**Logotipo

Descripción generada automáticamente**

**Desarrollo de modelo predictivo para determinar el costo total proyectado de un futuro beneficiario de Sapiencia.**

Genaro Alfonso Aristizabal Echeverri - Euler Leonardo Cuarán Rosero

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2025

|  |  |
| --- | --- |
| **Cita** | (Aristizabal Echeverri & Cuarán Rosero, 2025) |
| **Referencia**  **Estilo APA 7 (2020)** | Aristizabal Echeverri, G. A., & Cuarán Rosero, E. L. (2025). [*Desarrollo de modelo predictivo para determinar el costo total proyectado de un futuro beneficiario de sapiencia]*. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteIV.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Dedicatoria**

**Agradecimientos**

Agradecimientos a Dios, nuestros padres, profesores y al alma mater.

**Tabla de contenido**

[Resumen 7](#_heading=h.3rdcrjn)

[Abstract 8](#_heading=h.26in1rg)

[1. Descripción del problema 9](#_heading=h.35nkun2)

[1.1. Problema de negocio 9](#_heading=h.1ksv4uv)

[1.2. Aproximación desde la analítica de datos 10](#_heading=h.44sinio)

[1.3. Origen de los datos 11](#_heading=h.z337ya)

[1.4. Métricas de desempeño 12](#_heading=h.3j2qqm3)

[2. Objetivos 14](#_heading=h.4i7ojhp)

[2.1. Objetivo general 14](#_heading=h.2xcytpi)

[2.2. Objetivos específicos 14](#_heading=h.3whwml4)

[3. Datos 15](#_heading=h.2bn6wsx)

[3.1. Datos originales 15](#_heading=h.qsh70q)

[3.2. Datasets 16](#_heading=h.3as4poj)

[3.3. Analítica descriptiva 20](#_heading=h.1pxezwc)

[Referencias 26](#_heading=h.49x2ik5)

[Anexos 28](#_heading=h.43ky6rz)

[Anexo 1. Autoarchivo en Repositorio y documentos de interés 28](#_heading=h.xvir7l)

**Lista de tablas**

[**Tabla**](#_heading=h.2u6wntf) **3.1** Descripción de variables más representativas

**Tabla 3.2.1** Identificación de valores nulos

**Lista de figuras**

**Figura 3.3.1** Distribuciones de variables numéricas

**Figura 3.3.2.** Matriz de correlación de las variables numéricas

**Figura 3.3.3.** Gráficos de cajas de dispersión de las variables numéricas

**Figura 3.3.4.** Distribución de las variables Comuna de Residencia y Convocatoria.

**Figura 3.3.5.** Distribución de las variables Género y Estrato

**Figura 3.3.6.** Distribución de las variables IES y Programa

# Resumen

Este estudio tuvo como objetivo desarrollar un modelo predictivo basado en datos históricos financieros de los beneficiarios de Sapiencia, correspondientes a las convocatorias 2019-1 hasta 2024-2. El propósito principal fue estimar el costo total proyectado de futuros beneficiarios, utilizando esta información para optimizar la asignación de recursos disponibles y maximizar el impacto del programa. El conjunto de datos, compuesto por más de 468,000 registros con información socioeconómica, académica y transaccional, fue sometido a un proceso de limpieza, anonimización y tratamiento de valores faltantes o inconsistentes. Se implementaron métricas de desempeño como RMSE, R² y MAE, las cuales permitieron evaluar la precisión del modelo predictivo, logrando resultados satisfactorios con un error inferior al 5% y un coeficiente de determinación superior a 0.8. Adicionalmente, se analizó la distribución de variables demográficas y académicas, encontrando que la mayor parte de los beneficiarios pertenecen a estratos 2 y 3, con una notable representación femenina. Los resultados obtenidos contribuyen a una mejor gestión de los recursos del programa, posibilitando la proyección de apoyos financieros de manera más eficiente.

*Palabras clave: Modelo predictivo, Proyección financiera, Machine Learning, Sapiencia.*

# Abstract

This study aimed to develop a predictive model based on historical financial data of Sapiencia beneficiaries, corresponding to the 2019-1 to 2024-2 calls. The main one was to estimate the total projected cost of future beneficiaries, using this information to optimize the allocation of available resources and maximize the impact of the program. The data set, composed of more than 468,000 records with socioeconomic, academic and transactional information, was subjected to a process of cleaning, anonymization and treatment of missing or inconsistent values. Performance metrics such as RMSE, R² and MAE were implemented, which made it possible to evaluate the precision of the predictive model, achieving satisfactory results with an error of less than 5% and a coefficient of determination greater than 0.8. In addition, the distribution of demographic and academic variables was analyzed, finding that most of the beneficiaries belong to strata 2 and 3, with a notable female representation. The results obtained contribute to better management of program resources, making it possible to project financial support more efficiently.

*Keywords*: Predictive model, Financial projection, Machine Learning, Sapience.

# Descripción del problema

El acceso a la educación superior ha aumentado en todas las regiones del mundo durante las últimas dos décadas (UNESCO , 2020, p.24). Sin embargo, la deserción postsecundaria es uno de los principales desafíos a los que se enfrentan los estudiantes que aspiran a culminar sus estudios universitarios. Ya sea de manera voluntaria, por motivos familiares, económicos, académicos o personales. A nivel nacional, tanto en las Instituciones de Educación Superior (IES) públicas como privadas, la falta de recursos económicos es una de las causas más comunes. Por tal motivo, en Colombia han sido creadas entidades encargadas de, por medio de créditos, aumentar el acceso y permanencia a la educación superior de los Colombianos, como es el caso del ICETEX, una entidad del Estado que promueve la educación superior a través del otorgamiento de créditos educativos y su recaudo, con recursos propios o de terceros, a la población con menores posibilidades económicas y buen desempeño académico (ICETEX, n.d.)..

En Medellín, también se han evidenciado las problemáticas antes mencionadas, es por ello que desde el año 1996 existe, adscrita al distrito de Medellín, la entidad Sapiencia que es la Agencia de Educación Postsecundaria de la Alcaldía de Medellín, encargada de liderar los proyectos y programas de la educación postsecundaria del Distrito de Medellín (Medellín, 2023). El reto actual se centra en mejorar la gestión de los recursos económicos disponibles, optimizando la estimación de los créditos educativos para maximizar su impacto y garantizar que los estudiantes con menores recursos puedan continuar y culminar sus estudios sin que la falta de financiamiento sea una barrera.

## Problema de negocio

La Agencia de Educación Superior de Medellín - Sapiencia, brinda oportunidades para el acceso y permanencia en la educación superior a estudiantes de bajos recursos económicos, mediante créditos que se pueden condonar total o parcialmente si se cumplen con requisitos basados en méritos académicos y prestación de servicio social (Medellín, 2023). El proceso de inscripción se realiza de manera virtual; con una periodicidad semestral, en la que a cada aspirante se le asigna un puntaje según los criterios de selección establecidos por el reglamento del Fondo. Con el presupuesto disponible, se determina un punto de corte en cada convocatoria, a partir del cual se seleccionan los beneficiarios.

Sin embargo, en Sapiencia se ha identificado una problemática referente a la gestión y ejecución de los fondos destinados a financiar el acceso y permanencia de estudiantes en la educación postsecundaria. A pesar de los esfuerzos para proporcionar créditos condonables que faciliten la inclusión educativa, en ocasiones la cobertura no ha alcanzado a beneficiar a muchos estudiantes de las comunas y corregimientos de Medellín, que buscan adelantar o continuar sus estudios postsecundarios. Lo cuál ha generado malestar en los representantes de dichas comunas y corregimientos, dado que, en ocasiones dichos recursos no se llegado a ejecutar en su totalidad, ya sea porque el aspirante no cumple con los requisitos necesarios para optar por el crédito o porque a la hora de surtir el proceso de legalización del crédito el recurso ya ha sido consumido por otros aspirantes del fondo que logran legalizar su crédito primero. (Agregar fuentes)

Actualmente, la problemática radica en la imprecisión en las estimaciones de las proyecciones financieras utilizadas para destinar el recurso en cada convocatoria para las distintas becas y créditos condonables bajo los dos fondos de financiación con los que se cuentan actualmente como lo son Presupuesto Participativo (PP) y Recurso Ordinario (RO). La estimación inadecuada de los montos necesarios para cubrir a los aspirantes adecuados lleva a una distribución ineficiente de los recursos, limitando la efectividad de las políticas de inclusión educativa y provocando una mala ejecución del presupuesto.

