Laboratorio de Datos

Trabajo Práctico 1: escuelas y bibliotecas

21/5/2025

Club de fans de la bahía de Samborombón

Francisco Catri

Ioel Failenbogen

Tomás Benjamín Caula

Resumen

En este trabajo presentamos los productos de una investigación que tomó como objetivo la determinación de la existencia o carencia de una relación estrictamente numérica entre la cantidad de establecimientos educativos y la cantidad de bibliotecas populares en cada una de las provincias (y departamentos) del país. Tal relación fue abordada con la ayuda de diversos datos extraídos de las fuentes proporcionadas por los docentes, la cual implicó un entendimiento refinado del estado de la distribución de estas instituciones a lo largo del territorio nacional.

Introducción

El objetivo general de este trabajo fue resolver si existe alguna relación entre la cantidad de establecimientos educativos en cada una de las provincias de la República Argentina y la cantidad de bibliotecas populares en ellas. Para esto, contamos con tres fuentes de datos, siendo cada una de ellas en definitiva una tabla que relaciona distintas métricas de cada departamento del país (una de establecimientos educativos, otra de bibliotecas populares y otra de población). El núcleo de nuestra tarea residió entonces en manipular e interpretar aquellos datos necesarios provenientes de estas tablas con tal de constatar la existencia (o la ausencia) de un vínculo entre las cantidades de las dos clases de instituciones.

La estrategia empleada para desentrañar esta búsqueda incluyó un proceso de identificación de los datos importantes (lo cual implicó un descarte de los innecesarios), la elaboración de un diagrama entidad-relación y un modelo relacional de los mismos, un breve análisis de las formas normales y la calidad de los datos de las fuentes facilitadas, y por último un proceso de importación de los datos necesarios a los esquemas creados, con tal de utilizarlos para la generación de una serie de consultas SQL y un apartado de visualización y análisis de datos con gráficos. La intención de lo restante de este informe es documentar todas estas actividades, para finalmente arribar a un veredicto fundamentado acerca de la naturaleza de la relación entre las cantidades de ambos tipos de instituciones que nos disponemos a examinar, o bien de su independencia mutua.

Procesamiento de Datos

Los datos que utilizamos para realizar los gráficos estuvieron contenidos dentro de los utilizados para las consultas SQL. Por lo tanto, partimos de la lista de información mínima necesaria para la elaboración de las consultas, considerándola también la lista de datos necesarios en todo el resto del trabajo.

Datos mínimos necesarios para las consultas SQL pedidas:

* Población de gente, total y según su nivel educativo (jardín, primario y secundario)
* Nivel educativo de los establecimientos educativos
* Bibliotecas fundadas desde 1950
* Cantidad de bibliotecas públicas y establecimientos educativos(solo de modalidad común)
* Dominio de emails asociados a las bibliotecas públicas
* Nombre de la provincia y departamento

Toda esta información debe depender de su departamento.

Antes de disponernos a utilizar los datos crudos, necesitamos estudiar su formato tal y como son presentados, así como corroborar la completitud de los mismos de acuerdo a estos formatos.

La tabla de Establecimientos Educativos presenta varios casos de más de una fila que hacen referencia a un mismo establecimiento, conservando todos los datos excepto por la dirección (esto generalmente ocurre con escuelas que tienen una o dos direcciones aparte de la principal, las cuales son computadas en la tabla en distintos registros). Esto representa un problema para nuestro dataframe, ya que necesitamos quedarnos con un solo registro por establecimiento. De esta manera, desechamos todos los casos repetidos en la tabla original, aprovechando que el dato de la dirección nos resulta irrelevante.

Análisis de formas normales:

Tanto la forma normal de la tabla Establecimientos Educativos como la de Bibliotecas Populares son 1FN, ya que solamente poseen valores atómicos dentro de su dominio.

Sin embargo, no son 2FN porque ambas contienen muchos atributos no primos que dependen parcialmente de alguna de las claves. Por ejemplo, en Bibliotecas Populares están los atributos de provincia (dependiente de *id\_provincia*), departamento (dependiente de *id\_departamento*), y localidad (dependiente de *cod\_localidad*), y en Establecimientos Educativos figuran Departamento (dependiente de Código de Departamento) y Localidad (dependiente de Código de Localidad).

