

WINE DATABASE

Proyecto SQL
para Ironhack Data Analytics

Alicia Caminero
Andrea Lafarga
Gema Villena
Silvia Alonso



ORGANIZACIÓN DEL PROYECTO

- 1 Elección del tema
- 2 Recopilación de datos.
- 3 Limpieza de datos.
- 4 Creación de tablas.
- 5 Creación del diagrama entidad-relacion.
- 6 Importar datos.
- 7 Creación de consultas.
- 8 Creación de gráficos.
- 9 Conclusiones.

FROM DATA TO INSIGHT

1. ELECCIÓN DEL TEMA

Nos decidimos por el tema del vino por ser un tema con muchos datos disponibles para su uso y consulta, y con preguntas interesantes que analizar.

Hipótesis iniciales

¿Son los vinos más caros los mejor valorados?

¿Qué zonas cuentan con mayor cantidad de bodegas?

¿Cuál es el precio medio de los vinos peor valorados?



2. RECOPILACIÓN DE DATOS

Tras explorar varios datasets y realizar algún web scrapping en páginas sobre vinos, decidimos utilizar los datos que provenían del dataset “Spanish Wine Quality Dataset” obtenido de “<https://www.kaggle.com/datasets/fedesoriano/spanish-wine-quality-dataset>”, por ser estos los más completos y adecuados para nuestro propósito.

3. LIMPIEZA DE DATOS

Se procedió a la limpieza de los datos usando Python para eliminar los nulos y las columnas que no nos interesaban.

4. CREACIÓN DE TABLAS

WINES

wine
year
id_winery
id_type
id_designation
id_wine

WINERIES

winery
country
region
id_winery

RATINGS

id_wine
rating
num_reviews
price
body
acidity
id_rating

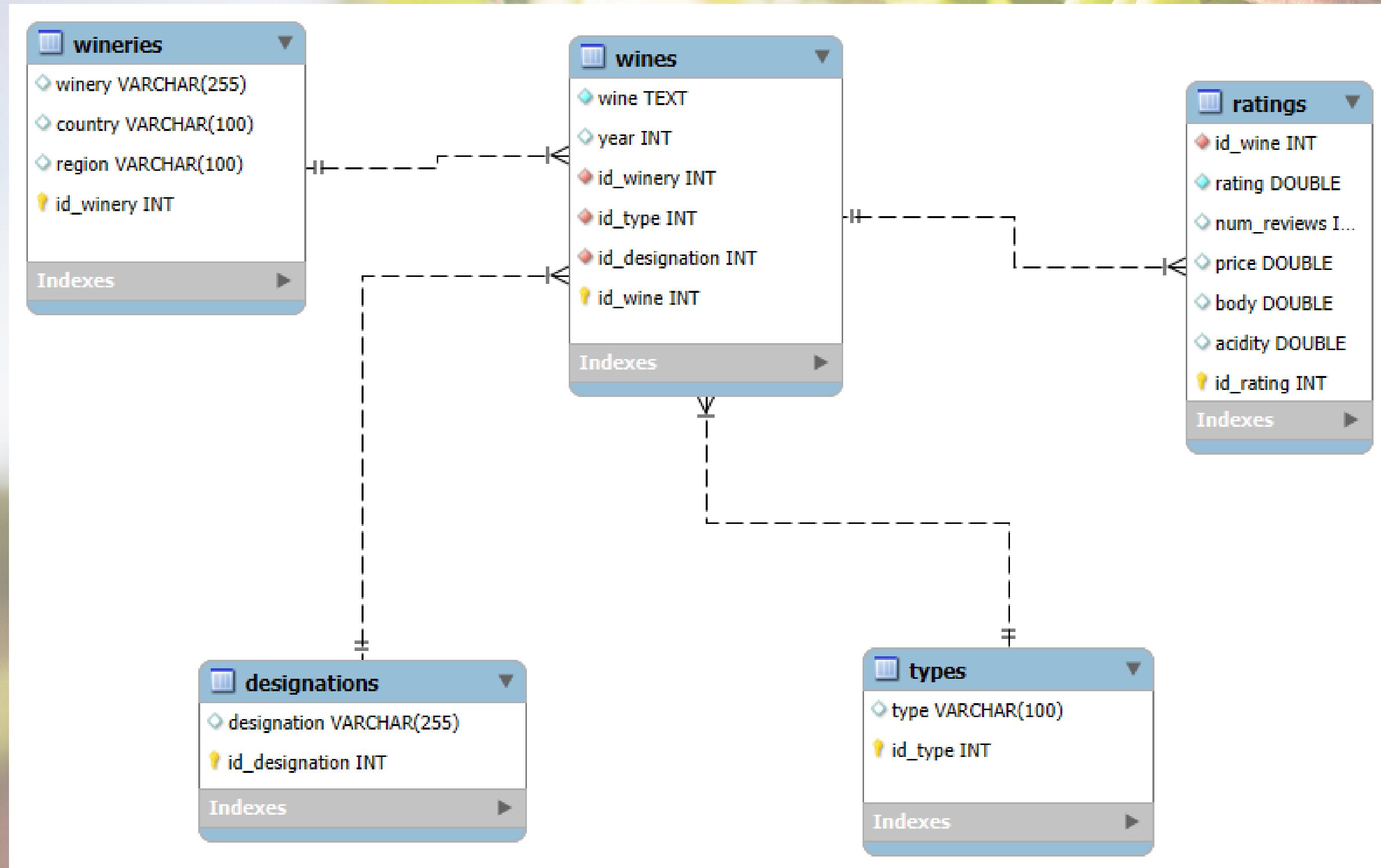
TYPES

type
id_type

DESIGNATIONS

designation
id_designation

5. MODELO ENTIDAD RELACIÓN (ERD)



6. IMPORTAR DATOS

Importamos los datos para nuestras tablas de los archivos csv en MySQL WorkBench usando la opción de la que dispone el programa para realizar esta acción concreta (Table Data Import Wizard).



