Big Data y Machine Learning (UBA) - 2025

Trabajo Práctico 2: Un primer encuentro con la EPH

Grupo: 10
Integrantes: Bautista Benetti, Luca D'adderio y Tongkun Weng

Según la información disponible en la página del INDEC, las personas desocupadas son aquellas que no tienen trabajo, están buscando activa mente trabajo, y están disponibles para comenzar a trabajar.

	CODUSU	ANO4	TRIMESTRE	NRO_HOGAR	COMPONENTE	H15	REGION	MAS_500	A
0	TQRMNOPUYHMOKPCDEGOIH00800329	2024.0	1	1.0	1	1	40	N	
1	TQRMNOPUYHMOKPCDEGOIH00800329	2024.0	1	1.0	2	1	40	N	
2	TQRMNOPUYHMOKPCDEGOIH00800329	2024.0	1	1.0	3	1	40	N	
3	TQRMNOPUYHMOKPCDEGOIH00800329	2024.0	1	1.0	4	1	40	N	
4	TQRMNOTSQHJKLMCDEGOIH00795856	2024.0	1	1.0	1	1	40	N	
...	...	...	...	...	...	...	...	...	
19087	288280	2004.0	1er. Trimestre	1.0	3.0	0.0	40	S	
19088	288280	2004.0	1er. Trimestre	1.0	4.0	0.0	40	S	
19089	288280	2004.0	1er. Trimestre	1.0	5.0	0.0	40	S	
19090	288337	2004.0	1er. Trimestre	1.0	1.0	Sí	40	S	
19091	288337	2004.0	1er. Trimestre	1.0	2.0	Sí	40	S	

19092 rows x 181 columns

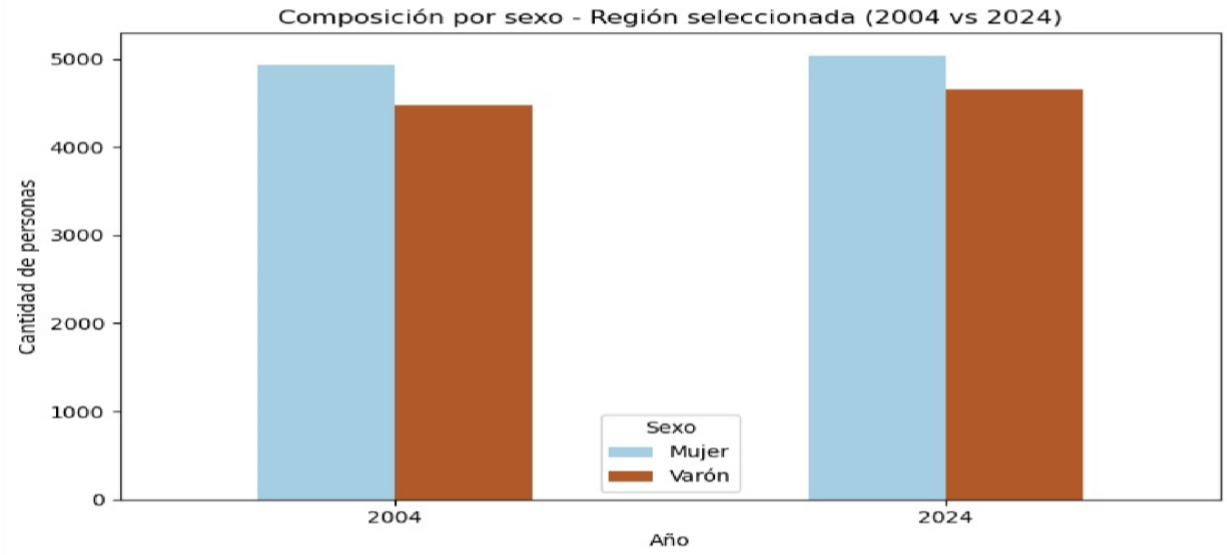
Valores faltantes en df_2024:

ANO4	0
REGION	0
PONDERA	0
CH03	0
CH04	0
CH08	0
CH06	0
CH07	0
NIVEL_ED	0
ESTADO	0
CAT_OCUP	0
CAT_INAC	0
PP04G	5436
PP07C	5436
PP08D1	5436

