**Proyecto BI Lead – Empresa ABC**

**Autor:** Jaime Orejarena  
**Rol:** Business Intelligence Lead  
**Fecha:** [fecha de entrega]  
**Ubicación del Proyecto:** Colombia – Industria Gambling

**Objetivo del Proyecto**

Realizar un análisis integral de los datos de la empresa para:

* Identificar oportunidades de optimización comercial y operativa.
* Preparar y transformar los datos utilizando SQL y Python.
* Visualizar los resultados clave mediante Power BI.
* Realizar una proyección de ingresos netos para los próximos dos años.
* Documentar y gestionar el proyecto en Notion para asegurar trazabilidad y claridad.

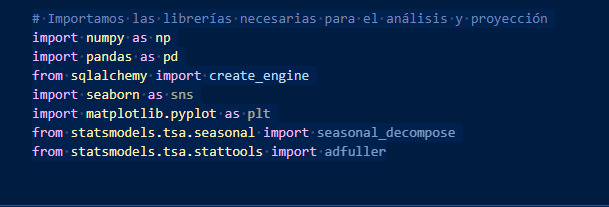
**Sección 1: Gestión de Datos y Análisis Exploratorio**

**Objetivo**

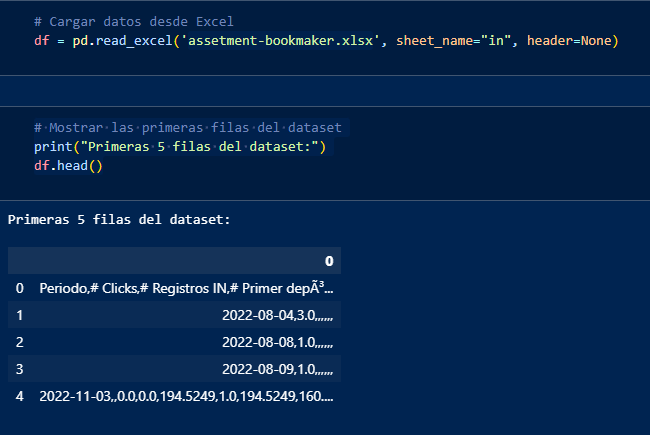
Realizar el análisis exploratorio de los datos disponibles en el archivo assetment-bookmaker.xlsx utilizando Python y SQL, con el fin de entender la estructura de los datos, identificar valores faltantes, tipos de variables y detectar posibles problemas de calidad de datos.

**1.1 Análisis Exploratorio en Python**

**Carga de Librerías**



**Carga del archivo Excel e impresion de las primeras filas**

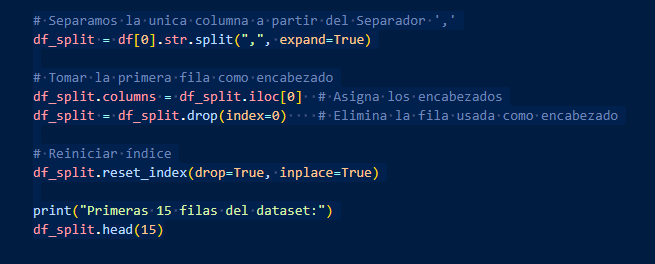


**Observación inicial**

El archivo Excel no está estructurado como una tabla tradicional. En su lugar, presenta todos los datos embebidos en una única columna, separados por comas, como si fuera un archivo .csv dentro de una celda de Excel.

**Transformación del DataFrame**

Se aplicó un proceso de "split" para descomponer la única columna en múltiples columnas usando str.split(), y se ajustó la primera fila como encabezado.

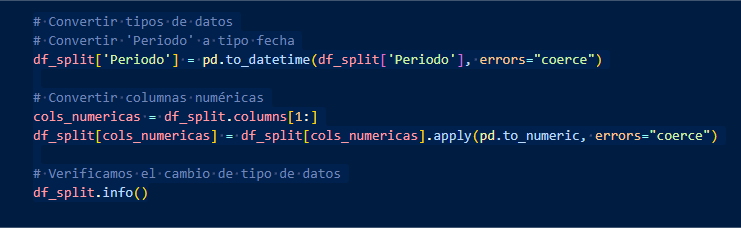


#### **Conversión de tipos de datos**

#### Todas las columnas eran del tipo object, incluyendo fechas y valores numéricos. Se aplicaron transformaciones para:

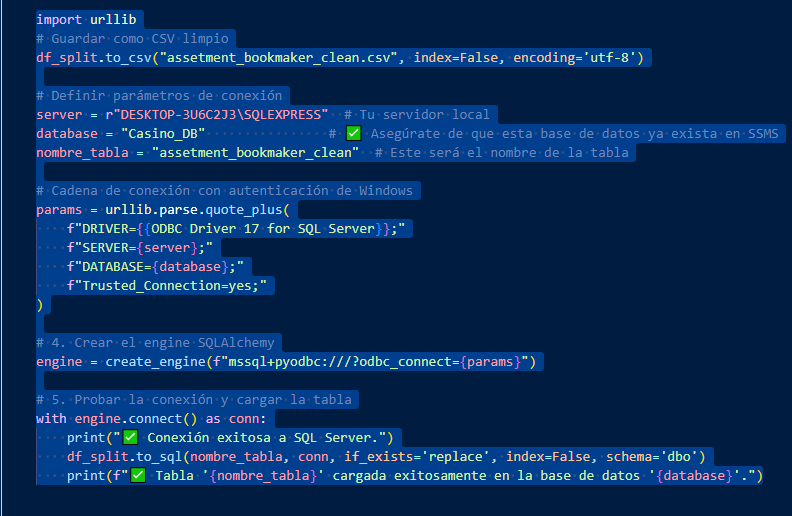
#### Eliminar caracteres especiales.

#### Convertir a formato numérico con pd.to\_numeric(errors='coerce').



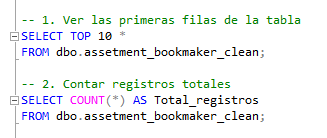
### **Carga del DataFrame limpio a SQL Server**

