

Selección y Justificación del Conjunto de Datos para el Proyecto de Visualización

Julio Úbeda Quesada



Imagen creada con <https://openart.ai/home>.

1. Justificación de la selección del conjunto de datos (10%)

He seleccionado el conjunto de datos obtenido mediante técnicas de *web scrapping* desde el [Registro General de la Flota Pesquera de la Unión Europea](#), y el [Registro Oficial de Buques Pesqueros del Ministerio de Agricultura, Pesca y Alimentación](#), porque constituye un proyecto personal en el que ya he trabajado en profundidad y cuya estructura, calidad y alcance conozco detalladamente. La elección también se fundamenta en motivos profesionales, ya que los datos relacionados con el sector marítimo-pesquero son de gran interés para análisis técnicos, históricos y de gestión, especialmente en el contexto de políticas pesqueras europeas (PPC), modernización de flotas y sostenibilidad del esfuerzo pesquero.

Además, al haber construido el dataset mediante *web scraping* ([Repositorio 1.](#)) y haber pasado por un proceso completo de **limpieza, imputación, transformación y modelado supervisado/no supervisado** ([Repositorio 2](#)), este proyecto permite continuar con una línea de trabajo coherente, mejorada y perfectamente documentada en las fases previas.

2. Relevancia del conjunto de datos en su contexto (10%)

El dataset es **altamente relevante** en su ámbito porque:

- Contiene **información oficial y pública** procedente del *Registro General de la Flota Pesquera de la Unión Europea* y del *Ministerio de Agricultura, Pesca y Alimentación de España*.
- Abarca un periodo muy amplio (**1987–2025**), lo que permite estudiar **casi 40 años de evolución de la flota pesquera española**, un sector clave para muchas comunidades costeras.
- Responde a necesidades reales de análisis: sostenibilidad, control del esfuerzo pesquero, envejecimiento de la flota, modernización tecnológica, etc.
- Permite analizar fenómenos que afectan a colectivos concretos, como **pescadores artesanales**, sectores industriales, o regiones enteras dependientes del sector.

3. Complejidad del conjunto de datos (25%)

Considero que el dataset tiene una complejidad muy adecuada para una práctica de visualización avanzada:

- **Número de registros:** 27.364 buques únicos (supera ampliamente el criterio de “miles de filas”).
- **Número de variables:** más de una veintena de columnas iniciales, y tras la transformación estructural se mantienen **trece de variables** relevantes.
- **Variedad de tipos de datos:**
 - **Categoricos:** material del casco, tipo de arte, estado administrativo...
 - **Cuantitativos continuos:** eslora, potencia, arqueo GT.

- **Cuantitativos discretos:** edad del buque.
- **Fechas:** fecha de alta y fecha de estado.
- **Texto:** nombre del buque.
- **Identificadores únicos:** CFR.
- **Variables geográficas:** puerto, provincia, comunidad autónoma.
- **Transformaciones adicionales:** separación de columnas, imputación de ceros ilógicos mediante KNN, estandarización, categorización del arte, normalización logarítmica de variables, etc.
- **Riqueza analítica:** el dataset permite trabajar con valores extremos, distribuciones sesgadas, correlaciones fuertes entre variables, diferentes escalas, e incluye información temporal y geográfica.

4. Originalidad del conjunto de datos (25%)

Este dataset cumple con creces los criterios de originalidad:

- **No es un dataset clásico** ni ampliamente explotado en prácticas formativas. A diferencia de temas típicos como COVID-19, ciberdelincuencia, accidentes o tráfico, este tema tiene poca presencia en ejercicios académicos.
- **Ha sido construido desde cero** a partir de *web scraping* de dos registros oficiales, lo cual incrementa su valor y singularidad.
- **Ya han pasado un proceso de depuración y transformación** mediante:
 - depuración profunda,
 - normalización de categorías,
 - creación de nuevas variables (edad del buque, comunidad, provincia, código postal),
 - separación estructural de columnas compuestas,
 - imputación avanzada,
 - estandarización y transformación logarítmica,
 - segmentaciones mediante *clustering*,
 - modelos predictivos para el tipo de arte.
- Existen estudios agregados de la UE (como [Joint Research Centre \(JRC\)](#) y [Annual Economic Report on the EU Fishing Fleet](#)), pero ninguno a este nivel de detalle y centrado en un análisis tan técnico a nivel español.

5. Cuestiones a responder mediante visualización y diccionario de variables (30%)

Propongo responder a las siguientes preguntas que considero relevantes, adecuadas al dataset, originales y conectadas con los puntos anteriores. Entre ellas destacan:

- **Evolución temporal de la flota:** ¿Cómo ha variado el número de buques, su eslora, potencia y arqueado desde 1987 hasta 2025?

- **Envejecimiento y renovación:** ¿Qué patrones presenta la edad de los buques? ¿Existen ciclos de renovación?
- **Distribución geográfica:** ¿Qué puertos o comunidades concentraron más buques? ¿Se observan diferencias regionales?
- **Modalidades de pesca:** ¿Qué tipos de arte son más comunes? ¿Existen diferencias por tamaño, material, potencia o zona geográfica?
- **Relaciones estadísticas:**
 - ¿Existe una relación entre la comunidad autónoma y el tipo de arte más frecuente?
 - ¿Existe una relación entre el tamaño del buque y el tipo de arte más frecuente?
- **Clustering:** ¿Qué tipologías de buques existen según sus características técnicas?

6. Contenido del dataset

El conjunto de datos incluye:

Variable	Tipo	Descripción	Rol
cfr	Categorico	Identificador único del buque	Identificador
nombre	Texto	Nombre del buque	Descriptiva
fc_alta_rgfp	Fecha	Fecha de alta en el registro	Hecho
fc_estado	Fecha	Fecha del estado administrativo	Hecho
estado_rgfp	Categorico	Activo, baja definitiva, etc.	Dimensión
eslora_total	Numérica continua	Longitud del buque (m)	Medida
arqueo_gt	Numérica continua	Capacidad (GT)	Medida
potencia_kw	Numérica continua	Potencia del motor (kW)	Medida
material_casco	Categorica	Madera, acero, poliéster, etc.	Dimensión
puerto	Categorica	Puerto de registro	Dimensión geográfica
provincia	Categorica	Provincia de registro	Dimensión geográfica
comunidad_autónoma	Categorica	Comunidad autónoma	Dimensión geográfica
tipo_arte	Categorica	Modalidad de pesca	Variable objetivo / dimensión
edad_buque	Numérica	Años de antigüedad	Medida derivada

7. Conclusión

El conjunto de datos seleccionado es:

- **Pertinente, complejo y rico en variables**, apto para visualizaciones avanzadas.
- **Actual, oficial y socialmente relevante**.
- **Original**, ya que ha sido creado, transformado y enriquecido desde cero.
- **Adecuado para responder preguntas analíticas sólidas**, fundamentadas estadísticamente y visualmente.