Capstone Project: A new gourmet wine shop in Rome

Business Introduction

As per our recent conversation I understood that Vinho Verde Distribution (VVD) wants to open a new gourmet wine shop in Rome.
You are specialized in selling Portuguese Vinho Verde wine, in the two variants, red and white. Your main needs are two:
- identify a neighborhood in Rome in which this kind of gourmet shop can be placed;
- identify a quick way to determine, starting from wine typical features, the quality of a wine, so to be able to quickly determine if a wine should be sold or not in your shop.

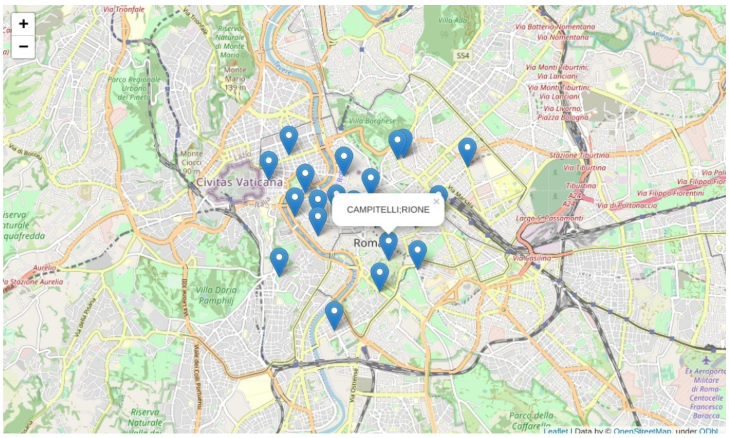
The neighborhood search should aim to find an area that is interested by tourism, and that can appreciate the value of a gourmet shop of this kind. However this neighborhood should not have yet too many gourmet shops, otherwise there would be no market for the new one. The ideal solution would be a neighborhood that is similar to gourmet-dense neighborhoods but that for now has not this kind of activity yet.

Regarding the wine quality classification, on the other hand, the aim is to create two classifiers, one for red and one for white wines. The classifiers should take as input the physical features of the wine and return a ‘good-poor’ label. Also, it would be useful to determine which of the physical features of the wine affect the most the final quality score.

Dataset used

Neighborhood search

Rome is a huge city, that is roughly divided in three big areas. The center, in which neighborhoods are called ‘Rione’, that is mostly touristic; the ring neighborhoods, which are called ‘Quartiere’ and in which the most Romans live, and finally the outer neighborhoods, which are called ‘Zona’ and which count neither tourism nor high-density population.

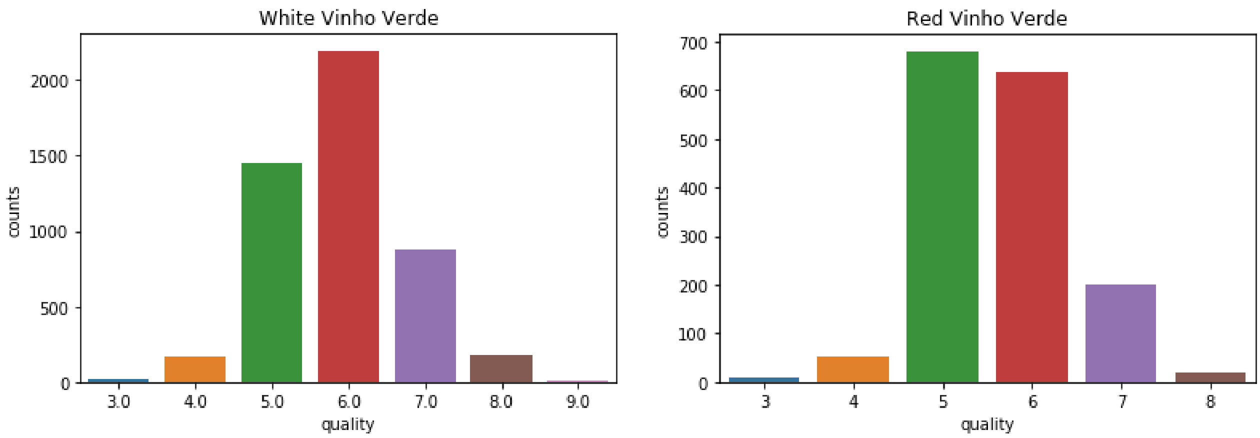


I used the Comprensori_Toponomastici.cvs [1] file, that lists all the neighborhoods in Rome. I focused on the one called Rione, so to select only the central neighborhoods of Rome. The lists counts 22 names, I searched for coordinates of each name using geopy.

