Texto, Pizarra

Descripción generada automáticamente

**Licenciatura en Ciencia de Datos**

**Trabajo practico Nro. 4**

**Seminario de practica en ciencia de datos**

**Aplicación de Ciencia de Datos a la industria vitivinícola**

**Autor: Donadío Gerónimo**

**Fecha de entrega: 17/11/2024**

Contenido

[Carga de datos: 2](#_Toc182740188)

[Tratamiento de datos faltantes: 3](#_Toc182740189)

[Interrogantes interesantes a explorar: 3](#_Toc182740190)

[¿ Cual es la variedad mas comúnmente resenada? 3](#_Toc182740191)

[¿ En que país la comunidad de aficionados al vino es mas activa? 4](#_Toc182740192)

[¿ Cual es son las variedades de vino mas caras, y cuales las mas baratas? 5](#_Toc182740193)

[¿ El origen del vino afecta en la calidad percibida? 6](#_Toc182740194)

[¿Cuáles son las bodegas mas resenadas? 7](#_Toc182740195)

[¿ A mas caro el vino mas puntos? 7](#_Toc182740196)

[¿ Como se distribuyen los vinos, cuales son sus precios? 8](#_Toc182740197)

[Conclusiones generales: 10](#_Toc182740198)

# 

# Carga de datos:

Carga del archivo: A la hora de cargar el archivo otorgado por el profesor se me presento un problema, el cual consistía en que no podía hacer por código que me quede en columnas, para solucionarlo, lo cargue en Excel y lo exporte como .xlsx, una vez en este formato no tuve problemas con la carga,

# Tratamiento de datos faltantes:

En cuanto a los valores faltantes empezare por los cuales decidí eliminar, estos son los que son menor al 1%, en las columnas country, province, variety, decidí e limarlos ya que son muy pocos como para que afecte a los resultados, no va a generar ningún tipo de distorsión.

Luego en el otro extremo están designation y región\_2, la primera tiene 30% y la segunda tiene 60% de null, esto es demasiado, decide aquí eliminar las columnas ya que perjudicarían el resto de las variables.

En el medio tenemos a Price con un 7%, aquí una buena opción es rellenar con la mediana, ya que Price no sigue una distribución normal. De las categóricas aquí tenemos a región\_1, taster\_name y taster\_twiter\_handle, aquí opte por llenar los valores null con desconocidos, ya que no afecta a mi análisis.

# Interrogantes interesantes a explorar:

## ¿Cuál es la variedad más comúnmente reseñada?

La variedad más comúnmente reseñada es el "Pinot Noir", seguida de cerca por "Chardonnay" y "Cabernet Sauvignon". Esto refleja probablemente las preferencias del mercado o el enfoque en estas variedades en las regiones productoras más reseñadas. Estos tres tipos son ampliamente reconocidos y consumidos a nivel global. La popularidad de variedades puede guiar decisiones sobre qué variedades producir en mayor volumen. También sugiere que hay un mercado competitivo, pero posiblemente lucrativo, para estas uvas en regiones donde aún no son predominantes. Para los consumidores estas variedades son las más probadas y evaluadas, lo que podría indicar una mayor confiabilidad en cuanto a la calidad. Sin embargo, también podría valer la pena explorar variedades menos comunes para encontrar opciones únicas y potencialmente más económicas.

Imagen que contiene Gráfico

Descripción generada automáticamente

## ¿En qué país la comunidad de aficionados al vino es más activa?

Estados Unidos domina abrumadoramente las reseñas de vinos, con más de 50,000 reseñas, mientras que Francia e Italia le siguen con cantidades significativamente menores. Esto puede indicar que la industria del vino en EE.UU. está más integrada con los sistemas de reseñas o que hay una mayor producción/distribución de vinos reseñados en ese país. El liderazgo de Estados Unidos en la cantidad de reseñas sugiere un mercado altamente activo y orientado hacia el consumidor. Esto representa una gran oportunidad para exportar vinos a este país, optimizando estrategias de marketing digital para plataformas populares de reseñas. Para los consumidores es de suma importancia también porque la abundancia de reseñas en Estados Unidos implica una gran cantidad de información disponible para elegir vinos, pueden aprovechar esto para tomar decisiones más informadas sobre calidad y relación precio-beneficio.

