Rosse Mathys
M2 EA

Analyse de nos vins français

2024-2025

SOMMAIRE

1. Préambule

Nous savons qu'il existe mille et une façon de définir un vin et de le décrire. Dans cette étude, nous nous efforcerons de déterminer ce qui pourrait différencier un vin dit de soif (entrée de gamme) d'un vin de garde (moyenne gamme) et d'un vin de prestige (haut de gamme). Nous nous appuierons sur les adjectifs choisis pour caractériser un vin, car ils jouent un rôle crucial dans sa perception et sa valorisation, tant par les consommateurs que par les professionnels du secteur.

Pour cela, notre problématique consiste à étudier les descriptions faites pour chaque vin afin d'analyser si ces descriptions ont une incidence significative sur les notes attribuées. Cette approche nous permettra non seulement d'identifier les particularités des différents types de vins, mais aussi de mettre en lumière des facteurs communs qui pourraient favoriser leur popularité. En parallèle, nous examinerons les relations entre les adjectifs utilisés et les prix des vins, ainsi que les corrélations avec les scores attribués, ce qui pourrait révéler des dynamiques intéressantes entre la qualité perçue et la valeur marchande.

Nous utiliserons différentes librairies disponibles via Python pour traiter notre base de données et faire parler les descriptions de vins fournies par nos goûteurs, qui sont des passionnés, des professionnels ou encore des vignerons. En particulier, la librairie **'SPACY'** sera employée pour le traitement linguistique de nos descriptions, permettant ainsi une analyse fine des adjectifs et des thèmes récurrents. En analysant ces données, nous espérons obtenir des perspectives précieuses sur la manière dont les mots façonnent notre compréhension et notre appréciation des vins.

2. Notre base de données

Pour étudier nos vins, nous avons choisi d'exploiter une base de données assez conséquentes de **150 000 lignes** au total, téléchargée sur le site LabelYourData. Cette base de données contient 10 colonnes, soit **10 variables** différentes.

Variables :

country	Pays de création du vin
description	Description du vin
designation	Nom du vin
points	Note sur 100
price	Prix
province	De quelle province vient le vin
region_1	Première région
region_2	Seconde région
variety	Variété
winery	Domaine vinicole

2.1 Variables pertinentes

Afin de rendre les calculs plus légers et l'analyse plus pertinente, de nos **10 variables**, nous sommes passées au nombre de **4** : country, points, price, description. Ces variables constituent l'essentiel de notre étude et vont permettre d'affiner notre recherche.

Nous avons filtré via la variable country les vins français afin de centrer notre étude, ce qui nous donne une base de données de **21 098 lignes.**

Les vins rouges, les vins blancs, les champagnes et autres breuvages n'ont pas les mêmes saveurs ni les mêmes profondeurs. Il serait inapproprié de tous les comparer, cela biaiserait notre étude. De ce fait, nous allons re-catégoriser chaque vin afin de créer une 5ème variable que l'on va appeler '**Type**'.
Pour cela, un processus de **'scraping'** (technique d'extraction automatisée de données depuis des sites web) a été réalisé afin de collecter des informations sur différents types de vins à partir de divers sites internet. Cela nous a permis de classer nos vins de manière plus efficace en utilisant des données fiables. Nous avons identifié **6 catégories de vins** : vin rouge, vin blanc, assemblage rouge, assemblage blanc, vin rosé et vin mousseux.
Un assemblage est un vin élaboré en mélangeant plusieurs cépages (variétés de raisin) ; si cet assemblage est rouge, il peut être classé comme vin rouge (de même pour le blanc).
Les occurrences seront donc les suivantes : « Vin rouge », « Vin blanc », « Vin rosé » et « Vin mousseux ». Nous orienterons bien sûr notre étude principalement sur les vins rouges.
Notons qu'il y a eu **253 vins** qui ont été notés comme **'Inconnu'** par notre algorithme, ces vins seront écartés de l'étude.

