Trabajo Práctico I:

Bibliotecas Y Escuelas

Informe

**Laboratorio De Datos (DC), Comisión Laje**

**2025**

Autores:

* **Cuestas, Martín Iván**
* **Nakasone, Julián**
* **Poli, Dante**



## 

## 

## **Resumen**

El presente informe consiste en el análisis de tres bases de datos públicas proporcionadas por el gobierno argentino: Establecimientos Educativos (EE), Bibliotecas Populares (BP) y Padrón Población (PP). Se propuso como objetivo el investigar las relaciones que podrían llegar a existir entre estos tres parámetros mencionados anteriormente, más en particular hallar respuestas a la pregunta: ¿Influye la cantidad de población de cierta provincia o departamento en la presencia de bibliotecas populares o establecimientos educativos?, y cuestiones derivadas de la misma.

Con este propósito, primero se realizó una evaluación de las *Formas Normales* (EE: *1FN*, BP: *2FN*), seguido de un diagnóstico de la calidad de las bases de datos a utilizar, particularmente para EE y BP; para cada uno se definió el tipo de problema padecido, y el atributo de calidad afectado por el mismo. Para respaldar tales observaciones se realizaron análisis *GQM*, es decir, métricas cuantitativas diseñadas para obtener un grado preciso de la gravedad del problema. Las observaciones a destacar para cada caso son las siguientes: En EE se destaca la falta de claridad en cuanto a la declaración de la falta de datos en sus atributos, siendo este un problema de “Modelado”. Para almacenar la modalidad y submodalidad a la que pertenece cada escuela, la tabla posee un sistema de variables dummy (1’s y vacíos) en cada instancia, descuidando la diferenciación entre una escuela que no tenga cierta modalidad o que el dato no haya sido guardado; añadiendo también la utilización excesiva de espacio en casilleros vacíos. En cuanto al atributo “Teléfono”, hay diversos símbolos que parecen indicar la falta del dato en cada instancia (17,36% inconcordancias del total de las instancias), pero no existe un consenso en cuál utilizar específicamente y tampoco diferenciación entre falta de información o que dicha escuela no posea teléfono en la realidad. También se hallaron problemas de instancia afectando la completitud, con mayor grado en “Mail” (20% faltantes del total de las instancias). En BP se encontró muy presente la falta de completitud, siendo este un problema de instancia, afectando mayormente a los atributos mail (46,27% faltantes del total de instancias), informacion\_adicional (100%), piso (100%), web (100%). Los últimos tres al ser completamente nulos se consideraron como un problema de modelado, de la misma forma que otros atributos con todas las instancias idénticas. Para cada caso se dieron sugerencias de cómo mejorar la situación.

Luego se construyó un Diagrama Entidad Relación (DER) y un Modelo Relacional derivado del mismo, con el propósito de modelar de la mejor forma el recorte de realidad que será de mayor utilidad para el trabajo a realizar, y que logre capturar de mejor manera el objetivo propuesto. Acto seguido, se procesaron las bases de datos originales convenientemente; tomando decisiones, descartando elementos y creando nuevos a partir de otros. Las decisiones más importantes a destacar son la toma de Modalidad escolar “Común” como la única de interés en el caso, y la agrupación de población por departamento por “Rango Etario”, en concordancia con las edades usuales de los asistentes a cada tipo de EE. Con las bases ya procesadas, se construyeron Data Frames en base al Modelo Relacional.

Luego de preparar todo, se realizaron análisis de los datos a través de consultas SQL.

El análisis continúa de forma gráfica, donde los patrones pudieron ser distinguidos de forma más clara. Se logró encontrar una relación entre la cantidad de personas habitantes de cada departamento y provincia, y la cantidad de EE y BP. En el caso de las provincias, en los gráficos (*V*), (*VI*) y (*VII*) se puede observar un claro aumento de la cantidad de escuelas en base a la cantidad de población por cada sitio. Lo mismo sucede “en general” con las Bibliotecas Populares, salvo que en este último caso se encuentran discrepancias en el crecimiento proporcional de ambos parámetros, siendo “La Pampa” y “CABA” ejemplos destacables, la primera por tener una población casi mínima en comparación a las demás provincias (puesto 22), pero igualmente estar alta en el ranking de bibliotecas populares (4ta), y la segunda por tener una población casi máxima (3era) y estar a su vez baja en el ranking (puesto 16). Por estas razones se conjeturó que los EE siguen esta relación de forma más concisa debido a que las modalidades tomadas como foco de este trabajo son algo en la mayoría de los casos “obligatorio” para la población de las correspondientes edades, por lo que deben existir suficientes establecimientos para cubrir esta necesidad; por otro lado las bibliotecas populares son creaciones vecinales, por lo que no deben necesariamente existir en base a su población, más mientras mayor población, mayor cantidad de vecinos que pueden organizarse para crearlas (sin contar diversos parámetros que de seguro influyen en esto). En el gráfico (*VIII*) y (I*X*) esta relación se corrobora, siendo más fuerte en establecimientos educativos que en bibliotecas populares. En el gráfico (*X*) se describieron los distintos comportamientos de la cantidad de escuelas por departamento en cuanto a la distribución de los datos en base a sus cuartiles y outliers, y se encontraron comportamientos similares a los mencionados anteriormente. En el (*XI*) y (*XII*), al analizar la relación entre cantidad de EE y BP cada 1000 habitantes por departamento, se pudo observar que en general las escuelas tienen mayor amplitud en el rango de los datos posibles, y sus valores son bastante mayores que en el caso de las bibliotecas. También, se halló que los departamentos con mayor población tienen menos EE y BP cada 1000 habitantes. En conclusión, se pudo encontrar una aparente relación en donde a mayor población en cada departamento y cada provincia, mayor es la cantidad total de escuelas públicas, y mayor es la cantidad de bibliotecas populares, aunque en el segundo caso este patrón no es tan claro como en el primero, y es más inconsistente, a pesar de que en general ocurre. Por otro lado, es notable que si se miden estas cantidades en cada departamento normalizarlas con muestras de 1000 habitantes, las poblaciones mayores poseen menos establecimientos de ambos tipos para cada grupo, y viceversa. Asimismo, se encontró una leve relación, no del todo consistente, entre la cantidad de BP y EE, siendo que la primera parece tener una tendencia a crecer levemente junto a la segunda, tal vez debido al parámetro en común que ambos elementos comparten: población.

## **Introducción**

Existen múltiples bases de datos públicas aportadas por el estado, obtenidas de la sociedad, que

le pertenecen en su conjunto y reflejan sus comportamientos. Este trabajo va a tener sus cimientos en la información disponible sobre Establecimientos Educativos (El artículo 122 de la Ley 26.206 estipula que la Institución Educativa es la unidad pedagógica del sistema responsable de los procesos de enseñanza-aprendizaje destinados al logro de los objetivos establecidos por dicha ley. - Ministerio de Capital Humano), Bibliotecas Populares (asociaciones civiles creadas por la iniciativa de grupos de vecinos de una comunidad, que ofrecen servicios y espacios de consulta, expresión y desarrollo de actividades culturales, de la lectura y de extensión bibliotecaria en forma amplia, libre y pluralista. - Secretaría de Cultura), y Padrón Población, datos tomados por el Censo Nacional de Población, Hogares y Viviendas - 2022.

El objetivo principal es encontrar relaciones entre la presencia de EE y BP en ciertas regiones y la cantidad de personas que viven allí. Para esta finalidad, se llevará a cabo como comienzo el planteamiento de un “Diagrama Entidad-Relación” junto con su “Modelo Relacional” correspondiente, a la vez que un diagnóstico de las bases de datos originales, para luego manipularlas y acercarlas a los modelos planteados. Estos modelos convertidos en Data Frames serán analizados y se realizarán estudios visuales para la búsqueda de relaciones y patrones subyacentes en los datos; con esto se obtendrán conclusiones y se buscará dar respuesta al objetivo planteado.

Todos los procedimientos relacionados a la manipulación de datos serán adjuntados en un script complementario.

