LABORATORIO DE DATOS

TRABAJO PRÁCTICO 01

Camila Molteni Ceccón

Luna Praino

Agustin Quintriqueo

SECCIÓN RESUMEN

En nuestro trabajo procuramos responder la problemática orientada a determinar si existe una relación entre la cantidad de sedes en el exterior que posee la Argentina en cada país y el flujo monetario neto de Inversión Extranjera Directa(IED) de ese país en nuestro país.

Con el fin de conseguir nuestro objetivo, descargamos las tablas con los datos correspondientes, las analizamos e hicimos reportes sobre la información que brindaban y al final graficamos los datos para su mayor comprensión.

Una vez adquirido el conocimiento sobre el objetivo, concluimos que si bien existe cierta relación entre los parámetros tomados, la misma no es muy específica pues no encontramos una proporción directa entre cantidad de sedes por país y el flujo monetario neto IED del mismo en nuestro país.

SECCIÓN INTRODUCCIÓN

El objetivo general del trabajo es determinar si existe una relación entre los flujos monetarios netos de Inversión Extranjera Directa (IED) de cada país, y la cantidad de sedes en el exterior que tiene Argentina en dicho país.

Para alcanzar dicho objetivo analizaremos la fuente de datos abiertos correspondiente a las Representaciones Argentinas en el exterior de la República Argentina y las inversiones extranjeras. Este mismo análisis lo haremos teniendo en cuenta las dependencias funcionales de cada tabla dada, la calidad de datos de ciertos atributos que nos interesen, y generando reportes mediante consultas de SQL. Estas consideraciones serán hechas en base a un DER realizado y a un cierto modelo relacional desprendido del mismo, donde se reflejen los datos necesarios para resolver la problemática.

SECCIÓN PROCESAMIENTO DE DATOS

En primer lugar vimos en qué forma normal se encuentran las tres tablas de “Representaciones Argentinas”, las cuales son: sedes_datos, lista_sedes y lista_secciones.

En el caso de sedes_datos, asumimos que la tabla no se encuentra en 1FN porque en el dominio del atributo “redes sociales” se encuentran valores multivaluados y por consecuencia, tampoco está en 2FN y 3FN.

En lista_sedes, pudimos observar que si se encuentra en 1FN debido a que todos los atributos de la tabla tienen un dominio formado por valores atómicos. También está en 2FN porque si consideramos a “sede_id” como clave, no hay columna que dependa parcialmente de la misma. Aun así, viendo la dependencia funcional “sede_desc_castellano” -> “sede_tipo”, la tabla no se considera que esté en 3FN porque “sede_desc_castellano” no es SK de la entidad y “sede_tipo” no es atributo primo en lista_sedes.

Por último, en lista_secciones , pudimos notar que no se encuentra en 1FN porque en la columna “telefono_principal” hay valores multivaluados y por ende, tampoco está en 2FN ni 3FN.

Para abordar los atributos afectados en las tablas, los problemas de las mismas y el modelo de calidad, usamos “flujos_monetarios”, “lista_secciones” y “lista_sedes_datos”. A continuación, mostramos lo enumerado anteriormente.

- **Secciones:** encontramos valores null en muchas columnas de la tabla e inconsistencia en los datos. Por ejemplo, en “sede_desc_ingles” hay datos en español o en “correo_electronico” hay valores nulos.

En esta oportunidad, analizaremos en profundidad los valores nulos. Creemos que el atributo de calidad afectado es la completitud y el problema corresponde al modelo de datos.

Para la realización del GQM utilizamos en particular la columna “correo_electronico”, queremos analizar la completitud de la misma:

GOAL: dato correspondiente al correo electrónico de cada sección.

QUESTION: ¿cuál es la proporción de “correo_electrónico” que tienen el valor null?

METRIC: $\frac{cantidad\ de\ registros\ con\ campo\ 'correo\ electrónico'\ vacío}{cantidad\ total\ de\ registros\ en\ esta\ columna} = \frac{12}{516} = 0.023$

Lo que podemos deducir del resultado de la métrica realizada (el cual es bajo) es que la calidad de datos respecto de completitud es buena. Para mejorar aún más la métrica, lo que hicimos fue cambiar los valores nulos por ceros.

Una vez mejorada la calidad del dato notamos una gran diferencia, debido a que antes con los valores nulos no se contaban ciertas filas y cuando lo reemplazamos con un cero si. Lo mismo da como resultado que la métrica sea excelente.

