Data integration and Data Profile for Eno-gastronomic Heritage Collection

Rachel Fanti Coelho Lima

Data Integration and Data Profiling Courses - 2021/2022

[1] Faculty of Computer Science, Free University of Bozen-Bolzano

Piazza Università, 1, 39100, Bolzano, BZ, Italy

rfanticoelholima@unibz.it

1. Contents

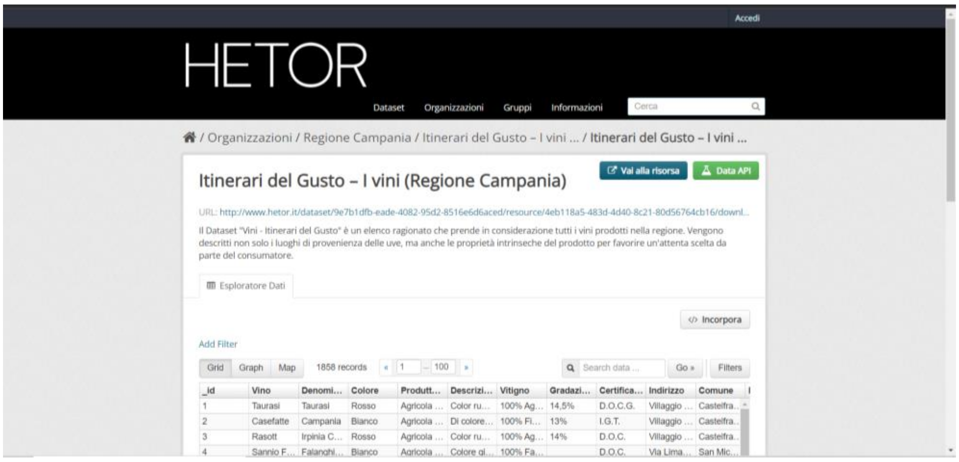
2. Introduction

This project is part of a larger project, which aims to integrate cultural, gastronomic, and natural data from the region of Campania in Italy. Specifically, this report will present analysis and methodologies of data profiling and data integration used for the integration of the Eno gastronomic data collection.

3. Description of the domain

According to Wikipedia, Italy occupies the first position in the list of the top wine-producing countries. Also in Campania, wine has long played an important role in the economy of the region.

The collection of data consists of a wider selection of wines produced in the region, including different types, made from different grapes. For each wine there is a description, considering the color, the alcohol content, and the organoleptic properties. Additionally, there are information about the certifications of these wines and the names of producers, their basic data and address.



/ Organizzazioni / Regione Campania / Itinerari del Gusto – I vini ... / Itinerari del Gusto – I vini ...

Itinerari del Gusto – I vini (Regione Campania)

URL: http://www.hetor.it/dataset/9e7b1dfb-eade-4082-95d2-8516e6d6aced/resource/4eb118a5-483d-4d40-8c21-80d56764cb16/downl...

Il Dataset "Vini - Itinerari del Gusto" è un elenco ragionato che prende in considerazione tutti i vini prodotti nella regione. Vengono descritti non solo i luoghi di provenienza delle uve, ma anche le proprietà intrinseche del prodotto per favorire un'attenta scelta da parte del consumatore.

_id	Vino	Denomi...	Colore	Produtt...	Descrizi...	Vitigno	Gradazi...	Certifica...	Indirizzo	Comune
1	Taurasi	Taurasi	Rosso	Agricola ...	Color ru...	100% Ag...	14,5%	D.O.C.G.	Villaggio ...	Castelfra...
2	Casefatte	Campania	Bianco	Agricola ...	Di colore...	100% Fi...	13%	I.G.T.	Villaggio ...	Castelfra...
3	Rasott	Irpinia C...	Rosso	Agricola ...	Color ru...	100% Ag...	14%	D.O.C.	Villaggio ...	Castelfra...
4	Sannio F...	Falanghi...	Bianco	Agricola ...	Colore gi...	100% Fa...		D.O.C.	Via Lima...	San Mic...

Figure 1 - Example of enogastronomic data source

4. Description of the selected data sources

The Eno-gastronomic Heritage Collection contain data from four data sources. In Table 1 is presented a description of each dataset, the number of registers, the format of them and the URL link for the ones that are available on the web. Furthermore, all data sources can be access on the google drive folder shared for the development of this project on page https://drive.google.com/drive/u/0/folders/1o-IKgdpY4W7xto9fdoFjkR1Ysv_PYpxx.

Data Source	# Registers	Type	URL
Wines-2	1858	CSV	-
Wines-3	1867	CSV	http://www.hetor.it/dataset/9e7b1dfb-eade-4082-95d2-8516e6d6aced/resource/4eb118a5-483d-4d40-8c21-80d56764cb16 (1858 records
Wines2016	251	CSV	-
Wines2016-2	971	CSV	-

Table 1 – Data sources considered

The attribute names of each data source are listed below.

Attributes		
Wines-2/ Wines-3	Wines2016/ Wines2016-2	Relevant attributes
Vino	Vino	Numerazione elenco ufficiale
Denominazione	Denominazione	Denominazione
Colore	Colore	Categoria di appartenenza
Produttore	Tipologia	Descrizione
Descrizione Organolettica	Produttore	Località - CSV
Vitigno	Descrizione Organolettica	Provincia
Gradazione Alcolica	Vitigno	Certificazione
Certificazione	Gradazione Alcolica	Tipo di certificazione
Indirizzo	Certificazione	Sigla certificazione per esteso
Comune	Indirizzo	Periodo tipico
Provincia	Comune	Stagione
Geolocalizzazione	Provincia	Tipologia
URL	Geolocalizzazione	Supplier category
Telefono	URL	Produttori
Fonti esterne	Telefono	Production manager
	Fonti esterne	Trasformazione prodotti
		Consorzi/ Presidium of belonging
		Contact person for producers
		Head of the garrison
		Eventi collegati
		Dove gustarlo
		Info tratte da
		Geolocalizzazione

Table 2 – Attributes of each data source

In the last column was included a list of relevant attributes (considered in Hetor dataset details). Some of them, are not present in the wine datasets, not being considered in the mappings; however, they will be considered in the UML Diagram and in the OWL2QL ontology since this can facilitate the integration of these data in the future.

Initially, equivalences between the attribute names among the tables were checked (Table 3 – Comparison between attributes of each data sourceTable 3). It is important to reinforce that although the attribute names are similar, the records may not be. This verification will be carried out later. In addition, some attributes with different names were highlighted to indicate that they maybe can be equivalent and need to be better analyzed further.

Attributes			
Wines-2/ Wines-3/Wines2016/ Wines2016-2		Relevant attributes	
English	Italian	English	Italian
		Official numerical code	Numerazione elenco ufficiale
Wine Code	Vino		
Name	Denominazione	Name	Denominazione
		Category	Categoria di appartenenza
		Description	Descrizione
Organoleptic Description	Descrizione Organolettica		
Colour	Colore		
Type	Tipologia	Type	Tipologia
		Supplier category	Categoria fornitore
Producer	Produttore	Producers	Produttori
		Production manager	Presidio sostenuto da
		Products transformation	Trasformazione prodotti
		Consorzi/Pres.of belonging	Presidio di appartenenza
		Contact person for producers	Referente dei produttori
		Head of the garrison	Responsabile del Presidio
Grape variety	Vitigno	Typical period	Periodo tipico
		Season	Stagione
		Certification	Certificazione
Alcohol content	Gradazione Alcolica	Certification type	Tipo di certificazione
Certification	Certificazione	Extended certification code	Sigla certificaz. per esteso
		Consortium	Consorzi
		Related events	Eventi collegati
		Tasting location	Dove gustarlo
		Location	Località - CSV
Address	Indirizzo		
Municipality	Comune	Province	Provincia
Province	Provincia	Geolocalization	Geolocalizzazione
Geolocalization	Geolocalizzazione		
Link	URL		
Phone number	Telefono	Source	Info tratte da
External sources	Fonti esterne		

Table 3 – Comparison between attributes of each data source

As the data sources are in Italian, a translation of the names for English is provided (Table 4).

Attributes	
English	Italian
Official numerical code	Numerazione elenco ufficiale
Wine Code	Vino
Name	Denominazione
Category	Categoria di appartenenza
Description	Descrizione
Organoleptic Description	Descrizione Organolettica
Colour	Colore
Type	Tipologia
Supplier category	Categoria fornitore
Producer	Produttore
Production manager	Presidio sostenuto da
Products transformation	Trasformazione prodotti
Consorzi/Pres.of belonging	Presidio di appartenenza
Contact person for producers	Referente dei produttori
Head of the garrison	Responsabile del Presidio
Grape variety	Vitigno
Typical period	Periodo tipico
Season	Stagione
Alcohol content	Gradazione Alcolica
Certification	Certificazione
Certification type	Tipo di certificazione
Extended certification code	Sigla certificazione per esteso
Consortium	Consorzi
Related events	Eventi collegati
Tasting location	Dove gustarlo
Address/ Location	Indirizzo/Località
Municipality	Comune
Province	Provincia
Geolocalization	Geolocalizzazione
Link	URL
Phone numbe	Telefono
Source	Info tratte da
External sources	Fonti esterne

Table 4 – Attributes translation English – Italian

5. Data Profile

5.1 Data preprocessing

Data preprocessing is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data.

Before do the preprocessing, some analyses were performed using Metanome.

5.1.1. Analysing the data sources - Metanome algorithms

The Metanome project is a project at HPI in cooperation with the Qatar Computing Reserach Institute (QCRI). Metanome provides a fresh view on data profiling by developing and integrating efficient algorithms into a common tool, expanding on the functionality of data profiling, and addressing performance and scalability issues for Big Data (https://hpi.de/naumann/projects/data-profiling-and-analytics/metanome-data-profiling.html, s.d.).

For the project, some data profiling methods will be used to efficiently analyze the data sources. The tables of the relational databases will be scanned to derive metadata, such as data types and value patterns, completeness and uniqueness of columns, keys and foreign keys, functional dependencies, and normalization. The algorithms used are listed below.

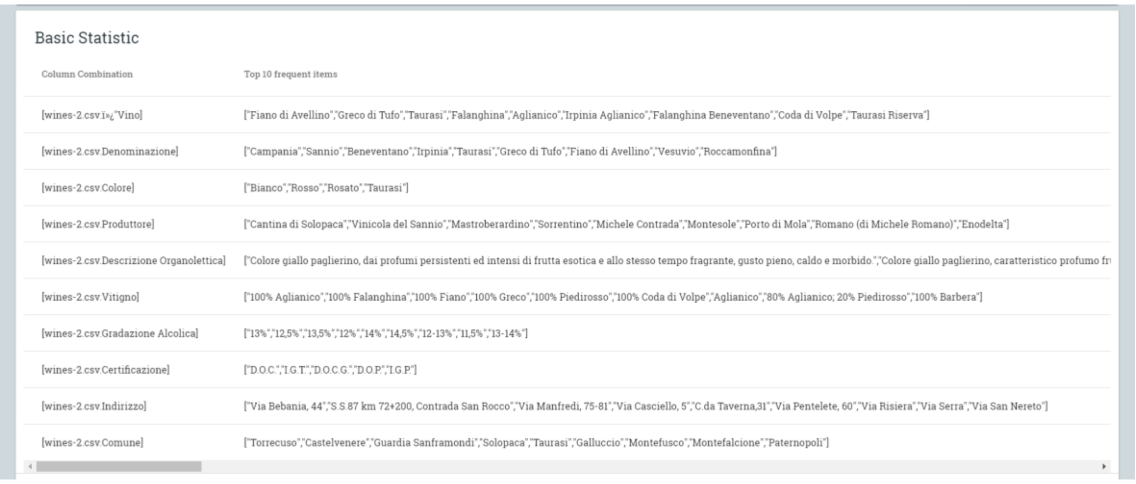
Algorithm	Description
SCDP-1.2-SNAPSHOT	Basic Statistics discovery
UCCs (DUCC or HYUCC)	Unique column combination discovery (Random Walk-based UCC discovery or Hybrid Sampling- and Lattice-Traversal-based UCC discovery)
Tane	Lattice Traversal-based FD discovery
Normalize	Schema normalization into BCNF using HyFD

Table 5– Algorithms for data profiling considered

Some examples of report are presented on the following.

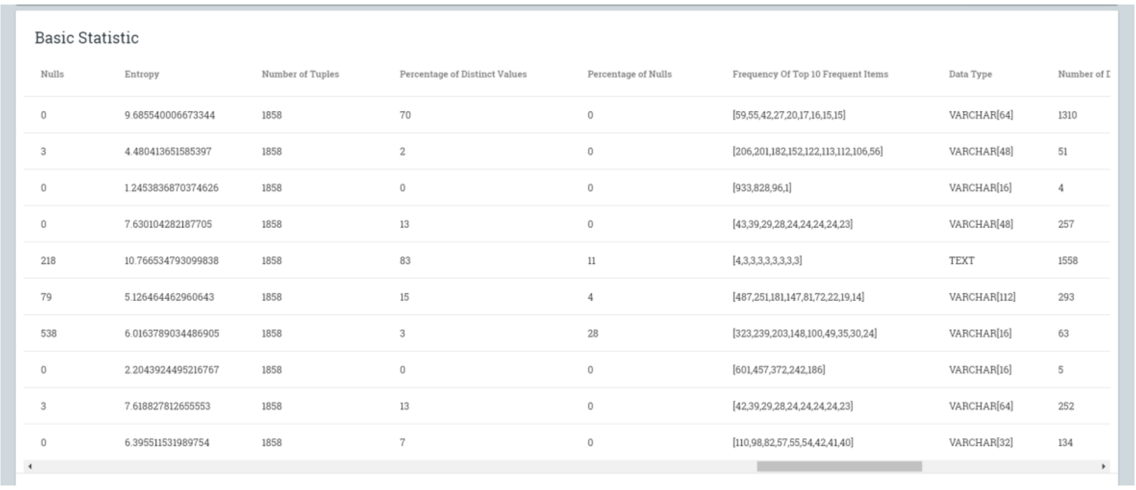
SCDP-1.2-SNAPSHOT – Example

Wine – 2 Dataset



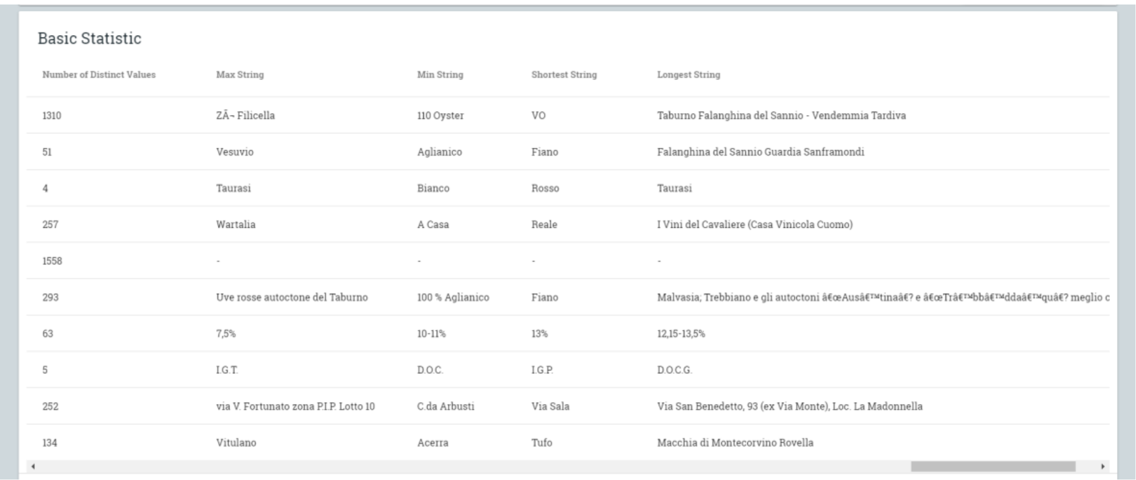
Basic Statistic

Column Combination	Top 10 frequent items
[wines-2.csv.ï»¿"Vino]	["Fiano di Avellino","Greco di Tufo","Taurasi","Falanghina","Aglianico","Irpinia Aglianico","Falanghina Beneventano","Coda di Volpe","Taurasi Riserva"]
[wines-2.csv.Denominazione]	["Campania","Sannio","Beneventano","Irpinia","Taurasi","Greco di Tufo","Fiano di Avellino","Vesuvio","Roccamonfina"]
[wines-2.csv.Colore]	["Bianco","Rosso","Rosato","Taurasi"]
[wines-2.csv.Produttore]	["Cantina di Solopaca","Vinicola del Sannio","Mastroberardino","Sorrentino","Michele Contrada","Montesole","Porto di Mola","Romano (di Michele Romano)","Enodelta"]
[wines-2.csv.Descrizione Organolettica]	["Colore giallo paglierino, dai profumi persistenti ed intensi di frutta esotica e allo stesso tempo fragrante, gusto pieno, caldo e morbido.","Colore giallo paglierino, caratteristico profumo fr
[wines-2.csv.Vitigno]	["100% Aglianico","100% Falanghina","100% Fiano","100% Greco","100% Piedirosso","100% Coda di Volpe","Aglianico","80% Aglianico; 20% Piedirosso","100% Barbera"]
[wines-2.csv.Gradazione Alcolica]	["13%","12,5%","13,5%","12%","14%","14,5%","12-13%","11,5%","13-14%"]
[wines-2.csv.Certificazione]	["D.O.C.","I.G.T.","D.O.C.G.","D.O.P.","I.G.P."]
[wines-2.csv.Indirizzo]	["Via Bebania, 44","S.S.87 km 72+200, Contrada San Rocco","Via Manfredi, 75-81","Via Casciello, 5","C.da Taverna,31","Via Pentelete, 60","Via Risiera","Via Serra","Via San Nereto"]
[wines-2.csv.Comune]	["Torrecuso","Castelvenere","Guardia Sanframondi","Solopaca","Taurasi","Galluccio","Montefusco","Montefalcione","Paternopoli"]



