Análisis sobre la Relación entre Establecimientos Educativos y Bibliotecas Populares en Territorio Nacional

25/05/2025

Trabajo Práctico Nro 1

1er Cuatrimestre 2025

Facultad de Ciencias Exactas y Naturales - UBA - Ciudad Universitaria

Dorogov Cristina - Pucciarelli Francisco Lautaro - Salto Julián

Resumen

En este trabajo se propone analizar la relación entre bibliotecas populares y establecimientos educativos en cada departamento en cada una de las provincias de Argentina. Para eso, se recurrió a tres fuentes de datos abiertos de información oficial, "*Padrón oficial de Establecimientos Educativos*","*Padrón de bibliotecas populares*" y "*Padrón población*" siendo las tres de ellas del año 2022. La hipótesis planteada es que a mayor población se presenta mayor cantidad de establecimientos educativos y bibliotecas populares. Los análisis que se van a realizar son ver en qué medida aumentan los establecimientos educativos según la distribución de la población en sus departamentos y cómo aumentan las bibliotecas populares conforme al aumento de escuelas. Para eso, se va a implementar procesamiento de los datos para llevarlos a su forma normal y facilitar un análisis posterior en el cual mediante consultas a la base y gráficos se pueda abordar la pregunta planteada.

introducción

Se pretende resolver si a mayor población se observa mayor cantidad de bibliotecas populares y establecimientos educativos. Para ello se analizaron bases de datos en Pandas y se buscó que las mismas se encuentren en tercera forma normal ya que de esta manera sólo se observa dependencias funcionales completas y se evita tener información redundante y atributos multivariados en una misma tabla.

En la sección *Procesamiento de Datos*, se cuenta sobre cómo se encontraban las fuentes de datos en su estado original y se realizan dos análisis de calidad de datos y su respectivo diagnóstico detallando el problema donde se sugieren mejoras de acción. También, plantea un diagrama entidad-relación y el modelo relacional, el cual contiene todas las tablas que se utilizaran para el posterior análisis. En la sección *Decisiones Tomadas*, se explican los problemas enfrentados y cómo se los decidió solucionar. En *Análisis de datos*, se muestran diversos gráficos y sus correspondientes estudios. Se buscó poder relacionar diversos gráficos entre sí. Por último, la sección *Conclusiones* contiene las conclusiones del trabajo realizado. En la misma se encuentra que a mayor población, si hay mayor cantidad de establecimientos educativos pero no ocurre lo mismo con bibliotecas populares. En el *Anexo* se encuentra el mapeo del diagrama entidad-relación y gráficos adicionales que ayudaron en la investigación.

Procesamiento de Datos

Diagrama entidad-relación y modelo relacional

El primer paso realizado fue comprender con qué información se estaba trabajando. Se quiso conocer cuáles eran las entidades participantes y sus atributos. Para ello, se buscó identificar en las tablas dadas los atributos propios de cada entidad, los que eran de interés para el trabajo, y diferenciarlos de las claves foráneas ya estipuladas, permitiendo así armar el siguiente diagrama entidad-relación (DER)

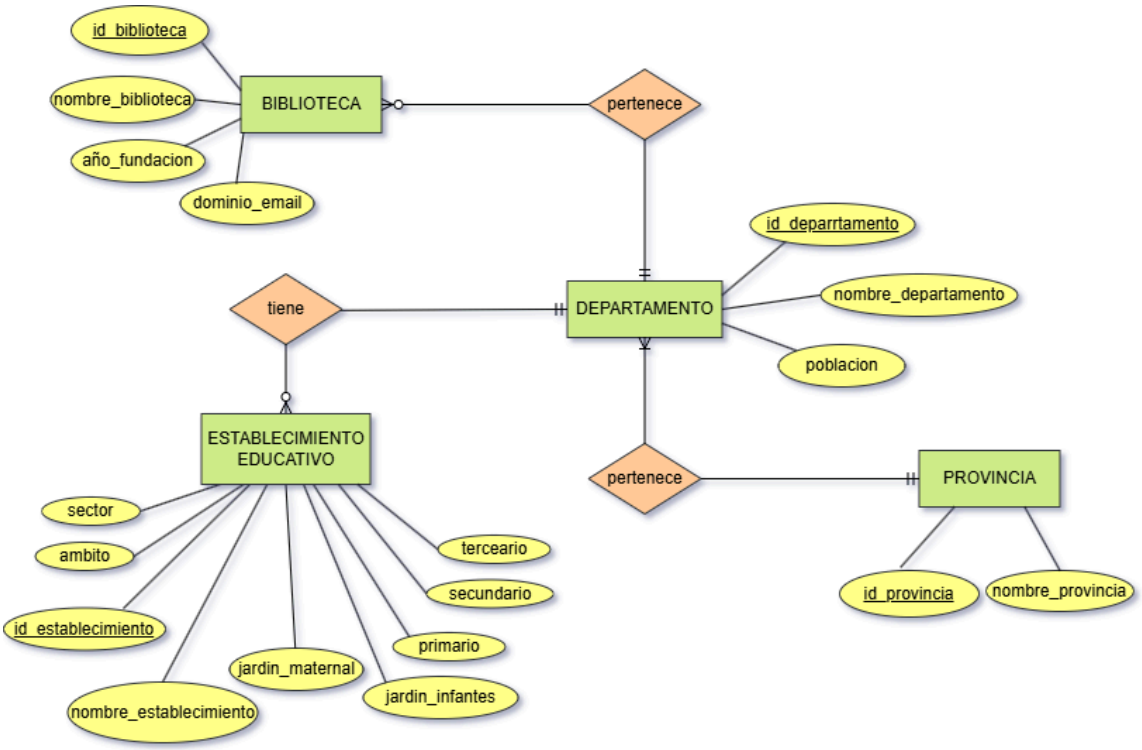


Figura 1. Diagrama entidad-relación (DER)

El siguiente paso fue armar el modelo relacional (MR). Para confeccionarlo se utilizó la plataforma en línea dbdiagram que utiliza como base el lenguaje DBML. Primero, se realizó un mapeo[1] del DER y luego a través de la observación del tipo de relaciones y sus respectivas cardinalidades se armaron nuevas relaciones y a otras se les insertó claves foráneas. Estas relaciones fueron las utilizadas para realizar el trabajo y se encuentran en tercera forma normal.

