Rapport de projet

Compte-rendu du projet d'étude

Introduction aux traitements de données

MAM 3
Année 2024 - 2025

Polytech Nice Sophia

6 juin 2025

Rédigé par :
Yseult Canac–Pons, Amiel Metier, Lucas Fert

Sommaire

1 Introduction **2**
Introduction 2

2 Team management **3**
2.1 Comment on a travaillé 3
2.2 Emploi du temps 3

3 Étude du jeu de données **4**
3.1 Description des données 4
3.2 Tri des données 4

4 Analyse des données **6**
4.1 Corrélation entre les données explicatives et la qualité 6
4.2 Matrice de corrélation 10
4.3 Régression linéaire 10

5 Conclusion **12**

1 Introduction

Dans un contexte où les données constituent une ressource stratégique que des plateformes comme **Kaggle** jouent un rôle essentiel pour apprendre la science des données. Kaggle propose en effet un grand nombre de jeux de données publiques, permettant d'aborder des problèmes très variés, concrets et d'actualité.

Pour ce projet, nous avons choisi un sujet à la fois original et scientifiquement pertinent : la qualité du vin rouge. Le vin est un produit complexe, influencé par de nombreux facteurs chimiques et physiques. Comprendre ce qui fait la qualité d'un vin rouge à partir de ses caractéristiques chimqiue est un défi intéressant.

Notre problématique principale est donc : Quelles caractéristiques mesurables d'un vin rouge influencent significativement sa qualité, et peut-on modéliser cette qualité à partir de ces données ?

Afin de répondre à cette question, nous allons adopter une démarche rigoureuse inspirée du travail d'un data analyste. Dans un premier temps, nous effectuerons un prétraitement des données, afin d'en assurer la qualité, d'éliminer les valeurs aberrantes et de mieux comprendre leur structure. Cette phase exploratoire sera suivie d'une analyse descriptive, visant à identifier les tendances générales et les corrélations éventuelles entre les différentes variables.

Nous poursuivrons ensuite avec une régression linéaire, afin de modéliser notre modèle estimant la qualité d'un vin rouge. Enfin, nous discuterons les limites de notre approche, notamment en ce qui concerne la subjectivité de la notion de qualité gustative, et la représentativité du jeu de données.

2 Team management

2.1 Comment on a travaillé

Dans ce projet, nous avons adopté une approche collaborative répartie sur plusieurs phases : recherche d'un jeu de données, nettoyage des données, analyse statistique et modélisation. Chacun des membres de l'équipe a pris en charge une ou plusieurs étapes spécifiques. Un partage de code via un dépôt GoogleColab nous a permis de suivre l'évolution du projet de manière fluide et efficace.

2.2 Emploi du temps

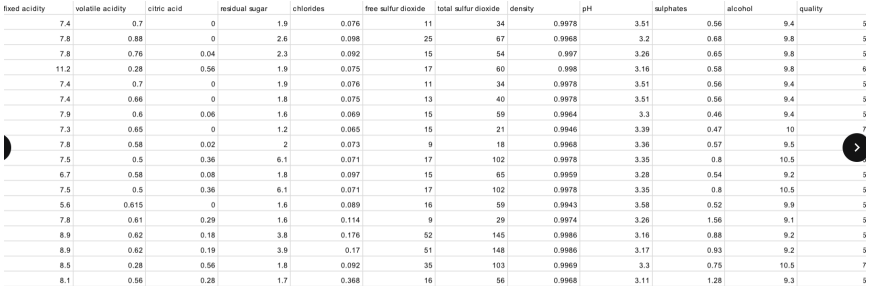
Le projet a été organisé sur 4 jours. Nous avons réparti les tâches en quatres grandes étapes :

— **Jour 1** : Recherche et nettoyage des données.
— **Jour 2** : Analyse statistique et visualisation.
— **Jour 3** : Régression linéaire, interprétation et diaporama.
— **Jour 4** : Rédaction du rapport.

