

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Construcción y depuración de una base de datos como insumo para el desarrollo de un modelo de segmentación de clientes.

INFORME DE PASANTÍA PARA OPTAR POR EL TÍTULO DE MATEMÁTICO
PROYECTO CURRICULAR DE MATEMÁTICAS

Paula Andrea Baquero Gómez
Profesional designado: Adriano Polo Losada
Docente director: Luis Alejandro Masmela Caita

Bogotá DC
Mayo de 2023

Índice

1. Introducción **2**

2. Propuesta de pasantía **3**
2.1. Objetivos . . . 3
2.1.1. Objetivo general . . . 3
2.1.2. Objetivos específicos . . . 3
2.2. Plan de trabajo . . . 5

3. Descripción de resultados **6**
3.1. Conceptos previos . . . 6
3.2. Desarrollo de base de datos . . . 8
3.2.1. Procesamiento de datos . . . 8
3.2.2. Estrategias de prevención de duplicidad de datos . . . 13
3.2.3. Limpieza de datos . . . 14
3.2.4. Cálculos de algunas variables . . . 16
3.2.5. Tabla resultante . . . 17

4. Análisis exploratorio de resultados **18**
4.1. Valores faltantes . . . 19
4.2. Frecuencias . . . 21
4.3. Valores mínimos y máximos . . . 24

5. Herramientas y Tecnologías Utilizadas **24**

6. Evaluación y cumplimiento de objetivos **25**

7. Conclusiones **26**

1. Introducción

El desarrollo de la presente pasantía se dio en DataKnow empresa de consultoría analítica en la cual, dada su razón de ser, provee de servicios profesionales enfocados en dar valor a los datos mediante el uso de distintas herramientas. DataKnow es una empresa joven de Medellín, la cual nace en 2018, cuenta con un equipo de profesionales altamente calificados en tres áreas de los datos: la arquitectura, la ingeniería y la ciencia de datos. DataKnow colabora con empresas reconocidas en el sector tecnológico como Microsoft y SAS. Además, ha sido reconocida como una de las empresas más innovadoras en tecnologías de la información en Colombia, ocupando un lugar destacado entre las 15 mejores del país.[1]

Dentro del desarrollo de los diversos proyectos que se ejecutan en la empresa, surge uno en el cual el reto es hacer un modelo de Pricing, enfocado al sector bancario. De manera general un modelo Pricing, está enfocado en una metodología en la cual se define el valor que se le asignará a un determinado producto para posteriormente comercializarse. En este proyecto, se busca determinar la tasa de interés óptima para los diferentes productos del banco, tanto de tipo activo como pasivo, mediante la segmentación de los clientes con ayuda del análisis de datos. Con el objetivo de ofrecer un producto más personalizado, se requiere consolidar información del cliente en una base descriptiva acorde a cada producto, que se entregará como insumo al área encargada para entrenar modelos y generar categorías según las características de cada cliente en el mercado.

De manera general los clientes en el banco se clasifican en dos grandes grupos los existentes (al menos cuentan con un producto activo en el banco) y los nuevos. Dentro de estos grupos viene el tipo de producto a solicitar, dentro de la familia del activo están créditos hipotecarios, libranza, tarjetas de crédito, rotativos entre otros. Y dentro de los pasivos, se encuentran CDT, cuentas de ahorro y corrientes. Con base a lo descrito anteriormente, el proceso se dividirá en tres partes iniciales para encaminar el proyecto en el desarrollo del modelo de segmentación. La primera consistirá en la construcción de una base de datos que incluirá información de riesgo, datos sociodemográficos, datos transaccionales e información de productos para clientes nuevos y existentes. En la segunda parte, se revisará la base de datos construida para validar los cruces utilizados, evaluar el comportamiento de los clientes durante los diferentes periodos y verificar la coherencia de los datos. La tercera parte se enfocará en mantener la calidad de los datos, abordando los valores faltantes y las distribuciones de los datos.

Para el desarrollo de la pasantía se usará el software SAS® Guide (SAS Institute, 2021), un software estadístico para la gestión de datos, análisis avanzado, análisis multivariante, inteligencia comercial, investigación criminal y análisis predictivo. Adicionalmente, se utilizará SQL, un lenguaje de programación utilizado para la gestión de bases de datos, con el fin de realizar consultas y manipulaciones de los datos necesarios para el análisis. También se utilizará Python versión 3.9.7 para desarrollar la exploración de los datos generados y generar

[1] Mas información en: https://www.portafolio.co/innovacion/empresas-y-startups-con-mas-abierta-del-pais-segun-100-open-startups-556022

alertas sobre la fiabilidad de los mismos.

2. Propuesta de pasantía

2.1. Objetivos

2.1.1. Objetivo general

Crear una base de datos detallada para proporcionar una mejor comprensión de los clientes y sus comportamientos financieros. La información recopilada en la base de datos se utilizará como insumo para el equipo encargado de la segmentación, quien será responsable de analizar los datos y crear los perfiles de clientes correspondientes con el fin de ofrecer tasas de interés competitivas y personalizadas basadas en el perfil de cada cliente.

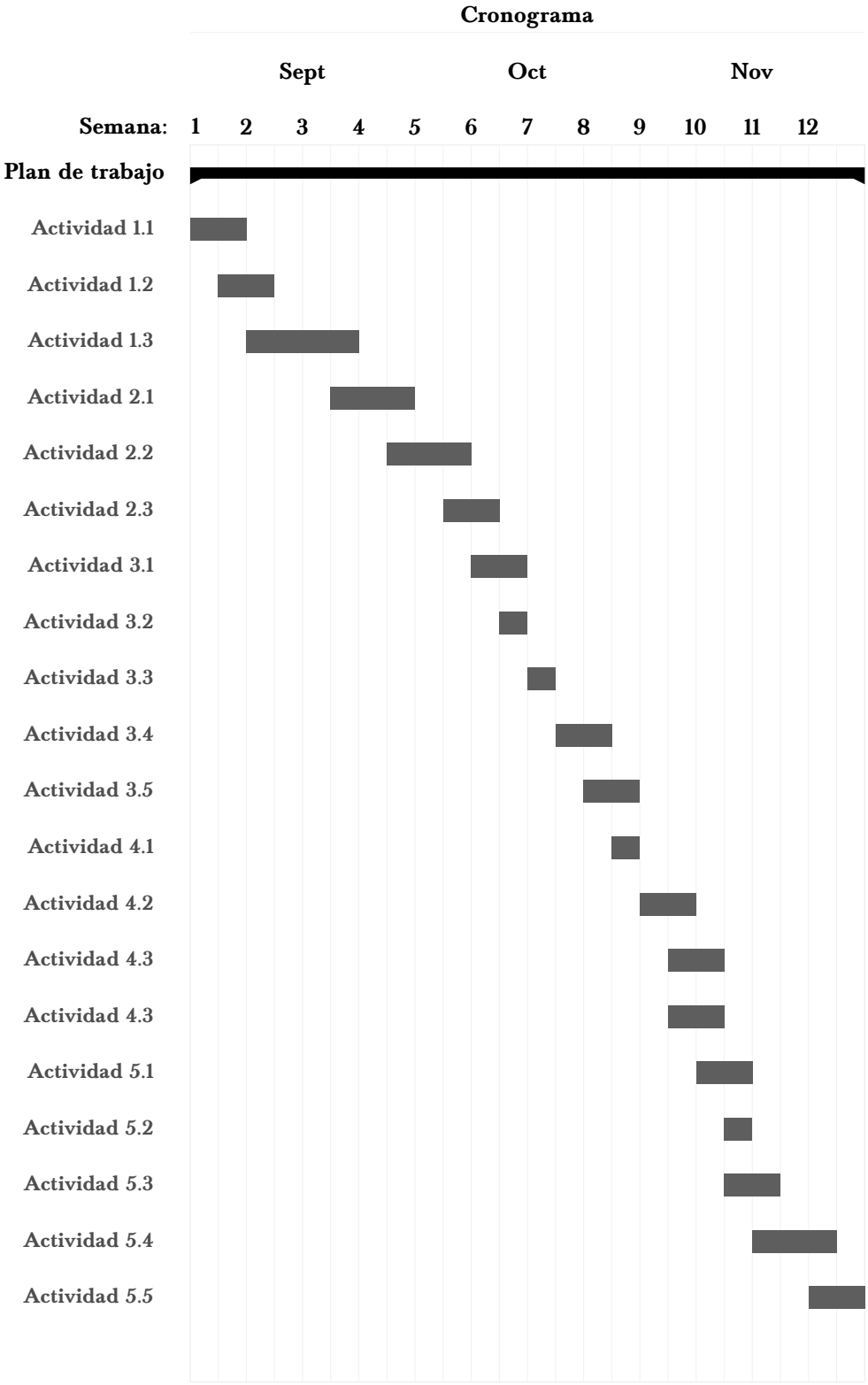
2.1.2. Objetivos específicos

- Recopilar y unificar la información disponible de todos los clientes en una base de datos general utilizando herramientas de procesamiento de datos como Python, SQL y SAS, lo que permitirá una gestión más eficiente y una mejor comprensión del comportamiento del mercado.
- Implementar estrategias para evitar la duplicidad de valores para asegurar la integridad de la información.
- Garantizar la calidad y precisión de la información en la base de datos mediante la aplicación de procesos de limpieza y preprocesamiento de datos.
- Efectuar un seguimiento periódico de la proporción de datos y el comportamiento de las variables a lo largo del tiempo. Esto permitirá detectar cualquier cambio significativo en las proporciones y tomar medidas para corregirlo, garantizando así la calidad y consistencia de los datos en la base de datos.
- Realizar un análisis detallado de los datos para identificar valores faltantes, errores y posibles inconsistencias en la información, a fin de garantizar la calidad y la fiabilidad de los datos recolectados.

