DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico I

Manejo de datos y su Visualización

26 de Febrero de 2024 — Laboratorio de Datos

Integrante	LU	Correo electrónico
Teplizky, Gonzalo	201/20	gonza.tepl@gmail.com

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

Resumen

El presente trabajo tiene por objetivo descubrir si es posible establecer una relación entre el Producto Bruto Interno (PBI) por persona de cada país al año 2022 y la cantidad de sedes en el exterior que la Argentina tiene en dichos paises. Para ello se comparó desde distintos enfoques la información extraída de las fuentes de datos provistas pero no se consiguió encontrar una relación definida con los datos procesados, por lo que se infirió que se necesita más información para poder realizar una conclusión sobre si la relación que investigamos existe.

Palabras clave: *Datos*, *Relación*, *Sedes Diplomáticas*, *Argentina*, *PBI*, *2022*, *Visualización*, *Países*, *DER*, *Esquema*, *Representación*, *Consultas*.

Índice

1. Introducción - Presentación del problema

El establecimiento de sedes en el exterior presenta una necesidad inminente para un país, pues le permite enlazar y profundizar relaciones con otros países y expresar sus intereses políticos y económicos con los mismos. El objetivo de este informe es analizar si existe una relación entre las representaciones argentinas en el exterior y el PBI per cápita al año 2022 de los paises donde se encuentran las sedes. Para ello deberemos analizar los datos de fuentes de datos provistas por el Banco Mundial y el Ministerio de Relaciones Exteriores, Comercio Internacional y Culto. Las mismas serán tratadas con el fin de que los datos sean fieles a la realidad y luego se extraerán de ella los datos de interés para el informe. Luego se analizará la información procesada y se elaborará una conclusión respecto a los resultados.

2. Procesamiento de Datos

En esta sección se detallará el proceso de adquisición de los datos, junto con los tratamientos que se realizaron para aumentar la calidad de los mismos y cómo se los organizó de acuerdo al objetivo del informe.

En primer lugar, al analizar la forma normal de las tablas de Representaciones Argentinas se pudo determinar que:

La tabla sede-datos-basicos se encuentra en FN2, pues cumple con FN1 (es decir, todos sus atributos tienen valores atómicos y no cuenta con relaciones compuestas) y para todas sus dependencias funcionales vale que si un conjunto de atributos Y depende de un conjunto de atributos X, entonces X siempre es la clave primaria y el conjunto Y esta formado por atributos no primos, pero no llega a cumplir con FN3, pues existen conjuntos de atributos W, V y Z tales que V depende funcionalmente de W y Z depende funcionalmente de V, donde W es la clave primaria y ni V ni Z forman parte de ninguna clave candidata. Como no cumple con la tercer forma normal se puede inferir que tampoco cumple con la norma formal Boyce-Codd, que es más estricta.

Tanto la tabla lista-secciones-sedes como la tabla sede-datos-completos no cumplen con ninguna forma normal, pues cuentan con atributos multivaluados. Además, la tabla lista-secciones-sedes no cuenta con clave primaria, por lo que no existe ningún conjunto de atributos que permita identificar sus tuplas.

Teniendo en cuenta los datos provistos por los datasets de PBI per cápita de los países y Representaciones Argentinas se diseñó un diagrama entidad-relación, presentado en la figura 1, que contiene la información que se consideró relevante para el análisis de datos del informe. Para armarlo se tuvieron en cuenta las entidades País, Regioón_geográfica, Sede y Red_social_de_sede.

