UNIVERSIDAD SIGLO 21

La educación evoluciona

Carrera: Lic. Ciencia de Datos.

Materia: Seminario de Práctica en Ciencia de Datos.

Título: Análisis de las variedad y calidad en vinos.

Entrega N°: 4

Prefesor: Virgolini, Pablo Alejandro.

Alumno: Altamirano, Nicolás.

Fecha de entrega:

17/11/2024

Comentarios:

Contenido

Descripción de la temática de los datos (Contexto)

Para el presente trabajo, se tomarán datos sobre reseñas de vinos de WineEnthusiast que contienen detalles de país de origen, variedad de uva, bodega, calificación de puntos, precio, detalles geográficos, y el nombre y usuario de Twitter de la persona que lo degustó. Este dataset permitirá considerar la relación entre el precio y la calidad, las diferencias geográficas o regionales, y las tendencias de calificación.

Selección, depuración, exploración y filtrado de datos

Se procedió a la lectura del dataset que contiene el conjunto de datos, el cual fue previamente depurado y organizado utilizando Microsoft Excel. En esta fase de preprocesamiento, los datos fueron transformados adecuadamente en columnas y reorganizados para mejorar su estructura. Originalmente, el archivo estaba en formato CSV, sin embargo, este formato presentaba una complicación adicional: el carácter de coma (,) no solo funcionaba como separador de columnas, sino que también aparecía dentro de ciertos campos de datos, lo que generaba desorganización. Por lo tanto, fue necesario realizar una depuración cuidadosa para evitar interpretaciones erróneas durante la carga del archivo en el entorno de análisis.

Librerías utilizadas

Se utilizó Python para la carga del archivo a través de la librería pandas, lo que permitió revisar el formato de los datos y su correcta importación. Además, se incluyeron diversas bibliotecas para facilitar el análisis y tratamiento de los datos: NumPy (cálculo numérico y operaciones con arreglos multidimensionales), matplotlib.pyplot y seaborn (creación de gráficos y visualización de datos, proporcionando herramientas para analizar patrones y distribuciones de las variables), StandardScaler y MinMaxScaler (preprocesamiento empleados para escalar los datos y normalizarlos, mejorando así la precisión en modelos posteriores),

SimpleImputer (imputar valores faltantes en el conjunto de datos mediante estrategias como la media, mediana o moda).

Carga del archivo

```
#Para poder realizar un análisis más completo y organizados, debemos ayudarnos
# de las diferentes librerías que tenemos a disposición.
#Carga de todas las librerías a utilizar:

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.impute import  SimpleImputer
```

Leemos el archivo donde tenemos los datos. Este xlsx fue depurado anteriormente con Excel: Se transformaron los datos en columnas y se organizaron mejor dichos datos, porque el archivo original era un CSV, pero las "," no solo separaban las columnas, sino que también había datos en donde se la utilizaba

```

df = pd.read_excel("../BD/Entregable2-Depurado.xlsx", index_col="ID")
```

Fuente: Elaboración Propia (TP4.py)

Verificación de datos

La verificación de datos reveló que el conjunto de datos fue cargado correctamente, confirmando que cuenta con 129.971 registros y 13 columnas. Se identificaron diversas variables, tanto numéricas como categóricas, entre las cuales destacan las columnas: *country*, *description*, *designation*, *points y price*. La mayoría de las columnas son del tipo *object* (string), pero también se encontraron tipos numéricos para las variables correspondientes a *points* y *price*. Durante el proceso de verificación, se detectaron valores vacíos en el conjunto de datos, los cuales serán abordados en etapas posteriores del análisis.

Verificación del archivo

```
1 df.head()
```

ID	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
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	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	150.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	140.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm
3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	130.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian
4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	650.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks

Fuente: Elaboración Propia (TP4.py)

```
1 df.info()
```

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

```
1 df.shape
```

```
(129971, 13)
```

```
1 df_countNA = df.isnull().sum()
2 df_countNA.sort_values()
```

```
description                0
points                     0
title                      0
winery                     0
variety                    1
country                   63
province                  63
price                   8996
region_1               21247
taster_name            26244
taster_twitter_handle  31213
designation            37465
region_2               79460
dtype: int64
```

Resumen de la Verificación del archivo

El archivo fue cargado correctamente, pudiendo verificar que cuenta con 129.971 filas y 13 columnas. Los tipos de datos fueron en su mayoría del tipo object (string), pero también se encontraron tipos numéricos, tanto para los puntos como para los precios. Se encontraron valores vacíos, que más adelante se tomara una decisión sobre ellos.

Fuente: Elaboración Propia (TP4.py)

Análisis exploratorio de datos

¿Qué es? El Análisis Exploratorio de Datos (EDA por su sigla en inglés) es una etapa fundamental en el proceso de análisis de datos, que tiene como objetivo principal comprender las características y la estructura del conjunto de datos antes de aplicar métodos estadísticos o modelos de machine learning. Este análisis permite identificar patrones y tendencias a través de visualizaciones y resúmenes estadísticos, descubriendo relaciones y patrones relevantes entre las variables. Además, ayuda a detectar valores atípicos, identificando y evaluando datos que se desvían significativamente de los valores esperados, lo que puede influir en los resultados del análisis.

También revisa la calidad de los datos, permitiendo detectar errores, valores faltantes e inconsistencias, lo que facilita su tratamiento adecuado.

Gracias a este primer análisis, se identificaron las siguientes variables, las cuales se analizarán para descubrir información relevante.

- **country**: País de origen del vino.
- **description**: Reseña del vino.
- **designation**: Viñedo específico dentro de la bodega.
- **points**: Puntuación de calidad percibida.
- **price**: Precio de una botella de vino.
- **province**: Provincia o estado de origen.
- **region_1 y region_2**: áreas vinícolas específicas.
- **taster_name y taster_twitter_handle**: información sobre el catador que realizó la reseña.
- **variety**: Variedad de uva utilizada.
- **winery**: Bodega productora.

Resumen estadístico inicial:

En cuanto a las puntuaciones, estas oscilan entre 80 y 100, lo que confirma que el sistema de evaluación se limita a incluir vinos de calidad moderada a excelente. La puntuación

media es de 88.45, indicando que la mayoría de los vinos evaluados son considerados de buena calidad. La desviación estándar es de 3.04, lo que muestra que las puntuaciones no están muy dispersas y tienden a concentrarse alrededor de la media, con pocas desviaciones hacia puntuaciones extremas. En conclusión, la distribución está sesgada hacia el extremo superior (valores mayores de 80), reflejando la naturaleza del sistema de puntuación que solo publica vinos con buenas calificaciones.

Respecto a los precios, estos varían ampliamente, desde 40 USD hasta 33,000 USD, lo que sugiere una gran diversidad en términos de accesibilidad y mercado objetivo. El precio promedio de los vinos es de 353.63 USD; sin embargo, este valor puede estar influenciado por algunos precios extremadamente altos. La desviación estándar de 410.22 USD indica una alta variabilidad en los precios, reforzando la idea de que algunos vinos son significativamente más caros que la mayoría. En conclusión, la distribución de precios parece estar sesgada hacia valores altos debido a algunos vinos extremadamente caros, siendo el máximo de 33,000 USD muy superior al resto. La mayoría de los vinos se encuentran en un rango de precios mucho más accesible, entre 170 y 420 USD.

Análisis exploratorio

```
# En una primera impresión veremos que dicen los datos:

# Descripción general de las columnas numéricas (points and price)
print("\nDescripción estadística de las columnas numéricas:")
df.describe()
```

```
Descripción estadística de las columnas numéricas:
```

	points	price
count	129971.000000	120975.000000
mean	88.447138	353.633891
std	3.039730	410.222177
min	80.000000	40.000000
25%	86.000000	170.000000
50%	88.000000	250.000000
75%	91.000000	420.000000
max	100.000000	33000.000000

Fuente: Elaboración Propia (TP4.py)

En cuanto al país, se registraron 129,908 vinos con una asociación geográfica. El conjunto de datos incluye 43 países diferentes, siendo Estados Unidos (US) el país más común, con 54,504 vinos, lo que sugiere una sobre-representación de vinos de esa nación.

Respecto a las descripciones, se registraron 129,971 reseñas, lo que indica que prácticamente todos los vinos tienen una descripción. Hay 119,955 descripciones únicas, lo que sugiere que la mayoría de los vinos cuentan con reseñas detalladas y específicas, lo que será útil para un análisis textual. La descripción más común solo aparece tres veces, lo que indica una gran variedad en el contenido.

En lo que respecta a la variedad de uva, casi todos los vinos (129,970) tienen información sobre esta variable. Existen 731 variedades diferentes, siendo Pinot Noir la más común, con 13,272 vinos.

Finalmente, todos los vinos tienen información sobre la bodega, con un total de 17,111 bodegas diferentes. La bodega más común es Wines & Winemakers, que produce 222 vinos.

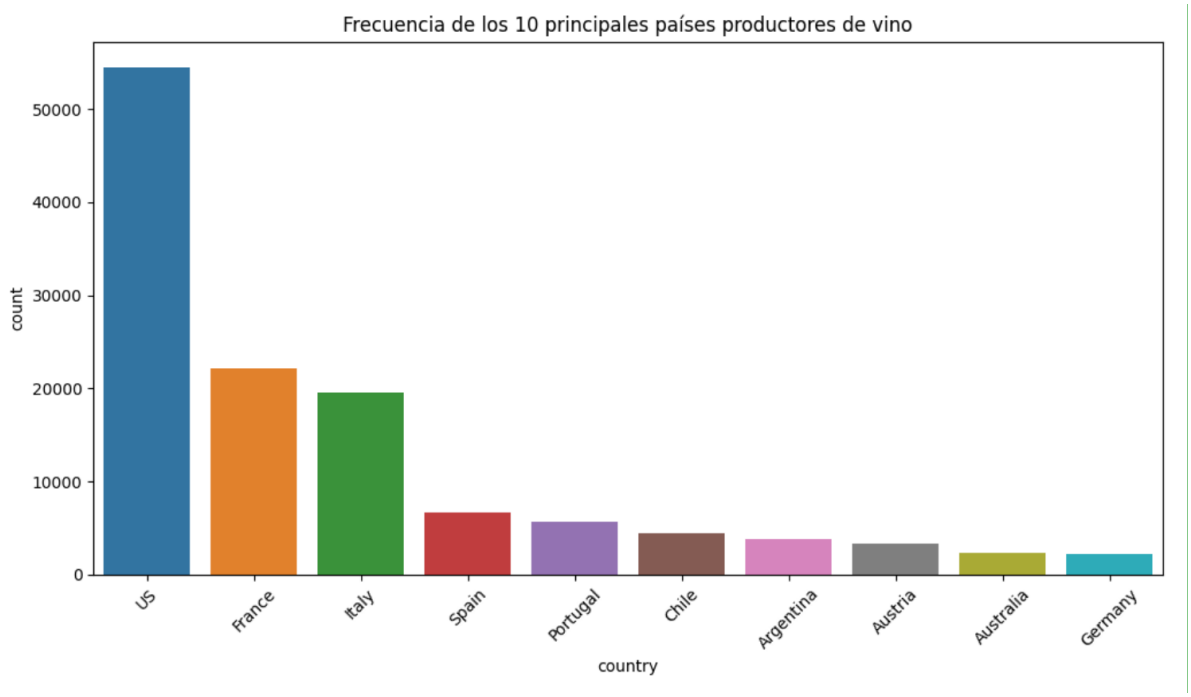
A continuación, se presentan algunas imágenes ilustrativas que ayudan a comprender mejor los datos analizados.

```
# Descripción general de las columnas categóricas
print("\nDescripción de las columnas categóricas:")
df.describe(include=['O'])
```

