


Trabajo Práctico 1: escuelas y bibliotecas

Enunciado

Los docentes de la materia Laboratorio de Datos se han encontrado con una fuente de datos abiertos correspondientes a los Establecimientos Educativos y las Bibliotecas Populares de la República Argentina. En particular, están interesados en saber si existe alguna relación entre la cantidad de establecimientos educativos en cada una de las provincias y la cantidad de Bibliotecas Populares. A continuación se detallan los datos con los que se cuenta.

Fuentes de datos

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. **Bibliotecas Populares (BP).** Padrón de Bibliotecas Populares.
https://datos.gob.ar/dataset/cultura-mapa-cultural-espacios-culturales/archivo/cultura_01c6c048-dbeb-44e0-8efa-6944f73715d7
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 Generales

Queremos evaluar las capacidades de manejo de datos y visualización por parte de cada uno de los alumnos. Para esto, se espera que los estudiantes cumplan con los siguientes puntos:

- Plantear claramente el objetivo de la investigación.
- 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 en tiempo y forma.
- Responder preguntas durante el coloquio.

Ejercicios

Primeros Pasos

- 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.
- Estudiar las fuentes de datos y analizar dónde se encuentra toda la información necesaria para cumplir con los objetivos.

Procesamiento de Datos

- 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 (**no olvidar ninguno de estos puntos**) 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)
- Analizar las formas normales en que se encuentran las tablas de **Establecimientos Educativos** y **Bibliotecas Populares**. Justificar de manera concisa.
- Revisar la calidad de los datos de las siguientes dos tablas originales: **Establecimientos Educativos** y **Bibliotecas Populares**. 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.
- 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.
- Sugerimos que resuelvan esta sección en Python nativo y/o Pandas.

Consultas SQL

- A partir de los esquemas con datos del punto anterior, generar los siguientes reportes **utilizando sólo consultas SQL**:
 - i) Para cada departamento informar la provincia, el nombre del departamento, la cantidad de EE de cada nivel educativo, considerando solamente la modalidad común, y la 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 Jardin	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.

- ii) Para cada departamento informar la provincia, el nombre del departamento y la cantidad de BP fundadas desde 1950. El orden del reporte debe ser alfabético por provincia y dentro de las provincias, descendente por cantidad de BP de dicha capacidad.

Provincia	Departamento	Cantidad de BP fundadas desde 1950
Buenos Aires	Avellaneda	8
Buenos Aires	La Plata	5
...

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

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

Provincia	Departamento	Cant_EE	Cant_BP	Población
Córdoba	CAPITAL	1415	30	1498060
Santa Fe	Rosario	1263	36	1337958
...	

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

- iv) Para cada departamento, indicar provincia, el nombre del departamento y qué dominios de mail se usan más para las BP.

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

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

Visualización y análisis de datos

Mostrar, utilizando herramientas de visualización, la siguiente información:

- Cantidad de BP 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 BP cada mil habitantes y de EE cada mil habitantes por departamento.

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 BP 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.

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**. El informe debe contener:

- **Carátula**, con el nombre de la materia y del TP del que se trata, fecha, nombre del grupo y nombres de los miembros del grupo.
- **Sección Resumen**, que sintetice el objetivo, el trabajo realizado y las conclusiones a las que arribaron.
- **Sección Introducción**, en donde se presente el problema a resolver, el objetivo general, las actividades a realizar para alcanzar dicho objetivo, 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 SQL (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**, donde describan qué conocimiento produjeron en relación al objetivo planteado.



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, interlineado simple, márgenes de 1 cm). Se evaluará que el documento 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 que en el informe), 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

Pueden utilizar esta herramienta para evaluar la calidad del trabajo a medida que lo realizan y antes de enviarlo: [TP1-Autoevaluacion](#). Realicen lo siguiente:

- Copiar la siguiente planilla de autoevaluación (una sola a nivel grupal) a una carpeta personal.
- Leerla **antes** de comenzar a trabajar.
- Completarla al finalizar el trabajo (antes de enviar).
- Descargarla como pdf y agregarla al envío.

El trabajo práctico (documento con el informe, código, ambos directorios con los archivos de datos, y el documento de autoevaluación) deberá subirse al campus en formato .zip (lo subirá el responsable del grupo encargado del envío). El nombre del archivo deberá ser *TP1-Apellidos_En_Orden_Alfabético_Con_Guión_Bajo.zip*. La fecha límite para subir el TP es el miércoles **21 de mayo a las 23:50 hs**.