DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 01

Evaluación del manejo de datos y su visualización

27 de febrero de 2024 — Laboratorio de Datos

Grupo : El Peligroso

Integrante	LU	Correo electrónico
Cocú, Dante	1119/22	dcocu19@gmail.com
Navarro, Solana	906/22	solanan3@gmail.com
Said, Tomás Uriel	170/23	saidtomasur@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

1. Resumen

Este proyecto se enfoca en la gestión efectiva de datos, utilizando modelos y representaciones abstractas para organizar la información de manera clara y concisa, seguido de la creación de visualizaciones que faciliten la comprensión, identificación de relaciones y tendencias, y extracción de conclusiones. Se trabajó con diversas bases de datos, incluyendo datos sobre el Producto Interno Bruto (PIB) per cápita de varios países obtenidos del Banco Mundial y datos sobre las representaciones argentinas en el exterior proporcionados por el Ministerio de Relaciones Exteriores y Culto.

---

2. Introducción

El objetivo de este proyecto es llevar a cabo la correcta gestión de datos, realizando modelos y representaciones abstractas para organizarlos de manera clara y concisa, y posteriormente elaborar una visualización que nos ayude a comprenderlos, encontrar relaciones y tendencias y sacar conclusiones. Para lograr esto, nos involucramos con diversas bases de datos. Por un lado, utilizamos los datos sobre el Producto Interno Bruto (PIB) per cápita de diferentes países, que fueron obtenidos del Banco Mundial. Por otro lado, empleamos la información acerca de las representaciones argentinas en el exterior proporcionada por el Ministerio de Relaciones Exteriores y Culto.

El informe sigue una serie de actividades detalladas, comenzando por la comprensión del contenido de las fuentes de datos y la descarga de los mismos. Luego, desarrollamos un Diagrama Entidad-Relación (DER) para modelar los datos necesarios y procedimos a crear el Modelo Relacional basado en este DER. Posteriormente, generamos los DataFrames en Python siguiendo nuestro modelo relacional, asegurándonos de que estén en Tercera Forma Normal (3FN) y corrigiendo problemas de calidad de datos.

Utilizamos la técnica GQM (Goal Question Metrics) para mejorar la calidad de las bases de datos, identificando problemas como URLs incorrectas y registros faltantes. Además, tuvimos que tomar decisiones como definir las redes sociales relevantes y realizamos análisis de datos para cumplir con diversas consignas, tal como realizar consultas SQL y armar gráficos con datos pedidos.

Finalmente, creamos visualizaciones para facilitar la comprensión de los datos, incluyendo gráficos de barras para mostrar la cantidad de sedes por región geográfica, boxplots para analizar la distribución del PIB per cápita por región y un scatter plot para buscar una relación entre la cantidad de sedes argentinas en un país y su PIB per cápita. Estas visualizaciones ayudan a identificar patrones y tendencias importantes en los datos analizados. Redactamos una conclusión sustentada en los análisis previamente hechos.

3. Procesamiento de Datos

Para realizar este trabajo, se nos proporcionaron 4 bases de datos que contienen:

- Los datos resumidos de las sedes (*sede-basico.csv*)
- Los datos completos de las sedes (*sede-completo.csv*)
- Los datos de las secciones de las sedes (*sede-secciones.csv*)
- La información del PIB Per Cápita de todos los países desde el año 1960 hasta el 2022 (*pibpercapita.csv*)

Cabe destacar que los datos de las tres primeras tablas provienen del Gobierno Nacional, mientras que los datos del PIB Per Cápita son suministrados por el Banco Mundial.

La primera problemática que debemos abordar es la referente a la calidad de los datos que poseemos. En primer lugar, resulta evidente que los datos de la tabla *sede-completo* no cumplen con la Primera Forma Normal (1FN), dado que los registros de la columna *redes_sociales* no son atributos atómicos. Por lo tanto, será necesario encontrar la manera de convertirlos en atributos atómicos para garantizar que nuestra base esté en Tercera Forma Normal (3FN). Además, será crucial trabajar en el filtrado y la mejora de calidad de los datos en esas tablas. Para lograr esto, desarrollamos diagramas y modelos que nos proporcionan una comprensión clara de lo que se necesita.

