Parcial 1 - Análisis exploratorio de datos

Estudiante: Kevin Daniel Sierra Castro

Docente: Alexander Agudelo Cardenas

Electiva II Profesional

Ingeniería de Sistemas

Corporación Universitaria Rafael Núñez

Cartagena - Colombia

2024

Contenido

Introducción

Este trabajo se centra en el análisis de datos relacionados con solicitudes de aplicación y aplicaciones previas. Se utilizan dos conjuntos de datos: uno correspondiente a las solicitudes actuales y otro a las aplicaciones previas realizadas por los mismos clientes. El objetivo principal es explorar y visualizar las características numéricas y categóricas de los datos, así como identificar patrones que podrían estar relacionados con el incumplimiento de pagos.

Desarrollo

- Carga de datos

Se utilizan dos archivos .CSV para la carga de datos: Datos de aplicación, Contiene información sobre las solicitudes actuales. Previas Aplicaciones, Incluye detalles de aplicaciones anteriores de los mismos clientes.

Se utiliza la biblioteca “pandas” para leer los archivos y cargar los datos en DataFrames.

```
import pandas as pd
import matplotlib.pyplot as plt

datos_aplicacion = pd.read_csv('assets/Datos de Aplicacion.csv')
previas_aplicaciones = pd.read_csv('assets/Previas Aplicaciones.csv')
```

- Exploración inicial de los datos

Se exploran ambos DataFrames para comprender la estructura de los datos, tipos de columnas, y la presencia de valores faltantes.

```
print("Datos de Aplicacion:")
print(datos_aplicacion.info())
print("\nPrevias Aplicaciones:")
print(previas_aplicaciones.info())
```

1. Resumen de datos de aplicacion

   Número de Entradas: 307,511 registros con 122 columnas.
   Tipos de Datos: Contiene 65 columnas de tipo `float64`, 41 de tipo `int64`, y 16 de tipo `object`.
   Uso de Memoria: Aproximadamente 286.2 MB.

2. Resumen de previas aplicaciones

   Número de Entradas: 1,670,214 registros con 37 columnas.
   Tipos de Datos: 15 columnas de tipo `float64`, 6 de tipo `int64`, y 16 de tipo `object`.
   Uso de Memoria: Aproximadamente 471.5 MB.

- Primeras filas y columnas

Se examinan las primeras filas y las columnas de cada DataFrame para familiarizarse con los datos.

```
print("\nPrimeras filas de Datos de Aplicacion:")
print(datos_aplicacion.head())
print("\nPrimeras filas de Previas Aplicaciones:")
print(previas_aplicaciones.head())
```

1. Primeras filas datos de aplicacion

   - Incluyen información como `SK_ID_CURR`, `TARGET`, `NAME_CONTRACT_TYPE`, `CODE_GENDER`, `FLAG_OWN_CAR`, y `AMT_INCOME_TOTAL`.
   - Las variables numéricas incluyen montos como `AMT_CREDIT`, `AMT_ANNUITY`, y `AMT_GOODS_PRICE`.

2. Primeras filas de previas aplicaciones

   - Incluyen `SK_ID_PREV`, `SK_ID_CURR`, `NAME_CONTRACT_TYPE`, `AMT_APPLICATION`, y `WEEKDAY_APPR_PROCESS_START`.
   - Algunas columnas contienen valores faltantes (`NaN`), como `AMT_DOWN_PAYMENT` y `RATE_INTEREST_PRIMARY`.

Estos datos reflejan información sobre los préstamos y el estado del cliente.

- Columnas

```
print("Columnas en Datos de Aplicación:")
print(datos_aplicacion.columns)
print("\nColumnas en Datos de Aplicación Previa:")
print(previas_aplicaciones.columns)
```

Ambos conjuntos de datos tienen columnas con nombres bastante descriptivos, por lo tanto debe ayudar a identificar qué información está contenida en cada uno.

