Trabajo Práctico 01

Manejo y Visualización de Datos

Fecha de entrega: 23-02-2025 **Laboratorio de Datos**

Pepitas

Nombre	LU
Caporaletti, Gonzalo	707/22
Carbonel, Victoria	978/23
Gonzalez Dardik, Micaela Natali	1143/22

EXACTAS UBA

Contenidos

1 Resumen

Este proyecto tiene como objetivo analizar la distribución de Establecimientos Educativos (EE) y Centros Culturales (CC) en Argentina para identificar patrones y evaluar si existe una relación entre la cantidad de ambos tipos de instituciones. A través del uso de datos abiertos de fuentes oficiales, se realizó un análisis cuantitativo y territorial que permite comprender la disponibilidad de EE y CC en distintas regiones del país y detectar posibles desigualdades en su acceso.

A través de un análisis riguroso de la calidad de los datos y la creación de un modelo relacional, se obtuvieron conclusiones sobre la accesibilidad a la educación y la cultura en Argentina. Los resultados se presentaron mediante visualizaciones gráficas, como gráficos de barras y boxplots, que resaltan las diferencias en la cobertura educativa y cultural a nivel departamental, provincial y nacional.

2 Introducción

Se va a investigar la relación entre la cantidad de establecimientos educativos y centros culturales en las provincias de Argentina, para determinar si existe una correlación significativa entre estas dos variables. La educación y la cultura son pilares fundamentales para el desarrollo social, y entender su distribución puede ayudar a abordar desigualdades existentes.

El análisis comenzará con la recopilación y exploración de conjuntos de datos provenientes de fuentes oficiales, evaluando su contenido y estructura. A continuación, se diseñará un modelo de base de datos mediante un Diagrama Entidad-Relación (DER) y se construirá un esquema relacional que facilite el análisis. Para garantizar la calidad de los datos, se aplicarán técnicas de limpieza y validación que aseguren su confiabilidad.

Mediante consultas SQL, se podrán consultar temas particulares, y se explorarán posibles relaciones entre la distribución de los establecimientos y factores poblacionales y socioeconómicos. El documento finalizará presentando los hallazgos obtenidos, complementados con visualizaciones gráficas que permitirán evaluar tendencias y diferencias en la cobertura educativa y cultural en el país.

3 Procesamiento de Datos

Para realizar este trabajo, se nos proporcionaron las siguientes bases de datos:

- **centros_culturales.csv**: Padrón de Centros Culturales de la República Argentina.
- **2022_padron_oficial_establecimientos_educativos.xlsx**: Padrón Oficial de Establecimientos Educativos 2022 de la República Argentina.
- **padron_poblacional.xlsx**: Datos de población por Departamento de la República Argentina.

Cabe destacar que las tres fuentes provienen de distintos entes del Gobierno Nacional de la República Argentina, lo que provoca que su estructura, códigos y/o estándares no sean necesariamente iguales.

La primera problemática a abordar es la calidad de los datos. Es necesario verificar que las tablas cumplen con los principios de normalización. En caso contrario, será necesario transformar los datos para que cumplan con la **Tercera Forma Normal (3FN)**.

Además, se realizó un proceso de filtrado y mejora de calidad en las tablas proporcionadas. Para lograr esto, elaboramos **diagramas y modelos** que nos permitieron comprender mejor la estructura de los datos y definir una base de datos optimizada para nuestro análisis, eliminando información redundante o innecesaria.

3.1 Formas Normales

Las formas normales establecen reglas para estructurar bases de datos minimizando redundancias y asegurando integridad:

- **Primera Forma Normal (1FN)**: Garantiza datos atómicos y la existencia de una clave primaria.
- **Segunda Forma Normal (2FN)**: Asegura que todos los atributos dependan completamente de la clave primaria.
- **Tercera Forma Normal (3FN)**: Elimina dependencias transitivas para optimizar la estructura.

A continuación, se explica la transformación realizada a las bases de datos para alcanzar la 3FN. Una vez normalizadas las bases, se elaboraró un Diagrama de Entidad-Relación (DER), y su respectivo esquema de Entidad-Relación (RE) para representar la estructura final del sistema.

3.1.1 Primera Forma Normal (1FN)

Evaluación de 1FN en las bases de datos proporcionadas

Ambas bases de datos analizadas violan **1FN**:

- `centros_culturales.csv`: La columna **Mail** almacena múltiples correos en una misma celda.
- `establecimientos_educativos.csv`: Columnas como **modalidad** y **teléfono** contienen valores separados por "/" o "o", infringiendo la atomicidad.

3.1.2 Segunda Forma Normal (2FN)

Evaluación de 2FN en las bases de datos proporcionadas

Ambas bases de datos analizadas violan **2FN**:

- `centros_culturales.csv`: Además de no cumplir con **1FN**, contiene dependencias parciales como **Departamento** respecto a **ID_DEPTO**.
- `establecimientos_educativos.csv`: Además de no cumplir con **1FN**, **Jurisdicción** y **Departamento** dependen del **Código de Localidad**.

3.1.3 Tercera Forma Normal (3FN)

Evaluación de 3FN en las bases de datos proporcionadas

- Ninguna cumple con **1FN** ni **2FN**, por lo que tampoco están en **3FN**.
- Presentan dependencias transitivas en la estructura de sus atributos.

