
# “Análisis de base de datos: Wine reviews”

## Tarea final

Curso: “Introducción a la Ciencia de Datos”

Año: 2023

Grupo: 5

Autores:

- Agustín Porley Santana
- Julián Rodríguez

## 1. Introducción a la base de datos

La base de datos a analizar tiene el nombre de "Wine reviews". La misma es una base en formato csv compuesta de las siguientes diez columnas: country, description, designation, points, price, province, region_1, region_2, variety, winery.
En la columna "country" se puede apreciar el país de origen del vino, la columna "description" contiene la reseña del vino realizada por un sommelier, "designation" contiene el nombre fantasía del vino, "points" indica el puntaje que obtuvo el vino, "price" indica el precio del mismo en dólares, "province" informa sobre la provincia donde está la bodega, "region_1" indica la región en la cual se cosecha la uva, "region_2" dentro de la primera región indica una sub área especifica de donde proviene el vino, "variety" indica la variedad de uva de la cual proviene y por último, "winery" indica la bodega.
La base de datos cuenta con 150.930 registros de vinos de todo el mundo.

## 2. Calidad de datos

Realizando una inspección visual se aprecia que la base de datos posee problemas de calidad de datos.
Como primer acercamiento a la misma, se decide analizar la cantidad de datos faltantes. En la Figura 1 se aprecia el porcentaje de datos disponible de cada columna en la base de datos.

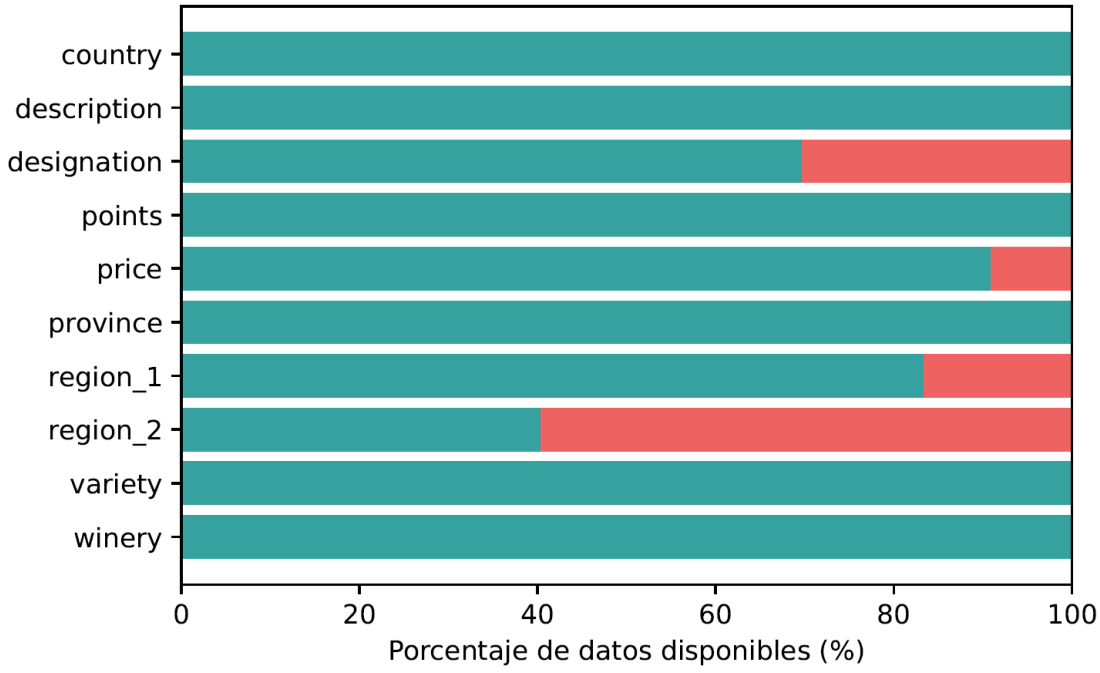


**Figura 1: Porcentajes de datos disponibles por columna**

Se observa como hay faltantes de datos importantes en algunas columnas, aunque no son críticos para los análisis que se pretenden realizar.