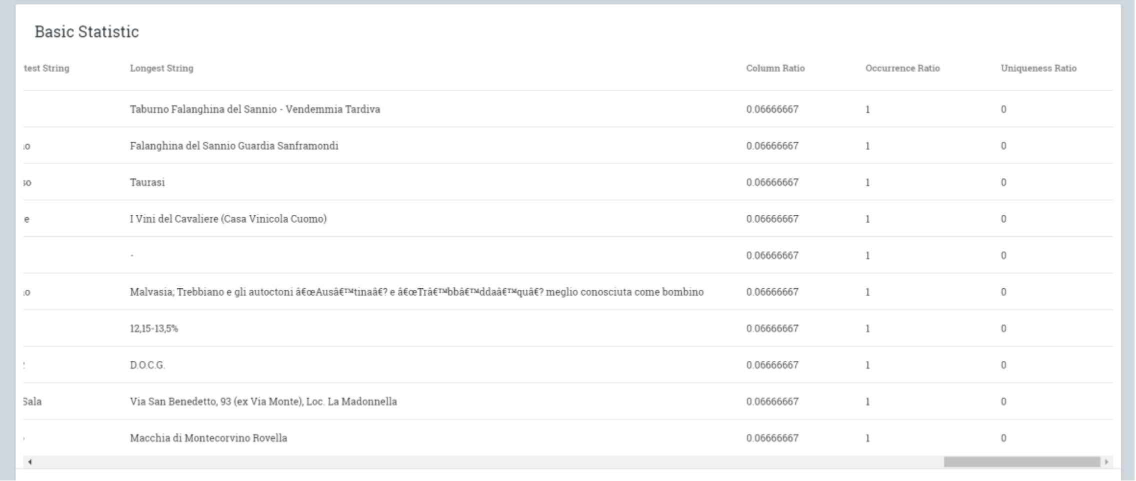
Basic Statistic

Nulls	Entropy	Number of Tuples	Percentage of Distinct Values	Percentage of Nulls	Frequency Of Top 10 Frequent Items	Data Type	Number of D
0	9.685540006673344	1858	70	0	[59,55,42,27,20,17,16,15,15]	VARCHAR[64]	1310
3	4.480413651585397	1858	2	0	[206,201,182,152,122,113,112,106,56]	VARCHAR[48]	51
0	1.2453836870374626	1858	0	0	[933,828,96,1]	VARCHAR[16]	4
0	7.630104282187705	1858	13	0	[43,39,29,28,24,24,24,24,23]	VARCHAR[48]	257
218	10.766534793099838	1858	83	11	[4,3,3,3,3,3,3,3,3]	TEXT	1558
79	5.126464462960643	1858	15	4	[487,251,181,147,81,72,22,19,14]	VARCHAR[112]	293
538	6.0163789034486905	1858	3	28	[323,239,203,148,100,49,35,30,24]	VARCHAR[16]	63
0	2.2043924495216767	1858	0	0	[601,457,372,242,186]	VARCHAR[16]	5
3	7.618827812655553	1858	13	0	[42,39,29,28,24,24,24,24,23]	VARCHAR[64]	252
0	6.395511531989754	1858	7	0	[110,98,82,57,55,54,42,41,40]	VARCHAR[32]	134



Basic Statistic

Number of Distinct Values	Max String	Min String	Shortest String	Longest String
1310	ZÃ¬ Filicella	110 Oyster	VO	Taburno Falanghina del Sannio - Vendemmia Tardiva
51	Vesuvio	Aglianico	Fiano	Falanghina del Sannio Guardia Sanframondi
4	Taurasi	Bianco	Rosso	Taurasi
257	Wartalia	A Casa	Reale	I Vini del Cavaliere (Casa Vinicola Cuomo)
1558	-	-	-	-
293	Uve rosse autoctone del Taburno	100 % Aglianico	Fiano	Malvasia; Trebbiano e gli autoctoni â€œAusâ€™tinaâ€? e â€œTrâ€™bbâ€™ddaâ€™quâ€? meglio c
63	7,5%	10-11%	13%	12,15-13,5%
5	I.G.T.	D.O.C.	I.G.P.	D.O.C.G.
252	via V. Fortunato zona P.I.P. Lotto 10	C.da Arbusti	Via Sala	Via San Benedetto, 93 (ex Via Monte), Loc. La Madonnella
134	Vitulano	Acerra	Tufo	Macchia di Montecorvino Rovella



Basic Statistic

test String	Longest String	Column Ratio	Occurrence Ratio	Uniqueness Ratio
	Taburno Falanghina del Sannio - Vendemmia Tardiva	0.06666667	1	0
o	Falanghina del Sannio Guardia Sanframondi	0.06666667	1	0
so	Taurasi	0.06666667	1	0
e	I Vini del Cavaliere (Casa Vinicola Cuomo)	0.06666667	1	0
	-	0.06666667	1	0
o	Malvasia; Trebbiano e gli autoctoni â€œAusâ€™tinaâ€? e â€œTrâ€™bbâ€™ddaâ€™quâ€? meglio conosciuta come bombino	0.06666667	1	0
	12,15-13,5%	0.06666667	1	0
:	D.O.C.G.	0.06666667	1	0
Sala	Via San Benedetto, 93 (ex Via Monte), Loc. La Madonnella	0.06666667	1	0
	Macchia di Montecorvino Rovella	0.06666667	1	0

Figure 2 – Example of results of Basic Statistics Discovery Algorithm (SCDP/Metanome)

These analses were realized for all data sets. Among the statistics, it is possible to observe the percentage of distinct values and the percentage of nulls for each attribute.

UCCs

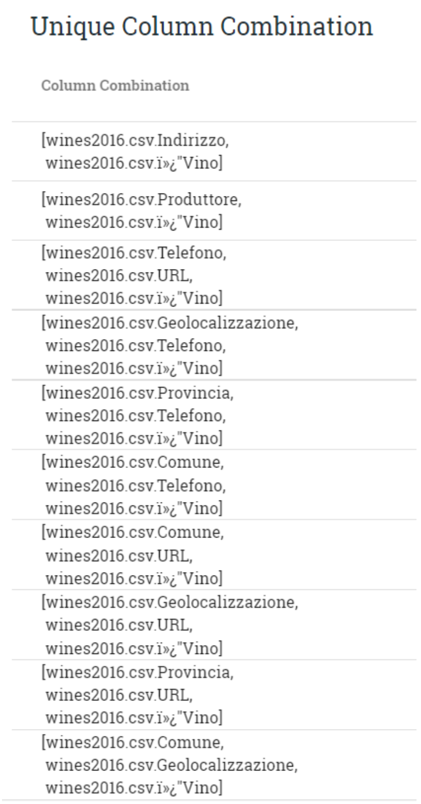
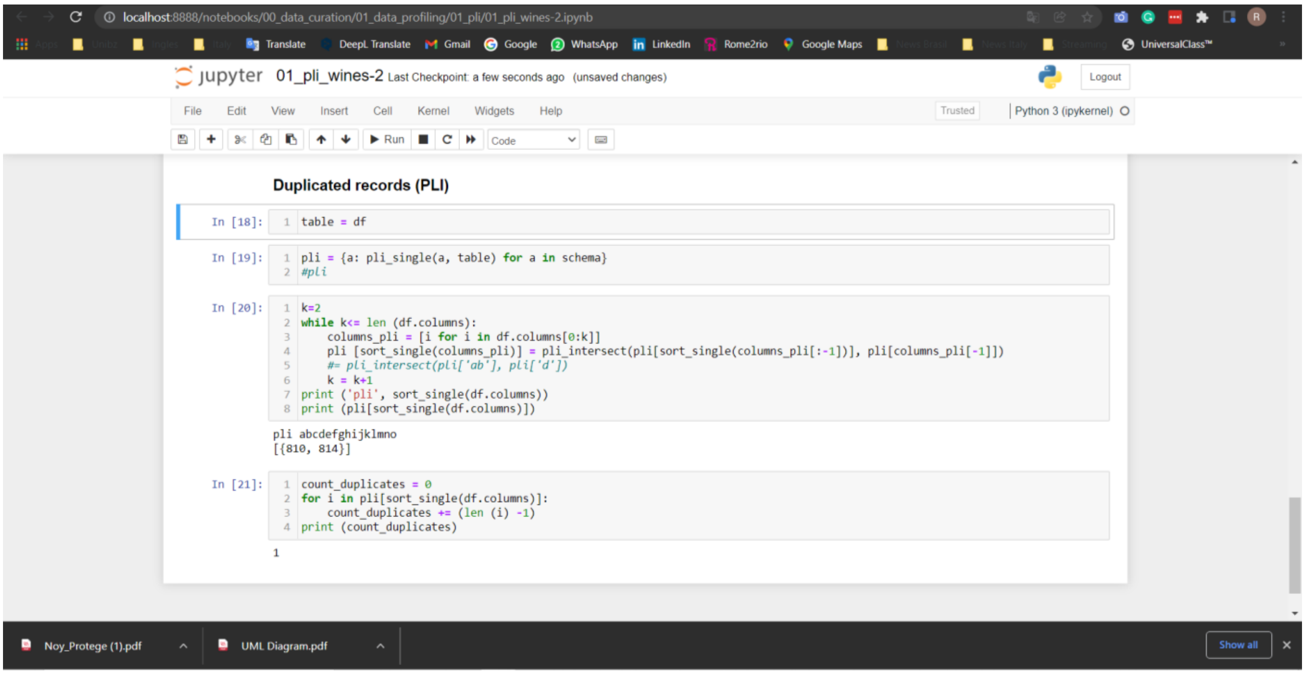
Unique column combination			
Wines-2	Wines-3	Wines2016	Wines2016-2
No results found!	**No results found!**	 Unique Column Combination Column Combination [wines2016.csv.Indirizzo, wines2016.csv.ï»¿"Vino] [wines2016.csv.Produttore, wines2016.csv.ï»¿"Vino] [wines2016.csv.Telefono, wines2016.csv.URL, wines2016.csv.ï»¿"Vino] [wines2016.csv.Geolocalizzazione, wines2016.csv.Telefono, wines2016.csv.ï»¿"Vino] [wines2016.csv.Provincia, wines2016.csv.Telefono, wines2016.csv.ï»¿"Vino] [wines2016.csv.Comune, wines2016.csv.Telefono, wines2016.csv.ï»¿"Vino] [wines2016.csv.Comune, wines2016.csv.URL, wines2016.csv.ï»¿"Vino] [wines2016.csv.Geolocalizzazione, wines2016.csv.URL, wines2016.csv.ï»¿"Vino] [wines2016.csv.Provincia, wines2016.csv.URL, wines2016.csv.ï»¿"Vino] [wines2016.csv.Comune, wines2016.csv.Geolocalizzazione, wines2016.csv.ï»¿"Vino]	**No results found!**

Table 6 – Example of results of Unique Column Combination Discovery Algorithm (DUCC/Metanome)

As shown in the table, no results were found in some datasets after running the algorithm of UCCs in Metanome, probably due to duplicated columns. To identify and remove these duplicates from each data source, was developed an algorithm based on PLI (position location information) in Jupyter Notebook, considering all attributes and calculating the number of repeated records.



localhost:8888/notebooks/00_data_curation/01_data_profiling/01_pli/01_pli_wines-2.ipynb
jupyter 01_pli_wines-2 Last Checkpoint: a few seconds ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Noy_Protege (1).pdf
UML Diagram.pdf
Show all

Duplicated records (PLI)

```
In [18]: table = df
```

```
In [19]: pli = {a: pli_single(a, table) for a in schema}
         #pli
```

```
In [20]: k=2
         while k<= len (df.columns):
             columns_pli = [i for i in df.columns[0:k]]
             pli [sort_single(columns_pli)] = pli_intersect(pli[sort_single(columns_pli[:-1])], pli[columns_pli[-1]])
             #= pli_intersect(pli['ab'], pli['d'])
             k = k+1
         print ('pli', sort_single(df.columns))
         print (pli[sort_single(df.columns)])

pli abcdefghijklmno
[{810, 814}]
```

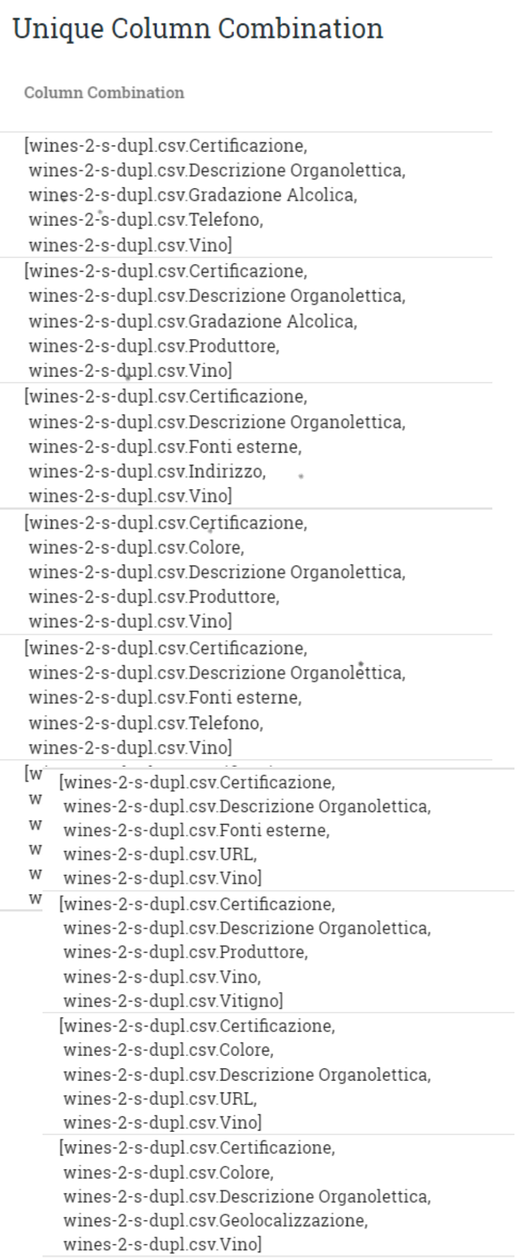
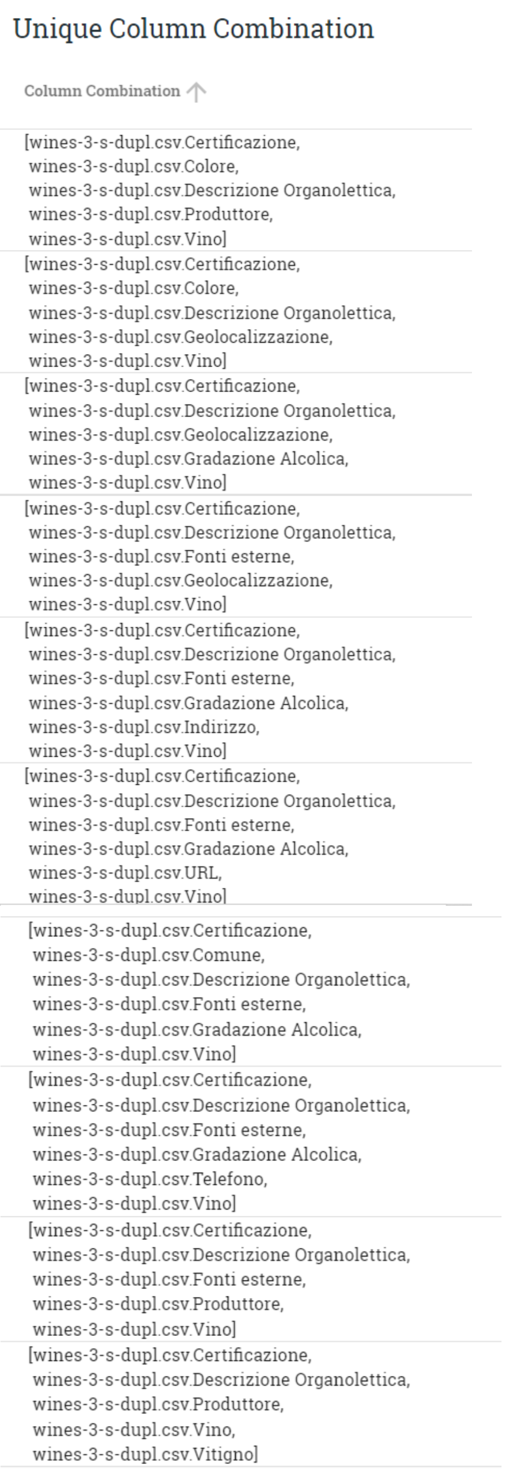
```
In [21]: count_duplicates = 0
         for i in pli[sort_single(df.columns)]:
             count_duplicates += (len (i) -1)
         print (count_duplicates)

1
```

Figure 3 – Eliminating duplicated records using algorithm based on PLI

Duplicates				
Wines-2	Wines-3	Wines2016	Wines2016-2	Consolidate
1 tuple removed	1 tuple removed	No duplicates	3 tuples removed	1089 tuples removed