## Aproximación desde la analítica de datos

En la actualidad, se cuenta con estudios que han explorado los factores que influyen en los valores de los créditos utilizando herramientas estadísticas como la econometría paramétrica. Generalmente, se han empleado modelos logit, probit y tobit para realizar las estimaciones correspondientes (Greene, 1989; Barone, 2006; Belfield, 2013). Sin embargo, son escasos los estudios que recurren a técnicas de *Machine Learning* para prever el monto total asignado en el ámbito de los créditos educativos, aunque algunas investigaciones han analizado este enfoque en sectores como hipotecas de vivienda y créditos de consumo ( Zhou y Wang, 2012).

Incorporar modelos predictivos basados en *machine learning* (ML) en la gestión de créditos educativos podría mejorar significativamente la precisión de las proyecciones financieras y la asignación de recursos. Estos modelos, a diferencia de los enfoques tradicionales como la econometría paramétrica, permiten manejar grandes volúmenes de datos y encontrar relaciones complejas entre variables que podrían no ser evidentes a simple vista. Por ejemplo, en el ámbito de los créditos de consumo, el uso de algoritmos como los árboles de decisión o redes neuronales ha permitido predecir con mayor exactitud el riesgo de incumplimiento, utilizando factores como el historial crediticio, la relación deuda-ingresos y el comportamiento de pago (Noriega et al., 2023).

En el caso específico de los créditos educativos, un modelo de ML podría predecir no solo quién es más probable que solicite o reciba un crédito, sino también cuánto se debe asignar a cada beneficiario para maximizar el impacto de los fondos disponibles. Usando datos históricos de beneficiarios anteriores, el modelo podría identificar patrones y hacer proyecciones más ajustadas a la realidad económica actual, considerando factores como la situación financiera, el rendimiento académico y las tendencias de endeudamiento (Noriega et al., 2023).

Es por ello que, para abordar los desafíos antes planteados y mejorar la eficacia en la gestión de los fondos, es crucial desarrollar una metodología avanzada y confiable. Un modelo de machine learning, basado en datos históricos de beneficiarios, podría proporcionar estimaciones más precisas sobre las proyecciones financieras realizadas una vez el beneficiario legaliza su crédito. Además, identificará la cantidad de beneficiarios que se pueden impactar con el recurso que se tiene disponible

## Origen de los datos

Se tienen un conjunto de datos histórico de los beneficiarios de Sapiencia desde el año 2019 hasta el 2024 extraídos de las bases de datos de la agencia con su respectivo tratamiento siguiendo los requerimientos establecidos en el *habeas data*, anonimizando y omitiendo los datos más sensibles. Los registros cuentan con la información socioeconómica, académica y transaccional concernientes a los giros desembolsados a lo largo de su ciclo académico.

## Métricas de desempeño

Para los modelos de predicción realizados que predicen el costo total proyectado de un futuro beneficiario se obtuvieron las siguientes métricas de desempeño:

**1.4.1. Métricas principales**

**Error cuadrático Medio (RMSE)**

Medida que permitió evaluar la precisión de las proyecciones financieras en términos monetarios, ya que dicha penaliza los error de estimación, donde se obtuvieron errores inferiores al 5%.

**Coeficiente de determinación (R^2)**

El coeficiente de determinación muestra la variabilidad en la predicción que es explicada por los modelos, donde entre más se acerque el dicho valor 1 más acertadas resultan ser las predicciones. En el caso de estudio, se evidenció como los modelos mostraron tener un coeficiente mayor a 0.8.

**Error Absoluto Medio (MAE)**

Esta métrica nos proporciona una medida más directa respecto al error de predicción en el costo total proyectado donde se obtuvieron errores menores al 5%.

**1.4.2 Métricas de negocio**

**Tasa de incremento anual del salario mínimo**

Esta métrica fue utilizada en el proceso de cálculo de proyección financiera desarrollado semestre a semestre hasta culminar el total de giros pactados. Esta tasa fue del 7% anual constante para todos los modelos de predicción utilizados.

**Índice de Precios al Consumidor - IPC**

Se utilizó el IPC como indicador de variación anual como medida de referencia respecto a variables exógenas como el aumento del salario mínimo y el incremento de la inflación. En nuestro caso de estudio se realizó con un índice ubicado en el 12% cómo inicial para el año de legalización de los beneficiarios del año 2024.

**1.4.3. Valores mínimos aceptados**

Para cada uno de los modelos predictivos realizados, se establecieron los siguientes umbrales técnicos:

* R² ≥ 0.80
* RMSE ≤ 5%
* MAPE ≤ 5%

Las convenciones antes planteadas, permitieron dar una certeza a la entidad respecto a la fiabilidad de los modelos utilizados, y con ello, tomar decisiones financieras respecto a la optimización de recursos con el objetivo de abarcar todos los topes presupuestales establecidos en cada respectiva convocatoria.

# Objetivos

## Objetivo general

Desarrollar un modelo predictivo que utiliza información financiera histórica de los beneficiarios de Sapiencia desde la convocatoria 2019-1 hasta la 2024-2, con el fin de predecir el costo total proyectado de un futuro beneficiario. A partir de ello, identificar la cantidad de beneficiarios que se pueden impactar con el recurso que se tiene disponible, para que la entidad, fuera del alcance del proyecto, optimice la gestión del recurso económico.

## Objetivos específicos

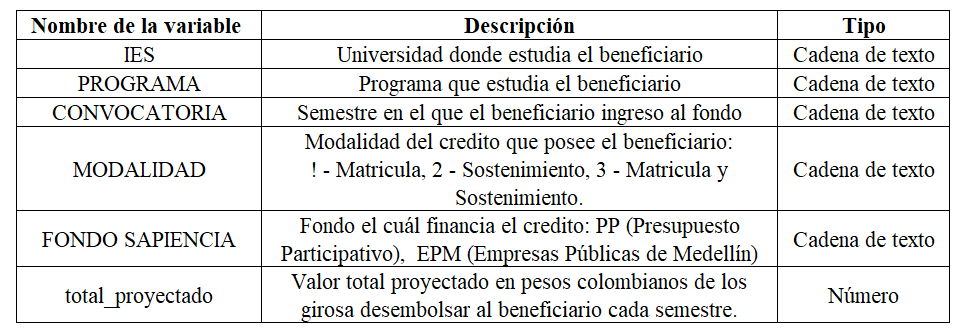
* Consolidar información financiera de los beneficiarios de Sapiencia desde la convocatoria 2019-1 hasta la convocatoria 2024-2.
* Preparar el conjunto de datos de información financiera con las variables más relevantes para la estimación del costo total proyectado de futuros beneficiarios.
* Desarrollar un modelo de predicción financiera que permita estimar el costo total proyectado de futuros beneficiarios.
* Identificar la cantidad de beneficiarios que se pueden impactar con el recurso que se tiene disponible.

# Datos

## Datos originales

Describe los datos crudos de los que se dispone, incluyendo formatos, cómo se distribuyen los datos en ficheros, descripción de las columnas, formatos y resoluciones de imágenes, etiquetado de los mismos (para aprendizaje supervisado), tamaño de los datos (número de registros, tamaño total en MB o GB), etc.

Los datos originales se extrajeron de las bases de datos SQL Server de los servidores de Sapiencia, a través de consultas a las distintas tablas referentes a proyecciones financieras, formularios de inscripción, montos y estados históricos de renovación de los beneficiarios. La base de datos resultante de dichas consultas se exportó en formato Parquet, un tipo de almacenamiento columnar, con un peso aproximado de 8 MB que contiene 468626 filas y 65 columnas..

Tabla 3.1. Descripción de variables más representativas

## Datasets

A partir de los datos originales se identificaron variables sensibles, las cuales, por cuestiones de privacidad y propósito del estudio, no fueron utilizadas. Por lo cual se eliminaron 19 columnas y se codificó la variable “id\_usuario” que hace referencia al identificador único del usuario en la agencia, todo ello con el objetivo de tener un *dataset* anonimizado. El *dataset* resultante se denominó **Data\_Sapiencia**  con un peso aproximado de 7 MB que contiene 468626 filas y 46 columnas..

