



## Objetivo del Trabajo Práctico 01


Evaluar el manejo de datos y su visualización por parte de cada uno de los alumnos.

## Enunciado

Los docentes de la materia Laboratorio de Datos se han encontrado con una fuente de datos abiertos correspondientes a los Establecimientos Educativos y los Centros Culturales de la República Argentina. En particular, están interesados en saber si existe cierta relación entre la cantidad de establecimientos educativos en cada una de las provincias, y la cantidad de centros culturales. A continuación se detallan los datos con los que se cuenta. Previo a arribar a una conclusión, los docentes desean conocer cierta información adicional de las fuentes de datos.

## Datos

### Fuentes

1. **Establecimientos Educativos (EE).** Padrón Oficial de Establecimientos Educativos 2022. Disponible en:  
<https://www.argentina.gob.ar/educacion/evaluacion-e-informacion-educativa/padron-oficial-de-establecimientos-educativos>
2. **Centros Culturales (CC).** Padrón de Centros Culturales.  
[https://datos.gob.ar/dataset/cultura-mapa-cultural-espacios-culturales/archivo/cultura\\_0e9a431c-b4f7-455b-aa1a-f419b5740900](https://datos.gob.ar/dataset/cultura-mapa-cultural-espacios-culturales/archivo/cultura_0e9a431c-b4f7-455b-aa1a-f419b5740900)
3. **Población.** Datos de población por Departamento. Se pueden obtener de los datos del censo de 2022, sección **Estructura por edad de la población**. Está disponible en:  
<https://www.indec.gob.ar/indec/web/Nivel4-Tema-2-41-165>  
Descargar el archivo xlsx generado por la consulta siguiendo este enlace:  
 padron\_poblacion.xlsx

## Objetivos

Se espera que para resolver el problema los estudiantes cumplan con los siguientes puntos:

- Plantear bien el objetivo general del trabajo solicitado.
- Dado que existen actividades que van a requerir de datos para alcanzar el objetivo, en primer lugar deberán realizar actividades para comprender el contenido de las fuentes de datos. Luego, deben leer todo el enunciado del TP, analizarlo y definir bien qué actividades deberán realizar y qué datos de las fuentes de datos deberán retener para llevar a cabo cada una de ellas (consultas, visualizaciones, etc.).

- Una vez definidas dichas actividades, deberán armar un diagrama conceptual de los datos (DER) que sea adecuado para los objetivos del trabajo, utilizando (solamente) los datos necesarios para resolverlo. No es necesario armar un DER por cada fuente de datos original (previa a procesar) ya que varios atributos quizás no sean relevantes para resolver el problema. Luego, deberán implementar un modelo relacional basado en el DER, decidir de dónde van a obtener los datos (de qué fuente de datos) y finalmente alimentarlos con los datos (limpios).
- Realizar las actividades solicitadas.
- Redactar el informe y realizar la entrega.

## Primeros Paso

- Descargar los datos de las fuentes de datos. En general, para comprender en detalle los datos, las páginas de descarga suelen contener documentación acerca de las fuentes (en algunos casos más detallada y en otros menos).
- Plantear el objetivo general del trabajo.

## Procesamiento de Datos

- Analizar las formas normales en que se encuentran las tablas de **Establecimientos Educativos** y **Centros Culturales**. Justificar de manera concisa.
- Revisar la calidad de datos de las fuentes. Describir los problemas de calidad de datos detectados en los datasets con los que trabajan. No es necesario que describan todos los problemas, pero sí al menos uno por cada fuente de datos utilizada. No puede ser el mismo problema para todas las fuentes (elegir al menos uno distinto para las tres tablas originales utilizadas). De esta manera, para cada uno de los datasets y cada problema de calidad deben mencionar:
  - el atributo de la calidad afectado,
  - si el problema corresponde a modelo y/o a instancia,
  - una medida concreta acerca de la magnitud del problema (usar el método GQM de manera estricta, es decir, mencionando de manera explícita el objetivo, las preguntas y las métricas).

En caso de corregir los datos, describir en cada caso qué criterios utilizaron para corregirlos y cómo impacta en la calidad (por ejemplo, cómo cambian los valores en las métricas).

- Generar un Diagrama Entidad-Relación (DER) que permita modelar de manera conceptual solamente los datos necesarios para resolver los problemas planteados en el presente trabajo práctico.
- Armar esquemas correspondientes al modelo relacional del DER del punto anterior. Todos ellos deben estar en 3FN. Para cada uno de ellos (no olvidar ninguno de estos puntos) definir:
  - i) Clave primaria (PK)
  - ii) Dependencias funcionales (DF). En lo posible, se desea que no escriban la totalidad de ellas sino un conjunto minimal de las mismas
  - iii) Claves foráneas (Foreign keys)

- Importar los datos (ya limpios) a los esquemas. Cada esquema del modelo relacional debe estar representado en un DataFrame de igual nombre, y con las mismas columnas. **Documentar en el informe** desde qué fuentes de datos se está importando la información de los DataFrames.

