Laboratorio de Datos
Trabajo Práctico 01
Cuatrimestre Verano - 2024

TP 01: Manejo de datos y su visualización

Integrantes del grupo:

- *Carlos Pórcel,*
- *Valentino Pons,*
- *Ricardo Javier Suarez*

Sección Resumen

En el siguiente documento se pretende recolectar información de varios conjuntos de datos que corresponden a “PBI por países” y “Representaciones argentinas en el exterior”.
Con dicha información se observará si existe cierta relación entre el PBI (Producto Bruto Interno) por persona de cada país (año 2022) y la cantidad de sedes en el exterior que tiene Argentina en dicho país.
Para cumplir dicho objetivo se realizó:

- Extracción y utilización de datos Open Source de internet
- Análisis de los datos crudos y de la Calidad de Datos
- Desarrollo de DER
- Procesado de datos
- Análisis de datos limpios (gráficas y tablas)

Con estos pasos se llegó a la conclusión de que existe una relación no tan visible o marcada pero que existe al fin, entre el PBI de cada país y la cantidad de sedes en el exterior que tiene Argentina. Dicha relación la iremos desarrollando y detallando a lo largo del documento.

Sección Introducción

<u>Objetivo:</u>
Saber si existe cierta relación entre el PBI (Producto Bruto Interno) por persona de cada país (año 2022) y la cantidad de sedes en el exterior que tiene Argentina en dicho país, mediante la utilización de datos del Banco Mundial y Ministerio de Relaciones Exteriores, Comercio Internacional y Culto.
A continuación, se detallan las fuentes de donde se analizaron y se extrajeron los datos.

Fuentes

1. PBI per cápita de los países (PBI en inglés es GDP, por Gross Domestic Product). Se puede obtener del sitio del Banco Mundial:
https://data.worldbank.org/indicator/NY.GDP.PCAP.CD, descargando los csv y accediendo al archivo API_NY.GDP.PCAP.CD_DS2_en_csv_v2_6298251.csv

2. Representaciones Argentinas. El responsable de estas fuentes de datos es el actual Ministerio de Relaciones Exteriores, Comercio Internacional y Culto, y pueden ser obtenidas del sitio que se detalla a continuación:
https://datos.gob.ar/dataset/exterior-representaciones-argentinas. En dicho sitio podrán acceder a los siguientes datos:

A. Datos básicos de las sedes
B. Datos completos de las sedes
C. Datos completos de las secciones de las sedes

Primer vistazo a los datos

A continuación, una breve explicación y análisis de los datos crudos obtenidos.

Análisis datos del Banco Mundial

Descripción info. de Banco Mundial:

El PIB per cápita es el producto interno bruto dividido por la población a mitad de año. El PIB es la suma del valor agregado bruto de todos los productores residentes en la economía más los impuestos depreciación de activos fabricados o por agotamiento y degradación de recursos naturales. Los datos están en dólares estadounidenses actuales.

*CSV -> **Nombre para el análisis** -> Nombre variable en el código:*

- API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73 -> **PBI per cápita (USD**) -> pbi_paises
- Metadata_Country_API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73 -> **Código de los Paises** -> datos_paises
- Metadata_Indicator_API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73 -> **Metadata** -> metadata

Info. tablas:

PBI per cápita:

- Descripción columnas:
  - 'Country Name': Nombre del País
  - 'Country Code'': Código del País
  - 'Indicator Name': Nombre del indicador
  - 'Indicator Code': Código del indicador
  - ['1960', '1961', ... ,'2022']: PBI per cápita del país en ese año
  - ‘'Unnamed: 67': Columna sin datos

- Tipos de datos de las columnas:
  - Categóricos:
    - Nominal: 'Country Name', 'Country Code', ''Indicator Name', 'Indicator Code'

- Cuantitativos:
  - Continuos: Las columnas de los años: ['1960', '1961', ... ,'2022']

- Formato datos en columnas:
  - str: 'Country Name', 'Country Code', 'Indicator Name', 'Indicator Code'
  - float: '1960', '1961', ... ,'2022'

- - Observaciones:
  - Columna 'Unnamed: 67' todos sus datos son nulos.
  - Sin filas repetidas.
  -

Código de los Países:

- Descripción columnas:
  - 'Country Code': Código del País

    - 'Region': Región en inglés a la que pertenece el país
    - 'IncomeGroup': Grupo de ingresos
    - 'SpecialNotes': Nota sobre el país
    - ‘'TableName': Nombre que lleva el país en la tabla PBI per cápita(USD)
    - 'Unnamed: 5': Columna sin datos

- Tipos de datos de las columnas:
    - Categóricos:
        - Nominal: Todas las columnas, excepto 'IncomeGroup' y 'Unnamed: 5'
        - Ordinal: 'IncomeGroup'

- Formato datos en columnas:
    - str: 'Country Code', 'Region', 'IncomeGroup', 'SpecialNotes', 'TableName'
- Observaciones:
    - Columna 'Unnamed: 5' todos sus datos son nulos.
    - Sin datos duplicados

Metadata:

- Descripción columnas:
    - 'INDICATOR_CODE': Código del indicador
    - 'INDICATOR_NAME': Nombre del Indicador
    - 'SOURCE_NOTE': Nota sobre el indicador
    - 'SOURCE_ORGANIZATION': Organizacion/es que recolecaton los datos
    - 'Unnamed: 4': Columna sin datos

- Tipos de datos de las columnas:
    - Categóricos:
        - Nominal: Todas las columnas, excepto 'Unnamed: 4'

- Formato datos en columnas:
    - str: Todas las columnas, excepto 'Unnamed: 4'
- Observaciones:
    - Solo hay una tupla
    - 'Unnamed: 4' todos sus datos son nulos.

Análisis datos Ministerio de Relaciones Exteriores, Comercio Internacional y Culto

Descripción info.:

Datos de las Representaciones Argentinas en el exterior.

*CSV -> **Nombre para el análisis** -> Nombre variable en el código:*

- lista-secciones -> **Secciones** -> secciones_orginal
- lista-sedes -> **Sedes** -> sedes_original
- lista-sedes-datos -> **Datos de Sedes** -> datos_sedes

Info. tablas:

Secciones:

- Descripción columnas:
  - https://datos.gob.ar/dataset/exterior-representaciones-argentinas/archivo/exterior_01.03

- Tipos de datos de las columnas:
  - Fecha y hora: 'atencion_dia_desde', 'atencion_dia_hasta', 'atencion_hora_desde', 'atencion_hora_hasta', 'comentario_del_horario'
  - Categóricos:
    - Nominal: Todas las filas restante, excepto ‘temas’
- Formato datos en columnas:
  - str : Todas las columnas, excepto la columna 'temas'
- Observaciones:
  - Columna 'temas' sin datos
  - Sin columnas repetidas

Sedes:

- Descripción columnas:
  - https://datos.gob.ar/dataset/exterior-representaciones-argentinas/archivo/exterior_01.01

- Tipos de datos de las columnas:
  - Categóricos:
    - Nominal: Todas las columnas,excepto 'estado'
    - Binario: ‘estado'
- Formato datos en columnas:
- str: Todas las columnas
- Observaciones:
  - Sin observaciones

Datos de Sedes:

- Descripción columnas:
  - https://datos.gob.ar/dataset/exterior-representaciones-argentinas/archivo/exterior_01.02

- Tipos de datos de las columnas:
  - Fecha y hora: 'atencion_dia_desde', 'atencion_dia_hasta', 'atencion_hora_desde', 'atencion_hora_hasta', 'atencion_comentario'
  - Categóricos:
    - Binario: 'estado'
    - Nominal: Todas las columnas restantes
- Formato datos en columnas:
  - int: 'pais_codigo_telefonico', 'ciudad_zona_horaria_gmt', 'ciudad_codigo_telefonico'

    - str : Todas las columnas restantes.
- Observaciones:
    - La columna 'sitios_web_adicionales' no tiene datos

Con el primer vistazo dado se concluye que es necesario una selección de datos necesarios y útiles para cumplir el objetivo, además de una buena limpieza de los datos antes de su utilización. A continuación se realizará un análisis más profundo de los datos crudos, y se presentará la documentación del DER y su representación en el modelo relacional, y una descripción del proceso de importación.

Sección Procesamiento de Datos

(donde se mencione en qué forma normal se encontraban las fuentes de datos originales (ejercicio b), qué procesos se siguieron para aumentar la calidad a los datos (ejercicio f), la documentación del DER y su representación en el modelo relacional (ejercicios d y e), y una descripción del proceso de importación (ejercicio g).)

