“Proyecto final: Limpieza, normalización y análisis de la pesca mexicana”

Bases de Datos

Trabajo elaborado por:

Eduardo García Cortez 207480

Rafael Harry Gomar Dawson 208999

Diego Alejandro Méndez Uzcátegui 208400

Maestro: José Antonio Lechuga Rivera

15/5/2024

Índice

Introducción al conjunto de datos y problema a estudiar

Este trabajo consiste en un análisis de exploración y normalización de un set de datos de pesca obtenido de la CONAPESCA, órgano desconcentrado de la SAGARPA.[1] Para el trabajo se seleccionó utilizar los datos de los años 2020, 2021 y 2022.

Nuestro objetivo es estudiar la fauna acuática capturada por la industria de la pesca mexicana tanto en su modalidad de acuacultura y captura. De esa forma analizar qué especies son las más capturadas, cuales valen más dinero, en que se especializada cada localidad, en qué se especializa cada entidad, cuáles son las que más producen, entre otras cosas.

Cabe recalcar que estos datos de la pesca no toman en cuenta la captura ilegal, pues todos los registros se llevan a cabo en oficinas gubernamentales donde obviamente no van a registrar los que lo hacen ilegalmente, pero es una parte de los datos que creemos falta para tener toda la imagen de la pesca mexicana.

Para replicar el proyecto y ver el código de SQL de cada sección dirigirse al repositorio en GitHub.[2]

[1] *Producción Pesquera - datos.gob.mx/busca.* (n.a.). Gob.mx. Recuperado mayo 15, 2024, de https://datos.gob.mx/busca/dataset/produccion-pesquera

[2] https://github.com/EduardoGC21/BasesDeDatosPesca/tree/main

Carga inicial y análisis preliminar

En el transcurso del proyecto se usaron tres archivos en formato CSV, correspondientes a los años 2020, 2021 y 2022. Cada archivo encapsula datos referentes a un mes de actividades pesqueras realizadas en distintas regiones del país. La estructura de la información se presenta en un formato tabular que incluye las siguientes columnas: EJERCICIO FISCAL, ENTIDAD FEDERATIVA, NOMBRE OFICINA, MES DE CORTE, ORIGEN, NOMBRE COMÚN, PESO DESEMBARCADO (en kilogramos), PESO VIVO (en kilogramos), VALOR (en pesos), NOMBRE PRINCIPAL, CLAVE DE ENTIDAD y CLAVE DE OFICINA. El volumen total de datos acumulados asciende a 636,747 entradas, distribuidas de la siguiente manera: 59,591 correspondientes al año 2022, 513,922 al 2021 y 63,237 al 2020.

Durante la revisión inicial de los datos, se identificó que una pequeña parte de uno de los archivos CSV presentaba corrupción. Específicamente, 37 registros tenían un carácter incorrecto de separación (¬ en lugar de coma).

EJERCICIO FISCAL	ENTIDAD FEDERATIVA	NOMBRE OFICINA	MES DE CORTE	ORIGEN	NOMBRE COMUN	PESO DESEMBARCADO KILOGRAMOS	PESO VIVO KILOGRAMOS	VALOR PESOS	NOMBRE PRINCIPAL	CLAVE DE ENTIDAD	CLAVE OFICINA	NUMERO DE ATUNES
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬622	622	24,880	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬535	535	21,400	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬489	489	19,560	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬399	399	15,960	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬340	340	13,600	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬290	290	11,600	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬264	264	10,560	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬235	235	9,400	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬205	205	8,200	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬181	181	7,240	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬130	130	5,200	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬115	115	5,980	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬99	99	3,960	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	ENERO	ACUACULTURA	MEJILL¬20	20	800	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	FEBRERO	ACUACULTURA	MEJILL¬776	776	31,040	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	FEBRERO	ACUACULTURA	MEJILL¬599	599	23,960	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	FEBRERO	ACUACULTURA	MEJILL¬459	459	18,360	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	FEBRERO	ACUACULTURA	MEJILL¬367	367	14,680	OTRAS	2	203		
2021	BAJA CALIFORNIA	ENSENADA	FEBRERO	ACUACULTURA	MEJILL¬335	335	13,400	OTRAS	2	203		

Para solucionar este problema, se reemplazó el carácter erróneo por la coma correcta. Posteriormente, se utilizó SQL para llevar a cabo las tareas de limpieza adicionales, asegurando que la integridad de los datos estuviera preservada para las fases siguientes del análisis.

Limpieza de datos

En la etapa inicial de manipulación de datos con SQL, se creó una tabla que incorpora doce atributos, todos del tipo texto, para albergar la totalidad de las entradas recopiladas.

Posteriormente, se inició el análisis de los datos, durante el cual se identificaron diversas necesidades de limpieza y optimización. En el proceso de limpieza general del texto, se implementaron expresiones regulares para eliminar espacios múltiples en secuencia dentro de varias columnas, reemplazándolos por un único espacio. Asimismo, se procedió a eliminar paréntesis y el contenido dentro de los mismos en ciertas columnas, aplicando esta limpieza únicamente donde era necesario.

En lo que concierne a la limpieza de los nombres de las ciudades, se corrigieron errores ortográficos y se reemplazaron formas abreviadas por nombres completos mediante la utilización de estructuras condicionales CASE en SQL. Además, se estandarizó el uso de 'Estado de México' en lugar de 'México' en una columna específica para reflejar de manera más precisa la entidad federativa correspondiente.

Se detectó un error de información en el que se asignaron incorrectamente códigos de oficina que no correspondían al estado adecuado. Por ejemplo, el código 3010 de la ciudad “Tuxpan”, apareció en registros para el estado de Nayarit, a pesar de que el prefijo “30” indica que debería pertenecer al estado de Veracruz. Este error probablemente se originó debido a la existencia de ciudades homónimas en diferentes estados. Al ingresar los datos, no se mantuvo una distinción clara entre los códigos de las ciudades con nombres idénticos, dependiendo del estado en el que se ubicaban, lo que condujo a inconsistencias en la base de datos.

En relación con las subespecies y especies, se efectuaron correcciones en los nombres, eliminando caracteres especiales erróneos y reintroduciendo letras ñ faltantes para corregir los nombres correctamente. Este proceso incluyó la eliminación de duplicados de frases y la corrección de errores tipográficos en nombres específicos.