3.2 Análisis de calidad de las base de datos

Para garantizar la calidad y estructura adecuada de la base de datos, evaluamos su cumplimiento con las normas de normalización. Con este propósito, aplicamos el método **GQM** (Goal-Question-Metric) de Basili, que nos permite definir objetivos claros, formular preguntas relevantes y establecer métricas cuantitativas para evaluar su cumplimiento. Este enfoque se desarrolla en tres niveles: definición de objetivos, formulación de preguntas y establecimiento de métricas, seguido por tres fases de aplicación: definición, recolección de datos y análisis. De esta manera, aseguramos la alineación entre las necesidades organizacionales y las mejoras en la calidad de los datos.

3.2.1 Fuente de datos: `centros_culturales.csv`

Problema: Inconsistencias en Departamento, Localidad y Piso

Las columnas más propensas a contener inconsistencias son Provincia, Departamento, Localidad y Piso. Para analizar si existen inconsistencias, se empleó el siguiente análisis:

1. **Atributo de la calidad comprometido**: **Consistencia**.
2. **Tipo de problema**: **Instancia**, aunque puede sugerir un problema de **Modelo** si no se estableció un sistema de referencia o un diccionario de localidades/departamentos.
3. **GQM**
   - **Goal**: Evaluar el grado de estandarización en los valores de Provincia, Departamento, Localidad y Piso.
   - **Questions**:
     - ¿Qué porcentaje de registros presenta diferencias ortográficas o formatos distintos para la misma provincia/departamento/localidad?
     - ¿Qué valores adopta la variable Piso?
     - ¿Qué porcentaje de registros se encuentra vacío en la columna Piso?
   - **Metrics**:
     - M1: $100 \times \left(\frac{Cantidad de ID\_\_DEPTO con mltiples nombres de Departamento asociados}{Cantidad total de ID\_\_DEPTO}\right)$.
     - M2: $100 \times \left(\frac{Cantidad de ID\_\_PROV con mltiples nombres de Provincia asociados}{Cantidad total de ID\_\_PROV}\right)$.
     - M3: $100 \times \left(\frac{Cantidad de Cod\_\_Loc con mltiples nombres de Localidad asociados}{Cantidad total de Cod\_\_Loc}\right)$.
     - M4: Valores que adopta el atributo Piso.
     - M5: $100 \times \left(\frac{Cantidad de registros faltantes en Piso}{Cantidad total de registros}\right)$.
   - **Métricas:**
     - **M1**: $100 \times \left(\frac{2}{1067}\right) \approx 0.19\%$
     - **M2**: 0%
     - **M3**: $100 \times \left(\frac{19}{1067}\right) \approx 1.8\%$
     - **M4**: { P.B., 2°, s/d, 1°, Timbre 4, P.A., Null, PB, PA, 4° }
     - **M5**: La proporción de **Piso** faltantes es de aproximadamente 72%
   - **Criterio de corrección:**
     - Establecer un único valor para cada una de las claves. Este valor será el correspondiente a la **moda** de los valores asociados a la clave (el valor que más se repite).
     - **Eliminar la columna "Piso"**. Dado que la mayor parte de los registros de **Piso** no están disponibles, no es trivial obtenerlos y no juega un papel importante en el análisis correspondiente a este trabajo práctico. Por lo tanto, lo más conveniente es eliminar este atributo.
   - **Impacto en la calidad:**
     - En el caso de las primeras tres métricas, el impacto en la calidad es muy bajo, ya que el porcentaje de registros afectados es bajo.
     - Para la cuarta y quinta métrica, el impacto en la calidad es considerable, ya que casi la totalidad de los registros se ve afectada.
     - Luego de aplicar los criterios de corrección, el impacto en la calidad pasa a ser nulo, ya que es posible corregir todos los errores sin mayor inconveniente.

3.2.2 Fuente de datos: `2022_padron_oficial_establecimientos_educativos.xlsx`: Padrón Oficial de Establecimientos Educativos 2022 de la República Argentina.

Problema: Celdas vacías en la sección de niveles educativos de establecimientos de modalidad "Común"

1. **Atributo de la calidad comprometido**: **Completitud**.
2. **Tipo de problema**: **Instancia**, dado que se observan celdas vacías en los registros correspondientes a los niveles educativos.
3. **GQM**
   - **Goal**: Analizar los establecimientos educativos de modalidad "Común" y garantizar la calidad de los datos en la sección de niveles educativos, identificando y corrigiendo los espacios vacíos para mejorar la base de datos.
   - **Questions**:
     - ¿Cuántas celdas vacías existen en la sección de niveles educativos?
   - **Metrics**:
     - M1: Número de celdas vacías en las columnas correspondientes a los niveles educativos.
   - **Métricas obtenidas:**
     - **M1**: Se identificaron **286,106** celdas vacías en las columnas de niveles educativos.
   - **Criterio de corrección:**
     - Convertir los valores `NULL` a `s/d` en la sección de niveles educativos.
   - **Impacto en la calidad:**
     - Inicialmente, el impacto en la calidad es significativo debido a la alta cantidad de celdas vacías.
     - Luego de aplicar el criterio de corrección, el impacto en la calidad se reduce a **nulo**, ya que se corrigen los errores sin inconvenientes.

