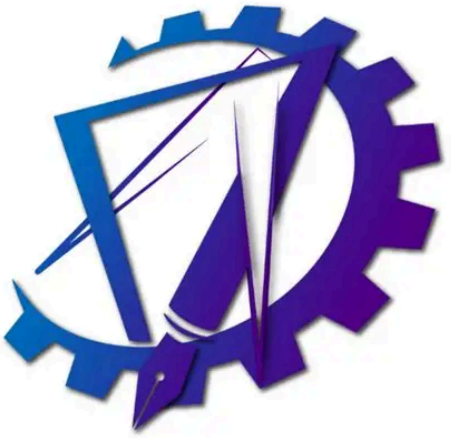


Instituto Industrial Luis A. Huergo
Desarrollo de Sistemas 2025
Cuatrimestre Nº2
Trabajo Práctico N°1

Alumnos: Santino Nicolas ANDREATTA , Pedro VILLARINO & Galo FERNANDEZ ACHILLE

Índice

Resumen

En este trabajo práctico analizamos la relación entre el PBI per cápita de los países y la cantidad de sedes diplomáticas argentinas en el exterior. Se procesaron datos del Banco Mundial sobre PBI per cápita en 2023 y datos del Ministerio de Relaciones Exteriores sobre representaciones argentinas en el exterior.

El análisis incluye la limpieza y procesamiento de datos, generación de reportes estadísticos y visualizaciones que permiten comprender la distribución geográfica de las sedes argentinas y su relación con indicadores económicos. Los resultados muestran una correlación débil positiva (0.107) entre el PBI per cápita y la cantidad de sedes, sugiriendo que factores políticos, históricos y estratégicos influyen más en la ubicación de las representaciones diplomáticas que el nivel económico de los países.

Introducción

Problema a resolver

Trabajamos con fuentes de datos abiertos sobre las Representaciones Argentinas en el exterior y el PBI de los países. El objetivo principal es determinar si existe una relación entre el PBI per cápita de cada país (año 2023) y la cantidad de sedes diplomáticas que Argentina mantiene en dicho país.

Los datasets los sacamos de las siguientes fuentes:

- https://datos.gob.ar/dataset/exterior-representaciones-argentinas.
- https://data.worldbank.org/indicator/NY.GDP.PCAP.CD.

Objetivo general

Analizar la relación entre el PBI per cápita de los países y la cantidad de sedes diplomáticas argentinas en el exterior, mediante el procesamiento de datos, generación de reportes estadísticos y visualizaciones que permitan comprender los patrones de distribución geográfica de las representaciones argentinas.

Actividades realizadas

1. **Limpieza de datos**: Limpieza, normalización de datos y corrección de problemas en los datasets
2. **Creación de diagramas e importación de datos:** Creamos diagramas conceptuales e Importamos los datos importantes a data frames creados por nosotros mismos
3. **Generación de reportes**: Creación de 4 reportes estadísticos específicos
4. **Visualizaciones**: Desarrollo de gráficos para análisis exploratorio
5. **Análisis de correlación**: Evaluación de la relación entre variables económicas y diplomáticas

Resumen de la resolución

El proyecto se estructura en módulos especializados: procesamiento de datos (`core/`), limpieza de datos (`data_limpia/`), generación de reportes (`reportes/`) y visualizaciones (`graficos/`). Se implementó una clase `CustomDataFrame` para manejo eficiente de datos y se desarrollaron funciones específicas para cada análisis requerido.

Procesamiento de datos

1) LIMPIEZA DE DATOS

Cuando se nos presentaron los datos, tenían varias inconsistencias, campos vacíos y errores que hubo que corregir para mejorar la calidad del dataset. Los problemas que detectamos y solucionamos fueron los siguientes:

1. DATAFRAME PBI MUNDIAL

1) Comentarios innecesarios
Este CSV tiene comentarios al principio que bloquean el nombre de las columnas. Para arreglarlo, borramos los comentarios y las columnas se arreglan.

- **Tipo de problema**: Información no confiable (formato de archivo no estándar que impide su lectura correcta)
- **Goal (Meta)**: Asegurar una correcta carga y estructura de los datos desde el archivo CSV.
- **Question (Pregunta)**: ¿El archivo CSV contiene encabezados válidos y detectables por las herramientas de análisis?
- **Metric (Métrica)**:
    - Número de líneas de comentarios al inicio del archivo.
    - Resultado de df.columns tras la carga del CSV.

