Progetto Data Manipulation and Visualization

Studente: <u>Alessandro Busà</u>

Indice

- Introduzione
- Dataset Utilizzato
- Librerie Utilizzate
- Funzioni
- Statistiche di Base
- Prezzo Medio dei Vini per Paese, Commento + Spiegazione
- Prezzo Medio dei Vini per le Principali Varietà, Commento + Spiegazione
- Le 10 Varietà di Vino più Rappresentate, Commento + Spiegazione
- Vini con prezzi superiori a 500\$ o punteggi sopra i 98
- Distribuzione dei Vini per Paese, Commento + Spiegazione
- Heatmap di correlazione tra prezzo e punteggio, Commento + Spiegazione
- Distribuzione del prezzo del vino, Commento + Spiegazione
- Distribuzione dei punteggi dei vini, Commento + Spiegazione
- Distribuzione dei Punteggi dei Vini per Paese con Overlay di Punti, Commento + Spiegazione
- Conclusioni e Raccomandazioni
- GitHub

Introduzione

Nel presente progetto, ho analizzato un dataset contenente circa 130.000 recensioni di vini. L'obiettivo è sviluppare una strategia per un marketplace che colleghi piccoli produttori locali con acquirenti globali, mettendo in risalto l'unicità e la qualità dei loro prodotti. L'analisi dei dati è stata suddivisa in varie fasi: pulizia e preparazione, esplorazione delle caratteristiche dei vini e delle cantine, e analisi delle relazioni tra punteggi e prezzi.

Dataset Utilizzato

Il dataset (file .csv) contiene circa 130.000 righe e 14 colonne. Ecco una breve descrizione delle colonne:

- u*nnamed: 0*: Indice numerico.
- country: Paese di origine del vino (alcuni valori mancanti).
- *description*: Descrizione del vino.
- <u>designation</u>: Designazione del vino (etichetta specifica, con molti valori mancanti).
- *points*: Punteggio assegnato al vino.
- *price*: Prezzo del vino (in alcune righe mancante).
- *province*: Provincia o regione del vino nel paese di origine.
- <u>region 1</u>: Regione geografica specifica (molti valori mancanti).
- <u>region 2</u>: Sotto-regione (molti valori mancanti).
- taster name: Nome del critico/degustatore (molti valori mancanti).
- taster twitter handle: Handle Twitter del degustatore (molti valori mancanti).
- <u>title</u>: Nome completo del vino.
- *variety*: Varietà di uva utilizzata.
- *winery*: Nome della cantina produttrice.

Il dataset contiene informazioni utili per l'analisi di vini, inclusi punteggi, prezzi, descrizioni e informazioni geografiche ed alcune colonne hanno molti valori mancanti (ad esempio, designazione e regioni).

Librerie Utilizzate

import pandas as pd

Importa la libreria pandas, utile per la manipolazione dei dati.

Libreria per la creazione di grafici

import matplotlib.pyplot as plt

import seaborn as sns

Libreria per visualizzazioni avanzate basate su Matplotlib

Funzioni

Funzione per aggiungere etichette in grassetto sui barplot (valori con una cifra decimale)

Funzione per aggiungere etichette in grassetto sui barplot (valori interi)

Funzione per aggiungere etichette in grassetto sui barplot orizzontali

Statistiche di Base

```
# Statistiche descrittive per i prezzi dei vini
price_mean = round(df_clean['price'].mean(), 2)  # Media dei prezzi
price_median = round(df_clean['price'].median(), 2)  # Wediana dei prezzi
price_var = round(df_clean['price'].var(), 2)  # Varianza dei prezzi

# Statistiche descrittive per i punteggi dei vini
points_mean = round(df_clean['points'].mean(), 2)  # Media dei punteggi
points_median = round(df_clean['points'].median(), 2)  # Mediana dei punteggi
points_var = round(df_clean['points'].var(), 2)  # Varianza dei punteggi
```

Spiegazione:

<u>.mean</u>: Consente di calcolare la media (Prezzi e Punteggi) <u>.median</u>: Consente di calcolare la mediana (Prezzi e Punteggi) <u>.var</u>: Consente di calcolare la varianza (Prezzi e Punteggi)

Commento:

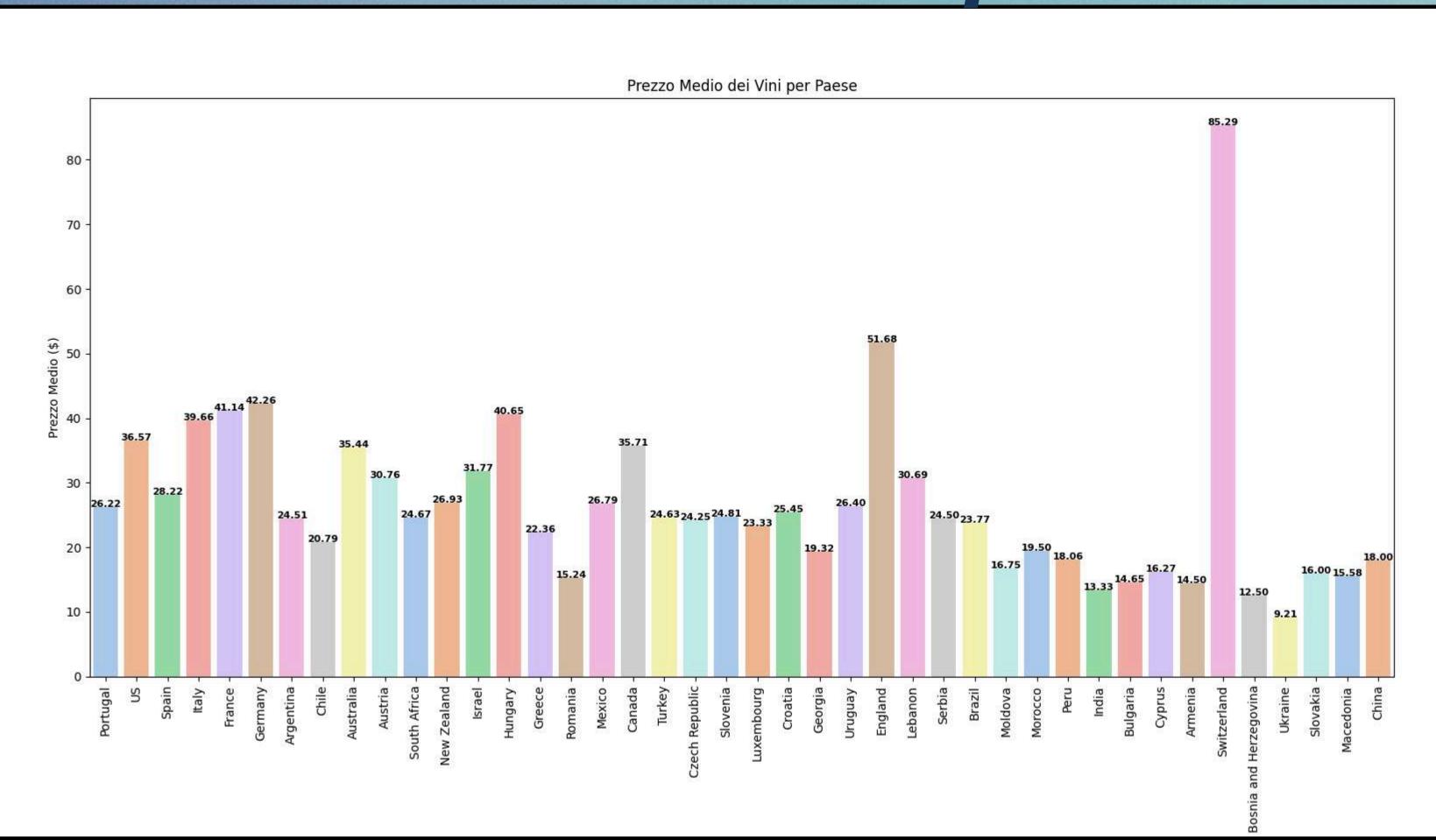
Prezzo medio dei vini: 35.37\$, con una mediana di 25.00\$.

