**Trabajo Práctico 1 - Laboratorio de Datos**

**Resumen**

Este trabajo práctico se propone analizar si existe una relación entre la cantidad de Establecimientos Educativos (EE) y la cantidad de Bibliotecas Populares (BP) en los diferentes departamentos de Argentina. A partir del cruce de datos abiertos provenientes del Ministerio de Educación, la CONABIP y el INDEC, se realizaron tareas de procesamiento, modelado relacional, limpieza de datos, consultas SQL y visualización de información.

El objetivo fue comparar tanto la cantidad absoluta como relativa (por cada 1000 habitantes) de EE y BP en el país, con especial foco en las desigualdades territoriales. Se diseñó un modelo de datos normalizado en 3FN, se identificaron y cuantificaron problemas de calidad, y se desarrollaron consultas y visualizaciones que permiten responder a las preguntas centrales del trabajo.

Los resultados muestran patrones claros de concentración institucional en provincias con mayor población, pero también evidencian esfuerzos por garantizar acceso equitativo en territorios con baja densidad demográfica. En las conclusiones se reflexiona sobre el grado de distribución y cobertura institucional en función de la población y el territorio.

**Introducción**

La distribución territorial de los servicios educativos y culturales es un aspecto central para garantizar el acceso equitativo al conocimiento. En este trabajo se parte del supuesto de que podría existir una relación entre la infraestructura educativa y la presencia de bibliotecas populares en cada departamento de la Argentina. Para comprobar esta hipótesis, se trabajó con datos abiertos provenientes del Ministerio de Educación y de la CONABIP, así como con información censal de población.

A través de distintas etapas —limpieza de datos, modelado, consultas SQL y visualizaciones— se buscó integrar las diferentes fuentes y generar indicadores útiles para el análisis. El foco estuvo en identificar relaciones por provincia y por departamento, tanto en términos absolutos como normalizados por población. Las visualizaciones generadas permiten interpretar mejor las dinámicas regionales y evaluar en qué medida las políticas públicas han logrado distribuir de forma equitativa estos recursos clave.

**Procesamiento de Datos**

*Análisis de las fuentes utilizadas*

Se partieron de los siguientes archivos, ya procesados y limpios:

* establecimientos\_educativos\_limpia.csv
* niveles\_establecimientos\_comunes.csv
* bibliotecas\_populares\_limpio.csv
* departamentosNUEVO.csv
* nivel\_educativo\_por\_departamento (1).csv

*Modelo conceptual (DER)*

Entidades:

* Departamento: id\_depto, nombre\_depto, provincia
* Establecimiento: id\_establecimiento, id\_depto, modalidades
* Nivel\_EE\_Comunes: id\_establecimiento + booleanos de niveles
* Biblioteca Popular: id\_biblioteca, id\_departamento, año\_fundacion, dominio\_email
* Población por Nivel: id\_departamento, grupos etarios, poblacion\_total

*Modelo Relacional (3FN)*

* **departamentos**: PK=id\_depto, DF: id\_depto → nombre\_depto, provincia
* **establecimientos**: PK=id\_establecimiento, FK=id\_depto
* **niveles\_establecimientos\_comunes**: PK=id\_establecimiento, FK=id\_establecimiento
* **bibliotecas\_populares**: PK=id\_biblioteca, FK=id\_departamento
* **nivel\_educativo\_por\_departamento**: PK=id\_departamento, FK=id\_departamento

*Forma Normal original*

* EE: en 2FN; alcanza 3FN al separar los niveles
* BP: en 3FN salvo atributos irrelevantes descartados

