

PROGETTO PYTHON CON PANDAS

ALESSANDRO SMAJLOVIC



PROGETTO FOOD

IN QUESTO DATASET TROVERAI 130MILA RECENSIONI DI VINI, DI CUI VENGONO INDICATI VARIETÀ, PROVENIENZA, VIGNA, PREZZO E DESCRIZIONE.

IMMAGINA ORA DI VOLER CREARE UN MARKETPLACE DI VINI PER METTERE IN CONTATTO I PICCOLI PRODUTTORI LOCALI CON ACQUIRENTI DA TUTTO IL MONDO

POTRESTI PROVARE A CAPIRE QUALI SONO LE VARIETÀ E LE VIGNE PIÙ APPREZZATE.

UN POSSIBILE OUTPUT DEL TUO LAVORO POTREBBE ESSERE LA PROPOSTA DI UNA STRATEGIA PER L'ASSORTIMENTO DA CUI PARTIRE PER IL MARKETPLACE DI VINO CHE VORRESTI CREARE.



IMPOSTAZIONE DEL PROGETTO

EXPLORATION



DATA CLEANING



DATA ANALYSIS



VISUALIZATION

COME PRIMA COSA SONO ANDATO A IMPORTARE LE VARIE LIBRERIE PIU' IL FILE CSV SU JUPYTER



```
: # standard libraries
import numpy as np
import pandas as pd

# data visualization libraries
import seaborn as sns
from matplotlib import pyplot as plt
import plotly.express as px

# data cleaning library
import re

df=pd.read_csv('winemag-data-130k-v2.csv')
df.copy()
```

Unnamed: 0		country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	Wt
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Pc
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	F
3	3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	
4	4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	F
...	
129966	129966	Germany	Notes of honeysuckle and cantaloupe sweeten th...	Brauneberger Juffer-Sonnenuhr Spätlese	90	28.0	Mosel	NaN	NaN	Anna Lee C. Iijima	NaN	Dr. H. Thanisch (Erben Müller-Burggraef) 2013 ...	
			Citation is given as										Citation 2004

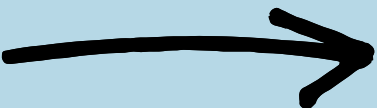
CON L'APERTURA DEL FILE POSSIAMO DARE UNA PRIMA OCCHIATA AL NOSTRO DATAFRAME.

**INIZIO LA FASE DI ESPLORAZIONE.
PRIMA GUARDANDO IN CHIARO TUTTE LE COLONNE A DISPOSIZIONE.**



```
In [9]: df.columns
Out[9]: Index(['Unnamed: 0', 'country', 'description', 'designation', 'points', 'price', 'province', 'region_1', 'region_2', 'taster_name', 'taster_twitter_handle', 'title', 'variety', 'winery'], dtype='object')
```

**RESTITUISCO UNA SERIE DI STATISTICHE RIASSUNTIVE PER
CIASCUNA COLONNA DEL DATAFRAME.**



```
df.describe()
```

	Unnamed: 0	points	price
count	129971.000000	129971.000000	120975.000000
mean	64985.000000	88.447138	35.363389
std	37519.540256	3.039730	41.022218
min	0.000000	80.000000	4.000000
25%	32492.500000	86.000000	17.000000
50%	64985.000000	88.000000	25.000000
75%	97477.500000	91.000000	42.000000
max	129970.000000	100.000000	3300.000000

OTTENGO UNA PANORAMICA DELLE INFORMAZIONI.



```
#Columns types and non null count
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129971 entries, 0 to 129970
Data columns (total 14 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Unnamed: 0                            129971 non-null int64
1   country                               129908 non-null object
2   description                           129971 non-null object
3   designation                           92506 non-null  object
4   points                                129971 non-null int64
5   price                                 120975 non-null float64
6   province                              129908 non-null object
7   region_1                              108724 non-null object
8   region_2                              50511 non-null  object
9   taster_name                           103727 non-null object
10  taster_twitter_handle                 98758 non-null  object
11  title                                 129971 non-null object
12  variety                               129970 non-null object
13  winery                                129971 non-null object
dtypes: float64(1), int64(2), object(11)
memory usage: 13.9+ MB
```



```
# Definisco una funzione per estrarre l'anno da una stringa usando regex
def estrai_anno_da_stringa(testo):
    year = re.findall(r'\b\d{4}\b', testo) # Trova tutti i pattern YYYY nel testo
    if year:
        return int(year[0]) # Prende il primo anno trovato come intero
    else:
        return None

# Applico la funzione al campo 'title' per estrarre l'anno
df['year'] = df['title'].apply(estrai_anno_da_stringa)

# Gestisco i valori NaN e infiniti sostituendoli con 0
df['year'].fillna(0, inplace=True)

df['year'] = df['year'].astype(int)

# Dropping unnecessary columns
df.drop(['Unnamed: 0', 'province', 'region_1', 'region_2', 'taster_name', 'taster_twitter_handle','title'], axis=1, inplace=True)
#Visualizza il Dataframe con l'anno estratto
df
```

**PROSEGUO CON UN PO' DI PULIZIA DEI DATI ,
ESTRAENDO L'ANNO PER CREARE UNA NUOVA
COLONNA APPOSITA E SUCCESSIVAMENTE
ELIMINANDO TUTTE LE COLONNE NON DI MIO
INTERESSE.**

**COSÌ HO SNELLITO LE COLONNE PER
AVERE UNA MAGGIORE
COMPRENSIONE E FACILITÀ
DELL'ANALISI.**

	country	description	designation	points	price	variety	winery	year
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	White Blend	Nicosia	2013
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Portuguese Red	Quinta dos Avidagos	2011
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Pinot Gris	Rainstorm	2013
3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Riesling	St. Julian	2013
4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Pinot Noir	Sweet Cheeks	2012
...
129966	Germany	Notes of honeysuckle and cantaloupe sweeten th...	Brauneberger Juffer-Sonnenuhr Spätlese	90	28.0	Riesling	Dr. H. Thanisch (Erben Müller-Burggraef)	2013
129967	US	Citation is given as much as a decade of bottl...	NaN	90	75.0	Pinot Noir	Citation	2004
129968	France	Well-drained gravel soil gives this wine its C...	Kritt	90	30.0	Gewürztraminer	Domaine Gresser	2013
129969	France	A dry style of Pinot Gris, this is crisp with ...	NaN	90	32.0	Pinot Gris	Domaine Marcel Deiss	2012
129970	France	Big, rich and off-dry, this is powered by inte...	Lieu-dit Harth Cuvée Caroline	90	21.0	Gewürztraminer	Domaine Schoffit	2012

129971 rows x 8 columns

CORRELAZIONE TRA PUNTEGGIO E PREZZO

IL GRAFICO MOSTRATO È UN GRAFICO SCATTER PLOT (A DISPERSIONE) CHE RAPPRESENTA LA CORRELAZIONE TRA PUNTEGGI E PREZZI DEI VINI NEL DATAFRAME.

