

## Trabajo Práctico nº 1

### Laboratorio de Datos

Nombre del grupo: Data Wizards

Integrantes: Valentina Morrone, Carlos Rafael Chaves Lopez, Jimena Jofré

LU: 35/24 - 19/23 - 696/23

Mails: valenmorrone@hotmail.com, rafael.chaves.lopez1@gmail.com,  
jimееjofee@gmail.com



*Facultad de Ciencias Exactas y Naturales,  
Universidad de Buenos Aires.  
Ciudad Universitaria - (Pabellón I/Planta Baja)  
Intendente Guiraldes 2610 -C1428EGA  
Ciudad Autónoma de Buenos Aires*



## **Resumen**

En este informe analizamos la relación entre la cantidad de establecimientos educativos y la cantidad de centros culturales en las provincias de Argentina. Para esto, recopilamos las fuentes de datos brindadas en la consigna, describimos su estructura y los problemas que genera cada una, creamos un modelo de datos apropiado para nuestro análisis, hicimos una limpieza de datos, y por último, creamos nuevas tablas y gráficos para intentar responder el interrogante. Concluimos que si bien no existe una correlación marcada entre ambas, las dos tienen un comportamiento creciente a medida que aumenta la cantidad de población en cada provincia.

## **Introducción**

Queremos analizar la existencia de una relación entre la cantidad de establecimientos educativos en cada una de las provincias de Argentina, y la cantidad de centros culturales. Para esto, utilizamos las tres tablas dadas en la consigna, que tienen información de los centros culturales, establecimientos educativos y del censo del 2022 realizado en Argentina. Luego de revisar la calidad de datos y los problemas que cada dataset presenta, diseñamos un diagrama de entidad-relación (DER) en tercera forma normal y su modelo relacional sólo con aquellos atributos que nos resulten útiles. Esto nos permite una representación clara y estructurada de la información, para poder entender mejor la problemática y luego trabajar en base a esto. Utilizamos herramientas de Python para realizar tareas de limpieza y mejora de los datos. Una vez que contamos con los datasets limpios y el DER, pasamos a realizar consultas y representaciones visuales que ayudan a cumplir con el objetivo del trabajo, utilizando librerías como Pandas y Matplotlib.

## **Procesamiento de Datos**

Lo primero que hacemos es importar las librerías necesarias para poder realizar el trabajo. Luego leemos las tablas, dadas en la consigna, en python y las almacenamos para poder utilizarlas más adelante. Las tablas importadas fueron centros culturales, establecimientos educativos y censo. Cada una contiene información acorde a su temática.

## **Análisis de las formas normales**

- Centros Culturales

La tabla de Centros Culturales no se encuentra en primera forma normal (1FN) ya que si bien no hay redundancia de datos, hay algunas columnas que no contienen valores



atómicos. Por ejemplo, el atributo *Mail* tiene dos mails asociados, separados por barras o comas. También podemos ver que en esta tabla hay dependencias transitivas, esto quiere decir que hay atributos no clave que dependen de otros atributos no claves. Por ejemplo, Provincia depende de ID PROV, Departamento depende de ID DEPTO y Localidad depende de Cod Loc. Al no estar en 1FN, tampoco está en 2FN, ni en 3FN. Por lo tanto, podemos concluir que esta tabla no se encuentra en ninguna de las tres formas normales.

- Establecimientos Educativos

La tabla de Establecimientos Educativos no se encuentra en primera forma normal (1FN) porque si bien no hay redundancia de datos, hay columnas que contienen valores atómicos. El atributo *Teléfono* tiene en varias ocasiones dos números de teléfono asociados separados por barras. Esta tabla también tiene dependencias transitivas, esto quiere decir que hay atributos no clave que dependen de otros que tampoco son clave, como por ejemplo Localidad depende de Código de localidad y C.P depende del Domicilio. Al igual que en la tabla anterior, podemos concluir que no se encuentra en ninguna de las tres formas normales.

