**1. Introducción**

El presente informe documenta el análisis realizado sobre un conjunto de datos que contiene información de vinos de distintas regiones del mundo. El objetivo es limpiar, transformar, analizar estadísticamente y visualizar la información para extraer conclusiones relevantes sobre la relación entre calidad, precio y procedencia.

**2. Fuentes de datos**

* **Fuente 1:** Dataset de vinos (fuente original: Kaggle Wine Reviews).
* **Fuente 2:** Información geográfica/regional complementaria unida al dataset principal.
* **Dataset final:** 280.832 filas × 10 columnas, tras limpieza y unión de las dos fuentes iniciales. Añadiendo 10 columnas más, dentro de la herramienta Power Bi.

**3. Transformación y limpieza de datos**

Pasos realizados:

1. **Normalización de nombres de columnas:** a minúsculas y sin espacios.
2. **Manejo de valores nulos:**
   * designation, region\_1, title → mantienen nulos para no eliminar información valiosa. Se imputarán como "Unknown" si es necesario para análisis categórico.
3. **Conversión de tipos:** variables categóricas (country, province, region\_1, variety, winery, designation, title) convertidas a tipo string.
4. **Tratamiento de outliers:** se identificaron precios extremos (hasta 3300 USD); para análisis estadístico se aplicó transformación log(price+1).
5. **Creación de variables derivadas:**
   * quality\_price = points / price.
   * price\_bin = categorías por rango de precio (<=10, 11-20, 21-50, 51-100, >100).
   * points\_bin = categorías por rango de puntuación (<=80, 81-85, 86-90, 91-95, 96-100).

**4. Análisis estadístico**

**4.1 Estadísticas descriptivas**

| **Variable** | **Media** | **Mediana** | **Mín** | **Máx** | **Desv.Est.** |
| --- | --- | --- | --- | --- | --- |
| Puntos | 88.15 | 88 | 80 | 100 | 3.15 |
| Precio | 33.44 | 25 | 4 | 3300 | 37.11 |

* Distribución de puntos concentrada entre 86y 90 puntos.
* Distribución de price altamente sesgada a la derecha (asimetría ≈ 15.96).

**4.2 Correlaciones**

* **Pearson:** 0.415 → relación moderada entre puntos y precio
* **Spearman:** 0.576 → relación monotónica más fuerte.
* Ambas con p-value ≈ 0 → significativas estadísticamente.

**4.3 Comparación vinos económicos vs. premium**

* Definición: Económico (≤20 USD), Premium (>100 USD).
* Test Mann–Whitney U: U=177314.0, p≈0.0 → diferencias significativas en la puntuación media.
* Conclusión: vinos más caros tienden a tener puntuaciones más altas.

**4.4 Países principales**

* **Top 3 por número de registros:** EE.UU., Francia, Italia.
* EE.UU. lidera en volumen y precio medio (~40 USD) con puntuación media de 88.6 puntos.

**5. Visualizaciones**

Se elaboraron en Power BI y Python:

* Histogramas de precio y puntuación.
* Transformación logarítmica de precios para reducir asimetría.
* Dispersión precio vs. puntuación (tendencia positiva).
* Dashboard interactivo en Power BI con filtros por país, variedad y rango de precio.

**6. Conclusiones y recomendaciones**

* Existe una correlación positiva entre el precio y la puntuación de los vinos, aunque no perfecta: hay vinos económicos con buena puntuación.
* El sesgo de precio obliga a aplicar transformaciones logarítmicas para análisis robustos.
* La procedencia y la variedad son factores que pueden influir fuertemente en la percepción de calidad y el precio.
* Recomendación para compradores: explorar vinos en el rango 20–50 USD que presenten puntuaciones ≥90, dado que ofrecen una buena relación calidad-precio.