INFORME DE ANÁLISIS DE DATOS

ANÁLISIS DE RELACIÓN ENTRE

BIBLIOTECAS PÚBLICAS Y

ESTABLECIMIENTOS EDUCATIVOS

BASADA EN DATOS OFICIALES

Resumen

En este trabajo intentamos hallar una relación entre la cantidad de establecimientos educativos y bibliotecas públicas en las distintas provincias argentinas. A partir de padrones oficiales, evaluamos su calidad y consistencia de datos, sus formas normales y diseñamos un modelo relacional para facilitar el procesamiento. El análisis se basó en la búsqueda de patrones de distribución mediante consultas SQL y visualización de los datos. Concluimos que existe una tendencia positiva entre ambas, aunque factores geopolíticos también influyen en su distribución.

Introducción

En el presente trabajo queremos determinar si existe alguna relación entre la cantidad de establecimientos educativos y bibliotecas públicas en las diferentes provincias de la Argentina. Ver si la oferta de bibliotecas acompaña la cantidad de escuelas parece relevante en estas épocas de revolución digital, ya que cada vez menos personas recurren a la información física.

En principio esto se cumpliría si existiera una proporcionalidad entre las cantidades de ambas instituciones. Para ello, analizamos los datos provenientes de padrones oficiales evaluando diferentes escenarios.

En primer lugar nos familiarizamos con las fuentes de datos originales y analizamos su calidad utilizando la técnica GQM. Luego, estudiamos cuál era la información necesaria de cada tabla para encontrar relaciones subyacentes que cumplan con el objetivo y lo plasmamos en un diagrama conceptual de los datos (DER). Basándonos en este esquema, hicimos una abstracción de los datos originales transformándolos en nuestras tablas de trabajo. Por último, utilizamos la sintaxis de SQL para relacionarlas y visualizar los resultados.

Procesamiento de Datos

Fuentes:
Las tablas de este trabajo fueron adquiridas de fuentes oficiales de padrones argentinos:

- Establecimientos Educativos (EE.) Padrón Oficial de Establecimientos Educativos del año 2022. https://www.argentina.gob.ar/educacion/evaluacion-e-informacion-educativa/padronoficial-de-establecimientos-educativos
- Bibliotecas Populares (BP). Padrón de Bibliotecas Populares. https://datos.gob.ar/dataset/cultura-mapa-cultural-espacios-culturales/archivo/cultura_01c6c048-dbeb-44e0-8efa-6944f73715d7
- Población. Datos de población por Departamento del censo de 2022, estructurado por edad de la población. https://www.indec.gob.ar/indec/web/Nivel4-Tema-2-41-165

Observación: Los únicos archivos que hay que ejecutar son el archivo “Principal.py” y “GQM.py”, y solo se tiene que modificar la variable path del primer archivo para que todo se ejecute correctamente.

Análisis de formas normales:
Vamos a comenzar analizando las formas normales de las tablas originales de “Bibliotecas Públicas” y “Establecimientos Educativos” para entender su funcionamiento y complejidad a la hora de trabajar con ellas.

En el caso de la tabla de “Bibliotecas Públicas” la 1FN se cumple ya que todos sus atributos están compuestos por valores atómicos como cadenas de string o números. La 2FN establece que todos los atributos deben depender de la clave primaria, esto se cumple ya que se establecio una clave única no compuesta (“Nro_Conabip”) la cual le define un único valor a los demás atributos. Para la 3FN no tendrían que existir dependencias funcionales transitivas entre los atributos que no son claves de las entidades. En este caso,esta entidad cuenta con los atributos id_provincia, id_departamento, cod_localidad, lo que genera la siguiente dependencia que viola la 3FN ya que cod_localidad no pertenece a una clave candidata de la entidad: {cod_localidad → id_departamento → id_provincia}

En la tabla de “Establecimientos Educativos” ninguno de los atributos de la entidad está formado por valores no atómicos,y además se estableció nuevamente una clave primaria no compuesta(“Cueanexo”), por lo que todos los atributos dependen de ella. Esto hace que la tabla cumpla la primera y segunda forma normal. En el caso de esta tabla el atributo “Código Postal” hace que no se cumpla la 3FN, ya que en Argentina el código postal depende de la localidad a la que perteneces. Entonces tendrías la dependencia funcional {CP → Localidad} pero “CP” no pertenece a ningún conjunto de claves candidatas.

Análisis de calidad de datos GQM:
Empezando el análisis con la tabla de **Bibliotecas Populares**, procedimos a usar el atributo de calidad **completitud** como criterio de análisis.

GOAL: Los datos sobre todos los atributos de la tabla están completos.

QUESTION: ¿Cuál es la proporción de datos faltantes en cada atributo?
¿Cuál es la proporción de atributos que directamente no tienen ningún dato?

METRIC:
M1) Proporción de datos vacíos en cada atributo, es decir,
(cantidad de registros vacíos del atributo / cantidad total de registros atributo)

M2) proporción de atributos sin datos en toda la tabla, es decir,
(cantidad de atributos sin datos / total de atributos)

VALORES OBTENIDOS CON M1:

Atributo analizado	*Porcentaje faltante de datos*
Mail	46.27%
Telefono	8.86%
cod_tel	8.86%
fecha_fundacion	2.10%
localidad	0.05%
departamento	0.05%

Encontramos también 5 columnas totalmente vacías: informacion_adicional, observacion, piso, subcategoria y web. Lo que nos da un **19% de incompletitud** de columnas la métrica M2

En cuanto al **tipo de problema** para nosotros se mezclan varios causantes lo que complica definir exactamente una única clasificación. Pudimos encontrar razones para incluir problema de software o de modelado de datos.