### **Revisión de la calidad de datos de las fuentes y descripción de los problemas de calidad de datos encontrados.**

#### **Problemáticas en el dataset de Centros Culturales**

Queremos analizar la completitud de los valores del atributo Departamento asociados a la tabla de Centros Culturales:

1. El atributo de calidad afectado es la Completitud, ya que en esta columna se encuentra un valor vacío (null/nan) y esto hace que la tabla esté incompleta y que sea más difícil manipularla.
2. Este problema corresponde a instancia, ya que hubo un error en la creación o manipulación de un objeto/valor en el atributo Departamento. A pesar de que el modelo está bien definido, este objeto en particular tiene un null debido a un error en el código que lo crea o modifica.
3. Medida concreta de la magnitud del problema usando GQM:  
**G:** Goal (Objetivo): Que el dato correspondiente al Departamento donde se encuentra el Centro Cultural está completo



**Q:** Question (Pregunta): ¿Cuál es la proporción de Centros Culturales que tienen el dato correspondiente a Departamento vacío (null/nan)?

**M:** Metric (Métrica): M1: Proporción de registros con campo Departamento vacío en la tabla de Centros Culturales

$$\left( \frac{\text{Cantidad de registros de Centros Culturales con campo Departamento vacío}}{\text{Cantidad total de registros de Centros Culturales}} \right) = 0,0009$$

#### Otros problemas encontrados con la tabla de Centros Culturales son

- Los valores del atributo *Categoría* son todos iguales (*Centro Cultural*), lo que es redundante porque es una tabla que solo tiene centros culturales.
- Algunos valores del atributo *Capacidad* aparecen en nulls o 0. Asumimos que es un error en la carga de datos.
- El atributo Departamento que contiene los valores correspondientes a la Provincia o Localidad *Ciudad Autónoma de Buenos Aires* está completo con el mismo nombre. Esto no aporta información sobre el departamento, y es inconsistente con la tabla de Centros Educativos, donde el atributo *Departamento* tiene el nombre de la Comuna asociada.

#### Problemáticas en el dataset de Centros Educativos

Queremos analizar la estandarización de los valores del atributo *Teléfono* asociados a la tabla de Centros Educativos

1. El atributo de calidad afectado es Consistencia ya que en esta columna se encuentran valores que no son atómicos y esto hace que sea más difícil de manipular la tabla.
2. El problema de tener valores que no son atómicos corresponde a un problema que puede ser tanto de modelo como de instancia.
  - Problema de Modelo: Si la columna permite almacenar más de un mail, es porque el modelo no define correctamente las reglas de validación, entonces es un problema de modelo. Esto ocurre cuando el modelo no restringe el formato de los datos.
  - Problema de Instancia: Si la columna está bien definida en el modelo (por ejemplo, sólo permite un único mail por establecimiento), pero algunos



objetos tienen valores incorrectos, entonces es un problema de instancia. Esto ocurre cuando los datos se ingresan o modifican incorrectamente, ya sea por un error de código o por una entrada manual incorrecta.

3. Medida concreta de la magnitud del problema usando GQM:

**G:** Goal (Objetivo): Ver si el dato correspondiente al Teléfono del Establecimiento educativo está bien cargado.

**Q:** Question (Pregunta): ¿Cuál es la proporción de Establecimientos Educativos que tienen el dato correspondiente al Teléfono bien cargado (solo con valores atómicos)?

**M:** Metric (Métrica): M1: Proporción de registros con valores atómicos en el campo Teléfono en la tabla de Establecimientos Educativos.

$$\left( \frac{\text{Cantidad de registros con valores atómicos en el campo Teléfono en la tabla de Establecimientos Educativos.}}{\text{Cantidad total de registros de Establecimientos Educativos}} \right) = 0,4$$

Otro problema encontrado con la tabla de Establecimientos Educativos:

- Los valores faltantes en el atributo teléfono no están estandarizados, en ocasiones tienen letras, están rellenos con ceros o no tienen valor (null/nan).

Problemáticas en el datasets de Censo

La problemática que encontramos en este dataset es que en las tablas, el porcentaje de completitud de la población no es consistente. Asumimos que esto se debe a un redondeo de solamente dos decimales. A partir de un determinado rango etario, el acumulado comienza a cambiar pero el porcentaje de cada edad es 0,00%.

Queremos analizar la consistencia (o correctitud) de los valores de la columna “%” asociados a la tabla de Censo.

1. El atributo de calidad afectado sería la Consistencia ya que al final de esta columna, donde se encuentra el total, aparece un valor diferente al que debería aparecer. Si haces la cuenta algunas columnas dan 99% y en el valor total aparece 100%. Esto ocurre porque los valores porcentuales están acotados solo a dos decimales.
2. Este problema corresponde principalmente al modelo, aunque también puede tener implicaciones en las instancias.



