**Actividad Guiada en IBM Watson**

**🎯** Proyecto: Predicción de Reclamaciones Fraudulentas en Seguros de Automóviles

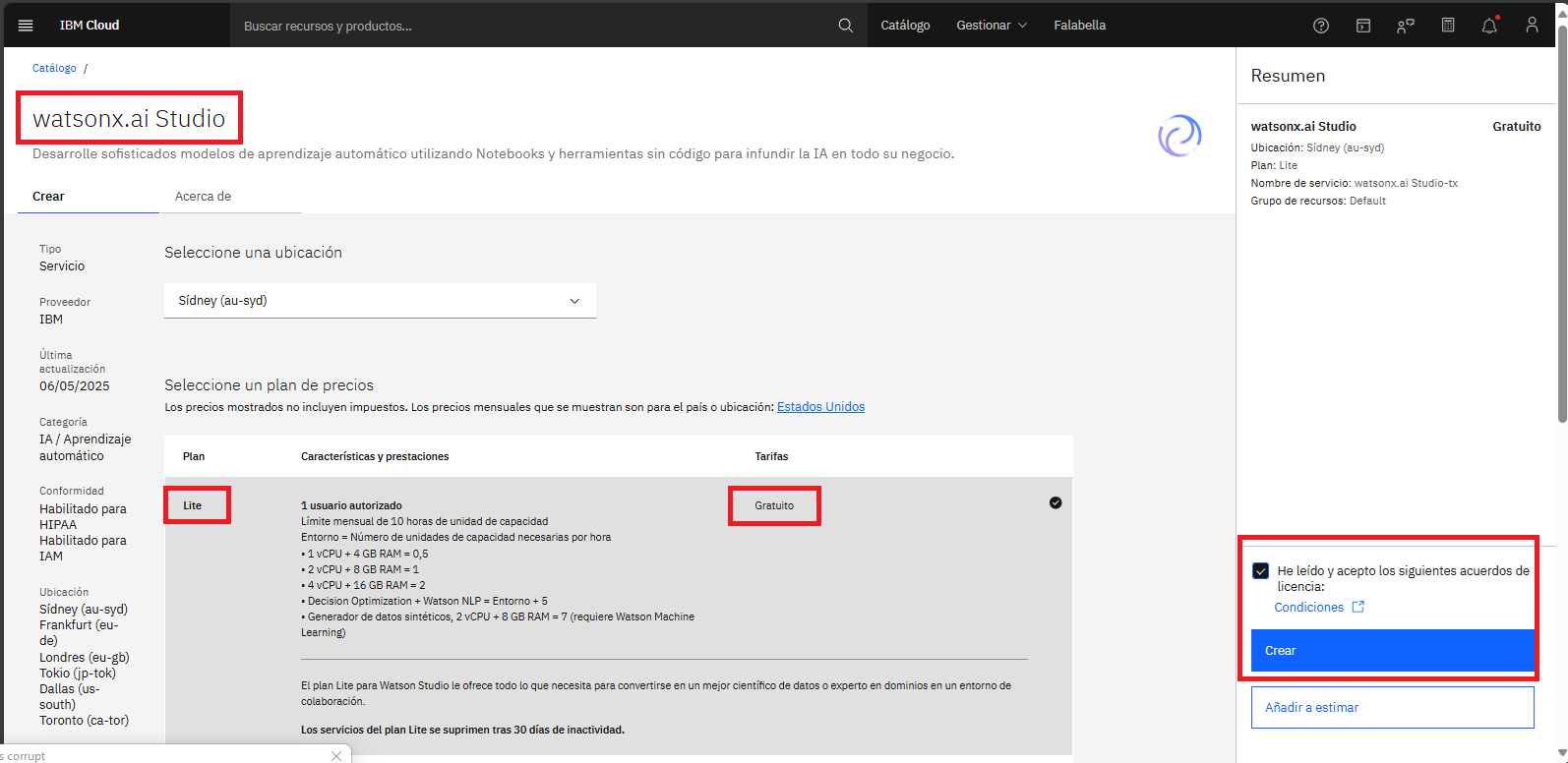
Objetivo del Proyecto:  
Utilizar IBM Watson Studio para realizar un análisis end-to-end de datos de reclamaciones de seguros, desde la importación hasta la visualización e interpretación, con el fin de validar una hipótesis de negocio y generar insights accionables para detectar fraude.  
Hipótesis de Negocio: *“Las reclamaciones superiores a $10,000 pueden ser fraudulentas.”*

**✅** Sesión Única Integrada: Guía Paso a Paso

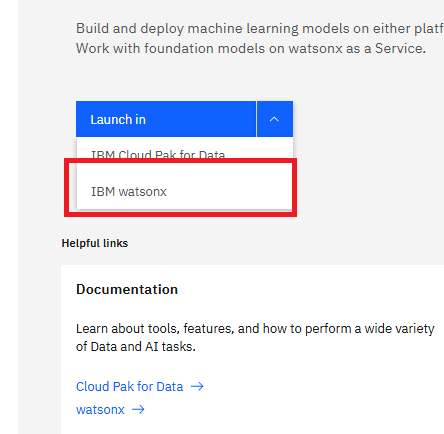
Paso 1: Provisión del Servicio IBM Watson Studio

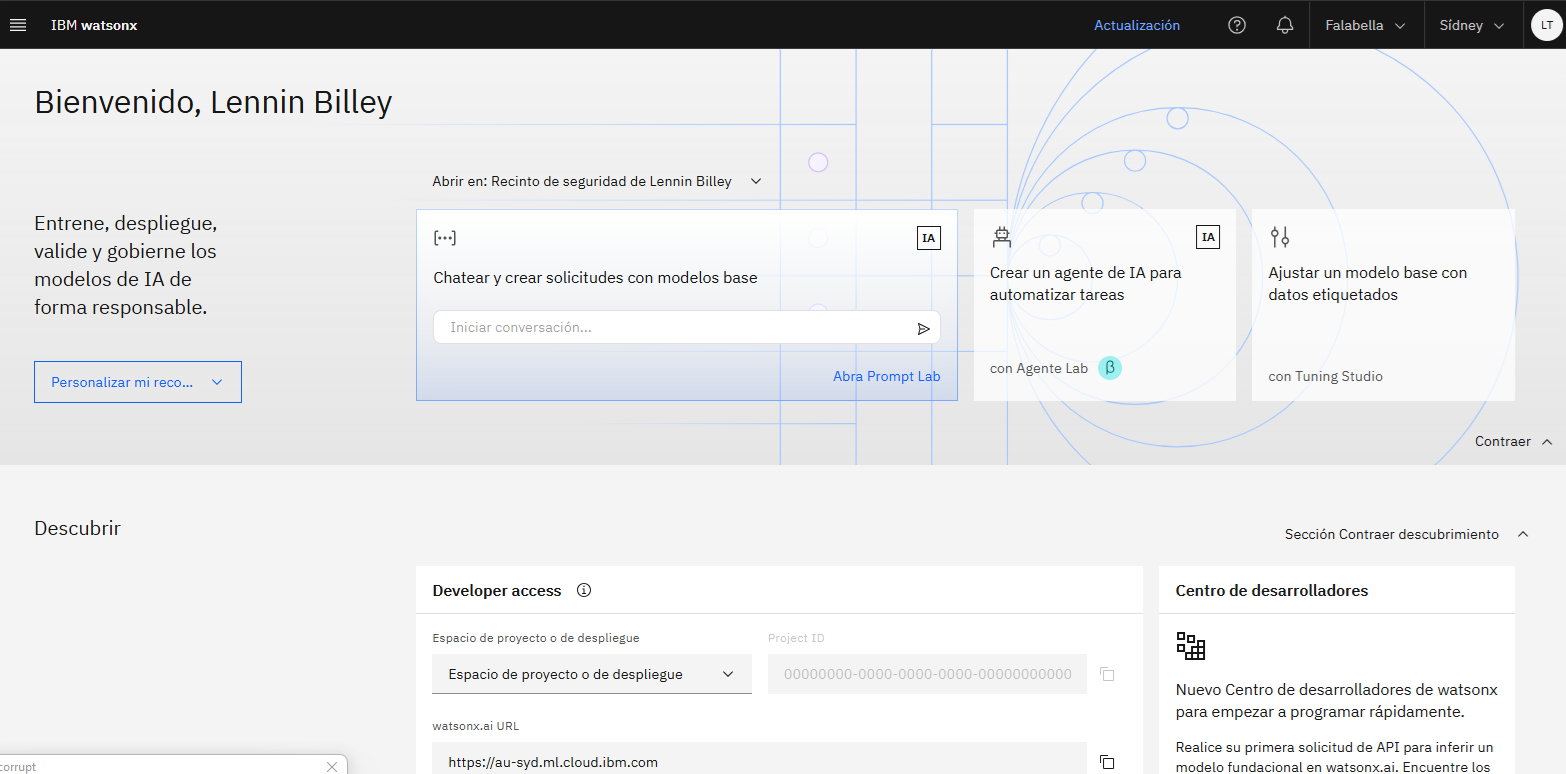
*Objetivo:* Configurar el entorno de trabajo en la nube para realizar el análisis de datos.

1. Inicia sesión en tu cuenta de IBM Cloud ([https://cloud.ibm.com](https://cloud.ibm.com/) ).
2. Navegar al Catálogo de Servicios (Service Catalog).
3. Busca IBM Watson Studio y selecciónalo.
4. Elegir el plan Lite (gratis) — es suficiente para este ejercicio.
5. Haz clic en Crear para aprovisionar el servicio.



1. Una vez creado, haz clic en Lanzar Watson Studio para acceder al entorno de trabajo.



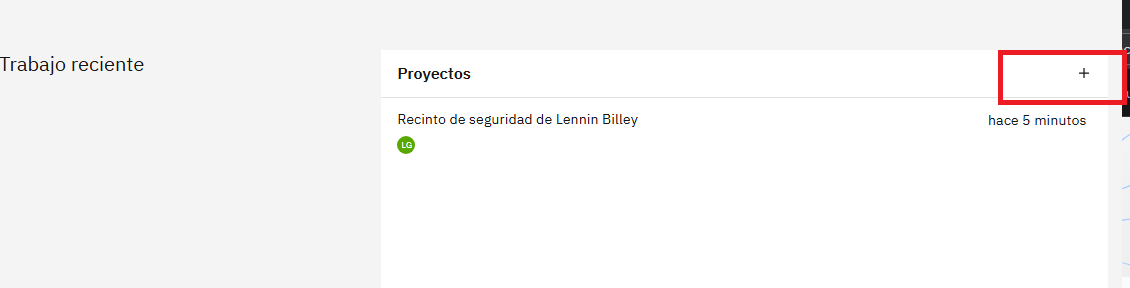


✅ *Nota del Instructor:* Watson Studio es una plataforma integrada que permite gestionar datos, limpiarlos, construir modelos y visualizar resultados en un solo entorno. Esto elimina la necesidad de cambiar entre herramientas, lo que reduce errores y mejora la eficiencia.

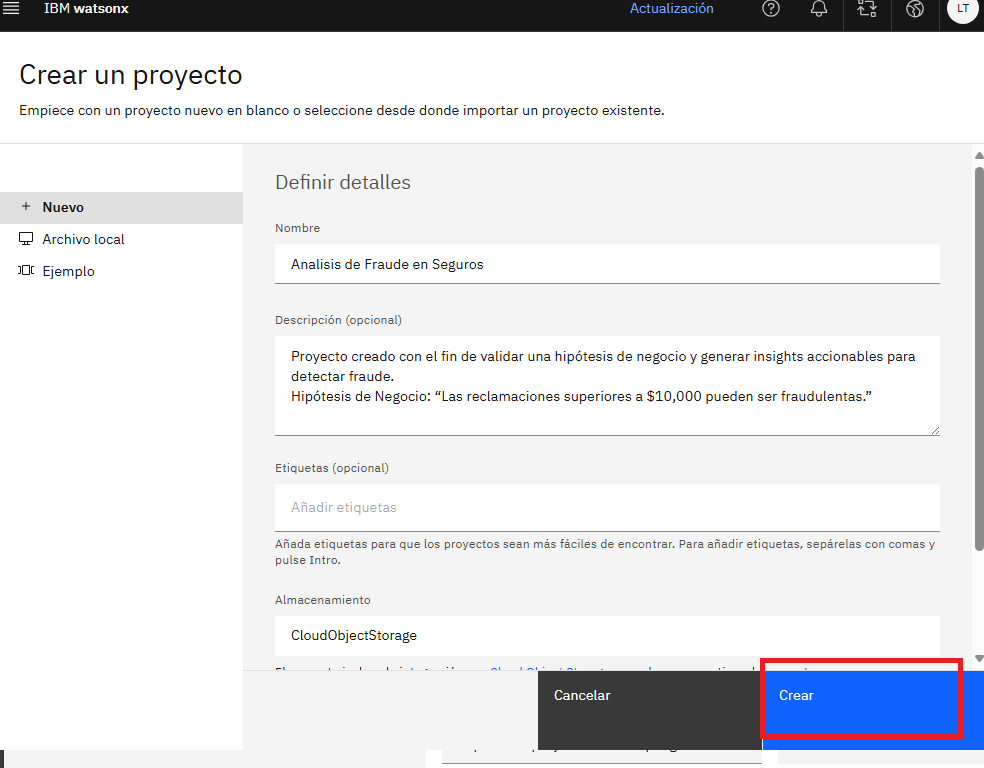
Paso 2: Crear un Nuevo Proyecto de Análisis

*Objetivo:* Organizar todos los Activos del proyecto en un espacio de trabajo estructurado.

