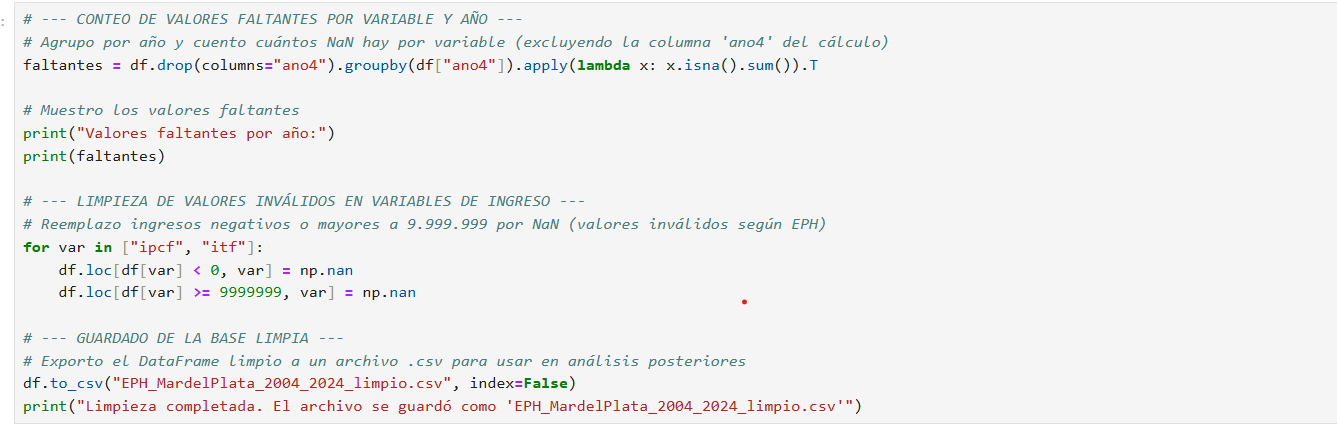
Big Data y Machine Learning – Trabajo Práctico 2

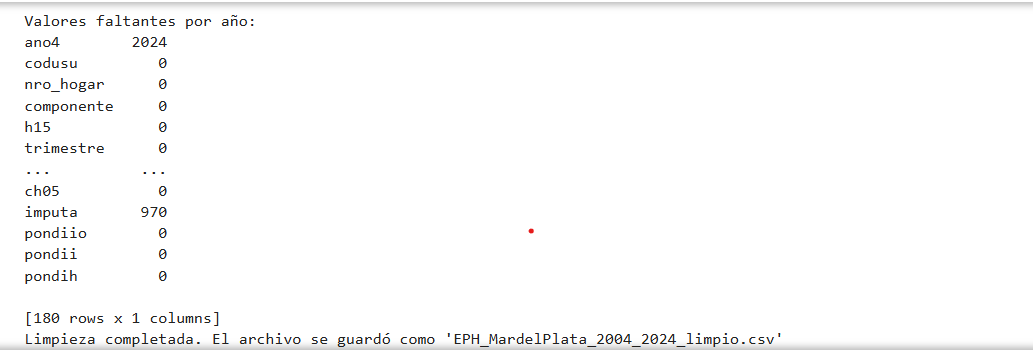
Parte I: Familiarizandonos con la base EPH y limpieza

En esta oportunidad, como grupo hemos elegido la región de Mar del Plata para analizar los datos correspondientes al primer trimestre de los años 2004 y 2024.

En este bloque de código se seleccionan 15 variables de interés provenientes de la Encuesta Permanente de Hogares (EPH), relacionadas con características sociodemográficas y económicas como edad, sexo, nivel educativo, condición de actividad, tipo de ocupación, ingresos y región. Estas variables son convertidas a minúsculas para asegurar compatibilidad con los nombres de las columnas del DataFrame. Finalmente, se filtra el DataFrame original para conservar únicamente estas variables junto con la columna que indica el año de relevamiento ("ano4"), dejando preparado el conjunto de datos para su posterior análisis.

