

Big Data y Machine Learning (UBA) - 2025

Trabajo Práctico 2: UN PRIMER ENCUENTRO CON LA EPH

Materia: Big Data y Machine Learning

Profesora: Romero Noelia

Grupo:18

Integrantes:

911748	44486233	MAMONE FACUNDO
915896	46278154	ROMERO MARTIN
916370	96256981	VARGAS VILLAROON RICARDO

Parte I: Familiarización con la base EPH y limpieza

1. Según el sitio online de la INDEC, basando nuestra información a partir de la Encuesta Permanente de Hogares -EPH- este grupo forma parte de la población económicamente activa. Las personas desocupadas son aquellas que no trabajaron ni si quiera una hora durante la semana de referencia (no poseen empleo en la actualidad), que se encuentren disponibles para trabajar, esto quiere decir que pasando 7 días de la semana de referencia puedan comenzar a ejercer dicha actividad y por último todas aquellas que hayan buscado trabajo de manera constante en los últimos 30 días.
2. Tal como se pedía en el enunciado se cargaron las bases del primer trimestre del 2004 y 2024 en formato .xls (usu_individual_t124.xls) y .dta (usu_individual_t104.dta).

Las 15 variables elegidas fueron: 'P47T' (ingreso total por persona), 'IPCF' (índice de precios de consumidor), 'CH04' (genero del individuo), 'CH06' (edad del individuo), 'CH08' (estado civil del individuo), 'P21' (nivel educativo del individuo), 'CAT_INAC' (catálogo de inactividad), 'PP04D_COD' (código de ubicación geográfica), 'PP04A' (código de aglomerado), 'TRIMESTRE', 'REGION', 'AGLOMERADO', 'IDECCFR' (índice de desarrollo socioeconómico y calidad de vida), 'ANO4' (año), 'ESTADO'.

La región elegida fue la Patagonia, en 2004 aparece como "patagónica" y en el de 2024 es =43.

B. Podemos observar que obtuvimos 14.410 observaciones para el 2024 mientras que 3.264 para el año 2004.

-Tabla de valores faltantes por año

Valores faltantes por año:

ANO4	2004	2024
P47T	0	15
IPCF	0	0
CH04	0	0
CH06	51	0
CH08	0	0
P21	0	0
CAT_INAC	0	0
PP04D_COD	0	8108
PP04A	0	8108
TRIMESTRE	0	0
REGION	0	0
AGLOMERADO	0	0
IDECCFR	0	0
ANO4	0	0
ESTADO	0	0

Cantidad de valores NaN en ingresos después de limpiar:

P47T: 2040

IPCF: 0

En esta tabla podemos visualizar que para el año 2004 obtenemos 51 valores faltantes para la variable CH06, la cual indica la edad. Pasando al año 2024 observamos una módica suma de 15 variables faltantes para P47T, el cual es el ingreso total, esto se puede entender ya que hay una posibilidad de que no se sientan cómodos ni en confianza de reportar dicha información. Para las variables PP04D_COD, código de ubí geográfica y PP04A, código de aglomerado, hay un alto número de valores faltantes, esto se puede dar por la misma razón que lo anterior, muchos encuestados no proporcionan información la cual no les parece fácil anunciar, por diversos temas como la privacidad, entre otros.

C. Debajo de todo podemos observar como se identificaron ingresos negativos en P47T (ingreso total) e IPCF (índices de precios de consumidor) esta última no ha arrojado ningún valor negativo. De la primera variable mencionada podemos observar como 2.040 valores se han reemplazado por NaN para un correcto análisis.

