Étude de marché - Vin

Domaine des Croix

Antoine SAVOURNIN

Contexte & Problématique

Vous souhaitez lancer vos vins sur le marché américain, un marché compétitif avec une forte demande pour des vins de qualité. Pour être compétitif, il est crucial de fixer un prix adapté à vos produits. Vous avez collecté un jeu de données comprenant des informations sur les cépages, les régions de production, les millésimes, les notes des oenologues, et le prix moyen des vins similaires sur le marché américain.

Notre objectif est d'analyser ces données pour comprendre quels facteurs influencent le prix et vous recommander un prix compétitif pour vos vins, tout en vous expliquant clairement la démarche, sans complexité technique.

Présentation des données

- country : Pays d'origine du vin.
- description : Courte description textuelle.
- designation : Le nom spécifique du vin ou de la cuvée. [drop]
- points : La note attribuée au vin sur 100.
- prix : Le prix de vente en dollars.
- province : La province ou l'état où le vin a été produit
- region_1: La 1ère région viticole dans laquelle le vin a été produit.
- region_2 : La 2e région viticole dans laquelle le vin a été produit. [drop]
- taster_name : Le nom de l'oenologue ayant dégusté le vin et attribué une note.
- taster_twitter_handle : Son identifiant twitter. [drop]
- title : Le titre du vin, souvent une combinaison de variety, winery et region_1.
- variety : Le cépage utilisé pour produire le vin.
- winery : Le nom de la cave ou du domaine viticole où le vin a été produit.

Les colonnes jugées non pertinente pour cette étude sont directement supprimées
Une colonne 'millesime' correspondant à l'année de production du vin a été créée

Analyse exploratoire des données

country	129894	non-null	object
description	129957	non-null	object
points	129957	non-null	int64
price	120964	non-null	float64
province	129894	non-null	object
region_1	108710	non-null	object
taster_name	103713	non-null	object
title	129957	non-null	object
variety	129956	non-null	object
winery	129957	non-null	object
millesime	125325	non-null	float64

On voit qu'il y a de nombreuses valeurs manquantes dans le dataset que l'on va devoir gérer, colonne par colonne.

Les valeurs manquantes dans les colonnes 'country' et 'province' ne représentent que 0.05% chacune du total des données, donc on va les supprimer pour une étude plus propre mais toujours pertinente. Même chose pour la colonne 'millesime' qui n'a que 3.56% de valeurs manquantes

Pour la colonne 'price', comme il y a moins de 7% de valeurs manquantes, je vais préférer les supprimer pour quand même garder une étude pertinente. En fonction de ce que vous souhaitez on pourra les garder et les remplacer par la médiane(pas la moyenne car trop de valeurs aberrantes).

Pour les 'region' et 'taster_name' manquants, comme cela n'aura pas une énorme influence sur l'étude je vais les remplacer par une valeur 'inconnu'

Méthodologie, outils et langages utilisés

La méthodologie adoptée pour cette étude repose sur une approche data-driven, visant à analyser les facteurs influençant le prix des vins et à traiter les valeurs manquantes.

- Extraction et transformation des données : Nous avons commencé par explorer et nettoyer le dataset fourni à l'aide de Google Colab, en appliquant des techniques de prétraitement et d'enrichissement des données
- Analyse et visualisation : Les données ont ensuite été intégrées dans Power BI pour concevoir un dashboard interactif, permettant une exploration intuitive et une prise de décision éclairée.
- Livrables et présentation : La présentation des résultats a été réalisée via Google Slides.

Outils utilisés :

- Google Colab : Extraction, transformation et export des données.
- Power BI : Visualisation et analyse interactive des données.
- Google Slides : Création du support de présentation.
- Freepik : Génération d'illustrations pour embellir les livrables.

Langages utilisés :

- Python (traitement et analyse des données).
- DAX (modélisation et mesures dans Power BI).

