DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico Final

Calidad de datos en la información de producción de pozos de gas y petróleo

4 de diciembre de 2024 Calidad de Datos

Grupo : 12

Integrante	LU	Correo electrónico
Navarro, Solana	906/22	solanan3@gmail.com
Suarez, Ines	890/22	ine.suarez22@gmail.com
Wittmund Montero, Lourdes	1103/22	lourdesmonterochiara@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
http://www.exactas.uba.ar

Índice

1. Introducción

La Secretaría de Energía pone a disposición conjuntos de datos detallados sobre la producción de pozos, tanto convencionales como no convencionales, permitiendo el análisis y diagnóstico de su calidad.

Este trabajo práctico tiene como objetivos principales aplicar los conocimientos adquiridos en la materia para evaluar la calidad de dichos datos y desarrollar mecanismos que permitan identificar y gestionar posibles desvíos en la información. A través de un enfoque integral, se buscará garantizar que los datos sean consistentes, únicos y relevantes para los análisis necesarios.

En este contexto, el trabajo no solo contribuirá a mejorar la interpretación de los datos actuales, sino que también servirá de base para establecer buenas prácticas en el manejo de grandes volúmenes de información en sectores estratégicos.

Las bases de datos que vamos a utilizar se pueden encontrar en la pagina Produccion de petroleo y gas, de alli nos interesan las siguientes tablas de informacion:

- Capitulo IV - Pozos: Cuenta con la informacion y caracteristicas de todos los pozos del pais, incluyendo ubicacion geografica, formacion, profundidad, tipo de pozo, entre otros. Funciona como la fuente maestra en la cual deberiamos tener todos los pozos detallados en las siguientes bases de datos.
- Produccion de Pozos de Gas y Petroleo No Convencional: Cuenta con la informacion registrada por mes y por año de la produccion de los pozos con tipo de recurso No Convencional.
- Produccion de Pozos de Gas y Petroleo - 2024: Cuenta con la informacion registrada por mes (desde enero hasta octubre) del año 2024 de la produccion de los pozos en su mayoria con tipo de recurso Convencional.

2. Análisis

De aca en adelante nos vamos a referir al dataset Capitulo IV - Pozos como **Pozos** y al de Produccion de Pozos de Gas y Petroleo No Convencional como **No Convencionales**.El otro dataset, el de Produccion de Pozos de Gas y Petroleo - 2024, lo llamaremos **Convencionales** dado que para realizar todo el trabajo decidimos en esa tabla quedarnos solo con los registros con tipo de recurso Convencional, para asi tener una separacion entre convencionales y no convencionales que nos serviria para realizar el analisis adecuado y pedido. Desarrollaremos mas sobre esto en la seccion de decisiones tomadas.

2.1. Descripción

Para este análisis, comenzaremos realizando una descripción del conjunto de datos.

- **Pozos**

Columna	Informacion
sigla	Una abreviatura para identificar el pozo
idpozo	ID del pozo
area	Nombre del area en la que se encuentra el pozo
cod_area	Codigo que representa el area
empresa	Nombre de la empresa que opera en el pozo
yacimiento	Nombre del yacimiento en el que de encuentra el pozo
cod_yacimiento	Codigo del yacimiento
formacion	Formacion geologica donde se perfora el pozo
cuenca	Cuenca en la que se ubica el pozo
provincia	Provincia en la que se ubica el pozo
cota	Altitud del pozo en la superficie con respecto al nivel del mar (en metros)
profundidad	Profundidad del pozo (en metros)
clasificacion	Clasificacion del pozo (ej explorado, explotado)
subclasificacion	Sub-clasificacion del pozo
tipo_recurso	Tipo de recurso extraido (ej convencional, no convencional)
sub_tipo_recurso	Sub-tipo de recurso
gasplus	Indica si el pozo es parte del programa Gas Plus
tipopozo	Tipo del pozo (ej petrolifero, gasifero, inyeccion)
tipoextraccion	Metodo de extraccion
tipoestado	Estado actual del pozo
adjiv_fecha_inicio_perf	Fecha de inicio de la perforacion
adjiv_fecha_fin_perf	Fecha de fin de perforacion
adjiv_fecha_inicio_term	Fecha de inicio de terminacion
adjiv_fecha_fin_term	Fecha de fin de terminacion
geojson	Coordenadas geograficas del pozo
geom	Datos geometricos para la ubicacion del pozo

Tabla 1: Descripcion del dataset Capitulo IV - Pozos

Este conjunto de datos contiene información detallada sobre pozos petroleros y gasíferos, con un enfoque integral en diversos aspectos clave de su perforación y explotación, cubre todos los pozos sin separación temporal, registrando cada pozo solo una vez.

En cuanto a la información de identificación y ubicación, cada pozo tiene identificadores únicos, como sigla e idpozo, que permiten su distinción y seguimiento. Además, se incluye la ubicación geográfica precisa, representada mediante formatos como geojson y geom, lo que facilita la visualización y análisis espacial. También se registran detalles sobre su localización administrativa, incluyendo la provincia, el área y el yacimiento en el que se encuentra.

En relación con las características técnicas de los pozos, la base de datos proporciona datos sobre la profundidad de perforación, la cota (altitud del pozo), el tipo de extracción utilizado (por ejemplo, convencional o no convencional), y el estado operativo del pozo, lo que es esencial para evaluar la viabilidad y la actividad del pozo en cuestión.

La información temporal en esta base se limita a las fechas de perforación y terminación de cada pozo, lo que permite seguir la cronología del desarrollo y la evolución de cada pozo sin la división mensual o anual presente en los siguientes conjuntos de datos.

Desde el punto de vista de los aspectos operativos, la base incluye información sobre el tipo de recurso extraído (petróleo, gas, agua) y el método de extracción utilizado, lo que proporciona una visión clara de las operaciones realizadas. También se detalla el estado operativo actual de cada pozo, lo que indica si está en producción, en reserva o en algún otro estado específico.

Finalmente, en cuanto a la información empresarial, se identifica a las empresas operadoras responsables de cada pozo, lo que facilita el seguimiento de las actividades y la rendición de cuentas a nivel corporativo.

- **No Convencionales**

Columna	Informacion
idempresa	ID de la empresa que opera el pozo
anio	Año en el que se registraron los datos
mes	Mes en el que se registraron los datos
idpozo	ID del pozo
prod_pet	Produccion de petroleo
prod_gas	Produccion de gas
prod_agua	Produccion de agua
iny_agua	Cantidad de agua inyectada en el pozo
iny_gas	Cantidad de gas inyectado en el pozo
iny_otro	Cantidad de otros fluidos inyectados en el pozo
tef	Tiempo efectivo de funcionamiento del pozo (en dias u horas)
vida_util	Vida util estimada del pozo (en dias)
tipoextraccion	Tipo de extraccion utilizada
tipoestado	Estado operativo del pozo (ej extraccion efectiva, inactivo)
tipopozo	Clasificacion del pozo
observaciones	Notas o comentarios adicionales sobre el pozo
fechaingreso	Fecha en la que se ingresaron los datos al sistema
rectificado	Indicador de si los datos han sido rectificados
habilitado	Indicador de si el pozo esta habilitado
idusuario	ID del usuario que ingreso los datos
empresa	Nombre de la empresa operadora
sigla	Sigla asociada al pozo
formprod	Formacion productiva del pozo
profundidad	Profundidad del pozo (en metros)
formacion	Formacion geologica del pozo
idareapermisoconcesion	ID del area de permiso o concesion
areapermisoconcesion	Nombre del area de permiso o concesion
idareayacimiento	ID del area del yacimiento
areayacimiento	Nombre del area del yacimiento
cuenca	Cuenca geologica donde se encuentra el pozo
provincia	Provincia donde se encuentra el pozo
coordenadax	Coordenada X (longitud) del pozo
coordenaday	Coordenada Y (latitud) del pozo
tipo_de_recurso	Tipo de recurso explotado
proyecto	Proyecto asociado al pozo
clasificacion	Clasificación del pozo (ej explotación, exploración)
subclasificacion	Subclasificación del pozo
sub_tipo_recurso	Subtipo del recurso explotado
fecha_data	Fecha de los datos registrados

Tabla 2: Descripcion del dataset Produccion de Pozos de Gas y Petroleo No Convencional

Lo que se puede ver en base a esta descripcion es que la base de datos proporciona una visión integral que abarca tanto aspectos operativos como geográficos relacionados con la producción y explotación de algunos recursos. Esta base contiene información detallada sobre la producción de petróleo, gas y agua, lo que permite realizar un análisis preciso del rendimiento de los pozos y su contribución a la producción general. Además, incluye datos sobre los procesos de inyección, como el uso de agua, gas, CO2 y otros fluidos, esenciales para los procesos de recuperación mejorada de hidrocarburos y el manejo de la presión en los yacimientos.

La base también registra información técnica relevante de los pozos, como su profundidad de perforación, el tipo de extracción utilizado, y la vida útil estimada de cada pozo, factores fundamentales para evaluar la eficiencia operativa y el ciclo de vida de los pozos. En cuanto a la ubicación geográfica, se detalla la provincia, la cuenca y el yacimiento donde se encuentran los pozos, lo que facilita el análisis geológico y estratégico sobre la distribución de los recursos.

En el ámbito administrativo, la base de datos incluye información sobre los permisos y concesiones otorgados a las empresas operadoras, lo que asegura el cumplimiento normativo y proporciona un marco legal para la explotación de los pozos.

La base de datos tiene una organización temporal que abarca año, mes y fecha, lo que facilita el seguimiento histórico de la producción y la inyección. Además, se incluye información sobre el estado operativo de los pozos, indicando si están activos, en reserva o en estudio, lo cual permite tener una visión clara sobre su disponibilidad para la producción.

Tambien la base contiene datos tanto de pozos petrolíferos como de pozos gasíferos, con detalles específicos sobre las características de cada tipo de recurso, lo que permite un análisis diferenciado. Un sistema de trazabilidad con identificadores únicos para empresas, pozos, áreas y usuarios asegura la integridad y el seguimiento de los datos, garantizando un acceso eficiente y el control de la información a lo largo del tiempo.