## **Procesamiento de Datos**

### Bases de datos a utilizar:

* Padrón Oficial de Establecimientos Educativos - 2022

<https://www.argentina.gob.ar/educacion/evaluacion-e-informacion-educativa/padron-oficial-de-establecimientos-educativos>

* Bibliotecas Populares - (Último cambio 2024)

[https://datos.gob.ar/dataset/cultura-mapa-cultural-espacios-culturales/archivo/cultura\_01c6c048](https://datos.gob.ar/dataset/cultura-mapa-cultural-espacios-culturales/archivo/cultura_01c6c048-dbeb-44e0-8efa-6944f73715d7)

[-dbeb-44e0-8efa-6944f73715d7](https://datos.gob.ar/dataset/cultura-mapa-cultural-espacios-culturales/archivo/cultura_01c6c048-dbeb-44e0-8efa-6944f73715d7)

* Padrón Población - 2022

*padron\_poblacion.xlsX* (Dado en Enunciado)

### Formas Normales de datos originales:

Definiciones generales:

* *1FN*: Prohíbe relaciones dentro de relaciones o relaciones como valores de atributo dentro de tuplas. El dominio de un atributo debe incluir sólo valores atómicos (simples e indivisibles).
* *2FN*: Un esquema R está en *2FN* si todo atributo *No Primo A* de *R* no depende parcialmente (de manera funcional) de ninguna clave de *R*.
* *3FN*: Un esquema *R* está en *3FN* si, para toda dependencia funcional no trivial *X* → *A* de *R*, se cumple alguna de las siguientes condiciones.

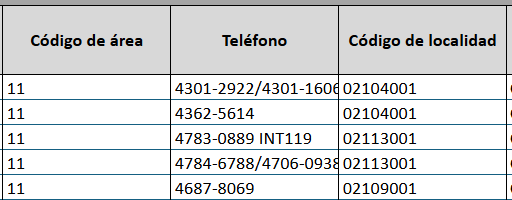
1. *X* es *SK* de *R.*
2. *A* es *Atributo Primo* de *R.*

Listas las definiciones, se pasa a analizar cada relación.

Para verificar en qué *Forma Normal* se encuentran las relaciones, debe verse progresivamente, es decir, de uno en uno, debido a que para que esté en cierto *FN*, se requiere a su vez que esté en el anterior, y por consecuencia si no cumple los requisitos de uno, no podrá nunca estar en el siguiente.

Padrón Oficial de Establecimientos Educativos - 2022

*1FN:* No cumple con las condiciones, ya que en el atributo “Teléfono” no siempre hay valores atómicos, en múltiples casos se encuentran más de un número de teléfono por instancia separados por “/”.



*Captura de la tabla original “2022\_padrón\_oficial\_establecimientos\_educativos.” (I)*

Luego, por consecuencia, no podrá ser *2FN* ni ninguna *Forma Normal* mayor.

Bibliotecas Populares

*1FN:* Cumple con los requisitos, no hay relaciones dentro de relaciones o relaciones como valores de atributos dentro de tuplas.

Para las siguientes *Formas Normales* debemos definir las *Dependencias Funcionales (DF)* presentes en la relación.

* *DF1*: nro\_conabip → {fecha\_fundacion, longitud, latitud, mail, teléfono, domicilio, nombre}
* *DF2*: domicilio → cp
* *DF3*: localidad → cod\_tel
* *DF4*: cod\_localidad → localidad
* *DF5*: id\_departamento → departamento
* *DF6*: id\_provincia → provincia
* *DF7*: id\_departamento → id\_provincia

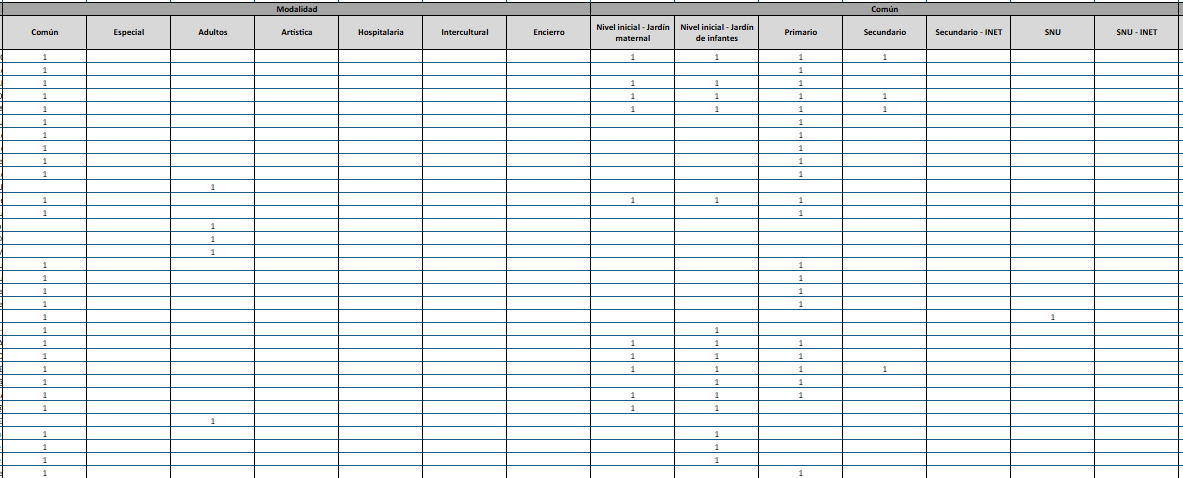
*2FN:* Luego de ver las *DF*, se puede concluir que la relación se encuentra en *2FN*, porque todo atributo depende completamente de *PK* (nro\_conabip), o directamente no depende de ese mismo.

*3FN*: En este caso, se pueden mencionar a *DF2, DF3, DF4, DF5, DF6, DF7* como contraejemplos para ver que no se encuentra en *3FN*. Todas estas dependencias se tratan de atributos no primos que dependen de atributos que no forman una *SuperClave.*

### Calidad de Datos

Padrón oficial de Establecimientos Educativos -2022

Esta base de datos presenta un problema de calidad relacionado al *Modelo* de la relación. La “Modalidad” de cada establecimiento se encuentra divida en 6 distintas secciones, “Común”, “Especial”, “Adultos”, “Artística”, “Hospitalaria”, “Intercultural”, “Encierro”; donde cada instancia de cada atributo tiene un 1 si dicho establecimiento tiene efectivamente dicha modalidad, o un espacio vacío si no la tiene. A su vez, hay 7 atributos asociados a la modalidad “Común”, 3 a “Artística”, 5 a “Especial”, 5 a “Adultos” y 3 a “Hospitalaria”, como así un atributo restante para “Servicios complementarios”; de la misma forma, cada instancia tiene un 1 o vacío dependiendo de si dicho establecimiento cumple con cada atributo. El riesgo que esto trae, es que es indistinguible si un establecimiento no posee una modalidad o submodalidad en específico, o si el dato no fue tomado, es decir que exista un dato faltante para modelar la realidad de forma precisa.



*Captura de fragmento de “Padrón oficial de Establecimientos Educativos - 2022.” a modo de ejemplo. (II)*

Por esto mismo, es recomendable establecer una distinción entre un Null o un “no posee dicha modalidad”, que podría ser usando un 1 para las escuelas que cumplan con la modalidad, y un 0 para las que no.

Otro problema que presenta es la falta de completitud de ciertos atributos, en particular “Domicilio”, “C.P.”, “Código de Área”, “Teléfono”, y “Mail”. La naturaleza de este problema está asociada a la *Instancia*, debido a que el modelo permite la obtención de dichos datos, pero de todas formas no se encuentran presentes. Vale la pena agregar, que no queda claro si la falta de dato representa falta de información sobre la realidad, o que dicho establecimiento de verdad no tiene dicho atributo asociado, lo que podría considerarse a su vez en parte un problema de *modelo,* ya que no se tiene una forma de especificar esta importante diferencia.

Para evaluar de forma más precisa y cuantificable la magnitud de este problema, se elaboran métricas de calidad *GQM*.

* *Goal:* Los atributos “Domicilio”, “C.P.”, “Teléfono”, “Código de Área”, y “Mail” están completos.
* *Question:* ¿Cuál es el porcentaje de celdas que no aportan información sobre dichos atributos? Uno por cada atributo.
* Metric: Se recorre cada atributo mencionado en *Goal*, y se cuentan cuántas instancias no tienen datos (buscando NULLs), luego el total es dividido por la cantidad de instancias totales en la relación, y se multiplica por 100 para tener un porcentaje. De esta forma se consigue una métrica para cada atributo que va del 0 al 100 (Ambos inclusive, siendo 0 más completo y 100 más incompleto) cuyo valor es un grado preciso de la No-Completitud de cada atributo.

Resultados (Métricas, porcentajes de Incompletitud) por atributo:

“Domicilio”: 1,44%

“C.P.”: 4,21%

“Teléfono”: 6,14%

“Código de Área”: 9,93%

“Mail”: 19,12%

Por diferencia, “Mail” es el atributo con mayor cantidad de espacios vacíos, seguido por “Código de Área”, “Teléfono”, “C.P.”, y “Domicilio”. Para solucionar esto, en los casos más incompletos como “Mail”, se podría armar directamente otra tabla que cuente con el cuanexo de cada establecimiento con mail asociado, evitando así tener que usar un atributo “Mail” en la tabla principal con casi 20% de incompletitud. Sin embargo, esta decisión queda a merced del uso que se le quiera dar a la relación, y a cómo se espera operar con ella a futuro. También podría considerarse hacer uso de un símbolo para diferenciar la falta de información con la no-existencia de esta en la realidad. Cabe considerar que el atributo “Teléfono” parece tener otras formas de demostrar su falta de información (ya sea por no tenerla o no existir) aunque sean deficientes; sin embargo para esta métrica sólo se consideran los espacios vacíos como incompletitud, debido a que parece no existir un consenso de cómo expresar dicha ausencia de datos de otra forma, por lo que tal vez se tomaría la falta de teléfono en la realidad como dato faltante, lo cual no sería del todo cierto. Lo que lleva al siguiente problema.