La varianza ci mostra una certa variabilità nei prezzi, suggerendo la presenza di vini sia molto economici che di alta fascia.

Il punteggio medio assegnato ai vini è 88.42, con una mediana di 88.00.

Anche qui, la varianza riflette che ci sono vini con punteggi eccezionali, ma anche alcuni con valutazioni più basse.

Prezzo Medio dei Vini per Paese



Spiegazione:

"plt.figure": Imposta le dimensioni della figura

"sns.barplot": Crea il barplot con prezzo medio per paese

"plt.title":Titolo del grafico

"plt.xticks":Ruota le etichette dell'asse X per una migliore leggibilità

"plt.ylabel": Etichetta asse Y

"grassetto": Aggiunge etichette in grassetto con i valori sui barplot

"plt.show": Mostra il grafico

Commento:

1. Paesi con i prezzi più elevati:

Il paese che si distingue maggiormente è la Svizzera, con un prezzo medio per bottiglia di vino che arriva a 85.29\$. Questo valore è sorprendentemente alto rispetto a tutti gli altri paesi, suggerendo che in Svizzera il mercato del vino è dominato da bottiglie di alta fascia, probabilmente per via della qualità superiore, della produzione limitata o dei costi di importazione. Altri paesi con prezzi medi elevati includono Inghilterra (51.68\$), Germania (42.26\$) e Francia (41.14\$), che riflettono la produzione di vini di alta qualità e la loro fama internazionale.

2. Paesi con prezzi intermedi:

Tra questi, troviamo paesi come l'Australia, l'Italia e gli Stati Uniti, con prezzi medi tra i 35\$ e i 39\$.

Questi paesi offrono una grande varietà di vini, dai più economici a quelli di lusso.

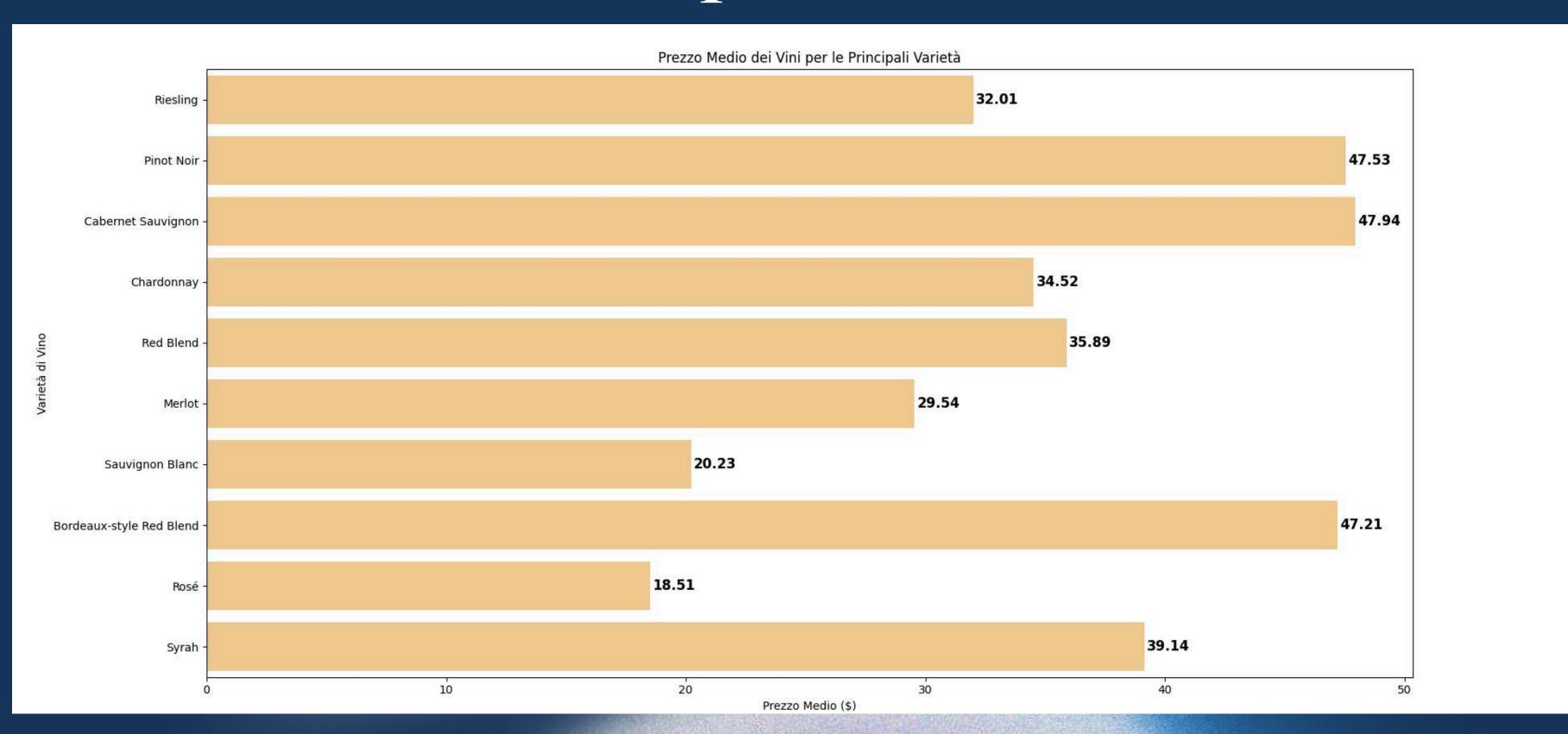
3. Paesi con prezzi più bassi:

Dall'altra parte dello spettro, vediamo paesi come l'Ucraina, con un prezzo medio di 9.21\$, seguita dalla Romania e la Slovacchia, con prezzi attorno ai 15-16\$. Questi paesi producono vini accessibili, probabilmente destinati principalmente al consumo interno. Paesi emergenti come la Cina e l'India presentano anch'essi prezzi relativamente bassi, riflettendo mercati vinicoli in crescita.

