Aumasson Yanis
Bohelay Corentin

Compte-rendu : Data-Mining

1) Présentation du Dataset

Le dataset que nous avons choisi provient d'une collecte de données effectuée sur le site WineEnthusiast en juin 2017. On peut y trouver de nombreuses informations détaillées sur divers aspects des vins listés, incluant les notes, les prix, les variétés, et plus encore, reflétant une image globale du marché du vin à ce moment.

Détails des variables :

- Country : Pays d'origine du vin.
- Description : Description sensorielle du vin.
- Designation : Nom du vignoble ou de la parcelle.
- Points : Note attribuée au vin, elles vont de 80 à 100.
- Price : Prix du vin en dollars.
- Province : Province ou état de production du vin.
- Region_1 et Region_2 : Zones spécifiques où le vin est produit.
- Taster_Name et Taster_Twitter_Handle : Nom du critique et son identifiant Twitter.
- Title : Titre complet du vin, incluant le millésime et le nom.
- Variety : Cépage du vin.
- Winery : Nom de la cave ou du producteur.

2) Nettoyage des données

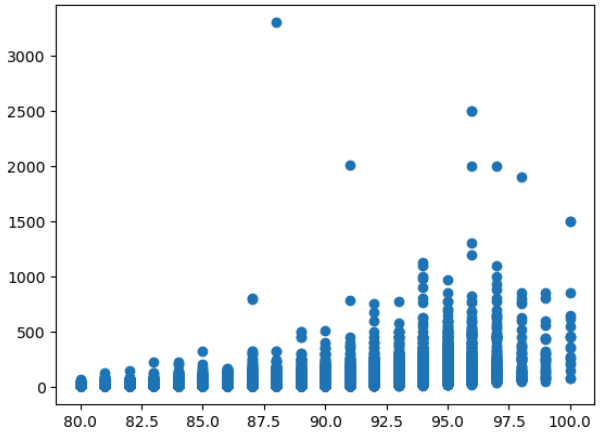
On va ici faire une exploration en préambule, des affichages divers pour s'approprier les données. On commence par drop les NaN, on considère que ne pas avoir de prix ou de score (les 2 seules catégories numériques que nous possédons) est disqualifiant pour une étude.

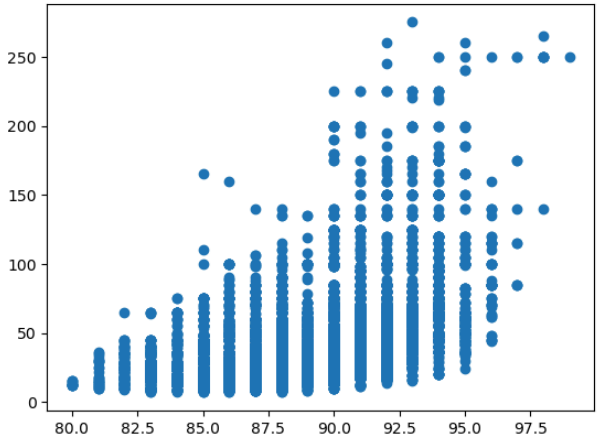
	points	price
count	129971.000	120975.000
mean	88.447	35.363
std	3.040	41.022
min	80.000	4.000
25%	86.000	17.000
50%	88.000	25.000
75%	91.000	42.000
max	100.000	3300.000

Premier élément d'analyse, on observe que les notes ne vont que de 80 à 100, le créateur du dataset a spécifiquement voulu analyser les meilleurs vins au monde. Même en ayant un dataset tronqué, on va remarquer que celui-ci possède des distributions assez intéressantes.

