

INFORME TÉCNICO – EDA EN PYTHON

Limpieza, Transformación y Análisis Exploratorio de Datos

 Proyecto Final – Máster en Data Analytics

 Autora: Patricia Merinero

 Diciembre 2025

Estructura del informe EDA en Python.

1. Introducción
2. Datos y calidad
3. Principales transformaciones
4. Análisis visual (8 gráficos) con insights y recomendaciones
5. Conclusión general del proyecto y utilidad
 - Síntesis global del flujo de trabajo
 - Principales hallazgos del análisis.
 - Utilidad del proyecto para negocio y para análisis de datos.
 - Cierre

1. Introducción

Objetivo del análisis

Realizar un **análisis exploratorio de datos (EDA)** sobre un conjunto de **operaciones financieras** para:

- **Identificar patrones, relaciones y tendencias** asociados a **fraude vs. no fraude**.
- Entender cómo variables como **comercio, tipo de comercio, nivel de riesgo (alto/medio/bajo), antigüedad, tipo de operación, tipo de tarjeta, resultado de la operación** se relacionan con la probabilidad de fraude.
- Generar **insights accionables** que ayuden a **priorizar revisiones, afinar reglas y mejorar la prevención y la gestión operativa del fraude**.

Preguntas guía:

- ¿Cuál es la **prevalecia de fraude** global y por **segmentos** (merchant / tipo de comercio / nivel de riesgo / tipo de operación)?
- ¿Existen **picos temporales** de fraude (por día/mes/hora o por antigüedad de la cuenta)?
- ¿Qué **combinaciones de variables** incrementan el riesgo?
- ¿Qué diferencias hay por variables **geográficas** (*region = región del cliente; country = país donde se realiza la operación*)?
- ¿Qué **insights** pueden convertirse en **acciones concretas**?

Estas preguntas actúan como guía del análisis y orientan la búsqueda de patrones en los datos.

Cada una de ellas se abordará a lo largo del informe mediante gráficos, métricas y conclusiones.

Contexto del dataset

El análisis parte de **dos archivos originales**:

1. **Clientes (original)** → contiene información básica del cliente: customer_id, nombre, email, phone, region, country, created_at, risk_core.

2. **Transacciones (original)** → recoge el detalle de operación: transaction_id, customer_id, transaction_time, amount, currency, merchant, merchant_category, arn, entry_mode, card_bin6, card_last4, card_masked, card_tipe, card_expiry, transation_result, is_fraud.

Ambos archivos se **unieron mediante una columna en común** (customer_id), dando lugar a un **único dataset consolidado** que integra tanto el perfil del cliente como la información transaccional.

El resultado de esta unión se sometió posteriormente a un proceso de **limpieza y transformación**, del cual surgió el archivo final **dataset_limpio_y_transformado.csv**. Este último es el que se utiliza en el análisis exploratorio y en todo el informe.

2. Datos y calidad

2.1. Origen y consolidación

El análisis se ha desarrollado a partir de un conjunto de **datos sintéticos** que simulan operaciones financieras reales.

El proyecto parte de **dos fuentes iniciales**:

1. **Clientes** → información general del cliente (identificador, correo, teléfono, país, etc.).
2. **Transacciones** → operaciones financieras con detalle de comercio, importe, tipo de tarjeta, modo de entrada, nivel de riesgo y etiqueta de fraude.

Ambos conjuntos se **unieron mediante una columna en común (customer_id)**, generando un **dataset consolidado** que integra el comportamiento transaccional con los atributos del cliente.

Posteriormente, se llevó a cabo un proceso de **limpieza, transformación y normalización**, cuyo resultado final se guardó en el archivo:

dataset_limpio_y_transformado.csv

Este archivo constituye la **fuente única de verdad** sobre la que se ha desarrollado todo el análisis exploratorio.

2.2. Estructura y volumen

Antes de comenzar con la limpieza y transformación de datos, se partía de dos ficheros sintéticos independientes:

- clientes_sinteticos.csv
- transacciones_sinteticas.csv

Ambos contienen información complementaria relativa al perfil del cliente y a las operaciones realizadas.

Tras unir ambos datasets utilizando la clave común correspondiente, se obtuvo la siguiente estructura inicial:

- **Filas totales (tras la unión): 60.000**
- **Columnas totales: 23**

Este volumen inicial cumple sobradamente con los requisitos del proyecto, que exigían trabajar con un dataset mayor a **50.000 observaciones**.

El resultado final, tras la limpieza y transformación completa:

- **Filas: 39.906**
- **Columnas: 34**
- **Tasa global de fraude (is_fraud): 1,10 %**

Variables relevantes:

- is_fraud → indicador principal (1 = fraude, 0 = no fraude).
- risk_level → clasificación de riesgo (alto, medio, bajo).
- entry_mode → modo de entrada de la transacción (*chip, contactless, magstripe, wallet, ecommerce*).
- transaction_result → resultado de la operación (*approved, declined, pending*).
- card_type → tipo de tarjeta (*VISA, MASTERCARD, AMEX, DISCOVER*).
- merchant_category → categoría o tipo de comercio (*electronics, fashion, grocery, travel, etc.*).
- amount → importe de la transacción.
- risk_score → puntuación numérica de riesgo asignada a la transacción.
- region_normalized → región de residencia o país del cliente.
- country_normalized → país donde se ha realizado la operación.
- customer_days_active → número de días que el cliente llevaba activo en el sistema.

2.3. Completitud y nulos

- **Nulos totales en el dataset: 0.**
 - **Nulos por columna:** ninguna columna presenta valores nulos tras la consolidación.
-

2.4. Integridad y duplicados

- **Duplicados en transaction_id:** 0 → cada transacción es única.
 - **Relación cliente–transacción:** consistente 1-a-N (sin pérdidas evidentes en la clave).
-

2.5. Tipos y rangos (checks básicos)

- **Importe (amount):**
 - **mín:** 2,35 · **q25:** 111,76 · **mediana:** 187,69 · **q75:** 261,12 · **máx:** 656,13
 - **Importes negativos:** 0
 - **Divisa (currency):** EUR (única).
 - **Categorías principales:**
 - **risk_level:** Alto, Medio, Bajo
 - **entry_mode:** chip, contactless, ecommerce, magstripe, wallet
 - **transaction_result:** approved, declined, pending
 - **card_type:** AMEX, DISCOVER, MASTERCARD, VISA
 - **merchant_category:** electronics, fashion, gaming, grocery, pharmacy, restaurant, travel
-

2.6. Fechas y coherencia temporal

- **transaction_date:** de 2024-10-11 a 2025-10-12
- **created_at:** de 2024-06-17 a 2026-02-02

Chequeo de coherencia (transaction_date vs created_at):

- Registros con **transacción anterior a created_at:** 19.899.

Interpretación: este comportamiento sugiere que **created_at** no representa necesariamente la “fecha de alta inicial” de la cuenta/tarjeta, sino **otro hito** (p. ej., fecha de alta del perfil en sistema actual, migración o actualización).

Por ello, para medir antigüedad operativa se utiliza **customer_days_active**, no la comparación directa **transaction_date < created_at**.