#### Una vez procesado y limpiado el dataset, se cargó en SQL Server Management Studio para continuar el análisis.



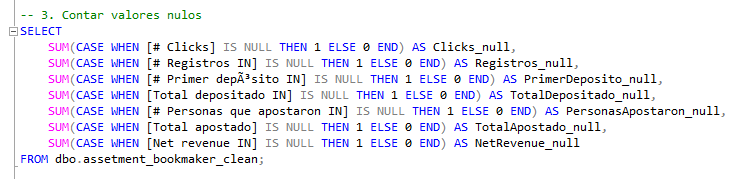
**1.2 EDA en SQL**

**Validaciones iniciales**



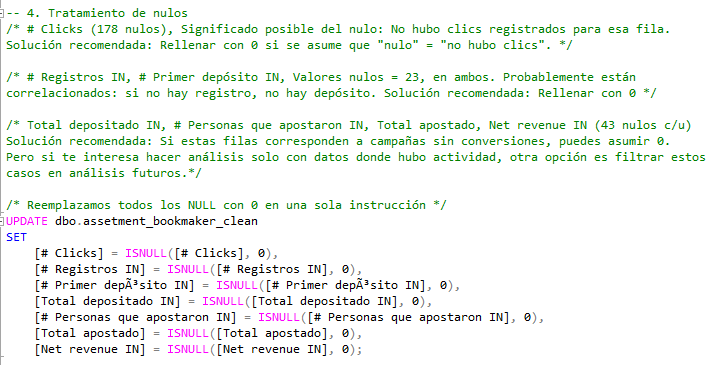
**Revisión de valores nulos**

Conteo de valores nulos por columna



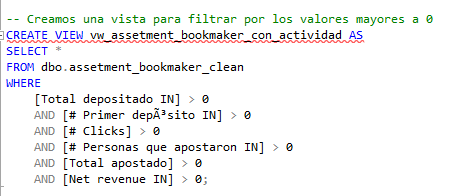
**Reemplazo de valores nulos**

Sustitución de valores nulos por 0



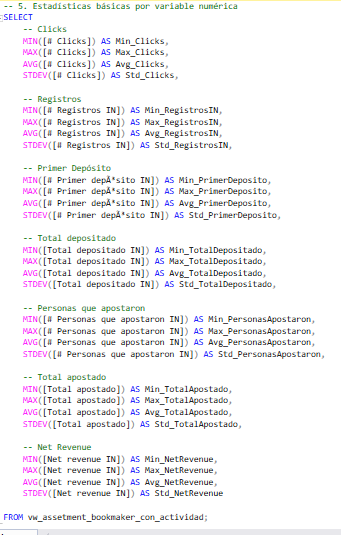
**Creación de Vista: Filtrado por Actividad Real**

Filtrar solo filas con actividad

****

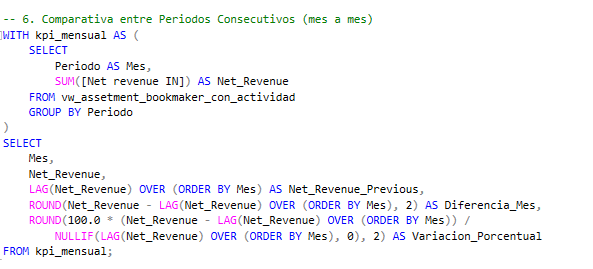
**Estadísticas Básicas por Métrica**

Estadísticos descriptivos para columnas numéricas

****

**Análisis de Tendencia Mensual**

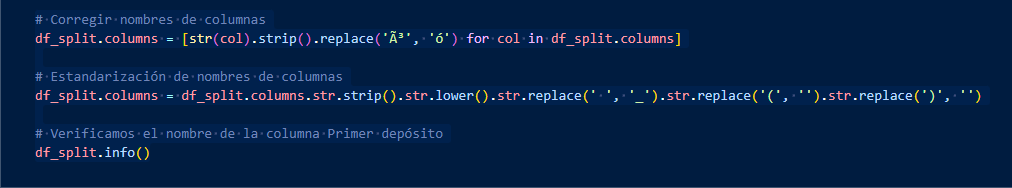
Variación mensual del ingreso neto



**Limpieza y Transformación de Datos en Python**

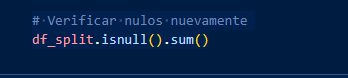
**1. Corrección y estandarización de nombres de columnas**

Se corrigieron errores de codificación y se estandarizaron los nombres a formato snake\_case.

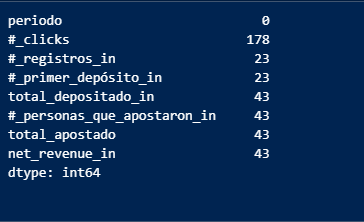


**2. Verificación de datos nulos**

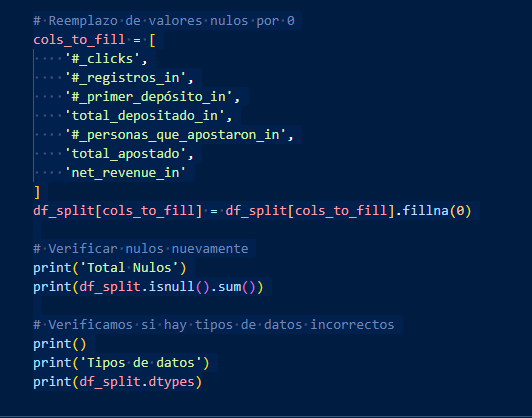
Se identificaron y reemplazaron por 0.



y los Hallazgos fueron:

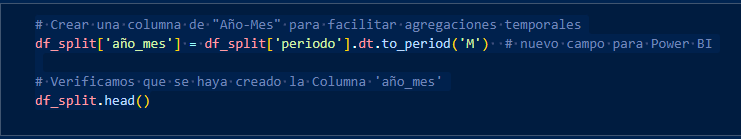


hacemos el reemplazo por numeros ‘0’



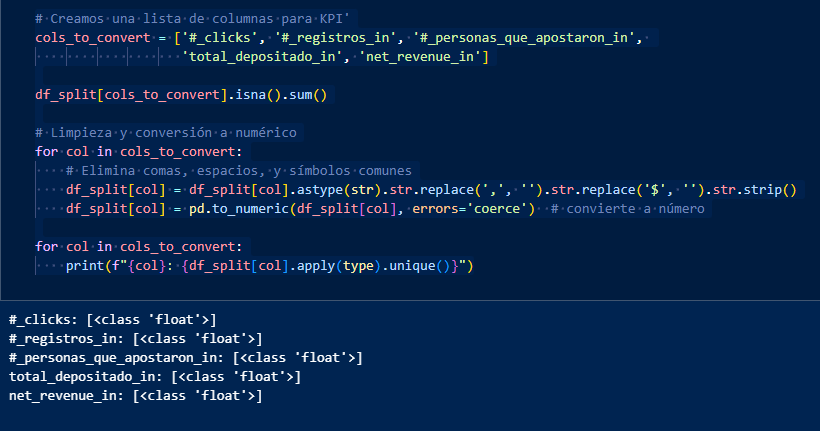
**3. Creación de columna Año-Mes**

Para análisis temporal se creó una columna Año-Mes a partir del campo de fecha.



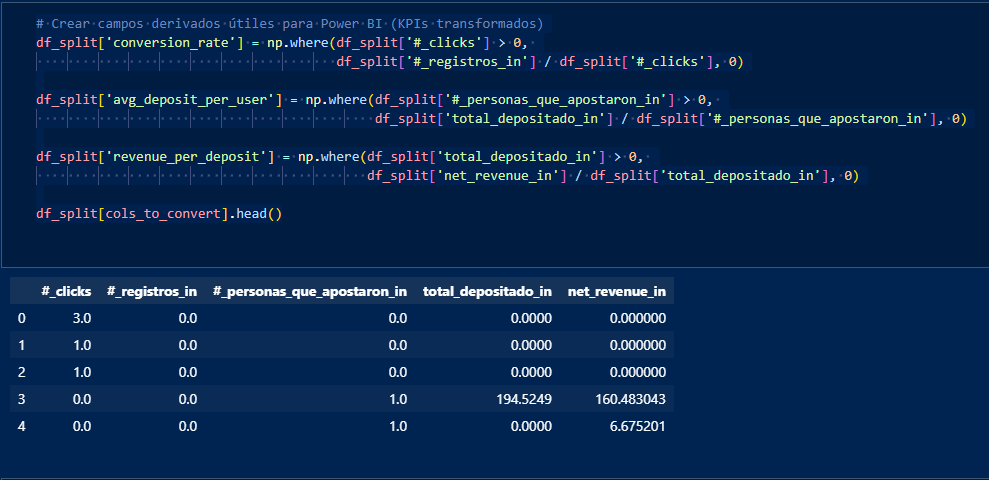
#### **4. Estandarización de KPIs**

* Eliminación de comas, símbolos y espacios.
* Conversión forzada a tipo float.