Finalmente, se llevó a cabo la limpieza y conversión de números. Se eliminaron comas de los valores numéricos para facilitar su conversión y manejo, tratando además casos especiales donde los datos numéricos estaban mal formateados o eran inexistentes. Para asegurar la consistencia, se eliminaron espacios al final de los valores en todas las columnas. A su vez, se transformaron varias columnas de texto a tipo numérico y se ajustó la columna de año a formato de fecha.

Cada uno de estos pasos contribuye significativamente a que la base de datos no solo esté más limpia, sino también más organizada y fácil de analizar, asegurando así la precisión y consistencia de los datos a la hora de la normalización.

Normalización hasta cuarta forma normal

Tabla Original
La tabla inicial denominada pesca contiene los siguientes atributos:
Año
Entidad Federativa
Nombre de Localidad
Mes de Corte
Origen
Nombre Común
Peso Desembarcado (kg)
Peso Vivo (kg)
Valor (Pesos)
Nombre Principal
Clave de Entidad
Clave de Oficina

Paso 1: Primera Forma Normal (1FN)
La Primera Forma Normal (1FN) requiere que los valores en cada columna de una tabla sean atómicos y que la tabla tenga una llave primaria única. La tabla original ya cumple con estos requisitos, dado que no presenta grupos repetitivos ni columnas con múltiples valores.

Se estableció la siguiente llave primaria: {Año, Mes de Corte, Clave de Oficina, Nombre Principal, Nombre Común}.

Paso 2: Segunda Forma Normal (2FN)
La Segunda Forma Normal (2FN) se enfoca en eliminar la redundancia de datos causada por dependencias funcionales parciales. Una dependencia funcional parcial ocurre en una base de datos relacional cuando el valor de un atributo o conjunto de atributos depende únicamente de una parte (y no de la totalidad) de una clave primaria compuesta. Entonces aseguramos que cada atributo no clave dependa completamente de la llave primaria compuesta.

Para alcanzar la 2FN, se identificaron y extrajeron las siguientes dependencias:

Dependencias sobre entidades geográficas:
{Clave de Oficina} → {Nombre de Oficina, Clave de Entidad}
{Clave de Entidad} → {Entidad Federativa}

Con una descomposición simple, se crearon tablas separadas para entidad y localidad, cada una con su propia llave primaria, eliminando así las dependencias parciales y posibles anomalías de inserción y borrado.

Paso 3: Tercera Forma Normal (3FN)
La Tercera Forma Normal (3FN) requiere la eliminación de dependencias transitivas (cuando un atributo no clave depende de otro atributo no clave que, a su vez, depende de la clave primaria de la tabla) entre los atributos. En una evaluación inicial, aplicamos una descomposición a la

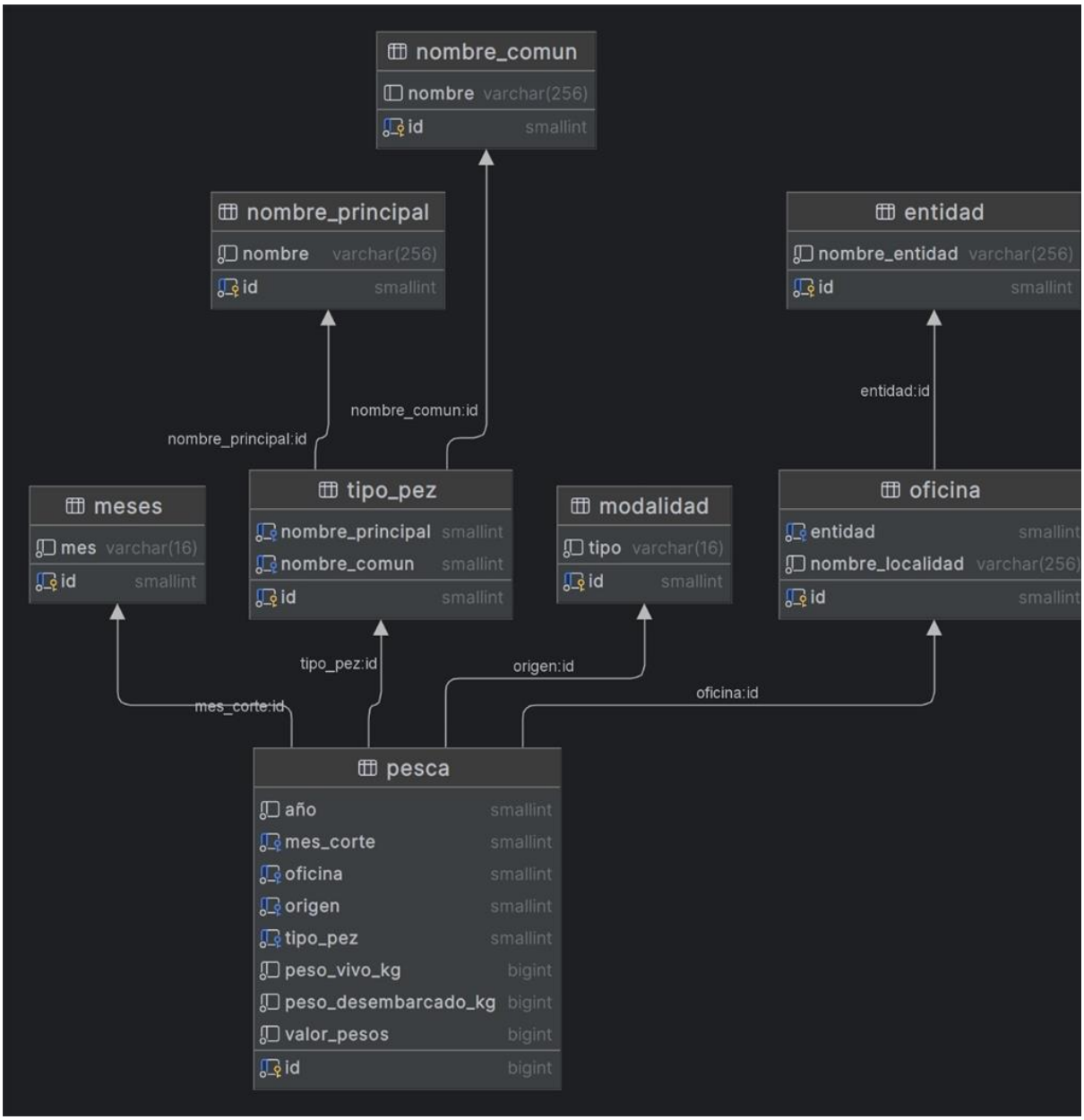
dependencia entre Nombre Común y Nombre Principal, considerando una relación tipo subespecie-especie. Sin embargo, se observó duplicación de datos en la inserción a la base de datos "normalizada". Un análisis más detallado reveló casos donde un Nombre Común podría asociarse con varios Nombres Principales y viceversa, invalidando nuestra suposición inicial de dependencia transitiva. En consecuencia, se determinó que la base de datos ya estaba en 3FN, ya que no existen dependencias transitivas entre atributos no clave.

