Limpieza y análisis preliminar del dataset

Preparación inicial del archivo

El archivo descargado desde la FAO fue renombrado como “Dataset_Europa_crudo”, conservando su estado original. Para trabajar sobre él, se generó una copia llamada “Dataset_Europa”, sobre la que se aplicaron todas las modificaciones y limpiezas.

Eliminación de columnas irrelevantes

Se identificaron varias columnas que no aportaban valor al análisis y que solo incrementaban el tamaño del archivo. Estas se eliminaron para simplificar el dataset y mejorar su manejabilidad en Excel:

Columna eliminada	**Motivo**
Código del ámbito	Valor único ("QCL") en toda la columna.
Ámbito	Solo contiene "Cultivos y productos de ganadería".
Código del área (M49)	Redundante con la columna de nombre del país.
Código del elemento	Codifica valores ya representados en texto ("Área cosechada", "Rendimiento", etc.).
Código del producto (CPC)	Igual que el caso anterior, pero para productos.
Código del año	Repite la información del año en formato numérico.
Descripción del símbolo	Contiene metadatos sobre la fuente del dato. No aporta valor directo al análisis en Excel. No es un studio “real” por lo que los consideraremos fiables.
Notas	Muy pocas entradas; no añade información útil.

Estas acciones permitieron reducir el número de columnas de 15 a solo 6, y el tamaño del archivo bajó de 13,6 MB a menos de 4 MB, lo que representa una reducción cercana al 70 %.

> 💡 Nota: Cabe destacar que en la página de descargas de la FAO se pueden seleccionar muchos más datos de los que hemos elegido. Esto explica que algunas columnas tengan una sola entrada: si hubiéramos descargado más variables, esas columnas habrían contenido valores distintos.
>
> En cuanto a los códigos, tienen sentido en contextos como el aprendizaje automático, modelos predictivos o procesos automatizados. Pero en nuestro caso, centrado en la visualización y análisis en Excel, no aportan valor y solo añaden ruido al dataset.

Estandarización de valores

Se realizaron algunos ajustes menores para mejorar la estética y coherencia del dataset:

- “Aceitunas, olivas” → “Aceituna” (5.294 celdas modificadas).
- “Colza o semillas de colza” → “Colza” (5.294 cambios).
- “Países Bajos (Reino de los)” → “Países Bajos” (1.323 cambios).
- “Bélgica-Luxemburgo” → “BELUX” (1.323 cambios).
- La columna “Área” fue renombrada como “Países” para mayor claridad.

> 😉 Nota: Aunque estos cambios no afectan al tamaño, sí mejoran la visualización y el diseño de tablas y gráficos.

Exploración del contenido del dataset

Se revisaron las siete columnas restantes para validar su contenido y detectar posibles problemas:

- Países: no contiene celdas vacías.
- Elemento: contiene “Área cosechada”, “Rendimiento” y “Producción”. Sin vacíos.
- Producto: sin celdas vacías; incluye cultivos como maíz, arroz, triticale, cebada, etc.
- Año: cubre el periodo 1961–2023. Sin vacíos.
- Unidad: contiene “ha”, “kg/ha” y “toneladas”. Tiene 15.007 celdas vacías, esperables por la ausencia de datos en ciertos años o países.
- Valor: campo numérico con 15.692 celdas vacías. Aparecen muchos valores “0,0” especialmente a partir de 2018.
- Descripción: indica la fuente o calidad del dato. Será eliminada más adelante mediante Power Query.

Profundizando en las celdas vacias y ceros

Durante la revisión del dataset se identificaron múltiples celdas vacías y valores iguales a “0,0” en la columna **Valor**. Ambas situaciones pueden responder a distintos significados:

- Un **valor vacío** podría indicar que no se produjo ese cultivo, que el dato no está disponible o simplemente que no fue registrado.
- En cambio, un **cero explícito** sugiere que sí hay constancia de que no hubo producción en ese país, año y cultivo específico.

Al explorar los datos por año, se observó que antes de **2018** apenas se utilizaban ceros, mientras que desde ese año su frecuencia aumenta considerablemente. Esto podría deberse a un cambio en los criterios de registro por parte de la FAO o de las agencias nacionales responsables.

Aunque esta diferencia no se abordará desde un punto de vista técnico profundo —por tratarse de un proyecto centrado en el uso de Excel— conviene tenerla presente al interpretar los resultados. La ausencia de datos no siempre equivale a una ausencia de producción.

> 💭 Nota: El manejo adecuado de datos faltantes —o valores que podrían resultar inconsistentes o extremos— constituye una de las partes más importantes en la preparación de cualquier dataset. En este caso, la buena calidad del archivo original simplifica considerablemente la labor de depuración y permite centrarse en su análisis con mayor confianza.

Recuento total

El archivo contiene 37.045 registros, una cifra que roza el límite óptimo para trabajar cómodamente en Excel sin perder fluidez. Su riqueza no está solo en el volumen, sino en la profundidad: cubre más de 60 años de datos agrícolas europeos, lo que permite observar tendencias claras y construir visualizaciones con verdadero valor analítico.

> 😬 Nota: Realmente Excel puede tratar dataset mucho mas grandes, pero por comodidad, practicidad, limpieza y tamaño (el trabajo hay que enviarlo), debemos considerar “no pasarnos” con el tamaño

Trabajar con Europa añade complejidad por la diversidad de países y cultivos, pero también aporta coherencia geopolítica al proyecto. Aunque existen otros actores relevantes en el sector —como Marruecos, Argentina…—, el enfoque europeo responde al contexto del estudio y permite delimitar bien el universo de análisis.

🔧 Con esta base sólida, entraremos ahora en materia con la limpieza técnica del dataset.