- **Flujos monetarios**: encontramos nombres de países mal escritos y valores null.

Analizamos aquellos casos en que los países tienen sus nombres mal escritos. Creemos que el atributo de calidad afectado es la consistencia de los datos y el problema corresponde a la instancia.

Para la realización del GQM utilizamos en particular la columna “paises”, donde queremos analizar la consistencia de sus datos:

GOAL: determinar cuántos nombres mal escritos hay en esta columna.

QUESTION: ¿Cuál es la proporción de países que tienen el dato erróneo?

METRIC: $\frac{cantidad\ de\ registros\ con\ campo\ 'paises'\ erróneo}{cantidad\ total\ de\ registros\ en\ esta\ columna} = \frac{3}{173} = 0.017$

Lo que podemos deducir del resultado de la métrica realizada es que la proporción de países con su nombre mal escrito es muy baja, por lo que se podría concluir en que la columna mencionada cuenta con la mayoría de sus datos correctos, es decir, la calidad de datos es alta teniendo en cuenta la consistencia. Para corregir estos datos erróneos sobre los nombres de los países, fuimos cambiando caso por caso, de manera que todos estén bien escritos.

Una vez mejorada la métrica, observamos que el impacto en la calidad de datos con respecto a la proporción de países con su nombre mal escrito se volvió cero. Es decir, la consistencia se volvió alta.

- **Sedes Datos:** en esta tabla encontramos valores null, valores multivaluados e inconsistencia en los datos junto con poca disponibilidad de los mismos, como por ejemplo, en “código postal”, se encuentran valores como: números, palabras (“no existe”, “s/c”), guiones y nulls.

Para la realización del GQM utilizamos en particular la columna “codigo_postal”, donde queremos analizar la disponibilidad de los mismos relacionados a cada sede:

GOAL: dato relacionado al código postal de cada dirección de su sede.

QUESTION: ¿cuál es la proporción de datos en “codigo_postal” que no están disponibles en la columna?

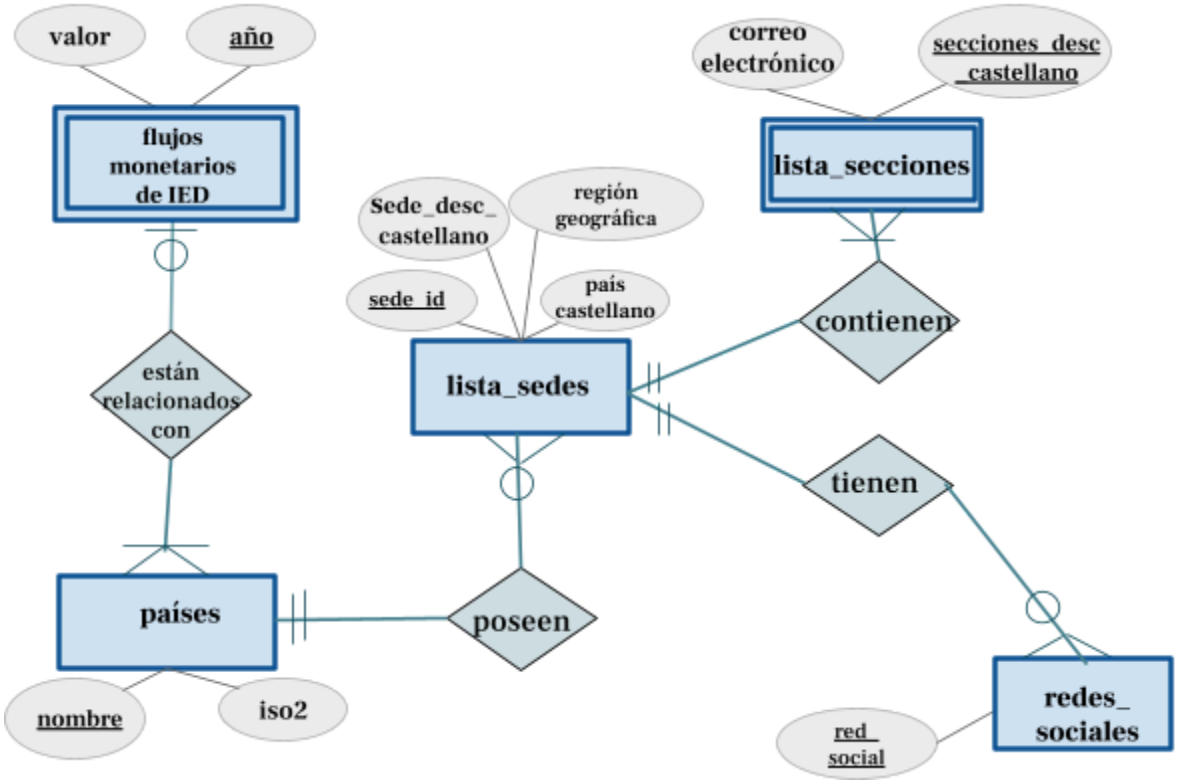
METRIC: $\frac{cantidad\ de\ registros\ con\ campo\ 'código\ postal'\ distinto\ a\ su\ código\ postal\ perteneciente}{cantidad\ total\ de\ registros\ en\ la\ columna}$

