DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo practico 01

26 de febrero de 2024 Laboratorio de Datos

pip install grupo_1

Integrante	LU	Correo electrónico
Rocca, Santiago	152/23	santiagorocca17@gmail.com
Moguilevsky, Agustin	951/23	agumogui@gmail.com
Pina, Martin	320/23	martinpina04@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

INTRODUCCIÓN Y FUENTES DE DATOS

El objetivo general de este trabajo consiste en la investigación y el análisis de la posible relación entre el PBI per cápita de distintos países durante el año 2022 y las Representaciones Argentinas en dichos países. Utilizando los datos proporcionados por el Banco Mundial y el Gobierno Nacional Argentino, se busca determinar si existe una correlación significativa entre estas variables. De esta manera podremos comprender si la presencia diplomática de Argentina en el extranjero podría estar relacionada con el nivel de desarrollo económico de esos países. Para llevar esto a cabo nos encontramos con escasez y poca accesibilidad a la información en las fuentes de datos, en consecuencia, utilizamos técnicas de gestión de tablas y mejorando la calidad de estos, esquematizamos una nueva base de datos definida en el modelo relacional para poder cumplir con el objetivo.

Las fuentes de datos para llevar adelante este trabajo (archivos CSV) fueron las siguientes: 'API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73' (https://rb.gy/2p726m) donde se encuentra un listado de países con su respectivo PBI desde 1960 a 2022, 'lista-sedes' y 'lista-sedes-completo' que contienen información de las Representaciones Argentinas (id, nombre completo del titular, ubicación de la misma, etc.) y por ultimo 'lista-secciones', que proporciona las diferentes secciones de cada una de las representaciones (https://datos.gob.ar/dataset/exterior-representaciones-argentinas).
Observación: renombramos 'API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73' cómo 'pais_pbi.csv'.

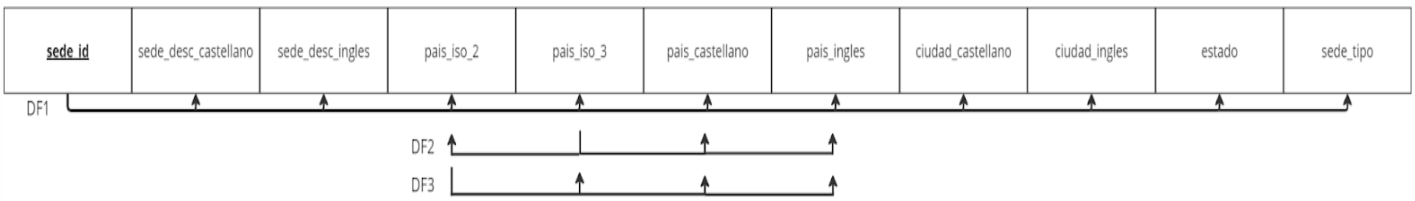
Formas Normales:

- **Forma normal de 'lista-sedes-completo':**
  En esta tabla se pueden observar atributos que poseen datos no atómicos, lo que implica que no cumple con la Primera Forma Normal (1FN). Un ejemplo de esto se da en el atributo "redes_sociales", donde algunas sedes poseen más de un link y es por ello que no es atómico.

- **Forma normal de 'lista-secciones':**
  Se pueden observar atributos que poseen datos no atómicos lo que implica que tampoco cumple con la 1FN. Por ejemplo, el campo "telefonos_adicionales" abarca en algunas secciones más de un número telefónico.