2) Datos faltantes de PBI
En 2023 (y en los otros años también) hay datos faltantes para el PBI de ciertos países, estos países son:

```
11                American Samoa
50                          Cuba
69                       Eritrea
72                      Ethiopia
84                     Gibraltar
93                          Guam
108                   Isle of Man
110               Not classified
147     St. Martin (French part)
164     Northern Mariana Islands
172                New Caledonia
193    Korea, Dem. People's Rep.
212                    San Marino
216                   South Sudan
254                 Venezuela, RB
255       British Virgin Islands
```

```
256        Virgin Islands (U.S.)
```

Esto lo encontramos ejecutando **get_empty_columns** de la clase **CustomDataFrame**. Deberíamos dejar solo el PBI de 2023 porque es lo que nos importa para nuestro caso de uso aunque una posible solución es usar el último dato anterior a este año, siendo 2022, o 2021, así respectivamente.

- **Tipo de problema**: Valores faltantes
    - **Extra**: Nuestra solución aún persiste el problema de tener **valores desactualizados**. Podríamos actualizarlos manualmente o mediante una API de algún tercero en el caso ideal.
- **Goal (Meta):** Contar con datos de PBI completos para todos los países relevantes en el año 2023.
- **Question (Pregunta):**
    - ¿Qué países no tienen datos de PBI en 2023?
    - ¿Es posible recuperar el dato de años anteriores (2022, 2021, etc.)?
- **Metric (Métrica):**
    - Cantidad de valores nulos en la columna de 2023.
    - Cantidad de valores que pueden ser imputados desde años anteriores.

Al hacer esto quedan 4 columnas vacías las cuales borraremos:

- {'columna_vacia': '2023', 'clave_primaria': 'Gibraltar', 'numero_registro': 85}
- {'columna_vacia': '2023', 'clave_primaria': 'Not classified', 'numero_registro': 111}
- {'columna_vacia': '2023', 'clave_primaria': "Korea, Dem. People's Rep.", 'numero_registro': 194}
- {'columna_vacia': '2023', 'clave_primaria': 'British Virgin Islands', 'numero_registro': 256}

2. DATAFRAME LISTA SEDES :

Por otro lado, hay un problema con la estructuración en el listado de redes sociales en el campo redes_sociales. Si bien todos siguen la estructura de un link seguido de un espacio, ‘//’ y otro espacio para separar las distintas redes sociales, hay algunos que contienen el nombre de usuario con un @ u otras formas de escribirlo. La solución que encontramos a esto fue realizar un análisis particular de los mismos y considerarlos simplemente como ‘otro’, en la sección del tipo de red social en el reporte.

- **Tipo de problema:** Este es un problema típico de Valores No Estandarizados.
- **Goal (Meta):** Contar con datos limpios y taggeados sobre las redes sociales de cada país.
- **Question:**
    - ¿Existen registros que no contienen URLs válidas sino solo handles o texto libre?
    - ¿Se respeta en todos los casos el separador estándar //?
    - ¿Es posible inferir de forma automática la red social a partir de esos formatos alternativos (ej. IG: @usuario → Instagram)?
- **Metrics (Métricas):**
    - Número de tokens no reconocidos como URL.
    - Número de elementos clasificados como “otro”.

- Cantidad de patrones distintos encontrados (URL, @usuario, prefijo: usuario).
- Porcentaje de registros con separador no estándar.

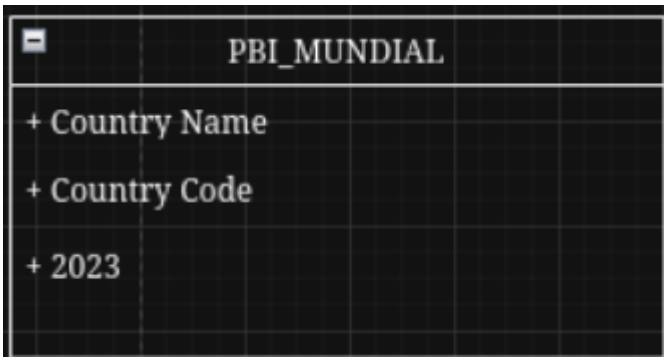
EISRA	twitter	https://twitter.com/argenisrael
EKUWA	facebook	https://www.facebook.com/ArgentinaenKuwait
REPAL	facebook	https://www.facebook.com/ArgEnPalestina
EQATR	otro	https://@embargenqatar
EQATR	otro	https://@embargenqatar
CCOCH	facebook	https://www.facebook.com/ArgentinaEnCochabamba/
CSCRS	facebook	https://www.facebook.com/profile.php?id=100069410160459&mibextid=ZbWKwL