- Problema de software: Una explicación que encontramos a los de datos faltantes es que no eran campos obligatorios cuando se subieron los datos al sistema. Y por lo tanto muchas personas no los completaron, como puede verse en los mails o los teléfonos.
- Modelado de datos: los atributos que directamente no tienen datos podría considerarse que están mal modelados ya que al ser ambiguos es difícil interpretar qué representa cada campo o si hay diferencias entre ellos y terminan generando mayor complejidad de la que se necesita. Por ejemplo observación, subcategoría, información adicional

Para la tabla de **Establecimientos Educativos (EE)** analizaremos principalmente la **consistencia** de datos .

GOAL: Los datos sobre todos los atributos de la tabla son consistentes en los aspectos que enumeramos a continuación

- Unicidad de identificadores clave: Cada establecimiento debería tener un identificador único (ID, CUE, etc.). No debería haber duplicados para columnas clave que deberían ser únicas.

- Uniformidad de formatos y tipos de datos: Todos los valores en una columna deben seguir el mismo formato o tipo

- Consistencia semántica categórica: Los términos usados deben tener un significado claro y ser usados de forma consistente.
- Dependencias funcionales válidas entre codigo_de_localidad→localidad y codigo_de_localidad→departamento

QUESTION: ¿Cuál es la proporción de datos inconsistentes en cada atributo?

METRIC:
Proporción de datos inconsistentes en cada atributo, es decir:
(cantidad de registros inconsistentes del atributo / cantidad total de registros del atributo)

VALORES OBTENIDOS:

Atributo analizado	*Tipo de consistencia*	*Inconsistencia*
Cuanexo	Unicidad (sin repetidos)	0%
Domicilio	Tipo de Dato	0.13%
C.P.	Tipo de Dato	38.28%
Telefono	Tipo de Dato	44.9%
Mail	Tipo de Dato	0.03%
Sector	Semantica (categorica)	0%
jurisdicción	Semantica (categorica)	0%
Ámbito	Semantica (categorica)	0%
Departamento	Dependencia funcional cod_loc→Dpto	0%
Localidad	Dependencia funcional cod_loc→loc	0%

Aclaraciones: Solamente ilustramos los datos que nos parecieron más importantes. También analizamos la completitud de la tabla EE y al igual que en el caso de BB atributos como mail, teléfono y C.P no se encontraban completos. Creemos que esto se podría explicar por las mismas razones ya nombradas en el análisis de la tabla anterior. Por lo que no merece la pena profundizar en esto más que dejarlo como observación.

Tipo de Problema

El único aspecto que tiene inconsistencias notables es el tipo de dato de los C.P. y de los Teléfonos. Para nosotros esto puede ser el reflejo de un mix de pequeñas imperfecciones en la captura, estandarización y consolidación de datos, más que de grandes fallos estructurales. Por esto consideramos que se trata de un problema quizás más asociado a **procesos.** Por decir algunos ejemplos:

- El 0.13% de inconsistencia en los domicilios podría explicarse por algún descuido de comillas que transformaran un string de números a un tipo de dato numérico o álbumes de datos provenientes de distintos sistemas (bases SQL, CSV, formularios web) pueden tener preajustes de tipos distintos. Que podría darse en lugares con domicilio numérico como el caso de La Plata.

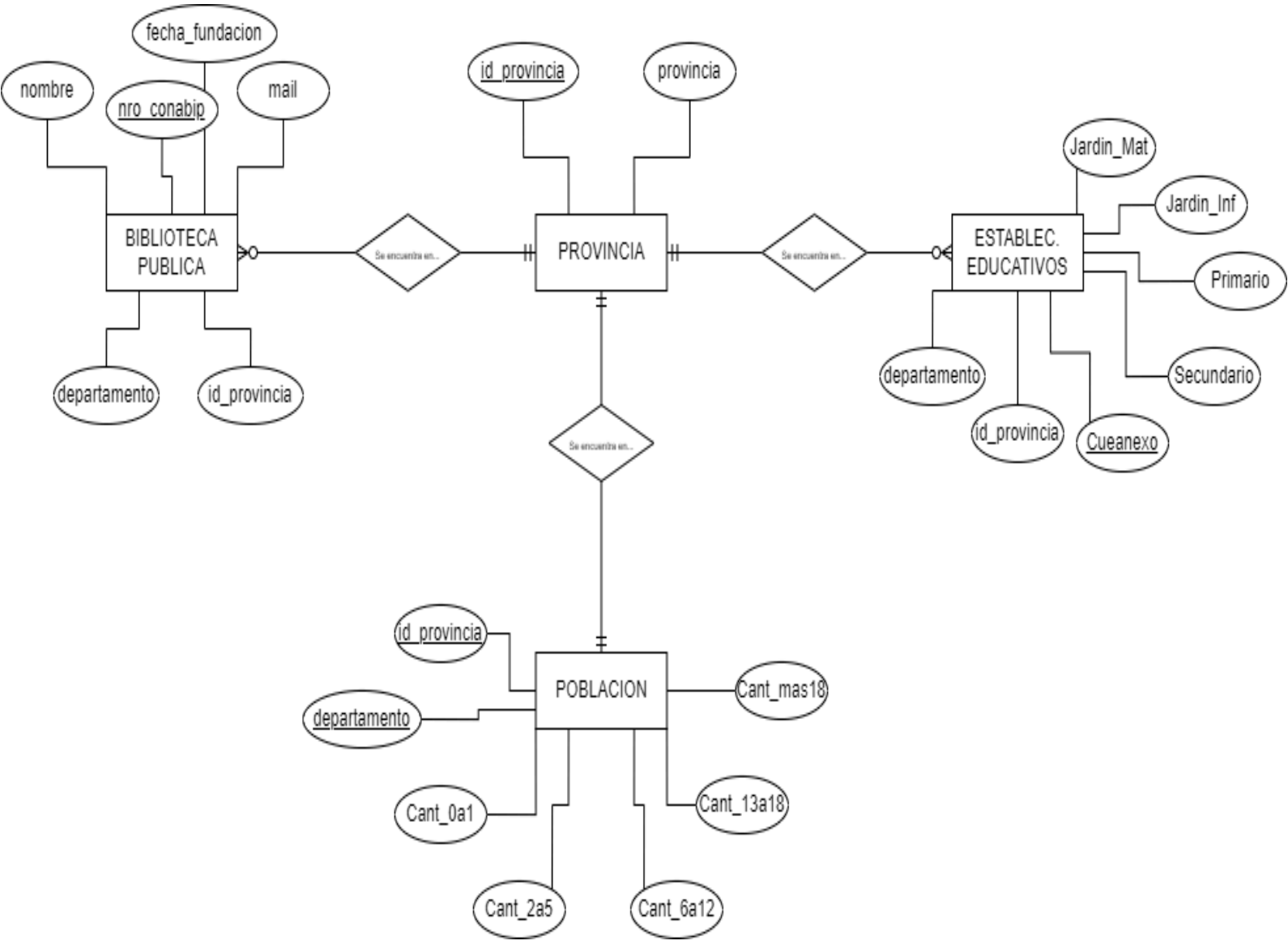
- El caso de CP y teléfonos podría ser por origen de datos heterogéneos. Algunos más o menos específicos que otros o con formatos diferentes y que se unificaron sin tener en cuenta la consistencia. Por ejemplo: algunos padrones que incluyen la letra inicial del CP y otros que no dando como resultados “B1407” y “4752” o valores con varios números telefónicos o con símbolos que no se permiten.