- **Forma normal de ‘lista-sedes’:**

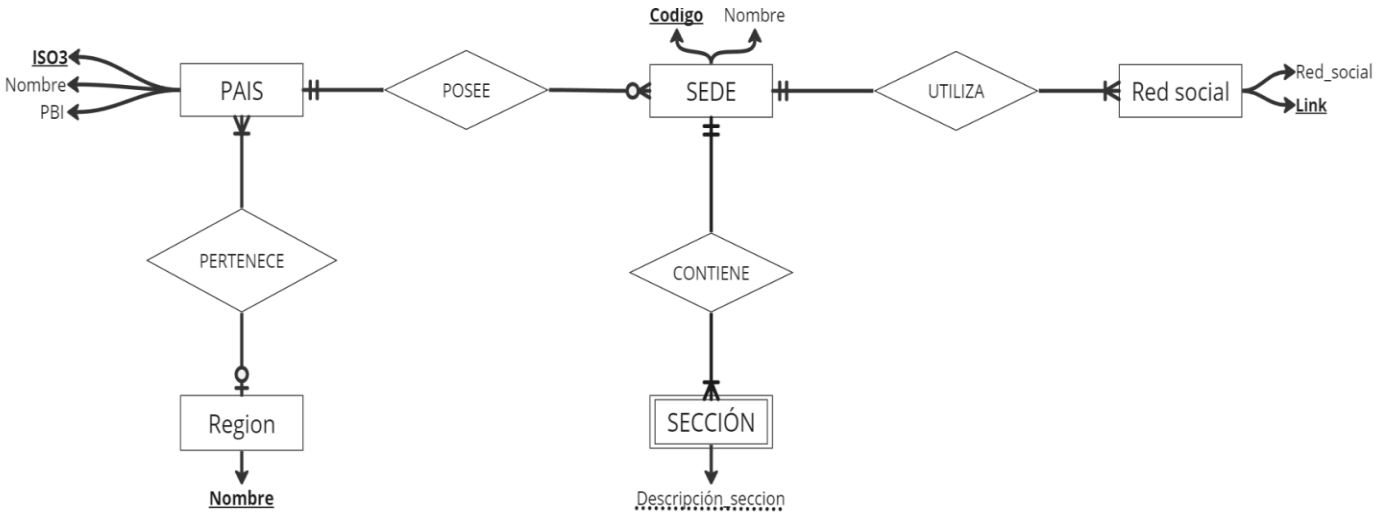
  Todos los datos son atómicos. De esta manera podemos afirmar que está en 1FN. Para determinar si se encuentra en 2FN planteamos este esquema a través de la semántica de los atributos:



La clave primaria es: “**sede id**”. Al no existir dependencias funcionales parciales (“sede_id” determina todos los atributos no claves), se cumplen los requisitos para afirmar que ‘lista-sedes’ se encuentra en Segunda Forma Normal (2FN). Sin embargo, como existen dos relaciones transitivas (DF1 a DF2 y DF1 a DF3), no se cumple la definición de Tercera Forma Normal (3FN). Por lo tanto, esta relación se encuentra en 2FN.

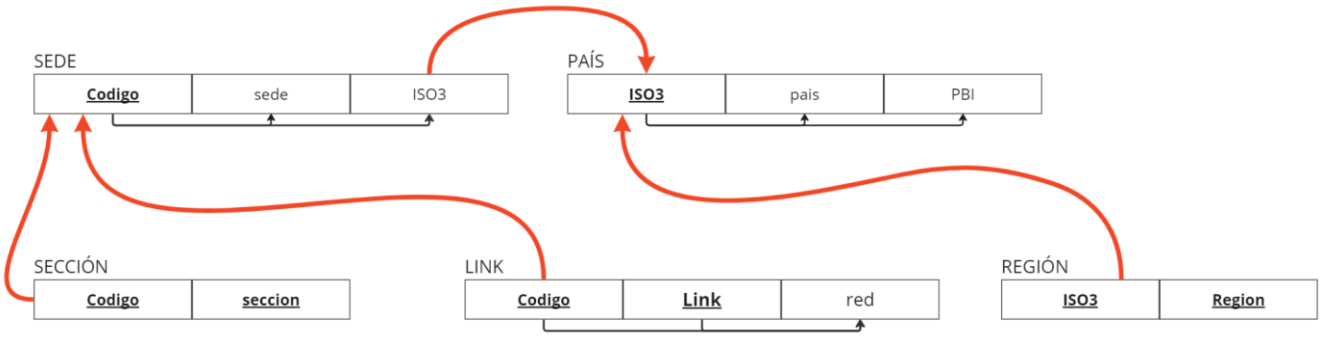
DISEÑO DE ESQUEMA RELACIONAL (DER)

Luego de descargar los datos de sus respectivas fuentes, diseñamos el siguiente modelo con respecto al objetivo planteado. Consideramos que este esquema es adecuado para abordar el trabajo. Hemos identificado las entidades principales como País, Sede, Región y Red social, mientras que hemos clasificado a Sección como una entidad débil, debido a que su existencia depende de la presencia de una sede.