$= \frac{14}{163} = 0.085$

Lo que podemos deducir de la métrica realizada es que la calidad de datos respecto al atributo mencionado es buena. Luego para corregir los datos erróneos, se podría unificar todos los casos, poniéndoles un cero.

Una vez corregida la métrica, obtendríamos que la misma es muy buena ya que su valor sería igual a cero.

Siguiendo con el procesamiento de datos, ahora abordaremos el Diagrama de Entidad Relación (DER) para contemplar el modelado de datos necesarios teniendo en cuenta el objetivo de nuestra problemática.

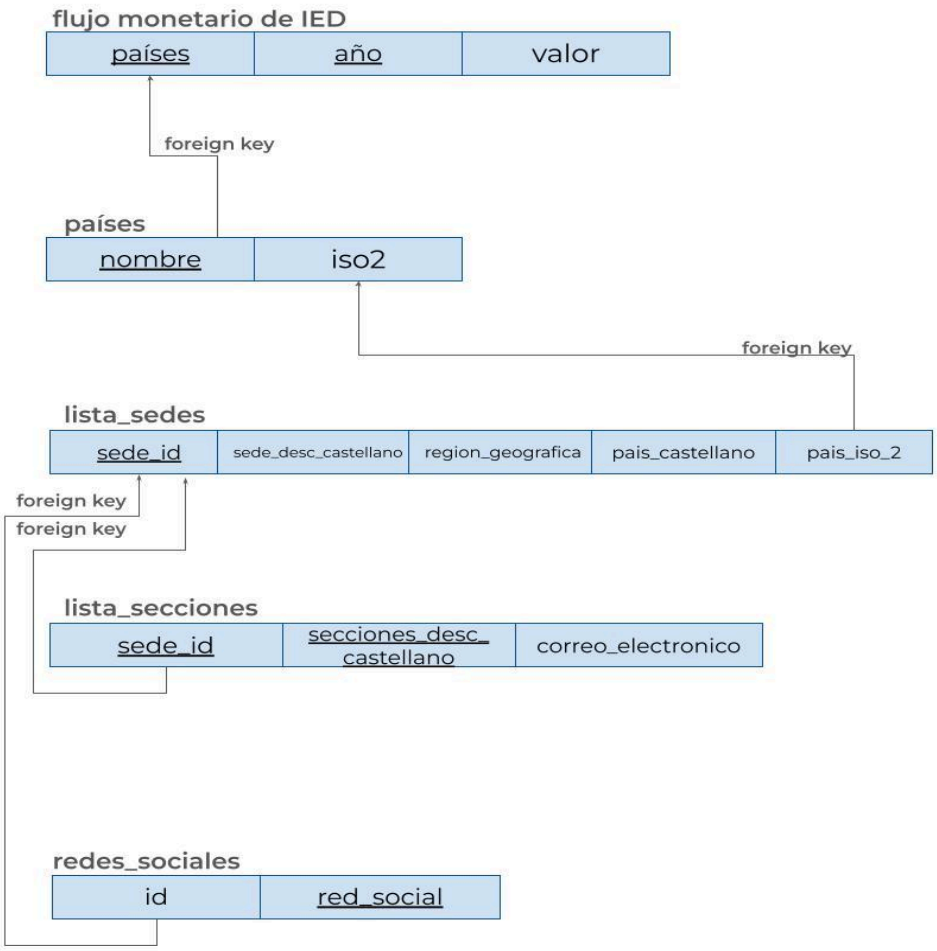


Suposiciones que hicimos a la hora de realizar el DER:

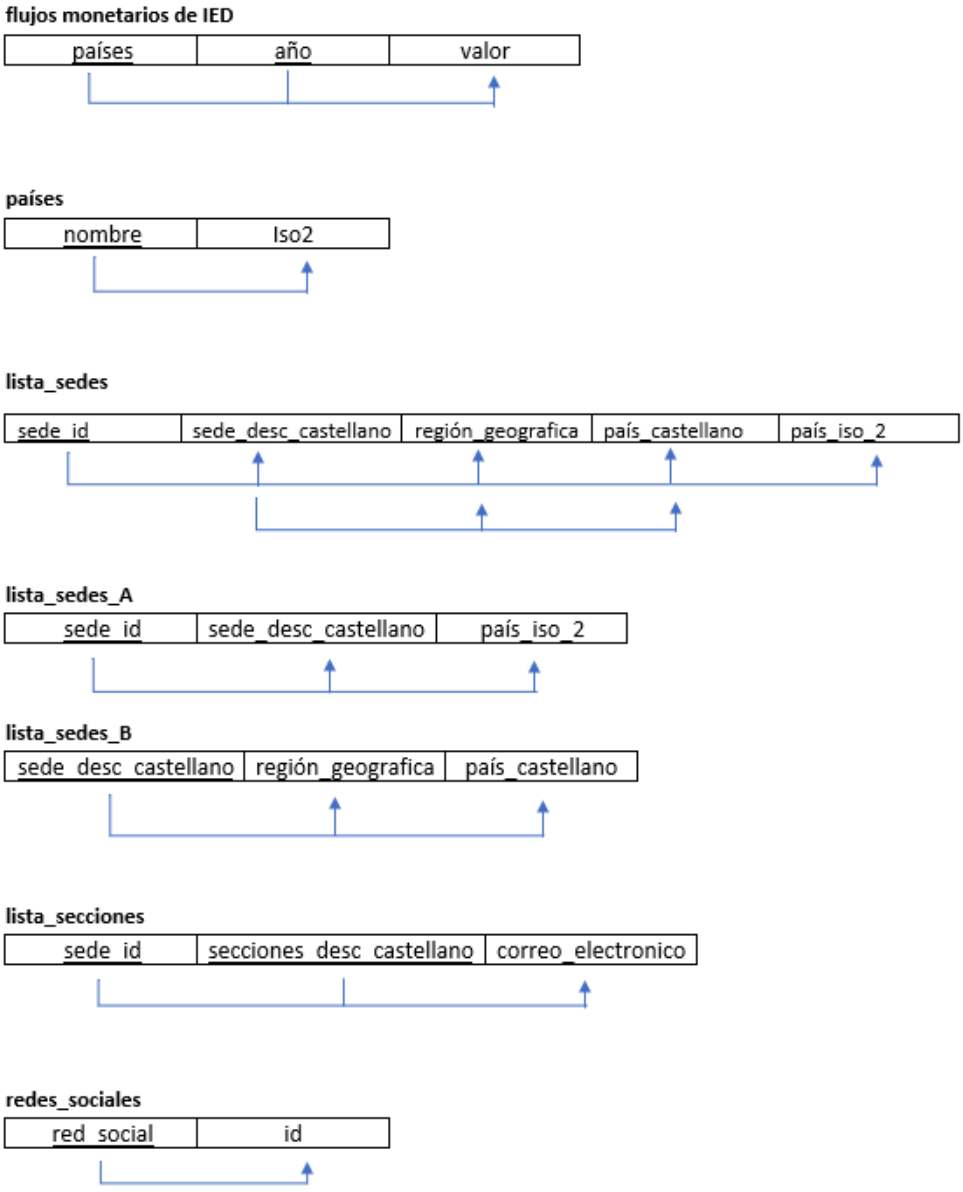
- En primer lugar consideramos que fue necesario unificar la tabla sedes y sedes datos ya que notamos que la segunda tabla era una derivación de sedes.
- En segundo lugar creamos una entidad aparte sobre las redes sociales la cual tiene como atributos los links de cada una de ellas.
- En tercer lugar, consideramos que existen dos entidades débiles, una llamada “flujos monetarios de IED” que depende de “países”, y otra llamada “secciones” que depende de “sedes”.
- En último lugar, nos parece que las cardinalidades correspondientes a cada relación entre entidades son:

- ➔ “Los flujos monetarios de IED están relacionados con varios países o al menos uno”. Además, “Los países son relacionados con uno o ningún flujo monetario”.
- ➔ “Los países poseen varias o ninguna sede”. Además, “las sedes poseen un único país”.
- ➔ “Las sedes contienen varias o al menos una sección”. Además,”Las secciones están contenidas por una única sede”.
- ➔ “Las sedes tienen varias o ninguna red social”.Además, “Las redes sociales tienen un única sede”.

Siguiendo con el procesamiento, plasmaremos el Modelo Relacional que realizamos.



Y en base al modelo relacional plasmado, a continuación dejamos exhibidas las dependencias funcionales asociadas a cada uno de sus esquemas.



En base a la imagen, podemos notar que para cada esquema del modelo, se vio reflejada la dependencia total de los atributos no primos sobre su correspondiente clave, por lo que cada tabla está en segunda forma normal.

Además, para cada esquema se vio evidenciada la condición (en este caso) de que si existe una dependencia funcional de la forma “X -> A ” entonces X es SK del esquema. Excepto, para “lista_sedes” la cual fue descompuesta en “lista_sedes_A” y “lista_sedes_B”, tablas las cuales si se encuentran en tercera forma normal, junto con el resto.

SECCIÓN DECISIONES TOMADAS

1. En primer lugar tomamos como decisión, que por razones de prolijidad a la hora de realizar el código hicimos dos archivos.py.

   El primero que se debe cargar se llama ”t1_liempiza_de_tablas”, en el mismo se encontrará el código correspondiente a las métricas(GQM), la corrección de errores y la selección de columnas para cada una de las tablas que consideramos necesarias para la problemática del trabajo.

   El segundo se llama “t1_codigo_consultasSQL_y_visualizacion”, en el mismo se encontrará el código que hace referencia a los reportes (consultas SQL) y además las herramientas de visualización (gráficos) correspondientes.

2. En base a la métrica que realizamos sobre la tabla “lista_sedes_datos” si bien, dejamos constancia de lo que haríamos para resolver los errores del atributo mencionado en el punto y por consecuencia, que mejore su métrica. Como para nuestro propósito no era relevante esa columna y además lo consultamos con los profesores, decidimos no codificar esa resolución.
3. En último lugar, modificamos algunos nombres del modelo relacional para mayor practicidad en la futura codificación, entre ellos:
   - En la tabla flujo monetario de IED en lugar de poner “paises.nombre”, pusimos “nombre” para la FK.
   - En la tabla “lista_sedes” cambiamos “paises.iso2” y decidimos poner “pais_iso_2”.
   - En la tabla “lista_secciones” cambiamos “Lista_sedes.Sede_id” y decidimos poner “sede_id”.
   - En la tabla “redes_sociales” cambiamos “Lista_sedes.sede_id” por “id”.

SECCIÓN ANÁLISIS DE DATOS

Pasando a un análisis más profundo de las tablas utilizadas durante toda la experiencia, llevamos a cabo una serie de consultas de SQL y herramientas de visualización para relacionar diferentes datos encontrados en las mismas.

En la primera consulta informamos la cantidad de sedes y secciones promedio de las mismas para cada país, además de su flujo monetario neto de IED del año 2022.

La segunda consulta consta de dos incisos. En el primero de ellos reportamos la cantidad de países en los cuales argentina contempla una o más sedes, y estos mismos fueron agrupados por región geográfica.

Índice	region_geografica	Paises_con_sedes_argentinas
0	AMÉRICA CENTRAL Y CARIBE	14
1	AMÉRICA DEL NORTE	16
2	AMÉRICA DEL SUR	44
3	ASIA	30
4	EUROPA CENTRAL Y ORIENTAL	8
5	EUROPA OCCIDENTAL	35
6	OCEANÍA	3
7	ÁFRICA DEL NORTE Y CERCANO ORIENTE	5
8	ÁFRICA SUBSAHARIANA	8

Luego, en el siguiente inciso se tuvo en cuenta la misma tabla del resultado pero, con la diferencia de que se relaciono cada región geográfica con el promedio del IED del año 2022

