EL PODER DE LOS DATOS EN LA INDUSTRIA VITIVINICOLA

JUAN MAXIMILIANO FEDERICI

Noviembre 2024
Trabajo Práctico Nº 4
Seminario de Práctica en Ciencia de Datos

Índice

1. Introducción 2
2. Selección Y Depuración De Datos 2
3. Contextualización Y Preguntas Clave 2
4. Creación E Interpretación De Visualizaciones 4
5. Conclusiones 16

1. Introducción

La creciente demanda de vinos de alta calidad y la necesidad de diferenciarse en un mercado cada vez más competitivo han impulsado a las bodegas a explorar nuevas tecnologías. La ciencia de datos ofrece un amplio abanico de oportunidades para mejorar la eficiencia, reducir costos y desarrollar productos personalizados. En este trabajo, analizaremos cómo los datos pueden convertirse en una ventaja competitiva para las empresas vitivinícolas.

2. Selección Y Depuración De Datos

Para el presente práctico se va utilizar el archivo *csv* resultante del TP2 (se entrega comprimido junto con este informe), en el cual ya se realizó la siguiente depuración de los datos:

- Eliminación de datos duplicados
- Borrado de campos que no aportan información (*ID*) o que tenían demasiados datos faltantes (*region2* y *tester_twitter*)
- Se completaron datos faltantes en *country* y *province*
- Se estimaron con distintos criterios los *price* faltantes
- Se eliminaron los registros en los cuales los *price* estaban fuera del rango intercuartil (outliers)

Los principales campos con los que se va a trabajar son: *country*, *description*, *points*, *price*, *province*, *tester_name*, *title*, *variety* y *winery*. Además, se crea un campo nuevo llamado "*año*" con información extraída de *title*, que contiene el año de cosecha del vino reseñado, a efectos de poder realizar un análisis respecto de las distintas añadas.

3. Contextualización Y Preguntas Clave

La industria vitivinícola, tradicionalmente arraigada en la tradición y la subjetividad de la cata, está experimentando una transformación digital impulsada por la ciencia de datos. Este campo ofrece una herramienta poderosa para analizar la vasta cantidad de información disponible sobre los vinos, desde las características de

la uva hasta las opiniones de los consumidores. A través del análisis de reseñas, podemos desentrañar patrones complejos en los gustos de los consumidores que van más allá de las simples preferencias por variedades o regiones.

Para los productores, esta información es un muy valiosa. Al comprender qué características sensoriales (taninos, acidez, aromas frutales, etc.) son más valoradas por los consumidores, pueden ajustar sus procesos de elaboración para satisfacer estas demandas. Además, pueden identificar nichos de mercado y oportunidades de innovación, desarrollando vinos que se adapten a paladares específicos.

Para los consumidores, la ciencia de datos brinda una guía invaluable en el laberinto de opciones que ofrece el mercado del vino. Al analizar las reseñas, pueden descubrir vinos que se ajusten a sus gustos personales, incluso si desconocen las características técnicas de un vino. Las plataformas de recomendación basadas en datos pueden sugerir maridajes ideales, ayudar a los consumidores a explorar nuevas variedades y regiones, y ofrecer una experiencia de compra más personalizada.

En este trabajo, se explora cómo la ciencia de datos puede arrojar luz sobre las preferencias de los consumidores de vino, revelando tendencias y patrones que pueden ser de gran utilidad tanto para los productores como para los consumidores. A través del análisis de un conjunto de datos de reseñas de vinos, se responde a una serie de preguntas clave que permiten comprender mejor este complejo y fascinante mundo[1].

A continuación, se desarrollan una serie de preguntas sobre la industria del vino, las cuales se intenta responder con los gráficos del siguiente punto.

Si bien no hay un criterio único, algunas de estas preguntas están más orientadas a la perspectiva de los productores, principalmente las que consideran la variable precio, ya que reflejan una valoración económica; y otras preguntas están más enfocada a los comercializadores y/o consumidores, que serían las que contemplan la puntuación, dado que están directamente relacionados con los gustos y preferencias.

[6] Google AI. (2024). Gemini (versión 16 de noviembre). [Modelo de lenguaje de gran tamaño], https://gemini.google.com/

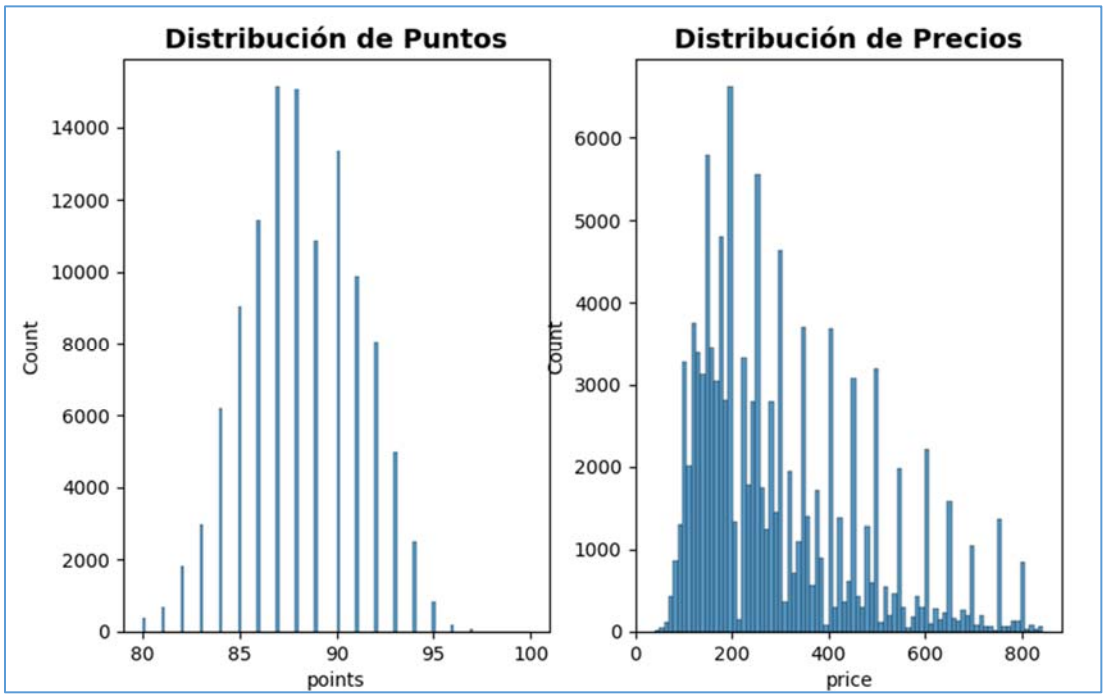
- ¿Cómo se distribuyen las 2 principales variables analizadas (precios puntos)?
- ¿Existe una correlación entre el precio y la puntuación?
- ¿Cómo han evolucionado la puntuación y los precios promedio de los vinos a lo largo del tiempo?
- ¿Cómo se ha mantenido en el tiempo la puntuación promedio de los principales países?
- ¿Existen diferencias significativas en la puntuación de vinos de diferentes países?
- ¿Qué características sensoriales son las más valoradas por los catadores según las reseñas?
- ¿Qué países producen los vinos más valoradas por los consumidores?
- ¿Cuáles son las bodegas más apreciados en términos de puntuación?
- ¿Cuáles son las variedades de uva más competitivas en términos de precios?
- ¿Cuáles son los países que producen los vinos malbec más caros?
- ¿Cuáles son las bodegas argentinas que producen los malbecs mejor puntuados?
- ¿Los catadores tienen distintos criterios de puntuación?
- ¿Qué variedades de uva producen los vinos mejor puntuados según el país de origen?
- ¿Qué variedades de uva producen los vinos más caros según la bodega que los elabora?

