DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

TP1

Relación PBI - Cantidad de sedes Argentinas en el exterior

February 23, 2024 Laboratorio de Datos

GRUPO 100

Integrante	LU	Correo electrónico
Chapana Puma, Joselin Miriam	1197/21	yoselin.chapana@gmail.com
Martinelli, Lorenzo	364/23	martinelli.lorenzo12@gmail.com
Padilla, Ramiro Martin	1636/21	ramiromdq123@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

1 Resumen

Estuve pensando y aca podriamos contar que fuimos pensando a lo largo del trabajo, los problemas con los que nos encontramos y como fue la dinámica, mas que nada porque la explicacion de la metodologia a seguir está en la parte de introducción.

2 Introducción

2.1 Objetivo y Fuente

El objetivo principal de este trabajo es encontrar una relación entre la cantidad de sedes de Argentina en un país y su PBI, si este será mayor, menor o si influirá la cantidad de secciones que una sede posea. Para esto, trabajaremos con los siguientes datos,

- PBI per cápita de los paises (1)
- Representaciones Argentinas en el exterior, donde tenemos, Datos básicos de las sedes, Datos completos de sedes y secciones (2)

(1) Extraido de la pagina del Banco Mundial, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

(2) Obtenidas del Ministerio de Relaciones Exteriores, Comercio Internacional y Culto, https://datos.gob.ar/dataset/exterior-representaciones-argentinas

2.2 Procedimiento

Este trabajo tendrá varias etapas, comenzando con el planteo de un Diagrama de Entidad Relacional (DER) adecuado al objetivo de nuestro trabajo, es decir, a partir de los datasets mencionados anteriormente, nos quedaremos solo con aquellos datos necesarios para resolver nuestro problema. Luego, pasaremos nuestro DER al modelo relacional, el cual se encontrará en tercera forma normal y especificará claves primarias (PK), claves candidatas (CK), claves foráneas (FK) y dependencias funcionales.

Una vez tengamos nuestro modelo planteado, pasaremos a Python donde realizaremos una limpieza de los datos tomando ciertas decisiones que estarán explicadas hacia el final de este informe. Finalmente, con los datos ya limpios, a traves de distintas librerias como Pandas, Matplotlib, Inlinesql, entre otras nos encargaremos de consultar, manipular y visualizar los datos necesarios para dar con nuestro objetivo.

3 Procesamiento de Datos

Antes de comenzar con el procesamiento de datos, nos encontramos con las fuentes de datos originales en distintas formas normales. Comenzando con **lista-sedes-datos** y **lista-secciones**, tenemos que ninguna de ambas se encuentra en primera forma normal puesto que el atributo *redes-sociales* de la primera y, el atributo *telefonos_principales* de la segunda tienen valores no atómicos. Al no estar en 1FN, tampoco se encontraran en 2FN.

Por otro lado, la tabla que contiene el pbi de los paises, está en primera y segunda forma normal, sin embargo no se encuentra en tercera forma normal puesto que tiene dependencias transitivas, en particular, la DF {Country Code $\rightarrow$ Indicator Name} es transitiva mediante el atributo *Indicator Code*, el cual no es clave candidata pero forma parte de la DF {Indicator Code $\rightarrow$ Indicator Name}, ademas tomamos *Country Code* como unica clave.

Y, por último, la tabla **lista-sedes** está en segunda forma normal, pero no está en tercera forma normal puesto que tiene dependencias transitivas como por ejemplo, la DF {sede_id $\rightarrow$ pais_iso_2} es transitiva mediante el atributo *pais_iso_3*. Esto es asi por que consideramos *sede_id* como clave primaria

3.1 Limpieza de Datos

Buscando mejorar la calidad de datos, encontramos problemas de instancia, por ejemplo, datos inconsistentes, problemas de proceso, puesto que hay diferentes criterios en la carga de los datos. Ahora, veamos cada dataset en detalle,

1. lista-sede-datos, en ella, encontramos un problema asociado a instancia en el atributo *'redes_sociales'* que influye en su consistencia puesto que no encontramos un criterio unificado a la hora de cargar distintas redes sociales. Por ejemplo, tenemos datos cargados en forma de URL donde es facil identificar a que red social pertenece, y por otro lado, tenemos datos que contienen, intuimos, nombres de usuarios, pero no podemos discernir su red social. Para tener una medida concreta acerca de la magnitud del problema utilizamos el método GQM de la siguiente manera:

   - Goal : Evitar que haya datos que no se puedan identificar con alguna red social.
   - Question : ¿Cual es la cantidad de elementos en redes_sociales que no podemos identificar?
   - Metrica : Proporción de registros sin campo redes_sociales que son Urls o comienzan con @.

   $$\text{M1: } \frac{\text{Cantidad de registros con datos en redes\_sociales que no son Urls o comienzan con @}}{\text{Cantidad de registros totales}}$$

   Para corregir esta tabla, además, seleccionamos columnas que consideramos relevantes para la resolución de nuestro problema principal. Puesto que encontramos datos redundantes, como dos nomenclaturas de codigo (pais_iso_3 / pais_iso_2) para un mismo pais, o la traducción al inglés de los nombres de estos paises. También, consideramos mantener unicamente los datos de redes sociales que eran consistentes, es decir, que eran Urls o comenzaban con @.