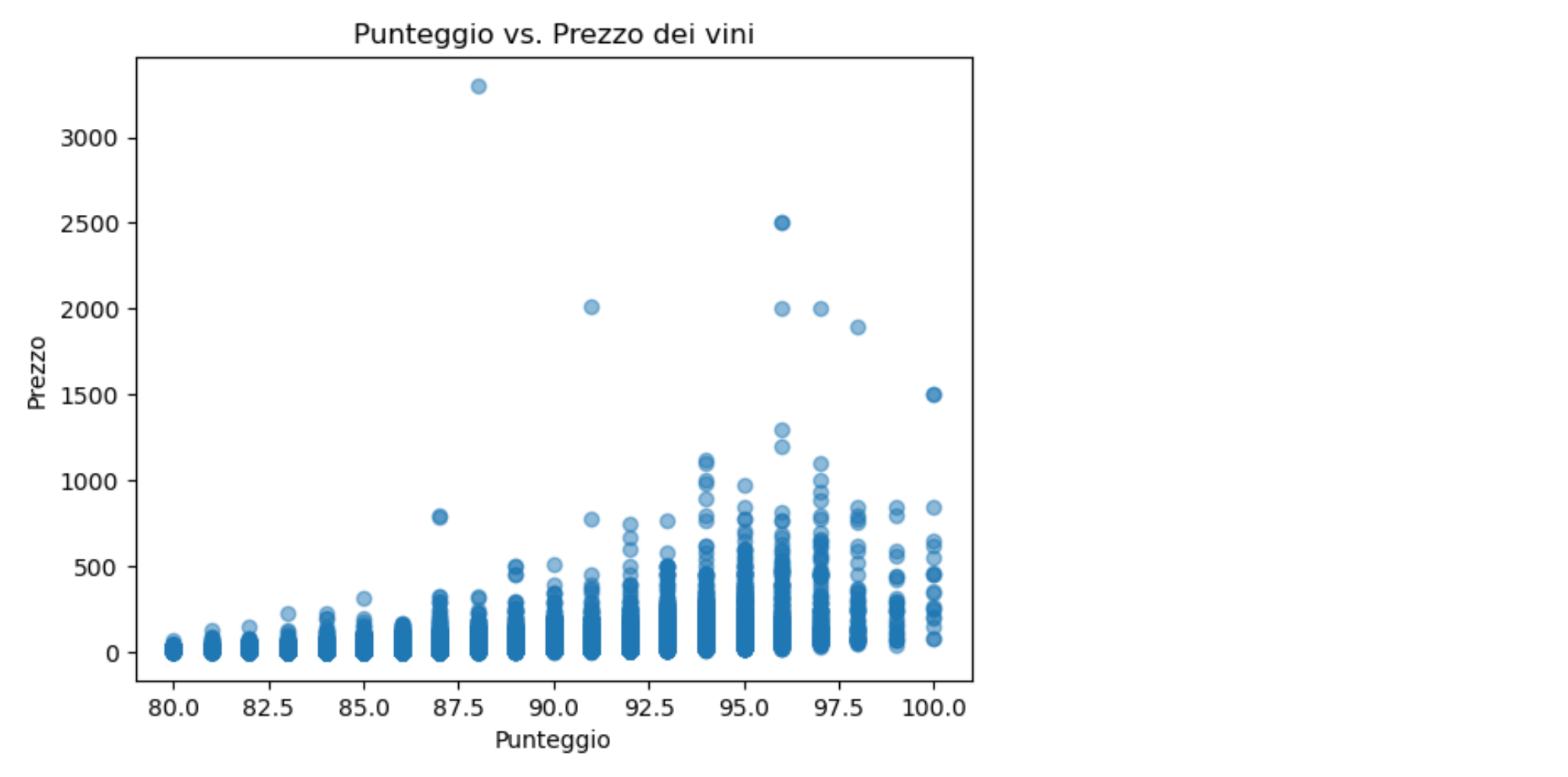
SULL’ASSE X NEL GRAFICO È RAPPRESENTATO IL PUNTEGGIO, MENTRE SULL’ASSE Y IL PREZZO.

IL GRAFICO DI DISPERSIONE PERMETTE DI VISUALIZZARE LA RELAZIONE O IL MODELLO TRA LE DUE VARIABILI.

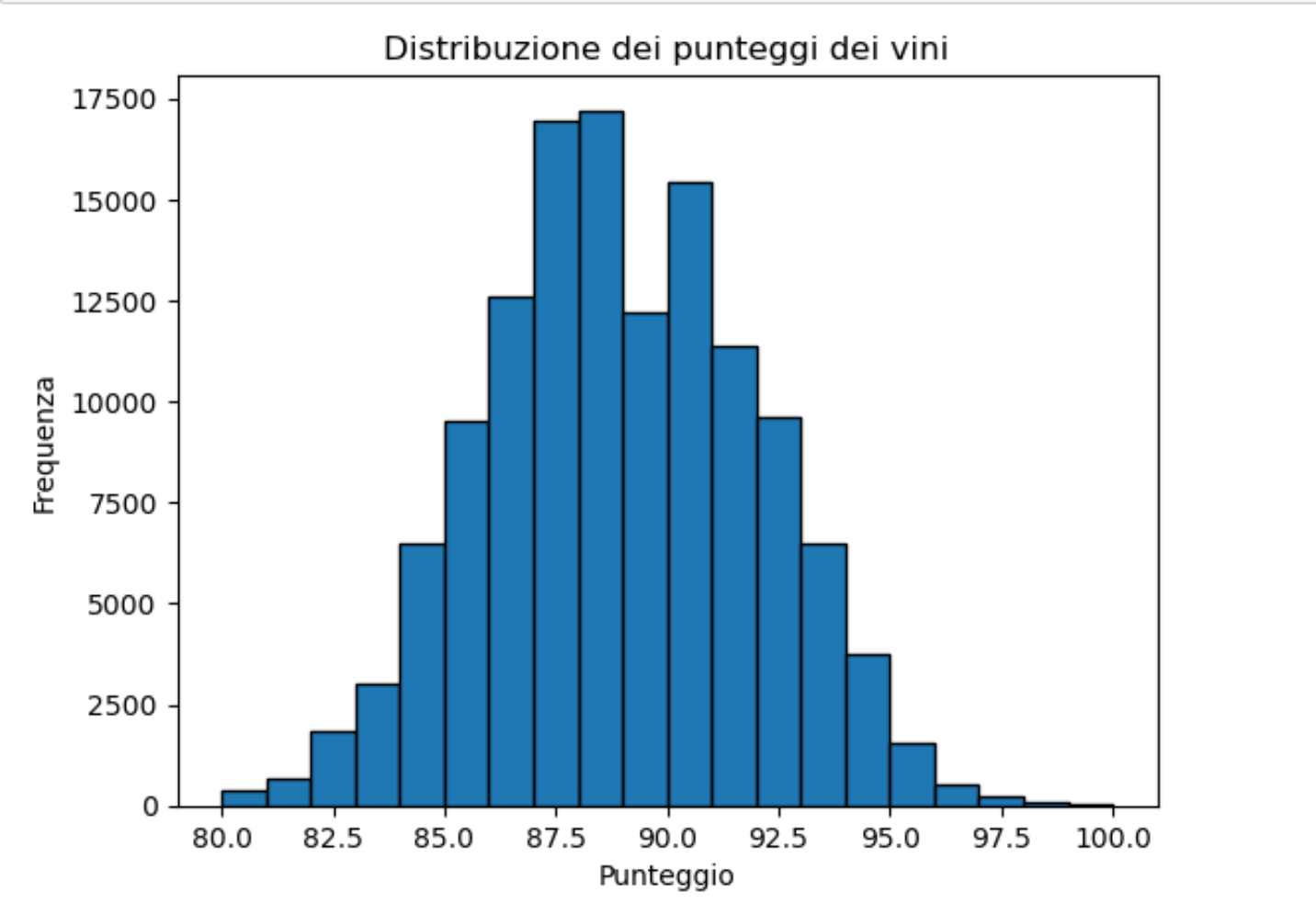
I PUNTI NEL GRAFICO MOSTRANO LA DISTRIBUZIONE DEI DATI, INDICANDO OGNI PUNTO COME UN VINO.