2.2 Nettoyage de la base de données

Pour poursuivre le traitement de notre base de données, il est essentiel d'effectuer un nettoyage des valeurs nulles. Les lignes contenant des valeurs nulles dans nos variables pertinentes ont été supprimées afin d'assurer la fiabilité et la qualité de nos données. Les valeurs nulles, identifiées par « **Nan** », n'étaient présentes que dans la colonne '**price**', avec 6313 occurrences manquantes, qui ont donc été exclues.

Après ce traitement, notre base de données comptera **14 532 lignes**. En ne sélectionnant que les vins rouges, nous avons abouti à une base de données finale de **6 603 lignes.**

3. Données textuelles

4 étapes majeures ont été identifiées afin de pouvoir solliciter nos descriptions de vins.

3.1 Normalisation

Étant donné que les descriptions sont constituées de phrases ponctuées et saisies manuellement sur la machine, nous devons uniformiser le texte brut en une simple séquence de mots séparés par des espaces. Pour ce faire, nous supprimons la ponctuation en utilisant la bibliothèque **'string'**, qui permet d'accéder à la constante **'string.punctuation'**, contenant tous les caractères de ponctuation, afin de les retirer de chaque description. Les majuscules ont été converties en minuscules grâce à la méthode **lower()**. Ce traitement aboutit à une nouvelle colonne que nous avons appelée '**description_normalisee**'.

3.2 Stopwords

Pour supprimer les stopwords, nous avons utilisé la fonction '**`is_stop`**' de la bibliothèque '**spaCy'** qui est particulièrement adaptée à l'analyse de texte à grande échelle. Le modèle '**`en_core_web_sm`**' est employé pour segmenter le texte en **tokens** (on découpe la description). Chaque token représente un mot et contient plusieurs informations, dont un indicateur '**`is_stop`**' qui permet de savoir si le token est un stopword ou non. La fonction parcourt chaque tokens et seuls les tokens qui ne sont pas identifiés comme des stopwords sont conservés. Une fois cette filtration effectuée, les tokens restants sont rassemblés en une nouvelle chaîne de caractères, avec les mots séparés par des espaces, offrant ainsi une version épurée du texte d'origine. Ce traitement a été effectué sur la colonne '**description_normalisee**' qui a abouti à la création d'une nouvelle colonne que nous avons appelée '**description_stopwords**'.

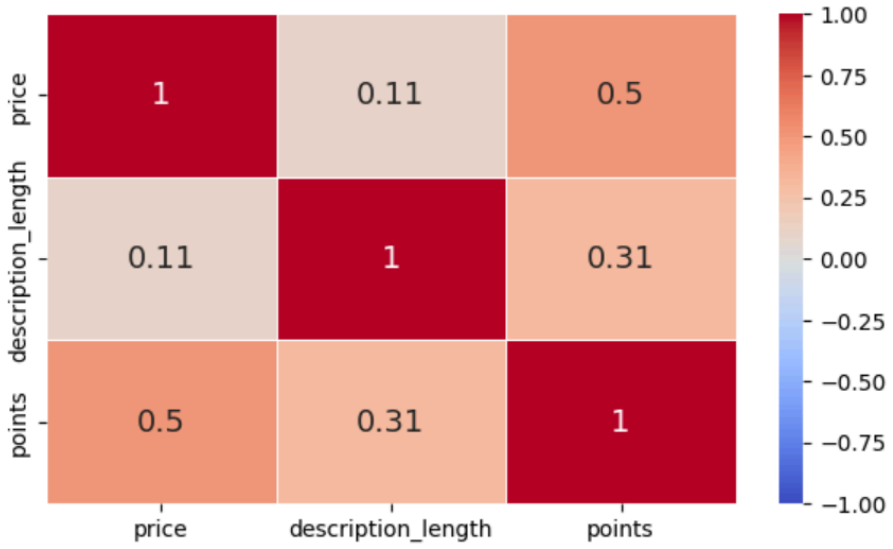
3.3 Lemmatisation & Tokenization

Avec la même démarche et grâce au modèle mentionné précédemment via spaCy, nous avons mis au point une fonction qui parcourt les tokens et identifie ceux appartenant à la

catégorie grammaticale des adjectifs (**ADJ**). Les adjectifs ainsi sélectionnés sont ensuite lemmatisés, c'est-à-dire réduits à leur forme de base, puis renvoyés sous forme de liste (ce qui est en fait notre tokenization). Cette méthode a été appliquée à la colonne **'description_stopwords'**, permettant de normaliser et de structurer les adjectifs dans les descriptions. Cela facilite leur analyse ultérieure, les résultats étant stockés dans une nouvelle colonne appelée **'description_lem_tok'** qui ne contient donc que des adjectifs.