One of the name could not be located, so I dropped it from the list.
Finally, using FourSquare API, I was able to retrieve venues for each Rione that matched the query ‘wine’. I retrieved 112 venues, but for two of the neighborhoods (Ludovisi and Sallustio) there were no category name associated with the venues, and for other two (Testaccio and Ripa) no venues were found. Given that these information are needed for the clustering, I dropped these neighborhoods too, ending up with 98 overall venues for 17 neighborhoods.

Wine quality classifier

To build the two wine classifiers I used the datasets from UCI [2]. These reports physical properties for red and white vinho verde wines, labeled with a quality score between 0 and 10.
The white wines dataset contains information for 4898 different wines, while the red wines dataset has entries for 1599 different wines.



The input variables, based on physio-chemical tests are the following:

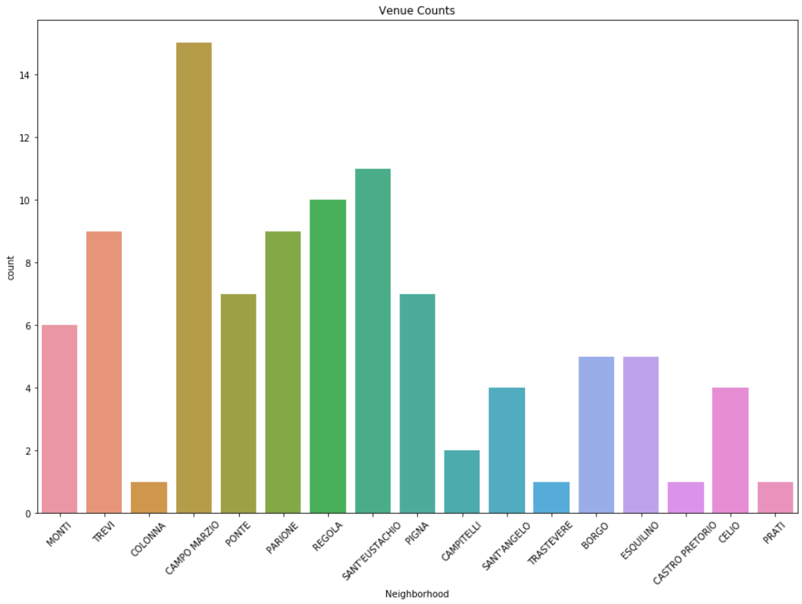
- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulfates
- alcohol

The output variable, based on sensory data, is a quality score between 0 and 10.