Recherche d'outliers :

Avant/Après filtrage par LocalOutlierFactor (n_neighbor = 20) :

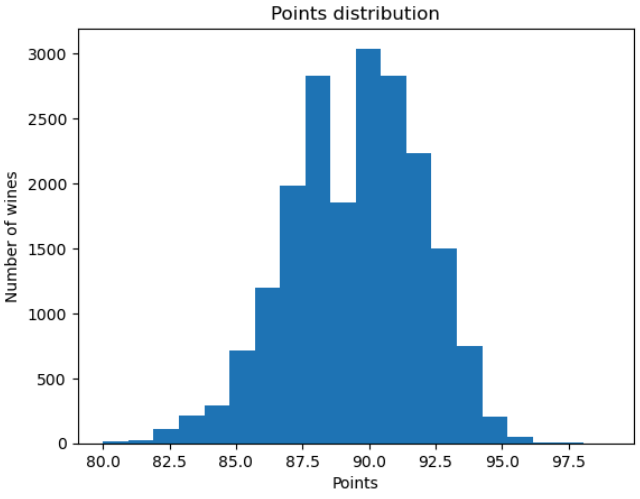




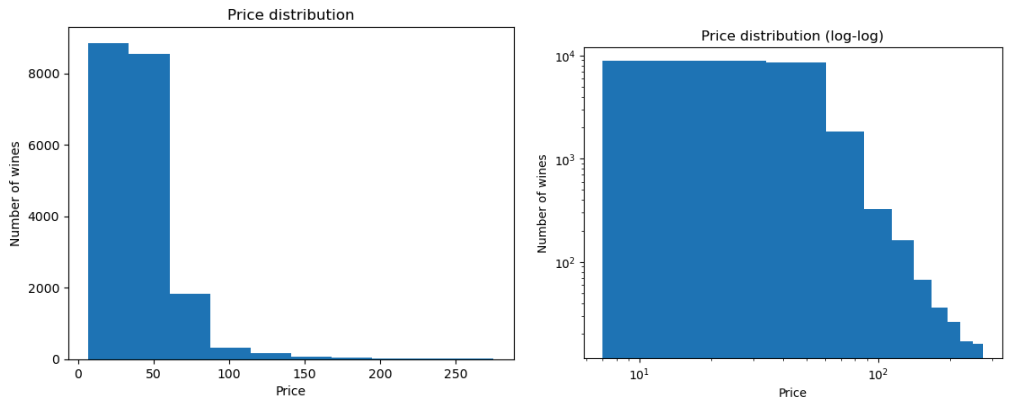
En plottant, on a des outliers visibles en ratio Prix/Note, sans les outliers nous avons une très jolie distribution presque linéaire entre prix et note. Cela pourrait exprimer une vraie corrélation entre prix et qualité, et non pas un simple effet "luxe" d'un vin plus cher. Inversement cette corrélation pourrait aussi signifier que les reviewer sont biaisés et on tendance a noter mieux des vins plus chers. La conclusion de cette analyse dépend donc de la confiance qu'on accorde à l'objectivité des reviewers.

Quoi qu'il en soit, les outliers ne sont pas extravagants, ce sont des vins a caractéristiques exceptionnelles, de niche. Ils seront gardés dans la plupart des analyses, sauf celles de pca et distribution, car elles faussent énormément l'affichage, et nous empêchent d'avoir une bonne image du marché.

Visualisations du dataset :



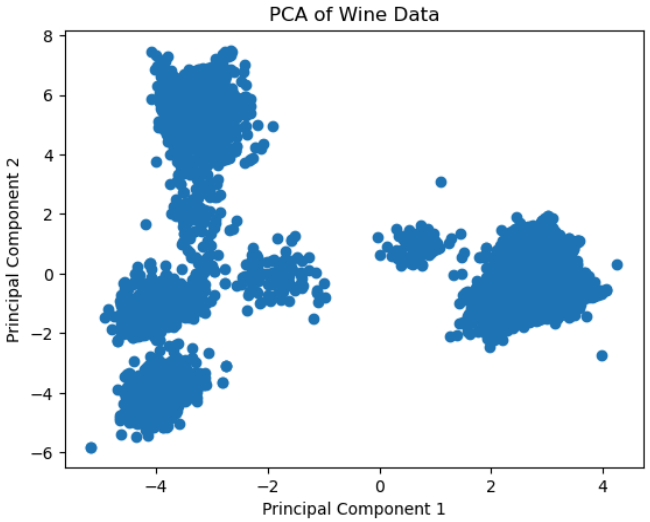
On observe une Bell Curve presque parfaite, ce qui est approprié pour un système de notation. Comme mentionné précédemment, le dataset est tronqué est ne considère que les meilleurs vins du site de reviews, cette distribution montre donc qu'au sein d'un sous ensemble assez particulier on a quand même une répartition équitable des notes. Les reviewers ont donc l'air de fournir des réflexions mûres et justes aux vins d'une certaine qualité. Le dataset complet aurait été approprié pour comparer, mais il n'était pas accessible sans scraping.