Paso 4: Cuarta Forma Normal (4FN)
La Cuarta Forma Normal (4FN) se alcanza cuando no existen dependencias multivaluadas no triviales que no sean una dependencia funcional. En este caso, se consideró la relación MVD:

{Nombre Principal} ↠ {Nombre Común}.

Se observó que, para un mismo Nombre Principal, existen múltiples Nombres Comunes asociados independientemente de otros factores en la base de datos. Esta relación se representó en una tabla separada denominada tipo_pez, almacenando todas las posibles combinaciones de nombres que podrían utilizarse.

Tablas Finales y Catálogos
Para optimizar la eficiencia y reducir la redundancia de texto, se crearon catálogos para Mes y Origen, así como para Nombre Principal y Nombre Común, facilitando la implementación y mejora de la velocidad en búsquedas y mantenimiento de la base de datos.

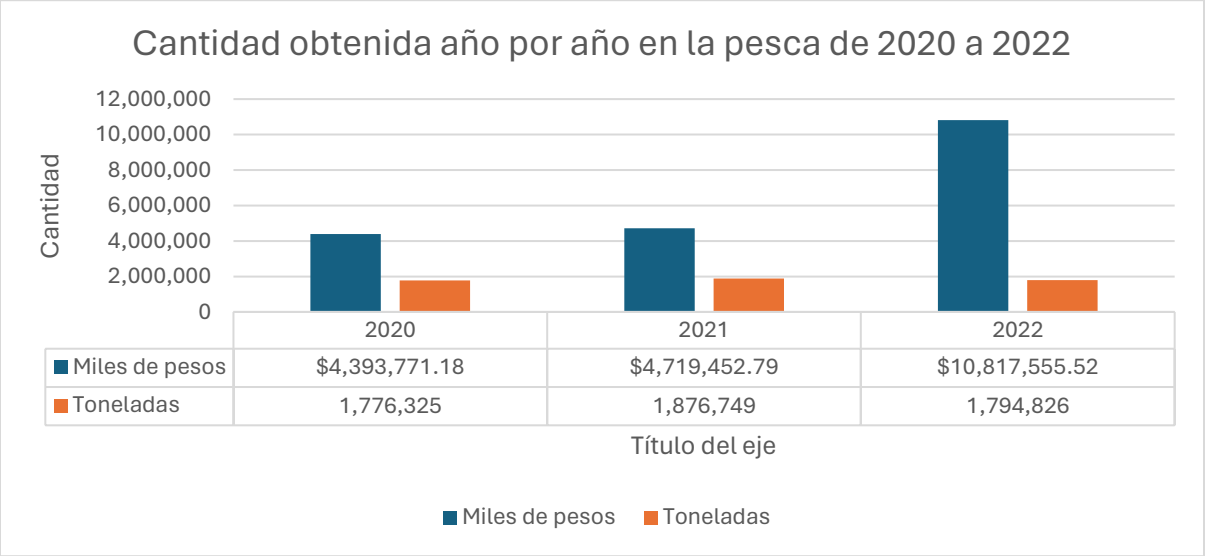


Esta estructura minimiza la redundancia de datos y optimiza la integridad del sistema de gestión de bases de datos al abordar eficazmente las anomalías que podrían surgir en la versión no normalizada. Además, permite una escalabilidad más controlada y una gestión eficiente de los datos a medida que crecen en volumen y complejidad. También se crearon índices para todas las FK.

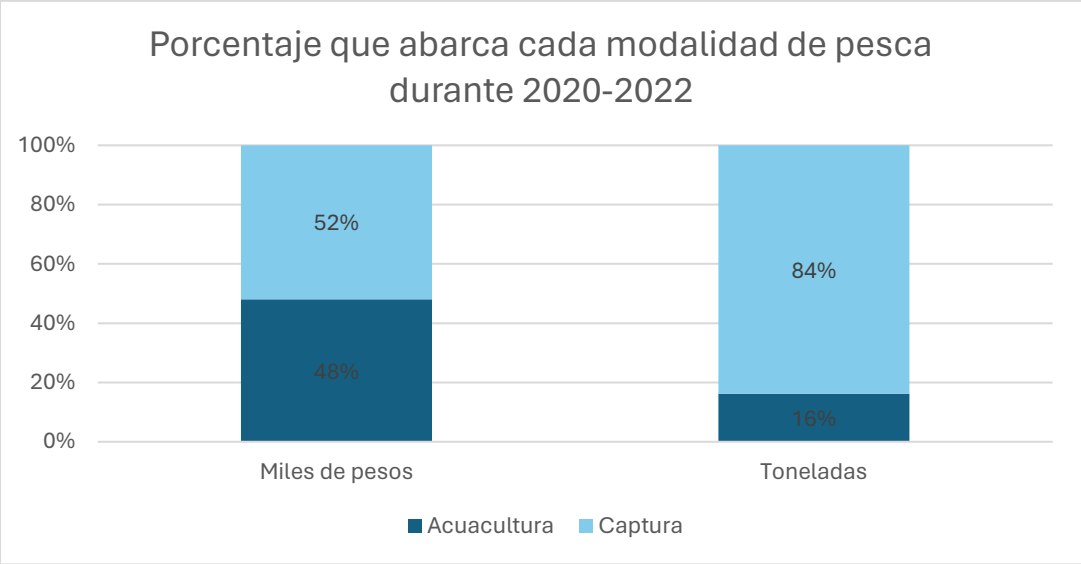
Análisis de datos a través de consultas SQL

Como se normalizó y categorizó todo, esto dificultaba el análisis gráfico, por lo cual se optó por crear una vista con todo unido para obtener los nombres y no los FK (números) en específico se creó una vista "todo_unido" y otras específicas de fauna acuática como "camarón", "atun", "sardina" y "pulpo". Se realizaron distintos querys, pero cabe aclarar que no todos se graficaron, a continuación, se visualizan los graficados, donde el número corresponde al número del query en SQL:

1. Primero se hizo un análisis general de la pesca en México, y podemos observar que está creciendo la industria de la pesca.