1. En la página de inicio de Watson Studio, haz clic en Crear proyecto o Nuevo proyecto.



1. Selecciona el tipo de proyecto: Proyecto de análisis (*Analytics Project*).
2. Asigna un nombre claro: “Análisis de Fraude en Seguros”.
3. Selecciona un servicio de almacenamiento existente (por defecto IBM Cloud Object Storage) o crea uno nuevo. Este es donde se guardarán tus archivos CSV, visualizaciones y modelos.
4. Haz clic en Crear.

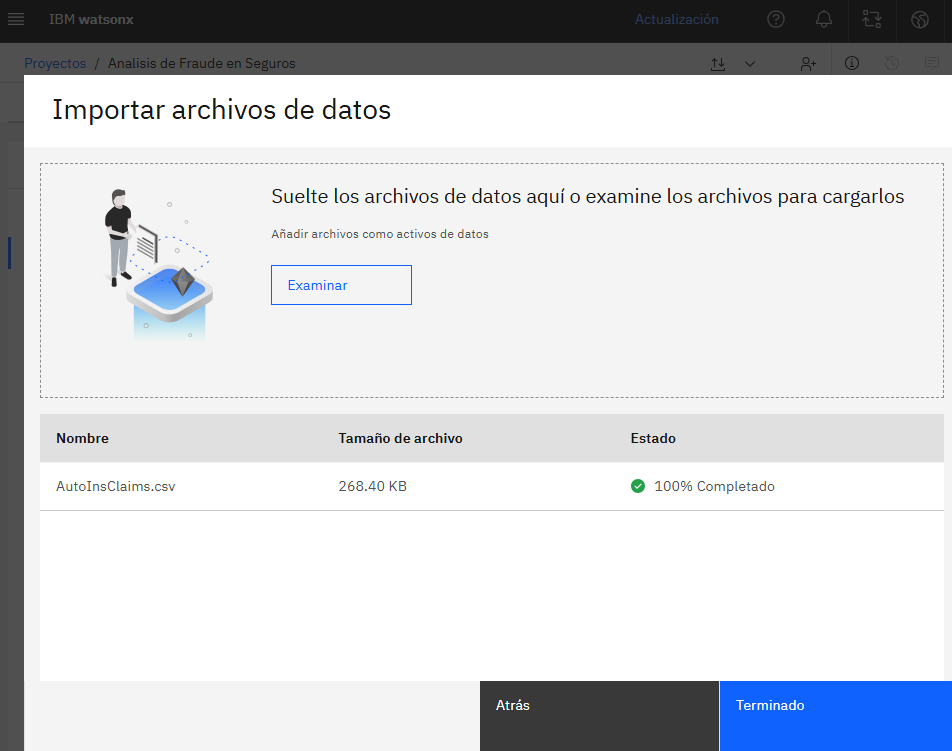


✅ *Nota del Instructor:* Un proyecto bien nombrado y estructurado facilita la colaboración y la reutilización futura. Este proyecto contendrá: datos, notebooks, flujos de ETL y visualizaciones.

Paso 3: Importar el Conjunto de Datos

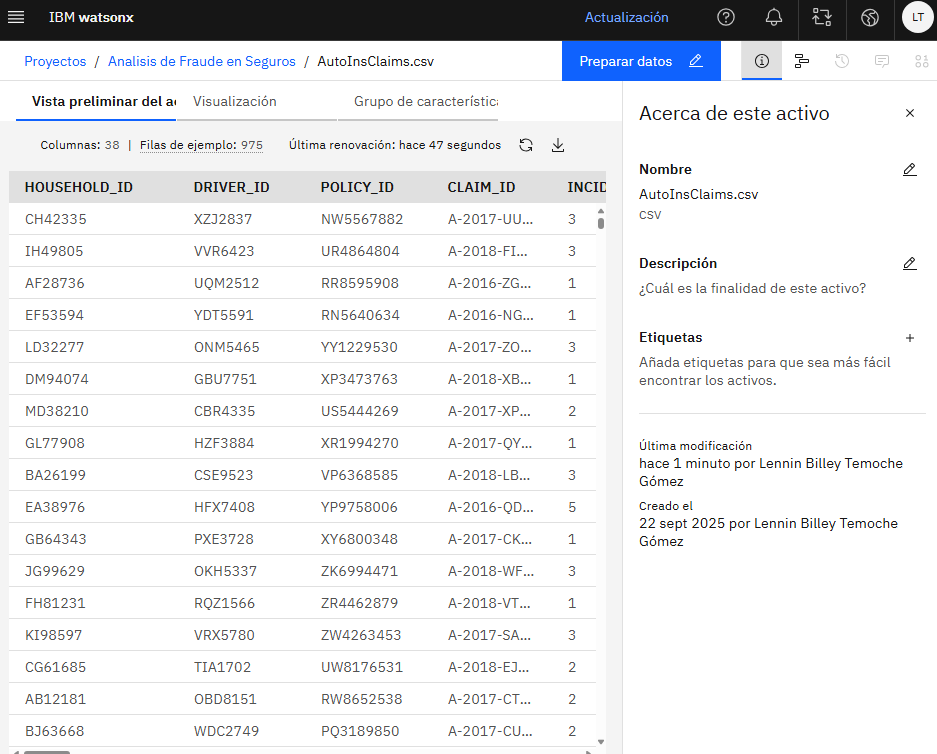
*Objetivo:* Cargar los datos de reclamaciones de seguros para su análisis.

1. Dentro de tu proyecto, ve a la pestaña Activos (*Assets*).
2. Haz clic en Importar activos.
3. Dentro de Importar archivos, haz clic en Archivo local, luego en Activo de datos.
4. Luego, Arrastra y suelta el archivo CSV del ejercicio (archivo “AutoInsClaim.csv”).



*Si no tienes el archivo original, busca un conjunto de datos público similar en Kaggle (ej. “Auto Insurance Claim Fraud Dataset”) para practicar.*

1. Una vez cargado, verás el archivo en la lista de activos. Asegúrate de que el nombre sea legible y correcto.
2. Haz clic en el nombre del archivo para ver información acerca de este activo.

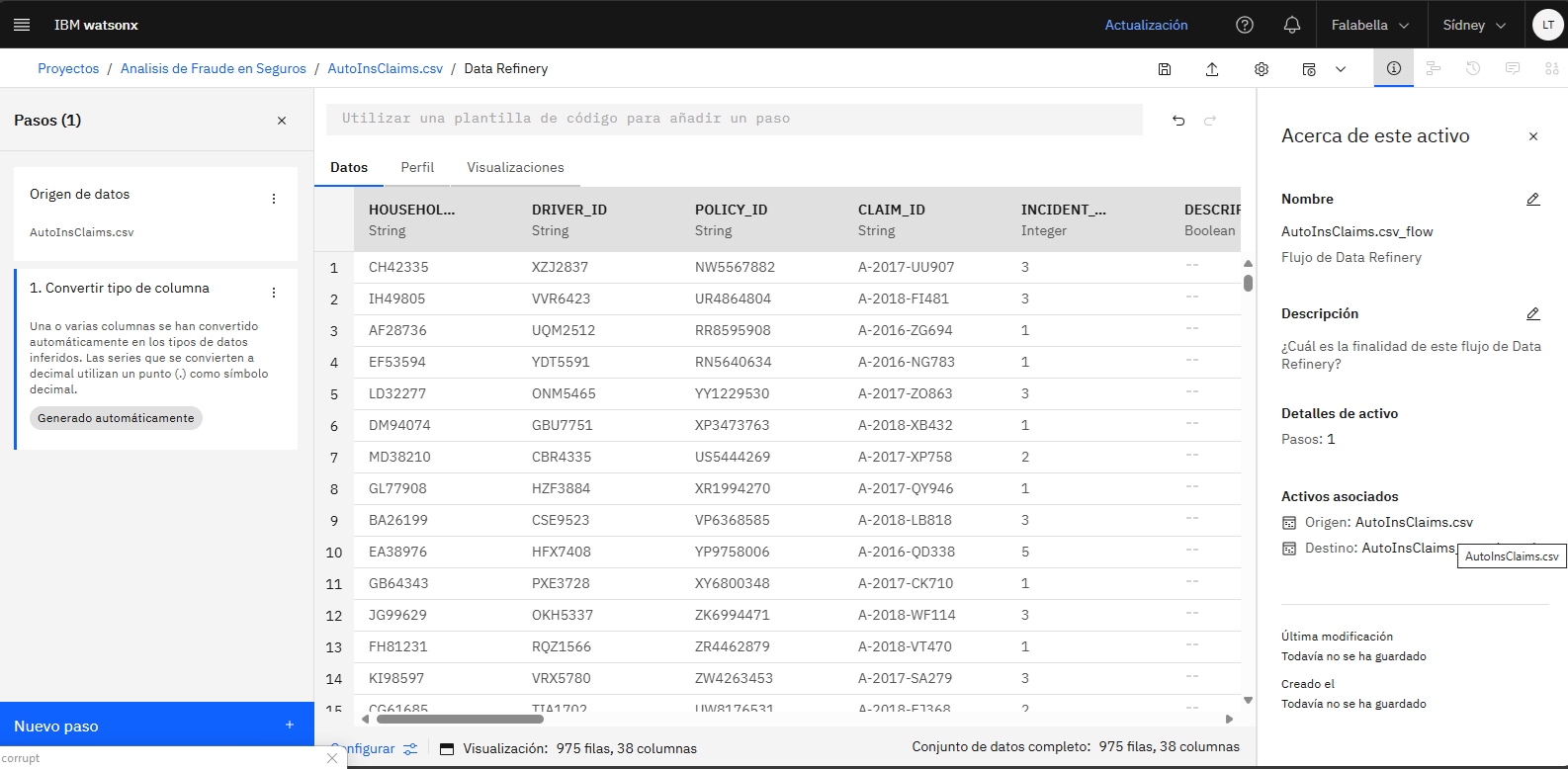


✅ *Nota del Instructor:* El conjunto de datos original contiene 38 columnas. Algunas son identificadores, otras están vacías o son redundantes. La limpieza será fundamental.

Paso 4: Limpiar y Preparar los Datos con IBM Data Refinery

*Objetivo:* Transformar los datos brutos en un conjunto limpio, confiable y listo para análisis predictivo.

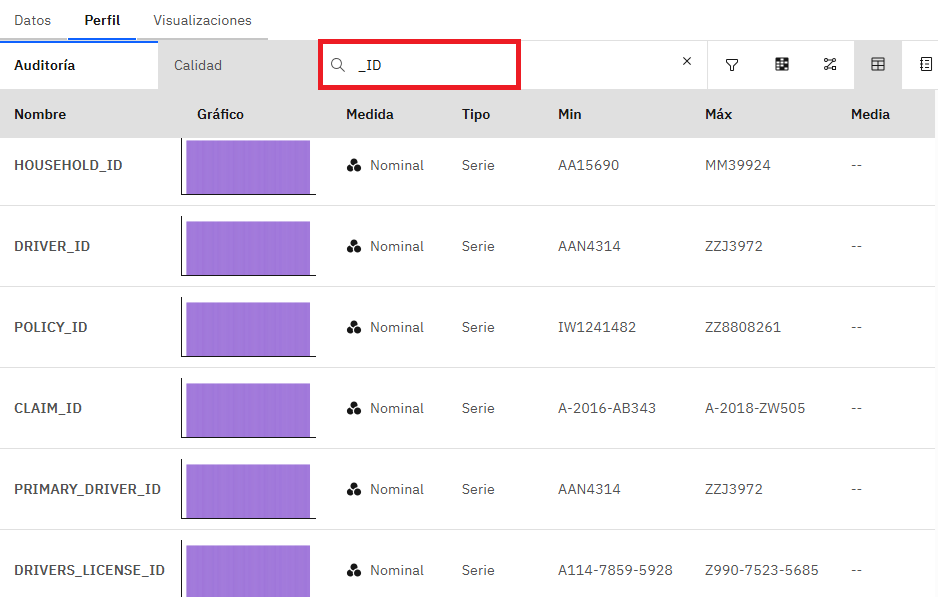
1. En la pestaña Activos, haz clic en el archivo CSV recién cargado.
2. En el panel derecho, haz clic en Abrir en Data Refinery o en Vista Preliminar Preparar Datos.
3. Espera unos minutos mientras Data Refinery procesa los datos. Verás tres pestañas principales: Datos, Perfil y Visualizaciones.