3.2.3 Fuente de datos: `padron_poblacion.csv`

Problema: Formato y Estructura

El archivo original presentaba un desorden en la estructura de los datos. Se registraron nombres de áreas en la columna "Edad" y comunas en la columna "Población", generando inconsistencias que afectaron la calidad y la claridad del dataset. Además, existían filas innecesarias, como totales y encabezados.

1. **Atributo de la calidad comprometido**: Consistencia.
2. **Tipo de problema**: Modelo.
3. **GQM**
   - **Goal**: Evaluar la calidad y consistencia de la estructura de datos, asegurando que cada fila contenga información válida y coherente en las columnas correspondientes.
   - **Questions**:
     - ¿Cuántas filas contenían valores de área y comuna en la columna "Edad" antes del proceso de limpieza? ¿Cuántas de estas filas fueron corregidas?
   - **Métrica:**

– **Número de filas incorrectas:** Se determinará la cantidad de filas que contenían áreas registradas en la columna "Edad" antes de la corrección mediante el siguiente código:

```
filas_incorrectas =
df[df["Edad"].astype(str).str.contains("AREA #")].shape[0]
```

– **Número de filas corregidas:** Se calculará el número de filas donde la columna "Edad" ahora solo contiene valores numéricos, lo que indicará la efectividad de la limpieza:

```
filas_corregidas = df["Edad"].notna().sum()
```

Corrección aplicada:

– Se añadieron nuevas columnas para volcar la información que aparecía como título en la columna de "Edad"; se recorrió el dataset para asignar correctamente esta información que no estaba en la columna correcta.

– Se estableció un criterio para que la columna "Edad" contuviera únicamente valores numéricos. Con esto se eliminaron filas incorrectas como aquellas que en la columna Edad decian 'Totales', y era informacion irrelevante.

3.3 Modelado del Diagrama de Entidad-Relación

3.4 Diagrama de Entidad-Relación

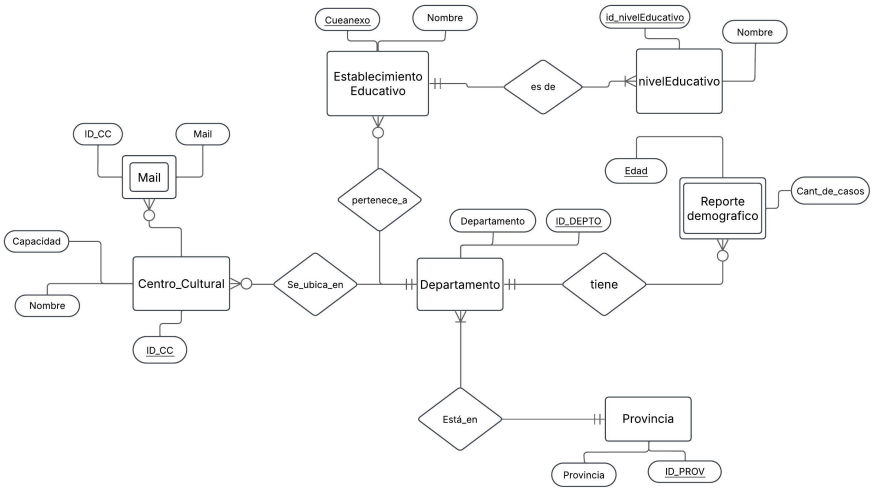

Figure 1: Diagrama Entidad-Relación

En el proceso de diseño del Diagrama Entidad-Relación (DER) (Ver figura1), primero identificamos las entidades, que son: `Centros_Culturales`, `Departamentos`, `Establecimientos_Edu` `Mails`, `Nivel_Educativo`, `Nivel_Educativo_de_ee`, `Reporte_Demográfico` y `Provincias`. Luego, evaluamos los atributos que describen a cada entidad.

- `Centros_Culturales`:
  – `ID_CC`: Identificador único del centro cultural.
  – `Nombre`: Nombre del centro cultural.
  – `Capacidad`: Número máximo de personas que puede albergar el centro cultural.
  – `ID_DEPTO`: Identificador del departamento al que pertenece el centro cultural.

- `Mails`:
  - `ID_CC`: Identificador del centro cultural asociado al correo electrónico.
  - `Mail`: Dirección de correo electrónico del centro cultural.
- `Establecimientos_Educativos`:
  - `Cueanexo`: Código único del establecimiento educativo.
  - `Nombre`: Nombre del establecimiento educativo.
  - `ID_DEPTO`: Identificador del departamento donde se ubica el establecimiento educativo.
- `Nivel_Educativo`:
  - `id_Nivel_Educativo`: Identificador único del nivel educativo.
  - `Nombre`: Nombre del nivel educativo (por ejemplo, primario, secundario, etc.).
- `Nivel_Educativo_de_ee`
  - `Cueanexo`: Código único del establecimiento educativo.
  - `id_Nivel_Educativo`: Identificador del nivel educativo correspondiente al establecimiento.
- `Reporte_Demográfico`:
  - `Edad`: Rango de edad de la población en el reporte.
  - `ID_DEPTO`: Identificador del departamento al que se refiere el reporte demográfico.
  - `Poblacion`: Cantidad de personas en el rango de edad especificado en el reporte.
- `Departamentos`:
  - `ID_DEPTO`: Identificador único del departamento.
  - `Departamento`: Nombre del departamento.
  - `ID_PROV`: Identificador de la provincia a la que pertenece el departamento.
- `Provincias`:
  - `Provincia`: Nombre de la provincia.
  - `ID_PROV`: Identificador único de la provincia.