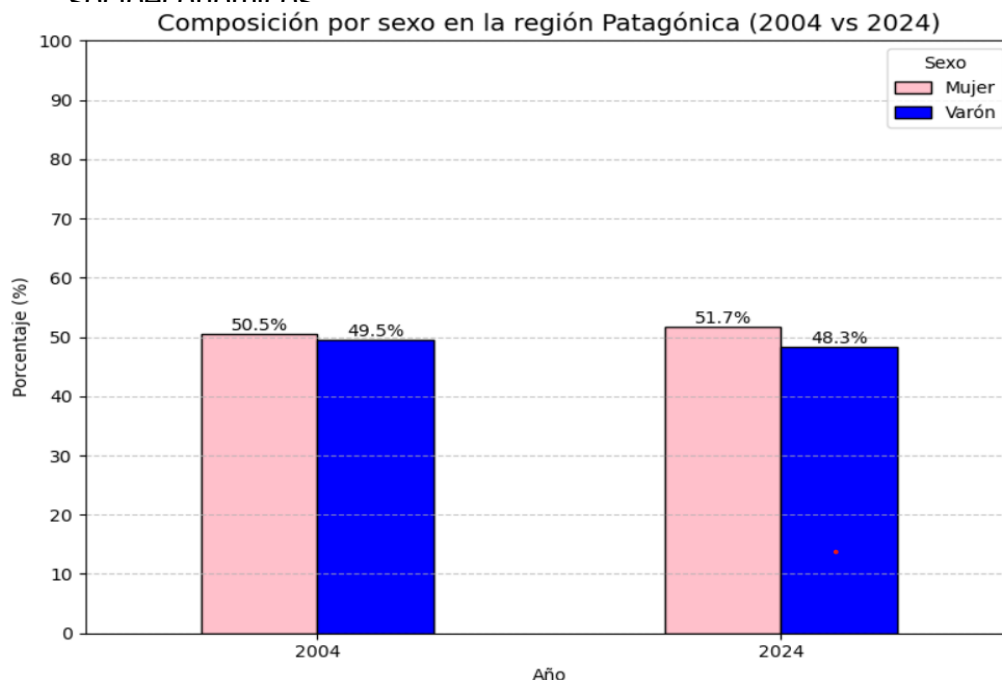
Parte II: Primer Análisis Exploratorio

3. Análisis de la composición por sexo en la región patagónica (2004 vs 2024)

-En el gráfico podemos observar como este ilustra de acuerdo con los datos trabajados que la proporción entre hombres y mujeres para el año 2004 era prácticamente igual, muy equilibrada, con solo un porcentaje del 0,5% de diferencia. El género femenino generaba el 50,5% de la región mientras que el masculino el 49,5%.

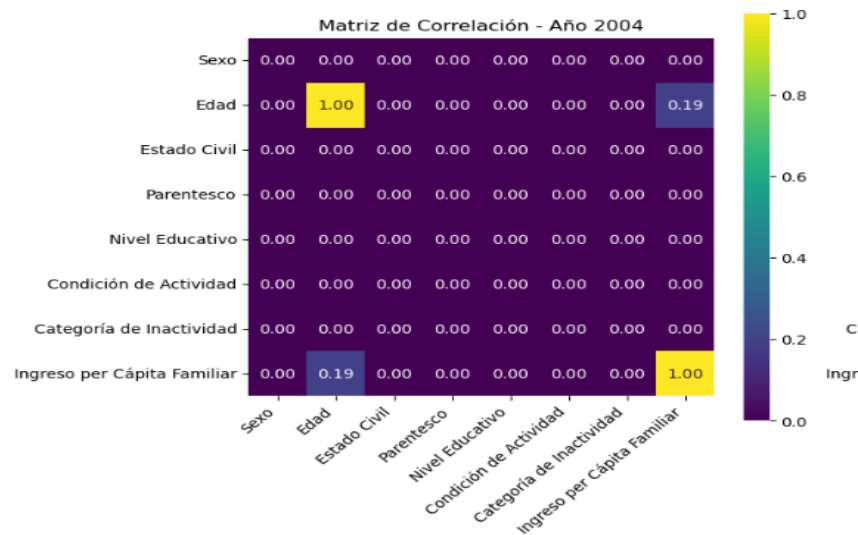
-Este porcentaje cambia, no en gran medida, para el año 2024. Las mujeres siguen siendo mayoría, esta vez la región está compuesta por un 51,7% por ellas vs un 48,3% de los hombres.

-Este cambio porcentual de la población de la Patagonia se puede dar por diversos factores, como demográficos o incluso la esperanza de vida del género femenino en la región así también como distintos factores socioeconómicos.



