**Big Data and Machine Learning: Trabajo Práctico 2.**

**Integrantes:**

* Kevin Luque Benegas (915007).
* Carlos Ezequiel Martinez (911822).

Parte I: Familiarizándonos con la base EPH y limpieza.

1. Nos dirigiremos a [www.indec.gob.ar](https://www.indec.gob.ar). Luego iremos a la sección “Trabajo e ingresos” y seleccionaremos el apartado “Encueta permanente de Hogares (EPH), y obtendremos los documentos con los datos necesarios para este punto.

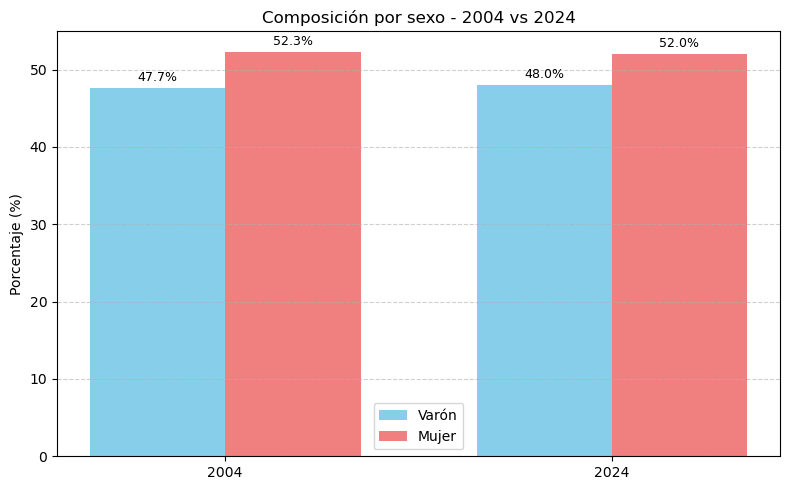
A-) Primero conseguiremos los archivos pedidos en este punto, para poder cargarlos en nuestro código, y también tendremos a mano los diccionarios de cada respectivo año para poder conseguir los códigos que necesitamos. Empezamos el código importando las librerías que necesitaremos, y cargamos los archivos que ya descargamos de 2004 y 2024, importante es recordar que necesitamos únicamente los que dicen individual, por lo que especificamos esto, luego vamos a estandarizar las columnas de ambos años y las vamos a filtrar para conseguir solo los datos de nuestra región seleccionada, ahora buscamos las columnas que tienen en común para evitar errores y por ultimo las unificamos. Es importante antes de terminar que hemos puesto correctamente nuestra dirección de directorio con nuestros archivos para evitar errores.

B-) Seleccionamos 15 variables al azar que podemos obtener del diccionario que teníamos guardado, hacemos que nuestro código haga que se quede únicamente con esas variables y creamos unas tablas con dichas variables y sus respectivos valores faltantes.

C-) Como parte final de nuestro código en este punto N\*1 vamos a eliminar todos esos valores que no tienen sentido, usaremos numpy para facilitarnos las cosas y volveremos a seleccionar nuestras variables, luego crearemos una copia de nuestra base para trabajar en ella la cual tendrá los datos limpios. Debido a que CH06 es una variable un poco mas conflictiva que las otras vamos a darle un tratado especial, y especificamos en el código que solo nos de valores numéricos entre 0 y 100, y ahora si vamos a poder limpiar las variables que son inconsistentes especificando en nuestro Código que no queremos valores negativos ya que es absurdo. Por último mostraremos los valores faltantes por variable después de la limpieza.

