Contenido

PARTE 1 DEL PROYECTO-> Transformación y limpieza de datos 1
PARTE 2 Del Proyecto-> Análisis descriptivo a través de la visualización de datos 2
1. Distribución de la edad (nombre de la variable age) 2
2. Duración de llamadas (duration) 3
VISUALIZACIÓN DE DATOS E INTERPRETACIÓN 5
CONCLUSIÓN 14

ANÁLISIS DEL PROYECTO

El objetivo de este proyecto es realizar un Análisis Exploratorio de Datos (EDA) a partir de dos conjuntos de datos relacionados con campañas de marketing directo de una institución bancaria portuguesa. Los datos incluyen información sobre clientes, descripciones socioeconómicas y detalles de campañas telefónicas realizadas entre 2018 y 2020.

PARTE 1 DEL PROYECTO-> Transformación y limpieza de datos

El primer paso es la carga de archivos en el fichero

- Se cargó el archivo bank-additional.csv mediante read_csv().
- Se cargó el archivo customer-details.xlsx leyendo sus tres hojas, correspondientes a los años 2018, 2019 y 2020.
- Posteriormente, se concatenaron las tres hojas en un único dataframe de clientes.

El segundo paso fue revisar:

- Tipos de datos (dtypes)
- Dimensiones (shape)
- Valores nulos (isnull().sum())
- Valores únicos en variables categóricas

Esto permitió detectar:

- Columnas numéricas almacenadas como cadenas (strings)
- Columnas con valores faltantes esporádicos
- Columnas como date, contact_month o contact_year susceptibles de transformaciones de tipo fecha

Limpieza aplicada

- Conversión de tipos erróneos (dates, enteros y floats cuando correspondía).
- Eliminación de espacios en variables categóricas.
- Imputación / revisión de valores nulos según su impacto.
- Unión final entre la tabla de clientes y la tabla de campaña utilizando el identificador id_.

Estas transformaciones generaron una versión depurada del dataset, lista para análisis descriptivo.

PARTE 2 Del Proyecto-> Análisis descriptivo a través de la visualización de datos

El propósito principal es comprender el comportamiento y características de los clientes, identificar patrones relevantes y detectar posibles problemas de calidad en los datos para su posterior uso en modelización o toma de decisiones.

Como parte de la fase de limpieza y exploración inicial de los datos, generamos diagramas de caja (*boxplots*) para identificar valores atípicos (*outliers*) en dos variables numéricas relevantes del dataset:

- Edad del cliente
- Duración de la llamada en segundos

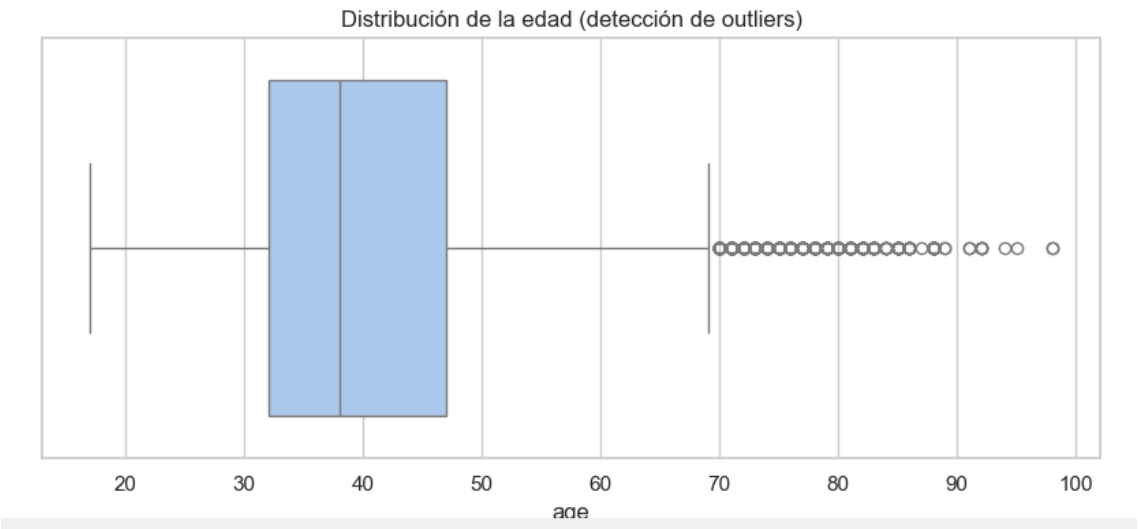
Estos gráficos permiten visualizar la dispersión de los datos, los cuartiles y posibles observaciones extremas.

