A new gourmet wine shop in Rome

Report of the analysis

Agenda

- Business context and objectives
- Neighborhood search
    - Dataset
    - Methodology
    - Results
- Wine quality classifier
    - Dataset
    - Methodology
    - Results
- Conclusions

Business Context and Objectives

Vinho Verde Distribution is willing to open a new shop in Rome.
This analysis aims to answer the following questions:

- Neighborhood search: identify a central neighborhood in Rome to be preferred to open the new shop. It should be a place in which a gourmet wine shop can be sucessful, and that is not crowded with many other gourmet shops yet.

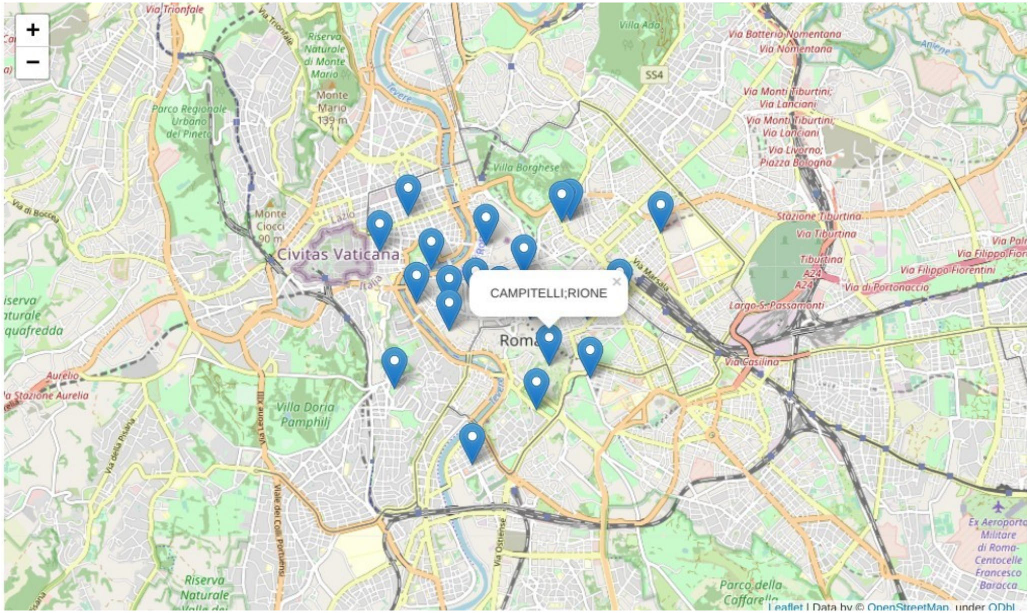
- Wine quality classifier: build a classifier for both white and red wines that is able to determine the quality of the wine by taking in input some physio-chemical features, and identify most informative features.

Neighborhood search: dataset

The dataset used contains all the neighborhoods in Rome*. The central one are indicated as Rione, and are the only one considered in the analysis.