7. CREACIÓN DE CONSULTAS

Promedio de los precios por tipo de vino

```
SELECT wine_type.wine_type, AVG(ratings.price) AS avg_price  
FROM wine_type  
JOIN wines ON wine_type.id_type = wines.id_type  
JOIN ratings ON wines.id_wine = ratings.id_wine  
GROUP BY wine_type.wine_type;
```

Bodegas que producen más de un tipo de vino

```
SELECT wineries.winery, wine_type.wine_type, COUNT(wines.id_wine) AS total_wines  
FROM wineries  
JOIN wines ON wineries.id_winery = wines.id_winery  
JOIN wine_type ON wines.id_type = wine_type.id_type  
GROUP BY wineries.winery, wine_type.wine_type  
HAVING COUNT(wines.id_wine) > 1  
ORDER BY total_wines DESC;
```

Los 10 vinos con mejor relación calificación-precio y su calificación media

```
SELECT wines.wine, ratings.rating, ratings.price, AVG(ratings.rating) AS avg_rating
FROM wines
JOIN ratings ON wines.id_wine = ratings.id_wine
GROUP BY wines.wine, ratings.rating, ratings.price
ORDER BY ratings.rating / ratings.price DESC
LIMIT 10;
```

Vinos más antiguos por denominación de vino y su calificación media

```
SELECT designations.designation, wines.wine, wines.wine_year, AVG(ratings.rating) AS avg_rating
FROM designations
LEFT JOIN wines ON designations.id_designation = wines.id_designation
LEFT JOIN ratings ON wines.id_wine = ratings.id_wine
GROUP BY designations.designation, wines.wine, wines.year
ORDER BY wines.year ASC;
```

Total de vinos y calificación promedio por región

```
SELECT  
    wi.region AS region_name,  
    COUNT(w.id_wine) AS total_wines,  
    AVG(r.rating) AS avg_rating  
FROM wines w  
JOIN wineries wi ON w.id_winery = wi.id_winery  
JOIN ratings r ON w.id_wine = r.id_wine  
GROUP BY wi.region  
ORDER BY total_wines DESC;
```

Promedio de los precios por tipo de vino

```
SELECT ratings.body, ratings.acidity, AVG(ratings.rating) AS avg_rating  
FROM ratings  
GROUP BY ratings.body, ratings.acidity  
ORDER BY avg_rating DESC;
```

Distribución de precios y relación calidad-precio por tipo de vino

```
SELECT
    wt.wine_type,
    MIN(r.price) AS min_price,
    MAX(r.price) AS max_price,
    AVG(r.price) AS avg_price,
    AVG(r.rating / r.price) AS avg_quality_price_ratio
FROM wine_type wt
JOIN wines w ON wt.id_type = w.id_type
JOIN ratings r ON w.id_wine = r.id_wine
GROUP BY wt.wine_type
ORDER BY avg_quality_price_ratio DESC;
```