Aplicación

Descripción generada automáticamente con confianza media

## ¿Cuál es son las variedades de vino mas caras, y cuales las más baratas?

Ramisco" destaca notablemente como la variedad con el precio más alto, separándose significativamente de otras variedades como "Terrantez" y "Francisca". Esto sugiere que "Ramisco" podría ser una variedad extremadamente rara o producida bajo condiciones específicas que justifican su alto costo. Además, otras variedades como "Rosenmuskateller" y "Malbec-Cabernet" también presentan precios elevados, aunque no alcanzan los niveles de "Ramisco". Esta información puede ser valiosa para los productores interesados en posicionar su producto en el mercado premium, resaltando la exclusividad o las características únicas de estas variedades. Para los consumidores, estas variedades representan opciones de lujo que probablemente ofrezcan experiencias únicas, aunque no necesariamente sean accesibles para todos.

Imagen que contiene piano

Descripción generada automáticamente

## ¿El origen del vino afecta en la calidad percibida?

La distribución de puntos en los vinos varía ligeramente según el país. En general, la mediana de los puntos para la mayoría de los países está alrededor de 87-89, lo cual indica una consistencia en la calidad de los vinos. Sin embargo, hay ciertos países, como Francia y Australia, con distribuciones más amplias, lo que podría reflejar una gama más variada de calidades de vino. Para los consumidores países con distribuciones amplias de calidad ofrecen opciones tanto premium como accesibles, permitiendo a los consumidores experimentar sin romper el presupuesto.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

## ¿Cuáles son las bodegas más reseñadas?

"Wines & Winemakers" es la bodega más reseñada, lo cual puede sugerir una fuerte presencia en el mercado o estrategias de marketing exitosas. Las bodegas más mencionadas están repartidas por distintas regiones productoras, destacando una mezcla de bodegas locales y más reconocidas globalmente. Estar entre las bodegas más reseñadas no solo indica popularidad, sino también visibilidad. Invertir en estrategias de marketing podría ser una opción interesante ya que la visibilidad de la compañía es de lo más importante. Para los consumidores las bodegas más reseñadas suelen ser una apuesta segura en cuanto a calidad y consistencia, pero explorar opciones menos conocidas podría revelar joyas ocultas.

Imagen que contiene Gráfico

Descripción generada automáticamente

## ¿A mas caro el vino más puntos?

A medida que aumenta el precio, también tienden a subir los puntos otorgados por los evaluadores. Sin embargo, esta relación no es completamente lineal ni determinante. En los rangos de precios más altos, hay una considerable dispersión en los puntajes, lo que indica que no todos los vinos caros son percibidos como de alta calidad. Por otro lado, en los rangos de precios bajos y medios, los puntajes suelen concentrarse entre 85 y 90, lo que implica que es posible encontrar vinos bien evaluados sin necesidad de gastar grandes cantidades de dinero. Para los productores, esto resalta la importancia de mantener la consistencia en la calidad percibida, independientemente del precio, ya que la percepción de calidad puede justificar precios altos solo si se mantiene constante. Para los consumidores, la dispersión en los puntajes a precios altos sugiere que deben investigar cuidadosamente antes de adquirir un vino caro, ya que el precio por sí solo no garantiza una experiencia superior. En cambio, en los precios medios, hay buenas oportunidades para disfrutar de vinos bien evaluados a precios razonables

Texto

Descripción generada automáticamente con confianza media

## ¿Como se distribuyen los vinos, cuáles son sus precios?

La mayoría de los vinos tienen precios concentrados entre 40 y 250, con una distribución que se inclina hacia la derecha, es decir, la distribución es asimétrica con una "cola" más larga a la derecha. Esto indica que mientras la mayoría de los vinos tienen precios relativamente bajos o medios, hay algunos con precios mucho más altos, pero son menos frecuentes. Mas adelante veremos este sesgo a la derecha o positivo y como tratarlo. El pico más alto de la distribución está alrededor de los 150-200, lo que sugiere que este es el rango de precio más común entre los vinos de este conjunto de datos. A continuación, un gráfico para poder de forma más clara esta concentración.