3.4 : Relations entre nos variables

On pourrait supposer que des vins avec de longues descriptions et de nombreux adjectifs peuvent recevoir des scores élevés, car les descriptions complexes pourraient suggérer une analyse approfondie et donc un vin de haute qualité. Pour explorer cette hypothèse naturelle qui nous parvient instinctivement, nous avons épluché les corrélations entre nos 3 variables **quantitatives** qui sont le **prix**, le **score**, et la longueur des descriptions qui sera en réalité le nombre d'adjectifs descriptifs qui seront utilisés pour discuter d'un vin ; cette nouvelle colonne sera intitulée **'description_length',** avec une moyenne de 5 mots par description.
Nous avons calculé une matrice des corrélations grâce à la fonction '**corr()**' et nous avons créé une visualisation de nos résultats avec la fonction '**heatmap**' de la librairie '**seaborn**'.
Les corrélations varient de -1 à 1 : une corrélation de 1 signifie une relation positive parfaite (quand une variable augmente, l'autre aussi), -1 indique une relation négative parfaite (quand l'une augmente, l'autre diminue), et 0 signifie qu'il n'y a pas de relation linéaire entre les deux variables.



Finalement, nous constatons que la longueur des descriptions a bien un impact sur le score du vin, mais n'a presque aucune influence sur le prix (corrélation de 0.11). En revanche, le prix est fortement corrélé au score (0.5). Cela nous amène à nous demander : un vin très bien noté n'est-il pas toujours plus cher ? Toutefois, la relation entre prix et score pourrait ne pas être **strictement linéaire**. En effet, des vins de gamme moyenne peuvent obtenir des scores proches de ceux des vins haut de gamme, car le prix ne reflète pas toujours directement la qualité. Des éléments comme la rareté, la région, ou **d'autres facteurs** peuvent également entrer en jeu. Il est possible qu'une relation linéaire entre score et longueur des descriptions existe jusqu'à un certain seuil, au-delà duquel une hausse de la

longueur des descriptions n'entraîne plus forcément une augmentation proportionnelle du score.

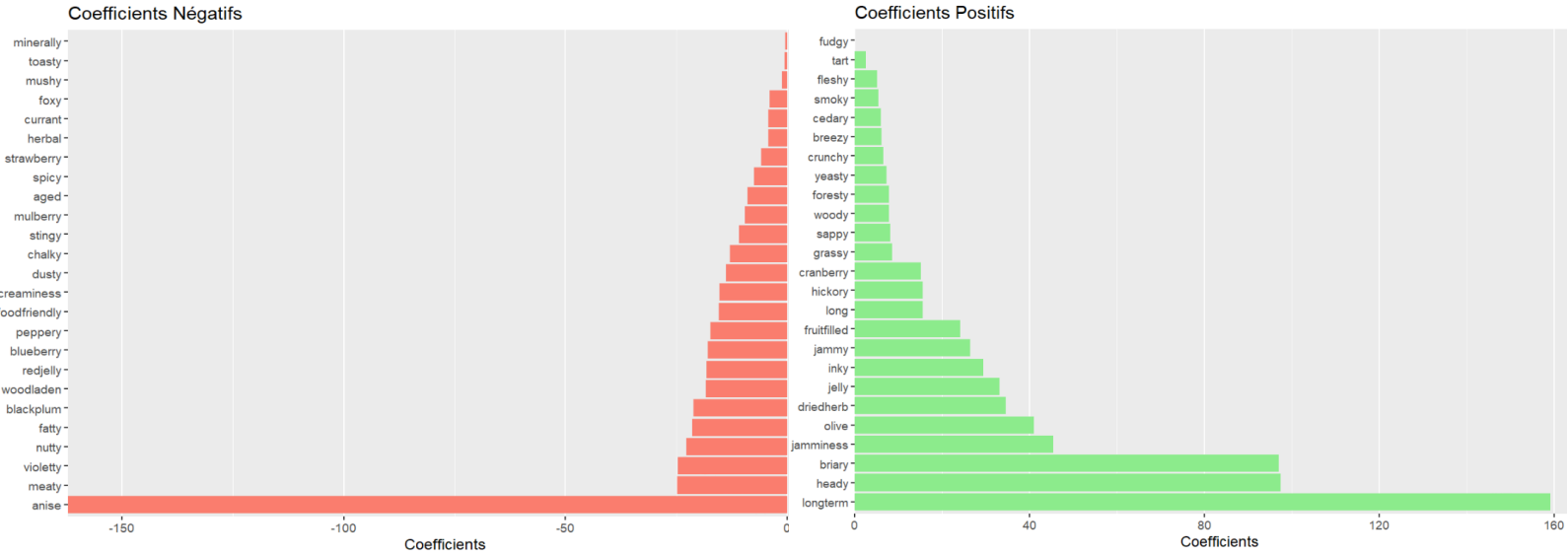
C'est justement ici que l'étude du champ lexical peut éclairer ce phénomène : certains mots pourraient-ils contribuer à faire grimper le prix d'un vin ?

