DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

TP1

Relación PBI - Cantidad de sedes Argentinas en el exterior

February 24, 2024 Laboratorio de Datos

GRUPO 100

Integrante	LU	Correo electrónico
Chapana Puma, Joselin Miriam	1197/21	yoselin.chapana@gmail.com
Martinelli, Lorenzo	364/23	martinelli.lorenzo12@gmail.com
Padilla, Ramiro Martin	1636/21	ramiromdq123@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

Estuve pensando y aca podriamos contar que fuimos pensando a lo largo del trabajo, los problemas con los que nos encontramos y como fue la dinámica, mas que nada porque la explicacion de la metodologia a seguir está en la parte de introducción.

2 Introducción

2.1 Objetivo y Fuente

El objetivo principal de este trabajo es encontrar una relación entre la cantidad de sedes de Argentina en un país y su PBI, si este será mayor, menor o si influirá la cantidad de secciones que una sede posea. Para esto, trabajaremos con los siguientes datos,

- PBI per cápita de los paises (1)
- Representaciones Argentinas en el exterior, donde tenemos, Datos básicos de las sedes, Datos completos de sedes y secciones (2)

(1) Extraido de la pagina del Banco Mundial, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

(2) Obtenidas del Ministerio de Relaciones Exteriores, Comercio Internacional y Culto, https://datos.gob.ar/dataset/exterior-representaciones-argentinas

2.2 Procedimiento

Este trabajo tendrá varias etapas, comenzando con el planteo de un Diagrama de Entidad Relacional (DER) adecuado al objetivo de nuestro trabajo, es decir, a partir de los datasets mencionados anteriormente, nos quedaremos solo con aquellos datos necesarios para resolver nuestro problema. Luego, pasaremos nuestro DER al modelo relacional, el cual se encontrará en tercera forma normal y especificará claves primarias (PK), claves candidatas (CK), claves foráneas (FK) y dependencias funcionales.

Una vez tengamos nuestro modelo planteado, pasaremos a Python donde realizaremos una limpieza de los datos tomando ciertas decisiones que estarán explicadas hacia el final de este informe. Finalmente, con los datos ya limpios, a traves de distintas librerias como Pandas, Matplotlib, Inlinesql, entre otras nos encargaremos de consultar, manipular y visualizar los datos necesarios para dar con nuestro objetivo.

3 Procesamiento de Datos

Antes de comenzar con el procesamiento de datos, nos encontramos con las fuentes de datos originales en distintas formas normales. Comenzando con **lista-sedes-datos** y **lista-secciones**, tenemos que ninguna de ambas se encuentra en primera forma normal puesto que el atributo *redes-sociales* de la primera y, el atributo *telefonos_principales* de la segunda tienen valores no atómicos. Al no estar en 1FN, tampoco se encontraran en 2FN.

Por otro lado, la tabla que contiene el pbi de los paises, está en primera y segunda forma normal, sin embargo no se encuentra en tercera forma normal puesto que tiene dependencias transitivas, en particular, la DF {Country Code $\rightarrow$ Indicator Name} es transitiva mediante el atributo *Indicator Code*, el cual no es clave candidata pero forma parte de la DF {Indicator Code $\rightarrow$ Indicator Name}, ademas tomamos *Country Code* como unica clave.

Y, por último, la tabla **lista-sedes** está en segunda forma normal, pero no está en tercera forma normal puesto que tiene dependencias transitivas como por ejemplo, la DF {sede_id $\rightarrow$ pais_iso_2} es transitiva mediante el atributo *pais_iso_3*. Esto es asi por que consideramos *sede_id* como clave primaria

3.1 Limpieza de Datos

Buscando mejorar la calidad de datos, encontramos problemas de instancia, por ejemplo, datos inconsistentes, problemas de proceso, puesto que hay diferentes criterios en la carga de los datos. Ahora, veamos cada dataset en detalle,

1. lista-sede-datos, en ella, encontramos un problema asociado a instancia en el atributo *'redes_sociales'* que influye en su consistencia puesto que no encontramos un criterio unificado a la hora de cargar distintas redes sociales. Por ejemplo, tenemos datos cargados en forma de URL donde es facil identificar a que red social pertenece, y por otro lado, tenemos datos que contienen, intuimos, nombres de usuarios, pero no podemos discernir su red social. Para tener una medida concreta acerca de la magnitud del problema utilizamos el método GQM de la siguiente manera:

   - Goal : Evitar que haya datos que no se puedan identificar con alguna red social.
   - Question : ¿Cual es la cantidad de elementos en redes_sociales que no podemos identificar?
   - Metrica : Proporción de registros sin campo redes_sociales que son Urls o comienzan con @.

   $$\text{M1: } \frac{\text{Cantidad de registros con datos en redes\_sociales que no son Urls o comienzan con @}}{\text{Cantidad de registros totales}}$$

   Para corregir esta tabla, además, seleccionamos columnas que consideramos relevantes para la resolución de nuestro problema principal. Puesto que encontramos datos redundantes, como dos nomenclaturas de codigo (pais_iso_3 / pais_iso_2) para un mismo pais, o la traducción al inglés de los nombres de estos paises. También, consideramos mantener unicamente los datos de redes sociales que eran consistentes, es decir, que eran Urls o comenzaban con @.

   Una vez hecha la limpieza de datos, procedemos a comparar las metricas antes y despues de dicha limpieza:

   - Antes: $M1 = 0.23$, es decir que un $\%23$ de registros son no deseados
   - Después: $M1 = 0$, es decir que un $\%0$ de registros son no deseados

2. En tabla correspondiente al PBI, API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73.csv, tenemos también un problema de instancia, que afecta en particular, la completitud del atributo '2022'. Asi como también, otro problema de instancia es que encontramos muchos datos que no corresponden a paises. Estos fueron descartados antes de elaborar la siguiente métrica,

   - Goal : Tener el dato '2022' que refiere al Pbi de cada país completo.
   - Question : ¿Es relevante la proporción de paises con el dato '2022' vacío?
   - Metric: Proporción de paises con el campo '2022' vacío.

   $$\text{M1: } \frac{\text{Cantidad de registros con dato vacio en el campo '2022'}}{\text{Cantidad de registros totales}}$$

   Para corregir este dataset, en primeria instancia, descartamos todas las columnas que no aportaban a nuestro objetivo, como por ejemplo, todas aquellas correspondientes a años anteriores a 2022. En segunda instancia, eliminamos aquellos datos que no correspondian con paises, por ejemplo 'Africa'.

   Una vez hecha la limpieza de datos, procedemos a comparar las metricas antes y despues:

   - Antes: $M1 = 0.09$, es decir, un $\%9$ de registros contienen NULL en el campo '2022'
   - Después: $M1 = 0$, es decir, un $\%0$ de registros son no deseados

   En último lugar, consideramos que al ser tan baja la proporcion de nulls, era conveniente descartarlos.

3. Por último, en lista-secciones, nos quedamos unicamente con las dos primeras columnas, sede_id y descripción, en las cuales no encontramos ningun problema de calidad.

3.2 Diagrama de Entidad Relacional

Una vez planteado nuestro objetivo, nos encargamos de ver que datos necesitabamos para alcanzarlo, y como estarian representados. Para esto, elaboramos el siguien diagrama de entidad relacional.