QUANDO I PUNTI SI CONCENTRANO SU UNA PARTE DI UNA LINEA SIGNIFICA CHE C’È UNA CORRELAZIONE.

```
# Creo un grafico a dispersione tra punteggi e prezzi, eseguo un scatter plot tra punteggi e prezzo (points vs. price):
plt.scatter(df['points'], df['price'],alpha=0.5)
plt.xlabel('Punteggio')
plt.ylabel('Prezzo')
plt.title('Punteggio vs. Prezzo dei vini')
plt.show()
```

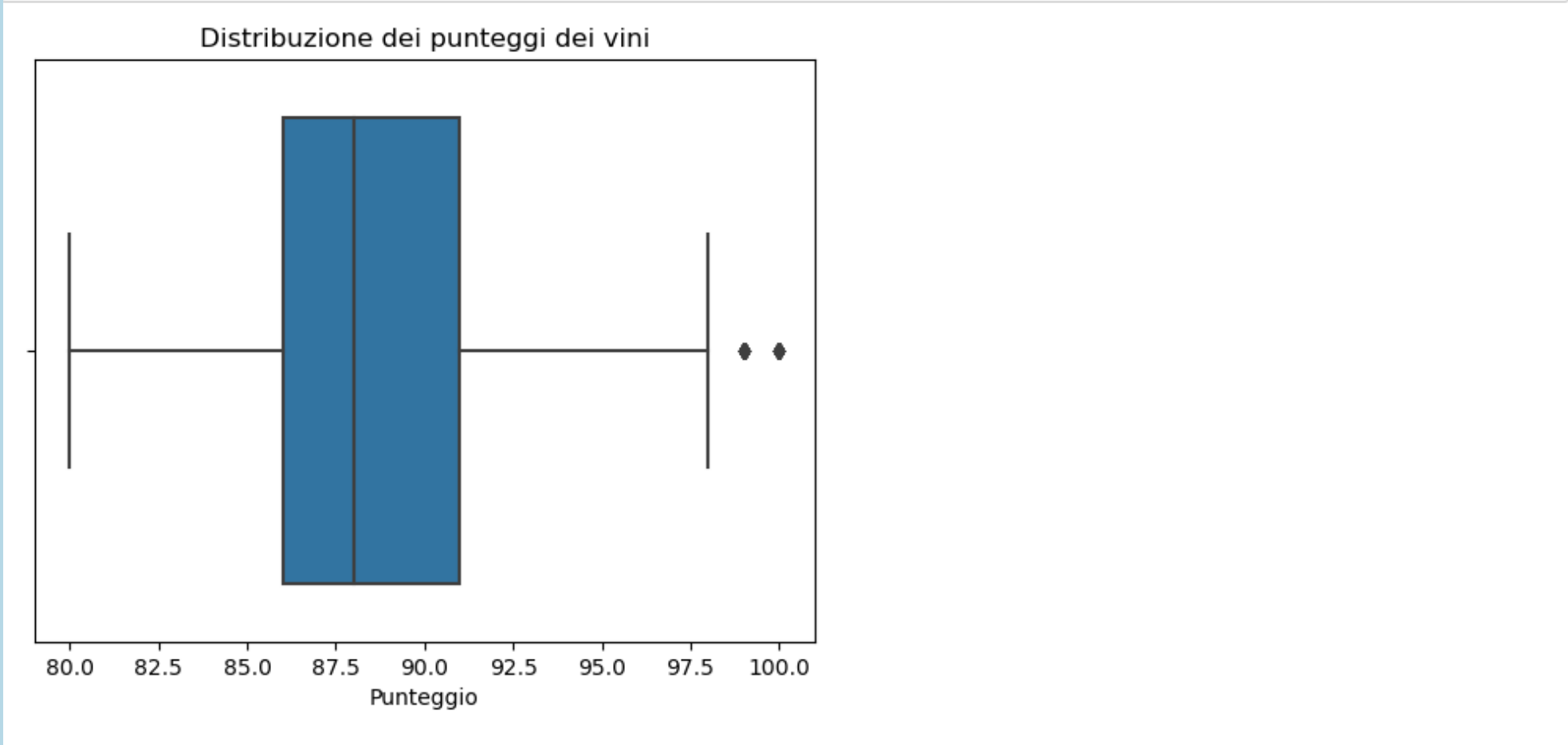


```
# Creo un istogramma dei punteggi
plt.hist(df['points'],bins=20,edgecolor='k')
plt.xlabel('Punteggio')
plt.ylabel('Frequenza')
plt.title('Distribuzione dei punteggi dei vini')
plt.show()
```



UN ISTOGRAMMA RAPPRESENTA LA DISTRIBUZIONE DEI DATI UTILIZZANDO BARRE VERTICALI, DOVE CIASCUNA BARRA RAPPRESENTA UN INTERVALLO DI VALORI. L’ALTEZZA DELLE BARRE INDICA LA FREQUENZA O IL CONTEGGIO DEI DATI ALL’INTERNO DI QUELL’ INTERVALLO.

```
# Crea un box plot dei punteggi, Un box plot fornisce
# una panoramica della distribuzione dei punteggi e permette di visualizzare la presenza di eventuali valori anomali (outliers).
sns.boxplot(x=df['points'])
plt.xlabel('Punteggio')
plt.title('Distribuzione dei punteggi dei vini')
plt.show()
```

