Taller No. 2. Análisis univariado.

Farid Sandoval
Mateo Silva
Jhonattan Reales
Josué Cobaleda

1) Introducción

1.1. Generalidades del reto y problema de interés.

Este análisis se enmarca en el Reto de Analítica de Datos, Inteligencia Artificial y Machine Learning del Banco W, cuyo objetivo es desarrollar un modelo que permita identificar y clasificar a los clientes y agencias con alto riesgo financiero en microcréditos. Para ello, se aplicarán técnicas de análisis de datos e IA, que permitan mejorar la toma de decisiones en la asignación de analistas de auditoría.

El proceso de auditoría en microcréditos requiere evaluar el riesgo de los créditos otorgados en campo, identificando patrones de conducta inusual, fraudes y perfiles de alto riesgo en clientes y agencias. En el sector de las microfinanzas, la gestión del riesgo financiero es fundamental para garantizar la sostenibilidad de las entidades prestamistas y minimizar pérdidas derivadas de incumplimientos crediticios.

Definición de microcréditos: Son préstamos de pequeña cuantía otorgados a personas de bajos ingresos o con dificultades de acceso a la banca tradicional, con el objetivo de fomentar el autoempleo y el desarrollo de pequeños negocios.

Características de los microcréditos: tienen montos reducidos, plazos de devolución cortos, tasas de interés accesibles y una evaluación del riesgo crediticio basada en la capacidad y voluntad de pago del solicitante, en lugar de garantías tangibles. En Colombia, un microcrédito no puede exceder los 120 salarios mínimos legales vigentes (SMLV). Este tipo de crédito está dirigido a microempresarios y personas independientes.

1.2. Objetivo claro - Pregunta SMART

"¿Cómo podemos desarrollar un modelo basado en análisis de datos e inteligencia artificial para identificar créditos, analistas y agencias con alto riesgo financiero en la cartera de microcréditos del banco W, utilizando características del crédito, historial de auditorías y factores de estabilidad financiera, con el fin de priorizar las auditorías internas y mejorar la eficiencia del proceso en un período de seis meses?"

¿Por qué es SMART?

Específica: Se especifica que se usará análisis de datos e inteligencia artificial. Se centra en créditos, analistas y agencias de alto riesgo, dentro del contexto de auditoría de microcréditos.

Medible: Se establece que el modelo debe identificar casos de alto riesgo y mejorar la eficiencia de auditorías.

Alcanzable: Puede resolverse con análisis exploratorio de datos (EDA modelos de IA, y se cuenta con los datos del área de auditoría del banco.

Relevante: Es clave para mejorar la gestión del riesgo financiero de microcréditos y la selección de auditorías internas.

Con límite de tiempo: Se define un horizonte de seis meses para la implementación. Puede abordarse en el plazo del curso.

1.3. Revisión de la bibliografía

El análisis del riesgo crediticio de cartera ha sido ampliamente estudiado en el contexto de la auditoría financiera y la auditoría bancaria. La investigación destaca que la auditoría interna es importante en la gestión del riesgo crediticio, permitiendo a las entidades financieras mitigar pérdidas y mejorar sus procesos de toma de decisiones (Berisha et al., 2023). Estos estudios se enfocan en el uso de metodologías cuantitativas, como el análisis univariado, para identificar patrones de riesgo dentro de la cartera crediticia.

Por un lado, Moposita y Ramírez (2016) presentan un marco de auditoría para cooperativas de ahorro y crédito, donde se destaca la importancia de realizar exámenes detallados sobre el comportamiento de los clientes. Este enfoque permite establecer relaciones entre variables clave como el monto del crédito, la morosidad y la probabilidad de incumplimiento.

Por otro lado, Hernández Bautista (2023) introduce una herramienta de análisis financiero que segmenta clientes según su nivel de riesgo. En este estudio, se resalta que el análisis de variables individuales, como el monto del crédito, permite predecir con mayor precisión la estabilidad financiera de un cliente, facilitando procesos de auditoría y toma de decisiones estratégicas.