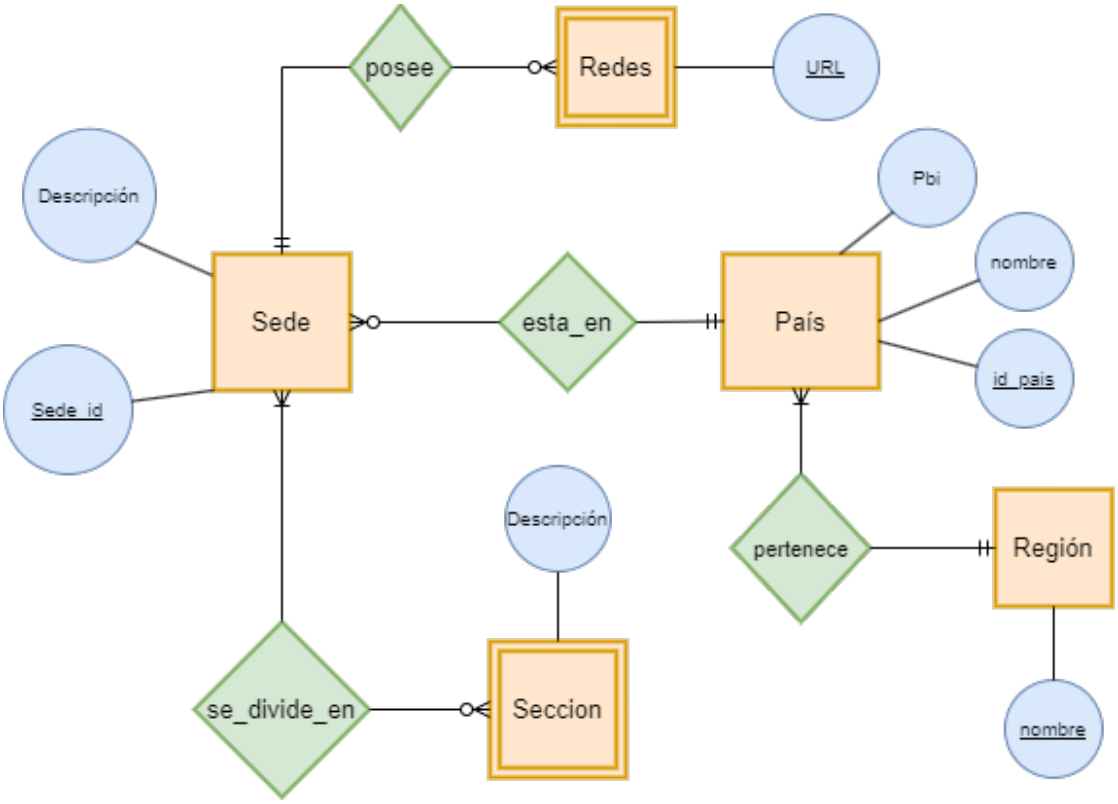


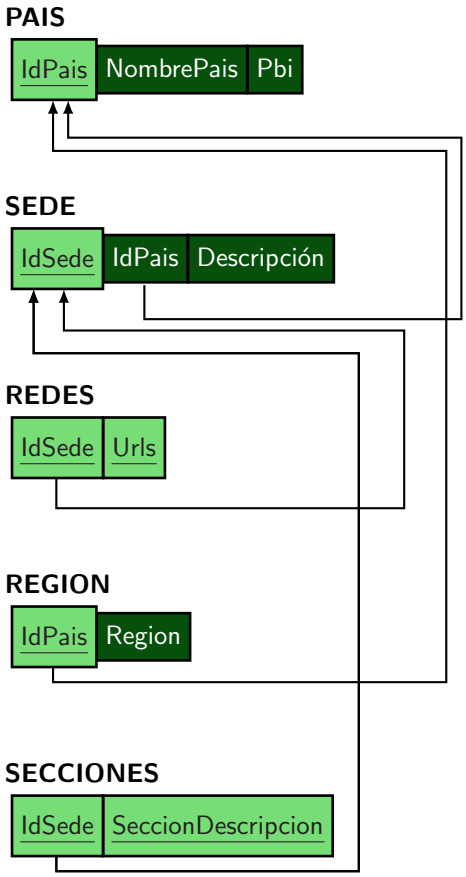
Figure 1: Diagrama de Entidad Relacional

Como se puede ver en la imagen anterior, consideramos que,

- Una sede esta en un país y solo en uno.
- Un País puede tener muchas o ninguna sede.
- Las secciones existen pues existen las sedes, entonces, lo consideramos una entidad debil.
- Una sección puede estar en una o muchas Sedes, por ejemplo, la Embajada en Brasil y en Chile tiene su sección Administración.
- Una sede tiene al menos una sección.
- Cuando hablamos de Redes, hablamos mas de un perfil en una red social, por lo tanto, una red puede pertenecer a una y solo una sede.

3.3 Modelo Relacional

Una vez que tenemos nuestro esquema gráfico, pasamos al planteo del modelo relacional. Notar que, las flechas representan las Foreign Keys y aquellos atributos subrayados representan las Primary Keys. En todas las relaciones, exceptuando a País, las Claves coinciden con las claves candidatas, esto se debe a que en País consideramos también a nombrePais como una posible clave.

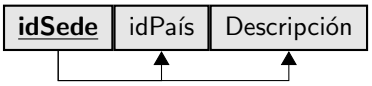


Ahora, solo resta mostrar las dependencias funcionales de este modelo relacional para dejar en claro que se encuentra en la forma normal deseada (3FN) y, también tener completo nuestro esquema para asi comenzar a manipular los datos.