Présentation de la partie technique

```
df = df.drop(['designation', 'region_2', 'taster_twitter_handle'], axis = 1)
```

```
df['millesime'] = df['title'].str.extract(r'\b(19[4-9][0-9]|20[0-2][0-9])\b')
df['millesime'] = df['millesime'].astype('float')
```

```
df = df[(df['country'].isna() == False) & (df['province'].isna() == False) & (df['variety'].isna() == False) & (df['millesime'].isna() == False)]
```

```
df = df[df['price'].isna() == False]
```

```
df['region_1'] = df['region_1'].fillna('unknown')
df['taster_name'] = df['taster_name'].fillna('unknown')
```

```
df.to_csv("donnees_vin.csv", index = False)
```

```
def categorize_price(price):
  if price < 10:
    return "moins de 10$"
  elif 10 <= price < 20:
    return "entre 10$ et 20$"
  elif 20 <= price < 30:
    return "entre 20$ et 30$"
  elif 30 <= price < 50:
    return "entre 30$ et 50$"
  elif 50 <= price < 100:
    return "entre 50$ et 100$"
  else:
    return "Plus de 100$"


df['price_range'] = df['price'].apply(categorize_price)
```

La majorité de l'exploration et du nettoyage a été effectué en Python sur un notebook Google colab.

- La méthode .drop() pour supprimer les colonnes.
- le .str.extract(r'\b(19[4-9][0-9]|20[0-2][0-9])\b') permet d'extraire les dates de millésime grâce au regex
- Les valeurs manquantes sont supprimées en filtrant le df avec isna() == False dans les colonnes définies
- la méthode .fillna('unknown') permet de changer les valeurs manquantes en 'unknown'
- la méthode .to_csv permet d'exporter le df propre en csv pour ensuite le charger dans power BI
- Création d'une fonction pour catégoriser les prix en ajoutant une colonne 'price_range'

Présentation de la partie technique

Création d'un WordCloud pour enrichir le dashboard:

- Utilisation de la colonne 'description'
- Téléchargement et préparation des ressources NLTK nécessaires, incluant les outils de tokenisation, les listes de stopwords et les corpus populaires pour l'analyse textuelle
- Tokenisation, suppression ponctuation, stopwords et lemmatisation

```
import nltk
from nltk.corpus import stopwords

nltk.download('popular', quiet=True)
nltk.download('punkt_tab')
nltk.download('stopwords')
```

```
def supprimer_ponctuation(texte):
  for caractere in texte:
    if caractere.isalpha() == False and caractere != ' ':
      texte = texte.replace(caractere, '')
  return texte.lower()
```

```
import spacy
nlp = spacy.load("en_core_web_sm")

def stopword_lemma(x):
  rendu = []
  doc = nlp(" ".join(x))
  for word in doc:
    if word.text not in stopwords.words("english"):
      rendu.append(word.lemma_)
  return rendu
```