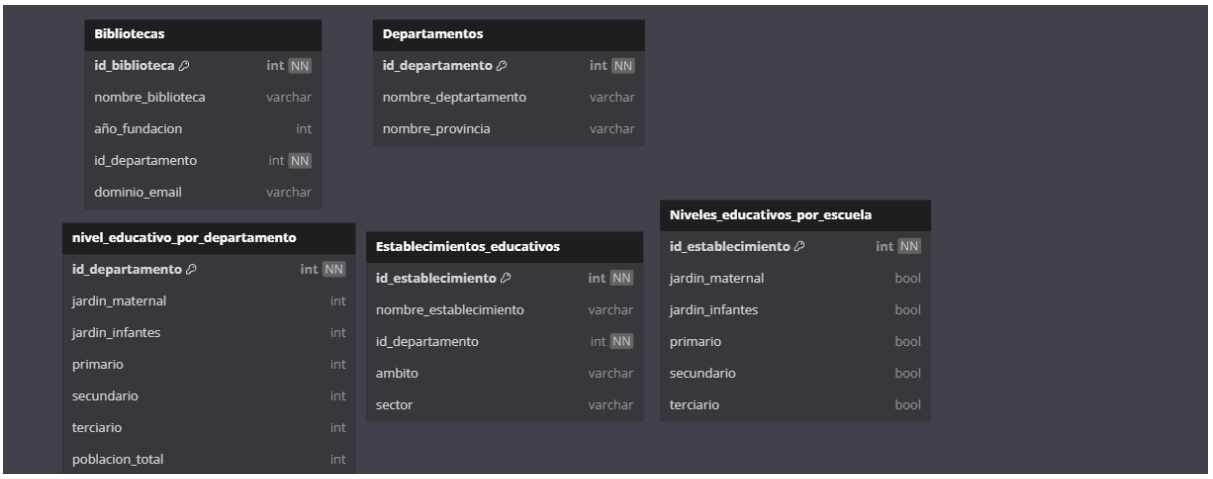


Figura 2. Modelo Relacional (MR) y sus claves primarias y foráneas. El primer atributo de cada relación es la clave primaria (PK), y en donde dice NN quiere decir "Not Null" no vacío.

[1] El mapeo fue incluido en el anexo

Algunos ejemplos de claves foráneas definidas son:

id_depto en la relación *Establecimientos_educativos* y *nivel_educativo_por departamento*
id_establecimiento en la relación *Establecimientos_educativos*

A continuación se muestra de manera minimal las dependencias funcionales.

Bibliotecas:
id_biblioteca → {nombre_biblioteca, año_fundacion, id_departamento, dominio_email}

Departamentos:
id_departamento → {nombre_departamento, nombre_provincia}

Población_por_nivel_educativo:
id_departamento→{jardin_maternal,jardin_infantes,primario,secundario,terciario, poblacion_total}

Establecimientos_educativos:
id_establecimiento → {nombre_establecimiento, id_departamento, ambito, sector}

Niveles_educativos_por_escuela:
id_establecimiento → {jardin_maternal,jardin_infantes, primario, secundario, terciario}

Tablas originales

En el archivo excel *"2022_padron_oficial_establecimientos_educativos*" se observaba a simple vista reiterados valores "Null" y además contenía relaciones anidadas. Las mismas están prohibidas en primera forma normal. Como consecuencia, la tabla no estaba normalizada.

Común						
Nivel inicial - Jardín maternal	Nivel inicial - Jardín de infantes	Primario	Secundario	Secundario - INET	SNU	SNU - INET
1	1	1	1			
		1				
1	1	1				
1	1	1	1			
1	1	1	1			
		1				
		1				
		1				
		1				
		1				
1	1	1				
		1				

Figura 3. Ejemplo de relación anidada con valores null

Para revertir esto, se propuso reagrupar todas las modalidades anidadas bajo el atributo "*modalidades*" el cual puede adquirir los valores *Común, Especial, Adultos, Artística, Hospitalaria, Intercultural* y *Encierro.* De esta manera se logra evitar relaciones anidadas y valores null.
Con el fin de poder conocer un poco más sobre la estructura de la tabla, si bien la misma al no encontrarse en primera forma normal no se iba a encontrar ni en segunda ni tercera forma normal, se quiso realizar de todas maneras dependencias funcionales para conocer si había información redundante. Se observa que hay atributos no primos que dependen de otros atributos no primos, implicando así transitividad y redundancia.

Cueanexo|nombre|sector|ambito|departamento|codigo localidad|localidad|email|modalidad

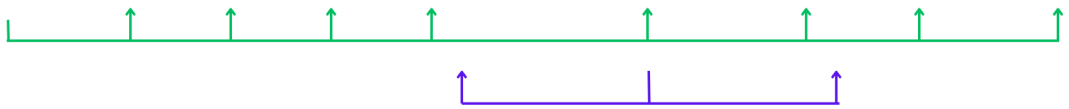


Figura 4. Sea Cueanexo la clave primaria, se observa la dependencia de atributos no primos de otros atributos no primos.

