Big Data & Machine Learning - Trabajo Práctico N°2

Grupo: 1
Integrantes:

-Benjamin Rodolfo Ayay Quispe (N° Registro 912662)

-Jesús David Ochochoque Mendoza (N° Registro 915690)

Región seleccionada: Gran Buenos Aires

Introducción

Este trabajo práctico tiene como objetivo familiarizarse con la base de datos de la Encuesta Permanente de Hogares (EPH) del INDEC, realizar una limpieza de datos y analizar indicadores sociodemográficos y laborales en la región de Gran Buenos Aires. Utilizamos las bases de microdatos del primer trimestre de 2004 y 2024, enfocándonos en variables como sexo, edad, nivel educativo, condición de actividad e ingreso per cápita familiar. El análisis incluye la identificación de valores faltantes, corrección de datos inválidos y un estudio comparativo de la población económicamente activa (PEA), la población en edad de trabajar (PET) y la desocupación entre ambos años.

Parte I: Familiarización con la Base EPH y Limpieza

1. Identificación de personas desocupadas

Según el INDEC, las personas desocupadas son aquellas que no trabajaron ni una hora en la semana de referencia, buscan activamente empleo y están disponibles para trabajar inmediatamente (INDEC, 2025). En la base EPH, esto se refleja en la variable ESTADO = 2.

2. Carga, filtrado y unificación de datos

Se cargaron las bases de 2004 (usu_individual_T104.dta) y 2024 (usu_individual_T124.xlsx), filtrando por la región Gran Buenos Aires (REGION = 'Gran Buenos Aires' en 2004, REGION = '1' en 2024). Se seleccionaron 15 variables comunes: CH04 (sexo), CH06 (edad), CH07 (estado civil), CH08 (cobertura de salud), NIVEL_ED (nivel educativo), ESTADO (condición de actividad), CAT_INAC

(categoría inactividad), IPCF (ingreso per cápita familiar), P21 (ingreso ocupación principal), PP04D_COD (código de ocupación), P47T (ingreso total), REGION, AGLOMERADO, TRIMESTRE y ANO4 (año). Las bases se unificaron en un solo dataframe con 7647 observaciones para 2004 y 7051 para 2024.

3. Valores faltantes

Tabla 1: Valores faltantes por variable y año

```
Valores faltantes por año:
ANO4          2004  2024
CH04             0     0
CH06           135     0
CH07             0     0
CH08             0     0
NIVEL_ED         0     0
ESTADO           0     0
CAT_INAC         0     0
IPCF             0     0
P21              0     0
PP04D_COD        0  3827
P47T             0    41
REGION           0     0
AGLOMERADO       0     0
TRIMESTRE        0     0
ANO4             0     0
Cantidad de valores NaN en ingresos después de limpiar:
P47T: 1333
IPCF: 0
```

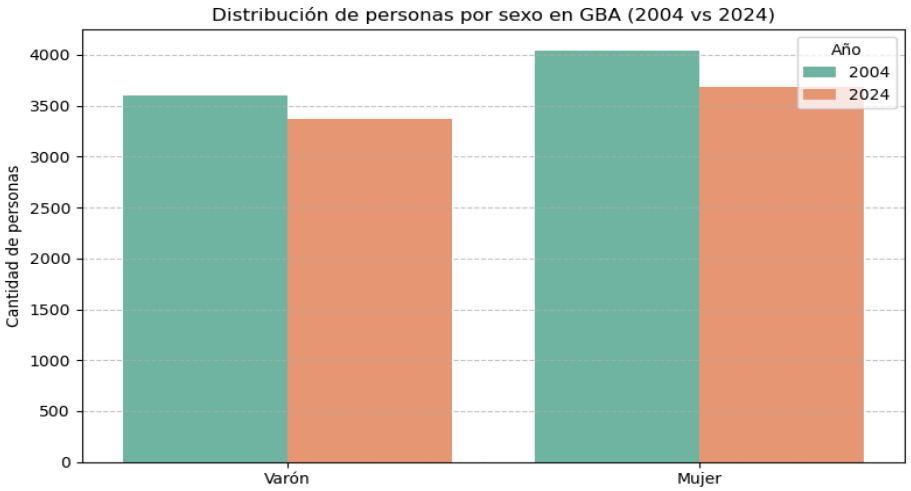
La Tabla 1 *presenta el número de valores faltantes para cada variable en 2004 y 2024. En 2004, la variable CH06 (edad) tiene 135 valores faltantes, mientras que en 2024, PP04D_COD (código de ocupación) tiene 3827 valores faltantes y P47T (ingreso total) tiene 41. El alto número de valores faltantes en PP04D_COD en 2024 sugiere que muchos encuestados no proporcionaron información sobre su ocupación, posiblemente por preocupaciones de privacidad o inaplicabilidad.*