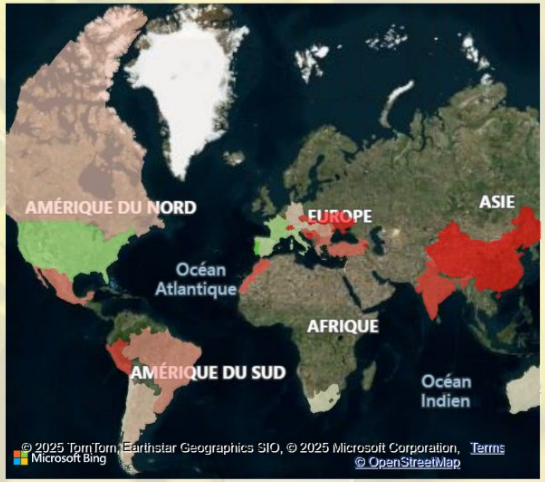
```
from wordcloud import WordCloud
import numpy as np
import matplotlib.pyplot as plt
from PIL import Image


text = " ".join(df['description'].apply(lambda x: " ".join(x)))
all_text = re.sub(r'\bwine\b', '', text)

mask_image2 = np.array(Image.open("/content/wordcloud.png"))


colors = ["#7B002C", "#73B761"]


wordcloud = WordCloud(
    width=1000,
    height=400,
    background_color="#FEFBE0",
    mask=mask_image2,
    max_words=50,
    contour_color="#7B002C",
    contour_width=3,
    max_font_size=100,
    color_func=lambda *args, **kwargs: np.random.choice(colors)
).generate(all_text)

plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

- Les listes de mots nettoyés sont toutes concaténées en un gros texte auquel on retire le mot "wine"
- Affichage du wordcloud en utilisant une image de bouteille qui servira de masque, et en utilisant les codes des couleurs du dashboard.
- Ceci sera fait 3 fois pour chaque page du dashboard : une fois pour les données globales, une pour la France et une pour les pinots noirs.


International
France
Pinot Noir
Domaine de la Croix
88,74
Note moyenne
38,12
Prix moyen ($)
1902
Prix maximum ($)
3
Prix minimum ($)
Millésime
1947
2022
Région
Tout
Cépage
Tout
Appellation
Tout
Œnologue
Tout
entre 10$ et 20$
entre 30$ et 50$
moins de 10$
entre 20$ et 30$
entre 50$ et 100$
Plus de 100$
La structure de chaque page restera identique, avec les filtres sur la gauche et les indicateurs fixes en haut à droite. Une barre de boutons en haut permettra de naviguer facilement entre les différentes pages.

International | France | Pinot Noir | Domaine de la Croix

Pays les plus millésimés



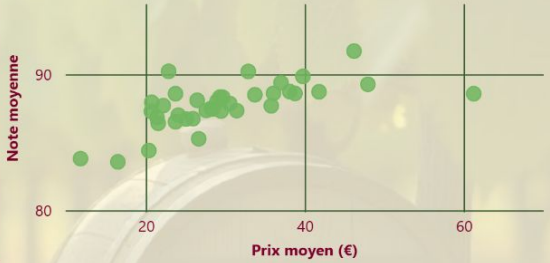
88,46
Note moyenne

35,65
Prix moyen ($)

1902
Prix maximum ($)

2
Prix minimum ($)

Prix & Notes Moyens par Pays



Millésime

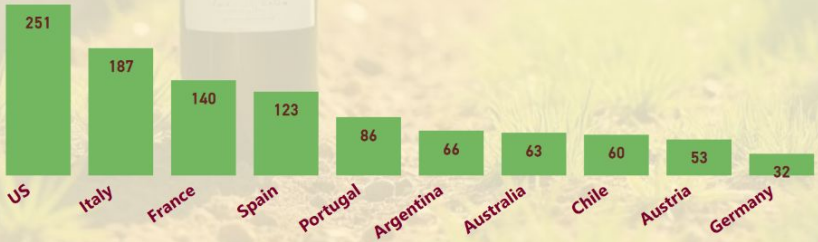
1943 — 2023

Région

Tout

entre 10$ et 20$	entre 50$ et 100$
entre 20$ et 30$	moins de 10$
entre 30$ et 50$	Plus de 100$