4. Creación E Interpretación De Visualizaciones

Se comienza mostrando un histograma de las 2 únicas variables numéricas con las que se cuenta, ya que las mismas se utilizan para realizar casi todos los gráficos del presente trabajo.

Para los puntos asignados por los catadores a cada vino, hay que aclarar que solo se consideran los valores que están entre 80 y 100 puntos, y se trata de una variable discreta, ya que no se utilizan decimales para la puntuación. En este caso se observa que la distribución presenta una forma casi normal, con un pequeño sesgo hacia la izquierda, lo cual indica que hay mayor cantidad de observaciones con menores puntajes.

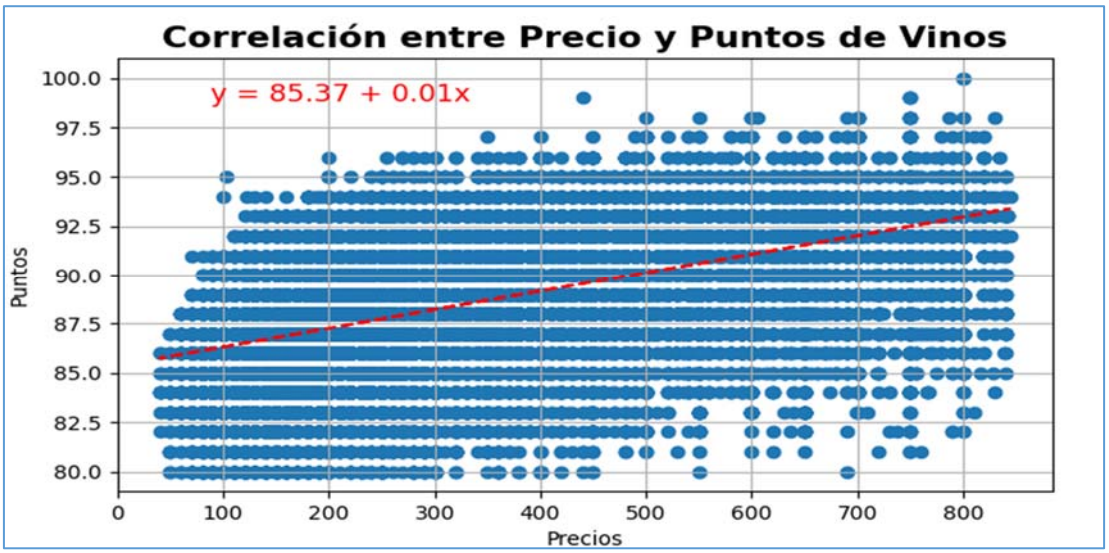
Para la variable precios los valores están entre 40 y 844, ya que como se explicó anteriormente, se eliminaron los registros con valores *outliers*. En este caso se observa un claro sesgo hacia la izquierda, donde se ubican en mayor cantidad los vinos más económicos.



Para analizar la relación entre precio y puntuación se realiza el siguiente gráfico de correlación, el cual proporciona evidencia visual de una relación directa entre ambas variables, lo cual se confirma con la línea de tendencia que se muestra en rojo, al igual que la ecuación que la representa.

En el contexto de los vinos, el precio suele ser considerado un indicador de diversos factores que influyen en la calidad percibida, como la variedad de uva, la región de origen, el proceso de elaboración y la añada, es por ello que un precio elevado puede generar expectativas más altas en los consumidores, lo que influye en su percepción de la calidad al momento de la degustación.

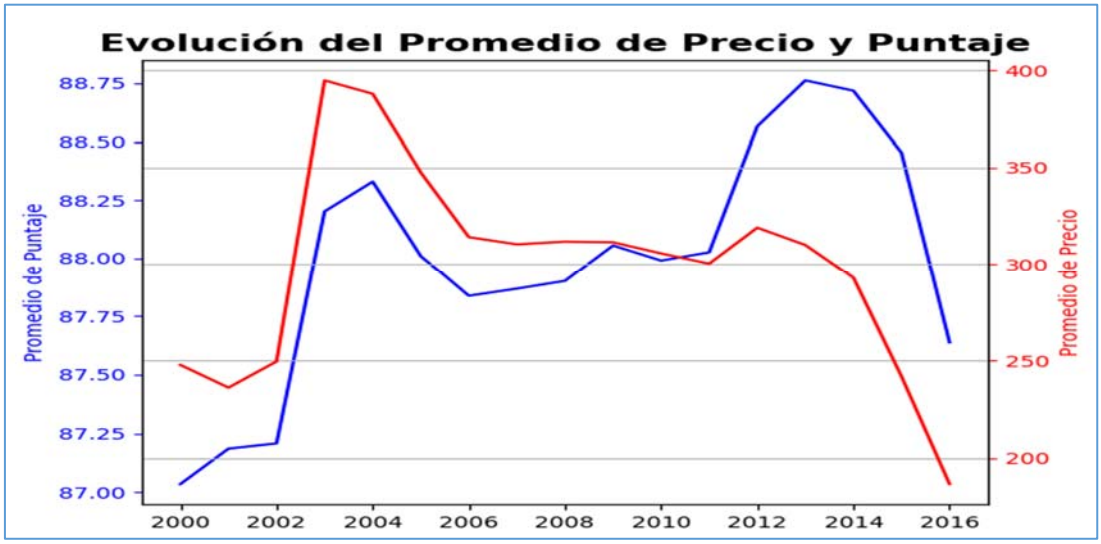
Otro factor que puede influir es que las bodegas suelen posicionar sus vinos de alta gama a precios más elevados para transmitir una imagen de exclusividad y superioridad, lo cual influye en las puntuaciones obtenidas por ese tipo de vinos.



