Laboratorio de datos
Trabajo práctico 01

2do cuatrimestre 2025

Nombre de Grupo: Labo de Datos

Integrantes:

- Nombre y Apellido : Gabriel Duham Lopez
  Legajo : 615/23
  Mail : gabriellopezdu@gmail.com

- Nombre y Apellido : Benjamin Francisco Vales
  Legajo : 156/23
  Mail : benja.vales@gmail.com

- Nombre y Apellido : Diego Pariona Escalante
  Legajo : 559/24
  Mail : parionadiego51tm@gmail.com

Resumen

Introducción

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, utilizando como base de datos (DB) datos abiertos correspondientes a los Establecimientos Educativos, Establecimientos Productivos y de Población de la República Argentina

Para esto, en primer lugar se hace un procesamiento de los datos de estas fuentes, determinando la calidad de los mismos mediante el análisis de sus formas normales y la realización y evaluación de métricas de calidad por el método Goal-Question-Metric (GQM). En función de los resultados, se toman medidas para mejorar la calidad, reducir incoherencias y eliminar información inútil de las DB, conservando sólo la información necesaria para resolver el problema. Luego, en función de estas DB, se diseña un modelo de datos gráficamente mediante un diagrama entidad-relación (DER) que explicite las relaciones lógicas entre las distintas entidades del modelo, y se representa este modelo como un modelo relacional. En este proceso una serie de decisiones son tomadas a fines de simplificar el uso e interpretación de las DB, explicadas en el apartado "Decisiones tomadas"

A continuación, se implementa este modelo en Python utilizando dataframes de Pandas para representar las entidades y relaciones. Sobre estos, se realiza una serie de consultas SQL que nos permiten analizar estos datos bajo ciertos criterios, detallados en el apartado "Análisis de datos". Además. Se incluyen algunos gráficos que brindan más información acerca del vínculo entre los datos.

Por último, se dan las Conclusiones del trabajo y se analiza si con la información obtenida, basta para responder el objetivo del trabajo, y, si no, enumerar qué información sería necesario incluir que no esté incluida y mostrar los resultados.

Procesamiento de datos

En esta sección se realiza un análisis de la calidad de las bases de datos originales, y en consecuencia se procesan para obtener un modelo óptimo a fines de la resolución del objetivo.

La sección se encuentra dividida por las fuentes de datos originales, y cada apartado cuenta con:

- Resumen del diseño de la tabla
- Análisis de formas normales (FN)
- Análisis de calidad: GQM y observaciones
- Procesos de limpieza y combinación
- Representación en el modelo (DER)

Luego, se incluye el DER con todas las partes correspondientes y su modelo relacional asociado. Finalmente, se da una descripción de importación de tablas del modelo relacional.

GQM(Goal, Question, Metric)

1. EP anio

- Atributo afectado : anio.
- Modelo o instancia u otro : instancia.
- Goal (objetivo) : el valor de anio debería ser 2022.
- Pregunta : ¿Cuál es la proporción de registros que tienen el campo anio distinto a 2022?
- Métrica : cantidad de registros con el valor 2021 en anio sobre la cantidad de registros totales.

- Valores obtenidos : 0.49326643447838153 es la proporción obtenida.
- Diagnóstico : casi la mitad de los registros con el año 2021, la solución a tomar sería hacer una consulta SQL donde solo me quedaría con los registros que quiero.
- Luego de la consulta el problema es solucionado, no hay ningún registro con el año 2021. Ver dataframe solucion_anio.

2. EE Teléfono

- Atributo afectado: Teléfono.
- Modelo o instancia u otro : instancia.
- Goal (objetivo) : el campo Teléfono debería estar completo.
- Pregunta : ¿Cuál es la proporción de registros que tienen el campo teléfono vacío?
- Métrica : cantidad de campos de teléfonos vacíos sobre la cantidad de registros totales (los campos "vacíos" están "llenas" con ceros 0, y espacios " " en el excel pero cuando pasamos a DataFrame los “hace” null).
- Valores obtenidos : 0.02042962200462903 es la proporción obtenida con la métrica antes mencionada.
- Diagnóstico : la cantidad de registros con el campo Teléfono vacío es mínima pero no se puede negar su existencia y eso afecta a la calidad de datos de este excel.

3. EE Domicilio

- Atributo afectado: Domicilio.
- Modelo o instancia u otro : instancia.
- Goal (objetivo): el nombre del domicilio debería estar completo.
- Pregunta : ¿Cuál es la proporción de registros que tienen el campo Domicilio vacío?
- Métrica : cantidad de domicilios vacíos sobre la cantidad de registros totales (los campos "vacíos" están "llenos" con espacios " " en el excel pero cuando pasamos a DataFrame los “hace” null).
- Valores obtenidos : 0.002889196929557274 es la proporción obtenida con la métrica dicha.

