Objetivo del Trabajo Práctico 01

Evaluar el manejo de datos y su visualización.

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 y el salario promedio que perciben los trabajadores del sector privado en cada departamento de las provincias argentinas. A continuación se detallan los datos con los que cuentan.

Datos

Fuentes primarias

1. **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. **Salarios del sector privado**. Esta fuente de datos contiene el salario bruto mediano de los trabajadores registrados del sector privado, por departamento/partido y clase, con frecuencia mensual y desde 2014. El responsable de dichos datos es el Ministerio de Desarrollo Productivo. Unidad Gabinete de Asesores. Dirección Nacional de Estudios para la Producción. Los datos fueron obtenidos de: https://www.datos.gob.ar/fa_IR/dataset/produccion-salarios-por-departamentopartido-sector-actividad/archivo/produccion_515b41b2-d008-42fa-a9d7-8a1bb26d04ab.

Fuentes secundarias

1. **Listado de las localidades censales según la base de datos censales del** INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS (**INDEC**), cuyo responsable es la Jefatura de Gabinete de Ministros. Secretaría de Innovación Pública. Subsecretaría de Servicios y País Digital. Dicha fuente se puede obtener de: https://datos.gob.ar/ar/dataset/jgm-servicio-normalizacion-datos-geograficos/archivo/jgm_8.12.

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. **Diccionario de departamentos.** El responsable de dichos datos es el Ministerio de Desarrollo Productivo - Unidad Gabinete de Asesores - Dirección Nacional de Estudios para la Producción. La fuente contiene los códigos utilizados por el INDEC para caracterizar los departamentos/partidos y provincias, con su correspondiente descripción. En el caso del código de CABA es un un código ficticio. Dicha fuente se puede obtener de:
https://datos.produccion.gob.ar/dataset/puestos-de-trabajo-por-departamento-partido-y-sector-de-actividad/archivo/125bdc76-0205-417a-bf20-76d34dbe184b
Esta fuente permite asociar a la fuente primaria “Salario del sector privado” los datos de departamento. Por si es de ayuda, hemos detectado algunas inconsistencias entre esta fuente (Diccionario de departamentos) y la fuente de Listado de las localidades censales. A modo de ejemplo, parecería ser que Ushuaia en esta tabla tiene el código 94014 y en la otra 94015. Esto puede ser un problema a la hora de unir las tablas. Deberán corregirse estos problemas.

3. **Diccionario de clases**, cuyo responsable es el Ministerio de Desarrollo Productivo - Unidad Gabinete de Asesores - Dirección Nacional de Estudios para la Producción. Dicha fuente contiene los nomencladores utilizados por AFIP para clasificar actividades con su correspondiente descripción. Dicha fuente, que puede obtenerse de
https://www.datos.gob.ar/fa_IR/dataset/produccion-salarios-por-departamentopartido-sector-actividad/archivo/produccion_8c7e4f21-750e-4298-93d1-55fe776ed6d4,
permite asociar a la fuente primaria “Salario del sector privado” los datos de actividades.

Ejercicios

a) Descargar los datos de las fuentes de datos mencionadas.
b) ¿Todas las tablas descargadas se encuentran en primera forma normal? Justificar. En caso de no encontrarse en 1FN adaptarlas para que lo estén.
c) Con las tablas resultantes del punto anterior (todas ya se encuentran en 1FN), definir las dependencias funcionales y escribirlas. En lo posible, se desea que no escriban la totalidad de ellas sino el conjunto minimal de las mismas.
d) Descomponer los esquemas con los que trabajaron el punto anterior para que todos ellos se encuentren en 3FN y sin perder información (deben cumplir con la propiedad de lossless join). ¿Se perdieron dependencias funcionales? ¿Cuáles?

e) A partir de estos esquemas, construir un modelo conceptual de los datos. Llevar esta tarea adelante utilizando el DER como herramienta.
f) A partir del DER definir un modelo relacional que se encuentre en 3FN. No olvidar definir las claves primarias y las foreign keys de cada una de las relaciones.
g) Generar en python los dataframes correspondientes al modelo relacional, conteniendo los datos de las fuentes primarias y secundarias.
h) Para poder llevar adelante el punto anterior se van a dar cuenta que algunos de los datasets 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 dimensión de la calidad afectada (y si corresponde a modelo y/o a instancia) y den una medida concreta acerca de la magnitud del problema. Describan qué criterios utilizaron para corregir los datos.
i) Responder las siguientes consultas a través de consultas SQL:
    i) ¿Existen provincias que no presentan Operadores Orgánicos Certificados? ¿En caso de que sí, cuántas y cuáles son?
    ii) ¿Existen departamentos que no presentan Operadores Orgánicos Certificados? ¿En caso de que sí, cuántos y cuáles son?
    iii) ¿Cuál es la actividad que más operadores tiene?
    iv) ¿Cuál fue el salario promedio de esa actividad en 2022? (si hay varios registros de salario, mostrar el más actual de ese año)
    v) ¿Cuál es el promedio anual de los salarios en Argentina y cual es su desvío?, ¿Y a nivel provincial? ¿Se les ocurre una forma de que sean comparables a lo largo de los años? ¿Necesitarían utilizar alguna fuente de datos externa secundaria? ¿Cuál?
j) Mostrar, utilizando herramientas de visualización, la siguiente información:
    i) Cantidad de Operadores por provincia.
    ii) Boxplot, por cada provincia, donde se pueda observar la cantidad de productos por operador.
    iii) Relación entre cantidad de emprendimientos certificados de cada provincia y el salario promedio en dicha provincia (para la actividad) en el año 2022. En caso de existir más de un salario promedio para ese año, mostrar el último del año 2022.
    iv) ¿Cuál es la distribución de los salarios promedio en Argentina? Realicen un violinplot de los salarios promedio por provincia. Grafiquen el último ingreso medio por provincia.
k) Todo el proceso realizado para modificar los datos tiene que estar en el código. Para realizarlo, podrán utilizar funciones de SQL (algunas de ellas no incluídas en los pdf de la clase) como son UPPER, REPLACE, etc.

Finalmente, se desea que intenten mostrar si existe " … cierta relación entre el desarrollo de la actividad y el salario promedio que perciben los trabajadores del sector privado en cada departamento de las provincias argentinas.". ¿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.

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, 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**, que explique las transformaciones que tuvieron que realizar a los datos (procesos para normalizar -ejercicios b,c,d, e y f-) y documentación del DER resultante. Aquí deben incluir la descripción realizada del trabajo de curado de datos (ejercicio h).
- **Sección de Análisis de datos**, en la que se encuentren las respuestas a las preguntas planteadas en los ejercicios i y j.
- **Sección de Conclusiones**

El largo total del informe (sin contar la carátula) no debe exceder las 10 páginas A4. Se evaluará la concisión y la completitud y correctitud de escritura del mismo.
Deberán entregar también el código generado en python (archivo .py).

Al comienzo del código deben incluir el nombre de los integrantes del grupo. 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 nombre representativos. Las tablas originales y las resultantes del proceso de normalización y limpieza deberán entregarlas con el resto del TP. Cada una deberá estar en formato .csv. Aquellas originales deberán estar en la carpeta TablasOriginales y aquellas limpias, en el directorio TablasLimpias.

El trabajo práctico (informe, código y ambos directorios con los archivos de datos) deberán subirse al campus en formato .zip. El nombre del archivo deberá ser *nombredelgrupoTP1.zip*. La fecha límite para subir el TP es el **15 de mayo a las 23:59 hs.**