Objetivo	Meta	Actividad
1.	Recopilar y unificar la información.	1. Identificación de fuentes necesarias. 2. Mapeo de fuentes. 3. Generación del código de extracción.
2.	Implementación de estrategias para evitar la duplicidad de datos.	1. Identificación de posibles causas. 2. Establecer criterios de detección. 3. Desarrollo de estrategia para evitar duplicidad.
3.	Limpieza y preprocesamiento de datos.	1. Verificación de consistencia de datos. 2. Estructuración de columnas. 3. Estandarización de nombres de las columnas. 4. Estandarización de texto para homologar los valores categóricos. 5. Revisión de tipos de datos.
4.	Establecer un proceso de revisión periódica del comportamiento en el tiempo.	1. Definir los indicadores a revisar. 2. Definir la frecuencia de revisión. 3. Establecer un proceso de análisis de tendencias.
5.	Análisis detallado de los datos para identificar faltantes, errores e inconsistencias en la información.	1. Identificar columnas con valores faltantes. 2. Identificar valores atípicos. 3. Corrección de inconsistencias en los casos que aplique. 4. Documentación de cambios. 5. Verificar integridad de los datos posterior a los cambios.

Cuadro 1: Tabla de objetivos, metas y actividades para la creación de una base de datos unificada.

2.2. Plan de trabajo



3. Descripción de resultados

3.1. Conceptos previos

Antes de iniciar con el desarrollo de los objetivos de la pasantía, es fundamental revisar algunos conceptos previos que permitan entender en detalle la construcción de una base de datos de calidad y eficiente. En este sentido, las bases de datos son herramientas esenciales en la gestión de información. Una base de datos es un conjunto de datos relacionados entre sí y organizados de tal manera que sea fácil acceder y recuperar información de manera eficiente. Las bases de datos se utilizan en todo tipo de organizaciones y en muchos campos, desde la gestión empresarial hasta la investigación científica. En la pasantía, se ha enfocado en la construcción de una base de datos para un modelo analítico, lo cual implica la recopilación y organización de información de diversas fuentes. Para garantizar la calidad, integridad y eficiencia de los datos en la base de datos, es fundamental tener en cuenta algunos conceptos previos, como las bases de datos, las uniones, las llaves primarias, los modelos de datos y los atributos, entre otros. En esta sección, se presentarán en detalle estos conceptos previos y su aplicación en la construcción de la base de datos.

En cuanto a la calidad de los datos, es fundamental contar con estrategias de limpieza y preprocesamiento de datos que permitan identificar y corregir valores faltantes, errores y posibles inconsistencias en la información recopilada. Estas estrategias incluyen la eliminación de registros duplicados, la corrección de errores de entrada y la normalización de los datos. En cuanto a la integridad de los datos, se deben implementar estrategias para evitar la duplicidad de valores y asegurar la coherencia de los datos en toda la base de datos. Por ejemplo, se puede utilizar una llave primaria para cada registro que permita la identificación única de cada registro y se pueden establecer restricciones para evitar la introducción de datos inválidos, este concepto se abordará en detalle mas adelante. En cuanto a la eficiencia de los datos, es importante implementar estrategias de optimización de consultas y de diseño de la base de datos que permitan un acceso rápido y eficiente a la información. Estas estrategias incluyen la utilización de índices y la selección de las técnicas de cruces o join más apropiadas para cada caso.

Para construir una base de datos, es fundamental entender el **modelo entidad-relación** (E-R), una herramienta utilizada en la gestión de bases de datos para representar las entidades relevantes de un sistema y las relaciones entre ellas. El modelo se compone de **entidades**, **atributos y relaciones** que se organizan en un diagrama para facilitar la comprensión de la estructura de la base de datos. Las entidades representan objetos o conceptos del mundo real, mientras que las relaciones representan las asociaciones entre estas entidades. Los atributos son características o propiedades que describen a las entidades o relaciones.[5, p. 55]

Antes de construir una base de datos, es necesario definir una **llave primaria**. Esta es un atributo o conjunto de atributos que identifica de forma única cada registro en una tabla. Es decir, no puede haber dos registros con la misma llave primaria. La llave primaria se utiliza para establecer relaciones entre tablas y mantener la integridad de los datos. En la mayoría de los casos, la llave primaria se define como un campo autoincremental, que toma valores únicos y consecutivos para cada registro. También es posible utilizar una combinación de campos para definir la llave primaria, en cuyo caso se habla de una llave primaria compuesta [11, p. 24].

El modelo entidad-relación es una técnica útil para diseñar y planificar una base de datos antes

de su implementación. Por ejemplo, en el caso de un banco, el conjunto de todos los clientes sería un conjunto de entidades, pero este conjunto de entidades no necesariamente debe ser disjunto; es decir, pueden existir empleados que sean a su vez clientes. Los atributos de cada entidad, en el caso de los clientes, podrían ser el ID de cliente, documento, nombre y algunos datos demográficos. Las relaciones pueden establecerse entre entidades, por ejemplo, entre clientes y préstamos, mediante una llave que permita generar la relación. Por ejemplo, el cliente con ID 123 tiene un préstamo hipotecario H-20.

En el modelo de entidad-relación, la **correspondencia de cardinalidades** se utiliza para indicar la cantidad de instancias de una entidad que pueden estar asociadas con una instancia de otra entidad. Por ejemplo, una instancia de la entidad “cliente"puede estar asociada con ninguna o varias instancias de la entidad “préstamo", pero una instancia de la entidad ”préstamo"solo puede estar asociada con una instancia de la entidad “cliente". Dicha correspondencia de cardinalidades debe ser una de las siguientes:

- **Uno a uno**: Implica que una entidad en A se asocia a lo sumo con una entidad en B, y una entidad en B se asocia con a lo sumo una entidad en A
- **Uno a varios**: Una entidad en A se asocia con cualquier número de entidades en B (ninguna o varias). Mientras que una entidad de B se puede tan solo asociar a una entidad en A.
- **Varios a uno**: Una entidad en A se asocia a lo sumo con una entidad en B. Mientras que una entidad en B se asocia con cualquier número de entidades (ninguna o varias) en A.
- **Varios a varios**: una entidad en A se asocia con cualquier cantidad de entidades de B y de igual manera B con A.

Los anteriores conceptos, se pueden ver gráficamente como en la Figura 1. [11, p. 23] .

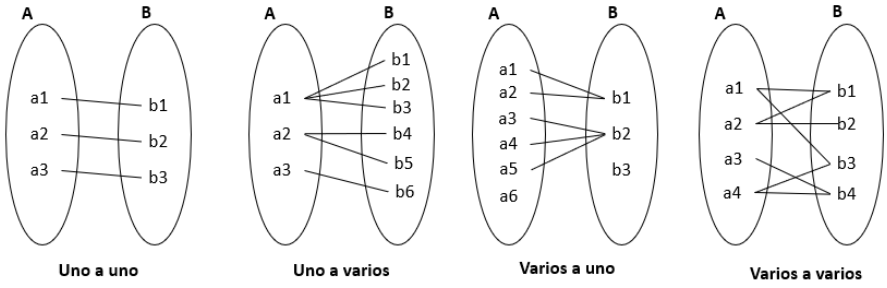


Figura 1: Gráfico de correspondencia de cardinalidades, donde se representa mediante flechas la relación entre dos conjuntos de datos.

Una vez que se han establecido las relaciones entre entidades, es posible combinar tablas utilizando los tipos de uniones o joins. El cual permite combinar dos o más tablas en una sola, utilizando una columna en común como referencia. Es necesario tener un buen conocimiento de estos tipos de uniones y cómo aplicarlos para lograr una correcta combinación de información y obtener resultados precisos. Entre los tipos de unión más utilizados, se encuentran:

- **Unión izquierda (Left Join)**: Si se tienen dos conjuntos A y B, a la izquierda y derecha respectivamente, esta operación devolverá todo lo del conjunto A, incluso si no existen valores coincidentes para registros en el conjunto B.
- **Unión derecha (Right join)**: Si se tienen dos conjuntos A y B, a la izquierda y derecha respectivamente, esta operación devolverá todo lo del conjunto B, incluso si no existen valores coincidentes para registros en el conjunto A.
- **Unión interna (Inner Join)**: Si tenemos los conjuntos A y B, esta operación devolverá solo los registros que coincidan en ambos conjuntos, es decir, la intersección.

Entendiendo como conjuntos A y B las tablas que se van a relacionar, en la Figura 2 se tiene gráficamente los tipos de uniones [9].

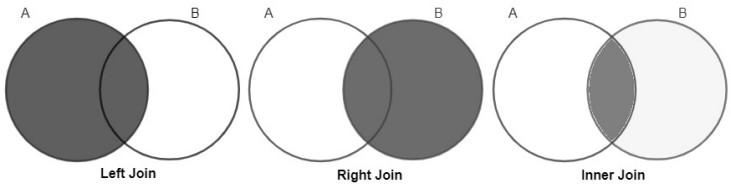


Figura 2: Representación gráfica de los tipos de uniones en bases de datos.

Una vez que se han establecido las relaciones entre entidades, es posible combinar tablas utilizando los tipos de uniones o joins. Estos tipos de uniones son fundamentales en el proceso de ETL (Extracción, Transformación y Carga) en la construcción de una base de datos. El proceso de ETL se refiere a la extracción de datos desde diferentes fuentes, su transformación para asegurar la consistencia y coherencia, y finalmente, la carga de los datos en el sistema de destino. [10, p. 122]. Durante la pasantía, se ha aplicado el proceso de ETL en SAS Guide utilizando una guía de referencia para la implementación del código. Esta guía, titulada [Título del libro], ha sido una herramienta invaluable para llevar a cabo las tareas de extracción, transformación y carga de datos de manera eficiente y precisa. El conocimiento de los tipos de uniones es esencial para lograr una correcta combinación de información y obtener resultados precisos en el proceso de ETL.