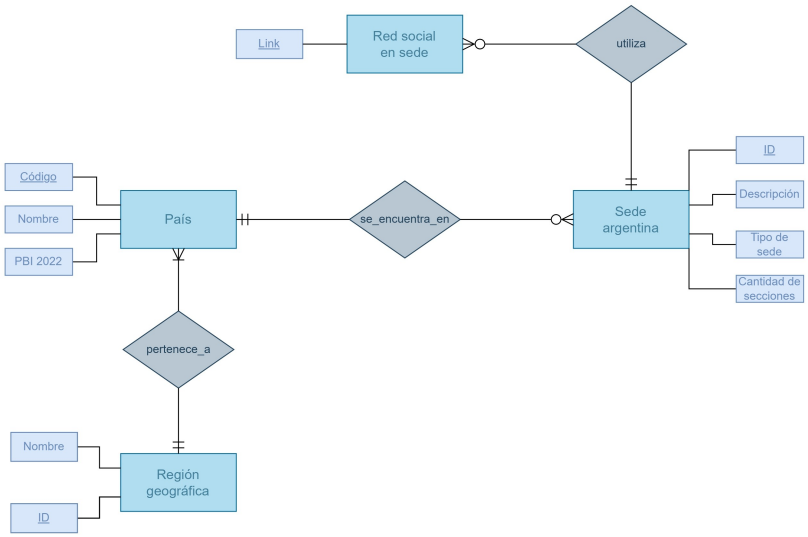

Figura 1: *Diagrama entidad-relación.*

Para poder implementar este esquema en el análisis de datos se realizó un mapeado del mismo al modelo relacional. Debido a que todas las relaciones del diagrama eran del tipo uno a muchos (*1:N*) se creó una relación por entidad, donde se agregó como clave foránea o *foreign key* (FK) la clave primaria o *primary key* (PK) de la relación correspondiente. A su vez se señalaron las dependencias funcionales (DFs) de cada relación con el fin de asegurar que las tablas cumplan con la tercer forma normal:

REGION_GEOGRAFICA(ID (PK), Nombre)

DF :

ID → Nombre

PAIS(Código(PK), Nombre, PBI_2022, Región_Geográfica_ID(FK))

DFs :

Código → { Nombre, PBI_2022, Region_Geografica_ID }

RED_SOCIAL_DE_SEDE(Link (PK), Sede_ID (FK))

DFs :

Link → Sede_ID

SEDE_ARGENTINA(ID (PK), Descripción, Tipo_Sede, Cant_Secciones, Pais_Código (FK))

DFs :

ID → {Descripción, Tipo_Sede, Pais_Código, Cant_Secciones }

Debido a la presencia de atributos de calidad afectados, una vez armado el modelo de datos

que se utilizará para resolver la problemática se llevaron a cabo procesos de calidad de datos con el objetivo de que la información que se analice sea lo más fiel posible a la realidad. Para esto se siguió el enfoque *GQM* (*Goal Question Metric*). Para cada atributo de calidad afectado que se encuentra se plantea un objetivo, una o más preguntas cuyas respuestas satisfagan el objetivo y se proponen métricas para cada una que permitan responder las mismas.

En el dataset sedes-datos-completos se encontró que el atributo redes sociales carece de relevancia. Se considera que es un problema de instancia, ya que hay datos que carecen de la precisión necesaria (por ejemplo, se cargó un nombre de usuario pero no se especificó a qué red social pertenece).

Analizando la *relevancia* del atributo redes sociales del dataset sedes-datos-completos.csv:

Goal → El dato correspondiente a la/s Red/es Social/es de cada Sede es relevante.

Question → ¿Qué proporción de Redes Sociales son links de plataformas de redes sociales?

Metric → M: Proporción de registros con campo Redes Sociales que contenga el nombre de una red social (Twitter, Instagram, Facebook, LinkedIn, Youtube, Flickr, Gmail, es decir

$$\frac{\text{Cantidad de registros de redes sociales de sedes con campo Redes Sociales que contenga una de las principales redes sociales}}{\text{Cantidad total de registros de redes sociales de sedes}}$$

Solución: como en nuestro esquema la columna redes_sociales está en una tabla donde es clave primaria, las tuplas que no cumplan con la condición serán borradas para evitar la presencia de *NULLs*.

En el dataset sedes-datos-basicos.csv se encontró que el atributo país_castellano carece de consistencia. Es más bien un problema de instancia o de procesos, la información se debe cargar de otras fuentes de datos o la cargan distintas personas y por eso surgen inconsistencias.

Analizando la *consistencia* del atributo país_castellano del dataset sedes-datos-basicos.csv:

Goal → El dato correspondiente al nombre del país en castellano de cada Sede es consistente.

