

# ***Proyecto Final: Desempleo y sus Efectos en el Desarrollo Económico y Social en México***

## **1. Contextualización del problema: La Persistencia del Desempleo y su Impacto en México**

En México, el desempleo representa una de las formas más claras de exclusión económica y vulnerabilidad social. En México, aunque se han implementado políticas para promover la inclusión laboral, persisten desafíos como altos niveles de informalidad, desajustes entre la oferta y demanda de habilidades, disparidades regionales, y un crecimiento económico que no siempre se traduce en generación de empleo formal. Esta situación afecta de manera desproporcionada a jóvenes, mujeres y personas con menor nivel educativo.

El desempleo actúa como una barrera crítica para que millones de personas accedan a ingresos dignos, seguridad social y oportunidades reales de desarrollo. En este contexto, entender su comportamiento y anticipar su evolución es clave para diseñar respuestas efectivas desde las políticas públicas.

### **1.1 Indicador: Tasa de Desocupación**

La tasa de desocupación mide el porcentaje de personas en edad y disposición de trabajar que buscan empleo activamente sin encontrarlo. En México, este indicador se ha registrado de forma continua desde 2005 mediante la Encuesta Nacional de Ocupación y Empleo (ENOE), lo que permite construir una serie de tiempo lo suficientemente extensa para analizar tendencias de largo plazo, estacionalidades y reacciones a choques externos como la pandemia por COVID-19 o el huracán Otis.

El análisis de esta serie mediante modelos estadísticos (como ARIMA o SARIMA) permite pronosticar escenarios futuros, evaluar el impacto de políticas previas y anticipar situaciones críticas que requieren intervención. Así, el indicador no solo describe el estado actual del mercado laboral, sino que se convierte en una herramienta estratégica para la toma de decisiones públicas.

### **1.2 Contribución del indicador a los Objetivos de Desarrollo Sostenible (ODS)**

El análisis de este indicador apunta a dos objetivos de desarrollo sostenible (ODS) en particular:

### ***ODS 1: Fin de la pobreza***

El desempleo es uno de los principales factores que perpetúan la pobreza. Quienes carecen de empleo formal enfrentan mayores dificultades para obtener ingresos estables, protección social y satisfacer sus necesidades básicas. El seguimiento y análisis de la tasa de desocupación permite:

- Identificar grupos en mayor riesgo de caer en pobreza por pérdida de empleo.
- Evaluar la efectividad de programas de inclusión laboral y subsidios.
- Diseñar políticas que reduzcan la precariedad y promuevan la movilidad social.

### ***ODS 8: Trabajo decente y crecimiento económico***

Este objetivo promueve no solo la creación de empleo, sino que sea inclusivo, digno y productivo. La tasa de desocupación actúa como un estimador directo del mercado laboral. Su análisis contribuye a:

- Detectar obstáculos en la generación de empleo.
- Medir el impacto real de las políticas económicas sobre la población.
- Planificar programas de formación y reconversión laboral ajustados a la realidad del mercado.

La medición y análisis de la tasa de desocupación no solo cumple una función estadística, sino que se alinea directamente con los objetivos de desarrollo sostenible, contribuyendo a la formulación de políticas públicas más justas y efectivas.

## **1.3 Descripción de los datos**

Para el análisis del desempleo en México se utilizarán los datos provenientes de la Encuesta Nacional de Ocupación y Empleo (ENOE), generada por el Instituto Nacional de Estadística y Geografía (INEGI). Como se mencionó, esta encuesta se ha aplicado de manera continua desde 2005 y constituye la principal fuente oficial de información sobre el mercado laboral en el país.

La variable de interés principal es la tasa de desocupación nacional, la cual se define como el porcentaje de la Población Económicamente Activa (PEA) que indica durante el periodo de referencia, el porcentaje de la población que no tenía empleo pero estaba buscando activamente uno. Esta variable se encuentra disponible con frecuencia mensual, lo que permite construir una serie de tiempo suficientemente larga y detallada para realizar un buen análisis.

Las observaciones comprenden el periodo de enero de 2005 hasta la fecha más reciente disponible, que es marzo de 2025, abarcando así distintos contextos económicos, incluyendo:

- La crisis financiera global de 2008-2009,
- La pandemia de COVID-19 (2020-2022),

- Eventos recientes como el huracán Otis (2023).

El conjunto de datos se encuentra desagregado por sexo, grupos de edad y entidad federativa, sin embargo, para este análisis se considerará la tasa de desocupación total nacional, sin distinción de subgrupos, a fin de concentrarse en los patrones generales de comportamiento.

## 2. Extracción, Limpieza y Preprocesamiento de Datos

Como se mencionó en la sección anterior, el objetivo es analizar la evolución de la tasa de desocupación en México. Para esto se utilizaron datos mensuales provistos por el INEGI, que abarcan el periodo de enero de 2005 a marzo de 2025. El archivo original fue importado manualmente en formato CSV.

### 2.1 Importación de datos

Inicialmente, se importaron las bibliotecas necesarias para el análisis, visualización y modelado de series de tiempo. Posteriormente, se cargó el archivo “Tasa\_de\_Desocupacion\_INEGI.csv”, el cual contiene dos columnas:

- **Periodo:** correspondiente al mes y año de la observación, en formato de texto y abreviado en español (por ejemplo, "2020 Ene").
- **Porcentaje:** representa la tasa de desocupación mensual como un valor porcentual.

Se verificó que el conjunto contenía 244 observaciones sin valores nulos, pero ambas columnas habían sido leídas como tipo objeto, por lo que fue necesario convertirlas a sus respectivos formatos adecuados para análisis temporal y numérico.

### 2.2 Conversión de fechas

Dado que la columna de fechas estaba en español y no era reconocida directamente por las funciones de conversión de fechas de Python, se implementó una traducción de las abreviaturas de los meses del español al inglés. Una vez traducidos, los valores de año y mes se unificaron en una nueva columna, la cual fue convertida exitosamente al formato de fecha (datetime) de pandas, asignando por defecto el primer día del mes.

### 2.3 Conversión de variables numéricas

La columna que contenía los porcentajes de desocupación también fue transformada al tipo numérico adecuado automáticamente, permitiendo su manipulación matemática y estadística.

## 2.4 Construcción del conjunto final

Se construyó un nuevo Data Frame que contenía únicamente las dos variables relevantes:

- **Periodo:** en formato de fecha, ordenado cronológicamente de la más antigua a la más reciente.
- **Porcentaje:** tasa de desocupación mensual ya convertida en valores numéricos.

Con esta estructura final, los datos quedaron listos para iniciar el análisis exploratorio y modelado.

## 3. Análisis Exploratorio de Datos (EDA)

En esta sección se realiza un análisis exploratorio para comprender la estructura, distribución y comportamiento general de los datos sobre la tasa de desocupación mensual en México entre enero de 2005 y marzo de 2025. El objetivo es detectar posibles problemas en los datos, obtener un panorama general de la variable principal y visualizar tendencias temporales relevantes.

### 3.1 Análisis de datos nulos

Se examinó la cantidad de valores faltantes por columna en la base de datos:

- Se confirmó que no existen valores nulos o ausentes en ninguna columna del dataset.
- Por lo tanto, no fue necesario realizar ningún tipo de imputación o limpieza adicional relacionada con datos faltantes.

### 3.2 Estadísticas descriptivas generales

Para obtener una primera caracterización de la variable **tasa de desocupación (%)**, se calcularon las estadísticas descriptivas básicas como como media, mediana, desviación estándar, mínimo, máximo y percentil:

Estadísticas Descriptivas de la Tasa de Desocupación (%)	
Estadística	Porcentaje

Conteo	243.000000
Media	4.031646
Desviación estándar	0.900073
Valor mínimo	2.220000
25%	3.360000
50%	3.880000
75%	4.785000
Valor máximo	6.340000

- El conjunto de datos contiene un total de **243 observaciones** completas.
- La tasa promedio de desocupación en el periodo analizado fue de **4.03%**, lo que indica una tasa moderada, que la mayoría de las personas que buscan empleo lo encuentran, aunque podría reflejar subempleo o personas que dejaron de buscar trabajo.
- La desviación estándar de **0.9%** sugiere una variabilidad moderada a lo largo del tiempo.
- Los valores mínimo y máximo fueron **2.22%** y **6.34%**, respectivamente, probablemente relacionados con crisis económicas o eventos inesperados.
- Los percentiles indican que el 50% de los datos se encuentran alrededor del 3.88%, con el 25% de los datos por debajo de 3.36% y el otro 25% por encima de 4.79% (rango intercuartílico de 3.36% a 4.79%).

Esta información inicial proporciona una visión clara de la distribución y comportamiento de la variable de interés.

### 3.3 Estadísticas descriptivas por año

Se calcularon nuevamente las estadísticas descriptivas, esta vez agrupando los datos por año para analizar la evolución anual de la tasa de desocupación, lo que permite identificar patrones, picos y posibles ciclos económicos.

<i>Estadísticas Descriptivas de la Tasa de Desocupación (%) por Año</i>								
Periodo	Tasa de Desocupación (%)							
	conteo	media	desv est.	min	25%	50%	75%	max
2005	12.0	3.525000	0.410554	2.70	3.3700	3.635	3.7225	4.06
2006	12.0	3.564167	0.344475	2.93	3.3075	3.565	3.8850	4.05

2007	12.0	3.656667	0.324719	3.17	3.3400	3.780	3.8500	4.16
2008	12.0	3.902500	0.383693	3.28	3.5325	4.055	4.1525	4.50
2009	12.0	5.395000	0.570685	4.68	5.0600	5.220	5.9125	6.34
2010	12.0	5.308333	0.343163	4.78	4.9925	5.360	5.6225	5.83
2011	12.0	5.195000	0.406437	4.48	4.9825	5.250	5.4525	5.78
2012	12.0	4.926667	0.263485	4.43	4.7875	4.940	5.0325	5.34
2013	12.0	4.942500	0.339307	4.27	4.7600	5.040	5.1700	5.38
2014	12.0	4.838333	0.424175	3.76	4.7475	4.840	5.0875	5.49
2015	12.0	4.350833	0.282922	3.86	4.2225	4.430	4.5175	4.72
2016	12.0	3.877500	0.265300	3.38	3.7300	3.950	4.0375	4.26
2017	12.0	3.420000	0.150151	3.15	3.3275	3.445	3.5275	3.60
2018	12.0	3.320000	0.166406	2.94	3.2425	3.360	3.4075	3.58
2019	12.0	3.490833	0.252387	2.91	3.3875	3.530	3.6950	3.78
2020	12.0	4.431667	0.815273	2.91	3.7875	4.520	5.1625	5.49
2021	12.0	4.136667	0.370462	3.51	3.9300	4.100	4.3850	4.70
2022	12.0	3.273333	0.316783	2.76	3.0150	3.320	3.4550	3.74
2023	12.0	2.795833	0.200384	2.39	2.6950	2.785	2.9375	3.13
2024	12.0	2.670833	0.234616	2.28	2.4875	2.630	2.8675	3.04
2025	3.0	2.476667	0.241730	2.22	2.3650	2.510	2.6050	2.70

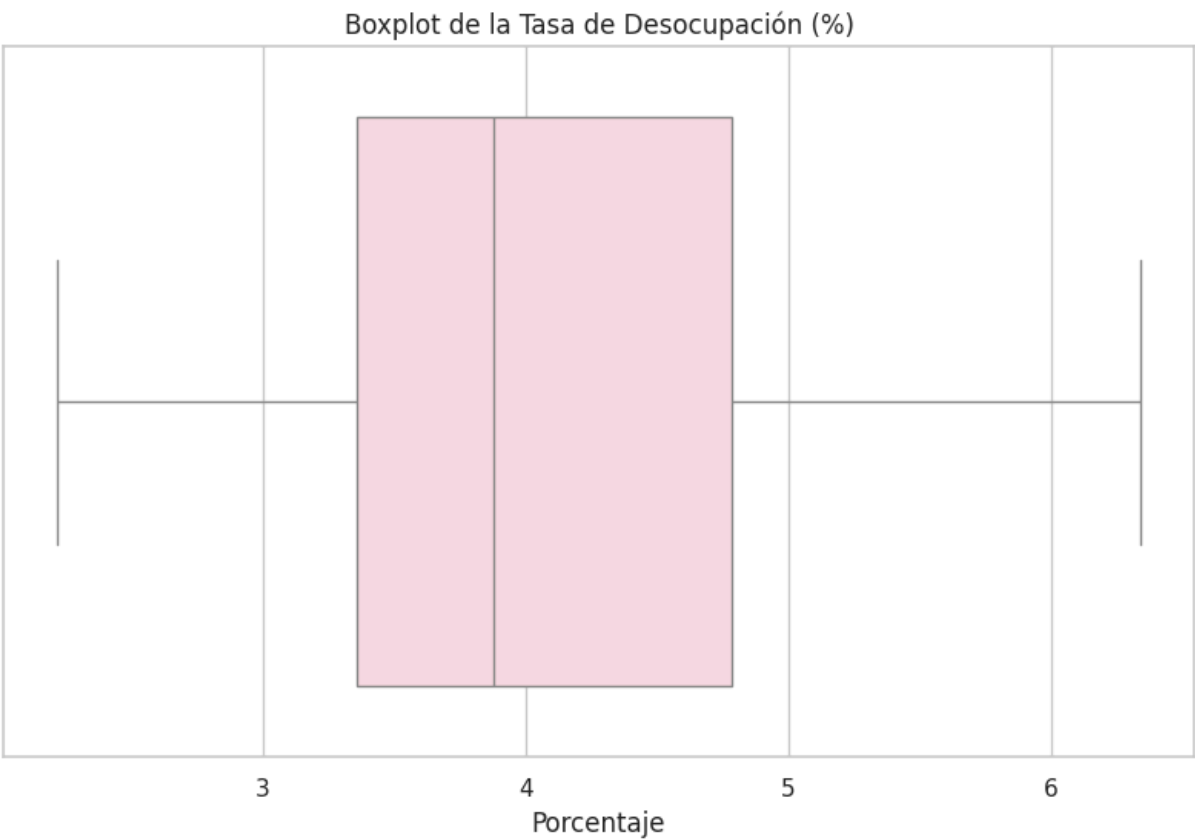
- **2009–2011 (Crisis financiera global):** Se observó un aumento significativo en la tasa de desocupación, alcanzando un máximo histórico de 6.34% y la media anual más alta de 5.39%. Este periodo coincide con la crisis financiera global de 2008-2009.
- **2012–2019 (Estabilización):** Posterior a la crisis, la tasa mostró una tendencia de estabilización con medias anuales cercanas a un mínimo histórico de 3.32% previo a la pandemia.
- **2020 (Efecto pandemia):** La tasa de desocupación se incrementó abruptamente a una media anual de 4.43% y un valor máximo de 5.49% en el año 2020, reflejando el impacto económico causado por la COVID-19.
- **2021–2022 (Recuperación):** Se observa una rápida recuperación, con tasas descendiendo a 4.13% y 3.27% respectivamente.

