DEPARTAMENTO DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 01: *Gestión y Visualización de Datos*

Verano 2024 *Laboratorio de Datos*

Integrante	L.U.	Correo Electrónico
Ibarra, Abril	945/23	Abrilibarra3095@gmail.com
Vassolo, Francisco	1121/23	vassolofran@gmail.com
Domínguez, Rocío	798/22	rociodominguezcpm@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires.
Ciudad Universitaria
http://www.exactas.uba.ar

Trabajo Práctico 01
Gestión de Datos y Visualización

Ibarra, Abril; Vassolo, Francisco; Dominguez, Rocio

Febrero 2024

1 Resumen

En este informe se detalla el trabajo realizado siguiendo las consignas del trabajo práctico, el cual se enfocó en analizar la relación entre el Producto Bruto Interno (PBI) per cápita de diferentes países y la cantidad de sedes que Argentina posee en cada territorio. Se realizó un proceso continuo de limpieza y procesamiento de los conjuntos de datos (datasets). A partir de las tablas limpias investigamos los datos para decidir cuáles conservar, cómo organizarlos e identificar los elementos (entidades, relaciones, atributos) que nos permitían modelar los datos para entender la información que necesitabamos para nuestro objetivo. Este proceso se realizó de manera reiterada a medida que descubríamos nuevos planteamientos de mejora.

El análisis de datos se centró en visualizar y observar patrones, destacando la cantidad de sedes en cada país y explorando la relación con el PBI per cápita. Se implementó el Método GQM (Goal-Question-Metric) con el fin de abordar problemas específicos en cada conjunto de datos y obtener cuantificaciones de su importancia en la calidad de los datos. Adicional a estos problemas particulares, identificamos muchos más que nos impedían estandarizar los datos y trabajar con seguridad sobre los mismos, por ello redactamos un apartado que sintetiza las decisiones que tomamos respecto a ello. Al final de este informe, se presentaron conclusiones basadas en el análisis realizado.

2 Introducción

En el presente informe, se documenta el proceso, realizado por los alumnos, al trabajar con los datasets proporcionados por los docentes de la materia.

Para ello, buscamos definir un objetivo de trabajo claro: *analizar la posible relación existente entre el Producto Bruto Interno (PBI) per cápita de diferentes países y la cantidad de sedes que Argentina posee en cada territorio*

En busca de alcanzar nuestro objetivo, nos encontramos ante el desafio de trabajar con conjuntos de datos que contenian errores, ambiguedades e incluso columnas incompletas, es por ello que nos avocamos a la tarea de comprender cada dataset y específicamente, nuestro objetivo. ¿Qué es lo que estamos analizando? ¿Qué estamos buscando? ¿Qué datos tienen relavancia y cuáles no?
Todo esto, se encuentra documentado en el presente informe, pero primero es importante entender y definir las palabras clave que se nombran en el objetivo.:

¿Qué es el P.B.I?

El PBI, o Producto Interno Bruto, es una medida cuantitativa que representa el valor total de los bienes y servicios producidos dentro de un pais durante un período específico de tiempo. Es uno de los indicadores más importantes y ampliamente utilizados para evaluar la salud económica de un país. Esta medida, se calcula sumando el valor monetario de todos los bienes y servicios finales producidos en un país en un año determinado.

El PBI per cápita, es una medida económica fundamental que refleja el ingreso promedio por habitante en un país. Se calcula dividiendo el Producto Interno Bruto entre la población estimada a mediados de año. A su vez, representa la suma del valor añadido bruto por todos los productores residentes de la economía, incluyendo impuestos sobre productos y excluyendo subvenciones no consideradas en el valor de los productos. Los datos se presentan en dólares actuales de los Estados Unidos.

3 Metodología

Con el propósito de alcanzar nuestro objetivo, comenzamos con la limpieza de los datasets proporcionados. Los mismos estaban llenos de columnas de datos que consideramos irrelevantes a nuestro objetivo, como por ejemplo el código telefónico del país donde se encontraba una sede, su zona horaria o datos de PBI de años previos al que se queria analizar, el 2022; también nos encontramos con datos en inglés y español, como es el caso de el nombre de la ciudad donde se hallaba la sede o su descripción. Para estos ultimos datos, tuvimos en cuenta cual traducción, nos serviría en caso de necesitar buscar en otras tablas.

