DEPARTAMENTO DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Trabajo práctico 01

Verano 2025

Laboratorio de Datos

12 de febrero de 2025

Grupo: "Datitos"

Integrante	LU	Correo electrónico
Badii, Marina	732/24	marinabadii19@gmail.com
Ballera, Alexander	668/24	alexballera@gmail.com
Roko, Tomas	262/23	tomas.e.roko@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón 0+∞)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
http://www.exactas.uba.ar

Resumen

Este informe analiza la relación entre la cantidad de establecimientos educativos y centros culturales en las provincias de Argentina. Para ello, se realiza un procesamiento de los datos disponibles, identificando problemas de calidad y aplicando transformaciones para mejorar su utilidad. Posteriormente, se generaron visualizaciones que facilitan la interpretación de los resultados, llegando a la conclusión de que no hay una relación clara entre la cantidad de establecimientos educativos y centros culturales en las provincias de Argentina.

Introducción

El presente trabajo tiene como objetivo analizar la posible existencia de una relación entre la cantidad de establecimientos educativos y la cantidad de centros culturales en las distintas provincias de Argentina, ya sea en cantidades totales por provincia o en proporciones por habitantes en cada provincia. Para dicho fin, se realizará un procesamiento de los datos disponibles y se diseñará un modelo relacional acorde a lo que fue considerado necesario para cumplir con el objetivo. Finalmente, se presentarán diferentes visualizaciones a partir de las cuales se intentará llegar a una conclusión.

Procesamiento de datos

- **Análisis de formas normales:**

  Las tablas **Establecimientos Educativos** y **Centros Culturales** no se encuentran en **1ra forma normal** debido a que el dominio de los atributos 'Mail', en ambas tablas, incluyen valores múltiples. Más precisamente, en ambas tablas encontramos registros con más de una dirección de correo electrónico. De este modo, queda descartada también la posibilidad de que se encuentren en 2FN, 3FN o BCFN.

  Para que cumplan la primera forma normal podríamos dividirlas. Por un lado, crear una tabla con todos los datos consignados excluyendo al atributo mail y otra con la primary key del establecimiento educativo o el centro cultural y los diferentes mails. De este modo, conservamos la cualidad de "lossless join" y garantizamos que no existan tuplas espúreas, ya que si realizáramos un natural join entre ambas tablas, en caso de que el establecimiento educativo o centro cultural posea dos o más correos electrónicos, habrá filas repetidas excepto por la dirección de correo electrónico.

- **Calidad de los datos:**

  Centros Culturales:
  Para este dataset encontramos algunos problemas de calidad de datos como inconsistencias en los nombres de las columnas o la presencia de nulls en algunos atributos de las instancias. El atributo *Mail* en Centros Culturales presentaba un espacio al final, 'Mail ', lo que generaba problemas al procesarlo. Además, había valores como 's/d' o '-' que no aportaban información y tampoco estaban estandarizados. Para mejorar la calidad de estos datos, se eliminaron los espacios adicionales en los nombres de columnas y se crearon tablas auxiliares para normalizar los correos electrónicos, asegurando que cada instancia tenga solo un mail válido. De este modo, un ejemplo de problema de la calidad de datos en la tabla de Centros Culturales sería el concerniente al atributo ***Mail*** con los atributos de calidad comprometidos ***completitud*** y ***disponibilidad.*** El primer problema mencionado es un problema asociado a **Errores de software*,*** mientras que el segundo problema está asociado al ***Modelo de Datos,*** al no tener los usuarios un único modo de expresar que no poseen mail o no lo quieren consignar.

  Para los tres casos vamos a medir la calidad de los datos y los atributos de interés siguiendo la metodología GQM (Goal, Question, Metric), donde vamos a asignar una medida para la calidad de ciertos atributos relacionados a cada tabla. En este caso:

  Goal: El dato correspondiente a Mail asociado a Centros Culturales esté completo
  Question: ¿Cuál es la proporción de Centros Culturales que tiene el dato correspondiente a Mail vacío/null?

Metric: Proporción de registros con campo Mail vacío o null en la tabla Centros Culturales, es decir,

$$\frac{Cantidad\ de\ Centros\ Culturales\ con\ campo\ Mail\ vacío/null}{Cantidad\ total\ de\ Centros\ Culturales} \cdot 100 = 28.6785\ \%$$

Establecimientos educativos:

Para este dataset nos encontramos con muchas irregularidades. Hay campos del atributo ***'Teléfono'*** vacíos, con textos cómo: 'SE CREA POR RESOL. 1707/2022 MECCyT FECHA:27/04/22' o 's/inf' o 'RED OFICIAL 978'.

El atributo afectado sería ***Teléfono*** y los atributos de calidad comprometidos serían la ***consistencia*** y la ***completitud.*** El problema mencionado es un problema asociado a los ***Procesos.***

Para el análisis de la calidad de los datos relacionados al atributo 'Teléfono' tuvimos las siguientes consideraciones para determinar si un número de teléfono es válido:

- Si es un número de teléfono legal tanto para teléfonos fijos cómo para teléfonos móviles. Con legal nos referimos a que hay un número al que se pueda intentar llamar, o sea que no sea un '000000000', '0', '1', 'sn', 's/n', etc.

- Si es legible. Debido a que el atributo es de tipo string, encontramos muchos campos completados con, por ejemplo, dos teléfonos celulares separados por '/'. Y a pesar de los problemas que esto presenta para el procesamiento de los datos, como es un atributo que no vamos a usar en nuestros datasets ya limpios, decidimos para este análisis considerar dos números de teléfono (legales) separados por '/' cómo un teléfono válido para el propósito de la métrica que usamos a continuación. Otros casos son por ejemplo, '(0381) 4930039', que es un teléfono perfectamente legible y válido según el reglamento.

De este modo, para tener una medida sobre cuántos Establecimientos Educativos no tienen un número de teléfono al que llamar, consideramos la siguiente métrica:

Goal: El dato correspondiente a número de teléfono esté completo con números de teléfono válidos
Question: ¿Cuál es el porcentaje de Establecimientos Educativos con números válidos ?
Metric: Porcentaje de Establecimientos Educativos con números de teléfono inválidos (Cantidad de registros - Cantidad de números válidos),

$$\frac{Cantidad\ de\ Establecimientos\ Educativos\ con\ campo\ Teléfono\ inválido}{Cantidad\ total\ de\ Establecimientos\ educativos} \cdot 100 = 15.7881\%$$

Padrón Población:

Para este dataset encontramos la mayor dificultad a la hora del trabajo con los datos, debido a la distribución de los mismos, que prescindía de un orden a la hora de diferenciar las áreas correspondientes a cada sub-set de mediciones censales. Por cada registro correspondiente a un área, teníamos 5 filas con Nulls y algunos datos dispersos útiles, cómo el número de área y la descripción de la misma.

Otra dificultad que encontramos fue que no todas las áreas habían registrado los datos de las mismas edades, es decir, encontramos registros con personas de hasta 110 años y otros con personas de hasta 108 como máximo. Lo que nos dificultó separar los registros por área.

Para estandarizar los datos y poder trabajar con ellos, definimos dos funciones:

La primera se llama calcular_largo_areas que se devuelve una lista con el largo en filas de todas las áreas registradas, y la segunda se llama extraer_bloques_variable_longitudes que se encarga de extraer cada "bloque" de registros correspondientes a un área, agregar una columna con el número de área y agregar otra columna con la descripción de dicha área. De esta manera es que el dataset ya modificado de padrón nos permite obtener fácilmente los registros para cada área.