Considerando los atributos especificados en las consignas, creamos el **Diagrama Entidad-Relación**, que nos servirá como guía para establecer una base de datos comprensible y completa para nuestro propósito, eliminando datos innecesarios.

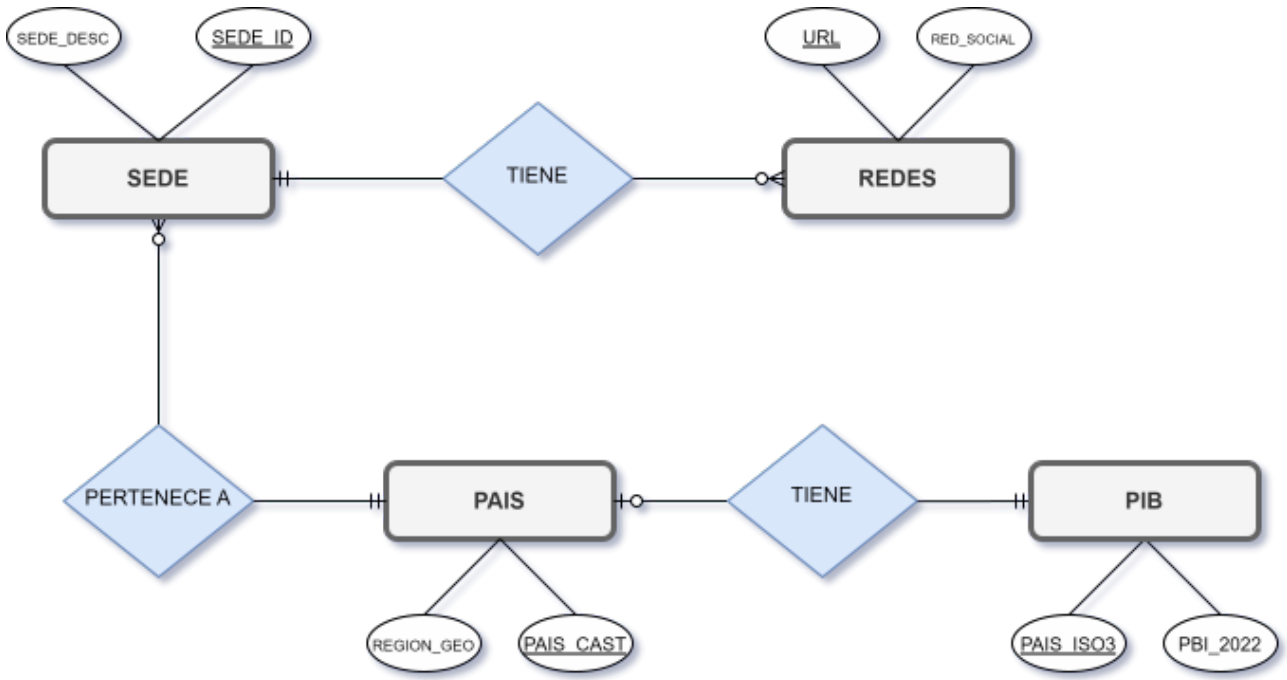

Figura 1: DER

En el proceso de diseño del DER, en primer lugar identificamos las entidades **pais**, **sede**, **redes** y **pib**. Luego, evaluamos qué atributos son los que describen a cada entidad.

- **pais**: *region_geografica*: la región geográfica en la que se encuentra el país.
  pais_cast: el nombre completo del país en castellano.
- **sede**: *sede_desc*: la descripción de la sección de la sede.
  sede_id: el identificador de 5 letras de la sede.
- **redes**: *url*: el link a la cuenta de alguna red social de una sede.
  red_social: el nombre de la red social que utiliza la sede.
- **pib**: *pais_iso_3*: el identificador de 3 letras del país.
  pib_2022: el valor del PIB per cápita del país en dólares estadounidenses en el año 2022.