- **2023–2025 (Tasas mínimas):** Finalmente, las tasas más bajas del periodo se registraron entre 2023 a la actualidad, con promedios entre 2.47% y 2.79%, probablemente debido a la reactivación económica y ajustes metodológicos post pandemia.

Este análisis anual confirma la influencia clara de eventos económicos globales y nacionales en la evolución del desempleo.

## 2.4 Análisis de valores atípicos mediante boxplot

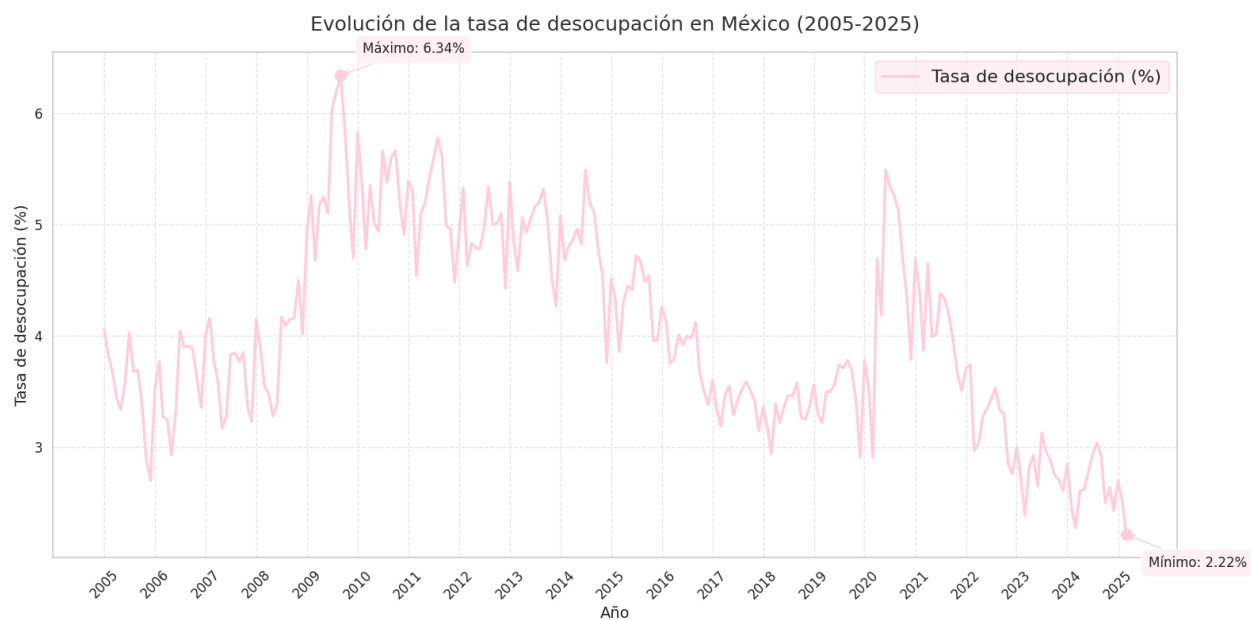
Se generó un boxplot para la variable tasa de desocupación (“Porcentaje”) con el fin de identificar posibles valores atípicos.



- No se detectaron valores fuera de los rangos esperados, lo que indica ausencia de valores atípicos en la serie.
- Los valores máximos y mínimos corresponden a eventos económicos identificados (crisis y pandemia) por lo que son consistentes con el contexto histórico.
- La mediana se encuentra cercana al 4%, en línea con las estadísticas descriptivas calculadas.

### 3.5 Visualización de la serie temporal

Finalmente, se graficó la serie temporal de la tasa de desocupación mensual en México desde 2005 hasta 2025 para visualizar su evolución, tendencias y eventos relevantes.



Los comportamientos identificados en las estadísticas descriptivas analizadas se reflejan al graficar la serie de tiempo:

- Se observa claramente el aumento abrupto entre 2009 y 2011, consecuencia de la crisis financiera global.
- Posteriormente, la tasa muestra estabilización hasta 2019, seguida por un aumento significativo en 2020 debido a la pandemia.
- En los años posteriores, se aprecia una rápida recuperación con una tendencia descendente hacia el nivel más bajo registrado en 2025.
- Se resaltan visualmente los puntos correspondientes al valor máximo y mínimo para facilitar su identificación.

Esta visualización complementa el análisis cuantitativo previo y confirma la presencia de ciclos económicos y eventos extraordinarios en la evolución del desempleo en México.

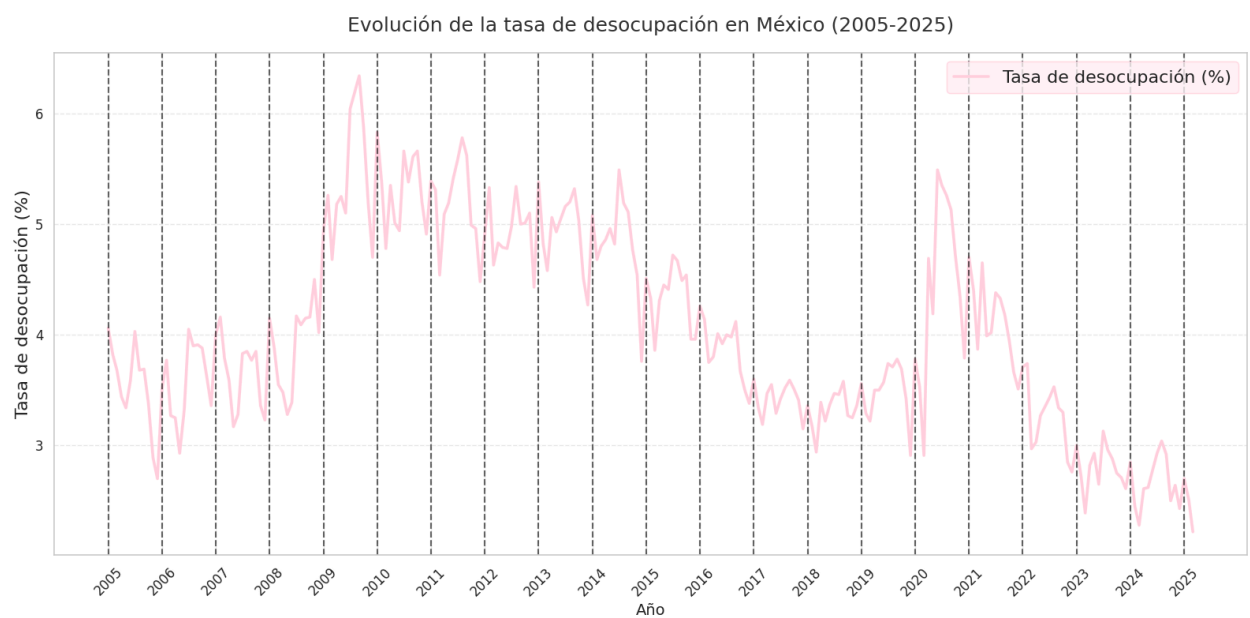
### 4. Análisis de la Estructura de la Serie de Tiempo

Con el fin de comprender la dinámica subyacente de la tasa de desocupación, en esta sección se analiza la estructura interna de la serie temporal, especialmente la presencia de componentes como **tendencia**, **estacionalidad** y **ruido**. Este análisis constituye un paso crucial previo a la modelación, ya que permite establecer supuestos adecuados sobre la naturaleza de los datos.



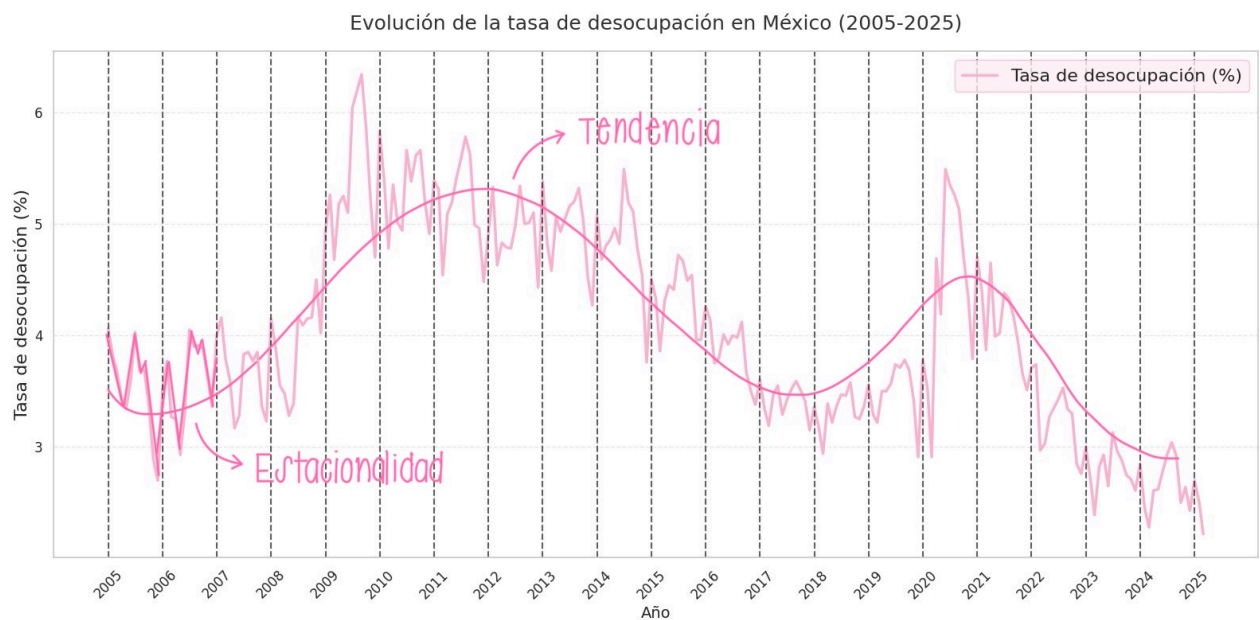
## 4.1 Análisis Visual Exploratorio

Para facilitar la identificación de patrones temporales, se graficó la serie temporal de la tasa de desocupación entre 2005 y 2025, añadiendo líneas verticales que marcan el inicio de cada año (1ro de enero). Esta representación permite observar de forma intuitiva la posible presencia de patrones recurrentes a lo largo de los años, lo cual podría evidenciar estacionalidad en la serie.



### Observaciones clave:

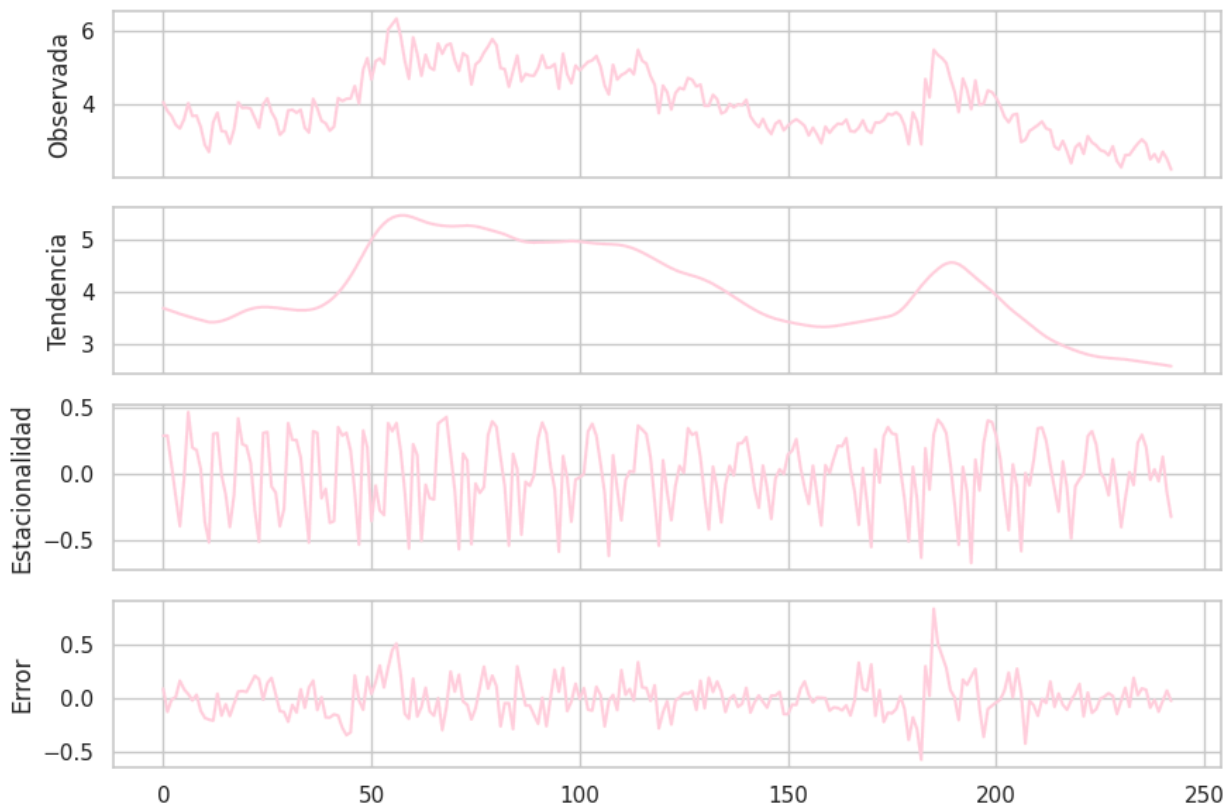
- Al analizar los intervalos entre las líneas, se identifica un patrón relativamente recurrente que tiende a repetirse anualmente: la tasa de desocupación suele ser relativamente alta al inicio del año, disminuye durante los meses siguientes, presenta algunos picos intermedios, y vuelve a aumentar ligeramente hacia el fin de año. Aunque este comportamiento no es marcadamente fuerte, su recurrencia sugiere la existencia de una **componente estacional**.
- Además, la trayectoria a largo plazo de la serie muestra una **tendencia decreciente de largo plazo**, con ciertos **picos abruptos** que coinciden con eventos económicos críticos, como la crisis financiera global o la pandemia por COVID-19. Esta evolución no lineal podría asemejarse a una función polinómica de orden superior o a un patrón suavemente oscilatorio.
- Dado que la serie muestra una tendencia general decreciente visible en vez de una tendencia constante como una línea recta horizontal a lo largo del tiempo sugiere que la serie **no es estacionaria**, lo que refuerza la necesidad de descomponerla para estudiar sus componentes individualmente.



## 4.2 Descomposición de la Serie Temporal

Para analizar en mayor detalle la estructura interna de la serie, se aplicó un proceso de descomposición aditiva mediante STL (Seasonal-Trend decomposition using Loess), con una periodicidad de 12 meses, adecuada para datos mensuales. Esta técnica permite descomponer la serie original en tres componentes fundamentales:

- **Tendencia:** captura la evolución a largo plazo de la tasa de desocupación.
- **Estacionalidad:** identifica patrones que se repiten con periodicidad anual (frecuencia mensual).
- **Residuo:** representa las fluctuaciones aleatorias no explicadas por la tendencia ni la estacionalidad.