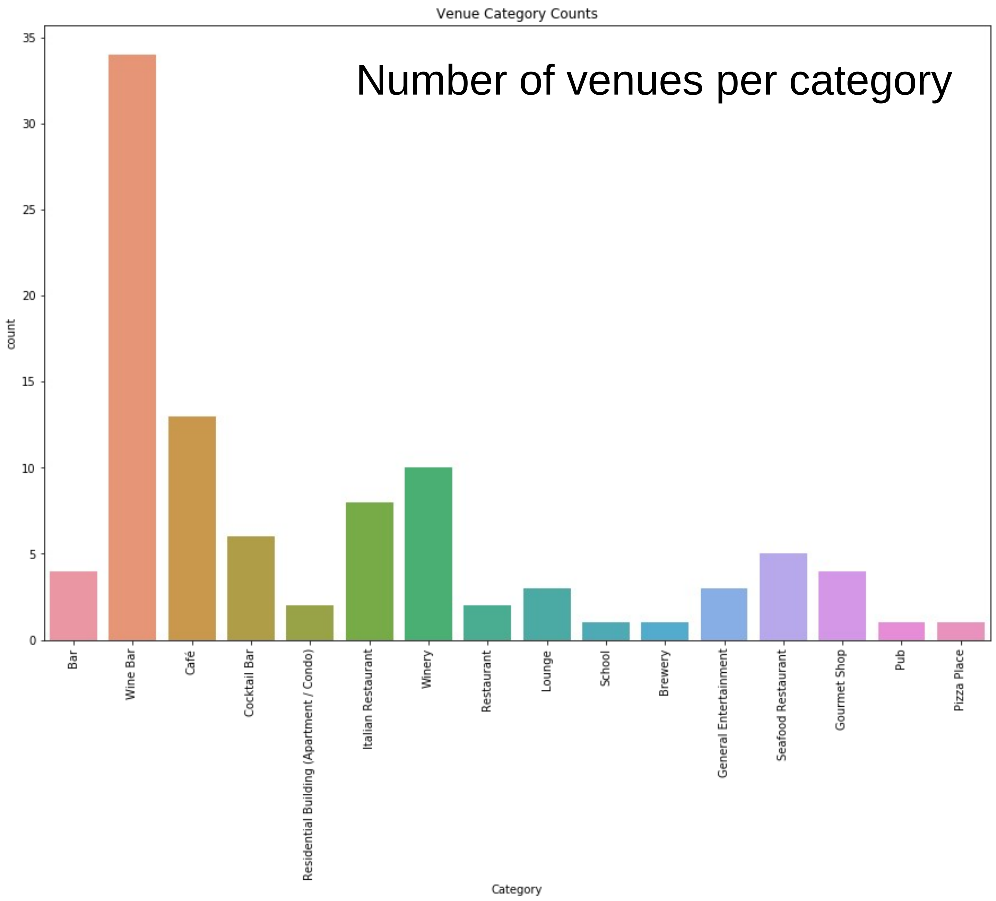
The dataset was enreached with coordinates for each neighborhood and, using FourSquare API, with the venues responding to a ‘wine’ query.

The overall count is of 98 venues distributed over 17 neighborhoods.