El proximo paso es distinguir las claves de cada entidad:

- Para la entidad **pais** elegimos *pais_cast*, ya que todos los paises tienen nombres diferentes.
- Para la entidad **sede** elegimos *sede_id*, ya que es el identificador que creó el gobierno para diferenciar las sedes.
- Para la entidad **redes** elegimos *url*, ya que a partir de este podemos deducir en qué red social se encuentra la página (porque dentro de los links está el nombre de la red social)
- Para la entidad **pib** elegimos **pais_iso_3**, ya que es el código identificador de los países, es decir a cada pais le corresponde un código distinto.

Por último, analizamos las relaciones entre las entidades.

- **sede *pertenece a* pais**: Esta relacion surge de ver a qué país pertenece una sede. Es una relación de uno a muchos ya que cada país puede tener muchas sedes argentinas o ninguna, mientras que a cada sede le corresponde exactamente un país.
- **sede *tiene* redes**: Esta relación se da porque hay sedes que tienen registradas redes sociales. Es una relación de uno a muchos ya que una sede puede no tener redes o puede tener una o varias. Por otra parte, cada una de las páginas en las redes le corresponde a una sola sede.
- **pais *tiene* pib**: Esta relación ocurre porque cada país tiene un y solo un código de país y valor de pib. Por otro lado, un pib corresponde a un o ningún pais ya que hay datos en **pib** que no están en **pais**.

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

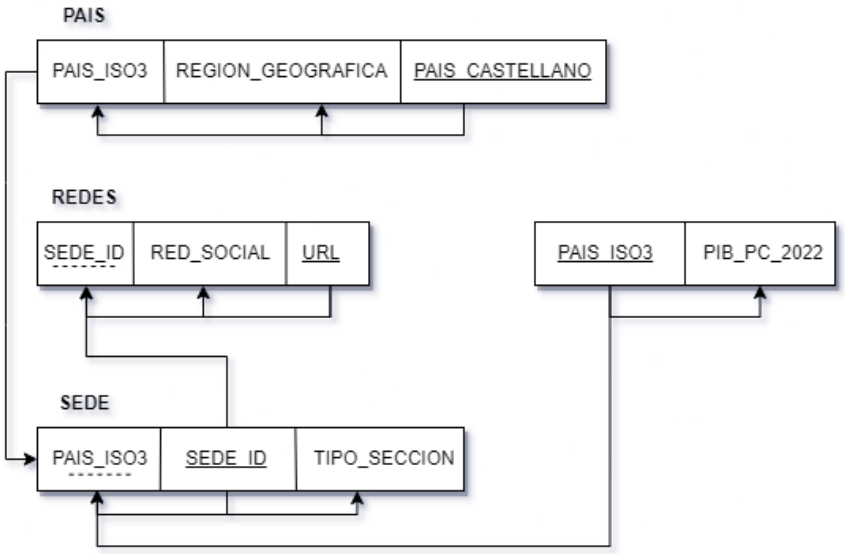

Figura 2: Modelo Relacional

Análisis de la normalización de nuestro modelo para una mejora en el rendimiento de las consultas:

- **1FN**: Nuestro modelo esta en 1FN si todos los atributos de la tabla son atómicos. En nuestro caso, teníamos el *url* de las sedes como atributos no atómicos, pero luego de la limpieza de **redes**, se puede decir que **sede**, **pais**, **redes** y **pib** no tienen atributos atómicos, por lo que están en Primera Forma Normal.
- **2FN**: Nuestro modelo esta en 2FN si esta en 1FN y además todo atributo (no clave) depende de manera completa a la clave primaria. En nuestro caso, podemos ver que todos los atributos de **pais** dependen completamente de *pais_castellano*. Lo mismo ocurre en **redes**, **sede** y **pib** con el *url*, *sede_id* y **pais_iso_3** respectivamente. Entonces, se puede decir que están en Segunda Forma Normal.
- **3FN**: nuestro modelo esta en 3FN si esta en 2FN y además ningún atributo depende transitivamente de la clave primaria. En nuestro caso podemos ver que no hay ninguna dependencia transitiva por lo que está en Tercera Forma Normal.