3 Étude du jeu de données

3.1 Description des données

Le jeu de données utilisé contient 1599 observations et 12 colonnes représentant différentes caractéristiques physico-chimiques du vin rouge, telles que l'acidité, la teneur en alcool ou encore la quantité de dioxyde de soufre.



fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5

FIGURE 3.1 – Aperçu du jeu de données brut

3.2 Tri des données

Les outliers (valeurs aberrantes) peuvent fausser l'analyse statistique, notamment lors de la modélisation linéaire. Afin de les supprimer, nous avons utilisé la méthode de l'IQR (Interquartile Range).

$$IQR = Q3 - Q1$$

$$\text{Seuils} = [Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR]$$

Les observations situées en dehors de cet intervalle ont été considérées comme aberrantes et supprimées. Ce nettoyage a permis d'éliminer 405 observations, soit environ 25% du jeu de données initial.

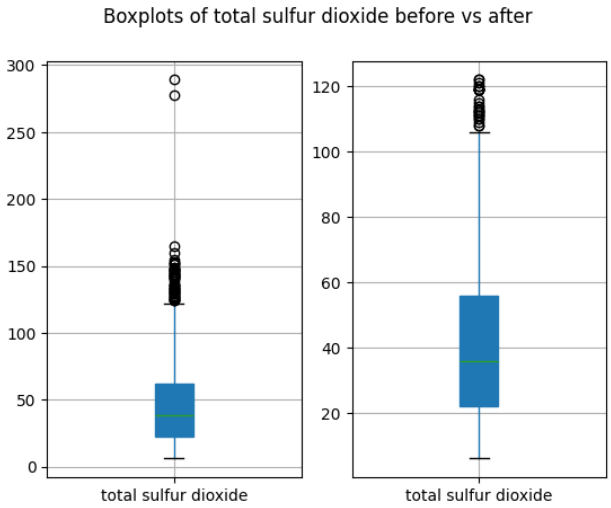


FIGURE 3.2 – Diagramme à boîtes avant/après suppression des outliers pour le dioxyde de soufre total

Comme on le voit ci-dessus, la suppression des valeurs extrêmes permet une visualisation plus claire et plus réaliste de la distribution des données.

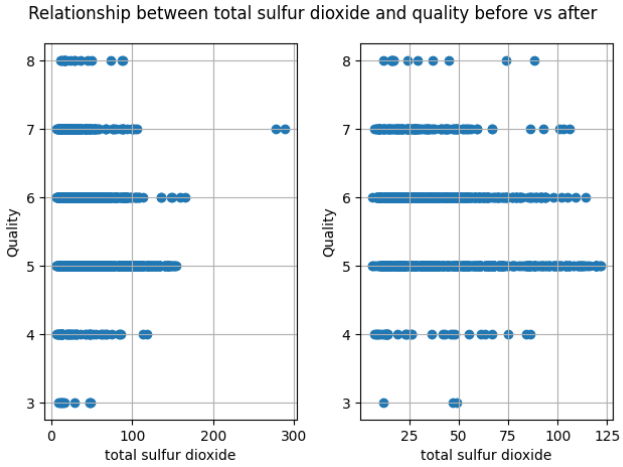


FIGURE 3.3 – Nuage de points : qualité vs dioxyde de soufre total (après nettoyage)

Le graphique montre une distribution plus homogène des points, rendant les relations entre variables plus lisibles.

4 Analyse des données

4.1 Corrélation entre les données explicatives et la qualité

Avant toute modélisation, il est pertinent d'examiner la répartition des scores de qualité.

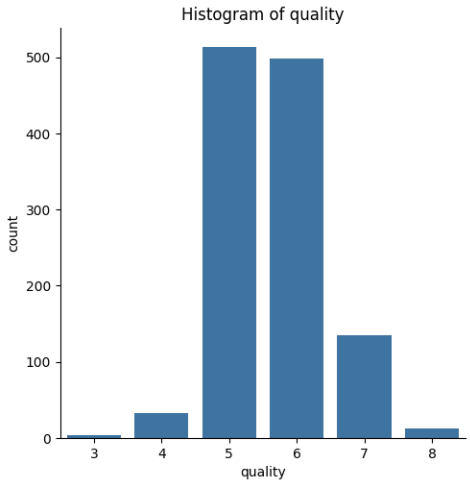


FIGURE 4.1 – Distribution des niveaux de qualité

On constate une forte concentration autour des qualités 5 et 6, correspondant à des vins de qualité moyenne. Les vins très bons ou très mauvais sont peu représentés, ce qui peut biaiser l'apprentissage du modèle.

