INFORME DE ANÁLISIS EXPLORATORIO DE DATOS

Detección de fraude en transacciones

Carlos Hernando
Repositorio: C4RL0S1515

Índice

1. Introducción

Este proyecto estudia un conjunto de transacciones realizadas con el objetivo de identificar patrones y factores asociados al fraude. El análisis se apoya en dos fuentes principales, el primer dataset contiene el histórico de transacciones, con información del importe, el comercio, la localización y atributos del cliente, y el segundo, un dataset socioeconómico agregado por código postal, que incorpora métricas de renta e indicadores derivados a nivel geográfico. La integración de ambas fuentes permite analizar el fraude no solo desde la perspectiva transaccional, sino también considerando el contexto económico del área donde se produce la operación.

La detección del fraude es un problema habitual en entornos financieros, donde la mayoría de operaciones son legítimas y los casos fraudulentos representan una proporción reducida, pero con alto impacto económico. Por ello, resulta clave entender cuándo y dónde se concentra el fraude, en qué tipos de comercios aparece con mayor frecuencia y si existen perfiles o condiciones de contexto que aumenten el riesgo. Este enfoque permite extraer conclusiones útiles a nivel descriptivo y, además, preparar una base sólida para etapas posteriores orientadas a la interpretación o a la construcción de modelos de detección.

El trabajo se organiza como un flujo completo de análisis exploratorio. En primer lugar, se realiza una revisión inicial de la estructura y calidad de los datos, tipos de variables, nulos, duplicados y coherencia. Posteriormente, se aplica un proceso de limpieza y estandarización para garantizar consistencia en formatos especialmente fechas y variables categóricas y se integran ambas fuentes mediante la clave zip.

Tras la unión, se ejecutan validaciones de integridad, control de discrepancias geográficas, verificación de registros, ausencia de duplicados y se construye un dataset final preparado para el EDA. Finalmente, se generan variables derivadas para reforzar el análisis, incluyendo variables temporales, demográficas y ratios económicos junto con segmentaciones.

La variable objetivo del estudio es **is_fraud**, definida como 1 para transacciones fraudulentas y 0 para transacciones legítimas. Dado el fuerte desbalance esperado en este tipo de problemas, las comparativas se interpretan principalmente mediante tasas y proporciones, evitando conclusiones basadas únicamente en recuentos absolutos.

El enfoque metodológico combina estadística descriptiva y visualización, priorizando medidas robustas para describir el comportamiento central sin que los valores extremos distorsionen la interpretación. En variables geográficas y categóricas de alta cardinalidad, las comparaciones se realizan controlando el volumen de observaciones, con el fin de evitar conclusiones condicionadas por categorías con pocos registros.

El alcance de esta fase se centra en caracterizar patrones temporales, geográficos, transaccionales y socioeconómicos asociados al fraude, dejando el dataset preparado para etapas posteriores de interpretación más profunda o modelización predictiva.

2. Objetivos del proyecto

2.1 Objetivo general

Desarrollar un análisis exploratorio de datos sobre un conjunto integrado de transacciones bancarias y variables socioeconómicas agregadas por código postal, con el fin de identificar patrones y factores asociados al fraude. El objetivo es caracterizar el fenómeno desde varias dimensiones y dejar un dataset final consistente, documentado y preparado para fases posteriores.

2.2 Objetivos específicos

Análisis preliminar y calidad del dato

Revisar la estructura de ambas fuentes, dimensiones, tipos de variables, rangos y formatos, detectando posibles inconsistencias que puedan afectar a cálculos y visualizaciones. Además, de comprobar la calidad del dataset en términos de valores nulos, duplicados y coherencia general, con el objetivo de asegurar que el análisis posterior se realiza sobre información fiable.

Limpieza, estandarización e integración de fuentes

Estandarizar variables clave como fechas, categorías, tipos numéricos y eliminar atributos no informativos o redundantes para reducir ruido analítico. La integración del histórico de transacciones con la información socioeconómica por zip, validando la consistencia del merge y la coherencia geográfica cuando corresponda. Por ultimo generar un dataset final depurado que conserve trazabilidad del proceso y facilite la reproducibilidad del proyecto.

Creación de variables derivadas y segmentación analítica

Se crean variables derivadas temporales mes, día de la semana, hora, franja horaria y fin de semana vs laborable para estudiar patrones de ocurrencia del fraude. También se crear variables demográficas edad y grupos de edad y ratios socioeconómicos taxable_share, amt_vs_avg_agi, income_level para capturar contexto económico y comparativas relativas, y verificar que las transformaciones y ratios no introducen valores inválidos que comprometan la consistencia del análisis.