1. Distribución de la edad (nombre de la variable age)

Qué hicimos en el código

Generamos un boxplot de la variable age para observar su rango, mediana y posibles outliers:

- Se utilizó **seaborn.boxplot()**
- Se ajustaron etiquetas y título
- La gráfica se usó como herramienta de detección de valores extremos



Qué representa esta gráfica

Un boxplot muestra:

- La mediana(línea dentro de la caja)
- Los quartiles Q1 y Q3 (límites de la caja)
- El rango normal de los datos (*bigotes*)
- Los outlier representados como puntos aislados más allá de 1.5×IQR

Resultados observados

En la gráfica podemos ver:

- La mayoría de las edades se concentran entre 30 y 50 años.
- La mediana está alrededor de 40 años, lo que indica que la mitad de los clientes tiene menos de esa edad.
- Existe un grupo de valores atípicos por encima de 70 años, llegando incluso a casos cercanos a 100 años.
- No se observan outliers significativos por la parte baja (edades extremadamente jóvenes).

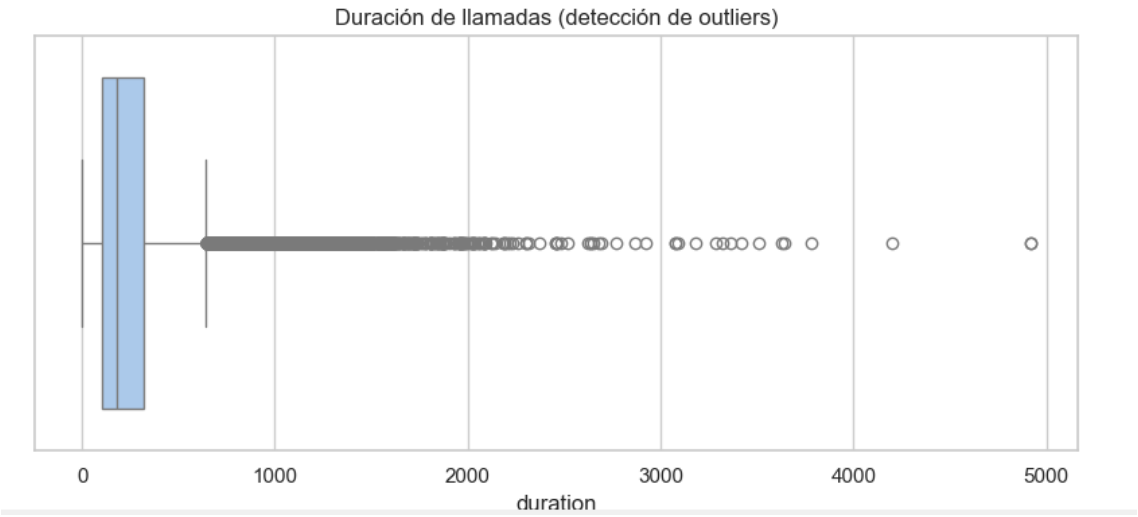
Interpretación

Los datos muestran una población adulta con una distribución relativamente concentrada, pero con algunos clientes de edad muy avanzada que deberán evaluarse según su impacto en el análisis.
Estos outliers pueden ser reales y no errores, pero conviene analizarlos antes de construir modelos predictivos.

2. Duración de llamadas (duration)

Qué hicimos en el código

De igual forma, generamos un boxplot para la variable duration, que representa el tiempo en segundos de cada llamada realizada durante la campaña de marketing.



Qué representa esta gráfica

El boxplot permite visualizar:

- La distribución de las duraciones típicas
- La presencia de llamadas extremadamente largas
- Posibles valores atípicos que puedan influir en el análisis

Resultados observados

En esta gráfica se observa:

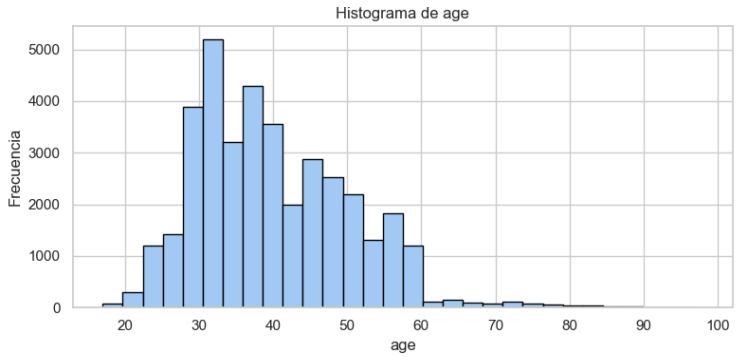
- La mayor parte de las llamadas dura menos de 300 segundos (5 minutos).
- Sin embargo, hay una cantidad muy significativa de outliers:
    - Llamadas que superan los 1.000 segundos
    - Otras que alcanzan más de 2.000 o incluso 4.000 segundos (más de una hora)
- La caja de datos es muy estrecha en comparación con el rango total, lo que indica una distribución fuertemente sesgada hacia valores bajos.