2.7. Normalización y geografía

- **País/Región:**
 - region_normalized (residencia del cliente) y country_normalized (país de operación) están **normalizados** (nombres consistentes).
 - Existe bandera **is_international** para detectar operaciones fuera del país del cliente.
 - **Codificaciones en minúsculas/estandarizadas** en: entry_mode, transaction_result, merchant_category.
-

2.8. Conclusión

El dataset final presenta **excelente completitud (0 nulos)**, **unicidad por transacción**, valores monetarios **dentro de rangos plausibles** y categorías **bien estandarizadas**.

La única consideración de calidad es la **interpretación de created_at**, que no debe usarse como “fecha de alta original” al comparar con transaction_date; para antigüedad se dispone de **customer_days_active**.

Con estas garantías, el conjunto es adecuado para analizar el **comportamiento del fraude** por riesgo, tarjeta, comercio, geografía y resultado de la operación.

3. Principales transformaciones

3.1. Consolidación de fuentes

- **Unión** de los dos orígenes (clientes + transacciones) mediante **customer_id**.
 - Se conservó **toda la casuística transaccional** (join 1–N desde transacciones), generando un único dataset consolidado.
 - Resultado de la consolidación: **dataset_limpio_y_transformado.csv**.
-

3.2. Conversión de tipos

- **Fechas** a datetime: transaction_date, created_at.
 - **Numéricos**: amount y risk_score a tipo numérico.
 - **Categóricas**: risk_level, entry_mode, transaction_result, card_type, merchant_category, region_normalized, country_normalized.
-

3.3. Normalización y estandarización de categorías

- **Etiquetas consistentes** en:
 - risk_level → **Alto / Medio / Bajo**
 - entry_mode → **chip / contactless / ecommerce / magstripe / wallet**
 - transaction_result → **approved / declined / pending**
 - card_type → **AMEX / DISCOVER / MASTERCARD / VISA**
 - **Geografía normalizada:**
 - region_normalized = país/región del cliente (normalizado).
 - country_normalized = país de la operación (normalizado).
-

3.4. Variables temporales derivadas

- Desde transaction_date se generaron:
 - **month** (1–12), **weekday** (0–6), **hour/transaction_hour**, **month_year** (etiqueta Y–M).
 - Objetivo: habilitar **series temporales**, estacionalidad y análisis por franjas horarias.
-

3.5. Antigüedad operativa del cliente

- **customer_days_active**: días de actividad del cliente disponibles en el dataset y utilizados para los análisis por madurez.
 - **Nota técnica:** Se intentó calcular la antigüedad restando transaction_date - created_at, pero se detectaron valores **negativos** en una parte significativa de los registros. Esto ocurre porque created_at no representa la fecha de alta original, sino una fecha más reciente (posiblemente de migración o actualización de sistema). Por ello, se decidió **no utilizar esa resta directa** y en su lugar emplear la variable customer_days_active, que refleja correctamente los días de actividad acumulados del cliente.
-

3.6. Campos de tarjeta y comercio

- **Tarjeta:** card_type (marca), card_bin6, card_last4, card_masked, card_expiry (soporte analítico/operativo).
 - **Comercio:** merchant y **merchant_category** (electronics, fashion, gaming, grocery, pharmacy, restaurant, travel).
-

3.7. Controles de calidad posteriores a las transformaciones

Se identificaron **filas incompletas** en las que únicamente figuraba el campo TransactionID, mientras que el resto de columnas se encontraban vacías. Estas filas se eliminaron al no aportar información analítica.

- **Nulos:** 0 nulos en el dataset (todas las columnas).
 - **Duplicados** en transaction_id: 0.
 - **Importes (amount):** rango [2.35 ; 656.13], sin valores negativos.
 - **Fechas:** transaction_date en [2024-10-11 ; 2025-10-12]; created_at en [2024-06-17 ; 2026-02-02].
Se documenta la **posible desalineación semántica** de created_at (no usar como “alta original”).
-

3.8. Filtros y decisiones de conservación

- Se conservaron **todas las categorías** de entry_mode, transaction_result, card_type y merchant_category para mantener representatividad.
-

3.9. Agrupaciones clave empleadas en el EDA (sin modificar el dataset)

En este notebook se desarrollaron distintos **análisis cruzados y resúmenes estadísticos** orientados a identificar patrones de fraude en la base de datos.

Cada cruce se diseñó con una finalidad analítica específica, priorizando la relación entre variables clave como is_fraud, merchant_category, entry_mode, risk_level, y risk_score.

Tablas y cruces analíticos realizados

Tabla / Variable derivada	Descripción	Métricas calculadas	Enfoque
res_merchant	Agrupación por categoría de comercio (merchant_category).	Número total de operaciones, fraudes, tasa de fraude, importe medio/media no, riesgo medio, peso de operaciones y fraudes.	Identificar los sectores con mayor exposición al fraude .
res_merchant_by_type	Agrupación doble por tipo de comercio y tipo de transacción (fraude/no fraude) .	Importe medio, importe mediano, riesgo medio.	Comparar el comportamiento económico de fraudes vs no fraudes por sector.
top_by_rate	Ranking de las categorías con mayor tasa de fraude .	Tasa de fraude ordenada descendente.	Determinar los comercios prioritarios en vigilancia antifraude .
plot_df (comparativa fraude vs no fraude)	Filtrado de los 10 principales comercios por volumen.	Importe medio agrupado por tipo de transacción.	Analizar la variación de importes medios entre fraudes y operaciones legítimas.
fraude_por_entry	Agrupación por modo de entrada (entry_mode) .	Porcentaje de fraude por tipo de operación.	Evaluar qué canales presentan mayor incidencia de fraude .

Tabla / Variable derivada	Descripción	Métricas calculadas	Enfoque
freq_entry	Frecuencia total de operaciones por modo de entrada.	Porcentaje de uso.	Relacionar el volumen operativo con la tasa de fraude .
analisis_entry	Unión de fraude_por_entry y freq_entry.	Tasa de fraude (%) y frecuencia de uso (%).	Permite visualizar la relación entre popularidad del canal y su vulnerabilidad .
tabla_entry_risk	Cruce entre entry_mode y risk_level mediante tabla cruzada.	Distribución porcentual de niveles de riesgo (Alto, Medio, Bajo) por canal.	Determinar qué modos de entrada concentran mayor proporción de riesgo alto o medio .
resumen_entry	Síntesis final de prevalencia y volumen por entry_mode.	Número de operaciones y tasa de fraude.	Combinar exposición (volumen) y prevalencia (fraude) en una misma visión global.

Objetivo de las agrupaciones y cruces

El propósito de las agrupaciones desarrolladas en este notebook es ofrecer una **visión analítica integral del fraude**, combinando indicadores de frecuencia, valor económico y riesgo operativo.

🎯 Objetivos específicos:

1. Detectar patrones de fraude por tipo de comercio

- Identificar los sectores con **mayor tasa y volumen de fraudes**.
- Determinar si los fraudes se concentran en operaciones de alto importe o riesgo.

2. Evaluar la influencia del canal de entrada (entry_mode) en la probabilidad de fraude

- Analizar qué métodos (wallet, magstripe, chip, etc.) presentan **mayor vulnerabilidad**.
- Relacionar el uso masivo de ciertos canales con la efectividad de los controles.