Forma normal tablas de representaciones Argentina:

b)
Representaciones argentinas estaba formada por tres tablas:
Datos Sedes
La tabla datos sedes no se encuentra en primera forma normal ya que contiene algunos campos con datos no atómicos , un requerimiento necesario para que esté en 1FN. Un caso que ocurre lo mencionado es la columna de ‘redes_sociales’ la cual contienen campos que tienen distintos tipos de ‘url’ separados por un “//”. Por ejemplo, la primera tupla de la tabla (índice 0) tiene tres urls en un mismo campo.

Sedes
Esta tabla se encuentra en 1FM ya que los valores de sus datos son atómicos.
Como PK se eligió la columna ‘sedes_id’. La tabla además se encuentra en segunda forma normal ya que la tabla al encontrarse en 1FN y utilizar una clave primaria de un solo atributo, esto ya es suficiente para que esté en 2FN. Sin embargo no se encuentra en tercera forma normal ya que hay atributos no primos los cuales generan otros atributos, por ejemplo: de ‘sede_desc_castellano’ se puede deducir ‘sede_desc_ingles’ pero la primera no es parte de la clave primaria.

Secciones
Los datos no son atómicos entonces no se encuentran en 1FN. En la columna ‘telefono_principal’ hay algunos campos los cuales tienen múltiples números de teléfonos separados por ‘/’ .Por ejemplo la tupla con índice 236 sufre esta condición.

Calidad de datos

Al analizar los datos originales, seleccionamos aquellos de importancia y para saber cuales son los campos necesarios teniendo en cuenta las pautas del trabajo práctico. Es por esto que a partir de las tablas originales, solo extrajimos las columnas que consideramos relevantes de cada una de ella las cuales verán a continuación:

- pbi_paises: ‘Country Code’, ‘2022’
- datos_sedes: 'sede_id','sede_desc_castellano',’pais_castellano’,’region_geografica’, 'pais_iso_3',’redes_sociales’
- secciones_original: 'sede_id','sede_desc_castellano'

- Notamos que la columna ‘sede_desc_castellano’ de la tabla cruda ‘secciones_original’ en realidad representaba el nombre de la sección, así que le cambiamos el nombre a esa columna y le pusimos simplemente ‘nombre_seccion’’.
Por lo tanto sedes_original la podemos pensar como:
secciones_original(‘sede_id’,’nombre_seccion’).
- Observamos que los datos de los registros de la tabla ‘sedes_original’ están incluidos en la tabla ‘datos_sedes’, y también que ambas tablas tienen la misma cantidad de filas. Por este motivo decidimos descartar la tabla sedes_original.
De esta manera los datos crudos de las tablas originales que utilizaremos serán:
- pbi_paises(‘Country Code’, ‘2022’)
- datos_sedes('sede_id','sede_desc_castellano',’pais_castellano’,’region_geografica’,' pais_iso_3',’redes_sociales’)
- secciones_original('sede_id','nombre_seccion')

Limpieza tabla Redes Sociales (redes_sociales)

i) Para esta tabla solo importamos las columnas de datos_sedes llamadas: **‘sede_id’** y **‘redes_sociales’** la cual esta última pasa a llamarse **‘url’**. **‘url’** tiene un problema, entre otros , el cual es que sus datos no son atómicos. Podemos encontrar campos los cuales contienen más de una red social (siempre separadas por un ‘ // ’). El atributo de calidad afectado en este caso es el de cantidad adecuada por lo explicado anteriormente

ii) El problema anteriormente mencionado corresponde a un problema de modelo y diseño de la tabla de datos. El simple hecho de que nuestro diseño admita o permita campos los cuales sus datos no estén de forma atómica indica que el problema proviene del modelo en sí y no de que los datos no son concisos u otras características más comunes de un problema de instancia.
iii)Para resolver este problema se va a utilizar la técnica GQM
Goal: Mejorar la calidad de los datos atomizando las celdas correspondientes a la columna
Question: ¿Cuántas filas con datos no atómicos hay? ¿Qué sedes tienen más de un url? ¿cuántos url hay en total ?
Metric: Números de urls separados , Lista de sedes con más de un url , Número de campos no atomizados

valores_no_atomicos = 85
total_valores = 126
Metrica = $(valores\_no\_atomicos\ /total\_valores)\ *\ 100\ =\ (85/126)\ *\ 100\ =\ 66,67\%$