4. Paesi emergenti:

Infine, mercati come il Libano (30.69\$) e la Israele (31.77\$) stanno iniziando a farsi notare sul mercato vinicolo internazionale. I loro prezzi medi indicano un crescente interesse per i loro vini unici, con tradizioni vinicole che risalgono a secoli fa.

Prezzo Medio dei Vini per le Principali Varietà



<u>Commento + Spiegazione</u>

Spiegazione:

"<u>plt.figure</u>": Imposta le dimensioni della figura

"<u>sns.barplot</u>": Crea il barplot con prezzo medio per paese

"plt.title":Titolo del grafico

"plt.xlabel": Etichetta dell'asse X

"<u>plt.ylabel</u>": Etichetta dell'asse Y

"grassetto Oriz": Aggiunge etichette in grassetto con i valori sui barplot

"plt.show":Mostra il grafico

Commento:

1. Varietà con i prezzi più elevati:

Le varietà di vino con i prezzi medi più alti includono il Pinot Noir (47.53\$), il Cabernet Sauvignon (47.94\$) e il Bordeaux-style Red Blend (47.21\$).

Questi risultati riflettono la domanda di vini premium associati a queste varietà, spesso legati a processi di produzione complessi o a regioni viticole di pregio.

Queste varietà sono apprezzate a livello globale per la loro qualità e complessità, il che spiega i prezzi più elevati.

2. Varietà con prezzi intermedi:

Tra le varietà con prezzi intermedi troviamo il Red Blend (35.89\$), il Syrah (39.14\$) e lo Chardonnay (34.52\$).

Queste varietà sono molto diffuse, con ampie fasce di prezzo che vanno dai vini da tavola a quelli di alta qualità. Tuttavia, la loro versatilità e popolarità portano a una media di prezzo leggermente inferiore rispetto ai vini più pregiati.

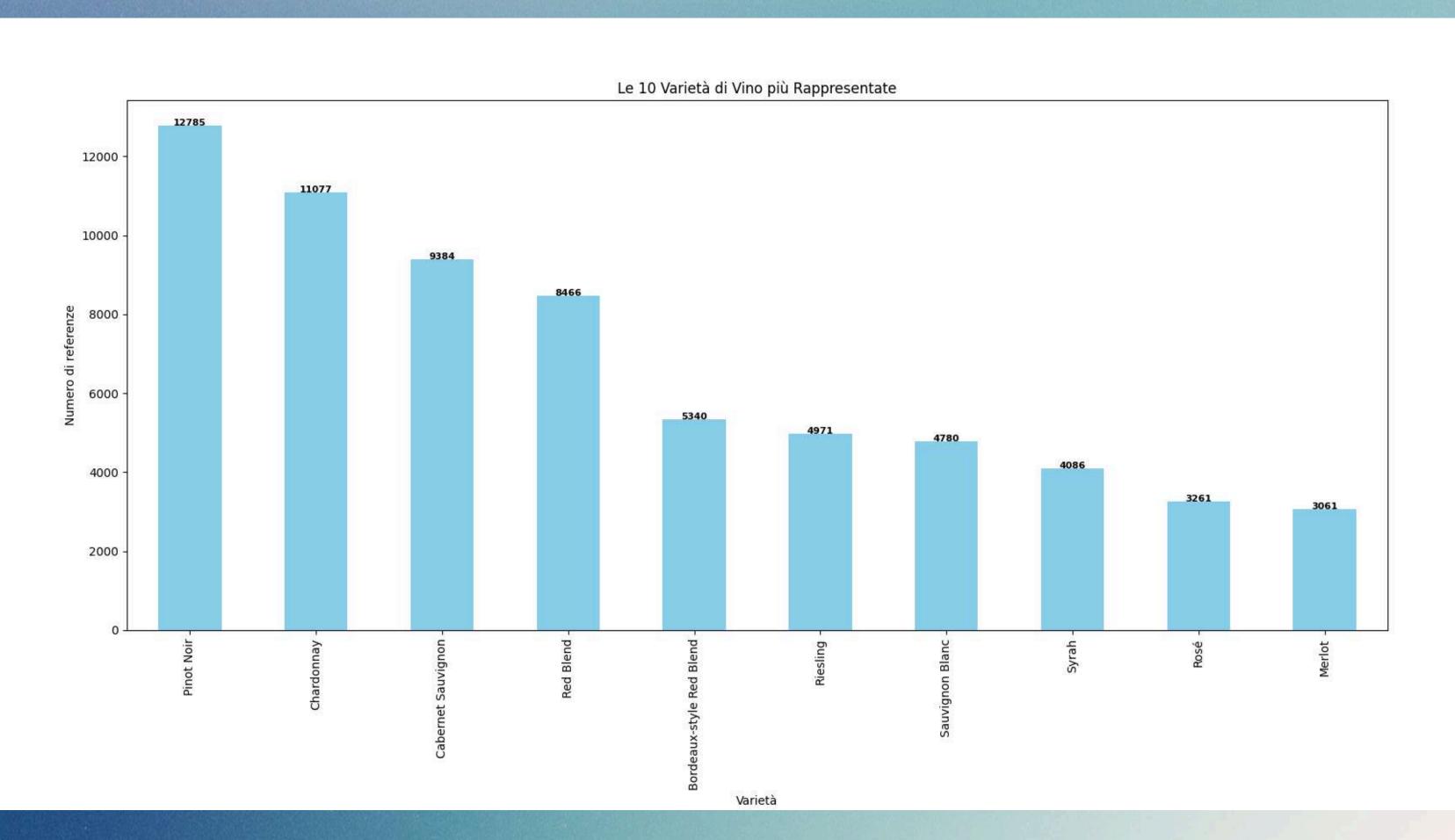
3. Varietà con prezzi più bassi:

Infine, varietà come il Merlot (29.54\$), il Sauvignon Blanc (20.23\$) e il Rosé (18.51\$) presentano prezzi medi più bassi.

Il Rosé, in particolare, ha il prezzo medio più basso, probabilmente a causa della sua produzione più economica e del suo consumo spesso stagionale.

Anche il Merlot, nonostante la sua popolarità, ha un prezzo medio inferiore rispetto a varietà più costose come il Pinot Noir e il Cabernet Sauvignon, riflettendo una maggiore disponibilità e accessibilità di questa varietà.