VISUALIZACIÓN DE DATOS E INTERPRETACIÓN

Una vez analizado los outliers, se generaron gráficos más fáciles de leer y entender como histogramas y gráficos circulares para a ver un análisis de las variables más destacable:

Edad (age)

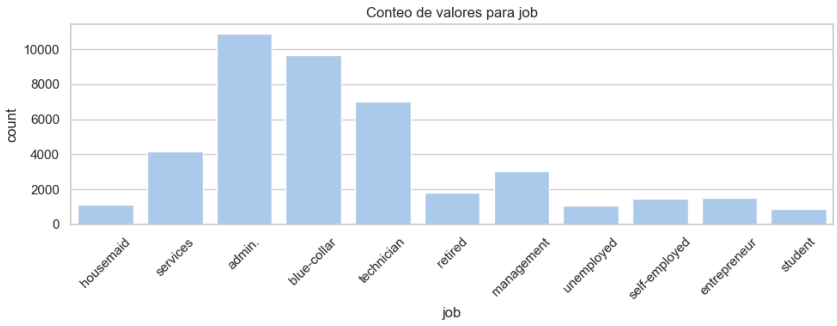
El boxplot muestra que la mayoría de clientes se encuentran entre **30 y 50 años**, con una mediana cercana a **40**.
Aparecen algunos outliers por encima de **70 años**, pero no parecen errores de registro.



Puesto de trabajo (job)
El gráfico de barras muestra la distribución de los diferentes tipos de empleo entre los clientes.
Las categorías más frecuentes son admin. y blue-collar, que representan la mayoría de la base de datos.
También destacan los grupos technician y services, con una presencia considerable.

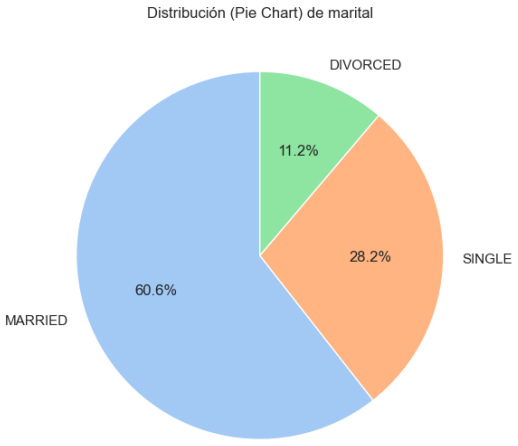
Por otro lado, se observa una menor representación en categorías como student, entrepreneur, self-employed y unemployed, lo que indica que estos perfiles son menos

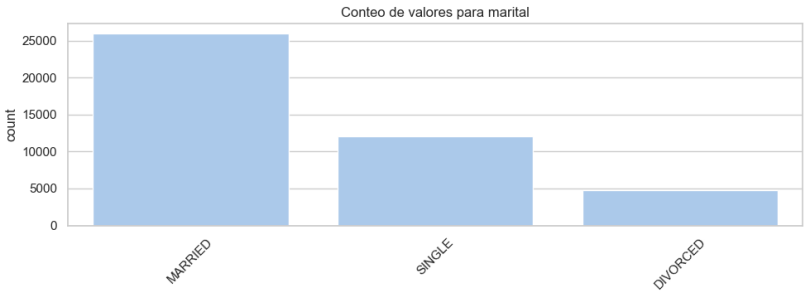
comunes dentro del dataset.
La distribución sugiere que el banco trabaja principalmente con clientes de empleos administrativos, técnicos y trabajados manuales.



Estado civil (Marital)
La categoría dominante y mayoritaria es MARRIED (Casado), que representa el 60.6% de toda la base de datos. En términos de conteo, esta categoría supera los 25.000 clientes.La segunda categoría más grande es SINGLE (Soltero), que abarca el 28.2% del total, con un conteo aproximado de 12.000 clientes.La categoría con menor representación es DIVORCED (Divorciado), que constituye el 11.2% de la clientela, con un conteo de aproximadamente 5.000 clientes.

La distribución es fuertemente asimétrica y está dominada por el segmento de clientes casados.Esto implica que el banco trabaja principalmente con clientes que se encuentran en el estado civil de casados, lo que puede tener implicaciones para productos financieros conjuntos, préstamos hipotecarios y estrategias de riesgo familiar. Los clientes solteros y divorciados constituyen segmentos secundarios, pero importantes, de la clientela.





