TRABAJO PRÁCTICO NRO 1

Escuelas y bibliotecas

Integrantes: Iván Pérez, Camila Lambot y Nicole Lamblot
Nombre del grupo: Grupo TP01 – 14
Materia: Laboratorio de Datos
Fecha: 25/05/25

RESUMEN

A lo largo de este informe realizamos consultas SQL y gráficos a nuestra base de datos creada a partir del DER y modelo relacional establecidos, buscando analizar la relación existente entre la cantidad de escuelas y bibliotecas populares, considerando el factor poblacional. Pudimos denotar un fuerte vínculo entre las grandes capitales o zonas densamente pobladas y la cantidad de establecimientos, y se evidenció una profunda disparidad en el acceso a bibliotecas populares y escuelas en zonas rurales.

INTRODUCCIÓN

En el contexto del análisis educativo y territorial, la distribución de los establecimientos escolares y culturales es un factor clave para comprender el acceso a derechos fundamentales como la educación y la información. Argentina presenta marcadas diferencias regionales en términos de densidad poblacional y disponibilidad de recursos, lo que motiva la necesidad de estudios que articulen ambos aspectos.

En este informe trabajamos con datos públicos provenientes del Censo Nacional 2020, registros de escuelas y bibliotecas populares (CONABIP), con el objetivo de explorar posibles relaciones entre la cantidad de establecimientos educativos y bibliotecas populares en distintos departamentos del país. Para ello, realizamos la limpieza de las bases de datos, la creación de nuevos datasets a partir de un modelo relacional propio, y el análisis a través de consultas SQL y representaciones gráficas.

Como resultado, pudimos ver cómo se distribuyen ambos tipos de instituciones por provincia y departamento, teniendo en cuenta la población, y sacamos conclusiones que detallamos más adelante.

El informe continúa con la descripción y explicación del modelo de datos junto a las decisiones tomadas en el proceso y, finalmente, la presentación de los resultados obtenidos.

PROCESAMIENTO DE DATOS

OBJETOS Y ATRIBUTOS

Teniendo en cuenta nuestro objetivo, notamos que es necesario crear una relación espacial entre los establecimientos educativos y las bibliotecas populares. Por lo que identificamos que principalmente debíamos definir un objeto Establecimientos Educativos, un objeto Bibliotecas Populares y uno de Departamentos (unidad mínima de representación geográfica en común entre los datasets).

Para realizar este paso debimos leer a conciencia todos los datasets a analizar, identificar las columnas necesarias para nuestro análisis, las columnas dispensables, y demás.

FORMA NORMAL DE LAS TABLAS ORIGINALES

Comencemos analizando la forma normal de los datasets originales:

Comenzando por el dataset de *Establecimientos Educativos*, podemos notar de antemano que no se encuentra en la primera forma normal. Tomemos de ejemplo a las columnas de ‘Mail’ y/o ‘teléfono’. Vemos que existen casos en los que a un mismo establecimiento le corresponden dos o más mails/teléfonos distintos que se almacenan en el mismo registro. Sin embargo, para que la tabla esté en primera forma normal, debería existir un objeto aparte, por ejemplo, ‘contacto’, que relacione al establecimiento con los distintos mails en tuplas diferentes. Lo mismo sucede con 'teléfono’. Al no estar en primera forma normal, tampoco está en segunda y, por lo tanto, tampoco en tercera.

Por otro lado, si miramos en dataset de *Bibliotecas Populares*, notemos que de hecho sí se encuentra en primera forma normal, pues todas las celdas de la tabla contienen un único valor atómico. Asimismo,

tenemos una clave primaria compuesta de un único atributo, que es 'nro_conabip', por lo tanto, no existen dependencias parciales y está en segunda forma normal. No obstante, existen dependencias transitivas entre atributos no primos y la clave primaria. Por ejemplo, 'nro_conabip' define 'id_departamento', y a su vez, este último define a 'departamento'. Pero, como 'id_departamento' no es un atributo primo, el dataset no se encuentra en tercera forma normal.

ANÁLISIS DE CALIDAD DE DATOS (GQM)

Revisando las distintos bases de datos (*Bibliotecas Populares* y *Establecimientos Educativos*), pudimos observar distintos problemas de calidad en cada uno. Para describirlos, utilizaremos la técnica GQM (Goal, Question, Metric).

Comencemos con la tabla de *Establecimientos Educativos*:

En primer lugar, echemos un vistazo a las últimas tres columnas de nuestro dataset: 'talleres-artística', 'servicios complementarios' y 'validez_títulos'. A simple vista, notamos que la mayoría de sus registros son nulos. Esto indica que, independientemente del objetivo del análisis, estas columnas aportan poca información. Además, en relación con el enfoque específico de nuestro trabajo, no brindan datos relevantes. Una posible causa del problema podría deberse a que el modelo de datos contempla atributos que en la práctica no se están utilizando.

Para evaluar la relevancia como atributo de calidad del dataset, podemos construir el siguiente GQM:

G: Todas las columnas deben aportar información relevante al registro.

Q: ¿Cuál es la proporción de establecimientos que no tiene participación en ninguna de estas tres categorías?

M: Nro de tuplas que tienen las celdas correspondientes a las 3 columnas vacías/Nro total de tuplas

$$\frac{62135}{64711} = 0.9601922393410703$$

```
#1ra GQM

ee.rename(columns={
    'Unnamed: 37': 'talleres_artistica',
    'Unnamed: 38': 'servicios_complementarios',
    'Unnamed: 39': 'validez_titulos'
}, inplace=True)

celdas_vacias = ee[["talleres_artistica", "servicios_complementarios", "validez_titulos"]].isna().all(axis=1).sum()
print(celdas_vacias)

total_celdas = ee.shape[0]
print(total_celdas)

GQM_1 = celdas_vacias/total_celdas
print(GQM_1)
```

Otro problema que distinguimos es que en algunas columnas en específico se dificulta la accesibilidad a ciertos datos. Como ya mencionamos anteriormente, existen establecimientos con dos o más mails en un mismo registro, así como sucede con los teléfonos. Esto representa un esfuerzo innecesario para quien desee hacer uso de estos datos, dado que debería, por lo menos, separarlos de alguna manera. Esto sugiere una falla en la estructura del modelo de datos, que no permite reflejar correctamente la multiplicidad del atributo.

Para hacer el siguiente GQM, reducimos el problema a la columna 'mail', pero debería hacerse también para 'teléfono'.

G: Todos los registros de la columna 'mail' deben contener un único valor atómico.

Q: ¿Cuál es la proporción de establecimientos que tienen más de un mail?

