Desafío - Variables

- Para realizar este desafío debes haber estudiado previamente todo el material disponible correspondiente a la unidad.
- Una vez terminado el desafío, comprime la carpeta que contiene el desarrollo de los requerimientos solicitados y sube el `.zip` en el LMS.
- Desarrollo del desafío: individual

Habilidades a evaluar

- Hacer uso de métodos de pandas para segmentar columnas y filas.
- Hacer uso de los métodos `iterrows` e `iteritems` para implementar loops en pandas.
- Implementar `enumerate` en loops.
- Conocer las principales convenciones en la visualización de resultados en histogramas, gráficos de punto y barras.
- Generar simulaciones de la distribución normal.
- Conocer las principales aplicaciones de las distribuciones.
- Calcular e interpretar puntajes z.

Descripción

La empresa en la cual usted trabaja tiene como solicitud el análisis del Índice de Desarrollo Humano (IDH) de diversos países para elaborar un informe para el Ministerio de Relaciones Exteriores a fin de poder contar con la data necesaria para poder comenzar a trabajar en relaciones diplomáticas más estrechas con diversos países. Para ello, usted queda a cargo de poder revisar la base de datos *Quality of Government*, de la Universidad de Gotemburgo.

Las unidades de medición en esta base corresponden a 194 países, recolectando los últimos datos de enero del 2018. Además, se le recuerda que la base de datos es un compendio de alrededor de 1900 variables que miden las causas y consecuencias de un buen gobierno en materias económicas, salud pública, medio ambiente, salud institucional, corrupción, entre otras.

Requerimientos

A continuación, revisaremos los requerimientos y acciones que el Ministerio de RREE te pide realizar.

1. **Genere una muestra de casos (1 punto).** Para ello debes considerar:

   - Utilice los últimos 4 dígitos de su rut como semilla pseudoaleatoria.
   - Seleccione el 50% de los casos.
   - Cada base generada debe contener los siguientes elementos:
     - El índice de desarrollo humano (`undp_hdi`).
     - El nombre del país (`ccodealp`).
     - La región a la que pertenece (`ht_region`).
     - El PIB per cápita. (`gle_cgdpc`).
     - El total de la población (`imf_pop`).

   - Si su apellido está entre la A y la M, escoja las siguientes variables del módulo Educación:
     - `ffp_hf`: Human Flight and Brain Drain.
     - `wef_qes`: Quality of the educational system.
     - `wdi_expedu`: Government expenditure on education, total (% of GDP).
     - `wdi_ners`: School enrollment, secondary (% net).

   - Si su apellido está entre la N y la Z, escoja las siguientes variables del módulo Salud:
     - `wef_imort`: Infant mortality, deaths/1000 live births.
     - `who_alc2000`: Alcohol consumption per capita (2000-).
     - `who_tobt`: Current smoking of any tobacco product (Total).
     - `wdi_exph`: Government expenditure on health, total (% of GDP).
     - Guarde esta tabla procesada en un nuevo objeto.
     - Renombre las categorías de la variable `ht_region` de números a regiones.

2. **Genere una función que ingrese su objeto y devuelva lo siguiente (2 puntos):**

   - Por cada variable existente en su objeto, calcule las medidas descriptivas para los casos continuos.
   - Para cada variable discreta, que calcule la frecuencia.
   - Reporte las estadísticas descriptivas para `gle_cgdpc`, `undp_hdi`, `imf_pop`.
   - Compare las estadísticas con algún compañero. ¿Ve alguna diferencia sustancial en alguna de ellas?

3. **Genere una función que liste las observaciones perdidas de una variable (2 puntos)**

- La función debe contener los siguientes argumentos:
  - `dataframe`: La función debe ingresar un objeto DataFrame.
  - `var`: Variable a inspeccionar.
  - `print_list`: Opción para imprimir la lista de observaciones perdidas en la variable. Debe ser False por defecto.
- La función debe retornar la cantidad de casos perdidos y el porcentaje correspondiente.
- Cuando `print_list = True`, debe retornar la lista de casos.
- Analice todas las variables y sus casos perdidos.
  Para las 3 variables con un mayor porcentaje de casos perdidos, solicite la lista de países con ausencia de datos.

4. **Grafique histogramas indicando medias muestral y total (2 puntos)**

- Genere una nueva función que grafique un histograma de una variable entregada para un DataFrame de muestra. El gráfico debe además señalar las medias de la variable entregada, tanto para el `DataFrame` de muestra entregado, como para el `DataFrame` completo correspondiente.
- La función debe incluir los siguientes argumentos:
  - `sample_df`: La base de datos donde se encuentran los datos específicos (muestra).
  - `full_df`: La base de datos donde se encuentran todos los datos (contiene los datos de la muestra).
  - `var`: La variable a graficar.
  - `sample_mean`: Booleano. Si es verdadero, debe generar una recta vertical indicando la media de la variable en la selección muestral (`sample_df`). Por defecto debe ser `False`.
  - `true_mean`: Booleano. Si es verdadero, debe generar una recta vertical indicando la media de variable en la base de datos completa (`full_df`).
- Implemente las funciones para las 4 variables seleccionadas según su grupo. ¿En qué variables la media de la muestra es mayor a la de los datos completos?

5. **Genere una función que devuelva un dotplot con las medias por región para una variable entregada (2 puntos)**

- Cada "punto" del dotplot debe representar la media, o mediana, de una variable para una región específica.
- La función debe contener los siguientes parámetros:

- `dataframe`: La tabla de datos donde buscar las variables.
- `plot_var`: Corresponde a una columna del dataframe entregado, de la cual se desea obtener la métrica (puede ser media o mediana).
- `plot_by`: Corresponde a otra columna del dataframe entregado. Es la columna por la cual se quiere agrupar el `dataframe`, para acceder luego a la columna entregada en `plot_var`.

Tip: Revise el uso del método `groupby` de `pandas`.

- `statistic:` Debe presentar dos opciones; `"mean"` para la media y `"median"` para la mediana. Por defecto debe ser `"mean"`.
- `global_stat`: Booleano. Si es **`True`**, debe graficar la media (o mediana, según lo indicado en `statistic`) de la variable `plot_var` entregada, sin agrupar (para todos los datos entregados en `dataframe`). Por defecto debe ser **`False`**.

- Implemente la función en las 3 variables con una menor cantidad de datos perdidos.

6. Guarde la base de datos (1 punto)

- La submuestra creada tiene un método llamado `to_csv`. Acceda a este y guarde la base de datos con la siguiente nomenclatura: `subsample_<iniciales>_demo.csv`.
  (Subela a la plataforma, junto al desafío)