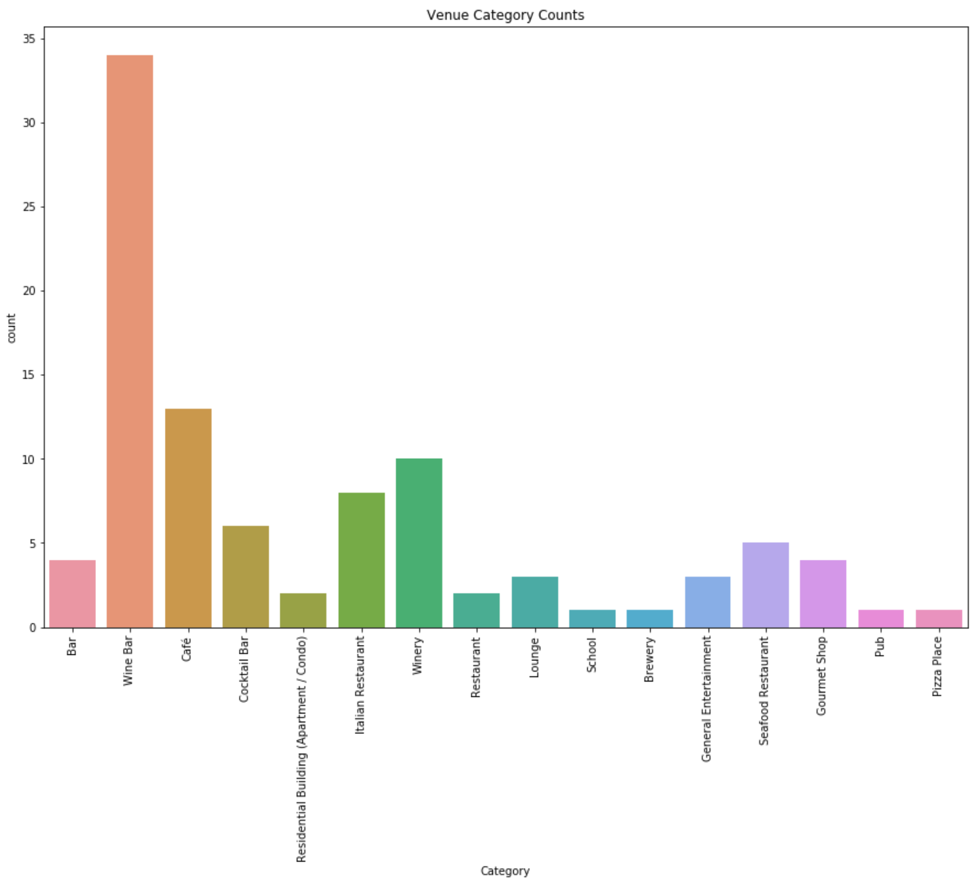
Methodology

Neighborhood search

I have created a dataset that contains wine-related venues for 17 central neighborhoods of Rome. Let's see how many venues I found for each neighborhood:



Among these venues I can count 16 different categories. I proceed with one hot encoding to use each of this category as a feature for my new dataset.
Next plot shows how many venues for each category are there in the dataset:



Next transformation computes the 5 most common category for the venues in each neighborhood. I selected the 5 most common, because the median value of the venue counts for the neighborhoods is 5. The head of the dataset looks like this:

	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
20	PRATI;RIONE	41.908329	12.464388	4	Italian Restaurant	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

I am now ready to run a K-means clustering algorithm. I choose to group the neighborhoods in 5 clusters.

Wine quality classifier

I perform some exploratory analysis on the white and red wine dataset. I check for null values and for wrong types but the datasets show high quality.
Here I report some descriptive analysis for the two.

