Laboratorio de Datos

Primer Cuatrimestre 2025

Trabajo Práctico N° 1

El Trabajo Práctico deberá ser resuelto en grupos tres integrantes (excepcionalmente dos integrantes). No se aceptarán entregas individuales. La entrega se realizará a través del campus (pestaña Trabajos Prácticos). La fecha límite es el 19/05 a las 23:59. Deben entregar un Notebook con los nombres de les integrantes del equipo, la resolución de los ejercicios y los informes pertinentes (si utilizaron otros archivos de datos, pueden también incluirlos).
Se valorará que el Notebook y el código tengan un formato prolijo: ejercicios separados por títulos (Ejercicio 1, Ejercicio 2, etc.), nombres descriptivos para las variables, comentarios, etc. Recomendamos que antes de entregar el TP, corroboren que todas las celdas corren sin errores:
`Kernel -> Restart Kernel and Run All Cells`

Trabajaremos con el dataset `indicadores-proceso-trayectoria-sobreedad-2012_2022.csv`[1] que contiene mediciones de variables educativas de la escuela primaria y secundaria en la Provincia de Buenos Aires (PBA) desde 2012 a 2022. En este link pueden consultar la descripción de cada variable.

```
df_educacion =
   pd.read_csv('indicadores-proceso-trayectoria-sobreedad-2012_2022.csv')
```

Observación: en este DataFrame los nombres de los municipios están sin acento.

Procesamiento de datos [2 pts.]

En esta primera sección, vamos implementar algunas acciones básicas de procesamiento de nuestros datos.

1. Describir el tipo de variable de cada columna.
2. ¿La base de datos contiene datos faltantes? ¿Cuántos?
3. Además de la división por municipio, nos interesa saber a que zona de la provincia pertenece cada municipio (Zona Norte, Zona Oeste , Zona Sur e Interior). Para esto:
   (a) Implementar una función `determinar_zona` que, dado el nombre de un municipio, devuelva a que zona pertencece: 'Norte', 'Oeste', 'Sur' o 'Interior'.
   (b) Agregar la columna `zona` a `df_educacion` que contenga la zona a la que pertenece el municipio.
   Sugerencia: tener en cuenta el método `apply`
4. Quisiéramos analizar si en general la tasa de repitencia es más alta en primaria que en secundaria. Para esto:
   (a) Agregar una columna `mayor_repitencia_primaria` a `df_educacion` que indique si ocurre esta situación.
   (b) Mostrar en una Serie de `pandas` cuantas veces entre los años 2017 y 2022 (inclusive) la repitencia es más alta en primaria que en secundaria para cada municipio.

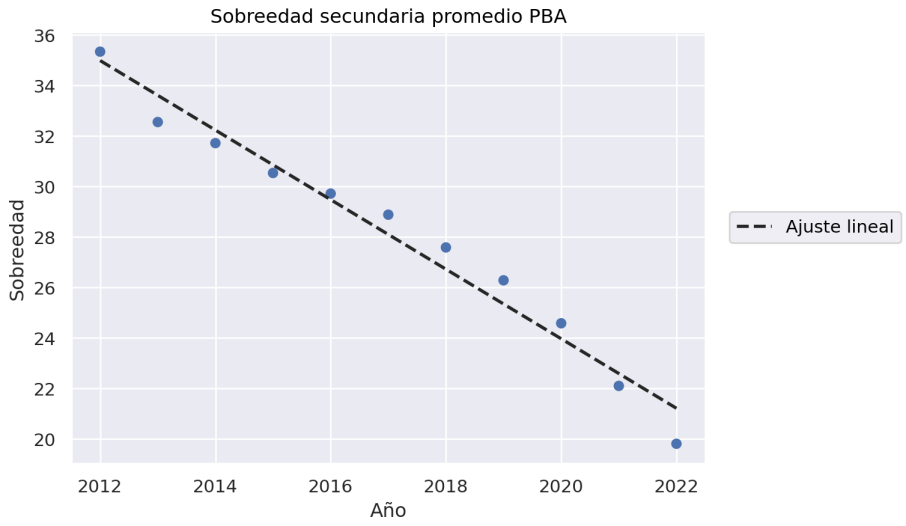
[1] Fuente: https://catalogo.datos.gba.gob.ar/dataset/indicadores-de-trayectoria

(c) Mostrar en un DataFrame la media de repitencia en primaria y la media de repitencia en secundaria agrupados por año y por zona.

Regresión Lineal [2 pts.]

En esta sección, sugerimos que, una vez realizada una Regresión, utilicen los valores de β_0 y β_1 para predecir, en vez del método `predict` de `scikit-learn`.

1. Reproducir el siguiente gráfico sobre la evolución de la sobreedad en secundaria promedio en la Provincia de Buenos Aires a lo largo de los años:



2. Realizando una Regresión Lineal, predecir para el año 2025 la sobreedad en secundaria promedio en la Provincia de Buenos Aires.

3. Para cada municipio del Área Metropolitana, mediante Regresión Lineal, predecir la sobreedad en secundaria para el 2025. Elaborar un DataFrame indexado por el nombre del municipio, con dos columnas: `sobreedad_2025` con el resultado de la predicción para 2025 y `menor_que_promedio` que indique con un booleano si la predicción es menor al valor calculado en el ítem anterior.

   Sugerencia: puede basarse en la siguiente estructura

```
prediccion_data = []
menor_que_promedio_data = []

for partido in area_metropolitana:
    ##########################
    # Realizar la regresion #
    ##########################

    prediccion_data.append(???)
    mejor_que_promedio_data.append(???)

predicciones_df = pd.DataFrame(
    data = {'prediccion_2025': ???, 'mejor_que_promedio': ???},
    index = ???
```

)

Visualización con datos geográficos [2 pts.]

