DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 01

Práctica Aproximación a la Calidad De Datos

December 8, 2024 Calidad de Datos

Grupo: LutSaRod

Integrante	LU	Correo electrónico
Said, Tomás Uriel	170/23	saidtomasur@gmail.com
Tag, Lucio	876/22	luciotag2011@gmail.com
Mizrahi, Rafael	282/22	rafamizrahi30@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

Contents

1. Caso de estudio y objetivo del análisis

La calidad de los datos es un pilar fundamental para cualquier análisis significativo. En este trabajo, el foco está puesto en la información publicada por la Secretaría de Energía sobre la producción de gas y petróleo, tanto de fuentes convencionales como no convencionales. Este caso de estudio tiene como objetivo evaluar la calidad de estos conjuntos de datos y proponer herramientas para mejorar su uso en análisis posteriores.

El análisis se centrará en dos conjuntos principales: uno que describe los pozos en el país y otro que detalla la producción de gas y petróleo, específicamente en el contexto de la producción no convencional. Estos datos serán sometidos a una serie de pruebas para garantizar su unicidad, consistencia interna y concordancia entre fuentes. Además, se evaluará cómo la incorporación de información geográfica puede enriquecer los resultados.

Los datos de baja calidad pueden introducir errores significativos en las decisiones basadas en ellos, lo que puede llevar a interpretaciones erróneas, pérdida de recursos y conclusiones imprecisas. En el caso del gas y petróleo, asegurar que los datos sean completos, consistentes y precisos es crucial no solo para realizar diagnósticos confiables, sino también para detectar irregularidades en la producción y prever posibles anomalías.

Un enfoque riguroso en la calidad de los datos permite detectar inconsistencias y desvíos que, de otro modo, podrían comprometer la validez de los análisis.

En este contexto, se utilizarán reglas de validación y técnicas descriptivas para identificar casos anómalos, haciendo énfasis en la integración de datos complementarios y validaciones geográficas que fortalezcan las conclusiones.

Este trabajo no solo busca diagnosticar problemas en los datos, sino también proponer estrategias para mejorar su calidad, asegurando que estén preparados para futuros análisis críticos en el sector energético.

2. Conozcamos los Datos

Para el análisis de los datos, utilizamos el lenguaje Python junto con Jupyter Notebook, herramientas ampliamente utilizadas en el análisis científico de datos. Dentro de este ecosistema, la librería Pandas se ha utilizado como base para la manipulación y el tratamiento de los datos. Otras librerías de Python, instaladas mediante el gestor de paquetes pip, complementaron ciertas tareas específicas.

Pandas nos permite leer datos desde diversas fuentes. En nuestro caso, hemos trabajado con archivos en formato CSV correspondientes a los datasets provistos. Comenzamos explorando la estructura de los datos, determinando la cantidad de filas y columnas presentes en cada conjunto:

2.1 Dataset: Capítulo IV: Pozos

- **Filas:** 84.332
- **Columnas:** 26

2.2 Dataset: Producción de pozos de gas y petróleo no convencional

- **Filas:** 324.077
- **Columnas:** 40

2.3 Dataset: Producción de pozos de gas y petróleo 2024

- **Filas:** 818788
- **Columnas:** 38

Esta exploración inicial es clave para estructurar y planificar los análisis subsecuentes, identificando tanto las variables relevantes como los posibles problemas de calidad presentes en los datos.

A continuación, se presentan los campos de las tres tablas: **Pozos**, **Producción** y **Producción No Convencional**. Cada tabla incluye una breve descripción de cada campo.

Campo	Descripción
sigla	Sigla del pozo
idpozo	Identificador único del pozo
area	Área geográfica del pozo
cod area	Código del área
empresa	Empresa propietaria del pozo
yacimiento	Yacimiento asociado al pozo
cod yacimiento	Código del yacimiento
formacion	Formación geológica
cuenca	Cuenca donde se encuentra el pozo
provincia	Provincia donde se encuentra el pozo
cota	Altitud o cota del pozo
profundidad	Profundidad del pozo
clasificacion	Clasificación del pozo
subclasificacion	Subclasificación del pozo
tipo recurso	Tipo de recurso extraído
sub tipo recurso	Subtipo de recurso extraído
gasplus	Información relacionada al gasplus
tipopozo	Tipo de pozo
tipo extraccion	Tipo de extracción en el pozo
tipo estado	Estado del pozo
adjiv fecha inicio perf	Fecha de inicio de perforación
adjiv fecha fin perf	Fecha de fin de perforación
adjiv fecha inicio term	Fecha de inicio de terminación
adjiv fecha fin term	Fecha de fin de terminación
geojson	Información geoespacial en formato GeoJSON
geom	Información geoespacial en formato geométrico

Table 1: Campos de la **Tabla Pozos**

Pasemos a ver los campos de la siguiente tabla, que detalla la información de producción asociada a los pozos. Esta tabla incluye datos sobre las empresas, los recursos extraídos y las características relacionadas con las actividades de producción.

Campo	Descripción
idempresa	Identificador único de la empresa
anio	Año de la producción
mes	Mes de la producción
idpozo	Identificador único del pozo
prod pet	Producción de petróleo
prod gas	Producción de gas
prod agua	Producción de agua
iny agua	Inyección de agua
iny gas	Inyección de gas
iny co2	Inyección de CO2
tef	Tiempo efectivo de funcionamiento
vida util	Vida útil estimada del pozo
tipo extraccion	Tipo de extracción realizado
tipo estado	Estado operativo del pozo
tipopozo	Clasificación del pozo
empresa	Nombre de la empresa operadora
formacion	Formación geológica asociada
cuenca	Cuenca asociada
provincia	Provincia donde se encuentra el pozo
tipo de recurso	Tipo de recurso producido
clasificacion	Clasificación general del pozo
subclasificacion	Subclasificación específica
fecha data	Fecha de ingreso de los datos

Table 2: Campos de la **Tabla Producción**

Por último, veamos la tabla de **Producción No Convencional**, que incluye campos similares a los de la tabla de Producción, pero enfocados en pozos de recursos no convencionales. Se incorporan además datos geográficos específicos y características únicas de este tipo de producción. La única diferencia entre ambas tablas radica en la inclusión de las columnas coordenadax y coordenaday, que especifican las coordenadas geográficas del pozo.

3. Uso de las Capas SIG

En esta sección, describimos los usos potenciales de las capas SIG proporcionadas por el *Instituto Geográfico Nacional* (IGN) y cómo pueden contribuir a la validación y análisis de los datos de pozos de petróleo y gas. Estas capas geoespaciales permiten realizar cruces y verificaciones con las coordenadas y las áreas administrativas incluidas en los datasets.

3.1 Usos Potenciales de las Capas SIG

1. **Validación de Coordenadas Geográficas:** Utilizar las capas SIG del IGN para verificar que las coordenadas de los pozos (`coordenadax`, `coordenaday`) estén dentro de las áreas permitidas, como provincias, cuencas o áreas de yacimientos. También permite detectar valores fuera de los límites geográficos válidos, como coordenadas invertidas (latitud y longitud mal asignadas) o errores de escala.

2. **Cruce con Unidades Administrativas:** Validar que la provincia asignada (`provincia`) corresponda correctamente a las coordenadas geográficas de cada pozo. Este cruce ayuda a identificar inconsistencias, como pozos registrados en una provincia que en realidad se encuentran en otra.

3. **Validación de Áreas de Producción:** Confirmar que las áreas de concesión o permisos (`areapermisoconcesion`, `idareapermisoconcesion`, `areayacimiento`) coincidan con los límites definidos en las capas del IGN. También es posible verificar que los pozos estén efectivamente ubicados dentro de sus áreas asignadas.

4. **Creación de Reportes Geoespaciales:** Generar mapas interactivos para visualizar los pozos y las áreas de concesión, y compararlos con las capas SIG. Este enfoque permite identificar patrones o irregularidades geográficas, como pozos agrupados en áreas de baja producción.

3.2 Resultados del Análisis Geoespacial

Si bien intentamos usar las capas SIG del IGN para realizar estos análisis, el procesamiento resultó ser muy pesado debido a la gran cantidad de registros en los datasets. En consecuencia, optamos por utilizar un mapa base de Argentina para visualizar la ubicación de los pozos.

Al realizar esta validación, identificamos más de 500 registros que inicialmente parecían estar fuera del país. Sin embargo, al analizar el gráfico generado, se observó que estos registros realmente se encuentran dentro de los límites de Argentina. Este hallazgo destaca la importancia de combinar herramientas geoespaciales con validaciones visuales para garantizar la precisión de los datos.

