Unsupervised Machine Learning – Project

IBM Machine Learning Professional Certificate

By: Bernardita Štitić

Date: March 10, 2022

TABLE OF CONTENTS

1. OBJECTIVE AND BENEFITS 2

2. BRIEF DESCRIPTION OF THE DATA SET AND ITS ATTRIBUTES 2

3. DATA EXPLORATION, DATA CLEANING AND FEATURE ENGINEERING 4

3.1. Data exploration 4

3.2. Data cleaning and feature engineering 7

4. CLUSTERING MODELS 10

4.1. K-Means 10

4.2. Agglomerative Hierarchical Clustering 12

4.3. DBSCAN 15

5. MODEL DISCUSSION AND RECOMMENDATION 16

6. KEY FINDINGS AND INSIGHTS 17

7. NEXT STEPS 18

REFERENCES 19

1. OBJECTIVE AND BENEFITS

In this project, the developed models will be focused on clustering. Moreover, the main objective of this analysis is to compare the performance of three different clustering algorithms to select the best performing model based on the available data. In this context, this refers to a model which can properly assign data samples to separate clusters. This topic will be addressed later Section 2 when explaining the number of chosen clusters (two) and how a variable of the data set (wine color) was used to understand the clustering capabilities of each model. In addition, with respect to the selected algorithms, the number of clusters can be optimized for in only two cases. This aspect of the project will be explained in Section 2 and 4.

It should be mentioned that the data set chosen for this project is the wine quality data set of the course, which will be described in the next section. In addition, the main benefits provided to the business or stakeholders of the wine quality data set are clustering models with different degrees of performance as mentioned earlier. From the obtained results, it is possible to select a model which can serve as a preliminary basis for further development. Additionally, deeper insight into the data can be derived from the models, especially from the third model, which does not require the number of clusters to be passed as a hyperparameter. Ultimately, the approach and models presented here can also serve as an initial basis for similar or related data sets.

2. BRIEF DESCRIPTION OF THE DATA SET AND ITS ATTRIBUTES

As already mentioned, the data set that was chosen for this project is the wine quality data set. This data set was explored and preprocessed to a large extent in the file “04c_LAB_Clustering_Methods.ipynb” (course laboratory) in this course. As explained in the laboratory, this data set contains several chemical properties of wine like acidity, alcohol, sugar and pH. In addition, it includes a quality metric on an ordinal scale which was already encoded using integers (3-9, where 9 is the highest quality) and a color variable (“red” or “white” wine).

Therefore, this data set includes both numeric and categorical variables. In total, there are 13 features, where 11 are numeric, continuous variables, 1 is a discrete variable (the quality metric) and 1 is a categorical, nominal variable. Specifically, the categorical feature is wine color, whose value can be either “white” or “red”. To summarize, Table 2.1 shows the information obtained

using the .info() Pandas method on a Pandas DataFrame which contains the data set. The table shows the names, non-null values and types of the data set attributes.

Row index	Column name	Non-null values	Column data type
0	fixed_acidity	6497	float64
1	volatile_acidity	6497	float64
2	citric_acid	6497	float64
3	residual_sugar	6497	float64
4	Chlorides	6497	float64
5	free_sulfur_dioxide	6497	float64
6	total_sulfur_dioxide	6497	float64
7	Density	6497	float64
8	pH	6497	float64
9	Sulphates	6497	float64
10	Alcohol	6497	float64
11	Quality	6497	int64
12	Color	6497	Object

Table 2.1. Relevant information of a Pandas DataFrame with the wine quality set.

As Table 2.1 shows, this data set consists of 6,497 observations with 13 variables and no missing values. In total, the Pandas DataFrame has 11 float64 columns (chemical properties), 1 int64 column (quality metric) and, finally, 1 object column (wine color). After the preprocessing steps which will be detailed in the next section, achieving satisfactory clustering performance based on the 11 continuous, numerical features was the main goal of this analysis. One of the reasons why the numerical features were chosen was to meet the requirements of Scikit-learn (continuous features to develop models).