Análisis exploratorio orientado a objetivos y preguntas clave

El análisis exploratorio se estructura para buscar responder de forma explícita a las siguientes preguntas:

- **¿Cuándo ocurre más fraude?** Se analizan las variaciones por mes, hora, día de la semana y franja horaria, priorizando tasas y proporciones frente a recuentos absolutos.
- **¿Dónde ocurre más fraude?** Se evalúan diferencias por estado y ciudad, con control de volumen, para evitar conclusiones sesgadas por categorías con pocos registros.

- **¿En qué tipo de comercio ocurre más fraude?** Se analiza el fraude por categoría del comercio y, cuando proceda, por nombre del comercio, aplicando filtros de volumen y comparando tasas de fraude por grupo.
- **¿En qué perfiles aparece con mayor frecuencia?** Se revisan patrones por edad, género, ocupación y variables de contexto como población de la ciudad, comparando distribuciones entre clases.
- **¿Influye el contexto económico?** Explorar la relación entre fraude y variables como avg_agi, avg_total_income, avg_taxable_income, income_level, y ratios como amt_vs_avg_agi y taxable_share.

Síntesis y conclusiones

Se resumen los hallazgos más relevantes, destacando variables y patrones con mayor capacidad descriptiva del fraude. Así como documentar limitaciones del análisis desbalance de clase y alta cardinalidad, dejando preparado el informe final y las fases predictiva posteriores.

3. Metodología

3.1 Enfoque general

El proyecto se desarrolla siguiendo un flujo analítico estructurado para garantizar trazabilidad y reproducibilidad, análisis preliminar de las fuentes, limpieza y estandarización, integración de datasets, creación de variables derivadas y, finalmente, análisis exploratorio de datos mediante técnicas univariantes, bivariantes, correlacionales y temporales.
El EDA se orienta a responder preguntas clave del problema: cuándo ocurre más fraude tendencia mensual, patrón horario y semanal, dónde se concentra estado, ciudad, código postal, controlando el volumen, en qué tipo de comercio aparece con mayor frecuencia en qué perfiles se observa mayor riesgo edad, género, ocupación, tamaño de ciudad y si el contexto económico aporta alguna señal.

3.2 Fuentes de datos e integración

El análisis se basa en dos fuentes principales:

- **Dataset del histórico de transacciones**: incluye variables transaccionales y de contexto, importe, comercio, localización, atributos del cliente y variable objetivo is_fraud.
- **Dataset socioeconómico agregado por ZIP:** comprende métricas de renta y variables derivadas a nivel geográfico.

La integración se realiza mediante la clave zip, aplicando un left join para conservar todas las transacciones, aunque no exista información socioeconómica para algún

código postal. Tras la unión, se ejecutan validaciones para asegurar consistencia e integridad del dataset consolidado.

Después de la integración por zip, se realizan varias validaciones para garantizar la calidad del dataset consolidado, se comprueba la coherencia geográfica entre el estado de la transacción y el estado asociado al zip en la tabla socioeconómica, cuantificando posibles discrepancias y aplicando correcciones cuando procede, se verifica el número de filas antes y después del merge para confirmar que no se pierden transacciones, se revisa la existencia de duplicados que pudieran haberse generado durante la unión y por último, se analiza la presencia de valores nulos, tomando decisiones de filtrado en aquellas variables socioeconómicas consideradas esenciales para el análisis.

3.3 Preparación y controles de calidad

Antes del análisis exploratorio se revisa la calidad del dato para evitar errores en cálculos e interpretaciones:

- **Valores nulos:** identificación y cuantificación por columna.
- **Duplicados:** comprobación de filas repetidas.
- **Tipos de dato:** revisión de formatos para asegurar coherencia en operaciones numéricas, categóricas y temporales.

Además, se realiza un control específico en variables de tipo ratio para prevenir valores inválidos.

3.4 Ingeniería de variables y segmentaciones para el EDA

Con el objetivo de mejorar la interpretabilidad y permitir análisis segmentados, se generan variables derivadas en tres bloques:

Variables temporales

- trans_date (fecha de transacción), month (mes), day_of_week (día de la semana), hour (hora).
- moment_of_day (franja horaria: night, early morning, morning, afternoon, evening y late night.) para agrupar horas con un criterio interpretable.
- is_weekend para comparar días laborables vs fin de semana.

Variables demográficas

- age, calculada a partir de dob y la fecha de transacción.
- age_group, construida mediante intervalos para análisis por segmentos.

