Data Manipulation and Visualization

Progetto di
Salvatore Nizza

start2impact
UNIVERSITY

Progetto

Creare un marketplace di vini per mettere in contatto i piccoli produttori locali con acquirenti da tutto il mondo provando a capire quali sono le varietà e le vigne più apprezzate, se ci sono outlier di prezzo, se c'è correlazione tra le variabili

Proporre una strategia per l'assortimento da cui partire per il marketplace di vino che vorresti creare tenendo a mente l'obiettivo di aiutare gli stakeholder principali a capire a fondo il settore in cui operano e a formulare delle strategie appropriate

La prima operazione che faccio è quella di importare le librerie con le quali esplorare i dati

```
import numpy as np
import pandas as pd
from sklearn.preprocessing import LabelEncoder
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
import seaborn as sns
```

Il dataset del seguente studio è disponibile sulla piattaforma Kaggle nel quale è possibile trovare circa 130mila recensioni di vini, di cui vengono indicati varietà, provenienza, vigna, prezzo e descrizione

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

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
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)	White Blend	Nicosia
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)	Portuguese Red	Quinta dos Avidagos
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)	Pinot Gris	Rainstorm
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 ...	Riesling	St. Julian
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...	Pinot Noir	Sweet Cheeks
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
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 ...	Riesling	Dr. H. Thanisch (Erben Müller-Burggraef)
129967	129967	US	Citation is given as much as a decade of bottl...	NaN	90	75.0	Oregon	Oregon	Oregon Other	Paul Gregutt	@paulgwine	Citation 2004 Pinot Noir (Oregon)	Pinot Noir	Citation
129968	129968	France	Well-drained gravel soil gives this wine its c...	Kritt	90	30.0	Alsace	Alsace	NaN	Roger Voss	@vossroger	Domaine Gresser 2013 Kritt Gewurztraminer (Als...	Gewürztraminer	Domaine Gresser
129969	129969	France	A dry style of Pinot Gris, this is crisp with ...	NaN	90	32.0	Alsace	Alsace	NaN	Roger Voss	@vossroger	Domaine Marcel Deiss 2012 Pinot Gris (Alsace)	Pinot Gris	Domaine Marcel Deiss
129970	129970	France	Big, rich and off-dry, this is powered by inte...	Lieu-dit Harth Cuvée Caroline	90	21.0	Alsace	Alsace	NaN	Roger Voss	@vossroger	Domaine Schoffit 2012 Lieu-dit Harth Cuvée Car...	Gewürztraminer	Domaine Schoffit


Exploratory data analysis
EDA

La prima fase dello studio consiste nell'esplorare il dataset (Exploratory Data Analysis - EDA) cercando di capirne le caratteristiche

1 - Dimensione dataset

```
data.shape
```

```
(129971, 14)
```

Il dataset è costituito da 129.971 righe e 14 colonne

2 - Verifica del tipo di dati e del numero dei valori mancanti

```
data.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
```

```
print(data.isnull().sum())
```

```
Unnamed: 0                 0
country                   63
description                0
designation            37465
points                     0
price                   8996
province                  63
region_1               21247
region_2               79460
taster_name            26244
taster_twitter_handle  31213
title                      0
variety                    1
winery                     0
```

3 - Dopo aver eliminato le colonne che non mi interessano, riempito i valori nulli con il valore '0' e convertito la variabile categorica *variety* in numerica si ottiene il dataset finale

```
data.drop(columns=['Unnamed: 0','description','designation', 'region_2','taster_name','taster_twitter_handle','title','region_1'], axis=1, inplace=True)
data.fillna(0, inplace=True)
```

```
data['variety'] = data['variety'].astype(str)
data['variety_target'] = LabelEncoder().fit_transform(data['variety'])
```

```
data = data[['variety','winery','price','points','country', 'province', 'variety_target']]
data
```

	variety	winery	price	points	country	province	variety_target
0	White Blend	Nicosia	0.0	87	Italy	Sicily & Sardinia	691
1	Portuguese Red	Quinta dos Avidagos	15.0	87	Portugal	Douro	451
2	Pinot Gris	Rainstorm	14.0	87	US	Oregon	437
3	Riesling	St. Julian	13.0	87	US	Michigan	480
4	Pinot Noir	Sweet Cheeks	65.0	87	US	Oregon	441
...	...	...	...	...	...	...	...
129966	Riesling	Dr. H. Thanisch (Erben Müller-Burggraef)	28.0	90	Germany	Mosel	480
129967	Pinot Noir	Citation	75.0	90	US	Oregon	441
129968	Gewürztraminer	Domaine Gresser	30.0	90	France	Alsace	210
129969	Pinot Gris	Domaine Marcel Deiss	32.0	90	France	Alsace	437
129970	Gewürztraminer	Domaine Schoffit	21.0	90	France	Alsace	210

4 - Verifichiamo nuovamente dimensione e tipo di dati e vediamo che adesso ci sono 10 colonne e non valori nulli

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129971 entries, 0 to 129970
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   variety         129971 non-null  object 
 1   winery          129971 non-null  object 
 2   price           129971 non-null  float64
 3   points          129971 non-null  int64  
 4   country         129971 non-null  object 
 5   province        129971 non-null  object 
 6   variety_target  129971 non-null  int64  
```

```
data.shape
```

```
(129971, 7)
```

5 - Verifica di alcune misure statistiche come la media (mean) e la mediana (50%)

```
print(data.describe().round(2))
```

	price	points	variety_target
count	129971.00	129971.00	129971.00
mean	32.92	88.45	353.53
std	40.58	3.04	195.97
min	0.00	80.00	0.00
25%	15.00	86.00	126.00
50%	25.00	88.00	441.00
75%	40.00	91.00	493.00
max	3300.00	100.00	707.00

6 - Controllo della distribuzione delle variabile *varietà* per vedere che non ci sia troppo squilibrio tra i valori

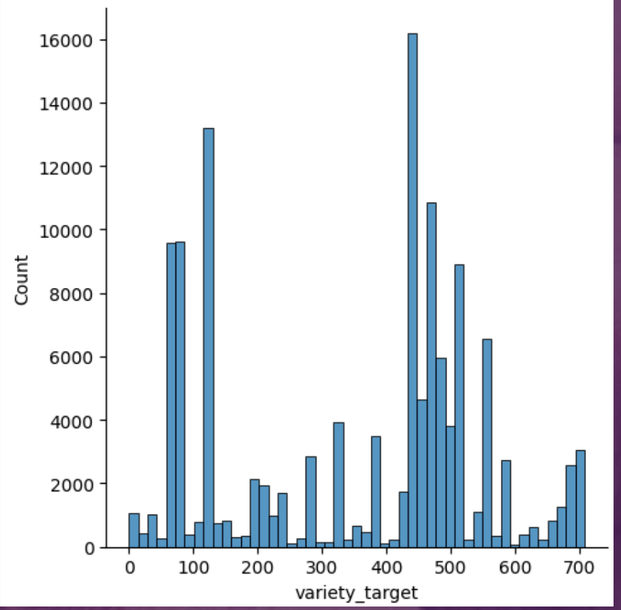
```
print(data['variety'].value_counts())
```

```
variety
Pinot Noir                    13272
Chardonnay                    11753
Cabernet Sauvignon             9472
Red Blend                      8946
Bordeaux-style Red Blend       6915
                              ...
Sauvignonasse                     1
Cabernet Sauvignon-Barbera        1
Meseguera                         1
Forcallà                          1
Bobal-Cabernet Sauvignon          1
```

e lo vediamo anche visivamente

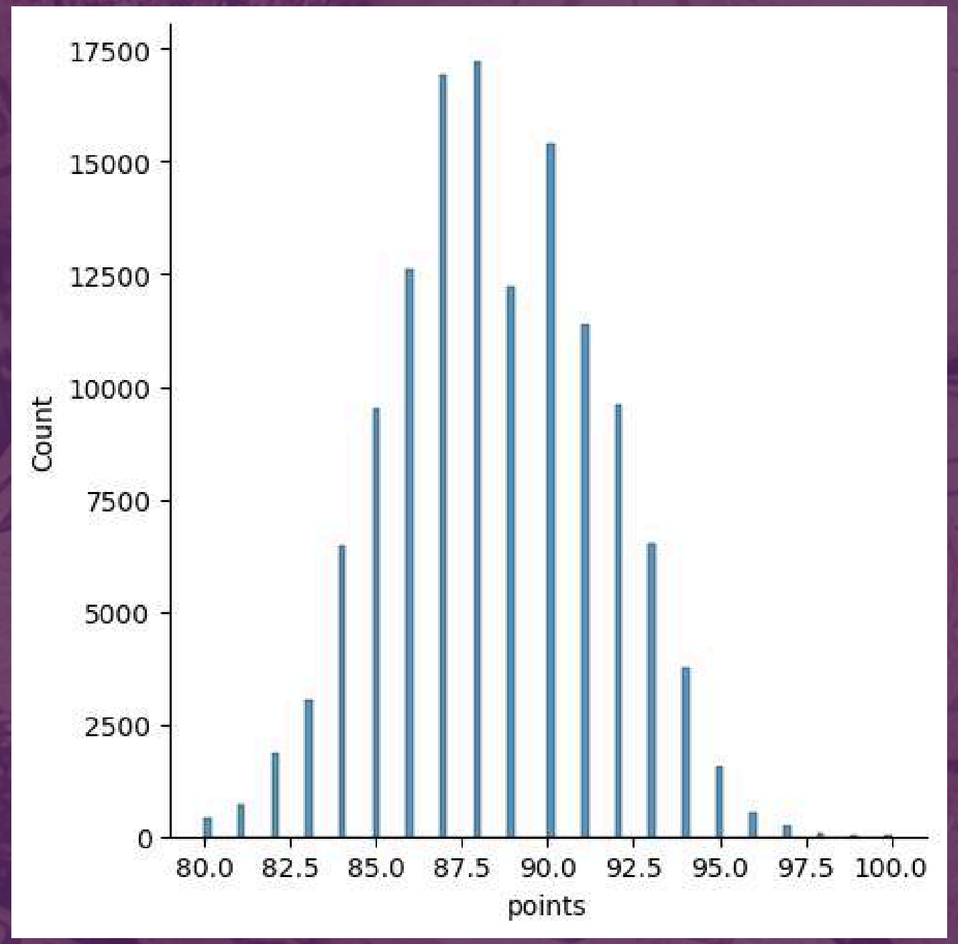
```
sns.displot(data['variety_target'])
```

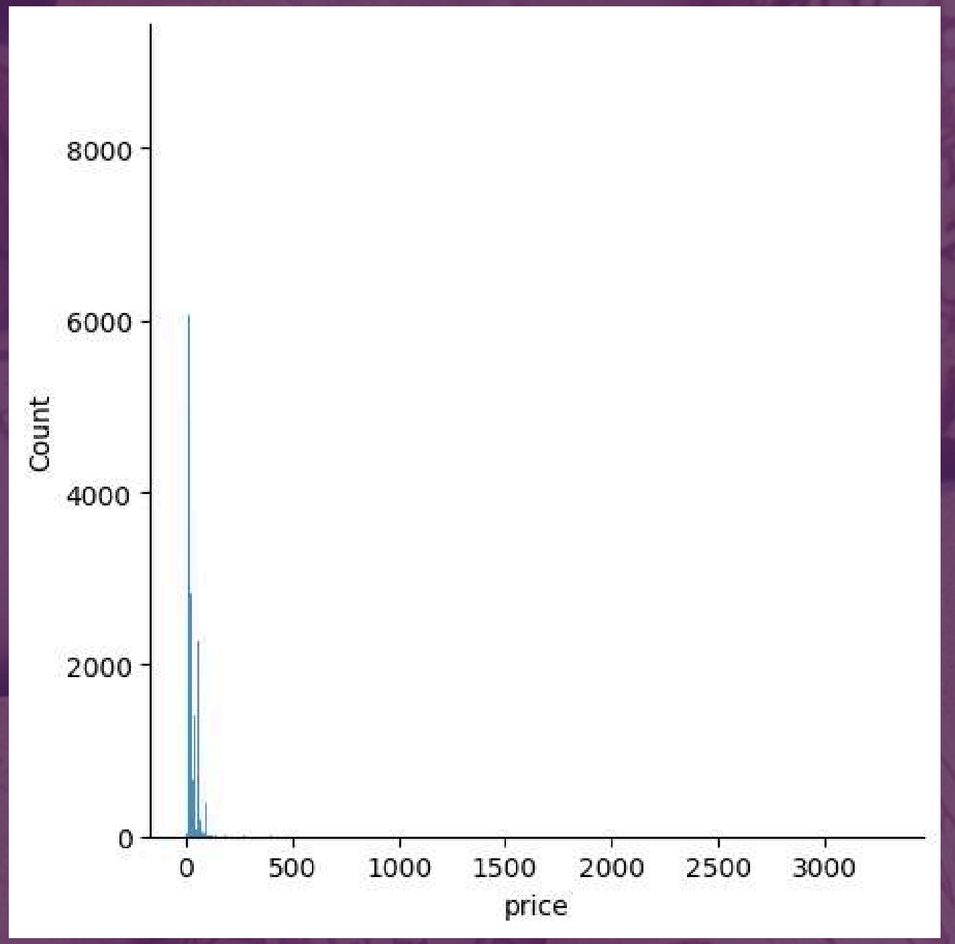
```
<seaborn.axisgrid.FacetGrid at 0x7fd3255711e0>
```



7 - Verifichiamo se i dati sono normalmente distribuiti oppure è presente qualche asimmetria (destra/sinistra)

