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 correspondientes al Padrón de Operadores Orgánicos Certificados de la República Argentina, y desean saber si existe cierta relación entre el desarrollo de la actividad de producción orgánica y la proporción de mujeres empleadas en establecimientos productivos (no necesariamente orgánicos) en cada departamento de las provincias argentinas. A continuación se detallan los datos con los que cuentan. Previo a arribar a una conclusión, los docentes desean conocer cierta información de las fuentes primarias de datos.

Datos

Fuentes primarias

1. **Operadores orgánicos. Padrón de Operadores Orgánicos Certificados,** cuyo responsable es la Dirección de Agroalimentos - Producción Orgánica, y fue obtenido del sitio que se detalla a continuación: https://datos.magyp.gob.ar/dataset/padron-de-operadores-organicos-certificados. Lamentablemente, es probable que el encoding de esta fuente no sea utf-8.

2. **Establecimientos productivos. Distribución geográfica de los establecimientos productivos.** Esta fuente de datos contiene coordenadas de los establecimientos productivos (no necesariamente orgánicos) con su respectiva actividad económica, su nivel de empleo, jurisdicción y proporción de mujeres. El responsable de dichos datos es el Ministerio de Economía. Secretaría de Industria. Dirección Nacional de Estudios para la Producción (CEP XXI). Los datos pueden ser obtenidos de: https://www.datos.gob.ar/fa_IR/dataset/produccion-distribucion-geografica-establecimientos-productivos/archivo/produccion_15d42a00-0d1f-480c-bea8-3257e34b7804

Fuentes secundarias

1. **Localidades. Localidades de la Base de Asentamientos Humanos de la República Argentina** (BAHRA - www.bahra.gob.ar). En noviembre de 2011 se celebró un Convenio Marco de Cooperación entre el Instituto Geográfico Nacional (IGN), dependiente del Ministerio de Defensa, el Instituto Nacional de Estadística y Censos (INDEC), dependiente del Ministerio de Hacienda, y el Programa Nacional

Mapa Educativo dependiente del Ministerio de Educación para el intercambio e integración de información y efectiva conformación de la Base de Asentamientos Humanos de la República Argentina. Dicha fuente se puede obtener de: https://wms.ign.gob.ar/geoserver/bahra/ows?service=WFS&version=1.0.0&request=GetFeature&typeName=bahra%3Alocalidad_bahra&outputFormat=csv.
Esta fuente permite asociar a la fuente primaria "Padrón de Operadores Orgánicos Certificados" con los datos de departamento. Lamentablemente la fuente primaria, en su campo departamento parece mezclar datos de departamento y ciudad, entre otras cosas. Esa fuente también tiene inconvenientes en cuanto al formato y escritura de los nombres (por ejemplo, no parecen contar con tildes, etc.). Deberán hacer lo necesario para curar y vincular los datos.

2. **CLAE. Diccionario de CLAE**, cuyo responsable es el Ministerio de Desarrollo Productivo. Unidad Gabinete de Asesores. Dirección Nacional de Estudios para la Producción (CEP XXI). Dicha fuente contiene los nomencladores utilizados por AFIP para clasificar actividades con su correspondiente descripción. Dicha fuente, que puede obtenerse de https://datos.gob.ar/lt/dataset/produccion_15211f62-04dc-42ed-acdd-ff2bbcaf4779/archivo/produccion_0c48706b-94e0-4ba9-abac-7a02a44cd840, permite asociar a la fuente primaria "establecimiento productivo" los datos de actividades.

Para resolver el problema los docentes aconsejan seguir la siguiente metodología:

a. Plantear bien el objetivo general del trabajo solicitado
b. Existen actividades que van a requerir de datos para alcanzar el objetivo. Deberán leer TODO el enunciado del TP, analizarlo y definir bien qué actividades deberán realizar y qué datos requerirán para llevar a cabo cada una de ellas (consultas, visualizaciones, etc.).
c. Una vez definidas dichas actividades, armar un DER con (solamente) los datos necesarios para resolver la totalidad del trabajo (no es necesario armar un DER por cada fuente de datos original, previa a procesar), decidir de dónde van a obtener los datos (de qué fuente primaria, secundaria, relaciones entre ambas, etc.), diseñar los esquemas, y finalmente alimentarlos con los datos (limpios).
d. Realizar las actividades solicitadas
e. Armar el informe y realizar la entrega