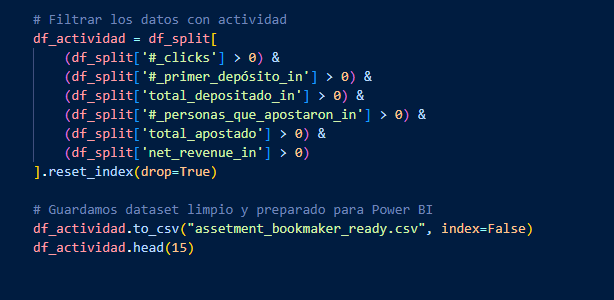
**5. Creación de KPIs Derivados**

Se agregaron métricas adicionales para análisis financiero y operativo.



**6. Filtrado de Datos con Actividad Significativa**

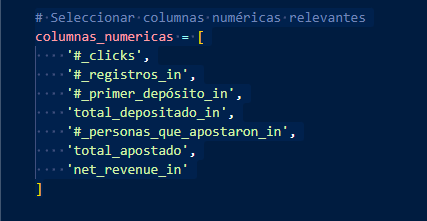
Solo se conservaron registros con actividad comprobable (ej. depósito, clics, apuestas).



## **Identificación y Manejo de Outliers**

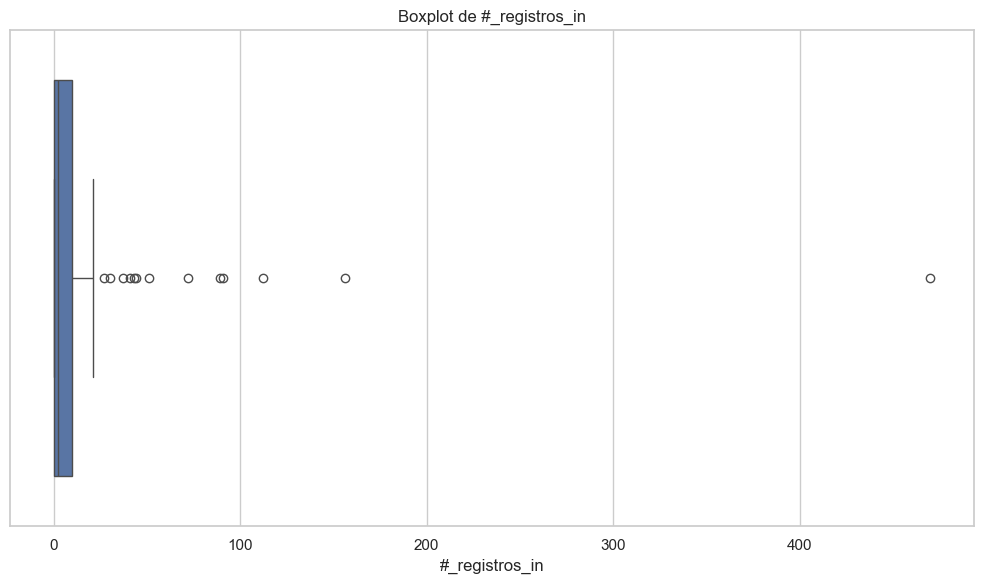
## **1. Selección de variables numéricas relevantes**

Métricas clave: número de apuestas, ingresos, depósitos, etc.



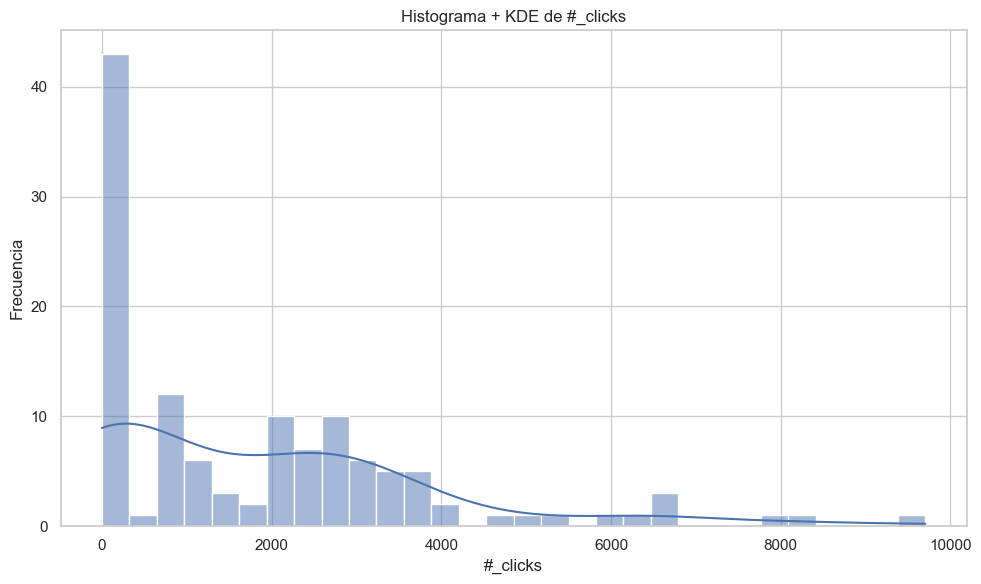
**2. Visualización y detección de outliers**

Se usaron boxplots para detectar valores fuera del rango intercuartílico (IQR).