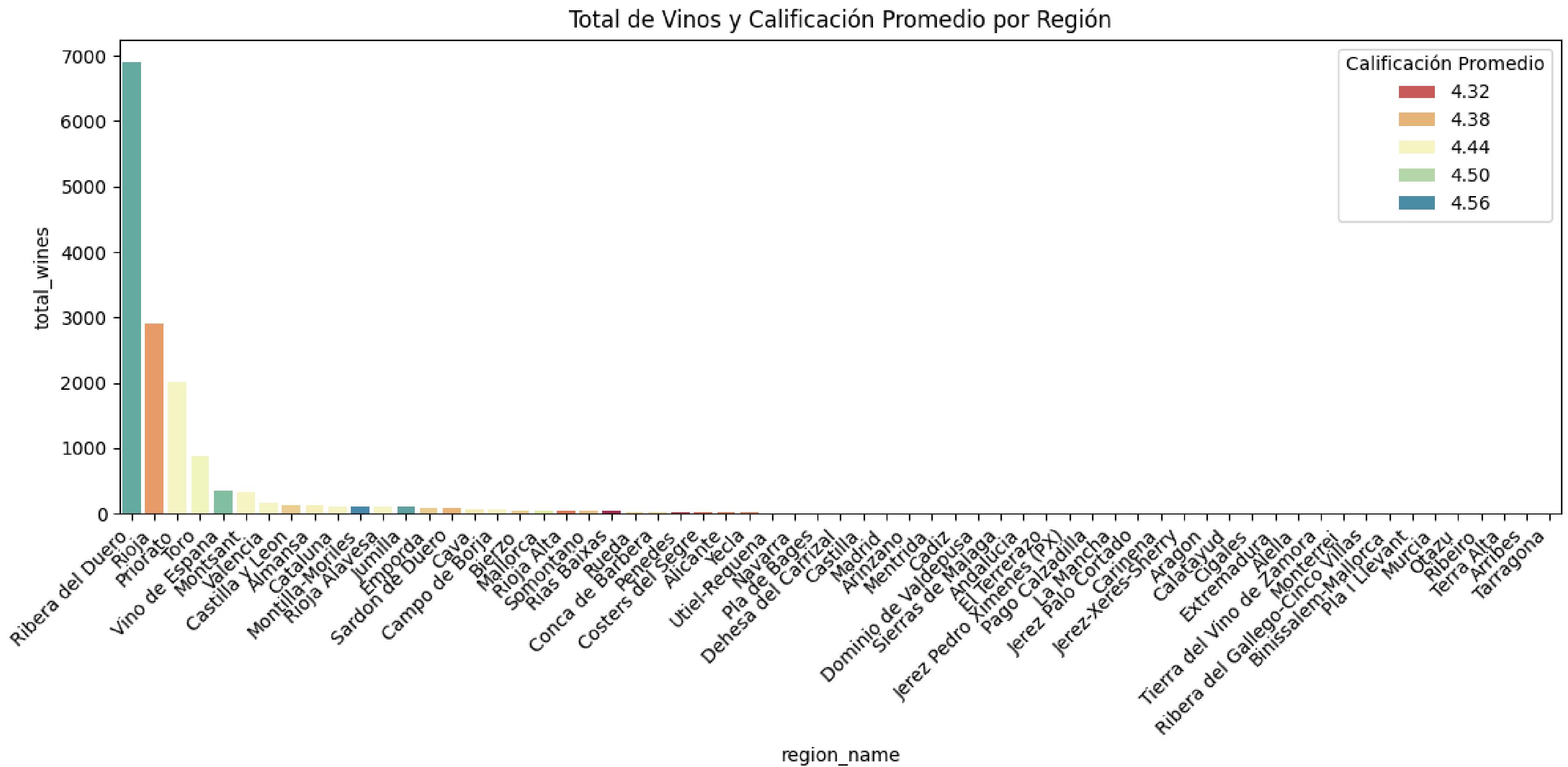
Años con mayor y menor producción de vinos

```
SELECT wines.year, COUNT(wines.id_wine) AS total_wines
FROM wines
GROUP BY wines.year
ORDER BY total_wines DESC;
```

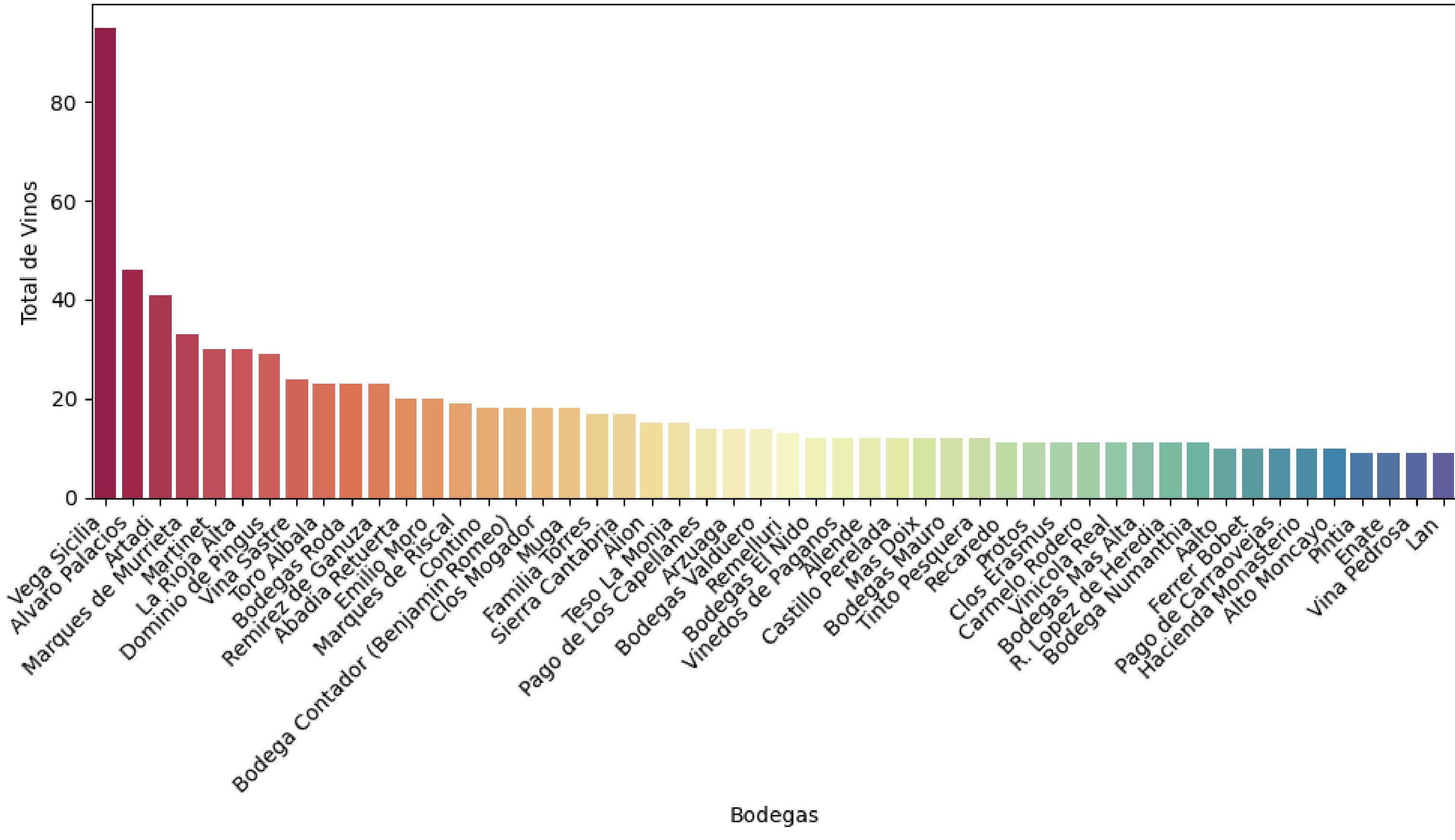
Análisis combinado de vinos por rango de calificaciones y precios

```
SELECT
CASE
    WHEN r.rating >= 4.5 THEN 'Excelente'
    WHEN r.rating >= 4 THEN 'Muy bueno'
    WHEN r.rating >= 3.5 THEN 'Bueno'
    WHEN r.rating >= 3 THEN 'Regular'
    ELSE 'Malo'
END AS rating_category,
CASE
    WHEN r.price >= 100 THEN 'Expensive'
    WHEN r.price >= 50 THEN 'Moderate'
    ELSE 'Affordable'
END AS price_category,
COUNT(*) AS total_wines,
ROUND(AVG(r.price), 2) AS avg_price,
ROUND(AVG(r.rating), 2) AS avg_rating,
ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM ratings), 2) AS percentage
FROM ratings r
JOIN wines w ON r.id_wine = w.id_wine
GROUP BY rating_category, price_category
ORDER BY percentage DESC;
```