4. Limpieza de datos

Se identificaron ingresos negativos en P47T (ingreso total) y IPCF (ingreso per cápita familiar), que según la documentación de la EPH indican no respuestas o errores de codificación. Estos valores se reemplazaron por NaN, resultando en 1333 valores faltantes en P47T y ninguno en IPCF. Esta limpieza asegura la consistencia de los datos para el análisis posterior.

Parte II: Primer Análisis Exploratorio1. Composición por sexo

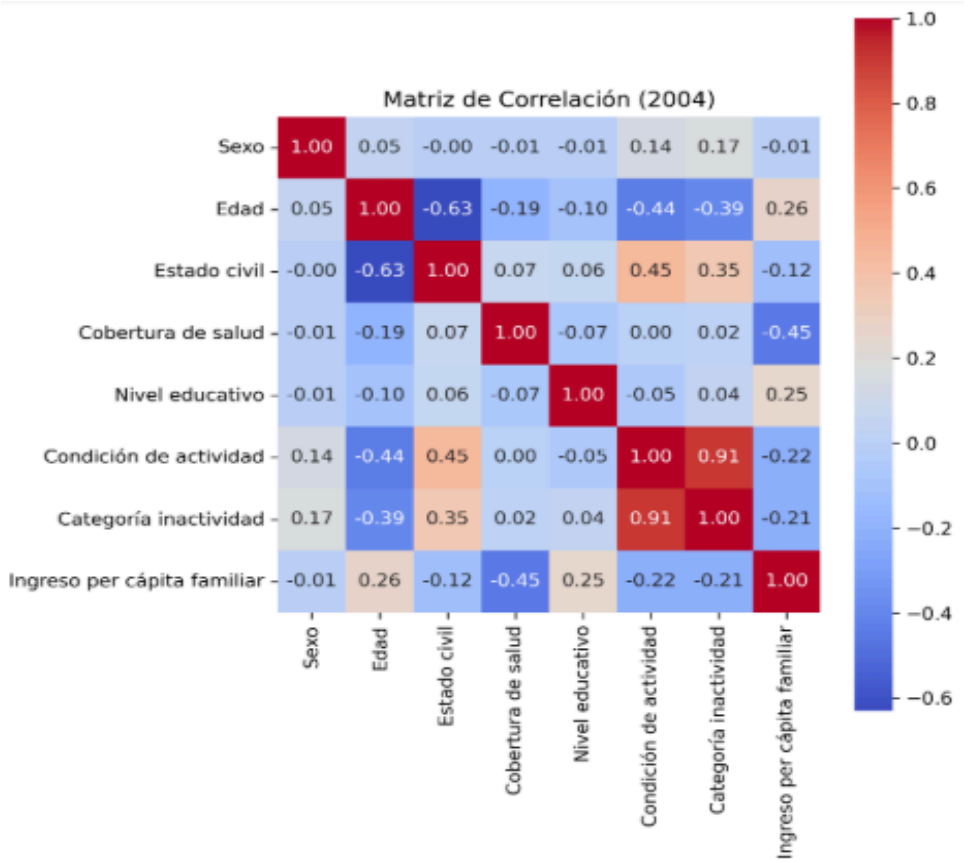
Figura 1: Distribución de personas por sexo en GBA (2004 vs 2024)