Para mejorar la calidad de nuestras bases de datos, vamos a utilizar la técnica llamada **GQM** (Goal Question Metrics) en la que se define un objetivo, se plantea una o más preguntas cuya respuesta permitirá saber si se satisface el objetivo, y se formula una o más métricas para cada una de las preguntas, cuya ejecución permitirá responderlas.

- **sede_completo:**

  - ***1. Objetivo:*** Limpiar la base de datos de URLs, eliminando aquellos que no son links, es decir, que si los pongo en el buscador no me llevan directamente a la página de la sede.

  - ***2. Pregunta:*** ¿Cuántos URLs no te llevan de forma directa a la página de la sede?

  - ***3. Metrica:*** Contamos la proporción de URLs que no son links. Son 37 de 272, es decir el 13,6 %.

  - ***4. Criterios de corrección:*** Eliminación de las filas que no representan links, lo que mejora la calidad de los datos al quedar un formato claro de registros. El atributo de calidad afectado sería la precisión o exactitud de los datos de las URLs.

- **sede_secciones:**

  - ***1. Objetivo:*** Identificar y eliminar los registros que no tiene asignado un sede_id.

  - ***2. Pregunta:*** ¿Cuantos datos de Sede_id estan vacíos?

  - ***3. Metrica:*** Calculamos la proporción de registros sin un sede_id asignado. El total seria 1 de 164 o el 0,006 %.

  - ***4. Criterios de corrección:*** Eliminación de la fila sin el sede_id asignado, ya que sin esta información, el dato no puede ser utilizado adecuadamente en análisis posteriores. El atributo de calidad afectado sería la integridad de los datos. La presencia de registros sin el atributo sede_id afecta a la base de datos, ya que esos registros estan en falta de información esencial.

- **pibpercapita:** En esta base de datos decidimos dejar los datos de pib en el 2022 con NULL, por lo tanto, no realizamos ninguna limpieza de datos.

Comentario: Ambos problemas de calidad corresponden a la instancia, ya que son errores como datos específicos o individuales dentro de una base de datos en el modelo de datos, pero que no cumplen con las expectativas o estándares de calidad.

Para construir los DataFrames utilizados en el proyecto, se extrajeron todos los datos necesarios de las fuentes proporcionadas. Aquí está la relación detallada de las fuentes de datos para cada DataFrame:

- **sede**: Los datos que conforman este DataFrame se obtuvieron de los archivos sede-secciones y *sede-completo.csv*.
- **pais**: Se recopilaron datos del archivo *sede-completo.csv*.
- **redes**: Los datos para este DataFrame se tomaron exclusivamente del archivo *sede-completo.csv*.
- **pib**: Se tomaron los datos del archivo *pibpercapita.csv*

---

4. Decisiones tomadas

Durante el avance del diseño de las tablas que utilizamos como información, nos encontramos con diferentes situaciones que requerían la toma de decisiones. Dejamos detalladas las mismas:

- Se toman como Redes Sociales: Facebook, Twitter, Instagram, Youtube, Linkedin y Flickr.
- Limpiamos los links que no te llevan directamente a la página. Había links que únicamente tenían el nombre o comenzaban con "@".
- En un principio utilizamos el atributo *sede_seccion* para tener el nombre de la seccion de una sede. Luego nos dimos cuenta que esa informacion estaba dada en *sede_descripción*.
- Decidimos no borrar los países que no tienen registrado su PIB Per Capita, ya que hay algunos, como Venezuela, en los que Argentina tiene sedes.
- En **sede** decidimos dejar las sedes que no tienen registrada la descripción de las secciones, dejando en la tabla el id_sede con descripción NULL. Esta decisión se debe a que dentro de los ejercicios nos piden realizar consultas en las que debemos contar la cantidad de estas, por lo que necesitamos tenerlas todas registradas —tengan o no descripción de sus secciones— para obtener datos precisos.