Asimismo, como sabemos que para que un modelo esté en 3FN es requisito que esté también en 2FN, de la misma manera establecemos que si no está en 2FN, tampoco puede estar en 3FN. Así podemos concluir que ambas tablas se encuentran en 1FN.

Los atributos que indican la ubicación de las entidades son los responsables de las dependencias parciales de la tabla de Bibliotecas Populares. Como el único dato necesario con respecto a la ubicación de la entidad es el departamento, podemos considerar solamente como importante el ID de departamento, omitiendo así en el análisis de las formas normales los datos restantes.

Con esta omisión vemos que esta vez las tablas sí estarían en la 3FN, las dependencias parciales se eliminarían y tampoco habrían dependencias transitivas

Para la tabla de Establecimientos Educativos no ocurre únicamente lo mismo que con la anterior, sino que también, por el formato en el que se plantea la información en la tabla, hay una dependencia entre la modalidad y su nivel educativo. Para indicar solamente la modalidad de la entidad se escribe un “1” en la columna correspondiente a ella, y para indicar el nivel educativo se escribe lo mismo en su columna correspondiente. Pero para un único nivel educativo existen varias columnas, cada una dependiendo de la modalidad. Entonces, no solo se indica el nivel educativo sino que a la vez se indica la modalidad. Esto crea una redundancia donde hay una dependencia parcial (nivel\_educativo → modalidad). Al solamente pedirse analizar aquellos establecimientos que sean de modalidad común, si excluimos los que no lo son, el estado de las columnas de nivel educativo de escuelas no comunes va a ser *null*, dándonos la posibilidad de eliminarlas sin deshacer información útil. Con el descarte de filas y columnas que no corresponden a instituciones comunes, y dejando como dato de ubicación el ID del departamento, la tabla estaría en la tercera forma normal.

Análisis de calidad de datos:

Dado que solo hay que considerar ciertos datos para cumplir con los objetivos del trabajo, el análisis de calidad va a ser sobre estos. Por eso vamos a mencionarlos según cada tabla y luego se les hará una evaluación a los problemáticos.

Aclaración: Como no se requieren los nombres de los EE y BP consideramos adecuado no agregarlos. Esto nos evitará el posible problema de nombres repetidos.

Establecimientos Educativos

Por lo explicado anteriormente, sobre las direcciónes, las métricas hechas en el estudio de esta tabla no serán sobre la cantidad registros sino sobre las cantidad de EE que nosotros consideramos reconocidos.

Datos necesarios para los objetivos: ID del departamento perteneciente, modalidad y nivel educativo del establecimiento.

* Problema identificado en la modalidad: hay establecimientos que no tienen ninguna asignada. De esta manera, el atributo de calidad afectado es la completitud, y el problema es de instancia, pues tiene su raíz en datos que no han sido almacenados con la precisión necesaria.

Utilizando la técnica GQM:

* + Objetivo: Que el dato correspondiente a la modalidad de cada Establecimiento Educativo esté completo
  + Pregunta: ¿Cuál es la proporción de establecimientos que no tienen escrito un “1” en ninguna de las tres modalidades (común, especial y adultos)?
  + Métrica: Porcentaje Completitud Modalidad =

= = 0.047750767716073982

Por lo que aproximadamente un 4,77% de los registros no tienen asignada ninguna modalidad. Estos podrían ser escuelas comunes que no vamos a considerar.

La única acción de mejora pertinente sería contactar directamente a los establecimientos en cuestión, consultar cuáles son sus modalidades y actualizar los datos en la tabla.

* Problema identificado: establecimientos comunes sin nivel educativo asignado. Por los mismos motivos que en el caso anterior, el atributo de calidad afectado es la completitud y el problema es de instancia.

Utilizando la técnica GQM:

* + Objetivo: Que cada EE tenga al menos un nivel educativo establecido
  + Pregunta: ¿Cuál es la proporción de EE que no tienen ningún nivel educativo?
  + Métrica: Porcentaje Completitud Nivel Educativo =

= = 0.06657446259

Por lo que aproximadamente un 6,65% de los registros no cuentan con un mail.

Nuevamente, la única acción de mejora posible sería contactar directamente a los establecimientos con tal de consultar su mail y actualizar los datos en la tabla.

Bibliotecas Populares

Datos necesarios para los objetivos: ID del departamento perteneciente, fecha de fundación y dominio del mail de la biblioteca.

* Problema identificado: faltantes en la columna “mail”. Por los mismos motivos que en el caso anterior, el atributo de calidad afectado es la completitud y el problema es de instancia.

