Trabajo Práctico N° 1

Análisis de la Relación entre Establecimientos Educativos y Centros Culturales en Argentina

Laboratorio de datos, verano 2025
Departamento de Computación, FCEyN, UBA

El presente trabajo tiene como objetivo analizar la probable relación entre la cantidad de establecimientos educativos y centros culturales en las diferentes provincias y departamentos de Argentina.

RESUMEN

Este informe tiene como objetivo analizar si existe una relación entre la cantidad de centros culturales y de establecimientos educativos en Argentina, para eso se contó con los datos del padrón de establecimientos educativos (EE), el padrón de centros culturales (CC) y el padrón de población, todos los datos de estas fuentes son del año 2022.

No está claro si existe una relación entre la cantidad de estos establecimientos y la densidad poblacional de cada región. Inicialmente, se realizó un análisis exploratorio de las fuentes para evaluar la calidad de los datos y definir un modelo relacional adecuado. Para ello, se diseñó un diagrama entidad-relación (DER) en tercera forma normal con el objetivo de garantizar la integridad y evitar redundancias. Luego se armó a partir del DER el modelo relacional y se realizaron consultas SQL para obtener información relevante, así como también gráficos para ver tendencias.

Finalmente se pudieron obtener conclusiones sobre la distribución de EE y CC en Argentina.

INTRODUCCIÓN

El objetivo del trabajo se basa en explorar la posible relación entre los EE y los CC en el país. Por esto en primer lugar se exploraron los datos para conocer cómo estaban organizados, en este proceso se encontraron muchos problemas de calidad de datos en los registros, los cuales detallaremos en la sección de procesamiento de datos.
Para tener una buena organización se realizó un DER que vincula las diferentes fuentes de datos, se propuso que el DER esté en tercera forma normal para mantener la integridad, evitar tuplas espurias y mejorar la consistencia evitando redundancias.
A partir del DER se armó el modelo relacional equivalente obteniendo un esquema claro a partir del cual trabajar. Luego pasamos a un proceso de filtrado de los datos originales para obtener las correspondientes relaciones del DER limpias, donde finalmente con ellas se trabajará para construir reportes, tablas y gráficos específicos en busca de obtener una respuesta a la pregunta.

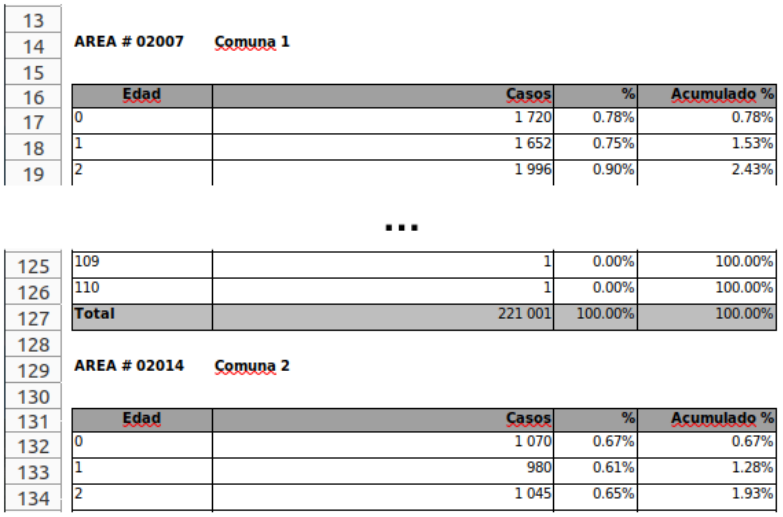
PROCESAMIENTO DE DATOS

Esta sección detalla el estado inicial de las fuentes de datos, los problemas de calidad detectados, los procesos de limpieza aplicados y la construcción del modelo relacional a partir del diagrama entidad-relación (DER). La primera etapa consistió en realizar un análisis exploratorio de las fuentes de datos, con el objetivo de evaluar las formas normales y la calidad de los datos. A continuación realizamos dicho análisis para cada fuente de datos.

Padron de poblacion

Esta información viene en un archivo excel, con los datos de cuántas personas con cada año de edad existen en cada departamento, sin embargo no está separado por id del departamento al que pertenecen, si no que en la primera columna hay un código de área y en la segunda el nombre del departamento. Además al terminar una tabla y separado por filas está la información de otro departamento (ver figura 1). El código de área guarda información sobre el departamento: los últimos 3 dígitos son el id del departamento, y los primeros 2 el id de la provincia en que se encuentra. Entonces cada código de área hace

referencia a la tabla que está debajo. Los datos de esta fuente no están en primera forma normal, ya que la información de cada departamento no está separada por filas y además en una misma columna se mezclan distintos tipos de datos, por ejemplo edades con códigos de área.



13				
14	**AREA # 02007**	**Comuna 1**		
15				
16	**Edad**	**Casos**	**%**	**Acumulado %**
17	0	1 720	0.78%	0.78%
18	1	1 652	0.75%	1.53%
19	2	1 996	0.90%	2.43%
	...			
125	109	1	0.00%	100.00%
126	110	1	0.00%	100.00%
127	**Total**	221 001	100.00%	100.00%
128				
129	**AREA # 02014**	**Comuna 2**		
130				
131	**Edad**	**Casos**	**%**	**Acumulado %**
132	0	1 070	0.67%	0.67%
133	1	980	0.61%	1.28%
134	2	1 045	0.65%	1.93%

Figura 1: se muestra el formato y estado en que se encuentra la tabla de datos original.

