UNIVERSIDAD DE SAN ANDRÉS - 2025

Ciencia de Datos

Trabajo Práctico N° 1

Un primer encuentro con la EPH

Fecha de entrega: 5 de septiembre 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. Medición de pobreza.

Modalidad de entrega

- Asegurense de haber creado una carpeta llamada TP1 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 **Ciencia_Datos_TP1_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 **Ciencia_Datos_TP1_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 TP1 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.

- Deben subir el trabajo práctico en el buzón del campus virtual **Entrega informe 1** con el título de entrega "**Ciencia Datos - TP 1 - Grupo #".**

- Además, deben adjuntar en el documento del informe el link del repositorio público del Github donde tienen los códigos que utilizaron para resolver el trabajo práctico.

- En resumen, la carpeta del repositorio debe incluir:

    - El codigo

    - Un documento PDF donde están 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 pobres.

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 **2005** y **2025** en formato .dta y .xls, respectivamente (una vez descargadas, las bases a usar deberán llamarse `usu_individual_T105.dta` y `usu_individual_T125.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 (los de 2005 y 2025). 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 y mantener dicha región en los próximos trabajos prácticos (ver variable `REGION`). Eliminen los datos de todas aquellas regiones que no se encuentren dentro de su región y unan ambos trimestres (2005 y 2025) en una sola base.[1]

    b. Asegúrense de que todas las variables tengan el formato correcto. Seleccione **15 variables de interés (entre las cuales tienen que estar: `CH04`, `CH06`, `CH07`, `CH08`, `NIVEL ED`, `ESTADO`, `CAT_INAC`, `IPCF`)** y reporten en una figura de heatmap la cantidad de valores faltantes para cada año (NA, o NaN en Python) en una tabla por cada año. Comenten qué variables de las 15 que seleccionaron tienen más valores faltantes y qué año.

    c. Corregir variables si notan valores sin sentido (como ingresos negativos) de acuerdo a la documentación de la EPH (puede ser

[1] *Hint*: Note el tipo de variables (string vs. byte or float) entre las dos bases de datos de 2005 y 2025. Deberán unificar todo con un solo tipo de variables de las 15 seleccionadas en el ítem **2.b**.

una codificación de no respuesta de los individuos) y eliminen estos valores extraños de sus 15 variables de interés. Comenten brevemente en el reporte dicho proceso de limpieza.

Parte II: Primer Análisis Exploratorio

3. Realicen un gráfico de barras mostrando la composición por sexo para 2005 y 2025 en su región. Comenten los resultados.

4. Realicen una matriz de correlación para 2005 y 2025 con las siguientes variables: `CH04`, `CH06`, `CH07`, `CH08`, `NIVEL ED`, `ESTADO`, `CAT_INAC`, `IPCF`. Crear las variables dicotómicas binarias necesarias (variables dummies) y renombrar dichas variables para que las etiquetas tengan sentido en el gráfico de correlación. Utilicen alguno de los comandos disponibles en este link para graficar la matriz de correlación. Comenten los resultados.[2]

Parte III: Conociendo a los pobres y no pobres

Los siguientes incisos apuntan a ver los resultados para su región seleccionada y comparando 2005 con 2025.

5. 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 ingreso total familiar (`ITF`). Las observaciones con `ITF=0` guárdenlas en una base bajo el nombre `norespondieron`.

6. Utilizando el archivo **tabla_adulto_equiv.xlsx**, agreguen a su base de datos una columna llamada `adulto_equiv` que contenga los valores de adulto equivalente de cada persona según su **sexo** y **edad** (por ejemplo, a un varón de 2 años le corresponde 0.46). Finalmente, con el comando `groupby` sumen esta nueva columna para las personas que pertenecen a un mismo hogar y guarden ese dato en una columna llamada `ad_equiv_hogar`[3].

---

[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.

[3] Por ejemplo, si una familia está compuesta por un varón de 40 años (adulto equiv= 1) y su esposa de la misma edad (adulto equiv= 0.77) con sus mellizos varones de 5 años (adulto equiv= 0.60 cada uno), a todos se les debería imputar en **ad_equiv_hogar** un valor igual a 2.97, qué es la cantidad de adultos equivalentes en ese hogar.

7. Sabiendo que la Canasta Básica Total para un adulto equivalente en el primer trimestre de 2025 es aproximadamente $365.177, agreguen a la base `respondieron` una columna llamada `ingreso_necesario` que sea el producto de este valor por `ad_equiv_hogar`. Para el primer trimestre de 2005 la Canasta Básica Total para un adulto equivalente era aproximadamente $205,07. Note que este es el valor mínimo que necesita ese hogar para no ser pobre.

8. Por último, agreguen a `respondieron` una columna llamada `pobre` que tome valor 1 si el ITF es menor al ingreso necesario que necesita esa familia, y 0 en caso contrario. ¿Cuántos pobres identificaron para cada año? ¿Qué porcentaje de la muestra representa?

9. Muestren estadísticas descriptivas relevantes de `pobre` en una tabla, comparando 2005 con 2025. Además, hagan 2 gráficos exploratorios a elección usando la variable `pobre`. Comenten. tulo