LABORATORIO
DE
DATOS

Trabajo Práctico #1
Escuelas y bibliotecas

25/05/2025

Hernán Blumenthal
Walid Sanchez Richani
Pedro Soldatich

Resumen:

El objetivo central de este trabajo es investigar si existe o no una relación entre la cantidad de establecimientos educativos y las bibliotecas populares por departamento en la Argentina. Para esto se utilizaron diversas fuentes de acceso público relevantes para responder al objetivo. Se analizó esta información y se procedió a generar un DER, y su modelo relacional correspondiente. Las bases de datos a las que se accedieron no eran de muy buena calidad, por lo que se tuvieron que limpiar y adaptar para generar los esquemas relacionales. Luego, se utilizó SQL para armar reportes en base a estos esquemas. Los reportes agrupan información útil por departamento la cual fue representada en distintos tipos de gráficos para su visualización y subsiguiente análisis cualitativo.

Se concluyó una débil relación positiva entre cantidad de EE y BP por departamento, y se recomendó hacer una análisis estadístico para confirmar esta tendencia.

Introducción:

El objetivo principal del presente trabajo práctico es determinar si existe una relación entre la cantidad de establecimientos educativos y la cantidad de bibliotecas populares en los departamentos del territorio nacional Argentino.

Esta labor se realizó a partir de múltiples fuentes de datos abiertas obtenidas de las páginas oficiales del INDEC, datos.gov.ar y argentina.gov.ar, en particular: datos de población por departamento (“Estructura por edad de la población”) del censo 2022 (DP), el Padrón de bibliotecas populares (BP) y el Padrón Oficial de Establecimientos Educativos 2022 (EE), datos de las provincias codificadas, y datos de los departamentos codificados. Para alcanzar dicho objetivos se debieron realizar las siguientes tareas:

1. Análisis preliminar de fuentes de datos: donde se estudiaron las formas normales en las que estaban los datos sin procesar, y se revisó la calidad de los datos originales. Este último punto incluye la identificación, descripción y diagnóstico de problemas de calidad con la técnica GQM sobre las tablas de EE y BP.
2. Procesamientos de datos: se esbozó el mapa conceptual, Diagrama-Entidad Relación (DER), donde se incluyeron únicamente los datos relevantes para alcanzar los objetivos propuestos. También se crearon los modelos relacionales los cuales implementan en esquemas las entidades vinculadas en el DER. Estos esquemas se diseñaron en 3FN, con sus claves primarias correspondientes (PK), dependencias funcionales (DF), y claves foráneas (FK) vinculantes.
3. Consultas SQL: se generaron los reportes pertinentes para el subsiguiente análisis de los datos. Se agruparon los datos por provincia y departamento sobre población total, población según grupo etario escolarizado, cantidad de jardines, primarios y secundarios, y cantidad de EE y BP, entre otros.
4. Visualización y análisis de datos: Para un análisis más ágil de los datos obtenidos se produjeron gráficos de tipo histograma, box plot y scatter plot según los requerimientos de la información a ser representada. A partir de ellos se revelaron algunas tendencias y reflexiones cualitativas que se informan en el análisis y las conclusiones.

El informe está divido en 6 secciones: (1) Resumen, (2) Introducción, (3) Procesamiento de Datos, (4) Decisiones tomadas (5) Análisis de Datos (6) Conclusiones. Continuamos en la siguiente página con la sección (3)

Procesamiento de Datos:

Formas normales (FN):

Bibliotecas Populares: Cada columna contiene valores atómicos, los cuales dependen de la clave primaria (nro_conabip), luego cumple con 1FN. Dado que la PK está compuesta de un solo atributo no son posibles las dependencias parciales, lo cual nos permite confirmar que también se encuentra en 2FN. Existen dependencias transitivas entre la PK y algunos atributos, como por ejemplo, PK → id_provincia → provincia, por lo tanto no está en 3FN.

Establecimientos Educativos: La base de datos de EE original, no se encontraba en 1FN, ya que esta prohíbe relaciones dentro de relaciones, en particular el caso de varios números de teléfono separados por comas en el atributo "teléfono". Las relaciones anidadas o expresadas en tuplas están prohibidas en 1FN.

Análisis de calidad:

GQM para EE

Con respecto a la calidad de EE, se encontró que el autor de la base de datos decidió incluir atributos que generaban una cantidad enorme de valores nulos, lo cual dificulta la visualización y análisis de los datos. Esto parece haberse dado para evitar relaciones anidadas, como por ejemplo tuplas como valores de instancias. Pero en ese proceso se incluyeron columnas con muy poca información útil, lo cual contribuye a una insuficiente *relevancia* (atributo de calidad afectado).