El principal atributo afectado es entonces el código de área, que tiene problemas de modelo ya que no separa bien el código de área de los demás atributos. Se puede usar el método GQM para evaluar este atributo.
Goal: Evaluar el modelo del padrón de población.
Question:
¿Qué porcentaje de registros en la tabla corresponde a filas que no contienen información de población, sino que funcionan como delimitadores entre bloques de datos (código de área, encabezados, filas vacías, etc.)?
Metric:
Se cuenta la cantidad de filas que no contienen información de la población y se calcula sobre el porcentaje de filas total.

$$\frac{Filas\ sin\ información\ sobre\ población}{Filas\ totales} = \frac{2643}{56703} = 4.7\ \%$$

Para corregirlo se armó la relación Población con los atributos id_prov, id_depto, Edad, Casos.

Padrón de centros culturales
La información de este archivo viene en formato CSV, posee varios atributos que no brindan información importante, como por ejemplo “Observaciones” e “InfoAdicional” los cuales poseen todas valores “nan”, o el atributo “Categoría” con un valor constante de “Centro cultural”. Posee además como atributos el id de departamento y el id de provincia, sin embargo tiene todos los datos de C.A.B.A englobados bajo un único id de departamento, no separando por comunas. También posee el nombre de provincia y departamento, algo que se podría separar en otra tabla para evitar redundancia. No posee un identificador único para cada centro cultural, la clave primaria de la tabla serían todos los atributos. Por este motivo se decidió generar una clave primaria numérica (id_cc).
Los atributos que se extraen son el id_prov, id_depto, id_cc, la Capacidad y el Mail. Se observó un formato inconsistente en los mails, en muchos casos se encontraron dos mails

en una misma celda, por lo que no está en primera forma normal, pero además para separar los mails, en muchos casos se usaron espacios y en otros comas, habiendo también problemas de instancia ya que se encontraron casos de mails inválidos como por ejemplo sin carácter arroba, o con caracteres de espacio dentro de la dirección de mail, lo cual no está permitido como dirección. Además hay celdas sin Mail indicadas por valores null pero también campos con “s/d” o “-” siendo ambiguo el significado. Se puede usar el método GQM para medir la calidad del atributo Mail.
Goal: Evaluar la calidad del atributo Mail viendo que los correos electrónicos sean válidos.
Question: ¿Qué porcentaje de registros en la tabla contiene mails inválidos o ambiguos?, ¿Cuántos registros presentan múltiples correos en una misma celda?

Metric: para los correos inválidos se cuenta el número de celdas donde el atributo Mail no contiene un "@", presenta espacios internos o es ambiguo. Para los casos con múltiples correos se calcula el porcentaje de celdas con más de un correo

$$\frac{Correos\ inválidos}{Total\ de\ tuplas} = 12.1\ \%$$

$$\frac{Múltiples\ correos}{Total\ de\ tuplas} = 1.5\ \%$$

Para corregirlo se decidió que los valores ambiguos o sin arroba dejarlos como null, en el caso de múltiples correos se tomó solo el primero, y en el caso de los que tenían espacios se decidió quitar el espacio.

Padrón de establecimientos educativos
La información de este archivo se encuentra en formato Excel, y contiene datos sobre cada EE, como su modalidad (común, especial, adultos, etc.) dentro de la cual puede haber distintos tipos de establecimientos. Luego hay más atributos como la jurisdicción, nombre, sector, ámbito, dirección, código de localidad, etc.

Para la construcción del modelo relacional, se seleccionaron los atributos necesarios: id_prov, id_depto, cueanexo, sector, ámbito, tipo_establecimiento (de modalidad común: jardín maternal, jardín de infantes, primario, secundario y secundario técnico).

La tabla original no está en 3FN, aunque cumple con 1FN (todos los atributos contienen valores atómicos) y 2FN (no hay dependencias parciales con la clave primaria cueanexo). Sin embargo existen dependencias funcionales transitivas, por ejemplo, el id del departamento junto al de provincia determinan el nombre del departamento.

Uno de los principales problemas de calidad es la incompletitud de los datos, los establecimientos educativos pueden tener varios tipos de establecimiento (jardín maternal, secundario, etc.) y por el diseño de la tabla ocurre que hay muchas celdas vacías. Se usó el método GQM para ver el impacto de esto.

Goal: Identificar el porcentaje que ocupan las celdas sin información sobre los tipos de establecimiento.
Question: ¿Cuál es el porcentaje de celdas que no aportan información?
Metrics: Para las columnas que indican la presencia de establecimientos de cada tipo (jardín maternal, jardín de infantes, primario, secundario, secundario técnico), se cuenta cuántas celdas no contienen un "1" (el cual indica un establecimiento de ese tipo), y se divide sobre la cantidad de celdas ocupadas.

$$\frac{Cantidad\ de\ no\ unos}{5*longitud\ de\ tabla} = 75.2\ \%$$

Viendo todo esto, para mejorar la estructura y evitar redundancia se propuso dividir la información en tres tablas:

- Establecimientos_educativos: cueanexo, id_prov, id_depto
- ee_tipo_establecimiento: cueanexo, id_tipo_establecimiento
- tipos_de_establecimientos: id_tipo_establecimiento, tipo_establecimiento

donde id_tipo_establecimiento se creó para identificar cada tipo_establecimiento.

DER

Se diseñó un DER para estructurar y organizar la información de manera clara, estableciendo las relaciones entre los datos provenientes de distintas fuentes. Esto permitió eliminar redundancias, garantizar la integridad de los datos y facilitar la posterior normalización. A continuación en la figura 2 se detalla cómo quedó realizado el DER, en la figura 3 se muestra el modelo relacional del DER, y en la figura 4 las dependencias funcionales de las relaciones del DER.