El próximo paso es distinguir las claves de cada entidad:

- Para la entidad `Centros_Culturales` definimos `ID_CC`: ya que la base de datos original no presentaba un Identificador único para cada centro cultural.
- Para la entidad `Departamentos` usamos `ID_DEPTO`.
- Para la entidad `Establecimientos_Educativos` usamos `Cueanexo`, ya que a cada Establecimiento Educativo lo identifica un único cueanexo.
- `Mails` al ser una entidad débil, no posee una clave única
- Para la entidad `Nivel_Educativo` definimos `id_Nivel_Educativo`, ya que identifica cada uno de los 7 Niveles con un número único a cada Establecimiento Educativo lo identifica un único cueanexo.
- `Nivel_Educativo_de_ee` también es una entidad débil, por lo que no posee una clave única
- `Reporte_Demográfico` también es una entidad debil, no posee clave.
- Para la entidad `Provincias` utilizamos`ID_PROV`

3.5 Modelado Relacional a partir del Diagrama de Entidad-Relación

En base al DER, realizamos el Modelo Relacional y de esta forma identificamos las claves foráneas.

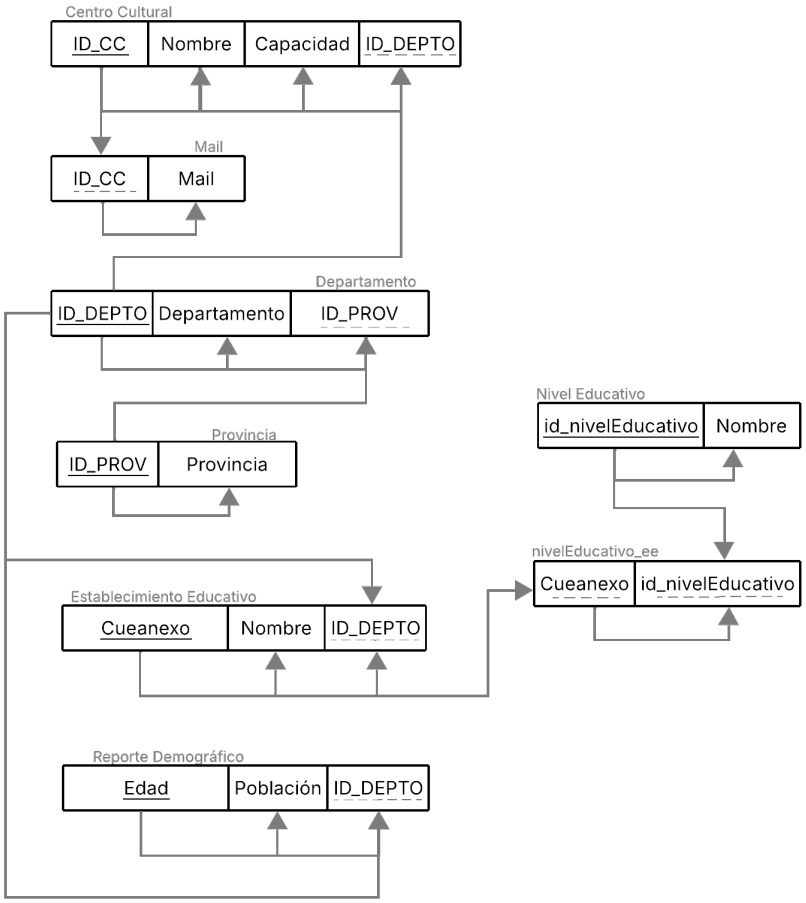

Figure 2: Modelado Relacional

3.6 Importación de Datos

Para elaborar los DataFrames empleados en el proyecto, se recopilaron los datos requeridos a partir de las fuentes. A continuación, se presenta un listado detallado de las fuentes de información utilizadas para cada DataFrame:

- **Centros Culturales:** Archivos centros_culturales.csv y padron_poblacion.xlsx para añadir el ID_DEPTO.
- **Establecimientos Educativos:** Archivos 2022_padron_oficial_establecimientos_educativos.xlsx y padron_poblacion.xlsx para añadir el ID_DEPTO.
- **Nivel Educativo:** Archivo 2022_padron_oficial_establecimientos_educativos.xlsx.
- **Nivel Educativo de EE:** Archivo 2022_padron_oficial_establecimientos_educativos.xlsx.
- **Departamentos:** Archivo padron_poblacion.xlsx y Archivo 2022_padron_oficial_establecimientos_
- **Provincias:** Archivo 2022_padron_oficial_establecimientos_educativos.xlsx.
- **Mails:** Archivo centros_culturales.csv.
- **Reporte Demográfico:** Archivo padron_poblacion.xlsx.

4 Análisis de Datos

En esta sección se van a extraer, para cada ejercicio, distintos DataFrames, a partir de consultas con SQL. Se presentan las primeras y ultimas filas con el objetivo de visualizar una muestra del DataFRame; sin embargo, los DataFrames completos se encuentran en

4.1 Consultas SQL

4.1.1 Establecimientos Educativos y Población por nivel para cada Departamento

Este reporte muestra, para cada departamento, la provincia, la cantidad de Establecimientos Educativos (EE) de modalidad común y la cantidad de habitantes por edad. Se organiza alfabéticamente por provincia y, dentro de cada una, de manera descendente según la cantidad de escuelas primarias. Esto permite visualizar la distribución de recursos educativos en relación con la población.