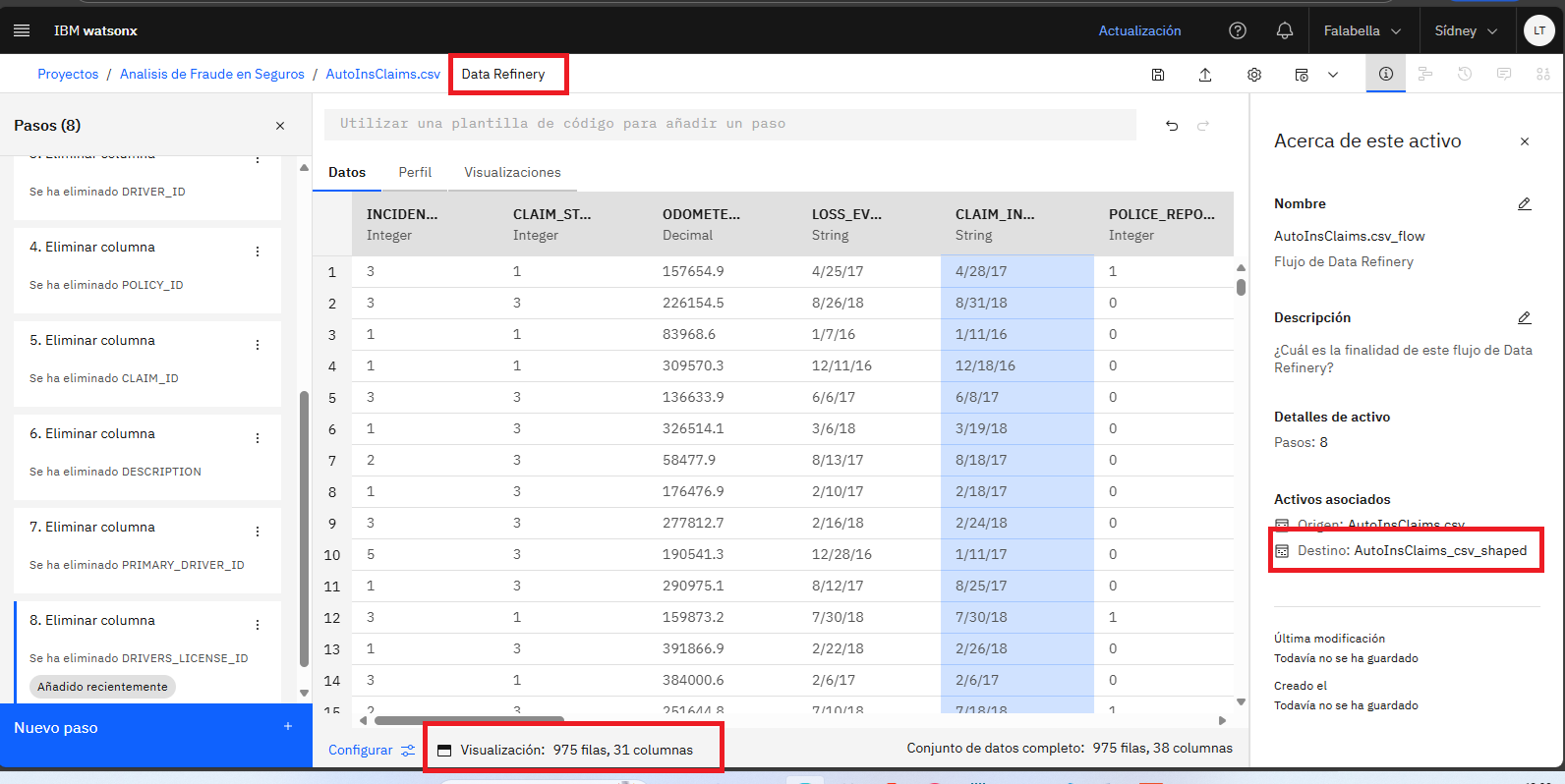
4.1. Identificación y Eliminación de Datos Inútiles (Columnas Ruidosas)

*Objetivo:* Eliminar columnas que no aportan valor predictivo (identificadores únicos y campos vacíos).

1. Ve a la pestaña Perfil.
2. Busca columnas con el mismo número de “eventos únicos” que el total de filas (975). Estas son columnas de identificación única. Sin embargo IBM Watsonx, al parecer ya identifica automáticamente este tipo de columnas únicas.



1. Regresa a la pestaña Datos.
2. Elimina las siguientes columnas:
   * HOUSEHOLD\_ID
   * DRIVER\_ID
   * POLICY\_ID
   * CLAIM\_ID
   * PRIMARY\_DRIVER\_ID *(es otro identificador)*
   * DESCRIPCIÓN *(está vacía según la transcripción)*
   * DRIVERS\_LICENSE\_ID (agregada para eliminar según IBM Watsonx)
3. *Resultado:* Se han reducido de 38 a 31 columnas.



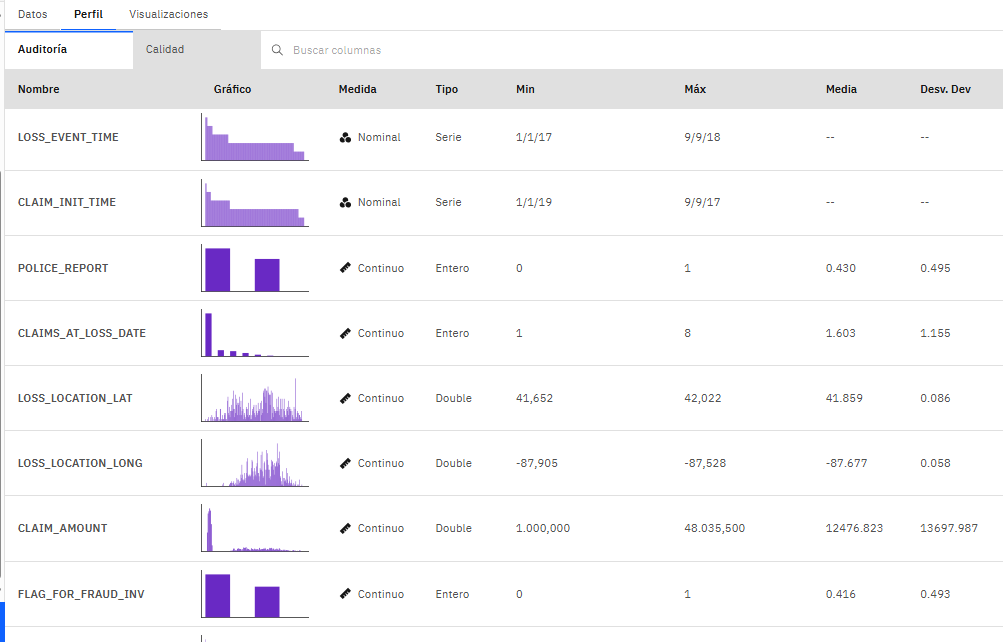
4.2. Análisis de Columnas Relevantes para la Predicción de Fraude

*Objetivo:* Identificar y conservar las variables con potencial predictivo.

Verifica en la pestaña Perfil y conserva las siguientes columnas:

| **Columna** | **Funcionalidad - Descripción** |
| --- | --- |
| INCIDENT\_CAUSE | Tiene variabilidad (desviación estándar > 0); puede indicar patrones de fraude (ej. colisión falsa). |
| CLAIM\_STATUS | Estado final de la reclamación (abierto, aprobado, pagado, fraude). Variable objetivo clave. |
| ODOMETER\_AT\_LOSS | Kilometraje del vehículo al momento del siniestro; puede indicar manipulación. |
| LOSS\_EVENT\_TIME | Fecha/hora del accidente. Crucial para calcular tiempos de denuncia. |
| CLAIM\_INIT\_TIME | Fecha/hora de presentación de la reclamación. Permite calcular el retraso entre evento y reclamación. |
| POLICE\_REPORT | Indica si hubo informe policial; los fraudes suelen evitarlo. |
| CLAIM\_AMOUNT | Monto reclamado. Variable central para la hipótesis de negocio. |
| FLAG\_FOR\_FRAUD\_INV | Variable de entrenamiento (target). Indica si la reclamación fue confirmada como fraude (1) o no (0).¡Conservar! |

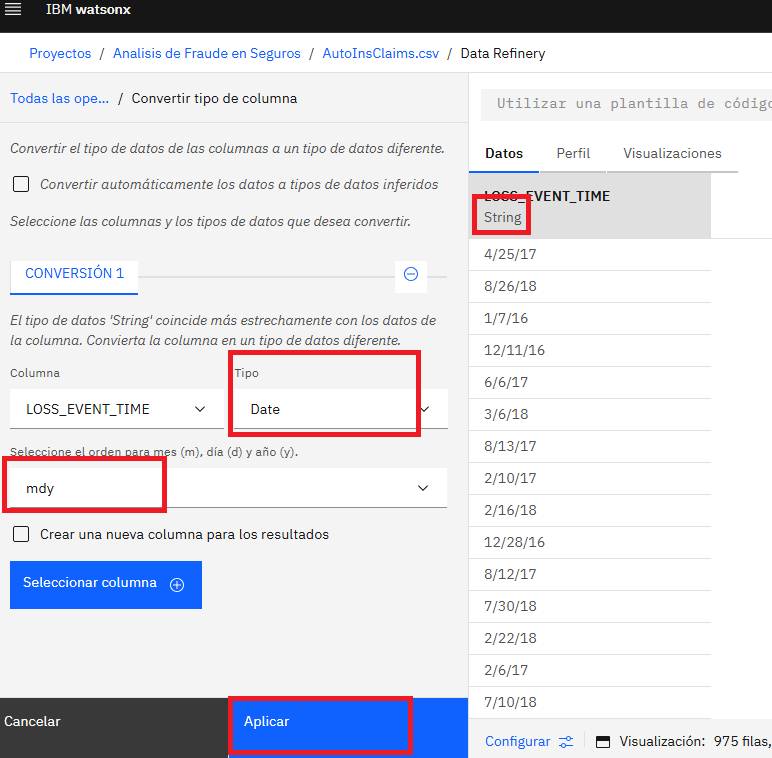
✅ *Nota del Instructor:* **FLAG\_FOR\_FRAUD\_INV** es tu etiqueta de clase. Sin ella, no puedes entrenar un modelo de aprendizaje supervisado. Nunca la elimines.



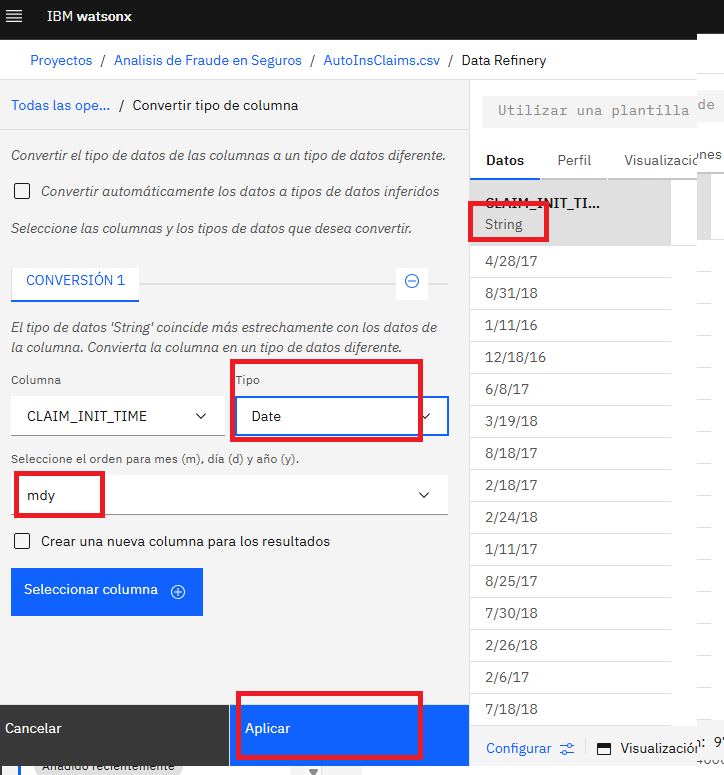
4.3. Corrección de Tipos de Datos

*Objetivo:* Asegurar que los cálculos temporales y numéricos sean precisos.

1. Localiza la columna LOSS\_EVENT\_TIME.
2. Observa que su tipo es Cadena (string) — esto impide realizar operaciones de fecha.
3. Haz clic en los tres puntos (⋯) junto al nombre de la columna.
4. Selecciona Convertir tipo de columna.
5. En el menú desplegable, cambia el tipo a Fecha.
6. Verifica que el formato sea DD/MM/YYYY (o el correcto según tus datos — ajusta si es necesario).
7. Haz clic en Aplicar.