Encore une fois une distribution qui apparaît naturelle, on a supposé une power law pour la répartition des prix (voir gauche). Après avoir mis en échelle log-log (droit), on ne voit pas exactement une power law, mais un plateau de prix assez constant, qui progresse en power law. Le marché aurait donc un certain nombre de tranches de prix standard, puis des prix qui décroissent de façon naturelle.

3) Clustering

Nous avons tenté en premier lieu une PCA pour obtenir une visualisation en 2 dimensions du dataset. Celle-ci a nécessité une dummification de la plupart des catégories, qui était textuelles.
Après extraction des composantes de PCA on observe 4 facteurs majeurs dans la formation de clusters, le vignoble, la variété de vin, la région d'origine, et le producteur.

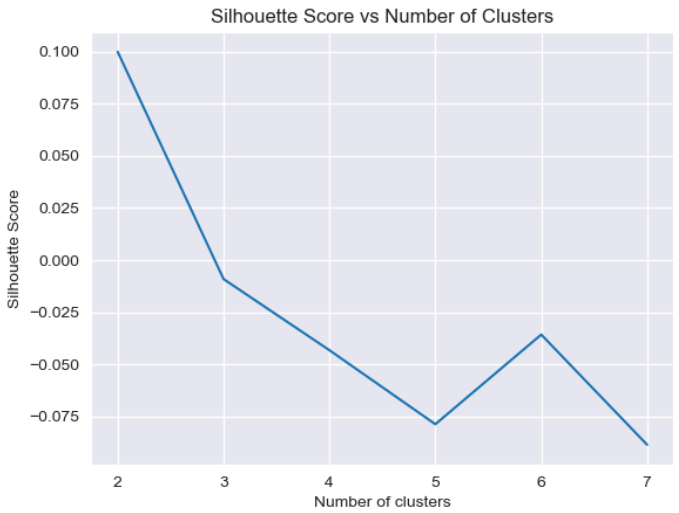


	PC1
winery	1.701
designation	1.000
region	0.669
title	-0.183
description	-0.159
province	-0.071
price	0.056
variety	-0.052
points	0.041
taster	-0.037
country	0.000

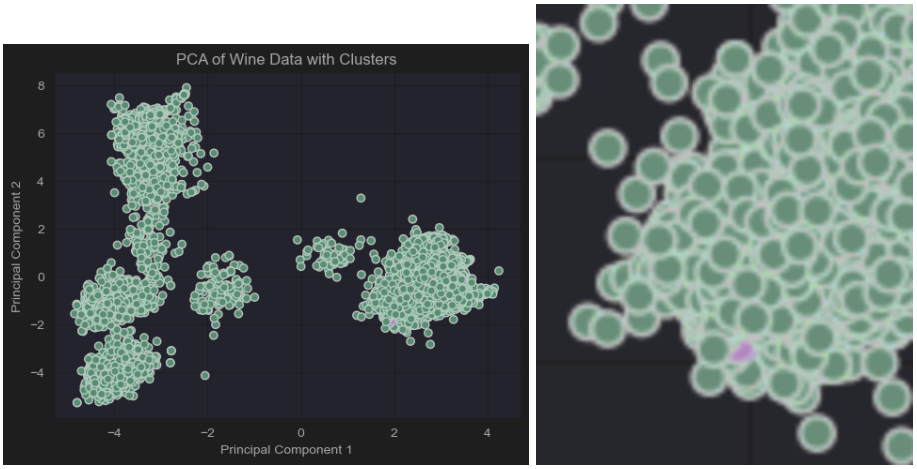
	PC2
designation	-0.559
variety	-0.282
region	0.180
title	0.143
description	0.136
province	0.053
points	0.031
taster	-0.029
price	0.028
winery	0.016
country	0.000