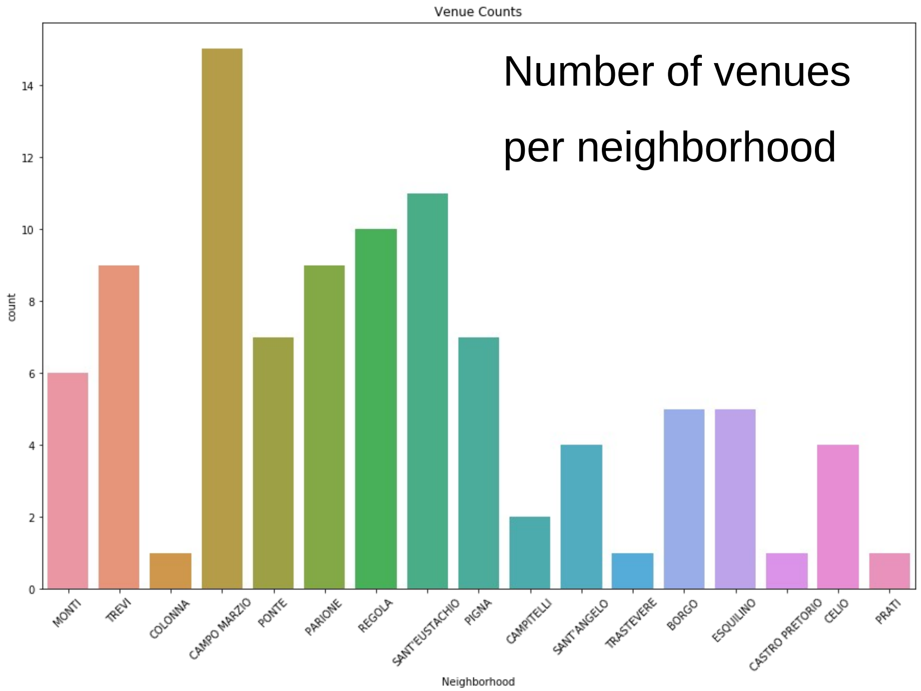


*https://www.sciamlab.com/opendatahub/dataset/c_h501_dts1580

Neighborhood search: methodology



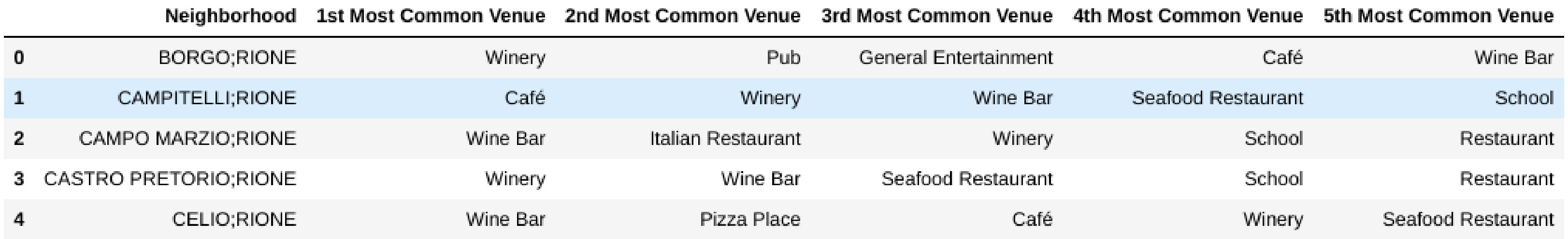
Number of venues per category



Number of venues per neighborhood

Neighborhood search: methodology

This is the head of the dataset used. It reports the 5 most common venues per neighborhood.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	BORGO;RIONE	Winery	Pub	General Entertainment	Café	Wine Bar
1	CAMPITELLI;RIONE	Café	Winery	Wine Bar	Seafood Restaurant	School
2	CAMPO MARZIO;RIONE	Wine Bar	Italian Restaurant	Winery	School	Restaurant
3	CASTRO PRETORIO;RIONE	Winery	Wine Bar	Seafood Restaurant	School	Restaurant
4	CELIO;RIONE	Wine Bar	Pizza Place	Café	Winery	Seafood Restaurant

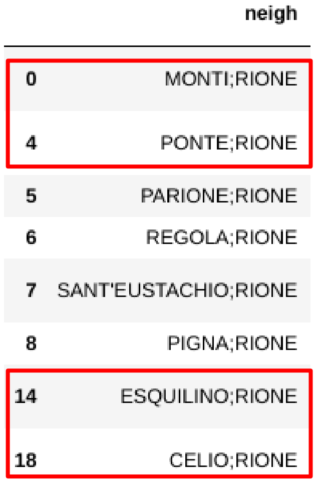
The algorithm used for clustering is a K-means algorithm looking for 5 different clusters.

Neighborhood search: results

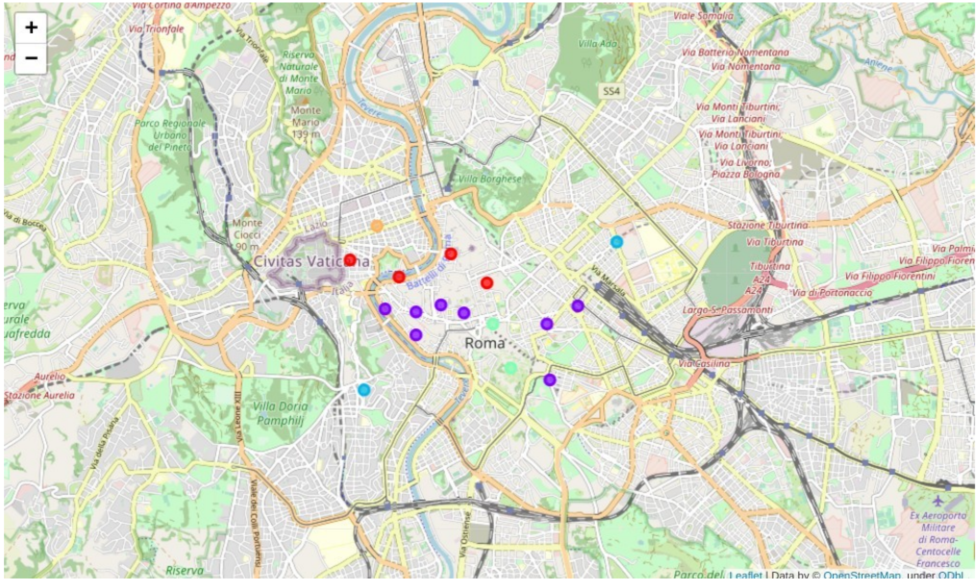
The 5 clusters identified are reported in the map below. The most interesting cluster is the one in purple.

Of the 8 neighborhoods of the clusters the four in red seems good places for the shop.