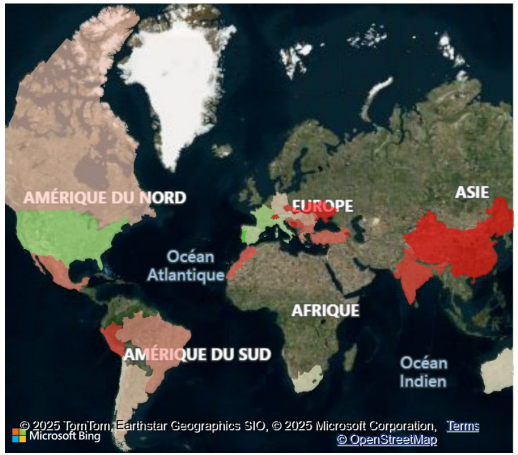
Top 10 nombre de cépages par pays



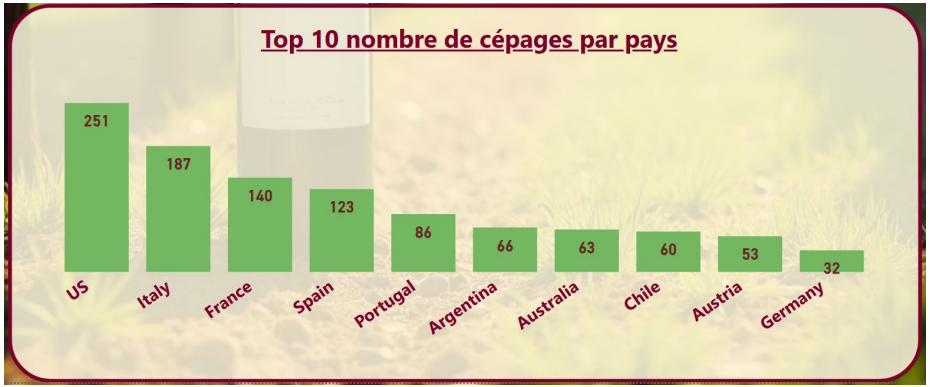
Prix & Notes Moyens par Pays

Pays	Prix moyen ($)	Note Moyenne /100
England	46,22	91,71
India	22,89	90,22
Austria	32,91	90,21
Germany	39,78	89,84
Canada	37,03	89,38
Hungary	47,96	89,27
France	38,12	88,74
Italy	41,82	88,72
US	36,08	88,60
Australia	38,79	88,60
Total	35,65	88,46

Analyse du marché international



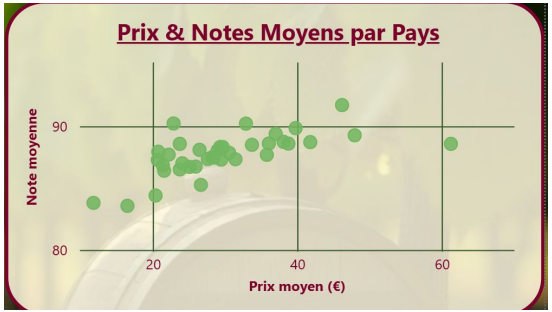
Le dataset révèle une diversité de vins provenant du monde entier. Toutefois, les zones affichant la plus grande variété de millésimes, et donc une production plus régulière, se situent principalement aux États-Unis, en Europe de l'Ouest, en Australie et dans le sud de l'Amérique du Sud. Cette tendance est confirmée par le graphique suivant, qui présente le top 10 des pays avec le plus grand nombre de cépages différents.



Analyse du marché international

Le wordcloud met en évidence une forte présence de termes liés aux fruits ainsi qu'au champ lexical de l'œnologie (comme structure, palais, riche...), de manière générale.

La matrice à droite révèle que les pays avec les prix moyens les plus élevés ne sont pas nécessairement ceux ayant la plus grande diversité de cépages. On note également une légère corrélation positive entre les prix et les notes, comme l'indique le nuage de points, suggérant une droite avec un coefficient de corrélation positif. Enfin, 8 des 10 pays ayant les prix les plus élevés figurent également parmi les 10 premiers en termes de notes.



Prix & Notes Moyens par Pays

Pays	Prix moyen ($)	Note Moyenne /100
England	46,22	91,71
India	22,89	90,22
Austria	32,91	90,21
Germany	39,78	89,84
Canada	37,03	89,38
Hungary	47,96	89,27
France	38,12	88,74
Italy	41,82	88,72
US	36,08	88,60
Australia	38,79	88,60

Prix & Notes Moyens par Pays

Pays	Prix moyen ($)	Note Moyenne /100
Switzerland	61,29	88,57
Hungary	47,96	89,27
England	46,22	91,71
Italy	41,82	88,72
Germany	39,78	89,84
Australia	38,79	88,60
France	38,12	88,74