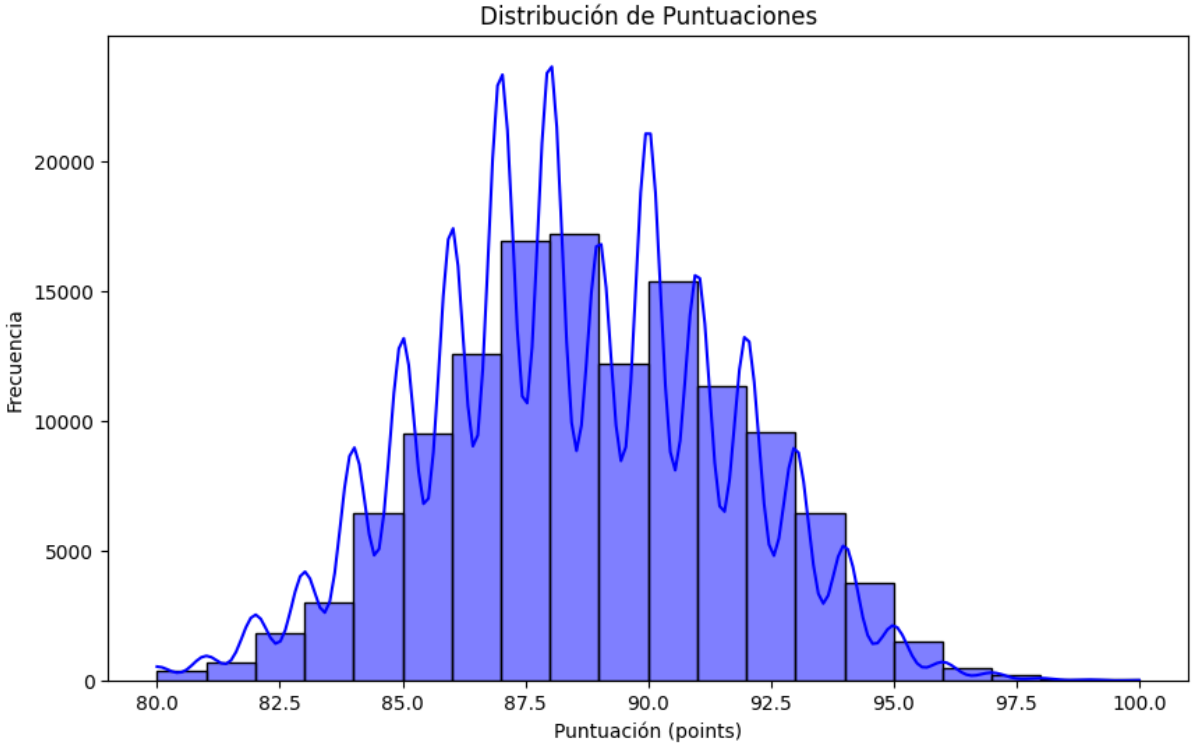
```
Descripción de las columnas categóricas:
```

	country	description	designation	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
count	129908	129971	92506	129908	108724	50511	103727	98758	129971	129970	129971
unique	43	119955	37880	436	1297	17	21	16	118866	731	17111
top	US	Seductively tart in lemon pith, cranberry and ...	Reserve	California	Napa Valley	Central Coast	Roger Voss	@vossroger	Gloria Ferrer NV Sonoma Brut Sparkling (Sonoma...	Pinot Noir	Wines & Winemakers
freq	54504	3	2009	36247	4480	11065	25514	25514	11	13272	222

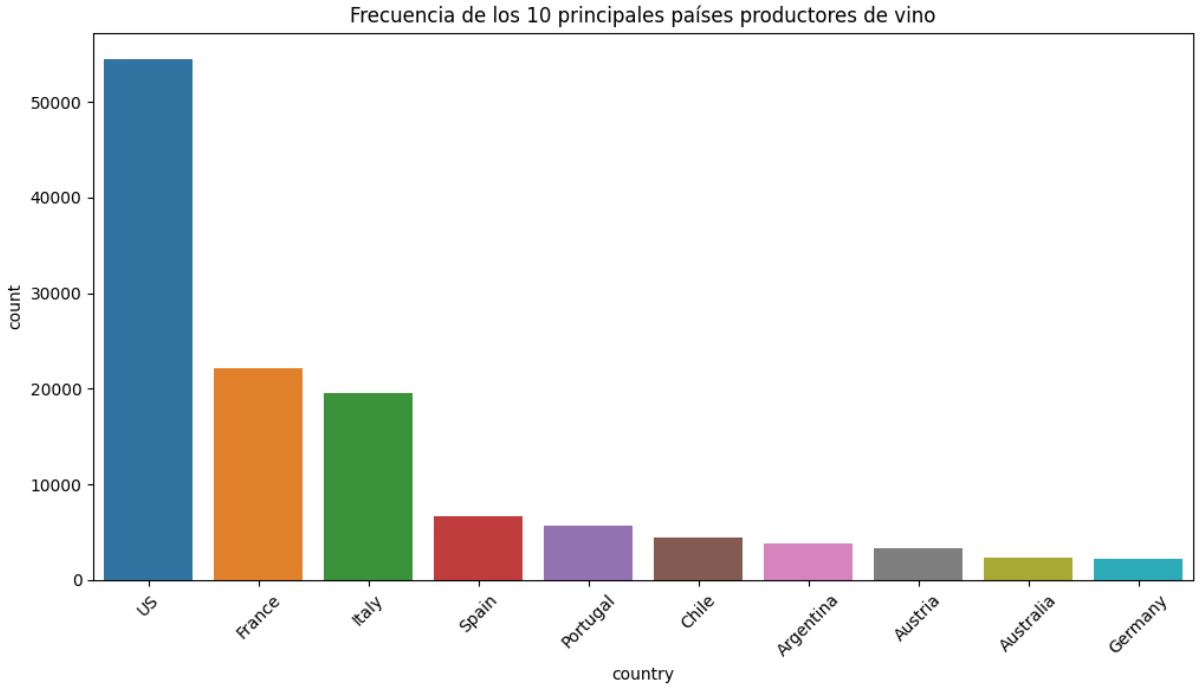
Fuente: Elaboración Propia (TP4.py)



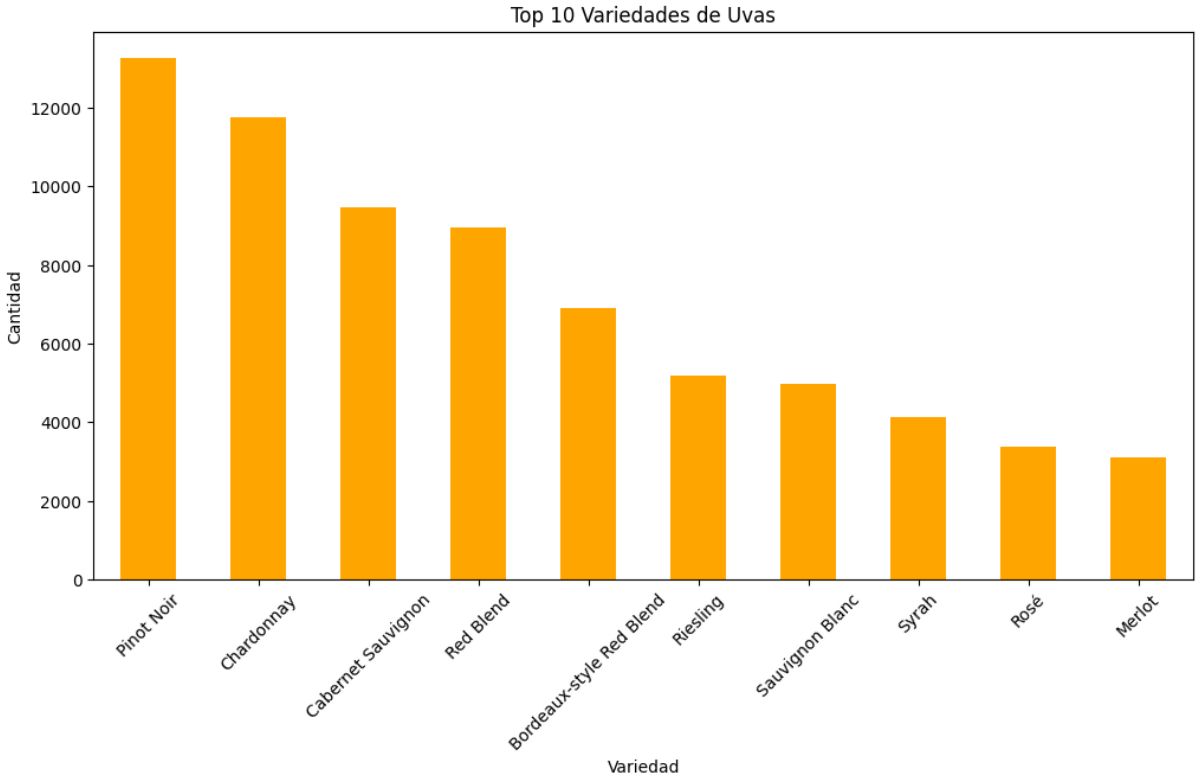
Fuente: Elaboración Propia (TP4.py)



Fuente: Elaboración Propia (TP4.py)



Fuente: Elaboración Propia (TP4.py)



Fuente: Elaboración Propia (TP4.py)

Tratamiento de datos faltantes

El tratamiento de datos faltantes se refiere a las técnicas y métodos utilizados para manejar los valores que no están disponibles en un conjunto de datos. Esto es importante porque los datos incompletos pueden afectar la calidad del análisis y las decisiones que se tomen basadas en esos datos. Las estrategias comunes incluyen eliminar las filas con datos faltantes, imputar valores (sustituyendo los faltantes por la media, mediana o moda), o usar técnicas más avanzadas para estimar los valores que faltan. La elección del método depende del contexto y del impacto que los datos faltantes puedan tener en el análisis.

Como se indicó en la *verificación de datos*, existen valores ausentes en diferentes variables. Para solucionarlos, iniciemos un análisis de cada variable con el fin de identificar la mejor estrategia para abordar este problema.

Tratamiento de datos faltantes

```
# Volvamos a ver como estan distribuidos los datos faltantes:
df_countNA.sort_values()
```

```
description                 0
points                      0
title                       0
winery                      0
variety                     1
country                    63
province                   63
price                    8996
region_1                21247
taster_name             26244
taster_twitter_handle   31213
designation             37465
region_2                79460
dtype: int64
```

```
df_Percent = df_countNA/129971*100
df_Percent.sort_values()
```

```
description              0.000000
points                   0.000000
title                    0.000000
winery                   0.000000
variety                  0.000769
country                  0.048472
province                 0.048472
price                    6.921544
region_1                16.347493
taster_name             20.192197
taster_twitter_handle   24.015357
designation             28.825661
region_2                61.136715
dtype: float64
```

Fuente: Elaboración Propia (TP4.py)

En la variable “variety”, se decidió optar por eliminar filas. Esta estrategia es adecuada cuando los datos faltantes son esporádicos o representan una pequeña proporción del total. En este caso, la reseña con ID 86909 es insignificante, por lo que se procederá a eliminar este dato.

```
df[df["variety"].isna() | (df["variety"] == "")]
```

ID	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
86909	Chile	A chalky, dusty mouthfeel nicely balances this...	NaN	88	170.0	Maipo Valley	NaN	NaN	NaN	NaN	Carmen 1999 (Maipo Valley)	NaN	Carmen

Eliminar filas: Esta estrategia es adecuada cuando los datos faltantes son esporádicos o representan una pequeña proporción del total. En este caso, la reseña ID 86909, es insignificante. Se procede a limpiar dicho dato

```
df_Limpio = df[~df["variety"].isna() & (df["variety"] != "")]
```

Fuente: Elaboración Propia (TP4.py)

En los valores faltantes de la columna "country", se observa que todos los datos relacionados con el lugar geográfico están ausentes. Si se eliminaran todos los registros con datos faltantes, se podría perder información valiosa, como la reseña en general, que apunta a una variedad específica. Sin embargo, la cantidad de datos faltantes es mínima, con solo 63 registros, lo que representa apenas el 0.048% del conjunto de datos. Por lo tanto, se procederá a eliminar estos datos faltantes.

```
df_Limpio[df_Limpio["country"].isna() | (df_Limpio["country"] == "")].head()
```

ID	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
913	NaN	Amber in color, this wine has aromas of peach ...	Asureti Valley	87	300.0	NaN	NaN	NaN	Mike DeSimone	@worldwineguys	Gotsa Family Wines 2014 Asureti Valley Chinuri	Chinuri	Gotsa Family Wines
3131	NaN	Soft, fruity and juicy, this is a pleasant, si...	Partager	83	NaN	NaN	NaN	NaN	Roger Voss	@vossroger	Barton & Guestier NV Partager Red	Red Blend	Barton & Guestier
4243	NaN	Violet-red in color, this semisweet wine has a...	Red Naturally Semi-Sweet	88	180.0	NaN	NaN	NaN	Mike DeSimone	@worldwineguys	Kakhetia Traditional Winemaking 2012 Red Natur...	Ojaleshi	Kakhetia Traditional Winemaking
9509	NaN	This mouthwatering blend starts with a nose of...	Theopetra Malagouzia-Assyrtiko	92	280.0	NaN	NaN	NaN	Susan Kostrzewa	@suskostrzewa	Tsililis 2015 Theopetra Malagouzia-Assyrtiko W...	White Blend	Tsililis
9750	NaN	This orange-style wine has a cloudy yellow-gol...	Orange Nikolaevo Vineyard	89	280.0	NaN	NaN	NaN	Jeff Jenssen	@worldwineguys	Ross-idi 2015 Orange Nikolaevo Vineyard Chardo...	Chardonnay	Ross-idi