Also, the wine color column was used to study how the algorithms clustered the available observations based on the numerical features (or chemical properties). Consequently, with respect to the 2 models which require the number of clusters to be passed a hyperparameter, the chosen number was 2 clusters. This was decided to analyze whether the chemical properties of wine could properly allow for the separation of the data into 2 different groups which, ideally, would be related to wine color. In consequence, results were compared against the color feature to assess clustering performance. Therefore, toward this purpose, the remaining integer variable (quality metric) was considered irrelevant in this analysis. Furthermore, the results from the models can serve to gain further insight into the data, specifically about other possible ways to cluster it based on interesting chemical combinations.

3. DATA EXPLORATION, DATA CLEANING AND FEATURE ENGINEERING

3.1. Data exploration

For the data exploration phase, the following actions were taken with respect to the wine quality data set with 6,497 observations and 13 variables:

1. Relevant statistics: using the Pandas `.describe()` method, relevant statistics of the numerical variables were obtained.
    - For numeric variables (`float64` and `int64`), the statistics included the number of non-null values, the sample mean, the sample standard deviation, the minimum value, relevant percentiles (25^{th}, 50^{th} and 75^{th}) and the maximum value. The results are shown in Figure 3.1.1. As the figure shows, features are not on the same scale; this is a critical aspect for distance based algorithms.
    - For the `object` type column (`color` variable), the statistics included the number of nun-null values, the unique values, the top value and its frequency. The results are shown in Figure 3.1.2.

	count	mean	std	min	25%	50%	75%	max
fixed_acidity	6497.0	7.215307	1.296434	3.80000	6.40000	7.00000	7.70000	15.90000
volatile_acidity	6497.0	0.339666	0.164636	0.08000	0.23000	0.29000	0.40000	1.58000
citric_acid	6497.0	0.318633	0.145318	0.00000	0.25000	0.31000	0.39000	1.66000
residual_sugar	6497.0	5.443235	4.757804	0.60000	1.80000	3.00000	8.10000	65.80000
chlorides	6497.0	0.056034	0.035034	0.00900	0.03800	0.04700	0.06500	0.61100
free_sulfur_dioxide	6497.0	30.525319	17.749400	1.00000	17.00000	29.00000	41.00000	289.00000
total_sulfur_dioxide	6497.0	115.744574	56.521855	6.00000	77.00000	118.00000	156.00000	440.00000
density	6497.0	0.994697	0.002999	0.98711	0.99234	0.99489	0.99699	1.03898
pH	6497.0	3.218501	0.160787	2.72000	3.11000	3.21000	3.32000	4.01000
sulphates	6497.0	0.531268	0.148806	0.22000	0.43000	0.51000	0.60000	2.00000
alcohol	6497.0	10.491801	1.192712	8.00000	9.50000	10.30000	11.30000	14.90000
quality	6497.0	5.818378	0.873255	3.00000	5.00000	6.00000	6.00000	9.00000

Figure 3.1.1. Relevant statistics of the numeric variables of the wine quality data set.

	count	unique	top	freq
color	6497	2	white	4898

Figure 3.1.2. Relevant statistics of the color variable of the wine quality data set.

2. Feature selection: as performed in the course laboratory on clustering methods, the features called quality and color were discarded. As explained earlier, the first was considered irrelevant for the analysis, while the second one was used to assess clustering performance as described in the previous section. In particular, in this project, a similar approach to the course laboratory on clustering was followed. Therefore, the columns of the previous two variables were not dropped. However, preprocessing and model development was performed only with the `float64` columns.
3. Cluster exploration: using `.value_counts()`, it was discovered that, out of the 6,497 data samples, 4,898 corresponded to white wine (75%) whereas 1,599 belonged to the red wine cluster (25%). To visualize this, a histogram was built using a similar approach to the one used in the clustering laboratory with `Matplotlib` and `Pandas`. This histogram is shown in Figure 3.1.3.