### ***Análisis de componentes:***

- **Serie Observada:**

Representa la serie de tiempo de la tasa de desocupación original.ç

- **Tendencia:**

Se observa una curva suavizada que sigue el comportamiento general de la serie. Esta componente muestra picos durante los periodos de crisis, una disminución gradual en años de recuperación, un repunte posiblemente asociado al periodo de pandemia y una rápida disminución reciente. La forma ondulante sugiere que una simple tendencia lineal no sería adecuada para modelar esta serie.

- **Estacionalidad:**

Se muestran patrones que, en general, tienden a repetirse cada año. Si bien es posible identificar un patrón estacional relativamente definido, este no se mantiene constante a lo largo de toda la serie. Específicamente, se observa que aproximadamente en la parte central de la gráfica, la magnitud del patrón estacional disminuye, para luego volver a incrementar.

- Esta variación en la amplitud del patrón podría indicar que la especificación **aditiva** utilizada en la descomposición no es del todo adecuada. En una descomposición aditiva, se asume que la estacionalidad es independiente de la tendencia (es decir, constante a lo largo del tiempo), lo cual no parece cumplirse en este caso. En cambio, lo observado

sugiere una posible relación **proporcional** entre la estacionalidad y la tendencia, lo que se asocia a un comportamiento **multiplicativo**.

- En modelos multiplicativos, la estacionalidad se amplifica o reduce en función del nivel de la tendencia (es decir, se modela como  $\text{tendencia} \times \text{estacionalidad}$ , en lugar de  $\text{tendencia} + \text{estacionalidad}$ ). Este tipo de comportamiento se refleja en patrones estacionales cuya amplitud varía conforme cambia la tendencia general de la serie, lo cual podemos observar en este caso.

- **Error:**

Este componente representa la parte de la variación que no puede explicarse ni por la tendencia ni por la estacionalidad. Idealmente, debería reflejar únicamente ruido aleatorio, es decir, variaciones sin un patrón específico.

- En este caso, si bien en la mayor parte de la serie parece observarse un comportamiento aleatorio, se identifica un pico notable en un punto particular de la gráfica. Este coincide con el periodo en el que también se reduce la magnitud del patrón estacional y la caída en la tendencia.

El análisis estructural revela que la serie presenta una tendencia descendente, estacionalidad anual relativamente definida, y residuos que podrían contener información no capturada. Para modelar correctamente comportamientos como los observamos sería necesario utilizar una descomposición multiplicativa u otras técnicas más avanzadas. Sin embargo, dado que estos métodos no forman parte del contenido de este curso, continuaremos con la metodología aditiva, considerando cómo este tipo de estructuras puede afectar el desempeño del modelo final.

## 5. Pruebas de Identificación del Modelo

Con el objetivo de modelar adecuadamente la serie de tiempo correspondiente al porcentaje mensual de desocupación, en esta sección evaluamos su estacionariedad, realizamos las transformaciones necesarias para estabilizar sus propiedades estadísticas, y analizamos la dependencia temporal mediante las funciones de autocorrelación y autocorrelación parcial. Estas herramientas nos permiten entender la estructura subyacente de la serie y orientar la elección del modelo más adecuado para su representación.

### 5.1 Identificación de Estacionariedad: Prueba ADF

En la sección anterior se observó visualmente que la serie presenta una tendencia en general decreciente, lo cual sugiere que no es estacionaria, ya que la media no se mantiene constante en el tiempo. Para confirmar formalmente esta hipótesis, se aplicó la **prueba aumentada de**

**Dickey-Fuller (ADF)**, una herramienta estadística utilizada para detectar la presencia de una raíz unitaria en la serie.

Para que una serie sea estacionaria, sus propiedades estadísticas deben ser invariantes en el tiempo, esto implica que:

- La media es constante en cualquier tiempo.
- La función de autocovarianza depende únicamente del retraso o diferencia entre observaciones, y no del tiempo absoluto.

La prueba ADF contrasta las siguientes hipótesis:

- **Hipótesis nula ( $H_0$ ):** La serie tiene una raíz unitaria, es decir, **no es estacionaria**.
- **Hipótesis alternativa ( $H_a$ ):** La serie **no tiene** raíz unitaria, es decir, **es estacionaria**.

El objetivo es rechazar  $H_0$ ; para ello, se requiere que el p-value sea menor a un nivel de significancia preestablecido (usualmente 0.05). Desde el punto de vista teórico, si el polinomio característico del proceso tiene raíces dentro del círculo unitario, el modelo puede representarse de diferentes maneras equivalentes (por ejemplo, como un modelo autorregresivo), lo que dificulta la identificación única de los parámetros del modelo.

Al aplicar la prueba ADF sobre la serie original correspondiente al porcentaje mensual de desocupación, se obtuvo un p-value de 0.5572, valor que excede el umbral de significancia. Por lo tanto, no se puede rechazar la hipótesis nula, lo que indica que la serie no es estacionaria. Este resultado coincide con el análisis visual previo, donde ya se observaba una tendencia presente en los datos.

## 5.2 Diferenciación

Dado que la serie no es estacionaria, se procedió a aplicar una diferenciación de primer orden, que consiste en calcular las diferencias entre valores consecutivos de la serie. Esta transformación tiene como objetivo eliminar la tendencia presente y estabilizar la media a lo largo del tiempo.

Tras realizar esta diferenciación, se volvió a aplicar la prueba ADF sobre la serie transformada para verificar si este procedimiento logró transformar la serie a una estacionaria. En esta ocasión, el p-value obtenido fue 0.0048, lo cual es menor al valor de significancia de 0.05. Esto indica que contamos con evidencia suficiente para rechazar la hipótesis nula, concluyendo que la serie sí se ha vuelto estacionaria tras la diferenciación de primer orden.

### 5.3 Función de Autocorrelación Muestral (ACF)

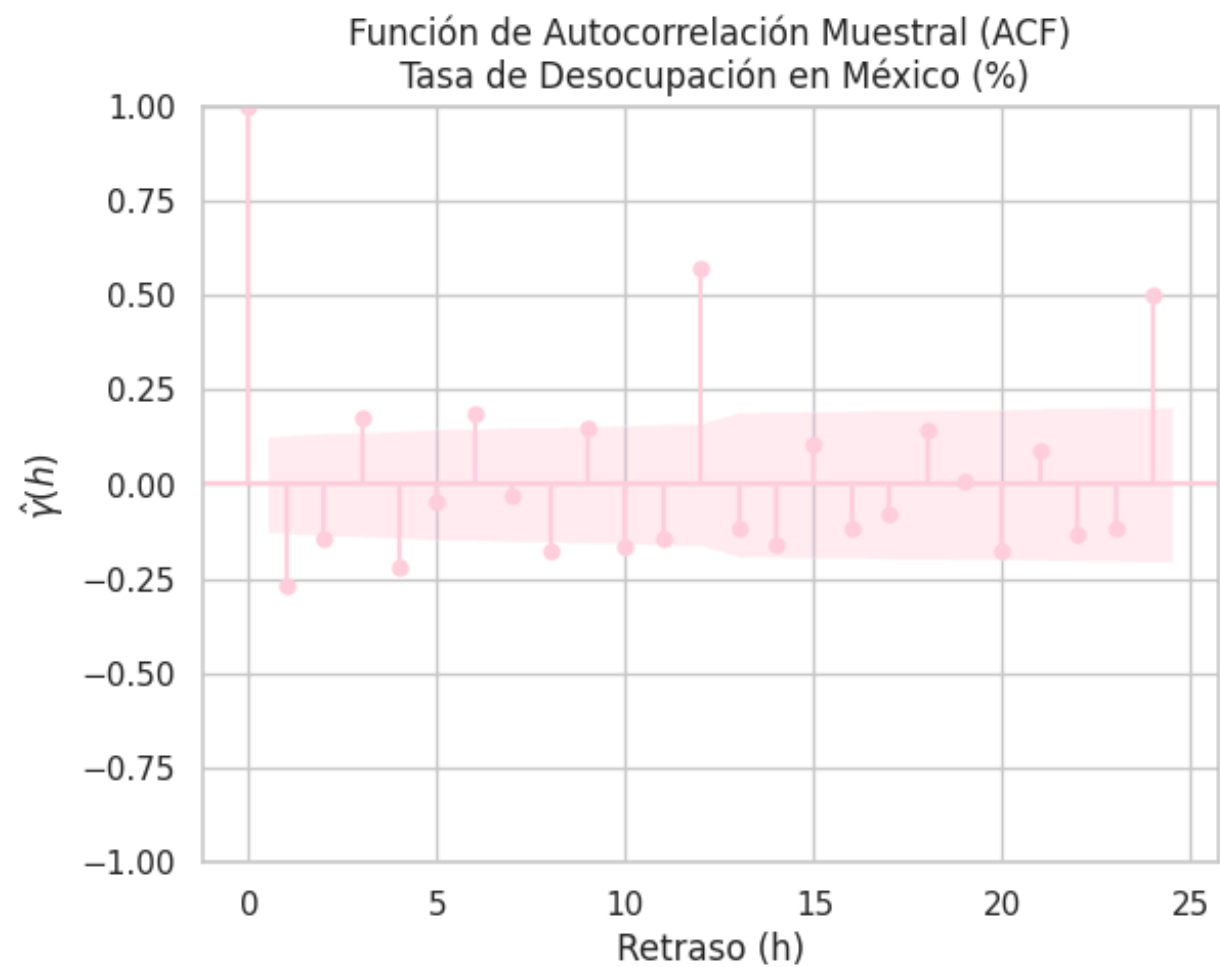
La **función de autocorrelación muestral (ACF)** nos permite analizar de forma práctica la dependencia entre observaciones separadas por distintos retrasos temporales (observaciones en distintos momentos del tiempo). Se define como:

$$r(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

donde:

- $\hat{\gamma}(h)$ : Es la estimación de la covarianza en el retraso  $h$ .
- $\hat{\gamma}(0)$ : Es la varianza muestral

Graficar esta función nos permite identificar a partir de qué retraso las observaciones dejan de estar correlacionadas significativamente. En modelos como los de media móvil,  $MA(q)$ , esta gráfica sugiere directamente el orden  $q$  del modelo. La gráfica de la ACF de la serie diferenciada es la siguiente:



Al graficar la ACF de la serie diferenciada se observa un comportamiento oscilante, donde los coeficientes suben y bajan de forma alternada. Puesto que varios coeficientes sobrepasan las

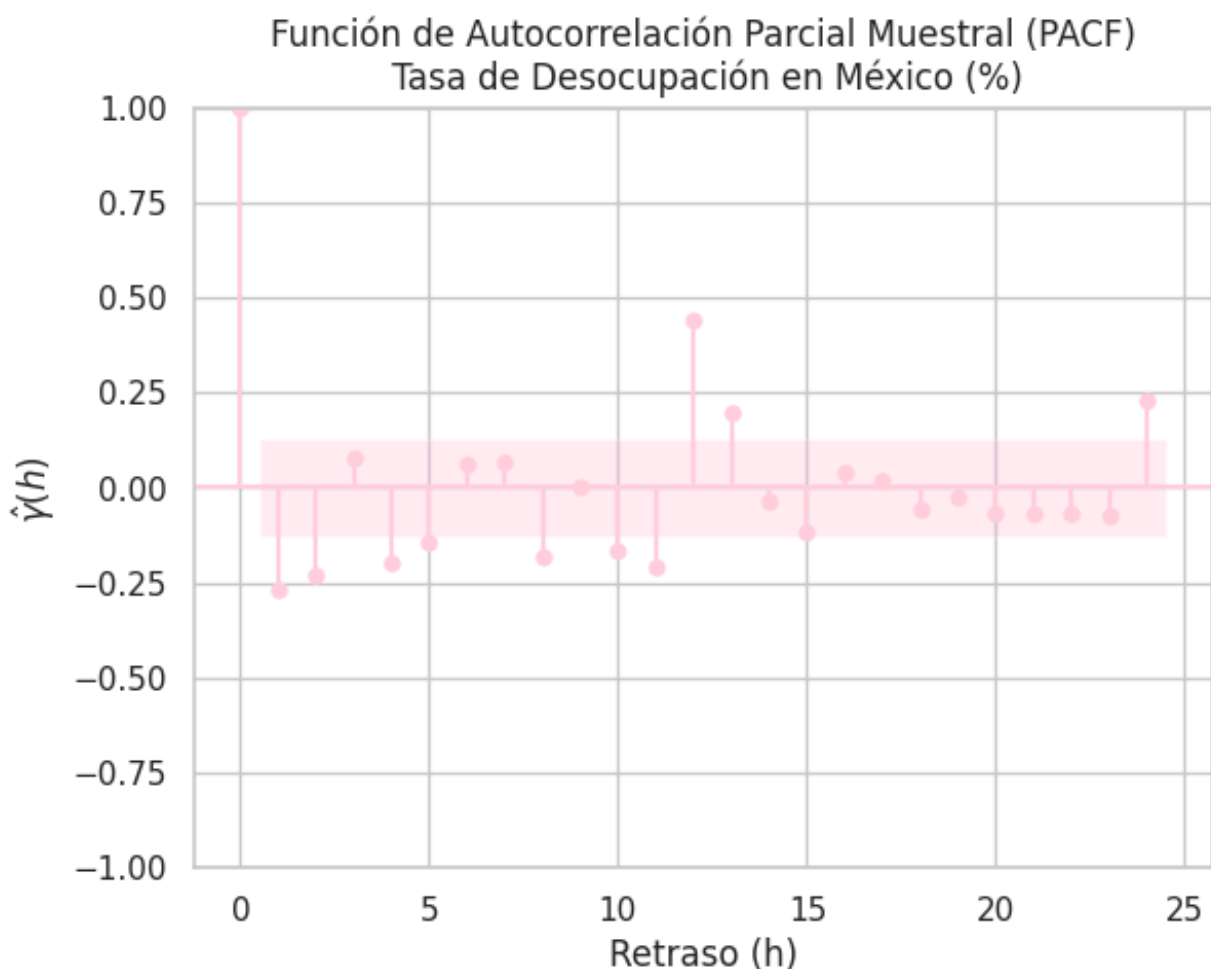
bandas de confianza, indica la presencia de autocorrelación significativa, lo que permite descartar que se trate de una caminata aleatoria, donde se esperaría que todos los coeficientes estén dentro de dichas bandas.

En modelos de media móvil,  $MA(q)$ , los coeficientes de autocorrelación suelen caer abruptamente dentro de las bandas de confianza (caen a cero) a partir del retraso  $q$ , lo cual no ocurre en este caso. Esto sugiere que un modelo puramente de media móvil no sería adecuado. Por otro lado, la ausencia de un decaimiento gradual claro o un patrón sinusoidal definido en la ACF también dificulta asumir un modelo autorregresivo,  $AR(p)$ , puro. Sin embargo, el hecho de que exista dependencia entre valores pasados sugiere la posible presencia de un componente autorregresivo.

Para determinar con mayor precisión la estructura del modelo, es necesario analizar también la función de autocorrelación parcial (PACF), ya que esta herramienta es clave para determinar si se trata de un proceso autorregresivo puro  $AR(p)$  o de una combinación de componentes autorregresivos y de media móvil.