Si bien múltiples instancias poseen de forma correcta uno o más teléfonos (separados por “/”), en otras se encuentran almacenados otros tipos de símbolos, tales como “s/n”, “0”, “1”, “S/N”, “-”, “sn”, y “s/inf. Parecen ser formas de indicar, tal vez, la falta de posesión de dicho dato, es decir, que no se sabe cuál es el o los teléfonos de dicho establecimiento; o también, que dicho establecimiento directamente no tiene un número de teléfono. Sin embargo, es claro que no existe un consenso en cómo representar esta falta de información, por lo que resulta en algo confuso y sin demasiada consistencia.

Es conveniente realizar un análisis *GQM* para cuantificar esta falta de consistencia.

* *Goal*: El atributo “Teléfono” es consistente.
* *Question*: ¿Qué porcentaje de inconsistencias hay en relación a la cantidad total de instancias de “Teléfono”?
* *Metric*: Se hace un conteo de todos los valores de “Teléfono” que no sean números de teléfono (debido a que no se sabe lo que significa cada símbolo), ya sean únicos o separados y se divide por la cantidad total de instancias de la relación. Finalmente se

multiplica por 100 para tener un porcentaje.

Resultados (Métrica, porcentaje de inconsistencias de “Teléfono”):

Se obtuvo un valor de 17,36%. Hay un grado relativamente alto de inconsistencias en cuanto al número de teléfono, casi ⅕ de “Teléfono” son símbolos o espacios vacíos sin significado claro, es decir que sus instancias son poco homogéneas, haciendo el análisis de datos más complejo e inefectivo. Podría solucionarse eligiendo a conciencia un único símbolo específico para representar la falta de datos, y otro para expresar que el establecimiento no tiene teléfono.

Bibliotecas Populares

Un problema muy presente en esta base de datos es la falta de completitud de ciertos atributos, en particular “mail”, “cod\_tel”, “telefono”, “informacion\_adicional”, “piso”, y “web”, donde algunos a simple vista tienen una gran cantidad de espacios vacíos, mientras que otros parecen directamente no tener ni un dato almacenado. Es un problema asociado a la *Instancia,* ya que el modelo permite el almacenamiento de todos estos datos, pero por alguna razón, no se encuentran almacenados, afectando el atributo de calidad “completitud”. A su vez, no se sabe si el dato efectivamente no se encuentra almacenado, o si en la realidad ese dato no existe, el modelo no permite dicha distinción.

Para evaluar de forma más precisa y cuantificable la magnitud de este problema, se elaboran métricas de calidad *GQM*.

* *Goal:* Los atributos “mail”, “cod\_tel”, “telefono”, “informacion\_adicional”, “piso”, y “web” están completos.
* *Question:* ¿Cuál es el porcentaje de celdas que no aportan información sobre dichos atributos? Uno por cada atributo.
* Metric: Se recorre cada atributo mencionado en *Goal*, y se cuentan cuántas instancias no tienen datos (buscando NULLs), luego el total es dividido por la cantidad de instancias totales en la relación, y se multiplica por 100 para tener un porcentaje. De esta forma se consigue una métrica para cada atributo que va del 0 al 100 (Ambos inclusive, siendo 0 más completo y 100 más incompleto), cuyo valor es un grado preciso de la No-Completitud de cada atributo.

Resultados (Métricas, porcentajes de Incompletitud) por atributo:

“mail”: 46,27%

“cod\_tel”: 8,68%

“telefono”: 8,68%

“informacion\_adicional”: 100%

“piso”: 100%

“web”: 100%

Sobre la completitud de "mail", por la métrica obtenida (46,27%, redondeada a dos decimales), se puede decir con certeza que "casi" la mitad de las instancias del atributo "mail" están vacías, dejando claro que está presente un problema grave en el atributo de calidad completitud. Para evitar los Null se podría construir otra relación en donde estén presentes solamente las bibliotecas con mails asociados, y sus respectivas direcciones de correo. Otra forma es dejarlo así y filtrar los datos por Null o No-Null. Todo depende de lo que se esté buscando realizar.

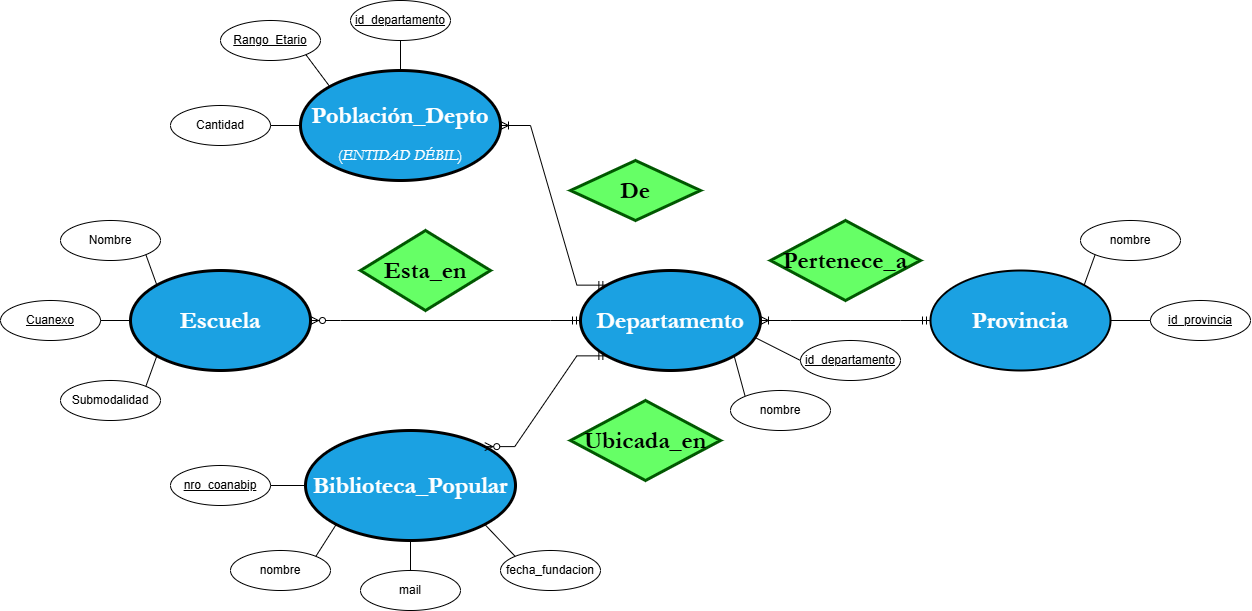
Sobre "cod\_tel" y "telefono", la métrica obtenida fue 8,68%, redondeada a dos decimales, por lo que se podría decir que la mayoría de instancias están completas. Este atributo podría quedar así sin tener grandes problemas, en general, aunque siempre es preferible evitar Nulls. En cuanto a "web" y "piso", como se mencionó anteriormente están completamente vacías (100% en la métrica), por lo que es recomendable eliminarlas de la relación en el caso de que no se esté planeando realizar algo con dicho atributo.

En estos casos podría también considerarse el uso de algún símbolo en concreto para diferenciar la falta de dato con la no-existencia del dato en la realidad.

Un problema similar al de los atributos vacíos es la presencia de atributos redundantes, es decir columnas en donde todas sus instancias son exactamente las mismas, almacenando así información sin sentido aparente. Esto parece indicar un problema asociado en gran parte al *modelo*, ya que se debería haber previsto que estas columnas iban a ser o completamente nulas, o completamente redundantes, por lo que debería haber una muy buena justificación para sostener esta decisión de modelado. Una recomendación es eliminar dichos atributos para evitar la redundancia.

