Análisis de la relación entre PBI de un país y sedes Argentinas

Resumen

En este informe nos propusimos entender la relación entre el PBI per cápita de un país y la cantidad de sedes que Argentina tiene en dicho país. Para ello hemos recopilado tablas que contengan datos acerca del PBI de los países del mundo como también información sobre ellos, por ejemplo la región a la que pertenecen, y también información acerca de las sedes que Argentina tiene en el mundo. A partir de éstas tablas hemos hecho una limpieza de los atributos que consideramos clave a la hora de analizar esta relación, hemos generado un DER, creado dataframes vacíos que corresponden al DER, y por último hemos importado los datos que consideramos relevantes.
A partir de estos datos, generamos reportes que muestran distintas relaciones entre los países, el PBI, su región y las redes que utilizan para su comunicación como también gráficos que nos ayuden a visualizar y entender la información que hemos recopilado y limpiado.
Estas herramientas nos han ayudado a analizar la información. Hemos llegado a la conclusión que a pesar que el PBI per cápita es un factor importante a la hora de analizar la cantidad de sedes de un país, esta relación no es lineal y no es el único factor relevante para entender la relación entre la cantidad de sedes que tiene un país.

Introducción

En este trabajo hemos planteado como objetivo analizar la relación entre el PBI per cápita de un país y la cantidad de sedes (embajadas, consulados, etc.) que Argentina tiene en dicho país. Queremos saber si es que esta relación existe y, en caso que así sea, observar cómo se comporta. Para alcanzar este objetivo, deberemos estudiar 6 tablas de dos bases de datos, ver qué información contienen y qué vamos a necesitar de ellas.Luego, limpiar los datos que nos sean necesarios, armar un DER y, a partir del modelo, crear nuestros propios dataframes con los que vamos a trabajar. Nuestros dataframes estarán en 3ra forma normal y la calidad de los datos será mayor a la presente en las tablas originales. Realizaremos diferentes reportes utilizando SQL y armaremos gráficos con el propósito de poder entender mejor la relación entre el PBI y las sedes, a partir de los cuales llegaremos a una conclusión. A continuación profundizaremos en las etapas que iremos recorriendo en este trabajo, como también nuestras reflexiones y decisiones que tomamos para avanzar hacia nuestro objetivo.

Procesamiento de Datos

Dejamos fuera la tabla “Metadata-Indicator” ya que no nos brindaba información relevante y analizamos el resto de las tablas. Comenzando con las tablas sobre las sedes de Argentina,

observamos que la tabla “lista-sedes” se encuentra en 2FN. Esto se debe a que cumple con los requisitos de la segunda forma normal, pero el atributo *pais_castellano* depende de *pais_iso*, donde *pais_iso* no es superclave y *pais_castellano* no es un atributo primo, por lo tanto la tabla no está en 3FN. De esta tabla tomamos los atributos *sede_id*, *pais_iso_3*, y *ciudad_castellano*. A su vez, hemos decidido utilizar aquellas filas cuyo valor del atributo *estado* sea “Activo”. No hemos encontrado problemas en la calidad de datos de esta tabla. Realizamos un GQM para ver la cantidad de NULLs.

Goal: Que los atributos *pais_iso_3* y *ciudad_castellano* estén completos.

Question: ¿Cuál es la proporción de NULLs con respecto a cada uno de estos atributos?

Metric: 0/164 = 0 y 0/164 = 0.

La tabla lista-sedes-datos no cumple 1ra forma normal debido a que el atributo *redes_sociales* tiene más de un valor por fila, por lo tanto los datos no son atómicos. De esta tabla tomamos los atributos *sede_id* y *redes_sociales*. La causa de este problema es de instancia ya que no se cumple la consistencia de los valores provenientes de distintas fuentes. Realizamos un GQM para analizar el problema:

Goal: Que los valores del atributo redes_sociales sean atómicos

Question: ¿Cuántos valores del atributo redes sociales no son atómicos respecto al total de filas?