En el gráfico debajo se estas líneas se presenta la evolución de los precios y puntaje entre los años 2000 y 2016, aunque cabe aclarar que los años se refieren al de la cosecha de las uvas, y no al momento en que se realiza la reseña.

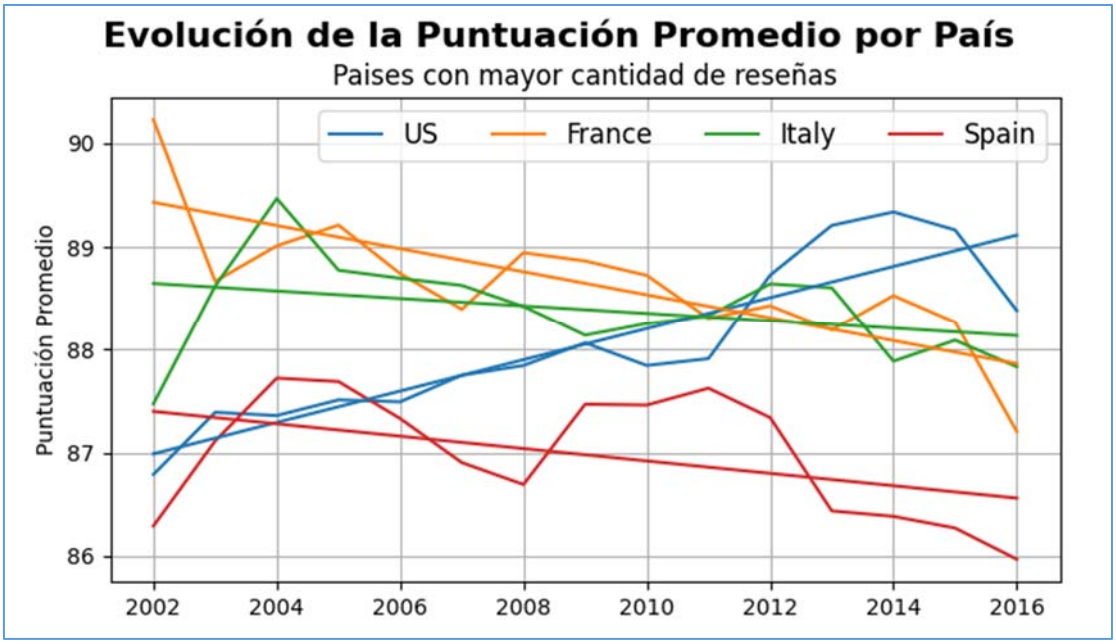
Ambas variables presentan un comportamiento fluctuante casi similar, lo cual es consistente con el gráfico anterior. Hay un fuerte crecimiento entre los años 2001 y 2004, en los años siguientes baja y se mantiene, aunque en el caso del puntaje vuelve a tener un aumento importante, pera después tanto precio como puntaje tienen una caída marcada hacia el final de la serie.

Este comportamiento se puede deber a distintos factores: económicos (por ejemplo, una recesión o periodos inflacionarios), climáticos (sequias o accidentes climáticos como granizo), ó factores de mercado (cambio en las modas o ingreso/egreso de nuevos competidores).



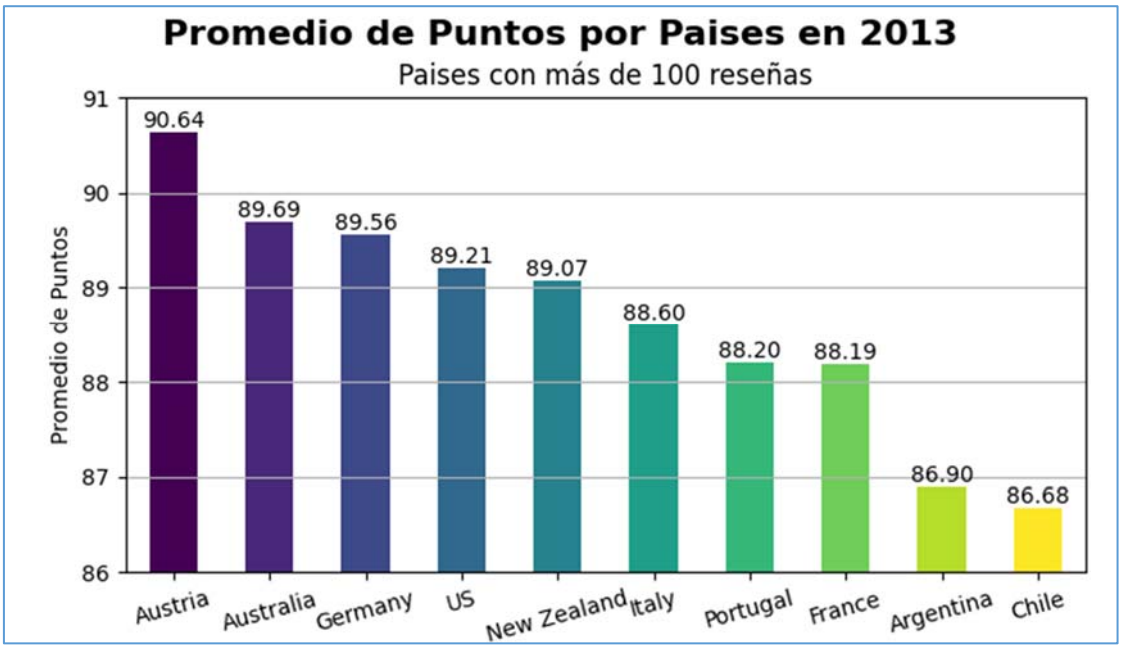
A continuación, se presenta un gráfico parecido al anterior, pero en este caso se muestra la evolución de la puntuación promedio para los 4 países con mayor cantidad de reseñas, permitiendo comparar sus desempeños relativos en términos de calidad percibida, según las puntuaciones otorgadas en las reseñas.

Para poder ver más fácilmente el comportamiento de cada país, es que se agregan las líneas de tendencia para cada país. Estados Unidos es el único país que mantiene una trayectoria en promedio creciente, los otros 3 países presentan puntuaciones en promedio decrecientes (en distintos niveles) a lo largo del periodo analizado, siendo Francia quien ha disminuido más rápidamente la valoración promedio de sus vinos.



Para ver cuáles eran los países que tenían los puntajes promedio más altos, se consideran solo aquellos que tuvieron al menos 100 reseñas de 2013, que fue el año que más observaciones hubo, y posteriormente se seleccionan los 10 con mayores promedios, con orden descendente.