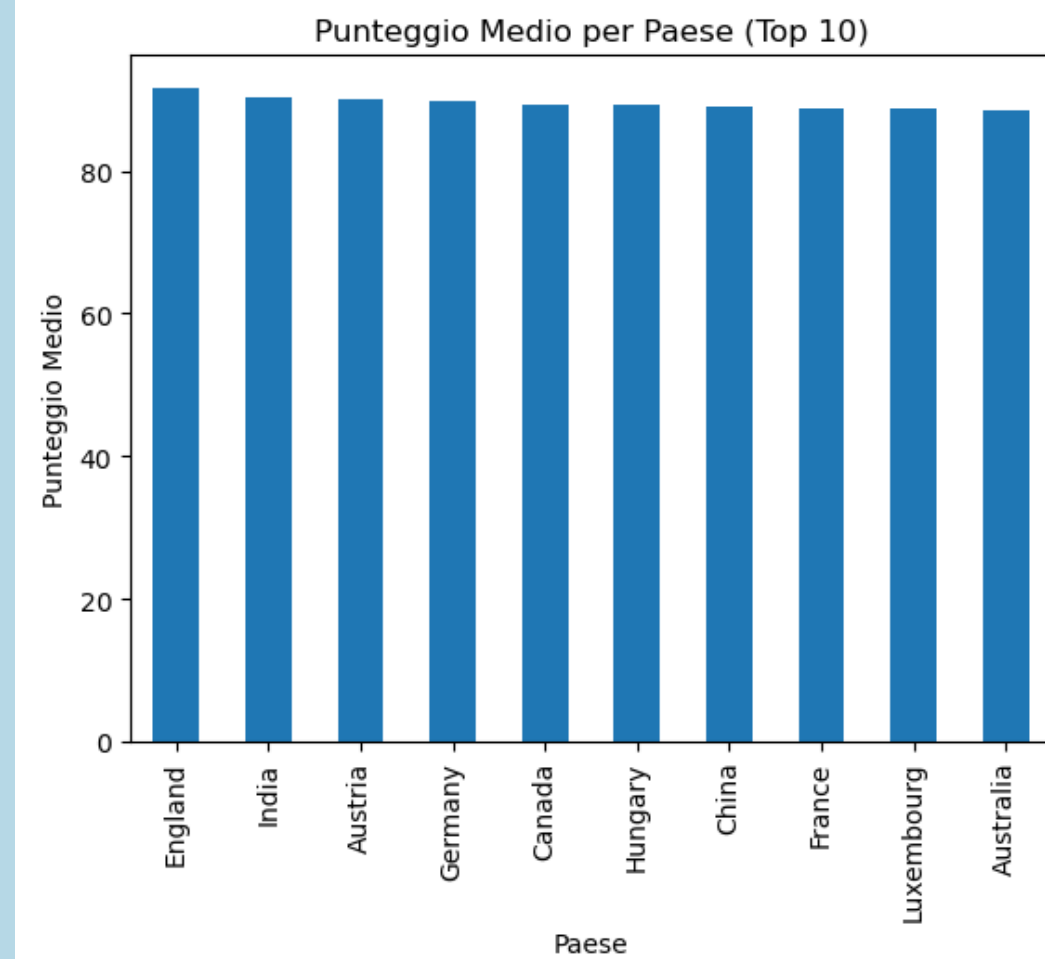


UN BOX PLOT È UTILE PER COMPRENDERE LA CENTRALITÀ DELLA DISTRIBUZIONE, LA DISPERSIONE, LA PRESENZA DI VALORI ANOMALI E LA SIMMETRIA DEI DATI

QUESTI GRAFICI (ISTOGRAMMA E BOXPLOT) IN ENTRAMBI I CASI RESTITUISCONO LA DISTRIBUZIONE CON LA FASCIA PIÙ ALTA DI PUNTEGGIO DEI VINI.

PUNTEGGIO MEDIO PER PAESE

```
# Grafico a barre che mostri i punteggi medi dei vini per ciascun paese (country):  
  
media_punteggi_per_paese = df.groupby('country')['points'].mean().sort_values(ascending=False)  
media_punteggi_per_paese.head(10).plot(kind='bar')  
plt.xlabel('Paese')  
plt.ylabel('Punteggio Medio')  
plt.title('Punteggio Medio per Paese (Top 10)')  
plt.show()
```



CALCOLANDO LA MEDIA,ATTRAVERSO UN GRAFICO A BARRE, SONO ANDATO A PRENDERE E RAFFIGURARE I PRIMI 10 PAESI COL PUNTEGGIO MEDIO MAGGIORE