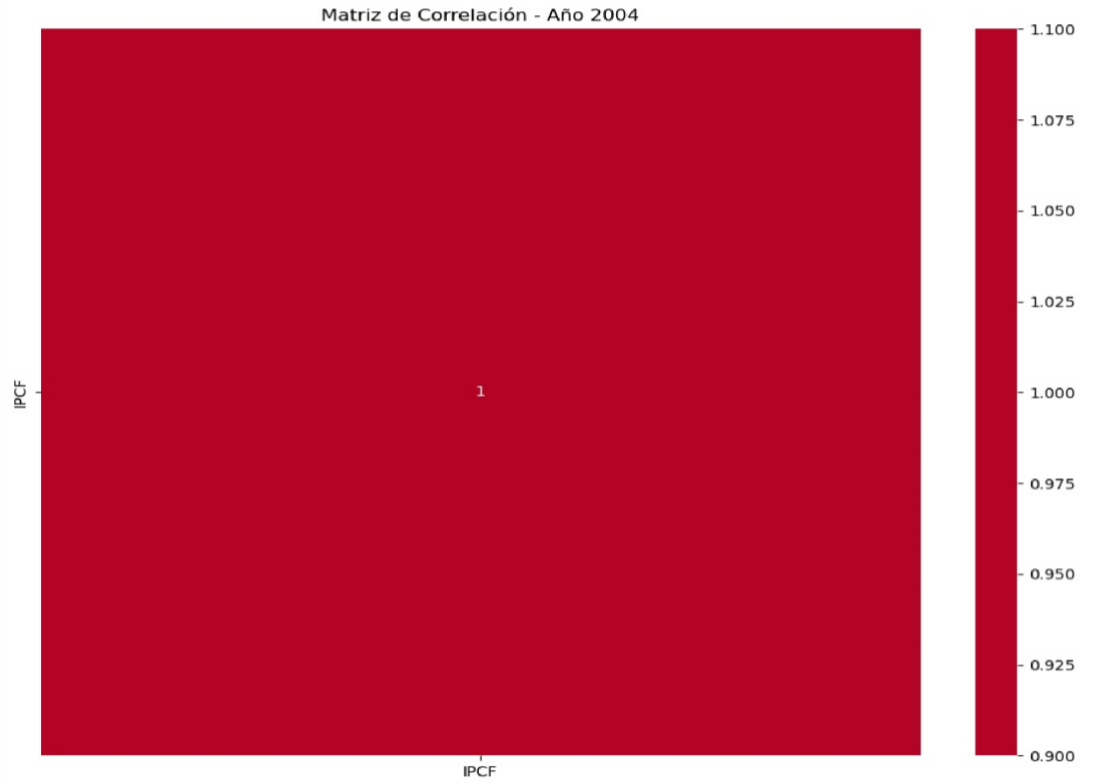
dtype: int64

Valores faltantes en df_2004:
ANO4 0
REGION 0
PONDERA 0
CH03 0
CH04 0
CH08 0
CH06 0
CH07 0
NIVEL_ED 0
ESTADO 0
CAT_OCUP 0
CAT_INAC 0
PP04G 0
PP07C 0
PP08D1 0
dtype: int64
Los variables PP04G, PP07C y PP08D1son los que tienen más valores faltantes, y son todos del año 2024.

Filas eliminadas por datos sin sentido (2024): 4263
Filas eliminadas por datos sin sentido (2004): 9393
Este proceso de limpieza consiste en eliminar todos los valores faltante y luego filtrar todos los datos que tengan valores razonables.



La cantidad de mujeres en el 2004 son superiores a las de los hombre y en el 2024 también.