. Los datos se encontraban disponibles para procesar? NO
La tabla de datos de padrón poblacional no tiene errores en los datos, de completitud o de exactitud. Sin embargo, los datos no se encontraban disponibles.

Goal:Los datos estén disponibles para procesar
Question: ¿Se pueden utilizar como están?
Metric: En este caso, la métrica es booleana y da negativo.

A continuación en la (Figura 1) presentamos el diagrama entidad relación con los Niveles Educativos separados en valores atómicos.

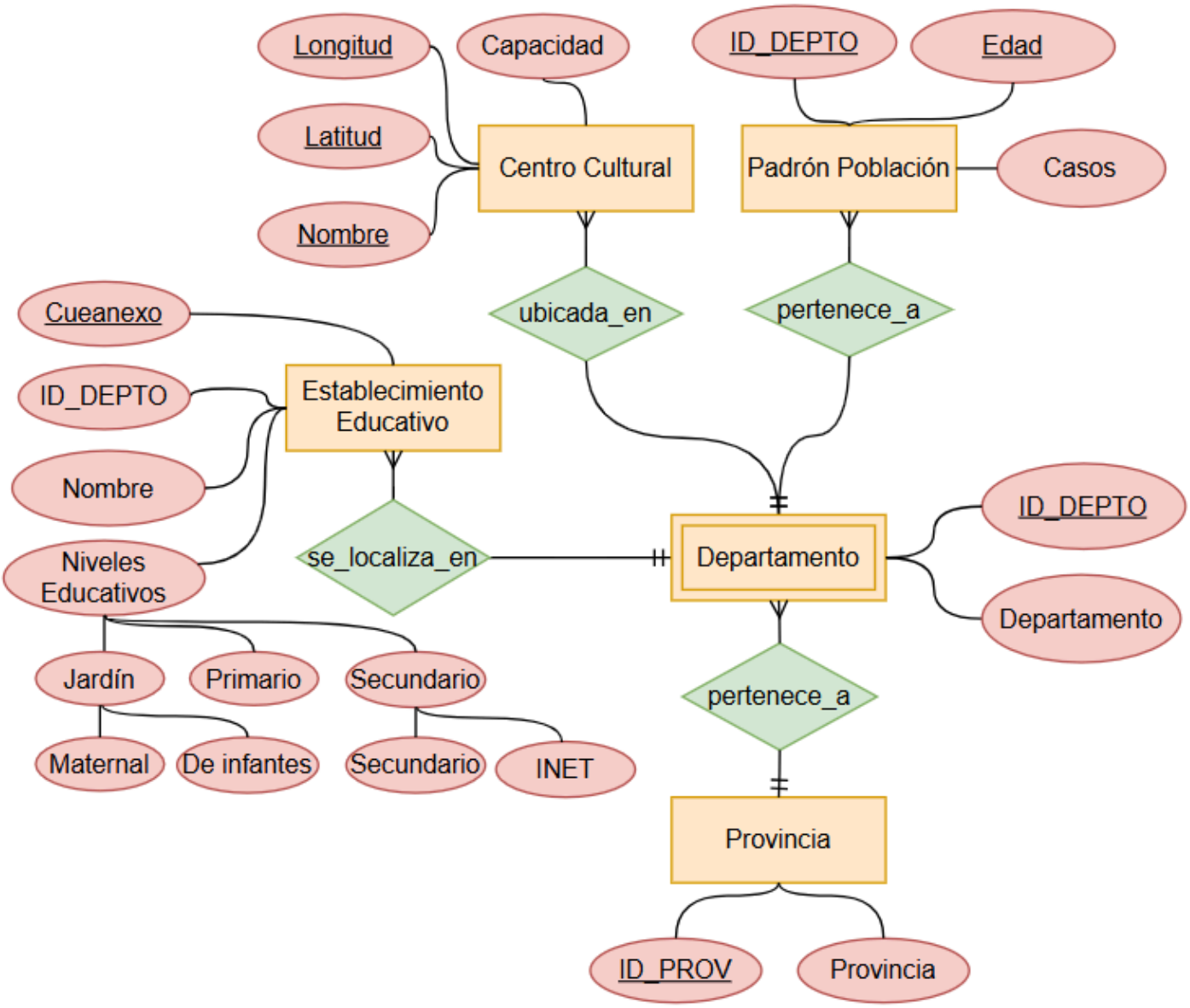

Figura 1: Diagrama Entidad-Relación.

Modelo Relacional:

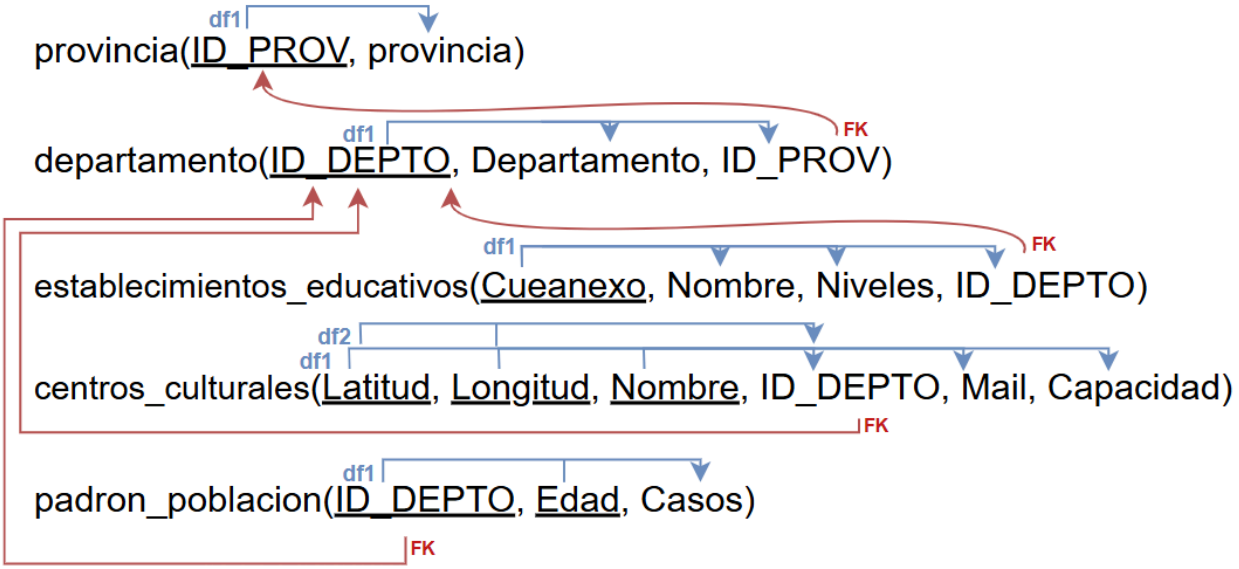

Figura 2: Modelo relacional sin normalizar.

Para el modelo relacional (Figura 2) resaltamos la dependencia funcional {Latitud,Longitud}-->{ID_DEPTO} pues es una dependencia parcial de la clave {lat,long,nombre}. También destacar que el atributo Mail no es atómico.

A continuación (Figura 3), el modelo relacional en 3ra forma normal. Notemos que, para centros_culturales la tabla mail_cc ahora contiene todas las direcciones de correo de forma atómica. También, la tabla localidad_cc ahora contiene latitud, longitud y el id de departamento, eliminando la dependencia parcial de la clave de centros_culturales y preservando la dependencia funcional df2.