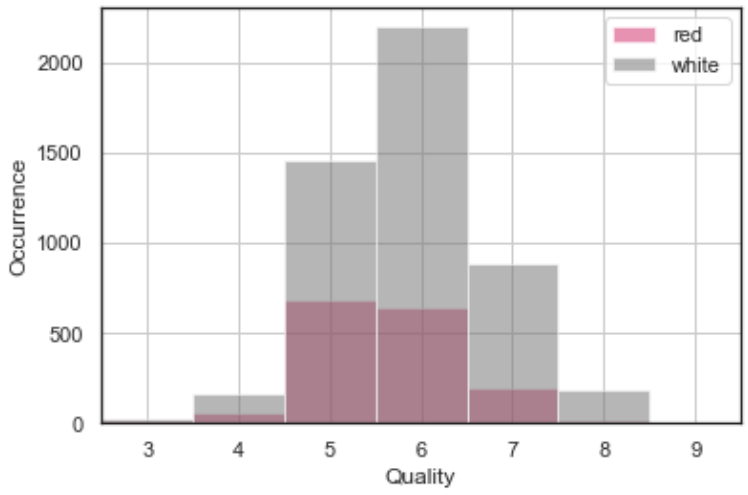

Figure 3.1.3. Histogram of the color variable distribution (wine quality data set).

4. Pair plot: a pair plot of the data set was obtained using the pairplot() function by Seaborn. It is shown in Figure 3.1.4. Also, the approach used in the clustering laboratory was followed to visualize which data points belonged to the white or red (potential) clusters. There as some apparently non-normal, skewed distributions that should be addressed during the preprocessing stage too. Furthermore, the different scales of the features can be appreciated on the plot. Finally, the density of the 2 clusters is also of interest (this will be further described in the next section).

Figure 3.1.4. Pair plot of the wine quality data set.

5. Correlation: due to the relevance of the curse of dimensionality in this case (unsupervised learning), correlation between the float64 variables was explored using the Pandas .corr() method. Strictly speaking, some assumptions should be verified, in addition to making additional scatter plots. However, like in the course laboratory on clustering, the .corr() method was enough for quickly exploring correlation values. These were not high enough to suggest that any float64 features should be excluded from the analysis. The correlation matrix is available in the Jupyter notebook.

3.2. Data cleaning and feature engineering

Data cleaning and feature engineering actions were strictly related to the float64 variables. To take adequate steps, a similar approach to the one presented in the course laboratory on clustering methods was followed. In particular, the following was performed:

1. Skewness: like in the laboratory, the skewness values of the float64 features were checked using the .skew() Pandas method to see which variables were more heavily skewed. Specifically, the same threshold of the laboratory was used so skewness values above 0.75 were considered relevant. The features chlorides, sulphates, fixed_acidity, volatile_acidity, residual_sugar and free_sulfur_dioxide met the previous condition so these were selected for further preprocessing. The skewness results are available in the Jupyter notebook of the project.
2. Logarithmic transformation: the previously selected variables were transformed using a logarithmic transformation, in particular .log1p() by NumPy.
3. Scaling: due to the relevance of number scales for clustering algorithms, the next step was to scale the float64 columns of the Pandas DataFrame. As a starting point, StandardScaler() by Scikit-learn was employed. Satisfactory results (regarding the scale of the features) were obtained and can be visualized in Figure 3.2.1.

	count	mean	std	min	25%	50%	75%	max
fixed_acidity	6497.0	5.276917e-15	1.000077	-3.629006	-0.644436	-0.106896	0.471461	5.049633
volatile_acidity	6497.0	-7.298900e-16	1.000077	-1.817681	-0.684343	-0.269292	0.443810	5.771086
citric_acid	6497.0	-1.753083e-16	1.000077	-2.192833	-0.472334	-0.059414	0.491146	9.231281
residual_sugar	6497.0	-1.453628e-15	1.000077	-1.687053	-0.866619	-0.343710	0.861368	3.783867
chlorides	6497.0	-7.225250e-16	1.000077	-1.463780	-0.543350	-0.262922	0.290773	13.734706
free_sulfur_dioxide	6497.0	3.034910e-15	1.000077	-3.927293	-0.574982	0.204386	0.717742	3.665722
total_sulfur_dioxide	6497.0	-9.658103e-16	1.000077	-1.941780	-0.685532	0.039907	0.712265	5.737257
density	6497.0	-4.487338e-15	1.000077	-2.530192	-0.785953	0.064489	0.764853	14.768791
pH	6497.0	3.086803e-15	1.000077	-3.100615	-0.674862	-0.052874	0.631312	4.923029
sulphates	6497.0	-4.748549e-15	1.000077	-2.432834	-0.699339	-0.105201	0.526688	7.387697
alcohol	6497.0	1.542248e-15	1.000077	-2.089350	-0.831615	-0.160823	0.677667	3.696231
quality	6497.0	5.818378e+00	0.873255	3.000000	5.000000	6.000000	6.000000	9.000000

Figure 3.2.1. Relevant statistics of the numeric variables of the preprocessed wine quality data set.