De acuerdo a esta información, los vinos de Austria fueron los más valorados, con una puntuación promedio de 90,64 puntos, seguido por Australia y Alemania, mientras que Argentina y Chile quedan al final del listado. El gráfico muestra una gran diversidad en la calidad de los vinos producidos en los diferentes países. Los factores que influyen en las puntuaciones son múltiples y complejos, y dependen de una combinación de factores climáticos, varietales, enológicos y de mercado.

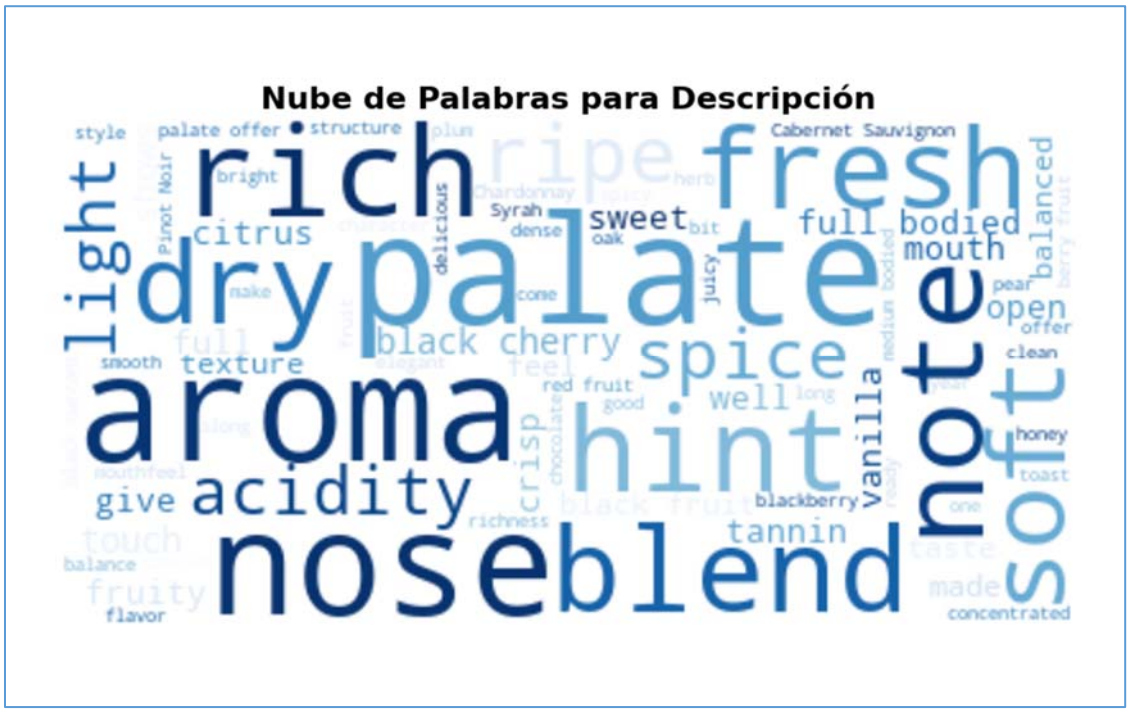


Una buena forma de conocer sobre los gustos de los catadores es analizando las palabras que más utilizan en la descripción de los vinos, es por ello que debajo se muestra una nube con las 80 palabras más mencionadas, aclarando que para mejorar la información se dejaron de lado palabras que no son representativas de los gustos, como “drink”, “wine” y otras.

La nube de palabras da una visión general de los términos más utilizados para describir los vinos. Al observar el tamaño y la frecuencia de cada palabra, podemos observar lo siguiente:

- *Características sensoriales*: aroma, flavor, palate y texture
- *Variedades de uva*: Cabernet Sauvignon, Pinot Noir y Syrah y Blend
- *Perfil de sabor*: fruity, sweet, spicy, dry, black cherry y acidity
- *Estructura del vino*: body, tannin y structure

Esta nube de palabras ofrece una visión general de los atributos más valorados en los vinos según las reseñas analizadas. Al comprender estos términos y sus relaciones, podemos obtener una mejor comprensión de las preferencias de los consumidores y las características que hacen que un vino sea mejor valorado.



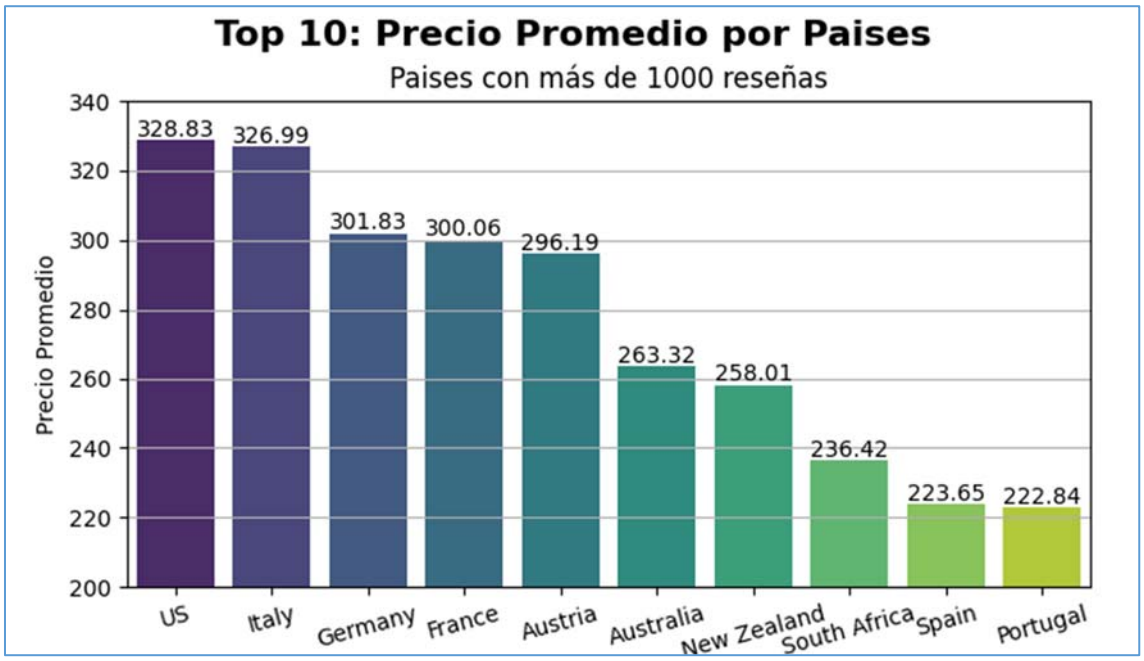
El siguiente gráfico muestra cuales son los 10 países que tienen los vinos que en promedio son más caros, pero para evitar distorsiones por pocos casos, solo se consideran aquellos países que tienen al menos 1000 reseñas.