Nous avons ensuite visualisé les relations entre certaines variables et la qualité (les noms perntiantes n'étant pas montrer dans le rapport) :

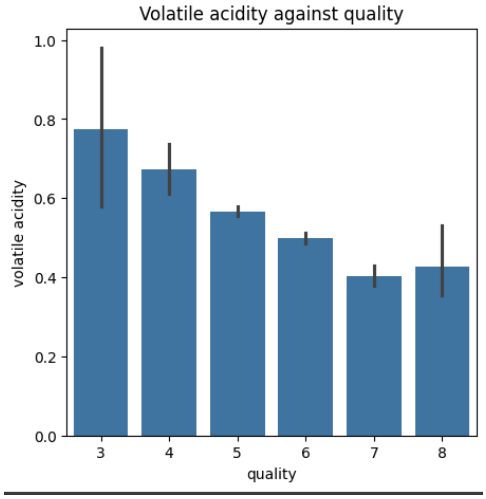


FIGURE 4.2 – Relation entre l'acidité volatile et la qualité

L'acidité volatile diminue lorsque la qualité augmente, indiquant une corrélation négative.

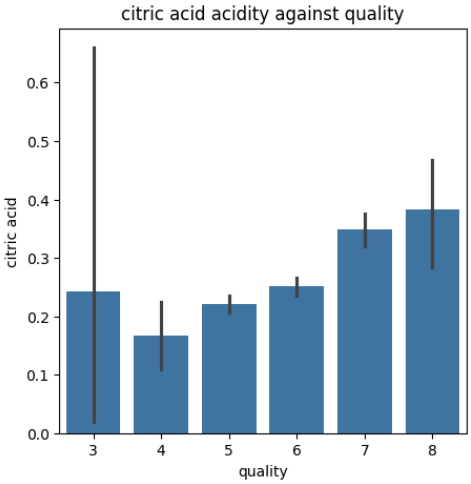


FIGURE 4.3 – Relation entre l'acide citrique et la qualité

On observe ici une légère corrélation positive entre l'acide citrique et la qualité.

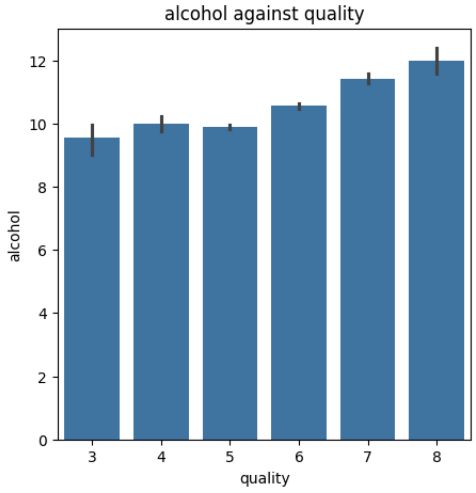


FIGURE 4.4 – Relation entre l'alcool et la qualité

Cette variable présente une corrélation plus marquée : les vins ayant une plus forte teneur en alcool semblent être mieux notés.

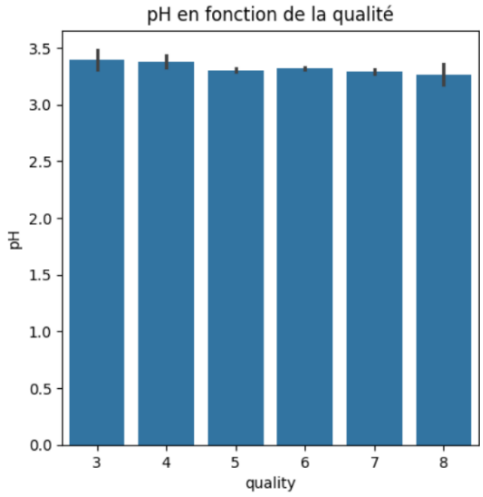


FIGURE 4.5 – Relation entre le pH et la qualité

La relation entre pH et qualité est plus difficile à interpréter, bien qu'une tendance puisse être perceptible.