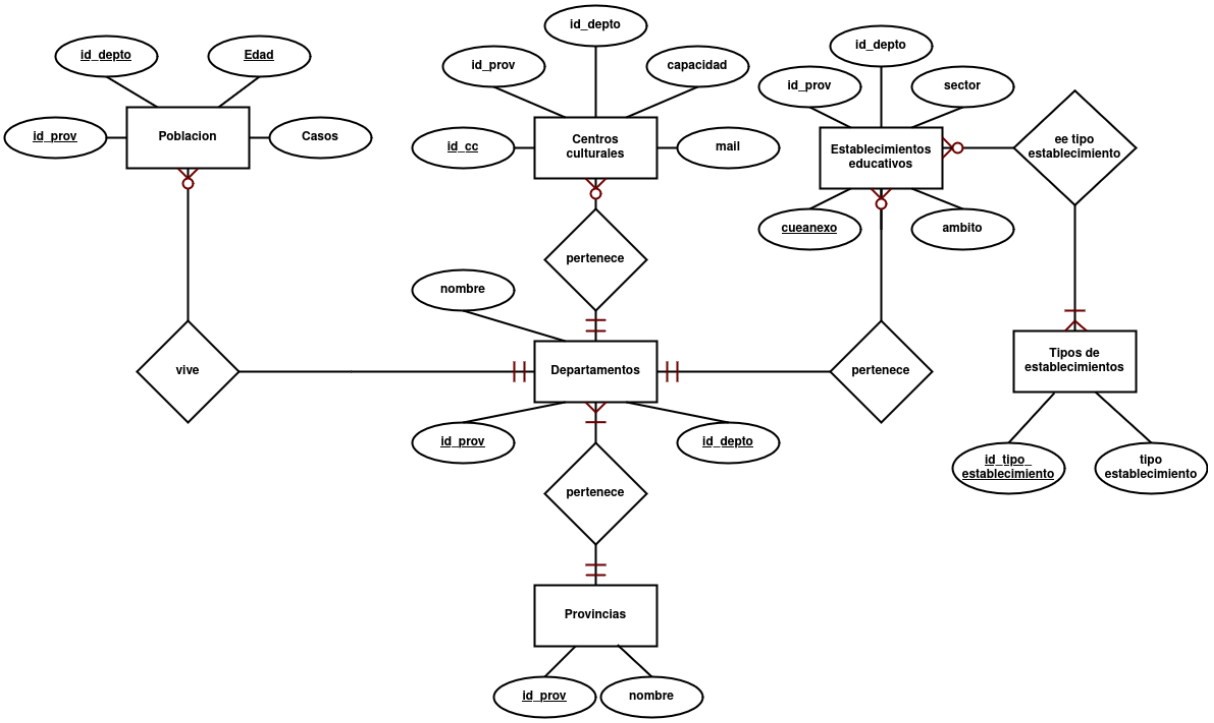


Figura 2: *Se puede observar el DER (Diagrama - Entidad - Relación)* modelado.