2) Creación de diagramas e importación de datos

Basándonos en esta consigna:

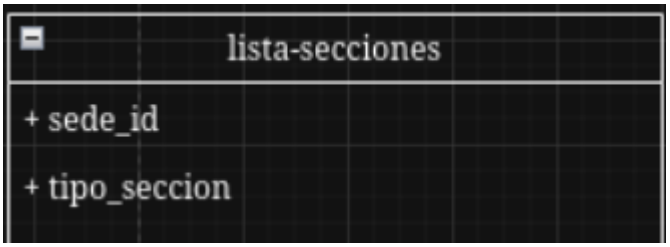
Una vez definidas dichas actividades, deberán armar un diagrama conceptual de los datos que sea adecuado para los objetivos del trabajo, utilizando (solamente) los datos necesarios para resolverlo. Cada fuente de datos original, previa a procesar, puede contar con varios atributos quizás no sean relevantes para resolver el problema. Luego, deberán decidir de dónde van a obtener los datos (de qué fuente de datos), diseñar los esquemas, y finalmente alimentarlos con los datos (limpios).

PBI MUNDIAL: importado de PBI MUNDIAL.csv



Este dataframe contenía los datos históricos desde 1960. Sin embargo, el trabajo solo necesitaba los datos de 2023. Por ende, esta es la única columna que dejamos.

LISTA SECCIONES: importado de lista-secciones.csv



En nuestro análisis de datos, solo necesitamos la ID de sede y el tipo de sección. El resto de datos son irrelevantes para nuestro análisis, puesto que son detalles de cada sección específica.
LISTA SEDES DATOS: lista-sedes-datos.csv

Al igual que en lista-secciones, solo nos centramos en los datos importantes de este data frame.
sede_id: Para identificar al pais
pais_castellano: Pequeña descripción para guiarnos sobre que país es
región geográfica: Para agrupar a cada país por región
pais_iso_3: Para vincularlo con la tabla PBI
estado: Para filtrar sedes activas o inactivas
redes_sociales: Listado de las redes sociales de cada sede

Decisiones tomadas

1) CLASE DATAFRAME CUSTOM

La clase custom dataframe la usamos como wrapper de pandas. Como en un principio nos costaba entender la sintaxis de pandas, decidimos hacer este wrapper para forzarnos a escribir nosotros mismos las funciones y poder reutilizar y encapsular la lógica de pandas.

2) PROBLEMAS DE LIMPIEZA DE DATOS

A la hora de agregar o remover datos, tomamos decisiones. Estas las mencionamos en el punto “**Procesamiento de datos**”. No las vamos a repetir para evitar la redundancia, pero simplemente las vamos a mencionar:

- Eliminación de comentarios adicionales
- Obtener PBI de años anteriores si éste faltaba
    - Si este dato faltaba, borramos el país correspondiente.
- Omitir las redes sociales que no siguen el formato :”https://”

Análisis de datos

PUNTO 7

Todos los reportes de este punto están en la carpeta reportes, cada uno con su archivo .csv y su archivo .py correspondientes en carpetas por separado.

Mostramos una parte de cada reporte, pero dado que no entra la información completa, recomendamos referirse a los .csv en el directorio mencionado anteriormente.

Ejercicio A:

	pais_castellano	Country Code	cantidad_sedes	promedio_secciones	pbi_per_capita_2023
1	REPÚBLICA FEDERATIVA DEL BRASIL	BRA	11	1.64	10377.5892792557
2	Estados Unidos de América	USA	9	3.22	82304.6204272866
3	REPÚBLICA ORIENTAL DEL URUGUAY	URY	8	0.62	23019.4221560046
4	ESTADO PLURINACIONAL DE BOLIVIA	BOL	7	2.14	3686.27996442176

Se puede observar cómo los países con más sedes son países ya sea cercanos territorialmente o en el caso de Estados Unidos con alto PBI y alta ventaja de relaciones exteriores.