En lo que respecta a la limpieza de datos, se usó la librería Pandas. Para simplificar el volumen de información, se decidió trabajar únicamente con la modalidad "*común*" junto con sus cinco niveles educativos, jardín maternal, jardín de infantes, primario, secundario y terciario. Cabe aclarar que SNU (educación superior no universitario) y SNU-INET (educación superior no universitario, instituto nacional de educación tecnológica) se los agrupó bajo el atributo "*Terciario*" y se agrupó a secundario y secundario-INET bajo el mismo atributo "*secundario*".

Este dataset permitió armar las tablas "*Departamentos*","*Establecimientos_educativos*", *"Niveles_educativos_por_escuela*" y "*nivel_educativo_por_departamento*", presentes en el modelo relacional de la figura 2. La relación *Departamentos* indica los departamentos existentes por cada provincia y se tuvo en cuenta la *jurisdicción* como *nombre_provincia* y el *departamento* como *nombre_departamento*. Para la relación *Establecimientos_educativos*, la cual muestra las diferentes escuelas que hay por cada departamento, se renombró al *cueanexo* como *id_establecimiento*, el nombre de la institución como *nombre_establecimiento*, ambito y sector se los dejó bajo el mismo nombre y, al igual que en la relación anterior, se importó el id_departamento de la misma tabla. La relación *Niveles_educativos_por_escuela* se nombró al cueanexo como id_establecimiento, y luego, los distintos niveles educativos presentan valores de 0 o 1, donde 0 es False y 1 es True que indican si un establecimiento educativo cuenta con jardin_maternal, jardin_infantes, primario, secundario y/o terciario.

El archivo bibliotecas-populares.csv no presentaba relaciones anidadas, por ende se encontraba en primera forma normal. Sin embargo, el mismo no se encuentra en tercera forma normal dado que el mismo presenta relaciones transitivas.

nro conabip|id_provincia|id_departamento|provincia|departamento|nombre|fecha_fundacion

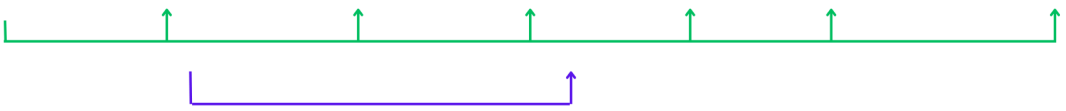


Figura 5. Sea nro_conabip la clave primaria, se observa al menos una relación transitiva por parte de id_provincia y provincia.

Para su limpieza, se usó la librería Pandas. Una primera inquietud era saber si existen dos tuplas iguales. Se verificó que el nro_conabip era diferente para cada una de las bibliotecas y en base a eso se verificó que no existían tuplas repetidas. Se optó eliminar los atributos de *latitud, longitud, tipo latitud longitud, teléfono, web, domicilio, observación, categoría, fuente,año actualización,código de teléfono, piso, código postal e información adicional* ya que los mismos no presentaban información relevante alguna para la investigación. Por ejemplo, *categoría* era común a todos, ya que siempre figuraba "*Biblioteca popular*" y el nombre del archivo csv era "*bibliotecas populares*". Otro dato a observar fue que las bibliotecas populares correspondientes a Ciudad Autónoma de Buenos Aires no estaban

clasificadas por departamento. Su localidad era igual a su departamento[2]. Siguiendo el modelo relacional planteado en la Figura 2, se armó una nueva tabla con los atributos renombrados como en el DER.

La tabla "*padron_poblacion"* presentó grandes inconvenientes en cuanto a su lectura[3] y trabajo. La misma tampoco estaba normalizada, ya que el dataset contiene distintas tablas consecutivas, y para revertir esto, se propuso reacomodar la estructura general del archivo. Uno de los principales problemas enfrentados fue que de los atributos incluidos en cada una de las tablas ninguno podría ser considerado clave primaria. El que sí podía ser considerado clave primaria, el código de área, no se encontraba explícitamente en la tabla. Además se cuenta con información redundante como son los porcentajes.

AREA # 02007 Comuna 1

Edad	Casos	%	Acumulado %
0	1 720	0.78%	0.78%
1	1 652	0.75%	1.53%
2	1 996	0.90%	2.43%
3	2 079	0.94%	3.37%
4	2 329	1.05%	4.42%
5	2 443	1.11%	5.53%

Figura 5. Se observa que el atributo más importante, el codigo de area, y el nombre quedan por fuera de la tabla

Para la limpieza del dataset, también se utilizó la librería Pandas y se omitieron los porcentajes que se observan en la figura 5. A base de esta tabla se hizo la relación nivel_educativo_por_departamento, la cual muestra cuántos habitantes de cada departamento participan de cada nivel educativo[4]. Los atributos de niveles educativos son los que se presentaban en el excel de establecimientos educativos y se agregó el atributo de id_departamento obtenido de la tabla de establecimientos educativos. También se decidió agregar el atributo correspondiente a la cantidad de habitantes en el departamento.

Calidad de datos

Para conocer la calidad de los datos, utilizó la técnica GOAL, QUESTION, METRIC (GQM).

Para Establecimientos Educativos

Goal: Analizar los números de teléfono de distintos establecimientos educativos.
Question: ¿Qué porcentaje de escuelas presenta más de un teléfono?

Metric: $\frac{100\ x\ cantidad\ de\ establecimientos\ con\ más\ de\ un\ teléfono}{Cantidad\ total\ de\ establecimientos\ educativos}$

El atributo de calidad afectado es la consistencia. Al presentarse dos teléfonos, lo cual también hace que sea un valor no atómico, no se sabe a cual llamar. Uno se podría preguntar cuál de los teléfonos

[2] Para mayor detalle ver la sección de decisiones tomadas
[3] Para conocer más sobre cómo se hizo para leerlo, se puede ver en los comentarios del código.
[4] ver sección Análisis de datos para más detalle

corresponde a jardín de infantes, primario o secundario o si da igual a cual llamar. Esto es un problema de instancia ya que no hay precisión para almacenar ese dato específico.
Al aplicar la métrica, se obtuvo un porcentaje de 3,94%. Si bien es un número relativamente bajo, se propone como mejora especificar que se escriba un único teléfono correspondiente a la institución educativa.