Estos estudios confirman que el análisis univariado es una técnica clave en la identificación y supervisión del riesgo crediticio, especialmente en auditorías de microcréditos. Su implementación permite a las instituciones financieras identificar patrones en los datos históricos y mejorar la calidad de su cartera crediticia.

1.4. Objetivo del taller

En este contexto, como equipo buscamos aplicar técnicas de análisis univariado, segmentación de clientes según el riesgo y detección de patrones anómalos para determinar qué factores son más relevantes en la identificación de clientes y agencias con alto riesgo financiero, contribuyendo a mejorar la gestión de riesgos en el Banco W.

Sin embargo, como primer paso, en este taller se hará un análisis univariado teniendo en cuenta que ya seguimos los anteriores pasos del análisis exploratorio de datos (EDA, por sus siglas en inglés) como lo son la pregunta smart y una vista general de las columnas.

Nota: en este caso no se cuenta con un diccionario de datos que nos permita determinar el significado real de algunas variables del dataset.

Para este punto buscamos establecer qué variables son más útiles y que nos pueden decir esta con respecto al análisis que queremos hacer y a la respuesta de nuestra pregunta smart, para ello se buscará información externa que nos pueda ayudar y además hacer algunos de los procesos que se deben seguir en el análisis univariado como datos faltantes, outliers y distribución.

2) Selección e Importancia de la variable escogida: Variable Monto.

2.1. Generalidades de la data a utilizar

Antes de elegir la variable, presentamos las dos bases de datos principales:

- **Histórico de Hallazgos:** Contiene información sobre auditorías, clientes evaluados, oficinas y zonas de colocación, tipo de crédito, estado del crédito y clasificación de hallazgos relacionados con la estabilidad financiera.
- **Colocación de Créditos:** Proporciona detalles sobre créditos otorgados, identificaciones de clientes, montos, tasas de interés, plazos y valores de cuota.

El histórico de auditorías desde 2022 a 2024 permite analizar tendencias en los hallazgos detectados, mientras que la base de colocación de créditos de octubre de 2024 permite identificar riesgos en la cartera más reciente.

Sin embargo, los datos presentaban:

- Cantidad significativa de valores faltantes, que no se pueden reemplazar fácilmente dado la cantidad.
- Muchas variables categóricas, en las que se podrían aplicar metodologías como One Hot encoding.
- Varios ID que no aportan información relevante, aunque se podrían identificar clientes mayores con cédulas de números más bajos y clientes más jóvenes con cédulas de número más altos.

2.2. Selección de la variable

Para abordar el reto desde un análisis univariado, consideramos diversas variables que podrían aportar a la evaluación del riesgo financiero en microcréditos, entre ellas:

- **Tipo de crédito:** Diferenciar entre créditos nuevos y renovaciones.
- **Estado del crédito:** Clasificación según su situación (cancelado, castigado, modificado, etc.).
- **Saldo:** Monto pendiente de pago, reflejando la deuda real del cliente.
- **Monto:** Valor total del crédito otorgado.
- **Tasa de interés:** Indicador del costo financiero y del perfil de riesgo del cliente.
- **Plazo del crédito:** Duración del préstamo (corto, mediano o largo plazo).

- **Ubicación de la oficina:** Posible relación entre riesgo y la distancia de las agencias a ciudades capitales.
- **Historial de hallazgos en auditorías:** Registro de clientes o agencias previamente categorizados como riesgosos.
- **Actividad económica del cliente:** Sectores con mayor o menor probabilidad de incumplimiento.

Sin embargo, la selección de la variable "Monto" se justifica por los siguientes criterios clave:

1. El monto es un indicador de exposición financiera: Créditos más grandes implican un mayor compromiso financiero, aumentando el riesgo de incumplimiento. Estudios previos han demostrado que los créditos de mayor monto suelen presentar un mayor riesgo, debido a la mayor carga financiera que representan para los prestatarios.
2. La variable "monto" del crédito es una de las más relevantes en el análisis de riesgo de cartera crediticia, ya que impacta directamente la probabilidad de incumplimiento y la estabilidad de la misma (Berisha et al., 2023).
3. Detección de anomalías y fraudes: Valores atípicos en los montos pueden ser señales de inestabilidad financiera o posibles irregularidades.
4. Respaldo en estudios previos:

   Investigaciones como la de Schreiner (2000) han demostrado que el monto del crédito es un factor determinante en la evaluación del riesgo crediticio.