La Figura 1 *ilustra la distribución de personas por sexo en Gran Buenos Aires para los años 2004 y 2024. El gráfico de barras muestra que en ambos años la distribución es relativamente equilibrada, con un ligero aumento en la proporción de mujeres de aproximadamente 50% en 2004 a 51% en 2024. Este cambio podría atribuirse a cambios demográficos, como una mayor esperanza de vida entre las mujeres.*

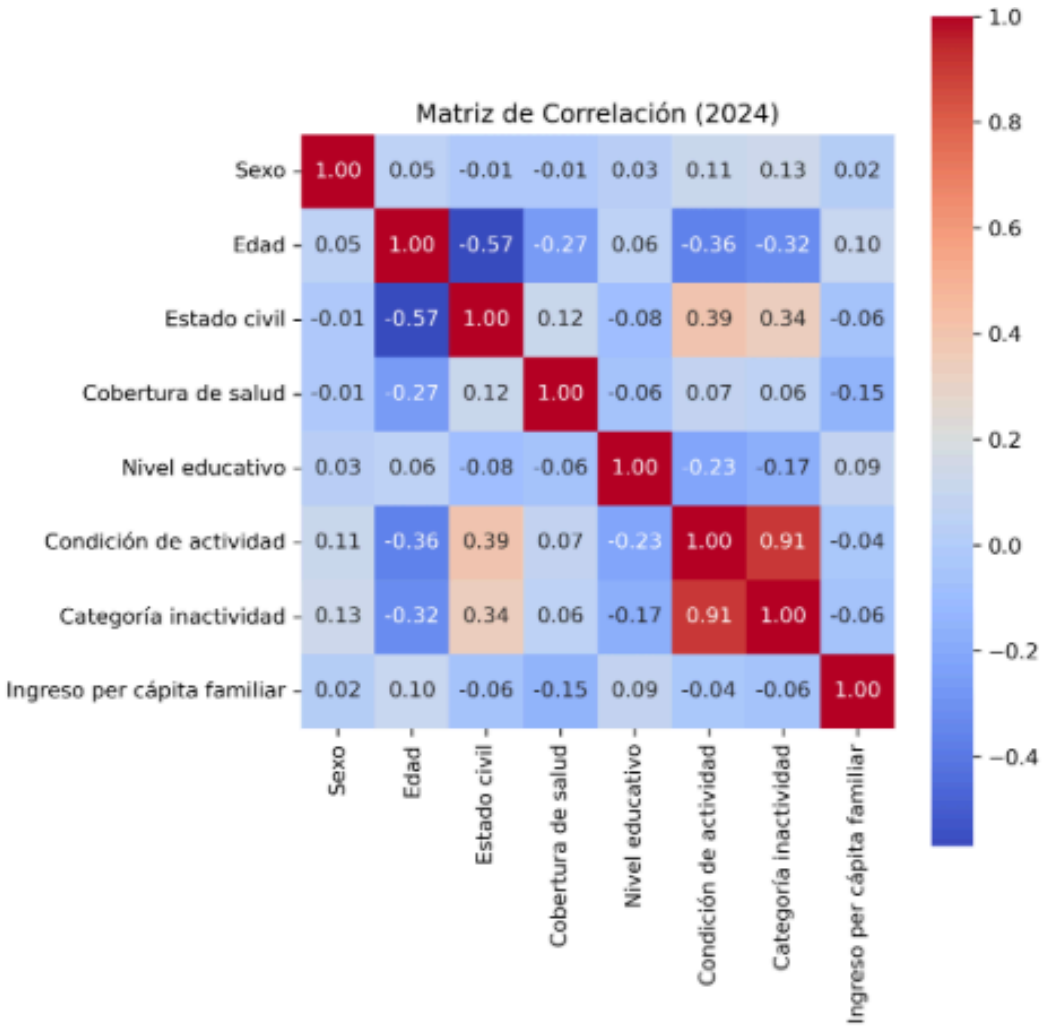
2. Matriz de correlación

Figura 2: Matriz de Correlación (2004)