Table 6 – Number of duplicated records found in each dataset

UCCs without duplicates

Unique column combination	
Wines-2	Wines-3
 Unique Column Combination Column Combination [wines-2-s-dupl.csv.Certificazione, wines-2-s-dupl.csv.Descrizione Organolettica, wines-2-s-dupl.csv.Gradazione Alcolica, wines-2-s-dupl.csv.Telefono, wines-2-s-dupl.csv.Vino] [wines-2-s-dupl.csv.Certificazione, wines-2-s-dupl.csv.Descrizione Organolettica, wines-2-s-dupl.csv.Gradazione Alcolica, wines-2-s-dupl.csv.Produttore, wines-2-s-dupl.csv.Vino] [wines-2-s-dupl.csv.Certificazione, wines-2-s-dupl.csv.Descrizione Organolettica, wines-2-s-dupl.csv.Fonti esterne, wines-2-s-dupl.csv.Indirizzo, wines-2-s-dupl.csv.Vino] [wines-2-s-dupl.csv.Certificazione, wines-2-s-dupl.csv.Colore, wines-2-s-dupl.csv.Descrizione Organolettica, wines-2-s-dupl.csv.Produttore, wines-2-s-dupl.csv.Vino] [wines-2-s-dupl.csv.Certificazione, wines-2-s-dupl.csv.Descrizione Organolettica, wines-2-s-dupl.csv.Fonti esterne, wines-2-s-dupl.csv.Telefono, wines-2-s-dupl.csv.Vino] [wines-2-s-dupl.csv.Certificazione, wines-2-s-dupl.csv.Descrizione Organolettica, wines-2-s-dupl.csv.Fonti esterne, wines-2-s-dupl.csv.URL, wines-2-s-dupl.csv.Vino] [wines-2-s-dupl.csv.Certificazione, wines-2-s-dupl.csv.Descrizione Organolettica, wines-2-s-dupl.csv.Produttore, wines-2-s-dupl.csv.Vino, wines-2-s-dupl.csv.Vitigno] [wines-2-s-dupl.csv.Certificazione, wines-2-s-dupl.csv.Colore, wines-2-s-dupl.csv.Descrizione Organolettica, wines-2-s-dupl.csv.URL, wines-2-s-dupl.csv.Vino] [wines-2-s-dupl.csv.Certificazione, wines-2-s-dupl.csv.Colore, wines-2-s-dupl.csv.Descrizione Organolettica, wines-2-s-dupl.csv.Geolocalizzazione, wines-2-s-dupl.csv.Vino]	 Unique Column Combination Column Combination ↑ [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Colore, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Produttore, wines-3-s-dupl.csv.Vino] [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Colore, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Geolocalizzazione, wines-3-s-dupl.csv.Vino] [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Geolocalizzazione, wines-3-s-dupl.csv.Gradazione Alcolica, wines-3-s-dupl.csv.Vino] [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Fonti esterne, wines-3-s-dupl.csv.Geolocalizzazione, wines-3-s-dupl.csv.Vino] [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Fonti esterne, wines-3-s-dupl.csv.Gradazione Alcolica, wines-3-s-dupl.csv.Indirizzo, wines-3-s-dupl.csv.Vino] [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Fonti esterne, wines-3-s-dupl.csv.Gradazione Alcolica, wines-3-s-dupl.csv.URL, wines-3-s-dupl.csv.Vino] [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Comune, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Fonti esterne, wines-3-s-dupl.csv.Gradazione Alcolica, wines-3-s-dupl.csv.Vino] [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Fonti esterne, wines-3-s-dupl.csv.Gradazione Alcolica, wines-3-s-dupl.csv.Telefono, wines-3-s-dupl.csv.Vino] [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Fonti esterne, wines-3-s-dupl.csv.Produttore, wines-3-s-dupl.csv.Vino] [wines-3-s-dupl.csv.Certificazione, wines-3-s-dupl.csv.Descrizione Organolettica, wines-3-s-dupl.csv.Produttore, wines-3-s-dupl.csv.Vino, wines-3-s-dupl.csv.Vitigno]

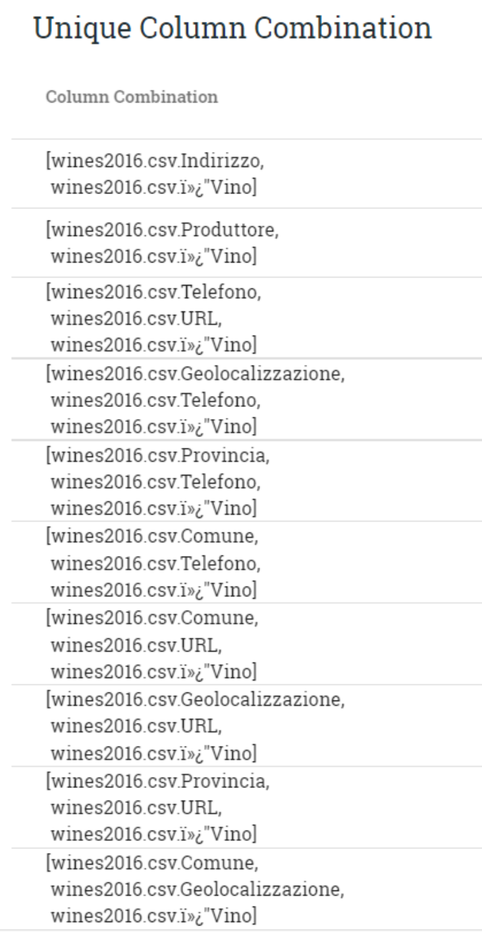
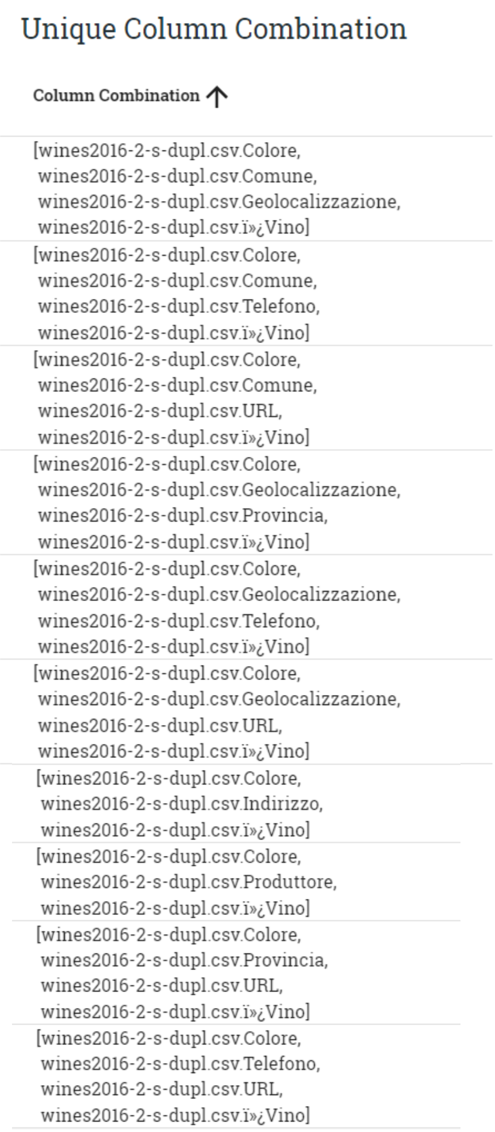
Unique column combination	
Wines2016	Wines2016-2
 Unique Column Combination Column Combination [wines2016.csv.Indirizzo, wines2016.csv.ï»¿"Vino] [wines2016.csv.Produttore, wines2016.csv.ï»¿"Vino] [wines2016.csv.Telefono, wines2016.csv.URL, wines2016.csv.ï»¿"Vino] [wines2016.csv.Geolocalizzazione, wines2016.csv.Telefono, wines2016.csv.ï»¿"Vino] [wines2016.csv.Provincia, wines2016.csv.Telefono, wines2016.csv.ï»¿"Vino] [wines2016.csv.Comune, wines2016.csv.Telefono, wines2016.csv.ï»¿"Vino] [wines2016.csv.Comune, wines2016.csv.URL, wines2016.csv.ï»¿"Vino] [wines2016.csv.Geolocalizzazione, wines2016.csv.URL, wines2016.csv.ï»¿"Vino] [wines2016.csv.Provincia, wines2016.csv.URL, wines2016.csv.ï»¿"Vino] [wines2016.csv.Comune, wines2016.csv.Geolocalizzazione, wines2016.csv.ï»¿"Vino]	 Unique Column Combination Column Combination ↑ [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Comune, wines2016-2-s-dupl.csv.Geolocalizzazione, wines2016-2-s-dupl.csv.ï»¿Vino] [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Comune, wines2016-2-s-dupl.csv.Telefono, wines2016-2-s-dupl.csv.ï»¿Vino] [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Comune, wines2016-2-s-dupl.csv.URL, wines2016-2-s-dupl.csv.ï»¿Vino] [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Geolocalizzazione, wines2016-2-s-dupl.csv.Provincia, wines2016-2-s-dupl.csv.ï»¿Vino] [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Geolocalizzazione, wines2016-2-s-dupl.csv.Telefono, wines2016-2-s-dupl.csv.ï»¿Vino] [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Geolocalizzazione, wines2016-2-s-dupl.csv.URL, wines2016-2-s-dupl.csv.ï»¿Vino] [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Indirizzo, wines2016-2-s-dupl.csv.ï»¿Vino] [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Produttore, wines2016-2-s-dupl.csv.ï»¿Vino] [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Provincia, wines2016-2-s-dupl.csv.URL, wines2016-2-s-dupl.csv.ï»¿Vino] [wines2016-2-s-dupl.csv.Colore, wines2016-2-s-dupl.csv.Telefono, wines2016-2-s-dupl.csv.URL, wines2016-2-s-dupl.csv.ï»¿Vino]

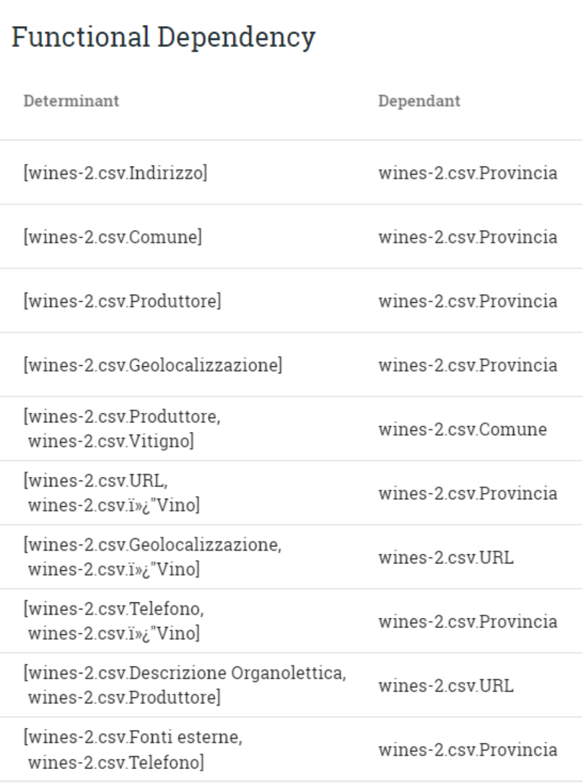
Table 7 – Example of results of Unique Column Combination Discovery Algorithm (DUCC/Metanome) after remove duplicates

In the table above we realized some problems in the data. For example, the attribute organoleptic description for the wines-2 dataset is present in all UCCs. However, there should be UCCs without this feature, probably, there are rows presents in the dataset that represents the same concept. Note that only identical data have been removed, but it is likely to have similar data, with small differences in writing. This will be analyzed further.

Tane - Examples

Functional Dependencies

Wines-2



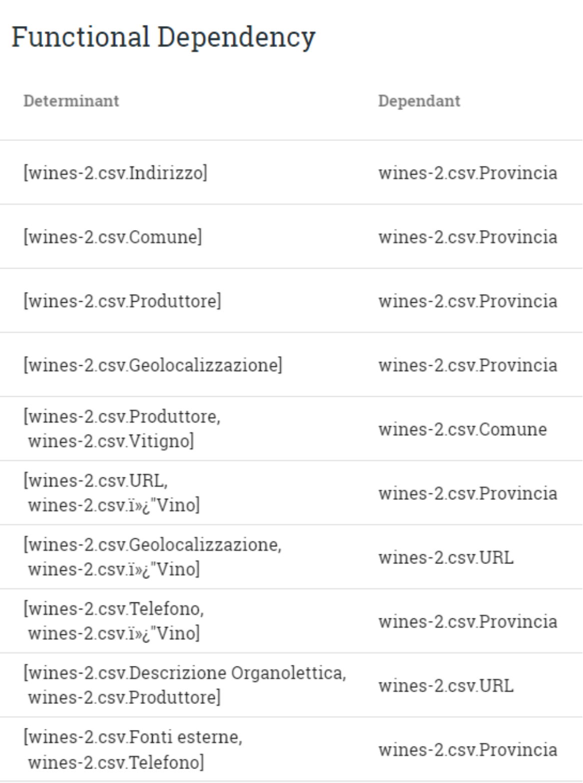
Functional Dependency

Determinant	Dependant
[wines-2.csv.Indirizzo]	wines-2.csv.Provincia
[wines-2.csv.Comune]	wines-2.csv.Provincia
[wines-2.csv.Produttore]	wines-2.csv.Provincia
[wines-2.csv.Geolocalizzazione]	wines-2.csv.Provincia
[wines-2.csv.Produttore, wines-2.csv.Vitigno]	wines-2.csv.Comune
[wines-2.csv.URL, wines-2.csv.ï»¿"Vino]	wines-2.csv.Provincia
[wines-2.csv.Geolocalizzazione, wines-2.csv.ï»¿"Vino]	wines-2.csv.URL
[wines-2.csv.Telefono, wines-2.csv.ï»¿"Vino]	wines-2.csv.Provincia
[wines-2.csv.Descrizione Organolettica, wines-2.csv.Produttore]	wines-2.csv.URL
[wines-2.csv.Fonti esterne, wines-2.csv.Telefono]	wines-2.csv.Provincia

…

(255 FDs)

Wines-3



Functional Dependency

Determinant	Dependant
[wines-2.csv.Indirizzo]	wines-2.csv.Provincia
[wines-2.csv.Comune]	wines-2.csv.Provincia
[wines-2.csv.Produttore]	wines-2.csv.Provincia
[wines-2.csv.Geolocalizzazione]	wines-2.csv.Provincia
[wines-2.csv.Produttore, wines-2.csv.Vitigno]	wines-2.csv.Comune
[wines-2.csv.URL, wines-2.csv.ï»¿"Vino]	wines-2.csv.Provincia
[wines-2.csv.Geolocalizzazione, wines-2.csv.ï»¿"Vino]	wines-2.csv.URL
[wines-2.csv.Telefono, wines-2.csv.ï»¿"Vino]	wines-2.csv.Provincia
[wines-2.csv.Descrizione Organolettica, wines-2.csv.Produttore]	wines-2.csv.URL
[wines-2.csv.Fonti esterne, wines-2.csv.Telefono]	wines-2.csv.Provincia

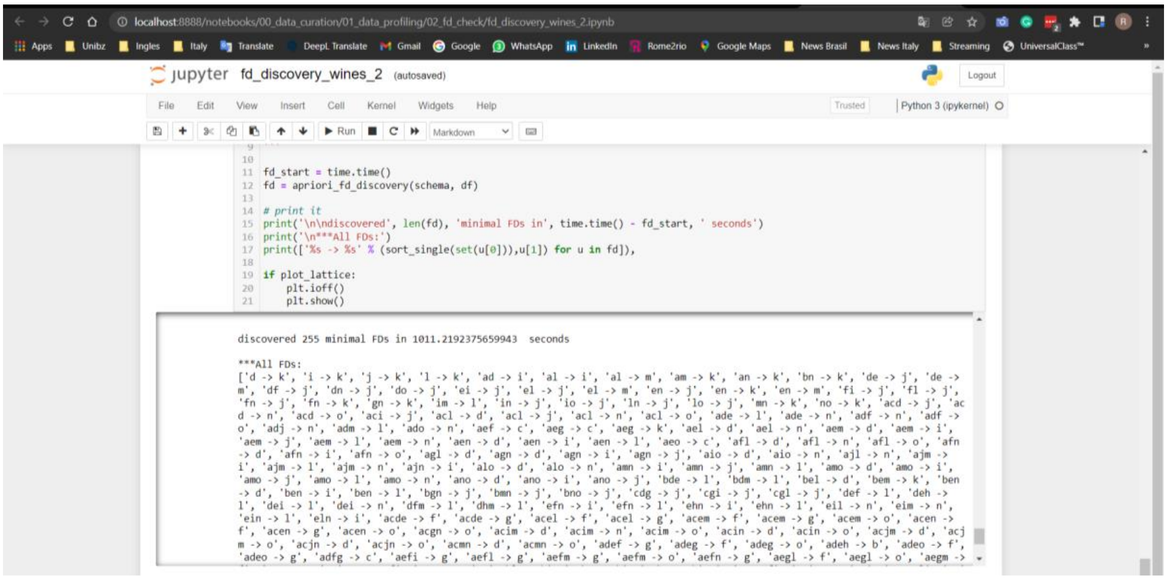
…

(191 FDs)