Question → ¿Qué proporción de nombres de países en castellano están escritos en mayúscula?

Metric → M: Proporción de registros con campo país_castellano que se encuentren en mayúscula, es decir

$$\frac{\text{Cantidad de registros de sedes-datos-basicos con campo país\_castellano que contenga una de las principales redes sociales}}{\text{Cantidad total de registros de sedes-datos-basicos}}$$

Solución: Estandarizar la columna asegurándose que todos los valores estén en mayúscula.

En el dataset API_NY.GDP.PCAP.CD_DS2_en_csv_v2_6298251.csv se encontró que el atributo 2022 carece de completitud. Se trata de un problema de modelo pues falta la información necesaria para calcular el PBI de ciertos países o subregiones en 2022 y por eso algunos campos se encuentran vacíos.

Analizando la *completitud* del atributo 2022 del dataset API_NY.GDP.PCAP.CD_DS2_en_csv_v2_6298251.csv:

Goal → El dato correspondiente al año 2022 del PBI de cada país está completo.

Question → ¿Qué proporción de países que tienen el dato correspondiente a 2022 está vacío?

Metric → M: Proporción de registros con campo 2022 que se encuentren vacíos en tabla de PBI per cápita de países, es decir

$$\frac{\text{Cantidad de registros de PBI per cápita de países con campo 2022 vacío}}{\text{Cantidad total de registros de PBI per cápita de países}}$$

Solución: Si se considera que la cantidad de datos que se pierden no es significativa (ejemplo: la métrica da menor al 10 %) se pueden borrar los registros de la tabla donde el campo 2022 está vacío.

También se utilizó la tabla lista-secciones-sede para generar la columna cant_secciones en la tabla Sede, sin embargo no se encontraron atributos de calidad afectados en dicha tabla que sean relevantes para la información que se precisaba.

Finalmente, se importaron los datos limpios que conforman el esquema relacional planteado previamente. Los mismos se extrajeron de las fuentes de datos de PBI per cápita de los países, Datos de sedes completos y Datos de sedes básicos. También se utilizaron los datos de la fuente de datos Lista de secciones de sedes para generar la columna Cantidad de secciones.

3. Decisiones tomadas

Al momento de pensar en el procesamiento de los datos provenientes de las fuentes que teníamos a disposición, decidimos tomar en consideración los siguientes puntos:

Con respecto a los datasets provistos por las fuentes de información y la extracción de los datos elegidos:

- Observamos que el dataset *lista_secciones-sedes* no contenía una forma unívoca de identificar a cada sección de una sede particular. Atributos como *sede_id, sede_desc_castellano, y tipo_seccion* fueron los elegidos como clave primaria en un principio por hallarse completos, pero los 3 presentaban duplicados que solamente podían diferenciarse recurriendo a otros atributos que no se encontraban para todos los registros por igual. Es por esto que decidimos reducir la entidad *Sección* al atributo *cant_secciones* de la relación *Sede*, dado que su uso final se explica en obtener la cantidad de secciones para cada una de las sedes analizadas.

- Para los valores pertenecientes a las relaciones creadas para modelar los datos que utilizamos para el análisis, definimos estandarizar todos los atributos pertenecientes al idioma español de los datasets originales. Esto respondió simplemente a querer mantener esta consistencia entre todos los valores de las tablas que involucramos.

Con respecto a la limpieza de estos datos y las auditorías realizadas:

- Para modelar las redes sociales optamos por mantener una entidad que hiciera referencia a las redes sociales por sede. A partir de cada una de estas, representada por su link correspondiente, extraemos la red social a la que este link hace referencia. Esto es lo que se ve reflejado en el reporte respectivo de la información de las redes sociales.

