



WINE VARIETY

Documentación del Proyecto de Análisis de Vinos



Contexto

Cliente: Empresa importadora de vinos

Objetivo general: Optimizar el portafolio de vinos importados, identificando oportunidades de alto valor.

Esto incluye:

- Vinos con excelente relación calidad/precio
- Regiones emergentes prometedoras
- Bodegas con potencial para posicionarse en el mercado premium internacional



Proceso de Limpieza de Datos



1. Selección del Dataset

El dataset original incluía las siguientes variables:

- **Categorías geográficas:** country, province, region_1, region_2
- **Producto:** variety, winery, designation
- **Valoración:** points, price
- **Catador:** taster_name, taster_twitter_handle
- **Descripción sensorial:** description



2. Limpieza General

a. Estandarización

- Renombrado de columnas para consistencia

- Conversión de tipos de datos

b. Eliminación de duplicados

- Se eliminaron entradas repetidas

c. Eliminación de columnas innecesarias

Columnas eliminadas por tener demasiados nulos o ser irrelevantes:

- region_1, region_2, designation, taster_twitter_handle



3. Tratamiento de valores nulos

price

- Imputación por mediana agrupada por combinación de variety + province
- Para casos sin agrupación posible: se imputó con la mediana general
- Valores extremos revisados manualmente: se comprobaron como válidos tras verificación

Otras columnas

- Valores nulos restantes (<0.6%) fueron eliminados por representar un riesgo bajo de sesgo



Extracción y Análisis de Sabores (NLP)



Técnica utilizada

- Procesamiento del texto de la columna description usando **spaCy**
- Tokenización y lematización
- Filtro por categorías gramaticales:
 - Solo se conservaron **adjetivos (ADJ)** y **sustantivos (NOUN)** con longitud > 3 letras

Objetivo

- Extraer términos relacionados con sabores, aromas o características sensoriales
- Convertirlos en variables estructuradas (dummies) para análisis descriptivo y predictivo

Resultados: 50 Sabores/Descriptorios más comunes

#	Sabor	Frecuencia
0	wine	82,911
1	flavor	68,737
2	fruit	63,373
3	palate	37,971
4	aroma	36,999
5	acidity	34,890
6	tannin	32,935
7	cherry	32,810
8	finish	32,594
9	black	28,980
...
46	chocolate	8,508
47	currant	8,492
48	character	8,448
49	vineyard	7,875




(ver documento completo para lista extendida)

Filtrado de términos no sensoriales

Se procederá a eliminar manualmente términos que no representen sabores u olores, tales como:

- wine, palate, note, nose, year, blend, vineyard, character, good, white, etc.

Próximos pasos

1.  **Filtrar sabores sensoriales reales** de la lista de palabras más frecuentes
2.  **Crear archivo CSV** con columnas dummy (1/0) por sabor, por vino
3.  Usar los sabores como variables de entrada para:
 - a. Análisis exploratorio: clusters, estilos
 - b. Modelos predictivos de precio o puntuación
 - c. Sistemas de recomendación personalizados

Como tenemos un problema al relacionar los datasets por tema de cardinalidad y valores repetidos en little cambiaremos la relación a un id que importaremos de unnamed 0 un index que se me creó sin querer que le cambiaremos el nombre

ANALISIS

? Preguntas clave de negocio

1. Relación calidad/precio

- ¿Qué países o regiones tienen los vinos con mejor puntuación ajustada al precio?
- ¿Qué variedades ofrecen mejor relación puntos/precio?
- ¿Existen vinos baratos con puntuaciones altas?

Insight esperado:

“Los vinos portugueses y argentinos ofrecen una relación calidad/precio un 20% mejor que la media global.”

2. Zonas emergentes y oportunidades geográficas

- ¿Qué regiones menos conocidas tienen vinos muy bien puntuados?
- ¿Qué países tienen vinos top 10% en puntuación y poca representación comercial?

Insight esperado:

“Valle de Uco (Argentina) y Navarra (España) tienen vinos con 90+ puntos a mitad del precio de Burdeos.”

3. Análisis de variedad de uva

- ¿Qué variedades tienen más presencia global? ¿Qué puntuación media obtienen?
- ¿Qué variedades destacan según país?
- ¿Existen variedades infravaloradas con alto potencial?

Insight esperado:

“El Tempranillo fuera de España está obteniendo puntuaciones consistentes por debajo de los \$20.”

4. Benchmarking de bodegas

- ¿Cuáles son las bodegas con mejor puntuación promedio?
- ¿Qué bodegas producen vinos económicos con buenas puntuaciones?
- ¿Existen bodegas jóvenes pero prometedoras?