M: Número de tuplas con más de un valor atómico en 'mail'/Número total de tuplas.

$$\frac{468}{64711} = 0.007232$$

```
#2da GQM

consultaSQL = """
                SELECT
                COUNT(*) AS total_tuplas,
                COUNT(CASE WHEN Mail LIKE '%/%' THEN 1 END) AS tuplas_con_mas_de_un_mail
                FROM ee
                """

tuplas_con_mas_de_un_mail = dd.sql(consultaSQL).df()

cant_mas_de_un_mail = tuplas_con_mas_de_un_mail['tuplas_con_mas_de_un_mail']
print(cant_mas_de_un_mail)

total_tuplas = tuplas_con_mas_de_un_mail['total_tuplas']
print(total_tuplas)

GQM_2 = cant_mas_de_un_mail/total_tuplas
print(GQM_2)
```

Por otro lado, en el dataset de *Bibliotecas_Populares*, podemos observar varios casos de columnas con un mismo valor en todas sus entradas, datos irrelevantes que no aportan profundidad a la información de los registros. Ejemplos de esto son la columna ‘categoría’, ‘subcategoría’, ‘tipo_latitud_longitud’ y ‘fuente’. Nuevamente atribuimos una posible causa del problema al modelo de datos, pues probablemente estos atributos se incluyeron sin hacer un análisis previo de su relevancia.

Para poder determinar la relevancia de estos registros usamos el siguiente modelo GQR:

G: Todas las columnas deben aportar información relevante al registro.
Q: ¿Cuánto varían los datos en cada registro de ese campo?
M: Suma de todos los valores distintos en la columna.
Con un solo count distinct ya pudimos observar que no hay valores distintos en ningún registro
count distinct = 1

Con respecto a la completitud, hay múltiples columnas que contienen registros nulos. Por ejemplo, la totalidad de los registros de las columnas ‘observación’ o ‘subcategoría’ tienen valores nulos. De todas formas, esto resulta de mayor utilidad para evaluar la relevancia de los datos, tal como hicimos antes, pero vale la pena mencionarlo. Asimismo, notamos que en la columna ‘mail’ también hay registros nulos. Dado que las consultas que debemos hacer más adelante involucran este atributo, nos concierne evaluar su completitud. En este caso, encontramos dos causas posibles para este problema. Podría tratarse de un error de software si el sistema permite que el campo ‘mail’ se cargue vacío cuando debería completarse obligatoriamente. Aunque, si por ejemplo algunas bibliotecas no tienen mail, entonces eso sería un problema de procesos.

G: El dato correspondiente al Mail de cada biblioteca debe estar completo.
Q: ¿Cuál es la proporción de bibliotecas que tienen el dato correspondiente a mail vacío?
M: Nro de registros nulos en la columna mail/Nro total de registros.

$$\frac{880}{1902} = 0.46267087276551$$

```
#4ta GQM

celdas_vacias = bp["mail"].isna().sum()
print(celdas_vacias)

total_celdas = bp.shape[0]
print(total_celdas)

GQM_4 = celdas_vacias/total_celdas
print(GQM_4)
```

DOCUMENTACIÓN DEL DER Y MODELO RELACIONAL

Antes de limpiar nuestras bases de datos, analizamos sus atributos para definir los nuevos datasets que armaremos para realizar las consultas de SQL. Debimos examinar cuáles eran las columnas de las que

podíamos prescindir y cuáles eran indispensables para nuestro objetivo. En base a esto, construimos un diagrama entidad-relación (DER) esquematizando las relaciones entre cada objeto y sus atributos.

‘Departamentos’:

Creamos una entidad *Departamentos* en el centro de nuestro DER, ya que es la entidad por la cual se relacionan todas nuestras bases de datos. Tiene como clave primaria a su id y como atributos el nombre, la cantidad de residentes que van al jardín, la cantidad de residentes que van a la primaria, y la cantidad de residentes que van a la secundaria. De esta forma incluimos la información que nos interesaba de la tabla *Padrón población*, y la relacionamos con los departamentos.

‘Provincias’

Asimismo, para indicar en qué provincia se encuentra cada departamento, creamos la entidad *Provincias* que tiene como clave primaria su id, y su nombre como atributo. Relacionamos las provincias con los departamentos a través de la relación uno a muchos (1:N) ‘*Se encuentra en*’. La participación por ambas partes es total. Esto quiere decir que un departamento se encuentra en una única provincia, pero en una misma provincia hay uno o más departamentos. Es así como logramos preservar la 3FN, ya que si hubiéramos puesto al nombre de las provincias como un atributo de *Departamentos*, habría una dependencia transitiva entre el id del departamento y el nombre de provincia (‘id_departamento’ define ‘departamento’, ‘departamento’ define ‘provincia’, por lo tanto, ‘id_departamento’ define ‘nombre_provincia’).

‘Escuelas de modalidad común’

Con respecto al dataset de establecimientos educativos, creamos una entidad llamada *Escuelas de modalidad común* para representarlo. Definimos como clave primaria al cueanexo, pues es el único dato que no se repite en ninguna tupla. Como atributos dejamos el nombre de escuela, el ámbito y el sector que le corresponde. Luego, relacionamos *Departamentos* con *Escuelas de modalidad común* a través de una relación que designamos como ‘*Queda en*’ de uno a muchos (1:N) con participación total a parcial. Esto se debe a que en un departamento puede haber ninguna o una o más escuelas, pero cada escuela reside en un único departamento.