<u>Le 10 Varietà di Vino più Rappresentate</u>



```
plt.figure(figsize=(10, 6))
ax = top_varieties.plot(kind='bar', color='skyblue')
plt.title('Le 10 Varietà di Vino più Rappresentate')
plt.xlabel('Varietà')
plt.ylabel('Numero di referenze')
plt.xticks(rotation=90)
grassetto2()
plt.show()
```

Spiegazione:

"plt.figure": Imposta le dimensioni della figura

".plot": Crea un grafico a barre per le varietà di vino più rappresentate

"plt.title":Titolo del grafico

"plt.xlabel": Etichetta asse X

"plt.ylabel": Etichetta asse Y

"plt.xticks":Ruota le etichette dell'asse X per una migliore leggibilità

"grassetto2": Aggiunge etichette in grassetto con i valori sui barplot

"plt.show":Mostra il grafico

Commento:

Notiamo che il Pinot Noir, con 12,785 referenze, è la varietà più rappresentata nel mercato, seguito da Chardonnay e Cabernet Sauvignon. Lenza per prodotti di alta rotazione. Queste varietà offrono un'ampia gamma di prezzi e qualità, evidenziando la diversità delle scelte disponibili

Le varietà di vino più rappresentate nel nostro campione sono:

Pinot Noir: 12.785 11.077 **Chardonnay**: 9.384 **Cabernet Sauvignon**: Red Blend: 8.466 **Bordeaux-style Red Blend**: 5.340 Riesling: 4.971 **Sauvignon Blanc**: 4.780 4.086 Syrah: Rosé: 3.261

3.061

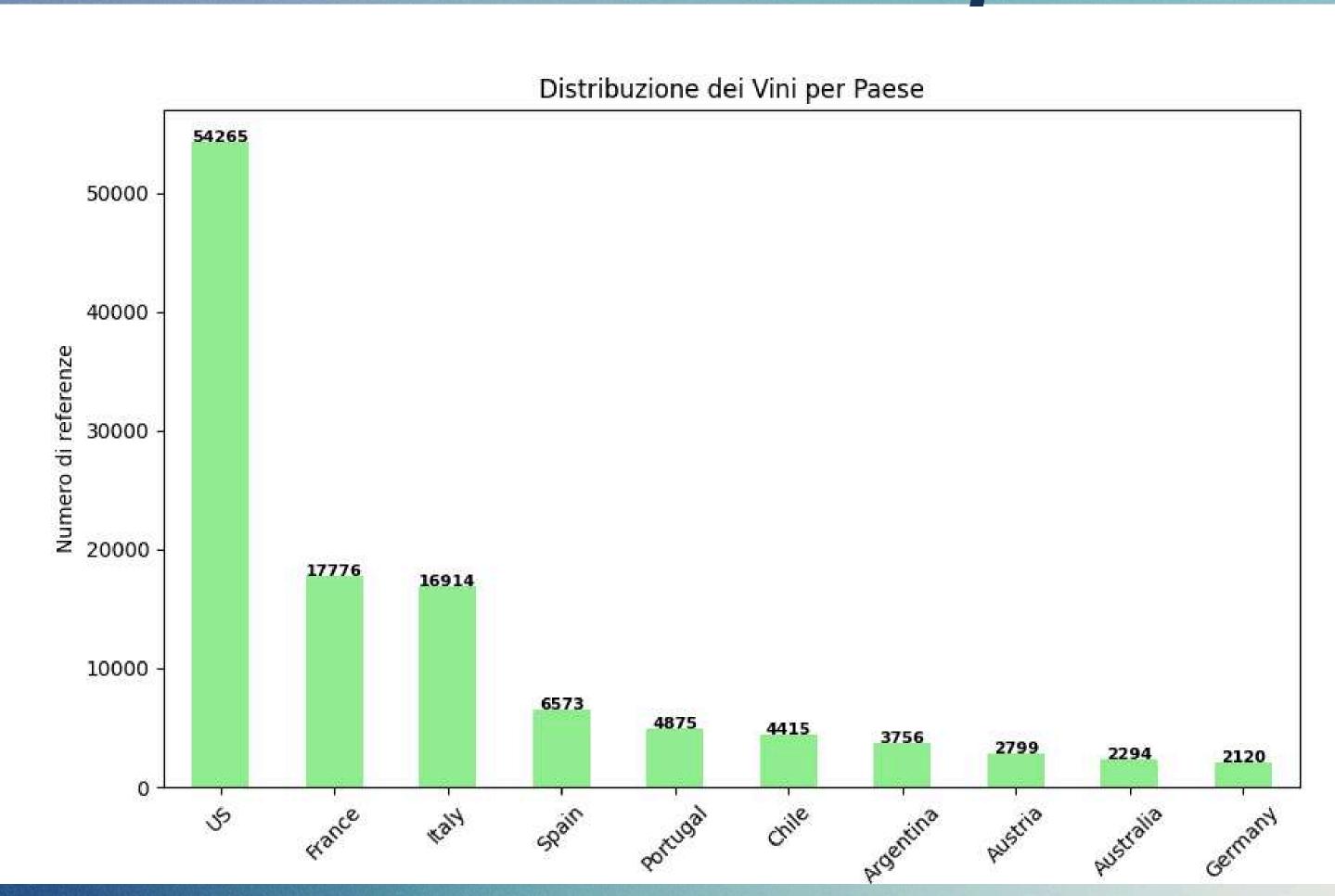
Merlot:

Vini con prezzi superiori a 500\$ o punteggi sopra i 98

```
outlier_vini = df_clean[(df_clean['price'] > 500) | (df_clean['points'] > 98)]
```

Questo comando restituisce tutti i vini fuori dal comune, con prezzi elevati o punteggi eccezionali, spesso rappresentano etichette di nicchia o riserve speciali, cherketplace, rendendolo più esclusivo. (precisamente rilascia tutte le righe con le rispettive colonne che rientrano in questo range)

<u>Distribuzione dei Vini per Paese</u>



```
plt.figure(figsize=(10, 6))
ax= country_distribution.plot(kind='bar', color='lightgreen')
plt.title('Distribuzione dei Vini per Paese')
plt.xlabel('Paese')
plt.ylabel('Numero di referenze')
plt.xticks(rotation=90)
grassetto2()
plt.show()
```

Spiegazione:

"plt.figure": Imposta le dimensioni della figura

"country_distribution.plot": Crea il barplot con prezzo medio per paese

"plt.title":Titolo del grafico

"plt.xlabel":Etichetta asse X

"plt.ylabel": Etichetta asse Y

"grassetto2": Aggiunge etichette in grassetto con i valori sui barplot

"plt.show":Mostra il grafico

Commento:

Il grafico illustra la distribuzione dei vini per paese, con il numero di referenze per ciascuno.

Gli Stati Uniti dominano la scena con 54.265 referenze, distaccandosi nettamente dagli altri paesi.