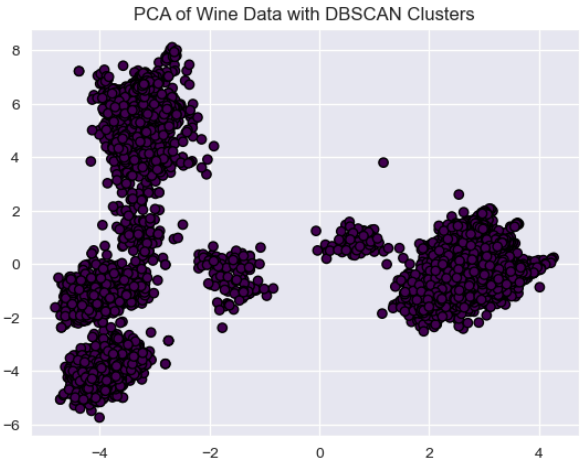
On voit que l'axe 1 de la pca place une forte corrélation positive a l'origine du vignoble et l'appellation, ou la réputation. L'axe 2 a l'inverse negationne l'appellation et même la variété spécifique du vin et corréle positivement la région. On peut donc imaginer que les clusters sont des regroupements de vins de vignobles similaires, et se distinguant par leurs positions géographiques.

On utilise ensuite le silhouette score pour déterminer le nombre optimal de clusters k-means:



Les silhouette scores sont peu indicatifs et mêmes confuses. Ils tournent autour de 0, donc random et indiquent donc une erreur dans le clustering, ou même dans les données, la PCA, je n'ai pas réussi à trouver la source de l'erreur.
Avec 2 ou 4 clusters de K-means nous obtenons ce type de clustering avec des points superposés, peu lisible.

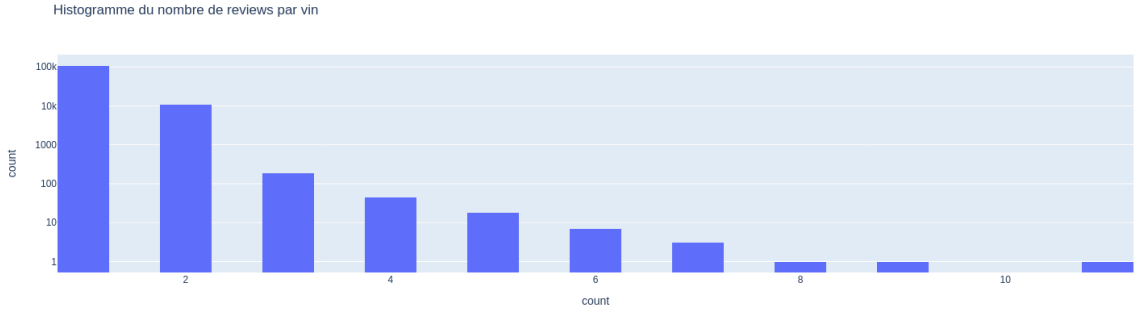




Également DBScan ne donne pas de résultats concluants. L'hypothèse est que nous avons mélangé les labels à un moment.

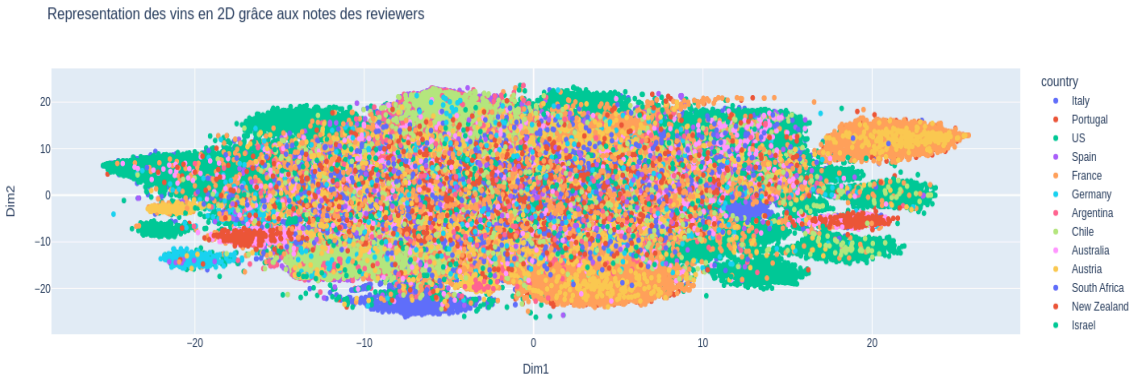
4) Recommendations

Pour mettre en place l'analyse par recommandations, nous avons commencé par créer un histogramme pour examiner la distribution du nombre de critiques par vin. Cette visualisation initiale nous a permis de constater que la majorité des vins sont évalués une seule fois, ce qui pose un défi en termes de densité de données pour les recommandations. Toutefois, nous avons tout de même choisi d'essayer cette approche.