Ejercicio B:

	region_geografica	cantidad_paises	promedio_pbi_per_capita
1	OCEANÍA	2	56745.41
2	EUROPA OCCIDENTAL	16	54953.35
3	AMÉRICA DEL NORTE	3	50117.03
4	ASIA	23	20970.53

En este reporte se puede observar las regiones que poseen menor cantidad de sedes Argentinas. Además, estos países están ordenados por promedio de pbi per cápita.

Ejercicio C:

	pais	nombre	cantidad_redes_distintas
1	AGO	REPÚBLICA DE ANGOLA	3
2	ARE	EMIRATOS ÁRABES UNIDOS	2
3	ARM	REPÚBLICA DE ARMENIA	4
4	AUS	AUSTRALIA	4

Este reporte no dice absolutamente nada relevante para el objetivo general ya que no vemos correlación entre el PBI, la cantidad de sedes y sus redes comprobables.

Ejercicio D:

	país	nombre	sede_id	tipo_red	url
2	AGO	REPÚBLICA DE ANGOLA	EANGO	facebook	https://www.facebook.com/ArgentinaEnAngola/
3	AGO	REPÚBLICA DE ANGOLA	EANGO	instagram	https://www.instagram.com/embargentinaenangola/
4	ARE	EMIRATOS ÁRABES UNIDOS	EEARB	facebook	https://www.facebook.com/ArgentinaEnEmiratosArabesUnidos

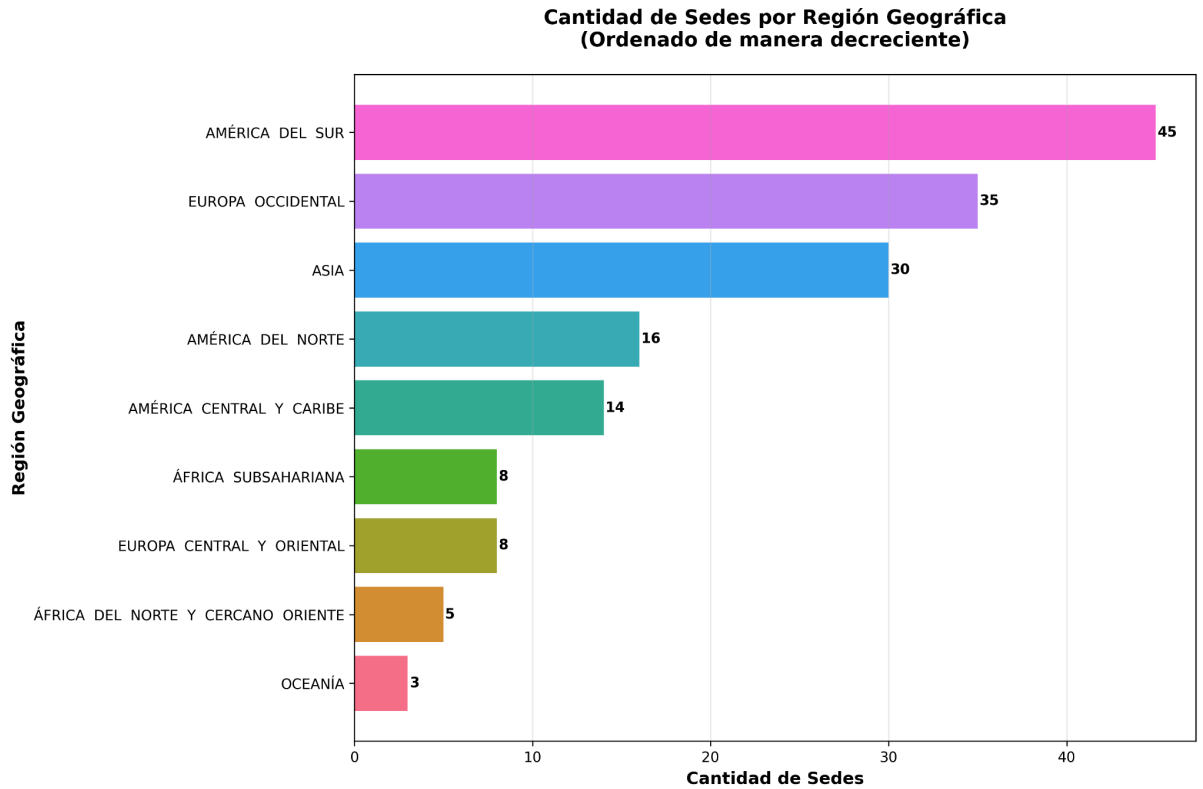
Este reporte no dice absolutamente nada relevante para el objetivo general. Para las redes que no se tiene información específica se pone “otra” en tipo de red.