Functional Dependencies			
Wines2016		Wines2016-2	
Functional Dependency		Functional Dependency	
Determinant	Dependant	Determinant	Dependant
[wines2016.csv.Indirizzo]	wines2016.csv.Fonti esterne	[wines2016-2.csv.Indirizzo]	wines2016-2.csv.Provincia
[wines2016.csv.Produttore]	wines2016.csv.Fonti esterne	[wines2016-2.csv.Produttore]	wines2016-2.csv.Comune
[wines2016.csv.ï»¿"Vino]	wines2016.csv.Colore	[wines2016-2.csv.Indirizzo]	wines2016-2.csv.Comune
[wines2016.csv.Indirizzo]	wines2016.csv.Telefono	[wines2016-2.csv.Produttore]	wines2016-2.csv.Provincia
[wines2016.csv.Geolocalizzazione]	wines2016.csv.Fonti esterne	[wines2016-2.csv.Comune]	wines2016-2.csv.Provincia
[wines2016.csv.Indirizzo]	wines2016.csv.Comune	[wines2016-2.csv.Telefono, wines2016-2.csv.ï»¿"Vino]	wines2016-2.csv.Fonti esterne
[wines2016.csv.Indirizzo]	wines2016.csv.Geolocalizzazione	[wines2016-2.csv.Indirizzo, wines2016-2.csv.ï»¿"Vino]	wines2016-2.csv.Telefono
[wines2016.csv.Produttore]	wines2016.csv.Geolocalizzazione	[wines2016-2.csv.Produttore, wines2016-2.csv.ï»¿"Vino]	wines2016-2.csv.Telefono
[wines2016.csv.ï»¿"Vino]	wines2016.csv.Tipologia	[wines2016-2.csv.Produttore, wines2016-2.csv.ï»¿"Vino]	wines2016-2.csv.Certificazione
[wines2016.csv.Telefono]	wines2016.csv.Fonti esterne	[wines2016-2.csv.Descrizione Organolettica, wines2016-2.csv.Telefono]	wines2016-2.csv.Fonti esterne
... (268 FDs)		... (377 FDs)	

Table 8 – Example of results of Lattice Traversal-based FD discovery (Tane/Metanome)

The results of TANE for each dataset were compared to the results of the Functional Dependency Discovery algorithm developed in Jupyter Notebook.


localhost:8888/notebooks/00_data_curation/01_data_profiling/02_fd_check/fd_discovery_wines_2.ipynb
jupyter fd_discovery_wines_2 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown

```

fd_start = time.time()
fd = apriori_fd_discovery(schema, df)

# print it
print('\n\ndiscovered', len(fd), 'minimal FDs in', time.time() - fd_start, ' seconds')
print('\n***All FDs:')
print(['%s -> %s' % (sort_single(set(u[0])),u[1]) for u in fd]),

if plot_lattice:
    plt.ioff()
    plt.show()
```

```
discovered 255 minimal FDs in 1011.2192375659943  seconds

***All FDs:
['d -> k', 'i -> k', 'j -> k', 'l -> k', 'ad -> i', 'al -> i', 'al -> m', 'am -> k', 'an -> k', 'bn -> k', 'de -> j', 'de ->
m', 'df -> j', 'dn -> j', 'do -> j', 'ei -> j', 'el -> j', 'el -> m', 'en -> j', 'en -> k', 'en -> m', 'fi -> j', 'fl -> j',
'fn -> j', 'fn -> k', 'gn -> k', 'im -> l', 'in -> j', 'io -> j', 'ln -> j', 'lo -> j', 'mn -> k', 'no -> k', 'acd -> j', 'ac
d -> n', 'acd -> o', 'aci -> j', 'acl -> d', 'acl -> j', 'acl -> n', 'acl -> o', 'ade -> l', 'ade -> n', 'adf -> n', 'adf ->
o', 'adj -> n', 'adm -> l', 'ado -> n', 'aef -> c', 'aeg -> c', 'aeg -> k', 'ael -> d', 'ael -> n', 'aem -> d', 'aem -> i',
'aem -> j', 'aem -> l', 'aem -> n', 'aen -> d', 'aen -> i', 'aen -> l', 'aeo -> c', 'afl -> d', 'afl -> n', 'afl -> o', 'afn
-> d', 'afn -> i', 'afn -> o', 'agl -> d', 'agn -> d', 'agn -> i', 'agn -> j', 'aio -> d', 'aio -> n', 'ajl -> n', 'ajm ->
i', 'ajm -> l', 'ajm -> n', 'ajn -> i', 'alo -> d', 'alo -> n', 'amn -> i', 'amn -> j', 'amn -> l', 'amo -> d', 'amo -> i',
'amo -> j', 'amo -> l', 'amo -> n', 'ano -> d', 'ano -> i', 'ano -> j', 'bde -> l', 'bdm -> l', 'bel -> d', 'bem -> k', 'ben
-> d', 'ben -> i', 'ben -> l', 'bgn -> j', 'bmn -> j', 'bno -> j', 'cdg -> j', 'cgi -> j', 'cgl -> j', 'def -> l', 'deh ->
l', 'dei -> l', 'dei -> n', 'dfm -> l', 'dhm -> l', 'efn -> i', 'efn -> l', 'ehn -> i', 'ehn -> l', 'eil -> n', 'eim -> n',
'ein -> l', 'eln -> i', 'acde -> f', 'acde -> g', 'acel -> f', 'acel -> g', 'acem -> f', 'acem -> g', 'acem -> o', 'acen ->
f', 'acen -> g', 'acen -> o', 'acgn -> o', 'acim -> d', 'acim -> n', 'acim -> o', 'acin -> d', 'acin -> o', 'acjm -> d', 'acj
m -> o', 'acjn -> d', 'acjn -> o', 'acmn -> d', 'acmn -> o', 'adef -> g', 'adeg -> f', 'adeg -> o', 'adeh -> b', 'adeo -> f',
'adeo -> g', 'adfg -> c', 'aefi -> g', 'aefl -> g', 'aefm -> g', 'aefm -> o', 'aefn -> g', 'aegl -> f', 'aegl -> o', 'aegm ->
```

Figure 4 – Functional Dependency Discovery algorithm developed

# of FDs				
	Wines-2	Wines-3	Wines2016	Wines2016-2
TANE	255	191	268	377
Developed algorithm	255	191	258	377

Table 9 – Duplicated records

Both algorithms obtained the same results. Observe that for the table Wines2016, the Metanome algorithm resulted in 10 more FDs (Table 10– Differences between Metanome and algorithmTable 10). However, it is possible to see that these 10 FDs are redundant comparing with previous FDs already identified.

Differences	
Metanome	**FDs redundant**
1,11,13->5	11,13->10 1,10->5
1,11,14->5	11,14->15 10->15 1,10->5
1,12,14->5	12,14->16 10->16 1,10->5
1,12,14->10	12,14->16 10->16 1,10->10
1,13,14->5	13,14->15 10->15 1,10->5
1,11,15->5	11,15->10 1,10->5
1,13,15->5	13,15->10 1,10->5
1,12,15->5	1,10->5 10->15 1,15->5
1,12,15->10	1,10->10 10->15 1, 15>10
1,14,15->5	14,15->10 1,10->5

Table 10– Differences between Metanome and algorithm

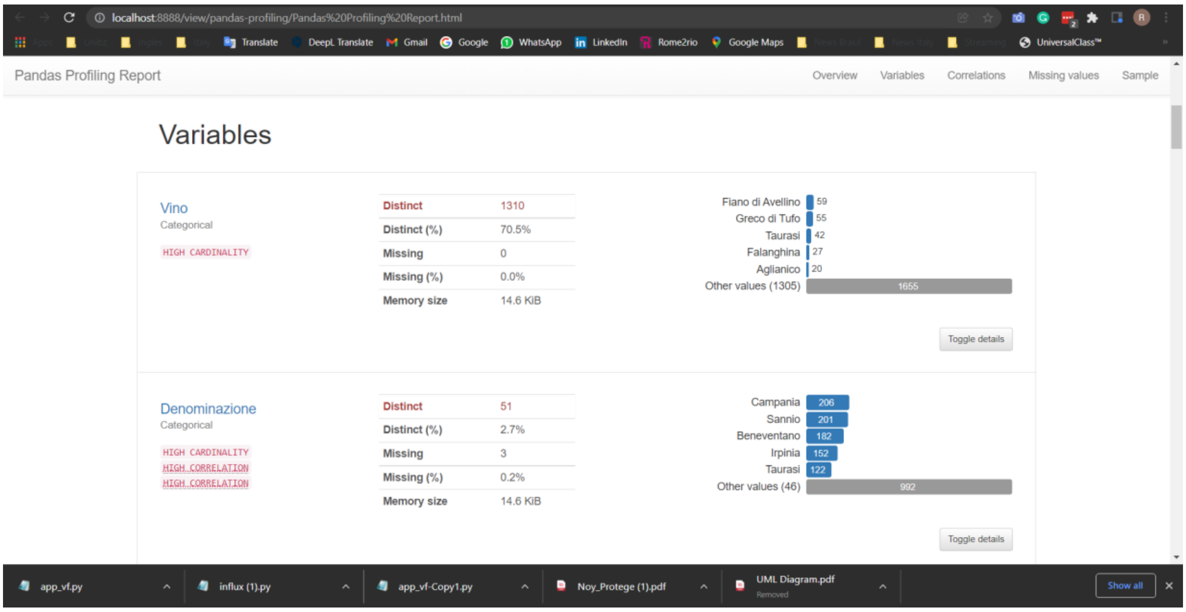
Normalization

Schema normalization into BCNF	
Wines-2	Wines--3
[wines-2.csv.Descrizione Organolettica, wines-2.csv.Vitigno, wines-2.csv.Denominazione, wines-2.csv.Gradazione Alcolica, wines-2.csv.ï»¿"Vino, wines-2.csv.Certificazione, wines-2.csv.Colore, wines-2.csv.Geolocalizzazione] PK* (All above) FK wines-2.csv.ï»¿"Vino, wines-2.csv.Colore, wines-2.csv.Geolocalizzazione	[wines-3.csv.Denominazione, wines-3.csv.Certificazione, wines-3.csv.Geolocalizzazione, wines-3.csv.ï»¿"Vino, wines-3.csv.Gradazione Alcolica, wines-3.csv.Colore, wines-3.csv.Descrizione Organolettica, wines-3.csv.Vitigno] PK* (All above) FK wines-3.csv.ï»¿"Vino, wines-3.csv.Colore, wines-3.csv.Geolocalizzazione
[wines-2.csv.ï»¿"Vino,* wines-2.csv.Colore,* wines-2.csv.Geolocalizzazione,* wines-2.csv.Produttore] PK * (3/4) FK wines-2.csv.ï»¿"Vino, wines-2.csv.Colore, wines-2.csv.Produttore	[wines-3.csv.Geolocalizzazione,* wines-3.csv.ï»¿"Vino,* wines-3.csv.Colore,* wines-3.csv.Produttore] PK * (3/4) FK wines-3.csv.ï»¿"Vino, wines-3.csv.Colore, wines-3.csv.Produttore
[wines-2.csv.Comune,* wines-2.csv.Telefono, wines-2.csv.ï»¿"Vino,* wines-2.csv.Produttore]* PK * (3/4) FK wines-2.csv.ï»¿"Vino, wines-2.csv.Produttore	[wines-3.csv.Comune, wines-3.csv.ï»¿"Vino,* wines-3.csv.Colore,* wines-3.csv.Produttore,* wines-3.csv.Fonti esterne] PK * (3/5) FK wines-3.csv.ï»¿"Vino, wines-3.csv.Produttore wines-3.csv.Comune
[wines-2.csv.ï»¿"Vino,* wines-2.csv.Geolocalizzazione,* wines-2.csv.URL] PK * (2/3) FK –	[wines-3.csv.Comune,* wines-3.csv.ï»¿"Vino,* wines-3.csv.Telefono, wines-3.csv.Produttore]* PK * (3/4) FK wines-3.csv.ï»¿"Vino, wines-3.csv.Produttore
[wines-2.csv.ï»¿"Vino,* wines-2.csv.Indirizzo, wines-2.csv.Produttore]* PK * (2/3) FK –	[wines-3.csv.Geolocalizzazione,* wines-3.csv.ï»¿"Vino,* wines-3.csv.URL] PK * (2/3) FK –
[wines-2.csv.Comune, wines-2.csv.Provincia] PK wines-2.csv.Comune FK –	[wines-3.csv.Indirizzo, wines-3.csv.ï»¿"Vino,* wines-3.csv.Produttore]* PK * (2/3) FK –
	[wines-3.csv.Comune, wines-3.csv.Provincia] PK wines-3.csv.Comune FK –

Schema normalization into BCNF	
Wines2016	Wines2016-2
[wines2016.csv.Colore, wines2016.csv.Produttore, wines2016.csv.Indirizzo, wines2016.csv.Certificazione, wines2016.csv.Descrizione Organolettica, wines2016.csv.ï»¿"Vino, wines2016.csv.Denominazione, wines2016.csv.URL, wines2016.csv.Comune, wines2016.csv.Provincia, wines2016.csv.Tipologia, wines2016.csv.Vitigno, wines2016.csv.Telefono, wines2016.csv.Gradazione Alcolica] PK * (All above) FK wines2016.csv.Produttore	[wines2016-2.csv.Vitigno, wines2016-2.csv.Denominazione, wines2016-2.csv.Tipologia, wines2016-2.csv.Fonti esterne, wines2016-2.csv.URL, wines2016-2.csv.Certificazione, wines2016-2.csv.Geolocalizzazione, wines2016-2.csv.ï»¿"Vino, wines2016-2.csv.Indirizzo, wines2016-2.csv.Descrizione Organolettica, wines2016-2.csv.Telefono, wines2016-2.csv.Colore, wines2016-2.csv.Produttore, wines2016-2.csv.Gradazione Alcolica] PK* (All above) FK wines2016.csv.Produttore
[wines2016.csv.Produttore, wines2016.csv.Geolocalizzazione, wines2016.csv.Fonti esterne] PK wines2016.csv.Produttore,	[wines2016-2.csv.Produttore, wines2016-2.csv.Provincia, wines2016-2.csv.Comune] PK wines2016.csv.Produttore

Table 11– Example of results of Schema normalization into BCNF (HyFD/Metanome) before cleaning

Other tools were used for understanding the data and help to define the cleaning and pre-processing of the data, such as Pandas Data Profiling Report.



Pandas Profiling Report

Variables

Vino
Categorical
HIGH CARDINALITY

Distinct	1310
Distinct (%)	70.5%
Missing	0
Missing (%)	0.0%
Memory size	14.6 KiB

Fiano di Avellino	59
Greco di Tufo	55
Taurasi	42
Falanghina	27
Aglianico	20
Other values (1305)	1655

Denominazione
Categorical
HIGH CARDINALITY
HIGH CORRELATION
HIGH CORRELATION

Distinct	51
Distinct (%)	2.7%
Missing	3
Missing (%)	0.2%
Memory size	14.6 KiB

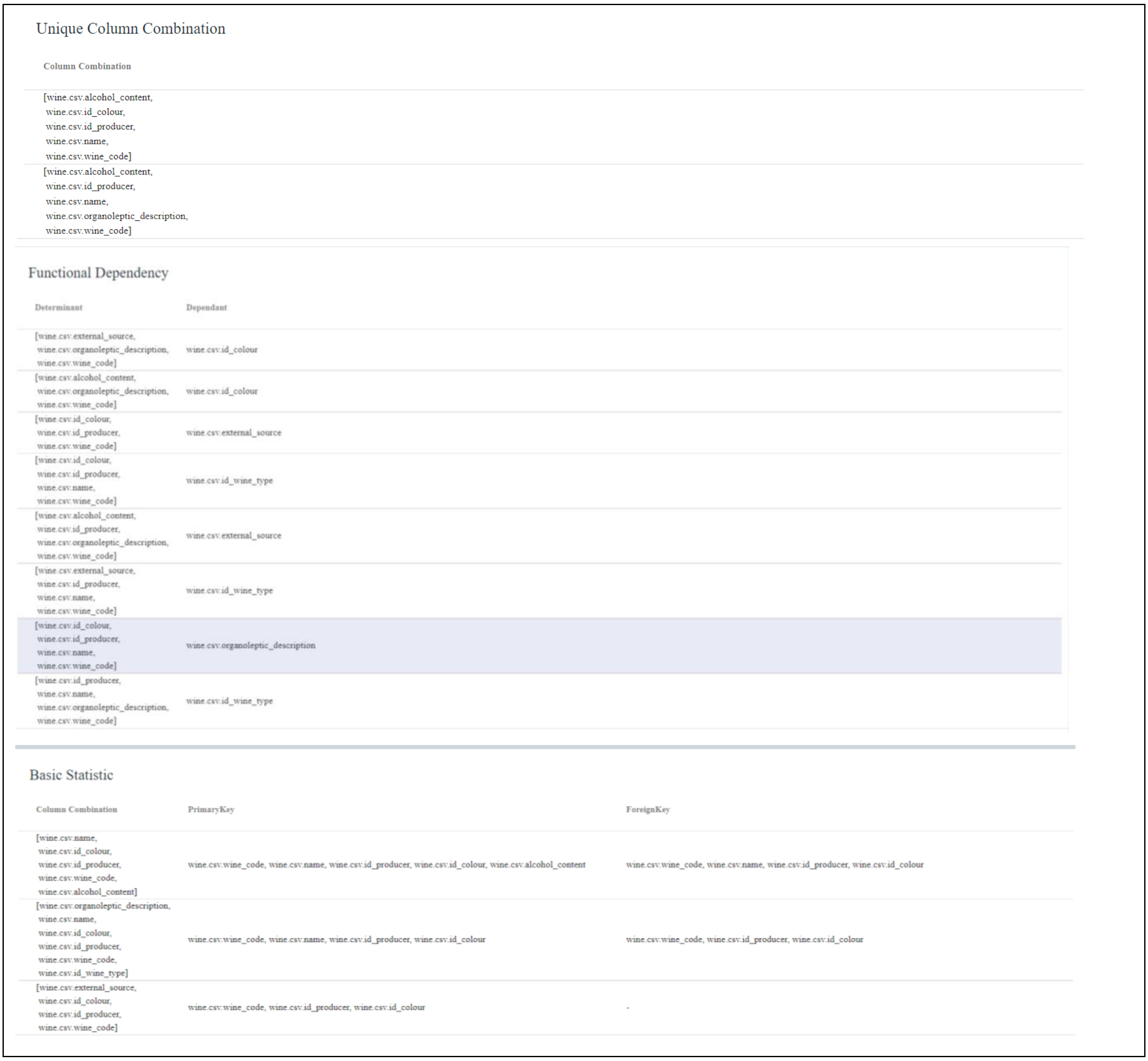
Campania	206
Sannio	201
Beneventano	182
Irpinia	152
Taurasi	122
Other values (46)	992

Figure 5 – Example of pandas data profiling report

The examples above illustrate analyses realized for better understanding of the data, however, most of them were reviewed and changed after cleaning the dataset.