Canada	37,03	89,38
US	36,08	88,60
Lebanon	35,80	87,69

International | France | Pinot Noir | Domaine de la Croix

Moyenne de note par région

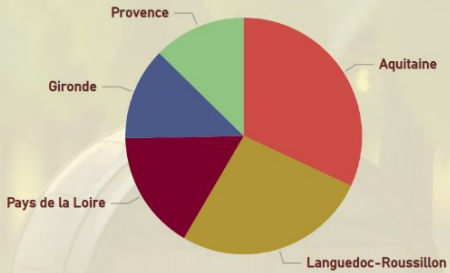
88,74
Note moyenne

38,12
Prix moyen ($)

1902
Prix maximum ($)

3
Prix minimum ($)

Régions ayant le plus de cépages



Millésime

1947 — 2022

Région

Tout

Cépage

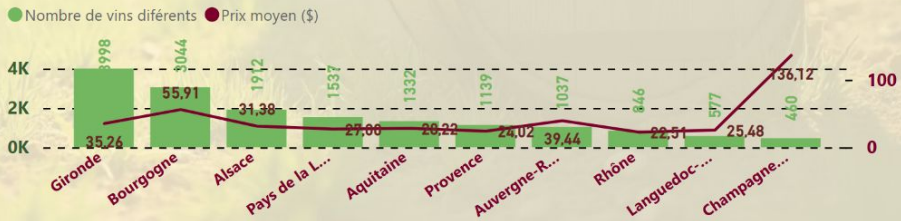
Tout

Appellation

Tout

entre 10$ et 20$	entre 30$ et 50$	moins de 10$
entre 20$ et 30$	entre 50$ et 100$	Plus de 100$

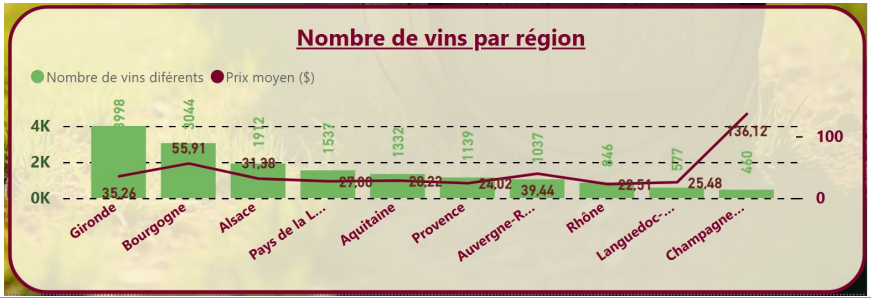
Nombre de vins par région



Analyse du marché français

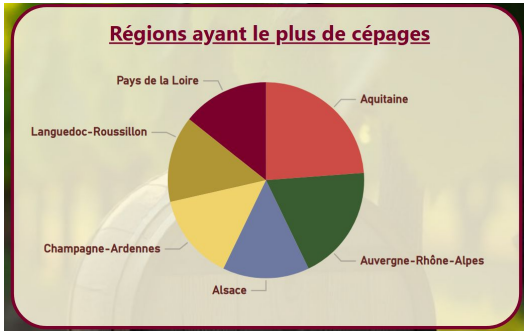
Il ressort que les vins français les mieux notés proviennent principalement de la Gironde et de la Bourgogne. Ces régions se distinguent également par une grande diversité de vins produits, avec la Bourgogne occupant la deuxième position en termes de prix moyen, juste après la Champagne. Cette dernière, en raison de la prestige du champagne, influe significativement sur la moyenne des prix. Les vins de Bordeaux, bien qu'ayant une moyenne de prix relativement élevée, restent plus abordables et dominent la production en termes de volume.

On peut donc voir que la Bourgogne et la Gironde sont des pôles essentiels tant en qualité qu'en diversité, alors que la Champagne tire sa moyenne vers le haut grâce à son produit emblématique.

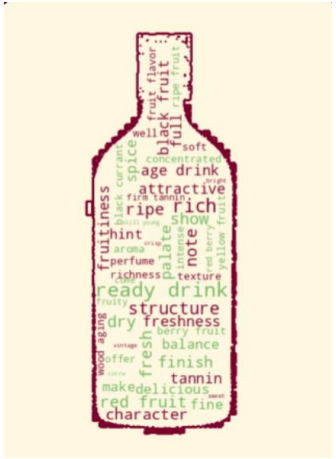


Analyse du marché français

En ce qui concerne les vins de luxe, ceux dépassant les 100$ la bouteille, la Bourgogne se distingue en tête avec les meilleures notes. Cependant, la région Aquitaine, bien que présentant des notes plus modestes, regroupe davantage de cépages dans cette gamme de prix. Globalement, l'analyse montre que la majorité des vins français présents dans le dataset proviennent de la moitié sud de la France.



Le WordCloud français met désormais davantage en avant le vocabulaire associé au vin de luxe, comme les termes 'structure', 'personnalité' ou 'tanins', en contraste avec une prédominance antérieure des mots liés aux fruits.