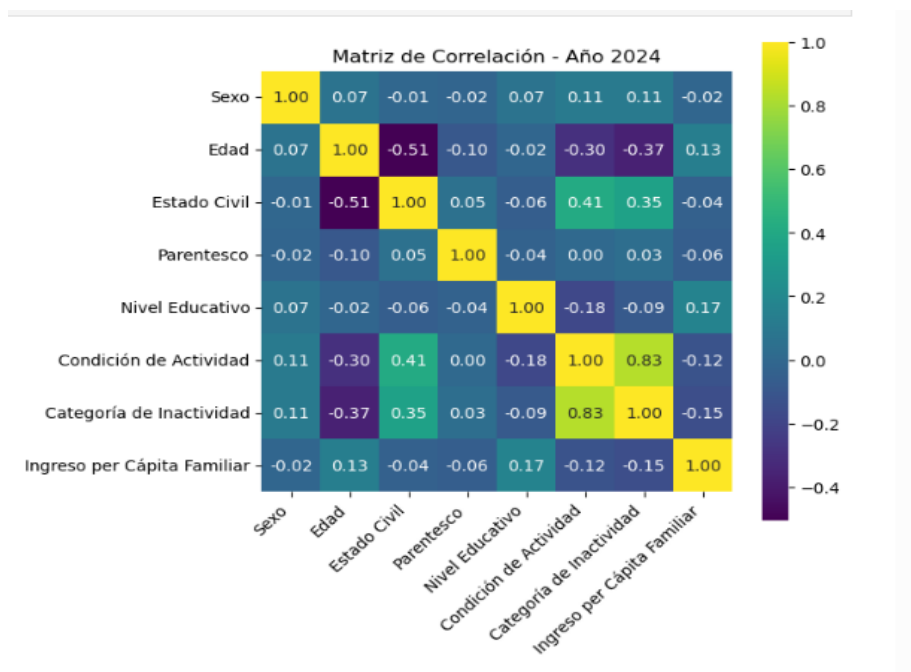
4. Matriz de correlación año 2004

-En esta matriz visualizamos correlaciones cercanas a 0 entre si, esto indica que para el año trabajado no hay relación estadística aparente entre las variables, según los datos obtenidos. El único vínculo relevante que se da es entre la edad y el ingreso per cápita familiar, aunque es débil.