En 2004, destaca la fuerte correlación positiva entre la condición de actividad y la categoría de inactividad (r = 0.91), lo cual es esperable, ya que ambas variables están conceptualmente vinculadas en torno a la participación en el mercado laboral. Además, se observa una correlación moderadamente negativa entre cobertura de salud e ingreso per cápita familiar (r = -0.45), lo que podría sugerir que en ese periodo los hogares con menores ingresos dependían más de sistemas de salud pública.

Figura 3: Matriz de Correlación (2024)



En 2024, la correlación entre condición de actividad y categoría de inactividad sigue siendo elevada (r = 0.91), aunque se nota una menor asociación entre cobertura de salud e ingreso per cápita (r = -0.15), lo que podría reflejar una mayor cobertura sanitaria universal o desvinculación del acceso a la salud del nivel de ingreso. Además, se observa una reducción general en la intensidad de las correlaciones, lo que podría indicar una mayor diversificación en los perfiles socioeconómicos.

Parte III: Conociendo a los Ocupados y Desocupados

1. Análisis de desocupados, inactivos y media de IPCF

Tabla 2: Media de IPCF por estado y año

```
Media de IPCF por ESTADO y AÑO:
ESTADO    Desocupado  \
ANO4
2004      224.231970
2024    85019.145466

ESTADO  Entrevista individual no realizada (no respuesta al cuestion  \
ANO4
2004                                                52.533333
2024                                                      NaN

ESTADO       Inactivo  Menor de 10 años  No respondió        Ocupado
ANO4
2004       315.891856        246.259032           NaN     476.064755
2024    130704.601499     104353.663296           0.0  207644.844045
```

```
Conteo por ESTADO y AÑO:
ESTADO  Desocupado  \
ANO4
2004           528
2024           311

ESTADO  Entrevista individual no realizada (no respuesta al cuestic
ANO4
2004                                                   10
2024                                                    0

ESTADO  Inactivo  Menor de 10 años  No respondió  Ocupado
ANO4
2004        2800              1230             0     3079
2024        2662               813            41     3224
```

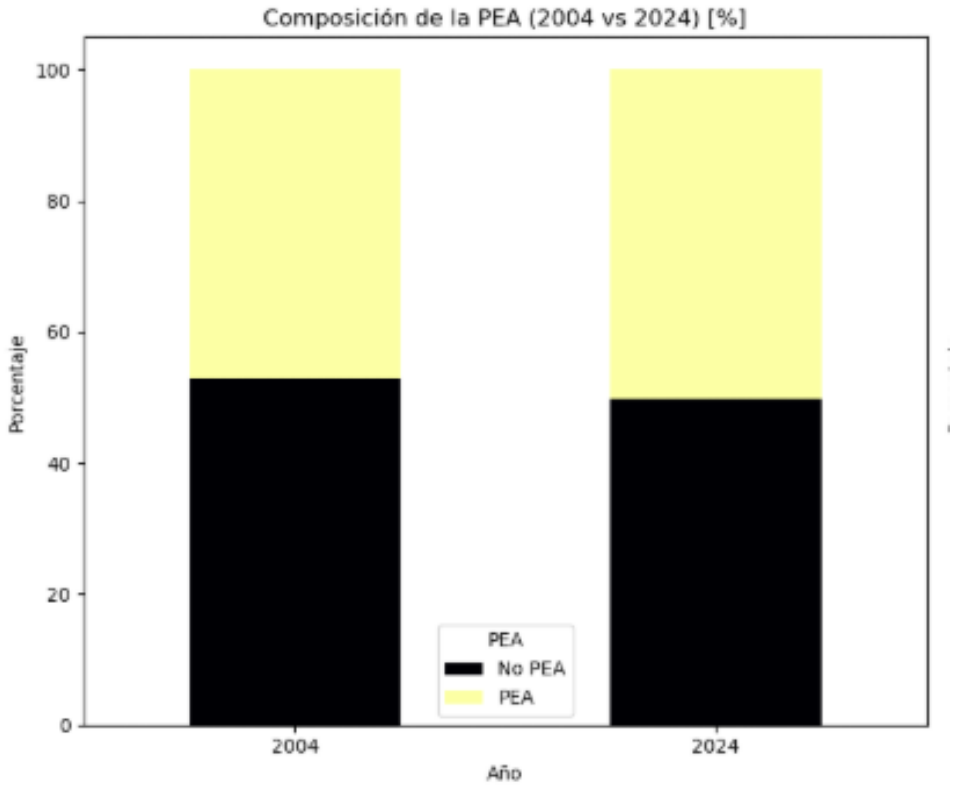
La Tabla 2 muestra el ingreso per cápita familiar promedio (IPCF) para diferentes estados de actividad en 2004 y 2024. En 2004, los ocupados tuvieron el IPCF promedio más alto (~476), mientras que los desocupados tuvieron un promedio de 224. En 2024, el IPCF promedio para los ocupados aumentó significativamente a 207,644, reflejando la inflación y cambios económicos. Los desocupados en 2024 tienen un IPCF promedio de 85,019, mucho mayor que en 2004, posiblemente debido a beneficios sociales u otras fuentes de ingresos