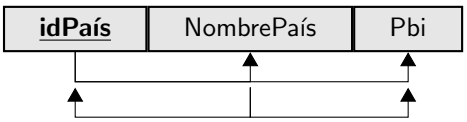
3.4 Dependencias Funcionales

A continuación, mostramos las dependencias funcionales de las relaciones de nuestro modelo relacional. En esta, se puede apreciar que todas se encuentran en segunda y tercera formal. Luego, con la limpieza de datos, garantizaremos que todos sus atributos sean atómicos y por ende se encuentre también en primera forma normal.

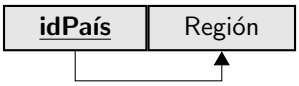
Sede



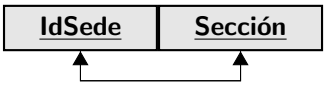
País



Región

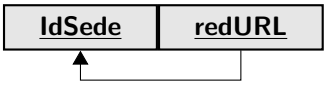


Sección



Observación: la única DF es {IdSede, Sección} $\rightarrow$ {IdSede, Sección}

Redes



3.5 Importación de Datos

Una vez ya limpios los datos, creamos con pandas un archivo csv para cada relación de nuestro modelo. Luego, los importamos a los dataframe vacíos creados anteriormente.

4 Decisiones tomadas

En los distintos procedimientos que llevamos acabo explicados en la sección anterior, encontramos necesario realizar tomas de decisiones para mejorar, mantener la calidad de datos y cumplir la normalización pedida en el informe. Esto debido a que encontramos atributos irrelevantes, que presentaban información confusa o datos ya presentados en otros atributos. Estas decisiones fueron las siguientes:

- En la lista-sede-datos encontramos muchos registros con datos correspondientes al atributo "redes_sociales" que no contaban con una forma consistente a la hora de cargarse. Por lo que, como fue explicado en la sección **3.1 Limpieza de datos**, mantuvimos los datos que efectivamente se podian deducir que eran URLs y aquellos que contenian "@" al principio. Esto último debido a que es la manera más frecuente de referirse a usuarios de una Red Social. En este caso especifico, tomamos la decisión de que estos casos sean adjudicados a la red social "Instagram". Consdieramos que, si los eliminamos, perdemos una cantidad importante de información.

- En el momento de crear el Diagrama de Entidad Relacioal y su correspondiente Esquema Relacional, tomamos la decisión de establecer como clave primaria aquellos atributos que son códigos. Por esto mismo, le dimos importancia a los id de las sedes en lugar de a su nombre. Asi también tomamos como clave el id de cada pais obtenida de la tabla que contenia los PBI de cada país, además porque en lista-sede-datos contamos con un atributo idéntico (pais_iso_3) que nos iba a permitir establecer una conexión entre tablas.

- Como en lista-sedes-datos encontramos la misma cantidad de registros y la misma información que en lista-sedes, tomamos la decisión de utilizar la primera tabla ya que contiene más atributos, y por ende, más información necesaria para analizar.

5 Análisis de datos

5.1 Consultas SQL

Para el analisis de datos, se elaboraron una serie de tablas mediante consultas SQL.

En primer lugar, se consideraron la cantidad de sedes, el número promedio de secciones por sede y el PBI per cápita correspondiente al año 2022 de cada país. De acuerdo con los resultados obtenidos en la tabla "pais_seccciones.csv" (consulte el Anexo para visualizarla), se destaca que Brasil y Estados Unidos son los países con mayor número de sedes, con PBIs de 8918 y 76330, respectivamente. Además, se observa que Mónaco, con un PBI de 240862, y Burundi, con un PBI de 259, son los países con el mayor y el menor PBI, respectivamente, a pesar de no tener sedes. Esto nos hace dudar, en que infulirá la cantidad de sedes argentinas.

En una segundo lugar, se analizó la tabla "region_pais_pbi", que incluye información sobre la cantidad de paises con alguna sede en distintas regiones geográficas y el promedio de su Pbi. Se destaca que Oceanía es la región con el mayor PBI, a pesar de tener la menor cantidad de sedes en Argentina. Sin embargo, podemos mirar también el porcentaje respecto a los Países totales. Oceanía posee 15 países según la ONU, es decir que argentina tiene presencia en el 13,3% de los paises, mientras que en África Subsahariana argentina cuenta con el 8% de paises cubiertos. Estos, no parecen porcentajes que nos permitan asumir algo respecto al Pbi, puesto que uno posee el promedio más alto y el otro, el mas bajo. Pero, podemos notar que desde Europa occidental, hasta América del Sur, el porcentaje de sedes Argentinas es muchisimo más alto, superando el 50%.