They are similar to places with gourmet shops, but does not have many gourmet shops right now.



	neigh
0	MONTI;RIONE
4	PONTE;RIONE
5	PARIONE;RIONE
6	REGOLA;RIONE
7	SANT'EUSTACHIO;RIONE
8	PIGNA;RIONE
14	ESQUILINO;RIONE
18	CELIO;RIONE

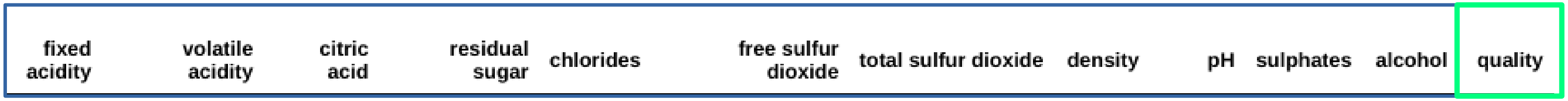


Wine quality classifier: datasets

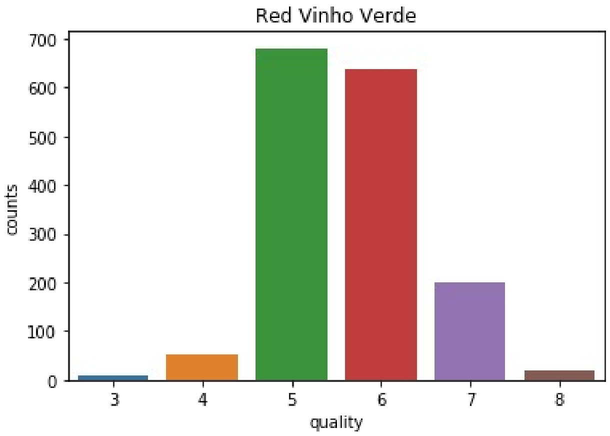
I used a dataset for red wines and one for white wines, from UCI repo*.

Phyisio-chemical variables

Target quality score

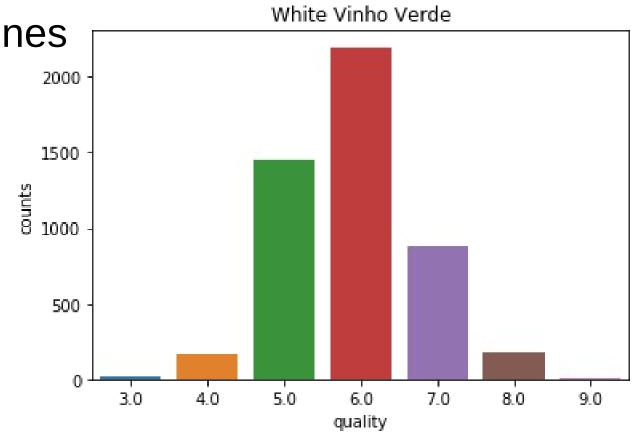


fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality



1599 wines records

4898 wines records

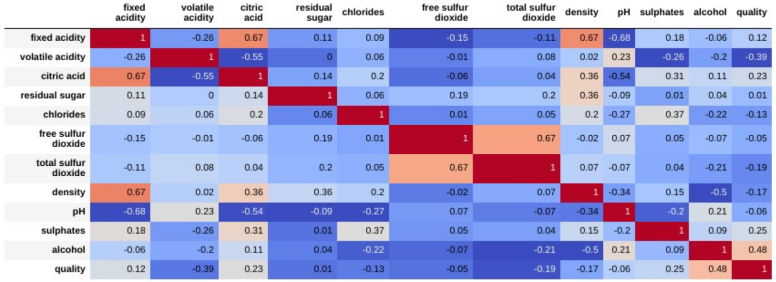


*https://archive.ics.uci.edu/ml/datasets/wine+quality

Wine quality classifier: methodology

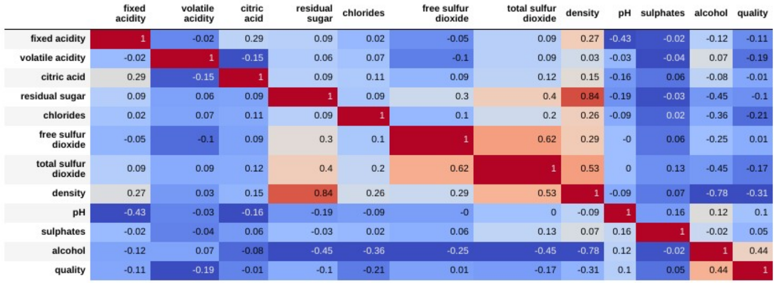
The datasets show weak correlations among variables.