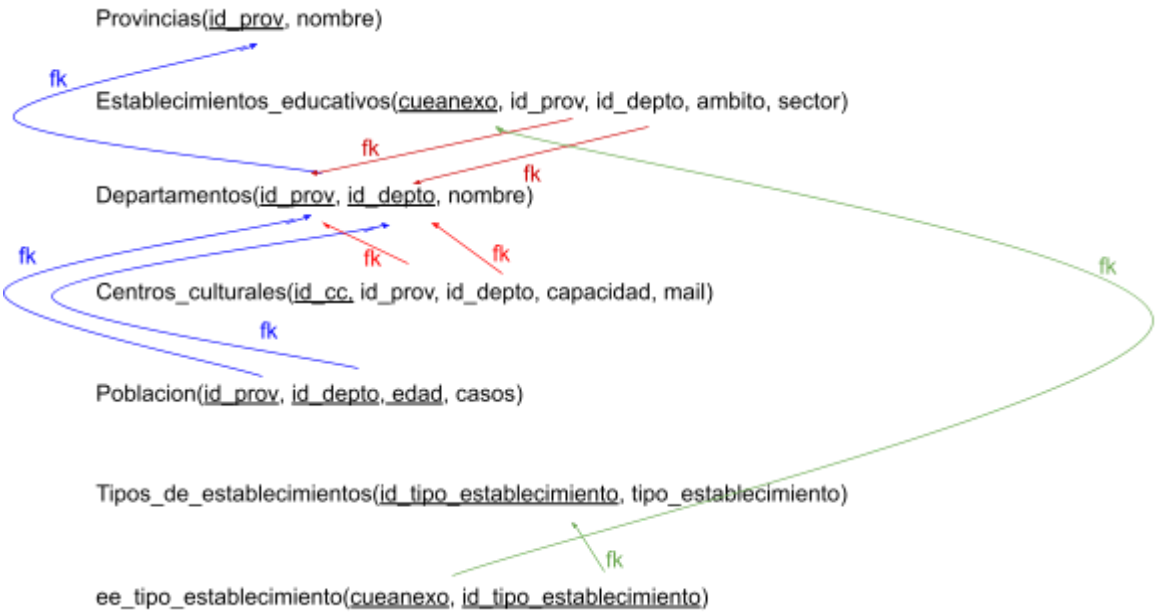


Figura 3: *Modelo Relacional equivalente para el DER de la figura 2.*

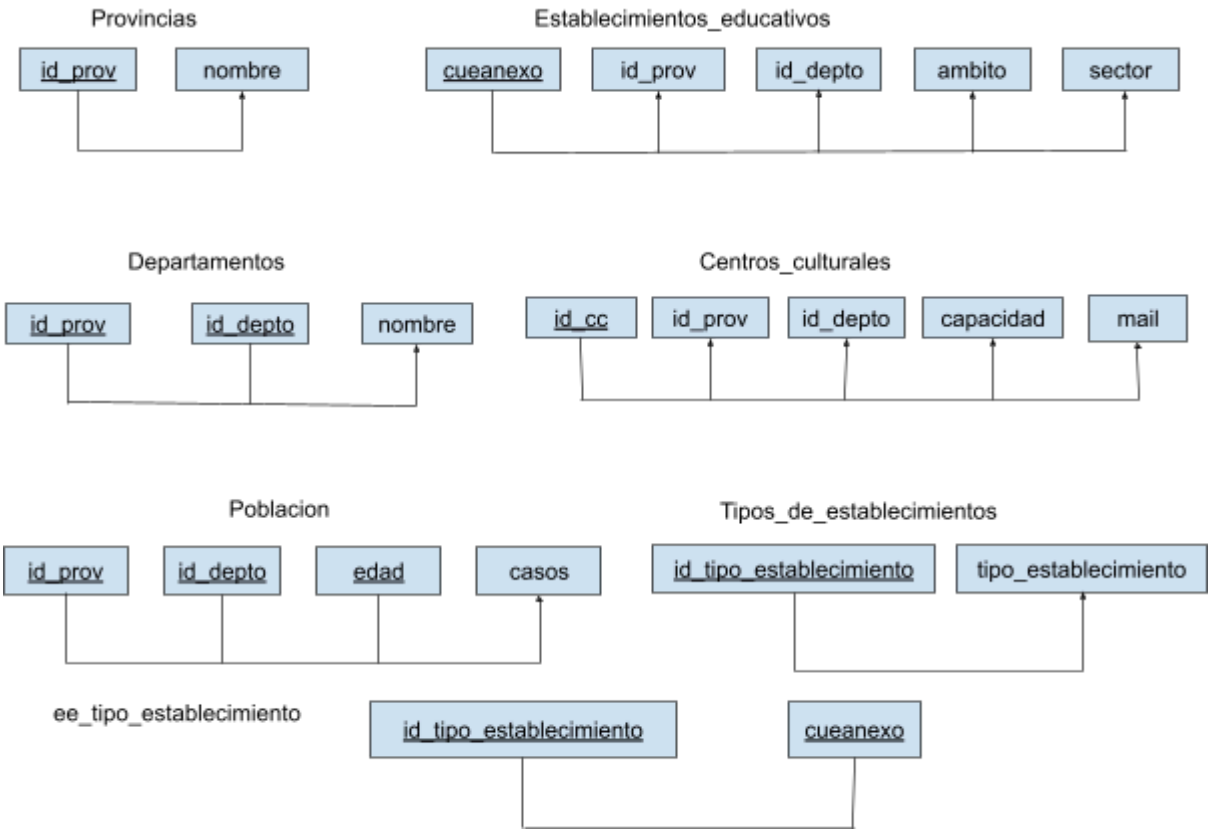


Figura 4: *Dependencias funcionales de los atributos.*

Para la construcción de las tablas, se realizaron diversas transformaciones a partir de las fuentes originales mediante consultas SQL y el uso de Pandas.

- **Tabla Población:** De la tabla padrón de población se separó la información del código de área en los atributos id_prov e id_depto, lo que permitió vincular los atributos Edad y Casos a cada departamento.

- **Tabla Centros Culturales:** Del atributo "ID_DEPTO" se extrajo id_depto, ya que contenía tanto el id de la provincia como el del departamento (siendo estos últimos, los últimos tres caracteres). También se incluyó id_prov, y por simplicidad se generó una clave primaria numérica id_cc, además se extrajeron los atributos de Mail y Capacidad. Para los casos en que las celdas contenían varios mails se dejó solo el primero, y donde se detectó que no estaba el carácter arroba se reemplazó por el valor null.
- **Tabla Tipos de Establecimientos:** Se generó una tabla que vincula cada tipo de establecimiento con un identificador único (id_tipo_establecimiento), esto con el fin de eliminar muchas celdas vacías del diseño original.
- **Tabla ee_info:** Se creó una tabla intermedia denominada ee_info, donde se extrajo id_prov e id_depto a partir del atributo Código de Localidad de la tabla establecimientos educativos. Además, se vincularon estos identificadores con el atributo Jurisdicción (que indica el nombre de la provincia o CABA) y con el nombre del departamento. También se extrajeron los atributos correspondientes a la modalidad común para jardines, primario y secundarios. De esta tabla se extrajeron los atributos para armar Provincias, Departamentos, Establecimientos educativos y ee_tipo_establecimiento.

DECISIONES TOMADAS

En esta sección vamos a aclarar las decisiones que tomamos para la realización de las consultas.
- En Argentina las divisiones de segundo orden consisten en 380 departamentos, 135 partidos (Buenos Aires) y 15 comunas (C.A.B.A), en el trabajo se las llamará a todas por departamento. Cada departamento posee un id, al igual que cada provincia, sin embargo en distintas provincias puede existir un mismo id de departamento, pero la combinación del id de provincia con el id del departamento es un identificador único. Por este motivo para vincular los datos se decidió usar estos ids.
- Para calcular la cantidad de establecimientos educativos, se simplificó la clasificación. Los jardines maternales y de infantes se agruparon bajo la categoría de jardines, mientras que los secundarios técnicos se unificaron con el nivel secundario. Para calcular la población en cada nivel educativo se usó esta clasificación, además se tuvo en cuenta la diferencia en la duración del ciclo primario y secundario según la provincia. En algunas provincias, el nivel primario dura 7 años, mientras que en otras dura 6. De manera similar, el nivel secundario varía entre 5 y 6 años según la jurisdicción. Además, se consideró la población de las escuelas técnicas, dado que su duración es de un año más que la del secundario común.
- Se detectó un problema de calidad en los id de departamento para Ushuaia y Río Grande, en Población tiene los valores 8 y 15 respectivamente mientras que en Establecimientos Educativos y Centros culturales poseen los id 7 y 14. Se cambiaron los id de Población para que coincidan.
- Decidimos generar una clave primaria numérica para la tabla de centros culturales por simplicidad, ya que no poseía una clave primaria para los atributos que íbamos a extraer que no sea la combinación de todos los atributos.