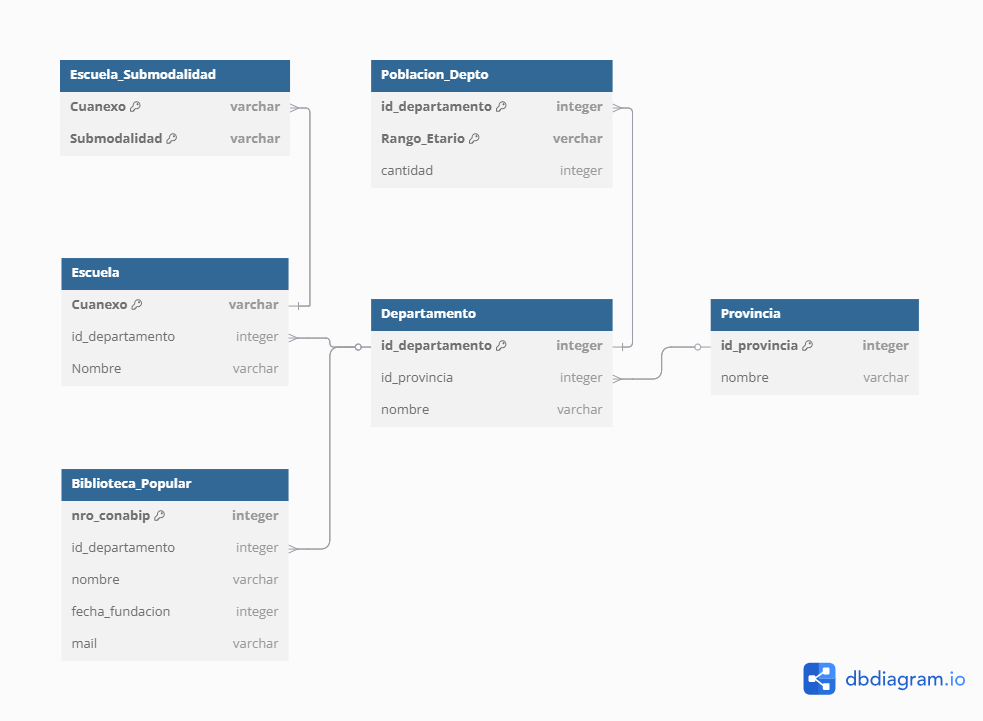
### Diagrama Entidad Relación

Se plantea y construye un Diagrama Entidad Relación (*DER*) con el objetivo de modelar el recorte de realidad que será necesario para responder a las preguntas de interés en el presente informe, seleccionando los fragmentos de las bases de datos originales necesarios, y descartando los irrelevantes.

*Diagrama Entidad Relación - Hecho en draw.io (III)*

### Modelo Relacional

A partir del *DER* se crean las tablas del *Modelo Relacional*, con el propósito de definir cómo serán los *Data Frames* a utilizar, incluyendo *Foreign Keys (FK), Primary Keys ( PK ),* y *Dependencias Funcionales (DF).*Para el proceso de conversión, en general para cada entidad del DER se establece una Relación (una tabla) con el mismo nombre y con los mismos atributos, sumadas las *Foreign Keys* que son colocadas en base a las relaciones entre entidades y su naturaleza. En este caso, al ser todas del tipo de cardinalidad (Muchos - Uno), y el tipo de participación se coloca la *PK* de la entidad que va a estar relacionada necesariamente con cada uno de los atributos de la otra (cada Poblacion\_Depto tiene necesariamente un único Depto, cada Depto necesariamente una única provincia, cada BP necesariamente un solo Depto). En el caso de Escuela, al ser el atributo “Submodalidad” compuesto (no atómico) y no ser *1FN*, se crea una nueva tabla con el nombre de la entidad más el del atributo (Escuela\_Submodalidad), para así convertirlo en una relación *3FN*.

**

*Modelo Relacional, las flechas representan Dependencias Funcionales (Ignorar símbolos en extremos) - Hecho en dbdiagram.io (IV)*

### Formas Normales de Modelo Relacional

*1FN* : Todas las relaciones se encuentran en *1FN* porque no hay relaciones dentro de relaciones o relaciones como valores de atributos dentro de tuplas, todos los valores son atómicos.

Dependencias Funcionales dentro de cada Relación:

Poblacion Depto:

* *1DF* {Rango\_Etario, id\_departamento} → Cantidad

Biblioteca\_Popular:

* *1DF* nro\_conabip → nombre
* *2DF* nro\_conabip → id\_departamento
* *3DF* nro\_conabip → fecha\_fundacion
* *4DF* nro\_conabip → mail

Escuela:

* *1DF* Cuanexo → id\_departamento
* *2DF* Cuanexo → Nombre

Escuela\_Submodalidad:

No tiene.

Departamento:

* *1FN* id\_departamento → id\_provincia
* *2FN* id\_departamento → nombre

Provincia:

* *1FN* id\_provincia → nombre

Viendo una a una cada *DF*, se puede concluir que las relaciones están en *2FN* , y en *3FN* . Todo atributo no primo depende de una superclave (*PK* en particular en todos los casos).

### Manipulación de las Bases de Datos Originales y Decisiones Tomadas / Generado de Data Frames en Base al Modelo Relacional

Con el fin de llegar a crear los Data Frames en base al Modelo Relacional de la forma más efectiva y ordenada posible, se realizan modificaciones en las bases de datos originales, y se crea una nueva a partir de otra base de datos oficial: Padrón población (Base: Viviendas particulares).

Se utilizan para esta tarea las librerías “Pandas”, y “SQL” de “DuckDB”.

Padrón Oficial de Establecimientos Educativos - 2022

Para las relaciones del Modelo Relacional “Escuela”, y “Escuela\_Submodalidad”, la modalidad de escuela con relevancia es la “Común”, el resto de información que respecta a las demás se considera irrelevante para el presente trabajo y, por lo tanto, se descarta sin inconvenientes. Para la primera, con consultas SQL se filtran los atributos necesarios (Cuanexo, id\_departamento, Nombre). En el caso de Id\_departamento, este coincide con los primeros 5 dígitos del atributo Código de Localidad, a excepción de los códigos de CABA, para los cuales hay que cambiar el tercer dígito por un “0”. Tolhuin no fue incluido debido a que no posee modalidad “Común”.  
 Para la segunda relación, se recorre instancia por instancia la tabla original modificada y se crea una nueva relación, donde cada instancia es un Cuanexo asociado a una submodalidad (si esa submodalidad de la modalidad “Común” tiene un 1 asociado a ese Cuanexo en la tabla original).

Por otro lado, a partir del Data Frame original sin modificar se obtienen las tablas del Modelo Relacional “Provincia” y “Departamento”, debido a que se calculó mediante consultas que esta es la base de datos con más información sobre este tópico entre las utilizadas en el trabajo, cubriendo a todas las demás. Para la primera tabla se toman los primeros 2 dígitos de Código de Localidad como “id\_provincia”, y Jurisdicción como “nombre”; para el segundo se toman los mismos id\_departamento asignados a “Escuela” (agregando Tolhuin), los mismos id\_provincia asignados a “Provincia”, y el nombre asociado a dicho id\_departamento en la base de datos original. También se decide incluir a “CABA” como un departamento adicional en la tabla “Departamento”, a pesar de no ser técnicamente un departamento y sus comunas ya estar incluidas en la tabla. Sin embargo, es de gran utilidad poder hacer uso de su información por separado debido a que en la tabla original Bibliotecas Populares CABA no tiene distinción por comuna, por lo que sus datos pueden ser obtenidos únicamente de esta forma, no por comuna.

Ver procesos más detallados en script.

Bibliotecas Populares

Se obtiene Biblioteca Popular del Modelo Relacional mediante un filtrado de atributos (nro\_conabip, id\_departamento, nombre, fecha\_fundacion, mail) de la base de datos original. Las fechas son convertidas al tipo de dato string para facilitar su manipulación, y el id\_departamento de Chascomús es cambiado de “6217” a “6218” para que coincida con los id de las demás bases de datos.

Padrón población (Base: Viviendas particulares)

De esta base de datos primero se recolectan los id\_departamento a partir de los códigos de “Área”. En general, el id\_departamento coincide con Área #, a excepción de algunos. En cuanto a los pertenecientes a CABA, los últimos 2 dígitos de cada código se deben dividir por 7, y ser reemplazados por el resultado (todos los resultados son enteros, pues los últimos 2 dígitos son múltiplos de 7). En los casos de 'Río Grande', 'Tolhuin' o 'Ushuaia', se decide reemplazar sus códigos de área por '94007', '94011' y '94014', para que coincidan con los id de departamento de la tabla de “Departamento”, que es tomada como referencia.

Una vez identificados los identificadores departamentales, se recorren todas las instancias de cantidad de población por edad y se agrupan en rangos etarios elegidos de forma conveniente para la resolución de este trabajo, en base a las submodalidades de las escuelas y que rango de edades correspondientes a cada nivel educativo.

## **Análisis de Datos**

### Consultas SQL (Resultados en código)

Realizadas a partir de los data frames derivados del modelo relacional.

