

## Objetivo del Trabajo Práctico 01

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

## Enunciado

Desde diversas áreas del conocimiento se ha propuesto que la educación está íntimamente relacionada al desarrollo productivo de una sociedad. Dicha afirmación ha sido sostenida por innumerables fuentes de datos y análisis. El objetivo del presente trabajo es explorar si en Argentina existe una relación entre la presencia de establecimientos educativos, y el desarrollo de actividades productivas. Para lo cual se dispone de fuentes de datos abiertos correspondientes a los Establecimientos Educativos, Establecimientos Productivos y de Población de la República Argentina. A continuación se detallan los datos con los que se cuenta, y los análisis propuestos.

## Datos

### Fuentes

1. **Establecimientos Educativos (EE).** Padrón Oficial de Establecimientos Educativos del año 2022. Disponible en:  
<https://www.argentina.gob.ar/educacion/evaluacion-e-informacion-educativa/padron-oficial-de-establecimientos-educativos>
2. **Establecimientos Productivos (EP).** Padrón de Establecimientos Productivos. Los establecimientos son locales, oficinas, plantas, talleres, sucursales, comercios, etc, en los que se desarrollan actividades económicas. Las actividades económicas son las distintas acciones o procesos que realizan las personas, empresas o instituciones para producir bienes o servicios, con el fin de satisfacer las necesidades de la sociedad. Estas generan valor económico y pueden clasificarse en diferentes categorías según el tipo de bienes o servicios que producen o la forma en que se llevan a cabo. Entre otras pueden ser: actividades primarias, que están relacionadas con la extracción u obtención directa de los recursos naturales y que no requieren de un proceso de transformación industrial complejo (como la agricultura, ganadería, pesca y minería); las actividades secundarias o industriales se refieren a la transformación de los recursos naturales en productos elaborados o bienes manufacturados y actividades terciarias, que son de servicios (comercio, educación, salud, etc).  
<https://datos.produccion.gob.ar/dataset/distribucion-geografica-de-los-establecimientos-productivos/archivo/16f0dfc2-a9ff-4696-b2e5-79831c5c0ec4>  
Descargar el dataset llamado: **Datos de establecimientos por departamento, actividad y género.**

Para interpretar los códigos de las actividades, descargar la siguiente tabla:


<https://datos.produccion.gob.ar/dataset/distribucion-geografica-de-los-establecimientos-productivos/archivo/d8ca32fa-a523-4583-96cc-97e4d21467aa>

(Descargar el dataset llamado Actividades)

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 Particulares

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 pasos que van a requerir de datos adicionales para alcanzar el objetivo, en primer lugar deberán realizar ciertas tareas 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, es decir antes de alimentarlos con los datos deberán evaluar la necesidad de hacer un trabajo de limpieza de los mismos y realizarlo).
- Realizar un análisis de los datos a través de consultas y visualizaciones.
- Una vez realizadas las actividades, redactar el informe y realizar la entrega en tiempo y forma.

## Ejercicios

### Primeros Pasos

- Plantear el objetivo general del trabajo, revisar los conceptos a explorar.
- 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).

- Investigar las fuentes de datos y analizar dónde se encuentra toda la información necesaria para cumplir con los objetivos.
- Refinar el objetivo general de acuerdo a los datos disponibles.

## Procesamiento de Datos

- Analizar las formas normales en que se encuentran las tablas de **Establecimientos Educativos** y **Establecimientos Productivos**. Justificar de manera concisa.
- Revisar la calidad de los datos de las tablas **Establecimientos Educativos** y **Establecimientos Productivos**. Para este trabajo se pide identificar y describir al menos un problema de calidad distinto en cada una de ellas. Para ello, elaborar métricas que permitan cuantificar la gravedad de cada problema utilizando la técnica **GQM (Goal, Question, Metric)**. Para cada problema identificado, indicar:
  - El **atributo de calidad** afectado.
  - Si se trata de un problema de **modelo**, de **instancia** u otro tipo.
  - Una **medida concreta** de la magnitud del problema, basada en GQM (deben explicitar el **objetivo**, las **preguntas** y las **métricas**).
  - Los **valores obtenidos** en las métricas propuestas.
  - Un **diagnóstico** del problema y posibles acciones de mejora. En caso de que consideren útil aplicar alguna corrección, pueden implementarla y reportar los resultados de las métricas luego de dicha mejora. Si no corresponde realizar cambios (por no ser necesarios o no ser viables), no es obligatorio hacerlo.
- Generar un Diagrama Entidad-Relación (DER) que permita modelar de manera conceptual solamente los datos necesarios para resolver los problemas y actividades planteados en el presente trabajo práctico.
- Definir los esquemas correspondientes al modelo relacional del DER del punto anterior. Todos ellos deben estar en **3FN**. Para cada uno de ellos definir:
  - Clave primaria (PK)
  - Dependencias funcionales (DF). En lo posible, se desea que no escriban la totalidad de ellas sino un conjunto minimal de las mismas
  - Claves foráneas (Foreign keys)
- Importar los datos (ya limpios) a los esquemas creados a partir del DER. 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 (después de cada ítem se muestran ejemplos de resultados de consultas; en ellos no necesariamente han sido tenidos en cuenta los datos reales de la fuente de datos):
  - i) Para cada departamento informar la provincia, el nombre del departamento, la cantidad de Establecimientos Educativos (EE) de cada nivel educativo, considerando solamente la modalidad común, y la cantidad de habitantes con edad correspondiente al nivel educativos listado. El orden del reporte debe