1. Repite el mismo proceso para la columna CLAIM\_INIT\_TIME (convertirla también a tipo Fecha).

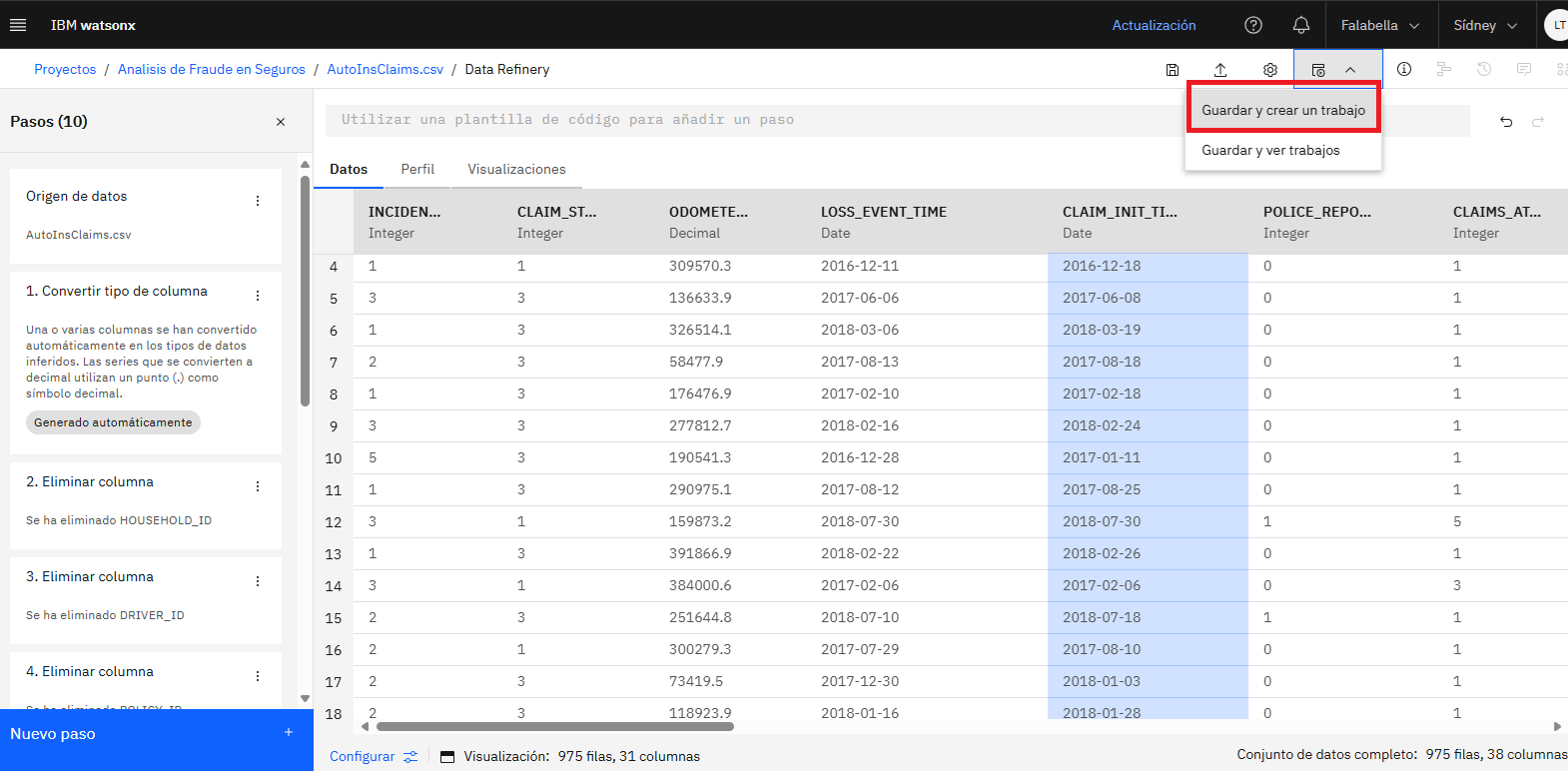


✅ *Nota del Instructor:* Un error común es dejar fechas como texto. Esto impide calcular diferencias de tiempo (ej. ¿cuánto tardaron en denunciar?). ¡Siempre verifica los tipos!

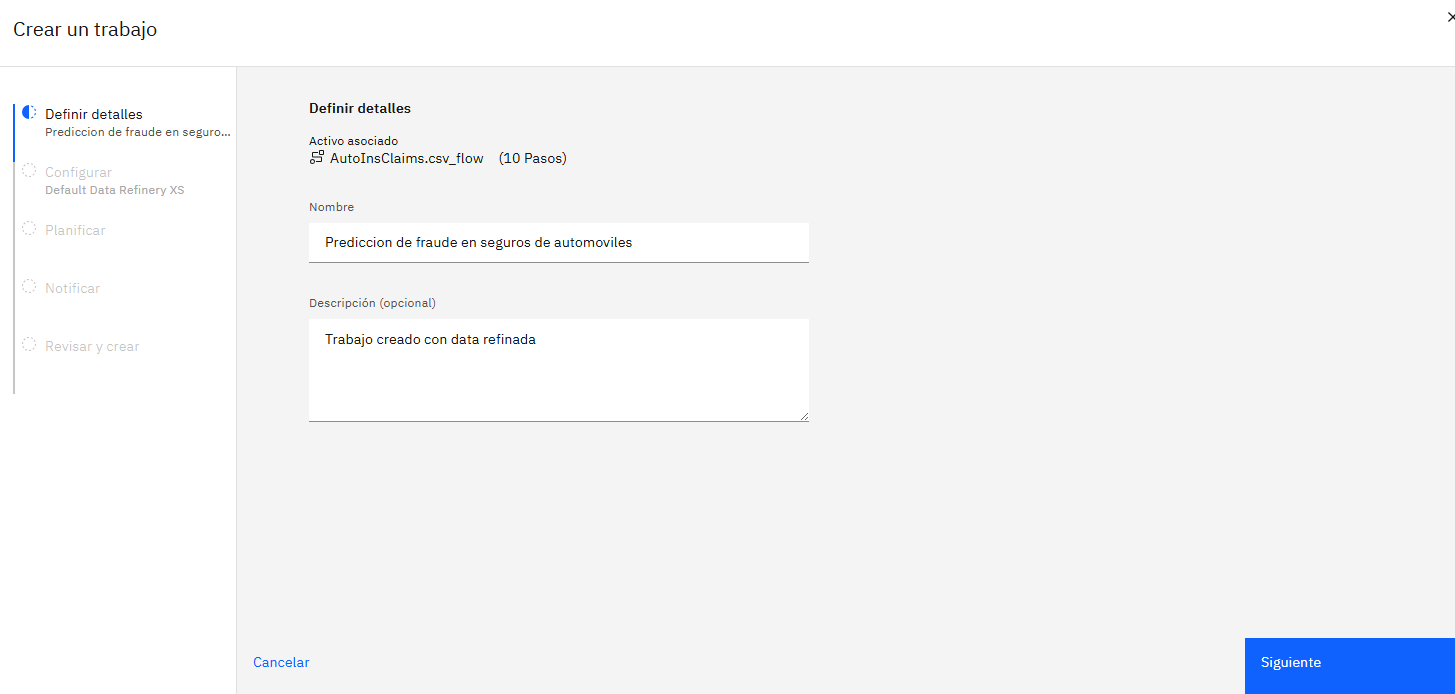
4.4. Guardar el Trabajo ETL y Crear el Archivo Limpio

*Objetivo:* Persistir todos los cambios en un nuevo archivo de datos limpio.

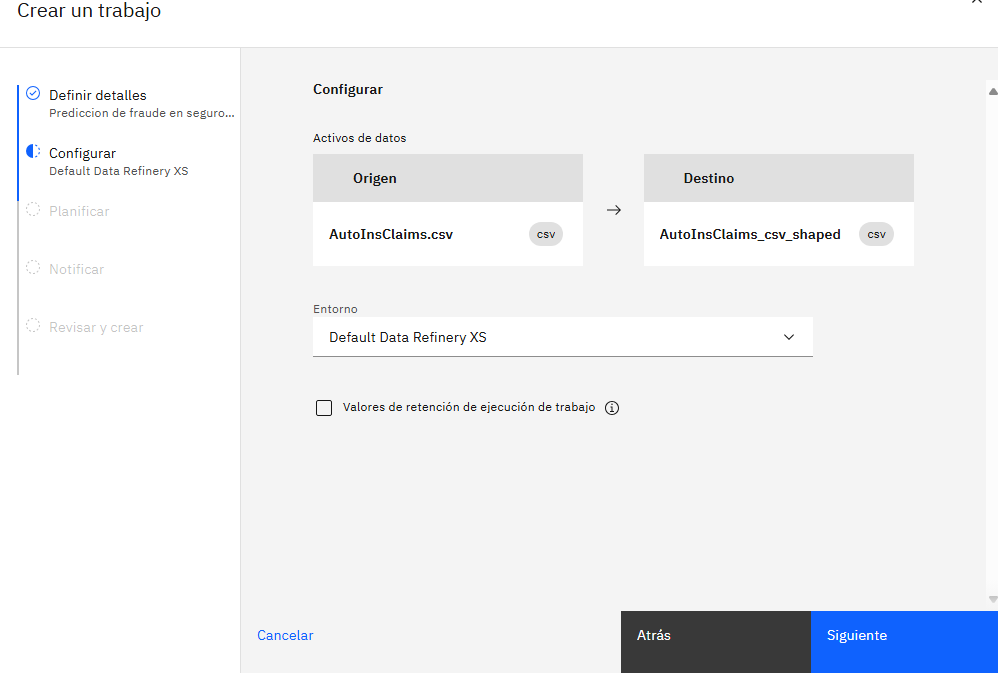
1. Haz clic en el ícono de Guardar (disco) en la esquina superior derecha.
2. Selecciona Crear trabajo (*Create Job*).



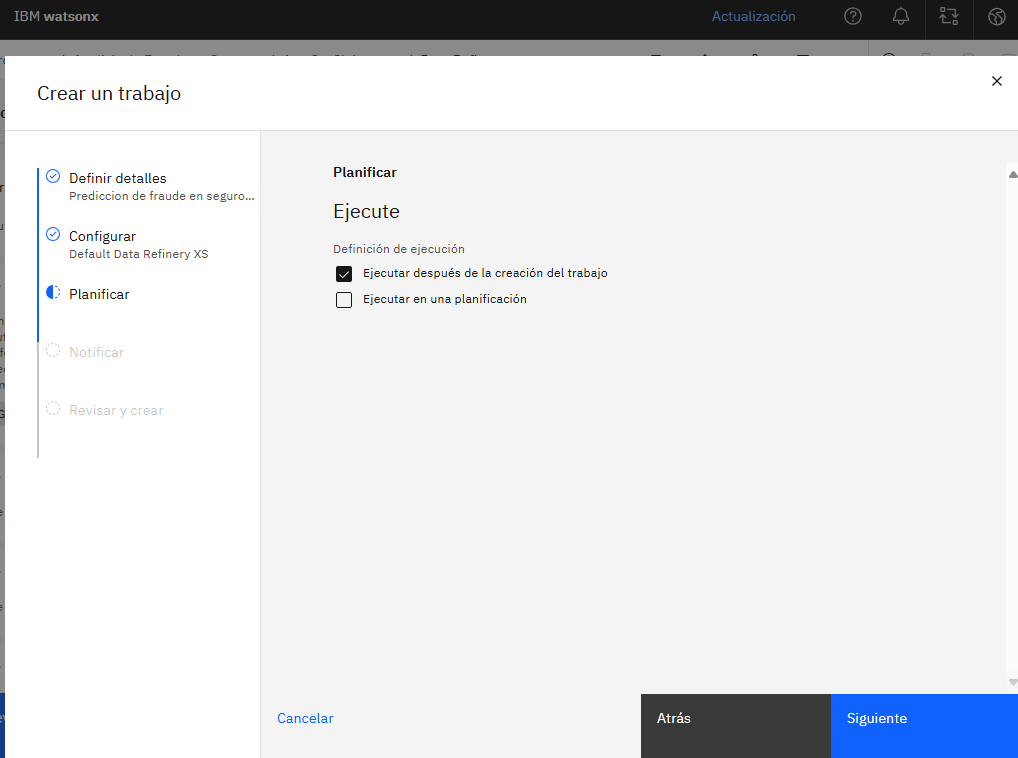
1. Nombre del trabajo: “Predicción de fraude en seguros de automóviles” (se omitió tildes).