   Una vez hecha la limpieza de datos, procedemos a comparar las metricas antes y despues de dicha limpieza:

   - Antes: $M1 = 0.23$, es decir que un $\%23$ de registros son no deseados
   - Después: $M1 = 0$, es decir que un $\%0$ de registros son no deseados

2. En tabla correspondiente al PBI, API_NY.GDP.PCAP.CD_DS2_en_csv_v2_73.csv, tenemos también un problema de instancia, que afecta en particular, la completitud del atributo '2022'. Asi como también, otro problema de instancia es que encontramos muchos datos que no corresponden a paises. Estos fueron descartados antes de elaborar la siguiente métrica,

   - Goal : Tener el dato '2022' que refiere al Pbi de cada país completo.
   - Question : ¿Es relevante la proporción de paises con el dato '2022' vacío?
   - Metric: Proporción de paises con el campo '2022' vacío.

   $$\text{M1: } \frac{\text{Cantidad de registros con dato vacio en el campo '2022'}}{\text{Cantidad de registros totales}}$$

   Para corregir este dataset, en primeria instancia, descartamos todas las columnas que no aportaban a nuestro objetivo, como por ejemplo, todas aquellas correspondientes a años anteriores a 2022. En segunda instancia, eliminamos aquellos datos que no correspondian con paises, por ejemplo 'Africa'.

   Una vez hecha la limpieza de datos, procedemos a comparar las metricas antes y despues:

   - Antes: $M1 = 0.09$, es decir, un $\%9$ de registros contienen NULL en el campo '2022'
   - Después: $M1 = 0$, es decir, un $\%0$ de registros son no deseados

   En último lugar, consideramos que al ser tan baja la proporcion de nulls, era conveniente descartarlos.

3. Por último, en lista-secciones, nos quedamos unicamente con las dos primeras columnas, sede_id y descripción, en las cuales no encontramos ningun problema de calidad.

3.2 Diagrama de Entidad Relacional

Una vez planteado nuestro objetivo, nos encargamos de ver que datos necesitabamos para alcanzarlo, y como estarian representados. Para esto, elaboramos el siguien diagrama de entidad relacional.