MODELO RELACIONAL Y DEPENDENCIAS FUNCIONALES

En base al DER, presentamos el siguiente modelo relacional:



Modelo relacional (las flechas rojas determinan las claves foráneas)

País pertenece a Región (relación uno a muchos):

No todos los países tienen una región asociada, por lo tanto, hemos establecido una relación llamada Región, donde asignamos a cada país su región correspondiente, evitando así campos vacíos.

País posee Sede (relación uno a muchos):

Cada país puede o no tener sedes, por lo tanto, hemos agregado la clave foránea ISO3 en Sede para vincular ambas relaciones.

Sede contiene Sección (relación uno a muchos):

Una sede contiene al menos una sección, entonces creamos una nueva relación llamada Sección con superclave: (Codigo_sede, nombre_seccion). Así aseguramos que la clave principal sea única y cumpla 3FN.

Sede utiliza red social (relación uno a muchos):

Una sede puede utilizar una o varias redes sociales; para esto, creamos una relación llamada Link con superclave: (codigo_sede, link). Así aseguramos que la clave principal sea única y cumpla 3FN de la misma forma.

GOAL-QUESTION-METRIC (GQM)

El paso previo a importar los datos a los dataframes fue garantizar su calidad. Utilizando el modelo GQM, definimos objetivos de calidad, formulamos preguntas específicas y seleccionamos métricas adecuadas para evaluar la disponibilidad, consistencia y completitud de ellos. A continuación, abordaremos los problemas de calidad de cada dataset de manera separada:

- **pais_pbi**

1)

El atributo de calidad afectado es la consistencia. Algunos de los datos que corresponden a “Country Name” son regiones. Consideramos que el problema reside en el modelo, ya que no existe un campo para definir si se trata de una región o un país.

Objetivo: El dato correspondiente a “Country Name” no sea una región.
Pregunta: ¿Solo aparecen países asociados a “Country Name”?
Métrica: Proporción de registro que no son países.

Los campos que no corresponden a un país son 71. Por lo tanto, la proporción de regiones es del 26.69%, en comparación con el 73.31% de los datos restantes.

2)

El atributo de calidad afectado es la completitud. No todos los datos de PBI (2022) están completos. Estos son fundamentales para desarrollar la conclusión del trabajo. De esta manera determinamos que este problema reside en la instancia. Posiblemente no se encuentren por situaciones políticas o por la caída fuerte de su economía.

Objetivo: El dato correspondiente al PBI de cada país esté completo.
Pregunta: ¿Cuál es la proporción de países que tienen el campo PBI vacío?
Métrica: Proporción de registro con campo 2022(PBI) vacío.

Desconocemos el PBI (2022) de 14 países. Por lo tanto, la proporción de campos NULL es de 6.69%, en comparación con el 93.31% de los datos restantes."

- **lista-sedes-completo**

  El atributo de calidad afectado es la disponibilidad. Esto se debe a que los campos que no son correspondientes a un link no son abordables. Este problema reside en la instancia, pues en este campo se encuentran datos correspondientes a nombres de usuarios (sin poder determinar a qué red social pertenece), correos electrónicos, entre otros.

  Objetivo: El dato correspondiente a “redes_sociales” sea un link.
  Pregunta:¿Cuál es la proporción de sedes sin redes sociales válidas?
  Métrica: Proporción de sedes con redes sociales válidas.

  El campo "redes_sociales" contiene información de un total de 126 sedes. De estas, únicamente 14 no corresponden a un link válido. Por lo tanto, la proporción de entradas inválidas es del 11.11%, en comparación con el 88,89% de los datos restantes que sí son válidos

  Observación: Consideramos válidos los enlaces con el siguiente formato:
  www.”nombre_red”.com\” …”.