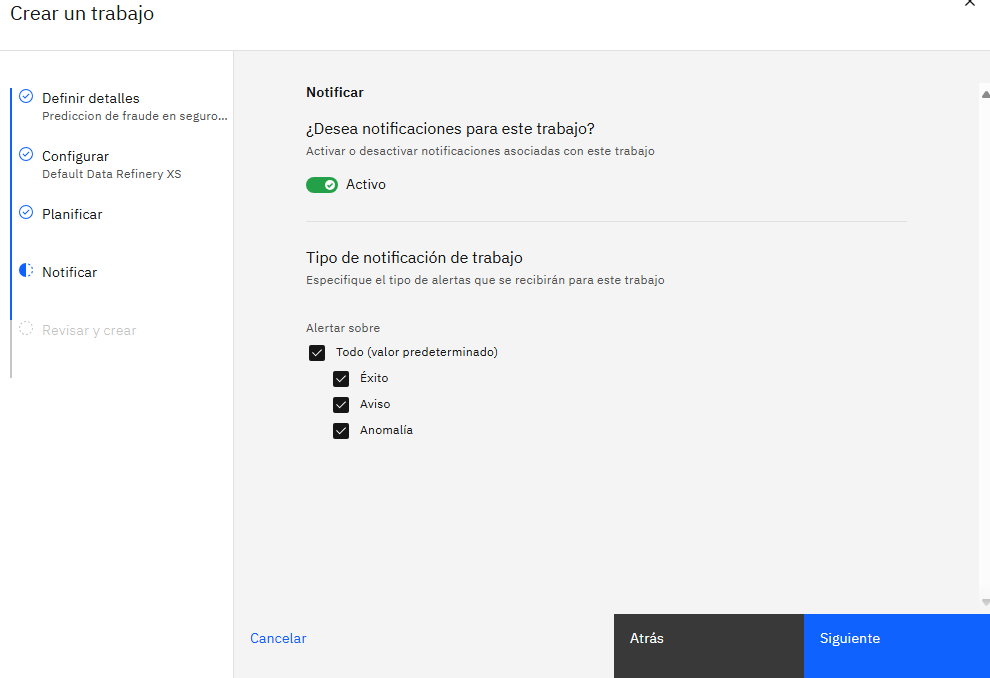
1. Acepta los valores predeterminados en las pantallas de configuración:
   * *Configurar* → Siguiente



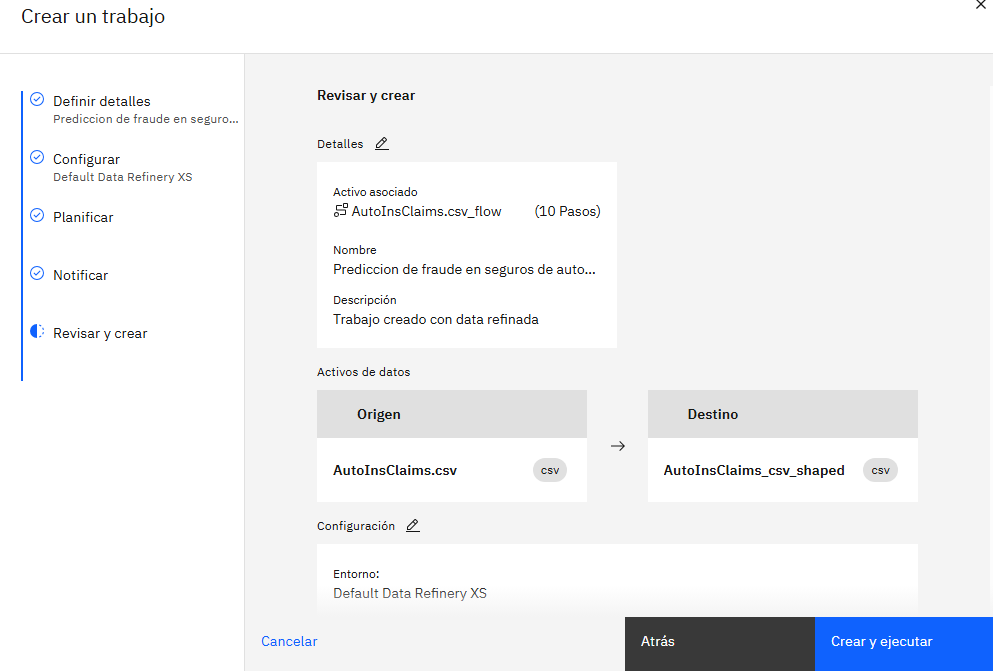
* + *Planificar* → Ejecutar después de la creación del trabajo → Siguiente



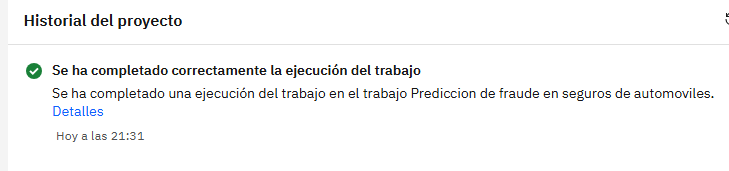
* + *Notificar* → Activo → Alertar Todo → Siguiente



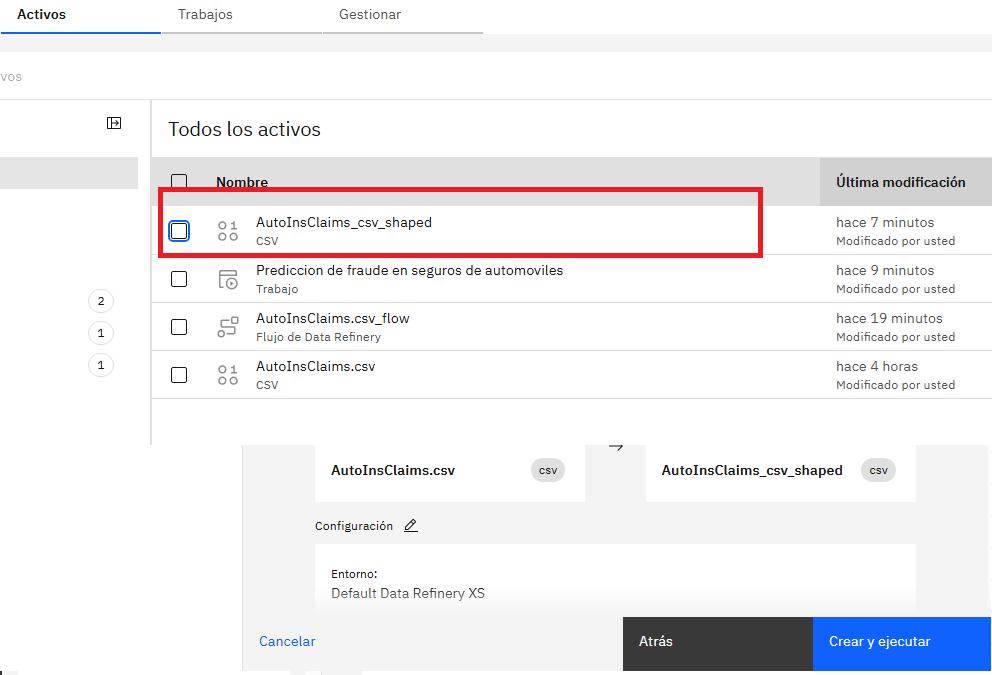
1. En la pantalla final Revisar y crear, verás:
   * Fuente (Source): AutoInsClaims.csv
   * Salida (Output): **AutoInsClaims\_csv\_shaped.csv** *(nuevo archivo con todos los cambios aplicados)*
2. Haz clic en Crear y ejecutar.



1. **Esperar** a que el trabajo se complete (aprox. 1-2 minutos).



*Cuando termine, verás el nuevo archivo AutoInsClaims\_csv\_shaped.csv en la pestaña Activos.*

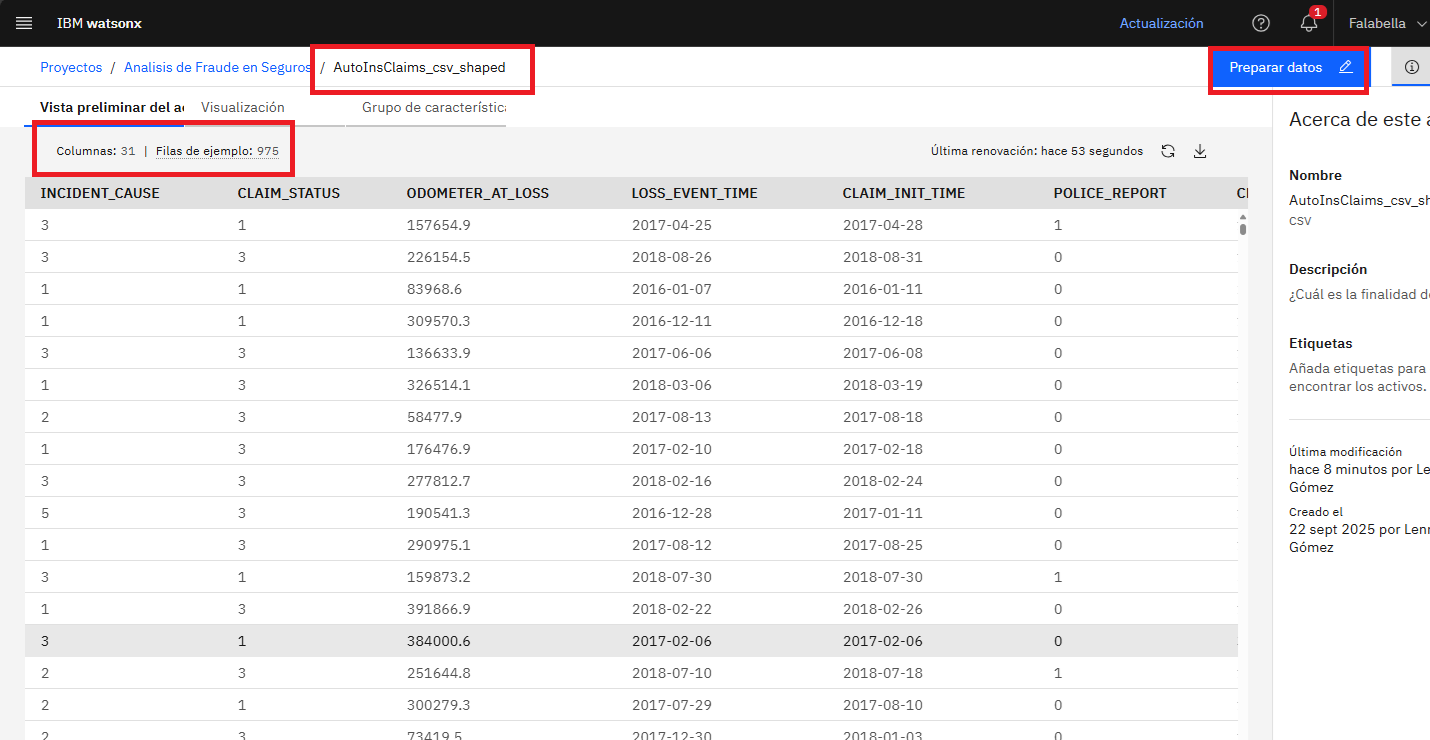


✅ *Nota del Instructor:* Este paso es el corazón del ETL (Extract, Transform, Load). Estás creando un *dataset limpio y reproducible* que será la base de todo el análisis futuro. Nunca trabajes sobre el archivo original (Este proceso se realizará en cada Preparación o nuevo Trabajo).

Paso 5: Refinar los Datos – Crear una Variable Derivada para Probar la Hipótesis

*Objetivo:* Transformar la hipótesis de negocio en una variable cuantificable para análisis.

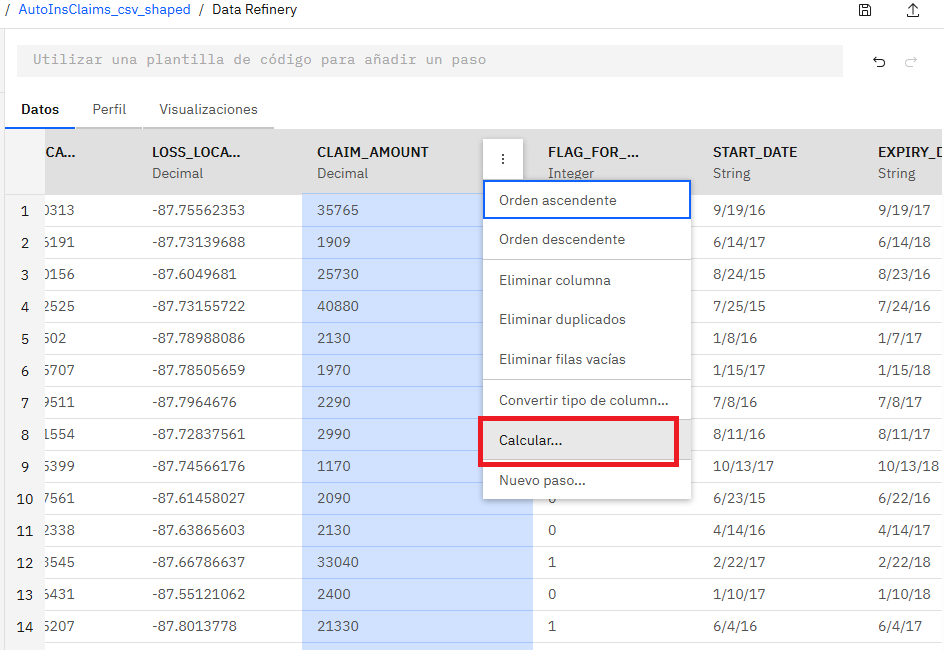
1. En la pestaña Activos, abre el archivo AutoInsClaims\_csv\_shaped.csv.
2. Haz clic en Abrir en Data Refinery.
3. Ahora trabajas con el conjunto limpio (31 columnas).



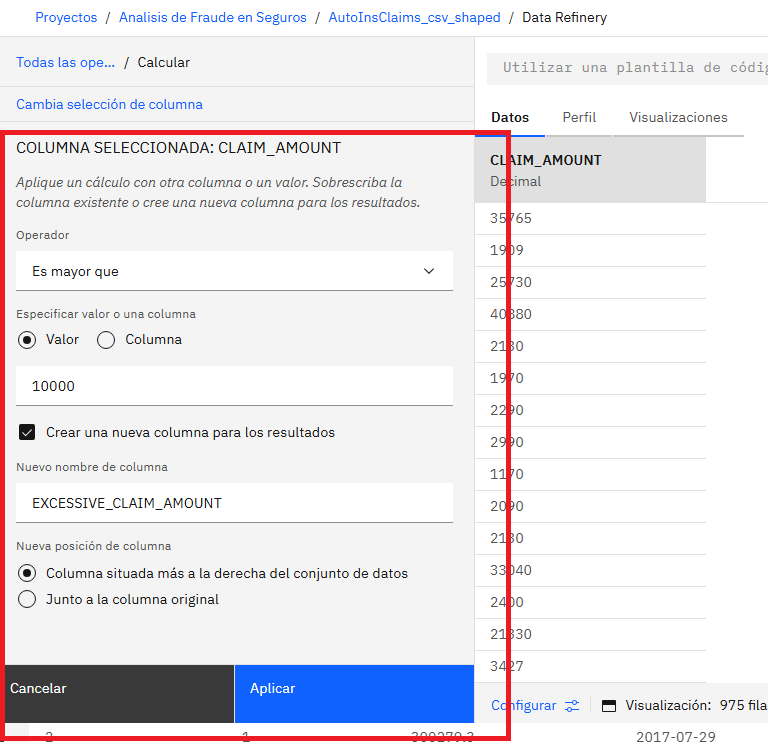
5.1. Crear la Columna EXCESSIVE\_CLAIM\_AMOUNT

*Objetivo:* Marcar si una reclamación supera los $10,000, como sugiere la hipótesis.

