



MAESTRÍA EN CIENCIA DE DATOS E INTELIGENCIA DE NEGOCIOS

MACHINE LEARNING NO SUPERVISADO

Proyecto final de curso, presentado por:

Alfredo Argueta Interiano - AI252944

Giovanni Alexander Escobar - EM252920

Ivo Luis Orellana Girón - OG252913

Marlon Alexander Palacios Díaz - PD252876

José Rodrigo López Torres - LT170438

Contenido

Introducción y objetivos del proyecto	3
Justificación y descripción detallada del dataset utilizado (origen, estructura, variables, etc.)	4
Metodología empleada	5
Análisis exploratorio de datos.....	7
Resultados de la reducción de dimensionalidad	14
Resultados del clustering.....	22
Resultados aplicando K-means para segmentar comercios	22
Resultados aplicando DBSCAN + Harvesine para clusterizar ubicaciones de comercios ..	26
Interpretación de resultados en el contexto del dominio de aplicación	30
Insights accionables y recomendaciones estratégicas	32
Conclusiones y trabajo futuro.....	33

Introducción y objetivos del proyecto

El presente proyecto aplica técnicas de aprendizaje no supervisado para analizar un conjunto de datos transaccionales de 3 meses provenientes de una plataforma fintech que opera en varios países de Centroamérica. Estas organizaciones procesan pagos mediante distintos canales (POS, enlaces de pago, códigos QR, integraciones por API y otros medios digitales) generando un gran volumen de transacciones que dificultan la detección de patrones sin herramientas analíticas avanzadas.

El objetivo central es utilizar métodos de reducción de dimensionalidad y clustering para descubrir estructuras internas, agrupar comercios según comportamientos comunes y generar conocimiento accionable que apoye decisiones estratégicas como segmentación comercial, monitoreo operativo y detección de comportamientos atípicos.

Los objetivos del proyecto son:

- Explorar la estructura de los datos transaccionales y detectar patrones naturales sin etiquetas.
- Reducir la dimensionalidad del dataset para visualizar mejor las relaciones entre comercios.
- Evaluar diversos algoritmos de clustering sobre representación reducida del espacio.
- Comparar la estabilidad y calidad de los grupos obtenidos usando métricas cuantitativas.
- Realizar una segmentación clásica basada en volumen transaccional para contrastar resultados.
- Complementar el análisis con clustering geoespacial mediante DBSCAN + distancia Haversine.
- Identificar segmentos útiles para decisiones comerciales y operativas.

Este proyecto demuestra cómo los métodos no supervisados pueden aportar una visión profunda del comportamiento comercial, incluso sin etiquetas ni clasificaciones previas.

Justificación y descripción detallada del dataset utilizado (origen, estructura, variables, etc.)

El siguiente set de datos muestra las transacciones de Agosto-Octubre 2025 de una Fintech Salvadoreña. Son datos de transacciones de 3 países, por lo que los montos vienen en diferentes monedas, así también la hora de la transacción ya está ajustada a la hora de dicho país.

La base suele usarse para segmentar tamaños de comercio en base a lo que transaccionan al mes, ver tendencias de volumen procesado mensualmente, categorizar montos transaccionados por Inbound/Outbound, ver cantidad de transacciones exitosas y rechazadas mensualmente y así también ver mayor razón de rechazo.

Dentro del set de datos se busca encontrar patrones relacionados a las transacciones con el fin de entender métricas clave y generar métricas significativas al comercio

El diccionario de datos es el siguiente:

- **id_comercio** es el id único para cada comercio que transacciona
- **company_created_at** estampa de tiempo de fecha de creación de comercio
- **categoriagiros del comercio**, es la categoría en la que opera el comercio
- **is_taxpayer** se refiere a que si el comercio que realiza la transacción es pagador de impuesto
- **banco_afiliacion_comercio** banco al que esta afiliado el comercio
- **id_transaccion** es el id único por cada transacción
- **monto** monto de la transacción
- **currency_iso** es el símbolo para distinguir el tipo de moneda
- **tipo_transaccion** es el tipo de transacción realizada
- **modo_lectura** se refiere a cómo el POS lee la tarjeta sea CHIP o NFC
- **pos_id** id único de cada POS
- **transaction_created_at** es la fecha y hora en la que se realizó la transacción
- **lat** latitud guardada de la transacción, solo en POS
- **log** longitud guardada de la transacción, solo en POS
- **card_brand** Marca de la tarjeta usada para el pago
- **país** país donde se realizó la transacción
- **vendedores** es el id único de los vendedores vinculados a ese comercio, esto quiere decir que si un comercio tiene un vendedor vinculado, es porque ese comercio se vinculó por un departamento Comercial y no por marketing puro, a estas dos categorías se le conoce como Outbound e inbound respectivamente
- **modelos_pos** son los diferentes modelos de POS

Metodología empleada

El proceso analítico completo se estructuró en cinco etapas principales:

1. Análisis Exploratorio de Datos (EDA)

Se examinó la estructura general del dataset, distribuciones, outliers, correlaciones y patrones iniciales.

Esta fase permitió comprender el comportamiento natural de los comercios antes de aplicar transformaciones o modelados.

2. Reducción de Dimensionalidad

Para comprender mejor la geometría del dataset se aplicaron múltiples técnicas:

- **PCA (Análisis de Componentes Principales)**
Captura relaciones lineales y explica qué porcentaje de la varianza es aportado por cada componente.
- **t-SNE**
Modelo no lineal que preserva la estructura local, ideal para visualizar clústeres orgánicos.
- **UMAP**
Similar a t-SNE pero más estable, más rápido y mejor para representar estructuras globales.

La combinación de métodos lineales y no lineales permitió obtener una visión robusta de la estructura interna de los datos.

3. Clustering sobre Espacios Reducidos

Sobre las proyecciones generadas (PCA, t-SNE, UMAP) se evaluaron múltiples algoritmos:

- K-Means
- Clustering Jerárquico
- HDBSCAN

Para comparar calidad y estabilidad se utilizaron métricas como:

- Silhouette Score
- Calinski-Harabasz Index
- Davies-Bouldin Score

Esto permitió identificar qué combinación de técnica de reducción + algoritmo de clustering ofrecía los grupos más claros y consistentes.

4. Segmentación basada en volumen transaccional

Como referencia, se aplicó un modelo clásico basado únicamente en el monto total transaccionado por cada comercio.

Se evaluaron dos variantes:

1. Monto en escala original
2. Monto transformado mediante logaritmo natural (*para reducir asimetría*)

El pipeline incluyó:

- Agregación por comercio
- Análisis de distribución (KDE, media, mediana)
- Estandarización
- Método del codo
- K-Means óptimo
- Boxplots por segmento

Este modelo genera segmentos intuitivos basados en el nivel de actividad económica.

5. Clusterización de Ubicaciones

Finalmente, para complementar el análisis transaccional se aplicó:

- Conversión de coordenadas a radianes
- Distancia Haversine
- DBSCAN
- Agrupación mensual por comercio

Esto permite identificar:

- zonas de operación recurrentes,
- comercios con múltiples sedes,
- comportamientos sospechosos o atípicos,
- áreas geográficas con mayor densidad de uso.

Análisis exploratorio de datos

El dataset original contiene con 403535 registros en total y 20 columnas/variables/features

vendedores	241455
long	56542
lat	56542
modo_lectura	54804
pos_id	54804
modelos_pos	39724
pos_ids	39724
card_brand	1968
banco_afiliacion_comercio	393

Las columnas banco_afiliacion_comercio, modo_lectura, pos_id , modelo_pos, lat log card_brand y vendedores contienen valores nulos (en las variables card_brand y las variables de POS (modo_lectura, pos_id, modelo_pos) es normal ya que no todas las transacciones se pagan con tarjetas de crédito, al igual id_vendedor puede ser nulo en las transacciones tipo Inbound pero los otros campos se debe investigar y analizar si es normal esos valores nulos o son valores faltantes).

NOTA No se observan registros duplicados

AGREGAR COLUMNAS ADICIONALES DE INTERES

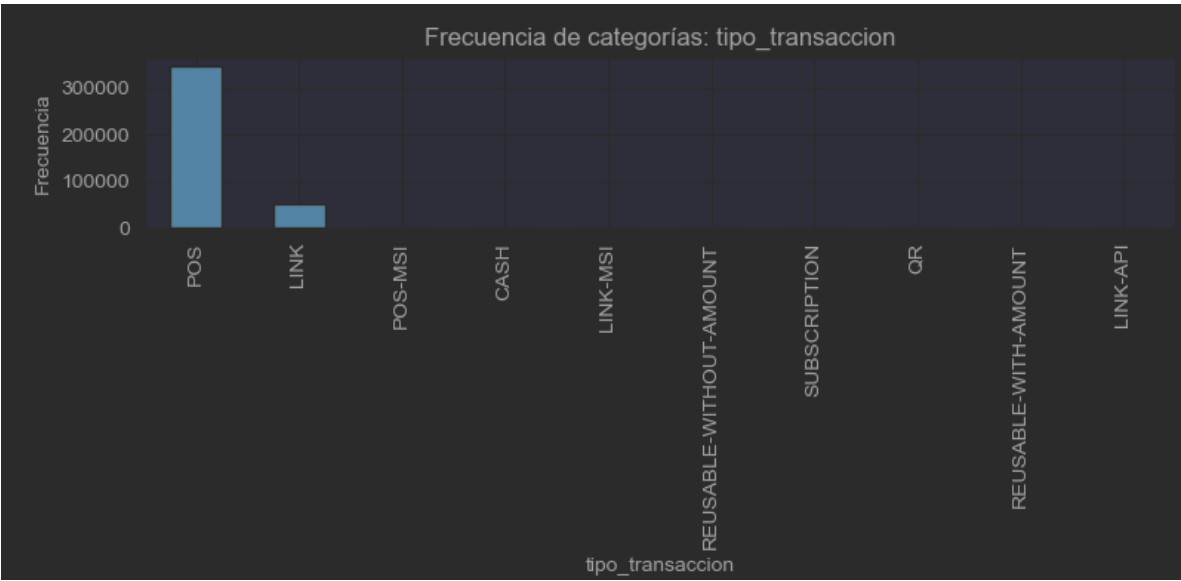
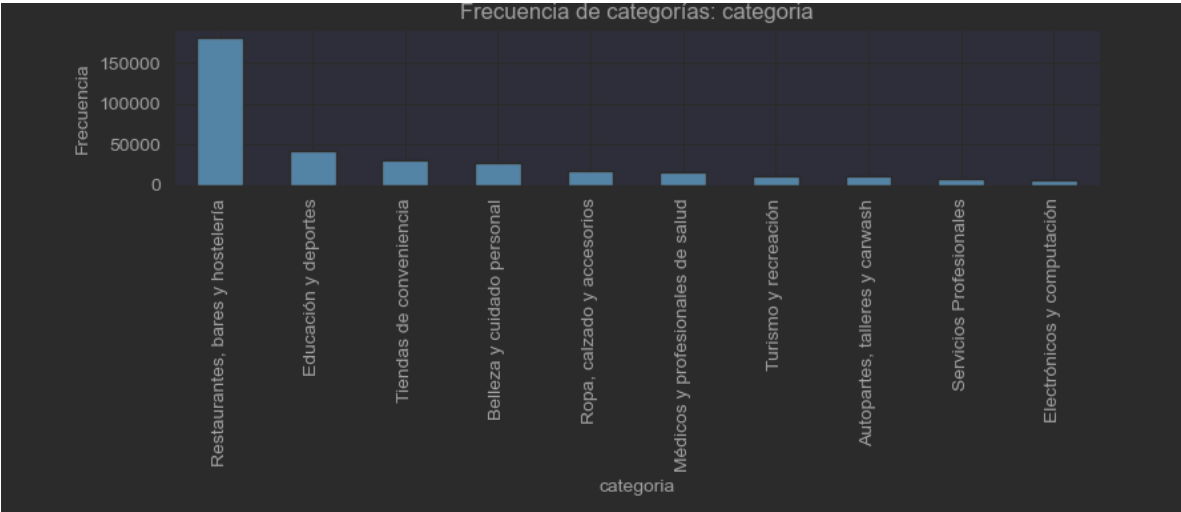
- Se calcula columna adicional para indicar el tipo de area outbound o inbound
- Como el monto esta en diferentes monedas es necesario homologar los valores a USD (Dólares Estadounidense) para realizar cualquier analisis comparativo eficiente. Por lo tanto se agregaran 2 columnas tasa_conversion y monto_usd
- Descomponer fecha en mes y dia de la semana y hora