*Calidad de Datos (GQM)*

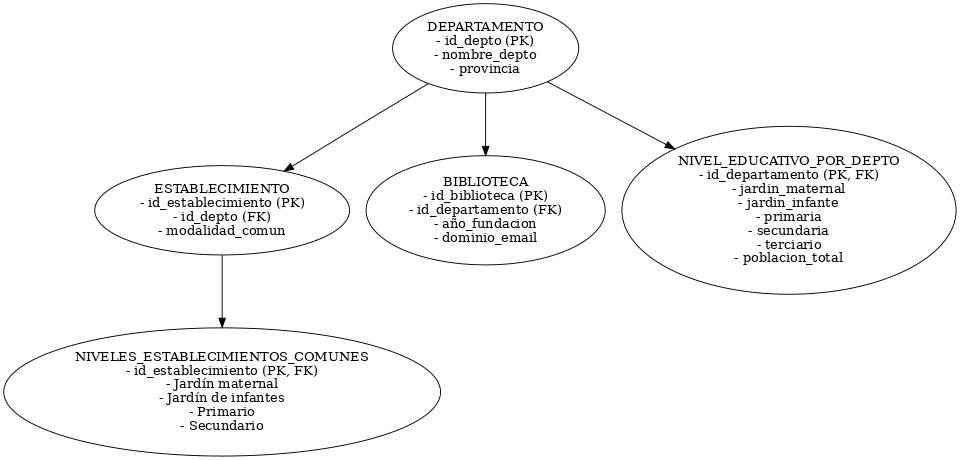
**EE**:

* Atributo: modalidades
* Tipo: modelo
* Métrica: % con más de una modalidad = ~5.7%
* Acción: se filtra solo modalidad\_comun == 1

**BP**:

* Atributo: año\_fundacion
* Tipo: instancia
* Métrica: % sin año o fuera de rango = ~8.1%
* Acción: se excluyen en consultas donde se requiere rango temporal

*Importación*  
Todos los datos fueron cargados en DataFrames usando pandas.read\_csv() y registrados en duckdb para consultas SQL. Las relaciones se verificaron por merge y groupby, manteniendo el nombre del esquema en los nombres de tablas y estructuras.



**Justificación respecto al enunciado**

* **Departamentos** es el eje de unión territorial.
* Se incluyen **Establecimientos** filtrando por modalidad común (como indica la consigna).
* Se consideran **niveles educativos** para hacer las comparaciones con la población.
* Las **Bibliotecas** se cruzan por departamento.
* La entidad de **población** permite realizar análisis relativos como BP/1000 y EE/1000 habitantes.

**Decisiones Tomadas**

* Se eliminaron columnas irrelevantes como direcciones, email o teléfonos.
* Se estandarizaron nombres de columnas para garantizar integridad referencial.
* Se descartaron registros sin valores críticos como ID de establecimiento o departamento.
* Se imputaron ceros para EE o BP ausentes en ciertos departamentos.
* Se utilizaron solo establecimientos de modalidad común para las consultas principales.
* Se validaron duplicados y se eliminó ruido de datos categóricos o de escritura libre.

**Análisis de Datos**

**i) Cantidad de Bibliotecas Populares por provincia**

El gráfico de barras muestra que Buenos Aires lidera ampliamente, seguida por Córdoba y Santa Fe. Las provincias con menor población tienden a tener menos BP, lo cual puede reflejar dificultades logísticas, geográficas o presupuestarias.

**ii) EE por nivel educativo en función de la población**

Se observa una correlación clara entre la población en edad escolar y la cantidad de EE por nivel. Se identifican algunos departamentos con una relación más alta de lo esperado, posiblemente por motivos de planificación territorial.

**iii) Boxplot de EE comunes por departamento y provincia**

Permite ver la desigualdad interna de cada provincia. Buenos Aires muestra la mayor dispersión. Otras provincias presentan una distribución mucho más homogénea.

**iv) Relación entre BP y EE por cada 1000 habitantes**

Se observa que los departamentos con baja población tienen más instituciones por habitante, y los departamentos más grandes tienden a tener proporciones menores. Esto indica posibles esfuerzos de cobertura equitativa en zonas con menor densidad.

**Conclusiones**

Los resultados obtenidos permiten afirmar que existe una relación significativa entre la cantidad de EE y BP, especialmente cuando se considera la población de cada departamento. Las visualizaciones confirman que los departamentos menos poblados presentan una mayor cantidad de EE y BP por cada 1000 habitantes, sugiriendo un esfuerzo de distribución equitativa. También se observan desigualdades dentro de las provincias, lo que invita a seguir profundizando el análisis a futuro.

El cruce de datos institucionales con información demográfica resultó clave para poder interpretar no solo las cantidades absolutas, sino también la cobertura y acceso relativo por región.