Variables económicas y ratios

- taxable_share, ratio entre renta imponible y renta total.
- amt_vs_avg_agi, ratio entre importe de transacción y renta media del zip.
- income_level, segmentación de renta media por cuartiles para comparar niveles socioeconómicos de forma ordenada.

Adicionalmente, se eliminan columnas con baja utilidad analítica o potencialmente redundantes identificadores, variables intermedias y campos no necesarios tras la construcción de derivadas, con el fin de reducir ruido en el EDA.

3.5 Estrategia de análisis exploratorio

El EDA se estructura en cuatro bloques complementarios:

1) Análisis univariante

- **Numéricas:** estadísticos descriptivos, histogramas y boxplots para identificar asimetrías y valores extremos.
- **Categóricas:** frecuencias y proporciones para detectar categorías dominantes y evaluar la distribución de cada variable.

2) Análisis bivariante respecto a is_fraud

- **Numéricas vs is_fraud:** boxplots por clase y tablas resumen por grupo para evaluar diferencias robustas entre fraude y no fraude.
- **Categóricas vs is_fraud:** tablas de contingencia normalizadas por categoría y gráficos de barras apiladas para variables de baja cardinalidad.

3) Correlaciones

Se construye una matriz de correlación para variables numéricas y un mapa de calor para identificar relaciones lineales relevantes y posibles dependencias entre variables.

4) Análisis temporal

Se estudia la evolución y patrones del fraude en tres escalas:

- **Mensual:** nº transacciones, nº fraudes y tasa de fraude mensual prioritaria para comparar meses con distinto volumen.
- **Intradía:** tasa de fraude por hora y por franja horaria.
- **Semanal:** tasa por día de la semana y comparación de fin de semana vs laborable.

3.6 Criterios para una interpretación robusta

Dado que el fraude es un evento minoritario, el análisis se interpreta principalmente con tasas y proporciones, evitando conclusiones basadas únicamente en recuentos absolutos.

En variables con distribuciones muy asimétricas o con outliers relevantes, se priorizan medidas robustas y, cuando es necesario, los valores extremos se limitan únicamente en los gráficos para mejorar la legibilidad y evitar que la distribución quede aplastada, sin modificar en ningún caso el dataset original utilizado para el análisis.

En variables categóricas de alta cardinalidad, se controla el volumen de observaciones para evitar interpretaciones sesgadas por categorías con pocos registros.

3.7 Herramientas y entorno de trabajo

El proyecto se ha desarrollado en Python, utilizando principalmente:

- **pandas y numpy** para manipulación y transformación de datos.
- **matplotlib** y **seaborn** para visualización.
- **Parquet** como formato de almacenamiento eficiente para datasets procesados y finales.

El flujo se organiza en notebooks por fases, generando versiones procesadas y un dataset final enriquecido **df_final_eda** para análisis posteriores y elaboración del informe.

Los valores extremos se limitan únicamente en los gráficos para mejorar la legibilidad y evitar que la distribución quede “aplastada”, sin modificar en ningún caso el dataset original utilizado para el análisis.

4. Datos utilizados

4.1 Fuentes

El análisis se basa en la integración de dos fuentes de información complementarias, diseñadas para estudiar el fraude desde una perspectiva transaccional y contextual.

1º fraudTrain.csv: Contiene el histórico de transacciones y recoge el detalle de cada operación, incluyendo información temporal con la fecha y hora, económica con el importe, comercial con la categoría comercial y nombre del comercio y geográfica con la ciudad, el estado y el código postal (zip), además de atributos del cliente como género y ocupación. Esta fuente contiene también la variable objetivo del proyecto **is_fraud**.

2º IRSIncomeByZipCode.xlsx: Contiene la información socioeconómica por código postal y aporta métricas agregadas a nivel de zip relacionadas con el contexto económico del área, como indicadores de renta media y de ingresos imponibles, que permiten contextualizar cada transacción según el entorno donde se produce.

4.2 Integración de fuentes y dataset final

Ambas tablas se unen mediante la clave **zip**, generando un dataset unificado **df_final** que conserva todas las transacciones y añade las variables socioeconómicas disponibles para cada código postal. Tras esta integración se aplican validaciones de calidad coherencia, duplicados y nulos, para asegurar consistencia.

4.3 Dataset para el EDA

Sobre df_final se construye una versión final para análisis **df_final_eda**, que incorpora variables derivadas orientadas a enriquecer el EDA, especialmente en la dimensión temporal y en ratios económicos. Entre ellas destacan:

- Variables temporales: trans_date, month, day_of_week, hour, moment_of_day, is_weekend.
- Variables demográficas: age, age_group.
- Variables económicas derivadas: taxable_share, amt_vs_avg_agi, income_level.