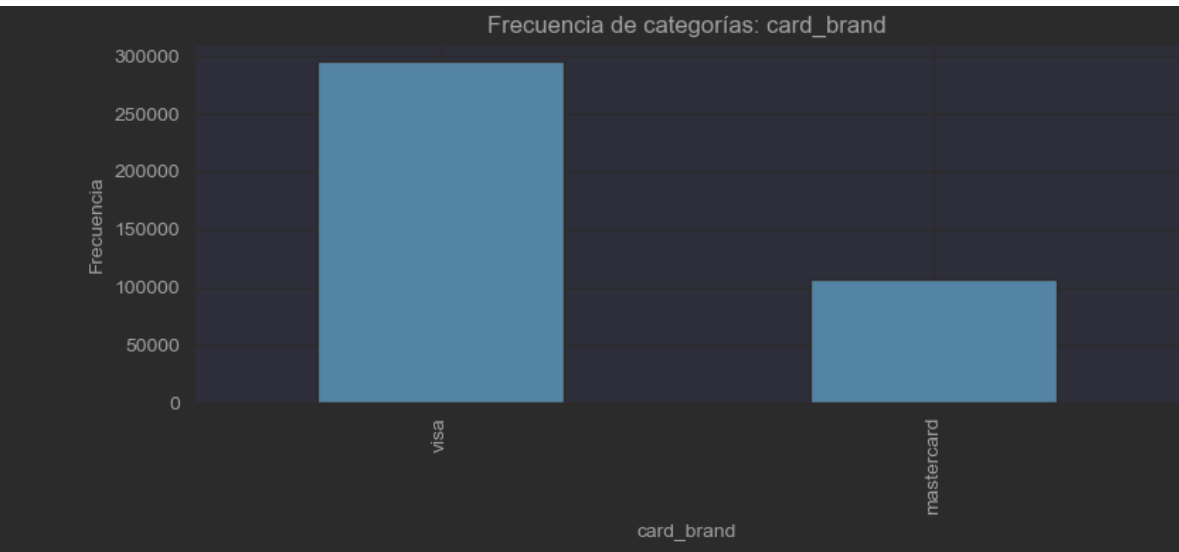
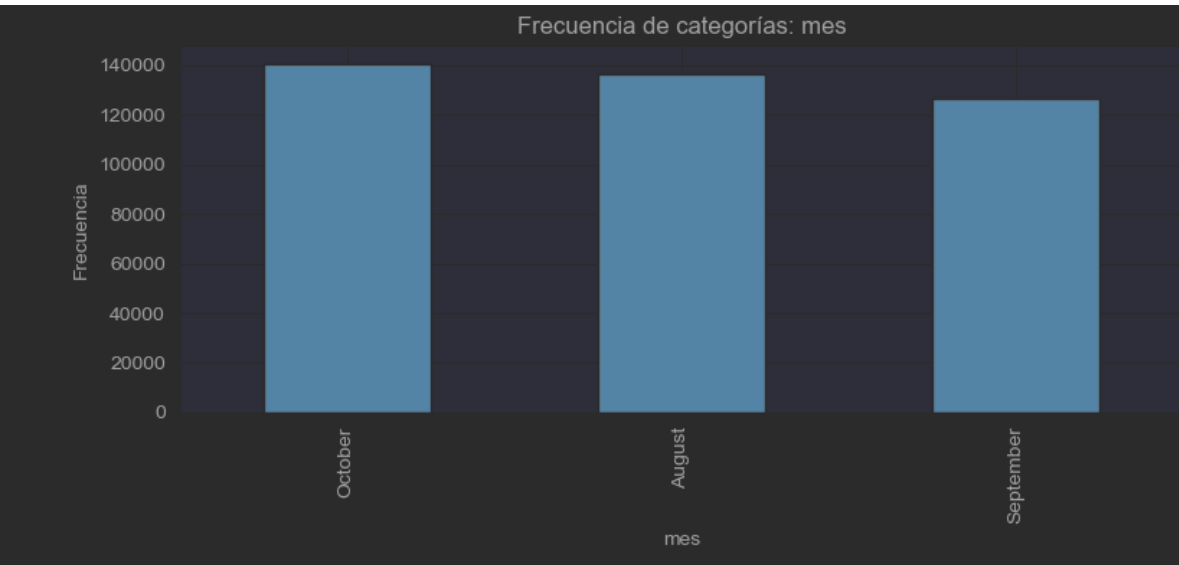
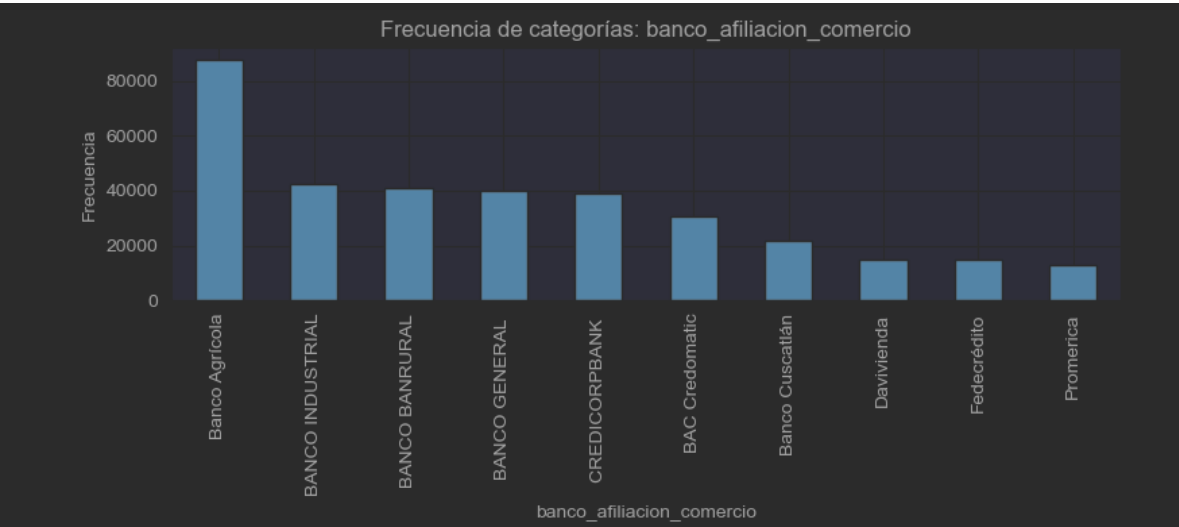
10 rows ▾ 10 rows × 3 cols

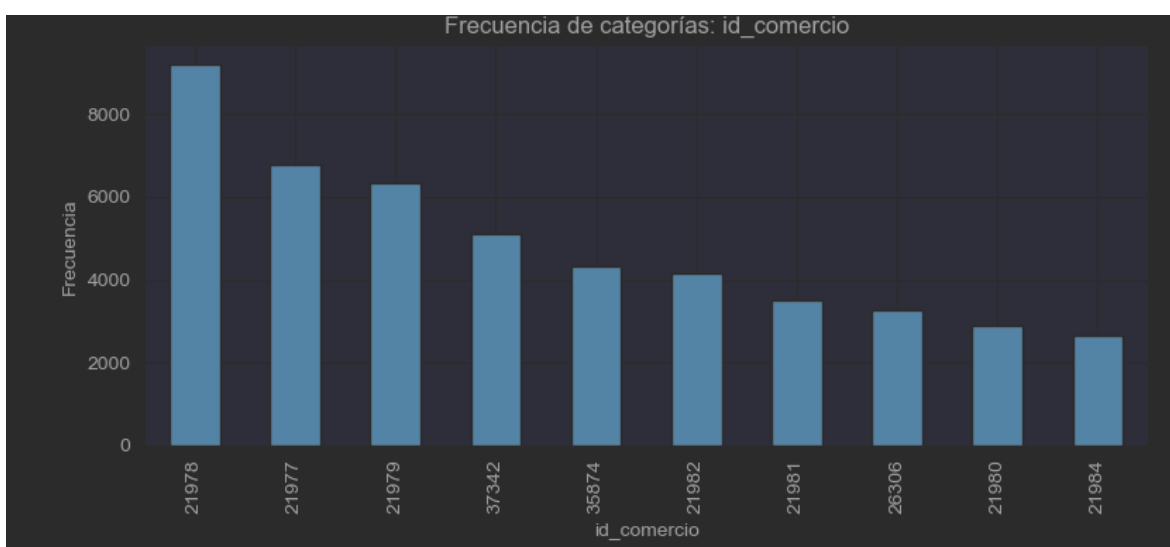
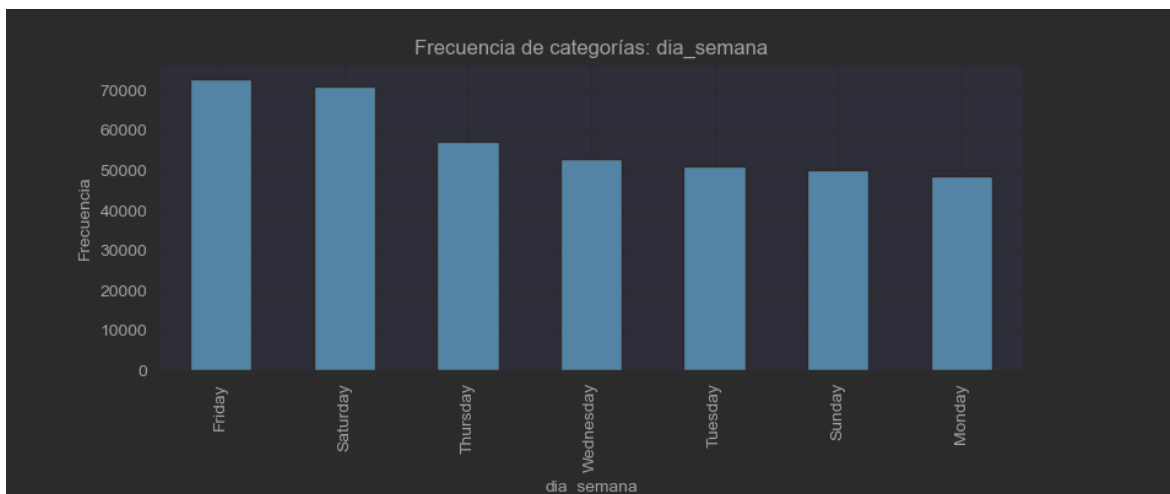
↕	pais	↕	currency_iso	↕	conteo	↕
2	El Salvador		Dolar estadounidense		193967	
13	Guatemala		Quetzal		115818	
16	Panamá		Dolar estadounidense		62959	
20	Panamá		Quetzal		9200	
17	Panamá		Lempira		6345	
18	Panamá		Mexican Peso		4717	
19	Panamá		Peso Dominicano		4139	
14	Panamá		Colon costarricense		3499	
15	Panamá		Cordoba		2890	
6	El Salvador		Quetzal		1	

Se observan diferencias de moneda incluso en un mismo país por lo que la mejor opción es convertir a USD (Dolares Estadounidense) todos los montos antes de analizar la variable

ANALISIS DE VARIABLES CUALITATIVAS







Interpretaciones:

El 86% de las transacciones se realizan por medio de POS, el 12% son realizadas por LINK y menos de un 1% son con CASH.