Una vez corregidos los algoritmos necesarios para la correcta limpieza de nuestra tabla podemos observar numerosos cambios. El número de campos no atomizados pasaría a

valer 0. Por lo tanto no queda ninguna sede la cual tenga más de un url. Lo que si va a pasar es que la tabla resultante sea más larga ya que por cada url distinto se va a generar una nueva tupla con el nombre de la sede y el url en cuestión.

valores_no_atomicos = 0
total_valores = 272
Metrica_no_atomicos = $(valores\_no\_atomicos\ /total\_valores) * 100 = 0\%$

Limpieza tabla secciones

i) Para esta tabla se importaron de la tabla secciones_originales las columnas sede_id y sede_desc_castellano. El atributo de calidad afectado en este caso es nuevamente el de cantidad adecuada solo que en este caso a diferencia de la tabla anterior es un problema de que tenemos tuplas repetidas. Si bien en el dataset secciones_originales esto no ocurre ya que hay otros atributos que las diferencian, nosotros usamos solamente dos columnas las cuales al ser importadas a nuestra tabla 'secciones' muchas tuplas quedan repetidas.
ii)Este problema corresponde a un problema de instancia y de modelo ya que por culpa de los datos importados de secciones_originales y del modelo de nuestra tabla se producían tuplas duplicadas.
iii)Tecnica GQM
Goal: Tener una tabla la cual no tenga tuplas repetidas
Question: ¿Cuántas tuplas repetidas hay?
Metric: Cantidad de tuplas repetidas a eliminar

valores_duplicados = 3
total_valores = 516
Metrica = $valores\_duplicados\ /total\_valores) * 100 = (3/516) * 100 = 0.58\%$

Optamos por la eliminación de los valores duplicados, dejando estos valores:

valores_duplicados = 0
total_valores = 513
Metrica_diplicados = $valores\_duplicados\ /total\_valores) * 100 = 0\%$

Limpieza tabla_paises

i) En esta tabla veíamos afectado el atributo de PBI, ya que algunos valores de esa columna se encontraban nulos.

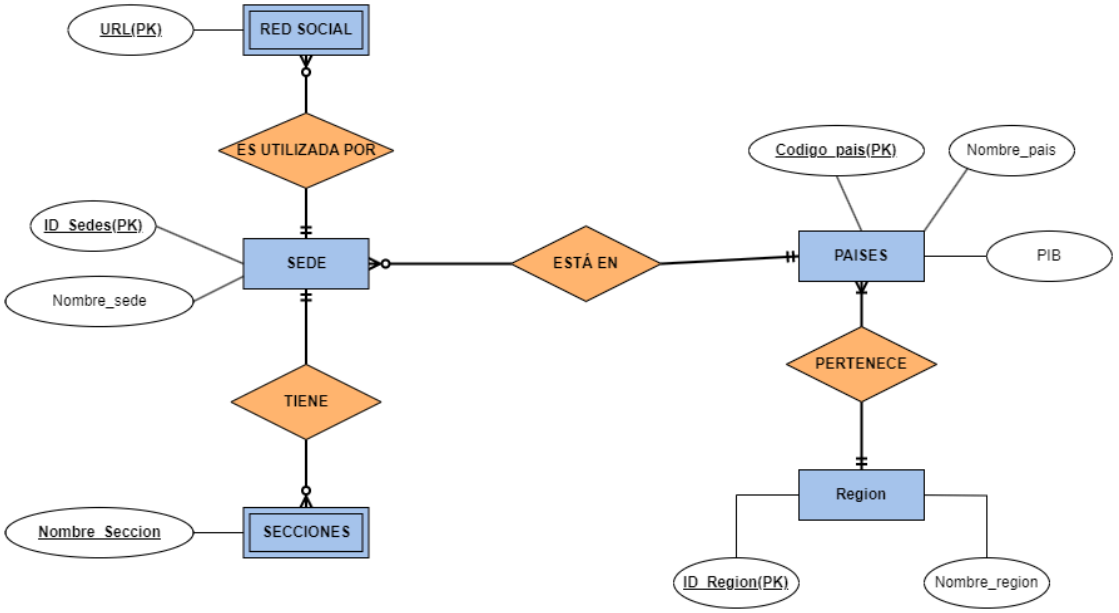
ii) En este caso teníamos un problema de instancia ya que al tratar con datos específicos, es decir, tomando países como Venezuela, en la tabla cruda de países original no figuraba su PBI, y lo mismo ocurría con otros países.