* + 1. **Identificación de valores duplicados**

Se realizó un análisis de valores duplicados sobre los datos originales donde no se encontraron registros duplicados.

* + 1. **Identificación de valores nulos y depuración del conjunto de datos**

Se realizó un análisis de valores nulos sobre las variables del dataset, donde se encontraron las siguientes variables con valores faltantes:

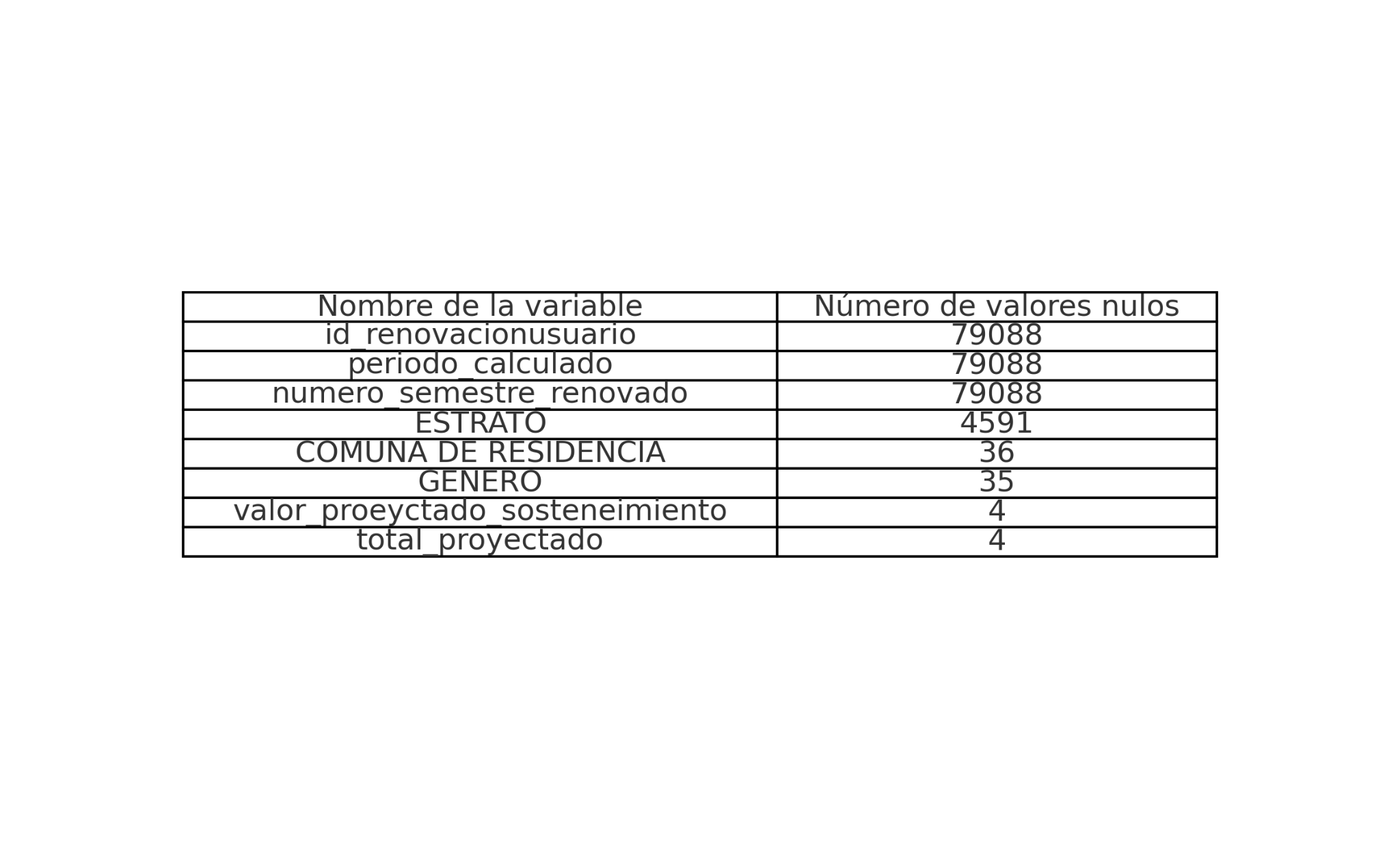
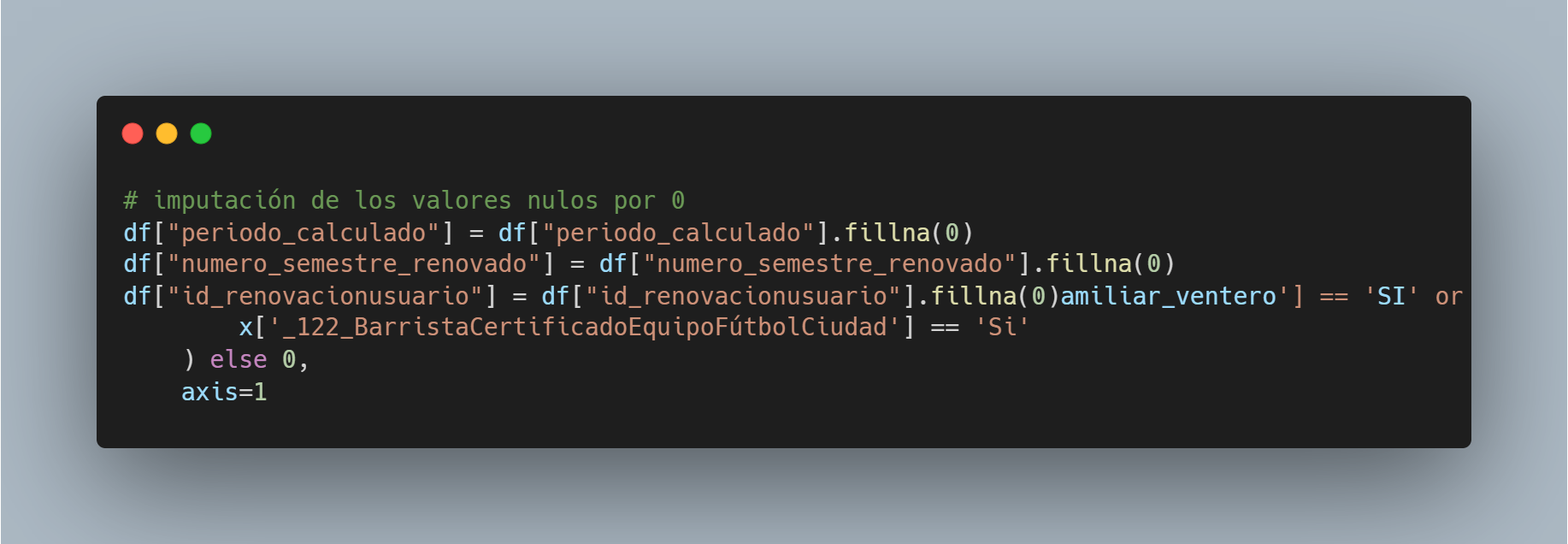
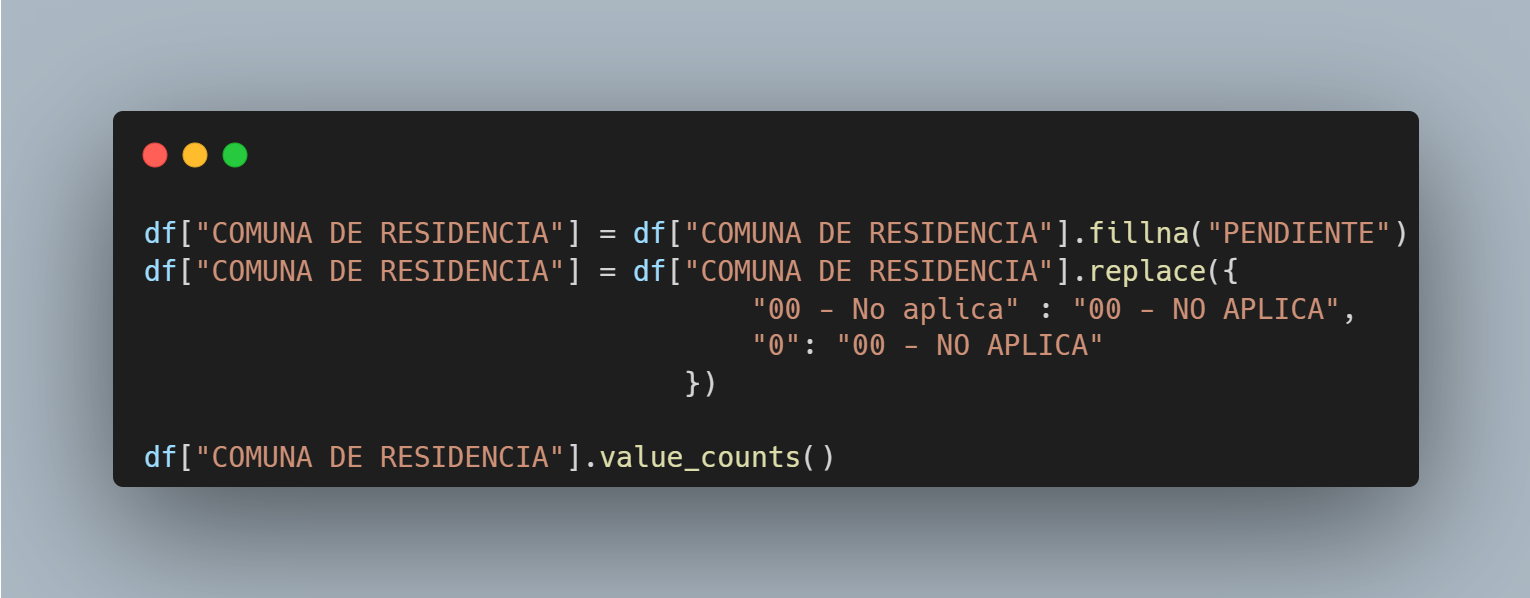


