SEMINARIO DE PRÁCTICA EN CIENCIA DE DATOS

Título: Trabajo Práctica N° 4

Nombre Autor: Lima, Jonatan Ezequiel

Fecha: 16/11/2024

Índice

1. Introducción

En esta entrega se realizaron análisis exploratorios de los datos con el objetivo de proporcionar información relevante que apoye la toma de decisiones, específicamente en el contexto de los comerciantes y productores de vino.

Para ello, se llevarán a cabo análisis estadísticos y descriptivos, examinando las relaciones entre las variables para identificar patrones o correlaciones significativas. Además, el análisis será complementado con gráficos que permitan visualizar y comprender mejor el comportamiento de los datos.

Para este trabajo práctico, se eligió utilizar el lenguaje de programación Python y se empleará el editor de código Visual Studio Code.

El conjunto de datos corresponde al archivo CSV proporcionado por kangle, url: Wine Reviews (kaggle.com).

2. Selección y depuración de datos

Para ejecutar los scripts, es necesario tener el archivo CSV 'winemag-data-130k-v2.csv' descargado desde la dirección proporcionada anteriormente. Este archivo debe colocarse en la misma carpeta que el archivo 'analisis_exploratorio_graficos_vinos.py' antes de proceder a su ejecución.

El primer paso consistió en la importaciones de las siguientes librerías:

- *pandas:* Utilizada para la manipulación y análisis de datos.
- *numpy:* Una biblioteca fundamental para el cálculo numérico en Python.
- *seaborn:* Una biblioteca basada en Matplotlib que facilita la creación de visualizaciones atractivas.
- *matplotlib.pyplot:* Una colección de funciones que permiten crear gráficos en 2D.
- *keyboard: Una biblioteca que permite controlar y monitorear eventos.*

En caso de no contar con las librerías, a continuación se presentan los comandos para su instalación directamente desde la línea de comandos (cmd):

- pip install pandas
- pip install numpy
- pip install seaborn
- pip install matplotlib
- pip install keyboard

Al igual que en las entregas anteriores, se eliminaron las columnas 'Unnamed: 0' y 'region_2'. La primera fue descartada por contener únicamente un índice innecesario, y la segunda, debido a que presentaba una gran cantidad de valores faltantes (NaN).

Figura 1: Total de datos nulos (NaN) de cada columna

Cantidad de valores NaN en cada variable:

	Variable	Total de NaN
0	region_2	79460
1	designation	37465
2	taster_twitter_handle	31213
3	taster_name	26244
4	region_1	21247
5	price	8996
6	province	63
7	country	63
8	variety	1
9	Unnamed: 0	0
10	points	0
11	description	0
12	title	0
13	winery	0

Al igual que en entregas anteriores, se eliminaron las filas con datos faltantes en las variables 'country' y 'province'. Esto se debe a que ambas variables están relacionadas, y la ausencia de datos en 'country' implica también su ausencia en 'province'. Esta eliminación afecta a un total de 63 filas, lo que representa el 0.05% del conjunto de datos.

Se realizó un análisis para identificar posibles filas duplicadas, encontrándose un total de 19,958 filas repetidas, lo que equivale al 15.36% del total. Por lo tanto, se

procedió a eliminar estas filas, asegurándose de que todos los campos coincidieran exactamente para evitar la eliminación de filas incorrectas.

Figura 2: Informe general del DataFrame, después de realizar la limpieza de datos

```
<class 'pandas.core.frame.DataFrame'>
Index: 119929 entries, 0 to 129970
Data columns (total 12 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   country                119929 non-null  object
 1   description            119929 non-null  object
 2   designation            85394 non-null   object
 3   points                 119929 non-null  int64
 4   price                  111538 non-null  float64
 5   province               119929 non-null  object
 6   region_1               100428 non-null  object
 7   taster_name            95012 non-null   object
 8   taster_twitter_handle  90483 non-null   object
 9   title                  119929 non-null  object
 10  variety                119928 non-null  object
 11  winery                 119929 non-null  object
dtypes: float64(1), int64(1), object(10)
memory usage: 11.9+ MB
```

Después de eliminar las 63 filas con valores faltantes en 'country' y 'province', las columnas 'region_2' y 'Unnamed: 0', así como las filas duplicadas, el DataFrame final quedó con un total de 119,929 filas y 12 columnas.

3. Contextualización y preguntas claves

En este análisis se considerarán dos grupos principales: comerciantes y productores. Para cada uno de ellos, se formularon preguntas clave con el objetivo de proporcionar datos e información útil para la toma de decisiones.

A partir de esto, se presentarán primero las preguntas correspondientes a cada sector, seguidas de los resultados obtenidos.

Comerciantes

1. *¿Cuáles son los países con los mejores puntajes promedio en vinos?*

   a. Identificar los países con las mejores puntuaciones promedio en vinos proporcionará información valiosa para promocionar estos productos como los de mayor calidad.

2. *¿Cuál es el país con la mejor relación precio-calidad en los vinos?*

   a. Identificar los vinos que ofrecen un precio accesible sin comprometer su calidad podría representar una ventaja competitiva significativa para los comerciantes.

3. *¿Cuáles son los vinos con más reseñas y cómo se distribuyen en términos de calidad y precio?*

   a. Analizar esta relación permite comprender las diferentes perspectivas de los catadores sobre un vino y evaluar si la percepción de su calidad varía significativamente según las reseñas realizadas.

4. *¿Qué relación existe entre el precio y el país de origen del vino?*

   a. Comprender esta relación podría ofrecer una visión sobre qué vinos son más accesibles en términos de costo y calidad en diferentes regiones.

Productores

1. *¿Cuáles son los varietales con los mejores puntajes globales?*

   a. Conocer los varietales mejor puntuados ayudará a tomar decisiones informadas sobre qué varietal elegir para invertir en su cultivo.

2. *¿Cuál es el varietal más utilizado por país y cuál es su puntaje promedio y variabilidad?*

   a. Conocer el varietal más utilizado por país proporcionará información sobre las preferencias y tendencias en los diferentes mercados de vino. Además, analizar el puntaje promedio y la variabilidad de estos varietales ofrecerá una visión más completa sobre la calidad y consistencia de los vinos.

3. *¿Cómo se relaciona el puntaje del vino con el año de producción?*

    a. Analizar el año de producción de cada vino y su relación con el puntaje obtenido será clave para entender si la evolución con el tiempo influye en la calidad del vino, según las reseñas.

4. *¿Existen variaciones significativas en el puntaje entre las provincias de cada país?*

    a. Comparar los puntajes promedio entre provincias podría revelar si algunas regiones se destacan en calidad, lo que sería útil para mejorar la segmentación del mercado.

Una vez formuladas las preguntas y explicado su propósito, se procederá a listar los resultados obtenidos de los diferentes análisis realizados:

Comerciantes

1. *¿Cuáles son los países con los mejores puntajes promedio en vinos?*

Figura 3: Países con los mejores puntajes en promedio.