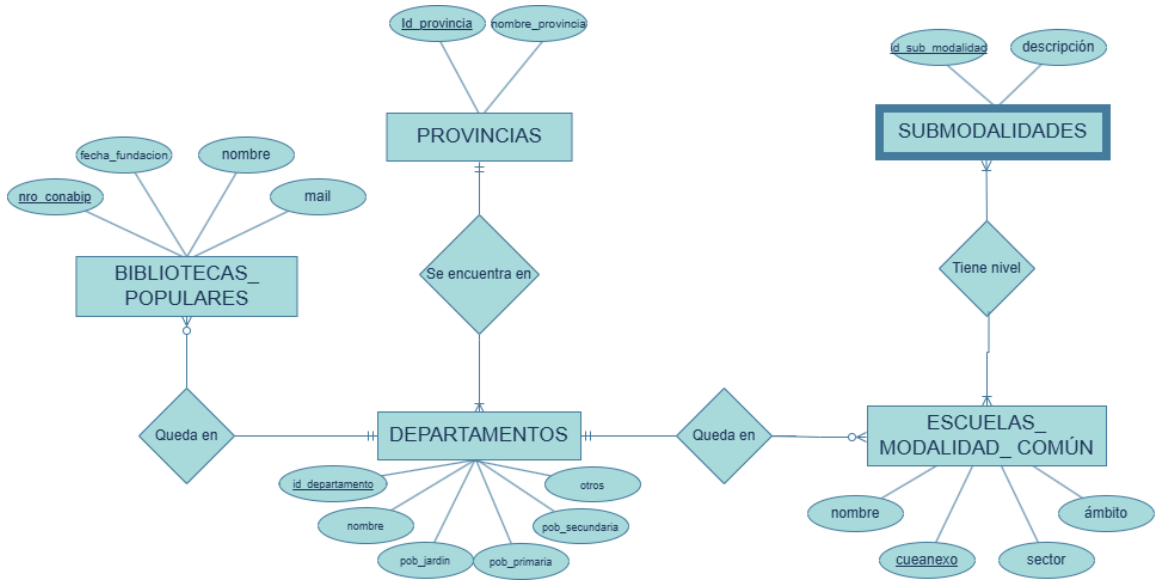
Dentro de la modalidad común existen diferentes submodalidades o niveles y cada establecimiento puede contener uno o más de ellos. Esto presenta un desafío al momento de mantener la tercera forma normal, pero para ello creamos una entidad débil *Submodalidades* que contiene todos los diferentes niveles. De esta manera, existe una relación de muchos a muchos (N:N) entre *Escuelas de modalidad común* y *Submodalidades*, con participación total de ambos lados. Esto significa que una escuela puede tener varias submodalidades y que una misma submodalidad puede corresponder a varias escuelas distintas. Denominamos esta relación ‘*Tiene nivel’*.

‘Bibliotecas Populares’

En cuanto a las Bibliotecas Populares, creamos una entidad con el mismo nombre para referirnos a ellas. Designamos como clave primaria al número de conabip, ya que, así como pasaba con el cueanexo en los establecimientos educativos, es el único valor que nunca se repite. Cuenta con los atributos nombre de cada biblioteca, fecha de fundación y el mail correspondiente. A diferencia de la entidad de *Escuelas de modalidad común*, cada biblioteca tiene un único mail, por lo que no es necesario hacer una entidad aparte.

Definimos la relación entre *Departamentos* y *Bibliotecas Populares* como ‘*Queda en’*, y es de uno a muchos (1:N), con participación total a parcial, al igual que en *Escuelas de modalidad común.* La representamos de esta forma ya que en cada departamento puede haber ninguna, una o más bibliotecas y cada biblioteca se encuentra en un único departamento.

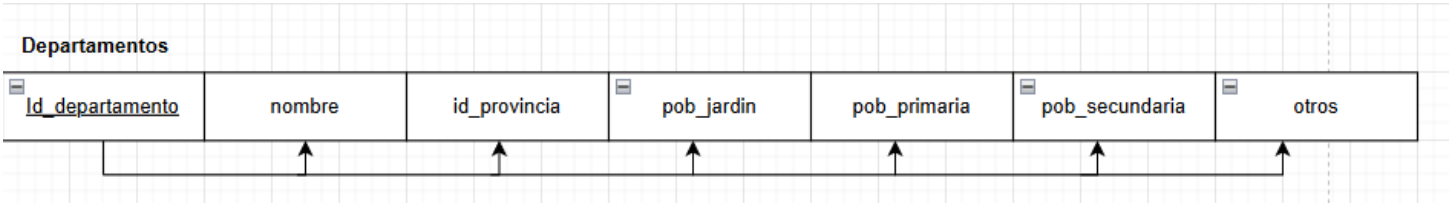
A continuación, se muestra cómo nos quedó el DER:

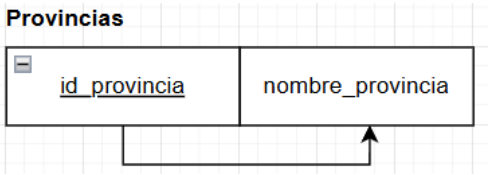


De esta manera, nos queda un diagrama entidad-relación que respeta y preserva la tercera forma normal, pues no existen dependencias transitivas. A partir de esto, podemos representarlo en un modelo relacional.

Modelo relacional:

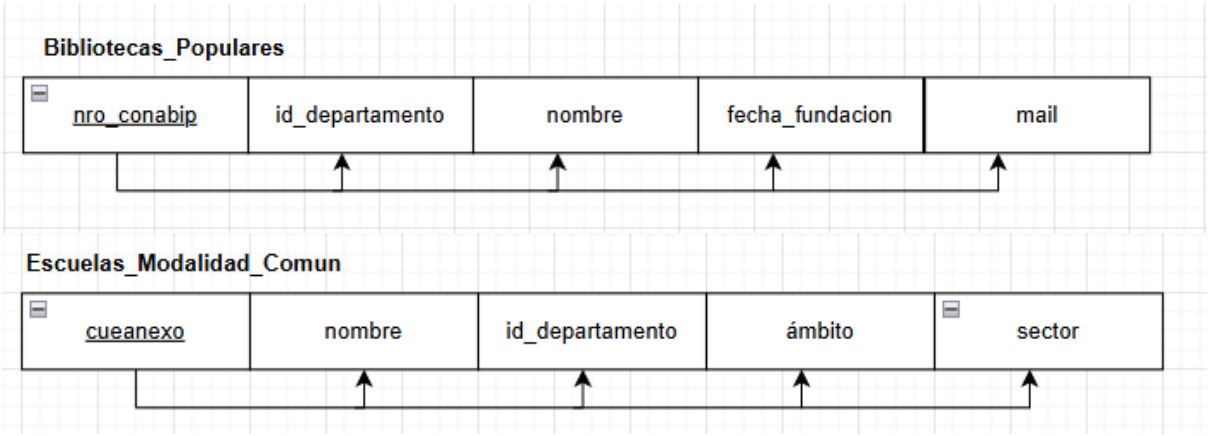
La entidad *Departamento*, que conecta a todo el resto, se mantiene exactamente igual a como la establecimos en el DER, pero además agregamos el atributo 'nombre_provincia'. En cambio, *Provincias* permanece igual sin ninguna modificación. Debido a que la relación entre *Departamento* y *Provincias* es de uno a muchos, como dijimos antes, decidimos organizar así los datos para preservar la 3FN. A continuación, se muestran las dependencias funcionales:



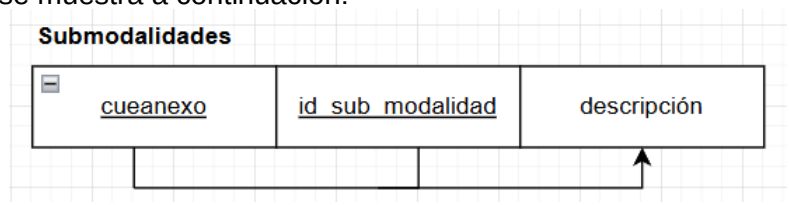


Como podemos ver, no existen dependencias parciales ni transitivas, por lo tanto, mantiene la 3FN.