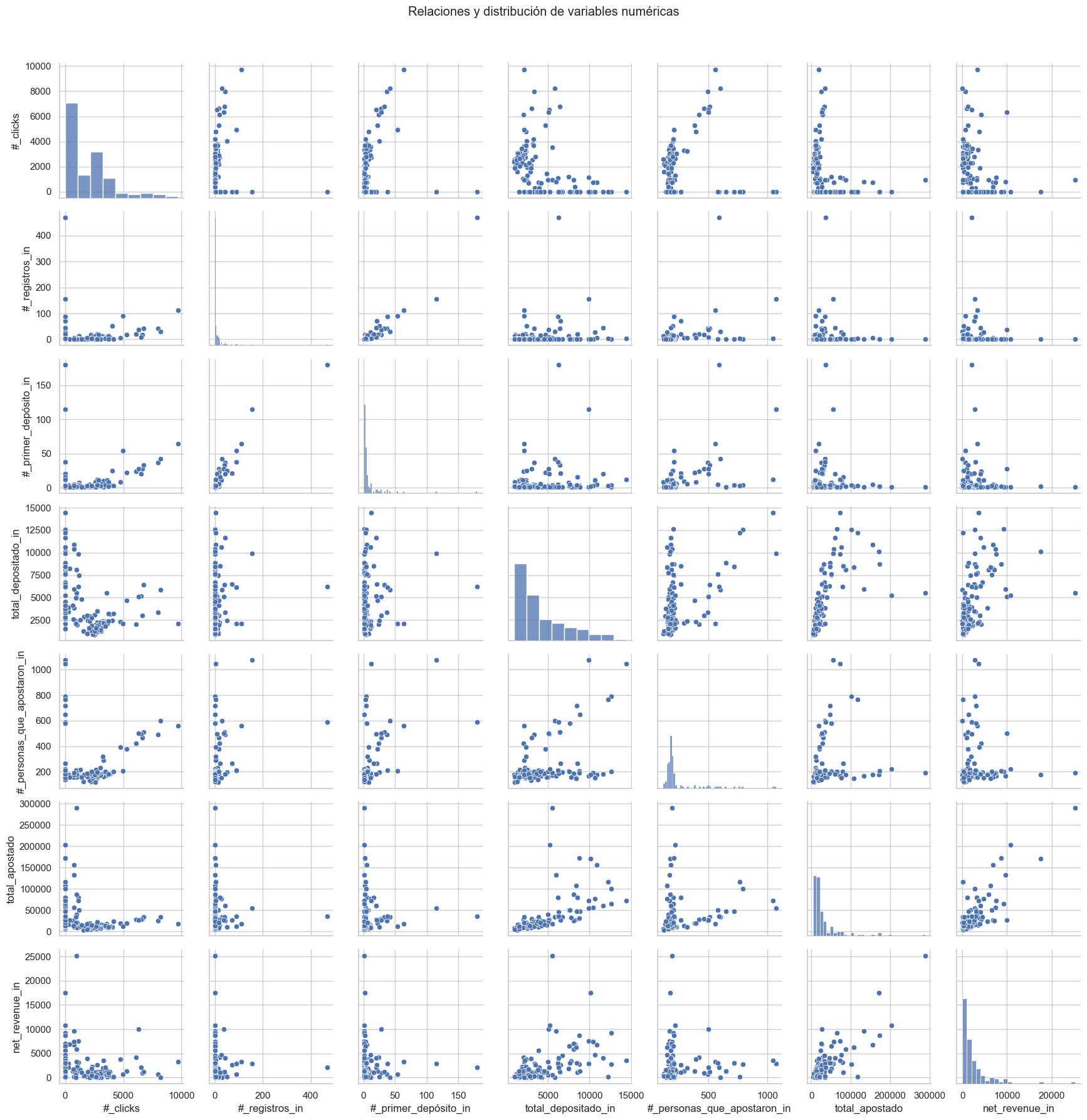
**3. Análisis de la distribución de los datos**

Se usaron histogramas con KDE, encontrando distribuciones sesgadas a la derecha.



El análisis reveló que la mayoría de las variables presentan una distribución sesgada a la derecha, es decir, con valores concentrados en los rangos inferiores y colas largas hacia la derecha, lo cual es característico en datos de comportamiento de usuarios online donde unos pocos generan un gran impacto en métricas clave como ingresos o montos apostados.

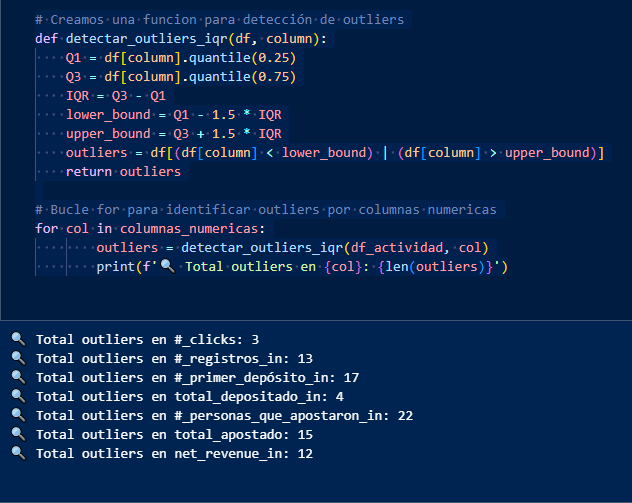
**4. Análisis de correlación entre variables**

pairplot mostró baja correlación entre métricas, útil para decidir el tipo de modelo a aplicar.

Los resultados mostraron una baja correlación entre las variables, lo que implica que los factores analizados no presentan una dependencia lineal significativa entre sí. Esto se debe considerar al momento de seleccionar los modelos estadísticos o de machine learning apropiados, priorizando algoritmos que no asuman relaciones lineales entre variables.

**5. Tratamiento Técnico de Outliers**

Se aplicó el método del IQR para ajustar o eliminar los extremos fuera del rango esperado.



El conjunto de datos contenía 123 observaciones (df\_actividad.shape = (123, n)), lo cual evidencia que los outliers representaban menos del 18% por variable.

### **Recomendación estratégica**

Dado que:

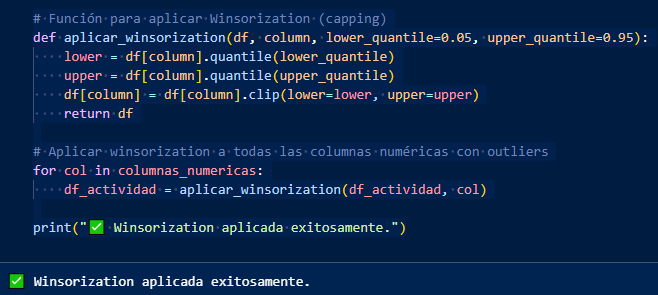
* El dataset es pequeño (123 filas).
* Los outliers no superan el 18% por variable.
* La proyección futura se basa en patrones reales que podrían estar en esos extremos.

**Recomendación técnica:** Aplicar **Winsorization** o **Capping** para conservar todos los registros y evitar eliminar datos relevantes. Este enfoque limita los valores extremos a los percentiles 5 y 95, preservando así la estructura del dataset sin alterar la escala mediante transformaciones logarítmicas.

En fases posteriores del modelado, si se evidencia sesgo o multicolinealidad elevada, se evaluará complementar con transformaciones logarítmicas específicas por variable.

### **Aplicación de Winsorization**

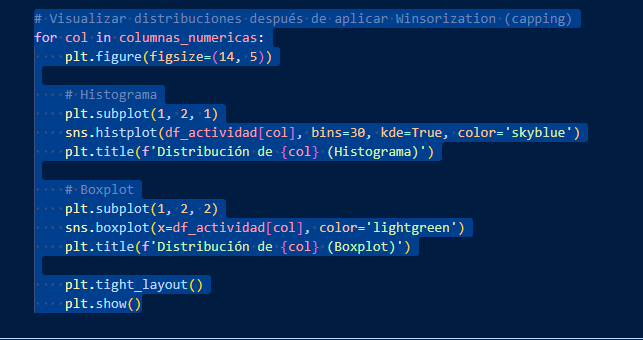
Se definió una función para aplicar Winsorization (también conocido como "capping") a cada columna numérica, Luego, se aplicó la función a cada variable con presencia de outliers.