<u>Red wine</u>

	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

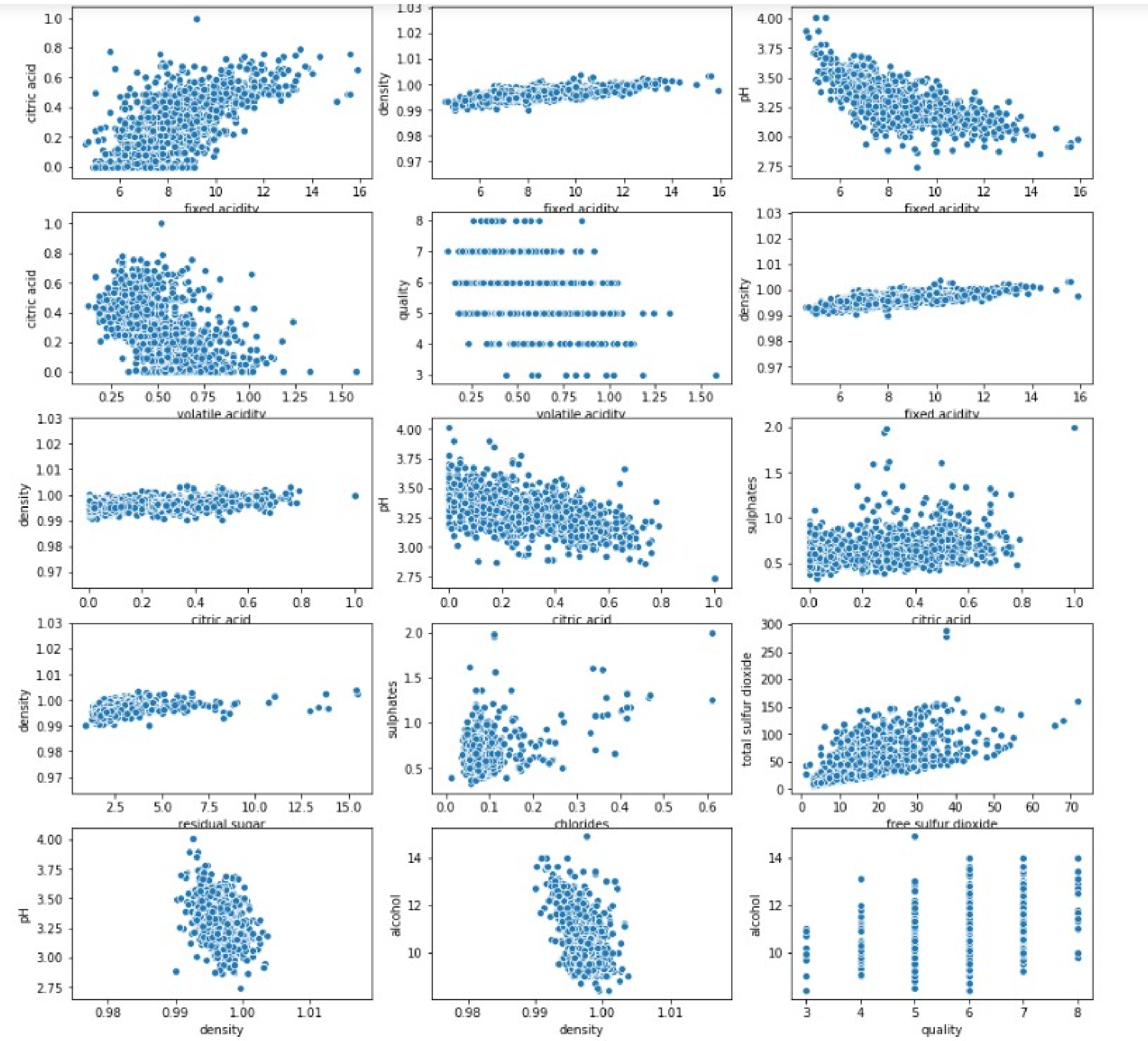
<u>White wine</u>

	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

I now try to determine possible correlations among variables through a correlation matrix and some scatter plots of the most correlated variables, to see if there are very strong dependencies that can affect with the results.