- Diagnóstico : no tener todos los domicilios llenos de los establecimientos educativos habla de lo mal que esta esté excel.

EE Mail

- Atributo afectado: Mail.
- Modelo o instancia u otro : instancia.
- Goal (objetivo): el campo de Mail debería estar completo.
- Pregunta : ¿Cuál es la proporción de registros que tienen el campo Mail vacío?
- Métrica : cantidad de registros con el campo Mail vacío sobre la cantidad de registros totales (los campos "vacíos" están "llenos" con espacios " " en el excel pero cuando pasamos a DataFrame los “hace” null).
- Valores obtenidos : La proporción que tenemos como resultado es 0.03843737073606633
- Diagnóstico : No tener campos llenos con los mails no es vital para esta tabla o por lo menos para nuestro objetivo, pero este defecto si toma importancia cuando hablamos de calidad de datos.

Establecimientos_educativos:

El archivo (2022_padron_oficial_establecimientos_educativos.xlsx) cuenta con una introducción y 7 categorías, divididas en subcategorías (columnas). Las filas representan los establecimientos educativos en particular, distinguidos por un código (*Cueanexo*) dentro de la primera categoría. La primera categoría brinda información sobre el establecimiento (Teléfono, mail, dirección, etc.), una columna por dato. La segunda sobre a qué modalidad pertenece, y las categorías consiguientes sobre el nivel al que corresponde, divididos por su modalidad. Un establecimiento tendrá un 1 en su categoría de *modalidad* y la columna con su modalidad correspondiente, caso contrario, la celda está vacía.

En primera instancia se acordó sólo tomar los datos respecto a las escuelas de modalidad “*Común*", por lo que descartamos toda información relativa a las otras modalidades (“*especial”*, “*adultos”*, “*artística”*, “*hospitalaria”*, “*intercultural”* y “*encierro”*).

FN: No es F1 porque 1) tiene relaciones dentro de la relación (Por ej: La sección “modalidad” es una relación) y 2) hay atributos que no son atómicos (Por ej: teléfono)

GQM

Así, es válido concluir que la calidad de esta fuente es relativamente pobre en perspectiva del trabajo que luego se debe realizar. Por esto, realizamos las siguientes acciones:

Dentro de la categoría común, sólo se conservan los datos respectivos al departamento y el nivel de cada institución educativa, y descartamos el nombre, sector, ambito, domicilio, C.P, código de área, teléfono, código de localidad, localidad y mail, a fines de eliminar información innecesaria y/o redundante para el trabajo, y mantener la base de datos lo menos cargada posible.

Establecimientos_productivos

En esta fuente se encuentran los datos de 2021 y 2022 de la cantidad de empleadas y empleados por actividad productiva (distinguidas por Clae6, Clae2 y letra) por departamento y provincia (con identificadores), y también de cuantas empresas y empresas exportadoras las/os emplean.

FN: Esta tabla cumple la condición de 1FN porque no tiene ningún atributo compuesto, pero no cumple 2FN ya que el atributo provincia depende de provincia_id, y a la vez provincia_id no es clave.

GQM

En general, la tabla no tiene las mejores características para nuestro objetivo, por lo tanto:

La información se filtró por el año que vamos a analizar (2022). En este marco, la información relevante por cada departamento en función de la actividad realizada y el sexo de los/as empleados/as es la provincia, la Clae6 de la actividad, la cantidad de empleados/as de ese sexo, la cantidad de establecimientos empleadores y la cantidad de empresas exportadoras empleadoras, y descartamos el id del departamento, el id provincial, y la clae2 y letra de la actividad, ya que son datos irrelevantes para el trabajo, y/o deducibles de la información conservada.

SECCIÓN DER

A continuación, mostramos detalladamente el DER creado con la descripción de entidades, atributos, y de más.

- ## Entidad Departamento

Definimos que esta relación es el centro de las entidades, relacionada tanto a Población, Establecimientos_Educativos, y Establecimientos_Productivos. Tendrá como atributos los siguientes: Provincia, y Nombre, ambos atributos clave.

- ## Entidad Población

Población es una entidad débil ya que depende totalmente de Departamento y no tiene ningún atributo clave[1]. Esta relación tiene como atributos, Cantidad_habitantes, Población_jardin, Población_primario, Población_secundaria.

- ## Entidad Establecimientos_Educativos

Establecimientos_Educativos es otra entidad débil en la que vamos a tomar como atributos Cantidad_jardines, Cantidad_primarios, Cantidad_secundarios, Cantidad_EE (refiriéndonos con EE a Establecimientos Educativos).