```
values = data.loc[:, ['points', 'price']]
values
for column in values:
     sns.displot(x=column, data=values)
```

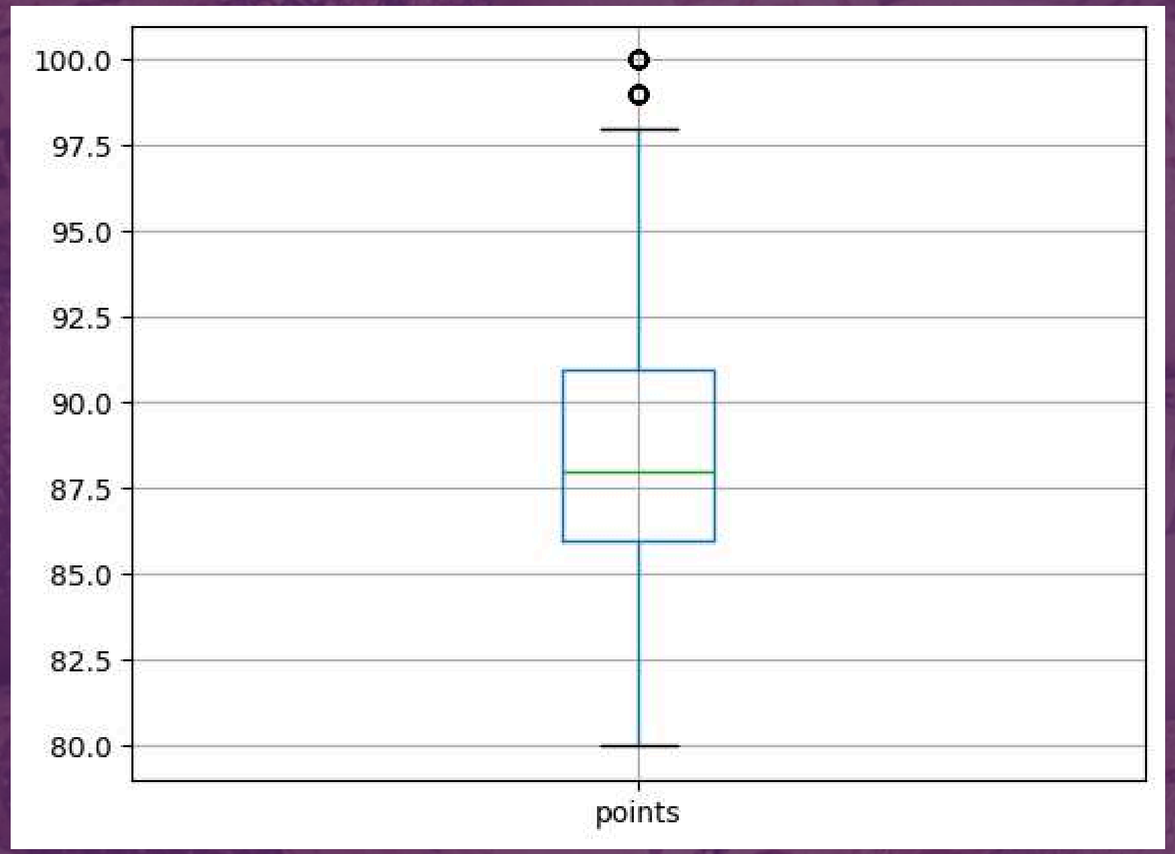


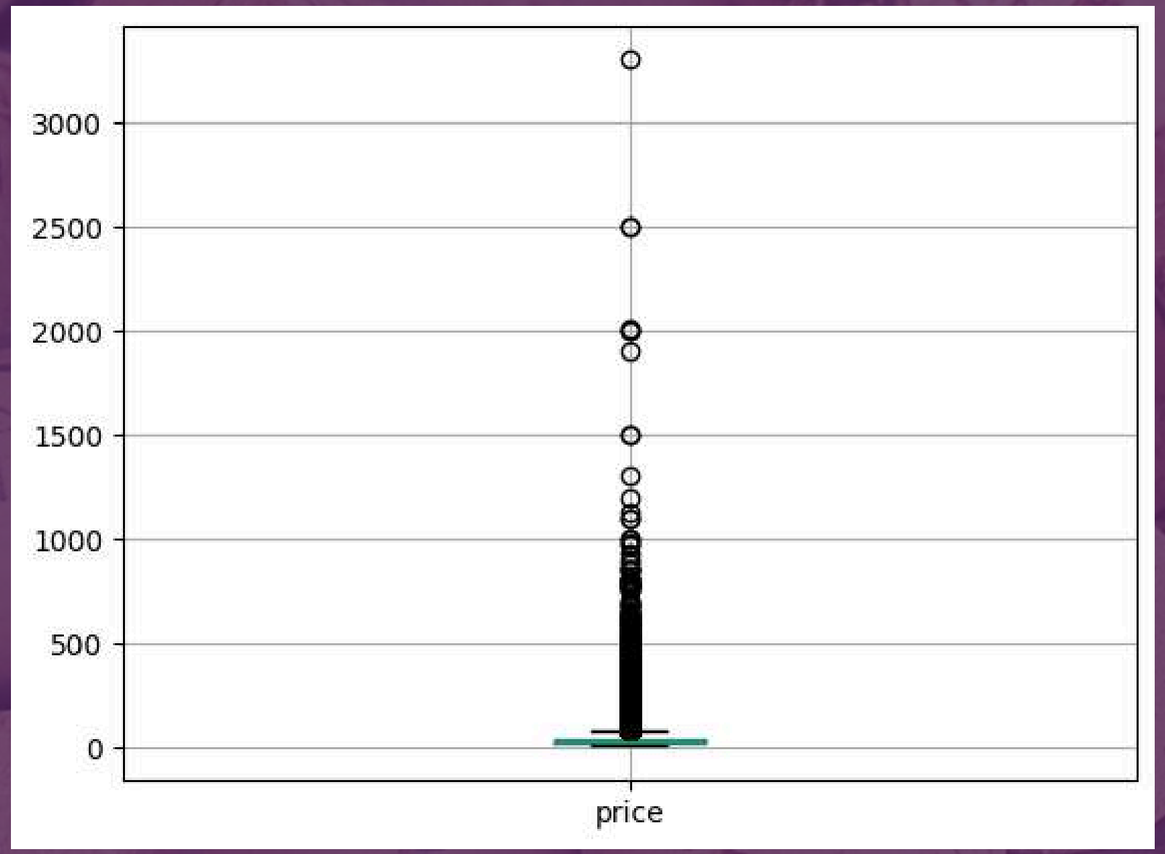


I punteggi sono pressoché normalmente distribuiti invece i prezzi non sono normalmente distribuiti ma presentano una simmetria sinistra

8 - Controllo dei valori outlier

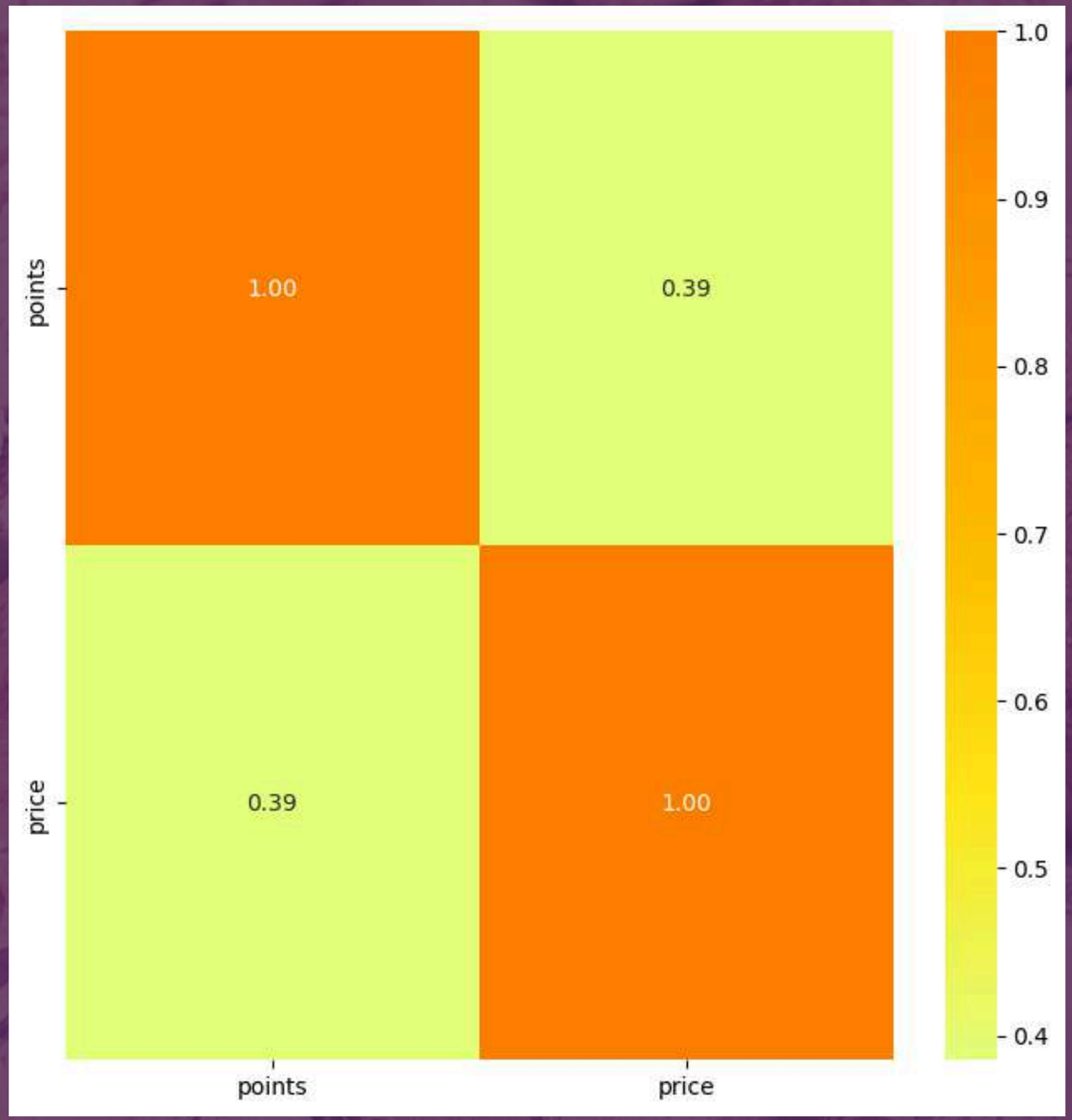
```
for column in values:
    plt.figure()
    data.boxplot([column])
```





I valori sopra il baffo superiore e sotto il baffo inferiore sono gli outlier

9 - Verifica della relazione tra prezzo e punteggio con la matrice di correlazione



Esiste un 39% di correlazione tra prezzo e punteggio

Sfoltimento dataset

+

EDA

La seconda fase dello studio consiste nell'eliminare outilier e rifare l'EDA sul dataset sfoltito

Come prima operazione eliminiamo gli outlier sulla colonna *points* utilizzando il modello Tukey

```
k = 1.5
q1 = np.percentile(data['points'], 20)
q3 = np.percentile(data['points'], 74)
iqr = q3 - q1
lower_point = q1 - k * iqr
upper_point = q3 + k * iqr

data2 = data[(data['points'] < upper_point) & (data['points'] > lower_point)]
```

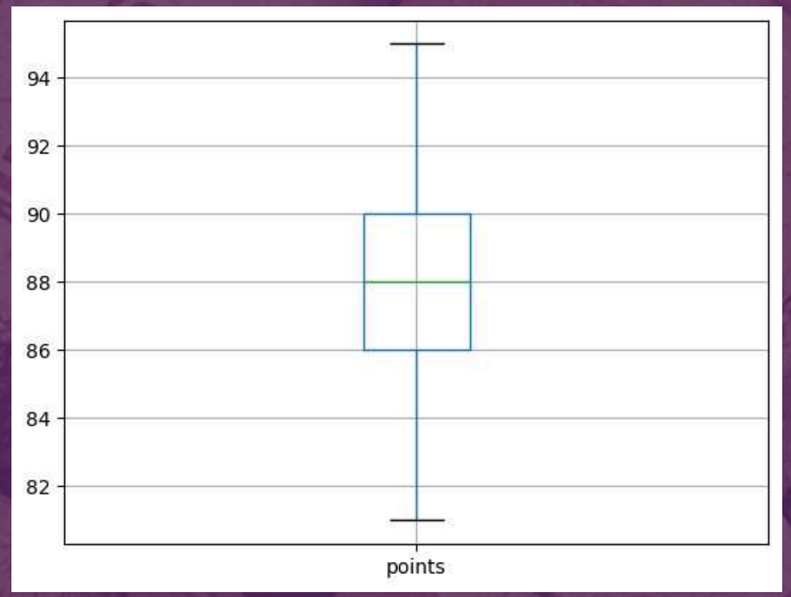
Successivamente facciamo la stessa operazione sulla colonna *price* del dataset privato degli outlier della colonna *points*

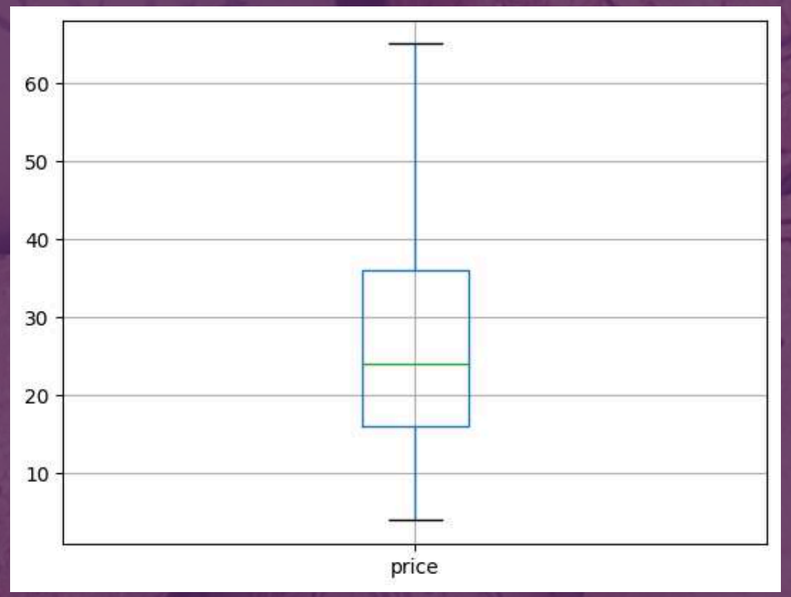
```
h = 1.5
p1 = np.percentile(data2['price'], 16.5)
p3 = np.percentile(data2['price'], 66.5)
ipr = p3 - p1
lower_price = p1 - h * ipr
upper_price = p3 + h * ipr

data3 = data2[(data2['price'] < upper_price) & (data2['price'] > lower_price) & (data2['price'] != 0)]
```

Il nuovo dataset ha 109.395 record

	variety	winery	price	points	country	province	variety_target
1	Portuguese Red	Quinta dos Avidagos	15.0	87	Portugal	Douro	451
2	Pinot Gris	Rainstorm	14.0	87	US	Oregon	437
3	Riesling	St. Julian	13.0	87	US	Michigan	480
4	Pinot Noir	Sweet Cheeks	65.0	87	US	Oregon	441
5	Tempranillo-Merlot	Tandem	15.0	87	Spain	Northern Spain	591
...	...	...	...	...	...	...	...
129965	Pinot Gris	Domaine Rieflé-Landmann	28.0	90	France	Alsace	437
129966	Riesling	Dr. H. Thanisch (Erben Müller-Burggraef)	28.0	90	Germany	Mosel	480
129968	Gewürztraminer	Domaine Gresser	30.0	90	France	Alsace	210
129969	Pinot Gris	Domaine Marcel Deiss	32.0	90	France	Alsace	437
129970	Gewürztraminer	Domaine Schoffit	21.0	90	France	Alsace	210

Graficamente vediamo che non ci sono più outlier





Per l'analisi vera e propria del dataset prendiamo in considerazione solo le varietà di vino con un numero di recensioni maggiori o uguali a 100

```
data3_count_variety = data3.groupby(['variety'])[['points']].count()
data3_count_variety1 = data3_count_variety.rename(columns={"points": "count_variety"})

# unione dataset
data3_merge = data3.set_index('variety').merge(data3_count_variety1, left_on = 'variety', right_on = 'variety', how = 'left').reset_index()

#varietà di vino con un numero di recensioni maggiori o uguali a 100
data3_merge = data3_merge[data3_merge['count_variety'] >= 100]
```

	variety	winery	price	points	country	province	variety_target	count_variety
0	Portuguese Red	Quinta dos Avidagos	15.0	87	Portugal	Douro	451	2087
1	Pinot Gris	Rainstorm	14.0	87	US	Oregon	437	1358
2	Riesling	St. Julian	13.0	87	US	Michigan	480	4626
3	Pinot Noir	Sweet Cheeks	65.0	87	US	Oregon	441	11071
6	Gewürztraminer	Trimbach	24.0	87	France	Alsace	210	930
...	...	...	...	...	...	...	...	...
109390	Pinot Gris	Domaine Rieflé-Landmann	28.0	90	France	Alsace	437	1358
109391	Riesling	Dr. H. Thanisch (Erben Müller-Burggraef)	28.0	90	Germany	Mosel	480	4626
109392	Gewürztraminer	Domaine Gresser	30.0	90	France	Alsace	210	930
109393	Pinot Gris	Domaine Marcel Deiss	32.0	90	France	Alsace	437	1358
109394	Gewürztraminer	Domaine Schoffit	21.0	90	France	Alsace	210	930

Adesso, dopo aver effettuato lo sfoltimento, andiamo a vedere come si presenta il dataset su cui faremo l'analisi vera e propria attraverso l'EDA

1 - Verifica della dimensione e del tipo di dati

```
data3_merge.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 102342 entries, 0 to 109394
Data columns (total 8 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   variety         102342 non-null  object
 1   winery          102342 non-null  object
 2   price           102342 non-null  float64
 3   points          102342 non-null  int64
 4   country         102342 non-null  object
 5   province        102342 non-null  object
 6   variety_target  102342 non-null  int64
 7   count_variety   102342 non-null  int64
```

```
data3_merge.shape
```

```
(102342, 8)
```

2 - Verifica di alcune misure statistiche come la media (mean) e la mediana (50%)

```
print(data3_merge.describe().round(2))
```

	price	points	variety_target	count_variety
count	102342.00	102342.00	102342.00	102342.00
mean	27.89	88.11	364.20	4750.80
std	14.34	2.82	193.91	3745.01
min	4.00	81.00	3.00	101.00
25%	16.00	86.00	126.00	1358.00
50%	25.00	88.00	441.00	3725.00
75%	38.00	90.00	493.00	7636.00
max	65.00	95.00	705.00	11071.00

Media e mediana della colonna *price* sono molto simili quindi successivamente possiamo prendere l'una o l'altra indistintamente per effettuare l'analisi vera e propria. Lo stesso discorso possiamo farlo per la media e la mediana della colonna *points*.

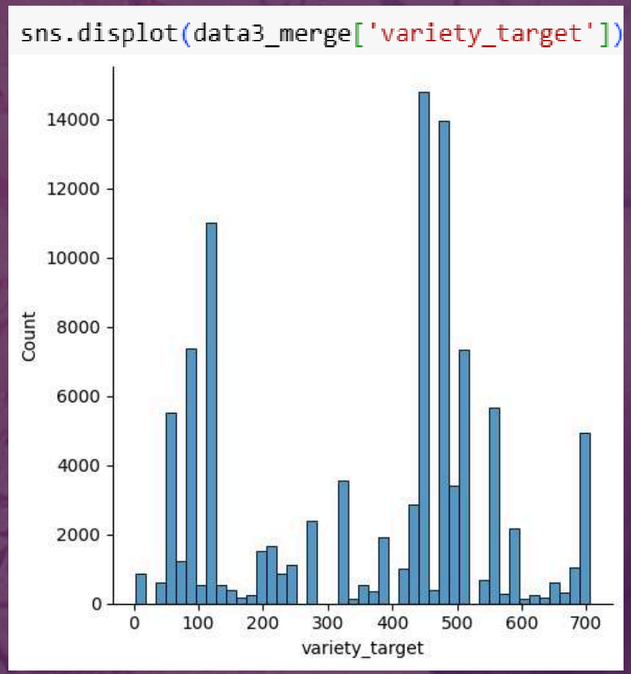
3 - Controllo della distribuzione delle variabile *varietà* per vedere che non ci sia troppo squilibrio tra i valori