Países con los mejores puntajes promedio y su desviación estándar:

	country	points_mean	points_std
0	England	91.75	1.84
1	India	90.25	1.83
2	Austria	90.13	2.54
3	Germany	89.90	2.50
4	Canada	89.31	2.45
5	Hungary	89.29	2.78
6	China	89.00	NaN
7	France	88.87	3.09
8	Luxembourg	88.80	0.84
9	Morocco	88.71	1.73
10	Italy	88.58	2.70
11	Australia	88.56	3.03
12	US	88.56	3.17
13	Switzerland	88.50	2.74
14	Israel	88.45	2.50
15	New Zealand	88.30	2.47
16	Portugal	88.26	3.05
17	Slovenia	88.13	1.74
18	Turkey	88.07	1.98
19	South Africa	88.03	2.46
20	Bulgaria	87.88	2.10
21	Serbia	87.62	1.30
...			
39	Brazil	84.63	2.40
40	Ukraine	84.07	1.59
41	Egypt	84.00	NaN
42	Peru	83.56	1.86

2. *¿Cuál es el país con la mejor relación precio-calidad en los vinos?*

Figura 4: Países y su relación precio-calidad.

Paises y su relación precio-calida:

	country	precio_calidad
0	Ukraine	9.67
1	Romania	8.96
2	Bulgaria	6.95
3	Bosnia and Herzegovina	6.94
4	India	6.90
5	Moldova	6.60
6	Armenia	6.04
7	Chile	5.78
8	Peru	5.63
9	Macedonia	5.59
10	Portugal	5.50
11	Cyprus	5.44
12	Slovakia	5.44
13	Georgia	5.19
14	Argentina	5.19
15	South Africa	5.03
16	Spain	4.98
17	Morocco	4.97
18	China	4.94
19	Greece	4.66
20	Slovenia	4.51
21	Czech Republic	4.50
...		
38	US	3.35
39	Canada	3.15
40	Switzerland	2.30
41	England	1.89

3. *¿Cuáles son los vinos con más reseñas y cómo se distribuyen en términos de calidad y precio?*

Figura 5: Lista de vinos con sus puntajes promedio y variaciones en puntos y precio

```
Resultados de los vinos con sus puntajes promedio obtenidos y la cantidad de reseñas realizadas:

                                                          title  \
51641              Gloria Ferrer NV Sonoma Brut Sparkling (Sonoma County)
97293                        Segura Viudas NV Extra Dry Sparkling (Cava)
97288             Segura Viudas NV Aria Estate Extra Dry Sparkling (Cava)
57903   J Vineyards & Winery NV Brut Rosé Sparkling (Russian River Valley)
51637                Gloria Ferrer NV Blanc de Noirs Sparkling (Carneros)
...                                                                  ...
118760       àMaurice 2011 Amparo Estate Malbec (Walla Walla Valley (WA))
118761              àMaurice 2011 Boushey Vineyard Syrah (Yakima Valley)
118762          àMaurice 2011 Fred Estate Syrah (Walla Walla Valley (WA))
118763       àMaurice 2011 Gamache Vineyard Malbec (Columbia Valley (WA))
5                         10 Knots 2006 Chardonnay (Santa Barbara County)

        mean_points  std_points  mean_price  std_price  num_reseñas
51641         88.56        2.13       21.33       1.41            9
97293         85.57        1.27       10.00       0.00            7
97288         84.14        2.12       13.43       0.98            7
57903         87.67        1.21       36.50       5.54            6
51637         88.33        2.07       22.00       1.26            6
...             ...         ...         ...        ...          ...
118760        88.00         NaN       47.00        NaN            1
118761        92.00         NaN       38.00        NaN            1
118762        89.00         NaN       45.00        NaN            1
118763        93.00         NaN       35.00        NaN            1
5             85.00         NaN       21.00        NaN            1

[118781 rows x 6 columns]
```

4. *¿Qué relación existe entre el precio y el país de origen del vino?*

Figura 6: País y su variación de precios

```
País y su variación de precios:

                         count   mean    std   min    25%   50%    75%  \
country
Argentina               3501.0  24.58  23.78   4.0  12.00  17.0  25.00
Armenia                    2.0  14.50   0.71  14.0  14.25  14.5  14.75
Australia               2169.0  35.69  50.14   5.0  15.00  21.0  38.00
Austria                 2536.0  31.34  28.25   7.0  19.00  25.0  38.00
Bosnia and Herzegovina     2.0  12.50   0.71  12.0  12.25  12.5  12.75
Brazil                    44.0  24.50  11.04  10.0  15.00  22.0  30.25
Bulgaria                 132.0  14.84   9.78   8.0  10.00  14.0  16.00
Canada                   224.0  35.78  19.80  12.0  21.00  30.0  45.00
Chile                   4130.0  20.85  22.15   5.0  12.00  15.0  20.00
China                      1.0  18.00    NaN  18.0  18.00  18.0  18.00
Croatia                   68.0  25.53  13.23  12.0  16.75  20.0  27.25
Cyprus                    10.0  16.50   2.88  11.0  15.25  16.5  17.75
Czech Republic            11.0  22.36  10.62  15.0  15.00  18.0  24.50
England                   59.0  52.68  15.41  25.0  43.00  50.0  60.00
France                 16286.0  41.99  76.55   5.0  16.00  25.0  45.00
Georgia                   74.0  19.36   7.58   9.0  14.00  17.5  25.00
Germany                 1951.0  43.46  65.16   5.0  18.00  27.0  43.50
Greece                   427.0  22.25  10.68   8.0  15.00  19.0  25.00
Hungary                  128.0  42.23  73.58  10.0  19.75  25.0  40.50
India                      8.0  13.75   3.65  10.0  12.00  12.0  14.50
Israel                   453.0  31.77  18.89   8.0  18.00  25.0  40.00
...
Turkey                   120.0
US                      2013.0
Ukraine                   13.0
Uruguay                  130.0
```

Productores

1. *¿Cuáles son los varietales con los mejores puntajes globales?*

Figura 7: Promedio de puntos por varietal

Promedio de puntos por varietal:

	variety	mean_points
0	Tinta del Pais	95.00
1	Terrantez	95.00
2	Gelber Traminer	95.00
3	Bual	94.14
4	Sercial	94.00
..	...	...
696	Aidani	82.00
697	Picapoll	82.00
698	Shiraz-Tempranillo	82.00
699	Airen	81.67
700	Chancellor	80.50

[701 rows x 2 columns]

2. *¿Cuál es el varietal más utilizado por país y cuál es su puntaje promedio y variabilidad?*

Figura 8: Varietal más usado por país y su puntaje

Varietal mas usado por pais y su puntaje:

country	variety	mean_points	std_points
England	Sparkling Blend	91.51	1.78
Austria	Grüner Veltliner	90.41	2.48
India	Shiraz	90.25	1.83
Germany	Riesling	90.08	2.47
Hungary	Furmint	89.52	2.81
Switzerland	Pinot Noir	89.33	0.58
Canada	Riesling	89.30	2.47
France	Bordeaux-style Red Blend	89.03	3.21
China	Cabernet Blend	89.00	NaN
Australia	Shiraz	88.99	2.92
US	Pinot Noir	88.89	3.19
Portugal	Portuguese Red	88.81	3.01
Israel	Cabernet Sauvignon	88.66	2.57
Italy	Red Blend	88.64	2.52
Morocco	Red Blend	88.57	2.03
Luxembourg	Sparkling Blend	88.50	0.58
Serbia	Riesling	88.25	1.50
New Zealand	Sauvignon Blanc	88.22	2.42
Slovenia	Sauvignon Blanc	88.19	1.89
South Africa	Chenin Blanc	88.10	2.40
Georgia	Saperavi	88.05	1.85
...			
Brazil	Sparkling Blend	84.50	2.30
Egypt	Grenache	84.00	NaN
Peru	Red Blend	83.73	2.00
Ukraine	Sparkling Blend	83.62	1.92