   Moposita y Ramírez (2016) resaltan que el monto del crédito es un indicador clave en auditorías financieras, ya que permite segmentar a los clientes según su nivel de exposición al riesgo. Su análisis a través de técnicas univariadas, como histogramas y boxplots, facilita la identificación de outliers y posibles patrones de incumplimiento.

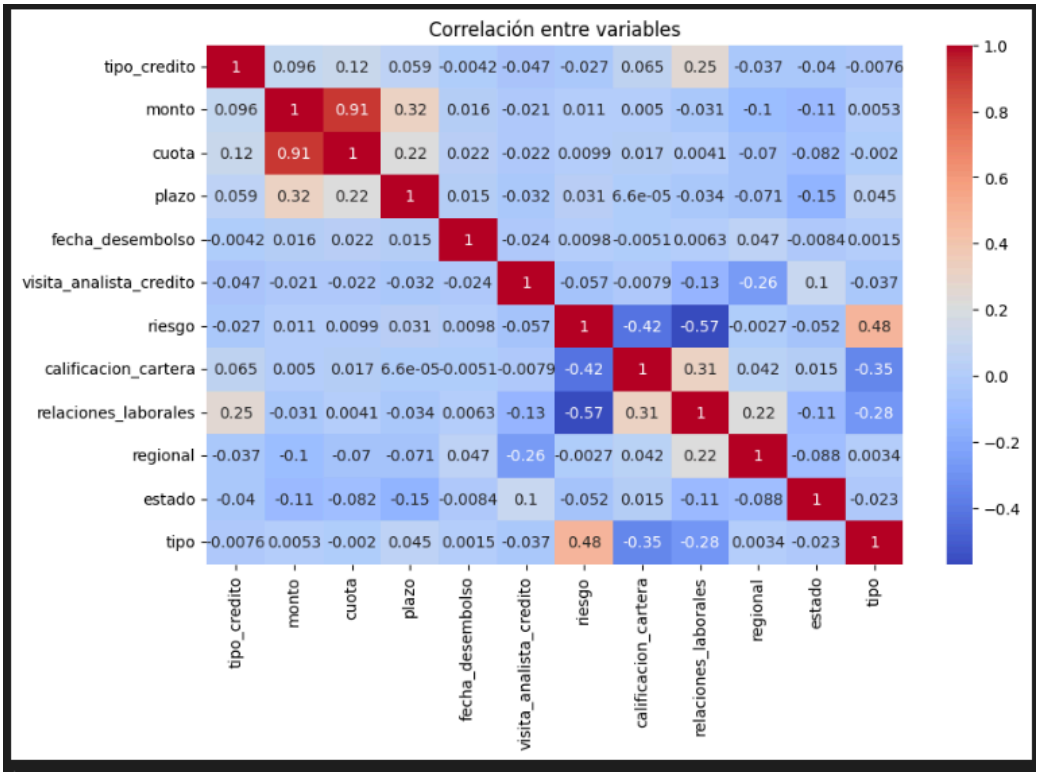
   Por su parte, Hernández Bautista (2023) establece que la evaluación de la variable "monto" en combinación con otras métricas financieras permite mejorar la segmentación del riesgo crediticio. En particular, su investigación demuestra que la relación entre el monto del crédito y el comportamiento de pago de los clientes es un factor determinante en auditorías de cartera.

5. Viabilidad para análisis avanzados: La Ley de Benford solo puede aplicarse a variables numéricas que se generan de forma natural, como el monto del crédito, permitiendo detectar posibles manipulaciones en los datos.
6. Segmentación de clientes según riesgo: La distribución del monto puede revelar patrones de riesgo y ayudar a identificar clientes con mayor probabilidad de incumplimiento.