Table 1: Informe por Departamento: Provincia, Cantidad de Escuelas por Nivel Educativo y Habitantes por Edad

Provincia	**Departamento**	**EE Inicial**	**Pob. Inicial**	**EE Primaria**	**Pob. Primaria**	**EE Secundaria**	**Pob. Secundaria**
Buenos Aires	La Matanza	53.00	87162.00	325.00	225872.00	333.00	181212.00
Buenos Aires	La Plata	59.00	29260.00	215.00	77998.00	199.00	67326.00
Buenos Aires	General Pueyrredon	52.00	23100.00	177.00	62565.00	169.00	57730.00
...	...	...	...	...	...	...	...
Tucumán	Famaillá	1.00	2322.00	24.00	5489.00	23.00	4607.00
Tucumán	La Cocha	0.00	992.00	17.00	2557.00	18.00	2282.00
Tucumán	Juan Bautista Alberdi	0.00	1586.00	0.00	4124.00	0.00	3462.00

El análisis del reporte muestra una desigualdad en la distribución de establecimientos educativos en relación con la población por edad en los diferentes departamentos. Aunque existe un ratio entre la población y los establecimientos educativos, algunos departamentos no cuentan con establecimientos suficientes para la población. Esta variabilidad en la disponibilidad de recursos educativos sugiere la necesidad de un enfoque más equitativo en la asignación de infraestructura educativa.

4.1.2 Cantidad de Centros Culturales con capacidad mayor a 100 personas, por departamento

Este reporte presenta, para cada departamento, la provincia y la cantidad de Centros Culturales (CC) con capacidad superior a 100 personas, ordenado alfabéticamente por provincia y de forma descendente según la cantidad de CC. Resalta la disponibilidad de espacios culturales significativos.

Table 2: Informe por Departamento: Provincia, Cantidad de Centros Culturales cuya capacidad es mayor a 100 personas.

Departamento	**Provincia**	**Cantidad de CC con cap > 100**
Avellaneda	Buenos Aires	20
La Plata	Buenos Aires	8
Lomas De Zamora	Buenos Aires	3
...	...	...
Juan Bautista Alberdi	Tucumán	0
Tafi Viejo	Tucumán	0
Trancas	Tucumán	0

El reporte analiza la cantidad de Centros Culturales (CC) con capacidad superior a 100 personas en cada departamento, organizado por provincia y en orden descendente. Se observa que algunas provincias tienen una mayor disponibilidad de espacios culturales, mientras que en departamentos de Tucumán no hay ningún CC. Esta desigualdad en la oferta cultural sugiere que, aunque existen áreas con una infraestructura cultural significativa, no es la prioridad de inversión en otras localidades. La distribución desigual de Centros Culturales puede afectar el acceso a actividades culturales, lo que podría limitar el enriquecimiento de la vida comunitaria en regiones con menos recursos culturales.

4.1.3 Cantidad de Centros Culturales, Establecimientos Educativos, y Población, por departamento

Se detalla, para cada departamento, la provincia, la cantidad de CC, la cantidad de EE de modalidad común y la población total, ordenado por la cantidad de EE y CC de manera descendente. Esto permite comparar la infraestructura educativa y cultural con la población total.

Table 3: Cantidad de Centros Culturales (CC) y Establecimientos Educativos (EE) por departamento, ordenados por cantidad de EE y CC.

ID_DEPTO	**Departamento**	**Provincia**	**Cantidad CC**	**Cantidad EE**	**Población Total**
112	Ciudad Autónoma De Buenos Aires	Ciudad Autónoma de Buenos Aires	296	1782	3095454.00
82	Capital	Córdoba	30	1136	1498060.00
253	La Matanza	Buenos Aires	2	977	1837168.00
...	...	...	...	...	...
280	Libertador General San Martín	San Luis	0	0	4691.00
479	Tolhuin	Tierra del Fuego, Antártida e Islas del Atlántico Sur	0	0	6027.00
241	Juan Bautista Alberdi	Tucumán	0	0	34745.00

Este análisis integral es clave para identificar disparidades en la disponibilidad de recursos educativos y culturales, proporcionando información valiosa para la toma de decisiones en la asignación de recursos y el desarrollo de políticas públicas más efectivas.

4.1.4 Dominios de Correo Electrónico más usados por los CC para cada departamento

Este reporte indica, para cada departamento, la provincia y los dominios de correo electrónico más utilizados por los Centros Culturales.

Table 4: Dominios de Correo Electrónico más usados por los Centros Culturales dentro de cada departamento.

Provincia	**Departamento**	**Dominio más frecuente**
Buenos Aires	9 De Julio	hotmail
Buenos Aires	Adolfo Alsina	adolfoalsina
Buenos Aires	Alberti	alberti
...	...	...
Tucumán	Cruz Alta	gmail
Tucumán	Lules	gmail
Tucumán	Yerba Buena	gmail

La identificación de los dominios de mail más comunes no solo refleja las prácticas de comunicación de los CC, sino que también puede indicar el nivel de modernidad y adaptación tecnológica de estas instituciones.

