**Introducción al Análisis de Datos**

**Segundo parcial**

****

| **Asignatura** | Introducción al Análisis de Datos |
| --- | --- |
| **Título del Trabajo** | TP EPH Aglomerados 2016-2025 |
| **Integrantes** | Corbalán Federico, Hermosilla Emanuel, Kevarkian Candela, Locaso Turiaci Santiago |
| **Profesor** | Fernández Luis N. |
| **Fecha de Entrega** | 25 de octubre de 2025 |

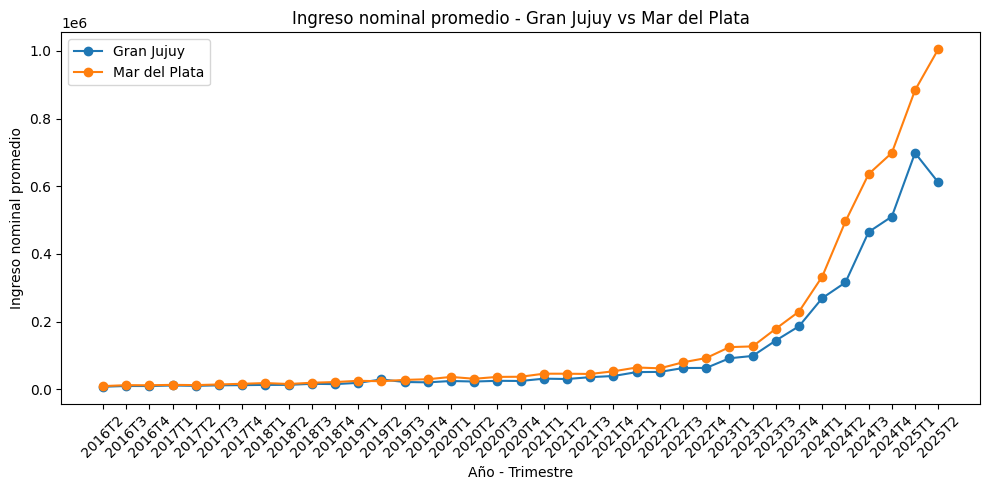
## **Análisis univariado**

Se realiza un análisis univariado utilizando estadísticas ponderadas según la metodología de la EPH. Dado que la serie presenta valores extremos —especialmente en 2020 y en los últimos años por inflación— se emplean medidas robustas como la mediana ponderada. No se eliminan observaciones ni trimestres, ya que se busca reflejar la evolución real de los indicadores para el período 2016–2025.

**1.1 Serie histórica de ingresos nominales**

Gran Jujuy - Ingreso nominal promedio

El ingreso nominal promedio muestra un crecimiento continuo entre 2016 y 2025. Durante los primeros años, el aumento es moderado, pero a partir de 2021 se acelera fuertemente y se vuelve muy pronunciado en 2023–2024, acompañado por el proceso inflacionario del período. En 2020 se observa una leve irregularidad asociada a los efectos de la pandemia sobre el mercado laboral: caída de la actividad, reducción de horas trabajadas y menor captación de ingresos por parte de la EPH. Hacia el final de la serie, en 2025, aparece una caída puntual del ingreso luego del pico alcanzado a fines de 2024, lo que constituye la principal baja observable en toda la trayectoria.



Mar del plata - Ingreso nominal promedio

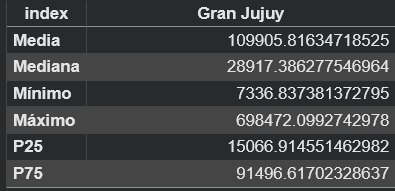
La serie de Mar del Plata presenta un comportamiento similar. Entre 2016 y 2020 el crecimiento es estable y moderado, pero a partir de 2021 se intensifica con fuerza. Al igual que en Jujuy, en 2020 aparece una pequeña discontinuidad vinculada a la disrupción económica del COVID-19.

Hacia 2024 se registra una leve irregularidad en la serie, una pequeña pausa o descenso puntual, antes de retomar el crecimiento en 2025, donde alcanza los valores más altos del período. En conjunto, la trayectoria muestra la misma dinámica nacional: el ingreso nominal crece aceleradamente en los últimos años, impulsado por la inflación.

**1.2 Medidas de tendencia central y de posición**

Medidas analizadas:

* Media - Mediana (50%) - Desvío estándar - Mínimo - Percentil 25 (25%) - Percentil 75 (75%) - Máximo

**Gran Jujuy (2016 - 2025)**

La media supera los 118.000 pesos y la mediana se ubica alrededor de 29.500. La diferencia entre ambas confirma una distribución asimétrica: los valores muy elevados de 2023–2025 empujan la media hacia arriba. El mínimo registrado es cercano a 7.300 (2016) y el máximo supera los 611.000 (2025). Los percentiles confirman que durante la mayor parte del período los ingresos se mantuvieron bajos (P25 ≈ 15.300; P50 ≈ 29.500), con incrementos abruptos recién en los últimos años.

Además de la caída registrada en 2020, la serie presenta un retroceso menor en 2023–2024, consistente con la desaceleración temporal del ingreso nominal observada en los gráficos. Esta baja es moderada y se revierte en 2025.

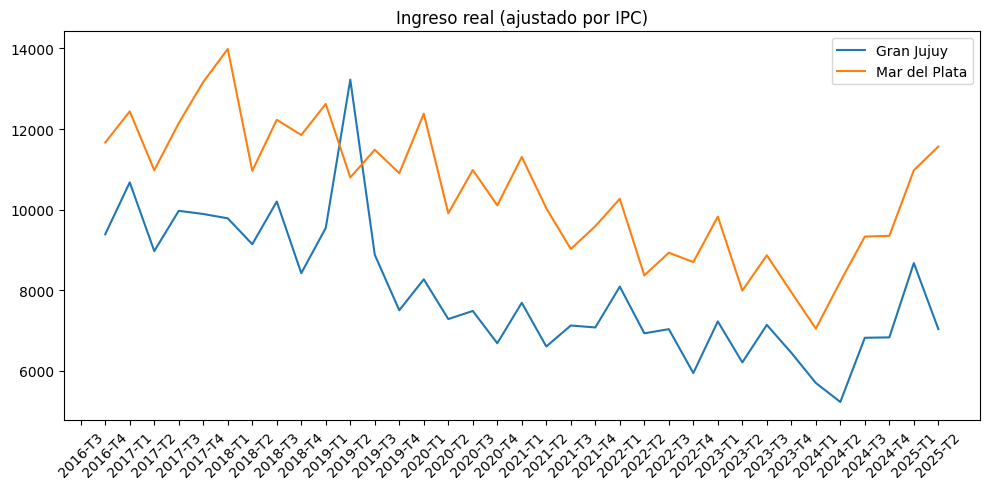


**Mar del Plata (2016 - 2025)**

La media ronda los 161.000 pesos y la mediana los 34.200. La asimetría vuelve a ser importante, por el crecimiento extremo de los últimos años. El mínimo es cercano a 8.700 y el máximo supera los 859.000. Los percentiles muestran ingresos mayores que en Jujuy durante casi todo el período (P25 ≈ 16.500; P75 ≈ 104.300).Al igual que en Jujuy, hacia 2023–2024 aparece una leve caída del ingreso nominal, previa al fuerte incremento de 2024–2025. Esta oscilación puede deberse a ajustes salariales que no acompañaron plenamente la inflación en esos trimestres.

Al igual que en Jujuy, **2020 presenta una distorsión atribuible al shock económico de la pandemia**, aunque menos marcada.

**1.3 Serie histórica de ingreso real**



Una vez ajustado por inflación, la dinámica del ingreso cambia notablemente. Mientras que el nominal crece de manera explosiva en los últimos años, el ingreso real se mantiene relativamente estable entre 2016 y 2019.