3.2. Desarrollo de base de datos

3.2.1. Procesamiento de datos

Después de comprender los conceptos básicos, podemos describir los pasos para la construcción de la base de datos general. En primer lugar, se llevó a cabo un análisis detallado de los requerimientos del cliente para identificar las entidades y relaciones necesarias en la base de datos. Luego, se diseñó un diagrama entidad-relación que sirvió como guía para la creación de las tablas y campos correspondientes en la base de datos. Cada tabla se creó cuidadosamente con una llave primaria definida y se establecieron las relaciones entre ellas.

En el desarrollo de la base de clientes, se comenzó con una tabla que contenía información de negocio sobre los clientes del banco en estados activos o inactivos durante un período de trece meses. A partir de ahí, se consultaron otras bases de datos para obtener información adicional, como la cantidad de transacciones débito y crédito por mes, información de créditos

hipotecarios, SPL, consumo, información referente a CDT y otros productos de otros bancos. Con el fin de extraer toda la información necesaria, se realizaron cruces de más de doce tablas. La construcción final de la tabla de clientes se basó en el análisis y el cruce de esta información.

En la Figura 3, se muestra el mapeo de la construcción de la base de datos y cómo se relacionan las tablas entre sí.

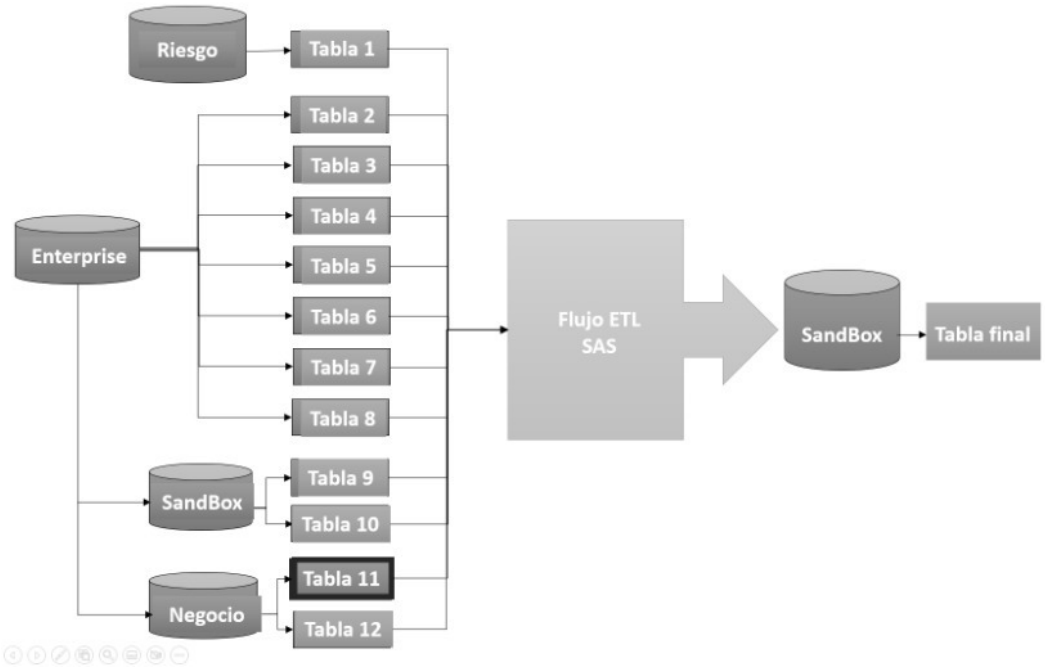


Figura 3: Flujo de alto nivel para la construcción de la base. El diagrama representa el proceso general de la construcción de la base de datos. Los cilindros representan las diferentes librerías de datos utilizadas y las cajas rectangulares representan las tablas de esas librerías. Las flechas indican el flujo de información entre las diferentes tablas y librerías durante el proceso de construcción de la base de datos.

En este caso, se extrajo la información de cuatro librerías a las que llamaremos Riesgo, Enterprise, SandBox y Negocio. La librería Riesgo contenía información que provenía de Experian[2]. La librería Enterprise alberga la información bancaria de los procesos automatizados y es la fuente principal de datos, de aquí se extrajo toda información referente a los productos como CDT, créditos libre destino, hipotecarios, información transaccional de tarjetas de crédito y débito, entre otros. La librería SandBox alberga los procesos en desarrollo, y se obtuvo información construida específicamente para este proceso, ya que hubo algunas variables que no se encontraron en ninguna otra base de datos y que se tuvieron que construir. Debido al gran procesamiento de los datos, se dejó como un proceso independiente en tablas alojadas en SandBox. Más adelante se ampliará sobre estas bases auxiliares construidas. Finalmente, la librería Negocio proporcionó dos tablas de las cuales la Tabla 11 de la Figura 3 fue nuestro punto de partida, ya que contiene el resumen de clientes activos e inactivos del banco desde junio 2021 hasta junio 2022, un período de historia de 13 meses. A partir de esta tabla, se extrajo toda la información requerida de las otras tablas, con ayuda de los cruces antes descritos. La idea

[2]Experian es una empresa multinacional de informes de crédito al consumo que recopila y agrega información sobre más de mil millones de personas y empresas.

siempre fue evitar la duplicidad de clientes y para ello se validó la cantidad de datos resultantes de cada cruce realizado, cuya cantidad siempre debe estar acorde a la Tabla 11 de partida.

Dentro de los desarrollos de la construcción de la base general, la mayoría de los cruces son del tipo uno a varios o uno a uno. Por ejemplo, partiendo de la tabla de negocio en la cual tenemos el registro de las personas con las que se iniciará el proceso, la tabla contaba con aproximadamente 4 millones de registros por mes. En esta tabla contábamos con información socio-demográfica, pero se requería la información consolidada de sus transacciones de tarjetas crédito y débito. Dicha tabla transaccional contiene la información de forma tal que por cada transacción realizada con algún producto aparecerá en una fila con todas sus descripciones. Por lo tanto, para el proyecto solo se requería los valores totales de las transacciones realizadas por cada cliente en un corte determinado, se debieron agrupar y definir que un cliente en la tabla transaccional apareciera una sola vez con su total de transacciones para evitar duplicidad de los datos. El flujo para el procesamiento de los datos se puede observar en los siguientes pseudocódigos:

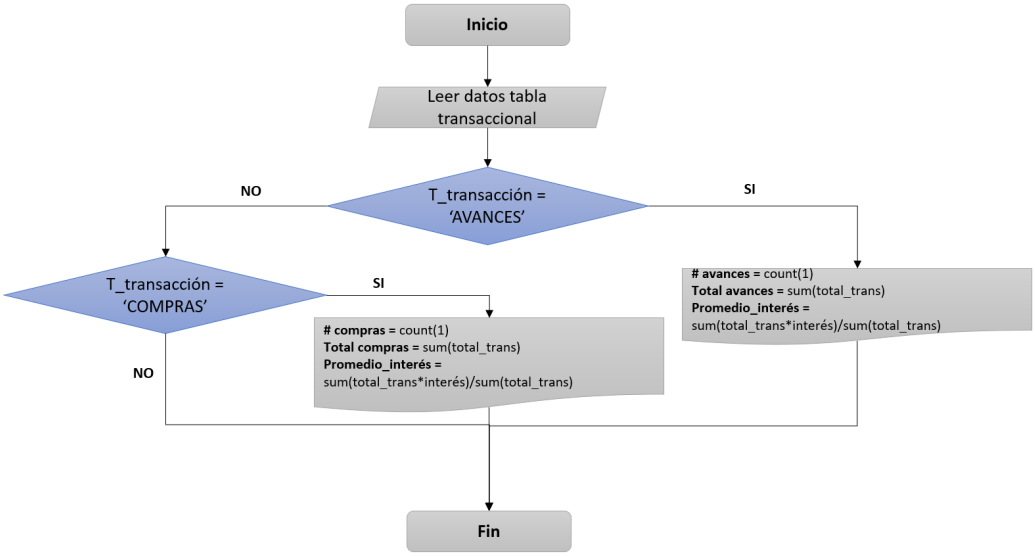


Figura 4: Diagrama de flujo de pseudocódigo que describe el proceso de conteo de transacciones y cálculo del promedio ponderado de la tasa de interés en base al tipo de transacción.

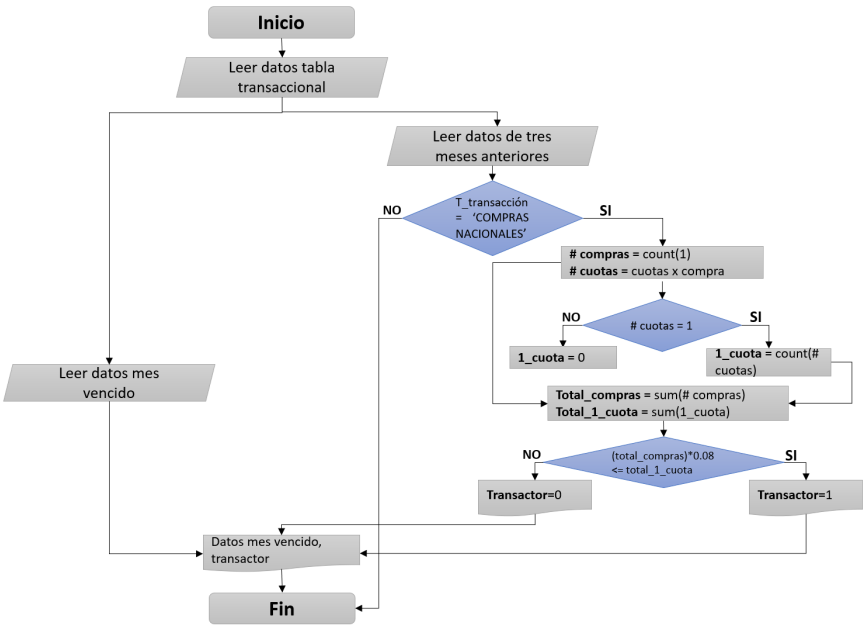


Figura 5: Diagrama de flujo de pseudocódigo que muestra el proceso para crear la variable cliente transactor. Un cliente es considerado transactor si en los últimos tres meses ha realizado compras nacionales y ha diferido más del 80 % del total de sus compras a una sola cuota.