Ejercicios

a) Descargar los datos de las fuentes de datos (primarias y secundarias). Para comprender en detalle los datos, notar que en las páginas de descarga suele haber documentación acerca de las fuentes (en algunos casos más detallada y en otros menos)
b) ¿En qué forma normal se encuentra cada tabla descargada? Justificar.
c) Plantear el objetivo general del trabajo

d) Generar un Diagrama Entidad-Relación (DER) que permita modelar conceptualmente 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 el conjunto minimal de las mismas
   iii) Claves foráneas (Foreign keys)

f) El siguiente punto debería ser Importar los datos (desde las fuentes de datos primarias, secundarias, etc.) a los esquemas 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 esta sea lo más parecida posible a la realidad. Describir los problemas de calidad de datos detectados en los datasets con los que trabajan. Para cada uno de los datasets y cada uno de los datos con problemas de calidad, mencionar:
   i) el atributo de la calidad afectado
   ii) si el problema corresponde a modelo y/o a instancia
   iii) dar una medida concreta acerca de la magnitud del problema (usar el método GQM)

   Finalmente, describir en cada caso qué criterios utilizaron para corregir los datos.

g) Importar los datos (ya limpios del modelo anterior) a los esquemas. Dejar documentado desde qué fuentes de datos se está importando.

h) Generar los siguientes reportes **utilizando sólo consultas SQL**:
   i) Para cada producto (producido por un productor orgánico) detallar en qué provincias se produce. El orden del reporte debe respetar la cantidad de provincias en las cuales se produce dicho producto (de mayor a menor). En caso de empate, ordenar alfabéticamente por nombre de producto. A modo de ejemplo, el resultado podría ser:

Producto	Provincia
CAÑA DE AZÚCAR	BUENOS AIRES
CAÑA DE AZÚCAR	ENTRE RÍOS
CAÑA DE AZÚCAR	JUJUY
CAÑA DE AZÚCAR	MISIONES
CAÑA DE AZÚCAR	SALTA
CAÑA DE AZÚCAR	SANTA FE
CAÑA DE AZÚCAR	TUCUMÁN
YERBA MATE	BUENOS AIRES
YERBA MATE	CIUDAD AUTÓNOMA DE BUENOS AIRES
YERBA MATE	CORRIENTES
YERBA MATE	MISIONES
…	…

Importante: Para el ejemplo no fueron tenidos en cuenta los datos de la fuente de datos (simplemente fueron inventados).

ii) ¿Cuál es el CLAE2 más frecuente en establecimientos productivos? Mencionar el Código y la Descripción de dicho CLAE2.

iii) ¿Cuál es el producto más producido (que lo producen más establecimientos de operadores orgánicos)?¿Qué Provincia-Departamento los producen?

iv) ¿Existen departamentos que no presentan Operadores Orgánicos Certificados? ¿En caso de que sí, cuántos y cuáles son?

v) ¿Cuál es la tasa promedio de participación de mujeres en cada provincia?¿Cuál es su desvío? En cada caso, mencionar si es mayor o menor al promedio de todo el país

vi) Mostrar por cada provincia-departamento cuántos establecimientos productivos y cuántos emprendimientos orgánicos posee

i) Mostrar, utilizando herramientas de visualización, la siguiente información:
   i) Cantidad de establecimientos productivos por provincia
   ii) Boxplot, por cada provincia, donde se pueda observar la cantidad de productos por operador
   iii) Relación entre cantidad de establecimientos de operadores orgánicos certificados de cada provincia y la proporción de mujeres empleadas en establecimientos productivos de dicha provincia. Para este punto deberán generar una tabla de equivalencia, de manera manual, entre la letra de CLAE y el rubro de del operador orgánico.
   iv) ¿Cuál es la distribución de los datos correspondientes a la proporción de mujeres empleadas en establecimientos productivos en Argentina? Realicen un violinplot por cada provincia. Mostrarlo en un solo gráfico.

Finalmente, se desea que intenten mostrar si existe " … cierta relación entre el desarrollo de la actividad orgánica y la proporción de mujeres empleadas en establecimientos productivos de las provincias.". 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.

Es importante documentar todo el proceso y que todos los integrantes se involucren en el mismo.

Grupos

Los grupos deben estar conformados por 3 (y sólo 3) integrantes. Ni más, ni menos. Deberán i) registrar la conformación del grupo en la siguiente planilla, y ii) definir quién va a ser el encargado del envío (debe ser uno y sólo uno de los integrantes del grupo): https://docs.google.com/spreadsheets/d/1h9omAxp_8P5xrG62Gpg5BlYWFP1GAEhz02Gmm72KUJs/edit?usp=sharing

Acerca de la entrega

La documentación deberá ser entregada en un informe. Este debe contener:

- **Carátula**, con el nombre de la materia y del TP del que se trata, y miembros del grupo.
- **Sección Resumen**, que resuma la problemática y el trabajo realizado.
- **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 Decisiones tomadas**, que explique las mismas en el caso de que hayan tenido que tomar alguna.
- **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 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**.
- **Sección de Conclusiones.**

El largo total del informe (sin contar la carátula y el material suplementario) no debe exceder las 10 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 el nombre de los integrantes del grupo, 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 (lo subirá el responsable del grupo encargado del envío). El nombre del archivo deberá ser *nombredelgrupoTP1.zip*. La fecha límite para subir el TP es el **martes 17 de octubre a las 23:59 hs.**