Modelo relacional normalizado:

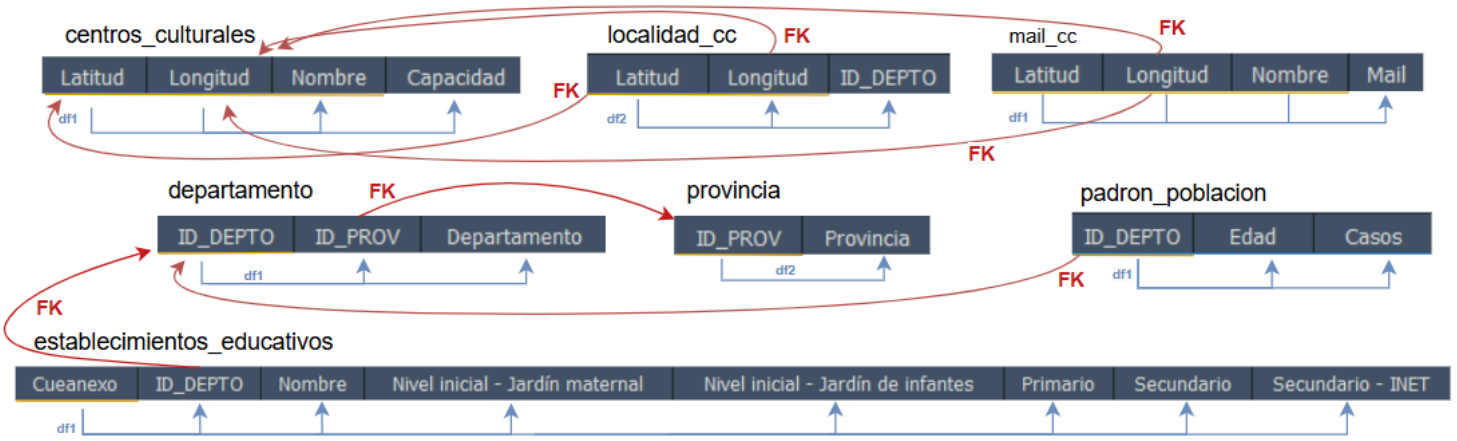

Figura 3: Modelo relacional en 3FN.

Decisiones tomadas

- **Omisión de datos innecesarios:** Para los tres datasets utilizados, decidimos quedarnos solo con aquellos atributos relevantes en función de los requerimientos del trabajo práctico.
- **Atributo 'Area' en padron_poblacion:** En el dataset original del padrón encontramos una mención al área y a la descripción de dicha área en la parte superior de cada sub-set de datos. Con el fin de poder usar este dato de manera sencilla, agregamos al dataset de padrón una columna con el número y descripción del área correspondiente a cada sub-set de datos. Que posteriormente normalizamos para eliminar la redundancia de la constante repetición del atributo 'Descripción'
- **Corrección de nombres de columnas:** Se eliminaron espacios adicionales en nombres como "Mail ".
- **Normalización de datos de contacto:** Se reestructuraron los datos de Mail y Teléfono en tablas auxiliares para evitar valores múltiples en un mismo campo.

- **Omisión de datos inválidos:** Se descartaron valores como "s/d" y "-" en Mail y Teléfono, ya que no aportaban información útil.

- **Cod_Loc:** debido a que Cod_Lod es la concatenación de id_provincia, id_depto y el código específico de la localidad, no es un atributo atómico, de modo que reemplazamos Cod_Loc por id_provincia e id_depto en la tabla de Establecimientos Educativos e hicimos tablas nuevas con el id_provincia y el nombre de la provincia y id_depto y el nombre del departamento.

- Al crear la tabla **departamento** que vincula un id_depto con el nombre del departamento, por un lado, tuvimos el problema de que en centros culturales había un único id_depto para toda CABA, considerando a toda CABA como un mismo departamento. En establecimientos educativos y en el padrón, CABA era dividida en diferentes departamentos, las comunas, que a su vez tenían un id_depto diferente del de los centros culturales para locaciones de CABA. Debido a que es más fácil convertir todas las comunas a un solo departamento, en vez de buscar todas las direcciónes de centros culturales y agregar a cada uno a qué comuna pertenece, **unificamos todos las comunas en un mismo departamento, CABA.** Por otra parte, los departamentos de Ushuaia y Río Grande aparecían intercambiados en algunas tablas. Además, aparecía un departamento en la Antártida entre los departamentos que se hallaban en la tabla de establecimientos educativos, debido a que hay una escuela en la base Marambio. Dado que este departamento no aparece en el padrón poblacional, no conocemos la población, ni la población estudiantil, o por lo menos, no aparece en la tabla de padrón poblacional ni mucho menos en la de centros culturales. **Por este motivo en las consultas de SQL, población y población estudiantil aparecen como Null,** porque desconocemos los valores.

- Para las visualizaciones ii e iii **utilizamos los datos de comunas**, omitidos para el análisis anterior. Ya que para ambos casos, debíamos trabajar solo sobre datos de establecimientos educativos y padrón, para los cuáles sí teníamos la información de las comunas y omitir dichos datos podría causar una distorsión de nuestras visualizaciones, pues todas las comunas se corresponderían con CABA en dicho caso, mostrando así una distribución más pobre de los datos.

- Al realizar la consulta iii de SQL, nos dimos cuenta de que en el **departamento de Tolhuin** aparecía el valor Null en establecimientos educativos. Buscando en el dataframe sin procesar de establecimientos educativos, encontramos que había escuelas que estaban localizadas en la ciudad de Tolhuin (que queda ubicada en el departamento de Tolhuin según una rápida búsqueda en Google) y sin embargo, en departamento aparecía Río Grande. El departamento de Tolhuin se creó en 2017, de modo que no es inusual que las personas sigan poniendo el departamento viejo en registros oficiales, sobre todo si la tabla es de 2022.

Análisis de datos

Del reporte del ejercicio i de consultas SQL y del segundo de los gráficos que nos pidieron, o sea el **Gráfico 1**, podemos observar que en general la población secundaria es mayor que la primaria que, a su vez, es mayor que la de jardín. En realidad, esto no es tan fácil de ver en el gráfico por lo que en el **Anexo** están los **gráficos 8 y 9** con diferentes superposiciones. Al observar los tres gráficos con diferente orden de superposición se aprecia mejor el área que ocupan los puntos de los diferentes niveles. El **Gráfico 2** muestra la cantidad de estudiantes por nivel educativo contra la población del departamento. En este gráfico es muy fácil ver que la relación es bastante lineal y en general hay más población secundaria que primaria que de jardín.