1) En esta consulta se integran las bases de datos sobre la población y los establecimientos educativos por cada departamento del país, destacando especialmente el nivel inicial (Jardines), Primarios y Secundarios. Se generó una tabla conteniendo, según la provincia y departamento, la cantidad de jardines, primarias y secundarias registradas, además de la población para cada nivel según su rango etario.

El objetivo es comparar, para cada departamento, la cantidad de establecimientos educativos disponibles en relación con la población que debería poder acceder a cada nivel, permitiendo así identificar posibles desigualdades de la infraestructura educativa en relación a la población.

Se realiza primero una consulta para unir los departamentos y provincias; luego a partir de la unión entre escuelas y submodalidades, se cuentan las instituciones por departamento para cada uno de los niveles, y usando los rangos etarios de poblacion\_depto, se asigna cada rango etario al nivel educativo correspondiente, y por último se unen todos los datos anteriores en una única tabla.

2) En este caso, el objetivo consta de comparar la cantidad de bibliotecas populares fundadas desde 1950 entre diferentes departamentos de la misma provincia. Para ello, se realiza una consulta que permita obtener únicamente aquellas bibliotecas populares que cumplan con el requisito de antigüedad. Luego se cuentan por departamento y, mediante los identificadores departamentales, se relacionan con una tabla que contiene todos los departamentos con los que contamos, incluidos aquellos de los cuales no se tiene información correspondiente a sus respectivas bibliotecas populares, dejándolos como valores NULL.

3) Se quiere realizar una consulta con la cantidad de establecimientos educativos, bibliotecas populares y la población por cada departamento. Al no tener los datos de bibliotecas populares por cada departamento (Comuna) de la Ciudad de Buenos Aires, se decide agregar a esta última como otra instancia más de la tabla del resultado; debido a que los dato que sí se poseen por estar así definido en la base de datos original BP, son los correspondientes a la totalidad de CABA (como un solo departamento). En esta nueva instancia, se suman todos los datos tanto de establecimientos educativos como de la población de todos los departamentos de la Ciudad de Buenos Aires.

4) Con el objetivo analizar el uso de correos electrónicos por parte de las Bibliotecas Populares (BP) de cada departamento, se realiza una consulta con la que se identifica la provincia correspondiente, el departamento y el dominio de correo electrónico más frecuente entre las bibliotecas que declararon un mail de contacto.   
Es conveniente agregar dos columnas adicionales que permitan contextualizar mejor la situación de cada departamento, al intentar reflejar la completitud de los datos proporcionados, y cómo esto puede influir en los resultados obtenidos en relación al mail.

* cantidad\_bibliotecas: indicando cuántas BP hay en el respectivo departamento.
* porcentaje\_BP\_con\_Mail: Donde se muestra el porcentaje de BP en el departamento que tienen un mail declarado.

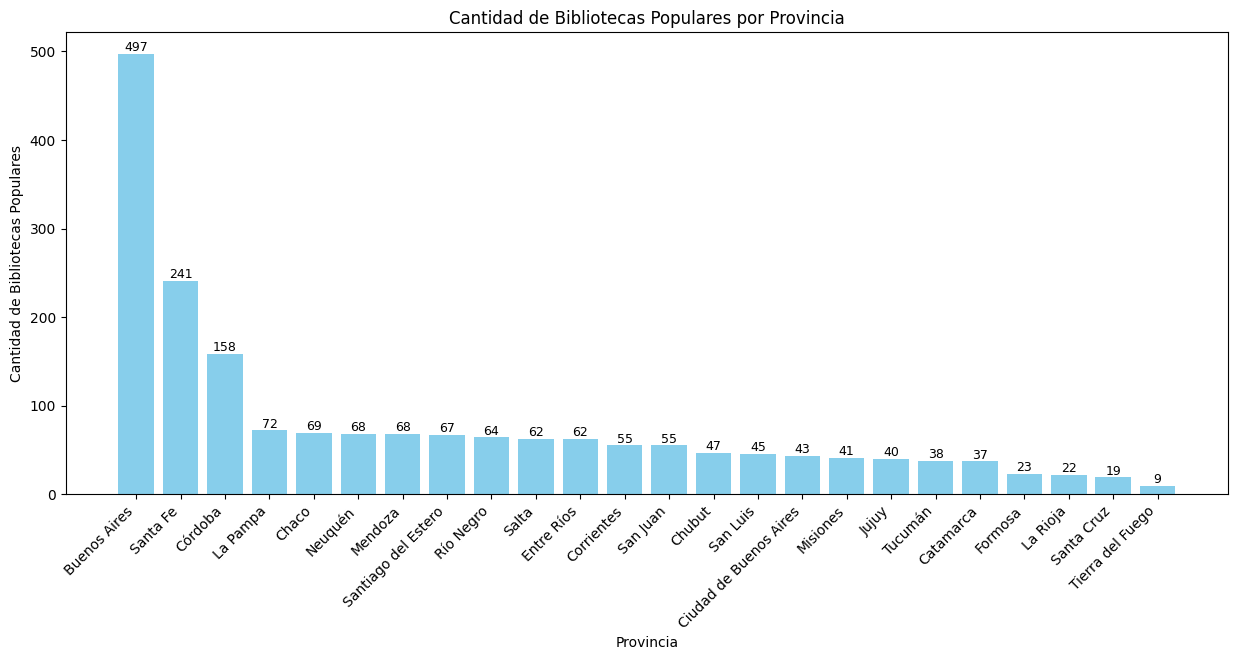
Estas mediciones complementarias permiten detectar departamentos donde pocas BP tienen contacto por correo electrónico, o donde no fueron registrados, teniendo de esta forma una muestra menor y convirtiendo el dominio más utilizado en un dato de menor “credibilidad”.

Gráficos

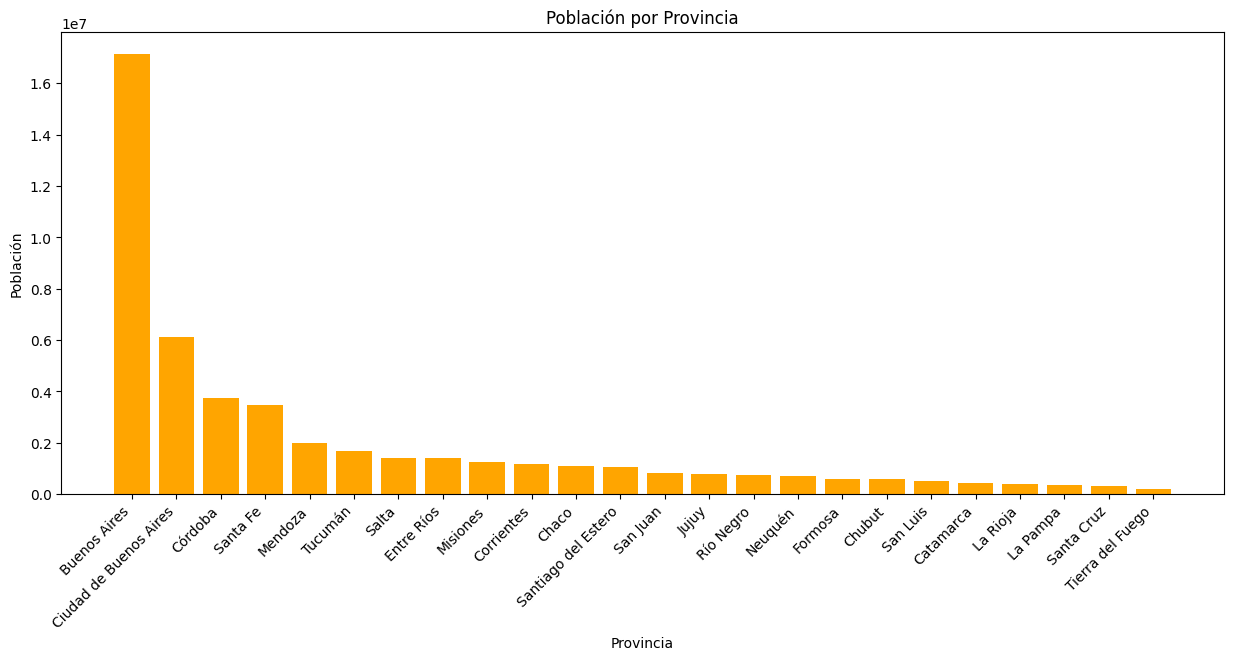
Sección de visualización de datos.

1)

Se busca conocer la cantidad de bibliotecas públicas establecidas por cada provincia, y que relaciones puede traer consigo este resultado.