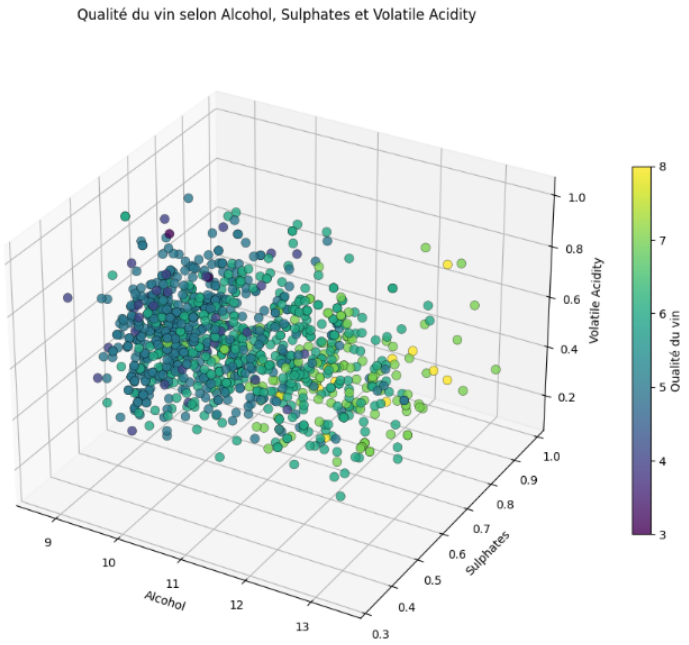


FIGURE 4.6 – Relation entre l'alcool, les sulfates, l'acidité volatile et la qualité du vin

Le graphique ci-dessus représente trois variables explicatives : l'alcool, les sulfates et l'acidité volatile en fonction de la qualité. La couleur des points correspond à la note de qualité du vin (de 3 à 8). On observe que :

— les vins avec une teneur plus élevée en alcool obitennent une meilleure note.

— les sulfates et l'acidité volatile jouent également un rôle positif.

On aperçoit une sorte de tendance linéaire où les plus mauvais vins sont à gauche et les meilleurs à droite.

Nous avons sélectionné les variables les plus significatives pour la suite de l'analyse : alcool, acide citrique, acidité volatile, pH, et un produit croisé que nous avons créé.