- Al comienzo del proceso de realizar el trabajo práctico, habíamos tomado *pib_2022* como un atributo de **pais**, pero al momento de graficar la relación entre el PIB per cápita y la cantidad de sedes argentinas nos dimos cuenta que no estaban en el gráfico los países sin sedes; esto se debe a que los países los estábamos sacando de la base de datos *sede_completo*, que solo tiene aquellos con sedes argentinas. Una solución a esto es obtener el identificador de país de la base *pibpercapita*, pero esto nos deja con los valores de *region_geografica* y *pais_castellano* en NULL para aquellos países sin sedes argentinas. Para evitar tener una base de datos con tantos valores en NULL, decidimos separar **pib** como una nueva entidad con clave primaria *pais_iso_3* y convertirla en clave foranea de **pais**

---

5. Análisis de datos

En esta seccion debemos conseguir distintos DataFrames dependiendo la consigna dada.

1. **Para cada país informar cantidad de sedes, cantidad de secciones en promedio que poseen sus sedes y el pib per cápita del país en 2022. El orden del reporte debe respetar la cantidad de sedes (de manera descendente). En caso de empate, ordenar alfabéticamente por nombre de país.**

   Obtuvimos la siguiente tabla:

Índice	pais_iso_3	sedes	secciones_promedio	pib_pc_2022
0	BRA	11	2.18	8917.6
1	USA	9	3.88	76329.5
2	URY	8	1.37	20975.04
3	BOL	7	2.85	3600.12
4	CHL	7	2.85	15355.48
...	...	...	...	...
239	WSM	0	0	3745.56
240	XKX	0	0	5340.27
241	YEM	0	0	650.272
242	ZMB	0	0	1456.9
243	ZWE	0	0	1676.82

Tabla 1: Tabla resultante del ejercicio I

2. **Reportar agrupando por región geográfica: a) la cantidad de países en que Argentina tiene al menos una sede y b) el promedio del pib per cápita 2022 de dichos países. Ordenar por el promedio del pib per Cápita.**

   Nos queda la siguiente tabla:

Índice	region_geografica	paises_con_sedes_arg	promedio_pib_per_capita
0	OCEANÍA	2	56759.2
1	EUROPA OCCIDENTAL	16	52978.1
2	AMÉRICA DEL NORTE	3	47581.3
3	ASIA	23	23375.3
4	EUROPA CENTRAL Y ORIENTAL	8	15425.6
5	AMÉRICA CENTRAL Y CARIBE	14	13722.9
6	AMÉRICA DEL SUR	11	9447.21
7	ÁFRICA DEL NORTE Y CERCANO ORIENTE	5	4508.71
8	ÁFRICA SUBSAHARIANA	7	2459.07

Tabla 2: Tabla resultante del ejercicio II

3. **Para saber cuál es la vía de comunicación de las sedes en cada país, nos hacemos la siguiente pregunta: ¿Cuán variado es, en cada el país, el tipo de redes sociales que utilizan las sedes?**

Indice	nombre_pais	redes_distintas
0	REPÚBLICA DE ARMENIA	3
1	REINO DE ESPAÑA	4
2	REPÚBLICA CHECA	2
3	UCRANIA	2
4	REPÚBLICA DE FINLANDIA	2
...	...	...
67	REPÚBLICA HELÉNICA	2
68	REPÚBLICA TUNECINA	4
69	REPÚBLICA DE SUDÁFRICA	3
70	ESTADO PLURINACIONAL DE BOLIVIA	1
71	REPÚBLICA DE COSTA RICA	1

Tabla 3: Tabla resultante del ejercicio III

4. **Confeccionar un reporte con la información de redes sociales, donde se indique para cada caso: el país, la sede, el tipo de red social y url utilizada. Ordenar de manera ascendente por nombre de país, sede, tipo de red y finalmente por url.**

Índice	nombre_pais	sede_id	red_social	URL
0	AUSTRALIA	CSIDN	Facebook	.../ArgentinaEnSidney/
1	AUSTRALIA	EAUST	Facebook	.../ArgentinaEnAustralia/
2	AUSTRALIA	EAUST	Twitter	.../ARGinAustralia/
3	BARBADOS	EBARB	Facebook	.../ArgentinaEnBarbados/
4	CANADÁ	CTORO	Facebook	.../ArgentinaEnToronto/
...	...	...	...	...
161	SANTA SEDE	ESSED	Twitter	.../ArgSantaSede
162	SANTA SEDE	ESSED	Youtube	.../EmbajadaArgentinaantelaSantaSede
163	UCRANIA	EUCRA	Facebook	.../ArgentinaEnUcrania
164	UCRANIA	EUCRA	Instagram	.../argenucrania/
165	JAPÓN	EJAPO	Facebook	.../ArgJapon/?ref=bookmarks