<u>Red wine</u>

	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

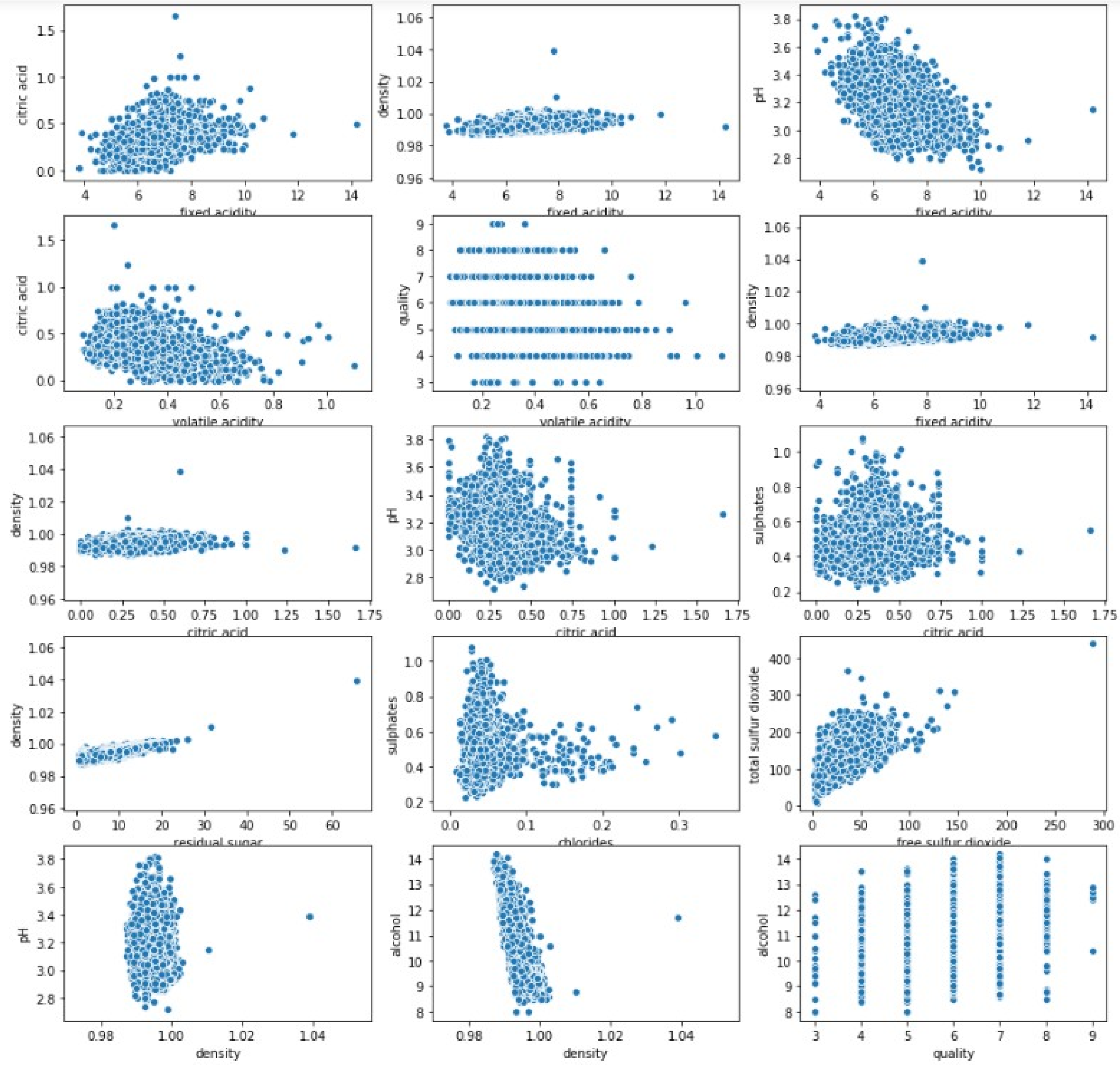


As one can easily see from the matrix and the plots there are some correlation but not that strong to justify further actions.