ANÁLISIS DE DATOS

Consultas en SQL

Para la primera consulta se tuvo en cuenta cada rango etario de la población por separado, es decir; Jardin por un lado (de 0 a 5 años) Primarios por otro (de 6 a, 12 o 13 dependiendo la provincia) y Secundario (de 13 o 14, a 19 también dependiendo la Provincia). Se observa que las ciudades con mayor población en las edades correspondientes a los niveles secundario, primario o jardín también presentan una mayor cantidad de establecimientos educativos, algo esperable. Además, se nota una cierta tendencia entre la cantidad de establecimientos y la cantidad de alumnos. Al analizar algunos valores, parece que los jardines tienen menos alumnos por establecimiento, mientras que las escuelas primarias suelen tener una mayor cantidad de estudiantes en comparación con los secundarios y jardines. Se pueden ver los primeros resultados de la consulta en la figura 5.

Índice	Provincia	Departamento	Jardines	Poblacion_jardin	Primarias	Poblacion_primaria	Secundario	Poblacion_secundaria
0	BUENOS AIRES	LA MATANZA	378	157034	333	225872	352	209596
1	BUENOS AIRES	LA PLATA	274	52075	199	77998	219	80263
2	BUENOS AIRES	LOMAS DE ZAMORA	201	50825	178	76967	205	75204
3	BUENOS AIRES	GENERAL PUEYRREDON	229	41427	169	62565	179	67623
4	BUENOS AIRES	QUILMES	203	47353	146	70881	163	69693
5	BUENOS AIRES	ALMIRANTE BROWN	156	43762	137	67913	153	67462
6	BUENOS AIRES	MORENO	127	50791	134	76357	143	69698
7	BUENOS AIRES	MERLO	114	48007	120	71541	129	69008
8	BUENOS AIRES	LANUS	146	28133	116	44506	124	46282
9	BUENOS AIRES	TIGRE	150	32611	111	51961	117	52412

Figura 5: *primeras filas de la tabla con la cantidad de Población de Jardin, Primario y Secundario y la Cantidad de EE de cada uno por Departamento. La información completa de la consulta se encuentra bajo el nombre Consulta1 en el código del Informe.*

Para la segunda consulta se observó que solo hay 56 departamentos en los cuales hay centros culturales con capacidad mayor a 100 (considerando todo CABA como un único departamento). Se pueden ver los primeros resultados en la figura 6.

Índice	Provincia	Departamento	Cantidad_mayor_100
0	BUENOS AIRES	AVELLANEDA	20
1	BUENOS AIRES	LA PLATA	8
2	BUENOS AIRES	LOMAS DE ZAMORA	3
3	BUENOS AIRES	GENERAL PUEYRREDON	2
4	BUENOS AIRES	ALMIRANTE BROWN	2
5	BUENOS AIRES	MERCEDES	1
6	BUENOS AIRES	BRAGADO	1
7	BUENOS AIRES	OLAVARRIA	1
8	BUENOS AIRES	TRES DE FEBRERO	1
9	CHACO	SAN FERNANDO	2

Figura 6: *primeras filas la tabla que cuenta departamentos con centros culturales con capacidad mayor a 100 personas. La información completa de la consulta se encuentra bajo el nombre Consulta2 en el código del Informe.*

Para la tercera consulta mostramos la cantidad de Centros Culturales, Establecimientos Educativos y población por departamento (ver figura 7). Se puede observar que no hay una relación muy clara entre la población y los centros culturales, por ejemplo se ve que en La Matanza solo hay 2, teniendo más población que Córdoba Capital donde hay 30. En cambio si se ve una relación del tipo mayor población implica más cantidad de EE.

Index	Provincia	Departamento	Cantidad_ee	Cantidad_cc	Personas
0	CIUDAD DE BUENOS AIRES	Ciudad Autónoma de Buenos Aires	1782	296	3.09545e+06
1	CÓRDOBA	CAPITAL	1136	30	1.49806e+06
2	BUENOS AIRES	LA MATANZA	977	2	1.83717e+06
3	SANTA FE	ROSARIO	817	36	1.33796e+06
4	BUENOS AIRES	LA PLATA	669	72	756074
5	BUENOS AIRES	LOMAS DE ZAMORA	548	17	685644
6	BUENOS AIRES	GENERAL PUEYRREDON	535	10	660569
7	BUENOS AIRES	QUILMES	464	10	631774
8	SANTA FE	LA CAPITAL	463	18	568259
9	CHACO	SAN FERNANDO	458	13	413764

Figura 7: *primeras filas de la Cantidad de CC, EE y Población para cada Departamento. La información completa está bajo el nombre Consulta3 en el código.*

Por último, para la cuarta consulta se vió que dominio de mail es el más frecuente para cada Centro Cultural de un determinado Departamento y Provincia, en caso de haber empate se muestran todos. De acá se llegó a la conclusión de que los más usados son gmail, hotmail y yahoo. En la figura 8 se pueden ver algunas filas de la consulta.

Índice	Provincia	Departamento	Dominio_mas_frecuente
0	BUENOS AIRES	GENERAL SAN MARTIN	gmail.com
1	BUENOS AIRES	DOLORES	gmail.com
2	BUENOS AIRES	GENERAL JUAN MADARIAGA	hotmail.com
3	BUENOS AIRES	BRANDSEN	gmail.com
4	BUENOS AIRES	QUILMES	gmail.com
5	BUENOS AIRES	ADOLFO ALSINA	adolfoalsina.gov.ar
6	CÓRDOBA	COLON	gmail.com
7	CÓRDOBA	UNION	gmail.com
8	CORRIENTES	CAPITAL	gmail.com

Figura 8: *primeras filas de la tabla con los mails más frecuentes por departamento. La información completa de la consulta se encuentra bajo el nombre Consulta4 en el código del Informe.*

Visualización de datos con SQL y Python

Para la etapa de visualización utilizamos librerías gráficas de python y consultas en SQL.