- Estadísticas descriptivas

```
print("\nEstadísticas descriptivas de Datos de Aplicacion:")
print(datos_aplicacion.describe())
print("\nEstadísticas descriptivas de Previas Aplicaciones:")
print(previas_aplicaciones.describe())
```

Las estadísticas descriptivas proporcionadas muestran un resumen estadístico de los datos en la tabla de "Datos de Aplicación"

- Valores faltantes

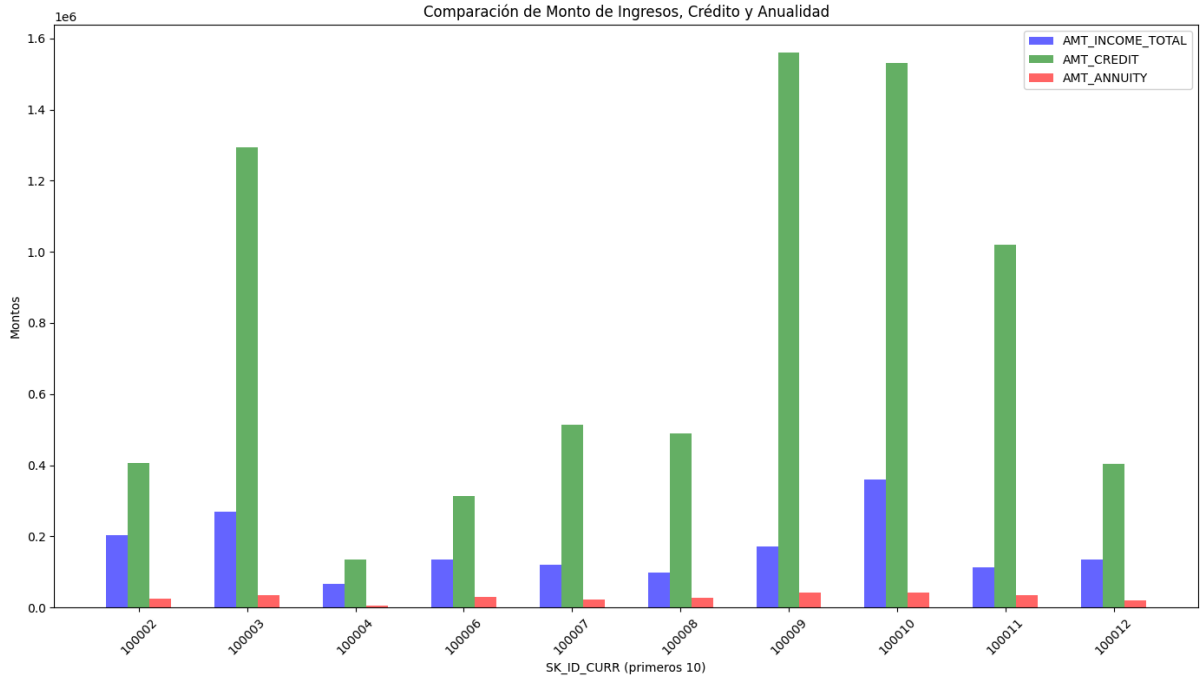
```
print("\nValores faltantes en Datos de Aplicacion:")
print(datos_aplicacion.isnull().sum())
print("\nValores faltantes en Previas Aplicaciones:")
print(previas_aplicaciones.isnull().sum())
```

En ambas tablas, algunas columnas tienen valores faltantes significativos, especialmente en las previas aplicaciones (AMT_DOWN_PAYMENT, RATE_INTEREST_PRIMARY, entre otras).

- Visualización

Para entender mejor las características numéricas de los datos, se realizó una visualización comparativa de los montos de ingreso total, crédito y anualidad de los clientes actuales. Esta visualización ayuda a identificar patrones y diferencias clave entre estos valores.