```
print(data3_merge['variety'].value_counts())
```

```
variety
Pinot Noir                    11071
Chardonnay                    10200
Red Blend                      7636
Cabernet Sauvignon             7262
Sauvignon Blanc                4741
                               ...
Grillo                          109
Cabernet Sauvignon-Merlot       107
Bonarda                         103
Viura                           101
Nerello Mascalese               101
```

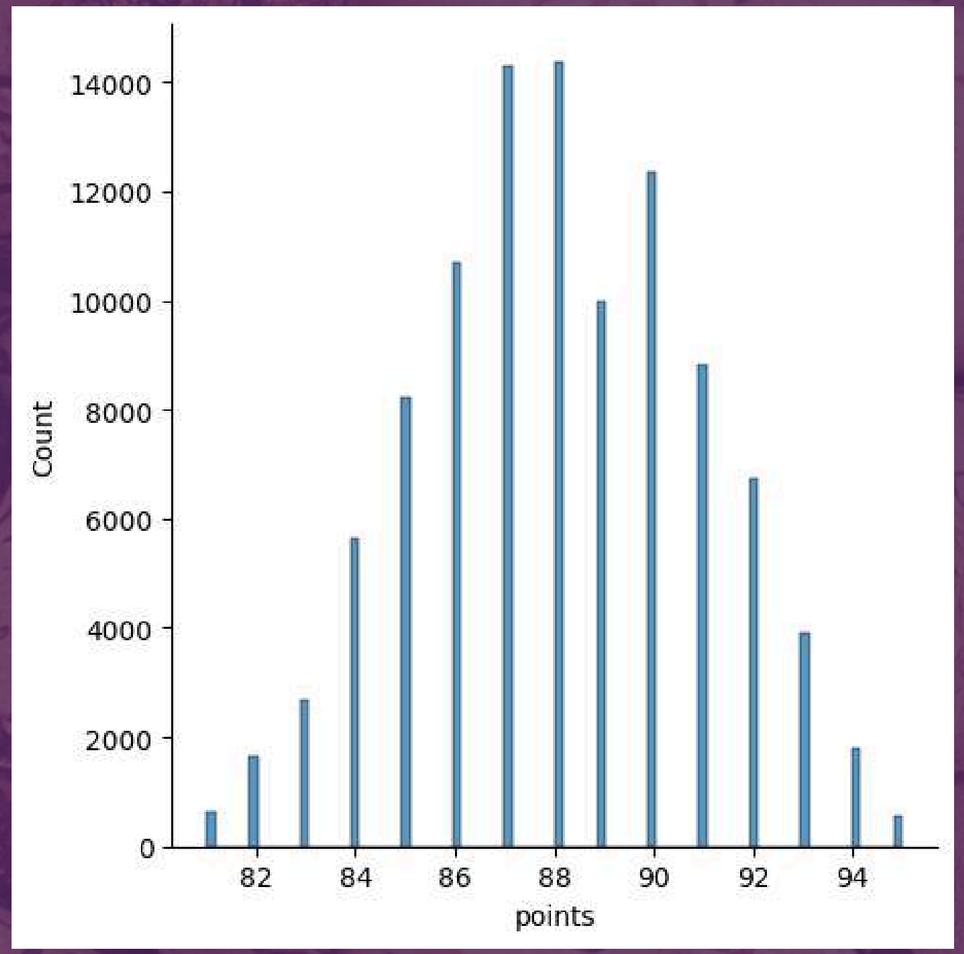
e lo vediamo anche visivamente

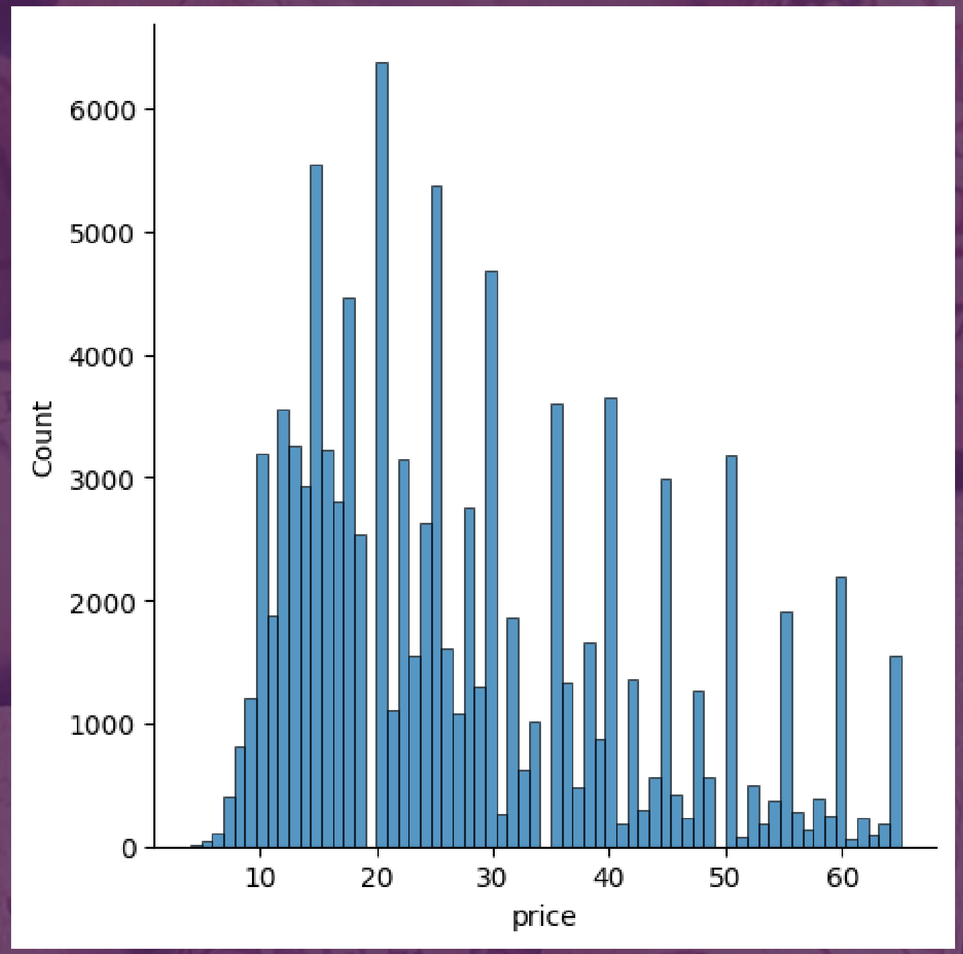
```
sns.displot(data3_merge['variety_target'])
```



4 - Verifica se i dati sono normalmente distribuiti oppure è presente qualche asimmetria (destra/sinistra)

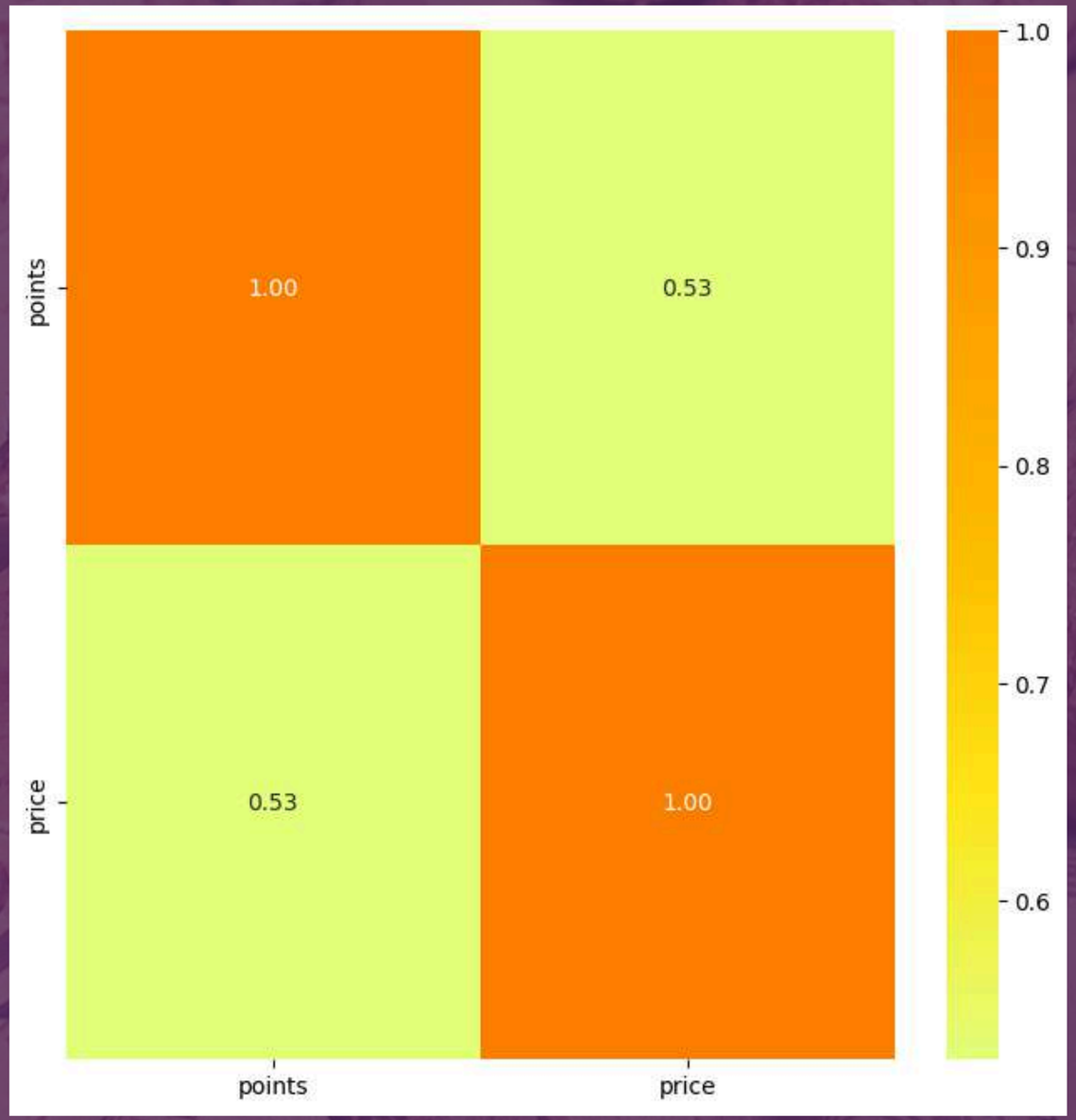
```
values2 = data3_merge.loc[:, ['points', 'price']]

for column in values2:
   sns.displot(x=column, data=values2)
```





I punteggi sono pressoché normalmente distribuiti invece i prezzi non sono normalmente distribuiti ma presentano una simmetria sinistra

4 - Verifica della relazione tra prezzo e punteggio con la matrice di correlazione



Esiste un 53% di correlazione tra prezzo e punteggio - è aumentata rispetto al dataset di partenza

Analisi dataset

La terza fase dello studio consiste nell'analisi vera e propria dei dati. Iniziamo con le varietà con prezzo più alto e più basso e con le varietà con prezzo medio più alto e più basso

Varietà con prezzo più alto

```
max_price = data3_merge[data3_merge['price'] == data3_merge['price'].max()]
print("Varietà con prezzo più alto: \n", max_price[['variety','price']])
```

```
Varietà con prezzo più alto:
                          variety  price
3                      Pinot Noir   65.0
383                    Pinot Noir   65.0
387     Bordeaux-style Red Blend    65.0
418                     Red Blend   65.0
429                    Pinot Noir   65.0
...                              ...   ...
109136                    Malbec    65.0
109140                    Malbec    65.0
109185         Cabernet Sauvignon   65.0
109259                 Pinot Noir   65.0
109263                 Chardonnay   65.0

[1554 rows x 2 columns]
```

Varietà con prezzo più basso

```
min_price = data3_merge[data3_merge['price'] == data3_merge['price'].min()]
print("\nVarietà con prezzo più basso: \n", min_price[['variety','price']])
```

```
Varietà con prezzo più basso:
                          variety  price
1663                        Syrah    4.0
17241               White Blend      4.0
26583                Chardonnay      4.0
49997                Chardonnay      4.0
51897    Cabernet Sauvignon         4.0
54288                    Merlot     4.0
92872                    Merlot     4.0
95152               Tempranillo     4.0
98816               White Blend     4.0
106153            Pinot Grigio      4.0
```

Varietà con prezzo medio più alto

```
max_price_mean = data3_merge.groupby(['variety'])['price'].mean().round(2).sort_values(ascending = False).reset_index()
print("\nVarietà con prezzo medio più alto: \n", max_price_mean.iloc[0])
```

```
Varietà con prezzo medio più alto:
 variety    Sangiovese Grosso
price                   46.37
```

Varietà con prezzo medio più basso

```
min_price_mean = data3_merge.groupby(['variety'])['price'].mean().round(2).sort_values(ascending = True).reset_index()
print("\nVarietà con prezzo medio più basso: \n", min_price_mean.iloc[0])
```

```
Varietà con prezzo medio più basso:
 variety    Torrontés
price           13.83
```

Proseguiamo con le varietà con punteggio più alto e più basso e con le varietà con punteggio medio più alto e più basso

Varietà con punteggio più alto

```
max_point = data3_merge[data3_merge['points'] == data3_merge['points'].max()]
print("Varietà con valutazione più alta: \n", max_point[['variety','points']])
```

```
Varietà con valutazione più alta:
            variety  points
295         Nebbiolo      95
296      Pinot Noir       95
297          Shiraz       95
1310     Pinot Noir       95
1311           Syrah      95
...              ...     ...
106278    Red Blend       95
106280     Riesling       95
106281       Merlot       95
107397   Pinot Noir       95
108319   Pinot Noir       95

[545 rows x 2 columns]
```

Varietà con punteggio più basso

```
min_point = data3_merge[data3_merge['points'] == data3_merge['points'].min()]
print("\nVarietà con valutazione più bassa: \n", min_point[['variety','points']])
```

```
Varietà con valutazione più bassa:
                variety  points
294          Pinot Noir      81
3098          Red Blend      81
3100             Merlot      81
3101               Rosé      81
3102         Pinot Noir      81
...                  ...     ...
105888       Chardonnay      81
105889       Chardonnay      81
107990        Carmenère      81
107991    Pinot Grigio       81
107992       Chardonnay      81

[633 rows x 2 columns]
```

Varietà con punteggio medio più alto

```
max_point_mean = data3_merge.groupby(['variety'])['points'].mean().round(2).sort_values(ascending = False).reset_index()
print("\nVarietà con valutazione media più alta: \n", max_point_mean.iloc[0])
```

```
Varietà con valutazione media più alta:
 variety    Sangiovese Grosso
points                  90.13
```

Varietà con punteggio medio più basso

```
min_point_mean = data3_merge.groupby(['variety'])['points'].mean().round(2).sort_values(ascending = True).reset_index()
print("\nVarietà con valutazione media più bassa: \n", min_point_mean.iloc[0])
```