En 2020–2021 se produce una **caída pronunciada del ingreso real**, coincidente con: Restricciones sanitarias, menor actividad económica, pérdida de empleo o reducción de horas trabajadas, problemas de captación en la EPH durante la pandemia. también presenta pequeñas oscilaciones en 2022–2023, con leves descensos intermedios antes de consolidar la recuperación final. Estas variaciones reflejan tanto la dinámica inflacionaria como los ciclos de actualización salarial. Desde 2022 se observa una recuperación parcial, aunque **sin volver completamente a los niveles pre-pandemia.**

## **1.4 Mediana ponderada del ingreso real (indicador principal)**

Es el indicador más robusto para describir la tendencia, ya que no se ve afectada por valores extremos ni por posibles errores de registro que pueden distorsionar la media.

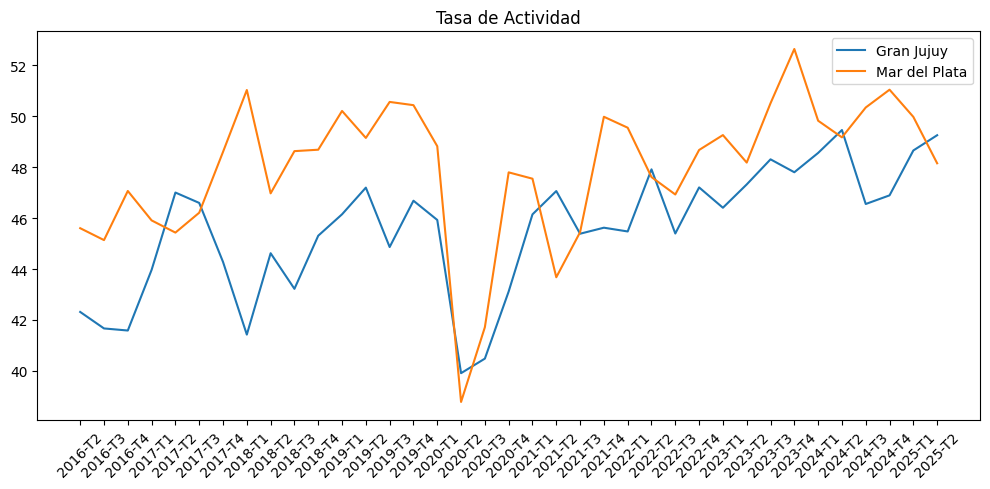
En ambos aglomerados la mediana real se mantiene relativamente estable entre 2016 y 2019, cae bruscamente en 2020–2021 por el impacto de la pandemia, y luego muestra una recuperación parcial hacia 2024–2025. Esta trayectoria representa mejor la evolución del poder adquisitivo que la media, especialmente en contextos con alta inflación.

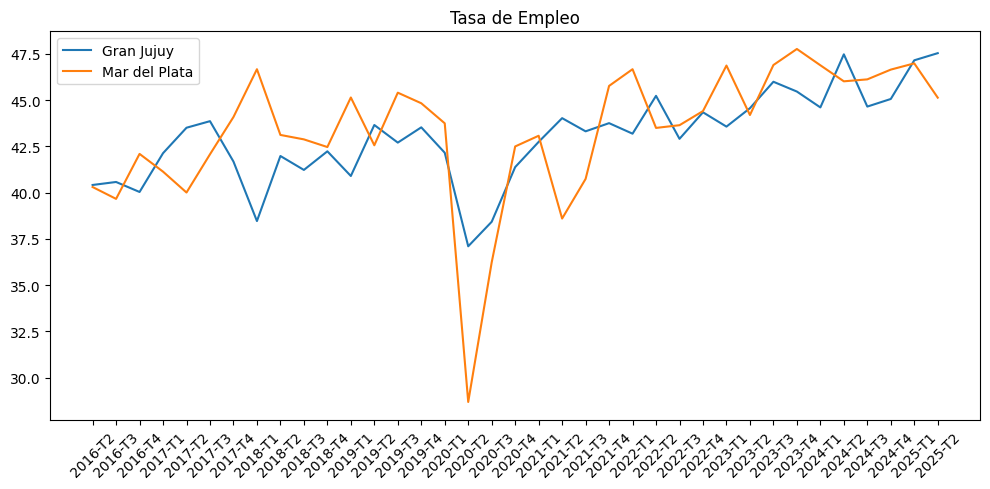
## **1.5 Evolución trimestral de tasas laborales**

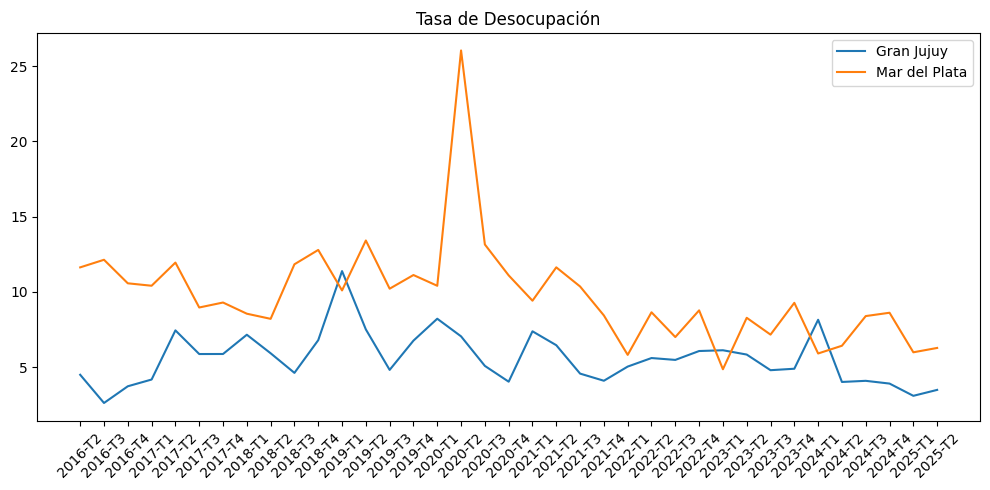
Las tasas laborales muestran una estacionalidad clara a lo largo de los trimestres. Las variaciones más destacadas son:

* 2020: caída de la actividad y el empleo debido al confinamiento, con un aumento excepcional de la desocupación.
* 2021–2022: recuperación progresiva, aunque no completamente simétrica entre ambos aglomerados.
* 2023–2025: estabilización en niveles superiores a los observados en 2016–2019.

El patrón es consistente con los efectos de la pandemia sobre la EPH, donde la interrupción de actividades formales e informales alteró las mediciones habituales del mercado laboral

****

****

****

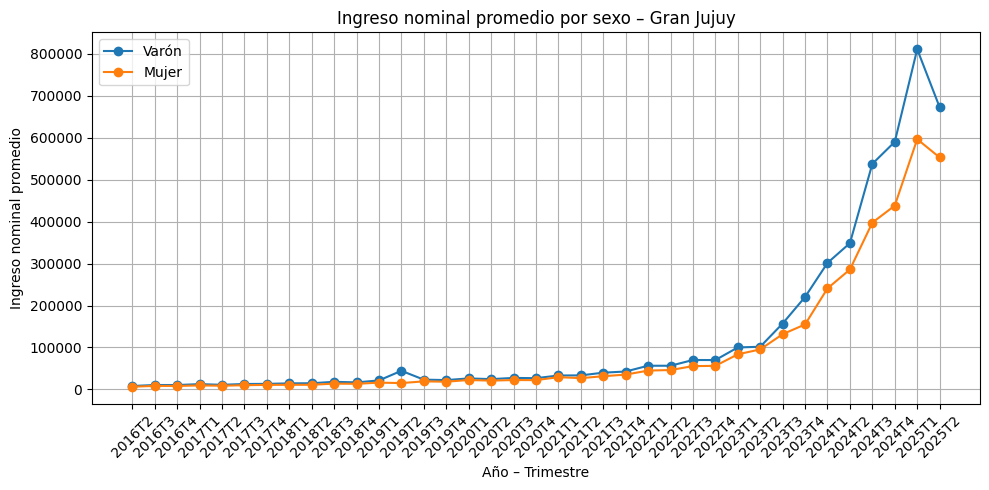
**Análisis multivariado**

Nos permite observar cómo las variables socioeconómicas se relacionan entre sí y cómo afectan de manera diferencial a distintos grupos de la población. Para ello, se estudian las tasas laborales y los ingresos —nominales y reales— según tres dimensiones clave: sexo, grupo de edad y nivel educativo.