Matriz de Correlación - Año 2004
IPCF
1
IPCF
1.100
1.075
1.050
1.025
1.000
0.975
0.950
0.925
0.900

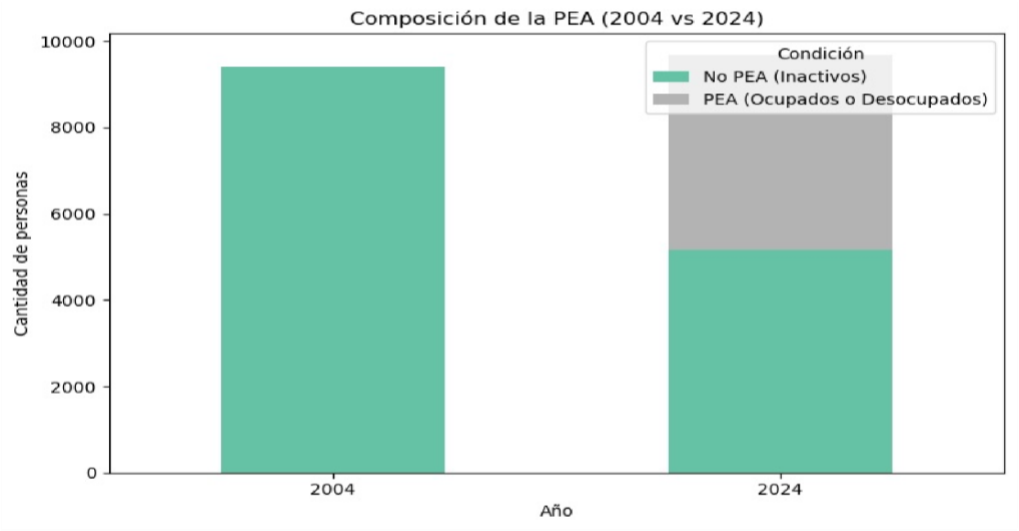

Matriz de Correlación - Año 2024
CH04
CH06
CH07
CH08
NIVEL_ED
ESTADO
CAT_INAC
IPCF
1 0.063 -0.0066 -0.024 0.073 0.096 0.11 -0.0094
0.063 1 -0.48 -0.18 -0.023 -0.34 -0.33 0.19
-0.0066 -0.48 1 0.081 -0.039 0.37 0.3 -0.052
-0.024 -0.18 0.081 1 -0.054 -0.02 0.035 -0.18
0.073 -0.023 -0.039 -0.054 1 -0.16 -0.066 0.16
0.096 -0.34 0.37 -0.02 -0.16 1 0.83 -0.13
0.11 -0.33 0.3 0.035 -0.066 0.83 1 -0.16
-0.0094 0.19 -0.052 -0.18 0.16 -0.13 -0.16 1
CH04
CH06
CH07
CH08
NIVEL_ED
ESTADO
CAT_INAC
IPCF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

La matriz de correlación del año 2004 muestra únicamente la variable P CF, por lo que el resultado refleja una autocorrelación perfecta (valor = 1). Esto indica que no hay otras variables en el conjunto de datos para ese año o que el filtrado previo dejó una sola variable numérica disponi ble. Para que la matriz de correlación aporte información útil sobre relaci ones entre variables, sería necesario contar con múltiples variables num éricas en el dataset.
En la matriz del 2024, muestra la correlación lineal (coeficiente de Pears on) entre distintas variables del dataset. Los valores van desde -1 (corre lación negativa perfecta) hasta 1 (correlación positiva perfecta). Los color es ayudan a visualizar: tonos rojizos indican correlaciones positivas, y az ulados, negativas.