- **Convencionales**

Columna	Informacion
idempresa	ID de la empresa que opera el pozo
anio	Año en el que se registraron los datos (2024)
mes	Mes en el que se registraron los datos
idpozo	ID del pozo
prod_pet	Produccion de petroleo en el pozo
prod_gas	Produccion de gas en el pozo
prod_agua	Produccion de agua en el pozo
iny_agua	Cantidad de agua inyectada en el pozo
iny_gas	Cantidad de gas inyectado en el pozo
iny_co2	Cantidad de CO2 inyectado en el pozo
iny_otro	Cantidad de otros fluidos inyectados en el pozo
tef	Tiempo efectivo de funcionamiento del pozo
vida_util	Vida util estimada del pozo
tipoextraccion	Tipo de extraccion utilizada en el pozo
tipoestado	Estado actual del pozo
tipopozo	Clasificacion del pozo segun su tipo
observaciones	Notas u observaciones adicionales sobre el pozo
fechaingreso	Fecha en la que se ingresaron los datos al sistema
rectificado	Indicador de si los datos han sido rectificados
habilitado	Indicador de si el pozo esta habilitado
idusuario	ID del usuario que ingreso los datos
empresa	Nombre de la empresa operadora
sigla	Sigla asociada al pozo
formprod	Formacion productiva del pozo
profundidad	Profundidad del pozo (en metros)
formacion	Formacion geologica del pozo
idareapermisoconcesion	ID del area de permiso o concesion
areapermisoconcesion	Nombre del area de permiso o concesion
idareayacimiento	ID del area del yacimiento
areayacimiento	Nombre del area del yacimiento
cuenca	Cuenca geologica donde se encuentra el pozo
provincia	Provincia donde se encuentra el pozo
tipo_de_recurso	Tipo de recurso extraido
proyecto	Proyecto asociado al pozo
clasificacion	Clasificación del pozo segun su estado o uso
subclasificacion	Subclasificación del pozo
sub_tipo_recurso	Subtipo del recurso extraido
fecha_data	Fecha de los datos registrados

Tabla 3: Descripcion del dataset Produccion de Pozos de Gas y Petroleo - 2024

Por ultimo esta base nos proporciona practicamente la misma informacion que la anterior pero solo para registros del año 2024 e incluyendo pozos de todo tipo de recurso (en un principio), mientras que en la anterior solo eran los pozos no convencionales.

A continuacion vamos a usar la informacion proporcionada para realizar analisis de calidad y luego de produccion de petroleo y gas para pozos no convencionales y convencionales.

2.2. Unicidad de los datos

Para empezar a tener una nocion de que tan alta es la calidad de estas fuentes vamos a empezar realizando una evaluacion de la unicidad de los datos en cada una de ellas.

Uno de los aspectos fundamentales a revisar en estas tablas es la unicidad de la columna idpozo, ya que representa un identificador único para cada pozo. Si encontráramos un ID duplicado, esto podría significar que un pozo está registrado más de una vez o que dos pozos distintos comparten el mismo identificador. Ambas situaciones generarían errores en el análisis, ya sea en la evaluación del rendimiento o en el manejo correcto de los recursos asociados.

Para el dataset de Pozos, el análisis se limita a verificar la unicidad de la columna idpozo, dado que cada registro representa un pozo único. Sin embargo, en los otros datasets, el análisis es más complejo debido a que contienen múltiples registros para un mismo pozo correspondientes a distintos momentos en el tiempo. Si solo verificáramos la columna idpozo, sería lógico encontrar IDs repetidos.

En el caso del dataset de Convencionales, analizamos los registros en función de combinaciones únicas de idpozo y mes. Esto nos permitió identificar cualquier duplicado que representara un problema real de unicidad en este contexto temporal.

Por otro lado, para el dataset de No Convencionales, el análisis incluyó las columnas idpozo, año y mes, ya que en este caso es posible tener varios registros para el mismo pozo y mes pero correspondientes a distintos años. Esto representa una diferencia clave con respecto al dataset de 2024, donde todos los registros pertenecen al mismo año y, por tanto, este problema no se presenta.

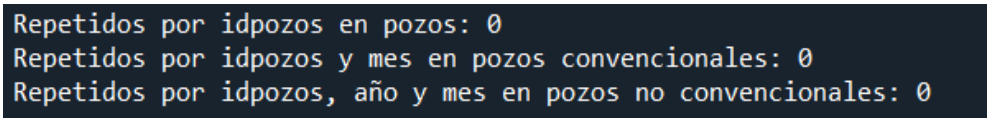


Figura 1: Cantidad de idpozo repetidos

Como se puede ver en la imagen, no hay ningún caso de repetición de ID, lo que es crucial para garantizar la integridad y precisión de los datos. La ausencia de IDs duplicados permite un seguimiento claro y único de cada pozo a lo largo de su ciclo de vida, lo cual es fundamental para realizar un análisis exhaustivo de su rendimiento, estado operativo y producción.

En los datasets que incluyen múltiples registros para un mismo pozo en distintos momentos (como en los casos de idpozo combinado con mes en el dataset de Convencionales, o con mes y año en el de No Convencionales), también se verificó que no existieran repeticiones indebidas de estas combinaciones clave. Esto asegura que no haya inconsistencias temporales en los datos, permitiendo un análisis preciso en contextos específicos.

Además, contar con identificadores exclusivos, ya sea a nivel de pozo o en combinación con información temporal, facilita la gestión de recursos, evita errores de asignación y mejora la eficiencia operativa en la toma de decisiones. Esta unicidad contribuye a que los informes sean más precisos y confiables, eliminando cualquier posibilidad de contar dos veces la misma información y, por lo tanto, asegurando una correcta evaluación de resultados.

Otro aspecto importante en términos de unicidad es la columna de coordenadas geograficas. Estas estan presentes en los datasets de Pozos y de No Convencionales por lo tanto vamos a estudiar que esta pasando en sus respectivas columnas. En principio, uno podría imaginar que no deberían existir dos pozos distintos en una misma ubicación.

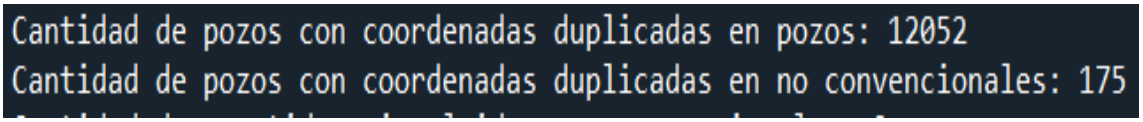


Figura 2: Cantidad de coordenadas repetidas

Sin embargo, como se observa en la imagen, efectivamente hay coordenadas repetidas, lo que contradice nuestra suposición inicial y el análisis previo, ya que habíamos concluido que no existían pozos duplicados. En consecuencia, no deberían existir coordenadas duplicadas tampoco. Ante este hallazgo, decidimos investigar más a fondo qué se entiende por "pozo" en este contexto. Tras indagar, llegamos a la conclusión de que, geográficamente, un mismo pozo puede contener diferentes formaciones, y en cada una de ellas se puede estar realizando una extracción distinta, lo que se considera como un pozo diferente con su propio ID. Aunque entendemos que esto puede ser interpretado de distintas maneras, decidimos adoptar este criterio a partir de ahora para identificar los diferentes pozos. Con esta nueva perspectiva, nos adentramos más en los datos para analizar su unicidad, ya que ahora contamos con un criterio distinto.

Un siguiente análisis se centró en identificar si existían registros duplicados en los datasets, considerando aquellos casos donde toda la información era idéntica en todas las columnas excepto en la columna de ID. Este tipo de duplicación podría ser indicativa de problemas de calidad, ya que reflejaría que un mismo pozo fue registrado más de una vez con diferentes identificadores.

idempresa	anio	mes	idpozo	prod_pet
PPSA	2024	8	10229	0
PPSA	2024	8	10230	0

(a) Convencionales

sigla	idpozo	area	cod_area	empresa
YPF.SC.ACBo.e-6	73652	CAMPO BOLEADORAS	CABO	COMPAÑÍA GENERAL DE COMBUSTIBLES S.A.
YPF.SC.ACBo.e-6	92265	CAMPO BOLEADORAS	CABO	COMPAÑÍA GENERAL DE COMBUSTIBLES S.A.

(b) Pozos

Figura 3: Registros con todo igual menos ID

Durante esta revision, detectamos varios casos de duplicados en los datasets de Pozos y de Convencionales, lo que sugiere que, por alguna razón, estos pozos fueron ingresados más de una vez con IDs diferentes. Esta situación podría deberse a errores en la etapa de carga de datos o a inconsistencias en los sistemas de registro originales. En la imagen se muestra un ejemplo de dos registros con información idéntica en todas sus columnas, excepto en la columna de ID. Para facilitar la visualización, no se incluyeron todas las columnas en la figura, pero la igualdad de datos se mantiene en todas ellas.

Por otro lado, al realizar la misma verificación en el dataset de No Convencionales, no encontramos registros que cumplieran con este patrón. Esto indica un nivel superior de consistencia en ese conjunto de datos, lo cual es un indicio positivo respecto a su calidad.

Es importante destacar que la presencia de duplicados puede afectar el análisis posterior, especialmente si estos registros no se gestionan adecuadamente. La decisión de eliminarlos o ajustarlos depende del contexto del proyecto y de cómo se planea utilizar la información en los análisis siguientes. Por ello, este aspecto deberá ser considerado cuidadosamente antes de proceder con el tratamiento de los datos. En la seccion de decisiones tomadas nos adentramos en esto.

2.3. Consistencia de los datos

Realizamos un análisis exhaustivo de los datasets Pozos, Convencionales y No Convencionales para evaluar la calidad y consistencia de los datos, identificando valores faltantes y detectando discrepancias entre registros. Este estudio nos permitió obtener un panorama sobre la integridad de la información disponible.

En primer lugar, evaluamos la cantidad de valores faltantes o nulos en cada dataset.

Convencionales:	
vida_util	799749
tipoextraccion	36
tipoestado	36
tipopozo	36
observaciones	759798
formprod	26800
formacion	27930
cuenca	30
clasificacion	168960
subclasificacion	168960
sub_tipo_recurso	778707

(a) Convencionales

No Convencionales:	
vida_util	317612
tipoextraccion	575
tipoestado	575
tipopozo	575
observaciones	306528
clasificacion	838
subclasificacion	838
sub_tipo_recurso	368

(b) No Convencionales

Pozos:	
empresa	897
formacion	2815
adjiv_fecha_inicio_perf	34139
adjiv_fecha_fin_perf	34284
adjiv_fecha_inicio_term	36594
adjiv_fecha_fin_term	36593

(c) Pozos

Figura 4: Cantidad de nulls por columnas de cada dataset

En el caso de los **pozos**, las columnas más afectadas fueron las relacionadas con las fechas de perforación y terminación, donde se encontraron hasta 36,594 valores faltantes. En el dataset de **no convencionales**, las variables como **vida útil**, con 317,612 valores faltantes, y **observaciones**, con 306,528, también presentaron un volumen considerable de datos ausentes. Finalmente, en el dataset de **convencionales**, las mayores ausencias se concentraron en variables como **vida útil** (799,749 valores faltantes) y **observaciones** (759,798). Este diagnóstico inicial permitió priorizar las columnas clave que requerían mayor atención para el procesamiento de los datos, las columnas mas afectadas por los valores nulos no eran tan relevantes para el analisis de la produccion por lo tanto no nos obstruyeron en nuestro trabajo, pero se podria tomar algun criterio general para tratar de evitarlos y poner otro tipo de valores en las columnas en las que no se cuenta con informacion disponible para agregar en lugar de no poner nada y que quede un null ya que son valores que bajan mucho la calidad de un dataset.

Posteriormente, realizamos una validación de las fechas relacionadas con las etapas de perforación y terminación de los pozos, asegurando que las fechas de inicio no fueran posteriores a las de finalización, y verificamos los casos en los que ambas fechas eran nulas. Como resultado, solo encontramos tres casos en los que las fechas eran nulas o contenían valores inválidos, lo que contribuyó a mejorar la fiabilidad de los datos.

