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 han encontrado fuentes de datos abiertos vinculados a la población de la República Argentina y a su salud. En particular, están interesados en saber si existe cierta relación entre las enfermedades que llevan a la muerte, variables demográficas y de acceso a la salud. A continuación se detallan los datos con los que se cuenta.

Datos

Fuentes

1. Población según censo 2010: cantidad de habitantes en base a provincia, edad, sexo, tipo de cobertura de salud.
   Obtenido por cruce múltiple en la categoría de viviendas particulares, y tomando como variables "cobertura de salud", "edad" y "sexo registrado al nacer"; considerando la provincia como corte de área.
   https://redatam.indec.gob.ar/binarg/RpWebEngine.exe/Portal?BASE=CPV2010A&lang=ESP
2. Población según censo 2022: cantidad de habitantes en base a provincia, edad, sexo, tipo de cobertura de salud.
   Obtenido por cruce múltiple en la categoría de viviendas particulares, y tomando como variables "cobertura de salud", "edad" y "sexo registrado al nacer"; considerando la provincia como corte de área.
   https://redatam.indec.gob.ar/binarg/RpWebEngine.exe/Portal?BASE=CPV2022&lang=ESP
3. Defunciones ocurridas y registradas en la República Argentina entre los años 2005-2022 en base a provincia, departamento, edad y sexo:
   Nota: Dado que el sitio https://datos.gob.ar/ presenta escaso mantenimiento y varios de sus enlaces de descarga se encuentran desactualizados o fuera de servicio, se optó por utilizar un enlace alternativo mantenido por la comunidad, disponible en:
   https://datos.nulo.lol/dump/https%3A%2F%2Fs3.g.s4.mega.io%2Fjosaimmknzzgy2h5johxxsmj7mfi5olydqhbd%2Fdatos-argentina%2Fdump-2025-02-03/datos.salud.gob.ar_data.json/dataset/27c588e8-43d0-411a-a40c-7ecc563c2c9f
   Aclaraciones: Notar que el código de defunción está compuesto por un primer caracter alfabético que representa la categoría de defunción y el mismo es seguido por dos caracteres numéricos que representan una subcategoría, esta última está detallada en la columna "clasificación ".

4. Establecimientos de salud asentados en el registro federal (REFES) Abril 2022 en base a provincia y departamento:
   Nota: Dado que el sitio https://datos.gob.ar/ presenta escaso mantenimiento y varios de sus enlaces de descarga se encuentran desactualizados o fuera de servicio, se optó por utilizar un enlace alternativo mantenido por la comunidad, disponible en: https://datos.nulo.lol/dump/https%3A%2F%2Fs3.g.s4.mega.io%2Fjosaimmknzzgy2h5johxxsmj7mfi5olydqhbd%2Fdatos-argentina%2Fdump-2025-02-03/datos.salud.gob.ar_data.json/dataset/336cf4d9-447a-44c4-8e34-0ba1fc293d55

Objetivos Generales

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 en tiempo y forma.

Ejercicios

Primeros Pasos

- Descargar los datos de la carpeta del campus. 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.
- Investigar 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).
  - Claves foráneas (Foreign keys)
- Analizar las formas normales en que se encuentran las tablas de Establecimientos de salud y Defunciones. Justificar de manera concisa.
- Revisar la calidad de los datos de las siguientes dos tablas originales: Establecimientos de salud y Defunciones. 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 guardarse en un archivo .csv 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, es decir los del modelo relacional construido, generar los siguientes reportes **utilizando sólo consultas SQL**:

  i) Cobertura de salud

  Para cada provincia, informar la cantidad de habitantes con y sin cobertura de salud, desagregada por grupo etario, para los años 2010 y 2022.