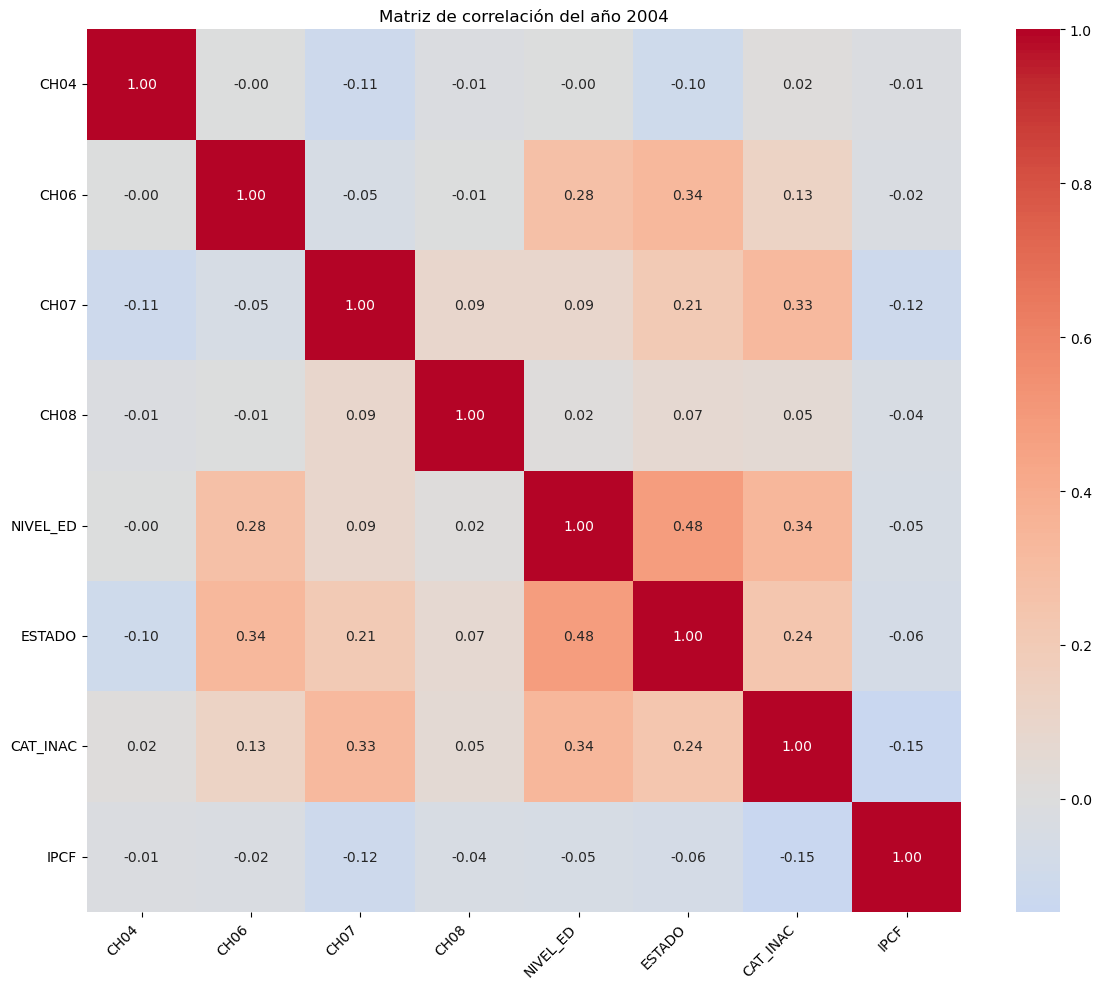
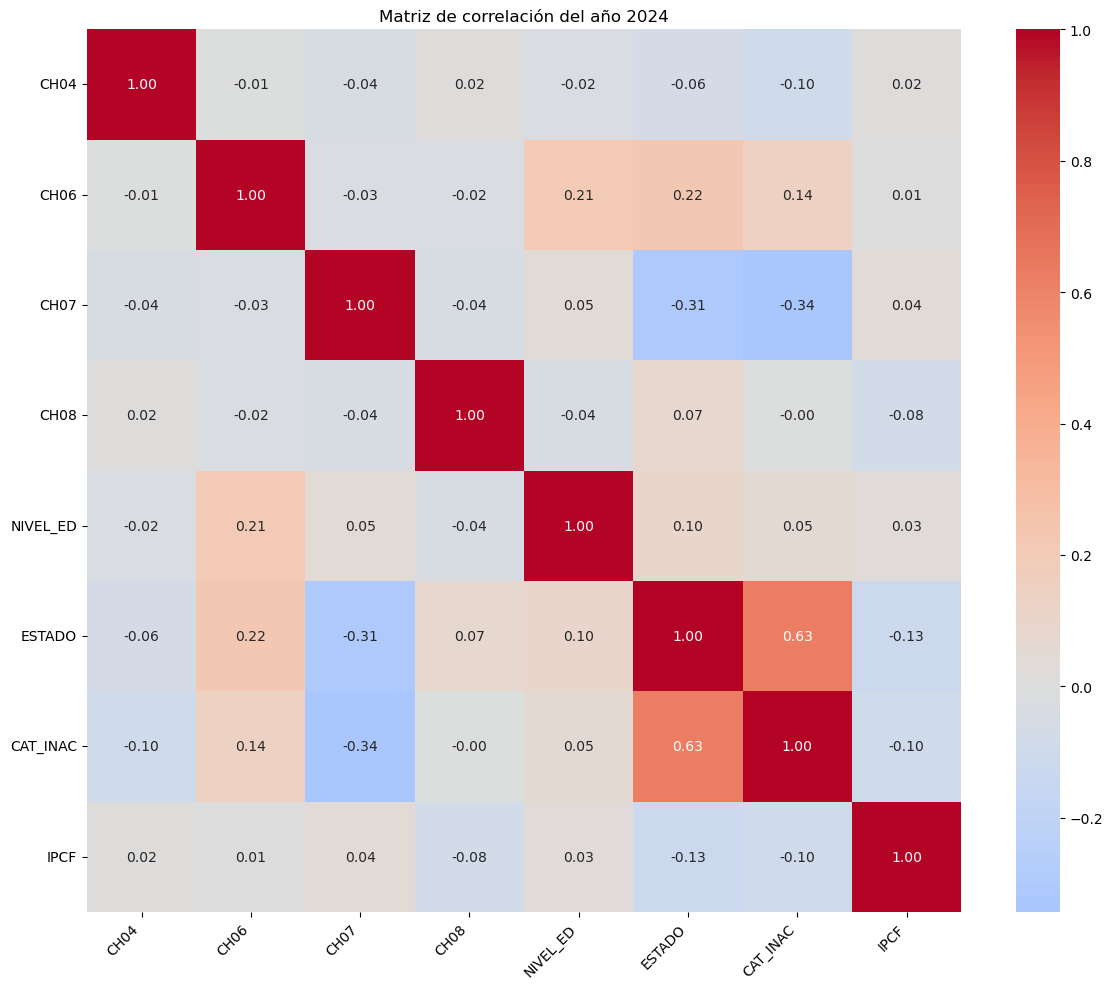
Parte II: Primer análisis exploratorio.