3. *¿Cómo se relaciona el puntaje del vino con el año de producción?*

Figura 9: Años y su relación con el promedio de puntos obtenidos.

```
Años y su relación con el promedio de puntos obtenidos:

       año  points
73  1821.0   85.75
21  1827.0   92.50
68  1845.0   87.00
22  1847.0   92.00
33  1848.0   89.50
..     ...     ...
37  2013.0   88.95
38  2014.0   88.89
43  2015.0   88.51
62  2016.0   87.63
74  2017.0   85.55

[78 rows x 2 columns]
```

Figura 10: Análisis de correlación entre el año y los puntajes.

```
Correlación entre el año de producción y el puntaje: 0.13

No hay una relación lineal clara entre el año y el puntaje
```

También se realizó un análisis entre el año (extraído del campo 'title') y el puntaje para observar si existe alguna relación. El resultado obtenido fue de 0.13 (calculado con el método .corr() de pandas), lo que indica que no hay una relación lineal clara entre el año y el puntaje. Esto sugiere que los cambios en el año no tienen un impacto predecible en los puntajes.

4. *¿Existen variaciones significativas en el puntaje entre las provincias de cada país?*

Figura 11: Variación de puntos entre provincias.

Variación de puntos entre provincias:

	country	province	mean_points	std_points
0	Argentina	Mendoza Province	86.77	3.27
1	Argentina	Other	85.96	2.76
2	Armenia	Armenia	87.50	0.71
3	Australia	Australia Other	85.47	2.21
4	Australia	New South Wales	87.72	2.57
..	...	...	...	...
420	Uruguay	Juanico	86.27	3.66
421	Uruguay	Montevideo	88.40	2.59
422	Uruguay	Progreso	86.82	2.18
423	Uruguay	San Jose	84.00	2.65
424	Uruguay	Uruguay	86.74	2.73

[425 rows x 4 columns]

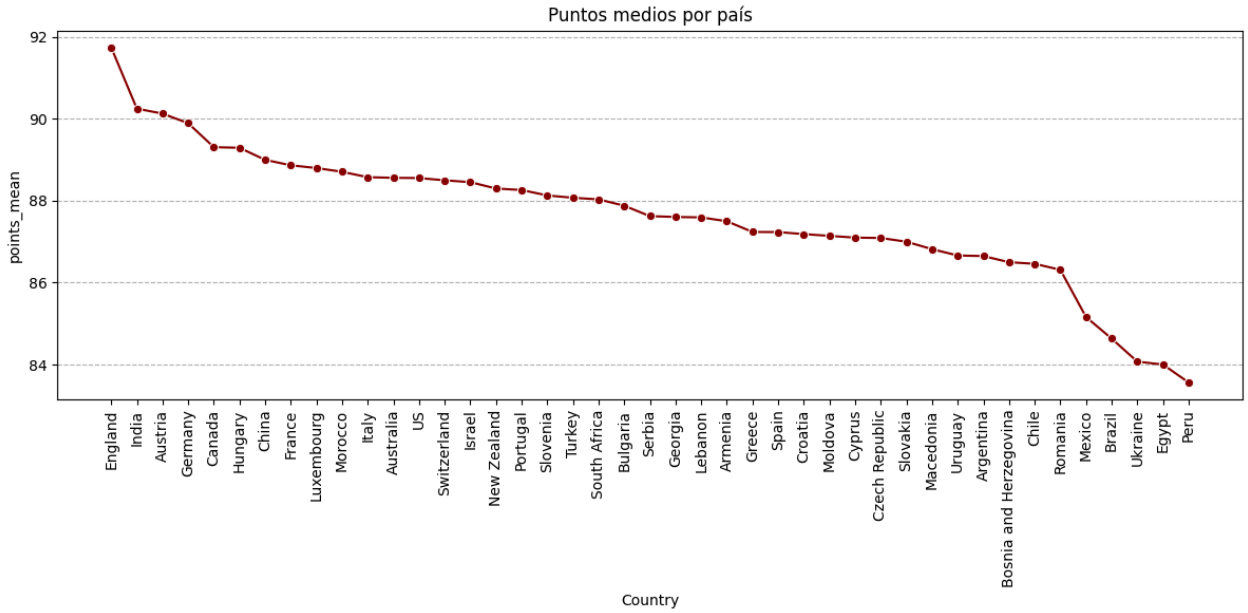
4. Creación de visualizaciones efectivas

Una vez completados los análisis para responder a las preguntas planteadas, se procederá a expresar los resultados mediante gráficos, lo que permitirá comprender de manera más sencilla y visual las posibles relaciones existentes. Comenzaremos con el sector de los comerciantes y luego pasaremos al de los productores. Además, cada visualización será acompañada de su respectiva interpretación.

Comerciantes

1. *¿Cuáles son los países con los mejores puntajes promedio en vinos?*

Figura 12: Puntos medios por país.

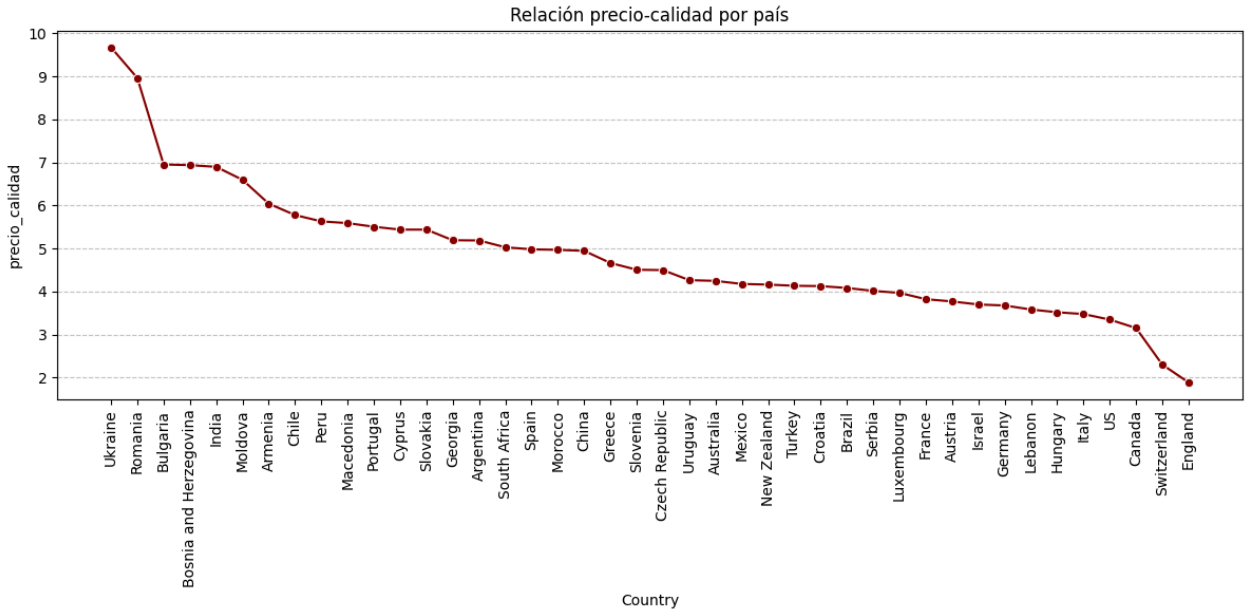


En el gráfico obtenido, podemos observar que el promedio de los puntajes oscila entre 80 y 100. Inglaterra, con un puntaje medio de 91.75, es el país con la mayor cantidad de puntajes altos, mientras que Perú, con un puntaje promedio de 83.56, tiene el puntaje medio más bajo de la lista. La media general se encuentra entre 87 y 88.