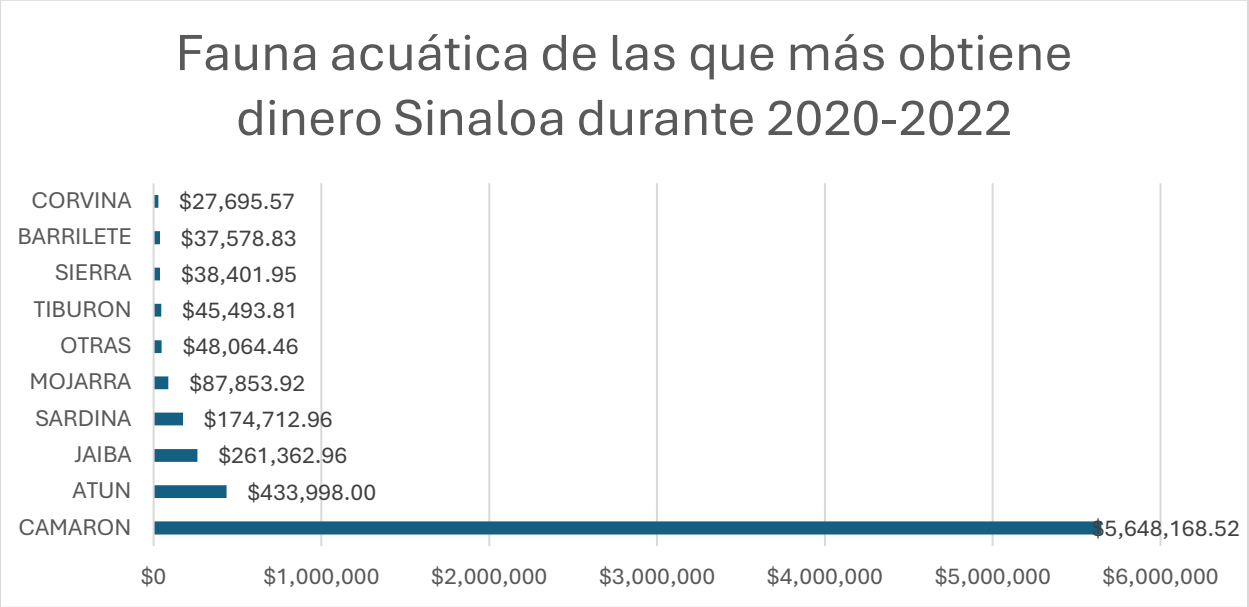
2. ¿Cómo se distribuye la pesca en la captura y acuacultura? aunque se obtiene muy pocas toneladas con acuacultura, rinde mucho más monetariamente, ya que se pueden especializar en cultivar cualquier pez que quieran, esto es, los más caros.



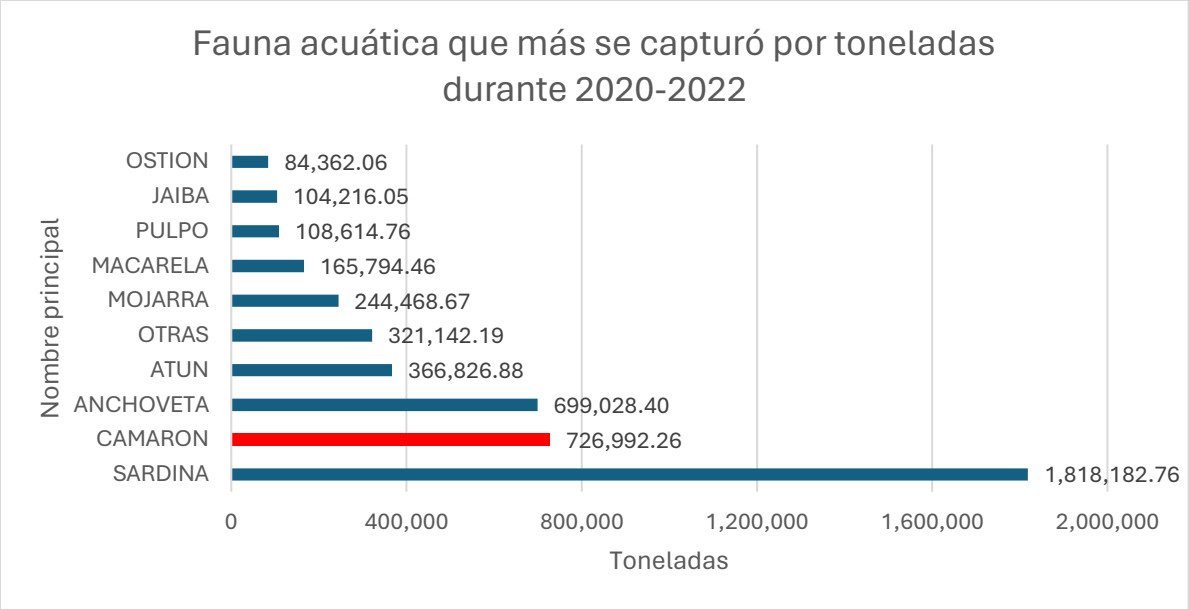
3. ¿Qué entidades producían más valor monetario en toda la república? Sinaloa fue la primera seguida por Sonora.

4. ¿Cuáles son las oficinas (ciudades) que más producen valor monetario? 5 de ellas son de Sinaloa, esto explica por qué produce tanto.

5. ¿En qué se especializa Sinaloa para obtener tanto valor monetario? El camarón es la clave para entender el papel de Sinaloa.