```
Varietà con valutazione media più bassa:
 variety    Viura
points      85.45
```

Adesso andiamo a vedere per ogni nazione il prezzo medio e il punteggio medio delle varietà e li visualizziamo su un grafico

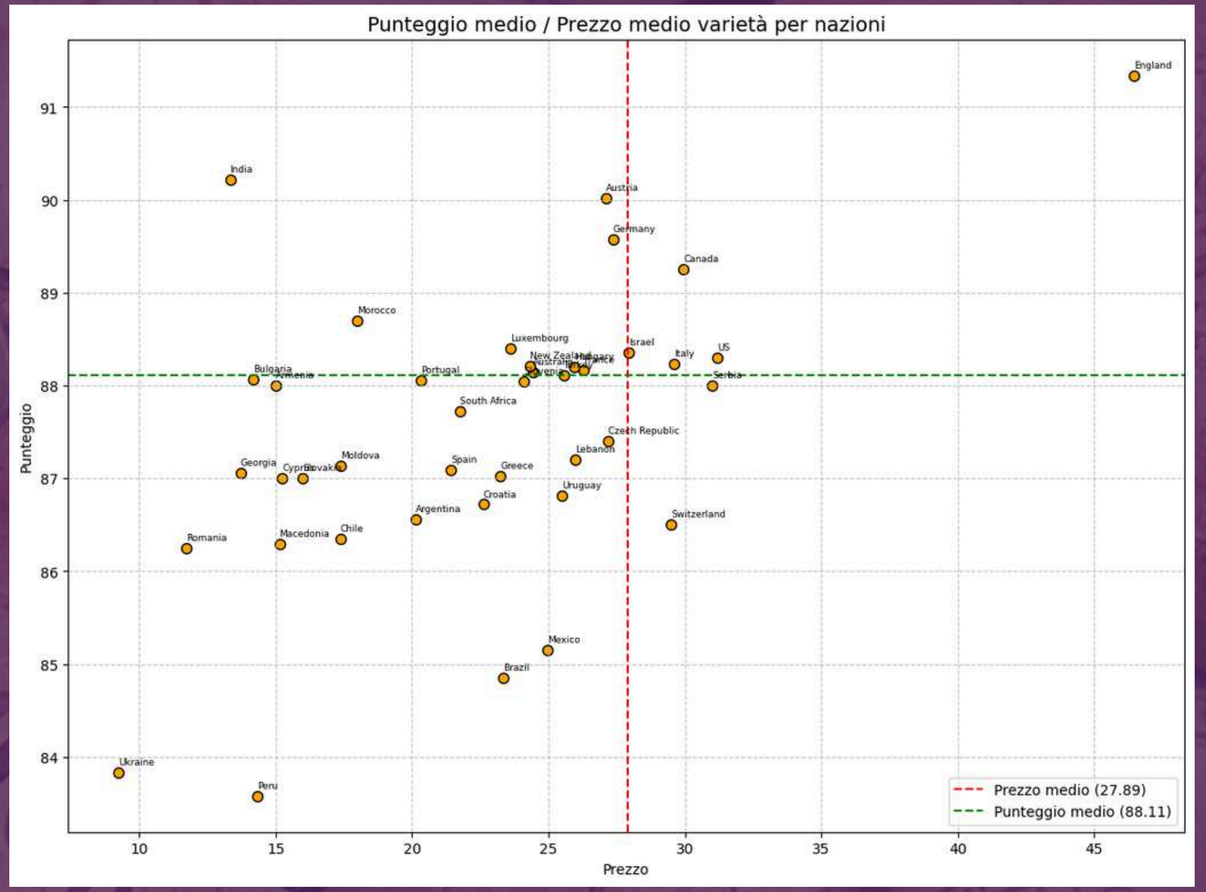
```
country_mean = data3_merge[data3_merge['country'] != 0].groupby(['country']).mean(['points', 'price']).round(2)[['price','points']]
```

```
country_mean_points = country_mean.sort_values('points', ascending=False)
mean_po = data3_merge['points'].mean().round(2)
print(f'Punteggio medio ({mean_po})')
print("Nazioni per punteggio medio\n", country_mean_points['points'])
```