Estos resultados nos permiten identificar los países con los mejores puntajes promedio. También se puede notar que el puntaje promedio más bajo es inferior a 84, lo que significa que la diferencia entre la media de puntajes de Inglaterra y Perú es de solo 8.19 puntos. Para un comerciante, esto ofrece claridad al elegir vinos según su país de origen, ya que la diferencia no es muy grande. Por lo tanto, se puede tener la tranquilidad de que los vinos tendrán una buena puntuación relativa.

2. *¿Cuál es el país con la mejor relación precio-calidad en los vinos?*

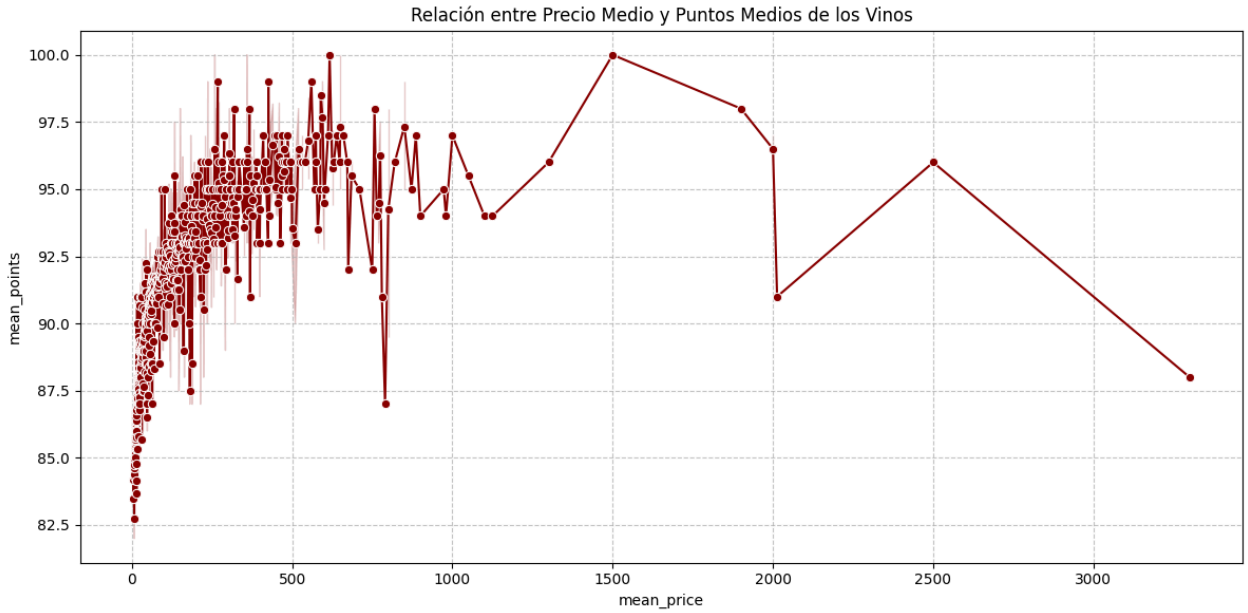
Figura 13: Relación precio-calidad por país.



Como se puede observar en el gráfico, el rango de precio-calidad oscila entre 1 y 10, siendo 1 el valor más bajo y 10 el más alto. Con esto en mente, podemos analizar la relación entre los países y esta variable de precio-calidad. Se observa que Ucrania lidera en este aspecto, mientras que Inglaterra ocupa el último puesto, a pesar de haber encabezado previamente la lista de los países con los mejores puntajes promedio. Esto sugiere que la variabilidad de los precios en Inglaterra es alta, lo que la convierte en uno de los países con la menor relación precio-calidad. Esta información puede ser muy útil tanto para comerciantes como para consumidores, ya que refleja una variabilidad en los precios según el país de origen.

3. *¿Cuáles son los vinos con más reseñas y cómo se distribuyen en términos de calidad y precio?*

Figura 14: Relación entre precio medio y puntos medios de los vinos.



Tendencia General

En general, se observa una tendencia ascendente: a medida que aumenta el precio promedio del vino, también lo hacen los puntajes promedio. Esto sugiere que los vinos más caros tienden a recibir mejores puntuaciones, lo cual es esperado, ya que los vinos de mayor precio suelen ser de mayor calidad.

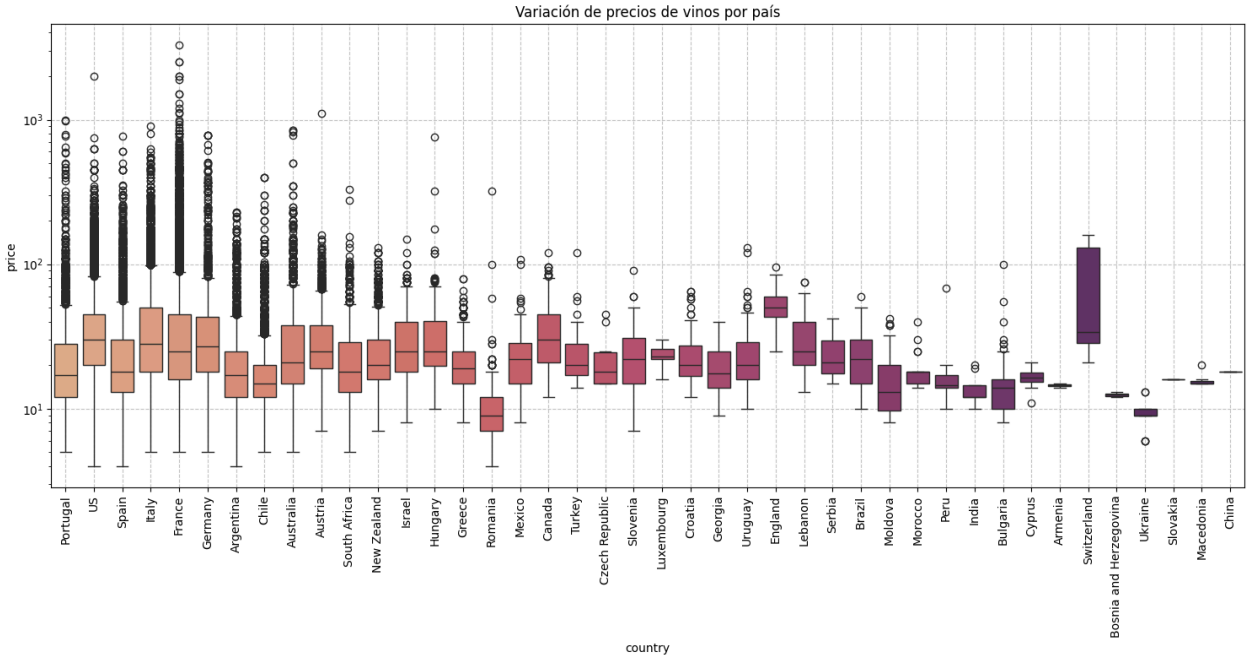
Variabilidad a precios altos

A partir de un precio promedio de aproximadamente 1000, la variabilidad en los puntajes promedio aumenta. Esto indica que, aunque algunos vinos caros reciben altas puntuaciones, no todos los vinos caros garantizan una alta calidad. Hay casos en los que vinos muy caros no obtienen puntuaciones tan altas, lo que puede deberse a factores como la percepción subjetiva de los catadores o a la alta variabilidad de los precios.