En conjunto, el dataset resultante permite analizar el fraude con enfoque temporal, geográfico, comercial, por perfil y por contexto económico, manteniendo trazabilidad entre las fuentes originales y las transformaciones aplicadas.

5. Preparación de los datos

Este apartado recoge el proceso seguido para dejar las dos fuentes en un estado coherente, limpio e integrable, para construir el dataset final sobre el que se realiza el EDA.

5.1 Revisión preliminar de las fuentes

Como primer paso se revisa la estructura de ambos datasets para entender su contenido y detectar posibles problemas que afecten al análisis posterior. En esta fase se comprueban:

- Dimensiones del dataset, número de filas y columnas, además de la vista inicial de registros.
- Tipos de variables y coherencia entre tipo y contenido.
- Presencia de valores nulos y duplicados.
- Cardinalidad en variables categóricas para anticipar cómo visualizarlas y analizarlas.

5.2 Limpieza y estandarización

A partir de la revisión inicial, se aplican transformaciones orientadas a asegurar consistencia:

- Conversión y normalización de tipos de dato.
- Revisión de variables que pueden actuar como identificadores o aportar poco valor analítico para evitar ruido.
- Comprobaciones de calidad básicas, ausencia de duplicados, formato homogéneo en categorías y consistencia general del dataset.

5.3 Integración por ZIP y validaciones

Una vez depuradas las fuentes, se realiza la unión por zip para incorporar el contexto socioeconómico a cada transacción. Para garantizar la integridad del dataset tras el merge, se aplican validaciones específicas:

- **Coherencia geográfica**: se contrasta el estado de la transacción con el estado asociado al ZIP en la tabla de ingresos, cuantificando discrepancias y corrigiendo cuando procede para homogeneizar el campo.
- **Control de integridad del merge**: se verifica que el número de transacciones se mantiene.
- **Duplicados post-unión**: se comprueba que la unión no genera registros repetidos inesperados.
- **Nulos en variables socioeconómicas**: se revisa el volumen y distribución de valores ausentes, y cuando afecta a variables consideradas esenciales se aplica filtrado para asegurar consistencia en los cálculos posteriores.

El resultado de esta fase es un dataset unificado **df_final** consistente y listo para el análisis exploratorio.

5.4 Creación de variables derivadas para el EDA

Con el dataset integrado, se generan variables adicionales para facilitar interpretaciones y análisis por segmentos:

- **Temporales**: trans_date, month, day_of_week, hour, moment_of_day e is_weekend.
- **Demográficas**: age y age_group.
- **Ratios económicos**: taxable_share y amt_vs_avg_agi, para medir el importe en términos relativos al contexto económico del ZIP.
- **Segmentación socioeconómica**: income.

Además, se realiza un control específico para confirmar que los ratios no generan valores problemáticos tras las divisiones.

5.5 Dataset final para el análisis

Tras completar la limpieza, integración y enriquecimiento del dataset, se guarda una versión final preparada para el EDA denominada **df_final_eda**, asegurando que:

- El dataset mantiene integridad estructural.
- No existen problemas de calidad relevantes.
- Las variables derivadas están correctamente calculadas y listas para análisis descriptivo, comparativo y temporal.

6. Análisis exploratorio (EDA)

En este apartado se presentan los principales resultados del análisis exploratorio sobre el dataset final, combinando estadística descriptiva y visualización para caracterizar el comportamiento de las transacciones y su relación con la variable objetivo **is_fraud**.

6.1 Variable objetivo: distribución y desbalance

Se analiza la distribución de is_fraud para cuantificar el grado de desbalance entre transacciones legítimas (0) y fraudulentas (1). Este punto es clave porque condiciona la interpretación del resto del EDA: los recuentos absolutos pueden inducir a conclusiones erróneas y, por ello, se priorizan comparativas en términos de tasa y proporciones.

6.2 Análisis univariante

Variables numéricas

Se describen medidas de tendencia central y dispersión, especialmente mediana y percentiles, además se visualizan distribuciones mediante histogramas y boxplots. En varias variables se observan distribuciones asimétricas y valores extremos, por lo que se emplean medidas robustas y, cuando es necesario para visualizar mejor, se aplican recortes solo con fines gráficos.

Variables categóricas

Se analizan frecuencias y proporciones para identificar categorías dominantes y posibles desequilibrios. En variables de alta cardinalidad se realizó priorizando el top categorías, evitando conclusiones basadas en categorías con bajo volumen.