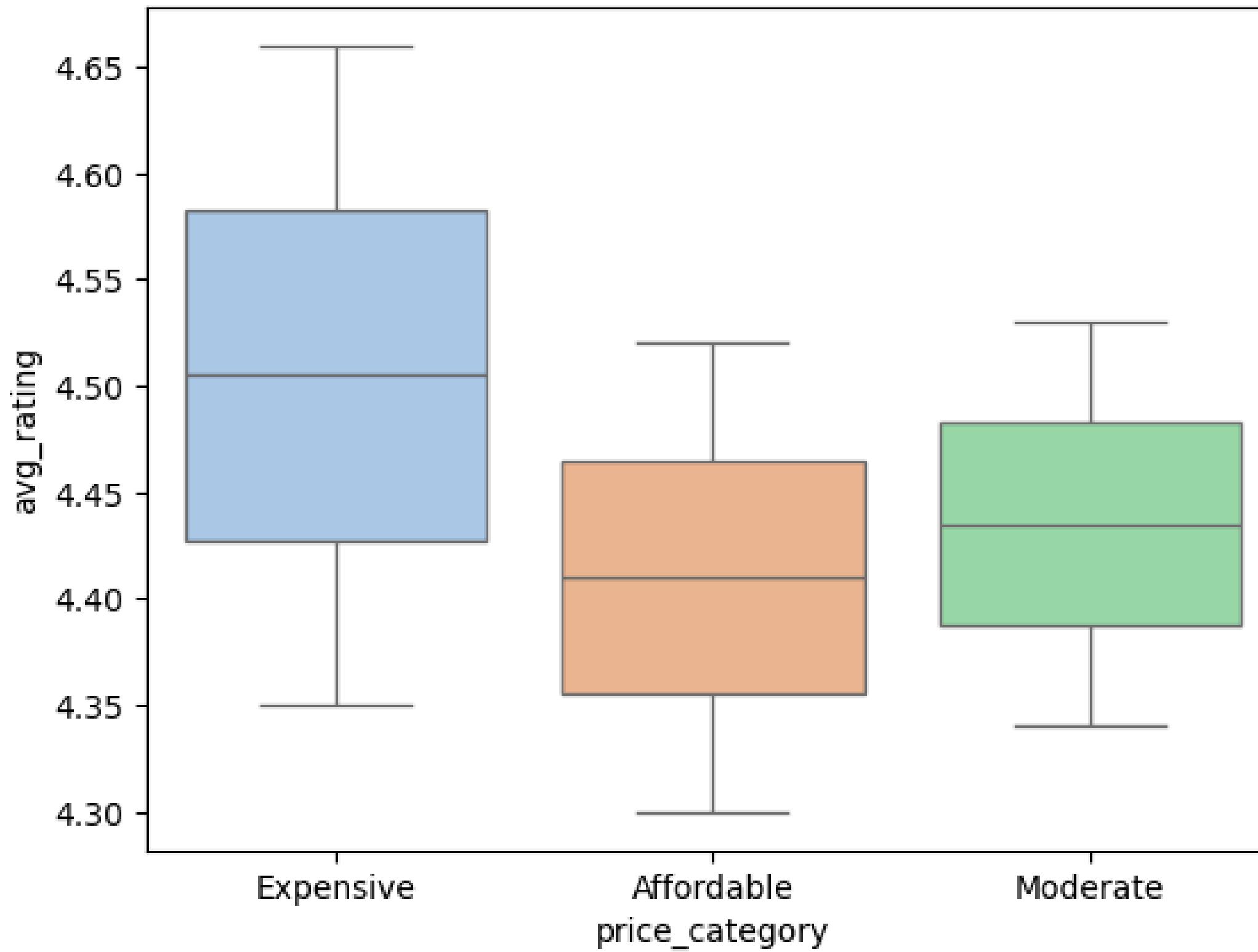
8. GRÁFICOS



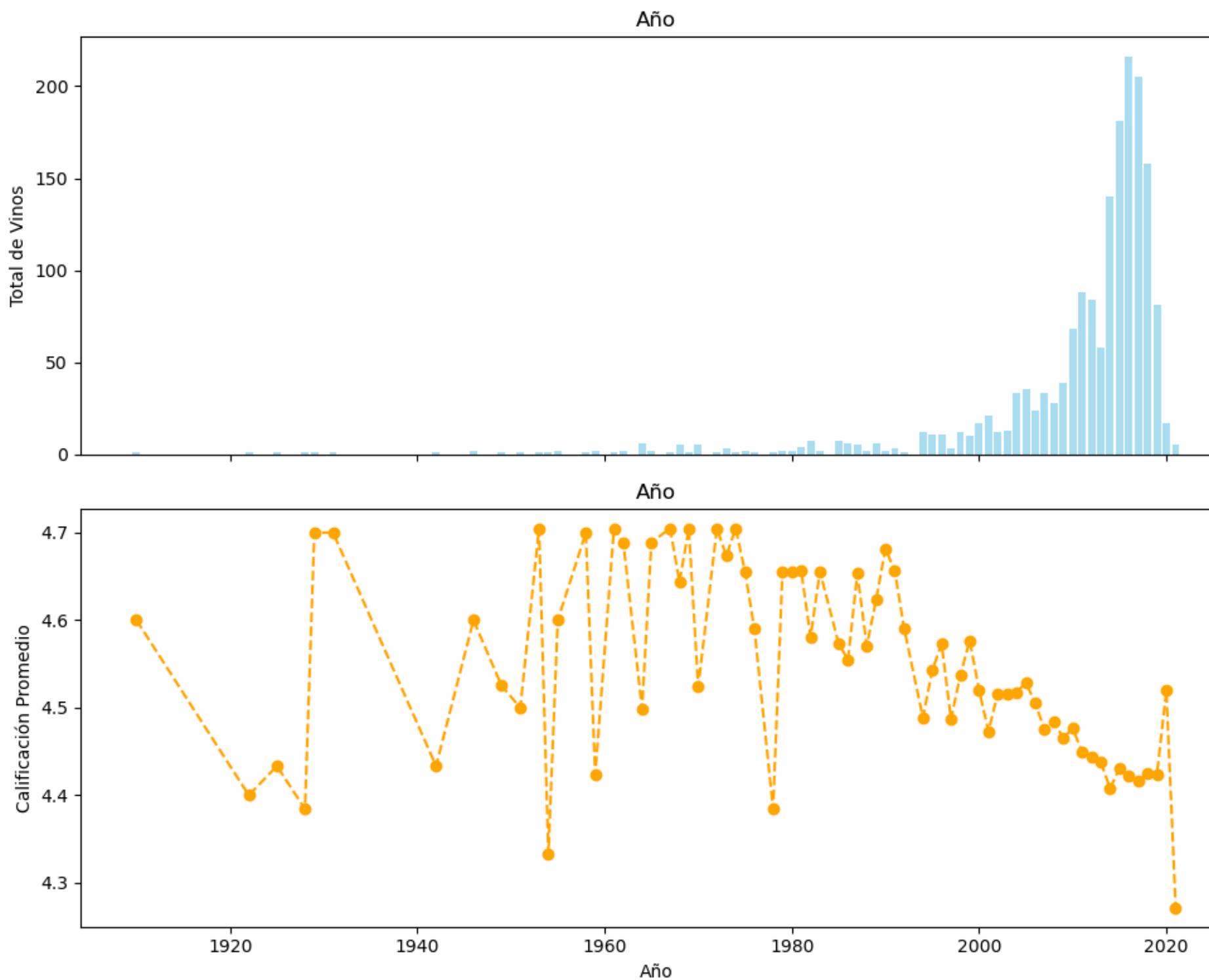
Top 50 Bodegas por Total de Vinos