Tabla 3.2.1 Identificación de valores nulos

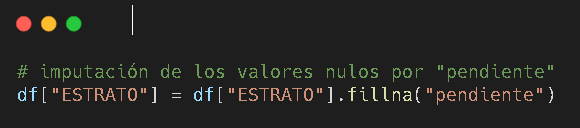
A continuación se realiza un análisis detallado de cada variable, esto con el fin de identificar el procesamiento adecuado para cada una de ellas::

* Las variables periodo\_calculado, numero\_semestre\_renovado e id\_renovacionusuario presentaban 79,088 valores nulos. Estas variables están relacionadas con la ejecución de los giros pactados en cada proceso de renovación. La presencia de valores nulos se debe a que dichas columnas se completan progresivamente a medida que cada beneficiario realiza su proceso de renovación semestral. Como muchos beneficiarios aún no han finalizado sus giros pactados, no cuentan con registros en estas variables, lo que genera los valores nulos. Por esta razón, los datos faltantes serán imputados con el valor "0".
* La variable "COMUNA DE RESIDENCIA" presenta un total de 36 valores nulos. Esto se debe a que, al diligenciar el formulario de inscripción para los créditos condonables, los beneficiarios del fondo EPM no están obligados a especificar la comuna de residencia. Como resultado, se registran estos valores nulos. Por tal motivo, los valores nulos serán incluidos en la categoría "00 - No Aplica".

Adicionalmente, se identificaron registros con los valores "0" y "00 - No Aplica", los cuales pueden haberse originado debido al proceso de unificación y migración de la información a la nube implementado a partir del año 2019. Este cambio en la estructura y el formato de los datos resultó en la aparición de ambas categorías.Por tal motivo, los registros etiquetados con "0" fueron recodificados y asignados a la categoría "00 - No Aplica" para asegurar la coherencia de los datos.



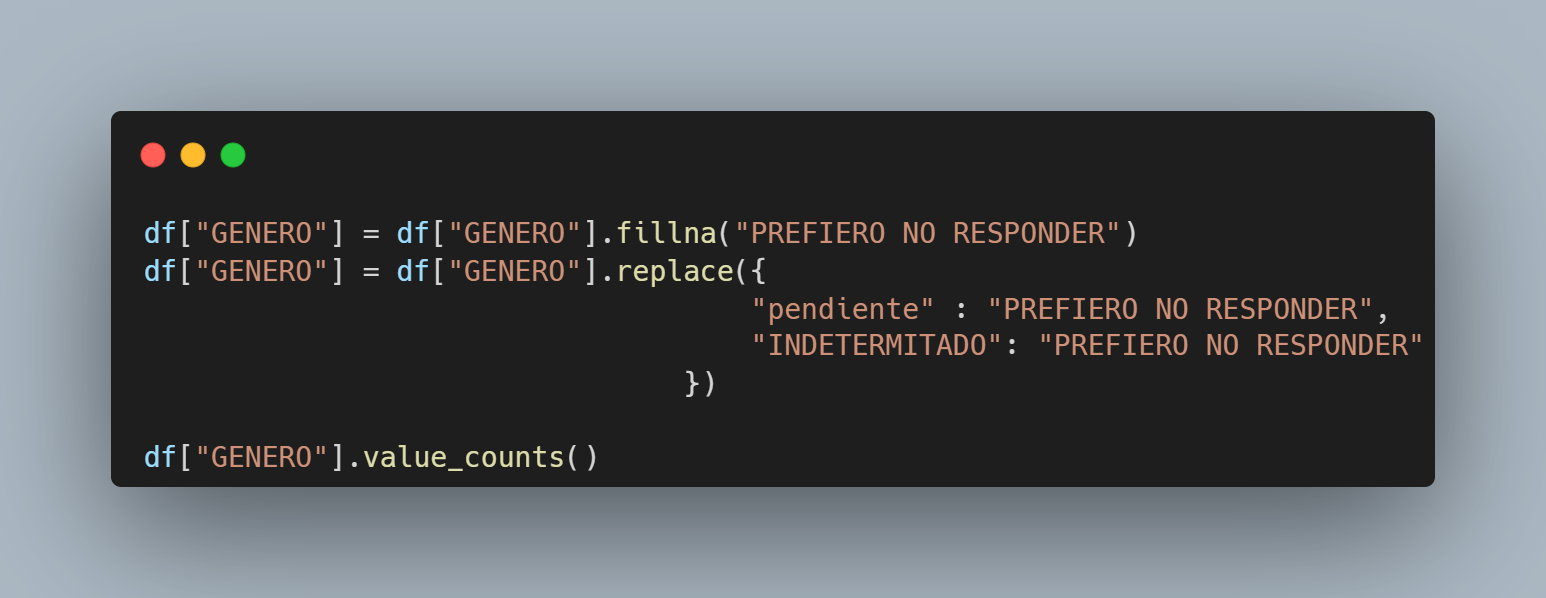
* La variable "ESTRATO" presenta un total de 4,591 valores nulos. Estos valores pueden corresponder a beneficiarios de convocatorias anteriores a 2019, cuando gran parte de la información se encontraba en archivos físicos. Por esta razón, los valores nulos serán imputados con la categoría "pendiente".



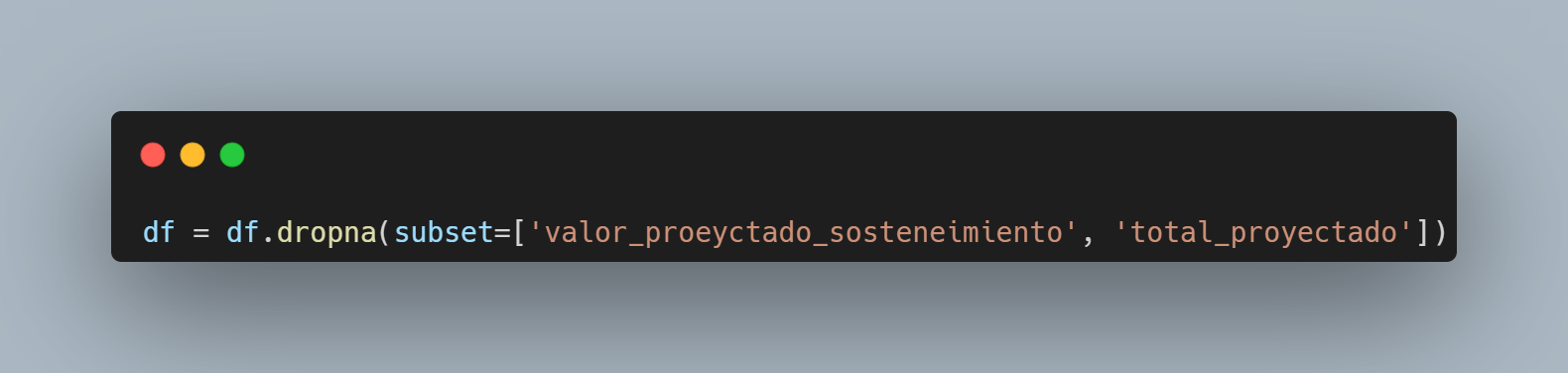
Adicionalmente, se realizó un proceso de homologación de categorías, debido a que algunas se encontraban en formato numérico y otras en formato de texto (strings). En consecuencia, todas las categorías fueron convertidas al formato de texto (strings) para asegurar la coherencia y consistencia en el manejo de los datos.