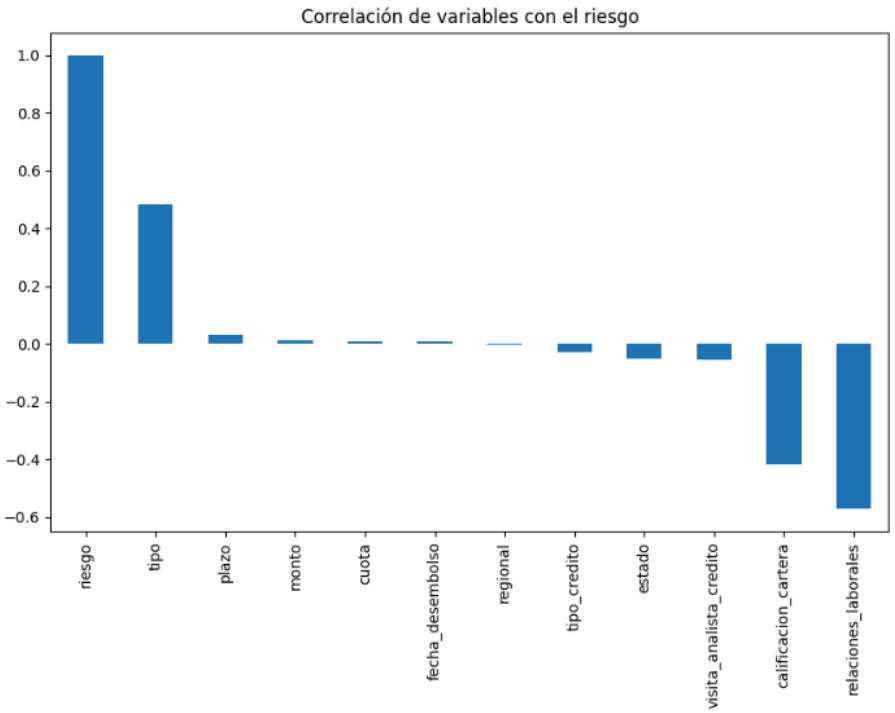
La selección de la variable "monto" en este análisis univariado está respaldada por fuentes investigativas como un criterio importante en la evaluación del riesgo crediticio. Su análisis proporciona información valiosa sobre la estabilidad financiera de los clientes y permite mejorar la precisión en la identificación de riesgos dentro del proceso de auditoría de una entidad bancaria, en este caso, Banco W.

2.2.1. Selección de la variable de un punto de vista cuantitativo.

Después de realizar un análisis de correlaciones obtuvimos los siguientes resultados.



y para la variable riesgo, la cual es nuestra variable objetivo, obtuvimos que las más correlacionadas son:



Para este punto, después de haber buscado información que nos fuera útil, nos dimos que cuenta de algo muy interesante, pues en nuestra investigación teórica vimos que este monto de los microcréditos podría estar muy relacionado con los créditos, pero haciendo un análisis univariado, nos dimos cuenta que la relación era de casi cero, lo cual no nos hacía sentido, es por esto que, después de ver que habian muchos datos atípicos, decidimos analizar esta variable para entender el porque está pasando este fenómeno.

2.4 ¿Cómo podría influir esta variable en el análisis o toma de decisiones?

Por último, la variable monto nos permitiría obtener información valiosa sobre el proceso y criterios de otorgación de microcréditos del banco. Un análisis univariado de la variable monto en este contexto de auditoría de riesgos, nos permitiría extraer información clave sobre la distribución, comportamiento y patrones de los montos otorgados.

Ahora bien, si la variable monto resulta importante para análisis posteriores y modelos de IA implementados, esto nos indicaría que dicha variable podría facilitar la toma de decisiones estratégicas en auditoría y la optimización del proceso de detección de riesgos.

¿Cómo influye en la toma de decisiones?

Mejor gestión del portafolio de créditos:

- Si el análisis muestra que prestamos grandes tienen más incumplimiento, se podrían ajustar los criterios de aprobación.
- Se podrían diseñar estrategias para ofrecer montos más adecuados según el perfil de riesgo del cliente.