Tabla 4: DataFrame del ejercicio IV

(a) *Aclaración: Los puntos suspensivos de URL es la pagina de la Red social, por ejemplo si la red es Facebook el URL empezaria con https://www.facebook.com*

Para poder analizar mejor estos datos, realizamos unos gráficos que nos ayudan a visualizar mejor la información. Fuimos siguiendo las consignas dadas:

1. **Cantidad de sedes por región geográfica. Mostrarlos ordenados de manera decreciente por dicha cantidad.**

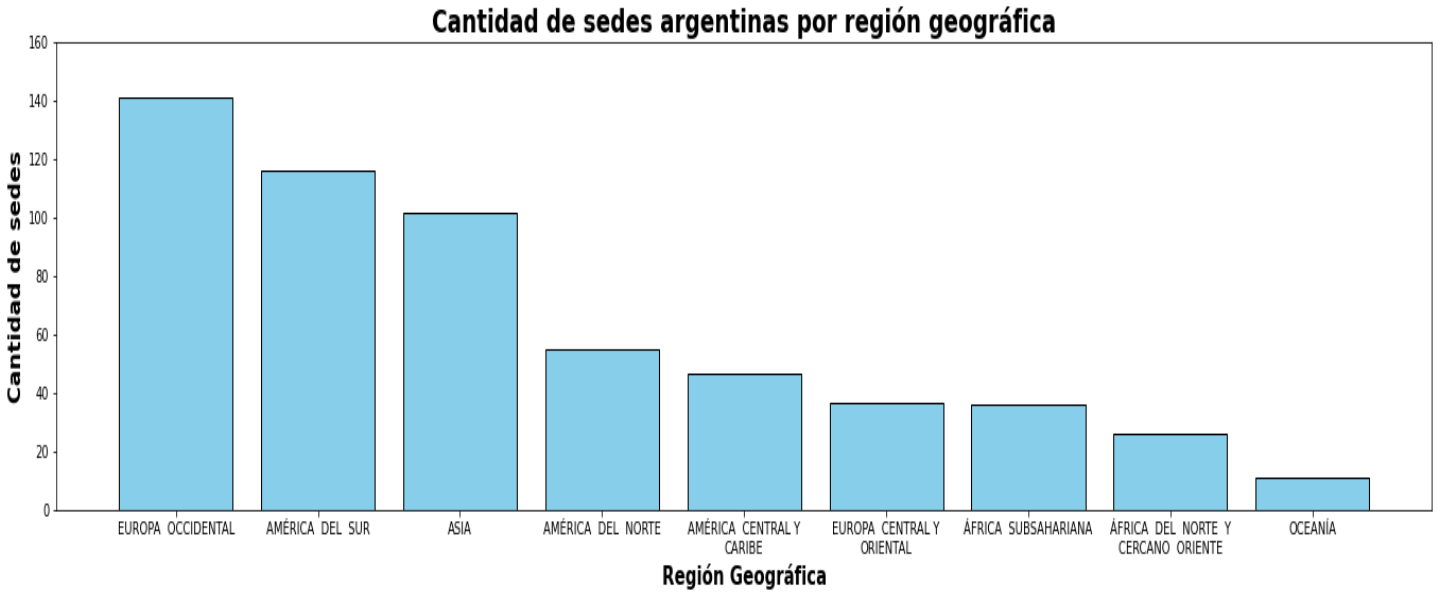

Figura 3: Cantidad de sedes por region

2. **Boxplot, por cada región geográfica, del pib per cápita 2022 de los países donde Argentina tiene una delegación. Mostrar todos los boxplots en una misma figura, ordenados por la mediana de cada región**