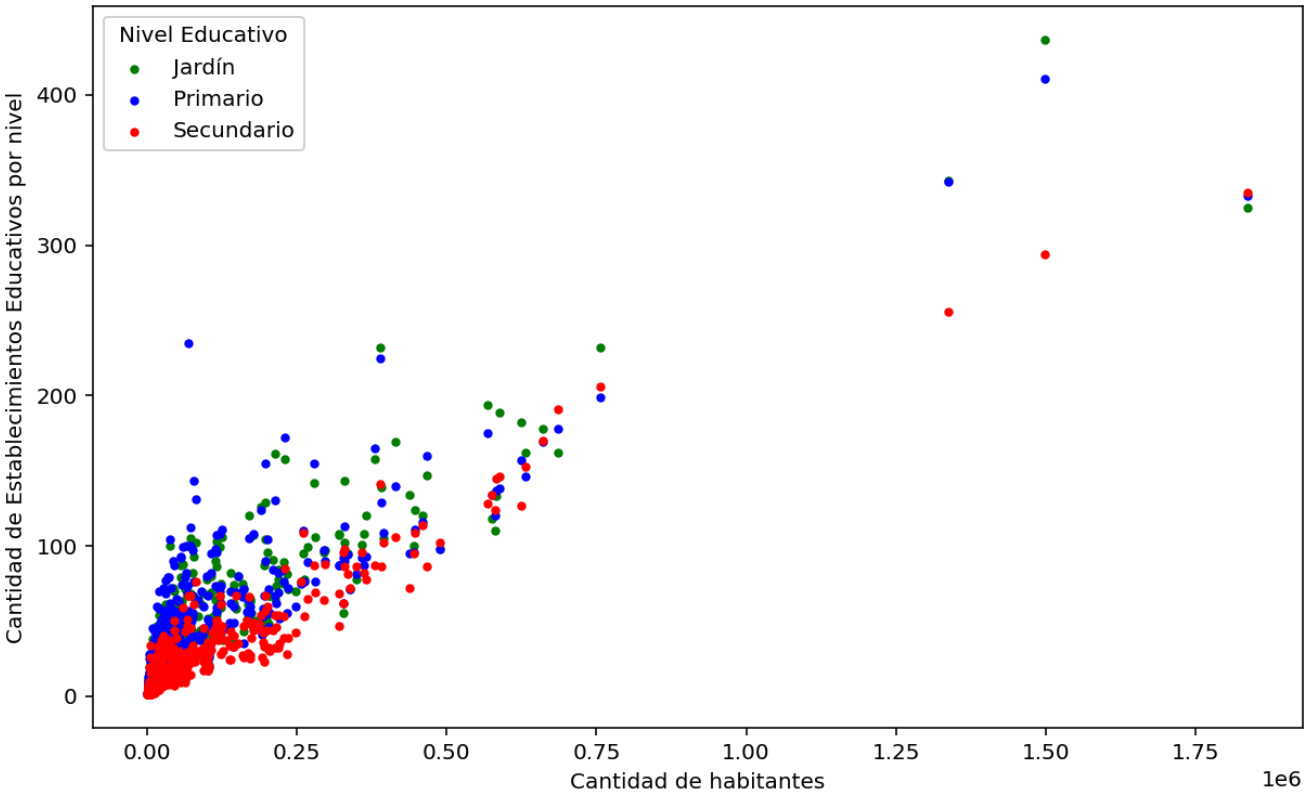

Gráfico 1: Cantidad de EE de los departamentos en función de la población, separando por nivel educativo y cantidad de población del correspondiente nivel

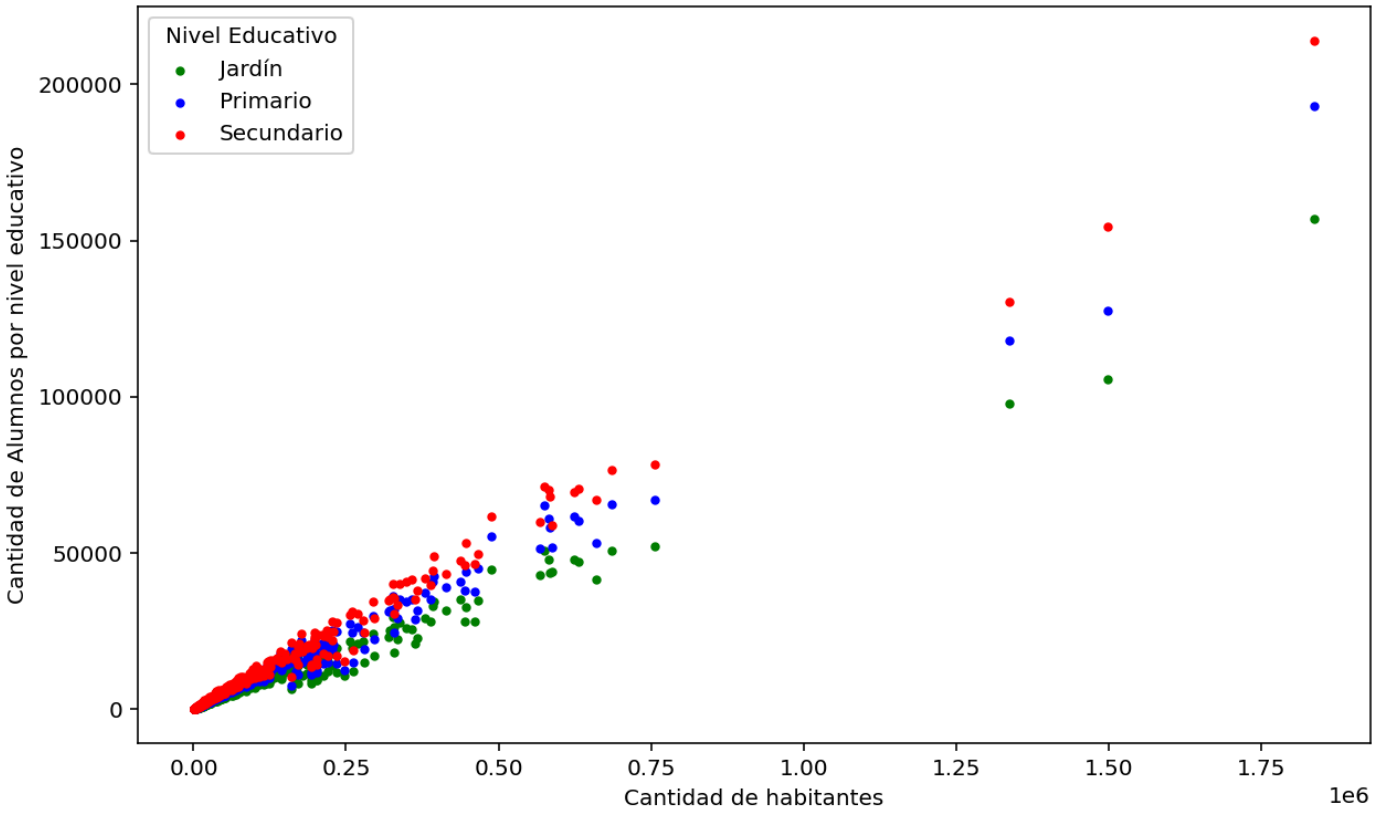

Gráfico 2: Cantidad de alumnos por nivel educativo vs cantidad de habitantes

De la comparación de los gráficos, podemos observar que en general y sobre todo en la agrupación de puntos del principio, hay más población en edad de ir a la secundaria que de ir a la primaria, y a su vez, esta población es mayor que la que está en edad de ir al jardín. Esto se puede notar por el hecho de que en el tercer gráfico, en donde los puntos rojos correspondientes a la secundaria están más arriba, éstos ocupan un área mayor hacia la derecha, indicando que el rango en el que se encuentran las poblaciones secundarias es más amplio hacia la derecha, o sea hay departamentos que tienen más población de ese nivel educativo. En el gráfico número II, se observa lo mismo comparado con el número I, hay departamentos que tienen más población en edad de ir a la primaria que en edad de ir al jardín. Esta situación, en la que hay más estudiantes de secundaria que de primaria, a pesar de tener en la mayoría del país la misma cantidad de años, podría deberse al descenso, en particular del periodo 2004-2016, de la tasa bruta de natalidad. Esta es la cantidad de nacidos vivos por año dividida la población medida a la mitad del año expresado por mil habitantes. Al tener la gente cada vez menos hijos, la población por edad va disminuyendo a medida que vemos edades más chicas. Otro factor que podría

influir son las migraciones.