## 5.4 Función de Autocorrelación Parcial Muestral (PACF)

A diferencia de la ACF, la **función de autocorrelación parcial muestral (PACF)** estima la correlación directa entre dos observaciones, eliminando el efecto de los retrasos intermedios. Esta función es fundamental para identificar el orden  $p$  de un modelo autorregresivo  $AR(p)$ . En estos modelos, se espera que los coeficientes significativos en la PACF desaparezcan abruptamente a partir del retraso  $p$ , revelando el orden verdadero del proceso. La gráfica de la PACF de la serie diferenciada es la siguiente:



En la PACF graficada para la serie diferenciada se observa nuevamente un **patrón ligeramente oscilante**, donde los coeficientes suben y bajan de forma alternada. El comportamiento no es lo suficientemente claro como para sugerir el uso de un modelo autorregresivo AR(p), donde se esperaría la caída abrupta de los coeficientes después del retraso p.

En consecuencia, tanto la ACF como la PACF no muestran patrones suficientemente definidos como para sugerir de forma concluyente el uso de un modelo AR(p) o MA(q) puros. En cambio, el comportamiento observado sugiere que podría estar involucrada una combinación más compleja de componentes autorregresivos y de media móvil.

### 5.5 Implicaciones para la Modelación

A partir del análisis conjunto de la ACF y la PACF, consideraremos tres enfoques de modelado:

#### 1. Modelo ARIMA(p, d, q):

Dado que ni la ACF ni la PACF presentan un patrón concluyente que indique un modelo puramente autorregresivo AR(p) o de media móvil MA(q), se propone explorar un modelo mixto ARMA(p,q) como una posible estructura subyacente de la serie. No obstante, es importante recordar que la serie de tiempo original no era estacionaria y requirió una diferenciación de



primer orden para alcanzar estacionariedad. Puesto que las gráficas ACF y PACF se realizaron con la serie diferenciada, es tras esta transformación que es posible modelarla con un proceso ARMA(p, q), por lo tanto, es natural considerar un modelo ARIMA(p, d, q) utilizando la serie original, donde:

- p: orden del componente autorregresivo (AR),
- d = 1: número de diferenciaciones necesarias para lograr estacionariedad,
- q: orden del componente de media móvil (MA).

Este modelo permite capturar tanto la dependencia de la serie con sus valores pasados como con sus errores pasados, sobre una serie no estacionaria.

## **2. Modelo SARIMA(p, d, q)(P, D, Q)m:**

Además, durante el análisis de la estructura de la serie de tiempo, se identificó un componente estacional relativamente definido en la serie. Por tanto, se propone utilizar un modelo SARIMA, que extiende el ARIMA para capturar también la estructura estacional. Este modelo incorpora:

1. Una parte no estacionaria: ARIMA(p, d, q),
2. Una parte estacional: (P, D, Q)\_m, donde:
  - P: orden autorregresivo estacional,
  - D: número de diferenciaciones estacionales,
  - Q: orden de media móvil estacional,
  - m: periodicidad o longitud del ciclo estacional (por ejemplo, 12 para datos mensuales con estacionalidad anual).

Este modelo es el más general y flexible, al considerar tanto la dinámica de corto plazo como los patrones cíclicos estacionales.

Para identificar la combinación de órdenes óptima en los modelos ARIMA y SARIMA, se aplicará la metodología Box-Jenkins. Esta metodología consiste en ajustar múltiples especificaciones de modelos y seleccionar aquella que minimice el criterio de información de Akaike (AIC). Posteriormente, se realizará un análisis de los residuos del modelo seleccionado para verificar que cumpla con los supuestos necesarios, lo cual es esencial para validar su uso en pronósticos.

## **3. Baseline Model: Media Histórica Mensual**

Adicionalmente, se construirá un modelo base (baseline model) mediante pronósticos ingenuos, con el objetivo de contar con un punto de referencia para evaluar el desempeño de los modelos

ARIMA y SARIMA. Este modelo consistirá en calcular la media histórica mensual a partir del conjunto de entrenamiento. Por ejemplo, para pronosticar la tasa de desocupación de un mes específico en el conjunto de prueba, se utiliza el promedio de dicho mes calculado con los datos históricos del conjunto de entrenamiento.

Opté por este enfoque en lugar de otros modelos base, como la predicción mediante el último valor observado, ya que este método permite capturar de manera simple la estacionalidad de la serie. No obstante, al tratarse únicamente de un promedio, no incorpora tendencias ni otras dinámicas temporales, por lo que se espera que modelos más sofisticados logren un mejor desempeño.

## 6. Implementación y Diagnóstico de Modelos

En esta sección se describe el proceso completo de construcción, evaluación y diagnóstico de los modelos de pronóstico aplicados a la serie temporal de la tasa de desocupación en México. Compararemos el desempeño de un modelo baseline sencillo con modelos más robustos como ARIMA y SARIMA.

Se comienza con la división de los datos en conjuntos de entrenamiento y prueba, definiendo las métricas de evaluación a utilizar. Posteriormente, se desarrolla un modelo de pronóstico estacional ingenuo como línea base, el cual servirá como punto de referencia. Seguido de esto, se implementan modelos ARIMA y SARIMA, siguiendo de forma rigurosa las etapas de identificación, estimación, selección del mejor modelo y análisis de residuos de la metodología Box-Jenkins , **buscando el modelo óptimo que capture las características de la serie.**

Cada modelo es evaluado con base en su capacidad para predecir el test set, utilizando el Error Porcentual Absoluto Medio (MAPE) como métrica de referencia. Finalmente, se entrenan los modelos finales con el total de la serie disponible y se generan pronósticos para los seis meses posteriores al último dato observado. Esta sección sienta las bases para la comparación de resultados, interpretación de los pronósticos y recomendaciones que se presentan en las siguientes secciones del informe.

### 6.1 División de Datos y Métricas a Utilizar

#### 6.1.1 División de Datos en Conjunto de Prueba y Entrenamiento

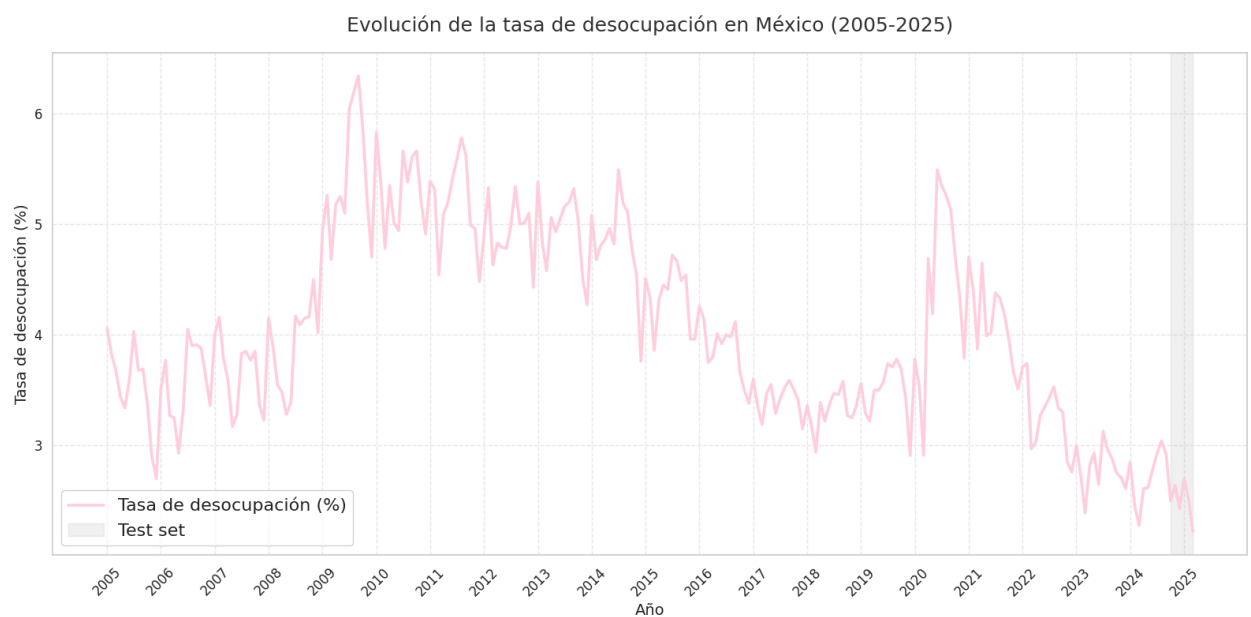
Para evaluar correctamente el desempeño de los modelos de pronóstico, es fundamental dividir la serie temporal en conjuntos de entrenamiento y prueba. La serie analizada contiene observaciones mensuales desde enero de 2005 hasta marzo de 2025. Se reservaron las últimas seis observaciones (abril a

septiembre de 2025) como conjunto de prueba, con el objetivo de validar la capacidad predictiva fuera del período de ajuste. El resto de los datos conforman el conjunto de entrenamiento, utilizado para ajustar los modelos. Los conjuntos se verían de la siguiente forma:

- **Conjunto de entrenamiento (training Set):** Incluye todas las observaciones disponibles, excepto las correspondientes a los últimos seis meses (de enero de 2004 hasta septiembre de 2024).
- **Conjunto de prueba (test set):** Contiene exclusivamente las observaciones de los últimos seis meses (octubre de 2024 hasta marzo de 2025).

### 6.1.2 Visualización de Conjuntos de Prueba y Entrenamiento

Para visualizar esta partición, se grafica la serie temporal completa destacando el período de prueba mediante una banda sombreada, lo que permite identificar claramente el intervalo donde se validarán los pronósticos.



### 6.1.3 Métrica de Evaluación (MAPE)

Para evaluar el desempeño de los modelos, utilizamos el MAPE (Mean Absolute Percentage Error), una métrica que mide el error absoluto promedio expresado como porcentaje del valor real. El MAPE compara directamente los valores observados con los pronosticados y nos indica qué tan lejos están, en promedio, nuestras predicciones de los datos reales.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

Donde:

- $y_i$ : valor real de la serie de tiempo.
- $\hat{y}_i$ : valor pronosticado.

El MAPE facilita la interpretación intuitiva del error promedio porcentual, y se interpreta según la siguiente regla general:

- **Menor a 10%:** Muy buena predicción.
- **Entre 10% y 20%:** Buena predicción.
- **Entre 20% y 50%:** Predicción aceptable.
- **Mayor a 50%:** Mala deficiente.

Un valor bajo de MAPE indica que, en promedio, las predicciones están muy cerca de los valores reales. Por el contrario, si el MAPE supera el 50%, el modelo tiene un desempeño tan pobre que sus predicciones no aportan más valor que el azar.

## 6.2 Baseline Model: Pronóstico Estacional Ingenuo

Para establecer un punto de referencia que nos permita evaluar la mejora que ofrecen los modelos ARIMA y SARIMA, se construyó un modelo baseline basado en un pronóstico estacional ingenuo. Este modelo es simple pero efectivo para capturar patrones estacionales presentes en la serie temporal.

La idea consiste en predecir el valor de la tasa de desocupación para un mes específico utilizando el promedio histórico de ese mismo mes, calculado a partir del conjunto de entrenamiento. Por ejemplo, para pronosticar la tasa de desocupación en enero, se toma el promedio de todas las tasas de desocupación observadas en enero durante el período del training set. Este procedimiento se repite para cada mes que se desea pronosticar en el test set.

El uso de este enfoque se fundamenta en que la tasa de desocupación mantiene patrones estacionales mensuales relativamente consistentes en el tiempo, por lo que el promedio histórico mensual es una estimación razonable para cada mes futuro.

Es importante señalar que se trabaja con los valores de la serie original y no con la serie diferenciada porque el promedio mensual tiene sentido interpretativo directo en la escala real de la tasa de desocupación. Si se usaran diferencias, los promedios representarían cambios promedio, no niveles, lo cual dificulta la interpretación y la aplicación como pronóstico directo.

### 6.1.1 Construcción del Baseline Model

Los pasos para construir este modelo fueron los siguientes:

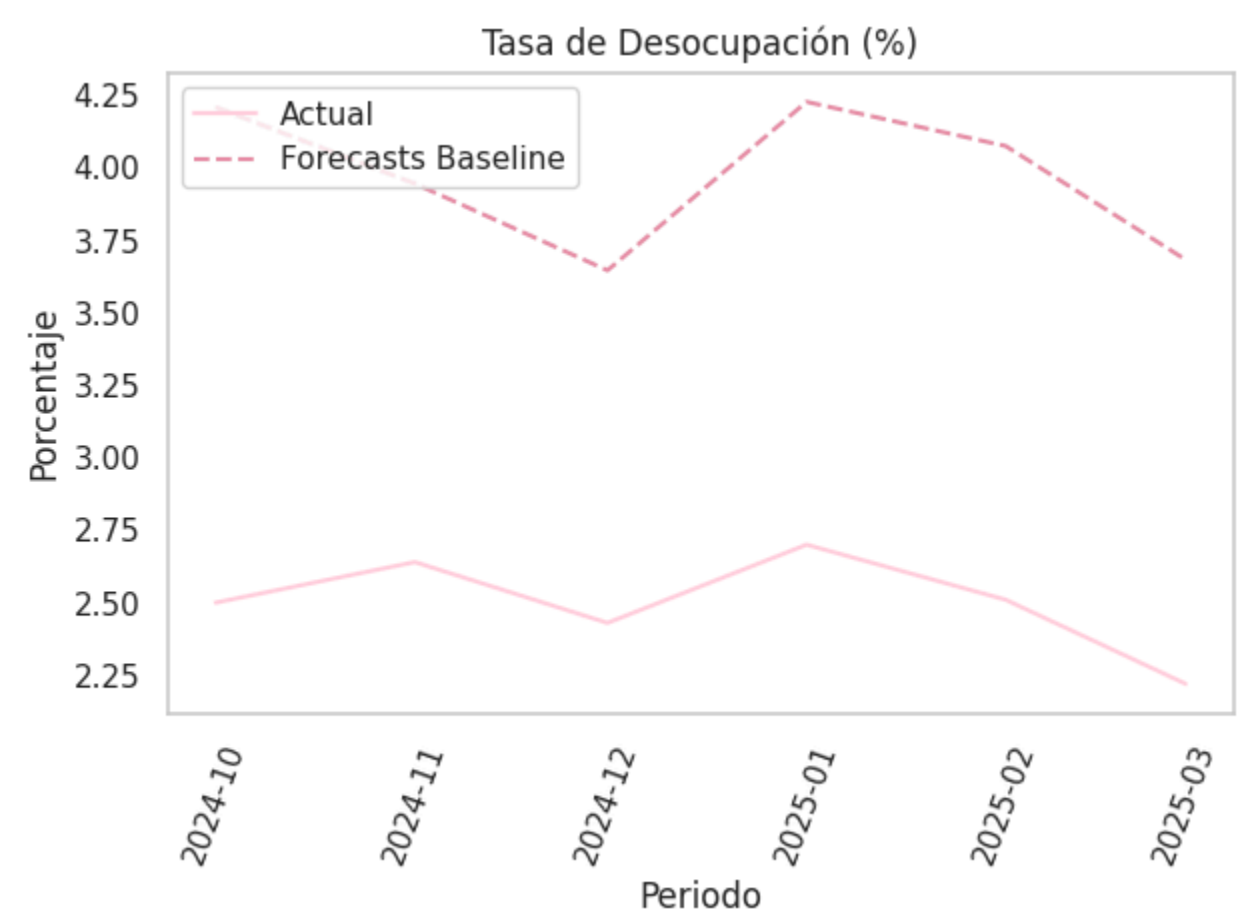
- 1. **Agrupación por mes:** Se agruparon las observaciones del training set según el mes del año (enero, febrero, etc.).
- 2. **Cálculo del promedio mensual:** Para cada mes, se calculó la media de la tasa de desocupación utilizando todas las observaciones correspondientes a ese mes en el training set.
- 3. **Asignación de pronósticos:** Para cada observación en el test set, se asignó como pronóstico el promedio histórico del mes correspondiente obtenido en el paso anterior.