En cuanto a las variables numéricas, establecimos rangos plausibles para evitar valores extremos que no se correspondieran con las características físicas reales de los pozos. Para la **cota**, limitamos los valores hasta 5,600 metros como maximo, mientras que para la **profundidad**, los restringimos entre 0 y 27,100 metros. Estos filtros permitieron eliminar registros anómalos y ajustar los datos a escenarios realistas. A continuación, mostramos un boxplot de las variables cota y profundidad, que nos permitió visualizar la distribución de los datos y detectar posibles valores extremos.

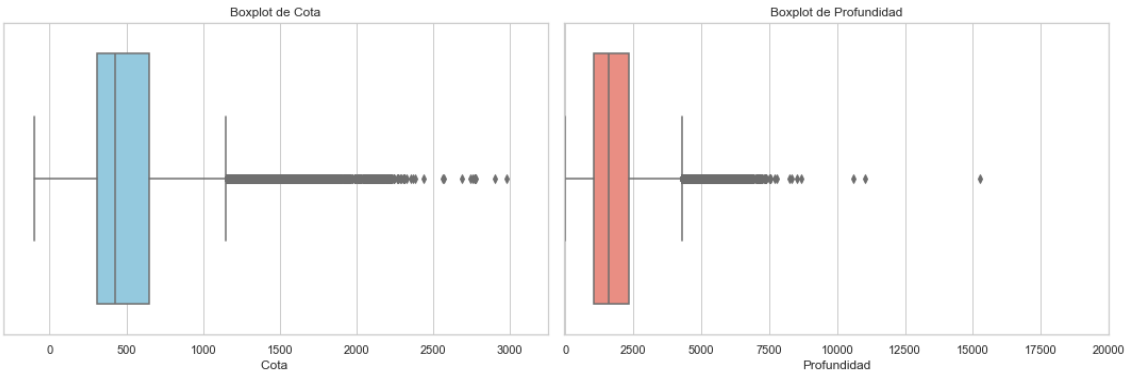


Figura 5: Distribucion de las variables cota y profundidad

Para analizar la distribución de estas variables, utilizamos boxplots que evidenciaron patrones consistentes con las condiciones esperadas. La mayoría de las cotas se concentraron entre 0 y 3,000 metros, lo cual refleja una distribución lógica dentro del terreno. Por otro lado, las profundidades de los pozos se encontraron principalmente por debajo de los 20,000 metros, aunque algunos valores cercanos al límite superior podrían representar pozos excepcionalmente profundos.

Después de completar el análisis de consistencia dentro de cada tabla, avanzamos al estudio de la consistencia entre ellas. Todas las tablas contienen datos sobre los mismos pozos, por lo que se espera que exista una correlación coherente entre la información que presentan.

En un primer paso, verificamos si algún pozo estaba presente simultáneamente en las tablas de pozos convencionales y no convencionales. Encontrar coincidencias de este tipo sería un error crítico, ya que un pozo no puede pertenecer a ambas categorías al mismo tiempo debido a que representan clasificaciones mutuamente excluyentes. Al poder detectar la superposición nula en estas tablas concluimos que este aspecto de consistencia se cumplía, permitiéndonos avanzar con confianza hacia las siguientes comparaciones.

El análisis continuó con la comparación entre las tablas de Pozos y Convencionales. En este caso, sí surgieron problemas de calidad. Para proceder de manera estructurada, seleccionamos únicamente el último registro disponible en la tabla de Convencionales para cada pozo, asumiendo que la tabla de Pozos reflejaba los datos más actualizados y, por tanto, representaba mejor la situación actual de los pozos. Después de consolidar ambas tablas, evaluamos cuántas filas mostraban datos inconsistentes y en qué columnas se encontraban esas discrepancias.

```
Cantidad de inconsistencias inicial entre Pozos y Convencionales: 20452
inconsistencia_empresa              23
inconsistencia_yacimiento            4
inconsistencia_cod_yacimiento        9
inconsistencia_cod_area              1
inconsistencia_area                  1
inconsistencia_tipo_recurso      20316
inconsistencia_tipopozo              7
inconsistencia_tipoextraccion       25
inconsistencia_tipoestado          240
```

Figura 6: Inconsistencias iniciales entre Pozos y Convencionales

Entre las inconsistencias detectadas, destacamos que la columna con mayor cantidad de valores conflictivos fue la correspondiente al tipo de recurso, con un número significativo (20,316) de filas afectadas. Esto refleja un problema de consistencia que requería atención inmediata. Una vez aplicado el tratamiento adecuado para resolver estos problemas, detallado más adelante en la sección de decisiones tomadas, se logró reducir drásticamente la cantidad de datos inconsistentes hasta que representaron una fracción pequeña del total. Esto permitió que los datos fueran suficientemente confiables para cumplir con los objetivos del análisis.

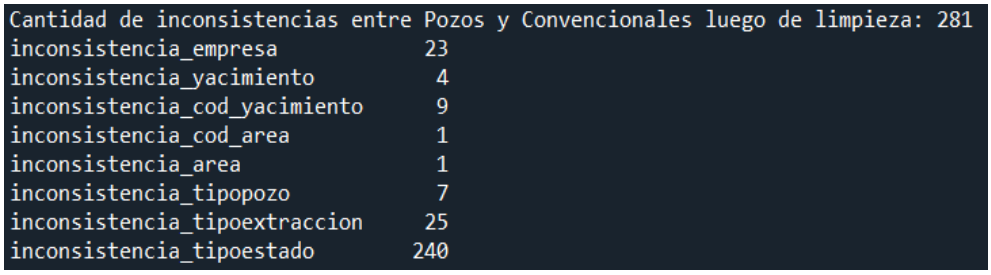


Figura 7: Inconsistencias entre Pozos y Convencionales luego de la limpieza

Finalmente, repetimos el procedimiento anterior para las tablas de Pozos y No Convencionales. Seleccionamos el último registro de cada pozo en la tabla de No Convencionales y analizamos las inconsistencias. A diferencia del caso anterior, no se detectaron discrepancias en la columna de tipo de registro, lo cual es un indicador positivo de coherencia en este aspecto. Sin embargo, la mayor cantidad de inconsistencias se encontró en la columna correspondiente a codigo yacimiento, con un total de 54 filas afectadas. Esto apunta a que el problema de consistencia en esta tabla es menos severo, pero aún significativo en ciertos campos específicos.

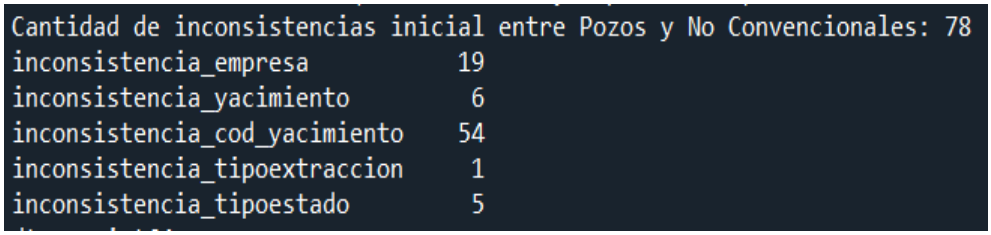


Figura 8: Inconsistencias entre Pozos y No Convencionales

En conclusión, tras la resolución de las inconsistencias principales, las discrepancias restantes representan un porcentaje muy reducido de los datos disponibles. Este resultado permite avanzar en el análisis sin que los problemas de calidad planteen un desafío significativo para los objetivos del estudio. Los pasos tomados para limpiar y consolidar estos datos se detallan a lo largo del informe.

2.4. Unidades geograficas

Esta sección del trabajo se enfoca en la validación de los datos geoespaciales de los pozos y de los registros de datos no convencionales mediante la verificación de su ubicación dentro de un sistema de coordenadas geográficas. El objetivo principal fue comprobar que la provincia registrada para cada pozo coincidiera con su ubicación geográfica real, garantizando así la precisión de los datos. Para ello, utilizamos la base de datos del Instituo Geografico Nacional, que proporciona los límites geográficos oficiales de las provincias argentinas.

El primer paso consistió en la normalización de los nombres de las provincias para corregir inconsistencias. Este proceso fue esencial, ya que encontramos discrepancias en cómo estaban escritos ciertos nombres, por ejemplo Río Negro y Tierra del Fuego.

provincia	nam
Rio Negro	Río Negro

(a) Ejemplo de variación a la hora de registrar - Río Negro

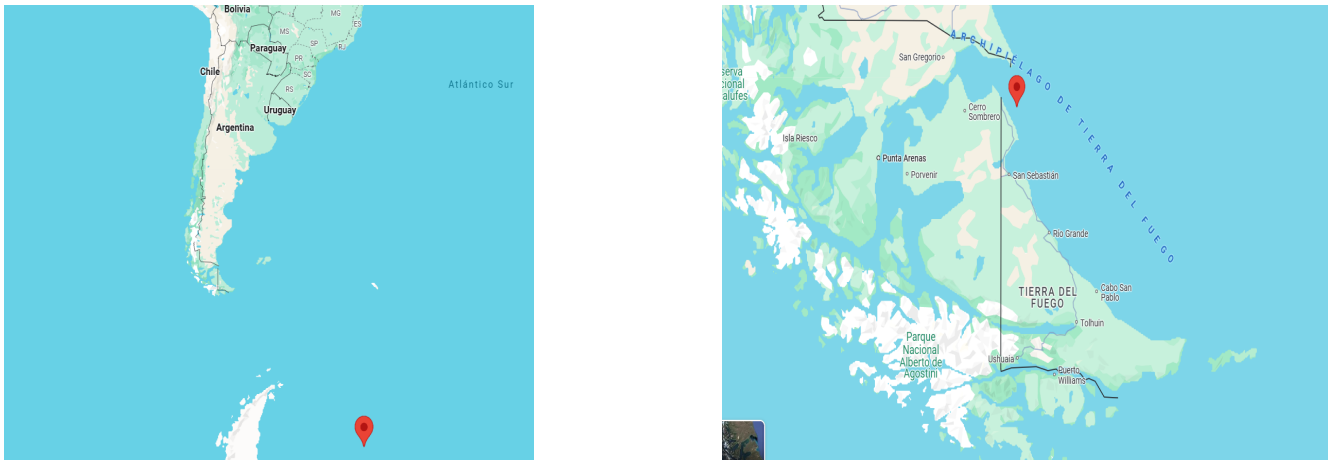
provincia	nam
Tierra del Fuego	Tierra del Fuego, Antártida e Islas del Atlántico Sur

(b) Ejemplo de variación a la hora de registrar - Tierra del Fuego

Figura 9: Provincia pertenece al dataset pozos; nam pertenece al dataset provincias.

En algunos casos, se detectaron caracteres especiales, tildes o diferencias en el uso de mayúsculas y espacios que podrían complicar el análisis al compararlos con otras fuentes de datos. En otros casos mas extremos se encontraron directamente palabras agregadas lo cual cambia la denominacion de la provincia por completo. La normalización permitió que todos los nombres de las provincias tuvieran un formato uniforme, facilitando su integración con las capas geográficas del IGN.