3.1 Procesamiento de Datos

En paralelo al proceso de limpieza de los datasets, creamos un Diagrama Entidad Relación (DER), que nos permitió modelar la información que queriamos representar y almacenar de acuerdo a nuestro objetivo:

- **Entidad:** Sede

  - **Atributos**:
    - **Id_sede** (Clave primaria) = identificador único de cada sede
    - **Tipo** = Existen distintos tipo de tipos de sede, por ejemplo Consulado o Embajada.
    - **Nombre** = Nombre propio de cada sede.

- **Entidad** : País
  - **Atributos** :
    - **Código** (Clave Primaria) = identificador único de cada sede, basado en el código ISO-3166-1 alfa-3, utilizado para representar países, territorios dependientes y zonas especiales de interés geográfico.
    - **Nombre** = Nombre de cada País
    - **Nivel-Ingreso** = Nivel de Ingreso promedio de los habitantes del país, pudiendo este atributo tomar los valores de : High income, Lower middle income y Upper middle income
    - **Región** = Región geográfica donde se encuentra ubicado el país, por ejemplo : Europe & Central Asia, es la región en la que se encuentra Armenia.
    - **PBI** =Producto Bruto Interno que registró el país en el año 2022

- **Entidad** : Red Social
  - **Atributos**:
    - **URL** (Clave Primaria) = URL (Localizador de Recursos Uniforme), única para cada red social de cada sede.
    - **Nombre** = Nombre de la red social, por ejemplo : Facebook.

- **Entidad debil** : Sección
  - **Atributos**
    - **Id_Sede** (Clave primaria): Identifica a la sede, Clave foránea a Sede.
    - **Descripción**(clave primaria):Refiere al tipo de sección. Por Ejemplo: Sección Administrativa. Junto con el Id_sede, conforman la clave de la entidad.

- **Relaciones** :
  - **está_en** (Sede-Pais) =
    - **Cardinalidad** : 1:N.
    - Una sede está solo en un país, pero un país tiene muchas sedes (o puede no tener ninguna).
  - **cuenta_con**(Sede -RedSocial) =

    - **Cardinalidad** : 1:N
    - Una sede puede tener muchas redes sociales (o no tener ninguna), pero cada cuenta de red social, pertenece a una unica sede.
- **compuesta_de**(Sede-Seccion) =
    - **Cardinalidad** : 1:N
    - Una sede puede tener muchas secciones, pero debe tener al menos un y una seccion pertenece solo a una sede.
    - **Clave Foranea** : 'Id_Sede' constituye parcialmente la clave primaria de la entidad y referencia a 'Id_Sede de la entidad Sede.