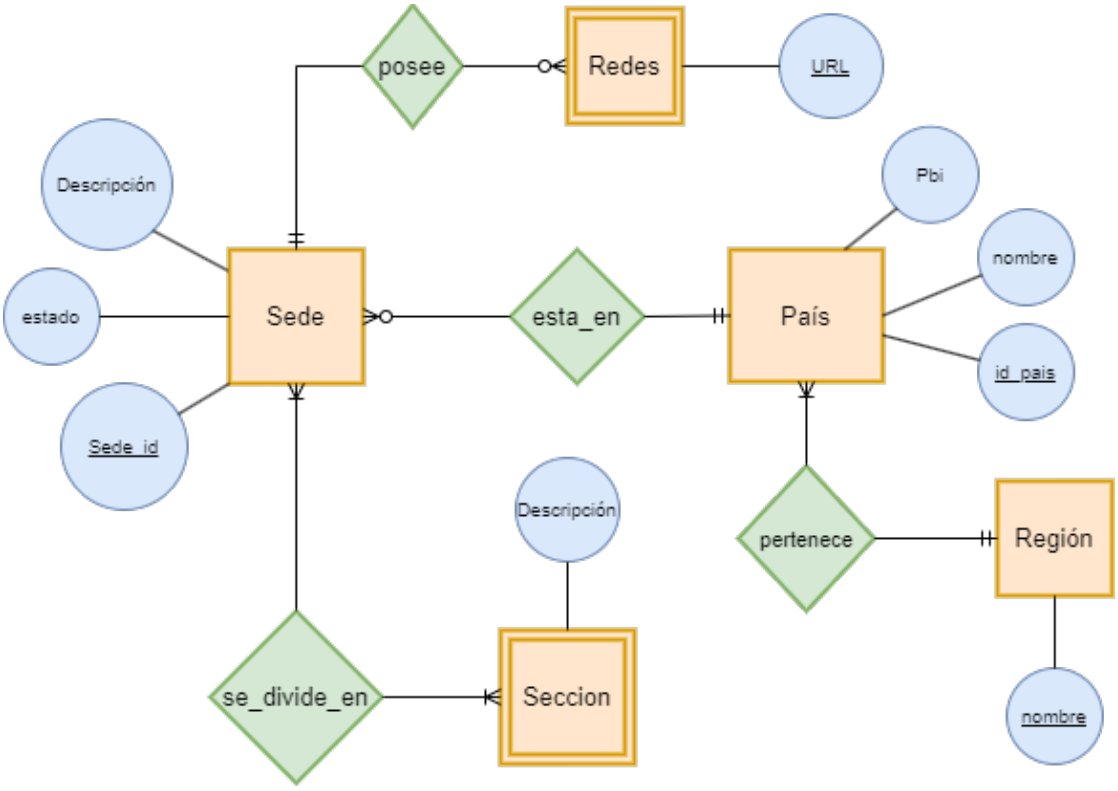


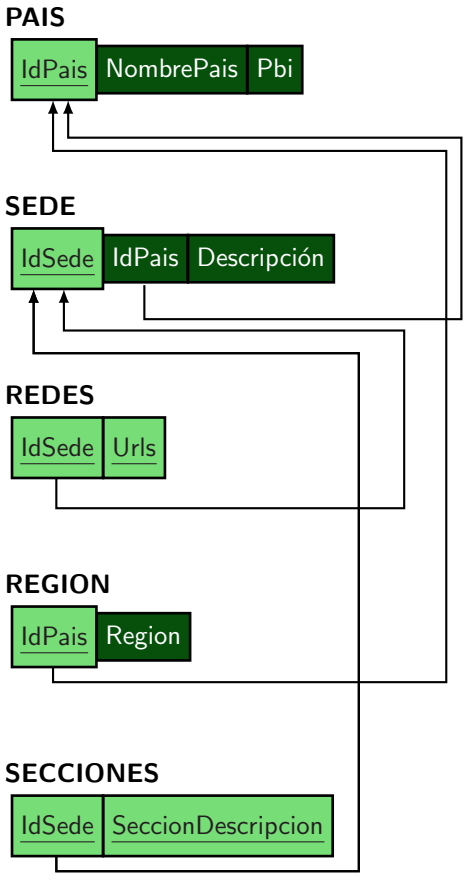
Figure 1: Diagrama de Entidad Relacional

Como se puede ver en la imagen anterior, consideramos que,

- Una sede esta en un país y solo en uno.
- Un País puede tener muchas o ninguna sede.
- Las secciones existen pues existen las sedes, entonces, lo consideramos una entidad debil.
- Una sección puede estar en una o muchas Sedes, por ejemplo, la Embajada en Brasil y en Chile tiene su sección Administración.
- Una sede tiene al menos una sección.
- Cuando hablamos de Redes, hablamos mas de un perfil en una red social, por lo tanto, una red puede pertenecer a una y solo una sede.

3.3 Modelo Relacional

Una vez que tenemos nuestro esquema gráfico, pasamos al planteo del modelo relacional. Notar que, las flechas representan las Foreign Keys y aquellos atributos subrayados representan las Primary Keys. En todas las relaciones, exceptuando a País, las Claves coinciden con las claves candidatas, esto se debe a que en País consideramos también a nombrePais como una posible clave.

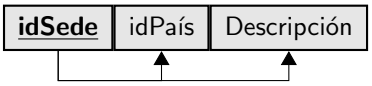


Ahora, solo resta mostrar las dependencias funcionales de este modelo relacional para dejar en claro que se encuentra en la forma normal deseada (3FN) y, también tener completo nuestro esquema para asi comenzar a manipular los datos.

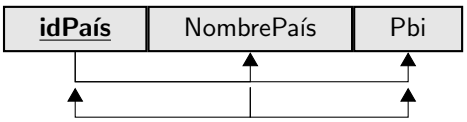
3.4 Dependencias Funcionales

A continuación, mostramos las dependencias funcionales de las relaciones de nuestro modelo relacional. En esta, se puede apreciar que todas se encuentran en segunda y tercera formal. Luego, con la limpieza de datos, garantizaremos que todos sus atributos sean atómicos y por ende se encuentre también en primera forma normal.

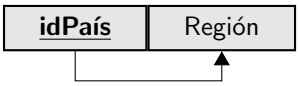
Sede



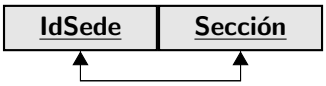
País



Región

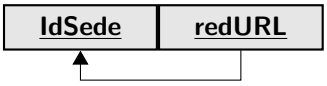


Sección



Observación: la única DF es {IdSede, Sección} $\rightarrow$ {IdSede, Sección}

Redes



3.5 Importación de Datos

Una vez ya limpios los datos, creamos con pandas un archivo csv para cada relación de nuestro modelo. Luego, los importamos a los dataframe vacíos creados anteriormente.

4 Decisiones tomadas

5 Análisis de datos

6 Conclusiones