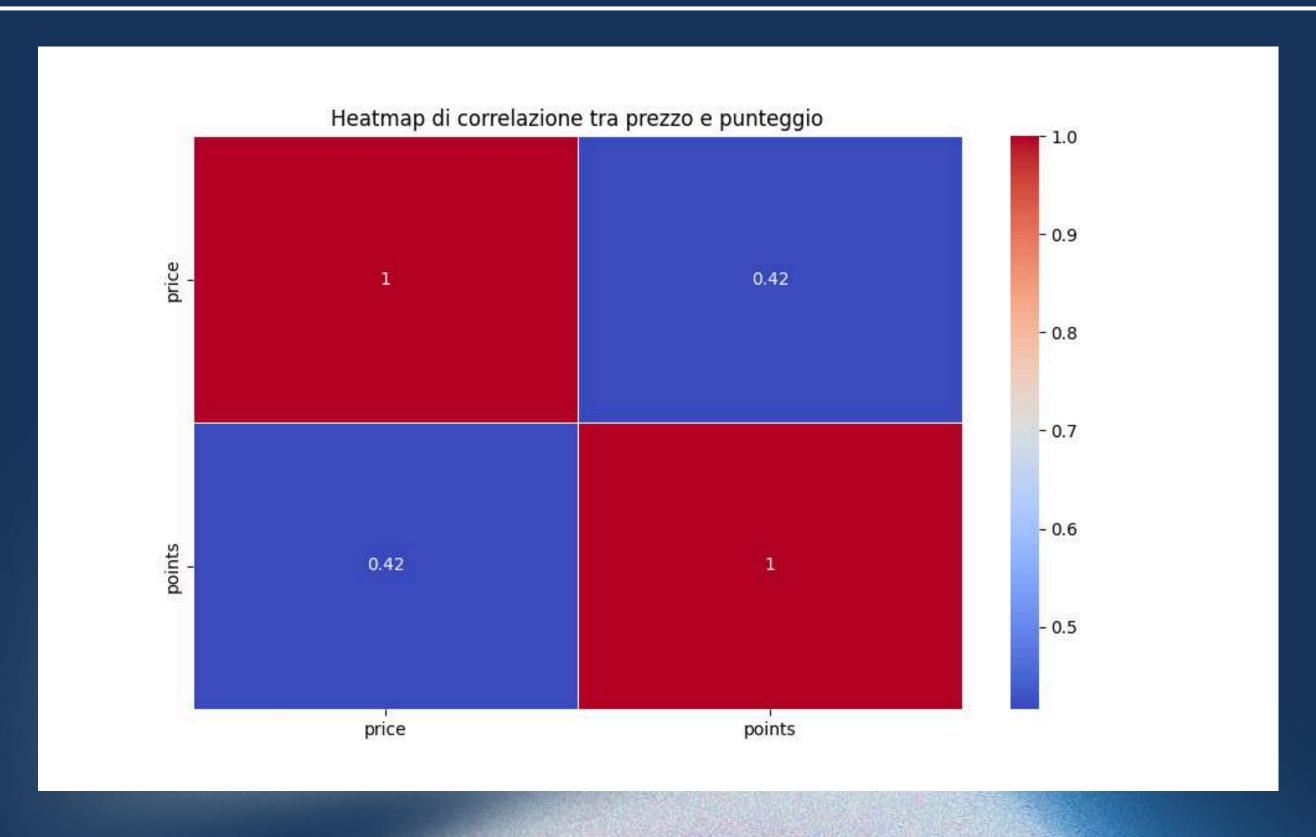
La Francia e l'Italia seguono rispettivamente con 17.776 e 16.914 referenze, sottolineando la loro fama mondiale nel settore vinicolo. La Spagna si piazza al quarto posto con 6.573 referenze, mentre il Portogallo ne conta 4.875.

Altri paesi produttori come il Cile, l'Argentina, l'Austria, l'Australia e la Germania registrano un numero inferiore di referenze, rispe Il grafico evidenzia chiaramente il predominio degli Stati Uniti nel mercato, seguiti dai tradizionali paesi europei noti per la produzione vinicola.

Distribuzione dei vini per paese:

<u>US:</u>	54.265
France:	17.776
<u>Italy:</u>	16.914
<u>Spain:</u>	6.573
Portugal:	4.875
Chile:	4.415
<u>Argentina:</u>	3.756
Austria:	2.799
Australia:	2.294
<u>Germany</u> :	2.120

Heatmap di correlazione tra prezzo e punteggio



```
plt.figure(figsize=(10, 6))
correlation_matrix = df_clean[['price', 'points']].corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Heatmap di correlazione tra prezzo e punteggio')
plt.show()
```

Spiegazione:

"<u>plt.figure</u>": Imposta le dimensioni della figura

".corr": Calcola la matrice di correlazione tra 'price' e 'points'

"<u>sns.heatmap</u>": Genera una heatmap utilizzando la matrice di correlazione.

"<u>plt.title</u>":Titolo del grafico

"plt.show":Mostra il grafico

Commento:

Correlazione tra Punteggio e Prezzo (0.42)

Questo valore è situato nella parte superiore destra della heatmap ed indica una correlazione positiva moderata tra il punteggio e il prezorrelazione forte, il che suggerisce che ci sono altri fattori che influenzano il prezzo oltre al punteggio, ovvero la presenza di vini con punteggi alti ma prezzi bassi.

Correlazione tra Prezzo e Prezzo (1):

Questa è una correlazione perfetta, come ci si aspetterebbe, dato che stiamo confrontando la variabile con se stessa. È indicata nella parte diagonale dell'heatmap e serve principalmente come punto di riferimento.

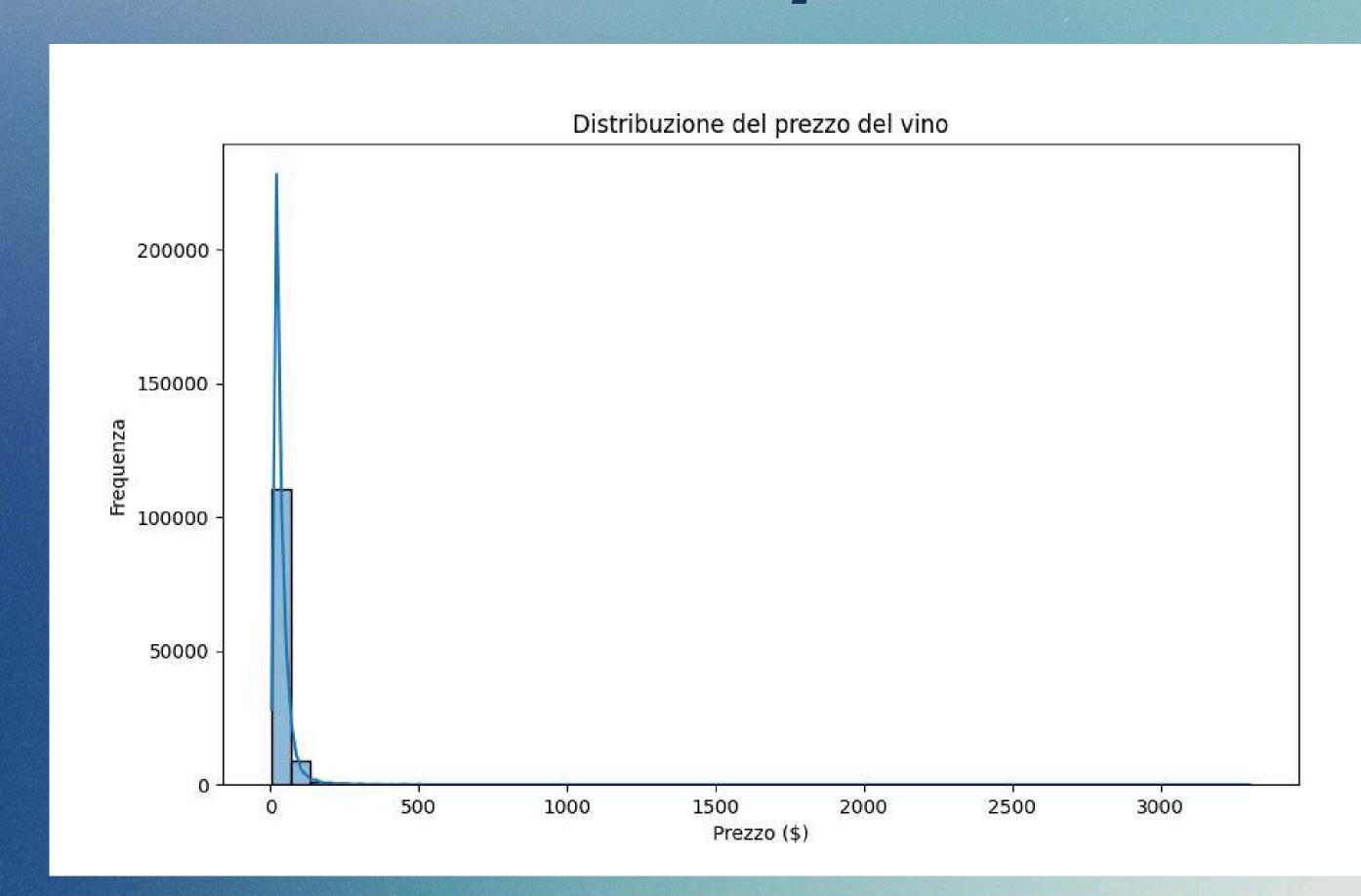
Correlazione tra Punteggio e Punteggio (1):