4. New pairplot: to visualize whether the scaling and normalization efforts were satisfactory, a new pair plot was made (Figure 3.2.2). As it can be seen, the previously taken steps did in fact improve the wine quality data set for clustering algorithm effects. It was relevant in this project to attempt to normalize and scale variables to avoid having heavily skewed data or features on different scales since these situations can negatively affect distance based algorithms. Also, the pair plot suggests that the separation between white and red wine is quite clean.

Figure 3.2.2. Pair plot of the preprocessed wine quality data set.

5. Skewness: the final skewness values of the `float64` variables were checked once more for verification purposes. The logarithmic transformation did in fact reduce the skewness of the features. This can be evidenced in the Jupyter notebook of the project, where these results can be found.

4. CLUSTERING MODELS

In this section, the developed models will be addressed. In particular, 3 clustering models were developed using the algorithms: K-Means, Agglomerative Hierarchical Clustering and DBSCAN. For all cases, the `.fit` method by `Scikit-learn` was called using all 6,497 preprocessed data samples and only the `float64` features. It is also relevant to highlight that, for the first two algorithms, the number of clusters was chosen to be two due to the comments about the pair plots concerning wine color in Section 3 and for the reasons explained in Section 2.

4.1. K-Means

As a first, exploratory approach, K-Means was chosen since there were not too many target clusters (only two). Also, `random_state=0` was used. Results are shown in Figure 4.1.1, which highlights how many data points that were assigned to each cluster by K-Means ("0" or "1") were actually red or white wine. This was also performed during the course laboratory.

		total
kmeans	**color**	
0	**red**	23
	white	4811
1	**red**	1576
	white	87

Figure 4.1.1. Results of the K-Means model with two clusters (preprocessed wine quality data set).

As shown in Figure 4.1.1, cluster "0" is dominated by white wine samples whereas cluster "1" mainly consists of red wine samples. It is also relevant to note that the actual number of white wine data points is 4,898 samples while for red wine the total is 1,599 samples (Section 3). In both cases, the cluster size of the clusters produced by K-Means was close to the real size of each wine group. It should be highlighted that, as explained in the course, generally K-Means is used to find a few clusters of roughly the same size. However, even though there were about 3 times more white wine samples, the model performed fairly well.

Furthermore, different K-Means models were trained using the same 6,497 preprocessed samples, `random_state=0` and a variable cluster size, which ranged from 1 to 20 clusters like in the laboratory. This was performed to build the elbow curve of inertia against cluster size to gain some insight into other possible ways of clustering the data and, possibly, to understand better how chemical properties could be combined to cluster wine. The resulting elbow curve is illustrated in Figure 4.1.2.

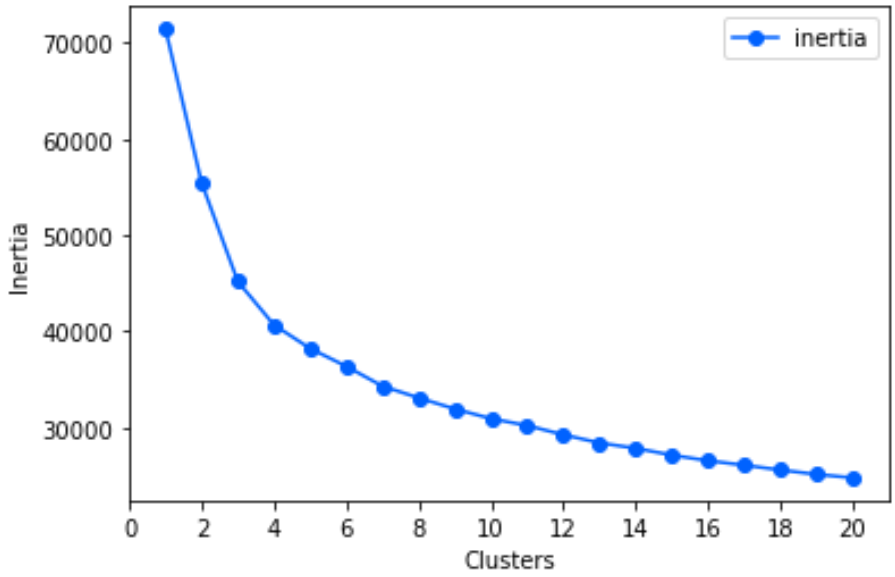

Figure 4.1.2. Elbow curve: inertia against cluster size (preprocessed wine quality data set).