* La variable "GENERO" presenta 36 valores nulos, los cuales pueden atribuirse a que la pregunta no es obligatoria en los formularios de inscripción, lo que permitió que algunos beneficiarios omitieran la información. Además, durante el proceso de unificación y migración de datos a la nube en 2019, algunos registros antiguos podrían no haber incluido esta variable o contener información incompleta. Por lo tanto, los valores nulos serán asignados a la categoría "Prefiero no responder", garantizando la coherencia y validez de los datos. Adicionalmente, se realizó la homologación de las categorías "pendiente" e "INDETERMINADO", dado que ambas hacen referencia a la misma información.

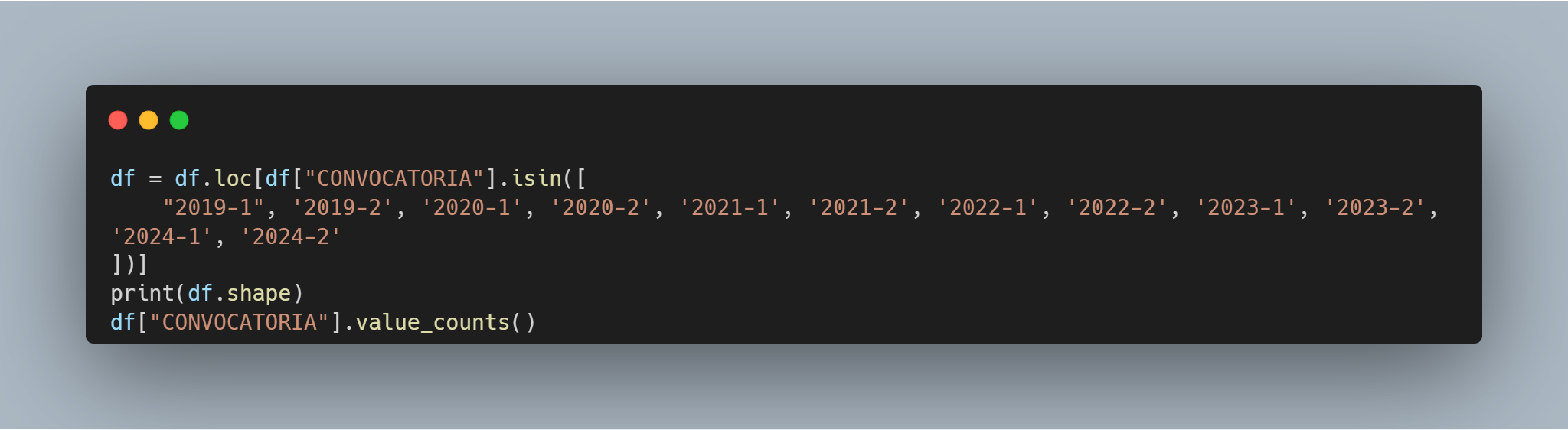


* Las variables “valor\_proyectado\_sostenimiento” y “total\_proyectado” presentan cuatro valores nulos cada una, lo que genera una inconsistencia en los datos. Esto se debe a que, una vez que un beneficiario legaliza su crédito, se debe calcular una proyección financiera que contemple todos los períodos pactados, lo que hace que la ausencia de valores en estas proyecciones no sea coherente. Esta situación probablemente ocurrió porque algunos beneficiarios iniciaron el proceso de legalización de su crédito, pero dicho proceso fue cancelado antes de completarse. En consecuencia, se procedió a eliminar estos valores nulos para asegurar la consistencia del conjunto de datos.

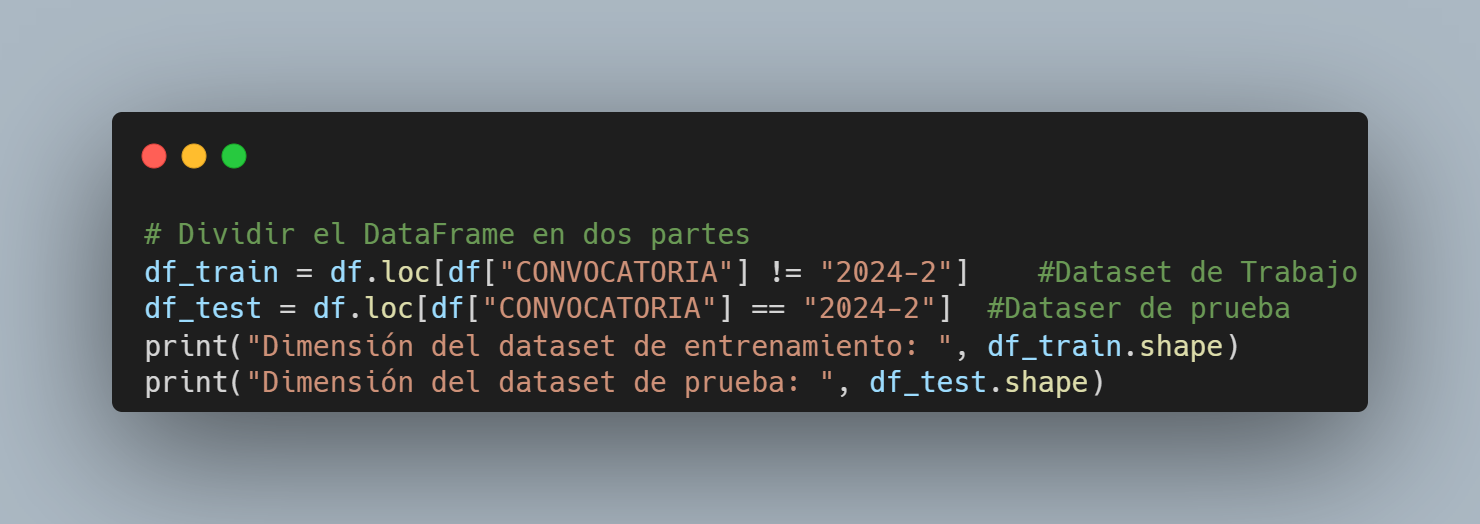


* + 1. **Construcción de datasets de entrenamiento y prueba**

Se observó una considerable cantidad de valores pendientes, los cuales pueden atribuirse a la dependencia histórica de archivos físicos utilizada por Sapiencia para el almacenamiento de la información. Este enfoque experimentó un cambio significativo a partir de la convocatoria 2019-1, cuando se decidió migrar toda la información histórica de los beneficiarios a la nube. Como resultado, la información más precisa y consistente respecto a la historia y proyección financiera de los beneficiarios está disponible a partir de la convocatoria 2019 en adelante.



A partir de los datos originales, se construyeron conjuntos específicos para el modelo de machine learning, abarcando el periodo comprendido entre la convocatoria 2019-1 y la convocatoria 2024-2. El dataframe se dividió en dos datasets: el primero, utilizado para el entrenamiento, incluyó los registros de las convocatorias 2019-1 a 2024-1, mientras que el segundo, destinado a la evaluación del modelo, correspondió a los registros de la convocatoria 2024-2. Esta división permitió entrenar el modelo con datos representativos y evaluarlo con información más reciente.