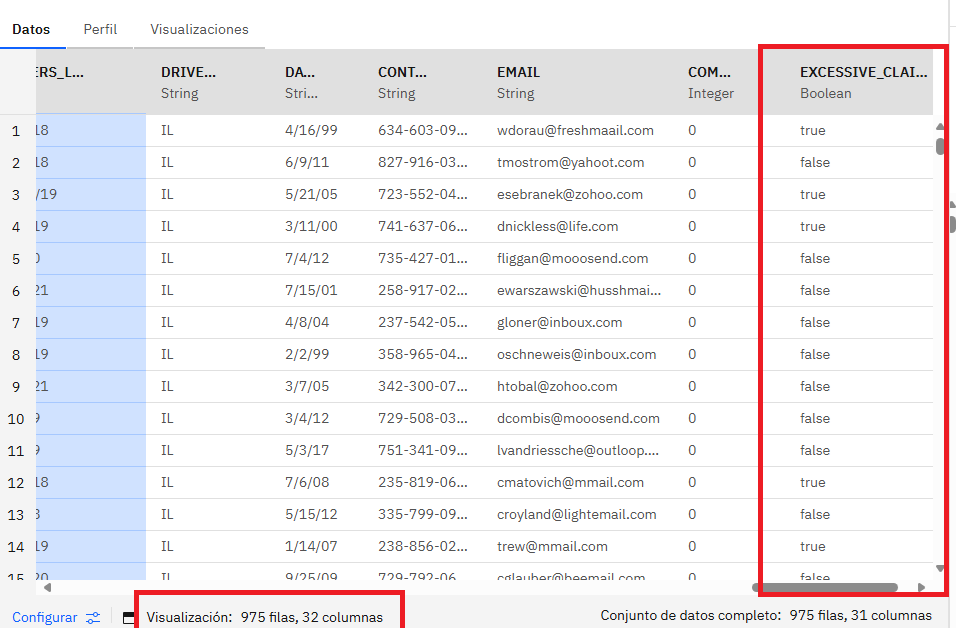
1. Localiza la columna CLAIM\_AMOUNT.
2. Haz clic en los tres puntos (⋯) junto al nombre de la columna.
3. Selecciona Nuevo paso... → Calcular.



1. Configura el cálculo:
   * Operador: Es mayor que
   * Valor: 10000
   * ✅ Marca la casilla: “Crear una nueva columna para los resultados” *(¡Crucial! No sobrescribas la original)*
   * Nombre de nueva columna: **EXCESSIVE\_CLAIM\_AMOUNT**
   * Posición: Deja en predeterminado (al final)
2. Haz clic en Aplicar.



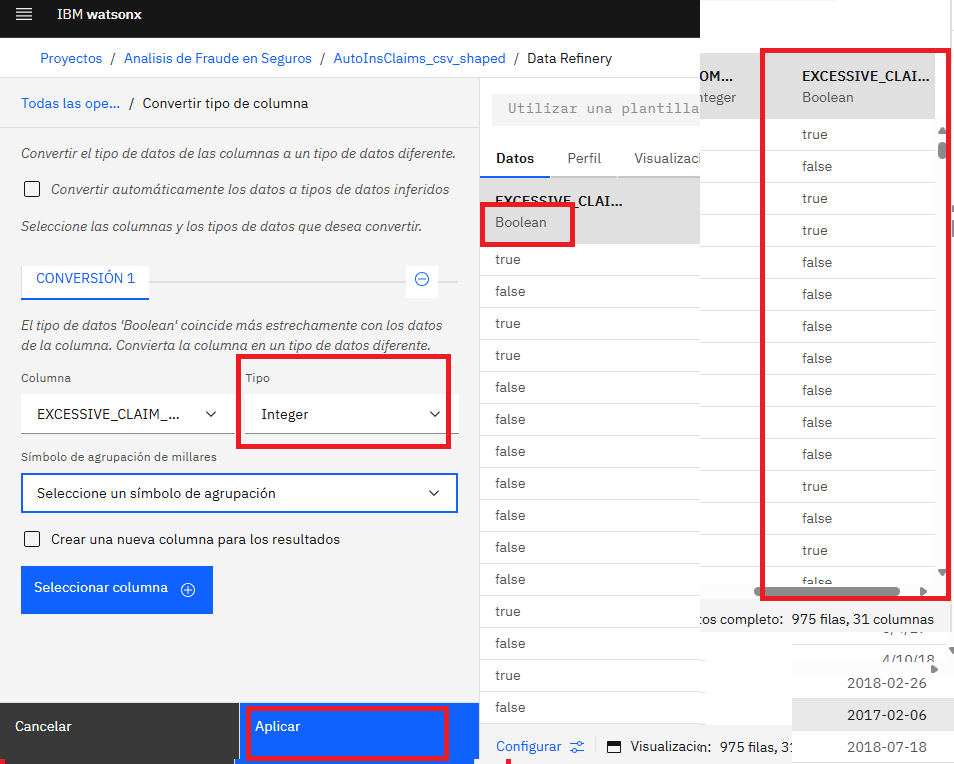
✅ *Resultado:* Se crea una columna booleana (32) con valores Verdadero (True) si CLAIM\_AMOUNT > 10000, Falso (False) en caso contrario.



5.2. Convertir la Nueva Columna a Entero (0/1)

*Objetivo:* Facilitar el análisis estadístico y la entrada en modelos de machine learning.

1. Localiza la nueva columna EXCESSIVE\_CLAIM\_AMOUNT (al final del dataset).
2. Observa que sus valores son Verdadero/Falso (tipo booleano).
3. Haz clic en los tres puntos (⋯) junto a EXCESSIVE\_CLAIM\_AMOUNT.
4. Selecciona Convertir tipo de columna...
5. En el menú desplegable Tipo, selecciona Entero.
6. Deja las demás opciones como están.
7. Haz clic en Aplicar.

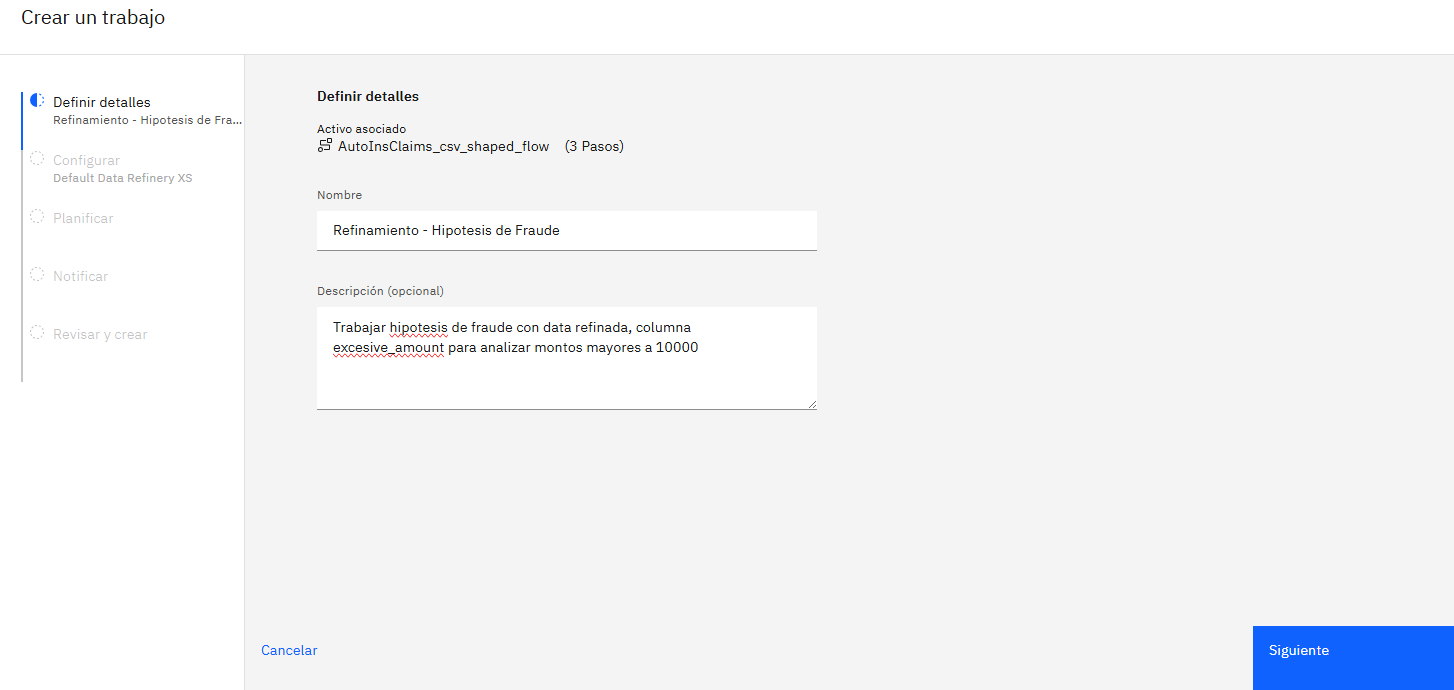


✅ *Resultado:*

* Falso → se convierte en 0
* Verdadero → se convierte en 1  
  *Esto permite sumar, promediar y usarla directamente en modelos como regresión logística o árboles de decisión.*

5.3. Guardar el Conjunto de Datos Refinado

1. Haz clic en el ícono de Guardar.
2. Crea un nuevo trabajo si se solicita (**Opcional en esta etapa, pero recomendado**).
   * Nombre: Refinamiento - Hipótesis de Fraude
   * Salida: **AutoInsClaims\_csv\_shaped\_shaped.csv**
3. Haz clic en Crear y ejecutar.
4. Espera a que se complete.

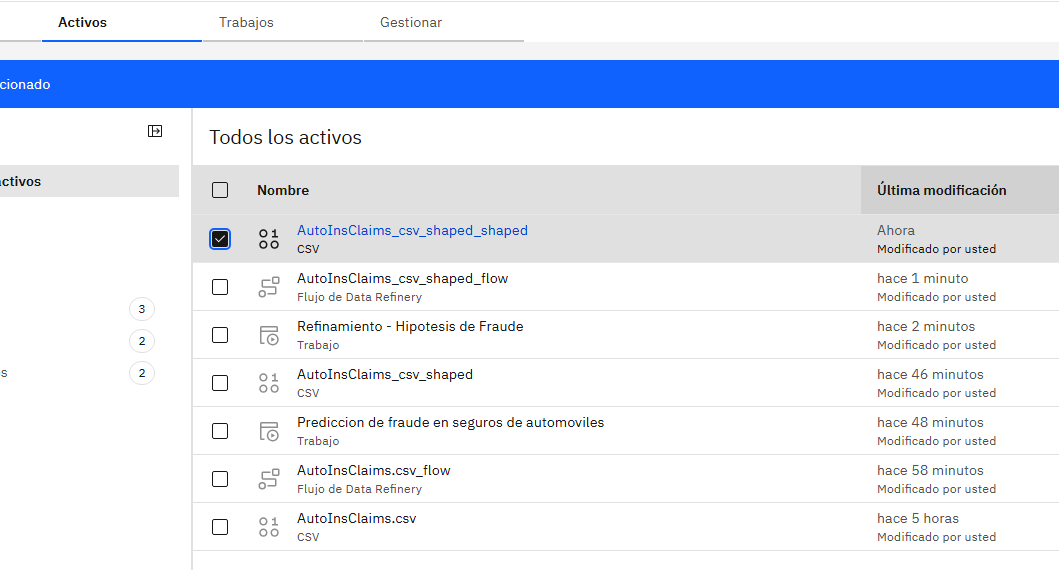


✅ *Ahora tienes tu dataset final: limpio, con tipos correctos y una variable derivada clave para la hipótesis.*