Se observa que Estados Unidos e Italia tienen los promedios de precios más altos, seguidos un poco más bajo por Alemania, Francia y Austria, y al final de la lista se ubican España y Portugal.

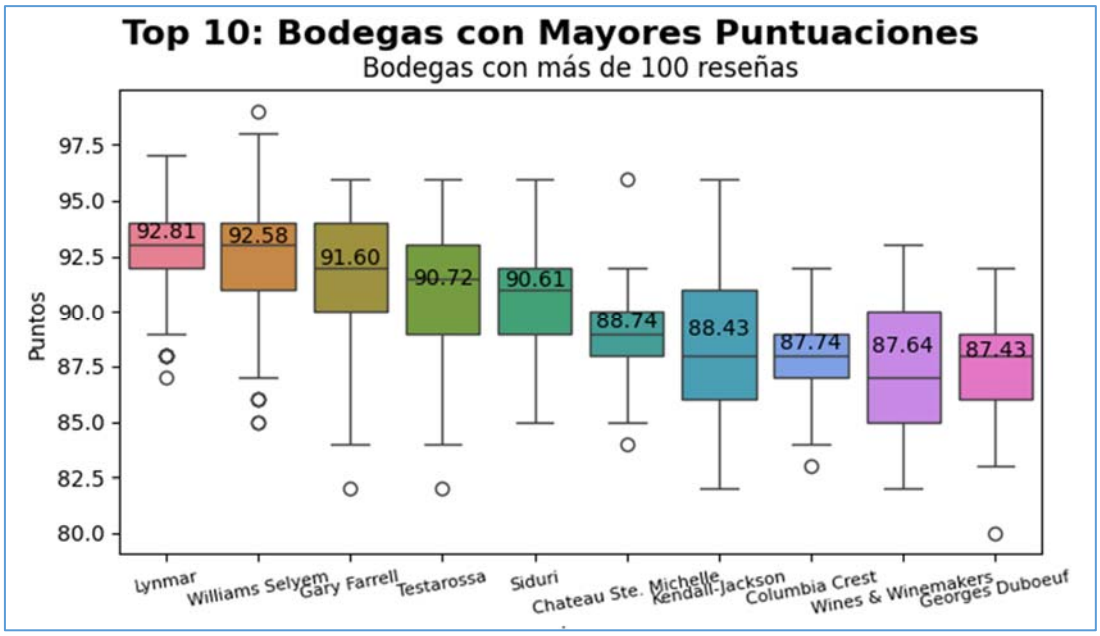
A simple vista se puede apreciar que existe una diferencia importante en los precios dependiendo de su país de origen. Pero para verificar si dicha afirmación es correcta, es que se realiza el test ANOVA, obteniéndose los siguientes resultados:

	sum_sq	df	F	PR(>F)
country	120664273	9	481,36	0
Residual	2886402130	103631		

Al ser PR(>F) < 0.05 se puede afirmar que si existen diferencias significativas en los precios promedios de los 10 países seleccionados.



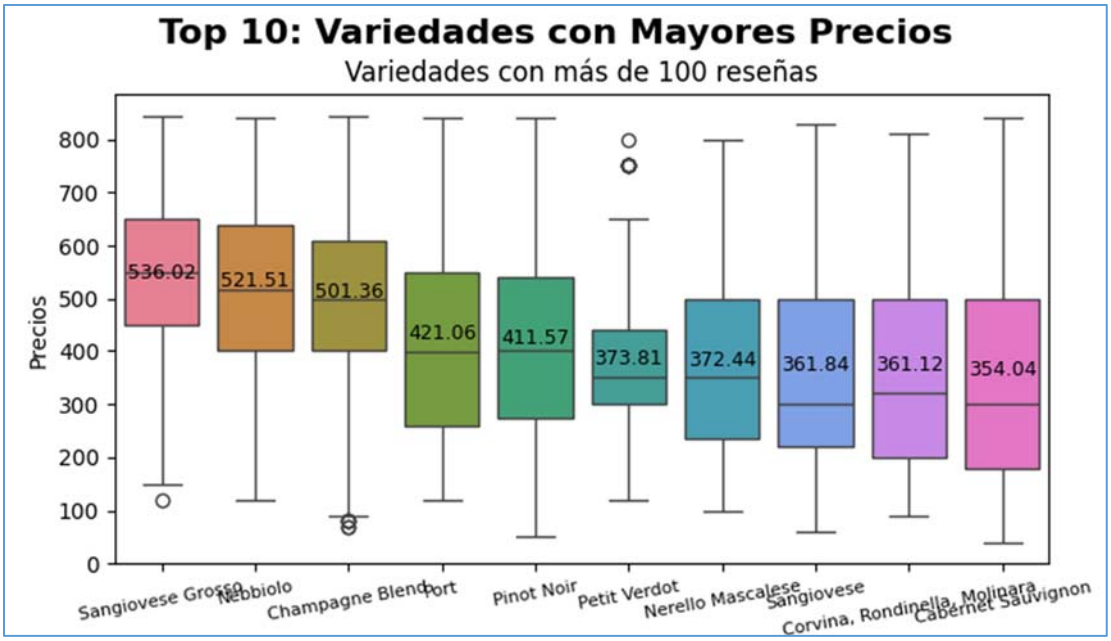
Una forma de conocer sobre los gustos de los consumidores, es analizando cuales son las bodegas que tienen las mejores calificaciones, es por ello que en el siguiente gráfico se presentan las 10 bodegas con mayores puntuaciones promedio (números dentro de las cajas). El diseño de caja y bigote permite apreciar la dispersión que tiene cada bodega en los puntos obtenidos, donde valores extremos distantes estarían indicando que no hay una clara consideración por la calidad de los vinos. Por el contrario, el caso de *Lynnar* muestra que además de tener el mayor promedio de puntos, tiene poca dispersión, aunque si tiene algunos valores fuera de rango.



Desde la perspectiva de los productores de vinos, conocer cuáles son las variedades por las cuales se pagan los precios más altos puede servirles para saber cuáles son los varietales que más les conviene producir. Es por ello que el gráfico que está debajo muestra en orden decreciente las 10 variedades con mayor precio promedio (números dentro de las cajas). Solo se consideran variedades que tengan más de 100 reseñas, ya que mientras más grande sea la muestra, proporciona una estimación más precisa del precio promedio y la distribución de los precios.