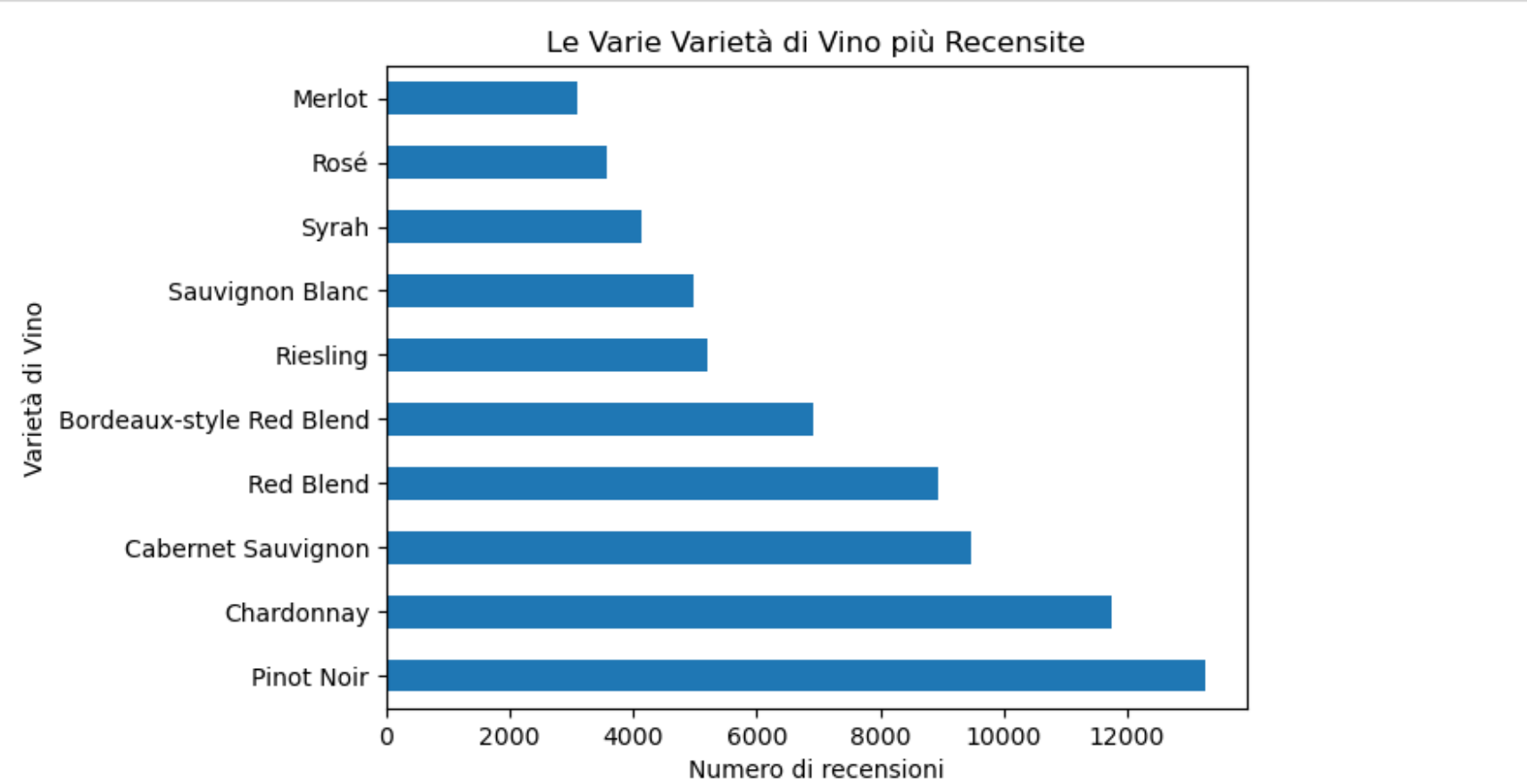
VARIETA' PIU' RECENSITE

QUESTO GRAFICO A BARRE ORIZZONTALE RAPPRESENTA LE PRIME 10 VARIETA' DI VINO PIU' RECENSITE.

```
# Calcolo il conteggio delle varietà di vino
conteggio_varietà = df['variety'].value_counts().sort_values(ascending=False)

# Ordino il conteggio in ordine decrescente
varietà_top = conteggio_varietà.head(10) # Mostra le prime 10 varietà più recensite

# Creo un grafico a barre delle varietà più recensite
varietà_top.plot(kind='barh')
plt.xlabel('Numero di recensioni')
plt.ylabel('Varietà di Vino')
plt.title('Le Varie Varietà di Vino più Recensite')
plt.show()
```



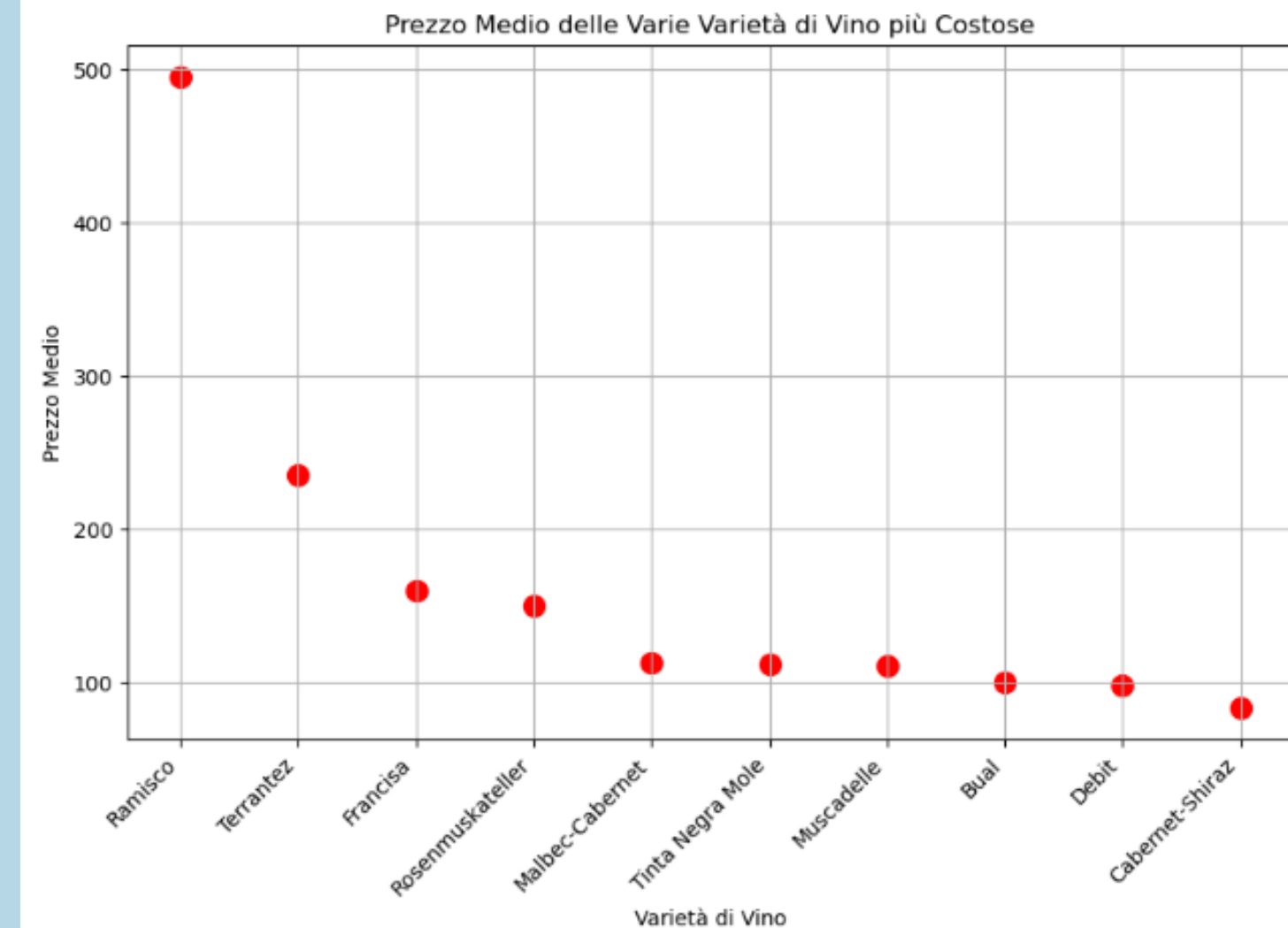
PREZZO MEDIO DELLE VARIETÀ DI VINO PIÙ COSTOSE

EFFETTUANDO UN ALTRA MEDIA PER OGNI TIPO DI VARIETA', ELENCO LE PRIME 10 VARIETA' COL PREZZO MEDIO PIU' ELEVATO RAPPRESENTATO IN QUESTO GRAFICO A DISPERSIONE.