Metric: 81/164 = 0.494

Para poder extraer los datos de redes sociales de lista-sedes-datos utilizamos funciones en python para convertir lo que era un string separados por espacios vacíos a las diferentes cuentas que usaba cada sede y pusimos los valores dentro de una lista, filtramos las ocurrencias de cadenas que no tenian informacion de la cuenta, sino que eran o espacios vacíos o ‘//’.
Por último removimos los valores faltantes y nos quedamos con registros de las cuentas en formato de lista con una función .explode() de pandas que transforma cada elemento de una lista en una fila, replicando los valores de los índices.
Posteriormente nos quedamos solamente con los atributos ‘sede_id’ y ‘redes_sociales’ de la tabla original y con funciones lambda de pandas construimos la columna ‘Nombre_red’ a partir de ‘Url’, parseando la url y capitalizando los nombres de cada valor.
Luego de realizar este procedimiento para limpiar los datos, al volver a realizar la misma métrica, los datos que no son atómicos son 0.

De la tabla lista-secciones tomamos solamente los atributos *id_sede* y *tipo_seccion*. No tomamos la descripción de las secciones puesto que había múltiples problemas en las instancias de este atributo, como la misma información escrita de formas diferentes, uso indistinto de género femenino y masculino, entre otros. Además, suponemos que una sede puede tener varias secciones, pero serán de diferentes tipos, por lo que tomar la sede y el tipo de sección debería alcanzar para individualizar todas las secciones. No vemos problemas de calidad de datos en los atributos elegidos.
Esta tabla se encuentra en 3FN puesto que sólo observamos dependencias funcionales del conjunto {id_sede, descripcion_castellano} a cada uno de los demás atributos. No podemos encontrar DF de descripcion_castellano a descripcion_ingles, por ejemplo, por las inconsistencias entre estas columnas. Y notamos que {id_sede, descripcion_castellano} es CK.

La tabla “(...)GDP(...)” donde se encuentran los PBI por año de los países está en 3FN. De esta tabla tomamos los atributos *Country Name*, *Country Code* y *2022*. En cuanto a la calidad de sus datos vemos que la causa del problema es de instancia, ya que el PBI de cada país cambia año a año, pero esto no siempre fue actualizado en la tabla (hay países donde hay información correspondiente al año 2021 pero no al 2022), que genera que no se cumple la calidad de vigencia del atributo. Realizamos un GQM para analizar el problema:

Goal: Que la información del PBI esté vigente al año que necesitamos (2022)

Question: ¿Cuál es la proporción de datos vigentes (tienen información sobre el PBI 2022) respecto al total?

Metric: 244/266 = 0.917

Como solución a este problema, debido a que la cantidad de países que no tienen el dato del PBI en 2022 son pocos respecto al total de países, decidimos eliminar éstos países ya que no nos sirven para lograr el objetivo del trabajo. Luego de realizar esto, la nueva métrica respecto a la cantidad de datos vigentes respecto del total es 244/244 = 1.

La tabla “Metadata_country(...)” se encuentra en 3FN. De esta tabla tomamos los atributos *Country Code* y *Region*. En sus datos observamos que el atributo *Region* no está completo. Revisamos los motivos de esta incompletitud, y vimos en la columna tableName (nombres de los países) que las filas con atributo Region nulo corresponden a grupos de países o a otros tipos de información, y no a un país. Vemos entonces un problema de consistencia en el atributo Country Code, ya que aquellos códigos no son de países como indica el nombre. Al mismo tiempo, en las mismas filas donde se produce este problema hay uno de completitud en el atributo Region. Entendemos que este es un problema de instancia ya que se incluye valores distintos a países en un atributo de país, y por otro lado quedan lugares incompletos de un atributo.

Planteamos entonces el siguiente GQM:

Goal: que el atributo Region esté completo.