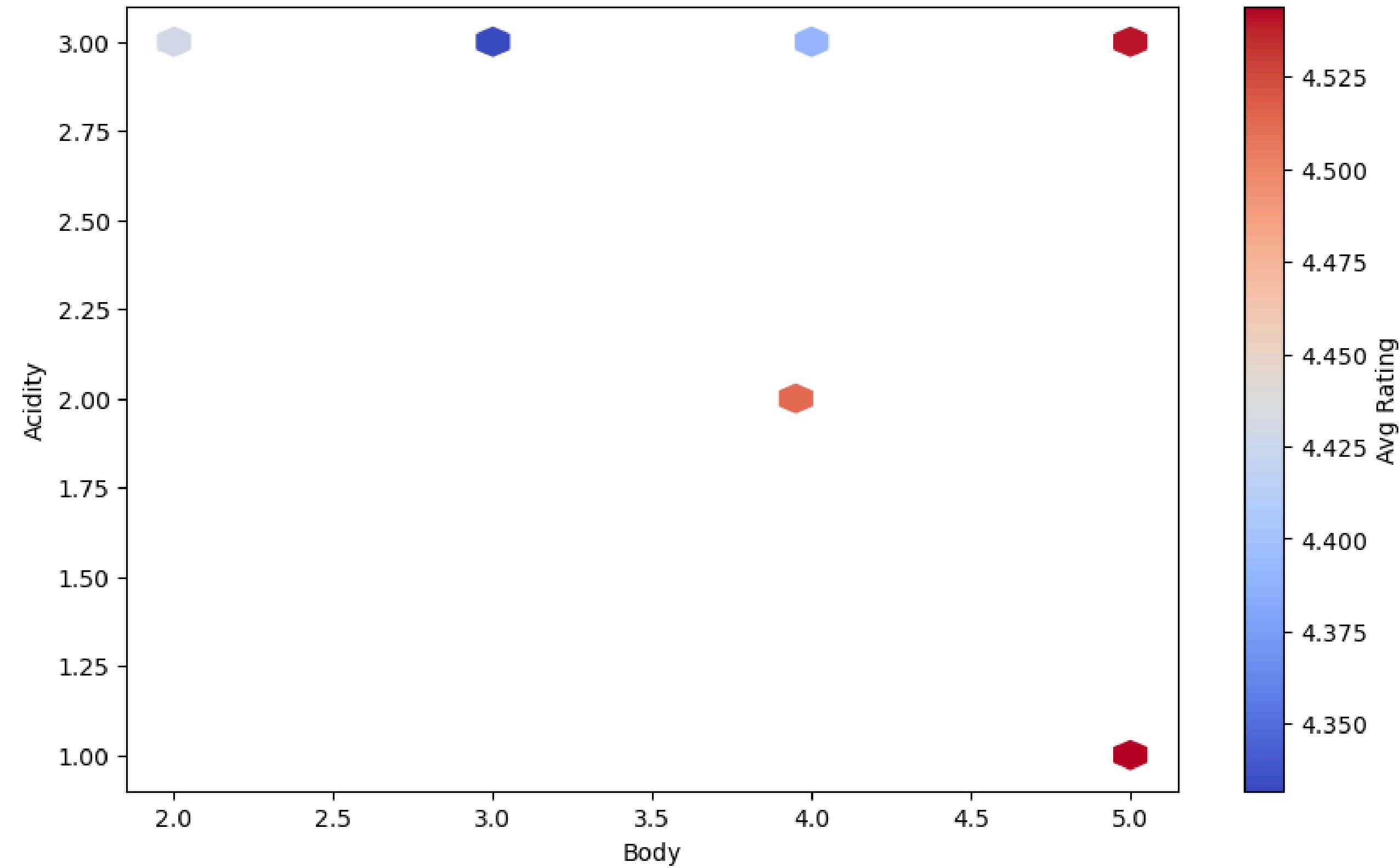
Relación entre la categoría de precios y el promedio de las calificaciones