El objetivo de esta parte del TP es visualizar datos sobre un mapa de Buenos Aires. Para esto, utilizaremos las librerías `geopandas`, que nos permite trabajar con un `.geojson` como si fuera un DataFrame de `pandas`, y `plotly` para generar mapas interactivos.

1. Cargar los datos de `partidos.geojson` utilizando `geopandas`:

   ```
   geo_df = gpd.read_file('partidos.geojson')
   ```

   Si imprimimos `geo_df` observamos que cada fila corresponde a un municipio con distintas características y, además, hay una columna `geometry` que tiene las coordenadas de los límites de los partidos. Pueden visualizar estos datos con `geo_df.plot()`.

2. Identificar qué columna se corresponde al número de identificación de cada municipio. Mostrar todos los municipios con números de identificación repetidos.

   Sugerencia: utilizar el argumento `keep` del método `duplicated`

3. Aplicar las siguientes modificaciones a `geo_df`:

   (a) eliminar las filas que correspondan a islas de partidos (por ejemplo, Islas Ramallo).

   (b) convertir la columna `cde` a numérica (recordar el método `astype`)

   (c) corregir los valores en la columna `cde` para los partidos de Chascomús y Lezama, para que coincidan con sus valores en la columna `municipio_id` de `df_educacion`.

   Sugerencia: puede basarse en el siguiente código para corregir los valores:

   ```
   geo_df.loc[geo_df[???] == ???, <columna a modificar>] = ???
   ```

   Finlamente, exportar `geo_df` como un GeoJSON llamado `partidos_limpio.geojson`:

   ```
   geo_df.to_file('partidos_limpio.geojson', driver='GeoJSON')
   ```

4. A partir del DataFrame `df_educacion`, crear un DataFrame `repit` que tenga tres columnas: `cde` con el código del municipio, `municipio_nombre` con el nombre del municipio y `repitencia_secundaria` con el promedio de repitencia en secundaria del municipio.

   Sugerencia: Partir de `df_educacion.groupby(['municipio_id', 'municipio_nombre'])`. Si obtienen una serie con multi-indices, pueden convertirlo a DataFrame con columnas para los índices utilizando `.reset_index()`. O si ingresan el nombre de la columna `repitencia_secundaria` entre doble corchetes, obtienen también un DataFrame con las tres columnas.

5. Con el siguiente código podemos elaborar un mapa interactivo:

   ```
   # Cargamos el GeoJSON
   with open('partidos_limpio.geojson', 'r') as f:
       partidos = json.load(f)

   # Elaboramos el mapa
   fig = px.choropleth_map(
       repit, # DataFrame con los datos a visualizar
   ```

```
    geojson=partidos, # archivo GeoJSON
    # los siguientes argumentos vinculan el GeoJSON con el
    # DataFrame: en featureidkey le indicamos que en el GeoJSON el
    # identificador del municipio es la propiedad cde
    # y con locations le indicamos que en el DataFrame
    # el indicador del municipio esta en la columna 'cde'
    featureidkey = 'properties.cde',
    locations='cde',
    # le indicamos la columna del DataFrame a visualizar
    color='repitencia_secundaria',
    # los siguientes argumentos determinan el zoom y las
    # coordenadas iniciales del mapa
    zoom = 4,
    center = {"lat": -34.61315, "lon": -58.37723},
)

# Mostramos el mapa interactivo
fig.show()
```

Agregar argumentos a `choropleth_map` para que:

(a) la escala de color sea `'Bluered'` (pueden ver otras escalas de color en este link)

(b) la opacidad del color sea de `0.7`

(c) que al pasar el cursor sobre un partido, se muestre también su nombre

Análisis exploratorio [4 pts.]

1. La idea de este ítem es que realicen un análisis exploratorio de los datos, aplicando las herramientas de visualización (`seaborn.objects`, `seaborn` y/o `matplotlib`), de resumen de datos (media, mediana, desvío estándar, operaciones sobre el DataFrame, etc.) y/o de Regresión.

   El objetivo es entender, comparar y/o estudiar aspectos en los indicadores educativos de los partidos de la Provincia de Buenos Aires. Algunas preguntas **disparadoras** pueden ser:

   - ¿Existe alguna relación entre la repitencia y los fondos enviados al partido para la educación?
   - ¿Podemos explicar por qué hay algunos partidos con repitencia muy superior al promedio?
   - ¿Qué relación hay entre la cantidad de habitantes y la cantidad de establecimientos educativos?

   No es necesario que respondan a cada una de esas preguntas (ni se limiten a eso), lo mejor es que exploren por donde se les ocurra. Alentamos que se planteen hipótesis y usen los datos para corroborarlas o rechazarlas. Pueden aplicar cualquiera de las herramientas que hemos visto hasta ahora. Asimismo, pueden centrarse en un conjunto de partidos, en un conjunto de indicadores educativos, etc.

   A continuación dejamos otros datasets que pueden serles de ayuda (su uso es opcional):

   - Establecimientos educativos
   - Población
   - Transferencias a Consejos Escolares

- Transferencias a municipios

También pueden explorar datasets del Gobierno Nacional o de la Ciudad de Buenos Aires (pero que el análisis esté centrado en PBA).

Importante: en el Notebook, las visualizaciones y resúmenes de datos que realicen deben estar acompañados por las conclusiones que obtengan a partir de ellos.