```
country_mean_price = country_mean.sort_values('price', ascending=False)
mean_pr = data3_merge['price'].mean().round(2)
print(f'\nPrezzo medio ({mean_pr})')
print("Nazioni per prezzo medio\n", country_mean_price['price'])
```

Punteggio medio (88.11)
Nazioni per punteggio medio
country

country	
England	91.34
India	90.22
Austria	90.02
Germany	89.57
Canada	89.25
Morocco	88.70
Luxembourg	88.40
Israel	88.35
US	88.30
Italy	88.23
New Zealand	88.21
Hungary	88.20
France	88.17
Australia	88.14
Turkey	88.11
Bulgaria	88.07
Portugal	88.06
Slovenia	88.04
Serbia	88.00
Armenia	88.00
South Africa	87.72
Czech Republic	87.40
Lebanon	87.20
Moldova	87.14
Spain	87.09
Georgia	87.06
Greece	87.03
Cyprus	87.00
Slovakia	87.00
Uruguay	86.81
Croatia	86.72
Argentina	86.56
Switzerland	86.50
Chile	86.35
Macedonia	86.29
Romania	86.25
Mexico	85.15
Brazil	84.85
Ukraine	83.83
Peru	83.58



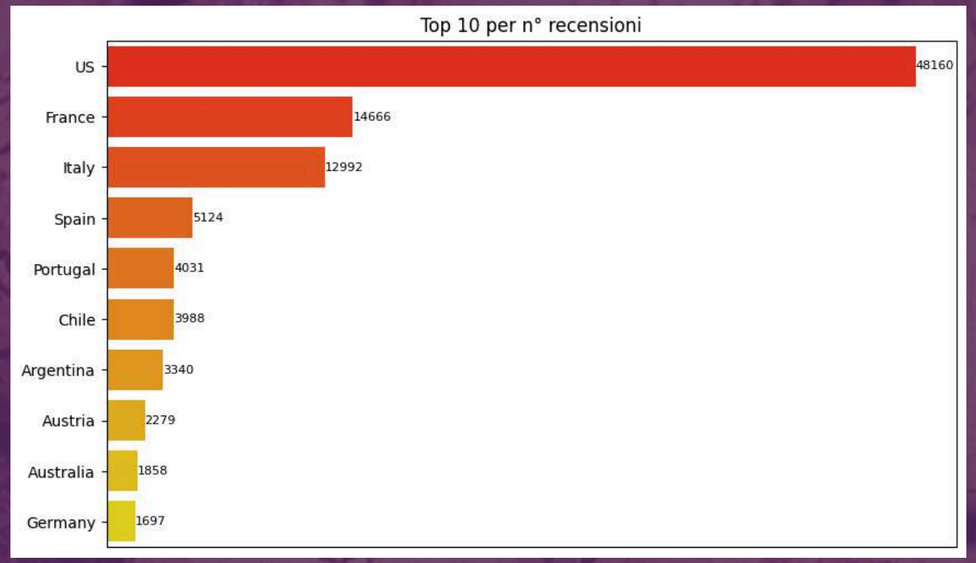
Prezzo medio (27.89)
Nazioni per prezzo medio
country

country	
England	46.45
US	31.19
Serbia	31.00
Canada	29.96
Italy	29.62
Switzerland	29.50
Israel	27.94
Germany	27.36
Czech Republic	27.20
Austria	27.10
France	26.30
Lebanon	25.97
Hungary	25.94
Turkey	25.55
Uruguay	25.49
Mexico	24.95
Australia	24.43
New Zealand	24.34
Slovenia	24.11
Luxembourg	23.60
Brazil	23.35
Greece	23.24
Croatia	22.61
South Africa	21.74
Spain	21.42
Portugal	20.32
Argentina	20.12
Morocco	18.00
Chile	17.38
Moldova	17.37
Slovakia	16.00
Cyprus	15.25
Macedonia	15.14
Armenia	15.00
Peru	14.33
Bulgaria	14.19
Georgia	13.71
India	13.33
Romania	11.73
Ukraine	9.25

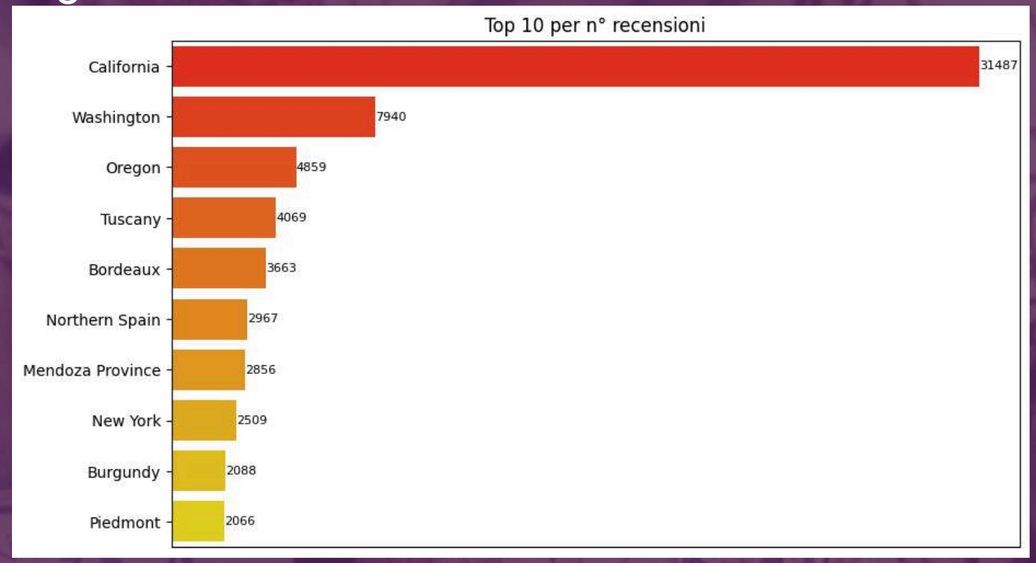
Sul grafico la linea tratteggiata verde rappresenta la media dei punteggi di tutte le varietà, mentre la linea tratteggiata rossa rappresenta la media dei prezzi di tutte le varietà. Possiamo vedere che le nazioni che sono sopra la media in termini di punteggio e prezzi sono Inghilterra (la migliore), Stati Uniti, Italia, Canada e Israele. Dall'altro canto, le nazioni che più di tutte sono sotto la media sono Ucraina e Perù.

Passiamo adesso ad analizzare il **numero di recensioni** per nazione, provincia, cantina e varietà visualizzando i più recensiti

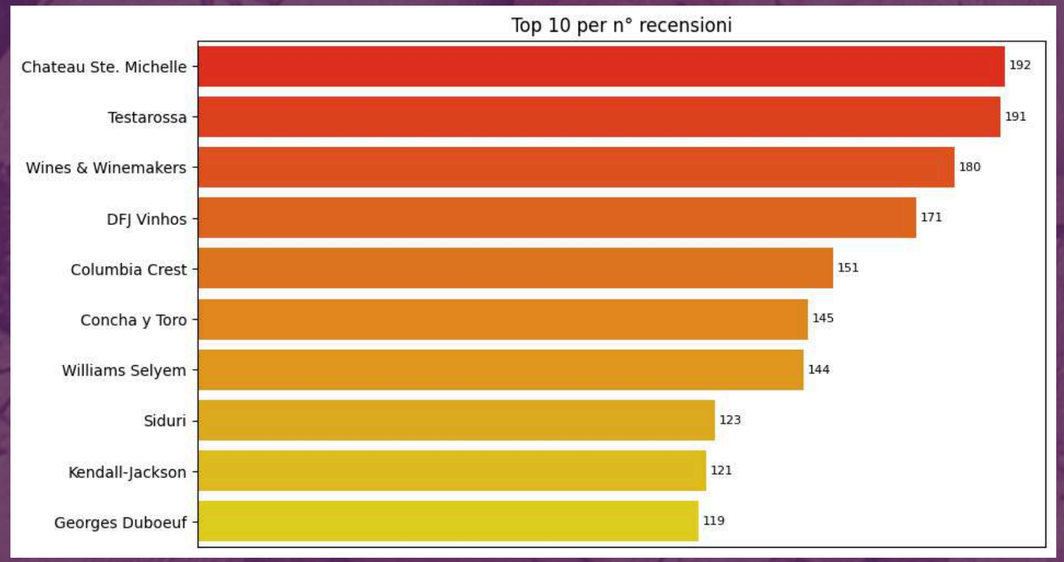
Nazioni



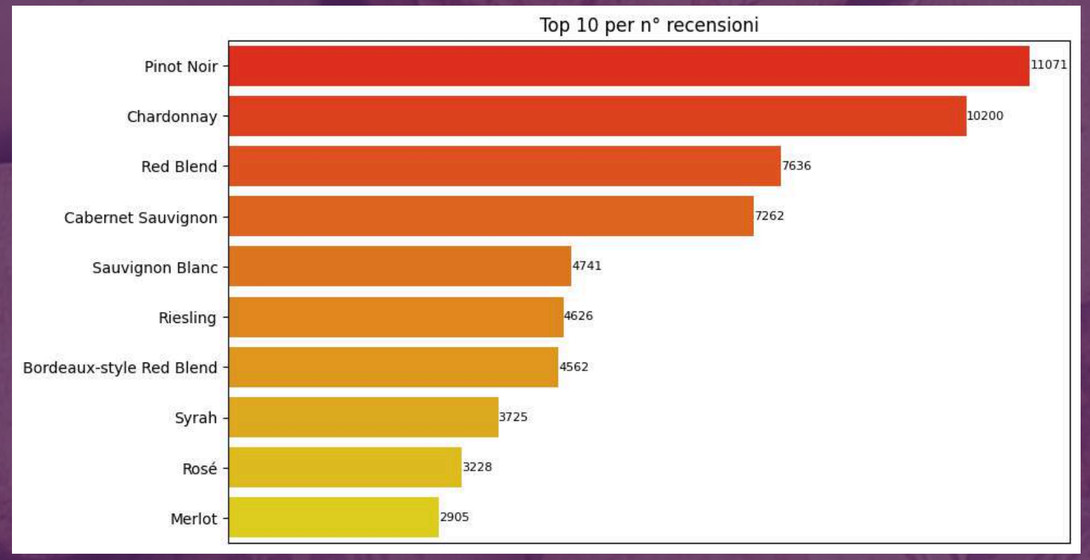
Regioni



Cantine

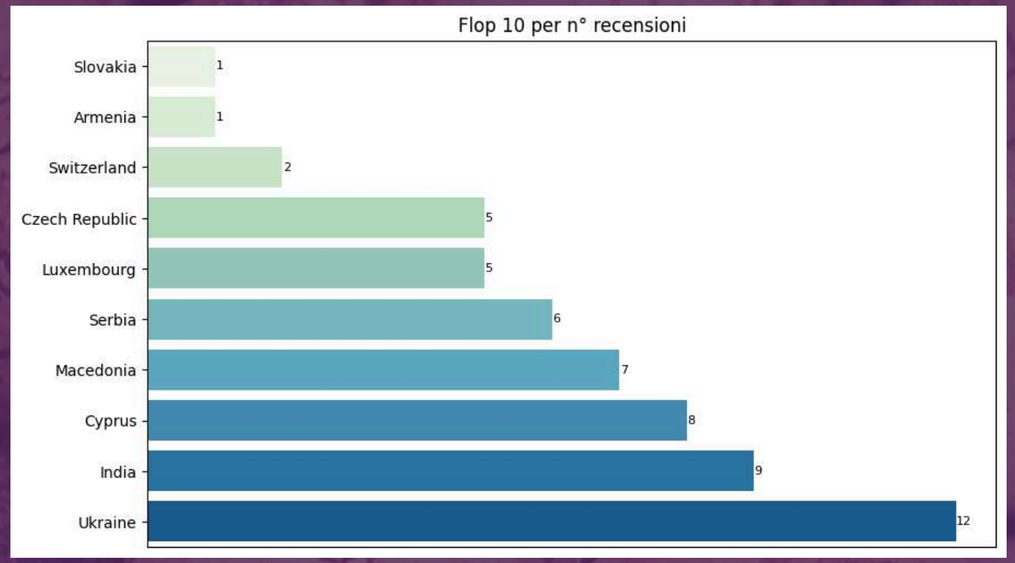


Varietà

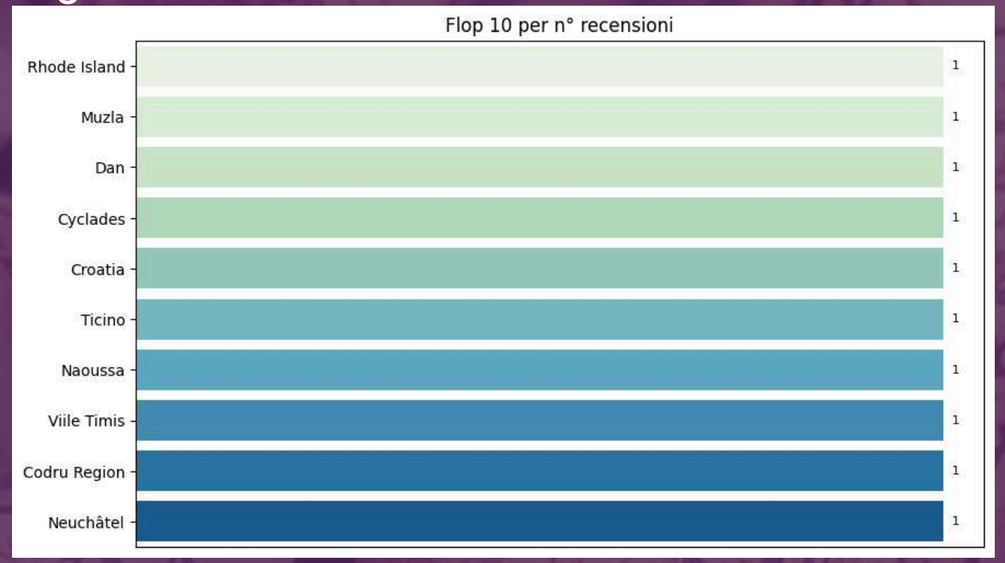


Passiamo adesso ad analizzare il **numero di recensioni** per nazione, provincia, cantina e varietà visualizzando i meno recensiti

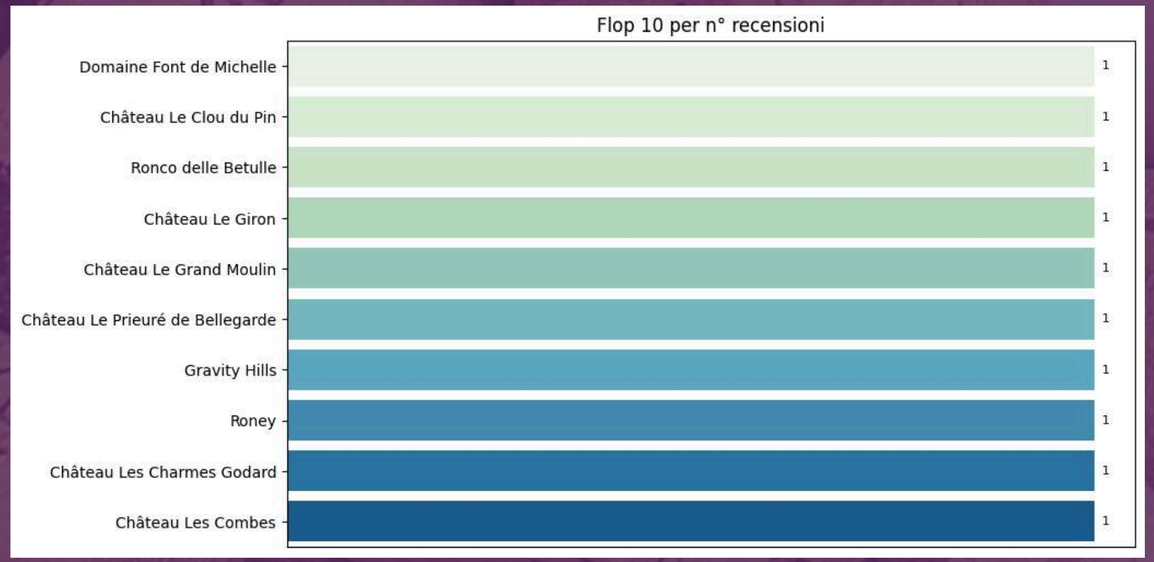
Nazioni



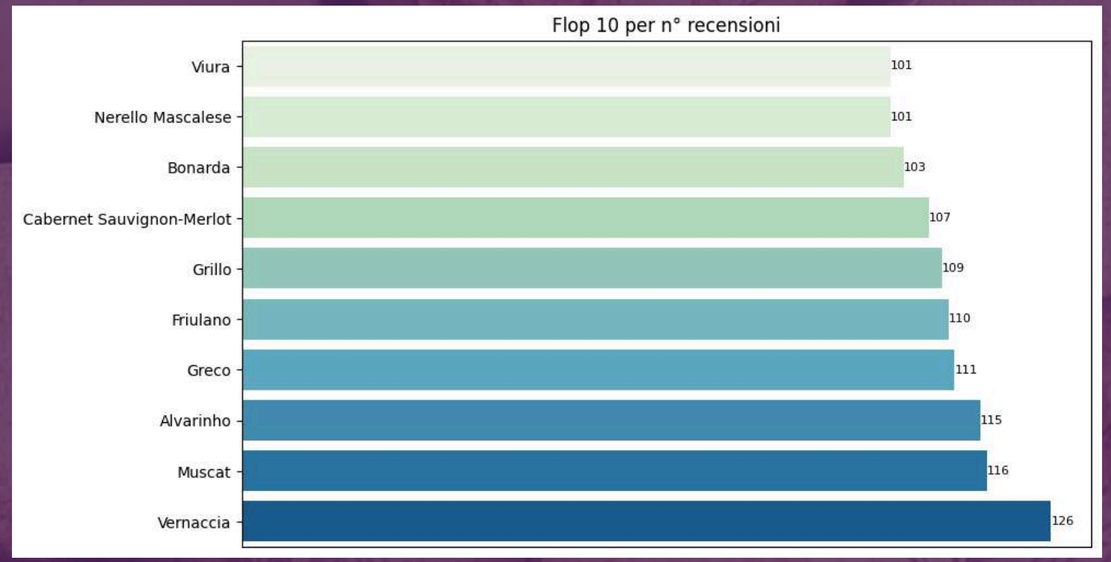
Regioni



Cantine

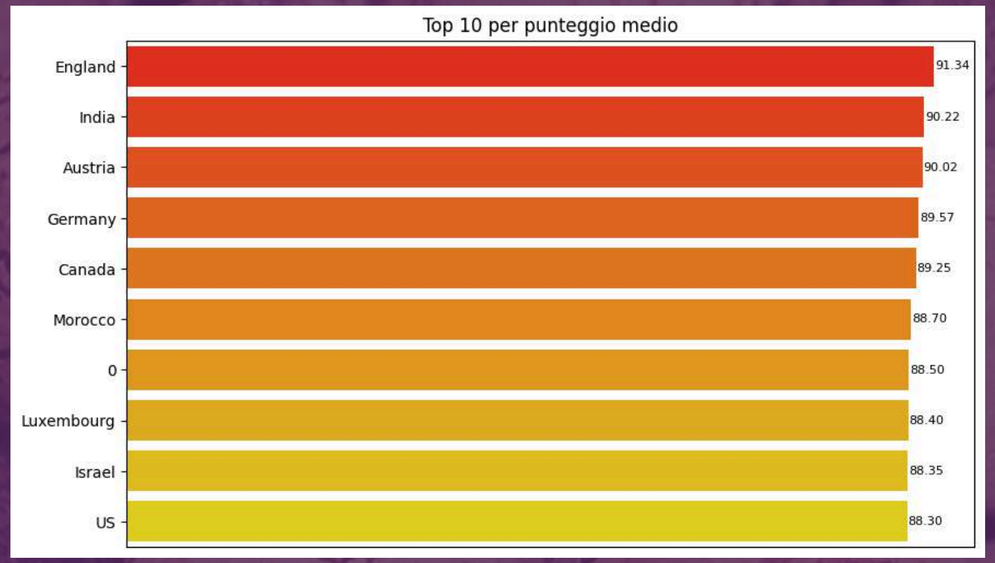


Varietà

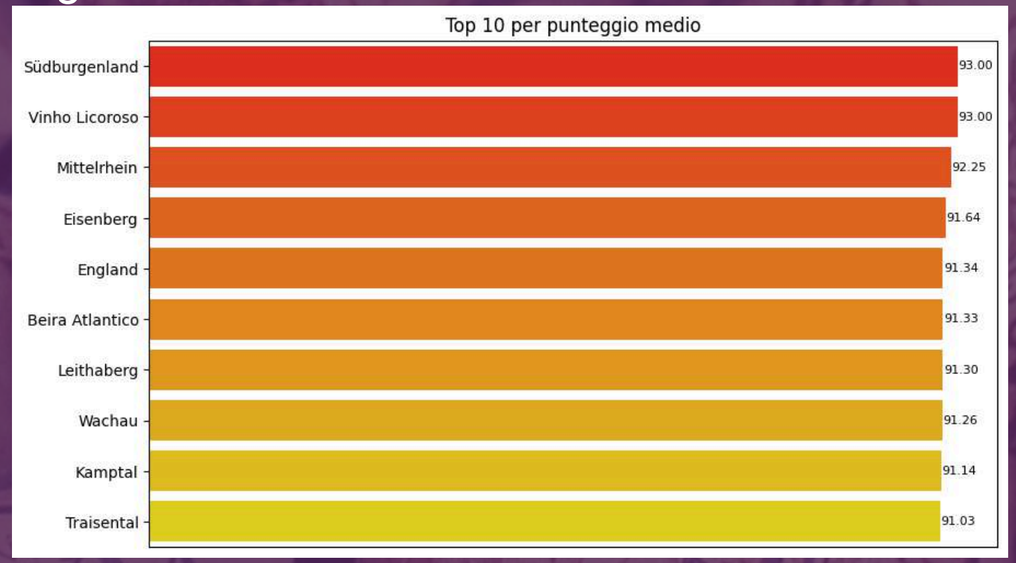