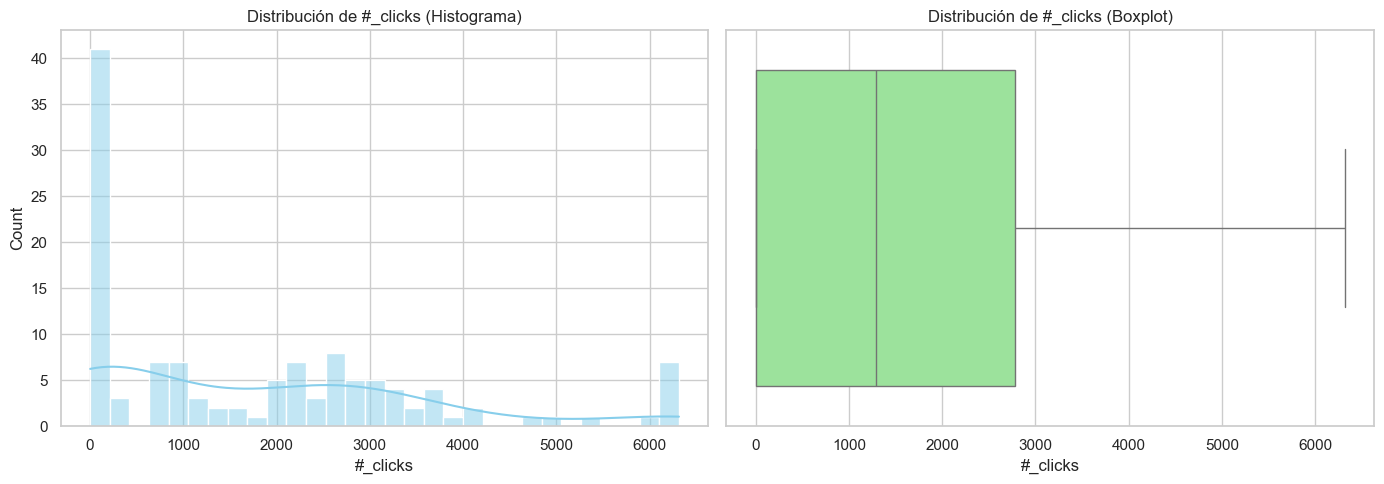


### 

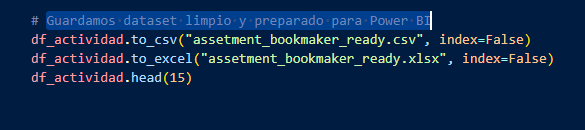
### **Verificación post-procesamiento**

Se visualizaron nuevamente los boxplots y distribuciones para confirmar que los valores extremos habían sido suavizados sin eliminar información:





**Guardamos el dataset limpio y preparado para power BI**

****

**Proyección de Ingresos Netos (net\_revenue\_in)**

**Objetivo**

Proyectar el ingreso neto esperado durante los próximos 24 meses utilizando un modelo estadístico adecuado que refleje las tendencias reales del negocio, garantizando precisión y fiabilidad.

**Tareas Realizadas**

* Selección de un modelo de proyección adecuado.
* Justificación técnica y de negocio del modelo seleccionado.
* Proyección del ingreso neto para los próximos 2 años.
* Validación del modelo con datos históricos.
* Garantizar que el modelo refleje fielmente las tendencias del negocio.

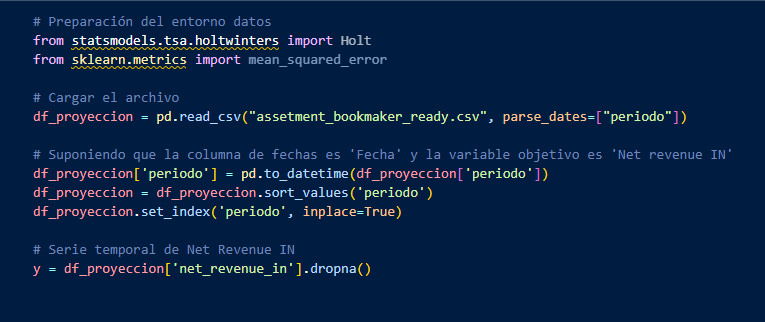
**Modelo Seleccionado: Holt (Exponential Smoothing)**

**Justificación del Modelo**

Se ha utilizado el modelo **Holt** (suavizado exponencial doble), una variante del modelo **Holt-Winters**, en su versión sin estacionalidad. Las razones para esta elección son las siguientes:

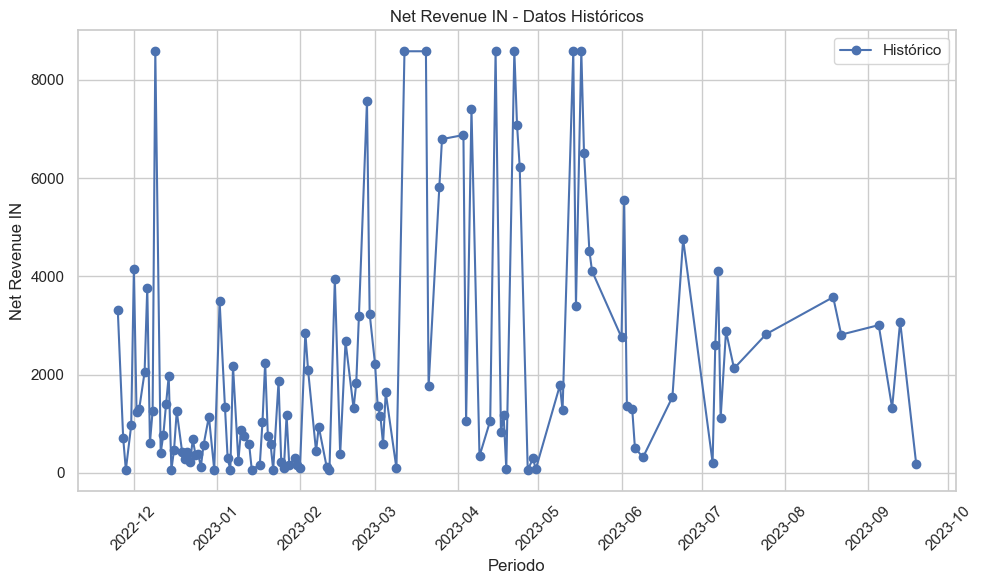
* **Baja disponibilidad de datos históricos**: la serie cuenta únicamente con 11 puntos de datos, lo cual limita el uso de modelos más complejos.
* **Captura de tendencia lineal**: el modelo Holt permite modelar de forma eficiente tendencias suaves en series temporales cortas sin requerir una estacionalidad explícita.
* **Simplicidad y robustez**: es fácil de interpretar para stakeholders no técnicos y reduce el riesgo de sobreajuste.
* **Adecuación al negocio**: refleja correctamente la evolución esperada del ingreso neto en función del comportamiento histórico observado.