Objetivo: Evaluar la cantidad de información real representada (relevancia) en las columnas a través de la cuantificación de valores nulos presentes en cada una de ellas.

Pregunta: ¿Cuántos valores nulos sobre el total existen en cada columna?

Métrica: Porcentajes de valores nulos en cada columna → *(valores nulos/total de filas) x 100*.

Resultados: El porcentaje de nulos en el atributo "Primarios" bajo la modalidad común dio 65%, el de "Secundario" bajo modalidad común dio 79%, el de "Nivel Inicial - Jardín Maternal") bajo modalidad común dio 93% y "Nivel Inicial - Jardín de Infantes" dio 68%. Como se puede ver, las métricas dieron resultados muy pobres, con un altísimo porcentaje de nulos. Esto pone en duda la utilidad de la inclusión de estos atributos de esta manera. Para resolver esta situación, durante nuestro diseño de los modelos, se procedió a reubicar esta relación "anidada" entre establecimientos educativos y modalidades (n:m) en un nuevo esquema.

GQM para BP

La base de datos de BP original, tiene los siguientes problemas: atributos vacíos, atributos que solo tienen un único valor repetido para todas las instancias (redundancia), y errores de veracidad (id_departamento inconsistentes), y valores faltantes (mail). Los atributos vacíos o de valores repetidos para todas las instancias habla de la pobre *relevancia* de la información contenida. Por otro lado, las filas con información errónea o contradictoria presentan un desafío para la *confiabilidad* y *consistencia* del data set. Luego, los valores faltantes generan un problema de *completitud*.

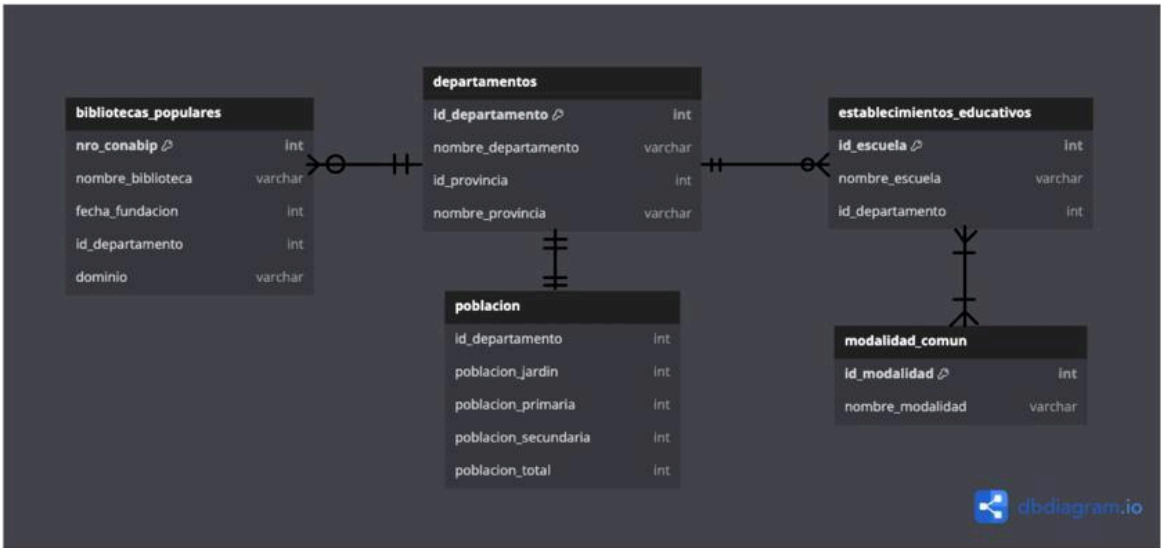
Objetivo: Evaluar la completitud de la información a través de la medición de la cantidad de valores nulos en la columna "mail".

Pregunta: ¿Cuántos valores nulos existen sobre el total en la columna "mail"?

Métrica: Porcentajes de valores nulos en la columna "mail" → *(valores nulos/total de filas) x 100*.

Resultados: El porcentaje de nulos en la columna "mail" dio 46%, lo cual representa casi la mitad de las BPs censadas con este campo vacío. Los resultados no fueron alentadores, lo cual afectó directamente la consulta del ejercicio de SQL (iv), ya que se produjo una subrepresentación de los dominios reales en cada departamento. Lamentablemente, no es viable la corrección de este problema, dado que requiere un trabajo de campo que excede las posibilidades del trabajo práctico.