International | France | Pinot Noir | Domaine de la Croix

Millésime

2002 — 2022

Région

Tout

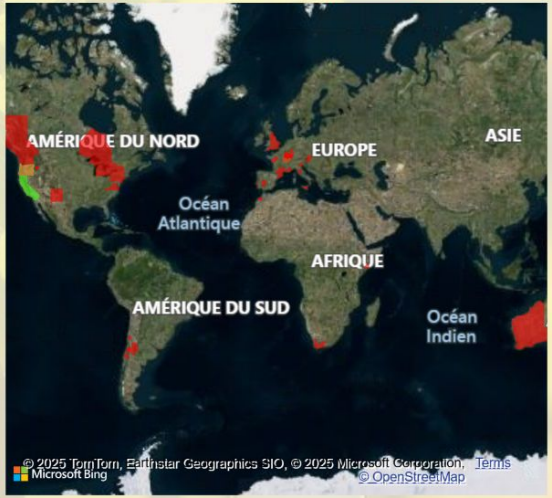
Appellation

Tout

Œnologue

Tout

Moyenne de note par région



89,42
Note moyenne

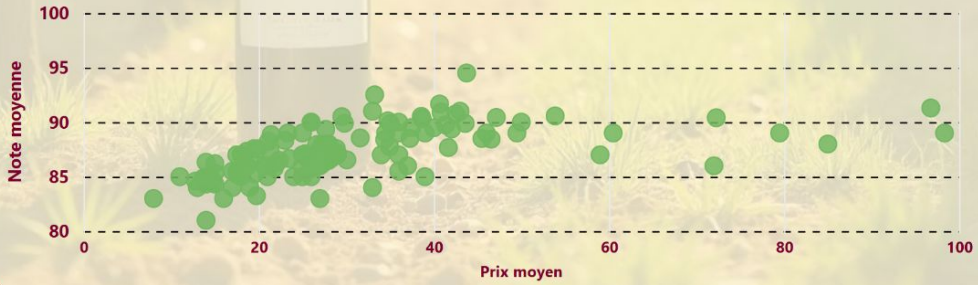
42,18
Prix moyen ($)

1902
Prix maximum ($)

4
Prix minimum ($)

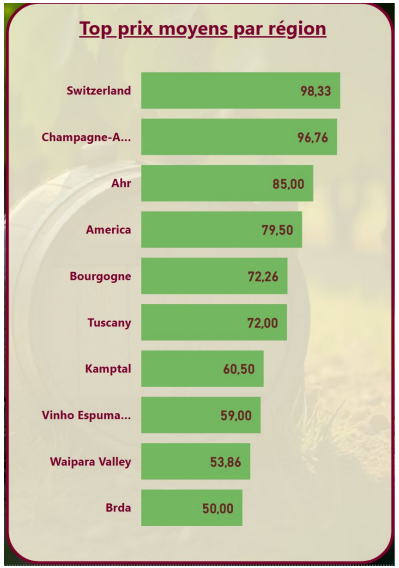
Top prix moyens par région

Relation Prix vs Notes



Analyse du marché du pinot noir

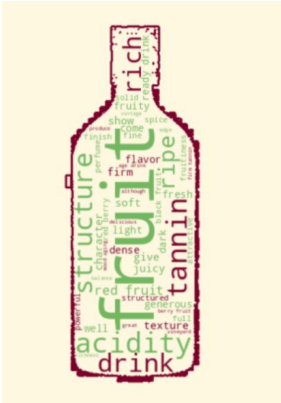
On constate que le Pinot Noir est cultivé à l'échelle mondiale, mais que ses meilleures notes proviennent de l'Amérique de l'Ouest, en particulier de Californie. Cela confirme la montée en puissance des vins californiens ces dernières années.



En ce qui concerne le top 10 des régions mondiales ayant les prix les plus élevés, la Suisse se distingue en restant en tête, comme dans l'étude globale mondiale. La Champagne se classe deuxième grâce à ses prix exceptionnellement élevés. La Bourgogne, quant à elle, occupe la cinquième place au niveau mondial en termes de prix et se distingue par des notes qui ne sont ni particulièrement élevées ni particulièrement faibles.

Le nuage de points montre que la légère corrélation entre le prix et la note reste présente à cette échelle.

Le WordCloud du Pinot Noir met en évidence cinq mots principaux : fruit, tannin, structure, richesse et acidité. Il serait pertinent de comparer ces résultats avec d'autres analyses d'oenologues pour vérifier si ces termes sont également fréquemment utilisés.



International | France | Pinot Noir | Domaine de la Croix

Millésime

2014 — 2020

Région

Tout

Cépage

Tout

Appellation

Tout

Moyenne de note par région

85
Prix médian

111,54
Prix moyen ($)

248
Prix maximum ($)

48
Prix minimum ($)

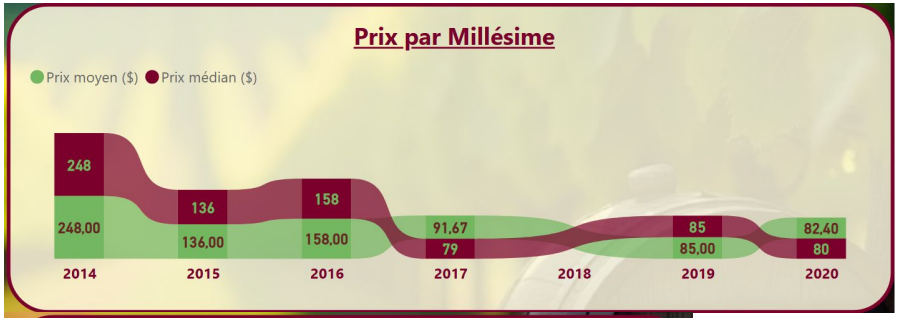
Prix par Millésime

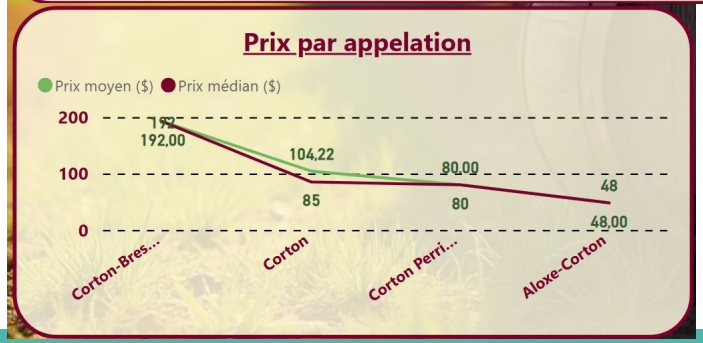
Prix par appelation

13
Bouteilles comparables