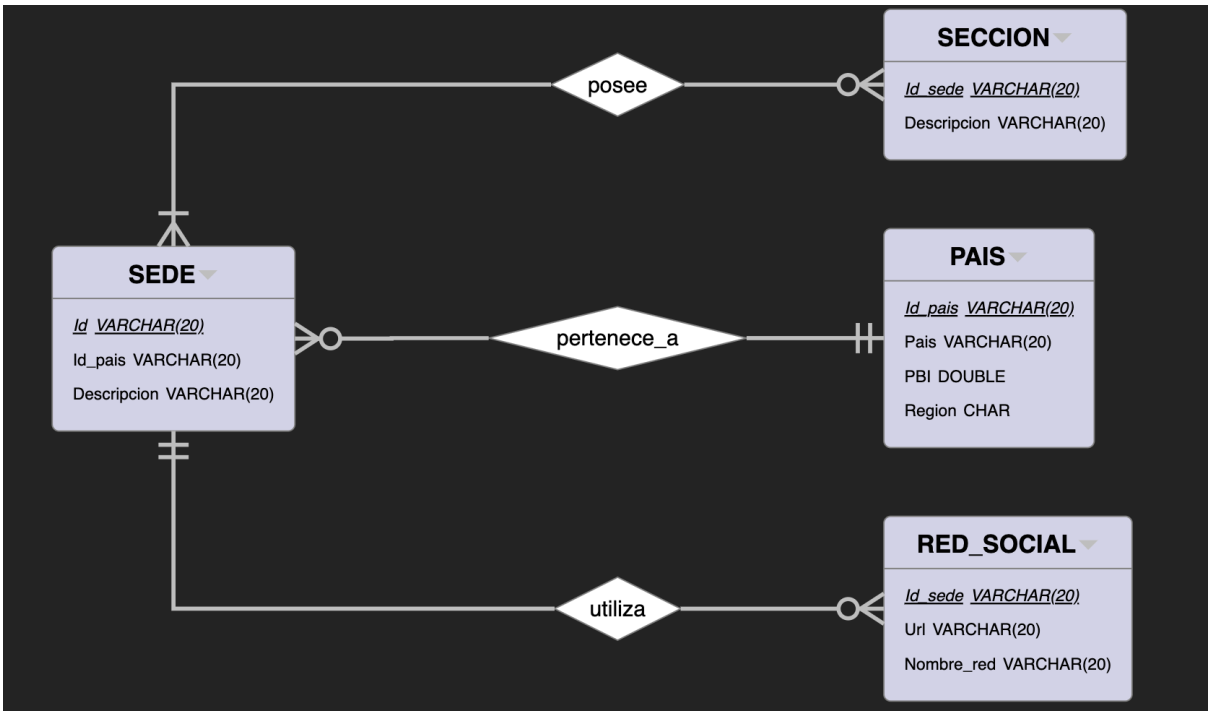
Question: ¿Cuál es la proporción de NULL en el atributo Region con respecto a las filas totales?

Metric: 48/265 = 0.181

Esto lo corregimos eliminando las filas que tienen NULL en Region, ya que al no corresponder a países no nos dan información útil para nuestros objetivos. Luego de eliminarlos, la cantidad de NULLs sobre el total de filas es 0/217 = 0. Verificamos con una consulta SQL que en todos los casos de Region de valor NULL, en la columna de país no había un nombre de país.