DER y modelo relacional:



El modelo relacional que implementa el DER es el siguiente (subrayada la PK):

- **Departamentos** (id_departamenos, nombre_departamentos, id_provincia, nombre_provincia)
- **Establecimientos_educativos** (id_escuela, nombre_escuela, id departamento)
- **Modalidad_común** (id_modalidad, nombre_modalidad)
- **Modalidad_escuela** (id_escuela, id_modalidad)
- **Bibliotecas_populares** (nro_conabip, nombre_biblioteca, fecha_fundacion, id_departamento, dominio)
- **Poblacion** (id_departamento, poblacion_jardin, poblacion_primaria, poblacion_secundaria, poblacion_total)

Glosario de atributos (este es el modelo relacional espejado, con los nombres como aparecen en las tablas implementadas, para facilitar la lectura de los mismos):

- **df_DEPARTAMENTOS_final** (DPTO, NOMDPTO, PROV, Provincia)
- **df_ee_FINAL** (Cueanexo, nombre, id_departamento)
- **df_modalidad_comun_ref** (id_modalidad, nombre_modalidad)
- **df_modalidad_escuela** (Cueanexo, id_modalidad)
- **df_bp_final** (nro_conabip, nombre, fecha_fundacion, id_departamento, dominio)
- **df_poblacion** (id_departamento, poblacion_jardin, poblacion_primaria, poblacion_secundaria, poblacion_total)

Claves primarias (PK), Claves foráneas (FK) y sus Dependencias funcionales (DF)
Departamentos (PK = id_departamentos): (DF) id_departamentos → nombre_departamentos, id_departamentos → id_provincia → nombre_provincia

Establecimientos_educativos (PK = id_escuela, FKs = id_departamento): (DF) id_escuela → nombre_escuela, id_departamento
Modalidad_comun (PK = id_modalidad): (DF) id_modalidad → nombre_modalidad
Modalidad_escuela (representa una relación): PK = Cueanexo, id_modalidad (está compuesta por ambas)
Bibliotecas_populares (PK = nro_conabip, FK = id_departamento): (DF) nro_conabip → nombre_biblioteca, fecha_fundacion, id_departamento, dominio
Poblacion (PK = FK = id_departamento: (DF) id_departamento → poblacion_jardin, poblacion_primaria, poblacion_secundaria, poblacion_total

A la hora de implementar el DER a los esquemas, se utilizó el criterio de practicidad según los objetivos a cumplir por el trabajo.

Proceso de limpieza y combinación de las fuentes de datos:

El proceso de limpieza y combinación realizado se enfocó en mejorar la calidad de los datos sobre las fuentes originales. Se trabajó con datos de población, establecimientos educativos, bibliotecas populares y departamentos, aplicando diversas técnicas para garantizar que los registros sean confiables y útiles.

Para mantener la consistencia entre los diferentes conjuntos de datos, se revisó la estructura y los formatos utilizados, asegurando que fueran uniformes. Se aplicaron validaciones cruzadas para verificar la correcta relación entre los esquemas de departamentos, establecimientos educativos, población y bibliotecas populares.

En primer lugar, se verificó la veracidad de los datos, asegurando que los identificadores de área y población correspondiera correctamente con las otras tablas. También se ajustaron los códigos de localidad, eliminando caracteres innecesarios para mantener la compatibilidad de claves en los cruces de información. Además, se implementó una limpieza, eliminando valores nulos y registros duplicados para garantizar la integridad de los datos. Estas acciones ayudaron a mejorar la calidad de las bases de datos, evitando posibles errores derivados de información incompleta.

Esquema Departamento:

Para este esquema se relaciono la información de la tabla donde están codificadas las provincias, con otra tabla que contiene los departamentos y sus códigos identificadores (INDEC - CENSO NACIONAL DE POBLACIÓN, HOGARES Y VIVIENDAS 2022).

En una etapa clave, se realizan ajuste especial a los registros con PROV == 2, consolidándose para evitar duplicaciones y representar la información de una manera más limpia, además algunos valores se modificaron (*"2000"* para CABA) para reflejar de manera más adecuada los datos agrupados, cómo reemplazar nombres de las comunas por un término más general (ej. “Comuna promedio”). Finalmente, se reorganizaron las filas para que la presentación sea más lógica, facilitando la lectura y análisis (más información sobre esto en la sección de decisiones tomadas).

Esquema Establecimientos Educativos:

Luego del análisis de calidad hecho sobre la fuente original de EE, se procedió a acotar la información en una entidad fuerte incluyendo los atributos de id_escuela, nombre_escuela, e id_departamento por un lado y armar otra entidad débil (dependiente de

la primera) que llamamos modalidad común. Esta última contiene las codificaciones de las modalidades (más información sobre esto en la sección de decisiones tomadas). Luego se generó un tercer esquema que asocia los id_escuela con sus respectivos id_modalidad (que llamamos modalidad_escuela). Esto fue necesario dada la relación muchos a muchos (n:m) entre las dos entidades.

