Nombre del Proyecto

Ingresos a través del tiempo: una mirada al futuro financiero

Introducción

El presente proyecto tiene como finalidad aplicar técnicas de series temporales para pronosticar los ingresos mensuales de una empresa retail. El dataset proviene de la plataforma Kaggle, que simula los datos de ventas e ingresos de una empresa, el cual incluye variables como la cantidad vendida, el costo promedio y la nómina anual promedio de la región

El análisis de este tipo de información resulta clave para instituciones financieras, ya que permite comprender mejor el comportamiento de clientes, ingresos, egresos, etc. para predecir riesgos y tomar decisiones más informadas.

Fuente: https://www.kaggle.com/datasets/podsyp/time-series-starter-dataset

Objetivos

- Modelar y predecir los ingresos mensuales de la empresa utilizando modelos de series de tiempo como Holt – Winters y SARIMA.
- Identificar patrones estacionales que impactan de forma recurrente en la variable de interés.
- Evaluar el desempeño del modelo mediante métricas como MAPE, MAE y RMSE para validar su precisión en escenarios de análisis a mediano plazo.

Herramientas

- Python
- Google Colaboratory

Contenido

1. Planteamiento del problema

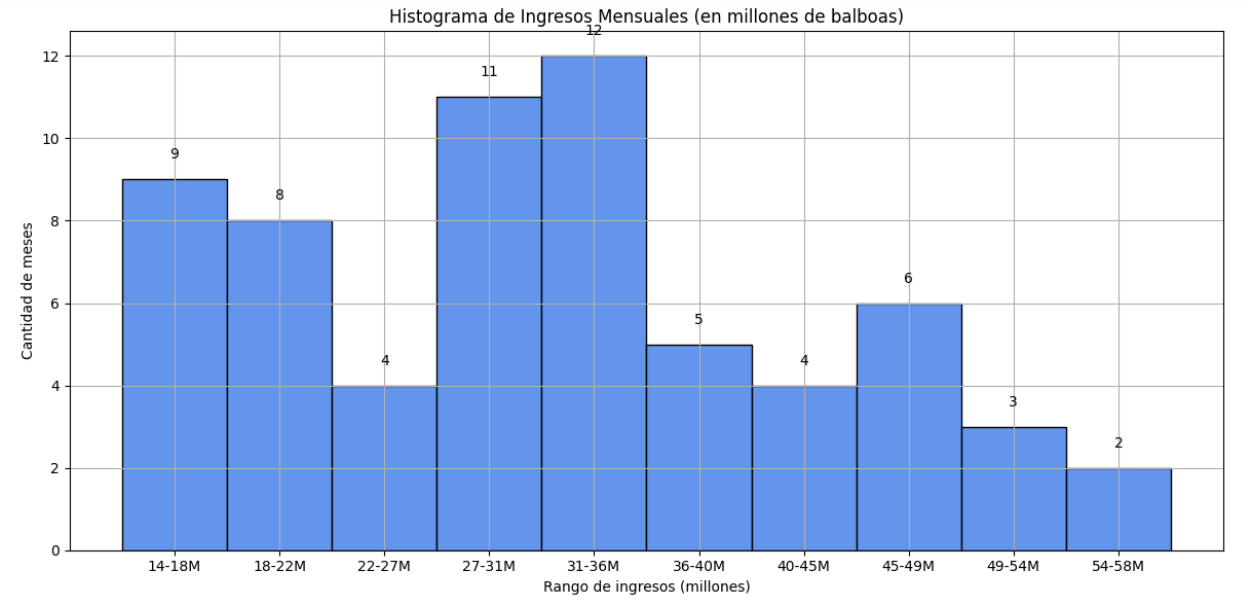
A lo largo del tiempo, los ingresos de una empresa pueden verse afectados por múltiples factores internos y externos, como la cantidad de ventas, los costos promedios y las condiciones económicas regionales. Sin embargo, la falta de un sistema de pronóstico confiable limita la capacidad de la organización para anticiparse a cambios, optimizar sus recursos y planificar estratégicamente y dificulta tomar decisiones informadas en áreas clave como la planificación financiera, el abastecimiento de inventario y la gestión del personal.

2. Conociendo la variable de interés

El conjunto de datos tiene una dimensión de 64 registros y 5 variables que incluye variables como el conteo de ventas, el costo promedio, y la nómina anual promedio de la región. La periodicidad de la serie es mensual y va desde enero del 2015 hasta abril del 2020.