IMPORTAR DATOS

Una vez completado el proceso de limpieza y mejora de calidad de datos, procedimos a importar los mismos a los dataframes correspondientes. Desde 'lista-sedes' importamos los datos a Sede, Link y Región, desde 'lista-secciones' a Sección, y finalmente, desde 'país_pbi.csv' a País a través de consultas SQL.
Para crear el dataframe esperado, desarrollamos un algoritmo en Python. El problema por resolver fue la presencia de más de un enlace en el campo 'redes_sociales' del archivo 'datos-completos-sedes.csv'. De esta manera, conseguimos una relación en la que el atributo 'link' tiene un solo elemento y además de que red social se trata.
Adjuntamos código:

```
archivo = open("Datos_Completos_Sedes.csv", encoding = "utf8")
filas = csv.reader(archivo)
next(filas)

links_nuevo = []

for linea in filas:
    if linea[5] != None:
        links_linea = linea[5].split(" // ")
        for link in links_linea:
            if link != " " and link != "":
                links_nuevo.append([linea[0],link, link.split(".")[1]])

archivo.close()

links_nuevo = pd.DataFrame(links_nuevo, columns = ["codigo_sede",
                                                   "link",
                                                   "red_Social"])
```

CONSULTAS SQL

Redujimos las consultas a una cantidad reducida de líneas. Sin embargo, para una mejor visualización de estas, entrar a la carpeta “consultas” donde se encuentran los dataframes en formato .csv.

Consulta 1:

Pais	sedes	Secciones_Promedio	PBI_per_Capita_2022_U$S
Brazil	11	3.6	8917.67
United States	9	7.5	76329.6
Uruguay	8	4	20795
Bolivia	7	7.5	3600.12
Chile	7	14	15355.5
Spain	7	5.33333	29674.5
China	5	5.66667	12720.2
Canada	4	3.5	54917.7

Con los valores de esta muestra podemos observar como Brasil es el país con mayor número de Representaciones Argentinas en el mundo, esto se puede dar debido a la cercanía y/o a las fluidas relaciones comerciales. Sin embargo, el promedio de secciones es relativamente bajo en comparación a los países con al menos 3 sedes. En segundo y tercer lugar, encontramos a Estados Unidos y China, dos potencias mundiales.

Consulta 2:

Region_geografica	Paises_con_Sedes_Argentinas	promedio_pbi
OCEANÍA	2	56759.2
EUROPA OCCIDENTAL	18	52978.1
AMÉRICA DEL NORTE	3	47581.3
ASIA	23	24354.6
EUROPA CENTRAL Y ORIENTAL	8	15425.6
AMÉRICA CENTRAL Y CARIBE	13	13776.1
AMÉRICA DEL SUR	11	9447.21
ÁFRICA DEL NORTE Y CERCANO ORIENTE	5	4508.71
ÁFRICA SUBSAHARIANA	7	2459.07

Cantidad de países con Sedes Argentinas por región y PBI promedio de la región

Consulta 3:

pais	cantidad_redes
Belgium	6
United States	6
Canada	5
Italy	4
Mexico	4
Spain	4
Switzerland	4
Tunisia	4
Algeria	3

Se puede diferenciar una gran cantidad de países que no poseen ninguna red social. Esto coincide en su mayor parte en los países que Argentina no posee representación alguna. Sin embargo, en los países donde sí existen sedes, la mayoría utiliza más de una red social.

Consulta 4:

Paises	sede	red	URL
Algeria	EARGE	Facebook	www.Facebook.com/ArgentinaEnArgelia
Algeria	EARGE	Instagram	www.Instagram.com/argenargelia
Algeria	EARGE	Twitter	www.Twitter.com/ARGenArgelia
Angola	EANGO	Facebook	www.Facebook.com/ArgentinaEnAngola/
Angola	EANGO	Instagram	www.Instagram.com/embargentinaenangola/
Armenia	EARME	Facebook	www.Facebook.com/ArgentinaEnArmenia
Armenia	EARME	Instagram	www.Instagram.com/arginarmenia/
Armenia	EARME	Twitter	www.Twitter.com/ARGinArmenia
Australia	CSIDN	Facebook	www.Facebook.com/ArgentinaEnSidney/
Australia	EAUST	Facebook	www.Facebook.com/ArgentinaEnAustralia/

Redes sociales de cada sede y su localización

VISUALIZACIONES

Si se requiere una mejor visualización de los gráficos, entrar a la carpeta gráficos donde se encuentran las imágenes exportadas del código.