ser alfabético por provincia y dentro de las provincias descendente por cantidad de escuelas primarias.

Ejemplo:

| 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                 |
| ...          | ...          | ...      | ...              | ...       | ...                | ...         | ...                  |

- ii) Para cada departamento informar la provincia, el nombre del departamento y la cantidad de empleados totales en ese departamento, para el año 2022. El orden del reporte debe ser alfabético por provincia y, dentro de las provincias, descendente por cantidad de empleados.

Ejemplo:

| Provincia    | Departamento | Cantidad total de empleados en 2022 |
|--------------|--------------|-------------------------------------|
| Buenos Aires | Avellaneda   | 25.841                              |
| Buenos Aires | La Plata     | 21.453                              |
| ...          | ...          | ...                                 |

- iii) Para cada departamento, indicar provincia, nombre del departamento, cantidad de empresas exportadoras que emplean mujeres (en **2022**), cantidad de EE (de modalidad común) y población total. Ordenar por cantidad de EE descendente, cantidad de empresas exportadoras descendente, nombre de provincia ascendente y nombre de departamento ascendente. No omitir departamentos sin EE o exportadoras con empleo femenino.

Ejemplo:

| Provincia | Departamento | Cant_Expo_Mujeres | Cant_EE | Población |
|-----------|--------------|-------------------|---------|-----------|
| Córdoba   | CAPITAL      | 115               | 2512    | 1498060   |
| Santa Fe  | Rosario      | 42                | 3140    | 1337958   |
| ...       | ...          | ...               | ...     | ...       |

- iv) Según los datos de 2022, para cada departamento que tenga una cantidad de empleados mayor que el promedio de los puestos de trabajo de los departamentos de la misma provincia, indicar: provincia, nombre del departamento, los primeros tres dígitos del CLAE6 que más empleos genera, (si no tiene 6 dígitos, agregar un 0 a la izquierda) y la cantidad de empleos en ese rubro.

Ejemplo:

| Provincia | Departamento | CLAE3 | Cant. empleos |
|-----------|--------------|-------|---------------|
| Córdoba   | CAPITAL      | 012   | 684           |
| Santa Fe  | Rosario      | 215   | 991           |
| ...       | ...          | ...   |               |

- Mostrar, utilizando herramientas de visualización, la siguiente información:
  - i) Cantidad de empleados por provincia, para 2022. Mostrarlos ordenados de manera decreciente por dicha cantidad.
  - ii) Graficar la cantidad de establecimientos educativos (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.
  - iii) 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.
  - iv) Relación entre la cantidad de empleados cada mil habitantes (para 2022) y de EE cada mil habitantes por departamento.
  - v) Las 5 actividades (CLAE6) con mayor y menor proporción (respectivamente) de empleadas mujeres, para 2022. Incluir en el gráfico la proporción promedio de empleo femenino.

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 "... es explorar si en Argentina existe una relación entre la presencia de establecimientos educativos, y el desarrollo de actividades productivas". 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/1ge8ESPJFSqIFG0Uaid\\_9cyl1xKsRRjScqZ9Z0NS89IY/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1ge8ESPJFSqIFG0Uaid_9cyl1xKsRRjScqZ9Z0NS89IY/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, nombres de los miembros del grupo y cuatrimestre y año.
- **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 (siempre 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/1DxCdCxwmbN14UH2UIHsApXV9U1ieJamVO4OfxWofiCg/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 miércoles **15 de Octubre a las 23:50 hs**. El día jueves 16 de Octubre, antes de las 12:00, 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.