At the beginning, it was difficult to understand which attributes represented a wine. For instance, for the same wine name there were several colours, types, grapes, etc. What were the actual attributes that identified the wine?

In the data sets there were many null values, incomplete data, and duplicates. In addition, there were small differences in the writing between rows that represented the same entity. For these reasons, these duplicates were not so easy to identify in the first moment. It was also only possible to understand unique combinations of columns after a good cleaning of the data. After that, it was observed that wine could be represented by the attributes wine_code, name, producer, colour and alcohol_content.



Unique Column Combination

Column Combination
[wine.csv.alcohol_content, wine.csv.id_colour, wine.csv.id_producer, wine.csv.name, wine.csv.wine_code]
[wine.csv.alcohol_content, wine.csv.id_producer, wine.csv.name, wine.csv.organoleptic_description, wine.csv.wine_code]

Functional Dependency

Determinant	Dependant
[wine.csv.external_source, wine.csv.organoleptic_description, wine.csv.wine_code]	wine.csv.id_colour
[wine.csv.alcohol_content, wine.csv.organoleptic_description, wine.csv.wine_code]	wine.csv.id_colour
[wine.csv.id_colour, wine.csv.id_producer, wine.csv.wine_code]	wine.csv.external_source
[wine.csv.id_colour, wine.csv.id_producer, wine.csv.name, wine.csv.wine_code]	wine.csv.id_wine_type
[wine.csv.alcohol_content, wine.csv.id_producer, wine.csv.organoleptic_description, wine.csv.wine_code]	wine.csv.external_source
[wine.csv.external_source, wine.csv.id_producer, wine.csv.name, wine.csv.wine_code]	wine.csv.id_wine_type
[wine.csv.id_colour, wine.csv.id_producer, wine.csv.name, wine.csv.wine_code]	wine.csv.organoleptic_description
[wine.csv.id_producer, wine.csv.name, wine.csv.organoleptic_description, wine.csv.wine_code]	wine.csv.id_wine_type

Basic Statistic

Column Combination	PrimaryKey	ForeignKey
[wine.csv.name, wine.csv.id_colour, wine.csv.id_producer, wine.csv.wine_code, wine.csv.alcohol_content]	wine.csv.wine_code, wine.csv.name, wine.csv.id_producer, wine.csv.id_colour, wine.csv.alcohol_content	wine.csv.wine_code, wine.csv.name, wine.csv.id_producer, wine.csv.id_colour
[wine.csv.organoleptic_description, wine.csv.name, wine.csv.id_colour, wine.csv.id_producer, wine.csv.wine_code, wine.csv.id_wine_type]	wine.csv.wine_code, wine.csv.name, wine.csv.id_producer, wine.csv.id_colour	wine.csv.wine_code, wine.csv.id_producer, wine.csv.id_colour
[wine.csv.external_source, wine.csv.id_colour, wine.csv.id_producer, wine.csv.wine_code]	wine.csv.wine_code, wine.csv.id_producer, wine.csv.id_colour	-

Table 12– Example of results of UCC, FD and Schema normalization after cleaning wine table

5.1.2.Cleaning and pre-processing of data

Some activities were performed as part of data pre-processing. The most important ones are listed below:

Actions for Cleaning and preprocessing the data

Number	Attribute	Actions
1	All	Removed duplicated records and special characters
2	Wine	Standardization of names
3	Name	Standardization of names
4	Colour	Standardization of names. Removed few wines name in this column.
5	Type	Standardization of names. Fill null values with data from the rows that have the same wine code/name/colour/alcohol content and producer
6	Producer	Standardization of names Standardization of names for producers with same address/ phone number (check on the internet)
7	Organoleptic description	Removed similar records, with only few differences in organoleptic description (few differences in text, but same description).
8	Grape variety	Transformation of this column into a relation n x n, with 3 columns: wine, percentage and grape variety. Eg: original cell: 33% Falanghina; 33% Fiano; 34% Greco Final: Id | Percentage | Grape variety 1 | 33% | Falanghina 1 | 33% | Fiano 1 | 33% | Greco Standardization of the names.
9	Alcohol content	Standardization of the numbers, correction of the number of decimal digits and format standardization
10	Certification	Standardization of the names, checking the acronyms in studies available on the internet and obtaining the meaning of the acronym
11	Address	Address segmentation into street, number and complement information. Fill null values with data from the rows that have the same producer/ geolocation.
12	Municipality	Standardization of the names, according to standard pattern in Eurostat (NUTS/ LAU). Matchings.
13	Province	Standardization of the names, according to standard pattern in Eurostat (NUTS/ LAU)
14	Geolocalization	Data standardization, keeping same number of decimal digits (6) Fill null values for cases that the same producers and address are in the dataset. Fill geolocation with data from the rows that have the same street/number.
15	URL	Data standardization (removed http://, https://, include in all "www", etc) Fill null values with data from the rows that have the same producer
16	Mobile Phone	Removed URLs that were in this column and placed in the correct one. Data standardization (removed characters that are not numbers, format standardization) Fill null values with data from the rows that have the same producer/address
17	External Source	Removed phone numbers that were in this column and placed in the correct one. Data standardization (removed http://, https://, include in all "www", etc)

Table 13– Actions for Cleaning and pre-processing the data

If such a base is made available in the future, it is suggested a better standardization of the names of wines, types, and grapes by a wine expert professional. As we do not have a depth knowledge in this area, some different names may be representing the same entity.

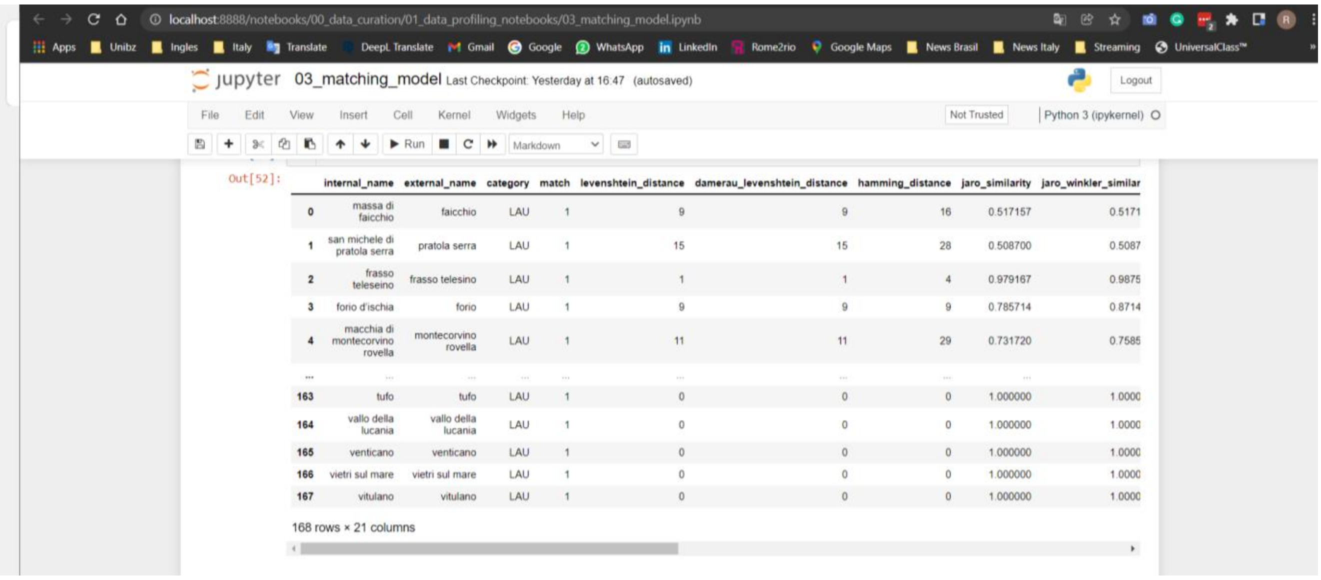
Also, some relations can be simplified if better understood. For instance, the relation wine-certificates (1 x n) maybe can become a relation 1 x 1, since there was only one register of wine with 2 certificates. Maybe, checking this directly with the producer, it is possible to remove one of them. The same, happens for alcohol-content. All wines (composed of the attributes wine_name, wine_code, colour, producer) have only one alcohol_content, except for 2 cases that there are 2 different alcohol content.

5.1.3. Matching

Although there were some attributes with values of the string type and it was necessary to standardize them, there was no need to use robust matching techniques to do it. During the cleanup, they were identified automatically, by grouping certain attributes and observing the relationships between them. However, some simplified matching solutions were tested.

For example, when standardizing the municipality reported in the data with the standard used by Eurostat, some names did not match. We tested to find it automatically, using Levenshtein distance and Jaccard score, and analyzed the results. Approximately 2/3 of the municipalities have been found using these solutions.

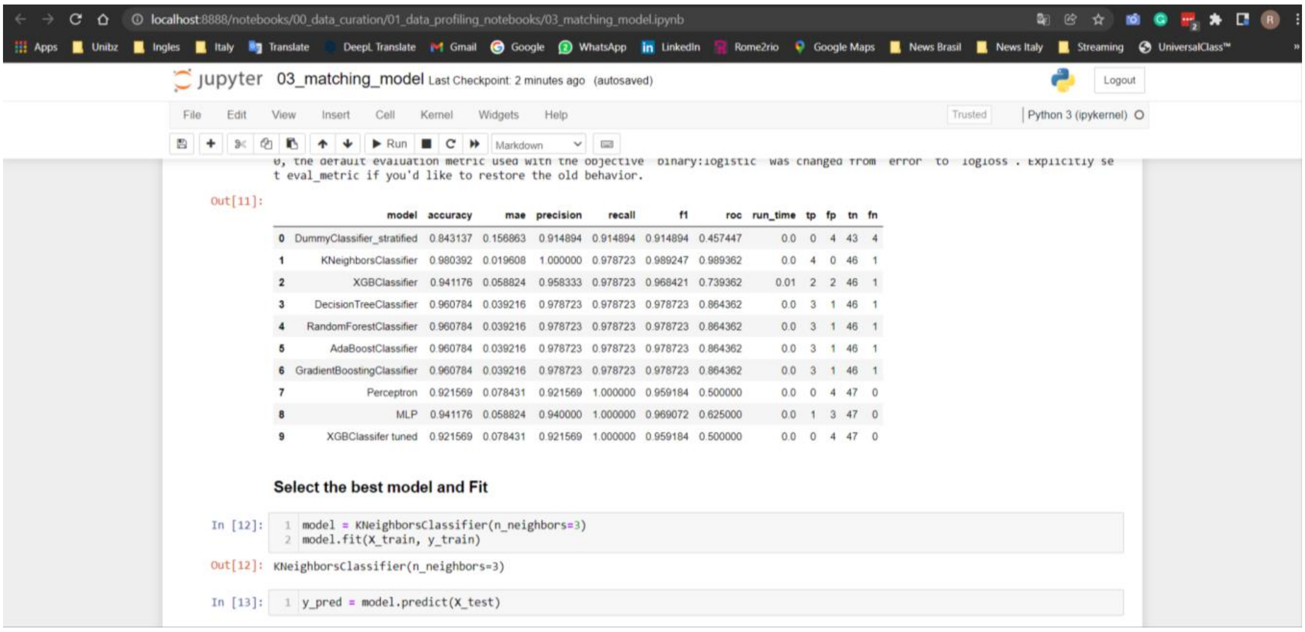
Subsequently, considering the errors obtained in this activity and the solutions, a model was trained and proposed to predict the correct municipality for the next time. The model considered a series of similarity and distance measures and was tested using several classifiers. The one with the best accuracy was chosen.



Out[52]:

	internal_name	external_name	category	match	levenshtein_distance	damerau_levenshtein_distance	hamming_distance	jaro_similarity	jaro_winkler_similar
0	massa di faicchio	faicchio	LAU	1	9	9	16	0.517157	0.5171
1	san michele di pratola serra	pratola serra	LAU	1	15	15	28	0.508700	0.5087
2	frasso teleseino	frasso telesino	LAU	1	1	1	4	0.979167	0.9875
3	forio d'ischia	forio	LAU	1	9	9	9	0.785714	0.8714
4	macchia di montecorvino rovella	montecorvino rovella	LAU	1	11	11	29	0.731720	0.7585
...	...	...	...	...	...	...	...	...	
163	tufo	tufo	LAU	1	0	0	0	1.000000	1.0000
164	vallo della lucania	vallo della lucania	LAU	1	0	0	0	1.000000	1.0000
165	venticano	venticano	LAU	1	0	0	0	1.000000	1.0000
166	vietri sul mare	vietri sul mare	LAU	1	0	0	0	1.000000	1.0000
167	vitulano	vitulano	LAU	1	0	0	0	1.000000	1.0000

168 rows × 21 columns



```
0, the default evaluation metric used with the objective  binary:logistic  was changed from  error  to  logloss . Explicitly se
t eval_metric if you'd like to restore the old behavior.
```

Out[11]:

	model	accuracy	mae	precision	recall	f1	roc	run_time	tp	fp	tn	fn
0	DummyClassifier_stratified	0.843137	0.156863	0.914894	0.914894	0.914894	0.457447	0.0	0	4	43	4
1	KNeighborsClassifier	0.980392	0.019608	1.000000	0.978723	0.989247	0.989362	0.0	4	0	46	1
2	XGBClassifier	0.941176	0.058824	0.958333	0.978723	0.968421	0.739362	0.01	2	2	46	1
3	DecisionTreeClassifier	0.960784	0.039216	0.978723	0.978723	0.978723	0.864362	0.0	3	1	46	1
4	RandomForestClassifier	0.960784	0.039216	0.978723	0.978723	0.978723	0.864362	0.0	3	1	46	1
5	AdaBoostClassifier	0.960784	0.039216	0.978723	0.978723	0.978723	0.864362	0.0	3	1	46	1
6	GradientBoostingClassifier	0.960784	0.039216	0.978723	0.978723	0.978723	0.864362	0.0	3	1	46	1
7	Perceptron	0.921569	0.078431	0.921569	1.000000	0.959184	0.500000	0.0	0	4	47	0
8	MLP	0.941176	0.058824	0.940000	1.000000	0.969072	0.625000	0.0	1	3	47	0
9	XGBClassifer tuned	0.921569	0.078431	0.921569	1.000000	0.959184	0.500000	0.0	0	4	47	0

Select the best model and Fit

In [12]:

```
model = KNeighborsClassifier(n_neighbors=3)
model.fit(X_train, y_train)
```

Out[12]: `KNeighborsClassifier(n_neighbors=3)`

In [13]:

```
y_pred = model.predict(X_test)
```

Figure 6 – Matching algorithm model for municipality

6. Data Integration

6.1 Ontology/mediated schema of the domain

Below, the domain is expressed in OWL2 QL, represented as a UML class diagram, with binary associations representing object properties, and class attributes representing data properties. A better resolution of the image is in the folder of the project, in the UML Diagram file.

As informed before, some relevant attributes are considered in the Ontology for allow the integration of these data in the future but are not present in the wine datasets. They were marked in the diagram with an “*”. Other suggestions were included, marked with “**”, which maybe can be considered in the future for populate this dataset.