4. Le champ lexical du vin

4.1 Régression

La régression appliquée à notre modèle nous en dit plus sur les adjectifs qui affectent le prix du vin par leur présence dans leurs descriptions.

La fonction '**CountVectorizer'** de la librairie '**scikit-learn'** transforme le texte en une matrice de présence/absence de mots (prenant la valeur 1 si le mot est présent dans la description du vin, 0 sinon) où chaque ligne représente un vin et chaque colonne un adjectif. Cette matrice va servir pour la régression qui aura comme variable à expliquer le prix du vin, et comme variables explicatives du prix, les adjectifs des descriptions. Une fois que l'on a initialisé et ajusté notre modèle aux données, nous avons les coefficients associés à chaque adjectif (ici choisis aléatoirement). Ces coefficients montrent si la présence de certains adjectifs dans une description est associée à un prix de vin plus élevé (ou plus bas).



Ici, le coefficient d'environ 160 (159.14) pour le mot **longterm** signifie que lorsque ce mot apparaît dans la description d'un vin, le modèle prédit que le prix du vin est **en moyenne plus élevé de 160 unités monétaires**, toutes choses égales par ailleurs.

Ces coefficients vont nous permettre de classifier les vins en trois catégories : bas, moyen et haut de gamme en les **sommant** via leurs apparitions ou non dans les descriptions. Cela nous aidera pour étudier la relation entre le vocabulaire associé aux vins haut de gamme et

celui des vins bas de gamme, ainsi qu'à déterminer si une catégorie intermédiaire relie les deux.

4.2 Catégorisation

Les statistiques descriptives montrent que 50% des vins sont en dessous de 25 unités monétaires, 25% entre 25 et 50 unités, et 25% au-dessus de 50. Ces données serviront à définir nos seuils pour nos groupes : bas, moyen et haut de gamme. Nous avons dès lors une nouvelle colonne dénommée '**Category**'.
Pour chaque catégorie de vin, nous avons extrait et trié les adjectifs par fréquence d'apparition par ordre décroissant, ce qui permet de visualiser les adjectifs les plus courants dans chaque catégorie.

Cependant, notre étude se concentre sur les sensations gustatives du vin, et des adjectifs trop généraux comme 'bon' ou 'excellent' ou encore 'black' qui ne nous apportent aucune information utile. C'est pourquoi nous avons mis en place une fonction visant à **filtrer** les adjectifs dans les descriptions, en ne conservant que ceux qui sont pertinents pour décrire les arômes, sensations et saveurs spécifiques du vin. Par exemple, nous avons conservé des mots dans le champ lexical des fruits de la forêt, ou encore simplement des adjectifs se terminant en **'y'** pour capter des termes comme 'dusty'. Il définit notre dernier filtre en plus du filtre sur la colonne gardant simplement les adjectifs. On passe de **5795** adjectifs uniques, à **1046,** ce qui nous fait une différence de **4749** jugés pertinents. Les **1046 adjectifs indésirables** seront stockés dans une liste appelée **'generic_adjectives'**, qui nous permettra d'exclure ces mots lors du calcul des fréquences d'apparition.