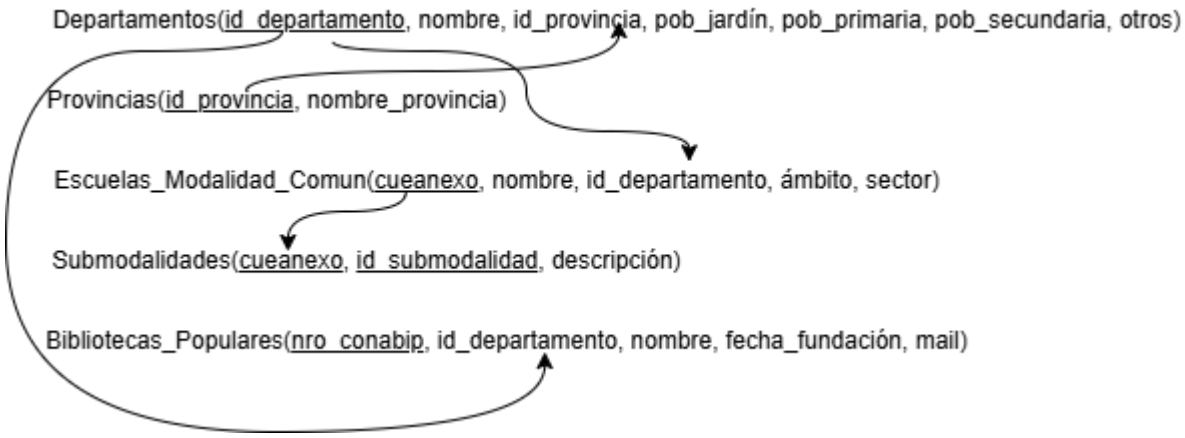
A su vez, las entidades principales, tales como *Escuelas de modalidad común* y *Bibliotecas Populares* también quedan iguales al DER, con sus respectivas claves primarias (subrayadas) y atributos, con la excepción de que les agregamos el atributo de 'id_departamento' como clave foránea para relacionarlas con *Departamento*. Notemos que es la situación inversa a lo que sucede entre *Departamento* y *Provincias*. En este caso, son muchas las escuelas y/o bibliotecas que pueden (o no) estar en un único departamento, mientras que a cada escuela/biblioteca le pertenece un único departamento. Es decir, esta también es una relación de uno a muchos, pero con los roles invertidos. Por esa razón, lo resolvemos haciendo lo mismo que antes, pero en vez de agregar 'cueanexo' o 'nro_conabip' a *Departamentos,* agregamos 'id_departamento' a cada relación. Así quedan entonces las dependencias funcionales, manteniendo la 3FN:



Ahora echemos un vistazo a las entidades débiles de nuestro DER. '*Tiene_nivel'* relaciona las escuelas con sus submodalidades o niveles. En este caso, la entidad débil es *Submodalidades*. Para representar esta relación en el modelo relacional, mantenemos la relación *Submodalidades* con sus atributos, pero agregamos 'cueanexo' como clave foránea al principio de todos sus atributos. Definimos entonces la clave primaria como la composición de los atributos 'cueanexo' y 'sub_modalidad'. Esta forma de separar los niveles de cada escuela de la relación *Escuelas con modalidad común* permite conservar la 3FN. La única dependencia funcional que existe se muestra a continuación:



Todas las claves primarias de cada tabla se encuentran subrayadas. Conectamos los distintos datasets a través de las mismas, señalando con flechas cuando la clave primaria de una tabla es clave foránea de otra u otras.



DECISIONES TOMADAS

Muchas de las decisiones que tomamos en esta etapa condicionaron directamente la estructura del modelo relacional presentado anteriormente. A continuación, detallamos los principales criterios adoptados para limpiar, unificar y transformar los datos, con el fin de que el modelo final respondiera de forma coherente a los objetivos planteados.

REDUCCIÓN Y LIMPIEZA DE COLUMNAS

Un desafío que enfrentamos fue cómo organizar la tabla de establecimientos educativos, ya que la información se presentaba según las categorías generales 'Establecimiento-Localización', 'Modalidad', 'Común', 'Adultos', 'Especial' y 'Hospitalaria'. A su vez, dentro de cada una de estas categorías existían subcategorías, lo cual iba a dificultarnos realizar las consultas que nos pide el enunciado. Es por eso que decidimos modificarla en función de nuestros objetivos.

En primer lugar, decidimos deshacernos de las categorías generales mencionadas anteriormente y dejar las más específicas, para facilitar el acceso a una columna en caso de que lo necesitemos. Optamos por prescindir de las columnas 'Jurisdicción', 'Departamento', 'Localidad', 'Código de localidad', 'Domicilio', 'C.P', 'Teléfono', 'Talleres-Artística', 'Servicios complementarios' y 'Validez-título'. Tomamos esta decisión dado que no necesitamos información tan precisa para abordar la geografía de los establecimientos: con conocer el departamento en el que residen nos basta para nuestro análisis (y con el código de departamento podemos relacionar a los establecimientos con los departamentos en la tabla *Departamento*, en la cual también se ve a qué provincia pertenece cada uno). En cuanto a 'Teléfono', 'Talleres-Artística', 'Servicios complementarios' y 'Validez-título'; notamos que no nos era de utilidad para las consultas conservar esa información, además de que las últimas tres columnas tenían casi todas sus entradas vacías.

En cuanto a la tabla *Bibliotecas Populares*, se encontraba mejor estructurada que la de *Establecimientos Educativos*, aunque tenía muchas columnas irrelevantes. Con respecto a la información geográfica nos pasó algo similar a la de las escuelas, de modo que aplicamos la misma estrategia: dejamos el código de departamento y borramos el resto. También vimos que había muchas columnas con el mismo valor en todas sus entradas, tales como fuente, fecha de actualización, web, categoría, subcategoría, entre otros; por lo que optamos por borrarlas ya que consideramos que era información redundante. A su vez, dejamos a un lado el teléfono ya que no es importante para nuestro objetivo.

Selección de modalidad educativa

Con respecto a las modalidades de los establecimientos educativos, consideramos que la información estaba organizada de manera poco eficiente. Las columnas correspondientes a los distintos niveles educativos según la modalidad contenían muchas celdas vacías, ya que no todos los establecimientos abarcan todos los niveles.