3. Examinar la relación entre riesgo y fraude

- A través de la tabla entry_mode vs risk_level, se identifican los **modos con mayor proporción de riesgo medio-alto**, reforzando el análisis predictivo.

4. Relacionar volumen de operaciones y prevalencia de fraude

- Mediante resúmenes combinados, se detectan canales con **alta actividad operativa y elevada incidencia de fraude**, priorizando la mitigación en esos puntos.

5. Visualizar hallazgos mediante gráficos y rankings claros

- Las visualizaciones (heatmaps, barras comparativas y rankings) aportan una **visión inmediata de las áreas críticas**, favoreciendo la toma de decisiones basada en datos.

Estos análisis constituyen la base de los **gráficos e insights** que se desarrollan posteriormente en el informe.

Resultado: tras estas transformaciones, el dataset quedó **coherente, completo (0 nulos)** y con variables **normalizadas y derivadas** que permiten explicar el fraude por **canal, riesgo, tarjeta, comercio, geografía, temporalidad y resultado** sin pérdida de información.

Visualizaciones Exportadas (EDA)

Este apartado recoge las visualizaciones generadas durante el análisis exploratorio del dataset de operaciones fraudulentas. Las gráficas se organizan por temática para facilitar una lectura coherente del comportamiento del fraude en el tiempo, por canal de entrada, por tipo de comercio y por segmentos relevantes.

Distribución temporal de operaciones

Operaciones fraudulentas por día

- Archivo: Operaciones_fraudulentas_dia.png
- Permite identificar patrones diarios en la actividad fraudulenta.
- Ayuda a detectar picos anómalos.

Operaciones fraudulentas por hora del día

- Archivo: Operaciones_fraudulentas_hora_dia.png
- Muestra las franjas horarias con mayor incidencia de fraude.

Operaciones fraudulentas por semana

- Archivo: Operaciones_fraudulentas_semana.png
- Evalúa tendencias semanales y posibles aumentos recurrentes.

Distribución de fraudes por mes

- Archivo: Distribucion_fraudes_por_mes.png
 - Mide la estacionalidad del fraude por mes.
 -
-

Análisis por modo de entrada (Entry Mode)

Distribución general por entry mode

- Archivo: Distribucion_general_entry_mode.png
- Muestra cómo se distribuyen las operaciones según el método de entrada.

Entry mode vs nivel de riesgo

- Archivo: entrymode_risklevel.png
 - Permite ver si ciertos modos de entrada concentran operaciones de mayor *risk_score*.
-

Análisis por tipo de comercio y categoría

Operaciones por tipo de tarjeta y comercio

- Archivo: operaciones_tipo_tarjeta_comercio.png
- Relaciona tipos de tarjeta con categorías de comercio.

Importe medio por categoría (Fraude vs No fraude)

- Archivo: importe_medio_fraude_vs_no_fraude_top10.png
- Compara los importes medios entre operaciones fraudulentas y no fraudulentas.

Fraude por tipo de tarjeta

- Archivo: fraude_por_tipo_tarjeta.png
- Identifica qué tarjetas presentan mayor proporción de fraude.

Top países por volumen de fraude

- Archivo: top_paises_volumen_operaciones_fraude.png
- Muestra los países con mayor concentración de operaciones fraudulentas.

Distribución geográfica del risk_score

- Archivo: geo_risk_score_violin.png
- Analiza la distribución del *risk_score* por región o país.

Comparativas de fraude y análisis avanzado

Comparativa general de fraude

- Archivo: comparativa_fraude.csv
- Resumen estadístico del comportamiento global de fraude vs no fraude.

Tasa de fraude por nivel de riesgo

- Archivos:
 - amount_por_risk_level.csv
 - amount_por_risk_level.png
- Relacionan riesgo y volumen económico.

Distribución del fraude

- Archivo: distribucion_fraude.csv
- Tabla descriptiva complementaria al análisis visual.

Tasa de fraude por merchant (Top N)

- Archivo: tasa_fraude_topN_merchant.png
- Identifica comercios con mayor tasa relativa de fraude.

Fraude por resultado de transacción

- Archivo: fraude_por_resultado_transaccion.png
 - Analiza el fraude según el resultado de la operación.
-



Resumen estadístico

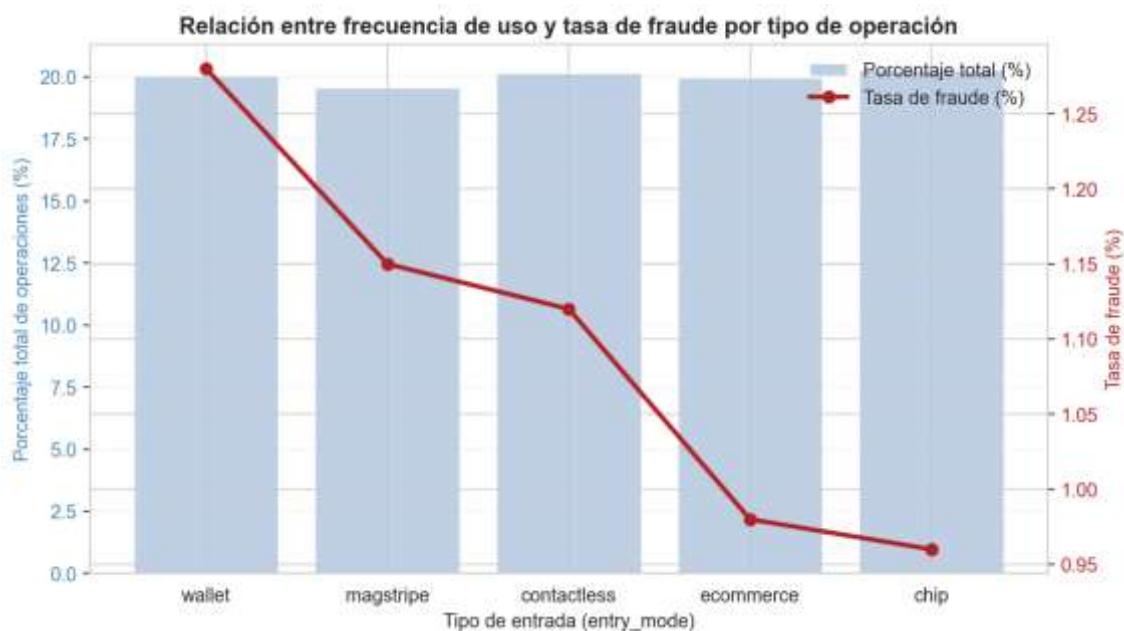
- Archivo: resumen_estadistico.csv
 - Incluye estadísticas descriptivas de las principales variables numéricas del dataset.
-
-

4. Análisis visualizaciones exploratorias

A continuación, se presentan las principales visualizaciones del análisis exploratorio, centradas en la detección de patrones de fraude según canal, tipo de tarjeta, riesgo, comercio, resultado y evolución temporal.

Cada gráfico incluye una breve interpretación (“insight”) basada en los datos del EDA.

-
1. Relación entre frecuencia de uso y tasa de fraude según el modo de entrada (entry_mode)



El gráfico muestra la **relación entre el porcentaje total de operaciones y la tasa de fraude** para cada tipo de modo de entrada (entry_mode), representando así tanto la **popularidad de uso** como el **riesgo relativo** asociado a cada canal de transacción.