4.2 Visualización de Datos

4.2.1 1er Gráfico

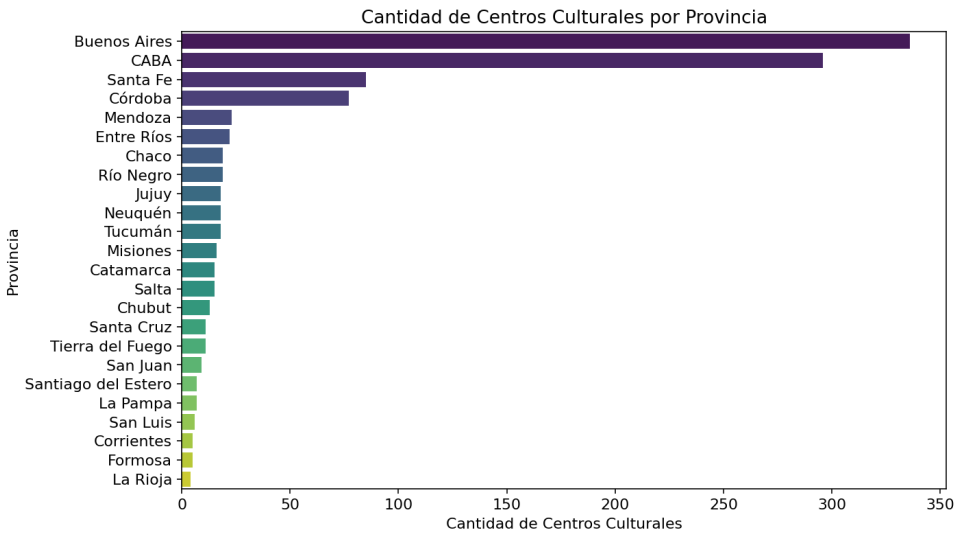

Figure 3: Cantidad de centros culturales por provincia

El análisis de la distribución de centros culturales (CC) en Argentina (ver figura 3) revela una fuerte concentración en Buenos Aires (338) y CABA (296), reflejando tanto la densidad poblacional como la facilidad de acceso en áreas urbanas con alta demanda.

En contraste, provincias como Formosa, Corrientes y La Rioja presentan una notable escasez de CC en relación con su extensión geográfica. Aunque menos pobladas, la mayor distancia entre centros dificulta el acceso a la cultura. Mientras en CABA hay un CC

cada pocas cuadras, en estas provincias los habitantes deben recorrer largas distancias, reduciendo sus oportunidades de participación.

4.2.2 2do Gráfico

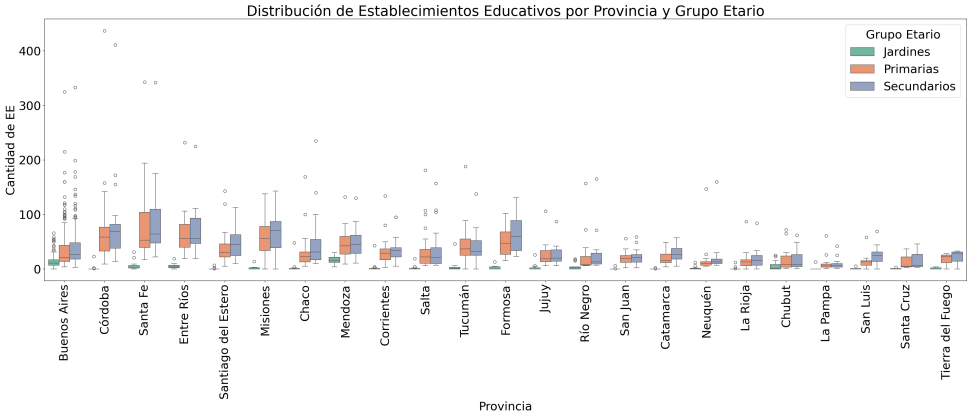

Figure 4: Cantidad de EE de los departamentos en función de la población, separados por nivel educativo y su correspondiente grupo etario

El gráfico presentado (ver figura 4) ofrece un análisis de la distribución de establecimientos educativos en Argentina, clasificados por nivel y departamento, en relación con la población en edad escolar. Esta visualización permite identificar patrones en la infraestructura educativa y su correspondencia con la demanda estudiantil.

El análisis de estos datos es crucial para evaluar la equidad en el acceso a la educación en Argentina. La distribución de establecimientos educativos por provincia y nivel educativo evidencia importantes disparidades regionales, con provincias como Buenos Aires, Córdoba y Santa Fe concentrando la mayor cantidad de instituciones.

Al diferenciar por nivel educativo (jardín, primaria y secundaria), se observa que los jardines de infantes son menos numerosos en comparación con las escuelas primarias y secundarias, lo que podría indicar una menor oferta de educación inicial en ciertas regiones. Además, la presencia de outliers en varias provincias sugiere la existencia de departamentos con una cantidad de establecimientos significativamente mayor que el promedio, lo que podría reflejar la concentración de la población o la presencia de políticas educativas locales específicas.

Cabe destacar que la Ciudad Autónoma de Buenos Aires no está incluida en el gráfico, debido a la dificultad de representar un único departamento en un gráfico de cajas. Además, su inclusión podría distorsionar la visualización, dado que presenta 1782 establecimientos educativos en total, siendo el distrito con la mayor cantidad de instituciones del país. Esta cifra destaca la disparidad entre CABA y el resto de las provincias, lo que refuerza las desigualdes planteadas.