En resumen, se agruparon las transacciones por tipo y se sumaron los valores totales para cada cliente. Además, se creó una variable para identificar si el cliente realizó alguna transacción durante el período de interés, lo que permitió eliminar aquellos clientes que no realizaron ninguna transacción y reducir el tamaño de la tabla.

Observemos cómo se transforman algunos de los registros de la tabla transaccional, basándonos en los pseudocódigos de la Figura 4 y la Figura 5.

- En primer lugar, tenemos la tabla transaccional, donde la columna KEY_BUSINESS es una llave que concatena el tipo de documento y el número. Por otro lado, COD_TRX es el código de la transacción, NO_PROD es el número del producto asociado a la tarjeta de crédito que realizó la transacción, DATE es la fecha de la transacción, TRANS_TYPE es el tipo de transacción realizada, TOTAL_TRANS es el monto total de la transacción, TENOR_CNT es el número de cuotas en las que se difirió la transacción y, finalmente, INT_RATE es el interés aplicado a la transacción. Esta tabla se puede representar de manera resumida como se muestra en la Cuadro 2.

KEY_BUSINESS	COD_TRX	NO_PROD	DATE	TRANS_TYPE	TOTAL_TRANS	TENOR_CNT	INT_RATE
CC100756321	145	1126	1-jun	PAGO	23000		
CC100756321	234	1126	11-jun	COMPRA	433000	1	2.2
CC100756321	453	1126	12-jun	AVANCE	500000	2	1.9
CC100756321	544	1126	12-jun	COMPRA	90000	1	2.2
CC111223422	123	1890	9-jun	COMPRA	1200000	12	2.2
CC111223422	332	1890	15-jun	COMPRA	950000	1	2.2
CC111223422	5622	1890	20-jun	AVANCE	1000000	3	1.98
CC129863265	5632	2643	3-jun	COMPRA	232000	1	2.2
CC129863265	789	2643	28-jun	COMPRA	30000	4	2.2

Cuadro 2: Ejemplo tabla inicial de transacciones con tarjeta de crédito. Donde cada cliente aparece según la cantidad de transacciones realizadas en el corte.

- Posteriormente, al clasificar el tipo de transacción realizada, se agrupan las transacciones por cliente. Se obtienen los promedios ponderados del interés de todas las compras y

avances, el promedio ponderado de cuotas a las que se difirieron las compras y los avances, la suma total de las compras y avances y, por último, la fecha de corte a la que pertenece la transacción. Es importante mencionar que la fórmula para calcular un promedio ponderado es:

$$PP = \frac{p_1 X_1 + p_2 X_2 + ... + p_n X_n}{p_1 + p_2 + ... + p_n},$$

donde $(X_1, X_2, ..., X_n)$ es el conjunto de datos, por ejemplo las tasas de interés de todas las compras, y $(p_1, p_2, ..., p_n)$ son los pesos, en este caso, los montos totales de las transacciones. Dado que se está agrupando por cliente, el resumen de datos transaccionales se reduce a una fila por cliente.

- En el pseudocódigo de la Figura 5, se obtiene el resumen de transacciones de los últimos tres meses y se realiza un condicional que es un contador que, cada vez que el cliente difiere a una cuota una compra nacional, suma 1. Esto se hace con el fin de obtener una variable que nos indique cuándo un cliente es transactor, es decir, todos los clientes que durante los últimos tres meses han diferido más del 80% de sus compras a una cuota. Para ello, se calcula inicialmente el total de todas las compras de los últimos tres meses y el total de compras diferidas a una cuota. Posteriormente, se valida que se cumpla la condición de que el 80% de las compras es menor o igual a la cantidad de compras a una cuota.
- Finalmente, los resultados obtenidos en cuanto a las agrupaciones realizadas en el pseudocódigo de la Figura 4 y el cálculo de la variable transactor en el pseudocódigo de la Figura 5 se unen por medio de un *left join* con una relación **uno a uno**. Recordemos que la variable FLG_TRANSAC, a diferencia de todas las otras, depende de un histórico de tres meses, por lo cual puede darse el caso en el que un cliente no realice compras en Junio, pero en Mayo y Abril sí, cuyo promedio de cuotas fue 1 en más del 80% de las transacciones. Despues de realizar los procesos anteriores obtenemos un resultado similar al presentado en el Cuadro 3

KEY_BUSINESS	NO_TRX_AVAN	NO_TRX_COMP	AVAN_TOTAL	AVAN_INT	AVAN_TENOR	COMP_TOTAL	COMP_INT	COMP_TENOR	FLG_TRANSAC
CC100756321	1	2	500000	1.9	2	523000	2.2	1	1
CC111223422	1	2	1000000	1.98	3	2150000	2.2	7.139534884	0
CC129863265	0	2	0	0	0	262000	2.2	1.34351145	0

Cuadro 3: Ejemplo tabla que muestra el resultado después de la agregación de datos y la adición de la variable FLG_TRANSAC, donde cada cliente aparece una sola vez con su resumen transaccional correspondiente.

Con el resultado obtenido en la tabla consolidada, se procede a agregar esta información a la tabla original para construir la base maestra. Debido a que el proceso anterior garantiza que cada cliente aparece solo una vez, esta adición se realizará mediante una relación uno a uno utilizando un *left join*, en el que la tabla izquierda es la tabla de negocio y la tabla derecha es la tabla consolidada transaccional construida en los pasos anteriores. Este es solo un ejemplo del tipo de procesamiento y anidamiento que se realizó para obtener la tabla consolidada con toda la información necesaria. No se muestra en detalle todo el proceso, ya que implica el cruce de más de 12 tablas.

3.2.2. Estrategias de prevención de duplicidad de datos

Durante la pasantía, uno de los principales objetivos fue garantizar la integridad de la información manejada en la organización. La duplicidad de valores en la base de datos puede ser un problema que afecta la calidad de la información y, en consecuencia, la toma de decisiones. Para evitar la duplicidad de valores, se implementaron diversas estrategias, incluyendo validaciones en la entrada de datos y definición de restricciones de clave única en la base de datos.

Es importante destacar que no toda duplicidad es un problema. Por ejemplo, en una tabla de clientes, no tiene sentido tener clientes repetidos, pero en una tabla de detalles de tarjetas de crédito, un cliente puede tener varias tarjetas asociadas a su número de identificación. Por lo tanto, es fundamental evaluar cada caso para determinar si la duplicidad es un problema y cómo abordarlo.

Para garantizar la integridad de los datos y evitar la duplicidad, es importante definir la llave primaria de cada tabla que se va a cruzar. La llave primaria debe ser única y no puede tener valores nulos. Además, en algunos casos, puede ser necesario realizar agrupaciones de los datos para eliminar duplicados y tener una vista más clara y precisa de la información.

En el caso específico de las transacciones, se decidió agrupar los registros por cliente para tener una vista más ordenada y simplificada de los datos. Para validar la duplicidad de datos, se emplearon diversos algoritmos y técnicas. Uno de ellos, que se describe a continuación, consistió en contar los valores de una columna y verificar que cada valor apareciera solamente una vez.

A continuación se muestra el pseudocódigo del algoritmo utilizado en SAS para validar la duplicidad de datos:

Algoritmo 1 Validar duplicidad de valores en una columna

```
i = 0
for cada fila en la tabla do
    if valor de la columna es igual al valor a contar then
        i = i + 1
        Poner el valor de i en la columna correspondiente de la fila
    end if
end for
```

El Algoritmo 1 contaba los valores de una columna y verificaba que cada valor apareciera solamente una vez. De esta manera, se utilizaba esta técnica para validar la integridad de los datos y prevenir la duplicidad en la base de datos. Si algún valor aparecía más de una vez, se generaba una alerta indicando la presencia de duplicidad en la columna considerada como llave.

Una vez analizados los resultados, se definía la estrategia para eliminar los registros duplicados. En algunos casos, bastaba con definir un *DISTINCT* en la consulta para eliminar los valores duplicados y garantizar la integridad de los datos. Esta cláusula se utiliza para seleccionar solo los valores únicos de una columna específica en una tabla, eliminando así los valores duplicados y permitiendo una vista más clara y precisa de la información. En otros casos, se debía revisar

las tablas de cruce y determinar si se debía realizar algún agrupamiento adicional en una de las tablas o considerar alguna regla adicional para evitar la duplicación de información.

3.2.3. Limpieza de datos

La limpieza de datos es una tarea fundamental en cualquier proyecto de datos. Durante el desarrollo de la pasantía, se trabajo con grandes cantidades de datos que presentaban errores e inconsistencias. Para garantizar la calidad de los datos, fue necesario llevar a cabo un proceso de limpieza riguroso y detallado. En primer lugar, se realizó una exploración de los datos para identificar posibles problemas y patrones en la información. Además, se corrigieron errores ortográficos y se unificaron los formatos de los datos. La limpieza de datos facilitó la interpretación y el manejo de la información. A continuación, detallaré los pasos específicos que se llevaron a cabo en el proceso de limpieza y las herramientas utilizadas para ello.

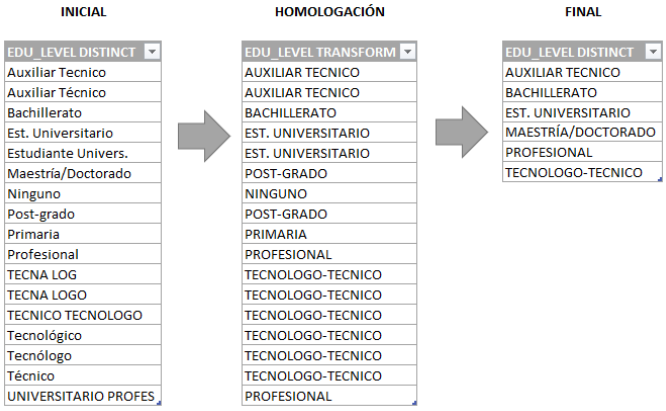


Figura 6: Proceso de limpieza de la variable EDU_LEVEL. En la primera tabla se presentan las categorías iniciales de la variable, seguidas por una segunda tabla que muestra la estandarización de todas las categorías en mayúsculas. Finalmente, en una tercera tabla se presentan las categorías finales luego de haber sido unificadas.