## Análisis de datos

- A partir de los esquemas con datos del punto anterior, generar los siguientes reportes **utilizando sólo consultas SQL**:
  - Para cada departamento informar la provincia, cantidad de EE de cada nivel educativo, considerando solamente la modalidad común, y cantidad de habitantes por edad según los niveles educativos. El orden del reporte debe ser alfabético por provincia y dentro de las provincias, descendente por cantidad de escuelas primarias.

Provincia	Departamento	Jardines	Población Jardín	Primarias	Población Primaria	Secundarios	Población Secundaria
Buenos Aires	Martínez	50	2000	60	3500	54	2770
Buenos Aires	Lanús	80	2200	50	3200	22	2900
...	...	...	...	...	...	...	...

Importante: Para el ejemplo no necesariamente han sido tenidos en cuenta los datos de la fuente de datos.

- Para cada departamento informar la provincia y la cantidad de CC con capacidad mayor a 100 personas. El orden del reporte debe ser alfabético por provincia y dentro de las provincias, descendente por cantidad de CC de dicha capacidad.

Provincia	Departamento	Cantidad de CC con cap >100
Buenos Aires	Avellaneda	20
Buenos Aires	La Plata	8
...	...	...

Importante: Para el ejemplo no necesariamente han sido tenidos en cuenta los datos de la fuente de datos.

- Para cada departamento, indicar provincia, cantidad de CC, cantidad de EE (de modalidad común) y población total. Ordenar por cantidad EE descendente, cantidad CC descendente, nombre de provincia ascendente y nombre de departamento ascendente. No omitir casos sin CC o EE.

Provincia	Departamento	Cant_EE	Cant_CC
Córdoba	CAPITAL	1415	30
Santa Fe	Rosario	1263	36
...	...	...	...

Importante: Para el ejemplo no necesariamente han sido tenidos en cuenta los datos de la fuente de datos.

- iv) Para cada departamento, indicar provincia y qué dominios de mail se usan más para los CC.

Provincia	Departamento	Dominio más frecuente en CC
Córdoba	CAPITAL	gmail
Santa Fe	Rosario	hotmail
...	...	...

- Mostrar, utilizando herramientas de visualización, la siguiente información:
  - Cantidad de CC por provincia. Mostrarlos ordenados de manera decreciente por dicha cantidad.
  - Graficar la cantidad de EE de los departamentos en función de la población, separando por nivel educativo y su correspondiente grupo etario (identificándolos por colores). Se pueden basar en la primera consulta SQL para realizar este gráfico.
  - Realizar un boxplot por cada provincia, de la cantidad de EE por cada departamento de la provincia. Mostrar todos los boxplots en una misma figura, ordenados por la mediana de cada provincia.
  - Relación entre la cantidad de CC cada mil habitantes y de EE cada mil habitantes.

Importante: En el informe, todos los reportes y gráficos deben ser acompañados por texto explicativo de lo observado en ellos y con las reflexiones que puedan desarrollar.

Finalmente, recordar que a modo de conclusión del trabajo se desea que intenten responder "... si existe cierta relación entre la cantidad de CC y EE en los departamentos del país". En caso de que aún no lo hayan hecho, ¿qué información les parece que deberían mostrar que aún no han mostrado? Enumerar y mostrar los resultados.

Es importante documentar todo el proceso y que todos los integrantes se involucren en el mismo.

## Grupos

Los grupos deben estar conformados por 3 (y sólo 3) integrantes. Ni más, ni menos. Deberán i) registrar la conformación del grupo en la siguiente planilla, y ii) definir quién va a ser el encargado del envío (debe ser uno y sólo uno de los integrantes del grupo):

[https://docs.google.com/spreadsheets/d/10Hf1JZ\\_boloeltCiecEo6kZoD3vPuBd1a5wpA2WfGl4/edit?usp=sharing](https://docs.google.com/spreadsheets/d/10Hf1JZ_boloeltCiecEo6kZoD3vPuBd1a5wpA2WfGl4/edit?usp=sharing)

## Acerca de la entrega

### Informe

La **documentación deberá ser entregada** en un informe. El mismo se debe entregar en formato pdf a través del **campus y** también una **versión impresa**. El informe debe contener:

- **Carátula**, con el nombre de la materia y del TP del que se trata, nombre del grupo y nombres de los miembros del grupo.
- **Sección Resumen**, que resuma la problemática, el trabajo realizado y las conclusiones a las que arribaron.
- **Sección Introducción**, en donde se introduzca el problema a resolver, el objetivo general, las actividades a realizar para alcanzar dicho objetivo y un resumen de la resolución y de cómo continúa el documento.
- **Sección Procesamiento de Datos**, donde se mencione en qué forma normal se encontraban las fuentes de datos originales, el análisis de calidad realizado, qué procesos se siguieron para limpiar y combinar las fuentes de datos, la documentación del DER y su representación en el modelo relacional, y una descripción del proceso de importación de datos mediante el cual se generaron las tablas asociadas al modelo relacional.
- **Sección Decisiones tomadas**, que explique las mismas en el caso de que hayan tenido que tomar alguna. Por ejemplo, omitir ciertas instancias por falta de valores en algún atributo determinado, imputación de datos faltantes, etc.
- **Sección de Análisis de datos**, en la que se encuentren las respuestas a las preguntas planteadas en los objetivos del Análisis de Datos. En el caso de reportes que involucren muchas filas, los mismos podrán ser incorporados en un **anexo como material suplementario o en un archivo csv, en el caso de las consultas (mencionando su ubicación)**. En estos casos, incluir en el informe las primeras filas de dicho reporte junto con la indicación de dónde se encuentra su versión completa.
- **Sección de Conclusiones**.

El largo total del informe (sin contar la carátula ni el material suplementario) no debe exceder las 14 páginas A4 (utilizando un formato de letra Arial 11). Se evaluará que el documento (en formato .pdf) sea **conciso**, además de considerar la completitud y correctitud de escritura del mismo.

## Código

Deberán entregar también el código generado en python (archivo .py). Al comienzo del código deben incluir un encabezado con el nombre de los integrantes del grupo, una descripción del contenido y otros datos que consideren relevantes.

El código debe tener comentarios donde se explique cada sección y debe poder correrse correctamente en cualquier máquina. Las variables usadas en el código y las tablas del modelo de datos tienen que tener nombres representativos. Al correr el código se deben generar correctamente los resultados que responden a todos los ejercicios. En particular, deben generarse las tablas asociadas a los esquemas del modelo relacional (con mismo nombre y atributos), así como también las tablas obtenidas con las consultas sql y los gráficos realizados en la sección de Análisis de Datos. Las tablas originales y las correspondientes a los esquemas del modelo relacional deberán entregarlas con el resto del TP. Aquellas originales deberán estar en una carpeta denominada `TablasOriginales` y aquellas asociadas al modelo relacional, que deben estar en formato csv, deben estar en una carpeta llamada `TablasModelo`.

## Autoevaluación

Al finalizar el trabajo, y **antes de enviar el TP-01**, realizar lo siguiente:

- Copiar la siguiente planilla de autoevaluación (una sola a nivel grupal) a una carpeta personal:  
[https://docs.google.com/spreadsheets/d/1igf\\_9g4vuya3GNAgeludjiW\\_y2-04JEmX6F7U-NOh3c/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1igf_9g4vuya3GNAgeludjiW_y2-04JEmX6F7U-NOh3c/edit?usp=sharing)
- Completarla.
- Descargarla como pdf y agregarla al envío virtual y en papel.

El trabajo práctico (documento con el informe, código, ambos directorios con los archivos de datos, y el documento de autoevaluación) deberán subirse al campus en formato .zip (lo subirá el responsable del grupo encargado del envío). El nombre del archivo deberá ser **TP01-nombredelgrupo.zip**. La fecha límite para subir el TP es el domingo **23 de febrero a las 23:50 hs**. El día martes 25 de mayo, antes de las 12:30, deben entregar el informe impreso junto con la autoevaluación.



## Anexo: Instrucciones en python que pueden ser de ayuda

Para más información pueden acceder a la documentación de cada biblioteca o usar los comandos 'help()' y el operador '?' en la consola de spyder.

- `pd.read_excel(sheet_name='...', skiprows=)` : comando para leer archivos tipo .xlsx, el atributo `skiprows` permite saltar las primeras n líneas del archivo. Requiere tener la biblioteca `openpyxl`.
- `df.dropna()` : Elimina las tuplas con valores nulos en alguna de las columnas del dataframe dado.
- `df.to_csv()` : Exporta un dataframe como archivo .csv.
- `fig.savefig('nombre.png')` : Exporta una figura de matplotlib como png.
- `np.where()` : Permite reemplazar los valores de una columna de un dataframe que cumplen con una condición dada.