4.2.3 3er Gráfico

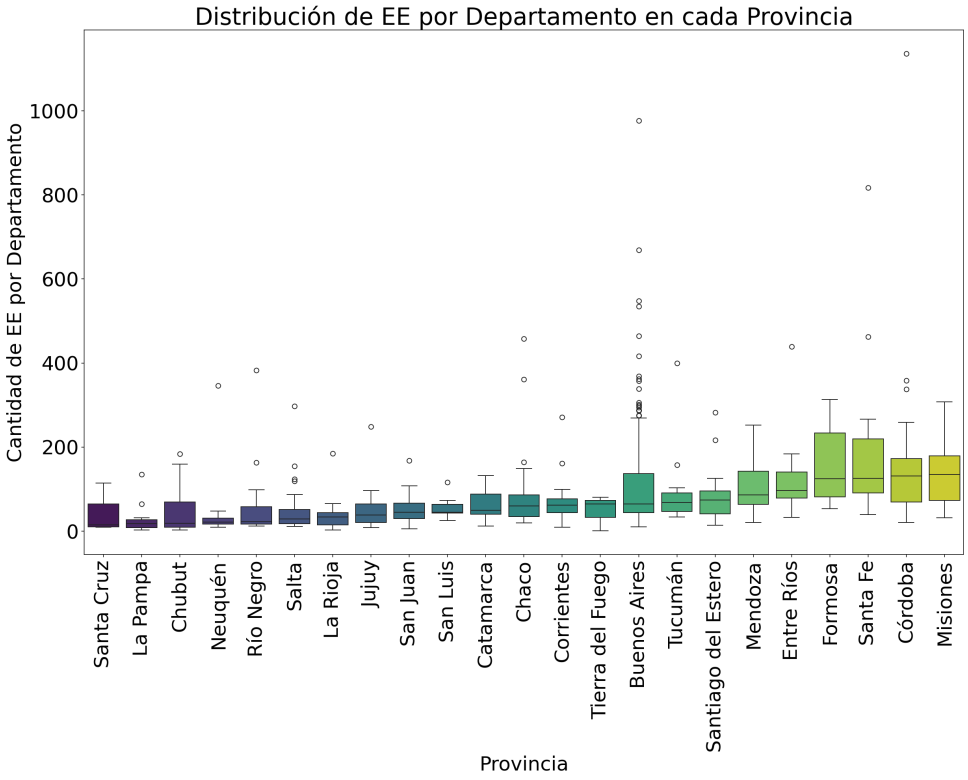

Figure 5: Boxplot por cada provincia, de la cantidad de EE por cada departamento de la provincia. Ordenados por la mediana de cada provincia.

El presente gráfico de cajas (ver figura 5) muestra la distribución de la cantidad de Establecimientos Educativos (EE) por departamento en cada provincia. Esta visualización permite analizar la variabilidad en la infraestructura educativa dentro de cada jurisdicción, destacando las provincias con una distribución más homogénea frente a aquellas con una marcada desigualdad entre sus departamentos.

Los boxplots han sido ordenados según la mediana de la cantidad de EE por provincia, lo que permite visualizar de manera estructurada cómo se distribuyen los recursos educativos en relación con los departamentos. Se observa que provincias como Misiones, Córdoba y Santa Fe presentan medianas más altas, lo que indica una mayor cantidad de establecimientos por departamento en comparación con provincias como Santa Cruz, La Pampa y Chubut, donde la cantidad de EE por departamento es menor.

Asimismo, la presencia de valores atípicos (outliers) en varias provincias sugiere la existencia de departamentos con una cantidad de EE significativamente mayor que el promedio provincial. Esto podría estar relacionado con la concentración de la población en determinadas áreas urbanas o con políticas de infraestructura educativa específicas.

Cabe destacar que una vez más, la Ciudad Autónoma de Buenos Aires no ha sido incluida en este gráfico, debido a que al ser un único distrito sin subdivisión en departamentos, su representación en este formato no sería adecuada. Además, su alta concentración de establecimientos educativos alteraría la escala del gráfico, dificultando la interpretación de los datos de otras provincias.

4.2.4 4toGráfico

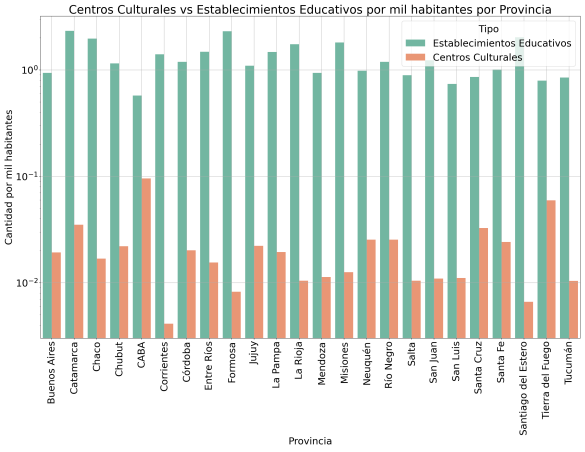

Figure 6: Cantidad de centros culturales y establecimientos educativos cada mil habitantes por provincia