Algunas de las variables categóricas, especialmente las que hacían referencia a datos sociodemográficos, requerían ser tratadas. Por ejemplo, la variable de nivel educativo (EDU_LEVEL) tenía 17 categorías que necesitaban ser unificadas, como se observa en la Figura 6. En el siguiente algoritmo (Algoritmo 2) se muestra el proceso de unificación de categorías.

Con lo cual, de 17 categorías iniciales, las cuales en su mayoría eran redundantes pasamos a tener 6 categorías finales. Otro proceso de limpieza de datos se originó con la edad del cliente, por temas de seguridad de la información no se debía proporcionar la edad del cliente sino el rango en el cual se encuentra, es decir si un cliente tenía 35 años, el valor a mostrar es: "30 a 40 años". En el Algoritmo 3 se muestra el proceso para el cálculo del rango para algunas de las edades.

El calculo era la diferencia de la fecha de nacimiento con respecto a la fecha de corte del cliente, no con la fecha actual ni con la fecha en la que se ejecuta el proceso. En total se obtuvieron más de 11 rangos de edad. Sin embargo, se encontró población en rangos de edad superiores a los 110 años, lo cual resulta algo atípico. La distribución de las edades obtenidas se muestra en la Figura 7.

Algoritmo 2 Unificación de categorías

```
Leer la variable categórica
Convertir todas las letras a mayúsculas
if la categoría contiene ”AUXILIAR” o ”TÉCNICO” then
    Asignar la categoría ”AUXILIAR TECNICO”
else if la categoría contiene ”UNIVERSI" then
    Asignar la categoría ”EST. UNIVERSITARIO”
else if la categoría contiene ”POST-GRADO” o ”MAESTRÍA” o ”DOCTORADO” then
    Asignar la categoría ’MAESTRÍA/DOCTORADO”
else if la categoría contiene ”PROFESIONAL” then
    Asignar la categoría ”PROFESIONAL”
else if la categoría contiene ”TECNÓ” o ”TÉCNI” then
    Asignar la categoría ”TECNOLOGO-TECNICO”
end if
```

Algoritmo 3 Rangos de edades

```
if (Fecha corte - Fecha nacimiento) <18  then
    "0-17"
else if (Fecha corte - Fecha nacimiento) =<30  then
    "18-30"
else if (Fecha corte - Fecha nacimiento) =<40  then
    "31-40"
. . .
else if (Fecha corte - Fecha nacimiento) =<90  then
    "81-90"
else if (Fecha corte - Fecha nacimiento) >91  then
    »91"
end if
```

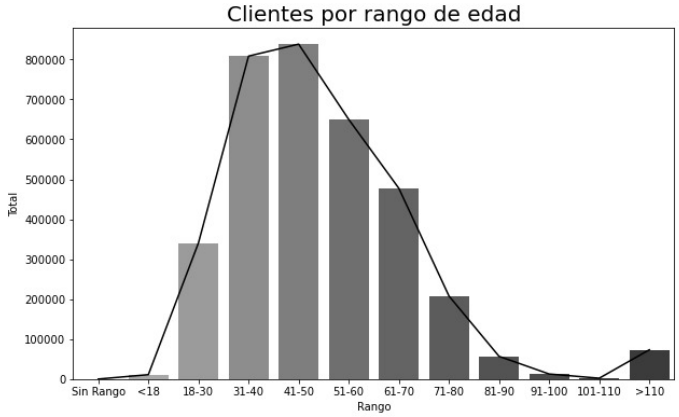


Figura 7: Gráfico de distribución de la variable rango de edad en promedio de los trece meses. En el eje horizontal se encuentran los diferentes rangos de edad y en el eje vertical se muestra la cantidad de datos correspondiente a cada rango.

Validando algunos de los casos de clientes mayores a 110 años, se observó que se debía a errores de digitación al momento de obtener los datos en oficinas u otros lugares donde se tuviese que digitar la información. También se evaluaron los rangos de edad con los que entrenaríamos el modelo, en caso de que la variable de edad fuera determinante para la segmentación. Dado el

conocimiento del problema de negocio, es probable que se limite la edad a los rangos donde se encuentre la mayoría de los clientes y donde se presente menor riesgo, por ejemplo, en la aprobación de préstamos. Durante la construcción de la base de datos, depuramos varias variables e incluso validamos sus categorías, como se hizo para el nivel educativo también se realizo para el estado civil, la actividad económica y el estrato socioeconómico, entre otras variables. En algunas ocasiones, las fuentes de información eran digitadas, lo que podía generar errores de clasificación o codificación.

Algunos de los procesos de limpieza de datos adicionales estuvieron relacionados con variables numéricas que por defecto eran digitadas con el valor de -99999 cuando el valor era nulo o vacío. Para nuestros propósitos, era más conveniente tener estos valores como nulos o vacíos, según correspondiera al caso, por lo que se llevó a cabo la corrección correspondiente para estas variables. Esto aseguró una mayor calidad y precisión en los datos utilizados para el análisis y modelado posterior.

3.2.4. Cálculos de algunas variables

Durante el proceso, se identificó la necesidad de construir algunas variables que no se encontraban disponibles en la base de datos original. Para ello, se realizaron cálculos que involucraban promedios ponderados y tasas de interés, entre otros.

En el caso de las tasas de interés, algunos productos financieros presentaban una tasa efectiva anual, mientras que otros tenían una tasa nominal anual. En el ámbito financiero, las tasas de interés se utilizan para expresar el costo del dinero. La tasa nominal es aquella que se pacta en un contrato, mientras que la tasa efectiva toma en cuenta el efecto de la capitalización (es decir, los intereses generados por los intereses). Por ejemplo, si una entidad bancaria nos presta \$5.000.000 a una tasa efectiva mensual del 2%, cada mes pagaremos un interés de \$100.000. Si no se pagan los intereses el primer mes, los \$100.000 se suman al capital, lo que significa que en el segundo mes se calcularán intereses sobre \$5.100.000, y así sucesivamente. En cambio, si el préstamo es por el mismo valor, pero a una tasa nominal mensual del 2%, los intereses serán siempre los mismos y no se capitalizarán como en el caso anterior. Por lo tanto, la tasa efectiva es por lo general mayor que la tasa nominal. Para garantizar la uniformidad de los datos, fue necesario transformar las tasas efectivas a tasas nominales, lo que se logra mediante la siguiente fórmula:

$$T_N = \left[(1 + T_E)^{\frac{1}{n}} - 1 \right] n,$$

donde T_N es la tasa nominal, T_E es la tasa efectiva y n es la periodicidad. Si se trata de una tasa anual, n será igual a 12 [3, p. 321].

Además de las tasas de interés, también se calcularon otras variables, como tasas de crecimiento, indicadores de rentabilidad y de eficiencia, entre otros. En algunos casos, fue necesario utilizar promedios ponderados para reflejar con mayor precisión la influencia de ciertos productos en el desempeño general de la entidad financiera.

En resumen, la construcción y cálculo de variables permitió obtener información adicional y más precisa sobre el desempeño de los productos financieros en la entidad estudiada.

3.2.5. Tabla resultante

Luego de aplicar los procesos de cruces y limpieza de datos, se obtuvo una tabla maestra con 86 variables. Para automatizar el proceso de extracción desde las diferentes fuentes de datos y hacerlo más eficiente, se creó una macro la cual es una herramienta de programación que permite automatizar tareas repetitivas o complejas, con el fin de obtener los trece meses de historia de las diferentes variables definidas. Aunque no se muestra el código de construcción de la base por temas de seguridad de la información, en el resumen de la tabla maestra se puede observar la cantidad de variables que contiene según su categoría.

Para tener una mejor comprensión de la distribución de las variables en la tabla maestra, se utilizó un gráfico de torta, el cual se presenta en la Figura 8. De acuerdo con la gráfica, se puede notar que las variables relacionadas con el banco, tales como ingresos, segmentos de clientes y puntajes, representan una proporción significativa del total de variables en la tabla. Además, se observa un número considerable de variables relacionadas con las tarjetas de crédito. De hecho, estas variables juntas conforman la mayor proporción en la tabla. Esto sugiere que la información de los clientes y su comportamiento financiero son elementos clave en el análisis y la evaluación del riesgo crediticio.

En general, la tabla maestra resultante del proceso de consolidación y limpieza de datos es una herramienta valiosa para el análisis y la toma de decisiones en el ámbito financiero, ya que contiene información relevante y útil sobre los clientes y su comportamiento financiero.