Índice	region_geografica	Paises_con_sedes_argentinas	Promedio_IED_2022
0	AMÉRICA DEL NORTE	16	168845
1	OCEANÍA	3	34584.1
2	ASIA	30	28983.7
3	AMÉRICA DEL SUR	44	15564.2
4	EUROPA CENTRAL Y ORIENTAL	8	9687.43
5	EUROPA OCCIDENTAL	35	7693.2
6	ÁFRICA DEL NORTE Y CERCANO ORIENTE	5	2868.69
7	AMÉRICA CENTRAL Y CARIBE	14	1366.44
8	ÁFRICA SUBSAHARIANA	8	1340.52

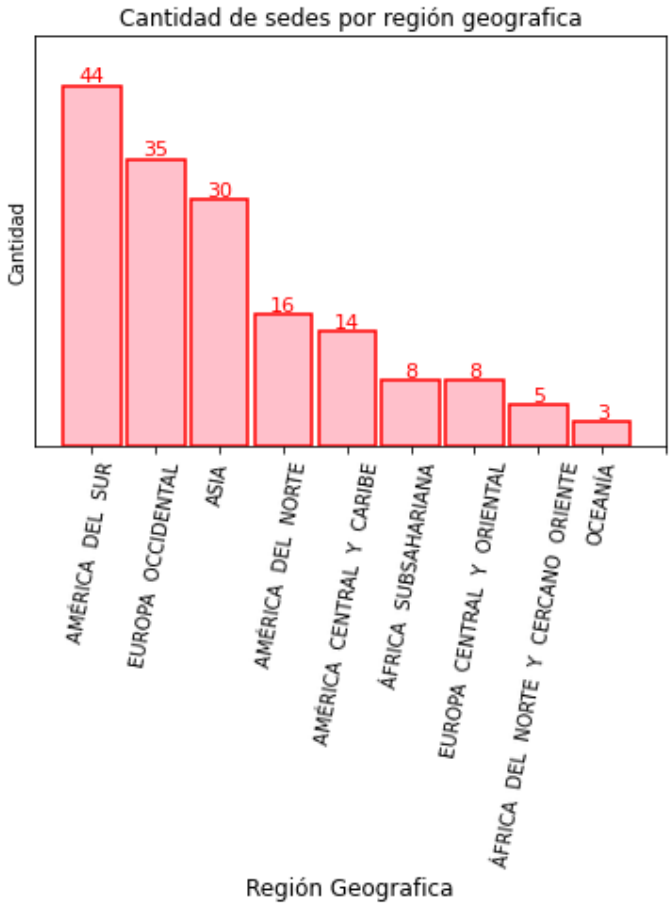
En la tercera consulta buscamos evidenciar las diferentes redes sociales utilizadas en las sedes por cada país.

Por último, en la cuarta consulta detallamos las redes sociales utilizadas por cada sede, incluyendo el país, la sede, el tipo y la URL asociada. En el caso donde su URL no estaba descripta como tal, sino como aquello que podría asumirse como un nombre de usuario, decidimos en las dos últimas columnas nombradas, dejar ese “nombre de usuario” correspondiente.

Todas las consultas pueden verse detalladas en el código adjunto con el presente informe.