[1] Ramez Elmasri y Shamkant B. Navathe, *Fundamentos de Sistemas de Bases de Datos* (5ta Ed., Pearson, 2007), 67-68.

- Entidad Establecimientos_Productivos

Al igual que las ultimas dos entidades Establecimientos_Productivos tambien es débil, y la conformamos con los atributos, Empleados, Establecimientos, Empresas_Exportadoras, Sexo, Clae6.

- Relación “tiene_EE”:

Un valor de Establecimientos Educativos debe estar (|) en solo un Departamento (|).
Un Departamento puede o no tener (o) solamente un valor de Establecimientos Educativos asociado (|).

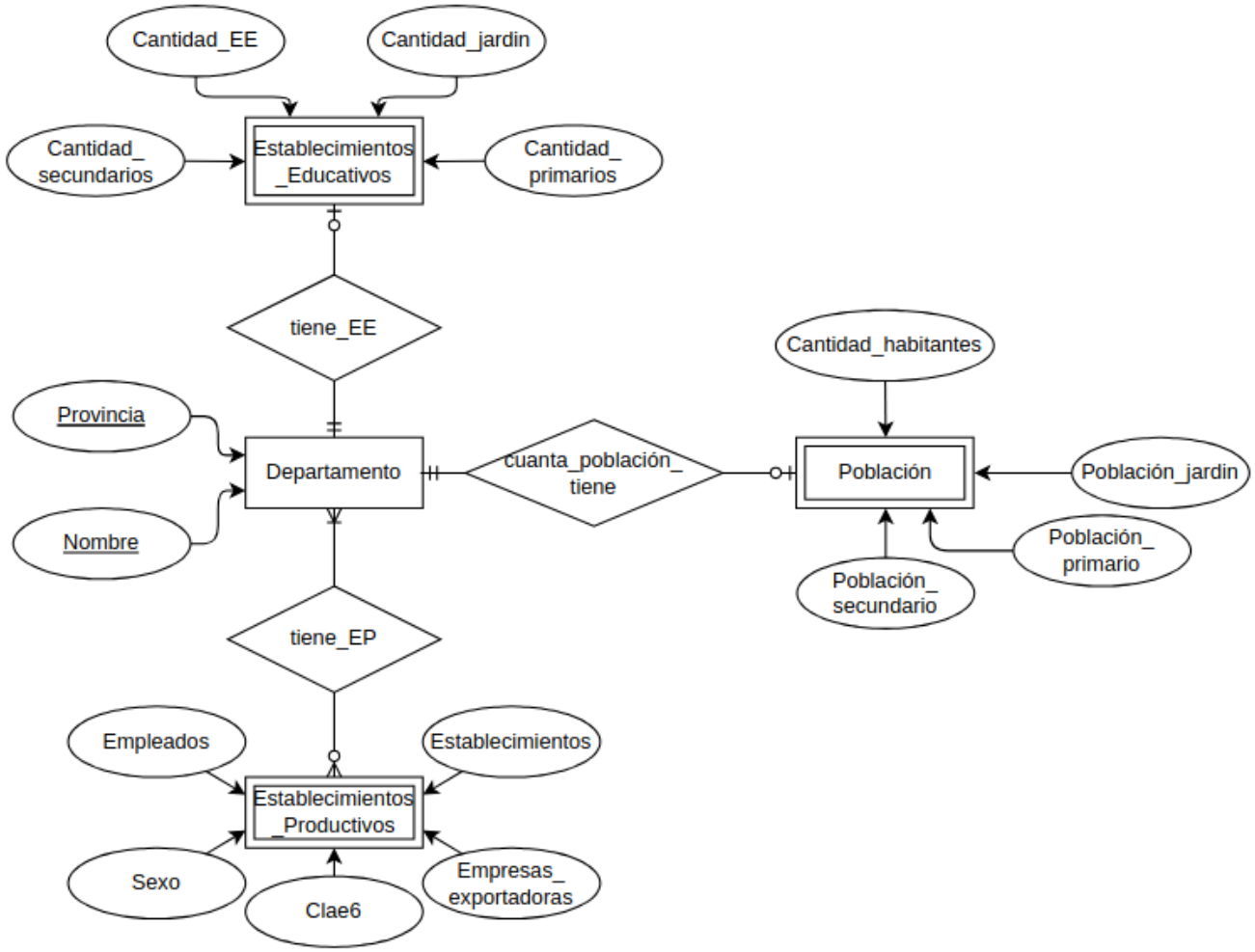
- Relación “tiene_EP”:

Un valor de Establecimientos_Productivos debe (|) estar estar asociado a un valor de Departamento, y puede estar a más de uno (<).
Un Departamento puede o no tener (o) varios valores de Establecimientos_Productivos asociados (<).