Aunque podemos afirmar que existen vinos de menor precio con mejores puntuaciones, no podemos saber con certeza si los precios están correctamente registrados o si están sujetos a distintas monedas, ya que el archivo CSV no menciona nada al respecto. Por lo tanto, este análisis podría ofrecer información incierta y con mucha incertidumbre al respecto.

4. *¿Qué relación existe entre el precio y el país de origen del vino?*

Figura 15: Variación de precios de vinos por país.



Como se puede observar, uno de los mejores gráficos para visualizar la variabilidad de los precios es el boxplot. En este gráfico, cada caja representa la distribución de precios de los vinos para un país específico, mostrando la mediana, los cuartiles y los valores atípicos (outliers).

Distribución de Precios

La mayoría de los países presentan una amplia variabilidad en los precios de sus vinos, lo que podría indicar una diversidad en la calidad y el tipo de vinos disponibles. Países como Francia, Italia y Estados Unidos muestran una gran dispersión en los precios, con valores atípicos que representan vinos de muy alto costo.

Mediana de Precios

La mediana de precios varía significativamente entre los países. Por ejemplo, Francia y Estados Unidos tienen medianas de precios más altas en comparación con

países como Argentina y Chile, lo que sugiere que los vinos en estos países tienden a ser más caros en promedio.

Valores Atípicos

Los valores atípicos (outliers) son especialmente notables en países como Francia, Italia y Estados Unidos, donde algunos vinos alcanzan precios extremadamente altos. Estos outliers podrían representar vinos de alta gama y ediciones limitadas, que son altamente valorados en el mercado.

Comparación entre Países

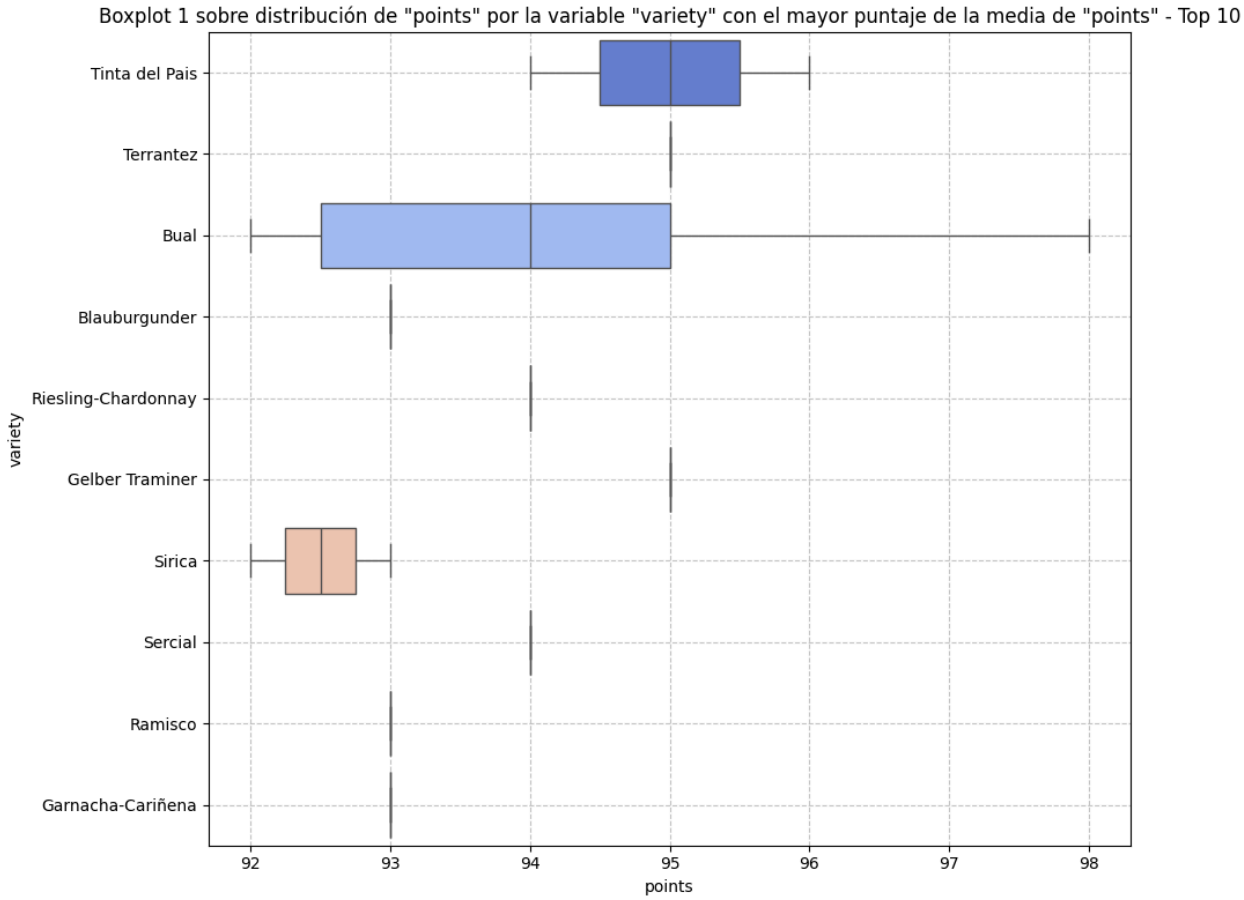
Países como Argentina y Chile muestran una menor variabilidad en los precios, con menos valores atípicos y una mediana de precios más baja. Esto puede indicar una oferta más homogénea de vinos en términos de precio, lo que podría ser atractivo tanto para consumidores como para comerciantes que buscan vinos de buena calidad con precios menos variables.

Como se mencionó anteriormente, no se puede confiar completamente en la variable ‘price’, por lo que el análisis realizado tiene incertidumbre al respecto.