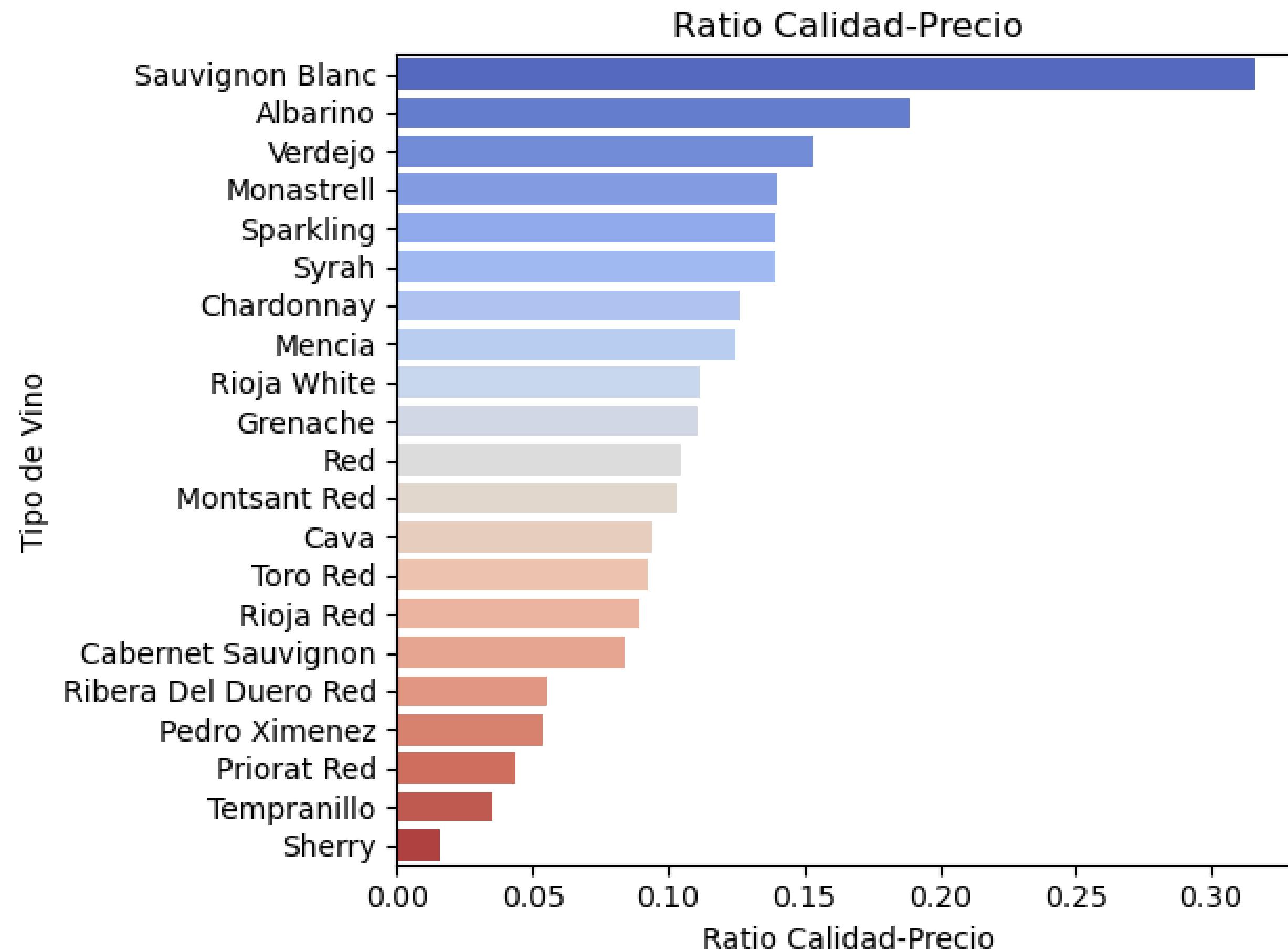


EVOLUCIÓN DEL TOTAL DE VINOS Y CALIFICACIÓN POR AÑO

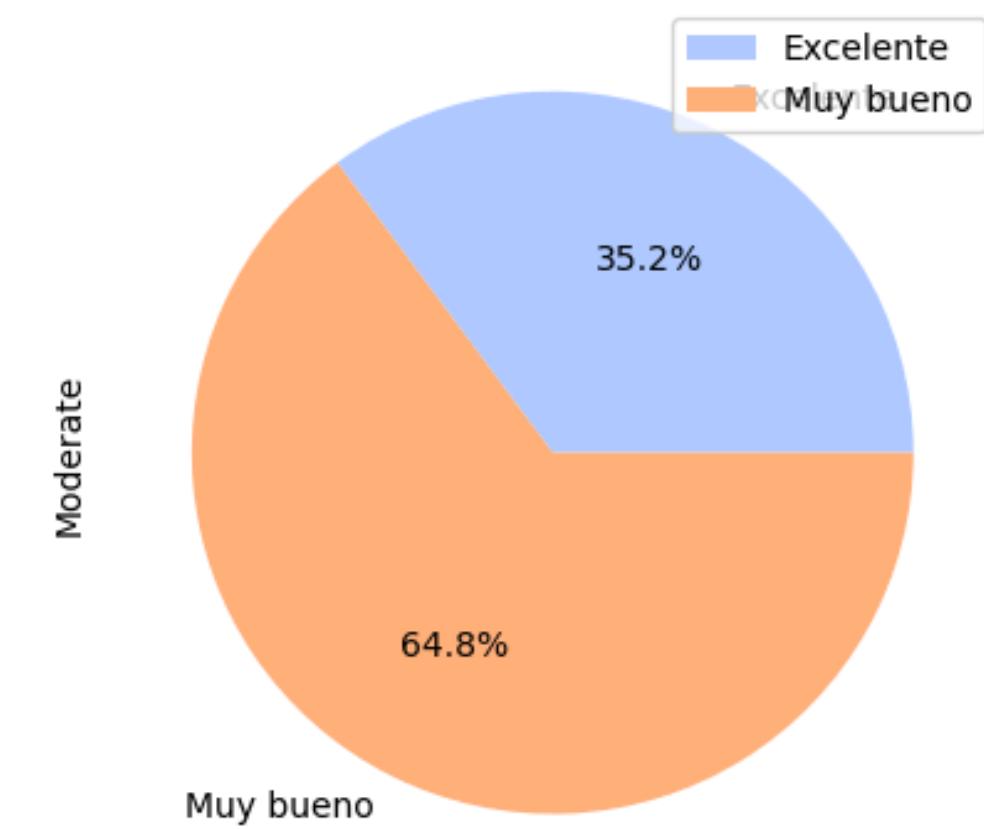
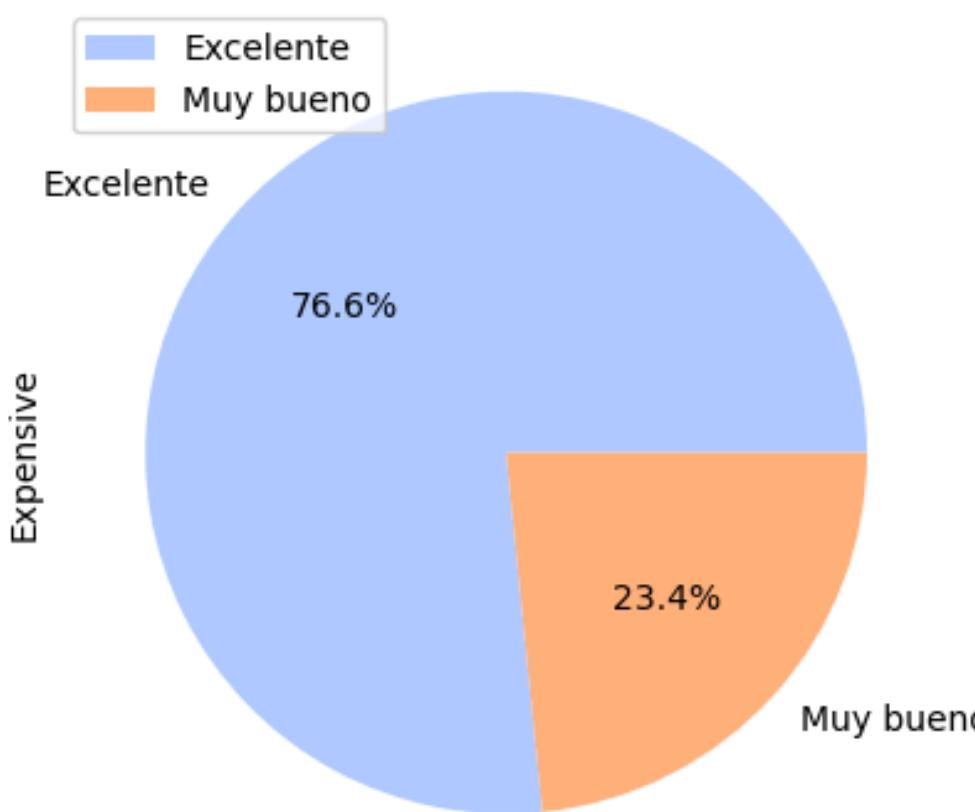
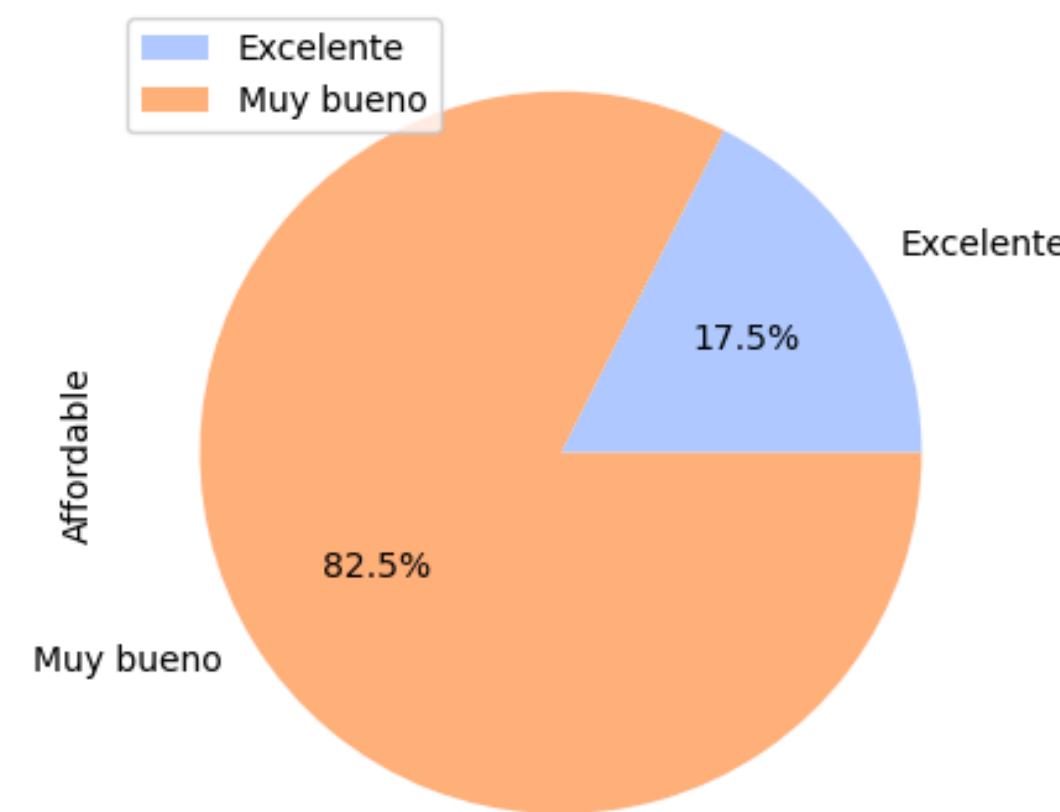


Relación entre el cuerpo, la acidez y el promedio de las calificaciones





Distribución de las calificaciones por el precio



9. CONCLUSIONES

La calidad de los datos es clave para obtener resultados significativos

La limpieza del dataset, incluyendo la imputación de valores faltantes y la eliminación de duplicados, fue un paso fundamental para garantizar la precisión y relevancia de los análisis.

El uso de SQL facilita la extracción de información valiosa

Las consultas implementadas permiten responder preguntas complejas de manera eficiente, proporcionando una base sólida para la toma de decisiones y el desarrollo de análisis más avanzados.

El proyecto abre puertas a exploraciones futuras

Este proyecto muestra la utilidad de las bases de datos para analizar tendencias, y también establece un punto de partida para incorporar datos adicionales en el futuro y añadir visualizaciones con Tableau o Power BI que enriquezcan aún más el análisis.



**iMUCHAS
GRACIAS!**

<https://github.com/datasilvia/WineDatabase>