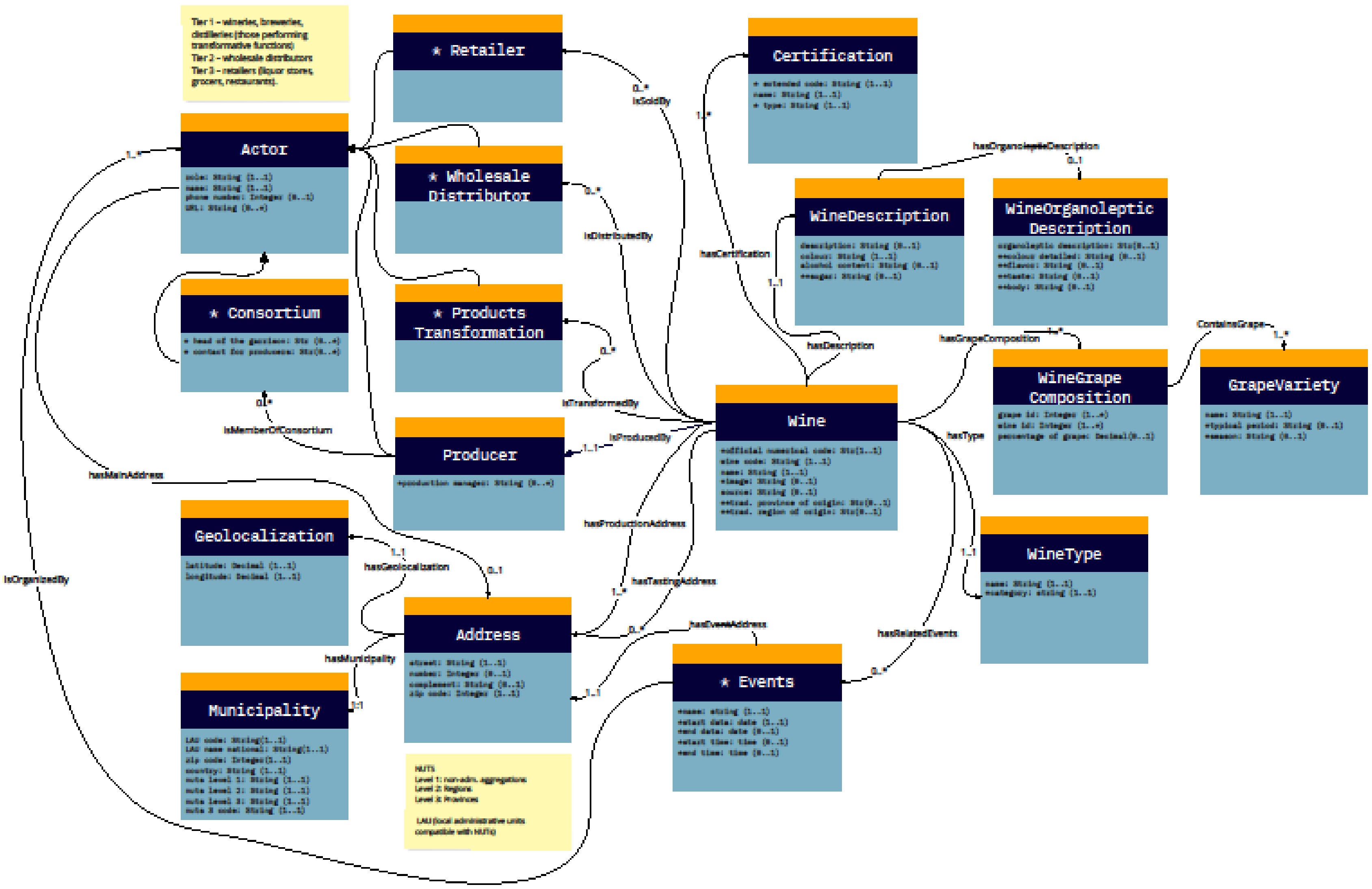


Figure 7 - Representation of the ontology/mediated schema using a UML class diagram

The formalization of the OWL2QL ontology is provided using the abstract syntax of Description Logics in the following order: set of classes, object properties, data properties, and then the ontology axioms that represent the above diagram.

Classes (12+5):

Wine, WineDescription, WineOrganolepticDescription, WineGrapeComposition, GrapeVariety, WineType, Certification, Actor, Retailer, WholesaleDistributor, ProductsTransformation, Producer, Consortium, Events, Address, Geolocalization, Municipality

Object Properties (19):

hasDescription, hasOrganolepticDescription, hasGrapeComposition, hasGrape, hasType, hasCertification, isSoldBy, isDistributedBy, isTransformedBy, isProducedBy, isMemberOfConsortium, hasAddress,
hasMainAddress, hasProductionAddress, hasTastingAddress, hasRelatedEvents, IsOrganizedBy, hasEventAddress, hasGeologalization, hasMunicipality

Data Properties (53):

The ontology contains one data property for each attribute of each class. To avoid conflicts between attributes with the same name in different classes, we qualify the name of the attribute by prefixing it with the initial letter(s) of the class name:

for class Wine (w): wOfficialNumericalCode, wWineCode, wName, wImage, wSource, wTraditionalProvinceOfOrigin, wTraditionalRegionOfOrigin
for class WineDescription (wd): wdDescription, wdColour, wdAlcoholContent, wdSugar
for class WineOrganolepticDescription (wod): wodOrganolepticDescription, wodColourDetailed, wodFlavor, wodTaste, wodBody
for class WineGrapeComposition (wgc): wgcGrapeId, wgcWineId, wgcPercentageOfGrape
for class GrapeVariety (gv): gvName, gvTypicalPeriod, gvSeason
for class WineType (wt): wtName, wtCategory
for class Certification (c): cExtendedCode, cName, cType
for class Actor (ac): acRole, acName, acPhoneNumber, acURL
for class Retailer (r): -
for class WholesaleDistributor (wsd): -
for class ProductsTransformation (pt): -
for class Producer (p): pProductionManager
for class Consortium (ct): ctHeadOfTheGarrison, ctContactForProducers
for class Events (e): eName, eStartDate, eEndDate, eStartTime, eEndTime
for class Address (a): aStreet, aNumber, aComplement, aZipCode
for class Geolocalization (g): gLatitude, gLongitude
for class Municipality (m): mLauCode, mLauNameNational, mzipCode, mCountry, mNUTSLevel1, mNUTSLevel2, mNUTSLevel3, mNUTS3Code

Ontology Axioms:

ISA and class hierarchies: Retailer $\sqsubseteq$ Actor, Wholesale Distributor $\sqsubseteq$ Actor, Products Transformation $\sqsubseteq$ Actor, Producer $\sqsubseteq$ Actor, Consortium $\sqsubseteq$ Actor

Property hierarchies:

hasMainAddress ⊑ hasAddress, hasProductionAddress ⊑ has Address, hasTastingAddress ⊑ has Address, hasEventAddress ⊑ hasAddress

Domain and range of object properties:

∃hasDescritpion ⊑ Wine	∃hasDescritpion− ⊑ WineDescritpion
∃hasOrganolepticDescription⊑ WineDescritpion	∃hasOrganolepticDescription−⊑ WineOrganolepticDescription
∃hasGrapeComposition ⊑ Wine	∃hasGrapeComposition− ⊑ WineGrapeComposition
∃hasGrape ⊑ WineGrapeComposition	∃hasGrape− ⊑ GrapeVariety
∃hasType ⊑ Wine	∃hasType− ⊑ WineType
∃hasCertification ⊑ Wine	∃hasCertification− ⊑ Certification
∃isSoldBy ⊑ Wine	∃isSoldBy− ⊑ Retailer
∃isDistributedBy ⊑ Wine	∃isDistributedBy −⊑ Wholesale Distributor
∃isTransformedBy ⊑ Wine	∃isTransformedBy− ⊑ Products Transformation
∃isProducedBy ⊑ Wine	∃isProducedBy− ⊑ Producer
∃isMemberOfConsotium ⊑ Producer	∃isMemberOfConsotium− ⊑ Consortium
*	∃hasAddress− ⊑ Address
∃hasMainAddress ⊑ Actor	∃hasMainAddress− ⊑ Address
∃hasProductionAddress ⊑ Wine	∃hasProductionAddress− ⊑ Address
∃hasTastingAddress ⊑ Wine	∃hasTastingAddress− ⊑ Address
∃hasRelatedEvents ⊑ Wine	∃hasRelatedEvents− ⊑ Events
∃IsOrganizedBy ⊑ Events	∃IsOrganizedBy− ⊑ Actor
∃hasEventAddress ⊑ Events	∃hasEventAddress− ⊑ Address
∃hasGeologalization ⊑ Address	∃hasGeologalization− ⊑ Geolocalization
∃hasMunicipality ⊑ Address	∃hasMunicipality− ⊑ Municipality

*In this case domain is a union among the classes (wine, actor, events), not being able to represent in Protégé.

Mandatory participation of classes to object properties (minimum multiplicity 1):

Wine ⊑ ∃hasDescritpion	WineDescritpion ⊑ ∃hasDescritpion−
-	WineOrganolepticDescription ⊑ ∃hasOrganolepticDescritpion−
Wine ⊑ ∃hasGrapeComposition	WineGrapeComposition ⊑ ∃hasGrapeComposition –
WineGrapeComposition ⊑ ∃hasGrape	GrapeVariety⊑ ∃hasGrape–
Wine ⊑ ∃hasType	WineType ⊑ ∃hasType−
Wine ⊑ ∃hasCertification	Certification ⊑ ∃hasCertification−
-	Retailer ⊑ ∃isSoldBy−
-	Wholesale Distributor ⊑ ∃isDistributedBy−
-	Products Transformation ⊑ ∃isTransformedBy−
Wine ⊑ ∃isProducedBy	Producer ⊑ ∃isProducedBy−
-	Consortium ⊑ ∃isMemberOfConsotium−
-	Address ⊑ ∃hasAddress−
Wine ⊑ ∃hasProductionAddress	Events ⊑ ∃hasRelatedEvents−
-	Actor ⊑ ∃IsOrganizedBy −
-	Geolocalization ⊑ ∃hasGeologalization−
Events ⊑ ∃IsOrganizedBy	Municipality ⊑ ∃hasMunicipality−
Events ⊑ ∃ hasEventAddress	
Address ⊑ ∃hasGeologalization	
Address ⊑ ∃hasMunicipality	

Domain of data properties (left column), their typing (central column), and mandatory value (right column):

∃wOfficialNumericalCode ⊑ Wine	∃wOfficialNumericalCode− ⊑ String	Wine ⊑ ∃wOfficialNumericalCode
∃wWineCode ⊑ Wine	∃wWineCode− ⊑ String	Wine ⊑ ∃wWineCode
∃wName ⊑ Wine	∃wName− ⊑ String	Wine ⊑ ∃wName
∃wImage ⊑ Wine	∃wImage− ⊑ String	WineDescription ⊑ ∃wdColour
∃wSource ⊑ Wine	∃wSource− ⊑ String	WineGrapeComp* ⊑ ∃wgcGrapeId
∃wTraditionalProvinceOfOrigin ⊑ Wine	∃wTraditionalProvinceOfOrigin−⊑ String	WineGrapeComp* ⊑ ∃wgcWineId
∃wTraditionalRegionOfOrigin ⊑ Wine	∃wTraditionalRegionOfOrigin− ⊑ String	GrapeVariety ⊑ ∃gvName
∃wdDescription ⊑ WineDescription	∃wdDescription− ⊑ String	WineType ⊑ ∃tName
∃wdColour ⊑ WineDescription	∃wdColour− ⊑ String	WineType ⊑ ∃tCategory
∃wdAlcoholContent ⊑ WineDescription	∃wdAlcoholContent− ⊑ String	Certification ⊑ ∃cExtendedCode
∃wdSugar ⊑ WineDescription	∃wdSugar− ⊑ String	Certification ⊑ ∃cName
∃wodOrganDescrip* ⊑ WineOrgDescrip*	∃wodOrganolepticDescription− ⊑ String	Certification ⊑ ∃cType
∃wodColourDetailed ⊑ WineOrgDescrip*	∃wodColourDetailed− ⊑ String	Actor ⊑ ∃acRole
∃wodFlavor ⊑ WineOrganolepticDescription	∃wodFlavor− ⊑ String	Actor ⊑ ∃acName
∃wodTaste ⊑ WineOrganolepticDescription	∃wodTaste− ⊑ String	Events ⊑ ∃eName
∃wodBody ⊑ WineOrganolepticDescription	∃wodBody− ⊑ String	Events ⊑ ∃eStartDate
∃wgcGrapeId ⊑ WineGrapeComposition	∃wgcGrapeId− ⊑ Integer	Address ⊑ ∃aStreet
∃wgcWineId ⊑ WineGrapeComposition	∃wgcWineId− ⊑ Integer	Address ⊑ ∃aZipCode
∃wgcPercOfGrape* ⊑ WineGrapeComp*	∃wgcPercentageOfGrape− ⊑ Decimal	Geolocalization ⊑ ∃gLatitude
∃gvName ⊑ GrapeVariety	∃gvName− ⊑ String	Geolocalization ⊑ ∃gLongitude
∃gvTypicalPeriod ⊑ GrapeVariety	∃gvTypicalPeriod− ⊑ String	Municipality ⊑ ∃mLauCode
∃gvSeason ⊑ GrapeVariety	∃gvSeason− ⊑ String	Municip* ⊑ ∃mLauNameNational
∃wtName ⊑ WineType	∃wtName− ⊑ String	Municipality ⊑ ∃mzipCode
∃wtCategory ⊑ WineType	∃wtCategory− ⊑ String	Municipality ⊑ ∃mCountry
∃cExtendedCode ⊑ Certification	∃cExtendedCode− ⊑ String	Municipality ⊑ ∃mNUTSLevel1
∃cName ⊑ Certification	∃cName− ⊑ String	Municipality ⊑ ∃mNUTSLevel2
∃cType ⊑ Certification	∃cType− ⊑ String	Municipality ⊑ ∃mNUTSLevel3
∃acRole ⊑ Actor	∃acRole− ⊑ String	Municipality ⊑ ∃mNUTS3Code
∃acName ⊑ Actor	∃acName− ⊑ String	
∃acPhoneNumber ⊑ Actor	∃acPhoneNumber− ⊑ String	
∃acURL ⊑ Actor	∃acURL− ⊑ String	
∃pProductionManager ⊑ Producer	∃pProductionManager− ⊑ String	
∃ctHeadOfTheGarrison ⊑ Consortium	∃ctHeadOfTheGarrison− ⊑ String	
∃ctContactForProducers ⊑ Consortium	∃ctContactForProducers− ⊑ String	
∃eName ⊑ Events	∃eName− ⊑ String	
∃eStartDate ⊑ Events	∃eStartDate− ⊑ Date	
∃eEndDate ⊑ Events	∃eEndDate− ⊑ Date	
∃eStartTime ⊑ Events	∃eStartTime− ⊑ Time	
∃eEndTime ⊑ Events	∃eEndTime− ⊑ Time	
∃aStreet ⊑ Address	∃aStreet− ⊑ String	
∃aNumber ⊑ Address	∃aNumber− ⊑ Integer	
∃aComplement ⊑ Address	∃aComplement− ⊑ String	
∃aZipCode ⊑ Address	∃aZipCode− ⊑ Integer	
∃gLatitude ⊑ Geolocalization	∃gLatitude− ⊑ Decimal	
∃gLongitude ⊑ Geolocalization	∃gLongitude− ⊑ Decimal	
∃mLauCode ⊑ Municipality	∃mLauCode− ⊑ String	
∃mLauNameNational ⊑ Municipality	∃mLauNameNational− ⊑ String	

$\exists \text{mzipCode} \sqsubseteq \text{Municipality}$ $\exists \text{mCountry} \sqsubseteq \text{Municipality}$ $\exists \text{mNUTSLevel1} \sqsubseteq \text{Municipality}$ $\exists \text{mNUTSLevel2} \sqsubseteq \text{Municipality}$ $\exists \text{mNUTSLevel3} \sqsubseteq \text{Municipality}$ $\exists \text{mNUTS3Code} \sqsubseteq \text{Municipality}$	$\exists \text{mzipCode}^{-} \sqsubseteq \text{Integer}$ $\exists \text{mCountry}^{-} \sqsubseteq \text{String}$ $\exists \text{mNUTSLevel1}^{-} \sqsubseteq \text{String}$ $\exists \text{mNUTSLevel2}^{-} \sqsubseteq \text{String}$ $\exists \text{mNUTSLevel3}^{-} \sqsubseteq \text{String}$ $\exists \text{mNUTS3Code}^{-} \sqsubseteq \text{String}$	

Disjointness assertions:

$\text{Wine} \sqsubseteq \neg\, \text{WineDescription}$
$\text{Wine} \sqsubseteq \neg\, \text{WineOrganolepticDescription}$
$\text{Wine} \sqsubseteq \neg\, \text{WineGrapeComposition}$
$\text{Wine} \sqsubseteq \neg\, \text{GrapeVariety}$
$\text{Wine} \sqsubseteq \neg\, \text{WineType}$
$\text{Wine} \sqsubseteq \neg\, \text{Certification}$
$\text{Wine} \sqsubseteq \neg\, \text{Actor}$
$\text{Wine} \sqsubseteq \neg\, \text{Events}$
$\text{Wine} \sqsubseteq \neg\, \text{Address}$
$\text{Wine} \sqsubseteq \neg\, \text{Geolocalization}$
$\text{Wine} \sqsubseteq \neg\, \text{Municipality}$

$\text{WineDescription} \sqsubseteq \neg\, \text{WineOrganolepticDescription}$
$\text{WineDescription} \sqsubseteq \neg\, \text{WineGrapeComposition}$
$\text{WineDescription} \sqsubseteq \neg\, \text{GrapeVariety}$
$\text{WineDescription} \sqsubseteq \neg\, \text{WineType}$
$\text{WineDescription} \sqsubseteq \neg\, \text{Certification}$
$\text{WineDescription} \sqsubseteq \neg\, \text{Actor}$
$\text{WineDescription} \sqsubseteq \neg\, \text{Events}$
$\text{WineDescription} \sqsubseteq \neg\, \text{Address}$
$\text{WineDescription} \sqsubseteq \neg\, \text{Geolocalization}$
$\text{WineDescription} \sqsubseteq \neg\, \text{Municipality}$