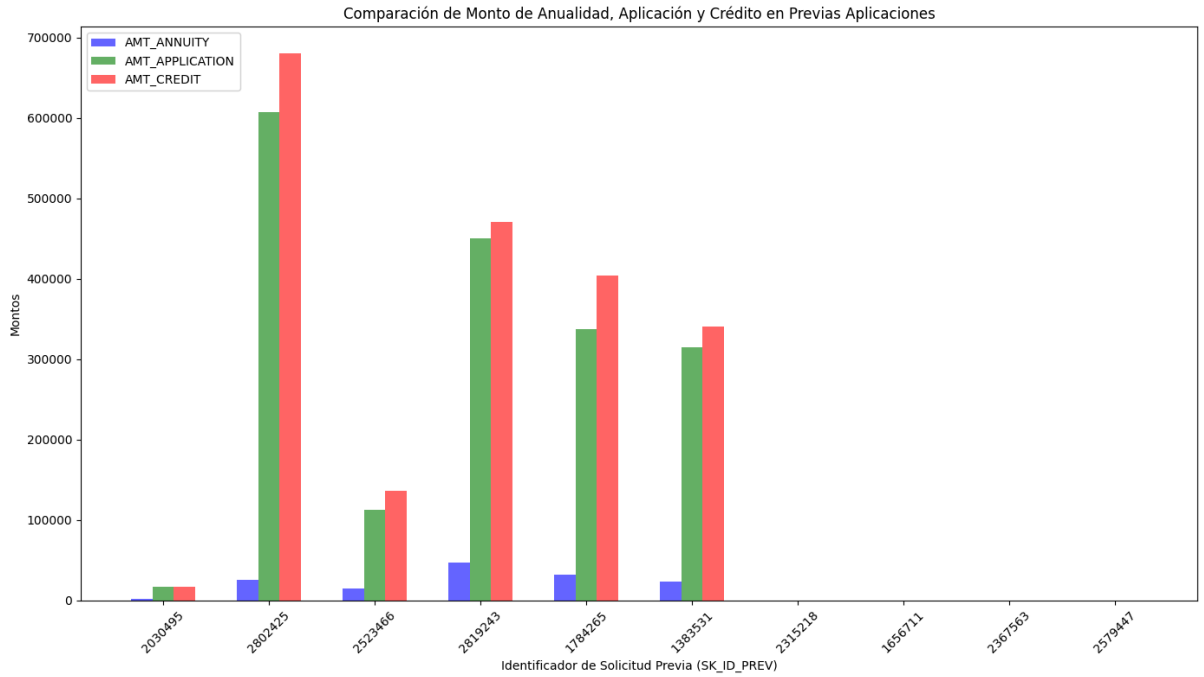
Datos de aplicacion:



- Descripción de la gráfica

    1. Se utiliza un gráfico de barras para visualizar las tres variables clave: `AMT_INCOME_TOTAL` (monto de ingresos totales), `AMT_CREDIT` (monto de crédito), y `AMT_ANNUITY` (monto de la anualidad).
    2. Esta comparación facilita la identificación de clientes con altos montos de crédito y cómo estos se relacionan con sus ingresos totales y anualidades, lo cual podría ser relevante para el análisis de riesgos y la probabilidad de incumplimiento de pagos.

Previas Aplicaciones:

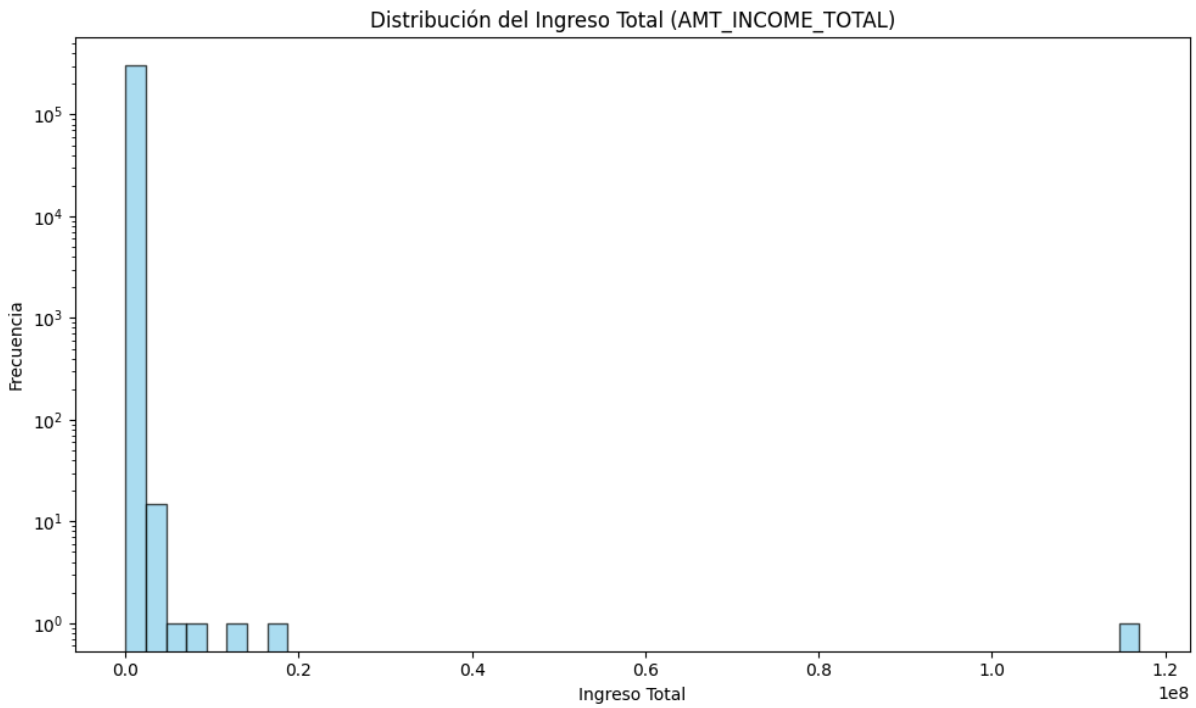


- Descripción de la gráfica

    1. Esta gráfica compara los montos de anualidad (`AMT_ANNUITY`), aplicación (`AMT_APPLICATION`), y crédito (`AMT_CREDIT`) para un subconjunto de previas aplicaciones (las primeras 10) identificadas por su `SK_ID_PREV`.
    2. La gráfica ayuda a identificar si existen diferencias significativas entre los montos solicitados y los concedidos, y cómo estas diferencias podrían estar asociadas a los riesgos de crédito en solicitudes previas.

- Distribución del Ingreso total

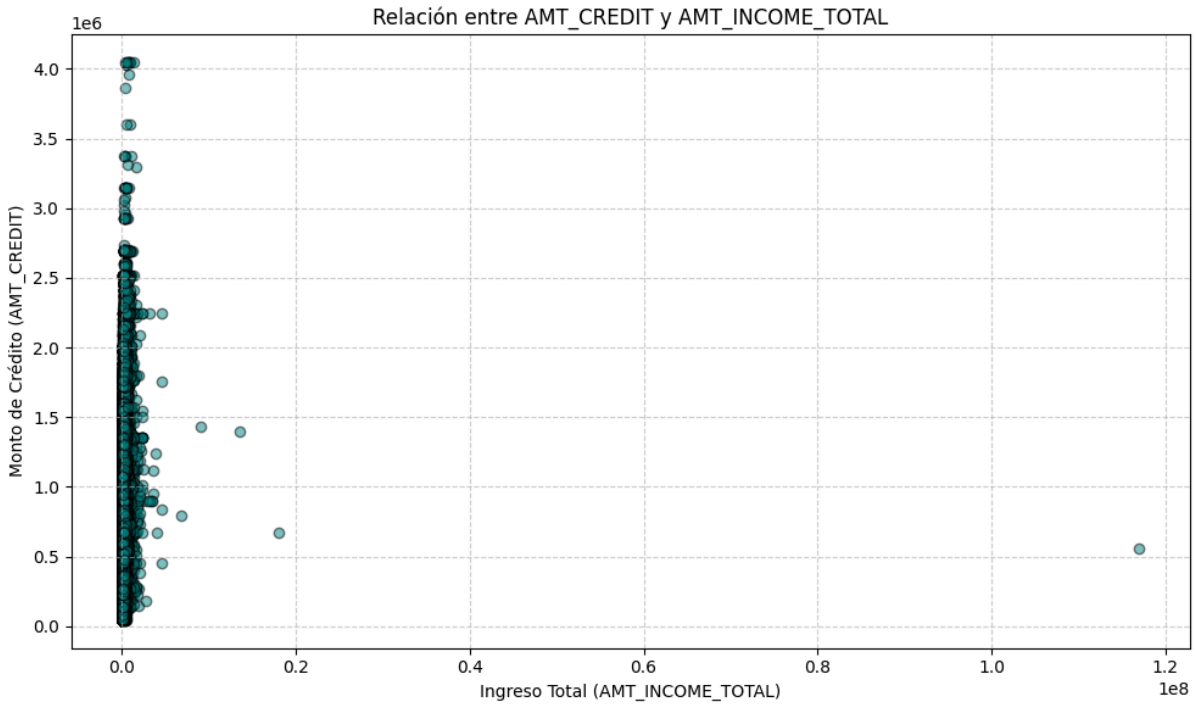
El siguiente histograma sirve para ver cómo se distribuyen los ingresos de los solicitantes.



