

Tarea de ETL con Python – Procesamiento Analítico de Datos

Se recibe un listado de alumnos en el archivo alumnos.xlsx, y se necesita obtener la siguiente información:

- Cantidad de alumnos matriculados y no matriculados por carrera
- Cantidad de alumnos con gratuidad por sexo
- Cantidad de alumnos por tipo de ingreso en cada facultad
- Cantidad de alumnos titulados según el año de ingreso para alumnos sin gratuidad.
- Carrera que tiene más cantidad de mujeres para años de ingresos posteriores al 2010.

Se tienen las siguientes situaciones:

- Existen algunos ruts faltantes en la planilla indicada arriba, pero por suerte se tiene una base de datos un poco más antigua con información que podría ayudar, esta se llama base_antigua.xlsx. Se puede hacer un cruce a través del número de matrícula.
- Existen algunos registros duplicados que se deben limpiar.
- Se entregan aparte las bases con el listado de carreras y listado de facultades en un archivo llamado facultades_y_carreras.xlsx.
- Existen alumnos que tienen información faltante para el sexo. Para aquellos casos debe etiquetarlos como “No definido”.

Nota:

- Considere que usted no puede hacer ningún cambio directo a los archivos a excel ya que estarán en una ruta de solo lectura en disco compartido que utilizan decenas de usuarios más.

Construya un ETL usando Python en conjunto con la librería Pandas y Matplotlib. Debe responder los elementos indicados en el primer párrafo. Puede utilizar pivot_table de Pandas.

Plazo máximo de entrega: 18 de enero a las 18:00 hrs.

Medio de entrega: Plataforma Moodle