**2.1 Ingreso nominal promedio por sexo**

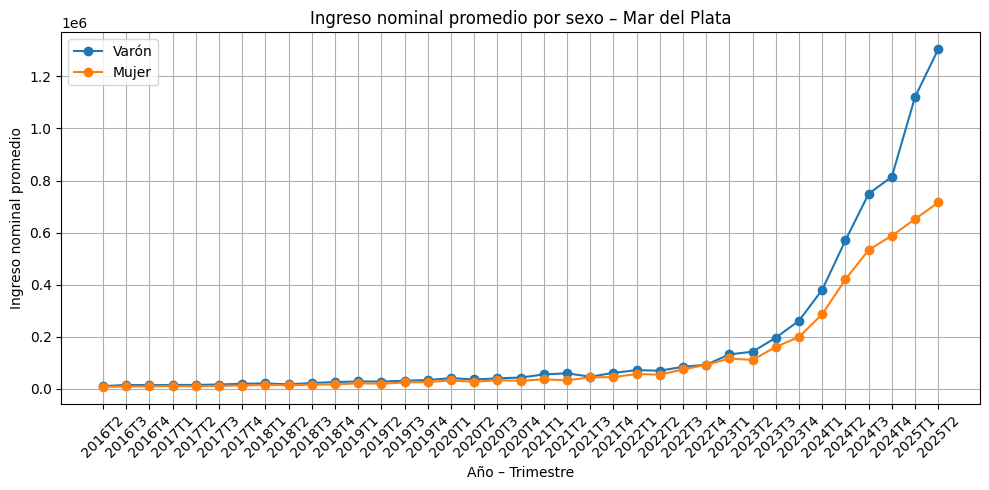
En ambos aglomerados se observan diferencias sistemáticas entre varones y mujeres.

Durante todo el período 2016–2025, los varones presentan ingresos nominales más altos que las mujeres, aunque ambas series se incrementan con fuerza a partir de 2022, acompañando el contexto inflacionario.



**Gran Jujuy**

La brecha de género se mantiene relativamente estable hasta 2021, pero se amplía entre 2023 y 2025 debido al mayor ritmo de crecimiento de los ingresos masculinos.

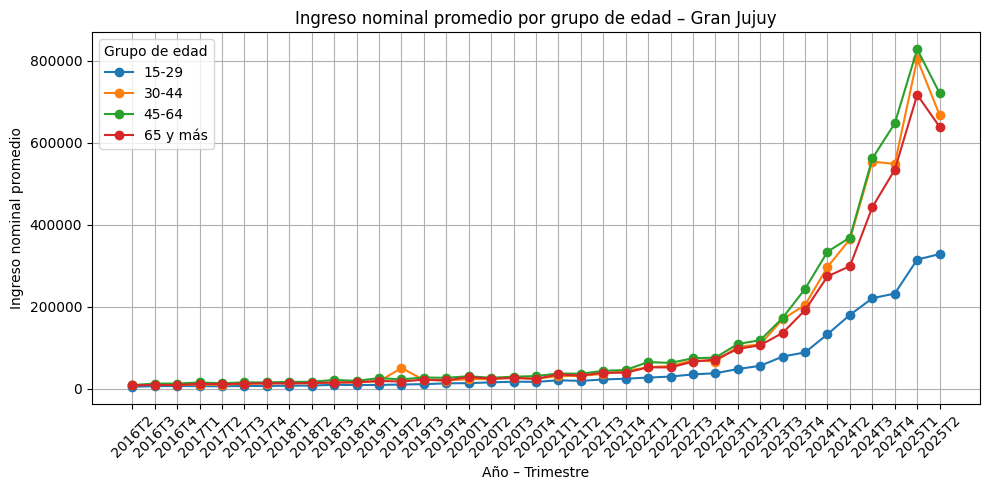


**Mar del Plata**

Presenta una brecha más marcada en los últimos años del período, especialmente en 2024–2025, donde el ingreso de los varones crece con mayor velocidad.

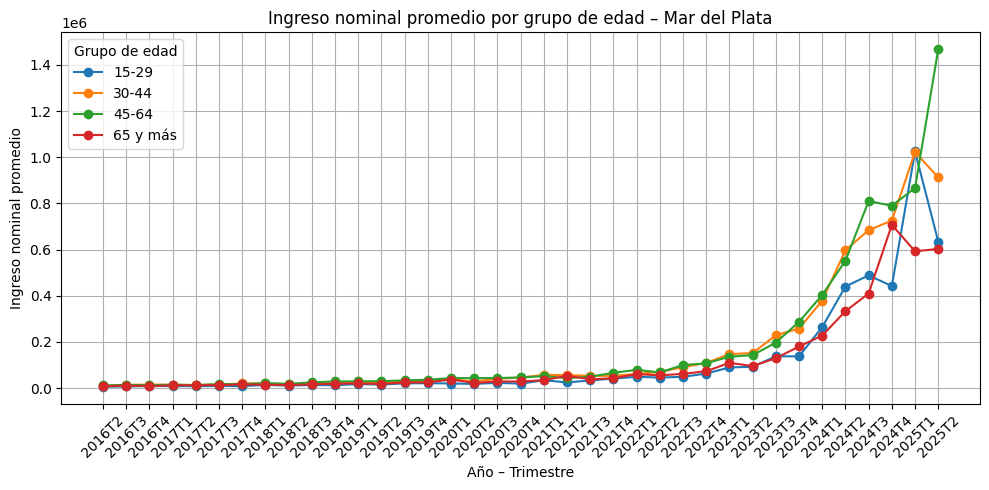
**2.2 Ingreso nominal promedio por grupo de edad**

Los ingresos muestran una clara asociación con la edad: los “adultos” y “adultos mayores” presentan ingresos promedio más elevados, mientras que los jóvenes tienen los valores más bajos. Esto es consistente con trayectorias laborales más largas y estabilidad ocupacional.



**Gran Jujuy**:

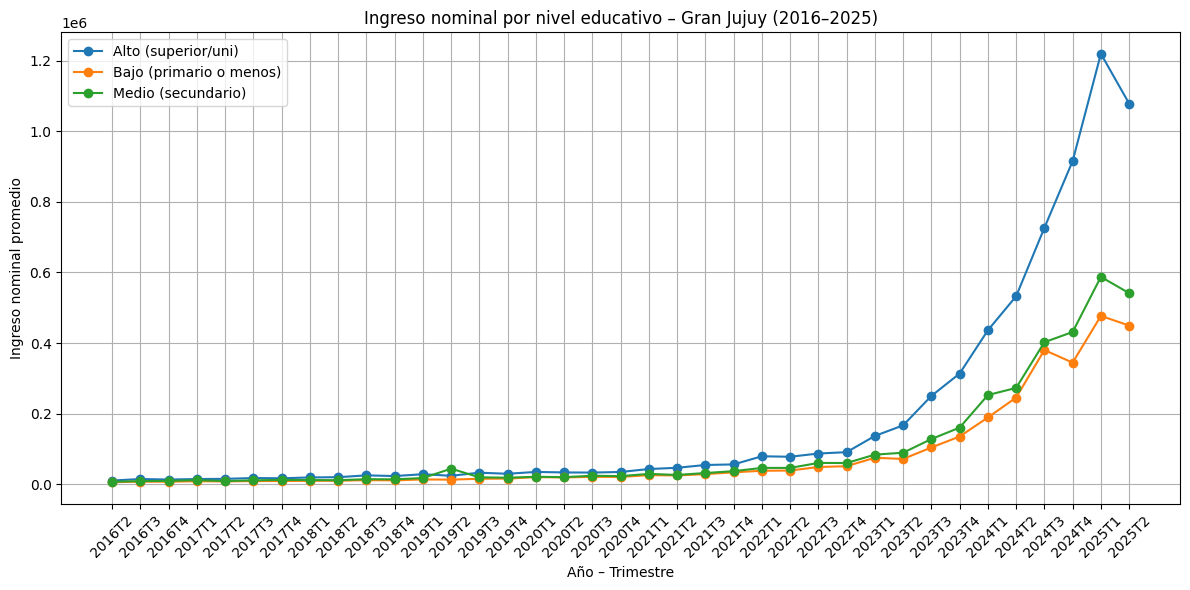
La distancia entre grupos se amplía notablemente a partir de 2022. Los jóvenes presentan un crecimiento más moderado.