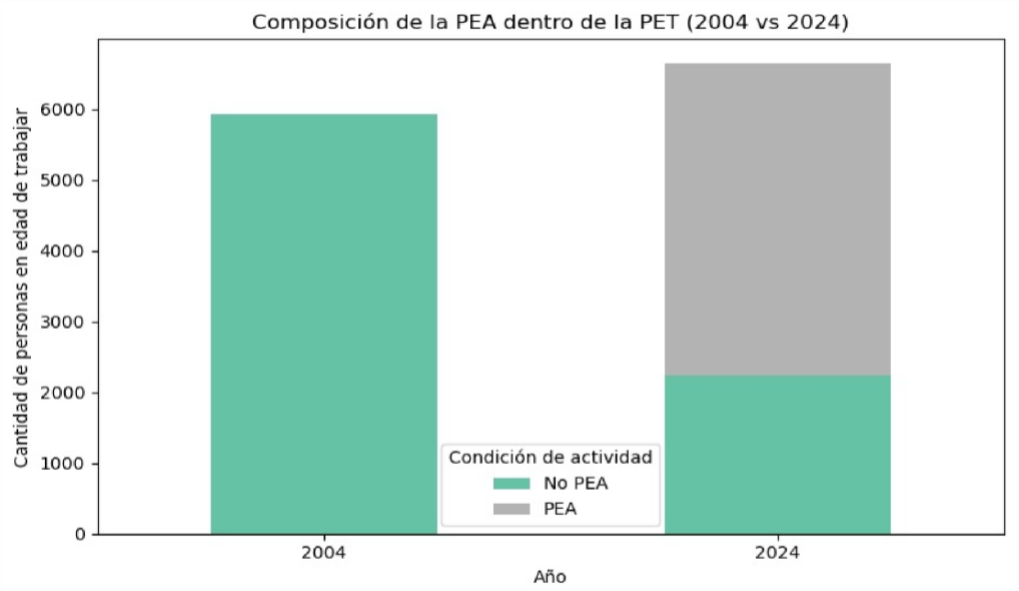
Cantidad de personas por estado:
ESTADO
Ocupado 4263
Inactivo 3999
Desocupado 256
Name: count, dtype: int64

Media de IPCF por estado:
ESTADO
Desocupado 105462.64
Inactivo 133618.53
Ocupado 168305.85
Name: IPCF, dtype: float64
Hay 256 de desocupados en la muestra, hay 3999 inactivos en la mues tra. También la media de ingreso per cápita per cápita familiar según es tado es por desocupado es 105462.64, de inactivo de 133618.53 y ocup ado 168305.85.

