Objetivo del Trabajo Práctico 01

Evaluar el manejo de datos y su visualización por parte de cada uno de los alumnos.

Enunciado

Los docentes de la materia Laboratorio de Datos se han encontrado con una fuente de datos abiertos correspondiente a las Representaciones Argentinas en el exterior de la República Argentina y otra sobre el PBI de los países. En particular, están interesados en 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. A continuación se detallan los datos con los que se cuentan. Previo a arribar a una conclusión, los docentes desean conocer cierta información de las fuentes de datos.

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

Se espera que para resolver el problema los estudiantes cumplan con los siguientes puntos:

- Plantear bien el objetivo general del trabajo solicitado.
- Dado que existen actividades que van a requerir de datos para alcanzar el objetivo, en primer lugar deberán realizar actividades para comprender el contenido de las fuentes de datos. Luego, deben leer todo el enunciado del TP, analizarlo y definir bien qué actividades deberán realizar y qué datos de las fuentes de datos deberán retener para llevar a cabo cada una de ellas (consultas, visualizaciones, etc.).
- Una vez definidas dichas actividades, deberán armar un diagrama conceptual de los datos (DER) que sea adecuado para los objetivos del trabajo, utilizando (solamente) los datos necesarios para resolverlo. No es necesario armar un DER por cada fuente de datos original, previa a procesar, ya que 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).

- Realizar las actividades solicitadas.
- Redactar el informe y realizar la entrega.

Ejercicios

a) Descargar los datos de las fuentes de datos. En general, para comprender en detalle los datos las páginas de descarga suelen contener documentación acerca de las fuentes (en algunos casos más detallada y en otros menos). En este caso dicha información es escasa.

b) ¿En qué forma normal se encuentran las tablas de Representaciones Argentinas? Justificar de manera concisa.

c) Plantear el objetivo general del trabajo.

d) Generar un Diagrama Entidad-Relación (DER) que permita modelar de manera conceptual **solamente** los datos necesarios para resolver los problemas planteados en el presente trabajo práctico.

e) Generar en python los dataframes (vacíos) correspondientes al modelo relacional del DER del punto anterior. Todos ellos deben estar en 3FN. Para cada uno de ellos definir (**no olvidar dejarlo documentado en el informe**):

   i) Clave primaria (PK)
   ii) Dependencias funcionales (DF). En lo posible, se desea que no escriban la totalidad de ellas sino un conjunto minimal de las mismas
   iii) Claves foráneas (Foreign keys)

f) El siguiente punto debería ser importar los datos (desde las fuentes de datos) a los esquemas vacíos generados en el punto anterior. Sin embargo, algunas de las fuentes de datos cuentan con problemas de calidad de datos y por lo tanto van a tener que llevar a cabo procesos para mejorar la misma, tratando de que ésta sea lo más parecida posible a la realidad. Describir los problemas de calidad de datos detectados en los datasets con los que trabajan. No es necesario que describan todos los problemas, pero sí al menos uno por cada fuente de datos utilizada. No puede ser el mismo problema para todas las fuentes (elegir al menos uno distinto para cada una). De esta manera, para cada uno de los datasets y cada problema de calidad deben mencionar:

   i) el atributo de la calidad afectado,
   ii) si el problema corresponde a modelo y/o a instancia,
   iii) una medida concreta acerca de la magnitud del problema (usar el método GQM de manera estricta, es decir, mencionando de manera explícita el objetivo, las preguntas y las métricas).

   Finalmente, describir en cada caso qué criterios utilizaron para corregir los datos y cómo impacta en la calidad (por ejemplo, cómo cambian los valores en las métricas).

g) Importar los datos (ya limpios) a los esquemas. **Documentar en el informe** desde qué fuentes de datos se está importando.

h) Generar los siguientes reportes **<u>utilizando sólo consultas SQL</u>**:

i) Para cada país informar cantidad de sedes, cantidad de secciones en promedio que poseen sus sedes y el PBI per cápita del país en 2022. El orden del reporte debe respetar la cantidad de sedes (de manera descendente). En caso de empate, ordenar alfabéticamente por nombre de país. A modo de ejemplo, el resultado podría ser:

País	sedes	secciones promedio	PBI per Cápita 2022 (U$S)
Brazil	11	1,6	8.917,67
Chile	7	2,0	15.355,47
...	...	...	...

<u>Importante:</u> Para el ejemplo no necesariamente han sido tenidos en cuenta los datos de la fuente de datos.

ii) Reportar agrupando por región geográfica: a) la cantidad de países en que Argentina tiene al menos una sede y b) el promedio del PBI per cápita 2022 de dichos países. Ordenar por el promedio del PBI per Cápita. Ejemplo:

Región geográfica	Países Con Sedes Argentinas	Promedio PBI per Cápita 2022 (U$S)
AMÉRICA DEL NORTE	3	9.532,61
OCEANÍA	2	8.712,47
...	...	...