Esta visualización permite identificar **qué modo de entrada concentra mayor volumen de uso y en cuáles se presenta una mayor proporción de fraude**.



Insights:

1. wallet (billeteras digitales)

- Es el canal con **mayor frecuencia de uso ($\approx 20\%$)** y también muestra la **tasa de fraude más elevada ($\approx 1.28\%$)**.
- Indica que, aunque es el medio más utilizado, **representa un foco importante de riesgo** y requiere un seguimiento reforzado.

2. magstripe y contactless

- Mantienen una participación similar ($\approx 19\text{--}20\%$ de operaciones), con tasas de fraude algo menores ($\approx 1.1\text{--}1.15\%$).
- Se consideran **canales moderadamente seguros**, aunque su alto volumen los convierte en **zonas potenciales de exposición**.

3. ecommerce

- A pesar de su **volumen de uso cercano al 20%**, presenta una **tasa de fraude inferior ($\approx 0.98\%$)**.
- Esto sugiere que **las medidas antifraude implementadas en pagos online están resultando efectivas**, o bien que las transacciones sospechosas son detectadas antes de completarse.

4. chip (EMV)

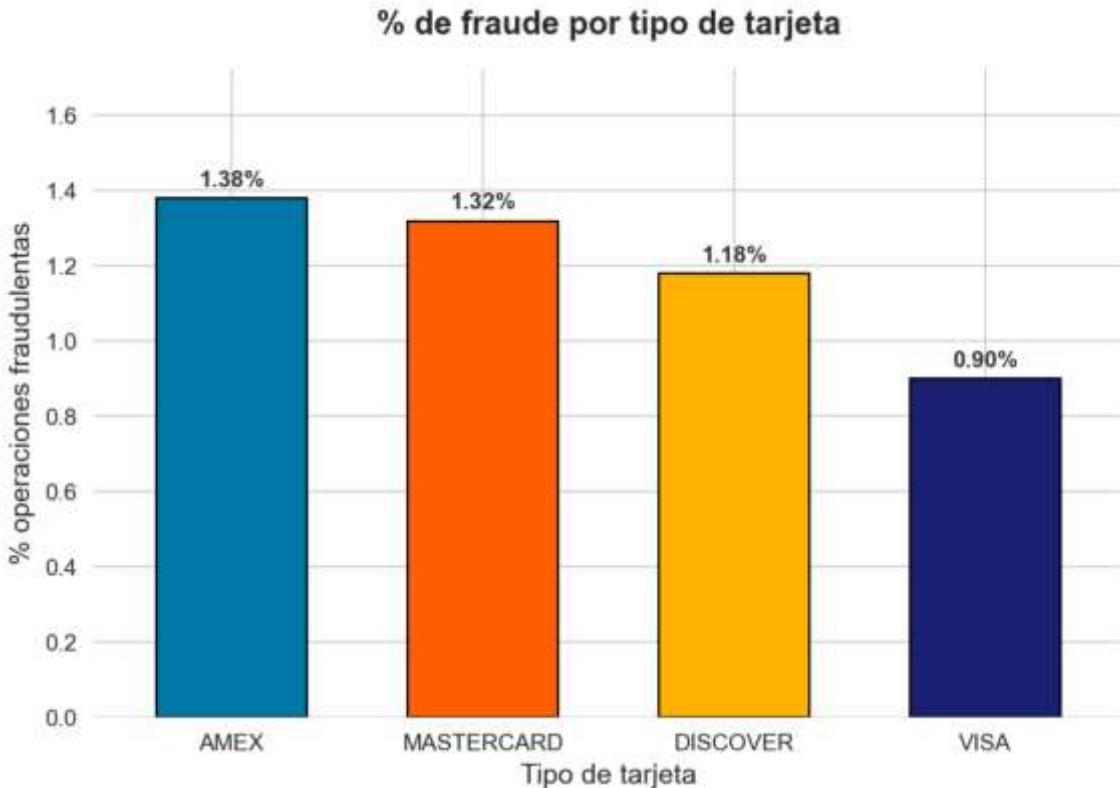
- Es el canal **más seguro**, con la **tasa de fraude más baja ($\approx 0.96\%$)**.
- Este resultado es consistente con la tecnología de chip, que incorpora **mecanismos criptográficos más robustos** frente a clonaciones y usos indebidos.



Conclusiones

- Existe una **tendencia descendente clara**: a mayor seguridad tecnológica del método de entrada, menor es la tasa de fraude.
- Sin embargo, los canales **más cómodos o rápidos para el usuario** (wallet y contactless) tienden a ser **más susceptibles a intentos de fraude**.
- Estos resultados pueden utilizarse para **priorizar controles o campañas de prevención** según el canal de entrada más expuesto.

2. Porcentaje de fraude por tipo de tarjeta (card_type)



El gráfico muestra la **tasa de operaciones fraudulentas (%)** según el tipo de tarjeta utilizada, permitiendo identificar **qué marcas presentan mayor vulnerabilidad** frente al fraude.

📈 Insights:

1. AMEX

- Registra la **tasa de fraude más alta ($\approx 1.38\%$)**, superando ligeramente al resto.
- Este resultado podría estar asociado a su **perfil de cliente más internacional** y al uso frecuente en **entornos digitales o de alto importe**, donde la exposición al fraude suele ser mayor.

2. MASTERCARD

- Presenta una tasa de fraude similar ($\approx 1.32\%$).
- Aunque ampliamente distribuida, su uso masivo en distintos comercios la convierte en un **objetivo habitual de intentos de fraude**.

3. DISCOVER

- Muestra una tasa intermedia ($\approx 1.18\%$), manteniendo un comportamiento relativamente estable.
- Indica una **exposición moderada**, posiblemente por su menor volumen de operaciones globales.