	A	B	C	D	E
1	Period	Revenue	Sales_quantity	Average_cost	The_average_annual_payroll_of_the_region
2	01.01.2015	16010072.12	12729	1257.763541	30024676
3	01.02.2015	15807587.45	11636	1358.507	30024676
4	01.03.2015	22047146.02	15922	1384.697024	30024676
5	01.04.2015	18814583.29	15227	1235.606705	30024676
6	01.05.2015	14021479.61	8620	1626.621765	30024676
7	01.06.2015	16783928.52	13160	1275.374508	30024676
8	01.07.2015	19161892.19	17254	1110.576805	30024676
9	01.08.2015	15204984.3	8642	1759.42887	30024676
10	01.09.2015	20603939.98	16144	1276.259909	30024676
11	01.10.2015	20992874.78	18135	1157.588904	30024676
12	01.11.2015	14993369.66	10841	1383.024597	30024676
13	01.12.2015	27791807.64	22113	1256.808558	30024676
14	01.01.2016	28601586.5	15365	1861.476505	27828571
15	01.02.2016	22367074.07	13153	1700.530226	27828571
16	01.03.2016	29738608.57	18339	1621.604699	27828571
17	01.04.2016	28351007.94	13909	2038.321083	27828571
18	01.05.2016	15264603.73	8553	1784.707557	27828571
19	01.06.2016	24385658.08	15101	1614.837301	27828571
20	01.07.2016	29486517.07	15695	1878.720425	27828571
21	01.08.2016	15270117.26	8314	1836.675157	27828571
22	01.09.2016	36141027.56	17764	2034.509545	27828571
23	01.10.2016	27915143.66	18969	1471.61915	27828571
24	01.11.2016	21272049.35	13433	1583.566541	27828571
25	01.12.2016	42014159.88	27029	1554.410444	27828571

Variable original	Traducción
Period	Periodo
Revenue	Ingresos
Sales_quantity	cantidad_ventas
Average_cost	costo_promedio
The_average_annual_payroll_of_the_region	promedio_anual_empleados_region



Interpretación: se puede observar que la mayoría de los ingresos empresariales están entre 27 y 36 millones, seguido, los ingresos entre 14 y 22 millones son los que más predominan.

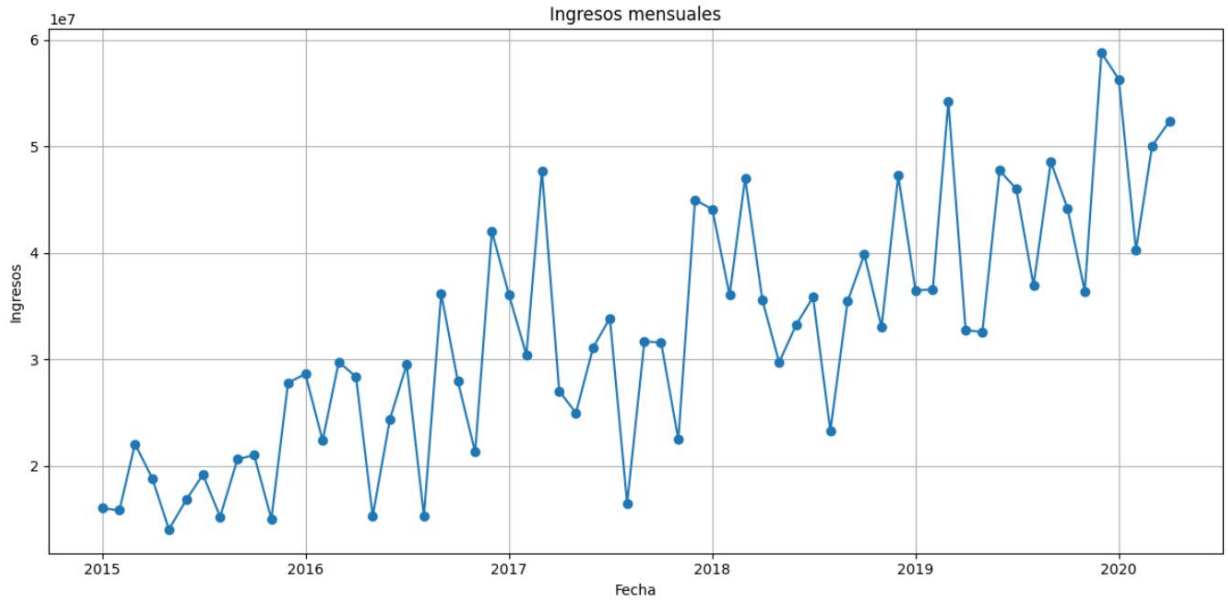
	Ingresos
count	64.00
mean	32360452.26
std	11641498.54
min	14021479.61
25%	22426546.79
50%	32090875.10
75%	39929985.09
max	58756473.66