Region geografica	Paises Con Sedes Argentinas	Promedio PBI per Cápita 2022 (U$S)
OCEANÍA	2	56759.22
EUROPA OCCIDENTAL	16	52978.08
AMÉRICA DEL NORTE	3	47581.26
ASIA	21	23375.33
EUROPA CENTRAL Y ORIENTAL	8	15425.61
AMÉRICA CENTRAL Y CARIBE	14	13722.88
AMÉRICA DEL SUR	10	9447.21
ÁFRICA DEL NORTE Y CERCANO ORIENTE	5	4508.71
ÁFRICA SUBSAHARIANA	7	2459.07

Figure 2: Cantidad de sedes argentinas y promedio PBI para cada region

En tercera instancia, nos planteamos la siguiente pregunta con respecto a las redes sociales: ¿Cuán variado es, en cada país, el tipo de redes sociales que utilizan las sedes?. Para abordar esta interrogante de manera precisa, se construyó una tabla, se puede consultar en el anexo como 'redes_por_pais.csv', que muestra los nombres de los países y la cantidad de redes sociales utilizadas por sus sedes. Al analizar esta tabla, se observó que algunos países como Bélgica y Estados Unidos tienen la máxima cantidad de redes sociales, alcanzando un total de 6. Por otro lado, se identificó que países como Argentina y Haití no tienen presencia en redes sociales en sus sedes.

Por ultimo, tenemos una tabla, 'pais_sede_red.csv' en el anexo, donde se han registrado las redes sociales de cada país junto con sus respectivas URL. Se observa que la mayoría de los países cuentan con presencia en Facebook, seguido por Twitter y YouTube, mientras que LinkedIn y Flickr tienen una menor presencia en comparación

5.2 Gráficos

Existen otras formas de visualizar datos las cuales mejoran la comprensión de la información. En este trabajo, se realizaron 3 tipos de gráfico, adecuados a los datos, histograma, boxplot, y scatter plot con el fin de entender más aquello que las tablas dicen. A continuación, en el gráfico de cantidad de sedes por región se puede observar que América del sur es la región con mayor cantidad de sedes, mientras que Oceanía es la que menos posee.