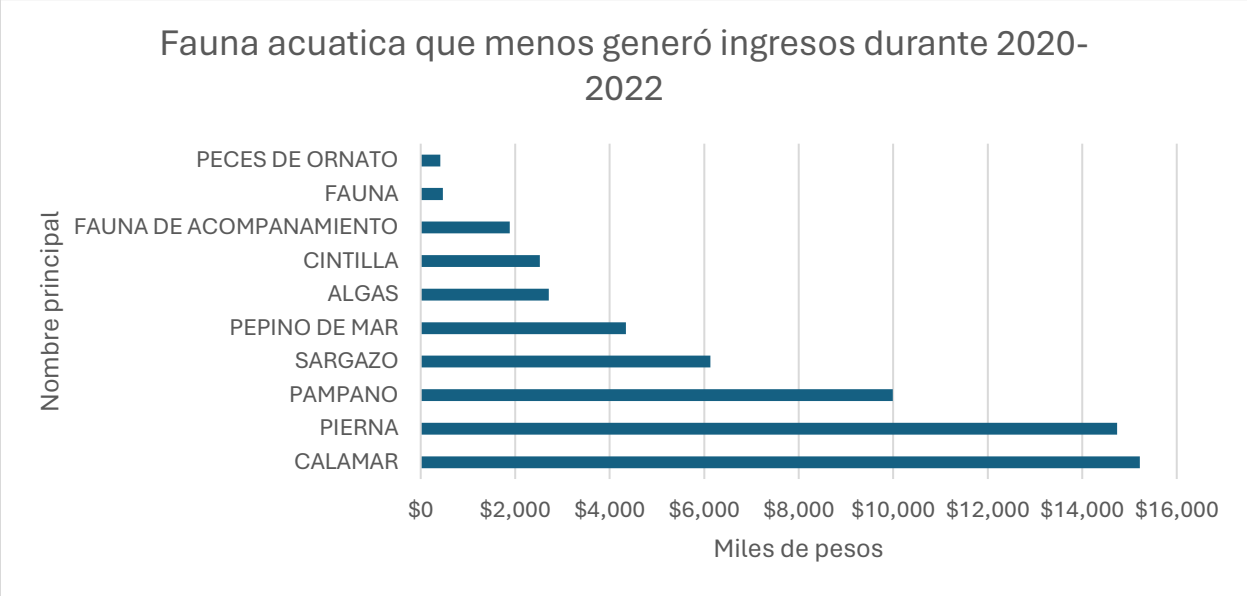


6. Ahora, ¿qué fauna acuática bajo nombre principal son las que más capturan?

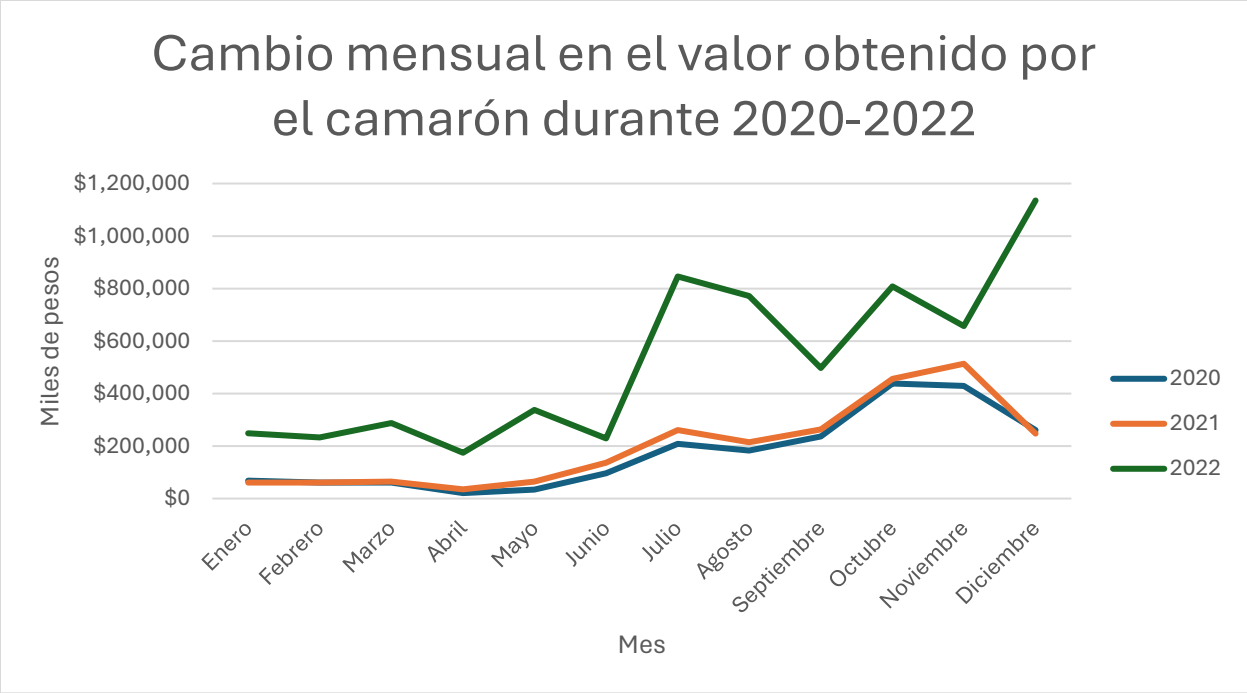


7. A pesar de que la Sardina es la que más se saca por toneladas, no da tanto dinero como el camarón.

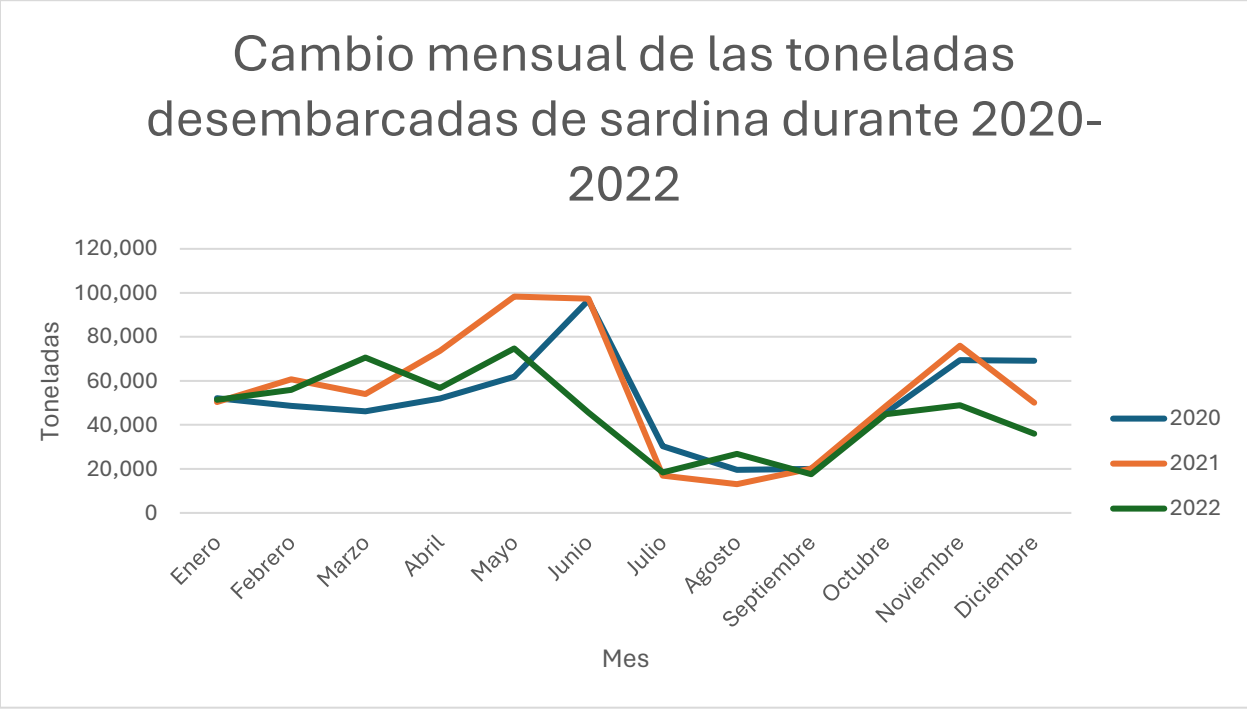
8. Y, ¿cuáles son los nombres principales que menos generan valor monetario?