Interpretación: el promedio de los ingresos mensuales (32.36 millones) y su mediana (32.09 millones) son muy similares, lo que indica una distribución pareja. Además, la desviación estándar (11.64 millones) es menor que la media, lo que sugiere que los ingresos no presentan una gran dispersión. El rango intercuartílico (IQR), que contiene el 50% central de los datos, abarca ingresos entre 22.4 y 39.9 millones de balboas, lo que refuerza la idea de una distribución bastante pareja.

3. Visualización y descomposición de la serie

Separo la variable de interés y la visualizo de forma general, después aplico descomposición para evaluar sus componentes como una serie temporal.

3.1 Gráfico general



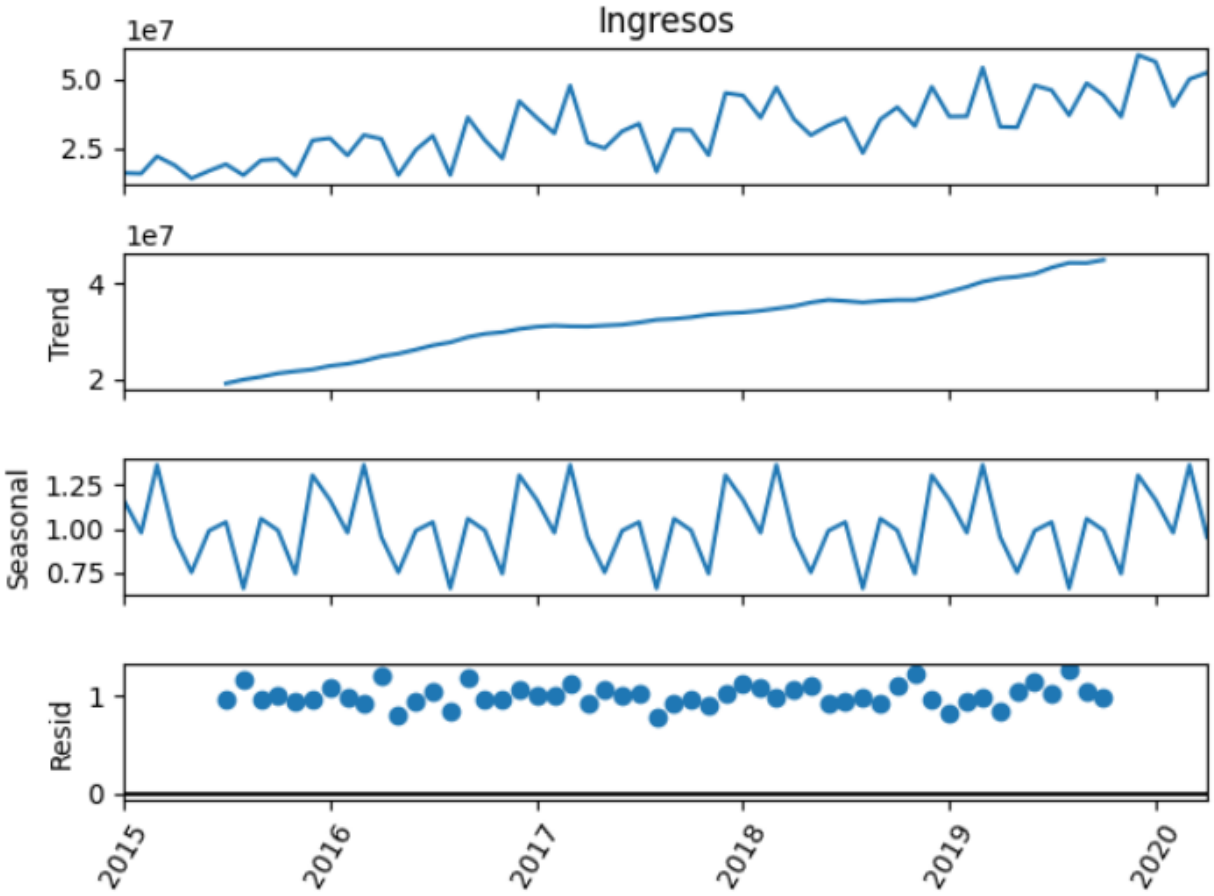
Interpretación: se observa cómo han ido cambiando los ingresos mensuales desde el año 2015 hasta principios del 2020.