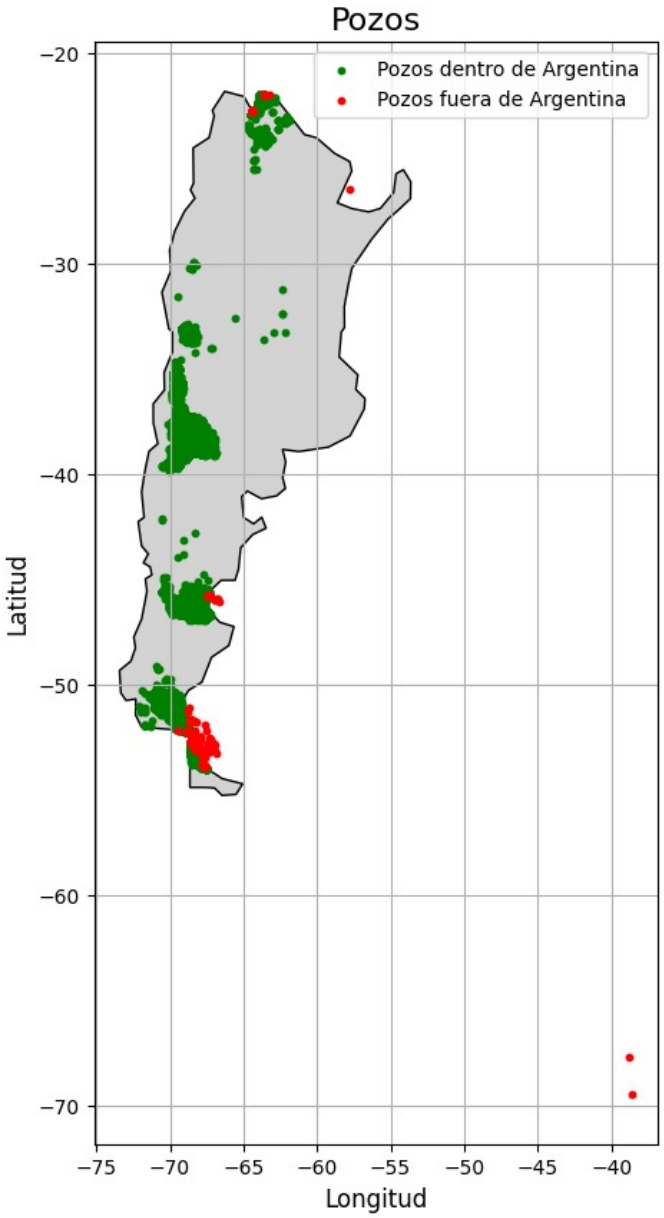

Figure 1: Enter Caption

3.3 Referencia a las Capas SIG del IGN

Para más información sobre las capas SIG utilizadas, consultar: Instituto Geográfico Nacional (Capas SIG).

4. Unicidad y Consistencia de los Datos

La importancia de chequear la unicidad y consistencia de los datos en los datasets radica en garantizar la calidad y fiabilidad de la información utilizada en el análisis. La unicidad permite identificar de manera inequívoca los registros, evitando duplicados que puedan distorsionar los resultados. Por su parte, la consistencia asegura que los datos sean

coherentes entre sí y cumplan con las reglas establecidas para su estructura y relaciones, previniendo errores y conflictos en los análisis o modelados posteriores.

4.1 Chequeo de la Unicidad de los Datos

En esta subsección, abordaremos la evaluación de la unicidad en los datos, comenzando por analizar si los identificadores clave, como el `idpozo` son únicos dentro de los datasets. Este chequeo nos permitirá garantizar que cada registro representa una entidad distinta y no hay duplicados que puedan comprometer la validez del análisis.

Para la tabla `Pozos`, realizamos un análisis de la unicidad del campo `idpozo`, que sirve como identificador único de los pozos. Los resultados obtenidos fueron los siguientes:

- **Cantidad de identificadores únicos de pozos (`idpozo`):** 84,332
- **Cantidad total de filas en la tabla:** 84,332

El análisis confirma que todos los registros de la tabla `Pozos` tienen un identificador único. Esto garantiza que el campo `idpozo` no contiene duplicados y puede ser utilizado como clave primaria para identificar de manera inequívoca cada pozo.

4.2 Chequeo de la Consistencia de los Datos

La consistencia de los datos es esencial para garantizar que la información en los datasets sea coherente y cumpla con las reglas lógicas esperadas. Este análisis permite identificar discrepancias, errores o valores anómalos que podrían comprometer la calidad del análisis.

En esta subsección, evaluaremos la consistencia de los datos en las tablas `Pozos`, `Producción` y `Producción No Convencional`. En primer lugar, analizamos la presencia de valores nulos en la tabla `Pozos`, ya que pueden indicar problemas de completitud en los datos. A continuación, presentamos los resultados obtenidos:

Valores Nulos en la Tabla Pozos

Identificamos que los valores nulos se concentran exclusivamente en las siguientes columnas:

- `empresa`: 897 valores nulos.
- `formacion`: 2815 valores nulos.
- `adjiv_fecha_inicio_perf`: 34,139 valores nulos.
- `adjiv_fecha_fin_perf`: 34,284 valores nulos.
- `adjiv_fecha_inicio_term`: 36,594 valores nulos.
- `adjiv_fecha_fin_term`: 36,593 valores nulos.

En total, los valores nulos representan aproximadamente la mitad del dataset, afectando más de 25,000 registros. Estas columnas contienen información relevante, como la empresa operadora, la formación geológica y las fechas de actividades de perforación y terminación de pozos.

Aun así, decidimos quedarnos con estos valores debido a la gran cantidad de registros afectados. Consideramos que eliminarlos podría generar una pérdida significativa de información. Sin embargo, tenemos en cuenta que estos valores nulos afectarán el análisis en aspectos relacionados con la producción de los pozos por fecha y actividades específicas, como las de perforación y terminación.

4.3 Chequeo de la Consistencia de los Datos

A continuación, se describen los chequeos realizados para garantizar la consistencia de los datos en los datasets, junto con las decisiones tomadas en cada caso:

- **Chequear si las fechas de inicio son anteriores a las fechas de fin:** Analizamos las columnas "adjiv_fecha_inicio_perf" y "adjiv_fecha_fin_perf". Encontramos que existen 3 registros con fechas inconsistentes. La decisión tomada fue eliminarlos, ya que representan un número bastante bajo.
- **Confirmar que las columnas de producción (`prod_pet`, `prod_gas`, etc.) y de inyección (`iny_agua`, `iny_gas`) no tienen valores negativos:** Identificamos valores negativos en las siguientes columnas:
  - `prod_pet`: 1 registro.

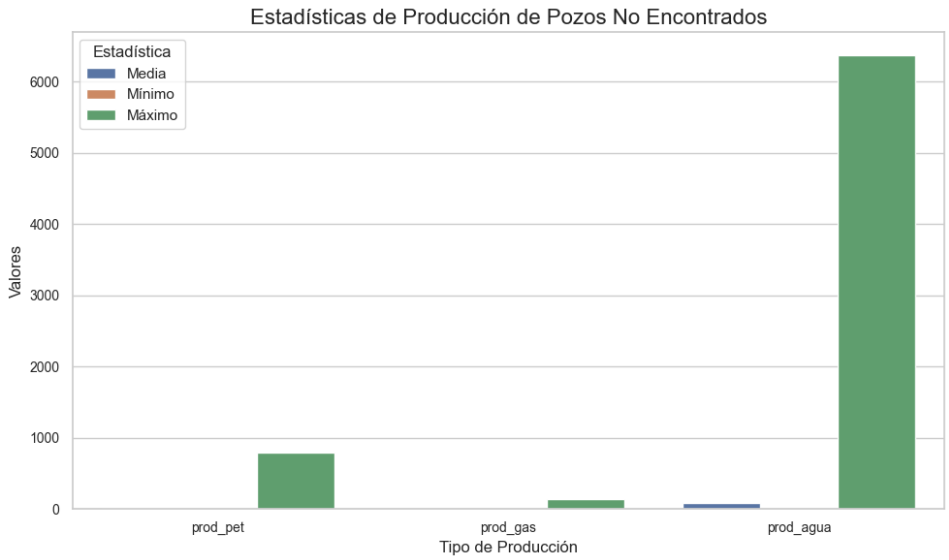

Figure 2: Estadísticas de Producción de Pozos No Encontrados en la Tabla Pozos

  - `prod_gas`: 1 registro.
  - `prod_agua`: 1 registro.
  - `iny_agua`: 0 registros.
  - `iny_gas`: 0 registros.

  La decisión fue eliminar esos tres registros con valores negativos.

- **Validar si la combinación de `anio`, `mes` e `idpozo` es única** en las tablas "produccion" y "produccion_no_convencional": Verificamos que la combinación es única, ya que la cantidad de combinaciones únicas (818,788) coincide con el total de registros en las tablas (818,788).
- **Validar si los meses (`mes`) están dentro del rango válido (1 a 12):** No se encontraron registros con meses fuera de rango.
- **Validar que todos los `idpozo` en `produccion` existen también en `pozos`:** Identificamos que 55 `idpozo` están presentes en la tabla `produccion` pero no en `pozos`. Para determinar si eliminarlos del dataset correspondiente, analizaremos qué ocurre con la producción asociada a estos pozos:
  - Analizamos la producción de los pozos no encontrados en `pozos` y encontramos los siguientes resultados estadísticos: Observamos que, aunque la media de producción no es particularmente alta, los valores máximos

Columna	Media	Mínimo	Máximo
`prod_pet`	15.78	0.00	787.65
`prod_gas`	1.41	0.00	134.85
`prod_agua`	78.14	0.00	6373.90

Table 3: Estadísticas de Producción de los Pozos No Encontrados en la Tabla Pozos.

    indican que algunos de estos pozos tienen una producción significativa (e.g., `prod_pet` con un máximo de 787.65). Por esta razón, decidimos mantener estos registros en el dataset, ya que eliminarlos podría afectar la representatividad del análisis.

Si bien en esta sección analizamos la consistencia de los tres datasets incluyendo todos los campos y todos los tipos de pozos, a continuación, nos vamos a centrar específicamente en los pozos de Petróleo y Gas. Este enfoque nos permitirá realizar un análisis más detallado y relevante para el objetivo del trabajo.