En los valores faltantes de la columna "country", se puede observar que faltan todos los datos relacionados con el lugar geográfico. Si se eliminan todos los datos los datos faltantes, se podrían perder información valiosa como lo es la reseña en general, y esta apunta a una variedad especifica. Sin embargo, la cantidad de datos faltantes es infimita (63), representando solo el %0.048 del dataset. Por ende, se procede a la eliminación de dichos datos faltantes

```
df_Limpio = df[~df["country"].isna() & (df["country"] != "")]
```

Fuente: Elaboración Propia (TP4.py)

Los datos faltantes en la columna de "price" son significativamente elevados, representando un 6.92% del conjunto de datos. Dado que esta columna es clave, eliminar esos valores podría resultar en la pérdida de información valiosa. Como la variable es numérica, se puede optar por imputar los datos faltantes utilizando medidas aritméticas, dependiendo de la distribución de los datos. Dado que los precios presentan una distribución sesgada hacia valores altos, sería más apropiado imputar con la mediana, ya que esta medida es menos sensible a los valores extremos.

```
:
# Imputar los valores faltantes en 'price' con la mediana
# Crear el imputador, en este caso usamos la estrategia 'median'
imputador = SimpleImputer(missing_values=np.nan, strategy='median')

# Aplicar el imputador a la columna 'price' y transformar los valores
df_Limpio['price'] = imputador.fit_transform(df_Limpio[['price']])
```

Fuente: Elaboración Propia (TP4.py)

En la columna "region_1", se observan 21,247 datos faltantes. Dado que esta columna es clave para el análisis, eliminar esos valores resultaría en la pérdida de información valiosa. Al tratarse de una variable categórica, se puede optar por imputar los datos faltantes utilizando la moda (el valor más frecuente). Por ejemplo, si la mayoría de los vinos en el conjunto de datos provienen de una región en particular, se podría asumir que los valores faltantes también pertenecen a esa misma región. Esta estrategia ayuda a reducir el sesgo introducido por el proceso de imputación y mantiene la coherencia en los datos.

```
# Imputar los valores faltantes en 'region_1' con la moda
mode_region_1 = df_Limpio['region_1'].mode()[0]
df_Limpio['region_1'].fillna(mode_region_1, inplace=True)
```

Fuente: Elaboración Propia (TP4.py)

En cuanto a las columnas "taster_name" y "taster_twitter_handle", actualmente no son de gran importancia para el análisis. Sin embargo, eliminarlas podría introducir sesgos en futuros estudios, y dado que representan más del 20% del total de datos, borrar las filas correspondientes no sería conveniente. Por lo tanto, se optará por completar "taster_name" con

el valor "Desconocido". De manera similar, se asignará el mismo valor "Desconocido" a taster_twitter_handle para mantener la coherencia en los datos y evitar la pérdida de información valiosa.

```
df_LimpioN = df_Limpio
df_LimpioN['taster_name'] = df_Limpio['taster_name'].fillna('Desconocido')
df_LimpioN['taster_twitter_handle'] = df_LimpioN['taster_twitter_handle'].fillna('@' + df_LimpioN['taster_name'])
```

Fuente: Elaboración Propia (TP4.py)

La columna “designation” proporciona un nivel adicional de detalle sobre el origen del vino, lo que la convierte en una variable importante para un análisis más profundo. Con casi un 30% de valores faltantes, la eliminación de estos registros podría reducir considerablemente el tamaño del dataset, lo que afectaría la robustez de cualquier análisis posterior.

Dado que “designation” es una columna categórica y no impacta directamente a las variables numéricas como “points” o “price”, agregar una categoría específica para los valores faltantes no alterará la estructura estadística del análisis. Al crear la categoría “No especificado”, se podría facilitar un análisis más completo. Por ejemplo, se podría examinar si los vinos sin designación presentan características distintas (en términos de calidad, precio, origen, etc.) en comparación con aquellos que sí tienen una designación conocida.

La categoría "No especificado" es neutral y no introduce sesgo adicional, lo que contribuirá a que el análisis sea más preciso y confiable.

```
df_LimpioN['designation'] = df_LimpioN['designation'].fillna('No especificado')
```

Fuente: Elaboración Propia (TP4.py)

La columna “region_2” es el atributo con la mayor proporción de datos faltantes, superando el 60%. Esta situación la convierte en un atributo que requiere un tratamiento cuidadoso. Para abordar este problema, se procede a analizar si existen diferencias significativas en los **precios** y **puntos** entre los vinos que tienen un valor asignado en “region_2” y aquellos que no lo tienen.

Este análisis puede proporcionar información valiosa sobre si la falta de datos en esta variable afecta otras métricas importantes, lo que puede influir en las decisiones de imputación o en la elección de estrategias de análisis posteriores.

```python
# Separar el conjunto de datos en dos subconjuntos: con y sin valores en 'region_2'
con_region_2 = df_LimpioN[df_LimpioN['region_2'].notna()]
sin_region_2 = df_LimpioN[df_LimpioN['region_2'].isna()]

# Comparar las medias de las variables 'points' y 'price'
media_con_region_2 = con_region_2[['points', 'price']].mean()
media_sin_region_2 = sin_region_2[['points', 'price']].mean()
```

```python
print("Media de 'points' y 'price' (con region_2):")
print(media_con_region_2)
```

```
Media de 'points' y 'price' (con region_2):
points     88.631625
price     370.063353
dtype: float64
```

```python
print("\nMedia de 'points' y 'price' (sin region_2):")
print(media_sin_region_2)
```

```
Media de 'points' y 'price' (sin region_2):
points     88.329622
price     331.524869
dtype: float64
```

```python
# Realizar pruebas estadísticas (prueba t) para ver si hay diferencias significativas
t_stat_points, p_val_points = stats.ttest_ind(con_region_2['points'], sin_region_2['points'], nan_policy='omit')
t_stat_price, p_val_price = stats.ttest_ind(con_region_2['price'], sin_region_2['price'], nan_policy='omit')

print(f"\nPrueba t para 'points': Estadístico t = {t_stat_points}, p-valor = {p_val_points}")
print(f"Prueba t para 'price': Estadístico t = {t_stat_price}, p-valor = {p_val_price}")

# Si los p-valores son mayores a 0.05, entonces no hay diferencias estadísticamente significativas entre los grupos.
```

```
Prueba t para 'points': Estadístico t = 17.47483794835146, p-valor = 2.668065211149369e-68
Prueba t para 'price': Estadístico t = 17.086902180757793, p-valor = 2.1909578246248457e-65
```

Fuente: Elaboración Propia (TP4.py)

Interpretación de las medias:

1. points:
    - Vinos con datos en region_2: 88.63
    - Vinos sin datos en region_2: 88.33

- La diferencia entre ambas medias es pequeña, pero notable: los vinos con información en region_2 tienen una media de puntos ligeramente mayor.

2. price:
    - Vinos con datos en region_2: 370.06
    - Vinos sin datos en region_2: 331.52

- Aquí también hay una diferencia más notable en el precio, con los vinos con datos en region_2 siendo, en promedio, más costosos

Prueba t:

- Para points, el valor p es 2.67e-68, lo cual es extremadamente bajo, indicando que la diferencia entre los puntos de los vinos con y sin region_2 es estadísticamente significativa. Aunque la diferencia en los puntos parece pequeña, la significancia estadística sugiere que region_2 puede tener una influencia sutil pero importante en la calidad percibida del vino.
- Para price, el valor p es 2.19e-65, igualmente muy bajo. Esto también indica que la diferencia en precios entre los vinos con y sin datos en region_2 es significativa, lo que sugiere que los vinos con información en region_2 tienden a ser más caros.

Conclusión del análisis de impacto:

- El análisis de impacto revela que region_2 tiene una influencia significativa tanto en la calidad (puntuación) como en el precio de los vinos. Dado que las diferencias son estadísticamente significativas, la columna region_2 no debe ser ignorada en el análisis.

Se procede a imputar los valores faltantes en region_2 basandose en otras variables geográficas como region_1 o country. Dado que region_1 y country suelen tener un vínculo directo con region_2, se utilizaran estas variables para inferir el valor más probable para los datos faltantes. Si no se consigue ninguna correlacion, se agrega la categoria "No especificado"

```python
# Imputar los valores faltantes basándote en otras variables geográficas (como region_1 o country).
df_LimpioUltima = df_LimpioN

# Función para imputar region_2 con base en region_1 y country
def imputar_region_2(row, df):
    # Si region_2 no está faltante, no hacemos nada
    if pd.notna(row['region_2']):
        return row['region_2']

    # Intentamos imputar con la moda de region_2 basada en region_1 y country
    mode_region_2 = df[(df['region_1'] == row['region_1']) & (df['country'] == row['country'])]['region_2'].mode()

    # Si encontramos una moda, la usamos, de lo contrario, devolvemos NaN
    if not mode_region_2.empty:
        return mode_region_2.iloc[0]
    else:
        mode_region_2 = df[df['country'] == row['country']]['region_2'].mode()
         if not mode_region_2.empty:
        return mode_region_2.iloc[0]
    else:
        return "No especificado"

# Aplicar la función de imputación fila por fila
df_LimpioUltima['region_2'] = df_LimpioUltima.apply(imputar_region_2, axis=1, df=df_LimpioUltima)

# Mostrar algunos ejemplos del resultado
print(df_LimpioUltima[['country', 'region_1', 'region_2']].head(10))

```

Fuente: Elaboración Propia (TP4.py)

Análisis de datos atípicos

El análisis de datos atípicos, también conocido como análisis de valores extremos o outliers, es una técnica estadística utilizada para identificar y evaluar observaciones que se desvían significativamente de otras en un conjunto de datos. Estos valores atípicos pueden influir en el análisis de datos, distorsionando resultados y conclusiones

- **Identificación de outliers**:
    - Se utilizaron gráficos de boxplot para las variables **price** y **points**, revelando una serie de valores atípicos, especialmente en **price**, donde se observaron valores extremadamente altos, con un máximo de **33,000**.

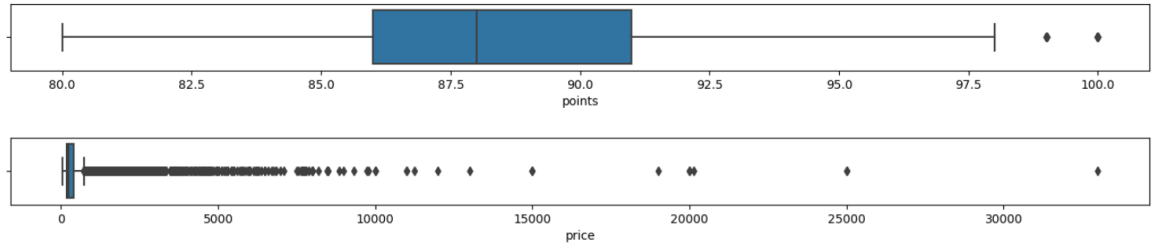
Análisis de atípicos

Graficamente

```
#cambiamos el dataset para trabahar mas facil
data = df_LimpioUltima
```

```
num_cols=data.select_dtypes(include='number').columns

for col in num_cols:
  plt.figure(figsize=(17,1))
  sns.boxplot(data=data[num_cols], x=col)
```



Numericamente

```
#Nº de Outliers usando rango IQR
outliers={}

for col in num_cols:
  Q1 = np.percentile(data[col], 25)
  Q3 = np.percentile(data[col], 75)
  IQR = Q3 - Q1
  lower_bound = Q1 - 1.5 * IQR
  upper_bound = Q3 + 1.5 * IQR
  outliers[col]= (data[col] > upper_bound).sum() + (data[col]< lower_bound).sum()

print(outliers)
```

```
{'points': 52, 'price': 9048}
```

```
#Nº de Outliers usando 3 std
outliers={}
for col in num_cols:
  mean=data[col].mean()
  std=data[col].std()

  outliers[col] = (data[col] > (mean + 3 *std)).sum() + (data[col] < (mean - 3 *std)).sum()

print(outliers)
```

```
{'points': 129, 'price': 1220}
```