Insight esperado:

“Bodega X, con solo 4 vinos registrados, tiene una media de 92 puntos por menos de \$25.”

5. Perfil de sabor (NLP sobre descripciones)

- ¿Qué sabores o descriptores aparecen más en vinos bien puntuados?

- ¿Podemos predecir puntuación o precio a partir de la descripción?
- ¿Qué perfil de sabor domina en ciertas regiones o variedades?

💡 **Insight esperado:**

“Vinos descritos como ‘mineral’, ‘elegante’ y ‘estructura firme’ tienen una media de +3 puntos frente al resto.”

6. Análisis de catadores

- ¿Qué catadores son más estrictos o generosos?
- ¿Existe sesgo por país o variedad?

💡 **Insight esperado:**

“Kerin O’Keefe tiende a puntuar vinos italianos con un promedio 1.8 puntos superior que otros críticos.”

Proyecto 1: Sommelier Virtual – Recomendador Basado en Perfil de Sabor

Contexto

- **Cliente:** Usuarios amantes del vino sin conocimiento experto
- **Objetivo:** Recomendador interactivo que sugiera vinos según el perfil de sabor del usuario.
- **Motivación:** Facilitar la exploración personalizada entre miles de vinos, permitiendo que el usuario encuentre nuevas etiquetas que se alineen con sus preferencias sensoriales.

Proceso de Preparación de Datos

- **Dataset:** wine_final_dataset.csv
- **Limpieza:**
 - Exclusión de columnas no útiles: title, country, variety, points, price, wine_id
 - Conversión de variables de sabor a float

- Estandarización y verificación de integridad de valores

Motor de Recomendación (Cosine Similarity)

- **Modelo:** Similitud del perfil de sabores entre vinos y preferencias del usuario
- **Input del usuario:** sliders por descriptor de sabor
- **Output:** top-N vinos con mayor similitud sensorial

Interfaz (Streamlit)

- Panel lateral para configuración de filtros y preferencias sensoriales
- Visualización de recomendaciones con:
 - Imagen ilustrativa
 - Similitud, país, variedad, puntuación y precio
 - Perfil de sabor en gráfico horizontal

Impacto Esperado

- Mejora la experiencia del consumidor
- Democratiza el acceso al mundo del vino

Proyecto 2: Sommelier AI – Recomendador por Vino Similar

Contexto

- **Cliente:** Amantes del vino que buscan etiquetas similares a las que ya conocen y disfrutan
- **Objetivo:** Recomendador que encuentre vinos similares a uno de referencia seleccionado por el usuario

Preparación y Modelado

- **Dataset:** wine_final_dataset.csv
- **Modelo:** Nearest Neighbors usando distancia de coseno (k=11)
- **Input:** nombre del vino preferido
- **Proceso:**

- Obtención del índice del vino
- Búsqueda de k vecinos más similares
- Cálculo de similitud inversa a la distancia

Interfaz (Streamlit)

- Selector de vino
- Muestra metadatos, puntuación, variedad y país
- Comparación de perfil de sabores entre el vino elegido y las recomendaciones
- Filtros avanzados (país, similitud mínima)

Visualización

- Comparación de perfiles de sabor mediante gráfico de barras dobles
- Pestañas interactivas para explorar cada recomendación

Valor Añadido

- Identificación rápida de etiquetas similares
- Educación progresiva del paladar del usuario





Proyecto 3: Wine Value Analyzer – Identificación de Vinos de Alto Valor

Contexto

- **Cliente:** Importadora de vinos y consumidores orientados al valor
- **Objetivo:** Detectar vinos infravalorados (alta calidad a bajo precio) o sobrevalorados (precio alto sin respaldo de puntuación)

Preparación de Datos

- **Filtros dinámicos:** precio, puntuación, país, variedad
- **Variables utilizadas:** points, price, country, variety
- **Modelado:**
 - Regresión lineal para estimar precio esperado
 - Cálculo de desviación entre precio real y predicho

- Clasificación en:
 -  Hidden Gem
 -  Alto Valor
 -  Inflado
 -  Precio Justo

Interfaz (Streamlit)

- Paneles para seleccionar umbrales de categorización
- Filtros avanzados con paginación y búsqueda
- Tabla dinámica con formato condicional
- Resultados limitados para rendimiento

Visualización y Análisis

- Tabla comparativa con columnas destacadas: puntuación, precio real, precio estimado, desviación, categoría
- Estilo visual según categoría para rápida identificación

Aplicación Estratégica

- Ideal para tomar decisiones de importación
- Facilita al consumidor elegir vinos que maximicen su inversión