5. Análisis Descriptivo de la Producción de Petróleo y Gas

En esta sección, efectuaremos un análisis descriptivo de la producción de petróleo y gas, considerando tanto los pozos convencionales como los no convencionales para el año 2024. Este análisis se realizará de manera separada para petróleo y gas, evaluando los principales indicadores y características de cada uno.

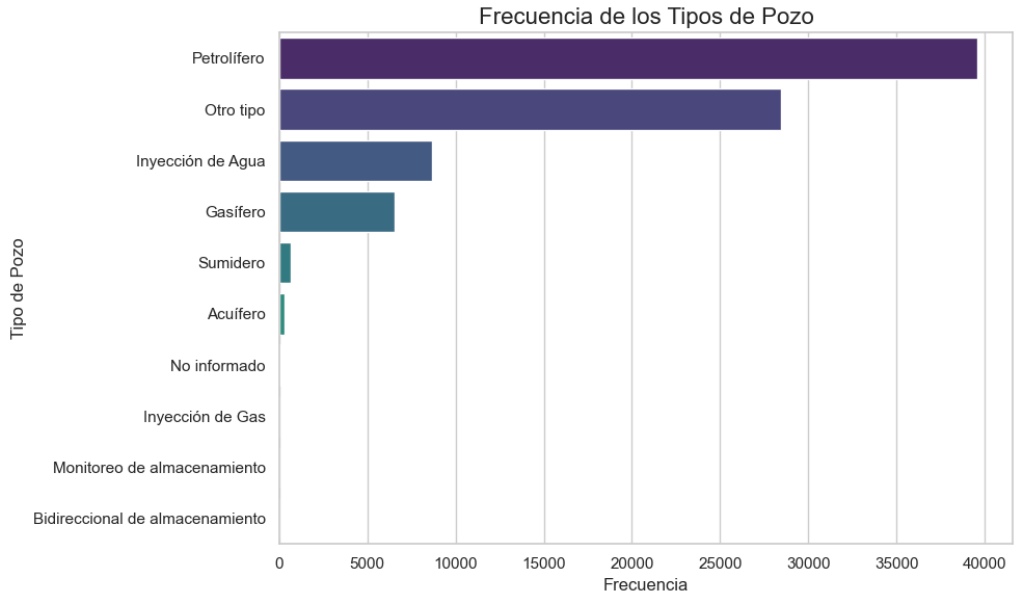

Figure 3: Frecuencia de los Tipos de Pozos

5.1 Frecuencia de los Tipos de Pozo

Analizamos la frecuencia de los valores únicos en la columna `tipopozo` para evaluar con qué tipos de pozos nos quedaremos en las próximas instancias, dado que nuestro interés principal radica en los pozos Petrolíferos y Gasíferos.

Tipo de Pozo	**Frecuencia**
Petrolífero	39,606
Otro tipo	28,466
Inyección de Agua	8,640
Gasífero	6,523
Sumidero	634
Acuífero	327
No informado	68
Inyección de Gas	14
Monitoreo de almacenamiento	2
Bidireccional de almacenamiento	2

Table 4: Frecuencia de los Tipos de Pozo en la Columna `tipopozo`.

Luego de este análisis, tomamos la decisión de quedarnos exclusivamente con los pozos de tipo `Petrolífero` y `Gasífero`, ya que son los relevantes para este trabajo. Este filtrado nos deja con un total de 46,129 registros, distribuidos en 26 columnas.

5.2 Análisis de Producción de Petróleo Convencional

En esta sección, comenzamos el análisis de la producción de petróleo. Primero seleccionamos las columnas que contienen información de la producción convencional y filtramos los datos correspondientes al año 2024. Este enfoque nos permitirá realizar futuras comparaciones con la producción no convencional, destacando las principales características de cada tipo de pozo.

Para analizar la producción de petróleo convencional en el año 2024, seleccionamos los pozos con mayor y menor producción. A continuación, se presenta una tabla que muestra los 5 pozos con mayor producción y los 5 con menor producción, junto con sus respectivas ubicaciones:

Explicación del Análisis

Del análisis, observamos que los pozos con mayor producción de petróleo están todos ubicados en la provincia de Neuquén, destacándose el pozo con ID `164941`, que aparece repetidamente con valores de producción significativamente altos, alcanzando un máximo de 19,826.739. Esto refleja una fuerte concentración de la producción en esta región.

ID Pozo	Producción de Petróleo	Provincia
164941	19,826.739	Neuquén
164941	18,034.332	Neuquén
164941	16,436.799	Neuquén
164941	14,645.788	Neuquén
164941	14,528.627	Neuquén
32186	0.000	Rio Negro
144117	0.000	Rio Negro
145614	0.000	Rio Negro
145615	0.000	Rio Negro
145626	0.000	Rio Negro

Table 5: Top 5 y Bottom 5 Pozos de Producción de Petróleo Convencional en 2024.

Por otro lado, los pozos con menor producción (0 en este caso) están todos ubicados en la provincia de Río Negro. Esto podría indicar pozos inactivos o que no están actualmente en operación, pero fueron registrados en el dataset. Estos datos resaltan la diferencia significativa en la actividad productiva entre las dos provincias analizadas.

Análisis de la Producción Total de Petróleo Convencional en 2024

Hacemos la suma de la columna correspondiente para obtener el total producido en pozos convencionales de petróleo durante el año 2024, obteniendo un total de **33,313,653.11 litros**.

En consecuencia, surgió la curiosidad de analizar en qué provincias se produce más petróleo. Para ello, agrupamos los datos por provincia y sumamos la producción total de petróleo en cada una. A continuación, presentamos un gráfico que muestra la distribución de la producción de petróleo convencional por provincia.

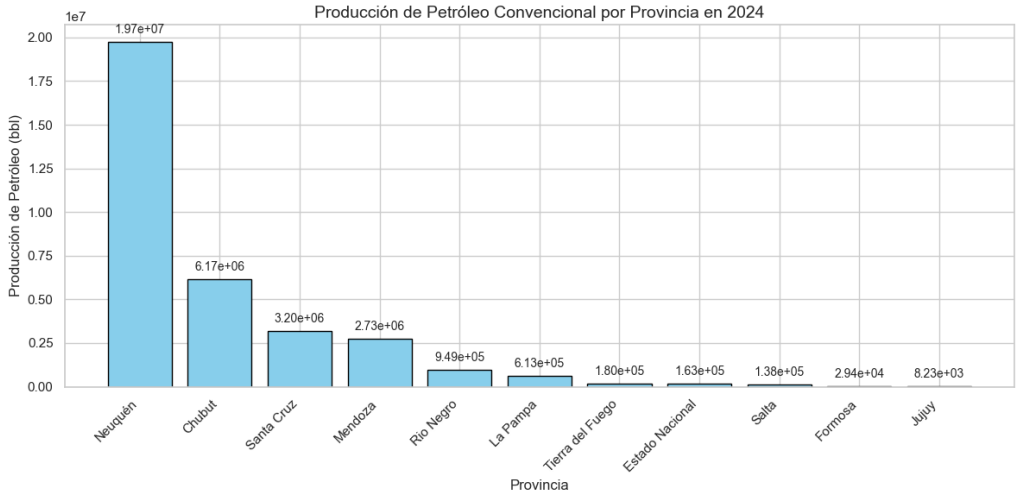

Figure 4: Producción de Petróleo Convencional por Provincia en 2024

El gráfico refleja que la provincia de **Neuquén** es la mayor productora de petróleo convencional en 2024, con una producción significativamente superior a las demás provincias, alcanzando casi 20 millones de litros. En segundo lugar se encuentra **Chubut**, seguida por **Santa Cruz** y **Mendoza**. Estas cuatro provincias concentran la mayor parte de la producción.

En contraste, provincias como **Jujuy**, **Formosa**, y **Salta** tienen producciones mucho menores, lo que destaca la disparidad en la actividad petrolera entre las regiones analizadas. Este análisis pone en evidencia la concentración de la producción en pocas provincias clave.

5.3 Análisis de Producción de Petróleo No Convencional

En esta sección, analizamos la producción de petróleo no convencional en el año 2024. Este análisis complementa el realizado previamente para la producción convencional, permitiendo comparaciones futuras entre ambos tipos de pozos.

Comenzamos calculando el total producido en los pozos no convencionales durante 2024. De esta manera, obtenemos un panorama general de la contribución de este tipo de explotación a la producción total de petróleo en el país. A continuación, presentamos los resultados del total producido en 2024:

- **Total producido en pozos no convencionales:** 18,630,057.34 litros.

Para analizar la producción de petróleo no convencional en el año 2024, seleccionamos los pozos con mayor y menor producción. A continuación, se presenta una tabla que muestra los 5 pozos con mayor producción y los 5 con menor producción, junto con sus respectivas ubicaciones:

ID Pozo	Producción de Petróleo (Litros)	Provincia
164941	19,826.739	Neuquén
164941	18,034.332	Neuquén
164941	16,436.799	Neuquén
164941	14,645.788	Neuquén
164941	14,528.627	Neuquén
159688	0.000	Rio Negro
156873	0.000	Neuquén
155540	0.000	Neuquén
155539	0.000	Neuquén
155525	0.000	Neuquén

Table 6: Top 5 y Bottom 5 Pozos de Producción de Petróleo No Convencional en 2024.