El 64% de las transacciones son en Dólares USD, 31% usan Quezales GTQ y el resto en otras divisas.

El 73% de las transacciones fueron pagadas con tarjeta VISA y el 26% con MASTERCARD.

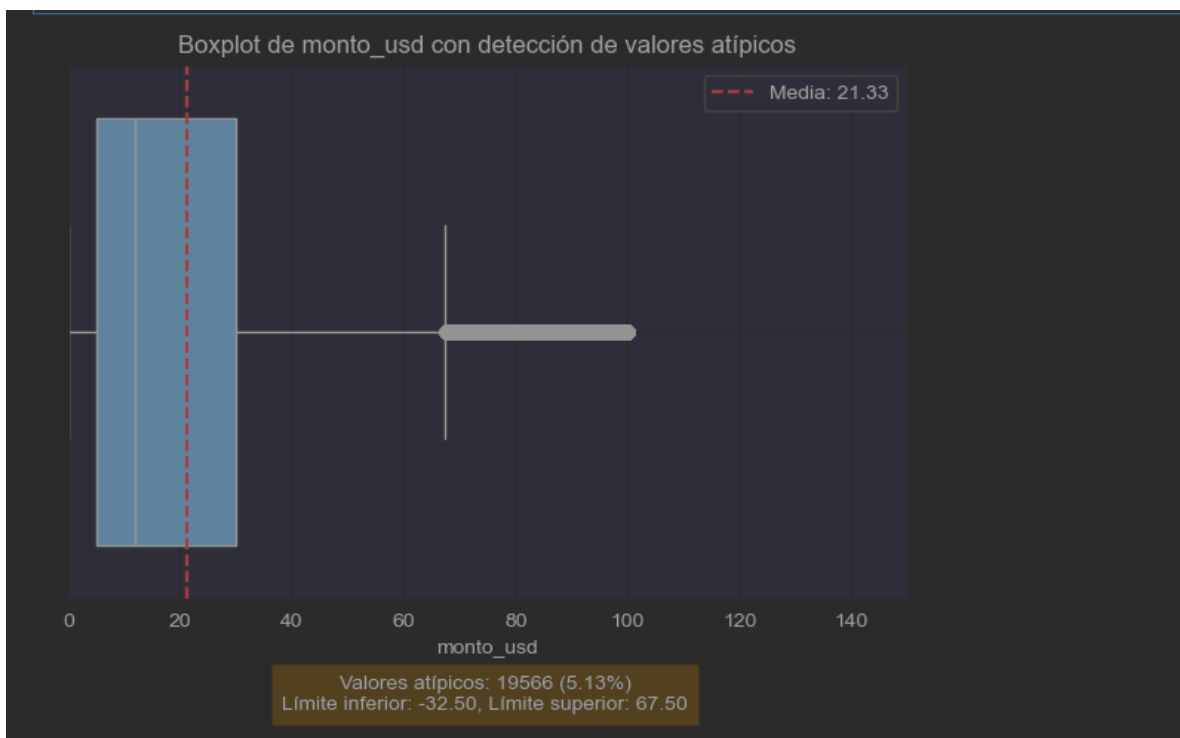
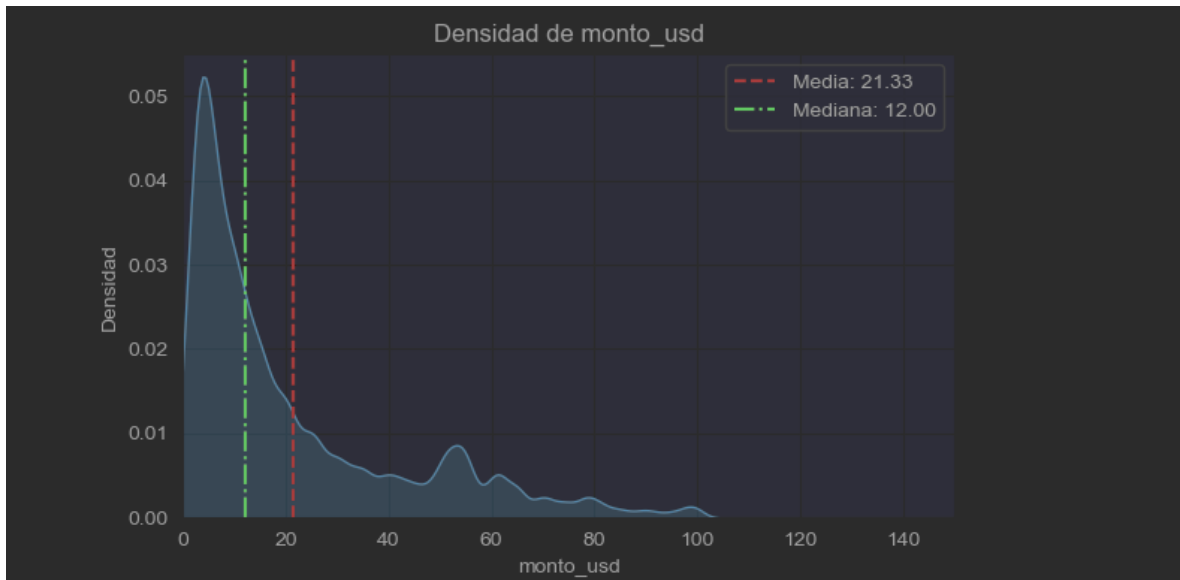
El Salvador aporta el 48% de las transacciones mientras que Guatemala un 28% y el resto es aportado por Panama con un 23%.

Las transacciones realizadas por mes tienen una tendencia similar siendo Octubre el mes con mas transacciones 140000 seguido de Agosto con 136000, en general el flujo de transacciones por mes es estable.

El rubro de empresas que mas genera transacciones es el de Restaurantes. Bares y hostelería aportando el 45% del total de transacciones.

La Empresa con mas transacciones realizadas es la compañía No.21978 con 21977 transacciones por 10 que se observa que existe diversidad de participación de empresas.

ANÁLISIS DE VARIABLES CUANTITATIVAS



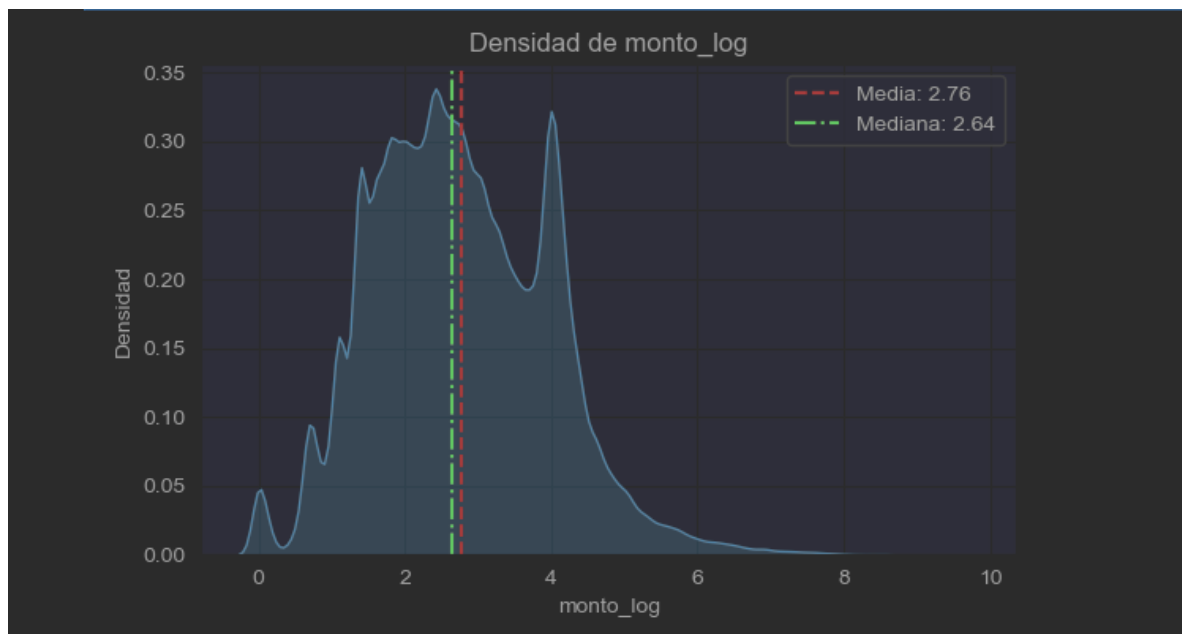
- La distribución sigue siendo fuertemente asimétrica. La Mediana es el centro de la actividad transaccional (12 USD) (el 50% de los datos). Su posición tan cerca del borde izquierdo de la caja confirmanuevamente la fuerte asimetría positiva.

- Es el Primer Cuartil (<10 USD). El 25% de las transacciones caen por debajo de este monto.
- Es el Tercer Cuartil (un poco menos de <40 USD). El 75% de las transacciones caen por debajo de este monto.
- El bigote inferior no es visible, ya que el mínimo de los datos (0.00 USD) es muy cercano al Q1.
- La longitud del bigote superior es mucho más larga que la distancia entre Q1 y Q3, 10 que visualmente confirma la asimetría a la derecha.
- Existe un 5.13% de atípicos en este segmento analizado cantidades superiores a 70 USD

SELECCION Y PREPROCESAMIENTO DE VARIABLES PARA ANALISIS NO SUPERVISAD

Se obtienen 17 variables relevantes para el analisis y/o aplicacion de modelos no supervisados

- Las variables continuas como lat y lomg no requieren tratamiento adicional pero se deben eliminar registros con valores nulos
- La variable monto_usd debido a que presenta un sesgo pronunciado a la derecha se recomienda aplicar una transformación logaritmica para evitar sesgos en el analisis debido a los valores atipicos.
- Tomamos la fecha de transaccion transaction_created_at en diferentes unidades de medida en los que nos interesa hacer analisis temporales
- Se calcula la antigüedad del comercio en días a la fecha de la transaccion company_created_at
- Se codifican el resto de las variables categoricas nominales con One-Hot Encoding