## Analítica descriptiva

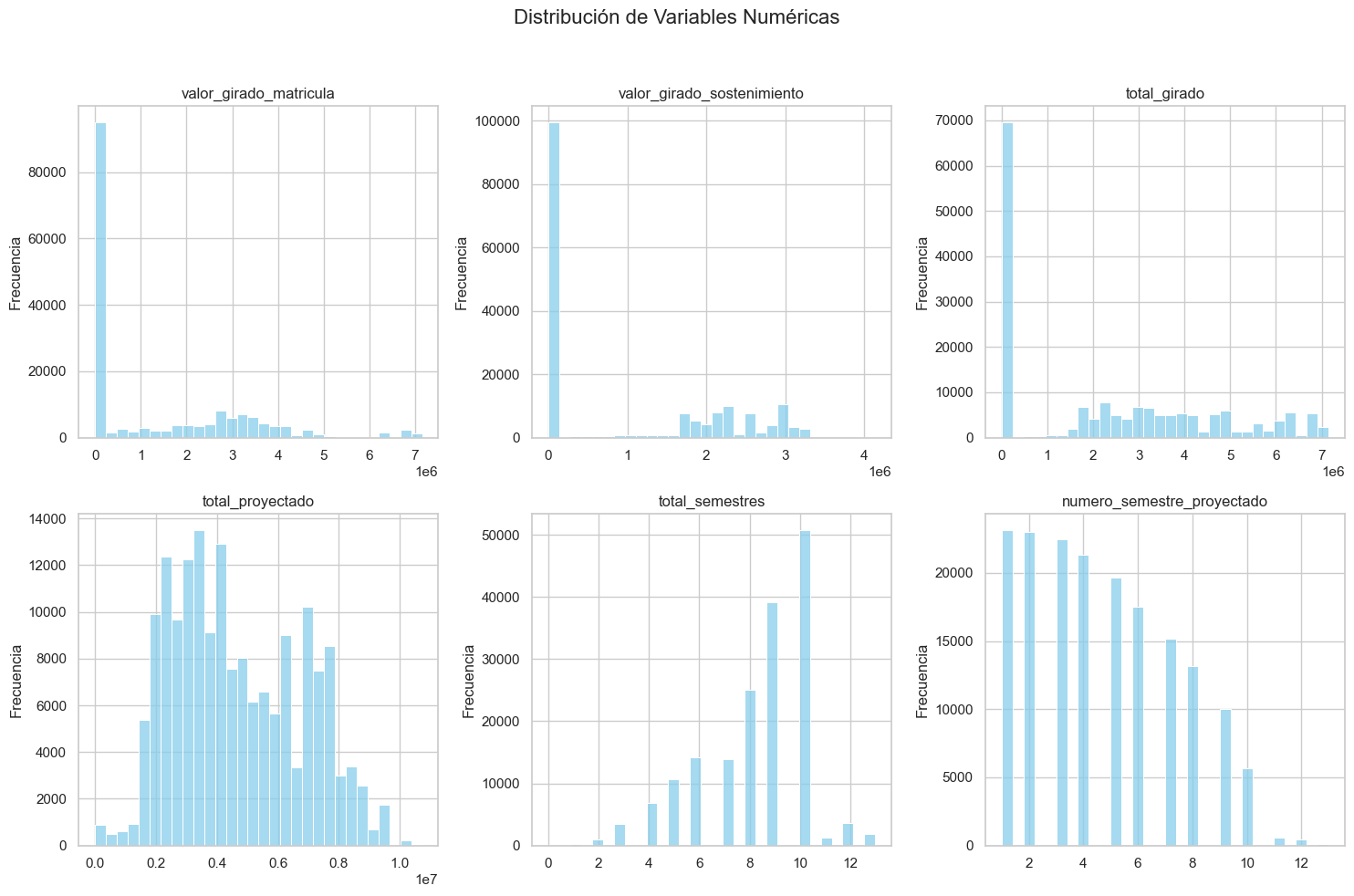
* A continuación, se presentan las distribuciones de las variables numéricas más relevantes (ver Figura 3.1.1). 

Figura 3.3.1 Distribuciones de variables numéricas

En la figura X, se puede observar que las distribuciones muestran que la mayoría de los beneficiarios reciben montos bajos en variables como **valor\_girado\_matricula**, **valor\_girado\_sostenimiento** y **total\_girado**, aunque existen algunos casos excepcionales con valores significativamente altos. Por otro lado, variables como **total\_semestres** y **numero\_semestre\_proyectado** tienen un pico alrededor de 10 a 12 semestres, lo que podría deberse a programas académicos de mayor duración, como medicina, que requieren más tiempo para completarse.

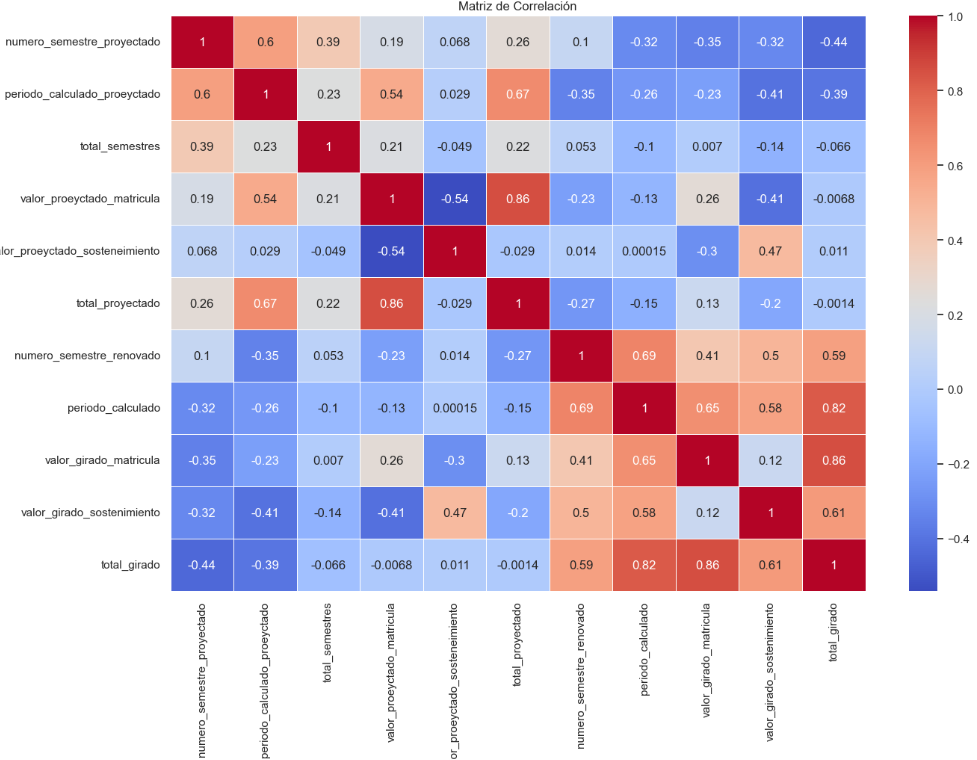
* Con el fin de identificar las relaciones entre las variables numéricas, se elaboró una matriz de correlación que muestra la intensidad y dirección de sus asociaciones. A continuación, se presenta la matriz de correlación de dichas variables. ( Ver Figura 3.3.2).

Figura 3.3.2. Matriz de correlación de las variables numéricas

En la Figura 3.3.2, se observa una fuerte correlación positiva entre el valor proyectado de matrícula y el total proyectado (0.86) lo que indica que el monto destinado a matrícula es un componente importante dentro del total proyectado. También se observa que el período calculado proyectado tiene una correlación positiva considerable con el total proyectado (0.67), lo que sugiere que los períodos influyen en las estimaciones. Además, se pueden apreciar variables como el valor proyectado de sostenimiento y el total de semestres, que presentan correlaciones más débiles o incluso negativas, evidenciando relaciones menos significativas.