Cantidad de personas que no respondieron su condición de actividad: 9

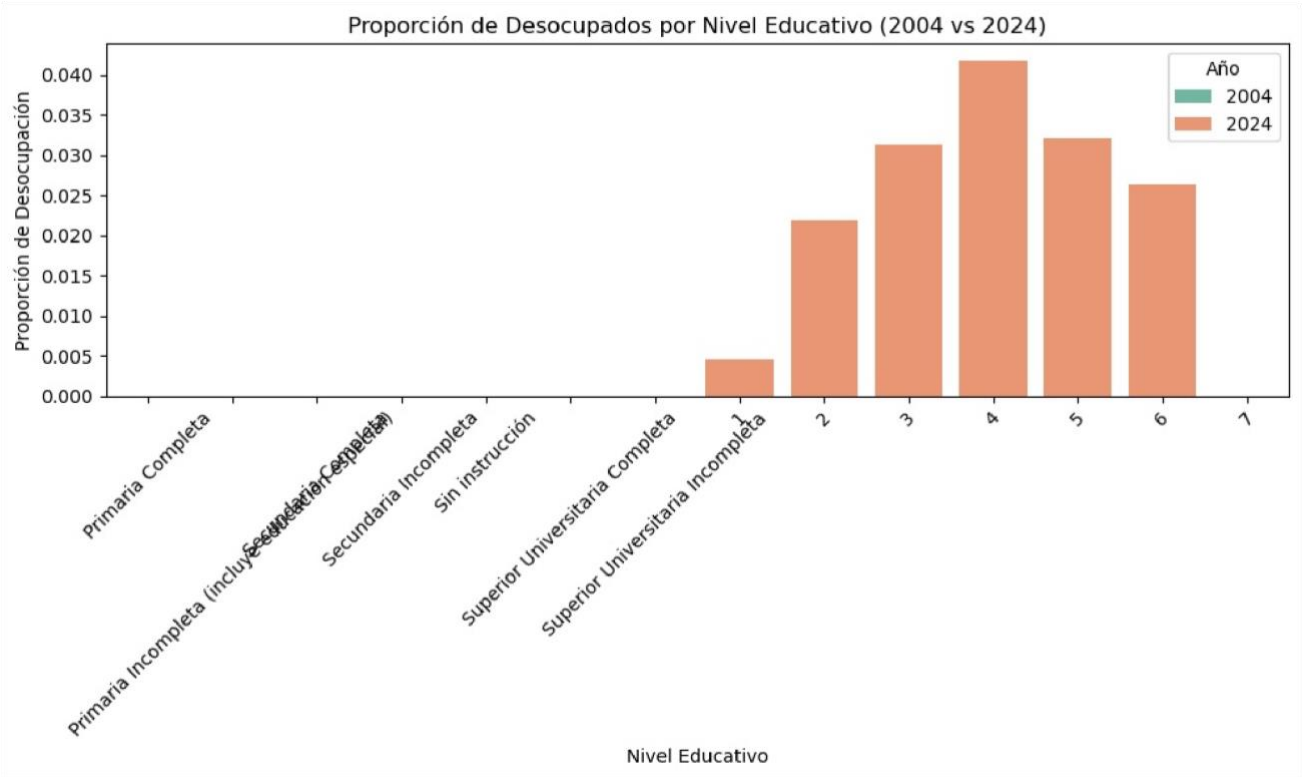


En el 2004 la cantidad de inactivos es total. Pero en el 2024 la PEA m ejor dicho ocupados o desocupados se divide, donde hasta 5000 person as son inactivos y 5000 son PEA.

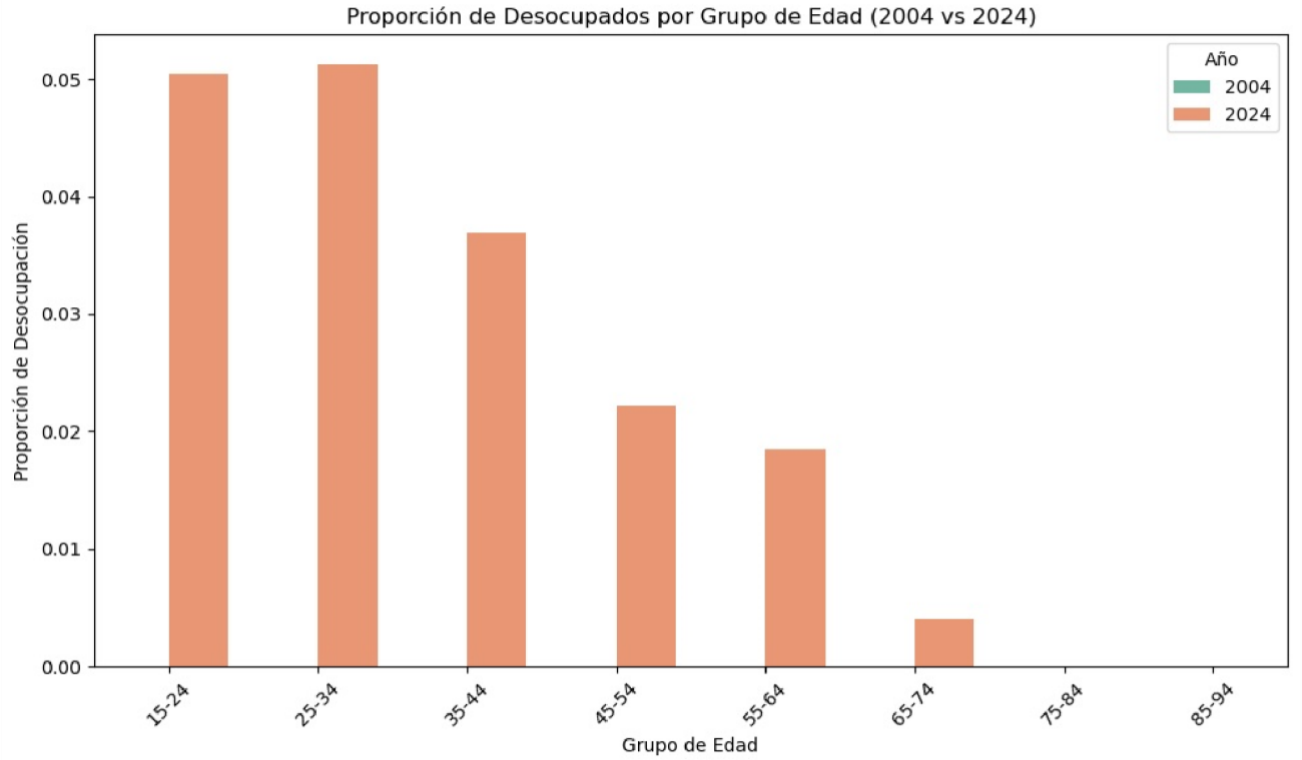


La cantidad de personas en edad de trabajar en el 2004 es total pero e n el 2024 es distinto debido a que hay 2000 que están en edad de tra bajar y 4000 que están ocupados o inactivos.