Luego, la inclusión de la jurisdicción en el esquema final de establecimientos educativos permite evaluar la distribución de la oferta escolar en distintos departamentos y su relación con la población y otras infraestructuras.

Esquema Población:

Trabajamos con la columna *"Unnamed:1"* para extraer los códigos de área. Aplicamos una separación basada en el carácter *"#",* lo que nos permite asociar cada registro con su correspondiente área, facilitando su identificación en nuestro modelo relacional.

Una vez estructurado el dataset, segmentamos la población en distintos rangos etarios:

- Población jardin (0-5 años)
- Población primaria (6-11 años)
- Población secundaria (12-18 años)

Para cada grupo, calculamos el total de casos por área y consolidamos la información en un único dataframe, lo que simplificó el análisis. Como previamente habíamos eliminado la columna "total" , filtramos las filas relevantes y sumamos los primeros 15 registros de CABA. Agrupamos esta información en un nuevo dataset permitiéndonos una integración más clara y representativa de los datos.

Posteriormente, se integró el total de población a esta tabla, sumando los valores de comunas urbanas y eliminando ceros innecesarios en los identificadores de departamento.

Esquema Biblioteca Populares:

A través de la normalización de identificadores de departamento y la extracción de dominios de correo electrónico, se ha logrado mejorar la calidad y compatibilidad del dataset. La identificación de estas características permite una mejor vinculación entre los datos de bibliotecas y su población respectiva, ofreciendo un panorama más estructurado para el objetivo a cumplir.

Decisiones tomadas:

En una primera instancia decidimos considerar la Ciudad Autónoma de Buenos Aires (CABA) como un único departamento, lo cual nos llevó a unificar los datos de sus comunas. Esto resultó en una sobrerrepresentación de CABA al compararlo con los demás departamentos del país. Por ello, se optó por otorgar un identificador único a CABA *("2000")* y calcular el promedio de los datos en aquellos casos donde la información es cuantificable, como la cantidad de establecimientos educativos o la población de jardines, entre otros (aclaración: en el ej. 2 de visualización optamos excepcionalmente por no promediar el valor absoluto de CABA, ya que daba menor a 1. Tener esto en cuenta a la hora de interpretar ese resultado).

Por otro lado, en los departamentos de la provincia de Buenos Aires consideramos sacar el *"0"* al identificador del mismo, en cada una de las relaciones donde aparece, esto nos

permitió reducir los errores en la claves foráneas o claves primarias para relacionarlas con otras entidades, facilitando la compatibilidad a la hora de cruzar los datos.

Muchas de las fuentes originales contenían columnas innecesarias para nuestro objetivo, por lo que las eliminamos, así como aquellas filas con valores nulos que no aportan información relevante. Sin embargo, hay algunas excepciones, como en el esquema de Bibliotecas Populares, donde decidimos conservar las filas con valores nulos en la columna *“mail”,* ya que dicha información era esencial para nuestro objetivo.

Siguiendo el análisis de calidad confeccionado sobre la tabla original de EE, procedimos a separar los atributos de modalidad para que siga la 3FN. Se generó un esquema “modalidad_común” el cual codifica numéricamente las modalidades de Jardín, Primario y Secundario en 1, 2 y 3. Luego se produjo otro esquema que representa la relación “escuela_modalidad”, donde se asocia la PK de EE con sus códigos de modalidad correspondientes. Esta estrategia nos permitió acomodar los datos en 3FN, lo cual ayudó a eliminar redundancias innecesarias.

Luego, por criterio de practicidad para la elaboración de los reportes derivados de las consultas de SQL, se dejaron los atributos nombre_prov y id_prov en el esquema de Departamentos. Esta decisión se tomó en base a los objetivos planteados, ya que todos los reportes requieren el atributo de "nombre de provincia” asociado al de "nombre de departamento". Por esto, nos pareció que tenía sentido dejarlos en la misma relación, aún sabiendo que esto desafiaba la 3FN, la cual no permite dependencias funcionales transitivas (id_departamento → id_prov → provincia).

Por falta de completitud en el atributo "mail” de la fuente de BP, decidimos omitir del cálculo sobre el atributo "dominio más frecuente” a las BPs sin esta información. Lo cual entendemos que subrepresenta la frecuencia real.