Relación con el Objetivo y posibles acciones de mejora

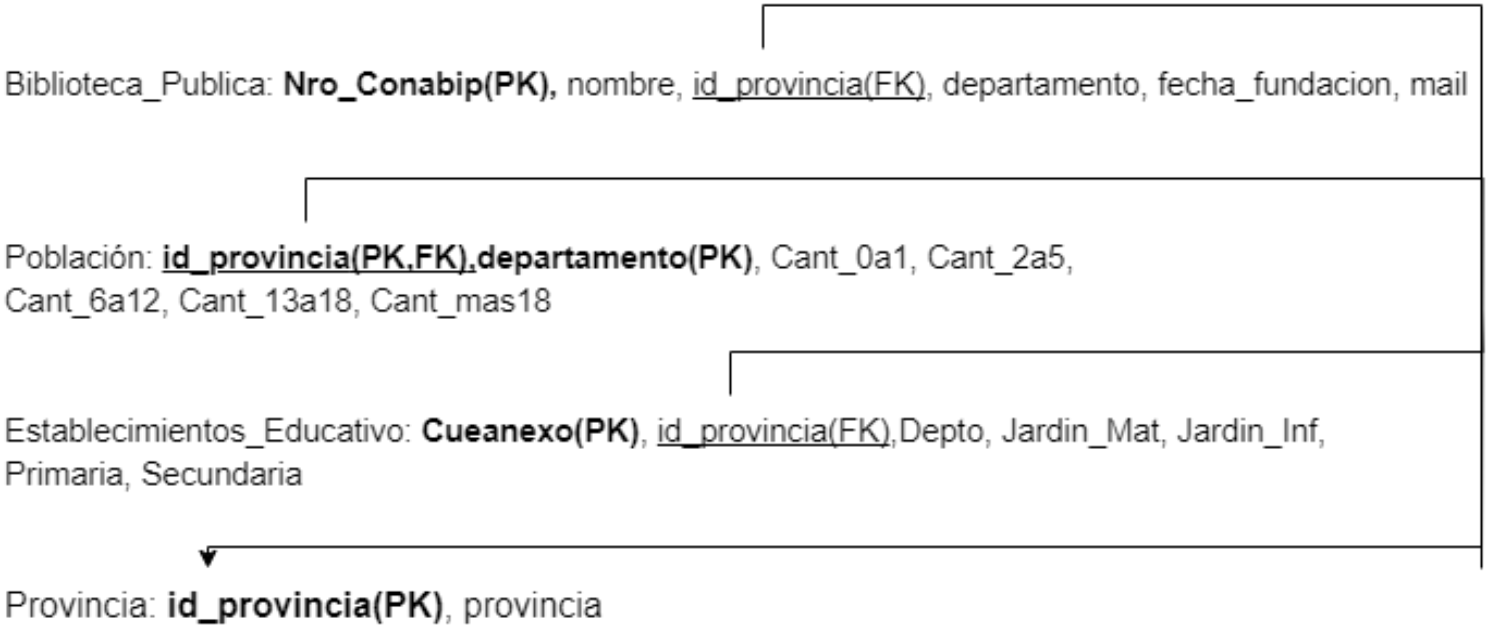
Si bien este diagnóstico revela problemas en varios atributos. Teniendo en cuenta nuestro objetivo de encontrar una relación entre establecimientos educativos y bibliotecas populares. Estos resultados nos resultan favorables, ya que los datos que mayor importancia tienen para nuestro propósito son aquellos que describen su ubicación (como provincia, localidad, departamento, etc.) y obviamente alguna clave primaria que nos ayude a identificarlas. Y justamente son los atributos que presentan la consistencia y la completitud que necesitamos para el análisis. Algunas posibles acciones de mejora sutiles pero útiles pueden ser:

- Eliminar columnas sin información o que no son relevantes para el análisis
- Normalizar el tipo de dato de cada atributo
- Completar de ser necesario valores faltantes usando otras fuentes

Diagrama Relacional(DER)



Modelo Relacional:



Según nuestro DER(Diagrama-Entidad-Relación) tenemos construidos cuatro dataframes (Bibliotecas,Establecimientos,Población,Provincias). Estos se encuentran en formato CSV en la carpeta “DataFrames_to_CSV” con sus respectivos nombres. (ej: “bibliotecas.csv”)

La entidad “Población” tiene como clave primaria una clave compuesta por el código de provincia al que pertenece, que lo obtuvimos quedándonos con los primeros dos dígitos del código de área, y el nombre del departamento, además de contar con una nueva separación de los datos por categorías de edad que corresponden a las edades de los niveles educativos. Para eso creamos un nuevo archivo csv con el que se nos hizo más fácil crear el dataframe que tiene como columnas las divisiones de edad nombradas anteriormente.

