CI-0115 Probabilidad y Estadística
Prof. Sebastián Ruiz Blais
I-2024
Grupos 2 y 3

Laboratorio 2: Python, preparación de datos y primeros análisis

Fecha de entrega: 12 de abril del 2024*, 23:55*
Trabaje en parejas. Entregue un archivo comprimido(.zip) que contenga el archivo Jupyter Notebook (extensión ipynb) y su base de datos (extensión csv), en el correspondiente espacio asignado en Mediación Virtual. En el notebook, incluya explicaciones generales del funcionamiento del código. Además, para cada línea de código, incluya explicaciones de qué hace.

Objetivos*:*
1. Descargar e instalar Python 3 y Jupyter Lab (o jupyter notebook, que es la versión previa)
2. Practicar a nivel básico la programación en Python, en el entorno Jupyter-lab
3. Realizar un análisis básico de datos usando Python y datos de su interés

Recursos:
- Para instalar Python y Jupyter:
https://www.youtube.com/watch?v=6j6L3feh1p4
- Tutorial de introducción de Python:
https://www.youtube.com/watch?v=a9UrKTVEeZA

1. (5%) Realice una gráfica escogiendo los datos de dos variables. Un ejemplo sería buscar la relación entre el peso de varios animales y cantidad de neuronas que tienen. Asegúrese que los datos del eje *x* se encuentran ordenados de manera creciente. (No olvide realizar las importaciones necesarias).
2. (5%) En su gráfica, coloque etiquetas para los ejes y un título significativo.
3. (5%) Descargue la siguiente base de datos en su carpeta de trabajo:
https://www.kaggle.com/datasets/arashnic/loneliness-and-social-connections. Se le solicitará crear una cuenta para realizar las descargas. Utilice la biblioteca Pandas para cargar la base de datos “one-person-households-vs-gdp-per-capita.csv” en un data frame. Realice los *imports* necesarios.
4. (5%) Explique en un comentario al inicio de su notebook qué consiste esta base de datos.
5. (5%) Imprima 3 valores de este data frame (por ejemplo, la posición 4 para alguna de sus columnas…)
6. (5%) Obtenga e imprima en pantalla los nombres de las columnas de la base de datos.
7. a. (5%) Modifique el nombre de las columnas correspondientes al GDP por persona y la proporción de hogares con una sola persona. Coloque los siguientes nombres: "GDP per

capita" y “Share 1p household”. Imprima para verificar que las columnas fueron modificadas satisfactoriamente.
b. (5%) Imprima el número de filas y columnas de la base de datos e imprima una muestra de la base de datos usando el comando head(100).
8. (10%) Seleccione los datos que correspondan a años más recientes que 2015. Además, seleccione los datos que únicamente tengan datos no nulos en las columnas "GDP per capita" y “Share 1p household”. Imprima una muestra de la base de datos.
9. (15%) Se tiene algunos países que aparecen varias veces en la muestra, con los años en desorden (a veces sale primero el año menor y a veces el año mayor). Para los países que aparecen varias veces, elimine las filas repetidas, correspondientes a los años más viejos. Por ejemplo, si Suecia aparece con 2016 y 2015, procure eliminar el 2015. Imprima una muestra de la base de datos y la cantidad de filas. Debería tener 43 filas.
10. (5%) Grafique los datos correspondientes a “Share 1p household” y "GDP per capita", como ejes *x* y *y*, respectivamente. Para ello, procure ordenar la base de datos con valores de menor a mayor respecto a la variable “Share 1p household”. Añada etiquetas para los ejes.
11. (5%) Calcule e imprima la correlación entre las variables “Share 1p household” y "GDP per capita". Responda como comentario en el notebook si cree que es una correlación fuerte (investigue para interpretar los números).
12. a. (10%) Hay algunos países que se comportan de manera excepcional. Para mejorar el análisis de datos, a veces es importante eliminar los casos excepcionales (*outliers*). Elimine los países que tengan un GDP per cápita mayor a $70,000 anuales y vuelva a graficar. ¿Aprecia las diferencias en la gráfica?
b. (5%) Vuelva a calcular la correlación entre las variables. Responda: ¿qué ocurrió con la correlación?
13. Calidad de las explicaciones: 10%