- Relación “cuanta_poblacion_tiene”:

Un Departamento puede o no tener población (o) y si tiene, debe ser solo una (|).
Una Población debe (|) estar asociada a un solo Departamento (|).

DER



Notamos que es un buen recurso hacer Entidades Débiles en nuestro DER, ya que al todas las entidades estar relacionadas a Departamento podíamos seleccionar esta estructura para un futuro mejor modelado de los datos.

Sección Mapeo a Modelo Relacional

Modelo Relacional

- Departamento(Nombre, Provincia).

La combinación de ambos atributos forman la clave de Departamento.

- Establecimientos_Educativos(Provincia, Departamento, Cantidad_Jardines, Cantidad_Primarios, Cantidad_Secundarios, Cantidad_EE)

Departamento.Establecimientos_Educativos es un clave, y aparte es Foreign Key de Nombre.Departamento.

- Población (Provincia, Departamento, Población_jardín, Población_primario, Población_secundario, Cantidad_habitantes).

Departamento.Población es una clave, además de ser Foreign Key de Nombre.Departamento.

- Actividades_Productivas(Provincia, Departamento, Clae6, Sexo, Empleados, Establecimientos, Empresas_exportadoras)

Departamento.Actividades_Productivas es Foreign Key de Nombre.Departamento, y los atributos Departamento, Clae6, y Sexo en conjunto son clave.

El hecho de que Departamento y Provincia formen parte de la clave en todas las relaciones se debe a que pueden haber más de un Departamento distinto con el mismo nombre, pero no pueden haber dos departamentos con el mismo nombre en una misma provincia. Razón por la cual la manera unívoca de identificar un Departamento es sabiendo su nombre, y su Provincia.
También vale aclarar que como todas las relaciones que no son Departamento son débiles, su mapeo al Modelo Relacional implica que esas relaciones mapeadas tendrán como Clave Foránea a los atributos Nombre.Departamento y Provincia.Departamento.

Dependencias Funcionales y Formas Normales

- Departamento(Nombre, Provincia)

Aquí la combinación de los dos atributos es la clave. Podemos garantizar que es primera forma normal porque no hay atributos compuestos, que es 2FN porque al no haber atributos no primos entonces no pueden haber dependencias funcionales parciales, y la relación es 3FN porque como solo estamos hablando de estos dos atributos clave en su conjunto no hay dependencias transitivas.

El tipo de dato de ambos atributos es String.

- Establecimientos_Educativos(Provincia, Departamento, Cantidad_Jardines, Cantidad_Primarios, Cantidad_Secundarios, Cantidad_EE)

Esta relación es 1FN porque no tiene atributos compuestos, 2FN porque todos los atributos no primos dependen de la clave en su conjunto, y es 3FN porque no hay dependencias transitivas, ya que Cantidad_EE no se compone de la suma de Cantidad_Jardines, Cantidad_Primarios, y Cantidad_Secundarios, sino que estos últimos tres atributos son la cantidad de Establecimientos que tienen cada modalidad. Por lo que la suma de estos tres atributos generalmente va a dar un número mayor que el valor de Cantidad_EE por cada Departamento, salvo que no hayan establecimientos con más de una modalidad en un Departamento en específico.

El tipo de dato de Provincia y Departamento es String, mientras que los demás atributos son de tipo Int.

Así se describe su dependencia funcional:

{Provincia, Departamento} → {Cantidad_Jardines, Cantidad_Primarios, Cantidad_Secundarios, Cantidad_EE}

- Población (Provincia, Departamento, Población_jardín, Población_primario, Población_secundario, Población_total).

Esta relación es 1FN porque no tiene atributos compuestos, 2FN porque los atributos no primos dependen totalmente de la clave, y es 3FN porqué Población_total no depende transitivamente de Población_jardín, Población_primario, Población_secundario, ni tampoco hay posibles dependencias transitivas.

Provincia y Departamento tienen tipo de datos String, mientras que los demas atributos son de tipo Int.

Parecido que en la anterior relación descrita, estas son sus dependencias funcionales: {Provincia, Departamento} → {Población_jardín, Población_primario, Población_secundario, Población_total}

- Actividades_Productivas(Provincia, Departamento, Clae6, Sexo, Empleados, Establecimientos, Empresas_exportadoras)

Esta relación cumple que es 1FN debido a que no tiene atributos compuestos, es 2FN porque los atributos no primos dependen funcionalmente de la clave completa, y es 3FN porque no se ve ninguna dependencia transitiva.