Construcción del DER

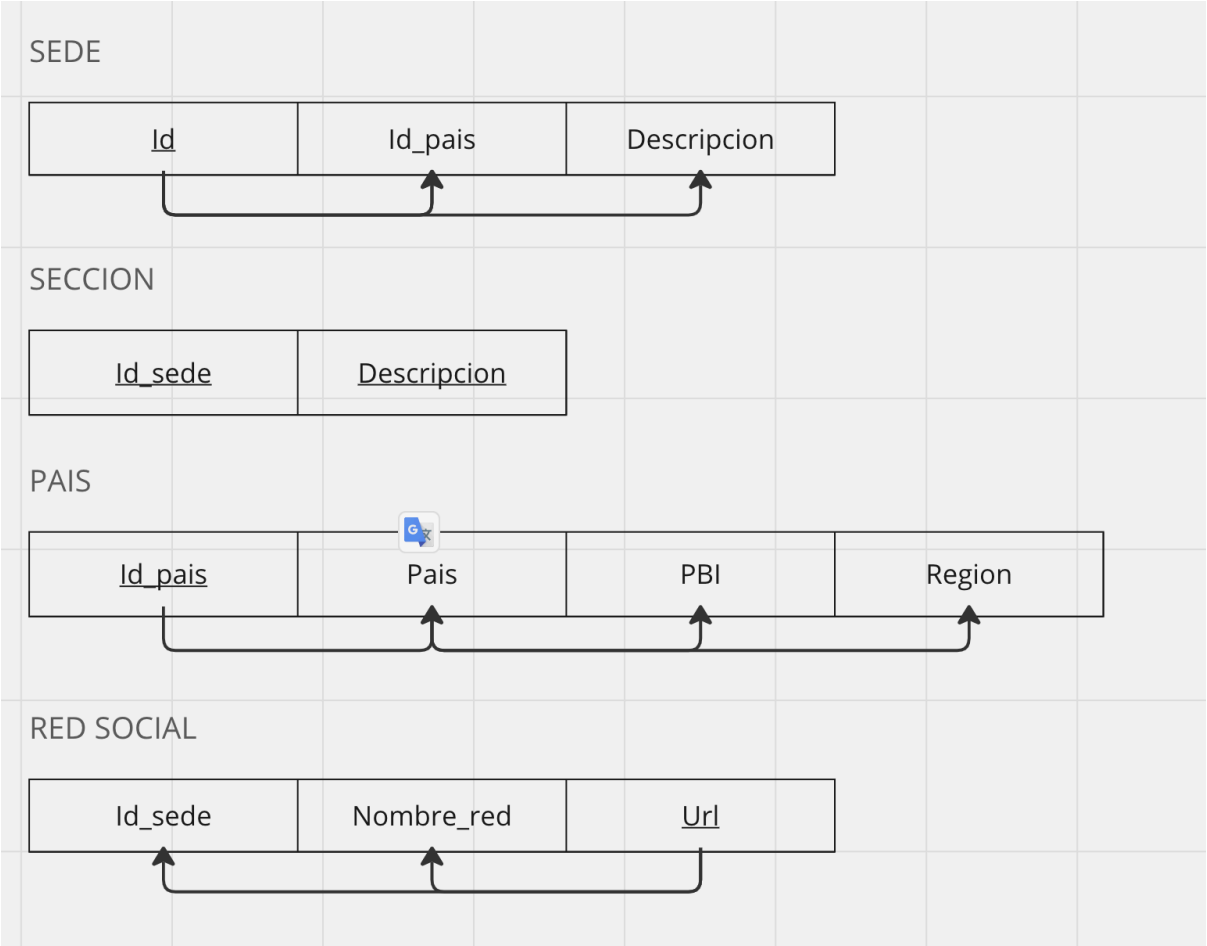
Definimos 4 entidades: sede, país, región y cuenta. La entidad cuenta tiene una url, que será clave porque identifica unívocamente la cuenta, y tiene el nombre de la red social correspondiente porque vamos a necesitar esa información. La entidad sección tiene la descripción, que corresponderá al tipo de sección de la tabla lista-secciones, como único atributo. Luego la sección se identificará de acuerdo a la sede donde se encuentra, ya que en la realidad cada sección es de una sede. La entidad país tiene iso, que será la clave, un nombre para describirlo, un PBI ya que cada país tiene el PBI 2022 y es único (o no lo tiene). El país tiene también el atributo región, ya que pertenece a una y sólo una. Por último, la entidad SEDE tiene id, que será la clave que la identifica, y una descripción para que sepamos a qué corresponde el id.



Entendemos que una sede corresponde a uno y sólo un país, mientras que un país puede tener varias o ninguna sede. De hecho, ambas cosas se verifican para los datos con los que trabajamos. Hay países con varias sedes, una o ninguna. Asimismo, cada sede puede tener varias cuentas o, aunque no lo comprobamos, en principio podría no tener ninguna. Por otro lado, una cuenta pertenece a una y sólo una sede, porque debe existir una sede que arme la cuenta y porque, teniendo cada cuenta información de alguna sede, debería ser únicamente de ésta. Por otro lado, cada sede puede tener ninguna o varias secciones (interpretamos a los efectos de este trabajo que una sede que no aparece en la tabla lista-secciones tiene 0 secciones). En cambio, una sección debe estar en por lo menos una sede, para que la existencia de esa sección tenga sentido, y puede haber varias sedes que tengan un mismo tipo de sección.

Modelo relacional y formas normales

Los atributos subrayados forman las PK.



Claves foráneas
iso_pais de PAÍS (iso) a SEDE; id_sede de SEDE (id) a CUENTA; id_sede de SEDE (id) a SECCIÓN.