- Problema de Modelo: El problema radica en cómo se define la estructura y las reglas del modelo. Si la columna de “%” está diseñada para almacenar sólo dos decimales, esta decisión es parte del modelo. La limitación de los decimales puede causar que, al sumar los valores, no se alcance exactamente el 100% debido a redondeos.
- Problema de Instancia: El problema también puede manifestarse en las instancias si los valores específicos de los objetos no se ajustan correctamente. En algunos casos hay filas con distinta cantidad de personas pero el porcentaje es el mismo.

3. Medida concreta de la magnitud del problema usando GQM:

**G:** Goal (Objetivo): Ver si el dato correspondiente al total en % en la tabla Censo, es correcto respecto a la suma de toda la columna

**Q:** Question (Pregunta): ¿Cuántos valores del porcentaje no son consistentes con respecto al total en “%”?

**M:** Metric (Métrica): M1: Proporción de registros con campo porcentaje no correlativo al total en “%” en la tabla de Censo

$$\left( \frac{\text{Cantidad de registros de porcentaje no correlativo al total en "\%"}}{\text{Cantidad total de registros del Censo}} \right) = 0,95$$

Otro problema encontrado en la tabla de Censo:

- Los datasets de Censo están empaquetados en una sola tabla, lo que dificulta su manipulación y análisis.

Limpieza de datos

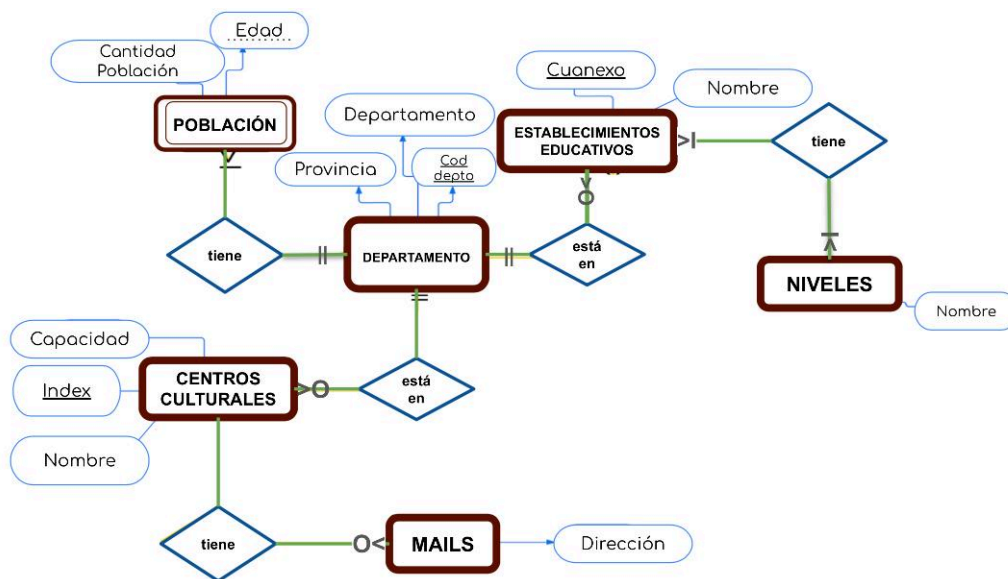
Pasamos a la parte de la limpieza de datos en la que creamos los DataFrames que vamos a utilizar según nuestro Modelo Relacional, las tablas con las que trabajamos nos fueron dadas en la consigna. El primer dataframe que creamos es el de Población, que está conformado por Edades y Cantidad de población por edad de cada departamento.

Para armar el dataframe de Establecimientos Educativos tomamos los cueanexos y el atributo Provincia (que eliminaremos más adelante). Para el dataframe de Centros Culturales nos quedamos con el atributo Capacidad, Provincia (que también eliminaremos más adelante) y le agregamos un Index que será la clave. Seguimos con la tabla de Mail, su atributo es Direcciones (las direcciones de mail de cada Centro Cultural). Luego creamos Departamento con los atributos Código de Depto, Departamento y provincia. Ya creada esta



tabla podemos eliminar provincia de centros culturales y establecimientos educativos. Por último creamos las tablas Niveles y Tiene, dónde niveles tiene el atributo nombre y Tiene es la relación entre Establecimientos Educativos y Niveles.