1. Primero vamos a usar print para ver los valores y datos de la variable CH04 de ambos años, con esos datos ahora vamos a hacer un gráfico con un estilo que hemos aprendido anteriormente “pyplot” el cual nos ayudara a hacer un gráfico sencillo pero fácil de entender.



Como verán en este grafico podemos comparar la repartición de población estadísticamente por sexo, en donde denotamos que se ha equilibrado un poco más la población, pero sigue habiendo un gran mayor porcentaje de mujeres que de varones.

1. Ahora no solo con la ayuda de Pyplot sino que también usaremos “Seaborn” para poder realizar la matriz de correlación para ambos años. Empecemos primero apoyándonos con la base unificada que habíamos conseguido puntos anteriores, y con los comandos conseguidos de los links proporcionados para este punto realizaremos un código para crear las matrices para cada año.



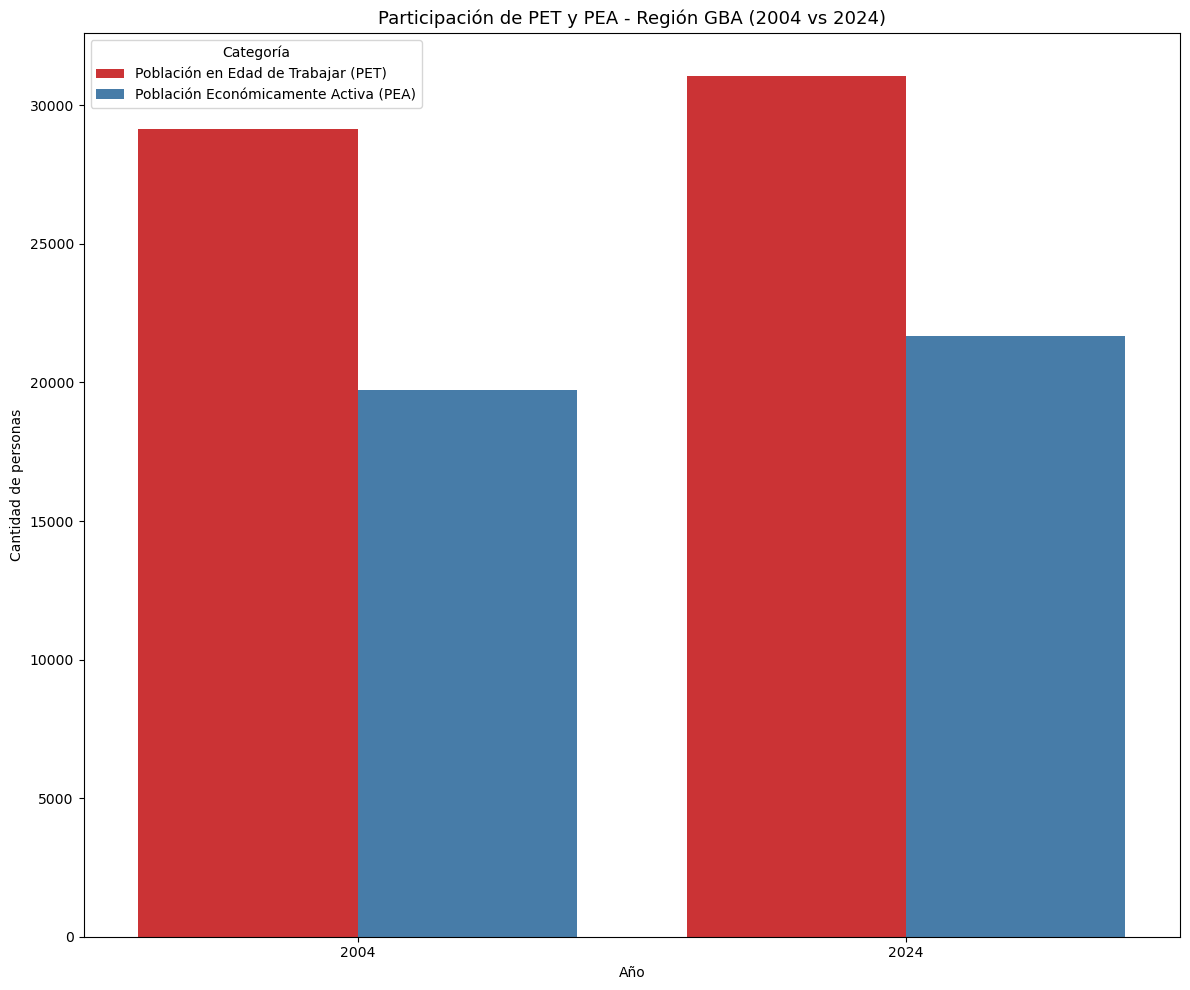
Podemos ver como los datos han cambiado drásticamente entre estos años, pero se pueden llegar a entender debido a la gran distancia entre años, observando que 2024 tiene en general valores más altos .

Parte III: Conociendo a los ocupados y desocupados.