Para Bibliotecas Populares

Goal: Revisar bibliotecas sin mail
Question: ¿Qué porcentaje de bibliotecas no cuentan con dirección de mail?

Metric: $\frac{100\ x\ cantidad\ de\ bibliotecas\ sin\ mail}{cantidad\ total\ de\ bibliotecas\ populares}$

Los atributos afectados son completitud y vigencia. Como hay ciertas bibliotecas que no registraron su email, ese dato presenta NULLs. También es posible que los datos no estén actualizados dado que el dataset provisto lleva dos años sin actualización. De todas formas, se detectó que es un problema de software, el email debería ser un dato obligatorio que no se asumió como tal y como resultado no se cargó.

Al aplicar la métrica se obtuvo un porcentaje de 46,27%. Casi la mitad de las bibliotecas no tiene el email registrado, es un número bastante grande que debería ser mejorado. Para eso se propone lo mencionado previamente, hacer del email un campo obligatorio.

Decisiones Tomadas

A los establecimientos educativos de Ciudad Autónoma de Buenos Aires (CABA) se le sacó la división por comunas ya que el dataset de bibliotecas populares no las tenía divididas de tal manera. Se agrupó a todas las escuelas bajo el mismo id_departamento. Esto fue hecho con el fin de facilitar las consultas. Sin embargo, a la hora de graficar, se tuvo en cuenta que cada departamento de CABA, o sea cada comuna, tiene la misma cantidad de bibliotecas y escuelas, ya que si no podría haber inconsistencias cuando se grafica por departamentos de provincias. Para ello, se dividió el total de escuelas (y los datos numéricos asociados a CABA) por quince (total comunas). Idem para bibliotecas.

Se observó que en las tablas de padron_población y bibliotecas_populares el id_departamento de Ushuaia no coincidía, difieren en un número, 94015 y 94014. Se optó por cambiarlo de manera manual con el fin de que coincidan. Lo mismo se hizo para Río Grande ya que se notó el mismo problema. Ambos departamentos de Tierra del Fuego.

En el dataset de establecimientos educativos, cuando una escuela se encuentra alojada en una capital provincial, en la columna “localidad” figura como valor CAPITAL. Si no fuera por el nombre de la provincia, no quedaría claro de qué capital se está hablando. Para que resulte más fácil de interpretar se decidió renombrar esos campos como nombre_provincia + CAPITAL. Por ejemplo, en vez de que diga CAPITAL dirá CATAMARCA CAPITAL.

En la base de datos correspondiente a nivel_educativo_por_departamento se asumió que de 0 a 3 años se participa en jardin_maternal, de 4 a 5, en jardin_infantes, de 6 a 12 primario, de 13 a 18 secundario y las personas entre 19 y 50 quedaron agrupadas en terciario. Cabe aclarar que se tuvo en cuenta que toda la población de esos rangos etarios participa de los niveles educativos y que no hay casos de abandono escolar.

Análisis de datos

A continuación se muestran diversos gráficos que permiten analizar los datos trabajados.

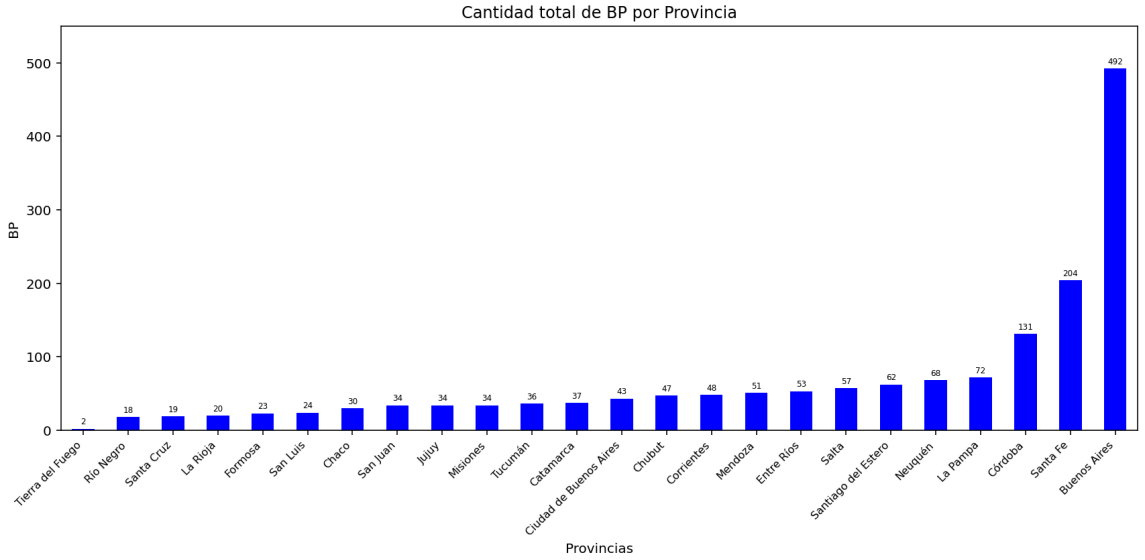


Figura 6. Muestra la cantidad de bibliotecas populares existentes en cada provincia.