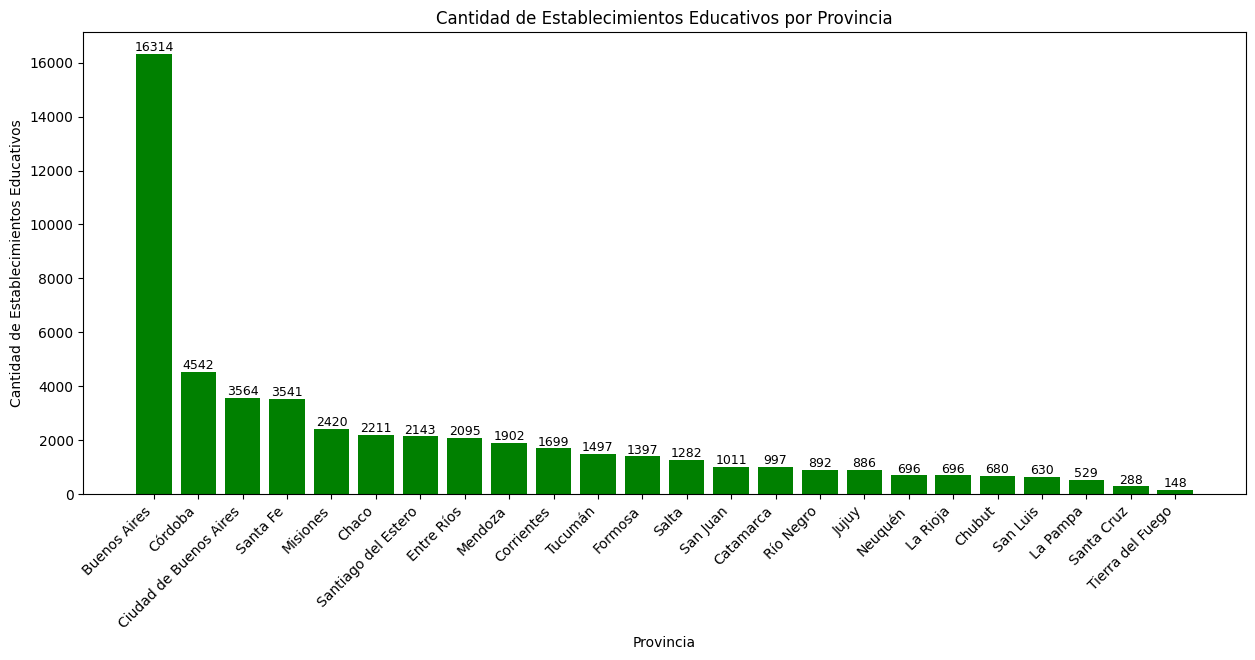
*Gráfico de barras, ordenado de forma descendiente por Cantidad de BP - Hecho con matplotlib (V)*

Se puede observar a simple vista en el gráfico obtenido una distinción clara en Buenos Aires, Santa Fe, y Córdoba, siendo estas tres las provincias con mayor cantidad de bibliotecas populares. Se puede conjeturar otra relación subyacente en los resultados obtenidos, y es que las tres provincias dichas anteriormente son las tres de las cuatro con mayor cantidad de Población (sin CABA), como puede observarse en el siguiente gráfico:  


*Gráfico de barras, ordenado de forma decreciente por Población - Hecho con matplotlib (VI)*

Puede verse también, que las provincias con menor cantidad de BP son aquellas con menor cantidad de habitantes, a excepción de La Pampa, que en el orden de población es anterior a la anteúltima, pero en cantidad de bibliotecas públicas se encuentra en el cuarto puesto. Otra excepción es CABA, que se encuentra segunda en población, pero 16ava en cantidad de BP. Estas excepciones son las más extremas en cuanto a discrepancia entre ambos parámetros, pero en general sucede también con otras provincias en menor medida y con menor discrepancia, lo que lleva a pensar que esta relación Población Provincia - Cant BP no es tan fuerte, si bien existe en general.

A su vez, se puede encontrar este mismo patrón con los Establecimientos Educativos (La Pampa y CABA dejan de ser excepciones), siendo observable en el siguiente gráfico:

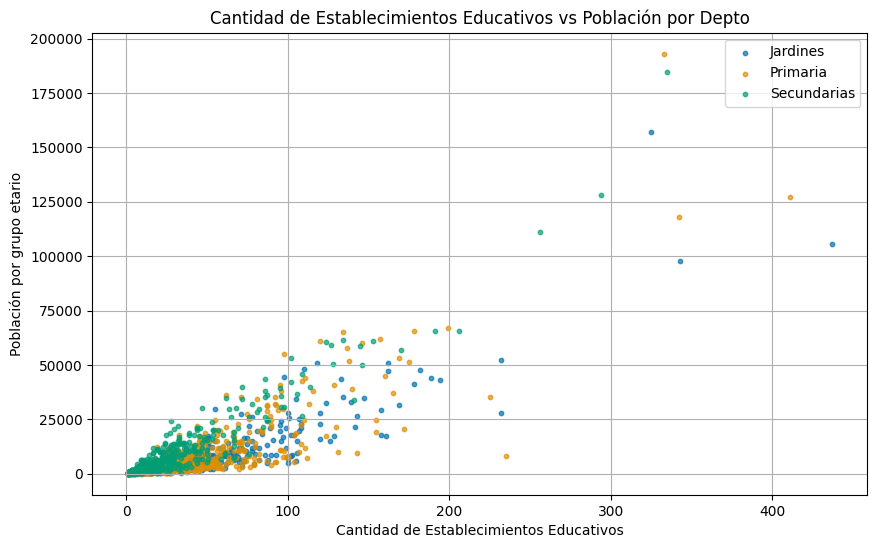


*Gráfico de barras, ordenado de forma descendiente por Cantidad de EE - Hecho con matplotlib (VII)*

Con estos tres resultados obtenidos, podría afirmarse con un alto nivel de seguridad que parece haber un patrón, donde la cantidad de escuelas públicas y bibliotecas populares suele ser mayor en provincias con mayor número de habitantes, y menor en el caso contrario. Esta relación es más clara en EE que en BP, y se puede conjeturar que es debido a que los establecimientos educativos son algo “necesario” y es menester que se adapten a la cantidad de habitantes debido a que la educación es algo obligatorio en la mayoría de las modalidades que se han tomado como foco de análisis en este informe; mientras que las Bibliotecas Populares son asociaciones civiles creadas por la comunidad y para la comunidad, por lo que no es obligatorio que existan acorde a la cantidad de población, pero mientras mayor es la población, mayor es la cantidad de vecinos disponibles para crearlas (sin contar diversos factores que de seguro influyen en el proceso).

2)

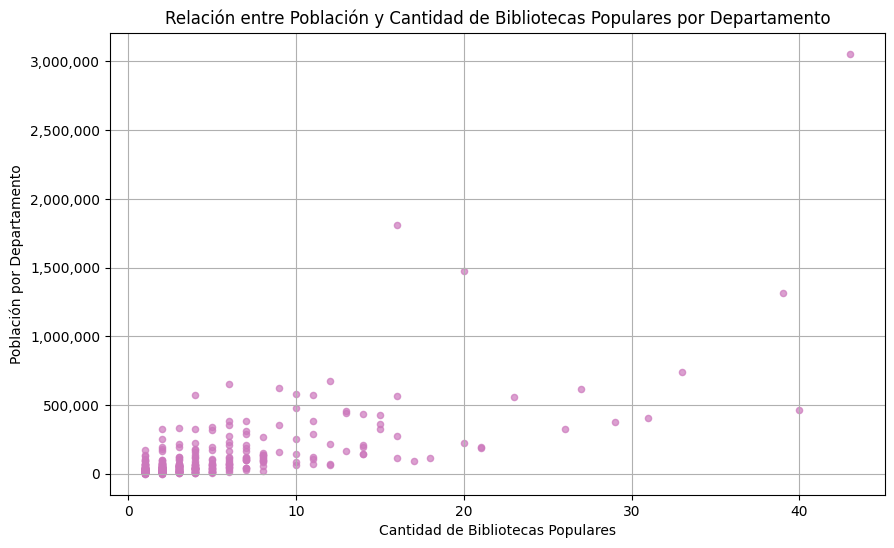
Para buscar relaciones entre la cantidad de establecimientos educativos por cada departamento, y su población, separada por rango etario y el nivel educativo correspondiente para mayor nivel de detalle, se realiza un scatter plot, o diagrama de dispersión.



*Diagrama de Dispersión, separado por color (colorblind friendly) según rango etario y nivel educativo - Hecho con matplotlib} (VIII)*

Se puede observar que hay una gran densidad de puntos de todas las tres clasificaciones en la parte inferior izquierda del gráfico, que es la zona en donde la cantidad de establecimientos educativos y la población es mínima. Luego, interpretando de izquierda a derecha parece haber un crecimiento proporcional de ambas variables, lo que estaría en concordancia con el anterior resultado (VI), que mostraba un patrón del mismo tipo, donde a mayor cantidad de habitantes, mayor cantidad de EE. En este caso la población se mide por departamento y en el anterior por provincia, pero los departamentos pertenecen a las provincias así que la relación entre ambas mediciones es compatible para ver tendencias generales.