1. Primero definiremos lo que queremos obtener de nuestras muestras, lamentablemente debido a que en 2004 no hay tantos datos tendremos problemas al conseguirlos, pero obtendremos todos los datos que nos puede ofrecer, para 2024 tendremos menos problemas, pero aun así pondremos lo mismo para ambos años, conseguimos los datos, y como podemos observar, los números para 2024 se dispararon para arriba en comparación con 2004, debido a todos nuestros problemas con el desarrollo y economía del país y la región.
2. Primero obtendremos los datos para las personas que no respondieron de la muestra que se hiso, y obtendremos que 41 personas no lo hicieron, lo cual muestra la fuerte presencia de labores en negro o ilícitos en el país.

Ahora como segunda parte guardaremos en 2 bases distintas las personas que “SI” respondieron y las que personas que “NO” respondieron.

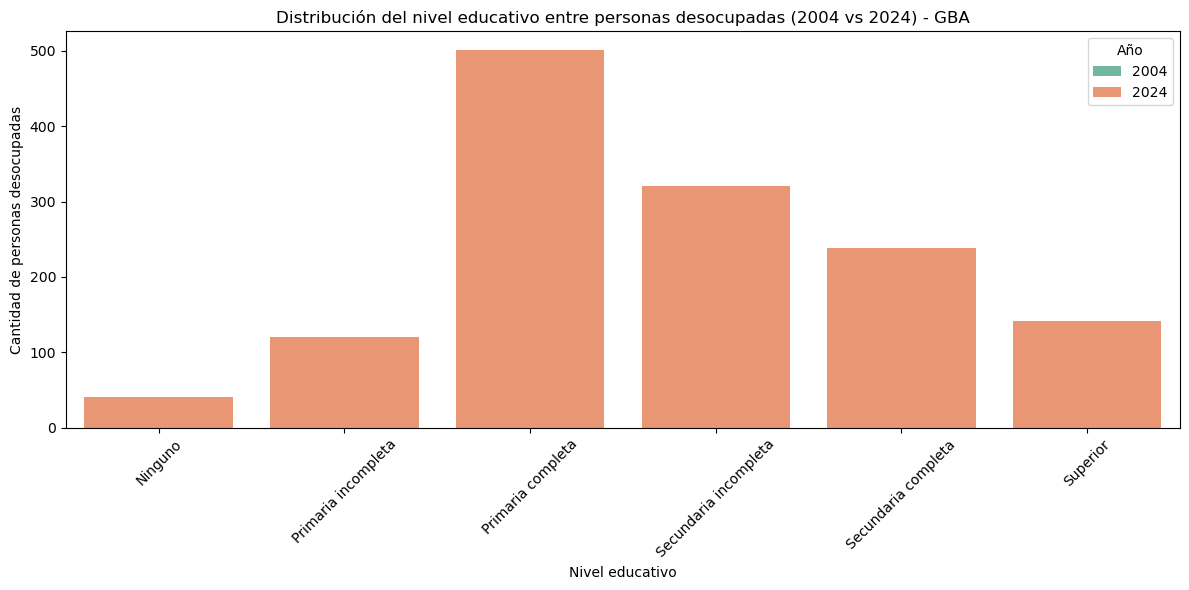
1. Usaremos las mismas herramientas que en el punto 4, haremos unas copias de cada año simplemente para evitar errores a la hora de correr el código. Empezamos filtrando los datos que queremos obtener y creamos la composición PEA, y creamos un gráfico para tener una mejor visualización en donde podremos observar que el PEA ha disminuido entre 2004 y 2024 mostrando un efecto positivo.
2. En este punto haremos lo mismo que en el anterior pero ahora con la composición PET, en el que podremos observar un incremento para 2024 en comparacion de 2004, pero esto tiene sentido ya que la población a crecido exponencialmente desde 2004.