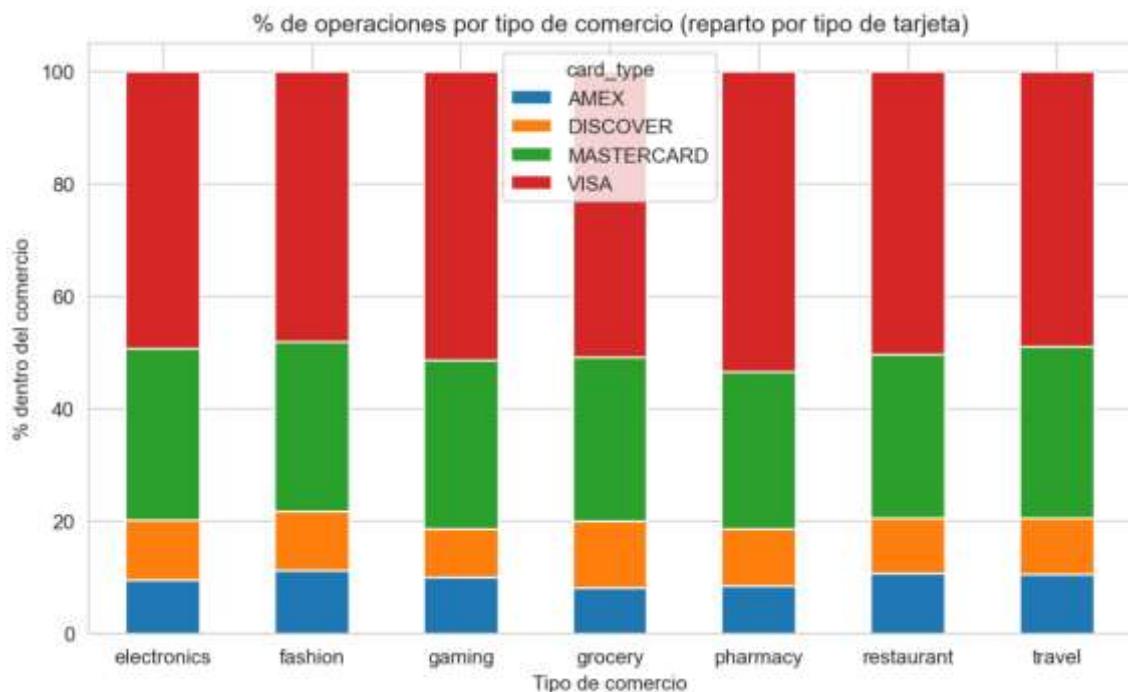
4. VISA

- Es la tarjeta **con menor tasa de fraude ($\approx 0.90\%$)**, lo que sugiere **mayor efectividad de sus controles antifraude y autenticación**.
- Puede relacionarse con una **amplia red de detección temprana** y adopción de estándares EMV en la mayoría de sus operaciones.

Conclusiones

- Se observa una **brecha de riesgo de aproximadamente 0.5 puntos porcentuales** entre la marca más vulnerable (AMEX) y la más segura (VISA).
 - Las diferencias pueden deberse tanto a **estrategias de control de riesgo propias de cada red** como a la **naturaleza de los clientes y comercios asociados**.
 - Este análisis permite **priorizar la vigilancia y los controles adicionales** en los tipos de tarjeta con mayor tasa de fraude detectada.
-

3. Tipo de comercio x tipo de tarjeta (merchant_category x card_type)



El gráfico muestra el **reparto porcentual de las operaciones por tipo de tarjeta dentro de cada categoría de comercio**, permitiendo analizar la **preferencia de uso de las distintas marcas** (AMEX, DISCOVER, MASTERCARD y VISA) según el sector.

Insights:

1. Predominio de VISA

- VISA concentra la **mayor proporción de operaciones** en todos los sectores analizados, superando aproximadamente el **45–50% del total de transacciones**.
- Esto la posiciona como la **tarjeta más utilizada por los clientes**, reflejando su amplia aceptación y cobertura internacional.

2. MASTERCARD con fuerte presencia

- Representa cerca de un **30–35% de las operaciones**, manteniendo un peso significativo en todos los tipos de comercio.
- Su cuota estable sugiere **una alta penetración en comercios generalistas y minoristas**.

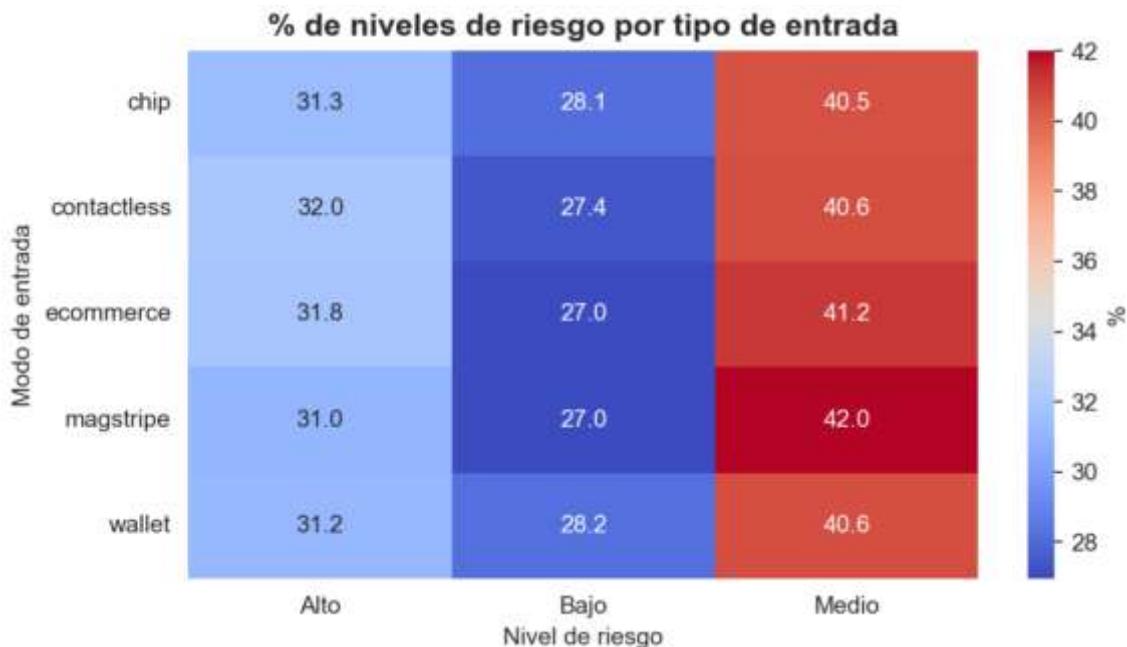
3. AMEX y DISCOVER con uso limitado

- Ambas presentan participaciones menores, en torno al **8–10% cada una**, con ligeras variaciones por sector.
- El uso de **AMEX** tiende a ser más elevado en **moda, viajes y restauración**, sectores asociados a **gasto medio-alto y clientes internacionales**.
- **DISCOVER**, en cambio, mantiene una presencia más homogénea pero reducida, concentrada en **pagos online o de nicho**.

Conclusiones

- El mercado muestra una **clara concentración de operaciones en las redes VISA y MASTERCARD**, lo que puede implicar una **mayor exposición agregada al riesgo** en estos emisores.
 - Los comercios de **viajes, restauración y moda** son los que presentan **mayor diversidad de marcas**, posiblemente por su **perfil de cliente internacional**.
 - Esta distribución es útil para **evaluar la exposición al fraude por tipo de producto y sector**, especialmente al cruzarla con la **tasa de fraude por tarjeta** obtenida en el análisis anterior.
-

4. Modo de entrada x nivel de riesgo (entry_mode x risk_level)



El mapa de calor muestra la **distribución porcentual de los niveles de riesgo (Alto, Medio, Bajo)** según el **modo de entrada de la transacción (entry_mode)**, ofreciendo una visión clara de qué canales presentan mayor concentración de riesgo.

Insights:

1. **Predominio del nivel de riesgo medio ($\approx 40\text{--}42\%$)**

- En todos los modos de entrada, **el riesgo medio es el más frecuente**, lo que sugiere una **exposición moderada generalizada** en las operaciones.
- Este patrón indica que, aunque no hay un riesgo extremo predominante, **los canales mantienen una vulnerabilidad constante**.