As shown in Figure 4.1.2, unfortunately there is no clear elbow. The same situation happened during the laboratory. However, based on the curve, perhaps 3 or 4 clusters could be good choices for other K-Means models. Therefore, another K-Means model was developed using the same 6,497 data samples, `n_clusters=4` and `random_state=0`. Results are shown in Figure 4.1.3.

		total
kmeans_4	color	
0	red	32
	white	2659
1	red	14
	white	2101
2	red	648
	white	53
3	red	905
	white	85

Figure 4.1.3. Results of the K-Means model with four clusters (preprocessed wine quality data set).

To properly interpret the results in Figure 4.1.3, it would be necessary to study the float64 columns further, likely ranges of these numerical values and work with subject matter experts. Nonetheless, it is possible to see that the way colors were distributed this time is a bit more balanced in comparison to the previous K-Means model.

4.2. Agglomerative Hierarchical Clustering

Next, due to the proportion of wine and red wine samples (about 3 times more white wine), an Agglomerative Hierarchical Clustering model was developed. Similarly as before, two clusters were used (n_clusters=2) with linkage='ward' and computer_full_tree=True like in the course laboratory on clustering methods. The model was developed using all 6,497 preprocessed data samples and only the float64 variables as well. The results are shown in Figure 4.2.1.

color	agglom	total
red	0	31
	1	1568
white	0	4755
	1	143

Figure 4.2.1. Results of the Agglomerative Hierarchical Clustering model with two clusters (preprocessed wine quality data set).

As illustrated in Figure 4.2.1, cluster “0” is dominated by white wine samples whereas cluster “1” mainly consists of red wine samples, like seen for the K-Means model (two clusters). In the case of Agglomerative Hierarchical Clustering, 4,755/4,898 white wine samples were clustered correctly in comparison to the 4,811/4,898 which were assigned to cluster “0” by K-Means (or 1.14% more). Furthermore, the Agglomerative Hierarchical Clustering model also properly clustered 1,568/1,599 red wine samples, in contrast to the 1,576/1,599 which were clustered correctly by K-Means (or 0.50% more).

Finally, Figure 4.2.2 shows a comparison of clustering performance between the Agglomerative Hierarchical Clustering and K-Means models with two clusters. Similarly to the course laboratory, results for each wine color seem fairly consistent. For its part, Figure 4.2.3 shows the dendrogram of the Hierarchical Clustering model. It is shown here for documentation purposes in case this model is chosen for further development later.

color	agglom	kmeans	total
red	0	0	18
		1	13
	1	0	5
		1	1563
white	0	0	4717
		1	38
	1	0	94
		1	49

Figure 4.2.2. Result comparison between the Agglomerative Hierarchical Clustering and K-Means models with two clusters (preprocessed wine quality data set).