6.3 Análisis bivariante respecto a is_fraud

Numéricas vs is_fraud

Se comparan distribuciones por clase mediante boxplots agrupados y se complementa con tablas de estadísticos por clase con medianas y percentiles. Este enfoque permite identificar variables con mayor separación entre fraude y no fraude, destacando aquellas relacionadas con el importe **amt** y con medidas relativas **amt_vs_avg_agi**.

Las diferencias más marcadas aparecen en variables ligadas al importe y a medidas relativas:

- amt mediana: **46,85** no fraude vs **362,28** fraude.
- amt_vs_avg_agi mediana: **0,86** no fraude vs **8,58** fraude.

Este resultado indica que el fraude tiende a concentrarse en transacciones mucho mayores, tanto en términos absolutos como respecto a la renta media del ZIP.

En cambio, los indicadores socioeconómicos medios por ZIP muestran separaciones pequeñas:

- avg_agi mediana: **48,57** no fraude vs **47,71** fraude
- avg_total_income mediana: **49,59** no fraude vs **48,72** fraude
- avg_taxable_income mediana: **40,44** no fraude vs **39,20** fraude

Por sí solos, estos promedios no diferencian claramente las clases.

También se aprecia señal temporal en variables numéricas:

- hour mediana: **14** no fraude vs **21** fraude

Sugiriendo mayor presencia relativa de fraude en horas tardías.

Categóricas vs is_fraud

Se emplean tablas de contingencia normalizadas por categoría para estimar la tasa de fraude dentro de cada grupo, y se visualiza con barras apiladas en variables de baja cardinalidad. En variables de alta cardinalidad se prioriza el análisis con control de volumen, ya que tasas extremas pueden deberse a categorías con pocos casos.

Categorías con diferencias más visibles

- category: algunas categorías presentan tasas de fraude superiores:
    - Shopping_net: **1,64%**
    - Misc_net: **1,53%**
    - Grocery_pos: **1,49%**

  Frente a otras con tasas claramente menores:
    - Health_fitness: **0,11%**.

Categorías con diferencias suaves

- gender: variación pequeña del fraude, **0,56%** mujeres vs **0,61%** hombres.
- income_level: rango estrecho del fraude, **0,53%** high hasta un **0,64%** en low, sin separación marcada.
- age_group: las tasas suben ligeramente en tramos altos donde de 76 a 85 años, el fraude es **0,97%** pero sin cambios extremos, ya que de 18 a 25 años el fraude es de **0,61%**.

Alta cardinalidad
En city, job, merchant o zip pueden aparecer tasas del **100%** en algunas filas, pero suelen corresponder a muy pocos registros **3** o **4**, por lo que no es evidencia sólida sin aplicar filtros por soporte mínimo.

6.4 Correlaciones entre variables numéricas

Se construye una matriz de correlación y un mapa de calor para identificar relaciones lineales relevantes y posibles dependencias entre variables. Este paso permite detectar colinealidad, relaciones muy altas entre variables de renta y entender qué variables podrían estar capturando información similar.

La matriz de correlación confirma relaciones fuertes entre variables construidas a partir de las mismas magnitudes:

- amt y amt_vs_avg_agi: **0,95**.
- avg_agi, avg_total_income, avg_taxable_income: correlaciones entre **0,99** y **1,00**, indicando que capturan información muy similar.

Respecto a la variable objetivo, las correlaciones lineales son moderadas:

- is_fraud con amt: **0,22**
- is_fraud con amt_vs_avg_agi: **0,21**

El resto se mantiene cerca de cero, lo que sugiere que muchas variables aportan señal débil de forma aislada, o que la relación no es lineal.

6.5 Análisis temporal

En el análisis temporal se estudia la evolución del fraude a lo largo del tiempo desde tres perspectivas complementarias, la primera muestra que el volumen de transacciones y el número de fraudes varían por mes, con picos asociados a periodos de alta actividad como por ejemplo el mes de diciembre. Sin embargo, la tasa de fraude mensual permite comparar meses con distinto volumen y evidencia que un aumento del fraude en número absoluto puede explicarse por mayor actividad sin que el riesgo relativo suba en la misma proporción.

A nivel intradía, el patrón del fraude es más marcado:

- Por franja horaria, moment_of_day:
    - late night: **1,92%**
    - night: **1,27%**

  Frente a franjas diurnas como morning con **0,10%** o afternoon con **0,13%**.
- Por hora, se confirman picos claros:
    - 22h: **2,92%**
    - 23h: **2,71%**

En cambio, por día de la semana las variaciones son más suaves de entre el **0,48%** y el **0,71%**, y la comparación fin de semana vs laborable muestra diferencias contenidas, coherentes con el patrón semanal, que muestra un **0,61%** de fraude en laborable y un **0,53%** el fin de semana.