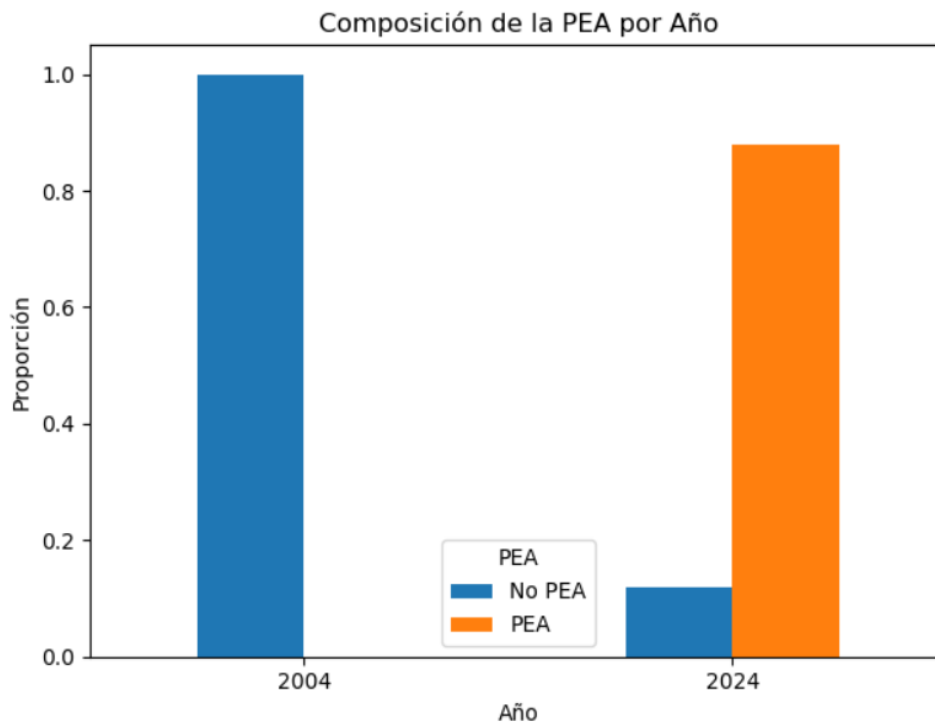
Matriz correlación año 2024

- En esta matriz podemos observar una alta correlación positiva entre la condición de actividad y la categoría de inactividad superando el 0.80, ya que al ser inactivo se sigue formando parte de una categoría específica de inactividad. Observamos una correlación un poco mas moderada entre estado civil y condición de actividad, así como edad y estado civil superando una el 0.40 y la ultima el 0.50. Si nos enfocamos en las negativas observamos la edad con condición de actividad, así como categoría de inactividad ambas en el rango de 3.0. Por ultimo las más débiles o casi nulas como sexo, parentesco, así como nivel educativo no superan el 0.20.