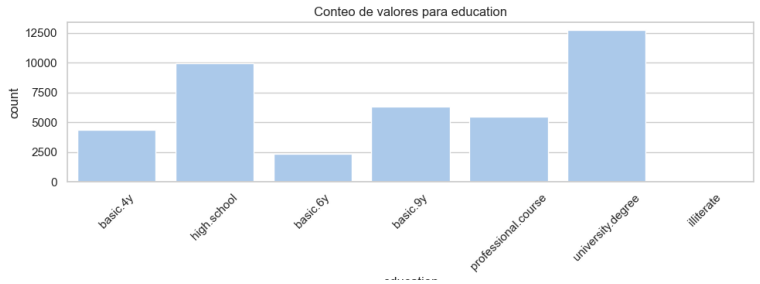
Nivel educacional del cliente (education)

Las categorías más frecuentes son university.degree y high.school. Estas representan la mayoría de la base de datos, con el nivel universitario siendo el más predominante (cerca de 12.500 clientes) y la secundaria completa como el segundo grupo más grande (cerca de 10.000 clientes).

También destacan los grupos con educación básica más avanzada (basic.9y, alrededor de 6.000) y aquellos que tomaron un professional.course (alrededor de 5.500), con una presencia considerable.

Por otro lado, se observa una menor representación en categorías de educación básica de menor duración (basic.6y y basic.4y). La categoría illiterate (analfabetos) tiene una presencia casi nula, lo que indica que estos perfiles son muy poco comunes dentro del dataset.

La distribución general sugiere que el banco trabaja principalmente con clientes que poseen un nivel educativo medio-alto, centrado en personas con títulos universitarios y educación secundaria completa



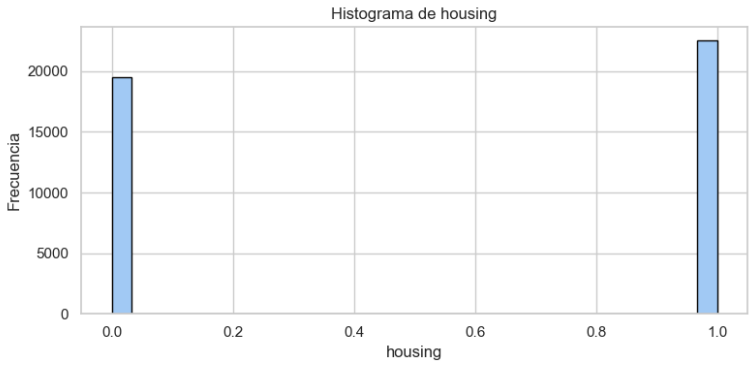
Préstamo hipotecario (Holding)

Se observa que la frecuencia de clientes que sí tienen un préstamo hipotecario (valor 1) es la más alta, alcanzando aproximadamente los 22.000 clientes.

La categoría de clientes que no tienen un préstamo hipotecario (valor 0) también es muy significativa, con una frecuencia cercana a los 19.500 clientes.

Esto implica que la mayoría de los clientes de la base de datos ya posee un préstamo hipotecario, lo que sugiere una alta penetración de este producto o una clientela ya consolidada.

A pesar del predominio de la categoría "Sí", la distribución es bastante equilibrada. El gran volumen de clientes en la categoría "No" indica que aún existe un segmento sustancial de la clientela que podría ser objetivo para futuros ofrecimientos de préstamos hipotecarios.



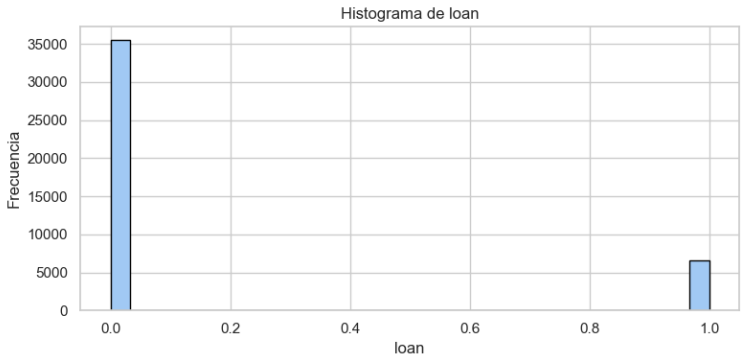
Otro tipo de préstamos (loan)

Se observa una marcada asimetría en la distribución de este producto, a diferencia del préstamo hipotecario.