Sección de Análisis de datos

Consultas SQL:

i)

df_res_ejercicio_1_final_comuna_agg_FINAL - DataFrame

Índice	Provincia	NOMDPTO	Jardines	poblacion_jardin	Primarios	poblacion_primario	Secundarios	poblacion_secundario
0	Buenos Aires	Almirante Brown	133	43762	137	58015	145	68125
1	Buenos Aires	Berazategui	101	25806	92	35310	96	41619
2	Buenos Aires	José C. Paz	55	29597	62	36232	62	40098
3	Buenos Aires	Pergamino	54	8676	56	10471	39	12182
4	Buenos Aires	Ituzaingó	50	10707	45	15026	49	18670
5	Buenos Aires	San Pedro	30	5781	42	7116	24	8280
6	Buenos Aires	Saladillo	25	2716	37	3213	13	3911
7	Buenos Aires	Patagones	24	2857	30	3677	15	3920
8	Buenos Aires	Rauch	18	1197	24	1501	7	1779
9	Chaco	Libertad	11	1282	12	1476	5	1563
10	Chaco	Tapenagá	5	394	10	437	5	510

(i) El texto explicativo y las correspondientes reflexiones de este reporte se hacen junto con su representación gráfica en el ejercicio ii) de visualización.

ii)

df_res_ejercicio_2_ii - DataFrame

Índice	Provincia	NOMDPTO	Cantidad de BP fundadas de 1950
235	Neuquén	Confluencia	25
261	Salta	Capital	20
94	Chaco	San Fernando	17
306	Santa Fe	Rosario	16
135	Córdoba	Capital	14
251	Río Negro	General Roca	14
82	Caba	Promedio Comunas	12
307	Santa Fe	Castellanos	12
0	Buenos Aires	Moreno	11
1	Buenos Aires	Tigre	11

(ii) La tabla se muestra ordenada de forma descendente, donde podemos notar que Neuquén es donde se encuentra la mayor cantidad de bibliotecas “modernas”.

iii)

df_res_ejercicio_3_final_comuna_agg_FINAL - DataFrame

Índice	Provincia	NOMDPTO	Cant_EE	Cant_BP	poblacion_total
0	Buenos Aires	Olavarría	174	11	122011
1	Buenos Aires	Adolfo Alsina	53	2	17344
2	Buenos Aires	Alberti	29	2	12906
3	Buenos Aires	General Lavalle	19	1	4846
4	Catamarca	Belén	127	5	30353
5	Catamarca	Pomán	45	3	12219
6	Catamarca	La Paz	97	2	26341
7	Corrientes	San Roque	61	2	22682
8	La Pampa	Loventué	20	3	9243
9	La Pampa	Catriló	15	3	8225
10	Mendoza	San Martín	168	4	139406

(iii) El texto explicativo y las correspondientes reflexiones de este reporte se hacen junto con su representación gráfica en el ejercicio iv) de visualización.

iv)