La escala logarítmica en el “eje y” ayuda a visualizar mejor los datos si hay valores muy extremos, como en tu conjunto de datos donde el ingreso máximo es muy alto comparado con la mayoría de los otros valores.

- Relación entre AMT_CREDIT y AMT_INCOME_TOTAL

Un scatter plot para identificar posibles correlaciones entre el monto del crédito y los ingresos.

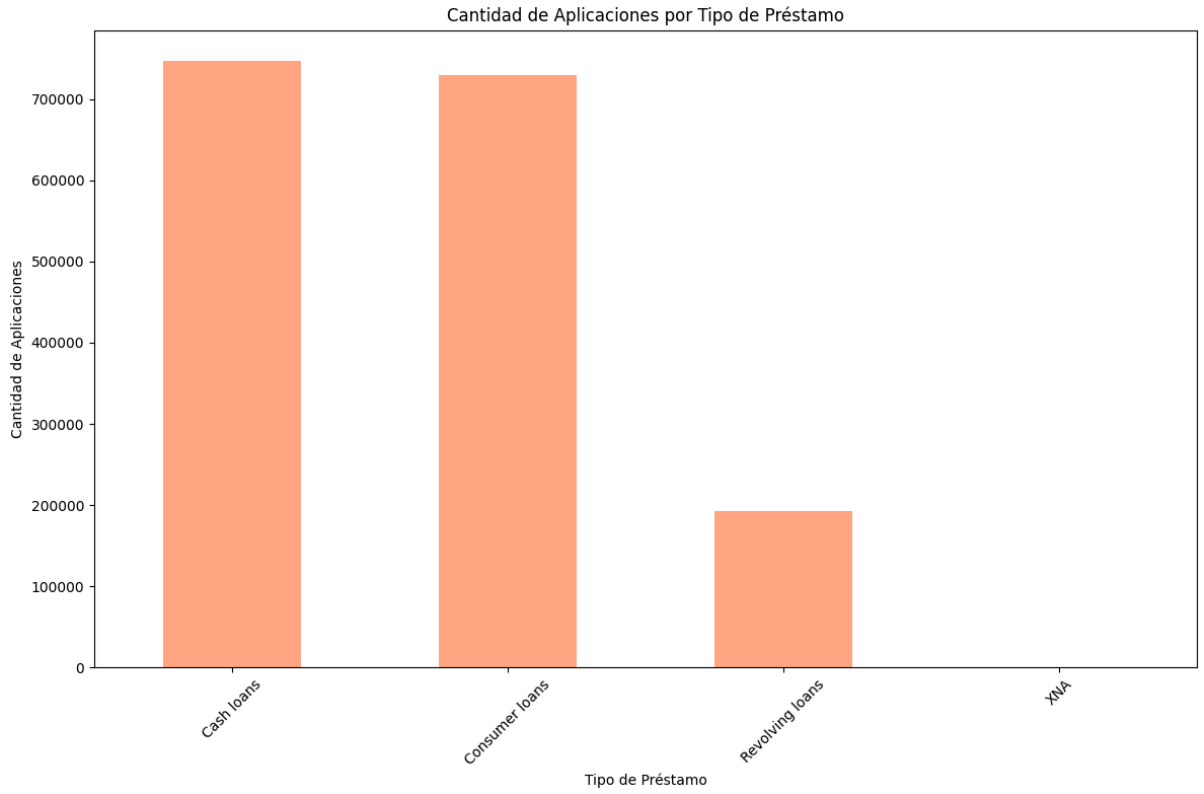


Nota: Cuando hay muchos puntos acumulados indica que hay una alta densidad de datos en el área específica del gráfico.

En el caso de AMT_CREDIT y AMT_INCOME_TOTAL, una tendencia ascendente podría indicar que a mayores ingresos, los montos de crédito solicitados también tienden a ser mayores.