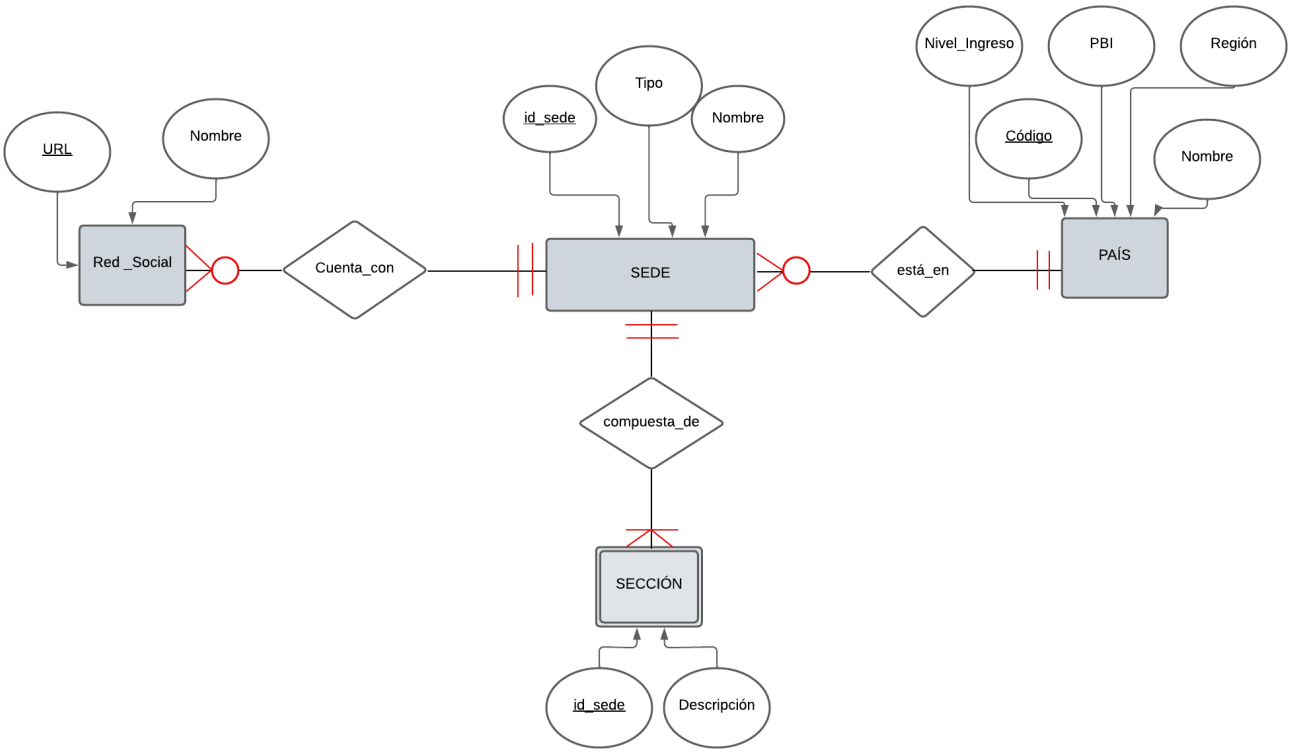


Figure 1: Diagrama Entidad Relación del Problema a modelar

En la siguiente figura 2, se observa el *mapeo* del Diagrama Entidad Relación (1) al Modelo Relacional propuesto por Edgar Codd en 1970.
En el, se plasman las entidades del DER, en un formato de tabla o de *relación*.

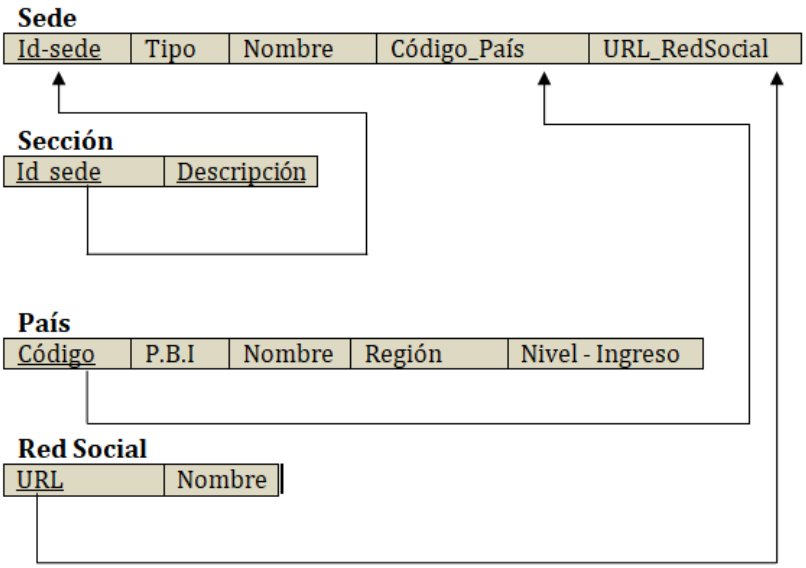


Figure 2: Esquema del Modelo Relacional asociado al DER ¨[1]

Cada relación, tiene subrayada su clave primaria, la cual identifica a cada tupla como única y además, flechas que hacen referencian a las claves foráneas, las cuales establecen relaciones entre las tablas.

- El atributo 'URL_RedSocial' en la tabla Sede,actúa como clave foránea y referencia a la clave primaria 'URL' en la tabla Red Social.
- El atributo 'Código_País' en la tabla Sede, actúa como clave foránea y referencia a la clave primaria 'Código' en la tabla Red Social.
- De igual manera, sucede con 'Id_Sede' en Sede y Sección.