De manera general, se puede determinar que:

- se determina que la serie sigue un modelo del tipo multiplicativo por el cambio de tamaño proporcional de cada pico,
- la empresa parece estar creciendo,
- los picos y valles se denotan fuertemente, como si cada año tuviera sus meses buenos y malos. Eso podría deberse a que el negocio tiene temporadas altas y bajas (por ejemplo, más ventas en ciertos meses del año).

3.2 Descomposición

La descomposición es un proceso en el que se aplican fórmulas a la serie para poder visualizar de forma separada la tendencia, estacionalidad y el ruido.



Interpretación:

En el primer gráfico se muestra la serie original

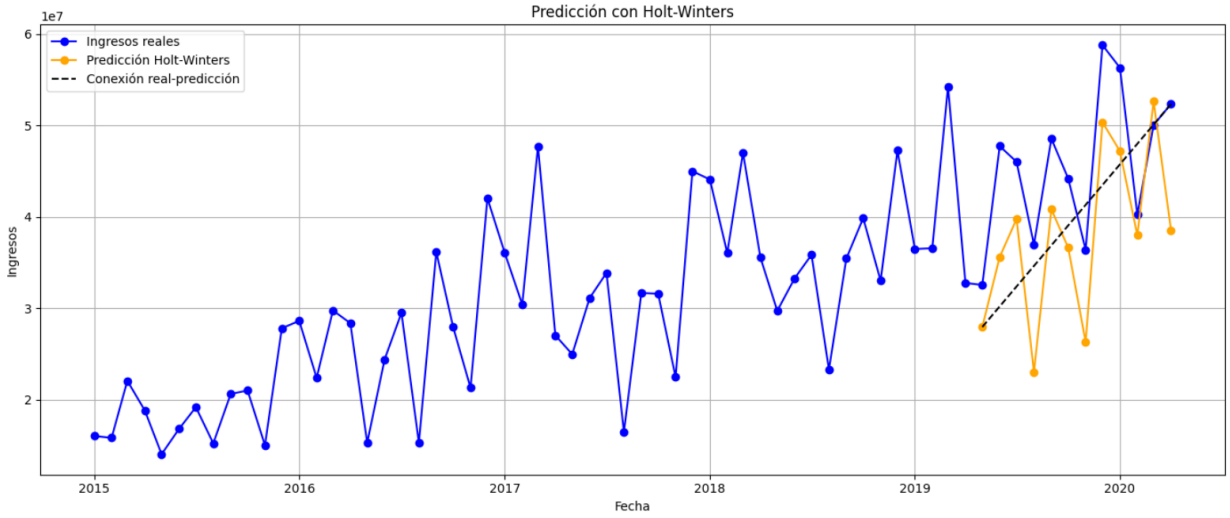
- el gráfico de tendencia muestra la tendencia ascendente marcada
- el tercer gráfico muestra que hay subidas y bajadas parecidas todos los años, lo que indica que el negocio tiene meses fuertes y meses flojos regularmente. Significa que hay estacionalidad.
- el gráfico de residuos indica los errores o parte de la data real que no puede ser explicada por la tendencia o la estacionalidad. La presentación uniforme indica que hubo muy pocos eventos externos que afectaron los ingresos.

4. Modelo Holt – Winter

Es un modelo de pronóstico que sirve cuando los datos presentan una tendencia clara y un patrón estacional. Ayuda a proyectar lo que viene con base en el comportamiento pasado.

4.1 Modelado

Primero dividí la serie en datos de entrenamiento y prueba tomando solo los últimos 12 meses como prueba. Seguido entreno el modelo y hago predicciones sobre los datos de prueba.