- Cantidad de aplicaciones por tipo de préstamos

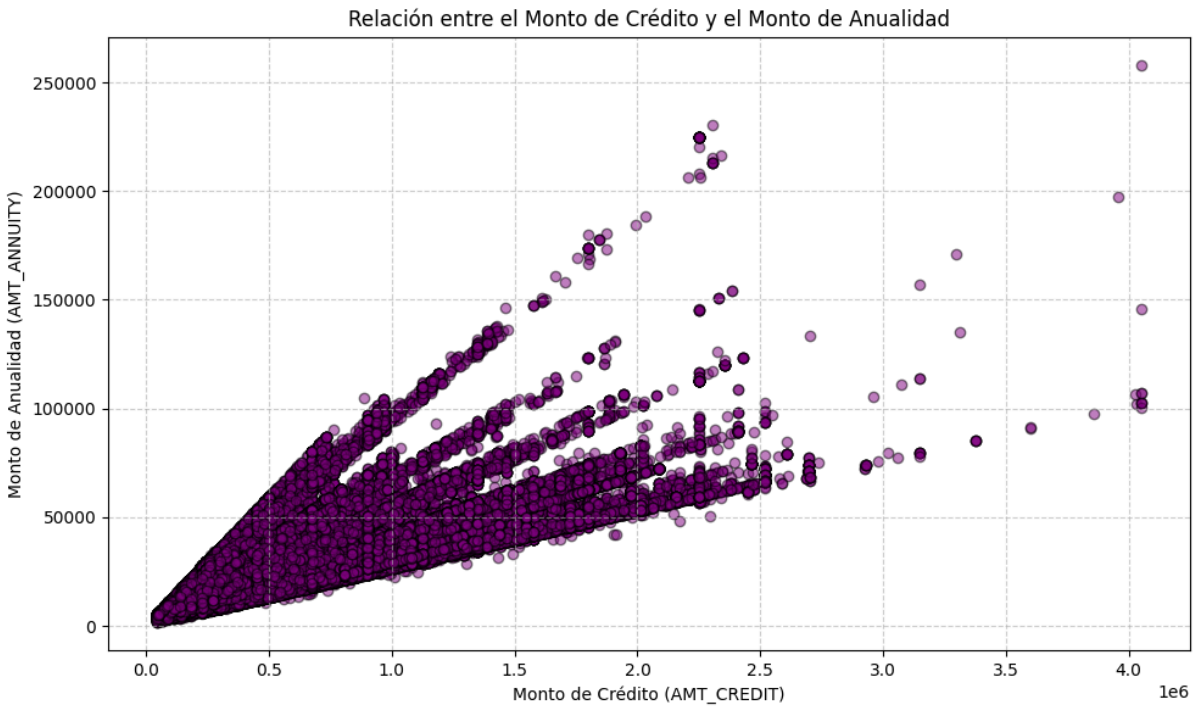
Se crea un gráfico de barras que muestra la cantidad de aplicaciones según el tipo de préstamo (NAME_CONTRACT_TYPE) en el conjunto de datos previas_aplicaciones.



Esto permite identificar qué tipos de préstamos son más comunes entre las aplicaciones previas.