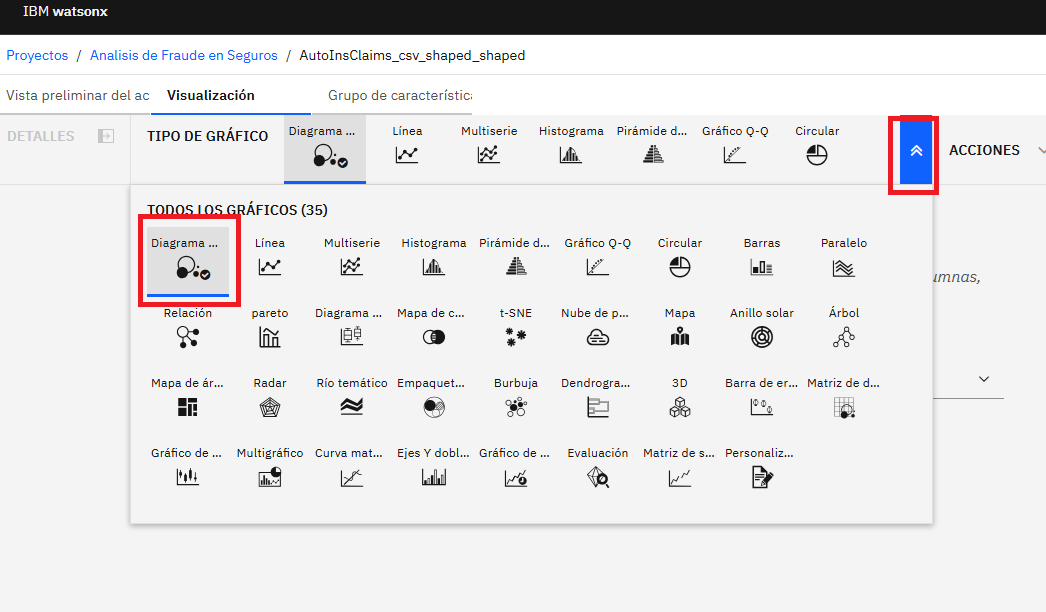
Paso 6: Visualización de Datos para Validar la Hipótesis de Negocio

***Objetivo:* Transformar los datos en una historia visual que responda a la pregunta del negocio: *“¿Las reclamaciones >$10,000 son fraudulentas?”***

1. En la pestaña Activos, abre el archivo *AutoInsClaims\_csv\_shaped\_shaped.csv*.

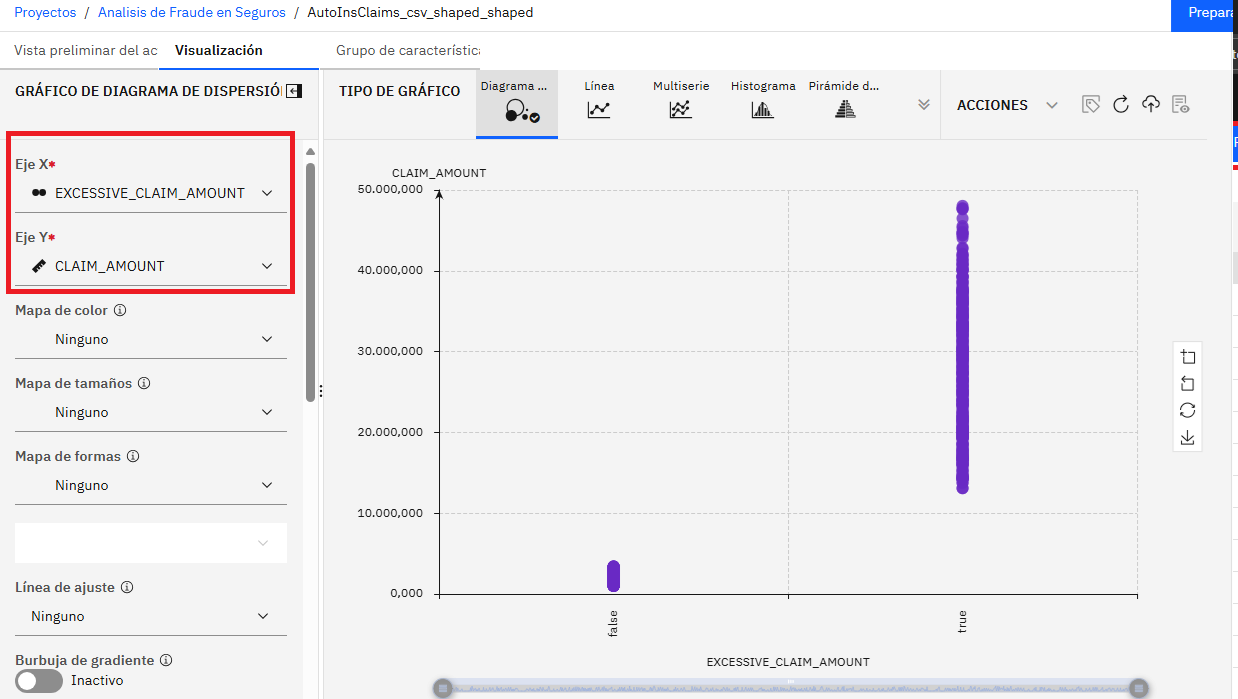


1. Selecciona la pestaña Visualizaciones.
2. Haz clic en el ícono de flecha doble (en la esquina superior derecha) para ver todas las opciones de gráficos.
3. Selecciona Gráfico de dispersión (*Scatter Plot*).



6.1. Configurar los Ejes del Gráfico

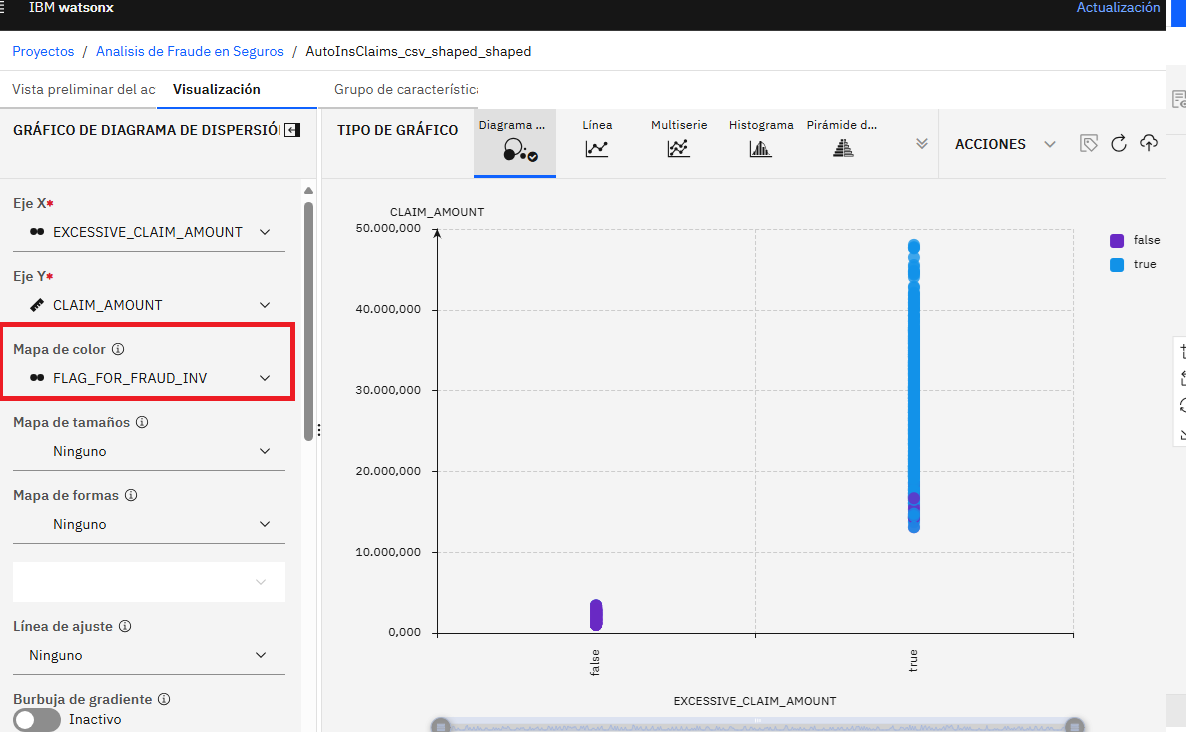
1. Eje X (Variable Independiente):  
   Selecciona EXCESSIVE\_CLAIM\_AMOUNT (0 = no excesiva, 1 = excesiva).  
   *Esto agrupará los puntos en dos columnas verticales: izquierda (≤$10K) y derecha (> $10K).*
2. Eje Y (Variable Dependiente):  
   Selecciona CLAIM\_AMOUNT (monto real de la reclamación).  
   *Esto muestra el valor real de cada reclamación, permitiendo ver su distribución.*
3. Observa el gráfico inicial:
   * Dos grupos de puntos en el eje X (0 y 1).
   * Los puntos en el eje Y van desde $0 hasta el máximo (posiblemente >$50,000).



6.2. Añadir Dimensión de Color: FLAG\_FOR\_FRAUD\_INV

*Objetivo:* Identificar qué reclamaciones son fraudes reales.

1. Busca la opción Mapa de colores (*Color by*) o similar.
2. Selecciona la columna FLAG\_FOR\_FRAUD\_INV.
3. El gráfico ahora mostrará:
   * Azul: Reclamaciones NO fraudulentas (0)
   * Rojo/Rosa/Morado: Reclamaciones fraudulentas (1)



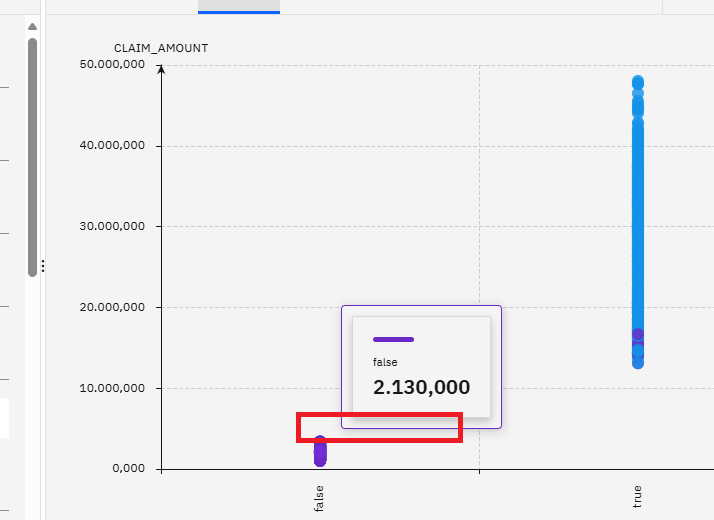
✅ *Nota del Instructor:* Este es el momento clave. Estás añadiendo la verdad del negocio (lo que realmente ocurrió) a tu visualización. Ahora puedes ver si tu variable derivada (EXCESSIVE\_CLAIM\_AMOUNT) tiene relación con la realidad.

6.3. Interpretación del Gráfico de Dispersión

*(Utiliza el gráfico interactivo: pasa el cursor sobre los puntos para ver detalles de cada reclamación)*

Insight 1: Fraude en Reclamaciones Pequeñas (<$10,000)

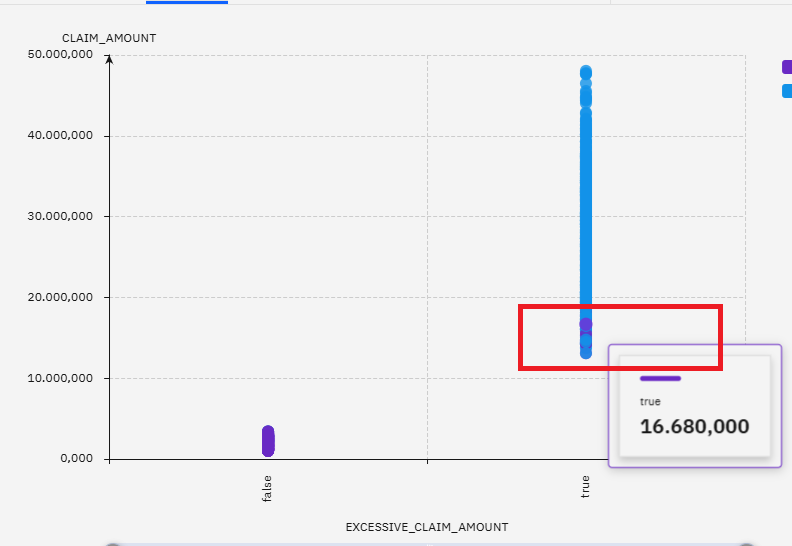
* En el lado izquierdo (Eje X = 0), hay puntos rojos/rosas (fraude) por debajo de $10,000.
* *Ejemplo:* Una reclamación de $2,130 marcada como fraude.



✅ Conclusión: El fraude no es exclusivo de las reclamaciones grandes. Hay fraudes de bajo monto que se escapan a la hipótesis.

Insight 2: Existe evidencia de Fraude en Reclamaciones Grandes (> $10,000)

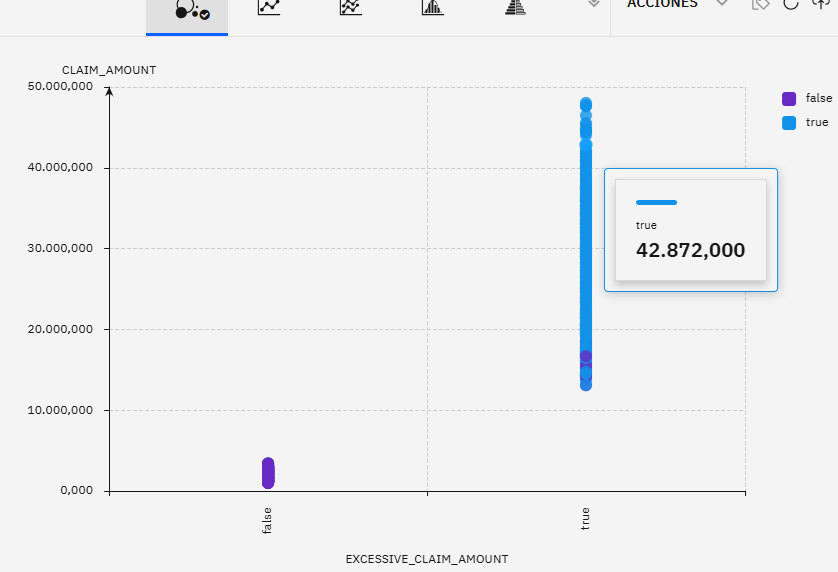
* En el lado derecho (Eje X = 1), hay una fuerte concentración de puntos rojos/rosas (fraude) entre los $10,000 y $20,000.