La entidad “Biblioteca Pública” tiene los atributos: nombre, código de provincia, departamento, fecha de fundación, mail y número de Conabip, siendo este último la clave primaria. Además, tiene agregada una columna “dominio”, donde se guarda únicamente el dominio del mail de cada biblioteca y se modificó la columna de “fecha_de_fundacion” para que solo contenga el año.

Para la entidad “Establecimientos Educativos” los atributos son los distintos niveles educativos que contiene, la clave primaria o “Cueanexo”, donde sus primeros dos dígitos son el código de la provincia a la que pertenecen, y el departamento. Las primeras filas de la tabla original fueron eliminadas porque contaban con información no redundante.

Como todas las entidades están relacionadas mediante el código de provincia, creamos una nueva entidad “Provincia” que tiene como clave primaria este mismo y además tiene como atributo su nombre. Este lo obtuvimos del archivo original de bibliotecas, eliminando los repetidos.

Ya teniendo los tres dataframes principales nos pusimos a darle consistencia a los datos de la columna “Departamento” ya que estos son los que vamos a utilizar para comparar la información. Como para nosotros la cantidad total de departamentos es la dada por el df de Población, tuvimos que modificar manualmente aquellos departamentos que estaban abreviados o tenían tildes en los df de establecimientos y bibliotecas. Por último pasamos las tres columnas a mayúsculas.

Todo lo mencionado anteriormente se puede encontrar en el archivo “DataFrames.ipynb”.

Dependencias Funcionales:
POBLACIÓN:
{id_provincia, departamento} → {Cant_0a2, Cant_3a5, Cant_6a12, Cant_13a17, Cant_mas18}

ESTABLECIMIENTOS EDUCATIVOS
Cueanexo → {id_provincia, departamento, Jardin_Mat, Jardin_Inf, Primaria, Secundaria}

BIBLIOTECAS PÚBLICAS
nro_conabip → {nombre,id_provincia, departamento, fecha_fundacion, mail}

UBICACIÓN
id_provincia→ {provincia}

Estas dependencias cumplen con la tercera forma normal ya que no hay dependencias entre los atributos que no forman parte de la clave primaria y además las dependencias de las claves primarias compuestas no se podrían dar con un subconjunto de las mismas.

Decisiones tomadas

En primer lugar, determinamos que para nuestras consultas íbamos a necesitar únicamente la información de los establecimientos de modalidad “común”, así que eliminamos todos los datos de los establecimientos que no pertenecían a esta modalidad y todas las columnas siguientes a la misma. Por lo que en la tabla solo quedaron los registros de las escuelas con esta modalidad y sus respectivos niveles educativos. Tuvimos la idea de separar los diferentes niveles educativos de la modalidad para evitar que se genere una gran cantidad de “Nulls”, pero concluimos que al momento de hacer las consultas de SQL no iba a generar un problema. Además, nos dimos cuenta que esta tabla no contaba con el código de provincia así que lo generamos a través del “Cueanexo”. Nuestra primera opción era quedarnos únicamente con los primeros dos dígitos de cada uno pero nos dimos cuenta que en una cantidad finita de establecimientos con código “02” y “06” se habían cargado sin el primer dígito “0”, por lo tanto optamos por cortarlos de atrás para adelante. Eso hizo que nos quedemos solo con un dígito por lo que reemplazamos a todos por el valor “2” y “6”.

Luego, observamos que la tabla de “Bibliotecas Públicas” no tenía las bibliotecas de CABA separadas por comuna como lo tenían las demás. Nuestra primera opción fue modificar los departamentos de los registros de las tablas “Población” y “Establecimientos Educativos” que tienen como provincia a “Ciudad Autónoma de Buenos Aires”. La segunda opción, era buscar todos los registros de la tabla de “Bibliotecas Públicas” que cumplían con lo pedido anteriormente para los otros registros y cambiar manualmente el nombre del departamento por la comuna correspondiente fijándonos por la ubicación del domicilio o por las coordenadas. Por último, nuestra idea definitiva fue crear dos nuevos dataframes que fuesen una copia de “Población” y “Establecimientos Educativos” y a esos aplicarlos nuestra primera opción. Por lo que ahora, tendríamos dos dataframes compatibles para poder comparar ambas instituciones por departamento sin perder la información por departamento de una provincia tan importante como CABA.