Provincia y Departamento tienen tipo de datos String, mientras que los demás atributos son de tipo Int.

Así se denota su dependencia funcional:
{Provincia, Departamento, Clae6, Sexo} → {Empleados, Establecimientos, Empresas_exportadoras}

Mostrado nuestro Modelo Relacional, la razón por la cual hicimos a las demás Entidades que no son Departamento débiles, fue en parte para modelar nuestra estructura de los datos, así quedan limpios con las cantidades de población, y Establecimientos Educativos. Con respecto a Establecimientos Productivos, no lo vimos conveniente, así que en ese caso seleccionamos simplemente los atributos que nos sirven para nuestro objetivo planteado.

La limpieza de datos se encuentra bien detallada en el archivo de Python que se encuentra en el directorio llamado “TP01”, y los datos utilizados provienen del INDEC, del Ministerio De Capital Humano, y del Ministerio de Economía de la Nación.

Decisiones tomadas

En esta sección se explica la toma de decisiones realizada en función de las bases de datos que utilizamos como fuente para realizar el trabajo. Cada apartado enuncia a qué

base de datos se refiere, y dentro se detallan las decisiones, se explica el razonamiento detrás de cada decisión, y qué parte de los datos se vio afectada.

Respecto a padron_poblacion:

En primera instancia, se considera que todo habitante pertenece a un y solo a un departamento, pues la base de datos sólo tiene esta información.

Para cada departamento, la población fue separada en 3 sectores; habitantes elegibles para jardín de infantes, para escuela primaria y para escuela secundaria, ya que las consultas que vamos a realizar sólo precisan esta información.

Definimos los sectores en función de la edad, siendo los habitantes de edad mayor o igual a 3 años y menor o igual a 5 años los elegibles para jardín de infantes, los de edad mayor o igual a 6 años y menor o igual a 12 años los elegibles para escuela primaria, y aquellos de edad mayor o igual a 13 y menor o igual a 18 años los elegibles para escuela secundaria.

Respecto a Departamentos

Vimos que con las mayúsculas y los acentos, los departamentos pueden tener muchas formas de referirse al mismo. Así que decidimos que los departamentos de Establecimientos Productivos sean los que durante el trabajo tengamos como referencia, es decir, las diferentes combinaciones o formas de escribir que detectamos en EE y Población, las modificamos y las reescribimos tal cual aparecen en EP.

Notamos tmb que hay un departamento que tiene EE y no EP, y tmb otro que está en EP y no en EE, los cuales son: Antártida Argentina, Tierra del Fuego está en EE y Tolhuin Tierra del Fuego está en EP.

Análisis de datos

Todas estas consultas SQL se encuentran en el directorio “/TP01-LabodeDatos/Anexo”

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 educativo listado. El orden del reporte debe ser alfabético por provincia y dentro de las provincias descendente por cantidad de escuelas primarias. Acá entra gráfico ii) y iii).

II. Datos de los nombres y cantidad de empleados por departamento por provincia, ordenado alfabéticamente por provincia, y luego por cantidad de empleados, junto con un gráfico de la cantidad de empleados por provincia. Acá entra el gráfico i).

Provincia	Departamento	Cantidad total de empleados en 2022
BUENOS AIRES	LA MATANZA	127962
BUENOS AIRES	VICENTE LOPEZ	120347
BUENOS AIRES	GENERAL PUEYRREDON	118660
...	...	...

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. Acá entra gráfico V).

IV. Datos de los departamentos cuya cantidad de empleados sea mayor que el promedio de los puestos de trabajo de los departamentos de la misma provincia, con los primeros tres dígitos del CLAE6 que más empleos genera, y la cantidad de empleos en ese rubro.

Provincia	Departamento	Clae3	Cant. empleos	
BUENOS AIRES	ALMIRANTE BROWN	851	4248	
BUENOS AIRES	AVELLANEDA	851	3280	
BUENOS AIRES	BAHIA BLANCA	471	1923	
...	...	...	...	

Se puede observar

V. Gráfico iv) no entra en ninguna de los anteriores, vale la pena ponerlo último porque es el más cercano a la conclusión.

Conclusiones

Llegado a este punto, pudimos procesar todos los datos y acomodarlos a nuestro parecer, para que luego podamos trabajar con ellos fluidamente y poder hacer las consultas y los gráficos pedidos. Lamentablemente no llegamos a ninguna conclusión sobre la relación entre establecimientos educativos y establecimientos productivos, ya que no pudimos llegar a hacer los últimos gráficos por tema de tiempo.