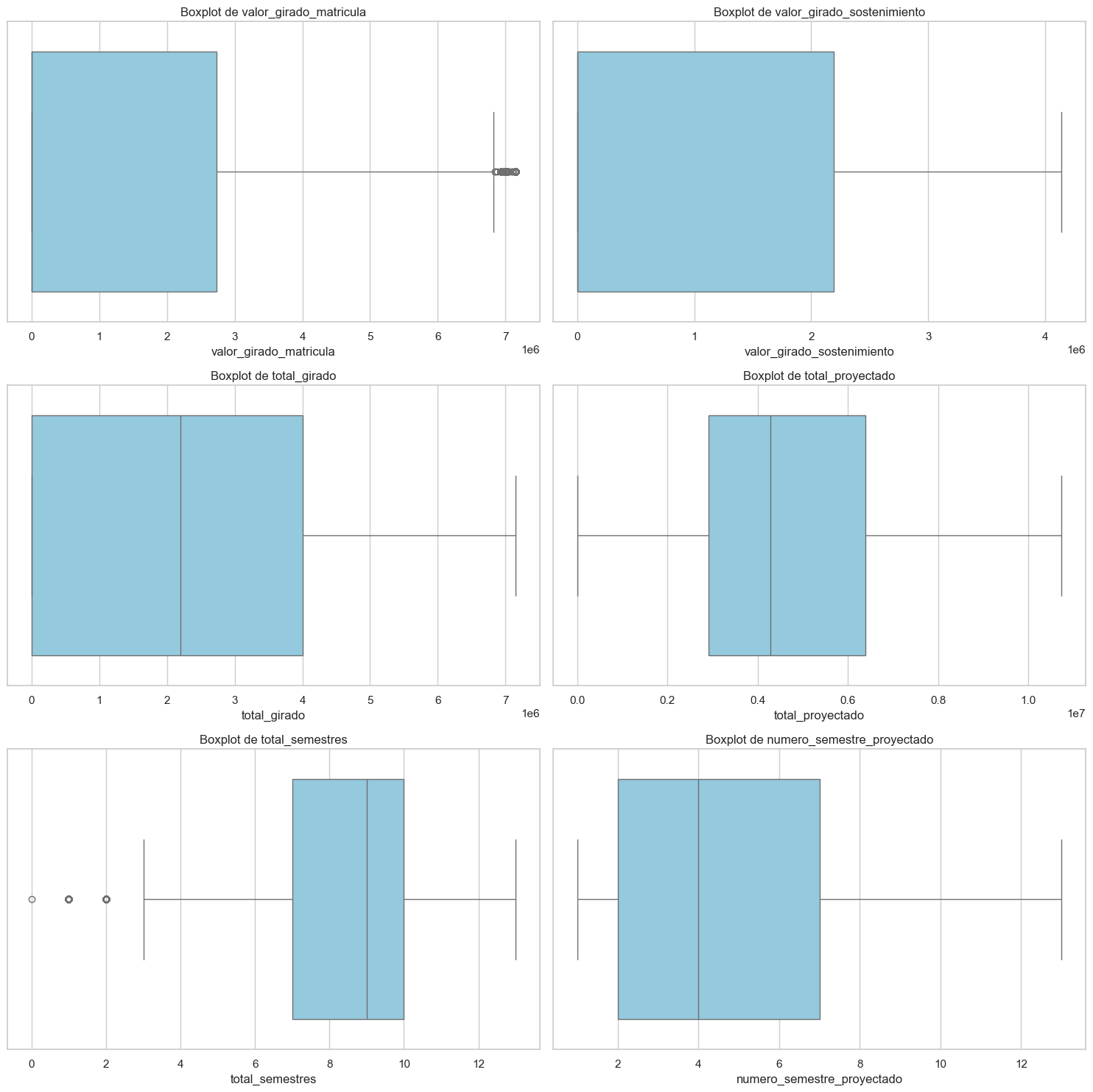
* Para analizar la distribución de las variables numéricas, se utilizaron gráficos de caja y bigotes, los cuales permiten identificar la dispersión, el rango intercuartil y posibles valores atípicos. Este enfoque facilita la detección de patrones relevantes y casos excepcionales que podrían influir en los resultados. A continuación, se presentan los boxplots de las principales variables (ver Figura 3.3.3).

Figura 3.3.3. Gráficos de cajas de dispersión de las variables numéricas

Las gráficas de caja muestran una alta dispersión y presencia de valores atípicos en varias variables. Por ejemplo, en valor girado matrícula, valor girado sostenimiento y total girado, se observan puntos extremos que indican montos significativamente más altos para ciertos beneficiarios. Adicionalmente, las variables total\_semestres y valor\_girado\_matricula presentan outliers. Estos valores atípicos deben analizarse para determinar si son casos legítimos o anomalías, ya que pueden influir en los modelos predictivos.

* A continuación, se presentan las distribuciones de las variables categóricas más representativas:

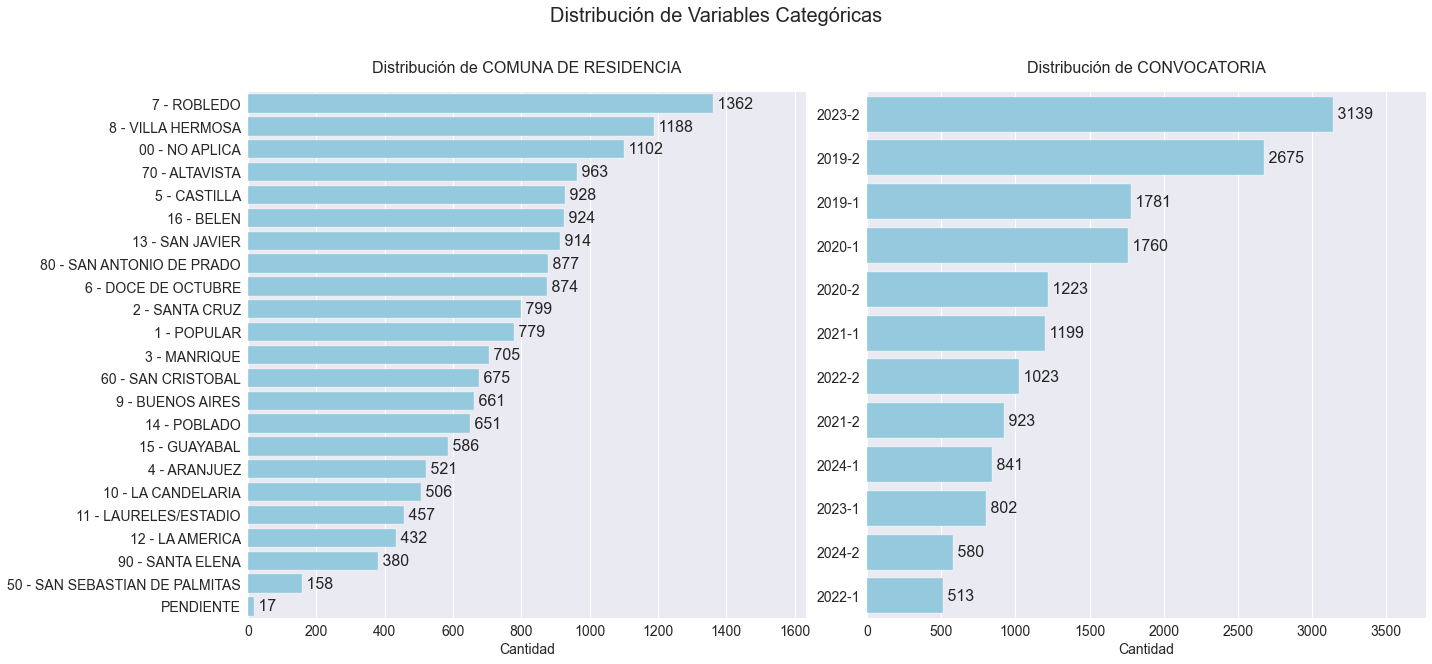


Figura 3.3.4. Distribución de las variables Comuna de Residencia y Convocatoria.

La variable **COMUNA DE RESIDENCIA** representa la comuna en la cual reside el beneficiario. Entre las comunas con más beneficiarios se encuentran 7 - ROBLEDO con 1362, 8- VILLA HERMOSA con 1188 y 70 - ALTAVISTA con 963 beneficiarios legalizados entre los años de 2019 a 2024.

Por otro lado, La variable **CONVOCATORIA**, corresponde con el semestre en el que el beneficiario legalizó su crédito. En este caso se observa como, actualmente, la convocatoria en la que se legalizaron más beneficiarios fue en el 2023-2 con 3139 estudiantes, seguida por la convocatoria 2019-2 con 2675 estudiantes legalizados.