La categoría dominante y más frecuente es 0 (No), que representa a los clientes sin otro préstamo, con una frecuencia muy alta que supera los 35.000 clientes.

La categoría 1 (Sí), que representa a los clientes con este tipo de préstamo, es la minoría, con un conteo que se sitúa alrededor de los 6.000 clientes.

Esto implica que la gran mayoría de la base de datos no posee un préstamo personal o de consumo, lo que sugiere una baja penetración de este producto en comparación con el préstamo hipotecario. Este segmento de clientes sin el producto (valor 0) podría representar un objetivo principal para campañas de venta de préstamos personales.

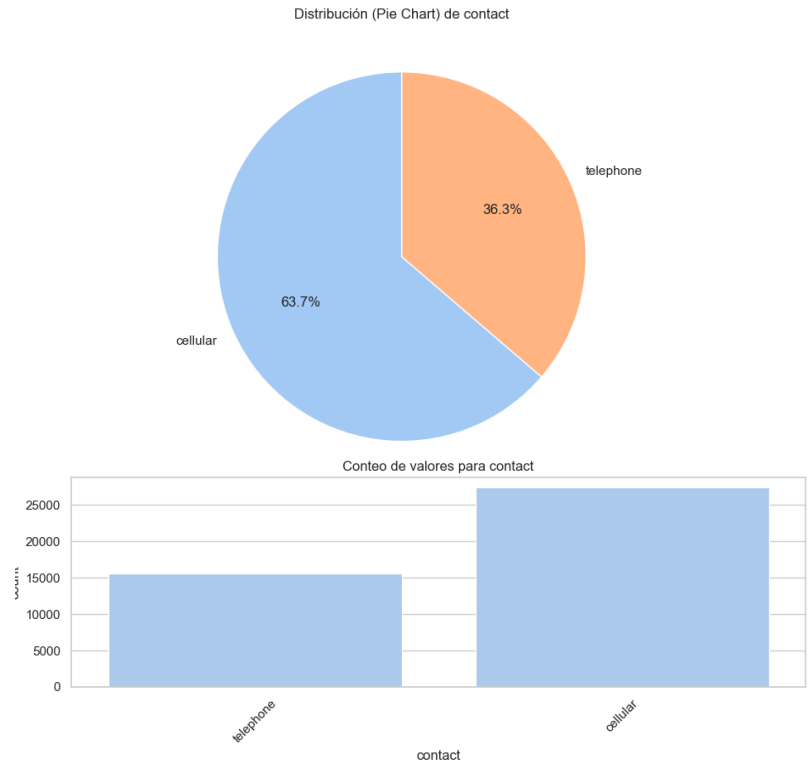


COMO CONTACTARON (CONTACT)

La categoría dominante es el contacto a través de cellular (teléfono móvil), que representa la mayoría de la base de datos con un 63.7% del total de interacciones. En términos de conteo, esta categoría alcanza aproximadamente los 26.000 clientes.

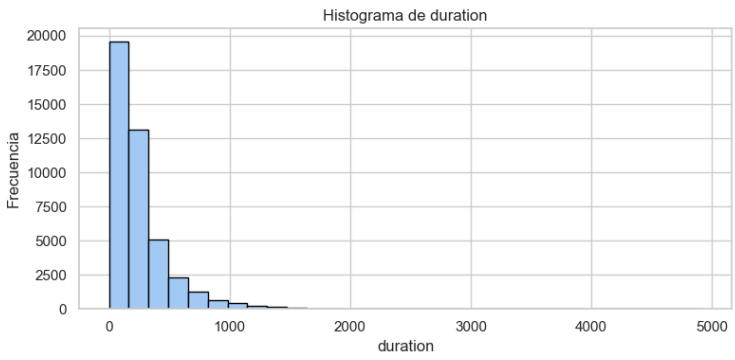
Por otro lado, el contacto vía telephone (teléfono fijo) es la minoría, representando el 36.3% del total, con un conteo aproximado de 15.500 clientes.

La distribución es claramente asimétrica. Esto implica que la estrategia de comunicación del banco se centra predominantemente en el uso del teléfono móvil. Este enfoque sugiere que el banco prioriza los métodos de contacto modernos y directos, probablemente para lograr mayor inmediatez y eficacia en la comunicación con su clientela.