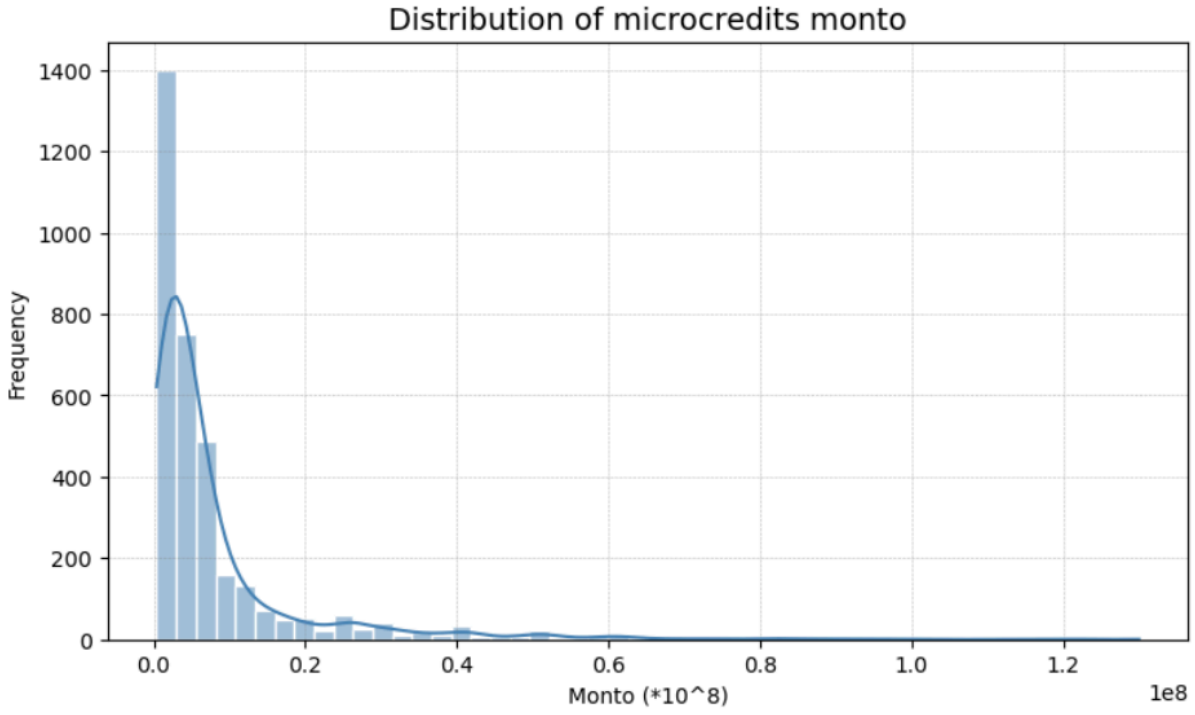
Definición de políticas de auditoría:

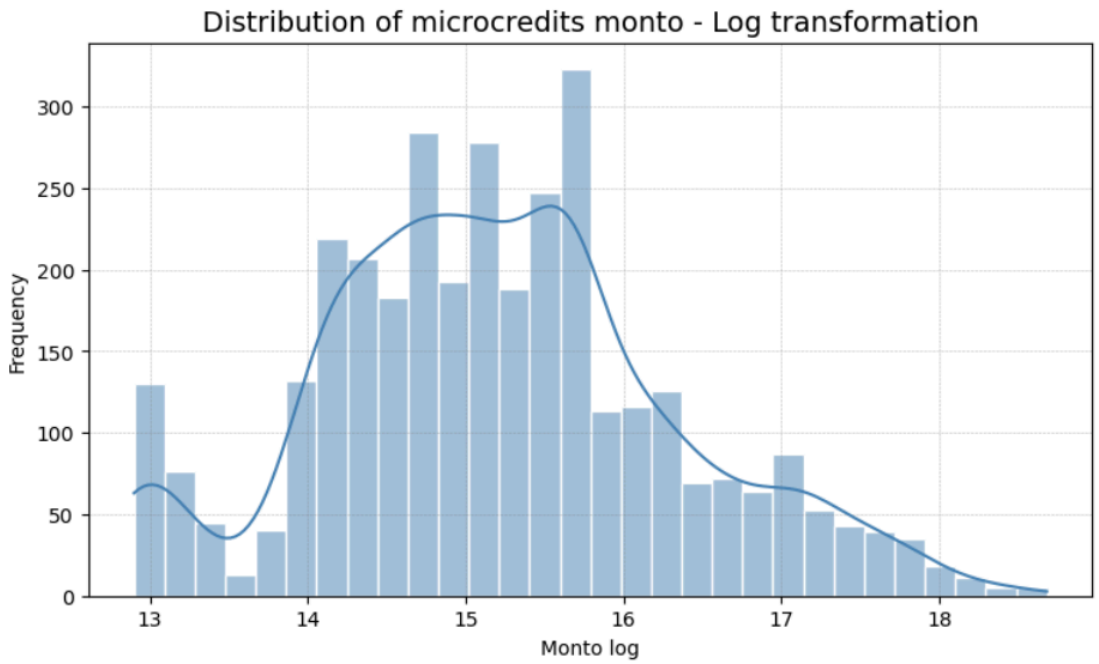
- Las auditorías pueden priorizar agencias que concedan créditos de alto monto o clientes con patrones de riesgos detectados.
- Se podría ajustar condiciones de pago en función del riesgo asociado al monto.

3) Análisis univariado en Python

METRICS AND DISTRIBUTION

```
¡Let's calculate some important metrics!


--Central tendency:
Mean: $7,931,928.0
Median: $3,736,415.0
Mode: $400,000
Min: $400,000
Max: $130,000,000


--Measures of dispersion:
standard deviation: $12,339,723.4
Variation coefficient: 155.6
Range: $129,600,000
IQR: $5,401,416.0

--Measures of distribution shape:
Skewness: 3.9
Kurtosis: 20.7

--Some percentiles:
['P10: $1,132,013', 'P25: $1,852,791', 'P50: $3,736,415', 'P75:
$7,254,207', 'P90: $19,691,704']
```





OUTLIERS CHECK

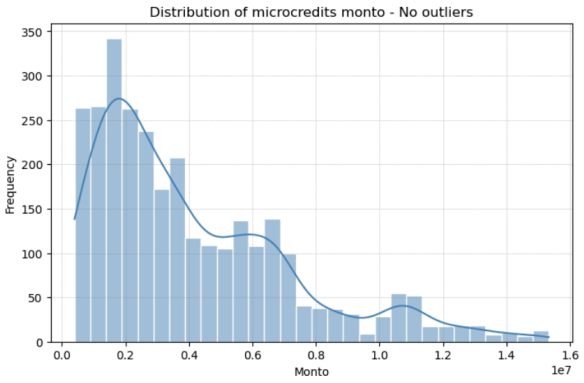
Lets calculate some outliers metrics:

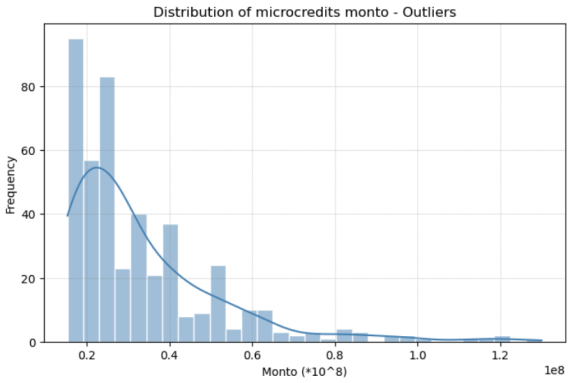
Outliers upper limit: $15,356,331.5
Outliers lower limit: $0
Number of microcredits consider as outliers: 447
% of outliers of the total observations: 13.09%