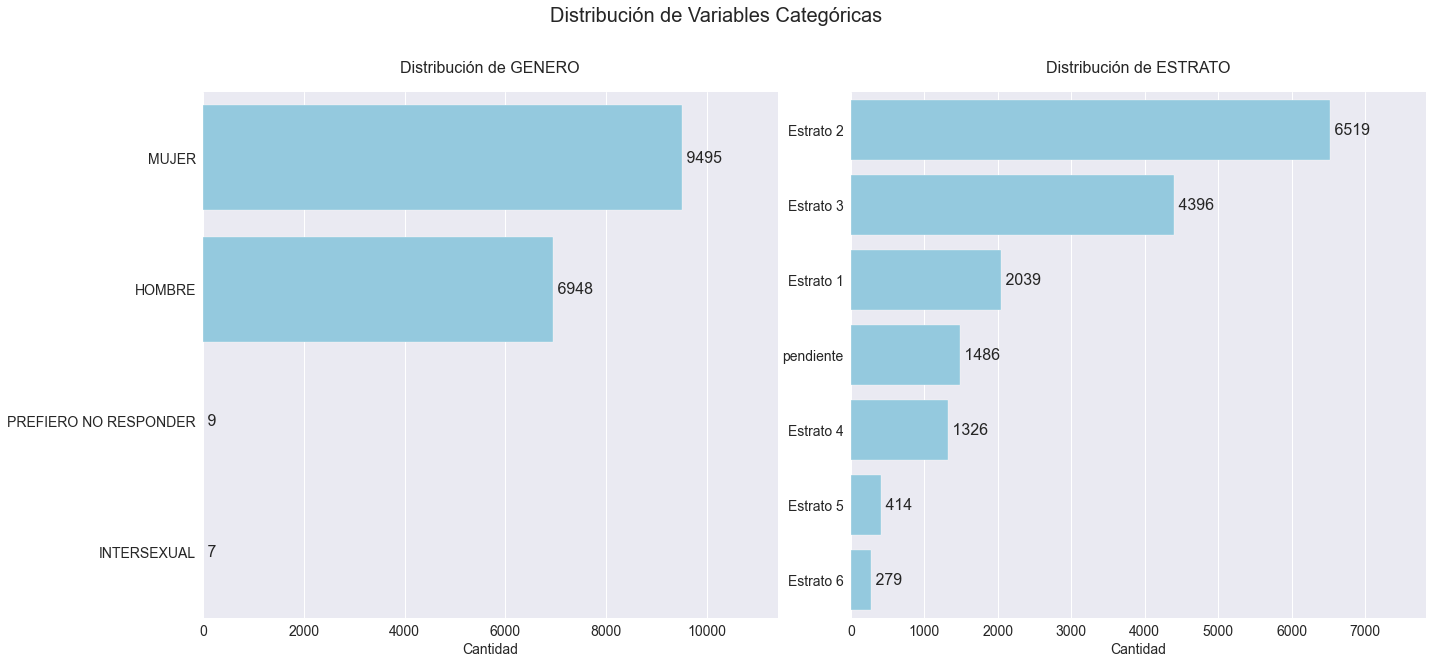


Figura 3.3.5. Distribución de las variables Género y Estrato

La variable **GENERO**, representa el género por el cual se identificaron los beneficiarios a la hora de legalizar su crédito. Se observa una gran diferencia entre hombres y mujeres que gozan de los beneficios ofrecidos por Sapiencia. Donde 9495 son mujeres mientras que 6948 son hombres, 7 son de género intersexual y 9 prefieren no responder. Esto refleja una representación femenina significativa en los créditos educativos condonables, lo cual podría estar relacionado con la demanda educativa o la estructura demográfica de la población estudiantil. Seguidamente, la variable **ESTRATO**, representa el estrato socioeconómico del beneficiario. Donde la mayoría de beneficiarios son de los estratos 2 y 3 con un total de 6519 y 4396 respectivamente.

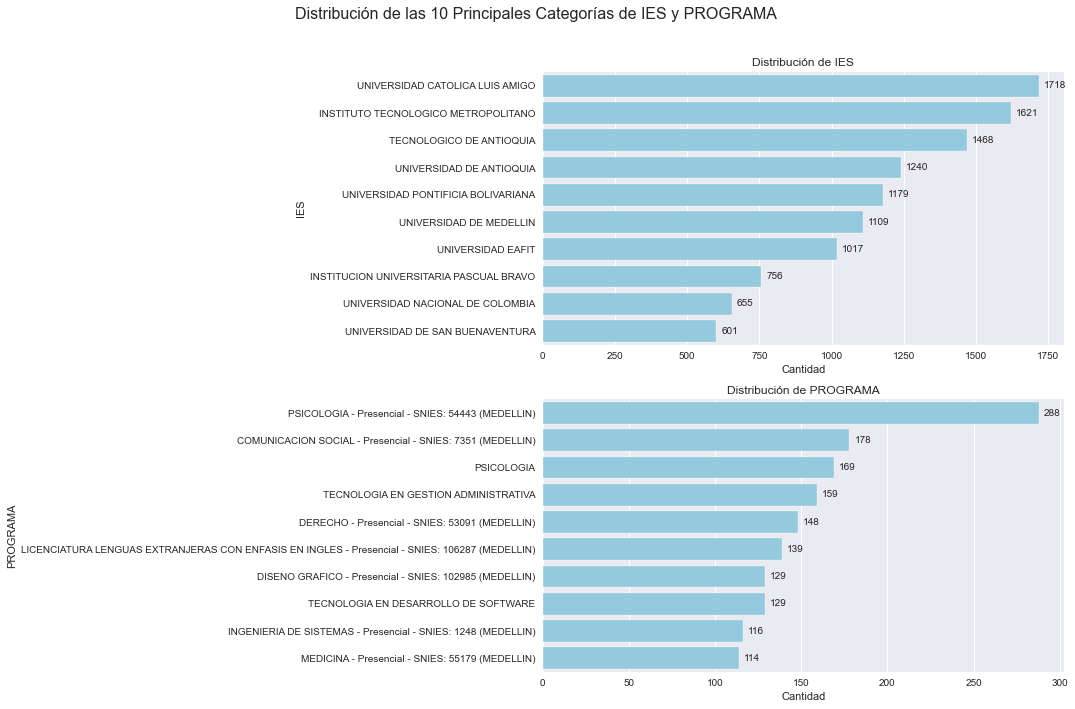


Figura 3.3.6. Distribución de las variables IES y Programa

La variable **IES**, representa la institución de educación superior en la cual estudian los beneficiarios. Se observa como las universidades que más beneficiarios de los fondos Sapiencia son la UNIVERSIDAD CATÓLICA LUIS AMIGÓ con 1718 estudiantes, el INSTITUTO TECNOLÓGICO METROPOLITANO (ITM) con 1621, el TECNOLÓGICO DE ANTIOQUIA (TDA) con 1486 y la UNIVERSIDAD DE ANTIOQUIA con 1240 beneficiarios.

La variable **PROGRAMA**, representa el nombre del programa que está estudiando el beneficiario. Se observa como, los programas que suelen ser más estudiados por los beneficiarios son PSICOLOGÍA con 469 y COMUNICACIÓN SOCIAL con 178 beneficiarios respectivamente.

# 

# 

# 

# 

# 

# 

**Referencias**

Barone, S. (2006). *Multivariate analysis of student loan defaulters at Prairie View A&M University*. TG Research and Analytical Services. https://files.eric.ed.gov/fulltext/ED542503.pdf

Greene, L. L. (1989). An Economic Analysis of Student Loan Default. *Educational Evaluation and Policy Analysis*, *11*(1), 61-68. https://doi.org/10.2307/1163716

ICETEX. (n.d.). *¿Quiénes somos?* Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior. Retrieved 10 20, 2024, from https://web.icetex.gov.co/el-icetex/informacion-institucional/quienes-somos

Medellín, S. (2023, enero 13). *Historia sapiencia - sapiencia*. Sapiencia Agencia de Educación Postsecundaria de Medellín. https://sapiencia.gov.co/historia/

Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine learning for credit risk prediction: A systematic literature review. *Data*, *8*(11), 169. https://doi.org/10.3390/data8110169

UNESCO. ((2020)). Hacia el acceso universal a la educación superior: tendencias internacionales. https://www.iesalc.unesco.org/wp-content/uploads/2020/11/acceso-universal-a-la-ES-ESPANOL.pdf

Zhou, L., & Wang, H. (2017). Prediction on large imbalanced data using random forests. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, *10*(6), 1519–1525.

# Anexos

## Anexo 1. Autoarchivo en Repositorio y documentos de interés

Repositorio de Github:

<https://github.com/genaroAE99/Seminario_UdeA/tree/main>

Decreto 032 de 2023 normativa oficial inscripciones créditos condonables linea pregrado en Sapiencia:

<https://sapiencia.gov.co/wp-content/uploads/2024/04/decreto-32-de-2023_compressed.pdf>