Duración de llamadas (duration)

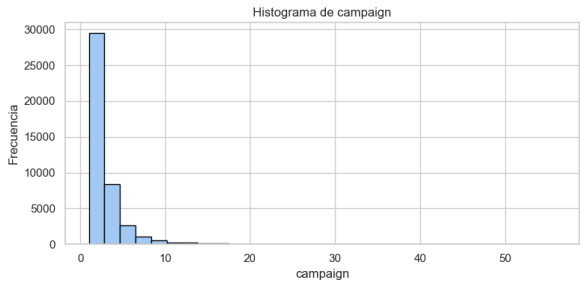
La mayor parte de las llamadas dura menos de 300 segundos, pero se detecta una gran cantidad de outliers con duraciones de hasta 4.000 segundos.
Esta variable está muy sesgada y contiene muchos valores extremos, por lo que debe manejarse con precaución en análisis posteriores.



El número de contactos realizados durante esta campaña para este cliente.(Campaign)

La distribución presenta una asimetría extrema hacia la derecha (sesgo positivo). Esto indica que la inmensa mayoría de los clientes fue contactada muy pocas veces. La frecuencia más alta se concentra en el primer bin (probablemente 1 o 2 contactos), que alcanza cerca de 30.000 clientes, siendo el modo de la distribución. La frecuencia de contactos cae drásticamente a partir del tercer contacto y continúa decreciendo

rápidamente. El número de clientes contactados más de 10 veces es marginal. Los valores en el extremo derecho de la cola (más de 20 contactos) son casi inexistentes.

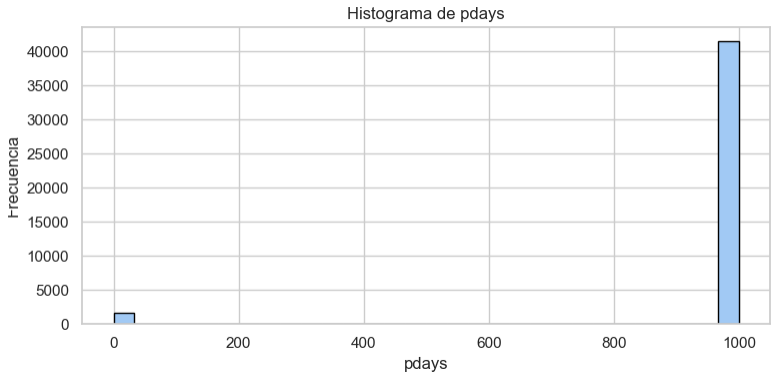


Numero de días que han pasado desde la última vez que se contacto (pdays)

El histograma de pdays muestra el número de días transcurridos desde que se contactó al cliente por última vez en una campaña previa (el valor 999 se utiliza para indicar que no hay un contacto anterior registrado).

La distribución es extremadamente asimétrica y se divide en dos grupos muy desiguales. El grupo dominante se encuentra en la barra extrema derecha, con una frecuencia de más de 41.000 clientes. Esto indica que la abrumadora mayoría de los clientes de esta base de datos nunca fue contactada en las campañas de marketing anteriores. Se observa un grupo muy pequeño de clientes en el extremo izquierdo (cercano a 0), con una frecuencia de apenas 1.000, lo que representa a aquellos que sí tienen un historial reciente de contacto en campañas pasadas.

Esta distribución sugiere que la campaña actual está enfocada principalmente en una base de clientes "nueva" para el esfuerzo de marketing, lo que implica que la mayoría de los contactos realizados en esta campaña son primeros contactos en el historial reciente del cliente.



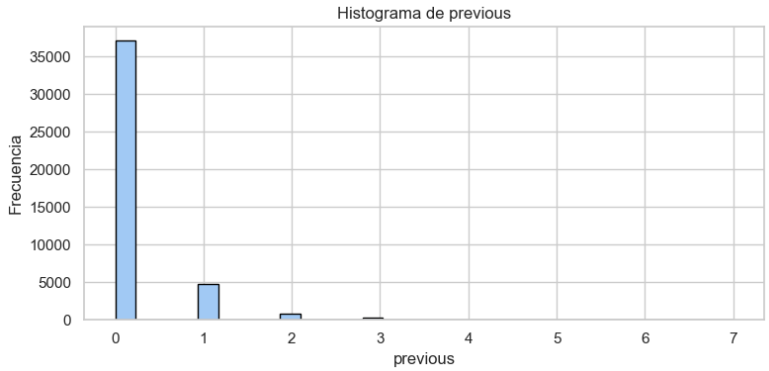
Número de veces que se ha contactado con el cliente antes de esta campaña (Previous)

El histograma muestra la distribución del número de contactos realizados al cliente antes de la campaña actual (variable previous).

La distribución es altamente asimétrica y está fuertemente concentrada en valores bajos, lo que indica que la gran mayoría de los clientes tiene muy pocos contactos previos.