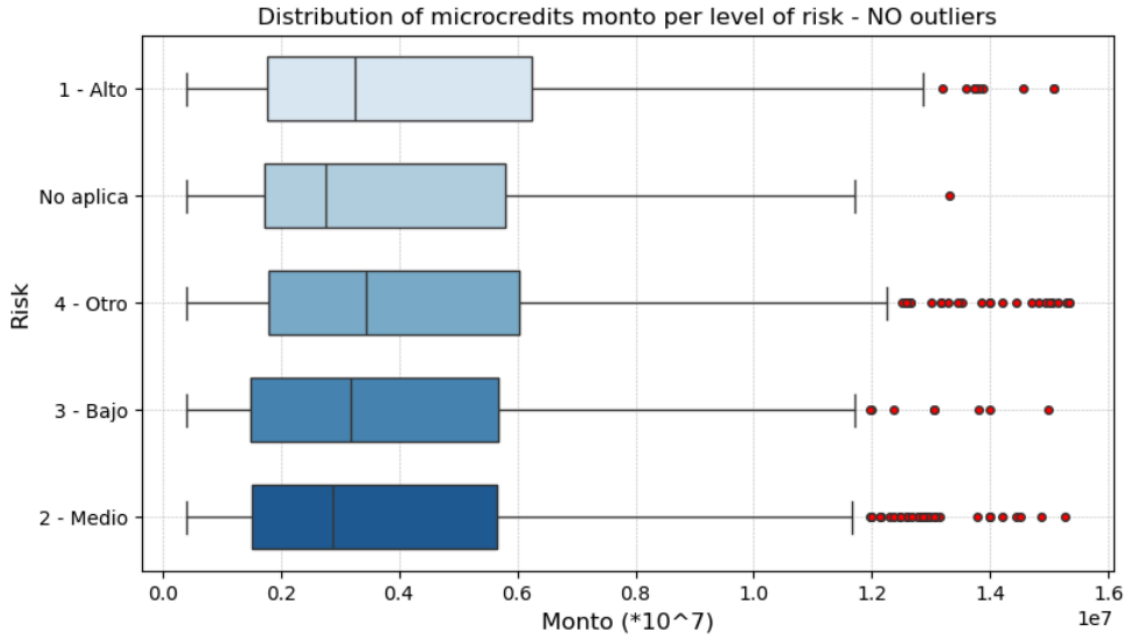
OTHER ANALYS AND PATTERNS

the most frequently requested loans are:

monto	
400000	123
500000	75
1000000	33
600000	32
1500000	31
1631982	21
2345852	14
800000	13
700000	12
2431982	11







4) Interpretación de resultados y hallazgos relevantes

A partir de las métricas, gráficos y análisis realizados sobre la variable monto, podemos resaltar los siguientes hallazgos:

- En primer lugar, notamos un rango muy amplio de monto de créditos, dado que, estos van desde $400.000 hasta $130.000.000. Lo anterior nos demuestra que la cartera de microcréditos cuenta con una variedad de perfiles de clientes amplia.

- La media de los montos ($7,931,928) es significativamente mayor que la mediana ($3,736,415), lo que sugiere la presencia de valores extremos u outliers, es decir, la presencia de créditos de muy alto monto los cuales mueven el promedio.

- El monto más solicitado (moda) es de $400,000, solicitado un total de 123 ocasiones, lo que indica que la mayoría de los créditos son pequeños.

- Al aplicar una transformación logarítmica a la variable monto, observamos que los datos tienden a comportarse como una campana de gauss. Esto es un comportamiento común en distribuciones muy sesgadas.

- Revisando las métricas de dispersión, notamos la presencia de una alta dispersión en los montos otorgados. La alta desviación estándar ($12,339,723) y el coeficiente de variación (155.6%) indican que los montos de los microcréditos son muy variables o muy distintos. Esto nos indica que los créditos y montos aprobados dependen de diversos factores y criterios del banco.

- La asimetría de los datos (3.9) nos indica que la distribución tiene un sesgo positivo fuerte. Lo que sugiere que la distribución de los montos presenta una cola larga hacia la derecha (montos altos poco frecuentes pero muy grandes). La curtosis (20.7) nos indica lo mismo, la presencia de un pico de datos pronunciado y de valores extremos.