Parte 3

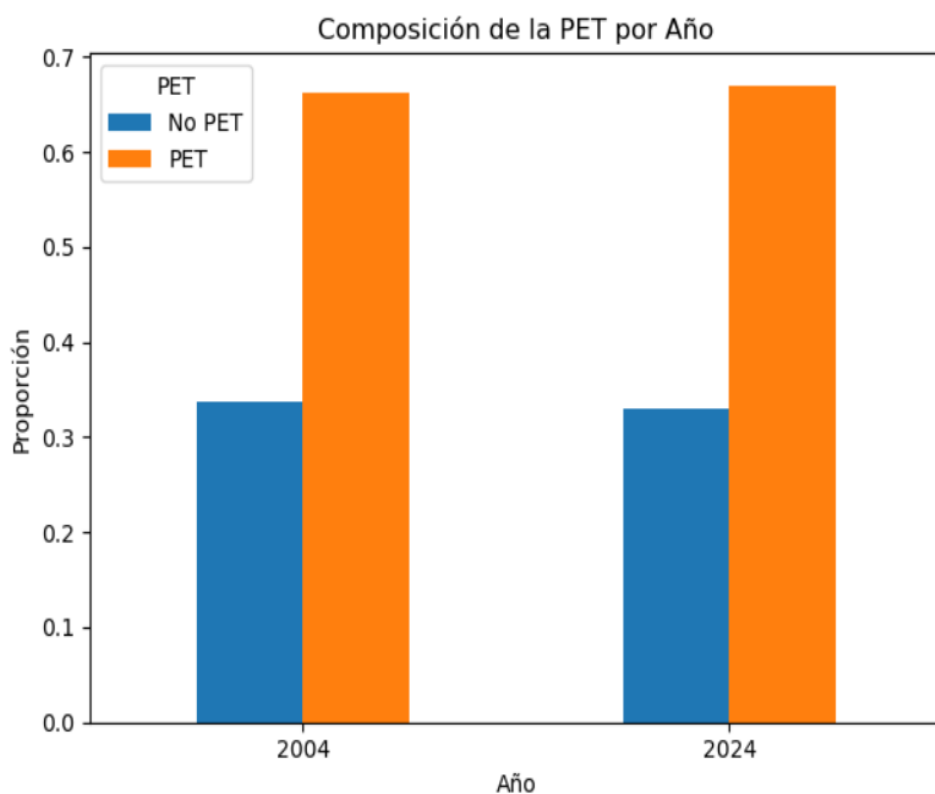
Composición de la PEA por Año



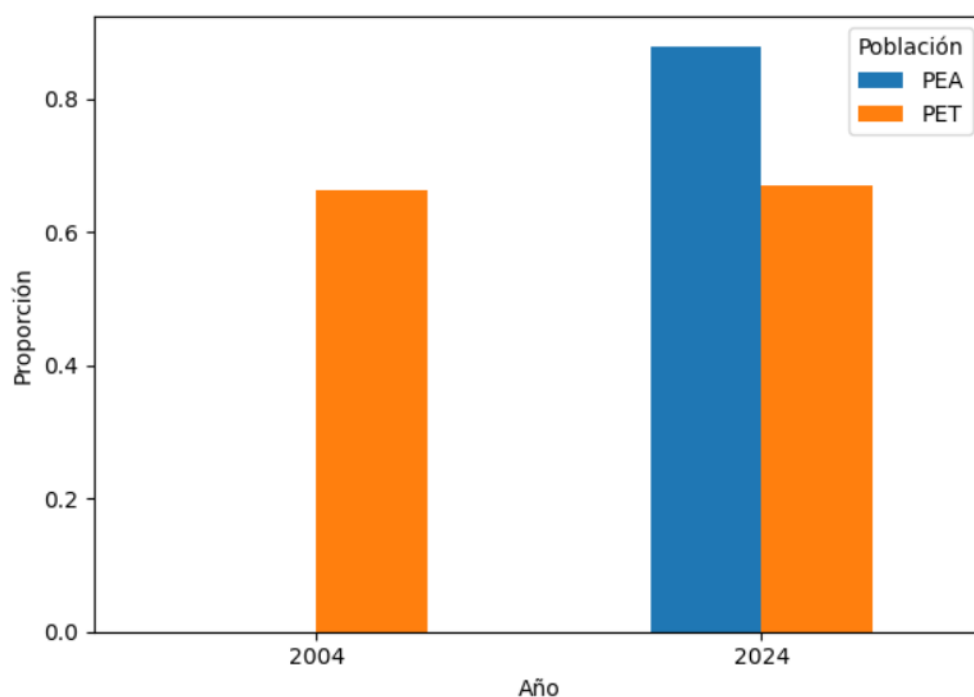
En el gráfico de la Población Económicamente Activa (PEA) observamos un cambio notable entre 2004 y 2024. En 2004, la proporción de personas dentro de la PEA representaba casi la totalidad de la población en edad de trabajar, mientras que en 2024 esa proporción disminuye considerablemente. Esto puede reflejar un aumento de la población inactiva (personas que no buscan empleo ni están disponibles para trabajar), posiblemente por factores como el envejecimiento poblacional, un aumento de estudiantes o cambios en el rol del trabajo dentro del hogar.

Composición de la PET por Año

En cuanto a la Población en Edad de Trabajar (PET), su proporción dentro de la población total se mantiene bastante estable entre ambos años. Esto indica que, demográficamente, la estructura de edades no ha cambiado significativamente. Sin embargo, al observar que la proporción de PEA dentro de la PET disminuye (como vimos en el gráfico anterior), se puede inferir que hay más personas en edad de trabajar que no forman parte del mercado laboral.