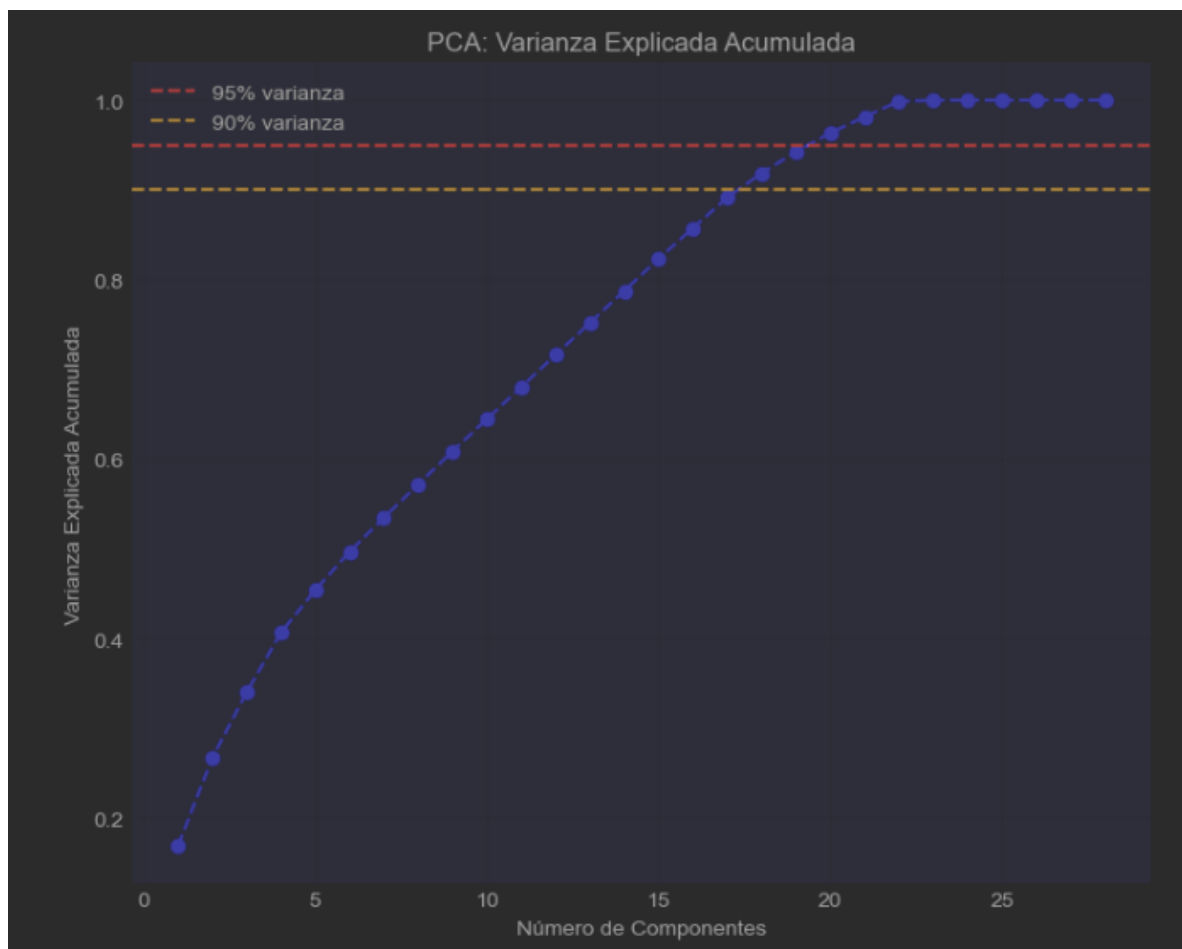
Resultados de la reducción de dimensionalidad

La reducción de dimensionalidad se aplicó con el propósito de identificar estructuras latentes en el comportamiento transaccional de los comercios y facilitar la aplicación de algoritmos de clustering. Se utilizaron tres técnicas principales: **PCA**, **t-SNE** y **UMAP**, cada una aportando una perspectiva distinta sobre los datos

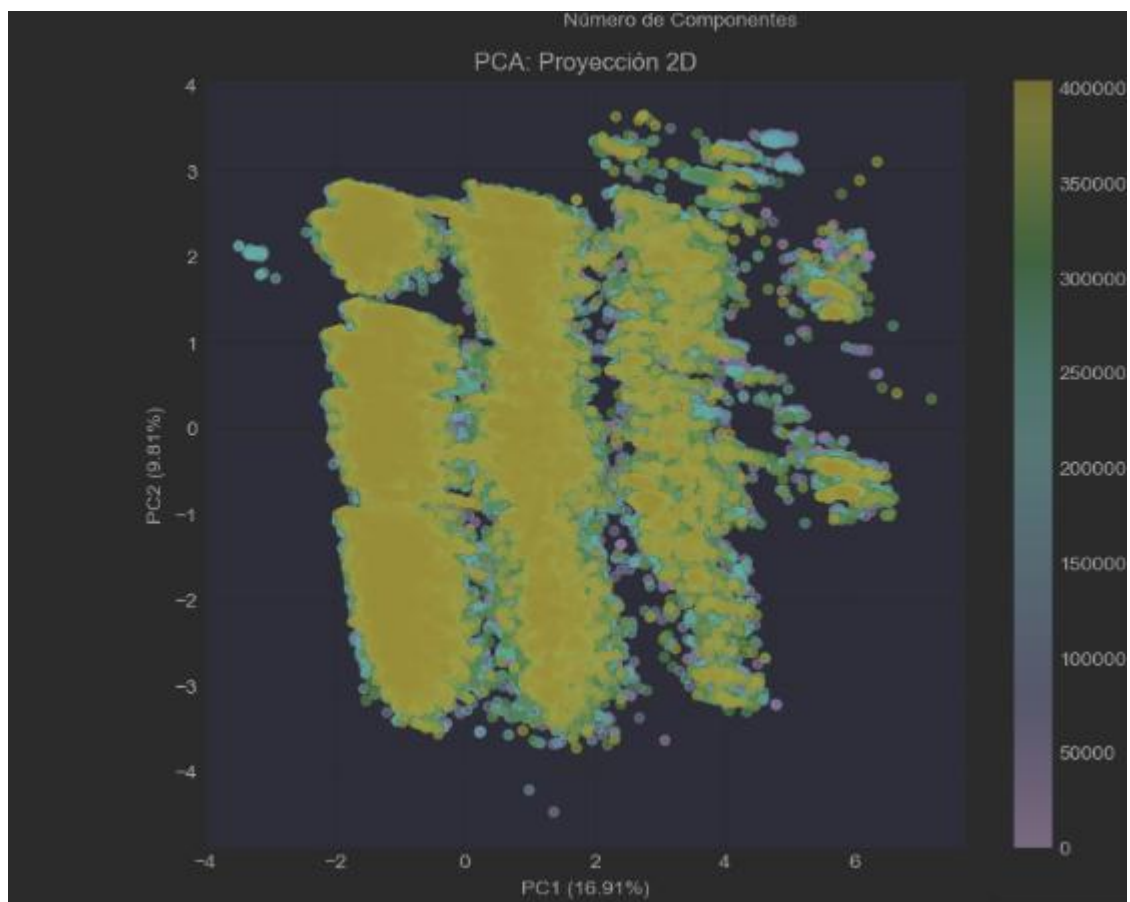
1. PCA – Perspectiva Lineal y Explicación de Varianza

El Análisis de Componentes Principales (PCA) reveló que:

Las primeras **15-18 componentes** capturan una proporción importante de la varianza del sistema.



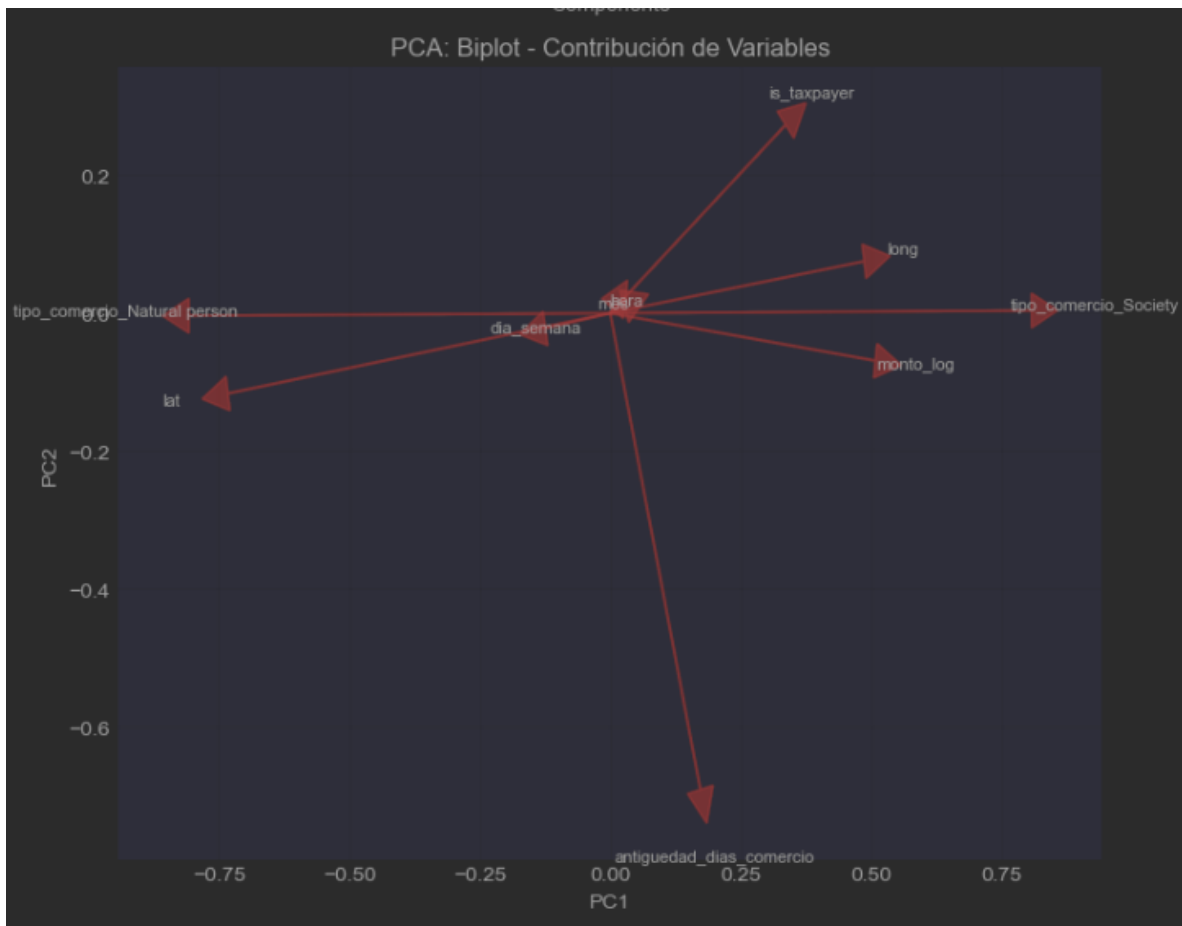
La varianza acumulada sugiere que el dataset contiene **estructura dispersa**, sin un eje dominante que condense la mayoría de la información.



El **biplot** muestra que variables como:

- antigüedad del comercio
- monto logarítmico
- tipo de contribuyente
- tipo de comercio

tienen mayor influencia en los primeros componentes.