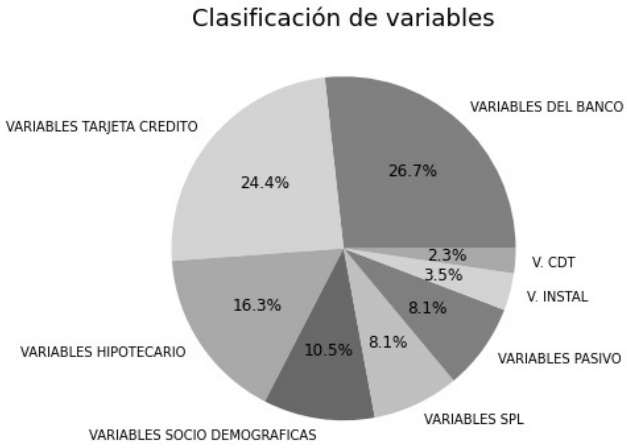
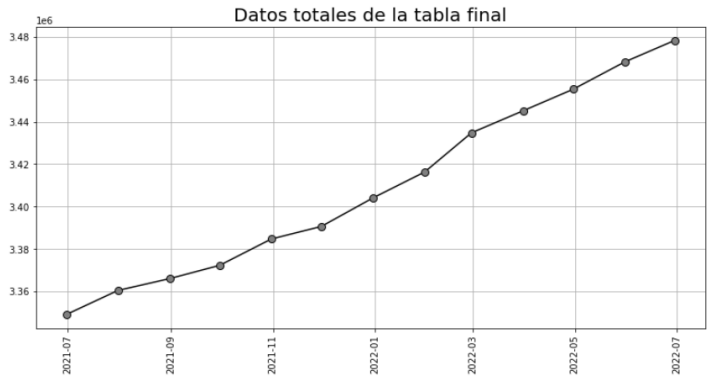


Figura 8: Gráfico de distribución de variables según su categoría en la base de datos. La gráfica de torta muestra la proporción de variables de diferentes categorías en la base de datos, incluyendo tarjetas de crédito, créditos hipotecarios, variables sociodemográficas, entre otras. Cada categoría está representada por un sector en la torta y su tamaño indica la proporción de variables que pertenecen a ella.

La cantidad total de datos es de 44.325.307 distribuidos en un período de 13 meses. Este volumen de información resalta la importancia de contar con herramientas y técnicas adecuadas para el manejo y procesamiento de grandes conjuntos de datos. Además, al estar distribuidos en un periodo de tiempo, se pueden identificar patrones y tendencias en el comportamiento de los datos.

Es importante destacar que la calidad de los datos es crucial para obtener conclusiones precisas a partir del análisis. En la Figura 9, se presenta una visualización del comportamiento de los

datos a lo largo del tiempo, y se puede apreciar un aumento en el número de clientes con el tiempo. En resumen, la cantidad y calidad de los datos en la tabla resalta la importancia de la gestión efectiva de los datos en el análisis y la toma de decisiones en el ámbito financiero. Además, la visualización del comportamiento de los datos a lo largo del tiempo puede proporcionar información valiosa para la evaluación del desempeño del negocio y la identificación de oportunidades de crecimiento.



(a) Cantidad de datos tabla final por mes

CORTE	CANTIDAD DATOS
2021-06-30	3’349.057
2021-07-31	3’360.434
2021-08-31	3’366.008
2021-09-30	3’372.266
2021-10-31	3’384.699
2021-11-30	3’390.623
2021-12-31	3’404.083
2022-01-31	3’416.316
2022-02-28	3’434.751
2022-03-31	3’445.162
2022-04-30	3’455.308
2022-05-31	3’468.224
2022-06-30	3’478.376

(b) Tabla resumen

Figura 9: Resumen tabla final. En la primera subgráfica se representa la cantidad de datos recopilados en la base general por mes, donde en el eje horizontal se encuentra el mes al que hace referencia y en el eje vertical la cantidad de datos. En la segunda subgráfica se muestra una tabla que contiene la cantidad de datos recopilados en la base general por mes.

4. Análisis exploratorio de resultados

En esta sección, se llevará a cabo un análisis exploratorio de los resultados obtenidos en la base de datos. Para ello, se revisarán aspectos como los valores faltantes, las distribuciones de las variables y los valores mínimos y máximos. El análisis de estos aspectos es fundamental para comprender la calidad de la información y entender el comportamiento de los datos, lo que permitirá tomar decisiones informadas en cuanto al análisis posterior y el desarrollo de modelos predictivos.

Los valores faltantes son aquellos que no están presentes en la base de datos para una variable en particular. La presencia de valores faltantes puede deberse a varios motivos, como errores en la recolección de datos, problemas técnicos en la base de datos o la ausencia de ciertos datos en la fuente original de información. La presencia de valores faltantes puede afectar la calidad del análisis de datos y la capacidad de los modelos para hacer predicciones precisas. Resulta importante validar valores ausentes, dado que entender su naturaleza permitirá saber si la ausencia es aleatoria o no aleatoria, si tiene una razón o no. La ausencia de datos puede generar molestias posteriormente en el desarrollo de modelos dado que algunos no permiten el entrenamiento con datos faltantes y se deben tomar decisiones sobre ellos, bien sea imputarlos o excluirlos.

Las distribuciones de las variables son patrones que describen cómo se distribuyen los valores de una variable en una muestra. En el análisis exploratorio de datos, es importante revisar las

distribuciones de las variables para identificar patrones y características de los datos, como la presencia de valores extremos, la presencia de valores atípicos, la asimetría de la distribución, entre otros. Estos patrones pueden ser relevantes para el análisis posterior y el desarrollo de modelos predictivos.

Los valores mínimos y máximos de una variable son los valores más pequeños y más grandes de dicha variable en una muestra. Revisar los valores mínimos y máximos de las variables es importante para detectar valores extremos que puedan afectar la calidad de los datos o el análisis posterior.

En el informe se presentarán los resultados obtenidos para algunas de las variables analizadas en cuanto a sus valores faltantes, distribuciones y valores mínimos y máximos.

4.1. Valores faltantes

De manera general, se observa que para cada variable, el porcentaje de valores ausentes se mantiene “constante"sin una variabilidad importante en el tiempo. Con la excepción de la variable DIGITAL_STATUS, que hace referencia a si un cliente está inscrito o no en plataformas digitales del banco. Para el periodo de mayo de 2022, todo queda sin marcar, y es por ello que se observa un pico del 100% de valores faltantes. Este caso fue una alerta para el proceso desarrollado, y se debía validar si era un error en la consulta construida o si se debía a un error en los datos de origen que alimentaban el proceso. Durante dicha validación, se observó que para este mes, efectivamente no se contaba con la información de las fuentes originales debido a un problema con las bases del banco. En la Figura 10 observamos algunas de las variables y sus valores ausentes.

Para la variable SEGMENT_DESC, el porcentaje de valores faltantes es muy bajo, en promedio del 0.01%, y con una variabilidad entre el 0% y el 0.02%. La ausencia del segmento está asociada a clientes que no están activos. Para la variable MARITAL_STATUS, el promedio del porcentaje de valores faltantes es del 26.90% en un rango entre el 26.80% y el 27.05%. Dado que es información que viene desde CRM y al ser digitada por el personal puede presentar inconsistencias e incluso no almacenarse la información, como se mencionó antes para la variable de rangos de edad del cliente.

En el caso de las variables NO_TRX_CC y NO_TRX_DC, las cuales hacen referencia al número de transacciones de compras y avances con tarjeta de crédito y al número de transacciones con tarjeta de débito en el periodo, se presentan valores faltantes del 82.23% para las transacciones de tarjeta de crédito, con una variabilidad entre el 81.5% y el 83.5%, tendiendo a disminuir. Para las transacciones de tarjeta de débito, tenemos un promedio del 81.2%, con una variabilidad entre el 80.6% y el 81.9%. A simple vista, parece una cantidad alta de valores faltantes, pero es importante tener en cuenta que la base contiene información de clientes con distintos productos, por lo que no todos los clientes tienen una tarjeta de crédito o débito con el banco, y mucho menos la utilizan. En este caso, los clientes transaccionales con tarjeta de crédito y débito representan aproximadamente el 17% y el 18%, respectivamente, de la población total de la base por mes.

En cuanto a la variable DIGITAL_STATUS, como se mencionó anteriormente, se encontraron inconsistencias en los datos obtenidos de la fuente en mayo de 2022. Para mitigar esta situación,

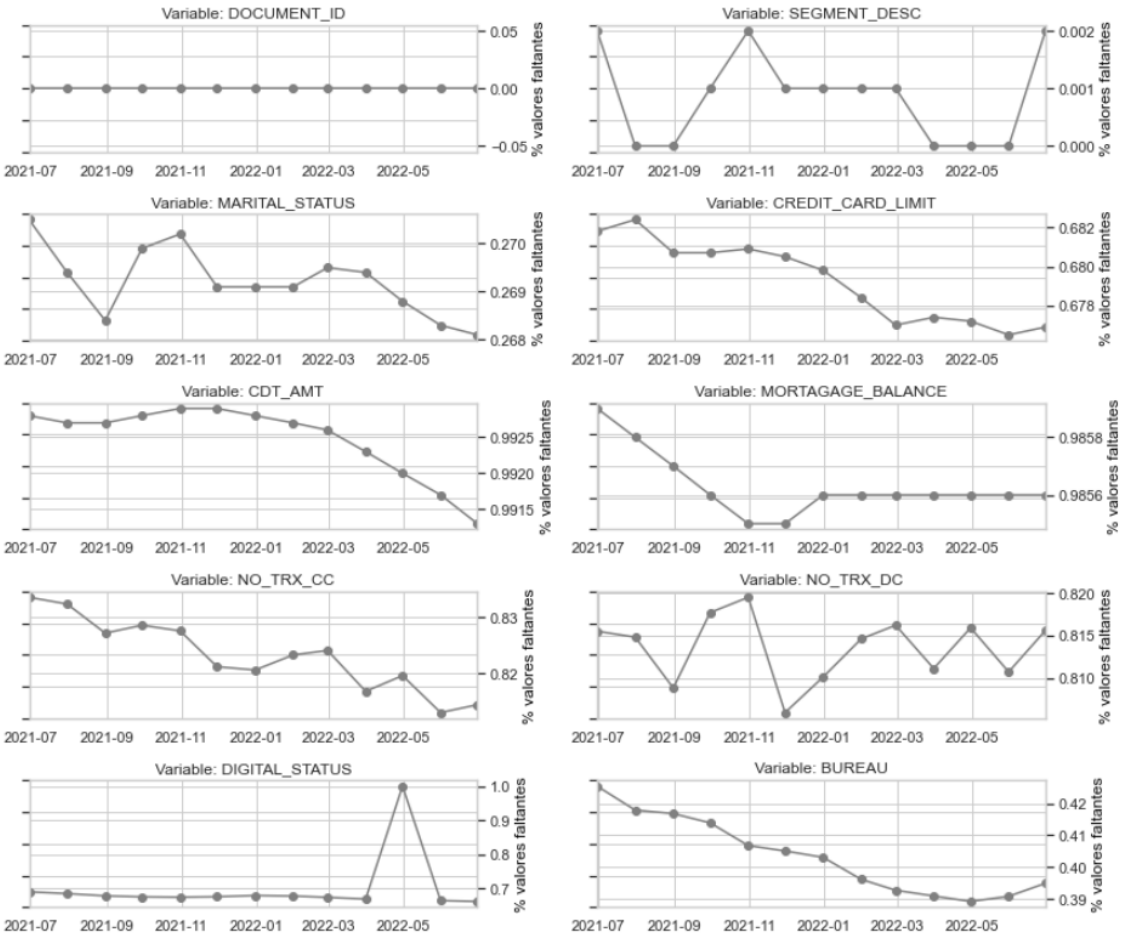


Figura 10: Comportamiento de los valores faltantes de las variables en la base general. El gráfico muestra la evolución mensual del porcentaje de valores faltantes para cada una de las 10 variables seleccionadas. Cada subgráfica representa una variable y en el eje vertical se encuentra el porcentaje de valores faltantes, mientras que en el eje horizontal se muestra el mes correspondiente