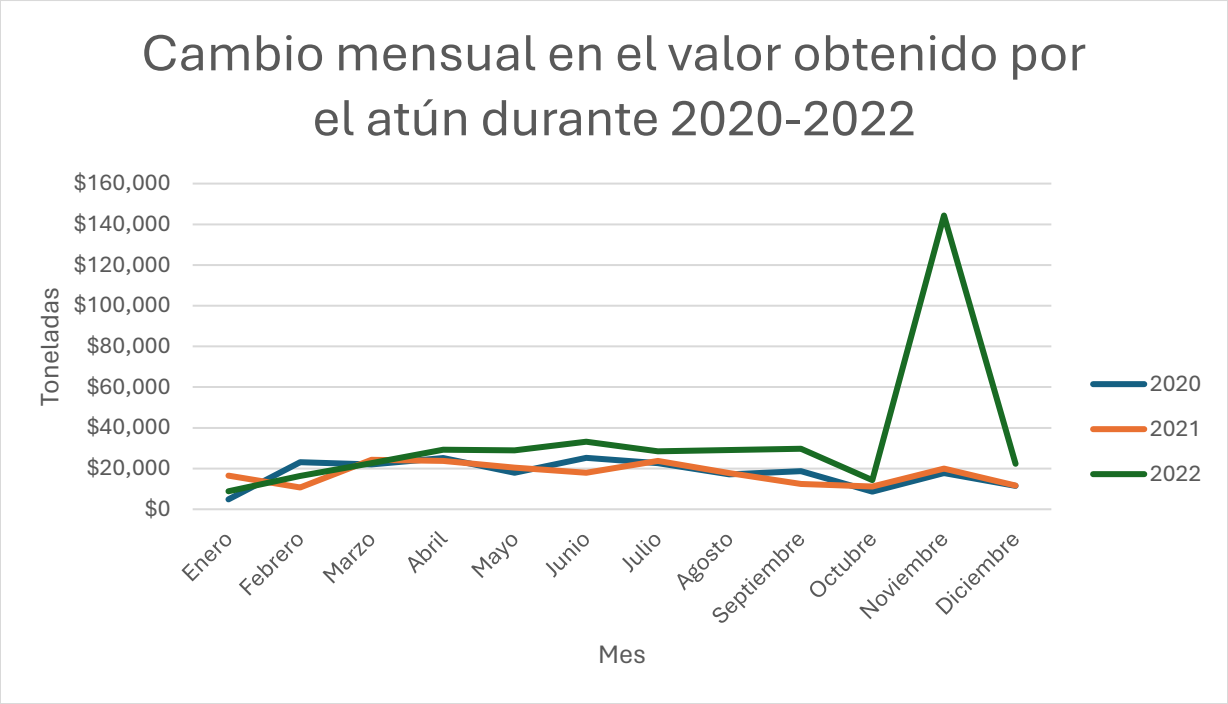
9. Nos interesó 4 especies en específico, camarón, atún, sardina y pulpo. A continuación se muestra el valor monetario de estos, mes por mes, para ver los picos donde valen más y como año por año cambia su valor monetario:



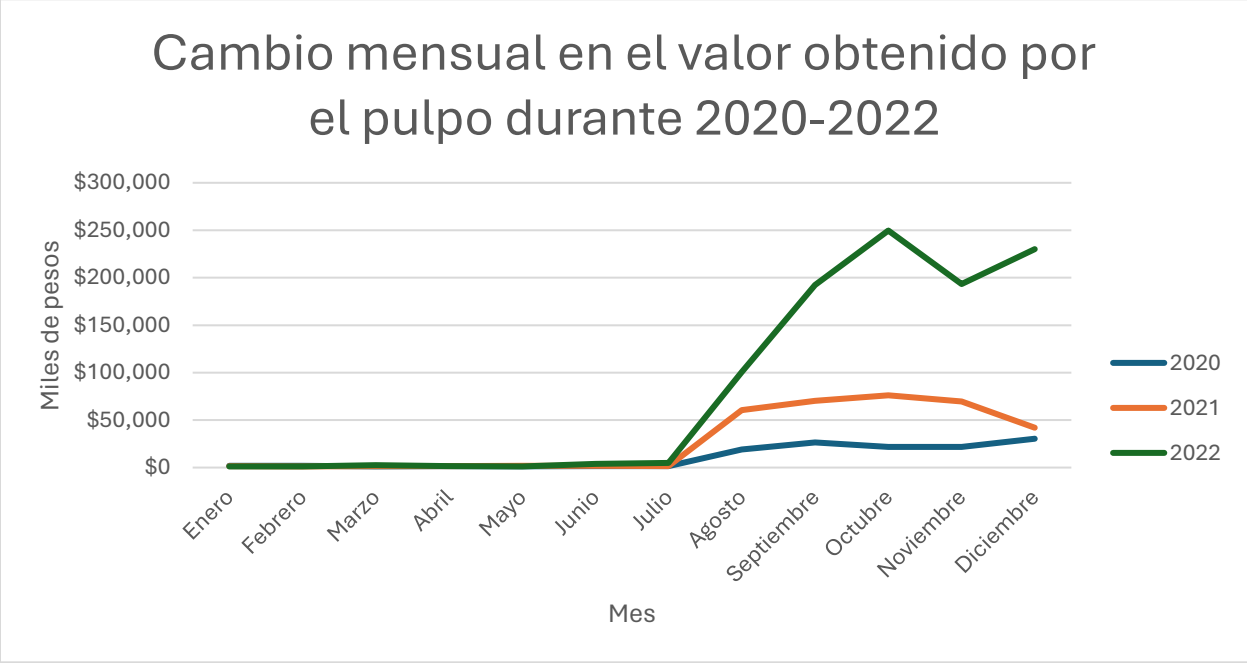
10. En la sardina se graficó su cantidad desembarcada en toneladas.



