UC | Chile

Ingeniería de datos

Tarea: Analizando datos con SQL y Pandas

¡Bienvenido(a)!

Te invitamos a realizar la tarea Analizando datos con SQL y Pandas.

- **Objetivo**: Analizar datos con SQL y Pandas.
- **Tipo de actividad**: Individual
- **Tipo de evaluación**: Sumativa (con calificación)
- **Ponderación**: 15% (Equivale al 15% de la nota final del curso)
- **Puntaje**: 60 puntos
- **Calificación**: Escala de 1 a 7, con una exigencia de 50%. La nota mínima para aprobar es 4.0.

Instrucciones

1. Antes de comenzar, debes haber revisado las siguientes clases y las lecturas hasta la semana 6.
2. Lee con atención el enunciado de la tarea que se adjunta a continuación
3. Debes entregar un archivo .ipynb con tus respuestas. Se adjunta un formulario de respuestas para que uses como template. Descargalo y desarrolla allí tu tarea.
4. Instrucciones:
    - Haz clic en el botón para agregar entrega. Se abrirá una nueva ventana que permite arrastrar el archivo y subirlo.
    - Comprueba que el archivo arrastrado es el correcto y presiona el botón para guardar cambios. El documento quedará guardado en la plataforma.

Enunciado

Introducción

En esta tarea continuaremos desarrollando nuestro caso de uso visto en las tareas anteriores. En particular, ahora que ya tenemos nuestra base de datos armada y analizada por sí misma, haremos un análisis de los datos cruzando datos que ya tenemos en nuestra base de datos, con unos datos coleccionados desde las Naciones Unidas para ves si podemos sacar algunas conclusiones sobre la participación en varios talleres en base indicadores del país dónde se estos organizan.

Descripción del problema

En esta tarea trabajaremos con el mismo esquema cómo en la Tarea 2. En particular, contamos con el siguiente esquema para nuestros datos:

Persona(rut,nombre,correo,teléfono)

TallerDeportivo(id,nombreTaller, nombreProfesor, fecha,hora,valor,duración,deporte,nrEquipos, nrPersonas)

TallerInstrumental(id,nombreTaller, nombreProfesor, fecha,hora,valor,duración,instrumento,nrInstrumentos)

Lugar(id,calle,número,código,nombreCiudad,nombrePaís)

Participa(rut,idTaller)

EstáEn(idTaller,idLugar)

Si necesitas acordarte del significado de cada una de estas tablas, debes consultar la descripción en Tarea 2.

Esta vez, adicionalmente contamos con un archivo .csv que contiene información sobre los indicadores de cada país. Los datos que se vinculan a un país en este archivo son: su nombre, población, pib (GDP), etc. Este archivo proviene de la información publicada por la ONU, y se puede encontrar, bajo la licencia CC0, en Kaggle: https://www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles El archivo lo

dejaremos disponible para descargar en un Google Drive interno igual cómo a la base de datos. Los datos en este archivo los explicaremos más adelante.

Idea de esta tarea es hacer un análisis del mercado, y ver si los indicadores de un país nos dicen algo sobre la participación en los talleres que se organizan en este país. En nuestro caso, nos concentraremos en los países en América Central y el Caribe.

Junto con esta tarea, te pasaremos un Jupyter notebook que descargará, de manera automática, la base de datos que ocuparemos, y el archivo .cvs que necesitamos. Adicionalmente, te pasaremos el archivo .csv por separado, por si queres analizar los datos por tu propia cuenta.

Pregunta 1

En esta primera pregunta, te pedimos ejecutar la primera celda en el notebook para configurar el entorno de SQL, Pandas, Python, y descargar todos los datos necesarios para la tarea. En particular, se descargará el archivo datos_T3.db, y country_profile_variables.csv.

Luego, te pedimos cargar el archivo .csv a un dataframe de Pandas, y en base de este dataframe, crear un dataframe (o arreglo de numpy) que consiste solo de países ubicadas en Centro América, o el Caribe. Para esto, el archivo .csv tiene una columna denominada “Region”, y los valores que buscamos aquí serán “CentralAmerica” (sin espacios), o “Caribbean” (qualquiera de los dos). Cómo referencia, debes recuperar 33 países distintos en estas dos regiones.

Pregunta 2

En esta pregunta debes conseguir la información de cual persona participa en cual taller, y en cual país se organiza este taller. Quiere decir, que, en base de nuestra base de datos, debes armar un cursor que ejecuta la consulta recuperando esta información (ver el notebook para más hints). En particular, te pedimos armar una consulta SQL que recupera, desde nuestra base de datos, la información de:

- Rut de persona
- Id del taller
- Y país de taller

Para cada inscripción registrada.

Luego debes cargar esta tabla/resultado del cursor en un dataframe de Pandas, y renombrar las columnas para que se llamen “rut”, “idTaller”, y “Pais”.