Una vez normalizados los nombres, vinculamos los datos de los pozos con las provincias correspondientes según sus coordenadas geográficas. Esto permitió verificar que cada pozo estuviera efectivamente ubicado dentro de los límites de una provincia específica. En este proceso, se detectaron varios casos problemáticos. Algunos pozos tenían coordenadas no válidas, lo que significa que no pertenecían a ninguna provincia o incluso estaban situados fuera de los límites del territorio nacional. Estos errores suelen derivar de problemas en el registro de coordenadas o la ausencia de datos completos. Asimismo, se identificaron pozos con provincias incorrectas registradas, un problema que podría generar errores importantes en análisis posteriores relacionados con la producción, regulación o categorización.



(a) Ejemplo 1 (b) Ejemplo 2

Figura 10: Ejemplos de coordenadas invalidas

```
La cantidad de pozos sin coordenadas validas son: 622
```

Figura 11: Coordenadas invalidas en Pozos

```
La cantidad de no Convencionales sin coordenadas válidas son: 187
```

Figura 12: Coordenadas invalidas en No convencionales

En las figuras 10 y 20, se ilustran ejemplos de coordenadas inválidas detectadas durante el análisis. En algunos casos, los pozos aparecían ubicados en el mar, lejos de cualquier área geográfica reconocida, lo que resalta la importancia de realizar este tipo de validación.

Este proceso de validación geoespacial es esencial para garantizar la calidad de los datos empleados en el análisis. Sin esta etapa de control, las coordenadas incorrectas o mal ubicadas podrían haber generado errores significativos, comprometiendo la confiabilidad de los resultados. Al asegurar que todos los pozos estén correctamente localizados dentro de las provincias correspondientes, podemos proceder con los análisis posteriores con mayor confianza y precisión, sabiendo que los datos reflejan fielmente la realidad geográfica del territorio argentino.

2.5. Deteccion de casos anomalos

La detección de casos anómalos en la producción es una parte fundamental de este análisis, ya que permite distinguir entre peculiaridades reales en los datos de producción y posibles errores en la carga que podrían invalidar la información. La identificación precisa de estas anomalías es esencial para garantizar la calidad y utilidad de los datos.

Para iniciar este análisis, decidimos centrarnos únicamente en los pozos cuyo estado operativo se clasifica como "Extracción Efectiva" o "Parado Transitoriamente". Esto se debe a que los demás pozos tienen producción igual a cero, no por características de su actividad extractiva en un momento específico, sino porque están en otras condiciones: abandonados, en estudio, o dedicados a actividades distintas, como la inyección de agua.

Estos valores de producción igual a cero introducían ruido en nuestro análisis, dificultando una interpretación clara. Si consideráramos todos los pozos, se generaría una falsa percepción de que una proporción significativa tiene producción nula. Esto sesgaría los resultados, ya que para la mayoría de los pozos no operativos, la producción igual a cero es consistente con su estado, pero no relevante para la detección de anomalías.

En consecuencia, incluir todos los pozos hubiera incrementado artificialmente la cantidad de casos detectados como anómalos, al interpretarse incorrectamente que cualquier valor distinto de cero en pozos no operativos sería atípico. Al enfocarnos únicamente en pozos operativos, eliminamos estas confusiones y logramos un análisis más preciso y representativo de la producción real.

En el análisis para la detección de casos anómalos en la producción de pozos, aplicamos el algoritmo K-means como una herramienta clave para identificar patrones y posibles irregularidades en los datos. Este enfoque permitió agrupar los pozos en función de características como la producción mensual de gas y petróleo, facilitando la identificación de comportamientos similares y posibles desviaciones significativas.

El método K-means fue seleccionado por su capacidad para dividir datos en clusters, cada uno representando un patrón característico de comportamiento en la producción. Los pozos fueron asignados a estos clusters según la similitud de sus niveles de producción. Posteriormente, analizamos las distancias de cada pozo respecto al centroide de su cluster. Aquellos pozos cuya producción estaba significativamente alejada del promedio del grupo al que pertenecían fueron identificados como potenciales casos anómalos.

Este análisis se realizó de manera separada para los pozos Convencionales y No Convencionales, debido a las diferencias inherentes en sus niveles de producción y características operativas. Separar estos grupos nos permitió ajustar mejor los parámetros del modelo y evitar confusiones entre tipos de pozos con comportamientos claramente distintos.

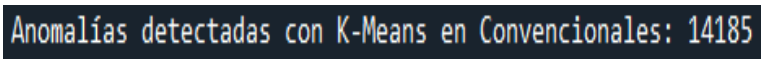


Figura 13: Cantidad de anomalias en Convencionales sin utilizacion de base de datos externa

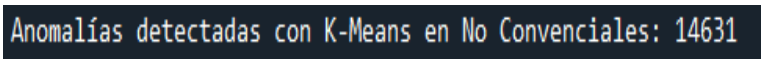


Figura 14: Cantidad de anomalias en no Convencionales sin utilizacion de base de datos externa

Como indican las imágenes, se detectaron anomalías tanto en pozos convencionales como en no convencionales. En el caso de los pozos convencionales, se identificaron 14,185 registros anómalos, mientras que en los pozos no convencionales, el número fue de 14,631 anomalías detectadas.

En los pozos convencionales, las anomalías pueden estar relacionadas con factores como una mayor diversidad operativa y geológica, o registros con valores extremos o inconsistencias. Por otro lado, en los pozos no convencionales, que suelen utilizar tecnologías modernas y operar bajo condiciones más específicas, las anomalías podrían asociarse a patrones operativos más definidos o variaciones específicas en la producción.

Es crucial interpretar cuidadosamente estas anomalías para determinar si se deben a eventos reales, como variaciones operativas o interrupciones, o a errores en los datos, como registros incorrectos o ausentes. Este análisis subraya la importancia de ajustar los criterios de detección según las características de cada tipo de pozo para obtener resultados más precisos y útiles en la toma de decisiones.

Este hallazgo subraya la importancia de contextualizar y validar las anomalías detectadas, utilizando información complementaria sobre los pozos y recurriendo a técnicas adicionales como análisis estadísticos o visualización de datos. Esto permitirá determinar si las anomalías son reflejo de condiciones reales de operación o simplemente desviaciones atribuibles a errores o limitaciones del modelo aplicado.

En el análisis de los pozos convencionales de 2024, utilizamos datos históricos de Produccion de Pozos de Gas y Petroleo - 2023 como fuente complementaria para la identificación de anomalías. A través de un modelo de Isolation Forest, donde se compararon las características de producción de agua, gas y petróleo del año actual con los patrones registrados en el año anterior. Esto nos permitio evaluar desviaciones significativas que pudieran indicar casos atípicos o anómalos.

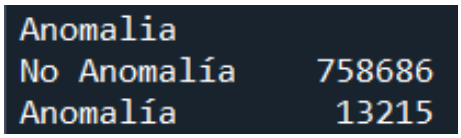


Figura 15: Cantidad de anomalias detectadas en Convencionales utilizando base de datos externa(convencionales2023)

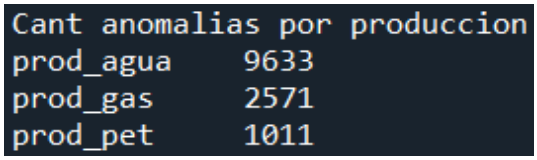


Figura 16: Cantidad de anomalias por produccion en Convencionales utilizando base de datos externa(convencionales2023)

El análisis arrojó que, del total de registros analizados para 2024, 758,686 fueron clasificados como normales, mientras que 13,215 se identificaron como anómalos. Aunque este porcentaje de anomalías es relativamente bajo, su impacto potencial en la operación y el análisis de datos es considerable. Al desglosar estas anomalías por tipo de producción, se observó que la

mayor cantidad ocurrió en la producción de agua, con 9,633 casos detectados, seguida por la producción de gas, con 2,571 casos, y finalmente la producción de petróleo, con 1,011 casos.

El uso de los datos históricos de 2023 fue crucial para contextualizar estas anomalías y determinar cómo se desviaron respecto al comportamiento promedio. Por ejemplo, en la producción de agua, los registros anómalos pueden estar vinculados a eventos como sobreextracción, fugas, problemas en el sistema de medición o cambios operativos imprevistos. En contraste, las anomalías detectadas en gas y petróleo, aunque menos frecuentes, pueden reflejar interrupciones puntuales, fluctuaciones en la extracción o problemas relacionados con la precisión de los datos registrados.

El resultado final subraya la importancia de combinar métodos avanzados de detección con datos históricos. Esto no solo mejora la precisión en la identificación de anomalías, sino que también permite interpretar las desviaciones en su contexto operativo y técnico, proporcionando herramientas más robustas para el manejo eficiente de recursos y operaciones.

En el análisis de pozos no convencionales, recurrimos a datos complementarios provenientes del Listado de pozos cargados por empresas operadoras. Esta fuente proporciona información adicional sobre los pozos y su producción basada en los registros mantenidos por las empresas propietarias y los reportes que estas presentan.

Para garantizar una base comparativa homogénea, seleccionamos los pozos registrados entre 2019 y 2024 en ambas fuentes, restringiendo el análisis a este periodo reciente. Este enfoque nos permitió comparar datos de un marco temporal coherente antes de realizar las evaluaciones y estudios correspondientes.

En el conjunto de datos proporcionados por las empresas operadoras, se especificaba la producción total acumulada por cada pozo en el periodo mencionado. En contraste, los registros de la tabla de No Convencionales detallaban la producción mes a mes y año por año. Por ello, realizamos un procesamiento de los datos para calcular la suma total de la producción registrada en esta última tabla para cada pozo, igualando así el formato de ambos conjuntos de datos.

El principio detrás de este análisis de anomalías se basó en la expectativa de que las cifras de producción entre ambas fuentes deberían coincidir exactamente o, en su defecto, mostrar discrepancias mínimas atribuibles a errores menores. Por tanto, cualquier diferencia significativa sería considerada una anomalía, ya que podría reflejar problemas en la carga de datos, en la transmisión de información o en otros procesos administrativos. En la siguiente imagen se puede ver un ejemplo de algo que categorizamos como anomalia y algo que no. Las primeras 5 filas muestran pozos para los cuales no se registro produccion en la tabla de No Convencionales pero en la tabla cargada por las empresas sí, indicando la presencia de casos anomalos en los que pozos que deberian tener produccion presentan una produccion nula. En cambio, en la ultima fila, se puede ver que el error en la carga de datos estuvo en el dataset de las empresas operadoras y por tanto, en nuestro dataset de No Convencionales que estamos analizando, no representan nignun caso fuera de lo comun. Estos ultimos casos no los contamos como anomalos.

idempresa	idpozo	d_gas_noc	id_pet_noc	d_agua_n	_pet_opera	_gas_opera	agua_oper
PCR	161305	0	0	0	138.055	0	1440.14
VIS	163138	0	0	0	73475.7	10070.4	16969.9
VIS	163139	0	0	0	53892.3	8195.83	24569.3
VIS	163140	0	0	0	68281.5	9753.52	14499
VIS	163141	0	0	0	58922.3	8533.27	27243
CGC	165743	113.249	1.415	40.851	0	0	0

Figura 17: Ejemplos de problemas de carga de datos en ambas tablas.
Observación: Las primeras 3 producciones pertenecen a la tabla noConvencionales, mientras que las ultimas 3 pertenecen a empresasoperadoras.