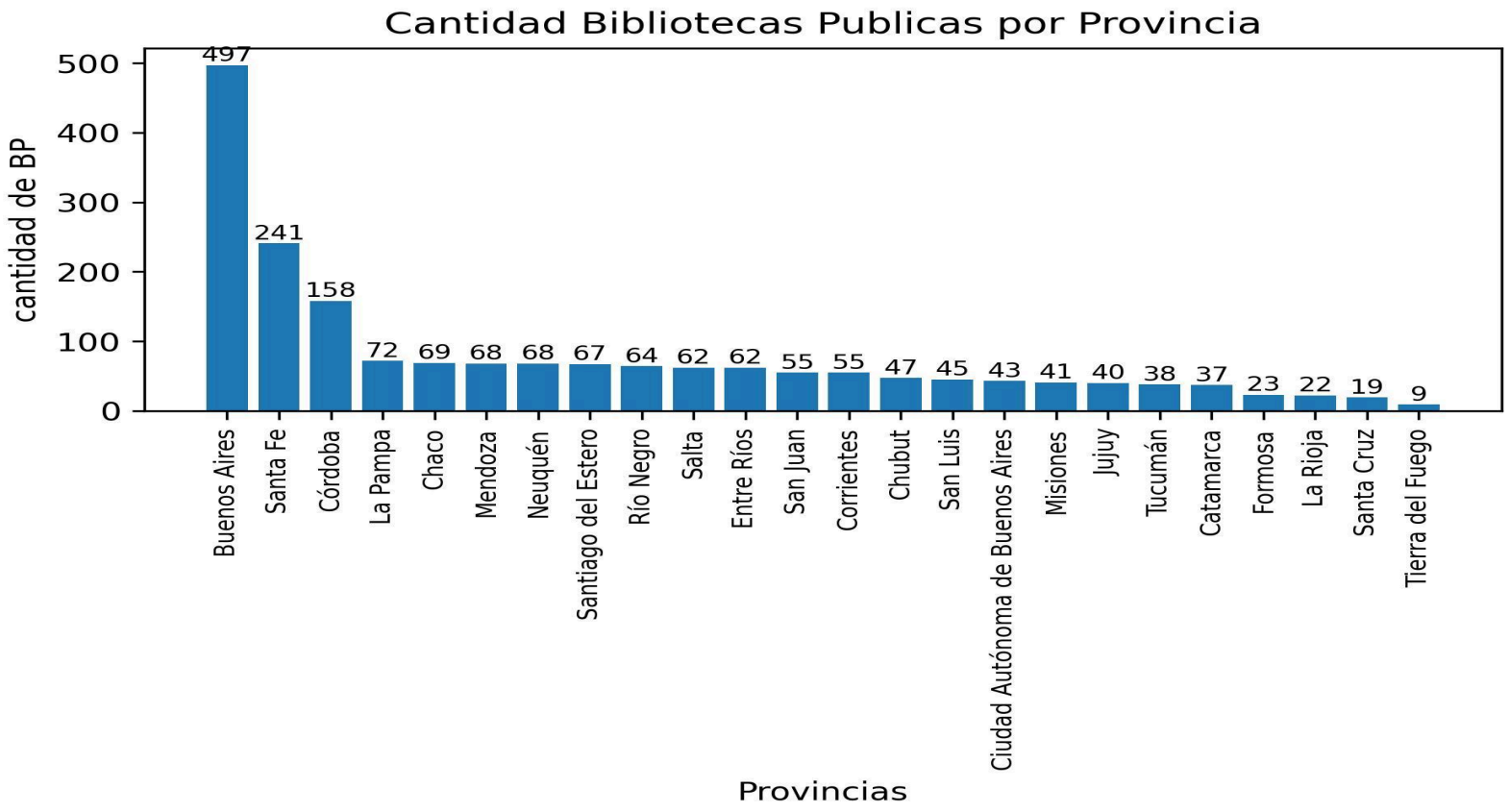
Como íbamos a comparar todas las tablas por código de provincia y solo la tabla original de bibliotecas lo contenía, tuvimos que modificar alguna columna de las demás tablas para conseguir este dato.

Con respecto a la visualización de datos, debido a que nuestro objetivo era ver si existe relación en las provincias, decidimos que todos nuestros gráficos sean provinciales y no respecto a los departamentos de cada provincia, ya que visualizar los 500 departamentos del país generaba gráficos ilegibles(Ver anexo 1). De igual manera, la información continua segmentada por departamentos ya que nos parece relevante tenerla guardada de esa manera por posibles futuras consultas.

Análisis de datos

En esta sección se presentan los principales resultados obtenidos a partir del procesamiento y exploración de los datos previamente normalizados y organizados en el modelo relacional. A través de consultas SQL y herramientas de visualización, se buscó identificar patrones, relaciones y particularidades en la distribución de establecimientos educativos y bibliotecas populares en las distintas provincias y departamentos del país. A continuación, se detallan los reportes y gráficos generados, junto con las interpretaciones correspondientes.