Los primeros 5 pozos coinciden con los de producción convencional, lo que nos llevó a considerar la posibilidad de que existan pozos que compartan características tanto de producción convencional como no convencional. Para explorar esta hipótesis, generamos el siguiente gráfico:

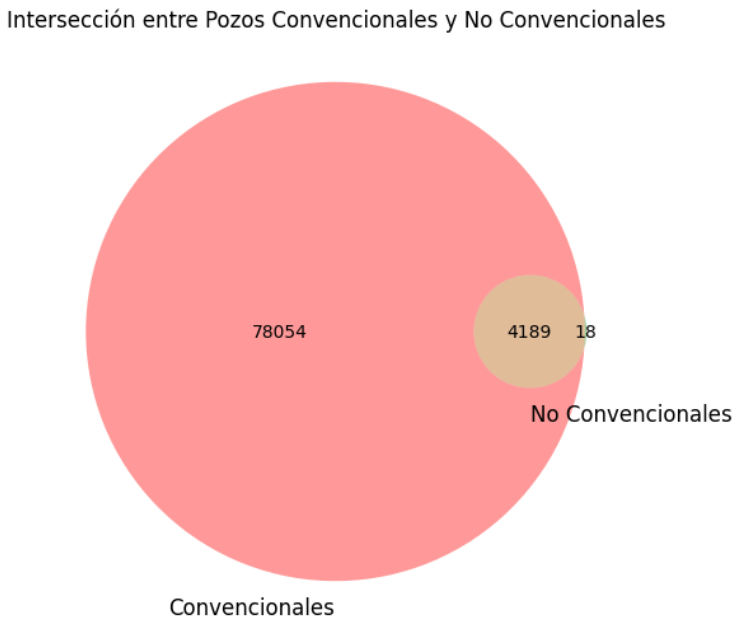

Figure 5: Intersección entre Pozos Convencionales y No Convencionales.

El gráfico muestra que **4,189 pozos** comparten características tanto de producción convencional como no convencional. Este hallazgo pone en evidencia la existencia de pozos que pueden operar bajo ambas modalidades, lo que podría responder a estrategias mixtas de extracción o a características geológicas específicas.

Validación de los Números de Producción

Al analizar los números de producción, observamos que tienen bastante sentido dentro del contexto de los datos disponibles. La producción total de petróleo convencional y no convencional es la siguiente:

- **Producción Convencional:** 33,313,653.11 litros.
- **Producción No Convencional:** 18,630,057.34 litros.

Notamos que la producción convencional es mayor, lo cual es esperable debido a la prevalencia histórica de este tipo de explotación. Sin embargo, dentro de la intersección entre pozos convencionales y no convencionales, encontramos los pozos de mayor producción, lo que explica en parte el número significativo de producción no convencional.

Es importante destacar que la suma de las producciones puede superponerse en pozos que comparten características convencionales y no convencionales. Esto podría llevar a una sobreestimación de la producción total si no se eliminan los duplicados. A continuación, presentamos un análisis que suma las producciones de pozos evitando dicha duplicación.

Validación de la Producción Total de Petróleo sin Duplicados

Al calcular la producción total de petróleo considerando únicamente los pozos únicos (es decir, eliminando duplicados entre pozos convencionales y no convencionales), obtenemos un total de **10,123,474.34 Litros**.

Este valor es significativamente menor que la suma directa de las producciones de petróleo convencional (**33,313,653.11 Litros**) y no convencional (**18,630,057.34 Litros**), que juntas dan un total de **51,943,710.45 Litros**.

Esta discrepancia pone en evidencia la importancia de eliminar duplicados al calcular la producción total, ya que algunos pozos están presentes en ambas categorías (convencionales y no convencionales). De no haber tenido en cuenta esta intersección, habríamos sobreestimado gravemente la producción total, generando un error significativo en los análisis y en la interpretación de los datos.

Por lo tanto, este análisis reafirma la necesidad de tener en cuenta las intersecciones entre categorías para evitar inconsistencias y garantizar que los resultados sean representativos de la realidad.

5.4 Análisis de Producción Convencional de Gas

En esta sección, comenzamos el análisis de la producción de gas. Primero seleccionamos las columnas que contienen información de la producción convencional y filtramos los datos correspondientes al año 2024. Este enfoque nos permitirá realizar futuras comparaciones con la producción no convencional, destacando las principales características de cada tipo de pozo.

Para analizar la producción de gas convencional en el año 2024, seleccionamos los pozos con mayor y menor producción. A continuación, se presenta una tabla que muestra los 5 pozos con mayor producción y los 5 con menor producción, junto con sus respectivas ubicaciones:

ID Pozo	Producción de Gas (Cf)	Provincia
165693	150,618.90	Estado Nacional
10464	85,016.22	Estado Nacional
10464	78,533.96	Estado Nacional
10464	73,944.48	Estado Nacional
10465	73,885.99	Estado Nacional
32186	0.00	Rio Negro
144117	0.00	Rio Negro
145614	0.00	Rio Negro
145615	0.00	Rio Negro
145626	0.00	Rio Negro

Table 7: Top 5 y Bottom 5 Pozos de Producción de Gas Convencional en 2024.

Con Estado Nacional se refiere a:

En Argentina, las minas de hidrocarburos fluidos y petróleo son bienes del dominio privado de la Nación o de las provincias, dependiendo del territorio en el que se encuentren. Esto significa que los pozos bajo la jurisdicción del "Estado Nacional" están administrados directamente por el gobierno federal, en áreas no asignadas a una provincia específica.

Fuente: Ley Nacional 12.161

Continuando con el análisis de la producción de gas convencional en 2024, la tabla presentada anteriormente muestra los pozos con mayor y menor producción junto con sus respectivas ubicaciones.

En el grupo de los pozos con mayor producción, todos están bajo la jurisdicción del **Estado Nacional**, con valores que van desde **150,618.90 Cf** hasta **73,885.99 Cf**. Esto resalta la importancia de las áreas administradas directamente por el gobierno federal en la producción de gas convencional.

Por otro lado, los pozos con menor producción (igual a 0 Cf) están ubicados en la provincia de **Río Negro**. Este fenómeno podría deberse a que estos pozos no están activos o no han reportado producción durante el periodo analizado.

Este contraste en la producción destaca la disparidad en la actividad productiva de los pozos, con una marcada concentración de la producción en pozos administrados por el Estado Nacional.

Distribución de Producción de Gas Convencional por Provincia

Hacemos la suma de la columna correspondiente para obtener el total producido en pozos convencionales de gas durante el año 2024, obteniendo un total de **42,317,680.84 Cf**.

En consecuencia, surgió la curiosidad de analizar en qué provincias se produce más gas. Para ello, agrupamos los datos por provincia y sumamos la producción total de gas en cada una. A continuación, presentamos un gráfico que muestra la distribución de la producción de gas convencional por provincia:

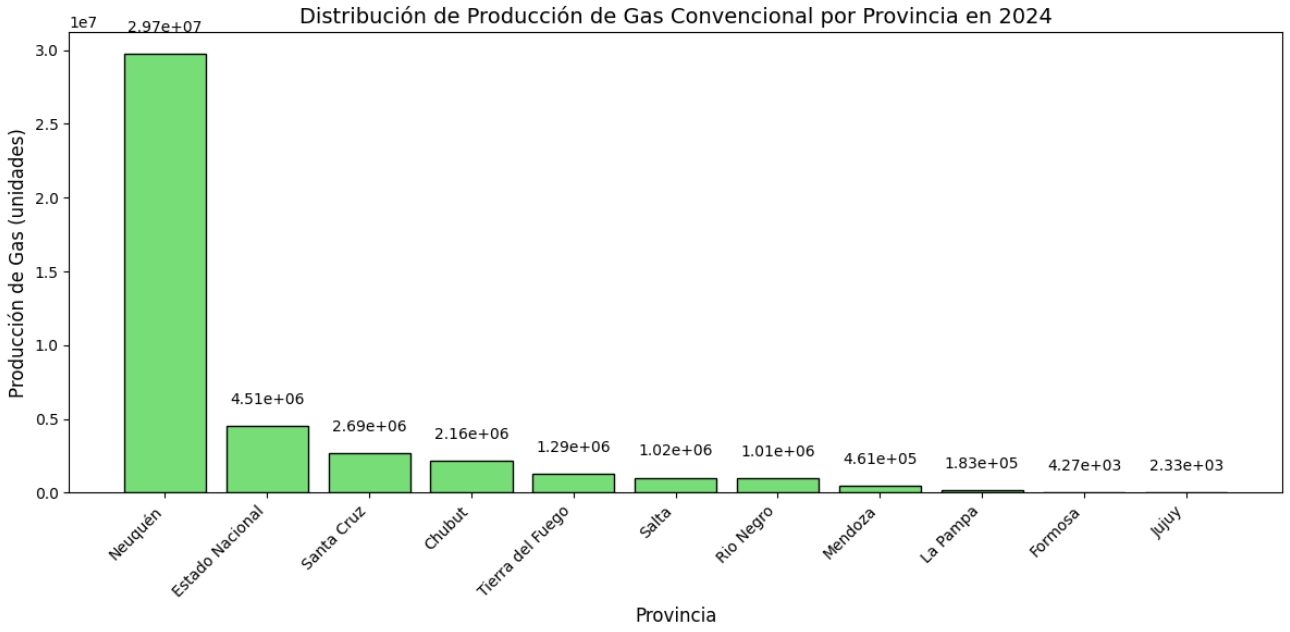

Figure 6: Distribución de Producción de Gas Convencional por Provincia en 2024

Distribución de Producción de Gas Convencional por Provincia

El gráfico presentado refleja la distribución de la producción de gas convencional por provincia en el año 2024. Se observan los siguientes puntos clave:

- **Neuquén:** Es la provincia con la producción más alta, representando una parte significativa del total nacional, con cerca de 30 millones en producción.
- **Estado Nacional:** Aunque no es una provincia, ocupa el segundo lugar en producción, con más de 4.5 millones, debido a pozos administrados directamente por el gobierno federal.
- **Otras provincias relevantes:** Santa Cruz y Chubut también tienen contribuciones importantes, con valores superiores a 2 millones de producción.
- **Provincias de menor producción:** Tierra del Fuego, Salta y Río Negro tienen contribuciones más modestas, mientras que provincias como Formosa y Jujuy muestran producciones marginales, reflejando su menor actividad en la explotación de gas convencional.