White wine

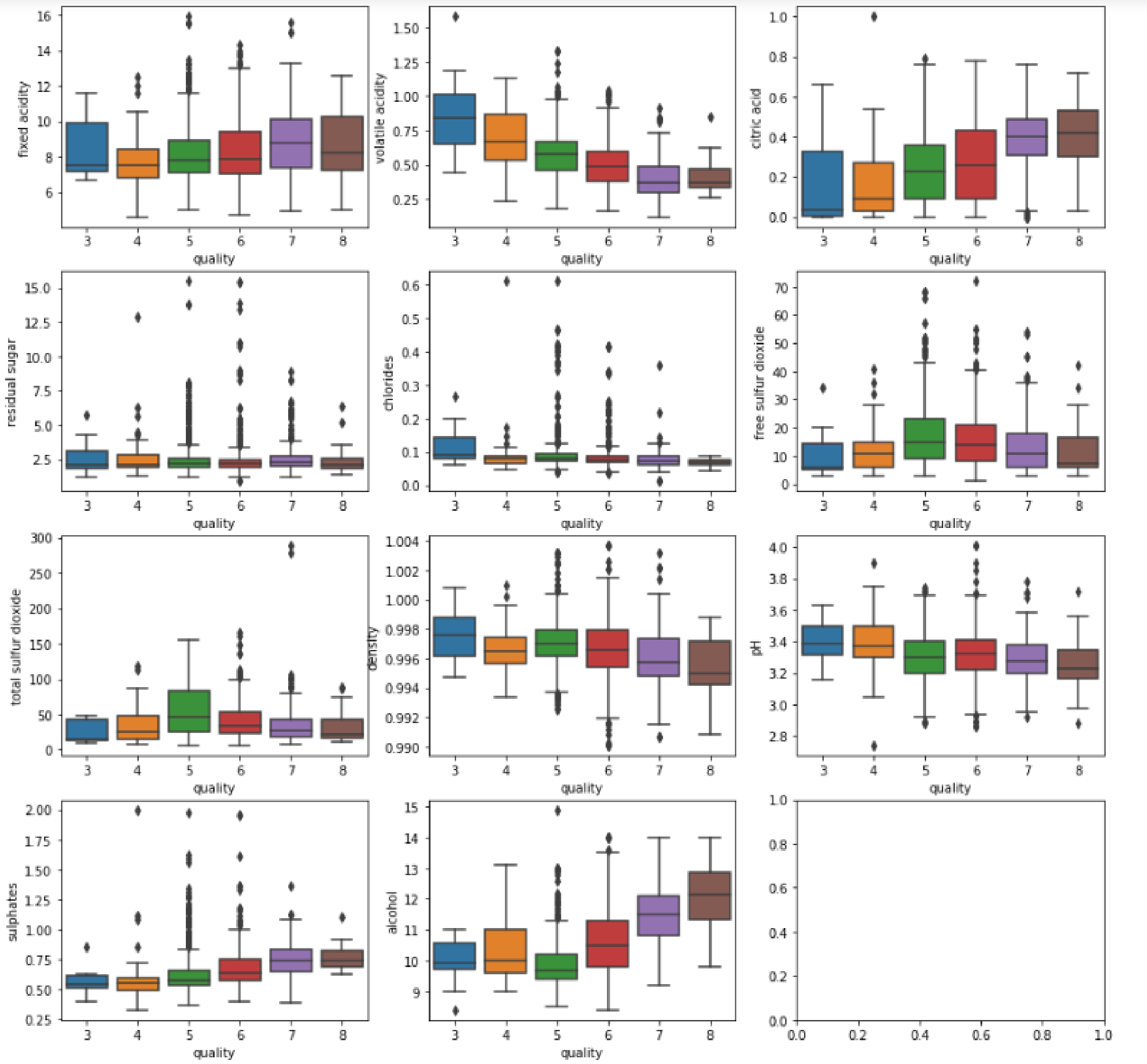
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12	-0.11
volatile acidity	-0.02	1	-0.15	0.06	0.07	-0.1	0.09	0.03	-0.03	-0.04	0.07	-0.19
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