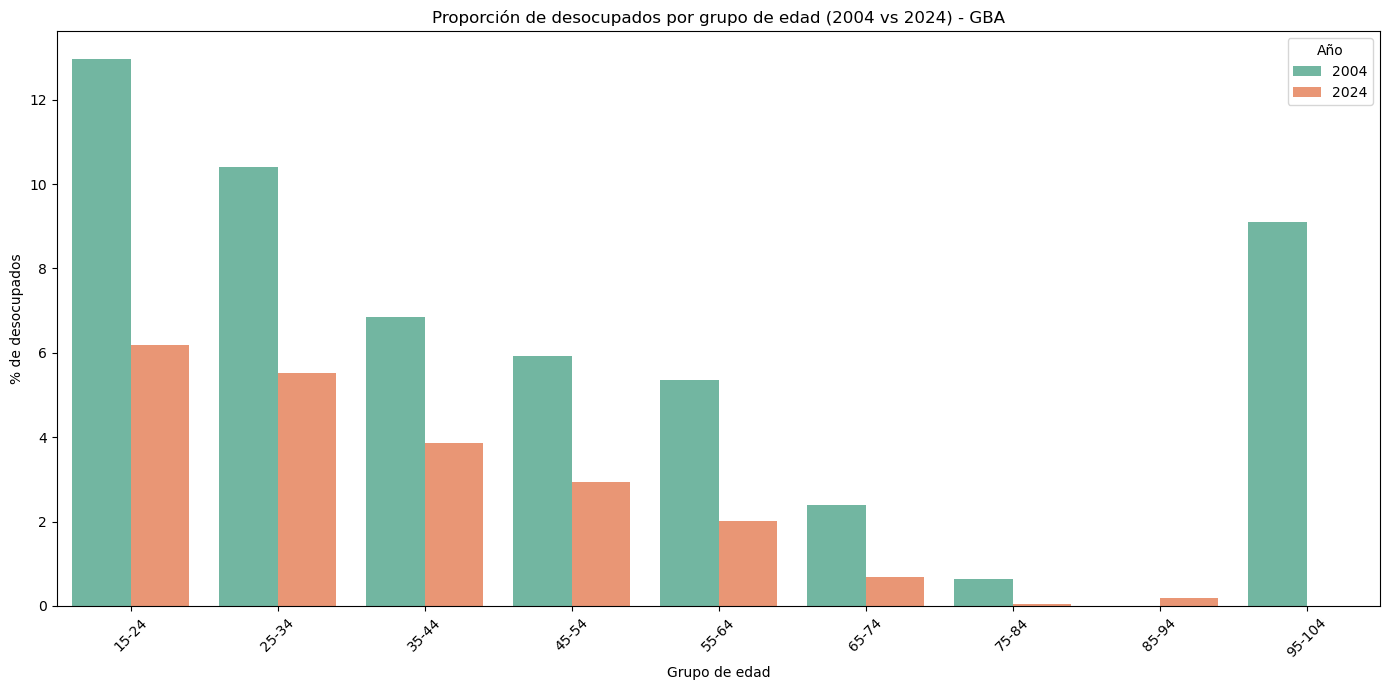
Y como veremos en este grafico tenemos una comparación para las composiciones PEA Y PET DE 2004 Y 2024 en donde ahora veremos más claramente sus diferencias y la representación de cuanta cantidad de gente de diferencia hay entre PEA Y PET, mostrando como no todo el mundo que tiene la edad para trabajar es participe activamente de la economía.

9.

A-) Observemos como los datos para la gente entre ambos años es que solo la gente termina la primaria en mayor medida, esto muestra como la sociedad argentina tiene precariedad en el sector de formación estudiantil y de aprendizaje.



B-) Vamos a hacer un mapeo del estado y crearemos la variable categórica que queremos para CH06, pondremos la categorización para cada 10 años entre 2004 y 2024, es decir 2 instancias y conseguiremos los datos buscados, para facilitar su entendimiento lo pondremos todo en un gráfico.



Veremos como una gran parte de la población de desocupados es claramente más alta entre gente joven y gente de avanzada edad.