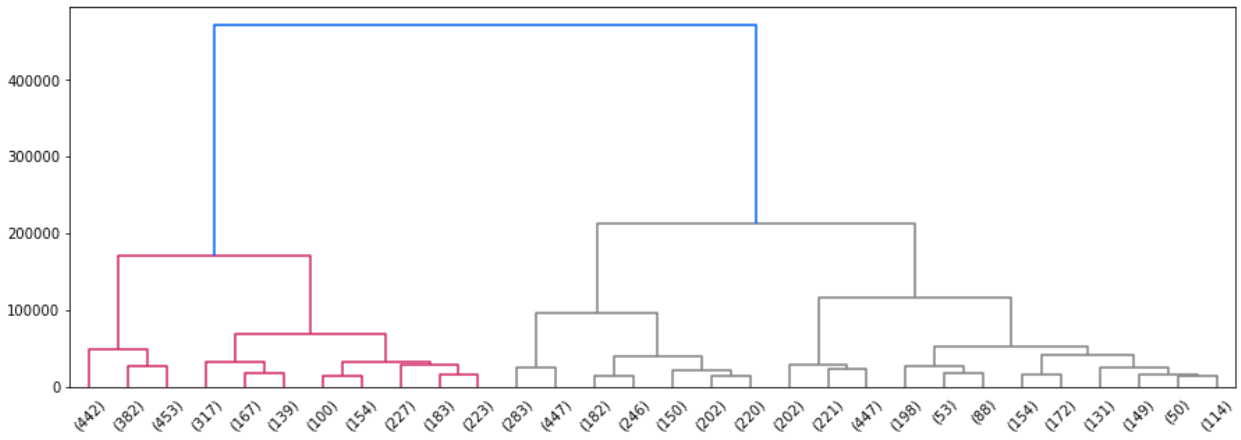

Figure 4.2.3. Agglomerative Hierarchical Clustering model: dendrogram (preprocessed wine quality data set).

4.3. DBSCAN

Based on the size of the clusters formed by the previous models and the observed behavior in the pair plots (Section 3), the clustering algorithm that was considered next for this project was DBSCAN, which also detects outliers. Optimizing for the hyperparameters of this algorithm is difficult so the values that were used were the same ones that were suggested in the course.

In particular, for the DBSCAN model developed for this project using Scikit-learn, eps=3 and min_samples=2. Furthermore, the usage of DBSCAN was considered to be exploratory, specifically to see whether further insight could be gathered about the clusters in the data set. To develop this model, the same 6,497 preprocessed samples were employed considering only the float64 columns. The results are shown in Figure 4.3.1.

color	dbscan	total
red	-1	5
	0	1590
	1	4
white	-1	10
	0	4886
	2	2

Figure 4.3.1. Results of the DBSCAN model (preprocessed wine quality data set).

As shown in Figure 4.3.1, some outliers were detected by DBSCAN (cluster “-1”), based on the definition of an outlier by the DBSCAN algorithm. In contrast to the previous models, DBSCAN formed three clusters: “0”, “1” and “2”. It is noticeable that only red wine samples belong to cluster “1” while only white wine samples belong to cluster “2”. However, the majority of both types of wine fall in cluster “0”.

Therefore, this suggests that, at least with respect to using the chemical properties to separate the data into distinct clusters based on wine color, the DBSCAN algorithm does not perform very well with the current hyperparameter selection. Toward this purpose, it will be

relevant to either optimize the values of the hyperparameters or study the density of both target clusters in more detail, as well as the separation between both clusters.

However, the results obtained using DBSCAN do suggest that there might be more clusters in the data (other than just two) and/or that there might be other ways of combining the chemical properties to cluster the data. Furthermore, this implies these clusters might not be based on wine color only.

5. MODEL DISCUSSION AND RECOMMENDATION

Based on the results that were reported and discussed in Section 4.1 (K-Means), Section 4.2 (Agglomerative Hierarchical Clustering) and Section 4.3 (DBSCAN), the recommended model is the K-Means model with two clusters. In particular, the K-Means model best suits the objective of this analysis, since it is the option that has the best clustering performance with respect to the preprocessed wine quality data set and the chosen number of clusters. Specifically, K-Means was able to separate data in two clusters in a sufficiently clean way. It created cluster “0”, which consisted mainly of white wine samples, and cluster “1”, which mostly consisted of red wine samples. Using the color variable, it was verified that 4,811/4,898 (98%) white wine data points were assigned to cluster “0” while 1,576/1,599 (99%) red wine data points belonged to cluster “1”.

These results were higher than those of Agglomerative Hierarchical Clustering with respect to the two clusters, which resulted to be based on wine color as this analysis hoped to find. Considering DBSCAN, this algorithm placed the majority of the white and red wine data samples in the same cluster (“0”) and created two more clusters (“1”, “2”) as addressed in Section 4.3. In this sense, this performance is not satisfactory with respect to clustering based on wine color, so K-Means is also preferable towards this purpose. However, as mentioned earlier, the results of DBSCAN do suggest that there might be other ways of clustering the data based on the chemical properties of wine, for which further study would be required (possible next step). Since non-color based clustering performance cannot be properly understood at the moment, the DBSCAN model is not recommended at this stage.