Also in this case I do not see too strong correlations, as also the scatter plots demonstrate.



To investigate more the possible dependency of quality from the other variables I prepare some box plots that show possible important variability (i.e. alcohol in both datasets seems to have some importance).

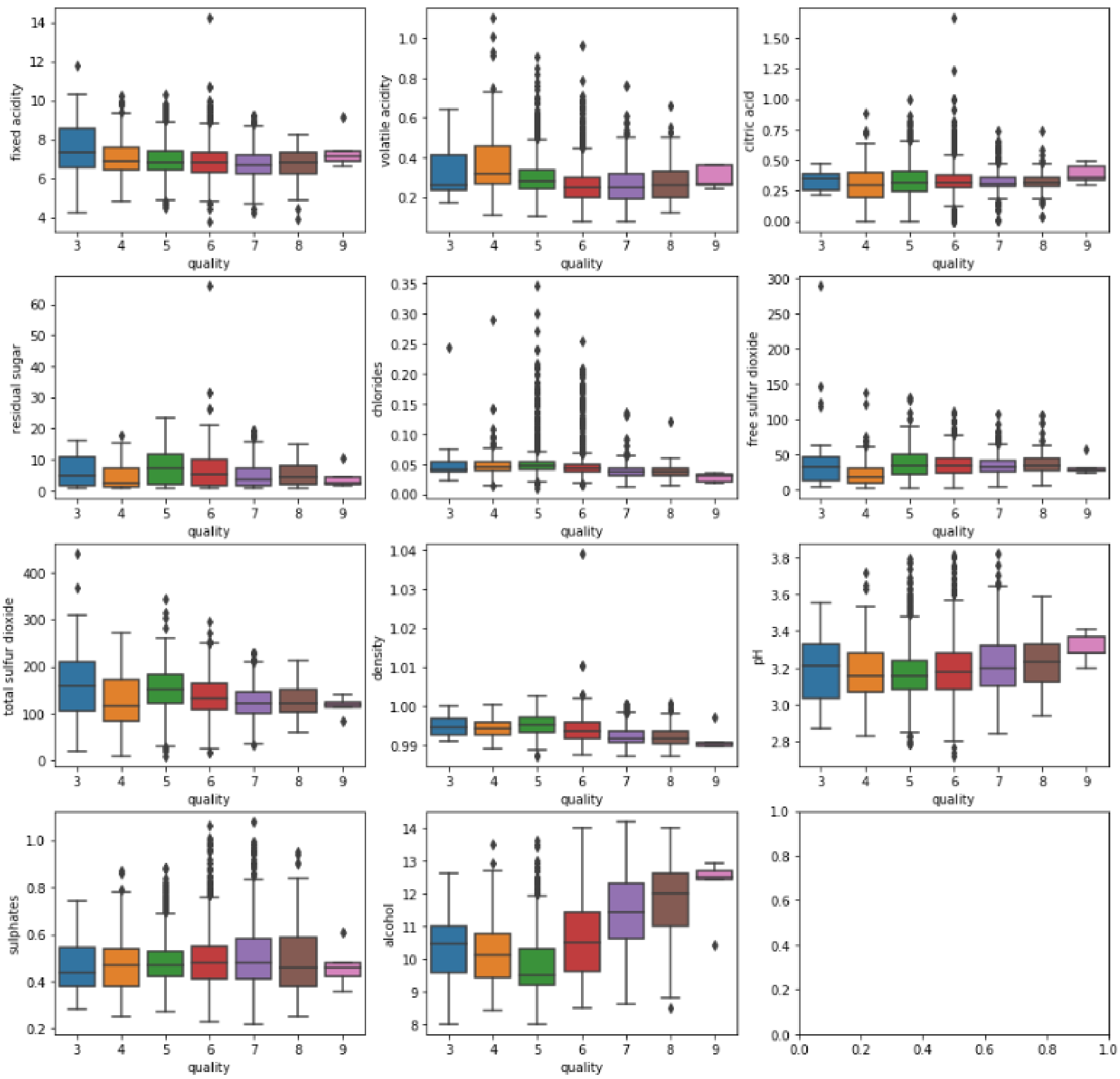
Red wine



The feature analysis, indeed, indicates that for red wine 4 variables are more important than others.
Here I report the ANOVA F-score (row 0) and the respective p-values (row 1). I decide to take in account for part of the analysis only the 4 variables with F-score higher than 20 (around 20% of the highest weight), each of which shows a very low p-value, indicating significance.

	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



Also for the white wine I find that 5 variables have a major effect with respect to others. I report the ANOVA F-score (row 0) and the respective p-values (row 1). I decide to take in account for part of the analysis only the 5 variables with F-score higher than 50 (around 20% of the highest weight).

	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

To proceed with the analysis I first associate the label 'good' to wines with a quality score higher than 6, and the label 'poor' to the others.

I split each dataset in a train-test. To determine the best K to use in the KNN classifier, however, I split again the training dataset in a training and an evaluation part.

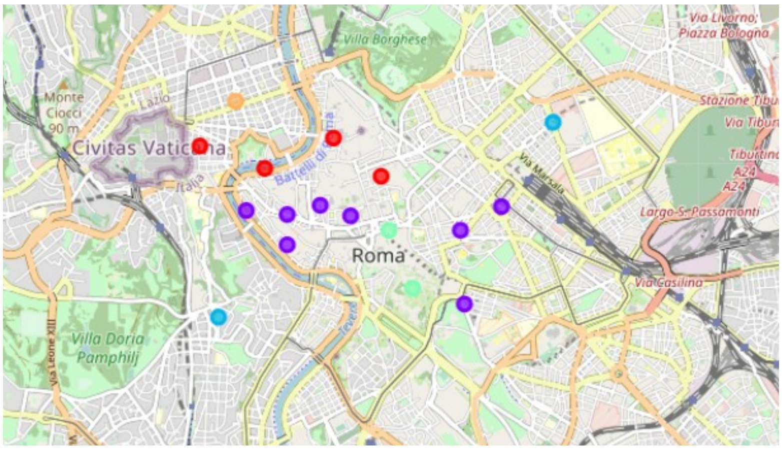
Once determined the best K, I try four classifiers using the original train-test split using KNN, Decision Tree, SVM, Logistic Regression as algorithms.

I make this analysis on the full dataset and on a dataset composed only by the 'selected features', to then compare the results.

Results

Neighborhood search

The results are depicted in the map:



It follows a description of the different clusters and their detailed composition.