Sur l'histogramme, l'échelle de l'axe Y est logarithmique pour mieux voir chaque barre, cependant la proportionnalité est perdue. Il faut donc bien prendre en compte la graduation de l'axe Y pour vraiment voir le déséquilibre.

Pour ce qui est du traitement des données, nous nous sommes concentrés exclusivement sur les notes attribuées aux vins. Pour ce faire, nous employons la méthode de décomposition en valeurs singulières (SVD), une technique de factorisation de matrices, pour obtenir 100 facteurs latents. Ensuite, nous utilisons l'algorithme t-SNE pour réduire ces facteurs à deux composantes principales, permettant ainsi une visualisation en deux dimensions ce qui facilite l'identification de groupes ou de similitudes dans les vins.

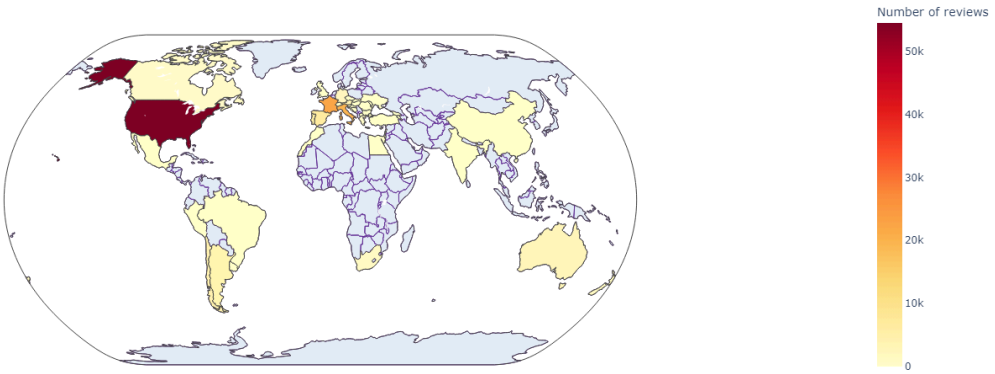


Le graphique en 2D ci-dessus représente la disposition des vins après réduction dimensionnelle, en utilisant les couleurs pour différencier les vins selon leur pays d'origine. Le centre du graphique apparaît assez homogène et moins distinctif, mais à plusieurs endroits, il est possible de distinguer des groupements spécifiques, notamment pour les vins allemands, français, italiens, néo-zélandais et américains. Par ailleurs, on observe que les vins français et autrichiens semblent présenter des similitudes, se regroupant souvent ensemble, ce qui suggère qu'ils pourraient être appréciés par des dégustateurs aux préférences similaires. Ces regroupements indiquent une certaine cohérence dans les caractéristiques ou les perceptions des vins selon leurs provenances. Ceci pourrait également s'expliquer par le fait qu'un critique de vins a tendance à évaluer plusieurs vins provenant du même pays.

5) Représentation spatiale

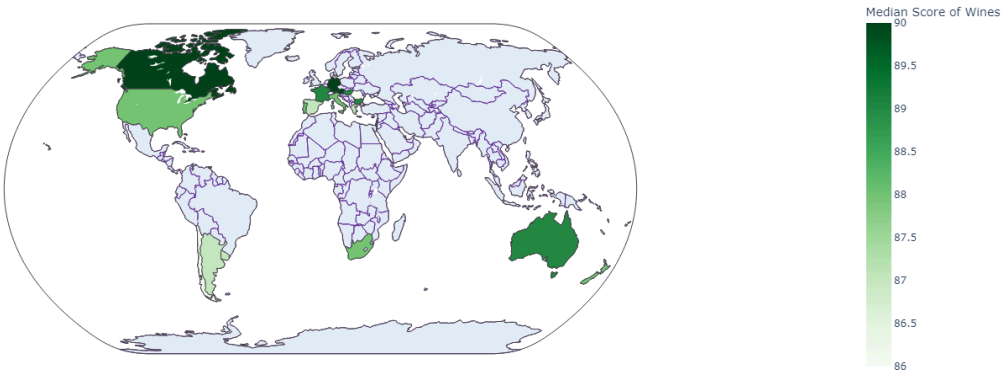
Disposant de données géographiques détaillées, nous avons décidé d'exploiter ces informations pour réaliser des représentations spatiales. Cette approche nous permet de visualiser de manière intuitive les tendances et les disparités de la critique de vins à travers le monde.