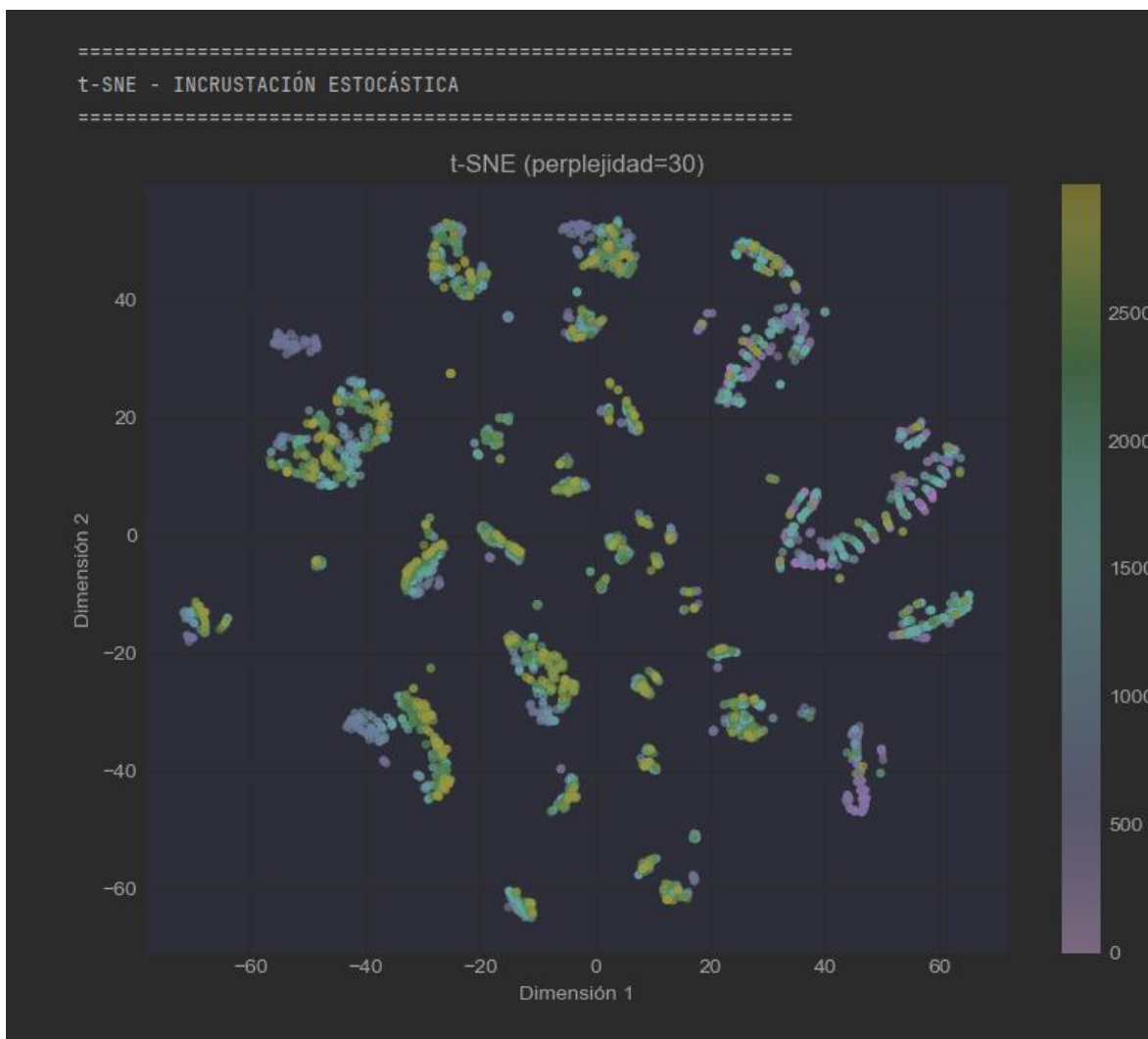


Sin embargo, la proyección 2D resultante evidencia que PCA **no separa claramente grupos naturales**, lo cual es coherente dado que los patrones son mayormente no lineales.

2. t-SNE – Estructuras No Lineales Claras

Al aplicar t-SNE con una perplejidad de 30, los datos mostraron:

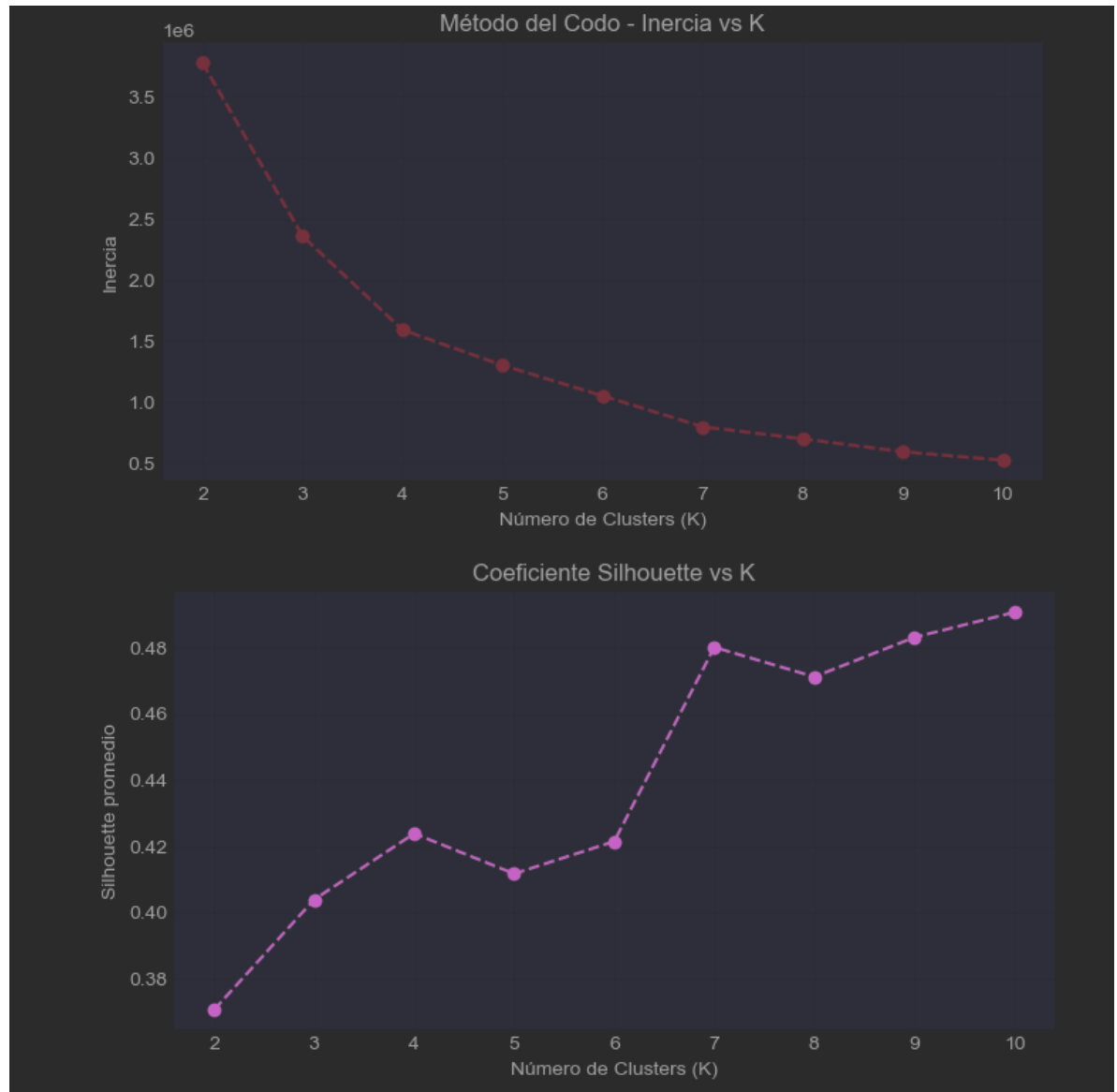
- Formación de **múltiples clústeres bien diferenciados**, compactos y con bordes definidos.
- Los grupos parecen reflejar patrones de comportamiento transaccional complejos.



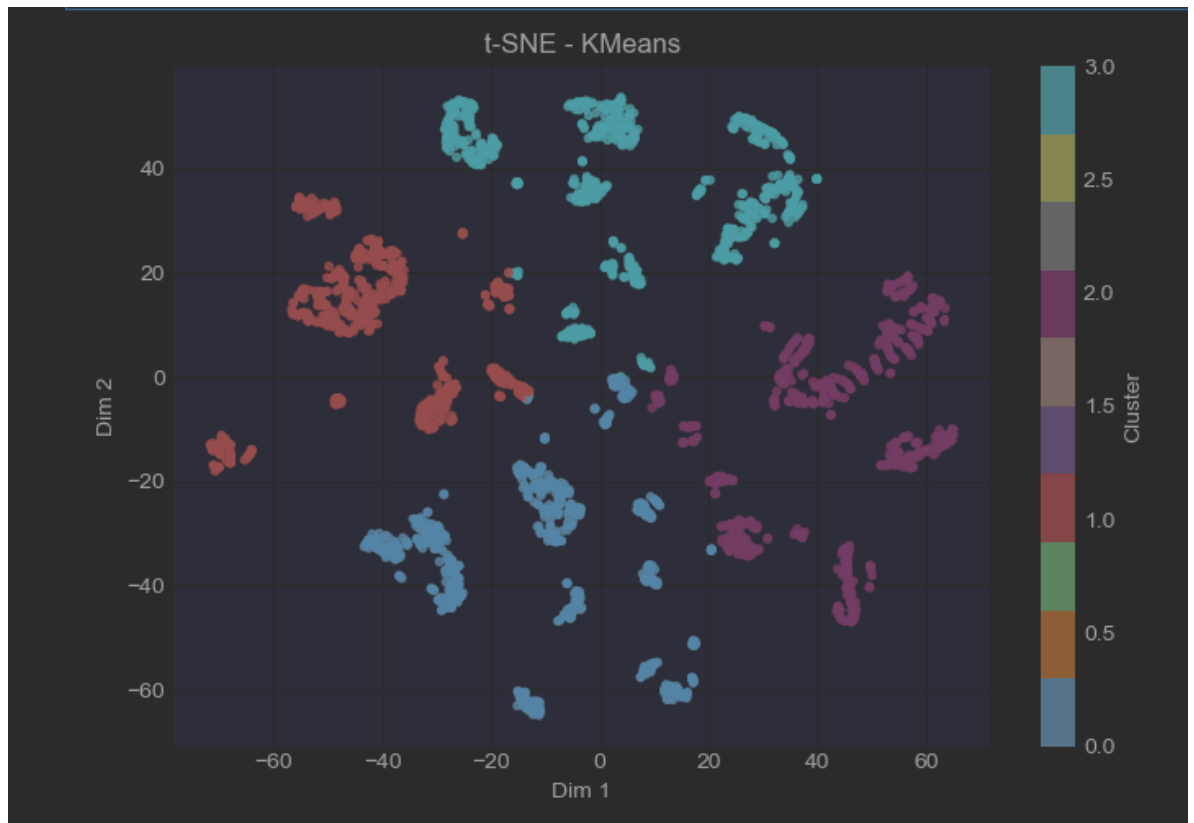
Esta técnica mostró **la mayor claridad visual** en la separación entre grupos, lo cual sugiere que el espacio transaccional tiene relaciones no lineales fuertes.

Al aplicar clustering sobre el espacio t-SNE:

- **K-Means (k=4)** presentó el mejor balance visual entre separación y coherencia de grupos.
- Aplicamos método del codo para seleccionar K



- El clustering jerárquico también separó bien los subconjuntos, aunque con transiciones menos definidas.