Fuente: Elaboración Propia (TP4.py)

- **Evaluación de impacto de outliers**:
  - **points**: Los valores atípicos no parecen impactar significativamente la distribución general, ya que los outliers están dentro de un rango razonable (80-100 puntos).
  - **price**: Los outliers en esta variable podrían distorsionar el análisis, dado que la media es **353.63**, pero existen vinos con precios inusualmente altos. Se podría

realizar un análisis adicional para tratar estos valores, por ende realizamos una separación de dataframe, para su posterior análisis .

```
# Crear diferentes Dataframe para agrupar en una primera instancia de mejor manera los outliers
df_precios_bajos = data[data['price'] < 250]
df_precios_medios = data[(data['price'] >= 250) & (data['price'] < 11000)]
df_precios_altos = data[data['price'] > 11000]
# Mostrar la cantidad de registros en cada DataFrame
print(f"Número de registros en precios bajos: {len(df_precios_bajos)}")
print(f"Número de registros en precios medios: {len(df_precios_medios)}")
print(f"Número de registros en precios altos: {len(df_precios_altos)}")
```

```
Número de registros en precios bajos: 55390
Número de registros en precios medios: 74504
Número de registros en precios altos: 12
```

```
df_precios_bajos.describe()
```

	points	price
count	55390.000000	55390.000000
mean	88.280520	163.223145
std	3.037886	43.208296
min	80.000000	40.000000
25%	86.000000	130.000000
50%	88.000000	160.000000
75%	90.000000	200.000000
max	100.000000	240.000000

```
df_precios_medios.describe()
```

	points	price
count	74504.000000	74504.000000
mean	88.570399	479.358692
std	3.035232	406.244143
min	80.000000	250.000000
25%	86.000000	280.000000
50%	88.000000	380.000000
75%	91.000000	540.000000
max	100.000000	10000.000000

```
df_precios_altos.describe()
```

	points	price
count	12.000000	12.000000
mean	90.750000	19031.666667
std	4.750598	6384.123537
min	85.000000	11250.000000
25%	87.000000	14500.000000
50%	89.000000	19500.000000
75%	96.000000	21347.500000
max	99.000000	33000.000000

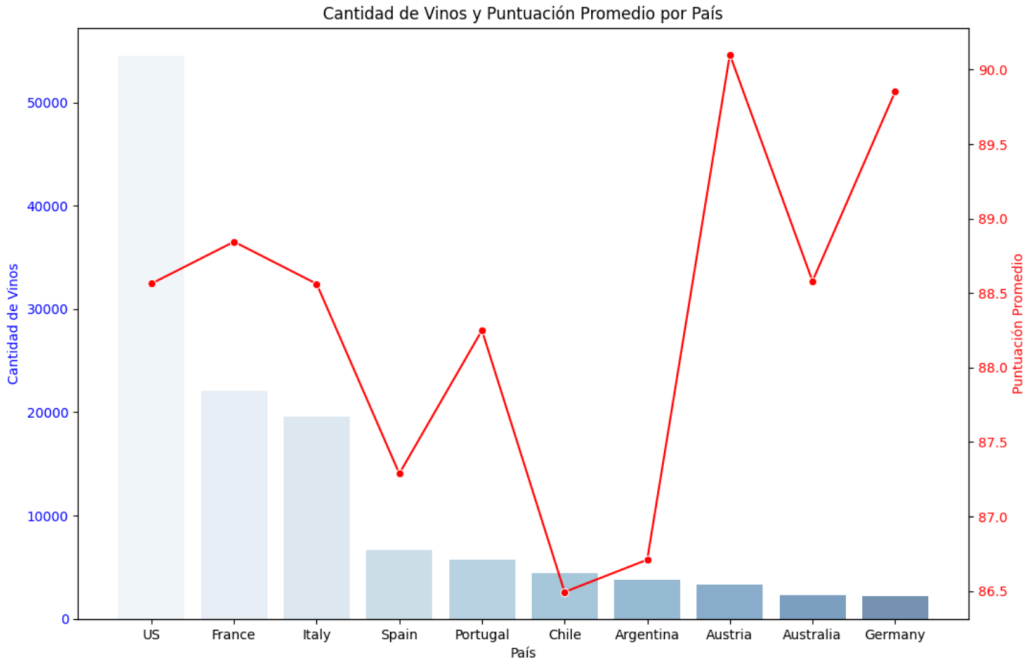
Fuente: Elaboración Propia (TP4.py)

Contextualización y preguntas clave

Una vez definidos y comprendidos los datos clave que queremos analizar para mejorar la productividad y los beneficios tanto de los productores como de los comerciantes de vinos, procedemos a un análisis más detallado. En esta etapa, es fundamental preguntarnos qué nos están diciendo los datos. Aunque puedan visualizarse de manera atractiva o con colores llamativos, los datos por sí solos no generan valor. Es necesario ubicarlos en un contexto adecuado para que puedan transformarse en *información* útil y accionable.

Preguntas Claves Cuantitativas:

¿Cuáles son los países que producen la mayor cantidad de vinos y cómo se comparan en términos de puntuación promedio? Este gráfico permite a los comerciantes y productores identificar no solo los países que tienen una gran variedad de vinos en el mercado, sino también aquellos que destacan por su calidad promedio, lo cual es útil para decisiones de importación, distribución y posicionamiento en el mercado.



Fuente: Elaboración Propia (TP4.py)

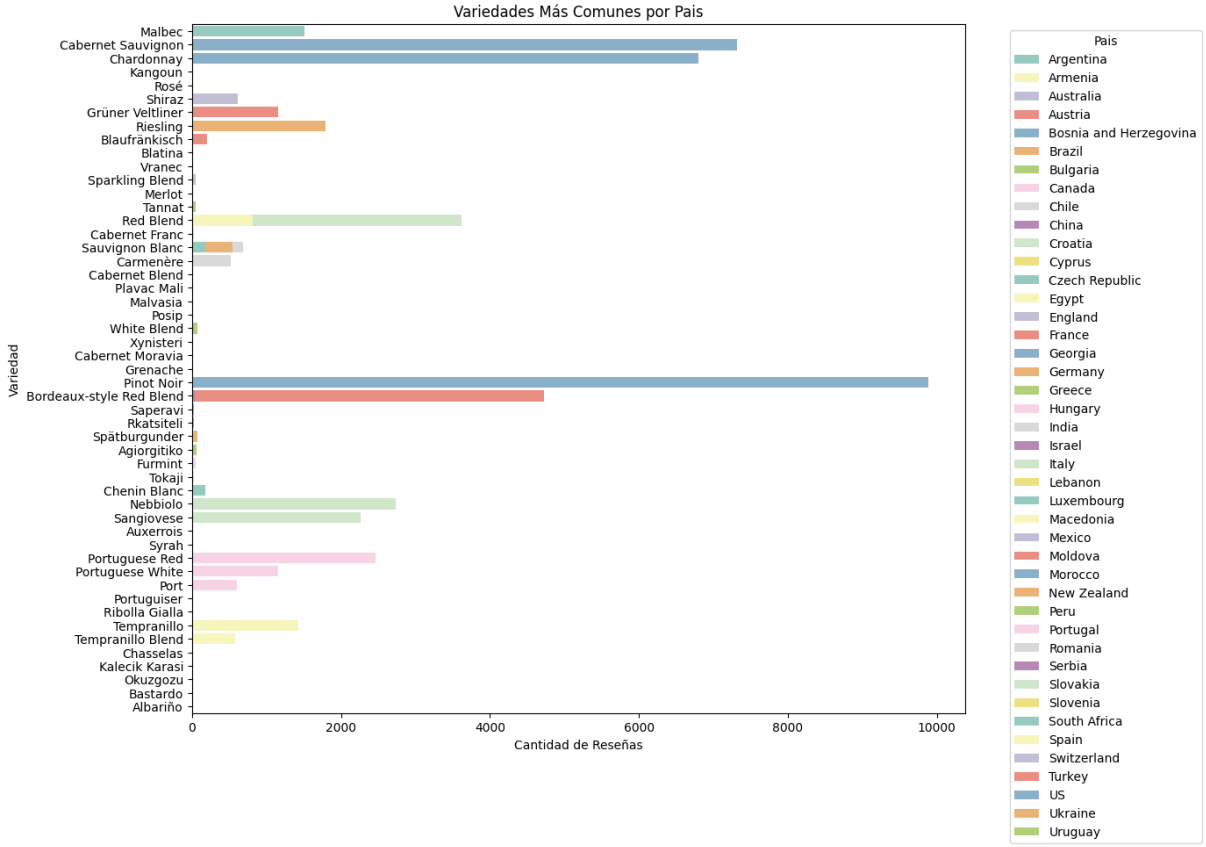
Información observada:

Cantidad de vinos: Estados Unidos es el país con la mayor cantidad de vinos en el conjunto de datos, seguido por Francia e Italia. Esto sugiere que Estados Unidos es un líder en volumen de producción de vinos o que, al menos, tiene una gran representación en las reseñas recopiladas.

Puntuación promedio: Aunque Austria y Alemania tienen una menor cantidad de vinos en comparación con países como Estados Unidos y Francia, su puntuación promedio es más alta, alcanzando los 90 puntos en promedio para Austria. Esto indica que, en promedio, los vinos de estos países son bien valorados por los críticos, a pesar de tener una representación menor.

Otros países: España, Portugal, Chile y Argentina tienen una menor cantidad de reseñas y una puntuación promedio en el rango de 87-89 puntos. Esto puede indicar que estos países tienen una buena reputación de calidad, aunque no destacan tanto en volumen de reseñas.

¿Qué variedades de uva son las más populares en cada país y que cantidad de reseñas tienen? Este gráfico proporciona información clave para productores y comerciantes al responder preguntas estratégicas sobre qué variedades de uva tienen mayor demanda por ende mayores reseñas.



Fuente: Elaboración Propia (TP4.py)

Información observada:

Para los comerciantes:

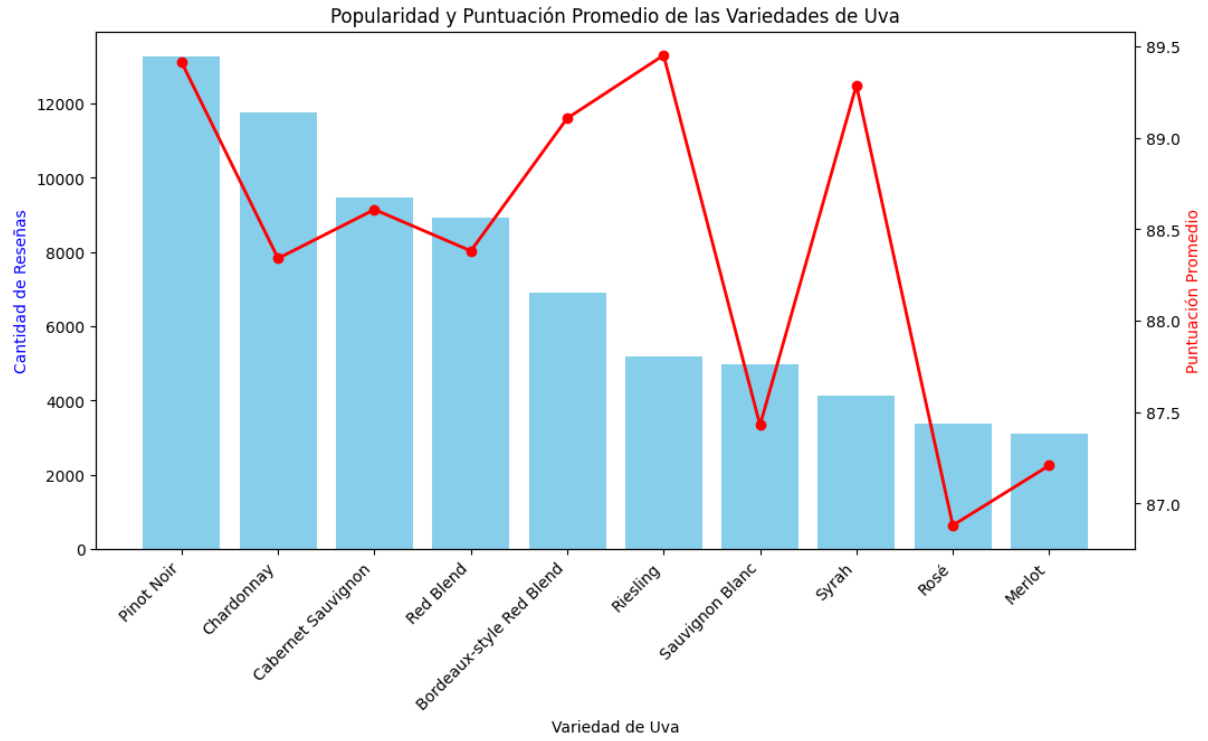
1. **Identificar variedades de alta demanda por país:**
    - Pueden enfocar sus esfuerzos en **importar y comercializar variedades populares** que ya cuentan con una alta aceptación en los mercados locales e internacionales.
    - Ejemplo: Promocionar **Malbec** de Argentina en mercados extranjeros o resaltar el **Bordeaux-style Red Blend** de Francia en tiendas especializadas.
2. **Valoración del precio según la calidad:**

- Al combinar la popularidad con la puntuación promedio, los comerciantes pueden justificar precios más altos para variedades que combinan **alta demanda y excelente calidad**.
- Ejemplo: Cobrar un precio premium por **Nebbiolo** italiano, que tiene calificaciones consistentemente superiores.