- Relación entre el Monto de Crédito y el Monto de Anualidad

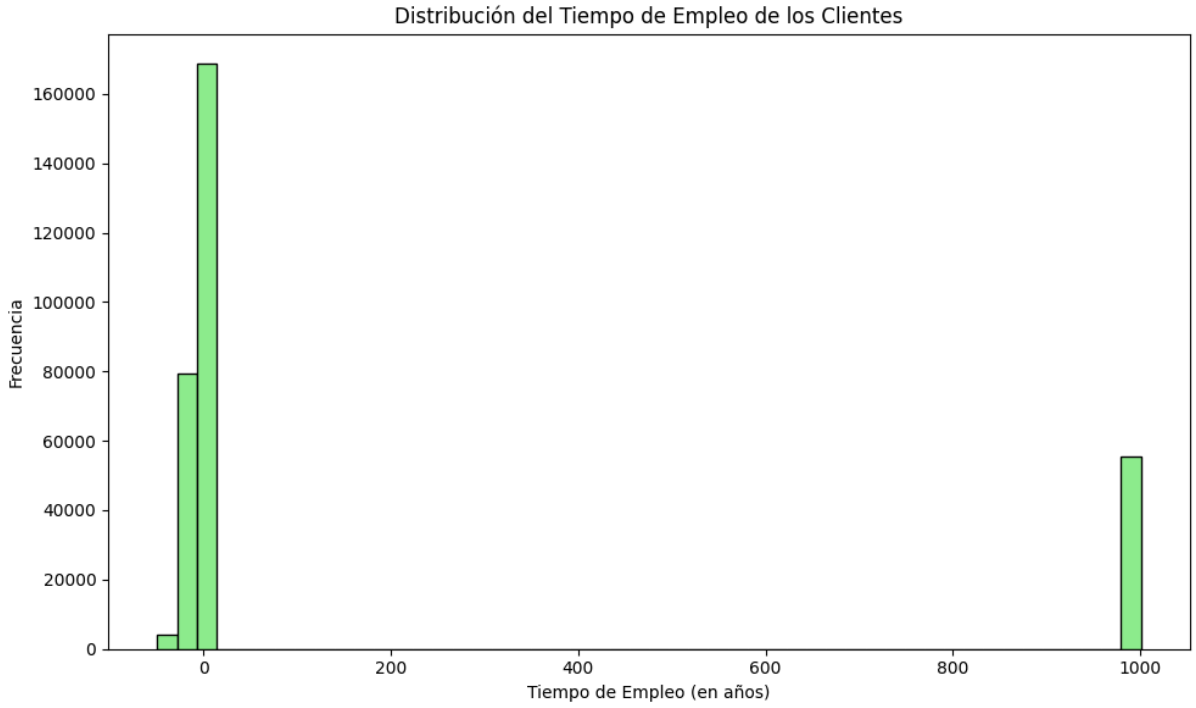
Se genera un gráfico de dispersión para explorar la relación entre el monto del crédito (AMT_CREDIT) y el monto de la anualidad (AMT_ANNUITY)



El objetivo es identificar si existe una correlación entre el crédito concedido y la anualidad correspondiente.

- Distribución del Tiempo de Empleo en años

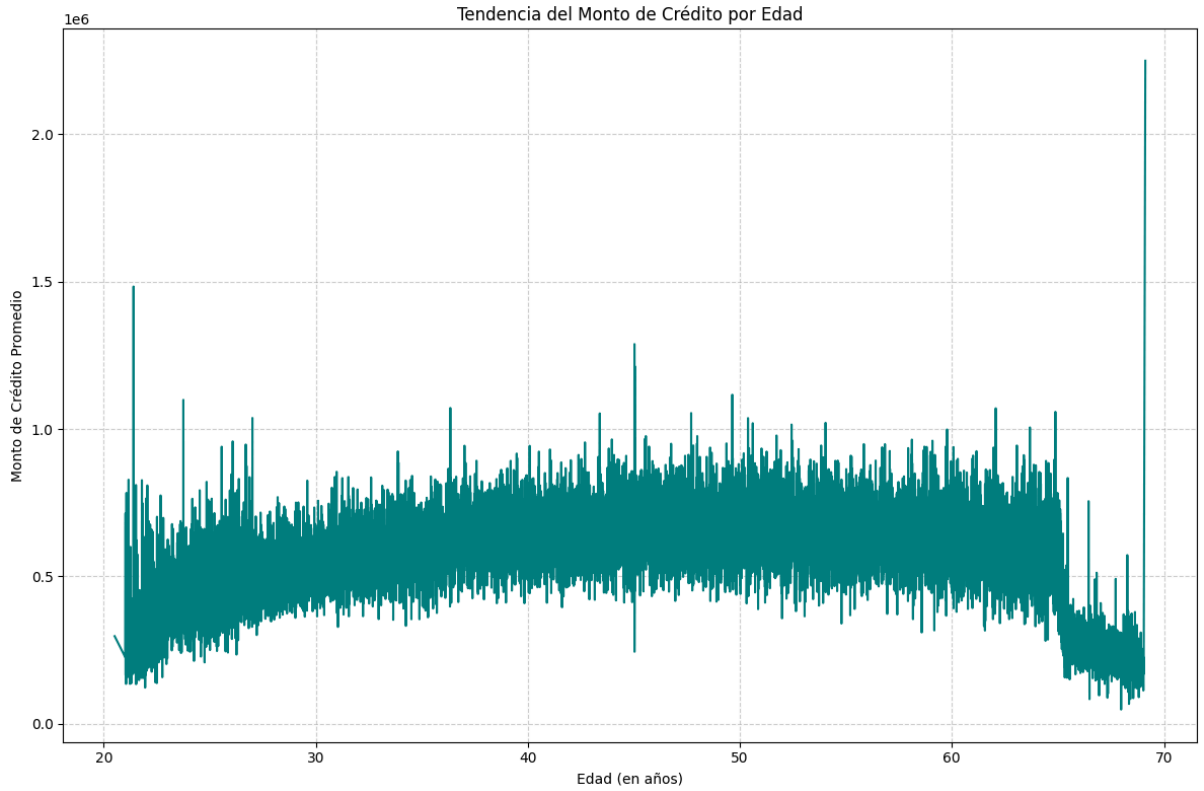
Aquí se muestra un histograma que representa la distribución del tiempo de empleo de los clientes en años, utilizando los datos de DAYS_EMPLOYED convertidos de días a años.