Red Wine



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12
volatile acidity	-0.26	1	-0.55	0	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.2	-0.39
citric acid	0.67	-0.55	1	0.14	0.2	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0	0.14	1	0.06	0.19	0.2	0.36	-0.09	0.01	0.04	0.01
chlorides	0.09	0.06	0.2	0.06	1	0.01	0.05	0.2	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.01	-0.06	0.19	0.01	1	0.67	-0.02	0.07	0.05	-0.07	-0.05
total sulfur dioxide	-0.11	0.08	0.04	0.2	0.05	0.67	1	0.07	-0.07	0.04	-0.21	-0.19
density	0.67	0.02	0.36	0.36	0.2	-0.02	0.07	1	-0.34	0.15	-0.5	-0.17
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1	-0.2	0.21	-0.06
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.2	1	0.09	0.25
alcohol	-0.06	-0.2	0.11	0.04	-0.22	-0.07	-0.21	-0.5	0.21	0.09	1	0.48
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05	-0.19	-0.17	-0.06	0.25	0.48	1

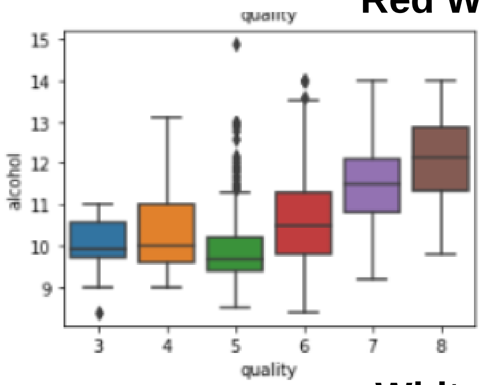
White Wine



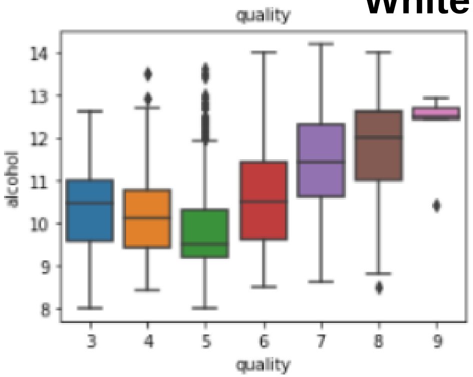
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12	-0.11
volatile acidity	-0.02	1	-0.15	0.06	0.07	-0.1	0.09	0.03	-0.03	-0.04	0.07	-0.19
citric acid	0.29	-0.15	1	0.09	0.11	0.09	0.12	0.15	-0.16	0.06	-0.08	-0.01
residual sugar	0.09	0.06	0.09	1	0.09	0.3	0.4	0.84	-0.19	-0.03	-0.45	-0.1
chlorides	0.02	0.07	0.11	0.09	1	0.1	0.2	0.26	-0.09	0.02	-0.36	-0.21
free sulfur dioxide	-0.05	-0.1	0.09	0.3	0.1	1	0.62	0.29	-0	0.06	-0.25	0.01
total sulfur dioxide	0.09	0.09	0.12	0.4	0.2	0.62	1	0.53	0	0.13	-0.45	-0.17
density	0.27	0.03	0.15	0.84	0.26	0.29	0.53	1	-0.09	0.07	-0.78	-0.31
pH	-0.43	-0.03	-0.16	-0.19	-0.09	-0	0	-0.09	1	0.16	0.12	0.1
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06	0.13	0.07	0.16	1	-0.02	0.05
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.02	1	0.44
quality	-0.11	-0.19	-0.01	-0.1	-0.21	0.01	-0.17	-0.31	0.1	0.05	0.44	1

One of the variables that correlates the most with quality seems to be alcohol.