Nombre de reviews par pays



La carte ci-dessus illustre la répartition géographique du nombre de reviews de vins par pays. Comme on peut le constater, les États-Unis se distinguent nettement avec le nombre le plus élevé de critiques, ce qui est cohérent avec le fait qu'ils sont non seulement l'un des plus grands producteurs de vin, mais également parce que WineEnthusiast, étant une plateforme américaine, a une tendance naturelle à se concentrer davantage sur les vins locaux.
Il est également intéressant de noter la très faible présence de la Chine sur la carte, malgré le fait que ces dernières années, elle n'a cessé de se développer dans ce secteur. Cette absence pourrait s'expliquer par divers facteurs, comme des différences culturelles dans la consommation de vin ou encore des difficultés pour la Chine à s'exporter et encore plus à conquérir le marché américain.

Note médiane des vins par pays



La carte ci-dessus affiche la médiane des scores attribués aux vins de différents pays, révélant les tendances qualitatives des principaux acteurs sur le marché mondial du vin. Nous avons opté pour la médiane comme mesure de tendance centrale pour évaluer les scores des vins, car elle est moins

sensible aux valeurs extrêmes, particulièrement pertinent dans notre contexte où le nombre de critiques varie significativement entre les pays. De plus, nous avons exclu les pays comptant moins de 100 critiques, jugés insuffisants pour fournir une estimation fiable, afin de préserver l'intégrité et la précision de nos résultats.

Comme attendu, les nations viticoles européennes traditionnelles ainsi que les États-Unis montrent une forte présence, ce qui est cohérent avec leur réputation et leur histoire dans la production de vins de qualité.

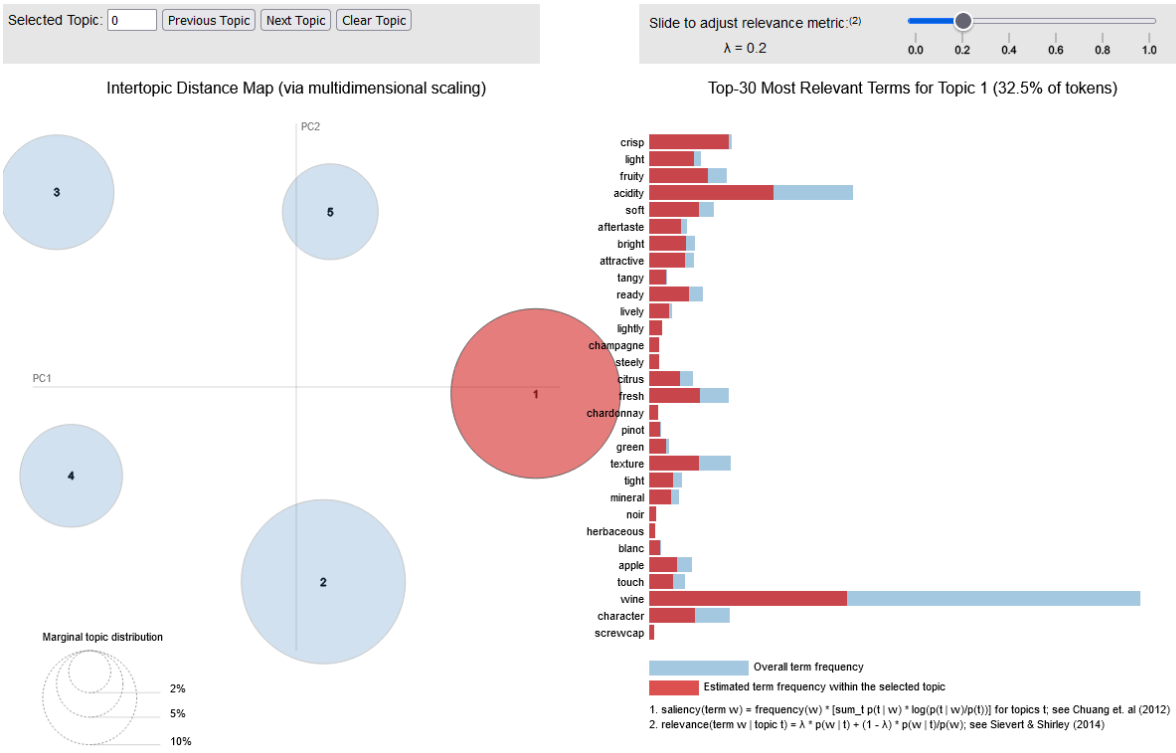
Une observation notable est la haute notation des vins provenant de l'Australie et du Canada, des régions qui continuent à gagner en estime sur la scène internationale. L'Australie, en particulier, est reconnue pour ses vins exceptionnels et a fréquemment été récompensée lors de compétitions internationales (seconde place aux Decanter World Wine Awards de 2024 derrière la France).