Para los productores:

1. **Reconocer la posición en el mercado global:**
   - Los productores pueden evaluar cómo su variedad insignia (por ejemplo, **Malbec** en Argentina) compite en términos de popularidad y calidad frente a otras variedades internacionales.
2. **Estrategias de mejora:**
   - Si la puntuación promedio de una variedad es inferior a la de competidores internacionales, esto indica oportunidades para **invertir en calidad** o técnicas de producción que mejoren la percepción del producto.

¿Cuáles son las variedades de uva más populares y cómo varía su puntuación promedio?
Este gráfico permite identificar oportunidades para maximizar las ganancias al equilibrar calidad percibida, demanda y estrategias de mercado.



Fuente: Elaboración Propia (TP4.py)

Información observada:

Para comerciantes:

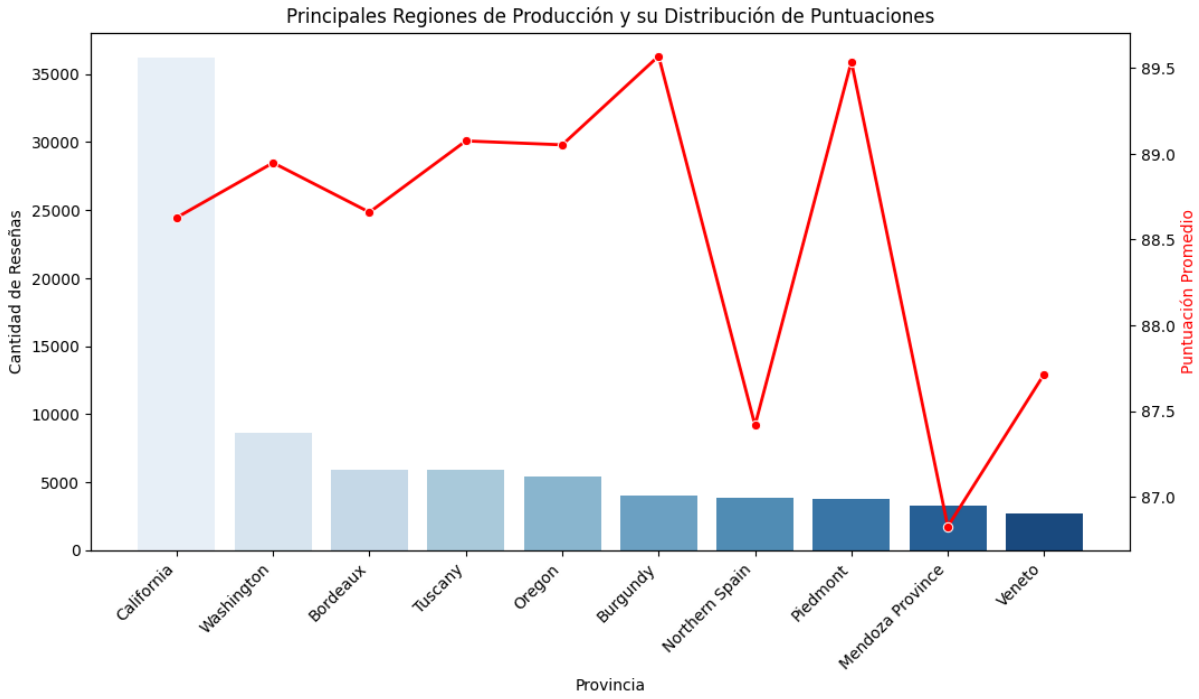
- **Promoción de vinos premium:**
  - **Bordeaux-style Red Blend** y **Riesling** pueden justificarse con precios más altos al destacar su excelente calificación promedio.
- **Saturación del mercado:**
  - Las variedades muy populares como **Pinot Noir** y **Chardonnay** podrían ser más competitivas en términos de precio, lo que requiere estrategias agresivas para diferenciar la oferta.

Para productores:

- **Enfoque en calidad:**
  - Mejorar la percepción de **Rosé** y **Merlot** a través de inversiones en técnicas de producción o branding.

- **Explotar oportunidades en variedades de nicho:**
    - Variedades como **Syrah** o **Riesling**, aunque menos populares, tienen calificaciones competitivas que pueden ser explotadas en mercados específicos.

¿Cuáles son las regiones principales de producción de vino, y cómo se distribuyen las puntuaciones de los vinos en estas regiones? Este gráfico resalta las diferencias entre volumen y calidad percibida, permitiendo decisiones estratégicas para posicionar productos en diferentes segmentos de mercado.



Fuente: Elaboración Propia (TP4.py)

Información observada:

Para comerciantes:

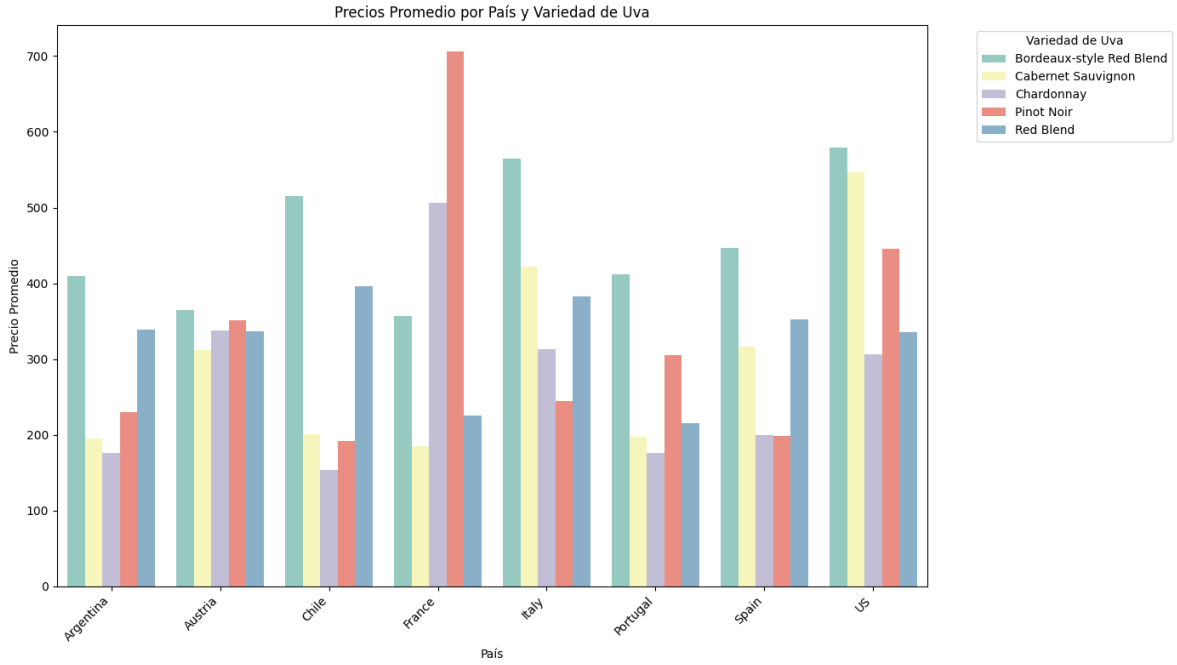
- **Regiones premium:**
    - **Bordeaux**, **Piedmont** y **Tuscany** son ideales para destacar en el segmento de lujo o vinos de alta gama, resaltando su alta calificación promedio.

- **Masividad y accesibilidad:**
    - Los vinos de **California** y **Washington** son más adecuados para estrategias de mercado masivo, debido a su alta popularidad y amplia aceptación.

Para productores:

- **Calidad percibida en Mendoza Province:**
    - Es crucial investigar por qué la percepción de calidad es más baja y considerar inversiones en enología, marketing, o enfocarse en mercados específicos.
- **Diferenciación de regiones con alto potencial:**
    - Destacar vinos de **Veneto** o regiones con menor volumen de reseñas pero buena calificación promedio puede atraer a consumidores interesados en experiencias únicas.

¿Cómo varían los precios de los vinos según su país y tipo de uva? Este gráfico ofrece una visión clara de cómo las variedades y los precios promedio varían según el país, ayudando a guiar decisiones estratégicas para maximizar el valor y las ganancias tanto para productores como para comerciantes.



Fuente: Elaboración Propia (TP4.py)

Información observada:

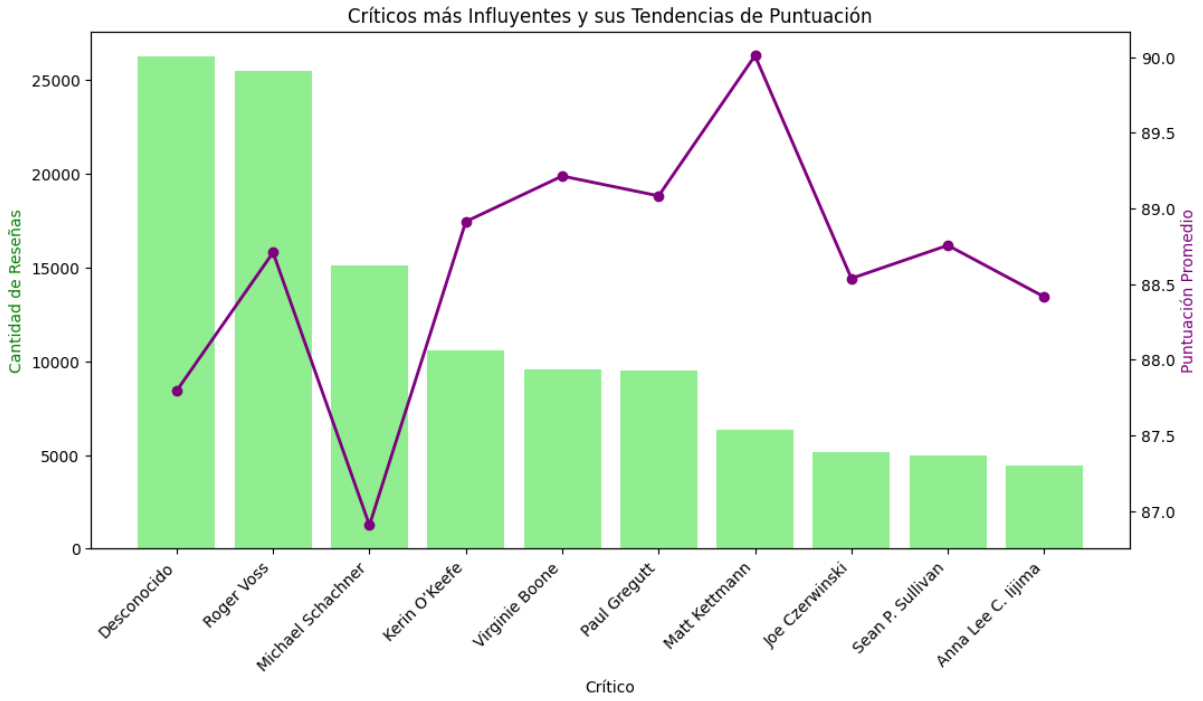
Para comerciantes:

- **Optimizar márgenes en variedades premium:**
  - En mercados como Francia e Italia, variedades como Bordeaux-style Red Blend y Pinot Noir justifican precios altos, lo que ofrece oportunidades para comercialización enfocada en segmentos de lujo.
- **Aprovechar mercados accesibles:**
  - En países como Portugal y Chile, donde los precios son más bajos, se pueden promover estas variedades como opciones de alta calidad a precios competitivos.