Como siguiente análisis, se buscan filas repetidas. Al hacer un recuento de datos, se obtiene como resultado 53.079 filas duplicadas. Casi un tercio de los datos se encuentran en esas condiciones. Para esta base de datos en particular se investiga también si no existen datos que sin ser duplicados, sólo se difieren en alguna de sus columnas. Como la columna "description" es una cadena de texto sin un formato estándar, podría darse el caso de vinos ingresados dos veces con reseñas diferentes, o que presenten alguna diferencia mínima en su texto. Por lo tanto, buscando datos que sean duplicados sin tener en cuenta la columna 'description', se encuentran 56.059 registros que cumplen con la mencionada condición. Esto quiere decir que existen 2.980 registros de vinos que podrían haber sido cargados más de una vez con una reseña diferente.

Las reseñas y los puntajes son datos subjetivos, de acuerdo a experiencias sensoriales y que pueden incluir sesgos, por lo que podría resultar en problemas de calidad de datos. Dos personas pueden escribir dos reseñas diferentes para un mismo vino. La información del lote de vino no está siendo brindada, por lo que dichos datos, además de ser subjetivos, pueden variar para un vino dependiendo del lote.

En particular la base de datos cuenta con vinos uruguayos, por lo que podemos verificar localidades que nos resultan familiares. Se observa que hay registros cuya columna 'province' presenta el valor 'Uruguay', lo que se entiende que no es una localidad válida. Se realiza la consulta sobre toda la base para identificar los registros que cuentan con la columna 'country' igual que la columna 'province' y se encuentran 386 filas con estas características, lo que debería corroborarse si no se trata de un problema de calidad.

## 3. Visualizaciones

Se puede realizar un histograma de precios para observar si la base cuenta con determinado rango de precios. Además, se podría obtener el puntaje promedio de cada rango de precio para conocer si realmente los vinos caros valen la pena.
Por otro lado, se puede generar un histograma normalizado de puntajes para observar el rango y el promedio general. Se puede solapar a dicho histograma uno que contenga los vinos que se producen en una región determinada. De esta manera, se puede apreciar la calidad de los vinos de la región con respecto a la media global. Por ejemplo se podría realizar, solamente con los tipos de vinos que se producen en Uruguay y ver cómo se posicionan a nivel mundial.

Para poder apreciar si existe una variedad de vino que tenga una mayor calidad se podría generar un gráfico de barras con el promedio de puntajes de cada variedad. Esto ayudaría a la hora de elegir qué variedad de vino comprar para un evento.
Para aquellos países que cuentan con más datos, se podría intentar visualizar el ranking de puntajes por localidad y poder deducir si existe alguna zona idónea para la producción transformando las localidades en coordenadas y ubicándolas en un mapa. Esta visualización aporta información relevante a posibles inversores que quieran apostar a la producción de vino. Además, permitiría generar políticas de desarrollo de turismo enológico en lugares que teniendo buenos vinos aún no han sido bien explotados.

## 4. PCA

Usando la metodología PCA con las columnas de precio y puntaje, se lograría obtener en la primer componente la relación general de precio-puntaje, y en la segunda componente se obtiene qué tan alejado se encuentra cierto vino de la relación general. Utilizando la segunda componente principal se puede determinar si un vino es mejor o peor con respecto a la generalidad. Esto podría ser un dato interesante para restaurantes que quieran ofrecer en su carta vinos de calidad a un buen precio, o incluso teniendo márgenes de ganancia altos.
En particular se podrían graficar los vinos uruguayos y ver cómo se comportan con respecto al mercado donde fueron relevados.

## 5. Entrenamientos de modelos para predicción

Se puede entrenar un modelo a partir de las reseñas que intente predecir qué variedad de vino es de acuerdo a su descripción. También se podrían generar rangos de puntajes, e intentar predecir qué puntaje tendría de acuerdo a su reseña.

Para ellos, se debería limpiar el texto de las reseñas extrayendo signos de puntuación y 'stop words'. Luego se usarían técnicas como 'bag of words' o 'td-idf' para generar representaciones de las palabras. A continuación, se podría entrenar un modelo, como por ejemplo Naive Bayes, para predecir en función de las palabras alguna etiqueta, ya sea variedad de vino o rango de puntuación.

## 6. Fuente

Base de datos : Wine reviews
Fuente: https://www.kaggle.com/datasets/zynicide/wine-reviews