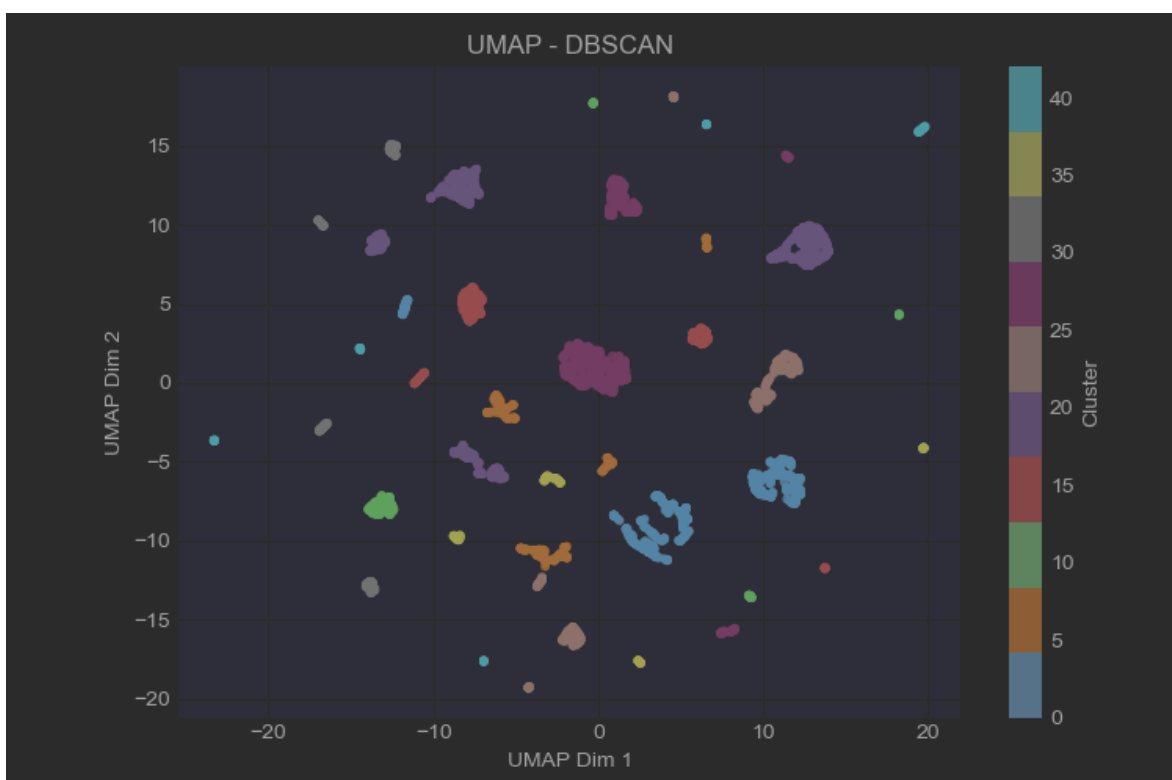


3. UMAP – Estructura Global + Separaciones Locales

La proyección UMAP entregó una distribución más **compacta**, con:

- grupos más redondeados y menos elongados que t-SNE,
- mejor preservación de distancias globales,
- menor dispersión dentro de cada estructura.

Esto lo convierte en un candidato natural para algoritmos basados en densidad, como DBSCAN.



3 rows 3 rows x 5 cols						
	Espacio	Método	Silhouette	Calinski-Harabasz	Davies-Bouldin	
0	t-SNE	KMeans	0.423874	2965.640450	0.760354	
1	t-SNE	Agglomerative	0.385596	2442.713031	0.828630	
2	UMAP	DBSCAN	0.755330	21394.573264	0.254173	

Estos indicadores revelan:

Este resultado indica que **UMAP + DBSCAN** captura mejor la geometría inherente de los comercios, detectando regiones densas y aislando valores atípicos sin depender de K.

Conclusión de la Sección de Reducción

- **PCA** sirvió para comprender la estructura lineal, pero no reveló clústeres significativos.
- **t-SNE** mostró una separación visual excelente y evidenció la presencia de múltiples patrones no lineales.
- **UMAP** equilibró estructura global y separación local, y permitió obtener el mejor resultado de clustering al aplicarlo con DBSCAN.

En conjunto, las reducciones demostraron que los comercios sí presentan patrones diferenciables, y que estos se manifiestan con mayor claridad cuando se emplean métodos no lineales.

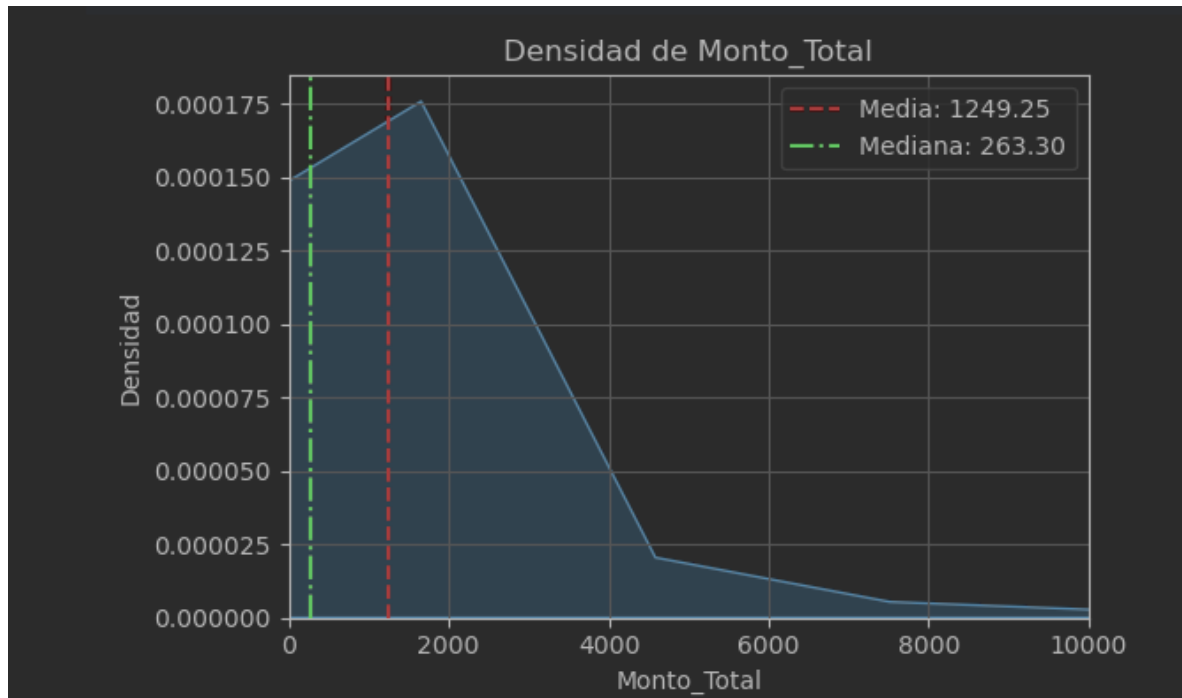
Resultados del clustering

Resultados aplicando K-means para segmentar comercios

Para complementar la segmentación obtenida mediante técnicas de reducción de dimensionalidad, se desarrolló un modelo paralelo enfocado únicamente en el comportamiento económico de cada comercio, medido a través del Monto Total transaccionado. Este enfoque unidimensional es común en análisis crediticio y en estrategias de categorización comercial, ya que permite identificar niveles de actividad de manera clara y cuantitativa.

Análisis inicial de la variable Monto_Total:

La distribución del monto total presentó un sesgo extremo a la derecha, característico de datos financieros: la mayoría de los comercios transaccionan valores relativamente bajos, mientras que un grupo reducido concentra montos muy altos.



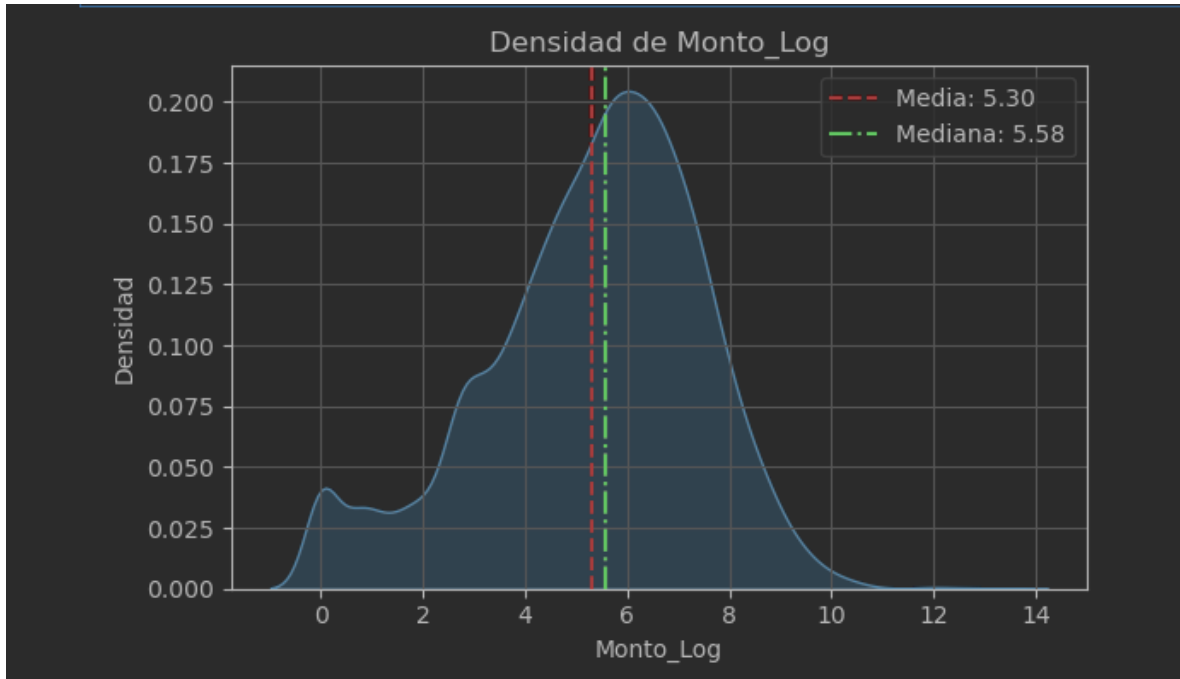
Este patrón se evidencia fácilmente al comparar:

- **Media \approx 1249 USD**
- **Mediana \approx 263 USD**

La diferencia entre ambos valores sugiere un comportamiento altamente asimétrico, pero está dentro de lo esperado debido a las operaciones de comercios.

Para mitigar el impacto de los outliers naturales y permitir que K-Means pueda formar grupos más estables, se aplicó la transformación logarítmica

Tras esta transformación, la distribución se vuelve mucho más simétrica, la media y mediana se acercan y se reducen los extremos que distorsionaban la distancia euclidiana en K-Means.

