UNIVERSIDAD DE BUENOS AIRES - 2025

BIG DATA & MACHINE LEARNING

TRABAJO PRÁCTICO N° 2

UN PRIMER ENCUENTRO CON LA EPH

Fecha de entrega: Abril 22 de marzo a las 13:00 hs.

Contenido*:* familiarización con la base de datos de la Encuesta Permanente de Hogares. Limpieza de datos, valores faltantes y análisis descriptivo.

Modalidad de entrega

- Asegurense de haber creado una carpeta llamada TP2 en el repositorio de GitHub de cada grupo.

- El informe debe subirse a dicha carpeta en repositorio del grupo en formato PDF con el nombre **Big_Data_TP2_Grupo#.pdf** (donde **#** es el número de grupo), incluyendo gráficos e imágenes dentro del mismo archivo. La extensión máxima es de 8 páginas (sin apéndices) y se espera una redacción clara y precisa.

- Se debe publicar el código con los comandos utilizados, indicando claramente a qué inciso corresponde cada uno. El nombre del archivo deberá ser **Big_Data_TP2_Grupo#**.
    - Al finalizar el trabajo práctico deben hacer un último commit en su repositorio de GitHub llamado “Entrega final del tp”.
    - El Jupyter Notebook y el correspondiente al TP2 deben estar dentro de esa carpeta.
    - La última versión en el repositorio es la que será evaluada. Por lo que es importante que:
        - No envien el correo hasta no haber terminado y estar seguros de que han hecho el *commit* y *push* a la versión final que quieren entregar.
        - No hagan nuevos *push* despues de haber entregado su versión final. Esto generaría confusión acerca de que versión es la que quieren que se les corrija.

- También deben enviar el link de su repositorio -para que pueda ser clonado y corregido- a mi correo 25RO35480961@campus.economicas.uba.ar . Usar de asunto de email

"**Big Data - TP 2 - Grupo #"** donde # es el número de grupo que le fue asignado.

- En resumen, la carpeta del repositorio debe incluir:
    - El codigo
    - Un documento Word (Parte A) donde esten las figuras y una breve descripción de las mismas.
- **Cualquier detección de copia o plagio será sancionada.**

Parte I: Familiarizandonos con la base EPH y limpieza

La Encuesta Permanente de Hogares (EPH) es un programa nacional de producción sistemática y permanente de indicadores sociales que lleva a cabo el Instituto Nacional de Estadística y Censos (INDEC), que permite conocer las características sociodemográficas y socioeconómicas de la población. Uno de los indicadores más valiosos sobre el mercado laboral que pueden obtenerse con los datos de esta encuesta es la tasa de desocupación.

1. Utilizando información disponible en la página del INDEC, expliquen brevemente cómo se identifica a las personas desocupadas.

2. Entren a la página https://www.indec.gob.ar/ y vayan a la sección *Servicios y Herramientas* *Bases de datos*. Descarguen la base de microdatos de la Encuesta Permanente de Hogares (EPH) correspondiente al primer trimestre de **2004** y **2024** en formato .dta y .xls, respectivamente (una vez descargadas, las bases a usar deberán llamarse `usu_individual_T104.dta` y `usu_individual_T124.xls`). En la página web, también encontrará un diccionario de variables con el nombre de "Diseño de registro y estructura para las bases preliminares (hogares y personas)". Descarguen el diccionario de cada año. En estos archivos se les indica qué significa cada variable que aparece en la base de datos, en particular, en la sección de Diseño de registros de la base Personas.

    a. A partir de ahora, cada grupo debe decidir trabajar con una región del país en específico (ver variable `REGION`). Eliminen los datos de todas aquellas regiones que no se encuentren dentro de su región y unan ambos trimestres (2004 y 2024) en una sola base.[1]

    b. Asegúrense de que todas las variables tengan el formato correcto. Selecciones 15 variables de interés y reporten la cantidad de valores faltantes (NA, o NaN en Python) en una tabla por cada año. Comenten qué variables de las 15 que seleccionarion tienen más valores faltantes y qué año.

    c. Si notan valores sin sentido (como ingresos negativos) corriganla de acuerdo a la documentación de la EPH (puede ser una codificación de no respuesta de los individuos) y eliminen estos valores extraños de sus 15 variables de interes. Comenten brevemente en el reporte dicho proceso de limpieza.

[1] *Hint*: Note el tipo de variables (string vs. byte or float) entre las dos bases de datos de 2004 y 2024. Deberán unificar todo con un solo tipo de variables.

Parte II: Primer Análisis Exploratorio

3. Realicen un gráfico de barras mostrando la composición por sexo para 2004 y 2024. Comenten los resultados.

4. Realicen una matriz de correlación para 2004 y 2024 con las siguientes variables: `CH04`, `CH06`, `CH07`, `CH08`, `NIVEL ED`, `ESTADO`, `CAT_INAC`, `IPCF`. Utilicen alguno de los comandos disponibles en este link o este link para graficar la matriz de correlación. Comenten los resultados.[2]

Parte III: Conociendo a los ocupados y desocupados

Los siguientes incisos apuntan a ver los resultados para su región seleccionada y comparando 2004 con 2024.

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

6. Uno de los grandes problemas de la EPH es la creciente cantidad de hogares que no reportan sus ingresos (ver por ejemplo el siguiente informe). ¿Cuántas personas no respondieron cuál es su condición de actividad? Guarden como una base distinta llamada `respondieron` las observaciones donde respondieron la pregunta sobre su condición de actividad (`ESTADO`). Las observaciones con `ESTADO=0` guárdenlas en una base bajo el nombre `norespondieron`.

7. Agreguen a la base `respondieron` una columna llamada "PEA" (Población Económicamente Activa) que tome 1 si están ocupados o desocupados en `ESTADO`. Realicen un gráfico de barras mostrando la composición por PEA para 2004 y 2024. Comenten los resultados.

8. Agreguen a la base `respondieron` una columna llamada "PET" (Población en Edad para Trabajar) que tome 1 si están la persona tiene entre 15 y 65 años cumplidos. Realicen un gráfico de barras mostrando la composición por PEA para 2004 y 2024. Comenten los resultados y comparen PET con PEA.

9. Agreguen a la base `respondieron` una columna llamada "desocupado" que tome 1 si la persona está desocupada. ¿Cuántas personas están desocupadas en 2004 vs 2024?

---

[2] Para todos los gráficos que presenten, recuerde tener presentes los tres principios de visualización de datos discutidos en la Clase 1. Referencia: † Schwabish, J. A. (2014). An economist's guide to visualizing data. *Journal of Economic Perspectives*, 28(1), 209-234.

a. Muestre la proporción de desocupados por nivel educativo comparando 2004 vs 2024. ¿Hubo cambios de desocupados por nivel educativo?

b. Cree una variable categórica de años cumplidos (`CH06`) agrupada de a 10 años. Muestre proporción de desocupados por edad agrupada comparando 2004 vs 2024. ¿Hubo cambios de desocupados por edad?

c. Seleccione **una (1)** variable de interés y muestre las diferencias 2004 vs 2024. ¿Hubo cambios de desocupados por edad?