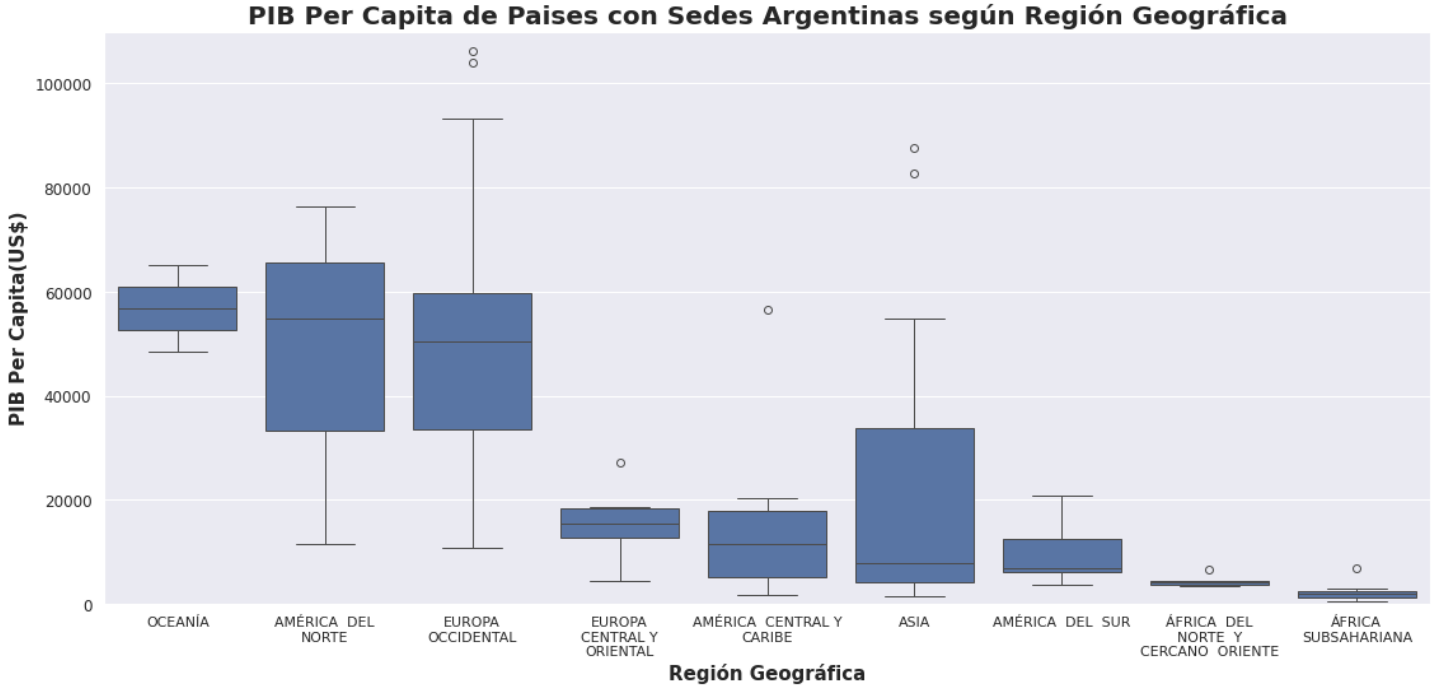

Figura 4: pib por Region Geografica

3. **Relación entre el pib per cápita de cada país (año 2022 y para todos los países que se tiene información) y la cantidad de sedes en el exterior que tiene Argentina en esos países.**