Dado que en el trabajo se nos pide realizar las consultas en torno a las escuelas de modalidad común únicamente, decidimos dejar sólo a esos establecimientos y descartar el resto, con el objetivo de reducir el volumen de datos y eliminar información irrelevante para nuestro análisis. A partir de esto, volvimos a filtrar las tuplas, pero esta vez de acuerdo a sus submodalidades y elegimos quedarnos con aquellas escuelas que ofrecen nivel jardín, primario y/o secundario, puesto que son los únicos niveles que nos interesan para realizar las consultas. De este modo, creamos una tabla aparte llamada 'Submodalidades', que relaciona a cada escuela con su nivel educativo (jardín, primario o secundario).

Agrupación de población

Por otro lado, en base a nuestro Modelo Relacional, construimos distintas bases de datos. Una de ellas es el correspondiente a *Establecimientos Educativos* y otro a *Bibliotecas Populares*. A diferencia de estas dos tablas, optamos por no armar un dataset específico de *Padrón Población*. En su lugar, agrupamos las edades de acuerdo al nivel educativo que típicamente pertenecen: 'pob_jardín' incluye a las edades de entre 2 y 5 años, 'pob_primaria' abarca de 6 a 12 años y 'pob_secundaria' de los 13 a los 18 años. El resto de edades no son relevantes para los fines de nuestro trabajo, por lo que descartamos todas las tuplas que no pertenezcan a dichos rangos de edad. Una vez definidos estos grupos etarios por departamento, incorporamos esta información al dataset de *Departamentos* como atributos adicionales.

Estandarización geográfica

A la hora de establecer la relación entre los distintos datasets, como ya explicamos antes, decidimos regirnos por los departamentos, puesto que es un dato geográfico que se encuentra en las tres bases de datos a analizar. Sin embargo, nos encontramos con que cada una de las tablas tenía un sistema diferente para identificar a los departamentos con sus respectivos códigos.

Por un lado, en la tabla de *Bibliotecas Populares*, CABA no está dividida en comunas tal como sucede en *Establecimientos Educativos* o *Padrón Población*. En cambio, figura como un único departamento asociado

al código 2000. A su vez, aunque en *Establecimientos Educativos* y en *Padrón Población* sí están las comunas diferenciadas, los códigos de las comunas no coinciden. Por ejemplo, el código de departamento de la comuna 1 en *Establecimientos Educativos* es 02101, mientras que en *Padrón Población* es 02007. Esto representaba un problema, ya que necesitábamos que las bases de datos tengan un atributo donde los datos sean comunes en las tablas para hacer los joins en nuestras consultas.

Otro punto importante es que no necesariamente hay escuelas o bibliotecas en todos los departamentos. Podría haber un departamento en el que haya varias escuelas y ninguna biblioteca, o al revés. También podría ocurrir que en un departamento no haya ninguna de las dos y que en otro hayan diversas escuelas y bibliotecas. Luego, si decidiéramos regirnos por los códigos de departamento de *Establecimientos Educativos* o los de *Bibliotecas Populares*, podríamos estar perdiéndonos de valiosa información para nuestro análisis. De modo que, si queremos un análisis más profundo y completo, deberíamos regirnos por los códigos de departamento de *Padrón Población*.

Teniendo todo esto en cuenta, decidimos adoptar los códigos del *Padrón de Población* como estándar, pero unificando todas las comunas de CABA en un único departamento general, con su código correspondiente siendo el 2000. Para el resto de departamentos mantuvimos los códigos igual.

Esta decisión requirió que debamos realizar algunos ajustes en cuanto a la población: sumamos la cantidad de personas por grupo etario de todas las comunas de CABA, para representar correctamente a la población total del departamento unificado.

Errores detectados y decisiones pragmáticas

Un inconveniente que encontramos mientras analizábamos el dataset, fue que había algunas bibliotecas que tenían un código de departamento erróneo. Por ejemplo, el código de departamento de Quilmes es 6658, pero tres veces aparece un código distinto. Sin embargo, dado que en total hay doce bibliotecas en Quilmes, consideramos que estos errores puntuales no tendrán un impacto significativo en los resultados de nuestro análisis.

Así como ocurre con algunos establecimientos de Quilmes ocurre con otros departamentos, pero son pocos los registros incorrectos. Corregir manualmente los códigos de todas las escuelas para garantizar que los datos sean correctos implicaría un esfuerzo considerable, que no se justifica dado que hay muy pocos casos afectados. Por este motivo, decidimos mantener los IDs tal como están y dejar constancia de esta decisión, aclarando que no impacta en el análisis que realizaremos.

Estandarización de nombres de columnas

Con el objetivo de poder realizar las consultas con la mayor facilidad posible, optamos por cambiar el nombre de las columnas que tenían en común nuestros datasets para que todos tengan los mismos nombres. Es así como para representar al código de departamento usamos ‘id_departamento’, por ejemplo.

Creación de bases de datos

BP: Seleccionamos las columnas necesarias del archivo *Bibliotecas Populares:* ‘Nro_Conabip’,’id_departamento’, ’nombre’, ‘fecha de fundacion’, ‘mail’. Al estar bastante limpio no fue necesaria una mayor intervención.

EE: Para esta tabla fue necesario estandarizar la columna de ‘id_departamento’, englobando todas las comunas de CABA con el mismo id (2000). Luego tomamos las columnas del archivo original: ‘nombre’, ‘ámbito’, ‘sector’ y ‘cueanexo’.

Departamentos: Para la creación de esta tabla fue necesario un primer análisis exploratorio del dataset de *Padrón Población*. Se recortaron las primeras filas del archivo, ya que no eran partes de tablas.

Un primer objetivo fue unificar las tablas, ya que se podían observar mas de 300 tablas distintas, pero todas con un mismo patrón de separación, por lo tanto, creamos un algoritmo que reconocía la fila en donde se mencionaba el nombre del departamento, lo guardaba en una variable, y a cada registro de ese mismo departamento, lo agregaba a una nueva lista con un nuevo atributo con el nombre del departamento.

Fueron necesarios varios ajustes debido a que, al final del archivo, se encuentran unas filas de resumen que no debían entrar en el nuevo Dataframe y fueron debidamente recortadas.

Otro ajuste importante fue detectar que las cifras de poblaciones tenían tipo de datos string y las cifras mayores a 999 contaban con un espacio entre la unidad de miles y la centena (por ej “1 000” en lugar de “1000”). Luego de corregir esto se pudo avanzar con el procesamiento.