2. Análisis de no respuesta en condición de actividad

Se encontraron personas que no respondieran su condición de actividad en 2024 se dicen q son 41 personas .(ESTADO = "No respondió"),

3. Población Económicamente Activa (PEA)

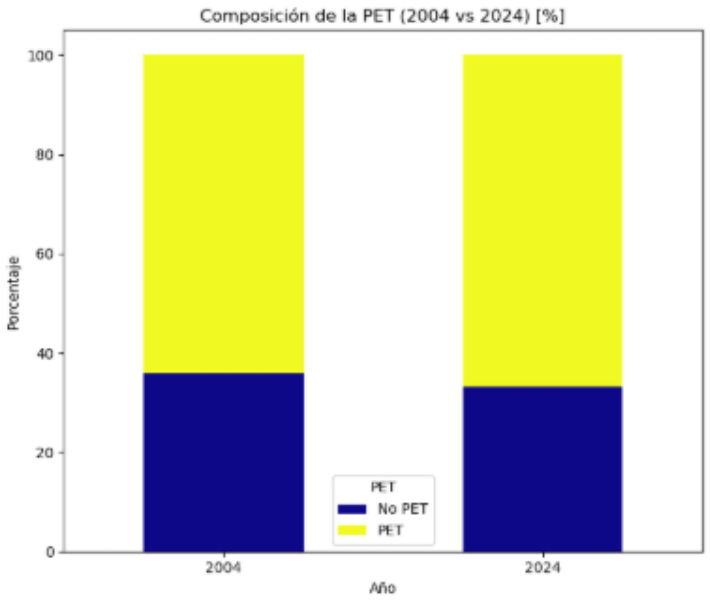
Figura 4: Composición de la PEA (2004 vs 2024) [%]



La Figura 4 muestra la composición de la Población Económicamente Activa (PEA) como porcentaje para 2004 y 2024. El gráfico de barras revela una disminución de la PEA del 53% en 2004 a aproximadamente el 50% en 2024. Esta disminución puede deberse a un aumento de la inactividad (estudiantes, jubilados) o al desaliento laboral.