**Preparación del Entorno de Datos**



**Visualización Inicial de la Serie Temporal**

****

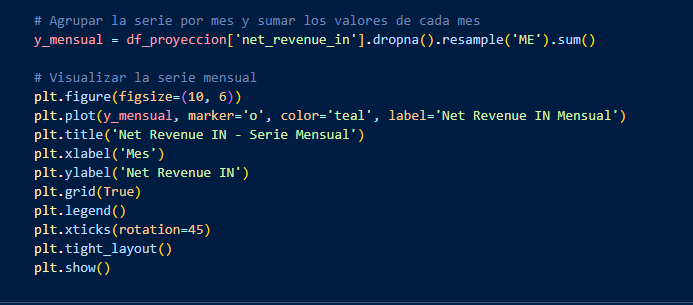
****

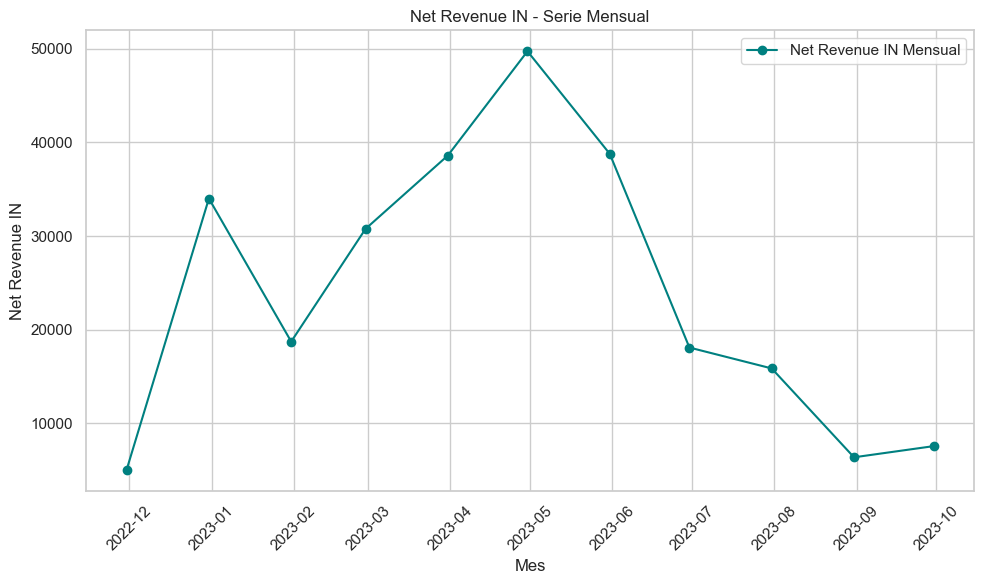
**Interpretación del Gráfico Histórico**

* La serie muestra una alta volatilidad, con picos muy marcados (ej. dic 2022, mar-abr 2023) y caídas abruptas, probablemente causadas por eventos puntuales como promociones o finales deportivas.
* No hay una tendencia clara, aunque a partir de julio de 2023 los ingresos parecen estabilizarse en un rango de 1.000 a 4.000.
* No se observa estacionalidad consistente, por lo que se descarta el uso de modelos como Holt-Winters completo.
* Algunos días sin datos generan saltos irregulares en la línea, indicando que los datos están a nivel diario, pero con huecos.

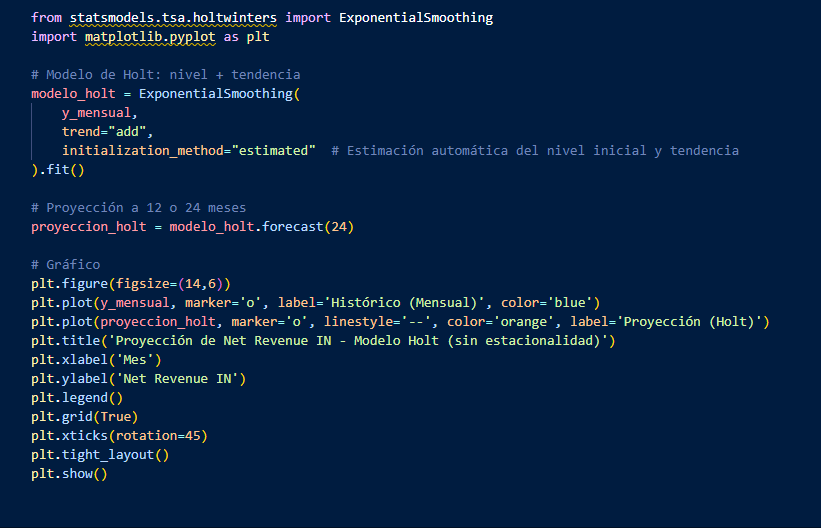
**Recomendación:**Dado el comportamiento irregular y la falta de estacionalidad, el modelo Holt sin estacionalidad es ideal. Captura la tendencia subyacente sin sobre ajustarse. Para la proyección de los próximos 24 meses, se utilizarán datos agregados mensualmente para suavizar la serie y reducir la influencia de outliers.