Establecimos un umbral de tolerancia: cualquier pozo cuya diferencia de producción entre ambas fuentes superara las 100 unidades sería clasificado como anómalo. Este criterio permite identificar posibles inconsistencias críticas que podrían afectar la fiabilidad de los datos y la calidad del análisis.

La importancia de este análisis radica en su capacidad para garantizar que los datos utilizados en estudios posteriores sean consistentes y confiables, evitando que errores o discrepancias introduzcan sesgos o conclusiones erróneas. Además, este tipo de auditorías cruzadas resalta la necesidad de integrar diferentes fuentes de datos y realizar controles de calidad exhaustivos para asegurar que las bases de datos reflejen con precisión la realidad operativa de los pozos.

Figura 18: Cantidad de anomalias detectadas en no Convencionales utilizando base de datos externa(empresasOperadoras)

Produccion	Cantidad de Anomalías
prod_gas	215
prod_pet	121
prod_agua	156

Figura 19: Cantidad de anomalias por produccion en no Convencionales utilizando base de datos externa(empresasOperadoras)

El análisis reveló un total de 275 anomalías en los registros de producción de pozos no convencionales. Al desglosar estos casos por tipo de producción, se observó que el gas fue el recurso con la mayor cantidad de irregularidades, con un total de 215 casos detectados. Le siguió la producción de agua, con 156 anomalías, y finalmente la de petróleo, con 121 registros atípicos.

La incorporación de fuentes externas, como los registros proporcionados por empresas operadoras, resultó fundamental para contrastar y validar los datos de producción. Este enfoque no solo permitió detectar y categorizar casos atípicos, sino que también sirvió para reforzar la calidad y consistencia de la información utilizada en el análisis. En última instancia, estas validaciones enriquecen la comprensión de las operaciones de los pozos no convencionales y aseguran que los resultados obtenidos sean más precisos y confiables, contribuyendo a una visión integral de la producción y las posibles áreas de mejora.

3. Decisiones tomadas

Una vez terminado el analisis de la calidad de los datasets tuvimos que decidir como resolver las cuestiones de limpieza de datos innecesarios, para eso tomamos las siguientes decisiones.

- **Transformar el dataset de 2024 en Convencionales:** Dado que para cumplir con los objetivos de este trabajo nos interesaba analizar los registros correspondientes a recursos convencionales en el año 2024, nos enfocamos en el dataset Producción de Pozos de Gas y Petróleo - 2024. Al revisar la distribución de los distintos tipos de registros, encontramos que la proporción de registros segun su tipo de recurso era la siguiente:

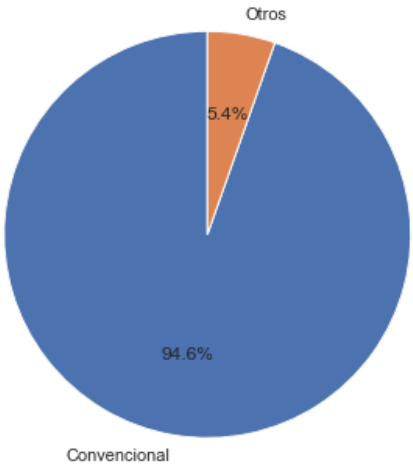


Figura 20: Proporcion de registros en base a su tipo de recurso

A partir de estos datos, y considerando que los registros con tipo de recurso convencional representaban más del 94 % del total de la tabla, decidimos eliminar aquellos con valores diferentes y "transformar"la base de datos para que contuviera únicamente pozos con recursos convencionales.

- **Unificar registros repetidos excepto en ID:** Debido a la aparición de pozos con información idéntica en todas las columnas, excepto en el ID, consideramos que estos registros corresponden al mismo pozo. Por lo tanto, unificamos en un solo registro aquellos casos en los que se presentaba esta problemática en las tablas Pozos y Convencionales, manteniendo únicamente la primera aparición de cada pozo repetido.

- **Eliminar registros con fechas nulas:** Parte del analisis de consistencia nos llevo a descubrir que algunas de las columnas con valores de fechas tenian valores nulos, al ser solamente 3 casos decidimos eliminarlos asi no obstruian nuestra investigacion y podiamos realizar un mejor analisis.

- **Acotar y eliminar valores numericos invalidos:** Inicialmente, eliminamos los registros con valores negativos en los campos que reflejan la producción, ya que no es posible producir una cantidad negativa, por lo que los consideramos datos inválidos. De manera similar, también descartamos los registros que presentaban una profundidad negativa. Posteriormente, decidimos eliminar los datos outliers encontrados en las columnas de profundidad y cota, ya que estos valores no mantenían consistencia con el resto de los datos numéricos del dataset.

- **Modificar valores nulos a No Informado:** Decidimos modificar en las tres tablas los valores nulos por "No Informado" dado que muchos de los registros tenian esa leyenda que representa lo mismo que un nan. Al realizar este ajuste, buscamos mejorar la calidad de los datos, garantizando la coherencia y la integridad de la base de datos, lo que nos permite tener un criterio unificado y facilita un análisis más preciso y consistente.

- **Modificar valores inconsistentes en la tabla Pozos:** Al encontrar una gran cantidad de discrepancias en el tipo de recurso entre la tabla Convencionales y la tabla Pozos, decidimos poner el tipo Convencional a todos los pozos con un valor distinto en la tabla Pozos dado que, si aparecian en la tabla de Convencionales, su tipo de recurso debe ser Convencional.

- **Eliminar pozos inconsistentes:** Para garantizar una consistencia total entre las tablas, procedimos a eliminar los registros correspondientes a los pozos que presentaban discrepancias en alguna de las columnas al comparar la tabla de Pozos con la tabla de Convencionales. Aplicamos el mismo criterio para los pozos que mostraban inconsistencias entre las tablas de Pozos y No Convencionales. Esta decisión se fundamentó en el hecho de que la proporción de pozos inconsistentes era extremadamente baja en relación con el total de datos, por lo que su eliminación no tendría un impacto significativo en los análisis posteriores.

- **Modificar provincias incorrectas:** Al realizar el análisis utilizando la delimitación de las provincias proporcionada por el Instituto Geográfico Nacional (IGN), identificamos que varios pozos estaban incorrectamente ubicados. En los datasets Pozos y No Convencionales se indicaba que pertenecían a determinadas provincias, pero al verificar su ubicación mediante las coordenadas geográficas, descubrimos discrepancias. Para corregir esta inconsistencia, actualizamos la columna correspondiente a "provincia", asignando a cada pozo su provincia correcta, según lo definido por sus coordenadas geográficas.
- **Eliminar pozos invalidos:** Otro aspecto que identificamos al analizar las coordenadas geográficas junto con la información proporcionada por el Instituto Geográfico Nacional (IGN) fue que algunos pozos no se encontraban dentro del territorio argentino o estaban ubicados en límites entre provincias, lo que impedía determinar con precisión su ubicación. Dado que estos casos representaban una proporción muy baja del dataset, optamos por eliminarlos para garantizar una mejor calidad en los datos y trabajar únicamente con los pozos correctamente ubicados.

4. Respuestas

4.1. Análisis descriptivo de la producción de petróleo y gas

En esta seccion se nos pide efectuar un análisis descriptivo de la producción de petróleo y gas tanto para los pozos convencionales como los no convencionales para el año 2024. Dicho análisis muestra una distribución interesante en cuanto a la cantidad de producción mensual y total. A continuación, se detalla el análisis por separado para gas y petróleo. Para hacer esto nos aseguramos de, nuevamente, usar para estos graficos nada mas los pozos con tipo de estado "Extracción Efectiva" o "Parado Transitoriamente" ya que eran los unicos que contaban con informacion verdadera sobre la produccion.

- **Produccion de Gas**

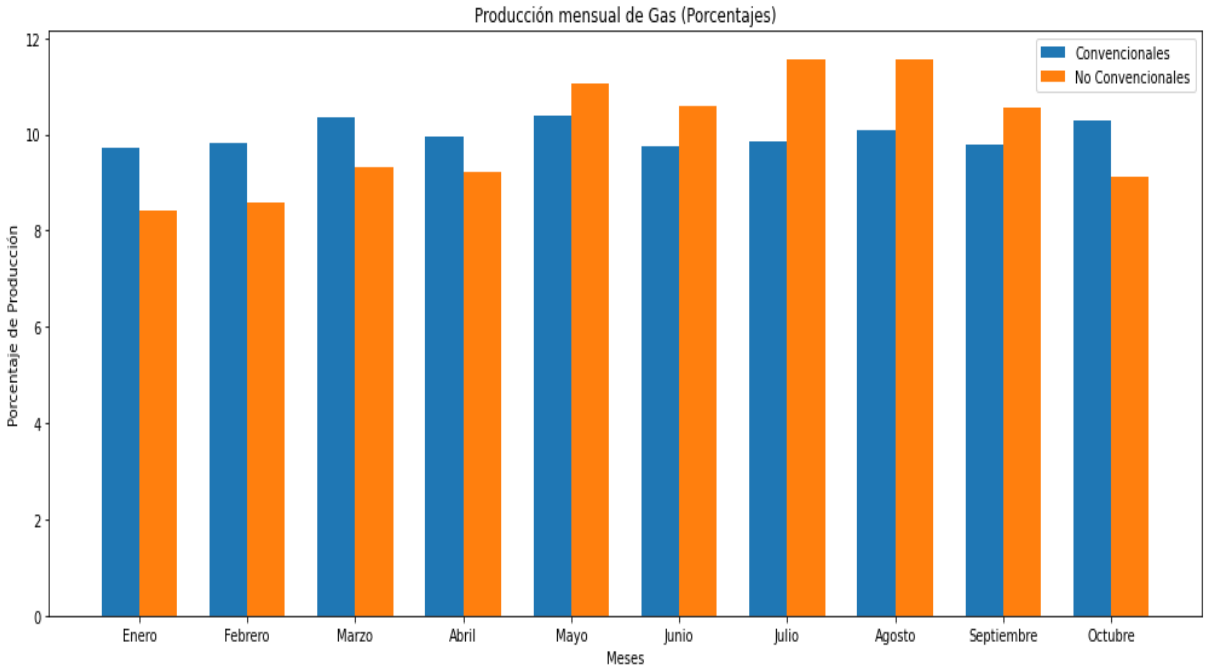


Figura 21: Produccion menusal de gas

Esta primera figura ilustra la producción mensual de gas en 2024, desglosada en porcentajes correspondientes a pozos convencionales y no convencionales. A lo largo del año, se observan variaciones significativas entre ambos tipos de pozos, lo que refleja la complejidad operativa y las dinámicas del mercado de hidrocarburos en Argentina. En los primeros meses del año, la producción general de gas muestra un crecimiento sostenido, con incrementos notables en mayo y junio, meses en los cuales los pozos no convencionales lideran la producción, superando por amplio margen a los convencionales. Este comportamiento podría estar relacionado con el aumento de la demanda energética estacional, asociada a factores climáticos, como la mayor necesidad de calefacción durante el invierno argentino. Durante el tercer trimestre, la producción tiende a estabilizarse, mientras que hacia el final del año se observan descensos puntuales en ambos tipos de pozos.