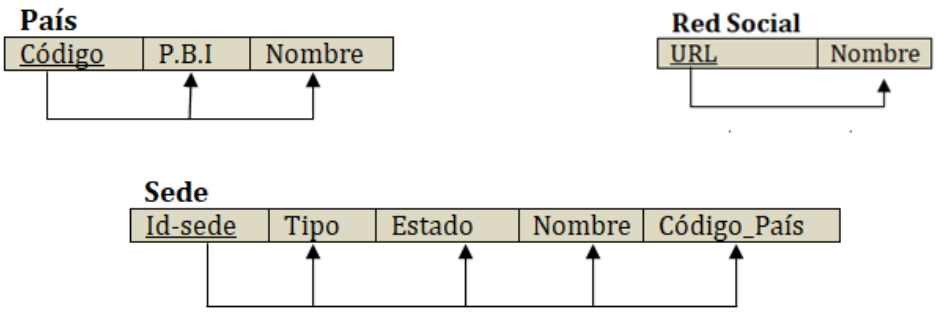


Figure 3: Dependencias Funcionales dadas en el Modelo Relacional [2]

En este esquema 3, podemos apreciar las dependencias funcionales dadas.
Todas ellas son dependencias funcionales completas, debido a que todos los atributos no primos, dependen de forma total de la clave primaria, puesto que existe una sola en cada caso.

Es importante mencionar que, en la Figura 3, no se observa el esquema de dependencias funcionales de la tabla Secciones, esto es debido a que dicha relación se compone únicamente de su clave primaria ({'Id_sede', 'Descripción'}) y no existe dependencia funcional no trivial alguna que destacar.

Basados en el Modelo Relacional creado [2], generamos en Python DataFrames vacios para almacenar y trabajar con la información. Para esta étapa del proyecto, fue donde nos encontramos con más desafíos.
Los problemas surgieron al identificar irregularidades y anomalías en los datasets proporcionados.

Varios de ellos se dieron debido a que las tablas de Representaciones Argentinas en el exterior, no estaban normalizadas, debido a que sus columnas poseian valores no atómicos y la Primera Forma Normal (1FN) no admite valores compuestos ni relaciones dentro de relaciones. Un ejemplo de esto, es la columna Redes Sociales en la Tabla que poseía la información completa de las Sedes de nuestro país en el Exterior. Cada sede, tenia varias cuentas, de distintas redes.
También, fue muy frecuente la presencia de valores "Null", dicho en otras palabra, filas y columnas que no poseían información de su sede correspondiente.

3.1.1 Exploración y Calidad de Datos

Para afrontar los problemas planteados en la sección anterior, nos vimos en la encrucijada de decidir que datos debíamos sacrificar, cuales tenian más importancia que otros y cuáles no eran relevantes para nuestro objetivo.

Por ello, realizamos una exploración de los datos, tabla por tabla, para comprender que datos nos serían de utilidad y cuales no tanto. Este fue un proceso de "ida y vuelta", debido a que al limpiar una tabla, debiamos cotejar en otra la información presentada, verificar datos y revisar incosistencias. Para luego volver a revisar la tabla original, al detectar datos faltantes. Resumimos nuestro proceso de la siguiente manera:

- **Primera exploración de datos**: Dados los dataset originales, se estudió y planteó la posible utilidad de cadda columna. Aquellas que no aportaban datos relevantes o eran redundantes, fueron descartadas
- **Segunda exploración de datos**: identificamos los objetos de nuestro interés: país y sede, juntos con datos que consideramos serían relevantes a nuestro objetivo, como por ejemplo, la clasificación por ingresos.

Además, tuvimos en cuentra que para realizar las consultas del punto H necesitamos conocer los datos de las redes sociales de cada sede y las secciones de cada sede, por lo que también fueron conservados estos datos.
En síntesis, nos quedamos con las siguientes tablas, las cuales fueron renombradas:

  - Sedes:
    * Dataset Original: Representaciones Argentinas, Datos completos Sedes
    * Fuente: Ministerio de Relaciones Exteriores y Culto
    * Atributos: sede_id, sede_desc_ingles, pais_iso_3, sede_tipo, redes_sociales
  - Secciones :
    * Dataset Original: Representaciones Argentinas, Datos secciones Sedes
    * Fuente: Ministerio de Relaciones Exteriores y Culto
    * Atributos: sede_id, sede_desc_castellano
  - pbi :
    * Dataset Original: GPD per cápita (current US$)
    * Fuente: World Bank national accounts data, and OECD National Accounts data files
    * Atributos: Country Code, 2022 PBI
  - Regiones:
    * Dataset Original: GPD per cápita (current US$)
    * Fuente: World Bank national accounts data, and OECD National Accounts data files
    * Atributos: Country Code, Region, IncomeGroup, TableName

- **Tercera exploración de datos**: Dado que Argentina no cuenta con sedes en todos los países registrados por el Banco Mundial, en la tabla de PBI y Región, filtramos aquellos que sí tienen sede de Argentina utilizando su código de país, verificando que exista en la tabla de sedes.Durante ese proceso, fueron encontrados 2 casos excepcionales: El primero se trata de un error en el código del país, en la tabla sedes, United Kingdom tiene código GRB y en la tablas PBI y Regiones, tiene código GBR. Este error se solucionó manualmente, reescribiendo el código en la tabla sedes como GBR. El segundo se trata de una sede ubicada en el Vaticano, cuyo codigo de país VAT no figura en las tablas del Banco Mundial, dado que no es representativo se resolvió eliminando la tupla correspondiente.

- **Cuarta exploración de datos**: Como última fase para estandarizar los datos, se observa que no hay datos del PBI de Venezuela, Siria y Líbano (VEN, SYR, LBN). Como se tratan de caso excepcionales, eliminamos las tuplas correspondientes. Además, eliminamos la tupla de la única sede que se encuentra en Argentina, puesto que nos interesan las sedes en el exterior.

- **Quinta exploración de datos:** realizamos visualizaciones para buscar patrones y lograr comprender mejor los datos.

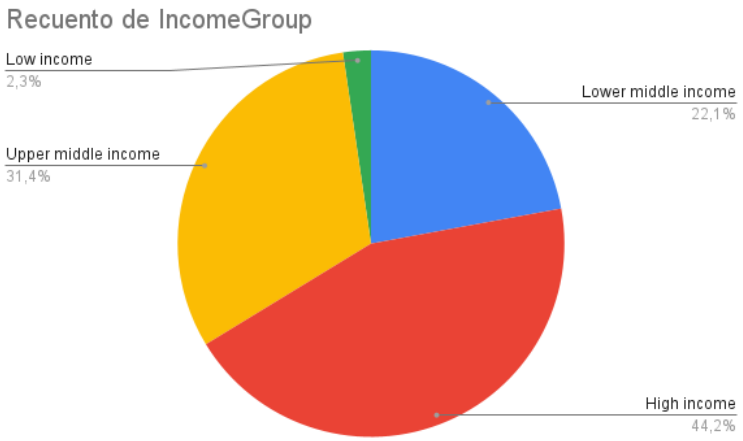


Figure 4: De la tabla Regiones, se visualiza el porcentaje de países que pertenece a cada grupo de nivel de ingresos

Simultáneamente, al analizar cada dataset y procesar que columnas servirían y cuáles no aportaban nada en lo absoluto, implementamos el llamado *Método GQM* (Goal-Question-Metric).
A continuación, se pueden observar los problemas particulares detectados en cada dataset.

Tabla Original	Problema Particular	Atributo de calidad afectado	Causa (clasificación
Datos completos sedes	Hay atributos multivaluados (redes_sociales)	Consistencia	Instancia y modelo de datos
API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73.csv	Datos innecesarios para el objetivo, solo necesitamos los datos correspondientes al año 2022 y de los países donde Argentina tiene sede.	Relevancia	Modelo de datos
Metadata_Country_API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73.csv	Atributos con valores vacíos, por ejemplo: faltan el dato de nivel de ingresos de Venezuela	Completitud	Instancia y modelo de datos

Figure 5: Tabla con los problemas identificados por dataset

En la siguiente tabla, se encuentra el método GQM, aplicado al problema particular detectado.
En el caso del dataset *secciones*, los problemas detectados son iguales a los ya planteados, afectando a los mismos atributos de calidad. Este no fue incluido en las tablas, ya que en las consignas de trabajo, se solicita específicamente, mencionar problemas distintos.