Para la primera consigna se utilizó una consulta en SQL bajo el nombre de *visualizacion1* que indica la cantidad de Centros Culturales por provincia de manera decreciente (Figura 9) y para visualizarlo se utilizó un gráfico de barras horizontal como el que se muestra a continuación. De este gráfico se puede concluir que en Buenos Aires es donde se encuentra la mayor cantidad de Centros Culturales y que contando a CABA es en donde están la gran mayoría de centros culturales.

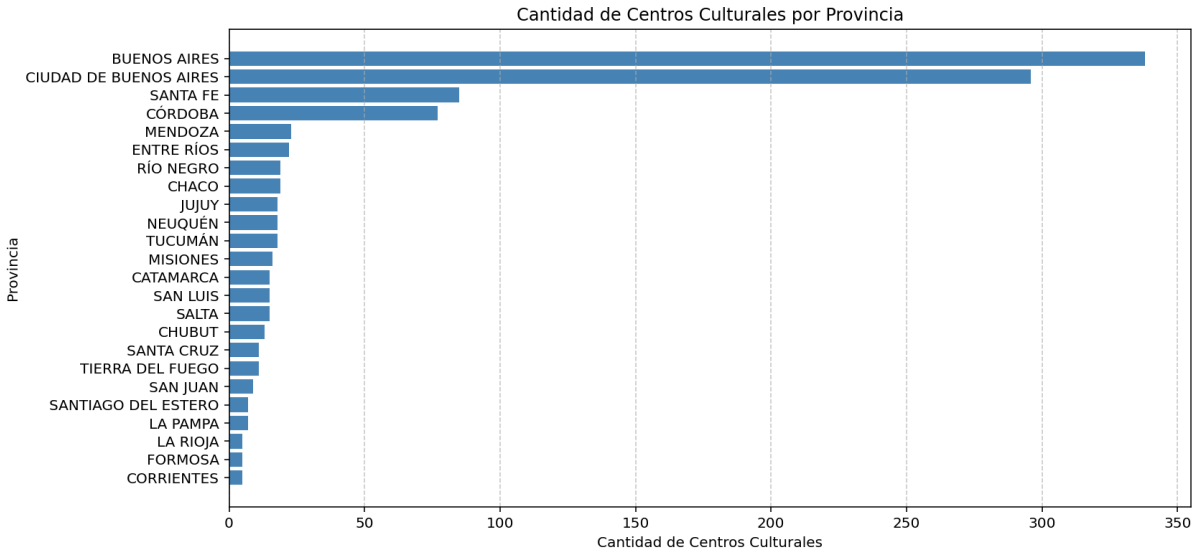


Figura 9: *Cantidad de Centros Culturales por Provincia*

A continuación se grafica la cantidad de EE en función de la población de cada nivel educativo (en colores) y para cada departamento del país. Así cada departamento posee un punto de cada color (rojo para jardín, verde para primario y azul para secundario).

Se modeló la relación con una recta para visualizar la tendencia en cada nivel educativo y analizar si existe una relación lineal entre la cantidad de establecimientos educativos y la población. La mayor pendiente de la recta de jardín indica que para una misma cantidad de población, hay más jardines que escuelas primarias o secundarias, y que se requieren menos primarias para albergar a un mayor número de alumnos (ya que tiene menor pendiente) si bien no hay una diferencia tan grande como con los jardines.

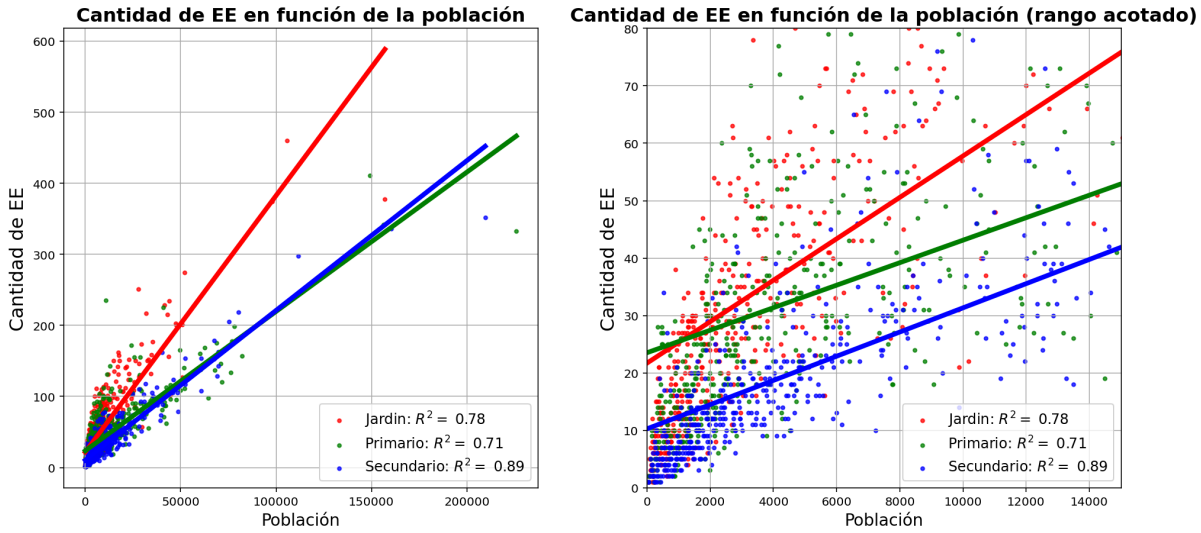


Figura 10: *El primero (izquierda) evalúa la cantidad de EE en función de la población y el segundo (derecha) es un aumento para un rango acotado (*0-15,000 habitantes) *.*

Para el siguiente gráfico (figura 11) utilizamos boxplots para mostrar la distribución de la cantidad de establecimientos educativos por departamento en cada provincia. Primero calculamos la cantidad de establecimientos educativos de cada departamento en todas las provincias, agrupamos los datos por provincia y obtuvimos la mediana. Luego, las provincias se ordenan según esta mediana. Si las cajas son más estrechas significa que la cantidad de establecimientos educativos en cada departamento es más parecida entre sí.