El histograma muestra la frecuencia de los diferentes tiempos de empleo, con las barras de color verde claro y bordes negros. Esto ayuda a visualizar la estabilidad laboral de los clientes en términos de duración del empleo.

- Tendencias de Aplicaciones a lo Largo del Tiempo

Se gráfica una línea para mostrar cómo varía el monto promedio de crédito en función de la edad de los solicitantes (DAYS_BIRTH convertido de días a años). Agrupa los datos por edad y calcula el monto promedio del crédito para cada grupo. Esto permite identificar patrones de crédito relacionados con la edad de los clientes.



- Identificar Clientes Potenciales para Préstamos

```
umbral_ingresos = 100000
umbral_credito = 500000

if 'AMT_INCOME_TOTAL' in df_combined.columns and 'AMT_CREDIT_x' in
df_combined.columns:
    clientes_potenciales = df_combined[
        (df_combined['AMT_INCOME_TOTAL'] > umbral_ingresos) &
        (df_combined['AMT_CREDIT_x'] < umbral_credito)
    ]
```

```
    clientes_no_potenciales = df_combined[
        (df_combined['AMT_INCOME_TOTAL'] <= umbral_ingresos) |
        (df_combined['AMT_CREDIT_x'] >= umbral_credito)
    ]

    print(f"\nClientes Potenciales: {len(clientes_potenciales)}")
    print(f"Clientes No Potenciales: {len(clientes_no_potenciales)}")

else:
    print("No se pueden identificar clientes potenciales porque las
columnas necesarias no están presentes.")
```

Se establecen criterios para identificar clientes potenciales para préstamos, basado en un umbral mínimo de ingresos (`umbral_ingresos = 100000`) y un umbral máximo de crédito (`umbral_credito = 500000`). Se verifica la presencia de las columnas necesarias antes de proceder con los filtros. Los clientes que cumplen con ambos criterios se consideran potenciales, mientras que los que no lo hacen se clasifican como no potenciales. Finalmente, se imprime la cantidad de clientes en cada categoría.

Conclusión

En este análisis exploratorio, se han examinado distintos aspectos de las solicitudes de crédito y aplicaciones previas de los clientes, utilizando técnicas de visualización para identificar patrones y relaciones clave en los datos. A través de gráficos de barras, dispersión, histogramas y líneas de tendencia, hemos obtenido una visión más clara de las características demográficas y financieras de los solicitantes, así como de sus comportamientos de aplicación.

Se identificaron diferencias significativas en la cantidad de aplicaciones por tipo de préstamo, la relación entre el monto de crédito y la anualidad, y la distribución del tiempo de empleo, lo cual podría ser indicativo de la estabilidad financiera de los clientes. Además, al analizar las tendencias del monto de crédito en función de la edad, se observó cómo estas variaciones pueden influir en las decisiones de otorgamiento de préstamos. La identificación de clientes potenciales para préstamos, basada en umbrales de ingresos y montos de crédito, permitió segmentar a los solicitantes en grupos más manejables para un análisis de riesgo más detallado.