Interpretación: se observa que el modelo capta la forma general de la serie (tendencia creciente y estacionalidad), aunque en los primeros 9 meses de predicción, se subestima sistemáticamente los ingresos reales. Es decir, el modelo predice valores más bajos de los que realmente ocurrieron. Solo en los meses 10 y 11, la predicción parece estar más alineada con los datos reales. Sin embargo, en el mes 12, el modelo nuevamente falla al predecir una caída fuerte que no se refleja en el dato real, lo que evidencia una subestimación significativa.

Ahora se aplican los errores de pronóstico, estas son métricas para evaluar la efectividad del modelo.

MAE (Mean Absolute Error): el **promedio de los errores absolutos** entre lo real y lo predicho (diferencia). Dice, en unidades (por ejemplo, balboas), cuánto se equivoca en promedio el modelo sin importar si se pasa o se queda corto. Es fácil de entender y comparar.

RMSE (Root Mean Squared Error): la **raíz del error cuadrático** medio es parecido al MAE, pero si el modelo se equivoca por mucho en algunos puntos, este valor sube bastante. O sea, da más importancia a los errores grandes. Es útil si queremos que el modelo no se equivoque demasiado en ningún momento, aunque puede exagerar si hay datos muy raros o extremos.

MAPE (Mean Absolute Percentage Error): el **error porcentual absoluto medio** mide el error en porcentaje, comparando el error con el valor real. Dice cuánto se equivoca el modelo en porcentaje. Es muy útil porque permite entender la precisión sin importar la escala de los datos.

MSE (Mean Squared Error): el error cuadrático medio es una forma de medir el error del modelo, parecida al RMSE, pero sin sacar la raíz cuadrada al final. Eso hace que el número final sea más grande y esté en unidades al cuadrado (por ejemplo: balboas^2), por eso es más difícil de interpretar directamente.

Resultados del modelo Holt - Winters

```
MAE  =  B/. 8207976.86
RMSE =  B/. 9023700.69
MAPE =  18.28%
MSE  =  B/. 81427174139721.86
```

Interpretación:

El MAE indica que el modelo, en promedio, se equivoca por unos 8.2 millones de balboas.

El RMSE muestra que los errores más fuertes pueden llegar a unos 9 millones de balboas.

El MAPE dice que el modelo se equivoca en un 18.28% respecto al valor real. Se considera que el modelo tiene un margen de error aceptable.

El MSE es un número muy grande porque eleva los errores al cuadrado, y aunque no se interpreta fácilmente, se usa para cálculos técnicos dentro del modelo.

5. Modelo SARIMA

Es un modelo que ayuda a predecir el futuro usando datos del pasado. Es muy útil cuando hay estacionalidad.

5.1 Transformación de datos

Como primer paso para aplicar SARIMA, es la transformación de los datos para llevarlos a una escala lo más uniforme posible. Primero aplico la prueba estadística Dicky – Fuller que ayuda a confirmar que serie es estable en varianza y tendencia.

```
H0: La serie no es estacionaria
H1: La serie es estacionaria

p-valor < 0.05; se rechaza H0
p-valor >= 0.05; se acepta H0

ADF Statistic: -0.26914893564362824
p-value: 0.9297615379617789
```

Interpretación: debido a que el p-valor es mayor a 0.05, se afirma que la serie tiene una varianza muy grande.

Hay que transformar los datos para hacerla más estable, por lo que se le aplica la técnica de transformación Box-Cox para hacerla más estable en varianza (hacer que sus valores no estén tan alejados unos de otros) y diferenciación para la tendencia. Y finalmente aplicar de nuevo la prueba Dicky – Fuller para evaluar su estacionariedad.