Analogamente, questa è una correlazione perfetta, indicata nella parte diagonale opposta. Anche in questo caso, è il risultato atteso.

Correlazione tra Price e Points (0.42):

Questo valore è situato nella parte inferiore sinistra della heatmap ed è una rappresentazione simmetrica della correlazione tra punteggio e prezzo, confermando i risultati precedenti.

Distribuzione del prezzo del vino



```
plt.figure(figsize=(10, 6))
sns.histplot(df_clean['price'], bins=50, kde=True)
plt.title('Distribuzione del prezzo del vino')
plt.xlabel('Prezzo ($)')
plt.ylabel('Frequenza')
plt.show()
```

Spiegazione:

"plt.figure": Imposta le dimensioni della figura

"sns.histplot": Crea un istogramma della distribuzione del prezzo.

"plt.title":Titolo del grafico

"plt.xlabel":Etichetta asse X

"plt.ylabel": Etichetta asse Y

"plt.show":Mostra il grafico

Commento:

La distribuzione dei prezzi dei vini mostra una chiara distribuzione asimmetrica positiva. La maggior parte dei vini si trova nella fascia di prezzo bassa, al di sotto dei 100\$, con una concentrazione molto alta in questa gamma.

Osservazioni principali:

La coda lunga a destra indica la presenza di vini che hanno prezzi molto più alti, probabilmente etichette premium o di nicchia, ma sono una parte estremamente ridotta del dataset. Questi valori sono considerati outlier rispetto alla maggior parte dei dati.

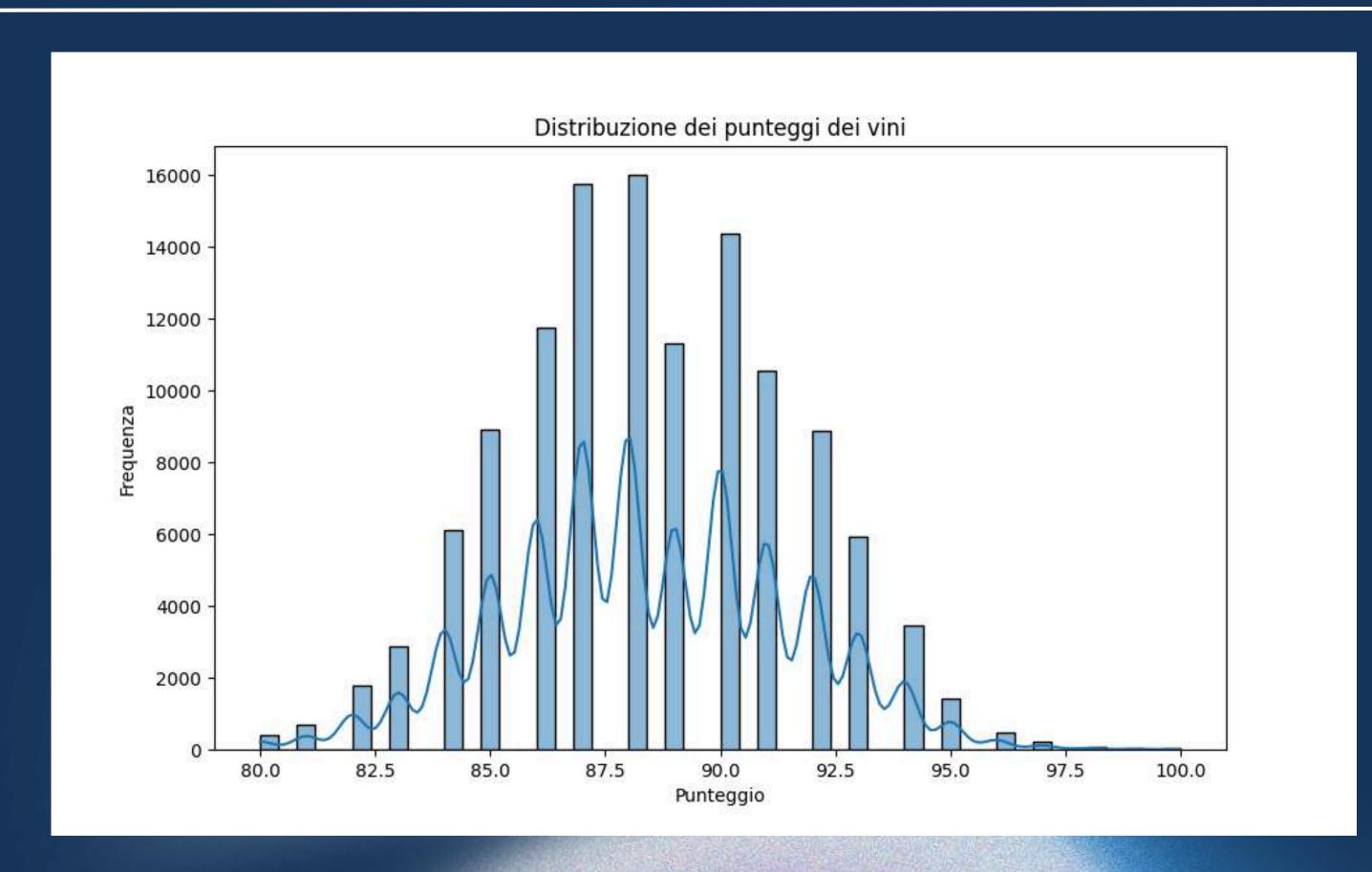
Mediana vs Media:

In un contesto come questo, la mediana è più indicativa della centralità del prezzo rispetto alla media. Infatti, poiché la media è influenzata dagli outlier (i vini molto costosi), essa potrebbe risultare più alta della mediana, distorcendo così l'interpretazione del prezzo tipico.