Del gráfico se desprende los siguiente:

- *Variabilidad en los Precios*: es evidente que existe una gran variabilidad en los precios de las diferentes variedades de uva, incluso dentro de una misma variedad. Esto puede deberse a factores como la región de origen, el productor, la añada, y la calidad percibida del vino.
- *Valores Atípicos*: la presencia de valores atípicos sugiere que algunos vinos de ciertas variedades pueden alcanzar precios significativamente más altos o más bajos que el promedio, lo que podría estar relacionado con factores como la rareza, la demanda o características únicas del vino.
- *Segmentación del Mercado*: el gráfico sugiere que existe una segmentación en el mercado de vinos, con algunas variedades posicionadas en el segmento premium (Sangiovese Grosso, Champagne Blend Port) y otras en un segmento más accesible (Sangiovese, Molinara, Cabernet Sauvignon).

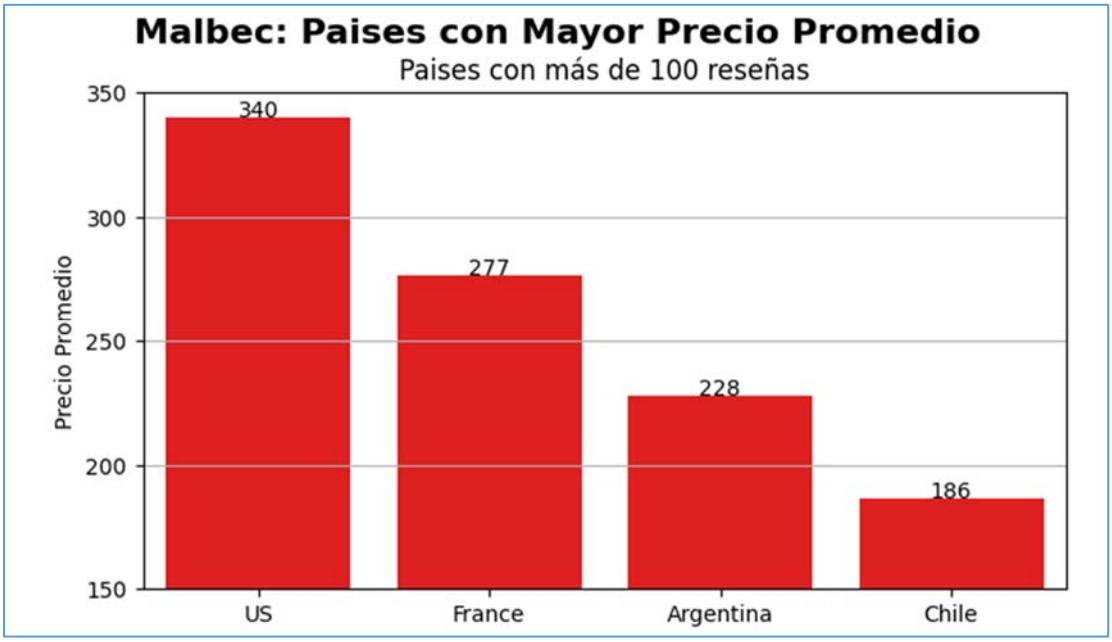


Si bien la variedad *malbec* no figura en los gráficos anteriores entre las más reseñadas, ni con mayores precios promedio, el siguiente gráfico es seleccionado por ser el varietal de mayor producción en Argentina (principalmente en Mendoza).

De los solo 4 países que tienen más de 100 reseñas de vinos malbec, se observa que por lejos Estados Unidos es quien tiene el precio promedio más alto, seguido por Francia, Argentina se ubica tercero, y finalmente Chile.

Las diferencias en los precios promedio de los malbec entre estos países pueden atribuirse a diversos factores, entre ellos:

- *Costos de producción*: varían significativamente entre países, influidos por factores como el terreno, el clima, la mano de obra, las regulaciones y las tecnologías utilizadas.
- *Reputación y prestigio*: países con una larga tradición vitivinícola y una reputación establecida por la calidad de sus vinos tienden a tener precios más altos.
- *Marketing y distribución*: las estrategias de marketing y los canales de distribución también influyen en los precios finales de los vinos.
- *Impuestos y aranceles*: las políticas fiscales y los aranceles pueden afectar los precios finales de los vinos en los diferentes mercados.



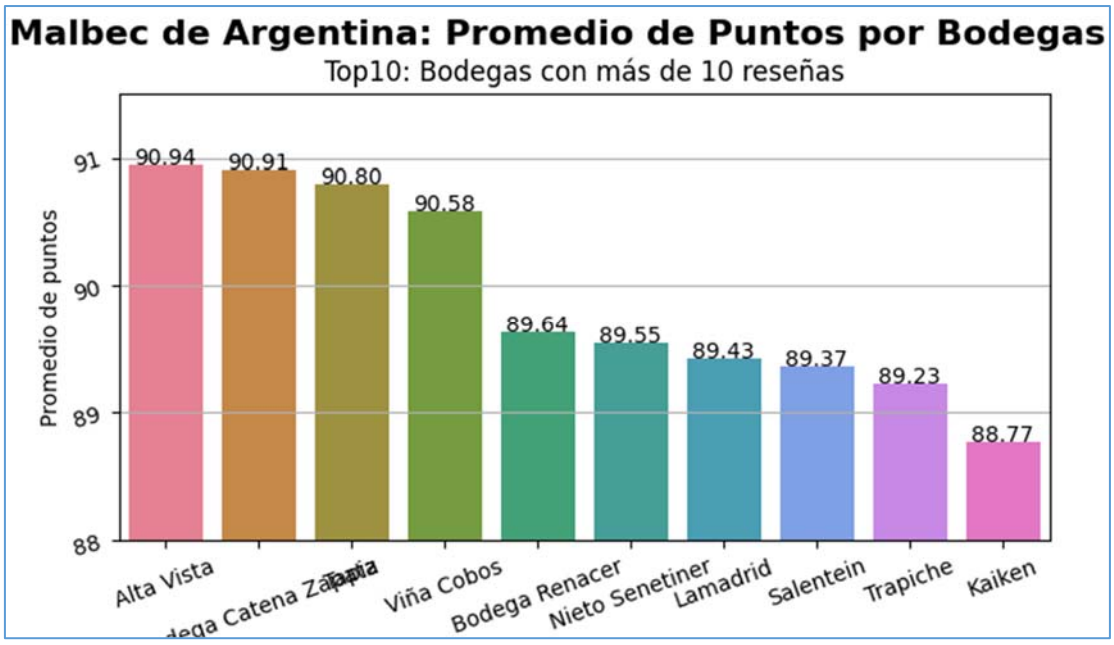
Siguiendo con el análisis del gráfico anterior, a continuación, se muestran las bodegas argentinas que producen vinos malbec, ordenadas decrecientemente según

puntaje promedio, aunque se debe aclarar solo se consideran las que tienen al menos 10 reseñas, dado que no se pudo obtener muestras más grandes.