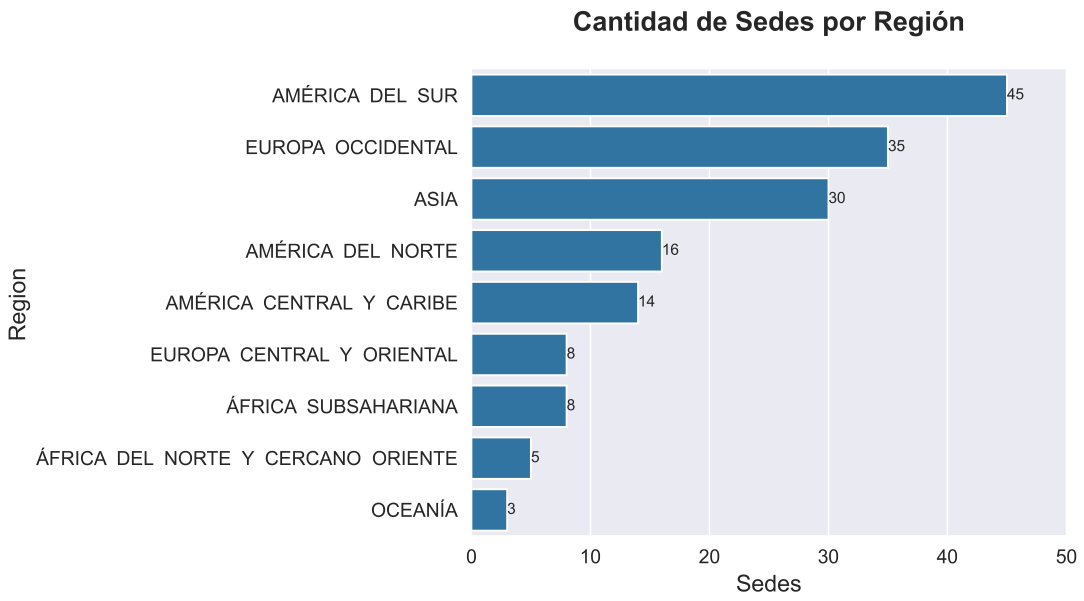


Figure 3: Cantidad de sedes por region

En el gráfico, se puede ver que Oceanía tiene la mediana más alta en relación con el PBI, mientras que África subsahariana tiene la mediana más baja. Por otro lado, se observa que Europa occidental tiene la sede con el PBI más grande, se puede ver claramente por el outlier que posee, el cual se encuentra más a la derecha en el grafico que los demás. También, podemos destacar que la mediana de Oceania, América del Norte y Europa Occidental es bastante similar, sin embargo, podemos ver que estos ulitmos dos presentan un "rango" de Pbi mucho más amplio.

Consideramos que el boxplot con todas las regiones juntas no aportaba toda la claridad necesaria, en particular, mirando aquellas dos regiones con el menor pbi. Por esto, se realizó otro grafico enfocado en África del norte y cercano oriente y África subsahariana. A continuación, se pueden ver ambos.

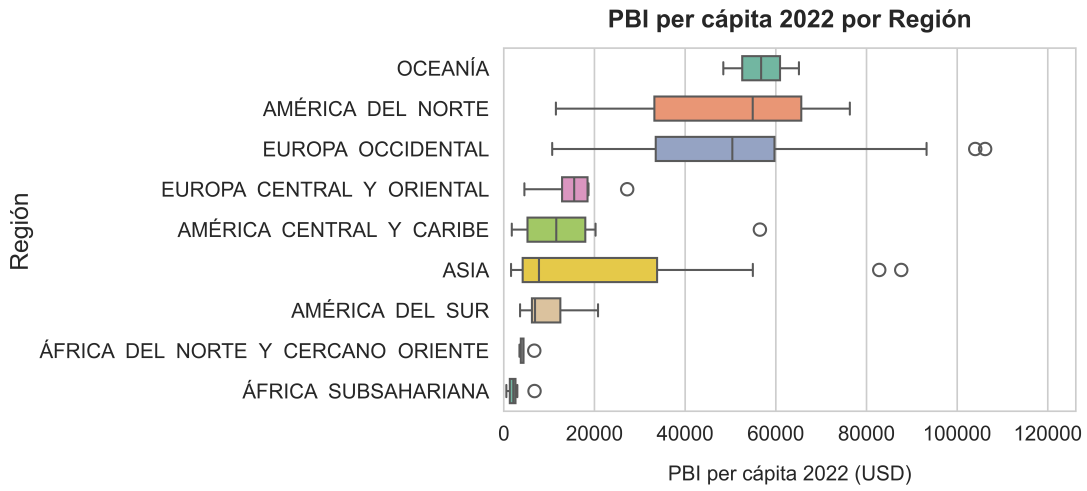


Figure 4: Pbi por región donde Argentina tiene una sede

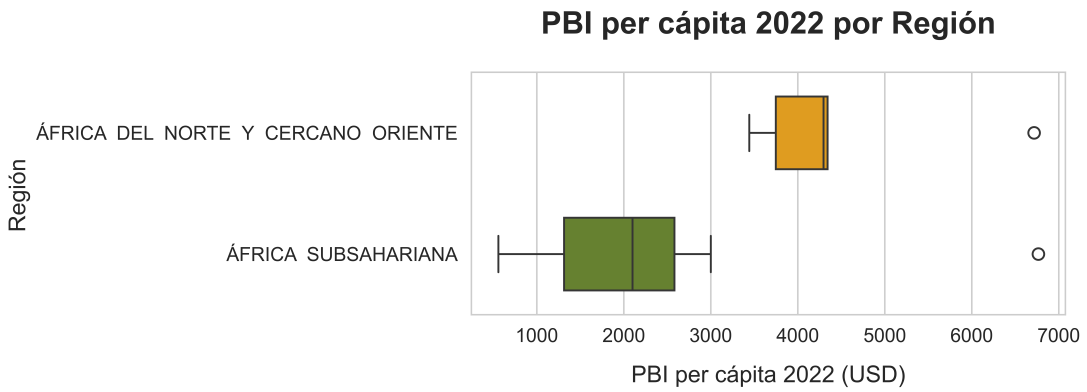


Figure 5: Pbi por region donde Argentina tiene una sede, en este grafico solo se muestra dos las regiones con menor PBI

Siguiendo con el último gráfico, se puede observar que los países que no tienen sedes poseen diversos niveles de PBI. Por un lado, hay países con valores altos de PBI, mientras que también se encuentran países con PBI bajos. Esto nos lleva a pensar que la cantidad de sedes no es un factor que influya en el PBI, consideremos mejor, comparar el promedio entre aquellos paises que tienen presencia Argentina contra aquellos que no.