se decidió utilizar el valor del mes anterior para los clientes que estuvieron activos tanto en el mes de abril como en el de mayo. Esta decisión se tomó considerando que no se observó un comportamiento rápido en los cambios del estado de la variable DIGITAL_STATUS de un mes a otro. Por otro lado, en el caso de la variable CDT_AMT se presentan valores faltantes cercanos al 100 %, con un rango de entre el 99.13 % y el 99.29 %. Sin embargo, gráficamente se puede observar que la cantidad de valores ausentes disminuye con el tiempo. Esto se debe a que, aunque el producto CDT no supera el 1% de los productos ofrecidos por el banco, ha captado más clientes con el tiempo.

Es fundamental enfatizar que, aunque las variables CDT_AMT y MORTGAGE_BALANCE presentan valores faltantes cercanos al 100 %, no debemos subestimar su importancia en nuestro análisis. Estas variables representan la presencia o ausencia de productos financieros específicos en la cartera de nuestros clientes.

A pesar de que estas variables constituyen una pequeña fracción de los productos ofrecidos por el banco y, en consecuencia, solo un reducido porcentaje de los clientes las posee, es esencial

reconocer que su adquisición ha ido en aumento con el tiempo. Esta tendencia se refleja en la disminución gradual de los valores ausentes a lo largo del período de estudio, lo cual indica un interés en constante crecimiento por parte de los clientes en estos productos financieros.

Por lo tanto, no debemos malinterpretar los valores faltantes como una debilidad en los datos construidos, sino más bien como una indicación de que no todos los clientes han adquirido estos productos de manera generalizada. Hace parte de un análisis futuro, explorar detalladamente las características y el comportamiento de los clientes que han optado por estos productos.

4.2. Frecuencias

Durante el proceso de análisis de los datos, es importante conocer las frecuencias de las variables para entender el comportamiento de los productos y los clientes. En esta sección, revisaremos algunas de las variables desarrolladas en la base maestra. Comenzamos con la variable que indica si un cliente está activo o no dentro del banco.

La variable ACTIVE_CUSTOMER es de tipo categórica y tiene dos posibles valores: NO, que indica que el cliente no es activo, y YES, que indica que el cliente es activo. Según la Figura 11, el 53% de los clientes son activos, mientras que el 47% son inactivos. Este comportamiento se mantiene constante a lo largo de los trece meses.

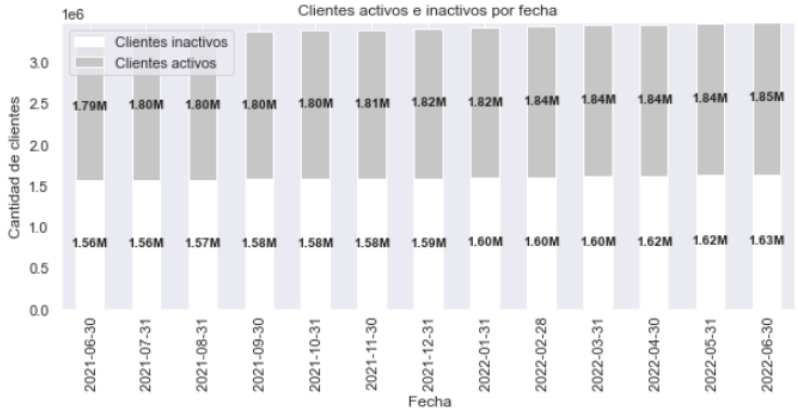


Figura 11: Gráfico de barras apilado muestra la cantidad de clientes activos e inactivos en cada mes. El eje vertical representa la cantidad de clientes, mientras que el eje horizontal representa los meses. Las barras de color claro representan la cantidad de clientes inactivos, mientras que las barras superpuestas de color oscuro representan la cantidad de clientes activos.

Ahora, consideremos la variable CDT_FLG, que también es categórica y tiene dos posibles valores: 1, que indica que el cliente tiene un producto CDT, y 0, que indica que el cliente no tiene este producto en el banco. Como se muestra en la Figura 12, la cantidad de clientes con producto CDT es muy cercana al 0%, lo cual va de la mano con el análisis previo de la variable CDT_AMT, que contiene la información del valor del CDT del cliente. Es decir, una cantidad considerablemente baja de clientes tienen este producto.

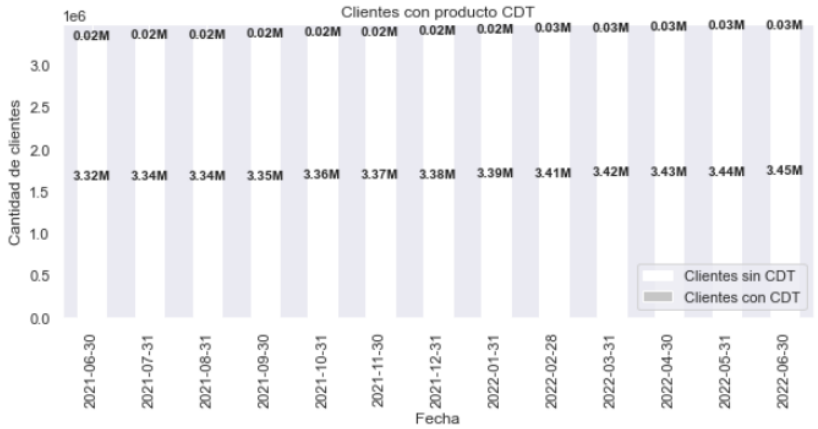


Figura 12: Gráfico de barras apilado muestra la cantidad de clientes con y sin producto CDT en cada mes. El eje vertical representa la cantidad de clientes, mientras que el eje horizontal representa los meses. Las barras de color claro representan la cantidad de clientes sin CDT, mientras que las barras superpuestas de color oscuro representan la cantidad de clientes con CDT.

En la Figura 13, tenemos el comportamiento de la variable CREDIT_CARD_FLG, la cual nos indica cuando un cliente tiene una tarjeta crédito. Observemos que aproximadamente un 34% de los clientes tienen este producto de los cuales por el análisis anterior en cuanto a su transaccionalidad con el producto podemos inferir que tan solo 50% de clientes que tienen tarjeta realizan transacciones con ella.



Figura 13: Gráfico de barras apilado muestra la cantidad de clientes con y sin tarjeta de crédito en cada mes. El eje vertical representa la cantidad de clientes, mientras que el eje horizontal representa los meses. Las barras más oscuras representan la cantidad de clientes con tarjeta de crédito, mientras que las barras de color claro representan la cantidad de clientes sin tarjeta de crédito.

En la Figura 14 se presentan 7 conjuntos de barras, cada uno asociado a una variable, que muestran los porcentajes promedio de cada categoría. La primera barra representa el porcentaje promedio de clientes que cumplen con la categoría, la segunda barra muestra el porcentaje promedio de casos que no cumplen con la categoría, y en la última barra se presenta el porcentaje promedio de valores vacíos.

En cuanto a las variables específicas, la variable MULTIPRODUCT_FLG, que indica si un cliente tiene más de un producto activo en el banco, muestra que, en promedio, el 16% de los clientes tiene más de un producto. La variable SPL_FLG, que indica si un cliente tiene al menos un crédito de libranza, muestra que, en promedio, el 7% de los clientes cuentan con este producto.

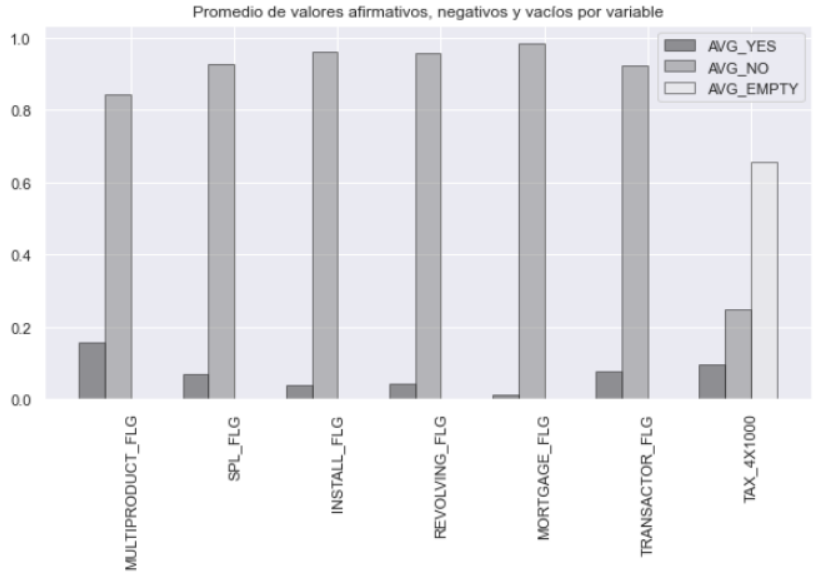


Figura 14: Gráfico de barras con el eje vertical representando los porcentajes promedio y el eje horizontal representando las 7 variables. Cada conjunto de barras tiene tres barras: una en gris oscuro que representa el porcentaje promedio de clientes que tienen el producto asociado a la variable, una en gris tono medio que representa el porcentaje promedio de clientes que no tienen el producto, y una en gris claro que representa el porcentaje promedio de valores vacíos para cada variable.