Para productores:

- **Incrementar el valor percibido:**
  - En regiones como Argentina y España, promover la narrativa de calidad y exclusividad de variedades como Bordeaux-style Red Blend puede justificar precios más altos.
- **Diversificación de oferta:**
  - Países con una distribución uniforme de precios, como Chile y Austria, podrían enfocarse en diversificar su portafolio con vinos premium para mercados más exigentes.

¿Qué críticos tienen la mayor influencia (cantidad de reseñas) y cuáles son sus tendencias de puntuación? Este gráfico revela el impacto de los críticos en el mercado del vino, destacando a los principales referentes y sus estilos de evaluación. La información permite a productores y comerciantes diseñar estrategias dirigidas a maximizar el valor de las reseñas, aumentar la exposición de sus vinos y construir confianza en los consumidores.



Fuente: Elaboración Propia (TP4.py)

Información observada:

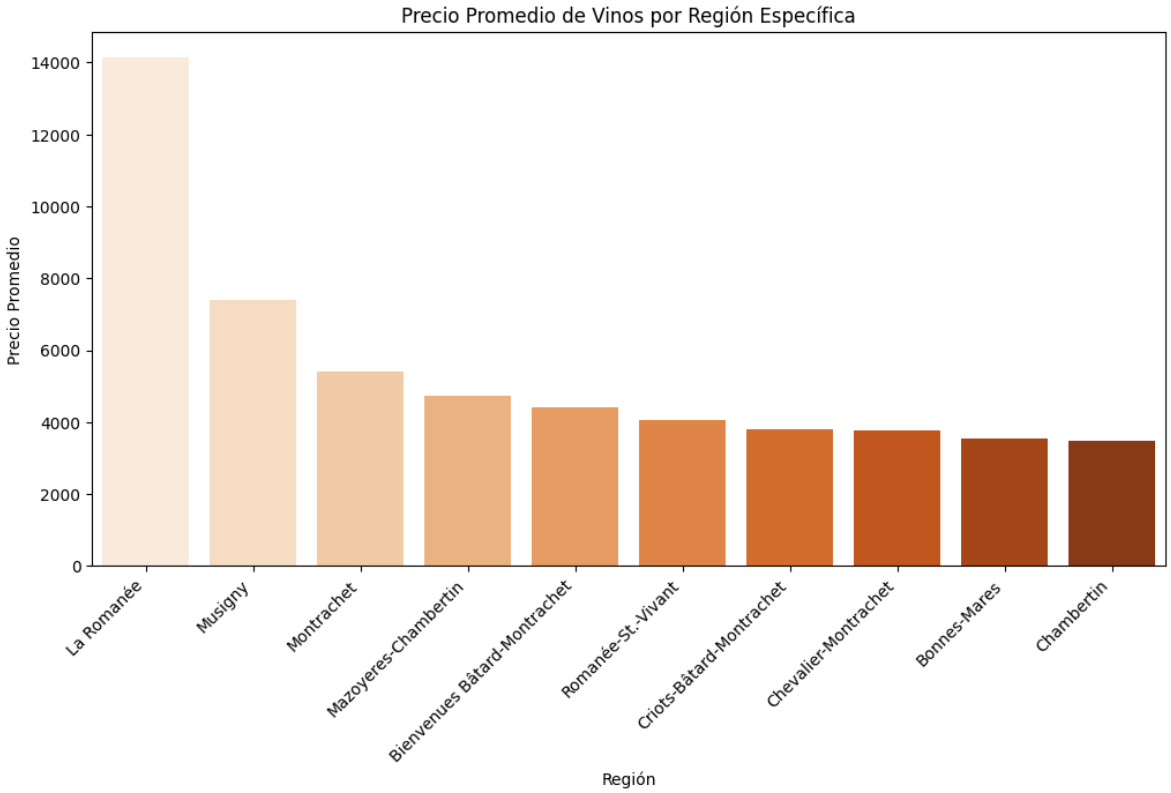
Para comerciantes:

- **Estrategias de marketing:**
  - Los vinos evaluados por críticos influyentes como **Roger Voss** o **Kerin O'Keefe** tienen un potencial de mercado más alto debido a la confianza y reconocimiento de estos críticos.
  - Resaltar etiquetas con evaluaciones de críticos con puntuaciones altas como **Matt Kettmann** puede justificar precios más altos.
- **Segmentación de mercado:**
  - Reseñas de críticos como **Michael Schachner**, quien parece evaluar una mayor variedad de vinos, pueden ser útiles para captar consumidores interesados en vinos accesibles o de calidad media.

Para productores:

- **Enfoque en críticos clave:**
    - Asegurar que los vinos lleguen a críticos influyentes como **Roger Voss** puede aumentar la visibilidad y reputación de la marca.
    - Trabajar en mejorar la calidad de los vinos que son revisados por críticos más estrictos (como **Michael Schachner**) puede ayudar a obtener mejores puntuaciones y aumentar la percepción de valor.
- **Identificar oportunidades:**
    - Críticos con menor volumen de reseñas, como **Anna Lee C. Iijima**, pueden ser una oportunidad para construir nuevas asociaciones en regiones o segmentos menos saturados.

¿Cuál es la relación entre las regiones específicas y el precio promedio de sus vinos? Este gráfico ofrece una guía para la **segmentación de mercado** y la **estrategia de precios**, subrayando las regiones clave donde se encuentran los vinos más caros y exclusivos del mundo.



Fuente: Elaboración Propia (TP4.py)

Información observada:

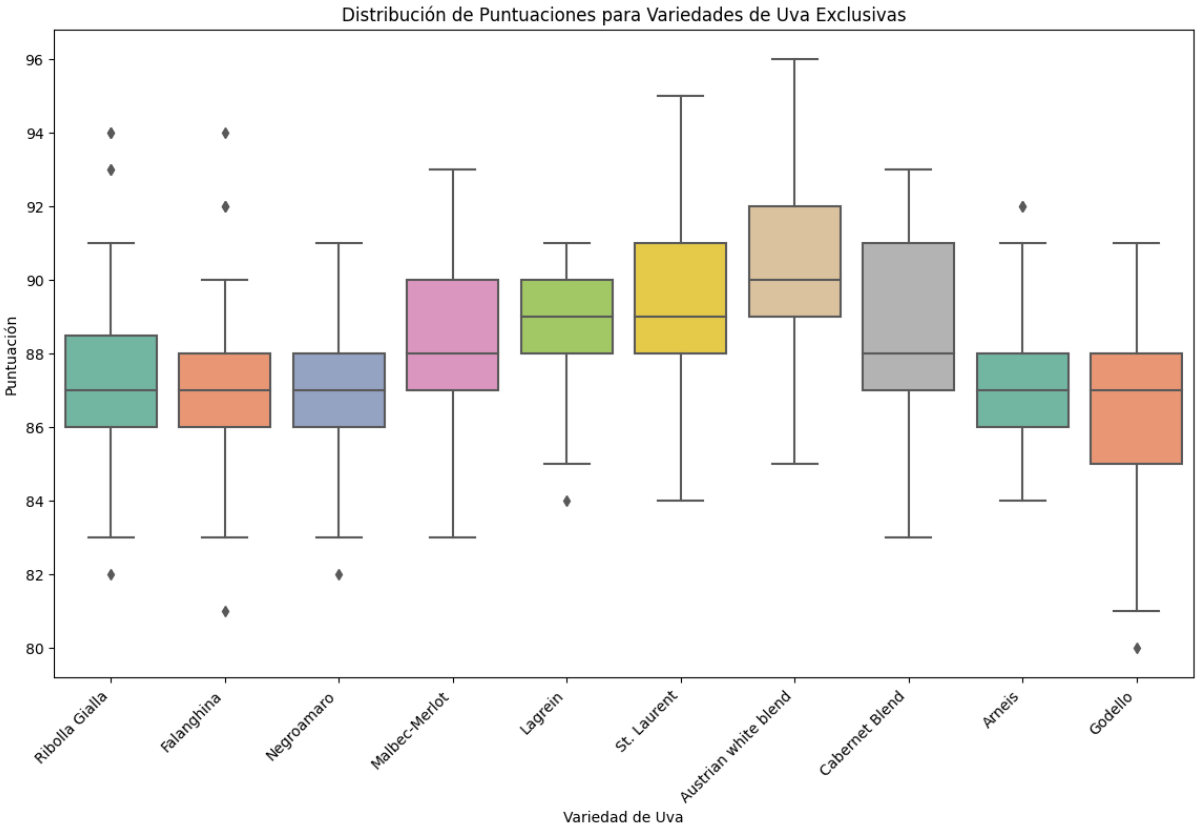
Para comerciantes:

- **Promoción de vinos de lujo:**
  - Vinos de regiones como **La Romanée**, **Musigny** y **Montrachet** son ideales para estrategias de comercialización orientadas al segmento de lujo. Resaltar su exclusividad y tradición puede justificar precios premium.
- **Segmentación del mercado:**
  - Ofrecer vinos de regiones como **Mazoyeres-Chambertin** y **Bienvenues Bâtard-Montrachet** puede atraer a consumidores interesados en vinos de alta gama, pero con un presupuesto ligeramente más moderado.

Para productores:

- **Mantener estándares de calidad:**
  - En regiones como **La Romanée** y **Musigny**, es esencial mantener la percepción de exclusividad mediante estrictos controles de calidad y baja producción.
- **Estrategias de marca:**
  - Productores de regiones menos costosas, como **Chambertin**, pueden buscar reposicionarse destacando características únicas que los diferencien de otras regiones premium.

¿Cómo se distribuyen las puntuaciones de los vinos en función de las variedades de uva más exclusivas? Este gráfico permite identificar oportunidades para maximizar el valor de las variedades exclusivas, destacando tanto fortalezas como áreas de mejora en la producción y comercialización.



Fuente: Elaboración Propia (TP4.py)

Información observada:

Para comerciantes:

- **Promoción de variedades exclusivas:**
  - Variedades como **Austrian white blend** y **Cabernet Blend** pueden ser posicionadas como opciones premium, destacando su alta puntuación promedio.
- **Diferenciación por consistencia:**
  - Variedades con puntuaciones más consistentes, como **Austrian white blend**, son ideales para consumidores que buscan calidad confiable.

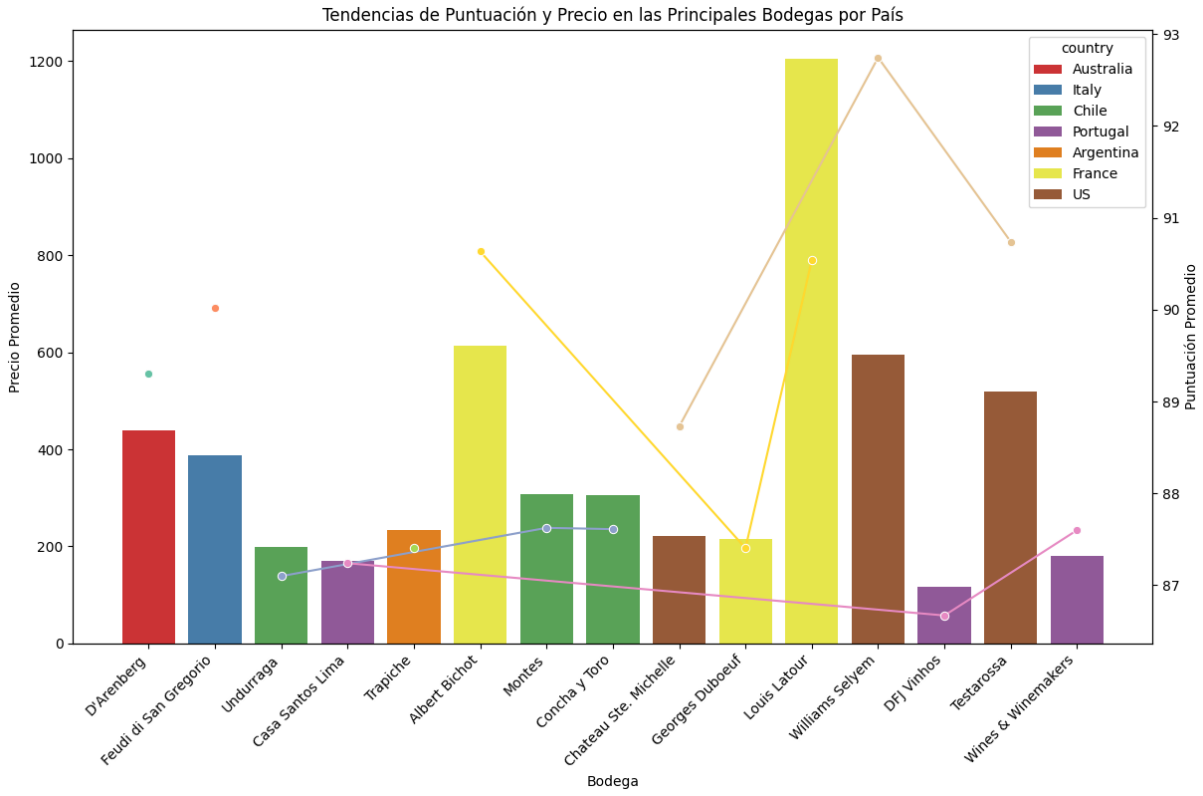
Para productores:

- **Mejorar la consistencia:**

- En variedades como **Lagrein** y **Arneis**, los productores deben enfocarse en estandarizar la calidad para reducir la variabilidad y mejorar la percepción general.