4. Análisis de datos

Para empezar el análisis, habría que aclarar que es importante tener presente el objetivo del trabajo e intentar limitar el enfoque con el que evaluamos los datos a responder las preguntas que de éste surgen. Es decir, orientar el estudio de los datos que obtuvimos de manera tal que éste

nos permita dar una respuesta concreta al problema que estamos tratando, evitando ahondar en cuestiones que no contribuyan a lograr este objetivo. Con esto en mente, realizamos a continuación el análisis sobre las tablas generadas por las consultas explicitadas en el enunciado, buscando patrones y relaciones que sean consistentes con el propósito del presente trabajo, remarcando las ventajas y limitaciones que tiene este medio (es decir, archivos .csv) de análisis de datos (comparado con un formato más visual, por ej, el gráfico) y finalmente, intentando dar respuestas al problema del TP.
Primer reporte: es muy evidente que, a primera vista, encontrar una relación entre el número de sedes argentinas localizadas en un país particular junto con su PBI en el año 2022 es dificultoso, ya sea por la forma en la que se decidieron representar los datos (ordenando decrecientemente por número de sedes, habiendo potencialmente mejores criterios de ordenamiento) o la utilización de un formato poco adecuado para representar el contenido en sí, aunque tampoco podemos descartar la posibilidad de que no haya relación entre estas dos variables. Retomaremos esta cuestión en la parte visual del análisis, viendo que no tenemos herramientas suficientes para establecer un vínculo.
Segundo reporte: si bien no encontramos una relación directa entre la cantidad de sedes por país y su PBI, podríamos reformular la pregunta y conjeturar si el PBI per cápita se comporta de alguna manera en particular cuando consideramos la cantidad de países con sedes argentinas, por región geográfica. Sin embargo, observando los datos, llegamos a la misma respuesta: no hay relación entre las variables (o, al menos, pareciera no haberla). Habiendo dicho esto, la situación no es exactamente la misma que en el primer reporte: la cantidad de tuplas en esta tabla es considerablemente menor. Este hecho por sí sólo facilita muchísimo el análisis, y sumado al comportamiento oscilatorio de la cantidad de sedes a medida que bajamos en la tabla, se puede pensar con más seguridad que la relación es nula (aunque, nuevamente, no podemos confirmar que esto sea así).
Tercer y cuarto reporte: en ambos casos, es difícil encontrar el común denominador entre cada reporte y el problema a responder, teniendo en cuenta que estamos trabajando con tablas que carecen de los datos necesarios para poder dar respuesta a la pregunta central del trabajo.

Podemos afirmar que, suponiendo que efectivamente existe una relación entre el PBI per cápita de un país en 2022 y la cantidad de sedes argentinas que éste tiene, las consultas realizadas para el ítem h del trabajo no permiten encontrar dicha relación, ya sea por tratarse de un formato poco adecuado para la interpretación de datos (como comentamos en el *primer reporte*), o por el simple hecho de que el contenido de las tablas es insuficiente y no está relacionado con la pregunta a la que queremos responder (como ocurre en el *tercer y cuarto reporte*).

A continuación realizamos la parte visual del análisis, nuevamente focalizándonos en los datos que son relevantes para el estudio de este trabajo.