Tabla Original	Goal	Question	Metric	Criterio de Corrección	Impacto en la calidad
Datos completos sedes (sin las tuplas cuyo dato en redes_sociales sea vacío o no contenga @ o formato de URL)	El dato correspondiente a redes_sociales para cada sede sea consistente (uniforme y coherente)	¿Cuál es el porcentaje de sedes que en el dato correspondiente a redes_sociales tienen más de una red social?	80 sedes registradas con más de una red social. 121 sedes registradas -> ≈66%	Se separan las redes en un mismo dato para crear una tupla por cada una de ellas (normalización a 1FN) utilizando el delimitador de la tabla original (' // ')	El porcentaje de sedes con más de una red social en el mismo dato pasa a ser 0%
API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73.csv (sin las columnas de años distintos a 2022. Tabla que	Las tuplas solo deben contener información de los países en donde Argentina tiene sede	¿Qué porcentaje de tuplas tiene como dato en Country Code el código de un país donde Argentina tiene sede?	86 registros con países con sede de Argentina / 265 registros -> ≈32.5%	Eliminamos las tuplas cuyo Country Code no esté en ninguno de los valores de pais_iso_3 de la tabla Datos completos sedes.	Mejora la métrica a un 100% de registros con países con sede de Argentina
Metadata_Country_API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73.csv	Todos los registros deben tener valor no vacío en la columna IncomeGroup	¿Cuál es el porcentaje de registros con valor vacío en IncomeGroup?	49 registros con valor vacío en IncomeGroup / 265 registros -> ≈18.5%	Eliminar aquellos registros cuyo valor en IncomeGroup es vacío, dado que no representan la realidad a modelar	El porcentaje desciende a 0% de registros con valor vacío en IncomeGroup

Figure 6: Tabla con el GQM de cada dataset

3.2 Decisiones tomadas

Manteniendo siempre el foco en analizar la relación existente entre el PBI per cápita de un país y el número de sedes Argentinas en el, fue que tomamos las siguientes decisiones:

- Deshacernos de columnas que aportaban información no necesaria. Por ejemplo, código postal.
- Borrar filas donde las sedes no tenian dato alguno.
- Eliminar filas con valores repetidos o que hacian alusión a lo mismo. Por ejemplo, el código del pais. Para este caso en particular, tuvimos en cuenta que el uso deL código ISO-3, era el usado en las otras tablas de las cuales necesitabamos información.
- Considerar casos aislados.Al analizar la columna ”Redes Sociales” del dataset ”lista-sedes-datos”, fue donde nos encontramos con uno de los mayores problemas de nuestro analisís. Había Sedes que no tenian redes, las cuales fueron eliminadas, algunas que poseían el correo electrónico ”gmail”, las cuales también fueron eliminadas debido a que es un servicio de mensajería electrónica y no una red social, y otras que solo poseían el nombre de usuario de la red. Para este último caso, se aislaron un total de 37 casos de sedes, los cuales fueron verificados, buscados ”a mano” en el sitio web de Embajadas y Consulados [5] y por último modificados en el dataset.

3.3 Análisis de datos

Para analizar exhaustivamente todos los datos recopilados hasta el momento, llevamos a cabo consultas SQL y utilizamos herramientas de visualización. Esto nos permitió interpretar los datos y realizar un análisis adecuado para intentar demostrar el objetivo principal de nuestra investigación.

País	Sedes	secciones promedio	PBI per Cápita 2022 (U$S)
Brazil	10	2.5	8917,674911
Spain	7	1.85714	29674,54429
United States	7	3.42857	76329,58227
Chile	5	1.4	15355,47974
Uruguay	5	1.6	20795,04235
Bolivia	4	1.75	3600,121635
Italy	4	2.25	34776,42323
Paraguay	3	1.66667	6153,055657
Australia	2	1.5	65099,84591
Belgium	2	3.5	49926,82543
Canada	2	3.5	54917,66252
France	2	1.5	40886,25327
Germany	2	1.5	48717,99114

Figure 7: Tabla de Relación entre Cantidad de Sedes y Secciónes con el PBI per cápita del País