**Mar del Plata**:

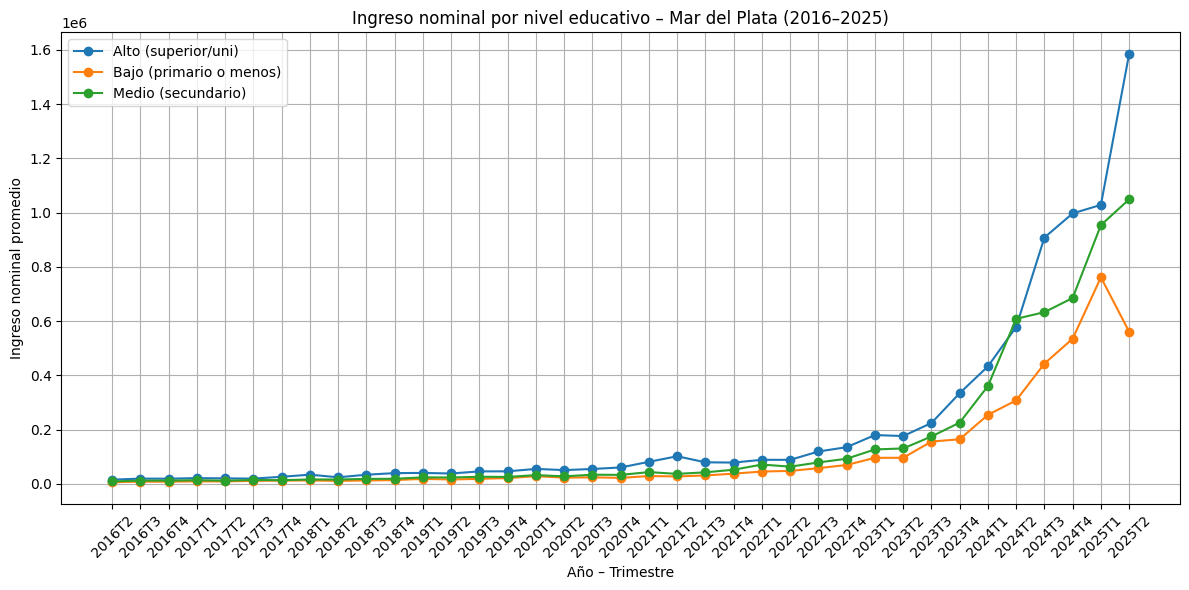
El grupo de mayores registra el salto más pronunciado hacia 2024–2025; la brecha con los jóvenes se vuelve más marcada.

**2.3 Ingreso nominal promedio por nivel educativo**



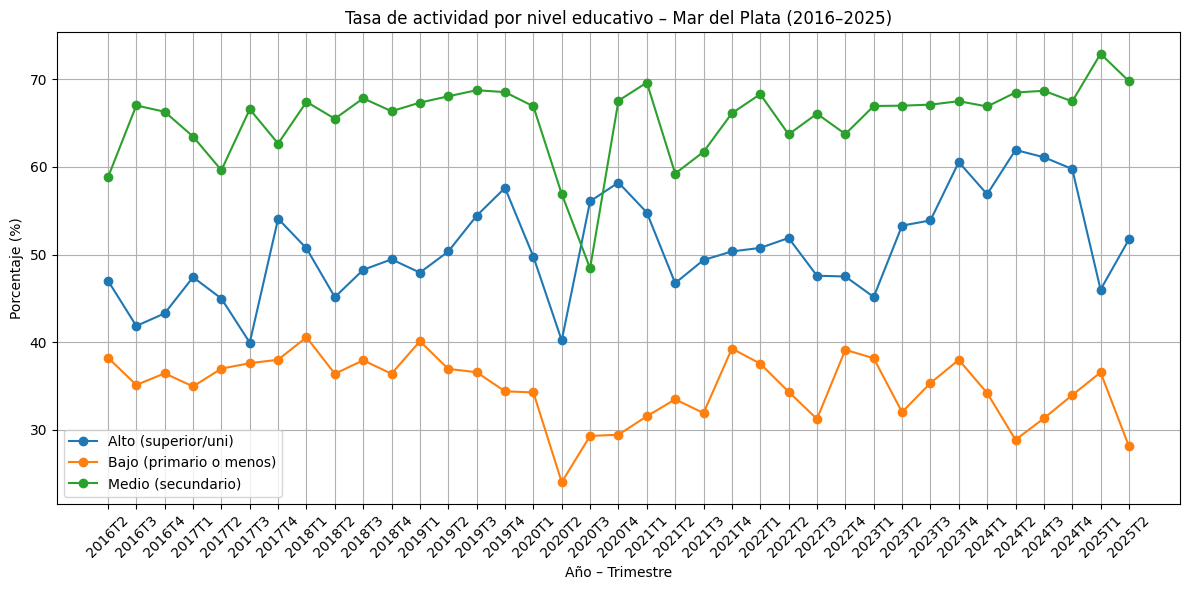
El nivel educativo es uno de los factores más determinantes del ingreso.

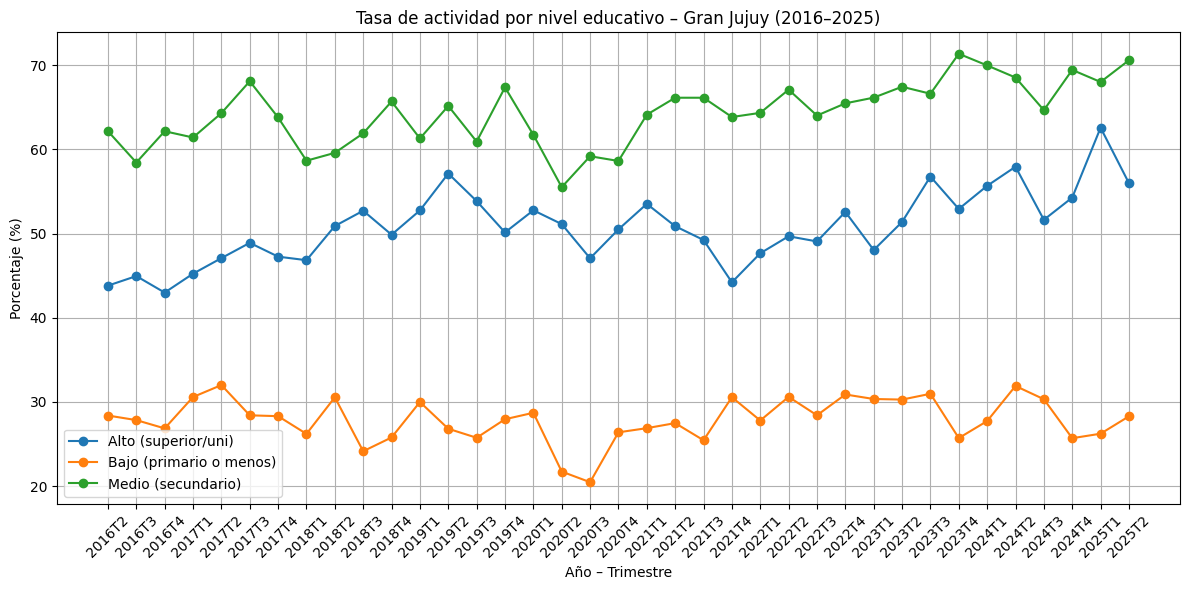
Las personas con **educación superior completa** registran ingresos que duplican o triplican a quienes tienen educación básica o incompleta.