Conclusione:

Il mercato del vino sembra dominato da vini di fascia medio-bassa. L'analisi dei vini di fascia alta potrebbe richiedere un approfondimento separato, poiché potrebbero essere prodotti destinati a un pubblico specifico.

Distribuzione dei punteggi dei vini



```
plt.figure(figsize=(10, 6))
ax = sns.histplot(df_clean['points'], bins=50, kde=True)
plt.title('Distribuzione dei punteggi dei vini')
plt.xlabel('Punteggio')
plt.ylabel('Frequenza')
plt.show()
```

Spiegazione:

"<u>plt.figure</u>": Imposta le dimensioni della figura

"<u>sns.histplot</u>": Crea un istogramma della distribuzione dei punteggi dei vini.

"<u>plt.title</u>":Titolo del grafico

"plt.xlabel": Etichetta dell'asse X

"<u>plt.ylabel</u>": Etichetta dell'asse Y

"plt.show":Mostra il grafico

Commento:

La distribuzione dei punteggi dei vini, segue una distribuzione quasi normale con un leggero spostamento verso destra.

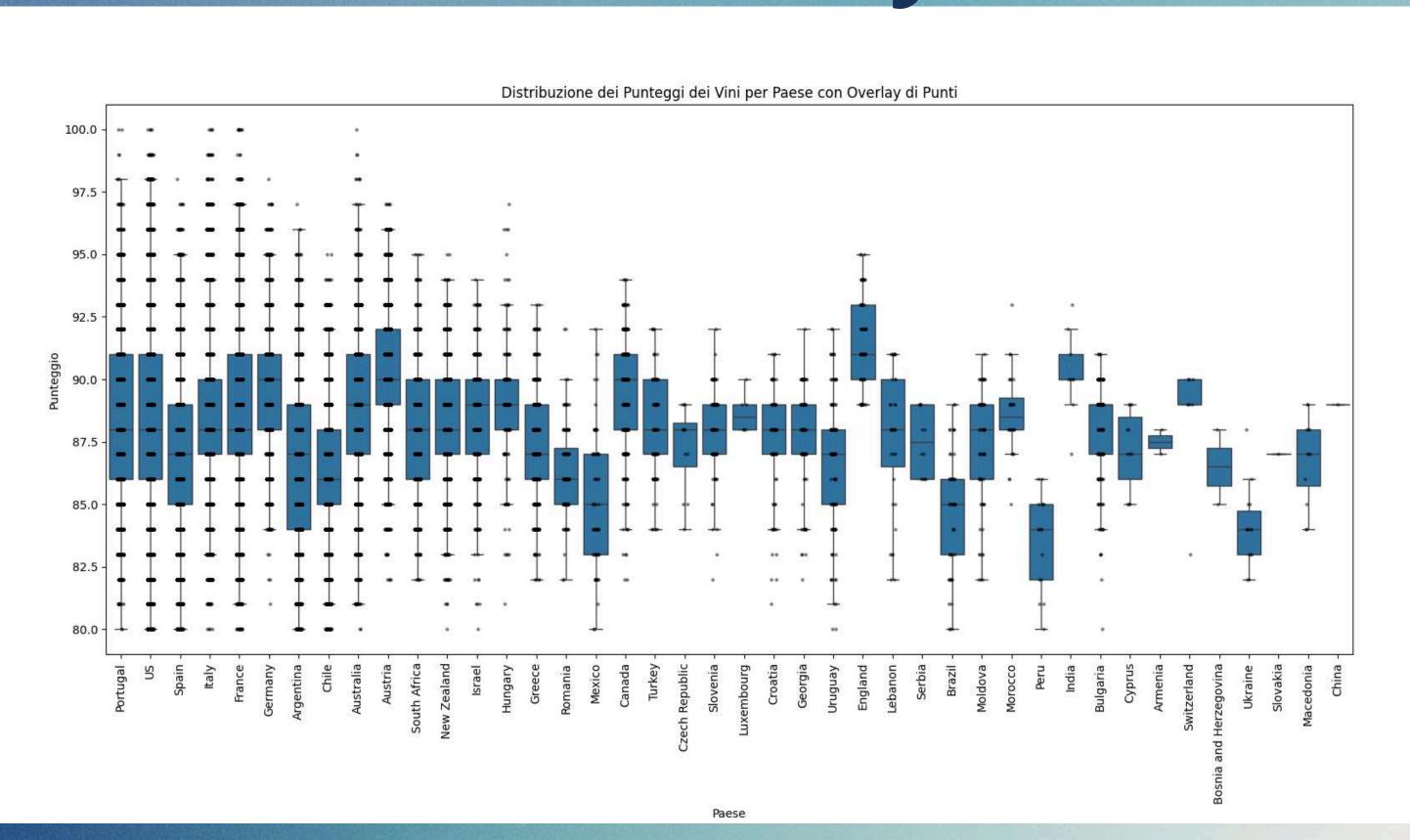
Osservazioni principali:

La maggior parte dei vini ha un punteggio tra 85 e 92, con un picco visibile attorno a 88. Questa distribuzione suggerisce che la maggior parte dei vini recensiti ha una qualità che rientra nella fascia medio-alta. I vini con punteggi estremamente bassi (< 85) o estremamente alti (> 95) sono relativamente rari.

Conclusioni:

La concentrazione dei vini nella fascia tra 85 e 92 indica che il mercato tende a recensire vini di una certa qualità standard. Tuttavia, sarebbe utile esplorare in che modo il punteggio incide sul prezzo e se ci sono correlazioni tra punteggio e le varietà o le regioni di produzione.

<u>Distribuzione dei Punteggi dei Vini per</u> <u>Paese con Overlay di Punti</u>



```
# Boxplot con overlay di punti per i punteggi dei vini per paese
plt.figure(figsize=(14, 7))
sns.boxplot(x='country', y='points', data=df_clean, showfliers=False)
sns.stripplot(x='country', y='points', data=df_clean, color='black', alpha=0.5, size=3)
plt.title('Distribuzione dei Punteggi dei Vini per Paese con Overlay di Punti')
plt.xticks(rotation=90)
plt.ylabel('Punteggio')
plt.xlabel('Paese')
plt.show()
```

Spiegazione:

"plt.figure": Imposta le dimensioni della figura

"sns.boxplot": Crea un boxplot per visualizzare la distribuzione dei punteggi per paese.

"sns.stripplot": Aggiunge i punti reali dei punteggi sopra il boxplot

"plt.title":Titolo del grafico

"plt.xticks":Ruota le etichette dell'asse X per una migliore leggibilità

"plt.xlabel":Etichetta asse X

"plt.ylabel": Etichetta asse Y

"plt.show":Mostra il grafico

Commento:

Il grafico boxplot rappresenta la distribuzione dei punteggi dei vini per paese, con un overlay di punti che mostra i dati effettivi.