- **Estrategias de calidad:**
  - Variedades con puntuaciones bajas como **Godello** y **Falanghina** representan oportunidades para mejorar las técnicas de producción o la percepción del mercado.

¿Cuáles son las tendencias de puntuación y precio en las principales bodegas de cada país? El gráfico destaca cómo las principales bodegas equilibran precio y calidad, proporcionando una guía clara para decisiones comerciales y de producción



Fuente: Elaboración Propia (TP4.py)

Información observada:

Para comerciantes:

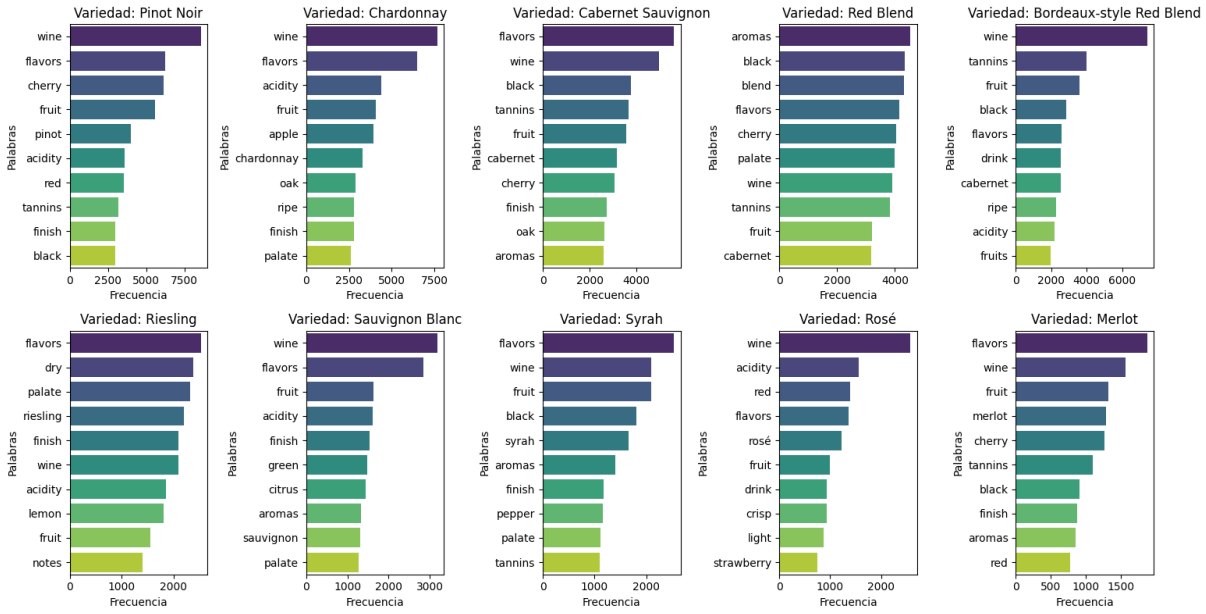
- **Promoción de vinos premium:**
    - Vinos de bodegas como **Louis Latour** y **Williams Selyem** pueden ser promovidos en el segmento de lujo, destacando su exclusividad y alta calidad.
- **Estrategias de mercado accesible:**
    - Bodegas como **Trapiche** y **Casa Santos Lima** son ideales para mercados más amplios, donde la relación calidad-precio es crucial para atraer consumidores.

Para productores:

- **Mejorar el posicionamiento:**
    - Bodegas como **Montes** y **Undurraga**, que ya tienen puntuaciones competitivas, pueden enfocarse en estrategias de branding para justificar un aumento en los precios.
- **Optimización de calidad:**
    - Bodegas con puntuaciones más bajas, como **DF Vinhos**, pueden invertir en mejorar procesos de producción o destacar atributos únicos para elevar su percepción de calidad.

Preguntas Claves Cualitativas:

¿Qué características se destacan más en cada variedad de vino? Este gráfico permite a productores y comerciantes comprender cómo perciben los críticos cada variedad de vino. Esto no solo ayuda a mejorar los procesos de producción, sino también a comunicar los puntos clave de venta de cada producto de manera más efectiva, alineándose con las expectativas del consumidor informado.



Fuente: Elaboración Propia (TP4.py)

Información observada:

Implicaciones para productores:

1. **Enfoque en las fortalezas de cada variedad:**
    - Los productores pueden usar esta información para refinar sus vinos y resaltar las características que los críticos consideran importantes. Por ejemplo, al producir Pinot Noir, deben enfocarse en mantener un equilibrio en la acidez y potenciar los sabores de frutas rojas.
2. **Desarrollo de estrategias de diferenciación:**
    - Si varias variedades comparten ciertos descriptores (como "fruit" o "flavors"), los productores pueden buscar atributos únicos para diferenciar sus productos en el mercado.

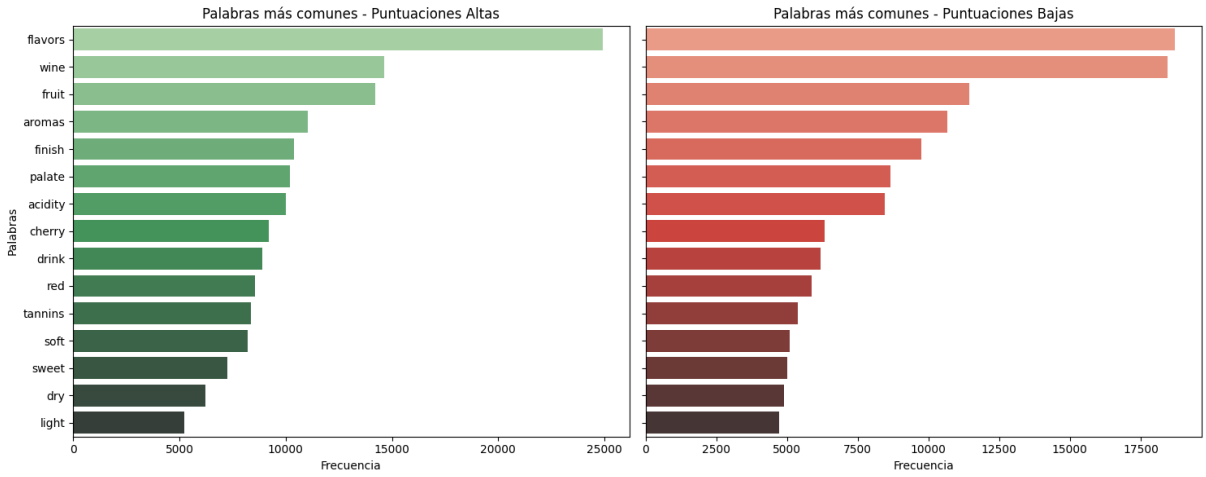
Implicaciones para comerciantes:

1. **Optimización del marketing:**
    - Las palabras más frecuentes pueden servir como guías para redactar descripciones atractivas en etiquetas o materiales promocionales. Por ejemplo,

destacar términos como "elegante acidez" o "taninos refinados" en descripciones de vinos puede atraer a consumidores informados.

2. **Segmentación de mercado:**
    - Los comerciantes pueden usar estos descriptores para orientar sus estrategias de venta a diferentes segmentos. Por ejemplo, los consumidores que prefieren vinos secos pueden ser dirigidos hacia variedades como Riesling o Sauvignon Blanc.

¿Qué descriptores están más asociados con vinos altamente puntuados? Este gráfico destaca cómo los críticos utilizan un lenguaje diferenciado según la calidad percibida de los vinos. Los productores y comerciantes pueden usar estos datos para refinar sus productos y estrategias de comunicación, asegurándose de que los descriptores clave estén alineados con las expectativas del consumidor objetivo.



Fuente: Elaboración Propia (TP4.py)

Información observada:

Implicaciones para productores:

1. **Refinamiento de atributos clave:**

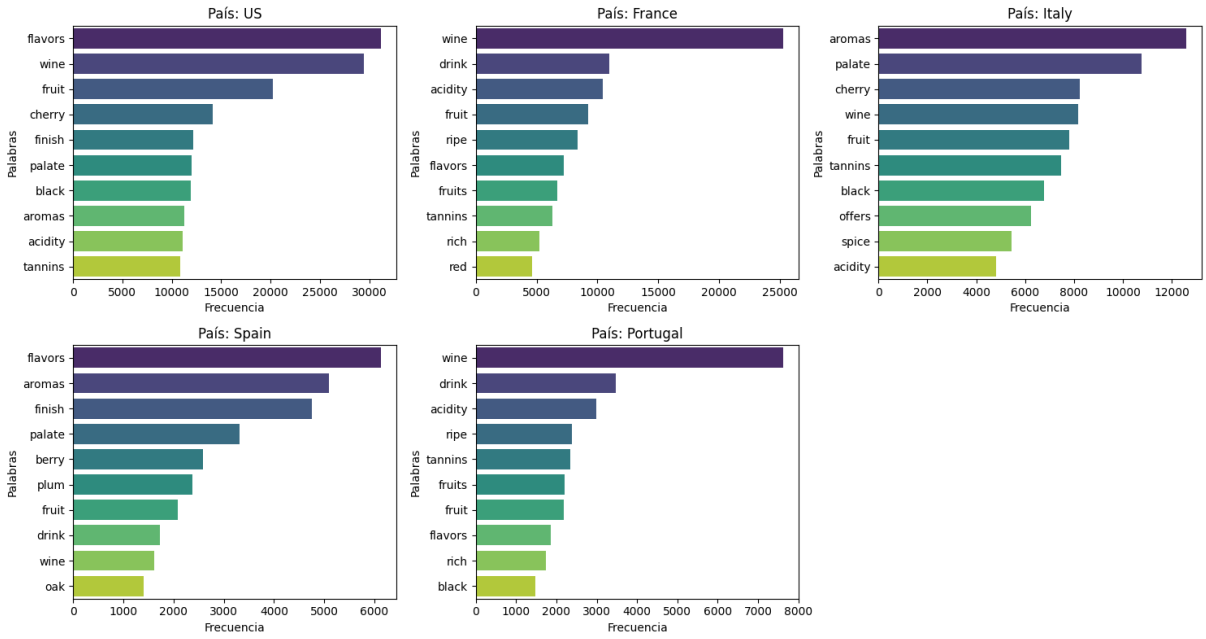
- Los productores de vinos de alta gama pueden enfocarse en mejorar los atributos asociados con puntuaciones altas, como un balance adecuado entre acidez y dulzura, perfiles aromáticos complejos y un final persistente.
- Los productores de vinos con puntuaciones más bajas pueden usar estos datos para identificar áreas de mejora, como agregar mayor complejidad o intensidad.

2. **Optimización del proceso de producción:**
   - Palabras como "finish" y "aromas" sugieren que los consumidores valoran la experiencia completa del vino, desde el aroma inicial hasta el sabor persistente, lo que podría guiar inversiones en técnicas de vinificación y selección de uvas.

Implicaciones para comerciantes:

1. **Segmentación del mercado:**
   - Los vinos con puntuaciones altas pueden ser promocionados utilizando términos como "complejo", "persistente" o "aromático" para atraer a consumidores premium.
   - Para vinos con puntuaciones más bajas, los comerciantes pueden enfocarse en términos como "ligero" o "refrescante", que podrían atraer a consumidores ocasionales o a quienes buscan vinos fáciles de beber.
2. **Estrategias de marketing:**
   - Utilizar términos positivos asociados con puntuaciones altas en etiquetas y descripciones puede mejorar la percepción de calidad y justificar precios más altos.
   - Identificar mercados específicos para vinos más ligeros y secos podría permitir aprovechar segmentos más accesibles del mercado.