Entre 2016 y 2021 las diferencias se mantienen estables, pero desde 2022 la brecha se amplía por completo: 

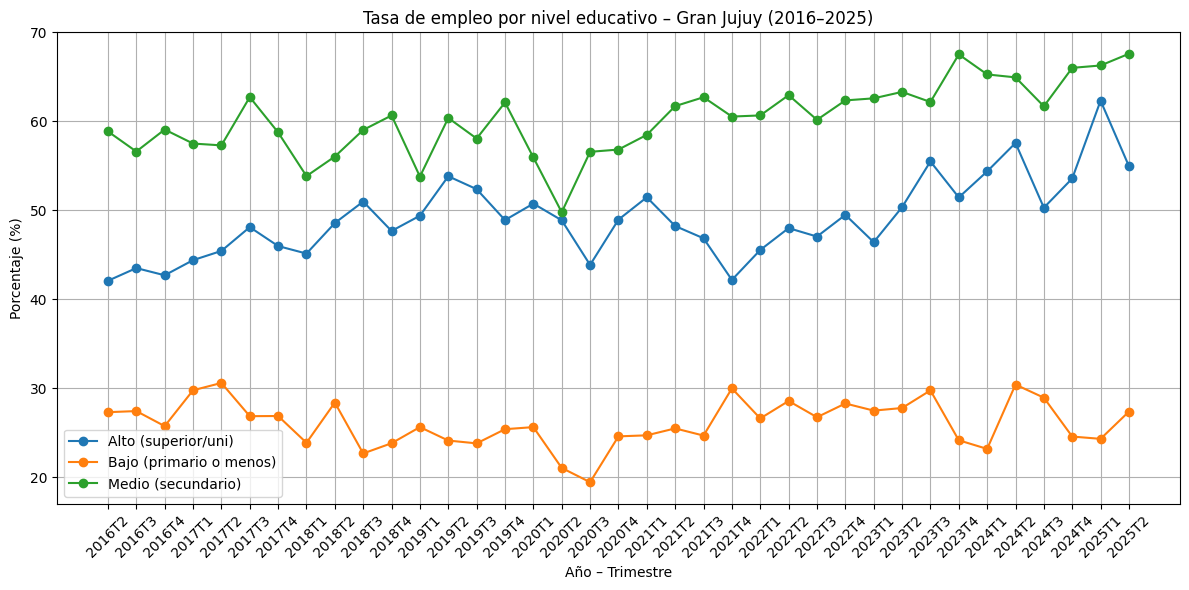
Los salarios de personas con estudios superiores crecen más rápido que los de niveles bajos. Esto evidencia la fuerte segmentación educativa del mercado laboral.

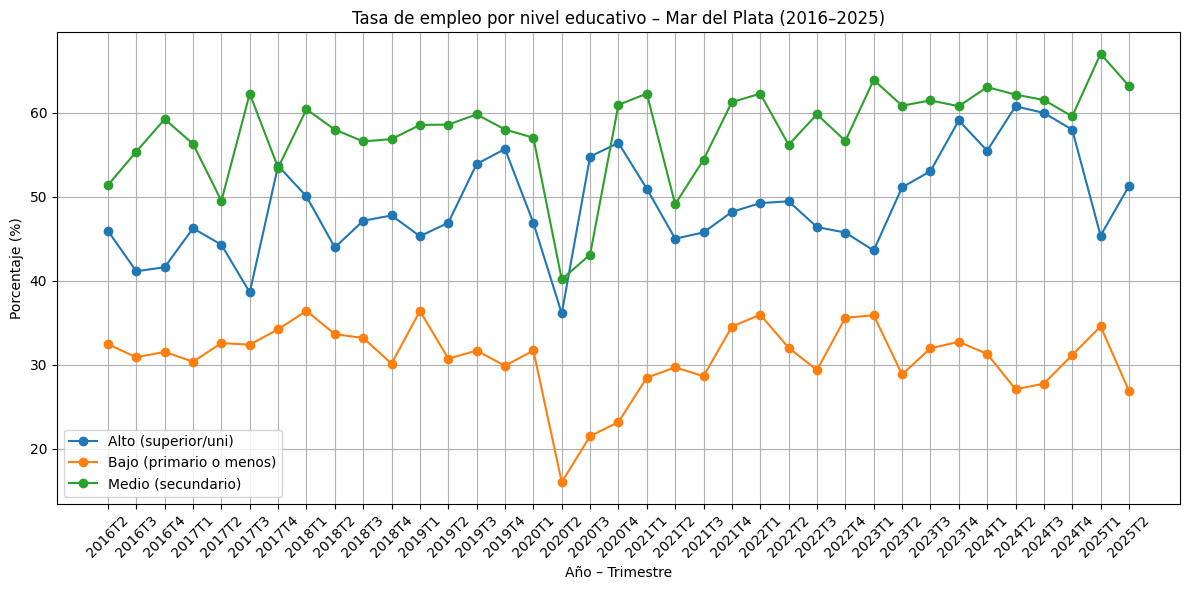
**2.4 Tasas laborales por nivel educativo**