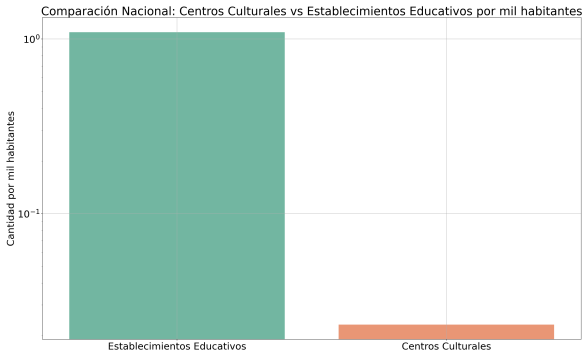

Figure 7: Cantidad de centros culturales y establecimientos educativos cada mil habitantes a nivel nacional

En el primer gráfico (ver figura 6), se observa que la cantidad de EE por mil habitantes es considerablemente mayor que la de CC en todas las provincias, lo que indica que la infraestructura educativa está más consolidada en comparación con los espacios culturales. Sin embargo, se evidencian diferencias significativas entre provincias, lo que sugiere disparidades en el acceso a estos recursos.

El segundo gráfico (ver figura 6) presenta una visión a nivel nacional, resaltando la marcada diferencia entre la cantidad de EE y CC por cada mil habitantes en todo el país. Se confirma que los EE tienen una presencia mucho mayor, mientras que los CC son significativamente menos frecuentes en relación con la población.

Estos resultados pueden interpretarse desde varias perspectivas. Una distribución equitativa de EE y CC podría reflejar una planificación educativa y cultural más equilibrada, favoreciendo el desarrollo integral de la población. No obstante, la baja cantidad de CC en comparación con EE podría indicar una falta de espacios complementarios para la educación no formal y el acceso a actividades culturales. Esta discrepancia pone de manifiesto la necesidad de fortalecer la oferta de centros culturales en diversas regiones para fomentar una educación más integral y accesible.

5 Decisiones Tomadas

5.1 Procesamiento de datos

- **Comunas de la Ciudad Autonoma de Buenos Aires**: Se decidió NO considerar a las comunas como departamentos independientes, sino agruparlas como parte de un mismo departamento/provincia que corresponde a la Ciudad Autonoma de Buenos Aires. A pesar de que las Comunas no son determinadas Departamentos, y la C.A.B.A. tampoco, se consideró dentro de qué categoría deberían clasificarse. Sin embargo en la base de datos de Centros Culturales, los centros culturales de la ciudad bonaerense, no tienen distinción alguna acerca del departamento al que pertenecían; por lo tanto, se definió agrupar las comunas.

5.2 Consultas SQL

- **Consulta 4**: Los departamentos que no tienen al menos un centro cultural **no aparecerán** en este DataFrame. Los centros culturales que **no tienen mail** no se considerarán para la determinación del dominio más utilizado. Si en un departamento **ninguno de sus centros culturales tiene mail**, dicho departamento **no será incluido** en la tabla resultante.

6 Conclusiones

El análisis de la cantidad de Establecimientos Educativos (EE) y Centros Culturales (CC) en cada provincia de Argentina sugiere que no existe una relación directa y uniforme entre ambas variables. A pesar de que ambas instituciones cumplen funciones educativas y culturales, su distribución territorial sigue patrones distintos, dependiendo principalmente de factores como la densidad poblacional y la urbanización.

Como se observa en el gráfico de cantidad de EE y CC por cada mil habitantes en las provincias, la oferta de EE tiende a ser más homogénea, con presencia en todo el territorio nacional, incluso en regiones rurales con baja densidad poblacional. En contraste, los CC muestran una fuerte concentración en áreas urbanas, particularmente en la provincia de Buenos Aires y en la Ciudad Autónoma de Buenos Aires.

Este fenómeno se debe a que los establecimientos educativos están diseñados para garantizar el acceso a la educación básica en todo el país, lo que obliga a su distribución equitativa en función de la población en edad escolar. En cambio, los centros culturales requieren una mayor concentración de público y demanda, lo que los hace más viables en ciudades con una alta densidad poblacional y mayor acceso a actividades culturales.

A nivel provincial, se identifican casos como Santiago del Estero y Formosa, donde la cantidad de EE por mil habitantes es elevada, pero la cantidad de CC sigue siendo baja. Esto indica que la infraestructura educativa en estas provincias está bien distribuida, pero la oferta cultural es limitada. En contraste, CABA y Buenos Aires concentran la mayor cantidad de CC del país, lo que refuerza la idea de que su distribución responde más a la demanda urbana que a un criterio poblacional uniforme.

Por lo tanto, se concluye que no hay una relación proporcional entre la cantidad de EE y CC en cada provincia. La educación formal se encuentra distribuida en función de la población escolar, mientras que la infraestructura cultural se concentra en las zonas con mayor densidad de habitantes y actividad cultural. Este análisis destaca la necesidad de políticas públicas diferenciadas para equilibrar el acceso a la educación y la cultura, asegurando que las provincias con menor cantidad de CC puedan ampliar su oferta sin depender exclusivamente de la densidad poblacional urbana.

A Anexo

Para facilitar la legibilidad del informe y la visualización de los datos, se presentaron exclusivamente las primeras y últimas 3 filas de cada reporte. Los reportes completos se encuentran en formato .csv adjuntos a la carpeta.