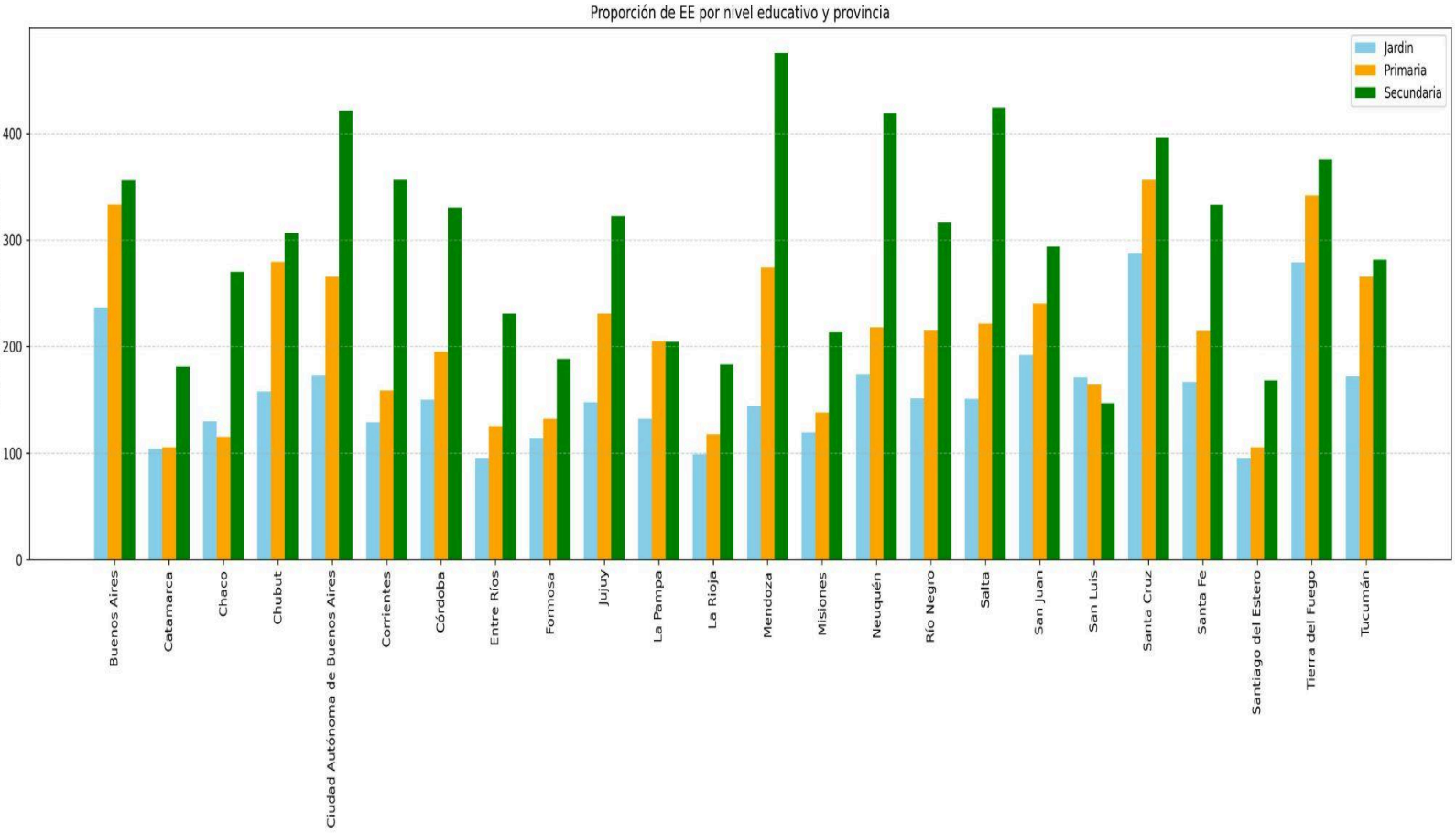
Figura 1:



Este gráfico muestra la cantidad total de Bibliotecas Públicas en cada provincia de Argentina. Se observa que una gran parte de estas instituciones están concentradas en solo tres provincias: Buenos Aires, Santa Fe y Córdoba.

Si bien es probable que esta distribución está influenciada por el tamaño de la población, este no parece ser el único factor determinante. Por ejemplo, la Ciudad Autónoma de Buenos Aires, que tiene una alta densidad poblacional, presenta una cantidad significativamente menor de bibliotecas públicas en comparación con otras provincias grandes.

Figura 2:

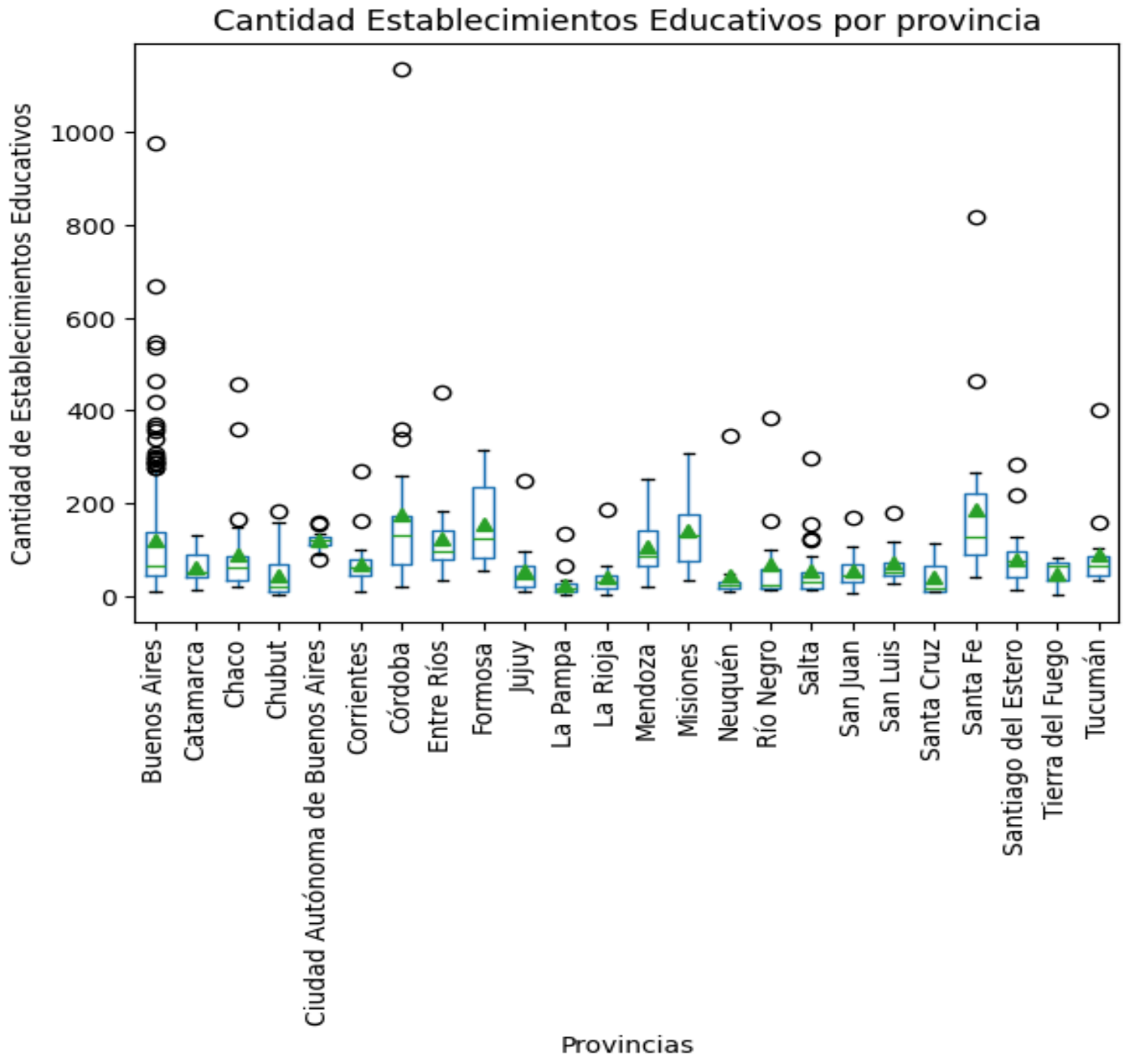


Este gráfico muestra la cantidad promedio de estudiantes por establecimiento educativo, diferenciada por nivel educativo (jardín – azul, primaria – naranja, secundaria – verde) en cada provincia de Argentina.

Se observa que, en general, los establecimientos de nivel secundario concentran la mayor cantidad de estudiantes. Esto sugiere que existen menos instituciones secundarias en relación con la cantidad de alumnos, lo que provoca una mayor concentración de estudiantes por escuela.

Por otro lado, las provincias con valores promedio más bajos suelen ser aquellas con menor población o con una mayor cantidad de establecimientos educativos distribuidos, lo que permite reducir la cantidad de estudiantes por institución.

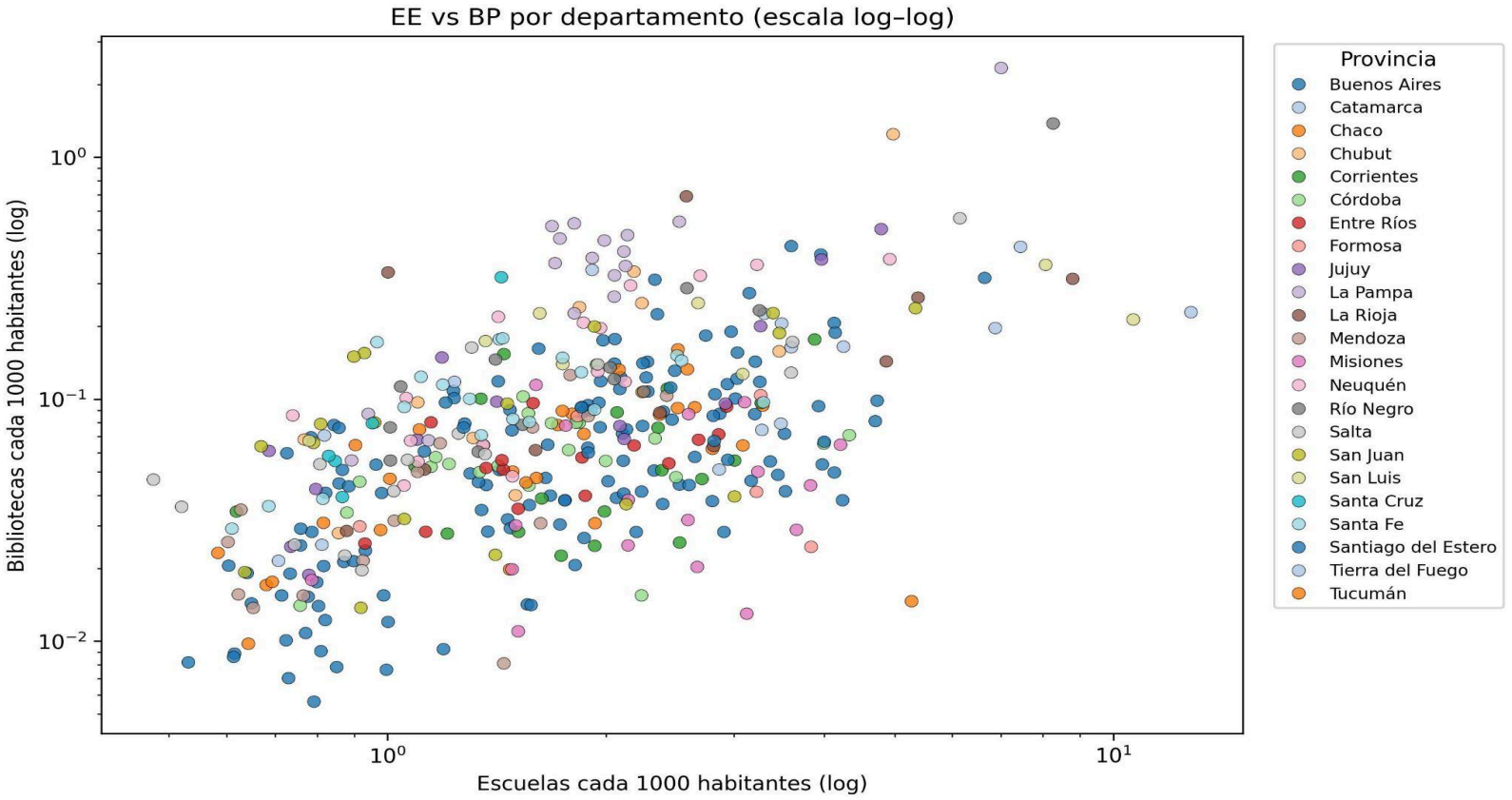
Figura 3:



Este gráfico utiliza boxplots para representar la distribución de la cantidad de EE por provincia. Se observa que varias provincias presentan una fuerte dispersión en la cantidad de escuelas por departamento, lo que indica desigualdades internas dentro de una misma provincia.