7. Aplicaciones prácticas para optimizar la detección del fraude

Desde un punto de vista operativo, los hallazgos del EDA permiten proponer medidas concretas para mejorar la detección sin necesidad de un sistema complejo, basándose en la priorización del riesgo, la reducción de falsos positivos y el uso eficiente de recursos.

Una primera aplicación inmediata es la priorización temporal del monitoreo. Dado que el riesgo relativo se concentra de forma clara en el tramo **22h** a **3h**, conviene elevar el nivel de control en esas horas. Esto puede traducirse en reglas como la revisión reforzada o autenticación adicional para transacciones realizadas a partir de cierta hora, especialmente cuando coinciden con otros factores de riesgo. En la práctica, esto ayuda a asignar más recursos humanos o computacionales en los momentos del día donde la probabilidad de fraude es mayor, y a relajar controles en franjas donde el riesgo es sistemáticamente bajo como la mañana y tarde, reduciendo fricción al usuario en operaciones de bajo riesgo.

En segundo lugar, el EDA sugiere que la señal más fuerte se encuentra en el importe y en su relación con el contexto económico del área. Por ello, una medida muy útil es el diseño de umbrales dinámicos basados en amt_vs_avg_agi, en lugar de umbrales fijos solo por amt. Un importe alto puede ser razonable en un ZIP con mayor renta media, pero el ratio capta precisamente cuándo una transacción resulta inusualmente grande respecto a ese entorno. Esto permite priorizar alertas que sean más comparables entre zonas, evitando que el sistema se limite a perseguir importes altos sin contexto. A nivel práctico, puede implementarse como reglas del tipo, si amt_vs_avg_agi supera cierto percentil y la transacción ocurre en horas nocturnas, elevar nivel de alerta. Este enfoque es especialmente útil porque combina dos señales consistentes del EDA, magnitud económica relativa + horario.

Tercero, los resultados por categoría de comercio permiten ajustar reglas de riesgo por tipo de operación. Categorías con tasas superiores pueden someterse a controles adicionales, mientras que categorías con tasas muy bajas pueden mantener un flujo más normal, siempre que no haya otras señales elevadas. Esto facilita construir una lógica de capas, la categoría no determina por sí sola el fraude, pero sí puede modular el umbral de alerta o la intensidad del control cuando se combina con importe y horario. En entornos reales, esta idea se usa para reducir falsos positivos, no todo importe alto tiene el mismo riesgo si ocurre en un tipo de comercio históricamente menos problemático.

Cuarto, en variables geográficas el análisis muestra que se deben evitar conclusiones basadas en categorías con pocos registros, pero sí se puede construir un enfoque práctico de listas de vigilancia con control de volumen. La aplicación profesional aquí no es etiquetar un ZIP como fraudulento por una tasa extrema, sino generar un ranking de zonas con, suficiente volumen de transacciones y una tasa por encima de la media. Esto produce un conjunto manejable de áreas para revisión o para calibración de umbrales, y reduce el riesgo de actuar sobre ruido estadístico. Con el mismo criterio, en variables de alta cardinalidad, una estrategia útil es centrarse en comercios con volumen suficiente y observar si mantienen tasas sistemáticamente superiores, si se detectan patrones persistentes, se pueden aplicar controles reforzados en esas entidades o revisar reglas específicas asociadas a ellas.

Quinto, el componente demográfico edad y género aporta una señal más suave, por lo que su aplicación debe ser prudente. El uso correcto no es tomar decisiones directas por género ni penalizar segmentos sin una separación clara. Sin embargo, la edad y los grupos de edad sí pueden utilizarse como variables auxiliares de calibración si en fases posteriores se confirma que añaden valor incremental.

Sexto, el análisis temporal mensual muestra por qué las decisiones operativas deberían basarse en tasas y no solo en recuentos. El caso de meses con picos de actividad como diciembre ilustra que puede haber muchos fraudes en términos absolutos simplemente porque hay muchas más transacciones. Una aplicación práctica es definir métricas de seguimiento basadas en la tasa mensual y, si se usan recuentos acompañarlos siempre de volumen total. Esto permite distinguir las situaciones de más fraude porque hay más transacciones de las situaciones de más fraude porque el riesgo ha subido, lo que es clave para activar alertas operativas de forma correcta.

Séptimo, los resultados de correlación sugieren una guía clara para el diseño de variables y selección en un sistema de detección, esto implica que conviene evitar introducir muchas variables que aportan la misma información, y priorizar un conjunto pequeño pero informativo.