¿Cómo varían las descripciones según la región geográfica? Este gráfico destaca las diferencias estilísticas en los vinos según su país de origen. Los productores pueden usar estos descriptores para afinar sus productos y resaltar las características más apreciadas. Los comerciantes, por otro lado, pueden utilizar esta información para crear campañas de marketing más efectivas y segmentar su oferta según el mercado objetivo.



Fuente: Elaboración Propia (TP4.py)

Información observada:

Implicaciones para productores:

1. **Adaptación al mercado local e internacional:**
    - Los productores pueden ajustar el perfil de sus vinos según los descriptores más valorados en su país. Por ejemplo, en Italia podrían resaltar aromas especiados y acidez en sus etiquetas y estrategias de marketing.
2. **Enfoque en fortalezas tradicionales:**
    - Países como Francia e Italia pueden continuar destacando atributos clásicos como taninos refinados y complejidad aromática para mantener su posición en el mercado global.

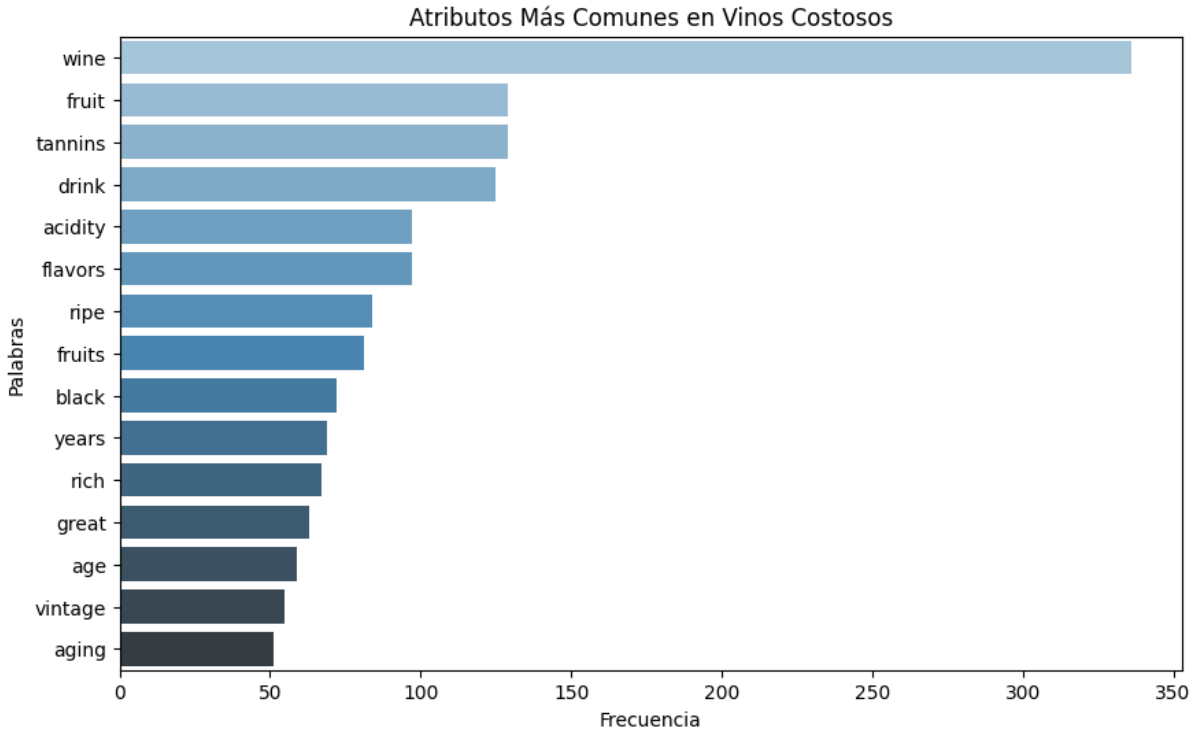
Implicaciones para comerciantes:

1. **Estrategias de marketing regionalizadas:**
    - Los comerciantes pueden destacar términos clave específicos de cada país para apelar a consumidores interesados en estilos particulares. Por ejemplo,

"sabores maduros y ricos" para vinos franceses o "notas especiadas y frescas" para vinos italianos.

2. **Segmentación de mercado:**
    - Este análisis permite identificar preferencias de los consumidores en diferentes mercados. Por ejemplo, los amantes de vinos frutales podrían ser más receptivos a vinos estadounidenses o españoles.

¿Qué atributos destacan los críticos en vinos más costosos? Este análisis resalta que los vinos costosos son apreciados por su complejidad, estructura y proceso de elaboración detallado. Los productores pueden centrarse en ofrecer características premium, mientras que los comerciantes pueden usar estos atributos como puntos clave para justificar el precio y atraer a consumidores exigentes.



Fuente: Elaboración Propia (TP4.py)

Información observada:

Implicaciones para productores:

1. **Enfoque en la calidad de la uva y el envejecimiento:**
    - Los productores deben priorizar el uso de uvas de alta calidad y procesos de envejecimiento adecuados para producir vinos con perfiles complejos y maduros, resaltando notas frutales y taninos equilibrados.
2. **Diferenciación mediante la añada:**
    - Resaltar la añada y el proceso de envejecimiento en las etiquetas y estrategias de marketing puede atraer a consumidores dispuestos a pagar un precio más alto.

Implicaciones para comerciantes:

1. **Estrategias de venta basadas en el envejecimiento:**
    - Los comerciantes pueden destacar en sus campañas términos relacionados con el envejecimiento ("vintage", "age", "aging") y notas de calidad ("rich", "ripe") para justificar los precios elevados de ciertos vinos.
2. **Segmentación del mercado:**
    - Identificar consumidores que valoren vinos con atributos técnicos y procesos refinados puede ayudar a optimizar estrategias de promoción y segmentación.

Conclusiones

El análisis del conjunto de datos de reseñas de vinos se llevó a cabo con el objetivo de identificar patrones, tendencias y relaciones clave entre las diferentes entidades involucradas, como regiones, variedades, bodegas, precio y críticos. Este informe busca sintetizar los insight claves obtenidos a partir del proceso de análisis. En una primera instancia desde el entendimiento del contexto, para saber a qué tipo de información queremos llegar y que datos tenemos disponible, posteriormente pasando por el Análisis Exploratorio de Datos (EDA por su sigla en inglés), para comprender las características y la estructura del conjunto de datos, permitiendo identificar patrones y tendencias a través de visualizaciones y resúmenes estadísticos, descubriendo relaciones y patrones relevantes entre las variables.

Análisis de Resultados

El análisis acerca de las reseñas de vinos permitió identificar patrones y relaciones clave que ofrecen información estratégica tanto para productores como para comerciantes. A nivel de países, Estados Unidos y Francia destacan como líderes en volumen de reseñas, mientras que Austria y Alemania sobresalen con puntuaciones promedio más altas, alrededor de los 90 puntos. Esto sugiere que, mientras los países con altos volúmenes representan oportunidades masivas, aquellos con altas calificaciones tienen un fuerte potencial para posicionarse como productores de vinos premium.

En cuanto a las variedades de uva, **Pinot Noir**, **Chardonnay** y **Cabernet Sauvignon** son las más populares, consolidándose como opciones tradicionales en el mercado. Sin embargo, variedades como **Bordeaux-style Red Blend** equilibran popularidad y calidad, justificando su inclusión en estrategias premium. Las regiones específicas también reflejan una clara segmentación del mercado. **La Romanée** y otras regiones francesas lideran con precios promedio superiores a los 14,000 USD, demostrando cómo el prestigio de una región impulsa los precios. Estas regiones exclusivas representan un modelo a seguir para otras que buscan establecer una reputación de lujo.

El análisis de los críticos reveló que figuras influyentes como **Roger Voss** tienen un impacto significativo en el mercado debido a su alto volumen de reseñas. Además, críticos como **Matt Kettmann**, con puntuaciones promedio más altas (~90 puntos), representan una oportunidad estratégica para mejorar la visibilidad de las marcas. Por otro lado, la distribución de puntuaciones en variedades de uva exclusivas, como **Austrian White Blend** y **Cabernet Blend**, reflejan consistencia en la calidad, mientras que variedades como **Lagrein** y **Godello** muestran mayor variabilidad, sugiriendo que los productores deben trabajar en estandarizar la calidad.

En cuanto a las bodegas, los datos muestran que, aunque Francia y Estados Unidos lideran en precios promedio, regiones como Chile y Portugal mantienen puntuaciones competitivas incluso con precios más bajos. Este equilibrio entre precio y calidad subraya la

importancia de posicionar estratégicamente vinos de alta calidad en segmentos de mercado accesibles.

El análisis de reseñas de vinos ha revelado importantes conclusiones que pueden guiar tanto a productores como a comerciantes en la optimización de sus estrategias. Una de las principales observaciones es la segmentación clara del mercado. Mientras que los mercados premium, como Francia y Estados Unidos, ofrecen oportunidades de ingresos significativos, es crucial justificar los precios elevados mediante estrategias de branding y calidad. Por otro lado, los mercados emergentes, como Portugal y Argentina, deben priorizar la percepción de calidad mientras mantienen precios competitivos para atraer consumidores.

Recomendaciones para Productores y Comerciantes

Este análisis proporciona una serie de recomendaciones clave que pueden ayudar tanto a productores como a comerciantes a optimizar sus estrategias y maximizar su impacto en el mercado.

Para los productores, es esencial estandarizar la calidad en variedades con alta dispersión en las puntuaciones, como **Lagrein** y **Godello**, invirtiendo en procesos de producción más consistentes que reduzcan la variabilidad y fortalezcan la confianza de los consumidores. Al mismo tiempo, variedades con calificaciones consistentes y altas, como **Austrian White Blend**, deben servir como referencia para replicar prácticas exitosas que garanticen una percepción de calidad uniforme.

Las regiones exclusivas, como **La Romanée** y **Musigny**, ofrecen un gran potencial para mantener su prestigio a través de controles estrictos de calidad y estrategias de marketing que refuercen su exclusividad. Por otro lado, para regiones menos reconocidas, como algunas en Argentina o Portugal, es crucial desarrollar una identidad única que permita diferenciarse en mercados internacionales. Además, colaborar con críticos influyentes, como **Roger Voss** o **Matt Kettmann**, puede ser una herramienta estratégica para aumentar la visibilidad de las marcas y mejorar su posicionamiento en el mercado global.

Los comerciantes, por su parte, deben enfocarse en segmentar el mercado de manera eficiente. Las variedades populares como **Pinot Noir** y **Cabernet Sauvignon** son ideales para mercados masivos, mientras que variedades de alta calificación como **Bordeaux-style Red Blend** pueden posicionarse en segmentos de lujo. En cuanto a las regiones, aquellas con prestigio histórico, como **Montrachet** o **Mazoyeres-Chambertin**, justifican estrategias de precios premium al destacar atributos únicos y altos puntajes promedio. Por el contrario, regiones emergentes como Mendoza o Alsacia ofrecen oportunidades para competir en el mercado a través de precios accesibles y una percepción creciente de calidad.

Además, es importante diseñar experiencias atractivas para los consumidores, como catas dirigidas o eventos centrados en variedades exclusivas, lo que no solo aumenta el interés en etiquetas menos conocidas, sino que también fomenta la fidelidad a la marca. Los comerciantes también deben aprovechar las reseñas de críticos reconocidos como una herramienta de marketing, destacando las calificaciones promedio en campañas para diferenciarse de la competencia y aumentar la confianza del consumidor.

Por último, la optimización del inventario se convierte en una prioridad. Es fundamental equilibrar la oferta de vinos de alta demanda, como **Chardonnay** y **Cabernet Sauvignon**, con opciones premium de alta calificación, asegurando así que ambos segmentos estén representados de manera adecuada para atender a diversos perfiles de consumidores.

Este análisis subraya la importancia de la toma de decisiones basada en datos. Visualizar patrones clave y comprender la relación entre variables como precio, puntuación, región y variedades permite a los actores de la industria del vino diseñar estrategias dirigidas y efectivas.

Links Importantes

Data Set

Se adjunta un enlace a Google drive con el Data Set sobre el que se realiza el proyecto: **DataSet.**

Archivo .py

Se adjunta un enlace a Google drive con el código en python sobre el que se realiza el proyecto: **Python.**

Archivo .ipynb

También se adjunta un enlace a Google drive con el código en ipynb, por si se quiere abrir en Google Colab o Jupyter Notebook sobre el que se realiza el proyecto: Colab**.**