### 6.1.2 Cálculo del MAPE

Con los pronósticos generados, se calculó el MAPE para cuantificar la precisión del modelo baseline en el conjunto de prueba, obteniendo así una medida inicial de desempeño. El modelo baseline obtuvo un **MAPE de 58.8175%**, lo cual indica que tiene un desempeño muy malo para realizar pronósticos. Este alto error resalta la necesidad de modelos más sofisticados como ARIMA/SARIMA.

### 6.1.3 Comparación de Pronósticos con los Datos Reales

Finalmente, se compararon visualmente los valores pronosticados con los observados en el conjunto de prueba, a fin de identificar patrones, desviaciones y características que permitan evaluar la utilidad y limitaciones del modelo ingenuo como base para modelos posteriores.



Se observa que los valores de la tasa de desocupación pronosticados por el modelo baseline presentan un desvío considerable respecto a los valores reales del conjunto de prueba. Esto indica que, aunque el modelo logra captar parcialmente el patrón estacional presente en la serie temporal, no reproduce con precisión la magnitud de las fluctuaciones observadas.

Por lo tanto, se espera que modelos más sofisticados, capaces de integrar tanto la tendencia como la estacionalidad y otras características temporales de la serie, proporcionan pronósticos más precisos y ajustados a los datos observados.

### 6.3 Modelo ARIMA(p, d, q)

Los modelos ARIMA(p, d, q) son una extensión de los modelos ARMA, ya que incorporan componentes autorregresivos (AR), de media móvil (MA) y una parte integrada (I), para tratar series no estacionarias.

- **AR (Autorregresivo):** Modela la dependencia entre un valor y sus observaciones pasadas de la serie.
- **I (Integrado o diferenciación):** Representa el número de diferenciaciones necesarias para volver la serie estacionaria.
- **MA (Media móvil):** Modela la dependencia entre un valor y errores pasados.

En esencia, un modelo ARIMA(p, d, q) ajusta un modelo ARMA(p, q) sobre una versión diferenciada de la serie original, donde  $d$  indica el número de veces que fue necesario diferenciar la serie para alcanzar estacionariedad. La estructura general de un modelo ARIMA puede expresarse como:

$$Y'_t = \rho_1 Y'_{t-1} + \dots + \rho_p Y'_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}$$

Donde:

- $Y'_t$  es la serie diferenciada
- $w_t$  es un ruido blanco, es decir,  $w_t \sim WN(0, \sigma^2)$ .

En notación con el operador retraso  $B$ , el modelo se expresa como:

$$\phi(B)Y'_t = \theta(B)w_t$$

Con:

- $\phi(B) = 1 - \rho_1 B - \dots - \rho_p B^p$ , es el polinomio característico autorregresivo.
- $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$ , es el polinomio característico de media móvil.

Este tipo de modelos es especialmente útil cuando, al diferenciar la serie, aún se observa una estructura significativa de dependencia temporal por las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF).

### **6.3.1 Construcción del Modelo ARIMA(p, d, q): Metodología Box-Jenkins**

Para ajustar un modelo ARIMA a nuestra serie de tiempo, seguimos el enfoque metodológico propuesto por Box y Jenkins, que consta de cuatro etapas: identificación, estimación, selección del mejor modelo y diagnóstico.

#### ***1. Identificación***

En esta etapa se analiza la serie para verificamos la presencia de estacionalidad y estacionariedad, y se determina el orden de diferenciación necesario.

- **Estacionariedad:**

Se aplicó la prueba aumentada de Dickey-Fuller (ADF) para detectar la presencia de raíz unitaria. Como se evidenció en la sección 5.1, el p-value de la prueba fue mayor a 0.05, por lo que no se rechazó la hipótesis nula, concluyendo que la serie no es estacionaria. Tras aplicar una diferenciación de primer orden, se logró alcanzar estacionariedad en la serie. En consecuencia, fijamos  $d = 1$  para el modelo.

- **Estacionalidad:**

Como se detalló en la sección 4.2, la serie presenta un comportamiento estacional anual relativamente definido, con picos recurrentes en los mismos periodos de cada año. Aunque este patrón sugiere que podría ajustarse un modelo SARIMA, en esta sección nos limitamos al análisis con un modelo ARIMA.

#### ***2. Estimación de Modelos***

Dado que se determinó que una diferenciación de orden 1 es suficiente, se procedió a ajustar modelos de la forma  $ARIMA(p, 1, q)$ , probando diversas combinaciones de los parámetros  $p$  y  $q$  dentro del rango  $[0, 4]$ . El criterio de selección de este rango de valores es arbitrario, pero esta selección permite considerar una amplia gama de estructuras: desde ruido blanco (cuando  $p = q = 0$ ), hasta procesos puramente autorregresivos, de media móvil o mixtos.

Para ello, se implementó una función que automatiza el ajuste y evaluación de modelos ARIMA, seleccionando aquellos con mejor desempeño según el criterio de información de Akaike (AIC). Este criterio penaliza la complejidad del modelo y aunque no mide la calidad absoluta del

modelo, permite comparar la calidad relativa entre diferentes especificaciones. A menor AIC, mejor es el modelo dentro del conjunto evaluado.

La función ejecuta los siguientes pasos:

1. Genera todas las combinaciones posibles de (p,q) dentro del rango definido.
2. Ajusta un modelo ARIMA(p, 1, q) para cada combinación.
3. Calcula el AIC correspondiente.
4. Devuelve un resumen con los modelos ordenados por desempeño.

El conjunto de valores seleccionados para variar los parámetros p, d, q del modelo ARIMA(p,d,q) fueron los siguientes:

$$p \in \{0, 1, 2, 3, 4\}, q \in \{0, 1, 2, 3, 4\}, d = 1$$

### ***3. Selección del Mejor Modelo***

Tras evaluar todas las combinaciones posibles de (p,q), se identificó que el modelo con el menor AIC fue el **ARIMA(3, 1, 3)**. Esto implica:

- **p = 3:** La parte autorregresiva (AR) es de orden 3, es decir, el modelo depende de los tres valores anteriores de la serie.
- **d = 1:** La parte integrada (I) indica que se aplicó una sola diferenciación para hacer la serie estacionaria.
- **q = 3:** La parte de media móvil (MA) también es de orden 3, lo que implica que el modelo utiliza los errores de predicción de los tres periodos anteriores.

Este modelo captura tanto la dinámica autorregresiva como la de media móvil de la serie diferenciada.

### ***4. Diagnóstico del Modelo: Análisis de Residuos***

Una vez ajustado el modelo ARIMA(3, 1, 3), es necesario evaluar si los residuos se comportan como un ruido blanco lo que implicaría que el modelo ha capturado toda la estructura sistemática de la serie y que lo que queda es meramente aleatorio. Este paso es fundamental para validar la calidad del ajuste y la fiabilidad de los pronósticos.

Para que los residuos puedan ser considerados ruido blanco deben cumplir con cumplen con las suposiciones de independencia, homocedasticidad y normalidad. Esto es:

1. Son independientes entre sí.

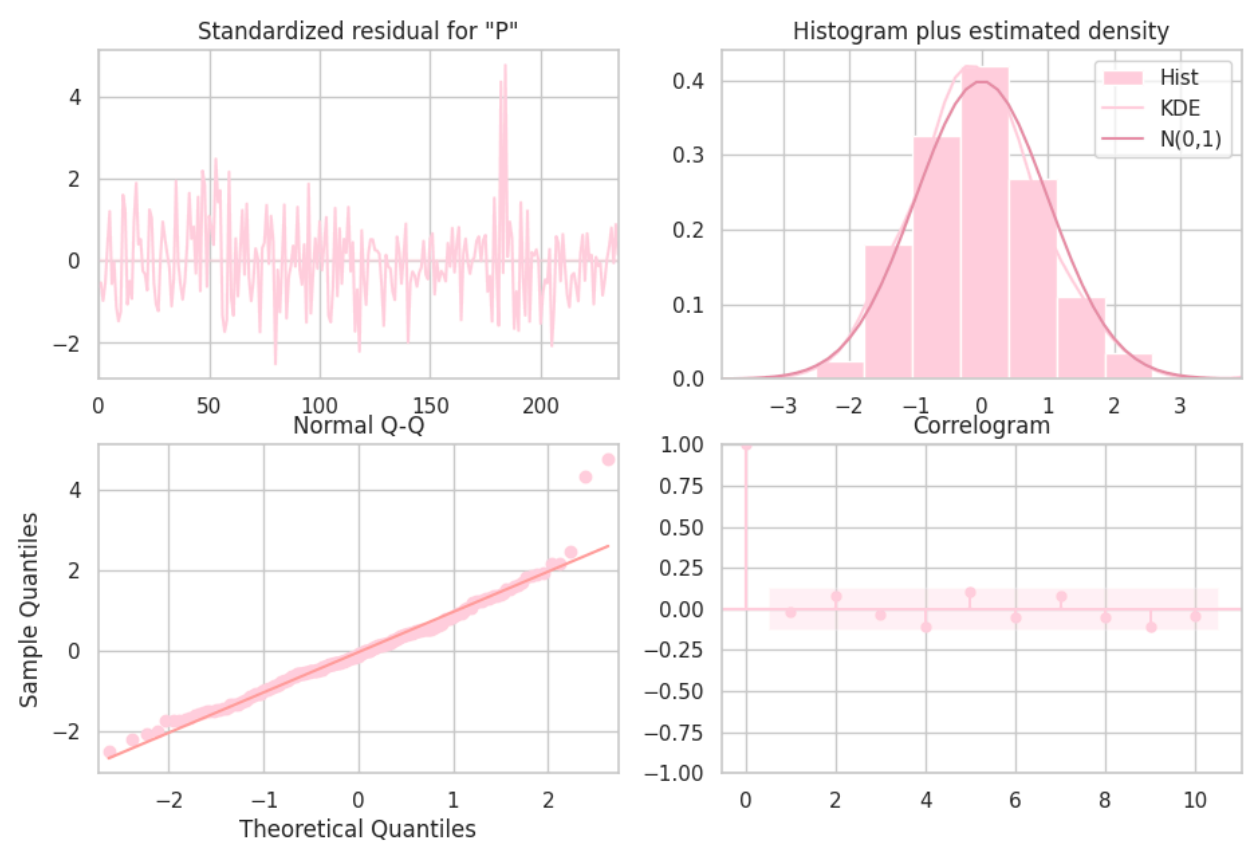


- 2. Tienen media cero.
- 3. Presentar varianza constante a lo largo del tiempo.

Para evaluar estas condiciones, llevamos a cabo dos tipos de diagnóstico:

1. Diagnóstico de Gráfico de Residuos

En este caso, los residuos fueron examinados mediante gráficos de residuos vs. tiempo, histograma de residuos, y un correlograma (ACF) para explorar posibles correlaciones temporales residuales.



a) Residuos estandarizados a lo largo del tiempo

Se graficaron los residuos estandarizados del modelo  $ARIMA(3, 1, 3)$  con el objetivo de examinar la posible presencia de patrones o variabilidad no capturada. En condiciones ideales, los residuos deben mostrar un comportamiento aleatorio en el tiempo, sin tendencias ni patrones visibles.

En la gráfica, si bien se observa un comportamiento en su mayoría aleatorio, Sin embargo, aunque en general los residuos parecen aleatorios, se observa un aumento notable en un periodo específico. Este comportamiento puede indicar que la varianza no es constante o que existe algún componente no capturado por el modelo que introduce variabilidad adicional.

Este hallazgo coincide con lo mencionado en la sección 4.2, donde se observó que la serie original presenta un patrón estacional relativamente definido, pero con un periodo en el que dicho patrón disminuye. Este cambio estructural se vuelve más evidente en los residuos del modelo. Además, como ya se mencionó, esto podría ser indicio de que una descomposición multiplicativa sería más adecuada que una aditiva, ya que la variabilidad parece estar relacionada con el nivel de la serie.

### **b) Histograma de los residuos**

El histograma permite visualizar la distribución de los residuos. En este caso, la forma del histograma se asemeja a la de una distribución normal.

Esto sugiere que los residuos presentan una distribución similar a distribución normal. Aunque no se busca comprobar una normalidad exacta, bajo este supuesto, podemos asumir que su media es cero.

### **c) Gráfico Q-Q Plot**

Esta gráfica compara los cuantiles observados de los residuos con los cuantiles teóricos de una distribución normal. Se observa que los puntos se alinean bastante bien sobre la línea diagonal, podemos concluir que los residuos siguen una distribución aproximadamente normal. Esto refuerza la hipótesis de que tienen media cero.

### **d) Correlograma (ACF) de los residuos**

En el correlograma, buscamos si existen autocorrelaciones significativas entre los residuos. Como en los análisis ACF anteriores, si los puntos caen dentro de las bandas de confianza, entonces las autocorrelaciones no son estadísticamente significativas. En este caso, los residuos parecen no presentar autocorrelación significativa, por lo que podemos considerar que son independientes.

### **Conclusión del diagnóstico visual**

A pesar de que la primer gráfica parece indicar que los residuos no tienen una varianza constante, las demás gráficas reflejan características de un ruido blanco. Por lo tanto, concluimos que en principio, los residuos del modelo cumplen con los supuestos básicos de aleatoriedad e independencia.

No obstante, para confirmarlo formalmente, realizaremos una prueba estadística de Ljung-Box, que nos permitirá evaluar si existe autocorrelación en los residuos. De haberla, entonces no se cumpliría el supuesto de independencia, y podría indicar problemas en la especificación del modelo.

## 2. Prueba de Ljung-Box:

Aplicamos la **prueba de Ljung-Box**, que permite verificar estadísticamente si los residuos del modelo pueden considerarse independientes, es decir, si constituyen ruido blanco. Esta prueba detecta la presencia de autocorrelación en distintos retrasos de la serie temporal.

Dado que es una prueba de tipo portmanteau, se plantea una hipótesis nula bien definida y una alternativa más general:

- **Hipótesis nula ( $H_0$ ):** No hay autocorrelación  $\rightarrow$  los residuos son independientes.
- **Hipótesis alternativa ( $H_a$ ):** Flexible, los residuos presentan autocorrelación.