4.2 Matrice de corrélation

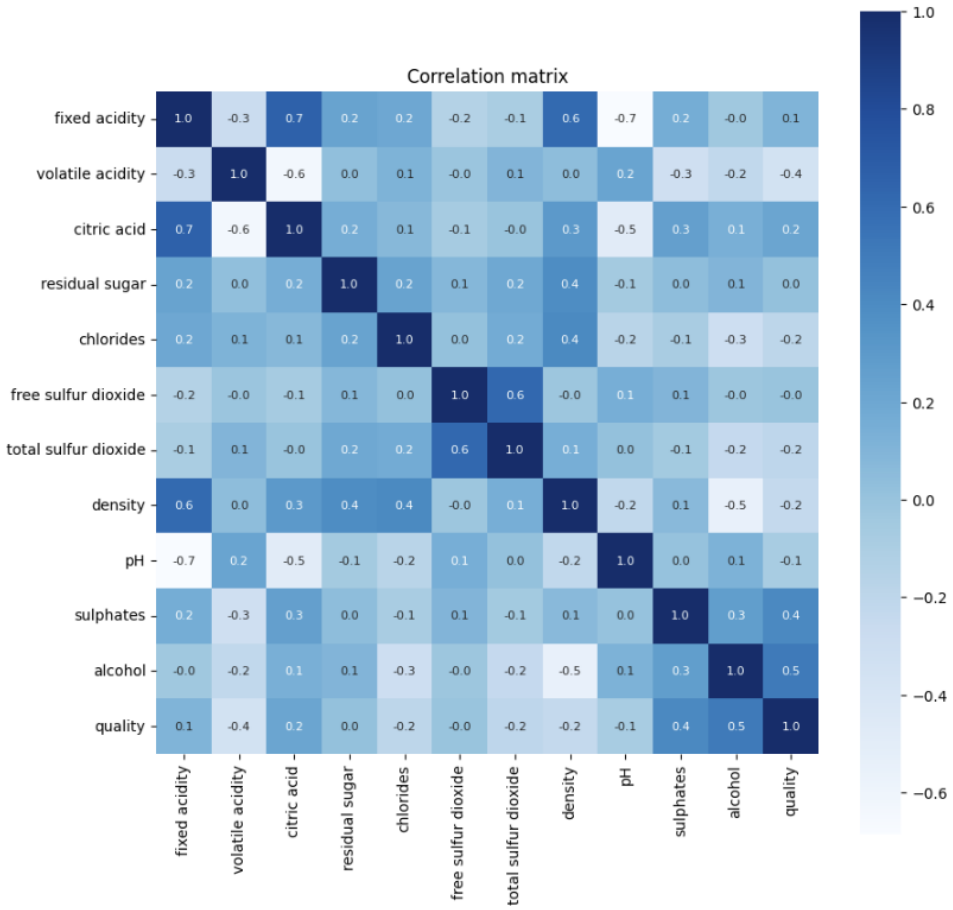


FIGURE 4.7 – Matrice de corrélation entre les variables

Cette matrice met en évidence les corrélations positives ou négatives entre les variables. Les plus fortes corrélations avec la qualité concernent l'alcool et l'acidité volatile, ce qui justifie leur inclusion dans la régression. Des corrélations internes apparaissent également (par exemple, entre acidité fixe et chlorure qui influent sur le pH permettant de faire varier plus rapidement celui-ci).

4.3 Régression linéaire

Nous avons entraîné un modèle de régression linéaire basé sur les variables les plus significatives.

Listing 4.1 – Régression linéaire sur les variables sélectionnées

```
wine_dataset['Fixed_Acidity X Chlorides'] = wine_dataset['fixed
    ↪  acidity'] * wine_dataset['chlorides']
X = wine_dataset[['alcohol', 'sulphates', 'volatile acidity', '
    ↪ pH', 'Fixed_Acidity X Chlorides']]
Y = wine_dataset['quality']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
    ↪ test_size=0.2, random_state=3)
model = LinearRegression()
model.fit(X_train, Y_train)
Y_pred = model.predict(X_test)
mse = mean_squared_error(Y_test, Y_pred)
n = X_test.shape[0]
p = X_test.shape[1]
r2 = r2_score(Y_test, (Y_pred))
r2_adjusted = 1 - ((1 - r2) * (n - 1)) / (n - p - 1)
```

Les résultats obtenus :

— **R^2 ajusté** : 0.402

— **Erreur quadratique moyenne (MSE)** : 0.387

Ils traduisent une performance peu fiable du modèle.

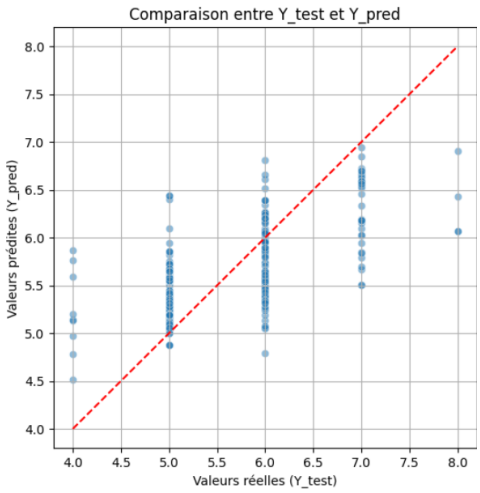


FIGURE 4.8 – Prédictions vs valeurs réelles

Le nuage de points montre que les prédictions sont très dispersées, montrant une faible capacité du modèle à estimer correctement la qualité des vins.

5 Conclusion

Notre modèle de régression linéaire n'est pas concluant. Plusieurs facteurs expliquent eela :

- **Distribution déséquilibrée** : très peu de vins de très bonne ou très mauvaise qualité, ce qui réduit la capacité du modèle à généraliser.
- **Manque de variables qualitatives** : les données sont exclusivement physico-chimiques. Or, la qualité d'un vin dépend aussi de critères subjectifs (arômes, texture, expérience gustative).

En réalité, la qualité d'un vin ne se limite pas à ses caractéristiques chimiques. Il est important de garder à l'esprit que des critères subjectifs jouent un rôle central dans l'évaluation gustative. Cela souligne la complexité d'un tel problème.