11. El atún tuvo un gran salto en 2022



12. En el pulpo ha estado creciendo mucho, año con año, su valor monetario.

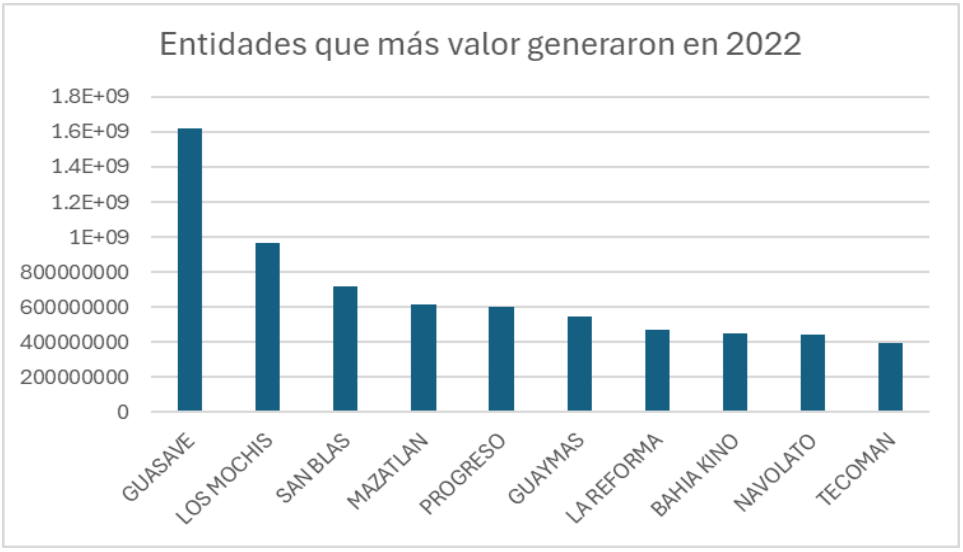


Creación de atributos para entrenamiento de modelos

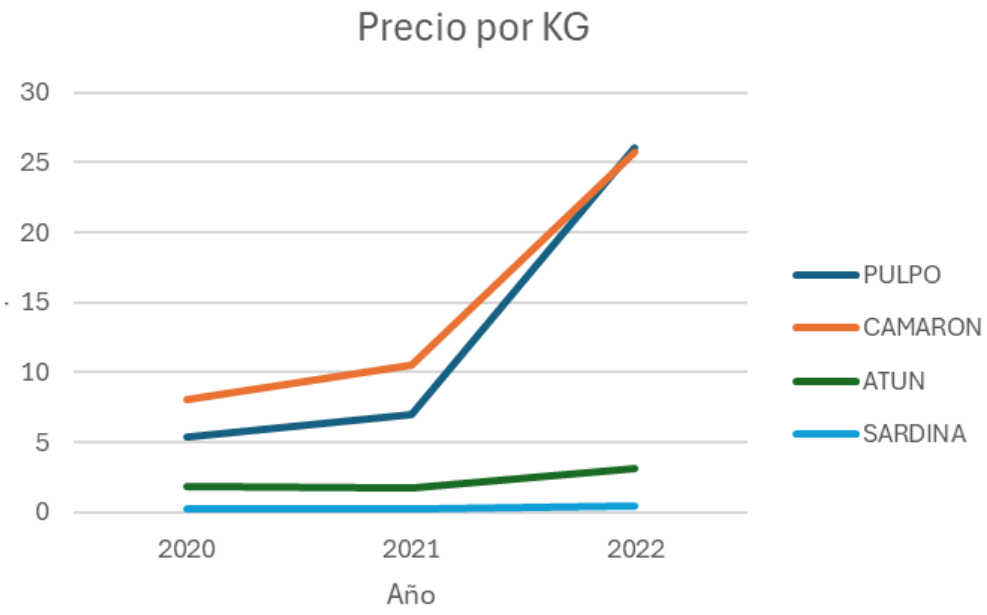
En esta sección se utilizaron Querys que implementaran funciones de ventana para sí poder conservar la integridad de los datos en las tuplas y al mismo tiempo poder visualizar la información requerida.

1. Nos interesó saber que ciudades de las mencionadas en los datos habían desembarcado la mayor cantidad de kilos de pesca en 2022.

2. También nos interesó saber que ciudades habían generado más valor por su pesca en el mismo año para poder compararlas y de ahí analizar los números. No eran las mismas ciudades que más KG pescaban, por lo que concluimos que estas pescaban especies de mayor valor.



3. Quisimos saber, para las mimas especies de animales acuáticos cual era su valor por kilo y como este había cambiado durante los años, por lo tanto, utilizamos una función de ventana para eso.



4. Después utilizamos este mismo parámetro para poder saber que pesces eran los que más valor generaban por año. El único movimiento de valor brusco fue el del pulpo en 2022.