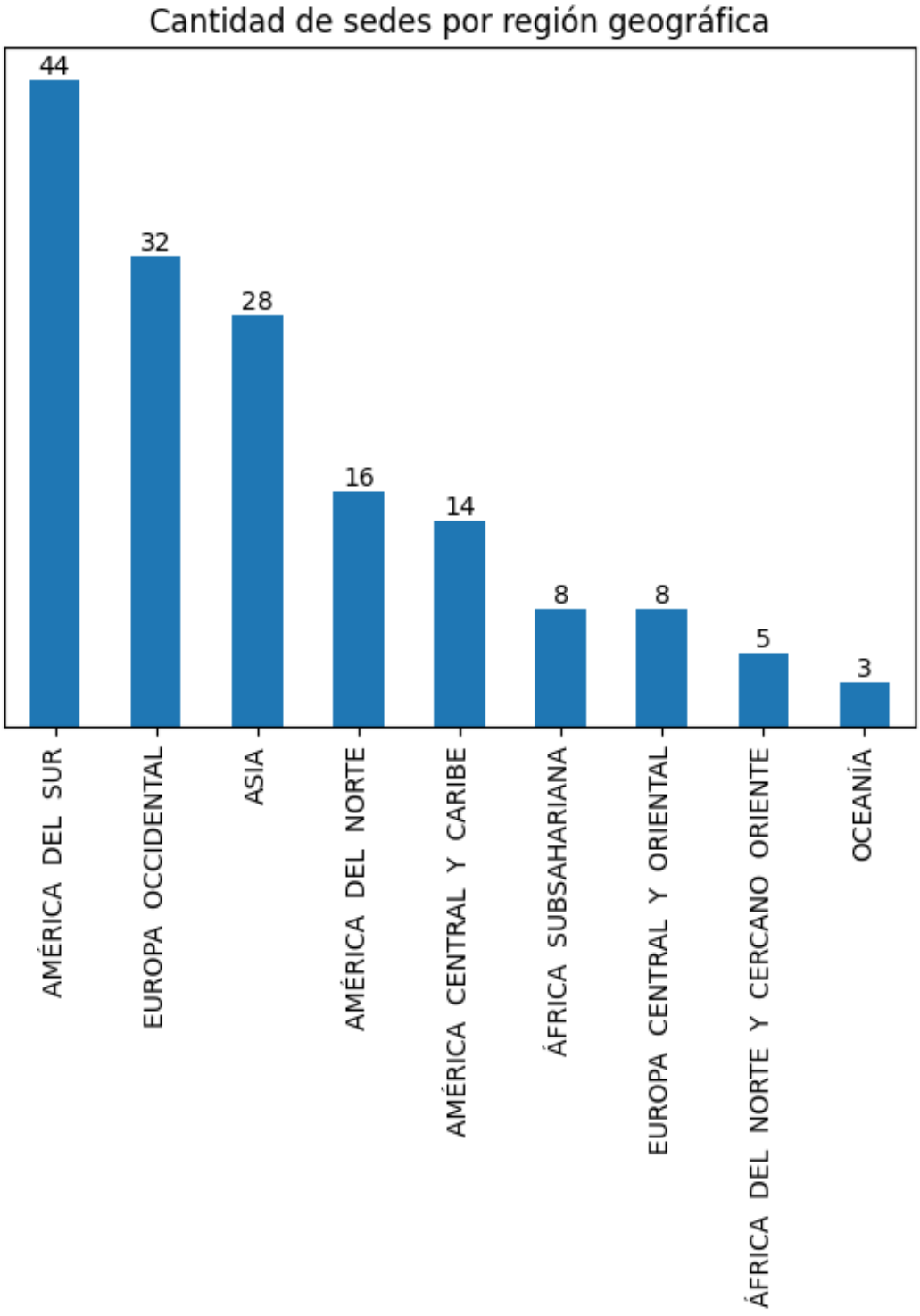

Figura 2: *Cantidad de sedes por región geográfica*

Aspectos relevantes a tener en cuenta para el análisis de la figura 2:

→ Proximidad geográfica: notamos que la región que más sedes tiene es América del Sur; considerando que Argentina forma parte de ésta, podemos pensar que la cercanía territorial es un factor importante en relación a esto, y determina, en cierta medida, el número de sedes localizadas

en un país. Sin embargo, no es el único criterio determinante en este sentido, viendo que Europa Occidental supera ampliamente el número correspondiente a América Central/Caribe. A raíz de esto, tuvimos en cuenta otra posible causa que explica la discrepancia en cantidad entre las distintas regiones: las relaciones políticas que mantiene la Argentina con los estados de cada región.
→ Relaciones políticas: consideramos que esto justifica coherentemente la diferencia cuantitativa de sedes entre, por ejemplo, Europa Occidental y África Subsahariana, o América del Norte y América Central/Caribe: se ve inmediatamente que, en ambos casos, la cercanía geográfica no puede dar una respuesta lógica a este fenómeno. Pensamos a la influencia y los vínculos políticos como posibles factores alternativos que, en este contexto, dan una mejor respuesta a esta incógnita.