En este bloque de código se realiza un control para identificar columnas completamente vacías (es decir, que contienen solo valores nulos) en los datasets correspondientes al primer trimestre de los años 2004 y 2024. Se observa que el dataset de 2004 presenta una gran cantidad de columnas vacías, lo cual podría deberse a diferencias en la estructura del archivo o a errores en la carga de los datos. En contraste, el dataset de 2024 contiene muy pocas columnas vacías, lo que indica una estructura más completa y posiblemente más depurada. Este paso es fundamental para decidir qué variables conservar y cuáles descartar en el análisis posterior.

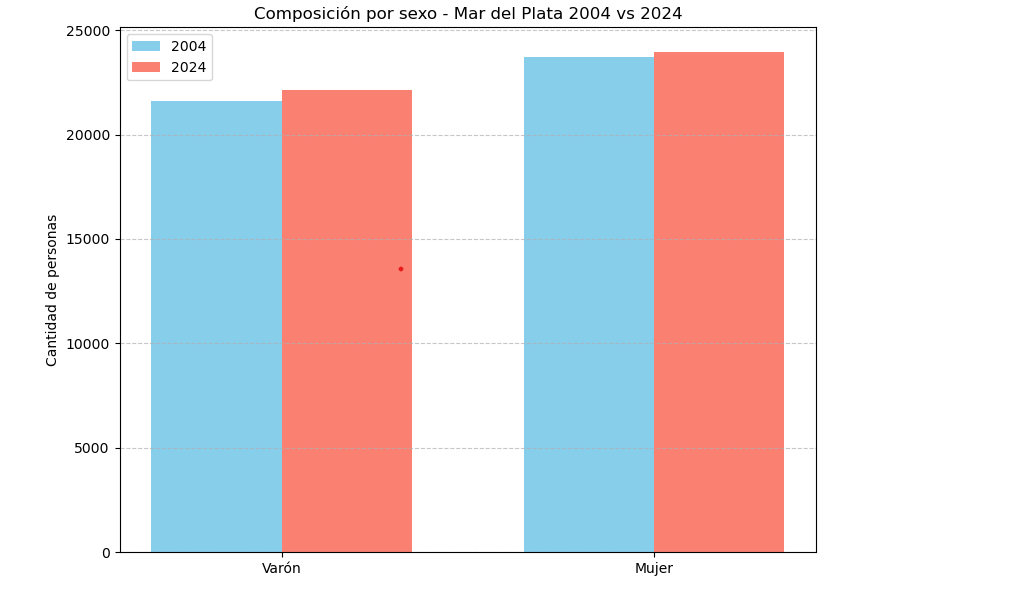




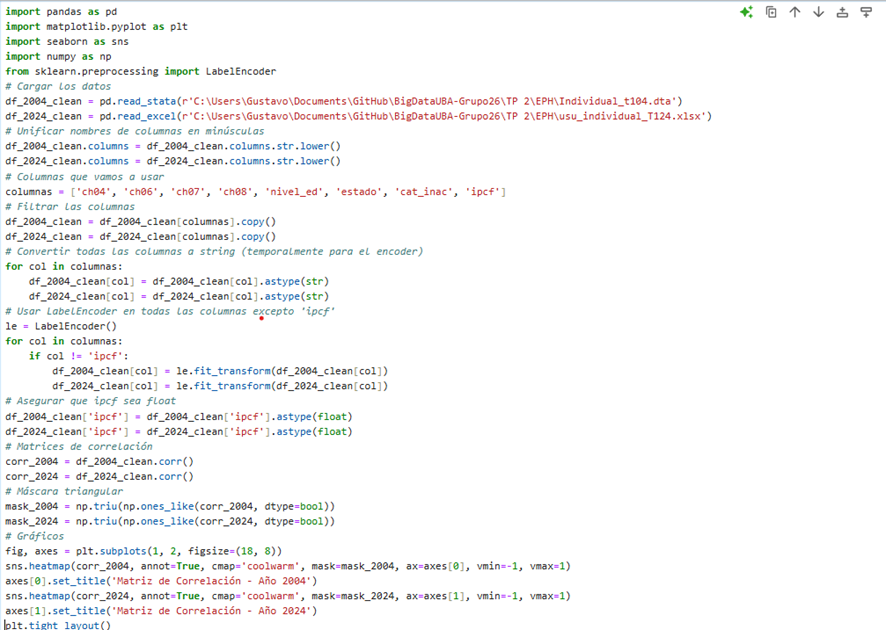
Parte II: Primer Análisis Exploratorio