$\text{WineOrganolepticDescription} \sqsubseteq \neg\, \text{WineGrapeComposition}$
$\text{WineOrganolepticDescription} \sqsubseteq \neg\, \text{GrapeVariety}$
$\text{WineOrganolepticDescription} \sqsubseteq \neg\, \text{WineType}$
$\text{WineOrganolepticDescription} \sqsubseteq \neg\, \text{Certification}$
$\text{WineOrganolepticDescription} \sqsubseteq \neg\, \text{Actor}$
$\text{WineOrganolepticDescription} \sqsubseteq \neg\, \text{Events}$
$\text{WineOrganolepticDescription} \sqsubseteq \neg\, \text{Address}$
$\text{WineOrganolepticDescription} \sqsubseteq \neg\, \text{Geolocalization}$
$\text{WineOrganolepticDescription} \sqsubseteq \neg\, \text{Municipality}$

$\text{WineGrapeComposition} \sqsubseteq \neg\, \text{GrapeVariety}$
$\text{WineGrapeComposition} \sqsubseteq \neg\, \text{WineType}$
$\text{WineGrapeComposition} \sqsubseteq \neg\, \text{Certification}$
$\text{WineGrapeComposition} \sqsubseteq \neg\, \text{Actor}$
$\text{WineGrapeComposition} \sqsubseteq \neg\, \text{Events}$
$\text{WineGrapeComposition} \sqsubseteq \neg\, \text{Address}$
$\text{WineGrapeComposition} \sqsubseteq \neg\, \text{Geolocalization}$
$\text{WineGrapeComposition} \sqsubseteq \neg\, \text{Municipality}$

GrapeVariety ⊑ ¬ WineType
GrapeVariety ⊑ ¬ Certification
GrapeVariety ⊑ ¬ Actor
GrapeVariety ⊑ ¬ Events
GrapeVariety ⊑ ¬ Address
GrapeVariety ⊑ ¬ Geolocalization
GrapeVariety ⊑ ¬ Municipality

Type ⊑ ¬ Certification
Type ⊑ ¬ Actor
Type ⊑ ¬ Events
Type ⊑ ¬ Address
Type ⊑ ¬ Geolocalization
Type ⊑ ¬ Municipality

Certification ⊑ ¬ Actor
Certification ⊑ ¬ Events
Certification ⊑ ¬ Address
Certification ⊑ ¬ Geolocalization
Certification ⊑ ¬ Municipality

Actor ⊑ ¬ Events
Actor ⊑ ¬ Address
Actor ⊑ ¬ Geolocalization
Actor ⊑ ¬ Municipality

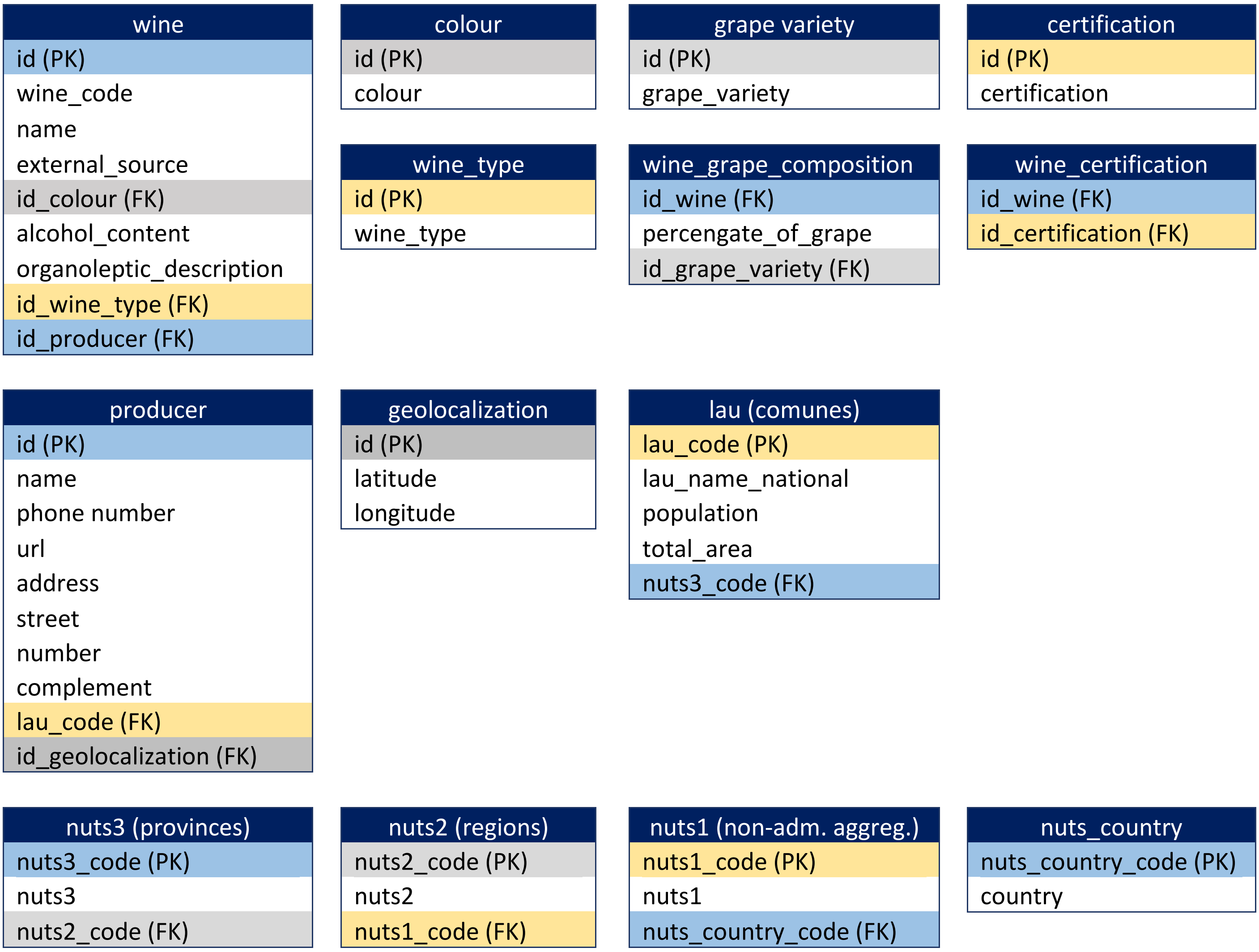
Events ⊑ ¬ Address
Events ⊑ ¬ Geolocalization
Events ⊑ ¬ Municipality

Address ⊑ ¬ Geolocalization
Address ⊑ ¬ Municipality

Geolocalization ⊑ ¬ Municipality

Table 14 - Abstract syntax of Description Logics for the formalization of the OWL2QL ontology

6.2 A graphical representation of the relational database in pgAdmin 4/PostgreSQL



wine
id (PK)
wine_code
name
external_source
id_colour (FK)
alcohol_content
organoleptic_description
id_wine_type (FK)
id_producer (FK)

colour
id (PK)
colour

grape variety
id (PK)
grape_variety

certification
id (PK)
certification

wine_type
id (PK)
wine_type

wine_grape_composition
id_wine (FK)
percengate_of_grape
id_grape_variety (FK)

wine_certification
id_wine (FK)
id_certification (FK)

producer
id (PK)
name
phone number
url
address
street
number
complement
lau_code (FK)
id_geolocalization (FK)

geolocalization
id (PK)
latitude
longitude

lau (comunes)
lau_code (PK)
lau_name_national
population
total_area
nuts3_code (FK)

nuts3 (provinces)
nuts3_code (PK)
nuts3
nuts2_code (FK)

nuts2 (regions)
nuts2_code (PK)
nuts2
nuts1_code (FK)

nuts1 (non-adm. aggreg.)
nuts1_code (PK)
nuts1
nuts_country_code (FK)

nuts_country
nuts_country_code (PK)
country

Figure 8 - A graphical representation of the relational database, with the attributes, keys, and foreign keys

6.3 Specification of the mappings between the mediated schema and the relational schema

Vkg mappings were design to connect the ontology to the relational schema using the Ontop plugin for Protégé (figure x). The file with all mappings used is in the folder of the project, as wine_project_vf.obda

Mapping	Target	source
MAPID-wine	:wine/wine/{id} a :Wine ; :wWineCode {wine_code}^^xsd:string ; :wName {name}^^xsd:string ; :wSource {external_source}^^xsd:string ; :hasDescription :wine/wine/{id} .	select id, wine_code, name, external_source, id_producer, id_wine_type from wine
MAPID-description	:wine/wine/{w_id} a :WineDescription ; :wdColour {colour}^^xsd:string ; :wdAlcoholContent {alcohol_content}^^xsd:string ; :hasOrganolepticDescription :wine/wine/{w_id} .	select w.id as w_id, w.id as wod_id, c.colour as colour, w.alcohol_content as alcohol_content from wine w left join colour c on w.id_colour = c.id
MAPID-organoleptic_description	:wine/wine/{id} a :WineOrganolepticDescription ; :wodOrganolepticDescription {organoleptic_description}^^xsd:string .	select id, organoleptic_description from wine
MAPID-wine_type	:wine/wine_type/{id} a :WineType ; :wtName {wine_type}^^xsd:string .	select id, wine_type from wine_type
MAPID-certification	:wine/certification/{id} a :Certification ; :cName {code}^^xsd:string ; :cExtendedCode {name}^^xsd:string .	select id, code, name from certification
MAPID-has_certification	:wine/wine/{w_id} :hasCertification :wine/certification/{c_id} .	select w.id as w_id, c.id as c_id from wine w, certification c, wine_certification wc where w.id = wc.id_wine and c.id = wc.id_certification
MAPID-grape_variety	:wine/grape_variety/{id} a :GrapeVariety ; :gvName {grape_variety}^^xsd:string .	select id, grape_variety from grape_variety
MAPID-has_grape_composition	:wine/wine/{w_id} :hasGrapeComposition :wine/wine_grape_composition/{wgc_id} .	select w.id as w_id, wgc.id as wgc_id from wine w, wine_grape_composition wgc where w.id = wgc.id_wine
MAPID-actor	:wine/producer/{id} a :Actor ; :acRole {role}^^xsd:string ; :acName {producer}^^xsd:string ; :acPhoneNumber {phone_number}^^xsd:string ; :acURL {url}^^xsd:string ; :hasMainAddress :wine/producer/{id} .	select id, 'Producer' as role, producer, phone_number, url from producer
MAPID-address	:wine/producer/{id} a :Address ; :aStreet {street}^^xsd:string ; :aNumber {number}^^xsd:integer ; :aComplement {complement}^^xsd:string .	select id, street, number, complement from producer
MAPID-geolocalization	:wine/geolocalization/{g_id} a :Geolocalization ; :gLatitude {latitude}^^xsd:decimal ; :gLongitude {longitude}^^xsd:decimal .	select id as g_id, latitude, longitude from geolocalization

MAPID-has_geolocalization	:wine/producer/{a_id} :hasGeolocalization :wine/geolocalization/{g_id} .	select a.id as a_id, g.id as g_id from producer a, geolocalization g where a.id_geolocalization = g.id

MAPID-is_produced_by	:wine/wine/{w_id} :isProducedBy :wine/producer/{p_id} .	SELECT w.id as w_id, p.id as p_id
		FROM wine w, producer p
		WHERE w.id_producer = p.id

MAPID-has_type	:wine/wine/{w_id} :hasType :wine/wine_type/{wt_id} .	select w.id as w_id, wt.id as wt_id from wine w, wine_type wt where w.id_wine_type = wt.id
MAPID-municipality	:wine/lau/{lau_code} a :Municipality ; :mLauNameNational {lau_name_national}^^xsd:string ; :mCountry {country}^^xsd:string ; :mNUTSLevel1 {nuts1}^^xsd:string ; :mNUTSLevel2 {nuts2}^^xsd:string ; :mNUTSLevel3 {nuts3}^^xsd:string ; :mNUTS3Code {nuts3_code}^^xsd:string .	select l.lau_code as lau_code, l.lau_name_national as lau_name_national, c.country as country, n1.nuts1 as nuts1, n2.nuts2 as nuts2, n3.nuts3 as nuts3, n3.nuts3_code as nuts3_code from lau l, nuts1 n1, nuts2 n2, nuts3 n3, nuts_country c where c.nuts_country_code = n1.nuts_country_code and n1.nuts1_code = n2.nuts1_code and n2.nuts2_code = n3.nuts2_code and n3.nuts3_code = l.nuts3_code
MAPID-has_municipality	:wine/producer/{a_id} :hasMunicipality :wine/lau/{lau_code} .	select a.id as a_id, l.lau_code as lau_code from producer a, lau l where a.lau_code = l.lau_code
MAPID-wine_grape_composition	:wine/wine_grape_composition/{id} a :WineGrapeComposition ; :wgcPercentageOfGrape {percentage_of_grape}^^xsd:decimal ; :wgcGrapeId {id_grape_variety}^^xsd:integer ; :wgcWineId {id_wine}^^xsd:integer .	select id, id_wine, id_grape_variety, percentage_of_grape from wine_grape_composition
MAPID-has_grape	:wine/wine_grape_composition/{wgc_id} :hasGrape :wine/grape_variety/{gv_id} .	select wgc.id as wgc_id, gv.id as gv_id from wine_grape_composition wgc, grape_variety gv where wgc.id_grape_variety = gv.id

Figure 9 – Viking Mapping Design using Ontop Plugin Protégé

6.4 SPARQL queries

Some SPARQL queries were used for testing and for the developed application. During the tests, the connections between classes, object properties and data properties were verified. Moreover, the number of records were checked, observing attributes that in the database relational have null values.

Name	Sparql Query

Query 1
wine_description
(1907 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?cod ?nam ?sour ?col ?alc ?org_desc
	WHERE {
	?w :wWineCode ?cod;
	:wName ?nam;
	:hasDescription ?d.
	?d :wdColour ?col.
	?d:hasOrganolepticDescription ?od.
	OPTIONAL {?w:wSource ?sour.}
	OPTIONAL {?d :wdAlcoholContent ?alc.}
	OPTIONAL {?od :wodOrganolepticDescription ?org_desc.}}

Query 2a
wine_grape_
composition_all
(2415 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?w ?gc ?perc ?gv ?grap
	WHERE {?w a :Wine.
	optional {?w :hasGrapeComposition ?gc.
	?gc :hasGrape ?gv.
	?gv :gvName ?grap. optional {?gc :wgcPercentageOfGrape ?perc.}}}

Query 2b
wine_grape_
composition_existent
(2331 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?w ?gc ?perc ?gv ?grap
	WHERE {?w :hasGrapeComposition ?gc.
	?gc :hasGrape ?gv.
	?gv :gvName ?grap.
	optional {?gc :wgcPercentageOfGrape ?perc.}}

Query 3a
wine_type_all
(1907 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?w ?wt ?t
	WHERE {?w a :Wine.
	optional {?w:hasType ?wt. ?wt :wtName ?t.}}

Query 3b
wine_type_existent
(973 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?w ?t
	WHERE {?w a :Wine;
	:hasType ?wt.
	?wt :wtName ?t.}

Query 4a
wine_certification_all
(1908 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?w ?c ?cert ?ext_cert
	WHERE {?w a :Wine.
	optional {?w :hasCertification ?c.
	?c :cName ?cert;
	:cExtendedCode ?ext_cert.}}

Query 4b
wine_certification_ existent
(1077 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?w ?c ?cert ?ext_cert
	WHERE {?w a :Wine;
	:hasCertification ?c.
	?c :cName ?cert;
	:cExtendedCode ?ext_cert.}

Query 5
Producer
(252 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?ac ?r ?n ?p ?url
	WHERE {
	?a a :Actor;
	:acRole ?r;
	:acName ?n.
	optional {?a :acPhoneNumber ?p.}
	optional {?a :acURL ?url}}

Query 6
Wine_producer
(1907 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?wn ?p ?r ?n ?ph ?url
	WHERE {?w :isProducedBy ?p;
	:wName ?wn.
	?p a :Actor;
	:acRole ?r;
	:acName ?n.
	optional {?p :acPhoneNumber ?ph}.
	optional {?p :acURL ?url}}

Query 7
address_geolocalization
(252 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	SELECT ?ad ?g ?lat ?long
	WHERE {
	?ad :hasGeolocalization ?g.
	?g :gLatitude ?lat ;
	:gLongitude ?long.}

Query 8
Address_municipality
(252 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	
	SELECT ?address ?mun ?country ?reg ?prov
	WHERE {?address :hasMunicipality ?m.
	?m :mLauNameNational ?mun;
	:mCountry ?country;
	:mNUTSLevel2 ?reg;
	:mNUTSLevel3 ?prov.}

Query 9
prod_geo_mun
(252 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	
	SELECT ?p ?r ?lat ?long ?lau ?reg ?prov
	WHERE {?ac :hasMainAddress ?d;
	:acName ?p.
	?d :hasGeolocalization ?g.
	?g :gLatitude ?lat;
	:gLongitude ?long.
	?d :hasMunicipality ?m.
	?m :mLauNameNational ?lau;
	:mNUTSLevel2 ?reg;
	:mNUTSLevel3 ?prov.}

Query 10
wine_desc_
prod_geo_mun
(1907 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	
	SELECT ?cod ?wn ?sour ?col ?alc ?org_desc ?p ?lat ?long ?lau ?reg ?prov
	WHERE {?w :isProducedBy ?ac;
	:wWineCode ?cod;
	:wName ?wn;
	:hasDescription ?d.
	?d :wdColour ?col.
	?d:hasOrganolepticDescription ?od.
	?ac a :Actor;
	:hasMainAddress ?ad;
	:acName ?p.
	?ad :hasGeolocalization ?g.
	?g :gLatitude ?lat;
	:gLongitude ?long.
	?ad :hasMunicipality ?m.
	?m :mLauNameNational ?lau;
	:mNUTSLevel2 ?reg;
	:mNUTSLevel3 ?prov.
	OPTIONAL {?w:wSource ?sour.}
	OPTIONAL {?d :wdAlcoholContent ?alc.}
	OPTIONAL {?od :wodOrganolepticDescription ?org_desc.}}