Productores

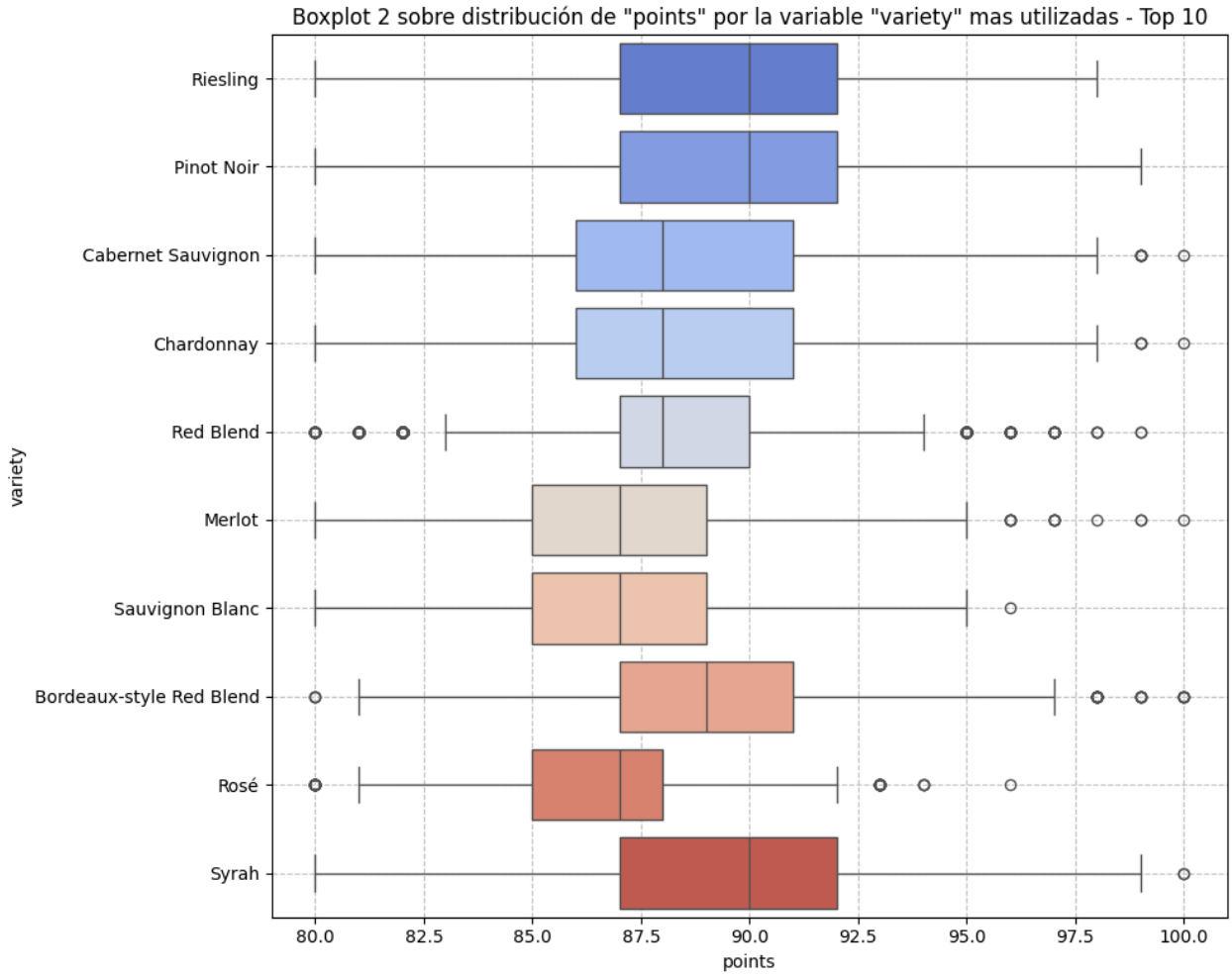
1. *¿Cuáles son los varietales con los mejores puntajes globales?*

Figura 16: Boxplot sobre la distribución de ‘points’ por la variable ‘variety’ con el mayor puntaje de la media de ‘points’.



Como vimos anteriormente, el gráfico boxplot nos ayuda a interpretar cómo varían los valores de una variable. En este caso, podemos observar que no siempre los mejores varietales, es decir, aquellos que tienen el mejor promedio de puntajes a nivel global, son los más utilizados. Esto puede deberse a varios factores, como la exclusividad del varietal para la elaboración de vinos de alta puntuación, la dificultad para cultivar las uvas necesarias o la disponibilidad limitada de ciertas variedades, entre otros. Por lo tanto, se ajustará el análisis observando el comportamiento de los varietales más utilizados y su relación con los puntajes obtenidos.

Figura 17: Boxplot sobre la distribución de 'points' por la variable 'variety' más utilizadas.



En este caso, dado que tenemos un total de 701 varietales, como se obtuvo en el análisis anterior, se procede a seleccionar los 10 varietales más utilizados en los vinos, lo cual arroja el siguiente resultado:

Mediana de Puntos

La mediana de puntajes para cada varietal se encuentra en un rango alto, generalmente entre 85 y 95 puntos. Esto indica que, en promedio, estos varietales de vino son bien valorados.

Variabilidad en los Puntos

Algunos varietales, como el Pinot Noir y el Cabernet Sauvignon, muestran una mayor variabilidad en los puntajes, lo que sugiere una amplia gama de calidades dentro de estos varietales. Por otro lado, varietales como el Riesling y el Rosé presentan una variabilidad menor, lo que indica una calidad más consistente.

Valores Atípicos

Los valores atípicos son visibles en varios varietales, lo que indica la presencia de vinos excepcionalmente buenos o malos dentro de esas categorías. Por ejemplo, el Pinot Noir y el Cabernet Sauvignon presentan varios valores atípicos altos, lo que sugiere la existencia de vinos de muy alta calidad en estas varietales.