PUNTO 8

De manera similar al punto anterior, todos los gráficos de este punto se pueden encontrar en la carpeta “graficos”. De vuelta, cada subcarpeta tiene dentro su respectivo gráfico en formato .png y el respectivo archivo .py.

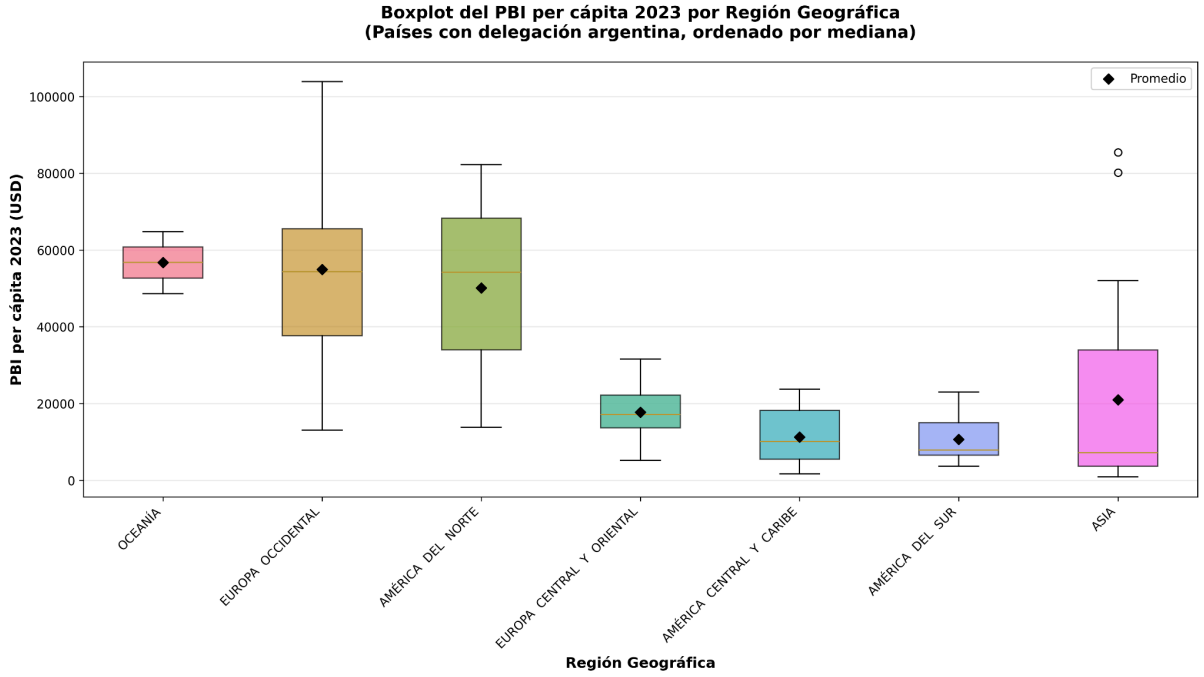
Mostramos los gráficos de cada ejercicio, aunque para un análisis más amplio, se deben utilizar los archivos que están en el directorio mencionado anteriormente.