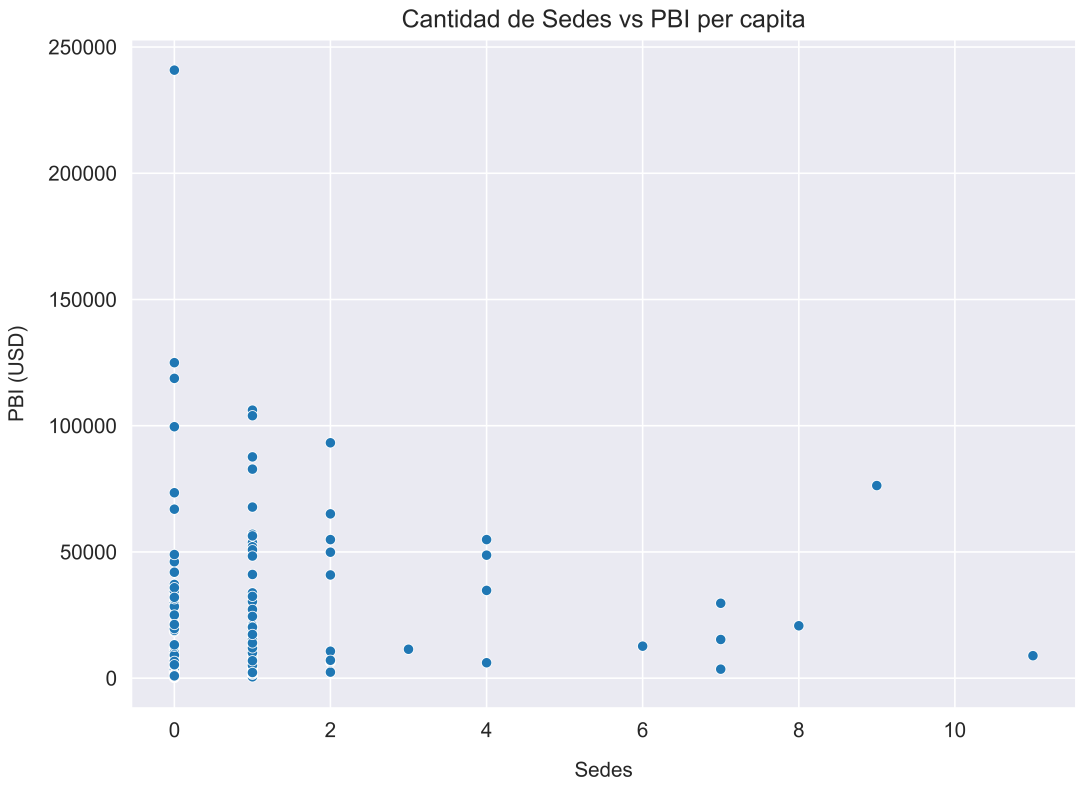


Figure 6: Pbi por region donde Argentina tiene una sede, en este grafico solo se muestra dos las regiones con menor PBI

5.3 Influencia de la Presencia Argentina

Como vimos antes, no pudimos terminar de determinar si la cantidad de sedes total tenia una influencia directa sobre el Pbi per Capita de un pais. Entonces, reformulemos nuestra pregunta un poco y veamos también si la muestra de paises por ejemplo con 11 sedes, por ejemplo, es significativa.

Sedes	Paises	Promedio PBI
11.0	1.0	8918.0
0.0	110.0	16433.0
9.0	1.0	76330.0
2.0	10.0	33747.0
8.0	1.0	20795.0
7.0	3.0	16210.0
1.0	64.0	21619.0
4.0	4.0	36141.0
6.0	1.0	12720.0
3.0	1.0	11497.0

Figure 7: La tabla muestra la cantidad de paises con determinada cantidad de sedes y el pbi promedio de estos

Se puede empezar a vislumbrar que en las filas con una muestra signficativa de paises hay un patrón...