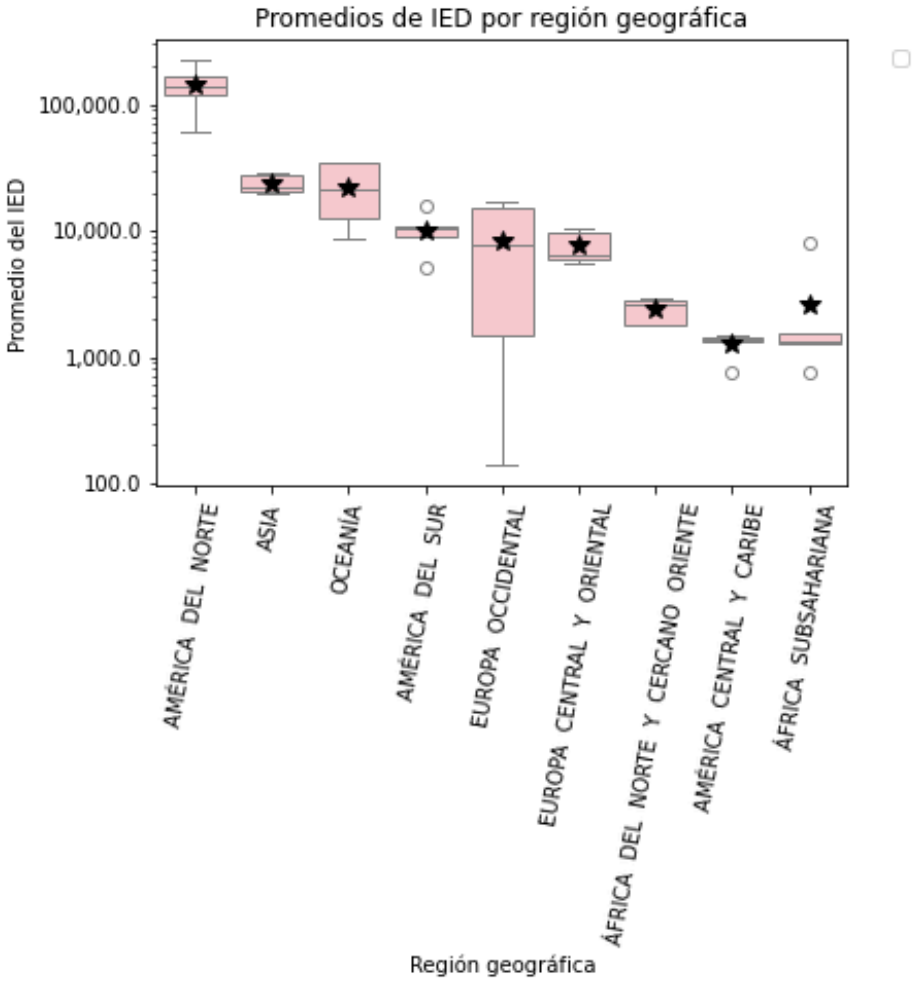
Además, para visualizar las tablas de las consultas 1,3 y 4, como tienen muchas filas, las pusimos como un archivo.csv en un anexo. El mismo es una carpeta que se encuentra dentro de “TP01-CPQ.zip” con el nombre de “Anexo”.

Luego, a partir de las herramientas de visualización conocidas, pasamos a vincular diferentes datos. Incluímos a continuación los gráficos obtenidos.