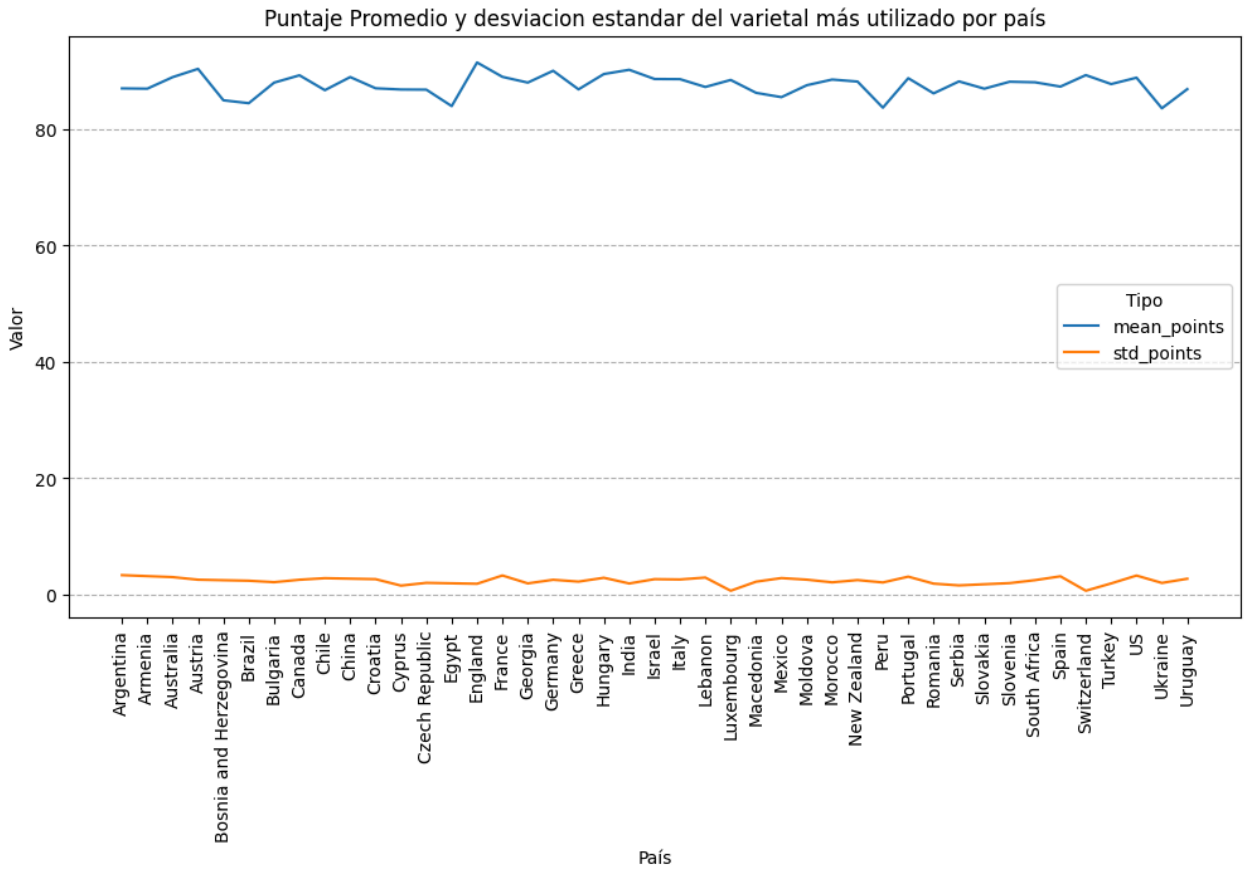
Comparación entre Variedades

En general, las varietales tintas, como el Pinot Noir, el Cabernet Sauvignon y el Bordeaux-style Red Blend, tienden a tener puntajes más altos y una mayor variabilidad en comparación con las varietales blancas, como el Riesling y el Sauvignon Blanc. Esto puede reflejar las preferencias de los catadores y la diversidad en la producción de vinos tintos.

Aunque solo se han considerado 10 varietales, para un productor de vino estos datos serán útiles para saber cuáles son los varietales más preferidos y los más demandados.

2. *¿Cuál es el varietal más utilizado por país y cuál es su puntaje promedio y variabilidad?*

Figura 18: Puntaje Promedio y desviación estándar del varietal más utilizado por país.



La línea azul representa los puntajes promedio, los cuales se mantienen consistentemente alrededor de 80 en todos los países. La línea naranja representa la desviación estándar, que permanece cercana a cero en todos los países.

Puntaje Promedio Consistente

Los puntajes promedio de los varietales más utilizados por país son bastante consistentes, rondando los 80 puntos. Esto sugiere que, independientemente del país, los varietales más populares tienden a recibir una valoración similar en cuanto a calidad.

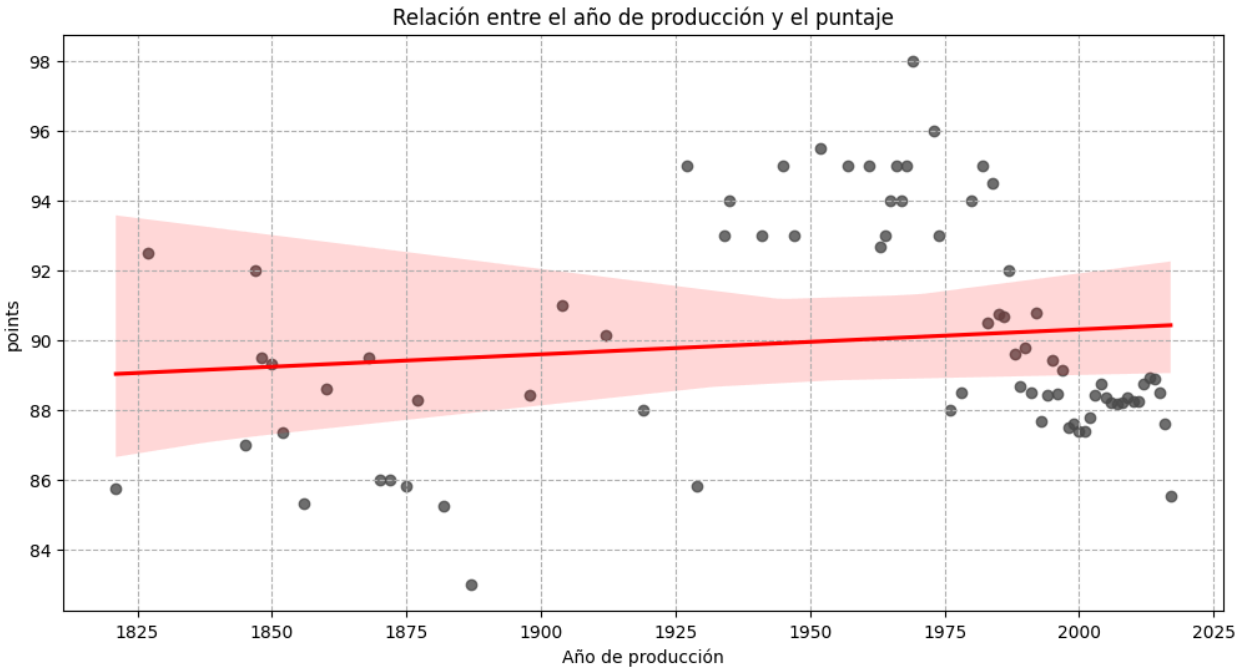
Baja Variabilidad

La desviación estándar es muy baja en todos los países, lo que indica que hay poca variabilidad en los puntajes de los varietales más utilizados. Esto sugiere que los puntajes son bastante consistentes y no varían mucho dentro de cada país.

Este gráfico es útil para entender la consistencia en la calidad de los varietales más utilizados por país. La baja variabilidad en los puntajes sugiere que los productores pueden esperar una calidad similar en los varietales más populares, independientemente del país de origen. Esto puede ser una ventaja, ya que pueden confiar en la consistencia de la calidad de estos varietales al tomar decisiones estratégicas.

3. *¿Cómo se relaciona el puntaje del vino con el año de producción?*

Figura 19: Relación entre el año de producción y el puntaje.



El siguiente gráfico muestra la relación entre el año de producción de los vinos (extraído del campo ‘title’) y sus puntajes. Los años varían desde 1821 hasta 2017, según los análisis anteriores, y los puntajes van desde 85.75 hasta 98.00 puntos,

también según los análisis previos. Cada punto negro en el gráfico representa un dato individual. Además, hay una línea de tendencia roja con un intervalo de confianza sombreado, que indica una ligera correlación positiva entre el año de producción y el puntaje.

Tendencia General

La línea de tendencia roja sugiere una ligera correlación positiva entre el año de producción y el puntaje. Esto significa que, en general, los vinos más recientes tienden a recibir puntajes ligeramente más altos.

Distribución de Puntos

La mayoría de los puntos se agrupan en torno a los años más recientes, lo que indica que hay más datos disponibles para los vinos producidos en los últimos años. Esto puede deberse al aumento en el registro de reseñas en tiempos recientes.