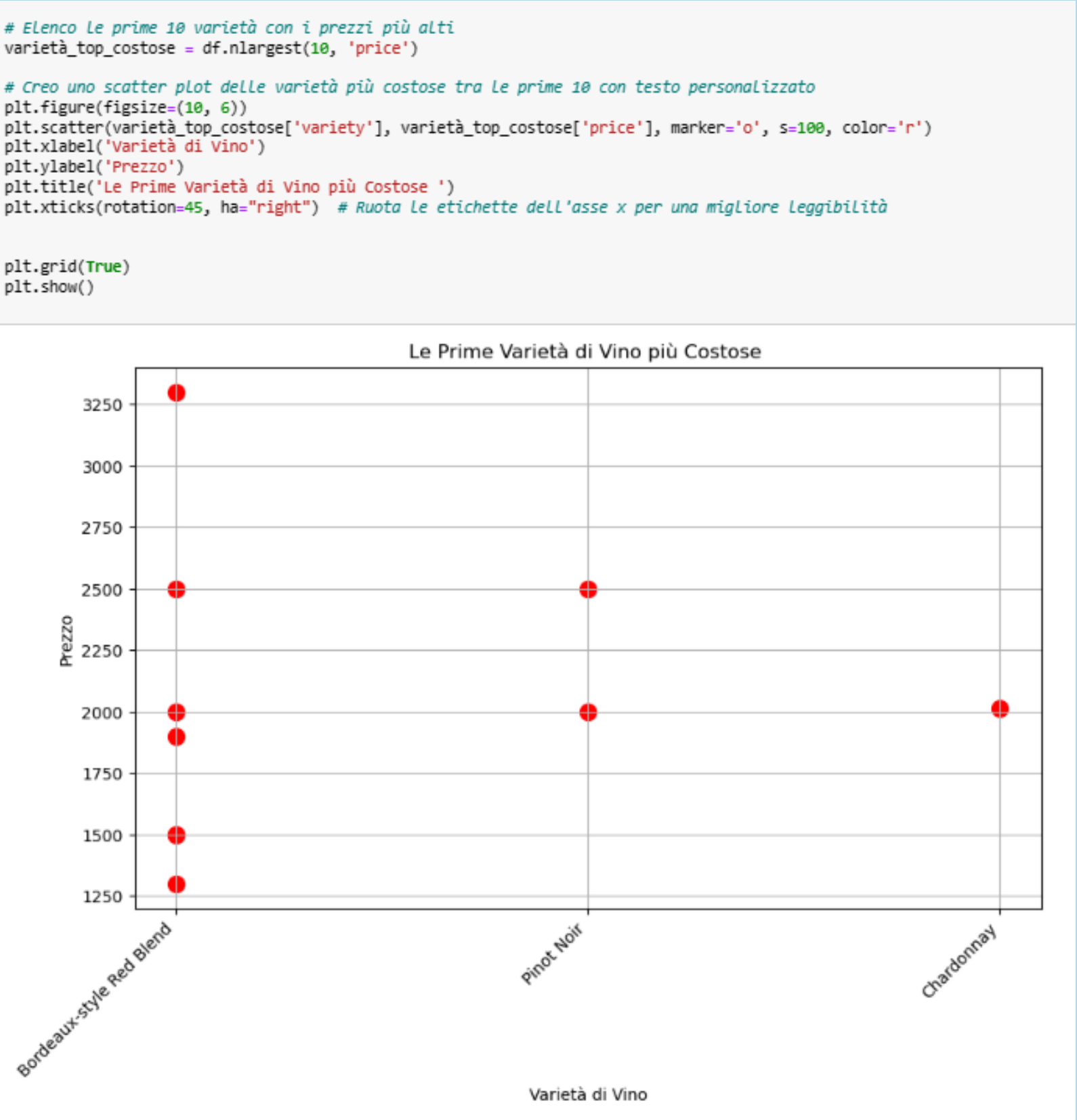
```
# Calcolo la media dei prezzi per ciascuna varietà di vino
media_prezzi_varietà = df.groupby('variety')['price'].mean()

# Elenco le prime 10 varietà con i prezzi medi più elevati
varietà_top_costose = media_prezzi_varietà.nlargest(10)

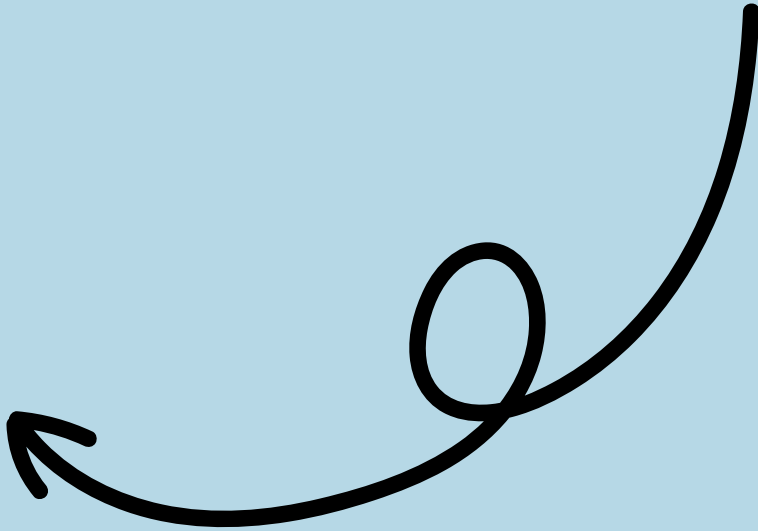
# Creo un grafico a dispersione delle varietà più costose
plt.figure(figsize=(10, 6))
plt.scatter(varietà_top_costose.index, varietà_top_costose.values, marker='o', s=100, color='r')
plt.xlabel('Varietà di Vino')
plt.ylabel('Prezzo Medio')
plt.title('Prezzo Medio delle Varie Varietà di Vino più Costose')
plt.xticks(rotation=45, ha="right") # Ruota le etichette dell'asse x per una migliore leggibilità
plt.grid(True)
plt.show()
```




VARIETÀ PIÙ COSTOSE




AL PRIMO POSTO DELLA VARIETÀ DI VINI PIÙ COSTOSI, SI CLASSIFICA IL BORDEAUX RED BLEND.