Gráfico 3: tasa bruta de natalidad por mil habitantes. Fuente: Banco Mundial

Por otra parte, también podemos observar que, a pesar de volverse más cuantiosa la población a medida que se alcanzan niveles educativos, en general son más abundantes las primarias.

De la consulta número ii de SQL, podemos observar que son pocos los departamentos que tienen centros culturales, 53 de 514 departamentos y, entre los que los tienen, solamente dos tienen una cantidad que podríamos considerar “alta”, CABA y Avellaneda. De 53 departamentos con centros culturales, 31 tienen 1 solo centro cultural y el resto oscilan entre 8 y 2, siendo más escasos los departamentos a medida que aumenta la cantidad de centros culturales.

En la consulta número iii, volvemos a observar que muy pocos departamentos poseen centros culturales. Además, a priori, no observamos ninguna relación lineal entre establecimientos educativos y centros culturales. En las primeras 5 líneas, vemos una gran varianza entre la cantidad de centros culturales pero estos no aparecen ordenados, de modo que al disminuir la cantidad de establecimientos educativos no implica que la cantidad de centros culturales vaya a seguir el mismo comportamiento. Además, si observamos la tabla desde el final y vamos subiendo, hasta cierto número por lo menos, puede crecer la cantidad de establecimientos educativos y que no aparezca ningún centro cultural. Si seguimos subiendo, comenzamos a observar de vez en cuando algún departamento con centros culturales pero la cantidad no parece determinada por la cantidad de establecimientos educativos.

De la consulta número iv, observamos que en la columna de dominio más frecuente suele aparecer gmail, hotmail y yahoo. También aparecen algunos dominios extraños y hay muchos departamentos sin dominio más frecuente ya sea porque no tienen centros culturales o bien no se ha consignado el mail.

En el primer gráfico solicitado, el de cantidad de centros culturales por provincia, se puede ver que hay dos provincias o jurisdicciones que tienen una cantidad exorbitante de centros culturales camaradas con el resto: Buenos Aires y CABA. Les siguen otras dos que también sobresalen por sobre el resto, Santa Fe y Córdoba. Luego, el resto decrece lentamente. Podemos pensar que las cuatro provincias con más centros culturales tal vez lo tengan por la cantidad de población que poseen y los grandes centros urbanos. Sin embargo la relación entre cantidad de centros culturales por cantidad de habitantes no es muy clara. Para eso, podemos ver un gráfico de cantidad de habitantes por provincia.