4. Población en Edad para Trabajar (PET)

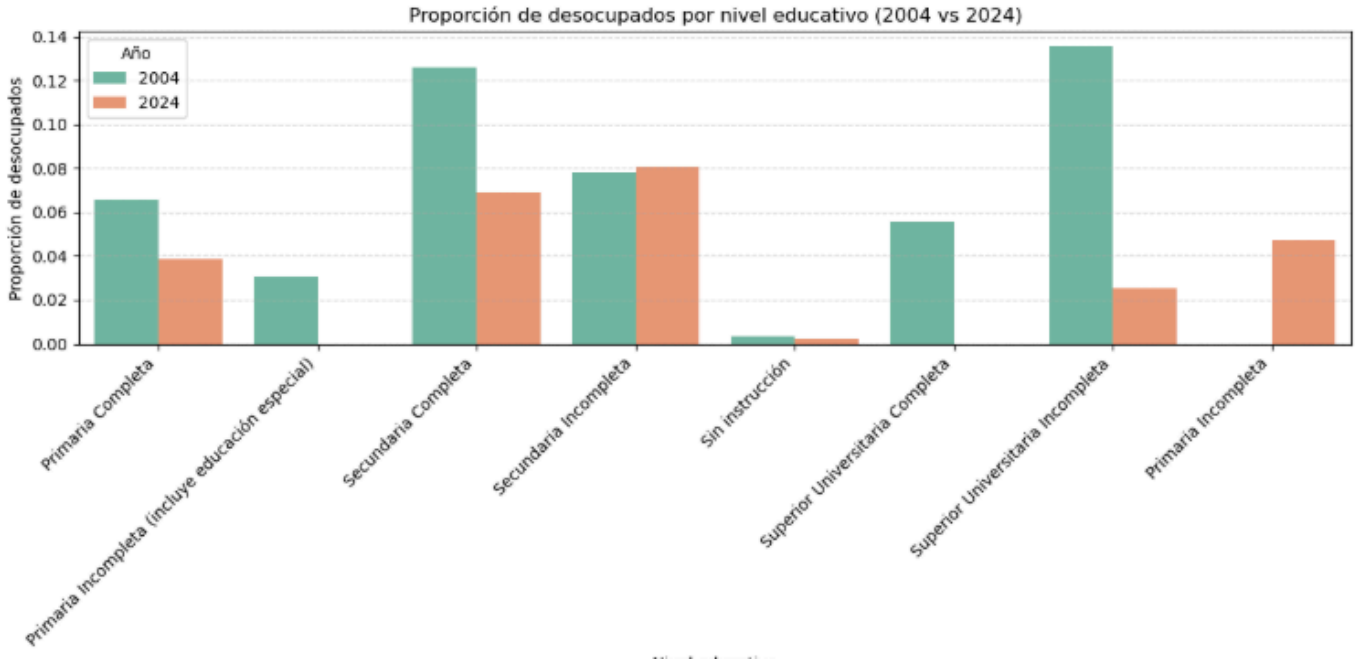
Figura 5: Composición de la PET (2004 vs 2024) [%]



La Figura 5 ilustra la composición de la Población en Edad para Trabajar (PET) para 2004 y 2024. La PET se mantuvo estable en ~65-66%, indicando que no hubo cambios significativos en la estructura etaria. Sin embargo, al compararla con la disminución de la PEA, sugiere que menos personas en edad de trabajar participan en el mercado laboral, lo que apunta a posibles desafíos en la inserción laboral.