PRIMARY FLAVORS




Black Currant




Black Cherry



Graphite



Chocolate



Dried Herbs

TASTE PROFILE

Bone-dry

Full Body

High Tannins

Medium Acidity

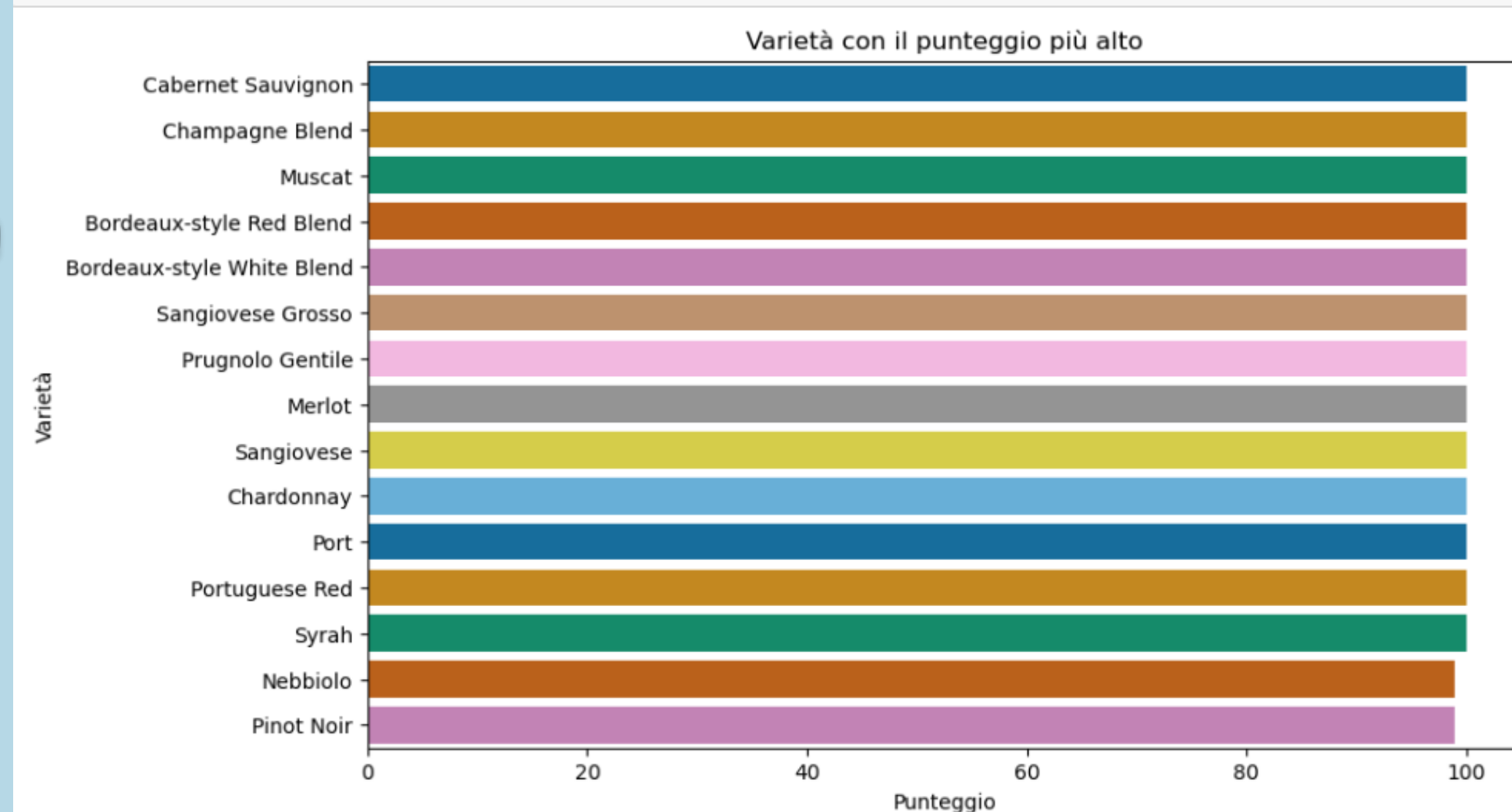
13.5–15% ABV

VARIETÀ CON ALTO PUNTEGGIO

IN QUESTO CASO METTIAMO IN EVIDENZA LE VARIETÀ CON IL PUNTEGGIO PIÙ ALTO, SEGUITO DAI VINI PIÙ ECONOMICI...

```
# Selezione le varietà di vino con il punteggio più alto (ad esempio, le prime 15)
higher_p = df.groupby(['variety'])['points'].max().sort_values(ascending=False).to_frame()[:15]

# Grafico che restituisce le varietà con il punteggio più alto
plt.figure(figsize=(10, 6))
ch1 = sns.barplot(x=higher_p['points'], y=higher_p.index, palette='colorblind')
ch1.set_xlabel('Punteggio')
ch1.set_ylabel('Varietà')
ch1.set_title('Varietà con il punteggio più alto')
plt.show()
```

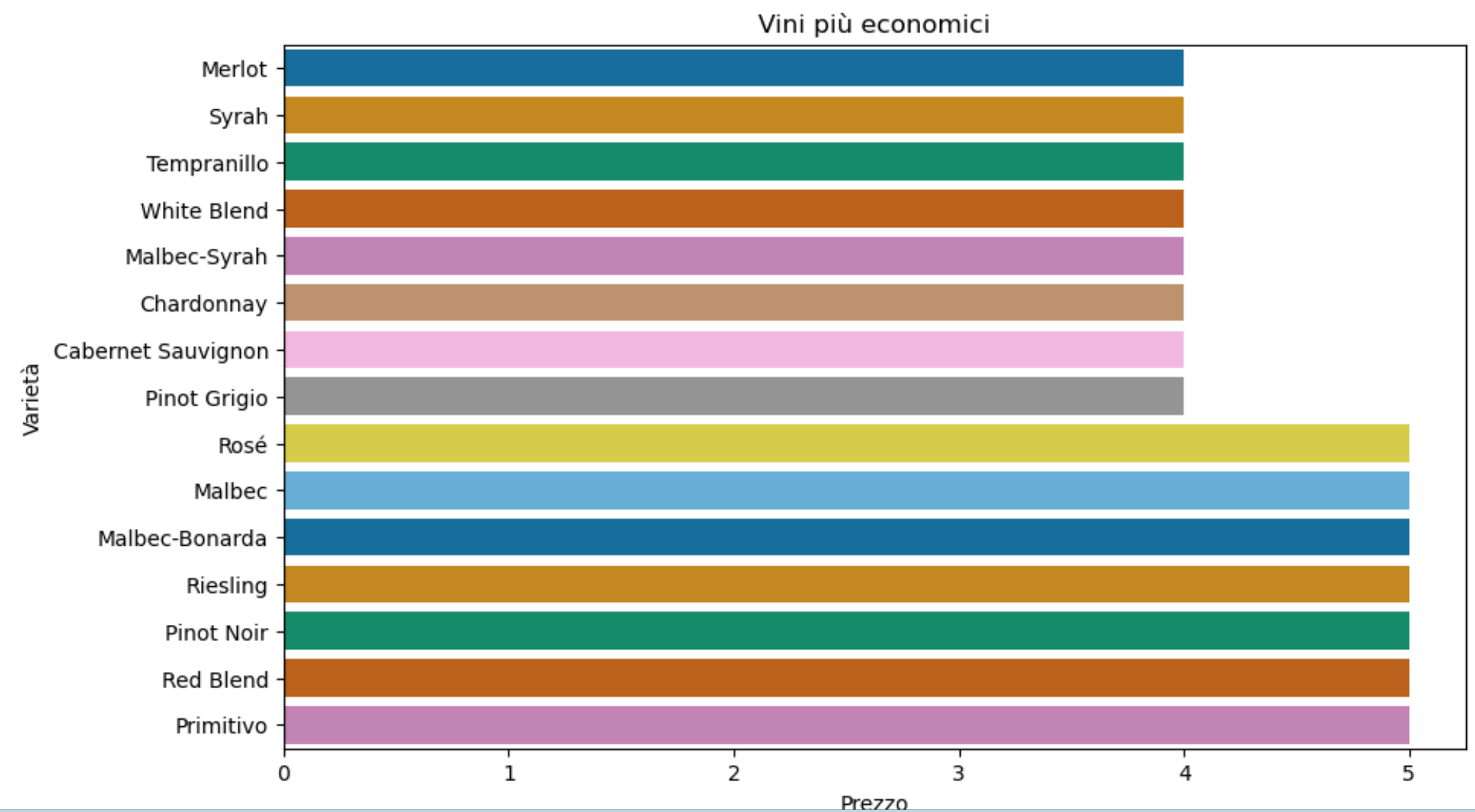


VINI PIÙ ECONOMICI

COSÌ FACENDO VEDIAMO I VINI PIÙ ECONOMICI, MA ANCHE CON UN ALTO PUNTEGGIO.

```
# Seleziona le varietà di vino con i prezzi più bassi (ad esempio, le prime 15)
ch_price = df.groupby(['variety'])['price'].min().sort_values().to_frame()[:15]

# Crea un grafico a barre orizzontali
plt.figure(figsize=(10, 6))
ch2 = sns.barplot(x=ch_price['price'], y=ch_price.index, palette='colorblind')
ch2.set_xlabel('Prezzo')
ch2.set_ylabel('Varietà')
ch2.set_title('Vini più economici')
plt.show()
```



APERTURA MARKETPLACE DI VINI INTERNAZIONALI.