Sedes	Paises	Promedio PBI
0.0	110.0	16433.0
1.0	64.0	21619.0
2.0	10.0	33747.0

Figure 8: Muestra del pbi promedio por cantidad de sedes donde hay más de 10 paises

Ahora si, todo parece indicar que la presencia de sedes argentinas influye de manera positiva en el pbi. Veamos también, solamente para dejar en claro esto la comparación entre el Pbi promedio de aquellos paises sin presencia Argentina contra aquellos con. En este, se verá claramente una tendencia que nos permitirá arribar a nuestro objetivo, encontrar una relación ente la cantidad de sedes argentinas y el Pbi.

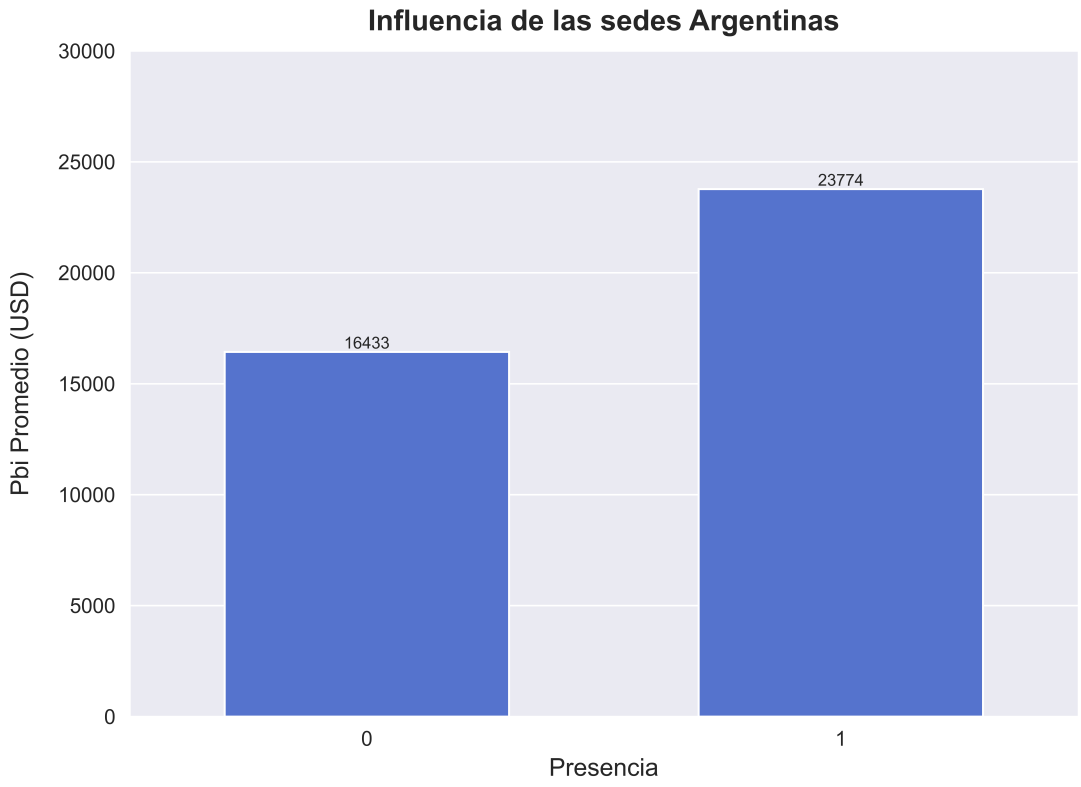


Figure 9: 0 y 1 indican la presencia de sede

6 Conclusiones

A raiz de lo visto en la sección anterior, 'Analisis de Datos', consideramos que la respuesta a nuestro problema es clara. Aquellos paises que poseen alguna sede Argentina, en promedio, tienen un Pbi mayor que aquellos que no. Más aún, vimos que cuando dividimos a los paises en cantidad de sedes por paises y tomamos aquellos grupos con muestras de paises significativas, el Pbi tendia también a ser mayor. Como todo, hay paises que se escapan de estas métricas, como por ejemplo, Monaco quien no posee sedes y aún asi es el país con mayo Pbi per cápita del mundo. También, creemos que esta tendencia se debe a que argentina tiene una cantidad baja, por no decir nula, de sedes en áfrica, mientras que en regiones historicamente mas ricas como europa occiddental o norteamerica tiene una fuerte presencia.