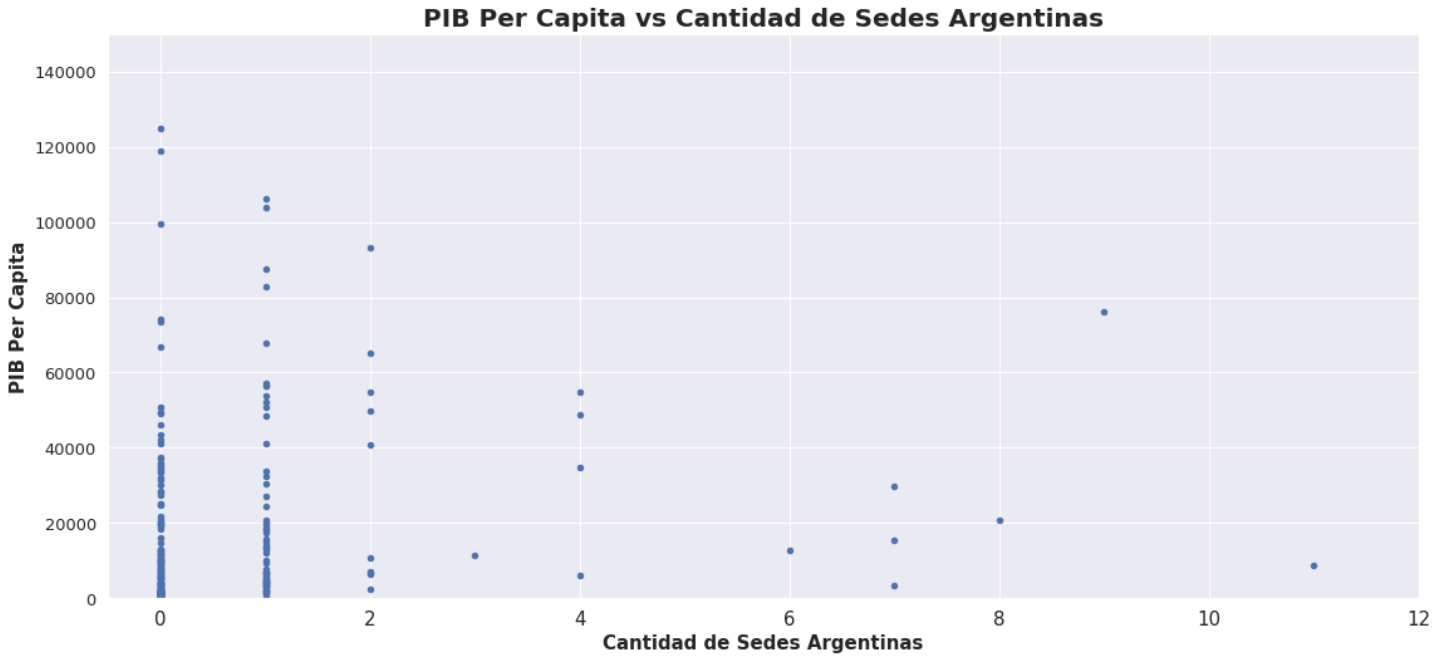

Figura 5: Relación entre pib y cantidad de sedes

---

6. Conclusiones

En general, los gráficos y tablas proporcionan una representación clara y efectiva de los resultados obtenidos en el análisis de datos, permitiendo una comprensión más profunda de la distribución de sedes, el promedio de secciones por país, la diversidad de redes sociales utilizadas y el PIB per cápita en diferentes regiones geográficas.

Una de las conclusiones que entendemos a partir de la Figura 3 es que la región geográfica con mayor cantidad de sedes argentinas es Europa Occidental. Esto se explica ya que la ascendencia de gran parte de la población de nuestro país es de esta región, por lo que hay gran cantidad de gente que necesita realizar trámites de documentación tanto para viajar como para emigrar. América del Sur es el segundo en cantidad de sedes por cercanía geográfica y buenas relaciones diplomáticas con los países de la zona. En tercer lugar, se encuentra Asia por los numerosos tratados económicos que tenemos con los países de aquella región.

Por otra parte, la Figura 4 nos muestra que la región geográfica con mayor distibución de PIB per cápita es Europa Occidental, extendiéndose desde poco más de 10000 US$ hasta poco menos de 100.000US$ sin contar los 2 outliers (Noruega con 106.000US$ e Irlanda con 103.000US$). Por otro lado, Asia también cuenta con una distribución diversa, ya que allí hay tanto países muy pobres (como Bangladesh) como muy ricos (Qatar y Singapur). No analizamos detalladamente Oceanía y América del Norte ya que tienen 2 y 3 países con sedes argentinas respectivamente.

Por último, en la Figura 5 podemos ver que la mayoría de los países no tienen sedes Argentinas o tienen una única sede. Tan solo Brasil supera las 10 sedes. No encontramos ninguna relacion entre el PIB per cápita y la cantidad de sedes Argentinas.