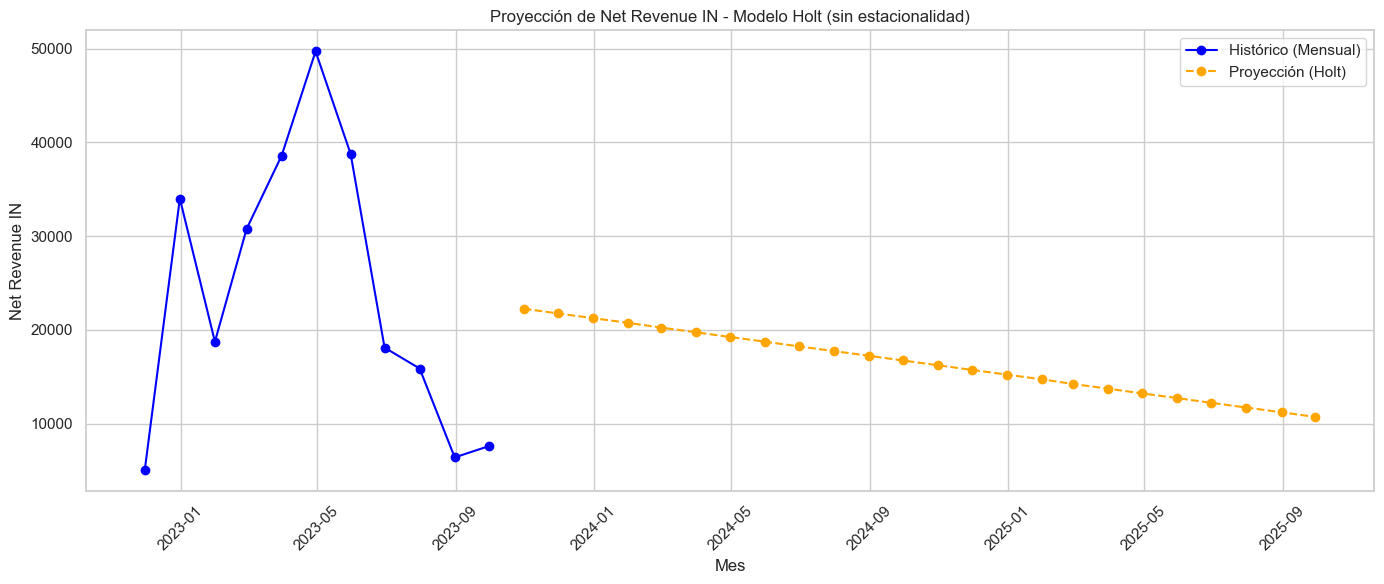
**Visualización Mensual**

****

****

**Aplicar el Modelo Holt**

****

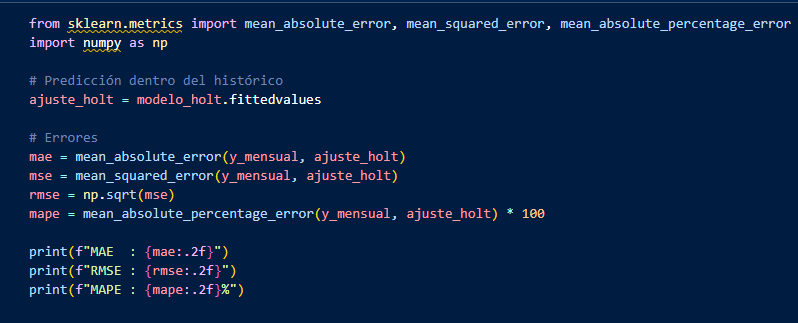
****

**Evaluación Visual del modelo**

**Observaciones:**

* El modelo proyecta una tendencia decreciente constante en la serie.
* El histórico muestra un patrón bastante volátil y no claramente lineal, con un pico pronunciado en el primer semestre de 2023 y una caída posterior.
* No se aprecia estacionalidad clara (además solo 11 observaciones), por lo tanto, tiene sentido usar el modelo Holt sin componente estacional.

**Errores y precisión del modelo**

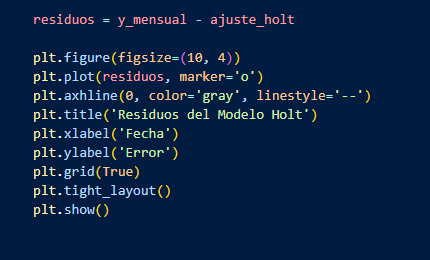
****

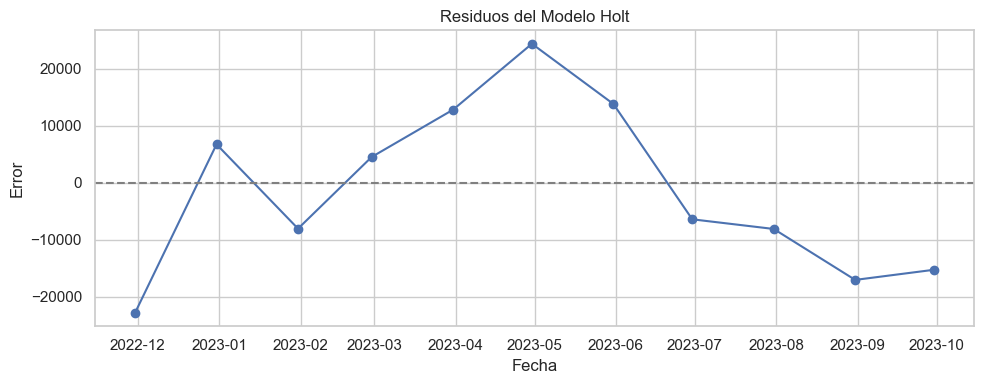
**MAE:** 12729.15

**RMSE:** 14247.36

**MAPE:** 109.19%

**Análisis de residuos**

****

****

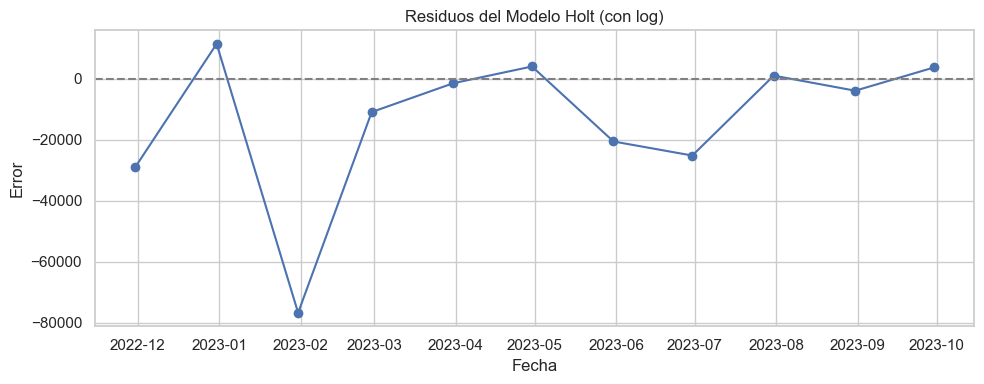
**Modelo Holt con logaritmo + proyección 24 meses**