En líneas generales, se observa que la distribución de establecimientos educativos es más equilibrada entre provincias que la distribución de bibliotecas públicas. Esto podría interpretarse como un reflejo de que el sistema educativo formal recibe una atención y planificación más sistemática, mientras que el acceso a bibliotecas puede estar más sujeto a factores políticos, presupuestarios o culturales.

Figura 4:



Este gráfico muestra la cantidad de Establecimientos Educativos y Bibliotecas Públicas por cada mil habitantes en cada departamento de las provincias de Argentina, utilizando una escala logarítmica en ambos ejes.

En líneas generales, se observa una tendencia positiva: a medida que aumenta la cantidad de escuelas por habitante, también lo hace la cantidad de bibliotecas. Sin embargo, también se destacan casos particulares en los que esta relación no se cumple, ya que hay departamentos con muchas escuelas y pocas bibliotecas, o viceversa. Esto sugiere que, si bien existe una correlación general, factores locales pueden influir en la distribución de estas instituciones.

Además se hicieron 4 consultas que no se utilizaron para visualizar (Estas se encuentran en el archivo “ConsultasSQL.py”. La primera consulta, no se utilizó explícitamente para graficar pero se usó una modificación de las misma para la Figura 2. Dos de ellas, la segunda y la cuarta, hacen referencia a atributos específicos de las Bibliotecas Populares, las cuales son irrelevantes para nuestra investigación. Una proponía analizar los dominios de mail más utilizados en el departamento, y la segunda buscar la cantidad de bibliotecas fundadas a partir del año 1950. Y por último, de la tercera consulta se pueden obtener conclusiones interesantes para el objetivo de este informe.

Figura 5:

	provincia	departamento	cant_BP	cant_EE	Poblacion
0	Ciudad Autónoma de Buenos Aires	CIUDAD AUTONOMA DE BUENOS AIRES	43	1782	3095454
1	Córdoba	CAPITAL	21	1136	1498060
2	Buenos Aires	LA MATANZA	15	977	1837168
3	Santa Fe	ROSARIO	39	817	1337958
4	Buenos Aires	LA PLATA	33	669	756074
...	...	...	...	...	...
508	La Rioja	SANAGASTA	1	3	2994
509	Chubut	FLORENTINO AMEGHINO	0	3	1782
510	Chubut	MARTIRES	0	3	744
511	Chaco	O'HIGGINS	1	0	21544
512	Tierra del Fuego	TOLHUIN	0	0	6027

La consulta permite observar la relación entre el número de bibliotecas populares (BP), establecimientos educativos (EE) de modalidad común, y la población total en cada departamento del país.

El análisis evidencia patrones interesantes y también desigualdades significativas:

- Alta correlación poblacional – institucional: Los departamentos con mayor población presentan los valores más altos tanto en cantidad de EE como de BP. Esto refleja una cierta relación entre concentración poblacional y oferta institucional.

- Departamentos sin bibliotecas: Se identifican casos donde no hay ninguna biblioteca popular registrada, pero sí existen EE.

- Departamentos con muchas escuelas pero pocas bibliotecas: En algunos casos la cantidad de EE es muy alta pero la presencia de bibliotecas populares es baja en proporción, lo que sugiere que no siempre la oferta educativa está acompañada por una oferta cultural equivalente.

En conclusión, aunque existe una tendencia general a que los departamentos más poblados concentren más instituciones, la distribución no es equitativa ni sistemática.

Las tablas completas de todas las consultas hechas en SQL a partir de las cuales se hizo todo el análisis de datos se encuentran en la carpeta “Consultas SQL”

Conclusiones

Este trabajo nos permitió integrar información de Establecimientos Educativos y Bibliotecas Populares para analizar su vínculo con criterios poblacionales, etarios y geopolíticos. A partir del procesamiento de datos, análisis y visualización, se obtuvo el siguiente conocimiento:

- La cantidad de bibliotecas públicas está fuertemente concentrada en pocas provincias (Buenos Aires, Santa Fe, Córdoba). Sin embargo, esta concentración no guarda una relación directa con la población, como lo evidencia el caso de CABA, que a pesar de su densidad poblacional, posee relativamente pocas bibliotecas.

- En promedio, los establecimientos secundarios concentran más estudiantes por escuela que los niveles primario e inicial, lo que sugiere una menor cantidad de escuelas secundarias disponibles para la demanda existente.

- La distribución de establecimientos educativos por departamento es desigual dentro de muchas provincias, especialmente en las más extensas o densamente pobladas. Esto revela desbalances internos importantes en el acceso a la educación.

- En términos generales, la distribución de EE entre provincias es más homogénea que la de bibliotecas, lo que sugiere que el sistema educativo formal recibe mayor atención y planificación sistemática, mientras que la implementación de bibliotecas depende más de iniciativas culturales, políticas locales o disponibilidad presupuestaria.

A partir del análisis realizado, se observa una relación positiva general entre la cantidad de establecimientos educativos y bibliotecas públicas, tanto en términos absolutos como relativos a la población. No obstante, esta relación no es perfectamente proporcional, y presenta variaciones significativas entre provincias y departamentos.

En resumen, la oferta de bibliotecas no siempre acompaña de forma directa a la de escuelas, pero se identifican patrones que permiten pensar en una vinculación parcial y condicionada por el contexto regional.