Comparación de PEA y PET por Año

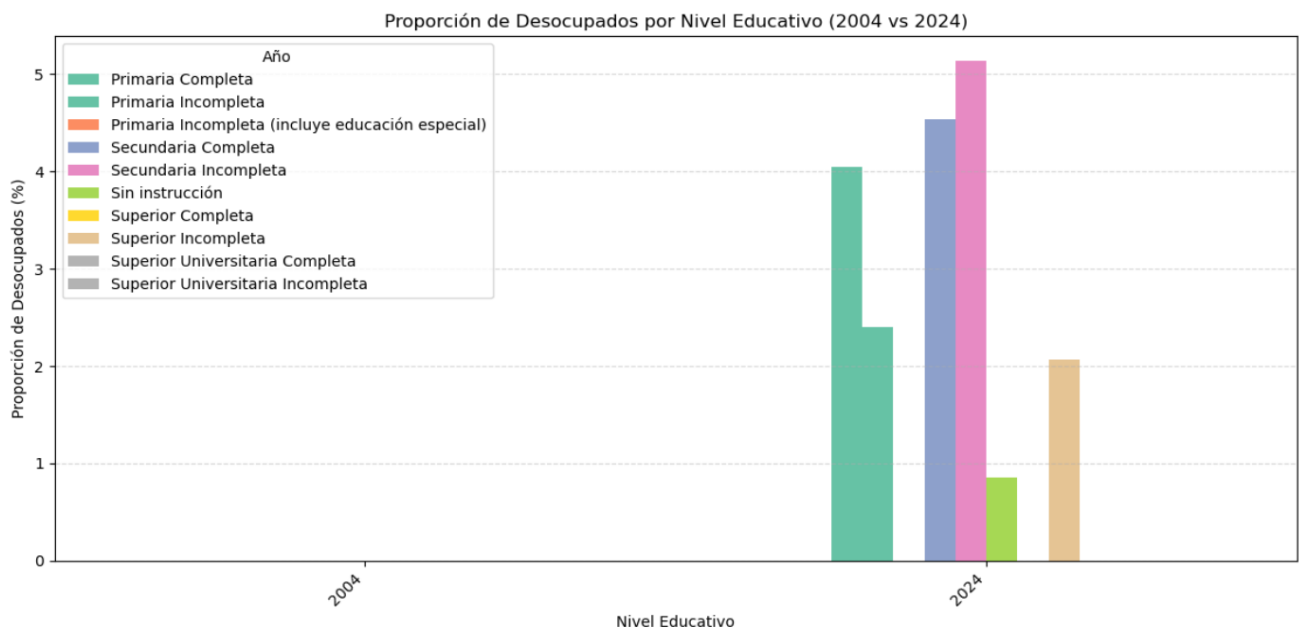


Este gráfico permite comparar directamente la proporción de PEA y PET para ambos años. En 2004, ambas proporciones eran muy similares, lo que indica que casi toda la PET estaba económicamente activa. En 2024, en cambio, la diferencia entre PET y PEA se amplía, lo que reafirma la hipótesis de un

aumento en la población económicamente inactiva. Este cambio puede estar relacionado con factores como:

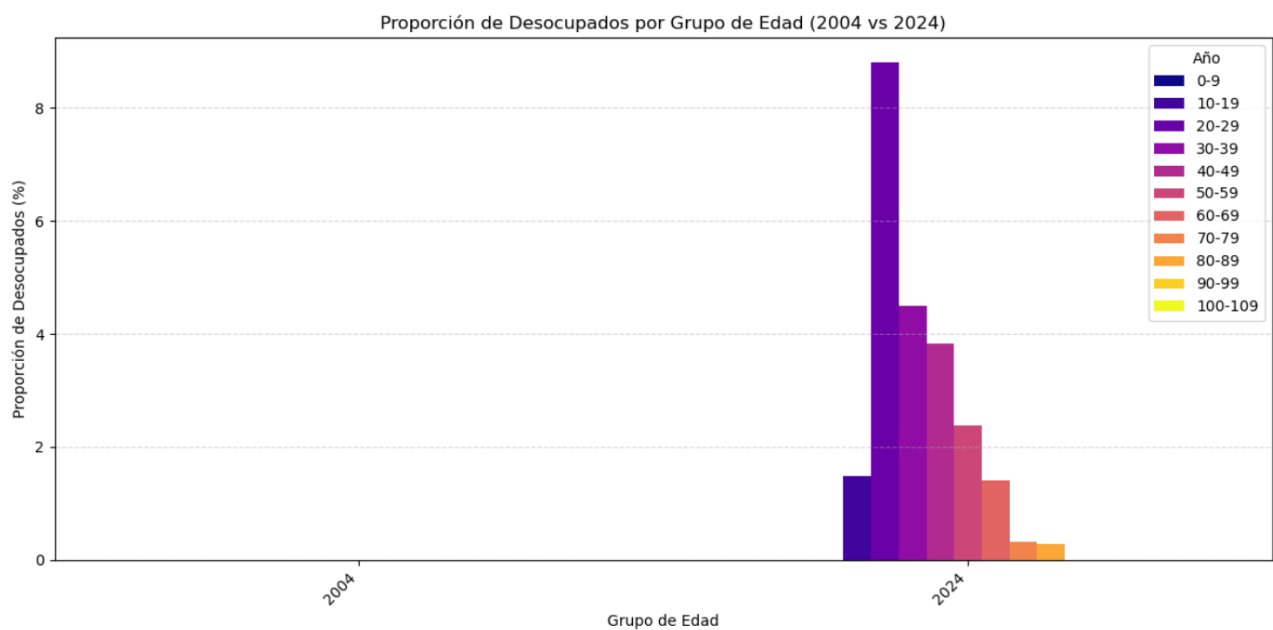
- Mayor cantidad de estudiantes o personas en capacitación.
- Envejecimiento de la población y aumento de personas jubiladas.
- Aumento de tareas no remuneradas (como cuidado del hogar).
- Cambios socioeconómicos que desincentivan la participación laboral.