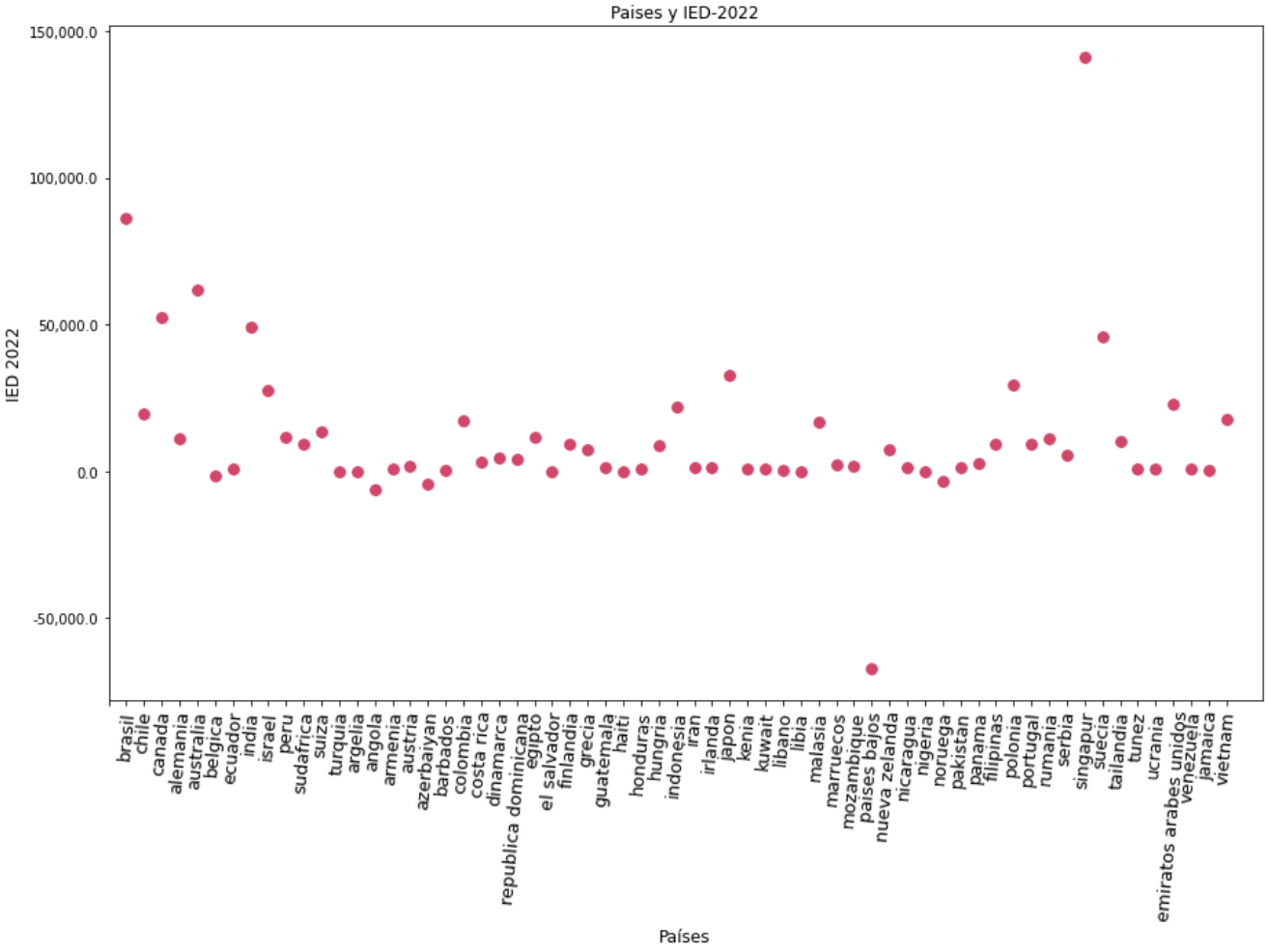
En este caso buscábamos ver la cantidad de sedes por región geográfica, ordenados de manera decreciente. Para esto, optamos por un gráfico de barras, donde en el eje vertical (y) tenemos el total de sedes correspondientes a las regiones relativas al eje horizontal (x). Podemos observar claramente que América del Sur es quien tiene mayor cantidad de sedes y Oceanía, la menor. Además que hay dos regiones que tienen la misma cantidad de sedes, las mismas son África Subsahariana y Europa central y oriental.

Posteriormente, utilizamos un boxplot para evidenciar, por cada región geográfica, el valor correspondiente al promedio de IED de cada una, abarcando el intervalo de años 2018-2022, de los países donde Argentina tiene una sede. Mostramos todos los boxplot en una misma figura para poder realizar comparaciones.



En este caso pudimos ver claramente que la mediana mayor del promedio de IED en este período fue, América del Norte con respecto a todas las regiones y, particularmente, en comparación a África Subsahariana.

Finalmente, para relacionar el IED de cada país en el año 2022 y la cantidad de sedes en el exterior que tiene Argentina en dichos países, usamos un diagrama de puntos.



Esta vez analizamos los datos de 63 países. Se puede ver claramente que hay dos valores atípicos, es decir, que están bastante alejados del resto del IED de cada país (un valor menor a -50000 y otro mayor a 100000). Sin embargo, en general, se mantienen en un rango de IED mayor a -50000 y menor a 50000.

CONCLUSIÓN

A manera de conclusión podemos decir que a lo largo del análisis realizado, si bien hemos encontrado cierta relación entre IED y las sedes de la Argentina en el exterior, lo cierto es que esta no es proporcionada.

Podemos evidenciar tal situación con los resultados arrojados por la tabla “flujo_monetarialimpia” a la que nos remitimos.

Ahora bien si tomamos un caso en particular, en el año 2022, encontramos una relación donde los países que tienen entre una y dos sedes, tienen un flujo monetario similar, a excepción de algunos casos como el de Brasil (con once sedes y un IED regular o bajo) o Singapur (con el mayor flujo monetario en ese año a pesar de tener una sola sede).

Por lo tanto, si bien se puede encontrar cierta relación, no hay un vínculo tan marcado entre los países, su cantidad de sedes y su flujo monetario neto de Inversión Extranjera Directa.

Para finalizar, como equipo consideramos que hemos podido cumplir el objeto sin adicionar mayor información.