Red Wine

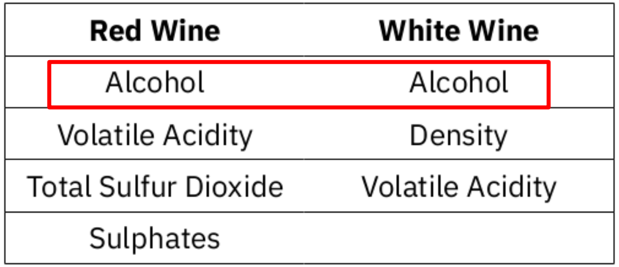


White Wine



Wine quality classifier: results

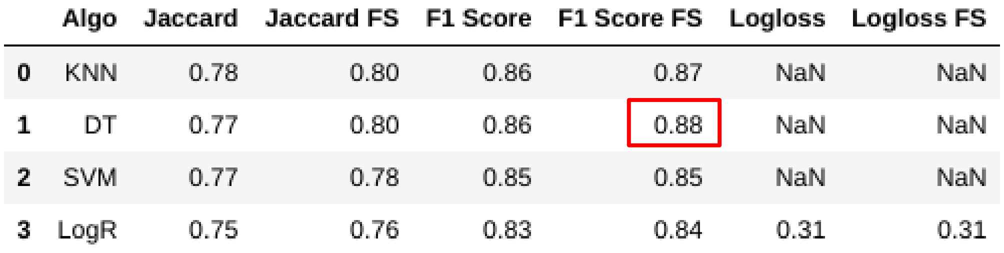
The feature analysis revealed that the most informative feature to determine wine quality are the following:



Red Wine	White Wine
Alcohol	Alcohol
Volatile Acidity	Density
Total Sulfur Dioxide	Volatile Acidity
Sulphates	

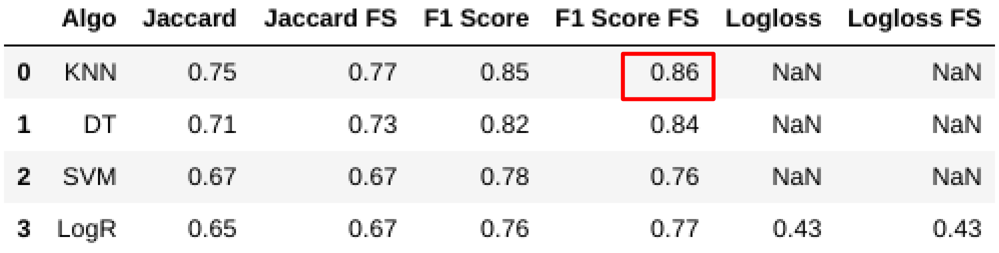
I tested four algorithms with full feature set and reduced feature set.

Red Wine: Decision Tree Classifier, Reduced feature set



	Algo	Jaccard	Jaccard FS	F1 Score	F1 Score FS	Logloss	Logloss FS
0	KNN	0.78	0.80	0.86	0.87	NaN	NaN
1	DT	0.77	0.80	0.86	0.88	NaN	NaN
2	SVM	0.77	0.78	0.85	0.85	NaN	NaN
3	LogR	0.75	0.76	0.83	0.84	0.31	0.31