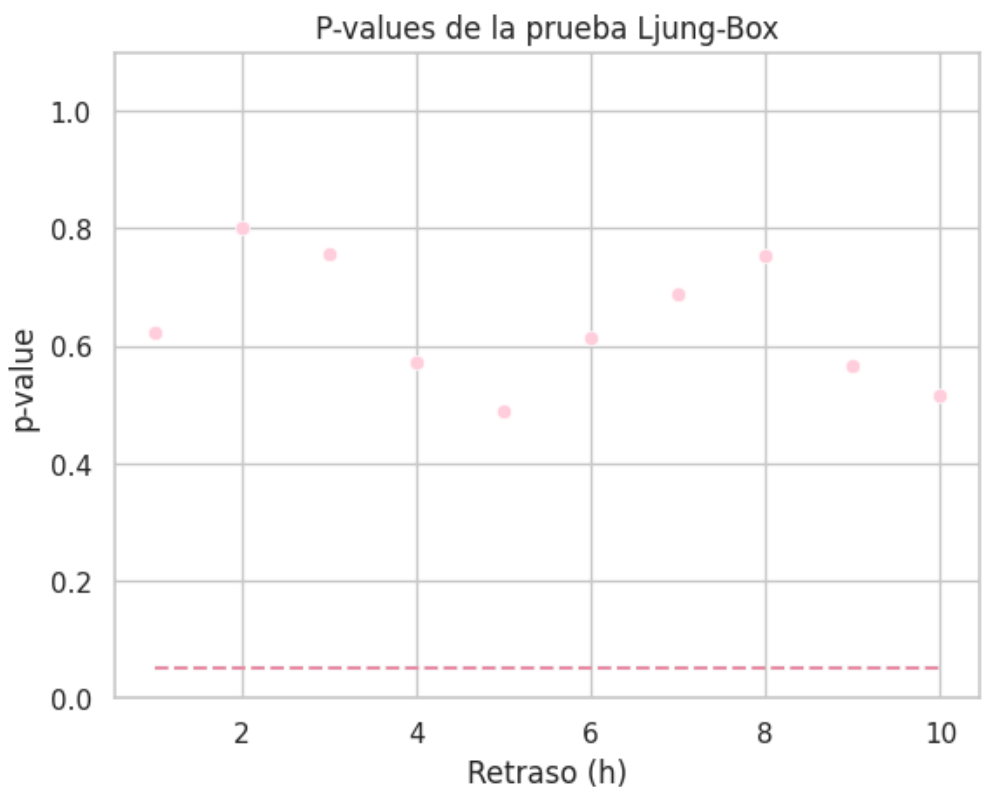
La prueba calcula un estadístico  $Q_m$  para diferentes retrasos  $m = 1,2,3,\dots$ , generando un p-valor asociado a cada uno. Estos p-valores pueden graficarse para facilitar su interpretación.

### *Criterio de interpretación:*

- Si el p-valor es **mayor a 0.05**, no se rechaza  $H_0$ : los residuos son independientes
- Si el p-valor es **menor a 0.05**, se rechaza  $H_0$ : los residuos presentan autocorrelación.

A diferencia de pruebas como ADF, aquí esperamos obtener p-valores grandes, ya que eso respalda que el modelo ha capturado adecuadamente la estructura temporal de la serie.

Los **p-valores** obtenidos se graficaron junto con una línea de referencia en 0.05 para facilitar la interpretación.



La gráfica muestra que todos los p-valores están por encima de 0.05, lo que implica que no se rechaza la hipótesis nula en ningún retraso. Esto sugiere que no hay evidencia estadística de autocorrelación en los residuos, por lo que pueden considerarse independientes.

Podemos concluir que los residuos del modelo cumplen con una de las principales condiciones de validez del ARIMA(3, 1, 3): comportarse como ruido blanco, es decir, presentar independencia, media cero y (suponemos) varianza constante.

### **6.3.2 Pronósticos**

#### **1. Pronósticos del Conjunto de Prueba**

Tras aplicar la metodología de Box-Jenkins y validar que los residuos del modelo ARIMA(3, 1, 3) se comportan como ruido blanco, concluimos que este modelo es adecuado para realizar pronósticos confiables de la tasa de desocupación. Con esto, procedemos a generar predicciones tanto para los datos de prueba (test set) como para los próximos seis meses no observados en la serie original.

##### ***1. Pronosticar Elementos del Conjunto de Prueba***

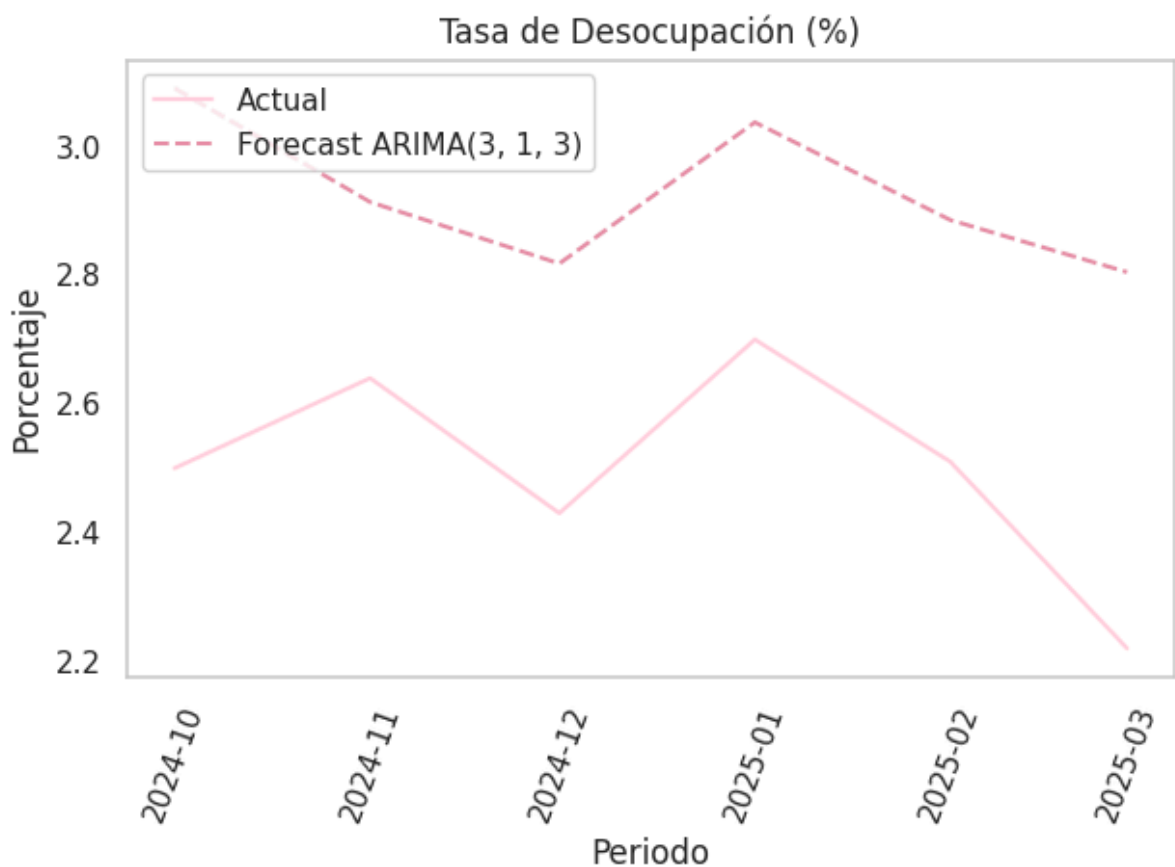
En primer lugar, evaluamos el desempeño del modelo al predecir los valores del test set, el cual abarca los últimos seis meses disponibles en la serie. Esta comparación con datos reales nos permite validar la capacidad predictiva del modelo antes de proyectar valores a futuro. Utilizamos la función de predicción correspondiente para obtener las estimaciones, y estas se agregaron directamente al test set para facilitar su análisis.

##### ***2. Calcular MAPE***

Para cuantificar la precisión del modelo, calculamos el MAPE, el cual arrojó un valor de 17.3%, indicando que el modelo ofrece un buen nivel de exactitud. Este resultado sugiere que el modelo logra capturar adecuadamente el comportamiento general de la serie temporal y proporciona estimaciones útiles.

##### ***3. Comparación de Pronósticos con Datos Reales***

Adicionalmente, realizamos una visualización comparativa entre los valores reales del test set y los pronósticos del modelo ARIMA.



Aunque el error absoluto medio es relativamente bajo, la gráfica revela una discrepancia visible entre los valores reales y los pronosticados. En particular, se observa una sobreestimación por parte del modelo. Esto puede deberse a ciertas características de la serie, como la posible presencia de varianza no constante observada en el análisis de residuos, lo que podría derivar en una leve sobreestimación en los pronósticos. También cabe considerar factores de escala como una posible fuente de esta diferencia.

**2. Pronósticos a Futuro**

Una vez evaluada la capacidad predictiva del modelo ARIMA(3, 1, 3) mediante el test set, se procede a su aplicación para realizar predicciones a futuro de la tasa de desocupación. En particular, se generan pronósticos para los próximos seis meses no observados en la serie: abril, mayo, junio, julio, agosto y septiembre de 2025.

Estos pronósticos incluyen tanto estimaciones puntuales como intervalos de confianza del 95%, permitiendo así acotar un rango plausible en el que podría ubicarse el valor real de la tasa de desocupación durante cada uno de estos meses.

**1. Entrenar Modelo Final**

Para generar pronósticos más robustos, se entrena nuevamente el modelo ARIMA(3, 1, 3), esta vez utilizando la serie completa disponible. Al no realizar la partición en conjuntos de

entrenamiento y prueba, se aprovecha toda la información histórica para maximizar la capacidad predictiva del modelo.

**2. Realizar Pronósticos a Futuro**

Una vez entrenado el modelo con todos los datos, se generaron predicciones para los seis meses siguientes a la última observación. Las proyecciones incluyen tanto los valores puntuales estimados como los intervalos de confianza del 95% para cada mes, permitiendo evaluar la variabilidad esperada alrededor de cada pronóstico.

**3. Resultados**

Se construyó una tabla con los resultados obtenidos, donde se presenta:

- El mes correspondiente a cada periodo pronosticado.
- El valor puntual estimado por el modelo ARIMA.
- El límite inferior y superior del intervalo de confianza del 95%.

	Periodo	Pronóstico ARIMA(3, 1, 3)	Límite Inferior	Límite Superior
243	2025-04-01	2.542597	1.902204	3.182991
244	2025-05-01	2.440753	1.636290	3.245217
245	2025-06-01	2.337889	1.440384	3.235395
246	2025-07-01	2.603489	1.635134	3.571843
247	2025-08-01	2.472028	1.444440	3.499615
248	2025-09-01	2.354265	1.276800	3.431729

Esto nos proporciona una visión cuantitativa y clara del comportamiento esperado de la tasa de desocupación, junto con el grado de incertidumbre de las estimaciones. Por ejemplo, para el mes de abril de 2025, el modelo estima una tasa de desocupación cercana al 2.54%. Sin embargo, considerando la incertidumbre, se puede afirmar con un 95% de confianza que el valor real se encontrará entre 1.90% y 3.18%.

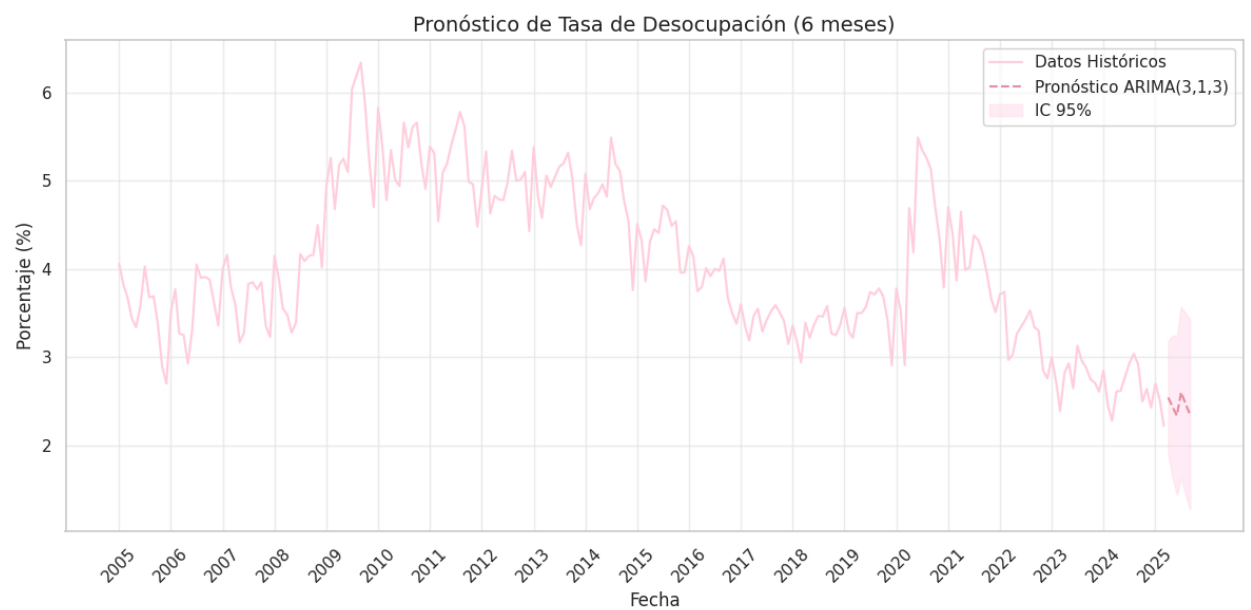
**4. Visualizar Serie de Tiempo y Pronósticos a Futuro**

Se elaboró una gráfica que muestra tanto la serie histórica como los pronósticos generados para los seis meses siguientes. En dicha visualización se destacan:

- Los datos históricos observados.
- La trayectoria esperada según el modelo ARIMA(3,1,3)

- Los intervalos de confianza al 95% para cada predicción.

Esta representación permite observar la evolución estimada de la tasa de desocupación y la incertidumbre asociada a cada punto futuro.



Los pronósticos generados por el modelo ARIMA(3,1,3) reflejan una dinámica que guarda coherencia con el comportamiento reciente de la serie.

Durante los primeros tres meses del periodo proyectado (abril, mayo y junio de 2025), se observa una leve tendencia al alza en la tasa de desocupación, lo que podría indicar un periodo de enfriamiento en el mercado laboral.

A partir de julio, el modelo proyecta una disminución gradual, lo cual sugiere una posible estabilización o ligera recuperación hacia el final del periodo pronosticado.

Los intervalos de confianza permiten contextualizar cada predicción, mostrando que, si bien existe una variabilidad natural en las estimaciones, los rangos obtenidos son razonables y consistentes con el comportamiento observado en los datos previos.

