The region keywords are the biggest challenge. These region descriptors can be more or less vague. The objective being to map this information to coordinates, we need an additional data source. To do this, ideas entail combining fields like Vineyard, Region and Country into a single search string, removing generic non-location keywords like “Winery”, and standardizing terms such as Bourgogne/Burgundy.

Datasource 2 : The OpenStreetMap geocoding API Nominatim https://nominatim.org/
By cross referencing the extracted data from Vivino, we can find coordinates from the API. Main challenges will be formatting the data from the region keywords into parsable information for the API. Doing some form of recursive

Datasource 3 : the opendatawine website https://www.opendatawine.fr/ , which provides open source geospatial informations about french AOC labelled wine (name of village), we could also extract from that source GIS delimitations that go with a french wine name.

Assessing data quality : Random wine sampling with the coordinates for example on folium is a way to ensure the wines and their coordinates are in the correct locations.

Data storage solution : A relational database such as PostgreSQL would be a good fit for our structured data, and it ensures data integrity and avoids redundancy. PostgreSQL also has PostGIS, which is very useful for querying geospatial data. Data can directly then be manipulated through python packages like SLalchemy and GeoAlchemy2. If time permits, building a frontend for the database with a map would be a nice addition.

Further ideas : restricting wines for a specific region such as France, and using databases or APIs such as the SNCF API to suggest the shortest route from X to Y. This would give us a proper extra table for which we could do a relational database on.