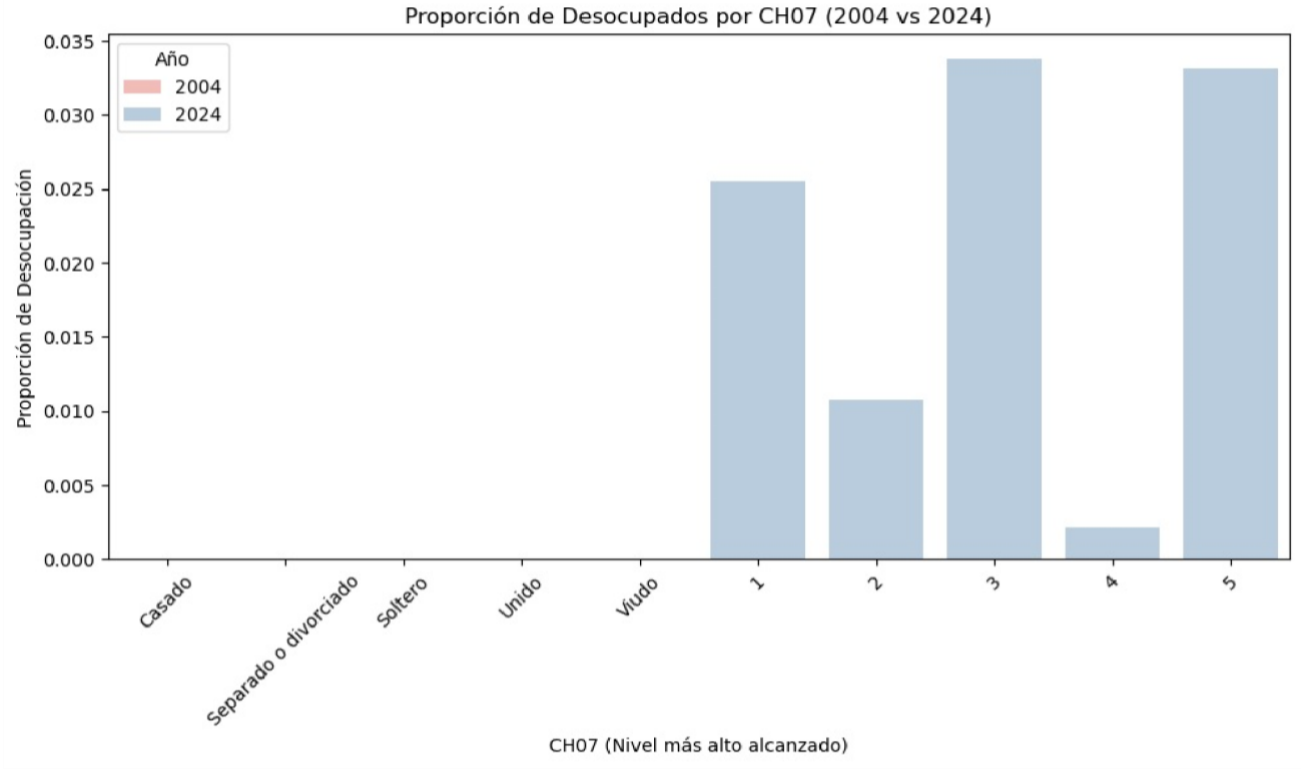
Cantidad de personas desocupadas por año:
AÑO
2004 0
2024 256
Name: desocupado, dtype: int64
En el 2004 la cantidad de personas desocupadas es 0 por ende no hay pero en el 2024 se ve un gran aumento donde pasan a ser 256.



Sí, hubieron cambios de desocupados por nivel educativo.



Sí, hubo cambios de desocupados por edad.



Sí, hubo cambios de desocupados por edad.