- La categoría dominante y más frecuente es 0 contactos previos, con una frecuencia que supera los 36.000 clientes. Esto significa que más del 85% de la base de datos no tiene ningún registro de contacto en campañas anteriores.
- Existe una presencia secundaria, pero significativa, en 1 contacto previo (alrededor de 4.500 clientes).
- Las frecuencias caen drásticamente a partir de 2 contactos previos (menos de 1.000 clientes), y son prácticamente nulas o insignificantes para 3 o más contactos.

Esta distribución confirma la observación hecha con la variable *pdays*: la campaña actual se dirige principalmente a una base de datos "nueva" o no contactada previamente. Esto implica que, para la mayoría de los clientes, la campaña actual constituye su primera o una de sus primeras interacciones con el esfuerzo de marketing del banco.



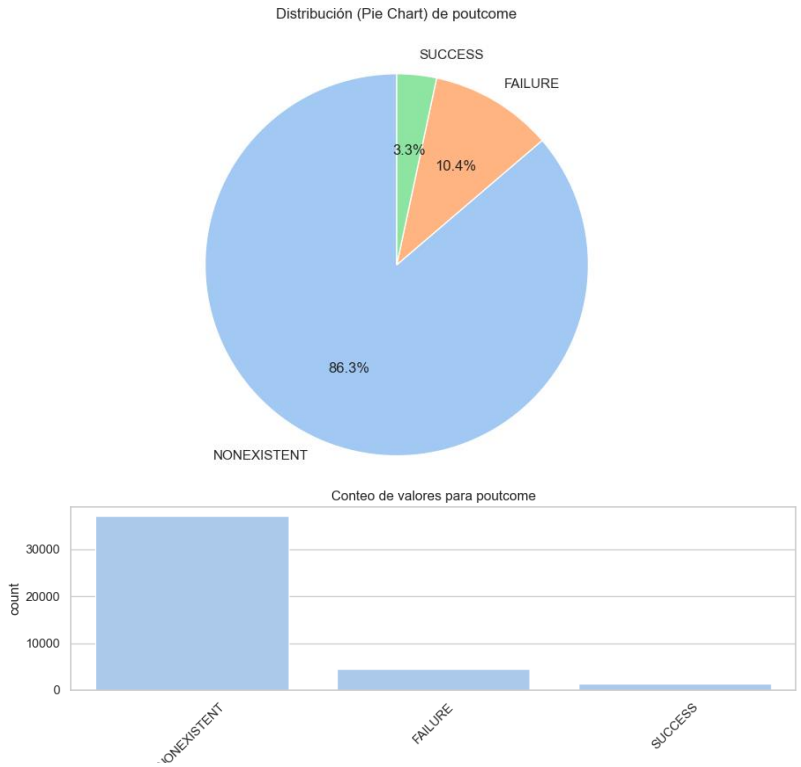
Resultado de la campaña de marketing anterior. (poutcome)

Los gráficos muestran la distribución del resultado de la campaña de marketing anterior para cada cliente, categorizados como NONEXISTENT (no existió contacto), FAILURE (fallo) o SUCCESS (éxito).

- La categoría NONEXISTENT es la abrumadora mayoría, representando el 86.3% del total de la base de datos. En el gráfico de conteo, esta categoría supera los 35.000 clientes.
- La segunda categoría más grande es FAILURE (Fallo), con el 10.4% del total.
- La categoría SUCCESS (Éxito) es la más pequeña, representando solo el 3.3% de los resultados.

Esta distribución confirma, por tercera vez (junto con *pdays* y *previous*), que la mayor parte de la base de datos no tiene un historial de contacto con las campañas de marketing del banco.

Para el grupo que sí fue contactado previamente (el 13.7% restante), el fracaso fue el resultado más común de esas campañas anteriores, con el éxito siendo un evento muy poco frecuente. Esto sugiere que las campañas anteriores tuvieron un rendimiento bajo en términos de conversión.



El índice de precios al consumidor. (cons.price.idx)

El gráfico de barras muestra la distribución de los clientes contactados en relación con el Índice de Precios al Consumidor (IPC) registrado en ese momento.

La distribución es multimodal (con varios picos) y no uniforme, lo que indica que los contactos de la campaña se concentraron fuertemente en períodos específicos, definidos por el nivel del IPC.