✅ Conclusión: La hipótesis es parcialmente válida: las reclamaciones grandes tienen mayor probabilidad de ser fraudulentas.

Insight 3: No Todas las Reclamaciones Grandes son Fraude

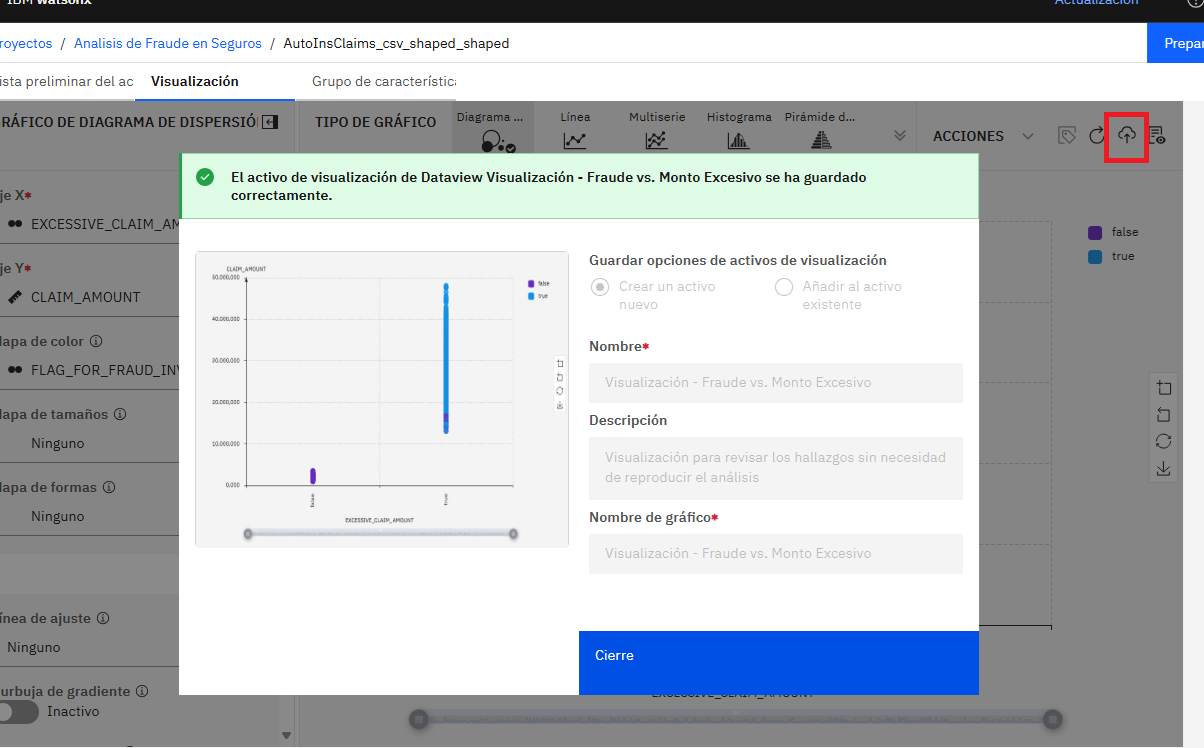
* También hay puntos azules (no fraude) por encima de $10,000.
* *Ejemplo:* Una reclamación de $42,000 por daño real en un camión de carga.



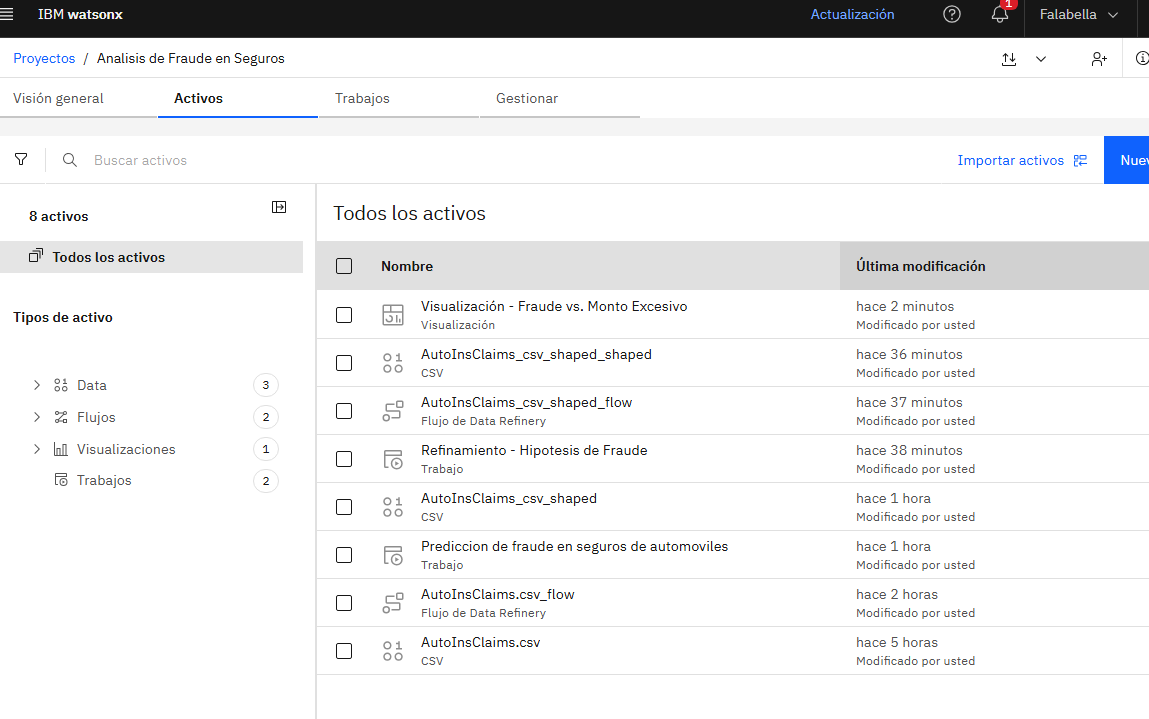
✅ Conclusión: Una regla simple como “todo >$10,000 es fraude” generaría muchos falsos positivos, afectando a clientes legítimos y generando insatisfacción.

6.4. Guardar la Visualización

1. Haz clic en Guardar visualización en el proyecto (ícono de disco en la esquina superior derecha).
2. Asigna un nombre claro: “Visualización - Fraude vs. Monto Excesivo”.
3. Confirma.



✅ *Nota del Instructor:* Guardar visualizaciones es esencial. Permitirá que los stakeholders revisen los hallazgos sin necesidad de reproducir el análisis. (Es parte del ciclo de gobernanza de datos).



*Fin de Práctica Guiada.*

**Práctica Guiada actualizada con Gemini Flash 2.5 (Instructor)**

Paso 7: Comunicación de Insights a la Patrocinadora de Negocio

*Objetivo:* Transformar los hallazgos técnicos en una narrativa clara, accionable y persuasiva.

¡Presentación Recomendada para la Patrocinadora!

*“Buenos días [Nombre de la Patrocinadora],*  
*He analizado los datos de reclamaciones para validar su hipótesis: ‘Las reclamaciones superiores a $10,000 pueden ser fraudulentas’. Los resultados son claros y tienen implicaciones estratégicas.*

1. Su hipótesis es PARCIALMENTE CORRECTA — y es un indicador poderoso.  
Los datos muestran una fuerte concentración de fraudes confirmados en reclamaciones mayores a $10,000. Esto confirma que el monto alto es un fuerte predictor y debe ser una regla de alerta automática en su sistema de detección.\*

2. El fraude no se limita a las reclamaciones grandes.  
Vemos claramente que hay fraudes en montos de $2,000 a $9,000. Esto significa que nuestro sistema de detección no puede enfocarse solo en montos altos. Necesitamos incluir otros factores, como el tipo de incidente, el retraso en la denuncia o la ausencia de informe policial.\*

3. No todas las reclamaciones grandes son fraudulentas.  
Hay muchas reclamaciones >$10,000 que son legítimas. Si aplicamos una regla rígida (“todo >$10K es fraude”), estaremos falsamente acusando a clientes honestos, lo que daña la reputación y genera costos por revisión innecesaria.\*

Recomendaciones Inmediatas:

* ✅ Utilice EXCESSIVE\_CLAIM\_AMOUNT como una variable clave en su modelo de machine learning.
* ✅ Combine esta variable con otras: POLICE\_REPORT, INCIDENT\_CAUSE, y la diferencia entre LOSS\_EVENT\_TIME y CLAIM\_INIT\_TIME.
* ✅ No implemente una regla simple de umbral. Implemente un modelo predictivo (como Random Forest o Regresión Logística) que combine múltiples señales para reducir falsos positivos y captar fraudes de bajo monto.\*

*En resumen: La hipótesis es una excelente pista, pero no una regla. La ciencia de datos nos permite ir más allá de las suposiciones — y construir sistemas inteligentes que detectan fraude con precisión, no con ruido.*  
*Desarrolle un modelo predictivo con estos datos en la próxima fase.*

**✅** Checklist Final de Comprensión.

| **Aprendizaje** | **Resultado esperado** |
| --- | --- |
| ¿Entiendo por qué se utiliza IBM Watson Studio? | Sí. Es una plataforma integrada que permite importar, limpiar, transformar, visualizar y  modelar datos en un solo entorno, eliminando la fragmentación de herramientas. |
| ¿Cuál es la importancia del refinamiento de datos? | Los datos crudos son ruidosos. Limpiar columnas, corregir tipos y crear variables  derivadas garantiza que los modelos sean precisos y los insights confiables. |
| ¿Cómo la visualización ayuda a resolver el fraude? | Permite ver patrones complejos (relación entre monto, fraude y otras variables)  de forma intuitiva, evitando conclusiones erróneas basadas solo en números. |
| ¿La hipótesis se validó completamente? | No. Se validó que las reclamaciones >$10,000 son un fuerte indicador,  Pero no son la única fuente de fraude, y no todas son fraudulentas. |
| ¿Qué me dice el color de los puntos? | El color (FLAG\_FOR\_FRAUD\_INV) representa la verdad real del negocio: qué reclamaciones  fueron investigadas y confirmadas como fraude. Es la etiqueta que valida todo el análisis. |
| ¿Cuál es el valor de esta visualización? | Transforma datos en historia. Permite tomar decisiones con evidencia, no con suposiciones.  Evita costosos errores de negocio (como acusar a clientes legítimos). |
| ¿Qué haría en la siguiente fase? | Entrenar un modelo de machine learning supervisado (por ejemplo, Random Forest) usando  EXCESSIVE\_CLAIM\_AMOUNT, POLICE\_REPORT, INCIDENT\_CAUSE, CLAIM\_INIT\_TIME - LOSS\_EVENT\_TIME, y FLAG\_FOR\_FRAUD\_INV como variable objetivo. |

Bitácora de Avance

| Proceso | Descripción | Fecha |
| --- | --- | --- |
| ✅ Análisis de Datos | Proceso Completo de Análisis de Datos en IBM Watson Studio  Importación → Limpieza → Refinamiento → Visualización → Comunicación | [Lunes 22/09/2025] |

Recomendaciones Finales del Instructor

1. Capturas de Pantalla:
   * Inserta una captura de cada paso clave:
     + Eliminación de columnas
     + Conversión de tipo de dato (fecha)
     + Creación de EXCESSIVE\_CLAIM\_AMOUNT
     + Gráfico de dispersión final con colores
     + Ventana de “Guardar visualización”
   * Etiqueta cada imagen claramente: *“Figura 1: Eliminación de COLUMNAS IDENTIFICADORAS”*
2. Documentación:
   * Guarda este documento como Actividad\_Guiada\_Fraude\_de\_Seguros.docx
   * Crea una carpeta dentro de tu proyecto en Watson Studio llamada Documentación y sube aquí este documento y las capturas.
3. Próxima Etapa (Opcional pero Recomendada):
   * Usa AutoAI en Watson Studio para entrenar automáticamente un modelo predictivo con este dataset.
   * Compara la importancia de variables: ¿es EXCESSIVE\_CLAIM\_AMOUNT la más importante? ¿y POLICE\_REPORT?
   * Genera una matriz de confusión y métricas de rendimiento (precisión, recall, F1-score).

**💡 Conclusión Final del Instructor**

“Los datos no hablan por sí solos. La ciencia de datos es el arte de hacer que los datos cuenten una historia que el negocio entienda, confíe y actúe.”

En este proyecto, no solo aprendiste a usar IBM Watson Studio.  
Aprendiste a pensar como un científico de datos:

* Hacer preguntas correctas.
* Transformar hipótesis en variables.
* Validar con evidencia visual.
* Comunicar con claridad y humildad.

Esto es lo que diferencia a un analista de un líder de datos.