En Argentina, la producción de gas proviene principalmente de dos tipos de yacimientos: convencionales y no convencionales. Los pozos convencionales extraen gas de reservorios de alta permeabilidad, donde los hidrocarburos fluyen naturalmente hacia la superficie. En 2024, la producción de gas convencional continuó siendo un componente estable, pero en declive debido al agotamiento progresivo de yacimientos maduros, como los ubicados en la Cuenca Austral y la Cuenca Neuquina. Aunque los costos operativos y la tecnología requerida para estos pozos son menores en comparación con los no convencionales, su capacidad de expansión es limitada sin la implementación de técnicas de recuperación secundaria o terciaria. Por otro lado, los pozos no convencionales, como los ubicados en Vaca Muerta, se caracterizan por la extracción de gas atrapado en formaciones de baja permeabilidad mediante técnicas avanzadas, como el fracturamiento hidráulico y la perforación horizontal. En 2024, los pozos no convencionales fueron responsables del mayor crecimiento en la producción de gas en el país. Este comportamiento se explica por una constante inversión en infraestructura, la optimización de las técnicas de extracción y un marco regulatorio favorable que atrajo capital extranjero.

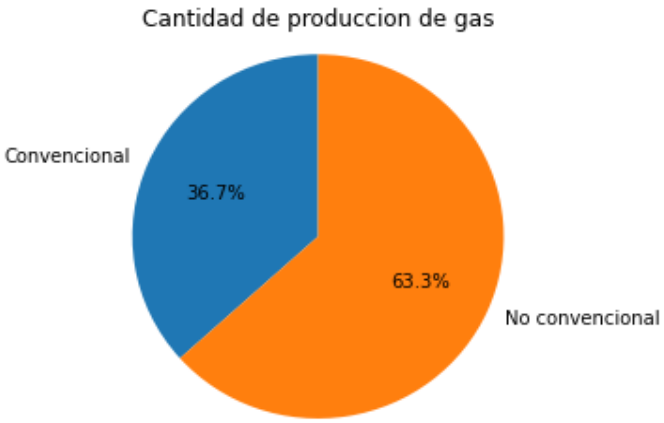


Figura 22: Produccion anual de gas

La segunda figura complementa el análisis con una visión de la producción anual total de gas en 2024. Este valor consolida la contribución de los pozos convencionales y no convencionales, destacando que los no convencionales, caracterizados por técnicas como el fracking, son los principales responsables del volumen total producido. Esto se debe a su capacidad para extraer recursos atrapados en formaciones geológicas más complejas, lo que asegura un aporte significativo a la producción nacional.

Durante el año 2024, varios factores influyeron en la producción de gas en Argentina. Por un lado, la demanda de gas se mantuvo alta debido a la dependencia del país de este recurso como principal fuente de energía para uso residencial, industrial y generación eléctrica. Además, el precio internacional del gas natural licuado (GNL) también influyó en la competitividad del gas producido localmente. Por otro lado, hubo una fuerte inversión en infraestructura para aumentar la capacidad de transporte desde las regiones productoras, como la construcción del gasoducto Néstor Kirchner, que mejoró significativamente la logística de distribución del gas no convencional. La producción fue impulsada también por avances tecnológicos que permitieron una mayor eficiencia y menores costos de extracción en los pozos no convencionales. A esto se sumaron políticas públicas que incentivaron la producción, como subsidios y contratos de precios diferenciados que estimularon la inversión en este tipo de yacimientos.

- **Produccion de Petroleo**

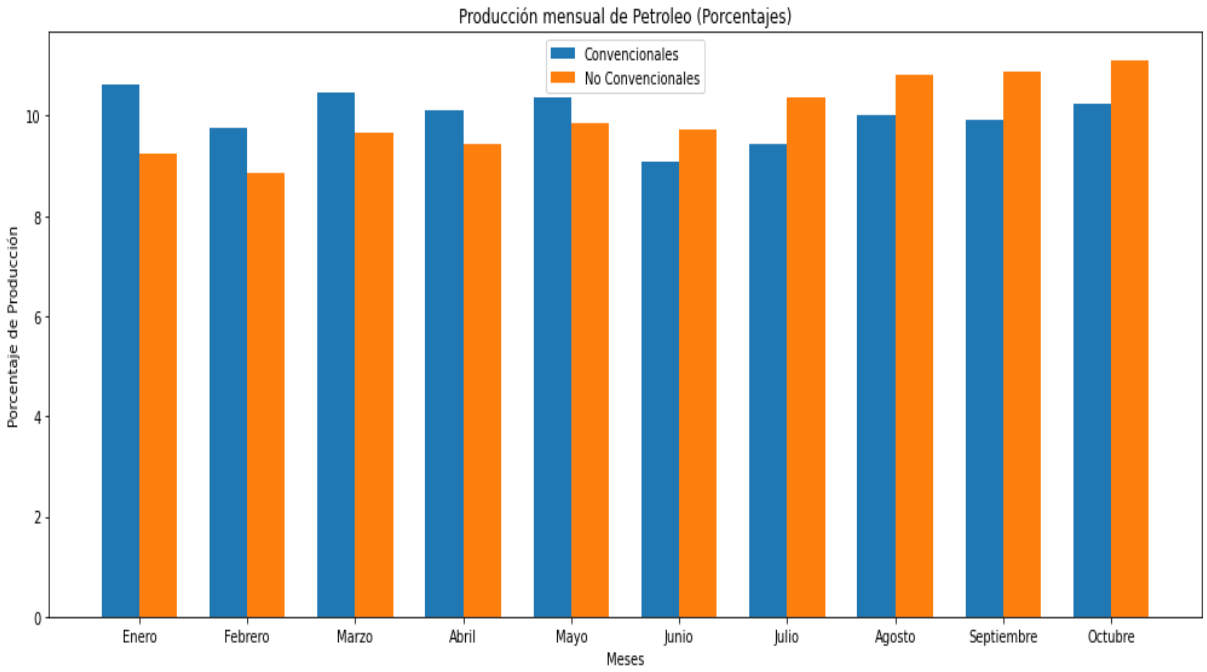


Figura 23: Produccion menusal de petroleo

La figura 23 ilustra la producción mensual de petróleo en 2024, desglosada en porcentajes correspondientes a pozos convencionales y no convencionales. A lo largo del año, se observa una tendencia que combina estabilidad y variaciones puntuales entre ambos tipos de pozos. Durante los primeros meses, los pozos convencionales lideran la producción, destacándose en enero y marzo. Sin embargo, a partir de junio, los pozos no convencionales toman protagonismo,

alcanzando valores superiores en la segunda mitad del año. Este comportamiento podría estar relacionado con una optimización de las operaciones en los yacimientos no convencionales y con una mayor participación de las tecnologías avanzadas que permiten incrementar la productividad de estos pozos.

En el caso de Argentina, la producción de petróleo proviene de dos tipos de yacimientos que responden a dinámicas operativas y tecnológicas distintas. Los pozos convencionales, ubicados en reservorios de alta permeabilidad, han sido históricamente los principales productores de petróleo en el país. Sin embargo, en 2024, muchos de estos yacimientos, particularmente los de la Cuenca del Golfo San Jorge y la Cuenca Neuquina, continúan en etapas maduras de explotación. Esto significa que, aunque siguen aportando una fracción significativa a la producción total, su rendimiento ha comenzado a declinar sin intervenciones más avanzadas, como la inyección de agua o gas para mantener la presión del reservorio. Por el contrario, los pozos no convencionales, impulsados principalmente por el desarrollo de Vaca Muerta, se han convertido en un motor clave para el crecimiento de la producción. Estos pozos emplean tecnologías como la fractura hidráulica y perforación horizontal, que permiten acceder a formaciones geológicas complejas, liberando recursos previamente inaccesibles.

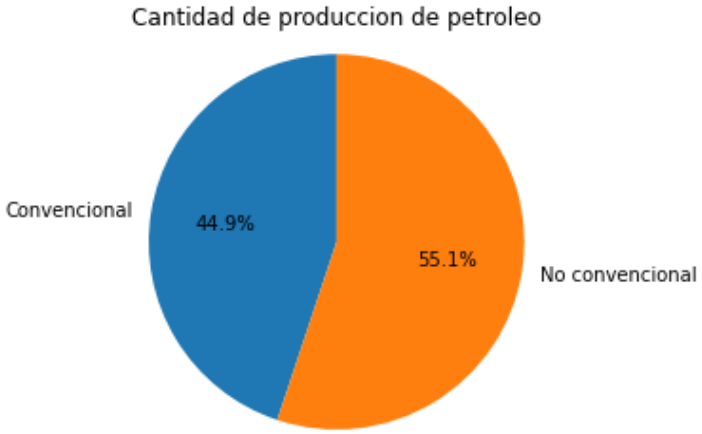


Figura 24: Produccion anual de petroleo

El desempeño de la producción de petróleo en 2024 estuvo influenciado por diversos factores. En términos de demanda, la necesidad de crudo se mantuvo alta tanto a nivel nacional como internacional, impulsada por la reactivación económica y los precios competitivos del barril. En el ámbito operativo, la capacidad para mejorar los procesos en los yacimientos no convencionales permitió aumentar los volúmenes extraídos de manera más eficiente. Esto estuvo respaldado por inversiones significativas en infraestructura, como oleoductos y plantas de procesamiento, que facilitaron el transporte y la comercialización del petróleo producido en las regiones más alejadas. Además, las políticas gubernamentales jugaron un papel fundamental, con incentivos dirigidos a promover la producción en yacimientos no convencionales mediante contratos de exportación preferencial y estabilidad fiscal para las empresas productoras.

En resumen, la producción mensual de petróleo en 2024 refleja un cambio estructural en el sector energético argentino, donde los pozos no convencionales están tomando un rol cada vez más relevante frente a los convencionales. Este cambio está impulsado por el avance tecnológico, políticas de apoyo y un contexto de mercado favorable que ha permitido maximizar el potencial de formaciones como Vaca Muerta. Este escenario subraya la importancia de continuar invirtiendo en infraestructura y tecnología para consolidar el liderazgo de Argentina en la producción de petróleo no convencional en la región.

4.2. Anomalias

En este trabajo nos encontramos con distintos tipos de anomalias, vamos a revisar sus caracteristicas y ver que justificaciones encontramos para ellas.

Para empezar analizaremos las anomalias de los pozos No Convencionales. Las primeras son las que encontramos mediante el metodo de k-means. Los pozos reconocidos por este metodo como anómalos presentan producciones muy altas, especialmente en gas y agua, con promedios de 1,876.01 m^3 de petróleo, 3,576.52 miles de m^3 de gas y 1,476.34 m^3 de agua, en comparación con los valores significativamente más bajos de los pozos no anómalos.

La distribución geográfica de las anomalías muestra que la mayoría de estos pozos se ubican en la provincia de Neuquén, con 14,354 registros, una región clave en la explotación de recursos no convencionales. También se detectaron anomalías, aunque en menor cantidad, en provincias como Río Negro, Santa Cruz, Salta, Mendoza y Chubut. Estas ubicaciones coinciden con zonas de alta actividad en desarrollos como la formación geológica Vaca Muerta, conocida por sus recursos de shale y tight. La localización precisa de los pozos más anómalos se concentra en coordenadas cercanas a (-69.426361, -38.587111), lo que indica que estan ubicados en áreas intensivas de producción, explicando los valores superiores al promedio en el resto del pais.