### Diagrama Entidad-Relación



### Aclaraciones del DER

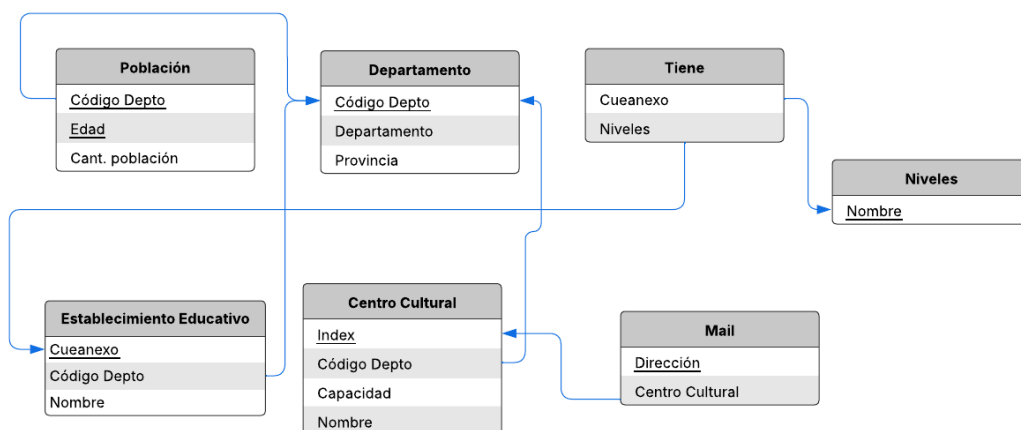
- Población es una entidad débil que se relaciona con Departamento, sus atributos son cantidad de población y su clave (está subrayada con línea punteada) que es Edad.
- Departamento se relaciona con Población, Establecimientos Educativos y Centros Culturales ya que cada uno de estos tiene departamentos asociados. Los atributos de Departamento son: Departamento, Código de Departamento y Provincia.
- Establecimientos Educativos se relaciona con Departamento y con Niveles. Se relaciona con niveles ya que esta entidad guarda los niveles educativos que puede tener cada establecimiento educativo. El único atributo que tiene EE es cuanexo que es un identificador único para cada establecimiento.
- Centros Culturales se relaciona con Departamento y con Mail. Mail es un atributo que guarda el/los mails de cada centro cultural.



## **Modelo Relacional**

En el siguiente diagrama vemos las entidades y sus respectivos atributos. Las claves primarias (PK) están subrayadas y las foreign keys (FK) están indicadas con flechas azules. Las dependencias funcionales (DF) son las siguientes:

- Población (Código Depto, Edad) → Cant. Población
- Departamento Código Depto → Departamento, Provincia
- Establecimiento Educativo Cueanexo → Código Depto
- Centro Cultural Index → Código Depto, Capacidad
- Mail Dirección → Centro Cultural



Con esta organización de los datos, vemos que los atributos son atómicos. También, dependen completamente de una PK. Al estar en 2FN y no tener dependencias transitivas, nos aseguramos de que todo está en tercera forma normal .

## **Decisiones tomadas**

Dado que en la base de datos dada de los centros culturales, no se especifican los código de departamento para las distintas comunas como en las bases de datos del padrón del censo y de los establecimientos educativos, decidimos modificar todos los códigos de departamentos, en las tablas creadas, Establecimientos Educativos y Población, reemplazando, los códigos correspondientes a la ciudad autónoma, por el valor 2000; el cuál luego asociamos al departamento Ciudad Autónoma de Buenos Aires en la tabla creada, Departamento. Para los gráficos consideramos a CABA como una provincia.





Como en Establecimientos Educativos y en Centros Culturales existían inconsistencias al nombrar a CABA y a Tierra del fuego, una lo hacía como “Ciudad Autónoma de Buenos Aires” y “Tierra del Fuego, Antártida e Islas del Atlántico Sur” mientras que la otra como “Ciudad de Buenos Aires y Tierra del Fuego”, decidimos unificar las apariciones llamando a todas como “Ciudad Autónoma De Buenos Aires y Tierra del Fuego”.

También, dado que Establecimientos Educativos no contenía los departamentos 94008 y 94015 asociados a la provincia de Tierra del Fuego para que estén correctamente en Departamento modificamos su aparición, asignando a la columna de provincia el valor de Tierra del Fuego.