Pour une raison de lisibilité, nous affichons que les **10 adjectifs les plus fréquents par catégories :**

	Bas de gamme	Moyenne gamme	Haut de gamme
1	tart	firm	heady
2	plummy	jelly	briary
3	early	long	long
4	cedary	inky	chunky
5	fleshy	creamy	jamminess
6	dry	dry	longterm
7	juicy	spicy	panopoly
8	dusty	ageworthy	smoky
9	herbal	olive	stingy
10	earthy	jammy	plenty

En somme, les mots associés aux vins sont révélateurs des perceptions de qualité et de valeur des testeurs par leurs coefficients associés. Les adjectifs ayant un impact **négatif** sur le prix, se retrouvent souvent chez les plus grandes fréquences d'apparitions pour les vins bas de gamme et contrastent avec ceux qui sont valorisés dans les catégories moyenne et haut de gamme.

Les résultats de la régression confirment l'importance de ces adjectifs dans l'évaluation des prix et soulignent le rôle de certaines caractéristiques perçues comme essentielles à la qualité (comme '**long**' par exemple). Une observation intéressante pour les vins de **'bas de gamme'** est la forte présence de l'adjectif **'early'**, qui pourrait indiquer que ces vins sont perçus comme trop jeunes ou insuffisamment mûrs.

En revanche, dans la catégorie '**moyenne gamme**', les adjectifs comme **'ageworthy'** suggèrent un vin dont le potentiel de vieillissement est reconnu, marquant une évolution vers des vins plus matures. Pour les vins **'haut de gamme'**, bien que l'adjectif **'aged'** ne figure pas dans la liste principale, il est mentionné plus loin, renforçant l'idée d'une continuité de maturation et d'amélioration avec l'élévation de la gamme.