Provincia	Grupo etario	Habitantes con cobertura en 2010	Habitantes sin cobertura en 2010	Habitantes con cobertura en 2022	Habitantes sin cobertura en 2022
Buenos Aires	0 a 14	857347	163119	921253	160242
Buenos Aires	15 a 34	531841	234218	623514	216214
…	…	…			

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

ii) Establecimientos de salud con terapia intensiva
Para cada provincia, informar la cantidad de establecimientos de salud que cuentan con terapia intensiva, agrupados por tipo de financiamiento (estatal o privado).

iii) Causas de muerte
Para cada grupo etario y sexo, informar las cinco categorías de defunción más frecuentes y las cinco categorías de defunción menos frecuentes. Ordenar en forma ascendente por grupo etario y por sexo.

iv) Tasa de mortalidad por provincia
Para cada provincia y grupo etario, calcular la tasa de mortalidad correspondiente al año 2022, normalizada cada 1000 habitantes.

v) Cambios en las causas de defunción
Para cada categoría de defunción, calcular la diferencia en la cantidad de defunciones entre los años 2010 y 2022. Ordenar de manera descendente por el valor de la diferencia.

- Mostrar, utilizando herramientas de visualización, la siguiente información:

i) Cantidad de habitantes por provincia
Realizar un gráfico que muestre la cantidad de habitantes por provincia para los años 2010 y 2022.

ii) Defunciones por categoría a lo largo del tiempo
Realizar un gráfico que muestre la cantidad total de defunciones por categoría en función del tiempo.
Dado que existen muchas categorías, se recomienda agruparlas o utilizar más de un gráfico para facilitar la lectura.

iii) Tasa de mortalidad por provincia en 2022
Realizar un gráfico que muestre la tasa de mortalidad por provincia en el año 2022, ordenada de menor a mayor.

Complementar el gráfico anterior de la siguiente manera: Realizar un gráfico que muestre la tasa de mortalidad por provincia en el año 2022, diferenciada según otra variable a elección (por ejemplo: grupo etario, sexo o categoría de defunción).
Indicar brevemente qué variable fue elegida y por qué.

iv) Defunciones por grupo etario y sexo
Realizar un gráfico que muestre la cantidad de defunciones en el año 2022 por grupo etario, diferenciadas por sexo.

Las defunciones deben estar normalizadas por la cantidad de habitantes de cada grupo etario.

v) **Distribución de establecimientos de salud**
Para cada provincia, realizar un boxplot que muestre la distribución de la cantidad de establecimientos de salud por departamento.
Utilizar este gráfico para comparar cómo se distribuyen los establecimientos dentro de cada provincia.

vi) **Gráfico a elección**
Realizar un gráfico adicional que explore alguna relación, comparación o patrón que haya surgido durante el análisis de los datos y que resulte de interés para el grupo.
Explicar brevemente:
- *qué se decidió mostrar,*
- *por qué se eligió ese gráfico,*
- *y qué se puede observar a partir del mismo.*

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 las enfermedades que llevan a la muerte, variables demográficas y de acceso a la salud, y si hubo variaciones a lo largo del tiempo". 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):
TP-01-Grupos

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 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.** Retomar el objetivo general del trabajo y discutir, a partir de los análisis realizados, si se observan relaciones entre las causas de muerte, las variables demográficas y el acceso al sistema de salud, y si dichas relaciones presentan cambios entre 2010 y 2022. En caso de identificar limitaciones en los datos o en el análisis, mencionarlas brevemente.

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:

a. Copiar la siguiente planilla de autoevaluación (una sola a nivel grupal) a una carpeta personal:
   TP-01-Autoevaluacion
b. Completarla.
c. Descargarla como pdf y agregarla al envío virtual y en papel.
   Aclaraciones:
   – La autoevaluación no será utilizada para la asignación de la nota, sino como una instancia de reflexión sobre el trabajo realizado.
   – Es importante que completen la columna “Comentarios”, incorporando observaciones, aclaraciones o reflexiones que consideren relevantes.

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 martes 17 de febrero a las 23:50 hs. El día miércoles 18, 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 saltear 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.