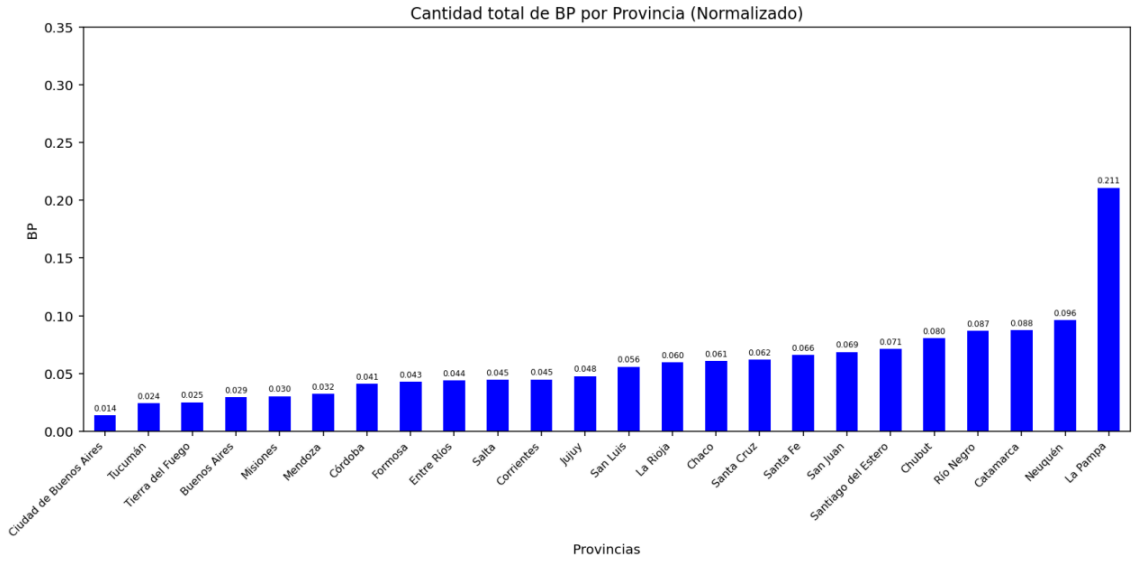


Figura 7. Muestra la cantidad de bibliotecas populares existentes en cada provincia de forma normalizada.

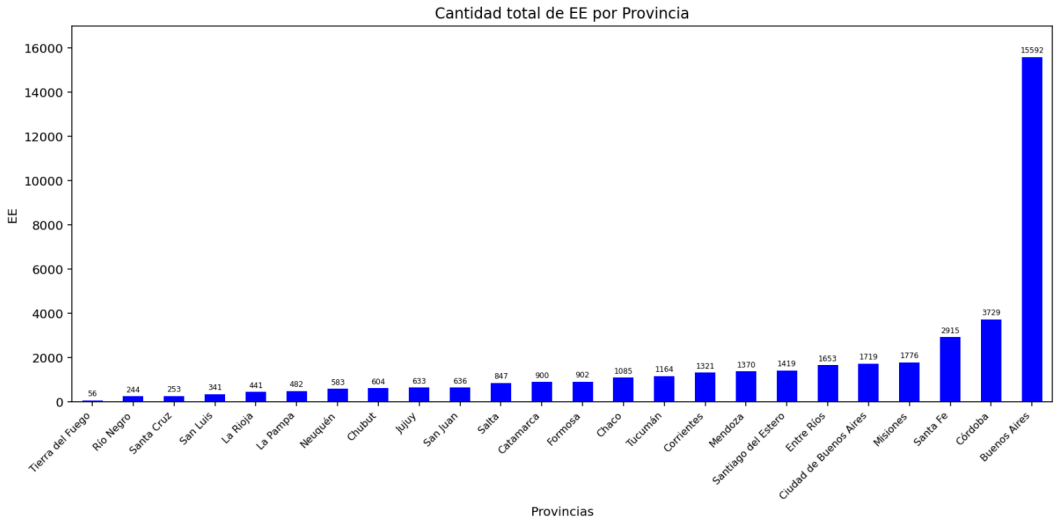


Figura 8. Cantidad de establecimientos educativos por provincia ordenados de forma ascendente.

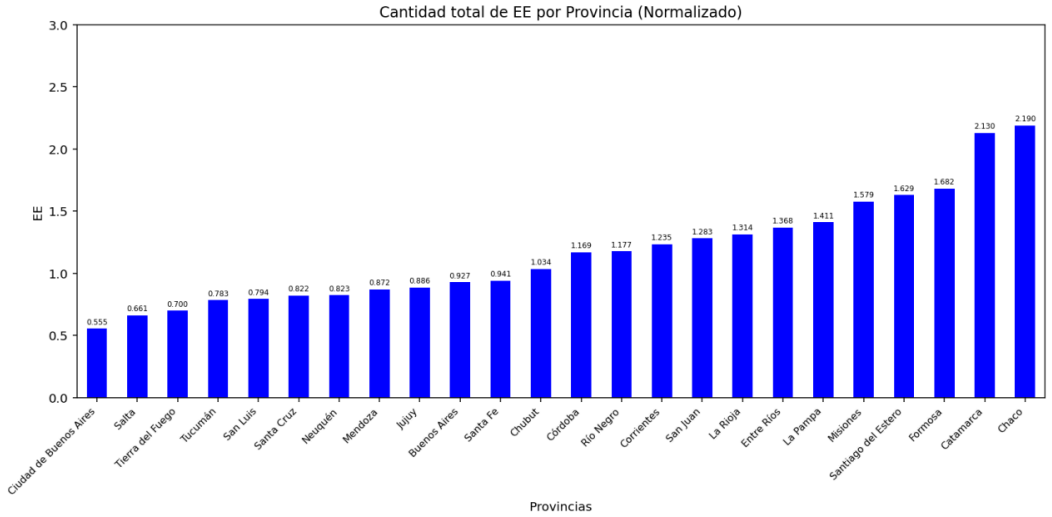


Figura 9. Cantidad de establecimientos educativos por provincia de forma normalizada