2. **magstripe muestra el mayor riesgo relativo (42% medio)**

- Las transacciones con **banda magnética** concentran el porcentaje más alto de operaciones en nivel de riesgo medio.
- Esto confirma que **los métodos más antiguos presentan mayor vulnerabilidad** frente a fraude o anomalías.

3. ecommerce y wallet mantienen un perfil equilibrado

- Ambos presentan distribuciones muy próximas al promedio (\approx 41% medio, 31% alto, 27–28% bajo).
- Muestran un **riesgo moderado y estable**, propio de canales digitales que aplican validaciones adicionales.

4. chip y contactless exhiben perfiles de riesgo controlados

- Con una proporción similar de riesgo alto (\approx 31–32%) y bajo (\approx 27–28%).
- Esto respalda la **efectividad de las tecnologías EMV y sin contacto** en la reducción del riesgo operativo.

Conclusiones

- Existe una **consistencia en la distribución de niveles de riesgo entre canales**, con predominio del **riesgo medio**, lo que sugiere una exposición homogénea.
 - **Magstripe** sigue siendo el **modo de entrada más sensible**, lo que justifica **acciones preventivas específicas o migración tecnológica** hacia métodos más seguros.
 - Este análisis permite **priorizar controles antifraude por canal**, alineando los esfuerzos de mitigación con la exposición al riesgo observada.
-

💡 5. Porcentaje de fraude por resultado (transaction_result)



El gráfico representa el **porcentaje de operaciones fraudulentas** según el **resultado final de la transacción** (aprobada, pendiente o declinada).

Permite evaluar la **efectividad de los filtros de autorización** y entender en qué fase se concentran las tentativas de fraude.

📈 Insights:

1. **Transacciones declinadas → mayor tasa de fraude (~2.07%)**

- La mayoría de las operaciones fraudulentas se concentran en las **transacciones rechazadas**.
- Esto indica que **los controles antifraude están actuando correctamente**, detectando comportamientos sospechosos antes de que la operación se autorice.

2. Transacciones pendientes → riesgo intermedio ($\approx 1.28\%$)

- Los casos en espera representan una fracción relevante de operaciones potencialmente fraudulentas.
- Es un grupo a **monitorizar con especial atención**, ya que podrían incluir transacciones legítimas bajo revisión o intentos de fraude no confirmados.

3. Transacciones aprobadas → menor tasa de fraude ($\approx 1.01\%$)

- El porcentaje más bajo se encuentra entre las **operaciones autorizadas**, lo que evidencia una **eficiente capa de prevención y filtrado previo**.
- Sin embargo, sigue existiendo una fracción de fraude residual que **consigue superar los filtros iniciales**.

Conclusiones

- El análisis demuestra una **buenas eficacia de los mecanismos de detección temprana**, que bloquean la mayoría de los intentos antes de su aprobación.
 - No obstante, el **1% de fraude aprobado** evidencia la necesidad de **reforzar controles en tiempo real y post-autorización**.
 - En conjunto, este comportamiento refleja una **operativa madura de prevención**, pero con **margen de mejora en los casos limítrofes** (pendientes o borderline).
-

⌚ 6. Evolución temporal de transacciones (día)



El gráfico presenta la **distribución diaria de las operaciones fraudulentas dentro de cada mes**, lo que permite analizar la **variabilidad del fraude a lo largo del ciclo mensual** y detectar posibles patrones de comportamiento.

📈 Insights:

1. Mayor concentración entre los días 1–5 y 24–27 del mes

- Los **inicios y finales de mes** muestran un aumento notable de operaciones fraudulentas (picos de 19 a 21 casos).
- Este patrón puede relacionarse con **cierres de facturación, ciclos de nómina o movimientos financieros habituales**, momentos donde los defraudadores buscan camuflar sus operaciones entre transacciones legítimas.

2. Descensos pronunciados hacia los días 10–12 y 18–22

- En la parte media del mes se observan varios descensos (con mínimos de 3 y 8 casos).
- Esto puede indicar **periodos de menor actividad económica** o una **mejor eficacia de los sistemas de control** en esos intervalos.

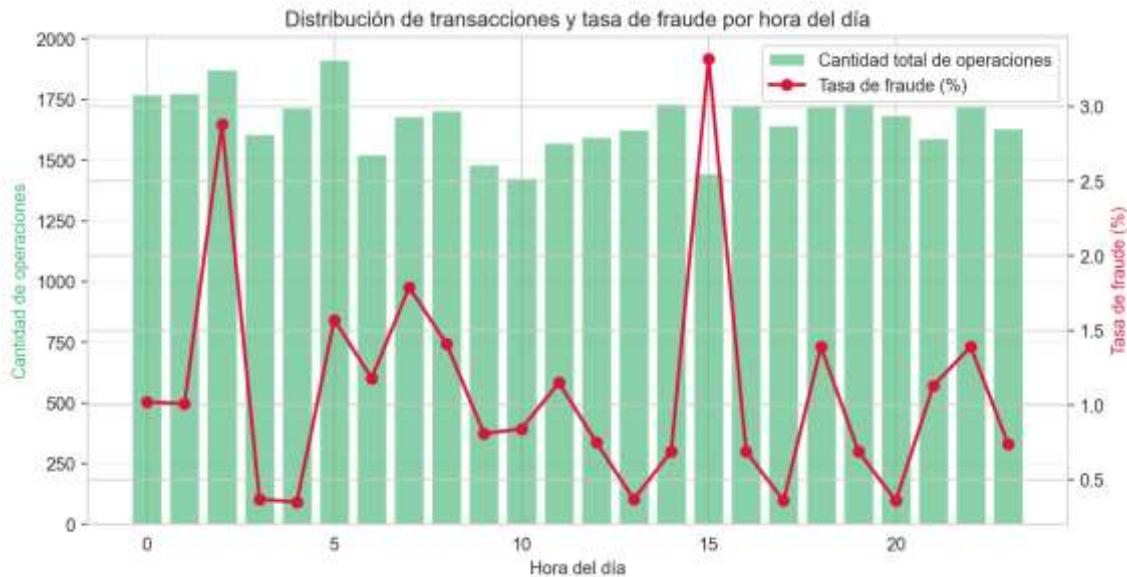
3. Actividad irregular pero con repeticiones cíclicas

- Se aprecia un **ritmo de repunte cada 8–10 días**, lo que podría reflejar **patrones de prueba o “testing” de fraudes recurrentes**.

Conclusiones

- El fraude presenta una **dinámica cíclica dentro del mes**, con **mayor intensidad en los extremos del periodo**.
 - Este comportamiento sugiere que los atacantes **aprovechan los momentos de mayor volumen transaccional** (cobros, pagos, cierres contables).
 - Se recomienda **reforzar los controles automáticos y alertas preventivas** durante los **primeros y últimos cinco días del mes**, cuando la exposición al riesgo es más elevada.
-

⌚ 7. Evolución temporal de transacciones (hora/día)



El gráfico combina la **cantidad total de operaciones por hora del día** (barras verdes) con la **tasa de fraude (%)** asociada (línea roja).

Este enfoque permite identificar **horas críticas del día** donde el volumen operativo y el fraude no siguen el mismo patrón.