Passiamo adesso ad analizzare il **punteggio medio** per nazione, provincia, cantina e varietà visualizzando dapprima la top 10

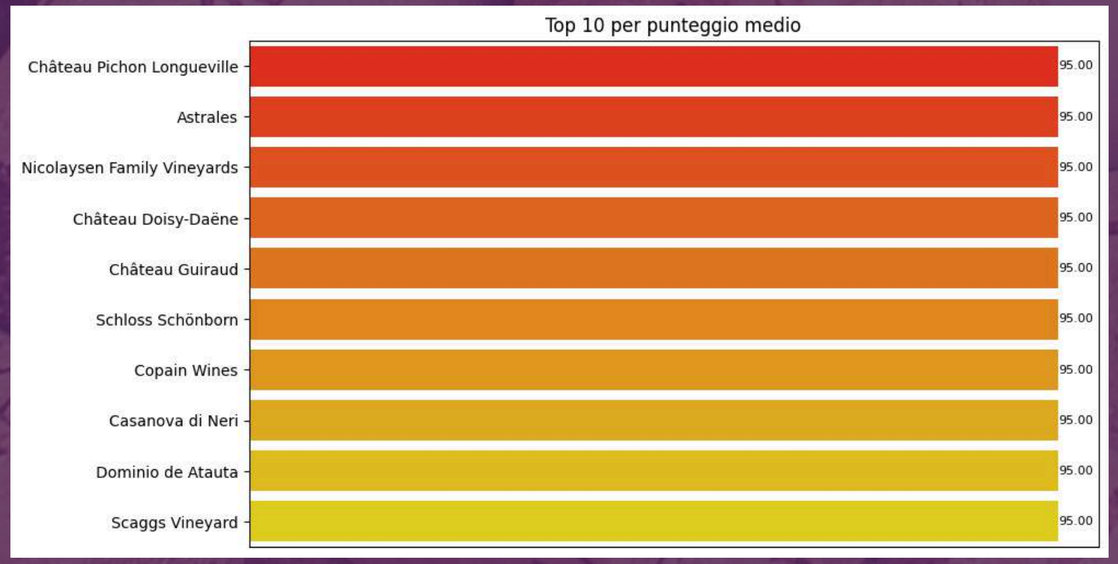
Nazioni



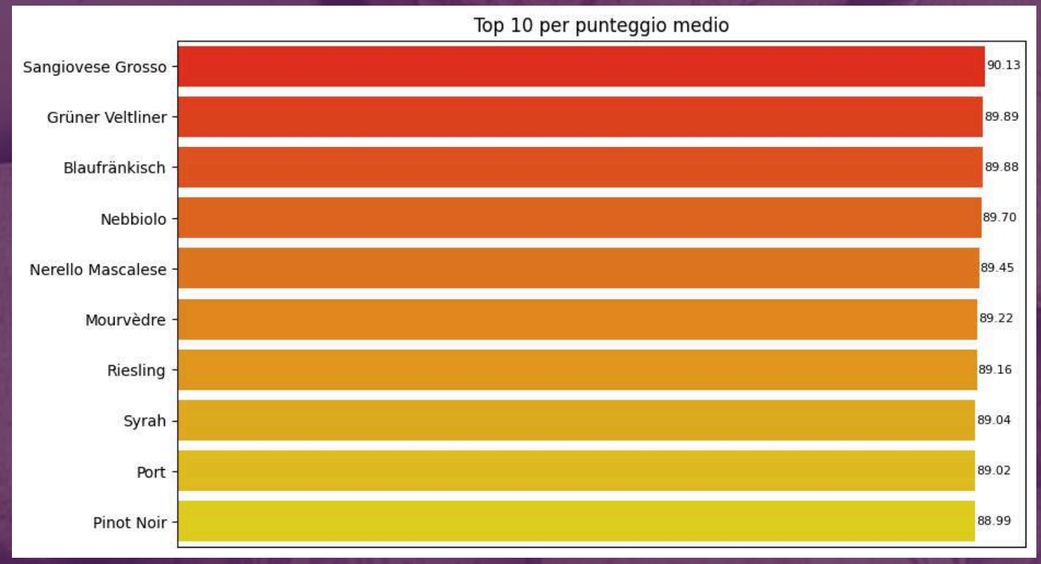
Regioni



Cantine

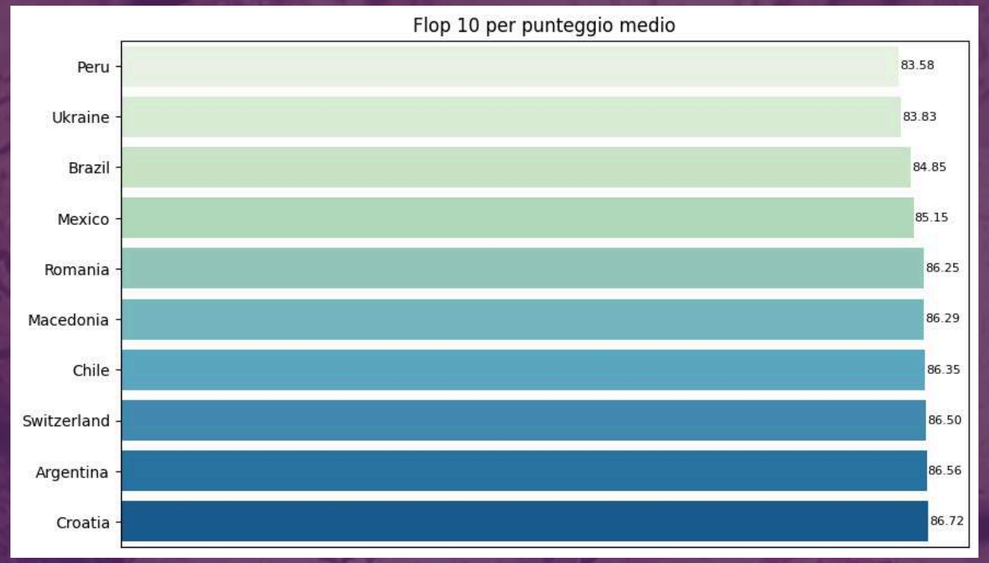


Varietà

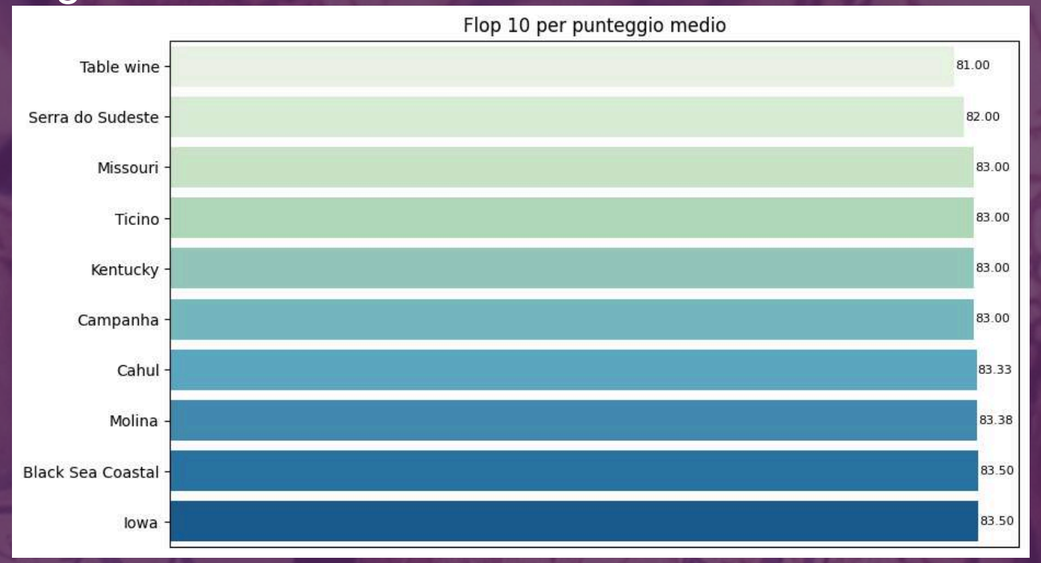


Passiamo adesso ad analizzare il **punteggio medio** per nazione, provincia, cantina e varietà visualizzando infine la flop 10

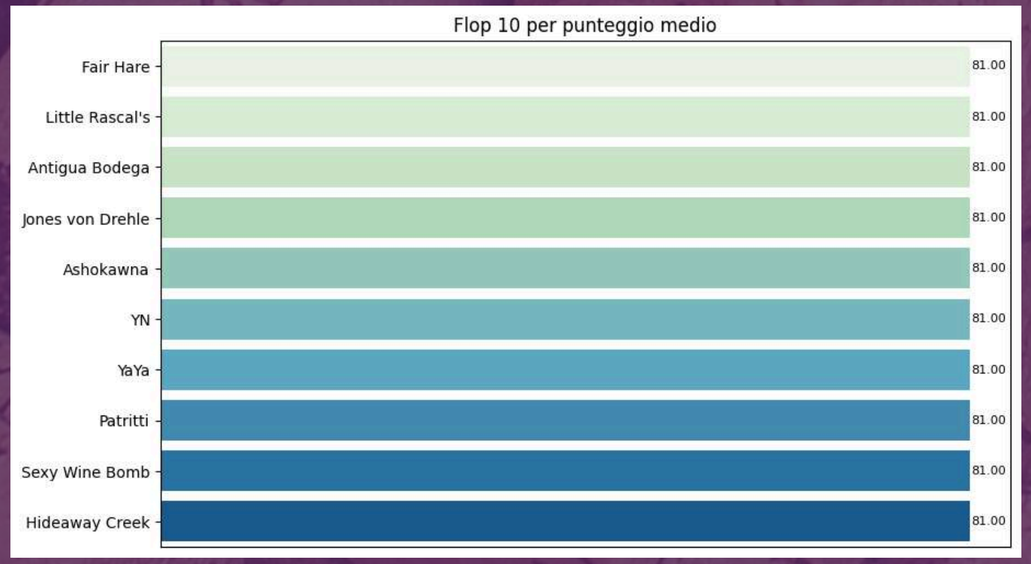
Nazioni



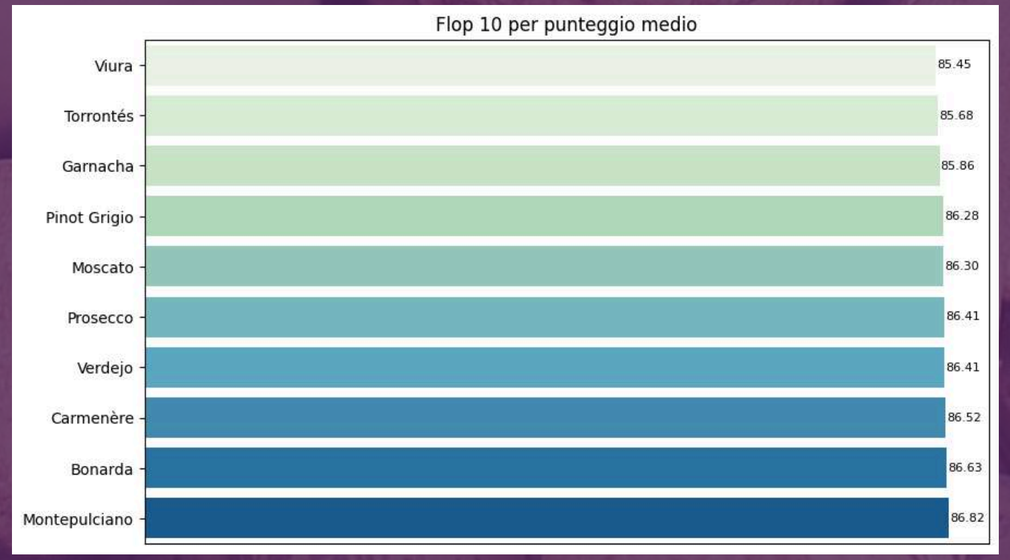
Regioni



Cantine

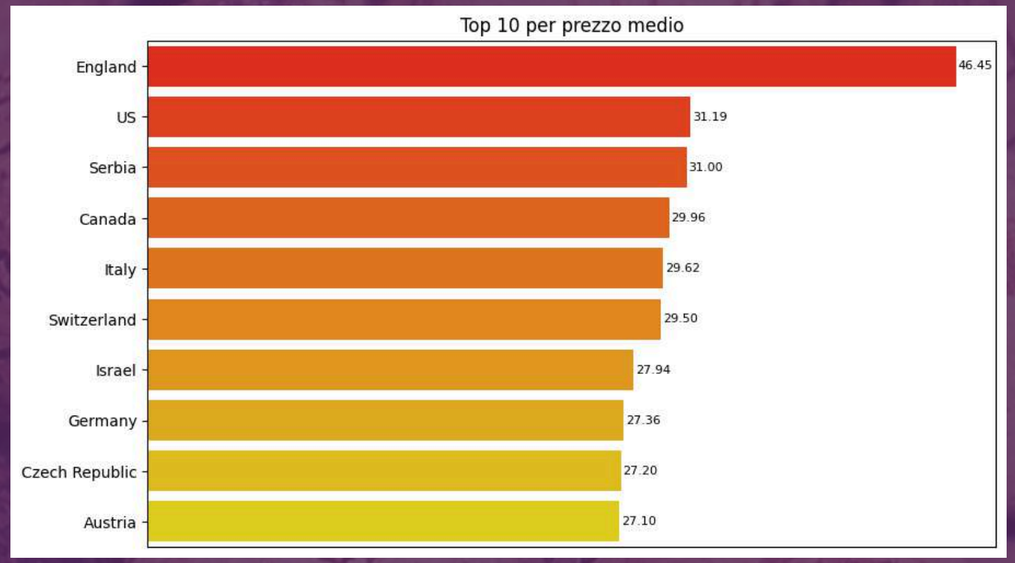


Varietà

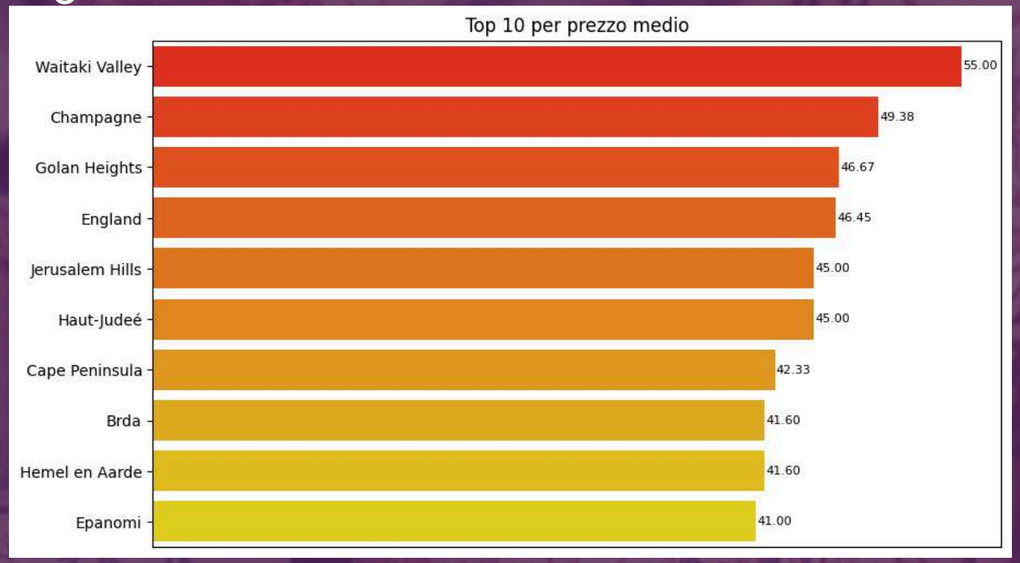


Passiamo adesso ad analizzare il **prezzo medio** per nazione, provincia, cantina e varietà visualizzando dapprima la top 10

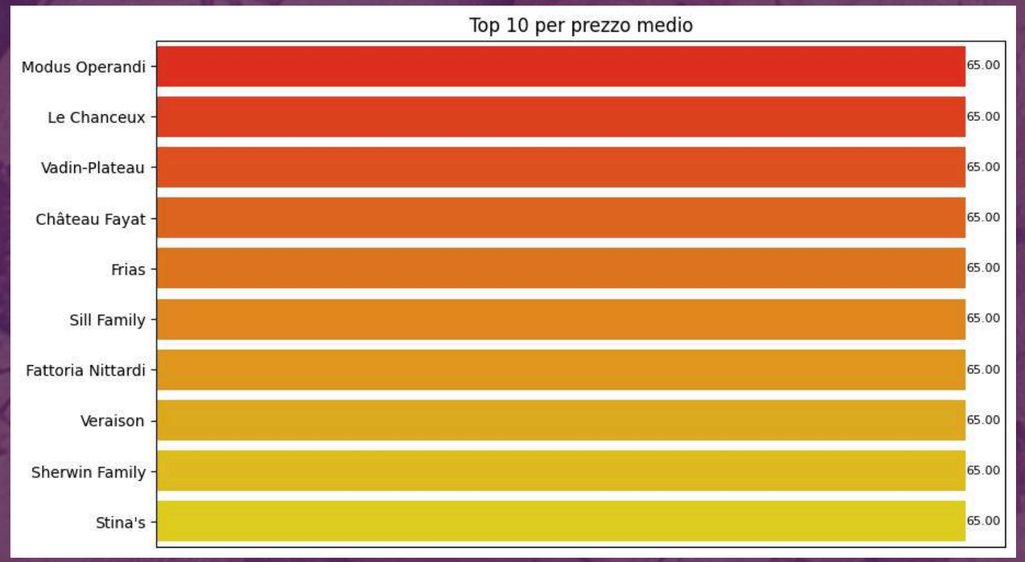
Nazioni



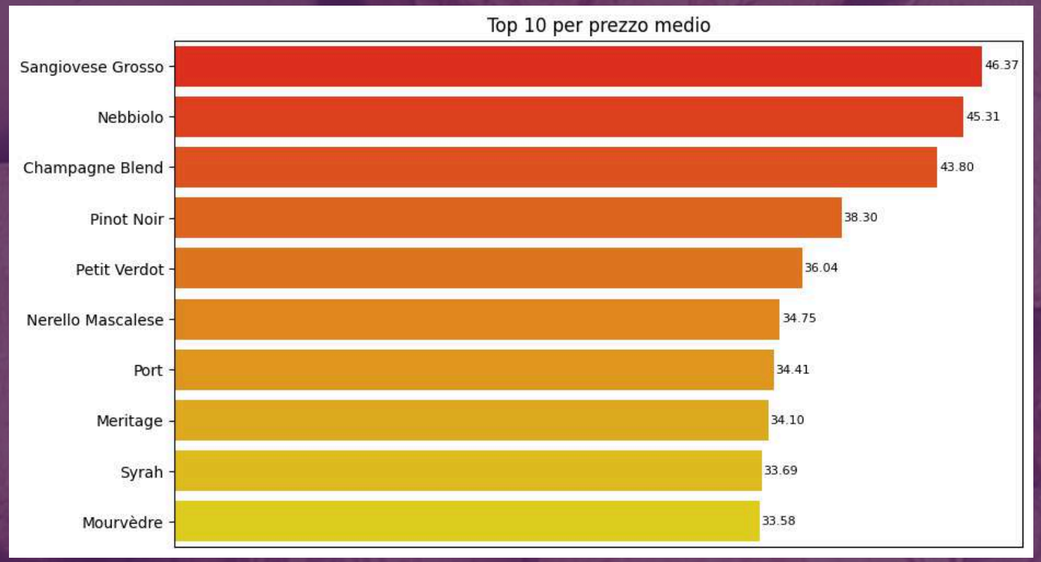
Regioni



Cantine

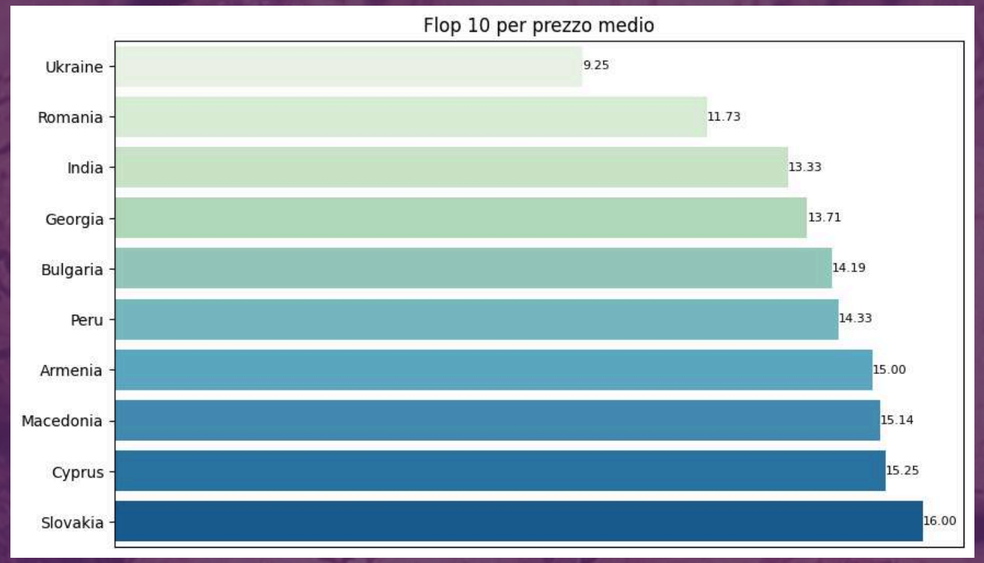


Varietà

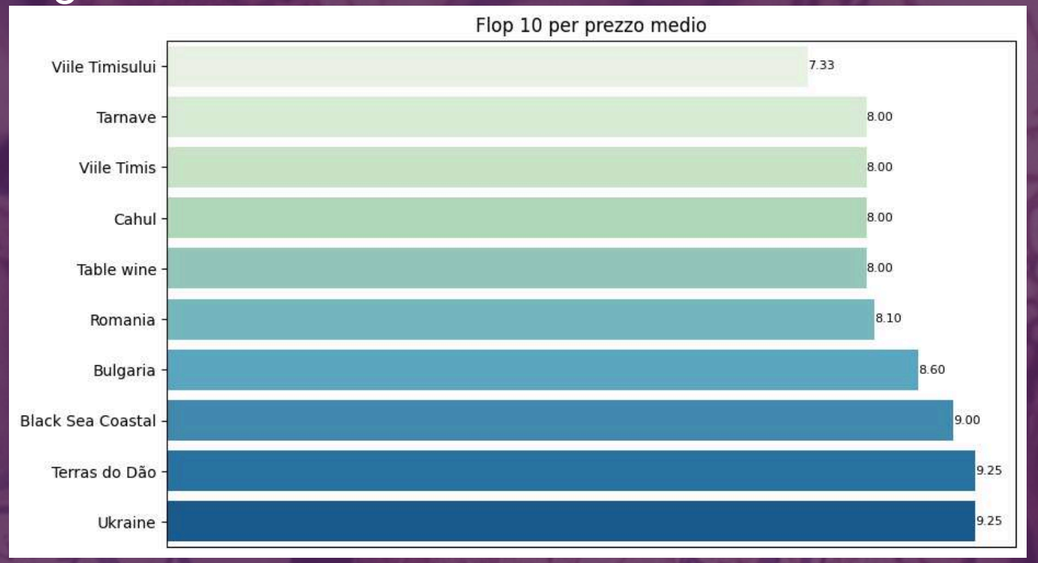


Passiamo adesso ad analizzare il **prezzo medio** per nazione, provincia, cantina e varietà visualizzando infine la flop 10

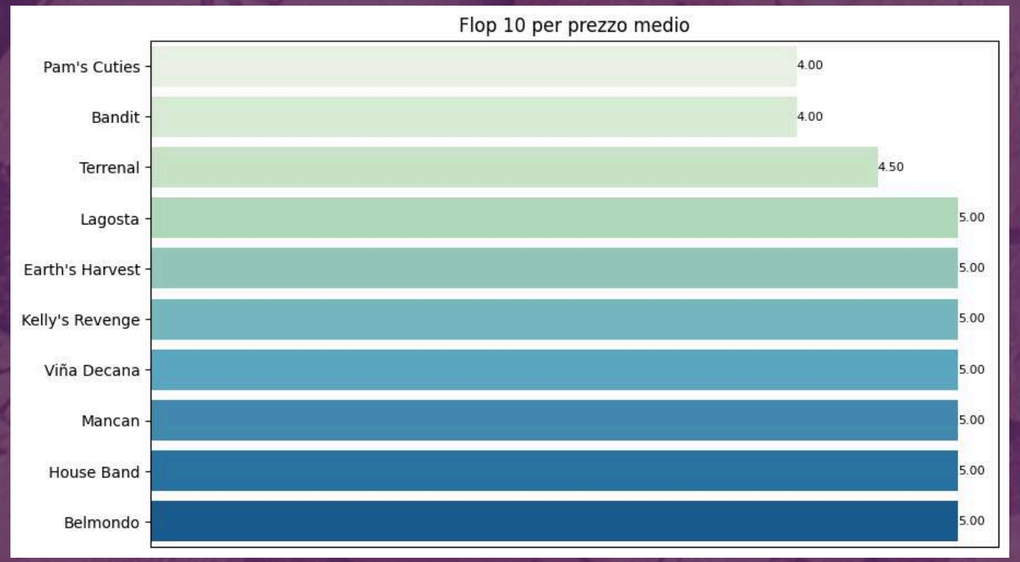
Nazioni



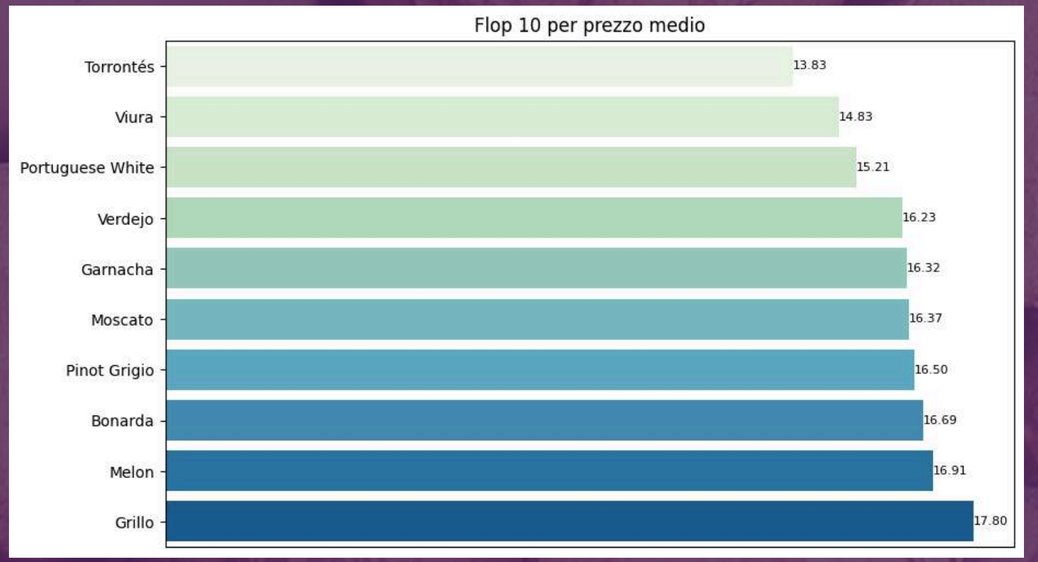
Regioni



Cantine

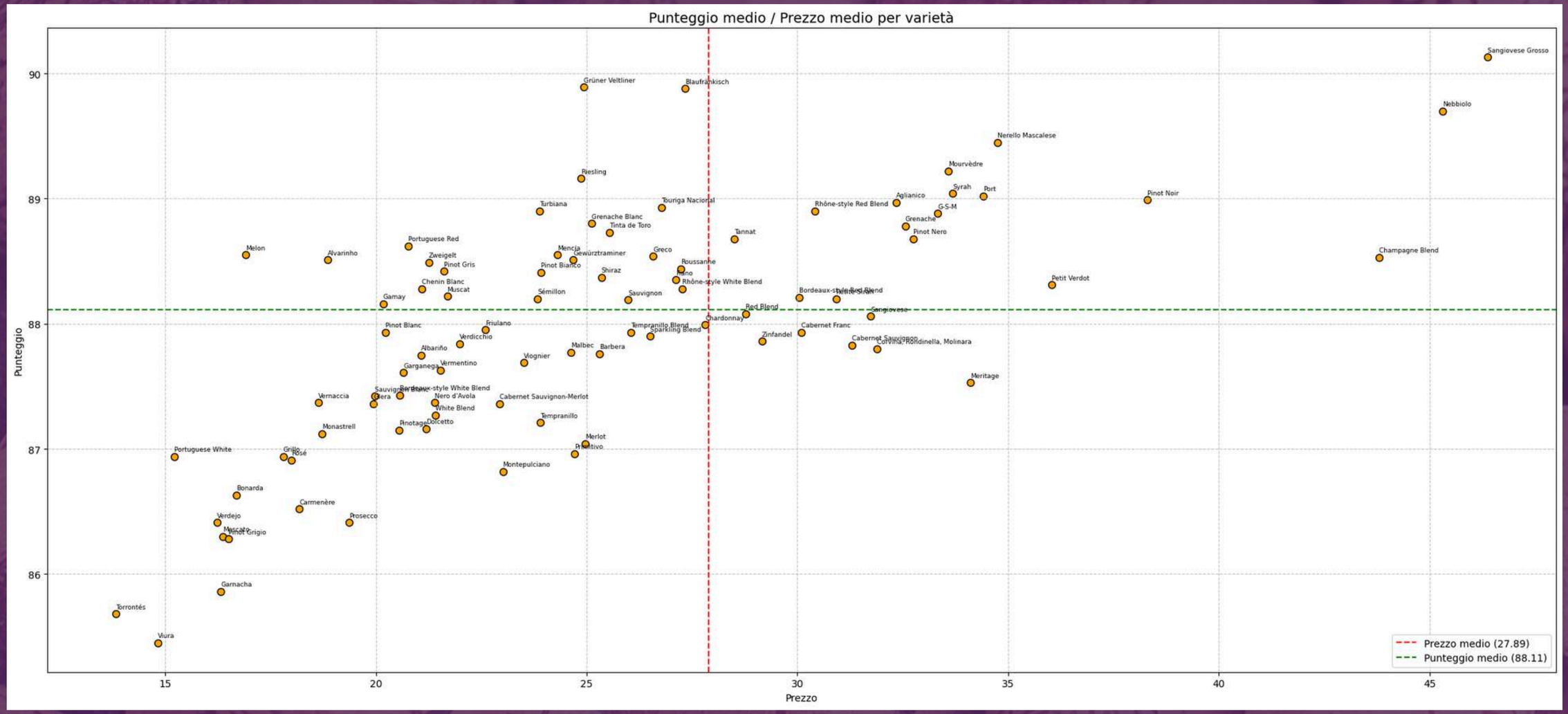


Varietà



Proposta Marketplace

Per poter avanzare una proposta di marketplace partiamo dal mettere su un grafico a dispersione il punteggio medio e il prezzo medio di ogni varietà prese in esame



Il passo successivo è quello di dividere il dataset in 3 fasce di prezzo: **basso** (< 24,68), **medio** (24,68 - 35,52) e **alto** (> 35,52)

Fascia prezzo basso

variety	price	points
Turbiana	23.89	88.90
Portuguese Red	20.77	88.62
Melon	16.91	88.55
Mencía	24.30	88.55
Alvarinho	18.86	88.51
Zweigelt	21.25	88.49
Pinot Gris	21.62	88.42
Pinot Bianco	23.92	88.41
Chenin Blanc	21.09	88.28
Muscat	21.69	88.22
Sémillon	23.83	88.20
Gamay	20.17	88.16
Friulano	22.60	87.95
Pinot Blanc	20.22	87.93
Verdicchio	21.99	87.84
Malbec	24.62	87.77
Albariño	21.07	87.75
Viognier	23.51	87.69
Vermentino	21.53	87.63
Garganega	20.65	87.61
Bordeaux-style White Blend	20.57	87.43
Sauvignon Blanc	19.97	87.42
Nero d'Avola	21.40	87.37
Vernaccia	18.63	87.37
Cabernet Sauvignon-Merlot	22.94	87.36
Glera	19.94	87.36
White Blend	21.41	87.27
Tempranillo	23.90	87.21
Dolcetto	21.19	87.16
Pinotage	20.55	87.15
Monastrell	18.72	87.12
Portuguese White	15.21	86.94
Grillo	17.80	86.94
Rosé	17.99	86.91
Montepulciano	23.01	86.82
Bonarda	16.69	86.63
Carmenère	18.18	86.52
Prosecco	19.36	86.41
Verdejo	16.23	86.41
Moscato	16.37	86.30
Pinot Grigio	16.50	86.28
Garnacha	16.32	85.86
Torrontés	13.83	85.68
Viura	14.83	85.45