Utilizando la técnica GQM:

* + Objetivo: Que el dato correspondiente al mail de cada Biblioteca Popular esté completo
  + Pregunta: ¿Cuál es la proporción de Bibliotecas Populares cuyo dato correspondiente a su mail está vacío?
  + Métrica: Porcentaje Completitud Mail =

= = 0.46267087276

Por lo que aproximadamente un 46,26% de los registros no cuentan con un mail. Entonces existe una cantidad significativa de dominios de mail que no se van a considerar, disminuyendo la fiabilidad de la información sobre dominios de mail que obtendremos más tarde.

Nuevamente, la única acción de mejora posible sería contactar directamente a las bibliotecas populares necesarias con tal de consultar su mail y actualizar los datos en la tabla.

* Problema identificado: faltantes en la columna “fecha\_fundacion”. Por los mismos motivos que en el caso anterior, el atributo de calidad afectado es la completitud y el problema es de instancia.

Utilizando la técnica GQM:

* + Objetivo: Que el dato correspondiente a la fecha de fundación de cada Biblioteca Popular esté completo
  + Pregunta: ¿Cuál es la proporción de Bibliotecas Populares cuyo dato correspondiente a su fecha de fundación está vacío?
  + Métrica: Porcentaje Completitud Fecha de Fundación =

= = 0.02103049421

Por lo que aproximadamente un 2,10% de los registros no cuentan con una fecha de fundación.

Nuevamente, la única acción de mejora posible sería contactar directamente a las bibliotecas populares necesarias con tal de consultar su fecha de fundación y actualizar los datos en la tabla.

De esta manera y habiendo identificado y analizado varias inconsistencias en los datos de las tablas de establecimientos educativos y bibliotecas populares, nos encontramos en posición de afirmar que ambas presentan una calidad de datos más bien pobre, contando con varios ejemplos tanto de unos que no respetan un formato claro establecido, como de otros directamente faltantes.

Sin embargo, nuestra ocupación ahora es copiar aquellos datos que consideramos necesarios y reestructurarlos para crear nuestros propios esquemas de información que nos facilitarán continuar con nuestro análisis.

Planteamiento del DER

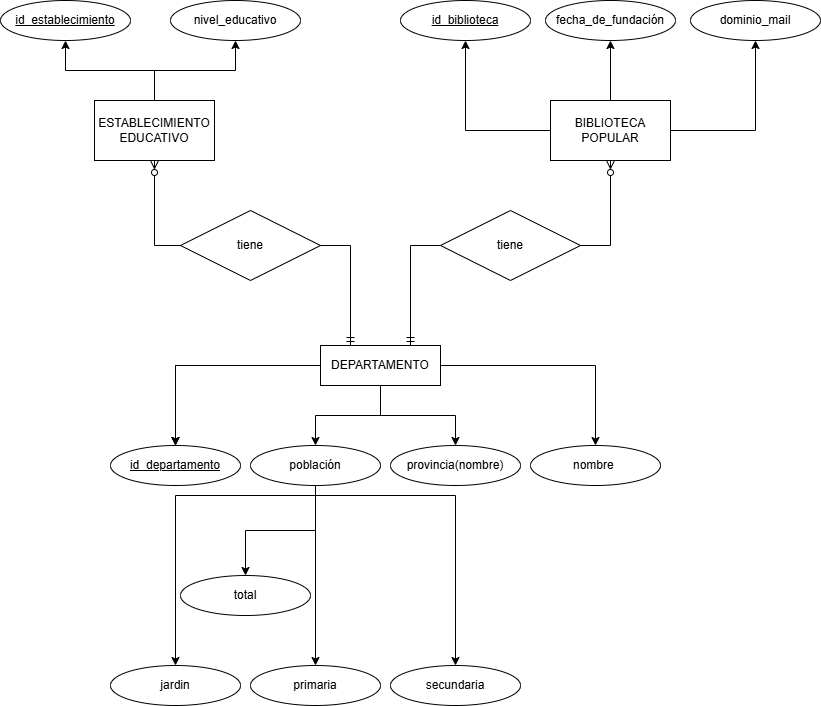
Aclaración: Al no agregarse los nombres de los EE y BP, creamos un ID único entre los tipos de entidad como atributo clave.