White Wine: KNN Classifier, Reduced feature set

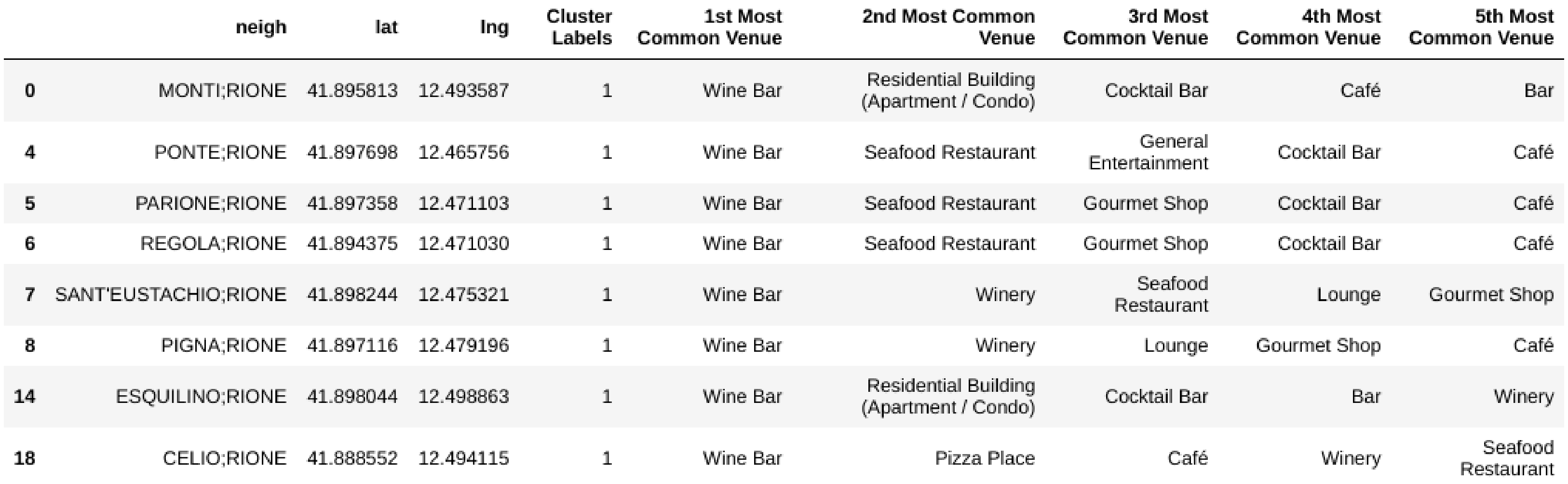


	Algo	Jaccard	Jaccard FS	F1 Score	F1 Score FS	Logloss	Logloss FS
0	KNN	0.75	0.77	0.85	0.86	NaN	NaN
1	DT	0.71	0.73	0.82	0.84	NaN	NaN
2	SVM	0.67	0.67	0.78	0.76	NaN	NaN
3	LogR	0.65	0.67	0.76	0.77	0.43	0.43

Conclusions

- Neighborhood search:
  - Select one among Monti, Ponte, Celio, Esquilino
  - **We could investigate further, using other features that can be important for you, which of these four would be a better fit for your shop.**
- Wine quality classifier:
  - Use Decision Tree with 4 features for red wines
  - Use KNN with 3 features for white wines
  - Accuracy is not perfect, always taste the wine before buying
  - **We could add more features to try to make the classifiers more accurate.**

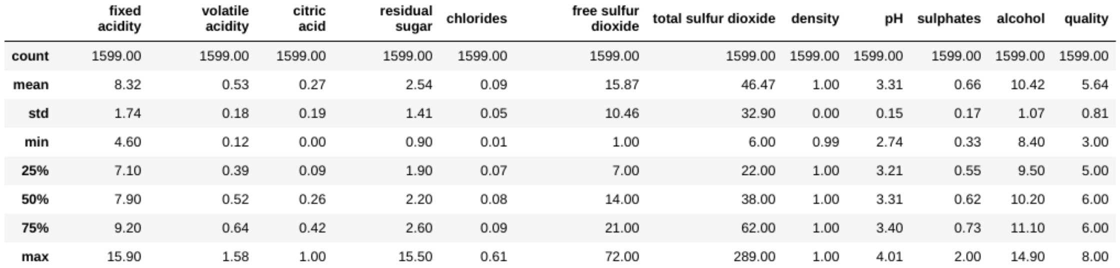
Appendix: Purple cluster



	neigh	lat	lng	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	MONTI;RIONE	41.895813	12.493587	1	Wine Bar	Residential Building (Apartment / Condo)	Cocktail Bar	Café	Bar
4	PONTE;RIONE	41.897698	12.465756	1	Wine Bar	Seafood Restaurant	General Entertainment	Cocktail Bar	Café
5	PARIONE;RIONE	41.897358	12.471103	1	Wine Bar	Seafood Restaurant	Gourmet Shop	Cocktail Bar	Café
6	REGOLA;RIONE	41.894375	12.471030	1	Wine Bar	Seafood Restaurant	Gourmet Shop	Cocktail Bar	Café
7	SANT'EUSTACHIO;RIONE	41.898244	12.475321	1	Wine Bar	Winery	Seafood Restaurant	Lounge	Gourmet Shop
8	PIGNA;RIONE	41.897116	12.479196	1	Wine Bar	Winery	Lounge	Gourmet Shop	Café
14	ESQUILINO;RIONE	41.898044	12.498863	1	Wine Bar	Residential Building (Apartment / Condo)	Cocktail Bar	Bar	Winery
18	CELIO;RIONE	41.888552	12.494115	1	Wine Bar	Pizza Place	Café	Winery	Seafood Restaurant