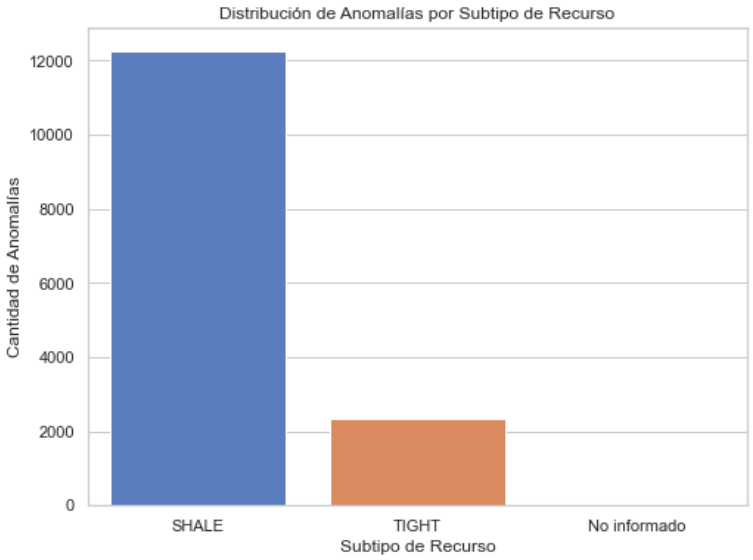


Figura 25: Distribucion del subtipo de recurso en las anomalias

Al evaluar las características específicas de los pozos anómalos, se encontró que los recursos tipo shale representan la mayoría de los casos, con un total de 12,270 registros, seguidos por pozos tipo tight, con 2,348. Además, la gran mayoría de estos pozos no están asociados a proyectos específicos reportados, con 13,521 registros clasificados como "Sin Proyecto". Solo un porcentaje menor pertenece a iniciativas identificadas, como GAS PLUS. La ausencia de proyectos claros podría implicar que estas anomalías están relacionadas con exploraciones experimentales o nuevas técnicas y por eso nos encontramos con casos anomalos.

El análisis temporal de las anomalías cruzado con datos climáticos reales muestra que eventos extremos, como las sequías de 2022-2023 y las lluvias intensas de finales de 2023 asociadas al fenómeno de El Niño, pudieron haber influido en la producción. Durante estos eventos, se observó un aumento en la producción de agua, posiblemente debido a cambios en la presión de los reservorios o a una mayor interacción con acuíferos superficiales. La producción de petróleo también se incrementó en estos períodos, lo que podría explicarse por decisiones operativas que priorizan pozos más productivos en condiciones adversas. Por otro lado, la producción de gas permaneció relativamente estable, lo que indica una mayor resiliencia en este recurso frente a las variaciones climáticas.

Estos resultados coinciden con investigaciones que sugieren que el impacto directo del clima en la geología profunda de la Cuenca Neuquina es moderado, pero que las operaciones en superficie y los costos asociados, como el acceso al agua y los insumos, sí pueden verse afectados. Además, los desafíos económicos globales, como la volatilidad de los precios del petróleo y las inversiones limitadas en mitigación climática, podrían haber influido en la capacidad de las empresas para operar con normalidad durante estos eventos.

Luego, siguiendo con los pozos No Convencionales y cruzando informacion con la fuente del listado de pozos cargado por las empresas operativas, usando validaciones hechas a mano en base a la diferencia entre la informacion para las tablas, obtuvimos la siguiente informacion.

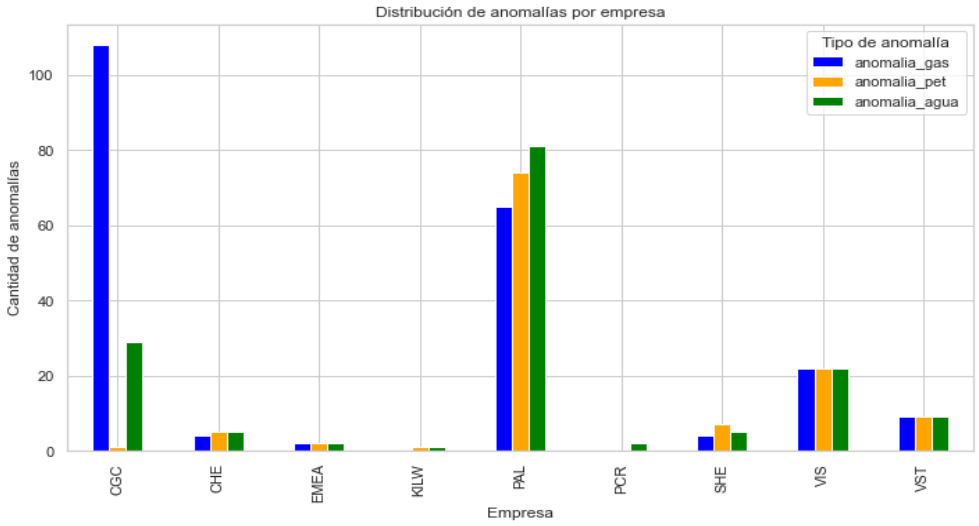


Figura 26: Cantidad de anomalias por empresa

El análisis de las anomalías en los datos de producción reveló que CGC es la empresa con mayor cantidad de discrepancias, con 108 anomalías relacionadas con el gas y 29 con el agua. PAL también muestra un alto número de anomalías, en petróleo (74), gas(65) y agua (81). Por su parte, VIS y VST presentan cantidades similares en todas las categorías, con aproximadamente 22 y 9 anomalías cada una respectivamente. El resto de las empresas tienen una cantidad muy baja de anomalias.

A nivel geográfico, Santa Cruz es la provincia con el mayor número de anomalías en gas, acumulando 108 casos, mientras que Neuquén lidera en anomalías relacionadas con petróleo (120) y agua (125). Mendoza, en contraste, tiene una incidencia mucho menor, con solo 2 anomalías, principalmente en agua.

idempresa	idpozo	d_gas_noc	d_pet_noc	d_agua_noc	_pet_opera	_gas_opera	agua_oper
CGC	160298	86917.5	3626.68	5399.28	3616.96	86682	5385.35
PAL	160455	8745.22	42331.3	31580.2	42153	8705.99	31430.4
CGC	160558	24664.4	1024.12	1144.46	1012.89	24382.3	1134.93
CGC	160559	36753.5	1359.69	2133.88	1350.86	36524.9	2122.25
CGC	160560	47375.9	1252.6	3014.94	1238.07	47200.4	2992.79

Figura 27: Primeras 5 anomalias de la tabla no Convencionales

Observación: Las primeras 3 producciones pertenecen a la tabla noConvencionales, mientras que las ultimas 3 pertenecen a empresasoperadoras.

En términos de diferencias de producción, las anomalías en gas reflejan un excedente promedio de 364 unidades en comparación con los datos del operador. Para petróleo y agua, las diferencias son negativas, con promedios de -6785 y -2405 unidades respectivamente, lo que sugiere un subregistro en los sistemas de consolidación.

Todas estas discrepancias podrían explicarse por diversos factores. En Santa Cruz, donde opera CGC, las anomalías de gas podrían estar relacionadas con problemas en la medición o en los reportes asociados a las cuencas específicas. En Neuquén, la intensa actividad en recursos no convencionales, como los desarrollos de tight y shale, podría estar vinculada a dificultades en el reporte de datos durante las etapas iniciales de desarrollo. Las diferencias negativas en petróleo y agua, en general, podrían reflejar subregistros por parte de los operadores o discrepancias en las metodologías de consolidación.

Las diferencias entre los datos cargados en ambas tablas también podrían deberse a inversiones recientes y al uso de tecnologías no convencionales. Empresas como CGC y las operadoras en Neuquén (YPF, Vista, Shell) han incrementado la producción mediante fractura hidráulica y otras tecnologías avanzadas. Estas prácticas pueden dar lugar a errores de reporte, especialmente durante las etapas iniciales de producción, cuando las estimaciones pueden diferir de los valores estabilizados. Además, puede haber reprocesamiento de datos, ya que las operadoras ajustan los volúmenes reportados para reflejar mejor las producciones reales.

Factores geográficos y administrativos también juegan un rol importante. En Santa Cruz, CGC podría estar enfrentando desafíos para consolidar datos entre los sistemas de la operadora y las autoridades nacionales. En Neuquén, el rápido crecimiento de Vaca Muerta ha generado posibles discrepancias debido a la integración de datos entre distintos sistemas operativos. Finalmente, las metodologías de medición podrían ser un factor crucial, ya que las empresas tienden a basarse en estimaciones iniciales, mientras que los sistemas de consolidación utilizan datos ajustados por auditorías posteriores. Estas complejidades reflejan la necesidad de mejorar la integración de datos y las metodologías de reporte en las principales cuencas productoras del país.

Las anomalías en pozos convencionales, detectadas mediante el método de K-Means, presentan características distintivas en las producciones de petróleo, gas y agua que las diferencia claramente de los pozos no anómalos. Estas observaciones destacan por operar bajo condiciones excepcionales o representar casos extremos dentro del contexto operativo general.

En cuanto a la producción de petróleo, los pozos anómalos registran un promedio de 317.11 barriles por día, con valores que alcanzan hasta 5,428.62 barriles por día. En comparación, los pozos no anómalos presentan un promedio considerablemente inferior de 39.36 barriles por día. Este patrón se repite con la producción de gas, donde los pozos anómalos tienen una media de 636.37 metros cúbicos por día, alcanzando máximos de 150,618.9 metros cúbicos por día, frente a los 24.71 metros cúbicos diarios de los pozos normales. De manera similar, la producción de agua en los pozos anómalos alcanza un promedio de 4,170.88 metros cúbicos por día, con valores máximos de 31,903.92 metros cúbicos diarios, superando ampliamente los 782.61 metros cúbicos por día registrados en los pozos no anómalos.

Estas anomalías podrían estar asociadas con pozos ubicados en yacimientos de alta presión o con características geológicas que facilitan producciones inusualmente altas. Asimismo, muchos de estos pozos parecen operar con métodos de extracción avanzados, como el bombeo mecánico y la cavidad progresiva, que son comunes entre las observaciones anómalas. Geográficamente, las regiones de Chubut y Santa Cruz destacan por concentrar una proporción considerable de estas anomalías, lo que sugiere una influencia significativa de las condiciones geológicas específicas de estas zonas.

Además, los pozos clasificados como petrolíferos representan la mayoría de las anomalías detectadas. Esto es consistente con el perfil de producción de los pozos anómalos, que se caracterizan por una alta producción de petróleo en comparación con gas y agua. Este predominio puede estar relacionado con el diseño y propósito inicial de estos pozos, enfocados en maximizar la extracción de petróleo, lo que podría también explicar los extremos en las métricas de producción observadas. La alta producción puede deberse a factores técnicos (como equipos avanzados de extracción), geológicos (yacimientos ricos en petróleo) o incluso a momentos específicos del ciclo de vida del pozo, donde la productividad tiende a ser máxima.