### 6.4 Modelo SARIMA(p, d, q)(P, D, Q)m

Como se mencionó anteriormente, detectamos en nuestra serie de tiempo un patrón estacional relativamente definido. En estos casos, el modelo ARIMA simple resulta insuficiente, ya que no puede capturar adecuadamente las fluctuaciones que se repiten de forma periódica. Por ello, recurrimos al modelo SARIMA, el cual extiende el ARIMA para incorporar explícitamente esta componente estacional.

El modelo SARIMA(p,d,q)(P,D,Q)m incluye dos partes fundamentales:

- Una **parte no estacionaria**, equivalente al modelo ARIMA(p,d,q), que capta las dinámicas de corto plazo.
- Una **parte estacional**, caracterizada por los parámetros P, D, Q y el período estacional m, que modela las repeticiones a intervalos regulares.

Matemáticamente, este modelo se expresa como:

$$\Phi(B^m)\phi(B)\nabla^d\partial^DY_t = \Theta(B^m)\theta(B)w_t$$

donde:

- $\phi(B)$  y  $\theta(B)$  son los polinomios no estacionales autorregresivo y de media móvil, respectivamente.
- $\Phi(B^m)$  y  $\Theta(B^m)$  son los polinomios estacionales.
- $\nabla^d = (1 - B)^d$  representa la diferenciación no estacional.
- $\partial^D = (1 - B^m)^D$  es la diferenciación estacional.
- $B$  es el operador retraso, y  $w_t$  es un ruido blanco.

Este modelo es especialmente útil cuando el valor de la serie en un momento dado no solo depende de los meses recientes, sino también de los valores de la misma estación en años anteriores.

#### 6.4.1 Construcción del Modelo SARIMA(p, d, q)(P, D, Q)m: Metodología Box-Jenkins

Aplicamos la metodología Box-Jenkins para seleccionar el modelo SARIMA(p,d,q)(P,D,Q)m que mejor se ajusta a nuestra serie de tiempo.

##### 1. Identificación

- **Estacionariedad:**

Como vimos anteriormente, la serie original no era estacionaria, lo cual fue confirmado mediante la prueba ADF. Aplicando una primera diferenciación ( $d = 1$ ), se logró alcanzar la estacionariedad.

- **Estacionalidad:**

La serie presenta un patrón estacional claro con periodicidad anual (dado que los datos



son mensuales), por lo que se define  $m = 12$ . Dado que una sola diferenciación fue suficiente para estabilizar la media, también se establece que  $D = 0$ .

## ***2. Estimación de Modelos***

Con base en la identificación, se procede a estimar varios modelos SARIMA con distintas combinaciones de parámetros  $p, q, P, Q$ , manteniendo fijo  $d = 1, D = 0$  y  $m = 12$ .

Para ello, se consideraron rangos amplios para los parámetros no estacionales ( $p$  y  $q$  de 0 a 2) y estacionales ( $P$  y  $Q$  de 0 a 3), con el objetivo de explorar un espacio suficiente sin un costo computacional excesivo.

Cada modelo fue ajustado iterativamente y evaluado mediante el criterio de información de Akaike (AIC), que penaliza la complejidad para favorecer modelos con buen ajuste.

Esta búsqueda sistemática permitió identificar el modelo SARIMA con el menor AIC, es decir, el modelo que mejor describe la serie según el balance entre ajuste y parsimonia.

## ***3. Selección del Mejor Modelo***

Los resultados indicaron que el modelo óptimo para la serie es un SARIMA(2, 1, 2)(1, 0, 1)<sub>12</sub>, que incluye:

- Orden autorregresivo no estacional  $p = 2$ , considerando dos retardos previos,
- Diferenciación no estacional  $d = 1$  para lograr estacionariedad,
- Orden de media móvil no estacional  $q = 2$ , que incorpora errores de dos periodos anteriores,
- Orden autorregresivo estacional  $P = 1$ , que capta la dependencia con el mismo mes en años previos,
- Sin diferenciación estacional ( $D = 0$ ),
- Orden de media móvil estacional  $Q = 1$ ,
- Y periodicidad estacional  $m = 12$ , acorde a la frecuencia mensual con estacionalidad anual.

Con esta combinación, el modelo logra capturar tanto la dinámica no estacional como la estacional observada en la serie.

Con base en los parámetros seleccionados, se ajustó el modelo SARIMA correspondiente sobre el conjunto de entrenamiento. Se obtuvieron los estimadores de los coeficientes autorregresivos y de media móvil tanto estacionales como no estacionales, y se analizaron los valores de significancia y la convergencia del modelo.

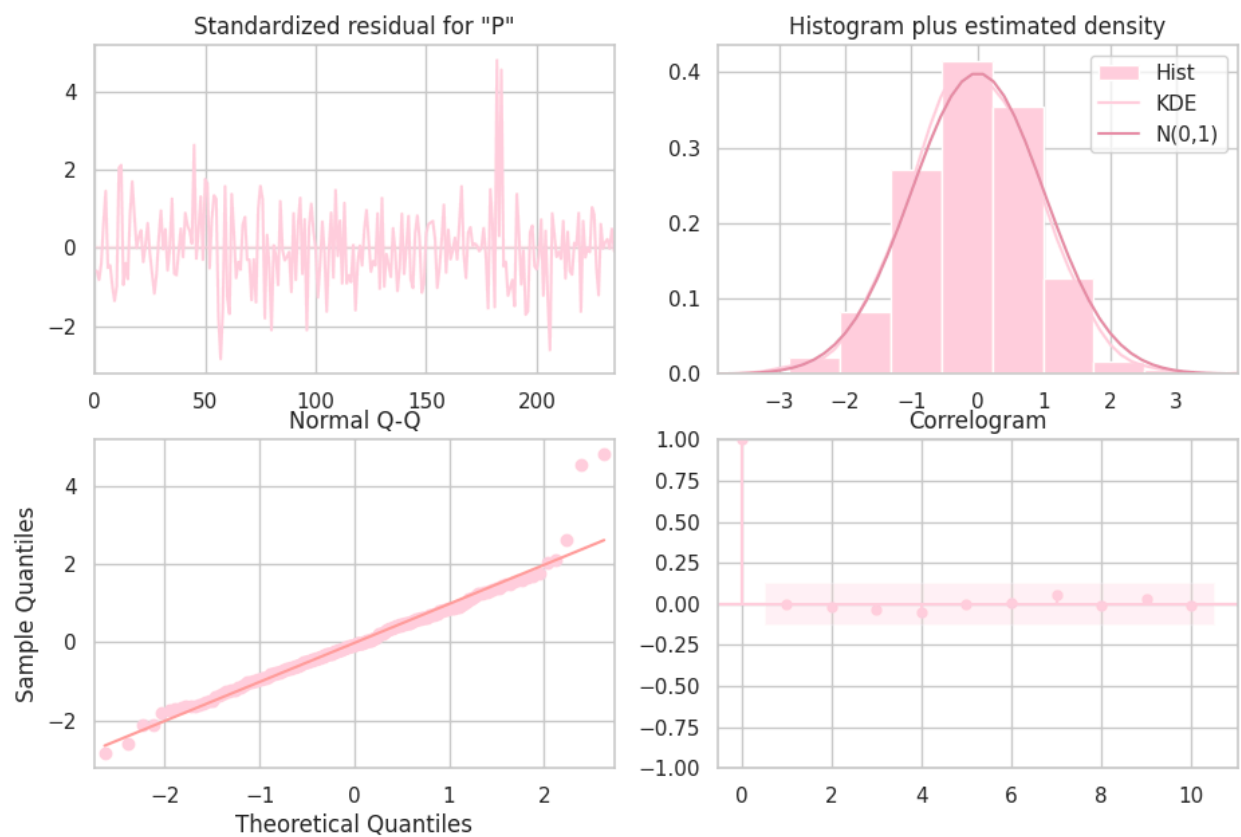
Este ajuste permite realizar predicciones que consideran la estructura temporal y estacional de la serie, mejorando la precisión frente a modelos que no incorporan estacionalidad explícitamente.

4. Diagnóstico del Modelo: Análisis de Residuos

Una vez ajustado el modelo SARIMA, es fundamental evaluar la calidad del ajuste mediante el análisis de sus residuos. Esto nos permite verificar que el modelo capture adecuadamente la dinámica de la serie y que los supuestos subyacentes se cumplan.

1. Diagnóstico de Gráfico de Residuos

Evaluamos los residuos del modelo SARIMA(2,1,2)(1,0,1)12 para verificar si se comportan como ruido blanco.



En el análisis de los residuos estandarizados del modelo SARIMA, se observa un comportamiento generalmente aleatorio, similar al modelo ARIMA(3,1,3). Sin embargo, durante un periodo específico, los residuos presentan un aumento notable que coincide con la componente estacional de la serie original. Este comportamiento sugiere que la varianza no es completamente constante o que el modelo no captura algún componente estructural, lo que indica que una descomposición multiplicativa podría ser más adecuada que una aditiva.

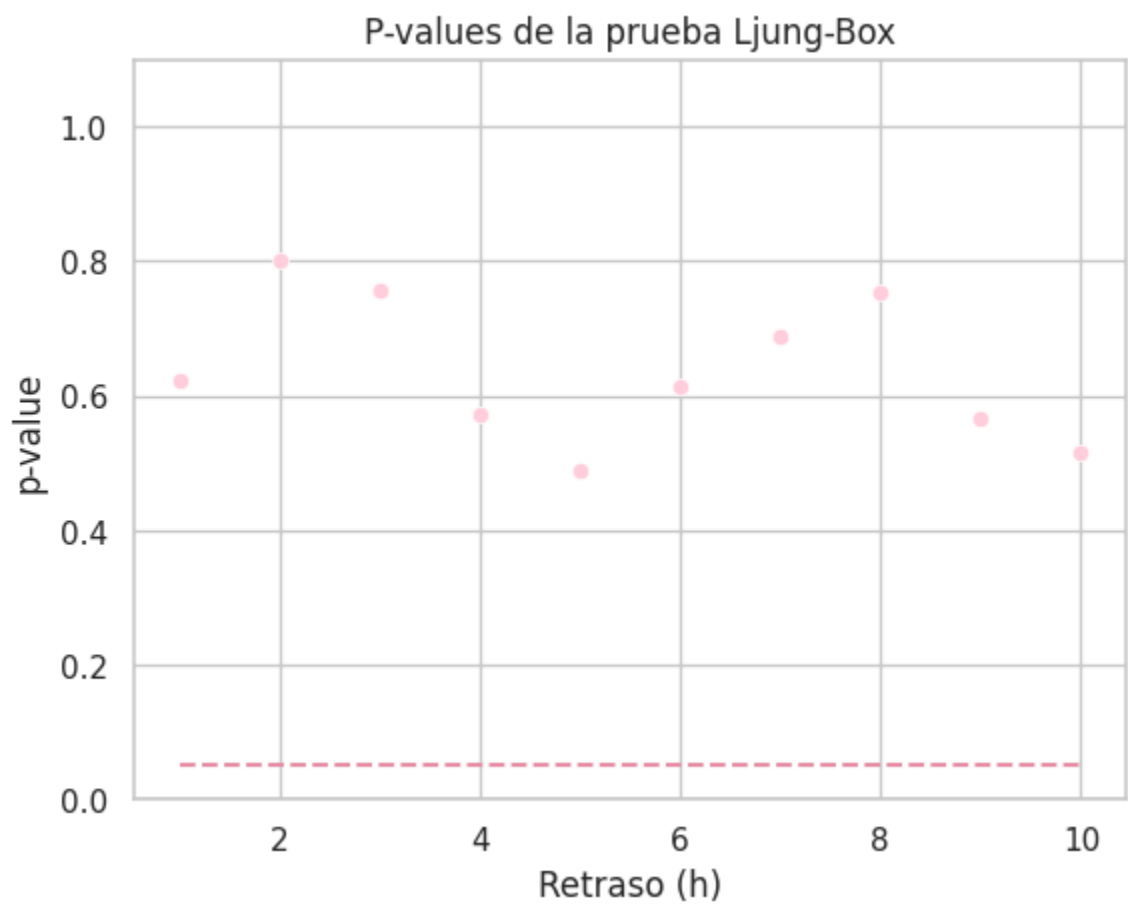
Al examinar la distribución de los residuos mediante el histograma y el Q-Q plot, se observa que éstos tienen una distribución simétrica que se aproxima bastante a la normalidad, más aún que en el modelo ARIMA. Esta normalidad en los residuos permite asumir que su media es cero, condición importante para validar la adecuación del modelo.

El análisis del correlograma ACF muestra que los coeficientes de autocorrelación de los residuos permanecen dentro de las bandas de confianza, lo que indica ausencia de autocorrelación significativa y permite concluir que los residuos pueden considerarse independientes.

En general, aunque se detecta una posible heterocedasticidad en un periodo puntual, las evidencias gráficas y estadísticas sugieren que los residuos del modelo SARIMA cumplen en su mayoría con las condiciones de ruido blanco, lo que valida el uso del modelo para la serie analizada y para la generación de pronósticos.

**2. Prueba de Ljung-Box:**

Para complementar el diagnóstico visual, realizamos la prueba de Ljung-Box para evaluar si los residuos del modelo SARIMA presentan autocorrelación significativa.



Al igual que en el modelo ARIMA, la mayoría de los p-valores se encuentran por encima del umbral, por lo que no se rechaza la hipótesis nula, indicando que los residuos del modelo SARIMA son independientes. Podemos concluir que los residuos del modelo se comportan como ruido blanco: son independientes, tienen media cero y asumiremos que su varianza es constante.

## 6.4.2 Pronósticos

Dado que los residuos del modelo  $SARIMA(2,1,2)(1,0,1)_{12}$  cumplen con las características de ruido blanco, consideramos que este modelo es adecuado para generar predicciones confiables.

Al igual que con el modelo  $ARIMA(3,1,3)$ , primero realizamos pronósticos para los últimos seis meses observados (test set), para evaluar qué tan bien el modelo predice datos conocidos. Luego, usaremos el modelo para proyectar la tasa de desocupación durante los siguientes seis meses (abril a septiembre de 2025), es decir, un pronóstico hacia el futuro.

La principal ventaja del modelo SARIMA es que, a diferencia del ARIMA que solo modela tendencia y ruido, el SARIMA incorpora explícitamente la estacionalidad, capturando patrones periódicos que se repiten cada año.

### 1. Pronósticos del Conjunto de Prueba

Después de aplicar el modelo  $SARIMA(2,1,2)(1,0,1)_{12}$  y confirmar que sus residuos cumplen con las características de un ruido blanco, concluimos que este modelo es adecuado para realizar pronósticos confiables de la tasa de desocupación.

Para validar su desempeño, generamos pronósticos para el conjunto de prueba, que incluye los últimos seis meses observados de la serie temporal. Esta comparación con los datos reales permite evaluar la capacidad predictiva del modelo antes de proyectar valores a futuro.