*Tasa de actividad*

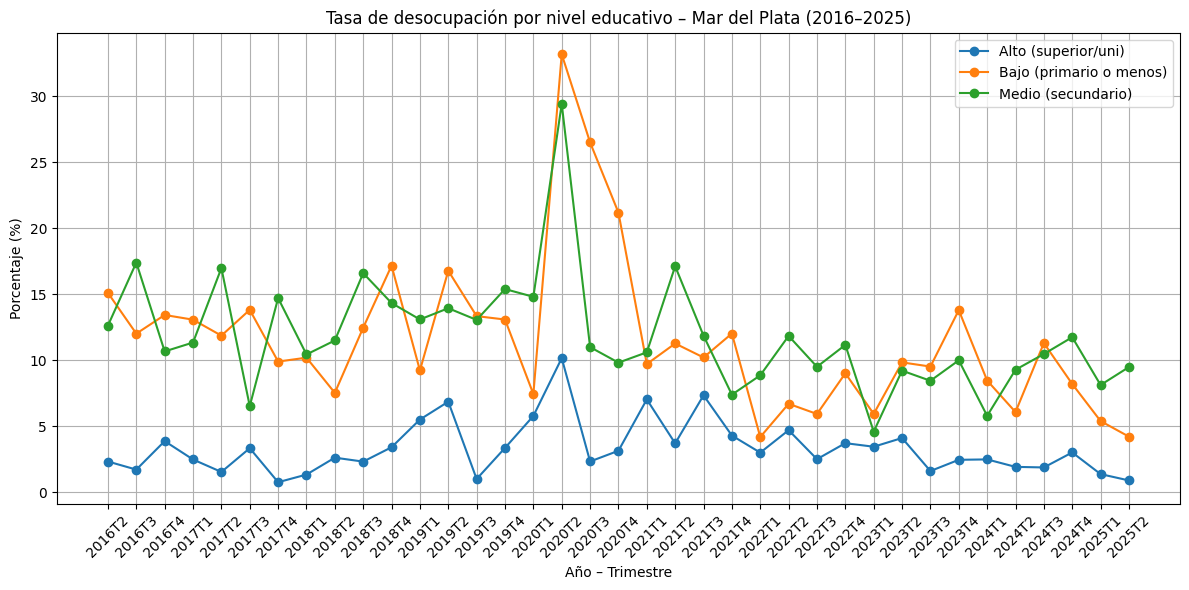


*Tasa de empleo*



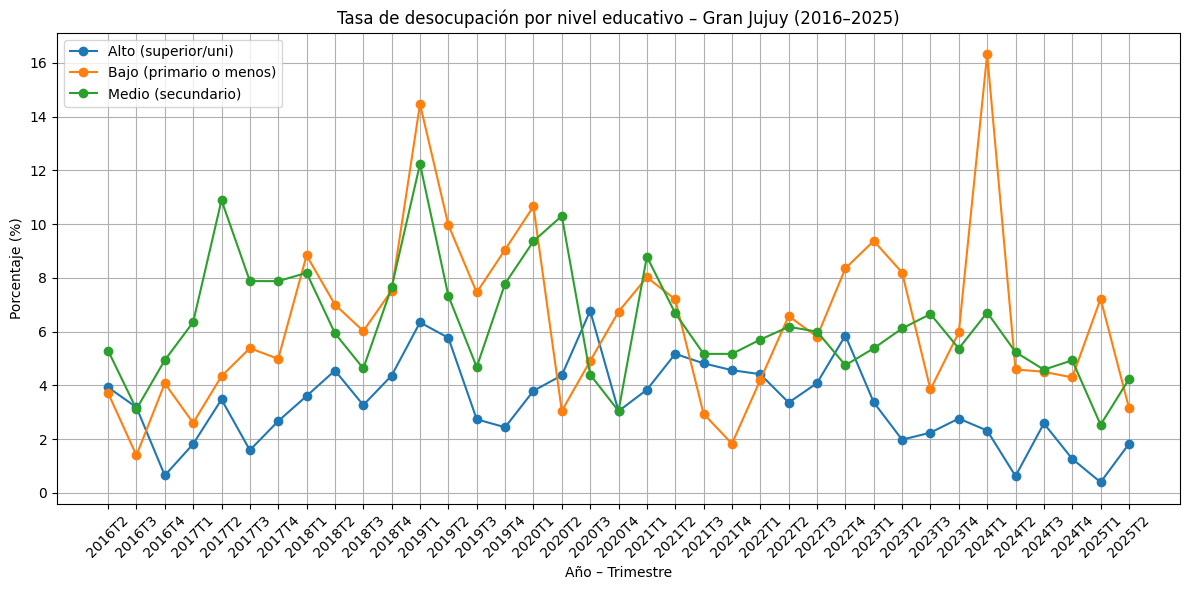


*Tasa de desocupación*



Las tasas laborales muestran patrones muy diferenciados según el nivel educativo.

Las personas con mayor educación exhiben mayores tasas de actividad y empleo, y menores tasas de desocupación.

Los niveles bajos presentan mayor vulnerabilidad laboral, especialmente durante 2020–2021, cuando la pandemia impacta con fuerza: la desocupación aumenta de forma marcada en estos grupos.

Desde 2022, la recuperación es más rápida entre los niveles educativos altos, lo que amplía la brecha laboral.

**2.5 Comparativa sexo + educación + edad (síntesis)**

El análisis combinado de sexo, edad y educación confirma que la educación es el factor con mayor efecto estructural sobre los ingresos y las tasas laborales.

El sexo introduce una brecha persistente que no se reduce con el tiempo, mientras que la edad refleja el ciclo laboral: los adultos en edades centrales alcanzan los mayores ingresos.

Durante 2020–2021 se observan distorsiones en todas las variables, pero afectan especialmente a jóvenes y a personas con bajo nivel educativo. Desde 2022, la recuperación favorece más a los grupos de mayor educación, lo que profundiza desigualdades preexistentes.

**2.6 Síntesis del análisis multivariado**

Sexo: Persiste la brecha salarial a favor de los varones.  
Edad: El ciclo laboral influye fuerte; los adultos mayores ganan más y tienen mayor estabilidad.  
Educación: Es el determinante más importante; mayor educación implica mejores ingresos y tasas laborales.  
Pandemia: 2020–2021 impacta más a jóvenes y personas con bajo nivel educativo.  
Tendencia reciente: Desde 2022 aumentan los ingresos nominales, pero el poder adquisitivo no se recupera completamente.

**Imputación de la no respuesta a ingresos.**

**4.1 Modelo utilizado**

Se utiliza un modelo de regresión lineal múltiple, donde el ingreso depende de factores estructurales como: Nivel educativo, Edad, Sexo, Aglomerado urbano, Condición laboral (ocupado / desocupado / inactivo)  
Este tipo de modelo permite estimar el “retorno” de cada variable sobre el ingreso , Obtener un valor predicho continuo, Realizar una imputación consistente con el resto de los datos, Evaluar su desempeño mediante métricas como MAE y RMSE.

## **4.2 Preparación de la base**

Se divide la base en dos subconjuntos: df\_train → personas con ingreso declarado (P47T > 0) y df\_missing → personas sin ingreso declarado, para imputar.

Se construyen variables explicativa: • Dummies de nivel educativo; Bajo (primario o menos), Medio (secundario), Alto (superior / universitario) — categoría base • Dummies de edad en grupos; 30–44, 45–64, 65+, Jóvenes como categoría base • Dummy de sexo (Varón) • Aglomerado (19 o 34)

## **4.3 Estimación del modelo**

El modelo OLS se ajusta sobre las observaciones con ingreso declarado.

### **Coeficientes principales del modelo**

Los coeficientes estimados permiten identificar cuáles variables tienen mayor influencia sobre el ingreso. El signo y magnitud indican cómo cambia el ingreso esperado cuando la persona pertenece a cada categoría, manteniendo las demás constantes.

* **Educación baja** → Reduce fuertemente el ingreso esperado.
* **Educación media** → También reduce el ingreso en comparación con educación superior.
* **Edad 45–64** → Incrementa el ingreso respecto de los más jóvenes.
* **Edad 65+** → Incremento moderado.
* **Sexo varón** → Presenta ingresos esperados más altos.
* **Aglomerado** (Mar del Plata) → Efecto positivo leve.

### **Métricas generales del modelo**

* **MAE:** 130.186,89
* **RMSE:** 271.876,44

Estas métricas reflejan la variabilidad natural de los ingresos y la presencia de valores altos, pero aun así el modelo captura razonablemente bien la estructura de los datos.

**4.4 Imputación de los ingresos faltantes**

Con el modelo estimado, se predicen los ingresos faltantes de las observaciones en **df\_missing**. Luego se unifica toda la base en una sola variable: **ingreso\_final** = ingreso declarado (si existía) **ingreso\_imputado** (si no existía)