2020

Fauna Acuática	Precio por KG	Posición
ABULON	52.29157965	1
LANGOSTA	32.92741605	2
LOBINA	14.32041507	3
LANGOSTINO	8.889187803	4
ROBALO	8.778987125	5

2021

Fauna Acuática	Precio por KG	Posición
ABULON	42.97081644	1
LANGOSTA	30.55676561	2
LOBINA	20.06002426	3
RUBIO	10.78818682	4
ROBALO	9.701367595	5

2022

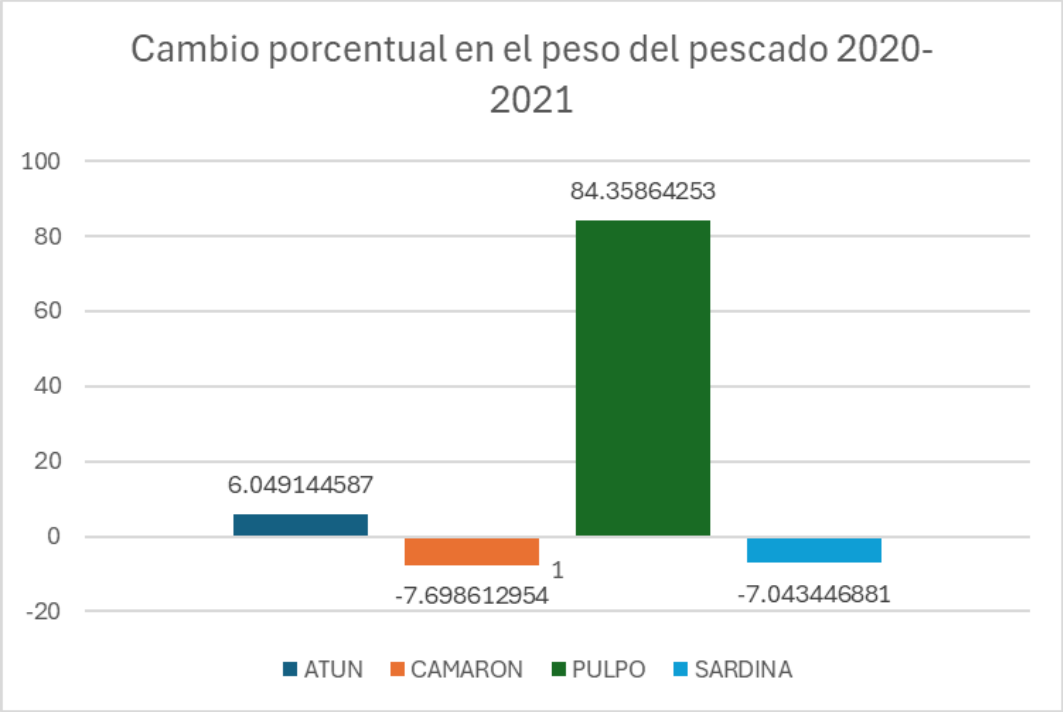
Fauna Acuática	Precio por KG	Posición
LANGOSTA	39.07340924	1
ABULON	54.2242953	2
PULPO	26.07059302	3
CAMARON	25.7553321	4
LOBINA	19.9052195	5

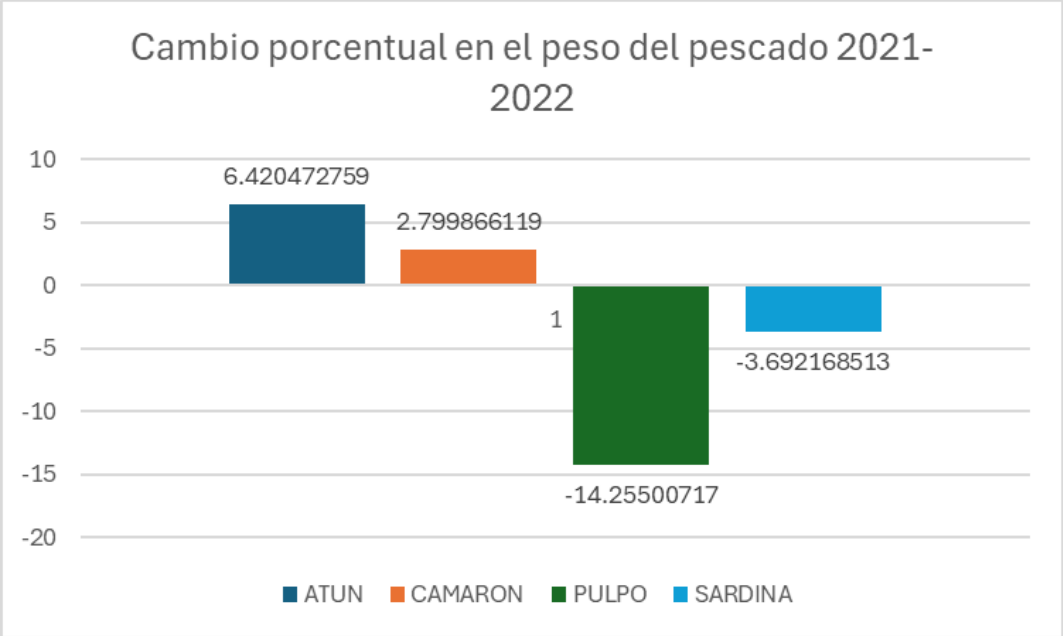
5. De igual manera nos interesó conocer cuáles eran las tres especies que menos valor generaban por kilo pescado, y encontramos que para los tres años en el registro eran las mismas tres especies.

2022

Fauna Acuática	**Precio por KG**	**Posición**
ANCHOVETA	0.449446497	55
SARDINA	0.462437416	54
MACARELA	0.629349255	53

6. Por último, nos llamó la atención ver el cambio anual entre los pesos totales pescados en KG de las cuatro especies que habíamos analizado antes: pulpo, camarón, atún y sardina.





Conclusiones

En conclusión, este set de datos nos dio la posibilidad de aprender de la pesca mexicana de forma muy específica. Fue un proyecto muy interesante pues aprendimos a normalizar, a limpiar un set de datos y a hacernos preguntas valiosas que nos ayudarían a entender este mercado y como sacarle más provecho. Los resultados obtenidos se podrían usar incluso para la sostenibilidad de la pesca.

Para futura investigación, se podría implementar un programa de machine learning que, dado ciertos parámetros, como peso, mes de captura, año, entidad y oficina, pudiera predecir que tipo de pez es.