### Análisis de Datos

1)

Provincia	Departamento	Jardines	Población Jardín	Primarias	Población Primaria	Secundarias	Población Secundaria
Buenos Aires	La Matanza	325	157034	333	258149	335	148935
Buenos Aires	La Plata	232	52075	199	88950	206	56374
Buenos Aires	Lomas de Zamora	162	50825	178	88046	191	54178
Buenos Aires	General Pueyrredón	178	41427	169	71739	170	48556
Buenos Aires	Quilmes	162	47353	146	81186	153	49780
Buenos Aires	Almirante Brown	133	43762	137	77813	145	48327
Buenos Aires	Moreno	118	50791	134	87151	134	49565
Buenos Aires	Merlo	110	48007	120	81849	124	49410
Buenos Aires	Lanús	120	28133	116	51082	114	33279

En la tabla podemos ver a qué provincia pertenece cada departamento y sus respectivas poblaciones. Segmentamos a la población y a los establecimientos educativos según los niveles educativos. Podemos analizar el ratio de EE por cantidad de población, además de comparar las provincias con respecto a la cantidad de EE que tienen.

2)

Provincia	Departamento	Cantidad de CC con cap > 100
Buenos Aires	Avellaneda	20
Buenos Aires	La Plata	8
Buenos Aires	Lomas de Zamora	3
Buenos Aires	General Pueyrredón	2
Buenos Aires	Almirante Brown	2
Buenos Aires	Mercedes	1
Buenos Aires	Tres de Febrero	1



Laboratorio de Datos  
Trabajo Práctico 01  
Verano - 2025



El reporte muestra a qué departamento corresponde cada provincia, y la cantidad de centros culturales con capacidad de más de 100 personas. Podemos entender la utilidad de cada CC complementando esta información con el primer reporte, donde indicamos la población según cada nivel educativo.

3)

Provincia	Departamento	Cant_CC	Cant_EE	Cantidad_Poblacion
Ciudad Autónoma de Buenos Aires	Ciudad Autónoma de Buenos Aires	0.0956241	0.896799	3095.45
Córdoba	Capital	0.0200259	0.934542	1498.06
Santa Fe	Rosario	0.0269067	0.932765	1337.96
Buenos Aires	La Matanza	0.00108863	0.645559	1837.17
Buenos Aires	La Plata	0.0952288	1.10042	756.074
Chaco	San Fernando	0.0314189	1.95764	413.764
Santa Fe	La Capital	0.0316757	1.28111	568.259
Buenos Aires	General Pueyrredón	0.0151385	1.00519	660.569
Buenos Aires	Lomas de Zamora	0.0247942	0.943638	685.644

En el siguiente reporte analizamos la cantidad de CC y de EE que tiene cada provincia, con su respectivo departamento. Además, podemos visualizar la cantidad de población que cada espacio tiene. Si tomamos en cuenta el segundo reporte, podemos entender la capacidad que tiene cada CC con respecto al total de población.

4)

Provincia	Departamento	Direccion
Buenos Aires	General Pueyrredón	gmail
Buenos Aires	Saavedra	gmail
Buenos Aires	San Fernando	yahoo
Buenos Aires	General Juan Madariaga	hotmail

En esta tabla analizamos para cada departamento y su respectiva provincia, qué tipo de dirección de mail es la más utilizada.



## Visualización

1.

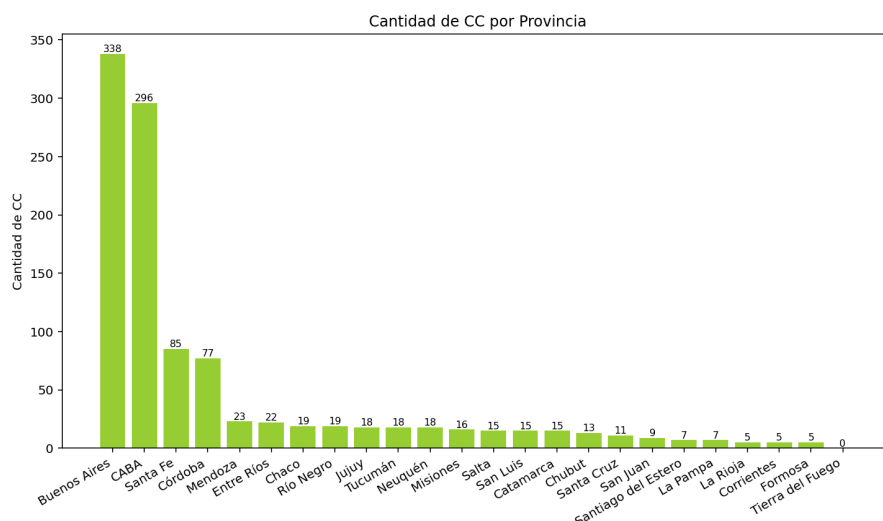


Figura I. Gráfico de la cantidad de CC en función de las provincias argentinas

En este gráfico podemos observar que Buenos Aires es la provincia con más centros culturales de Argentina, seguida de CABA. Santa Fe y Córdoba también tienen un número importante, comparable con respecto a las demás provincias. Podemos ver que La Rioja, Corrientes y Formosa son las que menos tienen, sin contar Tierra del fuego en la que no hay ni un solo Centro Cultural.