Appendix: statistics and feature selection

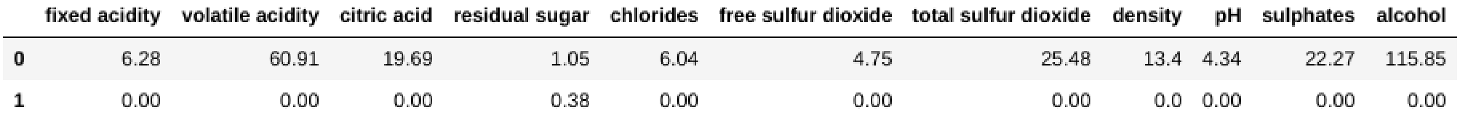
Red Wine



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00
mean	8.32	0.53	0.27	2.54	0.09	15.87	46.47	1.00	3.31	0.66	10.42	5.64
std	1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.00	0.15	0.17	1.07	0.81
min	4.60	0.12	0.00	0.90	0.01	1.00	6.00	0.99	2.74	0.33	8.40	3.00
25%	7.10	0.39	0.09	1.90	0.07	7.00	22.00	1.00	3.21	0.55	9.50	5.00
50%	7.90	0.52	0.26	2.20	0.08	14.00	38.00	1.00	3.31	0.62	10.20	6.00
75%	9.20	0.64	0.42	2.60	0.09	21.00	62.00	1.00	3.40	0.73	11.10	6.00
max	15.90	1.58	1.00	15.50	0.61	72.00	289.00	1.00	4.01	2.00	14.90	8.00

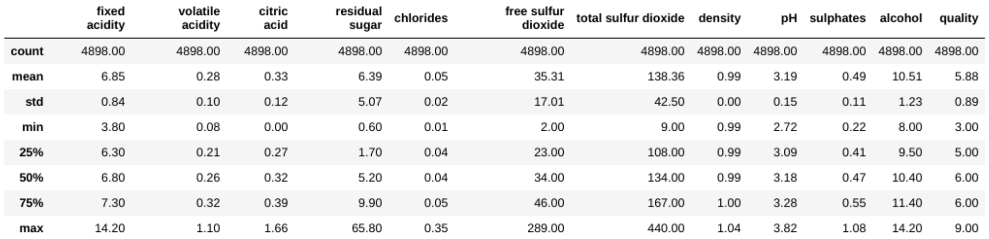
ANOVA F-Score

p-value



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	6.28	60.91	19.69	1.05	6.04	4.75	25.48	13.4	4.34	22.27	115.85
1	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.0	0.00	0.00	0.00

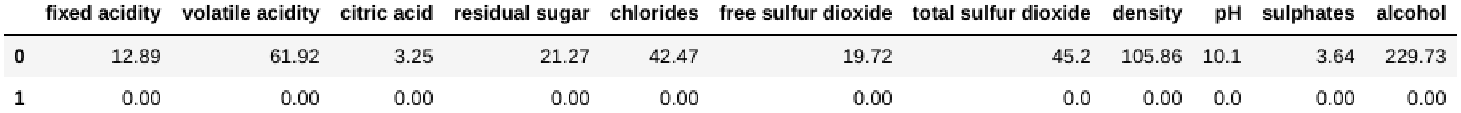
White Wine



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00
mean	6.85	0.28	0.33	6.39	0.05	35.31	138.36	0.99	3.19	0.49	10.51	5.88
std	0.84	0.10	0.12	5.07	0.02	17.01	42.50	0.00	0.15	0.11	1.23	0.89
min	3.80	0.08	0.00	0.60	0.01	2.00	9.00	0.99	2.72	0.22	8.00	3.00
25%	6.30	0.21	0.27	1.70	0.04	23.00	108.00	0.99	3.09	0.41	9.50	5.00
50%	6.80	0.26	0.32	5.20	0.04	34.00	134.00	0.99	3.18	0.47	10.40	6.00
75%	7.30	0.32	0.39	9.90	0.05	46.00	167.00	1.00	3.28	0.55	11.40	6.00
max	14.20	1.10	1.66	65.80	0.35	289.00	440.00	1.04	3.82	1.08	14.20	9.00

ANOVA F-Score

p-value



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	12.89	61.92	3.25	21.27	42.47	19.72	45.2	105.86	10.1	3.64	229.73
1	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.0	0.00	0.00