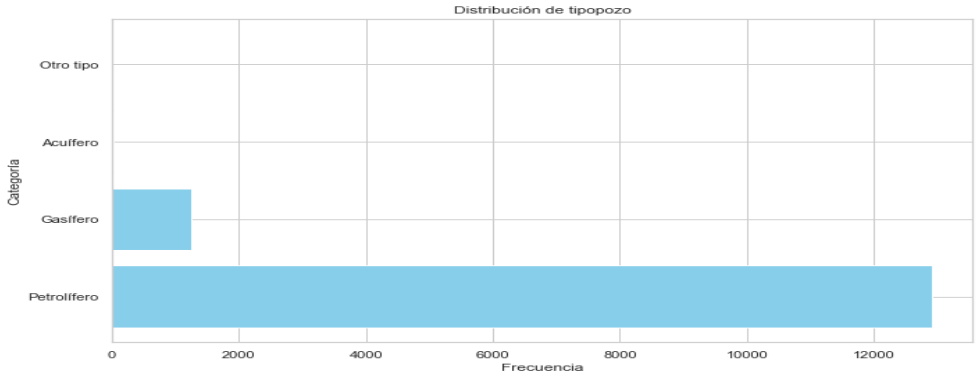


Figura 28: Proporcion de petroliferos en anomalias

El hecho de que los pozos asociados a proyectos clasificados como "*SinProyecto*" concentren casi todas las anomalías puede interpretarse de varias maneras. Esta categoría podría reflejar una falta de información detallada o una clasificación genérica que engloba pozos con características diversas y heterogéneas. Es posible que algunos pozos en esta categoría no estén vinculados a iniciativas estructuradas o grandes desarrollos, lo que podría derivar en un monitoreo menos riguroso, contribuyendo a que presenten valores extremos no controlados. También es plausible que esta clasificación incluya pozos independientes o de menor escala, operados fuera de los parámetros habituales de los grandes proyectos, lo que podría aumentar la probabilidad de registrar valores atípicos.

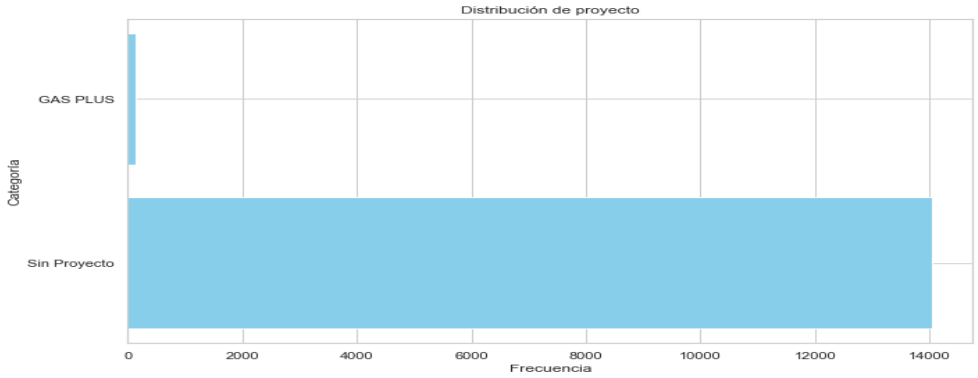


Figura 29: Proporcion de Sin proyecto en anomalias

Las anomalías en los pozos convencionales, identificadas inicialmente con K-Means y posteriormente confirmadas con el modelo Isolation Forest utilizando datos de 2023, muestran características consistentes. Aunque la cantidad de casos anómalos detectados es menor con Isolation Forest, las tendencias generales permanecen. La mayoría de los pozos anómalos siguen siendo petrolíferos y están asociados a la categoría "Sin Proyecto".

Esto puede ocurrir ya que entre 2023 y 2024, la industria petrolera en Argentina mantuvo una tendencia estable, con pocos cambios significativos en la producción y características de los pozos anómalos identificados. Esto se explica en gran medida por la naturaleza del sector, que es menos sensible a las variaciones climáticas en comparación con otros como la agricultura. Mientras que eventos como la sequía afectaron gravemente la economía agroexportadora, la producción de hidrocarburos, impulsada principalmente por la formación no convencional de Vaca Muerta, continuó creciendo a niveles récord. En 2024, esta región contribuyó con más del 50 % de la producción nacional de petróleo y gas, lo que aseguró estabilidad operativa y un aumento sostenido de la actividad en el sector.

A nivel global, la industria también se benefició de precios internacionales relativamente estables gracias a las decisiones de la OPEP, que ajustó la oferta mediante recortes en la producción para mantener un equilibrio favorable. Esto permitió a Argentina posicionarse mejor en el mercado exportador, asegurando ingresos estables y consolidando su producción sin alteraciones significativas en estos dos años. Por otro lado, la infraestructura y el desarrollo continuo en pozos no convencionales reforzaron la estabilidad, permitiendo que las anomalías observadas en los pozos petrolíferos y la categoría "Sin Proyecto"se mantuvieran en línea con las características ya detectadas en análisis previos.

En conclusión, la estabilidad observada entre 2023 y 2024 en las características de los pozos anómalos refleja un sector petrolero que opera bajo dinámicas bien establecidas, donde factores geológicos, tecnológicos y de infraestructura desempeñan un papel central. La concentración de estas anomalías en pozos asociados a la categoría "Sin Proyecto"sugiere que la operación independiente o con menor supervisión estructurada podría estar relacionada con las producciones extremas detectadas. Esto se ve reforzado por la influencia de condiciones geológicas particulares y el uso de tecnologías avanzadas de extracción, como las empleadas en regiones clave como Chubut y Santa Cruz.

Estas áreas continúan siendo epicentros de valores extremos en la producción de petróleo, gas y agua, reafirmando el impacto significativo de las características locales. En un contexto global de estabilidad de precios y crecimiento sostenido en Vaca Muerta, esta continuidad no solo subraya la resiliencia del sector en Argentina, sino también la capacidad de estas regiones y tecnologías para mantener altos niveles de productividad frente a un entorno de mercado y operativo sin mayores alteraciones.

5. Conclusion

En conclusion, el proceso de limpieza y análisis realizado sobre el dataset de pozos ha sido esencial para mejorar la calidad y precisión de la información disponible. A través de la revisión y estandarización de los datos, se lograron eliminar registros ambiguos, valores vacíos e inconsistentes, lo que permitió obtener una representación más fiel de los tipos de pozos y sus producciones. Este proceso de depuración fue crucial para identificar patrones claros en la producción de los pozos y comprender mejor los factores que influyen en la misma.

Aunque el dataset contenía un gran porcentaje de datos correctos desde el principio, lo que refleja una buena calidad general de la información, el análisis permitió identificar áreas que aún pueden mejorarse y ajustarse para optimizar la confiabilidad de los datos en el futuro. Este trabajo no solo contribuye a una representación más precisa de la industria energética, sino que también establece las bases para futuras investigaciones y análisis sobre la producción de petróleo y gas, destacando la importancia de mantener una base de datos depurada y consistente.

A lo largo de este análisis, se utilizaron varias reglas de validación de los datos que permitieron asegurar la integridad y precisión de la información. Las reglas aplicadas fueron:

1. **Valores de Producción no Negativos:** Es imposible que la producción de petróleo, gas o agua sea negativa, ya que esto contradice las leyes físicas y la lógica operativa. Errores como estos fueron encontrados tanto en la tabla de Convencionales como en la de No Convencionales "prod_agua", "pros_gas" y en "prod_pet".

idempresa	anio	mes	idpozo	*prod_pet*	prod_gas	prod_ag...
PSD	2024	4	134542	-0.01	-0.01	-0.51

Figura 30: Valores negativos en la produccion de pozos Convencionales

idempresa	anio	mes	idpozo	prod_pet	*prod_gas*	prod_ag...
PLU	2020	5	153228	1.67	-12.267	0
PLU	2020	5	153227	1.02	-7.519	0

Figura 31: Valores negativos en la produccion de pozos No Convencionales

   Este tipo de error suele deberse a problemas en la entrada de datos, como valores invertidos o mal ingresados. Detectar estos errores asegura que los análisis posteriores no estén distorsionados por valores físicamente imposibles.

2. **Coherencia Temporal en Fechas de Inicio:** La fecha de inicio de un pozo debería ser anterior o igual al año de producción registrado. Si un pozo está registrado como operativo en un año previo a su fecha de inicio, esto indica una inconsistencia en los datos. Errores como este fueron encontrados en la tablas Pozos. Este análisis permite identificar problemas en la asignación de datos temporales y asegurar que la línea de tiempo de producción sea coherente y confiable.

3. **Rangos Plausibles para Cota y Profundidad:** La validación de los datos de cota y profundidad es crucial para garantizar que los registros sean coherentes con valores físicamente posibles y técnicamente alcanzables:

   Cota: El valor de cota de un pozo debe ser inferior a 5600 metros. Cotas más altas podrían indicar errores de entrada o datos inverosímiles que no representan la realidad geográfica de los pozos en la región.

   Profundidad: La profundidad de un pozo debe estar dentro del rango plausible, es decir, no negativa (mayor o igual a 0) y menor a 27,100 metros. Valores fuera de este intervalo suelen ser errores de medición , entradas incorrectas o outliers, ya que exceden los límites prácticos de perforación convencional y no convencional.

4. **Consistencia entre Bases de Datos:** Las bases con las que trbajamos deberían coincidir razonablemente con los valores registrados en los datos de pozos con los Convencionales y los No Convencionales. Una discrepancia significativa (por ejemplo, superior al 10 %) podría indicar errores de registro o reportes inexactos. Esto fue analizado en la seccion 2.3 de consistendia de los datos donde fuimos confirmando que no había superposición entre las categorías de pozos convencionales y no convencionales. Sin embargo, se detectaron discrepancias en columnas específicas, como el "tipo de recursoz el çódigo de yacimiento", las cuales fueron significativamente reducidas tras el proceso de limpieza. Revisar estas diferencias ayuda a identificar errores en el cruce de datos entre diferentes fuentes y fortalece la confianza en las cifras reportadas.

5. **Verificación de Coordenadas Geográficas Válidas:**Las coordenadas geográficas asociadas a cada pozo deben estar dentro de los límites geográficos de las provincias de Argentina. Donde en la seccion 2.4 de analisis de unidades geograficas encontramos 622 pozos y 187 de los pozos No Convencionales con coordenadas geograficas invalidas. Esto asegura que cada pozo esté correctamente asignado a una provincia válida y evita inconsistencias en los datos de ubicación. Si las coordenadas de un pozo no coinciden con los límites de ninguna provincia o están mal asignadas, esto indica un error que debe ser corregido.

La aplicación de estas cinco reglas de validación permitió detectar varias anomalías y errores en los datos de producción de los pozos de gas y petróleo, mejorando la calidad de la información utilizada para el análisis. Estas validaciones son fundamentales para garantizar la precisión de los informes y la toma de decisiones en el sector energético. La implementación de estas reglas no solo mejora la fiabilidad de los datos, sino que también contribuye al desarrollo de mecanismos más robustos para la detección de desviaciones en futuros conjuntos de datos.