Una vez unificado en una sola tabla se agrupó a la población por rangos etáreos correspondientes a los distintos tipos de establecimientos educativos, consiguiendo de esta forma la tabla departamentos con la columna ‘id_departamento’, ‘nombre_departamento’, ‘id_provincia’, ‘pob_jardín’, ‘pob_primaria’ y ‘pob_secundaria’.

Provincias: Utilizamos el DataFrame de *Bibliotecas Populares*, seleccionamos las columnas ID provincia y provincia, y sacamos duplicados. De esta forma obtuvimos una lista de las provincias argentinas que estarán relacionadas con la tabla de departamentos.

Submodalidades: Construida de forma manual contiene 3 registros de las posibles submodalidades que pueden tener los establecimientos educativos objetos de estudio. Se conforma por una columna id y otra de submodalidad.

Submodalidades_EE: representa la relación muchos a muchos entre *Submodalidades* y *Establecimientos Educativos*. Para poder armar esta tabla se seleccionó las columnas ‘cueanexo’ y las 3 submodalidades. Luego, con el método melt, se unificó la información de las 3 columnas de submodalidades en una sola, representando por registro una institución con una modalidad.

ANÁLISIS DE DATOS

PRIMERA CONSULTA

Índice	Provincia	Departamento	Jardines	Poblacion_Jardin	Primarias	Poblacion_Primaria	Secundarias	Poblacion_Secundaria
0	Buenos Aires	La Matanza	333	113641	335	225872	336	181212
1	Buenos Aires	La Plata	218	37646	201	77998	211	67326
2	Buenos Aires	Lomas de Zamora	168	37017	179	76967	191	65257
3	Buenos Aires	General Pueyrredón	177	29666	168	62565	173	57730
4	Buenos Aires	Quilmes	158	34636	145	70881	151	60085
5	Buenos Aires	Moreno	120	36738	138	76357	134	60359
6	Buenos Aires	Almirante Brown	139	32199	137	67913	146	58227
7	Buenos Aires	Merlo	108	34982	120	71541	125	59718
8	Buenos Aires	Lanús	116	20686	117	44506	113	39855
9	Buenos Aires	Pilar	108	24609	112	49944	105	41528

Primeras diez filas de la consulta (Puede verse la consulta completa en la sección Anexos al final del informe).

A partir de los datos obtenidos en esta consulta, podemos observar una marcada concentración de escuelas en los departamentos con mayor población, tales como La Matanza, La Plata o Lomas de Zamora. En cambio, aquellas zonas con menor densidad poblacional cuentan con una cantidad de establecimientos significativamente más baja. Esto indica una relación proporcional entre la cantidad de escuelas y la población de entre 2 y 18 años en cada departamento.

Por esta razón es lógico que las capitales provinciales suelan concentrar la mayor cantidad de establecimientos educativos y de población escolar de todos los niveles por sobre el resto de los departamentos dentro de la provincia. Una excepción a esta tendencia, por ejemplo, es CABA, pero al haberla normalizado según su cantidad de comunas, es esperable que no muestre valores tan altos.

También se destaca que la población en edad de cursar el nivel primario es, en general, más numerosa que la de nivel inicial o secundario, siendo este último, en general, el que cuenta con menos establecimientos que ofrecen dicho nivel.

SEGUNDA CONSULTA

Índice	Provincia	Departamento	Cantidad de BP fundadas desde 1950
0	Buenos Aires	La Matanza	15
1	Buenos Aires	La Plata	15
2	Buenos Aires	Moreno	13
3	Buenos Aires	Bahía Blanca	12
4	Buenos Aires	Tigre	12
5	Buenos Aires	Lomas de Zamora	10
6	Buenos Aires	General San Martín	10
7	Buenos Aires	Almirante Brown	10
8	Buenos Aires	Tandil	10
9	Buenos Aires	Avellaneda	9

Primeras diez filas de la consulta (Puede verse la consulta completa en la sección Anexos al final del informe).

En base al análisis de los datos que obtuvimos en la segunda consulta, vemos que existe una distribución muy desigual de bibliotecas fundadas a partir de 1950. Encontramos varios departamento con gran cantidad de bibliotecas, así como lo son Confluencia en Neuquén (39), San Fernando en Chaco (26), Capital en Salta (26), y General Roca en Río Negro (25). Cabe destacar a La Plata y La Matanza de la Provincia de Buenos Aires, los cuales cuentan ambos con 15 bibliotecas cada uno. Recordemos que estos dos departamentos en especial eran los que poseían mayor concentración de establecimientos educativos y población escolar sobre el resto del país, lo cual podría estar asociado a una mayor necesidad —y por ende presencia— de espacios como las Bibliotecas Populares.

Por otro lado, un número importante de departamentos en distintas provincias no registran la fundación de ninguna Biblioteca Popular desde 1950. Esto se observa especialmente en Buenos Aires, donde más de 40 departamentos no cuentan con ninguna.

En líneas generales, se observa una tendencia a que los departamentos con mayor densidad poblacional concentren también una mayor cantidad de bibliotecas.

TERCERA CONSULTA

Índice	Provincia	Departamento	Cant_EE	Cant_BP	Población
0	Córdoba	Capital	1064	20	1498060
1	Buenos Aires	La Matanza	969	16	1837168
2	Santa Fe	Rosario	754	39	1337958
3	Buenos Aires	La Plata	619	33	756074
4	Buenos Aires	Lomas de Zamora	527	12	685644
5	Buenos Aires	General Pueyrredón	506	6	660569
6	Buenos Aires	Quilmes	443	9	631774
7	Santa Fe	La Capital	429	23	568259
8	Entre Ríos	Paraná	420	11	388716
9	Buenos Aires	Almirante Brown	417	11	583209

Primeras diez filas de la consulta (Puede verse la consulta completa en la sección Anexos al final del informe).

Los datos obtenidos en esta consulta nos muestran que la Provincia de Buenos Aires lidera ampliamente en términos de población y cantidad de escuelas y bibliotecas. Eran resultados esperables teniendo en cuenta la gran densidad poblacional de la provincia. Las provincias de Córdoba y Santa Fe también se destacan, mostrando números altos en los tres atributos especialmente en sus respectivas capitales. Córdoba, capital, por ejemplo, tiene 1024 establecimientos educativos, siendo este el valor más alto registrado en un solo departamento en todo el país.

Vemos que pareciera haber una correlación entre la cantidad de escuelas y la población de cada departamento: a mayor población, mayor cantidad de escuelas. Con respecto a las bibliotecas, si bien también hay cierta correlación con la población, aparenta ser más moderada. Por ejemplo, Confluencia, Neuquén posee la mayor cantidad de bibliotecas populares de todo el país (40), superando incluso a departamentos mucho más poblados.