La variable TRANSACTOR_FLG, una variable construida, indica que, en promedio, alrededor del 8% de los clientes difiere más del 80% de sus compras a una cuota. Por último, la variable TAX_4X1000, que indica si un cliente tiene el producto de tarjeta débito marcado como exento del 4X1000, muestra que aproximadamente el 9% de los productos se encuentran exentos, alrededor del 23% está marcado como no exento y el resto no tiene información referente a la variable.

Todo lo anterior se hizo para validar la consistencia de las variables de tipo categórico. En el caso de las variables numéricas como BUREAU, que indica el score crediticio del cliente, se revisó su distribución para asegurarse de que los datos fueran acordes al conocimiento que se tiene de la variable, en la Figura 15, tenemos el comportamiento de la variable, en la cual el valor mínimo es cero, el promedio aproximadamente 700 puntos y valor máximo 999. Lo cual va acorde a la información que asigna esta escala.

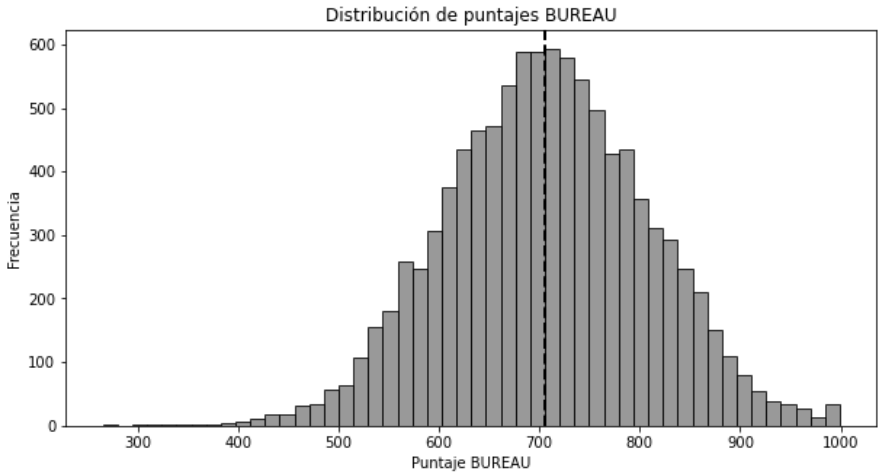


Figura 15: Distribución de la variable BUREAU en función de la cantidad de datos. El eje vertical representa la cantidad de datos, mientras que el eje horizontal representa los puntajes Bureau.

4.3. Valores mínimos y máximos

Observar el comportamiento de las variables numéricas en cuanto a los valores máximos y mínimos aporta una visión del comportamiento general de los datos. En la variable INCOME_RISK, la cual hace referencia a los ingresos del cliente, resulta llamativo que existan valores extremos muy grandes en esta variable, llegando a valores de 10^{14}. Estos valores atípicos nos generaron inicialmente sospechas de posibles errores en la fuente de datos o inconsistencias en su registro. Sin embargo, tras un análisis exhaustivo, no se logró determinar que fueran una inconsistencia en la fuente de datos en sí.

La presencia de estos valores extremos podría estar influyendo en el cálculo del valor promedio de esta variable, que es de aproximadamente 10^{10}. Es importante tener en cuenta que los valores extremos pueden afectar significativamente las medidas de tendencia central, como el promedio, y en consecuencia, es necesario prestar atención a su presencia y determinar su veracidad antes de realizar cualquier análisis posterior. Aunque estos datos nos parecieron atípicos en cuanto al valor extremo del máximo, no se encontraron indicios concluyentes de que fueran errores o inconsistencias en la fuente de datos. Por tanto, se decidió mantenerlos en el análisis y considerar su posible influencia en las conclusiones obtenidas.

5. Herramientas y Tecnologías Utilizadas

Durante el desarrollo de la pasantía, se emplearon diversas herramientas y tecnologías para llevar a cabo el procesamiento y análisis de los datos obtenidos de diferentes fuentes, así como la revisión de la base de datos generada. A continuación, se detallan las principales herramientas utilizadas y su función en el proyecto:

- **Python [1]**
    - Se utilizó el lenguaje de programación Python junto con las librerías Pandas, Matplotlib y Seaborn.

- Con Pandas, se realizó la revisión de la base de datos generada a partir de la consolidación de las diversas fuentes. Se exploraron aspectos como el comportamiento de los valores faltantes, las frecuencias de las variables, los valores mínimos y máximos, y la revisión de los datos en el tiempo [7].
- Matplotlib y Seaborn se emplearon para generar visualizaciones gráficas y facilitar la comprensión de los patrones y tendencias presentes en la información [6], [12].

- **SAS® Guide**

  - El software SAS® Guide fue utilizado para llevar a cabo el proceso de Extracción, Transformación y Carga (ETL) de los datos [2].
  - se realizaron tareas de limpieza, transformación y consolidación de los datos provenientes de diferentes fuentes. Esto incluyó la eliminación de registros duplicados, corrección de errores en la información recolectada y la combinación de datos para crear una base general.
  - Además, se implementaron flujos de trabajo utilizando SAS® Guide, siguiendo una guía de referencia para garantizar la correcta implementación del código y optimizar las consultas realizadas [8], [4].

6. Evaluación y cumplimiento de objetivos

Objetivo	**Tareas**	**¿Realizado?**
1	Identificación de fuentes necesarias.	Sí
1	Mapeo de fuentes.	Sí
1	Generación del código de extracción.	Sí
2	Identificación de posibles causas.	Sí
2	Establecer criterios de detección.	Sí
2	Desarrollo de estrategia para evitar duplicidad.	Sí
3	Verificación de consistencia de datos.	Sí
3	Estructuración de columnas.	Sí
3	Estandarización de nombres de las columnas.	Sí
3	Estandarización de texto para homologar los valores categóricos.	Sí
3	Revisión de tipos de datos.	Sí
4	Definir los indicadores a revisar.	Sí
4	Definir la frecuencia de revisión.	Sí
4	Establecer un proceso de análisis de tendencias.	Sí
5	Identificar columnas con valores faltantes.	Sí
5	Identificar valores atípicos.	Sí
5	Corrección de inconsistencias en los casos que aplique.	Sí
5	Documentación de cambios.	No
5	Verificar integridad de los datos posterior a los cambios.	Sí

7. Conclusiones

Durante la pasantía, se logró el objetivo de crear una base de datos unificada que reúne información sociodemográfica, transaccional, de segmento y de riesgo de los clientes de la empresa durante un periodo de trece meses. Esta base de datos será entregada como insumo para que otra área pueda generar un modelo de segmentación de clientes con el fin de otorgar tasas preferenciales. La creación de esta base de datos fue un proceso exhaustivo que involucró la recolección, limpieza y unificación de datos de diversas fuentes y formatos. A continuación, se presentan las principales conclusiones obtenidas durante la pasantía y los aprendizajes adquiridos en el proceso.

- La duplicidad de valores en una base de datos puede ser un problema que afecta la calidad de la información y la toma de decisiones. Es importante evaluar cada caso para determinar si la duplicidad es un problema y cómo abordarlo. Se utilizaron algoritmos y técnicas para validar la duplicidad de datos, y en algunos casos se realizó agrupación de los datos para eliminar duplicados y tener una vista más clara y precisa de la información.
- La limpieza de datos es una tarea fundamental en cualquier proyecto de datos para garantizar la calidad de la información.
- La visualización del comportamiento de los datos a lo largo del tiempo permite identificar patrones y tendencias en el comportamiento de los datos, lo que es útil para la toma de decisiones en el ámbito financiero.
- La calidad de los datos es crucial para obtener conclusiones precisas a partir del análisis. Por lo tanto, es importante contar con una gestión efectiva de los datos en el análisis y la toma de decisiones en el ámbito financiero.
- Comprender el problema de negocio que se desea abordar y entender los requerimientos del proyecto son pasos fundamentales para asegurar el éxito. Al tener una comprensión clara de lo que se necesita lograr, se pueden enfocar los recursos de manera efectiva y evitar desperdiciar tiempo y esfuerzos en tareas innecesarias o mal dirigidas.

Referencias

[1] Python 3.9.7. Software, 2021.

[2] SAS Guide 8.3. Software, 2021.

[3] Eugene F. Brigham and Joel F. Houston. *Fundamentals of Financial Management.* South-Western College Pub, 2003.

[4] Ron Cody. *Learning SAS® by Example: A Programmer's Guide.* SAS Institute Inc., 2007.

[5] Ramez Elmasri and Shamkant B. Navathe. *Fundamentos de Sistemas de Bases de Datos.* Informática. Pearson Educación S.A, 2007.

[6] John D. Hunter et al. Matplotlib. Software, 2021.

[7] Wes McKinney et al. Pandas. Software, 2021.

[8] Derek P. Morgan. *The Essential Guide to SAS® Dates and Times.* SAS Institute Inc., 2006.

[9] Bojan Petrovic. Introducción y resumen de la cláusula de “sql join”, 2019.

[10] Shazia Sadiq. *Handbook of Data Quality*. Springer, 2013.

[11] Abraham Silberschatz, Henry F. Korth, and S. Sudarshan. *Fundamentos de Bases de Datos*. McGraw-Hill, 2002.

[12] Michael Waskom et al. Seaborn. Software, 2021.