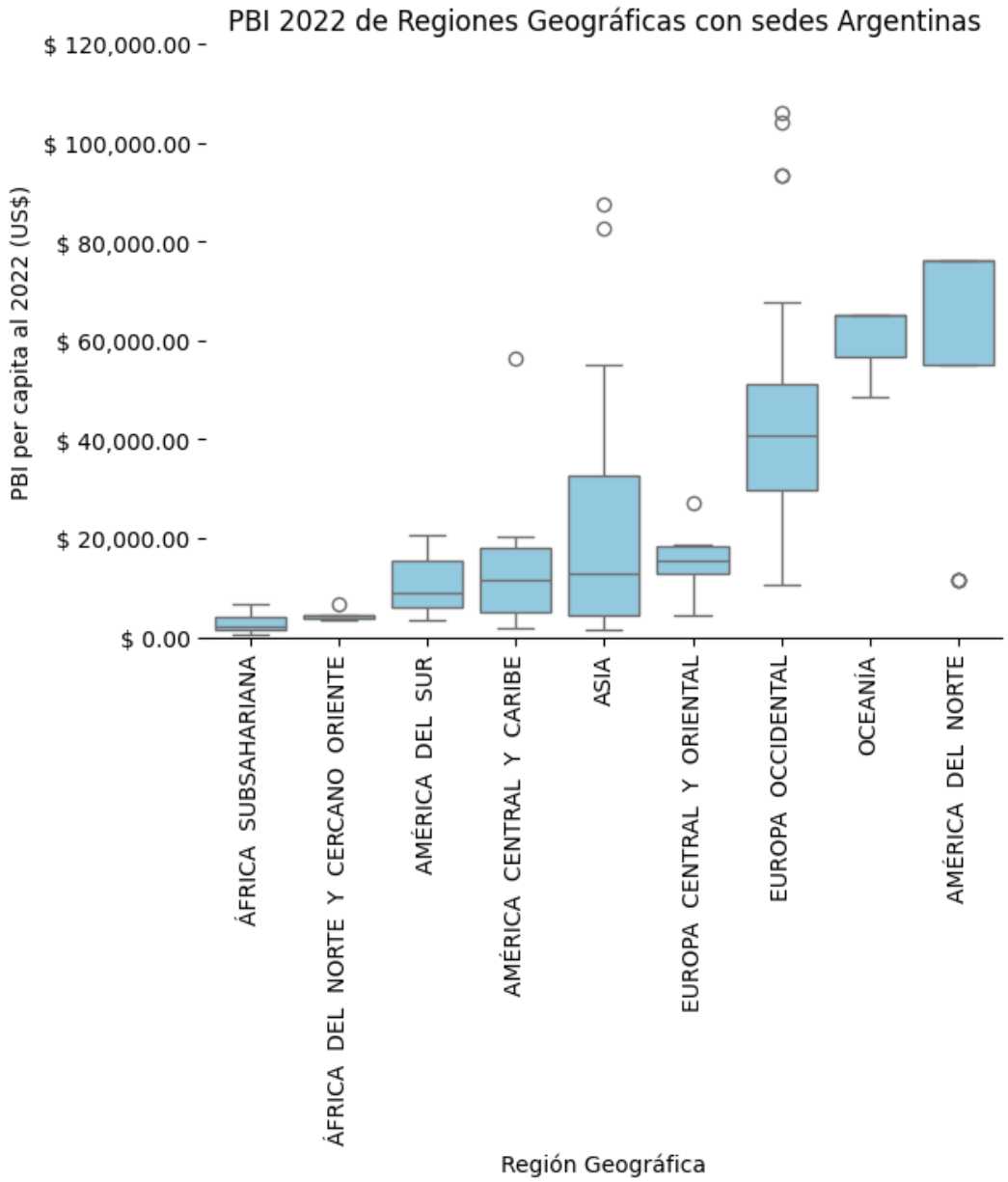

Figura 3: *Gráfico Boxplot de los PBI per cápita de regiones geógraficas con sedes argentinas.*