CUARTA CONSULTA

Índice	provincia	departamento	Dominio más frecuente en BP
0	Buenos Aires	Adolfo Alsina	hotmail.com
1	Buenos Aires	Adolfo Gonzales Chaves	yahoo.com.ar
2	Buenos Aires	Adolfo Gonzales Chaves	hotmail.com
3	Buenos Aires	Alberti	hotmail.com
4	Buenos Aires	Alberti	live.com.ar
5	Buenos Aires	Almirante Brown	yahoo.com.ar
6	Buenos Aires	Arrecifes	yahoo.com.ar
7	Buenos Aires	Avellaneda	yahoo.com.ar
8	Buenos Aires	Azul	bibliotecaronco.com.ar
9	Buenos Aires	Azul	yahoo.com.ar

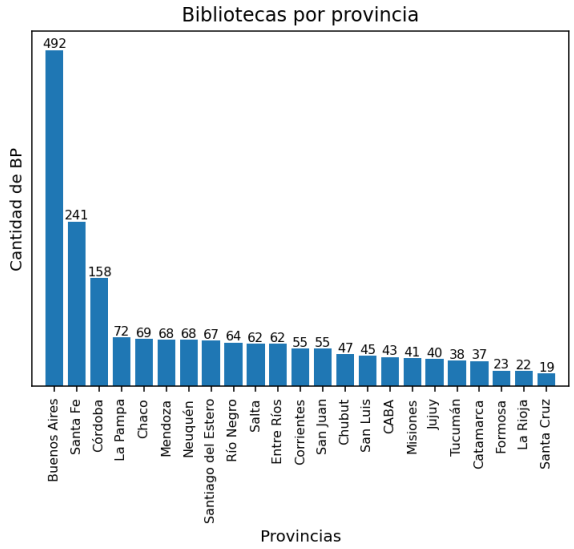
Primeras diez filas de la consulta (Puede verse la consulta completa en la sección Anexos al final del informe).

Para esta consulta, en caso de empate en la frecuencia de dominios dentro de un mismo departamento, se optó por conservar todas las opciones con igual número de menciones.

En base al análisis realizado, se observa que el dominio de correo electrónico mayormente utilizado en las bibliotecas populares es yahoo.com.ar. Luego, le sigue hotmail.com y gmail.com se lleva el tercer puesto. Nótese que dominios propios o personalizados es poco probable que sean compartidas por varias bibliotecas, predominando el uso de servicios de correos generales.

CONCLUSIONES DE LOS GRÁFICOS

PRIMER GRÁFICO



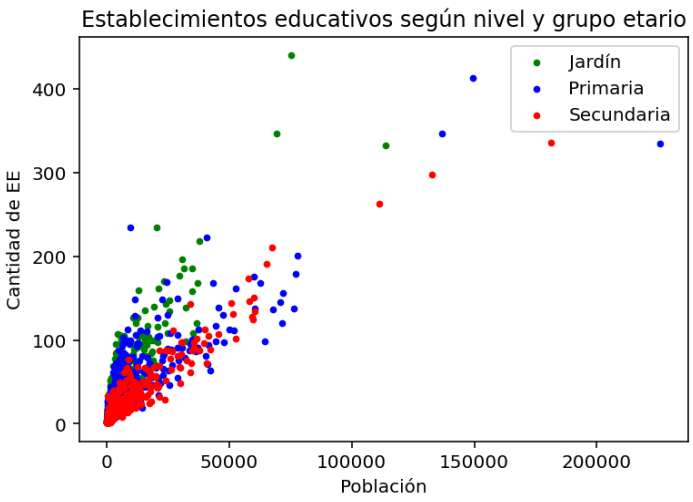
En el gráfico se muestra que la Provincia de Buenos Aires tiene, por lejos, la mayor cantidad de bibliotecas populares, seguida por Santa Fe y Córdoba. A partir de ahí, vemos que la cantidad de bibliotecas disminuye considerablemente, siendo La Pampa quien se lleva el cuarto puesto aun teniendo menos de la mitad que Córdoba.

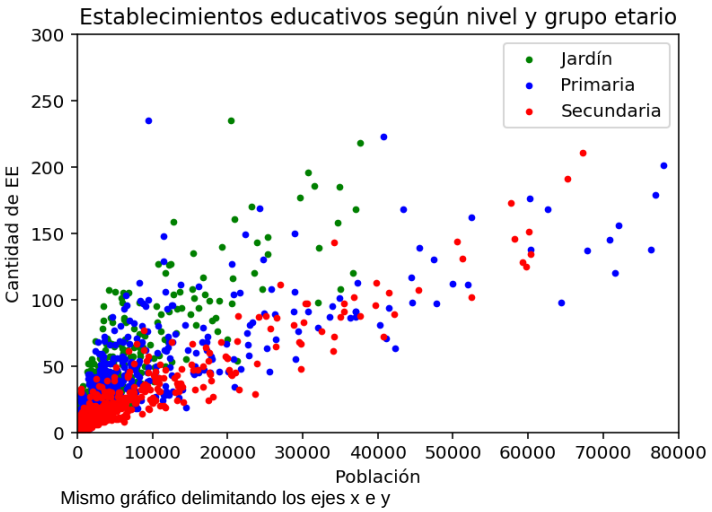
Es curioso que, aunque Buenos Aires tiene muchos departamentos sin bibliotecas (según la segunda consulta SQL), en este gráfico vemos que se destaca ampliamente por sobre el resto del país. Esto muestra una distribución desigual en la provincia, con más cantidad de bibliotecas, en general, en las zonas más pobladas.

Por otro lado, luego de Córdoba parece haber una distribución más equitativa entre el resto de las provincias, con cantidades entre 62 y 72 bibliotecas. Otras provincias tienen entre 37 y 55, y son Formosa, La Rioja y Santa Cruz las provincias que cuentan con la menor cantidad de BP.

En términos generales, podemos decir que las provincias con mayor cantidad de habitantes y desarrollo urbano tienden a concentrar más bibliotecas.