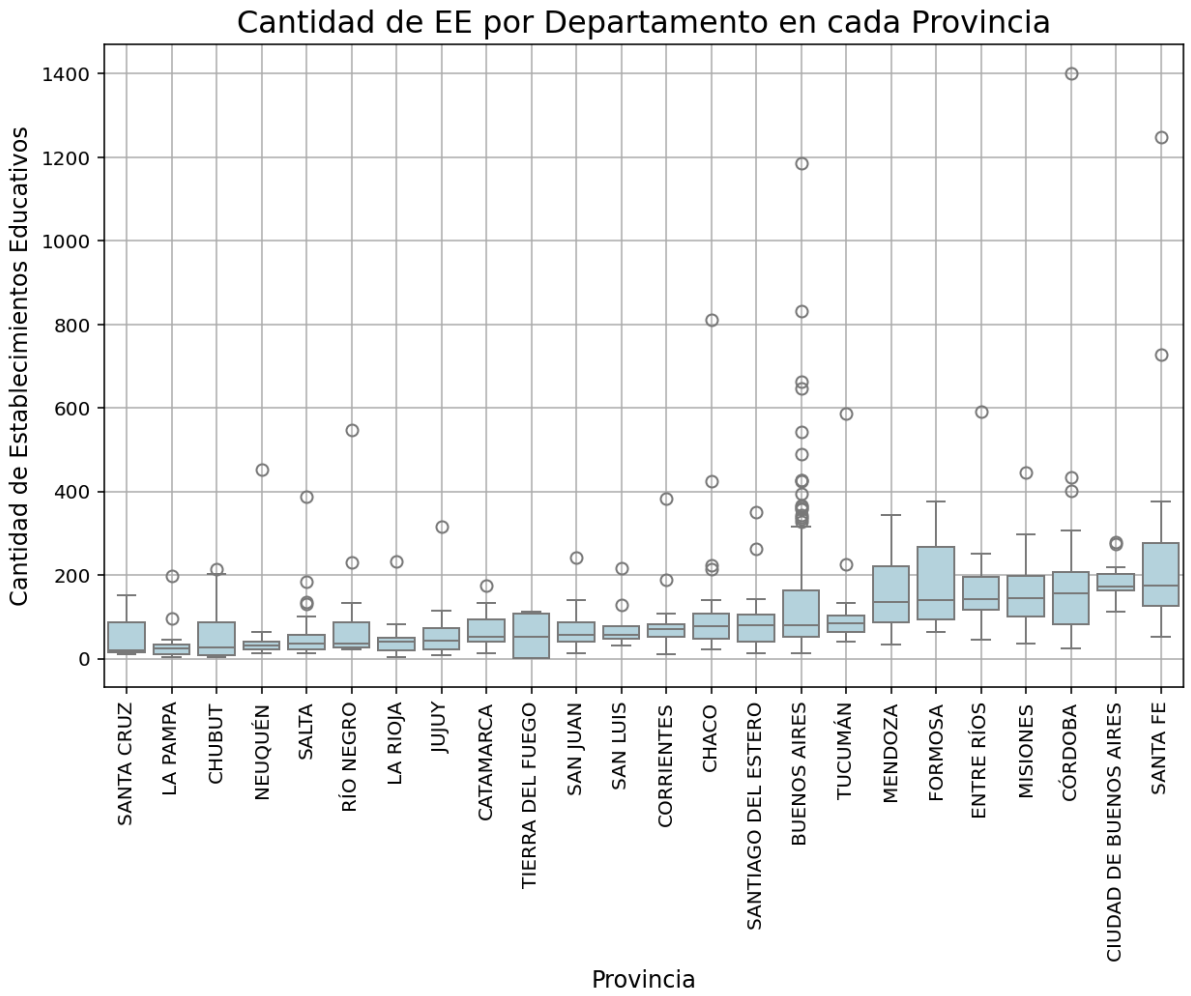


Figura 11: *Cantidad de Establecimientos Educativos por Departamento en cada Provincia.*

Para analizar la cantidad de EE y CC cada mil habitantes, realizamos una serie de consultas en SQL. Primero, obtuvimos la población total de cada provincia. Luego, contamos la cantidad de establecimientos educativos y centros culturales en cada provincia. Con estos datos, calculamos la cantidad de EE y CC por cada 1000 habitantes y ordenamos los resultados según la cantidad de EE. Finalmente, generamos dos gráficos de dispersión: el primero muestra la relación entre la cantidad de EE y CC (calculada como la cantidad de EE dividido por la cantidad de CC), y el segundo compara la cantidad de EE y CC por cada provincia (ver figura 12).