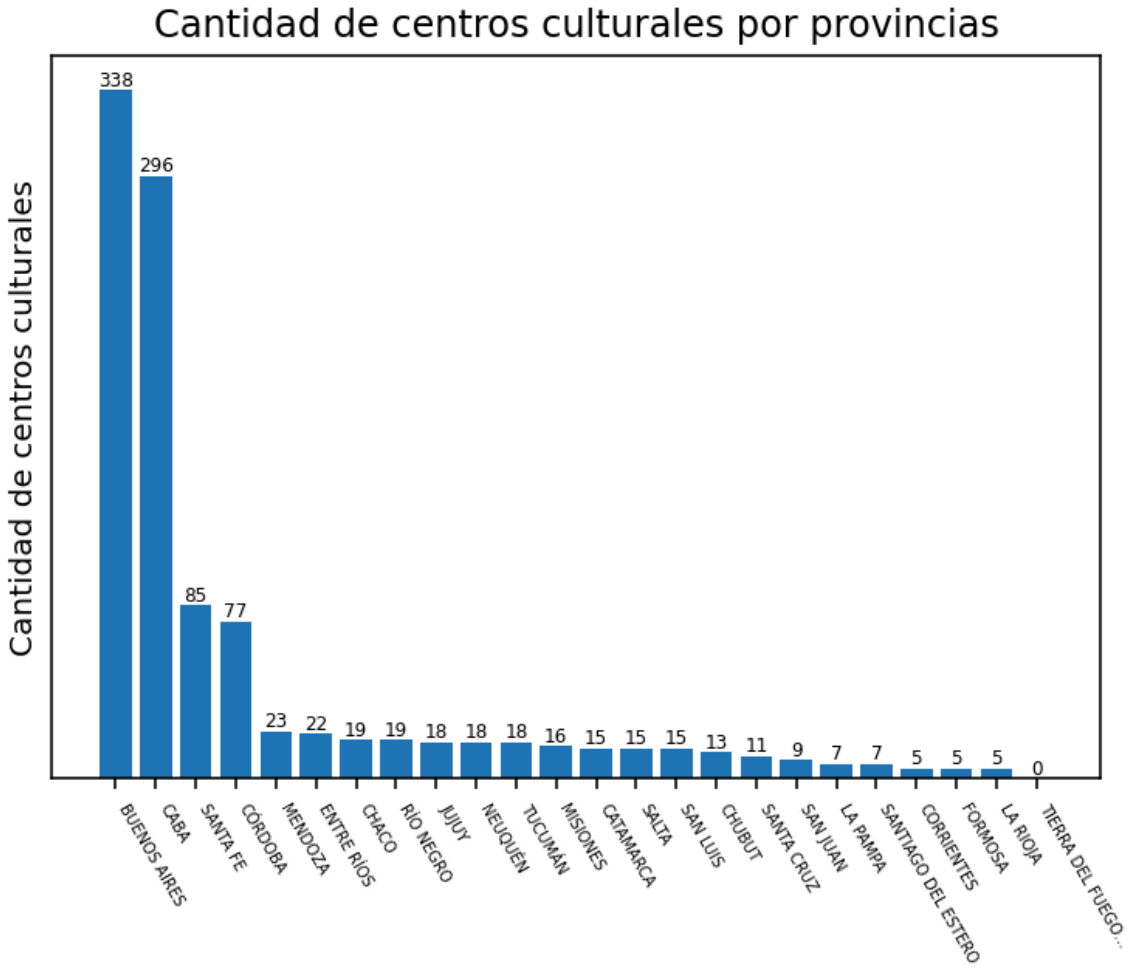

Gráfico 4: Centros culturales por provincia

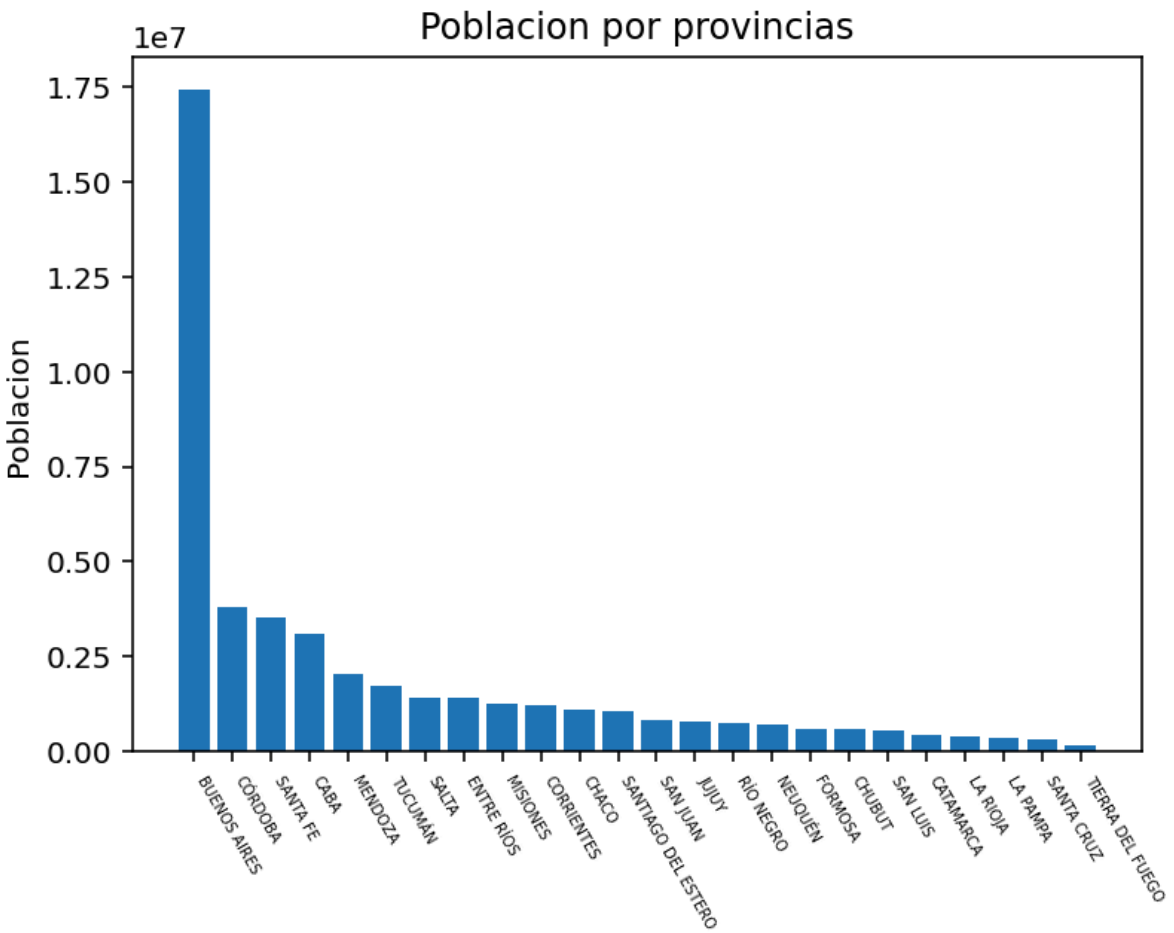

Gráfico 5: Población por provincias

Al segundo gráfico ya lo analizamos junto a la primer consulta de SQL

En el tercer gráfico, a simple vista no se observa ninguna relación clara. Podemos observar que Santa Fe, Formosa, Córdoba y Misiones tienen una cantidad más dispersa de centros educativos por departamento, dado que las cajas en donde está el 50% del medio de los números son más largas. O sea, la distribución es más dispersa. Por otra parte, en general, el bigote inferior es más chico que el superior, excepto en el caso de Entre Ríos y Santa Fe. Buenos Aires presenta muchos outliers y un bigote superior marcadamente más largo que el resto de provincias. Esto puede significar que hay algunos outliers muy alejados del resto, o sea un par de departamentos con muchas escuelas, tal vez a causa de una mayor población. No pudimos ver una relación en general que determine la distribución de las provincias ni el orden de las medianas.

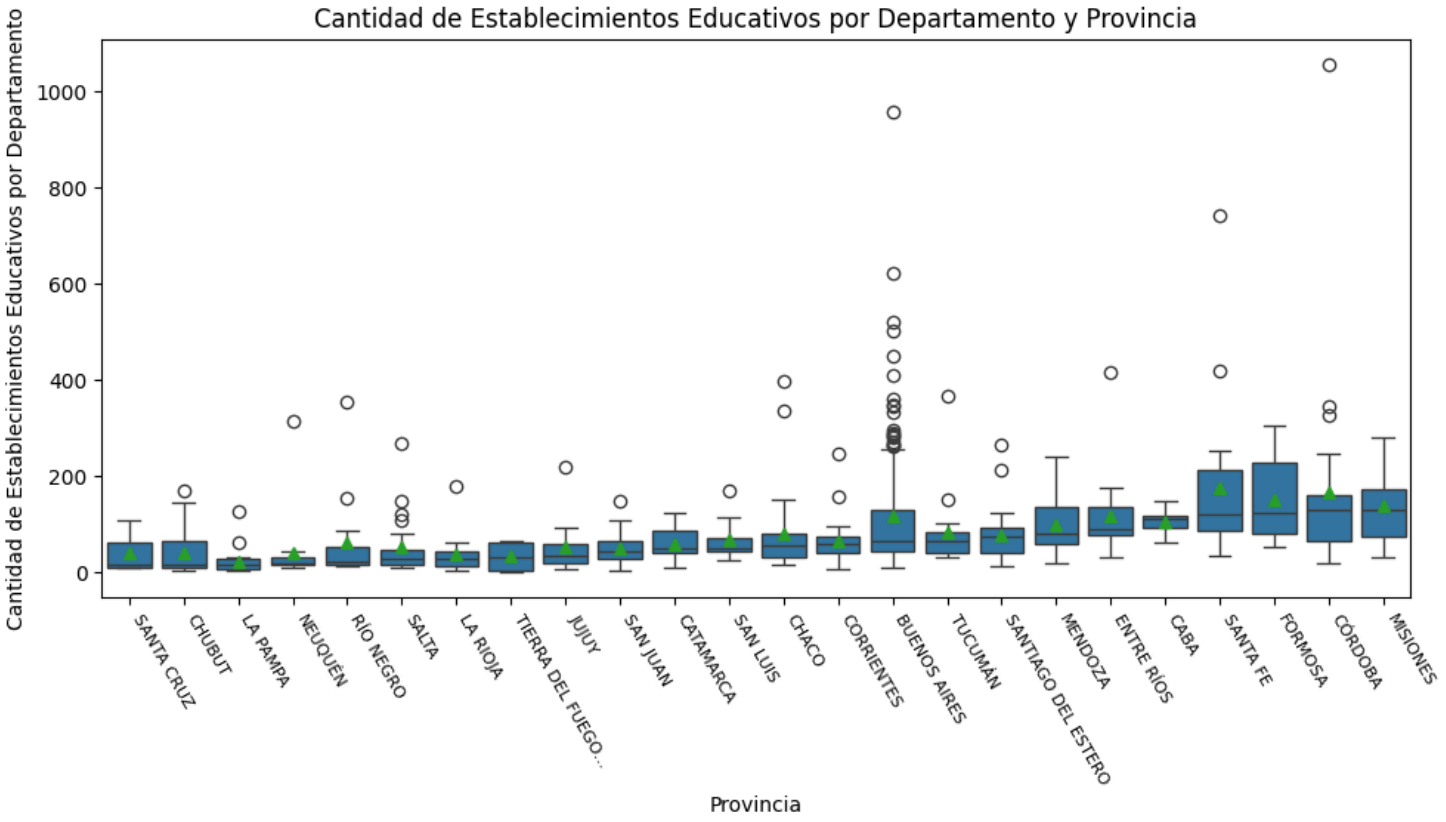

Gráfico 6: Cantidad de establecimientos educativos por departamento y provincia.