Analyse du marché Domaine de la Croix

Cette dernière page du dashboard présente les filtres correspondant aux vins ayant les mêmes caractéristiques que votre vin cible : un Pinot Noir de Bourgogne, noté 94/100 par Roger Voss et portant l'appellation Corton. Cela permet de comparer 13 bouteilles afin de déterminer le prix de la vôtre.





Sur la période 2014-2020, les quatre appellations Corton retenues présentent de grandes disparités de prix dans l'étude. Les prix varient de 48 à 258€ la bouteille, avec un prix moyen de 111,54€ et une médiane de 85€, qui sont relativement éloignés. On observe également une tendance à la baisse des prix au fil du temps, avec des bouteilles de 2016, comme la vôtre, se vendant autour de 158€. Cependant, la plupart des appellations similaires se situent plutôt autour de 100 € la bouteille.

Conclusion

À partir des données collectées sur les appellations Corton pour la période 2014-2020, il est possible de définir différentes fourchettes de prix selon la gamme que vous souhaitez viser :

Entrée de gamme : Si vous recherchez un vin accessible, vous pouvez vous positionner autour de la médiane des prix, soit 85€. Cela vous permettra d'être très compétitif sur le marché.

Milieu de gamme : Pour un vin de gamme supérieure mais à prix raisonnable, la majorité des vins des appellations Corton se situent autour de 100€. Vous pouvez également vous rapprocher de la moyenne des prix en vous plaçant entre 100€ et 110€.

Haut de gamme : Si vous souhaitez viser le haut de gamme, vous pouvez vous aligner avec les vins Corton de 2016, dont le prix se situe entre 155€ et 160€. Cette fourchette est pertinente compte tenu de la note obtenue et du millésime 2016, qui reste parmi les plus valorisés.

En fonction de la gamme que vous souhaitez offrir, vous avez une large gamme de prix possibles. Que ce soit pour un vin plus accessible ou un cru haut de gamme, vous trouverez des options qui conviennent à vos objectifs tout en étant compétitif sur le marché.


MERCI POUR VOTRE ATTENTION !