La bodega *Alta Vista* se destaca con el puntaje promedio más alto, lo que sugiere que sus vinos malbec son consistentemente valorados por los críticos y consumidores. Le siguen de cerca *Bodega Catena Zapata, Tapiz* y *Viña Cobos*, consolidando a estas cuatro bodegas como las de mayor prestigio en términos de malbec. Del quinto al décimo lugar, encontramos un grupo de bodegas con puntajes muy cercanos, lo que indica una alta competencia en la producción de malbec de calidad en Argentina.

Las diferencias en los puntajes promedio entre las bodegas pueden atribuirse a diversos factores, por ejemplo:

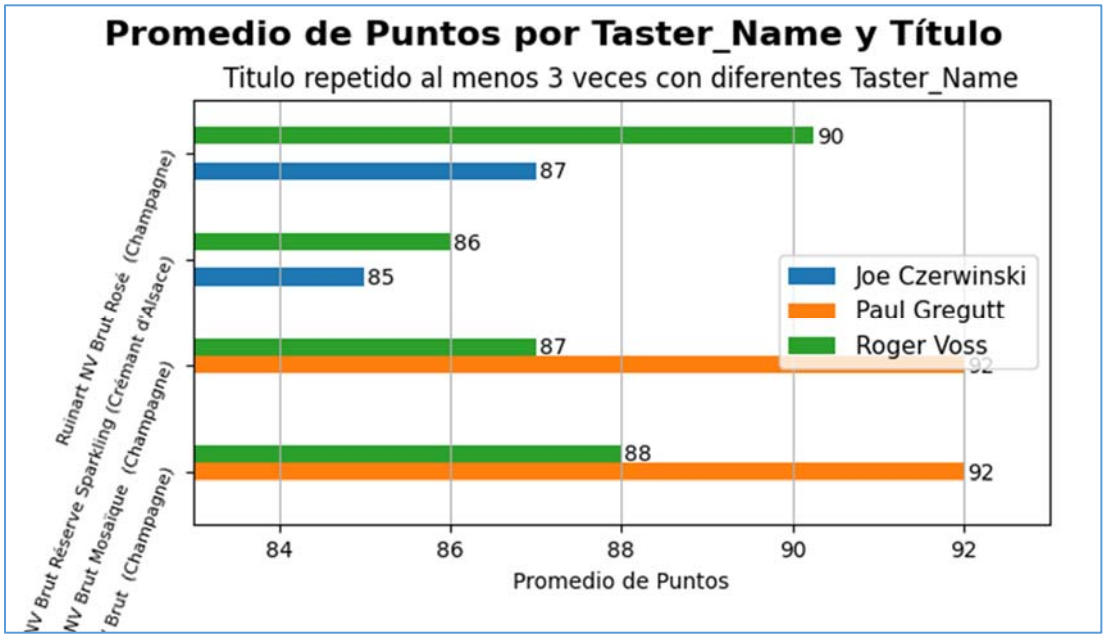
- *Terroir y viñedos*: la ubicación de los viñedos, las características del suelo y el clima influyen directamente en la calidad de la uva y, por ende, del vino.
- *Proceso de elaboración*: las técnicas de vinificación, la selección de las uvas, la crianza en barrica, la experiencia del enólogo y otros factores enológicos pueden marcar la diferencia en el resultado final.
- *Percepción subjetiva*: los puntajes asignados por los críticos y consumidores son en parte subjetivos y pueden variar según los gustos personales y las experiencias previas.



En relación a la *percepción subjetiva* antes mencionada, es que se presenta el siguiente gráfico, con el cual se pretende analizar cómo un mismo vino puede tener puntajes muy disimiles para 2 catadores distintos.

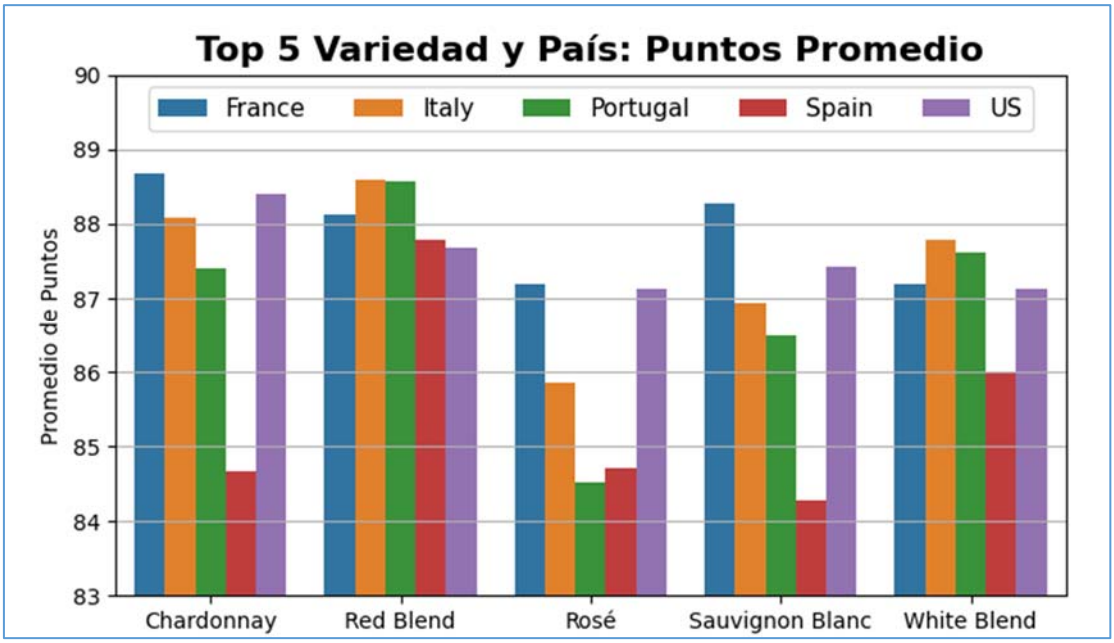
Se toma como referencia el campo *Title* ya que contiene el tipo de uva, la bodega y año de elaboración, lo cual permite identificar con cierta precisión un vino en articular. El otro dato que se utiliza es el campo *Taste_name* para identificar a cada catador. El gráfico se construye buscando aquellos *Title* que se repiten al menos 3 veces, y que además tengan al menos 2 *Taste_name* distintos (que tenga al menos 4000 observaciones), entonces se calcula el promedio de puntos para los distintos catadores.