5. Desocupación por nivel educativo

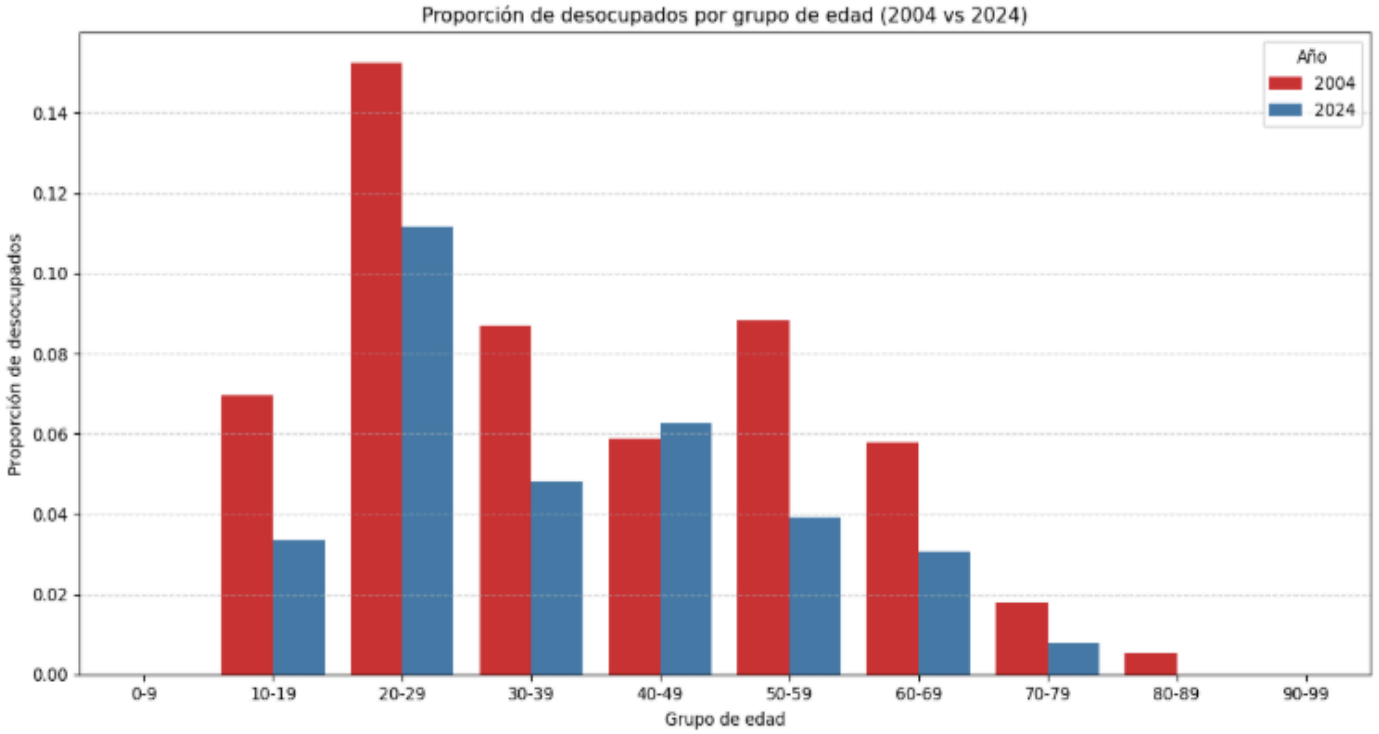
Figura 6: Proporción de desocupados por nivel educativo (2004 vs 2024)



La Figura 6 muestra la proporción de personas desocupadas por nivel educativo para 2004 y 2024. En 2004, las personas con educación secundaria completa y universitaria incompleta tuvieron tasas de desempleo más altas (12.6% y 13.5%, respectivamente). Para 2024, estas tasas disminuyeron a 6.9% y 2.5%, respectivamente, lo que podría reflejar una mayor demanda de trabajadores calificados o mejoras en el sistema educativo.