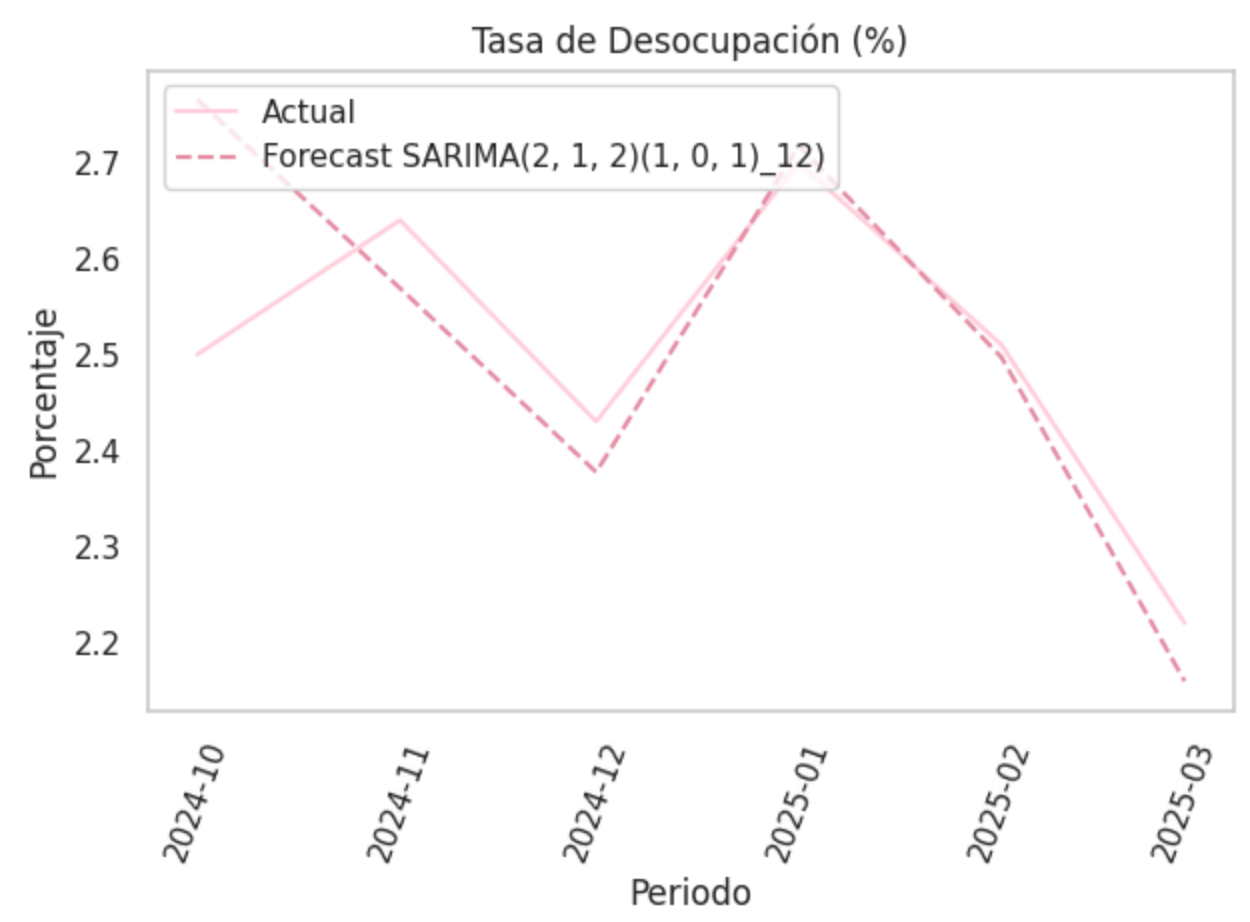
Se agregaron las predicciones generadas por el modelo SARIMA al conjunto de prueba, facilitando el análisis comparativo.

#### Cálculo del MAPE

Se calculó el MAPE para cuantificar la precisión del modelo en el conjunto de prueba. El valor obtenido fue de 3.23%, lo que indica un excelente nivel de exactitud. Este bajo error sugiere que el modelo captura con precisión tanto la tendencia como la estacionalidad presentes en la serie temporal.

#### Comparación Visual de Pronósticos y Datos Reales

Se realizó una visualización que compara los valores reales con los pronósticos generados por el modelo. A diferencia del modelo  $ARIMA(3, 1, 3)$ , las curvas de los pronósticos y los datos reales prácticamente se superponen, evidenciando un ajuste más preciso y sin sesgos evidentes, como la sobreestimación observada previamente.



## 2. Pronósticos a Futuro

Con base en la validación favorable en el conjunto de prueba, procedimos a generar pronósticos para los próximos seis meses no observados: abril a septiembre de 2025. Este pronóstico incorpora tanto los valores puntuales como los intervalos de confianza del 95%, lo que permite estimar un rango plausible para la tasa de desocupación en cada mes.

### Entrenamiento del Modelo Final

Para maximizar la capacidad predictiva, el modelo SARIMA(2,1,2)(1,0,1)12 se entrenó nuevamente utilizando la serie completa disponible, sin partición en conjuntos de entrenamiento y prueba.

### Resultados de los Pronósticos a Futuro

	Periodo	Pronostico SARIMA(2, 1, 2)(1, 0, 1)_12	Limite Inferior	Limite Superior
243	2025-04-01	2.581479	2.074254	3.088703
244	2025-05-01	2.580192	1.979286	3.181097
245	2025-06-01	2.669029	1.950169	3.387890

246	2025-07-01	2.842435	1.998744	3.686126
247	2025-08-01	2.855294	1.933867	3.776722
248	2025-09-01	2.772505	1.769814	3.775197

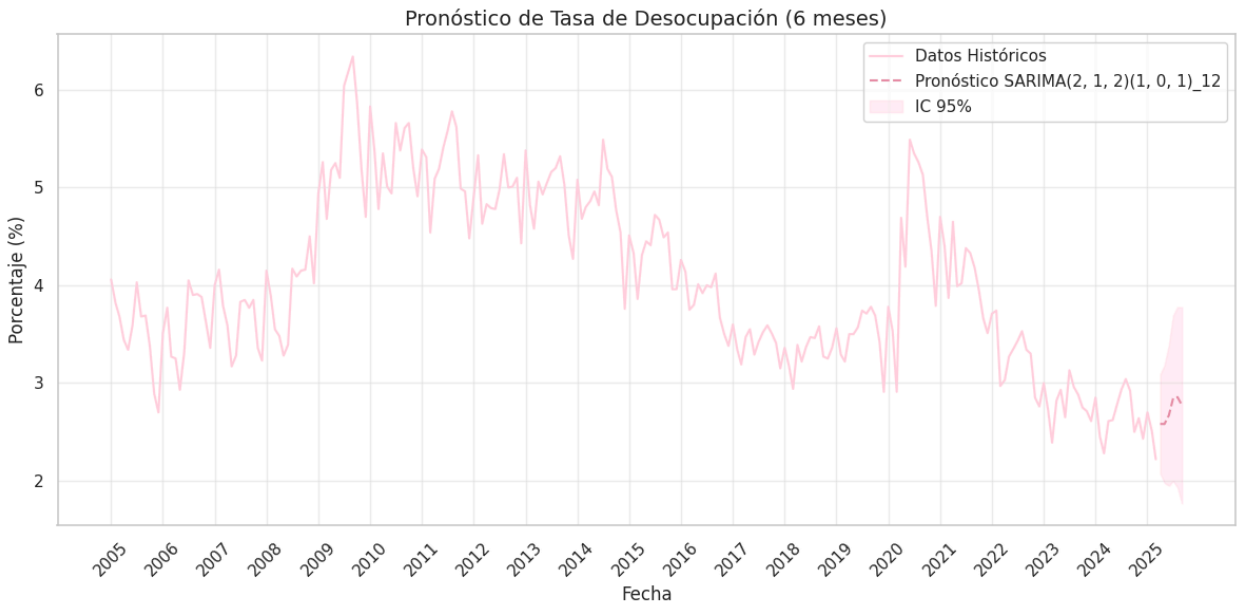
Los pronósticos muestran una tendencia clara que combina los efectos de la estacionalidad y la dinámica de la serie histórica. Por ejemplo, se observa que la tasa de desocupación podría presentar una ligera tendencia al alza durante los primeros meses del periodo proyectado, seguida de una estabilización hacia finales de septiembre de 2025.

Se construyó una tabla con los valores puntuales y sus intervalos de confianza, brindando una visión cuantitativa y confiable del comportamiento esperado, junto con la incertidumbre asociada a cada estimación.

### 3. Visualización de la Serie y Pronósticos

Se elaboró una gráfica que incluye:

- La serie histórica observada.
- Los pronósticos generados por el modelo SARIMA(2,1,2)(1,0,1)<sub>12</sub>.
- Los intervalos de confianza al 95% para cada punto pronosticado.



Esta visualización permite observar cómo el modelo capta patrones recurrentes estacionales y muestra la incertidumbre inherente a las predicciones.

## 7. Comparación de Resultados

En esta sección se presenta una comparación integral entre los modelos desarrollados: el modelo baseline, el modelo ARIMA(3, 1, 3) y el modelo SARIMA(2, 1, 2)(1, 0, 1)<sub>12</sub>. El objetivo es evaluar su desempeño predictivo mediante métricas cuantitativas y análisis visuales, que permitan establecer cuál modelo ofrece mejores pronósticos para la tasa de desocupación en México.

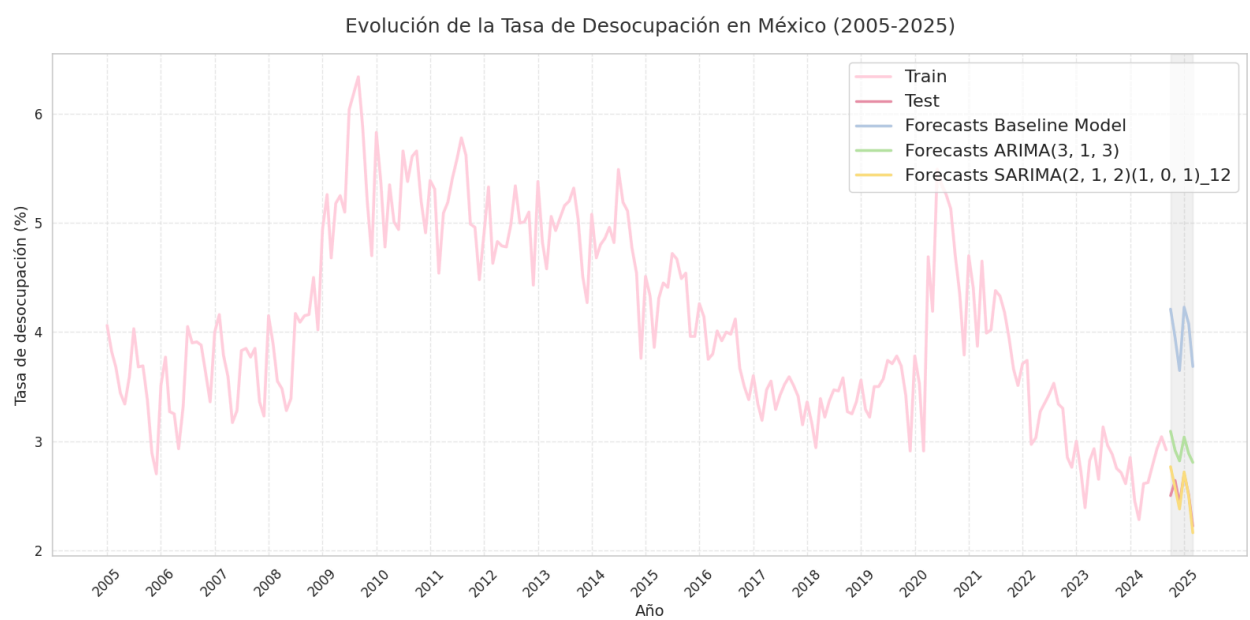
### 7.1 Comparación de Métricas obtenidas

Como se ha mencionado anteriormente, para evaluar la precisión de los pronósticos, se calculó el MAPE en cada modelo. Este indicador expresa en porcentaje, en promedio, la desviación entre los valores reales y los pronosticados, facilitando la comparación directa entre modelos.

Comparación de Métrica MAPE de cada Modelo		
Modelo Baseline	ARIMA(3, 1, 3)	SARIMA(2, 1, 2)(1,0, 1) <sub>12</sub>
58.8%	17.3%	3.23%

El modelo baseline, basado en promedios simples, obtuvo un MAPE de aproximadamente **58.8%**. Este elevado error confirma que dicho modelo no es adecuado para pronósticos confiables, ya que su desempeño es similar a una predicción aleatoria. La alta desviación sugiere que el modelo no logra capturar ni la tendencia ni la estacionalidad presentes en la serie temporal. En contraste, el modelo ARIMA(3, 1, 3) presentó un MAPE considerablemente menor, de alrededor de **17.3%**, lo que indica que este modelo captura mejor la dinámica de la serie y es capaz de generar pronósticos con una precisión aceptable para aplicaciones prácticas. Este nivel de error refleja que el ARIMA incorpora correctamente la tendencia y la autocorrelación en los datos. Finalmente, el modelo SARIMA(2, 1, 2)(1, 0, 1)<sub>12</sub>, que incluye componentes estacionales logró una mejora significativa en la precisión de los pronósticos, alcanzando un MAPE de **3.23%** inferior al del ARIMA. Esto confirma que la inclusión de componentes estacionales permite modelar más fielmente las fluctuaciones de la tasa de desocupación.

## 7.2 Comparación de Visualizaciones y Pronósticos



Al examinar las curvas de predicción junto con los valores reales, es evidente que el modelo SARIMA(2, 1, 2)(1, 0, 1)<sub>12</sub> tiene una capacidad predictiva para esta serie de tiempo superior a la de los modelos ARIMA(3, 1, 3) y el modelo baseline. Se observa una superposición casi perfecta entre los pronósticos realizados por el modelo SARIMA y los valores reales observados, lo que implica que el modelo no solo logra capturar la tendencia general de la serie, sino que también logra capturar de forma efectiva los patrones estacionales presentes. Esto probablemente le permitió al modelo evitar el sobreajuste de los pronósticos como fue el caso en modelos anteriores, lo cual sugiere que el modelo SARIMA(2, 1, 2)(1, 0, 1)<sub>12</sub> es una herramienta robusta para pronósticos en este contexto.

Además, el buen desempeño del modelo indica que los parámetros seleccionados para las componentes autorregresivas, de media móvil y estacionales están bien ajustados. Esto se traduce en pronósticos más confiables y estables, fundamentales para la toma de decisiones basadas en la serie temporal analizada.

## 10. Referencias y Repositorio

### 10.1 Referencias:

- [1] Instituto Nacional de Estadística y Geografía (INEGI), "Tabulados interactivos," 2025. [En línea]. Disponible: <https://www.inegi.org.mx/app/tabulados/default.html?nc=622>. [Accedido: 9-Jun-2025].
- [2] Naciones Unidas, "Crecimiento económico," 2025. [En línea]. Disponible: <https://www.un.org/sustainabledevelopment/es/economic-growth/>. [Accedido: 9-Jun-2025].



## 10.2 Repositorio:

- Los datos utilizados y el código reproducible se encuentran en el siguiente repositorio:  
<https://drive.google.com/drive/folders/1zeW9YT0OfNTEI0ROKBvIuI4FX8g07SrY?usp=sharing>