Los nombres de los departamentos sí son requeridos pero no todos tienen distinto nombre; por ejemplo existen 5 departamentos llamados “9 de Julio”, por eso usaremos sus IDs como atributos clave y no nos preocuparemos por futuras dependencias funcionales.

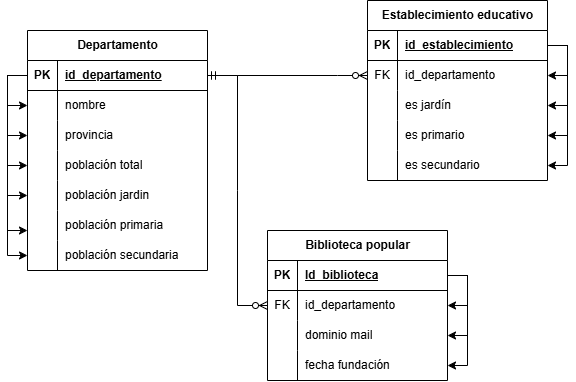
A partir de los datos requeridos, es conveniente separar la información total en tres entidades:

* Departamento: información de la población y nombre de su provincia.
* Establecimientos Educativos (EE): indicando su nivel educativo e ID del departamento al que pertenece.
* Bibliotecas Populares (BP): sus dominios de emails asociados, fecha de fundación e ID del departamento al que pertenece.

El DER concluyente es:



De esta manera tenemos un diagrama que permite modelar de manera conceptual solamente los datos necesarios para resolver los problemas y actividades planteados en el presente trabajo. A partir de este creamos el modelo relacional:



El conjunto minimal de dependencias funcionales está indicado con una flecha dentro del diagrama mostrado.

Como todas las consultas y análisis de los datos son dependiendo de los departamentos o provincia, decidimos usar como clave foránea los *id\_departamento* . Así se pueden relacionar los datos de una manera directa con los departamentos y luego las provincias.

Para normalizar a la tercera forma normal fue necesario dividir el atributo multivalor *nivel\_educativo* en tres atributos diferentes. El atributo compuesto *población* fue dividido en sus 4 atributos atómicos.

Decisiones tomadas

Dentro de la página del INDEC con publicaciones relacionadas al Censo Nacional de 2022 (tercer link de la consigna) figura una gran variedad de cuadros de datos de distintos temas, entre los que se encuentran salud y previsión social, identidad de género, fecundidad, etc. Decidimos que gran parte de estas cuestiones no representa un interés para nosotros en términos del cumplimiento del objetivo del trabajo, ya que la búsqueda de una relación estrictamente numérica entre bibliotecas populares y establecimientos educativos poco tiene que ver con temáticas de matices poblacionales como situación económica o salud, sino que se vincula más bien con las cantidades netas de bibliotecas públicas y establecimientos educativos.

El archivo generado por la consulta en el enunciado del trabajo práctico tiene información de la cantidad de habitantes por edad de cada departamento. Está formado por varias tablas cada una representando un departamento donde se informan los habitantes que hay por edad, qué porcentaje representan cada una y el acumulado de edades menores. De cada tabla únicamente nos interesa la población total del departamento y la de tres grupos etarios, 0 a 6 años, 6 a 12 años y 12 a 18 años, cada uno representando la edad de un nivel educativo distinto. Decidimos almacenar esa información en un archivo .csv generado con *python* que se representa en una tabla donde cada fila es un departamento con su ID y cada columna es la cantidad de habitantes de los 4 atributos mencionados, generada con el programa a través de lo original.

En el archivo crudo se observa que arriba de cada tabla se menciona “AREA # (ID del departamento) (nombre del departamento)” . Resulta que en los códigos de los departamentos de la Ciudad Autónoma de Buenos Aires, el último dígito está multiplicado por 7, haciendo así erróneo el dato del ID del departamento, indispensable para las relaciones de los datos poblacionales con las BP y EE de la ciudad. En consecuencia, fue necesario agregar al código la solución a este error, dividiendo por 7 los últimos 3 dígitos en el traspase de datos del censo al archivo .csv.