```
group1 = data3_merge_variety[data3_merge_variety['price'] <= (min+((max-min)/3))].sort_values('points', ascending=False)
print(f'Fascia prezzo basso\n', group1)
```

```
group2 = data3_merge_variety[(data3_merge_variety['price'] >= (min+((max-min)/3))) & (data3_merge_variety['price'] <= (max-((max-min)/3)))].sort_values('points', ascending=False)
print(f'\nFascia prezzo medio\n', group2)
```

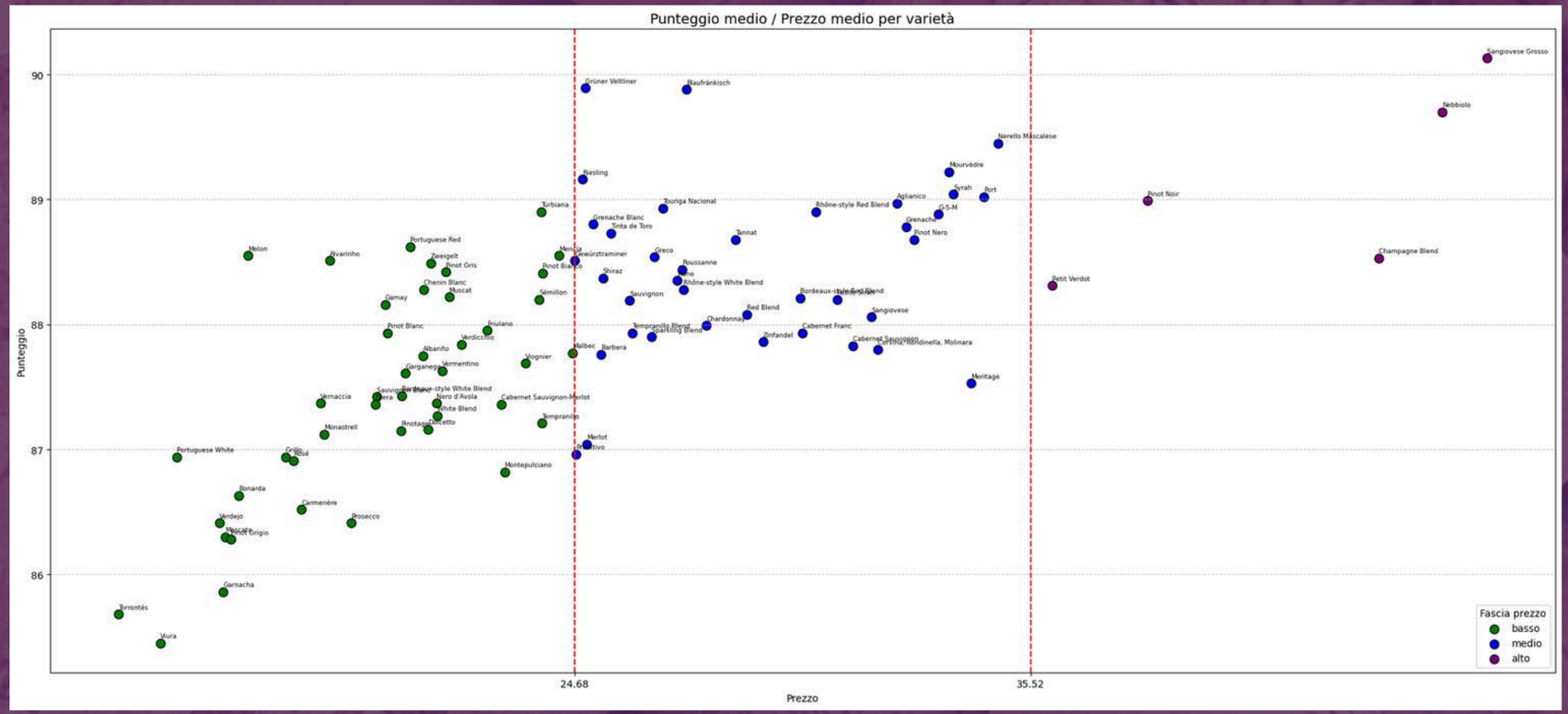
Fascia prezzo medio

variety	price	point
Grüner Veltliner	24.93	89.89
Blaufränkisch	27.34	89.88
Nerello Mascalese	34.75	89.45
Mourvèdre	33.58	89.22
Riesling	24.87	89.16
Syrah	33.69	89.04
Port	34.41	89.02
Aglianico	32.34	88.97
Touriga Nacional	26.78	88.93
Rhône-style Red Blend	30.42	88.90
G-S-M	33.33	88.88
Grenache Blanc	25.12	88.80
Grenache	32.56	88.78
Tinta de Toro	25.54	88.73
Tannat	28.50	88.68
Pinot Nero	32.75	88.68
Greco	26.58	88.54
Gewürztraminer	24.68	88.51
Roussanne	27.23	88.44
Shiraz	25.35	88.37
Fiano	27.12	88.35
Rhône-style White Blend	27.27	88.28
Bordeaux-style Red Blend	30.04	88.21
Petite Sirah	30.93	88.20
Sauvignon	25.98	88.19
Red Blend	28.78	88.08
Sangiovese	31.74	88.06
Chardonnay	27.81	87.99
Tempranillo Blend	26.05	87.93
Cabernet Franc	30.10	87.93
Sparkling Blend	26.50	87.90
Zinfandel	29.16	87.86
Cabernet Sauvignon	31.30	87.83
Corvina, Rondinella, Molinara	31.88	87.80
Barbera	25.30	87.76
Meritage	34.10	87.53
Merlot	24.97	87.04
Primitivo	24.71	86.96