iii) Aplicamos el método GQM para resolver este inconveniente
Goal: Obtener datos que sean consistentes y limpios para luego utilizarlos
Question: Teniendo en cuenta la cantidad de filas totales de la tabla cruda 'pbi_paises',

Es relevante, es decir, ¿son muchas las filas cuyos países no posean PBI en el año 2022? ¿Es este atributo una columna crucial para llegar al objetivo del trabajo?
Metric: En este caso vemos que solo hay 22 países con el PBI 2022 nulos. Decidimos que son pocos los países con el dato nulo, ya que la tabla consta de 266 filas, y en este caso eliminamos 22 filas ya que como mencionamos, estas filas corresponden a países que no figura su PBI del año 2022.
En proporción esto sería: (266 filas total) / (22 filas con PBI nulo) = 12.1%
Esto quiere decir que solo un 12.1% de datos son nulos con respecto a la tabla original. Lo cual es muy poco.
Para ir cerrando, eliminamos estos datos nulos con lo que obtenemos un 0% de datos nulos, que era el objetivo de la limpieza de esta tabla paises
Por último, este atributo si es fundamental y crucial para poder llegar al objetivo del trabajo, porque lo que queremos trabajar con datos totalmente limpios con respecto al PIB y como son pocas las filas a eliminar, decidimos sacarlas de la tabla.

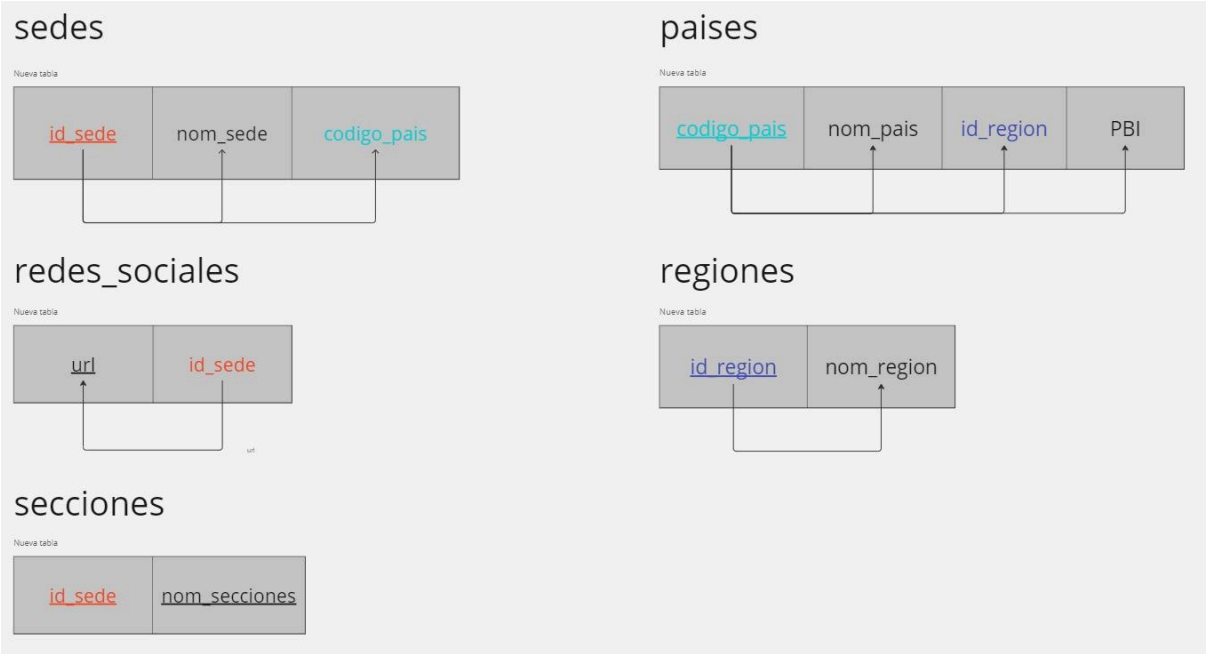
DER

El Diagrama Entidad-Relación (DER) que permita modelar de manera conceptual solamente los datos necesarios para lograr el objetivo propuesto sería el siguiente:



Esquema del DER correspondiente al Modelo Relacional

Dependencias funcionales minimales:
F = {
DF1: id_sede → {nom_sede, codigo_pais, url, nom_seccion}
DF2: codigo_pais → {nom_pais,id_region, PBI}
DF3: id_region → nom_region
}



Sección Decisiones tomadas

Secciones
- eliminamos duplicados

Países
- Solo importamos países que tienen sedes, es decir, los países los sacamos de la tabla ‘datos_sedes’
- eliminamos países sin datos en su pbi para 2022

Redes sociales
- Se eliminaron las tuplas con urls de mala calidad
  Se consideraron urls de mala calidad aquellos que no empezaban con “http” ," www" o alguna red social seguida de un ".com", ejemplo instagram.com.

Sección de Análisis de datos

Análisis mediante tablas

h1)
En principio, se observan resultados muy disparejos si se intentan relacionar los datos de sedes, secciones y pbi. Sin embargo, pareciera ser que la mayoría de los países de la tabla que más PBI poseen son aquellos que poseen secciones en promedio mayores o iguales a 2.5 y también aquellos que poseen solo una sede. Esto no quiere decir que si tienen una sola sede automáticamente su PBI es alto ya que hay países con una sola sede que tienen PBI bajo, solo significa que del conjunto de países con mayor PBI, la mayoría posee solo una sede argentina en el exterior.

h2)
Como era de esperarse debido a la cercanía, nuestro continente americano es el que tiene mayor cantidad de países con sedes argentinas (28). Sin embargo, también tenemos casi la misma cantidad de sedes argentinas en el continente Europeo, con 26.
Al parecer se le da primero mayor relevancia diplomática al sector de Europa Occidental que a Europa Central y Oriental.
El continente africano cuenta con muy pocas sedes argentinas, pero no tanto como Oceanía que cuenta con solo 2 sedes argentinas.
Sería interesante que Argentina lograra mayor contacto con el continente de Oceanía en este aspecto.

h3)
El dataset tabla_h_3 nos muestra por un lado el país y por el otro la cantidad de redes sociales distintas que tiene. Países como los Estados Unidos de América son de los que más tipos de redes distintas poseen sus sedes, en total 6. Por contraparte otros como BARBADOS sólo poseen una red. El promedio calculado a todos los respectivos países de la tabla ronda los 2.3 tipos de redes sociales.

h4)
La tabla ‘reporte’ generada en este punto nos permite observar, a diferencia de la del punto h3) , las redes sociales específicas que cada país utiliza. Por ejemplo Estados Unidos de América que en la otra tabla nos daba 6 tipos distintos de redes , ahora podemos observar cuales son {facebook , instagram , twitter , linkedin , youtube , flickr}

Importante: los resultados de las consulta se encuentran en archivos csv en la carpeta ‘TablasConsultas’

Visualizaciones

Cantidad de Sedes por Región(Figura1)

Esta visualización se corresponde con la cantidad de sedes agrupadas por regiones geográficas.

Se observa que Argentina posee mayor cantidad de sedes en América del Sur que en el resto de las demás regiones geográficas, esto puede deberse entre otras cosas a la proximidad de Argentina con los demás países.
En primer lugar, el continente americano cuenta con 74 sedes, seguido de Europa con 42 sedes. Sucede algo curioso y es que Argentina tiene muchísimas más sedes en la parte de Europa Occidental con 34 sedes, mientras que en el sector Central y Oriental de Europa solo posee 8 sedes.
Finalmente, Oceanía es la región que menos sedes posee ya que cuenta con solo 3.

Gráfico PBI per cápita por Región(Figura2)

Tomando en cuenta que la región con más sedes es América Latina con 45 queda en en el séptimo lugar, y Oceanía con solo 3 sedes es la primera en el gráfico.
También es posible ver que regiones con outliers muy alejadas de sus medianas tienen más sedes, como por ejemplo América del Sur, Europa Occidental y Asia.