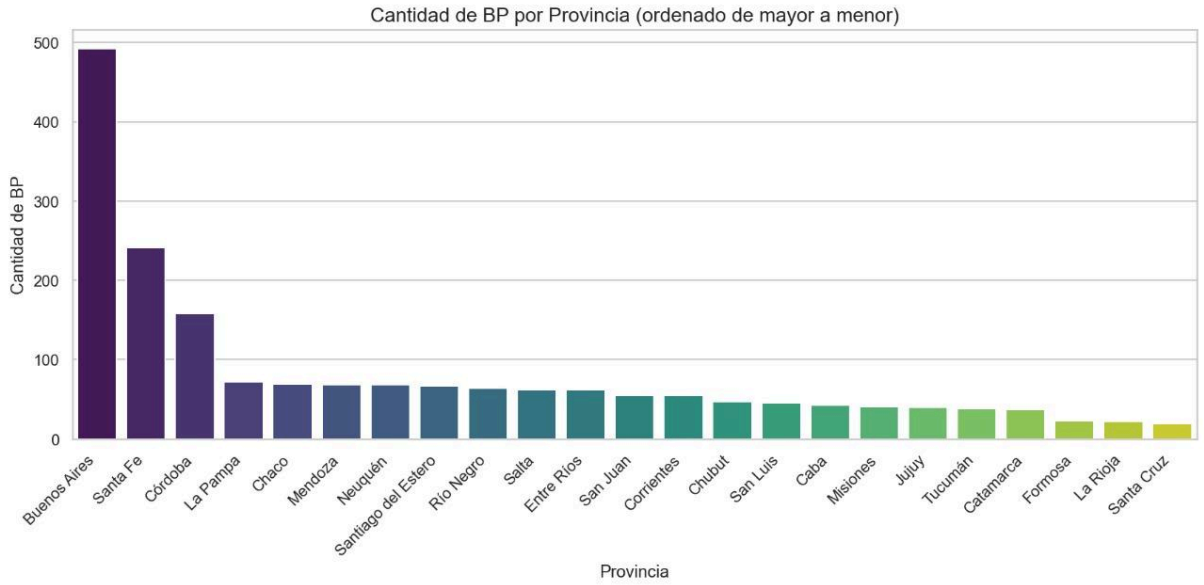
df_res_ejercicio_4_ix - DataFrame

Índice	Provincia	NOMDPTO	Dominio mas frecuente en BP
0	Buenos Aires	Esteban Echeverría	gmail
1	Tucumán	Capital	gmail
2	Entre Ríos	Diamante	yahoo
3	Mendoza	Tunuyán	gmail
4	Río Negro	Bariloche	gmail
5	La Pampa	Maracó	yahoo
6	Chubut	Rawson	hotmail
7	Córdoba	Calamuchita	gmail
8	San Luis	Junín	gmail
9	Buenos Aires	Bahía Blanca	yahoo
10	Misiones	Capital	gmail

(iv) Para este reporte solo se consideraron los dominios de gmail, hotmail y yahoo, dado que la frecuencia de dominios personalizados era tan baja que no incidían en los resultados. Como mencionamos con anterioridad existe una subrepresentación de dominios por departamento dada la incompletitud de la base de datos.

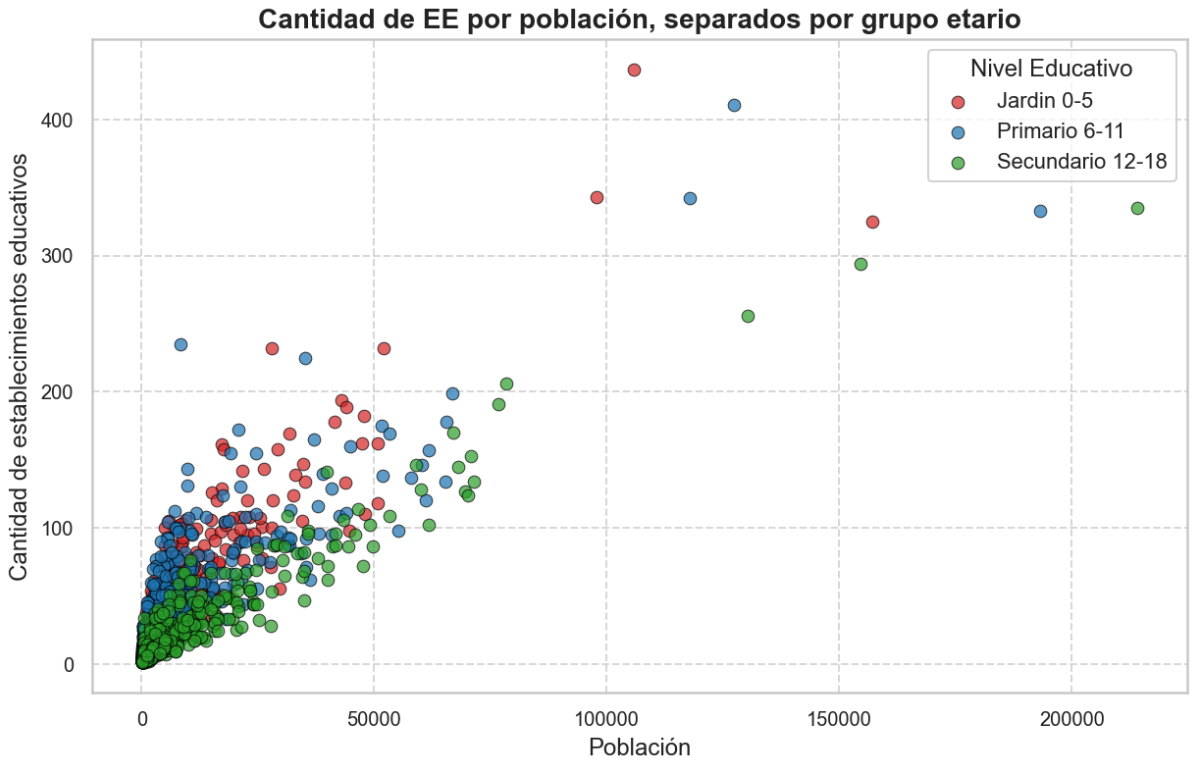
Visualización de datos:

i)