HO FILTRATO I DATI PER UNA SELEZIONE DI VINI INTERNAZIONALI CHE SODDISFANO I CRITERI SPECIFICATI PER LA VENDITA NEL MARKETPLACE. I PUNTEGGI SUPERIORI A 90 INDICANO LA QUALITÀ SUPERIORE DEL VINO. I PREZZI COMPRESI TRA 100 E 500 RAPPRESENTANO UN RANGE DI PREZZI ACCESSIBILE PER I VINI INTERNAZIONALI DI ALTA QUALITÀ.

wine = df[(df['points'] > 90) & (df['price'] > 50) & (df['price'] < 500)] wine												
	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title
119	119	France	Medium-gold in color. Complex and inviting nos...	Schoenenbourg Grand Cru Vendanges Tardives	92	80.0	Alsace	Alsace	NaN	NaN	NaN	Dopff & Irion 2004 Schoenenbourg Grand Cru Ven...
120	120	Italy	Slightly backward, particularly given the vint...	Bricco Rocche Prapó	92	70.0	Piedmont	Barolo	NaN	NaN	NaN	Ceretto 2003 Bricco Rocche Prapó (Barolo)
130	130	Italy	At the first it was quite muted and subdued, b...	Bricco Rocche Brunate	91	70.0	Piedmont	Barolo	NaN	NaN	NaN	Ceretto 2003 Bricco Rocche Brunate (Barolo)
133	133	Italy	Einaudi's wines have been improving lately, an...	NaN	91	68.0	Piedmont	Barolo	NaN	NaN	NaN	Poderi Luigi Einaudi 2003 Barolo
134	134	US	Give this young Cab time in the cellar to come...	NaN	91	78.0	California	Napa Valley	Napa	NaN	NaN	Clark-Clauden 2007 Cabernet Sauvignon (Napa Va...
...
129931	129931	France	A powerful, chunky wine, packed with solid tan...	NaN	91	107.0	Burgundy	Grands-Echezeaux	NaN	Roger Voss	@vossroger	Henri de Villamont 2005 Grands-Echezeaux
129932	129932	Argentina	Andeluna's top wines tend to be ripe and	Pasionado	91	55.0	Mendoza Province	Uco Valley	NaN	Michael Schachner	@wineschach	Andeluna 2004 Pasionado Red (Uco Valley)

INVESTIMENTO DELLE BOTTIGLIE

L'OBIETTIVO DI QUESTO CODICE È CALCOLARE IL NUMERO DI BOTTIGLIE DI VINO DA ACQUISTARE E L'INVESTIMENTO TOTALE IN BASE A CRITERI SPECIFICI DI PUNTEGGIO E PREZZO.

```
# Imposto le soglie per il prezzo e il punteggio
prezzo_minimo = 50
prezzo_massimo = 500
soglia_punteggio = 90

# Definisco il costo per bottiglia
investimento_per_bottiglia = 100

# Filtro le bottiglie con prezzo compreso tra le soglie specificate e punteggio maggiore o uguale a 90
bottiglie_selezionate = wine[(wine['price'] >= prezzo_minimo) & (wine['price'] <= prezzo_massimo) &
                             (wine['points'] >= soglia_punteggio)]

# Calcolo il numero di bottiglie da acquistare e l'investimento totale
numero_di_bottiglie_da_acquistare = bottiglie_selezionate.shape[0]
investimento_totale = (numero_di_bottiglie_da_acquistare * investimento_per_bottiglia)

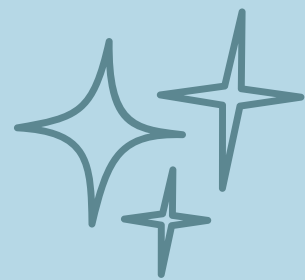
print(f'Numero di bottiglie da acquistare: {numero_di_bottiglie_da_acquistare}')
print(f'Investimento totale: {investimento_totale} euro')
```

```
Numero di bottiglie da acquistare: 12873
Investimento totale: 1287300 euro
```

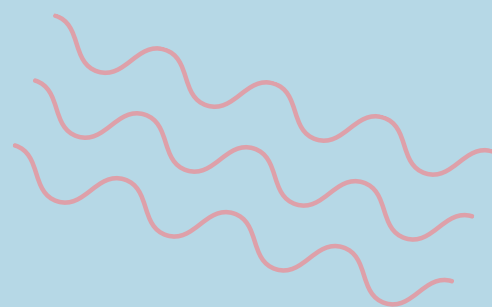
CONCLUSIONI

TRAMITE L'ANALISI DI QUESTO FILE ABBIAMO POTUTO CAPIRE COME NON NECESSARIAMENTE UN VINO COSTOSO HA UN PUNTEGGIO ALTO, E COME UN VINO ECONOMICO PUÒ AVERE UN PUNTEGGIO ELEVATO. ABBIAMO VISTO QUALI SONO I PUNTEGGI MEDI PER PAESE, LE VARIETÀ PIÙ RECENSITE E IL PREZZO MEDIO DELLE VARIETÀ PIÙ COSTOSE. INFINE ABBIAMO APERTO UN MARKETPLACE CALCOLANDO LA SPESA SECONDO I CRITERI SCELTI. QUESTA ANALISI CI HA AIUTATO A PRENDERE DECISIONI CONSAPEVOLI SULL'ACQUISTO DI BOTTIGLIE DI VINO IN BASE AI NOSTRI GUSTI E AL NOSTRO BUDGET. È STATO UN ESEMPIO DI COME L'ANALISI DEI DATI PUÒ ESSERE UTILIZZATA PER PRENDERE DECISIONI NEL MONDO REALE. IN CONCLUSIONE, QUESTO PROGETTO DIMOSTRA IL POTENZIALE DELL'ANALISI DEI DATI NEL PRENDERE DECISIONI INFORMATE E PERSONALIZZATE.

THANK YOU



CRAZIE PER LA VISIONE.



start2impact
UNIVERSITY