Para ver si lo mismo sucede con las bibliotecas públicas se realiza otro gráfico.

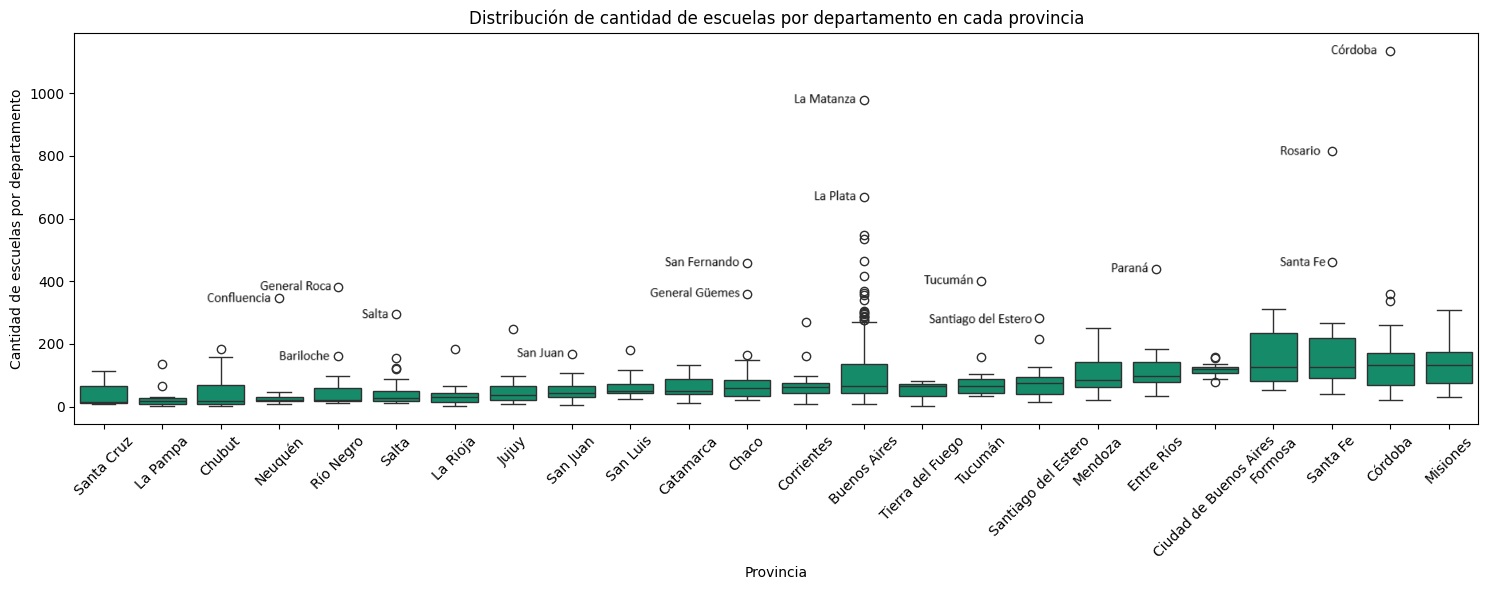


*Diagrama de Dispersión - Hecho con matplotlib} (IX)*

En este caso no se puede observar un patrón creciente proporcional entre ambos parámetros de forma tan clara como sucede con Escuelas. Hay una gran cantidad de puntos en el área inferior izquierda que crecen de forma aparentemente constante, es decir que crecen la cantidad de bibliotecas populares pero la población de los departamentos no aumenta necesariamente. Esto indica que hay muchos departamentos con igual población, pero cuya cantidad de BP varía considerablemente, yendo desde menos de 10 a 40 bibliotecas para deptos con población similar. Sin embargo, es importante decir también, que los puntos más elevados, es decir, los departamentos con mayor población, se encuentran hacia la derecha del gráfico, teniendo una gran cantidad de bibliotecas públicas en contraste con los demás, siguiendo la lógica de que a mayor cantidad de población suelen encontrarse mayores cantidades de BP (no necesariamente pasa lo contrario).

3)

Para observar la variación de la cantidad de escuelas de modalidad común por departamento en cada provincia, y detectar las provincias en las que haya alguna desigualdad notable en la distribución de las escuelas entre sus departamentos, se realiza el siguiente gráfico.

*Boxplot ordenado por mediana - Hecho con seaborn (X)*

Lo principal que destaca a simple vista es que Buenos Aires tiene muchos outliers, implicando que existen múltiples departamentos pertenecientes a esta provincia con picos extremos en cantidad de escuelas; lo mismo ocurre con la mayoría de las provincias, aunque en menor medida. Al identificar estos outliers, se

puede observar que coinciden con los departamentos con mayor población de cada provincia, lo que tiene sentido, considerando que la cantidad de establecimientos educativos debe ser la suficiente para garantizar el alojamiento de la cantidad esperada de posibles estudiantes.

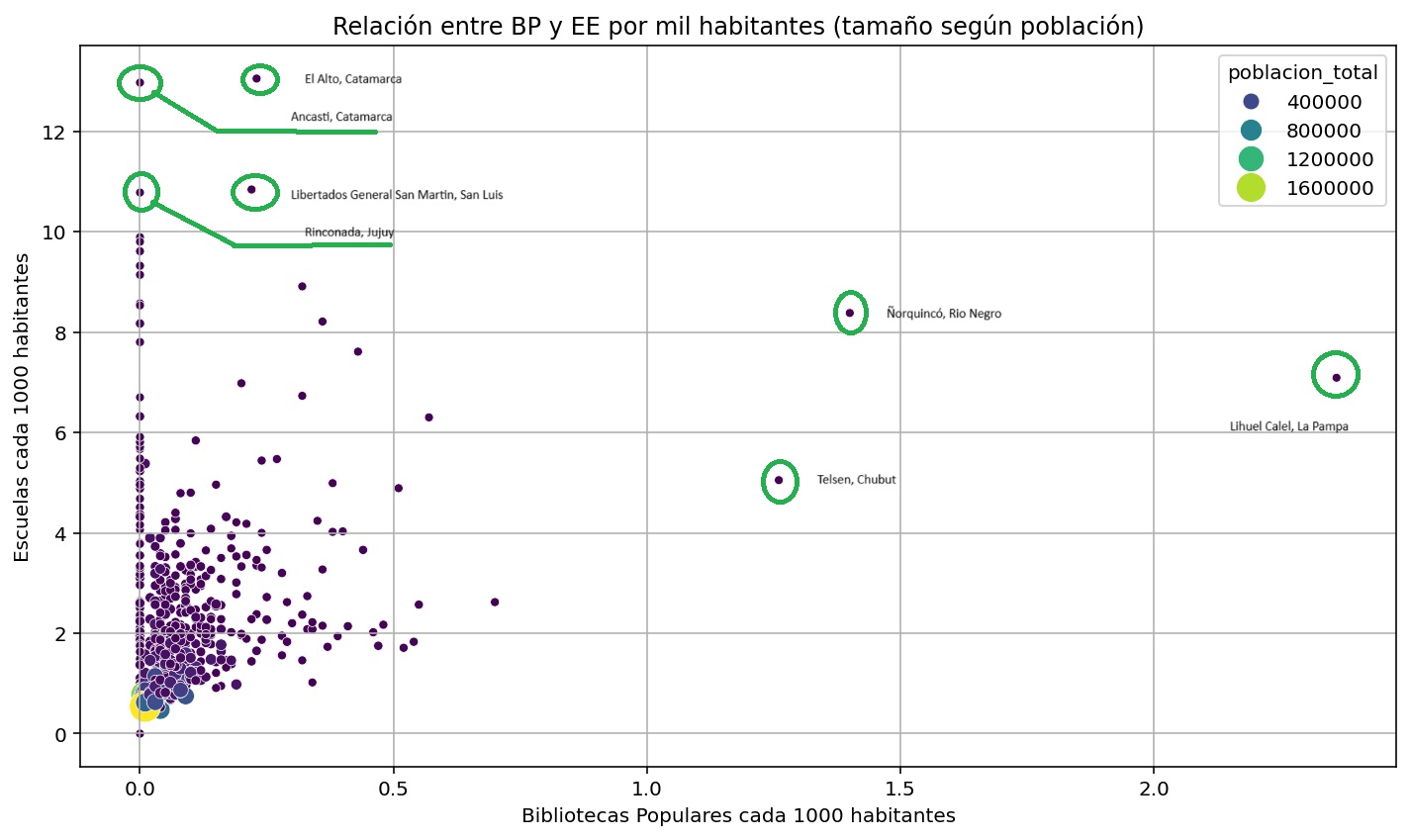
Analizando con más detalle el boxplot (visualización con gran cantidad de información), se pueden encontrar diversas cajas (provincias) con una gran dispersión de datos en la zona superior a la mediana (segundo cuartil), lo que es denominado asimetría positiva; estas son: Santa Fe, Formosa, Entre Ríos, Mendoza, Catamarca, San Luis, San Juan, Jujuy, Salta, Río Negro, Neuquén, Chubut y Santa Cruz. La información que esto ofrece, es que las cantidades de escuelas en dichas provincias suelen estar distribuidas de forma más consistente en los valores por debajo de la mediana, mientras que, por arriba, la variación es mayor.