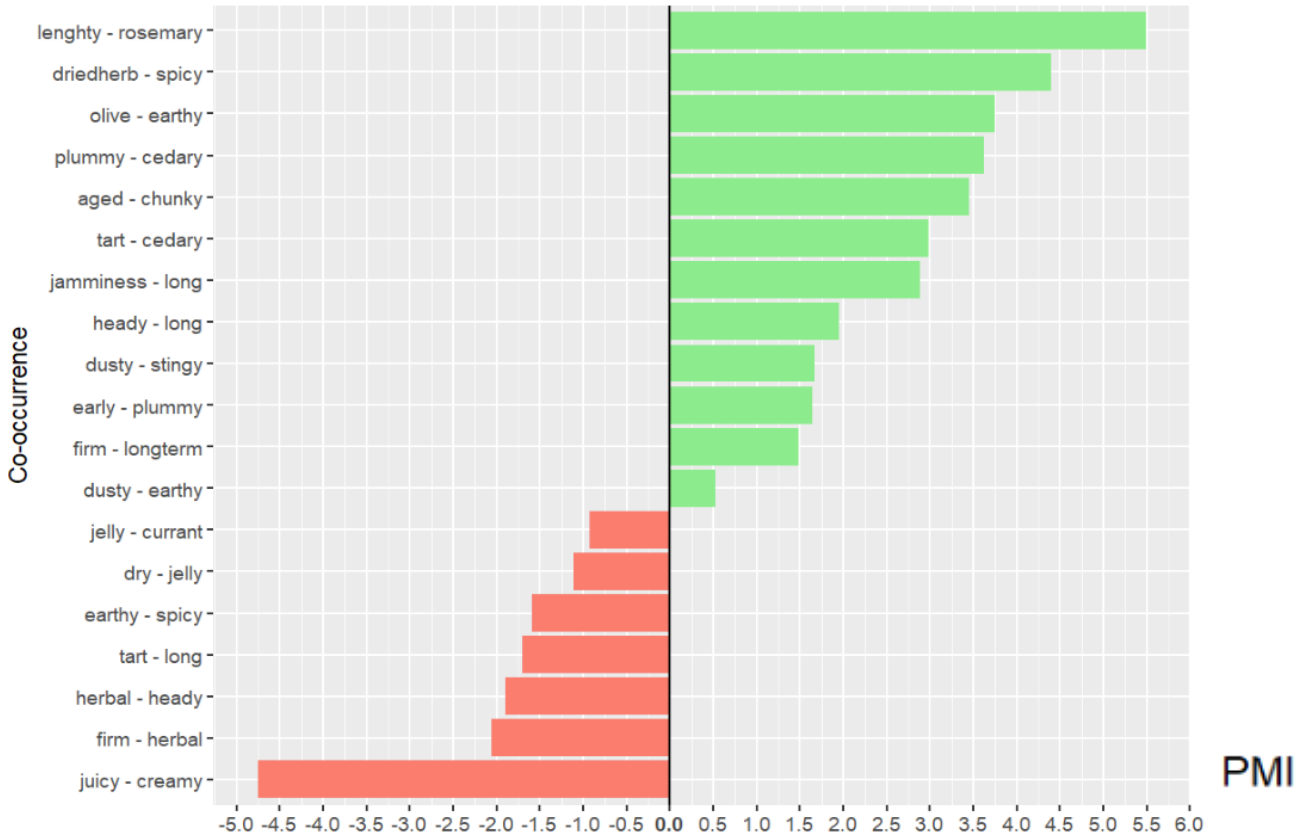
4.2 Co-occurences & PMI

On va ensuite mettre en place une fonction chargée d'analyser les adjectifs utilisés pour décrire les vins en calculant leur **co-occurrence**, c'est-à-dire combien de fois deux adjectifs apparaissent ensemble dans une même description.

Ensuite, on utilise une mesure appelée **PMI (Pointwise Mutual Information)** pour identifier les adjectifs qui ont des associations fortes.

La PMI utilise ces co-occurrences pour mesurer la force de l'association entre deux éléments en comparant leur probabilité conjointe à leurs probabilités individuelles. Plus la PMI entre deux adjectifs est élevée, plus ils sont souvent utilisés ensemble de manière significative. L'objectif est d'analyser si le vocabulaire utilisé pour décrire les différentes catégories de vins est distinct. En d'autres termes, nous cherchons à déterminer si le vocabulaire des vins haut de gamme se croise ou non avec celui des vins bas de gamme, par exemple.

L'exportation de la **matrice de calculs de la PMI** vers **Excel** facilitera l'analyse en permettant l'utilisation de filtres. Pour plus de clarté, nous avons sélectionné les relations les plus pertinentes à présenter :



Les PMI révèlent des **différences** marquées entre les adjectifs utilisés pour les vins bas, moyens et hauts de gamme. Les adjectifs des vins bas de gamme tendent à se caractériser par des termes plus **simples** ou déséquilibrés (comme "tart" (acidulé) ou "herbal"), tandis que ceux des vins haut de gamme soulignent la complexité et la **persistance des saveurs** ("heady", "long"). La moyenne gamme quant à elle, joue un rôle de **transition** où des adjectifs comme "firm" ou "olive" peuvent se retrouver tant dans des vins plus simples que dans des vins plus élaborés.

Les associations négatives ou faibles entre des adjectifs de bas et haut de gamme illustrent bien la **rupture qualitative** entre ces catégories et les différentes profondeurs apportées, alors que certaines associations positives dans la moyenne gamme signalent un équilibre et une complexité intermédiaires.

Finalement, la manière dont le vocabulaire et les adjectifs utilisés dans les descriptions de vins influencent non seulement la perception de la **qualité**, mais aussi le **prix** et le **score** attribués. Les adjectifs jouent un rôle crucial dans la catégorisation des vins et reflètent des caractéristiques communes dans chaque gamme. Les associations positives ou négatives entre les adjectifs révèlent également des dynamiques intéressantes, notamment la transition subtile mais significative entre les vins de moyenne et haut de gamme.

Bien que cette étude ne soit pas une science exacte, elle peut servir de base pour explorer de nouvelles pistes dans de futurs projets d'analyse de vins !