<u>Red Cluster:</u>

	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
1	TREVI;RIONE	41.900978	12.483285	0	Italian Restaurant	Winery	Wine Bar	Restaurant	Lounge
3	CAMPO MARZIO;RIONE	41.904647	12.477055	0	Wine Bar	Italian Restaurant	Winery	School	Restaurant
10	SANT'ANGELO;RIONE	41.901758	12.468148	0	Seafood Restaurant	Italian Restaurant	General Entertainment	Café	Winery
13	BORGO;RIONE	41.903900	12.459657	0	Winery	Pub	General Entertainment	Café	Wine Bar

It contains mostly restaurants, cafè, bar and entertainment place. No gourmet retailer in this cluster.

<u>Blue Cluster:</u>

	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
12	TRASTEVERE;RIONE	41.887263	12.462117	2	Winery	Wine Bar	Seafood Restaurant	School	Restaurant
17	CASTRO PRETORIO;RIONE	41.906298	12.505559	2	Winery	Wine Bar	Seafood Restaurant	School	Restaurant

In this cluster we can find wineries, seafood restaurants and schools,but no trace of gourmet shops.

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

It is the most populated cluster, it contains some gourmet shops, along with wine bar, cocktail bar and seafood restaurants.

Green 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
2	COLONNA;RIONE	41.895821	12.484269	3	Café	Winery	Wine Bar	Seafood Restaurant	School
9	CAMPITELLI;RIONE	41.890085	12.487416	3	Café	Winery	Wine Bar	Seafood Restaurant	School

This cluster is similar to the blue one, but counts as most common venue the cafè.

Orange 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
20	PRATI;RIONE	41.908329	12.464388	4	Italian Restaurant	Winery	Wine Bar	Seafood Restaurant	School

This cluster contains only a neighborhood.

Wine quality classifier

The results obtained are the following.

Red wine

	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

Looking at the F1 score, the best classifier in this case seems to be the Decision Tree with reduced features, with a score of 0.88. An important point is that all the F1 scores but the SVM's one increase using the reduced features. The selection of the most informative features is able in this case to reduce the noise.

I report the confusion matrix for the most accurate classifier:

	true good	true poor
predicted good	26	21
predicted poor	17	256

<u>White wine</u>

	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

Looking at F1 score, the best classifier in this case seems to be the KNN, with an F1 score of 0.86. Also in this case all the algorithms are performing better with less features, but the SVM. Nonetheless this dataset is bigger than the red wine dataset, its overall F1 score is lower than the previous.

I report the confusion matrix for the best model, the KNN with selected features.

	true good	true poor
predicted good	141	63
predicted poor	74	702

Discussion

The results of the neighborhoods analysis are quite easy to interpret. I would suggest to select one of the neighborhoods inside the purple cluster that still not have many gourmet shops (Monti, Ponte, Celio, Esquilino). Based on the analysis, indeed, this neighborhoods are similar to ones in which gourmet shops are frequent, so a business of this kind should be successful there. However, because of these neighborhoods not having so many gourmet shops yet, you should find it convenient to open there, so to have less competitors.

From the analysis on the wine quality classifiers I can also bring some takeaways.
First of all the features that affect the most the wine quality are, ordered by importance:

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

I built several classifiers, and I suggest to use:
- for red wine the Decision Tree Classifier with only the features reported above, that shows a performance score F1=0.88;
- for white wine the KNN Classifier with only the features reported above, that shows an F1=0.86.

In any case, when using the classifiers, be aware that they are quite good in identifying poor wines, while they sometimes fail in recognizing good wines. My suggestion would be to use the classifiers to lower the number of wines of your interest, but to always check with wine-tasting the real quality of the products that the classifiers suggested you as good.

Conclusion

In conclusion I can add some suggestions to bring the analysis a step forward in the future.

For the neighborhood selection we can identify together more features that can be of your interest and try to select the best fit from the 4 neighborhoods indicated in this analysis.

About the wine classification, while I am quite sure that wine-tasting would be always needed to take a final decision, we can refine the analysis by working on other possible features that I can add to the dataset. Maybe using different features we can grow the performance of the classifiers and end up with less false positive/negative values.

I am open to deepen this analysis for you in one of the ways suggested or in any other way of your interest.

References to datasets

[1] https://www.sciamlab.com/opendatahub/dataset/c_h501_dts1580
[2] https://archive.ics.uci.edu/ml/datasets/wine+quality