El análisis destaca que la producción de gas convencional está fuertemente concentrada en Neuquén y algunas otras regiones clave.

5.5 Análisis de Producción de Gas No Convencional

En esta subsección, analizamos la producción de gas no convencional en el año 2024. Este análisis complementa el realizado previamente para la producción convencional, permitiendo comparaciones futuras entre ambos tipos de pozos.

Comenzamos calculando el total producido en los pozos no convencionales durante 2024. De esta manera, obtenemos un panorama general de la contribución de este tipo de explotación a la producción total de gas en el país. A continuación, presentamos los resultados del total producido en 2024:

- **Total producido en pozos no convencionales:** 27,284,972.41 Cf.

Para analizar la producción de gas no convencional en el año 2024, seleccionamos los pozos con mayor y menor producción. A continuación, se presenta una tabla que muestra los 5 pozos con mayor producción y los 5 con menor producción, junto con sus respectivas ubicaciones:

ID Pozo	Producción de Gas (Cf)	Provincia
165073	27,460.550	Neuquén
165073	25,451.760	Neuquén
164006	22,909.539	Neuquén
164006	22,508.177	Neuquén
165073	22,493.020	Neuquén
159688	0.000	Rio Negro
156873	0.000	Neuquén
155540	0.000	Neuquén
155525	0.000	Neuquén
155537	0.000	Neuquén

Table 8: Top 5 y Bottom 5 Pozos de Producción de Gas No Convencional en 2024.

La tabla presentada muestra los 5 pozos con mayor producción de gas no convencional y los 5 con menor producción en el año 2024. Podemos observar lo siguiente:

- **Top 5 Pozos:** Todos los pozos con mayor producción de gas no convencional están ubicados en la provincia de Neuquén, con valores que oscilan entre 22,493.02 Cf y 27,460.55 Cf. Esto resalta el papel predominante de Neuquén en la explotación de gas no convencional.
- **Bottom 5 Pozos:** Los pozos con producción igual a 0 están distribuidos entre Neuquén y Río Negro. Esto podría reflejar pozos inactivos o que no han reportado producción durante el período analizado.

Dado este panorama, surge la curiosidad de analizar la distribución geográfica de los pozos de gas no convencional en Argentina. A continuación, se presenta un Pie Chart:

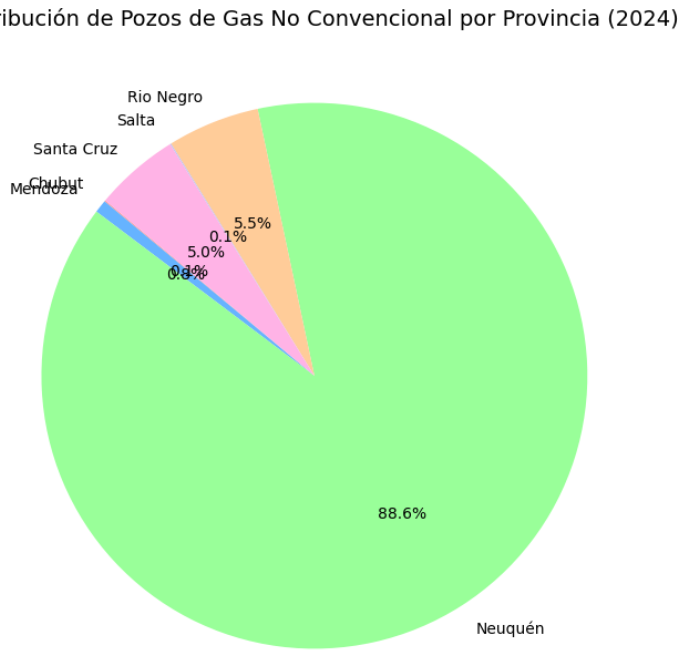

Figure 7: Distribución de Pozos de Gas No Convencional por Provincia (2024)

Siguiendo con el análisis

Siguiendo con el análisis, quisimos ver nuevamente si existe una intersección entre los pozos de gas convencional y no convencional. Este enfoque nos permite identificar si algunos pozos están siendo explotados bajo ambas categorías, lo que podría tener implicancias importantes en la interpretación de los datos de producción.

- **Cantidad de pozos únicos de gas convencional:** 21,696.
- **Cantidad de pozos únicos de gas no convencional:** 3,507.
- **Cantidad de pozos que son tanto convencionales como no convencionales:** 3,507.

Estos resultados indican que **todos los pozos no convencionales también están categorizados como convencionales**. Esto sugiere que, similar al caso del petróleo, existe una superposición completa entre estas categorías en lo que respecta a los pozos de gas.

Implicaciones:

- Al sumar la producción total de gas, es crucial evitar contar dos veces la producción de estos pozos compartidos. De lo contrario, los valores totales reportados serían incorrectos y podrían sobreestimar la producción real.
- Esta situación refuerza la necesidad de realizar un manejo cuidadoso de los datos para garantizar que las métricas reportadas reflejen la realidad de la explotación de gas en el país.

5.6 Producción Convencional vs. Producción No Convencional

En esta sección, comparamos la producción de gas y petróleo convencional con la producción no convencional en el año 2024. El objetivo es evaluar las contribuciones de ambos tipos de explotación al total producido, identificar patrones en su distribución y analizar la importancia relativa de cada uno.

Análisis Gráfico de la Producción Convencional y No Convencional

Para comenzar esta sección, presentamos un análisis gráfico que compara la producción convencional y no convencional de petróleo y gas en el año 2024. Este gráfico permite visualizar de forma clara las diferencias y similitudes en los volúmenes producidos, destacando las contribuciones de cada modalidad a la producción total de hidrocarburos.

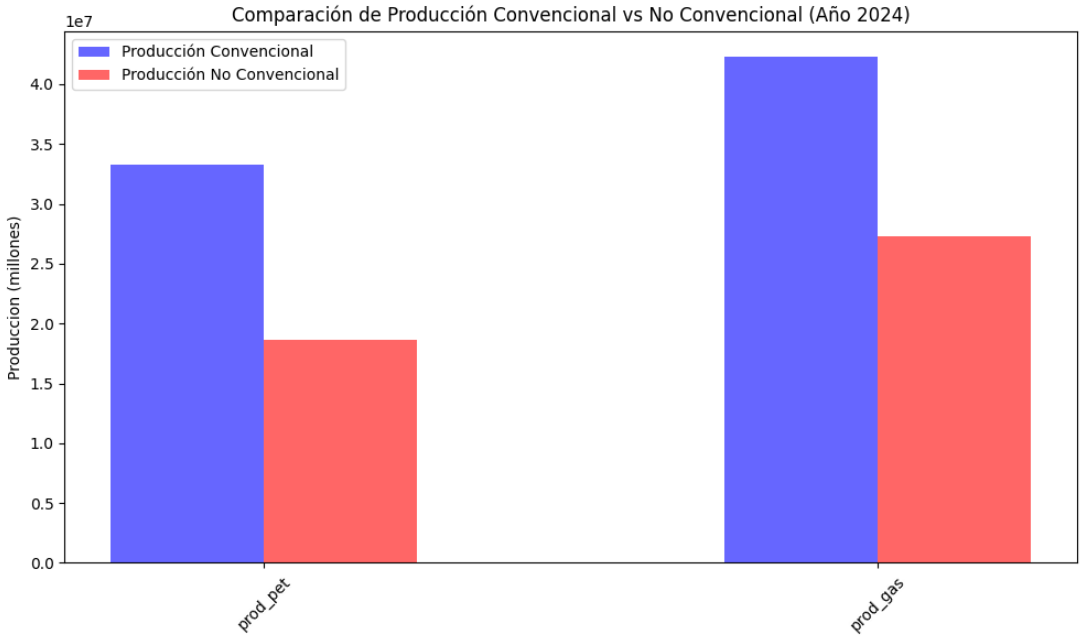

Figure 8: Producción Convencional vs. No Convencional de Petróleo y Gas en 2024.

En el gráfico, observamos que la producción convencional supera ampliamente a la no convencional. Esto se debe, en gran parte, a la predominancia de pozos convencionales en comparación con los pozos no convencionales, los cuales representan una proporción menor dentro de la matriz productiva.

Producción de Gas por Provincia

A continuación, presentamos un gráfico que muestra la producción de gas por provincia, reflejando lo discutido en las secciones anteriores. Este análisis destaca las regiones con mayor actividad en la explotación de gas convencional y no convencional, confirmando las diferencias geográficas y de volumen previamente mencionadas.