Query 11
wine_desc_grape_
prod_geo_mun
(2415 rows)
	PREFIX : <http://www.semanticweb.org/rachel/ontologies/2022/0/untitled-ontology-30#>
	
	SELECT ?cod ?wn ?sour ?col ?alc ?org_desc ?perc ?grap ?p ?lat ?long ?lau ?reg ?prov
	WHERE {?w :isProducedBy ?ac;
	:wWineCode ?cod;
	:wName ?wn;
	:hasDescription ?d.
	?d :wdColour ?col.
	?d:hasOrganolepticDescription ?od.
	?ac a :Actor;
	:hasMainAddress ?ad;
	:acName ?p.
	?ad :hasGeolocalization ?g.
	?g :gLatitude ?lat;
	:gLongitude ?long.
	?ad :hasMunicipality ?m.
	?m :mLauNameNational ?lau;
	:mNUTSLevel2 ?reg;
	:mNUTSLevel3 ?prov.
	OPTIONAL {?w:wSource ?sour.}
	OPTIONAL {?d :wdAlcoholContent ?alc.}
	OPTIONAL {?od :wodOrganolepticDescription ?org_desc.}
	optional {?w :hasGrapeComposition ?gc.
	?gc :hasGrape ?gv.
	?gv :gvName ?grap. optional {?gc :wgcPercentageOfGrape ?perc.}}}

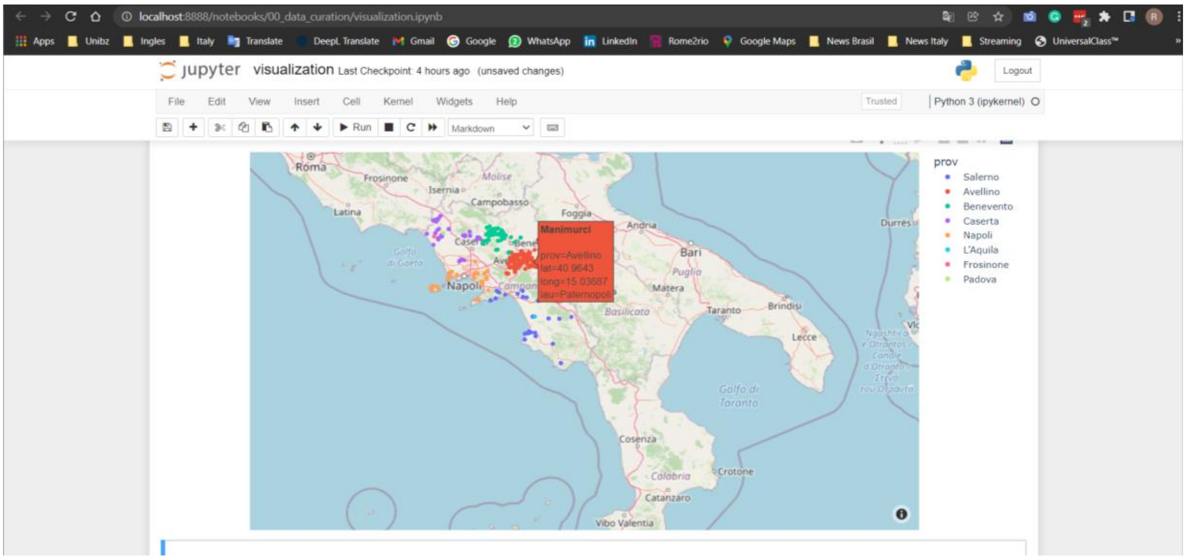
6.5 Description of the main functionalities of the application

A Jupyter Notebook application was developed for this domain, which makes use of Ontop as a SPARQL endpoint to query the database through the ontology, extracting relevant information and allowing some interactive visualizations.

For do these visualizations we used:
- Plotly, an open-source graphing libraries, to create interactive charts and maps using Python;
- Dash to build a web app with interactive charting capabilities where frontend and backend is handled with the tools that Dash provides.

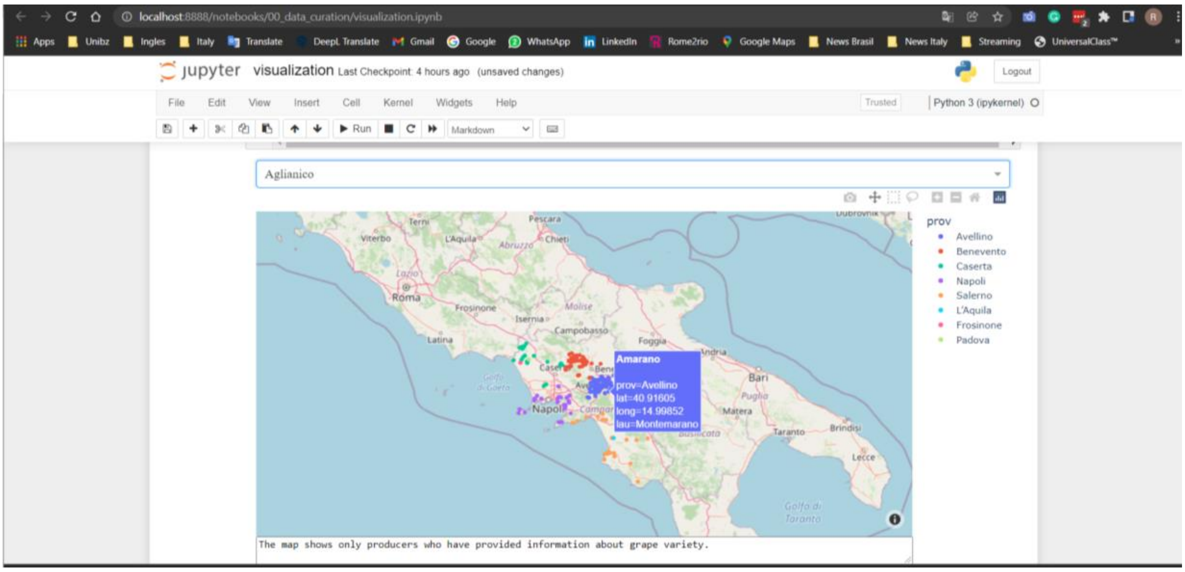
The visualizations with the results of the queries are presented below.

Producers - location map



Manimurci
prov=Avellino
lat=40.9643
long=15.03687
lau=Paternopoli

Producers by grape variety - location map

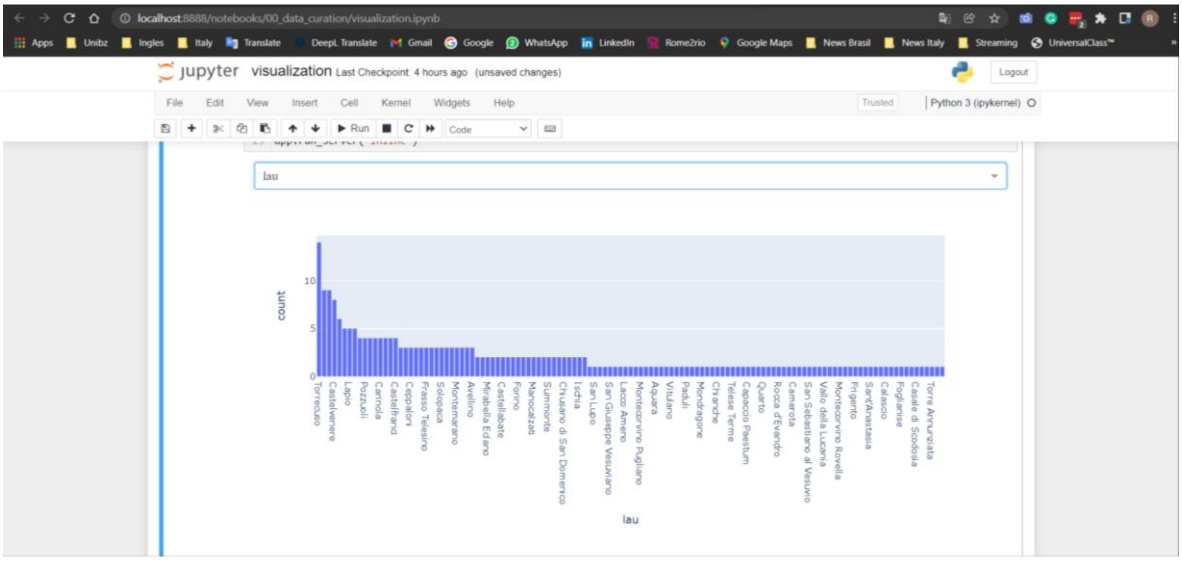


Aglianico

Amarano
prov=Avellino
lat=40.91605
long=14.99852
lau=Montemarano

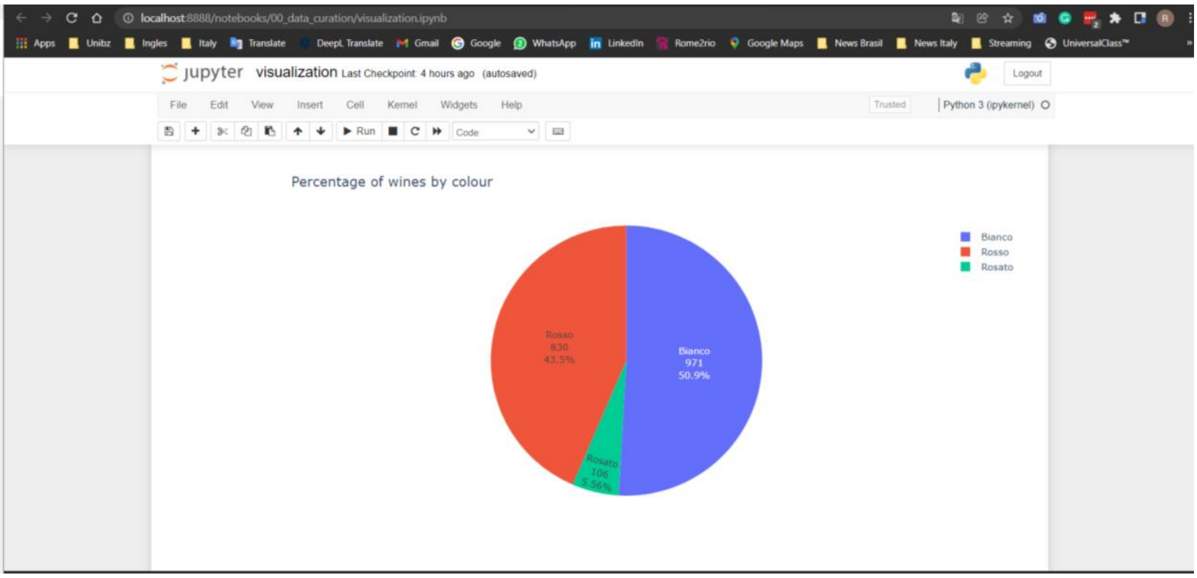
```
The map shows only producers who have provided information about grape variety.
```

Number of producers by Location

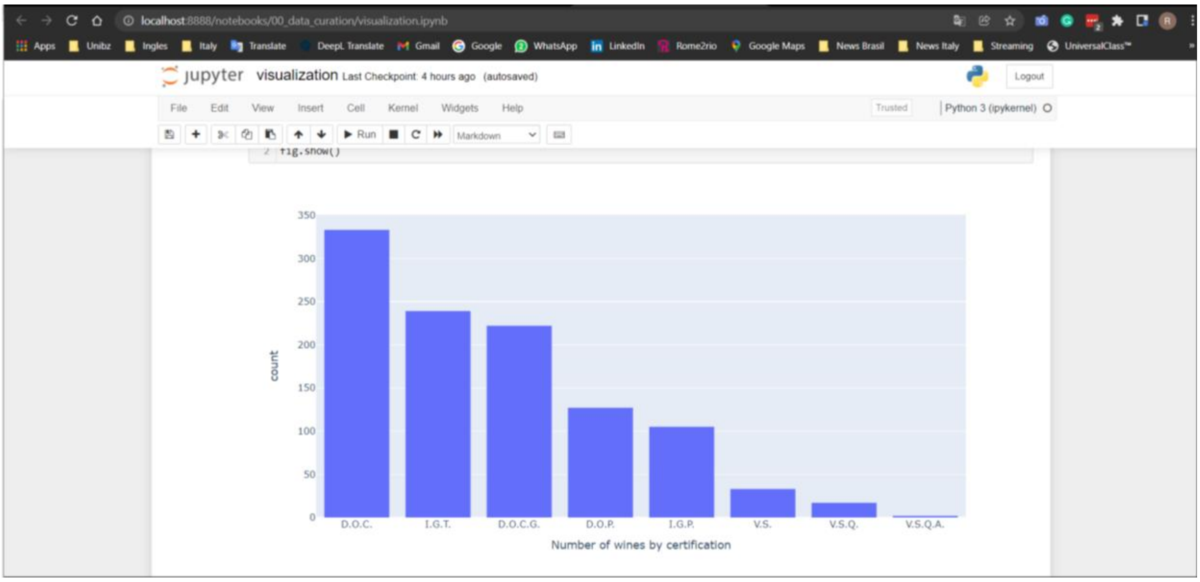


lau

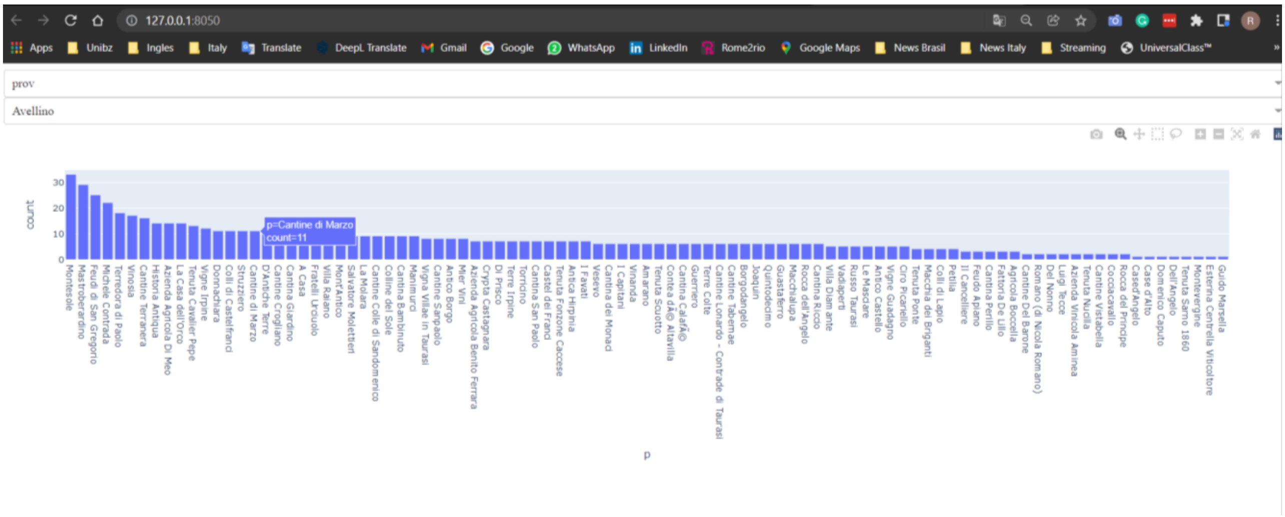
Percentage of wines by colour


localhost:8888/notebooks/00_data_curation/visualization.ipynb
Apps Unibz Ingles Italy Translate DeepL Translate Gmail Google WhatsApp LinkedIn Rome2rio Google Maps News Brasil News Italy Streaming UniversalClass™
jupyter visualization Last Checkpoint: 4 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Percentage of wines by colour
Bianco
Rosso
Rosato
Rosso
830
43.5%
Bianco
971
50.9%
Rosato
106
5.56%

Number of wines by certification


localhost:8888/notebooks/00_data_curation/visualization.ipynb
Apps Unibz Ingles Italy Translate DeepL Translate Gmail Google WhatsApp LinkedIn Rome2rio Google Maps News Brasil News Italy Streaming UniversalClass™
jupyter visualization Last Checkpoint: 4 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown
fig.show()
count
350
300
250
200
150
100
50
0
D.O.C.
I.G.T.
D.O.C.G.
D.O.P.
I.G.P.
V.S.
V.S.Q.
V.S.Q.A.
Number of wines by certification

Number of wines by producer and locality



7. The documentation

The documentation of the project includes:

1. data_curation_project_report: this document.
2. UML Diagram: file in a better resolution.
3. data_profiling_notebooks_html: the .html algorithms for find duplicates based on PLI, functional dependences and the matchings.
4. data_profiling_notebooks_html: the .ipynb algorithms for find duplicates based on PLI, functional dependences and the matchings.
5. ontop_vkg_specification: the .owl, .obda, and .properties files that make up the Ontop VKG specification of the project.
6. sparql_queries: some SPARQL queries that have been used either for testing or in the developed application (file with extension .q).
7. schema_and_dataset_dump: the export of the PostgreSQL database schema, generated (with extension sql) and the dump of the SQL database.
8. Visualizations: the .html and .ipynb with the visualizations.