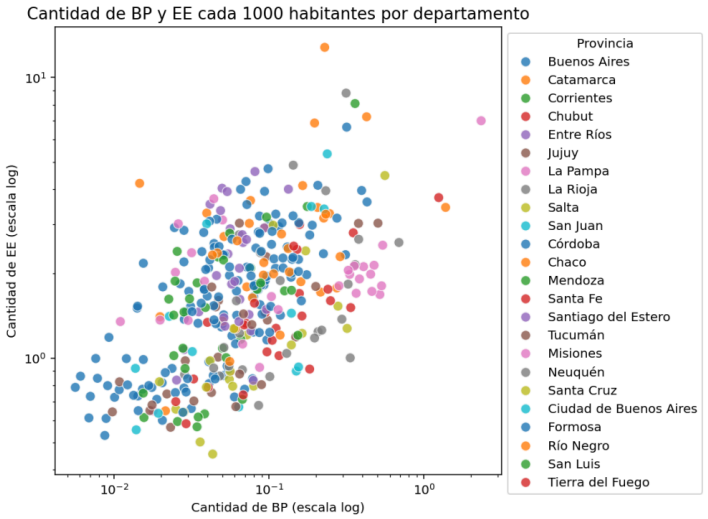


Figura 10. Relación entre cantidad de bibliotecas populares y establecimientos educativos cada mil habitantes.

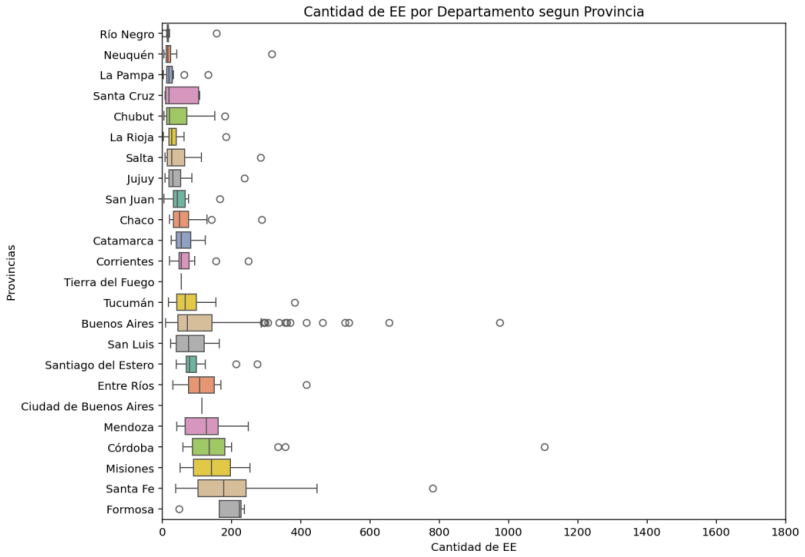


Figura 11. Muestra la cantidad de establecimientos educativos por departamento de cada provincia ordenados según la mediana de cada provincia. Notar que Ciudad de Buenos Aires presenta un único valor dado a que fue considerada como un único departamento.

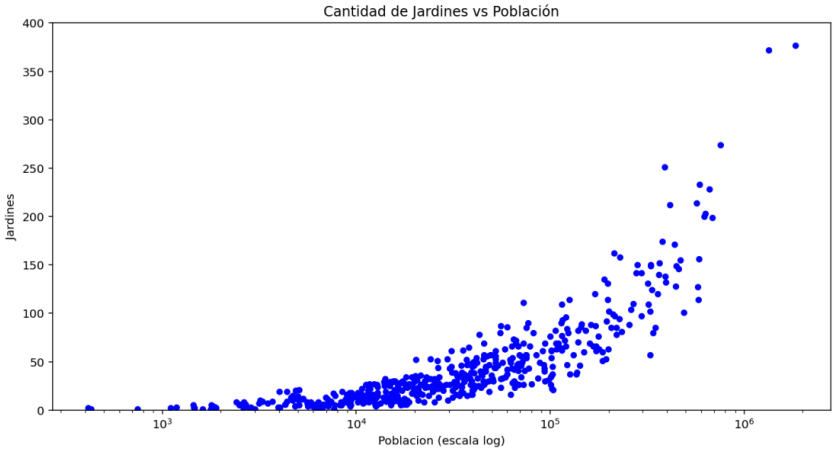


Figura 12. Cantidad de jardines maternales e infantiles por departamento en función de la población.

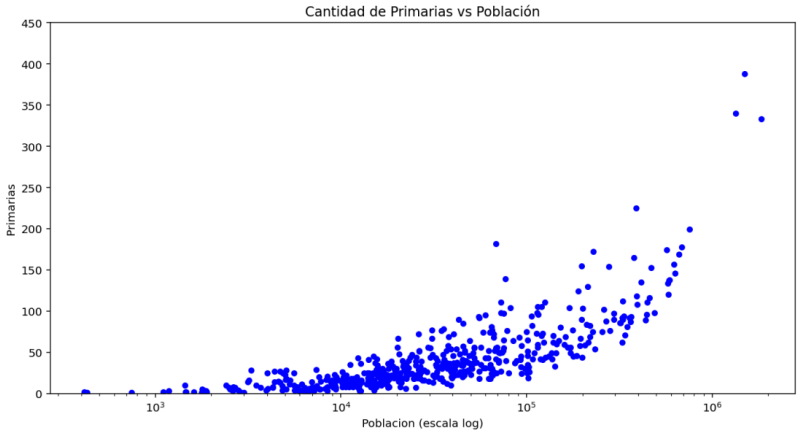


Figura 13. Cantidad de escuelas primarias por departamento en función de la población.