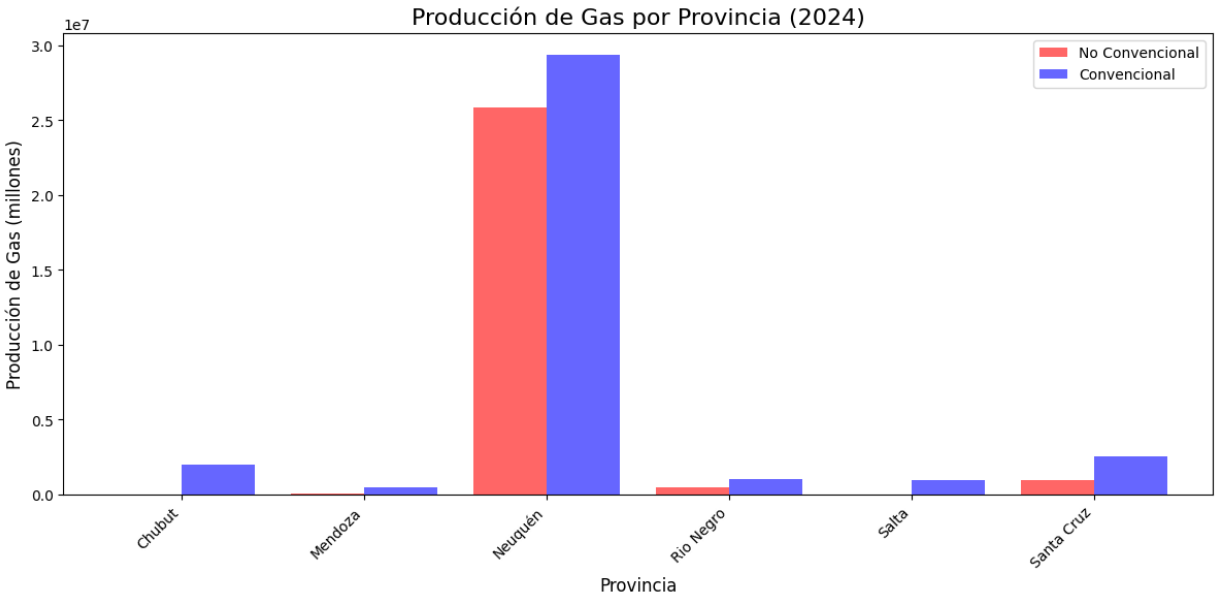

Figure 9: Producción de Gas por Provincia en 2024.

Producción de Petróleo por Provincia

Al igual que con el gráfico anterior, el siguiente gráfico muestra la producción de petróleo por provincia en el año 2024. Este análisis complementa el de gas, destacando las regiones con mayor actividad en la explotación de petróleo convencional y no convencional, tal como se discutió en las secciones previas.

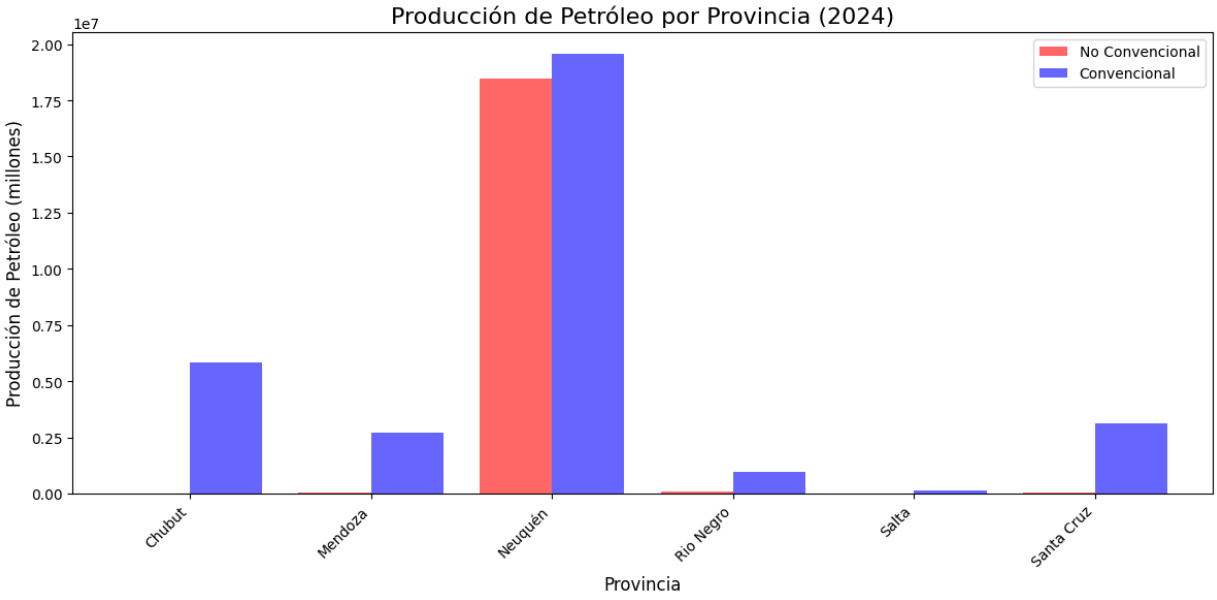

Figure 10: Producción de Petróleo por Provincia en 2024.

6. Verifiquemos la existencia de Casos Anómalos

En esta sección, analizamos la presencia de posibles casos anómalos en los datos de producción de petróleo y gas. Los casos anómalos pueden representar errores en los datos, problemas en la medición o eventos atípicos que merecen ser investigados más a fondo. Detectar y manejar estas anomalías es crucial para garantizar la calidad y fiabilidad de los análisis realizados.

6.1 Métodos Propuestos para Detectar Anomalías

Detección de Anomalías con K-Means

En esta sección, presentamos un enfoque basado en la distancia al centroide utilizando el algoritmo de *K-Means* para la detección de anomalías en los datos de producción de petróleo (`prod_pet`) en el año 2024. Este método nos permite identificar registros cuya distancia al centroide de su clúster excede un umbral predefinido, en este caso, el percentil 95 de las distancias.

- **Metodología:**
  - Los datos se agrupan en 3 clústeres utilizando *K-Means*.
  - Se calcula la distancia de cada punto al centroide de su clúster asignado.
  - Se define como anomalías aquellos registros cuya distancia supera el percentil 95 de todas las distancias.
- **Resultados:**
  - Se identificaron 40,327 registros como anómalos, con valores de producción que oscilan entre 125.76 y 19,826.74 Cf.
  - El gráfico muestra la distribución de las distancias al centroide, donde los puntos en rojo representan los registros considerados como anómalos.

A continuación, se presenta el gráfico que ilustra este análisis:

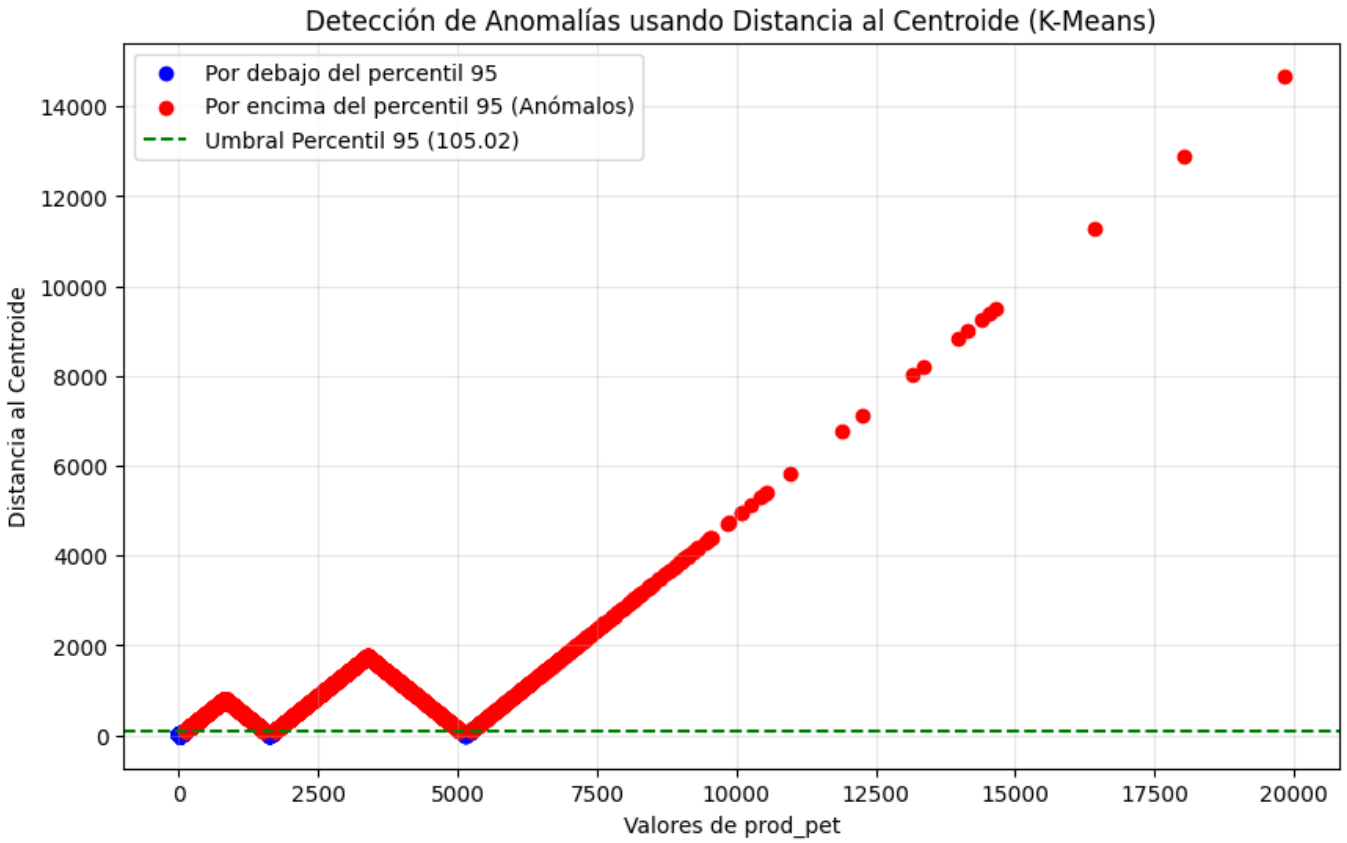

Figure 11: Detección de Anomalías usando Distancia al Centroide (K-Means).