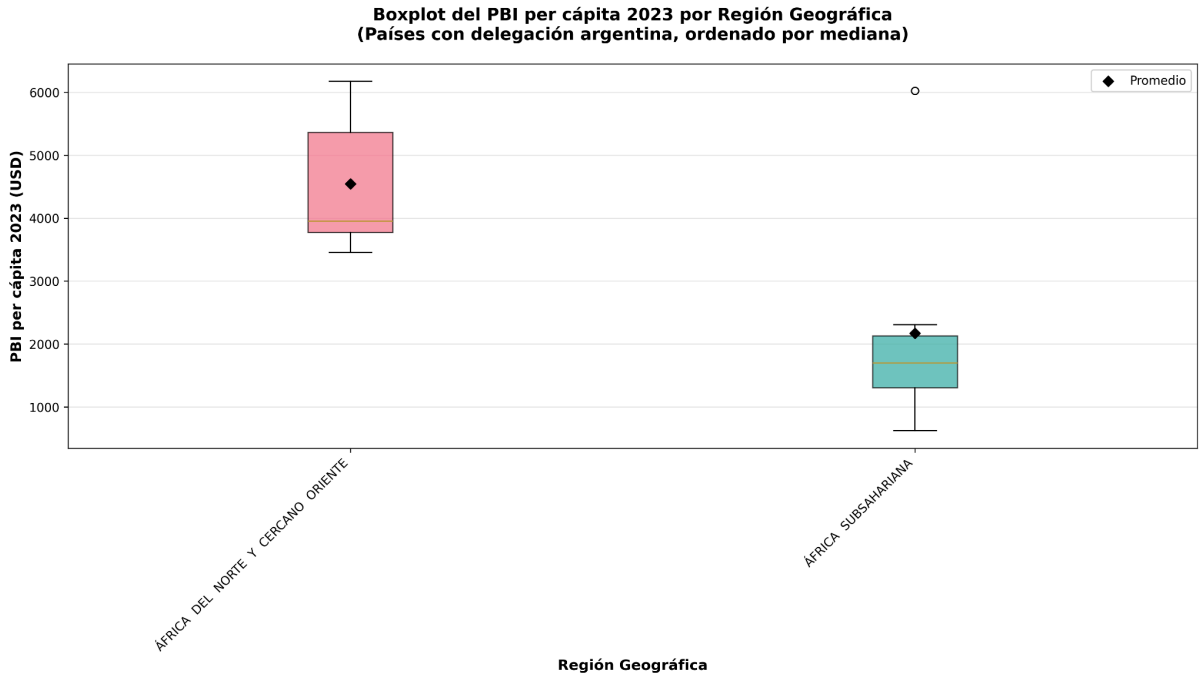
Ejercicio A:



En este gráfico se puede observar que en la región donde más sedes hay es en América del Sur. Este gráfico es muy sencillo de entender, no consideramos necesario analizarlo más profundamente. Se puede ver como las regiones con más sedes son América del Sur (por estrategia territorial), Europa (porque hay muchos países en Europa, entonces tiene sentido que tiendan a haber muchas sedes), Asia lo mismo y el resto se nota que hay menos.

Ejercicio B:



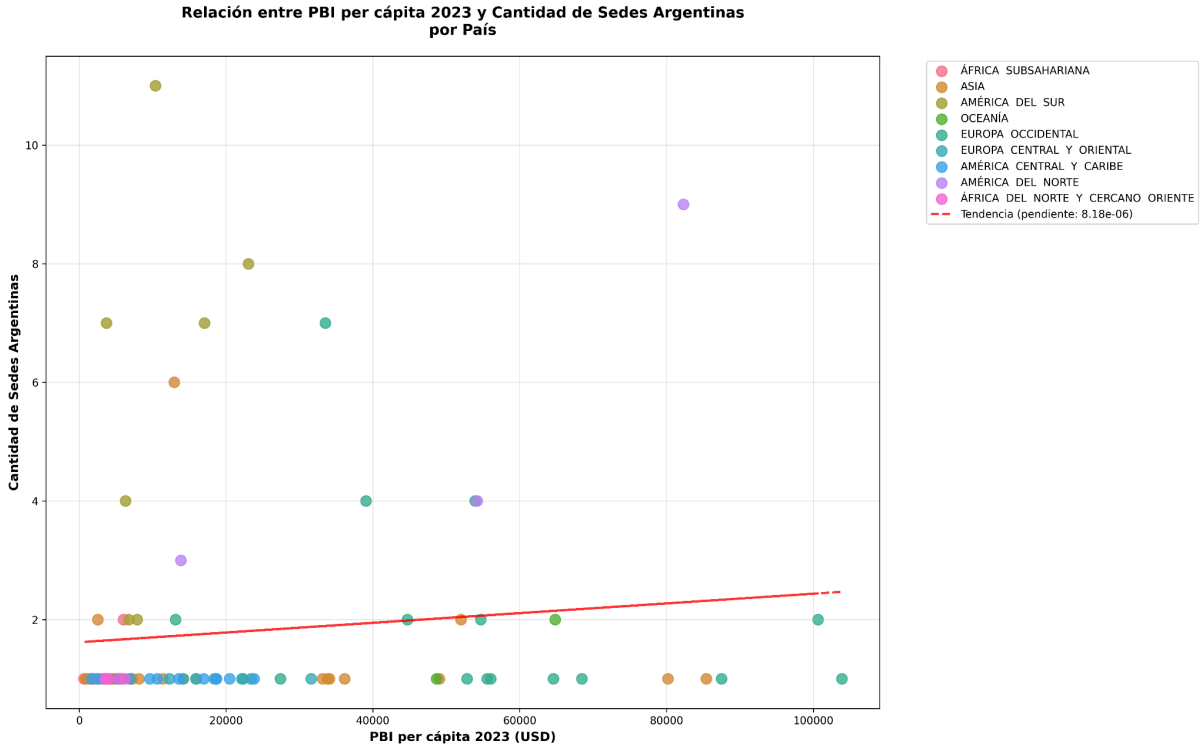


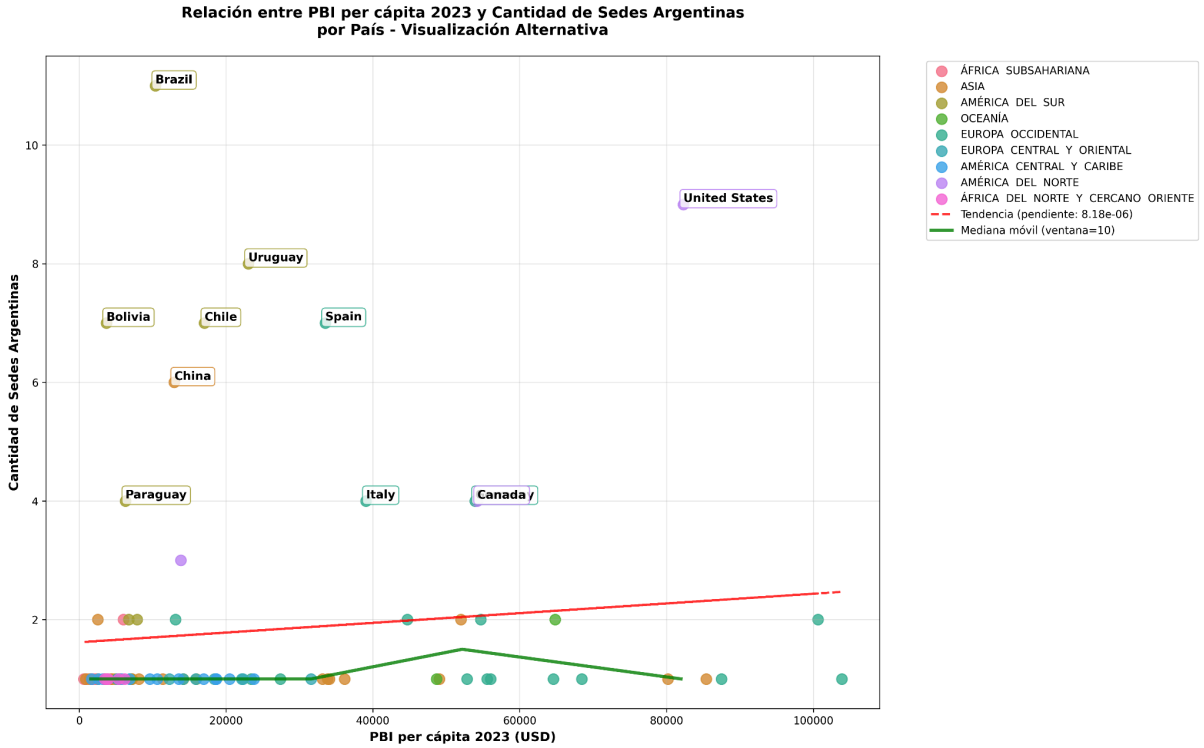
Algunas conclusiones:

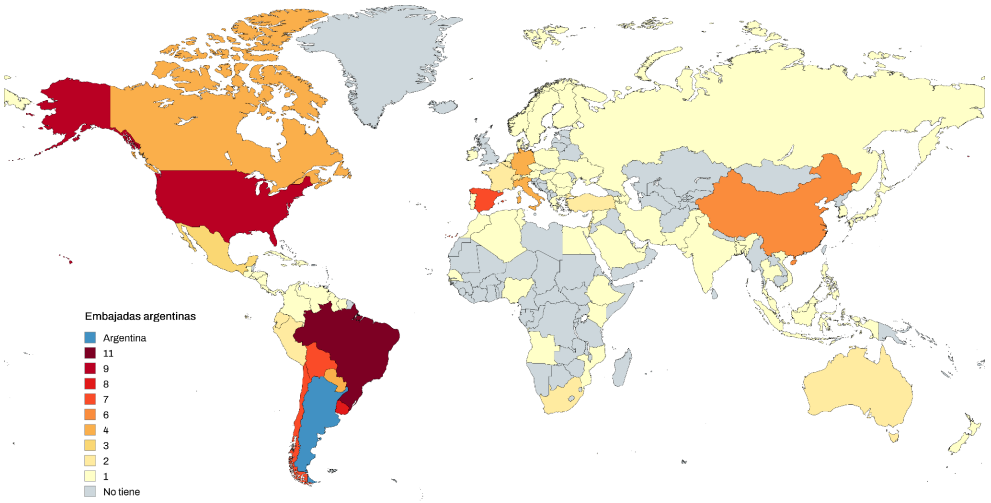
- Las regiones con mayor cantidad de países desarrollados (Oceanía, América del Norte, Europa Occidental) presentan los PBI per cápita más altos. Y las regiones con mayor cantidad de países muestran una diferencia entre máximo, mínimo y mediana, y promedio más grande.