Dependencias funcionales (primero un conjunto minimal y luego las que se deducen):

- SEDE

id → descripción, descripción → Id_pais.
Se deduce: id → Id_pais.
Está en 3FN porque id y descripción son CK.

- CUENTA

Url → Id_sede, Url → Nombre_red.
No se deducen otras dependencias.
Está en 3FN porque Url es CK.

- SECCIÓN

No tiene dependencias funcionales, luego está en 3FN.

- PAÍS

Id → Nombre, Nombre → PBI, Nombre → Region.
Se deduce: Id→ PBI, Id → Region.

Está en 3FN porque Idy Nombre son CK.

De dónde se importa los atributos

De la tabla Metadata-Country tomamos el iso_3 de los países como Id, y tomamos el nombre y región de los países. De ahí que los países y regiones estarán nombrados en inglés. Elegimos el iso_3 en lugar del iso_2 porque tiene la ventaja de que también es el iso que identifica a los países en las tablas de representaciones argentinas. Además, tomamos los iso y nombres de Metadata-Country porque esta se puede considerar la lista completa de países, a diferencia de las tablas de representaciones argentinas donde sólo aparecerán los países que tienen al menos una sede. De la tabla GDP tomamos el atributo 2022 de los países, en los casos en que existía, como atributo PBI.

De la tabla lista-sedes tomamos los Id de las sedes, y tomamos como atributo Descripción el atributo sede_desc_castellano.

De la tabla lista-sedes-datos seleccionamos el atributo redes_sociales como url, previa separación de los grupos de links que aparecían en cada fila. Luego de separar las cuentas e importarlas al atributo url, procedimos a eliminar las filas que no tenían una url. El criterio para hacerlo fue eliminar toda fila donde el atributo url no contuviera '@' ni '.com'. El atributo red_social (la red a la que pertenece la cuenta) fue deducido de las url.

De la tabla lista-secciones elegimos el atributo id, y como descripción seleccionamos el atributo tipo_seccion para tener menos problemas de calidad de datos, y porque interpretamos que cada sección estará identificada unívocamente por el tipo de sección y la sede a la que pertenece. De hecho, lo comprobamos mediante una consulta SQL y sólo hay tres repeticiones entre más de 500 pares, por lo que eliminamos estas y entendemos que los resultados no se verán afectados en mayor medida.

Análisis de Datos

Reportes generados con SQL:

1. Tabla donde se puede observar la cantidad de sedes, promedio de la cantidad de secciones y el PBI por país, ordenado de manera descendente por cantidad de sedes.

	Pais	sedes	secciones_promedio	PBI
0	Brazil	11.0	1.636364	8917.674911
1	United States	9.0	3.333333	76329.582265
2	Uruguay	8.0	0.500000	20795.042354
3	Bolivia	7.0	2.142857	3600.121635
4	Chile	7.0	2.000000	15355.479740
...	...	...	...	...
191	Uzbekistan	0.0	0.000000	2255.151155
192	Vanuatu	0.0	0.000000	3231.351300
193	Yemen, Rep.	0.0	0.000000	650.272218
194	Zambia	0.0	0.000000	1456.901570
195	Zimbabwe	0.0	0.000000	1676.821489

196 rows × 4 columns

Se puede observar que los países donde Argentina tiene más de una sede, son países limítrofe o cercanos geográficamente, (como es el caso de Brasil, Chile, Uruguay, Bolivia, etc.) países con una gran relación comercial con Argentina, como es el caso de Estados Unidos y China, países con una importante relación cultural, como España, Italia e Israel, además de un alto PBI como es el caso de estos últimos junto a Francia, Suiza, etc.

2. Tabla donde se puede observar el PBI per cápita y la cantidad de países que tienen sedes respecto a cada región.

	Region	cantidad_paises_con_sede	PBI_promedio_U$S
0	North America	2	65623.622394
1	Europe & Central Asia	26	37916.702419
2	East Asia & Pacific	11	27876.872317
3	Middle East & North Africa	12	24902.518923
4	Latin America & Caribbean	24	11877.417324
5	Sub-Saharan Africa	7	2459.067444
6	South Asia	3	2229.357783

Podemos observar que a pesar que hay cierta relación entre el promedio del PBI por región y la cantidad de países con sede (notar que en North America solo hay dos países) donde cuanto más alto el PBI más países con sedes tiene Argentina, esta relación no es lineal ya que en Latin America & Caribbean se encuentran la mayor cantidad de países donde Argentina tiene sede. Entendemos que esto se debe a la cercanía geográfica, lazos económicos y culturales.