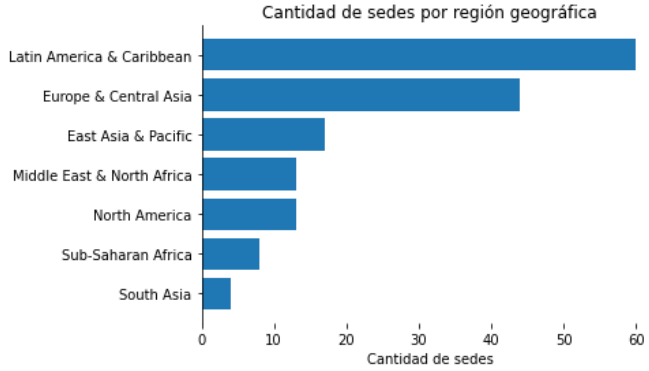
En esta figura, que representa solo una fracción de la tabla completa, se aprecia que no existe una clara relación entre la cantidad de sedes y el PBI per cápita del país.

Región geográfica	Países Con Sedes Argentinas	Promedio PBI per Cápita 2022 (U$S)
North America	2	65623.6
Europe & Central Asia	27	38220.7
East Asia & Pacific	11	27876.9
Middle East & North Africa	12	24902.5
Latin America & Caribbean	24	11851.6
Sub-Saharan Africa	7	2459.07
South Asia	3	2229.36

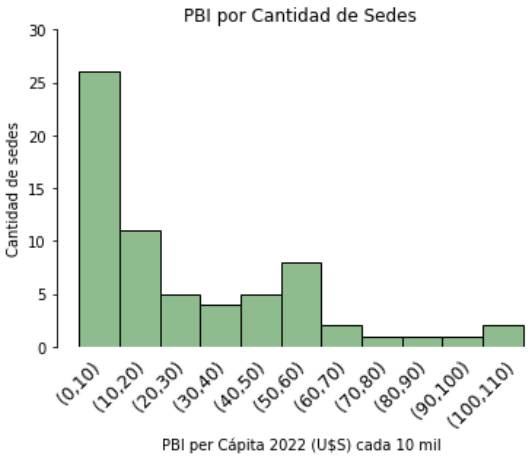
Figure 8: Tabla de Relación entre Sedes y PBI per capita, agrupada por región

En vista de los datos presentados en esta tabla adicional, la incertidumbre respecto

a la pregunta de nuestro objetivo principal se intensifica aún más. Aún así, esta tabla puede resultar muy engañosa, dado que algunas regiones contienen un número muy reducido de países, como es el caso de Norteamérica, lo que lleva a que diga que en la región con el mayor PBI, tiene solo sedes, pero no indica que hay solo dos países en dicha región.



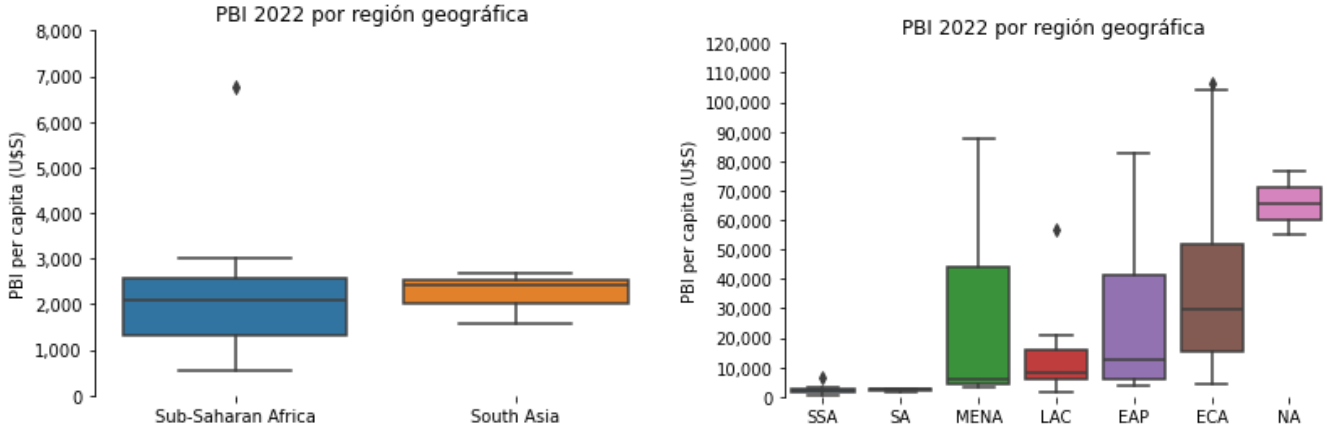
(a) Gráfico con la cantidad de sedes por region