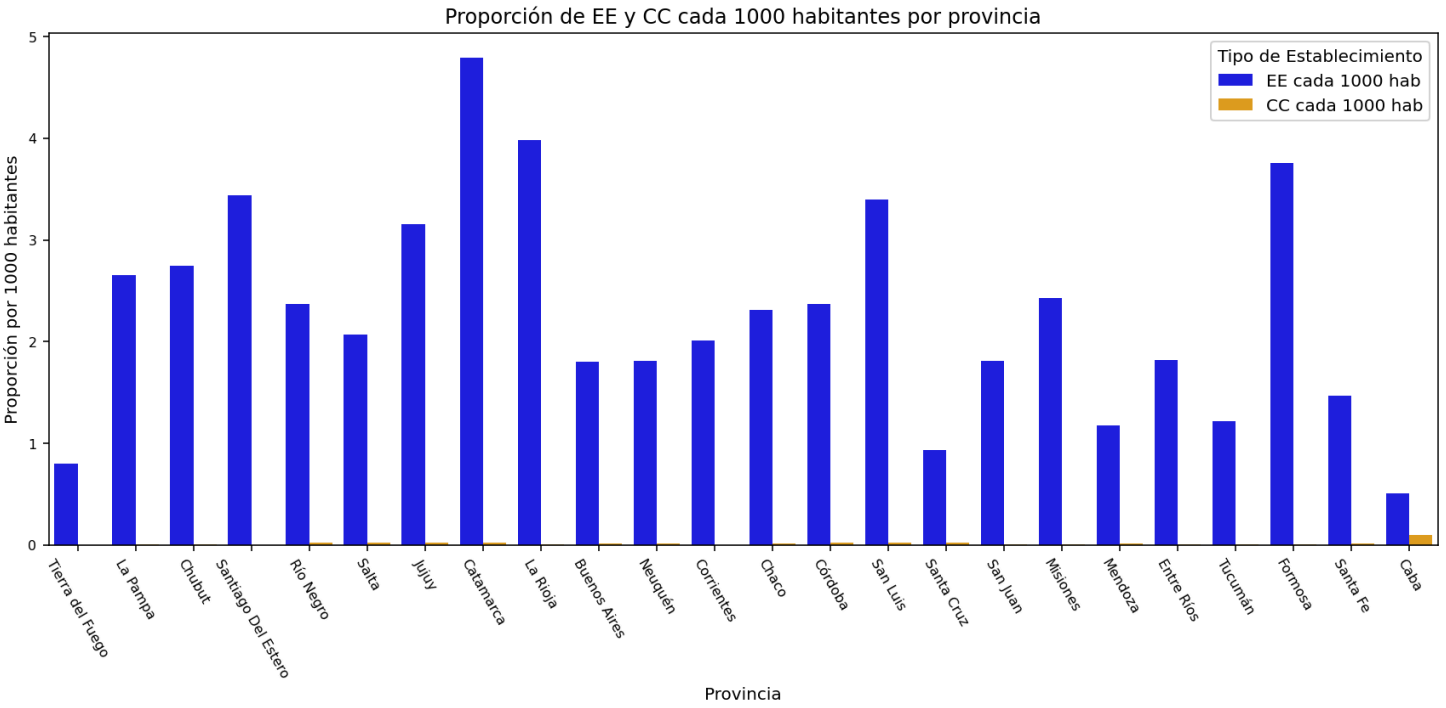

Gráfico 7: proporción de establecimientos educativos y proporción de centros culturales cada mil habitantes por provincia

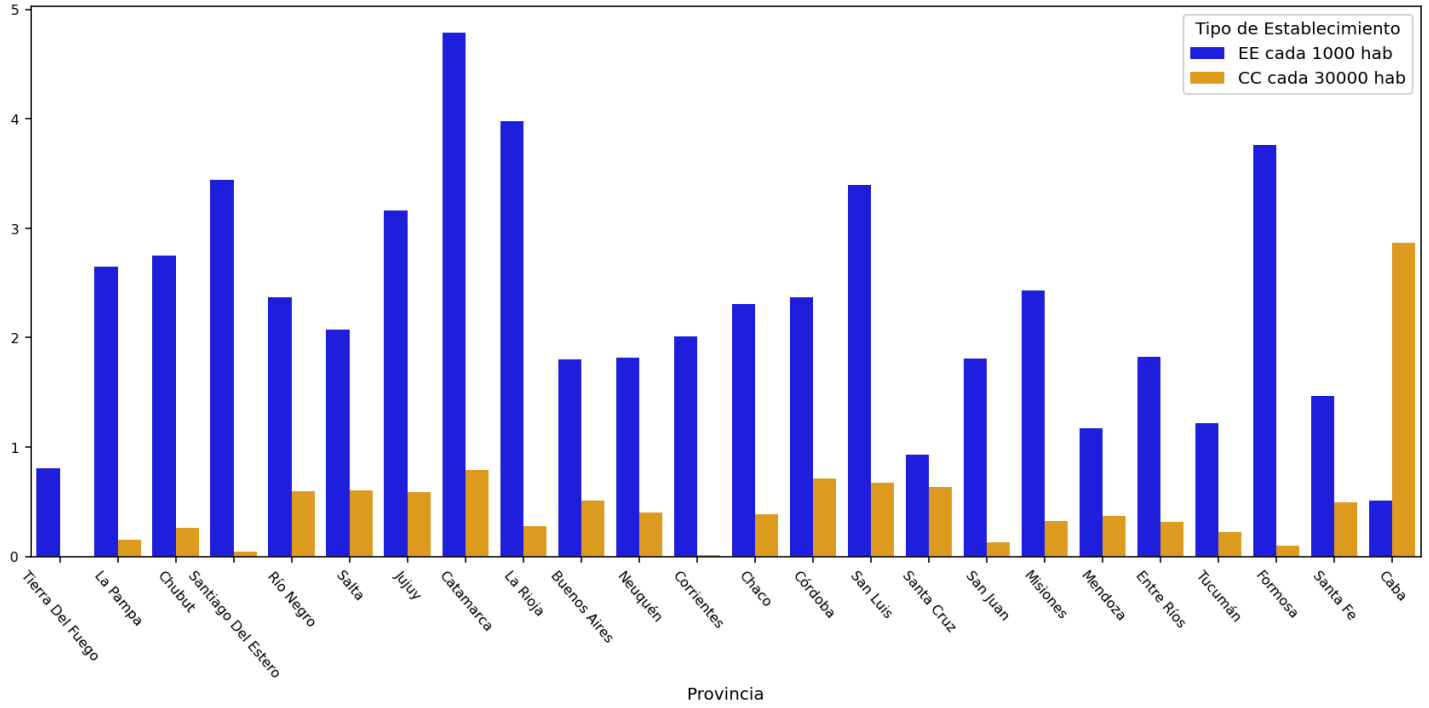

Gráfico 8: proporción de establecimientos educativos cada mil habitantes y proporción de centros culturales cada 30 mil habitantes por provincia

Conclusiones

El análisis realizado no encontró una relación clara entre la cantidad de establecimientos educativos y centros culturales en las provincias. Mientras que los primeros están más uniformemente distribuidos por necesidad de educación, los centros culturales se concentran en ciertas provincias, especialmente en áreas con alta densidad poblacional y mayores centros urbanos.