2.

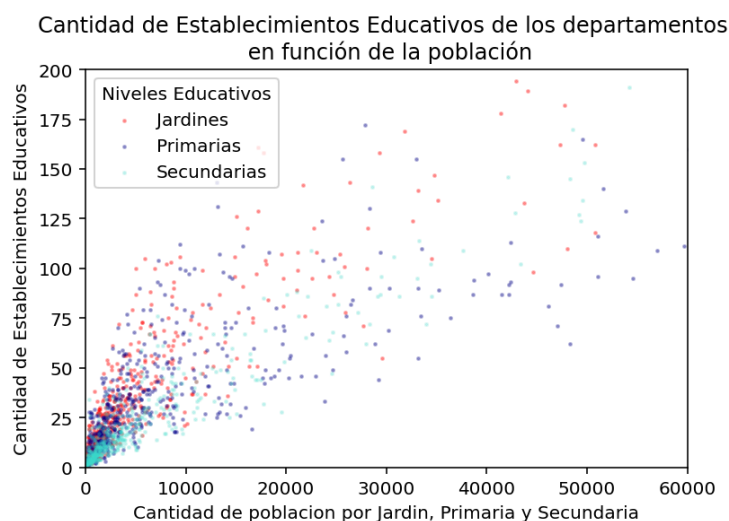


Figura II. Gráfico de la cant. de EE por departamento en función de la población

En este gráfico podemos observar cómo la mayoría de departamentos tienen una pequeña población en instancia educativa, y por consiguiente, pocos establecimientos educativos. Con una tendencia creciente, a medida que aumenta la población, aumenta significativamente la cantidad de establecimientos educativos. Son solo dos departamentos los que tienen más de 250.000 habitantes y más de 800 centros culturales.

3.