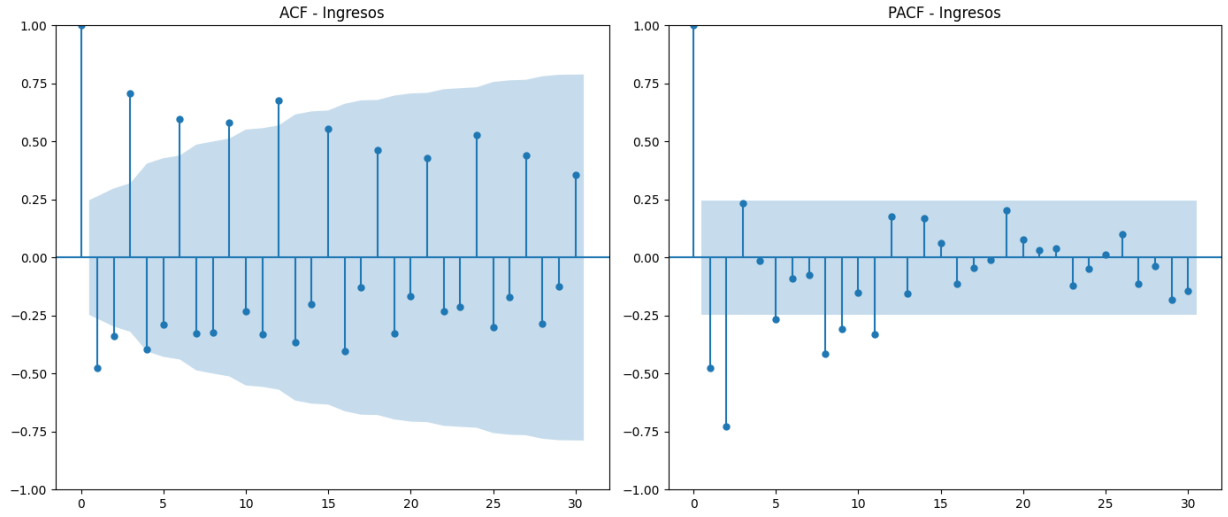
```
H0: La serie no es estacionaria
H1: La serie es estacionaria

p-valor < 0.05; se rechaza H0
p-valor >= 0.05; se acepta H0

ADF Statistic: -5.640943860760583
p-value: 1.0368850009076197e-06
```

Interpretación: el p-valor es extremadamente pequeño y menor a 0.05 por lo que se confirma que la serie ya es estacionaria.

5.2 Gráficos de Autocorrelación y Autocorrelación Parcial

La ACF (Autocorrelación) muestra cómo los datos actuales se relacionan con sus valores pasados, considerando todas las influencias anteriores. En cambio, la PACF (Autocorrelación Parcial) mide la relación directa entre un dato actual y uno pasado, eliminando el efecto de los intermedios.

Este paso es fundamental para determinar los rezagos o hasta cuántos meses anteriores debemos indicarle al SARIMA.



Interpretación:

ACF: Se observa que los ingresos del mes actual tienen una relación significativamente fuerte con los ingresos de hace 3 meses, aunque esta relación parece estar influenciada por una correlación inversa con los ingresos de los dos meses anteriores. Entonces se puede tomar hasta 2 errores anteriores, ya que se muestran significativas (rebasan el intervalo de confianza).

PACF: muestra que los ingresos actuales están influenciados por algunos meses pasados, especialmente los últimos 2 meses, y también posiblemente por lo que pasó hace 11 meses. Para este caso estaría usando hasta 2 meses anteriores porque presentan una correlación significativa con el mes actual.

5.3 Modelado

Para este proyecto se prueba un modelo SARIMA con los parámetros (2, 1, 1) (1, 1, 0, 5)

Decido predecir a mediano plazo, por lo que elijo una estacionalidad de 5 meses.

Componente no estacional (ARIMA):

- Rezagos autorregresivos (AR): elijo el valor de hace 2 meses
- Diferenciaciones (I): resto un valor para minimizar tendencia.
- Media Móvil (MA): selecciono el error del mes pasado.

Componente estacional (SARIMA):

- Rezagos autorregresivos estacionales (SAR): valores de hace **s** meses pasados
- Diferenciaciones estacionales (I): una diferencia del valor actual con el de hace **s** periodos.
- Media Móvil Estacional (MA): para este modelo no selecciono ningún error.
- s: para este caso, selecciono una estacionalidad de 5 meses.