Este método es particularmente útil para detectar valores atípicos en un conjunto de datos con múltiples clústeres, especialmente en dominios donde las distribuciones pueden ser no lineales. Sin embargo, es importante combinar este enfoque con otros métodos para validar las anomalías detectadas y comprender su impacto en el análisis global.

Detección de Anomalías con Isolation Forest

Para la detección de anomalías en la producción de petróleo y gas, empleamos el algoritmo **Isolation Forest**, el cual es particularmente efectivo en la identificación de valores atípicos en grandes conjuntos de datos. Este método se basa en la idea de que los puntos anómalos son fácilmente "aislables" en un conjunto de datos, mientras que los puntos normales requieren un mayor número de particiones para ser aislados.

- **Paso 1: Cálculo del cambio porcentual**: En primer lugar, calculamos el cambio porcentual en la producción de cada pozo de gas entre el mes actual y el promedio de los tres meses anteriores. Este valor nos permite identificar incrementos o disminuciones anómalas en la producción de cada pozo. La fórmula utilizada para este cálculo es la siguiente:

$$CambioPorcentual = \frac{ProducciónActual - PromedioDeLosTresMesesAnteriores}{PromedioDeLosTresMesesAnteriores} \times 100$$

- **Paso 2: Aplicación de Isolation Forest**: Utilizando los valores de cambio porcentual obtenidos en el paso anterior, aplicamos el algoritmo para identificar los pozos que presentan un comportamiento inusual, es decir, aquellos cuyo cambio porcentual es significativamente diferente de los demás. Este algoritmo etiqueta a los puntos anómalos con un valor de -1, mientras que los puntos normales reciben un valor de 1.

- **Paso 3: Ajuste de la tasa de anomalías**: Para obtener un número más controlado de anomalías, ajustamos el parámetro de "contaminación" del modelo, que representa el porcentaje de puntos que esperamos sean anómalos. En este caso, establecimos este parámetro en un valor bajo, lo que permite reducir el número de casos anómalos detectados.

Veamos el grafico resultante:

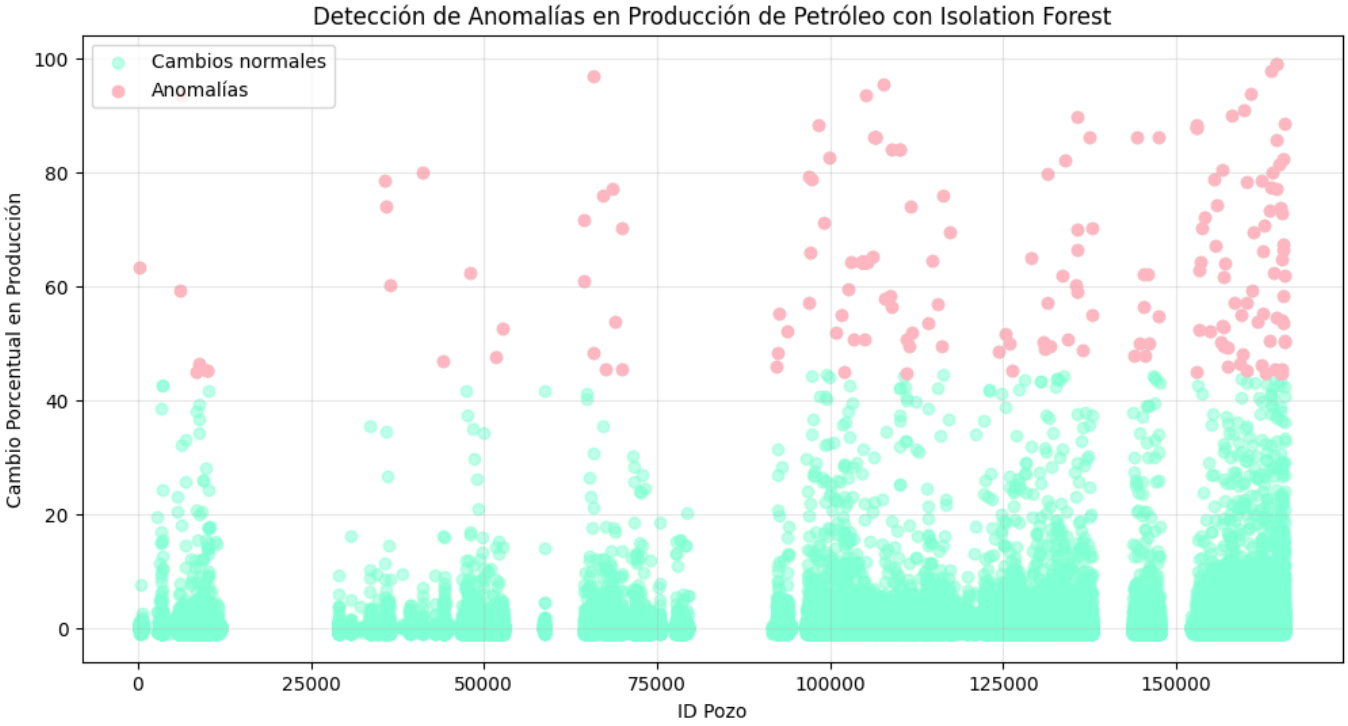

Figure 12: Enter Caption

Como pueden ver, los resultados del modelo fueron visualizados en un Scatter Plot, donde cada punto representa un pozo y su respectiva producción. Los puntos que fueron etiquetados como anómalos por el modelo fueron resaltados en rojo, mientras que los puntos normales se mostraron en verde. Este gráfico facilita la identificación visual de los pozos con producción inusualmente alta o baja.

6.2 Factores Técnicos que Influyen en la Productividad de los Pozos

Para un mayor entendimiento del caso de uso, consultamos la fuente complementaria *"Problemas de productividad en pozos petroleros"* publicada por **NRGI Broker** [1], la cual detalla las principales causas que pueden afectar la productividad en pozos de petróleo y gas. Según este recurso, problemas como la presión del yacimiento, la permeabilidad de las rocas, las condiciones ambientales y las técnicas de extracción pueden influir significativamente en el comportamiento de los pozos. Este conocimiento es clave para interpretar las posibles anomalías detectadas en nuestros datos.

Basándonos en esta información, realizamos un análisis de los datos disponibles en nuestros conjuntos de datos para identificar campos que puedan estar relacionados con estas problemáticas. A continuación, describimos los campos más relevantes y su posible vínculo con las anomalías en la producción:

- **Presión y características del pozo:**
  - `profundidad`: Este campo puede estar relacionado con problemas técnicos como cambios en las condiciones del yacimiento.
  - `tipoestado`: Indica si el pozo está operativo, en mantenimiento o cerrado, lo cual puede reflejar problemas técnicos o de productividad.
  - `tipoextraccion`: Permite analizar el método de extracción utilizado , lo que puede indicar dificultades relacionadas con la presión o la permeabilidad.
- **Condiciones ambientales y operativas:**
  - `provincia` y `cuenca`: La ubicación geográfica del pozo puede influir en su comportamiento debido a condiciones como la composición de los fluidos o la corrosión.
  - `adjiv_fecha_inicio_perf` y `adjiv_fecha_fin_perf`: Estas fechas permiten analizar si el pozo fue recientemente perforado o si lleva tiempo operativo, lo que podría estar relacionado con su productividad.
  - `tipopozo`: Diferencia entre pozos petrolíferos, gasíferos o de inyección, lo que puede reflejar la estrategia de recuperación aplicada.
- **Producción y rendimiento:**
  - `prod_pet` y `prod_gas`: Representan los valores de producción de petróleo y gas respectivamente, y son esenciales para identificar comportamientos atípicos.
  - `vida_util`: Proporciona una estimación del tiempo que un pozo puede permanecer productivo, siendo útil para analizar la longevidad de los pozos en relación con su producción.
- **Condiciones específicas del yacimiento:**
  - `formacion` y `clasificacion`: Están vinculados a las características geológicas del yacimiento, como la permeabilidad de las rocas, que puede afectar la facilidad para extraer hidrocarburos.

Este análisis nos permitió identificar qué campos del dataset podrían estar vinculados a problemáticas técnicas y operativas, proporcionando un marco más completo para interpretar las anomalías detectadas.