3.Este código genera un gráfico de barras comparando la cantidad de personas según el sexo (varón y mujer) en la región de Mar del Plata para los años 2004 y 2024. Se utilizan dos barras para cada categoría (una por año), colocadas lado a lado para facilitar la comparación. El gráfico incluye etiquetas, colores diferenciados, leyenda y una grilla en el eje Y para mejorar la visualización.



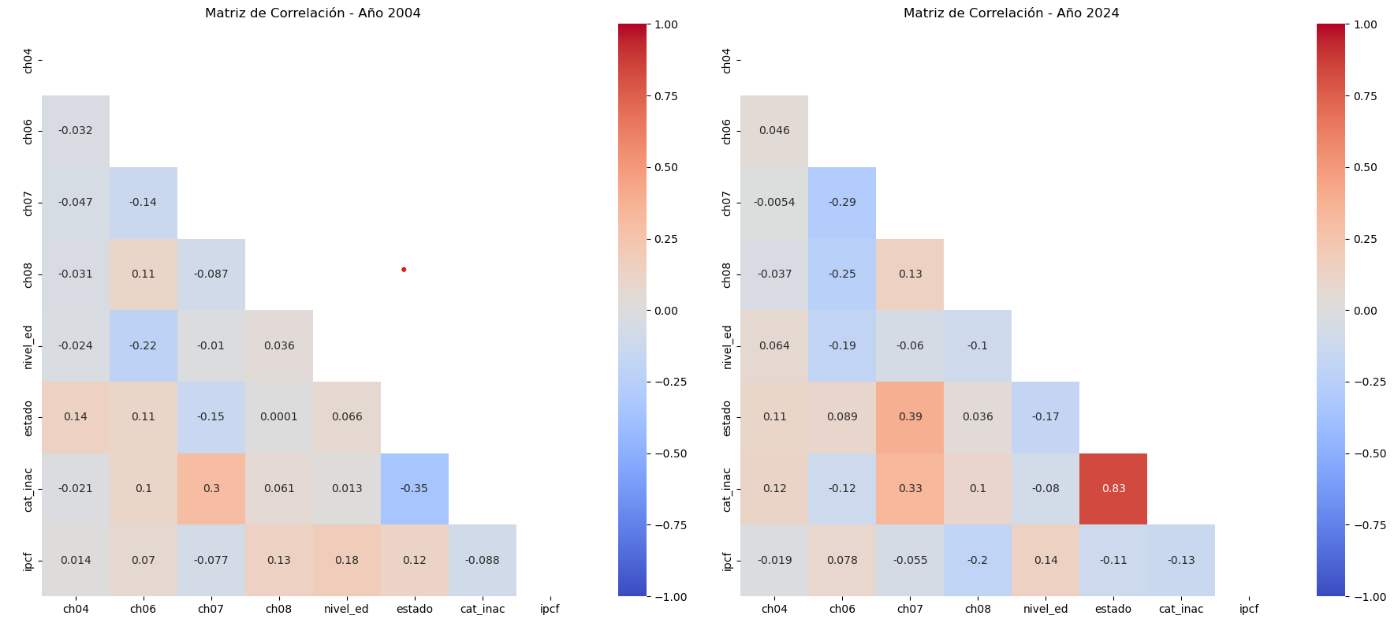
En el gráfico se observa que, tanto en 2004 como en 2024, la cantidad de mujeres en Mar del Plata es ligeramente superior a la de varones. En ambos años, la diferencia no es muy pronunciada, pero se mantiene constante. Además, puede notarse un leve crecimiento en la cantidad total de personas de ambos sexos entre 2004 y 2024, lo cual podría reflejar un aumento poblacional o una mayor cobertura de la encuesta. Esta estabilidad en la distribución por sexo indica que no hubo grandes cambios demográficos en términos de género durante el período analizado.