Gráfico 1:

En este gráfico se visualiza las Representaciones Argentinas agrupadas por sus regiones.
América del Sur lidera con un total de 39, seguida por Europa occidental, donde se encuentran 32 sedes. En Asia, América del Norte y América Central se observa una alta proporción de sedes en relación con la cantidad de países pertenecientes a estas regiones. Mientras que se dispone de una menor cantidad en las regiones africanas y en Oceanía.

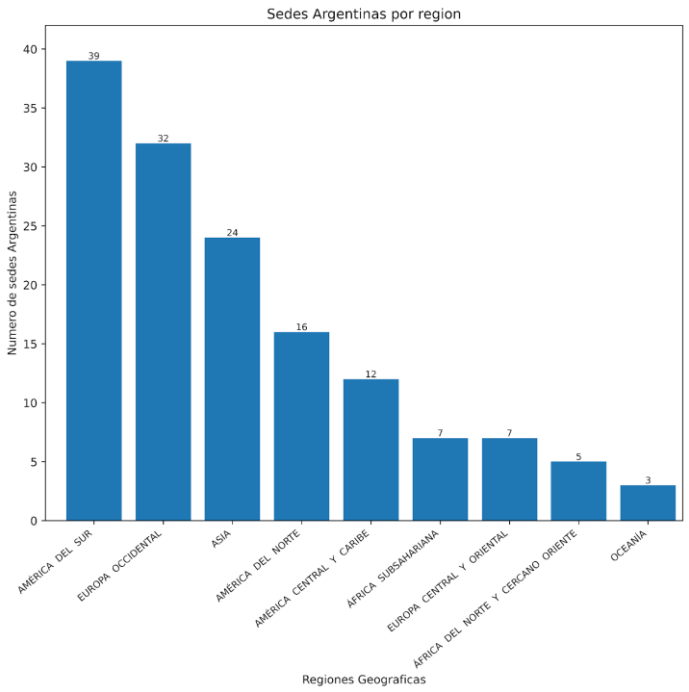


Gráfico 2:

En este gráfico se visualiza la distribución del PBI por región. Se observa una gran varianza en los percentiles 75 a 100 del PBI tanto en Europa Occidental, Asia y América del Norte. Oceanía posee una menor varianza, con un gran PBI en contraposición con las regiones africanas, las cuales presentan baja varianza y un PBI bajo. Entre el gráfico anterior y este podemos observar que Argentina posee una gran cantidad de sedes en países vecinos pese al bajo PBI que presentan los países de América del Sur.