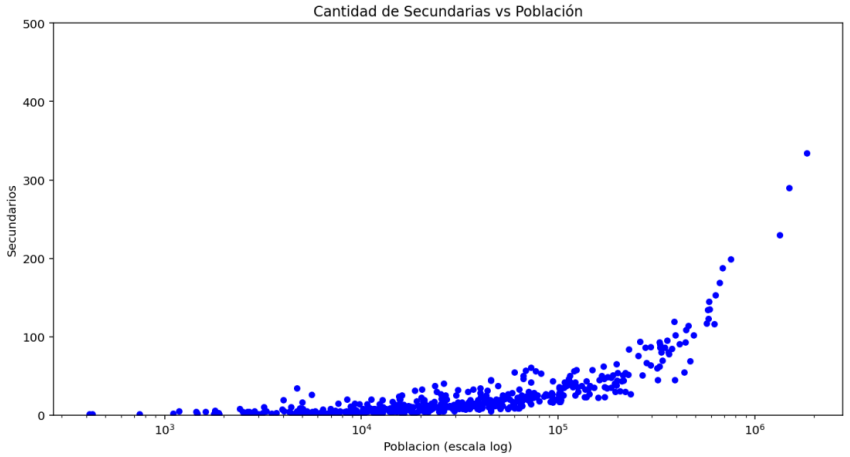


Figura 14. Cantidad de escuelas secundarias por departamento en función de la población.

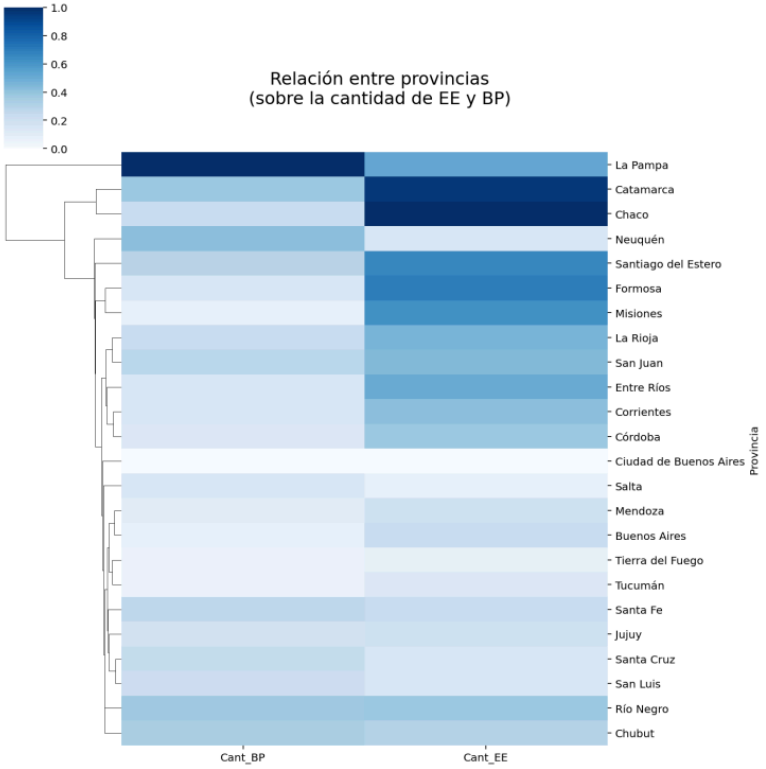


Figura 15. Muestra una relación entre la cantidad de establecimientos educativos y bibliotecas populares entre las diferentes provincias. A mayor oscuridad, mayor es el acercamiento de la provincia al promedio en cuanto a esa variable o más lejano (si es mayor al promedio)

Con los gráficos realizados, en las figuras 12, 13 y 14 cada punto representa un departamento. Se muestra una tendencia positiva entre la cantidad de distintos niveles educativos y la población. A medida que la población es mayor, mayor es la cantidad de niveles educativos. Es de esperarse que haya más puntos concentrados alrededor del centro ya que la mayoría de los departamentos del país cuentan con menos de cien mil habitantes y menos de 100 escuelas en ellos. En la figura 10 se observa una baja cantidad de bibliotecas populares en lo que respecta a la cantidad de establecimientos educativos. Las figuras 6 y 8 presentan una correlación en ambos extremos de los dos gráficos. Se mantiene el orden en las primeras y últimas provincias y, en algunos casos, en provincias intermedias. Este gráfico no alcanza para decir que a mayor cantidad de establecimientos educativos, mayor es la cantidad de bibliotecas populares, ya que, los resultados al no estar normalizados están muy influenciados por el número de población de cada provincia. Para ello, se decidió realizar gráficos normalizados, presentados en las figuras 7 y 9, que no tienen en cuenta la variable población. No se observa que las bibliotecas populares aumenten de forma similar a establecimientos educativos en cada provincia, de hecho, presentan comportamientos bastante diferentes. La figura 15 brinda un análisis