6) Topic Modelling

Pour approfondir notre analyse, nous nous sommes penchés sur les descriptions des vins présents dans notre dataset, nous avons mis en œuvre une analyse de topic modeling en utilisant l'algorithme Latent Dirichlet Allocation (LDA) avec la librairie pyLDAvis. Cette technique nous a permis de découvrir des thèmes ou des sujets récurrents à partir des descriptions textuelles des vins.

Nous avons commencé par extraire uniquement les descriptions des vins de notre dataset. Afin de simplifier le traitement du texte, nous avons éliminé toute ponctuation, ce qui aide à réduire les erreurs de parsing et à normaliser les entrées pour l'analyse.

Nous avons choisi de configurer notre modèle LDA pour identifier cinq topics différents. Bien que, généralement, il soit recommandé d'explorer un nombre plus élevé de topics pour capturer une plus grande variété de thèmes, dans ce cas spécifique, limiter le nombre à cinq s'est avéré être une approche plus pratique et suffisante pour notre analyse.



Interprétations des résultats :

Topic 1: Caractéristiques générales et sensorielles des vins blancs

- Termes clés : *crisp, light, fruity, acidity, soft, aftertaste, bright, tangy, lively, lightly, champagne*
- Ce topic se concentre sur les qualités sensorielles souvent associées aux vins blancs, tels que la fraîcheur, la légèreté et l'acidité.

Topic 2: Propriétés structurales et descripteurs des vins rouges

- Termes clés : *tannins, firm, black, dark, structure, dense, powerful, wood, structured, smoky, tannic*
- Ce topic se penche sur les attributs structuraux des vins rouges, en particulier la présence de tannins et la complexité perçue, souvent associée à des arômes boisés ou fumés.

Topic 3: Terroir et provenance

- Termes clés : *vines, vineyard, estate, cru, family, old, produced, grand, premier, owned, parcel*
- Ici, l'accent est mis sur le terroir, les appellations et la provenance des vins, ce qui est crucial pour l'identification et la classification des vins de qualité.

Topic 4: Profil aromatique et expérience gustative

- Termes clés : *grenache, syrah, medium, bit, lead, finish, medium-bodied, mouthfeel, aromas, silky, scents*
- Ce topic explore le profil aromatique et les sensations en bouche, avec une attention particulière aux arômes et à la texture perçue des vins.

Topic 5: Descripteurs spécifiques et expériences sensorielles uniques

- Termes clés : *palate, lovely, lemony, nose, slender, wonderfully, appetizing, purity, residual, mirabelle*
- Le dernier topic se concentre sur des descriptions très spécifiques et des qualités uniques qui peuvent être utilisées pour décrire des expériences de dégustation particulières.

7) Répartition du travail

Travail réalisé	Membre(s)
clustering.ipynb (nettoyage des données + PCA and clustering + Outlier detection + correlations)	Corentin Bohelay
recommendations.ipynb	Yanis Aumasson
countries-map.ipynb	Yanis Aumasson
topic-modelling.ipynb	Yanis Aumasson
Rédaction du rapport	Corentin Bohelay et Yanis Aumasson