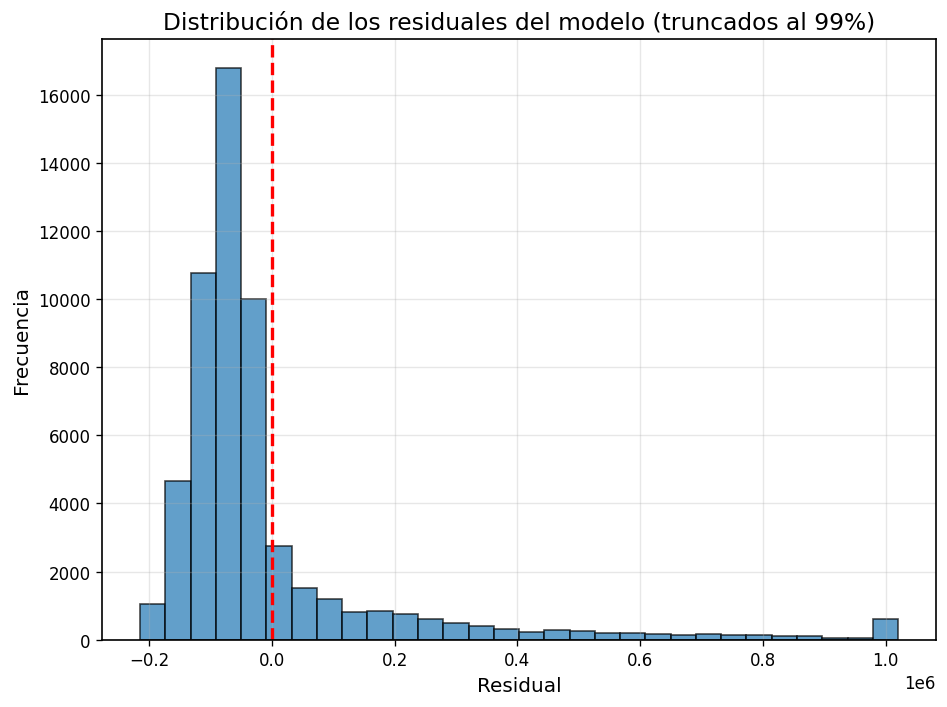
### **Resultado final**

* **Observaciones con ingreso declarado:** 55.782
* **Observaciones imputadas:** 37.298
* **Total de casos completos luego de la imputación:** **93.080**

Esto garantiza una base completa y homogénea para análisis posteriores, manteniendo la estructura original de la EPH.

**4.5 Evaluación gráfica del modelo**

### **Ingreso observado vs ingreso predicho - Residuales vs ingreso predicho - Histograma de los residuales**

En el **primer gráfico** se evalúa la capacidad del modelo para aproximar los valores reales. Dado que el ingreso presenta valores extremos y una fuerte asimetría, se utilizó una transformación logarítmica (log10), lo que permite visualizar mejor la relación entre lo observado y lo predicho. La línea roja representa la predicción perfecta. Se observa que el modelo captura bien el patrón general, aunque presenta mayor dispersión en los ingresos más altos, comportamiento típico en modelos de ingreso.

El **segundo gráfico** (residuales vs ingreso predicho) se utiliza para verificar: homocedasticidad, patrón de dispersión, presencia de outliers. Los residuales muestran más variabilidad a medida que aumenta el ingreso predicho, algo esperable por la naturaleza del ingreso laboral. No aparecen patrones sistemáticos ni curvaturas, lo que indica que la relación lineal es razonable. Para mejorar la visualización, los residuales fueron truncados al percentil 99.

El **último gráfico** permite evaluar la distribución del error. El histograma está fuertemente sesgado hacia la derecha (colas largas), típico de modelos de ingreso. No obstante, la mayor parte de los residuales se concentra cerca de cero, lo que confirma que el modelo funciona adecuadamente para la imputación.

## **4.6 Síntesis**

Se aplicó un **modelo de regresión lineal múltiple**.

Las variables con mayor impacto fueron **nivel educativo**, **edad** y **sexo**, lo que concuerda con la teoría económica y los resultados del análisis multivariado.

El modelo mostró errores altos esperables para ingresos, pero **suficientes para usarse en imputación**.

Se imputaron correctamente **más de 37 mil casos**, completando la base para el análisis posterior.

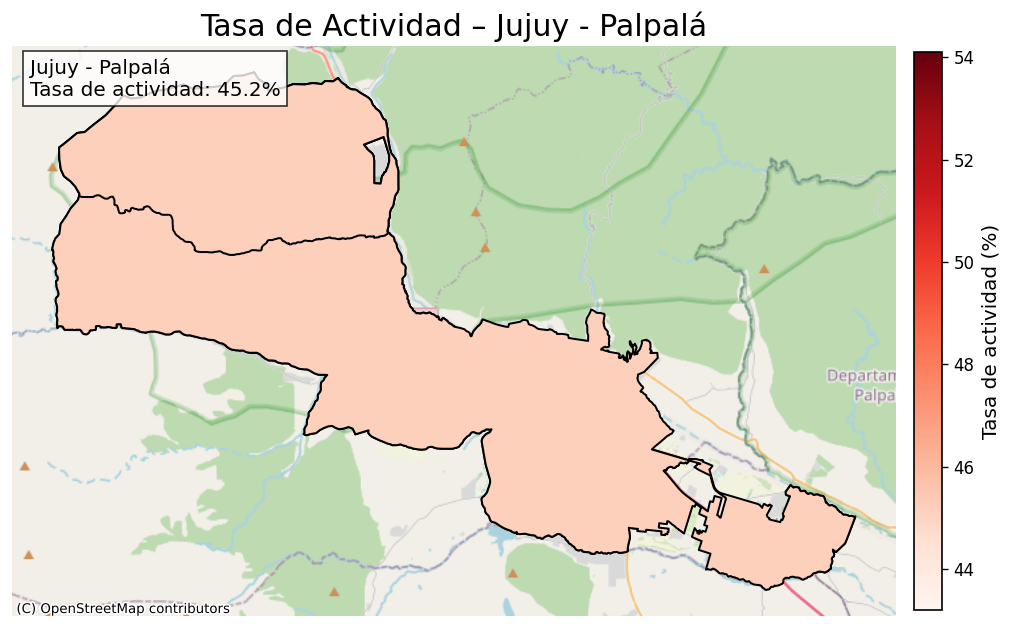
Los diagnósticos gráficos confirman que el modelo es adecuado y no presenta problemas graves de especificación.

**Análisis espacial de los aglomerados (promedios EPH 2016–2025)**

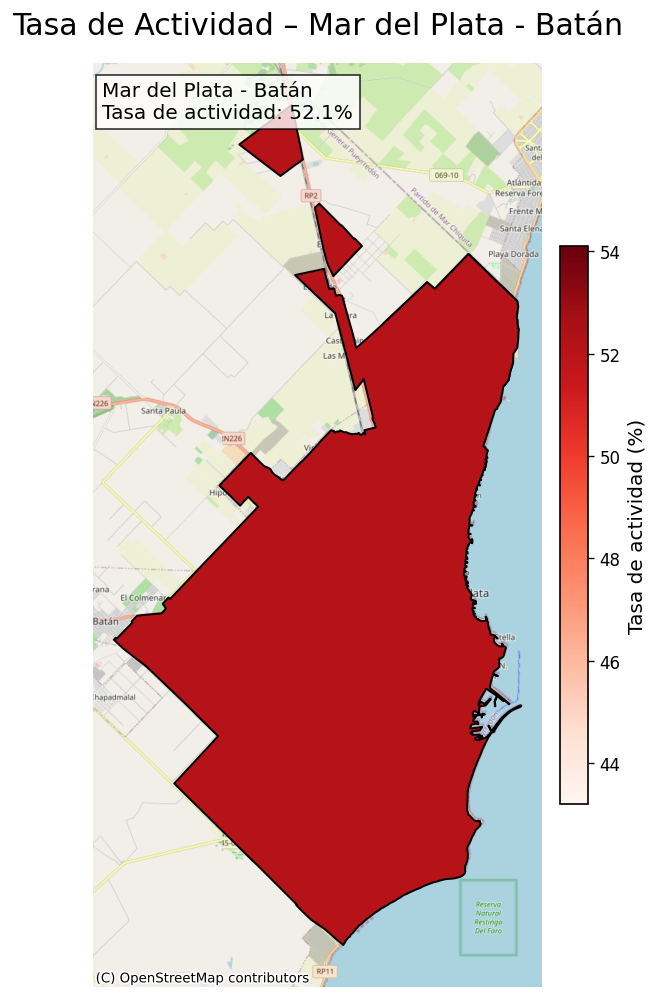
**5. Análisis espacial de los indicadores socioeconómicos**

Para el análisis espacial se utilizaron **los valores promedio de las tasas de actividad y desocupación calculados a partir de la EPH para el período 2016–2025**, abarcando todos los trimestres disponibles. No se representan valores de un trimestre específico, sino **un promedio estructural** que resume el comportamiento laboral de cada aglomerado a lo largo de toda la serie.