(b) Gráfico con la cantidad de sedes de acuerdo al PBI

Figure 9: Gráficos con la cantidad de sedes según distintos criterios

En el gráfico 9a se aprecia que las regiones con mayor cantidad de sedes son tanto Latinoamérica como Europa. Este gráfico, junto con la tabla 8, permite observar que Latinoamérica es la región con más sedes, pero también con uno de los PBI más bajos. Por último, el gráfico 9b muestra claramente que cuanto menor PBI, tiende a ser mayor la cantidad de sedes. Todo esto sumado, parece indicar que existe relación entre el PBI per capita y la cantidad de sedes, y es que a cuanto menor PBI, mayor cantidad de sedes. Es importante aclarar que en 9b, los resultados del gráfico son acumulativos, cada bin (barra) representa la cantidad de sedes que se obtiene al sumar el numero de sedes que tienen los paises clasificados por su nivel de PBI.



(a) (b)

Figure 10: Gráficos de PBI por región geográfica

En 10b, los acrónimos hacen referencia a las siguientes regiones:

- SSA = Sub-Saharan Africa
- SA = South Asia
- MENA =Middle East & North Africa
- LAC = Latin America & Caribbean
- EAP = East Asia & Pacific
- NA =North America

En los boxplot [10] (o también llamados diagramas de caja), se encuentran plasmados los datos de PBI del año 2022 clasificados por región geográfica y ordenados por mediana. En 10b, podemos observar como los valores de PBI de Norte America, son mayores a los de las demás regiones, siguiendo nuestro objetivo, esto nos lleva a relacionarlo con 9a, donde Norte América es una de las regiones que menos sedes tiene. Siguiendo
También observamos, que la región Latinoamerica y Caribe tiene una mediana que se ubica bastante más por debajo de la de Norte America y en 9 es la región que más sedes tiene.
Por lo tanto, no se observa a simple vista, una relación lineal entre el PBI per capita de una región y la cantidad de sedes en ella.

4 Conclusiones

Recordemos que nuestro objetivo era *analizar la posible relación existente entre el Producto Bruto Interno (PBI) per cápita de diferentes paises y la cantidad de sedes que Argentina posee en cada territorio*.
Tras un exhaustivo análisis, llegamos a la conclusión de que si bien pareciera ser que a menor PBI, mayor es la cantidad de sedes debido a la figura 9b, no es posible afirmar

que se da una relación lineal entre ambos factores, ya que la figura muestra como a medida que aumenta el PBI, inicialmente disminuye la cantidad de sedes, pero luego vuelve a aumentar. Por los gráficos 9, podemos intuir que el factor de cercanía del país con Argentina influye también en la aparición de sedes en él.

Junto con este informe, adjuntamos una carpeta llamada *Reportes*, la cual contiene archivos csv con los reportes generados con las consultas sql, en los cuales basamos nuestro análisis y fueron comentados en la sección Análisis de Datos [3.3].

5 Fuentes

Las siguientes fuentes, fueron proporcionadas por los docentes de la materias, siendo las bases de datos de dominio público.

- **PBI per cápita de los países :**
  `https://data.worldbank.org/indicator/NY.GDP.PCAP.CD`
- **Representaciones Argentinas :**
  `https://datos.gob.ar/dataset/exterior-representaciones-argentinas`

Fuentes consultadas:

- **Guías de aprendizaje para QMF:**
  `https://www.ibm.com/docs/es/qmf/12.1.0?topic=cics-tutorials-qmf-tso`
- **Python:**
  `https://docs.python.org/3/tutorial/index.html`
- **Embajadas y Consulados**
  `https://cancilleria.gob.ar/es/representaciones`
- **Uso de Overleaf-Latex :**
  `https://www.overleaf.com/learn`
- **Fuente consultada para acortar miles en gráficos:**
  `freecodecamp.org/espanol/news/tutorial-de-f-strings-en-python-formato-de-cadenas-en-python-explicado-con-ejemplos/`