4.En este bloque de código se realiza un análisis comparativo de la correlación entre variables socioeconómicas en los años 2004 y 2024 usando datos de la EPH. Primero, se importan las librerías necesarias para manipular datos (pandas, numpy), visualizar resultados (matplotlib, seaborn) y codificar variables categóricas (LabelEncoder). Luego, se cargan los archivos correspondientes a cada año, y se estandarizan los nombres de columnas a minúsculas para evitar errores. A continuación, se seleccionan ocho variables de interés: el sexo (ch04), edad (ch06, ch07, ch08), nivel educativo (nivel\_ed), situación laboral (estado), categoría de inactividad (cat\_inac) y el ingreso per cápita familiar (ipcf). Todas las variables se convierten temporalmente a texto para aplicar el LabelEncoder, que transforma las categorías en números, excepto la variable de ingreso (ipcf), que se convierte a tipo numérico (float). Con los datos ya codificados, se calculan las matrices de correlación de Pearson para ambos años, lo cual permite observar las relaciones lineales entre variables. Para facilitar la visualización, se utiliza una máscara triangular superior (dado que la matriz es simétrica) y se generan dos mapas de calor (heatmaps) con Seaborn, uno para 2004 y otro para 2024. En estos gráficos se muestran las correlaciones con una escala de color que va del rojo (correlación negativa) al azul (positiva), junto con los valores numéricos. Este análisis permite identificar cómo han cambiado las relaciones entre características como educación, empleo e ingresos a lo largo del tiempo en Mar del Plata.