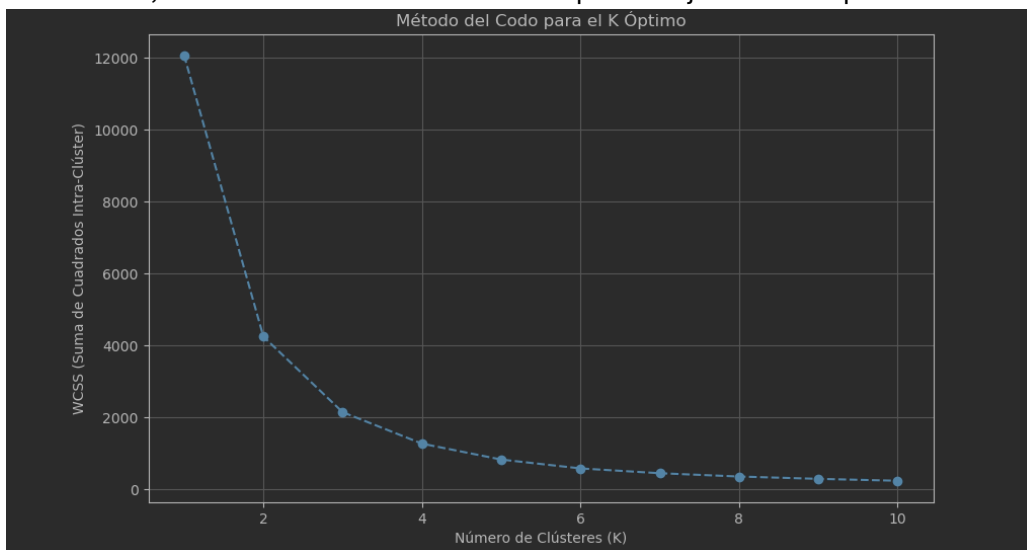


3. Selección del número óptimo de clusters (K)

Se utilizó el **Método del Codo** para evaluar la curva de reducción de inercia en función del número de clusters.

Los resultados muestran caídas pronunciadas hasta **K = 4**.

Por lo tanto, se seleccionó **K = 4** como valor óptimo bajo este enfoque.



4. Resultados del modelo K-Means

Al entrenar el modelo con la variable logarítmica estandarizada, obtenemos cuatro segmentos bien diferenciados. Los boxplots en escala logarítmica muestran que cada grupo concentra niveles de actividad económica claramente distintos:

Segmento 0 – Comercios de bajo volumen

Montos muy reducidos. En su mayoría comercios pequeños o de reciente incorporación.

Segmento 1 – Comercios de volumen medio

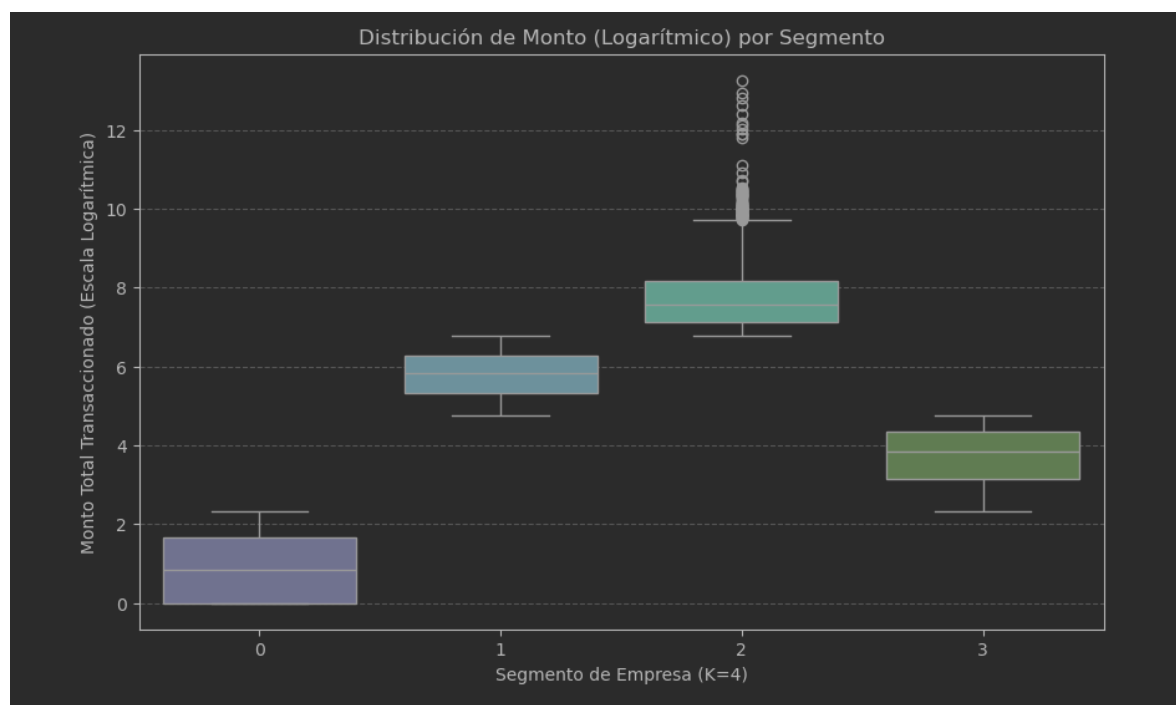
Actividad más estable, transaccionan de forma regular pero sin alcanzar volúmenes altos.

Segmento 2 – Comercios de alto/muy alto volumen

El clúster con mayor capacidad transaccional, concentración significativa de ingresos, esto representa una zona clave para estrategias de retención y riesgo, ya que son comercios que aportan alto volumen transaccional

Segmento 3 – Comercios de volumen bajo-medio

Podrían representar negocios estacionales o con fluctuaciones. No todos los comercios presentan mismos patrones, por su naturaleza existen muchos que son intermitentes o unos que dejan de usarse en meses.



Conclusiones de K-means:

Estos segmentos identificados aportan gran valor a la empresa para detectar clusters de mayor importancia, a los cuales se puede focalizar estrategias como mejoras de comisiones, paquetes premium o renovación de equipo.

Dependiendo de la estrategia implementar podemos analizar los clusters en el tiempo, y ver los comercios que se desplazan de un cluster a otro, lo que ayudaría a la empresa a mejorar la retención.

Resultados aplicando DBSCAN + Harvesine para clusterizar ubicaciones de comercios

La empresa al dedicarse al rubro transaccional, también debe de poder clusterizar las ubicaciones de transacciones de los distintos comercios, esto con el fin de poder generar estrategias de negocio definidas en sectores específicos y además, poder detectar patrones inusuales en las transacciones.

Un comercio puede movilizarse con su POS, y si se logran ver patrones extraños en distancia de transacciones, puede dar sospechas de fraude, esto ayuda a la empresa a fortalecer métricas a comercios sospechosos.

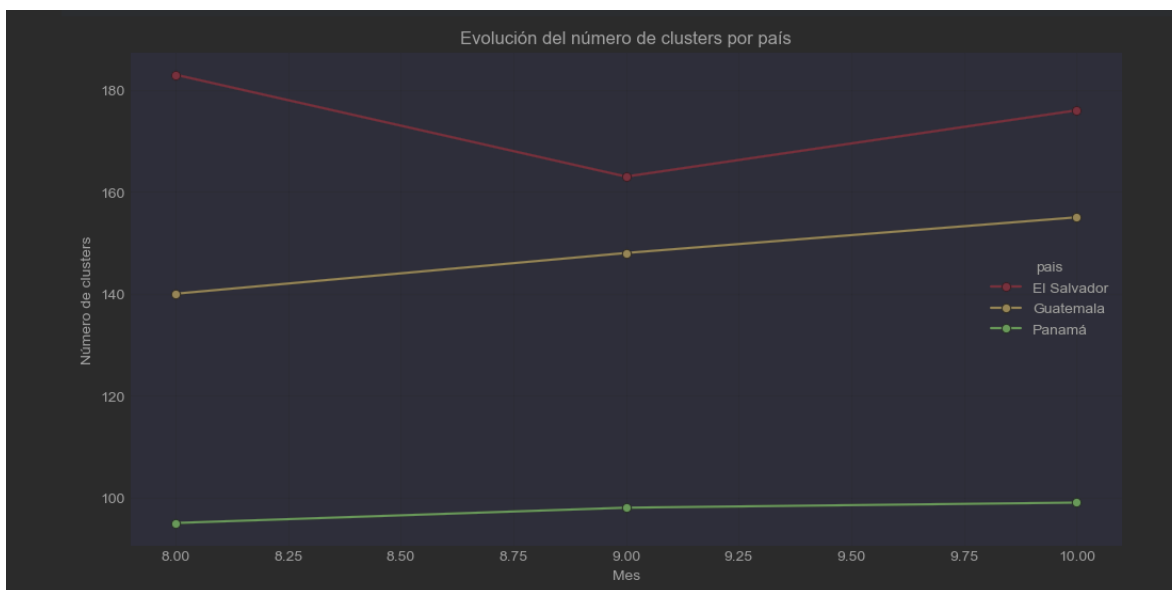
En este análisis usamos DBSCAN con la métrica de Harvesine para poder clusterizar las ubicaciones de las transacciones.

Solo se usarán las transacciones de tipo POS ya que son las que guardan la latitud y longitud de la transacción, además, se debe de clusterizar en base a los países, así no mezclamos transacciones de Guatemala con El Salvador o viceversa.

Una regla de negocio aplicada acá es sobre el la cantidad de transacciones mínimas y la distancia entre transacciones que se usan para poder definir un “cluster”. Ya que el análisis que se busca hacer es a nivel MACRO, la cantidad de puntos mínimos se dejo en 5 y el valor de epsilon (la distancia en metros) es de 1.5 km.

Epsilon puede parecer un valor alto, pero como se busca clusterizar a nivel macro, evidenciando zonas a nivel Regional, 1.5km ha sido la métrica a usar.

análisis de resultados:



El comportamiento de DBSCAN revela patrones geográficos y operativos interesantes:

El Salvador

- Presenta el mayor número de clusters, consistentemente entre 160 y 185 por mes.
- Esto refleja una base comercial altamente diversa y con comportamientos densos en zonas urbanas.
- También puede indicar mayor dispersión en montos y frecuencia de transacciones respecto a los otros países.