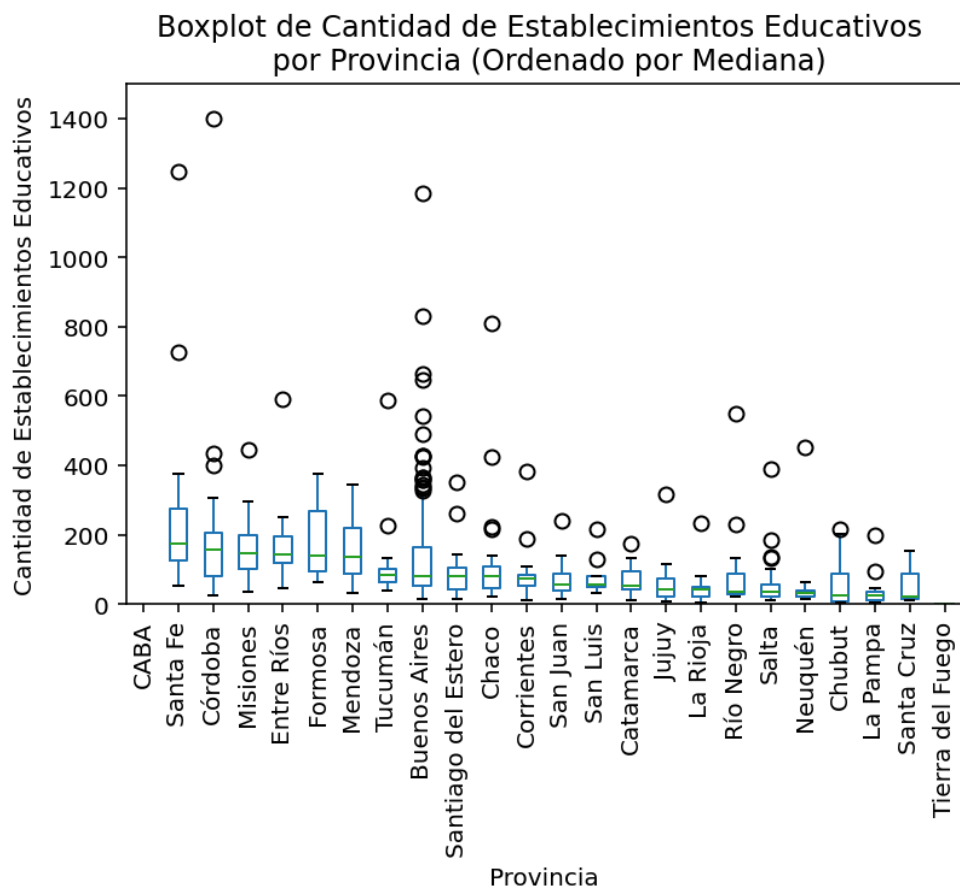


Figura III. Boxplot de la cant. de EE por cada Departamento de las Provincias.

Decidimos omitir CABA ya que es el departamento con más cantidad de EE, para conseguir una mejor visualización del gráfico. Como el gráfico está ordenado por mediana, sabemos que CABA, Santa Fé y Córdoba son las provincias con mayor concentración de EE por departamento. Las provincias que tienen rangos intercuartiles más amplios indican mayor dispersión en la cantidad de EE por departamento (Formosa, Santa Fe y Mendoza). Por otro



lado, provincias como Tierra del Fuego, La Pampa y Catamarca tienen valores mucho más bajos y menos dispersos, lo que indica que la cantidad de EE es más homogénea en sus departamentos.

4.

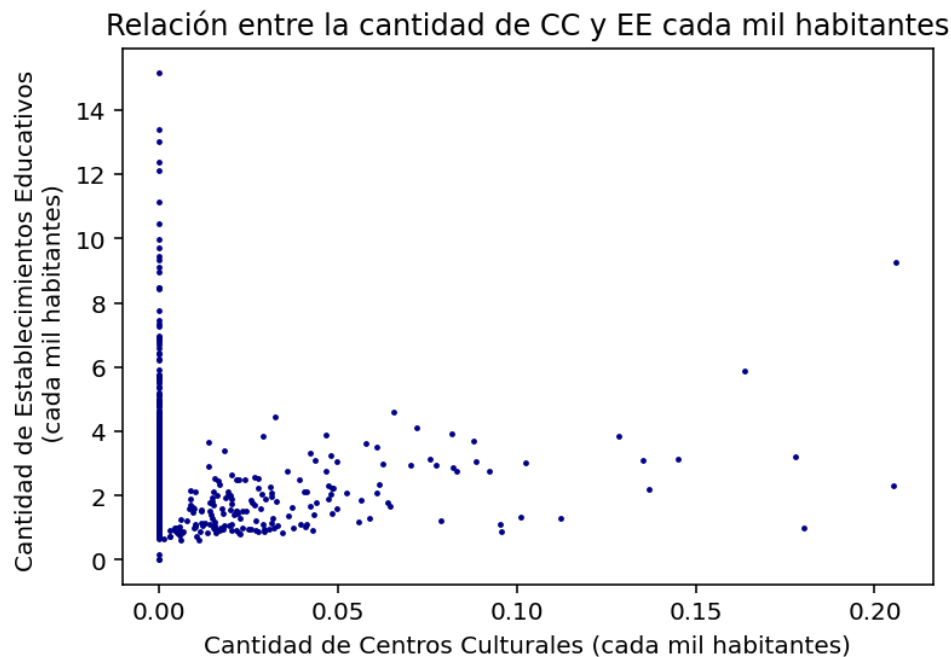


Figura IV. Gráfico de la relación entre la cant. de CC y EE cada mil habitantes

Si bien en la siguiente figura no vemos una relación marcada porque hay mucha dispersión, es evidente que a menor cantidad de CC, existe una menor cantidad de EE. Se puede ver una relación lineal de tendencia creciente con respecto a la población, sin embargo es muy evidente que hay muchos casos de departamentos que no tienen centros culturales pero sí establecimientos educativos.

El resto de outliers no tiene un patrón claro, solamente se encuentran dispersos. Esto es correlativo a lo que visualizamos en la figura I y III, donde provincias como Formosa tienen gran cant. de EE, pero muy poca cant. de CC. En las provincias donde hay EE pero no hay CC, podemos ver que la cantidad de EE es muy variable, o bien puede tener muchos EE o pocos, pero aun así no tienen CC.