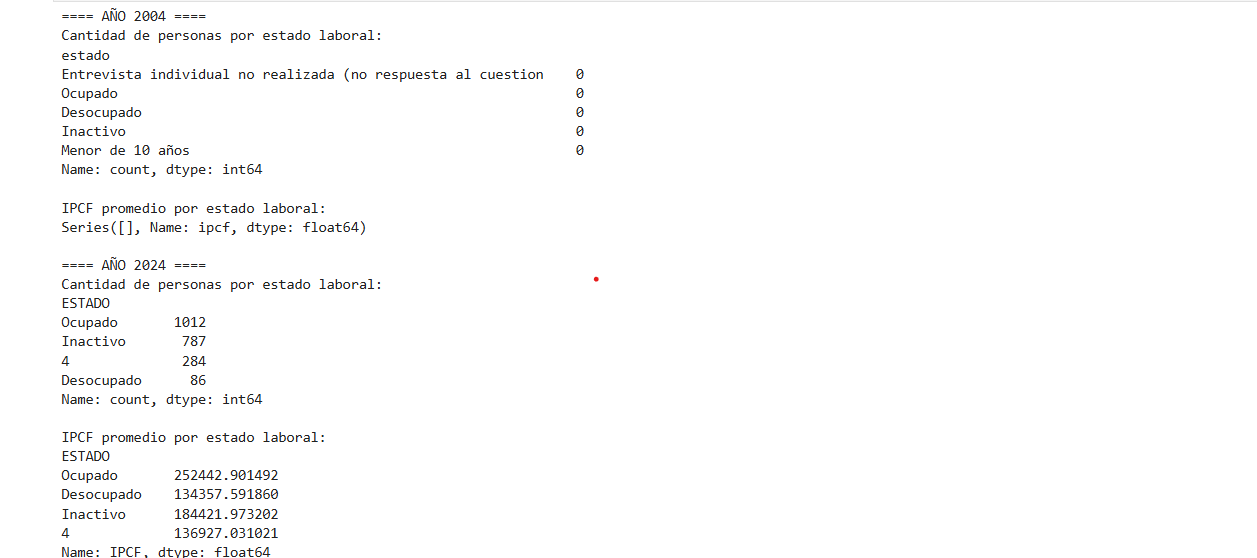


Las matrices de correlación permiten visualizar la relación lineal entre distintas variables socioeconómicas de la población de Mar del Plata para los años 2004 y 2024. En 2004, las correlaciones entre las variables son en general débiles. Se destaca una leve correlación negativa entre el nivel educativo (nivel\_ed) y la edad (ch06), lo cual puede deberse a que las personas mayores, en promedio, poseen menores niveles educativos. También se observa una débil correlación negativa entre la categoría de inactividad (cat\_inac) y el ingreso per cápita familiar (ipcf), lo cual sugiere que los hogares donde predominan personas inactivas tienden a tener ingresos más bajos. Sin embargo, los valores son bajos en todos los casos, lo que indica relaciones muy tenues.

En 2024, en cambio, emergen algunas correlaciones más marcadas. La variable estado (relacionada al empleo) presenta una correlación positiva de 0.39 con la edad (ch07), lo que sugiere que la participación en el empleo formal aumenta con la edad, posiblemente hasta cierto punto. También hay una correlación positiva significativa entre estado y cat\_inac (0.83), que podría interpretarse como un efecto de codificación, ya que ambas variables están relacionadas al estatus laboral y podrían estar representando aspectos superpuestos (por ejemplo, trabajadores activos vs. inactivos). En cuanto a los ingresos (ipcf), no se observan correlaciones fuertes con ninguna variable, aunque se mantienen algunas relaciones débiles, como con estado (0.14), lo que sigue indicando que el hecho de estar ocupado o no tiene un leve impacto en el nivel de ingreso familiar per cápita.

En resumen, entre 2004 y 2024 se observan cambios en las relaciones entre variables, particularmente en el fortalecimiento de algunas correlaciones relacionadas al mercado laboral. Esto podría estar reflejando transformaciones en la estructura ocupacional, educativa o en la forma en que se registra la información en la encuesta.

Parte III: Conociendo a los ocupados y desocupados

5.¿Cuántos desocupados hay en la muestra? ¿Cuántos inactivos? ¿Cuál es la media de ingreso per cápita familiar (IPCF) según estado (ocupado, desocupado, inactivo)?

Año 2004:

No se registraron observaciones válidas para los estados laborales "ocupado", "desocupado" o "inactivo". Por lo tanto, no es posible determinar la cantidad de personas en cada una de estas categorías ni calcular la media del ingreso per cápita familiar (IPCF) según estado laboral. Esto puede deberse a una falla en la codificación o filtrado de los datos durante el procesamiento de la base correspondiente a ese año.

Año 2024:

* Desocupados: 86 personas
* Inactivos: 787 personas
* Ocupados: 1012 personas

Además, hay 284 personas bajo una categoría identificada como "4", que no está claramente definida.

IPCF promedio por estado laboral en 2024:

* Ocupados: $252.442
* Desocupados: $134.358
* Inactivos: $184.422
* Categoría 4: $136.927

Estos valores muestran una clara diferencia de ingresos según la situación laboral. Las personas ocupadas presentan el mayor ingreso per cápita familiar promedio, seguidas por los inactivos y, en último lugar, los desocupados, lo que refleja la vulnerabilidad económica asociada a la falta de empleo.

6. ¿Cuántas personas no respondieron cuál es su condición de actividad?



7. El gráfico compara la distribución de la Población Económicamente Activa (PEA) en Mar del Plata entre los años 2004 y 2024. La PEA está compuesta por personas que están ocupadas o buscando activamente trabajo, es decir, que forman parte del mercado laboral.