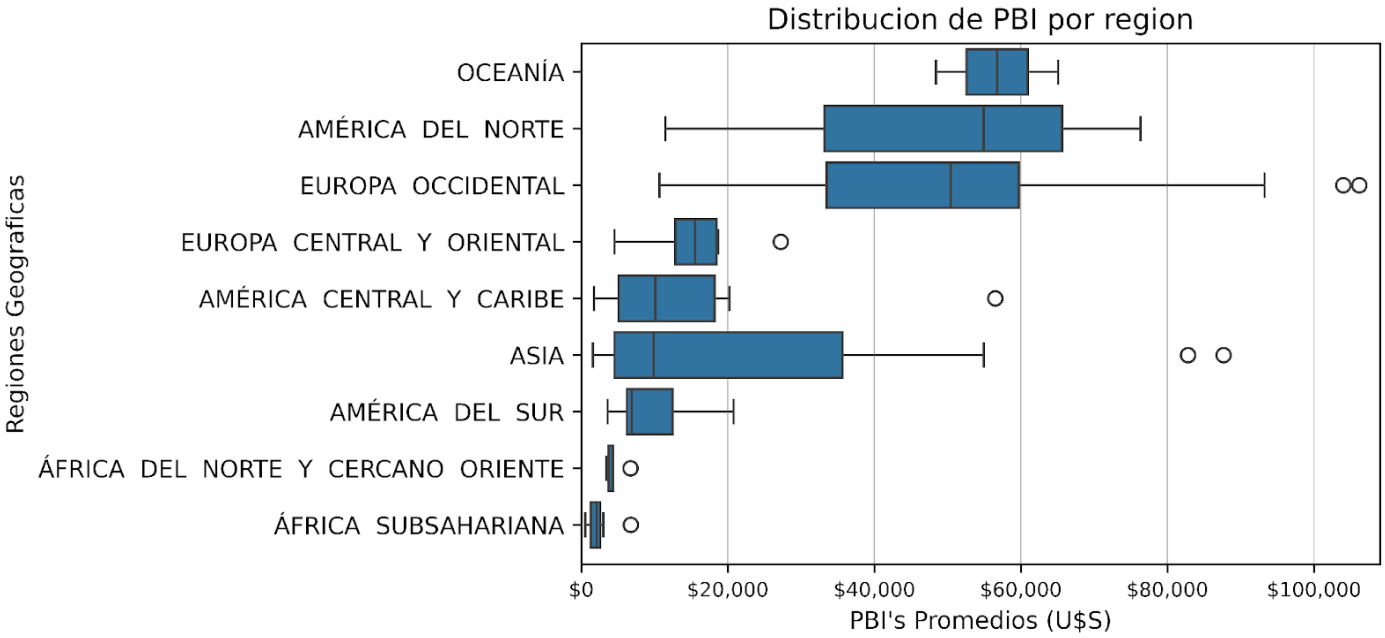


Gráfico 3:

En el siguiente gráfico se observa cuántas sedes posee cada país con su PBI. Se puede determinar que a medida que aumenta el PBI per cápita, también aumenta la cantidad de sedes argentinas.
Sin embargo, no se puede apreciar una relación directa entre estas variables.

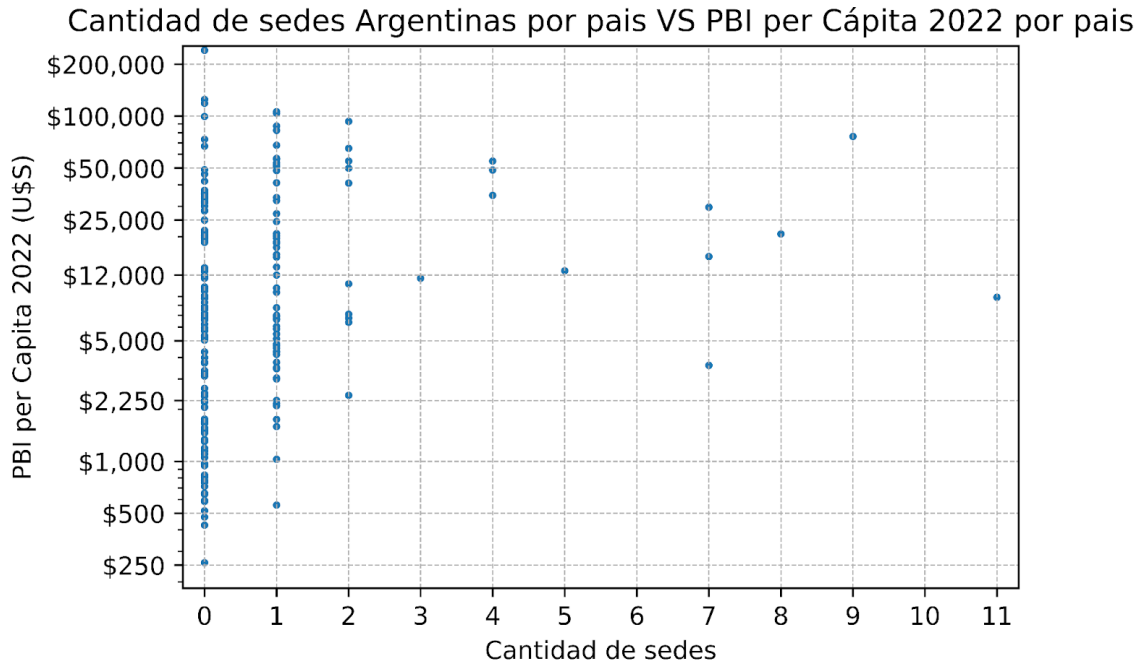
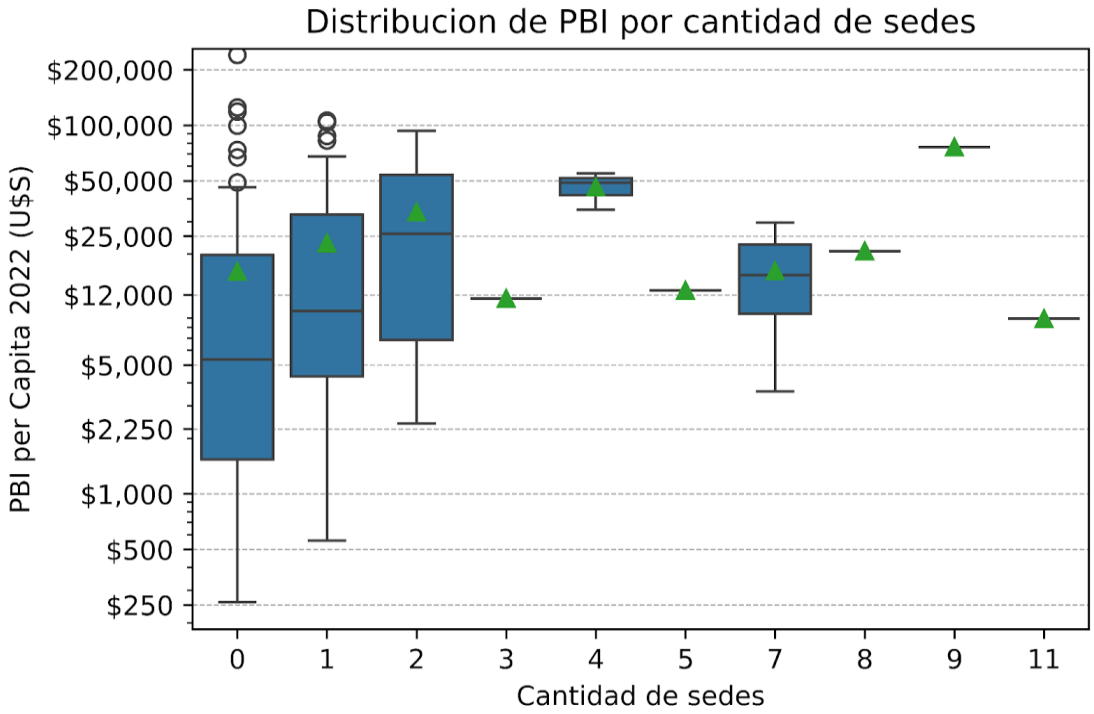