En cuanto a la modalidad de los establecimientos educativos, la consigna deja perfectamente claro que solo necesitaremos almacenar los datos de aquellos de modalidad común (de acuerdo a la tabla de la consigna de EE), por lo que los de modalidad especial y para adultos serán directamente descartados de la base de datos, y a su vez, las columnas asociadas únicamente a estas. La información requerida por EE, es el ID del departamento y su nivel educativo. Como cada una va a ser identificada por un ID propio generado por nosotros en *python*,descartamos el nombre de la entidad y el resto de columnas que no representan lo requerido. La metodología para aislar la información necesaria es la misma que antes: con un código transferimos y generamos lo necesario en un archivo .csv. También, las escuelas con varias direcciones asignadas en la tabla original generaban duplicados, por eso pusimos solo un registro por nombre.

En relación con las bibliotecas públicas, el archivo proporcionado es una tabla con cada fila representando una biblioteca. Se obtiene información de la ubicación de las bibliotecas, en las que nuevamente solo nos importa el ID del departamento, mail y fecha de fundación, el resto de la información es irrelevante. Por eso, en el archivo .csv que de nuevo creamos, en cuestión del ID de las BP repetimos lo mismo que hicimos con el de los EE: creamos uno y lo relacionamos con el dominio del correo electrónico, extraído del mismo, el ID del departamento y fecha de fundación. A pesar de haber tantas bibliotecas sin correo asociado, decidimos no hacer nada al respecto, porque no lo consideramos importante para el objetivo final del trabajo. Aún así, reconocemos su importancia en una de las consultas SQL.

Nuevamente encontramos un error asociado al ID de los departamentos de la Ciudad Autónoma de Buenos Aires. El identificador indicado en la fuente no es el del departamento (comuna en este caso) de la entidad, sino que es el código de la ciudad. Al ser solo 42 las afectadas decidimos *hardcodear* su ID correspondiente en el código de *python* que extrae la información necesaria para pasarla a un csv. La información incrustada ha sido encontrada en internet a través de la ubicación indicada en la columna *domicilio* de la fuente.

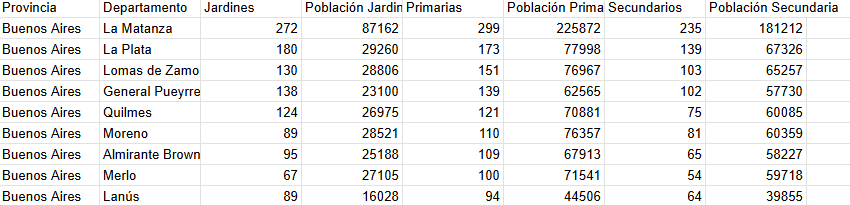
Así generamos tres archivos, cada uno con una tabla que corresponde a las 3 entidades planteadas anteriormente con su información necesaria aislada.

Análisis de datos

Las decisiones tomadas para separar los datos necesarios no fueron muy útiles para la simplificación de las consultas SQL requeridas en la consigna. Sin embargo, estas consultas sirvieron para las visualizaciones y análisis de los datos pedidos.

Cada consulta es uno de los archivos anexados a la entrega: *consulta1.csv*, *consulta2.csv*, *consulta3.csv* y *consulta4.csv*.

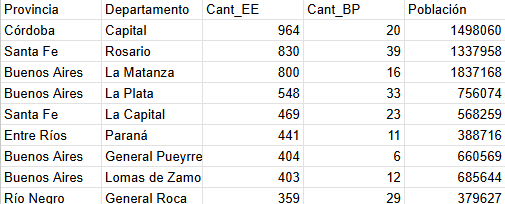
Primeras 10 filas de la consulta 1:



Primeras 10 filas de la consulta 2:



Primeras 10 filas de la consulta 3:

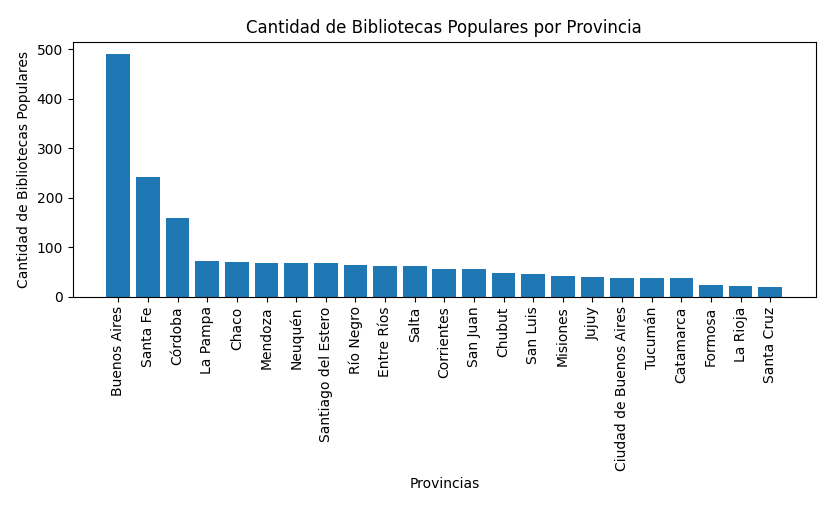


Primeras 10 filas de la consulta 4:



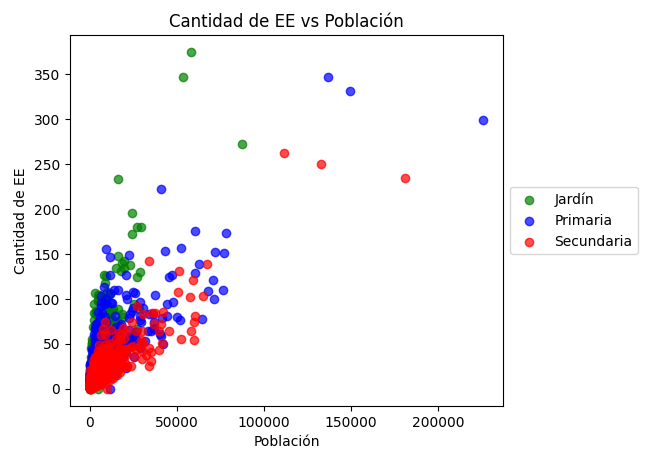
**Gráfico 1**: Compara la cantidad de Bibliotecas Populares de cada provincia. La provincia con más bibliotecas públicas, con una amplia diferencia con la segunda, es Buenos Aires, mientras que la que menos bibliotecas tiene es Santa Cruz. Este gráfico nos podría dar una idea de la inversión en Bibliotecas populares a lo largo de las provincias.

Un dato no menor que queremos destacar es que en este faltan datos cuantitativos y no exhibe datos no menores como podría ser la superficie de la provincia o la población de la misma. Por ejemplo, si consideramos cuántas bibliotecas públicas hay cada 100.000 habitantes en la Provincia de Buenos Aires vs Santa Fe, notamos que hay alrededor de 6 en esta última y 2 en la primera. Aún así, hay otros datos que tampoco estaríamos considerando, como la densidad de la población en cada una.



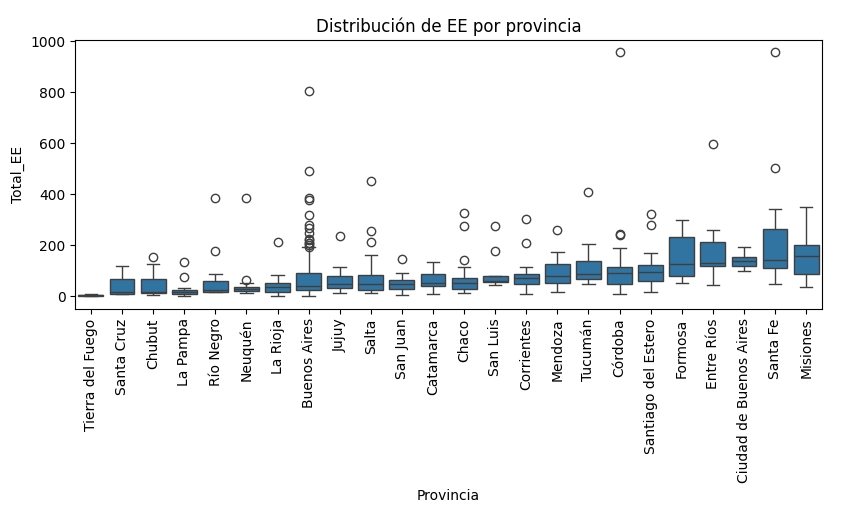
**Gráfico 2**: Compara la cantidad de establecimientos educativos de cada departamento con la población total del mismo. Presenta un crecimiento bastante lineal y similar entre establecimientos educativos y población, ya sea en jardines, primarios o secundarios, pero con un desarrollo todavía más parecido entre primarios y secundarios (los jardines muestran una cantidad ligeramente mayor de instituciones para la misma cantidad población).

A grandes rasgos, lo que nos muestra este gráfico es que la cantidad de establecimientos educativos en cada departamento suele estar ampliamente ligada a la población del mismo.