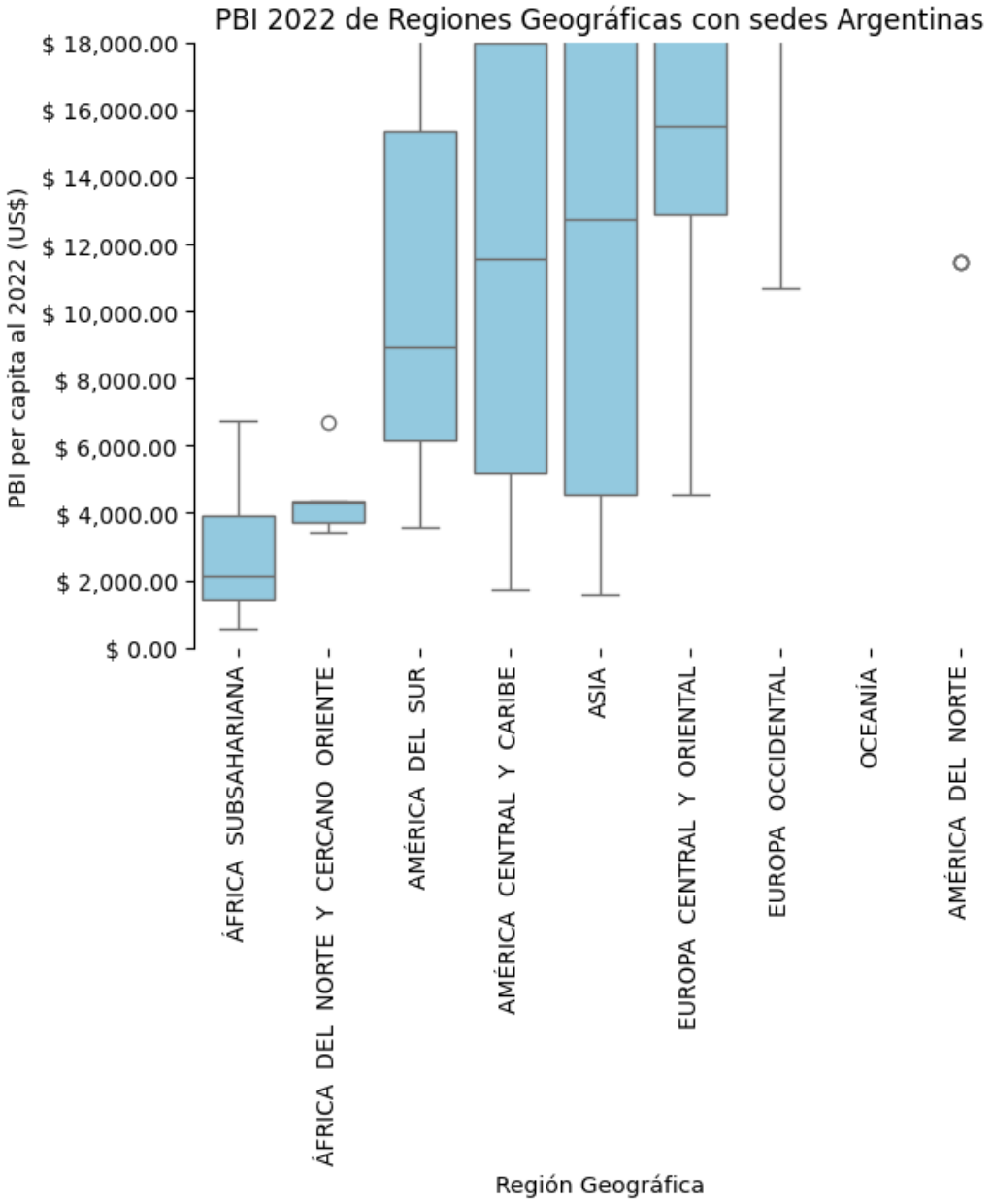

Figura 4: *Gráfico Boxplot de los PBI per cápita de regiones geógraficas con sedes argentinas (con zoom para poder apreciar mejor algunos datos).*

Si observamos la figura 3 (o en la figura 4, donde se pueden apreciar mejor algunos boxplots) podemos notar que las medianas de los boxplots en algunas regiones se encuentran muy alejadas de sus límites. Esto se debe a que si bien algunos paises de una región tienen un PBI per cápita muy alto, en general la mayoría suele ser mucho más bajo, lo que hace que la mediana se acerque mucho más a su límite inferior. En otros casos sucede algo similar pero al revés: la mediana se encuentra mucho más cerca del PBI más alto de la región, por lo que se puede concluir que hay una gran brecha entre los PBI per cápita de los países de cada región.