SARIMAX Results

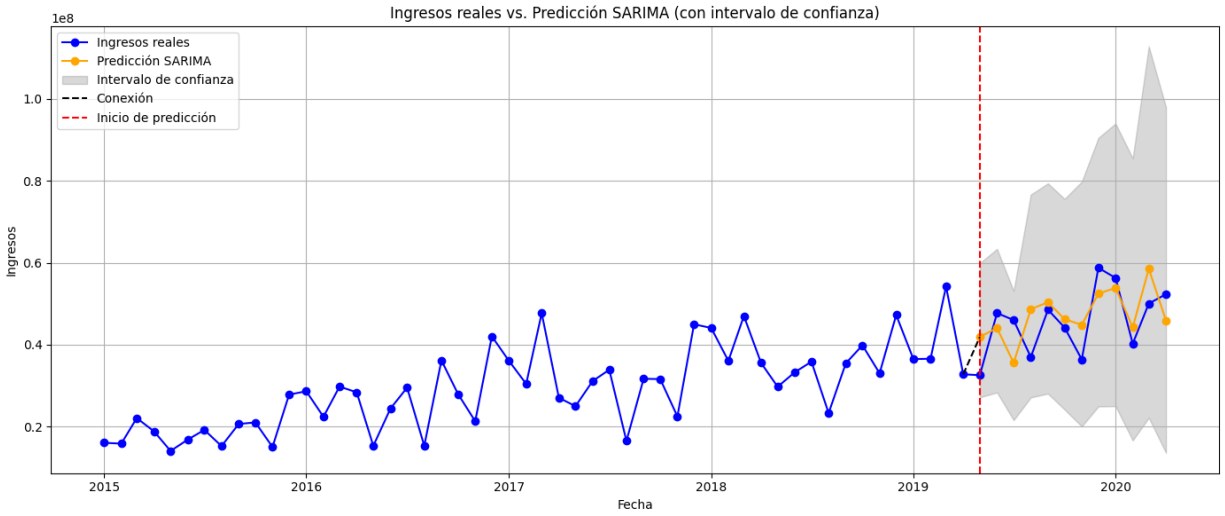
Dep. Variable:	Ingresos_lambda_optimo	No. Observations:	64
Model:	SARIMAX(2, 1, 1)x(1, 1, [], 5)	Log Likelihood	-424.186
Date:	Sat, 12 Apr 2025	AIC	858.372
Time:	21:13:08	BIC	868.031
Sample:	01-01-2015	HQIC	862.063
	- 04-01-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.9727	0.093	-10.451	0.000	-1.155	-0.790
ar.L2	-0.8885	0.071	-12.575	0.000	-1.027	-0.750
ma.L1	0.2882	0.210	1.374	0.170	-0.123	0.699
ar.S.L5	-0.7292	0.141	-5.187	0.000	-1.005	-0.454
sigma2	9.763e+05	2e+05	4.891	0.000	5.85e+05	1.37e+06

Ljung-Box (L1) (Q):	0.35	Jarque-Bera (JB):	0.23
Prob(Q):	0.55	Prob(JB):	0.89
Heteroskedasticity (H):	0.37	Skew:	-0.15
Prob(H) (two-sided):	0.05	Kurtosis:	2.88

Interpretación: Este modelo fue entrenado para predecir los ingresos mensuales, considerando tanto los valores recientes como los que ocurrieron hace 5 meses (ciclo estacional). Los coeficientes de los rezagos (ar.L1, ar.L2 y ar.S.L5) son significativos (muy cercanos a cero), lo que indica que esos valores pasados realmente ayudan a mejorar las predicciones.

Resultado del entrenamiento




```
MAPE dentro de la muestra: 11.04%
MAE: 941.27
RMSE: 1,198.63
MSE: 1,436,706.33
```

Interpretación: El modelo SARIMA con estacionalidad de 5 meses captura bien el comportamiento general de la serie, tanto en tendencia como en los ciclos. Presenta un buen desempeño dentro del conjunto de entrenamiento, y aunque las predicciones fuera de la muestra tienen mayor incertidumbre (lo cual es natural), siguen la trayectoria general de la serie.

Las métricas de error confirman que es un modelo aceptable para análisis a mediano plazo:

MAPE: 11.04%, indica que el modelo se equivoca un 11%, lo cual es aceptable.

MAE, RMSE y MSE: estos errores no presentan valores muy altos.

Conclusiones

- El modelo SARIMA con estacionalidad de 5 meses logró captar tanto la tendencia como los ciclos estacionales de los ingresos mensuales de la empresa.
- Se evaluó el desempeño de los modelos con la **media del error porcentual absoluto** del SARIMA (11.04%) contra la del Holt – Winter (18.28%) por lo que se confirma que SARIMA tiene mejor efectividad.
- El análisis puede apoyar decisiones estratégicas en planificación financiera, manejo de inventario y gestión del personal.
- Prever los ingresos mensuales también aporta valor en la gestión del riesgo de mora, ya que permite anticipar posibles caídas en la liquidez de los clientes.