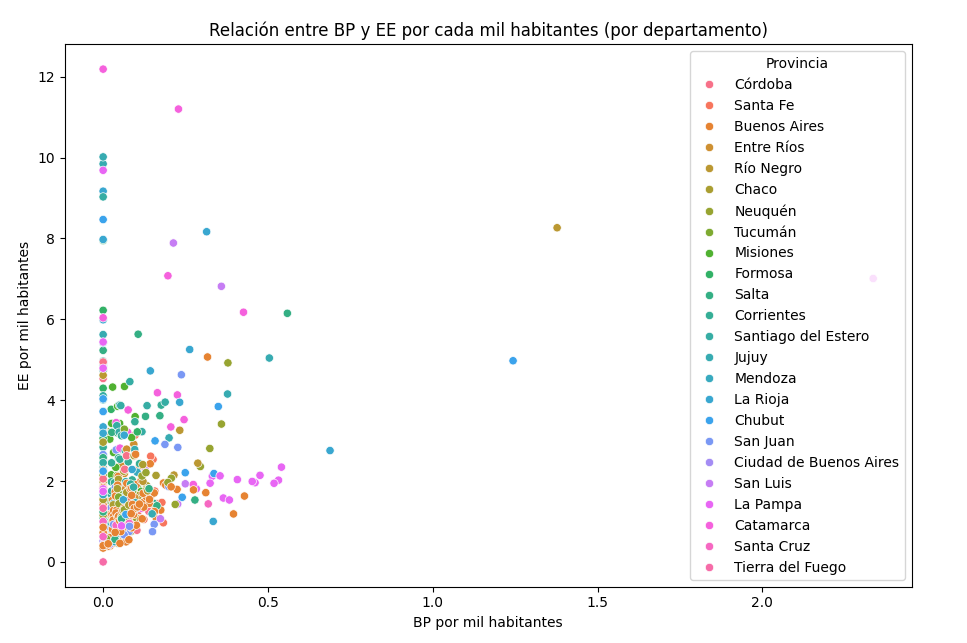


**Gráfico 3**: Muestra la cantidad de EE por departamento, separados por cada provincia. Vemos que la Provincia de Buenos Aires tiene una cantidad excepcional de departamentos con un número de establecimientos muy por encima de la norma, que provincias como Córdoba, Santa Fe, Misiones o Formosa tienen una distribución de establecimientos alta y estable a lo largo de sus departamentos, y que otras como Neuquén, La Pampa o Tierra del Fuego tienen un número bajo, pero extremadamente estable a lo largo de sus jurisdicciones.

En definitiva, este gráfico nuevamente se puede explicar en gran medida (aunque no completamente) en base a las diversas densidades poblacionales del país: la Provincia de Buenos Aires es la más poblada y particularmente tiene una gran cantidad de departamentos más densamente poblados que los demás; esta es seguida por Córdoba y Santa Fe en población, mientras que otras como La Pampa o Tierra del Fuego tienen poca población.



**Gráfico 4**: Cantidad de BP vs cantidad de EE, ambas cada mil habitantes y por departamento, el cual es identificado por color según su provincia. Vemos que las BP son mucho menos numerosas que los EE: casi todos los departamentos tienen menos de media BP cada mil habitantes, e incluso hay un número notable de departamentos (la franja vertical de la izquierda de todo), de distintas provincias y distintas cantidades de EE, que no tienen ninguna BP.



Conclusiones

Luego de observar los gráficos y las consultas, concluimos que no existe una relación numérica entre la cantidad de establecimientos educativos y bibliotecas populares del país. Esto se debe a que los primeros crecen principalmente en respuesta a la población de un departamento o provincia dada, mientras que las segundas deben su desarrollo no solamente a la población, sino también a un conjunto de factores mucho más diversos y desconocidos para nosotros. Por esto es que su presencia es muy independiente de la población, sobre todo más allá de las tres provincias más pobladas (gráfico 1). Por ejemplo, en La Pampa hay más bibliotecas populares que en la Ciudad Autónoma de Buenos Aires, con tan solo un décimo de los habitantes.

La relación que hay entre establecimientos educativos y población, por otro lado, es lineal (como podemos observar en el gráfico 2). Por ende, los establecimientos educativos y las bibliotecas populares deben su presencia numérica en una región dada a distintos factores, o, lo que es lo mismo, no están relacionados.