En el gráfico se observa en el primer caso como el *Ruinart NV Rosé (Champagene)* tiene 2 valoraciones distintas: *Roger Voss* en promedio le asigna 90 puntos, mientras que *Joe Czerwinski* solo le atribuye 87. Algo similar se repite en distintas medidas con los otros 3 casos analizados, lamentablemente no se cuenta con el tiempo suficiente para poder seguir profundizando este tipo de situaciones.



A continuación, se presenta un gráfico que contiene información de los 5 países y las 5 variedades con mayor cantidad de reseñas, considerando los puntajes promedio a efectos de indagar sobre los gustos de los consumidores relacionados a las variables antes mencionadas.

En el gráfico se observa que Francia y Estados Unidos dominan destacan en varias variedades, lo que refleja su importancia en la industria vitivinícola mundial. Por otro lado, cada país tiene sus fortalezas en determinadas variedades y estilos de vino.

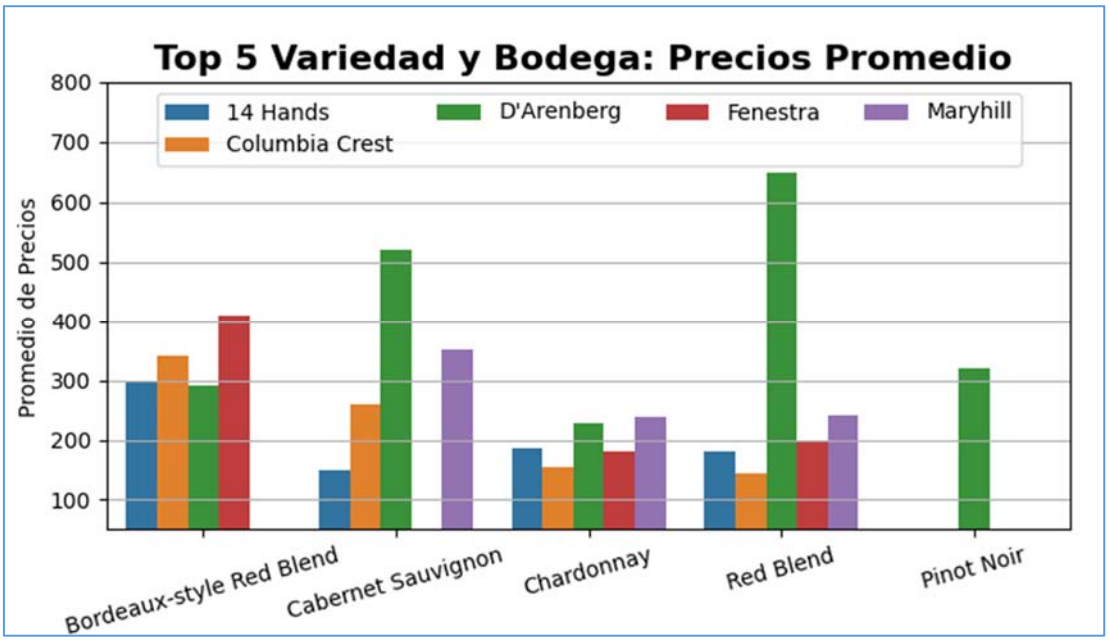


Este último grafico es similar al anterior, solo que en lugar de países de analizan bodegas, y en lugar de punto se consideran los precios promedios.

El gráfico revela que *D'Arenberg* es la bodega que, en general, ofrece los vinos con precios promedio más altos en las variedades analizadas. Esto sugiere que esta bodega posiciona sus vinos en un segmento de mercado premium, posiblemente debido a la calidad percibida, las técnicas de elaboración utilizadas y las estrategias de marketing.

Por otro lado, se observa una cierta especialización por parte de cada bodega. Por ejemplo, *Fenesra* destaca en la variedad de *Bordeaus-style Red Blend*, mientras que *Maryhill* tiene una fuerte presencia en la categoría de *Cabernet Sauvignon*.

Es importante tener en cuenta que los precios de los vinos son influenciados por una multitud de factores, y este análisis se basa únicamente en los datos proporcionados en el dataset.



En conclusión, este gráfico nos ofrece una visión general de los precios promedio de diferentes variedades de vino producidas por cinco bodegas específicas. Sin embargo, para una interpretación más detallada, se requeriría un análisis más profundo considerando factores adicionales y una mayor cantidad de datos.

5. Conclusiones

El análisis revela una compleja interacción entre las calificaciones de los vinos y sus precios a lo largo del tiempo. Si bien en los primeros años hubo una tendencia general al alza en ambas variables, las fluctuaciones posteriores sugieren que factores como las condiciones económicas, el clima y las tendencias del mercado influyen significativamente en las percepciones de los consumidores y en los precios. La tendencia a la baja en las calificaciones de algunos países, incluida Francia, puede indicar un cambio en las preferencias de los consumidores o una mayor competencia de las nuevas regiones vinícolas.

Hay una variación significativa de precios entre diferentes países, variedades de uva y bodegas, lo que sugiere que factores como el terroir, los costos de producción y la imagen de marca juegan un papel importante en los precios. El análisis de las combinaciones de países y variedades más populares revela que ciertos países se destacaron en variedades específicas.

El conjunto de datos destaca la diversidad del mercado mundial del vino. Austria surge como uno de los países con mejor desempeño en términos de calificaciones promedio, mientras que Estados Unidos domina tanto en términos de calificaciones como de precios. El Cabernet Sauvignon y el Pinot Noir son consistentemente populares entre los consumidores, y sus precios reflejan su alta demanda.

El malbec argentino tiene una fuerte presencia en el mercado global y el análisis mostró que bodegas argentinas, como Alta Vista y Catena Zapata, están produciendo malbec de alta calidad. Sin embargo, los precios promedio más altos para el malbec se encontraron en los Estados Unidos.

El análisis de la nube de palabras proporciona información sobre las preferencias de los consumidores, y términos como "afrutado", "picante" y "tanino" aparecieron con frecuencia. Esto sugiere que los consumidores están cada vez más interesados en vinos con perfiles de sabor complejos.

El análisis de múltiples calificaciones para el mismo vino destaca la naturaleza subjetiva de la cata de vinos y el potencial de una variación significativa en las calificaciones entre diferentes catadores.

En resumen, este trabajo proporciona una base sólida para comprender el complejo mundo del vino y las preferencias de los consumidores. Al combinar análisis estadísticos y visualización de datos, se han obtenido insights valiosos que pueden ser utilizados por productores, distribuidores y consumidores para tomar decisiones más informadas.