3. Tabla donde se pueden observar las vías de comunicación de las sedes en cada país.

	Pais	cant_de_redes_diferentes
0	Colombia	2
1	Barbados	1
2	Kenya	2
3	Switzerland	4
4	Pakistan	1
...	...	...
70	Jamaica	2
71	Mozambique	1
72	Nigeria	2
73	Azerbaijan	3
74	Israel	2

75 rows × 2 columns

Observamos que no hay un consenso sobre las redes sociales que deben usarse como vías de comunicación que deben tener las sedes sino que se usan según disponga cada sede, puesto que las diferencias en la cantidad de redes indican que algunas se usan en algunos países y en otros no.

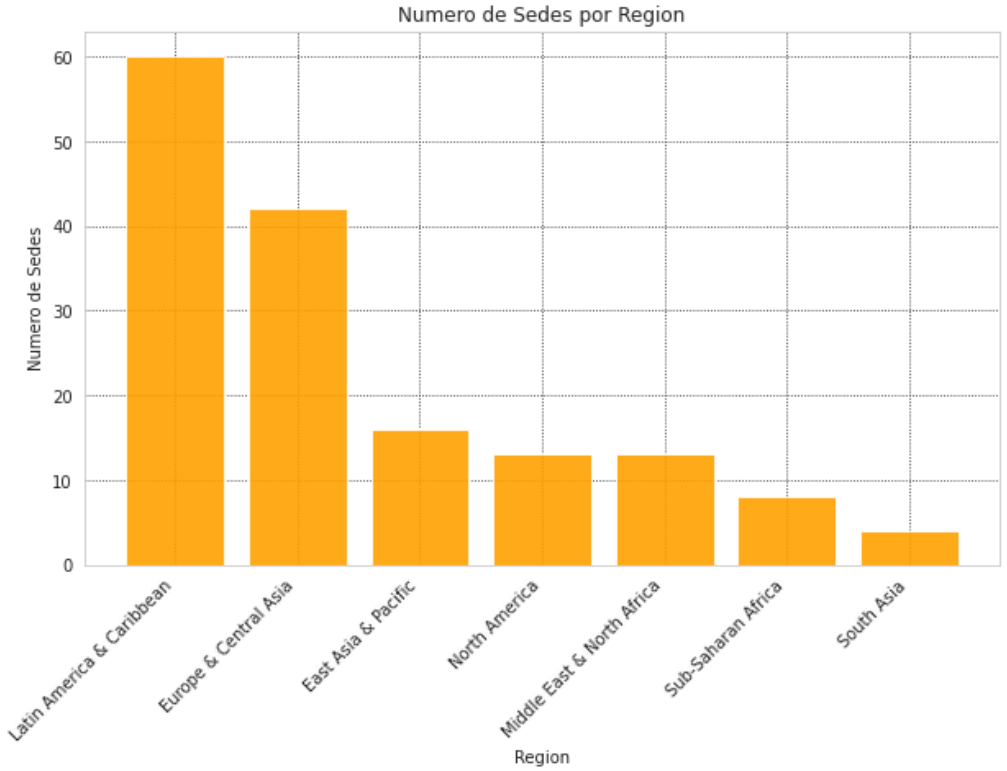
4. Tabla que muestra qué red social usan todas las sedes de cada país como también su URL.