- Las regiones latinoamericanas y africanas se concentran en la parte baja.
- Asia es la más heterogénea, con países pobres y ricos al mismo tiempo.
- Esto permite concluir que la distribución económica global es muy desigual y que, aunque Argentina mantiene sedes en todas las regiones, las diferencias de nivel económico son evidentes en los boxplots.

Ejercicio C:







La línea de tendencia que aparece en el gráfico (esa línea roja discontinua) muestra cómo, en general, se relacionan dos variables: el PBI per cápita de los países y la cantidad de sedes argentinas en ellos. Aunque la línea sube un poco (lo que sugiere que a mayor PBI per cápita hay una leve tendencia a tener más sedes) esa pendiente es tan pequeña (0.00001) que prácticamente no tiene peso. Además, el coeficiente de correlación es apenas 0.107, lo que confirma que la relación entre estas variables es muy débil. Entonces, ¿por qué hay tantos puntos agrupados entre países con PBI entre 0 y 40.000 dólares, y casi todos con pocas sedes? Hay varias razones que se explican mirando el gráfico y el mapa que hicimos nosotros a mano:

1. **La distribución global del PBI**: La mayoría de los países del mundo tienen un PBI per cápita por debajo de los 40.000 dólares. Solo unos pocos países muy desarrollados superan ese valor, así que es normal ver más puntos concentrados en ese rango. No aporta ningún valor para nuestro análisis.
2. **Estrategia diplomática argentina**: Argentina no decide dónde abrir sedes simplemente en función de la riqueza del país. Factores como cercanía geográfica, vínculos históricos, comercio bilateral o afinidad política influyen mucho más en esa decisión. Por ejemplo se observa que en países como Brasil, Bolivia están porque son cercanos por más que su PBI sea bajo otorgan otro tipo de valor estratégico para las relaciones internacionales de Argentina. Lo mismo para Estados Unidos, China y que por más que no están cerca son los países que controlan el mundo. Y por ejemplo España e Italia tienen vínculos históricos.
3. **Presencia mínima en muchos países**: En muchos países del mundo, especialmente aquellos que no son clave estratégicamente, Argentina mantiene una sola sede (por ejemplo, una embajada). Esto es algo común en las relaciones diplomáticas.

Aunque en teoría los países con más riqueza podrían tener más sedes argentinas, en la práctica esa relación es muy débil. La ubicación de las sedes responde a muchos otros factores, y por eso el gráfico muestra esa gran concentración de países con PBI medio y pocas sedes.

Conclusiones

En relación con la pregunta central del análisis, ¿existe una correlación entre el PBI per cápita y la cantidad de sedes diplomáticas argentinas?, la respuesta es no. La evidencia muestra que la distribución de sedes no depende principalmente del nivel económico de los países, sino de factores geográficos, históricos y culturales. La cercanía territorial, los vínculos políticos y la relevancia estratégica explican en mayor medida la ubicación de las representaciones argentinas. Solo en algunos casos puntuales (como Estados Unidos, China y Canadá) puede observarse una correspondencia más clara entre poder económico y número de sedes.

Este trabajo práctico también nos permitió adquirir experiencia en la limpieza y el filtrado de datos para un caso de uso concreto, aplicando herramientas de Python como pandas y matplotlib (junto con seaborn para las visualizaciones). Además, el ejercicio nos llevó a integrar no solo el análisis numérico, sino también consideraciones culturales y contextuales, lo que enriqueció la interpretación más allá de los datos estrictamente cuantitativos proporcionados por los datasets.