SEGUNDO GRÁFICO





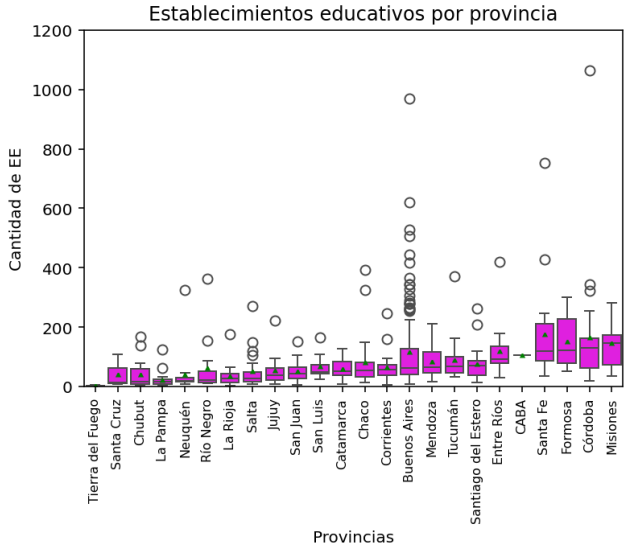
Mismo gráfico delimitando los ejes x e y

A partir de este gráfico podemos sacar varias conclusiones. En primer lugar, vemos que a medida que aumenta la población del grupo etario correspondiente, hay mayor cantidad de establecimientos educativos.

Por otro lado, los puntos verdes y azules (que representan los niveles inicial y primario) son los que tienen valores más altos en el eje Y, lo cual indica que son los niveles que tienen la mayor cantidad de escuelas por departamento. En cambio, aunque el nivel secundario sigue la misma tendencia creciente, cuenta con menos establecimientos que los otros dos.

Asimismo, podemos ver que hay algunos departamentos atípicos. Algunos cuentan con mucha población y una considerable cantidad de escuelas, así como otros con bastantes habitantes y pocas escuelas.

TERCER GRÁFICO

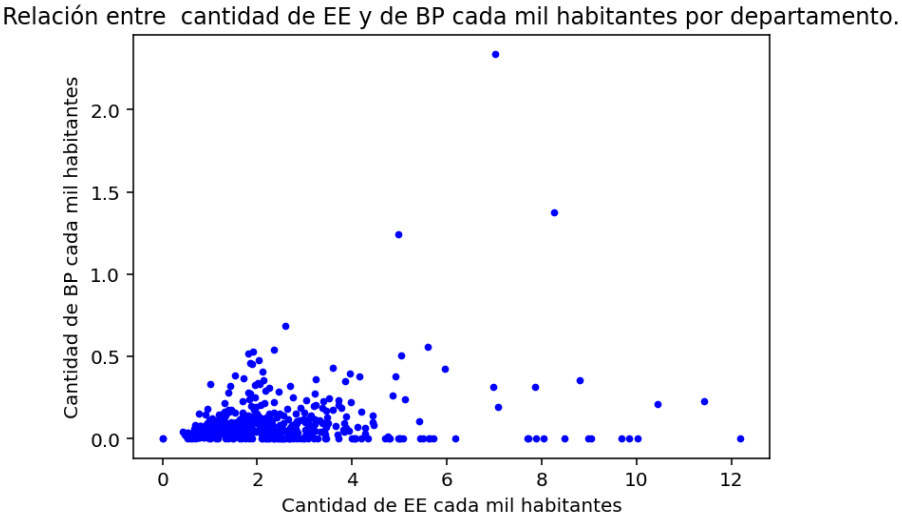


En este gráfico podemos ver la distribución de la cantidad de escuelas por departamento dentro de cada provincia. Al estar ordenados ascendentemente según la mediana de la cantidad de establecimientos en cada provincia, se facilita la identificación de provincias con mayor desigualdad interna en la distribución de escuelas entre sus departamentos.

Gracias a su estructura de boxplot, vemos que Buenos Aires, Córdoba y Santa Fe presentan numerosos outliers, indicando una alta dispersión y departamentos con valores muy por encima de la media. No es casualidad que suceda en provincias con departamentos o ciudades de considerable densidad de población.

En contraste, en provincias como Tierra del Fuego, San Luis o La Rioja, los boxes son más bajos y compactos, además de que cuentan con escasos o nulos outliers. Esto refleja una distribución más equitativa de EE dentro de cada provincia.

CUARTO GRÁFICO



Mirando el gráfico notamos que la mayoría de los departamentos se agrupan en valores bajos tanto de BP como de EE cada 1000 habitantes. Esto quiere decir que, en general, hay poca cantidad de ambas instituciones respecto a la población total. Sin embargo, los puntos suelen extenderse más por el eje X que por el Y, lo que indica que hay más escuelas que bibliotecas.

Además, al observar el scatter vemos que no hay una relación clara entre la cantidad de bibliotecas y establecimientos educativos cada 1000 habitantes. Hay departamentos con alta densidad de EE pero baja densidad de BP, mostrando que la presencia de escuelas no necesariamente va acompañada de una cantidad de bibliotecas proporcional.

CONCLUSIONES FINALES

Comenzamos este análisis con el objetivo de estudiar la relación entre la cantidad de establecimientos educativos y bibliotecas populares, considerando cómo influye la densidad poblacional. El segundo gráfico logró plasmar la relación intrínseca entre las escuelas y la población, en donde al aumentar el número de personas con la necesidad de escolarización, es necesario la construcción de más escuelas.

La densidad demográfica tiene un efecto similar sobre las bibliotecas populares, pero en una medida significativamente menor. Vemos que se observa un gran margen en la cantidad total de bibliotecas por provincia.

Provincias con una gran extensión, pero con poca densidad demográfica, cuentan con pocas bibliotecas populares, dejando afuera de su servicio a muchas personas. Un caso posible es el de Santa Cruz que cuenta con más de 400.000 personas y una extensión de 240.000km² pero con solamente 19 bibliotecas populares y 280 escuelas (teniendo en cuenta todas las modalidades).

Concluimos que, si bien la densidad demográfica influye en la cantidad de establecimientos y bibliotecas populares, no existe una relación directa entre estas instituciones; de manera que su distribución por departamentos debe vincularse con distintos factores.

ANEXOS

En este espacio agregaremos todos los datos complementarios al informe, tanto un acercamiento a las tablas creadas para del modelo, como las consultas hechas en el trabajo.

TABLAS

A continuación, se encuentran las tablas de nuestra base de datos de forma completa:

- Escuelas Modalidad Común
- Bibliotecas Populares
- Provincias
- Departamentos
- Submodalidades
- Submodalidades_EE

CONSULTAS SQL

A continuación, se encuentran los archivos con las consultas enteras:

- 1era consulta
- 2da consulta
- 3ra consulta
- 4ta consulta

GLOSARIO

Con el fin de realizar consultas más limpias renombramos las tablas de nuestra base de datos. En esta sección definiremos los nombres de las tablas en las consultas, para un mejor entendimiento de las mismas.

ee : Tabla de Establecimientos Educativos Modalidad Común.

bp : Tabla de Bibliotecas Populares.

deptos : Tabla de Departamentos.

prov : Tabla de provincias.

submodalidades : Tabla de la relación entre submodalidades y Establecimientos Educativos.

submo : tabla de las submodalidades.