6.3 Distribución de Problemas en los Pozos

Nos interesamos por entender los diferentes tipos de problemas técnicos y operativos que pueden estar presentes en los pozos. Para ello, analizamos la columna `tipoestado`, que clasifica el estado operativo de los pozos. A continuación, presentamos una tabla con la distribución de los diferentes tipos de problemas:

[1] `https://nrgibroker.com/problemas-de-productividad-en-pozos-petroleros/?utm_source`

Tipo de Estado Operativo	Cantidad de Pozos
Extracción Efectiva	26,908
Abandonado	19,973
En Estudio	8,354
En Reserva para Recuperación Secundaria/Asistida	8,069
En Inyección Efectiva	6,195
Parado Transitoriamente	5,571
A Abandonar	2,622
En Espera de Reparación	1,834
Abandono Temporario	1,407
Otras Situación Inactivo	1,385
Parado Alta Relación Agua/Petróleo	1,006
En Reserva de Gas	565
Otras Situación Activo	213
Parado Alta Relación Gas/Petróleo	104
En Reparación	43
No informado	43
Mantenimiento de Presión	40

Table 9: Distribución de los Tipos de Problemas en los Pozos por Estado Operativo.

A continuación, se presentan dos gráficos que analizan los problemas por tipo de pozo y por provincia. Estos gráficos complementan la tabla y permiten visualizar de manera clara cómo se distribuyen los problemas en diferentes categorías y ubicaciones geográficas.

Figure 13: Problemas por Tipo de Pozo.

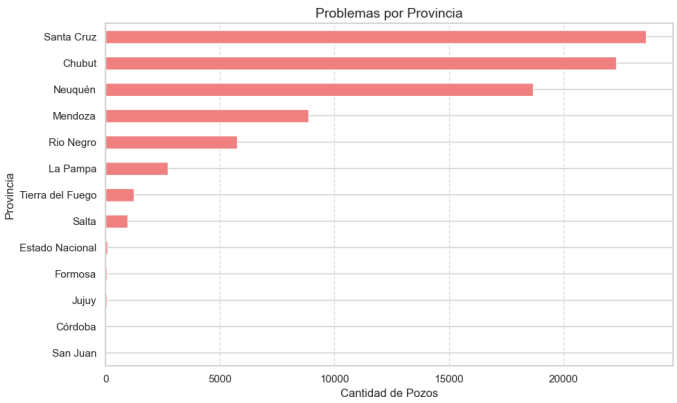

Figure 14: Problemas por Provincia.

6.3.1 Análisis de Problemas por Provincia

La figura 13 muestra la distribución de problemas técnicos y operativos según la provincia. Se observa que:

- Las provincias con mayor cantidad de pozos problemáticos son **Santa Cruz** y **Chubut**, seguidas de cerca por **Neuquén**.
- Las provincias **San Juan** y **Córdoba** presentan la menor cantidad de problemas, con solo seis pozos cada una.
- Esta distribución resalta la concentración de actividad petrolera y gasífera en las principales cuencas productoras del país.

6.3.2 Análisis de Problemas por Tipo de Pozo

La figura 12 muestra la cantidad de problemas según el tipo de pozo. Podemos destacar que:

- Los pozos **petrolíferos** concentran la mayor cantidad de problemas, seguidos por los clasificados como **otro tipo**.

- Los pozos **gasíferos** y de **inyección de agua** presentan menos problemas en comparación, reflejando diferencias en su operación y mantenimiento.
- Los pozos de almacenamiento (bidireccional y de monitoreo) tienen una cantidad mínima de problemas, posiblemente debido a su menor cantidad en el conjunto de datos.

Estos análisis destacan la necesidad de priorizar la gestión de problemas en las provincias con mayor actividad productiva y en los pozos petrolíferos, que representan la mayor proporción de problemas técnicos.

6.4 Conclusiones Generales de las Anomalías detectadas

El análisis de los datasets utilizados reveló varias anomalías que afectan la integridad, consistencia y calidad de los datos. Estas anomalías se clasifican en diferentes categorías, según su naturaleza, y se describen a continuación:

- **Duplicación de Pozos en Producción:** Se detectaron pozos que aparecen tanto en el dataset de producción convencional como en el de producción no convencional. Este hallazgo es incoherente con las clasificaciones mutuamente excluyentes de los pozos, ya que un pozo no puede ser simultáneamente convencional y no convencional. Este problema podría reflejar inconsistencias en la asignación de categorías o errores en el registro.
- **Inconsistencias Temporales:** Algunos registros presentan fechas de inicio de operación que son posteriores a las fechas de finalización. Esto representa un error lógico en los datos, que podría deberse a errores de digitación o registros incompletos. Este tipo de anomalías afecta la posibilidad de realizar análisis temporales confiables.
- **Datos Faltantes:** Se identificaron valores nulos en varias columnas críticas, como `tipoestado`, `tipoextraccion` y `provincia`. Estas ausencias dificultan el análisis y la interpretación de los datos, ya que eliminan información clave necesaria para comprender el comportamiento de los pozos.
- **Producción Negativa:** Se encontraron valores negativos en las columnas relacionadas con la producción, como `prod_pet` y `prod_gas`. Esto no es físicamente posible y puede deberse a errores en la entrada de datos. Estas anomalías requerirán una limpieza y corrección cuidadosa antes de proceder con análisis más avanzados.
- **Pozos Sin Información Relacionada:** En algunos casos, los pozos que aparecen en los datasets de producción no tienen una correspondencia en el dataset maestro de pozos. Esto indica omisiones o errores en los registros maestros, lo que complica la trazabilidad y validación de los datos.

Como se mencionó anteriormente, se utilizaron dos métodos complementarios para detectar anomalías en los datos de producción:

- **Método de K-medias:** Este algoritmo agrupó los registros en clusters y marcó como anómalos aquellos cuya producción de petróleo se encontraba lejos del centroide del cluster más cercano. Los registros cuya distancia superaba el percentil 95 fueron considerados anómalos, identificando **40,327 registros** con patrones de producción atípica, tanto altos como bajos.
- **Algoritmo Isolation Forest:** Para complementar el análisis, se aplicó este modelo no supervisado que detecta outliers al medir el nivel de aislamiento de cada registro. Este método permitió identificar anomalías adicionales y corroborar patrones inusuales en los datos, específicamente en ciertas regiones y pozos.

La combinación de ambos enfoques fortaleció el análisis de anomalías al abordar diferentes perspectivas: una basada en la distancia dentro de clusters y otra enfocada en el nivel de aislamiento de los datos.

7. Reglas de Validación

En esta sección, presentamos un conjunto de reglas de validación que aplicamos para garantizar un análisis efectivo y coherente de los datos. Estas reglas fueron clave para detectar errores, inconsistencias y anomalías, asegurando la calidad del análisis. Además, recomendamos seguir estos pasos en futuros estudios para evitar problemas similares.

7.1 Reglas de Validación Propuestas

1. **Filtrar datos de producción:** Es importante revisar las columnas de producción (`prod_pet`, `prod_gas`, etc.) y eliminar cualquier valor negativo, ya que estos resultados serían físicamente imposibles y afectarían el análisis.
2. **Consistencia en los pozos y su identificación:** Un mismo pozo no debería clasificarse simultáneamente como convencional y no convencional. Este tipo de inconsistencias afecta gravemente los resultados y genera datos erróneos.

3. **Verificación de casos anómalos:** Recomendamos establecer un umbral, como el percentil 95, para identificar posibles anomalías. Los valores fuera de este rango deben analizarse cuidadosamente para determinar si corresponden a errores en los datos o si son características reales del dataset.

4. **Ubicación geográfica válida:** Todos los pozos deben estar asociados a provincias, formaciones geológicas, coordenadas y demás datos geográficos que sean reales y verificables dentro de Argentina. La falta de información precisa en este aspecto puede comprometer la utilidad de los datos.

5. **Distribución temporal de la producción:** La producción de petróleo, gas y agua debe mostrar una distribución consistente a lo largo del tiempo. Si se identifican períodos de producción nula o inconsistencias temporales, es necesario investigar estos casos y validar los datos afectados.

Estas reglas de validación no solo aseguran la calidad del análisis, sino que también previenen problemas que podrían surgir en estudios posteriores. Su implementación contribuye a una mayor confiabilidad y precisión en los resultados obtenidos.

8. Conclusiones Generales

En este trabajo se abordó un análisis integral de los datos de producción de petróleo y gas en Argentina, considerando tanto los pozos convencionales como los no convencionales. Se evaluaron la calidad y consistencia de los datasets, así como la detección de casos anómalos mediante técnicas estadísticas y de aprendizaje automático. A lo largo del análisis se destacaron las principales características y problemáticas asociadas a los datos, así como las limitaciones inherentes a su manejo.

Entre los hallazgos más importantes se destacan las inconsistencias en las clasificaciones de pozos, los valores negativos en las producciones reportadas y la duplicación de registros entre categorías mutuamente excluyentes. Estas anomalías resaltan la importancia de implementar reglas de validación robustas y procesos de limpieza de datos antes de llevar a cabo cualquier análisis.

Por último, este estudio no solo permitió identificar patrones relevantes en la distribución geográfica y temporal de la producción, sino que también propuso estrategias prácticas para la detección de errores y para la mejora de la calidad de los datos. Este enfoque puede servir como base para análisis futuros, asegurando la integridad y confiabilidad de los resultados en un sector estratégico para el país.