6. Desocupación por grupo de edad

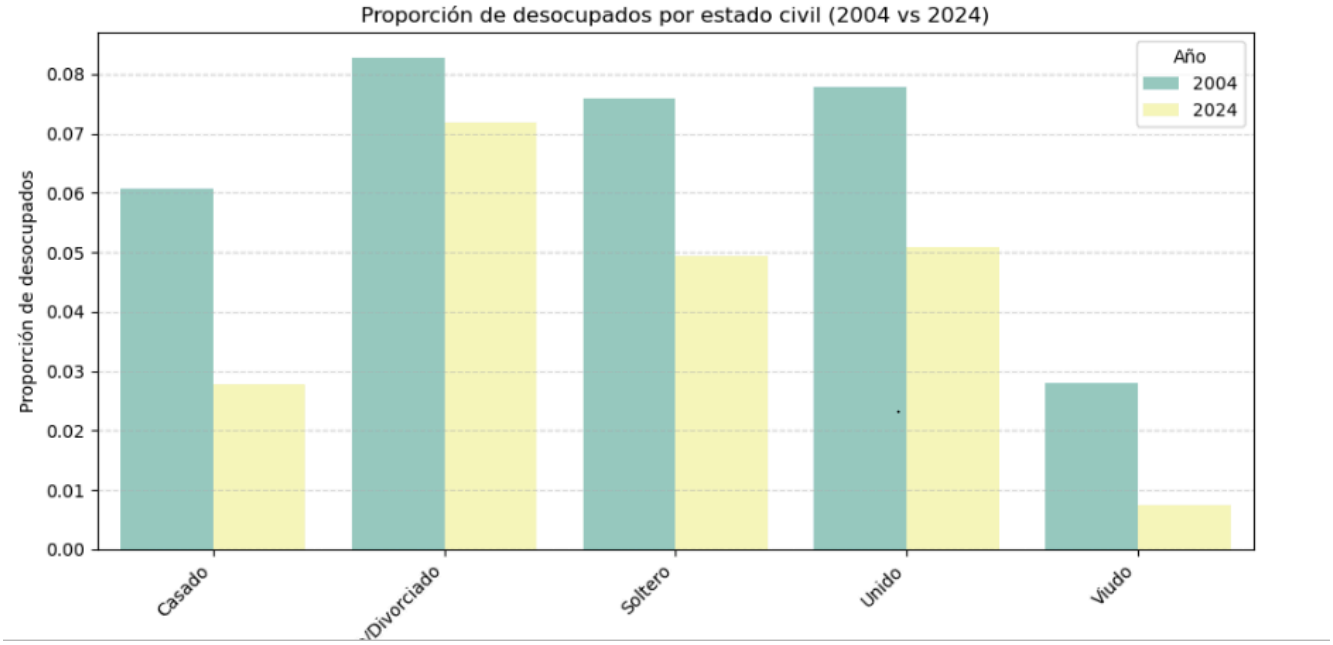
Figura 7: Proporción de desocupados por grupo de edad (2004 vs 2024)



La Figura 7 presenta la proporción de personas desocupadas por grupo de edad para 2004 y 2024. Los grupos de 20-29 y 30-39 años tuvieron las tasas de desempleo más altas en ambos años, pero estas tasas disminuyeron en 2024 (por ejemplo, de 15% a 10% para 20-29 años). Esto sugiere una mejor absorción de trabajadores jóvenes en el mercado laboral con el tiempo.

7. Desocupación por estado civil

Figura 8: Proporción de desocupados por estado civil (2004 vs 2024)



La Figura 8 muestra la proporción de personas desocupadas por estado civil para 2004 y 2024. Los solteros tuvieron las tasas de desempleo más altas en ambos años (~10% en 2004 y 7% en 2024), probablemente porque son más jóvenes y están en etapas iniciales de sus carreras. La disminución general de las tasas de desempleo en 2024 indica condiciones mejoradas en el mercado laboral.

Conclusión

El análisis de la EPH para Gran Buenos Aires revela una disminución de la desocupación y de la PEA entre 2004 y 2024, mientras que la PET se mantuvo estable. Los ingresos per cápita crecieron significativamente, reflejando la inflación. Los grupos más vulnerables al desempleo son los jóvenes, los solteros y aquellos con educación secundaria completa o universitaria incompleta, aunque las tasas de desempleo disminuyeron en 2024. Estos hallazgos son relevantes para diseñar políticas públicas que fomenten la inserción laboral, especialmente para los jóvenes y los sectores con menor educación formal.

Referencias

- *INDEC (2025). Encuesta Permanente de Hogares: Metodología. Disponible en: https://www.indec.gob.ar/*
- *Schwabish, J. A. (2014). An economist's guide to visualizing data. Journal of Economic Perspectives, 28(1), 209-234.*