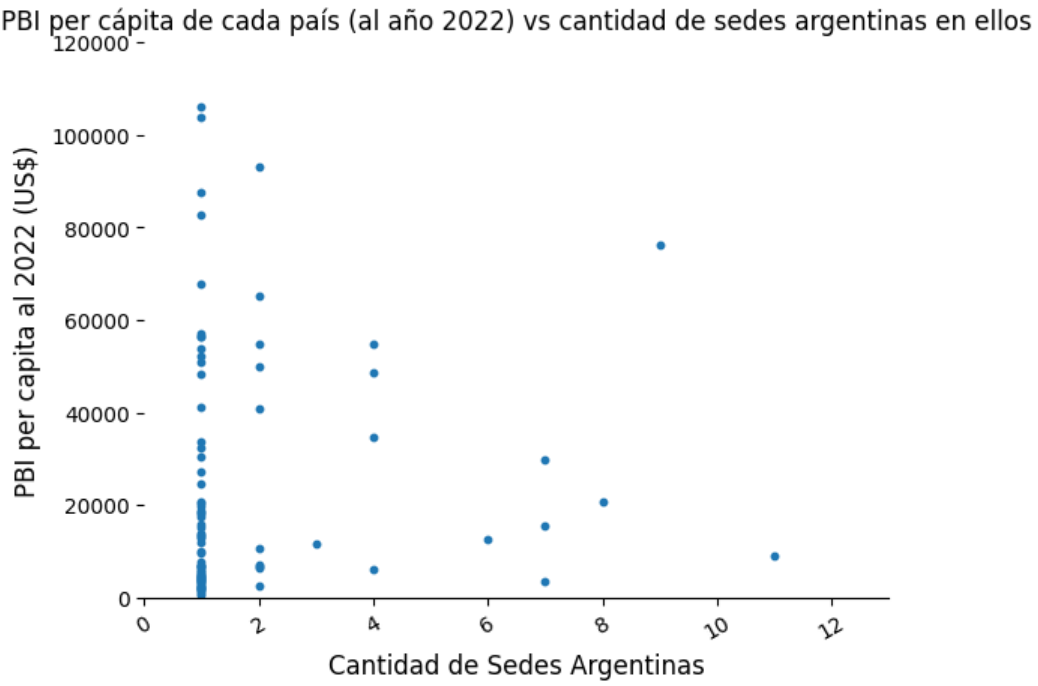

Figura 5: *PBI per cápita de cada país al año 2022 vs Cantidad de sedes argentinas en ellos.*

Como se puede observar en la figura 5 hay muchos países cuyas cantidades de sedes argentinas coinciden pero el rango de PBI per cápita es muy amplio, por lo que no es posible discernir una relación entre la cantidad de sedes argentinas en un país y su PBI per cápita al año 2022. Tal vez sería más lógico investigar la cantidad de sedes argentinas en promedio en países agrupados según su PBI per cápita (menor, mediano o mayor). También puede resultar útil analizar la cantidad de embajadas respecto a la cantidad de consulados argentinos en un país y su PBI, ya que si bien ambos son sedes diplomáticas el objetivo de los consulados suele ser principalmente la representación y apoyo de argentinos en el exterior, mientras que las embajadas suelen manejar el sector ecónomico y político de la relación entre Argentina y dicho país.

Otro análisis que podría brindar información sobre las relaciones argentinas con un país y el PBI per cápita del mismo puede ser considerar cuántas redes sociales hay en promedio de sedes argentinas en países con cierto PBI, o incluso cuáles redes sociales predominan y si estas coinciden con las redes sociales más populares.

5. Conclusiones

Si bien se analizaron varios puntos que conciernen las representaciones argentinas y el PBI per cápita de países en 2022, no se pudo encontrar una relación entre éstos. Esto puede significar que no se cuenta con suficiente información sobre las sedes argentinas en el exterior y nuestros vínculos económicos y/o políticos con ciertos países, o que la relación en sí entre estos factores no existe.