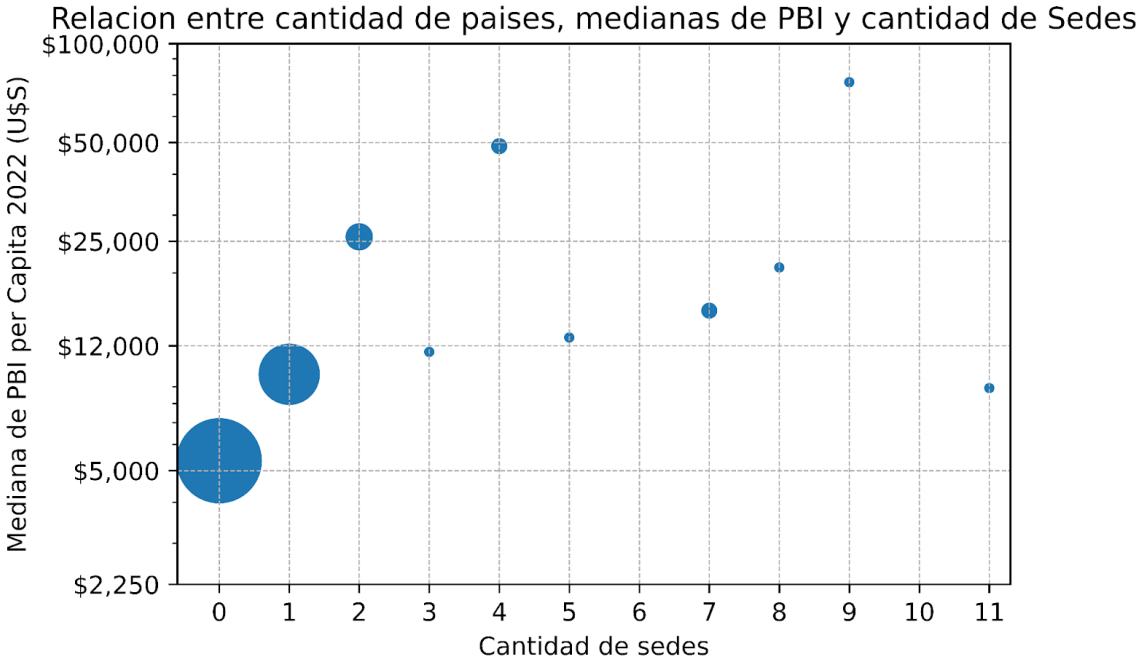
Gráfico 4 y 5:

Con el objetivo de visualizar mejor la relación mencionada anteriormente generamos los siguientes gráficos (ambos poseen una escala logarítmica para una mejor visualización):



Donde se puede observar que las distribuciones en donde Argentina posee sedes hay un mayor PBI que en donde no, se puede ver por las medianas este efecto. No tanto por el promedio, en donde no hay sedes, debido a que como existen una gran varianza y puntos outliers, no es tan fiable esta característica como el percentil 50.

Es por ello por lo que decidimos graficar un gráfico de dispersión, donde el tamaño está determinado por la cantidad de países que tienen cierta cantidad de sedes argentinas, los ejes por la mediana del PBI y la cantidad de sedes:



Aquí claramente se puede observar la correlación a determinar por este trabajo, sin embargo, no es perfecta por lo que podría haber otros indicadores los cuales determinan, además del PBI , la cantidad de sedes argentinas en otros países.

DECISIONES TOMADAS

- Modificamos los archivos originales, eliminando los campos que eran irrelevantes para llevar adelante el trabajo. Por ejemplo, el dato correspondiente al PBI de años distintos a 2022.
- Optamos por utilizar los datos de las sedes la cuales se encuentran activas en el dataset lista-sede-completa.
- Modificamos la estructura de los atributos de país PBI ya que el dataset original posee irregularidades en cuanto al formato del encabezado.
- Decidimos no utilizar el dataset llamado “lista-sedes” ya que los datos que albergaba estaban cubiertos por lista-sedes-completo. Por lo tanto, no limpiamos tal dataset, si no que mejoramos la calidad de los datos de listas-sedes-completo directamente.
- En base a los resultados obtenidos de GQM, decidimos eliminar:
    - 71 tuplas del archivo ‘pais_pbi.csv’ ya que no corresponden a países, sino que a regiones.
    - 14 tuplas que no contenían información del PBI 2022 en ‘país pbi.csv’.
    - los datos que no corresponden a un link en “redes_sociales”.
- Modificamos los datos que eran enlaces, pero no cumplían con el formato válido.
- Llevamos el “gráfico 3” a una escala logarítmica, con el objetivo de visualizar mejor los datos de menor PBI.
- Creamos dos gráficos adicionales, con el objetivo de tener mayor claridad a la hora de la conclusión

CONCLUSIONES:

Nuestro análisis revela una correlación moderada entre el PBI per cápita de un país y la cantidad de sedes que Argentina tiene en él. Sin embargo, esta correlación no parece ser directa. La relación entre estas dos variables no es perfecta, lo que sugiere que otros factores también influyen en la cantidad de sedes argentinas por país, más allá del PBI per cápita. Es notable la extensa distribución geográfica de las sedes argentinas, abarcando los principales países del mundo, con un aumento de sedes en América, Asia y Europa.

Los países con un PBI per cápita más alto suelen albergar un mayor número de sedes de Argentina, mientras que aquellos con un PBI per cápita más bajo generalmente poseen una única sede o ninguna. Sin embargo, encontramos una excepción en los países con los que Argentina comparte región, donde se presenta un considerable aumento de sedes por país.

Para una mejor comprensión y una relación más clara, recomendamos considerar la incorporación de nuevas fuentes de datos. Además, una actualización de los datos sería necesaria, ya que la información sobre las sedes se basaba en datos del año 2018.