📈 Insights:

1. Mayor tasa de fraude entre las 2:00 y las 3:00 h ($\approx 3\%$)

- A pesar del bajo volumen de operaciones, el porcentaje de fraude es muy alto.
- Este comportamiento sugiere **actividad fraudulenta concentrada en horas de baja supervisión**, posiblemente aprovechando la menor carga operativa de control.

2. Segundo pico relevante hacia las 15:00 h ($\approx 3\%$)

- Coincide con un horario de **alta actividad comercial**, lo que indica que los intentos de fraude **aumentan durante períodos de mayor flujo de transacciones**.

3. Franja de menor riesgo: entre las 10:00 y las 13:00 h

- Durante las horas laborales estándar, tanto el volumen como la tasa de fraude desciden.
- Esto puede asociarse a **mejoras en los controles activos o mayor atención de los equipos de revisión manual**.

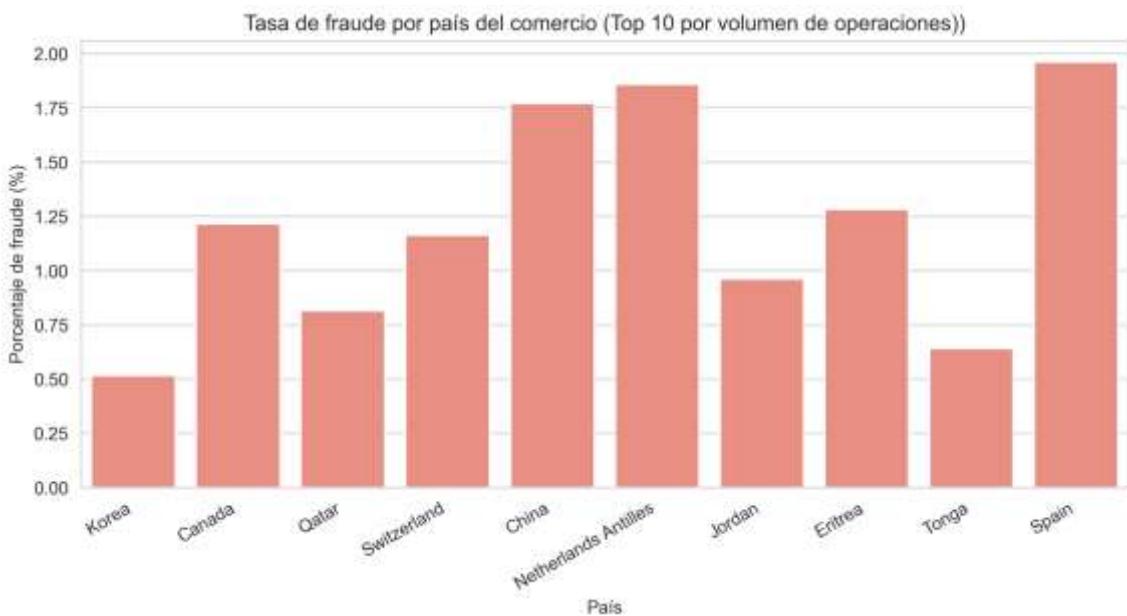
4. Alta actividad de operaciones entre 8:00 y 20:00 h, pero sin correlación directa con el fraude.

- La mayor parte del volumen se concentra en horas diurnas, aunque la tasa de fraude **permanece moderada**.
- Indica que **el riesgo no depende solo del volumen**, sino de **comportamientos específicos en franjas concretas**.

Conclusiones

- El fraude muestra **dos patrones horarios bien diferenciados**:
 - **Nocturno (2–3 h)**: fraude oportunista en baja actividad.
 - **Vespertino (15 h)**: fraude en picos de alto tráfico comercial.
 - Se recomienda **reforzar los sistemas de detección en tiempo real** durante estas franjas críticas, aplicando controles dinámicos por hora.
 - El análisis demuestra que **la vigilancia continua no debe ser homogénea**, sino **adaptativa según el riesgo horario**.
-

⌚ 8. Tasa de fraude por país del comercio.



El gráfico muestra el **porcentaje de operaciones fraudulentas por país del comercio**, considerando los **10 países de los comercios con mayor volumen de transacciones**. El objetivo es identificar **regiones con mayor propensión al fraude**, así como evaluar si existe **correlación entre el volumen transaccional y la incidencia delictiva**.

📈 Insights:

1. Mayores tasas de fraude en:

- **Netherlands Antilles (1.9%), Spain (1.95%) y China (1.75%).**
 - Estos países presentan **niveles de fraude muy por encima del promedio** del grupo.
 - En el caso de España, la cifra elevada podría estar relacionada con **el alto volumen operativo local y la diversidad de canales de pago**.
 - En China y Antillas Neerlandesas, podría influir la **exposición a transacciones internacionales y comercio electrónico**.

2. Riesgo medio en:

- **Canada (1.2%), Switzerland (1.15%), Eritrea (1.28%) y Qatar (0.8%).**
 - Estos países mantienen una **actividad moderada pero estable**, lo que indica **riesgo controlado pero persistente**.

3. Menor exposición en:

- **Korea (0.5%), Tonga (0.65%) y Jordan (0.95%).**
→ Estos mercados reflejan **baja incidencia de fraude**, posiblemente por **marcos regulatorios más estrictos o menor presencia transaccional internacional**.

Conclusiones

- El comercio con fraude **no se distribuye de forma uniforme entre los países**: algunos mercados presentan **tasas elevadas sin necesariamente tener más volumen**, lo que evidencia **factores específicos de vulnerabilidad**.
 - Se recomienda **analizar la naturaleza de las operaciones internacionales** en países de alto riesgo (p. ej., **España, China y Antillas Neerlandesas**) para identificar **posibles patrones de fraude transfronterizo**.
 - Este análisis geográfico es clave para **priorizar controles adaptados al contexto regional**, reforzando la detección temprana en los países de los comercios con **mayor exposición o comportamiento atípico**.
-
-

5. Conclusión general del proyecto y utilidad

1. Síntesis global del flujo de trabajo

El proyecto se ha estructurado en cuatro notebooks encadenados que conforman un flujo de trabajo completo y reproducible de Análisis Exploratorio de Datos (EDA) aplicado a operaciones financieras con etiqueta de fraude:

1. **01_EDA_PRELIMINAR**

- Carga de los ficheros originales de clientes y operaciones.
- Verificación de la raíz del proyecto y definición de rutas de trabajo.
- Comprobación de columnas comunes y unión de los datasets a través de la clave de cliente.
- Generación de un primer dataset consolidado y guardado en la capa DATA_RAW_OUTPUT.

2. **02_EDA_LIMPIEZA_TRANSFORMACION**

- Revisión sistemática de tipos de datos, nulos y valores atípicos.
- Normalización de valores vacíos en variables categóricas ("", ' ', null, None, NA, etc.) a NaN.
- Eliminación o tratamiento de filas sin información útil (p.ej., registros con solo transaction_id).
- Conversión de columnas de fecha y numéricas a tipos adecuados, garantizando coherencia temporal y monetaria.
- Creación de variables derivadas (por ejemplo, categorización de niveles de riesgo y otras variables auxiliares).
- Ordenación y homogeneización del orden de columnas.
- Guardado del **dataset limpio y transformado** en la capa DATA_OUTPUT/EDA.