Por último, una optimización clave en detección de fraude es el diseño de un flujo de decisión escalonado, basado en lo encontrado. Primero, identificar transacciones de bajo riesgo. Segundo, identificar transacciones con riesgo y aplicar acciones graduales, desde autenticación reforzada hasta revisión manual o bloqueo preventivo según el nivel de riesgo. Esta estrategia concentra recursos donde el riesgo es mayor y evita saturar con falsos positivos en el volumen principal de transacciones legítimas.

8. Resultados y conclusiones

El EDA se ha realizado sobre un dataset integrado de **325.090** transacciones que combina el histórico transaccional con variables socioeconómicas agregadas por ZIP. Tras la depuración e integración, el conjunto queda listo para el análisis, ya que no se detectaron nulos ni duplicados en el dataset final y, en particular, los ratios creados no presentan valores indefinidos, lo que nos permite utilizarlos en el análisis sin riesgo de errores en cálculos o visualizaciones.

La variable objetivo **is_fraud** confirma un escenario con un desbalance muy acusado típico del fraude, los casos positivos son minoritarios. En los resultados se contabilizan **1.901** fraudes sobre **325.090** transacciones, lo que equivale a una tasa global aproximada del **0,58%**. Este desbalance condiciona la lectura del análisis, ya que los recuentos absolutos pueden aumentar simplemente por un mayor volumen de actividad. Por este motivo, en las comparativas se ha priorizado el uso de tasas y proporciones, complementando los conteos cuando es necesario.

En relación con **cuándo ocurre más fraude**, el análisis temporal muestra que el componente más informativo es el patrón intradía. La tabla de tasa por hora identifica picos muy marcados en las últimas horas del día, a las **22h** con un **2,92%** y a las **23h** con un **2,71%**, además de un tramo elevado durante la madrugada, entre las **0** y las **3h** observándose una tasa del **1,48%** y **1,61%** respectivamente. Esto es coherente con el resumen bivariante, donde la mediana de hour pasa de **14** en no fraude a **21** en fraude, sugiriendo concentración del fraude en horarios más tardíos.

En cambio, el patrón por día de la semana es más suave, las tasas oscilan aproximadamente entre el **0,48%** el Lunes y **0,71%** el Viernes, con valores cercanos el Miércoles al **0,70%** y Jueves al **0,67%**. La comparación fin de semana vs laborable no introduce un cambio drástico adicional y resulta consistente con lo observado por día mostrando que el sábado y domingo no concentran sistemáticamente el máximo.

En la dimensión mensual, los gráficos de volumen y de número de fraudes presentan picos de actividad, destacando el de diciembre de 2019, pero observando la tasa mensual no se replican necesariamente esos picos, esto refuerza que los aumentos en fraudes absolutos pueden explicarse por un aumento masivo de transacciones y no siempre por un incremento proporcional del riesgo, esta diferencia entre recuento y tasa justifica que la comparación mensual se centre en porcentajes para interpretar riesgo.

Respecto a **dónde ocurre más fraude**, las variables geográficas state, city, zip requieren una lectura controlada por volumen. Aparecen tasas relativamente más altas en algunos estados como, District of Columbia **0,91%** o Arizona **0,86%**, pero la interpretación debe hacerse comparando contra la tasa global **0,58%** y evitando conclusiones a partir de categorías con pocos registros. Esto se ve claramente en ciudades con valores extremos por tamaño muestral pequeño como es el caso de Albany con un **100%** pero solo con 4 transacciones. Por tanto, el análisis geográfico es útil para localizar focos y generar hipótesis, pero debe aplicarse con filtros mínimos de transacciones para que las comparaciones sean robustas.

En cuanto a en qué **tipo de comercio ocurre más fraude**, la variable category ofrece una señal clara al ser de baja cardinalidad. Destacan tasas superiores en categorías como

Shopping_net **1,64%**, Misc_net **1,53%** y Grocery_pos **1,49%**, frente a categorías con tasas mucho menores como Health_fitness **0,11%**, Home **0,20%** o Food_dining **0,19%**. A nivel de merchant, también aparecen comercios con tasas más elevadas, pero dado que hay **693** distintos nombres de comercios y mucho volumen está distribuido, el análisis debe hacerse imponiendo umbrales de soporte para no sobrerrepresentar casos con pocos registros.