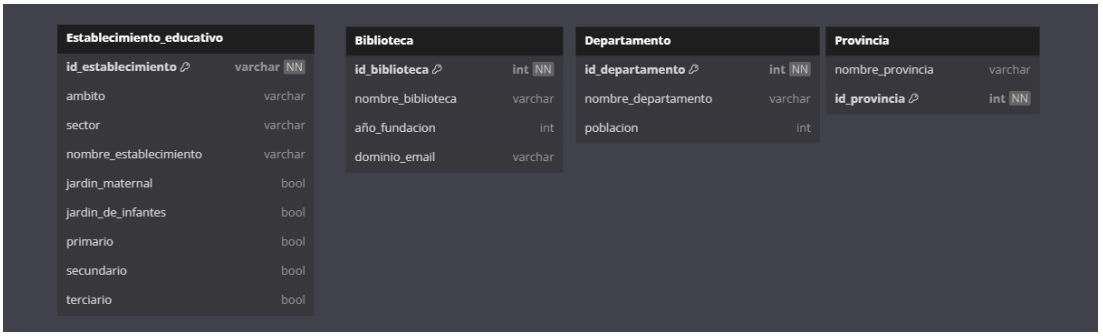
interesante y se observa en el claramente la relación entre gráficos de las figuras 7 y 9. La Pampa presenta valores muy altos de bibliotecas populares y establecimientos educativos en relación al resto de las provincias, mientras que Chaco y Catamarca presentan gran cantidad de establecimientos educativos, superior al resto (en términos normalizados, no totales). CABA presenta valores relativamente bajos en ambas variables. Provincias tales como Chubut y Río Negro estan bastante equiparadas ambas de sus variables y Santiago del Estero, formosa, Misiones, Corrientes, La Rioja, Entre Rios y Córdoba presentan mayor promedio de establecimientos educativos que bibliotecas populares. En Neuquén se da el caso opuesto, hay mayor promedio de bibliotecas que de establecimientos educativos.

Conclusiones

Al principio de la investigación se planteó la hipótesis que a mayor población hay mayor cantidad de escuelas y bibliotecas populares. Tras finalizar el trabajo observamos que la misma se verifica de manera parcial. A mayor cantidad de habitantes hay mayor cantidad de establecimientos educativos, visto claramente en las figuras 12,13 y 14, pero que aumente la población no implica que aumente la cantidad de bibliotecas populares. La población de Neuquén es casi el doble que la de La Pampa, y (Neuquén) presenta menor cantidad de bibliotecas populares.[5] Similar ocurre con Ciudad de Buenos Aires, que cuenta con más de tres millones de habitantes y se encuentra muy por detrás de La Pampa que cuenta con menos de cuatrocientos mil habitantes. Otra manera de visualizar esto fue con el gráfico de la figura 16[6]. El mismo se podría entender de la siguiente manera. Ya se probó que las escuelas aumentan con respecto a la población. Siguiendo la hipótesis uno podría tender a pensar que al aumentar la cantidad de establecimientos educativos, también debería aumentar la cantidad de bibliotecas populares. Esto se contradice en el gráfico mismo ya que hay provincias que tienen el mismo volumen de establecimientos educativos pero distinta cantidad de bibliotecas populares. Entonces, se concluye que a mayor población, hay mayor cantidad de establecimientos educativos. En lo que respecta al aumento o disminución de bibliotecas populares, consideramos que deben haber otros factores que hacen que varíe su cantidad y otro estudio debería ser realizado.

Anexo

A continuación se puede observar el mapeo del diagrama entidad-relación con sus respectivas claves primarias (PK)



Mapeo del DER

[5] Esto se observa en la consulta SQL número tres.
[6] Ver anexo

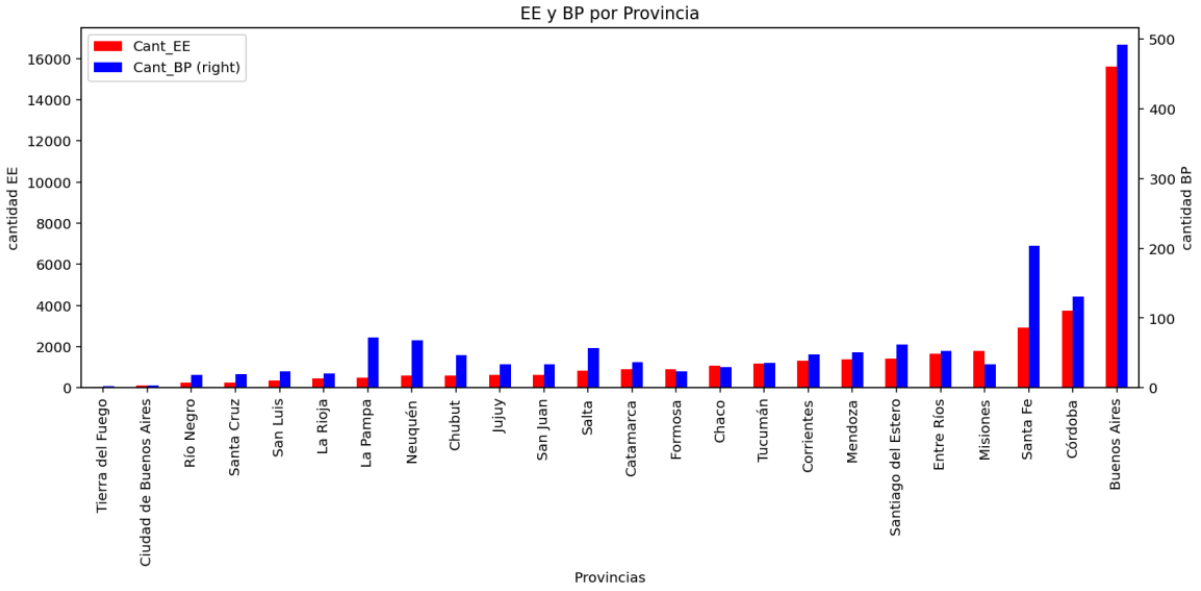


Figura 16. Bibliotecas populares y establecimientos educativos en función de provincias. A la izquierda se observa el eje de los establecimientos educativos y a la derecha el eje de las bibliotecas populares.

Este gráfico permite visualizar que a medida que aumenta una variable, no significa que la otra también aumente.