Aquellas que tienen asimetría negativa son: Misiones, CABA, Santiago del Estero, Tierra del Fuego, Corrientes, y La Rioja. Esto indica una mayor dispersión de los datos por debajo de la mediana. Esto indica lo contrario al caso anterior, hay mayor variación por debajo de la mediana y mayor consistencia por arriba.

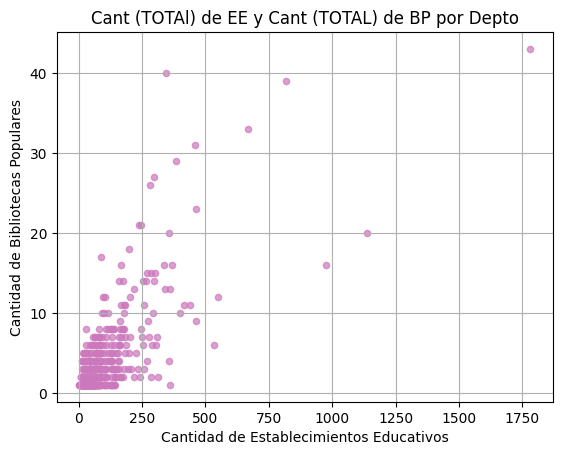
Por otro lado, Tucumán, Chaco, y La Pampa se acercan más a cierta simetría, es decir, que su mediana parece estar centrada.

En cuanto al tamaño de las cajas, se puede observar que Formosa, Santa Fe, Córdoba, Misiones, Buenos Aires son las que parecen tener un mayor tamaño, es decir mayor rango intercuartílico (entre tercer y primer cuartil), lo que demuestra una mayor variación en el valor de los datos. En términos de departamentos y cantidad de EE, esto quiere decir que hay departamentos con muchas más escuelas que otros, y viceversa.

4) Con el propósito de observar relaciones entre la cantidad de escuelas de modalidad común y bibliotecas públicas por habitantes en cada departamento, se normaliza utilizando muestras de 1000 habitantes, y se compara en un diagrama de dispersión.



*Diagrama de Dispersión, coloreado según cantidad de habitantes - Hecho con matplotlib (XI)*

**

*Diagrama de Dispersión - Hecho con matplotlib (XII)*

En cuanto al primer gráfico (*X*), se visualiza una gran concentración de puntos en la esquina inferior izquierda, en particular entre 0.5 y 4.5 Escuelas comunes cada 1000 habitantes, y 0.0 y 0.5 bibliotecas populares cada 1000 habitantes (más en particular, pegado al eje de escuelas). Las excepciones (por gran diferencia) son;

* Para bibliotecas Populares Ñorquincó, con aproximadamente 1.3 BP/1000 hab., Telsen con aproximadamente 1,25 BP/1000 hab., y Lihuel Calel, con casi 3 BP/1000 hab, siendo con amplia diferencia la mayor en este parámetro.
* Para escuelas comunes, Rinconada con aprox 11 EE/1000 hab., Libertador General San Martín con aprox 11 EE/1000 hab., Ancasti con aprox 13 EE/1000 hab. y El Alto, con aprox 13 EE/1000 hab.

En el caso de escuelas, se encuentran también (aunque en menor concentración y diferencia de la masa central), múltiples departamentos por encima del valor 4.5, es decir que hay una dispersión más grande que en el caso de las bibliotecas. Estos valores atípicos pueden deberse tanto a una gran cantidad de escuelas o bibliotecas, como a una baja cantidad de población; al ver que ningún departamento mencionado se trata de uno de mucha población (imagen (*X*)), es altamente posible que sea a causa de la segunda razón.

Todo esto permite observar que, en general, en cada departamento hay una cantidad mucho mayor de escuelas de modalidad común por cada 1000 habitantes, ya sea en concentración como en picos, que bibliotecas populares.

No podría decirse que exista una relación proporcional clara entre la cantidad de EE y BP que posee cada depto por cada grupo de personas, cualquiera sea la cantidad de escuelas, las bibliotecas están fuertemente acotadas por un umbral de 0,75, salvo las 3 excepciones mencionadas. Sin embargo, donde las proporciones de estas últimas tienden a ser más altas es entre los valores 1 y 7 de las EE/1000 hab.

Al tomar esta relación en cuanto a cantidades totales sin embargo (gráfico (*XI*)), se deja ver una leve inclinación a la derecha de los puntos, lo que habla de cierta relación tendencial en donde las bibliotecas aumentan ligeramente a medida que lo hacen las escuelas; fenómeno tal vez causado por el parámetro en común que poseen ambos elementos, el cual es la población, que tiene una relación ascendente más proporcional con ellos al tomar las cantidades de forma total, como fue mostrado en los gráficos previos.

Volviendo al gráfico anterior, los departamentos con mayor concentración poblacional están resaltados mediante tamaño y color, y se puede observar con claridad, que todos ellos se encuentran concentrados únicamente en la parte inferior izquierda del gráfico, lo que indica que las zonas más pobladas al compararlas con otros departamentos parecen tener menor cantidad de Escuelas de modalidad común y bibliotecas populares cuando es medido cada 1000 personas, a pesar de que en los anteriores resultados era al revés.

## **Conclusión**

En conclusión, tras analizar los datos obtenidos mediante los procesos detallados durante el informe, se pudieron encontrar relaciones subyacentes, y de esta forma dar respuesta a la pregunta troncal y disparadora: “¿Influye la cantidad de población de cierta provincia o departamento en la presencia de bibliotecas populares o establecimientos educativos?”. En definitiva, se observó que mientras mayor es la población de un departamento o una provincia, mayor es la cantidad total de establecimientos educativos. Este patrón fue claro y se pudo observar mediante diversos enfoques en visualizaciones. También sucede lo mismo con las bibliotecas públicas, pero en menor medida y de forma menos consistente. Si bien suelen aumentar en número al aumentar la población, existen una cantidad considerable de contraejemplos para que no sea algo “absoluto”, más bien una generalidad. No se encontró una relación clara y particular al comparar la cantidad de BP y EE por departamento; pudo verse que los valores de las escuelas pueden variar en gran amplitud, mientras que al mismo tiempo los de las bibliotecas permanecer dentro de umbrales similares. Sin embargo, las BP siguen cierta incipiente tendencia, tal vez debido al factor en común de ambos población, a aumentar en cantidad junto con los EE, aunque no sea abrupto, lineal ni en extremo directo. Como conjetura, este fenómeno podría deberse a la obligatoriedad de existencia de estos establecimientos en cada sector. Las BP son asociaciones civiles creadas por los vecinos, que, dicho sea, mientras más sean estos, mayor es el número de candidatos de posibles creadores o impulsores del proyecto (también influyen parámetros no representados en el informe, como cultura, compromiso social, contexto e historia de la zona); y las escuelas una necesidad imperativa y obligatoria en la mayoría de las submodalidades estudiadas en este informe (donde también influyen más parámetros). A su vez, es valioso mencionar que al hacer el análisis de BP y EE en cantidad no total, si no normalizada en grupos de población (por departamento, cada 1000 personas en este caso), puede observarse que a mayor población menor es la cantidad de estos establecimientos por cada uno de dichos grupos, y en zonas poco habitadas este número es altísimo en comparación a los demás, siendo esta una relación inversa a la que se estuvo mencionando. Se puede conjeturar que esto se debe a que una baja población eleva considerablemente esta proporción, aunque los establecimientos no sean demasiados, mientras que, en el caso contrario, la cantidad gigante de habitantes “diluye” esta medida; añadiendo que probablemente al tener poblaciones más dispersas y menor capacidad por escuela son necesarias más escuelas por habitante para poder dar los servicios adecuados a la población. Es lógico pensar que la forma de asegurarse de que las necesidades de la población están cubiertas no es por cantidad total de habitantes, si no por grupos normalizados, más en este análisis no fueron tomados otros parámetros de importancia, como puede ser la capacidad de los establecimientos, para llevar a mayor profundidad este tema y así comprender precisamente este fenómeno.