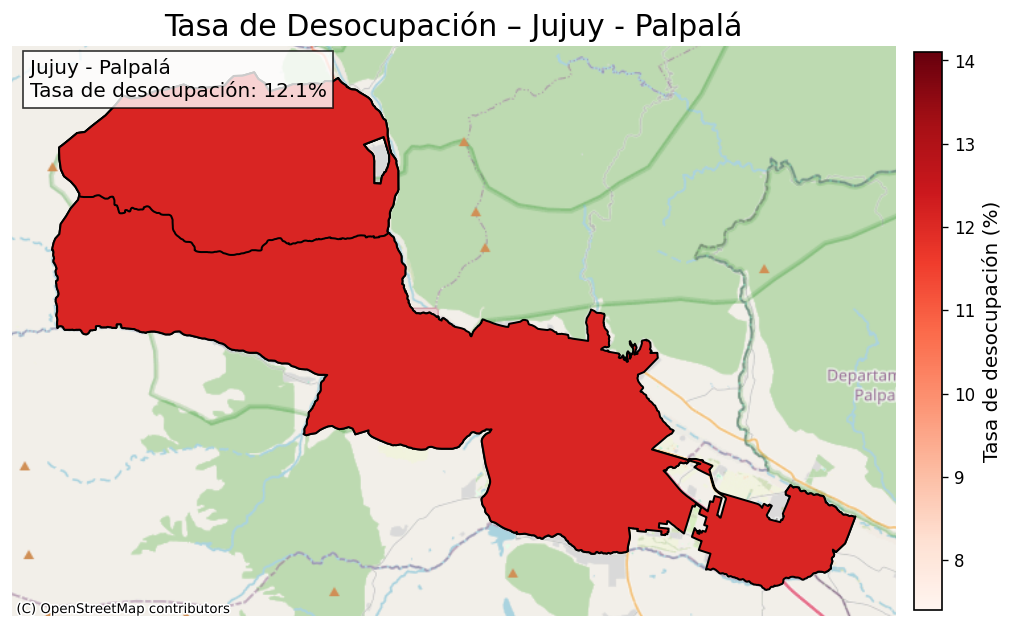
Los mapas fueron elaborados para los dos aglomerados en estudio —**Jujuy–Palpalá** y **Mar del Plata–Batán**— utilizando la **misma escala de colores**, permitiendo una comparación directa entre ambos territorios.

**5.1. Tasa de Actividad**

La tasa de actividad refleja la proporción de la población en condiciones de participar del mercado laboral (ocupados + desocupados).

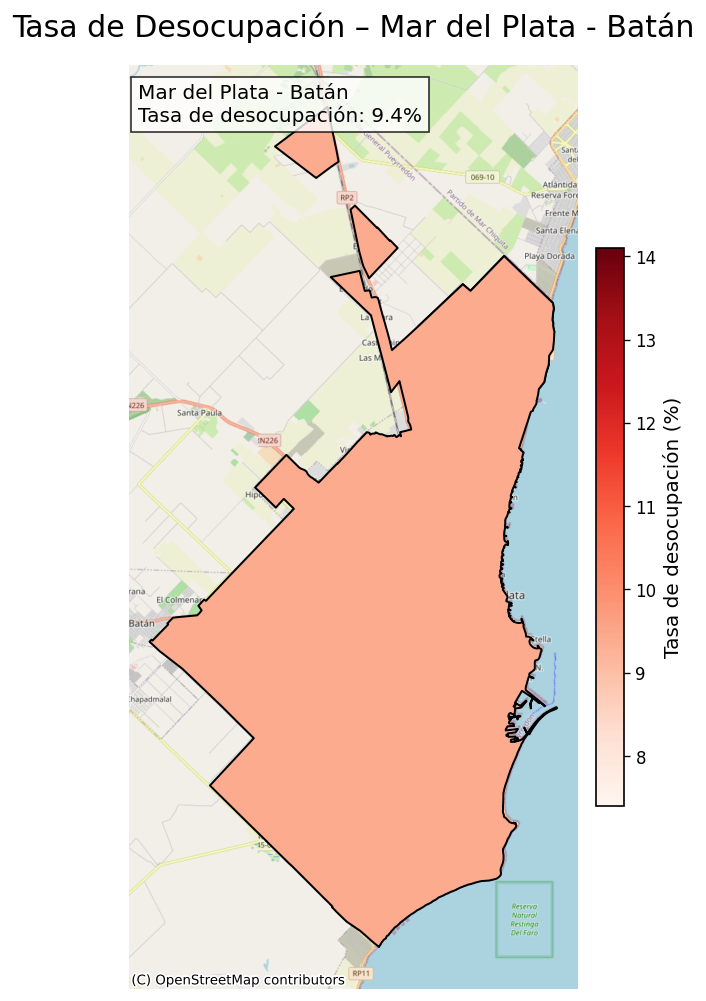
* **Jujuy – Palpalá** presenta una tasa de actividad de 45,2%, ubicándose en el rango medio–bajo de la escala utilizada.  
  En el mapa, esto se visualiza mediante un tono rojizo más claro, lo cual indica una menor incorporación de la población al mercado laboral.
* **Mar del Plata – Batán** alcanza una tasa significativamente más alta: 52,1%, superando ampliamente a Jujuy. Su mapa registra tonalidades rojas más intensas, lo que refleja una mayor participación laboral.

La mayor tasa de actividad de Mar del Plata puede asociarse al perfil económico de la ciudad (servicios, comercio, turismo y actividades urbanas formales), mientras que Jujuy – Palpalá muestra una menor integración laboral, posiblemente vinculada a estructuras productivas más limitadas y menor diversificación económica.



## **5.2. Tasa de Desocupación**

Este indicador revela diferencias importantes entre ambos aglomerados:

* **Jujuy – Palpalá presenta una desocupación del 12,1%**, muy superior al promedio nacional.  
   El mapa la muestra con una intensidad de color marcada, indicando un nivel crítico del fenómeno.
* **Mar del Plata – Batán registra 9,4%**, también elevada pero claramente menor. La tonalidad es más suave en comparación con Jujuy, reflejando una situación relativamente menos grave.

Si bien ambos aglomerados enfrentan problemas de desempleo, la situación de Jujuy es más severa. Esto se relaciona con un mercado laboral más reducido, con menor oferta de empleo privado y una informalidad históricamente más alta. Mar del Plata, aún con altos niveles de desocupación estructural, muestra mejores condiciones relativas.

## 

## **5.3. Lectura comparada**

La comparación conjunta de ambos indicadores permite observar:

* **Mar del Plata combina mayor tasa de actividad con menor desocupación**, señal de un mercado laboral más dinámico.
* **Jujuy** muestra **menor participación laboral y mayor desocupación**, lo cual evidencia una estructura económica más restringida.
* En ambos casos, los mapas permiten identificar **aglomerados urbanos completos con un solo valor promedio**, siendo útil para comparar regiones pero no para análisis intraurbanos.

## **5.4. Conclusiones generales del análisis espacial**

1. Existe una **brecha clara** entre los dos aglomerados: Mar del Plata se ubica sistemáticamente mejor en ambos indicadores.
2. Jujuy combina **baja actividad + alta desocupación**, lo que apunta a mayores dificultades de inserción laboral.
3. Mar del Plata presenta **mayor dinamismo**, aunque con un nivel de desocupación aún preocupante.
4. La representación cartográfica facilita la comparación visual y evidencia cómo ambos territorios, aun siendo urbanos, exhiben dinámicas laborales muy distintas.