Ogni boxplot indica la distribuzione del punteggio dei vini per ciascun paese, evidenziando la mediana, l'intervallo interquartile e gli eventuali outlier.

Si osserva una notevole variabilità nei punteggi tra i diversi paesi. Ad esempio, i vini di paesi come il Portogallo, la Francia, l'Italia e gli Stati Uniti tendono a mostrare punteggi medi elevati, con valori che raggiungono spesso i 95 punti o più. Paesi come la Svizzera, il Libano, e il Brasile mostrano boxplot con punteggi medi più bassi e un intervallo più ampio, indicando una maggiore variabilità nella qualità dei vini prodotti.

L'overlay di punti neri sovrapposti ai boxplot aiuta a visualizzare la distribuzione reale dei punteggi, rendendo evidente che alcuni paesi, come il Cile e la Spagna, presentano punteggi che si concentrano in una gamma specifica, mentre altri paesi come la Georgia o l'India mostrano una dispersione più ampia.

Conclusioni e Raccomandazioni

1. Posizionamento del Marketplace

L'obiettivo del marketplace sarà quello di connettere piccoli produttori locali con acquirenti globali, enfatizzando l'unicità e la qualità dei loro vini. Dato l'interesse crescente per i vini artigianali e biologici, il marketplace si posizionerà come una piattaforma che valorizza la storia e l'autenticità dietro ogni etichetta.

2. Varietà di Offerta

L'analisi dei dati ha rivelato che il Pinot Noir, il Chardonnay e il Cabernet Sauvignon sono le varietà più rappresentate e richieste. Pertanto, il marketplace dovrebbe assicurarsi di avere una selezione equilibrata di queste varietà, insieme a vini di nicchia provenienti da regioni meno note, come il Portogallo, per attrarre intenditori e collezionisti.

3. Strategia di Prezzo

La maggior parte dei vini si colloca nella fascia di prezzo medio-bassa, con una concentrazione significativa al di sotto dei 100\$. Si raccomanda di:

<u>Stabilire diverse fasce di prezzo</u>: Offrire vini a prezzi accessibili accanto a opzioni premium per attrarre un pubblico più ampio.

<u>Evidenziare i vini outlier</u>: I vini con punteggi elevati o prezzi sopra i 500\$ dovrebbero essere ben pubblicizzati come prodotti esclusivi, in grado di attirare collezionisti e intenditori.

4. Analisi del Mercato

L'analisi ha mostrato una correlazione positiva tra prezzo e punteggio (0.42). Questo suggerisce che i vini di alta qualità tendono ad avere prezzi superiori, ma non è sempre garantito. Pertanto:

Educazione del cliente: Creare contenuti che aiutino gli acquirenti a comprendere il valore dei vini, inclusi fattori come la varietà, la regione, e le pratiche di produzione.

Recensioni e valutazioni: Includere recensioni dettagliate da esperti e utenti per rafforzare la fiducia del consumatore e migliorare l'esperienza d'acquisto.

Conclusioni e Raccomandazioni

5. Target di Clientela

Focalizzarsi su diverse categorie di clienti:

Appassionati di vino: Offrire vini di alta qualità e informazioni dettagliate sulle varietà e le pratiche di produzione.

Collezionisti: Creare una sezione dedicata ai vini rari e costosi, con opzioni di acquisto limitate o esclusive.

Nuovi acquirenti: Fornire guide e consigli per i principianti, con sezioni di vini a prezzi accessibili e abbinamenti suggeriti.

6. Promozione e Marketing

Utilizzare strategie di marketing mirate per attrarre diversi segmenti di clientela:

Collaborazioni con influencer: Coinvolgere esperti e appassionati di vino per promuovere il marketplace sui social media.

Eventi di degustazione online: Organizzare eventi virtuali per presentare vini selezionati e interagire direttamente con i clienti.

Contenuti educativi: Creare blog e video che esplorano le varietà di vino, le tecniche di degustazione e le tendenze del settore.

7. Analisi dei Dati

Implementare strumenti di analisi per monitorare le vendite, le recensioni dei clienti e le tendenze del mercato:

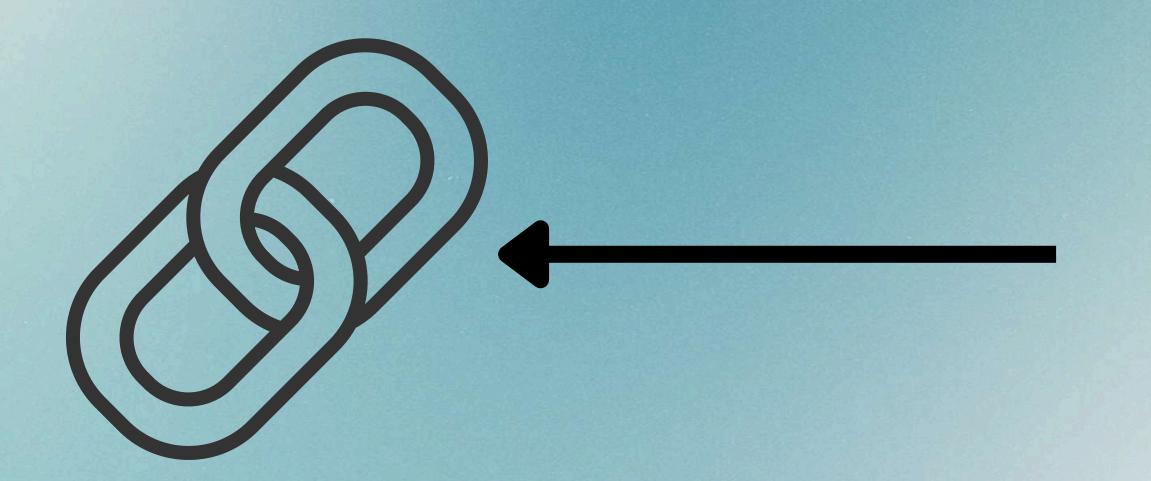
Raccolta di feedback: Utilizzare sondaggi e recensioni per migliorare continuamente l'offerta e l'esperienza del cliente.

Aggiornamenti dell'inventario: Monitorare le vendite per adattare l'inventario in base alla domanda e alle tendenze emergenti.

Conclusione Finale

Un marketplace dedicato ai vini dovrebbe mirare a creare un ambiente dove i produttori locali possano mostrare la loro unicità e qualità, mentre gli acquirenti possono scoprire e apprezzare un'ampia gamma di vini. La chiave per il successo sarà un equilibrio tra varietà, prezzo, e promozione, supportato da un'analisi continua del mercato e delle preferenze dei consumatori. Con queste strategie, il marketplace non solo attrarrà clienti, ma contribuirà anche a valorizzare il patrimonio vitivinicolo locale.

GitHub



Clicca qui per visualizzare il Codice