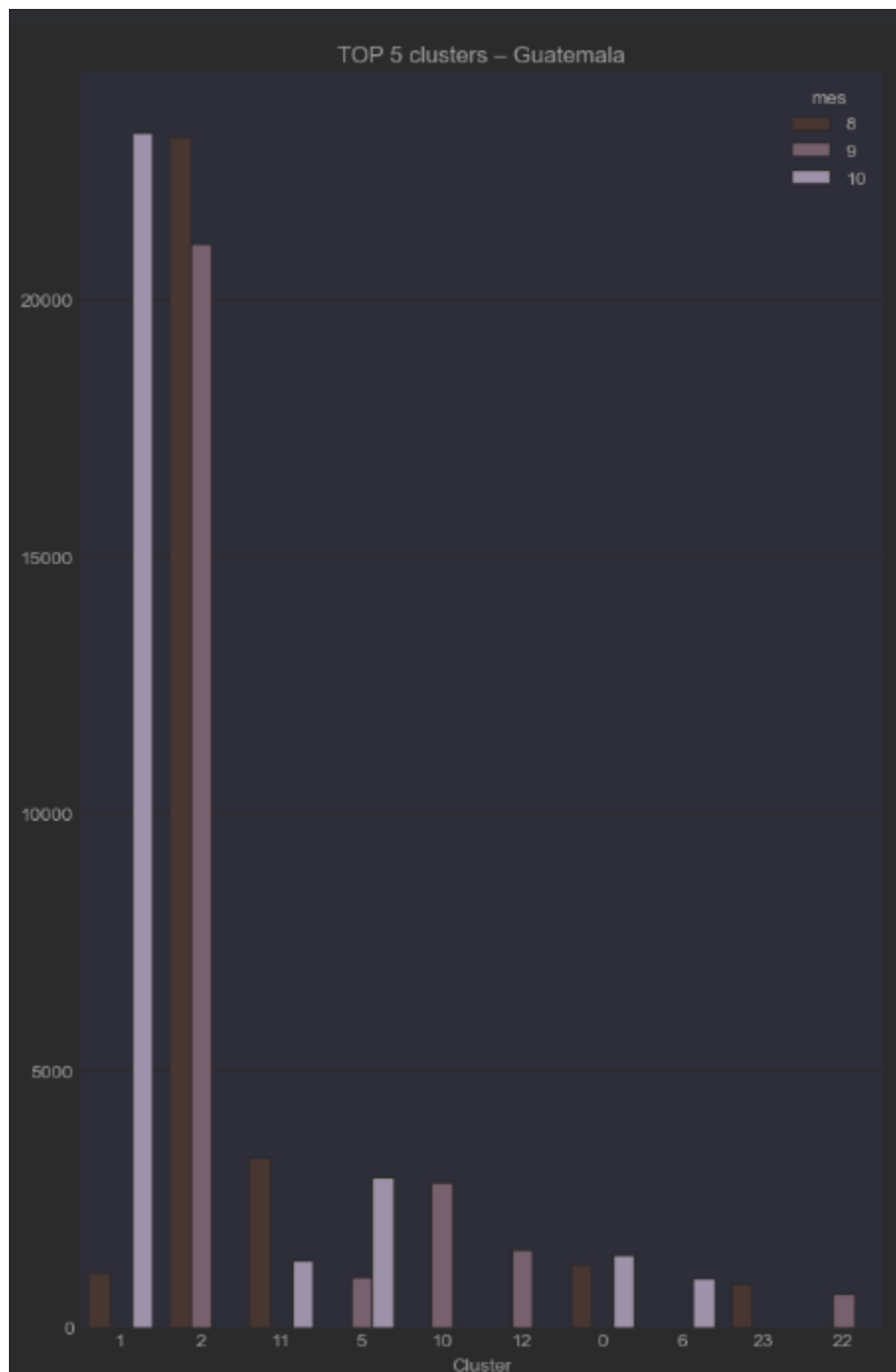
Guatemala

- Muestra una tendencia creciente: de ~140 a ~155 clusters.
- La estructura es estable, pero con un incremento progresivo de densidad transaccional.
- Sugiere crecimiento real del uso del servicio, lo cual refuerza campañas de mejora aplicadas

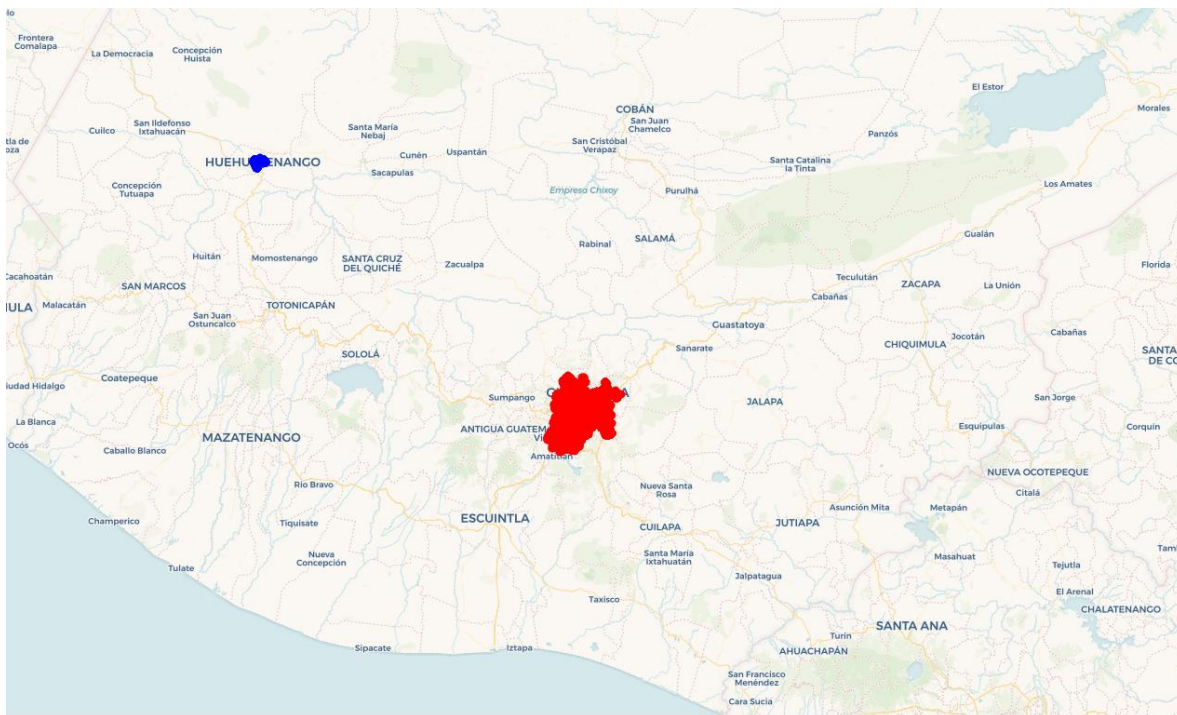
Panamá

- Clara estabilidad entre 95 y 100 clusters.
- Menor densidad de comercios y patrones agregados más uniformes.
- Representa un mercado más concentrado y homogéneo.

Un ejemplo de un análisis importante e insight imperativo a ver , es la cantidad de transacciones por cluster. En el caso de Guatemala tenemos lo siguiente:



Si graficamos los dos clusters más importantes de Guatemala (cluster 1 y 2)



Podemos evidenciar como la Zona central de GT abarca la gran cantidad de transacciones hechas. Como se mencionó anteriormente, la idea es realizar un análisis Macro de las transacciones, una vez identificado los Clusters Padres (rojo) podemos seleccionar solo esas transacciones y volver a aplicar DBSCAN con un número de puntos y eps distritos, así habremos sub-clusterizado un cluster padre en segmentos más pequeños.

Interpretación de resultados en el contexto del dominio de aplicación

Los resultados obtenidos mediante las técnicas de reducción de dimensionalidad y clustering fueron interpretados basado en el comportamiento transaccional real de los comercios que usan la plataforma de pagos. En este contexto, cada cluster representa un grupo de comercios con características operativas similares, agrupados no por categorías declaradas, sino por patrones observados directamente en las transacciones.

1. Clusters como segmentos operacionales de comercios

Los modelos revelan segmentos diferenciados de comercios según su actividad real:

- Comercios de alto volumen y alta frecuencia, que concentran la mayor parte del GPV.
- Comercios de actividad media, con patrones regulares pero menor densidad.
- Comercios intermitentes o estacionales.
- Comercios de bajo volumen, posiblemente recién afiliados o con uso eventual del servicio.

Esta segmentación permite entender cómo se distribuye la actividad económica dentro del ecosistema y cuáles comercios aportan mayor estabilidad.

2. Resultados relevantes por país

El análisis muestra diferencias claras entre los mercados:

- **El Salvador** presenta mayor fragmentación y diversidad operativa, es decir, más clusters.
- **Guatemala** evidencia crecimiento en densidad y nuevos segmentos emergentes.
- **Panamá** mantiene una estructura más homogénea y estable.

Estas diferencias son importantes para adaptar estrategias comerciales o de soporte según el país.

3. Interpretación operativa de los resultados

Los clusters identificados ayudan a responder preguntas del negocio como:

- ¿Qué grupos de comercios generan la mayor parte de las transacciones?
- ¿Dónde se concentran los comercios de riesgo por comportamiento atípico?
- ¿Qué segmentos muestran patrones ideales para campañas de retención?

Interpretación final:

En conjunto, los resultados permiten construir una segmentación basada en datos que, refleja el comportamiento real del ecosistema de la fintech. Es útil para priorizar recursos, mejorar la toma de decisiones y entender dinámicas operativas. También Sientan la base para futuros modelos de churn, detección de fraude, recomendación y scoring de comercios.

Estos modelos aplicados ayudan a tomar decisiones basadas en datos, la empresa ahora puede segmentar sus estrategias, focalizar o reforzar segmentos que tienen alta probabilidad de perdida. Esto abre un mundo enorme a la Fintech para poder gestionar sus tareas operacionales y comerciales, es un paso grande a una mejora continua.

Insights accionables y recomendaciones estratégicas

A partir de estos resultados, emergen varias líneas de acción útiles para la estrategia de la empresa.

En primer lugar, la segmentación transaccional mostró que los comercios no forman un grupo homogéneo, sino que se distribuyen en niveles bien diferenciados de actividad. Se observan comercios con volúmenes altos y consistentes, otros con actividad media pero estable, un grupo notablemente estacional y, finalmente, un conjunto amplio de comercios de bajo volumen.

Cada uno de estos segmentos requiere un tratamiento distinto: los de alto volumen deben priorizarse en estrategias de retención y fidelización, mientras que los de volumen medio representan la mayor oportunidad de crecimiento si reciben capacitación y acompañamiento adecuado.

Al observar las proyecciones con t-SNE y UMAP, quedó claro que el comportamiento real de los comercios no sigue patrones lineales simples. Más bien existen “microcomunidades” que comparten hábitos muy específicos, desde horarios hasta montos o modos de operar.

Estas técnicas mostraron que las relaciones entre comercios son más complejas de lo esperado, lo cual abre la puerta a análisis más precisos y personalizados en el futuro. De hecho, UMAP resultó particularmente útil porque conserva muy bien la estructura local de los datos, por lo que se vuelve una excelente herramienta preliminar para tareas como segmentación avanzada, detección temprana de anomalías o incluso modelos de recomendación interna.

En conjunto, estos hallazgos ofrecen una visión más profunda del portafolio de comercios. Permiten pasar de un enfoque general a uno verdaderamente estratégico, donde las acciones pueden adaptarse a cada tipo de comercio de manera más informada. Esto no solo mejora decisiones comerciales y operativas, sino que también puede integrarse con modelos predictivos futuros, reduciendo incertidumbre y anticipando comportamientos clave dentro del ecosistema.

Conclusiones y trabajo futuro

El análisis aplicando técnicas de reducción de dimensionalidad y clustering permitió descubrir patrones relevantes en el comportamiento de los comercios, evidenciando que existen grupos bien diferenciados según su actividad transaccional y su comportamiento geoespacial. Estos hallazgos confirman que la base de clientes no es homogénea y que la analítica no supervisada aporta una visión más profunda que la simple agregación de datos.

Sin embargo, el estudio se limitó a un periodo de tres meses, lo cual restringe la capacidad de detectar estacionalidad anual, tendencias prolongadas o efectos comerciales de largo plazo. Asimismo, se identificaron comercios con ubicaciones dispersas o fuera del país, lo que abre necesidades de análisis más detallado y potenciales alertas operativas.

Este trabajo representa un punto de partida sólido, pero con amplio espacio para evolucionar hacia análisis más robustos y predictivos.

Hallazgos relevantes:

- Existen segmentos claros de comercios según volumen transaccional, comportamiento operativo y patrones geoespaciales.
- Se detectaron comercios con actividad atípica o ubicaciones inusuales, útiles para análisis de riesgo o validación operativa.

Trabajo futuro recomendado:

- Ampliar el análisis a un año completo o más para capturar estacionalidad y tendencias.
- Evolucionar hacia modelos predictivos: churn, potencial de crecimiento, alertas tempranas.
- Incorporar estos hallazgos en la estrategia comercial y operativa para priorizar esfuerzos y optimizar recursos.

En síntesis, este proyecto abre la puerta a un uso más estratégico de los datos y establece la base para análisis avanzados que pueden transformar la gestión del portafolio de comercios.