<u>Importante:</u> Para el ejemplo no necesariamente han sido tenidos en cuenta los datos de la fuente de datos.

iii) Para saber cuál es la vía de comunicación de las sedes en cada país, nos hacemos la siguiente pregunta: ¿Cuán variado es, en cada el país, el tipo de redes sociales que utilizan las sedes? Se espera como respuesta que para cada país se informe la cantidad de tipos de redes distintas utilizadas. Por ejemplo, si en Chile utilizan 4 redes de facebook, 5 de instagram y 4 de twitter, el valor para Chile debería ser 3 (facebook, instagram y twitter).

iv) Confeccionar un reporte con la información de redes sociales, donde se indique para cada caso: el país, la sede, el tipo de red social y url utilizada. Ordenar de manera ascendente por nombre de país, sede, tipo de red y finalmente por url. Ejemplo:

País	Sede	Red Social	URL
Brazil	CFDIG	Twitter	https://twitter.com/argfoz
Chile	CCONC	Facebook	https://www.facebook.com/pmont
Chile	CCONC	Twitter	https://twitter.com/argpmont
...	...	...	...

<u>Importante:</u> Para el ejemplo no necesariamente han sido tenidos en cuenta los datos de la fuente de datos.

Importante: En aquellos reportes que resumen información no deben mostrar únicamente los listados sino que en el informe también deben comentar los resultados observados.

i) Mostrar, utilizando herramientas de visualización, la siguiente información:
   i) Cantidad de sedes por región geográfica. Mostrarlos ordenados de manera decreciente por dicha cantidad.
   ii) Boxplot, por cada región geográfica, del PBI per cápita 2022 de los países donde Argentina tiene una delegación. Mostrar todos los boxplots en una misma figura, ordenados por la mediana de cada región.
   iii) Relación entre el PBI per cápita de cada país (año 2022 y para todos los países que se tiene información) y la cantidad de sedes en el exterior que tiene Argentina en esos países.

Importante: No deben mostrar únicamente los gráficos sino que en el informe también deben comentar lo observado. Recordar al mostrar los ejes de los gráficos agregar separador de miles.

Finalmente, recordar que a modo de conclusión del trabajo se desea que intenten responder "… si existe cierta relación 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." En caso de que aún no lo hayan hecho, ¿qué información les parece que deberían mostrar que aún no han mostrado? Enumerar y mostrar los resultados.

Acerca de la entrega

La **documentación deberá ser entregada** en un informe. El mismo se debe entregar en formato pdf a través del **campus y** también una **versión impresa**. El informe debe contener:

- **Carátula**, con el nombre de la materia y del TP del que se trata.
- **Sección Resumen**, que resuma la problemática, el trabajo realizado y las conclusiones a las que arribaron.
- **Sección Introducción**, en donde se introduzca el problema a resolver, el objetivo general (ejercicio c), las actividades a realizar para alcanzar dicho objetivo y un resumen de la resolución y de cómo continúa el documento.
- **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).

- **Sección Decisiones tomadas**, que explique las mismas en el caso de que hayan tenido que tomar alguna.
- **Sección de Análisis de datos**, en la que se encuentren las respuestas a las preguntas planteadas en los ejercicios h e i. En el caso de reportes que involucren muchas filas, los mismos podrán ser incorporados en un **anexo** como **material suplementario o en un archivo csv, en el caso de las consultas (mencionando su ubicación)**.
- **Sección de Conclusiones.**

El largo total del informe (sin contar la carátula y el material suplementario) no debe exceder las 14 páginas A4 (utilizando un formato de letra Arial 11). Se evaluará que el documento (en formato .pdf) sea conciso, además de considerar la completitud y correctitud de escritura del mismo. Deberán entregar también el código generado en python (archivo .py).

Al comienzo del código deben incluir un encabezado con una descripción del contenido y otros datos que considere relevantes. El código debe tener comentarios donde se explique cada sección y debe poder correrse en cualquier máquina. Las variables usadas en el código y las tablas del modelo de datos tienen que tener nombres representativos. Las tablas originales y las resultantes del proceso de importación (al finalizar el ejercicio g) deberán entregarlas con el resto del TP. Cada una deberá estar en formato .csv. Aquellas originales deberán estar en una carpeta denominada `TablasOriginales` y aquellas limpias, en una carpeta llamada `TablasLimpias`.

El trabajo práctico (documento con el informe, código y ambos directorios con los archivos de datos) deberán subirse al campus en formato .zip. El nombre del archivo deberá ser *TP01-nombreyapellido.zip*. La fecha límite para subir el TP es el **lunes 26 de febrero a las 23:50 hs.**