Sobre en qué **perfiles ocurre más fraude**, las diferencias son moderadas en variables demográficas. En género, la variación es pequeña Female muestra un **0,56%** mientras Male muestra **0,61%**, lo que nos indica que no existe una diferencia marcada. En la edad, se aprecia un desplazamiento ligero, la mediana pasa de **43** en no fraude a **48** en fraude, y por age_group se observan tasas relativamente más altas en tramos mayores, como de **76** a **85** con **0,97%**, de **56** a **65** con un **0,84%** y **86+** con un **0,79%**, aunque sin cambios extremos. Las variables como job y city_pop aportan menos señal directa en esta fase, ya que, la mediana de city_pop es prácticamente igual entre clases **4056** en el caso de no fraude, mientras que el fraude presenta un valor de **4046**, y en job aparecen porcentajes muy altos asociados a categorías con muy pocos registros, por lo que en este caso la interpretación también depende de aplicar filtros de volumen para extraer conclusiones sólidas.

La pregunta de si **influye el contexto económico** se responde de forma matizada. Los indicadores socioeconómicos medios por ZIP avg_agi, avg_total_income, avg_taxable_income muestran valores centrales muy parecidos entre clases, en el caso de avg_agi la mediana es **48,57** en no fraude frente a **47,71** en fraude y patrones similares se observan en avg_total_income con **49,59** en no fraude y **48,72** en fraude, y en avg_taxable_income también se presenta un escenario similar con **40,44** en no fraude y un **39,20** en fraude. También taxable_share apenas difiere **0,63** en no fraude frente a un **0,62** en fraude. Sin embargo, la señal más fuerte no está en el nivel socioeconómico en bruto, sino en el importe y su magnitud relativa respecto a la renta del ZIP.
La mediana de amt pasa de **46,85** en no fraude a **362,28** en fraude, y el ratio amt_vs_avg_agi sube de **0,86** a **8,58**, lo que indica que el fraude tiende a concentrarse en transacciones muy elevadas tanto en términos absolutos como respecto al contexto económico del área.

La matriz de correlación refuerza estas conclusiones, amt y amt_vs_avg_agi presentan una correlación muy alta de **0,95**, lo que es esperable porque el ratio incorpora el importe. Frente a is_fraud, las correlaciones más relevantes son precisamente amt con **0,22** y amt_vs_avg_agi con **0,21**, mientras que el resto se mantiene cerca de cero. Además, los indicadores de ingresos avg_agi, avg_total_income, avg_taxable_income aparecen prácticamente redundantes entre sí con correlaciones cercanas a **1**, por lo que aportan información similar del nivel medio del ZIP.

El escenario con mayor concentración de fraude, según los resultados obtenidos, se da en transacciones realizadas en horarios nocturnos y de madrugada, especialmente en las últimas horas del día, la tasa alcanza sus máximos a las **22h** con el **2,92%** y **23h** con **2,71%**, y se mantiene elevada entre **0h** y las **3h** con valores entre **1,48%** y **1,61%.** Este patrón coincide con la segmentación por franja horaria, donde late night y night presentan valores claramente superiores al resto con **1,92%** y **1,27%** respectivamente, y

donde la mediana de hour presenta un valor de **14** en no fraude frente **21** en fraude. Además, el fraude tiende a asociarse a operaciones de importe alto, con una diferencia muy marcada en amt con una mediana de **46,85** en no fraude vs **362,28** en fraude y, sobre todo, en el ratio amt_vs_avg_agi con una mediana de **0,86** en no fraude respecto del **8,58** en fraude, lo que indica transacciones desproporcionadas respecto al nivel de renta medio del ZIP.

En cambio, el escenario de menor fraude aparece en franjas diurnas, donde las tasas son muy bajas con morning **0,10%**, afternoon **0,13%**, evening **0,11%**, y también en variables donde la diferencia entre clases es mínima, como es el caso de género donde la variación es reducida **0,56%** en no fraude frente a **0,61%** de fraude y por día de la semana las tasas se mueven en un rango estrecho entre **0,48%** de no fraude y **0,71%** de fraude, lo que sugiere que el componente semanal aporta menos separación que el horario. A nivel socioeconómico medio del ZIP, los valores centrales de avg_agi, avg_total_income y avg_taxable_income son muy similares entre clases por lo que el riesgo no parece depender tanto del nivel de ingresos del área como del importe relativo de la operación frente a ese contexto.

En conclusión, el análisis exploratorio confirma que el fraude no se explica por un único factor, sino por la combinación del importe en valor absoluto y relativo y el contexto temporal, con una señal especialmente clara en el horario nocturno. Las categorías de comercio también aportan información relevante, mientras que los perfiles demográficos y algunos indicadores socioeconómicos agregados muestran diferencias más contenidas cuando se analizan de forma aislada. El resultado final es un dataset consistente, enriquecido con variables derivadas útiles para describir el fenómeno y preparado para una fase posterior, ya sea de interpretación más profunda o de modelización, donde el fuerte desbalance deberá tratarse con métricas y estrategias adecuadas.