### **Conclusión:**


A partir del análisis realizado en este informe, podemos concluir que no existe relación entre Centros Culturales y Establecimientos Educativos. Se puede notar una ligera tendencia creciente con respecto a la población, pero por la gran dispersión de casos y la gran cantidad de departamentos que no tienen centros culturales pero sí establecimientos educativos, concluimos que no se puede determinar una relación entre las dos variables.

Llegamos a la misma conclusión observando los gráficos de las figuras I y III. Podemos ver que la cantidad de centros culturales y de establecimientos educativos no son correlativas entre sí. Hay provincias como Formosa o Corrientes que si bien tienen una cantidad significativa de establecimientos educativos, no tienen muchos centros culturales. Entendemos de este análisis que una provincia tenga muchos establecimientos educativos no implica que también tenga muchos centros culturales, ni vice versa. Consideramos que esto se debe a otros factores que escapan del informe y de los cuales no tenemos acceso. Por ejemplo, las políticas particulares de cada provincia y hacia dónde deciden destinar su dinero. Nos resultó interesante también ver la gran diferencia entre CABA o Buenos Aires y las demás provincias de la Argentina. Mucho más con respecto a Tierra del Fuego, donde ni siquiera hay presencia de centros culturales y hay muy pocos establecimientos educativos.

Sin dudas este informe nos deja ponderando sobre el irregular acceso a la cultura y a la educación que existe en un mismo país. Por esto es muy importante tener en cuenta las realidades propias de cada espacio a analizar. Si bien nosotros conocemos que Argentina tiene una gran variabilidad demográfica y política entre provincias, si este análisis se quiere replicar en otros países deberíamos tener en cuenta sus propias cuestiones políticas, históricas y económicas. De lo contrario, podemos caer en un análisis sesgado o en realizar relaciones erróneas.




### **Anexo:**

 tablaConsigna1.xlsx

[https://docs.google.com/spreadsheets/d/1obpvzQzOUFBqbTLujCKcm1LvY07t9FND/edit?usp=drive\\_link&oid=109948414481168507707&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1obpvzQzOUFBqbTLujCKcm1LvY07t9FND/edit?usp=drive_link&oid=109948414481168507707&rtpof=true&sd=true)

 tablaConsigna2.xlsx

[https://docs.google.com/spreadsheets/d/1fGnPrGaJVz9gf4zWhlAyg1ps6-QiYNem/edit?usp=drive\\_link&oid=109948414481168507707&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1fGnPrGaJVz9gf4zWhlAyg1ps6-QiYNem/edit?usp=drive_link&oid=109948414481168507707&rtpof=true&sd=true)

 tablaConsigna3.xlsx


[https://docs.google.com/spreadsheets/d/1MEOHQluC-TtLLBLdLlwDsM2yruzUevoU/edit?usp=drive\\_link&oid=109948414481168507707&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1MEOHQluC-TtLLBLdLlwDsM2yruzUevoU/edit?usp=drive_link&oid=109948414481168507707&rtpof=true&sd=true)

 tablaConsigna4.xlsx

[https://docs.google.com/spreadsheets/d/1ul\\_FzbzvZQZyYwb\\_IYgv\\_ASA9n8-uL1G/edit?usp=drive\\_link&oid=109948414481168507707&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1ul_FzbzvZQZyYwb_IYgv_ASA9n8-uL1G/edit?usp=drive_link&oid=109948414481168507707&rtpof=true&sd=true)

### **Bibliografía**

1. Establecimientos Educativos (EE). Padrón Oficial de Establecimientos Educativos 2022. Disponible en:  
<https://www.argentina.gob.ar/educacion/evaluacion-e-informacion-educativa/padron-oficial-de-establecimientos-educativos>
2. Centros Culturales (CC). Padrón de Centros Culturales.  
[https://datos.gob.ar/dataset/cultura-mapa-cultural-espacios-culturales/archivo/cultura\\_0e9a431c-b4f7-455b-aa1a-f419b5740900](https://datos.gob.ar/dataset/cultura-mapa-cultural-espacios-culturales/archivo/cultura_0e9a431c-b4f7-455b-aa1a-f419b5740900)
3. Población. Datos de población por Departamento. Se pueden obtener de los datos del censo de 2022, sección Estructura por edad de la población. Está disponible en:  
<https://www.indec.gob.ar/indec/web/Nivel4-Tema-2-41-165>

 padron\_poblacion.xlsX

4. Clases subidas al campus.