3. **03_EDA_ANALISIS DESCRIPTIVO**

- Carga del dataset limpio y configuración global de parámetros de visualización.
- Definición de funciones de utilidad para guardar gráficos (.png) y tablas (.csv) de forma consistente.
- Análisis descriptivo por bloques:
 - **Distribución temporal** (día, hora, semana, mes).
 - **Entry mode / canal de entrada** y su relación con el riesgo.
 - **Tipo de comercio y tipo de tarjeta** (distribuciones cruzadas, porcentajes, top categorías).
 - **Distribución geográfica** y análisis de países con mayor volumen de fraude.

- **Comparativas fraude vs no fraude** (tanto en volumen como en importes medios).
- **Evaluación del risk_score** mediante deciles y niveles, analizando si el modelo discrimina bien el fraude.
- Exportación ordenada de todas las salidas a DATA/DATA_OUTPUT/EDA.

4. 04_INFORME_TECNICO_EDA PYTHON

- Integración narrativa de todo el trabajo previo en un informe estructurado.
- Descripción del origen de los datos, criterios de limpieza y principales transformaciones.
- Inclusión de las visualizaciones exportadas y su interpretación cualitativa.
- Redacción de síntesis y apartados explicativos orientados a negocio y a la toma de decisiones.

En conjunto, los cuatro notebooks constituyen un pipeline claro: **de datos crudos a informe analítico listo para stakeholders**, con trazabilidad completa de cada paso.

2. Principales hallazgos del análisis

A partir del dataset limpio y del conjunto de visualizaciones generadas, se pueden destacar las siguientes conclusiones de alto nivel:

1. Distribución temporal del fraude

- El desglose por día, semana, hora y mes permite detectar picos de actividad fraudulenta y patrones recurrentes.
- Estos patrones temporales son clave para reforzar recursos de monitorización en ventanas de mayor riesgo (franjas horarias concretas, determinados días de la semana o períodos del año).

2. Importancia del modo de entrada (entry mode)

- El análisis por entry_mode muestra que no todos los canales tienen el mismo comportamiento de riesgo.
- Los modos de entrada asociados a operaciones no presenciales o con menor verificación suelen concentrar una proporción relativamente mayor de fraude, lo que justifica reglas y límites específicos por canal.

3. Diferencias según tipo de comercio y categoría

- El estudio por categorías de comercio y por tipo de tarjeta permite identificar sectores especialmente sensibles al fraude.

- Las comparativas de **importe medio (fraude vs no fraude)** por categoría revelan que, en algunas verticales, las operaciones fraudulentas se asocian a tickets medios distintos de los regulares (más altos en ciertos segmentos, similares o incluso inferiores en otros), lo que ayuda a diseñar umbrales de alerta más precisos.
- El análisis de comercios con mayor tasa relativa de fraude (top N merchants) facilita la priorización de revisiones y medidas de mitigación específicas por comercio.

4. Dimensión geográfica del riesgo

- Las visualizaciones por país y la distribución geográfica del risk_score muestran que el fraude no se distribuye de forma homogénea.
- Algunos países o regiones concentran un mayor volumen de operaciones fraudulentas o un perfil de riesgo más elevado, lo que sugiere la necesidad de políticas diferenciadas por geografía (controles adicionales, límites por país, etc.).

5. Evaluación crítica del risk_score y sus deciles

- El análisis por deciles de risk_score y por niveles de riesgo revela que la tasa de fraude **no crece de forma perfectamente monótona** del decil 1 al decil 10.
- Se observan deciles intermedios (por ejemplo, el segundo y algunos deciles altos) con tasas de fraude comparables o incluso superiores a las del decil máximo.
- Esta falta de escalada perfecta sugiere que:
 - el modelo de scoring puede requerir recalibración,
 - podrían faltar variables explicativas relevantes, o
 - el patrón de fraude ha cambiado con el tiempo (drift).
- El resultado es especialmente valioso porque no solo describe el comportamiento del fraude, sino que **evalúa la calidad del modelo de riesgo existente**.

6. Comportamiento global fraude vs no fraude

- La comparación sistemática entre operaciones fraudulentas y no fraudulentas (en volumen, importe medio, distribución por canales, países, categorías y resultados de transacción) permite construir un perfil diferencial del fraude.
- Este perfil sirve como base para diseñar reglas, modelos predictivos y dashboards que se centren en los factores con mayor poder discriminante.

3. Utilidad del proyecto para negocio y para análisis de datos

Este proyecto no se limita a un ejercicio académico de EDA, sino que aporta una **base práctica y reutilizable** para equipos de operaciones, riesgo y analítica:

1. Base de datos limpia y documentada

- El dataset final se encuentra **limpio, normalizado y enriquecido** con variables derivadas (niveles de riesgo, categorías, etc.).
- La documentación de cada transformación facilita su reutilización en futuros proyectos (por ejemplo, modelos de clasificación de fraude, reporting recurrente o integración con herramientas como Power BI).

2. Repositorio de visualizaciones clave

- La carpeta DATA/DATA_OUTPUT/EDA actúa como un **catálogo de gráficos y tablas** que cubre las principales dimensiones del fraude: tiempo, canal, comercio, país, tipo de tarjeta, resultado de la operación y score de riesgo.
- Estas salidas pueden incorporarse directamente a informes ejecutivos, presentaciones internas o dashboards interactivos.

3. Marco analítico para la toma de decisiones

- El análisis muestra qué combinaciones de variables (entry mode, país, categoría de comercio, tipo de tarjeta, nivel de riesgo, etc.) son más relevantes para entender el fraude.
- Esto permite:
 - redefinir reglas de negocio,
 - ajustar límites de autorización,
 - priorizar investigaciones manuales,
 - y focalizar recursos en segmentos de mayor impacto.

4. Evaluación y mejora de modelos de riesgo

- El trabajo sobre los deciles de risk_score y la comparativa de tasas de fraude por nivel de riesgo ofrece una **evaluación objetiva del desempeño del score actual**.
- A partir de estos resultados se pueden plantear:
 - recalibraciones del modelo,
 - incorporación de nuevas variables,

- o incluso el diseño de modelos alternativos (por ejemplo, usando técnicas de machine learning sobre el dataset ya preparado).

5. Reproducibilidad y buenas prácticas de Data Analytics

- La separación en cuatro notebooks (preliminar, limpieza/transformación, análisis descriptivo e informe) sigue un esquema profesional de proyecto de datos.
- El uso de rutas relativas, funciones de guardado, estructura de carpetas y sintaxis clara en Markdown facilita que otro analista pueda entender, ejecutar y extender el proyecto sin depender de conocimiento tácito.

4. Cierre

En resumen, el proyecto consigue:

- transformar datos dispersos de clientes y operaciones en un **dataset integrado y de calidad**,
- realizar un **análisis exploratorio profundo** que revela patrones temporales, geográficos, de canal y de negocio en el fraude,
- y proporcionar una **base sólida para la toma de decisiones y el diseño de futuros modelos y dashboards**.

La combinación de limpieza rigorosa, análisis visual y evaluación crítica del risk_score convierte este trabajo en una herramienta de gran valor para cualquier área de fraude y riesgo que quiera **entender mejor su operativa y mejorar sus estrategias de prevención**.