```
group3 = data3_merge_variety[data3_merge_variety['price'] >= (max-((max-min)/3))].sort_values('points', ascending=False)
print(f'\nFascia prezzo alto\n', group3)
```

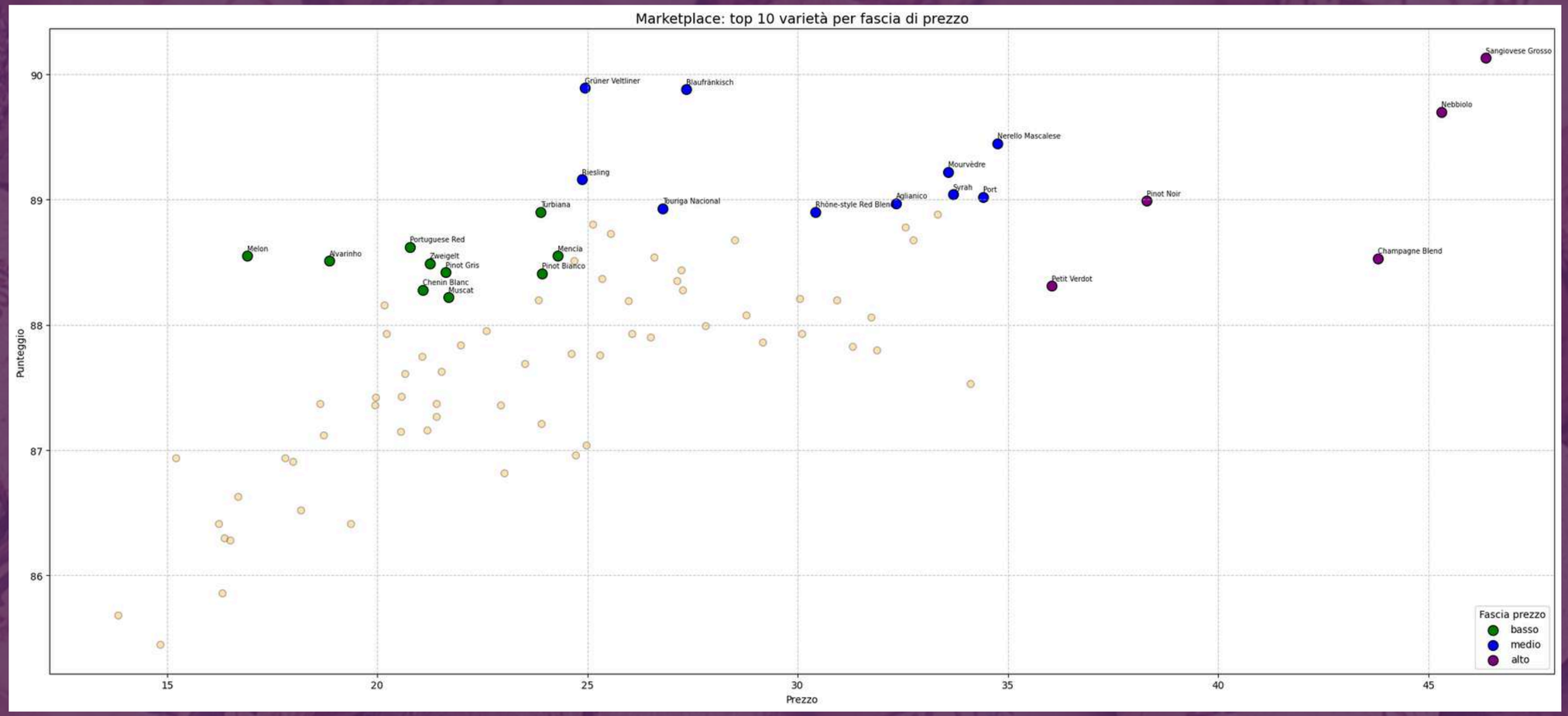
Fascia prezzo alto

variety	price	points
Sangiovese Grosso	46.37	90.13
Nebbiolo	45.31	89.70
Pinot Noir	38.30	88.99
Champagne Blend	43.80	88.53
Petit Verdot	36.04	88.31



Dopodiché per ogni fascia di prezzo prendiamo le prime 10 varietà con il punteggio più alto.
Ed ecco qua il nostro marketplace!

variety	price	points
Melon	16.91	88.55
Alvarinho	18.86	88.51
Portuguese Red	20.77	88.62
Chenin Blanc	21.09	88.28
Zweigelt	21.25	88.49
Pinot Gris	21.62	88.42
Muscat	21.69	88.22
Turbiana	23.89	88.90
Pinot Bianco	23.92	88.41
Mencía	24.30	88.55
Riesling	24.87	89.16
Grüner Veltliner	24.93	89.89
Touriga Nacional	26.78	88.93
Blaufränkisch	27.34	89.88
Rhône-style Red Blend	30.42	88.90
Aglianico	32.34	88.97
Mourvèdre	33.58	89.22
Syrah	33.69	89.04
Port	34.41	89.02
Nerello Mascalese	34.75	89.45
Petit Verdot	36.04	88.31
Pinot Noir	38.30	88.99
Champagne Blend	43.80	88.53
Nebbiolo	45.31	89.70
Sangiovese Grosso	46.37	90.13

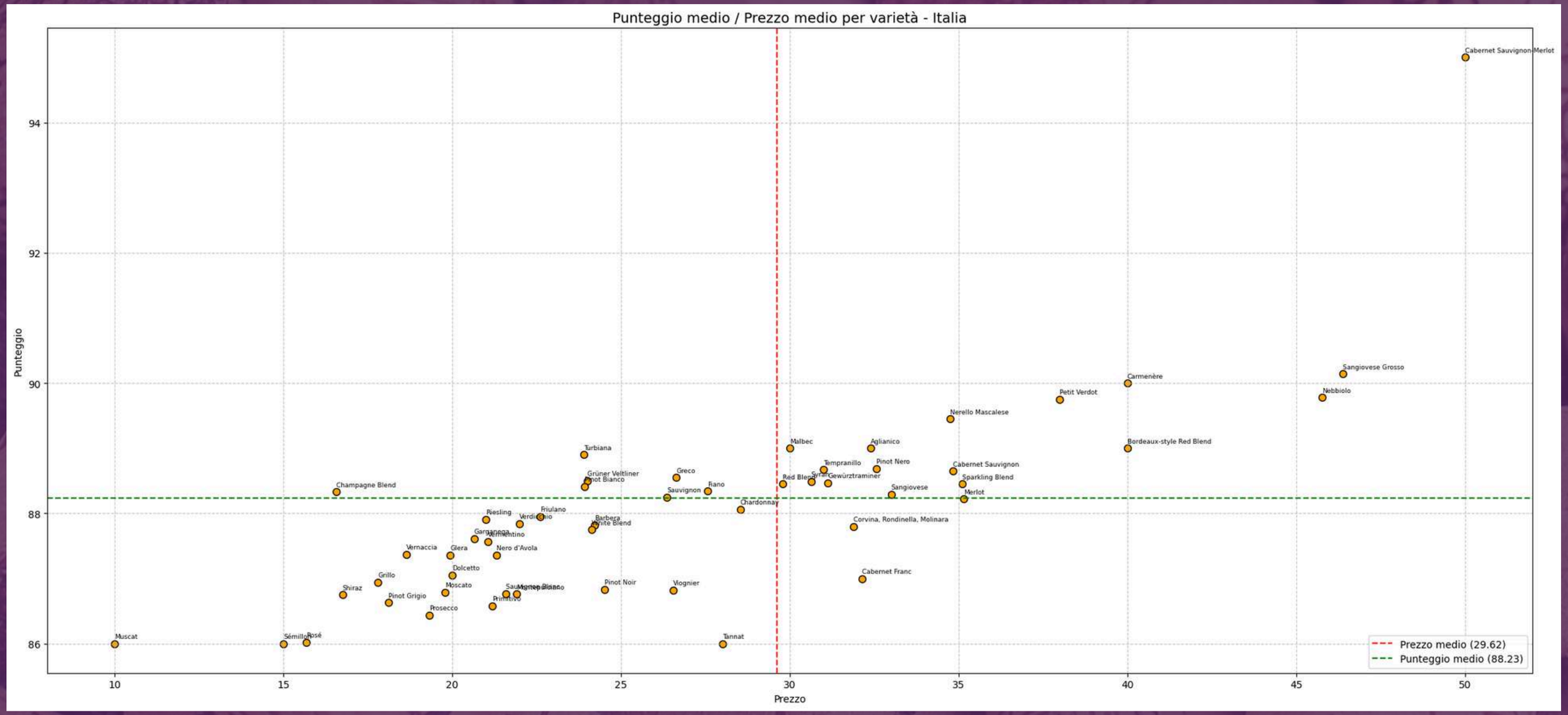


Sono stati individuati complessivamente 25 vini*

* nella fascia di prezzo alto sono presenti solamente 5 vini

E in Italia?

Come fatto precedentemente, mettiamo su un grafico a dispersione il punteggio medio e il prezzo medio di ogni varietà in l'Italia



Dividiamo il dataset in 3 fasce di prezzo: **basso** (< 23,33), **medio** (23,33 - 36,67) e **alto** (> 36,67)

Fascia prezzo basso - Italia

variety	price	points
Champagne Blend	16.56	88.33
Friulano	22.60	87.95
Riesling	21.00	87.91
Verdicchio	21.99	87.84
Garganega	20.65	87.61
Vermentino	21.06	87.57
Vernaccia	18.63	87.37
Nero d'Avola	21.30	87.36
Glera	19.94	87.36
Dolcetto	19.99	87.05
Grillo	17.80	86.94
Moscato	19.79	86.79
Sauvignon Blanc	21.59	86.77
Montepulciano	21.91	86.76
Shiraz	16.75	86.75
Pinot Grigio	18.11	86.63
Primitivo	21.19	86.58
Prosecco	19.32	86.44
Rosé	15.67	86.02
Muscat	10.00	86.00
Sémillon	15.00	86.00

```
group1_italy = data3_merge_italy[data3_merge_italy['price'] <= (min_italy+((max_italy-min_italy)/3))].sort_values('points', ascending=False)
print(f'Fascia prezzo basso - Italia\n', group1_italy)
```

```
group2_italy = data3_merge_italy[(data3_merge_italy['price'] >= (min_italy+((max_italy-min_italy)/3))) & (data3_merge_italy['price'] <= (max_italy-((max_italy-min_italy)/3)))].sort_values('points', ascending=False)
print(f'\nFascia prezzo medio - Italia\n', group2_italy)
```

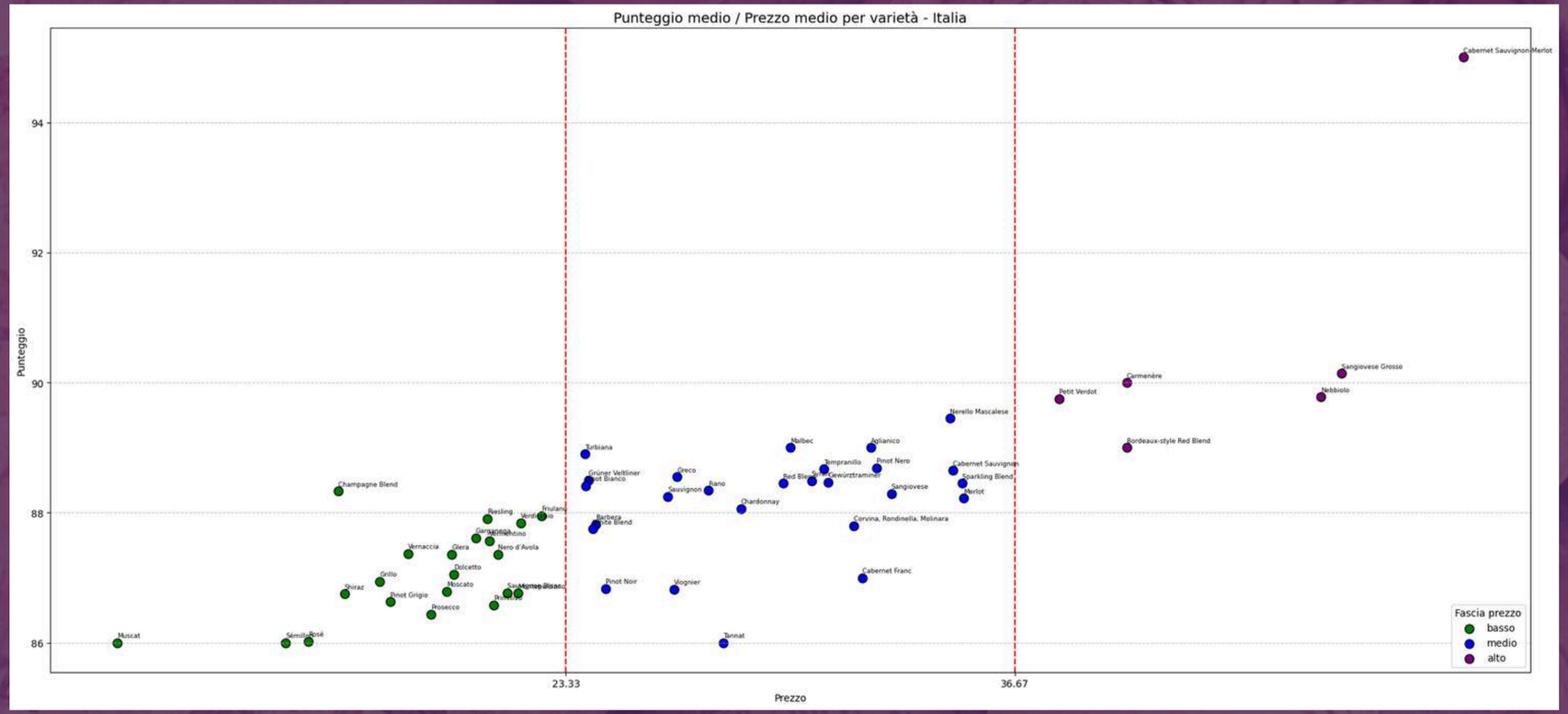
Fascia prezzo medio - Italia

variety	price	points
Nerello Mascalese	34.75	89.45
Aglianico	32.40	89.00
Malbec	30.00	89.00
Turbiana	23.89	88.90
Pinot Nero	32.57	88.69
Tempranillo	31.00	88.67
Cabernet Sauvignon	34.84	88.65
Greco	26.64	88.55
Grüner Veltliner	24.00	88.50
Syrah	30.63	88.49
Gewürztraminer	31.13	88.47
Sparkling Blend	35.11	88.45
Red Blend	29.79	88.45
Pinot Bianco	23.92	88.41
Fiano	27.56	88.34
Sangiovese	33.00	88.29
Sauvignon	26.36	88.25
Merlot	35.16	88.22
Chardonnay	28.53	88.06
Barbera	24.22	87.82
Corvina, Rondinella, Molinara	31.88	87.80
White Blend	24.12	87.75
Cabernet Franc	32.14	87.00
Pinot Noir	24.50	86.83
Viognier	26.55	86.82
Tannat	28.00	86.00

```
group3_italy = data3_merge_italy[data3_merge_italy['price'] >= (max_italy-((max_italy-min_italy)/3))].sort_values('points', ascending=False)
print(f'\nFascia prezzo alto - Italia\n', group3_italy)
```

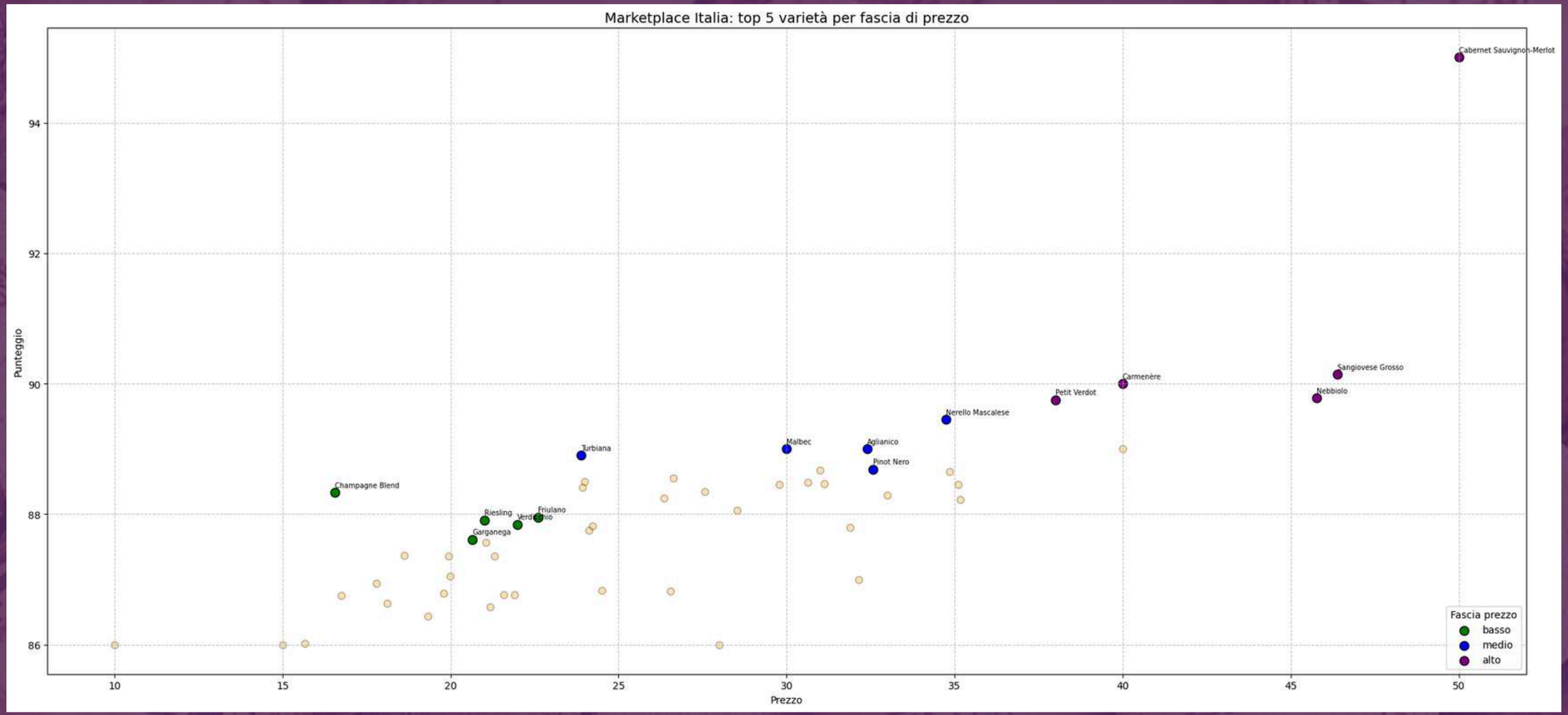
Fascia prezzo alto - Italia

variety	price	points
Cabernet Sauvignon-Merlot	50.00	95.00
Sangiovese Grosso	46.38	90.14
Carmenère	40.00	90.00
Nebbiolo	45.77	89.78
Petit Verdot	38.00	89.75
Bordeaux-style Red Blend	40.00	89.00



Dopodiché per ogni fascia di prezzo prendiamo le prime 5 varietà con il punteggio più alto.
Ed ecco qua il nostro marketplace per l'Italia!

variety	price	points
Champagne Blend	16.56	88.33
Garganega	20.65	87.61
Riesling	21.00	87.91
Verdicchio	21.99	87.84
Friulano	22.60	87.95
Turbiana	23.89	88.90
Malbec	30.00	89.00
Aglianico	32.40	89.00
Pinot Nero	32.57	88.69
Nerello Mascalese	34.75	89.45
Petit Verdot	38.00	89.75
Carmenère	40.00	90.00
Nebbiolo	45.77	89.78
Sangiovese Grosso	46.38	90.14
Cabernet Sauvignon-Merlot	50.00	95.00



Sono stati individuati complessivamente 15 vini

Conclusioni

Conclusioni

L'analisi è stata effettuata su 102.342 record dopo aver sfoltito il dataset iniziale attraverso la conversione dei valori nulli in "0", l'eliminazione degli outlier sulla colonna *points* e *price* e l'eliminazione delle varietà con un numero di recensioni inferiore a 100

La media (27,89) e la mediana (25) sulla colonna *price* sono molto simili per cui per l'analisi si è deciso di prendere come riferimento la media. Lo stesso è avvenuto per la media (88,11) e la mediana (88) sulla colonna *points*

I valori della colonna *price* presentano una leggera asimmetria sinistra, mentre i valori della colonna *points* sono pressoché normalmente distribuiti (entrambi confermati numericamente dai valori di media e mediana come sopra)

Esiste circa il 53% di correlazione tra i valori di *price* e *points*

Correlazione tra valori di *price* e *points* sufficientemente confermata andando a visualizzare i grafici di pag. 31 e pag. 35, rispettivamente nel caso generale e nel caso dell'Italia

Per la proposta di marketplace per vini recensiti in generale sono stati individuati complessivamente 25 vini divisi in 3 fasce di prezzo (basso, medo, alto). Per la proposta di marketplace per vini recensiti in Italia sono stati individuati invece 15 vini altresì divisi in 3 fasce di prezzo (basso, medio, alto)

Per ulteriori approfondimenti si potrebbe estendere lo studio dividendo i vini per colore (bianchi, neri e rosati) oppure in base all'annata


Buona Degustazione!