6. KEY FINDINGS AND INSIGHTS

In summary, after performing this analysis, the key findings and insights are:

- Considering the results presented in Section 4, K-Means and Agglomerative Hierarchical Clustering with two clusters had the best clustering performance. This was verified using the color variable. Moreover, in this context, this means that these two models clustered the wine data in a sufficiently clean way according to wine color (this is related to the main objective of this project).
    - As this analysis hoped to find, the two clusters were, in fact, based on wine color. Therefore, it can be said that the chemical properties of wine can be used to relate data samples to wine color (at least for this data set).
    - When comparing the K-Means and Agglomerative Hierarchical Clustering models, the preferred model is K-Means (Section 5). In particular, K-Means was able to assign 4,811/4,898 (98%) white wine data points to cluster “0” and 1,576/1,599 (99%) red wine data points to cluster “1”. These results were higher than the ones obtained using Agglomerative Hierarchical Clustering. In other words, K-Means was the model that was better able to separate red from white wine samples based on the chemical properties of wine (features).
    - It should be noted that Agglomerative Hierarchical Clustering offers a good degree of explainability of the clustering process (see dendrogram in Figure 4.2.3).
    - On the contrary, the DBSCAN model was not able to properly cluster data based on wine color.
- Based on the results obtained using DBSCAN (Section 4.3), some outliers were detected. Furthermore, DBSCAN clustered data into three clusters, not two.
    - As mentioned in Section 4.3, this suggests that there might be other possible ways of combining the chemical properties of wine to cluster the data and/or that there might be other clusters which are not purely based on wine color.
    - At this stage, as explained in Section 5, the criteria used by DBSCAN to cluster the data is not completely understood yet so it cannot be recommended as the best model with respect to the objective of this analysis. Also, its clustering performance with respect to wine color is not satisfactory.
- Based on the results of Figure 4.1.2 (K-Means elbow curve of inertia against number of clusters) and Figure 4.1.3 (K-Means with four clusters), it can be seen that it is

possible to form more clusters in the data. In addition, as seen in Figure 4.1.3, at least with respect to wine color these four clusters were a bit more balanced. However, to properly interpret those results (Figure 4.1.3), it will be necessary to study the `float64` columns further.

- With respect to the chosen data preprocessing steps, these seem to have been appropriate based on the results, the chosen algorithms and the objective of this analysis (Section 1).

7. NEXT STEPS

The recommended model at this point, as stated in Section 5 and 6, is K-Means with two clusters. However, this model should be understood as a preliminary basis for further development. Moreover, some possible next steps are the following:

- An important next step is to increase the size of the data set to further validate the clustering results of this project. A possible flaw of the recommended model (K-Means with two clusters) is, precisely, that its performance might be affected by the fact that the data set is not very big (6,497 samples). Therefore, it is possible that the satisfactory results might be linked to this particular data set. In consequence, it is necessary to validate the clustering performance observed so far with more data.
- It would be interesting to expand this analysis and go deeper into metrics like inertia.
    - This would be especially relevant if considering a K-Means model with more clusters, for instance 3 or 4. As seen in Section 4.1, using 4 clusters resulted in a more balanced distribution of the wine color samples. Whether something similar will be done later will depend, however, on the new objective of a future version of this project.
- Studying the chemical properties of wine further will aid in the understanding of the results of the K-Means model with four clusters. Specifically, studying ranges of the chemical properties or working with subject matter experts could be useful.
- Deepening the understanding of the data set would also aid in the understanding of the possible clusters of the data set. As suggested by the results of DBSCAN, it is possible that there could be new ways of clustering the wine quality data set. Furthermore, these new ways need not be based on wine color only.

    - Depending on future requirements, perhaps having more than two clusters could be more useful so properly understanding the nature of the data set is a very relevant issue.
- Depending on the objective or focus of a future version of this project, if the DBSCAN algorithm is chosen, it will be necessary to optimize its hyperparameters or study the density of the target clusters in more detail, as well as the separation between said clusters (Section 4.3).
- In a future version of this analysis, it would be beneficial to properly check for the calculation requirements of the correlation coefficients using hypothesis testing (Zach, 2021).

REFERENCES

Zach. (November 17, 2021). The Five Assumptions for Pearson Correlation [Blog entry]. Statology. https://www.statology.org/pearson-correlation-assumptions/