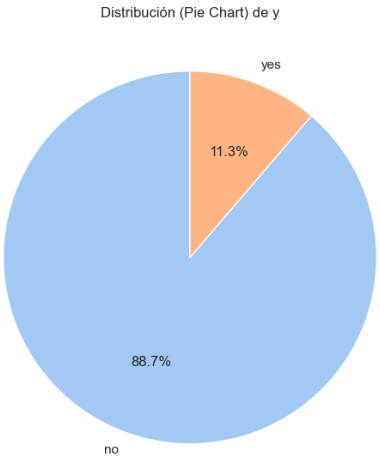
- Los contactos se agruparon principalmente en torno a dos rangos clave, creando los picos más altos de la distribución:
    - El pico más alto se encuentra en el valor 93.994, con cerca de 8.000 clientes contactados.
    - Otros picos significativos ocurren alrededor de 93.918 (cerca de 7.000 contactos) y 92.893 (cerca de 6.000 contactos).
- El resto de los valores del IPC, especialmente aquellos hacia la derecha de la distribución, presentan frecuencias de contacto muy bajas.

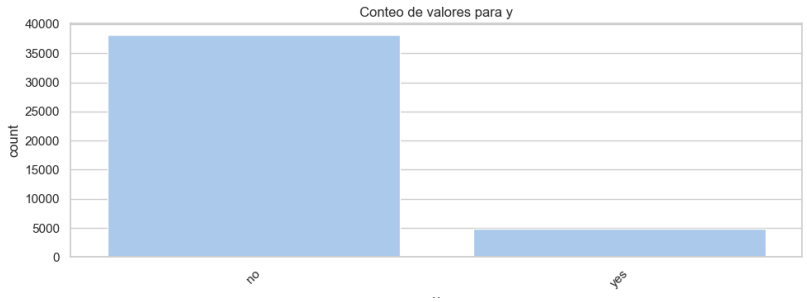
Esto implica que la estrategia de marketing del banco no fue continua a lo largo del tiempo, sino que se programó en meses o trimestres específicos donde el IPC reflejaba ciertas condiciones económicas (por ejemplo, períodos de alta inflación o estabilidad), priorizando las campañas en estos momentos concretos.

Indica si el cliente ha suscrito un producto o servicio (Sí/No). (y)

Los gráficos (circular y de conteo) muestran la distribución de la variable y, que indica si el cliente suscribió el depósito a plazo bancario (Yes/Sí) o no (No).

La distribución es extremadamente desequilibrada (desbalanceada), con una marcada asimetría hacia la categoría "No". La categoría No (No Suscribió) es la inmensa mayoría, representando el 88.7% del total de la base de datos.La categoría Yes (Sí Suscribió), que representa el éxito de la campaña, es la minoría, con solo el 11.3% del total de clientes.





CONCLUSIÓN

El análisis de la base de datos de marketing bancario se centró en perfilar la clientela, la estrategia de contacto y el rendimiento histórico, con el objetivo de optimizar futuras campañas para la suscripción de depósitos a plazo.

1. Resultado Global y Desafío Principal

El hallazgo más crítico es la tasa de éxito, que es baja y altamente desequilibrada. Solo el 11.3% de los clientes contactados suscribió, mientras que el 88.7% no lo hizo. Este desequilibrio define el principal desafío del modelado predictivo y resalta la necesidad de identificar con precisión el perfil del cliente propenso a convertir.

2. Perfil de la Base de Clientes

La clientela del banco se caracteriza principalmente por ser Casada (60.6%) y poseer un nivel educativo alto, dominado por personas con título universitario y educación secundaria completa.

En cuanto a productos financieros, aunque la mayoría de los clientes ya posee un préstamo hipotecario (la distribución es relativamente equilibrada), hay una penetración muy baja del préstamo personal, lo que sugiere un área de oportunidad para la venta cruzada de este último producto.

3. Estrategia y Historial de la Campaña

La estrategia de contacto se centra en el teléfono móvil (63.7%), privilegiando un método de comunicación directo y moderno.

Respecto al esfuerzo de contacto, se observaron dos puntos clave:

- Base Mayoritariamente "Nueva": La gran mayoría de los clientes (más del 86%) no tiene un historial de contacto con campañas previas
- Baja Persistencia: La campaña actual mostró poca persistencia, contactando a la mayoría de los clientes solo una o dos veces antes de cesar el intento.

4. Implicaciones Estratégicas

Los datos sugieren que el banco está utilizando su campaña actual para probar una base de datos prácticamente virgen en términos de marketing telefónico. El bajo éxito global (11.3%) puede deberse a la baja persistencia o a que los contactos se realizaron en periodos económicos específicos definidos por el IPC, que no necesariamente se alinearon con la mejor ventana de oportunidad para el cliente.

Para optimizar el rendimiento, futuras estrategias deben enfocarse en segmentar y priorizar a los clientes según su potencial de conversión, utilizando el perfil demográfico (educación, estado civil) y financiero, e investigando si un nivel de persistencia ligeramente mayor podría mejorar la tasa de éxito entre la amplia base de clientes que solo fue contactada una o dos veces.