Pregunta 3

En esta pregunta debes conseguir número de inscripciones a talleres en cada país recuperado en Pregunta 1. Quiere decir, para cada país de Centro América y del Caribe, debes calcular cuantas inscripciones hay en talleres que se organizan en este país. Nótense que aquí si una misma persona está inscrita en dos talleres distintos, esto cuenta cómo dos inscripciones (estamos calculando la participación, y no analizando el número de personas distintas que se inscriben). ***Si en un país no se organiza ningún taller, este no debe aparecer en tu resultado.***

Para esto debes ocupar el dataframe de la Pregunta 1, junto con el dataframe de la Pregunta 2, y hacerles un merge (mismo cómo join en SQL). Después, ocupando la agregación de Pandas, debes contar cuantas inscripciones hay en talleres de cada país de Centro América y Caribe.

En la clase presencial explicaremos cómo hacer joins con Pandas. Para ayudarte, igual incluimos una pequeña descripción al final de este enunciado (después de la última pregunta). Adicionalmente, puedes buscar “cómo hacer joins con Pandas” en Google para ver varios tutoriales sobre el tema. Si prefieres, puedes ocupar los comandos de Python para esto igual (dado que Jupyter Notebook corre Python).

Pregunta 4

En esta pregunta te pedimos armar un gráfico de barras (bar chart) de los datos obtenidos en la Pregunta 3. Quiere decir, tu gráfico debe tener en un eje el nombre del país, y en el otro, el número de inscripciones a talleres ene este país. De nuevo, esto es solo para los

países de Centro América y del Caribe, y no es necesario graficar los países dónde no se organiza ningún taller.

Pregunta 5

Ahora queremos ver si hay alguna conexión entre indicadores de un país con la cantidad de inscripciones a talleres que se organizan en este país. Para esto, debes conseguir, desde nuestro archivo .csv, o uno de los dataframes que armaste antes, la población del país (en el archivo .csv esta columna se llama “Population in thousands (2017)”), y ver, para cada país de Centro América y del Caribe, si la población más grande significa más inscripciones. ¿Qué puedes concluir aquí? ¿Hay alguna conexión? Con tu respuesta, también debes entregar el código qué genera esta información (de nuevo, para países de Centro América y Caribe dónde hay talleres).

Pregunta 6

Aquí debes hacer el mismo ejercicio de la Pregunta 5, pero ahora ver si hay alguna conexión entre el ÏB del áis y el número de inscripciones. Para esto te sirve la columna “GDP per capita (current US$)” del archivo .csv. ¿Qué puedes concluir aquí? ¿Hay alguna conexión? Con tu respuesta, también debes entregar el código qué genera esta información (de nuevo, para países de Centro América y Caribe dónde hay talleres).

Cómo hago join (merge) con dataframes de Pandas

Cuando tenemos dos dataframes de Pandas, los podemos pensar cómo dos tablas en el modelo relacional. Por lo tanto, muchas veces nos gustaría hacer el join entre estas dos tablas. Pandas tiene varios métodos que nos permiten hacer esto. Nosotros nos enfocaremos en el método **merge**.

El método merge de Pandas tiene muchas funcionalidades, pero nosotros nos enfocaremos en las más básicas.

Definamos dos dataframes:

Import pandas as pd

```
df1 = pd.DataFrame({'A': ['b', 'b', 'a'],
                    'B': [ 2 ,  7 ,  4 ]})
df2 = pd.DataFrame({'C': ['a', 'b', 'd', 'a'],
                    'D': [ 6 ,  7 ,  2 ,  4 ]})


pd.merge(df1, df2, left_on='B', right_on='D')
```

Pensando en df1 cómo la tabla con el esquema df1(A,B) y df2 cómo la tabla con el esquema df2(C,D), el comando merge de arriba realiza el join de df1 con el df2, con la condición que los valores de los atributos B y D sean iguales. En nuestro ejemplo, esto corresponde a una consulta SQL del estilo

```
SELECT *
FROM df1, df2
WHERE df1.B = df2.D
```

Aquí el parámetro ‘left_on’ dice cual atributo de la primera relación (df1 n nuestro caso) se ocupará para comparar con ‘right_on’ de df2 (la segunda relación). En nuestro ejemplo, el resultado de este merge/join será el siguiente dataframe:

	A	B	C	D
0	b	2	d	2
1	b	7	b	7
2	a	4	a	4

El método merge permite muchas funcionalidades y formas de hacer joins. Por ejemplo, si los dataframes tienen atributos con el mismo nombre, el join se hace sobre estos atributos. Para más información, puedes consultar:

- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html
- https://github.com/IIC2413/Syllabus-2020-2/blob/master/Notebooks/DataScience/DataFrames_Pandas.ipynb

Aspectos formales

Considera los aspectos formales que se describen a continuación:

- Entrega un archivo .ipynb con las celdas ejecutadas

¡Mucho éxito!

Importante: la fecha de entrega está indicada en el calendario del curso. Cuidar la redacción y la ortografía. Si tienes alguna duda sobre los contenidos o sobre cómo realizar esta actividad, puedes utilizar la herramienta "Mensajes" y enviar tu pregunta. Recibirás la respuesta de su tutor con las orientaciones correspondientes