Relación PBI y Número de Sedes argentinas por país(Firgura3)

Observando el gráfico obtenido (figura 3) nos da entender que no existe una relación directa entre el pbi del país y la cantidad de sedes argentinas en dicho país a simple vista.
Si es verdad que existe una tendencia a que los países con menos de 60000 usd de pib tengan una sola sede argentina, hay bastantes casos en los que esto no se cumple. Todos los países, excepto dos, que tienen igual o más de 2 sedes argentinas tienen un pib menor a 60000 usd y 5 están por arriba de los 60000 usd y aún así contienen solamente una sede.

Sección de Conclusiones.

En primer lugar, durante el desarrollo del trabajo, tuvimos varios debates sobre cómo y de qué manera crear las tablas y qué atributos de las tablas originales utilizar, también en el transcurso de la limpieza de datos y finalmente a la hora de observar los reportes y visualizaciones, en este marco se dieron a conocer diferentes opiniones e ideas sobre cómo llevar a cabo el desarrollo del trabajo práctico, debatiendo y decidiendo en consenso cuál estrategia se adecua más al objetivo solicitado.
Finalmente se pudo concretar el objetivo del trabajo que era decidir si existía relación alguna entre el PBI (Producto Bruto Interno) per cápita de cada país (año 2022) y la cantidad de sedes en el exterior que tiene Argentina en dicho país.
Teniendo en cuenta la figura 3, llegamos a la conclusión de que si excluimos a Suiza y Estados Unidos, cuyos valores de PBI son un poco atípicos en nuestro gráfico, la mayoría de los países con mayor PBI efectivamente se encuentran con solo una sede Argentina en el exterior.
Hacemos hincapié una vez más como ya se había mencionado anteriormente, que esto no quiere decir que si un país posee una sola sede Argentina automáticamente posee un PBI alto por encima de la media, sino que si tenemos en cuenta el conjunto de los países que superan la media de todos los PBI de los países, la mayoría de estos se encuentran con una sola sede Argentina en el exterior.
Puede que sea una relación leve o no tan visible pero al fin y al cabo existe.