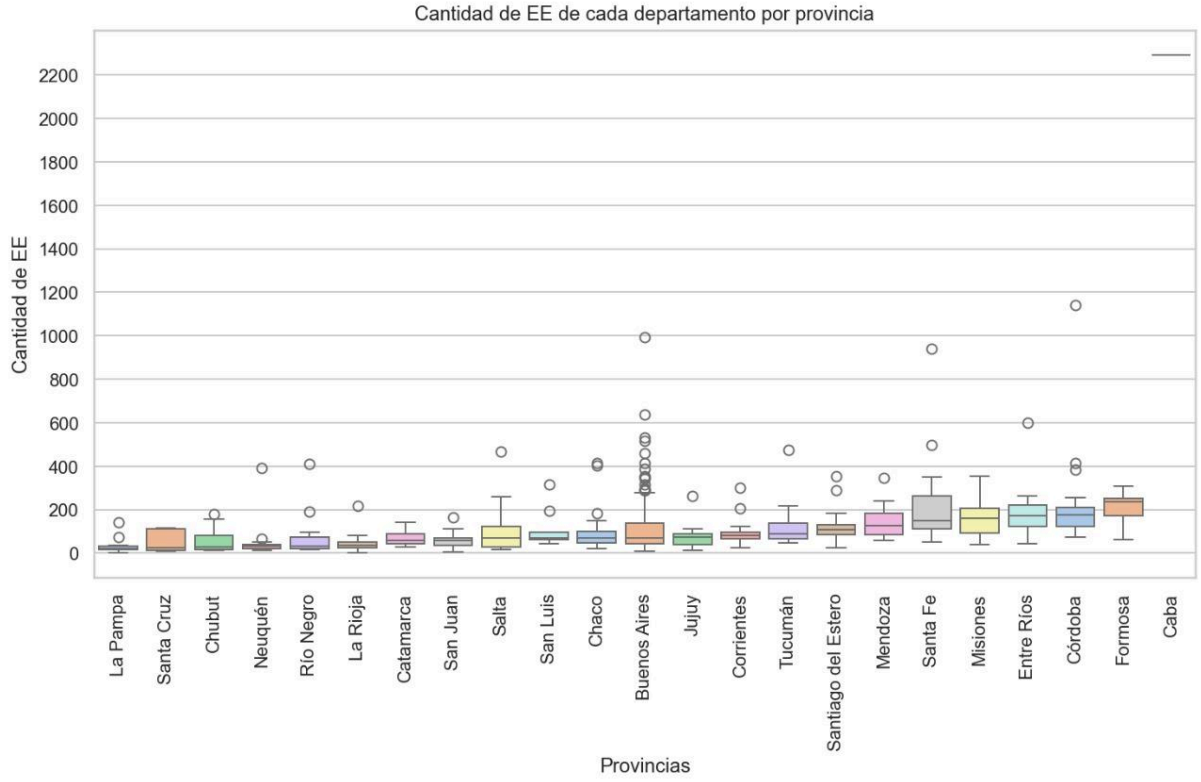
(i) Encontramos una desproporcionada cantidad de BPs en las provincias de Buenos Aires, Santa Fe y Córdoba con respecto al resto, lo cual coincide con las concentraciones demográficas más altas del país.

ii)