Desocupación por Nivel Educativo



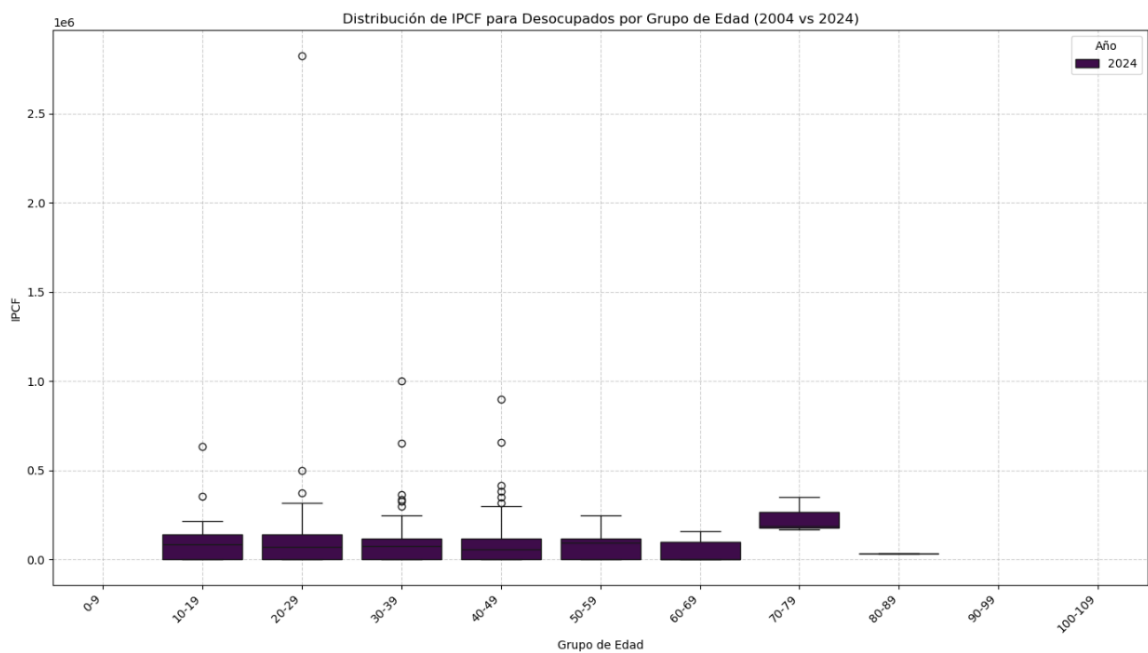
Este gráfico muestra cómo se distribuye la desocupación según el nivel educativo. En 2024, la proporción más alta de personas desocupadas se encuentra entre quienes tienen educación secundaria incompleta y completa. Esto puede indicar una mayor dificultad de inserción laboral para quienes no acceden a niveles educativos superiores. También se observa un aumento en la desocupación entre personas con nivel superior incompleto. En cambio, los niveles de desocupación para quienes no tienen instrucción o solo educación primaria son menores, posiblemente por su menor participación en la PEA directamente.

Distribución del IPCF entre Desocupados por Edad



El boxplot de IPCF (Ingreso Per Cápita Familiar) para personas desocupadas muestra que, en 2024, los grupos de edad más jóvenes (20-29 y 30-39 años) concentran mayores desigualdades y casos extremos de ingreso familiar. Este fenómeno puede estar relacionado con mayor precariedad o dependencia económica en etapas tempranas de la vida laboral. También podría reflejar que muchos jóvenes desocupados viven en hogares con ingresos familiares altos, lo que modera su urgencia de insertarse en el mercado laboral.

Proporción de Desocupados por Grupo de Edad



En este gráfico se destaca que, para el año 2024, los grupos etarios con mayor proporción de desocupación son los de 20-29 y 30-39 años, lo que puede evidenciar una mayor vulnerabilidad laboral en la población joven-adulta. Las tasas más bajas se observan a partir de los 60 años, en parte por el retiro del mercado laboral. Comparado con 2004, estas diferencias sugieren un aumento en la desocupación juvenil.