Material Adjunto

Gráficos

Figura 1

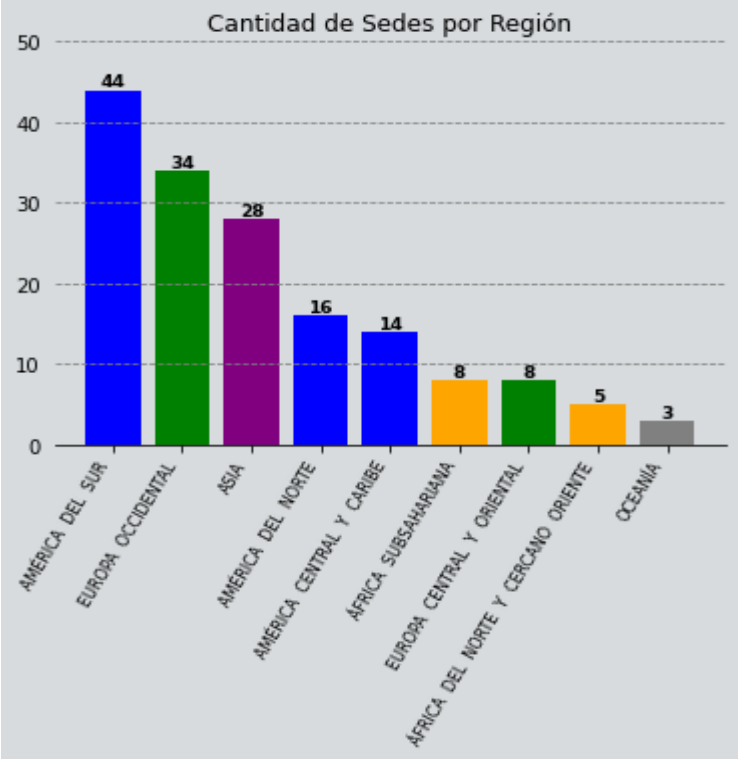


Figura 2

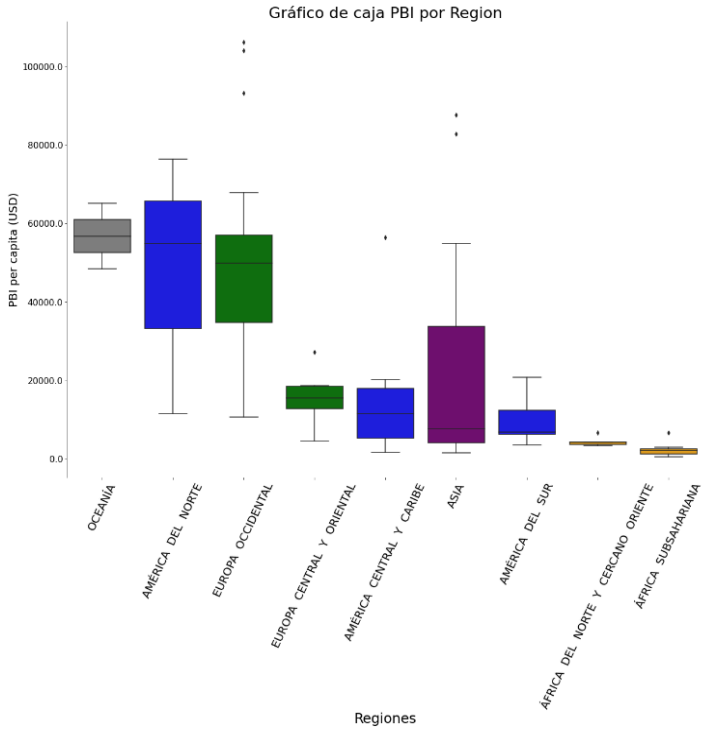
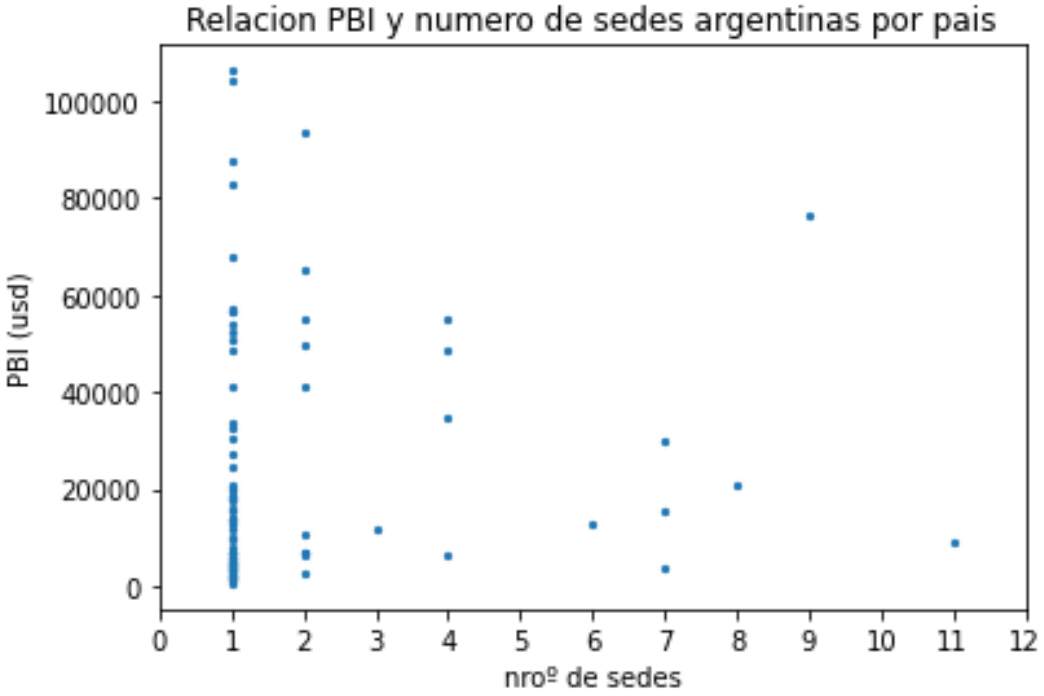


Figura 3


Relacion PBI y numero de sedes argentinas por pais
PBI (usd)
nroº de sedes
0
20000
40000
60000
80000
100000
0
1
2
3
4
5
6
7
8
9
10
11
12