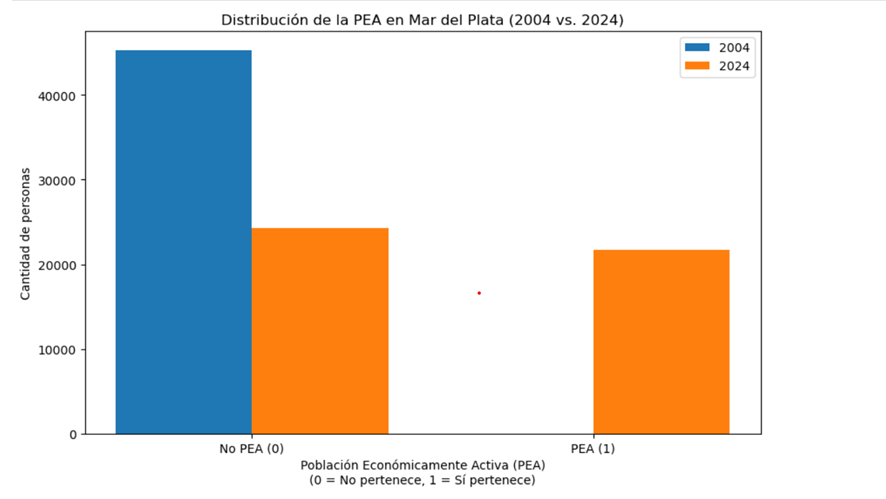
En 2004, se observa una gran mayoría de personas que no formaban parte de la PEA (valor 0), mientras que no hay datos representados para quienes sí pertenecían a la PEA (valor 1). Esto podría deberse a la falta de respuestas sobre el estado laboral en ese año, como se evidenció anteriormente (7.850 personas no respondieron en 2004).

En 2024, se aprecia una distribución más equilibrada entre quienes pertenecen y quienes no pertenecen a la PEA, con aproximadamente la misma cantidad de personas en cada grupo (aunque sigue siendo levemente mayor el grupo "No PEA").

Este cambio puede reflejar:

Mejoras en la calidad de los datos recogidos en 2024, con menos personas sin respuesta.Transformaciones en el mercado laboral, como mayor participación económica de sectores que antes no lo hacían (por ejemplo, mujeres o jóvenes).

Una mejor definición y captación de los conceptos ocupacionales por parte de los encuestadores del INDEC en 2024.



8. El gráfico muestra la composición de la Población en Edad de Trabajar (PET) en Mar del Plata para los años 2004 y 2024. Se considera PET a las personas con edad suficiente para participar del mercado laboral, típicamente mayores de 10 o 14 años, según el criterio del INDEC.

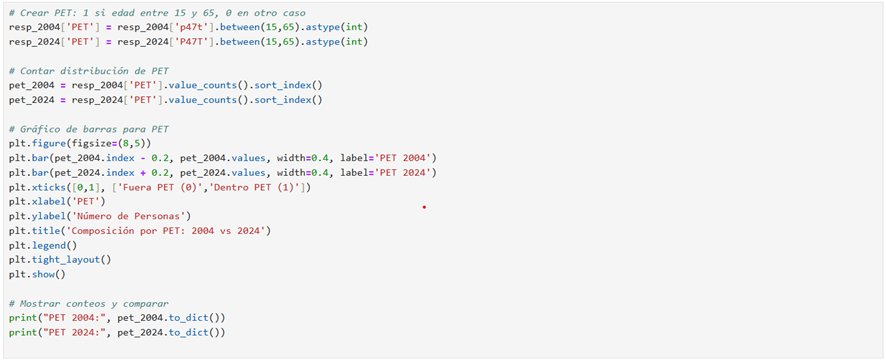
En ambos años, la gran mayoría de personas se encuentran fuera de la PET (valor 0):