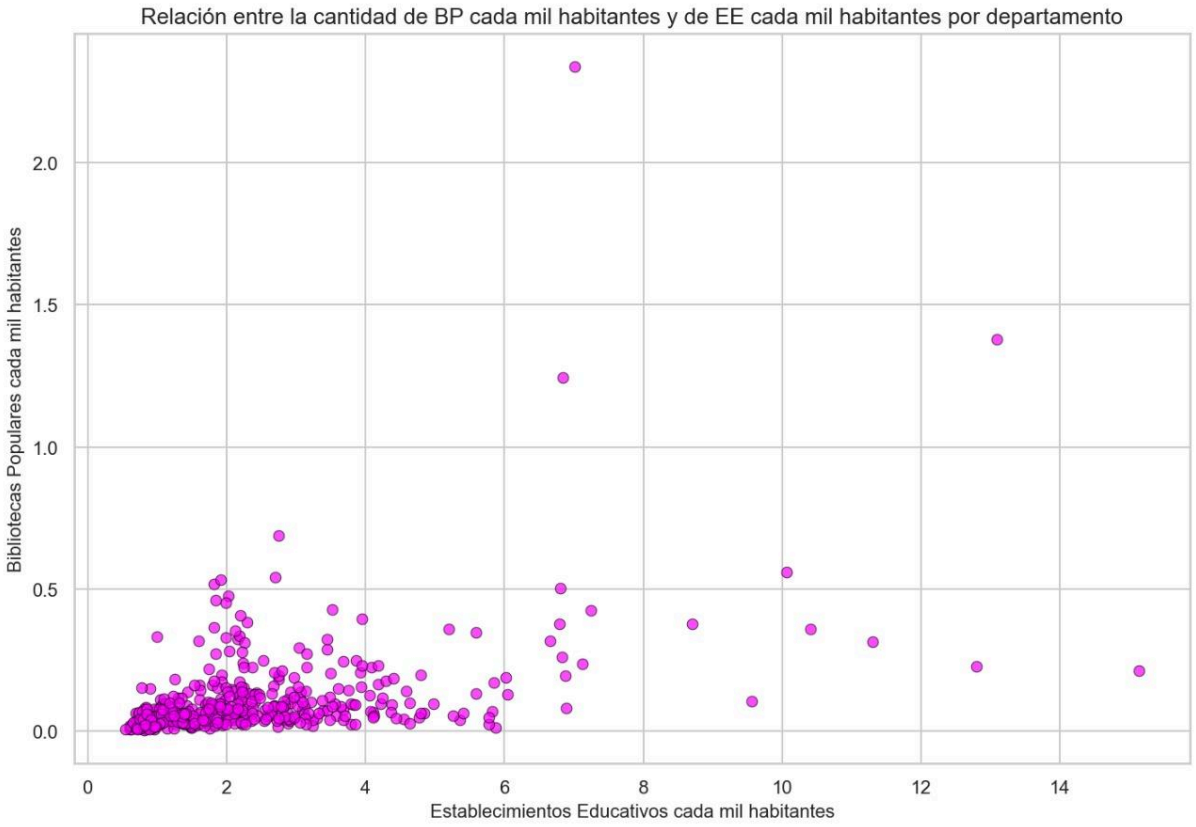
(ii) Parece haber una relación positiva más fuerte entre la cantidad de población y cantidad de EE en el grupo jardín y primario, y una relación más débil en el grupo secundario. Hay algunas razones que podrían explicar esta diferencia en el contexto argentino, entre ellas se puede destacar que la educación primaria es obligatoria y universal por ley desde hace décadas, no así la educación secundaria (comenzó a ser obligatoria a partir de la sanción de la Ley de Educación Nacional del 2006). Por otro lado, los establecimientos secundarios suelen tener una capacidad más grande que sus contrapartes primarias, lo que disminuye su número por cantidad de población.

iii)



(iii) Parece haber una cantidad mayor de EE en las regiones centrales y del noreste, y una menor cantidad en el noroeste y sur del territorio argentino. Tras consultar un mapa de densidad de poblaciones, podemos confirmar que el resultado obtenido en nuestro gráfico parece coincidir con la distribución de la densidad demográfica por regiones del país. Lo que hace a la coherencia de nuestra interpretación original.

iv)



(iv) La inspección visual del scatter plot parece indicar una falta de correlación entre las variables de interés, o a lo sumo una débil correlación positiva. Los data points se encuentran concentrados en el cuadrante inferior izquierdo, pero esta observación no parece ser muy representativa, ya que la escala de los gráficos se generó para incluir todos los datos, aun los que están muy por fuera de la media. Si hacemos omisión de estos outliers y trazamos una recta, seguramente esta tendría una leve pendiente positiva. Para confirmar esta hipótesis habría que hacer un análisis estadístico de los datos.

Conclusiones:

En respuesta a la pregunta que motivó este trabajo: Existe o no cierta relación entre la cantidad de BP y los EE en los departamentos del país? Primero tenemos que hacer un comentario sobre el alcance de la herramienta analítica aquí utilizada, puntualmente el de un análisis cualitativo. El beneficio de este tipo de análisis es la posibilidad de encontrar rápidamente una tendencia en los datos presentados. Es muy útil para generar hipótesis, que luego deben ser confirmadas con análisis cuantitativos. En otras palabras, es un eslabón importante del proceso analítico, pero por sí mismo es limitado. Esto queda demostrado en el gráfico del ejercicio (ii) de visualizaciones, donde hay una superposición de datos que impide encontrar diferencias entre poblaciones de jardín y primaria.

Para el manejo y análisis de grandes cantidades de datos, en general no basta con un análisis cualitativo. El gold-standard para estos casos es un análisis estadístico de los datos obtenidos, y así poder discernir coeficientes de correlación y su significancia estadística. Por ejemplo, la inspección visual del último gráfico no permite diferenciar si verdaderamente existe una relación entre la cantidad de EE y BP por departamento. En cambio, si trazamos una curva, podríamos conocer su pendiente aproximada, la cual nos daría una idea más precisa de la naturaleza de la relación (o falta de).

En conclusión, la inspección visual del gráfico (iv) sugiere una débil relación positiva entre las BP y los EE en los departamentos del país, pero para confirmar o rechazar esta hipótesis habría que hacer un análisis estadístico.