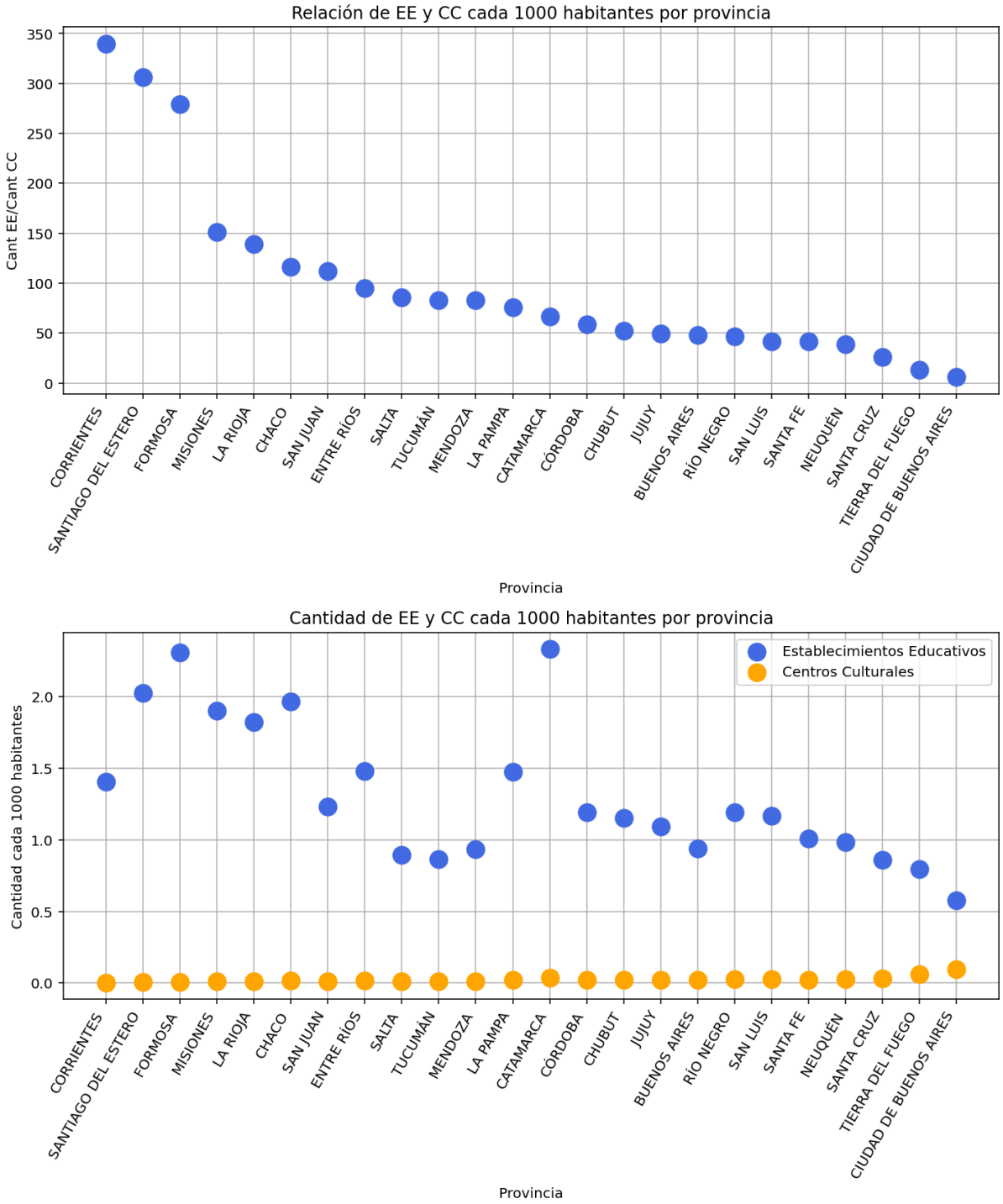


Figura 12: *El primero (arriba) muestra la relación EE/CC cada 1000 habitantes por provincia mientras que el segundo (abajo) muestra la cantidad de EE y CC cada 1000 habitantes.*

En ambos gráficos se puede ver que la cantidad de Establecimientos Educativos es considerablemente mayor a la de Centros Culturales, en el primero por el cociente y en el segundo en magnitud. Se ve que en provincias donde hay población más rural hay menos centros culturales, y que hay más centros culturales donde la población está más concentrada como en ciudad autónoma de buenos aires o tierra del fuego que tiene toda su población en 2 departamentos prácticamente.

CONCLUSIONES

A partir del análisis realizado, se concluye que existe una relación clara entre la cantidad de establecimientos educativos y la población de cada provincia, lo que era esperable dado que la demanda educativa crece con el número de habitantes. Sin embargo, no se observa la misma correlación con la cantidad de centros culturales, ya que su distribución no sigue un patrón poblacional definido. Esto se evidencia en que por ejemplo en La Matanza, donde pese a haber mucha población apenas hay CC, mientras que en Córdoba Capital hay muchos centros culturales.

Los gráficos de dispersión muestran que los jardines tienen una mayor cantidad de establecimientos por población en comparación con los niveles primario y secundario, lo que sugiere que requieren una distribución más densa para atender a los niños en etapas tempranas. Asimismo, las escuelas primarias suelen contar con más alumnos por establecimiento en relación con los secundarios, lo que se refleja en la pendiente de las rectas de tendencia.

El análisis de boxplots permitió visualizar la variabilidad en la cantidad de establecimientos educativos por departamento en cada provincia. Se observa que algunas provincias presentan una distribución más homogénea, mientras que otras tienen valores atípicos con departamentos que concentran muchos más establecimientos educativos que el resto.

Por otro lado, el estudio de la relación entre establecimientos educativos y centros culturales mostró una disparidad significativa entre provincias. En lugares como Tierra del Fuego y CABA, la proporción entre ambos tipos de instituciones es más equilibrada, mientras que en provincias como Santiago del Estero o Corrientes, la cantidad de establecimientos educativos supera ampliamente a la de centros culturales. Esto podría indicar diferencias en las políticas de infraestructura y el nivel económico de la provincia.