	Pais	Sede	Red_Social	URL
0	Algeria	EARGE	www.facebook.com	facebook.com/ArgentinaEnArgelia
1	Algeria	EARGE	www.instagram.com	https://instagram.com/argenargelia
2	Algeria	EARGE	www.twitter.com	https://twitter.com/ARGenArgelia
3	Angola	EANGO	www.facebook.com	https://www.facebook.com/ArgentinaEnAngola/
4	Angola	EANGO	www.instagram.com	https://www.instagram.com/embargentinaenangola/
...	...	...	...	...
248	Venezuela, RB	EVENE	www.instagram.com	https://www.instagram.com/argenvenezuela/
249	Venezuela, RB	EVENE	www.twitter.com	https://twitter.com/argenvenezuela?lang=es
250	Viet Nam	EVIET	www.facebook.com	https://www.facebook.com/ArgentinaEnVietnam/
251	Viet Nam	EVIET	www.instagram.com	https://www.instagram.com/argenvietnam/
252	West Bank and Gaza	REPAL	www.facebook.com	https://www.facebook.com/ArgEnPalestina

Observamos que más allá del país, las redes más utilizadas son Twitter, Instagram y Facebook, aunque también algunas sedes utilizan YouTube o incluso Flickr.

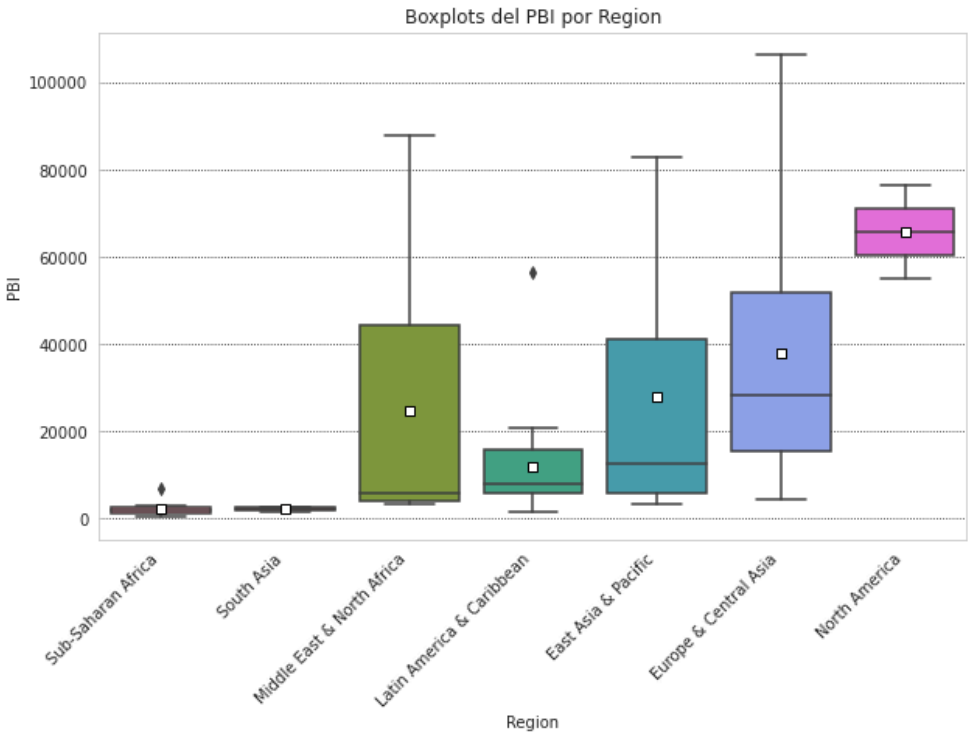
Gráficos:

1. Cantidad de sedes por región geográfica.



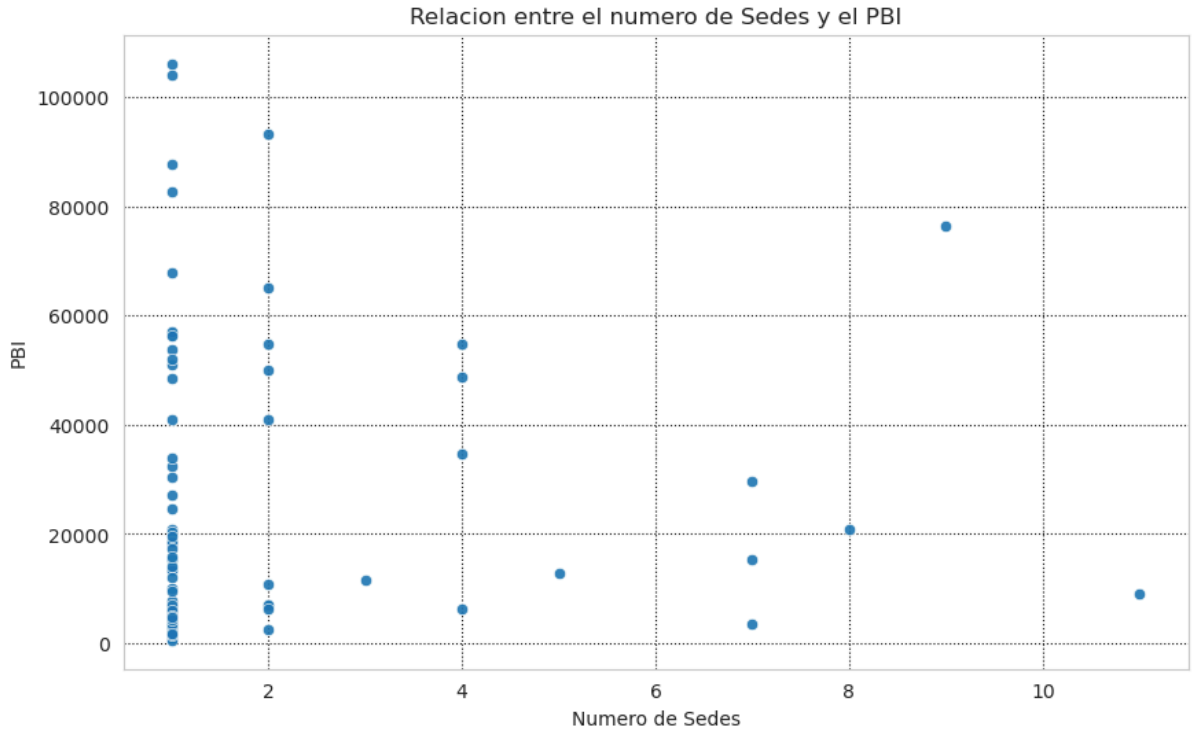
Podemos notar de manera mucho más clara lo visto en la Tabla 1. Vemos que la mayor cantidad de sedes están en Latinoamérica y el Caribe a pesar de no ser la región ni los países con mayor PBI, sino los más cercanos a nivel geográfico, económico y cultural, seguido de países con un alto PBI y menos relación cultural y, por último, regiones con bajo PBI y que están geográfica y culturalmente lejanos a Argentina.

2. Boxplot, por cada región geográfica,del PBI per cápita 2022 de los países donde Argentina tiene una delegación



Nuevamente podemos observar que el PBI en Latinoamérica y el Caribe es bajo en comparación al resto de las regiones donde hay un número significativo de sedes de Argentina. Teniendo en cuenta esto, para el resto de regiones tiene sentido trazar una relación entre el PBI de las regiones y la cantidad de sedes que éstas poseen.

3. Relación entre el PBI per cápita de cada país del 2022 y la cantidad de sedes en el exterior que tiene Argentina en esos países.



Extrañamente hay una relación inversamente proporcional entre la cantidad de sedes y el PBI del país, viendo que los países que tienen mayor cantidad de sedes son los que tienen menor PBI.

Conclusiones

Finalmente, teniendo en cuenta los reportes armados y los gráficos que nos ayudaron a entender mejor la información obtenida y teniendo en cuenta que el objetivo de trabajo es analizar el PBI per cápita de un país y las sedes en éste, podemos afirmar que no hay una relación lineal entre el PBI per cápita de un país y la cantidad de sedes Argentinas que posea. Es decir, sabiendo únicamente el dato del PBI per cápita de un país no podemos predecir correctamente la cantidad de sedes que éste tendrá y viceversa.
Con esto en mente, hemos analizado qué otros factores son de mucha influencia a la hora de observar la cantidad de sedes en un país. Estos son: la cercanía geográfica a la Argentina, entendemos que los países limítrofes y latinoamericanos en general suelen tener sedes argentinas más allá de su PBI per cápita; lazos comerciales, como es el caso de Estados Unidos y China; y lazos culturales, como es el caso de España, Italia e Israel.
En conclusión, afirmamos que el PBI per cápita es uno de los factores más importantes a la hora de observar las sedes argentinas en un país, mas no el más importante ni el único a ser considerado si se quiere entender esta relación en profundidad.