Gráfico, Histograma

Descripción generada automáticamente

después de la transformación logarítmica, los precios, que antes estaban extremadamente sesgados, ahora están distribuidos de manera más uniforme, con la mayor parte de los precios logarítmicos en el rango de 5.5 a 6.5 (en escala log). Esto indica que, tras la transformación, los precios están en su mayoría dentro de un rango más controlado, y los outliers tienen menos impacto en la estructura general de los datos.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

La transformación logarítmica es útil para analizar este tipo de distribuciones, ya que permite visualizar mejor la concentración de precios en el rango bajo-medio y la influencia de los precios extremadamente altos, esto nos muestra más claramente que la mayor parte de los vinos tienen precios relativamente bajos, y que los precios más altos son mucho menos comunes, pero tienen un efecto importante en la distribución general.

Para corregir este sesgo podemos aplicar una transformación logarítmica, es necesario corregir ya que el sesgo original nos da un valor de 18, arriba de 2 ya no es aceptable, para después correr modelos de machine learning, con la transformación bajamos el sesgo a 0,6 que es significativo, pero muchísimo mejor que antes, a continuación, el grafico de cómo queda la variable después del tratamiento.

Gráfico, Histograma

Descripción generada automáticamente

La concentración de precios en rangos medios (como 40-250) puede guiar estrategias de fijación de precios. Por otro lado, el análisis de outliers podría revelar oportunidades en el mercado de lujo. La transformación logarítmica permite a los consumidores comprender mejor las opciones de precio. Los precios en el rango bajo-medio ofrecen una excelente calidad relativa sin el costo de los vinos premium.

# Conclusiones generales:

El análisis de los datos sobre vinos nos permite extraer conclusiones significativas tanto para la industria como para los consumidores. En términos generales, hemos visto cómo la ciencia de datos aplicada a la industria vitivinícola puede generar valor al identificar patrones en precios, puntajes y preferencias de los consumidores. Primero, se destaca la importancia de ciertas variedades como el "Pinot Noir", "Chardonnay" y "Cabernet Sauvignon", que lideran en reseñas, reflejando la demanda global y la necesidad de los productores de centrarse en estas uvas en regiones estratégicas. Sin embargo, también se identificaron variedades menos comunes, como "Ramisco", que, a pesar de su rareza y alto costo, ofrecen oportunidades en mercados exclusivos. La actividad de la comunidad de aficionados al vino varía entre países, con Estados Unidos liderando, lo que refleja un mercado vibrante y conectado digitalmente. Esto representa una oportunidad significativa para los productores interesados en exportar y aumentar su visibilidad en plataformas de reseñas. En cuanto al precio y la calidad, el análisis muestra que, aunque los vinos más caros tienden a recibir mejores puntuaciones, esto no siempre es una regla estricta. Los vinos en rangos de precios medios ofrecen una excelente relación calidad-precio, lo que brinda a los consumidores opciones accesibles sin comprometer la experiencia. La transformación logarítmica aplicada a los precios demostró ser una herramienta valiosa para manejar el sesgo en la distribución, permitiendo un análisis más claro y facilitando el uso de modelos predictivos. Este tratamiento también ayuda a los consumidores a entender mejor la relación entre precio y calidad. Finalmente, el análisis desarrollado en este trabajo puede ser utilizados como base para sistemas de recomendación personalizados y estrategias de marketing orientadas a segmentos específicos del mercado. Esto no solo mejora la experiencia del consumidor, sino que también optimiza las decisiones comerciales de los productores. En conclusión, la aplicación de técnicas de ciencia de datos a este sector no solo permite profundizar en las dinámicas de mercado, sino también ofrece herramientas prácticas para mejorar la competitividad de los productores y la satisfacción de los consumidores. Este trabajo demuestra el potencial transformador de los datos en una industria tan tradicional como la del vino.