- Al revisar los percentiles de la data, observamos que el 75% de los créditos son menores a $7,254,207 y solo el 10% de los créditos superan los $19,691,704. Esto muestra que la mayoría de los créditos están en rangos relativamente bajos, mientras que hay algunos montos elevados que pueden representar riesgos mayores.

- Al desarrollar el análisis de valores outliers, notamos que más del **13%** de los microcréditos están por encima del límite de outliers. Estos montos atípicos podrían estar asociados a un mayor riesgo de fraude, deficiencias en la asignación del crédito o condiciones especiales que justifican montos más altos. Es importante analizar si estos créditos presentan más hallazgos de auditoría en comparación con los montos más bajos.

- Al analizar los valores de monto más frecuentes, encontramos que los microcréditos más comunes están en un rango por debajo de $2,000,000.

- Por último, podemos apreciar que al dividir la data entre observaciones por debajo del umbral de valores outliers y por encima de este umbral, la forma y distribución para cada conjunto de datos es muy similar, con un sesgo pronunciado hacia la derecha, lo cual evidencia que la mayoría de los créditos son de bajo monto.

5) Conclusiones del análisis

A manera de conclusión, el análisis univariado sobre la variable monto nos brindó los siguientes hallazgos claves con respecto a la data y el contexto a trabajar:

- **Alta dispersión y presencia de valores extremos en los montos de crédito:**
  La variable monto presenta un sesgo positivo y una alta dispersión, dado que, los créditos de alto monto son pocos. Este aspecto se debe tener en cuenta a la hora de analizar el riesgo de la cartera de microcréditos. Si se llegase a identificar qué créditos de alto monto son propensos a presentar mayor riesgo, esta variable sería de mucho interés para cualquier modelo y/o análisis inferencial.

- **Identificación de un 13% de créditos como outliers y su posible relación con fraude:**
  Notamos una alta cantidad de outliers bajo el criterio del 1.5*IQR. Cerca del 13% de los montos se podrían considerar outliers. Esta población de altos montos, pueden ser interesantes para mayores y más profundos análisis de riesgos. Se pueden desarrollar reglas diferenciadas para evaluar y modelar riesgos en montos bajos y altos.

- **Concentración de créditos en montos bajos y necesidad de analizar otros factores de riesgo:**
  La mayoría de los créditos están en rangos bajos, los créditos más comunes rondan entre los $400.000 y $1,500,000. Si dentro de este segmento se encuentran diferentes niveles de riesgo, esto sugiere que existen otras condiciones y factores del crédito que influyen en la probabilidad de incumplimiento, más allá del monto otorgado.

En cuanto a próximos pasos, es de mucho valor desarrollar análisis exploratorios más profundos a la base de datos de hallazgos. En primer lugar, es importante realizar análisis univariados para las demás variables del conjunto de datos proporcionado. A continuación, podemos ganar mucho conocimiento llevando a cabo análisis bivariados y multivariados entre la variable monto, el riesgo y las otras variables de interés para el contexto.

Podemos hallar métricas claves de la variable monto, como percentiles, medidas de tendencia y dispersión, agrupado por otras variables categóricas como las oficinas, zonas, analistas, riesgo del crédito, categoría de los hallazgos, año y actividad económica. Para cada relación, podemos desarrollar diagramas de tipo boxplots, los cuales nos ayudan a identificar patrones o sesgos en las observaciones. Con lo anterior, podemos recopilar

mayores insights del contexto de los datos y generar preguntas o hipótesis que nos guíen en el desarrollo de las soluciones.

Adicionalmente, en el análisis bibliográfico hemos identificado que existen metodologías y reglas matemáticas que se le pueden aplicar a los montos de créditos para obtener mayor información sobre posibles actividades fraudulentas, tal es el caso de la regla llamada “Ley de Benford”. Como siguiente paso, podríamos emplear la Ley de Benford sobre el dato de monto para crear nuevas variables para un posible modelo de IA, o para realizar análisis descriptivos más rigurosos con respecto al perfil de riesgo de los créditos.