Aunque en algunos casos ambos tipos de establecimientos pueden crecer en paralelo, la relación no es lineal ni consistente en todas las provincias. Factores como la inversión en cultura podrían influir más en la distribución de los centros culturales que la cantidad de establecimientos educativos

Anexo:

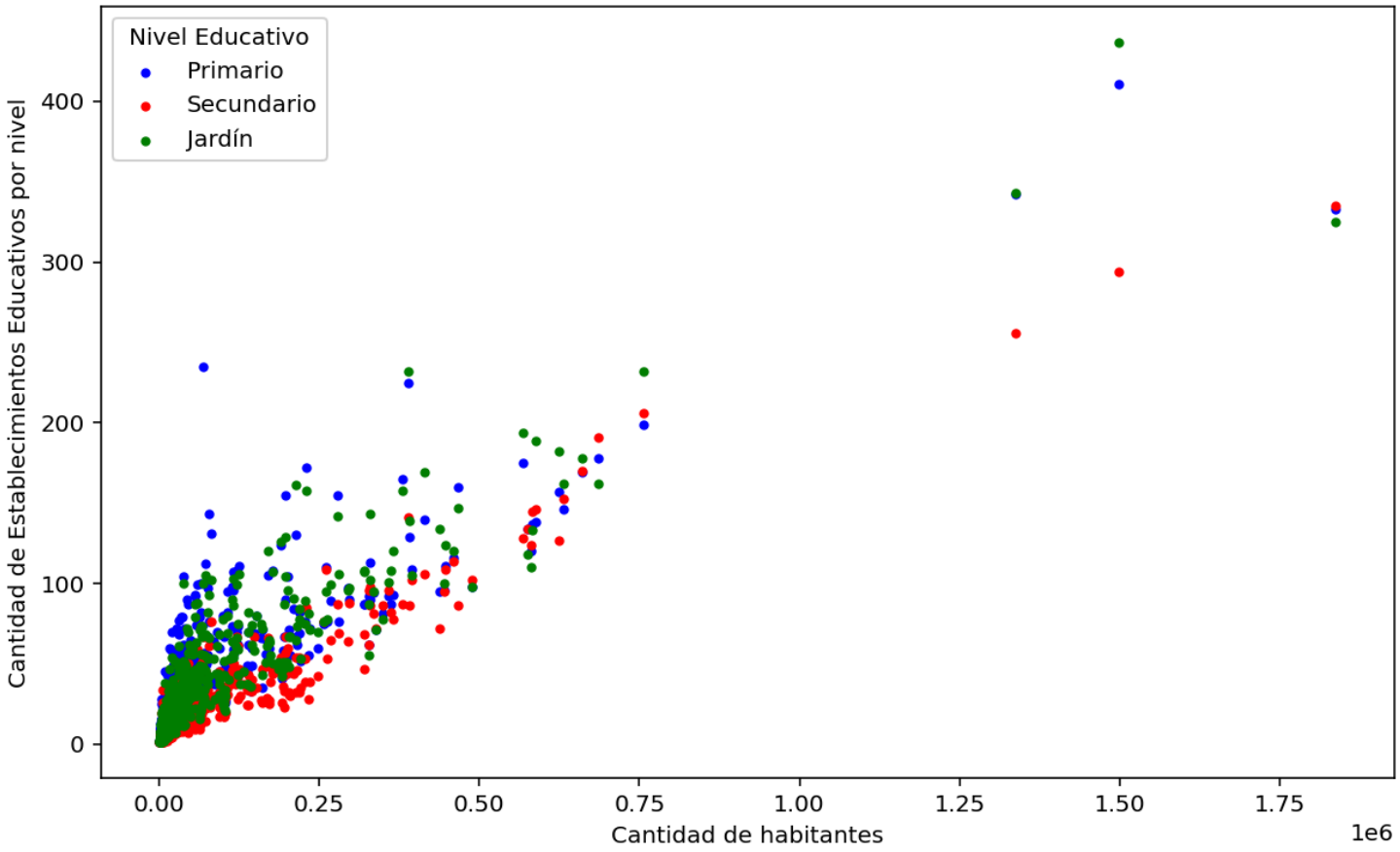

Gráfico 9: Cantidad de EE de los departamentos en función de la población, separando por nivel educativo y cantidad de población del correspondiente nivel (Orden de color 2)

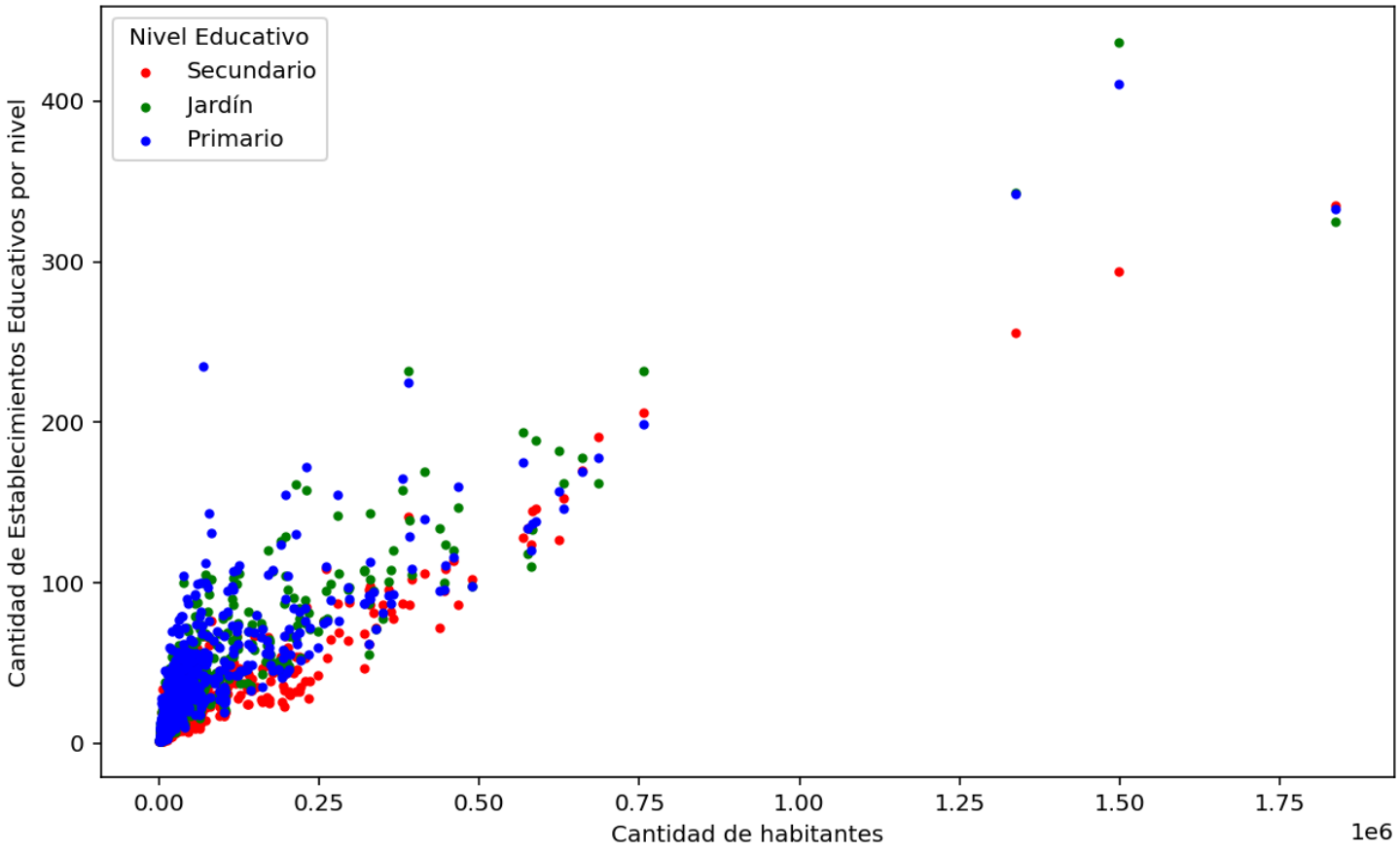

Gráfico 10: Cantidad de EE de los departamentos en función de la población, separando por nivel educativo y cantidad de población del correspondiente nivel (Orden de color 3)