Variabilidad en los Puntajes

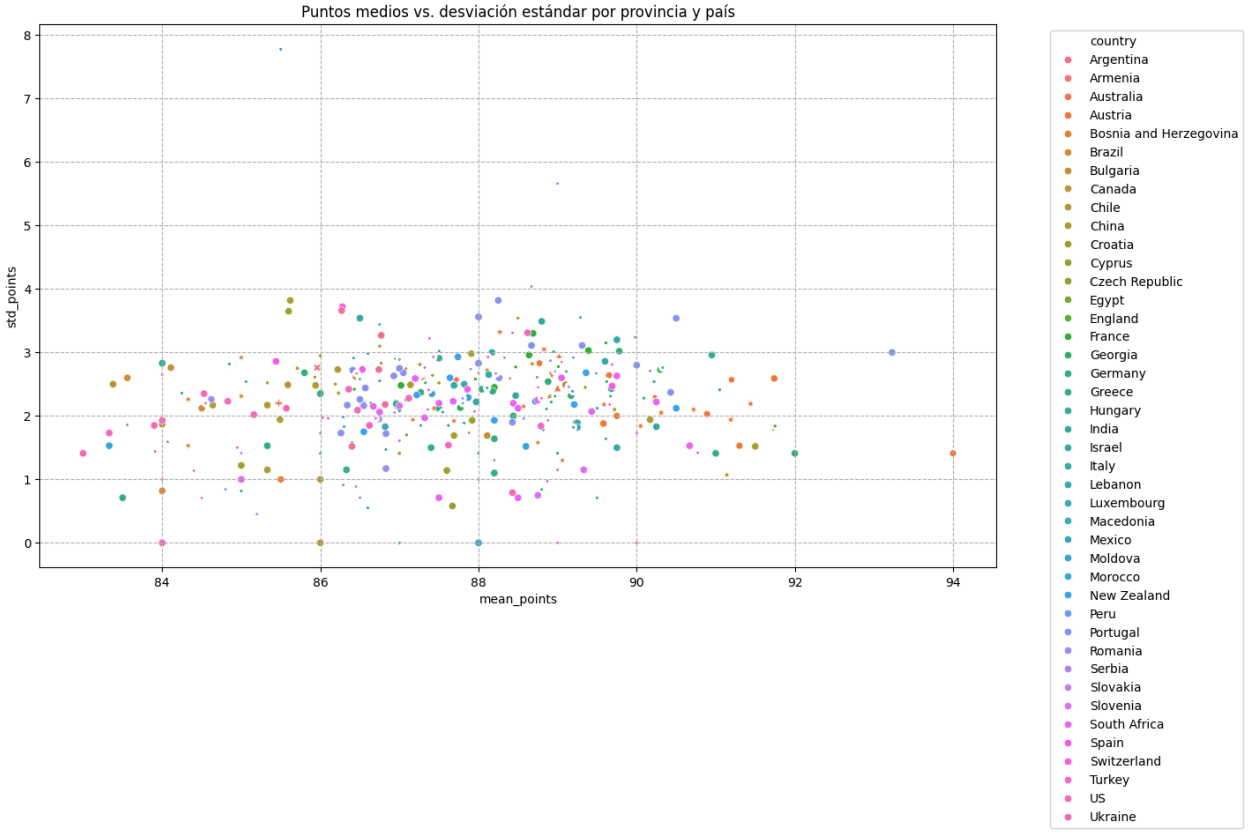
Aunque hay una tendencia general al alza, la variabilidad en los puntajes es notable. Esto sugiere que, aunque los vinos más recientes tienden a recibir mejores puntajes, hay una amplia gama de calidades en cada año de producción.

La ligera correlación positiva indica que los vinos más recientes tienden a ser mejor valorados, lo que puede reflejar mejoras en las técnicas de producción y enología. Sin embargo, la variabilidad en los puntajes también sugiere que no todos los vinos recientes son de alta calidad y que existen excepciones notables.

El valor de correlación de 0.13 refuerza esta interpretación, indicando que la relación entre el año de producción y el puntaje es positiva pero débil. Esto significa que, aunque existe una tendencia a que los vinos más recientes tengan mejores puntajes, la relación no es lo suficientemente fuerte como para hacer predicciones precisas basadas únicamente en el año de producción.

4. *¿Existen variaciones significativas en el puntaje entre las provincias de cada país?*

Figura 20: Puntos medios vs. desviación estándar por provincia y país.



El gráfico de dispersión muestra la relación entre los puntos medios (mean_points) y la desviación estándar (std_points) de los vinos de diferentes provincias y países. Cada punto en el gráfico está codificado por colores según el país que representa, como se indica en la leyenda a la derecha del gráfico.

Relación Genera

El gráfico muestra una dispersión de puntos que indica la variabilidad en los puntajes medios y las desviaciones estándar de los vinos por provincia y país. No se observa una tendencia clara que sugiera una relación directa entre los puntos medios y la desviación estándar.

Países Destacados

Francia e Italia: Estos países tienen varios puntos con altos puntajes medios y bajas desviaciones estándar, lo que sugiere una calidad consistente en sus vinos.

Estados Unidos: Muestra una mayor variabilidad en los puntajes medios y las desviaciones estándar, indicando una diversidad en la calidad de los vinos producidos en diferentes provincias.

España y Portugal: También presentan una buena relación entre puntajes medios altos y bajas desviaciones estándar, lo que indica una calidad consistente en sus vinos.

Variabilidad en los Puntajes

La variabilidad en los puntajes medios y las desviaciones estándar sugiere que, aunque algunos países tienen vinos de alta calidad de manera consistente, otros presentan una mayor diversidad en la calidad de sus vinos.

6. Conclusiones

A lo largo de este trabajo práctico, hemos observado el comportamiento de las diferentes variables, como ‘points’, ‘price’, ‘country’, ‘variety’, ‘province’ y ‘año’, que fue extraído de la variable ‘title’. Al relacionarlas entre sí, pudimos observar lo siguiente:

Relaciones existentes

Se encontraron relaciones entre los varietales, los puntajes y los países, lo cual nos brinda una valiosa información acerca de qué tipo de varietales usar en caso de que un productor decida invertir en un nuevo vino. Esto ayuda a identificar qué varietales son más preferidos y su impacto en los diferentes países como favoritos. Por ejemplo, se observó que los varietales tintos son los más utilizados y, además, obtienen muy buenos puntajes, aunque estos pueden ser variables; sin embargo, esa variabilidad no es muy pronunciada.

Por parte de los comerciantes, se observa una dependencia entre el varietal y el país, lo que permitió obtener relaciones importantes para determinar en qué tipo de vino

invertir, dependiendo del varietal más consumido en ese país y también por provincias. De esta manera, se obtiene un análisis detallado de esta dimensión regional.

Relaciones inexistentes

Por otro lado, se detectaron relaciones en algunos casos muy débiles y otras en las que fue difícil encontrar una relación directa.

En el caso del precio, se observó que existe una gran dificultad para determinar su relación con otras variables, ya que los precios tienden a variar significativamente, lo que genera relaciones poco claras y confusas. Inicialmente se consideró la posibilidad de normalizarlos, pero al no contar con datos sobre el momento en que fueron capturados esos precios ni sobre la moneda en la que están expresados, resulta imposible llevar a cabo dicha normalización. Por lo tanto, cualquier análisis relacionado con los precios estará marcado por incertidumbre, lo que dificultará la interpretación de los resultados.

Dado que no es posible utilizar los precios para relacionar las distintas variables, no será posible responder adecuadamente a las preguntas formuladas previamente. Tanto para productores como comerciantes no se podrá realizar un análisis relacionado con los precios debido a la incertidumbre de los mismos.

Por otro lado, se pudo trabajar exitosamente con el año de cada vino, extraído de la variable 'title'. Se logró relacionarlo con la variable 'points', lo que arrojó un resultado que sugiere que podría existir alguna relación. Sin embargo, debido a que los resultados varían demasiado, se puede concluir que esta relación es muy débil para afirmar que existe una conexión directa. Como se mencionó anteriormente, relacionar el año con la variable 'price' solo produciría resultados con alta incertidumbre.