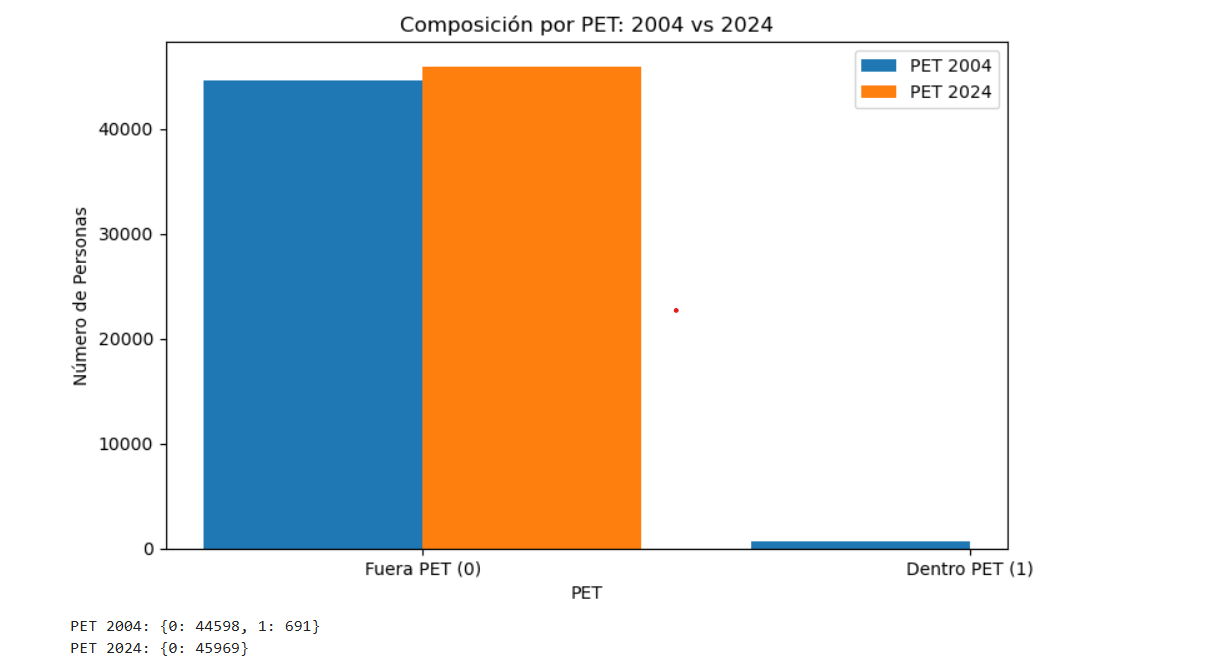
En 2004: 44.598 personas fuera de la PET y solo 691 dentro.

En 2024: 45.969 personas fuera de la PET y no se registran personas dentro.

Esto indica que, según los datos cargados, prácticamente no hay personas clasificadas como dentro de la PET en 2024, lo que sugiere un problema con la codificación o el filtrado de esta variable. Lo más probable es que haya un error en cómo se asignaron los valores 0 y 1, o que no se haya actualizado correctamente la lógica de clasificación en el dataset de 2024.

La leve presencia de personas “dentro de la PET” en 2004 (691) también sugiere que el filtro pudo haber sido mal aplicado o que no se consideraron adecuadamente los rangos etarios.





9.

¿Cuántas personas están desocupadas en 2004?

* Según los datos, solo el grupo con nivel educativo 0 tiene desocupación, y es del 100%.
* Los demás niveles muestran 0% desocupación o no tienen datos.
* Esto sugiere que la muestra es muy limitada, no se puede estimar una cantidad total realista.

¿Hubo cambios por nivel educativo?

* Sí, pero solo hay un valor con desocupación (nivel 0), así que no se puede analizar bien la diferencia entre niveles.}
* Parece que la desocupación afecta solo al nivel más bajo.