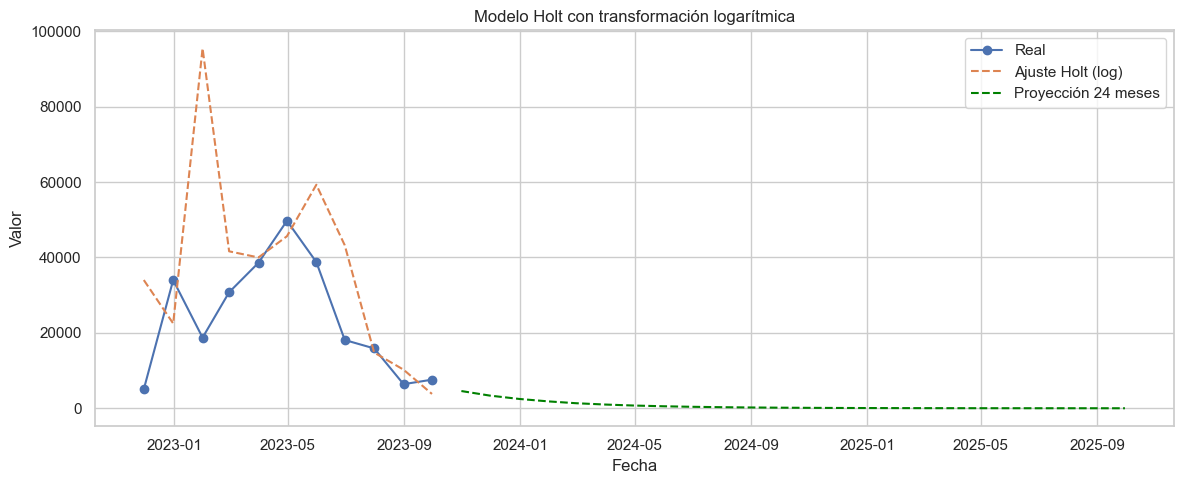
**Resultados:**

**MAE:** 17079.64

**RMSE:** 27089.80

**MAPE:** 124.72%

****

****

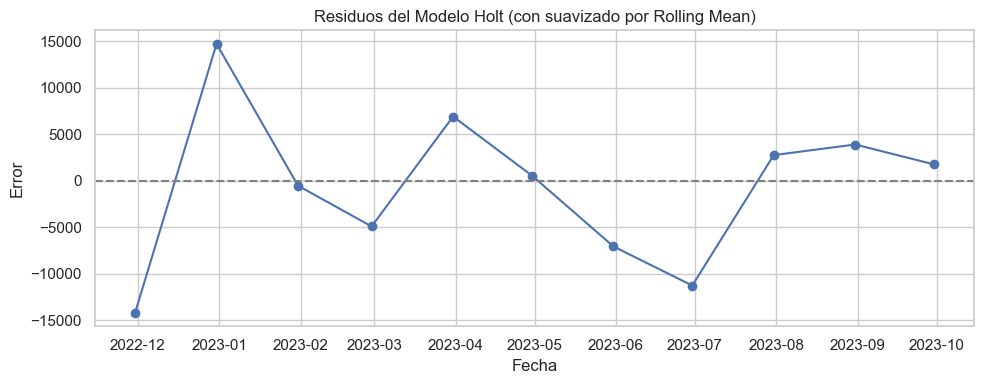
**Modelo Holt con Rolling Mean + proyección 24 meses**

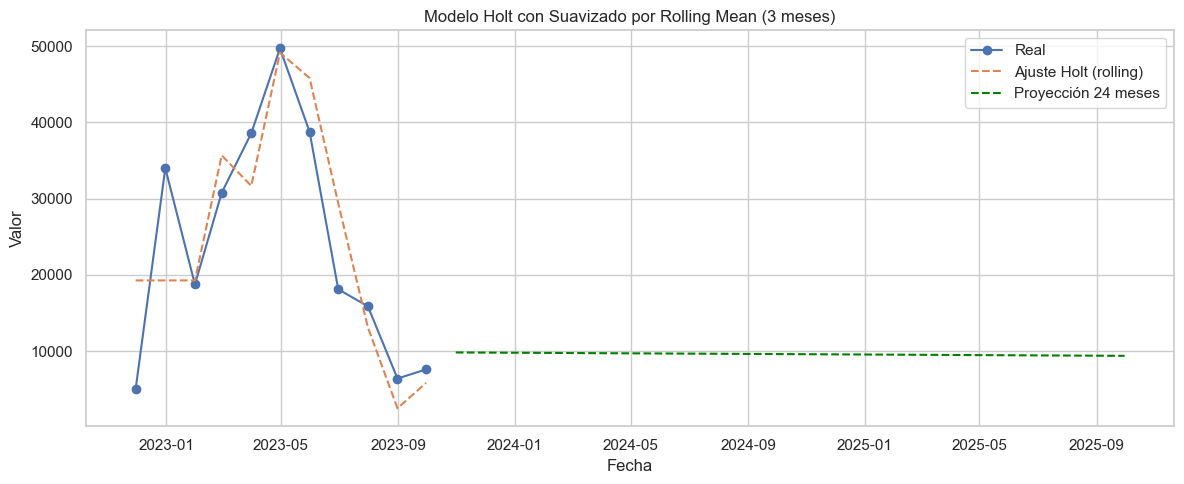
**Resultados:**

**MAE:** 6241.51

**RMSE:** 7949.28

**MAPE:** 49.52%

****

****

**Modelo Seleccionado: Holt con Rolling Mean**

Se evaluaron tres variantes de modelos de proyección sobre la serie net\_revenue\_in con solo 11 observaciones mensuales disponibles.

| **Modelo** | **MAE** | **RMSE** | **MAPE** |
| --- | --- | --- | --- |
| **Holt (simple)** | **12,729.15** | **14,247.36** | **109.19%** |
| **Holt + Logaritmo** | **17,079.64** | **27,089.80** | **124.72%** |
| **Holt + Rolling Mean (3 meses)** | **6,241.51** | **7,949.28** | **49.52% ✅** |

**Motivo de selección:**

* El modelo con *rolling mean* de 3 meses reduce la volatilidad, suaviza la serie y mejora sustancialmente las métricas de error.
* Es robusto, interpretable y adecuado para series cortas (≤12 datos).
* Mejora la estabilidad del *forecast*, evitando sobreajuste y proyecciones poco realistas.

**Proyección:**Se realizó la proyección de los próximos 24 meses (2 años), generando una tabla de resultados lista para visualizarse en Power BI.

**Archivo exportado: proyeccion\_net\_revenue\_24\_meses.csv**

**DASHBOARD POWER BI**

****

**Interpretación del Dashboard: “Desempeño & Proyección”**

**1. Métricas de Alto Nivel (KPIs principales):**

* **Total Clicks:** 219,55K
* **Total Registros:** 981,10
* **Total Depósitos:** 517,12K
* **Total Ganancias:** 263,62K

Estas cifras resumen el rendimiento general del funnel de conversión, desde clics iniciales hasta las ganancias generadas. Permiten una lectura rápida del volumen de actividad y monetización.

**2. Gráfico de Línea: Ganancias Totales vs. Ganancias Proyectadas**

* Muestra la evolución temporal de las ganancias y la comparación con las proyecciones futuras.
* Se aprecia una **alta ganancia en el primer semestre**, con una caída en los meses posteriores (consistente con el análisis previo de estacionalidad y volatilidad).
* La línea proyectada indica una tendencia descendente pero controlada, suavizada con el modelo aplicado Holt con Rolling Mean.

**3. Gráfico de Barras Apiladas: Clicks, Primer Depósito y Registros por Año\_Mes**

* Muestra la distribución mensual de eventos clave del embudo.
* Se observa claramente que los **clicks dominan en volumen**, seguidos por registros y depósitos.
* Hay una reducción progresiva en eventos en los últimos meses, lo que coincide con la caída en las ganancias.

**4. Gráfico de Dispersión / Burbujas: Personas vs Total Apostado vs Ganancias Totales**

* Relaciona la cantidad de usuarios, el monto apostado y las ganancias obtenidas por mes.
* Permite identificar meses con **alta eficiencia (alta ganancia con pocos usuarios)** o viceversa.
* Las burbujas más grandes indican meses de alta actividad.

**5. Filtros: Año y Mes**

* Permiten segmentar el análisis temporalmente, útil para identificar patrones o cambios por campañas puntuales.

**6. Tabla Detallada**

* Resume métricas como clicks, registros, depositos, ganancias y conversiones por periodo.
* Es útil para validación cruzada, exportación a Excel o como tabla fuente para reportes ejecutivos.

**Conclusión Técnica**

Este dashboard ofrece una **visión integral del funnel de conversión y sus resultados financieros**, combinando análisis histórico y proyecciones futuras. La estructura es sólida para:

* Evaluar desempeño por periodo.
* Identificar puntos de fuga en el funnel (muchos clicks pero pocos registros o depósitos).
* Analizar la efectividad de campañas.
* Proyectar resultados con base en modelos estadísticos.